TUM School of Computation, Information and Technology

Technische Universität München

Dissertation

Towards precision medicine in inflammatory skin diseases: Applying machine learning and statistics on spatial, single-cell, and bulk transcriptomics data

Christina Katharina Luise Hillig

March 2025

Technische Universität München
TUM School of Computation, Information and Technology

Towards precision medicine in inflammatory skin diseases: Applying machine learning and statistics on spatial, single-cell, and bulk transcriptomics data

Christina Katharina Luise Hillig

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Julia Schnabel

Prüfende der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Prof. Dr. Markus List

Die Dissertation wurde am 11.03.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 09.07.2025 angenommen.

Acknowledgements

I would like to express my gratitude to everyone who supported me throughout the years leading up to and during my doctoral research. First, I would like to thank my supervisor, Michael Menden, for his invaluable guidance throughout my PhD, for introducing me to numerous fascinating projects, and for creating an awesome group of kind and supportive people. A huge thanks to the MendenLab members Martin Meinel, Ines Assum, Ginte Kutkaite, Diyuan Lu, Daniel Garger, Maria Bordukova, Nikita Makarov, Phong Nguyen, Clara Meijs, Alexander Ohnmacht, Ali Farnoud, Ana Galhoz, Göksu Avar, Alina Arneth, and Anantharamanan Rajamani. I enjoyed our trips to Budapest, Stans, beer garden, and Christmas markets. I am grateful to have had the opportunity to work with such wonderful people like you. Thank you, Elisabeth Noheimer, for taking care of all the administrative work and being always supportive.

In addition, I want to thank our collaboration partners Stefanie Eyerich, Kilian Eyerich, Heydar Maboudi Afkham, Alexander Schäbitz, Natalie Garzorz-Stark, Peter Seiringer, Caroline Pilz, Menna Mubarak, Manja Jargosch, Jenny Thomas, Emmanuelle Scala, and Kunal Das Mahapatra. I have truly enjoyed working with you on these exciting and impactful projects. Thank you, Kilian Eyerich and IGSSE, for giving me this opportunity to go on a research exchange to the Karolinska Institute in Stockholm, Sweden.

I would also like to express my gratitude to my students Jigyansa Mishra, Iris Lade, Gülce Işıl Gökçe, Barbara Streibl, Niklas Lochmann, Deniz Onur Üreyener, Eva Briem, Anna Maguza, Tamina Dietl, and Serhat Beyaz. It has been a pleasure guiding you through your internships and theses, and I also want to thank you for sharing your knowledge about biology.

I also would like to express my gratitude to Fabian Theis for creating an environment conducive to growth and exchange. Furthermore, I would like to express my gratitude to Carlos Talavera-Lopez and Malte Lücken for the insightful discussions in my first PhD year.

In addition, a special thank you to Max for always being by my side. Your loving, supportive, and understanding nature has helped me through every phase of my doctoral journey. I would also like to thank your parents, Birgit and Oliver, for always supporting us. Finally, I would like to express my deep gratitude to my family. Your continuous belief in me and constant support have been invaluable throughout this journey. Thank you so much, Christine, Thomas, Simone, Luise, Peter, Kurt and Betty.

Abstract

Non-communicable chronic inflammatory skin diseases (ncISDs) present significant challenges in terms of diagnosis and treatment due to their complex underlying mechanisms and overlapping clinical presentations. Precision medicine has the potential to improve patient outcomes, as conventional diagnostics based on clinical phenotypes often lack specificity and are insufficient for reliably predicting therapy response. While emerging strategies have categorised diseases by immune response patterns and employed biomarkers to predict diagnoses, these methods often remain limited in granularity and specificity, challenging their ability to accurately predict drug response and refine patient stratification. To address these challenges, I present a hypothesis-free patient stratification approach using bulk RNA-sequencing data from a diverse set of ncISDs, linking specific groups to drug targets. I further demonstrate the potential of spatial transcriptomics and single-cell RNA sequencing to enhance the understanding of disease mechanisms. These findings support the development of tailored treatment options, advancing precision medicine in ncISDs.

In this thesis, the hypothesis-free patient stratification framework, including an automated gene selection pipeline, stratified 23 ncISDs into 13 endotypes. Their relevance was confirmed by the association of clinical phenotypes and biological processes, with metabolism and inflammation identified as primary differentiating features. Moreover, I hypothesised an association between specific endotype groups and drug response, utilising data from 34 psoriasis patients treated with IL - 23, IL - 17, or TNF - α inhibitors. The association exhibited certain trends that require further validation in larger cohorts. Using my feature selection method, GeneSTRIVE, I identified potential biomarkers and created gene-expression-based classifiers, which were validated using an independent cohort. To further explore disease mechanisms, I used spatial and single-cell transcriptomics data of the most prevalent ncISDs, focusing on the expression patterns of disease-driving cytokines within the skin. I also developed a density-based clustering algorithm, revealing that small amounts of these cytokine transcripts induce thousands of specific immune response transcripts in their vicinity, thereby initiating an inflammatory amplification cascade.

This thesis improves the understanding of ncISDs by leveraging bulk, single-cell, and spatial transcriptomics data and introducing new computational methodologies. Moreover, it establishes a foundation for precision medicine by identifying endotypes and potential biomarkers.

Zusammenfassung

Nicht-übertragbare, chronisch-entzündliche Hauterkrankungen (ncISDs) stellen aufgrund des begrenzten Verständnisses ihrer Mechanismen und der überschneidenden Krankheitsbilder eine Herausforderung in Diagnose und Behandlung dar. Die Präzisionsmedizin hat das Potenzial, die Behandlungsergebnisse zu verbessern, da die konventionelle Diagnostik auf klinischen Phänotypen oft nicht spezifisch genug ist, um das Therapieansprechen zuverlässig vorherzusagen. Aktuelle diagnostische Ansätze streben eine weitergehende Einteilung der Erkrankungen an, indem sie Immunantwortprofile und Biomarker verwenden. Diese wurden jedoch auf Grundlage klinischer Diagnosen identifiziert, wodurch die komplexen biologischen Mechanismen der ncISDs nur unzureichend abgebildet werden. Zur Lösung dieser Herausforderungen präsentiere ich einen hypothesefreien Ansatz zur Patienteneinteilung auf Basis von Bulk RNA-Sequenzierungsdaten zahlreicher ncISDs. Zudem zeige ich das Potenzial von räumlichen und Einzelzell-Transkriptomdaten, um die Krankheitsmechanismen besser zu verstehen. Diese Ergebnisse tragen zur Entwicklung verbesserter Behandlungsansätze und zur Förderung der Präzisionsmedizin in ncISDs bei.

Ich verwendete einen hypothesefreien Stratifizierungsansatz mit einer automatisierten Gen-Selektions-Pipeline, um 23 ncISDs in 13 Endotypen zu gruppieren. Ihre Relevanz wurde durch die Assoziation von klinischen Phänotypen und biologischen Prozessen bestätigt, wobei Metabolismus und Entzündung als Hauptunterscheidungsmerkmale identifiziert wurden. Zudem stellte ich die Hypothese einer Assoziation zwischen spezifischen Endotyp-Gruppen und dem Therapieansprechen auf. Diese basierte auf Daten von 34 Psoriasis Patienten, die mit IL - 23, IL - 17 oder TNF - α Inhibitoren behandelt wurden. Die Assoziation zeigte Trends, die einer weiteren Validierung in größeren Kohorten bedürfen. Mit meiner Gen-Selektions-Methode GeneSTRIVE identifizierte ich potenzielle Biomarker und erstellte Klassifikatoren, die an einer unabhängigen Kohorte validiert wurden. Zur weiteren Untersuchung der Krankheitsmechanismen, nutzte ich räumliche und Einzelzell-Transkriptomdaten der häufigsten ncISDs und analysierte die Expressionsmuster von krankheitsfördernden Zytokinen in der Haut. Zudem entwickelte ich einen dichte-basierten Clustering Algorithmus, der zeigte, dass geringe Mengen dieser Zytokin-Transkripte spezifische Immunantworten induzieren und eine entzündliche Amplifikationskaskade auslösen.

Diese Dissertation verbessert das Verständnis von ncISDs durch die Nutzung von Bulk-, Einzelzell- und räumlicher Transkriptomdaten und die Einführung neuer Methoden. Darüber hinaus schafft sie eine Grundlage für die Präzisionsmedizin durch die Identifizierung von Endotypen und potenziellen Biomarkern.

List of publications

The following projects are presented in this thesis.

- A. Schäbitz*, **C. Hillig***, M. Mubarak, M. Jargosch, A. Farnoud, E. Scala, N. Kurzen, A. C. Pilz, N. Bhalla, J. Thomas, M. Stahle, T. Biedermann, C. B. Schmidt-Weber, F. Theis, N. Garzorz-Stark, K. Eyerich*, M. P. Menden* & S. Eyerich*,† "Spatial transcriptomics landscape of lesions from non-communicable inflammatory skin diseases." *Nature Communications* **13.1** (2022): 7729. DOI: https://doi.org/10.1038/s41467-022-35319-w.

- N. Garzorz-Stark*, **C. Hillig***, P. Seiringer*, M. Meinel*, H. Maboudi Afkham, J. Mishra, M. Jargosch, A. Farnoud, S. Eyerich*, M. P. Menden*, K. Eyerich*,†. "Integrating phenomics and transcriptomics to identify clinically meaningful endotypes of non-communicable inflammatory skin diseases." (*In preparation*).

- **C. Hillig***, M. Meinel*, P. Seiringer*, N. Garzorz-Stark*, I. Harder, M. Hübenthal, N. Lochmann, J. Mishra, A. Farnoud, H. Maboudi Afkham, M. Jargosch, S. Weidinger, S. Eyerich*, K. Eyerich*, M. P. Menden*,†. "Minimal set of predictive biomarkers enable endotype classification for precision medicine in inflammatory skin diseases."(*In preparation*).

My doctoral research contributed to the following manuscripts, which are not the subject of this thesis. These are listed in reverse chronological order.

Peer-reviewed

- M. Jargosch*, J. Kuruvila*, E. Scala, J. Grosch, J. Eigemann, S. Wasserer, S. Lekiashvili, N. Trautwein, D. J. Kowalewski, A. Böhner, Y. Köseoglu, **C. Hillig**, J. Thomas, F. Lauffer, C. B. Schmidt-Weber, M. P. Menden, J. S. Walz, S. Kaesler, S. Eyerich, S. Blank, H. Rammensee, T. Biedermann, K. Eyerich, Z. Kurgyis*, L. K. Freudenmann*, N. Garzorz-Stark*,†. "Immunopeptidome analysis reveals SERPINB3 as an autoantigen driving eczematized psoriasis." *Science Advances* (2025). DOI: 10.1126/sciadv.adx0637.

- S. Schärli, F. Luther, J. Di Domizio, **C. Hillig**, S. Radonjic-Hoesli, K. Thormann, D. Simon, A. Thorsti Møller Rønnstad, I. Frier Ruge, B. G. Fritz, T. Bjarnsholt, A.

*Contributed equally

†Corresponding author

Vallone, S. Kezic, M. P. Menden, L. M. Roesner, T. Werfel, J. P. Thyssen, S. Eyerich, M. Gilliet, N. L. Bertschi, C. Schlapbach†. "IL-9 sensitizes human Th2 cells to proinflammatory IL-18 signals in atopic dermatitis." *Journal of Allergy and Clinical Immunology* (2025). DOI: https://doi.org/10.1016/j.jaci.2024.10.027.

- N. Kurzen*, M. Mubarak*, M. Jargosch, J. Eigemann, P. Seiringer, **C. Hillig**, M. P. Menden, T. Biedermann, C. B. Schmidt-Weber, K. Eyerich, S. Eyerich, F. Lauffer† "Death associated protein kinase 1 dampens keratinocyte necroptosis and expression of inflammatory genes in lichen planus." *Journal of Investigative Dermatology* (2024). DOI: 10.1016/j.jid.2024.11.017.

- P. Seiringer*, **C. Hillig***, A. Schäbitz, M. Jargosch, C. Pilz, S. Eyerich, A. Szegedi, M. Sochorová, F. Gruber, C. C. Zouboulis, T. Biedermann, M. P. Menden*, K. Eyerich*, D. Torocsik*,†. "Spatial transcriptomics reveals altered lipid metabolism and inflammation-related gene expression of sebaceous glands in psoriasis and atopic dermatitis." *Frontiers in Immunology* 15 (2024): 1334844. DOI: https://doi.org/10.3389/fimmu.2024.1334844.

- J. Ding, **C. Hillig**, C. W. White, N. A. Fernandopulle, H. Anderton, J. S. Kern, M. P. Menden, G. A. Mackay†. "CXCL17 induces activation of human mast cells via MRGPRX2." *Allergy* (2024). DOI: 10.1111/all.16077

- F. Fischer*, A. Doll*, D. Uereyener, S. Roenneberg, **C. Hillig**, L. Weber, V. Hackert, M. Meinel, A. Farnoud, P. Seiringer, J. Thomas, P. Anand, L. Graner, F. Schlenker, R. Zengerle, P. Jonsson, M. Jargosch, F. J. Theis, C. B. Schmidt-Weber, T. Biedermann, M. Howell, K. Reich, K. Eyerich, M. P. Menden, N. Garzorz-Stark, F. Lauffer*, S. Eyerich*,† "Gene expression based molecular test as diagnostic aid for the differential diagnosis of psoriasis and eczema in formalin fixed and paraffin embedded tissue, microbiopsies and tape strips." *Journal of Investigative Dermatology* 143.8 (2023): 1461-1469. DOI: https://doi.org/10.1016/j.jid.2023.02.015.

- D. Garger*, M. Meinel*, T. Dietl, **C. Hillig**, N. Garzorz-Stark, K. Eyerich, M. Hrabě de Angelis, S. Eyerich, M. P. Menden†. "The impact of the cardiovascular component and somatic mutations on ageing." *Aging Cell* 22.10 (2023): e13957. DOI: 10.1111/acel.13957.

*Contributed equally

Contents

1 Introduction **1**

1.1 Skin anatomy and its role 2

1.1.1 Skin structure and layers 3

1.1.2 Role as immune organ 6

1.2 Non-communicable, chronic inflammatory skin diseases 11

1.2.1 Clinical categorisation of skin diseases 12

1.2.2 Stratified medicine by immune response patterns 16

1.3 Transcriptomics 19

1.4 Advancements in non-communicable, chronic inflammatory skin diseases by transcriptomics 21

1.4.1 Diagnostic opportunities through transcriptomics and computational approaches 21

1.4.2 Skin disease models 23

1.4.3 Opportunities of single-cell RNA-sequencing and spatial transcriptomics 23

1.5 Challenges in non-communicable, chronic inflammatory skin diseases 25

1.6 Aims and research questions 27

1.7 Outline 28

2 Statistical background **31**

2.1 Sampling methods 31

2.2 Statistical testing 32

2.2.1 Non-parametric tests 35

2.2.2 Multiple hypothesis testing 38

2.2.3 Bayesian correlated t-test 39

2.3 Correlation and multicollinearity 41

2.3.1 Pearson's correlation coefficient 41

2.3.2 Spearman's correlation coefficient 42

2.3.3 Handling multicollinearity with variance inflation factor 43

2.4 Machine Learning 44

2.4.1 Dimension reduction techniques for visualisation 45

2.4.2 Unsupervised clustering 47

2.4.3 Supervised regression and classification 53

2.5 High-throughput sequencing techniques 72

2.5.1 Bulk RNA-sequencing 72
2.5.2 Single-cell RNA-sequencing 74
2.5.3 Spatial Transcriptomics 79
2.6 Exploratory data analysis 81
2.6.1 Differential gene expression analysis 81
2.6.2 Pathway enrichment analysis 84

3 Endotypes as alternative to the established disease ontology 87
3.1 Materials 89
3.2 Methods 90
3.2.1 Bulk RNA-sequencing data 90
3.2.2 Clinical attributes 93
3.3 Results 95
3.3.1 Diagnoses result in heterogeneous clinical and transcriptomics profiles 96
3.3.2 Framework to identify highly variable genes and endotypes 99
3.3.3 Endotypes capture heterogeneity of skin conditions 107
3.3.4 Biological interpretation of endotypes 110
3.3.5 Endotypes correlate with clinical hallmarks 120
3.4 Summary and discussion 123

4 Gene-expression-based psoriasis-like endotype classifiers 133
4.1 Materials 135
4.1.1 Therapy response cohort 135
4.1.2 Independent psoriasis cohort 136
4.2 Methods 137
4.2.1 Therapy response data 137
4.2.2 Psoriasis bulk RNA-sequencing cohort 138
4.2.3 Endotype classifiers 139
4.2.4 Independent psoriasis bulk RNA-sequencing cohort 141
4.2.5 Hypothesis evaluation for each drug target 142
4.3 Results 143
4.3.1 Trend of response patterns in psoriasis-like endotypes 143
4.3.2 Biological functions of hypothesised drug-target associated endotype groups 145
4.3.3 Advanced gene selection method for gene expression data 146
4.3.4 Triad feature selection pipeline including GeneSTRIVE 151
4.3.5 Gene-expression-based classifiers predicting groups of endotypes . . . 154

4.3.6 Application of classifiers on unseen endotypes 163
4.3.7 Prediction of endotypes in an independent cohort 166
4.3.8 Assessment of the endotype-therapy-response association 167
4.4 Summary and discussion 171

5 Spatial transcriptomics landscape of lesions of non-communicable, inflammatory skin diseases 177
5.1 Materials 179
5.2 Methods 179
5.2.1 Spatial transcriptomics dataset 179
5.2.2 Preparation of the single-cell RNA-sequencing data 183
5.2.3 Preprocessing of a public scRNA-seq dataset 185
5.2.4 Bulk RNA-sequencing data 186
5.2.5 Additional validation techniques 186
5.3 Results 188
5.3.1 Low expression of disease-promoting cytokines in lesion skin 191
5.3.2 Cytokines show up-regulation of immune response regulators 198
5.3.3 Cytokine expression is spatially correlated with immune response regulators 203
5.4 Summary and discussion 212

6 Discussion and Outlook 217

7 Conclusion 227

A Further statistical tests 229
A.1 Parametric tests 229
A.2 Non-parametric tests: Handling ties in ranks 230

B Further machine learning algorithms and metrics 231
B.1 Proof of relation between R_2 and r_p 231
B.2 Linear regression models 232
B.3 Model evaluation metrics 234

C Supplementary figures and tables 237
C.1 Molecular classifiers of specific groups of endotypes 237
C.2 Spatial transcriptomics landscape of skin lesions 242

Acronyms 245

List of Figures 249

List of Tables 251

Bibliography 253

Chapter 1

Introduction

Almost one fourth of the world's population suffers from severe skin conditions that extend beyond physical discomfort [Uji+22]. These non-communicable chronic inflammatory skin diseases (ncISDs) cause severe symptoms and can be accompanied by comorbidities such as depression, cardiovascular diseases, asthma, and allergies, significantly affecting quality of life and daily activities [Dal+15] [Boe+12] [Cve+06] [Gis+23] [Uji+22]. Current diagnostics and therapies, based on clinical pictures, amnesia and histological assessments, often often fail to account for disease heterogeneity, resulting in varied responses to treatment and increased patient suffering [Nai+21]. Thus, there is an urgent need for enhanced patient stratification and tailored treatments.

One promising approach to address these challenges in diagnosing and treating ncISDs is *precision medicine*, which pioneered in oncology and has transformed modern diagnostics [Men00] [Men13] [Lan21]. Precision medicine involves tailoring treatment strategies to individual patients based on comprehensive data collected from various sources, including genetic, environmental, and lifestyle factors [Cou+11] [MG16] [Kön+17]. This concept aims to provide effective treatments and minimise adverse effects, leading to improved patient outcomes. In this thesis, I propose strategies to realise precision medicine in ncISDs.

A related strategy to precision medicine is *drug repurposing*. It involves applying approved drugs to other diseases, thereby reducing costs and being also applicable to individual patients or to subpopulations with common characteristics or *biomarkers* [MG16] [Bon+21]. Drug repurposing can be particularly beneficial for individuals or subpopulations characterised by common traits or shared biomarkers, allowing for targeted interventions based on known efficacy. Associating specific patient groups with approved drugs thus further supports the implementation of precision medicine in ncISDs.

Biomarkers play a crucial role in precision medicine and drug repurposing, serving as diagnostic, prognostic, or predictive indicators. Diagnostic biomarkers confirm disease presence, prognostic biomarkers predict disease course, and predictive biomarkers forecast treatment responses [Lit19]. They are especially valuable for differentiating patients with overlapping entities [Kön+17]. Thus, biomarkers are essential in precision medicine to enable more nuanced patient stratification and optimise treatment.

Implementing precision medicine in ncISDs requires a deep understanding of underlying disease mechanisms, achievable through advanced data analysis, next generation sequencing (NGS), and artificial intelligence (AI)/machine learning (ML) [Nai+21]. Over the past few decades, increased computing power and storage capacity have led to the collection of data from a variety of sources, such as transcriptomics data and electronic health records. High-throughput NGS facilitates the study of transcriptomics, and ML enables to build customised models tailored to the underlying disease biology. Thus, advanced data tools and analytical approaches are essential to drive precision medicine in ncISDs.

One step towards precision medicine in ncISDs was accomplished by stratifying skin conditions based on their common immune response characteristics [EE18]. This approach is also known as *stratified medicine*, which refines patient group stratification and association of *drug targets* to each group [Kön+17]. However, this approach still relies on the subjective clinical phenotyping of patients and lacks robust biomarkers, granularity, and specificity. New approaches are needed, such as stratifying patients into *endotypes* by integrating clinical phenotypes and transcriptomics [AA+19]. They have already been identified in several diseases, including asthma [AA+19], tuberculosis [DiN+22], COVID-19 [Ran+21], and idiopathic pulmonary fibrosis [Kra+23b]. Studying endotypes can provide a more precise clinical and biological picture, enabling the identification of robust biomarkers and potentially advancing precision medicine in ncISDs.

This thesis aims to contribute to the advancement of precision medicine in ncISDs by presenting methods and biological insights that are gained through the integration of clinical phenotyping and genetic material. Groups with similar transcriptomics profiles (endotypes) are created and analysed. I use ML models to predict selected endotypes, which I hypothesise to be associated with drug response, thereby offering a glimpse of precision medicine. In addition, I provide new insights in the immune response of ncISDs using the spatial transcriptomics (ST) technology, Visium by 10x Genomics, to explore the landscape of ncISDs.

1.1 Skin anatomy and its role

The skin is the largest organ of the body and functions as a crucial barrier with complex biological mechanisms that are not yet fully understood. It serves as physical, chemical, and immunological barrier, protecting the body against UV light, loss of moisture, and regulates the body temperature [EE18]. As the outermost part of the human body, the skin is constantly challenged by environmental factors and thus, acts as the first line of defence. The skin itself is a complex system with a well understood anatomical structure, while the biological mechanisms remain to be elucidated. In ncISDs, the primary focus is on deciphering and characterising the skin's intricate biological functions and their role in disease manifestation.

The skin is composed of a variety of cell types, which are the smallest living units and fulfil distinct roles that support the skin's functions, including repair and immune defence. Cell types such as keratinocytes (KCs), Langerhans cells (LCs), melanocytes, fibroblasts, Dendritic cells (DCs), macrophages (Macs), and immune cells (Figure 1.1 a), collectively fulfil the roles of protectors, regulators and communicators, thereby maintaining skin health [YAS17]. Each cell type is characterised by a unique set of expressed genes and micro-environmental influences that define its phenotype. Both factors determine a cell's function, structure, and role within biological processes [Zen22]. In the context of ncISDs, studying the composition and interactions of these different cell types is essential to understanding how cell organisation, communication, and regulation contribute to disease pathogenesis. Comparing cell type composition across different skin conditions can provide insights into the mechanisms underlying various skin disorders.

1.1.1 Skin structure and layers

The skin is composed of three primary tissue layers, i.e. the epidermis, dermis, and hypodermis. Each layer performs specific functions essential for maintaining skin integrity, protection, and overall homeostasis (Figure 1.1 b). The following sections provide a detailed examination of the epidermis and dermis.

The epidermis, the outermost layer of the skin (Figure 1.1 b), serves as the primary protective barrier, formed and maintained through a well-regulated process of KC proliferation and differentiation. KCs are a type of cell that constitute the majority of the epidermis. The epidermis consists of four to five sublayers: stratum corneum, stratum lucidum (only in certain parts of the body), stratum granulosum, stratum spinosum, stratum basale [YAS17]. In this thesis, these layers are grouped as follows. The stratum corneum is referred to as *upper epidermis*, the stratum lucidum, stratum granulosum, stratum spinosum, are combined and referred to as *middle epidermis*. The deepest epidermal layer, the stratum basale, is referred to as the *basal epidermis*. Subsequent analyses in chapter 5 focus on complex mechanisms within these epidermal layers and the dermis relevant to the most common ncISDs.

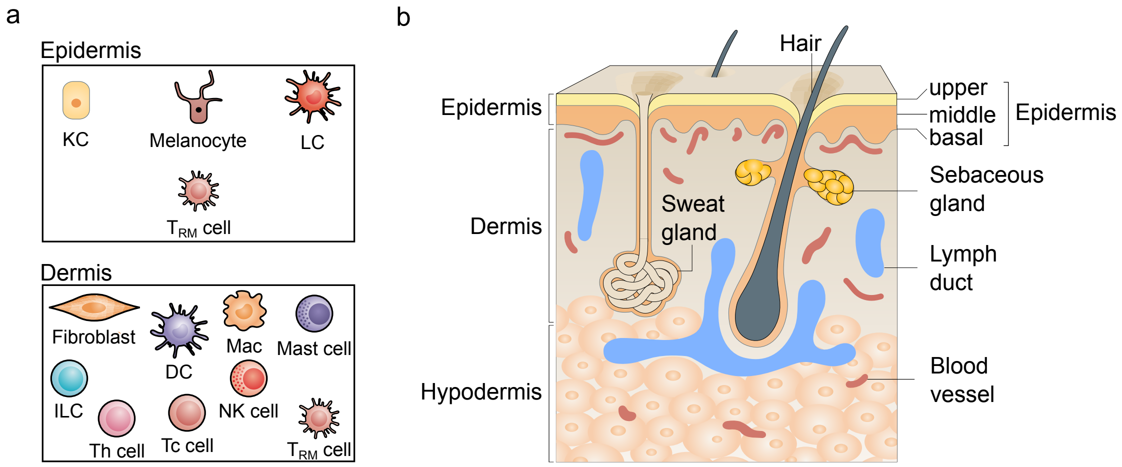


Figure 1.1: Structure and cell type composition of healthy human skin. a) Main cell types in the skin, including those in the epidermis (KCs, melanocytes, LCs, and resident memory T-cells (T_{RM}-cells)) and in the dermis (fibroblasts, DCs, Macs, mast cells, natural killer cells (NKs), innate lymphoid cells (ILCs), and T-cells such as T-helper (Th), cytotoxic T-cell (Tc cell), and T_{RM}-cells). Immune cells are T-cells, LCs, DCs, Macs, mast cells, NKs, and ILCs. **b)** Diagram showing the skin layers and associated structures. The skin is comprised of three main layers, i.e. epidermis, dermis, and hypodermis. The epidermis is further divided into upper, middle, and basal layers. The skin also contains other components such as hairs, sebaceous glands, sweat glands, lymph ducts, and blood vessels. Figure adapted from [Akh+22] and [Kab+19].

The upper epidermis is the outermost skin layer, acting as the first line of physical and immune defence against environmental factors and is the thickest layer of the epidermis. It is primarily composed of terminally differentiated KCs, termed corneocytes, which form the cornified envelope, a structure critical for maintaining skin integrity [CSM05]. This layer continually sheds dead cells, a process balanced by cell production in the basal layer to maintain skin homeostasis. This process is also known as keratinization (or cornification), a form of programmed cell death of the KCs [Gal+18]. Malfunctioning during keratinization can lead to skin disorders, such as ncISDs, marked by, e.g., hyperproliferation, parakeratosis, hyperkeratosis, and chronic wounds [Kom+22] [Sto+08]. As the skin's primary defence, the upper epidermis plays a key role in maintaining barrier integrity and can involve dysfunctions that manifest in ncISDs.

The middle epidermis comprises layers where KCs are in intermediate stages of differentiation. As KCs migrate towards the upper epidermis, they gradually lose their nuclei and acquire properties that prepare them for barrier formation. Additionally, this layer contains antigen presenting cells (APCs) that play a role in immune defence

(Section 1.1.2.1) [Mar+18]. The middle epidermis is a transitional layer crucial for preparing KCs for barrier function and initiating immune responses.

The basal epidermis is the deepest epidermal layer and is relatively thin compared to the upper and middle epidermis. Amongst its resident cell types are basal cells (precursors of KCs), melanocytes, Merkel cells, and T-cells [YAS17] [NS19]. Basal cells proliferate (multiply through cell division) and differentiate (specialise) into KCs thereby migrating upwards through the middle epidermis. Upon reaching the upper epidermis, these KCs eventually differentiate into corneocytes, thereby completing their lifecycle [CSM05]. The basal epidermis is a proliferative zone that replenishes the epidermis, forming the foundation for continual skin renewal.

The dermis, located beneath the epidermis, is a complex tissue layer with diverse cell types and functions (Figure 1.1). In comparison to the simply structured and densely packed epidermis, the dermis is more loosely organised and provides structural support and elasticity to the skin [AMP17] [NS19]. It is composed of various cell types such as DCs, Macs, mast cells, NKs, T-cells, ILCs, and fibroblasts [AMP17] [NS19]. The dermis is also characterised by an extracellular matrix (ECM), a network made of blood vessels and lymph ducts, enabling cell migration. The ECM is responsible for skin elasticity and strength, which is regulated by collagen formation produced by fibroblasts [Kab+19]. The dermis also contains skin appendage such as hair, muscles, and glands [HKS22] (Figure 1.1 b). They regulate the body temperature, skin dryness, and prevent sun damages [Kab+19]. In summary, the dermis provides structural integrity, supports immune functions, and contains appendages that contribute to essential physiological roles.

The hypodermis, the innermost layer, lies beneath the dermis and is composed of blood vessels, connective tissue and mainly adipose tissue (Figure 1.1 b). This layer prevents heat loss and connects the skin to the underlying fascia surrounding the muscles [YAS17] [NS19]. Together with the dermis, the hypodermis is essential for skin functioning, providing structure, elasticity, and interconnectivity to deeper tissues.

This thesis investigates the skin in two complementary ways. In Chapter 3 and 4, I analyse the skin as a whole, thereby focusing on general biological disease mechanisms. While in Chapter 5, I analyse the skin layers, focusing on spatial expression patterns of cells involved in the immune response. This approach of analysing general and spatial aspects of the skin supports a comprehensive understanding of the pathophysiology of ncISD.

1.1.2 Role as immune organ

The immune system is divided into two lines of defence: the innate and adaptive immunity (Figure 1.2). The former is activated first and provides a rapid, unspecific response to internal or external trigger mechanisms, while the latter is a secondary targeted response activated by the innate immune system. It is a slow but highly specific response to the threat, with the ability to memorise and thus provide a more effective immune response against the recurring threat [Mar+18]. Together, the innate and adaptive systems create a robust defence against diverse threats, with innate immunity serving as the rapid initial response and adaptive immunity providing targeted, long-term protection.

The cutaneous immune system belongs to both innate and adaptive immunity, acting as the skin's specific immune defence. It serves as the body's first line of defence against external influences and threats, such as pathogens, through a combination of chemical, physical, and immune barriers composed of both immune and non-immune cells [NS19]. Dysfunction of the immune system can cause inflammatory disorders, autoimmune diseases, immunodeficiency disorders, and hypersensitivity reactions [Mar+18]. In essence, the cutaneous immune system is essential for host defence and disease prevention.

The upper epidermis has specific roles and components to protect the skin. As a physical and chemical barrier, it consists primarily of corneocytes and biomolecules Biomolecules, such as antimicrobial peptides (AMPs) and lipids, are secreted by e.g., KCs, DCs, fibroblasts, and skin appendages (e.g., sebaceous glands) on the skin surface [NS19]. These biomolecules, together with corneocytes, prevent the penetration of foreign substances, help retain moisture, and unicellular organisms to settle down and grow uncontrollably on the skin [NS19]. Additionally, AMPs have immunological functions such as regulating the immune response and inflammation [Zha+22a] [DMSC23] [CA16]. This indicates that the upper epidermis also belongs to the immune barrier, which involves the entire skin, i.e. the epidermis and dermis [Zha+22a]. Thus, the upper epidermis serves as protective barrier by maintaining the skin's integrity and defending against external threats.

Immune cells are part of the immune barrier and originate from hematopoietic stem cells in the bone marrow (Figure 1.2) [She+23]. These stem cells differentiate into red and white blood cells, which have immunological functions. White blood cells or leukocytes are found in the blood and lymph nodes. In contrast to the red blood cells, white blood cells are precursors of immune cells [She+23]. White blood cells differentiate into functional immune cells such as Granulocytes (e.g., Neutrophils), Monocytes (DCs, Macs), and Lymphocytes

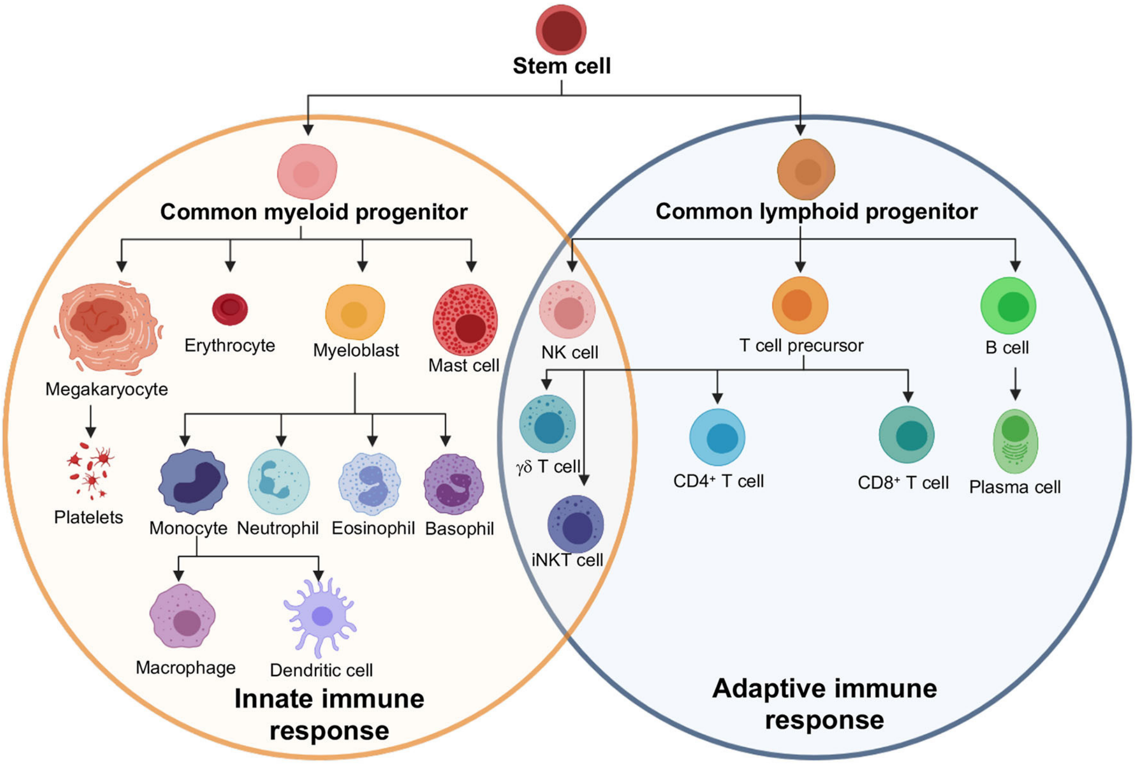


Figure 1.2: Lineage tree of immune cells. The hematopoietic stem cells differentiate into either myeloid or lymphoid progenitors, which are types of white blood cells. Upon activation, white blood cells or leukocytes differentiate into specific immune cells, which belong to the innate or adaptive immune response. Innate immune cells become first activated and assist in activating the cells of the adaptive immune response by, e.g., presenting specific antigens. While cells of the adaptive immunity provide a more specific response. Figure adapted from [Mur20].

(T-cells, B-cells, NKs) (Figure 1.2) [She+23] [Mur20]. Each of them fulfils a specific function in the immune response. Summarising, differentiated hematocytes, KCs, and fibroblasts are all components of the innate and/or adaptive immune system, jointly acting against external threats in the skin.

1.1.2.1 Cells of the innate immune response

The cells of the innate immune system are important in the skin's initial defence mechanisms and regulation of immune responses. The skin consists out of a variety of innate and adaptive immune cells [Zha+22a]. Innate immune cells, together with KCs, build the first layer of defence. Identifying and characterising these cell types in single-cell RNA-sequencing (scRNA-seq) and ST analyses allows for a deeper understanding of cell-type-specific gene expression patterns, cellular interactions and spatial organisation within the skin. Knowing cells involved in the immune system is essential for studying their contributions to inflammation and immune response in ncISDs.

KCs are the most abundant cell type in the epidermis, being key players in wound healing and immune response [NL21]. In response to an external trigger (e.g., injury or pathogens), they initiate inflammation cascades by accumulating neutrophils and activating skin resident T-cells [Orl+20]. They achieve this by expressing and secreting messenger molecules such as pro-inflammatory cytokines (e.g., IL - 36, IL - 1β, IL - 8, TNF - α), chemokines (e.g., CXCL8, CXCL10), and AMPs (e.g., S100 family) thus enabling them to interact with immune cells of the innate and adaptive immune system [CVS20] [Zha+22a] [Orl+20]. The activated immune cells produce cytokines, which in turn act on the KCs, thereby inducing a vicious cycle of inflammation in ncISDs [SEG21]. Moreover, precursors cells of KCs have been found to acquire an inflammatory memory during inflammation allowing for an enhanced immune response in the future [Nai+17] [Lar+21]. Thus, investigating the influence of KCs in skin disorders can reveal potential therapeutic targets to mitigate hyperinflammatory cycles.

LCs, epidermal-resident Macs with a DC-like phenotype, belong to the family of APCs, playing a pivotal role in the interaction between innate and adaptive immunity [PB06] [Kab+19] [Mar+18]. They continuously scan the surface of the skin for antigens. Upon the detection of an antigen, LCs present it to epidermal resident T-cells, thereby initiating a local immune response [CVS20] [Kab+19]. They can also migrate to the lymph nodes, triggering an immune cascade [PB06]. As primary APCs in the epidermis, understanding LCs expression profiles under various conditions can reveal how they contribute to persistent inflammation or immune tolerance.

The dermis hosts diverse structural and immune cells essential for skin defence and immune regulation. Fibroblasts detect pathogens and produce regulatory cytokines (e.g., IL - 10) that suppress T-cell proliferation [CVS20]. DCs and macrophages, such as LCs, function as APCs in the dermis, with macrophages also playing healing and anti-inflammatory roles [CVS20]. ILCs become more abundant in inflammation and differentiate based on immune response type, secreting cytokines such as IFN - γ, IL - 4, and IL - 17A [Mar+18] [Zha+22a]. Mast cells contribute to homeostatic immunity and are linked to atopic reactions, itching, and rashes [GT10]. Thus, dermis-resident cells are vital for pathogen detection, immune modulation, and homeostasis.

This thesis analyses the transcriptional profiles of these diverse innate immune cells using scRNA-seq and ST, enabling to identify the spatial relationships and molecular signatures unique to a cell type within the skin. This approach is important for investigating the cellular mechanisms underlying immune regulation and dysfunction, contributing to a more comprehensive understanding of ncISDs and providing insights for more targeted therapies.

1.1.2.2 Cells of the adaptive immune response

The innate immune system provides a rapid but unspecific defence, which may be insufficient against some pathogens. In contrast, the adaptive immune system learns, adapts, and remembers antigens, enabling a more targeted response [Mar+18]. Its activation is triggered by signals from the innate immune system, such as cytokines, cell-cell interactions, and antigen presentation, leading to a specific immune response by T-cells and B-cells [Mar+18].

In the lymph node, naïve T-cells become activated by an APC and begin to proliferate. Depending on the presented antigen by the APC and surrounding cytokine composition, these naïve T-cells differentiate into specific effector T-cells, i.e. Tc cells (CD8+ cells) or Th cells (CD4+ cells) (Figure 1.2) [Mar+18]. These circulate throughout the body and infiltrate various tissues, including the skin, where they perform their effector function upon encountering their antigen. Tc cells eliminate target cells, such as foreign or cancer cells, while Th cells coordinate the immune response. Both Th cells and Tc cells are integral to immune responses in ncISDs and are further investigated in Chapter 5.

Six main types of Th cell subsets have been identified to date: Th1, Th2, Th17, T regulatory cell (Treg), follicular Th cells, and Th22 [ZZ20] [Eye+09]. These subsets including both Tc cells and Th cells contribute to distinct types of immune responses. For instance, Tc/Th1, Tc/Th2, and Tc/Th17 cells participate in type 1, type 2 and type 3 immune response, respectively. They are characterised by their cytokine expression. Type 1 is associated with the expression of IFN - γ, type 2 with IL - 4/IL - 13, and type 3 with IL - 17/IL - 22 [Mar+18] [ARR15]. The type 1 to type 3 T-cell subsets are associated with specific ncISDs [EE18] [Mar+18] (Section 1.2.2) and will be further examined in Chapter 5.

In general, Th cells have many functions. For instance, during an immune response, Th1 cells trigger the ability of macrophages to fight bacteria and boost immunity, while Th2 cells activate mast cells. Th17 cells are, for example, responsible for combating external pathogens and regulating inflammatory responses [Mar+18]. In addition, defect Th1, Th2, and Th17 responses are associated with distinct immune disorders such as autoimmune diseases, atopic conditions, and chronic disorders, respectively [Mar+18].

After neutralising the threat, the majority of T-cells undergo apoptosis, with only a few remaining in the skin. Those serve as $\mathrm{T_{RM}}$ - cells and recall immune responses to the specific antigen, thereby enabling a specific and rapid immune reaction upon re-exposure to the same antigen [NS19]. By maintaining this immunological memory, $\mathrm{T_{RM}}$ - cells are essential in eliminating invaders, coordinating the immune response, and long-term immunity. Dysfunctions in these processes are associated with various immune disorders.

B-cells (Figure 1.2), part of the adaptive immune system, produce antibodies to target specific antigens. In contrast to T-cells, B-cells can directly capture antigens via B-cell receptors on their surface, leading to their activation [Mar+18]. B-cells also act as APCs, presenting antigens to T-cells. Cytokines from Th cells promote B-cell proliferation and differentiation into plasma cells, which produce antibodies, or memory B-cells, which provide rapid protection against future infections [IK24]. B-cells are involved in ncISDs such as atopic eczema, where they are found in the dermis [Sim+08] [NS19]. Thus, B-cells contribute to antibody production, long-term immunity, and immune defence.

Both, T-cells and B-cells, are essential to adaptive immunity. Antigen-specific T-cells assist in either eliminating (Tc cell cells) invaders or coordinating (Th cells) the immune response. B-cells are initially activated by capturing antigens with their B-cell receptor and are further stimulated through cytokine signalling of Th cells. This leads to the production of antibodies, which are crucial for the immune defence. B-cells also present these antigens to T-cells, resulting in their activation and an antigen specific immune response. Defects T-cells or B-cells functions are associated with immune disorders. Moreover, both cell types can form memory cells, allowing for rapid responses to previously encountered antigens. In summary, the adaptive immunity is complementary to the innate immune system as it recognises specific antigens, thereby providing a targeted immune response.

1.1.2.3 General processes during innate and adaptive immune response

The immune response in the skin is initiated by an internal or external trigger, such as skin injury, leading to potential infiltration of pathogens. These pathogens present specific antigens on their surface that initiate a cascade of biological events (pathways). KCs in the skin respond by expressing and secreting messenger molecules that activate the immune system. Epidermal APCs of the innate immune system detect, process, and present the antigens on their surface, thereby informing and activating T_{RM} - cell cells and naïve T-cells that of the adaptive immune system. Once activated, T-cells differentiate into CD8+ Tc cells, which target the specific threat, and CD4+ Th cells, which organise the immune response. Th cells further activate B-cells and, along with Tregs, ILCs, and mast cells, they coordinate and regulate the immune response.

The innate immune system provides a rapid, unspecific response while activating the slower, highly specific adaptive immune system via cytokines or cell interactions. Adaptive immunity regulates response strength through cytokine signalling, activating other immune cells. This interplay is essential for skin homeostasis, as immune cell function depends on spatial distribution [Mar+18]. Dysregulation of either system can lead to autoimmunity, chronic inflammation, or immunodeficiency [HK19] [Mar+18]. The following section examines ncISDs in more detail.

1.2 Non-communicable, chronic inflammatory skin diseases

The ncISDs are categorised into chronic inflammatory, autoimmune, autoinflammatory, and rheumatic diseases based on their pathogenesis [Uji+22] [EE18] [CBF12]. These conditions are often accompanied by comorbidities such as arthritis, asthma, and metabolic syndrome [Das+21]. The diagnosis of ncISDs usually relies on clinical phenotypes which include morphological, comorbidities, histological, and immunological observations of skin lesions [Gri+21] [Uji+22]. The morphology describes the appearance of skin lesions and histology quantifies the architecture of the skin, e.g., by correlating structure and function of the epidermis using microscopic images. Immunological observations involve the immune cell composition and active pathways [SGSE22]. ncISDs are complex disorders that require a multifaceted diagnostic approach.

The ncISDs can be triggered by various factors, such as genetic predisposition, smoking, obesity, and other disorders [SGSE22] [Uji+22], leading to variable symptom severity. Severity is assessed using metrics, such as Psoriasis Area and Severity Index (PASI), SCORing of Atopic Dermatitis (SCORAD), Dermatology Life Quality Index (DLQI), Physician Global Assessment (PGA), and Patient Global Assessment (PtGA). PASI scores range from 0 to 72, with values below 5 indicating mild psoriasis, 5-10 moderate, and above 10 severe cases [PP]. SCORAD for eczema ranges from 0 to 103, with lower scores indicating less severity [Oak09] [Cho+17]. More general metrics are the DLQI, PGA, and PtGA. The DLQI score (0-30) is calculated from a patient questionnaire and assesses the impact of a skin condition on the patient's quality of life. Lower scores indicate better quality of life [PP]. The PGA and PtGA use a five-point scale, where 0 denotes no symptoms and 5 the highest severity [PP]. These subjective scores, assessed by clinicians or patients, are also used to evaluate therapeutic efficacy, with changes in scores (e.g., PASI 90, PASI 75) reflecting treatment outcomes.

1.2.1 Clinical categorisation of skin diseases

The clinical classification of ncISDs is based on their underlying pathomechanisms, distinguishing chronic inflammatory, autoimmune, autoinflammatory, and rheumatic skin diseases. Chronic inflammatory diseases, the largest group, are marked by persistent inflammation [Liu+22]. Autoimmune skin diseases result from immune attacks on healthy tissue, whereas autoinflammatory disorders involve inflammation without a clear trigger or autoimmune component. Rheumatic skin diseases affect the joints and connective tissue. The classification remains challenging due to overlapping clinical and biological features, as seen in psoriasis, which exhibits characteristics of multiple categories [Uji+22] [Kan20]. The following sections provide a detailed overview of these categories.

1.2.1.1 Chronic inflammatory skin diseases

Psoriasis, a chronic inflammatory disorder affecting approximately 2 % - 3 % of the global population [Dam+21] [Sew+19], is mediated by T-cell-driven chronic inflammation. It is associated with a reduced life expectancy, cardiovascular diseases, and higher BMI [Kru12] [Dav+10] [Rei12] [Ede+19], and can be triggered by intrinsic factors (e.g., stress), extrinsic factors (e.g., lifestyle), or genetic predisposition, leading to a systemic inflammatory cascade. Psoriatic skin is characterised by erythema from T-cell infiltrates and angiogenesis, and by acanthosis resulting from hyperproliferation of KCs. Additionally, parakeratosis occurs due to a defect in KC terminal differentiation, leading to a thin or absent granular layer and KCs with nuclei in the upper epidermis [Ruc+11] [DRM11]. While the symptoms are typically visible on lesional (L) skin, Farber et al. (1985) suggested that non-lesional (NL) skin also displays pre-psoriatic characteristics, indicating that the entire skin is affected in psoriasis [FNS85].

Psoriasis is a disease spectrum with varying clinical and biological phenotypes, including chronic plaque and generalised pustular psoriasis at opposite ends [Gri+21]. Chronic plaque psoriasis, the most common form, is characterised by itchy, burning red plaques with silvery scales, typically on the scalp, trunk, and limbs, while pustular psoriasis is marked by pustules and erythema and differs by being driven by autoinflammatory rather than autoimmune pathways [Gri+21] [Lia+17]. This spectrum reflects the diverse clinical presentations and immune mechanisms underlying the disorder. Psoriasis is generally driven by a type 3 immune response, leading to hyperproliferation of KCs, neutrophil infiltration, and production of immune mediators like cytokines and chemokines [Gri+21]. IL - 17A is the hallmark cytokine, produced by various immune cells, including Th17, Th22, and Th1, which also secrete TNF, IL - 22, IL - 23, and IFN - γ [LSFK14] [Mar+18] [Dav+10]. KCs, activated by T-cells, produce TNF - α, pro-inflammatory cytokines of the IL - 1 family such as IL - 36, and chemokines such as CXCL1, CXCL8, and CCL20 [LSFK14] that sustain the chronic inflammatory cycle in psoriasis. In summary, psoriasis is a multifaceted skin disease spectrum, characterised by varied clinical phenotypes and complex interplays between immune cells and KCs.

Plaque psoriasis is often linked to mutations in CARD14, which increase chemokine production and immune cell recruitment, sustaining inflammation [LSFK14]. $\mathrm{T_{RM}}$ - cell cells contribute to relapse after remission [OS+20] [Mat+17]. Clinicians assess disease status and therapy responsiveness based on skin lesion characteristics [Gri+21]. Four approved drug targets, i.e. TNF - α, IL - 12/23, IL - 17, and IL - 23 inhibitors, reduce cytokine expression and associated comorbidities [AR20] [Zab+08]. In summary, plaque psoriasis is driven by CARD14 mutations and $\mathrm{T_{RM}}$ - cell cell activity, with treatments managing disease progression and reduce comorbidities.

Atopic eczema (or atopic dermatitis), affecting up to 30 % of children and 3 %- 10 % of adults, significantly impacts quality of life and can lead to depression [Uji+22] [SS18] [Kim+15]. Its symptoms include itchy, painful skin lesions, and its diagnosis is challenging due to heterogeneity. The condition is linked to genetic and environmental factors, including mutations in filaggrin (FLG), which impair the skin barrier and allow allergens to trigger immune responses [SS18]. Acute stages are dominated by type 2 immunity, while chronic phases involve mixed immune responses [Uji+22] [VS19]. Treatment focuses on restoring the skin barrier and regulating immune function, with biologics such as IL - 4 blockers being used [Uji+22] [Han80]. As prevalent and heterogeneous skin disorder, eczema will benefit from precision medicine, as effective therapies are urgently needed.

Lichen planus is a chronic inflammatory skin disorder affecting the skin, mucosa, and appendages, often triggered by side-effects of medications [Uji+22]. Cutaneous lichen planus is characterised by T-cell and macrophage infiltration in the upper dermis, leading to purple-red, itchy papules with whitish net-like lines [Sol+21] [Boc+21]. Diagnosis is based on clinical and histopathological observations, but distinguishing it from other skin conditions is challenging [GDF14]. Despite treatment advances with biologics targeting IL - 17, IL - 12/IL - 23, and IL - 23, effective therapies remain limited [Uji+22] [Boc+21]. The pathogenesis of lichen planus, particularly involving Th17 cells, suggests the potential for drug repurposing and precision medicine. Cutaneous lichen planus has a complex pathogenesis involving T-cell and macrophage infiltration, requires careful diagnosis, and thus will also benefit from precision medicine.

Granulomas are chronic inflammatory skin lesions that can result from infections, such as tuberculosis, or from unknown causes in conditions like sarcoidosis and granuloma annulare [Tim+16] [SHS17]. These lesions consist of immune cells, including macrophages, T-cells, and B-cells, and appear as annular clusters of papules and plaques [Chu21] [Tim+16]. Granulomatous diseases are rare, and while granulomas may resolve over time, persistent ones can cause organ damage due to fibrosis. TNF - α inhibitors are commonly used in sarcoidosis but require careful selection to avoid exacerbating the condition [Bau+06] [Tim+16] [SHS17]. The formation of granulomas is the result of an unidentified underlying cause and the selection of an appropriate treatment plan is challenging due to the potential for organ damage.

1.2.1.2 Autoimmune skin diseases

Bullous pemphigoid is one of the most common autoimmune skin diseases. It is characterised by blisters surrounded by itchy skin rashes on the trunk and extremities [SZ13]. Bullous pemphigoid can be triggered by radiation, thermal burn, scaring, mechanical irritation, and malfunction of Tregs [Stä+21] [Mai+18] [Mur+18]. Based on these clinical, histological, and immunological features the patients are diagnosed and subsequently treated. Targeting CD20 by specific antibodies is an effective treatment strategy and other biologics, which block IL - 4, are currently under investigation [Uji+22] [Hey20] [Abd+20]. The drug target IL - 4 is already approved for eczema patients and could be another candidate of a successful drug repurposing in ncISDs.

1.2.1.3 Autoinflammatory skin diseases

Autoinflammatory skin disease are defined by recurrent or continuous inflammation due to antigen-independent hyperactivation of the innate immune system [WSY21]. Examples are pyoderma gangrenosum and pustular psoriasis.

Pyoderma gangrenosum is characterised by ulcers, which are often open painful wounds with a reddish-purple border, and an accumulation of neutrophils in its skin lesions

[Mav+20]. Clinically, pyoderma gangrenosum is characterised by heterogeneous morphological features and additional laboratory analysis of skin biopsies taken from the ulcer border is required for a definitive diagnosis [Mav+20]. In addition to the high infiltration of neutrophils, it is suspected that T-cells are involved in the autoinflammatory response of the innate immune system [Mav+20]. However, the cause of pyoderma gangrenosum remains unknown [Mav+20]. Treatment plans of pyoderma gangrenosum include wound care and a combination of fast and slow acting biologics [Mav+20].

Pustular psoriasis, like pyoderma gangrenosum, is characterised by a high neutrophil count and active innate immune response [Lia+17]. It differs immunologically from plaque psoriasis, with IL - 36 contributing to neutrophil recruitment and inflammation. A defect in IL - 36RN leads to increased IL - 36 expression and neutrophil accumulation, contributing to a cycle of inflammation [Fur+18] [Lia+17]. Subforms include generalised pustular psoriasis, palmoplantar pustulosis, and Acrodermatitis continua of Hallopeau, with the former being life-threatening. Pustular psoriasis is treated with biologics targeting IL - 17, IL - 12/IL - 23, IL - 23, and TNF - α, although responses may vary from plaque psoriasis, indicating the need for alternative therapies [Fur+18]. Effective treatments are crucial for autoinflammatory diseases like pustular psoriasis, given their complex and varied responses.

1.2.1.4 Rheumatic skin diseases

Systemic sclerosis is an example of a rheumatic skin disease, which is also an autoimmune disease [Bai+21] [All+15]. It is characterised by three comorbidities, i.e. vascular disease, inflammation, and fibrosis, which result in a heterogeneous clinical profile [JD04] [DK17]. The skin lesions in form of oedema, scleroderma, and necrosis or organ damage are typical clinical characteristics of systemic sclerosis [Uji+22]. The pathophysiological mechanisms are not yet understood and effective therapies are still lacking. The identification of patient specific biomarkers is therefore urgently needed [Uji+22] [All+15]. Until now, the biomarkers interleukin 6 receptor, serum S100A6, calumenin, and cytohesin 2 have been found for systemic sclerosis [Bal+21] [DK17]. Systemic sclerosis is a complex disease with a diverse clinical profile and there is an urgent need for effective therapies, highlighting the necessity for precision medicine to manage this condition.

The categorisation of skin diseases is a challenging process due to their phenotypic complexity, making it difficult to find effective treatments. Therefore, alternative approaches are required. One step towards precision medicine is accomplished by stratifying ncISDs into immune response pattern (IRP), which are introduced in the subsequent section.

1.2.2 Stratified medicine by immune response patterns

Over 100 ncISDs have been classified based on clinical phenotypes [EE18]. Due to their heterogeneous nature, accurate diagnosis and appropriate therapy selection remain challenging. Incorrect treatment, particularly biologics, can be life-threatening for a patient. Therefore, the right diagnosis is important for an effective symptomatic therapy.

Advances in molecular technologies have revealed immunologically active pathways in ncISDs, but the majority of skin diseases is rather rare and their pathogenesis still hardly understood, complicating treatment development. Efforts towards stratified medicine group ncISDs by lymphocyte-driven IRPs, enabling the repurposing of treatments for rare diseases within the same IRPs [EE18]. Immunological diversity in ncISDs is influenced by factors like microbiomes, sex, genetics, and autoantigens, with T-cell diversity playing a pivotal role in pathogenesis. To date, six distinct IRPs have been identified, each linked to a specific lymphocyte immune response (Figure 1.3) [EE18] [SGSE22] [EZ14].

Lichenoid IRP 1 diseases, known as *interface dermatitis*, involve immune responses against basal epidermal KCs by type 1 immune cells (Tc1, Th1, ILC1, NK cells) [EE18]. Associated diseases, such as lichen planus, lupus erythematosus, and alopecia areata, show higher IFN - γ expression compared to other IRPs [SGSE22]. The cytokine IFN - γ induces apoptosis and necroptosis in KCs [EE18], and targeting it has been effective in treating IRP 1 diseases [SGSE22] [Sei+20]. In summary, IRP 1 is characterised by immune responses targeting KCs, with IFN - γ being its signature cytokine.

Eczematous IRP 2a diseases are characterised by an overactivated type 2 immune response, leading to impaired skin barriers and defence against allergens, microbes, and parasites (e.g., Staphylococcus). Cytokines involved are IL - 4, IL - 13, and IL - 31 , and IL - 5, which activate granulocytes and mast cells [EE18]. Symptoms of IRP 2a include itch, oedema, and dry skin. Diseases in IRP 2a include eczema and its subtypes. Targeting IL - 4 and IL - 13 has proven effective in treatment [EE18]. Overall, IRP 2a is characterised by skin barrier defects and inflammation, with IL - 4 and IL - 13 as therapeutic targets.

Blistering IRP 2b, like the eczematous pattern, involves an overactivated type 2 immune response. However, in IRP 2b, IL - 4 and IL - 5 cytokines, secreted by B-cells and plasma cells, lead to auto-antibody production against epidermal structural proteins [EE18]. These antibodies cause gaps in the skin, filled with immune cell infiltrates, resulting in blisters and erythema. Bullous pemphigoid is one fo the diseases associated with IRP 2b. An effec-

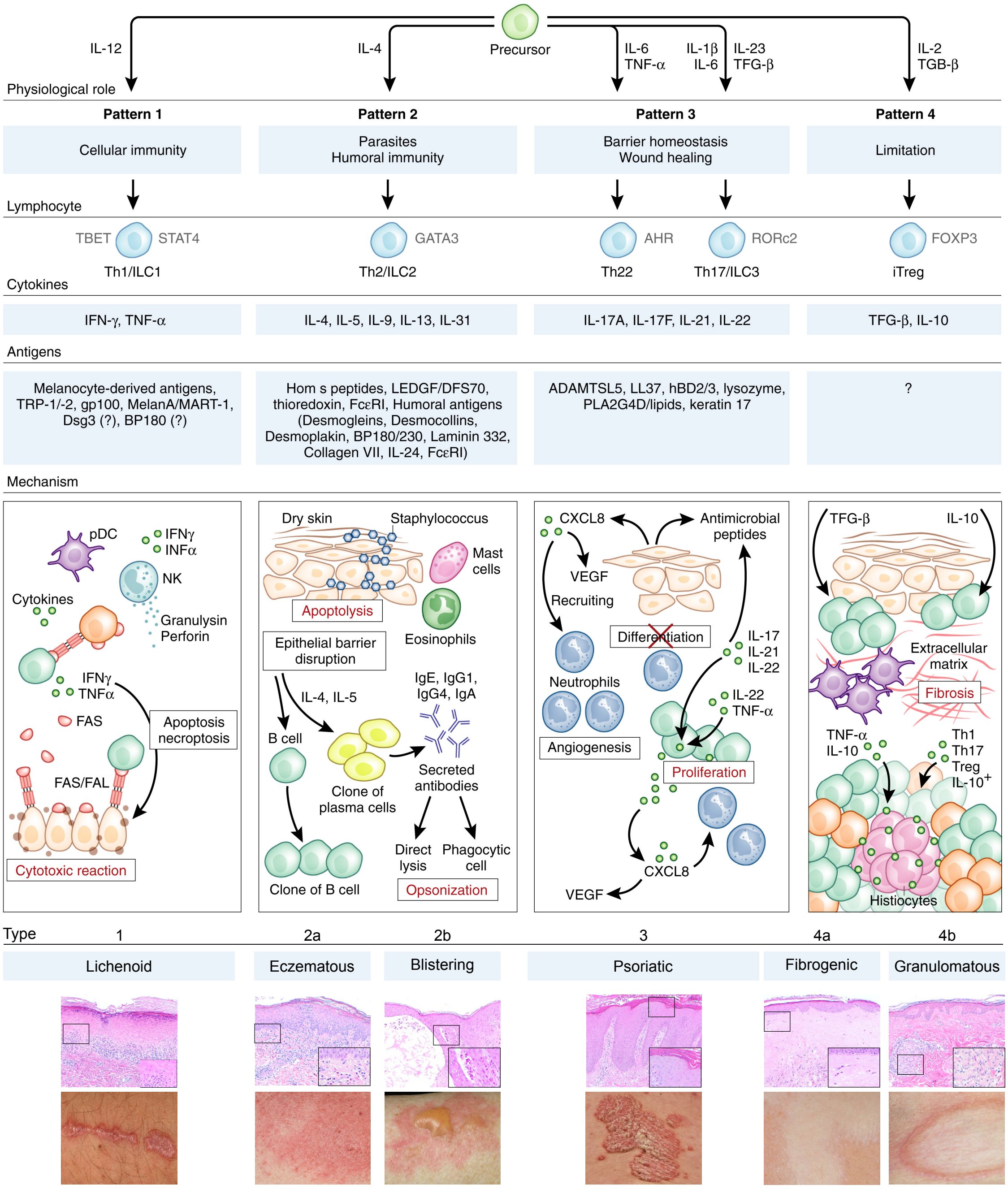


Figure 1.3: Existing IRPs and their mechanism. The IRPs 1-4 are characterised by specific lymphocyte-driven immune responses. The precursor cells differentiate into type 1-4 lymphocytes depending on the presented autoantigen. Each IRP is defined by specific pathways, cell types, signature cytokines, and antigens, along with mechanisms that drive these patterns. These differences result in immune infiltrates accumulating in distinct skin layers, leading to various histopathological changes such as papules, blisters, plaques, and granulomas. Figure adapted from [SGSE22] and [EE18].

tive therapy strategy is to reduce the amount of B-cells by targeting CD20 and IL - 4 inhibitors from IRP 2a may also be beneficial for IRPs 2a associated diseases [EE18]. In essence, IRP 2b immune responses are driven by IL - 4 and IL - 5, with therapies targeting B-cells and possibly blocking IL - 4.

Psoriatic IRP 3 is defined by a type 3 immune response, involving both innate and adaptive immune cells [EE18]. Upon activation, KCs secrete AMPs and chemokines, triggering the recruitment of type 3 immune cells, leading to uncontrolled KC proliferation and epidermal thickening. This cycle may involve autoinflammatory and autoimmune processes, resulting in acanthosis, hyperparakeratosis, and micro-abscesses. IRP 3 associated diseases are psoriasis and pityriasis rubra pilaris, with NOS2 as a known biomarker. Biologics approved for psoriasis are also effective for pityriasis rubra pilaris [EE18]. In summary, IRP 3 is driven by IL - 17 cytokines, and its associated diseases can be treated with psoriasis biologics.

Fibrogenic IRP 4a is characterised by type 4 immune cells, such as Tregs, which secreteIL - 10 and TGF - β. Prolonged Treg activity modulates the immune response, resulting in deeper skin immune infiltrates and less epidermal involvement One of the associated diseases with IRP 4a is systemic sclerosis. While approved therapies are lacking, targeting Treg-secreted cytokines is under investigation in clinical trials [EE18]. The IRP 4a leads to fibrogenic skin changes with deeper immune infiltrates, for which effective therapies remain under development.

In IRP 4b, granulomas form in the dermis to neutralise exogenous infiltrates, with the epidermis often unaffected. These granulomas, consisting of dead or living cells surrounded by Th1, Th17, and Tregs, are seen in diseases like granuloma annulare and sarcoidosis. TNF - α inhibitors are used, although evidence on their efficacy remains conflicting [EE18]. In summary, granulomatous IRP 4b involves complex immune interactions, with TNF - α inhibitors being a common yet controversial treatment for this IRP.

IRPs support stratified medicine by classifying ncISDs based on lymphocyte-driven immune responses. However, they may not fully capture the heterogeneity of ncISDs, as the molecular profile can vary with disease status (e.g., onset, early, chronic) and severity. Additionally, environmental, therapeutic, and psychosomatic factors influencing the immune response are neglected [EE18]. While IRPs are a step towards precision medicine, further methods are needed to enhance understanding of ncISDs pathogenesis.

1.3 Transcriptomics

In order to pursue finding novel therapeutic targets or biomarkers, microarray and NGS technologies have been developed over the last decades [SKK20]. They offer fast and high-throughput molecular profiling of skin biopsies, thus providing snap-shots of healthy or diseased tissue. This yields additional information to clinical phenotyping to improve diagnostics and response to therapy. It also advances the understanding of the pathogenesis of ncISDs [SKK20].

Cells are the fundamental units of life, comprised of a membrane, cytoskeleton, organelles, mitochondria, ribosomes, and a nucleus (Figure 1.4 a) [Cam+20]. The membrane controls molecular exchange, the cytoskeleton maintains structure, mitochondria are responsible for metabolism and energy production, and the nucleus contains genetic information stored on double helix DNA strands [FM10]. The DNA, composed of nucleotide sequences, encodes genes that serve as blueprints for protein synthesis [Cam+20], which enable specific cellular functions. Thus, cells are highly organised units, with each component contributing to cellular functions.

Cells regulate gene expression by adjusting transcription and translation efficiency based on their type and conditions, leading to variations between healthy and diseased skin. Gene expression, the process of building a protein, consists of *transcription* and *translation* (Figure 1.4 b). During transcription, a gene, i.e. a specific nucleotide sequence on the DNA, is transcribed into messenger RNA (mRNA). This involves unwinding the DNA double helix, using one strand as a template, and synthesizing a complementary RNA strand, with the coding strand serving as the reference for the gene sequence [Alb+02a]. The resulting mRNA serves as an intermediary for protein synthesis and is translated by ribosomes into a polypeptide chain, which folds into a functional protein [Alb+02b] [Cam+20]. To this date, the human *genome* comprises approximately $60,000$ genes, of which around $20,000$ encode proteins [Ama+23]. In essence, gene expression process involves the transcription of DNA into mRNA and translation of mRNA into proteins, a process essential for cellular functions.

In ncISDs, conventional diagnostics rely on the phenotyping by clinicians, which are therefore subjective. An objective, hypothesis-free alternative are high-throughput technologies such as microarray and NGS. Latter has been the focus of studies due to its advantage of higher sensitivity, quantitative accuracy, and measurement of other forms of RNA such as microRNA and RNA isoform [SKK20] [Qui14] [Kõk+16]. Both microarray

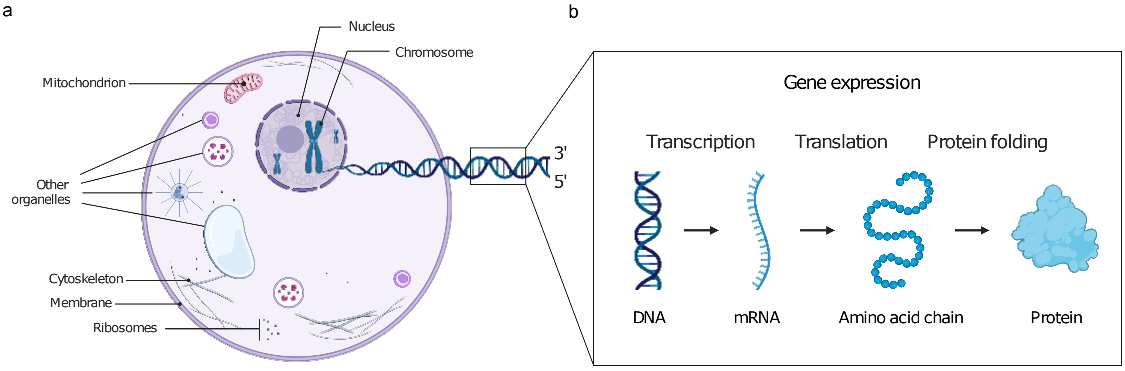


Figure 1.4: Cell organelles and the gene expression process. a) Structure of a cell showing the most important components. **b)** Gene expression is the process of transcribing and translating a gene into a protein: DNA → RNA → protein. A nucleotide sequence on the template DNA strand, the protein-coding gene, is transcribed into mRNA and then translated into a chain of amino acids, a polypeptide. Finally, the protein is created by folding the polypeptide into a specific shape giving the protein its characteristic function. Created with BioRender.com.

and NGS enable the study of complex, biological systems measuring genomics (DNA), transcriptomics (mRNA/transcripts) or proteomics (proteins). This thesis focuses on analysing transcriptomics of bulk, single cell, and spatially resolved RNA-sequencing (RNA-seq) data using NGS technologies (Chapters 3.1, 4.1, 5.1, 2.5). Essentially, RNA-seq is used to measure the amount of molecules expressed in a tissue on different resolutions in the form of transcribed genes.

To quantify transcripts, skin punch biopsies or tape stripping samples are collected from L and healthy controls. These are prepared for sequencing and aligned to a *human reference genome* (e.g., GRCh38 [Ama+23]). This process includes stochastic elements aiming to generate reliable and biological accurate results [Dob+13]. The output is a count matrix representing transcript abundance per sample, which serves as a measure of gene expression. Such data enable comparative analyses of disease states, cell communication, and micro-environment interactions, providing insights into ncISDs.

The transcriptome of ncISDs is comprised of skin cells and skin resident, infiltrated, and activated immune cells. RNA-seq enables the study of gene expression patterns in complex, biological systems like the skin. By examining mRNA levels, transcriptomics analysis can reveal pathways and mechanisms involved in ncISDs pathogenesis, potentially leading to novel prevention and treatment strategies.

1.4 Advancements in non-communicable, chronic inflammatory skin diseases by transcriptomics

Over the past decades, microarray and RNA-seq have been used to study the molecular profiles of ncISDs, particularly psoriasis and eczema [SKK20]. While analyses of *significantly differentially expressed genes (DEGs)* revealed both differences and similarities between these diseases, studies comparing DEGs produced also contradictory results [SKK20]. This could be due to various factors, such as methodological differences, cohort size, biopsy site, disease status, and diagnostic accuracy. Therefore, comparing studies at the functional level, focusing on differentially active biological pathways, is recommended, as it offers a more robust understanding of disease pathophysiology [SKK20]. This section discusses the current research, opportunities, and applications of transcriptomics in ncISDs.

1.4.1 Diagnostic opportunities through transcriptomics and computational approaches

To improve diagnostics, which currently relies on clinical phenotypes, amnesia, and histological assessments, transcriptomics data can define molecular disease profiles. In psoriasis, molecular profiling shows that KC and the cytokines IL - 17A and TNF - α play a major role [Swi+16]. Comparisons of healthy, L, and NL psoriatic skin confirm that the entire skin is involved, with NL psoriatic skin representing a pre-psoriatic state [Air+15] [Gud+09] [Nos+21]. In eczema, molecular profiling challenges the idea that FLG mutations impair the skin barrier, as these mutations occur only in a subset of patients [Blu+18] [SKK20]. Common gene sets have also been identified, revealing shared biological processes across diseases [SKK20] [Jab+14] [Swi+16]. Thus, transcriptomics provides a deeper understanding of ncISDs, offering more accurate diagnostics by revealing shared and disease-specific mechanisms.

Several studies have identified disease-specific genes for conditions such as psoriasis, eczema, and lichen planus [Gar+16] [Rei+19b] [Tia+12] [Ewa+15] [Gho+15] [Kam+10] [Tso+19]. In 2010, a set of seven genes (*CAII*, *NELL2*, *hBD2*, *IL1F9*, *CXCL8*, *CXCL10*, *CCL17*) was used to differentiate between healthy skin, psoriasis, eczema, and contact dermatitis [Kam+10]. Another classifier, based on four genes (*IL36G*, *CCL27*, *NOS2*, *C10ORF99*), distinguished psoriasis, eczema, contact dermatitis, and lichen planus [Rei+19b]. Tsoi et al. (2019) identified *IL17A*, *IL13*, and *IL36G* as markers distinguishing psoriasis from eczema. In addition, they revealed an overlap between these phenotypically different ncISDs hinting to common signalling pathways and molecular mechanism

[Tso+19]. Additionally, *NOS2* and *CCL27* have been found to differentiate psoriasis and eczema, even their subtypes [Qua+14] [Gar+16]. Notably, reducing over $20,000$ protein-coding genes to just a few for accurate disease classification is remarkable. Based on *NOS2* and *CCL27*, a classifier, commercialised as "PsorX" by Dermagnostix, was developed [Gar+16] [DER]. This demonstrates the potential of molecular profiling to simplify dermatological diagnostics.

Analysing transcriptomics and clinical assessments in conjunction can enhance the understanding of ncISDs. For instance, correlating transcriptomics with severity metrics, such as PASI or PGA score, could uncover genes and pathways associated with disease states such as "mild" and "severe" [SKK20]. Batra et al. (2020), demonstrated the benefits of integrating phenomics and transcriptomics by revealing previously unknown, clinically relevant biological processes [Bat+20]. Gathering and combining all available data sources in the age of AI can be crucial for driving precision medicine.

Molecular profiling can help identify disease subtypes and endotypes. Psoriasis, for instance, includes plaque psoriasis and generalised psoriasis pustulosa, which differ in *IL36G* expression [Lia+17]. Yet, they share common molecular features, suggesting they are subtypes of the same disease [Joh+17]. Further clustering of plaque psoriasis transcriptomes revealed molecular subtypes with distinct pathway activities, indicating different treatment responses [Ain+12]. Similarly, eczema is a heterogeneous disease, with a study identifying four endotypes characterised by distinct pathophysiological processes [Thi+17]. Understanding these subtypes and endotypes can guide patient-centred treatment strategies, advancing precision medicine.

Transcriptomics can assist in understanding treatment mechanisms and effects on a molecular level [SKK20]. A study, carried out by Johnson-Huang et al. (2012), elucidated the pairwise respective similarities in molecular profiles of NL samples and responders as well as between L samples and non-responders psoriasis patients to the given CD11a suppressor [JH+12]. However, comparing the transcriptomes of psoriasis patients before and after treatment with TNF - α blockers showed that some genes are still differentially expressed belonging to a "molecular scar" [Tia+12]. Hence, anti TNF - α biologics do not completely transition L into NL skin at the molecular level, even though histopathology shows complete recovery [SF+11]. In eczema, JAK inhibitors are promising candidates for an effective therapy [Tsa+21]. Thus, integrating transcriptomics into treatment evaluation can reveal molecular changes, assisting in optimising therapeutic strategies.

1.4.2 Skin disease models

Disease models, such as 2D or 3D KC, mouse, and canine models, offer alternatives to skin biopsies for studying ncISDs. They provide a simplistic and accessible representation of a skin condition and therefore do not fully reflect the biological mechanisms of the diseases. For example, the Imiquimod (IQM) mouse model, commonly used to mimic psoriasis, was found to more closely resemble allergic contact dermatitis at a molecular level [GS+18]. KC models can be used to study interactions between cells and effects of TNF - α and hallmark cytokines such as IL - 17A, potentially offering new insights into the disease mechanism and drug targets [Chi+11] [Rio+21]. Up to now, no standardised skin model exists [SKK20].

1.4.3 Opportunities of single-cell RNA-sequencing and spatial transcriptomics

The aforementioned findings and applications of transcriptomics were achieved by utilising either microarray or bulk RNA-seq, which operate on the population level of cells and, thereby providing an average gene expression profile per sample. In the following, insights are introduced that can be derived from scRNA-seq and ST.

1.4.3.1 Single-cell RNA-sequencing in skin diseases

Utilising scRNA-seq provides insights into the cell type composition and complex regulatory interactions between cells within human skin, particularly into the immune response in ncISDs. Using regulatory networks and cell-cell communication analysis, scRNA-seq can reveal signalling processes that drive immune responses in these skin disorders [IMBH19] [Jin+21] [Arm+21]. Moreover, clustering analyses of single and mixed disease datasets, including healthy controls, allow for identifying differences in the expression within and between cell types as well as unravel the heterogeneity in cell type composition [HLB18] [Rey+21]. This approach can also enhance the understanding of the cellular landscape of both common and rare skin diseases, such as systemic sclerosis [Xue+22]. These analyses enhance the understanding of cellular interactions and disease-specific differences within the skin, providing a foundation for more targeted treatment strategies for ncISDs.

The exploration of cell states, rare subpopulations, and cell lineages is enabled by scRNA-seq [Rey+21] [Wol+19] [Ber+20] [Lan+22]. For example, Reynolds et al. (2021) demonstrated potential developmental differences in the cutaneous immune system that already occur during the prenatal phase, which might reveal new drug targets [Rey+21]. Additionally, scRNA-seq can refine disease classifiers and focus on specific cellular subgroups, as shown by Liu et al. (2022), who identified a promising subgroup for more precise diagnostics and treatments [Liu+22]. Thus, scRNA-seq allows to elucidate developmental biological processes and to identify molecular markers, supporting the development of targeted therapies in ncISDs.

Despite its advantages over bulk RNA-seq, scRNA-seq has limitations. During preparation, samples are dissociated into individual cells, leading to loss of structural integrity, potential cell death, and altered gene expression due to stress or stimuli [Wil+22]. Since scRNA-seq lacks spatial information, it cannot capture how micro-environmental factors influence cell function and communication. Consequently, it primarily reflects dominant biological processes, potentially overlooking region-specific or less prominent functions. Integrating spatial information is essential for a more accurate understanding of cellular functions and micro-environmental influences.

In summary, scRNA-seq provides valuable insights into cell-specific functions, cell type composition, interaction, and developmental processes. In order to study a cell's behaviour within tissue, integrating spatial information for more accurate conclusions is important. This thesis uses scRNA-seq to explore ncISD on a single-cell resolution and to validate findings drawn from ST, thereby providing a more comprehensive view on the pathogenesis of the skin conditions.

1.4.3.2 Spatial transcriptomics in skin diseases

ST advances scRNA-seq by integrating spatial information and analysing intact tissue, thereby preserving gene expression profiles [Wil+22]. This technology quantifies mRNA in a spatial context, with early efforts in ST dating back to the 1960s using in situ hybridisation (ISH) [MP22]. ISH detects genetic material within cells using, e.g., fluorescence markers (FISH). This process is called hybridisation and enhances also the contrast with counter staining of the background, i.e. cells and other structures [HVVK18]. Recent ST technologies, such as Visium or Xenium from 10X Genomics [Stå+16] [Lee+15] [Ke+13], GeoMx or CosMx$^{\text{TM}}$ SMI from Nanostring [Mer+20] [He+22], and RNAScope [Wan+12] have become more accessible. While RNAscope, Xenium, and CosMx offer single-cell resolution but are limited to around $1,000$ RNA transcripts, Visium and GeoMX provide whole-transcriptome analysis at mini-bulk resolution, averaging molecular profiles from multiple cells in predefined spots across tissue sections [Stå+16] [He+22]. In this thesis, ST data is created using Visium to explore ncISDs on a spatial resolution (Chapter 5).

In addition to measuring spatially resolved transcripts, ST provides a microscopic image of the tissue being studied, allowing for the examination of the architecture and cellular organisation. Moreover, it enables to study tissue niches [FST23], cellular communication [Can+23] [Li+23], and spatial gene expression patterns in 2D. As certain analyses require single cell resolution, methods to recover the cell type compositions using spot deconvolution [Bia+21] [And+20] [Kle+22] and nuclei segmentation [Str+21] [Sch+18] [Ban+17] have been developed. These techniques recover cell type composition within tissues, revealing cellular heterogeneity and biological differences associated with tissue layers, structures, and disease states.

In ncISDs, ST can advance the understanding of cell states and tissue function in different disease states [Zha+22b]. Moreover, it enables to explore anatomical regions based on their spatial expression pattern, analysing neighbourhoods, and studying the immune response [Rao+21]. Thus, ST offers valuable insights into the molecular landscape of ncISDs on a spatial resolution.

Currently, ST data captures tissue snap shots, making it ideal for studying static systems. Dynamic systems or diseases, such as skin lesions and solid tumours, can also be investigated by taking serial samples, which is however, not always possible [Joh+22]. Further limitations, such as low capture area, resolution, and sensitivity, can lead to missing or low abundant transcripts, potentially impacting downstream analyses [Wan+23] [SK23]. By overcoming these limitations and reducing the costs, more and more datasets will be generated, leading to a better understanding of the disease pathogenesis.

In summary, ST is advantageous for studying ncISDs as it preserves spatial information, revealing localised immune response patterns, cell type composition, and providing insights into disease pathogenesis within tissue. This thesis presents one of the first Visium datasets of ncISDs (Chapter 5).

1.5 Challenges in non-communicable, chronic inflammatory skin diseases

Common ncISDs, including psoriasis, eczema, and lichen planus, have been studied at the phenotypic and transcriptomics levels, providing insights into their pathogenesis and supporting targeted therapy developments. However, challenges remain, particularly for rare ncISDs, with gaps in (1) patient stratification, (2) treatment response, and

(3) detailed molecular understanding. These are compounded by overlapping clinical phenotypes, limited knowledge, and difficulties in obtaining skin biopsies due to patient discomfort, scarring, and the need for local anaesthesia. As a result, molecular profiling data, especially for rare ncISDs, is scarce.

Challenge 1: Enhancing patient stratification. Diagnosis of ncISDs relies on clinical phenotyping, amnesia, and histological assessments, which are subjective and variable. Overlapping entities and intra-disease variability complicate diagnosis and treatment response. Current approaches group diseases based on lymphocyte IRPs [EE18] [SGSE22], but these neglect other factors influencing disease manifestation, limiting their effectiveness. Thus, more precise disease descriptions and refined patient stratification approaches are needed to improve diagnostics and treatment outcomes.

Challenge 2: Enhancing therapeutic precision through molecular insights. While molecular classifiers distinguish between psoriasis, eczema, and lichen planus, they do not ensure treatment success [Ain+12]. Drug responses vary even within subtypes, showing the limitations of current disease definitions. Although IRPs-based classifications improve stratification, they still rely on clinical phenotypes, resulting in variability in treatment response. To optimise efficacy, diagnostic tools must incorporate molecular markers beyond IRPs. Integrating these into therapeutic decision-making could improve treatment efficacy and minimise adverse effects, advancing precision medicine in ncISDs.

Challenge 3: Advancing molecular understanding of ncISDs. A deeper molecular understanding of ncISDs is essential for improving diagnosis and treatment. Sequencing techniques, such as ST and scRNA-seq, offer insights into cell composition, interactions, and tissue function. While recent studies have explored ncISDs at single-cell resolution [GR20] [Liu+22], further research utilising ST is required to fully characterise disease-driving cells and their micro-environmental interactions. This could enhance our understanding of pathogenesis and assist in identifying biomarkers to intervene in the vicious cycle of inflammation of these complex disease.

In summary, improving patient stratification in ncISDs requires large, diverse cohorts and integration of transcriptomics and clinical phenotypes. Enhancing therapeutic precision necessitates diagnostic tools based on molecular markers. Moreover, advanced sequencing techniques, such as ST and scRNA-seq, enable a deeper molecular understanding of ncISDs, allowing to characterise disease-driving cells and their micro-environment. Addressing these challenges will enable the implementation of precision medicine in ncISDs.

1.6 Aims and research questions

Addressing aforementioned challenges necessitates improved diagnostics, tailored biologic recommendations, and insights into the spatial organisation to realise precision medicine in ncISDs.

This thesis enhances patient stratification by applying a hypothesis-free approach on bulk RNA-seq data to form endotype clusters. I assess their relevance through pathway enrichment analyses and phenotype integration, hypothesising that endotypes better capture disease heterogeneity than clinical assessments alone. Moreover, to enhance therapeutic precision, I hypothesise a correlation between drug response and psoriasis-associated endotypes. I create classifiers, based on a minimal set of single predictive molecular markers, capable of distinguishing between specified groups of endotypes. Lastly, I use ST to explore the micro-environment of disease-driving immune cells. In particular, I show ST's advantage by comparing it to bulk RNA-seq and scRNA-seq, and hypothesise that it yields insights into the inflammatory micro-environment.

This thesis aims to address the following objectives (Figure 1.5) and research questions.

Aim 1: Utilising molecular profiling to derive biological meaningful endotypes:

(i) Can I group patients into endotypes?

(ii) How are endotypes and the current disease ontology related?

(iii) What are the clinical and biological characteristics of endotypes?

Aim 2: Linking endotypes to drug response to advance precision medicine:

(iv) Can I link endotypes with drug response?

(v) Can I classify new patients into specific endotypes based on a minimal set of single genes?

Aim 3: Leveraging ST to advance the understanding of the underlying disease mechanisms:

(vi) What is the spatial expression distribution of disease-driving immune cells?

(vii) What biologically characterises their micro-environments?

(viii) Can I determine the impact radius of disease-driving immune cells?

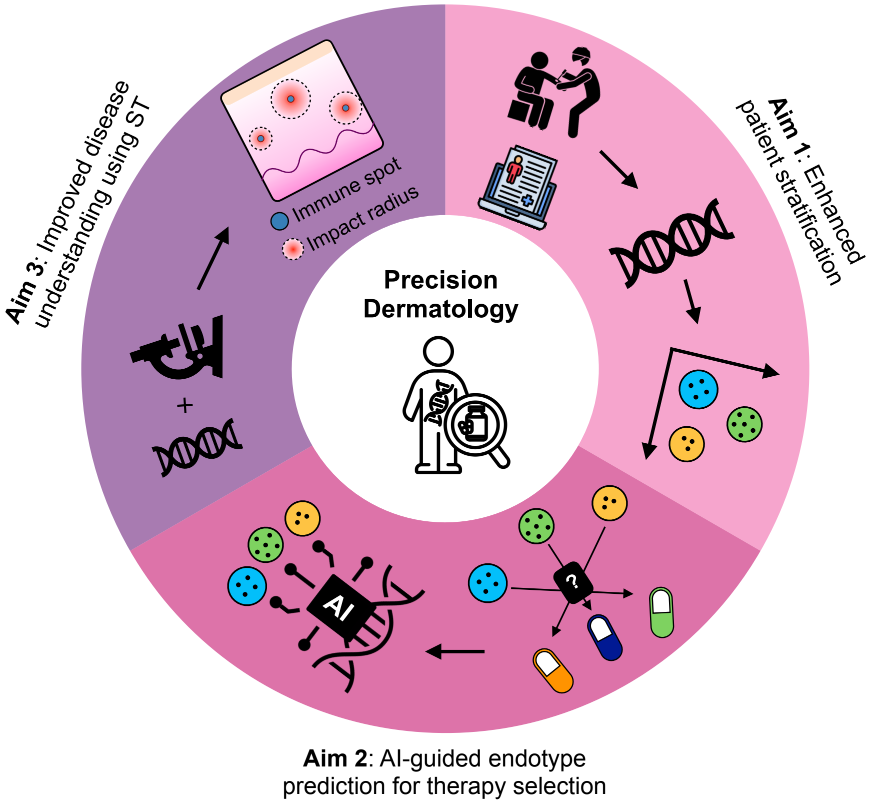

Aim 1: Enhanced patient stratification
Aim 2: AI-guided endotype prediction for therapy selection
Aim 3: Improved disease understanding using ST
Precision Dermatology
Immune spot
Impact radius
AI
?
+

Figure 1.5: Graphical abstract of research aims. Aim 1 enhances patient stratification by identifying molecularly distinct endotypes. Building on this, Aim 2 uses AI to classify patients into these groups, associated with drug responses, thereby improving therapy selection. Finally, Aim 3 utilises ST to improve the understanding of ncISDs, particularly through spatial immune interactions. Icons made by Freepik [Fre], HANIS [HAN], Parzival' 1997 [199], and Awicon [Awi] from www.flaticon.com.

1.7 Outline

The aim of this thesis is to provide methods and strategies to drive precision medicine in ncISDs. In particular, this will be addressed by suggesting alternatives to the established disease ontology and exploring the spatial landscape of ncISDs.

Chapter 2 provides an introduction to statistical tests, ML, and analysis methods that were utilised in Chapter 3, 4 and 5. This chapter sets the foundation for the analytical techniques applied throughout this thesis.

Chapter 3 aims for improving the current disease ontology of ncISDs. It identifies disease endotypes in ncISDs by integrating transcriptomics and phenotypes. Endotypes provide an objective, data-driven alternative to the current patient stratification approach by IRP, which is relies on subjective assessments. This chapter establishes the foundation for a more precise patient stratification approach that improves the understanding of ncISDs.

Chapter 4 aims to pave the way for enhanced therapy response by hypothesising an association between psoriasis endotypes and drug response. To achieve this, binary classifiers are created using a minimal set of predictive genes identified through a novel features selection approach. The hypothesis is tested by evaluating the classifiers in an independent cohort. This chapter is pivotal in providing predictive biomarkers for endotype classification, which may be linked to drug response, offering a potential approach for the realisation of precision medicine in ncISDs.

Chapter 5 aims to deepen the understanding of ncISDs by exploring their ST landscape. Using ST data, I investigate the spatial distribution of immune cell expression and hallmark cytokines in psoriasis, eczema, and lichen planus within the skin. This chapter provides novel insights into the inflammatory micro-environment induced by cytokines, highlighting their impact radius through a novel clustering algorithm, as well as the spatial localisation and heterogeneity of ncISDs.

In Chapter 6, I discuss the outcome of Chapter 3 - 5 and provide an outlook on potential applications and implications of my findings for the clinical practice and investigation of ncISDs.

Chapter 2

Statistical background

This chapter provides an overview and introduction to the utilised statistical tests, machine learning (ML) techniques, and the subsequent exploratory data analyses. All analyses were conducted in Python (version 3.7 and 3.8 [VRD09]) and R (4.0.0, 4.2.1, and, 4.2.2 [Tea22]). Throughout this thesis, it is always assumed that the default parameters of the packages or libraries were used, unless otherwise stated.

2.1 Sampling methods

In research, it is often infeasible to collect data from all eligible sources. Consequently, representative samples must be selected. The sampling approach determines the suitability of the representatives for the task and whether bias has been introduced. In addition, it affects the reliability of the results and, consequently, the integrity of the findings [Tur20]. Thus, the choice of sampling strategy is fundamental in ensuring that research findings are both valid and representative of the intended population.

Sampling methods are categorised into probability and non-probability sampling. Non-probability sampling, which relies on non-random selection, is not examined in this thesis [EN17]. Probability sampling methods, such as random sampling and stratified sampling, involve random selection. As probability sampling is useful for generalising sample-based findings to the census [EN17], it is introduced in more detail below.

In order to introduce the sampling techniques, it is necessary to define the terms population, sampling frame, and sample. A *population* refers to the entire set of elements, i.e. data points or observations, within the study's defined scope. The *sampling frame* is a list or description specifying the elements that qualify for inclusion in the population, while a *sample* is a subset of these elements taken from the population [Tur20]. The sample size depends on various factors, such as population size, the variation within the population, and the research objectives [EN17]. In general, the more representative a sample, the more likely it is that the findings will be generalisable.

In the context of the studies presented in this thesis, patients were selected based on specific criteria and asked to donate punch biopsies from both lesional (L) and non-lesional

(NL) skin. Statistically, each study functions as a sampling frame that defines which samples (i.e., biopsies) are drawn from the population (i.e., biobank). The sample size is therefore constrained by the number of donors. In addition, observations, representing the data points in a sample, are described by features such as the expression level of the genes or clinical traits. These sampling consideration ensure that the research findings are based on a well-defined subset of the population, contributing to the study's robustness and relevance.

Random sampling is the simplest probability sampling method. Each element in a population has equal chance to be selected [EN17] [Tur20]. However, it is not guaranteed that the selected samples are representative of the population. Especially, if the representatives are not equally represented or only parts of the whole population could be collected. Thus, more complex sampling methods should be considered to ensure the representativeness of the samples.

