Technische Universität München
TUM School of Life Sciences

# Challenges and Opportunities of Computational Biomarker Discovery in Cancer

Manuela Lautizi

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Frank Johannes

Prüfende der Dissertation:

1. Prof. Dr. Dietmar Zehn
2. Prof. Dr. Jan Baumbach

Die Dissertation wurde am 17.04.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 16.10.2023 angenommen.

# Abstract

Cancer is one of the leading causes of death with nearly ten million casualties per year. Vast intratumor heterogeneity makes the establishment of treatment strategies challenging. The development of customized healthcare based on large amounts of omics data is the basis for the emerging discipline of precision oncology. One of the main objectives in the analysis of omics data is the discovery of novel biomarkers that can serve as putative drug targets or surrogates for prognosis. In this thesis, three projects were implemented to investigate established signatures and identify novel candidate biomarkers.

One of the aims of biomarker discovery in oncology is to distinguish clinically relevant cancer subtypes which differ in morphologic and molecular configuration. Pancreatic ductal adenocarcinoma (PDAC) is one of the most clinically complex types of cancer. Four studies attempted PDAC stratification and established gene signatures that in this thesis were submitted to a comprehensive computational evaluation across data sets, revealing inconsistencies. Moreover, cellular decomposition identified the presence of healthy pancreas contamination in the data as a source of bias in the subtypes characterization. Predictions obtained from random genes and shuffled label subtypes achieved comparable performances. Thus, these gene signatures can not be generalized to other datasets. These findings suggest that the heterogeneity of the tumor composition should be taken into account as well as the different platforms and sample preparation used. In addition, signatures should be carefully validated on independent datasets to control overfitting.

The accumulation of not only genetic but also epigenetic changes can lead to cancer progression. DNA methylation occurs at the CpG level and across CpGs in differentially methylated regions (DMRs). The second project of this thesis presents Dimmer 2.0, a previously published tool for the identification of de novo DMRs. In this work, Dimmer 2.0 was extended to whole-genome bisulfite sequencing and multiple array platform integration. Additionally, the analysis of a pre-processed data matrix and methylation profiles of a non-CpG context is now possible. A new approach has been introduced to assign an individual p-value to every DMR accounting for their size. The capabilities of Dimmer 2.0 are evaluated in PDAC patients with liver metastasis, where we revealed methylation alteration in regions involved in zinc homeostasis and the regulation of toll-like receptor signaling pathways. Methylation profiles of meningioma samples were used to investigate potential epigenetic attributes that explain the progression from benign to malignant stage. DMRs were found associated with constitutive signaling generated by aberrant PI3K.

The scope of the third project is the study of biomarkers of tumor budding (TB). Tumor budding is defined as the presence of single cells and clusters of less than five cells detached from the tumor bulk and settled in the adjacent stroma. While a higher count is known to be linked to worse survival, the molecular characteristics of tumor budding have not been investigated. Gene and protein expression, DNA methylation and somatic mutations were jointly analyzed. Results show the expression of two oncogenes, KCNN4 and TNFR12A linked with increased budding activity. Our findings show enrichment in the epithelial-to-mesenchymal transition process and WNT signaling pathway.

In the scope of this thesis, machine learning algorithms and statistical analyses are applied to omics data in three projects with a focus on biomarker evaluation and identification. The findings presented in this thesis provide new insights for future studies to elucidate disease mechanisms needed for improving targeted treatments. Making sense of this kind of data is a prerequisite for advancement in the area of personalized medicine.

# Kurzfassung

Krebs ist eine der häufigsten Todesursachen mit fast zehn Millionen Todesopfern pro Jahr. Die enorme Heterogenität innerhalb des Tumors macht die Entwicklung von Behandlungsstrategien zu einer Herausforderung. Die Entwicklung einer maßgeschneiderten Gesundheitsfürsorge, basierend auf großen Mengen von Omics-Daten, bildet die Grundlage für die entstehende Disziplin der Präzisionsonkologie. Eines der Hauptziele bei der Analyse von Omics-Daten ist die Entdeckung neuartiger Biomarker, die als potenzielle Angriffspunkte für Medikamente oder als Indikatoren für die Prognose dienen können. In dieser Arbeit wurden drei Projekte durchgeführt, um etablierte Signaturen zu untersuchen und neue Biomarkerkandidaten zu identifizieren.

Eines der Ziele der Entdeckung von Biomarkern in der Onkologie besteht darin, klinisch relevante Krebs-Subtypen zu unterscheiden, die sich in ihrer morphologischen und molekularen Konfiguration unterscheiden. Das duktale Adenokarzinom der Bauchspeicheldrüse (PDAC) ist eine der klinisch komplexesten Krebsarten. In vier Studien wurde versucht, das PDAC zu stratifizieren und Gensignaturen zu erstellen. Diese wurden in der vorliegenden Arbeit einer umfassenden computergestützten Bewertung über verschiedene Datensätze hinweg unterzogen, wobei sich Unstimmigkeiten zeigten. Darüber hinaus wurde bei der zellulären Dekomposition das Vorhandensein von Kontaminationen mit gesunden Pankreaszellen als eine Quelle von Verzerrungen bei der Charakterisierung der Subtypen identifiziert. Vorhersagen, die auf der Grundlage zufälliger Gene und randomisierter Subtypannotationen erstellt wurden, erzielten vergleichbare Ergebnisse. Daher können diese Gensignaturen nicht auf andere Datensätze verallgemeinert werden. Diese Ergebnisse deuten darauf hin, dass die Heterogenität der Tumorzusammensetzung ebenso berücksichtigt werden sollte wie die unterschiedlichen Plattformen und die verwendete Probenvorbereitung. Darüber hinaus sollten die Signaturen sorgfältig an unabhängigen Datensätzen validiert werden, um sogenanntes Overfitting zu vermeiden. Die Anhäufung nicht nur genetischer, sondern auch epigenetischer Veränderungen kann zum Fortschreiten von Krebs führen. DNA-Methylierung findet auf CpG-Ebene und CpG-übergreifend in differentiell methylierten Regionen (DMRs) statt. Das zweite Projekt dieser Arbeit stellt Dimmer 2.0 vor, ein bereits veröffentlichtes Tool zur Identifizierung von de novo DMRs. In dieser Arbeit wurde Dimmer 2.0 auf Whole-Genome Bisulfite Sequencing und die Integration mehrerer Array-Plattformen ausgeweitet. Zusätzlich ist nun die Analyse einer vorverarbeiteten Datenmatrix und von Methylierungsprofilen eines Nicht-CpG-Kontextes möglich. Ein neuer Ansatz wurde eingeführt, um jeder DMR einen individuellen p-Wert zuzuweisen, der ihre Größe berücksichtigt. Die

Fähigkeiten von Dimmer 2.0 wurden bei PDAC-Patienten mit Lebermetastasen evaluiert, wobei Methylierungsveränderungen in Regionen nachgewiesen werden konnten, die an der Zink-Homöostase und der Regulierung von Toll-like-Rezeptor-Signalwegen beteiligt sind. Methylierungsprofile von Meningeom-Proben wurden verwendet, um potenzielle epigenetische Merkmale zu untersuchen, die das Fortschreiten vom gutartigen zum bösartigen Stadium erklären. Es wurden DMRs gefunden, die mit einer konstitutiven Signalübertragung durch aberrante PI3K in Verbindung stehen.

Das dritte Projekt befasst sich mit der Untersuchung von Biomarkern für Tumor-Budding (TB). TB ist definiert als das Vorhandensein einzelner Zellen und Cluster von weniger als fünf Zellen, die sich von der Tumormasse ablösen und sich im angrenzenden Stroma ansiedeln. Es ist zwar bekannt, dass eine höhere Anzahl solcher Cluster mit einer schlechteren Überlebensrate einhergeht, doch die molekularen Merkmale von TB sind noch nicht untersucht worden. In diesem Kontext wurden in der vorliegenden Arbeit Gen- und Proteinexpression, DNA-Methylierung und somatische Mutationen gemeinsam analysiert. Die Ergebnisse zeigen, dass die Expression von zwei Onkogenen, KCNN4 und TNFR12A, mit einer erhöhten TB-Aktivität verbunden ist. Unsere Ergebnisse zeigen eine Anreicherung im epithelialen-zu-mesenchymalen Übergangsprozess und im WNT-Signalweg.