Stratified random sampling ensures that each element has a known probability of selection [EN17]. It involves dividing a population into strata based on relevant characteristics, then selecting a random sample from each group, ensuring representation across subgroups. This method helps correct imbalances, such as when one patient group is significantly larger than another, preventing bias in results [EN17]. In summary, stratified sampling enhances the reliability of comparisons across groups by ensuring that each subgroup is represented in the strata in an equal and proportional manner, thereby supporting accurate conclusions.

In this thesis, subsamples are drawn from a representative sample of the population using either random or stratified sampling in the method presented in Chapter 4.

2.2 Statistical testing

In research, theories are developed based on data. In particular, hypotheses are formulated with the intention of either being confirmed or rejected. This is achieved through the application of statistical tests, whose underlying assumptions are met by the *test statistic* of the data. Statistical testing enables researchers to objectively evaluate hypotheses by comparing data-driven expectations with observed patterns.

In order to test a theory, a *null hypothesis* $\mathcal{H}_0$ and an *alternative hypothesis* $\mathcal{H}_a$ are formulated [Sid57]. Subsequently, a statistical test is chosen that aligns with research question and meets the data assumptions. After performing the statistical test, a *significance level* α is determined, which sets the threshold for rejecting $\mathcal{H}_0$. The most common values for α are 0.05 and 0.01 [Sid57]. In case of $\mathcal{H}_0$ being rejected, i.e., p-value $\leq \alpha$, the alternative hypothesis, $\mathcal{H}_a$, is accepted. In cases where $\mathcal{H}_0$ is retained, it is essential to exercise caution and indicates that there is not enough evidence to reject $\mathcal{H}_0$. Thus, the outcome of a statistical test provides insights into whether the data supports rejecting the null hypothesis.

A typical application of statistical testing is to determine whether two groups differ significantly using a two-tailed test. To illustrate, consider a study of responders and non-responders to a drug with additional clinical information such as age. A research question might be formulated as follows: Does age have an impact on the drug response? In order to answer this question, $\mathcal{H}_0$ and $\mathcal{H}_a$ have to be formulated. This includes two distinct hypothesis: firstly, that both groups are equally old ($\mu_0 = \mu_1$) and secondly, that they are not equally old ($\mu_0 \neq \mu_1$) (eq. 2.1a). A significant difference in terms of the population means, μ_0 and μ_1, is indicated if $\mathcal{H}_0$ is rejected. This, in turn, means that $\mathcal{H}_a$ is accepted and that age does, indeed, have an impact on the drug response. In the event that $\mathcal{H}_0$ is not rejected, it is imperative to consider the possibility that the assumption of $\mu_0 = \mu_1$ is false. Indeed, no conclusions can be drawn in this case and the hypothesis must therefore be reformulated. In summary, two-tailed statistical tests can determine whether there are statistically significant differences between groups.

$$\mathcal{H}_0: \ \mu = \mu_0; \ \ \mathcal{H}_a: \ \mu \neq \mu_0 \ \ \text{(two-tailed test)} \tag{2.1a}$$

$$\mathcal{H}_0: \ \mu \leq \mu_0; \ \ \mathcal{H}_a: \ \mu > \mu_0 \ \ \text{(left-tailed test)} \tag{2.1b}$$

$$\mathcal{H}_0: \ \mu \geq \mu_0; \ \ \mathcal{H}_a: \ \mu < \mu_0 \ \ \text{(right-tailed test)} \tag{2.1c}$$

In addition to testing for equality, a particular direction of difference can be tested using left- or right-tailed tests. These test whether the two groups have equal means (eq. 2.1b and eq. 2.1c). For example, a directional research question might be, whether drug responders are younger than non-responders. In this case, the null hypothesis $\mathcal{H}_0$ would state that either age does not have an impact on the drug response or that non-responders are younger ($\mu_0 \leq \mu_1$). The alternative hypothesis, $\mathcal{H}_a$, would be that drug responders are younger ($\mu_0 > \mu_1$). Therefore, one-tailed tests can provide further insights into the direction of difference between groups.

The significance level α determines the threshold for rejecting the null hypothesis $\mathcal{H}_0$ and accepting the alternative hypothesis $\mathcal{H}_a$. It also represents the probability of falsely

		Null hypothesis $\mathcal{H}_0$		
		True	False	
Decision about $\mathcal{H}_0$	not rejected	true negative (TN) (1 - α)	false negative (FN) p (Type II error) = β	$m - R$
	rejected	false positive (FP) p (Type I error) = α	true positive (TP) (1 - β)	R
Total		m_0	$m_a = m - m_0$	m

Table 2.1: Association between Type I and Type II error probabilities and the decision about $\mathcal{H}_0$. The number of performed tests is denoted as m, truly $\mathcal{H}_0$ tests is m_0, truly $\mathcal{H}_a$ is m_a, and rejected tests is R.

rejecting $\mathcal{H}_0$ [Sid57]. Two types of errors can arise during this decision process.

- Type I error occurs when $\mathcal{H}_0$ is wrongly rejected, with the likelihood of this error increasing as α increases.
- Type II error occurs when $\mathcal{H}_0$ is wrongly accepted, with its probability denoted by β. The power of a statistical test, defined as $1 - \beta$, reflects the test's ability to correctly reject a false $\mathcal{H}_0$.

Increasing sample size n can reduce both errors and enhance test power, although challenging in practice [Sid57]. The probabilities of committing any error and making the correct assumptions under $\mathcal{H}_0$ is shown in Table 2.1. Careful selection of α and consideration of sample size are essential to balancing the risk of errors in hypothesis testing.

Once the hypothesis is defined and samples of equal size have been randomly drawn from a population under the constraint of $\mathcal{H}_0$, an appropriate statistical test can be selected. The choice of test depends on the *test statistic*, which represents the distribution of the sample (Sections A.1, 2.2.1, 2.2.3). The test statistic can be determined either directly or indirectly by applying mathematical theorems such as the central-limit theorem [Sid57]. It is applicable for populations following a normal distribution with large sample sizes (n > 30), and for independent samples [Sid57]. Thus, the test statistic is essential for determining the appropriate statistical test and ensuring valid results.

Once the test statistic has been defined, the *sampling distribution* can be determined, which represents the probability distribution of the test statistic. It is differentiated between *one-sample tests*, *two-sample tests*, and *k-sample tests*. The one-sample test tests whether a sample was drawn from a specific population. In contrast, the two-sample test is used to test whether two samples originate from the same population, meaning it tests for differences between two samples. The k-sample test compares two or more samples simultaneously [Sid57]. Each test has a different requirement regarding the test statistic. In the following sections (Sections 2.2.1, 2.2.3), I introduce statistical tests, which have been applied in the conducted studies (Chapters 3, 4, 5). Parametric tests, which were considered in the analysis but were not used due to the violation of assumptions, can be found in the supplements A.1. Understanding the sampling distributions is important for choosing the correct statistical test and accurately interpreting results.

The degrees of freedom $df \in \mathbb{N}$ represent the number of independent observations available after being subjected to certain constraints. Depending on the context, the restrictions can refer to relations in the data, such as the mean, orthogonal dimensions, and parameter estimation. The former also occurs in statistical testing, as it is often tested whether the means of two samples are significantly different by rejecting the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$. In essence, degrees of freedom provide flexibility for estimating parameters and conducting meaningful hypothesis tests.

In summary, the sequence of events in statistical testing includes the formulation of a theory → collection of data → formulation of a hypothesis → selection of an appropriate statistical test → definition of a significance level → execution of the statistical test → and, finally, acceptance or rejection of the null hypothesis $\mathcal{H}_0$. The following sections will introduce the most relevant statistical tests.

2.2.1 Non-parametric tests

It is recommended to use non-parametric tests when the data does not meet the requirements of parametric tests, such as normality and having numeric values. In comparison to parametric tests, non-parametric tests offer the following advantages [Sid57]:

- Being able to handle ranked, interval, ratio, nominal, and ordinal data
- Being applicable to small sample sizes $n \leq 6$, if the exact distribution of population is not known
- Handling observations originating from multiple populations
- Giving the exact probabilities (most non-parametric tests)

In order to apply a non-parametric test, the data is provided in form of frequencies or categories. The latter requires to rank the observations before applying the statistical test

[Sid57]. During the rank assignment, it may happen that two observations are equal and would therefore have been given the same rank, thus violating the continuity requirement. These observations are called ties. In order to break ties, possible solutions are to build the average/mid/mean rank for those tied observations, thereby preserving the sum of ranks [Sid57].

The binomial test is a non-parametric, one-sample test that assumes a binomial distribution of the data. It requires binary (dichotomous) data, independent observations, a fixed sample size n, and equal probability for each outcome, which can be achieved through random sampling [Sid57]. In a two-class population, with proportions P and $Q = 1 - P$, the binomial test evaluates how likely it is that the observed proportion x of the first class matches the expected proportion P. The null hypothesis $\mathcal{H}_0$ is that the observed proportion equals the hypothesised population proportion [Sid57]. The probability of the two-tailed test is given by

$$p(x) = \left(\begin{array}{c} N \\ x \end{array} \right) P^x Q^{N-x} = \frac{N!}{x!(N-x)!} P^x Q^{N-x},$$

where N is the total number of objects, x is the number of objects in class one, and i is the number of objects in one class. Hence, $N - x$ is the number of objects in the second class. In the case of a directed research question and hence, a one-tailed test, the sampling distribution is defined by

$$p(x \leq \kappa) = \sum_{i=0}^{x \leq \kappa} \left(\begin{array}{c} N \\ i \end{array} \right) P^i Q^{N-i} = \sum_{i=0}^{x \leq \kappa} \frac{N!}{i!(N-i)!} P^i Q^{N-i}, \tag{2.2}$$

which is the sum of probabilities $p(x \leq \kappa)$ of the values x begin equally or even more extreme than the observed value κ. The number of objects in the first class is i while $N - i$ is the number of objects in the second class.

The Wilcoxon signed-ranks test compares whether the median of two matched independent samples differs significantly. It accounts for both the direction and magnitude of differences [Sid57]. The test involves calculating the signed differences between matched samples, ranking their absolute values, assigning average ranks for ties, and summing the positive and negative ranks to determine T [Sid57]. In this thesis, for large sample sizes ($n : n_1 = n_2$), the sampling distribution is approximately normal, allowing the use of a standard normal distribution (z-distribution), which is given by

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{n(n+1)}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} .$$

		Observed frequencies		Total
		group: 1	group: 2	
Nominal categories	category 1	$O_1 1$	$O_1 2$	m_1
	category 2	$O_2 1$	$O_2 2$	m_2
Total		n_1	n_2	N

Table 2.2: An example of a 2 × 2 contingency table.

The corresponding $p-$ value, for a one-sided test is derived from statistical tables, while for a two-tailed test, the $p-$ value is multiplied by two.

The χ^2 test can be used to test for significant difference between $k = 2$ independent samples (groups) across r categories measured at least on nominal scale [Sid57]. Under the null hypothesis $\mathcal{H}_0$, the data are arranged in an $r \times k$ contingency table. Each cell contains the observed frequency O_{ij} for the i-th category in the j-th sample. An example contingency table is shown in Table 2.2. The expected frequency E_{ij} of a cell can be calculated by

$$E_{ij} = \frac{n_i m_j}{N} \ ,$$

where n_i is the total number of observations in category i, m_j is the total number of observations in group j, and N is the overall sample size. The χ^2 statistic measures the discrepancy between observed and expected frequencies given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k=2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \ ,$$

where the sampling distribution follows approximately a χ^2 distribution with degrees of freedom $df = (r-1)(k-1)$ [Sid57]. Leveraging df and the χ^2 value, the $p-$ value can be determined using statistical tables. Finally, $\mathcal{H}_0$ rejected or accepted depending on the significance level α.

The Kruskal-Wallis test is the non-parametric analogue to the analysis of variance (ANOVA) test (Section A.1) [Sid57]. It assesses the null hypothesis that $k \geq 2$ independent samples come from the same population or populations by means of their average. The test assumes a continuous distribution of ordinal data. Before conducting the test, the data has to be transformed into ranks by assigning the lowest rank 1 to the smallest value and the highest rank N to the greatest value in the categorical data [Sid57]. Sometimes ties occur during the creation of ranks. In such cases the mean value of the tied ranks is assigned to the observations. In case of no ties this equation can be used

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \ ,$$

where N is the total number of independent observations in the k samples, n_j is the number of independent observations in jth sample, R_j is the sum of ranks in jth sample. The formula in case of ties observations can be found in the supplements (Section A.2, eq. A.4). The test statistic H in the Kruskal-Wallis test is χ^2 distributed with degrees of freedom of $df = k - 1$, where $k \geq 2$ and being the number of samples [Sid57]. The same procedure as in the χ^2 test is used to determine whether the null hypothesis $\mathcal{H}_0$ is rejected or accepted.

The Mann-Whitney U test can be seen as the non-parametric analogue to the t-test and requires the data to be at least ordinal [Sid57]. It can be applied in the two-sample test scenario to test whether two independent samples with size n_1 and n_2 come from the same population. Similar to the Kruskal-Wallis test, the data is transformed into ranks. However, in the Mann-Whitney U test, the observations from both samples are combined into one set and then ranked [Sid57]. Values are ranked and ties are treated as in Kruskal-Wallis test. Afterwards, the ranked data is ordered by increasing number. The U test statistic is given by

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$
$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 ,$$

where R_1 and R_2 are the sum of ranks assigned to the sample with size n_1 and n_2, respectively. The two formulas do not yield the same result and for statistical testing, it is suggested to use the smaller value [Sid57]. Using the formula $U = n_1 n_2 - U'$, where U' is either U_1 or U_2, it can be easily switched U_1 and U_2. For large sample sizes n, the sampling distribution of U is approximately normally distributed. In that case, it can bee compared to the z-distribution also known as standard normal distribution with mean $\mu = 0$ and unit variance $\sigma^2 = 1$ [Sid57]. Thus, the significance of U is given by

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} ,$$

where μ_U and σ_U are the mean and standard deviation of U, respectively. The formula for tied observations can be found in the supplements (Section A.2, eq. A.5). The probability value, $p-$ value, corresponding to z under $\mathcal{H}_0$ can be taken from statistical tables and is valid for a directed one-sided test. In order to determine the $p-$value for a two-tailed test, double the probability [Sid57].

2.2.2 Multiple hypothesis testing

In transcriptomics studies it is quite common to test for multiple null hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$, e.g., testing for differences between expression levels of multiple genes in parallel. However, performing $m \in (1, \infty)$ independent tests leads to higher Type I error rates [YB99]. In order to control for these false positive discoveries, for example, the Bonferroni or step-up Benjamini and Hochberg (BH) false discovery rate (FDR) correction can be used [Dun61] [BH95] [YB99]. In this thesis I use the latter, BH FDR, to control for the number of false discoveries. Considering m ordered p-values $p_{(1)} \ll \ldots \ll p_{(m)}$ corresponding to m performed statistical tests. The FDR is controlled by rejecting all null hypothesis $\mathcal{H}_0^{(i)}$ where the following applies [BH95]

$$p_{(i)} \ll \frac{i}{m}\alpha \quad \text{with } i \in [1, k] \ , \tag{2.3}$$

where k is the maximum number of rejected $\mathcal{H}_0^{(i)}$ [BH95]. The eq. 2.3 can be used to define FDR corrected p-values, i.e., p - adjusted (padj) value p_{adj}, as [YB99]

$$p_{adj}(p(i)) = \min_{i \leq k}\left(p_{(k)}\frac{m}{k}\right) \quad \text{with } k \in [i, m] \ .$$

A multiple hypothesis test $\mathcal{H}_0^{(i)}$ is considered significant if $p_{adj}(p(i)) \leq \alpha$. The BH FDR method is in comparison to other multiple test correction methods more powerful and less conservative [BH95] and is thus used to correct for FDR in this thesis.

2.2.3 Bayesian correlated t-test

The Bayesian correlated t-test [CB15] is introduced here, despite being classified as a non-parametric tests. It is employed to assess the performance of two ML models trained on the same dataset (Section 2.4.3 and Chapter 4). This test addresses the specific question: "What is the probability that one model is superior to the other, or how likely is $\mathcal{H}_0$?" [Ben+17]. The common null hypothesis statistical tests (NHST) is unable to respond to this question, as it provides the probability indicating the unlikelihood of the observed difference in performance between the two models occurring by random chance under the null hypothesis $\mathcal{H}_0$ [Ben+17].

The Bayesian correlated t-test is a superior choice for this task for the following reasons [Ben+17] [WL16]:

(i) Lack of uncertainty in NHST: NHST does not provide information about the uncertainty of its reported probability.

(ii) Non-identical model performance: The performance of two models is unlikely to be identical, thus $\mathcal{H}_0$ is never true.

(iii) Statistical significance and sample size: Statistical significance can be reached by using just enough samples.

(iv) Probabilistic nature of observed performances: Since observed performances are only estimates, the answer is probabilistic.

The Bayesian correlated t-test can be used to compare the cross-validation (CV) performance results of two models as these observations are dependent and correlated [CB15] [Ben+17]. Let $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ be the difference in CV performance results. Then, the data can be modelled as

$$\boldsymbol{x} = \mathbf{1}\mu + \boldsymbol{v} \ ,$$

where μ is the mean difference of the performance scores and $\boldsymbol{v}_{n\times 1}$ is a noise vector modelled by a multivariate normal distribution $\boldsymbol{v} \sim \mathrm{MVN}(0, \Sigma)$. Internally $\boldsymbol{v}$ considers the variance σ^2 and correlation r between performance results introduced by CV [CB15]. Then, the posterior distribution is then given by

$$p(x|\mu, \Sigma) = \frac{\exp(-\frac{1}{2}(\boldsymbol{x} - \mathbf{1}\mu)^T \Sigma^{-1}(\boldsymbol{x} - \mathbf{1}\mu))}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} \ , \tag{2.4}$$

which is the mean difference of performance between two models evaluated using CV. Considering the two classification models, A and B, the practical probabilities P that (i) $P(A \ll B)$ for $\boldsymbol{x} \in (-\infty, -\tau)$, (ii) $P(A = B)$ for $\boldsymbol{x} \in [-\tau, -\tau]$, and (iii) $P(A \gg B)$ for $\boldsymbol{x} \in (\tau, \infty)$ can be calculated by taking the integral over eq. 2.4. The area defined between $x \in [-\tau, -\tau]$ marks the region where both models are practically equal in their performance [KL15]. This region is also called rope and the interval depends on the evaluation metric and definition of practical equivalence of two models [KL15] [Ben+17].

The posterior distribution can also be used to calculate the uncertainty in the probability by integrating over certain regions of the data so that most of the information is still retained. In addition, like in NHST, a threshold on P can be applied to decide whether one model is superior to the other [Ben+17].

In summary, the Bayesian correlated t-test provides a more sophisticated, probabilistic approach for model comparisons, accounting for uncertainty and practical equivalence. Therefore, it is suited for assessing the relative performance of ML models trained on the same dataset.

Correlation Coefficient	Strength of Correlation
1	Perfect
0.7 - 0.9	Strong
0.4 - 0.6	Moderate
0.1 - 0.3	Weak
0	None

Table 2.3: Assessment of the absolute correlation coefficient score [Ako18].

2.3 Correlation and multicollinearity

In statistics, not only differences between variables are of interest, but also whether there is a linear relationship between two or more variables. Correlation measures the strength and direction of an association between two random samples. Established methods are Pearson's and Spearman's correlation coefficients for parametric and non-parametric data, respectively (Section 2.3.1 and Section 2.3.2).

Since correlation indicates whether there is a dependency between two observations (Table 2.3), it can also be used to remove redundant information in the data to reduce the dimension and increase predictive power in ML models (Section 2.4.3.1.3). A drawback of correlation is that it cannot be used to identify multicollinearity easily, which occurs when a random observation can be expressed by a linear combination of two or more other observations. This is especially true when, e.g., noise occurs in the data. Then these linear combinations might not be as obvious. One solution is to calculate the variance inflation factor (VIF), which will be introduced after introducing Pearson's and Spearman's correlation coefficients in section 2.3.3.

2.3.1 Pearson's correlation coefficient

Pearson's correlation coefficient (PCC) measures the strength and direction of the linear relation between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$. It can be applied if the data is at least continuous and normal distributed. The PCC is defined by [Ueb]

$$r_{xy} = \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x},\boldsymbol{x}} \cdot \sigma_{\boldsymbol{y},\boldsymbol{y}}} \,, \tag{2.5}$$

where cov is the covariance and σ is the standard deviation of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. The correlation coefficient r_{xy} can take values between -1 and 1. Values of $r_{xy} = -1$ or $r_{xy} = 1$ imply a strict linear dependency, while for a value of $r_{xy} = 0$ the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are not linearly dependent. In order to test whether the linear relationship defined by r_{xy} is significant across the population, a null hypothesis $\mathcal{H}_0 : r_{xy} = 0$ can be formulated

[BCB03]. Since its a requirement that both variables are independent and have normally distributed values, the sampling distribution can be approximated by a t-distribution. Thus, Student's t-test can be applied [Ueb] which is defined by

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \;, \tag{2.6}$$

where n is the total sample size. The null hypothesis $\mathcal{H}_0$ is rejected if the p-value is below the significance level α. Then, it can be concluded that there is indeed a linear relationship between the two variables.

Additionally to applying a statistical test, the *confidence interval (CI)* of the correlation value r_{xy} can be determined. It provides further information about the strength of the relationship [BCB03]. In order to calculate the CI, r_{xy} has to be transformed using the Fisher's z-transformation which is defined by [Fis+21] [AG88] [BW00]

$$z = \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) .$$

The distribution of z is approximately normally distributed with $\sigma_z = \frac{1}{\sqrt{n-3}}$ [AG88]. The general formula of the upper and lower bound CI of r_{xy} in z-space is given by

$$\mathrm{CI}_z = z \pm N_{1-\frac{\alpha}{2}} \cdot \sigma_z$$

where $N_{1-\frac{\alpha}{2}}$ corresponds the $100(1-\frac{\alpha}{2})$ percentile of the z-distribution [AG88]. A common chosen value of α is 0.05 which refers to the $\mathrm{CI}_{95\%}$. Using a statistical table, the value is approx. 1.96 [AG88]. Subsequently, the $\mathrm{CI}_{95\%}$ of the correlation value r_{xy} can be calculated by using the inverse Fisher's transformation [AG88] [BW00]:

$$\mathrm{CI} = \frac{\exp^{2z} - 1}{\exp^{2z} + 1} \;. \tag{2.7}$$

The value of $\mathrm{CI}_{95\%}$ can be interpreted as that 95% of the data is within the area of a normal distribution centred around zero.

2.3.2 Spearman's correlation coefficient

In case of non-parametric data the Spearman's correlation coefficient (SCC) can be used [Spe61] instead of PCC to examine the monotonic relation between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$. It is able to handle ordinal, ranked data, while PCC expects the data to be continuous. Therefore, the SCC is defined as the PCC (e.q. 2.5) except that it takes the ranked values $\tilde{\boldsymbol{x}} := \mathrm{rank}(\boldsymbol{x})$ and $\tilde{\boldsymbol{y}} := \mathrm{rank}(\boldsymbol{y})$ as input. Thus, the SCC is given by

$$\rho = r_{xy}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \frac{\mathrm{cov}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})}{\sigma(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}})\;\sigma(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{y}})} \;. \tag{2.8}$$

where $\rho \in [-1,1]$. Negative values indicate a negative correlation and vice versa for positive values. For $\rho = 0$, no correlation exists between the two variables. Conclusively, ρ enables to assess the strength and direction of the relation just like Pearson Correlation. In ranked data it can happen that ties occur. That is the case when two values are equal and would therefore be assigned the same rank. In the scenario of transcriptomics data, measuring the same expression level of gene in multiple observations is not a rare case, especially examining biopsies from patients suffering from similar conditions. The same procedure to correct for tied observations as mentioned in the section 2.2.1 is used. If ties are not assigned the same rank it would cause issues and bias in the statistical analysis. The above equation 2.8 is also able to handle tied ranks.

Hypothesis testing allows to determine whether a significant difference exists. In this case, the null hypothesis $\mathcal{H}_0 : \rho = 0$ and $\mathcal{H}_a : \rho \neq 0$ is tested meaning it is assumed that there is no relation between the two observations, $\boldsymbol{x}$ and $\boldsymbol{y}$ [Sid57]. In case of large samples size $n \geq 10$, the sampling distribution follows a t-distribution [Sid57]. Thus, the test statistic is the same as for PCC stated in eq. 2.6. This also applies to the CI (eq. 2.7).

In some instances, using a *weighted SCC* is more appropriate. For example, if you want to pay more attention to a particular attribute in the data, or between two genes whose expression levels have been aggregated across, e.g., multiple conditions. Then the sample size will not be reflected in the variables. Thus, weighting the variables by the number of conditions may be more appropriate. In order to determine weighted SCC, the *weighted rank* has to be defined first. For a variable $\boldsymbol{\xi}$, the weighted rank is given by [Bai+18]

$$\operatorname{rank}_j^{(w)}(\boldsymbol{\xi}) := \tilde{\boldsymbol{\xi}}^{(w)} = \underbrace{\sum_{i=1}^{n} w_i \mathbf{1}(\xi_i < \xi_j)}_{a_j} + \underbrace{\frac{n+1}{2}\overline{w}_j}_{b_j} \tag{2.9}$$

where a_j is the sum of all weights w_j with $j \in [1,\ldots,n]$ under the constraint $\xi_i < \xi_j$ and b_j handles tied ranks by taking the mean of all weights $\overline{w}_j$ [Bai+18]. The weighted SCC $\rho^{(w)}$ is received, where the superscript denotes the weighted version of the SCC, by inserting eq. 2.9 into eq. 2.8. In case of $w_j = 1\ \forall j$, eq. 2.8 can be used.

2.3.3 Handling multicollinearity with variance inflation factor

In order to determine whether there is collinearity in a data matrix $X^{n \times k+1}$, the VIF can be leveraged [MR19]. Based on the samples, the VIF tries to build linear combinations of the form

$$\boldsymbol{x}_j = \boldsymbol{\beta}_0\mathbf{1} + \boldsymbol{\beta}_1\boldsymbol{x}_1 + \ldots + \boldsymbol{\beta}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{\epsilon}\ , \tag{2.10}$$

where $\boldsymbol{x}_j$ with $j \in [1, k]$ is the jth independent variable described by the other $k-1$ independent variables in X, $\boldsymbol{\beta_0}$ is an intercept term, and $\boldsymbol{\epsilon}$ is an error term [MR19]. In the case of $k = 3$, the variance of the estimated coefficient $\hat{\beta}_3$ of the third variable is related to the correlation between x_1 and x_2. That means, for higher correlation values the variance also increases. For the general case (see eq. 2.10), the variance is given by

$$\sigma^2(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \cdot \frac{\sigma_\epsilon^2}{\sum x_j^2} \quad \text{with } j \in [1, k] \ ,$$

where R_j^2 is the coefficient of determination used in the context of evaluating the goodness of a fit in a regression task, σ_ϵ^2 is the variance of the error, and $\sum x_j^2$ is the sum of squared deviations of the jth independent variable in X [O'b07]. Regression models will be discussed in more detail in the following Section 2.4.3.2 and the proof that the square root of R^2 is the correlation coefficient can be found in the Supplements B.1. The VIF of the jth variable is given by [O'b07]

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \ . \tag{2.11}$$

The higher the values of VIF the more likely it is that there is a multicollinearity amongst the samples [MR19]. The interpretation of VIF values is as follows. A VIF $= 1$ means no correlation, $1 <$ VIF < 5 suggest moderate correlation, and VIF > 5 indicate high correlation [MR19] [O'b07].

In some applications, a strong relation between multiple variables leads to biases in the results or interpretations. To circumvent this issue, the VIF can be calculated and a threshold can be applied. Commonly, the threshold VIF > 5 is used indicating collinearity [MR19].

In summary, the VIF detects multicollinearity, which can result in biased outcomes if not addressed. By applying a VIF threshold, the integrity of the models and downstream analyses can be ensured. In this thesis, the VIF is implemented in a novel feature selection pipeline introduced in Chapter 4.

2.4 Machine Learning

While statistical tests are employed to test for differences or relations between variables, ML is used to reduce the amount of information in the data to its most essential ones and learning patterns to solve clustering and prediction tasks. ML was long a theoretical concept due to the lack of computational resources. Since 1990 [Sie95], computing became more and more efficient leading to the practical realisation and new subfields of ML such as Deep Learning.

ML is commonly divided into three types of learning strategies: unsupervised, supervised, and reinforcement learning. In *unsupervised learning*, the model seeks to identify patterns or structures in the data, with the objective of categorising or representing it in a compact format without knowing the sample labels. Examples are clustering and dimensionality reduction methods. In *supervised learning*, the sample labels are known, and the model learns its parameters by minimising the error between the predicted and ground truth labels. Common applications of supervised learning include classification and regression tasks. Examples for the former are the classification of dog vs. cat images, while for the regression tasks it is weather forecasting. Whereas, in *reinforcement learning*, an agent teaches itself how optimally perform a task by continuously interacting through a feedback-loop with an environment. The feedback is provided in the form of rewards, which the agent attempts to maximise [SB18]. Reinforcement learning is used in various fields, including natural language processing and gaming [Li19].

In this thesis, dimension reduction, clustering, and prediction tasks are carried out in the Chapters 3, 4, and 5 to reveal patterns hidden in the data, identify disease endotypes, construct classifiers on a feature subset, and identify spatial associations in non-communicable chronic inflammatory skin diseases (ncISDs).

2.4.1 Dimension reduction techniques for visualisation

In this thesis, the analysed transcriptomics data and clinical assessments are high-dimensional - at least transcriptomics data - due to the amount of measured expression levels of ten thousands of genes contained on the human genome and number of leveraged clinical traits. In order to reveal interpretable strong/dominant structures and connections in the data sets, an established approach is to project high-dimensional data into a lower dimensional space, while preserving the most important information. The top two or three dimensions, containing the most essential information, which are then visualised. This allows theories to be drawn, which can subsequently be tested.

High-dimensional data $X \in \mathbb{R}^{n \times f}$ is projected to a lower dimensional space $\hat{X} \in \mathbb{R}^{n \times q}$ with $q \ll f$ by creating embeddings containing the most essential information [Mur22]. The resulting data matrix $\hat{X}$ is an approximation of the original data. Data points that are closer to each other in the embedding space share characteristics and are more similar than data points that are more distant. In general, dimension reduction for visualisation is used to reveal hidden structures and drivers of variation in the data.

This section briefly introduces principal component analysis (PCA) [Hot33] and Uniform Manifold Approximation and Projection (UMAP) [MHM18]. Both techniques are used to visualise the analysed high-dimensional transcriptomics and clinical data. Other commonly used dimension reduction tools to visualise transcriptomics data are t-SNE (t-distributed Stochastic Neighborhood Embedding), diffusion maps and latent space of variational Autoencoder (VAE). For the interested reader, more details can be found in [MH08] [Coi+05] [HBT15] [Mur22] [Fos19].

2.4.1.1 Principal component analysis

PCA [Hot33] is a form of unsupervised learning and is applied to reduce the dimensions of, e.g., images, landmarks, or shapes [Gre+22] [Mur22] [WEG87] [TH09]. It aims for representing the data as uncorrelated, orthogonal, linear approximations (i.e. principal components (PCs)) in fewer dimensions that explain the variance in the data [Gre+22].

To compute the PCs, let $\boldsymbol{X} \in \mathbb{R}^{n \times f}$ be a data matrix with n observations and f features (i.e., genes, clinical traits). In order to project $\boldsymbol{X}$ in a lower dimensional space $\boldsymbol{X} \rightarrow \hat{\boldsymbol{X}}$ such that $q \leq n$, a matrix decomposition of $\boldsymbol{X}$ is computed using, e.g., singular value decomposition (SVD) given by [GK65] [Mur22] [WEG87]

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T ,$$

where $\boldsymbol{U}$ is an $n \times f$ orthogonal matrix with left singular vectors, $\boldsymbol{V}$ is a $f \times f$ orthogonal matrix with right singular vectors, and Σ is a $f \times f$ diagonal matrix containing singular values $\sigma \in \mathbb{R}^+$ in descending order. The singular values are the square root of variance, i.e., standard deviation, explained by the features in the data. Using the result of the SVD, the PCs are the left singular vectors scaled by their respective singular values [Gre+22]. In essence, PCA uses SVD to decompose the data matrix, deriving PCs that capture the maximum explained variance.

For visualisation purposes the first two to four dimensions are commonly examined, as they contain the majority of explained variance, assisting in elucidating, e.g., biological distinct groups.. However, for downstream analyses, keeping a higher number of PCs is advised to prevent loss of too much information [LT19]. Hence, by using only p PCs the data matrix can be reconstructed such that $\hat{\boldsymbol{X}} \approx \boldsymbol{X}$. Since PCs are linear approximations of the data, they are more intuitive to interpret than non-linear methods. In conclusion, the elbow method assists in determining the optimal number of PCs to retain.

In summary, PCA reduces data dimensionality, while preserving dominant features, enhancing both intuitive visual interpretation and robust data analysis.

2.4.1.2 Uniform Manifold Approximation and Projection

The UMAP is a non-linear dimension reduction technique, which uses manifold learning and relies on Riemannian geometry and algebraic topology [MHM18]. A manifold is a lower-dimensional topological space of the original data where the neighbourhood of each data point is describable by the euclidean space. UMAP can be used, e.g., for visualisation and as dimension reduction technique in ML tasks. Especially the fact that UMAP represents local structures and preserves the topology of the data can be handy in big data or data with many local topologies [MHM18]. Moreover, UMAP is able to handle big data and can, like PCA, be directly applied on the data without the need of prior dimension reduction [MHM18]. A weakness of UMAP is, that its embedding is not directly interpretable, meaning the gap between local structures does not reflect the true distance. In addition, the application of UMAP on small (< 500 samples), noisy datasets can result in distorted embeddings and should be thus considered with care [MHM18].

This thesis leverages UMAP to visualise high and low dimensional data. The embedding of the latter should be considered carefully. However, these datasets contain many biologically driven topological structures, thus UMAP can be used to reveal them.

2.4.2 Unsupervised clustering

Clustering is a hypothesis free approach to categorise data points into groups based on shared characteristics, while maximising dissimilarity in an unsupervised fashion. This method is especially helpful when the ground truth labels are not known, as it does not require prior knowledge of sample labels [Mur22]. Clustering is widely applied across different research disciplines, including anomaly detection and image compression [TH09].

Common clustering methods for transcriptomics data are partitioning-based, hierarchical, community-detection-based, and density-based clustering [HK99]. Examples are K-Means, Agglomerative, Leiden, and DBSCAN clustering, respectively. This thesis employs K-Means, hierarchical, and Leiden clustering, which are introduced in Section 2.4.2.2.

A main characteristic of clustering algorithms is the definition of clusters based on a metric, e.g., distance, correlation or modularity [Llo82] [BBC04] [TWVE19]. The choice of metric depends on the data and context. For instance, K-Means clustering should be preferably applied on datasets without outliers, while the Leiden algorithms is especially useful in detecting groups in graph-based data. Therefore, the metric determines the how the data points are grouped together and should be chosen based on the type of data.

Many algorithms, such as K-Means and Leiden clustering, require information about the expected number of clusters in the data, which is either provided by a parameter k corresponding to the number of clusters or, indirectly, by a resolution parameter γ. Since it is not known how many clusters are actually in the data, it is common practice to repeat the clustering procedure for various numbers of clusters. Thus, estimating the correct number of clusters is an iterative process and may depend on the clustering algorithm and characteristics of the data.

The decision on the optimal number of clusters can be further supported by clustering metrics such as Silhouette score [Rou87] and Davies-Bouldin Index (DBI) [DB79]. However, the result of these metrics should be considered carefully as they may not fully capture the complexity of the data or are affected by noise. In addition, the number of clusters often depends also on the context. Therefore, domain knowledge is required, as there is no objective ground truth information and no clustering metric answering the question of how many clusters are in the data. In essence, clustering algorithms rely on specific metrics to group data points and, depending on the algorithm, may require prior knowledge of the number of clusters to form.

Clustering transcriptomics data presents challenges due to sensitivity to initialisation, high-dimensionality, the presence of outliers, and noise from irrelevant features. Cluster algorithms, influenced by stochastic processes such as initialisation, yield variable results, necessitating multiple runs with different initial values to ensure stability. Identifying robust and well-separated clusters is further complicated by outliers and noise, as many clustering algorithms are sensitive to these factors [NBG22].While methods like DBSCAN can handle outliers, they are less effective in high-dimensional spaces [XP16]. Given that most clustering algorithms are optimised for low-dimensional data [HK99], dimensionality reduction is essential. The increasing availability of high-dimensional data exacerbates computational inefficiencies due to the "*curse of dimensionality*" [HK99] [KM17]. To address this, I will introduce feature selection methods to refine high-dimensional transcriptomics datasets.

2.4.2.1 Feature selection methods

The high dimensionality of biological data, such as transcriptomics data, can impede the efficacy of clustering algorithms in identifying meaningful patterns. In order to avoid the "curse of dimensionality", feature selection can be applied as a preprocessing step. This approach allows to identify of the most relevant features, reducing the influence of noise and computational complexity [LT19] . Feature selection methods are split into filter, wrapper, hybrid, and embedded approaches. For unsupervised clustering, filter methods are of interest, as they can handle high-dimensional data [Pud+22] [CS14] [BC+14]. All methods are discussed in more detail in the section 2.4.3.1.3 on feature selection in supervised tasks.

Filter methods are one of the most widely applied approaches in unsupervised settings. They assess the relevance of features based on certain statistical measures or metrics and select features independently of the downstream learning algorithm. Established of filtering approaches are similarity measurements, standard deviation, and controlling the mean-variance relationship. For instance, similarity-based feature selection methods such as maximal information compression index (MICI) [MMP02] and PCC select features based on their similarity to other features in the dataset.

However, filtering methods may not be ideal for high-dimensional, noisy data, such as the transcriptomics data analysed in this thesis, as they are sensitive to noise and outliers. Further challenges in transcriptomics datasets are the sparsity, redundancy, and interactions between genes. Ideally, a feature selection method would account for these challenges to improve clustering outcomes.

An established feature selection approach for single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics (ST) data is to control the mean-variance relationship. This method selects features (genes) expressing a high variability across samples, as they may contain meaningful biological information [LT19] [Bre+13]. Variations are described in [Stu+19] [Zhe+17] [Sat+15]. The method presented in Zheng et al. (2017) [Zhe+17] is briefly introduced. Let $\boldsymbol{X} \in \mathbb{R}^{s \times f}$ be a data matrix with features f and samples s. Then, the mean μ_f and variance σ_f^2 of each feature across all samples can be determined. Subsequently, the dispersion value can be calculated, which is defined as $\delta_f = \frac{\sigma_f^2}{\mu_f}$. Based on the mean μ_f, the features are placed into 20 bins, where each bin contains features with similar average values. Then, for each bin, the normalised dispersion value for each feature is calculated by

$$\delta_{il}^{\text{norm}} = \frac{\delta_i - \text{med}(\delta)_l}{\text{MAD}(\delta)_l} \quad \text{with } l \in [1, \ldots, 20] \ ,$$

where i denotes the ith feature in bin l, med is the median of all dispersion values in bin l, and MAD is the median absolute deviation (MAD) of all δ in bin l. The normalised dispersion values δ^{norm} are sorted and the top h features having the highest δ^{norm} are selected and denoted as highly variable genes (HVGs). This method is implemented in the python package SCANPY [WAT18].

2.4.2.2 Clustering algorithms

Clustering algorithms generally require a distance metric to measure the similarity between groups and samples, with Euclidean distance being one of the most commonly used. The selection of the appropriate clustering algorithm depends on the data characteristics, as factors such as outliers, noise, and high dimensionality can affect the clustering results. The following section presents the clustering algorithms used in this thesis.

K-Means clustering groups data points into k clusters by assigning each point to the nearest centroid and minimising the Euclidean distance or the variance within each cluster [Jai10]. The algorithm operates iteratively. In the first step (i), centroids are set randomly, then in step (ii) data points are assigned to the nearest centroid, and in step (iii) centroids are recalculated as the mean of their assigned points. Steps (ii-iii) are repeated until convergence. However, K-Means is non-deterministic, as its results depend on the initialisation of centroids, often converging to local minima. Running the algorithm multiple times can stabilise the clusters. Additionally, K-Means assumes spherical, similarly sized clusters, which may not match real-world data distributions. The method was independently introduced by [Ste+56] and [For65]. The default implementation in Scikit-learn is based on Lloyd's algorithm [Llo82] and improved to K-means++ by Arthur et al. (2007) [AV+07], enhancing centroid initialisation and convergence. As it does not scale for large datasets, Bahmani et al. (2012) addressed this limitation with K-means|| [Bah+12], which is not yet available in Scikit-learn [Ped+11]. In this thesis, K-means++ from Scikit-learn is used for gene clustering (Chapter 3).

There are two types of Hierarchical clustering, agglomerative and divisive [TH09]. Divisive clustering starts with all data points in a single cluster and iteratively divides it into smaller clusters based on similarity measures, progressing from one cluster to two, then to four, and so on, until each data point forms its own cluster. This method builds a tree-like structure where cluster divisions create pairwise branches, and is therefore known as a top-down approach. Conversely, agglomerative clustering operates in a bottom-up manner. Initially, each data point is treated as an individual cluster.

Pairs of clusters are then iteratively merged based on their similarity until all data points are combined into a single cluster. The resulting hierarchical relationships can be visualised using a *dendrogram*, which depicts the merging process in a tree-like diagram.

Agglomerative clustering is used to reveal relationships between clusters by iteratively merging pairs of similar clusters based on a dissimilarity measure and a chosen agglomerative, or *linkage*, method [TH09]. Common linkage methods are, for instance, Ward's method [WJ63], complete linkage, and single linkage. The complete linkage method determines cluster proximity based on the maximum distance between data points from different clusters, ensuring that all points within a merged cluster remain relatively close to one another. In contrast, the single linkage method considers the minimum distance between data points from two clusters [Joh67]. The objective of Ward's method is to minimise the data points total variance inside a cluster [WJ63]. Agglomerative hierarchical clustering in this thesis is applied using the `hclust` function from the R package "stats" [Tea22], the linkage method "ward.D2" and dissimilarities measured by $1-$cosine distance.

The Leiden algorithm [TWVE19] is an improved version of the Louvain algorithm [Blo+08], as it is faster and scales on datasets of any size. In addition, it guarantees to build well-connected groups [TWVE19] [HA+21]. Although the Leiden algorithm belongs to the agglomerative clustering methods, it differs from hierarchical agglomerative clustering in several key aspects. Leiden clustering requires the input data in the form of a network or graph, where nodes represent entities (e.g., cells) and edges represent relationships or interactions between them. It aims for detecting communities, or dense groups of nodes, within networks or graphs [For10] [POM+09]. To achieve this, Leiden optimises the modularity, which is a quality of function [NG04][TWVE19]. It controls the size and granularity of the clusters, i.e., the communities, by a user-defined resolution parameter γ. The Leiden algorithm is split into three main phases [TWVE19]:

1. Find partitions by moving nodes from one community to another.
2. Refine partitioning by optimising modularity.
3. Built a network by aggregating the refined partition.

These phases are repeated until no further refinements can be made. The Leiden algorithm is widely used in clustering high-dimensional data, such as scRNA-seq data [DRS18] [Fre+18] [WR16]. In this thesis, I use the Leiden clustering algorithm implemented in SCANPY [WAT18] and in the python module leidenalg [TWVE19] to cluster scRNA-seq and bulk RNA-sequencing (RNA-seq), respectively.

2.4.2.3 Clustering metrics

Clustering metrics are used to evaluate the performance of clustering models or to compare the results of two cluster algorithms. Implementations of the metrics are accessible via the module Scikit-learn [Ped+11] in Python.

The silhouette score assesses the similarity of data points within the same cluster to those outside in other clusters. It should be used if the aim is to maximise the compactness of a cluster and separation to other clusters [Rou87]. The silhouette score can be determined by

$$\begin{aligned} S(i) &= \frac{(b(i) - a(i))}{\max(a(i), b(i))} && \text{if} |C_I| > 1 \\ S(i) &= 0 && \text{if} |C_I| = 1 \ , \end{aligned}$$

where $a(i)$ is the average distance between a data point i and all other data points j with $j \neq i$ in the same cluster C_I while $b(i)$ is the smallest average distance between data point i and data points in other clusters. The number of data points in C_I is given by $|C_I|$. The silhouette score is zero if only one data is in cluster C_I as no distance $a(i)$ between $i \in C_I$ and other data points $j \in C_I$ can be calculated. The silhouette score $S(i)$ ranges from -1 to 1, with a low score indicating poor clustering of the data points and a high score indicating that the data points are well clustered. Any kind of distance metric can be used. However, the choice should be always tailored to the data - the euclidean distance may not be the right choice for high dimensional data as in higher dimensions the distances become uniform and it can fall for the "curse of dimensionality" [AHK01] [Bey+99].

Another clustering metric is the DBI, which was introduced by David L. Davies and Donald W. Bouldin [DB79]. Unlike the silhouette score, the DBI is the measure of ratio between inter- and intra-cluster distances. Let $\boldsymbol{x} \in \mathbb{R}^{1\times n}$ and $\boldsymbol{y} \in \mathbb{R}^{1\times n}$ be vectors assigned to clusters C_I and C_J, respectively, then the fraction between the inter- and intra-cluster distances can be built by

$$R_{I,J} = \frac{dist(\boldsymbol{x}_i, \overline{\boldsymbol{x}}) - dist(\boldsymbol{y}_j, \overline{\boldsymbol{y}})}{dist(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}})} \ ,$$

where $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ are the centroids of the clusters C_I and C_J, respectively. The DBI minimises the dispersion inside the clusters and at the same time maximises the separation between clusters and is defined by

$$\text{DBI} = \left(\frac{1}{k}\right) \sum_{i=1}^{k} \max_{i \neq j}(R_{ij}) \ ,$$

where k is the number of clusters in the data. A lower value implies a better clustering of the data points, while a higher value indicates the opposite.

A metric, to assess the similarity of two clustering results, is Adjusted Mutual Information (AMI). It is a variation of mutual information (MI) and accounts for detecting mutual agreement by chance [Rom+16]. Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a set of data points and $U = \{U_1, \ldots, U_k\}$ a random cluster partition with k clusters and $V = \{V_1, \ldots, V_n\}$ with n clusters. The probability $P_U(i)$ that a data point $s_i \in \mathcal{S}$ belongs to cluster U_i can be determined by

$$P_U(i) = \frac{|U_i|}{n} \ .$$

Subsequently, the entropy H of the cluster U can be calculated by

$$H(U) = -\sum_{i=1}^{k} P_U(i) \log(P_U(i)) \ .$$

Similarity, the probability and entropy of cluster V can be determined. Then, the MI can be corrected by

$$\text{AMI}(U, V) = \frac{\text{MI}(U, V) - E[\text{MI}(U, V)]}{\max(H(U), H(V)) - E[\text{MI}(U, V)]} \ .$$

Values of AMI are in the range of AMI $\in [0, 1]$, where higher values indicate a better agreement between the two clustering. It should be used when it is expected that the ground truth clustering contains clusters of different size and small clusters [Rom+16].

2.4.3 Supervised regression and classification

Supervised learning is a ML approach that utilises labelled data to train models to solve regression and classification tasks. In contrast to unsupervised learning, which identifies patterns or clusters without pre-defined labels, supervised learning requires ground truth values $y \in \mathcal{Y}$ corresponding to the data points $\boldsymbol{X} \in \mathbb{R}^{n \times f}$. Depending on the task, the label space $\mathcal{Y}$ varies. For classification, $\mathcal{Y} \in [1, \ldots, k]$ represents k classes, while for regression $\mathcal{Y} \in \mathbb{R}^n$ represents continuous values. Supervised learning enables the development of models that predict discrete classes or continuous outcomes by leveraging labelled data tailored to the specific task.

Supervised learning models aim to predict output variables by establishing a mathematical relationship between input data and their corresponding labels. A supervised learning model maps an independent (input) data point $\boldsymbol{x}_i$ to its response (output) variable y_i. For example, in a linear regression model, this relationship is expressed as $\boldsymbol{y} = \beta_1 \boldsymbol{x} + \beta_0$, where β_1 (slope) and β_0 (y-intercept or bias) are parameters learned during training. The training process aims to optimise these parameters to best describe the relationship between $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$. Learning parametrised relationships between inputs and outputs, supervised models achieve accurate predictions for regression and classification tasks.

The optimisation of supervised learning models involves minimising errors using an objective function. In real world application, models often fail to perfectly describe the data, leading to residual errors $\boldsymbol{\epsilon}$ between the true label $\boldsymbol{y}$ and the estimated value $\hat{\boldsymbol{y}}$ occur. The objective of model training is to minimise these errors through an optimisation process. This involves defining an objective function, which is either minimised or maximised, and is composed of a cost or loss function and a penalisation term [Mur22] [TH09]. The choice between a cost or loss function depends on whether the focus is on aggregated or single errors, respectively. In the objective function, the penalisation term, such as L2-regularisation, prevents the model from overfitting, ensuring its generalisability [Mur22] [TH09]. For example, the objective function $\sum_{i=1}^{n} \epsilon^2$ can be approximated by $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$, where $\lambda \sum_{i=1}^{n} \beta_i^2$ is the L2-regularisation, representing the penalisation term [TH09]. The optimal solution is found at the zero crossings of the derivative of the objective function. Thus, model optimisation balances error minimisation and regularisation, ensuring accurate and generalisable predictions.

Supervised models produce outputs that can be adapted for classification tasks using decision thresholds. The model's output $\hat{\boldsymbol{y}}$ is a vector of continuous values. These can be mapped to probabilities $P(\boldsymbol{x}_i) \in [0, 1]$, which are converted into binary outputs using a decision threshold θ. In the described scenario, the decision function can be defined as

$$G(\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } P(\boldsymbol{x}_i) \geq \theta \\ 0 & \text{if } P(\boldsymbol{x}_i) < \theta \ , \end{cases} \tag{2.12}$$

where an established decision threshold is $\theta = 0.5$. This transforms the regression into a classification model. Desired is a prediction probability outside a user-defined uncertainty area (e.g., ± 0.1 around $P(\boldsymbol{x}_i) = 0.5$) [TH09] In summary, applying decision thresholds to continuous model outputs, regression models can be adapted to classification tasks.

The performance of the trained model is evaluated using a test set, which contains data not exposed during training. Evaluation metrics such as accuracy, precision, recall, and F1-score, assess the model's ability to generalise to unseen data. To prevent overfitting, it is good practice to divide the dataset into a training, validation, and test set. The training set is used to estimate the model parameters by minimising a cost or loss function, while the validation set is used to monitor the model's performance during training to refine the model. Overfitting occurs, when the chosen model fits the training set perfectly, but performs poorly on unseen data. In addition, techniques such as CV use the validation set to select the best model (section 2.4.3.3). The test set, withheld during training, is reserved for the final evaluation, ensuring an unbiased measure of the model's generalisability.

This thesis applies supervised learning to binary classification tasks using both regression and classification models. The following sections introduce the relevant methods.

2.4.3.1 Data preparation

Data preparation is essential for ensuring ML models can process and learn from the provided information. In real world applications, data is provided in different forms such as strings, floats, and integers. Since most ML models operate only on numerical data, preprocessing steps are required to transform raw data into a format suitable for model input. These steps include encoding categorical variables, feature scaling, handling missing data, addressing class imbalance, and selecting relevant features. Proper data preparation ensures that models are trained effectively, yielding accurate and generalisable predictions. Many datasets include categorical variables, which consist of labels with a finite set of values. In order to enable ML models to interpret this information, these variables must be encoded into numerical representations. Common encoding techniques include one-hot encoding or label encoding [TH09]. This transformation preserves the information contained in categorical variables and enables the model to interpret and learn from the encoded information effectively.

Variations in measurement units or magnitudes between features can adversely affect model performance. Feature scaling or normalisation ensure uniformity across features. They transform variables in the data having different measurement units or orders of magnitude to the same scale, while preserving their relative relations and distributions. This ensures that all features contribute equally to the model. It thereby prevents the potential for extreme values to be considered as more important than others, as this could lead to systematic errors in the model [TH09]. In this thesis normalisation and min-max scaling are employed. The latter transforms the features of a variable between 0 and 1 by

$$\boldsymbol{x}_{scaled} = \frac{\boldsymbol{x} - \boldsymbol{x}_{min}}{\boldsymbol{x}_{max} - \boldsymbol{x}_{min}} \, . \tag{2.13}$$

Specific normalisation techniques for each type of analysed transcriptomics data is presented in the sections 2.5.1, 2.5.2, and 2.5.3.

Missing data is a common issue in medical records and other real-world datasets, and the majority of statistical and ML models require complete data. There are three categories of missing data [HT22] [BS14]:

- Missing completely at random (MCAR): Missingness is independent of observed and unobserved data.
- Missing at random (MAR): Missingness depends on observed variables.
- Missing not at random (MNAR): Missingness depends on unobserved variables or the value itself.

The first two categories, MCAR and MAR, can be handled by removing variables with a high proportion of missing values ($> 90\%$ [MD+19]) and by imputation techniques such as K-nearest neighbour (KNN) and regression [Zha12] [MD+19] [Kan13]. They replace the missing values using the information of the measured observations or correlated variables [Zha12] [MD+19]. Especially in the case of MCAR, the imputation does not introduce any bias as the variable with missing data is independent from the others [Kan13]. In case of MAR, the missing data can be explained by other variables [BS14]. MNAR scenarios, where missingness is tied to unobserved factors, requires further investigation and is more challenging to solve. Advanced ML models can also estimate missing values in an unbiased fashion, when dealing with MNAR scenarios [Kan13]. Thus, depending on the applicable scenario, the imputation method has to be chosen accordingly.

Class imbalance occurs when the dataset contains unequal representation of the k classes, with the majority class having significantly more instances than the minority class. This imbalance can impact model performance, as ML algorithms may be biased towards the majority class and hinder the model to generalise well. Strategies such as oversampling the minority class, undersampling the majority class, or using synthetic data generation techniques (e.g., Synthetic Minority Oversampling Technique (SMOTE) [Cha+02]) can address this issue and improve model generalisation.

Another crucial data preparation step is the selection of relevant features, as they strongly influence the performance of a model. Irrelevant or redundant features can increase computational complexity and lead to overfitting, while selected features can enhance model accuracy and interpretability. Techniques for feature selection, including filtering techniques [KR92], wrapper [SKZ10] [Gho+20] and embedded [Szy+09] feature selection methods, are explored in section 2.4.3.1.3.

Data preparation is essential for building robust and accurate ML models. By encoding variables, scaling features, handling missing data, addressing class imbalance, and selecting relevant features, it can be ensured that model performance is improved and that the resulting predictions are meaningful, interpretable, and generalisable. In the following sections, techniques applied in this thesis are presented.

2.4.3.1.1 Encoding of the dependent variable

Many ML algorithms, excluding, e.g., decision-tree based models, require categorical labels to be presented numerically in the form of zeros and ones. The term "categorical labels" refers to the dependent variable, which is defined as a finite set of values or classes. In order to transform these categorical labels into numerical representations, encoding techniques such as ordinal and one-hot encoding are applied [Mur22].

The encoding methods are suitable for specific types of categories. A distinction is made between ordinal and nominal categories. Ordinal variables contain information of the intrinsic ordering or ranking among their categories, which must be preserved during encoding. For these variables, techniques such as *Ordinal encoding* or *Label encoding* can be used. These methods assign a unique numerical value to each category, maintaining the original order of the variable. In contrast, nominal variables lack any intrinsic ordering among their categories. For these, *one-hot encoding* can be leveraged. This method creates a binary variable for each category, ensuring no ordinal relationship is implied. Each category is represented by a separate binary column (often referred to as dummy variables), and the number of new variables corresponds to the number of distinct categories in the original variable.

In this thesis, Label encoding is used, which is implemented in the Python package Scikit-learn [Ped+11] to transform the binary class labels into numerical values.

2.4.3.1.2 Handling imbalanced data

Imbalanced datasets, where classes are unequally represented, are a common challenge in real-world applications. Ideally, all classes in a dataset should have approximately the same number of samples to ensure balanced learning. When this condition is not met, ML models tend to prioritise the majority class, leading to poor performance on the minority class. Such dataset are called *imbalanced* and can severely affect a model's generalisation ability.

ML models are often optimised by maximising the predictive accuracy requiring all classes to be of equal size. However, for imbalanced datasets, alternative evaluation metrics, such as area under the curve (AUC) or the Receiver Operating Characteristics (ROC) convex hull, are better suited as they also optimise the low detection rate of the minority class [Swe88] [DHS01] [Bra97] [Lee00] [PF01].

Several strategies have been developed to mitigate the effects of class imbalance. These include resampling methods, such as over-sampling, under-sampling, SMOTE [Cha+02], K-Means SMOTE [DBL18], and Support Vector Machine (SVM) SMOTE [NCK11]. Another effective approach is to modify the loss function by, e.g., assigning class-specific weights during training [Aur+19]. These strategies aim to ensure adequate representation of all classes in the learning process.

Under-sampling reduces the amount of samples in the majority class, thereby increasing sensitivity of the minority class. However, this may lead to potentially discarding valuable information, as not all majority class samples are used. In contrast, over-sampling replicates minority class samples. While this addresses the imbalance, it does not add new information to the class and may lead to overfitting [MT14]. Thus, both techniques have limitations that may impact the model's overall performance.

As over- and under-sampling have inherent limitations, advanced methods such as SMOTE aim to overcome these by generating synthetic data to improve class representation. SMOTE is a statistical method, which generates synthetic samples to balance class representation more effectively. It combines under-sampling of the majority class and over-sampling of the minority class by creating artificial data points [Cha+02]. These synthetic samples are created by randomly choosing a sample from all observations. Then, the difference between the selected sample and its k-nearest neighbour observations is calculated. The difference is multiplied by a weight w between 0and 1 and added to the sample under consideration. The generation of the new synthetic samples is described by

$$\boldsymbol{s}_i^{(synthetic)} = \boldsymbol{s} + w \cdot (\boldsymbol{s}_i - \boldsymbol{s}) \qquad\qquad \text{with } i \in [1, \ldots, k] \ \wedge \ w \in [0, 1] \, .$$

Combining both, over- and under-sampling, overcomes the limitation of over-sampling and yields higher performance than using under-sampling alone [Cha+02]. Although SMOTE effectively addresses the between-class imbalance problem, it fails in cases of within-class imbalance and noise.

Advanced SMOTE variants, such as K-Means SMOTE [DBL18] and SVM SMOTE [NCK11], address specific limitations of the original technique. K-Means SMOTE applies clustering, filtering, and over-sampling to handle skewed and noisy data. Clustering identifies dense regions in the data, and filtering retains clusters with a high proportion of minority samples. SMOTE is then applied by generating synthetic samples specifically in clusters with sparse minority class samples.

SVM SMOTE [NCK11] extends SMOTE by generating synthetic samples along the decision boundary using support vectors from an SVM trained on the original dataset [HWM05] [Wan08]. This approach prioritises regions near the decision boundary, where classification accuracy is most crucial. In contrast to SMOTE, SVM SMOTE selects k-nearest neighbour samples along the decision boundary of the interpolated and extrapolated line between nearest neighbours support vectors of the minority class, ensuring more targeted over-sampling. Extrapolation is used to shrink the distance between minority and majority classes when the proportion of nearby majority samples is below 50% [NCK11]. Both K-Means SMOTE and SVM SMOTE address some of the challenges of SMOTE, such as noise and within-class imbalance, making them more robust for complex datasets.

In this thesis, SVM SMOTE, which is implemented in the python package Imbalanced-learn [LNA17], is used to account for class imbalance before performing ML-based feature selection and training of the model. By leveraging SVM SMOTE, the data preparation pipeline ensures that the classifier effectively learns patterns from both majority and minority classes.

2.4.3.1.3 Supervised feature selection

Feature selection is an essential preparatory step of predictive modelling. This is particularly challenging in biological datasets, where high-dimensional feature spaces and small samples sizes are common, and it is desired to avoid the "curse of dimensionality". Using all features can lead to overfitting, where the model captures noise rather than meaningful patterns. By identifying a relevant feature subset, feature selection enhances computational efficiency, reduces resource usage, and mitigates overfitting [Pud+22]. The optimal feature subset carries the most relevant information for the defined task and is given by the following definition [KJ97].

Definition 1. *Given a classifier model and a dataset with features $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a distribution drawn from the labelled samples, the optimal feature subset $\boldsymbol{X}_{opt}$ is defined by the maximal performance metric score of the classifier.*

The optimal feature set can be selected through various feature selection and reduction techniques, which also remove noise and redundant information. This avoids overfitting, improves the computational efficiency, and enhances interpretability of the ML model. Common feature selection techniques are filter, wrapper, and embedded methods.

Filter methods, such as correlation (Chapter 2.3), hypothesis tests (chapter 2.2) and Relief [KR92] [Kon94], assist in selecting feature subsets. by ranking class-dependent and -independent features. These methods are independent of the predictive model and scale well with high-dimensional data. However, class-independent methods produce general feature sets that are not tailored to specific prediction tasks [Pud+22]. In contrast, class-dependent methods enhance model performance by selecting task-specific features, but they increase computational costs and may overfit, particularly in small or imbalanced datasets [Pud+22]. Additionally, some filter methods, such as correlation and hypothesis tests, do not account for feature redundancy or interactions, which are important to consider in biological data. Alternative filter methods such as BOOST [Wan+10], CMIFSI [Lia+19] and FS-RRC [Li+20] address these limitations, focusing on pairwise interactions. Some newer methods explore higher-order interactions, albeit with greater computational complexity, as they do not scale to high-dimensional data [Pud+22]. Overall, filter methods are useful as initial feature reduction step due to their simplicity and interpretability.

Wrapper methods, which combine heuristic search strategies with classifier algorithms, offer a more tailored approach to feature selection. These methods iteratively evaluate feature subsets based on model performance [Pud+22], stopping once a predefined criterion, such as performance plateau or feature limit, is met. Compared to filter techniques, wrapper methods achieve superior performance [Inz+04] [Wah+18] [Gho+20] by considering feature interactions and redundancies, which are of significant importance in biological data. Wrapper methods require significant computational resources and are not suitable for high-dimensional data [CS14] [BC+14]. Additionally, the selected feature subset depends on the classifier model used [JKP94], which can bias the selection and lead to overfitting, while making it difficult to trace feature contributions. Examples of wrapper methods include sequential feature selection (SFS) [KJ97] and recursive feature selection with cross-validation (RFSCV) [Guy+02] [Zha+13], with the latter being more robust to overfitting due to its internal CV. While computationally intensive, wrapper methods provide superior performance, particularly for identifying task-relevant features in lower-dimensional datasets.

Embedded techniques are integrated in the classifier algorithm, thereby influencing a feature's contribution through techniques such as regularisation (e.g., L1-regularisation, Appendix B.2) and tree-based models (e.g., Random Forest, Chapter 2.4.3.2) to assign weights or penalties to features. These methods combine the advantages of filter and wrapper approaches, offering computational efficiency and superior performance due to their interaction with the classifier. While tree-based decision models are capable of considering higher-order interactions in datasets that are not high-dimensional, they cannot handle redundancy. In contrast, regularisation models are able to handle redundancy but are not suited to deal with interactions [BG+17] [Lun+04]. In essence, it is a decision between considering feature interactions or being resistant to redundancy. With regard to the application to high-dimensional data it is better to preselect the feature space before using embedded feature selection methods [Pud+22] [Szy+09] [SKZ10].

Hybrid approaches combine different feature selection methods, such as integrating filter and wrapper techniques. In general, they achieve higher performance and overcome individual limitations [Gho+20]. Feature reduction methods, which aim for reducing the dimension of the data by building linear or non-linear combinations of the features, include techniques such as are PCA [Hot33] [WEG87], UMAP [MHM18], and VAE [HZ93]. More detailed descriptions of PCA and UMAP can be found in section 2.4.1. Both, hybrid approaches and feature reduction methods, provide robust solutions for handling complex, high-dimensional datasets effectively.

In this thesis, multiple supervised feature selection approaches are combined to identify a relevant feature subset for binary classification tasks (Chapter 4.2.3).

2.4.3.2 Regression and classification models

Selecting an appropriate ML model is crucial for the overall performance and reliability of predictions. The choice of model should align with the specific task and characteristics of the data, as a mismatch can result in poor performance. For example, applying a non-linear model to linear data would lead to overfitting, where irrelevant features and noise are captured instead of meaningful patters. This reduces the model's ability to generalise to unseen data. Thus, understanding the data characteristics and application domains of different ML models is important for optimal performance.

ML models operate as functions mapping an input $\boldsymbol{X}$ to its corresponding output $\boldsymbol{y}$. The input matrix $\boldsymbol{X}$ represents the independent or predictor variables and is also referred to as design matrix, while the output vector $\boldsymbol{y}$ is the dependent or response variable. In order to find the best model, an optimisation problem is solved. Besides learning the coefficients β of a model, additional parameters can often be adjusted. These are called hyperparameters and define a model's complexity and influence the learning behaviour of the model. In general, ML models map inputs to outputs, with optimisation involving both model coefficients and hyperparameters to enhance performance.

This section provides a brief overview of regression and classification models. Multiple linear regression is introduced first due to its simplicity and forming the foundation of many other ML models, even though it is not explicitly applied in this thesis.

Similar to correlation analysis, linear regression examines the relationship between a dependent variable $\boldsymbol{y}$ and an independent variable [BCB03]. They express the relationship in form of equations, enabling the estimation of the regression coefficients $\beta \in \mathbb{R}^{f \times 1}$. These coefficients describe how the independent variables contribute to the dependent variables $\boldsymbol{X} \in \mathbb{R}^{n \times f}$, with the aim of minimising the error ϵ, defined as the difference between the fitted (predicted) variable $\hat{\boldsymbol{y}} \in \mathbb{R}^{n \times 1}$ and dependent (true) variable $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$, as given by $\boldsymbol{y} = \hat{\boldsymbol{y}} + \epsilon$. The linear regression model for multiple, independent variables (observations) is given by

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\beta + \epsilon \ , \tag{2.14}$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times (f+1)}$ is a design matrix containing the intercept term and observations of $\boldsymbol{X}$. The goal is to minimise the error ϵ between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ by

$$\min\left(\sum_{j=1}^{n} \epsilon_j^2\right) = \min\left(\sum_{j=1}^{n}(y_j - \hat{y_j})^2\right) = \min\left(\sum_{j=1}^{n}(y_j - \boldsymbol{x}_j\beta)^2\right) . \tag{2.15}$$

In other words, an optimal set of coefficients β is sought, which describes best the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. The minimum can be determined by the zero-crossing of the first derivative of eq. 2.15. Under the assumption that $\boldsymbol{X}$ is non-singular, the estimated coefficients β can be obtained by [TH09]

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \ .$$

Given the unique solution of $\hat{\beta}$, the value for the ith data point $\boldsymbol{x}_i$ can be predicted by $\hat{y}_i(\boldsymbol{x}_i) = \boldsymbol{x}_i\hat{\beta}$ [TH09]. The fitted value $\hat{y}_i(\boldsymbol{x}_i)$ can be mapped to $P(\boldsymbol{x_i}) \in [0, 1]$ and subsequently converted into class labels by applying a decision rule using eq. 2.12. For binary classification, the threshold θ is often set to 0.5. That means predicted values $\hat{y}_i \geq 0.5$ are assigned to class 1 and $\hat{y}_i < 0.5$ to class 0.

In order to determine the goodness of the fit, the coefficient of determination R^2 can be calculated. It assesses the proportion of explained variance of the fitted values $\hat{\boldsymbol{y}}$ to the overall variance [Ueb] by

$$R^2 = \frac{\sum_{j=1}^{n}(\hat{y_j} - \overline{y}_j)^2}{\sum_{j=1}^{n}(y_j - \overline{y}_j)^2} \ . \tag{2.16}$$

The values of R^2 can vary between 0 and 1 where $R^2 = 1$ indicates no variance between the estimated $\hat{\boldsymbol{y}}$ and true values $\boldsymbol{y}$. The square root of R^2 is the correlation coefficient. The proof can be found in the Appendix B.1. In summary, multiple linear regression enables to explore and quantify relationships between variables.

Logistic regression is a non-linear regression model used to predict the probability of a binary (dichotomous) variable $\boldsymbol{y}^{n\times 1} \in [0, 1]$ based on one or more independent variables $\boldsymbol{X} \in \mathbb{R}^{n\times(f+1)}$. The model is defined as

$$p(y = 1 \mid \boldsymbol{X}) = \frac{1}{1 + \exp(-\beta \boldsymbol{X})} \ .$$

This equation uses a sigmoid function to model the probability of class membership, which maps the output to the range $[0, 1]$ [Mur22]. To classify the outcome, a threshold θ is applied to the predicted probability (eq. 2.12), with values above θ assigned to one class and below to the other (e.g., $\theta = 0.5$). In order to prevent overfitting and potentially reduce the number of features, a regularisation term can be incorporated into the model [Lee+06] [Ng04] (see Appendix B.2). Overall, logistic regression provides probabilistic predictions, which can be further enhanced through regularisation to prevent overfitting. It is used in my feature selection pipeline, which is described in Chapter 4.2.3.

The KNN classifier is a widely-used algorithm in ML, due to its simplicity and effectiveness in classifying data points based on their similarity to existing samples. The model assigns the class of a given data point based on the properties of its neighbouring data points within a defined feature space. Specifically, the KNN algorithm classifies a data point $\boldsymbol{x}$ based on the majority class among its k-nearest neighbours with $k \in [1, \infty)$. These neighbours are defined as the k-closest samples, sharing similar features in an n-dimensional space. The similarity between two data points, $\boldsymbol{x}$ and $\boldsymbol{x'}$, is measured by the Mahalanobis-distance, which is defined as

$$d(\boldsymbol{x}, \boldsymbol{x'}) = \sqrt{(\boldsymbol{x} - \boldsymbol{x'})^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{x'})} \ ,$$

where the $d(\boldsymbol{x}, \boldsymbol{x'})$ quantifies the distance between the data points and $\boldsymbol{M}$ is a positive definite matrix [Mur22]. Notably, for $\boldsymbol{M} = \boldsymbol{I}$, the Mahalanobis-distance simplifies to the Euclidean distance. Given the distance between all data points, the sets $S_k(\boldsymbol{x})$ of k-nearest neighbours of each data point $\boldsymbol{x}$ can be found. Subsequently, the k-nearest neighbour fit can be conducted, which is defined as

$$p(y = c|\boldsymbol{x}) = \frac{1}{k} \sum_{n \in S_k(\boldsymbol{x})} \mathbb{I}(y_n = c) \ ,$$

to predict the average response of data point $\boldsymbol{x}$ to its k-nearest neighbours for class c [Mur22]. The KNN is used to add new data points to pre-existing clusters in Chapter 3. This application leverages the algorithm's capacity to identify samples with similar characteristics and associate them with the most appropriate group.

The Random Forest model is an ensemble learning method that combines multiple decision trees to enhance prediction accuracy. A decision tree is a hierarchical structure that can be envisioned as an upside-down tree, consisting of nodes and branches. Each node represents the decision/answer to a condition/question, with general decisions made at the highest level of the tree structure, while more specific conditions are queried at deeper levels. The final leaf node delivers the final decision based on a series of queries.

The Random Forest algorithm uses two techniques, bagging and feature bagging, to build uncorrelated decision trees. Bagging, or bootstrap aggregation, generates new datasets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ by randomly selecting samples from the original training set $\boldsymbol{X} \in \mathbb{R}^{n \times f}$ with replacement [TH09] [LUW20]. This technique yields more accurate predictions, as the results of the m decision trees are aggregated and majority voting, i.e., selecting the prediction of an input sample occurring the most, is applied. This takes into account the variance in the data [TH09]. Feature bagging involves repeatedly selecting random subsets of features for each decision tree, ensuring the creation of uncorrelated decision trees [Mur22].

The Random Forest model has three hyperparameters, i.e., number of trees, depth (number of nodes), and number of sampled features. These parameters control the complexity of the tree structure and influence the model's performance. In summary, the Random Forest algorithm is an ensemble of decision trees, where each tree is trained and evaluated on a different subset of data and features, improving robustness and prediction performance. In this thesis, the Random Forest classifier is used as part of an embedded feature selection method due to its ability to assess feature importance using metrics such as Gini Importance [TH09], see Chapter 4.

The XGBoost (short for extreme gradient boosting) algorithm is an ensemble tree of sequentially arranged weak learners, such as decision trees, minimising an objective function.

This objective function consists of a loss function and a regularisation term, preventing the model from overfitting [CG16]. XGBoost leverages gradient boosting to find the optimal ensemble of trees. Gradient boosting is an iterative optimisation technique that calculates the gradient of the loss function to reduce the residual error ϵ of the current ensemble model by adding new weak learners. In essence, each subsequent learner corrects the errors made by the previous ensemble models, leading to more robust predictions [Mur22] [CG16].

The complexity of the XGBoost model can be controlled via hyperparameters such as tree depth, the number of trees, and learning rate. Latter specifies the contribution of each tree to the ensemble and controls the rate at which the β coefficients are updated during training. The lower the value the slower the updates and vice versa. In this thesis, the XGBoost algorithm is used to build binary classifiers, as described in Chapter 4. The algorithm is available in the Python package XGBoost [CG16].

The Multi-layer Perceptron (MLP) is a neural network (NN) designed to learn a mapping function $f(x) : \mathbb{R}^f \rightarrow \mathbb{R}^c$ for classification, or $f(x) : \mathbb{R}^f \rightarrow \mathbb{R}$ for regression tasks. Here, f is the number of input features in a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times f}$, while c corresponds to the number of classes in the dependent output variable y. A standard NN consists of an input layer, one or more optional hidden layers, and an output layer. The hidden layers introduce non-linear transformations via activation functions such as the ReLU (rectified linear unit) or tanh (hyperbolic tangent). Theses enable the model to solve non-linear prediction tasks by constructing non-linear decision boundaries [Mur22]. The layers are interconnected by weights, which are iteratively learned during the training [Mur22].

Each layer consists of multiple nodes (or neurons), each assigned with an activation function. Briefly, a node processes the weighted outputs from the previous layer as input, applies the activation function, and propagates the outcome to the subsequent layer [Mur22]. The input and output layers have a fixed number of nodes, given by the number of features and classes, respectively. The number of nodes in the hidden layers is typically set between the number of features and classes. Instead of setting this number arbitrarily, pruning can be applied to determine this number more systematically [Hoe+21].

During training, the weights are optimised using backpropagation, which is an optimisation process that updates the weights to minimise the loss function [Mur22]. The batch size, a hyperparameter that controls the number of samples processed before updating the model weights, also affects the optimisation process and model performance. Smaller batch sizes allow for more frequent updates, while larger batches stabilise weight updates.

In this thesis, the MLP models are trained using cross-entropy as loss function to minimise the difference between two probability distributions, i.e., the true binary labels of the dependent variable $\boldsymbol{y}$ and the predicted probabilities $P(\boldsymbol{y})$ [Bro19]. The cross-entropy loss function for the ith data point is defined as

$$\mathcal{L}^{(i)} = -\sum_{j=1}^{c} y_j^{(i)} \cdot \log\Big(P(y_j^{(i)})\Big) \ .$$

In summary, the MLP is able to solve complex prediction tasks, due to its layered structure and non-linear activation functions. In this thesis, I used the MLP model available in the Python package Scikit-learn [Ped+11] to train a binary classifier.

SVM models transform non-linearly separable data into a higher-dimensional space where linear separation is possible, using the kernel trick [TH09]. However, some data points may remain non-separable even after applying the kernel trick. For a binary SVM classifier, the hyperplane is defined as

$$h(\boldsymbol{x}) = \text{sign}[\boldsymbol{x}^T\beta + \beta_0] \ ,$$

where are assigned to class +1 if above the hyperplane and to class −1 if below [TH09]. The SVM aims to maximise the margin between classes, achieved by optimising the following objective function:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i}^{n} \xi_i \tag{2.17}$$

$$\text{subject to } \xi_i \geq 0, \ \ y_i(\boldsymbol{x}_i^T\beta + \beta_0) \geq 1 - \xi_i \ \ \forall i \ , \tag{2.18}$$

where C is the cost parameter, and ξ is a slack variable for misclassification, making the SVM less sensitive to outliers [Mur22]. The Lagrangian dual form is used to rewrite the optimisation problem (eq. 2.17) as:

$$\mathcal{L}_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{i'=1} \alpha_i\alpha_{i'}y_iy_{i'}K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ and } \sum_{i}^{n} \alpha_i y_i = 0 \ .$$

Here, $K(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ is a kernel function, transforming the feature vectors [TH09]. The prediction is given by

$$\hat{y} = \text{sign}[\boldsymbol{x}^T\hat{\beta} + \beta_0] \ ,$$

with estimated coefficients

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \ \boldsymbol{x}_i \ ,$$

where $\hat{\alpha}_i$ corresponds to the support vectors, the closest data points to the hyperplane. The performance of an SVM classifier depends on hyperparameters such as the regularisation parameter C and the kernel function, which must be optimised for the application. In this thesis, the SVM model from Scikit-learn [Ped+11] is used in two contexts, i.e. as part of an embedded feature selection method and as a binary classifier in Chapter 4.

2.4.3.3 Model selection using Cross-Validation

Model selection involves identifying the most suitable model for a given task by assessing its performance and complexity. This process relies on techniques such as probabilistic and resampling methods, which assist in reducing the risks of overfitting and underfitting, while evaluating the model's performance on unseen data. Two approaches, probabilistic and resampling techniques, are often used for this purpose [TH09].

Probabilistic techniques use *in-sample* measures, such as performance metrics on the training set. The objective of fitting a model is to minimise the in-sample error by simultaneously considering the model's generalisability. To achieve this balance, penalisation terms are added to correct for the in-sample bias. Commonly used metrics are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [TH09]. However, probabilistic techniques are often based on prior assumptions about data distribution, are sensitive to outliers, and can lack interpretability [TH09].

In this thesis, resampling techniques, such as CV, are employed (Chapter 4). In contrast to probabilistic techniques, resampling techniques evaluate model performance on *out-of-sample* data, specifically on validation sets not used during training. This is achieved through a systematic process, which can be generalised into the following steps:

1) Partition data into training and validation / test sets.

2) Train the model on the training set.

3) Validate the model on the validation / test set.

4) Repeat steps 1-3 according to the chosen CV approach.

Most resampling techniques determine the optimal model and allow for hyperparameter tuning by repeating these steps 1-3 across multiple iterations [TH09]. These methods are particularly advantageous for small datasets, as they account for high variance and maximise the use of available data.

CV techniques can be broadly categorised into exhaustive and non-exhaustive approaches. In an *exhaustive* CV, all possible combinations of samples are used as training and validation samples in each iteration. An example is Leave-p-out CV, which uses p samples for validation and n-p samples as training set [Mur22]. A special case is $p = 1$, which is known as Leave-one-out CV. However, the Leave-p-out CV does not consider the class distributions and is therefore unsuitable for imbalanced data.

Non-exhaustive CV methods, such as hold-out, k-fold, stratified k-fold, and nested CV, use subsets of the training and validation data to estimate performance.

- Hold-out CV: This basic form of CV splits the dataset into a single training and test set. This form of CV does not repeat steps 1-3, which makes it computationally efficient. On the other hand, its results are sensitive to the specific data split, leading to potentially high variance in the performance.
- k-fold CV: In this method divides the dataset into k equally sized, non-overlapping subsets. In each iteration, k-th held-out subset is used for validation, while the remaining $k - 1$ partitions are used for training. In total k models are fit and their mean performance is reported.
- Stratified k-fold CV: This approach is a variation of k-fold CV and is designed to handle imbalanced class proportions. However, stratified and k-fold CV might result in biased performance estimates due to lucky or unlucky splits of the data.
- Repeated k-fold CV: To address the limitations of a single k-fold run, this approach repeats the k-fold process multiple times, i.e., step 1-4. In comparison to k-fold CV, the division into training and test set is randomised, therefore some samples might be never selected in the test set.
- Repeated stratified k-fold CV: As an extension of repeated k-fold CV, this method combines the benefits of stratification with repeated sampling. It is useful to address class imbalance, as it stratifies samples by class labels. However, due to the repeated CV, it is computationally more expensive.
- Nested CV: As a hierarchical approach, nested CV divides the process into an inner and outer loop. The inner CV is used for hyperparameter tuning to identify the best model, while the outer CV provides an unbiased estimate of the expected performance of the best model on unseen data [WC21].

In summary, CV assists in model selection by optimising hyperparameters and accounting for generalisability. While exhaustive CV methods are thorough, they are computationally

		Actual class		Total
		positive: 1	negative: 0	
Predicted class	positive : 1	true positives (TP)	false positives (FP) (Type I error)	m_1
	negative: 0	false negatives (FN) (Type II error)	true negatives (TN)	m_0
Total		n_1	n_0	N

Table 2.4: Confusion matrix for binary classification.

more expensive. In contrast, non-exhaustive CV techniques, such as k-fold and nested CV, aim for balancing between efficiency and reliability. Selecting the appropriate CV method depends on the dataset size, class distribution, and computational constraints. This thesis ensures the development of robust and generalisable models, by integrating these techniques into the modelling process (Chapter 3, Chapter 4). Additionally, an independent test sets were provided, which remained unused during training and validation, to provide an unbiased final evaluation of the model performance.

2.4.3.4 Model evaluation metrics

After selecting the model that best generalises to the underlying data distributions, its performance is assessed using evaluation metrics such as accuracy and F1-score. These metrics compare the predicted sample labels with their true labels. This evaluation process involves categorising predictions into TP, FP, FN, and TN samples, which are summarised using a confusion matrix.

The confusion matrix, see table 2.4, shows the relationship between predicted and actual labels. The diagonal cells of the confusion matrix represent correct predictions, while off-diagonal cells capture errors. In binary classification, these include the Type I (FP) and Type II (FN) error. Analogous to statistical testing (Chapter 2.2), these errors reflect instances where the Null hypothesis $\mathcal{H}_0$ is falsely rejected or accepted [CL15]. Thus, the confusion matrix provides insights into model's classification strengths and weaknesses.

In this thesis, the data sets are highly imbalanced. Hence, the following metrics are adjusted for use with such datasets by adding class weights.

Precision (eq. B.11) measures the proportion of predicted positives that are correctly classified as TP. The weighted precision incorporates class-specific weights. The weight w_i is the fraction of the number of samples n_i in class i to the total number of samples in the dataset. The weighted precision can be calculated by

$$\text{weighted precision} = \sum_{i=0}^{c-1} w_i \cdot \text{precision}_i \qquad \text{with } w_i = \frac{n_i}{\sum n_i} \ ,$$

where c represents the total number of classes, which is $c = 2$ throughout this thesis. Weighted precision should be used when there is a desire for certainty regarding the prediction of positive events.

The weighted recall (or sensitivity or true positive rate (TPR)) assesses how effectively the model identifies true positives. It is calculated by multiplying the recall (eq. B.12) by class-specific weights

$$\text{weighted recall} = \sum_{i=0}^{c-1} w_i \cdot \text{recall}_i \qquad \text{with } w_i = \frac{n_i}{\sum n_i} \ .$$

This metric should be used when maximising the positive event rate is desired. Depending on the goal, which is often to classify the more seldom events (TP), either precision or recall are more in focus for evaluating the performance of the model. For example, in the scenario of classifying skin diseases, the decision can have consequences, as FP predictions can lead to wrong treatment approaches. Especially in healthcare, a higher certainty about the TP events is desired. This can be achieved by focusing on a high precision scores. However, many patients might be missed that would have had the skin disease by solely pushing for a high precision value. These patients would have been detected if a high recall of the model would have been of desire. In essence, both precision and recall should be optimised together. Furthermore, the evaluation metric should be selected in a manner that aligns with the objective.

Balancing precision and recall is essential for robust model evaluation. The F1-score (eq. B.13) provides a harmonic mean of these metrics. For imbalanced datasets, the *weighted F1-score* is calculated by

$$\text{weighted F1-score} = \sum_{i=0}^{c-1} w_i \cdot \text{f1 - score}_i \qquad \text{with } w_i = \frac{n_i}{\sum n_i} \ . \tag{2.19}$$

The weighted F1-score should be used in scenarios prioritising both precision and recall, such as diagnostics, where misdiagnoses must be minimised without neglecting detection. By optimising this metric, the model ensures a balanced approach to performance evaluation, reducing risks of FP and FN.

In scenarios where both classes are equally important, accuracy (eq. B.10) is a common metric for binary classification. However, accuracy can be misleading in highly imbalanced datasets, as the model may achieve high accuracy by always predicting the majority class. To address this, *balanced accuracy* is used, which equally weighs performance across both classes, as given by

$$\text{balanced accuracy} = \frac{\text{recall} + \text{specificity}}{2} \ . \tag{2.20}$$

Balanced accuracy reduces the bias towards the majority class, therefore making it suitable for imbalanced classification tasks.

Another tool to evaluate the performance of binary classifier is the *ROC curve* (Supplements B.3). It illustrates the model's ability to distinguish between the positive and negative classes by comparing the true positive rate (TPR) against the false positive rate (false positive rate (FPR)). The FPR is defined as [Tha20]

$$FPR = \frac{FP}{FP + TN} \ .$$

In contrast to metrics based on a fixed decision threshold (e.g., $\theta = 0.5$), the ROC curve evaluates the classifier's performance across a range of threshold values, offering a more comprehensive analysis. In order to quantify the performance shown by the ROC curve, the *AUC* is calculated. It represents the area under the ROC curve and is defined as [CL15]

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x)) \; dx \ .$$

In a perfect scenario, the AUC would be 1, indicating perfect discrimination between the classes, whereas an AUC of 0.5 suggests no discriminatory power, equivalent to random guessing. In data sets with a high degree of class imbalance, the AUC can produce overly optimistic results, failing to reflect the classifier's performance on the minority (positive) class. In this case, the precision-recall (PR) AUC offers an alternative. It focuses more on the minority (positive) class by assessing the trade-off between precision and recall. The AUC can be similarly computed by exchanging TPR and FPR by precision and recall. In summary, both ROC and PR curves, along with their respective AUC metrics, range between 0 and 1, where higher values indicate better performance. Their use should be guided by the characteristics of the dataset and the specific objectives of the analysis.

In summary, model evaluation metrics provide insights into a model's performance, guiding decisions on model suitability for specific tasks. By incorporating metrics tailored to imbalanced data, such as weighted precision, recall, and F1-score, and employing balanced accuracy, this thesis ensures robust evaluation practices.

2.5 High-throughput sequencing techniques

In the introduction (Section 1.4.3), high-throughput next generation sequencing (NGS) techniques were briefly introduced. These technologies have revolutionised sequencing by enabling the parallel analysis of millions of DNA molecules, advancing our understanding of the molecular landscape of diseases. For instance, bulk RNA-seq enables the study of diseases at the population level, while scRNA-seq allows to investigate changes in expression profiles within the same cell types across diseases. Additionally, ST adds another layer of information, namely the location of the measured gene expression profile in the tissue.

2.5.1 Bulk RNA-sequencing

Bulk RNA-seq involves all methods which measure average gene expression levels of all cells present in a sample [Heg+22]. Its generated data, is thus predestinated to reveal differences in gene expression between conditions on the population level [WGS09]. In 2019, it was reported that the median number of RNA-seq samples in a study is eight [Ber+19]. In this thesis, I analyse a RNA-seq dataset with 727 samples from various skin conditions. Before, I will build a basis to understand RNA-seq results by briefly introducing its generation and computational preprocessing.

2.5.1.1 Experimental basics

To computationally analyse bulk RNA-seq data, tissue punch biopsies are obtained from donors, typically ranging from 3 to 6 mm in diameter and cutting deep into the skin [Zub02]. The tissue samples are then stored as fresh-frozen tissue (FFT), embedded using formalin-fixed paraffin-embedded (FFPE) material or RNA-stabilising solutions like RNAProtect. In the wet lab, the samples undergo NGS preparation, starting with library preparation. This involves sample dissociation to release RNA, followed by RNA isolation [Heg+22]. Ribosomal RNA depletion or mRNA enrichment is performed, and reverse transcriptase is used for complementary DNA (cDNA) synthesis, converting RNA into cDNA [Heg+22] [Ber+19]. Adapters are added to the cDNA for sequencing, determining whether single-end or paired-end sequencing is performed. Single-end sequencing involves reading the DNA strand once from one end to the other ($3'$ to $5'$ or vice versa), whereas paired-end sequencing reads the strand from both directions [Heg+22]. The cDNA is then amplified by polymerase chain reaction (PCR), and after Quality control (QC), the library is submitted to the sequencer, which processes the data and stores it in .fastq files [Ber+19].

2.5.1.2 Computational data processing

In order to generate a count matrix from the .fastq files which is the required format to apply computational analysis methods, alignment tools such as STAR can be used [Dob+13]. It aligns the reads against a reference genome, e.g., GRCh38 for human data which was released by the Genome Reference Consortium (GRC) [Sch+17] [Nur+22]. The mapping quality of any read alignment tool can be evaluated using FASTQC [And] and MULTIQC [Ewe+16]. The final product is a count matrix $\boldsymbol{X}^{(R)} \in \mathbb{N}_0^{s \times g}$ holding the information of relative abundances of all genes g across all samples s. The superscript denotes the processing status, e.g., (R) which refers to the raw count matrix.

During the QC, outliers are removed, as they negatively influence the analysis, leading to false discoveries. Samples having comparably low total number of reads, i.e., library size, are removed from the cohort. In order to avoid biases from sequencing depth and transcript length in downstream analyses, resulting in low expression of genes, I require at least one transcript per million (TPM) and raw counts across all samples to be measured. The TPM are defined as

$$\mathrm{TPM}_{gs} = \frac{y_{gs}}{l_{gs}^{(eff)}} \left(\frac{1}{\sum_g \frac{y_{gs}}{l_{gs}^{(eff)}}} \right) \cdot 10^6 \ ,$$

where $l_{gs}^{(eff)} = l_g - \hat{\bar{F}}_s + 1$ is effective length of a gene. In addition, filtering based on count - per - million (CPM) is applied. Moreover, I only keep protein coding genes, having HUGO name annotations, as they are directly interpretable and comparable with literature.

Samples have to be normalised, to account for technical biases, causing variations in the library size. In this thesis, I use trimmed mean of M-values (TMM) values to normalise the library sizes yielding effective library sizes. Those can be used to receive normalised counts, log2 counts-per-million (logCPM) counts, which are used in downstream analyses [RMS10].

Technical and biological artefacts can affect the measured relative abundances of genes, leading to variations (batch effects) across factors like sequencing batches or sex. Methods such as ComBat [JLR07], available in the sva package [Lee+12], correct gene counts by regressing out the effects of the variables causing these batch effects.

High-level summaries of RNA-seq data can be visualised using various embeddings. Commonly, PCA [WEG87] is used for this purpose, with variations such as the Biplot [Gab71] offering insights into which genes or observations are the dominant features in the dataset, potentially revealing batch effects. Additionally to PCA, UMAP can be applied [MHM18] to RNA-seq data, highlighting global or local structures in the data. However, UMAP is particularly useful for datasets with many samples or conditions, where PCA may fail to adequately represent the variations in two dimensions [Yan+21].

2.5.2 Single-cell RNA-sequencing

In 2013, scRNA-seq technology was awarded as the "method of the year" and revolutionised the understanding of various diseases [Scr]. In contrast to bulk RNA-seq, which analyses the transcriptome at a population level, it captures the transcriptome of individual cells within a biopsy. This enables to study heterogeneity of cell types between conditions such as healthy and diseased states, composition, interaction across cells, population dynamics. Furthermore, scRNA-seq facilitates the characterisation and discovery of rare or novel cell types. Major efforts are being made to create a Human Cell Atlas (`https://www.humancellatlas.org/`), as scRNA-seq has the potential to advance precision medicine [Reg+17] [Wie+19].

Various scRNA-seq platforms, such as 10x Chromium and Fluidigm C1, enable transcriptome analysis using microdroplet- and microfluidic plate-based approaches, respectively [Kas+20]. The choice of platform depends on study requirements, such as number of cells captured per sample, amplification bias, number of reads per cell, expected cell sizes, and sequencing library [Kas+20]. For instance, 10x Chromium captures $100 - 10,000$ cells up to a size of 40 μm with reduced amplification bias and supports mixed libraries [Kas+20]. In contrast, Fluidigm C1 processes a few hundred, smaller cells [Kas+20]. Thus, platform selection determines the resolution and applicability of scRNA-seq data.

In this thesis, scRNA-seq data, generated using the droplet-based platform 10x Chromium from 10x Genomics, is analysed. The subsequent paragraphs will briefly explain the experimental basics underlying this platform and the computational preprocessing steps involved. More details about frameworks and set parameters can be found in Chapter 5.