Im Rahmen dieser Arbeit werden Algorithmen des maschinellen Lernens und statistische Analysen auf Omics-Daten in drei Projekten angewandt, wobei der Schwerpunkt auf der Bewertung und Identifizierung von Biomarkern liegt. Die in dieser Arbeit vorgestellten Ergebnisse liefern neue Erkenntnisse für künftige Studien zur Aufklärung von Krankheitsmechanismen, die zur Verbesserung gezielter Behandlungen erforderlich sind. Diese Art von Daten sinnvoll zu nutzen, ist eine Voraussetzung für Fortschritte im Bereich der personalisierten Medizin.

# Contents

# 1 Introduction

## 1.1 Introduction to cancer biology

Cancer is one of the major causes of death in the world with approximately 19 million estimated new cases in 2020 [1]. All diseases induced by the uncontrolled division and multiplication of abnormal cells can be classified as cancer [2]. Cancer is a multistep process and even though the etiology of some tumors is unknown, there are common characteristics that contribute to the progression and advancement to more severe stages, also used as starting points for targeted therapies.

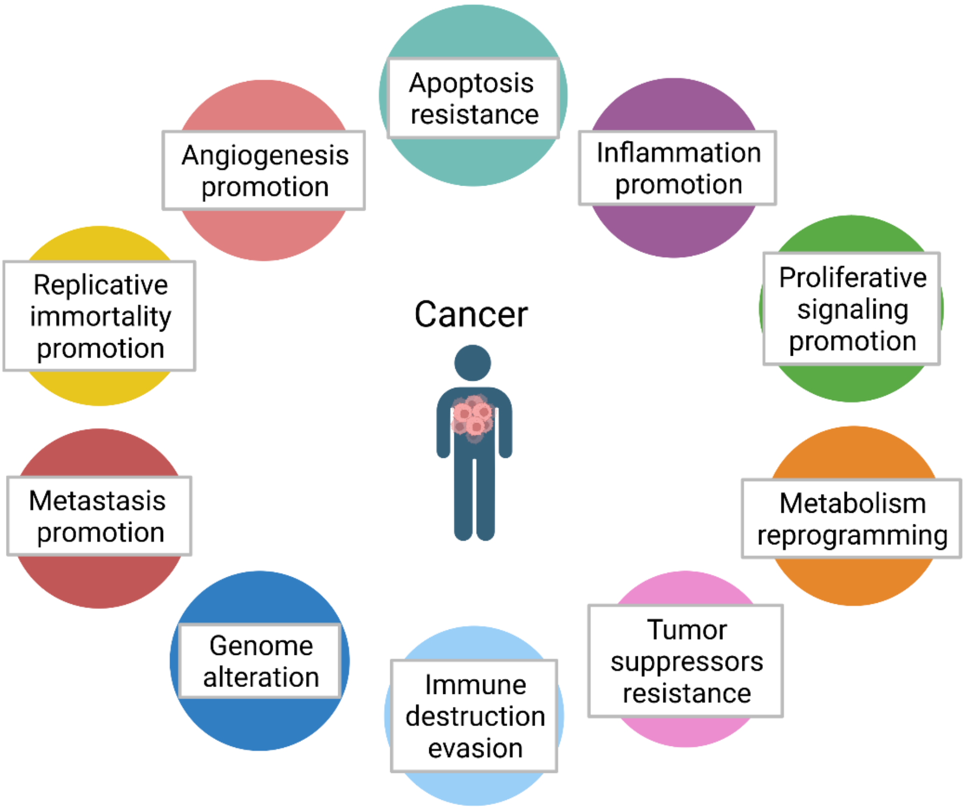


**Figure 1.1** The ten hallmarks of cancer as proposed by Hanahan and Weinberg in 2011. Figure readapted from [3].

The combination of the characteristics of cancer cells, summarized in Figure 1.1, constitute the so called hallmarks of cancer, first proposed in 2000 by Hanahan and Weinberg [4] with six main cancer features (apoptosis resistance, proliferative signaling promotion, tumor suppressor resistance, metastasis promotion, replicative immortality promotion, angiogenesis promotion), then expanded in 2011 [3] with four additional emerging hallmarks (immune destruction evasion, metabolism reprogramming, genome alteration, inflammation promotion) later scientifically consolidated and included in the list as official hallmarks.

When DNA is damaged it can alternate cellular function transforming healthy cells into cancerous ones. Damage can occur because of external causes (exogenous) such as radiation exposure or viruses, or internal to the cell (endogenous), for instance in cell metabolism or the DNA damage caused by free radicals [5]. The consequence of both exogenous and endogenous agents is the modification of the DNA structure hence mutation. A healthy organism is able to detect and repair genomic mutation, maintaining the regular state and containing the mutation occurrence. All healthy cells are subject to mutations, however, DNA in tumor cells mutates more frequently than in normal cells, generating accumulation of cancer mutations, favoring tumor proliferation and migration, angiogenesis and new unknown mutations. This mechanism is linked to the so-called "caretaker" genes, involved in blocking the rise of cancer as tumor suppressor genes, for instance [6, 7].

One main distinctive trait of cancer is the excessive hyperproliferation of cancerous cells, as a possible consequence of changes in signal transduction. Signal transduction consists in the transmission of a signal from the extracellular domain into the intracellular one through cell-surface receptors [8, 9]. The process most commonly starts when a signaling molecule named ligand binds to a receptor activating a kinase enzyme. Kinases function as regulatory elements that can to activate or deactivate proteins *via* phosphorylation. This event results in a cascade of pathways called signaling pathways that, when reaching the nucleus cell, are responsible for gene activity and ultimately cell behavior such as metabolism and mitosis stimulation [8]. Signal transduction is a fundamental process in normal cells but the presence of, for instance, gene mutations might create abnormal receptors or overexpression of receptors. Both of these changes can result in independent and excessive receptor activation, linked to aberrant signal activation and the triggering of malignant cell proliferation. Examples of signaling pathways commonly found altered in cancer are the MAPK pathway and the PI3K/AKT/mTOR pathway [8]. Cascades do not act independently but they interplay with one another, resulting in a complex signaling network.

Responsible for limiting replication in normal cells is the shortening after every replication of telomeres, DNA protein structures located at the chromosome extremities. In cancer cells, this process is inverted and telomeres are elongated thus preventing the cells to stop duplicating [10].

When cells in an organism are irreparable or dangerous, apoptosis comes into play to get rid of such unwanted cells. Apoptosis is a mechanism that kills cells by programming their death and, in the case of pathogens, contributes to limiting damaged cell replication [11]. Abnormal apoptotic activity is linked to carcinogenesis where malignant cells evade the self-disruption. The regulation of apoptosis depends on two pathways occurring in the extracellular or intracellular domain, named extrinsic and intrinsic pathways, respectively. The first one begins with a death signal while the second one, also called mitochondrial pathway, arises from a lack of survival signal. Both pathways conclude with a cascade of catalytic activation of the proteases caspase which ultimately controls cell death. DNA damage at the level of pro-apoptosis genes might lead to apoptosis inhibition that in the case of pathogenic cells increases their survival, linked to proliferation promotion [12, 13].

The new tumor requires an additional bloodstream to sustain its growth, possible thanks to the mechanisms of angiogenesis that stimulates the creation of new blood vessels from the physiological vascular system. One trait of tumoral cells is the ability to promote angiogenesis and thus sustain the tumor growth with oxygen and nutrients [14]

Another main aspect of cancer aggressiveness is that it is often not only limited to the original location but it can induce tissue invasion which leads to the initiation of new tumors. When malignant cells access the vascular or lymphatic systems, they are transported through the body and spread to different tissues forming metastasis [15].