2.5.2.1 Experimental basics

In order to isolate single cells from a specimen, the tissue is digested and dissociated into a cell suspension. However, this process can alter the gene expression profile of a cell due to stress and damage introduced during dissociation [Jov+22]. In the downstream analysis, only good samples should be included, allowing to confidently annotate the cell types. These samples should have a clean cell suspension and contain living cells with

$> 90\,\%$ viability [Hab23]. Thus, poor quality samples, containing compromised cells, cell aggregates, and other contaminants in the cell suspension should be not included in the analysis [Jov+22] [Hab23].

Following dissociation, the cells are separated using fluorescence activated cell sorting (FACS) to sort for living cells [Jov+22]. In this thesis, additional gating for immune cell markers was applied to increase their proportion in the samples. For library preparation, the Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Single Index) were used. These kits contain reagents to pack individual cells into beads to induce cell destruction, reverse transcription, and molecular barcoding to uniquely label each cell [Haq+17].

To address the limited starting material, amplification by PCR is performed to increase the amount of cDNA [Kas+20]. Since the amplification can introduce bias, unique molecular identifiers (UMIs) are attached to each molecule within a cell before amplification [Haq+17] [Kas+20]. These UMIs serve as barcodes, enabling accurate quantification of the original molecules. Moreover, a sample barcode is added to each amplified single cell mRNA to enable the pooling of samples derived from different sources, such as distinct patients or disease conditions.

Finally, the constructed library is sequenced using a NGS platform, such as Nova Seq$^{\mathrm{TM}}$ Illumina. This workflow ensures high-quality scRNA-seq data for computational analysis.

2.5.2.2 Computational data and procedures

Read mapping was performed by Thomas Walzthoeni using CellRanger [Zhe+17], which internally uses the STAR aligner [Dob+13] with a species-specific reference genome, such as GRCh38 for human data [Nur+22]. The output is a count matrix $\boldsymbol{X}^{(R)} \in \mathbb{N}_0^{c \times g}$, where c is the number of cells and g is the number of unique genes measured. This matrix provides the relative abundance of genes in cells as UMI-counts. Due to dropout events, the count matrix is often sparse, where a gene may be detected in one cell but missed in another of the same type due to low expression levels. In the analysis it needs to be accounted for this phenomenon [Qiu20]. Before downstream analysis, scRNA-seq data has to undergo multiple steps of preprocessing, QC, normalisation, HVG selection, batch correction, dimension reduction, and cell type annotation.

Technical and biological artefacts can influence scRNA-seq data and must be addressed before downstream analysis. This can be accomplished by applying QC metrics on the cells and genes, focusing on parameters such as number of counts (count depth) and genes, total sum of UMI-counts of one barcode across all genes (library size), and the fraction of mitochondrial (MT) transcripts recovered [Hon+22][LMM16]. The following filtering steps are applied to remove outlier cells.

- Cell level QC: In order to filter low quality cells, it is recommended to jointly asses the number of UMI-counts, number of genes, and MT RNA content per cell. They categorise cells into cells with high UMI-counts and number of genes and with low UMI-counts, few genes and high MT fraction. The latter is an indicator for a stressed or dying cell [GKK12][LT19]. It is recommended to remove cells exhibiting an increased MT fraction above 20 - 25 % [LT19]. However, the precise cut-off should be tailored to the specific experimental conditions and research objectives.
- Gene level QC: In order to reduce the amount of non-informative genes, lowly expressed genes across all cells are removed using a threshold [LT19].

By systematically applying these QC steps, it is possible to minimise artefacts and ensure the reliability of the downstream analysis.

In addition to measuring empty droplets, multiple cells can aggregate in a single droplet, forming doublets or multiplets. These aggregates may involve cells of the same type (homotypic) or different types (heterotypic) [DeP+19]. Doublets or multiplets occur by overloading the chip with too many cells and are characterised by higher expression levels and a hybrid transcriptomes [WLK19][DeP+19]. Doublets are technical artefacts that cause biologically erroneous findings in the downstream analysis [WLK19]. While QC removes already cells with unusually high number UMI-counts and genes, some doublets remain undetected. To address this, computational methods are applied to detect doublets based on their unique characteristics, transcriptomes similar to single cells, distinguishable gene expression profiles, and mixed cell states [WLK19][DeP+19]. Applying doublet detection removes doublets, thereby enhancing the accuracy of cell type annotation.

Normalisation accounts for differences in capture efficiency and library size across cells while preserving biological variation [LMM16]. A common assumption in normalisation tools is that most genes are not significantly differentially expressed between cells. The technical artefacts caused by variations in sequencing depth across sample can be addressed by applying a count depth scaling or CPM normalisation [LT19]. Originally for bulk RNA-seq data, CPM normalisation was extended to scRNA-seq data by Lun et al. (2016) [LLBM16]. This method accounts for dropout events by calculating pool-based size factors, which are estimated by summing gene expressions across multiple cell pools and then are deconvoluted into cell-specific size factors s_c. These size factors are used to normalise the data. This technique requires that the majority of genes is not differentially expressed across the whole data set. To weaken the assumption, Lun et al. (2016) recommend to clustering cells first, as cells within a cluster share similar expression profiles, hence no significantly differentially expressed genes (DEGs). The filtered raw counts are then normalised and log-transformed with a pseudo-count of 1

$$\tilde{y}_{gc} = log(\frac{y_{gc}}{s_c} + 1),$$

where y is the expression value of gene g in cell c. Log-transformation minimises the effect of high-abundance differences and forces the data to follow a Gaussian-like normal distribution. This variance-stabilising approach transforms heteroscedastic (variance is not constant) data into homoscedastic data [AEH23]. However, size factor normalisation and log-transformation may not fully stabilise variance, particularly for highly abundant genes with low UMI-counts [HS19]. Therefore, results should be interpreted with caution. In summary, normalisation mitigates technical biases, while preserving meaningful biological variation.

After normalisation and variance stabilisation, gene expression differences reveal biological variation, with genes contributing to this variation referred to as HVGs. These genes drive population differences, describe cellular phenotypes, and aid in identifying new cell types [LMM16]. HVGs should not be confused with marker genes, which are specific to cell types and serve as cell type identifiers. Identifying HVGs based on variance alone is insufficient, as count data is not fully heteroscedastic even after stabilisation [HS19]. Methods typically use the mean-to-variance relationship, with alternatives such as standard deviation, coefficient of variation [Che+16], and squared coefficient of variation [Bre+13]. Commonly, researchers select between $1,000$ and $6,000$ HVGs [LT19], depending on the dataset. Once HVGs are identified, they should be compared to housekeeping genes (HKGs), which are uniformly expressed across cells, maintain cell viability, and remain stable across conditions. This comparison serves as a QC step to validate HVG selection. The process ensures HVGs capture biologically relevant variability, enhancing the reliability of cell type identification and downstream analyses.

Technical batch effects in scRNA-seq can arise from factors such as differences in sample preparation, collection time point, or sequencing lane. These effects can be categorised into classical batch effects (within the same experiment) and data integration issues (across experiments or laboratories), both of which can confound biological interpretation and require computational correction [LT19]. Methods, including linear, graph-based, and deep learning approaches, have been developed to remove batch effects [LH22]. Amongst the top performers are Scanorama [HBB19] and scVI [Lop+18], and scANVI [Xu+21]. Scanorama, in particular, is effective in batch correction, as it embeds data linearly, and can handle large complex data while preserving biological variability [Lue+22] [HBB19]. In summary, batch correction methods reduce technical artefacts in multi-sample scRNA-seq studies.

The scRNA-seq data can be also visualised in lower-dimensional embeddings using, e.g., PCA and UMAP. PCA can be used to identify strong technical and biological artefacts that drive the separation of the samples in 2D/3D. In contrast, UMAP provides an insight into the global and local structures of the high-dimensional data. A recommended approach for visualising scRNA-seq data involves applying PCA followed by the construction of a KNN graph. This graph serves as the basis for generating the UMAP embedding, which represents the high-dimensional structure of the data in a comprehensible format [LT19]. Thus, using both, PCA and UMAP, offers a robust approach for visualising and interpreting the complex structures in scRNA-seq data.

One of the advantages of scRNA-seq data is the identification of known and new cell types, with various computational methods developed for this purpose. These include single - cell ANnotation using Variational Inference (scANVI) [Xu+21], cell type classification via anchor genes in Seurat [Stu+19], and marker gene enrichment tests provided by SCANPY [WAT18] or Seurat. The probabilistic annotation tool scANVI, an extension of single-cell variational inference (scVI) [Lop+18], uses NNs, stochastic optimisation, and variational inference to integrate multiple scRNA-seq datasets, mapping cell type labels from an annotated reference dataset onto unlabelled cells in a query dataset. Seurat uses anchor-based classification, where reference data is projected onto unlabelled data by identifying anchor points, enabling cell type annotation. Marker gene enrichment tests involve identifying cell clusters with similar transcriptomics profiles using clustering algorithms like Leiden [TWVE19] (Section 2.4.2.2), then evaluating marker genes for enrichment within each cluster. If the null hypothesis H_0 is rejected, the cluster can be annotated with the corresponding cell type. It is recommended to combine automatic methods with marker gene validation and domain expert confirmation for robust cell type annotation [Cla+21].

2.5.3 Spatial Transcriptomics

ST, a new technology developed by Ståhle P. et al. (2016) [Stå+16], has enabled the possibility of measuring transcriptomes in a spatially resolved manner from freshly frozen, intact tissue. In 2018, the company 10x Genomics commercialised and further refined the technique under the name Visium [Mar21]. In this thesis, the terms Visium and ST are used interchangeably.

Visium consists of an object slide with four capture areas, each containing barcoded spots that preserve the spatial architecture of the transcriptome. The design minimises lateral diffusion of transcripts, ensuring accurate localisation [Stå+16]. This technique allows for the analysis of thousands of spatial locations on a single tissue section, enabling the discovery of spatial gene expression patterns.

Due to the spot size ($\varnothing\, 55\, \mu m$), each spot captures approximately 1 - 10 cells, yielding a "mini-bulk" resolution. However, ST suffers from a low detection sensitivity ($(6.5 \pm 1.5)\%$) [Stå+16], compared to scRNA-seq (5 % - 40 %) [GO15], , leading to dropout events that must be considered in downstream analyses. Additionally, ST is more expensive than bulk or scRNA-seq methods. Despite these limitations, the advantages of ST outweigh the disadvantages, enabling researchers to investigate, for instance, the spatial expression patterns of genes, the enrichment of cell types and niches in tissues [Pal+22].

2.5.3.1 Experimental basics

ST involves cutting fresh frozen or FFPE tissue into thin slices of roughly $10\, \mu m$, corresponding to the diameter of one cell. The number of slices dependents on the required replicates, which is two in the presented study in Chapter 5. The tissue slices are placed onto object slides, containing four capture areas each measuring $6.5\, \mu m \times 6.5\, \mu m$. Each capture area contains 4992 uniquely barcoded spots with diameters of $55\, \mu m$ and a spacing of $100\, \mu m$ between spots. These spots are equipped with millions of barcoded oligonucleotides that act as primers to bind to the released mRNA in a spot [Gen].

After placement, the tissue slices are fixated and stained. In the study presented in Chapter 5, hematoxylin and eosin (H&E) staining is used, which colours nuclei in purple and everything else on the tissue in pink. A brightfield microscope can then be used to take images of the stained tissue section. In order to extract the mRNA, the fresh frozen tissue is permeabilised and spot-specific barcodes are added from the object slide (e.g., 10x Visium). This allows later to map back the measured gene expression to its original location in the tissue. Then the tissue is lysed to release the mRNA that binds to the barcoded oligonucleotides. In case of FFPE tissues, ligated probe pairs are released that bind to the barcoded oligonucleotides. Reverse transcriptase creates cDNA from the barcoded mRNA, which is then pooled for sequencing [Gen]. The sample-specific barcodes are added and the library is subjected to NGS. The experimental workflow of ST combines tissue slicing, spatially barcoded capture, and sequencing to generate spatially resolved transcriptomics data.

2.5.3.2 Computational data processing

After sequencing, reads are mapped back to a reference genome using 10x Genomics' software SpaceRanger, yielding a spatially resolved matrix of gene expression counts. The resulting data consists of a count matrix, $\boldsymbol{X}^{(R)} \in \mathbb{N}_0^{s \times g}$, where s is the number of spots and g is the number of unique genes measured. Additionally, SpaceRanger provides images and spot location information for each capture area.

After running SpaceRanger, QC and data preparation steps can be applied on the count matrix. Most of the scRNA-seq data preparation tools (Section 2.5.2.2), i.e., QC, normalisation, HVG selection, batch correction, and visualisation, can be applied on ST data as well. Although, the ST data in this thesis does not have single cell resolution, it behaves to some extend similarly to scRNA-seq. For confounder correction, Scanorama [HBB19], which was originally developed for batch correction in scRNA-seq data, was employed in this thesis. SCANPY [WAT18] was used for other preprocessing steps.

Due to the mini-bulk spatial resolution of ST, which leads to measuring the sum of transcripts in a spot from various cells, the number of cells and the cell types present in a spot cannot be directly assessed. Spot deconvolution algorithms aim for recovering the cell type composition in a spot and enable gaining insights into cell type related expression patterns in a tissue. Examples of deconvolution algorithms are Tangram [Bia+21], Stereoscope [And+20], and Cell2Location [Kle+22].

- Tangram [Bia+21] aligns an annotated scRNA-seq reference dataset with the mini-bulk transcriptomes of spots using a deep learning framework. It generates spatial maps that assign cell types to tissue locations, revealing the cell type composition in spots. An advantage of Tangram is that it corrects for the sensitivity of ST, improving performance on sparse data [Bia+21] [Li+22].

- Stereoscope uses a negative binomial distribution to model the data and estimate the proportion of cell types in a spot, using scRNA-seq dataset as reference. It determines the probability that cell types are present in a spot by finding combinations of cell types that best reflect the measured transcriptome profile of a spot [And+20].

- Cell2Location uses a similar model as Stereoscope and also requires a reference scRNA-seq dataset. Based on that, it estimates the cell type abundance in spatial locations. Moreover, Cell2Location accounts for technical artefacts [Kle+22].

A benchmark study by Li et al. (2022) showed, that Tangram, Stereoscope, and Cell2Location are amongst the top performers in recovering the cell type compositions in simulated data [Li+22]. Due to Tangram's superiority in dealing with highly sparse data, it was chosen as the deconvolution method in this thesis.

Data processing in ST involves preprocessing steps similar to those in scRNA-seq. In addition, deconvolution algorithms such as Tangram are used to infer the cell type compositions and spatial expression patterns.

2.6 Exploratory data analysis

Biologically focused exploratory data analysis aims to identify biological mechanisms and patterns within the data. In addition, they reveal technical artefacts and outliers in the data. It is an essential component of the data analysis and assists in formulating hypothesis. Visualisation techniques, such as PCA, boxplots, and heatmaps are employed. In the context of biologically driven tasks, differential gene expression (DGE) and pathway enrichment analyses can be leveraged as well. They permit insights into differences in the regulation of biological functions at the gene and pathway level between, e.g., healthy and diseased patients. This section presents methods for conducting DGE and pathway enrichment analyses.

2.6.1 Differential gene expression analysis

The DGE analysis aims to determine whether a gene is significantly differentially expressed between two experimental groups. This is accomplished by modelling the count data for each gene using a generalised linear model (GLM) and conducting a null hypothesis test $\mathcal{H}_0 : \mu = \mu_0$. This is repeated for all measured genes in an experiment. To ensure robustness against FP, corrections for multi-testing is applied to control the FDR. The resulting DEGs are characteristic of one group over the other, based on the observed effect sizes in the comparison. Various tools have been developed for bulk RNA-seq data such as DESeq2 [LHA14] and edgeR [RMS10].

Before conducting DGE analyses, low expressed genes have to be removed, as these can negatively impact the outcome [RMS10]. The tool edgeR has the option to filter based on the CPM values, taking into account the two groups selected for comparison. That means, a gene has to be expressed in at least k samples with a minimum CPM count (10 is the default value in edgeR), where k is the minimum group size [RMS10]. The CPM counts are defined as

$$\mathrm{CPM}_{gj} = \frac{y_{gj}}{N_j} \cdot 10^6 \ ,$$

where y_{gj} is count of a gene g in the data point j and $N_j = \sum_g y_{gj}$ is the sequencing depth. As an alternative to CPM-based filtering, thresholds can also be applied directly to raw counts or TPM values, depending on the specific requirements of the experiment and analysis pipeline.

In DGE analysis, the objective is to identify genes that are differentially expressed between two groups, where each group is represented by multiple samples. Bulk RNA-seq measures relative gene abundances rather than total expression levels, leading to the overexpression of certain genes expressed in the same sample. This can introduce biases when identifying DEGs, as the other genes are then incorrectly defined as downregulated. In addition, variations in sequencing depth resulting in different library sizes between samples can also lead to false discoveries. [LL13]. To mitigate these biases, library size normalisation techniques, such as TMM provided by edgeR, can be applied. TMM assumes that more than half of the genes are not differentially expressed [RMS10]. Another normalisation approach is to calculate size factors, which were introduced in section 2.5.2.2 for scRNA-seq. In DESeq2, size factors are computed for RNA-seq data without the pooling approach.

The objective of conducting a DGE analysis is to detect DEGs. Therefore, a negative binomial (NB) GLM is fitted, which is described by a model matrix [LHA14] [RMS10]. The model matrix extends the design matrix X introduced in section 2.4.3.2, by adding additional columns, which describe further dependencies. The model matrix is created by a *design function*, which holds the information about potential confounding variables as well as the group variable, here denoted as condition, for instance $y \sim 0 + \mathrm{batch} + \mathrm{condition}$. In this scenario, the expression levels of a gene between two groups would be compared, taking into account fluctuations caused by the batch variable.

The count data is modelled using a NB distribution to account for heteroscedasticity by

$$y_{gs} \sim \mathrm{NB}(\mu_{gs}, \alpha_g) \ ,$$

where y_{gs} is the observed count of a gene g in a sample s, $\mu_{gs} = M_s q_{gs}$ is the mean modelled by the true expression level q_{gs} as well as library size M_s, and α_s is the dispersion value. The dispersion parameter $\boldsymbol{\alpha}$ accounts for biological and technical variability in the gene counts. Estimating the dispersion is crucial as over- or underestimation can also lead to falsely identified DEGs [LL13]. More details on its computation can be found in Landau et al. (2013) [LL13]. Alternatively, the data can be modelled using a log-linear or quasi-likelihood model, rather than a NB distribution in edgeR [RMS10]. These distributions also assume that the underlying data follows a NB or Poisson distribution for $\alpha_g = 0$ [RMS10]. Besides estimating the dispersion, the true expression level q_{gs} has to be estimated for the DEG analysis [RMS10].

After fitting the NB GLM, statistical tests are applied to determine whether a gene rejects the null hypothesis $\mathcal{H}_0$. In DESeq2, the Wald test and likelihood ratio test (LRT) are used, while edgeR supports the empirical Bayes quasi-likelihood F-test and LRT [LHA14] [RMS10]. The derived p-values are corrected for FDR using BH. In addition, the effect sizes, such as log2 Fold Change (log2FC) factor (log-transformed difference in the expression level of a gene between the two groups), are calculated to quantify the magnitude of differential expression. Depending on whether a gene meets the thresholds set for the padj and log2FC, it is referred to either as DEG or non significant. The former applies if $\mathcal{H}_0$ is rejected and the latter if it is retained.

Due to dropout events, scRNA-seq and ST data require distinct modelling approaches. An alternative DGE analysis method is glmGamPoi [AEH20], which models a Gamma-Poisson distribution

$$y_{gc} = \text{GammaPoisson}(\mu_{gc}, \alpha_g) \ ,$$

where y_{gc} is the observed count of a gene, μ is the mean of the true expression of gene g in cell or spot c, and α_g is the dispersion estimate. A GLM is fitted and a LRT is used to test for DEGs [AEH20]. Besides comparing groups of single cells underlying a specific condition, glmGamPoi offers the option to create pseudo-bulk samples to identify DEGs between conditions on a sample level, e.g., comparing the samples of two diagnosis. In this thesis, the author compared conditions that occur at the single cell or spot level, therefore the option pseudo-bulk was not used.

The DGE analysis can be conducted using DESeq2 and edgeR for bulk RNA-seq data, while glmGamPoi is suitable for scRNA-seq and ST data. The results from DGE analysis can be further explored through downstream analyses, such as the identification of biological pathways, to gain deeper insights into the molecular mechanisms underlying the observed differences between experimental conditions.

2.6.2 Pathway enrichment analysis

Pathway enrichment analysis identifies biological pathways that are either activated or suppressed within a set of genes, with an occurrence higher than would be expected by chance [Rei+19a]. Biological pathways can be described as networks of interacting proteins that provide insights into the underlying biological functions within a specific condition. Identifying enriched pathways enhances the understanding of experimental conditions and can generate new hypotheses.

Significant efforts have been made to develop comprehensive databases, such as Reactome [Gil+22] and KEGG [KG00], assigning genes to their corresponding pathways. Reactome focuses on protein-receptors interactions, whereas KEGG includes molecular interaction, reaction and relationship networks. In order to find enriched pathways, two analysis methods have been employed in this thesis, i.e. over representation analysis (ORA) and gene set enrichment analysis (GSEA).

ORA involves testing whether a list of predefined DEGs are represented in a given biological pathway. Specifically, ORA assesses whether x or more genes from the input gene list X are found within a particular pathway more frequently than would be expected by chance [Kar+21]. This can be tested using the hypergeometric enrichment test:

$$P(X \geq x) = 1 - P(X \leq x-1) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \ ,$$

where M is the total number of genes in the pathway, N is the total number of genes measured in the experiment referred to as background or universe genes, and n is the number of genes associated with the pathway [Boy+04]. Applying the test on all pathways in a database, the derived p-values are corrected using a multi-testing method such as BH. A limitation of ORA is that it uses only the DEGs, which are determined based on arbitrary cut-offs for the log2FC and padj value derived from the DGE analysis. Consequently, small but potentially meaningful changes between phenotypes may not be captured by this approach.

GSEA addressed the limitation of ORA [CA22]. It evaluates all genes measured in an experiment, allowing for the detection of small but consistent phenotypic changes. The input to GSEA consists of a ranked list of genes, L, where the genes have been sorted in descending order by, e.g., the log2FC factor or signed padj value [Sub+05]. The algorithm calculates an enrichment score (ES), which reflects whether a pathway S is overrepresented at the top or bottom of the list L. Hence, the ES increases if a gene from the pathway S is present in the ranked list L and decreases otherwise. The ES can be described by a weighted Kolmogorov–Smirnov-like statistic [Sub+05] [HWC99]. Pathways that are overrepresented at the top or bottom of the ranked gene list will have a higher ES compared to gene sets with a uniform distribution of genes. The statistically significance of an ES is calculated by repeatedly randomly sampling from the list of ranked genes L creating new gene sets of the same size and performing the GSEA more than 1000 times. In other words, an empirical phenotype-based permutation test is performed for each pathway [Sub+05]. Additionally, to account for differences in pathway sizes, a normalised enrichment score (NES) is calculated, and multi-testing correction is applied to control for the FDR [Sub+05].

Pathway enrichment analysis, allows for the identification of biologically relevant pathways that are differentially activated or suppressed in experimental conditions. While ORA relies on a predefined set of DEGs, GSEA offers a more comprehensive approach by considering all genes and enabling the detection of subtle but consistent effects. As a result, GSEA has become the state-of-the-art method.

Chapter 3

Endotypes as alternative to the established disease ontology

Current clinical practice in diagnostics is to diagnose patients based on clinical assessments, including histopathology, symptoms, comorbidities, and laboratory tests. These assessments, which also guide treatment plans, are often subjective from the perspective of both, the patient and the clinician. This can lead to misdiagnosis, ineffective treatments, and increased healthcare costs. A progress towards improving the assignment of therapeutics is accomplished by grouping diseases into immune response patterns (IRPs) [EE18] [SGSE22]. This approach, which is also referred to as stratified medicine, can be seen as a precursor to precision medicine. Six IRPs have been identified so far, exhibiting distinct cytokine expression profiles and targeting specific pathways (Chapter 1.2.2) [EE18] [SGSE22]. Consequently, clinical assessments and the level of cytokine expression are used to assign patients to IRPs, thereby providing more objective and precise treatment options. However, diseases and their grouping into IRPs are not specific and granular enough to capture the heterogeneity of drug response.

For diseases, such as cutaneous lymphoma and parapsoriasis, the cytokine expression profile does not fit into the described IRPs, making an association challenging. Thus, the implementation of alternative treatment suggestions and drug repurposing for these diseases is not a straightforward process. Moreover, the categorisation of patients based on IRPs, does not guarantee drug response. One possible reason for this is the overlap of phenotypes between diseases such as eczema or eczematized psoriasis [LE23]. Such cases are relatively common and present a significant challenge in establishing an accurate diagnosis, which in turn makes it impossible to guarantee treatment success. Improving the stratification of patients by identifying endotypes is crucial for more targeted treatment and thus, driving precision medicine.

I aim to identify disease endotypes using machine learning (ML) and transcriptomics in order to enhance diagnostics, improve the established disease ontology, and advance the study of rare diseases. Moreover, I prove the clinical meaningfulness of endotypes by integrating patient phenotypes. This approach leads to an enhanced and data-driven patient stratification, thereby contributing to precision medicine in non-communicable chronic inflammatory skin diseases (ncISDs).

In this Chapter, objective disease endotypes are derived and characterised in terms of their gene expression profile and phenotype. In addition, I develop a feature selection strategy for bulk RNA-sequencing (RNA-seq) data that automatically identifies the most biologically relevant genes for clustering (Chapter 3.3.2). Subsequently, the approach is compared to two commonly used feature selection methods, the standard deviation and mean-variance relationship.

The deliverable "Utilising molecular profiling to derive biological meaningful endotypes" and its associated research questions, as outlined in Section 1.6, are addressed in this chapter:

- Objective (i), grouping patients into endotypes based on their molecular profiles (Section 3.3.2). Therefore, I introduce a pipeline and method to automatically identify the number of highly variable genes (HVGs) and the HVGs themselves. I use the HVGs to cluster patients into non-subjective disease endotypes.

- Objective (ii), comparing endotypes and the current disease ontology by investigating composition of diseases and IRPs (Section 3.3.3). In addition, I compare the explained variance of diseases, IRPs, and endotypes.

- Objective (iii), interpreting biological and phenotypic characteristics of endotypes by applying pathway (Section 3.3.4) and clinical attribute enrichment tests (Section 3.3.5).

This Chapter is related to the manuscript by Garzorz-Stark, Hillig, Seiringer, and Meinel et al. (In preparation).
Garzorz-Stark, Natalie[*] and **Hillig, Christina**[*] and Seiringer, Peter[*] and Meinel, Martin[*] and Maboudi Afkham, Heydar and Mishra, Jigyansa and Jargosch, Manja and Eyerich, Stefanie[*] and Menden, Michael[*] and Eyerich, Kilian[*†]. Integrating phenomics and transcriptomics to identify clinically meaningful endotypes of non-communicable inflammatory skin diseases." (*In preparation*).

In particular, my contributions to this study were as follows. I filtered, imputed, encoded, normalised, analysed, and visualised the clinical data. In order to identify clinically

[*] Contributed equally
[†] Corresponding author

meaningful endotypes, my co-author Heydar Maboudi Afkham and I conceptualised a feature selection pipeline, which is able to automatically determine the optimal number of HVGs for downstream analyses. In addition, he and I conceptualised a feature selection method, which I implemented and subsequently, compared to golden standard methods. Furthermore, I identified the endotypes and investigated the explained variance by each stratification approach. Moreover, I performed differential gene expression (DGE) analysis, gene set enrichment analysis (GSEA), and over representation analysis (ORA). In addition, I conducted literature research for the interpretation of the outcomes. I also performed statistical testing and calculated the effect sizes to find significant associations between endotypes and clinical attributes. All figures were designed and created by me. The results were interpreted together with my supervisor, Michael Menden, and our collaborators.

The bulk RNA-seq data generation, processing, and preprocessing were performed by our collaborators, the Helmholtz Munich bioinformatics core facility, and by my master student Jigyansa Mishra under my guidance, respectively. The clinical attributes were collected by Natalie-Garzorz Stark and Peter Seiringer.

3.1 Materials

In collaboration with Natalie Garzorz-Stark, Peter Seiringer, and Kilian Eyerich from the TUM - Department of Dermatology and Allergy, Karolinska Institutet - Department of Medicine Solna, and University of Freiburg - Department of Dermatology and Venerology, and Manja Jargosch and Stefanie Eyerich from the TUM - ZAUM, a substantial data repository was constructed, comprising bulk RNA-seq data and paired clinical traits. Skin punch biopsies of 23 inflammatory skin diseases from 408 patients were analysed. In total 727 samples from lesional (L) and non-lesional (NL) skin were collected. Of these, 313 patients provided both L and NL, 76 only L, and 19 only NL biopsies.

In order to investigate the association between phenotypes and endotypes, 86 clinical traits were collected. In particular, the diagnosis and IRPs, clinical information such as age, sex, four severity scores, 15 comorbidities, 10 laboratory assessments, 11 therapy associated information, 11 history associated information, 26 histology assessments, and seven morphology assessments, were collected. These either contain continuous, ordinal or nominal (dichotomous) variables.

Beyond the established IRPs, this study introduces an additional pattern, IRP 5. In contrast to the other IRPs, which are characterised by T-helper (Th) cell subsets, IRP 5 resembles diagnoses associated with autoinflammation. Autoinflammation is not yet fully understood, and is introduced in this study to be consistent with the terminology of IRP [SSS22] [CBF12] [PR+21].

3.2 Methods

3.2.1 Bulk RNA-sequencing data

Skin biopsies were prepared by Manja Jargosch and Stefanie Eyerich and analysed with an Illumina HiSeq 4000 sequencer. Processing of 727 samples from 408 patients was performed using the RNA-seq pipeline from nf-core v3.3-Bronze Bear [Pat+21] and the reference genome GRCh38 [Sch+17]. The pipeline was run by Xavier Pastor Hostench and Thomas Walzthöni from the Bioinformatics Core facility of the Helmholtz Center Munich. They set the parameters skip_dupradar, deseq2_vst, and skip_preseq to true. The mapping quality was assessed by Jigyansa Mishra, using FASTQC [And] and MULTIQC [Ewe+16], under my supervision. In total 721 samples passed the mapping quality check. The resulting gene expression matrix contained $60,666$ genes and 721 samples.

3.2.1.1 Data preprocessing

The preprocessing was performed by Jigyansa Mishra under my supervision. Quality control (QC) was applied to remove low expressed genes and samples with poor quality. On the gene level, all Y-chromosome genes and genes, which had less than one transcript per million (TPM) and raw counts across all samples, were removed. Y-chromosome genes were removed due to the large influence of sex on the gene expression level. In addition, only protein coding genes were selected. QC on the sample level included filtering for samples with low sequencing quality. This was done by calculating the sizefactor of each sample and removing those with a sizefactor within the first 0.1 quantile of all sizefactors. The final gene count matrix contained $17,816$ genes and 629 samples.

The data was normalised using size factors to account for fluctuations of the library sizes. In addition, variance of the data was stabilised to enforce the data to be homoscedastic. Batch correction, using the function `ComBat` from the Bioconductor package sva [Lee+12], was applied to correct for sex and batch. Using the dimension reduction technique Uniform Manifold Approximation and Projection (UMAP) [MHM18], the dimension of the data was reduced to 2D (Figure 3.3). The UMAP hyperparameters in R were set to n_components = 3, n_neighbors = 20, min_dist = 0.2, random_state = 123.

3.2.1.2 Endotype identification

To identify disease endotypes, I focused exclusively on L skin samples (n=342). Since most clustering algorithms are not designed for high-dimensional data, Heydar Maboudi Afkham and I conceptualised a pipeline that automatically determines the optimal percentage $x \in]0, 100]$ of HVGs to use in the analysis. This determination was guided by the Davies-Bouldin Index (DBI) and accuracy, which was calculated using ground truth labels as control. The pipeline includes flexible components that allow for alternative feature selection methods. In addition, the pipeline was implemented by me, providing a systematic evaluation of clustering performance across multiple resolutions.

The pipeline currently includes three methods to determine HVGs, i.e. (i) the mean-variance relationship implemented in the `highly_variable_genes` function from SCANPY [WAT18], (ii) a simpler method that selects genes based on their standard deviation, referred to as SD in this study, and (iii) a novel method developed by Heydar Maboudi Afkham and me, referred to as highly relevant gene (HRG).

In order to define HVGs, I employed HRG, which involves the following steps. First, each gene's counts are normalised by subtracting its mean value across all samples, resulting in so-called *gene profiles*. These gene profiles are then clustered using K-Means into $k = 50$ groups. The resulting clusters are sorted by their standard deviation, and only the top x % of genes are retained for further analysis. The selected data is subsequently normalised using library size normalisation and visualised in two dimensions using UMAP [MHM18], with the the random_state parameter set to 0.

Following feature selection, I applied Leiden clustering across a range of resolutions ($\gamma \in \{0.1, 0.2, \ldots, 1.5\}$). The performance of the clustering was evaluated on the test set by comparing the DBI values and the accuracy, which was calculated using the ground truth, here the IRP annotations. This evaluation was performed for each combination of resolution γ and percentage of genes to keep $x \in \{0.5, 1.0, 2.0, \ldots, 100\}$. The optimal value of x was identified as the percentage that yielded the lowest mean DBI and the highest mean accuracy across all resolutions and random seeds. Due to the high imbalance of the dataset and for greater granularity in the clustering than the IRPs, a resolution γ of 0.9 was selected for further analysis.

To assess the clustering stability for the resolutions $\gamma \in \{0.1, 0.2, \ldots, 1.5\}$, I calculated the Adjusted Mutual Information (AMI) [Rom+16] (chapter 2.4.2.3). Specifically, the AMI scores were calculated for each value of γ and for each initialisation value from the set of values $\{0, 42, 50, 100, 3210, 500, 300, 5000, 123\}$. Higher AMI scores indicate a better agreement between the clustering result, which is the desired outcome. Further, I calculated the DBI to assess the compactness and similarity between the clusters [DB79] (Chapter 2.4.2.3). Lower values imply a better clustering of the samples.

To uncover hierarchical relationships among the identified endotypes, I first normalised the data using trimmed mean of M-values (TMM) values, corrected for batch effects and sex, and scaled the data into z-scores. Then the mean of each cluster was taken. Finally, I used the hclust package [Tea22] in R with the linkage method “ward” on a predefined cosine distance matrix to create the Dendrogram. Based on the built branches I renamed the clusters from 0-12 to E1-E13.

3.2.1.3 Variance partition analysis

In order to perform the variance partition analysis, using the R package variancePartition [HS16], I prepared the count matrix by first calculating the normalisation factors using edgeR’s function `calcNormFactors` [RMS10]. Next, I set the design function in the DGEList object to ~ 1 and apply QC on the count matrix by requiring a gene to be expressed in at least two samples and having at least one read count. Finally, the counts were transformed to log2 counts-per-million (logCPM) using the function `voom` from the R package limma [Rit+15]. In addition, I checked whether any of the variables of interest were highly correlated, by computing the canonical correlation analysis, using the function `canCorPairs`. By conducting the variance partition analysis using the design function

$$y \sim \frac{1}{\text{diag}} + \frac{1}{\text{IRP}} + \frac{1}{\text{Endotype}} + \frac{1}{\text{batchID}} + \frac{1}{\text{sex}} + \text{age}\,, \tag{3.1}$$

I observed a high variance of 0.035 of the variance explained values of *batchID* compared to the other variables. Therefore, I corrected for technical artefacts originating from this factor. Afterwards, I repeated the analysis on the corrected residuals using the design

$$y \sim \frac{1}{\text{diag}} + \frac{1}{\text{IRP}} + \frac{1}{\text{Endotype}}\,. \tag{3.2}$$

3.2.1.4 Differential gene expression and pathway enrichment analysis

I run DGE analyses between all Endotypes, one endotype vs. all other endotypes, and all Dendrogram branches using edgeR [RMS10]. For each subset, the data was filtered for low expressed genes, as described above, and normalised using TMM values. The following design function was used for the comparisons between endotypes, one vs. all other endotypes, and the dendrogram splits

$$y_{gs} \sim \text{age} + \text{sex} + \text{batchID} + \text{Endotype} \ , \tag{3.3}$$

where y_{gs} is the normalised TMM value of a gene g in sample s. I required a padj value < 0.05 and $|\text{log2FC}| > 1$ to call a gene significantly differentially expressed between two conditions.

All pathway enrichment analyses were performed using the Bioconductor [Gen+04] packages ReactomePA [YH16] and org.Hs.eg.db [Car+19]. The enriched pathways were illustrated using the package Enrichplot [Yu21]. All measured genes in the dataset were provided to the function `enrichPathway` as universe parameter. P-values were false discovery rate (FDR) corrected using Benjamini and Hochberg (BH). Enriched pathways had to meet the significance level padj value < 0.05.

3.2.2 Clinical attributes

In total 86 clinical attributes were collected by Peter Seiringer and Natalie Garzorz-Stark for each patient. These include information about age, sex, severity scores, comorbidities, laboratory assessments, therapy associated information, history associated information, histology assessments, and morphology assessments (Section 3.1). I divided the 86 clinical attributes into 24 nominal, 30 ordinal, and 32 continuous attributes.

3.2.2.1 Processing of clinical data

Nominal categories were encoded using OneHotEncoder and ordinal categories were encoded using an OrdinalEncoder by leveraging the package Scikit-learn [Ped+11] in python. Further, I removed NaN categories. Through the encoding and removal of NaN categories, I received in total 26 encoded nominal categories. Attributes consisting out of a single category were removed. After encoding and filtering, I continued with in total 88 (26 nominal, 30 ordinal, and 32 continuous) clinical attributes. Continuous and ordinal data were scaled using `MinMaxScaler` function from Scikit-learn [Ped+11] in Python.

I further accounted for missing values. In order to apply the assumption that data is Missing completely at random (MCAR), I tested for possible relations between categorical attributes, using the one-way ANOVA test, and between categorical and continuous attributes, using the χ^2 test. Furthermore, I corrected for FDR using BH. I found relations between categorical attributes in 301 out of 1420 comparisons and between categorical and continuous in 121 out of 1398 cases. Combining this information with expert domain knowledge, I claimed that the missing data is MCAR. In order to account for data MCAR, the clinical attributes were imputed using the `KNNImputer` in Scikit-learn [Ped+11], with the parameter k set to one and the ordinal_categories parameter set to the names of the ordinal attributes.

3.2.2.2 Attribute enrichment in endotypes

In this analysis, I first removed the disease severity scores, degree and speed of therapy response, and the therapeutic target. The final number of encoded attributes tested for an enrichment is 77, consisting of 22 nominal, 24 ordinal, and 31 continuous attributes. Subsequently, clinical traits were encoded and missing values were imputed. The attribute with the highest percentage of 59.65 % missing data was Laboratory specific IgE.

In order to perform an enrichment test for an attribute between an endotype and the remaining endotypes, I categorised them into continuous, ordinal, and nominal phenotypes. Continuous attributes were first assessed for normality and homogeneity of variance. As all continuous attributes rejected the Shapiro-Wilk and subsequent Levene tests, I used the Kruskal-Wallis test to test for enrichment of an attribute in an endotype. Similarly, the Kruskal-Wallis test was used for ordinal categories. For nominal clinical attributes, the χ^2 test was used. An attribute was considered enriched if it met the significance level padj value < 0.05. In this thesis, I considered only those clinical attributes having a highly significant p-value below $1e\text{-}05$ in at least one endotype. The statistical tests were conducted using SciPy [Vir+20] in Python.

Effect sizes were calculated using Cohen's d, Cliff's Delta, and odds ratio for continuous, ordinal, and nominal variables, respectively. Cohen's d is defined within $-\infty$ to ∞. Absolute values of approximately 0.2, 0.5, and 0.8 mark small, medium, and large effects, respectively [Coh13]. For calculating Cliff's Delta, the Python package cliffsDelta was used [NE21]. Its values range from $[-1, 1]$. Values close to ± 1 indicate an absence of an overlap, and 0 indicates a complete overlap between two groups [MRL11]. The odds ratio takes values in the range $[0, \infty[$. An odds ratio of 1 indicates that the attribute is equally likely in both groups. Ratios less than 1 suggest that the attribute is less likely in the first group compared to the second group, while ratios greater than 1 suggest the opposite [Szu10].

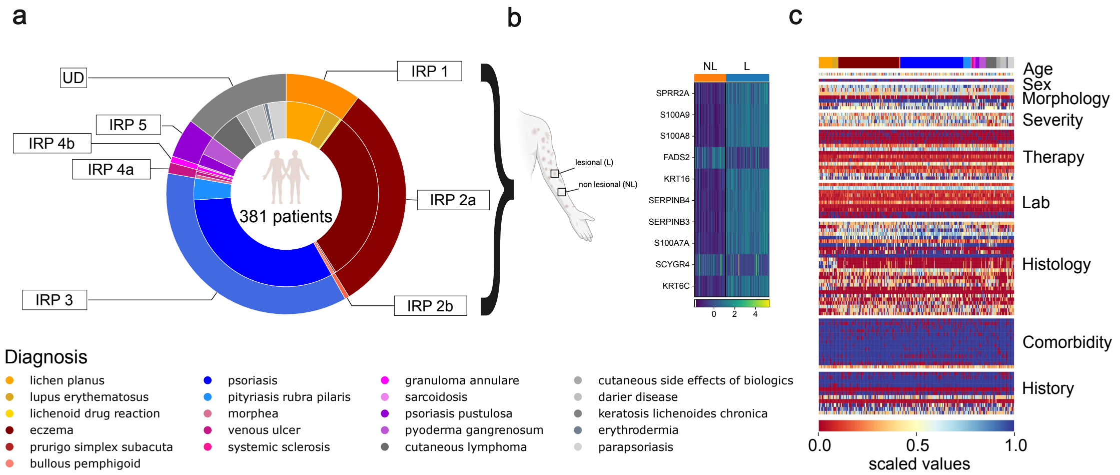


Figure 3.1: Rich data resource including gene expression data and clinical traits. a) Diagnosis and IRP annotations for each of the 381 patients (n_L=342, n_{NL}=287). In total 21 skin diseases associated with seven IRPs and one undefined (UD) IRP are in the final processed data. The inner ring shows the ncISDs that are associated with the IRPs shown in the outer ring. **b)** Gene expression profiles from L and NL skin are measured and **(c)** 86 clinical attributes are collected. Encoded, imputed and scaled attributes are shown in the heatmap.

3.3 Results

In order to categorise ncISDs into objective endotypes, with the aim of enhancing diagnostics and providing treatment options also for rare diseases, I leveraged a large transcriptomics landscape of 21 ncISDs and their corresponding IRPs (Figure 3.1 a). After performing the preprocessing, a total n=639 samples from 381 patients remained. Of these, 246 patients donated paired L and NL biopsies, 95 patients provided only L, and 40 only NL samples (Figure 3.1 a, b). Most samples were collected from patients with plaque psoriasis (n=193), eczema (n=180), and lichen planus (n=47), which made up two thirds of the cohort. Hereafter, I refer to plaque psoriasis as psoriasis. Additionally, 86 patient specific attributes such as morphology, comorbidities, and history of the patient were encoded, imputed, and filtered, resulting in 88 individual phenotypic fingerprints (Figure 3.1 c, Methods 3.2.2). This large and comprehensive cohort including transcriptomics and deep clinical phenotyping will assist me in defining endotypes, a refined patient stratification approach to advance precision medicine in ncISDs.

3.3.1 Diagnoses result in heterogeneous clinical and transcriptomics profiles

I investigated similarities between patients based on their clinical profiles by projecting the data in 2D using UMAP (Figure 3.2 a). The resulting embedding revealed apparent clusters of certain diseases, including psoriasis, lichen planus, and lupus erythematosus. While psoriasis samples were more widely distributed across the embedding, the majority appeared concentrated in the lower left corner. Diseases associated with IRP 1, i.e., lichen planus, lupus erythematosus, and lichenoid drug reaction, displayed distinct clinical characteristics that allowed them to be well differentiated from other ncISDs. In contrast, eczema and other diseases, such as pityriasis rubra pilaris, bullous pemphigoid, and cutaneous lymphoma, showed no apparent clusters and were distributed throughout the embedding. These observations reveal both the expected overlaps among diseases and the high heterogeneity within patients diagnosed with the same condition, strengthening the assumptions that the clinical assessments alone are insufficient to achieve a definitive diagnosis for a patient.

Examining three exemplarily chosen psoriasis patients (Figure 3.2 a, black circles) revealed distinct positions in the embedding. Two patients were located in close proximity in the embedding, indicating similar clinical fingerprints, with the primary difference being the attribute of morphology (Figure 3.2 e, h). In contrast, the third patient was positioned further away, closer to IRP 1, and exhibited a distinct clinical profile compared to the neighbouring psoriasis patients (Figure 3.2 b). An assumptions is that patients with similar clinical profiles are likely to exhibit comparable responses to an administered therapy, whereas those with dissimilar profiles would respond differently. However, considering the example images of psoriasis or psoriasis pustulosa patients before and after the treatment with an IL-17 inhibitor (Figure 3.2 d, g, j), demonstrated that this assumption does not consistently hold true. Conclusively, these findings indicate that clinical assessments do not provide sufficient evidence for an accurate diagnosis and treatment recommendation of a patient.

Given the high heterogeneity observed in the clinical profiles and the disease and IRPs annotations derived from these profiles, I investigated whether similar patterns could be identified at the gene expression level. To explore this, I utilised L and NL gene expression data and visualised it in two dimensions, highlighting diseases and IRP annotations (Figure 3.3). As anticipated, L and NL skin samples segregated into distinct clusters, with only a few outlier L samples located within the NL cluster. Interestingly,

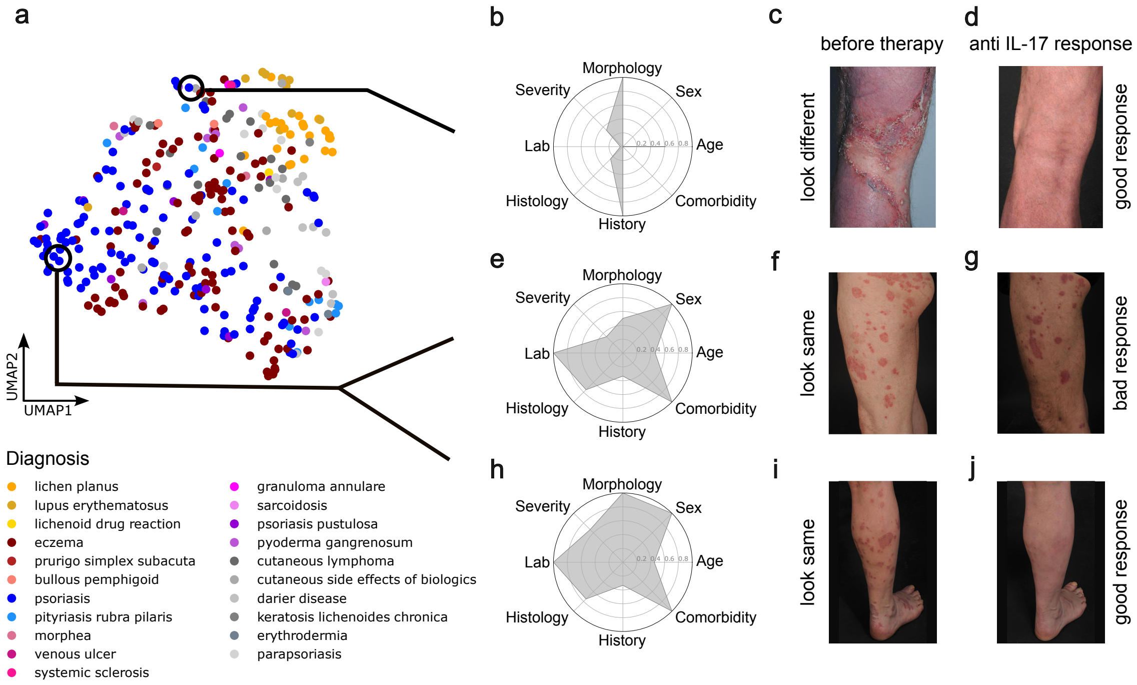


Figure 3.2: Clinical traits are not sufficient for diagnostic purposes. a) Clinical profile embedding revealing the heterogeneity within and between diseases. **b, e, h)** Radarplots of three psoriasis patients showing the mean over all attributes belonging to the same category. The **(b)** patient at the top has a distinct profile from the other **(e, h)** two psoriasis patients. **c, f, i)** Photos of three psoriasis or psoriasis pustulosa patients before treatment. The **(c)** patient in the top row has distinct visual symptoms compared with the patient **(f)** in the middle and **(i)** at the bottom. Both show the typical red, scaly psoriasis symptoms. **d, g, j)** The same patients after treatment with an IL-17 inhibitor. **(d)** The patient, having a distinct clinical profile and visual symptoms, responds equally well to the drug target as **(j)** the patient shown at the bottom. Interestingly, the **(g)** patient in the middle, with similar symptoms to the patient at the bottom, does not respond to treatment.

some NL samples were positioned between L and NL clusters. This may indicate that these samples represent an intermediate state between healthy and diseased skin or were collected from areas too close to L skin. These findings indicate that gene expression data reflects the heterogeneity observed in clinical profiles and highlights potential transitional states between NL and L skin, offering insights into the complexity of patient stratification.

There were also apparent clusters of lichen planus, pyoderma gangrenosum, eczema, and psoriasis and their corresponding IRPs 1, 5, 2a and 3. Both eczema and psoriasis exhibited particularly heterogeneous profiles, which was reflected in their widespread distribution in

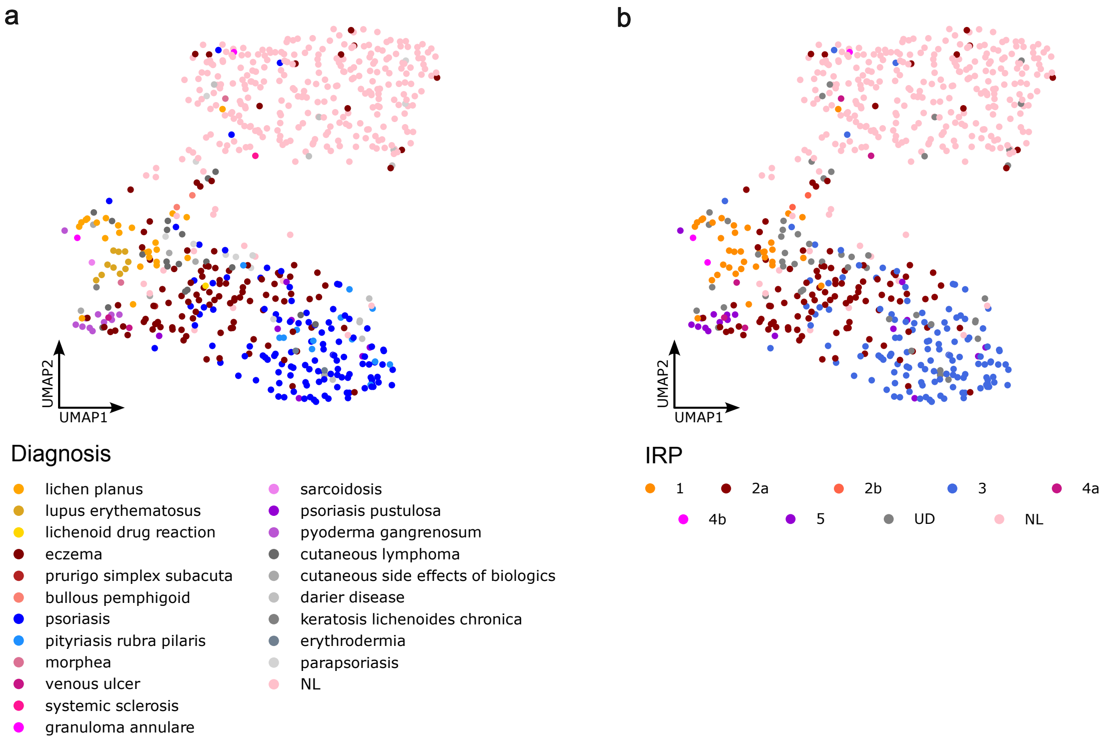


Figure 3.3: Gene expression embedding of 17, 816 genes reveals overlap between diagnoses and IRPs. UMAP of 342 L and 287 NL samples shows the transcriptomics profiles of patients and their similarities between **(a)** diagnoses and **(b)** IRPs.

the UMAP embedding. Notably, some patients did not align with their expected disease clusters. For instance, some patients diagnosed with eczema and psoriasis were observed to be located in each other's clusters. In addition, another apparent cluster contained patients from a range of diverse diseases. This was also evident at the IRP level. While most diseases are aligned to an IRPs, some patients' IRPs remained undefined. I refer to this undefined pattern as UD IRP. At the gene expression level, these patients frequently clustered with those belonging to already defined IRPs (Figure 3.3 b). This finding suggests that rare diseases, which are not yet assigned to specific IRPs, may benefit from a data-driven approach to disease endotype definition. In summary, the high heterogeneity observed in diseases and IRPs at the clinical level was mirrored in the gene expression data, suggesting that endotypes derived by a data-driven approach could improve patient stratification.

In summary, clinical fingerprints do not provide sufficient information about the underlying disease and therefore do not guarantee response to therapy. This was observed considering the heterogeneous gene expression profiles within diseases. In conclusion, an alternative approach to patient stratification is needed. I therefore hypothesise that gene expression could be used to groups patients into disease endotypes providing an objective view in addition to the current disease ontology. In addition, groups of patients may be created that include less diseases with heterogeneous phenotypes and rare skin conditions. In such cases, drug repurposing could be applied and patients diagnosed with a rare disease could be offered a therapeutic approach. In the following, I introduce a framework to determine HVGs, which enables the clustering of patients into endotypes using transcriptomics.

3.3.2 Framework to identify highly variable genes and endotypes

For the identification of endotypes, I applied unsupervised ML to cluster the transcriptomes of skin lesions. Due to the data's high-dimensional nature, the clustering algorithm might fall for the *curse of dimensionality.* To mitigate this issue, it is common practice to reduce the dimensionality by selecting HVGs.

The HVGs expression profile strongly varies across the dataset and assist in distinguishing between dissimilar patients or cells. Several methods [Stu+19] [Bre+13][Zhe+17] have been developed to identify HVGs, often leveraging statistical techniques to rank genes based on their deviation from an expected distribution. However, these approaches have a limitation, i.e. the number of HVGs to be selected must be manually set [LT19]. This is an arbitrary process, potentially leading to bias in the analysis. Addressing this limitation is essential for ensuring that feature selection remains objective and scientifically reproducible.

To address this limitation, I developed a method to determine the optimal percentage of HVGs for utilisation in the downstream analysis. This approach uses the DBI, and optionally the accuracy if ground truth labels are available. In addition, I proposed an alternative method for HVGs selection to the established dispersion- and variance-based approaches [Zhe+17] [Bre+13] [Stu+19]. Specifically, my feature selection method was designed to (i) form groups whose primary differentiating mechanism is not variance and (ii) avoid selecting genes that convey redundant information across samples. This approach prioritises shared information between genes and offers complementary insights. Selecting a small subset of biologically essential features that provide complementary information not only improves the performance of clustering algorithms but also enhances downstream analyses. This framework ensures better stratification and interpretability of patient data, contributing to the advancement of data-driven endotyping.

3.3.2.1 Methodological framework of HVG selection

The highly variable feature selection pipeline operates on a raw L gene expression matrix, requiring information on potential batch effects and, optionally, ground truth data. The dataset is divided into a training set (80 %) and a test set (20 %). Feature selection is performed exclusively on the training set, while evaluation is conducted on both subsets. Hence, the pipeline is designed to optimise feature selection by splitting the dataset a for robust evaluation.

To prepare the training set for feature selection, library size normalisation is performed (Section 2.5.1.2). This step ensures that the resulting logCPM counts are comparable across samples by correcting for differences in sequencing depth. Following normalisation, batch correction is applied to address technical and biological artefacts, thereby minimising unwanted variation and enhancing the biological signal.

Next each gene's counts are scaled. This is achieved by subtracting the mean expression of a gene across all samples

$$\widetilde{\boldsymbol{y}}_{\boldsymbol{g}} = \overline{y}_g - \frac{\sum_{i=0}^{n} y_{g,i}}{n},$$

where $\widetilde{\boldsymbol{y}}_{\boldsymbol{g}}$ is the scaled gene expression vector of a gene g, $\overline{y}_g$ is the mean expression of a gene over all samples, y_g contains the raw counts of a gene for each sample, and n is the total number of samples. Scaling gene counts standardises the data, enabling to compare gene expression values on the same scale.

Feature selection is applied to the training set to identify the most variable genes. Three feature selection tools are available, including (i) a dispersion-based method as described by [Zhe+17] and implemented in SCANPY [WAT18], (ii) a method based on standard deviation, or (iii) my proposed alternative feature selection method, HRG.

In the HRG approach, the scaled gene counts are sorted $\widetilde{y}_{g,1} \geq \cdots \geq \widetilde{y}_{g,n}$ (Figure 3.4). This sorting ensures that the genes are independent of sample order, revealing the underlying distribution of each gene. These distributions, referred to as gene profiles, exhibit distinct patterns that can be categorised as highly variable, intermediate, or constant. Subsequently, K-Means clustering is then applied to group the gene profiles into k gene profile clusters. For each cluster, the standard deviation (σ) is first computed for each gene. Then, the mean of these standard deviations is calculated, resulting in the mean σ_k for cluster k. This process is repeated for all clusters, which are then sorted in descending order based on their mean standard deviations ($\sigma_1 \geq \cdots \geq \sigma_k$). The potential

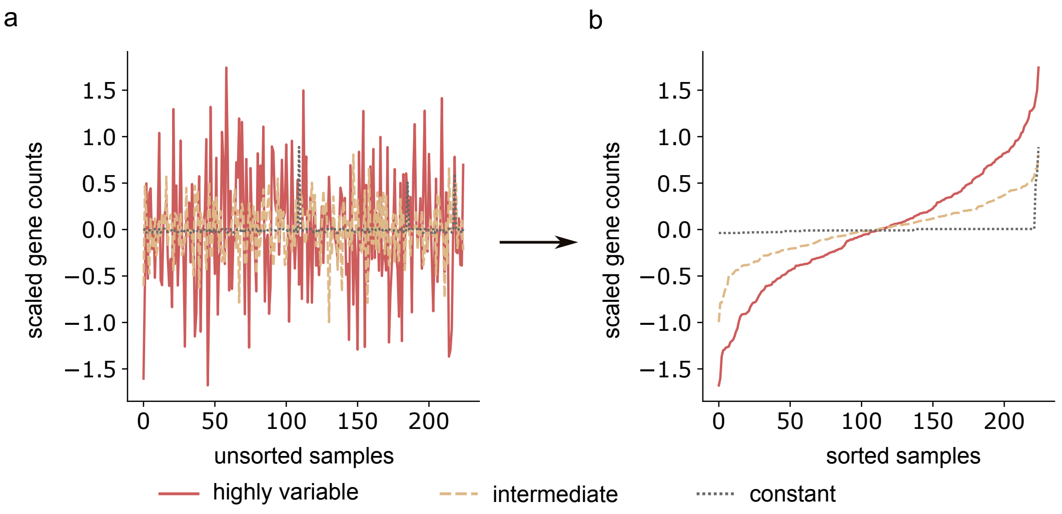


Figure 3.4: Example showing the distribution of genes before and after sorting. Examples of a highly variable, intermediate, and constant expressed gene profile shows their distribution **(a)** before and **(b)** after sorting each gene by its scaled expression.

HVGs are then selected by retaining a predefined percentage $x \in]0, 100]$ of genes from the ranked gene clusters. The HRG method identifies and ranks gene clusters based on variability, ensuring robust feature selection.

Leiden clustering is applied to group samples based on the HVGs, optimising resolution for biological relevance. Using the reduced training data with x selected features, a K-nearest neighbour (KNN) graph is constructed using the samples. Leiden clustering is applied, with the resolution parameter γ varied to match the dataset size and expected biological groupings. A gridsearch is performed by varying the values of x, γ, and the initialisation value (seed).

The clustering performance is evaluated by calculating the DBI and, optionally, the accuracy on both training and test set. The accuracy involves assigning ground truth labels to the Leiden clusters by identifying the most frequent label within each cluster. It is possible for a single label to be assigned to multiple clusters. The accuracy can then be calculated using the eq. B.10. As it requires true positive (TP) and true negative (TN) samples, I define samples, whose label matches the ground truth label, as TP, while TN samples mark the opposite. The optimal percentage of features x_{opt} is determined by the lowest mean DBI or highest mean accuracy across all resolutions γ and initialisation values. The selected features x_{opt} are the final HVGs. In summary, the pipeline ensures

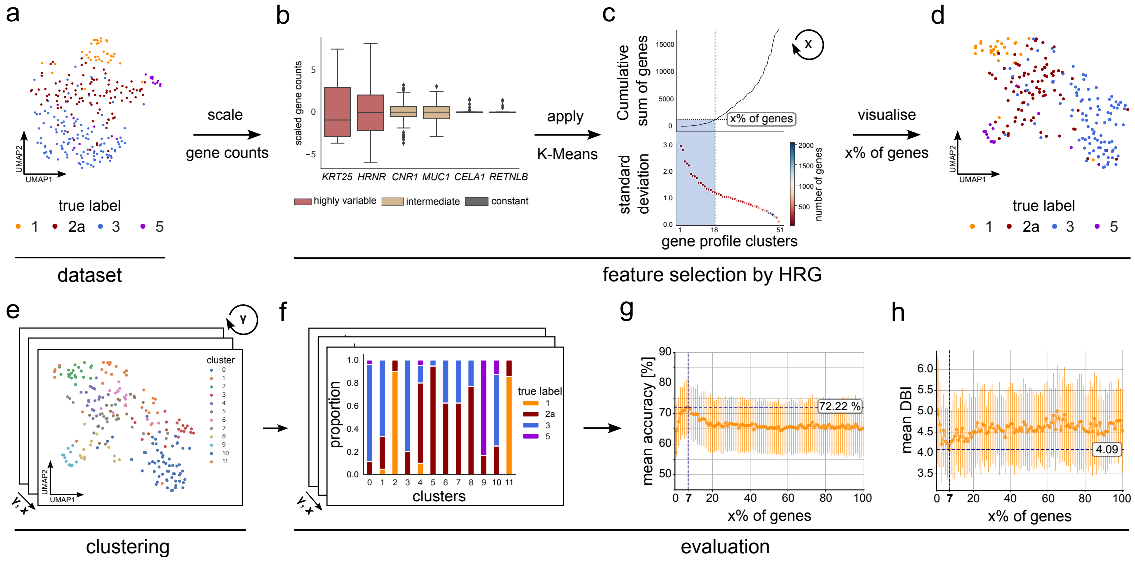


Figure 3.5: Workflow of determining HVGs. a) The most frequent IRPs in the dataset are used to identify endotypes. **b)** Samples are normalised. Each gene's counts are scaled and the genes are sorted by their standard deviation across all samples. Examples of highly variable, intermediate, and constant gene profiles are shown. These differ in their variance of expression and therefore, in the size of the standard deviation. **c)** Workflow of the feature selection method HRG. K-Means clusters the gene profiles into $k = 50$ clusters. The clusters are sorted by the standard deviation of the genes in a cluster. The number of genes per cluster increases as the standard deviation decreases and the cumulative sum of genes increases exponentially. The dashed, horizontal line marks the number of genes, i.e. HVGs, corresponding to $x\,\%$. The blue area and the dashed vertical line mark the gene profile clusters being considered for the gene selection. The selection procedure is repeated iteratively for $x \in \{0.5, 1, 2, \ldots, 100\}$. **d)** The selected HVGs are visualised in 2D using UMAP. **e)** For each value of x, the Leiden clustering, for the resolutions $\gamma \in \{0.1, 0.2, \ldots, 1.5\}$, is applied. **f-h)** Performance evaluation of the feature selection method. **f)** In order to calculate the accuracy, the number of TP and TN samples is defined by determining the true label composition of each Leiden cluster. Subsequently, the mean accuracy **(g)** and DBI **(h)** across all resolutions are calculated revealing the optimal percentage value for $x_{opt} = 7$ (vertical dashed lines).

that the final HVGs are optimally selected based on the robustness and accuracy for robust downstream analyses.

In summary, the pipeline provides a way to automatically identify HVGs. It incorporates normalisation, feature selection, clustering, and performance evaluation, providing a robust framework for downstream analyses.

3.3.2.2 Application of the feature selection pipeline

To identify HVGs, I selected the main IRPs 1, 2a, 3, and 5 and applied the developed framework to the L data (n=282) (Figure 3.5 a).

The first step involved defining gene profiles in the training set through scaling and sorting of gene counts (Figure 3.5 b). This process revealed distinct characteristics across gene categories. HVGs exhibited a large interquartile range and long whiskers, indicating high variability. In contrast, intermediate genes showed less variation, while constant genes displayed no variability across the population. Thus, the scaling and sorting of genes revealed distinct variability patterns of the genes.

Next, I applied my feature selection method, HRG (Figure 3.5 c). Gene profiles were grouped into $k = 50$ clusters using K-Means clustering and sorted based on the standard deviation of their expression profiles. Notably, the standard deviation indirectly correlated with the number of genes in a cluster. The cumulative sum of the number of genes across the clusters followed an exponential trend. In combination with the standard deviation, this could indicate that only a few genes have highly variable and distinct expression profiles compared to the majority of genes, which exhibited almost no variation across the population. The top x % of genes, as determined by HRG, were selected as HVGs. In the event that the requested number of genes did not encompass all genes within a cluster, as this number exceeds the required number of genes, the genes were selected according to their order within the cluster. As K-Means groups genes with similar profiles, it can be reasonably assumed that this selection approach did not affect the result. In essence, HRG selected HVGs by ranking gene clusters based on variability, focusing on genes with distinct expression patterns.

The selected HVGs were visualised in a 2D embedding (Figure 3.5 d). I repeated the feature selection process for $x \in \{0.5, 1, 2, \dots, 100\}$. For each percentage x, Leiden clustering was applied on the reduced dimension for $\gamma_x \in \{0.1, 0.2, \dots, 1.5\}$ (Figure 3.5 e). I repeated the feature selection and clustering procedure for different initial values (seeds), as the clustering algorithm is non-deterministic (Figure 3.5 c-e). This approach ensures to choose a percentage x, which provides a robust selection of HVGs and clustering result.

To optimise gene selection and IRP annotation, I identified the optimal percentage of genes x_{opt}, that maximised both accuracy and clustering quality. The optimal percentage of genes, $x_{opt} = 7$, resulted in the highest mean accuracy of $(72.22 \pm 9.30)\%$ and the lowest

k	**x**	**γ**	**initialisation value**
50	$\{0.5, 1, 2, \ldots, 100\}$	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5}	{0, 42, 50, 100, 3210, 500, 300, 5000, 123}

Table 3.1: Feature selection parameter to identify HVGs using HRG, SCANPY, or SD.

mean DBI of 4.09 ± 0.83 across all resolutions γ and seeds (Figure 3.5 g, h). Notably, the feature selection pipeline automatically determines which genes to define as HVGs by performing a gridsearch over the parameters x, γ, and seed. Thus, by considering both accuracy and DBI, the pipeline enables the construction of robust, biologically relevant clusters based on the identified HVGs.

In summary, using the developed feature selection pipeline, I identified the optimal percentage of HVGs in the L dataset ($x_{opt} = 7$). By varying the parameters, the pipeline ensures robust results, yielding biologically meaningful clusters, thereby paving the way for disease endotype identification.

3.3.2.3 Comparing HVG selection approaches

To evaluate the performance of my feature selection method HRG, I compared it against state-of-the-art tools, including the standard deviation, which I refer to as SD, and the mean-variance relationship implemented in SCANPY [WAT18]. Specifically, I compared (i) the metric scores, (ii) the number of selected genes required for clustering, and (iii) the robustness across initialisation values.

The comparison of metric scores involved the evaluation of the mean clustering accuracy and mean DBI (Figure 3.6 a, b). The one-sided Wilcoxon signed-rank test indicated that HRG demonstrates superior performance. It achieved a general higher average accuracy (HRG vs. SCANPY: p-value = $5.05e\text{-}15$, HRG vs. SD: p-value = $1.92e\text{-}02$) and lower mean DBI (HRG vs. SCANPY: p-value = $1.63e\text{-}04$, HRG vs. SD: p-value = 1.00) across all resolutions and initialisation values. HRG outperformed SCANPY and SD in metric scores, demonstrating higher accuracy and lower DBI.

The number of genes selected for clustering was evaluated based on the average accuracy and DBI across all initialisation values and resolutions γ (Figure 3.6 a, b; Table 3.1). For SCANPY, I identified the optimal percentages of $x_{opt} = 6$ and $x_{opt} = 24$, resulting in an accuracy of (69.59 ± 7.60) % and DBI of 4.36 ± 0.83, respectively. SD required the

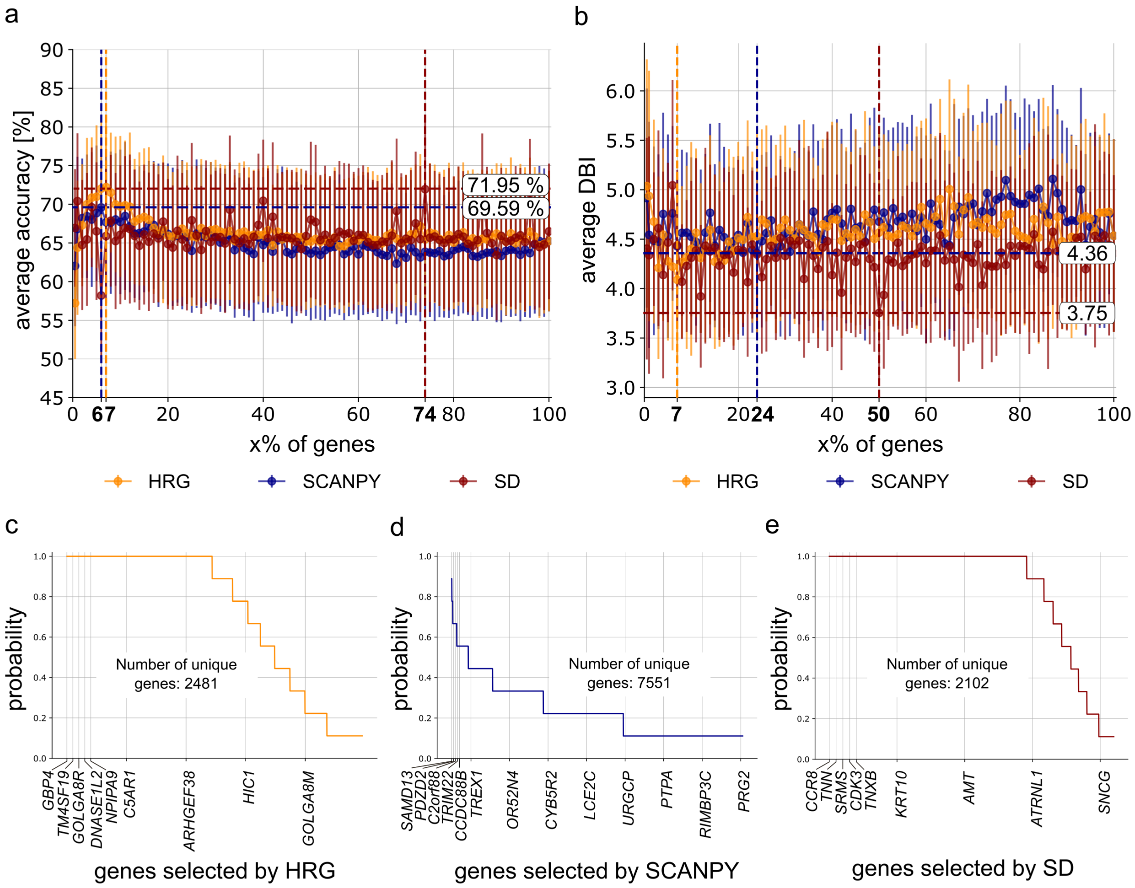


Figure 3.6: Comparison of feature selection methods to identify HVGs. a, b) Average **(a)** accuracy and **(b)** DBI across all resolutions γ and initialisation values. Dashed lines mark the highest and lowest scores of the mean accuracy and DBI, respectively. The colours indicate the method used for feature selection. The optimal percentage of HRG (orange) is 7 for both metrics, 6 and 24 for SCANPY (blue), and 74 and 60 for SD (red). **c-d)** Probability for $x = 10$ of a gene to be selected in each initialisation round by **(c)** HRG, **(d)** SCANPY, or **(e)** SD. Examples of genes for the same or different probability are shown. In total $2,481$, $7,551$, and $2,102$ genes are selected by HRG, SCANPY, and SD, respectively.

most genes, i.e. $x_{opt} = 74$ and $x_{opt} = 50$, for clustering. Its highest average accuracy was $(71.95 \pm 8.90)\,\%$ and lowest DBI was 3.75 ± 0.68. A comparison of the mean accuracy of the optimal percentage of genes to retain for each method revealed that HRG ($1,248$ genes) required a greater number of genes than SCANPY ($1,067$ genes) and less than SD ($13,184$ genes). However, considering the DBI, I found that HRG requires less genes than SCANPY and SD (SCANPY: $4,276$ genes, HRG: $1,248$ genes, SD: $8,908$ genes). Thus, HRG required less genes, which are sufficient for forming biologically meaningful clustering.

I was also interested in the robustness of feature selection across all initialisation values (Figure 3.6 c-e). Therefore, I investigated the probability of a gene being selected by setting the percentage of genes to be retained to $x = 10$. I observed that HRG is more robust in comparison to SCANPY and demonstrates comparable robustness to SD. This was reflected by the selection probability of 100 % for $1,219$ and $1,458$ genes for HRG and SD, respectively (Figure 3.6 c, e). In contrast, no gene occurred in each initialisation using SCANPY (Figure 3.6 d). Moreover, HRG selected $2,481$ unique genes across all initialisation values, similar to SD ($2,102$ unique genes), and three time less than SCANPY ($7,551$ unique genes). HRG showed superior robustness to SCANPY and was comparable to SD, selecting stable gene sets across different initialisation values.

In conclusion, HRG was the best-performing feature selection method for this dataset. It achieved superior metric scores with 7 % of genes, demonstrated greater robustness than SCANPY, and required fewer genes for clustering compared to SD. However, the performance of HRG may vary across datasets. Thus, I recommend benchmarking multiple feature selection approaches across diverse datasets to identify the most suitable method for a specific application. In summary, HRG outperformed SCANPY and SD in this dataset, but its generalisability requires further validation through benchmarking.

3.3.2.4 Identifying robust clusters of endotypes

In order to define endotypes, I accounted for the heterogeneity of the dataset and required the clustering to be more granular than the IRPs, including the UD IRP. Therefore, I aimed to define more than seven clusters, which required setting the clustering resolution γ above 0.5, as shown by the linear relationship between resolution and number of clusters (Figure 3.7 a).

To ensure stable clustering across different initialisation values, I utilised the AMI (Methods 3.2.1.2), achieving a score of 0.64 ± 0.01 for $\gamma \in \{0.6, 1.5\}$ (Figure 3.7 b). Additionally, I shortlisted potential cluster resolutions using the DBI (Figure 3.7 c). The lowest DBI of 5.81 ± 0.56 was observed at $\gamma = 0.6$, with an average increase of 0.13 ± 0.2 per 0.1 step in γ between 0.7 and 0.11. Despite using only 7 % of the genes ($x_{opt} = 7$), the resulting DBI might have been distorted, due to the limitations of euclidean distance in high-dimensional datasets [AHK01] [Bey+99]. Thus, I combined the objective assessment metrics with expert knowledge, and selected a resolution of $\gamma_{opt} = 0.9$ to determine the disease endotype clusters.

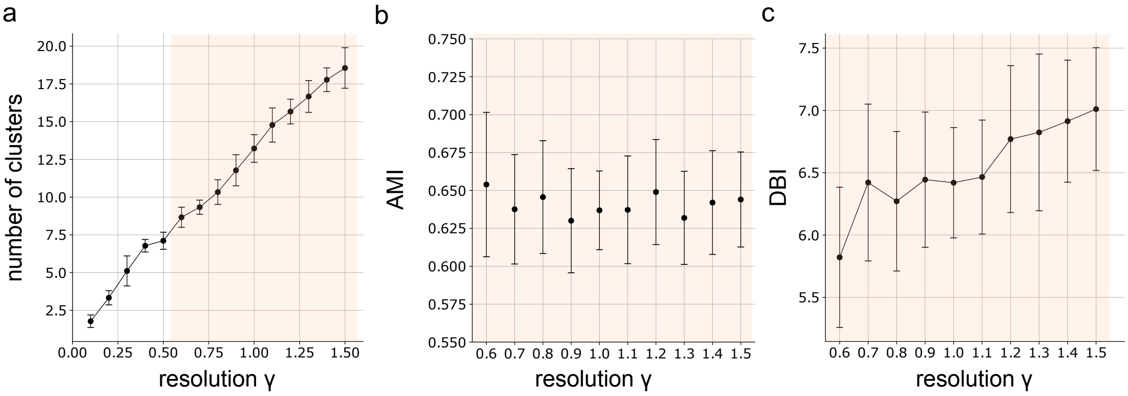


Figure 3.7: Determine cluster resolution using AMI and DBI. a) Linear relation between number of clusters and resolution parameter γ. **b)** Metric to assess the stability of clustering results for all resolutions using AMI. **c)**. DBI to assess the compactness of and similarity between clusters. Area of interest ($\gamma > 0.5$) is highlighted in orange.

Summarising Section 3.3.2, a feature selection pipeline was designed to automatically determine the percentage of HVGs. Additionally, I developed a feature selection method, HRG. Comparative analysis against established tools such as SCANPY and SD demonstrated that HRG performs best for this dataset. In particular, I found that HRG exhibits superior performance in comparison to SCANPY and SD by requiring less genes for clustering. HRG was also more robust in its feature selection than SCANPY. Leveraging my pipeline, I found that only 7 % of genes, i.e. HVGs, were required for the identification of disease endotypes. These findings suggest that the pipeline and HRG could significantly assist in the gene selection process for downstream analysis, potentially leading to more precise disease endotypes or subtypes.

Subsequently, I related the identified endotypes to the clinical diagnoses and IRPs, and further investigated whether the endotypes are superior towards the traditional disease ontology.

3.3.3 Endotypes capture heterogeneity of skin conditions

I identified 13 endotypes by running the Leiden clustering algorithm on HVG using the determined optimal resolution of $\gamma = 0.9$ (Figure 3.8 a). These endotype clusters represent distinct disease groupings based on their gene expression profiles, providing insights into the underlying heterogeneity of the skin conditions. In the following, I analyse the composition of each endotype in relation to clinical diagnoses and IRPs (Figure 3.8 b).

Several endotypes were predominantly characterised by specific diseases. Specifically, the endotypes **E1** and **E7** were primarily composed of eczema (E1: 69 % and E7: 71 %) patients. Endotype **E2** was characterised by a distinct gene expression profile dominated by pyoderma gangrenosum (62 %), which belongs to IRP 5. This skin condition was the most prevalent disease and IRP in E2 (Figure 3.8 b). Additionally, diseases associated with IRP 1, including lichen planus and lupus erythematosous, clustered separately in **E3** (67 % lichen planus:) and **E4** (82 % lupus erythematosous) (Figure 3.3, Figure 3.8 b). This shows that these conditions were well defined by clinical assessments. Endotype **E5** primarily consisted of pityriasis rubra pilaris (45 %), a condition suspected to be associated with psoriasis and IRP 3 (Figure 3.8 b). Notably, this condition did not cluster with the most psoriasis patients in E8 and E11-E13, indicating distinct characteristics on the gene expression level. Overall, these findings demonstrate that specific diseases dominate distinct endotypes, further emphasising the heterogeneity within ncISDs.

Some endotypes displayed more heterogeneous disease compositions, with rare disease associations providing further insights into the gene expression relationships across diverse clinical conditions. Endotype **E6** was mainly composed of Darier's disease (39 %), eczema (33 %), and samples from IRP 3 (17 %) (Figure 3.8 b). Endotype **E9** included eczema (36 %), cutaneous lymphoma (29 %), and parapsoriasis patients (21 %). Parapsoriasis, a potential precursor of skin cancer, may develop into T-cell lymphoma [Kik+93], and these diseases are clinically challenging to differentiate [EG99]. The inclusion of these conditions in a single endotype highlights the complexities of distinguishing between them at the clinical level. Endotype **E10** contained samples from a variety of diseases and IRPs, including eczema (43 %), IRP 4a (9 %), including systemic sclerosis and morphea, as well as IRP UD (26 %), comprising cutaneous lymphoma, Darier's disease, and parapsoriasis. Despite this diversity, eczema samples constituted the majority. These observations demonstrate the potential of using gene expression to refine the grouping of heterogeneous and rare diseases.

The endotypes divide psoriasis into multiple clusters. The clustering analysis revealed that psoriasis is a dominant condition within the endotypes E8, E11-E13, with subtle differences in disease composition (Figure 3.8 b). In endotype **E8**, psoriasis coexisted with eczema as the predominant conditions (psoriasis: 48 % and eczema: 44 %), reflecting their overlapping gene expression patterns. Endotype **E11** was primarily composed of psoriasis (81 %), followed by eczema (10 %) and cutaneous lymphoma (10 %) samples. Similarly, **E12** was primarily composed of psoriasis samples (73 %), along with a smaller proportion of eczema (12 %) and rare diseases (15 %) such as parapsoriasis, Darier's disease, and erythrodermia. In endotype **E13**, psoriasis pustulosa (10 %) and psoriasis (87 %) patients were clustered together due to similar characteristics. This was an expected behaviour, as both belong to the psoriasis disease spectrum. Although they are at opposite ends of the psoriasis spectrum, they share characteristics at the gene expression level [Joh+17]. This clustering observation raised the hypothesis that these patients may have different drug response profiles [Ain+12]. In essence, these findings emphasise the heterogeneity within psoriasis, illustrating that gene expression can be used to distinguish between different subtypes and inform potential treatment approaches.

In summary, the identification of 13 endotypes shows the complexity and heterogeneity within ncISDs. This confirms that the established disease ontology, which is based on clinical features, does not capture the full complexity of these conditions. Moreover, the ability to group rare diseases with more common conditions provides opportunities for drug repurposing. Thus, the endotypes may provide an enhanced approach for patient stratification.

In order to assess whether the endotypes better represent the underlying data distributions in the L transcriptomes, I compared the variance explained by endotypes, diagnoses, and IRPs. Despite some outliers, the endotypes had higher explained variances (11.55 ± 10.57) in comparison to diagnoses (4.95 ± 7.68), and IRPs (1.85 ± 4.43) (Figure 3.8 c, Methods 3.2.1.3). The effect size, given by Cohen's d, between endotype and diagnosis was 0.71, which is considered as a medium effect (Methods 3.2.1.3). In contrast, the effect size between endotype and IRP depicted a strong effect of 1.20. The effect size between IRP and diagnosis was 0.50 and is described as a low effect. This suggested that the objective and data-driven categorisation of ncISDs patients into endotypes more accurately reflects the underlying heterogeneity of the data. It is possible that the downstream analysis will be improved in terms of interpretability and quality of the identified target groups.

These 13 endotypes revealed the heterogeneity of diseases and IRPs at the gene expression level. They also grouped rare diseases with other more common diseases with known IRP. This could provide a basis for further investigation towards drug repurposing and precision medicine. In the following, I explored the relationships between endotypes using transcriptomics and phenomics.

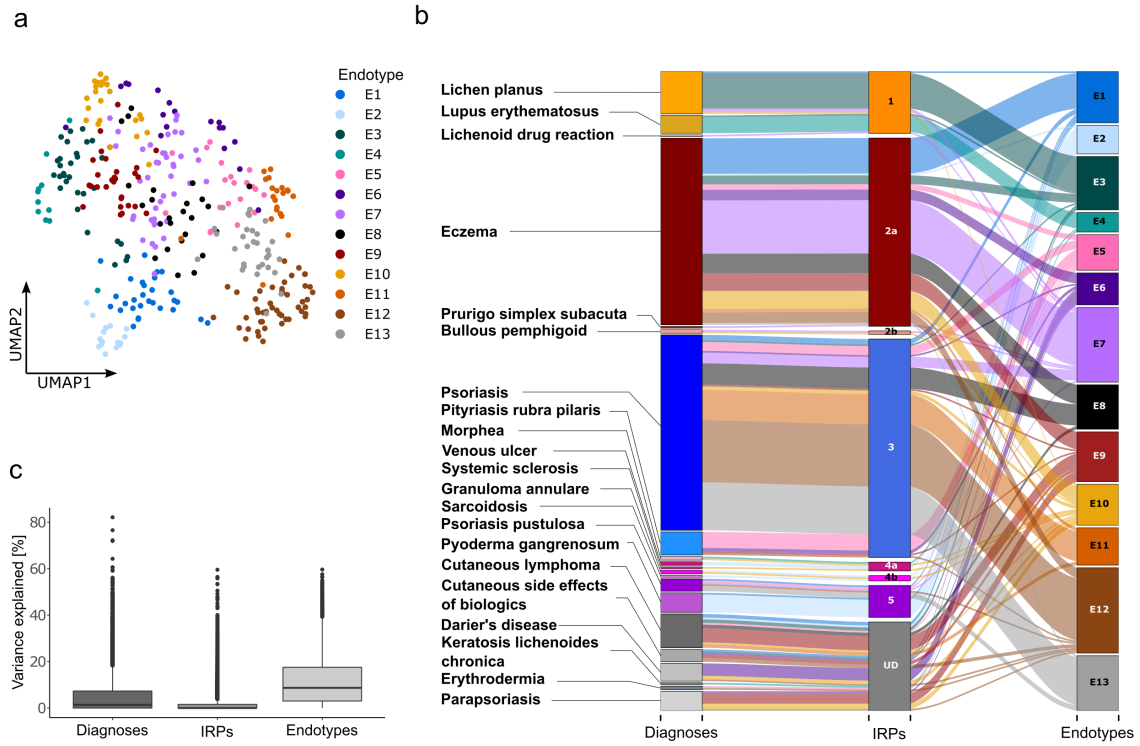


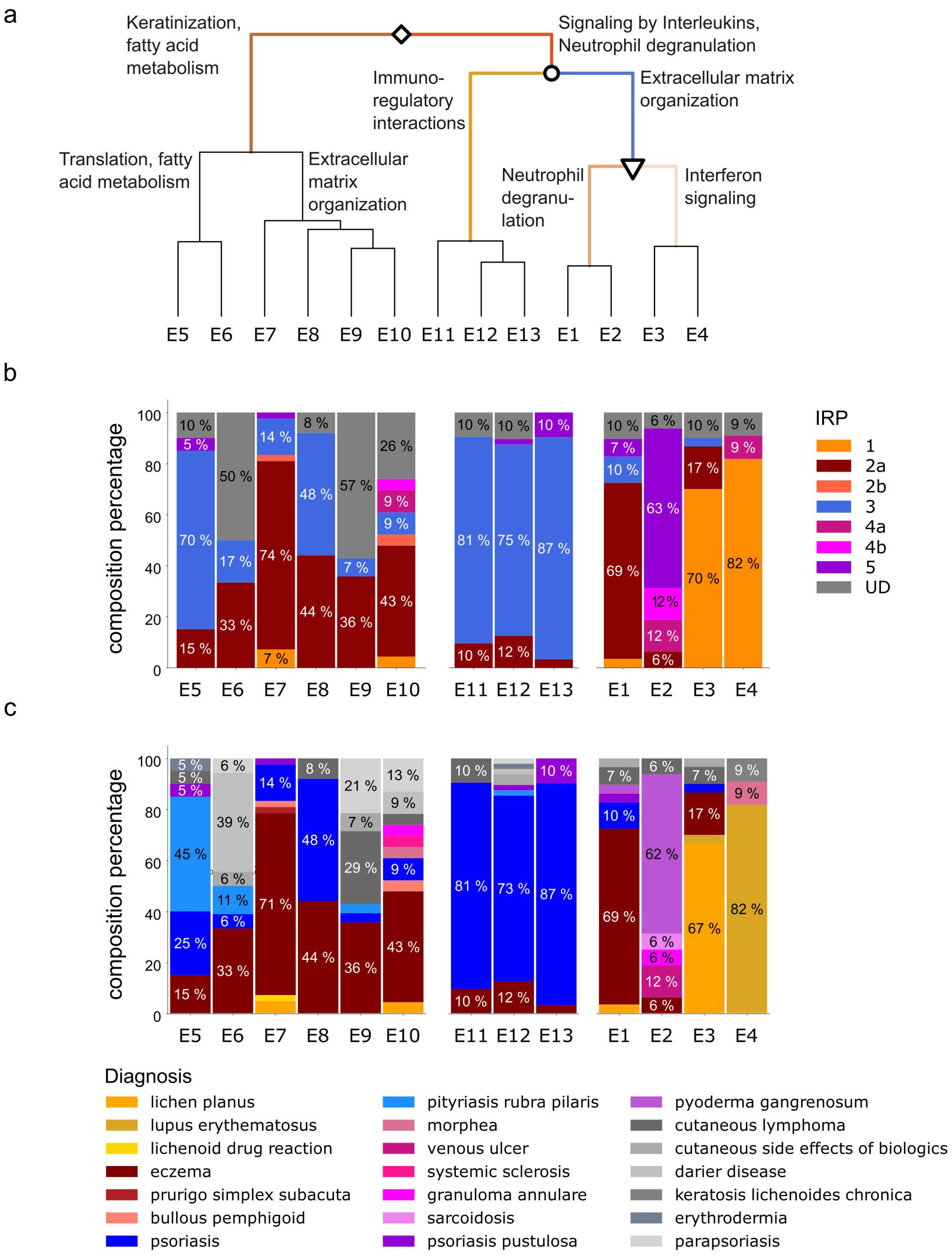
Figure 3.8: Identification of disease endotypes. a) 2D representation of identified disease endotypes. **b)** Sankey diagram shows grouping of ncISDs into IRPs and endotypes revealing the heterogeneity of the diseases that cluster into multiple endotypes. **c)** Variance partition analysis shows a tendency that endotypes explain more variance of the transcriptomics data than diagnoses and IRPs.

3.3.4 Biological interpretation of endotypes

The endotypes resulted in a comprehensive overview of ncISDs. In order to assess their biological meaningfulness, I investigated their hierarchical relationships. At a high level, the endotypes split into two primary groups, i.e. **E5 - E10** and **E11 - E13 & E1 - E4** (Figure 3.9 a). These branches were distinguished by different biological characteristics, involving processes such as metabolism, structure, and inflammation. In the following, the distinct branches were explored in more detail.

3.3.4.1 The metabolism and epidermal structure branch

The metabolism and structural organisation of the epidermis branch, **E5 - E10**, was associated with diseases linked to IRP 2a, 3 and UD (Figure 3.9 b, c). Notably, the distribution across multiple endotypes and over-representation of IRP 2a, indicates associated diseases



are likely driven by distinct biological processes. This branch was further subdivided into E5/E6 and E7 - E10 (Figure 3.10 a).

The endotypes E5/E6 in comparison to E7 - E10 were characterised by protein synthesis and "fatty acid metabolism" (Figure 3.11 j). Amongst the upregulated genes were *IL17C*,

Figure 3.9: Clustering of ncISDs into molecular endotypes provides objective granularity as compared to diseases and IRPs. a) Hierarchical relationships between endotypes based on molecular similarities and differences. The branches at the highest level (diamond) divide the structural and metabolism arm (brown) from the inflammation arm (red). The inflammation arm (circle) is further divided into two branches. The first one is the psoriasis-like endotype branch (orange) and the second branch comprises autoinflammatory and interface dermatitis diseases (blue). The autoinflammatory and interface dermatitis branch (triangle) is subdivided into an autoinflammatory (sandy brown) and interface dermatitis disease (bisque) branch. Each endotype is composed of different combinations of **(b)** IRPs and **(c)** diseases, highlighting the high level of heterogeneity within IRPs and diseases.

CCL20, and *PGLYRP2*, which are involved in the innate immune system (Figure 3.11 i). **E5** mainly contained pityriasis rubra pilaris patients, expressing the same IRP as psoriasis. In contrast, **E6** primarily consisted of Darier's disease patients and was characterised by "Mitochondrial Fatty Acid Beta-Oxidation" and metabolism (Figure 3.9 b, c; Figure 3.11 k, l). E5 and E6 were distinguished by protein synthesis and fatty acid metabolism, where E5 was associated with pityriasis rubra pilaris and E6 with Darier's disease.

The E7 - E10 branch contained activated pathways such as "Extracellular matrix organization", "Collagen formation", and "ECM proteoglycans" (Figure 3.11 i, j). Although the differences between the lower levels in the E7 - E10 branch were less pronounced, I observed hierarchical distinctions in their biological functions. Pathways enriched in E7 included barrier formation and immune response, while olfactory pathways were activated in E8 - E10 (Figure 3.11 m, n).