More energy is required for malignant tumor cells to grow, which sees an increased amount of glucose transporters on the cancer cell surface to increase nutrient supply. From the glucose, energy is produced through the Warburg effect where aerobic glycolysis is adopted by the cell to break down glucose into lactate, contrary to non-neoplastic cells where the presence of oxygen would lead to an increase in oxidative phosphorylation activity at the mitochondrial level [16]. An increase in the lactic acid production rate is linked to an extracellular pH level reduction responsible for tumor progression and apoptotic resistance [16, 17].

The progression of cancer is additionally supported by its resistance to tumor suppressor genes. One example is the tumor protein P53 (TP53), the most common gene mutation observed in certain types of cancer and is known for its tumor suppressor behavior [18]. TP53 controls cell proliferation in case of damaged DNA, communicates with other genes to repair the DNA damage or, in case the repair is not possible, causes cell death preventing cancer initiation. However, a mutation in TP53 inactivates its tumor suppression functionalities and inhibits DNA repair, allowing the growth and multiplication of cells containing damaged DNA. TP53 mutation status has been proven to be an accurate prognosis indicator hence it can be used as a predictor marker for the survival of cancer patients [19].

While healthy cells can replicate only a limited number of times (from 40 to 60 times) before senescence or apoptosis [20], cancerous cells manage to evade this limit and acquire the ability to replicate an indefinite number of times.

Another cancer hallmark is the evasion of the immune response. The immune system constantly scans the body to detect and knock down emerging tumors. While a large part of rising tumors is destroyed by this system, named immunosurveillance, certain cancer cells can evade or withstand it [21]. In a functioning cancer immunity cycle, dead cancer cells release antigens which are absorbed by antigen-presenting immune cells (such as dendritic cells) [22]. Such cells, with the antigens exposed on their surface, migrate to the lymphatic organs (such as lymph nodes). Through the T cell receptor, CD8+ T cells recognize the antigens presented by the major-histocompatibility complex (MHC) class 1 and, together with costimulatory signals, get activated [23]. Activated cytotoxic T lymphocytes (CTL) move from the draining lymph nodes to the tumor site, where they infiltrate. Once there, the MHC 1 present on the cancer cells surface can bind to the CD8+ cells, and this interaction leads to the killing of the formers and the release of more antigens to continue with the immune cycle. One or more of these steps can be evaded or inhibited through several mechanisms that elude the immune response. One of these mechanisms is the antigen loss. A decrease or loss of tumor antigens and their presentation makes cells not recognizable from the CTLs [24]. In this way, there is a reduction of CTL activation, and their consequent potential migration and infiltration. One known mechanism employed for immune evasion is the downregulation of MHC class 1 [25, 26]. When there is a decreased expression of MHC 1 on cancer cells, these cells are invisible to CTLs, preventing their binding and hence the CTLs activation. Another strategy that cancer cells can adopt is the active inhibition of the host immune components. This can be done with the use of immunological checkpoints, molecules that regulate the activation of some immune cells to prevent them from attacking self-cells. One example is the receptor/ligand programmed death-1 (PD-1) / PD-1 ligand (PD-L1) [27]. The activation of cytotoxic T lymphocytes is blocked by the engagement of PD-1, expressed on T cells, with PD-L1, expressed on cancer cells. The use of immune checkpoint inhibitors is one of the strategies used in immunotherapy. Another aspect to consider is the immunosuppressive TME, characterized by the presence of inflammatory cytokines such as TGF-beta and interleukin-10 (IL-10), and the involvement of immunosuppressive cells like myeloid-derived suppressor cells, T regulatory cells and tumor-associated macrophages [28, 29]. The inflammation generated by the immune response contributes to making the TME a desirable environment for tumor growth.

Despite these characteristics being the common denominators in cancer, different cancer types behave heterogeneously showing varying levels of resistance to metastatic power. Therefore, the hallmarks can show up in different ways depending on multiple factors, like the tissue type or the genetic history of the patient.

## 1.2 Overview on the omics data

The advancement in high-throughput assay allowed the measurement of large-scale molecular data simultaneously, generating the notion of omics data [30]. All the data with the name ending in *-omics*, including genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics, are part of it. The study of the omics data in their interactions takes the name of multi-omics [30, 31].

- **Genomics**. Genomics is the branch of biology that studies the complete set of DNA of an organism, also called genome. Genomics analysis aims to identify those genetic variations associated with a condition or disease. The implementation of genomics is commonly directed at the detection of a single position DNA variation, defined as single nucleotide polymorphisms

(SNPs) and small insert/deletions (indels). However, the interaction of two or more SNPs (named epistasis) is also a matter of study. This approach is called genome-wide association study (GWAS) [32] and it is possible thanks to the progression in sequencing technologies that made the acquisition of DNA sequence more affordable and accurate [33]. This was also possible thanks to the human genome project [34], one of the biggest research projects focused on determining the sequencing of the complete human genome for the first time. Launched in 1990 and concluded in 2003, the cost of sequencing the human genome at the end of the project was $10,000, a cost that nowadays dropped to $600 [35].

- **Epigenetics**. Epigenetics studies the alteration of gene expression or mechanisms that influence the phenotype without affecting the DNA sequence, contrary to genetics. Histone modification, DNA methylation, and chromatin remodeling are the main forms of epigenetic manifestation. DNA methylation is one of the most commonly studied mechanisms [36], involving chemical modification at the DNA level. Epigenetic changes can arise from environmental factors such as nutrition and smoke or exposure to pollutants, metals, or radiation.

- **Transcriptomics**. Transcriptomics studies are focused on the transcriptome, the whole set of RNA transcripts in a cell [37]. Besides messenger RNA (mRNA), the coding RNA that is translated into protein, transcriptomics covers also non-coding transcripts like small non-coding RNA (e.g. microRNA), and long non-coding RNA such as circular RNA (circRNA) [38]. The available methods retrieve transcriptomics data from both a qualitative and quantitative point of view, allowing the annotation of the genome by identifying thousands of transcripts, providing their location and function together with their intensity in terms of expression level. Transcriptomics can be applied for exploring splicing events, post-translational modifications, biomarker discovery, and the understanding of a disease' stage, a tissue status, or a cellular function [39].

- **Proteomics**. The whole collection of proteins present in a cell is analyzed by the field of proteomics. Proteins are complex molecules composed of a chain of amino acids bonded together and they are responsible for the functioning of cells, tissues and organs [40]. As stated in the central dogma of molecular biology, the nucleotides constituting the RNA sequence are taken in triplets (codons) and translated into a sequence of amino acids forming a polypeptide bond [41]. The folding of a polypeptide generates a protein and its structure determines the function the protein will perform. The study of proteins arise in complexity because of the tens of thousands of possible structures a polypeptide can fold into. Additional events such as alternative splicing and mRNA editing further contribute to the generation of new different proteins that make their quantification and identification challenging [42, 43]. The first choice instrument for proteome analysis is mass spectrometry (MS), an accurate technique used to identify the proteins expressed in a sample, determine the protein structure and the amino acid sequence, and identify post-translational modifications [44].

- **Metabolomics**. Metabolomics studies metabolites (e.g. vitamins, amino acids, and organic acids) and low-molecular-weight molecules for the understanding of cellular metabolic activity [45]. Metabolites are generated by the process of metabolism and they are employed in multiple activities like defense, energy production, signaling regulation, and regulation of other biomolecules [46]. Measurements of metabolites profile and abundance are often carried out by MS.

- **Microbiomics**. Microbiota is the collection of microorganisms, including bacteria, protozoa, viruses, and fungi, that live together in the organism. Microbial cells in a human are estimated to be up to 100 trillion [47]. Microbiome is the set of genes that arises from the microbiota. A microbial population can inhabit both inside and outside a living organism, such as the gut, throat, and skin. Nevertheless, the microbiome of the digestive tract is one of the most important given its direct association with the immune system and metabolism, affecting the organism's health [48]. The study of microbiome-derived diseases is one of the main use in bioinformatics. Alteration in the gut microbiota can lead to several diseases like type 2 diabetes, cancer, cardiovascular diseases, and obesity [49]. Furthermore, an association has been found with mental health diseases such as schizophrenia and depression [50, 51].