A higher activation of immune related pathways was observed in E8 compared to E9/E10. The endotype E8 consisted of 44 % eczema and 48 % psoriasis patients (Figure 3.9 b). It was differentiated from E9 and E10 by higher activation of cell cycle-related pathways and an up-regulation of significantly differentially expressed genes (DEGs), which are associated with immune response, such as *IL36G*, *DEFB4A*, and *PI3*, and immune regulation such as *MPRSS11D* and *VNN3*. In E9 and E10, I observed a higher activation of olfactory pathways (Figure 3.11 o, p). Thus, E8 was characterised by a greater activation of immune-related pathways, whereas E9 and E10 were dominated by the activation of olfactory pathways.

Immune regulation and skin structure pathways distinguish E9 from E10. At the lowest hierarchy level, I compared E9 and E10 (Figure 3.11 q, r). Among the upregulated DEGs

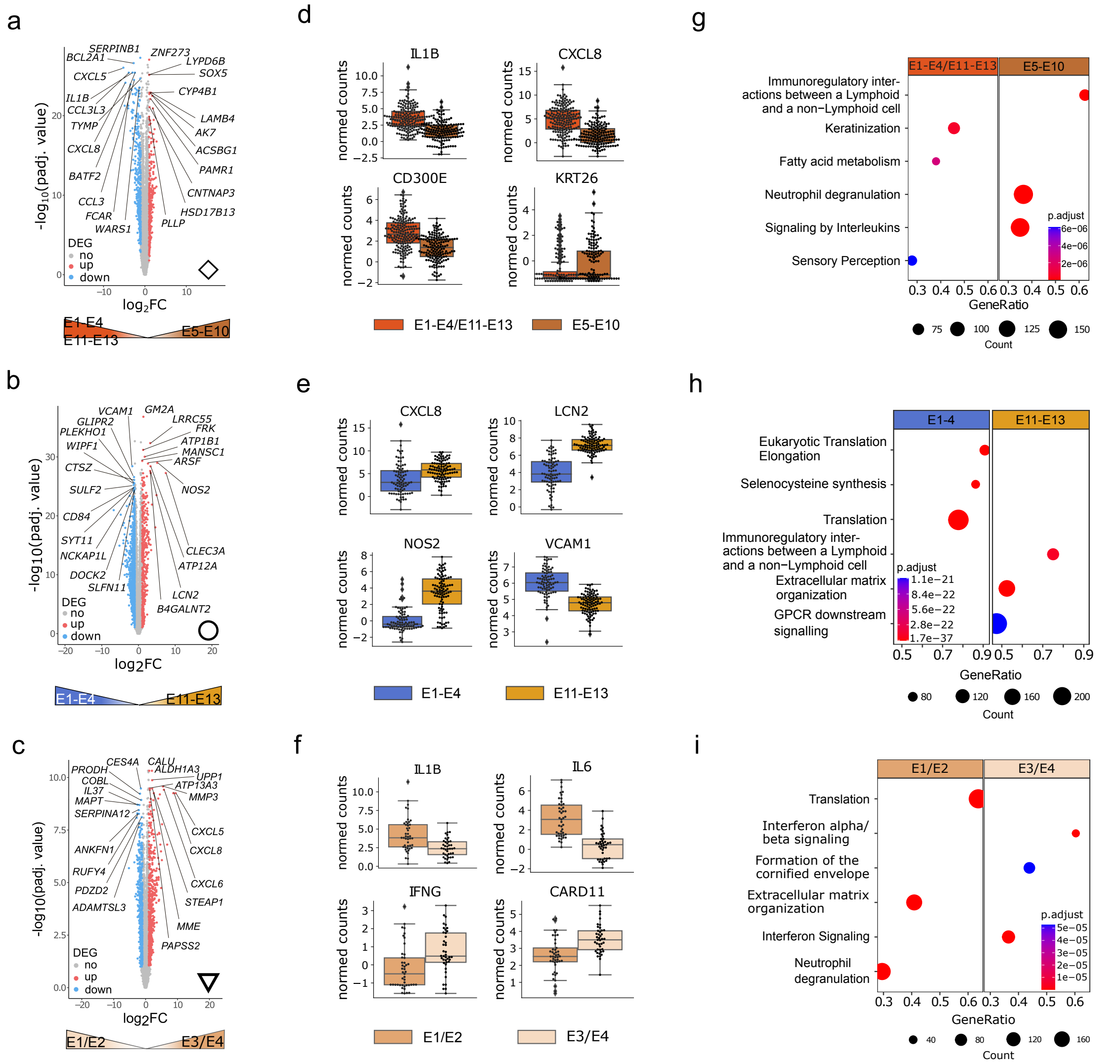


Figure 3.10: Biological interpretation of interesting endotype branches. a-c) DGE analysis results showing the label of the top 10 DEGs in each group. DEGs are required to meet the threshold of padj value < 0.05 and $|\log2FC| > 1$. **d-f)** Boxplots of interesting DEGs. **g-i)** Pathways are determined by a GSEA using all genes ranked by their log2FC value. The top three enriched pathways per group are shown.

in E9 were *IL13*, *CCL22*, and *LCE3D*. The first two were involved in immune regulations, and *LCE3D* was involved in skin structure. Moreover, I found cell cycle-related pathways activated in E9. In E10, I observed an enrichment of "Cardiac conduction", "Muscle contraction", and "Transcriptional regulation of white adipocyte differentiation" (Figure 3.11 r). In summary, E9 was associated with immune regulation and skin structure, and E10 with pathways related to muscle and adipocyte differentiation.

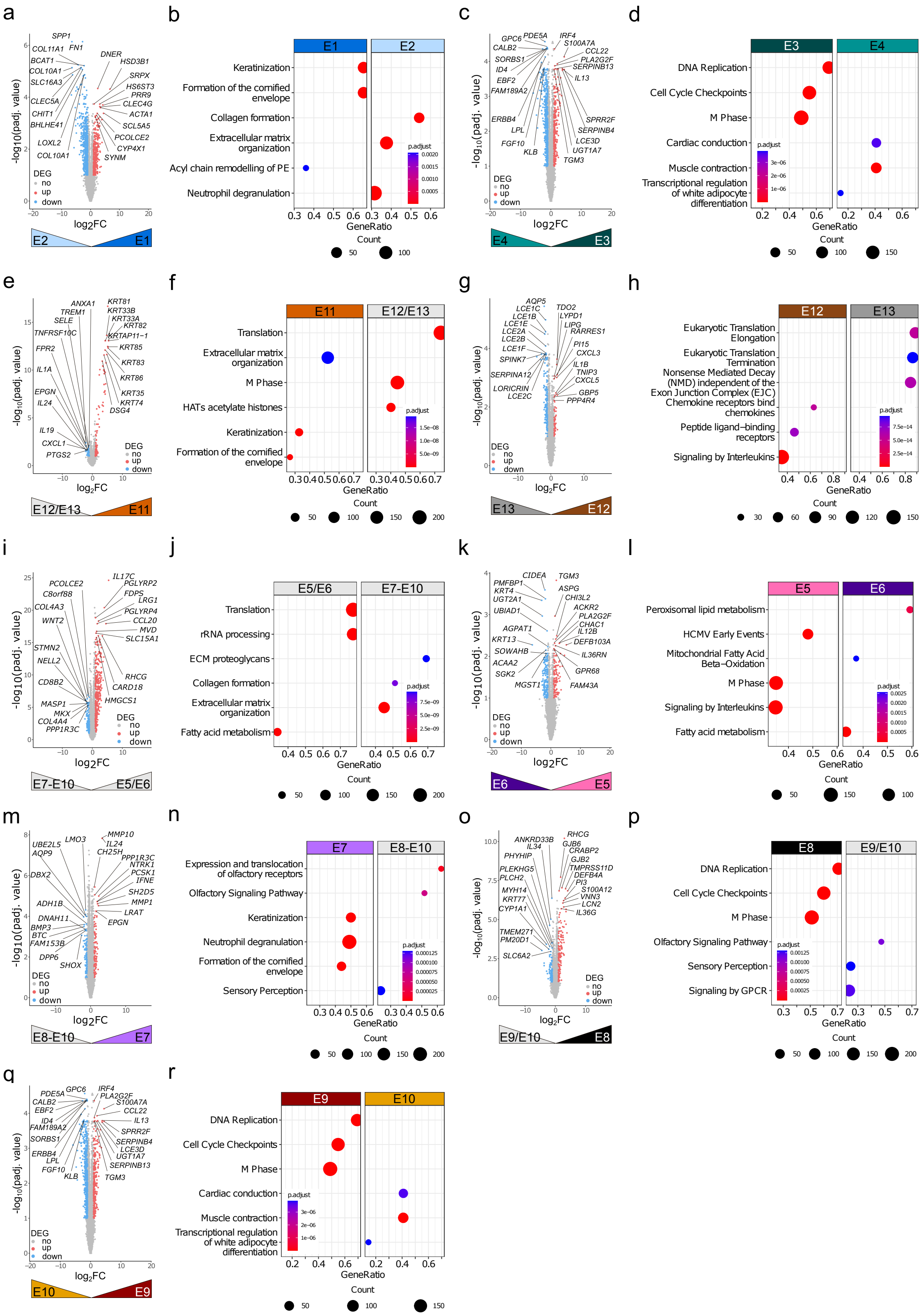

a
SPP1
COL11A1
FN1
BCAT1
COL10A1
SLC16A3
CLEC5A
CHIT1
BHLHE41
LOXL2
COL10A1
DNER
HSD3B1
SRPX
HS6ST3
PRR9
CLEC4G
ACTA1
SCL5A5
PCOLCE2
CYP4X1
SYNM
-log10(padj. value)
DEG
no
up
down
log2FC
E2
E1
b
E1
E2
Keratinization
Formation of the cornified envelope
Collagen formation
Extracellular matrix organization
Acyl chain remodelling of PE
Neutrophil degranulation
p.adjust
GeneRatio
Count
50
100
c
GPC6
PDE5A
IRF4
S100A7A
CALB2
CCL22
PLA2G2F
SORBS1
SERPINB13
ID4
IL13
EBF2
FAM189A2
SPRR2F
ERBB4
SERPINB4
LPL
LCE3D
FGF10
UGT1A7
KLB
TGM3
E4
E3
d
E3
E4
DNA Replication
Cell Cycle Checkpoints
M Phase
Cardiac conduction
Muscle contraction
Transcriptional regulation of white adipocyte differentiation
e
ANXA1
TREM1
SELE
TNFRSF10C
FPR2
IL1A
EPGN
IL24
IL19
CXCL1
PTGS2
KRT81
KRT33B
KRT33A
KRT82
KRTAP11−1
KRT85
KRT83
KRT86
KRT35
KRT74
DSG4
E12/E13
E11
f
E11
E12/E13
Translation
Extracellular matrix organization
M Phase
HATs acetylate histones
Keratinization
Formation of the cornified envelope
50
100
150
200
g
AQP5
LCE1C
LCE1B
LCE1E
LCE2A
LCE2B
LCE1F
SPINK7
SERPINA12
LORICRIN
LCE2C
TDO2
LYPD1
LIPG
RARRES1
PI15
CXCL3
IL1B
TNIP3
CXCL5
GBP5
PPP4R4
E13
E12
h
E12
E13
Eukaryotic Translation Elongation
Eukaryotic Translation Termination
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
Chemokine receptors bind chemokines
Peptide ligand−binding receptors
Signaling by Interleukins
30
60
90
120
150
i
PCOLCE2
C8orf88
COL4A3
WNT2
STMN2
NELL2
CD8B2
MASP1
MKX
COL4A4
PPP1R3C
IL17C
PGLYRP2
FDPS
LRG1
PGLYRP4
CCL20
MVD
SLC15A1
RHCG
CARD18
HMGCS1
E7-E10
E5/E6
j
E5/E6
E7-E10
Translation
rRNA processing
ECM proteoglycans
Collagen formation
Extracellular matrix organization
Fatty acid metabolism
k
CIDEA
PMFBP1
KRT4
UGT2A1
UBIAD1
AGPAT1
KRT13
SOWAHB
ACAA2
SGK2
MGST1
TGM3
ASPG
CHI3L2
ACKR2
PLA2G2F
CHAC1
IL12B
DEFB103A
IL36RN
GPR68
FAM43A
E6
E5
l
E5
E6
Peroxisomal lipid metabolism
HCMV Early Events
Mitochondrial Fatty Acid Beta−Oxidation
M Phase
Signaling by Interleukins
Fatty acid metabolism
m
UBE2L5
LMO3
AQP9
DBX2
ADH1B
DNAH11
BMP3
BTC
FAM153B
DPP6
SHOX
MMP10
IL24
CH25H
PPP1R3C
NTRK1
PCSK1
IFNE
SH2D5
MMP1
LRAT
EPGN
E8-E10
E7
n
E7
E8-E10
Expression and translocation of olfactory receptors
Olfactory Signaling Pathway
Keratinization
Neutrophil degranulation
Formation of the cornified envelope
Sensory Perception
o
ANKRD33B
IL34
PHYHIP
PLEKHG5
PLCH2
MYH14
KRT77
CYP1A1
TMEM271
PM20D1
SLC6A2
RHCG
GJB6
CRABP2
GJB2
TMPRSS11D
DEFB4A
PI3
S100A12
VNN3
LCN2
IL36G
E9/E10
E8
p
E8
E9/E10
DNA Replication
Cell Cycle Checkpoints
M Phase
Olfactory Signaling Pathway
Sensory Perception
Signaling by GPCR
q
PDE5A
GPC6
IRF4
CALB2
PLA2G2F
EBF2
S100A7A
ID4
CCL22
FAM189A2
IL13
SORBS1
SPRR2F
SERPINB4
ERBB4
LCE3D
LPL
UGT1A7
FGF10
SERPINB13
KLB
TGM3
E10
E9
r
E9
E10
DNA Replication
Cell Cycle Checkpoints
M Phase
Cardiac conduction
Muscle contraction
Transcriptional regulation of white adipocyte differentiation

Figure 3.11: Biologically distinct functionalities within hierarchical endotype relationships. a, c, e, g, i, k, m, o, q) DGE analysis results showing the label of the top 10 DEGs in each group. DEGs are required to meet the threshold of padj value < 0.05 and $|\log2FC| > 1$. **b, d, f, h, j, l, n, p, r)** Activated and suppressed pathways, which were determined by a GSEA. The top three enriched gene sets per group are shown.

In conclusion, the metabolism branch was divided into E5/E6 and E7 - E10. E5 and E6 were characterised by protein synthesis and fatty acid metabolism pathways, respectively. In contrast, E7 - E10 exhibited pathways associated with extracellular matrix (ECM) formation, cell cycle, and immune response.

3.3.4.2 The inflammatory branch

The E11 - E13 & E1 - E4 branch was predominantly driven by inflammation and included diseases associated with IRP 1, 3, and 5 (Figure 3.9 b, c). This branch was characterised by the activation of cell cycle and interleukin signalling pathways (Figure 3.12 n). Amongst the genes associated with the inflammatory branch, were pro-inflammatory cytokines (*IL1B*), chemokines (*CXCL8*), innate immune system (*CD300E*), and epidermal structure (*KRT26*) (Figure 3.10 a, d, g). The inflammatory branch was further subdivided into psoriasis-like endotypes E11 - E13 and interface dermatitis and autoinflammatory like endotypes E1 - E4 (Figure 3.10 a).

In the psoriasis-like endotypes **E11 - E13**, upregulated genes included Th17 markers such as *CXCL8* and *NOS2* (Figure 3.10 b, e). Another biomarker of inflammation was *LCN2*, which has recently been implicated in the pathogenesis of psoriasis [RX22]. E11 - E13 were also characterised by the activation of protein synthesis and cell cycle pathways (Figure 3.10 h; Figure 3.12 n). In general, the E11 - E13 endotypes were characterised by inflammation-related genes as well as protein synthesis and cell cycle regulation pathways.

The endotypes **E1 - E4** resembled the interface dermatitis and autoinflammatory endotype branch. This branch was defined by pathways such as "Immunoregulatory interactions between Lymphoid and non-lymphoid cells", "Extracellular matrix organisation" and "GPCR downstream signalling" (Figure 3.10 h). The top DEGs in E1 - E4 included *VCAM1*, *GLIPR2*, and *PLEKHO1* (Figure 3.10 b, e). The gene *VCAM1* is a marker for cardiovascular diseases [Hos+04], *GLIPR2* is a potential pan-cancer biomarker [Lin+24] [Zen+23], and *PLEKHO1* is likely associated with the viability of renal cell carcinoma cells [Yu+19]. In essence, the E1 - E4 branch was characterised by cell interaction and ECM organisation pathways, with DEGs linked to various diseases.

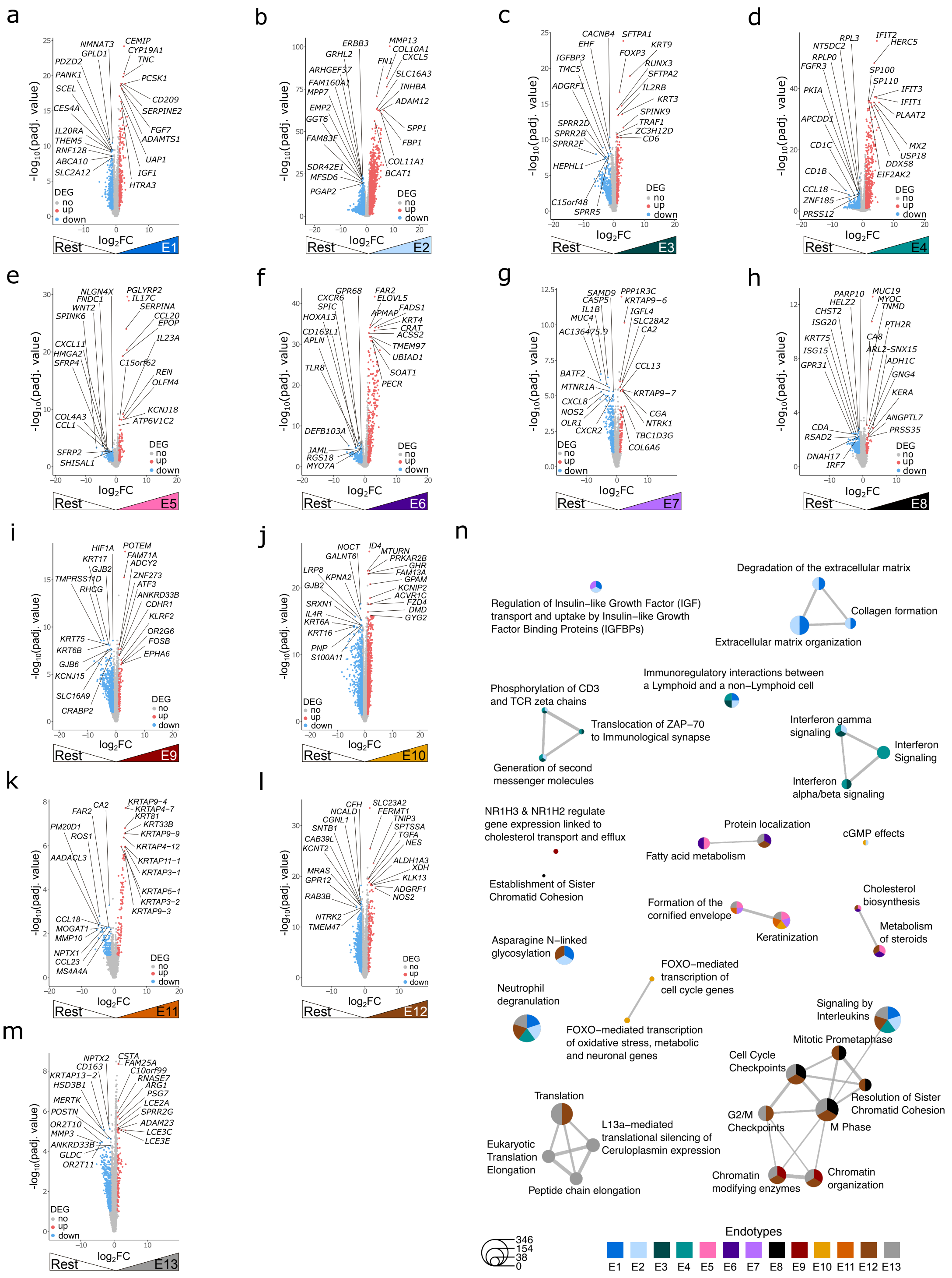

a
NMNAT3
GPLD1
PDZD2
PANK1
SCEL
CES4A
IL20RA
THEM5
RNF128
ABCA10
SLC2A12
CEMIP
CYP19A1
TNC
PCSK1
CD209
SERPINE2
FGF7
ADAMTS1
UAP1
IGF1
HTRA3
DEG
no
up
down
-log10(padj. value)
log2FC
Rest
E1
b
ERBB3
GRHL2
ARHGEF37
FAM160A1
MPP7
EMP2
GGT6
FAM83F
SDR42E1
MFSD6
PGAP2
MMP13
COL10A1
FN1
CXCL5
SLC16A3
INHBA
ADAM12
SPP1
FBP1
COL11A1
BCAT1
E2
c
CACNB4
EHF
IGFBP3
TMC5
ADGRF1
SPRR2D
SPRR2B
SPRR2F
HEPHL1
C15orf48
SPRR5
SFTPA1
KRT9
FOXP3
RUNX3
SFTPA2
IL2RB
KRT3
SPINK9
TRAF1
ZC3H12D
CD6
E3
d
RPL3
NT5DC2
RPLP0
FGFR3
PKIA
APCDD1
CD1C
CD1B
CCL18
ZNF185
PRSS12
IFIT2
HERC5
SP100
SP110
IFIT3
IFIT1
PLAAT2
MX2
USP18
DDX58
EIF2AK2
E4
e
NLGN4X
FNDC1
WNT2
SPINK6
CXCL11
HMGA2
SFRP4
COL4A3
CCL1
SFRP2
SHISAL1
PGLYRP2
IL17C
SERPINA
CCL20
EPOP
IL23A
C15orf62
REN
OLFM4
KCNJ18
ATP6V1C2
E5
f
GPR68
CXCR6
SPIC
HOXA13
CD163L1
APLN
TLR8
DEFB103A
JAML
RGS18
MYO7A
FAR2
ELOVL5
FADS1
APMAP
KRT4
CRAT
ACSS2
TMEM97
UBIAD1
SOAT1
PECR
E6
g
SAMD9
CASP5
IL1B
MUC4
AC136475.9
BATF2
MTNR1A
CXCL8
NOS2
OLR1
CXCR2
PPP1R3C
KRTAP9−6
IGFL4
SLC28A2
CA2
CCL13
KRTAP9−7
CGA
NTRK1
TBC1D3G
COL6A6
E7
h
PARP10
HELZ2
CHST2
ISG20
KRT75
ISG15
GPR31
CDA
RSAD2
DNAH17
IRF7
MUC19
MYOC
TNMD
PTH2R
CA8
ARL2−SNX15
ADH1C
GNG4
KERA
ANGPTL7
PRSS35
E8
i
HIF1A
KRT17
GJB2
TMPRSS11D
RHCG
KRT75
KRT6B
GJB6
KCNJ15
SLC16A9
CRABP2
POTEM
FAM71A
ADCY2
ZNF273
ATF3
ANKRD33B
CDHR1
KLRF2
OR2G6
FOSB
EPHA6
E9
j
NOCT
GALNT6
LRP8
KPNA2
GJB2
SRXN1
IL4R
KRT6A
KRT16
PNP
S100A11
ID4
MTURN
PRKAR2B
GHR
FAM13A
GPAM
KCNIP2
ACVR1C
FZD4
DMD
GYG2
E10
k
FAR2
CA2
PM20D1
ROS1
AADACL3
CCL18
MOGAT1
MMP10
NPTX1
CCL23
MS4A4A
KRTAP9−4
KRTAP4−7
KRT81
KRT33B
KRTAP9−9
KRTAP4−12
KRTAP11−1
KRTAP3−1
KRTAP5−1
KRTAP3−2
KRTAP9−3
E11
l
CFH
NCALD
CGNL1
SNTB1
CAB39L
KCNT2
MRAS
GPR12
RAB3B
NTRK2
TMEM47
SLC23A2
FERMT1
TNIP3
SPTSSA
TGFA
NES
ALDH1A3
XDH
KLK13
ADGRF1
NOS2
E12
m
NPTX2
CD163
KRTAP13−2
HSD3B1
MERTK
POSTN
OR2T10
MMP3
ANKRD33B
GLDC
OR2T11
CSTA
FAM25A
C10orf99
RNASE7
ARG1
PSG7
LCE2A
SPRR2G
ADAM23
LCE3C
LCE3E
E13
n
Regulation of Insulin−like Growth Factor (IGF) transport and uptake by Insulin−like Growth Factor Binding Proteins (IGFBPs)
Degradation of the extracellular matrix
Collagen formation
Extracellular matrix organization
Immunoregulatory interactions between a Lymphoid and a non−Lymphoid cell
Phosphorylation of CD3 and TCR zeta chains
Translocation of ZAP−70 to Immunological synapse
Generation of second messenger molecules
Interferon gamma signaling
Interferon Signaling
Interferon alpha/beta signaling
NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux
Protein localization
Fatty acid metabolism
cGMP effects
Establishment of Sister Chromatid Cohesion
Formation of the cornified envelope
Keratinization
Cholesterol biosynthesis
Metabolism of steroids
Asparagine N−linked glycosylation
FOXO−mediated transcription of cell cycle genes
FOXO−mediated transcription of oxidative stress, metabolic and neuronal genes
Neutrophil degranulation
Signaling by Interleukins
Mitotic Prometaphase
Cell Cycle Checkpoints
Resolution of Sister Chromatid Cohesion
G2/M Checkpoints
M Phase
Translation
L13a−mediated translational silencing of Ceruloplasmin expression
Eukaryotic Translation Elongation
Peptide chain elongation
Chromatin modifying enzymes
Chromatin organization
346
154
38
0
Endotypes
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13

Figure 3.12: Functional differences between one endotype and all others. a-m) DGE analysis results showing the label of the top 10 DEGs in each group. DEGs are required to meet the threshold of padj value < 0.05 and $|\log2FC| > 1$. **n)** Clustered pathways showing similarities between endotypes on a functional level. Pathways are determined by an ORA considering significantly upregulated genes from the pairwise comparisons of one endotype versus all other endotypes in combination.

Summarising, the inflammatory branch was divided into two distinct subgroups. The psoriasis-like endotypes E11 - E13 were characterised by inflammation-related pathways and genes, while the endotypes E1 - E4 were associated with biological events such as cell interaction and ECM organisation.

3.3.4.3 Dermatitis, autoinflammatory/granulomatous and interface dermatitis subdivisions

The E1 - E4 branch was further divided into E1/E2, a dermatitis and autoinflammatory/ granulomatous branch, and E3/E4, representing the interface dermatitis diseases lichen planus and lupus erythematosus (Figure 3.9 a).

I compared the dermatitis and autoinflammatory/granulomatous branch, involving **E1/E2**, to E3/E4. E1 was primarily composed of eczema patients and shared biological characteristics with pyoderma gangrenosum and venous ulcer. These were the autoinflammatory and granulomatous diseases representing E2 (Figure 3.9 c). This overlap was supported by the expression of immune response regulating cytokines, such as *IL1B* and *IL6*, metalloproteins, and other genes involved in the innate immune system (Figure 3.10 c, f). In addition, regulation of protein synthesis by "translation", "neutrophil degranulation" and "extracellular matrix organization" pathways were activated (Figure 3.10 i). The E1/E2 branch revealed shared biological mechanisms of eczema and rare diseases, such as innate immune system pathways and protein synthesis regulation.

The interface dermatitis branch **E3/E4** branch was characterised by the activation of "interferon signalling" and "formation of the cornified envelope" pathways (Figure 3.10 i). This finding aligned with the definition of the IRP 1, which involves the dominant diseases within these endotypes, namely lichen planus and lupus erythematosus. In this IRP, immune cells induce apoptosis of cells located in the basal epidermis [BCS22] [Lau+18]. Although the "formation of the cornified envelope" did not appear to play an active role in inducing cell death, it may indirectly contribute to apoptosis [Eck+13]. In line with this finding, *IFN-γ* and *CARD11* were upregulated in E3/E4, which is part of the "interferon signalling" pathway and involved in cell death, respectively (Figure 3.10 c, f). The E3/E4 branch showed an enrichment of interferon signalling and cell death associated pathways, supporting its association with interface dermatitis diseases.

In conclusion, the E1 - E4 branch was subdivided into an E1/E2 and E3/E4 clusters. The former included eczema and autoinflammatory/granulomatous diseases, driven by innate immune system pathways. In contrast, the E3/E4 branch, associated with interface dermatitis diseases, was characterised by interferon signalling and pathways linked to cell death. Moving forward, I was interested in the biological drivers (Section 3.3.4.4) of each endotype compared to all others.

3.3.4.4 Distinct and shared biological processes in endotypes

Comparative analyses of each endotype against all other endotypes revealed distinct gene expression profiles (Figure 3.12 a-m) alongside shared biological pathways (Figure 3.12 n), demonstrating the heterogeneity and overlapping features of endotypes.

Rare diseases with an UD IRP are associated with distinct metabolic and keratinisation-related pathways. One of my objectives is to drive the understanding of rare diseases with an UD IRP. I observed three endotypes that were mainly composed of diseases with an UD IRP, i.e. E6, E9 and E10 (Figure 3.9 b, c). In E6, DEGs, such as *ELOVL5* and *KRT4*, were found upregulated, which were involved in metabolism and keratinisation (Figure 3.12 f). In E9, olfactory signalling pathways were enriched with associated genes such as *OR2G6* Figure 3.12 i). In addition, I found "NR1H3 & NR1H2 regulate gene expression linked to cholesterol transport and efflux" uniquely activated in E9 (Figure 3.12 n). Notably, E10 exhibited a higher activation of both, keratinisation and olfactory signalling pathways, and an activation of FOXO transcription factors, which were involved in "FOXO-mediated transcription of cell cycle genes", "cGMP effects", and "FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes" (Figure 3.12 j, n). In summary, metabolic and keratinisation-related pathways distinguish E6, E9, and E10 from other endotypes.

Looking at the entire metabolism branch, I found that E5, E7, E9 and E10 shared skin barrier related pathways such as keratinisation (Figure 3.12 n). Moreover, endotypes E5 and E6 showed an higher activation of metabolism pathways (Figure 3.12 n). Intriguingly, in E5, whose dominant disease was pityriasis rubra pilaris, the upregulated DEGs corresponded to genes, which are assumed to be associated with this disease, such as *IL23A*, *CCL20*, and *IL17C* [Hay+23] (Figure 3.9 c; Figure 3.12 e). These results suggest shared skin barrier and metabolic pathways across the endotypes E5, E7, E9 and E10, with E5 showing also an upregulation of inflammatory markers of pityriasis rubra pilaris.

I explored the characteristic genes upregulated in the psoriasis-like endotypes E8, E11 - E13 and their activated pathways. In E11 - E13, I observed DEGs, which were involved in the formation of the cornified envelope and keratinization. Examples for E11 are *KRT81*, *KRTAP9-4*, *KRT33B*, for E12 *KLK13*, and for E13 *CSTA*, *LCE3E*, and *LCE2A* (Figure 3.12 k-m). Interestingly, endotype E11 did not show significant activation of cell cycle and interleukin signalling pathways, but shared characteristics with E5, E7, and E13, including epidermal structure-related pathways, such as "Keratinization" and "Formation of cornified envelope". E13 also showed a higher activation of translational pathways (Figure 3.12 n). Despite a 48 % of psoriasis patients, E8 exhibited distinct gene regulations, such as *PTH2R*, which is associated with G alpha (s) signalling events (Figure 3.12 h). The psoriasis-like endotypes E8, E12-E13 demonstrated both distinct an shared features, explaining the subdivision of psoriasis into multiple endotypes.

The interface and autoinflammatory like endotypes E1 - E4 were characterised by immune response and ECM associated pathways (Figure 3.12 a-d, n). The immune response pathways were primarily enriched in the interface like endotypes E3 and E4, which were mainly composed of lichen planus and lupus erythematosus, respectively. The ECM associated pathways were primarily observed in E1 and E2. Examples of genes involved in ECM were *TNC* for E1 and *MMP13*, *COL10A1*, and *SPP1* for E2. Immune response pathways dominated the interface-like endotypes E3/E4, while ECM pathways were more prominent in autoinflammatory-like endotypes E1/E2.

Summarising Section 3.3.4, I discovered a hierarchical relationship among the, hypothesis-free derived endotypes, clustering into two primary branches driven by metabolic and inflammatory pathways. The metabolism-driven branch was predominantly composed of eczema, pityriasis rubra pilaris, and diseases with UD IRPs, whereas the inflammation-driven branch was primarily composed of psoriasis and diseases associated with IRP 2a, IRP 5, and IRP 1. These findings further illustrated that despite heterogeneous transcriptomics profiles, these endotypes share common biological characteristics. Additionally, I revealed pathways and genes that were differentially expressed and activated within and independent of the hierarchical structure. In essence, this molecular characterisation of endotypes enhances the understanding of skin diseases and paves the way for developing more precise diagnostic and therapeutic approaches for ncISDs. Next, I was interested whether endotypes are associated with distinct clinical attributes.

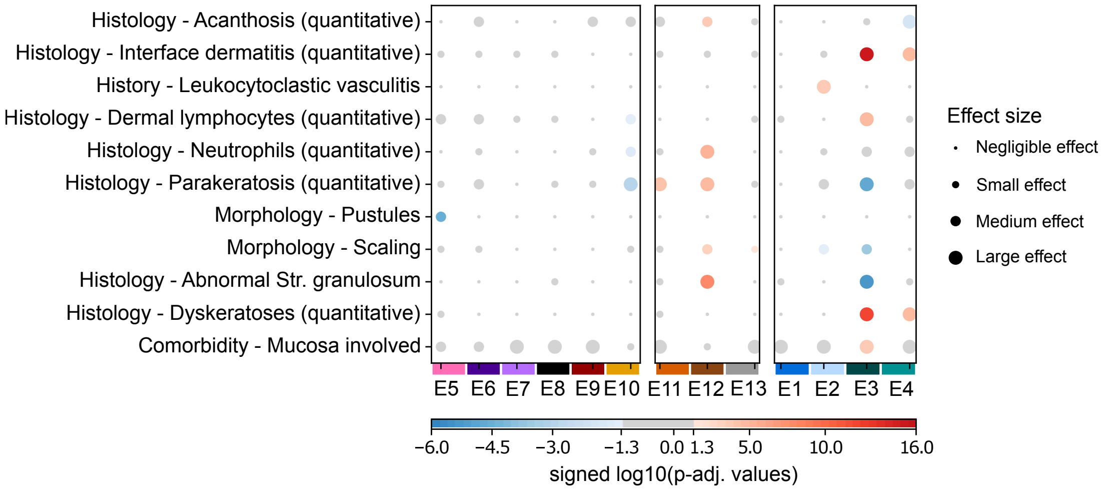


Figure 3.13: Clinical interpretation of endotypes. Dotplot showing the signed significance of selected clinical attributes in the endotypes. They are chosen based on the criteria of having a highly significant p-value below $1e\text{-}05$ in at least one endotype. Moreover, the column split shows the top three branches, as defined by the hierarchical relationship between the endotypes. The colour indicates whether the attribute is increased (red) or decreased (blue) in the respective endotype. Kruskal-Wallis test is conducted to test for enrichment in a one versus all fashion for continuous and ordinal data and the χ^2 test is applied on nominal traits. The results are presented in red or blue if they meet the threshold of a padj value of less than 0.05, otherwise they are shown in grey.

3.3.5 Endotypes correlate with clinical hallmarks

Clinical diagnosis typically relies on specific clinical hallmarks. In this analysis, I aimed to explore the correlation between endotypes and clinical attributes. Considering the imputed assessments (Methods 3.2.2), I observed some significant associations with the endotypes (Figure 3.13). In particular, I considered the clinical attributes that had a p-value below $1e\text{-}05$ in at least one endotype (Methods 3.2.2), providing a robust foundation to correlate the molecular endotypes to clinical assessments.

In the metabolism-driven endotype branch, distinct clinical markers assisted in differentiating the endotypes E5 and E10 from others. Four out of 66 attributes were significantly pronounced, with each endotype showing a reduced prominence of these clinical markers. Specifically, endotype E5 was significantly associated with "Morphology - Pustules" ($padj = 3.36e\text{-}05$, Cliff's Delta = -0.37), with a negative effect size indicating a reduction in the number of pustules in E5 in comparison to the mean number observed in the other endotypes. In E6-E9, no differentially pronounced attributes were observed, while

E10 represented a contrasting clinical phenotype to psoriasis. This is evidenced by the moderately less pronounced attribute "Histology - Neutrophils" ($padj$ = 2.41e-02, Cohen's d = -0.69) and highly less expressed clinical feature "Histology - Parakeratosis" ($padj$ = 1.22e-03, Cohen's d = -0.87) [EE18] [Orl+20]. In addition, I observed a moderately decreased number of "Histology - Dermal lymphocytes" ($padj$ = 5.06e-02, Cohen's d = -0.61). The clinical features associated with E5 and E10 suggest distinct markers, with reduced pustule formation and altered histological characteristics, setting these endotypes apart from others.

Endotypes within the psoriasis-dominated branch showed distinct clinical features associated with psoriasis. Specifically, E12 exhibited a significant enrichment of traits such as "Histology - Abnormal Str. granulosum" ($padj$ = 1.99e-08, Cliff's Delta = 0.56), "Histology - Parakeratosis" ($padj$ = 1.54e-05, Cohen's d = 0.89), "Histology - Acanthosis" ($padj$ = 1.66e-04, Cohen's d = 0.70), "Histology - Neutrophils" ($padj$ = 4.09e-06, Cohen's d = 1.00), and "Morphology - Scaling" ($padj$ = 1.12e-03, Cliff's Delta = 0.35). The presence of a more pronounced Histology - Abnormal Str. granulosum" indicates that this layer is either absent or thinner than normal skin. I defined the stratum granulosum as component of the middle epidermis in Section 1.1.1. E12's branch neighbours, E11 and E13, exhibited higher expression of at least one of these attributes, i.e. "Histology - Parakeratosis" (E11: $padj$ = 7.97e-05, Cohen's d = 1.15) and "Morphology - Scaling" (E13: $padj$ = 3.54e-02, Cliff's Delta = 0.32). The feature "Morphology - Scaling" was also more pronounced in E12 and is characteristic of psoriasis [Gri+21]. In summary, these observations support that E12 most closely resembles psoriasis, while E11 and E13 show partial overlap with psoriasis-related traits, suggesting a continuum of psoriasis-like features across these endotypes.

In the autoinflammatory and granulomatous endotype branch, specific clinical features helped differentiate E2 from other endotypes. In E1, I did not observe any distinct expressed clinical attribute, while in E2, "History - Leukocytoclastic vasculitis" ($padj$ = 2.41e-04, Cohen's d = 1.19) was stronger pronounced and "Morphology -Scaling" ($padj$ = 3.46e-02, Cliff's Delta = -0.43) was moderately more present in other endotypes. These findings highlight leukocytoclastic vasculitis as a distinguishing feature of the granulomatous endotype E2, providing a potential clinical marker for identifying this specific disease group.

The interface dermatitis branch, including E3 and E4, was defined by certain clinical traits. As expected in the interface dermatitis branch E3/E4, the clinical trait "Histology - Interface dermatitis" was most enriched in E3 ($padj = 4.45e$ - 16, Cohen's d = 2.29) and E4 ($padj = 2.13e$ - 05, Cohen's d = 2.22), which I previously defined as the interface dermatitis branch based on biological evidences (Figure 3.9 a-c). In line with this, the cell death associated clinical trait "Histology - Dyskeratosis" (E3: $padj = 5.13e$ - 13, Cliff's Delta = -0.60; E4: $padj = 2.34$ - 05, Cliff's Delta = 0.62) was also more dominant. The most prominent diseases within the endotypes E3 and E4 are commonly based on the presence of the cell death associated pathways [EE18]. In addition, I observed an absent of "Histology - Parakeratosis" in E3 ($padj$ = 2.34e - 05, Cohen's d = -0.95). The attributes "Histology - Acanthosis" (E4: $padj = 9.56e$ - 03, Cohen's d = -1.31), and "Histology - Parakeratosis" (E3: $padj = 2.34e$ - 05, Cohen's d = -0.95) were more expressed in other endotypes. In addition, E3, which represents lupus erythematosus (Figure 3.9 c), showed a significant decrease in "Morphology - Scaling" ($padj = 2.59e$ - 04, Cliff's Delta = -0.47) and "Histology - Abnormal Str. granulosum" ($padj = 4.09e$ - 06, Cliff's Delta = -0.58) in combination with an increase in "Comorbidity - Mucosa involved" ($padj$ = 1.66e - 04, odds ratio = 12.75) and "Histology - Dermal lymphocytes" (E3: $padj = 1.96e$ - 05, Cohen's d = 1.07). The feature "Histology - Abnormal Str. granulosum" is a typical histological feature of IRP 1 associated diseases, where the middle epidermis (stratum granulosum) is thicker compared to the normal state [Uji+22]. In summary, the observed features of E3 and E4, particularly their enrichment in interface dermatitis-related traits and cell death-associated pathways, align with literature about interface dermatitis diseases.

The result of the association of endotypes with clinical traits yielded specific findings for the endotype branches. In the metabolism branch E5 - E10, certain features were less prominent in E5 and E10 compared to all other endotypes. In the psoriasis-dominated endotype branch E11 - E13, clinical traits commonly associated with psoriasis were mainly enriched in E12, while I observed less enriched psoriasis-specific clinical traits in E11 and E13. Moreover, in the autoinflammatory and interface dermatitis branch E1 - E4, E3 and E4 exhibited characteristic phenotypes of interface dermatitis-associated diseases. These findings suggest, that endotypes exhibit a diverse relationships with clinical traits, with certain endotypes within a branch showing more distinct or stronger associations with the expected clinical features of the dominant disease in that endotype.

In summary, the analyses conducted in the Section 3.3.4 and Section 3.3.5 demonstrated the biological and clinical significance of the endotypes, highlighting that the hierarchical relationships between endotypes are driven by distinct pathways. In particular, I characterised endotypes composed of more common and rare diseases as well as identified shared and distinct biological features across endotypes. Furthermore, I identified an association matrix between clinical attributes and endotypes, further emphasising their clinical relevance. These findings provide evidence for the biological and clinical relevance of the endotypes in ncISDs.

3.4 Summary and discussion

In this chapter, the objective was to achieve an hypothesis-free categorisation of ncISDs. The established practice in clinics bases the diagnoses of patients solely on clinical characteristics, leading to misdiagnosis and non-responsiveness to therapies. To address this issue, I leveraged bulk RNA-seq comprised of 21 ncISDs with additional clinical information. By applying a clustering algorithm on the transcriptomics data, I identified 13 disease endotypes (objective (i)). I approached this task by developing a pipeline to determine HVGs and presented an alternative HVG method called HRGs. I demonstrated its competitive behaviour against state-of-the-art methods, such as filtering by standard deviation or the mean-variance relation. Additionally, I investigated the relationship between endotypes and the current disease ontology (objective (ii)). I explored also the biological and phenotypic similarities and differences between endotypes by comparing enriched genes, pathways, and clinical traits (objective (iii)). In summary, this chapter aimed to improve the patient stratification in ncISDs by employing a hypothesis-free approach that leverages transcriptomics data to identify endotypes.

My analysis showed that both clinical and transcriptomics profiles exhibit a high heterogeneity between and within diseases. Leveraging my HVG selection pipeline and subsequent clustering of transcriptomics profiles from 381 patients, I identified 13 endotypes. Comparing these endotypes with the current diagnoses and IRPs, I observed an almost complete regrouping of patients at the gene expression level. In addition, the endotypes have a higher level of explained variance, providing a more accurate representation of patient groups with similar profiles. In addition, I identified hierarchical relationships between endotypes, resulting in two main branches driven by metabolism and inflammatory pathways. The comparison of endotypes with clinical phenotypes revealed specific associations. These findings demonstrate that transcriptomics enhances patient stratification and show that endotypes provide a more precise categorisation of ncISDs, advancing precision medicine in this field.

In this study, I observed heterogeneity among patient profiles both within and across ncISDs. Clusters of IRP 1 showed distinct phenotypes and gene expression profiles, while other diseases appeared more dispersed across clinical attributes and transcriptomics embeddings, reflecting a broader spectrum of variability. Considering the clinical attributes embedding, more dominant disease clusters were anticipated due to the reliance of ncISDs diagnoses on these traits. This finding highlights the potential limitation of categorising patients solely based on clinical attributes. At the transcriptomics level, the distributed nature of disease clusters aligns with my hypothesis that the current disease ontology is not sufficient to determine the underlying disease. The observed heterogeneity in clinical profiles may partly reflect the subjective assessment of symptoms by patients and clinicians. Additionally, diseases with complex presentations, such as eczema, further complicate accurate diagnosis due to their varied clinical and biological characteristics [Uji+22]. These findings suggest to reconsider patient categorisation into ncISDs using solely clinical assessments and to integrate gene expression data alongside clinical attributes to enhance patient stratification.

I developed a HVG selection pipeline that integrates the HRG method, enabling automated dimensionality reduction. The pipeline reduced the dimension of the dataset from $17,816$ to $1,248$ genes, thereby improving clustering efficiency. Manual determination of the number of genes or cut-off parameters is a common challenge in feature selection methods, which can result in the inclusion of redundant features, compromising clustering performance. By automating this process, the pipeline mitigates such risks, providing more specific features. In addition, the HVGs selection pipeline ensures biologically meaningful and robust clustering results by providing accuracy and DBI as complementary metrics. While the accuracy reflects the alignment with ground truth labels, it can be misleading when patients with different diagnoses exhibit similar gene expression patterns. Focusing solely on the DBI, which evaluates intra-cluster compactness and inter-cluster separation [DB79], may lack biological relevance due to the dataset's complexity and heterogeneity. In combination, these two metrics enabled the identification of the $1,248$ most relevant genes and clustering resolution $\gamma = 0.9$ for grouping patients into endotypes. In summary, the pipeline automates HVG selection, ensuring robust clustering results, thereby offering advancements in feature selection for data-driven patient stratification.

In order to determine HVGs, I developed the method HRG to prioritise genes that convey unique and biologically relevant information. The HRG method provides a new approach to gene selection by ranking genes based on their expression levels across all samples. This enables the clustering of genes with similar expression profiles. The HRG method demonstrated robust feature selection across multiple initialisation values. Its performance was comparable to SD, while outperforming the mean-variance relationship approach implemented in SCANPY. Additionally, the method required less HVGs to achieve optimal clustering performance than both SD and SCANPY. Beyond its applications to bulk RNA-seq data, the HRG is also applicable to other diseases and could assist in automating the selection of HVGs. Although it has not yet been tested on single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics (ST) data, the method has been designed with these data types in mind, extending its potential scope. In essence, the HRG method, as part of the automated HVG selection pipeline, enhances clustering efficiency and robustness for diverse datasets.

The endotypes identified through transcriptomics data revealed the molecular heterogeneity within ncISDs. Using 1, 248 HVGs, I stratified 21 ncISDs into 13 endotypes and analysed their composition with respect to diagnoses and IRPs. Notably, each endotype was composed of multiple diseases, reinforcing that diagnostics based solely on clinical attributes are not sufficient. For instance, several endotype clusters were dominated by eczema, confirming eczema's heterogeneity as previously reported [Tso+19] [Thi+17]. Similarly, I identified four distinct psoriasis-like endotypes, which likely expand upon the two psoriasis subtypes identified by Ainali et al. (2012). This expansion may be attributed to the larger sample size in my study (125 psoriasis patients vs. 37 in Ainali et al. (2012)), which provides greater resolution for detecting disease heterogeneity [Ain+12]. These findings highlight the potential of endotype-based patient stratification to uncover shared biological mechanisms across both common and rare diseases, thereby advancing precision medicine in ncISDs and leading to novel therapeutic opportunities, including drug repurposing.

Rare diseases, such as cutaneous lymphoma and parapsoriasis, were clustered with more prevalent diseases, whose pathogenesis is better understood. This offered insights into shared biological processes, facilitating a better understanding of rare diseases. Drug repurposing could be applied to these endotypes, offering a potential therapeutic avenue for rare diseases.

The endotypes revealed discrepancies between clinical and transcriptomics profiles. In comparison to previous research, which focused on only two or three diseases [Gar+16] [Rei+19b] [Tia+12] [Ewa+15] [Gho+15] [Kam+10] [Tso+19], this study provides a more comprehensive perspective on shared and distinct molecular mechanisms. Notably, the identification of endotypes in rare diseases allows for the discovery of molecular similarities with more common, less heterogeneous ncISDs. For instance, cutaneous lymphoma and parapsoriasis clustered separately from one another but together with some patients diagnosed with eczema, reflecting distinct molecular signatures despite phenotypic similarities. Conversely, pityriasis rubra pilaris and psoriasis formed distinct endotypes, despite overlapping phenotypes and belonging to the same IRP [EE18]. These findings suggest that diseases with similar clinical presentations can differ significantly at the molecular level. Consequently, patient stratification by endotype rather than clinical diagnosis could enhance treatment outcomes.

In the metabolism branch, I found hierarchical structures driven by distinct signalling pathways. I also examined whether any specific clinical features were uniquely expressed in one endotype compared to the others. However, I did not observe any overlap with the expected clinical characteristics of the dominant disease within each endotype. This suggests a need for more specific clinical features to describe these endotypes. In addition, I observed that eczema was distributed across all endotypes within the metabolism branch. This is in line with the expected behaviour of this heterogeneous disease, as it is associated with multiple IRPs, ethnicities, and age groups [Mai+23] [KA22]. Therefore, clinical features alone are not sufficient to accurately describe the endotypes within the metabolism branch, highlighting the need for molecular markers.

I observed an upregulation of a potential biomarker in E5/E6. The gene *LRG1* exhibited higher expression in the E5/E6 cluster compared to E7 - E10. This gene has been suggested as a diagnostic biomarker, prognostic biomarker, and therapeutic target in, e.g., autoimmune diseases and cancer [Dri+22]. The elevated expression of *LRG1* in E5/E6 suggests a potential therapeutic response to LRG1-inhibiting drugs, indicating its potential as drug target for these endotypes.

The clustering analysis identified a distinct molecular profile for pityriasis rubra pilaris within E5, challenging its conventional classification within IRP 3. In E5, pityriasis rubra pilaris was the dominant condition. Contrary to expectations, this skin condition did not cluster with psoriasis-dominated endotypes. This shows the limitations of current stratification approaches relying solely on clinical attributes [Soe86], which group these two conditions within the same IRP 3 [EE18]. Additionally, E5 displayed an upregulation of the genes *IL17C* and *PGLYRP2*, which are known to be associated with pityriasis rubra pilaris [Hay+23]. These findings would suggest an enrichment of dyskeratoses, pustules, and dermal lymphocytes [Soe86]. However, these clinical traits were less prominent in E5 compared to other endotypes. Thus, these findings show potentials of using transcriptomics data to more accurately differentiate pityriasis rubra pilaris from psoriasis.

The endotype E6 comprised patients diagnosed predominantly with Darier's disease and eczema. This suggests that these patients may benefit from the biologics commonly utilised for the treatment of eczema. This observation indicates a potential for drug repurposing, whereby biologics developed for eczema could be explored as a therapeutic option for Darier's disease within this specific endotype, offering a targeted treatment approach.

The endotype E8 consisted of a mixture of eczema and psoriasis patients, with an enrichment of similar pathways to those found in other other psoriasis-like endotypes E11 - E13. This could be an indication that these patients actually have eczematised psoriasis, a condition that clinically resembles features of both psoriasis and eczema [LE23]. However, the autoantigens *SERPINB3* and *SERPINB4*, which have recently been identified as drivers of eczematised psoriasis, are not upregulated in E8 [Jar+24]. As eczematised psoriasis is a common but poorly understood condition, the identification of endotypes may shed light on its pathogenesis.

Endotype E9 was composed of complex and overlapping conditions. The grouping of eczema, cutaneous lymphoma and parapsoriasis into endotype E9 confirmed that these diseases are clinically difficult to distinguish. Given their symptom similarities, cutaneous lymphoma and eczema can be confused, potentially leading to the incorrect administration of drugs, which could have life-threatening consequences for patients [HAL19]. Furthermore, parapsoriasis can be a precursor of cutaneous lymphoma [Kik+93] [EG99]. Therefore, it is important to identify the disease state of parapsoriasis and find biomarkers to distinguish between these three diseases. The endotypes already showed that these patients are distributed across several endotypes in the metabolism branch, including E5 - E10. In summary, endotype specific biomarkers are needed to differentiate between these heterogeneous skin conditions within the metabolism branch, leading to enhanced treatment suggestions.

The endotypes E11 - E13 form the psoriasis-dominated endotype branch within the inflammatory arm were characterised by distinct biological functions and clinical features. Endotype E12 most closely resembles the classical type of psoriasis, with upregulation of the psoriasis vs. eczema classifier gene *NOS2* and a more prominent presentation of typical clinical features of psoriasis [Qua+14] [Gar+16]. In comparison, no significant differences were observed in the clinical characteristics abnormal stratum granulosum and neutrophil accumulation, in E11 and E13. In addition, parakeratosis was not significantly associated with E13, but I observed an upregulation of the cornified envelope gene *LCE3C* and the small proline-rich proteins gene *SPRR2G*. These two genes have been identified as risk factors for psoriasis [Ber+11] [Car+13]. These differences strengthen the concept of psoriasis endotypes driven by distinct biological processes.

The molecular and clinical variations observed across E11-E13 could be explained by disease progression and psoriasis subtypes. Evidence supporting the disease progression hypothesis is provided by the increased accumulation of neutrophils, absence of a granular layer, and more pronounced acanthosis and parakeratosis in E12. These are all features commonly associated with the advanced stage of psoriasis [DRM11]. In contrast, E11 and E13 showed less classic psoriasis features in comparison to E12. Supporting the psoriasis subtype hypothesis, Ainali et al. (2012) identified subtypes of psoriasis using microarray data [Ain+12]. The authors hypothesised that one subtype, which was enriched in TGFb and ERbB signalling pathways, might respond to therapies targeting these pathways. Consistently, I observed enrichment of cell cycle-related pathways associated with ERbB signalling pathway in E11 - E13, along with an upregulation of *TGF-α* in E12. Further research is needed to determine whether these molecular and clinical variations reflect distinct progression stages, disease subtypes, or therapeutic responses.

Distinct clinical features characterised the endotypes in the autoinflammatory and interface dermatitis branch. In endotype E1, eczema was the most prevalent disease, while E2 was dominated by pyoderma gangrenosum and venous ulcer. E2 also showed a higher incidence of leukocytoclastic vasculitis, which was exclusively enriched in this endotype, and a lower incidence of scaling. This indicates that the stratification of patients into endotype is also detectable using clinical assessments. Similarly, the interface dermatitis branch showed characteristics consistent with expected clinical features, with E3 (lupus erythematosus) and E4 (lichen planus) having well-defined and distinct clinical features even prior to the categorisation into endotypes. These findings demonstrate the alignment between E2, E3, and E4 with clinical observations, offering potential improvements in diagnostic and therapeutic approaches.

One limitation of this study is the inclusion of only 21 out of over hundreds of defined ncISDs [EE18] and the bias towards more common ncISDs. Therefore, increasing the variety of diseases and expanding the sample size could yield more endotypes. In addition, the dataset was generated by multiple biologists from the same laboratory. Although I accounted for technical bias, the integration of other cohorts from different laboratories may affect the result but may provide a more generalised view and therefore, an optimised stratification of patients into endotypes. Other factors influencing the outcome, are the choice of processing pipeline and reference genome, affecting the clustering and downstream analysis. In essence, more diverse datasets would enable the identification of shared and distinct characteristics in transcriptomics data, enhancing the robustness of findings.

The dataset composition also affects the HVG selection process. Currently, the developed method, HRG, selects global, highly variable features based on their gene expression profile. As the dataset predominantly includes psoriasis and eczema samples, the feature selection method might be biased towards these overrepresented diseases. Thus, some HVGs may not be detected, because they do not belong to the top global variation drivers in the whole dataset. To address this, one potential solution is to define hierarchical subgroups within the data and select features specific to each subgroup, as suggested by Tyler et al. (2024) [Tyl+24]. This would allow for the identification of features that might otherwise be overlooked in the dataset.

Another technical limitation of the HRG method is the use of the K-Means clustering algorithm, which is best suited for datasets with well-separated data points and hyperspherical clusters of equal size [CKV13] [SRN16] [RA18] [AR14]. In biological data, which is heterogeneous and noisy, these conditions do not always apply. To address these issues, future work could explore alternative clustering algorithms, such as DBSCAN [Est+96], which better handle noisy and complex data structures.

The DBI used in the HVG selection pipeline is sensitive to outliers [TPM13] and thus, might fail to capture important biological drivers of under-represented groups within imbalanced datasets. Evaluating metrics less sensitive to outliers could further improve clustering accuracy and robustness. A comprehensive benchmarking of the automated pipeline, including HRG, against other feature selection approaches across diverse datasets is also needed to validate its utility and generalisability. Incorporating alternative evaluation metrics can enhance the pipeline's reliability and applicability.

Limitations also arise from the approach used to account for missing values in the clinical attributes, necessitating cautious interpretation of the association between clinical characteristics and endotypes. The assumption that the data is MCAR and the subsequent choice of imputation technique can also impact the results. While the applied imputation approach relied on the KNN algorithm, alternative methods such as multiple imputation [VBGO11] [Lit88] or model-based approaches [DLR77] [SB12] may offer improved accuracy, particularly in cases where missingness is dependent on observed or unobserved factors. Future studies should also systematically assess the nature of missing data and verify the MCAR assumptions. Moreover, sensitivity analyses comparing different imputation techniques could help to assess the robustness of findings and minimise potential biases introduced by a single data imputation technique. Integrating multiple imputation techniques and validating assumptions about missingness may further improve the reliability of the association analysis between endotypes and clinical traits.

For future applications, the study of ncISDs could greatly benefit from the integration of datasets from various labs, expanding beyond the current set of 21 ncISDs. One promising approach to achieve this is federated unsupervised representation learning [Zha+23], where researchers collaborate globally without the need to share sensitive data. This could improve the representation of ncISDs and subsequently, allow for more comprehensive downstream analysis.

Refining the HVG selection appraoch further enhances the robustness and reliability of the selected HVG for downstream analyses. The pipeline provides a valuable approach to detect HVGs without the need to manually specify the number of expected HVGs in the dataset. Benchmarking the internally used feature selection method, HRG, against more methods and across additional datasets is necessary to demonstrate its competitiveness. Furthermore, other validation metrics may indicate the need for a different number of genes for optimal clustering. It is also important to note that the choice of initialisation value can significantly impact clustering results, as the algorithm is not deterministic. Thus, further benchmarking, validation, and assessment of initialisation effects are necessary to optimise the HRG method and enhance its applicability across diverse datasets.

Molecular endotype classifiers could enhance the accuracy of diagnosis. This study confirmed the heterogeneity of ncISDs and the complexity of linking clinical phenotypes to specific endotypes. Thus, there is a need to integrate gene expression information to improve diagnosis. Molecular classifiers that distinguish between diseases or endotypes could assist in stratifying patients [Gar+16] [Fis+23] [DER]. In addition, the classifiers will be beneficial for rare diseases or clinically challenging conditions. These classifiers could be pivotal for tailoring treatment strategies and improving diagnostic accuracy.

Looking forward, the integration of various data types will likely transform our understanding of ncISD. For instance, the integration of genetic data from GWAS (Genomewide Association Study) [Man10], could provide additional information on possible mutations in diseases or endotypes. For example, known mutations in *ATP2A2* for Darier's disease [Sak+99], *CARD14* for pityriasis rubra pilaris and psoriasis [IM18], and FLG for eczema [Blu+18] [SKK20], highlight the importance of understanding the genetic basis of ncISD. The integrated approach will lead to more precise stratification of patients and the development of targeted therapies. As more disease-specific biomarkers are identified, precision medicine will advance. Detecting and treating these diseases at the molecular level will enable earlier, more effective interventions, reducing the burden of these complex and heterogeneous diseases.

In summary, this study provides an approach to define and characterise ncISDs endotypes using gene expression data. By addressing the identified limitations and integrating additional data types, these endotypes offer a promising step towards precision medicine. They have the potential to revolutionise the management of these conditions in clinical settings, enabling more tailored and effective treatment strategies.

Chapter 4

Gene-expression-based psoriasis-like endotype classifiers

Overlapping entities and limited understanding of subtypes reduce the diagnostic accuracy and treatment success rate of non-communicable chronic inflammatory skin diseases (ncISDs). Psoriasis, one of the most prevalent ncISDs worldwide, immensely impacts the quality of life and is often accompanied by comorbidities such as cardiovascular diseases [Gri+21]. Over the past decades, antibody therapies have become available. Amongst these therapies are the commonly administered antibodies IL-23, IL-17, and TNF-α. Notably, drugs targeting IL-23 and IL-17 achieve efficacy rates of up to 90 % in three out of four patients within three months of therapy [Sbi+23]. Administered therapies depend on diagnosis or immune response pattern (IRP), which are based on a patient's clinical profile and do not take into account genetic information. This often results in misdiagnosis and prescribing ineffective or incorrect treatments, which can, in the worst case, have life-threatening consequences for the patient [Has+24] [WBB18]. My previously identified endotypes, E1-E13, showed high biological and clinical relevance (Chapter 3). Furthermore, these endotypes are objective, as opposed to the established subjective clinical characterisation of ncISDs. This makes psoriasis-like endotypes ideal candidates to complement state-of-the-art diagnostic methods for psoriasis, improving treatment outcomes and reducing the risk of adverse effects.

In order to approach these challenges, I utilise therapy response data of psoriasis patients and link them to the psoriasis-like endotypes E8, E11-E13. I hypothesise that these endotypes are correlated with therapy efficacy. In addition, I aim to identify molecular markers that can be used to distinguish between the endotypes, with the future objective of enhancing diagnostic accuracy. The classification based on biomarkers is already a well-established practice in cancer treatment and introduced in the field of ncISDs for psoriasis and eczema by Dermagnostix [Ama+20] [Gar+16]. Therefore, I develop a feature selection pipeline and a gene selection method, "GeneSTRIVE", to identify robust genes, serving as molecular markers for the psoriasis-like endotypes E8, E11-E13. These genes will be used to build gene-expression based classifiers that potentially support treatment guidance.

Molecular markers are identified by stratifying effectively between the defined groups of endotypes, which are hypothesised to be associated with therapy response. The classifiers and hypothesis are evaluated using an independent cohort, and the results demonstrate the generalisability of the classifiers. While the dataset size limits definitive conclusions regarding the hypothesis, my study provides a foundation for future validation using a larger cohort.

In this Chapter clinical diagnosis and endotypes are integrated with therapy response information. I aim to create classifiers that predict groups of psoriasis-like endotypes with hypothesised similar response profiles. This is achieved by identifying genes using my feature selection method, GeneSTRIVE. It identifies robust molecular markers using prior knowledge and is able to handle high-dimensional, noisy gene expression data.

Specifically, I address the deliverable "Linking endotypes to drug response to advance precision medicine" and its associated research questions, as outlined in Section 1.6:

- Objective (iv), identify endotypes, which are likely associated to with response to the drugs targeting IL - 23, IL - 17, TNF - α (Section 4.3.1). Therefore, I utilised a cohort of patients diagnosed with psoriasis and investigated their responsiveness based on the ΔPGA score at week 0 and 12.
- Objective (v), identify a minimal set of single genes for the selected groups of endotypes (Section 4.3.3 and Section 4.3.4), to build gene-expression-based psoriasis-like endotype classifiers (Section 4.3.5).

This study represents a continuation of the project introduced in Chapter 3:
Hillig, Christina[*] and Meinel, Martin[*], and Seiringer, Peter[*]and Garzorz-Stark, Natalie[*] and Harder, Inken and Hübenthal, Matthias and Lochmann, Niklas and Mishra, Jigyansa and Farnoud, Ali and Maboudi Afkham, Heydar and Jargosch, Manja and Weidinger, Stephan and Eyerich, Stefanie[*] and Eyerich, Kilian[*] and Menden, Michael P.[*,†]. "Minimal set of predictive biomarkers enable endotype classification for precision medicine in inflammatory skin diseases." (*In preparation*).

In particular, my contributions are as follows. I conceived, conceptualised, and developed a robust feature selection method called GeneSTRIVE. In addition, I designed a triad feature selection pipeline and applied nested cross-validation (CV) to find the best

[*] Contributed equally
[†] Corresponding author

classifier models. Subsequently, I applied the Bayesian correlated t-test to identify the best performing model for each classification task. In order to test my hypothesis, I preprocessed an independent test cohort, which was provided by Inken Harder, Matthias Hübenthal, and Stephan Weidinger. I created and designed almost all figure panels, except those from Figure 4.1 and Figure 4.2.

The following parts were carried out by co-authors. The therapy response information was collected by Peter Seiringer. Ali Farnoud designed a metric to determined whether a patient responded to a therapy. Martin Meinel imputed the severity scores and performed differential gene expression (DGE) analysis and gene set enrichment analysis (GSEA) on the predefined endotypes, which I hypothesised to be related to therapy response. The figure panels, showing the biological interpretation of therapy response information in the context of endotypes (Figure 4.2), was mainly created by Martin Meinel and I created the two Volcanoplots. The results were interpreted together with my supervisor Michael Menden and our collaborators.

4.1 Materials

4.1.1 Therapy response cohort

In this chapter, the processed gene expression data, identified endotypes, sex, age, and batch information from Chapter 3 were used. In addition, therapy response data of 34 psoriasis patients, including their administered drug targets - IL - 23, IL - 17, and TNF - α - and Physician Global Assessment (PGA) scores before treatment, i.e. week 0, and during treatment, week 12 and 20, were collected (Table 4.1). Patients were treated with the drug targets IL - 23 (n=14; guselkumab, risankizumab), IL - 17 (n=15; secukinumab, ixekizumab, brodalumab), and/or TNF - α (n=14; infliximab, adalimumab). Amongst the 34 psoriasis patients, nine patients had been successively treated with multiple biologics. Specifically, eight patients have been treated with biologics inhibiting IL - 17 / TNF - α and one with IL - 17 / IL - 23 antibodies.

I leveraged the unpublished ncISDs bulk RNA-sequencing (RNA-seq) cohort comprising 342 lesional (L) and 287 non-lesional (NL) samples from Chapter 3. In order to match the disease with the therapy response information, the conducted analysis was only applied to patients belonging to the psoriasis-like endotypes E8, E11-E13 and patients who only provided NL samples (n_L=125, n_{NL}=129). Additionally, information about sex, age, batch, and endotypes were provided. Moreover, the 34 psoriasis patients, who provided

Drug target	**Sex**	**Patients**	**Age**	**PGA score**		
				Week 0	**Week 12**	**Week 20**
IL-23	male	11	44.27±16.46	4.36±0.50	1.27±0.47	1.00±0.71
	female	3	53.00±12.53	3.67±1.53	1.67±1.15	1.11±0.84
IL-17	male	11	48.64±13.81	3.73±1.42	1.33±1.14	1.02±1.43
	female	4	56.25±10.90	3.25±1.71	1.85±1.70	0.38±0.48
TNF-α	male	11	48.27±10.00	4.09±0.54	2.30±1.05	2.01±1.02
	female	3	57.67±12.90	4.00±1.00	2.67±1.15	2.71±1.55

Table 4.1: Overview of the unpublished, interpolated therapy response information of 34 psoriasis patients. The data has been collected and provided by Peter Seiringer, Natalie Garzorz-Stark, and Kilian Eyerich from the Department of Dermatology and Allergy - Technical University of Munich, Division of Dermatology and Venereology - Karolinska Institute, and Department of Dermatology and Venerology, Medical Center - University of Freiburg.

Drug target	**Endotypes**					
	E6	**E7**	**E8**	**E11**	**E12**	**E13**
IL - 23	1	1	1	3	6	2
IL - 17	0	1	2	4	6	2
TNF - α	0	1	1	4	5	3

Table 4.2: Commonly administered drug targets linked to the previously defined endotypes (Chapter 3) of the 34 psoriasis patients.

information about their severity before and during treatment, were also included in the psoriasis bulk RNA-seq dataset. An overview of the number of patients per drug target and endotype can be found in Table 4.2.

4.1.2 Independent psoriasis cohort

A second psoriasis bulk RNA-seq cohort (n=22) was provided by Inken Harder, Matthias Hübenthal, and Stephan Weidinger from the University Hospital Schleswig-Holstein, Kiel. The material was loaded onto the first and second lanes of S4 flowcells and subsequently provided to sequencing using the NovaSeq6000 sequencer. All samples originated from the same batch. The clinical information includes the time of visit, age, sex, severity scores, and administered drug (Table 4.3). In total 11 patients were treated with biologics targeting IL - 23 (risankizumab), eight with IL - 17 (ixekizumab), and three with TNF - α

Drug target	**Sex**	**Samples**	**Patients**	**Age**	**PtGA score**	
					Week 0	**Week 14**
IL - 23	male	10 L	10	50.80±15.74	2.9 ± 0.7	1.11±0.57
	female	1 L	1	46	3	0
TNF - α	male	2 L	2	54.5 ± 7.78	3 ± 0	3 ± 1
	female	1 L	1	60	3	3
IL - 17	male	5 L	5	47.40±13.85	2.8 ± 0.4	0.75±0.43
	female	3 L	3	40.33 ± 4.73	3.67±0.47	1.0

Table 4.3: Overview of the unpublished RNA-seq cohort from Inken Harder, Matthias Hübenthal, and Stephan Weidinger from the University Hospital Schleswig-Holstein, Kiel.

(adalimumab). For 20 patients the Patient Global Assessment (PtGA) score before and after treatment was provided. For one patient treated with an IL - 23 and one with an IL - 17 inhibitor, I did not receive the information on the disease severity at week 14. Additionally, the raw, unfiltered count matrix, comprising 22 samples and $19,019$ genes, was supplied.

4.2 Methods

4.2.1 Therapy response data

The therapy response data included information about the therapy, drug target, and severity (PGA) scores at week 0, 12, and 20. Recapitulating, the highest PGA score of 5 indicates a severe form of the disease, while a value of 0 represents the absence of visible symptoms in the patient (Section 1.2). As not every patient visited the clinic at each time point, the data collected about the severity scores was relatively sparse. Thus, the data was imputed by Martin Meinel. He employed a linear interpolation technique to estimate the PGA scores between the closest predecessor and successor PGA scores, using the PGA scores of week 0, 12, and 20.

In order to determine whether a patient had responded to a prescribed treatment, Ali Farnoud described the change in response ΔPGA_{12} from week 0 to week 12 by

$$\Delta\mathrm{PGA}_{12} = \frac{\mathrm{PGA}_0 - \mathrm{PGA}_{12}}{\mathrm{PGA}_0}, \tag{4.1}$$

where PGA_{12} is the severity score at week 12 of a patient's visit at the clinic. To define a patient as a responder or non-responder a $\Delta\mathrm{PGA}_{12} \geq 0.5$ or $\Delta\mathrm{PGA}_{12} < 0.5$ was required.

In order to assess whether there are differences between and within the response to the drug targets IL - 23, IL - 17, and TNF - α within a psoriasis-like endotype, the two-sided Mann-Whitney-U test was applied [Sid57]. In addition, the effect size, given by Cohen's d, was calculated [Coh13]. Moreover, I run the two-sided Mann-Whitney-U test and calculated Cohen's d to assess whether specific endotype groups show significant differences in their response profile towards the same drug target. I set the significance threshold to 0.2. The values of Cohen's d are defined within -∞ to ∞. They can be interpreted based on the following definitions. Absolute values of Cohen's d of approximately 0.2, 0.5, and 0.8 mark small, medium, and large effects, respectively [Coh13].

4.2.2 Psoriasis bulk RNA-sequencing cohort

The same preprocessing steps, as outlined in Section 3.2.1.1, were applied. The final preprocessed count matrix contained 632 samples and $17,816$ genes from 381 patients of 21 ncISDs. In addition, I leveraged the identified endotypes E1-E13.

The DGE analysis between defined groups of endotypes was conducted using edgeR [RMS10]. The count matrix was filtered for zero count genes and transcript per million (TPM) values. Additionally, filtering on the count - per - million (CPM) values was applied using the formed groups of endotypes as group variable (Section 2.6.1). The following design for the DGE analysis was formulated

$$y_g \sim \text{age} + \text{sex} + \text{batchID} + \text{group} \ , \tag{4.2}$$

where y_g is the log CPM value of gene g and the *group* variable contains the two groups to compare. These are E8/E11 against E12/E13 and E12 against E13. The p-values were corrected for false discovery rate (FDR) using Benjamini and Hochberg (BH) [BH95]. In order to call a gene significantly differentially expressed gene (DEG), I applied the thresholds p - adjusted (padj) value < 0.05 and $|\text{log2FC}| > 1$. Subsequently, GSEA was performed using all genes sorted by their signed padj value ($\text{sign}[\text{log2FC}] \cdot \text{padj value}$). Since GSEA operates on gene entrezIDs, biomaRt [Dur+09] was used to convert the gene names into entrezIDs. The significance threshold of the p-value was set to 0.05 and BH was used as FDR correction method.

The Bioconductor package DOSE [Yu+14] was then applied to make the resulting entrezIDs in the GSEA object readable again. The Bioconductor [Gen+04] packages ReactomePA [YH16], org.Hs.eg.db [Car+19], and enrichplot [Yu21] were used.

4.2.3 Endotype classifiers

To develop two binary endotype classifiers, E12/E13 vs. E8/E11 (n=254) and E12 vs. E13 (n=175), the corresponding datasets containing L and NL samples were split into training (80 %) and test (20 %) sets in a stratified manner using `train_test_split` from the Python package Scikit-learn [Ped+11]. The dataset included 79 L samples for E12/E13, 46 L samples for E8/E11, 48 L samples for E12, and 31 L samples for E13. In addition, there were 129 NL samples for E12/E13 vs. E8/E11 and 96 NL samples for E12 vs. E13. The NL samples were only used in the developed feature selection method "GeneSTRIVE" (Section 4.3.3) and were excluded prior to the training and testing of the classifiers.

4.2.3.1 Triad feature selection pipeline

In this study, a triad feature selection pipeline was implemented to reduce the number of input features and optimise the performance of the model. The pipeline consists of three main parts, each serving a distinct purpose in identifying the most robust and discriminative features within the dataset.

Before training of the classifier, the number of input features ($17,816$) was reduced. Nested k-fold CV (Section 2.4.3.3) was performed to account for sample heterogeneity in the defined endotype groups and supports the selection of robust genes. The training set was split and shuffled using stratified k-fold CV with $k = 5$ partitions from the python package Scikit-learn. For each partition the following steps were executed.

In part one, a feature preselection was conducted using my feature selection method “GeneSTRIVE”, which was run $N = 1000$ times (Figure 4.3 c, step 1-2). The most frequent occurring feature candidates were selected by performing a binomial test on each gene and correcting the resulting p-values using BH. In “GeneSTRIVE”, this step is implemented as an item frequency analysis, "Dominance Ranking", in the function `Dominance Ranking` (Figure 4.3 c, step 3). The results were so called *class-specific genes.*

In the second part, the features were further reduced using variance inflation factor (VIF) and machine learning (ML)-based techniques. Therefore, I leveraged these genes and proceeded by using only L samples (Figure 4.3 d, step 4). An optional filtering method, VIF (Figure 4.3 d, step 5), to remove highly co-correlated genes, having a VIF > 5, and one of the following feature reduction techniques, L1-regularisation, recursive feature selection with cross-validation (RFSCV) or sequential feature selection (SFS), were applied (Figure 4.3 d, step 6). In the feature reduction technique, I accounted for class

Technique	Model	Parameter	Values
L1-regularisation	Logistic regression	C penalty	[0.1, 1, ... 10] L1
SFS	SVM	C kernel	[0.25, 0.5, 1.0, 2.0, 3.0, 4.0] [linear, poly, rbf, sigmoid]
RFSCV	Random Forest	n_estimators	[10, 20, ..., 100]

Table 4.4: Hyperparameters of the classifiers employed in the feature selection.

imbalance using the upsampling tactic Support Vector Machine (SVM)-SMOTE with the default sampling strategy. SVM-SMOTE is implemented in the python package imbalanced-learn [LNA17]. In addition, the classifier models of each feature selection approach were optimised using GridSearchCV from the package Scikit-learn [Ped+11] with stratified k-fold CV $k_{inner} = 3$ and the weighted F1-score (Section 2.4.3.4, eq. 2.19) for evaluating the CV results. A logistic regression model for L1-regularisation, a SVM model for SFS, and a Random Forest model for RFSCV were used. The corresponding hyperparameters are shown in Table 4.4.

Following the completion of parts one and two of the triad feature selection pipeline across all $k = 5$ partitions, the final step utilised an optimised classifier to refine the feature set. This step incorporated all unique features identified from the five partitions, alongside the entire training set, which comprised only L samples. An optional filtering method VIF and embedded or wrapper techniques were applied. The final feature set contained robust and stable features across all five partitions. Subsequently, these features were used to train the classifiers on the L training set and their performance was evaluated on the test set (Figure 4.3 e).

4.2.3.2 Classifiers and their evaluation

To construct optimal binary classifiers, I designed a pipeline that integrated feature selection, model selection, hyperparameter tuning, and an upsampling strategy SVM-SMOTE addressing class imbalances [LNA17] [NCK11] (Section 2.4.3.1.2). GridSearchCV [Ped+11] was employed for hyperparameter optimisation using repeated stratified k-fold CV, the number of folds $k_c = 3$ and 10 repeats. The weighted F1-score (Section 2.4.3.4, eq. 2.19) was used as evaluation metric to assess the classifiers' performance within GridSearchCV. An overview of the models and their hyperparameters is presented in Table 4.5. In the Multi-layer Perceptron (MLP) model, I set the parameter batch_size to 5. In addition, the random_state was set to 0 for all models. The final feature selection, including all five feature sets, the training, and the optimisation of the classifiers was conducted for multiple parameter combinations, yielding 18 models per classification task (Figure 4.4 a, b). The decision boundary for classification was set to 0.5.

In order to select the best model, I considered posterior probabilities of the Bayesian correlated t-test. This allows to compare the CV performances of all created classifiers in a pairwise fashion. The Bayesian correlated t-test is implemented in the python package baycomp [Ben+17]. In order to ensure consistency with the number of repeats used in the repeated stratified k-fold CV, the run parameter in the Bayesian correlated t-test was set to 10. Additionally, the rope parameter was set to 0.01, as I anticipate that an identical performance of two classifiers is unlikely to occur.

In order to provide more realistic predictions, I calibrated the final models using Platt's method implemented in the function `CalibratedClassifierCV` from Scikit-learn [Ped+11] by setting the parameter method to sigmoid and using repeated stratified k-fold CV. The two best calibrated models for each classification task were compared to a baseline classifier that always predicts the majority class. For this purpose, a dummy classifier from Scikit-learn was used.

I further evaluated the classifiers performances on samples from unknown classes. These are samples, which have not been seen by the classifier before, but potentially share characteristics with the training set. In total 217 and 263 samples from the endotypes E1-E7 and E9-E11 were used to assess the classifiers behaviour on unseen classes.

I also visually assessed the sample distribution by embedding the classifier features in 2D using Uniform Manifold Approximation and Projection (UMAP) by setting random_state to 0 [MHM18] [Ped+11].

4.2.4 Independent psoriasis bulk RNA-sequencing cohort

In order to compare the independent psoriasis cohort (n=22) to my dataset, the Ensembl IDs were converted to HUGO gene names using biomaRt [Dur+09].
Psoriasis samples from the Kiel cohort were assigned to the defined endotype clusters (Section 3.3.3). To achieve this, the highly variable genes (HVGs) used for clustering (Section 3.3.2.2) were selected. Missing HVGs values were imputed using the mean gene expression levels of the housekeeping genes (HKGs) *TBP*, *SDHAF2*, *TSR3*, and

Model	Parameter	Values
MLP	hidden_layer_sizes	[(12, 8, 6), (8, 6, 4), (10, 5), (6, 4), (8,), (7,), (6,), (5,), (4,), (3,), (2,)]
	activation	[identity, logistic, tanh, relu]
	alpha	[0.0001, 0.0005, 0.00001]
	learning_rate_init	[0.0001]
	sampling_strategy	[minority, majority, all]
SVM	C	[0.25, 0.5, 1.0, 2.0, 3.0, 4.0]
	kernel	[linear, poly, rbf, sigmoid]
	sampling_strategy	[minority, majority, all]
XGBoost	n_estimators	[10, 100, 500]
	eta	[0.1, 0.2, 0.3, 0.4]
	max_depth	[3, 4, 5, 6]
	sampling_strategy	[minority, majority, all]

Table 4.5: Hyperparameters of the MLP, SVM, and XGBoost model.

GAPDH. The Kiel psoriasis samples were subsequently integrated with the L ncISDs cohort from the Eyerich lab (Materials 3.1) through data normalisation, followed by batch correction for the covariates sex and batchID. To incorporate new samples into the graph object, the nearest existing graph node was identified using k-nearest neighbours with $k = 1$. Independent psoriasis samples were then assigned to their corresponding nearest endotype node. Finally, the UMAP embedding of the two integrated cohorts was generated.

To predict groups of endotypes using gene-expression-based classifiers, I determined the intersection between genes in my dataset $(17,816)$ and in an independent cohort $(19,019)$. In total 95 genes were found to be missing in the independent dataset. To facilitate normalisation, a preparatory step for the classifiers, the missing genes were imputed by replacing their values with the mean gene expression of the HKGs *TBP*, *SDHAF2*, *TSR3*, and *GAPDH*. Following this, the classifiers were applied to the independent dataset to predict the endotype label.

4.2.5 Hypothesis evaluation for each drug target

In order to assess whether there is a significant difference in the response profile of IL - 23, IL - 17, and TNF - α between and within the defined groups of psoriasis-like endotypes, the two-sided Mann-Whitney-U test was applied [Sid57]. The effect size, given by Cohen's d, was calculated [Coh13] and the significance threshold for the p-values was set to 0.2.

4.3 Results

To test whether psoriasis-like endotypes are associated with therapy response, I leveraged 204 L and 225 NL bulk RNA-seq samples and disease severity information of week 0, 12, and 20. The latter was collected from 34 psoriasis patients treated with IL - 23 (n=14), IL - 17 (n=15), and TNF - α (n=14) inhibitors. Amongst these patients are nine that have been successively treated with multiple antibodies. Specifically, eight have been treated with anti IL - 17/TNF - α and one with anti IL - 17/IL - 23 (Methods 4.1.1). Additionally, I used matched clinical assessments, endotype labels (Section 3.3.2, Table 4.2), and L and NL bulk RNA-seq samples (Section 4.2.2). This dataset enabled me to investigate the association between drug targets and psoriasis-like endotypes.

4.3.1 Trend of response patterns in psoriasis-like endotypes

In order to link the psoriasis-like endotypes to the three drug targets, IL - 23, IL - 17, and TNF - α (Table 4.1), I used the endotypes E8, E11, E12, and E13, as they grouped psoriasis into several clusters (Chapter 3, Figure 3.8). Treatment success was determined for each drug target by calculating the ΔPGA scores between week 0 and 12 using the imputed PGA scores (Methods 4.2.1). A mean ΔPGA score of 0.25 for IL - 23, 0.83 ± 0.11 for IL - 17, and 0.75 for TNF - α in E8 was measured. In E11, I observed a ΔPGA score of 0.58 ± 0.14 for IL - 23, 0.68 ± 0.34 for IL - 17, and 0.50 ± 0.20 for TNF - α. In the endotype E12, a mean ΔPGA score of 0.73 ± 0.12 for IL - 23, 0.40 ± 0.33 for IL - 17, and 0.15 ± 0.14 for TNF - α was observed. A mean ΔPGA score of 0.78 ± 0.04 for IL - 23, 0.46 ± 0.48 for IL - 17, and 0.62 ± 0.12 for TNF - α in E13 was measured. In essence, the mean suggested a trend of a stronger association of E12 and E13 with IL - 23 as well as TNF - α, and for E8 and E11 with IL - 17, indicating a possible endotype drug target correlation.

A comparative analysis of drug response within endotypes was conducted. I investigated whether there is a difference in the drug response between the drug targets within an endotype using the two-sided Mann-Whitney-U test and calculating Cohen's d (Figure 4.1, Methods 4.2.1). I observed no significant padj values for the endotypes E8, E11, and E13 and drug targets IL - 17, TNF - α, and IL - 23. For E12 I obtained significant padj values, indicating that patients treated with the biologics targeting IL - 23 resulted in a greater change in their severity compared to patients treated with biologics that target TNF - α and IL - 17 (IL - 23 vs. TNF - α: p-adj. = 0.06, Cohen's d = 4.54; IL - 23 vs. IL - 17: p-adj. = 0.13, Cohen's d = 1.34). This led to the conclusion that in E12, IL - 23 inhibitors may be more effective than biologics targeting TNF - α and IL - 17.

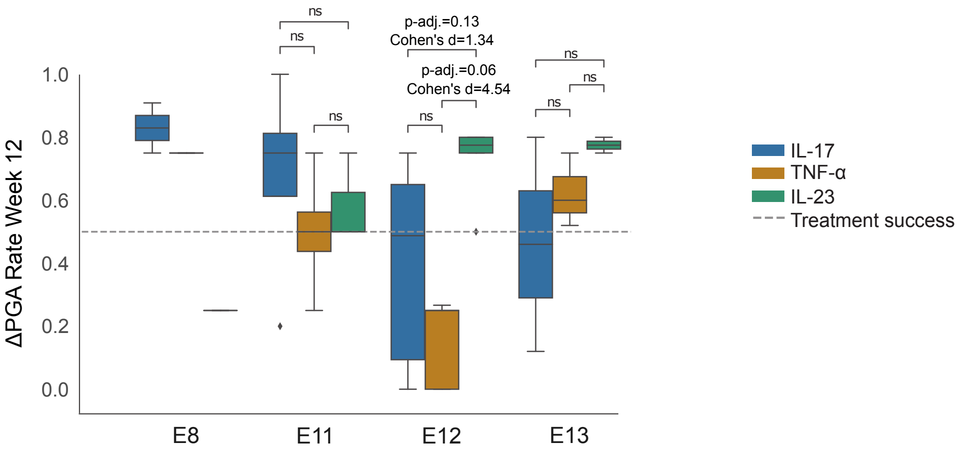


Figure 4.1: Responder patterns of commonly prescribed biologics that inhibit IL - 17, TNF - α, and IL - 23 in patients with of psoriasis. Boxplots showing the ΔPGA score for patients belonging to the psoriasis-like endotypes E8, E11 - E13 for the drug targets TNF - α, IL - 23, or IL - 17. The dashed line visualises treatment success, corresponding to a ΔPGA score ≥ 0.5. A two-sided Mann-Whitney-U test is performed between drug targets of the endotypes E11 - E13. The BH is used for multiple testing corrections. ns: padj > 0.2.

I further examined whether there is a positive drug response trend between specified groups of endotypes (Methods 4.2.1). A tendency of an effective treatment was visible for the following comparisons. For the drug target IL - 23, I found a significant result for E12/E13 against E8/E11 (p-value = 0.03, Cohen's d = 1.74), while for IL - 17, I observed a significant p-value for E8/ E11 compared to E12/E13 (p-value = 0.07, Cohen's d = 1.00). Moreover, for E13 against E12 I obtained a positive association to the drug target TNF - α (p-value = 0.03, Cohen's d = 3.54). The results indicated that specific drug targets may be more efficacious in certain endotypes.

Subsequently, I created the following psoriasis-like endotype groups of responder and non-responder for specific drug targets. Patients of endotype E8/E11 were declared as IL - 17 responders, E12/E13 as IL - 23 responders, E12 as TNF - α non-responders, and E13 as TNF - α responders. These new definitions increased the sample sizes for downstream analyses. Specifically, I associated 129 NL, 46 L E8/E11, and 79 L E12/E13 with IL - 17 responders and IL - 23 responders. In total 175 samples were related to TNF - α responders and non-responders. These samples split into 96 NL, 48 L E12, and 31 L E13 samples. This categorisation significantly enhanced sample size and offered valuable insights into the correlation between psoriasis-like endotypes and specific drug responses.

4.3.2 Biological functions of hypothesised drug-target associated endotype groups

To compare the biological functions of the new defined responder and non-responder endotype groups for each drug target, DGE analysis and GSEA were performed (Figure 4.2). These analyses aimed to identify molecular pathways and gene expression patterns that differentiate these endotype groups and offer insight into their respective drug responses.

The analysis revealed a higher expression in E12/E13 in comparison to E8/E11 of DEGs, such as *IL1B*, *NOS2*, *IL19*, *IL20*, *IL36A* (Figure 4.2 a). These genes are involved in type 3 immune response (Figure 4.2 b). Additional enriched pathways in E12/E13 compared to E8/E11 were "Translation", "rRNA processing", and "HATs acetylate histones". This suggests higher activities of protein synthesis and gene expression regulation. Conversely, the E8/E11 endotype group was characterised by keratin and collagen expression, as evidenced by the enrichment of the pathways "Keratinization", "Extraculluar matrix organization", and "Collagen chain trimerization". These findings highlight the molecular and pathway-level distinctions between E12/E13 and E8/E11, suggesting differing roles in inflammation, immune response, and tissue structure.

Furthermore, I compared E12 and E13, which are the TNF-α associated non-responder and responder endotypes, respectively. E12 showed an upregulation of inflammatory genes, such as *IL1B*, *CXCL5*, and *CXCL3*. This was supported by the top three enriched pathways "Chemokine receptors bind chemokine", "Peptide ligand-binding receptors", and "Signaling by Interleukins" (Figure 4.2 d). In comparison, E13 showed higher expression of the Late cornified envelope (LCE) genes family and *LORICRIN*, which are expressed in the cornified envelope (Figure 4.2 c). LORICRIN is a protein building the uppermost layer of the epidermis [CSM05]. Pathways enriched in E13 included "Eukaryotic Translation Elongation", "Eukaryotic Translation Termination", and "Nonsense Mediated Decay (NMD) independent of the exon junction complex" pathways, being involved in protein synthesis and its regulation (Figure 4.2 d). The DEGs and pathway analyses revealed differences in molecular events typically associated with psoriasis, highlighting a distinct emphasis on inflammatory signalling in E12 and structural protein synthesis and regulation in E13, which may explain their differential response to TNF - α inhibition and underlying drug response variability.

In summary, higher activation of pathways involved in the development of the epidermis were observed in the formed endotype group E8/E11, while in E12 and E13 inflammatory

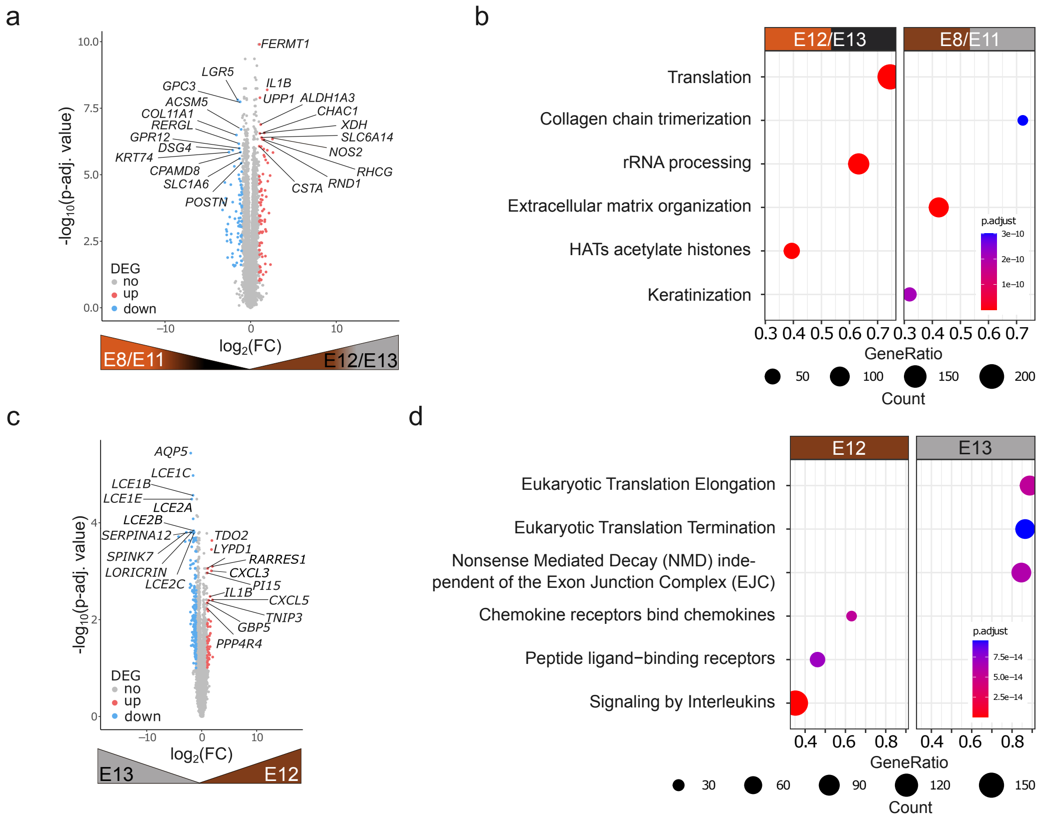


Figure 4.2: Molecular characterisation of psoriasis responders and non-responders to the drug targets IL-17, TNF-α, and IL-23. a, c) DGE analysis result comparing **(a)** E12/E13 against E8/E11 and **(c)** E12 against E13. DEGs are determined by requiring a padj value < 0.05 and log2FC < 1. The top DEGs on each side are annotated. **b, d)** GSEA result using the Reactome database.

or protein regulating pathways were higher regulated. Potentially, the observed differences in response to drug targets are attributed to the involvement of these distinct biological pathways. In order to utilise my proposed groups of endotypes for molecular diagnostics in clinics, I aimed for identifying biomarkers, which can be used as features in molecular classifiers. Therefore, I developed a gene selection method "GeneSTRIVE". Its resulting markers were used to create gene-expression-based classifiers that distinguish between E12/E13 vs. E8/E11 and E12 vs. E13.

4.3.3 Advanced gene selection method for gene expression data

In order to determine molecular markers for the gene-expression based classifiers, I had to identify meaningful genes (features) in the high-dimensional dataset, containing 17,816 genes. The selection process was constrained to identify a minimal set of single genes for each classifier, avoiding combinations of genes. Feature selection in ML is applied to identify the most informative features, thereby enhancing model performance and achieving better generalisation. Amongst the commonly used tools are reduction, filter, wrapper, and embedded methods (Section 2.4.3.1.3) [Pud+22].

Dimensionality reduction methods, such as principal component analysis (PCA) [Hot33], construct combinations of genes rather than selecting individual genes. In contrast, feature selection methods, such as filter, wrapper, and embedded approaches, identify single genes. Filter methods tend to yield general feature sets that are not specifically tailored to the prediction task, while wrapper and embedded methods are more task-specific. However, these may face challenges due to noise or dimensionality [Pud+22] [HK99].

Hybrid feature selection techniques address these challenges of filter, wrapper and embedded methods. They combine different feature selection techniques, for example filter and wrapper methods, enabling to operate on a high-dimensional feature spaces. This leads to improved performances in the prediction task [Gho+20]. Despite these advantages, hybrid approaches inherit the limitations of their filter-based components [Gho+20]. In conclusion, the need for a feature selection method that can handle high-dimensional, noisy data, is tailored to the prediction task, and provides a robust set of single features remains unmet.

In order to address these challenges, I introduce an advanced feature selection method called *GeneSTRIVE*. GeneSTRIVE is designed to operate on high-dimensional, noisy, and heterogeneous gene expression data, producing a list of robust single genes tailored to the prediction task.

GeneSTRIVE is designed to identify robust and generalisable genes from heterogeneous transcriptomics datasets through a structured, multi-step process. It employs stratified subsampling (Section 2.1), where a subset of samples is repeatedly selected from the classes of interest to ensure robust gene identification. GeneSTRIVE operates on a raw count matrix containing L and NL samples, their respective class labels, and a design function. Optional covariates, such as sex and age, can also be incorporated to account for additional variables influencing the analysis.

GeneSTRIVE contains three main steps. In step 1, the data is divided into stratified subsets based on class labels and provided categorical covariates. Subsequently, DGE analysis is performed either between L and NL samples or within L samples. The data stratification followed by DGE analyses are repeated $M \in \mathbb{N}$ times, allowing to identify robust genes accounting for the heterogeneity within the classes. In step 2, the identified DEGs are used to determine class-specific genes that are uniquely associated with each class. Steps 1 and 2 are repeated $N \in \mathbb{N}$ times, where the number of iterations N is dependent on the dataset size and its heterogeneity. Larger and more heterogeneous datasets typically require a higher number of iterations N to ensure comprehensive gene identification. Step 3 involves conducting an item frequency analysis on the resulting gene sets to finalise the list of robust and stable genes for each class. The final output is a collection of genes capable of effectively distinguishing heterogeneous classes. In summary, GeneSTRIVE employs a systematic, iterative approach to identify robust genes by accounting for class-specific heterogeneity in high-dimensional transcriptomics data.

4.3.3.1 Step 1: Identifying DGEs as feature candidates

In order to identify DGEs as candidates for the class-specific genes, a raw count matrix, class labels, optional covariates, and a design function must be provided. The following steps are executed for each iteration i of a total of N iterations.

The process begins by determining pairwise comparisons between the NL group and L of each class c_k, as well as between L samples between classes. Consequently, the input count matrix can be described as a set of M sub-count matrices $\mathcal{X} = \{X_1, \ldots, X_M\}$ with $X_m \in \mathbb{R}^{n \times g}$, where n is the number of samples and g the number of genes. Subsequently, so-called runs are executed. The number of runs is equal to the number of sub-count matrices M, and thus depends on the number of classes c. The number runs can be calculated using the following formula

$$M(c) = c + \frac{c!}{2! \cdot (c-2)!} \,. \tag{4.3}$$

A run is divided into three sequentially executed sub-steps (i-iii) and handles one pairwise comparison at a time, for example, comparing L vs. NL samples of one class or L samples of one class vs. another class.

(i) Sampling: Subsampling or stratified sampling is applied to account for the heterogeneity within the classes and class imbalance. Depending on whether one or more additional categorical covariates are provided, stratified sampling is used; otherwise, subsampling is applied. The sample size of the resulting sub-samples depends on the maximum number of samples that can be taken from the class with the fewest samples, considering additional covariates and the conditions of the subsampling or stratified sampling.

(ii) Gene filtering: The stratified or subsampled data is filtered to remove low expressed genes. A gene is retained if it is measured in at least two samples in one class with TPM values and counts exceeding one in sub-count X_m where $m \in [1, M]$.

(iii) DGE analysis: Using the edgeR package [RMS10], DGE analysis is performed based on the user-defined design matrix, which is of the form

$$y_{gs} \sim \text{covariate}_1 + \text{covariate}_2 + \text{condition} .$$

Here, y_{gs} represents normalised counts of gene g and samples s. The condition variable contains the pairwise comparison labels, while optional covariates adjust for confounding effects. FDR using BH is applied. DEGs are identified based on user-specific thresholds for log2FC and padj values. Up- and down-regulated DEGs are sorted by padj value and stored in a list for subsequent analysis.

Step 1 establishes a rigorous approach for identifying DEGs as feature candidates, leveraging stratified sampling, filtering, and DGE analysis. By iteratively performing pairwise comparisons, the process ensures the identification of genes that are both statistically significant and biologically relevant.

4.3.3.2 Step 2: Identification of class-specific genes

After M runs have been conducted in Step 1, the M lists of DEGs are used to identify class-specific genes. These are either up- or down regulated in comparison to all other classes but are not DEGs in other class comparisons. This means that they are uniquely associated with a specific class and show distinct expression patterns that distinguish them from the rest of the classes. Additionally, class-specific genes do not exhibit significant differences in expression levels when comparing any other pair of classes.

The process of determining class-specific genes can be divided into three sub-steps (a - c): In **sub-step (a)**, uniquely upregulated DEGs per class are identified across all L pairwise comparisons. In **sub-step (b)**, DEGs either uniquely up- or down-regulated in L vs. NL pairwise comparisons are defined. Subsequently, in **sub-step (c)**, the unique DEGs from step (a) and (b) are intersected. The result is a set of class-specific genes for each class.

The process of identifying class-specific genes is mathematically defined as follows. In step (a), sets $A_1, \ldots, A_c$ of upregulated genes per class j where $j \in [1, \ldots, c]$ from the L comparisons are defined. In step (b), sets $B_1, \ldots, B_c$ are defined, where the elements of B_j correspond to all up- and down-regulated DEGs of the L vs. NL comparison of a class j. In order to identify unique DEGs in step (a) and (b), the following definition is used.

Definition 2. *Let $S = \{S_1,\ \ldots,\ S_l\}$ be a set of l sets of DEGs. Then the unique elements of S_p with $p \in [1, \ldots, l]$ can be determined by*

$$S'_p = S_p \setminus \left(\bigcup_{k=1}^{l} S_k \right) \qquad \text{with } k \neq p.$$

The unique elements for each set A_j and B_j is given by A'_j and B'_j using the definition 2. Since I aim to identify sets of class-specific genes U_j for each class, I build the intersection of A'_j and B'_j by

$$U_j = A'_j \cap B'_j \qquad \text{with } j \in [1, \ldots, c] \ .$$

The resulting set of class-specific genes U_j is determined for each iteration $i \in [0, \ldots, N]$. These steps are repeated N times, resulting in $c \cdot N$ class-specific gene sets.

In summary, Step 2 identifies class-specific genes by isolating uniquely up- or down-regulated genes per class. The iterative process ensures reliability and specificity of these genes, enabling accurate characterisation of each class.

4.3.3.3 Step 3: Item frequency analysis

The item frequency analysis identifies genes that are overrepresented in the class-specific gene sets. It includes two methods, *Robust Rank Aggregation (RRA)* and *Dominance Ranking*, which assess the statistical significance of gene representation within the class-specific gene sets.

The RRA assesses whether genes are consistently highly ranked across class-specific gene sets [Kol+12]. This method assumes ranked input sets, a criterion met by the class-specific gene sets, where genes are ordered by adjusted padj values. The RRA procedure involves the following steps. For each gene, a rank vector $\boldsymbol{r} = \{r_1, \ldots, r_N\}$ is defined, where N denotes the number of class-specific gene sets U_j. Each element $r_j \in \boldsymbol{r}$ represents the gene's rank, normalised by the maximal rank l, corresponding to the total number of DEGs identified. To test whether all class-specific gene sets in U_j contain relevant rankings, I assume that informative elements in $\boldsymbol{r}$ come from a right-skewed distribution. A binomial test then evaluates the significance of observed rankings at the gene level [Kol+12]. The final score of $\boldsymbol{r}$ is given by

$$\zeta(\boldsymbol{r}) = \min_{l=1,\ldots,N} p_{l,N}(\boldsymbol{r}) \ ,$$

which is the robust rank score of a gene for class j in U_j. Since multiple tests are conducted, I correct for FDRs by applying BH on the scores ζ.

The Dominance Ranking method assesses whether a gene is observed more frequently amongst a group of specific gene candidates than expected by chance. It employs a one-tailed binomial test (eq. 2.2) with the alternative hypothesis $\pi > \pi_0$, where π represents the observed frequency of the gene, and π_0 denotes the expected frequency under the assumption of randomness. Since the test is applied to each gene, BH is applied to correct for FDRs.

Genes that meet the statistical significance threshold are considered robust candidates. The final result of Step 3 is a list of robust genes that are most distinctively expressed between the classes that can be used for downstream analysis.

In summary, GeneSTRIVE is a robust supervised feature selection method for identifying class-specific genes. It repeatedly applies DGE analysis on a sub-count matrices and performs an item frequency analysis to ensure the detection of robust genes with distinct expression patterns across different classes. GeneSTRIVE is capable of identifying single, robust features in noisy, high-dimensional, heterogeneous gene expression data.

4.3.4 Triad feature selection pipeline including GeneSTRIVE

Hybrid feature selection methods combine the strengths of two single techniques, leading to better performances [Gho+20]. Commonly, a filtering method is used as an initial step to reduce the high-dimensionality of the data. However, class-dependent filtering methods do not consider the class labels, producing overly general feature sets [Pud+22]. To address this limitation, I developed a triad feature selection pipeline, which includes GeneSTRIVE, an optional filtering method, and a subsequent wrapper or embedded feature selection technique (Figure 4.3, Methods 4.2.3). The pipeline is designed to handle imbalanced, heterogeneous transcriptomics data of any size, with the objective of identifying biomarkers that can be, e.g., employed in the diagnosis of diseases in clinical settings.

The triad feature selection pipeline is designed to handle imbalanced, heterogeneous transcriptomics data of any size. Its objective is to extract robust feature sets that

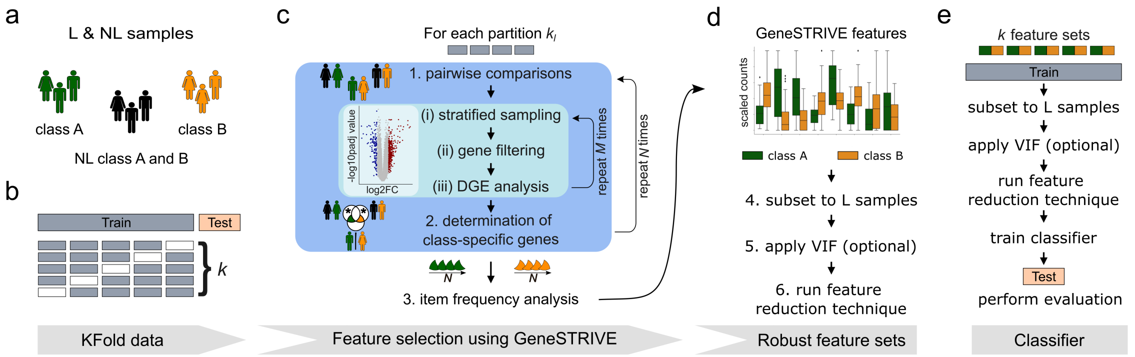


Figure 4.3: Triad feature selection pipeline including GeneSTRIVE. a) Data selection and annotation: Responders (green), non-responders (orange), and NL samples (black). **(b)** Data splitting: Division of dataset into training and test sets, with the training set being further partitioned into k folds using stratified k-fold CV. **c)** GeneSTRIVE Workflow: Step 1, pairwise comparisons: For each pairwise comparison (e.g., class A vs. class B) stratified sampling and DGE analysis are performed (steps i-iii). Step 2, class-specific genes: Identify DEGs specific to each class using VennDiagrams. Step 3, item frequency analysis: Determine frequently occurring features from N class-specific gene sets. **d)** Integration into triad feature selection pipeline: Step 4, reduction of training data: Filtered to L samples. Step 5, VIF filtering: Optional step. Step 6, reduction of features: Applied to each training set partition. **e)** final Steps: Generalise feature set, train classifier on entire L training samples, and evaluate on test set.

can address the heterogeneous nature of transcriptomics data while ensuring accurate classification. The pipeline has been applied to two binary classification tasks involving the psoriasis-like endotype groups E8/E11 vs. E12/E13 and E12 vs. E13. These groups are hypothesised to be associated with drug response (Section 4.3.1).

The input data includes L and NL samples, along with their respective class labels (Figure 4.3 a). Additionally, covariates such as age and sex are included, resulting in the following design function

$$y_{gs} \sim \text{age} + \text{sex} + \text{condition} \ ,$$

where age is a numerical variable, sex is categorical, and condition refers to either binary class labels or a class label and its corresponding NL label. The data is then divided into a training (80 %) and test (20 %) set in a stratified manner. To define a robust feature set for each classifier, nested k-fold CV is applied to account for partition bias. The training data is shuffled and split into $k = 5$ partitions using stratified k-fold CV (Figure 4.3 b).

The partitioned train sets are provided to my triad feature selection pipeline, which is divided into two parts, described in more detail below. Part 1 is my supervised feature selection method, GeneSTRIVE (Figure 4.3 c, step 1-3) and Part 2 includes additional feature selection approaches exclusively applied to L samples (Figure 4.3 d, step 4-6).

4.3.4.1 Part 1: Application of GeneSTRIVE

In **step 1** (Figure 4.3 c, step 1), combinations of pairwise comparisons are determined based on the number of classes $c = 2$ in the data, resulting in $M = 3$ pairs (eq. 4.3). I apply (i) stratified sampling, as it ensures balanced samples concerning covariates sex and condition. The latter represents the pairwise groups, for example, L samples of E12 and E13. Subsequently, for each pairwise comparison, lowly expressed genes are (ii) filtered and (iii) DGE analysis is performed. The steps (i-iii) are repeated M times.

In **step 2** (Figure 4.3 c, step 2), class-specific gene candidates are determined, which are both uniquely, differentially expressed between L and NL samples of the respective class and uniquely upregulated in the corresponding class from the L samples comparison of the two classes (Figure 4.3 c, green and orange area in VennDiagram). In order to identify DEGs, I require a gene to have a padj value smaller than 0.05 and |log2FC| above one. These thresholds are frequently employed and may be modified by the user. The steps 1-2 are repeated $N = 1000$ times, resulting in N class-specific gene sets.

In **step 3** (Figure 4.3 c, step 3), the item frequency analysis, Dominance Ranking, is applied to identify class-specific genes occurring with greater frequency than expected by chance. I require the resulting padj values to be smaller than 0.01.

4.3.4.2 Part 2: Further feature reduction methods

The resulting feature sets of each class defined by GeneSTRIVE are used to reduce the feature dimension of the L samples in **step 4** (Figure 4.3 d). Additional feature reduction tools, such as optional filtering by VIF and a feature reduction technique, are applied in **step 5 and 6**, respectively (Methods 4.2.3). This allows to further reduce the number of features and add the advantages of these selection methods. Recapitulating, the VIF can be employed to eliminate co-correlated features. Wrapper feature selection methods consider interactions and redundancies between features and embedded feature selection methods can address higher order interactions or handle redundancy (Section 2.4.3.1.3).

The final classifier features are determined by repeating the steps 1-6 for each partition, resulting in five ($k = 5$) feature sets. In order to combine these feature sets, a further round of optional filtering and feature reduction method is executed, using the entire train set (Figure 4.3 e). The final classifier model is then trained on a robust set of features, generally valid across the defined data for the corresponding prediction task. Subsequently, the performance evaluation is conducted on the test set (Figure 4.3 e).

In summary, the triad feature selection pipeline, incorporating GeneSTRIVE, provides a robust and scalable solution for handling imbalanced and heterogeneous transcriptomics datasets. The resulting robust feature sets enable accurate classification in high-dimensional, heterogeneous transcriptomics data. This approach has the potential to be applied more widely in biomedical research that could facilitate the realisation of precision medicine in any disease segment.

4.3.5 Gene-expression-based classifiers predicting groups of endotypes

To develop robust gene-expression-based classifiers for distinguishing psoriasis-like endotypes, I utilised the triad feature selection pipeline and optimised model performance. I created two binary classifiers, E12/E13 vs. E8/E11 ($\mathrm{n}_{\mathrm{E12/E13}}$=79, $\mathrm{n}_{\mathrm{E8/E11}}$=46) and E12 vs. E13 ($\mathrm{n}_{\mathrm{E12}}$=48, $\mathrm{n}_{\mathrm{E13}}$=31). Before selecting the final models, I trained multiple classifiers using CV, optimised hyperparameters, and tested various combinations of feature selection techniques (Methods 4.2.3, Figure 4.3 e). In total 18 classifiers were built using MLP, SVM, and XGBoost models, in conjunction with or without VIF and with L1-regularisation, SFS, or RFSCV (Figure 4.4 a-d). These classifiers enable the classification of psoriasis-like endotypes using minimal features, offering a promising foundation for future molecular diagnostics.

4.3.5.1 Best E8/E11 vs. E12/E13 classifiers

I evaluated the CV results of the E12/E13 vs. E8/E11 classifiers. The two top-performing classifiers were selected based the Bayesian correlated t-test probabilities. Both top models were MLPs with L1-regularisation, one incorporating the VIF filtering and the other excluding it (Figure 4.4 a, c, e).

The two top-performing classifiers exhibited comparable performances, as shown by the Bayesian probabilities of $P = 0.43$ and $P = 0.34$. Furthermore, both models had the same median CV score of 94.00 % and standard deviation of 19.00 % (Figure 4.4 c). Notably,

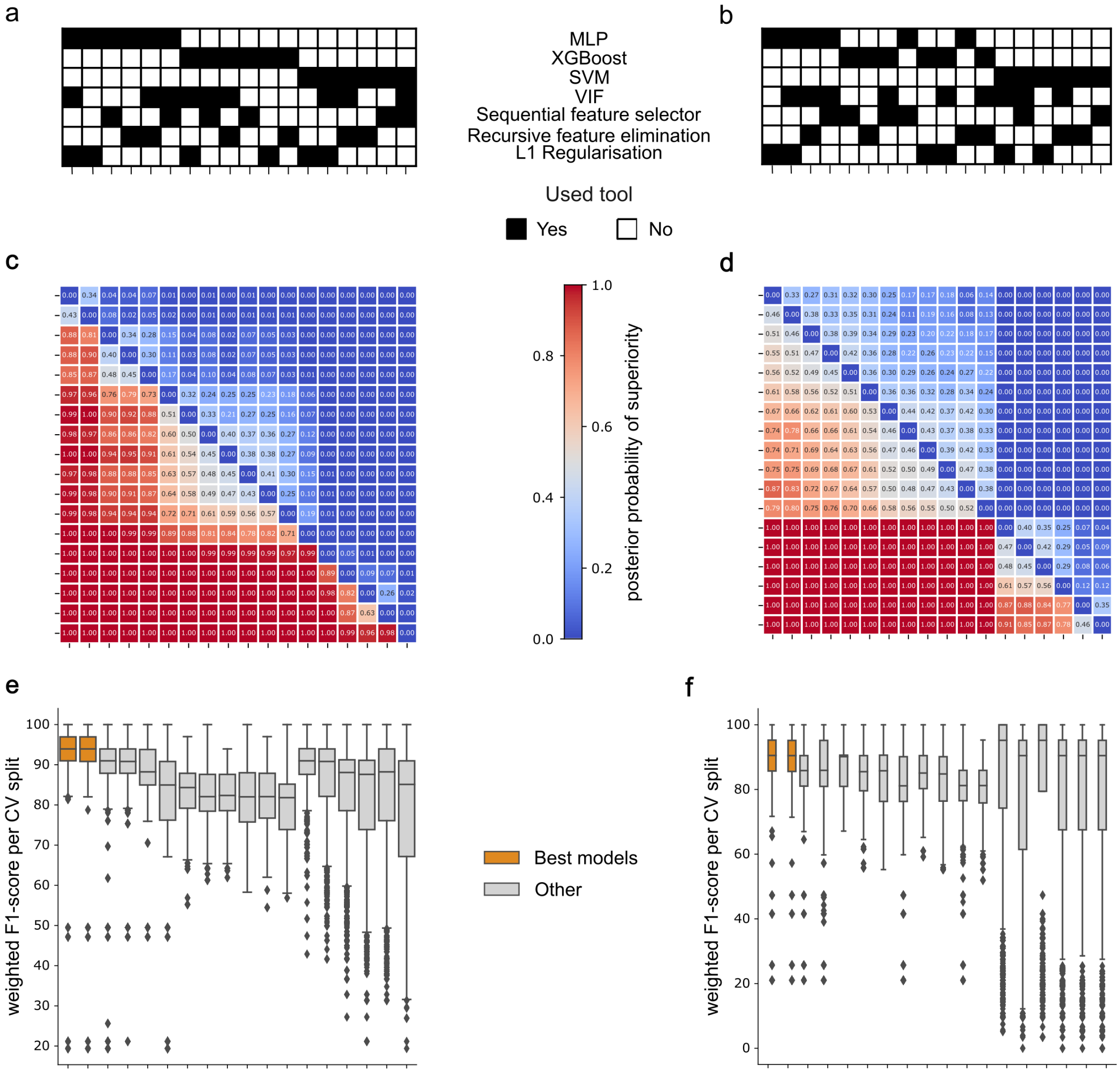


Figure 4.4: Models and CV performance of all E12/E13 vs. E8/E11 and E12 vs. E13 classifiers. a, b) Heatmaps showing the built models with their used feature selection methods for **(a)** E12/E13 vs. E8/E11 and **(b)** E12 vs. E13. The columns are sorted by the Bayesian correlated t-test posterior probabilities between the models. The model, which performs in general better than the others according to the probabilities, is presented first, followed by the second best model, and so on. **c, d)** Heatmaps showing the posterior probability of two models. Values above 0.5 indicate that the model shown on the x-axis is superior in its performance compared to the model on the y-axis. Starting from the top, models on the y-axis are in the same order as the models on the x-axis. **e, f)** Boxplots are sorted by the Bayesian correlated t-test posterior probabilities. The two best performing models are coloured orange. The x-axis of **(c-f)** are aligned to the parameter combinations from **(a)** and **(b)**.

the performance order of the models, given by the Bayesian correlated t-test probabilities, first showed all MLPs, followed by XGBoost models, and finally SVMs. Especially, the SVM models suffered from a high variance. The reasons for this are multifaceted. Given that I scaled the features, considered class imbalance, and observed more robust results in MLP and XGBoost models, a potential cause could be that the provided set of hyperparameters during gridsearch CV was too broad, resulting in a high variance in the validation set performance (Figure 4.4 e). The results suggest that MLP models offer a robust approach for classifying psoriasis-like endotypes that might be associated with drugs inhibiting IL - 23 and IL - 17.

Both top-performing classifiers employed 19 genes associated with specific biological processes. These genes were associated with barrier function, keratinization, inflammation, and immune response such as *CLDN17*, *KRT71*, *IL20*, and *FPR2*, respectively [TTT19] [Wei+06] [LK19] (Figure 4.6 a). Despite similar level of performance, the models exhibited variation in specific features used. Specifically, these were *GPR12*, *KRT71*, *AQP7*, *KRT73*, *ALAS2*, *LEP*, *RBP4*, *TREM1*, *IL1B*, *KRT27*, and *KRT85* (Figure 4.5). In addition, these classifiers shared features with 16 other models (Figure 4.5), with frequently selected genes including *CLDN17*, *IL20*, *SCG2*, *SLC1A6*, *HRNR*, and *TTC6* (Figure 4.5). The top-performing classifiers utilised 19 genes, with overlap in features across other models, showing the features' relevance for classification.

To evaluate feature effectiveness, I identified the MLP with VIF and L1-regularisation as the optimal model, with a Bayesian probability of $P = 0.43$ indicating superiority over the second-best model. I then projected the 19 genes in 2D using UMAP (Figure 4.6 d), which revealed a concentration of training and test set samples of E12/E13 at the bottom and E8/E11 samples at the top. However, the classes were not linearly separable. These findings suggest that while the best-performing model effectively differentiates the classes, the lack of linear separability indicates that the feature set may require refinement or expansion for better classification.

Regarding feature importance in the best model, the genes *KRT71* and *IL20* achieved the highest scores (Figure 4.6 a). To examine class distribution in the training and test sets, I analysed the normed counts of these top two genes (Figure 4.6 e). The degree of overlap between the two classes indicated that more than these features are required for accurate classification. Although the best classifier used 19 features, it remains plausible that a reduced gene set could achieve comparable performance.

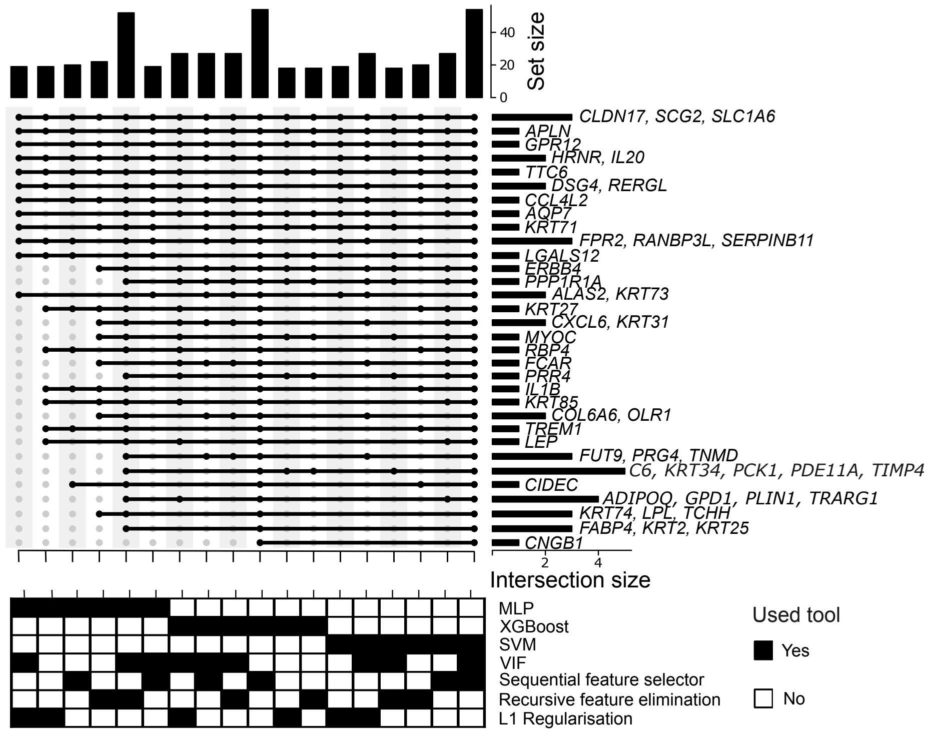


Figure 4.5: Feature sets of all E12/E13 vs. E8/E11 classifiers. UpsetPlot shows the feature set size per model (upper bars) and their intersection size (bars on the right) across the 18 E12/E13 vs. E8/E11 classifiers. Moreover, the black dots indicate, which models use the same specific genes in their feature set. The x-axes are aligned to the parameter combinations shown in the columns of the binary heatmaps below the UpsetPlot. The columns are sorted by the Bayesian correlated t-test posterior probabilities between the models. The classifier that generally performs better than the others in terms of probabilities is presented first, followed by the second best model, and so on.

Assessing the optimal set of hyperparameters for both top performing MLP models, I observed that they either used one hidden layer with 4 nodes or two hidden layers consisting of 10 and 5 nodes (Table 4.6). This was a reasonable choice, given the size of the dataset ($n = 125$) and number of features, as more layers would have led to overfitting.

I calibrated the classifiers using repeated stratified k-fold CV to enhance the reliability of the classifiers. This is particularly of importance when the classifier is applied to samples from unseen classes, as these deviate from the training distribution. The best calibrated E12/E13 vs. E8/E11 model had a log-loss of 0.46 and achieved a weighted accuracy of 82.64 % and a weighted F1-score of 84.00 % (Table 4.7, Figure 4.6 c). The prediction

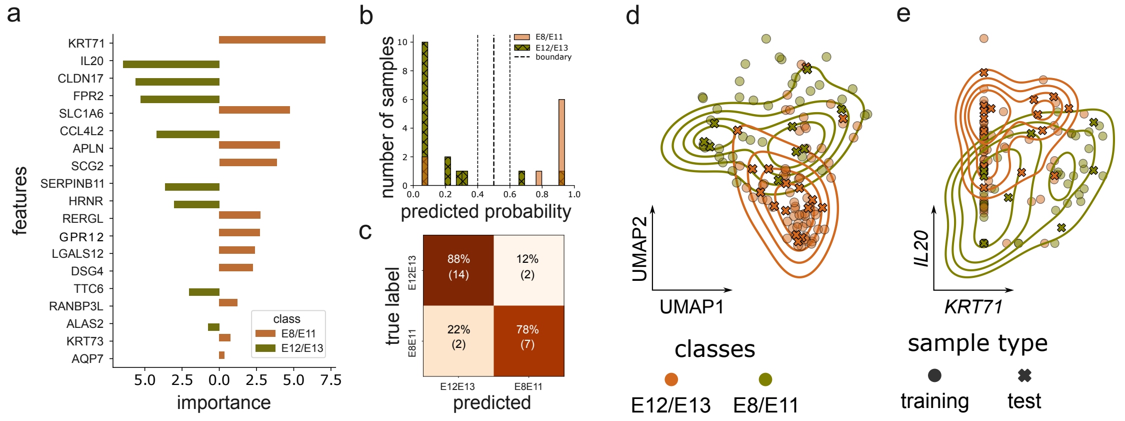


Figure 4.6: Best gene-expression-based classifier for E12/E13 vs. E8/E11. a) Feature importance scores, showing if the feature is either predictive for **(a)** E12/E13 (left, < 0) or E8/E11 (right, > 0). In total 19 features are selected for E12/E13 vs. E8/E11. **b)** Histogram shows the prediction probability of class E8/E11 of the classifier E12/E13 vs. E8/E11. The decision boundary is set to 0.5. **c)** Confusion matrix summarising the performance of the E12/E13 vs. E8/E11 classifier. **d)** UMAP embedding of the classifier features of E12/E13 vs. E8/E11. The contour lines are created using the test set. **e)** 2D representation of the two top features normed counts, showing the training and test set samples. The contour lines are created using the test set. Outermost contour lines in **(d, e)** represent the threshold value of 0.3, i.e. 30 % of the probability mass lies outside of these contour lines. Each contour line marks regions with same density values.

probabilities of all samples were outside the range of $[0.4, 0.6]$, indicating the E12/E13 vs. E8/E11 classifier was confident in its prediction (Figure 4.6 b).

The calibrated second best model (Table 4.7, Figure C.1 a-c) had a log-loss of 0.45. It achieved a weighted accuracy of 85.76 % and a weighted F1-score of 87.83 % on the test set, showing an equal confidence in its prediction as the best performing model. Both models had a higher weighted F1-score, precision, and recall than the baseline model (Table 4.7), predicting the most frequent class. The baseline model did not capture relationships in the data, which confirmed that more advanced models are needed capable of capturing important patterns in the data. The calibrated models performed comparably to the not calibrated ones and yielded more realistic predictions, indicating their potential effective application to other datasets.

In summary, the top two calibrated E12/E13 vs. E8/E11 classifiers performed similar on the CV and test set. Both met the criteria of employing a minimal set of single genes, as they use only 19 features for this task. These genes showed a non-linear separation in the 2D

Model	Parameter	Values
MLP	hidden_layer_sizes	(4,)
	activation	tanh
	alpha	0.0001
	learning_rate_init	0.0001
	sampling_strategy	majority
	VIF	True
	feature selection	L1-Regularisation
MLP	hidden_layer_sizes	(10, 5)
	activation	logistic
	alpha	$1e-05$
	learning_rate_init	0.0001
	sampling_strategy	minority
	VIF	False
	feature selection	L1-Regularisation

Table 4.6: Hyperparameters of the best models for E12/E13 vs. E8/11.

Classifier	Balanced accuracy	Weighted F1-score	Weighted Precision	Weighted Recall	AUC
Baseline model	50.00 %	49.95 %	40.96 %	64.00 %	50.00 %
MLP with VIF and L1-Regularisation	82.64 %	84.00 %	84.00 %	84.00 %	81.94 %
MLP without VIF and L1-Regularisation	85.76 %	87.83 %	87.97 %	88.00 %	82.64 %

Table 4.7: Test set performance of the E12/E13 vs. E8/E11 baseline model and best, calibrated classifiers.

UMAP embedding, explaining the selection of non-linear models for this task. In essence, the classifiers effectively distinguish between the hypothesised drug-target-associated endotype groups E12/E13 and E8/E11, thereby establishing a foundation for similar tasks.

4.3.5.2 Best E12 vs. E13 classifiers

To determine the best E12 vs. E13 models, I conducted a nested CV and compared the performance scores. The overall weighted F1-scores per CV split revealed that, while MLP and XGBoost were more robust overall, the SVM models achieved higher median weighted F1-score per CV split. The increased variance in SVM validation set performance could be attributed to overly broad hyperparameter ranges used during the grid searchCV process, leading to outliers (Figure 4.4 f). In essence, MLP and XGBoost models yielded higher performances and more robust results than SVM models in classifying E12 vs. E13.

I then identified two E12 vs. E13 endotype classifiers that achieved similar performance and outperformed other models (Figure 4.4 b, d; Figure 4.7). The best-performing classifier was an MLP model using no VIF filtering and L1-regularisation for feature selection. This model had a likelihood of $P = 0.46$ of outperforming the second-best model and achieved a median weighted F1-score per CV of 91 % with a standard deviation of 17 %. The second-best model, an MLP model using VIF filtering and L1-regularisation, exhibited a likelihood of $P = 0.33$ of surpassing the best model (Figure 4.4 b, d, f, Table 4.9, Table 4.8). It achieved a median weighted F1-score per CV of 90 % with a standard deviation of 20 %. Thus, both the best and second-best model were MLPs, with performance scores above 90 %.

The best E12 vs. E13 model utilised in total 15 genes, whereas the second-best model used 17 genes (Figure 4.8 a, Figure 4.7, Supplemental Figure C.1 d). Among the genes included in the best model, but not in the second-best, were *CIDEC*, *LORICRIN*, *VTCN1*, and *SLC464A2* (Figure 4.8 a). These genes are involved in cell death and metabolism, epidermal barrier integrity, T-cell activation, and immune response. The second-best model used *FLG*, which has been reported to be downregulated by TNF - α along with *LORICRIN* [Kim+11], and *FAIM2*, which is associated with a form of cell death (Supplemental Figure C.1 d). Both models shared features with the other 16 classifiers developed (Figure 4.7). In essence, the best models relied on features involved in cell death and immune response pathways.

To visually assess class separability of the best classifier, I embedded the 15 genes in 2D using UMAP (Figure 4.8 d). The visualisation revealed near-linear separation of E12 and E13, except for two E13 samples. The distribution of training and test set class samples was further examined using normed counts of the two most important features *GALNT13* and *FCAR* (Figure 4.8 a, e). While some overlap was observed, E12 samples consistently showed higher expression of both features compared to E13 samples. Feature importance analysis (Figure 4.8 a, Supplemental Figure C.1 d) suggested that the number of features used in the top-performing classifiers could be further minimised by eliminating genes exhibiting lower importance. These findings showed the potential for minimising the feature set, enabling the use of simpler, computationally efficient models without compromising performance.

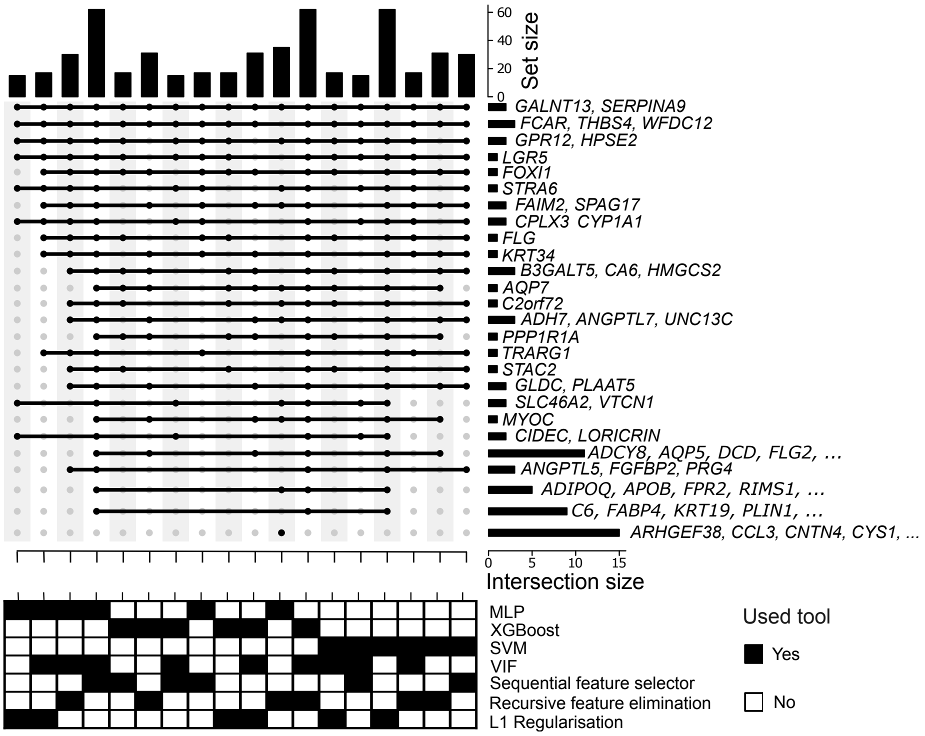


Figure 4.7: Feature sets of all E12 vs. E13 classifiers. UpsetPlot shows the feature set size per model (upper bars) and their intersection size (bars on the right) across the 18 E12 vs. E13 classifiers. Moreover, the black dots indicate, which models use the same specific genes in their feature set. The x-axes are aligned to the parameter combinations shown in the columns of the binary heatmaps below each UpsetPlot. The columns are sorted by the Bayesian correlated t-test posterior probabilities between the models. The classifier that generally performs better than the others in terms of probabilities is presented first, followed by the second best model, and so on.

The optimal hyperparameters (Table 4.8) of the two best E12 vs. E13 models focused on generalisation and on avoiding overfitting. The discrepancy in the number of selected nodes is negligible, with the best model using 6 and the second best mode 7 nodes. This demonstrates again the similarity of the two models. Overall, both models used hyperparameters that allowed them to generalise well.

To enhance the reliability of the classifiers, I calibrated the models using repeated stratified k-fold CV. After calibration, the best and second-best models achieved a log-loss of 0.37 and 0.41, respectively. The classifiers achieved an accuracy of 86.67 % and 83.33 % and a weighted F1-score of 87.50 % and 86.82 % on the test set for the best and

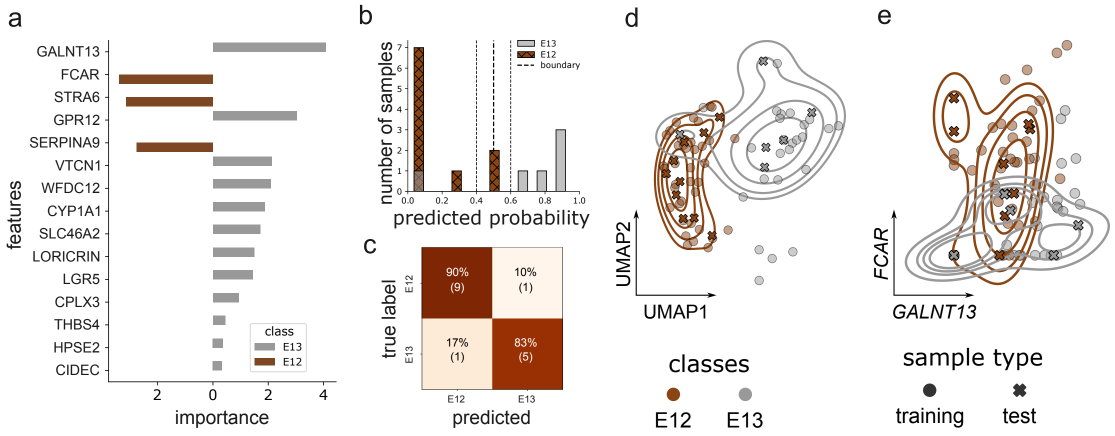


Figure 4.8: Best E12 vs. E13 gene-expression-based classifier. **a)** Feature importance scores, showing if the feature is either predictive for E12 (left, < 0) or E13 (right, > 0). In total 15 features are selected for E12 vs. E13. **b)** Histogram shows the prediction probability of class E13 of classifier E12 vs. E13. The decision boundary is set to 0.5. **c)** Confusion matrix summarising the performance of the E12 vs. E13 classifier. **d)** UMAP embedding of the classifier features. The contour lines are created using the test set. **e)** 2D representation of two top features normed counts of the training and test set samples. The contour lines are created using the test set. Outermost contour lines in **d, e)** represent the threshold value of 0.3, i.e. 30 % of the probability mass lies outside of these contour lines. Each contour line marks regions with same density values.

second best model, respectively (Table 4.9, Figure 4.8 c, Figure C.1 f). Most samples had prediction probabilities outside the defined uncertainty range of $[0.4, 0.6]$ (Figure 4.8 b, Figure C.1 e). In comparison to the baseline model, both calibrated models had a higher weighted F1-score, precision, and recall (Table 4.9). This indicates that advanced models, which account for feature relationships and class imbalance, lead to improved predictive performance. Thus, the calibrated the classifiers demonstrated good performances and surpassed the baseline model in all metrics.

In summary, the top two E12 vs. E13 calibrated classifiers performed similarly in the classification task on the CV and test set. In addition, they met the criteria of employing a minimal set of single genes for this task. The features of the best classifier showed an almost linear separation in the UMAP embedding. Although more data and testing is required, these classifiers represent a promising step towards improved patient stratification and enhanced treatment recommendations.

Model	**Parameter**	**Values**
MLP	hidden_layer_sizes	(6,)
	activation	logistic
	alpha	0.0005
	learning_rate_init	0.0001
	sampling_strategy	majority
	VIF	False
	feature selection	L1-Regularisation
MLP	hidden_layer_sizes	(7,)
	activation	logistic
	alpha	0.0001
	learning_rate_init	0.0001
	sampling_strategy	majority
	VIF	True
	feature selection	L1-Regularisation

Table 4.8: Hyperparameters of the best models for E12 vs. E13.

Classifier	**Balanced accuracy**	**Weighted F1-score**	**Weighted Precision**	**Weighted Recall**	**AUC**
Baseline model	50.00 %	48.08 %	39.06 %	62.50 %	50.00 %
MLP without VIF and L1-Regularisation	86.67 %	87.50 %	87.50 %	87.50 %	90.83 %
MLP with VIF and L1-Regularisation	83.33 %	86.82 %	89.58 %	87.50 %	85.83 %

Table 4.9: Test set performance of baseline model and best-performing, calibrated classifiers of E12 vs. E13.

4.3.6 Application of classifiers on unseen endotypes

In real-world applications the true label is often unavailable, making it necessary to understand a models behaviour on samples from unseen classes and to adapt it accordingly. Furthermore, evaluating a model's performance under these conditions is important for ensuring patient safety, robustness, generalisability to new patients, diagnostic confidence, and therapy efficacy. In order to simulate a realistic scenario, I evaluated the performance of the classifiers using 217 for the E12/E13 vs. E8/E11 classification task and 263 samples for the E12 vs. E13 classifier. These samples, associated with endotype classes E1 - E7 and E9 - E10 (Chapter 3), were excluded from the feature selection and training processes, ensuring they represented truly unseen data. Challenging the classifiers with samples from unseen endotype classes enables to understand the model's behaviour and to identify potential areas for improvement.

The best E12/E13 vs. E8/E11 classifier was tested on samples from unseen endotype classes to evaluate its predictive capability. For the majority of samples, the predicted probabilities were close to one, suggesting similarity to the E8/E11 endotype group (Figure 4.9 a). Endotypes E3 - E7, E9, and E10 exhibited expression patterns characteristic of E8/E11, whereas most samples from E1 and E2 had probabilities close to zero, indicating alignment with E12/E13. Given the classifier's high confidence in its predictions for the majority of samples, the next steps involved investigation potential causes for this behaviour in order to prevent it.

Therefore, I investigated how the training, test, and samples from unseen endotype classes aligned in 2D based on the 19 classifier features (Figure 4.9 b). This allowed me to examine whether the selected features were also expressed in other endotypes and their ability to distinguish unknown classes. I observed that unseen classes significantly overlapped with the regions marked by the training and test set samples of E12/E13 and E8/E11. Interestingly, the unseen classes appeared to form clusters within the E12/E13 and E8/E11 areas (Figure C.3 a-i). Endotypes E1 and E2 clustered within the E12/E13 region, while E3, E5 - E7, E9, and E10 mostly aligned with the E8/E11. Notably, E4 was the only endotype distributed across both classes, indicating shared characteristics. While the 19 classifier features did not separate known from unknown classes, they successfully distinguished E8/E11 and E12/E13 and revealed structured clustering of unseen endotypes, indicating their potential use for future multi-class classification tasks aimed at differentiating individual endotype.

The best E12 vs. E13 classifier was evaluated on unseen endotype classes to assess its predictive performance and feature specificity. Challenging the classifier on unseen classes revealed its predictive certainty for most samples (Figure 4.10 a). Notably, the majority of samples from endotypes E3, E6-E11 were incorrectly classified as E13, with prediction probabilities close to one. In contrast, most E2 samples had a prediction probability close to zero, suggesting that their profile is more similar to E12. Similar to the performance of the E12/E13 vs. E8/E11 classifier, the best E12 vs. E13 classifier showed high confidence in its predictions for the majority of unseen samples.

To investigate this behaviour, the alignment of training, test, and unseen samples was visualised in a 2D space using the 15 classifier features (Figure 4.10 b). This allowed me to

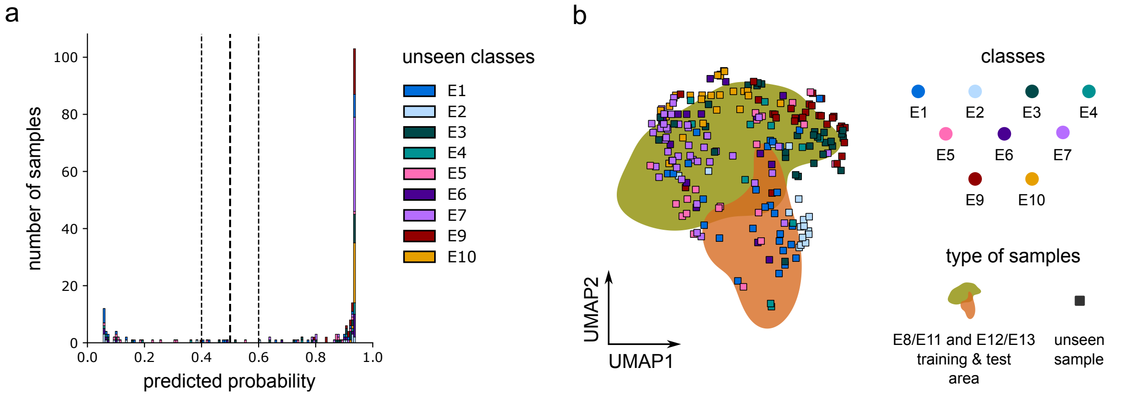


Figure 4.9: Evaluation of best E12/E13 vs. E8/E11 classifier on unseen class samples. a) Probability histogram of best E12/E13 vs. E8/E11 classifier of the samples from the unknown classes. **b)** UMAP embedding of 19 classifier genes showing the area spanned by the training and test set samples of the classes E8/E11 (orange) and E12/E13 (green) and data points of unknown class samples. The area contains 70 % of the probability mass.

understand, whether the selected features are not specific enough to differentiate between the endotypes, causing the high prediction probability for majority of unseen class samples. A similar trend, as shown in the predicted probability histogram (Figure 4.10 a), was visible. The endotype E2 aligned closely with E12, while E3-E9, and E11 aligned with E13 in the 2D UMAP embedding space (Figure C.4 a-i, k). E10 Samples visually separated from both E12 and E13, were still consistently predicted as E13 (Figure C.4 j). These findings show the classifier's limitation to differentiate between unseen class samples and those in the training or test set. Consequently, further efforts are required to incorporate strategies to manage unseen classes, to ensure the classifier's robustness and suitability for clinical applications.

In summary, the results indicated that further analysis is necessary to define more specific feature sets to address the issue of samples from unseen classes in the classifiers. Despite the expected but undesirable behaviour of the classifiers on samples from unseen classes, they were able to effectively distinguish samples belonging to the classes the classifier has seen during the training, by utilising only 19 and 15 features out of the original 17, 816 genes. Further optimisation is required in the feature selection process, model selection, and classification task to ensure the provision of a safe and robust application in the clinics.

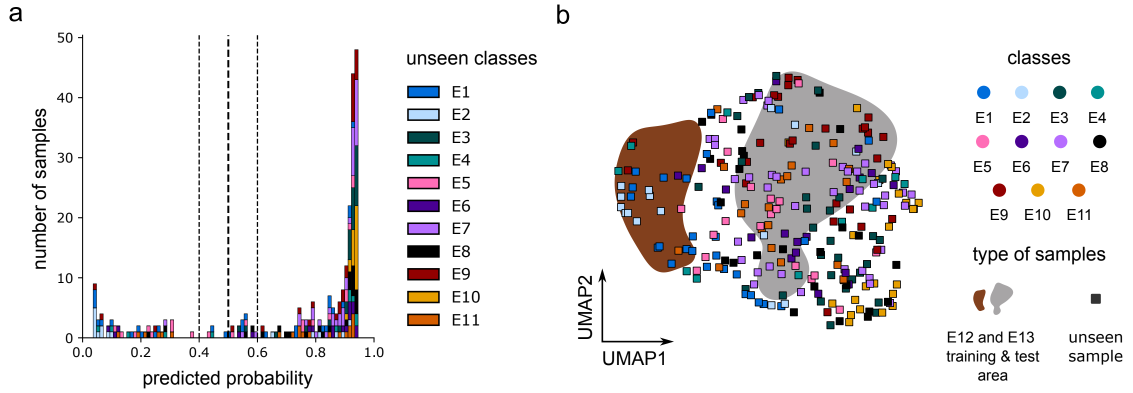


Figure 4.10: Evaluation of best E12 vs. E13 classifier on unseen class samples. **a)** Probability histogram of best E12 vs. E13 classifier of unseen class samples. **b)** UMAP embedding of 15 classifier genes showing the area spanned by the training and test set samples of the classes E12 (brown) and E13 (grey) and data points of samples from unseen classes. The area contains 70 % of the probability mass.

4.3.7 Prediction of endotypes in an independent cohort

Despite being aware of the importance of handling unseen class samples, to ensure robustness and reliability in the real-world applications, I sought to examine the classifiers performance in an independent test cohort, thereby assessing its generalisation ability.

Endotype labels were assigned to an independent psoriasis cohort from Kiel by integrating it with the ncISDs dataset (Materials 4.1.2, Methods 4.2.4). Most Kiel samples aligned with their respective endotype clusters (Figure 4.11 a), with more than half classified as psoriasis-like endotypes E11 (n=2), E12 (n=4), and E13 (n=7), while nine samples have been assigned to other endotypes. A moderate endotype cluster resolution was observed, with visually more overlap, potentially due to gene imputation, number of nearest neighbours ($k = 1$) used for the assignment, or unobserved artefacts. As the findings are based on the 2D UMAP embedding, which provides an approximation of the high-dimensional feature space, they should be interpreted with caution. In summary, the independent cohort showed good alignment with endotype clusters, though some variability and potential artefacts were observed.

The validation of two top-performing classifiers was performed using samples from the Kiel cohort, which have been assigned to the E8, E11-E13 endotype clusters (Figure 4.11 b-e). The E12/E13 vs. E8/E11 and E12 vs. E13 classifiers differentiated well between the specified groups of endotypes, despite the absence of one gene (*CPLX3*) in the E12

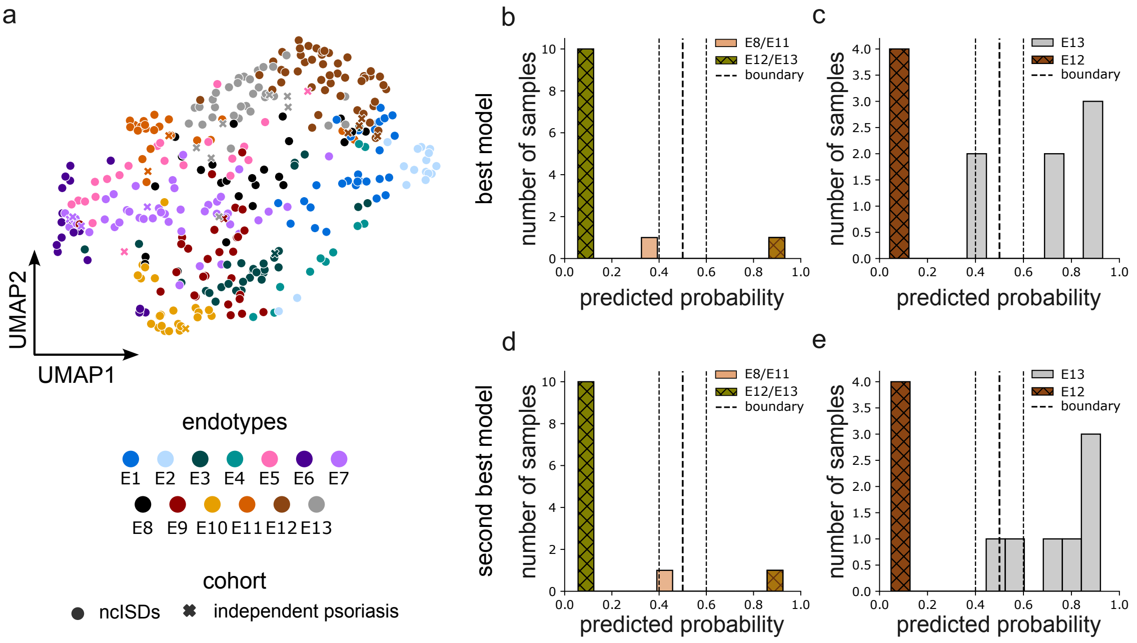


Figure 4.11: E12/E13 vs. E8/E11 and E12 vs. E13 classifier evaluation on independent cohort. a) Joint embedding of the ncISDs dataset and the independent psoriasis cohort from Kiel. Endotypes are transferred to the independent cohort. **b-e)** Best and second best classifiers for the classification task E12/E13 vs. E8/E11 and E12 vs. E13 perform similar.

vs. E13 classifier, which was supplemented using imputed values (Methods 4.2.4). The E12/E13 vs. E8/E11 classifiers achieved a weighted F1-score of 84.62 %, while the E12 vs. E13 classifiers reached a weighted F1-score of 82.12 % and 91.06 % for the best and second best model, respectively. These performance results indicate that the classifiers generalise well to new samples belonging to the same classes as those seen during training.

Further analysis of calibrated log-loss values revealed values of 0.32 for the E12/E13 vs. E8/E11 classifiers and 0.26 and 0.22 for the best and second-best E12 vs. E13 classifiers, respectively. Only a few samples (E12/E13 vs. E8/E11: n=2, E12 vs. E13: n=2) were either misclassified or fell within the uncertainty range of $[0.4, 0.6]$. This indicates that the classifiers confidently assign new samples from known classes to psoriasis-like endotypes, highlighting their robustness.

4.3.8 Assessment of the endotype-therapy-response association

As I hypothesised that certain groups of endotypes might be associated with treatment outcome, I used psoriasis samples with known severity changes over time. Specifically,

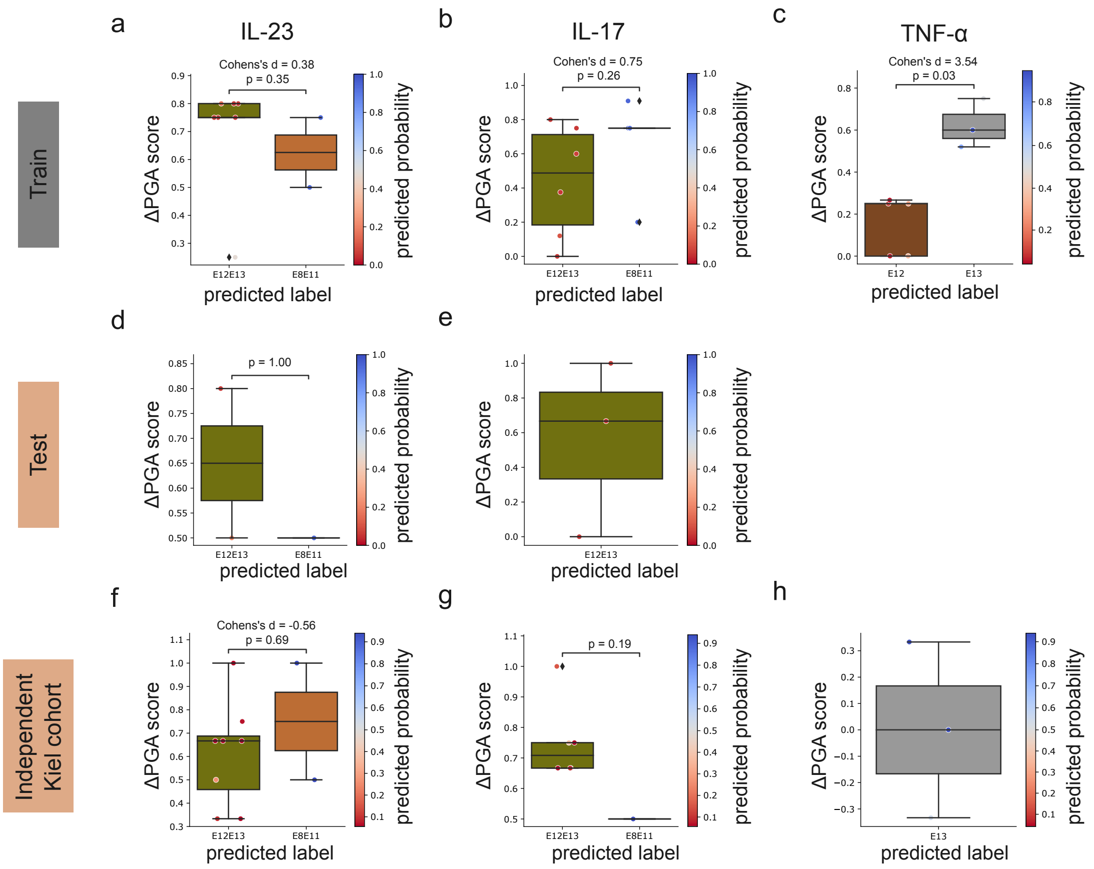


Figure 4.12: Hypothesis evaluation using the best classifiers for E12/E13 vs. E8/E11 and E12 vs. E13. a, b, d-g) Boxplots showing the predicted labels of classifier E12/E13 vs. E8/E11 and the corresponding ΔPGA scores of each patient separated by the drug targets **(a, d, f)** IL - 23 and **(b, e, g)** IL - 17. In total 9, 3, and 10 samples are provided for the training, test, and independent Kiel cohort for drug target IL - 23 and 11, 3, 7 samples are provided for drug target IL - 17. **c, h)** Boxplot showing the predicted label of classifier E12 vs. E13 and the ΔPGA scores of each patient for TNF - α inhibitors. In total 8 and 3 samples of the train set and independent cohort are provided. **(a-c)** show results of the training set, **(d, e)** of the test set, and **(f-g)** of the independent cohort. Statistical testing is performed using the Mann-Whitney-U test.

psoriasis samples with information on the change in severity from week 0 to week 12 were used to evaluate the endotype's association with therapy response (Methods 4.2.1, 4.2.5).

I first tested the hypothesis on the classifiers' training set. As these samples were seen by the E12/E13 vs. E8/E11 (IL - 23 and IL - 17: 20/26) and E12 vs. E13 (TNF - α: 8/8) models, the evaluation was inherently biased (Table 4.2). Considering the classifiers' predicted labels (Figure 4.12 a-c), p-values and effect sizes for the drug targets IL - 23 and IL - 17 changed from p-value = 0.03, Cohen's d = 1.74 to p-value = 0.35, Cohen's d = 0.38, and from p-value = 0.07, Cohen's d = 1.00 to p-value = 0.26, Cohen's d = 0.75, respectively. For the drug target TNF - α, the values remained unchanged. Given that most treatment response samples were used for training, definitive conclusions could not be drawn.

Next, I evaluated the hypothesis on the test set. Predicted labels from the E12/E13 vs. E8/E11 classifier provided mixed evidence regarding treatment response to IL - 23 and IL - 17 inhibitors. The E12/E13 vs. E8/E11 classifier predicted two out of three IL - 23 inhibitor-treated patients as E12/E13 and one as E8/E11. The latter and one predicted E12/E13 patient had a ΔPGA score of 0.5, meeting the lowest limit to be considered as responder (Figure 4.12 d, Table 4.10, Methods 4.2.1). Notably, all IL - 17 inhibitor-treated patients were confidently predicted as E12/E13 (Figure 4.12 e, Table 4.10). This was contrary to my expectation, as I had anticipated to observe higher ΔPGA scores in E8/E11, which would have supported my hypothesis. These findings contradicted my hypothesis about a potential endotype-drug-target association.

The independent cohort analysis yielded inconsistent evidence regarding endotype and therapy response association (Figure 4.12 f-h, Table 4.10, Material 4.1.2, Table 4.3, Methods 4.2.4). For IL - 23 inhibitors, the effect size (Cohen's d = -0.56) in the E8, E11 - E13 samples (n=10) contradicted the hypothesis, exceeding the significance threshold ($\alpha < 0.2$) (Figure 4.12 f, Table 4.10). For IL - 17 inhibitors, most treated samples (n=7) were classified as E12/E13, with a p-value below the threshold (Figure 4.12 g, Table 4.10). Despite moderate to high E12 vs. E13 classifier confidences, solely one TNF - α-treated patient with a negative ΔPGA score of -0.33 fell into the uncertainty range $[0.4, 0.6]$, representing a non-responder (Figure 4.12 h, Table 4.10). The second-best models supported these trends (Figure C.2 a-h, Table C.1). The Kiel cohort provided mixed and contradictory results, challenging the initial hypothesis.

In conclusion, while the hypothesis linking endotypes to therapy response could not be confirmed or completely disproved, results demonstrate inconsistencies across the training, test, and independent cohort. Factors such as limited data and classifier limitations may have influenced findings, requiring further exploration and hypothesis refinement.

Statistic	Drug Target	Class Label	Dataset		
			Training	Test	Kiel Cohort
Median	IL-23	E12/E13	0.75	0.65	0.67
		E8/E11	0.63	0.50	0.75
	IL-17	E12/E13	0.49	0.67	0.71
		E8/E11	0.75	-	0.50
	TNF-α	E12	0.25	-	-
		E13	0.60	-	0.00
Mean	IL-23	E12/E13	0.70	0.65	0.61
		E8/E11	0.63	0.50	0.75
	IL-17	E12/E13	0.44	0.56	0.75
		E8/E11	0.67	-	0.50
	TNF-α	E12	0.15	-	-
		E13	0.62	-	0.00
Min	IL-23	E12/E13	0.25	0.50	0.33
		E8/E11	0.50	0.50	0.50
	IL-17	E12/E13	0.00	0.00	0.67
		E8/E11	0.20	-	0.50
	TNF-α	E12	0.00	-	-
		E13	0.52	-	-0.33
Max	IL-23	E12/E13	0.80	0.80	1.00
		E8/E11	0.75	0.50	1.00
	IL-17	E12/E13	0.80	1.00	1.00
		E8/E11	0.91	-	0.50
	TNF-α	E12	0.27	-	-
		E13	0.75	-	0.33
Q1	IL-23	E12/E13	0.75	0.58	0.46
		E8/E11	0.56	0.50	0.63
	IL-17	E12/E13	0.18	0.33	0.67
		E8/E11	0.75	-	0.50
	TNF-α	E12	0.00	-	-
		E13	0.56	-	-0.17
Q3	IL-23	E12/E13	0.80	0.73	0.69
		E8/E11	0.69	0.50	0.88
	IL-17	E12/E13	0.71	0.83	0.75
		E8/E11	0.75	-	0.50
	TNF-α	E12	0.25	-	-
		E13	0.68	-	0.17

Table 4.10: Boxplot information of training, test, and Kiel cohort datasets for IL-23, IL-17, and TNF-α drug targets for the best classifiers.

In summary, the final gene-expression-based classifiers performed well, achieving weighted F1-scores of 84.00 % (E12/E13 vs. E8/E11) and 86.82 % (E12 vs. E13) on the test set. Dimensionality was effectively reduced from 17, 816 to 19 and 15 genes, respectively, with selected features involved in epithelial differentiation, keratinization, T-cell activation, and epidermal barrier integrity. While both classifiers generalised well to unseen data, their high confidence in predictions for unknown classes raises concerns about diagnostic applicability without prior endotype assumptions. Limited sample sizes restricted conclusions on therapy response. Thus, further validation in larger cohorts is needed to confirm clinical utility and endotype-therapy-response associations.

4.4 Summary and discussion

This chapter examined the potential association of psoriasis-like endotypes with drug response and identified predictive features for their classification Current diagnostic and therapy suggestions for psoriasis rely on clinical profiles, which can lead to misdiagnosis or reduced therapy efficacy. To address this challenge, I utilised therapy response information from 34 psoriasis patients and integrated these into the hypothesis-free derived endotypes identified in Chapter 3. This enabled the exploration of associations between psoriasis-like endotypes and drug responses (objective (iv)). Predictive features for these specified endotype groups were identified leveraging GeneSTRIVE, incorporated into a triad feature selection pipeline. Based on these features, binary gene-expression-based classifiers were constructed, predicting the psoriasis-like endotype groups (objective (v)). In summary, this chapter aimed to enhance psoriasis diagnosis and therapy response by classifying psoriasis-like endotypes.

The analysis revealed groups of psoriasis-like endotypes that are potentially linked to therapy response. Specifically, E8/E11 was significantly associated with positive response to the drug target IL - 17, E12/E13 with IL - 23, and E13 with TNF - α. The best classifiers utilised only 19 and 15 genes from the original 17, 816 genes to differentiate between E12/E13 vs. E8/E11 and E12 vs. E13, respectively. The hypothesis was evaluated in an independent cohort, but the small sample sizes for the treatment response groups (IL - 23: n=10, IL - 17: n=7, TNF - α: n=3) led to results that contradicted the previously identified associations. The limited statistical power and inconsistent results prevented conclusive validation of the hypothesised associations. While this study identified potential biomarkers and introduced gene-expression-based classifiers, larger cohorts are needed to explore the relationship between endotypes and therapy response in order to enable tailored treatment strategies for patients.

I investigated the relationship between psoriasis-like endotypes and their responses to drugs inhibiting specific drug targets. Significant associations were observed between the drug targets IL - 23, IL - 17, TNF - α and the psoriasis-like endotypes E8, E11 - E13. In particular, E8/E11 was associated with a positive drug response targeting IL - 17, E12/E13 with IL - 23, and E13 with TNF - α.

To understand the biological rationale underlying these associations, I analysed the molecular characteristics of the defined endotype groups. In E12/E13, higher expression of type 3 immune response associated genes and activation of protein synthesis related pathways were observed. E8/E11 displayed increased activity in keratinization and extracellular matrix (ECM) organisation, potentially explaining drug response differences. Additionally, E13 showed more active protein synthesis pathways and upregulation of LCE family genes compared to E12. Notably, E12, resembling non-responders to TNF - α inhibitors, exhibited an enrichment of immune response and cellular signalling pathways. These findings suggest that distinct immune and cellular processes contribute to variation in drug response across psoriasis-like endotypes and provide a foundation for endotype-specific therapeutic strategies to improve treatment outcomes.

In order to identify potential biomarkers for endotypes hypothesised to be associated with drug response, I developed GeneSTRIVE, a gene selection method for noisy and heterogeneous transcriptomics data. GeneSTRIVE enhances the robustness, generalisability, and biological relevance of DEGs by integrating subsampling, repeated DGE analysis, and contrastive comparisons between L and NL samples. In comparison the other methods using L samples, GeneSTRIVE detects systemic disease-associated changes rather than lesion-specific variations, ensuring more biologically meaningful features. By incorporating a user-defined design matrix, GeneSTRIVE accounts for class labels and covariates, supporting complex study designs and multi-class classification. This adaptability of GeneSTRIVE demonstrates its potential for biomarker discover and improved classification interpretability.

Compared to other feature selection techniques, GeneSTRIVE effectively addresses challenges inherent to high-dimensional transcriptomics data, such as the "curse of dimensionality" and noise [Pud+22] [HK99]. While tree-based ML models can handle high-dimensional data, they are prone to overfitting, and studies recommend reducing the feature space before applying them [Pud+22] [Szy+09] [SKZ10]. GeneSTRIVE mitigates these issues through repeated subsampling, ensuring the stability, generalisability, and robustness of selected genes across multiple iterations. In summary, GeneSTRIVE's emphasis on robust single-gene selection to improve patient stratification, and advance precision medicine in ncISDs and other diseases.

I developed a triad feature selection pipeline capable of reducing the dimensionality of transcriptomics data from $17,816$ to a minimal set of single genes, offering advantages over hybrid methods. It uses GeneSTRIVE for robust gene selection, followed by an optional filtering step. The pipeline further refines the output by incorporating Random Forest or linear regression models to capture gene interactions [CPB18] [Ook+21]. The pipeline also handles imbalanced and heterogeneous transcriptomics datasets, regardless of size or number of classes, ensuring broad applicability. In comparison to single-method approaches, such as Random Forest and RFSCV [Ser+22], which are prone to overfitting and struggle with redundancy in high-dimensional data [Pud+22] [CS14] [BC+14] [BG+17], my pipeline overcomes these challenges. It has the potential to advance the identification of single predictive genes and improve classification accuracy.

To differentiate patients responding to IL-23 inhibitors from those responding to IL-17 inhibitors, I developed classifiers distinguishing E12/E13 from E8/E11. The best models, MLPs, achieved weighted F1-scores of 84.00 % and 87.83 % for the best and second best model, respectively. The selected feature set included 19 genes, involved in immune response, GPCR downstream signalling, and keratinisation pathways. Amongst these were genes such as *KRT71*, *IL20*, and *CLDN17*, exhibiting the greatest impact on model predictions, showing up to seven times higher influence compared to the genes with least influence on the prediction such as *AQP7*, *KRT73*, and *ALAS2*. These findings suggest that further feature reduction could be explored by training classifiers focusing on the most predictive genes, potentially without compromising performance. In summary, in total 19 genes, involved immune response and keratinisation, were used by the best E12/E13 vs. E8/E11 models, which could be further minimised without compromising performance and using less computational resources.

In order to distinguish the hypothesised non-responders from responders to TNF-α inhibitors, I created E12 vs. E13 classifiers. The best-performing models, based on 15 and 17 genes, were MLPs, achieving a weighted F1-score of 87.50 % and 86.82 %, respectively. The most overall predictive genes included *GALNT13*, *FCAR*, and *STRA6*. Another predictive feature was *LGR5* for E13, which is known to be upregulated in various cancer types, including basal cell carcinomas [Tan+08] [McC+06]. The low importance scores of certain features suggest that the classifier could be further refined by focusing on the most predictive genes without compromising performance. In summary, the MLP models successfully differentiated between E12 and E13, with evidence suggesting that the feature set could be further reduced while maintaining predictive power.

The performance of my classifiers on samples from unknown endotype classes was assessed to evaluate their robustness and real-world applicability. Although these samples shared some characteristics with the training data, the analysis revealed false positives (FPs) and false negatives (FNs) due to insufficient feature specificity. Despite this limitation, the classifiers effectively differentiated between training and test samples, demonstrating GeneSTRIVE's ability to identify genes for the defined endotype groups. In summary, while the classifiers performed well on known data, their ability to generalise to unseen classes requires refinement to ensure clinical applicability and patient safety.

I integrated the independent psoriasis cohort (n=22) from Kiel with the ncISDs dataset and subsequently, endotype labels via nearest neighbour classification using HVGs were assigned. The 2D manifold confirmed alignment with the original endotype clusters, though some samples were assigned to endotypes beyond the psoriasis-like groups. Thus, future work should prioritise the development of a diagnostic tool capable of distinguishing all endotypes rather than focusing on specific subsets.

To validate my classifiers, I tested them on samples from the independent psoriasis cohort assigned to the psoriasis-like endotype clusters. The top-performing models, using minimal feature sets of 19 and 15 genes, demonstrated good generalisability (E12/E13 vs. E8/E11: weighted F1-score = 84.62 %; E12 vs. E13: weighted F1-score = 82.12 %) when tested on samples assigned to these endotypes. Notably, despite the absence of *CPLX3* in the Kiel cohort, the E12 vs. E13 classifier maintained high confidence, likely due to the gene's low importance. This supports refining the feature set to retain only the most predictive genes. In conclusion, both classifiers demonstrated generalisability across independent datasets, which also reinforces the relevance of the selected features as potential biomarkers.

I investigated a potential association between psoriasis-like endotypes and drug targets IL-23, IL-17, and TNF-α, using the predicted endotypes in an independent psoriasis cohort. Despite these efforts, the hypothesis could neither be confirmed nor entirely disproved, possibly due to limitations such as small sample size, the elevated significance level $\alpha = 0.2$ used to draw the conclusion in the original data, and reliance on transcriptomics data alone, which may not fully capture the underlying biological complexity. Additionally, imputing missing genes using the mean expression of four HKGs may have influenced classification results. While p-values did not reach significance, a trend towards rejecting the hypothesis was observed, suggesting the need for further investigation. In essence, larger datasets and integrative multi-omics approaches may be required to determine whether psoriasis-like endotypes are associated with drug response.

While GeneSTRIVE demonstrated good performance in handling noisy, high-dimensional transcriptomics data, certain limitations should be addressed to further enhance its utility and generalisability. Currently, GeneSTRIVE is designed for (pseudo-) bulk RNA-seq data and does not support single-cell RNA-sequencing (scRNA-seq) data without cell aggregation. Implementing glmGamPoi [AEH20] instead of edgeR [RMS10] could resolve this constraint. Additionally, GeneSTRIVE assumes similar NL skin expression profiles across diseases. This assumptions is not applicable in conditions such as psoriasis, where the entire skin is affected [Air+15] [Gud+09] [Nos+21]. While this does not impact the current study, it may affect future applications involving diseases with differing NL profiles. Further evaluation is needed to show GeneSTRIVE's performance in multi-class classification tasks. In summary, addressing these limitations and expanding compatibility to scRNA-seq will enhance GeneSTRIVE's utility and applicability in diverse settings.

Optimising hyperparameter spaces is crucial for improving the performance of binary classifiers, particularly to address dataset size limitations and heterogeneity, thus enhancing generalisability. Simplified models, such as logistic regression, alongside data augmentation techniques, can mitigate sample size constraints [Lu+24]. Additionally, Piccolo et al. highlighted that classifier performance is highly dependent on both the model and the performance metric used [Pic+22]. Future efforts should focus on refining these aspects to further improve model robustness and adaptability across varied datasets.

A notable limitation is the classifiers' inability to handle samples from unknown classes. Since psoriasis-like endotypes only represent a subset of psoriasis patients, it is important to account for samples from unseen categories to ensure generalisability and prevent FPs. Strategies to address this include anomaly detection, models for distinguishing known and unknown classes, and ensemble approaches [Yan+24] [Sel+21]. Alternatively, classifiers could assign low confidence to unknown samples or incorporate additional classes representing them [Hsu+20]. Another option is a hierarchical classification framework, starting with the identification of psoriasis-like endotypes, followed by further classification within these categories. Alternatively, a multi-class classifier for all 13 endotypes could be created. Incorporating strategies to handle unknown classes would enhance classifier reliability and clinical applicability.

The hypothesis, based on psoriasis patients assigned to endotypes E8, E11-E13, requires further validation due to the small sample size for the drug targets IL - 23 (n=12), IL - 17 (n=14), and TNF - α (n=14). Increasing the number of samples from both psoriasis patients and other ncISDs would provide a more comprehensive understanding and improve confidence in refining or rejecting the hypothesis. Further, the hypothesis was not supported in the independent psoriasis cohort from Kiel, potentially due to limited sample size (n=20), gene imputation, and methodological differences in read alignment and counting. While 95 missing genes were imputed in the independent cohort, a good alignment between datasets was observed, suggesting that imputation may not have been the primary limiting factor. In essence, larger, more diverse ncISDs datasets would enhance statistical power and robustness, providing clearer insights into the potential association between endotypes and drug response.

This study introduces a feature selection pipeline capable of identifying less than 20 robust features to classify psoriasis-like endotypes. The gene-expression-based classifiers demonstrated strong performance in distinguishing samples within the defined classes, with validation on an independent cohort confirming their efficacy. These findings demonstrate the pipeline's potential of identifying potential biomarkers for patient stratification. While the results are promising, improving the handling of samples from unknown classes remains part of future work to enhance the classifiers' reliability and generalisability. Although the hypothesis associating specific endotype groups to drug response could neither be confirmed nor rejected, the study provides tools for future research. Further efforts should focus on refining these classifiers and expanding their applicability to larger and more diverse datasets, thereby advancing precision medicine for psoriasis and related conditions.

Chapter 5

Spatial transcriptomics landscape of lesions of non-communicable, inflammatory skin diseases

In the past decade, studies of non-communicable chronic inflammatory skin diseases (ncISDs) enhanced disease categorisation [EE18], identified biomarkers [Gar+16], and advanced the understanding of molecular mechanisms [SKK20] on the population level using next generation sequencing (NGS). Yet, effective treatment options are lacking due to the incomplete understanding of their underlying pathogenesis on more granular resolutions. The reason is that ncISDs are heterogeneous and complex diseases, characterised by various biological mechanisms varying based on tissue structure and cellular interactions happening in the micro-environment. To address these challenges, more refined approaches are needed to study skin diseases at higher resolutions and explore cellular interactions within the tissue. The integration of high-resolution technologies is essential for uncovering cellular mechanisms that can inform the development of more targeted and effective therapies.

By preserving spatial information, spatial transcriptomics (ST) enables the exploration of gene expression in complex tissues, providing deeper insights into cellular interactions in diseases. Transcriptomics technologies, such as bulk and single-cell RNA-sequencing (scRNA-seq), offer insights into the biological functions and cell types involved in ncISDs. These methods require complete tissue dissociation, leading to the loss of spatial information. A new technology, Visium by 10x Genomics, combines ST and imaging to explore the whole transcriptome on a spatial resolution within the tissue. This provides an advantage over bulk RNA-sequencing (RNA-seq), scRNA-seq, and in situ hybridisation (ISH) technology [Wil+22] [Wan+12]. While Visium does not achieve single-cell resolution, it provides spatially resolved transcriptomes. This offers information about the location of cellular activities within the tissue section.

Understanding the spatial distribution of cells in ncISDs can reveal how different immune responses manifest in tissue architecture. The most prevalent and understood ncISDs are lichen planus, eczema, and psoriasis. Each has a distinct immune response, with lichen planus being driven by type 1, eczema by type 2, and psoriasis by type 3 immune response pattern (IRP) [Mar+18]. The characteristic hallmark cytokines are *IFNG*, *IL13*, and *IL17A* for lichen planus, eczema, and psoriasis, respectively [Mar+18]. Investigating these diseases on a spatial resolution elucidates tissue associated cellular interactions and may assist in identifying potential drug targets.

In this study, I characterise disease-driving leukocytes and investigate their impact on the direct micro-environment using ST of 31 lesional (L) and non-lesional (NL) skin lesions from lichen planus, eczema, and psoriasis patients. Only low numbers of hallmark cytokine transcripts were detected. Yet, they were capable of triggering tremendous immune response cascades in their local micro-environment. In conclusion, only a few leukocytes were found to promote inflammation in the skin lesions of patients with ncISDs.

This Chapter addresses the deliverable "Leveraging ST to advance the understanding of the underlying disease mechanisms" and its associated research questions, as outlined in Section 1.6:

1. Objective (vi), investigating the expression distribution of hallmark cytokines of the most common ncISDs using ST (Section 5.3.1). I leverage a ST dataset and compared the number of transcripts and expression pattern against established methods such as bulk RNA-seq, scRNA-seq, and ISH.

2. Objective (vii), characterising the micro-environment of cytokine transcript-positive spots (Section 5.3.2). Specifically, I investigate the differential gene expression (DGE) signatures of these spots and compare them to the expected outcome using scRNA-seq data. In addition, the cell type composition of cytokine transcript-positive spots is investigated.

3. Objective (viii), exploring the spatial inflammation pattern and investigating the impact radius of cytokines in the epidermis (Section 5.3.3). I identify the induced immune response signature of the hallmark cytokines. Furthermore, I develop a density-based clustering algorithm that assisted in identifying immune hotspots in the skin initiated by the hallmark cytokines.

The study presented in this Chapter is based on, and thus partly identical to, the following publication [Sch+22] and preprint on bioRxiv [Sch+21]:

A. Schäbitz*, **C. Hillig***, M. Mubarak, M. Jargosch, A. Farnoud, E. Scala, N. Kurzen, A. C. Pilz, N. Bhalla, J. Thomas, M. Stahle, T. Biedermann, C. B. Schmidt-Weber, F. Theis, N. Garzorz-Stark, K. Eyerich*, M. P. Menden* & S. Eyerich*†. "Spatial transcriptomics landscape of lesions from non-communicable inflammatory skin diseases." *Nature Communications* 13.1 (2022): 7729. DOI: https://doi.org/10.1038/s41467-022-35319-w.
The code is available on GitHub:
https://github.com/Chillig/ST_biostatistical_analysis [HFM22].

My contributions encompassed the preprocessing and analysis of the ST, scRNA-seq, and bulk RNA-seq data. In addition, I designed a method for determining a cytokine's radius of action, which was conceptualised in collaboration with Ali Farnoud and Michael Menden. I was responsible for implementing this method, which was published alongside with the analysis code. All raw versions of the plots, showing the analysis of the ST, scRNA-seq, and bulk RNA-seq data, were created by me, and the manuscript versions were further adjusted by all authors.

5.1 Materials

In order to address the research questions, a ST dataset was generated. In collaboration with Alexander Schäbitz, Natalie Garzorz-Stark, and Emanuele Scala from Karolinska Institute and Menatullah Mubarak, Manja Jargosch, Stefanie Eyerich from ZAUM multiple datasets were created. The ST cohort was published together with the manuscript "Spatial transcriptomics landscape of lesions from non-communicable inflammatory skin diseases" by Schäbitz & Hillig et al. (2022), in Nature Communications [Sch+22]. An overview of all utilised datasets is shown in the Table 5.1.

5.2 Methods

5.2.1 Spatial transcriptomics dataset

To generate the ST dataset (Tables 5.1 and 5.2), fresh frozen skin biopsies ($6\,mm$ and $4\,mm$) from 31 patients were sectioned into $10\,\mu m$ thick slices and placed on the capture areas of Visium Gene Expression object slides. A replicate was created from each sample. Prior to sequencing, slides were stained using hematoxylin and eosin (H&E) staining and imaged using scanning microscopes. Sequencing was performed on the Illumina NovaSeq6000 sequencer, yielding $71,606$ transcriptomes. Read alignment, conducted by

*Contributed equally
†Corresponding author

Affiliation	Samples	Technique	images (number)
Karolinska Institute	15 patients, 58 samples	ST	H&E (58)
ZAUM	16 patients, 32 samples	ST	H&E (16)
ZAUM	1 patients, 2 samples	scRNA-seq	- -
Derma TUM	330 samples	bulk RNA-seq	-
Karolinska Institute	6 samples	ISH	fluorescence (3)
Karolinska Institute	20 samples	IHC	brightfield (1)
ZAUM	52 samples	flowcytometry	-
ZAUM	-	in vitro analysis	-

Table 5.1: Overview of data set generated by Alexander Schäbitz, Menatullah Mubarak, Manja Jargosch, Natalie Garzorz-Stark, and Emanuele Scala.

Thomas Walzthoeni from the Helmholtz Bioinformatics Corefacility, utilised 10x Visium Space Ranger-1.0.0 [Zhe+17] and the GRCh38 reference genome. Afterwards, the quality of the 90 processed samples was assessed.

The spots were annotated in a blinded manner using the 10x Genomics Loupe Browser by the collaborating dermatopathologists and biologists Alexander Schäbitz, Kilian Eyerich, Manja Jargosch, and Stefanie Eyerich from the Karolinska Institute, University of Freiburg, and ZAUM. The epidermis was segmented into upper, middle, and basal layers, while the dermis was divided into seven depths (1-7), with depth 7 being the deepest and depth 1 the most superficial. Additionally, the dermal-epidermal junction was identified as a distinct layer between the basal epidermis and dermis [BWJ75].

The data preprocessing was conducted using SCANPY [WAT18], if not explicitly stated otherwise. After read processing, 90 samples comprised of $71,606$ spots and $20,613$ genes were subjected to Quality control (QC). From these 90 samples, 8 demonstrated a general low quality. The final QC was applied on 82 samples with in total $62,968$ spots. Since I was investigating ncISDs, an acute inflammation was expected, elevating the stress level in

Diagnosis	**Sex**	**No. samples**	**Age**	**Severity**
psoriasis	m	7	41.14 ± 14.74	PASI: 11.76 ± 4.97
	w	4	56.75 ± 9.54	PASI: 9.95 ± 3.74
eczema	m	8	49.38 ± 7.48	EASI(n=4): 22.35 ± 11.75 SCORAD(n=2): 58 ± 9
	w	1	21	EASI: 56
lichen planus	m	6	40.67 ± 10.39	-
	w	5	56.6 ± 7.8	-

Table 5.2: Overview of patient characteristics of the ST dataset. EASI is short for Eczema Area and Severity Index and is another severity score of eczema [Les+15]. Table adapted from [Sch+22].

the cells. Therefore, I applied a more conservative mitochondrial (MT)-fraction cut-off of $> 25\%$. Additionally, a spot was required to have at least 30 genes and unique molecular identifier (UMI)-counts between 50 and $500,000$. Further, genes had to be measured in at least 20 spots. After QC, the dataset was comprised of $59,319$ spots and $16,685$ genes from 82 samples.

The data was normalised by size factors using scran [LMM16] and subsequently $\log_{10}$ transformed after adding a pseudo count of one to each gene to avoid log-transformation of zero [LT19]. In total $4,000$ highly variable genes (HVGs) were selected, which were defined as HVGs in all specimens, using the `highly_variable_gene` function with flavor cellranger from SCANPY [WAT18].

I assessed potential confounding factors, such as specimen, capture area, batch, and sequencing project, by fitting a linear regression model on all principal components (PCs) individually, with the confounding factor in question as the dependent variable. This approach is described in detail in Büttner et al. (2019) [Büt+19]. The variable *project* yielded the highest variance in the data. Hence, I applied batch correction to correct for effects introduced by *project* using Scanorama [HBB19]. After batch correction, the top 15 PCs were identified and a K-nearest neighbour (KNN) graph was built. Based on the KNN graph, the data was embedded in a 2D manifold using Uniform Manifold Approximation and Projection (UMAP).

The filtered, raw UMI-counts of the cytokines *IFNG*, *IL13*, *IL17A*, and housekeeping gene (HKG) *GAPDH* were tested for higher expression levels in predefined tissue layers. First, the data was tested for normality using the Shapiro test implemented in the

Python package Scikit-learn [Ped+11], yielding in non-normality for all data distributions. Therefore, the Wilcoxon signed-rank test, also available in the package Scikit-learn, was used to test for significant differences between expression levels in specific tissue layers. I tested for the null hypothesis $\mathcal{H}_0 : \mu = \mu_0$, which is rejected when the p-value is below the significance level $\alpha = 0.05$.

In order to characterise disease-promoting cells in comparison to bystander cells, leukocytes-positive spots were first identified based on the expression of marker genes including *CD2*, *CD3D*, *CD3E*, *CD3G*, *CD247* (*CD3Z*), and *PTPRC* (*CD45*). These markers were required to be expressed with a minimum UMI-count of one within a spot, either individually or in any combination. Disease-promoting positive spots, were defined by the expression of at least one cytokine (*IFNG*, *IL13*, or *IL17A*) with a minimum UMI-count of one. DGE analysis was performed to identify genes characteristic of *IFNG*, *IL13*, or *IL17A* leukocyte-positive spots in comparison to leukocyte-negative spots. The raw, filtered UMI-counts and the size factors, which were calculated on the entire dataset, were provided to the R library glmGamPoi [AEH20]. Additional biological and technical effects, such as cellular detection rate (cdr), sequencing project, patient variability, and skin layer annotations, were incorporated into the model design, which was expressed as

$$y_{sg} \sim \text{cdr} + \text{project} + \text{patient} + \text{annotation} + \text{condition} \;, \tag{5.1}$$

where y_{sg} is the raw count of gene g in a spot s, and the condition variable differentiates between cytokine-positive and cytokine-negative leukocyte spots. The comparison was performed at the spot level. To account for multiple testing, p-values were adjusted for false discovery rate (FDR) using Benjamini and Hochberg (BH). A gene was considered significantly differentially expressed if it met the criteria of padj ≤ 0.05 and $|\text{log2FC}| \geq 1$.

Before over representation analysis (ORA) was performed, the entrezIDs were assigned to each gene, using the database org.Hs.eg.db [Car+19]. The significantly differentially expressed genes (DEGs) were subjected to the Bioconductor [Gen+04] R library ReactomePA [YH16], using all measured genes in the experiment as background. P-values were corrected for FDR using BH. A pathway was considered enriched if its padj value was below 0.05. For visualisation purposes, the R library enrichplot [Yu21] was used.

In order to determine cytokine responder gene signatures, Stefanie Eyerich stimulated for 16 h primary human keratinocytes in vitro with recombinant *IFNG*, *IL13*, or *IL17A* (Figure C.5 a). Afterwards, she performed whole genome expression arrays on the isolated RNA and subsequently, identified DEGs by requiring a padj value of $\alpha < 0.05$ and log2FC $>$ 1.5 for *IFNG* and *IL17A*, and a log2FC $>$ 1 for *IL13*. In order to build robust responder gene signature lists, the intersection was built from the in vitro and computationally derived DEGs. The latter were identified from the comparison of cytokine-positive leukocyte spots against cytokine-negative leukocyte spots, as described in the previous paragraph. Shared responder gene signatures between two or more cytokines were excluded from the list. The final list is comprised of 29, 4, and 21 responder gene signatures of the cytokines *IFNG*, *IL13*, and *IL17A*, respectively (Figure C.5 b-d). As the *IL13* responder gene list only included four genes, additional genes were added from the literature.