### 1.2.1 Transcriptomics

Two main techniques are used to quantify and identify transcripts at a specific moment, fluorescence hybridization-based named DNA microarray, and RNA-seq. Array-based techniques consist of a plastic or glass chip where hundreds of thousands of probes are bound [52]. Probes are typically oligonucleotides (DNA sequences) of different lengths that can go from 25 bp to 60 bp [53]. Their location on the chip is in clusters, where each cluster is a target transcript. RNA is then isolated from the study sample and exposed to the array slide to make the new RNA strands bind together with the complementary base of the probes, generating fluorescent light. Such light will be used to determine whether the gene is up-/down-regulated or unchanged and the presence of multiple probes per gene is fundamental to give consistent results. With RNA-seq it is possible to identify the nucleotide sequence of the RNA. Next-generation sequencing (NGS) platforms have been developed by multiple firms, among them, Roche, Thermo Fisher Scientific, and Illumina, with the latter one as one of the most popular [54–56]. The Illumina NGS method follows four main steps named library preparation, amplification, sequencing, and alignment [57–59]. RNA is converted into complementary DNA (cDNA) used as input in the NGS platform. This is done with the implementation of reverse transcription. cDNA is segmented and each segment is ligated to a sequencing adaptor. In the amplification step, segments are bound to chips and a complementary strand of each segment is made while the original is washed away. Segments are amplified through bridge amplification. The amplification starts folding segments over the chip to form bridges, followed by the synthesization of a complementary strand for each segment. Denaturation of the bridge will create two strands of single-stranded DNA. Bridges are cloned into millions of copies thanks to the polymerase chain reaction. The sequencing step is to determine the RNA reads, where lasers scan the chip after every amplification and a different fluorescent label is assigned to each nucleotide. The last step is alignment, where reads are mapped to a reference genome. The introduction of NGS technologies paved the way to make RNA-seq replace the DNA microarray. RNA-seq captures transcripts at higher coverage, not known *a priori* hence allowing the identification of novel transcripts.

The main steps of both technologies are illustrated in Figure 1.2.

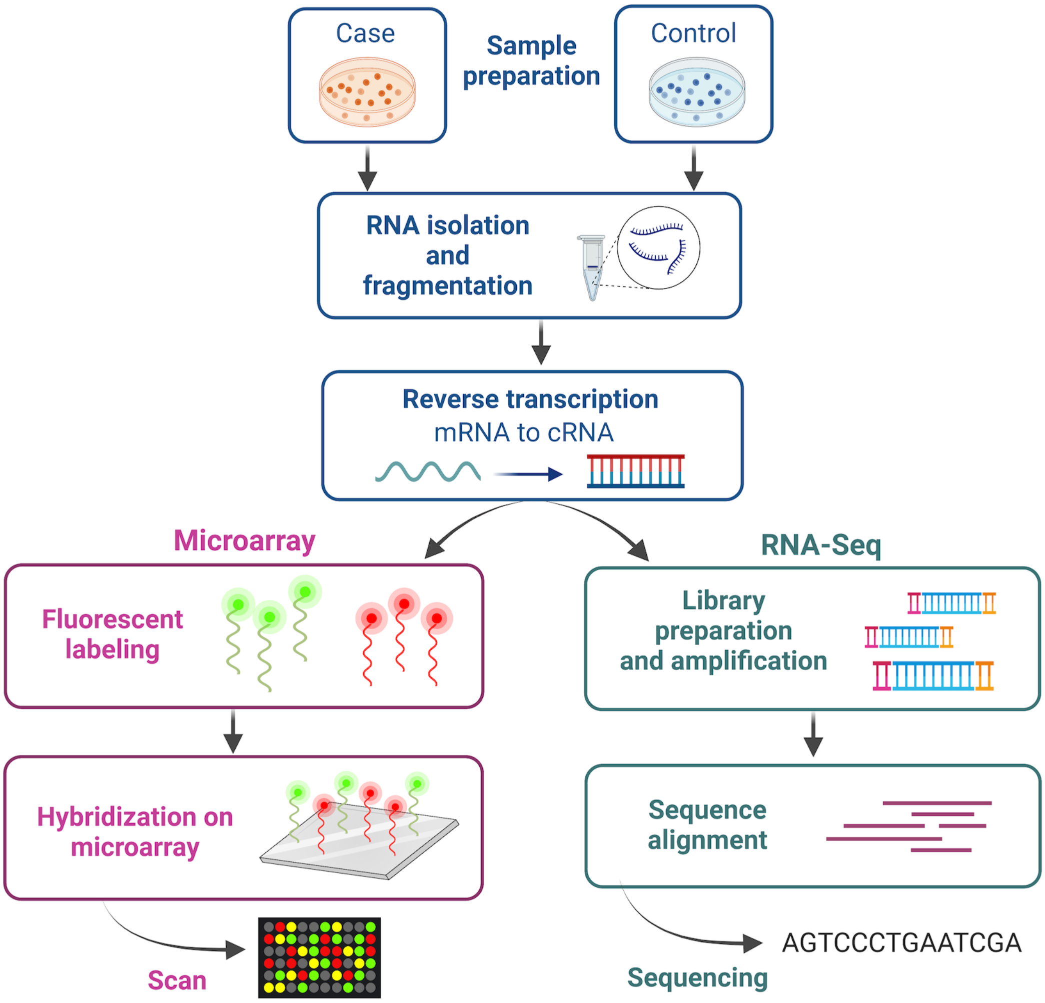


**Figure 1.2** Overview of microarray and RNA-Seq assays.

One of the major applications of sequencing and array technologies is the measurement of gene expression levels. Altered expression values of genes involved in cell cycle processes are one leading consequences of cancer. The alteration might inhibit the tumor suppressor role of a gene, turning it into a tumor promoter involved in cellular mechanisms like metabolism, cell growth, progression and multiplication, apoptosis. With the awareness that gene expression alteration might be associated with cancer initiation and development, it's important to observe the expression status of as many genes as possible to unravel not only the elements that triggered the tumor start but also the ones that enhance its proliferating activity.

### 1.2.2 Epigenetics

DNA methylation consists in a chemical modification that arises when the DNA methyltransferase (DNMT) transfers a methyl group from the S-adenyl methionine to the fifth carbon position of the a cytosine, forming a 5-methylcytosine. In mammals, 98% of DNA methylation occurs when, on the strand in direction 5′ to 3′, the nucleotides cytosine precede guanine and there is a phosphate group between the two, context that takes the name of CpG site [60, 61] (Figure 1.3 A). CpGs tend to group in dense clusters named CpG islands and are usually unmethylated and located in gene promoters. However, aberrant methylation of CpG islands is known to be responsible for regulating gene expression activity because of transcription prevention or promotion [60, 62, 63]. Hypermethylation of CpG islands in gene promoters prevents the binding of the transcription factor resulting in epigenetic gene silencing, leading to cancer initiation and therapy resistance in case the repressed gene is a tumor suppressor. Hypomethylated CpG islands situated in the promoter region are associated with gene activation and overexpression (Figure 1.3 B).

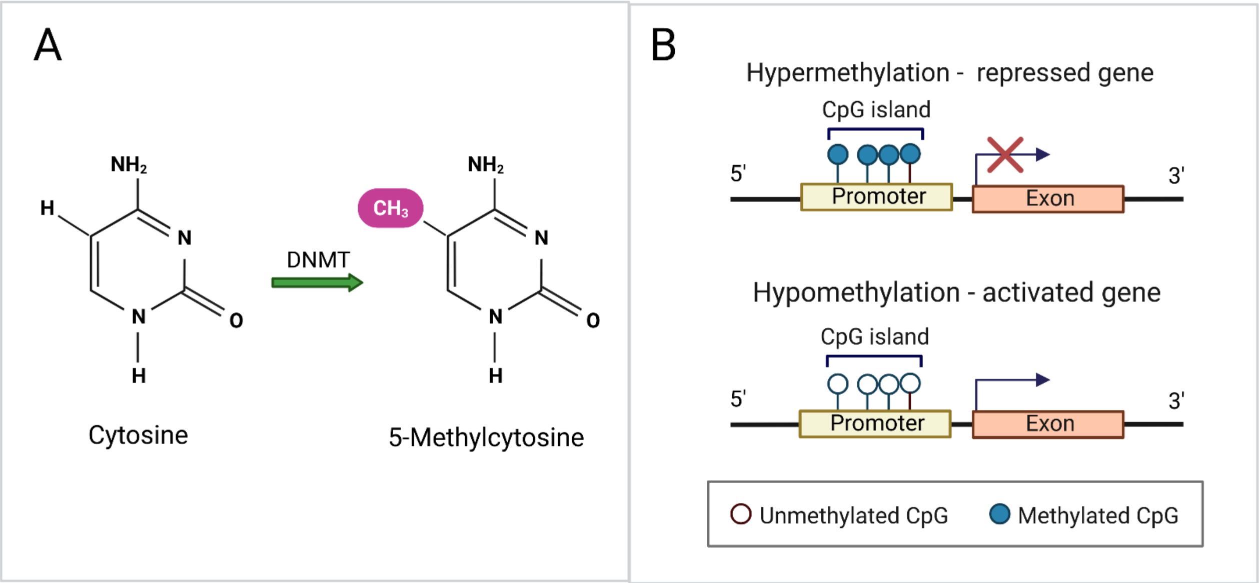


**Figure 1.3** A) Creation of a 5-Methylcytosine by the transfer of a methyl group to the fifth position of a cytosine. B) Hypermethylated and hypomethylated CpG islands in promoter regions are responsible for gene silencing and activation.

The DNA methylation's regulatory role can also be observed outside the promoter, for instance at the enhancer and silencer level or gene-body. Methylated enhancers have been linked to immune infiltration in lung cancer [64]. With the estimation of immune cell types, it has been shown the correlation between immune infiltration abundance and DNA methylation at the enhancer level, indicating that methylated enhancers might influence cancer-infiltrating immune cells. DNA methylation in the gene body has been found associated with an increase in gene expression [65]. It has been discovered that DNA methylation has a significant link to carcinogenesis and its role in disease evolution and growth has been widely investigated. [66–69]. Colorectal cancer is characterized by altered methylation activity combined with genetic modifications such as TP53, APC, BRAF, and KRAS. Colorectal cancer is a highly age-related disease and the presence of hypermethylated CpG islands is observed in aging patients. Indeed, colorectal mucosa has been linked to altered methylation in the promoter regions of the genes, which might be responsible for the neoplasm evolution [70–72]. The impact DNA methylation has on numerous diseases makes it the perfect candidate for biomarker discovery.

Measurement of DNA methylation levels is mostly carried out following two technologies, one based on genome sequencing, named whole genome bisulfite sequencing (WGBS), and one based on array technique. The main difference between the two methods is the depth of the information they can retrieve. Indeed, WGBS can measure  95% of the CpGs present in the whole human genome while the most recent array-based method, MethylationEPIC Array released by Illumina, can measure approximately 850,000 CpGs out of  28 million, hence  30% of the total CpGs in the genome [73, 74]. The proportion of CpGs distributed across gene regions and CpG context is also different between WGBS and EPIC, with WGBS showing a higher percentage of CpG islands covered ( 50% vs  18% of EPIC), CpGs in 5′UTR ( 17% vs  13% of EPIC), CpGs in the promoter ( 13% vs  7% of EPIC) and CpGs in FANTOM5 enhancer ( 6% vs  3% for EPIC) [75–77]. On the other hand, other gene regions are covered with proportional percentages. For its elevated in-depth analysis, WGBS is the preferred method and the advent of NGS allows DNA sample preparation and successive genome sequencing to be more affordable. However, array techniques are still the preferred choice considering the efficiency in terms of time and costs [78].

Popular measurement of methylation intensity at the CpG level, is given by a continuous value ranging between 0 and 1, also known as $\beta$ value. $\beta$ value is the ratio of methylated signal intensities (M) and the total intensities, which is the sum of both methylated (M) and unmethylated (U) signal intensities. Hence, $\beta$ can be defined as: $\beta = \frac{M}{M+U}$. The 0 to 1 interval allows easy interpretation of the methylation level of an individual CpG, where the extremes correspond to unmethylated CpG, in case of $\beta = 0$ or fully methylated CpG when $\beta = 1$.

## 1.3 The importance of biomarkers

Cancer initiation, development and spread to metastasis can be driven by a wide variety factors such as inherited genetic alteration, lifestyle, the environment the person is exposed to or a combination of them. The understanding of cancer increases in complexity when investigating the molecular pattern that triggers the malignant cell transformation and progression. The advent of NGS technologies changed the analysis of molecular cancer scenarios. Thanks to the high genomic coverage, NGS made large-scale data available at great detail. This will soon allow researchers to replace the one-fits-all solutions with approaches personalized to each patient. Indeed, if the discovery of cancer therapies and molecular drivers has always been performed for universal application, the last decade observed a substantial change on how tumor patients are studied and treated. The final aim is to elucidate genomic alteration to better delineate tumor progression and address a therapy which matches the patient tumor's molecular characteristics [79]. This leads to precision medicine, an innovative and promising holistic approach which takes into consideration each patient's information individually. To facilitate precision medicine, it is important to be aware of the medical state of the patient to make the most effective pharmacological choice and prolong the patient's survival expectation. The knowledge of the biological processes standing behind a pathogenic outcome is crucial to predict the drug response and thus to design the best intervention.

A biomarker, short for biological marker, is an indicator for such status, which the National Cancer Institute [80] defines as "a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease [...]". The role of biomarkers is diverse and is subject to different measurements, applicability and interpretations. Three main categories distinguish biomarkers: *predictive*, *prognostic* and *diagnostic* [81]. *Predictive* biomarkers are used to predict the therapeutic response hence to assess whether a patient can benefit from a specific medication. These biomarkers compare therapy to placebo or to other therapies to quantify the level of benefit from exposing a patient to that therapy. *Prognostic* biomarkers are indicators of the probability of an event such as death or disease recurrence. Hence, this kind of biomarker is commonly used to predict a patient's prognosis. *Diagnostic* biomarkers are used for finding out the presence of a disease in a patient or for delineating its subtype. It can be additionally used to confirm the presence of a disease. The advancement in high-throughput technologies facilitated biomarker discovery and their evaluation, with the final goal of clinical application. Biomarkers constitute a powerful tool for assisting doctors during the process of decision-making and play a key role in the development of personalized treatments.

### 1.3.1 Computational biomarker discovery

Identification of biomarkers can come from any measurable biological data such as exome and whole genome sequencing, expression profiling, proteomics and DNA methylation. Hence biomarkers can be SNPs, copy number variations (CNVs), genes or methylated cytosines. The computational biomarker discovery can be carried out following a sequence of steps, summarized in Figure 1.4.

Biomarkers can derive from multiple molecular data, taken individually or integrated. Besides molecular profiles, a phenotype of interest can be known *a priori*, like the knowledge of the cancer subtype labels for each sample. To ensure reliable results, molecular data is then cleaned and corrected. Quality control is mandatory to remove low-quality samples and features such as genes with low sequencing depth. Data can be subject to variability due to sample preparation, origin, transportation, and technical settings. Such unwanted variation takes the name of batch effect and it is important to correct the data to remove the technical variation and leave only the biological variation of interest. Omics data are large-scale data with thousands or even millions of features, and it is common practice to reduce the data dimension by removing less informative features with low variance or correlated features responsible for data redundancy. These are just a few of the data preparation steps. For increasing the model accuracy and the interpretability of the results, data can be further adjusted by removing the effect of confounding factors such as tissue heterogeneity, by estimating the sample's cell type composition, or other covariates like demographic or clinical factors. The choice of the model depends on the data characteristics, size, knowledge of the variable of interest. This last variable is what distinguish supervised from unsupervised models. Unsupervised can help find hidden molecular patterns that correspond to distinct cancer subtypes. The features that produce relevant clusters are further investigated as they might be potential biomarkers. On the other hand, a supervised approach can be employed to investigate the relationship between a phenotype and a genotype. Example of a supervised case is the training of a classification algorithm on Immunohistochemistry (IHC) obtained labels and predicting the labels on a different set of samples, a test data. These features most related to the variable of interest constitute potential biomarkers. Such potential biomarkers have to go through deep analytical validation. Measures like accuracy and false discovery rate are used to estimate the predictive performance of a set of features used as classifiers. It has