The spot deconvolution tool Tangram [Bia+21] was used to determine the proportions of cell types in each spot (see section 2.5.3.2). The publicly available dataset of Reynolds et al. (2021) [Rey+21] was used as reference. To align the diagnoses in the public dataset with those in the ST dataset, the L and NL samples from psoriasis and eczema patients were extracted. Tangram's provided model was configured with the mode set to "clusters", referring to the tissue layers in the ST dataset. The cluster_label parameter was defined as "cell types" and the density_prior was set to "rna count based". The model was trained on a CPU for a maximum of 500 epochs. Afterwards, the 42 cell types were summarised into the four coarse groups antigen presenting cells (APCs), lymphoid/mast cell, epidermal non-immune cell, and dermal non-immune cell. The APCs group contained Dendritic cells (DCs) and macrophages (Macs), while the lymphoid/mast cell group is comprised of T-cells, natural killer cells (NKs), innate lymphoid cells (ILCs), mast cells, and plasma cells. The epidermal and dermal non-immune cells included keratinocytes (KCs) and melanocytes for the former, and Schwann, pericyte, vascular endothelium (VE), lymphatic endothelium (LE) cells for the latter.

For the pseudo-bulk correlation analysis, two pseudo-bulk samples were build per specimen, containing the aggregated UMI-counts of each cytokine and the cytokines' corresponding responder signature genes. Additionally, the total number of cytokine transcript-positive spots was read out. A cytokine transcript-positive spot is defined by measuring at least one UMI-count of the respective cytokine. Subsequently, the weighted Spearman's correlation coefficient (SCC) was calculated between cytokine and responder transcripts using as weights the number of cytokines transcript-positive spots in the whole cohort.

5.2.2 Preparation of the single-cell RNA-sequencing data

The scRNA-seq data was generated from a single psoriasis patient. Sample preparation in the wet-lab was conducted by the collaborating former PhD student, Menatullah

Mubarak, from ZAUM. Two samples were derived from the donated L skin biopsy ($6\,mm$). The samples were lysed and stained for CD3 and CD45. The stained single cells were subsequently subjected to fluorescence activated cell sorting (FACS) to remove dead cells and doublets. Additionally, the cells were gated based on size and the expression of CD3 and CD45, a step undertaken to enrich the proportion of immune cells. The library was prepared using an equal ratio of keratinocytes, T-cells (CD45+, CD3+), and APCs (CD45+, CD3-). The Chromium Next GEM SingleCell 3' GEM Kit v3.1 was utilised for library preparation, and sequencing was performed on an Illumina HiSeq4000 platform. Further details are available in Schäbitz & Hillig et al. (2022) [Sch+22].

The read alignment of the two scRNA-seq samples was performed using an established pipeline of the Helmholtz Bioinformatics Core Facility, which utilises CellRanger v1.0.0. The preprocessing steps were carried out by me using SCANPY [WAT18], if not stated otherwise. First, the raw count matrix, comprised of $2,663$ cells and $20,613$, is subjected to QC. A minimum and maximum UMI-count of 600 and $25,000$, respectively, and at least 500 genes per cell were required for a cell to pass filtering. Genes had to be measured in at least 20 cells having a MT-fraction below $25\,\%$.

The doublet detection tool Scrublet [WLK19] was applied on both samples individually. The expected_doublet_rate parameter was set to 0.1, which is the anticipated doublet rate by 10x Genomics [Gen22] when $16,000$ cells are loaded on a chip. According to the scrublet score, no doublets were among the cells. After removing low quality cells, the dataset was comprised of $2,187$ cells and $13,304$ genes.

In order to normalise the data, cell specific size factors were determine by first applying a coarse Leiden clustering on the data. Subsequently, data was normalised using scran [LMM16] and a pseudo count of one was added to each gene. Then, the data was log-transformed to base 10. In total $4,000$ HVGs were selected using the flavor cellranger in SCANPY's `highly_variable_gene` function. Since both samples originate from the same donor and were identically processed in the lab, no batch effect was identified. For visualisation, principal component analysis (PCA) was applied and 7 PCs were used to create a KNN graph. The data was subsequently embedded in a 2D manifold using UMAP.

I clustered the scRNA-seq data using the Leiden algorithm. The number of clusters was determined by the maximum silhouette score (section 2.4.2.3) and biological priors, i.e., molecular markers enriched in the clusters. The silhouette score peaked at 0.54 for a resolution of 0.1. The final annotation was based on the enrichment of marker genes and expert knowledge. In total six different cell types were identified, including lymphocytes, APCs, fibroblasts, smooth muscle cells, granulocytes, and KCs.

As described for the ST dataset, leukocyte cells were identified by marker genes and selected for DGE analysis. Subsequently, I compared cytokine-expressing cells against cytokine-negative cells. The size factor for each cell was calculated on the entire dataset and provided to the DGE analysis tool, glmGamPoi [AEH20]. I accounted for additional biological and technical effects, such as the cdr and cell type annotations, in the design function

$$y_{cg} \sim \text{cdr} + \text{annotation} + \text{condition} \ , \tag{5.2}$$

where y_{cg} is the raw count of gene g in a cell c. The condition variable marks the cytokine-positive and cytokine-negative cells. After performing the DGE analysis, the p-values were corrected for FDR using BH. A gene was called DEG if it met the thresholds of p-value ≤ 0.05 and $|\text{log2FC}| \geq 1$.

Before performing the ORA, the entrezIDs were assigned to each gene by leveraging the database org.Hs.eg.db [Car+19]. The DEGs were then subjected to the Bioconductor [Gen+04] R library ReactomePA [YH16], using as background all measured genes in the experiment. The p-values were corrected for FDR using BH. A pathway was defined as enriched if its padj value was below 0.05. The R library enrichplot [Yu21] was used for visualisation purposes.

5.2.3 Preprocessing of a public scRNA-seq dataset

I analysed a public dataset from Reynolds et al. (2021), which I acquired from zenodo.org [Rey+21] [GR20] (Table 5.3). In total 96 samples from L and NL psoriasis and atopic dermatitis skin biopsies as well as from healthy donors were collected. The skin biopsies were donated by 5 healthy, 3 psoriasis, and 4 atopic dermatitis patients. The scRNA-seq repository contained $451,594$ cells of 42 different cell types and $33,538$ genes. Reynolds et al. (2021) [Rey+21] provided a preprocessed version of the data, which I subjected to a more stringent QC.

During QC, genes were required to have at least one UMI-count and to be detected in at least 30 cells. Moreover, a cell had to have at least 250 genes and a minimum of 500 and a maximum of $400,000$ UMI-counts. Additionally, a MT-fraction above 25 % and a ribosomal fraction in the range of 5% to 60% were required for a cell to pass the filtering. Potential doublets, detected by scrublet [WLK19], were removed if they had a scrublet

Diagnosis	Samples	Donors	Age	Severity	Location
healthy	40 samples 195,739 cells	5 patients	- -	- -	breast
psoriasis	24 samples 137,320 cells	3 patients	54.67 ± 12.6	18.6 ± 5.0	lower back
eczema	32 samples 118,535 cells	4 patients	48.50 ± 17.0	9.0 ± 3.6	lower back

Table 5.3: Overview of the publicly, available scRNA-seq dataset of Reynolds, Gary, et al. (2021) [Rey+21]. A total of 42 cell types were detected and their location in the skin layers, epidermis and dermis, was also provided. The severity scores refer to Eczema Area and Severity Index and Psoriasis Area and Severity Index (PASI) for eczema and psoriasis, respectively.

score above 0.6. After removing low quality cells and potential doublets, the dataset was comprised of 429,285 cells and 24,560 genes.

5.2.4 Bulk RNA-sequencing data

In this study, I applied a coarse QC using SCANPY [WAT18] on the raw bulk RNA-seq count matrix, having in total 330 samples and 28,515 genes. The samples were collected from 168 patients with 168 L and 162 NL samples. Patients were diagnosed with psoriasis (n_L=90, n_{NL}=88), lichen planus (n_L=30 , n_{NL}=28), and eczema (n_L=48, n_{NL}=46) (Table 5.4). Genes measured in less than 20 samples were filtered out. After QC, the dataset contained 330 samples and 24,298 genes. In order to visualise the data, I normalised the filtered count matrix, yielding the same total library size for each sample, and subsequently log-transformed, adding a pseudo-count of one. No batch correction was applied, as only the raw counts were used for validation. PCA was applied keeping the top 6 PCs, as determined by the elbow method. The PCs were leveraged to create a KNN graph, which was subsequently used to generate the UMAP embedding.

5.2.5 Additional validation techniques

The presented experiments were conducted by the collaborating biologists Stefanie Eyerich from the ZAUM and Emanuele Scala from the Karolinska Institute. Detailed description can be found in Schäbitz & Hillig et al. (2022) [Sch+22].

ISH was performed on six L formalin-fixed paraffin-embedded (FFPE) skin sections of psoriasis, eczema, and lichen planus patients. The tissue sections had a thickness of

Diagnosis	**Sex**	**Samples**	**Donors**	**Age**
lichen planus	male	13 L, 12 NL	13	55.13 ± 12.59
	female	17 L, 16 NL	17	61.67 ± 14.73
eczema	male	32 L, 31 NL	32	48.91 ± 17.94
	female	16 L, 15 NL	16	47.59 ± 22.56
psoriasis	male	52 L, 50 NL	52	52.34 ± 13.90
	female	38 L, 38 NL	38	57.36 ± 18.50

Table 5.4: Overview of the unpublished bulk RNA-seq dataset, which comprises a total of 330 samples.

5 μm. For each disease representative one positive and one negative control was created. The images were taken using a microscope scanner and visualised using the software QuPath-0.3.2 [Ban+17]. A cell was denoted as positive if the RNA signal matched the nuclear background staining.

For the immuno histochemistry (IHC), in total 20 L FFPE psoriasis samples were taken and cut into 5 μm tissue sections. The IL - 17A proteins were stained using IL - 17A antibodies. Afterwards, the slides were counter stained with hematoxylin. In addition to the positive samples, negative controls were generated. The number of positive cells was determined by counting their appearance in four to nine visual fields.

In the flow cytometry analysis, human L psoriasis skin biopsies (∅ 6 mm) were dissociated into single cells and gated for T-cells (n=52). The T-cells were used for flow cytometric analysis. They were stimulated for 5 h and subsequently stained for CD3, CD4, CD8, IL - 17A, IFN - γ, TNF - α, IL - 22, IL - 10 cytokines. The number of co-expressed cytokines of IL - 17A in T-cells were determined using the FlowJo^{TM} Software [BC20].

For the time course analysis, blood cells from three healthy donors (n=1 male, n=2 female, age=38 ± 7) were isolated by centrifugation. Primary human Pan T-cells were then separated into CD4 and CD8 positive subsets. One subset was stimulated to activate T-cell receptors for 10 min, 1 h, or 6 h, while the other remained unstimulated. Activated cells were collected at 10 min, 30 min, 1 h, 6 h, 12 h, or 24 h post-stimulation. RNA was extracted following cell lysis, and real-time polymerase chain reaction (PCR) was performed for counting the copy number of *IFNG*, *IL13*, and *IL17A* of the stimulated and unstimulated T-cells.

5.3 Results

I examined the pathogenic micro-environment of the most common ncISDs, including lichen planus, eczema, and psoriasis, using ST data from L and NL sections. This resulted in a dataset of 90 samples, including 31 L and 14 matched NL specimens, existing in duplicates. Overall 71,606 transcriptomes were captured (Figure 5.1 a, Methods 5.2.1), with a mean number of spots of 767 ± 293 per section. Notably, L sections contained more spots than NL ones (L: 823 ± 324, NL: 633 ± 125; p-value = $1.5e$-03), potentially due to inflammation-induced morphological changes, such as epidermal thickening in psoriasis [Ruc+11] [DRM11]. Additionally, the mean number of UMI-counts per spot was greater in L compared to NL skin sections (L: 3,189 ± 6,620, NL: 605 ± 613; p-value = $2.0e$-04) (Supplementary Table C.6). The ST data revealed significant differences between L and NL skin sections, suggesting inflammation-driven alterations in tissue architecture and transcriptional regulation in ncISDs.

A workflow was designed to differentiate between bystander and disease-promoting cells using two complementary analysis approaches (Figure 5.1 b-g). Given the crucial role of leukocytes in immune regulation, these methods aimed to explore their function and impact on the surrounding micro-environment (Section 1.1.2) [She+23] [Mur20].

The first part (Figure 5.1 b-d) investigated the biological rationale between hallmark cytokine transcript-positive and -negative spots in leukocytes. These spots were first identified and compared using DGE analysis, followed by pathway enrichment analysis to determine associated biological functions.

Complementing this approach, the second part (Figure 5.1 e-g) examined the spatial organisation of cytokine-producing leukocytes. A density-based clustering algorithm (Section 5.3.3) was applied to first identify cytokine transcript-positive spots (Figure 5.1 e), then group cytokine and responder transcript-positive spots within a user-defined radius (Figure 5.1 f). Finally, the weighted correlation between cytokines and their responders was calculated (Figure 5.1 g).

In order to proof the validity of my findings, multiple techniques were employed. Methods such as ISH, single cell and bulk RNA-seq, IHC, flow cytometry and cell culture analysis (Figure 5.1 i-n) provided additional confirmation. This comprehensive workflow, integrating both established and new analytical approaches and validated through various techniques, effectively differentiate between bystander and disease-promoting cells.

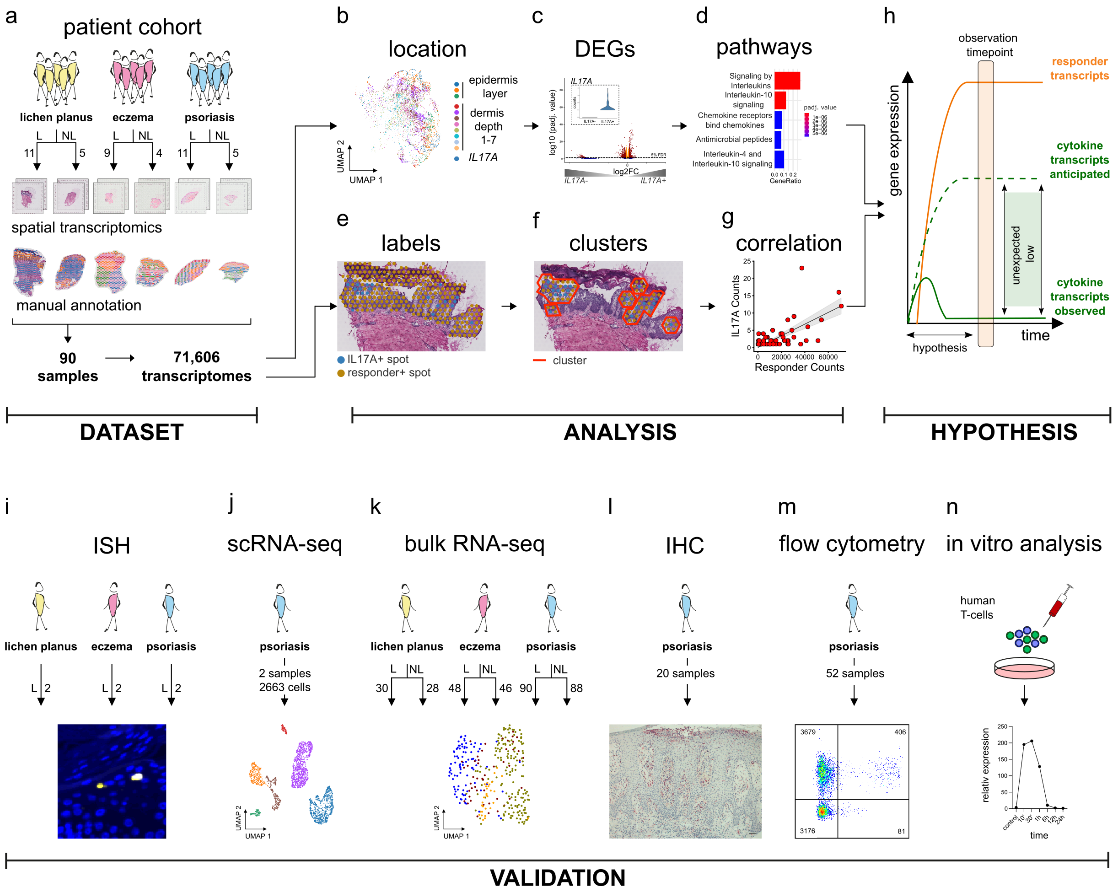


Figure 5.1: Analysis workflow and validation data. **a)** ST dataset comprised of 90 specimen (62 L and 28 NL) of 31 patients being placed on 72 capture areas. The number of capture areas is equal to the number of H&E images. In addition, each specimen has one replicate. In total 71,606 transcriptomes are measured and each is assigned to a skin layer. **b)** UMAP of immune cell containing spots showing the skin layer annotation and location of *IL17A* transcript positive spots. The latter requires at least one UMI-count of *IL17A* in the corresponding spot. **c, d)** Functional characterisation of cytokine transcript positive leukocyte spots leveraging DGE and pathway enrichment analysis. **e)** Preparation phase for the clustering, which requires the information whether a spots is cytokine and/or responder transcript positive. **f)** Spatial density clustering is applied to group cytokine and responder positive spots being co- or close localised in a user defined radius. **g)** The summed UMI-counts of a cytokine and its responder signature in each cluster are correlated leveraging the spatial spearman correlation. **h)** Hypothesis, expecting a higher number of cytokine transcripts than actually observed. **i-n)** Validation experiments, i.e. ISH, scRNA-seq, bulk RNA-seq, IHC, flow cytometry, and in *vitro* analysis, are leveraged to confirm the finding of low expressed cytokines in ncISDs. Figure adapted and modified from [Sch+22].

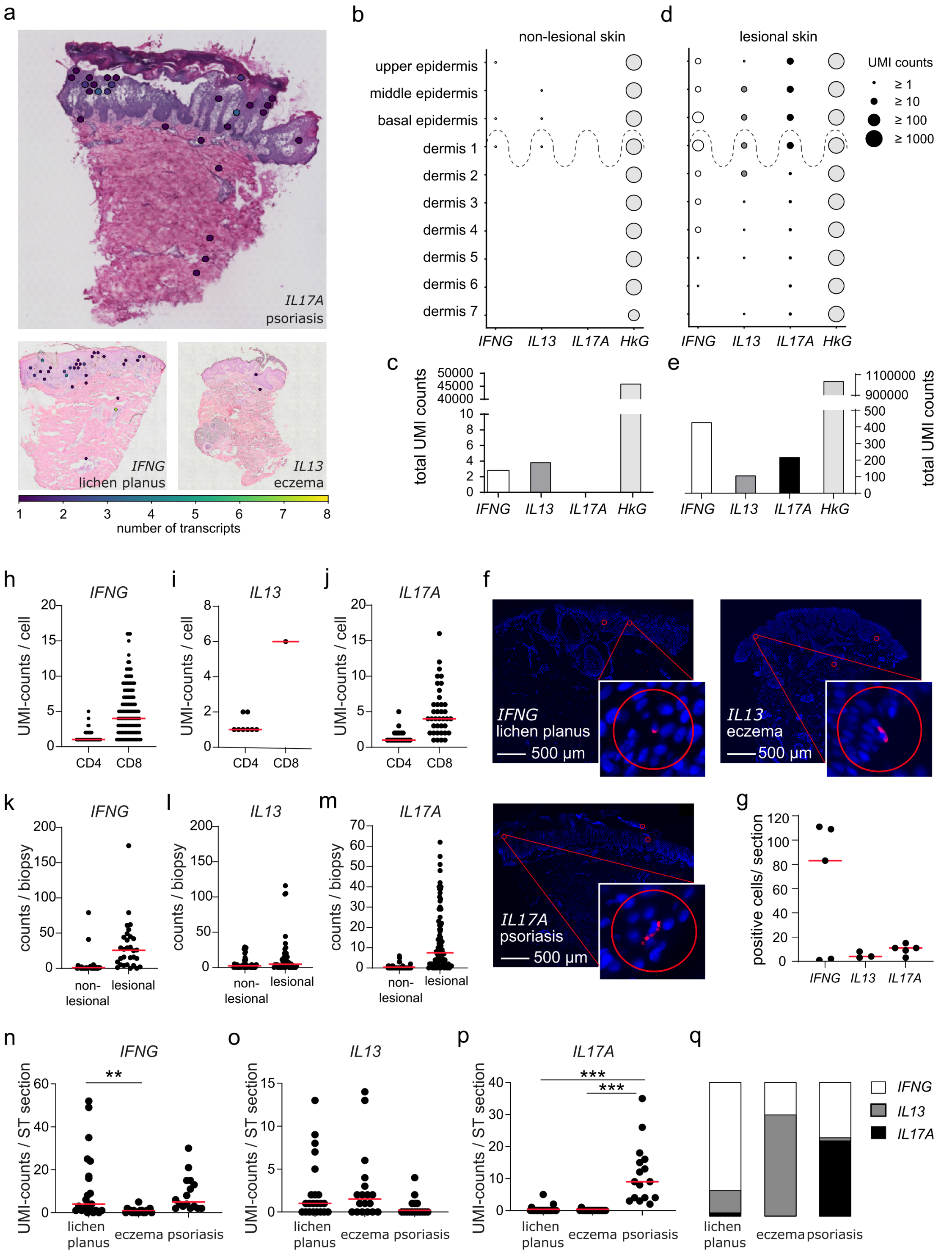

a
IL17A
psoriasis
IFNG
lichen planus
IL13
eczema
1
2
3
4
5
6
7
8
number of transcripts
b
non-lesional skin
upper epidermis
middle epidermis
basal epidermis
dermis 1
dermis 2
dermis 3
dermis 4
dermis 5
dermis 6
dermis 7
IFNG
IL13
IL17A
HkG
d
lesional skin
UMI counts
≥ 1
≥ 10
≥ 100
≥ 1000
c
total UMI counts
50000
45000
40000
10
8
6
4
2
0
e
1100000
900000
500
400
300
200
100
0
h
IFNG
UMI-counts / cell
20
15
10
5
0
CD4
CD8
i
IL13
8
6
4
2
0
j
IL17A
f
IFNG
lichen planus
500 µm
IL13
eczema
500 µm
IL17A
psoriasis
500 µm
k
counts / biopsy
200
150
100
50
0
non-lesional
lesional
l
m
70
60
50
40
30
20
10
g
positive cells/ section
120
100
80
60
40
20
0
n
UMI-counts / ST section
60
40
20
0
**
lichen planus
eczema
psoriasis
o
15
10
5
p
***
***
40
30
q
IFNG
IL13
IL17A

Figure 5.2: Hallmark cytokines found to be lowly expressed in all experiments. **a)** Representative ST tissue sections of lichen planus, eczema, and psoriasis showing the location and UMI-counts of their corresponding hallmark cytokines *IFNG*, *IL13*, *IL17A*, respectively. **b-e)** Total number of raw UMI-counts of cytokines and HKG, *GAPDH*, per and over all skin layers in NL **(b, c)** and L **(d, e)** skin of all specimen (n=82). **f)** Representative ISH staining sections of hallmark cytokines in lichen planus, eczema, and psoriasis. The size of the red circle is identical to the size of a Visium spot ($\varnothing$ 55 μm). **g)** Number of cytokine positive cells per ISH section of lichen planus (n=5), eczema (n=3), and psoriasis (n=5). The red colour marks signal of cytokine mRNA and blue colour the background. **h-j)** Raw UMI-counts per cell of *IFNG* (178 cells), *IL13* (9 cells), and *IL17A* (61 cells) in cytotoxic T-cell (Tc cell) (*CD8*+) and T-helper (Th) (*CD4*+) cells including both psoriasis samples (n=2). **k-m)** Raw bulk RNA-seq counts per biopsy in NL and L skin of lichen planus (n_{NL}=28, n_L=30), eczema (n_{NL}=46, n_L=48), and psoriasis (n_{NL} = 88, n_L=90). **n-p)** UMI-counts per ST section divided by disease (lichen planus: n=22, eczema: n=8, psoriasis: n=18). Statistical testing is conducted using one-way analysis of variance (ANOVA) and Tukey's HSD post hoc test without FDR correction. Stars indicate **: p-value < 0.01 and ***: p-value < 0.001. **q)** Composition of hallmark cytokines per disease. Figure adapted and modified from [Sch+22].

5.3.1 Low expression of disease-promoting cytokines in lesion skin

Disease-promoting cells play a crucial role in coordinating and regulating the immune response, as they secrete signalling molecules, known as cytokines. They regulate the strength of the immune response by altering the gene expression in the immune cells [She+23] [Mur20]. I investigated the corresponding hallmark cytokines *IFNG*, *IL13*, and *IL17A* of lichen planus, eczema, and psoriasis, respectively, in a spatially resolved manner using a ST dataset (Figure 5.2 a, Figure 5.1 b). In contrast to the anticipated high concentration of cytokines in inflamed tissue, only a few cytokine transcript-positive spots were observed in each of the selected L skin sections, which are representatives of the respective diseases. Furthermore, the maximum number of measured UMI-counts of *IFNG*, *IL13*, and *IL17A* per spot was 8, 3, and 4, respectively. Despite the pivotal role of cytokines in disease promotion, I observed unexpectedly low numbers of these cytokine transcripts in inflamed tissue in three randomly selected representatives of each disease.

The cohort was analysed to assess cytokine expression levels and spatial distribution within tissue sections. As anticipated, *IFNG*, *IL13*, and *IL17A* transcripts were scarcely detected in NL samples, with a mean UMI-count of 1, 1, and 0, respectively (Figure 5.2 b, c; Supplementary Table C.6). In contrast, L sections exhibited distinct spatial enrichment patterns (Figure 5.2 d). *IFNG* was higher expressed in the dermal-epidermal junction (p-value = $1.6e-22$), *IL13* was enriched in the middle/basal epidermis and upper dermis

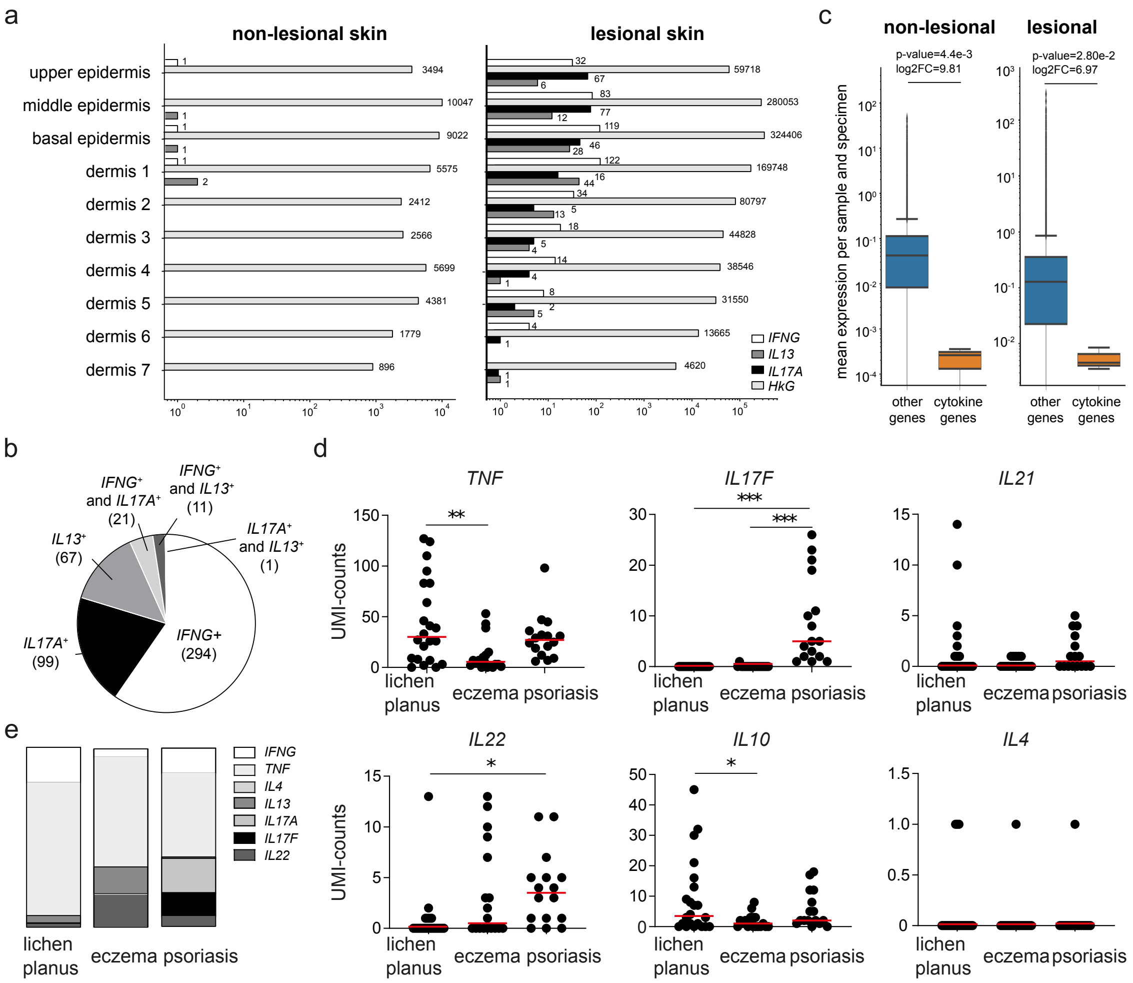


Figure 5.3: Low expression of cytokines in ncISDs. **a)** Total number of UMI-counts of the hallmark cytokines, i.e. *IFNG*, *IL13*, and *IL17A*, and the HKG *GAPDH* in NL and L skin. **b)** Proportion and number of single and double cytokine transcript-positive spots. **c)** Comparison of the average expression of all other genes and the hallmark cytokines in NL (n=24) and L (n=58) skin. Statistical testing is conducted using the two-sided Mann-Whitney-U test. Boxplot information of NL (others, cytokines): median (4.18*e*-01, 2.57*e*-04), mean (1.82*e*-01, 2.02*e*-04), min (0, 0), max (3*e*-01, 3.50*e*-04), Q1 (0, 1.29*e*-04), Q3 (1*e*-01, 3.04*e*-04). Boxplot information of L (others, cytokines): median (1.26*e*-01, 4.50-03), mean (6.87*e*-01, 5.49*e*-03), min (0, 0), max (8*e*-01, 8.45-03), Q1 (2.22*e*-02, 4.00-03), Q3 (3.51*e*-01, 6.48-03). **d)** Expression level of cytokines in lichen planus (n=32), eczema (n=26), and psoriasis (n=24). Each dot represents a specimen. Statistical testing is conducted using one-way ANOVA and Tukey's HSD post hoc test without FDR correction. Stars indicate *: p-value < 0.05, **: p-value < 0.01, and ***: p-value < 0.001. **e)** Disease cytokine composition. UMI-counts are normalised to 100 %. Figure adapted from [Sch+22].

layers 1-2 (p-value $= 2.41e\text{-}17$), and *IL17A* was mainly observed in the epidermis and dermis depth 1 (p-value $= 1.81e\text{-}27$). These findings align with reported cytokine localisation in the epidermis and upper skin layers [Gri+21]. In essence, I observed distinct spatial patterns of the hallmark cytokine in L skin, providing insights into their localised roles within tissue architecture.

The distribution and expression level of the hallmark cytokines were compared to the HKG *GAPDH*, as HKGs are uniformly expressed across cells [Jos+22]. I observed that the hallmark cytokines were lower expressed compared to *GAPDH* (Figure 5.2 d, e), which was evenly expressed and distributed across skin layers. The expression levels for *IFNG*, *IL13*, and *IL17A* ranged from 1 to 37, 1 to 12, and 1 to 27 per tissue section, respectively. In total, across L tissue sections, only 434, 144 and 224 UMI-counts were measured for *IFNG*, *IL13*, and *IL17A*, respectively. Considering the $44,034$ L spots analysed, these counts were distributed across only 372, 103 and 154 L spots, respectively (Figure 5.2 e, Figure 5.3 a, Supplementary table C.6). Overall I observed low expression and sparse distribution of hallmark cytokines in comparison to a HKG, indicating specific roles within distinct L micro-environments.

A disease specific expression pattern was observed in the ST dataset, despite the generally low expression of the cytokine transcripts. *IFNG*, a hallmark cytokine of lichen planus, was predominantly expressed in L skin from lichen planus patients (median of 9 UMI-counts per section) (Figure 5.2 n). The *IL13* was mainly expressed in L eczema skin (median of 1.5 UMI-counts per section) (Figure 5.2 o). *IL17A* was significantly more expressed in L psoriatic skin (median of 9 UMI-counts per section) than in lichen planus or eczema (Figure 5.2 p). These observations confirmed each cytokine was most prevalent in its respective disease (Figure 5.2 q). Similarly, disease-specific expression patterns were observed for other cytokines, including *IL17F*, *IL21*, *IL22*, *TNF*, *IL10*, and *IL4* (Figure 5.3 d, e). This supports the classification of lichen planus as type 1, eczema as type 2, and psoriasis as type 3 immune-driven diseases [EE18]. These results validated the ST dataset's reliability, as they align with known cytokine disease-specific expression patterns.

As another quality check of the ST data, I compared L and NL gene expression levels. The expression levels were investigated by comparing the mean hallmark cytokine transcripts per specimen against the other genes (Figure 5.3 c). The mean expression level was found to be 900 and 125 times lower in NL and L skin, respectively. These results align with literature indicating lower expression of cutaneous hallmark cytokines in NL compared to L skin.

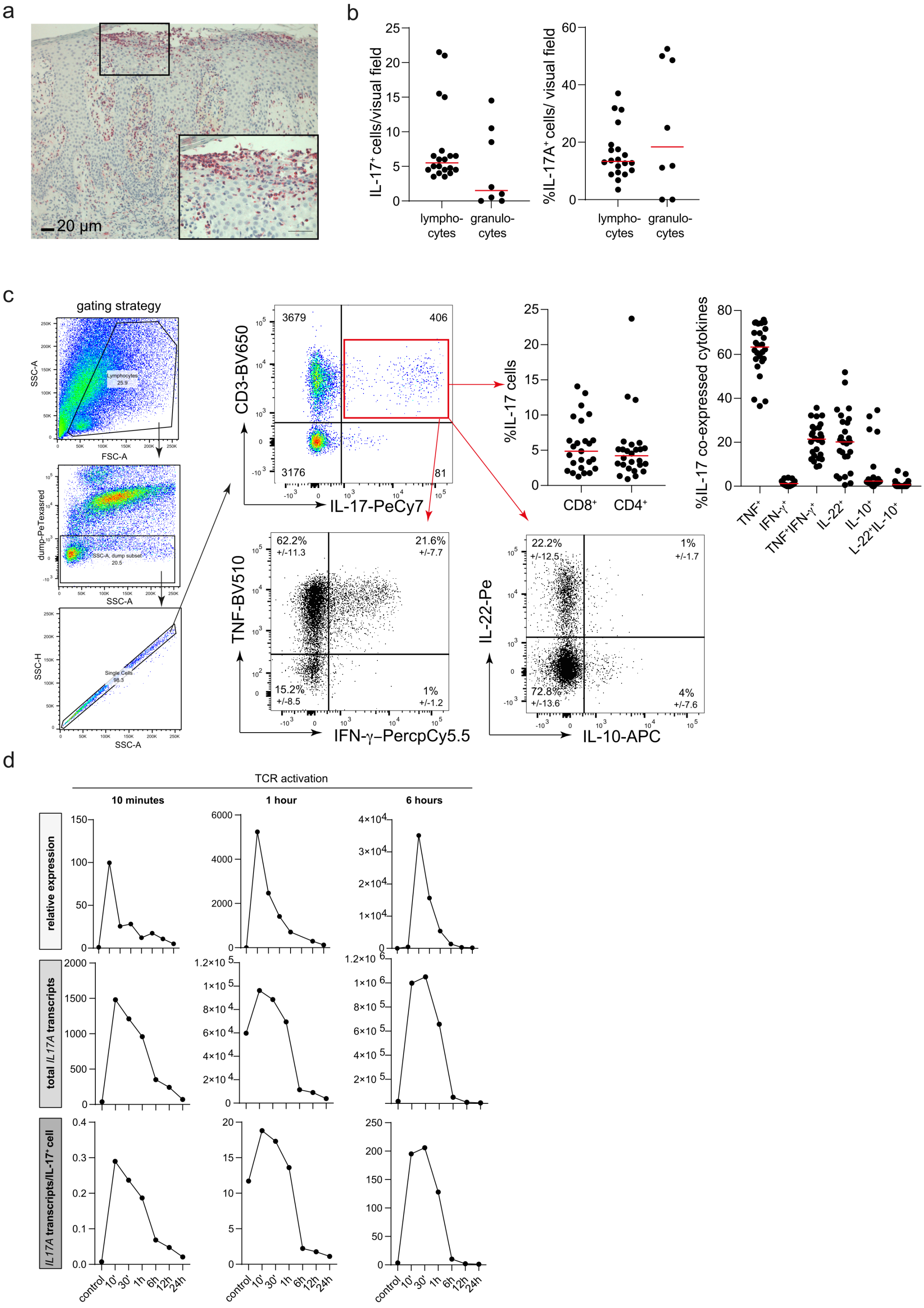

a
20 µm
b
IL-17+ cells/visual field
25
20
15
10
5
0
lympho-cytes
granulo-cytes
%IL-17A+ cells/ visual field
60
40
20
0
lympho-cytes
granulo-cytes
c
gating strategy
SSC-A
FSC-A
Lymphocytes 25.9
dump-PeTexasred
SSC-A, dump subset 20.5
SSC-A
SSC-H
Single Cells 98.3
SSC-A
CD3-BV650
3679
406
3176
81
IL-17-PeCy7
%IL-17 cells
25
20
15
10
5
0
CD8+
CD4+
%IL-17 co-expressed cytokines
80
60
40
20
0
TNF+
IFN-γ+
TNF+IFN-γ+
IL-22+
IL-10+
IL-22+IL-10+
TNF-BV510
62.2%
+/-11.3
21.6%
+/-7.7
15.2%
+/-8.5
1%
+/-1.2
IFN-γ–PercpCy5.5
IL-22-Pe
22.2%
+/-12.5
1%
+/-1.7
72.8%
+/-13.6
4%
+/-7.6
IL-10-APC
d
TCR activation
10 minutes
1 hour
6 hours
relative expression
150
100
50
0
6000
4000
2000
0
4×10^4
3×10^4
2×10^4
1×10^4
0
total IL17A transcripts
2000
1500
1000
500
0
1.2×10^5
1×10^5
8×10^4
6×10^4
4×10^4
2×10^4
0
1.2×10^6
1×10^6
8×10^5
6×10^5
4×10^5
2×10^5
0
IL17A transcripts/IL-17+ cell
0.4
0.3
0.2
0.1
0.0
20
15
10
5
0
250
200
150
100
50
0
control
10'
30'
1h
6h
12h
24h

Figure 5.4: Validation experiments on protein and RNA level of cytokine expressing T-cells in skin lesions. **a)** Representative IHC image of a L psoriasis section showing in red the stained IL - 17A protein. **b)** The number and percentage of IL - 17A positive cells per visual field is counted and calculated of all IHC images (n=20). **c)** Flow cell cytometry pipeline showing the gating strategy for and staining of T-cells isolated from L psoriasis skin biopsies (n=52). The $CD3_{+}IL-17_{+}$ cells are gated for co-production IL - 17A with TNF/IFN - γ and IL - 22/IL - 10. The percentage of IL - 17A secreting Th and Tc cell cells and the percentage of IL - 17A being co-expressed with either one or a combination of cytokines is shown. **d)** Half-life time study of in vitro stimulated (10 min, 1 h, 6 h) and unstimulated (control) Th cells isolated from human healthy blood donors. RNA is isolated and analysed for *IL17A* expression over 24 h using real time PCR. The relative expression of *IL17A* (first row), total number of *IL17A* transcripts (second row), and *IL17A* transcripts per cell (last row) are shown. Figure adapted and modified from [Sch+22].

To assess Visium's sensitivity in detecting cytokine transcripts, I validated the findings using ex vivo and in vitro methods (Figure 5.2 f-m). ISH (Section 1.4.3) was applied to detect cytokine mRNA on tissue sections, allowing direct comparison with Visium by counting mRNA molecules within circles matching the size of a spot ($\varnothing$ 55 μm) (Figure 5.2 f). In line with the ST observations, only a few cytokine mRNA signals were detected per tissue section, with a median number of 83, 4 and 11 for lichen planus, eczema, and psoriasis, respectively. The validation using ISH confirmed that Visium accurately detected the low cytokine transcript levels across tissue sections.

To further validate the detection of cytokine transcripts, I examined their presence in CD4+ and CD8+ cells using scRNA-seq and in bulk RNA-seq. In CD4+ (Th) cells, a median UMI-count of 1 for all hallmark cytokines was observed, while CD8+ (Tc cell) cells showed medians of 4, 6, and 4 for *IFNG*, *IL13*, and *IL17A*, respectively (Figure 5.2 h-j). In addition, the number of cytokine counts in the bulk RNA-seq cohort (n=330), with L (n=168) and NL (n=162) samples, were investigated (Figure 5.2 k-m). To compare the size of the punch biopsies with the Visium skin sections, one third of a 6 mm skin punch biopsy was analysed, which is 60 times the amount of genetic material in a 10 μm thick ST tissue section. A median of 1 and 25.5 counts/biopsy of *IFNG* in NL and L lichen planus skin, respectively, was observed (Figure 5.2 k). In eczema, a median of 2 and 4.5 counts/biopsy of *IL13* in NL and L skin, respectively, was measured (Figure 5.2 l). In psoriasis a median of 0 and 7.5 *IL17A* counts/biopsy in NL and L skin, respectively, was detected (Figure 5.2 m). Both scRNA-seq and bulk RNA-seq confirmed the low presence of cytokine transcripts, supporting Visium's sensitivity in detecting these molecules.

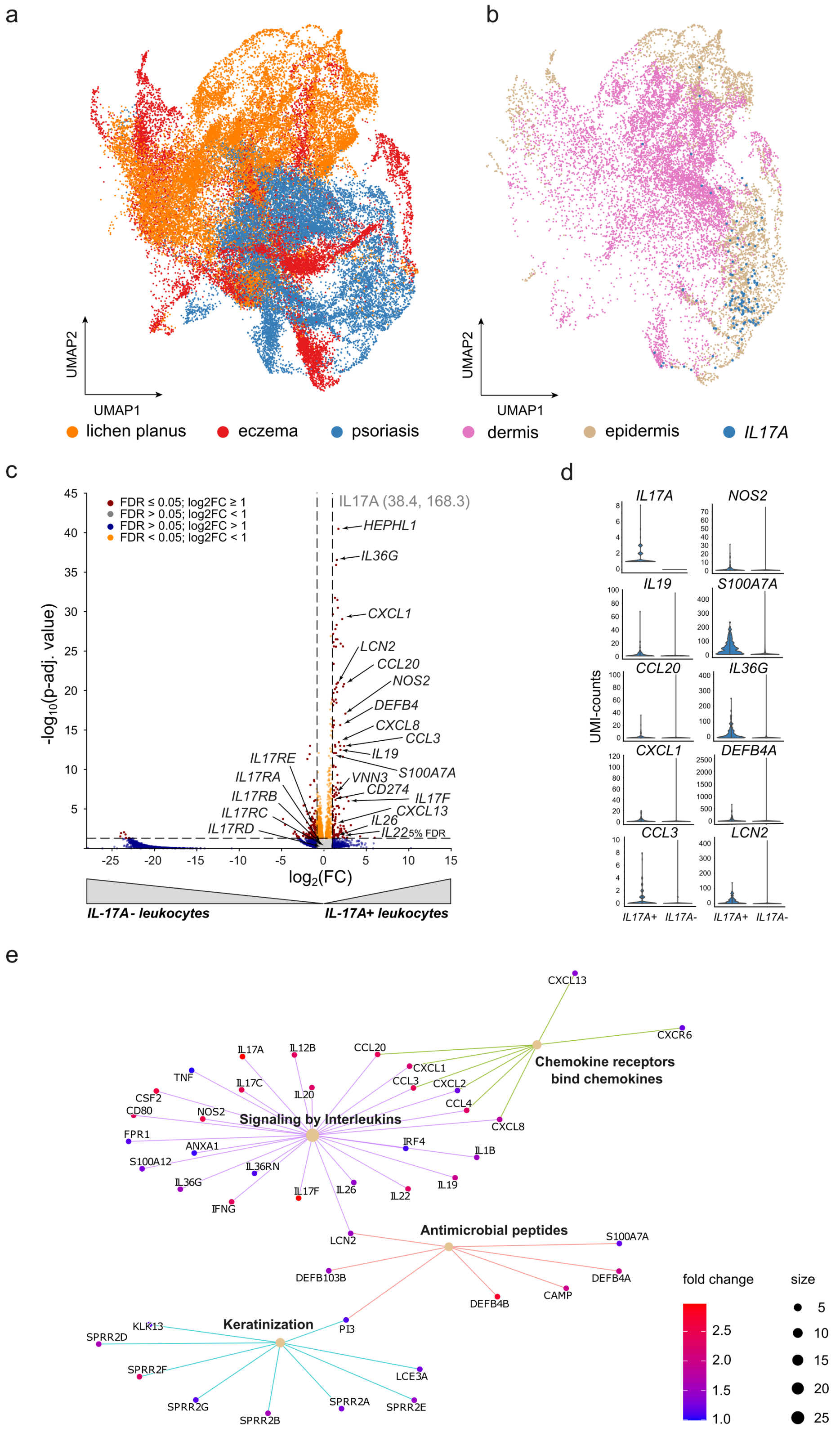

a
UMAP2
UMAP1
● lichen planus ● eczema ● psoriasis
b
UMAP2
UMAP1
● dermis ● epidermis ● IL17A
c
FDR ≤ 0.05; log2FC ≥ 1
FDR > 0.05; log2FC < 1
FDR > 0.05; log2FC > 1
FDR < 0.05; log2FC < 1
IL17A (38.4, 168.3)
HEPHL1
IL36G
CXCL1
LCN2
CCL20
NOS2
DEFB4
CXCL8
CCL3
IL19
S100A7A
VNN3
CD274
IL17F
CXCL13
IL26
IL22
5% FDR
IL17RE
IL17RA
IL17RB
IL17RC
IL17RD
-log$_{10}$(p-adj. value)
log$_2$(FC)
45 40 35 30 25 20 15 10 5
-25 -20 -15 -10 -5 0 5 10 15
IL-17A- leukocytes
IL-17A+ leukocytes
d
UMI-counts
IL17A
NOS2
IL19
S100A7A
CCL20
IL36G
CXCL1
DEFB4A
CCL3
LCN2
IL17A+ IL17A- IL17A+ IL17A-
e
CXCL13
CXCR6
Chemokine receptors bind chemokines
IL17A
IL12B
CCL20
CXCL1
TNF
IL17C
CCL3
CXCL2
IL20
CSF2
CCL4
CD80
NOS2
Signaling by Interleukins
CXCL8
FPR1
ANXA1
IRF4
IL1B
S100A12
IL36RN
IL36G
IL17F
IL26
IL22
IL19
IFNG
LCN2
Antimicrobial peptides
S100A7A
DEFB103B
DEFB4A
DEFB4B
CAMP
KLK13
Keratinization
PI3
SPRR2D
SPRR2F
LCE3A
SPRR2G
SPRR2A
SPRR2E
SPRR2B
fold change
2.5
2.0
1.5
1.0
size
5
10
15
20
25

Figure 5.5: Micro-environment of cytokine transcript-positive spots is characterised by type 3 immune response. **a)** Gene expression embedding of the whole ST data showing the disease annotations. **b)** Embedding of leukocyte containing spots colouring the tissue layer annotations and cytokine transcript-positive spots. **c)** DGE analysis between *IL17A* transcript-positive and -negative leukocyte spots. *IL17A* (38.4, 168.3) is not shown. **d)** Violinplots of selected upregulated DEGs (padj value < 0.05, log2FC > 1) in cytokine transcript-positive spots. Amongst them are gene regulated and induced by *IL17A*. **e)** ORA of upregulated DEGs (padj value < 0.05, log2FC > 1) showing the top three enriched pathways according to the padj values. Figure adapted and modified from [Sch+22].

In order to understand the role of cytokine protein-secreting cells in ncISDs, IHC and flow cytometric analysis were utilised. Given the severe observable symptoms of ncISDs on the skin and their treatment with biologics that suppress hallmark cytokines at the protein level, I was interested in examining the number of cytokine protein-secreting cells. Briefly recapitulating, IHC visualises the spatial localisation of cytokine proteins within tissue specimens, while flow cytometric analysis enables the quantitative assessment of cytokine proteins in skin-infiltrating T-cells [Mag+19] (Methods 5.2.5). The results from both methods were comparable and revealed similar numbers of cytokine-positive lymphocytes in the L skin (histology: 13.3 % IL-17A+ lymphocytes, flow cytometry: 4.2 % CD4+IL-17A+, 4.9 % CD8+IL-17A+) (Figure 5.4 a-c). These observations revealed a discrepancy between the measured number of cytokine proteins and observed mRNA in L skin.

To understand the discrepancy, a time course analysis of in vitro stimulated T-cells was conducted to explore the hypothesis that this discrepancy could be due to a short half-life time of cytokine mRNA. The time course analysis showed that short T-cell receptor stimulation of 10 min, 1 h, or 6 h results in a temporary cytokine mRNA production (Figure 5.4 d). It lasted less than 6 h and reached its maximum at 10 min and 30 min. Even though the number of mRNA transcripts per cell increased for longer T-cell receptor activation, the production peak was always at 10 min and 30 min, supporting the hypothesis of a short half-life time. These results suggest that the observed discrepancy between cytokine protein levels and mRNA transcripts in L skin is likely due to the short half-life of cytokine mRNA, as demonstrated by the time course analysis of T-cell receptor stimulation.

In summary, low numbers of disease-specific cytokine transcripts were observed, being expressed in a distinct spatial pattern. These transcripts were only produced by a few

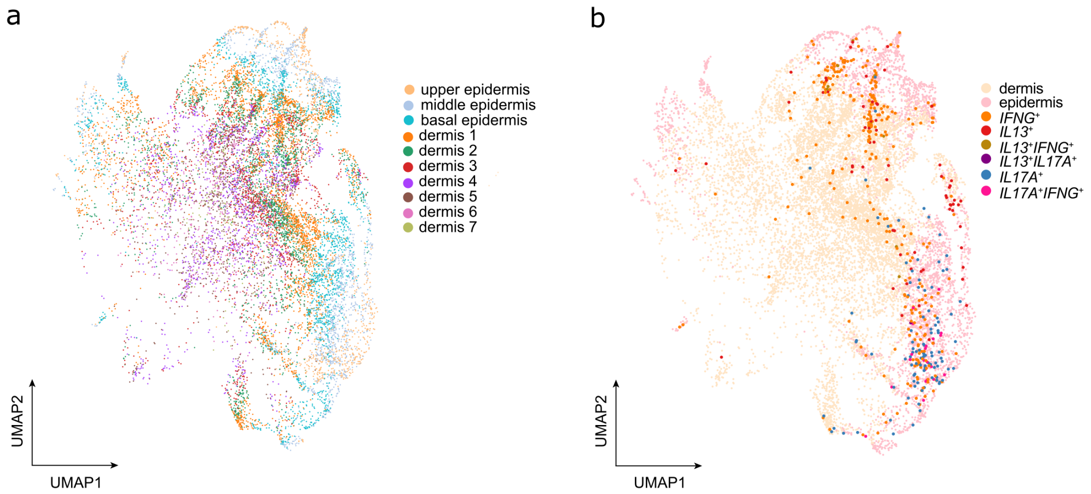


Figure 5.6: Gene expression embedding of leukocyte containing spots. **a)** Fine-resolved tissue layer annotations of each spot showing a similar structure like natural skin. Outermost data points are the spots located in the epidermis, while the innermost spots belong to the dermis. **b)** Location of single and double-positive cytokines in the gene expression embedding. Majority of cytokine transcript positive spots can be found in the epidermis. Figure adapted from [Sch+22].

skin-infiltrating T-cells. Additionally, I confirmed the reliability of the ST dataset by validating this finding in independent methods. My results indicate that the spatial expression pattern of the cytokine secreting immune cells may be important to drive the understanding of psoriasis, eczema, and lichen planus.

5.3.2 Cytokines show up-regulation of immune response regulators

The analysis of ST profiles revealed distinct gene expression and immune cell distribution patterns across skin diseases. The 2D manifold showed good separation of spots from lichen planus and psoriasis, while spots from eczema overlapped with both (Figure 5.5 a). Since immune cells were the focus, leukocyte-containing spots were identified and subsequently, the disease-promoting cytokine transcript-positive spots (Methods 5.2.1) . The 2D embedding of leukocyte-containing spots revealed a distribution of epidermal and dermal spots, aligning with the structural organisation of the skin (Figure 5.5 b, Figure 5.6 a, b). Specifically, the outermost spots belonged to the epidermis, while the inner, central spots belonged to the dermis. The distribution of the spots in the 2D manifold confirmed that ST effectively preserved the positional signature of each spot within the tissue.

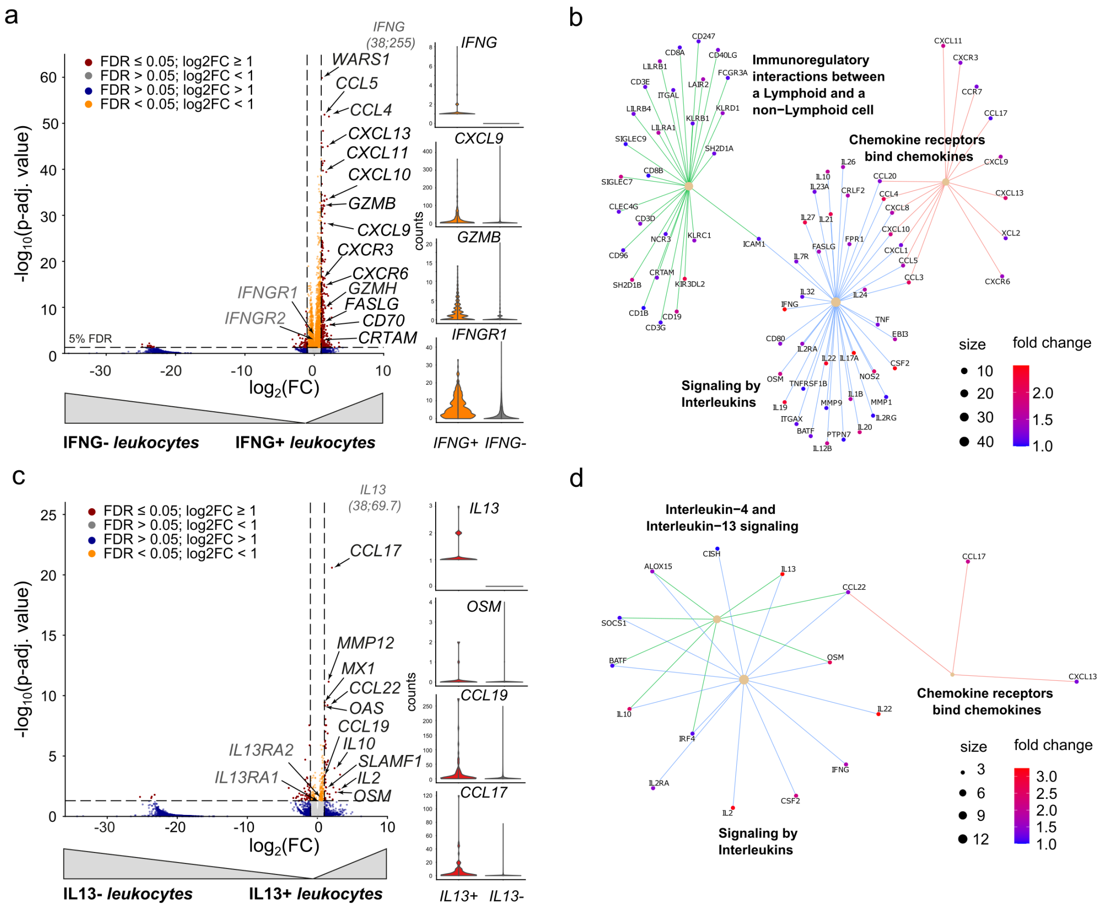


Figure 5.7: Characterisation of *IFNG* and *IL13* transcript-positive leukocyte spots. DGE analysis between *IFNG* **(a)** and between *IL13* **(c)** transcript-positive and -negative spots in ST specimens (n=82). The highlighted genes are either receptors of, induced by, or regulated by the respective cytokine. The following genes are not shown (log2FC, padj value): *IFNG* (38, 255) and *IL13* (38, 69.7). Violinplots showing the raw UMI-counts distribution of selected genes. ORA of upregulated DEG (padj value < 0.05, log2FC > 1) in *IFNG* **(b)** and *IL13* **(d)** transcript positive spots in Reactome pathways. The three most significant pathways are shown. Figure adapted from [Sch+22].

To characterise disease-promoting immune cells, DGE analysis was performed comparing *IFNG*, *IL13*, or *IL17A* transcript-positive spots to transcript-negative spots. The analysis revealed an upregulation of cytokine-associated genes and their responder signature genes. In *IFNG* transcript-positive spots, genes involved in type 1 immune cells, such as *GZMB*, *FASLG*, *CD70*, *CXCR3*, and *CXCR6*, showed significant upregulation. Additionally, *IFNG*-induced genes expressed in epithelial cells, including *CXCL9*, *CXCL10*, and *CXCL11*, were also significantly upregulated (Figure 5.7 a). *IL13* transcript-positive spots were characterised by genes associated with type 2 immune cells, such as *IL2*, *IL10*, and

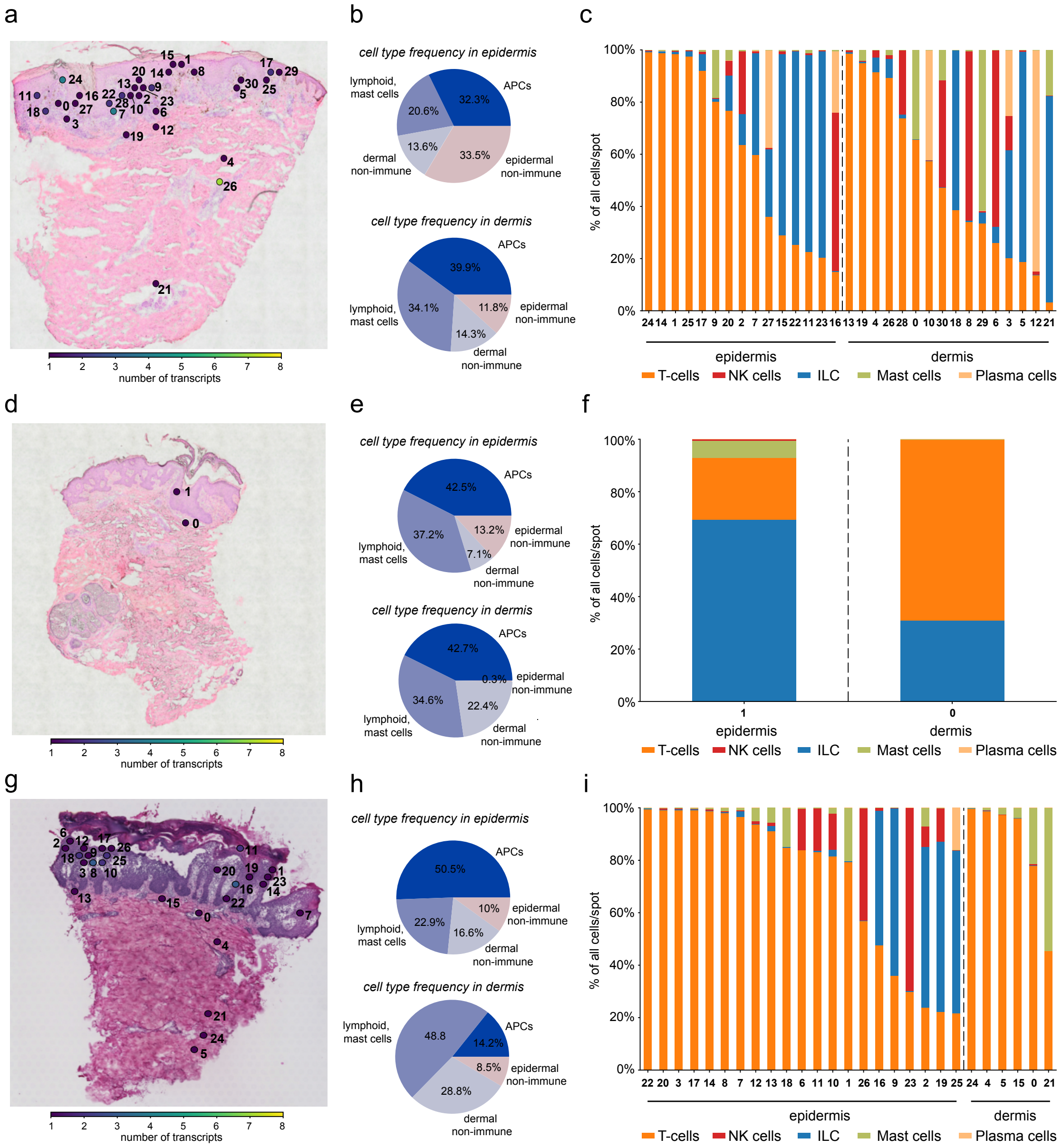


Figure 5.8: Deconvolution of ST spots. Position and numbering of *IFNG* (**a**), *IL13* (**d**), and *IL17A* (**g**) transcript positive leukocyte spots. The scale bar indicate the number of observed transcripts. Tangram is used to reveal the cell type composition in each spot. **b, e, h)** show the cell type proportions of the cytokine transcript-positive spots in the epidermis and dermis. Cell types in the lymphoid/mast cell group: ILC, NK, Tc cell, Th, T regulatory cell (Treg), mast and plasma cells; APCs: DC, Mac, Langerhans cell (LC) cells; epidermal non-immune: KC and melanocyte; dermal non-immune: Schwann, pericyte, VE, LE cells. **c, f, h)** The probabilities of a lymphoid/mast cell being present in a cytokine transcript-positive spot. Figure adapted from [Sch+22].

SLAMF1, which is in line with the literature [EE18]. Genes involved in tissue response, such as *CCL17*, *CCL22*, *MMP12*, and *OSM*, were also upregulated (Figure 5.7 c). Similar results were observed for *IL17A* transcript-positive spots. Genes associated with type 3 immune response, such as *IL17F*, *IL22* and *IL26*, and genes induced by *IL17A*, e.g., *IL36G*, *IL19*, *CCL20*, *CXCL1*, CCL3, LCN2, *NOS2*, *S100A7A*, *DEFB4A*, and *CXCL8* were upregulated in the *IL17A* transcript-positive spots. (Figure 5.5 c, d). The upregulated gene sets demonstrate that ST captured diverse cell types, including those activated by cytokines in their micro-environment.

To investigate the functional implications of the upregulated genes, I conducted an ORA. The analysis identified pathways associated with inflammation-driven signalling, such as "Signalling by Interleukins" and "Chemokine receptors bind chemokines" in *IFNG*, *IL13*, and *IL17A* transcript-positive spots (Figure 5.5 e, Figure 5.7 b, d). Additionally, pathways linked to tissue responses to inflammation, including "Keratinization" and "Antimicrobial peptides", were enriched in *IL17A* transcript-positive spots (Figure 5.5 e). In *IFNG* transcript-positive spots, genes involved in "Immune regulatory interactions between a Lymphoid and a non-Lymphoid cell" were identified (Figure 5.7 b). These findings demonstrate that ST enables to investigate the micro-environment of disease-driving immune cells on a functional level.

I was further interested in the cellular composition of cytokine transcript-positive spots, as a spot can contain multiple cell types. Therefore, I used Tangram [Bia+21] to deconvolute the leukocyte cytokine transcript-positive spots. The resulting cell type compositions showed that APCs, lymphoid cells, and mast cells predominantly express cytokine transcripts (Figure 5.8 a-h). In the epidermis, the percentages of APCs and lymphoid/mast cells were 32.3 %/20.6 %, 42.5 %/37.2 %, and 50.5 %/22.9 % for *IFNG*, *IL13*, and *IL17A*, respectively. In the dermis, the percentages of APCs and lymphoid/mast cells were 39.9 %/34.1 %, 42.7 %/34.6 %, and 14.2 %/48.8 % for *IFNG*, *IL13*, and *IL17A*, respectively (Figure 5.8 b, e, h). T-cells were the dominating cell types within the lymphoid/mast cell group for all hallmark cytokines (Figure 5.8 c, f, i). In summary, the identified cell types in the cytokine transcript-positive spots align with prior expectations, results, and the literature [EE18] [Mar+18] [NL21].

To explore additional signatures captured within cytokine transcript-positive spots, a psoriasis scRNA-seq dataset ($n = 2$) was analysed. Unsupervised clustering identified distinct cell types (Figure 5.9 a, Methods 5.2.2), confirming that *IL17A* and *IFNG* were almost ex-

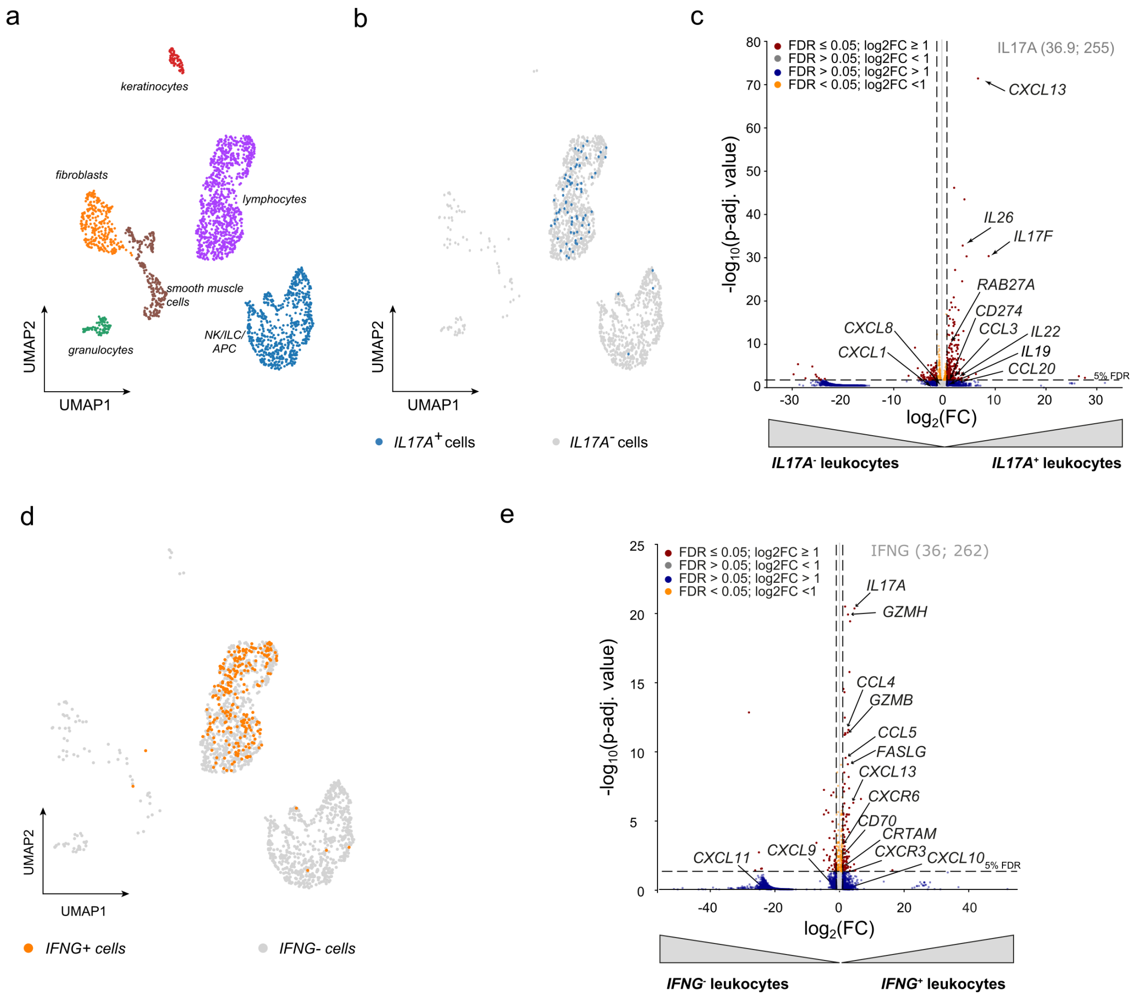


Figure 5.9: Characterisation of cytokine transcript-positive cells. a, b) Embedding of L psoriasis samples (n=2), showing **(a)** cell type annotations and **(b)** leukocytes. The majority of *IL17A*-expressing leukocytes (blue) are located within the lymphocyte cluster. **c)** DGE analysis between *IL17A* transcript-positive and -negative cells. *IL17A* (36.9, 255) is excluded from the plot as it lies outside the displayed range. **d)** Embedding of leukocytes, highlighting *IFNG* transcript-positive leukocytes in orange. **e)** DGE analysis of cytokine transcript-positive versus transcript-negative spots. DEGs that were also identified as DEGs in the ST data are annotated. Thresholds applied: padj value < 0.05 and absolute log2FC > 1. *IFNG* (36, 262) is omitted from the plot as it lies outside the displayed range. Figure adapted and modified from [Sch+22].

clusively expressed in lymphocytes, while *IL13* was not detected, consistent with the literature [Mar+18] (Figure 5.9 b, d). DGE analysis of cytokine-secreting lymphocytes revealed upregulated genes involved in inflammation-driven signalling pathways, which aligned with the DEGs observed in cytokine transcript-positive leukocyte spots

(Figure 5.9 c). As expected, genes associated with inflammatory tissue responses were not among the upregulated DEG in the scRNA-seq data, emphasising ST's ability to capture micro-environmental features of cytokine-secreting cells. Interestingly, *CCL4* and *CCL5*, which are induced by *IFNG* in epithelial cells, were also upregulated in *IFNG*-secreting lymphocytes, suggesting that these chemokines are also produced by lymphocytes (Figure 5.9 e). The analysis demonstrate that while cytokine-secreting lymphocytes predominantly express inflammatory genes, ST captures micro-environmental features surrounding them.

In summary, expected genes and pathways in cytokine transcript-positive spots were found in L skin. Additionally, gene signatures induced by cytokines, which are associated with immune response, were observed. Thus, ST enables to study the direct influence of cytokines on the immune response within the micro-environment in a spatial context.

5.3.3 Cytokine expression is spatially correlated with immune response regulators

Since ST also captured cytokine response signatures in the cytokine transcript-positive spots, I wanted to understand the effect of these cytokines on their close micro-environment in L skin. Therefore, I was interested in the correlation between cytokines and their induced response signatures (Figure 5.10).

Responder signatures for hallmark cytokines were defined by in vitro keratinocyte stimulation and ST DEGs. Since cytokine expression was concentrated in the epidermis and upper dermis (Figure 5.2 c), primary human keratinocytes were stimulated in vitro with recombinant *IFNG*, *IL13*, or *IL17A* (Methods 5.2.1). Upregulated DEGs from stimulated samples were identified using gene expression arrays (Supplemental Figure C.5 a) and intersected with the upregulated DEGs in cytokine transcript-positive spots in the ST data (Figure 5.5 c, Supplemental Figure C.5 a, c). This yielded specific response signatures for *IFNG* (29 genes), *IL13* (4 genes and 10 literature derived genes), and *IL17A* (21 genes) (Supplemental Figure C.5 b-d).

Subsequent enrichment analysis of the responder signatures (Supplemental Figure C.5 e-g) revealed that *IL17A* associated responder genes were almost uniformly expressed across the upregulated DEGs in *IL17A* transcript-positive spots (Supplemental Figure C.5 e). The responder signatures of *IFNG* showed a tendency to be more abundant amongst the top significantly, upregulated genes in *IFNG* (Supplemental Figure C.5 f). No enrichment

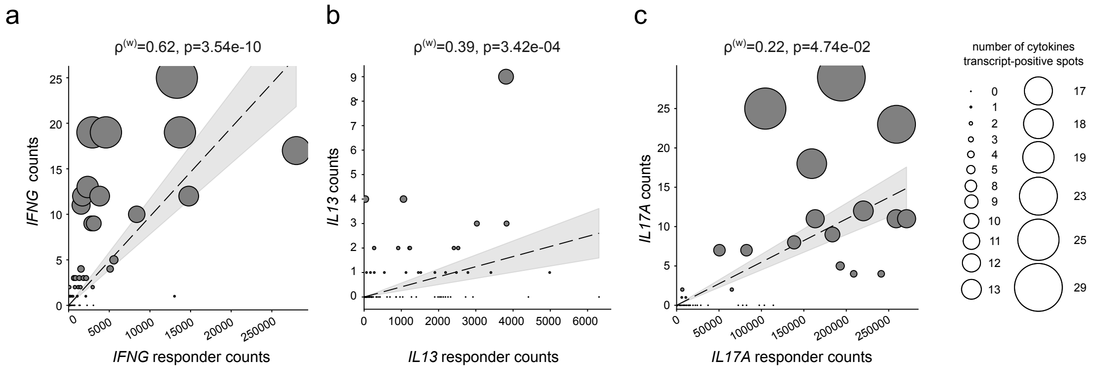


Figure 5.10: Correlation between cytokines and their corresponding response signature genes transcripts. Pseudo-bulk weighted Spearman correlation analysis between the aggregated transcripts of *IFNG* **(a)**, *IL13* **(b)**, or *IL17A* **(b)** and their corresponding responder signatures over all specimen (n=82). The size indicates the number of cytokine transcripts on a specimen. The black dashed line **(a-c)** is fit by an ordinary least square model. The grey area shows the confidence interval $(\mathrm{CI})_{95\%}$. Figure adapted from [Sch+22].

was found for *IL13* induced responders (Supplemental Figure C.5 g). Interestingly, the responder expression was 270 times higher compared to the combined expression of the hallmark cytokines (Supplemental Figure C.5 h). The analysis defined cytokine-specific responder signature sets and revealed a general higher expression level of responders in comparison to the cytokines *IL17A*, *IFNG*, and *IL13*.

To investigate the relationship between hallmark cytokines and their responder signatures, I created pseudo-bulk samples by summarising the counts of each cytokine and its corresponding responder signature genes within each sample. This analysis revealed weak to moderate correlation values between hallmark cytokines and their induced response genes (Figure 5.10 a-c). For *IFNG*, the correlation with its responder genes demonstrated moderate strength ($\rho^{(w)} = 0.62$; $p = 3.54e\text{-}10$). In contrast, I observed weaker correlations for *IL13* and its responder genes ($\rho^{(w)} = 0.39$; $p = 3.42e\text{-}04$)as well as for *IL17A* and its responders ($\rho^{(w)} = 0.22$; $p = 4.74e\text{-}02$). While the pseudo-bulk analysis indicated weak to moderate correlations between hallmark cytokines and their responder signature genes, it lacked spatial resolution, indicating the need for alternative methods to fully understand a cytokine's impact on its micro-environment in L skin.

5.3.3.1 Density-based spatial clustering algorithm

Building upon the spatial information provided by Visium technology, the next step was to examine a cytokine's radius of action on its direct micro-environment within the epidermis. Therefore, a density-based spatial clustering algorithm was developed. This method enabled the identification of spatially coherent clusters, which were then analysed to determine the relationship between cytokines and their micro-environment. To quantify the strength and direction of these relationships, a *spatial weighted correlation* was calculated. The details of the *spatial clustering algorithm* is described in the following.

For each tissue specimen, cytokine transcript-positive spots in the epidermis were identified and designated as set $\mathcal{A}$. A spot was classified as cytokine transcript-positive, if at least 1 UMI-count was observed. The same requirement was imposed to define responder signature genes transcript-positive spots. If $\mathcal{A}$ was non-empty, a K-D Tree was constructed. Otherwise, the next tissue specimen was processed.

A K-D Tree, similar to the Random Forest algorithm (Section 2.4.3.2), follows rules to partition the data in an optimal manner and enables efficient nearest neighbour search [BSW77]. It was used to identify all spots within a fixed distance ($d = 2$) of a cytokine transcript-positive spot, solving the *Fixed-radius Near Neighbor Search Problem.* Commonly, the point under consideration, i.e. the query point, is excluded from the set of nearest neighbours [BSW77]. The implemented version retained the query point within the identified cluster. This adjustment ensured that the cytokine transcript-positive spot under consideration remained included in the cluster definition 5.3.

Definition 3. ***(Fixed-radius Near Neighbor Search Problem)*** *Given a set of input points $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ with $a_i \in \mathbb{R}^d$, and a query point $b_j \in \mathbb{R}^d \wedge b_j \in \mathcal{B}$, where $\mathcal{A} = \mathcal{B}$, such that*

$$\|a_i - b_j\|_p \leq d \; , \tag{5.3}$$

output any point a_i in the set $\mathcal{A}$ within a distance d. The resulting family of subsets $\mathcal{C} = \{X_m \mid X_m \subseteq \mathcal{A}\}$ contains n subsets X_m with $m \in [1, n]$, where each X_m consists of the query point b_j and its nearest neighbour points $a_i \subseteq \mathcal{A}$.

The Fixed-radius Near Neighbor Search Problem is solved in a two-dimensional space, as the given H&E images of each specimen are two-dimensional. Hence, the eq. 5.3 can be simplified to the euclidean distance ($p = 2$). The nearest neighbour spots are determined for the entire set of query points $\mathcal{B}$. The resulting family of subsets $\mathcal{C}$ contains the nearest neighbour cytokine transcript-positive spots of each query point $b_j \in \mathbb{R}^2$ and the query point itself. If a query point does not have any nearest neighbour spots, the resulting subset X_m contains only a single element, the query point.

To connect the nearest neighbour cytokines transcript-positive spots $\mathcal{C}$, undirected graphs $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ are constructed. Let $G_i = (\mathcal{V}, \mathcal{E})$ be a connected graph, where $\mathcal{V}$ is a set of nodes derived from $\mathcal{C}$ and $\mathcal{E} \subseteq \{\{u_i, u_j\} \mid u_i, u_j \in \mathcal{V} \text{ and } i \neq j\}$ is a set of unordered edges. For each node $u \in \mathcal{V}$ in a graph G_i embedded in the two-dimensional tissue section space, a set of nearest neighbour responder signature transcript-positive spots $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$ is identified in the range of a user-defined radius $r^{(r)}$ by

$$\|s_m - u\|_2 \leq r^{(r)} \qquad\qquad \text{with } s_m \in \mathcal{S}.$$

A nearest neighbour spot s is added as a node to graph G_i if it is not already present in G_i, as denoted by the mathematical operation $G_i = (\mathcal{V} \cup \{s\}, \mathcal{E} \cup \{\cdot, s\})$.
To avoid unwanted artefacts in the correlation analysis due to nodes (spots) being present in multiple graphs, the graphs are merged, if they share subsets of nodes using the following definition:

Definition 4. *Let $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ be a set of undirected graphs and each graph consists out of at least one clique, i.e., a fully connected subgraph in a graph G_i. A union of two or more graphs, $G_i \cup G_j = (\mathcal{V}_i \cup \mathcal{V}_j, \mathcal{E}_i \cup \mathcal{E}_j)$, can be build if they share at least one common node. The set of joined graphs $\mathcal{G}^*$ is denoted as*

$$\mathcal{G}^* = \bigcup_{i=1}^{n} G_i$$

subject to the conditions

$$G_i \neq G_j \;\; \forall i \neq j$$
$$\mathcal{V}(G_i) \cap \mathcal{V}(G_j) \neq \emptyset \;\; \textit{or} \;\; \mathcal{E}(G_i) \cap \mathcal{E}(G_j) \neq \emptyset \;\; \forall i \neq j \; .$$

Then, $\mathcal{G}^*$ is the union of all sets of undirected graphs G_i that are not disjoint. This means none of the graphs are identical to each other and that they share at least one node or edge.

Conclusively from the definition 4, sets of separated graphs $\mathcal{G}'$ are defined by

$$\mathcal{G}' = \mathcal{G} - \mathcal{G}^*,$$

which updates the set of graphs $\mathcal{G}$ to $\mathcal{G}^\dagger$ by

$$\begin{aligned} \mathcal{G}^\dagger &= \mathcal{G}' \bigcup \mathcal{G}^* \\ &= \left(\mathcal{G} - \bigcup_{i=1}^{n} G_i\right) \cup \left(\bigcup_{i=1}^{n} G_i\right). \end{aligned}$$

Then the final set of graphs is given by $\mathcal{G}^{\dagger}$, containing cytokine transcript-positive spots and their corresponding responder signature in a predefined radius $r^{(r)}$ on a specimen. In addition, attributes, such as UMI-counts of a cytokine, sum of the UMI-counts of the responder genes signature, number of cytokines in a cluster, tissue layer, specimen, disease, and patient ID, are assigned to each node in $\mathcal{G}^{\dagger}$. The final set of graphs $\mathcal{G}^{\dagger}$ are denoted as spatial clusters. An example on the cluster construction for different radii is shown in Figure 5.11 a. The clusters are identified for each specimen individually. Moreover, the aggregated cytokine and responder signature genes UMI-counts in a cluster are used to calculate the weighted SCC (Section 2.3.2). Subsequently, the radius of action can be determined.

Determination of a cytokine's radius of action. In order to determine the radius of action, the weighted SCC has to be calculated for various radii $r^{(r)} = \{r_1^{(r)}, \ldots, r_n^{(r)}\}$. The weights w_j in the weighted SCC are set to the number of cytokine transcript-positive spots in a cluster $\mathcal{G}_j^{\dagger}$. The counts are then ranked using eq. 2.9. Subsequently, the weighted SCC and its p-value, for the radii $r^{(r)}$, are calculated.

A cytokine's radius of action $r^{(a)}$ is determined considering only the significant correlation coefficients. The radius, associated with the highest correlation value, is designated as the radius of action $r^{(a)}$ of a cytokine, as defined by the formula

$$r^{(a)} = \max_{r_i \in r^{(r)}} \rho^{(w)}(r_i^{(r)}) \quad \text{with } \rho^{(w)} = \{i \mid p_i \leq 0.05\}.$$

In summary, $r^{(a)}$ is the radius that results in the highest possible significantly weighted SCC, defining a cytokine's local impact radius on a specimen.

5.3.3.2 Application of spatial clustering identifies local hotspots

To investigate the impact radius of a disease hallmark cytokine in the epidermis, I applied the developed density-based spatial clustering. The clustering of *IFNG*, *IL13*, and *IL17A* their corresponding response signature genes was performed for the radii $r^{(r)} \in \{0, 1, \ldots, 9\}$. A radius of $r^{(r)} = 0$ considers a cytokine and its induced responder genes within the same spot, whereas a radius of $r^{(r)} > 0$ includes responder signals from neighbouring spot. Weighted Spearman correlation was then calculated for each cytokine and its associated responder signature across the defined radii. This approach allowed for a detailed characterisation of cytokine-specific spatial correlations and their micro-environmental reach.

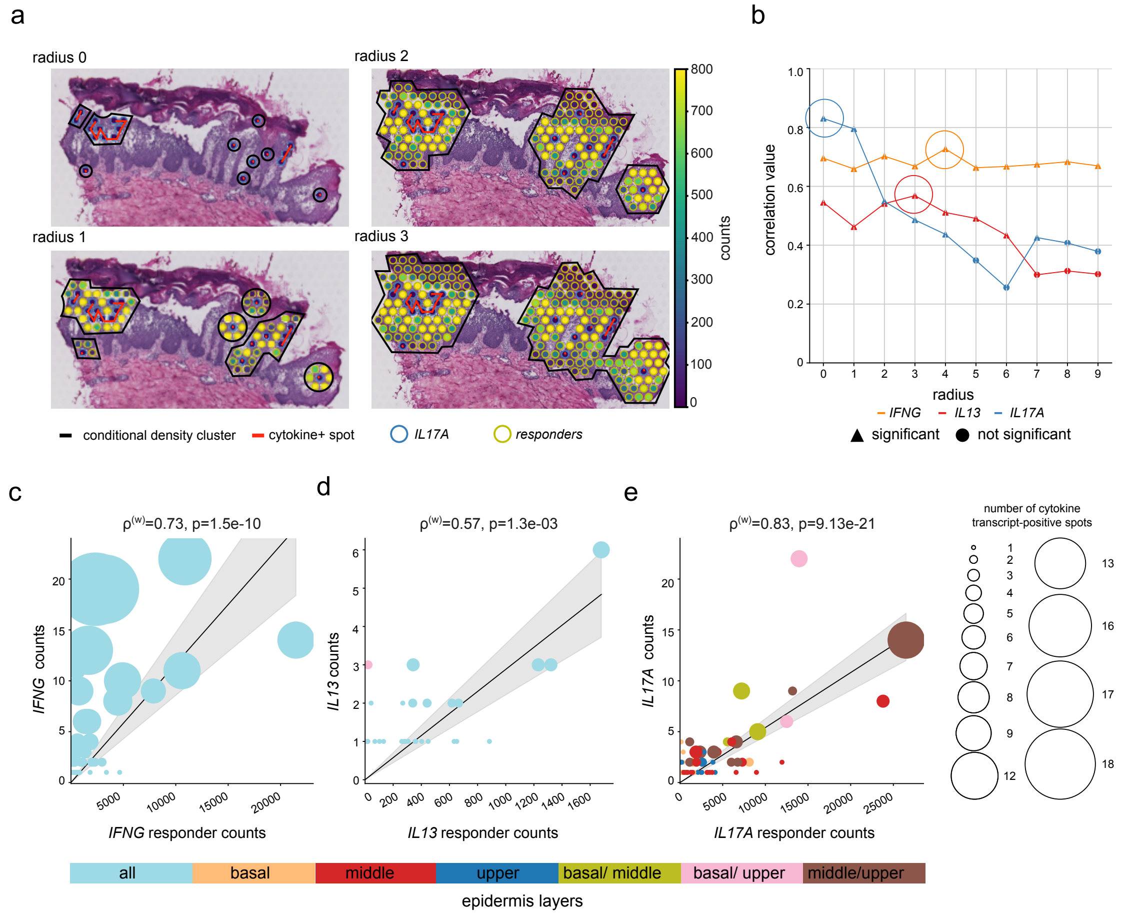


Figure 5.11: Spatial correlation between hallmark cytokines and their corresponding response signature genes transcripts. a) Example of the density-based clustering algorithm on a representative lesion psoriatic specimen for radii 0−3 and cytokine *IL17A*. A cluster is outlined by a black line and nearest neighbour *IL17A* transcript-positive spots are connected by a red line. The nearest neighbour responder signature transcript-positive spots are outlined by a yellow circle. In each spot the sum of the responder or *IL17A* transcripts is shown. The colour indicates the number of transcripts. **b)** Identification of the radius of action for each hallmark cytokine in the range of $0 - 9$. The highest significant weighted SCC value is highlighted by a circle and marks the radius of action. Weighted SCC analysis of *IFNG* **(c)**, *IL13* **(d)**, or *IL17A* **(e)** density clusters. Each data point represents a cluster and the size indicates the number of cytokine transcript-positive spots in a cluster and the colour the location on the skin section. Figure adapted and modified from [Sch+22]. The black solid **(c-e)** line is fit by an ordinary least square model. The grey area shows the $\mathrm{CI}_{95\%}$. Figure adapted from [Sch+22].

For each hallmark cytokine, specific radii of action were identified (Figure 5.11 b). A near-constant correlation between *IFNG* and its responder signature was observed, peaking at radius $r^{(r)} = 4$ with $\rho^{(w)} = 0.73; p = 1.5e\text{-}10$. The correlation of *IL13* exhibited a steady increase until its maximum at $r^{(r)} = 3$ with $\rho^{(w)} = 0.57;\ p = 1.3e\text{-}03$, before declining and stabilising at non-significant p-values by $r^{(r)} = 7$. Interestingly, *IL17A* showed the highest correlation of $\rho^{(w)} = 0.83; p = 9.13e\text{-}21$ in its direct micro-environment ($r^{(r)} = 0$), followed by a decline $r^{(r)} \leq 7$ and an increase for radii $r^{(r)} > 7$. The results of the correlation analysis, based on my density-based spatial clustering, revealed that the hallmark cytokines have distinct radii of action, indicating their specific influence on their micro-environment.

Additionally, I examined the distribution of the clusters within the epidermal layers using the optimal radii of 4, 3, and 0 for *IFNG*, *IL13*, and *IL17A*, respectively (Figure 5.11 c-e). While clusters of *IFNG* and *IL13* were primarily distributed throughout the entire epidermis, clusters of *IL17A* appeared in specific epidermal layers. Due to the naturally thicker epidermis in psoriatic lesions, more finely resolved clusters were found within the epidermis. In lichen planus and eczema, the epidermis was comparatively thin, corresponding to one or two rows of spots in a tissue section. Considering these observations, it became evident that in lichen planus and eczema, the clusters were distributed throughout the entire epidermis for radii $r^{(r)} \geq 2$. In essence, the spatial clusters of *IFNG* and *IL13* were spread throughout the epidermis, while *IL17A* clusters were confined to specific layers.

I further analysed the impact of the cytokine transcripts on the immune response within the identified spatial clusters. Interestingly, only up to 15 cytokine transcripts (*IFNG*: 1 to 8, *IL13*: 1 to 3, *IL17A*: 1 to 15 UMI-counts/spot) were measured in total in the identified clusters. These were able to induce up to $25,000$ responder transcripts in the close vicinity of *IFNG*, *IL13*, and *IL17A*. This indicates a notable amplification of the immune response by relatively low numbers of cytokine transcripts, leading to increased inflammation in the tissue.

In summary, leveraging the density-based clustering, I determined the radius of action for each hallmark cytokine, *IFNG*, *IL13*, and *IL17A*. These cytokines are able to induce localised immune response cascade.

5.3.3.3 Data-driven expansion of cytokine associated genes

The Visium technology provided insights into the direct micro-environment of cells, thereby enabling an assessment cytokine induced activity within the tissue. This raised the question of whether the clustering algorithm could reveal additional genes potentially associated

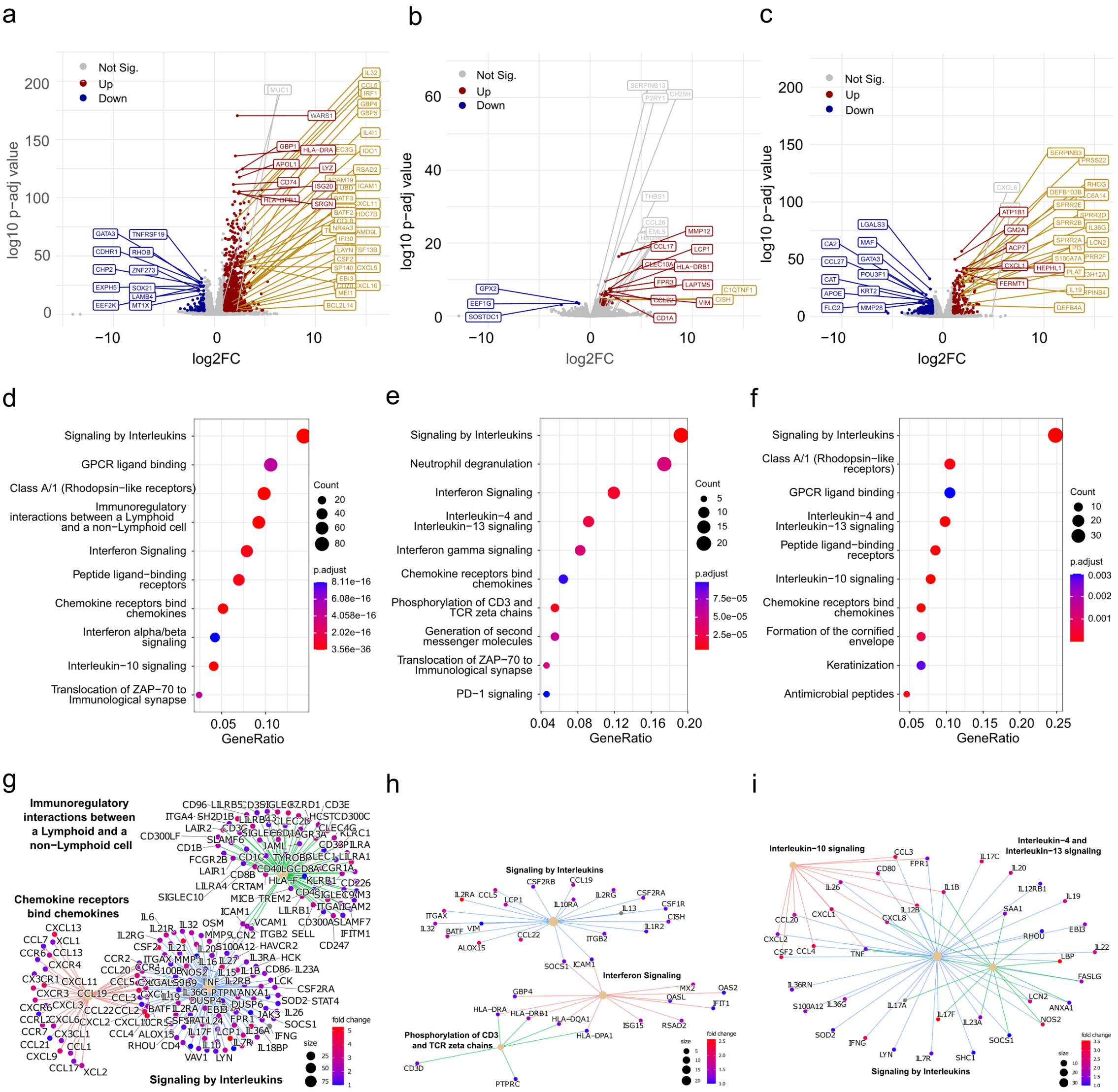


Figure 5.12: The spatial clustering algorithm identifies additional cytokine associated genes and potential drug targets. **a-c)** DGE analysis between cytokine transcript-positive spots and cytokine transcript negative-spot outside of the range of the radius of action of each hallmark cytokine, which is 4, 3, and 0 for *IFNG* (**a**), *IL13* (**b**), and *IL17A* (**c**), respectively. Golden and grey coloured genes mark are significantly upregulated and non-significant responder signatures, respectively. A gene is denoted as DEG if it meets the threshold of padj value < 0.001 and absolute log2FC > 1. **d-f)** Top 10 enriched pathways identified by ORA of upregulated DEGs (padj value < 0.05 , log2FC > 1) in *IFNG*, *IL13*, and *IL17A* transcript-positive spots. **g-i)** The top 3 pathways selected by padj value and their associated genes. Figure adapted from [Sch+22].

with cytokines beyond the predefined responder signature genes. To address this question, I provided raw counts from the preprocessed count matrix as input to the clustering algorithm along with the optimal radii 4, 3, and 0 for *IFNG*, *IL13*, and *IL17A*, respectively. The Density-based clustering was then applied to the epidermal spots, identifying cytokine transcript-positive clusters for downstream analysis. This data-driven approach demonstrated the potential to uncover additional genes within cytokine-enriched micro-environments.

To identify potential cytokine-associated genes, DGE analysis was performed, comparing cytokine transcript-positive spots with other epidermal spots outside of any respective cytokine clusters. This analysis identified 974 *IFNG*-related, 148 *IL13*-related, and 228 *IL17A*-related upregulated DEGs in the cytokine transcript-positive spots. Notable genes included *SRGN*, *LYZ*, and *CCL17* for *IFNG*; *CLEC10A* for *IL13*; and *GM2A* for *IL17A*, respectively (Figure 5.12 a-c). In addition, genes from cytokine response signatures derived from the literature and keratinocyte experiments were observed. This indicates that additional cytokine-associated genes could be indeed identified using a data-driven approach.

The upregulated DEGs were subjected to an ORA to elucidate pathways highly activated in cytokine transcript-positive spots. Beyond previously identified pathways (Figure 5.5 e, Figure 5.7 b, d), other enriched pathways were observed, such as "GPCR ligand binding" and "Translocation of ZAP-70 to Immunological synapse" for *IFNG*; "Neutrophil granulation", "Interferon Signaling", "Phosphorylation of CD3 and TCR zeta chains", "generation of second messenger molecules", and "PD-1 signaling" for *IL13*; and "GPCR ligand binding" and "Antimicrobial peptides" for *IL17A*, respectively (Figure 5.12 d-f). These results suggest that additional genes could be identified that are involved in the activation and regulation of T cells, the production of cytokines, and the immune response.

In summary, the combination of spatially resolved transcriptomes and my density-based clustering algorithm allowed to gain insight into the pathogenesis of cytokines and their radius of action in the immediate micro-environment. In general, a higher induction of the immune response was observed in epidermal regions, as there were more cytokine secreting cells measured compared to areas with less or no cytokines. In addition, the study showed that even low numbers of cytokine transcripts can trigger enormous inflammatory cascades in localised epidermal clusters, confirming the stated hypothesis.

5.4 Summary and discussion

This chapter focused on investigating the ST landscape of the most common ncISDs, namely lichen planus, eczema, and psoriasis. Previous molecular studies of ncISD have advanced biomarker discovery and elucidated general biological mechanisms, yet effective treatments are lacking due to the incomplete molecular understanding of cellular interactions within the complex tissue micro-environment. Addressing this challenge requires high-resolution approaches capable of uncovering these mechanisms to guide the development of targeted therapies.

To achieve this, I analysed a large repository comprised of 31 L and 14 NL samples using 10x Visium. The expression levels and spatial patterns of the hallmark cytokines *IFNG*, *IL13*, and *IL17A* were characterised and validated using bulk RNA-seq, scRNA-seq, and ISH (objective (vi)). The micro-environment of cytokine-expressing leukocytes was examined and compared to single-cell leukocyte profiles (objective (vii)). Building upon these insights, I developed a density-based clustering algorithm to identify spatial clusters of cytokine and responder signature transcript-positive spots. Each cytokine's radius of action was determined by maximising the correlation between cytokine and responder signature transcripts within these clusters (objective (viii)). The findings indicate that a small subset of immune cells sustains an immune cascade in ncISDs, producing cytokine transcripts that induce thousands of pro-inflammatory responder transcripts in spatially localised patterns. These insights provide a granular understanding of the ncISD pathogenesis and provide a foundation for developing therapies to treat skin lesions with greater precision.

I investigated the expression characteristics of three hallmark cytokines, *IFNG*, *IL13*, and *IL17A*, using ST. Despite their crucial roles in immune regulation, only 372, 103, and 154 of *IFNG*, *IL13*, and *IL17A* transcript-positive spots, respectively, were detected in L skin. Moreover, these were predominantly located in the upper dermis and epidermis layers in L skin [CVS20] [Zha+22a]. Specifically, *IFNG* was expressed in the upper dermis and basal epidermis, while *IL17A* was primarily detected in the epidermis, consistent with the literature [EE18] [Ruc+11] [DRM11] [Uji+22]. Validation through bulk RNA-seq, scRNA-seq, and ISH confirmed the general low expression of these cytokines, while functional verification via IHC and flow cell cytometry demonstrated their activity and disease relevance. In summary, while the measured cytokine transcripts were limited in number, they exhibited disease-specific spatial expression patterns, offering insights into localised immune responses in the skin.

To characterise disease-driving leukocytes in L skin, I compared *IL17A* transcript-positive and transcript-negative spots. In *IL17A* transcript-positive spots, I observed an expected upregulation of type 3 immune response genes, such as *IL17F* and *IL26*. Additionally, genes induced by *IL17A*, such as *NOS2*, *CXCL8*, and *DEFB4A* were upregulated, which are markers of oxidative stress, neutrophil migration, and antimicrobial peptides (AMPs). Amongst the top enriched pathways was "Keratinization", which is commonly accelerated in psoriatic epidermis [CSM05] [Gal+18], suggesting exacerbated inflammation symptoms near *IL17A*-producing leukocytes. Notably, response signatures induced by *IL17A*-expressing leukocytes were not upregulated in leukocytes in the scRNA-seq data, demonstrating the advantages of ST in preserving spatial context. Thus, ST revealed that *IL17A* induces specific immune responses and enriches distinct pathways in its micro-environment in inflamed tissue.

I compared *IFNG* transcript-positive and transcript-negative spots, to further characterise the disease-driving leukocytes. In *IFNG* transcript-positive spots, I observed an upregulation of type 1 immune response associated genes, including *CXCL9*, *CXCL10*, and cytotoxic markers. Pathway enrichment analysis confirmed a dominant presence of immune-response-related pathways. Additionally, genes such as *FASLG* and *GZMB*, which are associated with cell death in *IFNG* expressing leukocytes [Lau+18] [Sha+], were also upregulated. This findings were partially consistent with the scRNA-seq data, where these cell death markers were upregulated, but type 1 immune response associated genes were downregulated. These results indicate that *IFNG*-expressing leukocytes are primarily involved in type 1 immune responses and cell death pathways, showing their dominant role in driving immune responses in skin lesions.

A biological characterisation of disease-driving leukocytes expressing *IL13* was conducted using the ST dataset. Type 2 immune response markers such as *CCL17*, *CCL19*, and *CCL22* were observed in leukocytes *IL13* transcript-positive spots. Commonly they are induced by *IL13* and *IL4*. However, *IL4* was not detected in the ST dataset, which could be due to the low sensitivity of the technology [SK23]. These findings demonstrate that, like *IFNG* and *IL17A*, *IL13* induces specific immune responses and enriches particular pathways within its micro-environment.

Responder signatures for *IL17A*, *IFNG*, and *IL13* were identified by integrating DEGs from in vitro stimulated primary human KCs, ST data, and literature. In total 21, 29, and 4 epidermal responder signature genes were determined for *IL17A*, *IFNG*, and *IL13*, respectively. The impact radius of cytokines in the epidermis was assessed by correlating their transcripts with their induced response signatures using the density-based clustering algorithm. By varying the clustering radius, the optimal impact radius was determined as the one yielding the highest weighted SCC. Results showed that cytokine-induced responses were strongest in the close micro-environment of *IL17A*, *IFNG*, and *IL13*, with a few cytokine transcripts inducing a thousand fold higher immune response. This strengthened the finding that a few cytokines drive inflammation.

A limitation of ST is its inability to resolve individual cells due to the mini-bulk resolution of each spot ($\varnothing\, 55\, \mu m$), which often captures transcriptomes from multiple cells. To address this, I applied Tangram, a computational tool that predicts cell type compositions within spots and corrects for ST's low sensitivity [Bia+21]. While Tangram identified multiple cell types in cytokine transcript-positive leukocyte spots, the potential influence of neighbouring or co-localised cells on the observed immune response markers remains unresolved and requires further investigation. Additional limitations include the spatial distance between spots ($100\, \mu m$) and the lack of a 3D context. These challenges complicate the analysis of inter-cellular interactions, spatial dynamics of the immune response, and influence of neighbouring cells on each others expression profiles. Single-cell resolved ST (e.g., 10x Genomics Xenium) offers a promising alternative, providing greater granularity [Liu+24]. Furthermore, integration of z-stacks could further enhance insights into tissue structure and immune response dynamics [MP22]. Addressing these challenges will advance our understanding of immune response dynamics in ncISDs.

The limitations of the density-based clustering algorithm are partly linked to the challenges of ST. Correlations between cytokines and responder signatures may be affected by co-existing cell types, markers, and cytokine transcript-positive spots containing multiple cytokines. Incorporating additional attributes into graph node could enable to account for these effects, thereby improving clustering accuracy. Furthermore, the algorithm currently operates in 2D, limiting its ability to fully capture the immune response dimensions. Extending the algorithm to 3D could better identify disease-promoting networks and inflammation hotspots, informing targeted therapies such as CAR T-cell treatments [SS21]. Another limitation is the circular clustering formation, which includes entire rings of nearest neighbour responder signature transcript-positive spots, potentially overlooking irregular inflammatory patterns. A more adaptive approach that selectively adds individual neighbours could better represent irregularly shaped inflammatory regions. Enhancing the clustering algorithm to operate in 3D and refining neighbour selection would improve the its ability to identify disease-promoting networks and guide precision therapies.

This study reinforces the utility of cytokine-induced responder signatures as biomarkers, providing further evidence to support their theragnostic applications in ncISDs. Recent examples include molecular classifiers that distinguish between psoriasis and eczema using *NOS2* and *CCL27* [Gar+16] [Qua+14] [Fis+23], as well as predictive biomarkers such as serum IL - 19 levels for anti-IL - 17 therapy response [Kon+19] [Kol+17]. Additionally, correlations between disease severity and markers such as *DEFB4A* in psoriasis and *CCL17*/*TARC* in eczema demonstrate the diagnostic and prognostic potential of responder signatures [Bak+20]. In summary, the findings from the characterisation of cytokine transcript-positive leukocyte spots using ST could inform biomarker discovery and contribute to optimising treatment strategies in ncISDs.

The density-based clustering algorithm developed in this study provides a promising approach for detecting mediators and their specific antigens that activate disease-driving immune cells. Identifying these immune cell activators could support the development of antigen-specific immunotherapies. Furthermore, the clustering algorithm is applicable to other diseases and tissues, such as cancer, where it could contribute to identifying immune hotspots and biomarkers. This broad applicability positions the algorithm as potential tool for advancing precision medicine across various diseases.

Precision medicine in oncology has shown the importance of targeting disease-driving mutations. For instance, in malignant melanoma, focusing on specific mutations that drive tumour growth and metastasis has significantly improved patient survival rates [Icg]. This study suggest a parallel in ncISDs, where a small subset of cytokine-producing immune cells maintains inflammation. As these cytokines form localised epidermal clusters, adapting precision medicine principles from oncology to ncISDs could disrupt this cycle by selectively targeting disease-driving immune cells, offering a potential strategy for curative therapies.

In summary, this study provided insights into the functional relevance of sparse and low-expressed cytokine transcripts in comparison to bystander immune cells. The findings were validated across various technologies. Additionally, the results of my density-based clustering algorithm revealed that cytokines can induce an enormous immune response cascade in localised, epidermal clusters. These insights may pave the way for innovative, targeted therapeutic strategies and contribute to the advancement precision medicine approaches for ncISDs.

Chapter 6

Discussion and Outlook

In non-communicable chronic inflammatory skin diseases (ncISDs), the established disease ontology often fails to capture the heterogeneity and complexity, leading to diagnostic inaccuracies and ineffective treatments. These limitations can have serious consequences for patients and contribute to increased healthcare costs. Precision medicine is a promising solution, enabling accurate patient stratification and customised treatment plans. However, despite advancements in stratifying patients into immune response patterns (IRPs), some patients remain non-responders to treatment, demonstrating the need for further exploration of additional molecular information in ncISDs. This will enhance the understanding of disease heterogeneity, reveal additional biological mechanisms, and uncover spatially organised inflammatory processes that shape the tissue micro-environment in ncISDs.

This thesis integrates transcriptomics, statistical methods, and machine learning (ML) to enhance precision medicine in ncISDs. Using a hypothesis-free approach and an automatised highly variable gene (HVG) selection pipeline, I stratified patients into 13 biologically and clinically characterised endotypes, revealing inaccuracies in current clinical diagnostics (objectives i-iii). I further hypothesised an association between psoriasis-like endotypes and drug response, training and validating gene-expression-based classifiers to distinguish these groups using a minimal set of single predictive markers (objectives iv, v). Additionally, spatial transcriptomics (ST) enabled spatial analysis of ncISDs, identifying the impact radius of disease-driving cells on their micro-environment (objectives vi-viii). These methodologies refine patient stratification, identify predictive biomarkers, and enhance understanding of the inflammatory micro-environment, contributing to precision medicine.

Summary of outcomes

Molecular clustering revealed 13 distinct ncISDs endotypes, showing the limitations of current clinical classifications. To improve patient stratification, I developed an automated HVG selection pipeline incorporating the highly relevant gene (HRG) method to enhance clustering efficiency, reduce dimensionality, and retain biologically relevant transcriptomics signatures. This approach does not require user-defined thresholds, thereby improving reproducibility. By determining the optimal number of clusters and applying the Leiden algorithm, I identified 13 molecular endotypes across 21 ncISDs (objective i). This data-driven approach refines disease classification by capturing molecular heterogeneity in a reproducible manner.

Patient stratification into molecular endotypes revealed that clinical classifications fail to capture underlying biological differences (objective ii). Phenotypically similar diseases, such as psoriasis and pityriasis rubra pilaris, showed distinct gene expression profiles, while clinically distinct conditions, like cutaneous lymphoma and parapsoriasis, shared molecular features with eczema. The co-clustering of rare and common diseases suggests shared molecular pathways, potentially enabling drug repurposing. Notably, eczema exhibited high heterogeneity across multiple endotypes, and four distinct psoriasis-like endotypes (E8, E11 - E13) were identified. These findings highlight the limitations of current disease classification and emphasise the importance of molecular profiling for improved patient stratification.

Endotype separation was primarily driven by metabolic (E5 - E10) and inflammatory (E1 - E4, E11 - E13) processes (objective iii). Metabolism-driven endotypes were predominantly associated with eczema, pityriasis rubra pilaris, and diseases with undefined (UD) IRPs, suggesting shared skin barrier dysfunction. In contrast, inflammation-driven endotypes were linked to psoriasis and diseases characterised by specific IRPs (e.g., IRP 1, IRP 2a, IRP 5). Notably, rare disease-associated endotypes (E6, E9, E10) exhibited distinct metabolic and keratinisation pathways, reinforcing their connection to barrier abnormalities. These findings demonstrate the role of metabolic and inflammatory pathways in ncISD endotypes and their potential as therapeutic targets.

The clinical presentation of certain endotypes revealed inconsistencies in current diagnostic approaches (objective iii), reinforcing the need for molecular diagnostics. E5 exhibited reduced pustule formation despite its association with pityriasis rubra pilaris . In contrast, E12 was the only endotype with clear psoriasis-like features. E3 and E4 aligned more closely with interface dermatitis diseases, representing the only cases where clinical and molecular profiles matched. These findings show the limitations of diagnosis based solely on clinical features, reinforcing the necessity of molecular diagnostics for accurate patient stratification and targeted therapies.

I also explored the potential implications of endotypes in guiding therapy selection by examining the relationship between psoriasis-like endotypes and drug response. Trends of correlations between IL - 23, IL - 17, and TNF - α inhibitors and E12/E13, E8/E11 and

E13 were observed, respectively (objective iv). The differences in drug response can be potentially be explained by biological differences in immune activation, protein synthesis, and keratinisation pathways. While these findings provide a promising step towards linking endotypes with drug response and identifying biomarkers for patient classification, the relatively small sample sizes in both the discovery (n=34) and validation (n=22) cohorts limited statistical power and generalisability. Studies involving larger cohorts are needed to confirm endotype-drug-target associations.

Building on insights into potential endotype-specific drug responses, I developed classifiers using a minimal set of predictive genes (objective v). To enhance feature selection, I introduced GeneSTRIVE, a methodology designed to process noisy and heterogeneous transcriptomics data, improving the biological relevance and robustness of selected features. Integrated into a triad feature selection pipeline, GeneSTRIVE reduced dimensionality from $17,816$ to less than 20 genes, an essential factor for future clinical translation. The resulting classifiers achieved high predictive accuracy, including an E12/E13 vs. E8/E11 classifier (19 genes, weighted F1-score=84.00%) and an E12 vs. E13 classifier (15 genes, weighted F1-score=87.50%), demonstrating their potential in precision diagnostics.

To further enhance our understanding of disease mechanisms within skin, I analysed a ST dataset. Focusing on the hallmark cytokines *IFNG*, *IL13*, and *IL17A* of disease-driving immune cells, I observed tissue layer-specific expression patterns (objective vi). Despite their low number of transcripts, these cytokines significantly shaped local immune environments, influencing immune activation, epidermal changes, and inflammatory cascades (objective vii). Using the density-based clustering algorithm on the ST dataset, I identified distinct cytokine-driven immune clusters (objective viii), reinforcing that a few cytokine transcripts ranging from 1 to 37, 1 to 12, and 1 to 27 transcripts/section for *IFNG*, *IL13*, and *IL17A*, respectively, can induce profound inflammatory responses. These findings refine our understanding of ncISD pathogenesis and provide the potential for spatially resolved molecular profiling for enhanced therapeutic interventions.

In summary, this thesis establishes a molecular framework for endotyping, offering insights into cytokine-driven immune responses and disease heterogeneity. By integrating transcriptomics, statistical modelling, and ML-based methods, it advances patient stratification, biomarker discovery, and our understanding of ncISDs, paving the way for precision medicine in dermatology.

Impact of research

The identification of molecularly distinct endotypes has significant implications for patient stratification and treatment in ncISDs. Clinical diagnoses often fail to capture molecular heterogeneity, limiting treatment precision. This thesis addresses this gap by developing a transcriptomics-based strategy for endotype identification, advancing precision medicine in dermatology. Stratifying patients into biologically meaningful subgroups enables targeted therapies, reduces misdiagnoses, and improves treatment responses. Additionally, the molecular composition of endotypes supports classifier development to distinguish clinically similar diseases, such as eczema, cutaneous lymphoma, and parapsoriasis [EG99] [Kik+93], as well as psoriasis and pityriasis rubra pilaris [EE18]. This approach also facilitates drug repurposing, particularly for rare diseases sharing molecular features with better-characterised conditions. Integrating molecular endotypes into ncISDs diagnostics and theragnostics could transform disease classification, paving the way for personalised treatment strategies.

The endotypes were identified using a strategy similar to previous studies, which clustered patients from a dimensionality-reduced dataset [AA+19] [DiN+22] [Kra+23b] [KVR21]. To enhance clustering efficiency, I incorporated the HRG approach into my automated HVG selection pipeline, outperforming established methods based on the mean-variance relationship [WAT18] and standard deviation. The pipeline uses accuracy (optional) and Davies-Bouldin Index (DBI) to ensure a data-driven selection of HVGs and automatically determines the optimal number. This improves reproducibility, reduces bias, and enhances the accuracy of downstream analyses. The pipeline is also broadly applicable to RNA-sequencing (RNA-seq) studies and enables the identification of biologically meaningful clusters.

The identification of psoriasis-like endotypes linked to specific drug targets can improve therapy selection. The triad feature selection method, including GeneSTRIVE, demonstrated the feasibility of selecting a minimal, biologically meaningful set of single genes for classification, offering a robust approach for biomarker discovery. This pipeline overcomes limitations of other methods, such as the "curse of dimensionality", noise, and risk of overfitting [Pud+22] [HK99] [Szy+09] [SKZ10], improving classification robustness and interpretability in high-dimensional transcriptomics datasets. The triad selection method provides markers for molecular diagnostics and therapy selection, thereby contributing to precision medicine.

This research provides a foundation for the development of artificial intelligence (AI)-guided therapy selection through endotype classification, instead of purely classifying diseases [Gar+16] [Kam+10] [Tso+19] [Fis+23]. By utilising classifiers able to handle unseen classes, dermatologists will be better equipped to make informed decisions, leading to more effective therapies tailored to each patient's molecular profile. In essence, the classifiers are a step towards precision medicine in ncISDs.

Beyond patient classification on the bulk RNA-seq level, this thesis advances the molecular characterisation of ncISDs by studying the spatially and single-cell resolved characteristics of disease-promoting cells. The identification of cytokine-induced local, epidermal immune response clusters could lead to the discovery of spatial biomarkers and development of targeted treatment strategies. While clustering approaches classify patients based on transcriptomics profiles, ST provides insight into the tissue-specific structure of these molecular signatures. This complementary approach refines disease classification by revealing spatial patterns of immune activation, thereby enhancing therapy selection and precision medicine strategies.

The density-based clustering algorithm can be also applied in other disease contexts studied using ST. For instance, it could be used to investigate various tissue structures, including sebaceous glands [Sei+24], and tissue-specific disease processes such as granulomas [Kra+23a]. Beyond dermatology this methodology holds promise for the field of oncology [Aro+23] [Bra+24], broadening the translational impact of this algorithm beyond dermatology. These examples show the versatile usage of this method potentially informing biomarker discovery.

Challenges and limitations

While this study provides valuable insights into the molecular endotypes of ncISDs, several limitations have to be considered when interpreting the findings. These limitations primarily concern dataset composition, methodological constraints in feature selection and classification, sequencing resolution, and the robustness of linking endotypes to drug response. Some challenges have been already discussed in detail in Chapters 3.4, 4.4, 5.4. The following provides a summary of the most critical aspects and outline future directions.

A limitation of this thesis is the restricted diversity of datasets, which include only 21 or three of over 100 defined ncISDs [EE18], with a bias towards psoriasis, eczema, and lichen planus. While biologically meaningful endotypes were identified (objective i), the disease composition may have influenced clustering, limiting endotype discovery and alignment with existing disease ontology (objective ii). A comprehensive evaluation of endotype-therapy-response associations would require inclusion of all endotypes, potentially revealing opportunities for drug repurposing (objective iv). Expanding the investigation of the inflammatory micro-environment beyond psoriasis, eczema, and lichen planus using ST could have provided broader insights into immune cascade dynamics across multiple diseases (objective viii). The under-representation of rarer diseases may have affected generalisability, highlighting the need for future studies to incorporate a wider spectrum of ncISDs (objective iii). Thus, integrating datasets from diverse sources would enhance patient stratification and improve the robustness of these findings.

While effective, the automated feature selection pipeline and clustering approach have methodological limitations. The predominance of psoriasis and eczema samples may have biased feature selection, potentially overlooking markers in under-represented ncISDs (objective i). Additionally, the classification models were optimised for psoriasis-like endotypes but not designed for unseen patient categories, limiting generalisability (objective v). To reduce misclassification risks, alternative strategies such as anomaly detection [Yan+24] or hierarchical classification [Sel+21] should be explored. These methods could enhance adaptability, better capture ncISD heterogeneity, and improve clinically applicable predictive models.

Establishing a robust link between endotypes and drug response (objective iv) remains a challenge, primarily due to the limited sample size in the used cohorts. While initial findings indicate potential associations, the statistical power of these results is constrained, and differences in experimental design, choice of read alignment tool [REE19], and the missing data imputation approach in the validation cohort may have introduced variability. Addressing these issues will be crucial for enhancing the reliability of drug response predictions. Thus, future work should prioritise the standardisation of the whole data process chain and the inclusion of larger cohorts to validate these associations.

ST has provided valuable insights into immune cell distributions in psoriasis, eczema, and lichen planus, yet it is limited by its resolution. The mini-bulk resolution of ST captures transcriptomes from multiple cells [Gen] [Stå+16], potentially confounding cell-type-specific expression patterns and restricting precise characterisation of disease-driving immune cells (objective vi) and their micro-environment (objective vii). Despite computational deconvolution using Tangram [Bia+21], the impact radius of disease-driving cells (objective viii) remains constrained by spot size, spatial arrangement, and the lack of a 3D perspective. Higher-resolution ST technologies, including single-cell and 3D methods, could enhance spatial characterisation and provide deeper insights into immune cell interactions in the skin.

The density-based clustering algorithm used to identify immune response clusters is constraint by the presence of multiple cell types within transcript-positive spots. These may confound the correlation between cytokines and their corresponding responder signature genes as well as obscure individual immune cell contributions (objective vii). As the algorithm operates in two dimensions, limiting its capacity to capture the complexity of the tissue micro-environment. A 3D framework could provide a more accurate representation of inflammatory networks and disease progression. Moreover, the current clustering method adds entire rings of nearest-neighbour transcript-positive spots, which may not always align with the irregular shapes of inflammatory regions in biological tissues. Refining the algorithm to selectively include neighbouring spots would improve spatial clustering accuracy and enhance the identification of biologically relevant inflammatory networks.

While these limitations present challenges, this thesis contributes towards the realisation of precision medicine in ncISDs. Further research is required to refine the endotypes, enhance the generalisability of classification models, and improve the spatial characterisation of the immune response. Expanding datasets to include a broader spectrum of ncISDs, optimising feature selection methodologies and classifiers, as well as integrating higher-resolution transcriptomics technologies will enable developing more precise diagnostic and therapeutic strategies in dermatology.

Future opportunities

The identification of endotypes has significant implications for diagnosis, patient stratification, and therapy selection. This thesis presents an automatised HVG selection framework, enabling the clustering of patients into previously unrecognised molecular subgroups and revealing limitations in clinical classifications. These endotypes could refine diagnostic frameworks, enhance patient stratification in clinical trials, and inform therapy selection [Ozd+18]. Furthermore, the ability to distinguish molecularly distinct patient groups has direct implications for drug repurposing, particularly for rare diseases with molecular similarities to better characterised conditions. Collectively, these findings advance the understanding of ncISDs and contribute to a biologically informed classification system.

The complexity and heterogeneity of ncISDs highlight the necessity of molecular diagnostics. Endotype-based patient stratification has the potential to transform ncISDs management, shifting from descriptive diagnostics to molecular-driven precision medicine. As transcriptomics profiling becomes more accessible and cost-effective, molecularly defined endotypes could become integral to routine diagnostics, improving treatment selection while minimising adverse effects.

Future research could refine endotype classifications by integrating single-cell RNA-sequencing (scRNA-seq) data or multi-omics approaches, as suggest for asthma [RDW22] [SN17] and Parkinson's disease [Mih+24]. Expanding ncISDs endotype definitions to include data from other diseases may further identify cross-disease endotypes. For example, integrating cancer and ncISDs datasets could reveal shared molecular pathways, given the role of chronic inflammation in tumorigenesis [Cią+21]. This integrative approach may reveal common biological mechanisms and biomarkers, enabling drug repurposing and refining patient stratification. Thus, considering both shared and distinct molecular characteristics alongside clinical phenotypes, treatment strategies could be optimised to enhance therapeutic outcomes.

Endotype classification will further drive the realisation of precision medicine in ncISDs and inform therapy selection. By leveraging ML models trained on endotype-specific gene expression signatures, this research has shown the possibility of predicting patient subgroups that are more likely to respond to specific biologic treatments, such as IL - 23, IL - 17, and TNF - α inhibitors. After addressing the limitations of these classifiers, they could be integrated in clinical settings to assist in patient stratification and treatment selection. Predicting therapy response based on molecular profiles may improve treatment efficacy, reduce adverse effects, and lower healthcare costs [Mal+20] [Özm+19]. As these models are further refined and validated, their incorporation into the diagnostic process could enhance clinician confidence in patient centred treatment strategies.

The development of advanced ML models and the integration of diverse data sources hold great potential for advancing precision medicine in ncISDs. The increasing availability of RNA-seq data provides opportunities to adapt foundation models, such as CancerGPT [Li+24], to predict drug synergy and response in ncISDs. Integrating data sources, such as miRNA profiles [WD+18] [Zib+10], proteomics [Yan+23] [Man+21], scRNA-seq data [Wu+24], and time-course studies [Joh+20] [Mih+24], may enable the discovery of biomarkers and support drug repurposing [IS23]. This integration could enable the creation of digital twins for monitoring disease progression and treatment response

[VRK22]. However, clinical implementation requires data privacy measures and the standardisation of data acquisition, processing, and analysis protocols [Bla+18] [SKK20] [Man+21]. Once these challenges are addressed, these models could greatly benefit clinical practice. In summary, as data collection grows, the development of ML models and the identification of biomarkers will drive precision medicine in ncISDs.

ST and scRNA-seq can enhance endotype characterisation by providing insights into immune micro-environments and cytokine-induced immune response clusters. While bulk RNA-seq has identified molecular subtypes, it cannot capture cell-cell interactions or layer-specific immune activities in the skin. scRNA-seq addresses this limitation by providing cell type composition, either by transferring endotype labels from bulk RNA-seq or using deconvolution methods like MuSiC [Wan+19] and BayesPrism [Chu+22]. It also reveals intercellular communication [STM15] [Dim+22]. Additionally, ST refines endotype classification by offering insights into spatial immune environments, cytokine distributions, and tissue niches [FST23]. Combining scRNA-seq and ST may reveal previously undetectable disease subtypes, IRPs, and inflammatory mechanisms. In the short term, cross-referencing endotype signatures with single-cell and spatial immune interactions could identify biomarkers linked to therapy response, while classifiers based on bulk, scRNA-seq, and ST data may improve endotype prediction. These techniques could refine endotype definitions, highlighting specific cytokine clusters or interactions that distinguish therapy responders from non-responders, thus advancing patient stratification.

In the long term, integrating ST data into clinical workflows could transform dermatological diagnostics by enabling spatially informed therapy selection. As ST technologies advance, combining single-cell ST with multi-omics approaches like proteomics [Yan+23] [Man+21] and metabolomics [Gow+08] will offer a comprehensive molecular landscape of ncISDs, facilitating the discovery of therapeutic targets also for more rare subtypes. ST-based profiling, combined with advanced deep learning models, could enable longitudinal monitoring of immune dynamics [Zha+22b] [VRK22], allowing for timely treatment adjustments. Identifying spatial biomarkers could lead to spatial target-based drugs, enhancing diagnostics, therapy selection, and patient outcomes [Zha+22b]. Additionally, ST may help identify biomarker locations in the upper skin layers, potentially reducing the need for deep skin biopsies and enabling the use of less invasive sampling techniques such as tape strips or microbiopsies [Fis+23]. In the future, ST may bridge the gap between skin-layer resolved molecular profiling and actionable clinical insights, advancing precision medicine in dermatology.

Exploring the immune hotspots identified by my density-based clustering algorithm provides insights into cellular interactions and offers the potential for discovering biomarkers. The clustering algorithm could be also used to identify spatially resolved immune clusters of endotypes in the skin and inform about their distribution. Investigating tissue structures such as sebaceous glands, which contribute to inflammation in ncISDs [Sei+24], will further enhance disease understanding. Additionally, ST could be also used to investigate inflammatory environments, such as granulomas. In cancer, ST has identified tumour core and edge profiles, aiding in survival prediction and therapy response in oral squamous cell carcinoma and high-grade serous ovarian carcinoma [Aro+23] [Den+24]. Studies on 3D molecular alterations in pancreatic cancer [Bra+24] show the potential for early detection, which could also apply to recurrent ncISDs. In summary, ST provides valuable insights into the cellular composition and inflammatory micro-environment in ncISDs, advancing disease understanding and treatment strategies.

Chapter 7

Conclusion

Precision medicine remains an unmet need in the field of non-communicable chronic inflammatory skin diseases (ncISDs), as the established disease ontology does not fully capture their heterogeneity and complexity. To address this gap, I introduced approaches for patient stratification and artificial intelligence (AI)-guided therapy selection, while also exploring the inflammatory micro-environment of disease-promoting immune cells.

This thesis proposed a hypothesis-free, data-driven framework for patient stratification, identifying 13 endotypes. Additionally, I identified a minimal set of less than 20 robust molecular markers capable of distinguishing between psoriasis-like endotypes, which I hypothesised are linked to therapy response. Complementing these findings, the exploration of the spatial transcriptomics (ST) landscape provided insights into localised immune responses cascades, maintained by low numbers of cytokine transcripts. These findings contribute to advancing precision medicine in ncISDs, offering potential diagnostic and therapeutic applications.

The identified endotypes hold clinical potential, particularly in explaining therapy response variability, although further validation through larger, more diverse cohorts is necessary. These endotypes also deepen our understanding of rare disease subtypes and provide a foundation for molecular diagnostics. The application of single-cell and spatially resolved data will allow for a more comprehensive characterisation of these endotypes, revealing their cellular heterogeneity, intercellular communication, and the localised inflammatory mechanisms driving disease.

These findings enhance our understanding of ncISD heterogeneity, representing a step towards precision medicine. Further research will be essential to confirm the robustness of endotypes in diverse patient populations and to refine AI-guided therapy strategies. The integration of multi-omics, advanced AI models, and biomarkers will drive the development of targeted therapies and improve ncISD management through molecular diagnostics, facilitating the realisation of precision medicine in ncISDs. These advancements will enhance patient care by enabling more personalised, effective treatment strategies.

Appendix A

Further statistical tests

A.1 Parametric tests

The parametric test are used, when the data fulfils the criteria of normality and being numerical [Sid57].

The one-sample Student's t-test requires the data to be continuous and normality distributed. It tests whether mean μ of the sample, drawn from the population, is significantly different from the population mean μ_0 [Ueb]. The null hypotheses $\mathcal{H}_0$ can be tested, using the t-distribution. Hence, the t-test is defined by [Stu08]

$$t = \sqrt{n} \cdot \frac{\mu - \mu_0}{\sigma}\,, \tag{A.1}$$

where n is the number of samples in a population and σ is the standard deviation of the population. It should be noted that the t-test follows t-distribution with $n-1$ degrees of freedom.

The one-way analysis of variance (ANOVA) test analyses the variance differences between groups and can handle interactions between observations [Ayl25]. In order to apply the two-sample, one-way ANOVA test, the data must be continuous, normally distributed, have equal variance across groups, and consist of independent samples. The name one-way originates from the input which is a single, independent variable $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ containing at least two population groups ($k \geq 2$) and a dependent variable $\boldsymbol{y}$. It examines the effect of $\boldsymbol{X}$ on $\boldsymbol{y}$ and whether the means of the population groups differ, with the hypothesis:

$$\begin{aligned} &\mathcal{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k \\ &\mathcal{H}_1 : \text{Not all sample means are equal}\,. \end{aligned}$$

The test statistic is defined by

$$f = \frac{\sum_{j=1}^{k} \left(\frac{n_j(\mu_j - \mu)^2}{k-1}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{k} \left(\frac{(x_i - \mu_j)^2}{n-k}\right)}\,. \tag{A.2}$$

For two groups, both the t-test and one-way ANOVA can be used [Ayl25].

In case that two populations have unequal variance, the one-way ANOVA test cannot be applied. Instead, the Welch's t-test can be leveraged using the hypothesis

$$\mathcal{H}_0 : \mu_1 = \mu_2$$
$$\mathcal{H}_1 : \mu_1 \neq \mu_2 \ .$$

The Welch's t-test is defined by [Wel47]

$$t_w = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_{x_1}^2 - \sigma_{x_2}^2}} \ , \tag{A.3}$$

where σ_{x_i} is the standard deviation of sample $i \in [1, 2]$.

A.2 Non-parametric tests: Handling ties in ranks

For tied observations, the following formula of the Kruskal-Wallis test should be used [Sid57]

$$H = \frac{\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum t^3 - t}{N^3 - N}} \ , \tag{A.4}$$

where t is the number of tied observations in a tied group of scores.
It is assumed that the sampling distribution can be approximated as χ^2 with $df = k - 1$ for large samples sizes (n_j).

The Mann-Whitney U test assumes that the observations in the pooled set of both samples underlay a continuous distribution [Sid57]. In tied observations this assumption is violated. Ties can be handled by assigned the mean of the rank which they would have been assigned to in case of no ties. For tied observations the formula of the z-value is [Sid57]

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\left(\frac{n_1 n_2}{N(N-1)}\right)\left(\frac{N^3 - N}{12} - \sum \frac{t_3 - t}{12}\right)}} \ , \tag{A.5}$$

where N is the total samples size $n_1 + n_2$ and t is the number of observations tied for a given rank.

Appendix B

Further machine learning algorithms and metrics

B.1 Proof of relation between $\mathbf{R_2}$ and $\mathbf{r_p}$

Theorem 1 (Equivalence of R^2 and r_p)**.** *Let $\boldsymbol{x}$ be an independent and $\boldsymbol{y}$ a dependent variable. Their relationship is described by a simple linear regression model*

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j \quad \textit{with } j \in [1, \ldots, n] \ ,$$

where n is the samples size. The model parameters β_0 and β_1 can be estimated by minimising the error ϵ (eq. 2.15) which is similar to solving the ordinary least square equation. It aims for minimising the residual sum of squares (RSS) which is defined as

$$\begin{aligned} RSS &= \sum_{j=1}^{n} \left(y_j - \hat{y_j} \right)^2 \\ &= \sum_{j=1}^{n} \left(y_j - (\hat{\beta_0} + \beta_1 x_j) \right)^2 \end{aligned} \tag{B.1}$$

Then the coefficient of determination R^2 equals the squared Pearson's correlation coefficient (PCC) r_{xy}^2 between $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$R^2 = r_{xy}^2 \ . \tag{B.2}$$

Proof. The minimum of RSS is defined as the zero of the derivate function w.r.t β_0 which is $0 = 2\big(\sum_{j=1}^{n} \beta_1 x_j - y_j + \beta_0\big)$. Thus, the coefficient β_0 can be calculated by

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \ , \tag{B.3}$$

where the means are defined as $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ and $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$. In order to determine the β_1 coefficient the zero of the derivate of RSS w.r.t. β_1 is solved which is $0 = 2\big(\sum_{j=1}^{n} \beta_1 x_j^2 - x_j y_j + \beta_0 x_j\big)$. This can be resolved after β_1 by

$$\begin{aligned} \beta_1 &= \frac{\sum_{j=1}^{n} x_j y_j - \frac{1}{n} y_j \sum_{j=1}^{n} x_j}{\sum_{j=1}^{n} x_j^2 - \frac{1}{n} x_j \sum_{j=1}^{n} x_j} \\ &= \frac{\sum_{j=1}^{n} \left(x_j - \overline{x} \right) \left(y_j - \overline{y} \right)}{\sum_{j=1}^{n} \left(x_j - \overline{x} \right)^2} \end{aligned} \tag{B.4}$$

Recalling the correlation coefficient formula (eq. 2.5), which is defined by the covariance and the standard deviation of $\boldsymbol{x}$ and $\boldsymbol{y}$. It can be rewritten to

$$r_{xy} = \frac{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)\left(y_j - \overline{y}\right)}{\sqrt{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2} \sqrt{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2}} \,. \tag{B.5}$$

Rewriting eq. B.4 the relation to r_p becomes visible

$$\begin{aligned} \beta_1 &= \left(\frac{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)\left(y_j - \overline{y}\right)}{\sqrt{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2} \sqrt{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2}} \right) \cdot \left(\frac{\sqrt{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2}}{\sqrt{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2}} \right) \\ &= r_p \cdot \frac{\sqrt{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2}}{\sqrt{\sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2}} \\ &= r_{xy} \cdot \frac{\sigma_y}{\sigma_x} \end{aligned} \tag{B.6}$$

The solutions for the coefficients, eq. B.3 and eq. B.6, can be inserted into eq. 2.16 which rewrites to:

$$\begin{aligned} R^2 &= \frac{\sum_{j=1}^{n} \left(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_j - \overline{y}\right)^2}{\sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2} \\ &= \hat{\beta_1}^2 \frac{\frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2}{\frac{1}{n-1} \sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2} \\ &= \left(r_p \cdot \frac{\sigma_y}{\sigma_x}\right)^2 \cdot \frac{\frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \overline{x}\right)^2}{\frac{1}{n-1} \sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2} \\ &= r_{xy}^2 \end{aligned} \tag{B.7}$$

Thus, the coefficient of determination is equal to the squared PCC between x and y.

□

B.2 Linear regression models

Lasso regression, or Least Absolute Shrinkage and Selection Operator [TH09], is a linear regression model that applies shrinkage methods to reduce the contribution of less important features. In comparison to other feature selection methods, which focus on subsets of observations, shrinkage methods are less sensitive to noise [TH09]. Both approaches improve model interpretability and are considered embedded feature selection techniques. Lasso regression minimises the RSS while penalising it with the L1 lasso penalty. Let

$\boldsymbol{X} \in \mathbb{R}^{n \times f}$ be a data matrix and $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$. Then, the coefficients β can be determined by

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{n} |\beta_j| \right\} ,$$

where $\lambda \geq 0$ is the cost factor, controlling the amount of shrinkage, with larger values resulting in greater shrinkage. The term $\sum_{j=1}^{n} |\beta_j|$ is the L1 lasso penalty, also known as L1-regularisation [TH09].

Ridge regression, analogous to Lasso regression, minimises a penalised RSS. The coefficients β are defined by [TH09]

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \right\} , \tag{B.8}$$

where β_j^2 is the quadratic penalisation term, and $\lambda \geq 0$ controls shrinkage, with lager values resulting in greater shrinkage and reducing the contribution of features towards zero. This penalty is also known as L2-regularisation and is referred to as weight decay α in neural networks (NNs) [TH09].

The principal components (PCs) regression leverages principal component analysis (PCA) to create uncorrelated linear approximations of the data matrix $\boldsymbol{X}$ (Section 2.4.1), reducing its dimensionality from $\boldsymbol{X}^{n \times f}$ to $\hat{\boldsymbol{X}}^{n \times q}$, where $\boldsymbol{X} \approx \hat{\boldsymbol{X}}$. This ensures independence of features by using PCs and reduces overfitting.
Similar to Lasso regression, it regularises the model by using only q subsets of PCs. The number of PCs is determined as described in Section 2.4.1. PCs regression is a linear model for predicting output variables $\boldsymbol{y}^{n \times 1}$ using PCs as predictors, as described by [TH09]

$$\hat{y}_q = \bar{y}\mathbf{1} + \sum_{k=1}^{q} \hat{\theta}_k \hat{\boldsymbol{X}}_k , \tag{B.9}$$

where $\hat{y}_q$ is the predicted label using q PCs, $\bar{y}\mathbf{1}$ is the mean of the response variable, and $\hat{\theta}$ contains the estimates of the coefficients of PCs. The eq. B.9 can be rewritten to

$$\hat{y}_i = \theta_0 + \sum_{j=1}^{f} \left(\sum_{k=1}^{q} \hat{\theta}_k \phi_{jk} \right) x_{ij} = \sum_{j=1}^{f} \hat{\beta}_j x_{ij} ,$$

which is similar to the linear regression model (eq. 2.14). A limitation of PCs regression is that its prediction accuracy may be lower than other methods, as it uses an approximation rather than the full data.

B.3 Model evaluation metrics

Accuracy is the commonly used metric to assess the performance of a binary classifier by

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \, , \tag{B.10}$$

Accuracy is used when the classes are equally represented in the dataset, otherwise the weighted accuracy or another metric should be leveraged to assess a models performance [Tha20].

Precision makes a statement about how many of the positively labelled samples are actually TP by

$$\text{precision} = \frac{TP}{TP + FP} \, . \tag{B.11}$$

It should be used if one wants to be certain about the prediction of positive events [Tha20].

Recall or sensitivity makes statements about how reliably the model predicts the positive class

$$\text{recall} = \frac{TP}{TP + FN} \tag{B.12}$$

Recall should be used when the aim is to predict as many positive events as possible [Tha20].

Depending on the goal, particularly when aiming to classify rare events (TP), precision and recall should be used for model performance evaluation. For example, when determining if a patient should be treated with a drug, an incorrect decision (FP) could have severe consequences. In such cases, prioritising precision is crucial to ensure fewer cases of FP. However, this might lead to many actual responders being classified as FN. In essence, the balance between precision and recall should align with the specific objectives and conditions of the task [Tha20].

F1-score combines recall and precision

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \, . \tag{B.13}$$

The F1-score should be used if the focus is to keep recall and precision high [Tha20].

Specificity is considered as the opposite of recall and is the ratio of TN to all samples belonging to the negative class [Tha20]

$$\text{specificity} = \frac{TN}{TN + FP} \, . \tag{B.14}$$

The Receiver Operating Characteristics (ROC) curve displays the probability of separating the positive and negative classes [Mur22]. It plots the true positive rate (TPR) against the false positive rate (FPR) across decision thresholds $\theta_i \in [0, 1]$. Ideal performance is represented by a point at TPR = 1 and FPR = 0, indicating a perfect classifier [Mur22]. A random classifier has equal TPR and FPR across thresholds. The ROC curve can also be used to evaluate and optimise machine learning (ML) models through the area under the curve (AUC), where values closer to 1 suggest better performance, and an AUC of 0.5 indicates random guessing [Mur22].

Appendix C

Supplementary figures and tables

C.1 Molecular classifiers of specific groups of endotypes

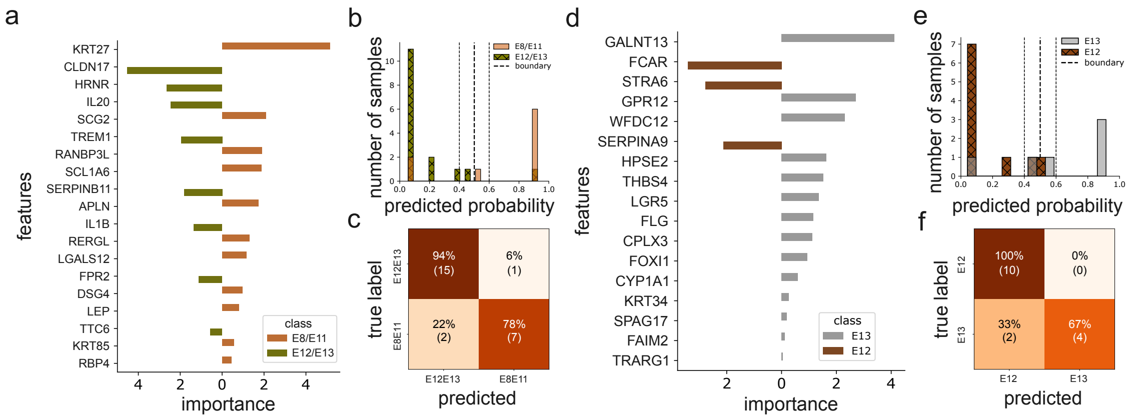


Figure C.1: Second best molecular classifiers for E12/E13 vs. E8/E11 and E12 vs. E13. a, d) Feature importance showing if the feature is either predictive for **(a)** E12/E13 (left, < 0) or E8/E11 (right, > 0) and **(d)** E12 (left, < 0) or E13 (right, > 0). **b, e)** Histogram showing the prediction probability of class 1 of **(b)** classifier E12/E13 (class 0) vs. E8/E11 (class 1) and of **(e)** classifier E12 (class 0) vs. E13 (class 1). The decision boundary is set to 0.5. **c, f)** Confusion matrix summarising the performance of the **(c)** E12/E13 vs. E8/E11 and **(f)** E12 vs. E13 classifier. Higher values in the diagonal are desired.

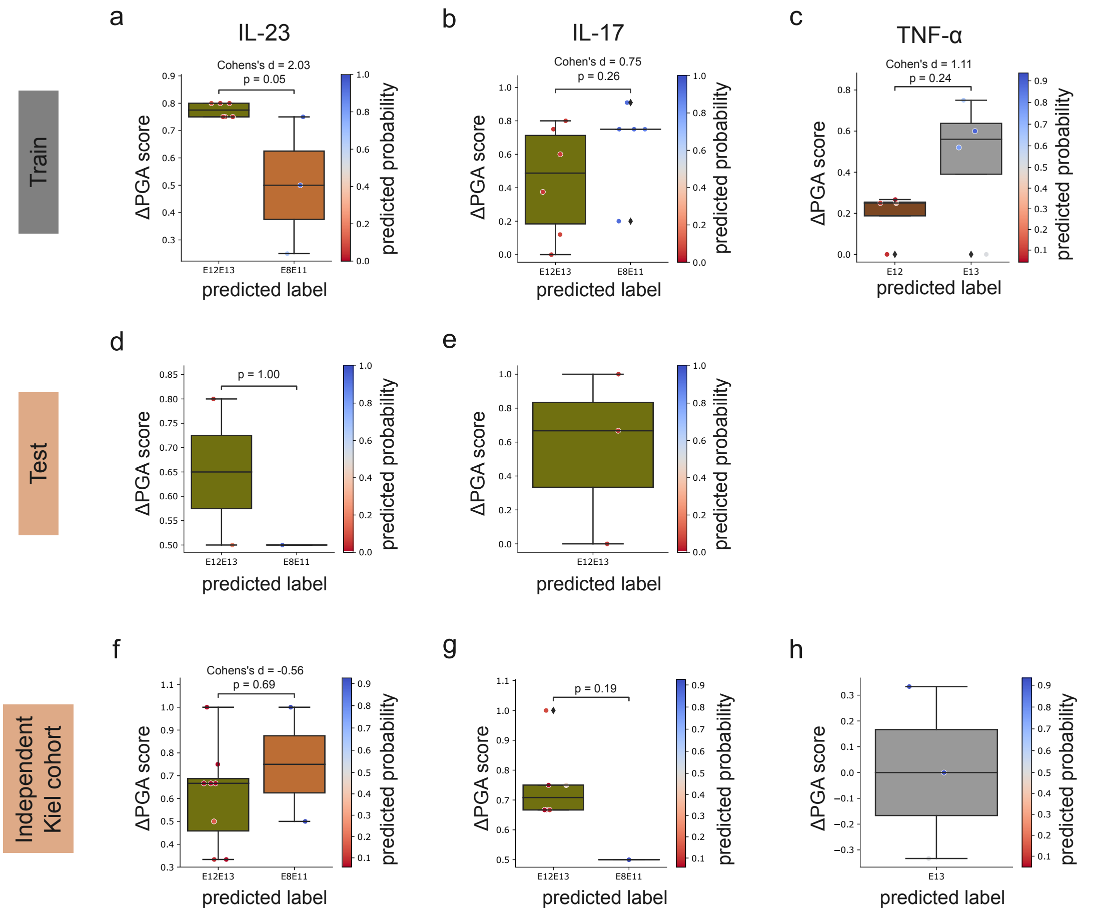


Figure C.2: Evaluation of the association of the second best classifiers for E12/E13 vs. E8/E11 and E12 vs. E13 with the treatment response. a, b, d-g) Boxplots showing the predicted labels of classifier E12/E13 vs. E8/E11 and the corresponding ΔPGA scores of each patient separated by the drug targets **(b, d, f)** IL-23 and **(b, e, g)** IL-17. **c, h)** Boxplots showing the predicted label of classifier E12 vs. E13 and the ΔPGA scores of each patient for drug target TNF-α. **(a-c)** show results of the train set, **(d, e)** of the test set, and **(f-g)** of the independent test cohort. Statistical testing is performed using the Mann-Whitney-U test.

Statistic	Drug Target	Class Label	Dataset		
			Training	Test	Kiel Cohort
Median	IL - 23	E12/E13	0.78	0.65	0.67
		E8/E11	0.50	0.50	0.75
	IL - 17	E12/E13	0.49	0.67	0.71
		E8/E11	0.75	-	0.50
	TNF - α	E12	0.25	-	-
		E13	0.60	-	0.00
Mean	IL - 23	E12/E13	0.78	0.65	0.61
		E8/E11	0.50	0.50	0.75
	IL - 17	E12/E13	0.44	0.56	0.75
		E8/E11	0.67	-	0.50
	TNF - α	E12	0.15	-	-
		E13	0.62	-	0.00
Min	IL - 23	E12/E13	0.75	0.50	0.33
		E8/E11	0.25	0.50	0.50
	IL - 17	E12/E13	0.00	0.00	0.67
		E8/E11	0.20	-	0.50
	TNF - α	E12	0.00	-	-
		E13	0.52	-	-0.33
Max	IL - 23	E12/E13	0.80	0.80	1.00
		E8/E11	0.75	0.50	1.00
	IL - 17	E12/E13	0.80	1.00	1.00
		E8/E11	0.91	-	0.50
	TNF - α	E12	0.27	-	-
		E13	0.75	-	0.33
Q1	IL - 23	E12/E13	0.75	0.58	0.46
		E8/E11	0.38	0.50	0.63
	IL - 17	E12/E13	0.18	0.33	0.67
		E8/E11	0.75	-	0.50
	TNF - α	E12	0.00	-	-
		E13	0.56	-	-0.17
Q3	IL - 23	E12/E13	0.80	0.73	0.69
		E8/E11	0.63	0.50	0.88
	IL - 17	E12/E13	0.71	0.83	0.75
		E8/E11	0.75	-	0.50
	TNF - α	E12	0.25	-	-
		E13	0.68	-	0.17

Table C.1: Boxplot information of training, test, and Kiel cohort datasets for IL - 23, IL - 17, and TNF - α drug targets for the second best classifiers.

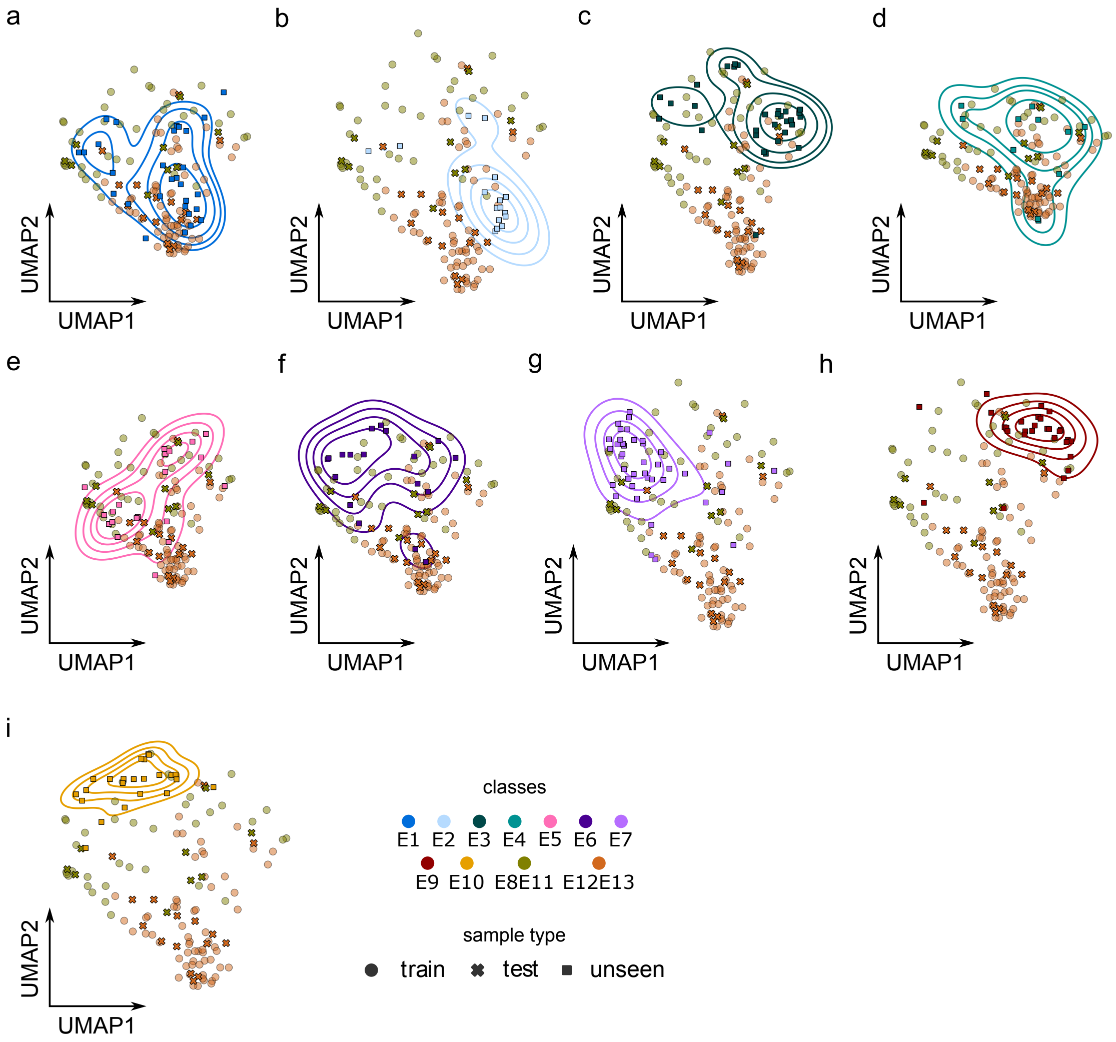


Figure C.3: 2D representation of best E12/E13 vs. E8/E11 classifier features highlighting unseen classes. a-i) Uniform Manifold Approximation and Projection (UMAP) embedding of 19 classifier genes showing the contour lines of each unseen class in comparison to the training and test set samples. The outermost contour line represents the threshold value of 0.3, i.e. 30% of the probability mass lies outside of this contour line.

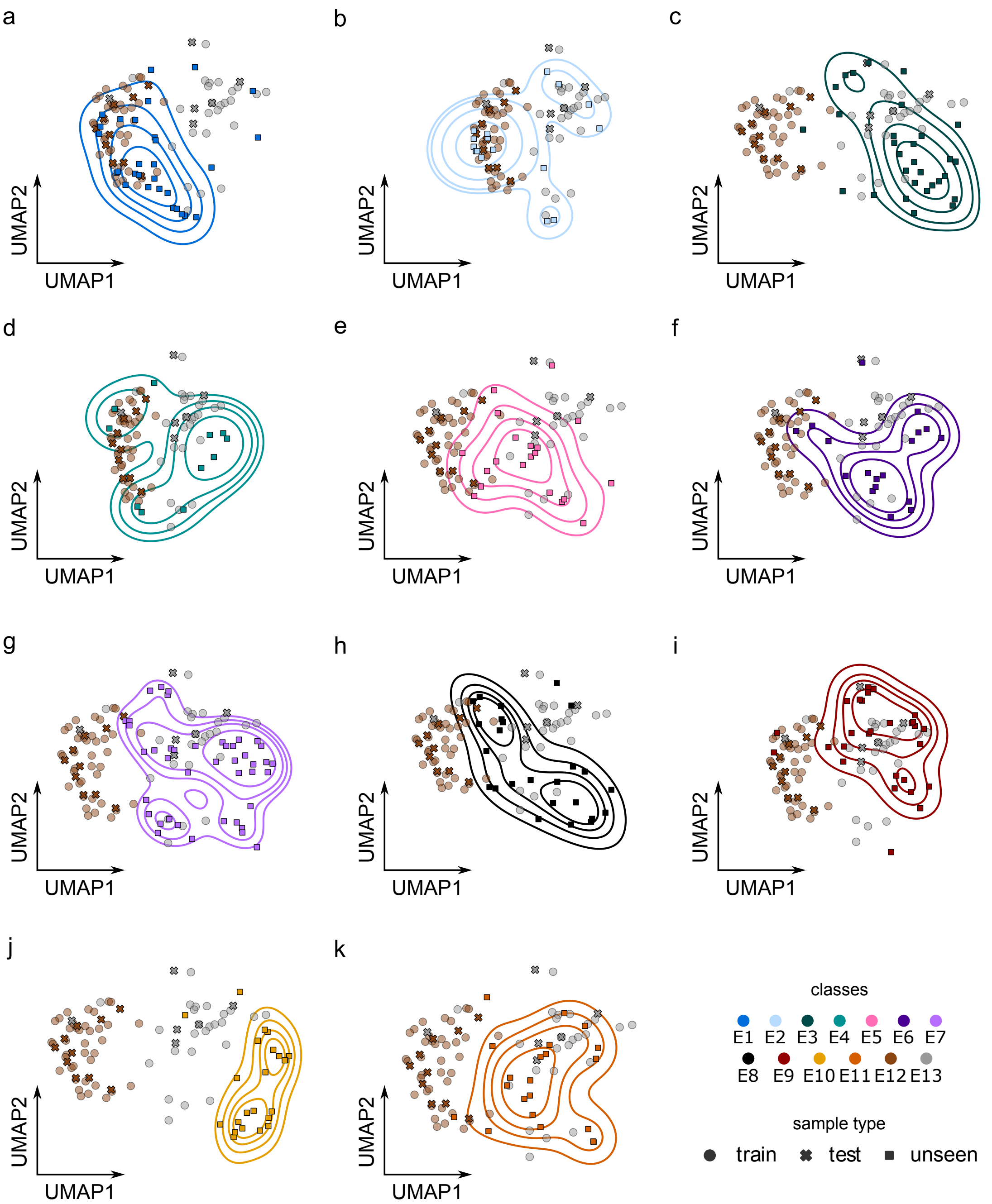


Figure C.4: 2D representation of best E12 vs. E13 classifier features highlighting unseen classes. a-k) UMAP embedding of 15 classifier genes showing the contour lines of each unseen class in comparison to the training and test set samples. The outermost contour line represents the threshold value of 0.3, i.e. 30% of the probability mass lies outside of this contour line.

C.2 Spatial transcriptomics landscape of skin lesions

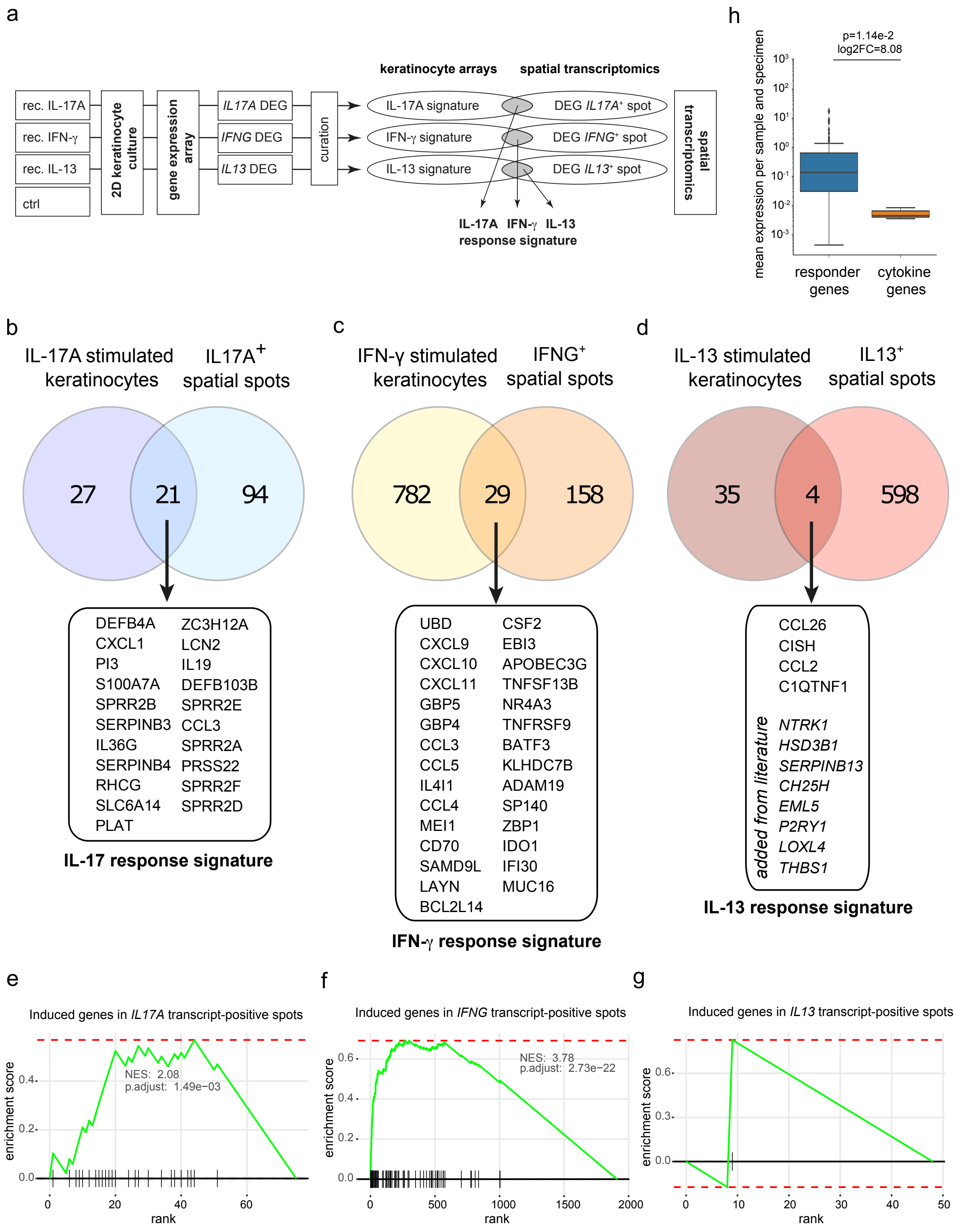
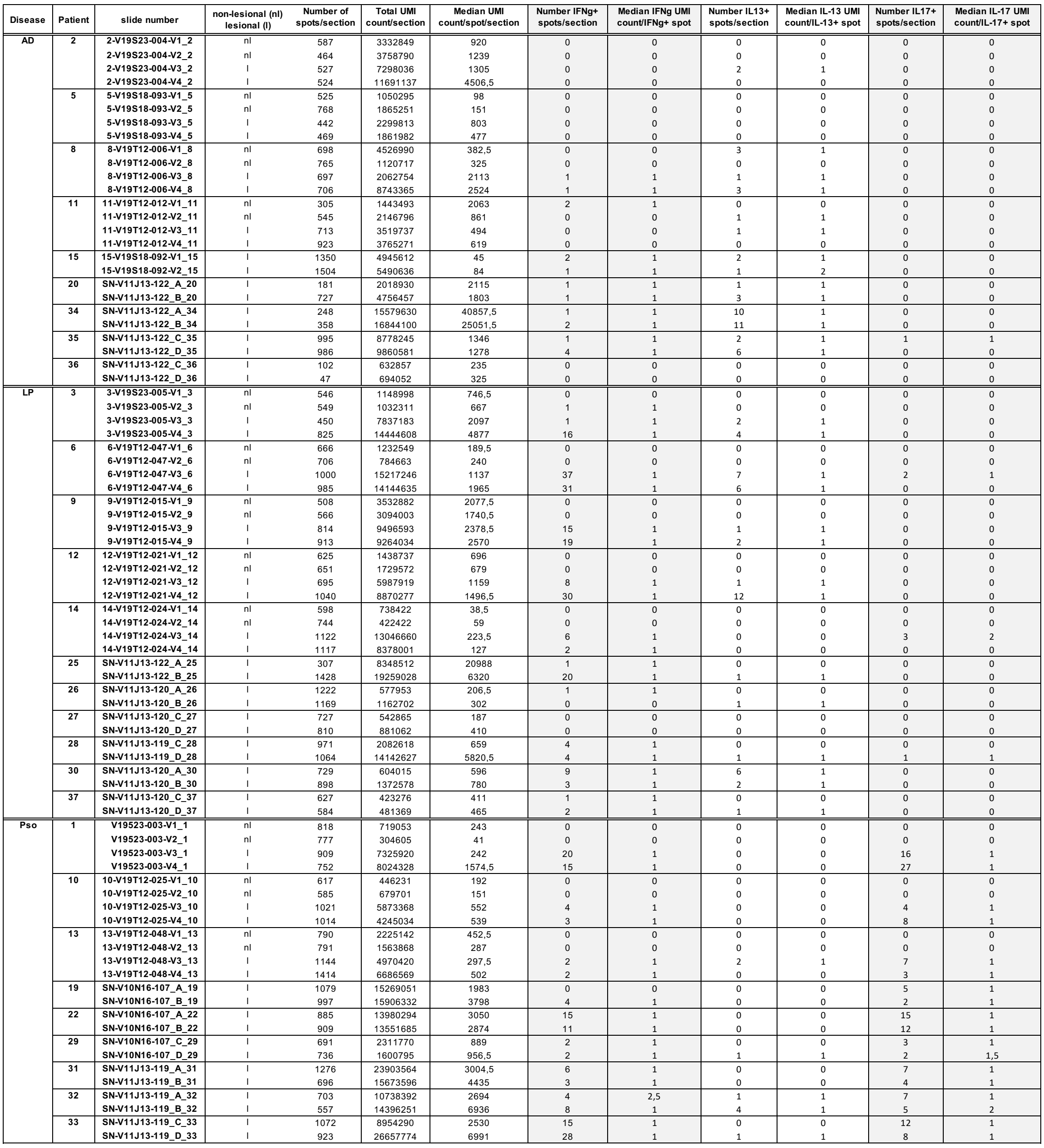


Figure C.5: Experimental and data driven definition of responder signatures of each hallmark cytokine. **a)** In vitro experiment with stimulated (for 16 h) and unstimulated primary human keratinocytes (KCs). differential gene expression (DGE) analysis is performed on generated gene expression data between stimulated and control group for each disease hallmark cytokine. significantly differentially expressed genes (DEGs) are identified by requiring a p - adjusted (padj) value less than 0.05 and a log2 Fold Change (log2FC) above 1.5 for *IL17A* and *IFNG* or a log2FC greater than 1 for *IL13*. Intersection between KC and spatial transcriptomics (ST) derived DEGs is built for each cytokine to define sets of specific responder genes for *IL17A* **(b)**, *IFNG* **(c)**, and *IL13* **(d)**. **e-g)** Enrichment analysis of responder signatures in the identified upregulated DEGs (padj values < 0.05, log2FC > 1) from the cytokine transcript-positive against negative spots DGE analysis. The black stripes mark the position of the responder signatures in the DEG list sorted by the signed padj value. **h)** Average expression of hallmark cytokines and their corresponding responder signature genes across all samples and specimens in lesional (L) skin samples (n=58). The two-sided Mann-Whitney-U test is performed. Boxplot information (responders, cytokines): median (1*e* - 01, 0), mean (1.48, 5.49*e* - 03), min (0, 0), max (1.3, 0), Q1(0, 0), Q3 (46*e* - 01, 0), and whiskers (0, 4.6*e* - 01). Figure adapted from [Sch+22].



Disease	Patient	slide number	non-lesional (nl) lesional (l)	Number of spots/section	Total UMI count/section	Median UMI count/spot/section	Number IFNg+ spots/section	Median IFNg UMI count/IFNg+ spot	Number IL13+ spots/section	Median IL-13 UMI count/IL-13+ spot	Number IL17+ spots/section	Median IL-17 UMI count/IL-17+ spot
AD	2	2-V19S23-004-V1_2	nl	587	3332849	920	0	0	0	0	0	0
		2-V19S23-004-V2_2	nl	464	3758790	1239	0	0	0	0	0	0
		2-V19S23-004-V3_2	l	527	7298036	1305	0	0	2	1	0	0
		2-V19S23-004-V4_2	l	524	11691137	4506,5	0	0	0	0	0	0
	5	5-V19S18-093-V1_5	nl	525	1050295	98	0	0	0	0	0	0
		5-V19S18-093-V2_5	nl	768	1865251	151	0	0	0	0	0	0
		5-V19S18-093-V3_5	l	442	2299813	803	0	0	0	0	0	0
		5-V19S18-093-V4_5	l	469	1861982	477	0	0	0	0	0	0
	8	8-V19T12-006-V1_8	nl	698	4526990	382,5	0	0	3	1	0	0
		8-V19T12-006-V2_8	nl	765	1120717	325	0	0	0	0	0	0
		8-V19T12-006-V3_8	l	697	2062754	2113	1	1	1	1	0	0
		8-V19T12-006-V4_8	l	706	8743365	2524	1	1	3	1	0	0
	11	11-V19T12-012-V1_11	nl	305	1443493	2063	2	1	0	0	0	0
		11-V19T12-012-V2_11	nl	545	2146796	861	0	0	1	1	0	0
		11-V19T12-012-V3_11	l	713	3519737	494	0	0	1	1	0	0
		11-V19T12-012-V4_11	l	923	3765271	619	0	0	0	0	0	0
	15	15-V19S18-092-V1_15	l	1350	4945612	45	2	1	2	1	0	0
		15-V19S18-092-V2_15	l	1504	5490636	84	1	1	1	2	0	0
	20	SN-V11J13-122_A_20	l	181	2018930	2115	1	1	1	1	0	0
		SN-V11J13-122_B_20	l	727	4756457	1803	1	1	3	1	0	0
	34	SN-V11J13-122_A_34	l	248	15579630	40857,5	1	1	10	1	0	0
		SN-V11J13-122_B_34	l	358	16844100	25051,5	2	1	11	1	0	0
	35	SN-V11J13-122_C_35	l	995	8778245	1346	1	1	2	1	1	1
		SN-V11J13-122_D_35	l	986	9860581	1278	4	1	6	1	0	0
	36	SN-V11J13-122_C_36	l	102	632857	235	0	0	0	0	0	0
		SN-V11J13-122_D_36	l	47	694052	325	0	0	0	0	0	0
LP	3	3-V19S23-005-V1_3	nl	546	1148998	746,5	0	0	0	0	0	0
		3-V19S23-005-V2_3	nl	549	1032311	667	1	1	0	0	0	0
		3-V19S23-005-V3_3	l	450	7837183	2097	1	1	2	1	0	0
		3-V19S23-005-V4_3	l	825	14444608	4877	16	1	4	1	0	0
	6	6-V19T12-047-V1_6	nl	666	1232549	189,5	0	0	0	0	0	0
		6-V19T12-047-V2_6	nl	706	784663	240	0	0	0	0	0	0
		6-V19T12-047-V3_6	l	1000	15217246	1137	37	1	7	1	2	1
		6-V19T12-047-V4_6	l	985	14144635	1965	31	1	6	1	0	0
	9	9-V19T12-015-V1_9	nl	508	3532882	2077,5	0	0	0	0	0	0
		9-V19T12-015-V2_9	nl	566	3094003	1740,5	0	0	0	0	0	0
		9-V19T12-015-V3_9	l	814	9496593	2378,5	15	1	1	1	0	0
		9-V19T12-015-V4_9	l	913	9264034	2570	19	1	2	1	0	0
	12	12-V19T12-021-V1_12	nl	625	1438737	696	0	0	0	0	0	0
		12-V19T12-021-V2_12	nl	651	1729572	679	0	0	0	0	0	0
		12-V19T12-021-V3_12	l	695	5987919	1159	8	1	1	1	0	0
		12-V19T12-021-V4_12	l	1040	8870277	1496,5	30	1	12	1	0	0
	14	14-V19T12-024-V1_14	nl	598	738422	38,5	0	0	0	0	0	0
		14-V19T12-024-V2_14	nl	744	422422	59	0	0	0	0	0	0
		14-V19T12-024-V3_14	l	1122	13046660	223,5	6	1	0	0	3	2
		14-V19T12-024-V4_14	l	1117	8378001	127	2	1	0	0	0	0
	25	SN-V11J13-122_A_25	l	307	8348512	20988	1	1	0	0	0	0
		SN-V11J13-122_B_25	l	1428	19259028	6320	20	1	1	1	0	0
	26	SN-V11J13-120_A_26	l	1222	577953	206,5	1	1	0	0	0	0
		SN-V11J13-120_B_26	l	1169	1162702	302	0	0	1	1	0	0
	27	SN-V11J13-120_C_27	l	727	542865	187	0	0	0	0	0	0
		SN-V11J13-120_D_27	l	810	881062	410	0	0	0	0	0	0
	28	SN-V11J13-119_C_28	l	971	2082618	659	4	1	0	0	0	0
		SN-V11J13-119_D_28	l	1064	14142627	5820,5	4	1	1	1	1	1
	30	SN-V11J13-120_A_30	l	729	604015	596	9	1	6	1	0	0
		SN-V11J13-120_B_30	l	898	1372578	780	3	1	2	1	0	0
	37	SN-V11J13-120_C_37	l	627	423276	411	1	1	0	0	0	0
		SN-V11J13-120_D_37	l	584	481369	465	2	1	1	1	0	0
Pso	1	V19523-003-V1_1	nl	818	719053	243	0	0	0	0	0	0
		V19523-003-V2_1	nl	777	304605	41	0	0	0	0	0	0
		V19523-003-V3_1	l	909	7325920	242	20	1	0	0	16	1
		V19523-003-V4_1	l	752	8024328	1574,5	15	1	0	0	27	1
	10	10-V19T12-025-V1_10	nl	617	446231	192	0	0	0	0	0	0
		10-V19T12-025-V2_10	nl	585	679701	151	0	0	0	0	0	0
		10-V19T12-025-V3_10	l	1021	5873368	552	4	1	0	0	4	1
		10-V19T12-025-V4_10	l	1014	4245034	539	3	1	0	0	8	1
	13	13-V19T12-048-V1_13	nl	790	2225142	452,5	0	0	0	0	0	0
		13-V19T12-048-V2_13	nl	791	1563868	287	0	0	0	0	0	0
		13-V19T12-048-V3_13	l	1144	4970420	297,5	2	1	2	1	7	1
		13-V19T12-048-V4_13	l	1414	6686569	502	2	1	0	0	3	1
	19	SN-V10N16-107_A_19	l	1079	15269051	1983	0	0	0	0	5	1
		SN-V10N16-107_B_19	l	997	15906332	3798	4	1	0	0	2	1
	22	SN-V10N16-107_A_22	l	885	13980294	3050	15	1	0	0	15	1
		SN-V10N16-107_B_22	l	909	13551685	2874	11	1	0	0	12	1
	29	SN-V10N16-107_C_29	l	691	2311770	889	2	1	0	0	3	1
		SN-V10N16-107_D_29	l	736	1600795	956,5	2	1	1	1	2	1,5
	31	SN-V11J13-119_A_31	l	1276	23903564	3004,5	6	1	0	0	7	1
		SN-V11J13-119_B_31	l	696	15673596	4435	3	1	0	0	4	1
	32	SN-V11J13-119_A_32	l	703	10738392	2694	4	2,5	1	1	7	1
		SN-V11J13-119_B_32	l	557	14396251	6936	8	1	4	1	5	2
	33	SN-V11J13-119_C_33	l	1072	8954290	2530	15	1	0	0	12	1
		SN-V11J13-119_D_33	l	923	26657774	6991	28	1	1	1	8	1

Figure C.6: Summary of general and cytokine expression in ST dataset. Table adapted from Schäbitz & Hillig et al. (2022) [Sch+22].

Acronyms

AI	artificial intelligence
AMI	Adjusted Mutual Information
AMP	antimicrobial peptide
ANOVA	analysis of variance
APC	antigen presenting cell
AUC	area under the curve
BH	Benjamini and Hochberg
cDNA	complementary DNA
cdr	cellular detection rate
CI	confidence interval
CPM	count - per - million
CV	cross-validation
DBI	Davies-Bouldin Index
DC	Dendritic cell
DEG	significantly differentially expressed gene
DGE	differential gene expression
DLQI	Dermatology Life Quality Index
ECM	extracellular matrix
ES	enrichment score
FACS	fluorescence activated cell sorting
FDR	false discovery rate
FFPE	formalin-fixed paraffin-embedded
FFT	fresh-frozen tissue
FLG	filaggrin
FN	false negative
FP	false positive
FPR	false positive rate
GLM	generalised linear model
GRC	Genome Reference Consortium
GSEA	gene set enrichment analysis
H&E	hematoxylin and eosin
HKG	housekeeping gene
HRG	highly relevant gene
HVG	highly variable gene
IHC	immuno histochemistry
ILC	innate lymphoid cell
IRP	immune response pattern
ISH	in situ hybridisation
KC	keratinocyte
KNN	K-nearest neighbour

L	lesional
LC	Langerhans cell
LCE	Late cornified envelope
LE	lymphatic endothelium
log2FC	log2 Fold Change
logCPM	log2 counts-per-million
LRT	likelihood ratio test
Mac	macrophage
MAR	Missing at random
MCAR	Missing completely at random
MI	mutual information
MICI	maximal information compression index
ML	machine learning
MLP	Multi-layer Perceptron
MNAR	Missing not at random
mRNA	messenger RNA
MT	mitochondrial
NB	negative binomial
ncISD	non-communicable chronic inflammatory skin disease
NES	normalised enrichment score
NGS	next generation sequencing
NHST	null hypothesis statistical tests
NK	natural killer cell
NL	non-lesional
NN	neural network
ORA	over representation analysis
padj	p - adjusted
PASI	Psoriasis Area and Severity Index
PCA	principal component analysis
PCC	Pearson's correlation coefficient
PCR	polymerase chain reaction
PCs	principal components
PGA	Physician Global Assessment
PR	precision-recall
PtGA	Patient Global Assessment
QC	Quality control
RFSCV	recursive feature selection with cross-validation
RNA-seq	RNA-sequencing
ROC	Receiver Operating Characteristics
RRA	Robust Rank Aggregation
RSS	residual sum of squares

scANVI	single-cell ANnotation using Variational Inference
SCC	Spearman’s correlation coefficient
SCORAD	SCORing of Atopic Dermatitis
scRNA-seq	single-cell RNA-sequencing
scVI	single-cell variational inference
SFS	sequential feature selection
SMOTE	Synthetic Minority Oversampling Technique
ST	spatial transcriptomics
SVD	singular value decomposition
SVM	Support Vector Machine
$\mathbf{T_{RM}}$ - cell	resident memory T - cell
Tc cell	cytotoxic T - cell
Th	T-helper
TMM	trimmed mean of M-values
TN	true negative
TP	true positive
TPM	transcript per million
TPR	true positive rate
Treg	T regulatory cell
UD	undefined
UMAP	Uniform Manifold Approximation and Projection
UMI	unique molecular identifier
VAE	variational Autoencoder
VE	vascular endothelium
VIF	variance inflation factor

List of Figures

1.1 Structure and cell type composition of healthy human skin. 4
1.2 Lineage tree of immune cells in the skin. 7
1.3 Existing IRPs and their mechanisms. 17
1.4 Cell organelles and the gene expression process. 20
1.5 Thesis aims. 28

3.1 Rich data resource including gene expression data and clinical traits. 95
3.2 Clinical traits are not sufficient for diagnostic purposes. 97
3.3 Gene expression embedding reveals overlap between diagnoses and IRPs. 98
3.4 Example showing the distribution of genes before and after sorting. 101
3.5 Workflow of determining HVGs. 102
3.6 Comparison of feature selection methods to identify HVGs. 105
3.7 Determine cluster resolution using AMI and DBI. 107
3.8 Identification of disease endotypes. 110
3.9 Hierarchical relationships of endotypes. 112
3.10 Biological interpretation of interesting endotype branches. 113
3.11 Biologically distinct functionalities within hierarchical endotype relationships. . 115
3.12 Functional differences between one endotype and all others. 117
3.13 Clinical interpretation of endotypes. 120

4.1 Responder patterns in psoriasis patients. 144
4.2 Molecular characterisation of responders and non-responder psoriasis patients. . 146
4.3 Triad feature selection pipeline including GeneSTRIVE. 152
4.4 Models and CV performance of all classifiers. 155
4.5 Feature sets of all E12/E13 vs. E8/E11 classifiers. 157
4.6 Best E12/E13 vs. E8/E11 gene-expression-based classifier. 158
4.7 Feature sets of all E12 vs. E13 classifiers. 161
4.8 Best E12 vs. E13 gene-expression-based classifier. 162
4.9 Evaluation of best E12/E13 vs. E8/E11 classifier on unseen class samples. . . . 165
4.10 Evaluation of best E12 vs. E13 classifier on unseen class samples. 166
4.11 Classifier evaluation on independent cohort. 167
4.12 Evaluation of the association of the best classifiers with treatment response. . . 168

5.1 Analysis workflow and validation data 189
5.2 Hallmark cytokines found to be lowly expressed in all experiments 191
5.3 Low expression of cytokines in ncISDs 192

5.4 Validation experiments on protein and RNA level of cytokine expressing T-cells in skin lesions 195
5.5 Micro-environment of cytokine transcript-positive spots is characterised by type 3 immune response 197
5.6 Gene expression embedding of leukocyte containing spots. 198
5.7 Characterisation of *IFNG* and *IL13* transcript-positive leukocyte spots. 199
5.8 Deconvolution of spots. 200
5.9 Characterisation of cytokine transcript-positive cells. 202
5.10 Correlation between cytokines and their corresponding response signature genes transcripts 204
5.11 Spatial correlation between hallmark cytokines and their corresponding response signature genes transcripts. 208
5.12 The spatial clustering algorithm identifies additional cytokine associated genes and potential drug targets. 210

C.1 Second best molecular classifiers. 237
C.2 Evaluation of the association of the second best classifiers with treatment response.238
C.3 2D representation of best E12/E13 vs. E8/E11 classifier features highlighting unseen classes. 240
C.4 2D representation of best E12 vs. E13 classifier features highlighting unseen classes. 241
C.5 Experimental and data driven definition of responder signatures of each hallmark cytokine. 243
C.6 Summary of ST dataset 244

List of Tables

2.1 Association between Type I and Type II error. 34
2.2 An example of a 2×2 contingency table. 37
2.3 Assessment of the absolute correlation coefficient score 41
2.4 Confusion matrix for binary classification. 69

3.1 Feature selection parameter to identify HVGs using HRG, SCANPY, or SD. . 104

4.1 Therapy response cohort of 34 psoriasis patients. 136
4.2 Drug targets associated with endotypes. 136
4.3 Overview of the unpublished RNA-seq cohort from Kiel. 137
4.4 Hyperparameters of the classifiers employed in the feature selection. 140
4.5 Hyperparameters of the MLP, SVM, and XGBoost model. 142
4.6 Hyperparameters of the best models for E12/E13 vs. E8/11. 159
4.7 Test set performance of the E12/E13 vs. E8/E11 baseline model and best, calibrated classifiers. 159
4.8 Hyperparameters of the best models for E12 vs. E13. 163
4.9 Test set performance of baseline model and best-performing, calibrated classifiers of E12 vs. E13. 163
4.10 Boxplot information of training, test, and Kiel cohort datasets for IL - 23, IL - 17, and TNF - α drug targets for the best classifiers. 170

5.1 Overview of data set generated by Alexander Schäbitz, Menatullah Mubarak, Manja Jargosch, Natalie Garzorz-Stark, and Emanuele Scala. 180
5.2 Overview of patient characteristics of the ST dataset. 181
5.3 Overview of the scRNA-seq dataset from Reynolds, Gary, et al. (2021). 186
5.4 Overview of the unpublished bulk RNA-seq dataset, which comprises a total of 330 samples. 187

C.1 Boxplot information of training, test, and Kiel cohort datasets for IL - 23, IL - 17, and TNF - α drug targets for the second best classifiers. 239

Bibliography

[199] Parzival' 1997. *Biotechnology icons*. https://www.flaticon.com/free-icons/biotechnology. Last visit: 03.02.2025.

[AA+19] Ioana Agache, Cezmi A Akdis, et al. "Precision medicine and phenotypes, endotypes, genotypes, regiotypes, and theratypes of allergic diseases". In: *Journal of Clinical Investigation* 129.4 (Mar. 2019), pp. 1493–1503. ISSN: 1558-8238. DOI: 10.1172/jci124611.

[Abd+20] Rana Abdat et al. "Dupilumab as a novel therapy for bullous pemphigoid: a multicenter case series". In: *Journal of the American Academy of Dermatology* 83.1 (2020), pp. 46–52. DOI: https://doi.org/10.1016/j.jaad.2020.01.089.

[AEH20] Constantin Ahlmann-Eltze and Wolfgang Huber. "glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data". In: *Bioinformatics* 36.24 (2020), pp. 5701–5702. DOI: 10.1093/bioinformatics/btaa1009.

[AEH23] Constantin Ahlmann-Eltze and Wolfgang Huber. "Comparison of transformations for single-cell RNA-seq data". In: *Nature Methods* (2023), pp. 1–8. DOI: https://doi.org/10.1038/s41592-023-01814-1.

[AG88] Douglas G Altman and Martin J Gardner. "Statistics in Medicine: Calculating confidence intervals for regression and correlation". In: *British medical journal (Clinical research ed.)* 296.6631 (1988), p. 1238. DOI: 10.1136/bmj.296.6631.1238.

[AHK01] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2001, pp. 420–434. ISBN: 9783540445036. DOI: 10.1007/3-540-44503-x_27.

[Ain+12] Chrysanthi Ainali et al. "Transcriptome classification reveals molecular subtypes in psoriasis". In: *BMC Genomics* 13.1 (Sept. 2012), pp. 1–15. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-472.

[Air+15] Daniel J Aires et al. "Reproducible Novel Transcriptional Differences Between Psoriatic Lesional and Non-Lesional Skin Show Increased Inflammation and Metabolism." In: *Journal of Drugs in Dermatology: JDD* 14.8 (2015), pp. 794–800.

[Akh+22] Leyla A Akh et al. "-Omics potential of in vitro skin models for radiation exposure". In: *Cellular and Molecular Life Sciences* 79.7 (2022), p. 390. DOI: 10.1007/s00018-022-04394-z.

[Ako18] Haldun Akoglu. "User's guide to correlation coefficients". In: *Turkish journal of emergency medicine* 18.3 (2018), pp. 91–93. DOI: 10.1016/j.tjem.2018.08.001.

[Alb+02a] Bruce Alberts et al. "From DNA to RNA". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[Alb+02b] Bruce Alberts et al. "From RNA to protein". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[All+15] Yannick Allanore et al. "Systemic sclerosis". In: *Nature reviews Disease primers* 1.1 (2015), pp. 1–21.

[Ama+20] Teresa M.S. Amaral et al. "Clinical validation of a prognostic 11-gene expression profiling score in prospectively collected FFPE tissue of patients with AJCC v8 stage II cutaneous melanoma". In: *European Journal of Cancer* 125 (Jan. 2020), pp. 38–45. ISSN: 0959-8049. DOI: 10.1016/j.ejca.2019.10.027.

[Ama+23] Paulo Amaral et al. "The status of the human gene catalogue". In: *Nature* 622.7981 (Oct. 2023), pp. 41–47. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06490-x.

[AMP17] Florence Abdallah, Lily Mijouin, and Chantal Pichon. "Skin Immune Landscape: Inside and Outside the Organism". In: *Mediators of Inflammation* 2017 (2017), pp. 1–17. ISSN: 1466-1861. DOI: 10.1155/2017/5095293.

[And] .

[And+20] Alma Andersson et al. "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography". In: *Communications biology* 3.1 (2020), p. 565. DOI: 10.1038/s42003-020-01247-y.

[AR14] Charu C Aggarwal and Chandan K Reddy. "Data clustering". In: *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra* (2014).

[AR20] April W Armstrong and Charlotte Read. "Pathophysiology, Clinical Presentation, and Treatment of Psoriasis: A Review". In: *JAMA* 323.19 (May 2020), pp. 1945–1960. ISSN: 0098-7484. DOI: 10.1001/jama.2020.4006.

[Arm+21] Erick Armingol et al. "Deciphering cell–cell interactions and communication from gene expression". In: *Nature Reviews Genetics* 22.2 (2021), pp. 71–88.

[Aro+23] Rohit Arora et al. "Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response". In: *Nature Communications* 14.1 (Aug. 2023), p. 5029. ISSN: 2041-1723. DOI: 10.1038/s41467-023-40271-4.

[ARR15] Francesco Annunziato, Chiara Romagnani, and Sergio Romagnani. "The 3 major types of innate and adaptive cell-mediated effector immunity". In: *Journal of Allergy and Clinical Immunology* 135.3 (Mar. 2015), pp. 626–635. ISSN: 0091-6749. DOI: 10.1016/j.jaci.2014.11.001.