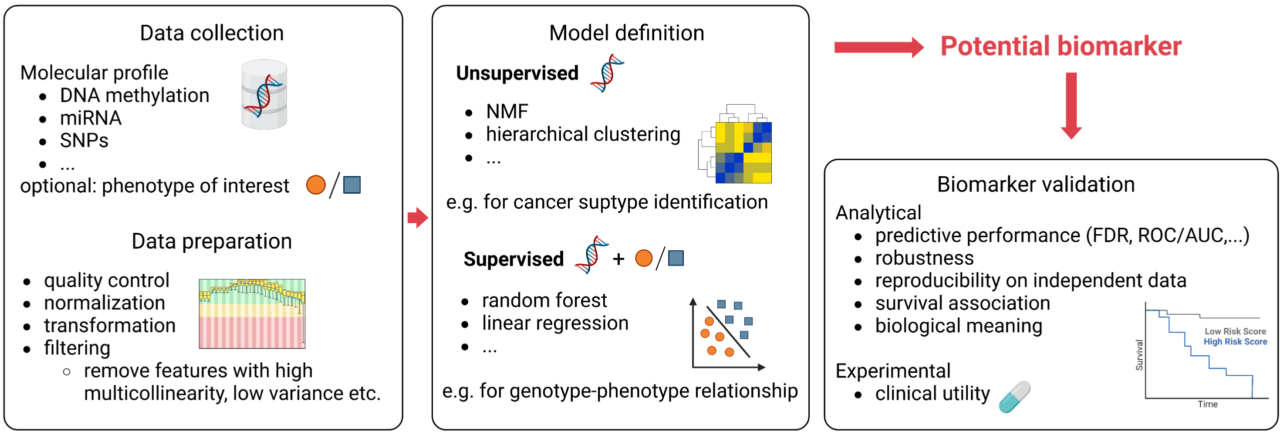


**Figure 1.4** Overview of the computational workflow for biomarker discovery. The first step is collecting molecular data and phenotype information, when available. Data is then subject to pre-processing where steps like quality control, normalization and features filtering are applied. Once the data are ready, the next step is building the model that will be used for the discovery of potential biomarkers. According to the presence or not of phenotypic data, the model will be supervised or unsupervised. A potential biomarker will be validated analytically and/or experimentally, with the design of a clinical study for its evaluation.

been shown how random features reach a predictive accuracy comparable to the one of established biomarkers [82]. This finding makes the comparison between potential biomarkers and random features a fundamental validation step that is often ignored. A biomarker to be applicable has to be generalizable and consistent, reaching comparable performance on independent data. In the case of prognostic biomarkers, the prognostic relevance is assessed by estimating the odds of survival with the gold standard method Kaplan-Meier estimator [83] or Cox regression [84]. To be used in clinical practice, biomarkers have to be validated through clinical trials to prove their validity. A biological experiment is designed and used to test whether the biomarker affects positively the patient's outcome or therapeutic response.

### 1.3.2 Gene signatures for cancer subtyping

Cancer is not considered as one single disease and even the same cancer type can be heterogeneous showing distinct morphologic and phenotypic characteristics. Understanding their differences is crucial for guiding treatment strategies and predicting the disease outcome. The study of the tumoral properties shows differentiation in the cellular traits and genomic landscape, suggesting the existence of subgroups within each class of tumor type. Identifying the correct subtype a sample belongs to, is fundamental for designing personalized care. However, cancer stratification is one of the biggest research challenges.

Researchers were able to develop computational tools for patient stratification in subtypes for numerous cancer types such as breast cancer [85], colorectal cancer [86, 87] and PDAC [88–91]. The classification in subtypes of breast cancer is the most straightforward due to the divergent plasticity of the subtypes [85, 92, 93]. Indeed, breast cancer subtypes can be divided into three main classes with distinct prognoses and treatment responsiveness while, for several cancer types, there is still inconsistency in determining both the number of classes and their characteristics.

Histology is the standard method for tumor subtype classification. In particular, immunohistochemistry (IHC) uses antibodies to bind antigens and detect specific proteins in a portion of tissue previously formalin-fixed and paraffin-embedded [94]. However, the use of IHC is extremely dependent on the pathologist who performs the analysis. Not only the interpretation of the results, but components such as reagents for antigen retrieval, duration of tissue fixation, and fixation agents are only a few of the choices a pathologist has to make, making IHC a non-standardized technique subject to noise. For these reasons, the classification in subtypes based on molecular data would advance precision oncology and contribute to a tumor stratification systematic and free of biases [95]. Intrinsic molecular cancer subtypes are subclasses of cancer identified thanks to the combination of molecular data with machine learning techniques. This new approach which started being adopted in recent years, is a systematic alternative to histology meant to speed up the classification, increase the accuracy and link molecular knowledge to the subtypes. Molecular subtypes are commonly based on transcriptome and identified based on gene expression values where unsupervised or supervised machine learning methods are applied, depending on the availability of pre-assigned histological labels. When a set of genes share the same expression pattern and relationship with a disease or condition it takes the name of gene signature [96, 97]. A gene signature is a potential composite biomarker and one of the diagnostic uses is the tumor subtype classification. The advancement in computational analysis along with large-scale omics data constitutes a powerful tool for making cancer sample classification fast, automated and bias-free. This is possible with the identification of signatures whose altered expression is linked to a specific phenotype, which constitute a potential cancer subtype classifiers.

### 1.3.3 Current challenges

With the implementation of computational techniques, bioinformaticians address research questions aimed at unraveling unknown mechanisms and contributing to the understanding of systems biology. Despite the great effort in developing new technologies and algorithms to assist biomarker discovery, there are several major challenges that researchers have to face [98]. In section 1.3.1 we review the principal steps implemented for the identification of novel biomarkers from a computational point of view, briefly mentioning limitations and obstacles. In this section, we dive deeper into the subject and cover the major challenges when performing biomarker discovery.

#### Data size and preprocessing

New technology platforms made molecular data available at a great detail. Nevertheless, dealing with big dimensional data is one of the major challenges in computational biomarker discovery. Molecular data are often characterized by the measurement of thousands or millions of features in a much smaller amount of samples. This data configuration, i.e. when the feature size is considerably bigger than the sample size, can lead to overfitting [99]. Overfitting occurs when the model is built to catch exactly the characteristics and minor details of the data used for training, resulting in the inapplicability on external data. Moreover, the use of a small sample size might lead to false discoveries and misleading conclusions due to the underrepresentation of a more realistic population of interest. If the goal is to find biomarkers to distinguish cancer subtypes, we need to acquire enough samples to be able to capture enough patient-to-patient variability for a robust distinction. With fewer samples, the risk is to only resemble a subpopulation or only one of the subtypes and miss out on relevant heterogeneity needed to extend the results to a broader real-world cohort. Nevertheless, a large data structure is not only complex to handle but also challenging to interpret, especially in the case of a multi-omics approach. There are two main approaches for studying multiple omics data [100], where the first one is targeted to specific molecular mechanisms or molecules where we know a priori the pathways to look at. The analysis proceeds by assessing the behavior and activity of the molecules involved, from a multi-omics point of view. A second approach is oriented to the discovery of novel markers thanks to the analysis of the omics data first individually and then in combination to identify overlap and meaningful correlations. Such an approach is very advantageous to gain novel insight yet complex given the extensive amount of data available, making the analysis computationally challenging and the results of hard interpretation. Indeed, the integration and interpretation of distinct levels of molecular data is not always straightforward [101]. Another important aspect to consider is the eventuality of technical and biological biases in the data introduced by sample handling or data generation, resulting in noisy and misleading signals. Preprocessing and data cleaning are crucial steps for removing data biases yet maintaining the wanted heterogeneity. Filtering out low variable, co-correlated features or outliers are some of the standard preprocessing approaches, followed by data standardization or normalization [102].

### Resources

Another important aspect is the limited resources available to researchers. Resources can be of different kinds, for instance, money budget, computational resources, or clinical and demographic data as support to the molecular data. A limited budget obstacles researchers to acquire the necessary amount of data and it limits the purchase and use of particular platforms and technologies. It is additionally involved in providing the required computational resources able to sustain the analysis of big data and computationally heavy algorithms. Nevertheless, many computations are challenging despite the resources because of the extremely high feature size [103]. Moreover, samples used for the analysis are often lacking additional useful information, such as demographics and clinical/medical data of the patients (e.g. age, medical history, previous treatments, eventual cancer relapse or refractory).

### Model

The model adopted for discovery is of fundamental importance and its design can be used to control for overfitting and prevent false discoveries [104]. Selecting the right model requires the understanding of the data and the desired output, however, it might provide an additional level of bias especially if it requires the selection of multiple hyperparameters [105]. Additionally, numerous novel sophisticated algorithms might be hard to implement or to interpret the outcome's biological relevance.

### Results validation

A meticulous data and model selection that led to a potential biomarker needs to go through one last phase which is the validation of the results [104]. This step involves the evaluation of the accuracy of the model as well as of the results. The major limitations of novel potential biomarkers are a lack of generalizability, reproducibility, and robustness. Biomarkers are generalizable when their discovery is not limited to representing a subpopulation but it catches variabilities that can resemble a larger population real-world-like. Another fundamental aspect is the reproducibility of the results.

One common issue in biomarker discovery is that, testing the potential biomarker on an independent cohort, it does not reproduce the outcome it is supposed to return. This is a common case of overfitting, when the model performs too well on the discovery cohort but it does not reproduce the same result on external data. Hence, testing the potential biomarkers on an independent dataset helps in understanding the biomarker applicability. One additional point is assessing whether the biomarkers obtained are robust enough, meaning that they carry true biological meaning, are reliable, and their discovery is not due to chance but they are real indicators of a phenotype or a condition. When we test for robustness we want to compare our potential biomarkers with random ones, to assess which is the chance of obtaining the same outcome when using random features [102]. This step is often omitted and, without it, we might ignore cases of false discoveries. One example is the study published by Venet and colleagues [106], where it was shown how most published breast cancer gene signatures lack of robustness and their predictive ability is comparable to random genes. Hence, we conclude emphasizing the results validation as a crucial step in the computational workflow, to ensure reliability and biologically significant results.

## 1.4 Pancreatic Ductal Adenocarcinoma (PDAC)

Pancreatic Ductal Adenocarcinoma (PDAC) is the most common type of pancreatic cancer and more than 90% of the neoplasm of the pancreas is PDAC [107]. Besides being the most prevalent pancreatic cancer type, PDAC also registers the highest aggressiveness among pancreatic malignancies with a five-year survival rate of less than 10% [108]. This is due to PDAC already being at an advanced stage at the moment of its detection. Indeed, the disease progresses slowly without visible symptoms, resulting in a delayed diagnosis that makes treatment challenging. Prediction of PDAC patients' prognosis and therapy performance depends on the stage of the disease when it is found. First-line treatment is pancreatectomy, the surgical resection of the tumor which guarantees the best outcome and combined with chemotherapy reaches a five-year survival rate of over 30% [109]. However, mainly because of the lack of symptoms, patients develop distant metastases and the majority die within 2 years from the diagnosis, with chances of surviving the first year after diagnosis lower than 20% [107].

PDAC is not longer studied as one disease but is instead analyzed considering the heterogeneity across different individuals. Different histological phenotypes in PDAC are due to different genetic modifications and cellular compositions, with an impact on the survival and metastases spread. Continuing from a histological point of view, PDAC is characterized by the presence of dense fibrotic stroma which surrounds the pancreatic glands [110]. The cellular landscape in PDAC displays not only intertumor but more importantly also intratumor heterogeneity. Indeed, the same neoplasm observed at different stages is characterized by a progressive change in the cellular landscape, at a histological level. A distinctive feature of PDAC is the substantial desmoplastic response which changes and develops together with the tumor, adapting its components to the TME and creating a stromal barrier around the tumor, preventing immune cells from penetrating, promoting tumorigenesis and enhancing therapy resistance [111]. TME of the PDAC is composed of extracellular matrix (ECM) proteins, immune cells, pancreatic stellate cells, and fibroblasts [111], as illustrated in Figure 1.5.

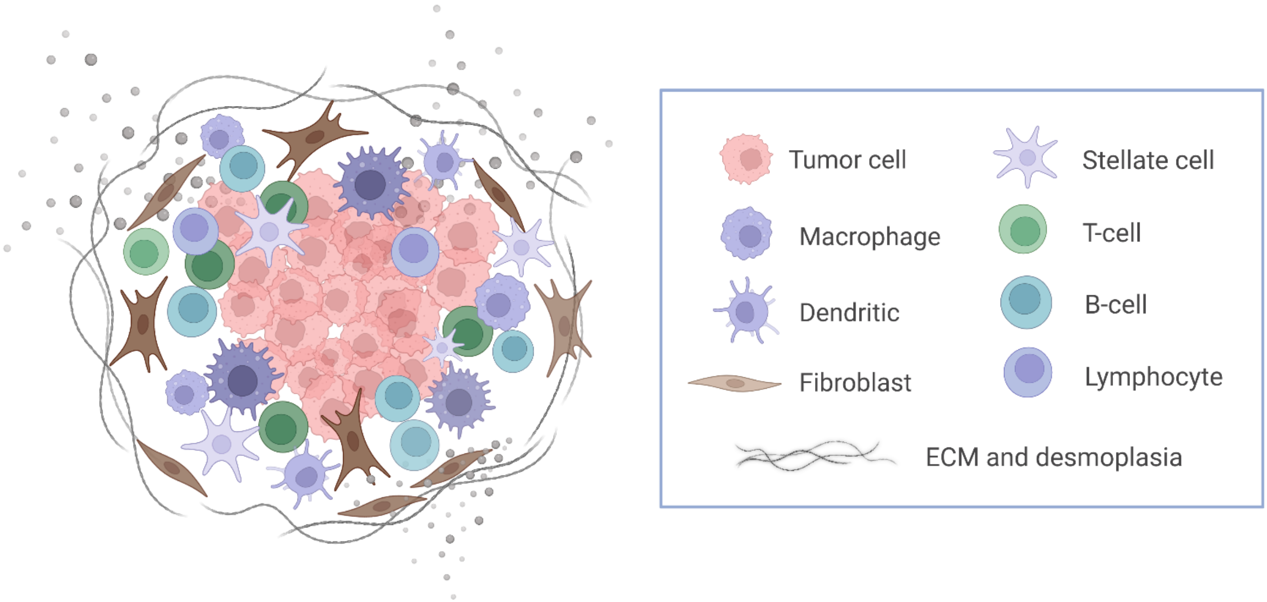


**Figure 1.5** Simplified overview on the elements in the PDAC microenvironment, made of tumor cells, stroma, immune cells, and ECM production together with the desmoplastic response.

PDAC arises from the epithelial cells of the pancreatic ducts, driven by the combination of mutational changes and then supported at a cellular level by the TME complexity. Four genes are the most commonly altered in PDAC, the oncogene KRAS and the three tumor suppressor genes CDKN2A, SMAD4, and TP53 [112]. KRAS as the most frequent (found mutated in more than 95% of the PDAC cases) is the early genetic alteration, contrary to TP53 and SMAD4 observed altered in later stages thus linked to tumor progression.

A deep understanding of the immune and stromal compartment together with the TME is crucial to help solve this oncological challenge.

### 1.4.1 PDAC published signatures

One of the goals of PDAC analysis is the arrangement of samples in subgroups based on their molecular pattern. Several transcriptomic-based classification schemes have been proposed where samples are mainly distinguishable for their morphology and prognosis [88–91]. However, still no clinical decisions have been made. Numerous studies contributed to finding the biological differences between PDAC samples. Some studies gained more attention than others from the related research field and are considered a gold standard in the literature. These three studies are Collisson *et al.* [89] in 2011, followed by Moffitt *et al.* [88] in 2015 and Bailey *et al.* [90] in 2016. A more recent study published in

2018 by Puleo *et al.* [91] also proposes promising molecular-based PDAC subtypes. We will refer to the studies by the name of their first author.