[Aur+19] Yuri Sousa Aurelio et al. "Learning from imbalanced data sets with weighted cross-entropy function". In: *Neural processing letters* 50 (2019), pp. 1937–1949. DOI: 10.1007/s11063-018-09977-1.

[AV+07] David Arthur, Sergei Vassilvitskii, et al. "k-means++: The advantages of careful seeding". In: *Soda*. Vol. 7. 2007, pp. 1027–1035.

[Awi] Awicon. *Patient record icons*. https://www.flaticon.com/free-icons/patient-record. Last visit: 03.02.2025.

[Ayl25] Fisher Ronal Aylmer. "Statistical Methods for Research Workers". In: (1925). DOI: 10.1093/oso/9780198522294.002.0003.

[Bah+12] Bahman Bahmani et al. "Scalable k-means++". In: *arXiv preprint arXiv:1203.6402* (2012). DOI: 10.14778/2180912.2180915.

[Bai+18] Paul Bailey et al. "Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas. AIR–NAEP Working Paper No. 2018-01. NCES Data R Project Series# 02." In: *American Institutes for Research* (2018).

[Bai+21] Majd Bairkdar et al. "Incidence and prevalence of systemic sclerosis globally: a comprehensive systematic review and meta-analysis". In: *Rheumatology* 60.7 (2021), pp. 3121–3133.

[Bak+20] Daphne S. Bakker et al. "EASI p-EASI: Predicting disease severity in atopic dermatitis patients treated with dupilumab using a combination of serum biomarkers". In: *Allergy* 75.12 (July 2020), pp. 3287–3289. ISSN: 1398-9995. DOI: 10.1111/all.14492.

[Bal+21] Paul Balanescu et al. "S100A6, calumenin and cytohesin 2 as biomarkers for cutaneous involvement in systemic sclerosis patients: A case control study". In: *Journal of Personalized Medicine* 11.5 (2021), p. 368. DOI: 10.3390/jpm11050368.

[Ban+17] Peter Bankhead et al. “QuPath: Open source software for digital pathology image analysis”. In: *Scientific Reports* 7.1 (Dec. 2017), pp. 1–7. ISSN: 2045-2322. DOI: 10.1038/s41598-017-17204-5.

[Bat+20] Richa Batra et al. “Integration of phenomics and transcriptomics data to reveal drivers of inflammatory processes in the skin”. In: *bioRxiv* (2020), pp. 2020–07. DOI: 10.1101/2020.07.25.221309.

[Bau+06] Robert P Baughman et al. “Infliximab therapy in patients with chronic sarcoidosis and pulmonary involvement”. In: *American journal of respiratory and critical care medicine* 174.7 (2006), pp. 795–802.

[BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. “Correlation clustering”. In: *Machine learning* 56 (2004), pp. 89–113. DOI: 10.1023/b:mach.0000033116.57574.95.

[BC+14] Verónica Bolón-Canedo et al. “A review of microarray datasets and applied feature selection methods”. In: *Information Sciences* 282 (Oct. 2014), pp. 111–135. ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.05.042.

[BC20] Dickinson Becton and Company. *FlowJoTM Software.* Version 10.7.1. 2020.

[BCB03] Viv Bewick, Liz Cheek, and Jonathan Ball. “Statistics review 7: Correlation and regression”. In: *Critical care* 7 (2003), pp. 1–9.

[BCS22] Eugene N Brooks, Michelle D Colbert, and Inbal B Sander. “Interface dermatitis: Delineating the diagnosis with adaptive immune markers”. In: *Journal of cutaneous pathology* 49.7 (2022), p. 669. DOI: 10.1111/cup.14236.

[Ben+17] Alessio Benavoli et al. “Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 2653–2688.

[Ber+11] Judith G.M. Bergboer et al. “Psoriasis Risk Genes of the Late Cornified Envelope-3 Group Are Distinctly Expressed Compared with Genes of Other LCE Groups”. In: *The American Journal of Pathology* 178.4 (Apr. 2011), pp. 1470–1477. ISSN: 0002-9440. DOI: 10.1016/j.ajpath.2010.12.017.

[Ber+19] Koen Van den Berge et al. “RNA sequencing data: hitchhiker’s guide to expression analysis”. In: *Annual Review of Biomedical Data Science* 2 (2019), pp. 139–173. DOI: 10.1146/annurev-biodatasci-072018-021255.

[Ber+20] Volker Bergen et al. “Generalizing RNA velocity to transient cell states through dynamical modeling”. In: *Nature biotechnology* 38.12 (2020), pp. 1408–1414. DOI: 10.1038/s41587-020-0591-3.

[Bey+99] Kevin Beyer et al. "When Is "Nearest Neighbor" Meaningful?" In: *Database Theory — ICDT'99*. Springer Berlin Heidelberg, 1999, pp. 217–235. ISBN: 9783540492573. DOI: 10.1007/3-540-49257-7_15.

[BG+17] Jose Barrera-Gómez et al. "A systematic comparison of statistical methods to detect interactions in exposome-health associations". In: *Environmental Health* 16.1 (July 2017), pp. 1–13. ISSN: 1476-069X. DOI: 10.1186/s12940-017-0277-6.

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[Bia+21] Tommaso Biancalani et al. "Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram". In: *Nature methods* 18.11 (2021), pp. 1352–1362. DOI: 10.1038/s41592-021-01264-7.

[Bla+18] Alessandro Blasimme et al. "Data sharing for precision medicine: policy lessons and future directions". In: *Health Affairs* 37.5 (2018), pp. 702–709. DOI: https://doi.org/10.1377/hlthaff.2017.1558.

[Blo+08] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

[Blu+18] Stefan Blunder et al. "Enhanced expression of genes related to xenobiotic metabolism in the skin of patients with atopic dermatitis but not with ichthyosis vulgaris". In: *Journal of Investigative Dermatology* 138.1 (2018), pp. 98–108. DOI: 10.1016/j.jid.2017.08.036.

[Boc+21] Katharina Boch et al. "Lichen planus". In: *Frontiers in medicine* 8 (2021), p. 737813.

[Boe+12] Dana Boehm et al. "Anxiety, depression and impaired health-related quality of life in patients with occupational hand eczema". In: *Contact dermatitis* 67.4 (2012), pp. 184–192.

[Bon+21] Fabio Boniolo et al. "Artificial intelligence in early drug discovery enabling precision medicine". In: *Expert Opinion on Drug Discovery* 16.9 (June 2021), pp. 991–1007. ISSN: 1746-045X. DOI: 10.1080/17460441.2021.1918096.

[Boy+04] Elizabeth I Boyle et al. "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (Aug. 2004), pp. 3710–3715. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth456.

[Bra+24] Alicia M. Braxton et al. "3D genomic mapping reveals multifocality of human pancreatic precancers". In: *Nature* 629.8012 (May 2024), pp. 679–687. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07359-3.

[Bra97] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159. DOI: 10.1016/s0031-3203(96)00142-2.

[Bre+13] Philip Brennecke et al. "Accounting for technical noise in single-cell RNA-seq experiments". In: *Nature methods* 10.11 (2013), pp. 1093–1095. DOI: https://doi.org/10.1038/nmeth.2645.

[Bro19] Jason Brownlee. *Probability for machine learning: Discover how to harness uncertainty with Python*. Machine Learning Mastery, 2019.

[BS14] Krishnan Bhaskaran and Liam Smeeth. "What is the difference between missing completely at random and missing at random?" In: *International journal of epidemiology* 43.4 (2014), pp. 1336–1339. DOI: 10.1093/ije/dyu080.

[BSW77] Jon L. Bentley, Donald F. Stanat, and E.Hollins Williams. "The complexity of finding fixed-radius near neighbors". In: *Information Processing Letters* 6.6 (Dec. 1977), pp. 209–212. ISSN: 0020-0190. DOI: 10.1016/0020-0190(77)90070-9.

[Büt+19] Maren Büttner et al. "A test metric for assessing single-cell RNA-seq batch correction". In: *Nature methods* 16.1 (2019), pp. 43–49.

[BW00] Douglas G Bonett and Thomas A Wright. "Sample size requirements for estimating Pearson, Kendall and Spearman correlations". In: *Psychometrika* 65 (2000), pp. 23–28. DOI: 10.1007/bf02294183.

[BWJ75] Robert A Briggaman and Clayton E Wheeler Jr. "The Epidermal-Dermal Junction". In: *Journal of Investigative Dermatology* 65.1 (July 1975), pp. 71–84. ISSN: 0022-202X. DOI: 10.1111/1523-1747.ep12598050.

[CA16] Maja-Lisa Clausen and Tove Agner. "Antimicrobial peptides, infections and the skin barrier". In: *Skin Barrier Function* 49 (2016), pp. 38–46.

[CA22] Davide Chicco and Giuseppe Agapito. "Nine quick tips for pathway enrichment analysis". In: *PLoS computational biology* 18.8 (2022), e1010348.

[Cam+20] Neil Campbell et al. *Biology: A Global Approach*. Vol. 12. Pearson, 2020. ISBN: 978-1-292-34163-7.

[Can+23] Zixuan Cang et al. "Screening cell–cell communication in spatial transcriptomics via collective optimal transport". In: *Nature methods* 20.2 (2023), pp. 218–228. DOI: https://doi.org/10.1038/s41592-022-01728-4.

[Car+13] Fernanda Carregaro et al. "Study of small proline-rich proteins (SPRRs) in health and disease: a review of the literature". In: *Archives of Dermatological Research* 305.10 (Oct. 2013), pp. 857–866. ISSN: 1432-069X. DOI: 10.1007/s00403-013-1415-9.

[Car+19] Marc Carlson et al. "org. Hs. eg. db: Genome wide annotation for Human". In: *R package version* 3.2 (2019), p. 3.

[CB15] Giorgio Corani and Alessio Benavoli. "A Bayesian approach for comparing cross-validated algorithms on multiple data sets". In: *Machine Learning* 100.2-3 (2015), pp. 285–304. DOI: https://doi.org/10.1007/s10994-015-5486-z.

[CBF12] E Contassot, HD Beer, and LE French. "Interleukin-1, inflammasomes, autoinflammation and the skin". In: *Swiss Medical Weekly* 142.2122 (May 2012), w13590–w13590. ISSN: 1424-3997. DOI: 10.4414/smw.2012.13590.

[CG16] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[Cha+02] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[Che+16] Hung-I Harry Chen et al. "Detection of high variability in gene expression from single-cell RNA-seq profiling". In: *BMC genomics* 17 (2016), pp. 119–128. DOI: https://doi.org/10.1186/s12864-016-2897-6.

[Chi+11] Andrea Chiricozzi et al. "Integrative responses to IL-17 and TNF-α in human keratinocytes account for key inflammatory pathogenic circuits in psoriasis". In: *Journal of Investigative Dermatology* 131.3 (2011), pp. 677–687. DOI: 10.1038/jid.2010.340.

[Cho+17] R Chopra et al. "Severity strata for EASI, mEASI, oSCORAD, SCORAD, ADSI and BSA in adolescents and adults with atopic dermatitis". In: *Br J Dermatol* 177 (2017), pp. 1316–1321.

[Chu21] Chia-Yu Chu. “New targets in treating granuloma annulare”. In: *Journal of Allergy and Clinical Immunology* 147.5 (2021), pp. 1646–1647. DOI: 10.1016/j.jaci.2021.03.003.

[Chu+22] Tinyi Chu et al. “Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology”. In: *Nature Cancer* 3.4 (2022), pp. 505–517. DOI: 10.1038/s43018-022-00356-3.

[Cią+21] Magdalena Ciążyńska et al. “Ultraviolet Radiation and Chronic Inflammation–Molecules and Mechanisms Involved in Skin Carcinogenesis: A Narrative Review”. In: *Life* 11.4 (Apr. 2021), p. 326. ISSN: 2075-1729. DOI: 10.3390/life11040326.

[CKV13] M Emre Celebi, Hassan A Kingravi, and Patricio A Vela. “A comparative study of efficient initialization methods for the k-means clustering algorithm”. In: *Expert systems with applications* 40.1 (2013), pp. 200–210. DOI: https://doi.org/10.1016/j.eswa.2012.07.021.

[CL15] Camilla Calì and Maria Longobardi. “Some mathematical properties of the ROC curve and their applications”. In: *Ricerche di Matematica* 64.2 (Oct. 2015), pp. 391–402. ISSN: 1827-3491. DOI: 10.1007/s11587-015-0246-8.

[Cla+21] Zoe A Clarke et al. “Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods”. In: *Nature protocols* 16.6 (2021), pp. 2749–2764. DOI: 10.1038/s41596-021-00534-0.

[Coh13] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, May 2013. ISBN: 9781134742707. DOI: 10.4324/9780203771587.

[Coi+05] Ronald R Coifman et al. “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps”. In: *Proceedings of the national academy of sciences* 102.21 (2005), pp. 7426–7431. DOI: 10.1073/pnas.0500334102.

[Cou+11] National Research Council et al. “Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease”. In: (2011).

[CPB18] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. “Random forest versus logistic regression: a large-scale benchmark experiment”. In: *BMC Bioinformatics* 19.1 (July 2018), pp. 1–14. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2264-5.

[CS14] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

[CSM05] Eleonora Candi, Rainer Schmidt, and Gerry Melino. "The cornified envelope: a model of cell death in the skin". In: *Nature Reviews Molecular Cell Biology* 6.4 (Apr. 2005), pp. 328–340. ISSN: 1471-0080. DOI: 10.1038/nrm1619.

[Cve+06] Rikke Skoet Cvetkovski et al. "Quality of life and depression in a population of occupational hand eczema patients". In: *Contact dermatitis* 54.2 (2006), pp. 106–111.

[CVS20] Emma S Chambers and Milica Vukmanovic-Stejic. "Skin barrier immunity and ageing". In: *Immunology* 160.2 (2020), pp. 116–125.

[Dal+15] Florence J Dalgard et al. "The psychological burden of skin diseases: a cross-sectional multicenter study among dermatological out-patients in 13 European countries". In: *Journal of Investigative Dermatology* 135.4 (2015), pp. 984–991.

[Dam+21] Giovanni Damiani et al. "The global, regional, and national burden of psoriasis: results and insights from the global burden of disease 2019 study". In: *Frontiers in medicine* 8 (2021), p. 743180.

[Das+21] Anupam Das et al. "Skin and metabolic syndrome: An evidence based comprehensive review". In: *Indian Journal of Dermatology* 66.3 (2021), p. 302. ISSN: 0019-5154. DOI: 10.4103/ijd.ijd_728_20.

[Dav+10] Batya B. Davidovici et al. "Psoriasis and Systemic Inflammatory Diseases: Potential Mechanistic Links between Skin Disease and Co-Morbid Conditions". In: *Journal of Investigative Dermatology* 130.7 (July 2010), pp. 1785–1796. ISSN: 0022-202X. DOI: 10.1038/jid.2010.103.

[DB79] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227. DOI: 10.1109/tpami.1979.4766909.

[DBL18] Georgios Douzas, Fernando Bacao, and Felix Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information sciences* 465 (2018), pp. 1–20. DOI: https://doi.org/10.1016/j.ins.2018.06.056.

[Den+24] Elena Denisenko et al. "Spatial transcriptomics reveals discrete tumour microenvironments and autocrine loops within ovarian cancer subclones". In: *Nature Communications* 15.1 (Apr. 2024), p. 2860. ISSN: 2041-1723. DOI: 10.1038/s41467-024-47271-y.

[DeP+19] Erica AK DePasquale et al. "DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data". In: *Cell reports* 29.6 (2019), pp. 1718–1727. DOI: 10.1016/j.celrep.2019.09.082.

[DER] DERMAGNOSTIX. URL: https://dermagnostix.com/de/labdisk-produkte/.

[DHS01] Richard O Duda, Peter E Hart, and David G Stork. "Pattern Classification. JohnWiley & Sons". In: *Inc.,* (2001).

[Dim+22] Daniel Dimitrov et al. "Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data". In: *Nature communications* 13.1 (2022), p. 3224.

[DiN+22] Andrew R. DiNardo et al. "Gene expression signatures identify biologically and clinically distinct tuberculosis endotypes". In: *European Respiratory Journal* 60.3 (Feb. 2022), p. 2102263. ISSN: 1399-3003. DOI: 10.1183/13993003.02263-2021.

[DK17] Christopher P Denton and Dinesh Khanna. "Systemic sclerosis". In: *The Lancet* 390.10103 (2017), pp. 1685–1699.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[DMSC23] Diana Ivonne Duarte-Mata and Mario César Salinas-Carmona. "Antimicrobial peptides immune modulation role in intracellular bacterial infection". In: *Frontiers in Immunology* 14 (2023), p. 1119574.

[Dob+13] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[Dri+22] Athina Dritsoula et al. "Angiopathic activity of LRG1 is induced by the IL-6/STAT3 pathway". In: *Scientific Reports* 12.1 (Mar. 2022), p. 4867. ISSN: 2045-2322. DOI: 10.1038/s41598-022-08516-2.

[DRM11] G. De Rosa and C. Mignogna. “THE HISTOPATHOLOGY OF PSORIASIS”. In: *Reumatismo* 59.s1 (Sept. 2011), pp. 46–48. ISSN: 0048-7449. DOI: 10.4081/reumatismo.2007.1s.46.

[DRS18] Angelo Duó, Mark D Robinson, and Charlotte Soneson. “A systematic performance evaluation of clustering methods for single-cell RNA-seq data”. In: *F1000Research* 7 (2018). DOI: doi:10.12688/f1000research.15666.1.

[Dun61] Olive Jean Dunn. “Multiple comparisons among means”. In: *Journal of the American statistical association* 56.293 (1961), pp. 52–64. DOI: 10.2307/2282330.

[Dur+09] Steffen Durinck et al. “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt”. In: *Nature Protocols* 4.8 (July 2009), pp. 1184–1191. ISSN: 1750-2799. DOI: 10.1038/nprot.2009.97.

[Eck+13] Leopold Eckhart et al. “Cell death by cornification”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1833.12 (2013), pp. 3471–3480. DOI: 10.1016/j.bbamcr.2013.06.010.

[Ede+19] Lihi Eder et al. “Cardiovascular Diseases in Psoriasis and Psoriatic Arthritis”. In: *The Journal of Rheumatology* 95 (June 2019), pp. 20–27. ISSN: 1499-2752. DOI: 10.3899/jrheum.190114.

[EE18] K Eyerich and S Eyerich. “Immune response patterns in non-communicable inflammatory skin diseases”. In: *Journal of the European Academy of Dermatology and Venereology* 32.5 (Jan. 2018), pp. 692–703. ISSN: 1468-3083. DOI: 10.1111/jdv.14673.

[EG99] Kathleen B Elmer and Rita M George. “Cutaneous T-cell lymphoma presenting as benign dermatoses”. In: *American family physician* 59.10 (1999), pp. 2809–2813.

[EN17] Mohamed Elfil and Ahmed Negida. “Sampling methods in clinical research; an educational review”. In: *Emergency* 5.1 (2017).

[Est+96] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[Ewa+15] David A Ewald et al. “Meta-analysis derived atopic dermatitis (MADAD) transcriptome defines a robust AD signature highlighting the involvement of atherosclerosis and lipid metabolism pathways”. In: *BMC medical genomics* 8 (2015), pp. 1–15. DOI: 10.1186/s12920-015-0133-x.

[Ewe+16] Philip Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

[Eye+09] Stefanie Eyerich et al. “Th22 cells represent a distinct human T cell subset involved in epidermal immunity and remodeling”. In: *The Journal of clinical investigation* 119.12 (2009), pp. 3573–3585.

[EZ14] Stefanie Eyerich and Christina E Zielinski. “Defining Th-cell subsets in a classical and tissue-specific manner: Examples from the skin”. In: *European Journal of Immunology* 44.12 (Nov. 2014), pp. 3475–3483. ISSN: 1521-4141. DOI: 10.1002/eji.201444891.

[Fis+21] Ronald Aylmer Fisher et al. “014: On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample.” In: (1921).

[Fis+23] Felix Fischer et al. “Gene Expression–Based Molecular Test as Diagnostic Aid for the Differential Diagnosis of Psoriasis and Eczema in Formalin-Fixed and Paraffin-Embedded Tissue, Microbiopsies, and Tape Strips”. In: *Journal of Investigative Dermatology* (2023). DOI: https://doi.org/10.1016/j.jid.2023.02.015.

[FM10] Daniel A Fletcher and R Dyche Mullins. “Cell mechanics and the cytoskeleton”. In: *Nature* 463.7280 (Jan. 2010), pp. 485–492. ISSN: 1476-4687. DOI: 10.1038/nature08908.

[FNS85] Eugene M Farber, Lexie Nall, and Aaron Strefling. “Psoriasis: A disease of the total skin”. In: *Journal of the American Academy of Dermatology* 12.1 (Jan. 1985), pp. 150–156. ISSN: 0190-9622. DOI: 10.1016/s0190-9622(85)70019-9.

[For10] Santo Fortunato. “Community detection in graphs”. In: *Physics reports* 486.3-5 (2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

[For65] Edward W Forgy. “Cluster analysis of multivariate data: efficiency versus interpretability of classifications”. In: *biometrics* 21 (1965), pp. 768–769.

[Fos19] D Foster. “Generative Deep Learning. Teaching Machines to Paint, Write, Compose and Play (2019)”. In: *Beijing-Boston-Farnham-Sebastopol-Tokyo, OREILLY* (2019), p. 330.

[Fre] Freepik. *Clustering icons.* https://www.flaticon.com/free-icons/clustering. Last visit: 03.02.2025.

[Fre+18] Saskia Freytag et al. “Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data”. In: *F1000Research* 7 (2018). DOI: doi:10.12688/f1000research.15809.1.

[FST23] David S Fischer, Anna C Schaar, and Fabian J Theis. “Modeling intercellular communication in tissues using spatial graphs of cells”. In: *Nature Biotechnology* 41.3 (2023), pp. 332–336. DOI: https://doi.org/10.1038/s41587-022-01467-z.

[Fur+18] K Furue et al. “Highlighting Interleukin-36 Signalling in Plaque Psoriasis and Pustular Psoriasis”. In: *Acta Dermato Venereologica* 98.1 (2018), pp. 5–13. ISSN: 0001-5555. DOI: 10.2340/00015555-2808.

[Gab71] Karl Ruben Gabriel. “The biplot graphic display of matrices with application to principal component analysis”. In: *Biometrika* 58.3 (1971), pp. 453–467. DOI: 10.2307/2334381.

[Gal+18] Lorenzo Galluzzi et al. “Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018”. In: *Cell Death & Differentiation* 25.3 (2018), pp. 486–541.

[Gar+16] Natalie Garzorz-Stark et al. “A novel molecular disease classifier for psoriasis and eczema”. In: *Experimental Dermatology* 25.10 (Sept. 2016), pp. 767–774. ISSN: 1600-0625. DOI: 10.1111/exd.13077.

[GDF14] Farzam Gorouhi, Parastoo Davari, and Nasim Fazel. “Cutaneous and mucosal lichen planus: a comprehensive review of clinical subtypes, risk factors, diagnosis, and prognosis”. In: *The Scientific World Journal* 2014.1 (2014), p. 742826.

[Gen] 10x Genomics. *Map the whole transcriptome within the tissue context.* URL: https://www.10xgenomics.com/products/spatial-gene-expression (visited on 08/28/2024).

[Gen+04] Robert C Gentleman et al. “Bioconductor: open software development for computational biology and bioinformatics”. In: *Genome biology* 5 (2004), pp. 1–16.

[Gen22] 10x Genomics. *Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 (Dual Index).* 2022. URL: https://cdn.10xgenomics.com/image/upload/v1668017706/support-documents/CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex__RevE.pdf (visited on 02/15/2024).

[Gho+15] Debajyoti Ghosh et al. "Multiple transcriptome data analysis reveals biologically relevant atopic dermatitis signature genes and pathways". In: *PloS one* 10.12 (2015), e0144316. DOI: 10.1371/journal.pone.0144316.

[Gho+20] Manosij Ghosh et al. "A wrapper-filter feature selection technique based on ant colony optimization". In: *Neural Computing and Applications* 32 (2020), pp. 7839–7857.

[Gil+22] Marc Gillespie et al. "The reactome pathway knowledgebase 2022". In: *Nucleic acids research* 50.D1 (2022), pp. D687–D692. DOI: 10.1093/nar/gkad1025.

[Gis+23] Paolo Gisondi et al. "Quality of life and stigmatization in people with skin diseases in Europe: A large survey from the 'burden of skin diseases' EADV project". In: *Journal of the European Academy of Dermatology and Venereology* 37 (2023), pp. 6–14.

[GK65] Gene Golub and William Kahan. "Calculating the singular values and pseudo-inverse of a matrix". In: *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 2.2 (1965), pp. 205–224. DOI: 10.1137/0702016.

[GKK12] Lorenzo Galluzzi, Oliver Kepp, and Guido Kroemer. "Mitochondria: master regulators of danger signalling". In: *Nature reviews Molecular cell biology* 13.12 (2012), pp. 780–788. DOI: https://doi.org/10.1038/nrm3479.

[GO15] Dominic Grün and Alexander van Oudenaarden. "Design and analysis of single-cell sequencing experiments". In: *Cell* 163.4 (2015), pp. 799–810. DOI: 10.1016/j.cell.2015.10.039.

[Gow+08] GA Nagana Gowda et al. "Metabolomics-based methods for early disease diagnostics". In: *Expert review of molecular diagnostics* 8.5 (2008), pp. 617–633. DOI: https://doi.org/10.1586/14737159.8.5.617.

[GR20] James Fletcher Elizabeth Poyner Muzlifah Haniffa Fiona Watt Sarah Teichmann Gary Reynolds Peter Vegh. *Data for Developmental cell programs are co-opted in inflammatory skin disease - filtered, annotated anndata object (1.5) [Data set]*. 2020.

[Gre+22] Michael Greenacre et al. "Principal component analysis". In: *Nature Reviews Methods Primers* 2.1 (2022), p. 100. DOI: 10.1038/s43586-022-00184-w.

[Gri+21] Christopher E M Griffiths et al. "Psoriasis". In: *The Lancet* 397.10281 (10281 Apr. 2021), pp. 1301–1315. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(20)32549-6.

[GS+18] Natalie Garzorz-Stark et al. “Toll-like receptor 7/8 agonists stimulate plasmacytoid dendritic cells to initiate TH17-deviated acute contact dermatitis in human subjects”. In: *Journal of Allergy and Clinical Immunology* 141.4 (2018), pp. 1320–1333. DOI: 10.1016/j.jaci.2017.07.045.

[GT10] Stephen J. Galli and Mindy Tsai. “Mast cells in allergy and infection: Versatile effector and regulatory cells in innate and adaptive immunity”. In: *European Journal of Immunology* 40.7 (June 2010), pp. 1843–1851. ISSN: 1521-4141. DOI: 10.1002/eji.201040559.

[Gud+09] Johann E Gudjonsson et al. “Global gene expression analysis reveals evidence for decreased lipid biosynthesis and increased innate immunity in uninvolved psoriatic skin”. In: *Journal of Investigative Dermatology* 129.12 (2009), pp. 2795–2804. DOI: 10.1038/jid.2009.173.

[Guy+02] Isabelle Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46 (2002), pp. 389–422.

[HA+21] Siti Haryanti Hairol Anuar et al. “Comparison between Louvain and Leiden Algorithm for Network Structure: A Review”. In: *Journal of Physics: Conference Series* 2129.1 (Dec. 2021), p. 012028. ISSN: 1742-6596. DOI: 10.1088/1742-6596/2129/1/012028.

[Hab23] Olivia Habern. *FAQs about single cell sample preparation (covering the basics)*. https://www.10xgenomics.com/blog/faqs-about-single-cell-sample-preparation-covering-the-basics. Accessed: 2024-03-27. 2023.

[HAL19] Emmilia Hodak and Iris Amitay-Laish. “Mycosis fungoides: A great imitator”. In: *Clinics in Dermatology* 37.3 (May 2019), pp. 255–267. ISSN: 0738-081X. DOI: 10.1016/j.clindermatol.2019.01.004.

[HAN] HANIS. *Medicine icons*. https://www.flaticon.com/free-icons/medicine. Last visit: 03.02.2025.

[Han80] Jon M Hanifin. “Diagnostic features of atopic dermatitis”. In: *Acta Derm Venereol.(Stockh)* 92 (1980), p. 236. DOI: 10.2340/00015555924447.

[Haq+17] Ashraful Haque et al. “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”. In: *Genome medicine* 9 (2017), pp. 1–12. DOI: 10.1186/s13073-017-0467-4.

[Has+24] Iraj Hasan et al. “Dupilumab therapy for atopic dermatitis is associated with increased risk of cutaneous T cell lymphoma: A retrospective cohort study”. In: *Journal of the American Academy of Dermatology* (2024).

[Hay+23] Dylan Haynes et al. “Pityriasis Rubra Pilaris Transcriptomics Implicate T Helper 17 Signaling and Correlate with Response to Ixekizumab, with Distinct Gene Expression Profiles in Nonresponders”. In: *Journal of Investigative Dermatology* 143.3 (Mar. 2023), pp. 501–504. ISSN: 0022-202X. DOI: 10.1016/j.jid.2022.09.005.

[HBB19] Brian Hie, Bryan Bryson, and Bonnie Berger. “Efficient integration of heterogeneous single-cell transcriptomes using Scanorama”. In: *Nature biotechnology* 37.6 (2019), pp. 685–691. DOI: https://doi.org/10.1038/s41587-019-0113-3.

[HBT15] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. “Diffusion maps for high-dimensional single-cell analysis of differentiation data”. In: *Bioinformatics* 31.18 (2015), pp. 2989–2998. DOI: 10.1093/bioinformatics/btv325.

[He+22] Shanshan He et al. “High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging”. In: *Nature Biotechnology* 40.12 (2022), pp. 1794–1806. DOI: 10.1038/s41587-022-01483-z.

[Heg+22] Jana-Charlotte Hegenbarth et al. “Perspectives on bulk-tissue RNA sequencing and single-cell RNA sequencing for cardiac transcriptomics”. In: *Frontiers in Molecular Medicine* 2 (2022), p. 839338. DOI: 10.3389/fmmed.2022.839338.

[Hey20] Warren R Heymann. “Bullous pemphigoid: Rituximab to the rescue?” In: *Journal of the American Academy of Dermatology* 82.5 (2020), pp. 1089–1090. DOI: https://doi.org/10.1016/j.jaad.2020.02.058.

[HFM22] Christina Hillig, Ali Farnoud, and Michael Menden. *Spatial transcriptomics landscape of non-communicable inflammatory skin diseases*. 2022.

[HK19] Allen W Ho and Thomas S Kupper. “T cells and the skin: from protective immunity to inflammatory skin disorders”. In: *Nature Reviews Immunology* 19.8 (Apr. 2019), pp. 490–502. ISSN: 1474-1741. DOI: 10.1038/s41577-019-0162-3.

[HK99] Alexander Hinneburg and Daniel A Keim. “Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering”. In: (1999).

[HKS22] Motaharesadat Hosseini, Karl R Koehler, and Abbas Shafiee. “Biofabrication of Human Skin with Its Appendages”. In: *Advanced Healthcare Materials* 11.22 (Sept. 2022), p. 2201626. ISSN: 2192-2659. DOI: 10.1002/adhm.202201626.

[HLB18] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14.

[Hoe+21] Torsten Hoefler et al. "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks". In: *Journal of Machine Learning Research* 22.241 (2021), pp. 1–124.

[Hon+22] Rui Hong et al. "Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data". In: *Nature Communications* 13.1 (2022), p. 1688. DOI: https://doi.org/10.1038/s41467-022-29212-9.

[Hos+04] Brett M Hosking et al. "The VCAM-1 Gene That Encodes the Vascular Cell Adhesion Molecule Is a Target of the Sry-related High Mobility Group Box Gene, Sox18". In: *Journal of Biological Chemistry* 279.7 (Feb. 2004), pp. 5314–5322. ISSN: 0021-9258. DOI: 10.1074/jbc.m308512200.

[Hot33] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of educational psychology* 24.6 (1933), p. 417. DOI: 10.1037/h0070888.

[HS16] Gabriel E Hoffman and Eric E Schadt. "variancePartition: interpreting drivers of variation in complex gene expression studies". In: *BMC Bioinformatics* 17.1 (Nov. 2016), pp. 1–13. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1323-z.

[HS19] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1 (2019), p. 296. DOI: https://doi.org/10.1186/s13059-019-1874-1.

[Hsu+20] Yen-Chang Hsu et al. "Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 10951–10960. DOI: 10.1109/cvpr42600.2020.01096.

[HT22] Martijn W Heymans and Jos WR Twisk. "Handling missing data in clinical research". In: *Journal of clinical epidemiology* 151 (2022), pp. 185–188. DOI: 10.1016/j.jclinepi.2022.08.016.

[HVVK18] Deborah Huber, L Voith Von Voithenberg, and Govind V Kaigala. "Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH?" In: *Micro and Nano Engineering* 1 (2018), pp. 15–24. DOI: 10.1016/j.mne.2018.10.006.

[HWC99] Myles Hollander, Douglas A Wolfe, and E Chicken. “Nonparametric statistical methods john wiley & sons”. In: *New York* 57 (1999), pp. 58–59.

[HWM05] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”. In: *Advances in Intelligent Computing*. Springer Berlin Heidelberg, 2005, pp. 878–887. ISBN: 9783540319023. DOI: 10.1007/11538059_91.

[HZ93] Geoffrey E Hinton and Richard Zemel. “Autoencoders, minimum description length and Helmholtz free energy”. In: *Advances in neural information processing systems* 6 (1993).

[Icg] “Pan-cancer analysis of whole genomes”. In: *Nature* 578.7793 (2020), pp. 82–93. DOI: https://doi.org/10.1038/s41586-020-1969-6.

[IK24] Takeshi Inoue and Tomohiro Kurosaki. “Memory B cells”. In: *Nature Reviews Immunology* 24.1 (2024), pp. 5–17. DOI: https://doi.org/10.1038/s41577-023-00897-3.

[IM18] Laura Israel and Mark Mellett. “Clinical and Genetic Heterogeneity of CARD14 Mutations in Psoriatic Skin Disease”. In: *Frontiers in Immunology* 9 (Oct. 2018), p. 401445. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02239.

[IMBH19] Giovanni Iacono, Ramon Massoni-Badosa, and Holger Heyn. “Single-cell transcriptomics unveils gene regulatory network plasticity”. In: *Genome biology* 20 (2019), pp. 1–20. DOI: 10.1186/s13059-019-1713-4.

[Inz+04] Iñaki Inza et al. “Filter versus wrapper gene selection approaches in DNA microarray domains”. In: *Artificial Intelligence in Medicine* 31.2 (June 2004), pp. 91–103. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2004.01.007.

[IS23] Tonci Ivanisevic and Raj N Sewduth. “Multi-omics integration for the design of novel therapies and the identification of novel biomarkers”. In: *Proteomes* 11.4 (2023), p. 34.

[Jab+14] Ali Jabbari et al. “Dominant Th1 and minimal Th17 skewing in discoid lupus revealed by transcriptomic comparison with psoriasis”. In: *Journal of Investigative Dermatology* 134.1 (2014), pp. 87–95. DOI: 10.1038/jid.2014.41.

[Jai10] Anil K Jain. “Data clustering: 50 years beyond K-means”. In: *Pattern recognition letters* 31.8 (2010), pp. 651–666. DOI: 10.1016/j.patrec.2009.09.011.

[Jar+24] Manja Jargosch et al. “SERPINB3/B4 is an autoantigen driving the clinical endotype of eczematized psoriasis”. In: *In preparation* (2024).

[JD04] Sergio A Jimenez and Chris T Derk. "Following the molecular pathways toward an understanding of the pathogenesis of systemic sclerosis". In: *Annals of internal medicine* 140.1 (2004), pp. 37–50.

[JH+12] Leanne M Johnson-Huang et al. "Post-therapeutic relapse of psoriasis after CD11a blockade is associated with T cells and inflammatory myeloid DCs". In: *PLoS One* 7.2 (2012), e30308. DOI: 10.1371/journal.pone.0030308.

[Jin+21] Suoqin Jin et al. "Inference and analysis of cell-cell communication using CellChat". In: *Nature communications* 12.1 (2021), p. 1088. DOI: 10.1038/s41467-021-21246-9.

[JKP94] George H John, Ron Kohavi, and Karl Pfleger. "Irrelevant Features and the Subset Selection Problem". In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 121–129. ISBN: 9781558603356. DOI: 10.1016/b978-1-55860-335-6.50023-4.

[JLR07] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

[Joh+17] Andrew Johnston et al. "IL-1 and IL-36 are dominant cytokines in generalized pustular psoriasis". In: *Journal of Allergy and Clinical Immunology* 140.1 (2017), pp. 109–120. DOI: 10.1016/j.jaci.2016.08.056.

[Joh+20] Kaitlyn E Johnson et al. "Integrating transcriptomics and bulk time course data into a mathematical framework to describe and predict therapeutic resistance in cancer". In: *Physical biology* 18.1 (2020), p. 016001. DOI: 10.1088/1478-3975/abb09c.

[Joh+22] Brett E Johnson et al. "An omic and multidimensional spatial atlas from serial biopsies of an evolving metastatic breast cancer". In: *Cell Reports Medicine* 3.2 (2022). DOI: https://doi.org/10.1016/j.xcrm.2022.100525.

[Joh67] Stephen C Johnson. "Hierarchical clustering schemes". In: *Psychometrika* 32.3 (1967), pp. 241–254. DOI: 10.1007/bf02289588.

[Jos+22] Chintan J Joshi et al. "What are housekeeping genes?" In: *PLOS Computational Biology* 18.7 (July 2022). Ed. by Christoph Kaleta, e1010295. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010295.

[Jov+22] Dragomirka Jovic et al. "Single-cell RNA sequencing technologies and applications: A brief overview". In: *Clinical and Translational Medicine* 12.3 (2022), e694. DOI: 10.1002/ctm2.694.

[KA22] Jihyun Kim and Kangmo Ahn. “Atopic dermatitis endotypes: Knowledge for personalized medicine”. In: *Current Opinion in Allergy and Clinical Immunology* 22.3 (2022), pp. 153–159. DOI: 10.1097/aci.0000000000000820.

[Kab+19] Kenji Kabashima et al. “The immunological anatomy of the skin”. In: *Nature Reviews Immunology* 19.1 (2019), pp. 19–30.

[Kam+10] M Kamsteeg et al. “Molecular diagnostics of psoriasis, atopic dermatitis, allergic contact dermatitis and irritant contact dermatitis”. In: *British journal of dermatology* 162.3 (2010), pp. 568–578.

[Kan13] Hyun Kang. “The prevention and handling of the missing data”. In: *Korean journal of anesthesiology* 64.5 (2013), p. 402. DOI: 10.4097/kjae.2013.64.5.402.

[Kan20] Nobuo Kanazawa. “Designation of autoinflammatory skin manifestations with specific genetic backgrounds”. In: *Frontiers in Immunology* 11 (2020), p. 475. DOI: 10.3389/fimmu.2020.00475.

[Kar+21] Peter D Karp et al. “Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics”. In: *BMC genomics* 22 (2021), pp. 1–11. DOI: 10.1186/s12864-021-07502-8.

[Kas+20] Yukie Kashima et al. “Single-cell sequencing techniques from individual to multiomics analyses”. In: *Experimental & Molecular Medicine* 52.9 (2020), pp. 1419–1427. DOI: 10.1038/s12276-020-00499-2.

[Ke+13] Rongqin Ke et al. “In situ sequencing for RNA analysis in preserved tissue and cells”. In: *Nature methods* 10.9 (2013), pp. 857–860. DOI: 10.1038/nmeth.2563.

[KG00] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.

[Kik+93] Arata Kikuchi et al. “Parapsoriasis en plaques: Its potential for progression to malignant lymphoma”. In: *Journal of the American Academy of Dermatology* 29.3 (Sept. 1993), pp. 419–422. ISSN: 0190-9622. DOI: 10.1016/0190-9622(93)70204-7.

[Kim+11] Byung Eui Kim et al. “TNF-α downregulates filaggrin and loricrin through c-Jun N-terminal kinase: role for TNF-α antagonists to improve skin barrier”. In: *Journal of investigative dermatology* 131.6 (2011), pp. 1272–1279. DOI: 10.1038/jid.2011.24.

[Kim+15] Si-Heon Kim et al. “Psychological distress in young adult males with atopic dermatitis: a cross-sectional study”. In: *Medicine* 94.23 (2015), e949.

[KJ97] Ron Kohavi and George H John. “Wrappers for feature subset selection”. In: *Artificial Intelligence* 97.1-2 (Dec. 1997), pp. 273–324. ISSN: 0004-3702. DOI: 10.1016/s0004-3702(97)00043-x.

[KL15] John K Kruschke and Torrin M Liddell. “The Bayesian new statistics: Two historical trends converge”. In: *SSRN Electronic Journal* 2606016 (2015), p. 26. DOI: 10.2139/ssrn.2606016.

[Kle+22] Vitalii Kleshchevnikov et al. “Cell2location maps fine-grained cell types in spatial transcriptomics”. In: *Nature biotechnology* 40.5 (2022), pp. 661–671. DOI: 10.1038/s41587-021-01139-4.

[KM17] Eamonn J Keogh and Abdullah Mueen. “Curse of dimensionality.” In: *Encyclopedia of machine learning and data mining* 2017 (2017), pp. 314–315. DOI: 10.1007/978-1-4899-7687-1_192.

[Kõk+16] Sulev Kõks et al. “Psoriasis-specific RNA isoforms identified by RNA-seq analysis of 173,446 transcripts”. In: *Frontiers in Medicine* 3 (2016), p. 46.

[Kol+12] Raivo Kolde et al. “Robust rank aggregation for gene list integration and meta-analysis”. In: *Bioinformatics* 28.4 (Jan. 2012), pp. 573–580. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr709.

[Kol+17] Frank Kolbinger et al. “β-Defensin 2 is a responsive biomarker of IL-17A–driven skin pathology in patients with psoriasis”. In: *Journal of Allergy and Clinical Immunology* 139.3 (Mar. 2017), pp. 923–932. ISSN: 0091-6749. DOI: 10.1016/j.jaci.2016.06.038.

[Kom+22] Mayumi Komine et al. “Keratinocytes in Skin Disorders: The Importance of Keratinocytes as a Barrier”. In: *Keratinocyte Biology - Structure and Function in the Epidermis*. IntechOpen, Sept. 2022, p. 87. ISBN: 9781803551005. DOI: 10.5772/intechopen.103732.

[Kön+17] Inke R König et al. “What is precision medicine?” In: *European respiratory journal* 50.4 (2017).

[Kon+19] Robert J Konrad et al. “Assessment and Clinical Relevance of Serum IL-19 Levels in Psoriasis and Atopic Dermatitis Using a Sensitive and Specific Novel Immunoassay”. In: *Scientific Reports* 9.1 (Mar. 2019), p. 5211. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41609-z.

[Kon94] Igor Kononenko. “Estimating attributes: Analysis and extensions of RELIEF”. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1994, pp. 171–182. ISBN: 9783540483656. DOI: 10.1007/3-540-57868-4_57.

[KR92] Kenji Kira and Larry A Rendell. “The feature selection problem: Traditional methods and a new algorithm”. In: *Proceedings of the tenth national conference on Artificial intelligence*. 1992, pp. 129–134.

[Kra+23a] Thomas Krausgruber et al. “Single-cell and spatial transcriptomics reveal aberrant lymphoid developmental programs driving granuloma formation”. In: *Immunity* 56.2 (2023), pp. 289–306. DOI: https://doi.org/10.1016/j.immuni.2023.01.014.

[Kra+23b] Luke M Kraven et al. “Cluster analysis of transcriptomic datasets to identify endotypes of idiopathic pulmonary fibrosis”. In: *Thorax* 78.6 (2023), pp. 551–558.

[Kru12] James G Krueger. “Hiding under the skin: A welcome surprise in psoriasis”. In: *Nature Medicine* 18.12 (Dec. 2012), pp. 1750–1751. ISSN: 1546-170X. DOI: 10.1038/nm.3025.

[KVR21] David Källberg, Linda Vidman, and Patrik Rydén. “Comparison of Methods for Feature Selection in Clustering of High-Dimensional RNA-Sequencing Data to Identify Cancer Subtypes”. In: *Frontiers in Genetics* 12 (Feb. 2021), p. 632620. ISSN: 1664-8021. DOI: 10.3389/fgene.2021.632620.

[Lan21] The Lancet. *20 years of precision medicine in oncology*. 2021. DOI: https://doi.org/10.1016/S0140-6736(21)01099-0.

[Lan+22] Marius Lange et al. “CellRank for directed single-cell fate mapping”. In: *Nature methods* 19.2 (2022), pp. 159–170. DOI: 10.1038/s41592-021-01346-6.

[Lar+21] Samantha B Larsen et al. “Establishment, maintenance, and recall of inflammatory memory”. In: *Cell Stem Cell* 28.10 (Oct. 2021), 1758–1774.e8. ISSN: 1934-5909. DOI: 10.1016/j.stem.2021.07.001.

[Lau+18] Felix Lauffer et al. “Type I Immune Response Induces Keratinocyte Necroptosis and Is Associated with Interface Dermatitis”. In: *Journal of Investigative Dermatology* 138.8 (Aug. 2018), pp. 1785–1794. ISSN: 0022-202X. DOI: 10.1016/j.jid.2018.02.034.

[LE23] Felix Lauffer and Kilian Eyerich. “Eczematized psoriasis – a frequent but often neglected variant of plaque psoriasis”. In: *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 21.5 (Feb. 2023), pp. 445–453. ISSN: 1610-0387. DOI: 10.1111/ddg.14991.

[Lee00] Sauchi Stephen Lee. “Noisy replication in skewed binary classification”. In: *Computational statistics & data analysis* 34.2 (2000), pp. 165–191. DOI: 10.1016/s0167-9473(99)00095-x.

[Lee+06] Su-In Lee et al. “Efficient l˜ 1 regularized logistic regression”. In: *Aaai*. Vol. 6. 2006, pp. 401–408.

[Lee+12] Jeffrey T Leek et al. “The sva package for removing batch effects and other unwanted variation in high-throughput experiments”. In: *Bioinformatics* 28.6 (2012), pp. 882–883.

[Lee+15] Je Hyuk Lee et al. “Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues”. In: *Nature protocols* 10.3 (2015), pp. 442–458. DOI: 10.17504/protocols.io.mgsc3we.

[Les+15] YA Leshem et al. “What the Eczema Area and Severity Index score tells us about the severity of atopic dermatitis: an interpretability study”. In: *British Journal of Dermatology* 172.5 (2015), pp. 1353–1357.

[LH22] single-cell best practices consortium Lukas Heumos Anna Schaar. *Single-cell best practices*. 2022.

[LHA14] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21. DOI: 10.1101/002832.

[Li19] Yuxi Li. “Reinforcement learning applications”. In: *arXiv preprint arXiv:1908.06973* (2019).

[Li+20] Chao Li et al. “A new feature selection algorithm based on relevance, redundancy and complementarity”. In: *Computers in Biology and Medicine* 119 (Apr. 2020), p. 103667. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2020.103667.

[Li+22] Bin Li et al. “Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution”. In: *Nature methods* 19.6 (2022), pp. 662–670. DOI: 10.1038/s41592-022-01480-9.

[Li+23] Zhuoxuan Li et al. “SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns”. In: *Nature communications* 14.1 (2023), p. 3995. DOI: https://doi.org/10.1038/s41467-023-39608-w.

[Li+24] Tianhao Li et al. “CancerGPT for few shot drug pair synergy prediction using large pretrained language models”. In: *npj Digital Medicine* 7.1 (Feb. 2024), p. 40. ISSN: 2398-6352. DOI: 10.1038/s41746-024-01024-9.

[Lia+17] Yun Liang et al. “Psoriasis: a mixed autoimmune and autoinflammatory disease”. In: *Current Opinion in Immunology* 49 (Dec. 2017), pp. 1–8. ISSN: 0952-7915. DOI: 10.1016/j.coi.2017.07.007.

[Lia+19] Jun Liang et al. “Feature Selection with Conditional Mutual Information Considering Feature Interaction”. In: *Symmetry* 11.7 (July 2019), p. 858. ISSN: 2073-8994. DOI: 10.3390/sym11070858.

[Lin+24] Wei Lin et al. “GLIPR2: a potential biomarker and therapeutic target unveiled – Insights from extensive pan-cancer analyses, with a spotlight on lung adenocarcinoma”. In: *Frontiers in Immunology* 15 (Feb. 2024), p. 1280525. ISSN: 1664-3224. DOI: 10.3389/fimmu.2024.1280525.

[Lit19] Thomas Litman. “Personalized medicine—concepts, technologies, and applications in inflammatory skin diseases”. In: *APMIS* 127.5 (May 2019), pp. 386–424. ISSN: 1600-0463. DOI: 10.1111/apm.12934.

[Lit88] Roderick JA Little. “Missing-data adjustments in large surveys”. In: *Journal of Business & Economic Statistics* 6.3 (1988), pp. 287–296. DOI: https://doi.org/10.2307/1391878.

[Liu+22] Yale Liu et al. “Classification of human chronic inflammatory skin disease based on single-cell immune profiling”. In: *Science Immunology* 7.70 (Apr. 2022). ISSN: 2470-9468. DOI: 10.1126/sciimmunol.abl9165.

[Liu+24] Quanlei Liu et al. “Single-cell, single-nucleus and xenium-based spatial transcriptomics analyses reveal inflammatory activation and altered cell interactions in the hippocampus in mice with temporal lobe epilepsy”. In: *Biomarker Research* 12.1 (2024), p. 103.

[LK19] Tim Lämmermann and Wolfgang Kastenmüller. “Concepts of GPCR-controlled navigation in the immune system”. In: *Immunological reviews* 289.1 (2019), pp. 205–231.

[LL13] William Michael Landau and Peng Liu. “Dispersion Estimation and Its Effect on Test Performance in RNA-seq Data Analysis: A Simulation-Based Comparison of Methods”. In: *PLoS ONE* 8.12 (Dec. 2013). Ed. by Lin Chen, e81415. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0081415.

[LLBM16] Aaron T L. Lun, Karsten Bach, and John C Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome biology* 17 (2016), pp. 1–14. DOI: https://doi.org/10.1186/s13059-016-0947-7.

[Llo82] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/tit.1982.1056489.

[LMM16] Aaron TL Lun, Davis J McCarthy, and John C Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor". In: *F1000Research* 5 (2016). DOI: 10.12688/f1000research.9501.2.

[LNA17] Guillaume LemaÃŽtre, Fernando Nogueira, and Christos K Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning". In: *Journal of machine learning research* 18.17 (2017), pp. 1–5.

[Lop+18] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature methods* 15.12 (2018), pp. 1053–1058. DOI: https://doi.org/10.1038/s41592-018-0229-2.

[LSFK14] Michelle A. Lowes, Mayte Suárez-Fariñas, and James G. Krueger. "Immunology of Psoriasis". In: *Annual Review of Immunology* 32.1 (Mar. 2014), pp. 227–255. ISSN: 1545-3278. DOI: 10.1146/annurev-immunol-032713-120225.

[LT19] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular systems biology* 15.6 (2019), e8746. DOI: https://doi.org/10.15252/msb.20188746.

[Lu+24] Diyuan Lu et al. "Enhancing Gene Expression Representation and Drug Response Prediction with Data Augmentation and Gene Emphasis". In: *bioRxiv* (2024), pp. 2024–05. DOI: https://doi.org/10.1101/2024.05.15.592959.

[Lue+22] Malte D Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature methods* 19.1 (2022), pp. 41–50. DOI: https://doi.org/10.1038/s41592-021-01336-8.

[Lun+04] Kathryn L Lunetta et al. "Screening large-scale association study data: exploiting interactions using random forests". In: *BMC genetics* 5 (2004), pp. 1–13.

[LUW20] Tae-Hwy Lee, Aman Ullah, and Ran Wang. "Bootstrap aggregating and random forest". In: *Macroeconomic forecasting in the era of big data: Theory and practice* (2020), pp. 389–429.

[Mag+19] Shino Magaki et al. "An introduction to the performance of immunohistochemistry". In: *Biobanking: methods and protocols* (2019), pp. 289–298. DOI: https://doi.org/10.1007/978-1-4939-8935-5_25.

[Mai+18] Yosuke Mai et al. "Bullous pemphigoid triggered by thermal burn under medication with a dipeptidyl peptidase-IV inhibitor: a case report and review of the literature". In: *Frontiers in immunology* 9 (2018), p. 542.

[Mai+23] Laura Maintz et al. "IL-13, periostin and dipeptidyl-peptidase-4 reveal endotype-phenotype associations in atopic dermatitis". In: *Allergy* (2023). DOI: 10.1111/all.15647.

[Mal+20] Eoghan R Malone et al. "Molecular profiling for precision cancer therapies". In: *Genome medicine* 12 (2020), pp. 1–19. DOI: https://doi.org/10.1186/s13073-019-0703-1.

[Man10] Teri A Manolio. "Genomewide Association Studies and Assessment of the Risk of Disease". In: *New England Journal of Medicine* 363.2 (July 2010). Ed. by W. Gregory Feero and Alan E. Guttmacher, pp. 166–176. ISSN: 1533-4406. DOI: 10.1056/nejmra0905980.

[Man+21] Matthias Mann et al. "Artificial intelligence for proteomics and biomarker discovery". In: *Cell systems* 12.8 (2021), pp. 759–770. DOI: 10.1016/j.cels.2021.06.006.

[Mar+18] Jean S Marshall et al. "An introduction to immunology and immunopathology". In: *Allergy, Asthma & Clinical Immunology* 14.2 (Sept. 2018), pp. 1–10. ISSN: 1710-1492. DOI: 10.1186/s13223-018-0278-1.

[Mar21] Vivien Marx. "Method of the Year: spatially resolved transcriptomics". In: *Nature methods* 18.1 (2021), pp. 9–14. DOI: 10.1038/s41592-020-01033-y.

[Mat+17] Tiago R. Matos et al. "Clinically resolved psoriatic lesions contain psoriasis-specific IL-17-producing $\alpha\beta$ T cell clones". In: *Journal of Clinical Investigation* 127.11 (Sept. 2017), pp. 4031–4041. ISSN: 1558-8238. DOI: 10.1172/jci93396.

[Mav+20] Emanual Maverakis et al. "Pyoderma gangrenosum". In: *Nature Reviews Disease Primers* 6.1 (Oct. 2020), p. 81. ISSN: 2056-676X. DOI: 10.1038/s41572-020-0213-x.

[McC+06] Terrill McClanahan et al. "Identification of overexpression of orphan G protein-coupled receptor GPR49 in human colon and ovarian primary tumors". In: *Cancer biology & therapy* 5.4 (2006), pp. 419–426. DOI: 10.4161/cbt.5.4.2521.

[MD+19] Paul Madley-Dowd et al. “The proportion of missing data should not be used to guide decisions on multiple imputation”. In: *Journal of clinical epidemiology* 110 (2019), pp. 63–73. DOI: 10.1016/j.jclinepi.2019.02.016.

[Men00] John Mendelsohn. “Blockade of receptors for growth factors: an anticancer therapy—the fourth annual Joseph H. Burchenal American Association for Cancer Research Clinical Research Award Lecture”. In: *Clinical Cancer Research* 6.3 (2000), pp. 747–753.

[Men13] John Mendelsohn. “Personalizing oncology: perspectives and prospects”. In: *Journal of clinical oncology* 31.15 (2013), pp. 1904–1911.

[Mer+20] Christopher R Merritt et al. “Multiplex digital spatial profiling of proteins and RNA in fixed tissue”. In: *Nature biotechnology* 38.5 (2020), pp. 586–599. DOI: 10.1038/s41587-020-0472-9.

[MG16] Scott McGrath and Dario Ghersi. “Building towards precision medicine: empowering medical professionals for the next revolution”. In: *BMC Medical Genomics* 9.1 (May 2016), pp. 1–6. ISSN: 1755-8794. DOI: 10.1186/s12920-016-0183-8.

[MH08] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[MHM18] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[Mih+24] Katarina Mihajlović et al. “Multi-omics integration of scRNA-seq time series data predicts new intervention points for Parkinson’s disease”. In: *Scientific Reports* 14.1 (2024), p. 10983. DOI: https://doi.org/10.1038/s41598-024-61844-3.

[MMP02] Pabitra Mitra, CA Murthy, and Sankar K. Pal. “Unsupervised feature selection using feature similarity”. In: *IEEE transactions on pattern analysis and machine intelligence* 24.3 (2002), pp. 301–312. DOI: 10.1109/34.990133.

[MP22] Lambda Moses and Lior Pachter. “Museum of spatial transcriptomics”. In: *Nature Methods* 19.5 (2022), pp. 534–546. DOI: 10.1038/s41592-022-01409-2.

[MR19] Katerina M Marcoulides and Tenko Raykov. “Evaluation of variance inflation factors in regression models using latent variable modeling methods”. In: *Educational and psychological measurement* 79.5 (2019), pp. 874–882.

[MRL11] Guillermo Macbeth, Eugenia Razumiejczyk, and Rubén Daniel Ledesma. "Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations". In: *Universitas Psychologica* 10.2 (2011), pp. 545–555.

[MT14] Giovanna Menardi and Nicola Torelli. "Training and assessing classification rules with imbalanced data". In: *Data mining and knowledge discovery* 28 (2014), pp. 92–122. DOI: 10.1007/s10618-012-0295-5.

[Mur+18] Ken Muramatsu et al. "Regulatory T-cell dysfunction induces autoantibodies to bullous pemphigoid antigens in mice and human subjects". In: *Journal of Allergy and Clinical Immunology* 142.6 (2018), pp. 1818–1830.

[Mur20] Charles D Murin. "Considerations of Antibody Geometric Constraints on NK Cell Antibody Dependent Cellular Cytotoxicity". In: *Frontiers in Immunology* 11 (July 2020), p. 1635. ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.01635.

[Mur22] Kevin P Murphy. *Probabilistic machine learning: an introduction.* MIT press, 2022.

[Nai+17] Shruti Naik et al. "Inflammatory memory sensitizes skin epithelial stem cells to tissue damage". In: *Nature* 550.7677 (Oct. 2017), pp. 475–480. ISSN: 1476-4687. DOI: 10.1038/nature24271.

[Nai+21] Nardeep Naithani et al. "Precision medicine: Concept and tools". In: *Medical Journal Armed Forces India* 77.3 (July 2021), pp. 249–257. ISSN: 0377-1237. DOI: 10.1016/j.mjafi.2021.06.021.

[NBG22] Agnieszka Nowak-Brzezińska and Igor Gaibei. "How the outliers influence the quality of clustering?" In: *Entropy* 24.7 (2022), p. 917. DOI: 10.3390/e24070917.

[NCK11] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification". In: *International Journal of Knowledge Engineering and Soft Data Paradigms* 3.1 (2011), pp. 4–21. DOI: 10.1504/ijkesdp.2011.039875.

[NE21] Prakhar Varshney Neil Ernst Tim Menzies. *cliffsDelta.* 2021.

[NG04] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* 69.2 (2004), p. 026113. DOI: 10.1103/physreve.69.026113.

[Ng04] Andrew Y Ng. "Feature selection, L 1 vs. L 2 regularization, and rotational invariance". In: *Proceedings of the twenty-first international conference on Machine learning.* 2004, p. 78.

[NL21] Xinhui Ni and Yuping Lai. "Crosstalk between keratinocytes and immune cells in inflammatory skin diseases". In: *Exploration of Immunology* 1.5 (Dec. 2021), pp. 418–431. ISSN: 2768-6655. DOI: 10.37349/ei.2021.00028.

[Nos+21] Audrey Nosbaum et al. "Psoriasis is a disease of the entire skin: non-lesional skin displays a prepsoriasis phenotype". In: *European Journal of Dermatology* 31.2 (Apr. 2021), pp. 143–154. ISSN: 1952-4013. DOI: 10.1684/ejd.2021.4015.

[NS19] Alan V Nguyen and Athena M Soulika. "The Dynamics of the Skin's Immune System". In: *International Journal of Molecular Sciences* 20.8 (Apr. 2019), p. 1811. ISSN: 1422-0067. DOI: 10.3390/ijms20081811.

[Nur+22] Sergey Nurk et al. "The complete sequence of a human genome". In: *Science* 376.6588 (2022), pp. 44–53.

[Oak09] Dr. Amanda Oakley. *SCORAD*. https://dermnetnz.org/topics/scorad. Last visit: 08.03.2024. 2009.

[O'b07] Robert M O'brien. "A caution regarding rules of thumb for variance inflation factors". In: *Quality & quantity* 41 (2007), pp. 673–690. DOI: 10.1007/s11135-006-9018-6.

[Ook+21] Tadao Ooka et al. "Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan". In: *BMJ Nutrition, Prevention & Health* 4.1 (Mar. 2021), pp. 140–148. ISSN: 2516-5542. DOI: 10.1136/bmjnph-2020-000200.

[Orl+20] Christian Orlik et al. "Keratinocytes costimulate naive human T cells via CD2: a potential target to prevent the development of proinflammatory Th1 cells in the skin". In: *Cellular & Molecular Immunology* 17.4 (2020), pp. 380–394.

[OS+20] Agnieszka Owczarczyk-Saczonek et al. "Immunological Memory of Psoriatic Lesions". In: *International Journal of Molecular Sciences* 21.2 (Jan. 2020), p. 625. ISSN: 1422-0067. DOI: 10.3390/ijms21020625.

[Ozd+18] Cevdet Ozdemir et al. "The concepts of asthma endotypes and phenotypes to guide current and novel treatment strategies". In: *Expert review of respiratory medicine* 12.9 (2018), pp. 733–743.

[Özm+19] Vahit Özmen et al. "Cost effectiveness of gene expression profiling in patients with early-stage breast cancer in a middle-income country, Turkey: results of a prospective multicenter study". In: *European Journal of Breast Health* 15.3 (2019), p. 183. DOI: 10.5152/ejbh.2019.4761.

[Pal+22] Giovanni Palla et al. “Squidpy: a scalable framework for spatial omics analysis”. In: *Nature methods* 19.2 (2022), pp. 171–178. DOI: 10.1038/s41592-021-01358-2.

[Pat+21] Harshil Patel et al. *nf-core/rnaseq: nf-core/rnaseq v3.3 - Bronze Bear*. 2021.

[PB06] A Karolina Palucka and Jacques Banchereau. “Langerhans cells: daughters of monocytes”. In: *Nature Immunology* 7.3 (Mar. 2006), pp. 223–224. ISSN: 1529-2916. DOI: 10.1038/ni0306-223.

[Ped+11] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[PF01] Foster Provost and Tom Fawcett. “Robust classification for imprecise environments”. In: *Machine learning* 42 (2001), pp. 203–231.

[Pic+22] Stephen R. Piccolo et al. “The ability to classify patients based on gene-expression data varies by algorithm and performance metric”. In: *PLOS Computational Biology* 18.3 (Mar. 2022). Ed. by Xing Chen, e1009926. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009926.

[POM+09] Mason Alexander Porter, Jukka-Pekka Onnela, Peter J Mucha, et al. “Communities in networks”. In: (2009).

[PP] Psoriasis-Praxisnetz. *Messverfahren zur Bestimmung des Schweregrades der Psoriasis*. https://www.psoriasisnetz.info/fileadmin/Media/PDF/Downloads_Patienten/Psoriasis/Messverfahren_zur_Bestimmung_des_Schweregrades_der_Psoriasis.pdf. Last visit: 08.03.2024.

[PR+21] A Peña-Rosado et al. “Autoinflammatory Keratinization Diseases”. In: *Actas Dermo-Sifiliográficas (English Edition)* 112.10 (Nov. 2021), pp. 891–900. ISSN: 1578-2190. DOI: 10.1016/j.adengl.2021.09.005.

[Pud+22] Nicholas Pudjihartono et al. “A review of feature selection methods for machine learning-based disease risk prediction”. In: *Frontiers in Bioinformatics* 2 (2022), p. 927312. DOI: 10.3389/fbinf.2022.927312.

[Qiu20] Peng Qiu. “Embracing the dropouts in single-cell RNA-seq analysis”. In: *Nature communications* 11.1 (2020), p. 1169. DOI: 10.1038/s41467-020-14976-9.

[Qua+14] Maria Quaranta et al. “Intraindividual genome expression analysis reveals a specific molecular signature of psoriasis and eczema”. In: *Science translational medicine* 6.244 (July 2014). ISSN: 1946-6242. DOI: 10.1126/scitranslmed.3008946.

[Qui14] David Quigley. “RNA-seq permits a closer look at normal skin and psoriasis gene networks”. In: *Journal of Investigative Dermatology* 134.7 (2014), pp. 1789–1791.

[RA18] LA Rasyid and S Andayani. “Review on clustering algorithms based on data type: towards the method for data combined of numeric-fuzzy linguistics”. In: *Journal of Physics: Conference Series*. Vol. 1097. 1. IOP Publishing. 2018, p. 012082.

[Ran+21] Benjamin L Ranard et al. “Identification of Endotypes of Hospitalized COVID-19 Patients”. In: *Frontiers in Medicine* 8 (Nov. 2021), p. 770343. ISSN: 2296-858X. DOI: 10.3389/fmed.2021.770343.

[Rao+21] Anjali Rao et al. “Exploring tissue architecture using spatial transcriptomics”. In: *Nature* 596.7871 (2021), pp. 211–220. DOI: https://doi.org/10.1038/s41586-021-03634-9.

[RDW22] Anuradha Ray, Jishnu Das, and Sally E Wenzel. “Determining asthma endotypes and outcomes: Complementing existing clinical practice with modern machine learning”. In: *Cell Reports Medicine* 3.12 (2022). DOI: 10.1016/j.xcrm.2022.100857.

[REE19] Isaac D Raplee, Alexei V Evsikov, and Caralina Marín de Evsikova. “Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research”. In: *Journal of personalized medicine* 9.2 (2019), p. 18. DOI: 10.3390/jpm9020018.

[Reg+17] Aviv Regev et al. “The human cell atlas”. In: *elife* 6 (2017), e27041. DOI: 10.1101/121202.

[Rei12] K Reich. “The concept of psoriasis as a systemic inflammation: implications for disease management”. In: *Journal of the European Academy of Dermatology and Venereology* 26.s2 (Feb. 2012), pp. 3–11. ISSN: 1468-3083. DOI: 10.1111/j.1468-3083.2011.04410.x.

[Rei+19a] Jüri Reimand et al. “Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap”. In: *Nature protocols* 14.2 (2019), pp. 482–517. DOI: 10.1038/s41596-018-0103-9.

[Rei+19b] Ene Reimann et al. “Multicomponent biomarker approach improves the accuracy of diagnostic biomarkers for psoriasis vulgaris”. In: *Acta Dermato Venereologica* 99.13 (2019), pp. 1258–1265. DOI: 10.2340/00015555-3337.

[Rey+21] Gary Reynolds et al. “Developmental cell programs are co-opted in inflammatory skin disease”. In: *Science* 371.6527 (Jan. 2021). ISSN: 1095-9203. DOI: 10.1126/science.aba6500.

[Rio+21] Genevieve Rioux et al. “Development of a 3D psoriatic skin model optimized for infiltration of IL-17A producing T cells: Focus on the crosstalk between T cells and psoriatic keratinocytes”. In: *Acta biomaterialia* 136 (2021), pp. 210–222.

[Rit+15] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47. DOI: 10.1093/nar/gkv007.

[RMS10] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *bioinformatics* 26.1 (2010), pp. 139–140.

[Rom+16] Simone Romano et al. “Adjusting for chance clustering comparison measures”. In: *Journal of Machine Learning Research* 17.134 (2016), pp. 1–32.

[Rou87] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[Ruc+11] Kriangsak Ruchusatsawat et al. “Parakeratosis in skin is associated with loss of inhibitor of differentiation 4 via promoter methylation”. In: *Human Pathology* 42.12 (Dec. 2011), pp. 1878–1887. ISSN: 0046-8177. DOI: 10.1016/j.humpath.2011.02.005.

[RX22] Kaixuan Ren and Yumin Xia. “Lipocalin 2 Participates in the Epidermal Differentiation and Inflammatory Processes of Psoriasis”. In: *Journal of Inflammation Research* (2022), pp. 2157–2166. DOI: 10.2147/jir.s358492.

[Sak+99] Anavaj Sakuntabhai et al. “Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease”. In: *Nature Genetics* 21.3 (Mar. 1999), pp. 271–277. ISSN: 1546-1718. DOI: 10.1038/6784.

[Sat+15] Rahul Satija et al. “Spatial reconstruction of single-cell gene expression data”. In: *Nature biotechnology* 33.5 (2015), pp. 495–502. DOI: 10.1038/nbt.3192.

[SB12] Daniel J Stekhoven and Peter Bühlmann. “MissForest—non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1 (2012), pp. 112–118. DOI: https://doi.org/10.1093/bioinformatics/btr597.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

[Sbi+23] Emilie Sbidian et al. "Systemic pharmacological treatments for chronic plaque psoriasis: a network meta-analysis". In: *Cochrane Database of Systematic Reviews* 2023.7 (July 2023). ISSN: 1465-1858. DOI: 10.1002/14651858.cd011535.pub6.

[Sch+17] Valerie A Schneider et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly". In: *Genome research* 27.5 (2017), pp. 849–864. DOI: 10.1101/gr.213611.116.

[Sch+18] Uwe Schmidt et al. "Cell Detection with Star-Convex Polygons". In: *Lecture Notes in Computer Science.* Springer International Publishing, 2018, pp. 265–273. ISBN: 9783030009342. DOI: 10.1007/978-3-030-00934-2_30.

[Sch+21] A Schäbitz et al. "Low numbers of cytokine transcripts drive inflammatory skin diseases by initiating amplification cascades in localized epidermal clusters". In: *bioRxiv* (2021), pp. 2021–06. DOI: 10.1101/2021.06.10.447894. URL: https://www.biorxiv.org/content/10.1101/2021.06.10.447894v1.full.

[Sch+22] Alexander Schäbitz et al. "Spatial transcriptomics landscape of lesions from non-communicable inflammatory skin diseases". In: *Nature Communications* 13.1 (2022), p. 7729. DOI: https://doi.org/10.1038/s41467-022-35319-w.

[Scr] "Method of the Year 2013". In: *Nature Methods* 1 (2014). DOI: https://doi.org/10.1038/nmeth.2801.

[SEG21] A. Schäbitz, K. Eyerich, and N. Garzorz-Stark. "So close, and yet so far away: The dichotomy of the specific immune response and inflammation in psoriasis and atopic dermatitis". In: *Journal of Internal Medicine* 290.1 (Jan. 2021), pp. 27–39. ISSN: 1365-2796. DOI: 10.1111/joim.13235.

[Sei+20] P Seiringer et al. "Tofacitinib in Hypertrophic Lichen Planus". In: *Acta Dermato Venereologica* 100.14 (2020), adv00220. ISSN: 1651-2057. DOI: 10.2340/00015555-3585.

[Sei+24] Peter Seiringer et al. "Spatial transcriptomics reveals altered lipid metabolism and inflammation-related gene expression of sebaceous glands in psoriasis and atopic dermatitis". In: *Frontiers in Immunology* 15 (Feb. 2024), p. 1334844. ISSN: 1664-3224. DOI: 10.3389/fimmu.2024.1334844.

[Sel+21] Laura Selicato et al. "A New Ensemble Method for Detecting Anomalies in Gene Expression Matrices". In: *Mathematics* 9.8 (Apr. 2021), p. 882. ISSN: 2227-7390. DOI: 10.3390/math9080882.

[Ser+22] Venice Servellita et al. "A diagnostic classifier for gene expression-based identification of early Lyme disease". In: *Communications Medicine* 2.1 (July 2022), p. 92. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00127-2.

[Sew+19] Philipp Sewerin et al. "Prevalence and incidence of psoriasis and psoriatic arthritis". In: *Annals of the Rheumatic Diseases* 78.2 (2019), pp. 286–287.

[SF+11] Mayte Suárez-Fariñas et al. "Resolved psoriasis lesions retain expression of a subset of disease-related genes". In: *Journal of Investigative Dermatology* 131.2 (2011), pp. 391–400. DOI: 10.1038/jid.2010.280.

[SGSE22] Peter Seiringer, Natalie Garzorz-Stark, and Kilian Eyerich. "T-Cell–Mediated Autoimmunity: Mechanisms and Future Directions". In: *Journal of Investigative Dermatology* 142.3 (Mar. 2022), pp. 804–810. ISSN: 0022-202X. DOI: 10.1016/j.jid.2021.04.032.

[Sha+] S Shao et al. *IFN-γ enhances cell-mediated cytotoxicity against keratinocytes via JAK2/-STAT1 in lichen planus. Sci Transl Med. 2019; 11 (511): eaav7561.*

[She+23] Daniil Shevyrev et al. "Hematopoietic Stem Cells and the Immune System in Development and Aging". In: *International Journal of Molecular Sciences* 24.6 (Mar. 2023), p. 5862. ISSN: 1422-0067. DOI: 10.3390/ijms24065862.

[SHS17] Shawn J Schmieder, Chelsea D Harper, and George J Schmieder. "Granuloma annulare". In: (2017).

[Sid57] Siegel Sidney. "NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES". In: *The Journal of Nervous and Mental Disease* 125.3 (July 1957), p. 497. ISSN: 0022-3018. DOI: 10.1097/00005053-195707000-00032.

[Sie95] Hava T Siegelmann. "Computation beyond the Turing limit". In: *Science* 268.5210 (1995), pp. 545–548. DOI: 10.1007/978-1-4612-0707-8_12.

[Sim+08] Dagmar Simon et al. "Anti-CD20 (rituximab) treatment improves atopic eczema". In: *Journal of Allergy and Clinical Immunology* 121.1 (2008), pp. 122–128. DOI: https://doi.org/10.1016/j.jaci.2007.11.016.

[SK23] Dominik Saul and Robyn Laura Kosinsky. “Spatial transcriptomics herald a new era of transcriptome research”. In: *Clinical and Translational Medicine* 13.5 (2023). DOI: 10.1002/ctm2.1264.

[SKK20] Julius Schwingen, Mustafa Kaplan, and Florian C Kurschus. “current concepts in inflammatory skin diseases evolved by transcriptome analysis: In-depth analysis of atopic dermatitis and psoriasis”. In: *International journal of molecular sciences* 21.3 (2020), p. 699.

[SKZ10] Daniel F. Schwarz, Inke R. König, and Andreas Ziegler. “On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data”. In: *Bioinformatics* 26.14 (May 2010), pp. 1752–1758. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq257.

[SN17] Sarah Svenningsen and Parameswaran Nair. “Asthma endotypes and an overview of targeted therapy for asthma”. In: *Frontiers in medicine* 4 (2017), p. 158.

[Soe86] Fred F Soeprono. “Histologic Criteria for the Diagnosis of Pityriasis Rubra Pilaris”. In: *The American Journal of Dermatopathology* 8.4 (Aug. 1986), pp. 277–283. ISSN: 0193-1091. DOI: 10.1097/00000372-198608000-00001.

[Sol+21] Farzan Solimani et al. “Lichen planus–a clinical guide”. In: *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 19.6 (2021), pp. 864–882. DOI: 10.1111/ddg.14565.

[Spe61] Charles Spearman. “The proof and measurement of association between two things.” In: (1961). DOI: 10.1037/11491-005.

[SRN16] T Sajana, C Sheela Rani, and KV Narayana. “A survey on clustering techniques for big data mining”. In: *Indian journal of Science and Technology* 9.3 (2016), pp. 1–12. DOI: 10.17485/ijst/2016/v9i3/75971.

[SS18] Ryan Sacotte and Jonathan I Silverberg. “Epidemiology of adult atopic dermatitis”. In: *Clinics in dermatology* 36.5 (2018), pp. 595–605. DOI: https://doi.org/10.1016/j.clindermatol.2018.05.007.

[SS21] Robert C Sterner and Rosalie M Sterner. “CAR-T cell therapy: current limitations and potential strategies”. In: *Blood Cancer Journal* 11.4 (Apr. 2021), p. 69. ISSN: 2044-5385. DOI: 10.1038/s41408-021-00459-7.

[SSS22] McKella Sylvester, Aran Son, and Daniella M Schwartz. “The Interactions Between Autoinflammation and Type 2 Immunity: From Mechanistic Studies to Epidemiologic Associations”. In: *Frontiers in Immunology* 13 (Feb. 2022), p. 818039. ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.818039.

[Stå+16] Patrik L Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". In: *Science* 353.6294 (2016), pp. 78–82. DOI: 10.1126/science.aaf2403.

[Stä+21] Sascha Ständer et al. "Prevalence and presumptive triggers of localized bullous pemphigoid". In: *The Journal of Dermatology* 48.8 (2021), pp. 1257–1261.

[Ste+56] Hugo Steinhaus et al. "Sur la division des corps matériels en parties". In: *Bull. Acad. Polon. Sci* 1.804 (1956), p. 801.

[STM15] Oliver Stegle, Sarah A Teichmann, and John C Marioni. "Computational and analytical challenges in single-cell transcriptomics". In: *Nature Reviews Genetics* 16.3 (2015), pp. 133–145. DOI: https://doi.org/10.1038/nrg3833.

[Sto+08] Olivera Stojadinovic et al. "Deregulation of keratinocyte differentiation and activation: a hallmark of venous ulcers". In: *Journal of Cellular and Molecular Medicine* 12.6b (Dec. 2008), pp. 2675–2690. ISSN: 1582-4934. DOI: 10.1111/j.1582-4934.2008.00321.x.

[Str+21] Carsen Stringer et al. "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature methods* 18.1 (2021), pp. 100–106.

[Stu08] Student. "The probable error of a mean". In: *Biometrika* (1908), pp. 1–25. DOI: 10.2307/2331554.

[Stu+19] Tim Stuart et al. "Comprehensive integration of single-cell data". In: *Cell* 177.7 (2019), pp. 1888–1902. DOI: 10.1016/j.cell.2019.05.031.

[Sub+05] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

[Swe88] John A Swets. "Measuring the accuracy of diagnostic systems". In: *Science* 240.4857 (1988), pp. 1285–1293. DOI: 10.1126/science.3287615.

[Swi+16] William R Swindell et al. "Cross-disease transcriptomics: unique IL-17A signaling in psoriasis lesions and an autoimmune PBMC signature". In: *Journal of Investigative Dermatology* 136.9 (2016), pp. 1820–1830. DOI: 10.1016/j.jid.2016.04.035.

[SZ13] Enno Schmidt and Detlef Zillikens. "Pemphigoid diseases". In: *The Lancet* 381.9863 (2013), pp. 320–332.

[Szu10] Magdalena Szumilas. "Explaining odds ratios". In: *Journal of the Canadian academy of child and adolescent psychiatry* 19.3 (2010), p. 227.

[Szy+09] Silke Szymczak et al. "Machine learning in genome-wide association studies". In: *Genetic Epidemiology* 33.S1 (Jan. 2009), S51–S57. ISSN: 1098-2272. DOI: 10.1002/gepi.20473.

[Tan+08] Keiji Tanese et al. "G-protein-coupled receptor GPR49 is up-regulated in basal cell carcinoma and promotes cell proliferation and tumor formation". In: *The American journal of pathology* 173.3 (2008), pp. 835–843. DOI: 10.2353/ajpath.2008.071091.

[Tea22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[TH09] Jerome Friedman Trevor Hastie Robert Tibshirani. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Vol. 2. Springer New York, NY, 2009.

[Tha20] Alaa Tharwat. "Classification assessment methods". In: *Applied Computing and Informatics* 17.1 (July 2020), pp. 168–192. ISSN: 2210-8327. DOI: 10.1016/j.aci.2018.08.003.

[Thi+17] Judith L Thijs et al. "Moving toward endotypes in atopic dermatitis: identification of patient clusters based on serum biomarker analysis". In: *Journal of Allergy and Clinical Immunology* 140.3 (2017), pp. 730–737. DOI: 10.1016/j.jaci.2017.03.023.

[Tia+12] Suyan Tian et al. "Meta-Analysis Derived (MAD) Transcriptome of Psoriasis Defines the "Core" Pathogenesis of Disease". In: *PLoS ONE* 7.9 (Sept. 2012). Ed. by H. Peter Soyer, e44274. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0044274.

[Tim+16] Wilhelmina Maria Cornelia Timmermans et al. "Immunopathogenesis of granulomas in chronic autoinflammatory diseases". In: *Clinical & translational immunology* 5.12 (Dec. 2016), e118. ISSN: 2050-0068. DOI: 10.1038/cti.2016.75.

[TPM13] Juan Carlos Rojas Thomas, Matilde Santos Peñas, and Marco Mora. "New version of Davies-Bouldin index for clustering validation based on cylindrical distance". In: *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)*. IEEE. 2013, pp. 49–53.

[Tsa+21] Hou-Ren Tsai et al. “Application of Janus kinase inhibitors in atopic dermatitis: an updated systematic review and meta-analysis of clinical trials”. In: *Journal of Personalized Medicine* 11.4 (2021), p. 279. DOI: 10.3390/jpm11040279.

[Tso+19] Lam C Tsoi et al. “Atopic dermatitis is an IL-13–dominant disease with greater molecular heterogeneity compared to psoriasis”. In: *Journal of Investigative Dermatology* 139.7 (2019), pp. 1480–1489. DOI: 10.1016/j.jid.2018.12.018.

[TTT19] Sachiko Tsukita, Hiroo Tanaka, and Atsushi Tamura. “The Claudins: From Tight Junctions to Biological Systems”. In: *Trends in Biochemical Sciences* 44.2 (Feb. 2019), pp. 141–152. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2018.09.008.

[Tur20] Dana P Turner. “Sampling Methods in Research Design”. In: *Headache: The Journal of Head and Face Pain* 60.1 (Jan. 2020), pp. 8–12. ISSN: 1526-4610. DOI: 10.1111/head.13707.

[TWVE19] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1 (2019), p. 5233. DOI: 10.1038/s41598-019-41695-z.

[Tyl+24] Scott R Tyler et al. “Anti-correlated feature selection prevents false discovery of subpopulations in scRNAseq”. In: *Nature Communications* 15.1 (Jan. 2024), p. 699. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43406-9.

[Ueb] *Faktorenanalyse. Eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozial- wissenschaftler.*

[Uji+22] Hideyuki Ujiie et al. “Unmet Medical Needs in Chronic, Non-communicable Inflammatory Skin Diseases”. In: *Frontiers in Medicine* 9 (June 2022), p. 875492. ISSN: 2296-858X. DOI: 10.3389/fmed.2022.875492.

[VBGO11] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of statistical software* 45 (2011), pp. 1–67. DOI: https://doi.org/10.18637/jss.v045.i03.

[Vir+20] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[VRD09] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[VRK22] Kaushik P Venkatesh, Marium M Raza, and Joseph C Kvedar. "Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation". In: *NPJ digital medicine* 5.1 (2022), p. 150. DOI: https://doi.org/10.1038/s41746-022-00694-7.

[VS19] Paras P Vakharia and Jonathan I Silverberg. "Adult-onset atopic dermatitis: characteristics and management". In: *American Journal of Clinical Dermatology* 20 (2019), pp. 771–779.

[Wah+18] Yap Bee Wah et al. "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy." In: *Pertanika Journal of Science & Technology* 26.1 (2018).

[Wan08] He-Yong Wang. "Combination approach of SMOTE and biased-SVM for imbalanced datasets". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE. IEEE, June 2008, pp. 228–231. DOI: 10.1109/ijcnn.2008.4633794.

[Wan+10] Xiang Wan et al. "BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies". In: *The American Journal of Human Genetics* 87.3 (Sept. 2010), pp. 325–340. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2010.07.021.

[Wan+12] Fay Wang et al. "RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues". In: *The Journal of molecular diagnostics* 14.1 (2012), pp. 22–29.

[Wan+19] Xuran Wang et al. "Bulk tissue cell type deconvolution with multi-subject single-cell expression reference". In: *Nature communications* 10.1 (2019), p. 380. DOI: https://doi.org/10.1038/s41467-018-08023-x.

[Wan+23] Ye Wang et al. "Spatial transcriptomics: Technologies, applications and experimental considerations". In: *Genomics* (2023), p. 110671. DOI: 10.1016/j.ygeno.2023.110671.

[WAT18] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome biology* 19 (2018), pp. 1–5. DOI: 10.1186/s13059-017-1382-0.

[WBB18] S. Weidinger, L.A. Beck, and T et al. Bieber. "Atopic dermatitis". In: *Nature Reviews Disease Primers* 4.1 (2018). DOI: https://doi.org/10.1038/s41572-018-0001-z.

[WC21] Jacques Wainer and Gavin Cawley. “Nested cross-validation when selecting classifiers is overzealous for most practical applications”. In: *Expert Systems with Applications* 182 (2021), p. 115222. DOI: 10.1016/j.eswa.2021.115222.

[WD+18] Dominika Wcisło-Dziadecka et al. “Psoriasis Treatment Changes the Expression Profile of Selected Caspases and their Regulatory MicroRNAs”. In: *Cellular Physiology and Biochemistry* 50.2 (2018), pp. 525–537. ISSN: 1421-9778. DOI: 10.1159/000494166.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52. DOI: 10.1016/0169-7439(87)80084-9.

[Wei+06] Chi-Chen Wei et al. “IL-20: biological functions and clinical implications”. In: *Journal of Biomedical Science* 13.5 (May 2006), pp. 601–612. ISSN: 1423-0127. DOI: 10.1007/s11373-006-9087-5.

[Wel47] Bernard L Welch. “The generalization of ‘STUDENT’S’problem when several different population varlances are involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35. DOI: 10.1093/biomet/34.1-2.28.

[WGS09] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1 (2009), pp. 57–63. DOI: 10.1038/nrg2484.

[Wie+19] Julia E Wiedmeier et al. “Single-cell sequencing in precision medicine”. In: *Precision Medicine in Cancer Therapy* (2019), pp. 237–252. DOI: 10.1007/978-3-030-16391-4_9.

[Wil+22] Cameron G Williams et al. “An introduction to spatial transcriptomics for biomedical research”. In: *Genome Medicine* 14.1 (2022), pp. 1–18. DOI: 10.1186/s13073-022-01075-1.

[WJ63] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244. DOI: 10.2307/2282967.

[WL16] Ronald L Wasserstein and Nicole A Lazar. *The ASA statement on p-values: context, process, and purpose.* 2016.

[WLK19] Samuel L Wolock, Romain Lopez, and Allon M Klein. “Scrublet: computational identification of cell doublets in single-cell transcriptomic data”. In: *Cell systems* 8.4 (2019), pp. 281–291. DOI: https://doi.org/10.1016/j.cels.2018.11.005.

[Wol+19] F Alexander Wolf et al. “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. In: *Genome biology* 20 (2019), pp. 1–9. DOI: 10.1186/s13059-019-1663-x.

[WR16] Lukas M Weber and Mark D Robinson. “Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data”. In: *Cytometry Part A* 89.12 (2016), pp. 1084–1096. DOI: doi:https://doi.org/10.1002/cyto.a.23030.

[WSY21] Di Wu, Min Shen, and Qingping Yao. “Cutaneous manifestations of autoinflammatory diseases”. In: *Rheumatology and Immunology Research* 2.4 (2021), pp. 217–225.

[Wu+24] Jiale Wu et al. “Integration of single-cell sequencing and bulk RNA-seq to identify and develop a prognostic signature related to colorectal cancer stem cells”. In: *Scientific Reports* 14.1 (2024), p. 12270.

[XP16] Qi Xianting and Wang Pan. “A Density-Based Clustering Algorithm for High-Dimensional Data with Feature Selection”. In: *2016 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*. IEEE. IEEE, Dec. 2016, pp. 114–118. DOI: 10.1109/iciicii.2016.0038.

[Xu+21] Chenling Xu et al. “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models”. In: *Molecular systems biology* 17.1 (2021), e9620. DOI: https://doi.org/10.15252/msb.20209620.

[Xue+22] Dan Xue et al. “Expansion of Fcγ receptor IIIa–positive macrophages, ficolin 1–positive monocyte-derived dendritic cells, and plasmacytoid dendritic cells associated with severe skin disease in systemic sclerosis”. In: *Arthritis & Rheumatology* 74.2 (2022), pp. 329–341.

[Yan+21] Yang Yang et al. “Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data”. In: *Cell reports* 36.4 (2021). DOI: 10.1016/j.celrep.2021.109442.

[Yan+23] Yaohua Yang et al. “Integrating genomics and proteomics data to identify candidate plasma biomarkers for lung cancer risk among European descendants”. In: *British Journal of Cancer* 129.9 (2023), pp. 1510–1515.

[Yan+24] Jingkang Yang et al. “Generalized Out-of-Distribution Detection: A Survey”. In: *International Journal of Computer Vision* (June 2024). ISSN: 1573-1405. DOI: 10.1007/s11263-024-02117-4.

[YAS17] Hani Yousef, Mandy Alhajj, and Sandeep Sharma. "Anatomy, skin (integument), epidermis". In: (2017).

[YB99] Daniel Yekutieli and Yoav Benjamini. "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics". In: *Journal of Statistical Planning and Inference* 82.1-2 (1999), pp. 171–196. DOI: 10.1016/s0378-3758(99)00041-5.

[YH16] Guangchuang Yu and Qing-Yu He. "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization". In: *Molecular BioSystems* 12.2 (2016), pp. 477–479. DOI: 10.1039/c5mb00663e.

[Yu+14] Guangchuang Yu et al. "DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis". In: *Bioinformatics* 31.4 (2014), pp. 608–609. DOI: 10.1093/bioinformatics/btu684.

[Yu+19] Zi Yu et al. "PLEKHO1 knockdown inhibits RCC cell viability in vitro and in vivo, potentially by the Hippo and MAPK/JNK pathways". In: *International Journal of Oncology* 55.1 (2019), pp. 81–92. DOI: 10.3892/ijo.2019.4819.

[Yu21] Guangchuang Yu. "Enrichplot: visualization of functional enrichment result". In: *R package version* 1.2 (2021).

[Zab+08] Lisa C. Zaba et al. "Amelioration of epidermal hyperplasia by TNF inhibition is associated with reduced Th17 responses". In: *The Journal of Experimental Medicine* 205.8 (July 2008), pp. 1941–1941. ISSN: 0022-1007. DOI: 10.1084/jem.20071094071408c.

[Zen22] Hongkui Zeng. "What is a cell type and how to define it?" In: *Cell* 185.15 (2022), pp. 2739–2755. DOI: 10.1016/j.cell.2022.06.031.

[Zen+23] Hong Zeng et al. "GLIPR2 emerges as a potential predictor of prognosis for renal clear cell carcinoma, exhibiting substantial relevance with cellular metastasis and CD8+ T cell infiltration". In: *Informatics in Medicine Unlocked* (2023), p. 101371. DOI: 10.1016/j.imu.2023.101371.

[Zha12] Shichao Zhang. "Nearest neighbor selection for iteratively kNN imputation". In: *Journal of Systems and Software* 85.11 (2012), pp. 2541–2552. DOI: 10.1016/j.jss.2012.05.073.

[Zha+13] Fan Zhang et al. "Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood". In: *BMC Medical Genomics* 6.1 (Jan. 2013), pp. 1–10. ISSN: 1755-8794. DOI: 10.1186/1755-8794-6-s1-s4.

[Zha+22a] Chenlu Zhang et al. “Skin immunity: dissecting the complex biology of our body’s outer barrier”. In: *Mucosal Immunology* 15.4 (Apr. 2022), pp. 551–561. ISSN: 1933-0219. DOI: 10.1038/s41385-022-00505-y.

[Zha+22b] Linlin Zhang et al. “Clinical and translational values of spatial transcriptomics”. In: *Signal Transduction and Targeted Therapy* 7.1 (2022), p. 111. DOI: https://doi.org/10.1038/s41392-022-00960-w.

[Zha+23] Fengda Zhang et al. “Federated unsupervised representation learning”. In: *Frontiers of Information Technology & Electronic Engineering* 24.8 (Aug. 2023), pp. 1181–1193. ISSN: 2095-9230. DOI: 10.1631/fitee.2200268.

[Zhe+17] Grace XY Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nature communications* 8.1 (2017), p. 14049. DOI: 10.1038/ncomms14049.

[Zib+10] John R. Zibert et al. “MicroRNAs and potential target interactions in psoriasis”. In: *Journal of Dermatological Science* 58.3 (June 2010), pp. 177–185. ISSN: 0923-1811. DOI: 10.1016/j.jdermsci.2010.03.004.

[Zub02] Thomas J Zuber. “Punch biopsy of the skin”. In: *American family physician* 65.6 (2002), pp. 1155–1158.

[ZZ20] Xiaoliang Zhu and Jinfang Zhu. “CD4 T Helper Cell Subsets and Related Human Immunological Disorders”. In: *International Journal of Molecular Sciences* 21.21 (Oct. 2020), p. 8011. ISSN: 1422-0067. DOI: 10.3390/ijms21218011.