The four studies identified PDAC subtypes, different in number but with similar characteristics in the composition. Collisson discovered three subtypes defined as Classical, Quasi-Mesenchymal (QM-PDA), and Exocrine-like. Moffitt established that PDAC can be divided into two main categories with opposite morphological and prognostic characteristics: Classical and Basal subtypes. Bailey proposed four subtypes: Squamous, Immunogenic, Pancreatic Progenitor, and Aberrantly Differentiated Exocrine (ADEX) while Puleo identified five subtypes named Desmoplastic, Pure Classical, Stroma Activated, Immune classical and Pure Basal-like. The subtypes discrimination has been linked to a list of genes whose expression status reflects the biological alteration responsible for tumor diversification.

Collisson employs microarray expression data from two groups of PDAC samples, one taken considering the tumor in its whole and the other one previously submitted to microdissection, where the epithelium was devoid of the stroma. After combining the two datasets, the authors identified three clusters with the use of consensus clustering-based NMF. A multi-class SAM analysis identified a panel of 62 genes to be used for classifying patients to their subtypes. The three subtypes proposed have been then validated via clustering analysis to test their reproducibility on independent datasets. One dataset originated from both human and mouse cell lines did not return any Exocrine-like subtype. The other three datasets from human samples were first merged into the dataset used for the subtype identification and then clustered using the signature genes. The clusters obtained matched with the three subtypes proposed by Collisson and colleagues. However, the same result was not observed when the clustering was applied on one of the validation datasets taken independently, without prior merging it with the discovery data. Additionally, a validation using 19 *in vitro* human PDAC cell lines to test the impact of two drug therapies was carried out. The response observed was different, with Erlotinib more beneficial for the Classical type contrary to Gemcitabine with a positive outcome on the QM-PDA subtype.

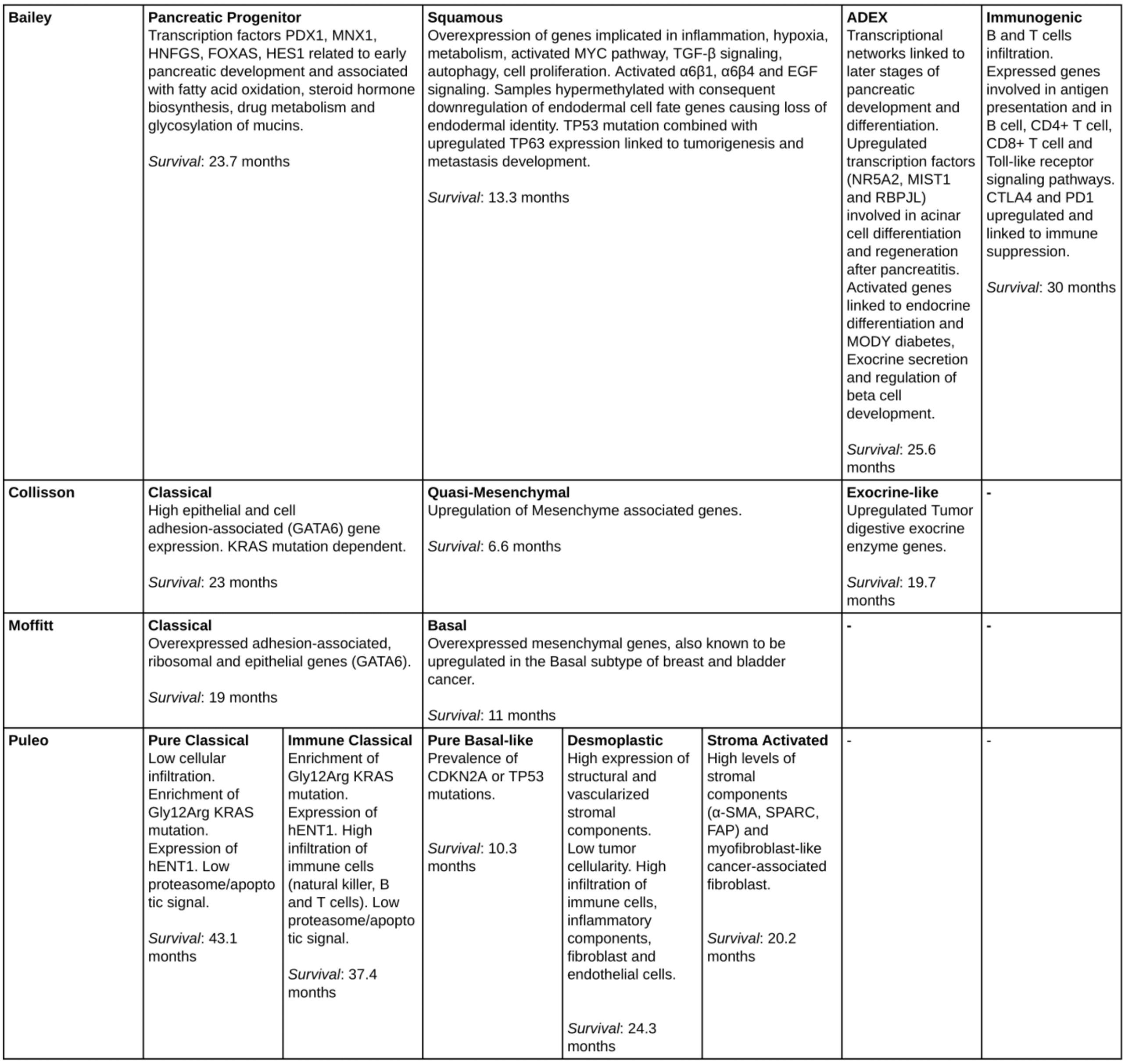
Moffitt combines multiple microarray datasets obtained from diverse sources and tissue types (primary and metastatic tumors, cell lines, normal pancreas and distant sites adjacent to normal samples). The authors decomposed the data into 14 biological components with the use of NMF. The components are used for performing virtual microdissection and isolating those categories of components linked to the stroma compartment and tumor tissue. The stroma-related genes and tumor-related genes, 48 and 50 respectively, were identified as representative of the sample's diversity and used to establish two distinct binary subtyping schemes. Samples are classified into Normal Stroma or Activated Stroma subtype, in the first case, and Classical or Basal, in the second classification schema. Validation of the subtypes was carried out by implementing clustering analysis on samples of primary PDAC, cell lines, patient-derived xenografts, and cancer-associated fibroblasts. The clusters obtained did not confirm the classification in Basal and Classical in the case of the cell lines, contrary to the use of patient-derived xenografts which confirmed the binary classification. The two tumor subtypes are linked to opposite prognoses confirmed via survival analysis, where Basal is associated with a worse outcome hence a more aggressive subtype, opposite of Classical which showed a better prognosis.

Differently from the previous two studies which used microarray data, Bailey used RNA-Seq gene expression data derived from sequencing samples of ductal histopathological subtypes as well as a few acinar cell carcinoma samples and intraductal papillary mucinous neoplasm with invasion. The technique used for grouping the samples was the same adopted by Collisson and Moffitt, this time without previous microdissection (neither via laser nor virtually) but filtering the most informative samples by keeping only the ones that registered tumor cellularity above 40%. Clustering the data via consensus clustering-based NMF returned the samples divided into four subtypes, where the optimal number was established with a cophenetic coefficient. The four subtypes proposed are different in prognosis and tumor cellular composition. The subtypes are called Squamous, Immunogenic, Pancreatic Progenitor and Aberrantly Differentiated Exocrine (ADEX) and their genomic hallmarks have been further investigated by Bailey and colleagues with a whole-genome and deep-exome sequencing analysis. 613 genes were found significantly differentially expressed between the four subtypes, identified via multi-class SAM. The reproducibility of the four classes on independent data was assessed employing a microarray dataset of human PDAC samples, this time not limited to only samples rich in tumor cellularity. Performing the same clustering analysis returned the same subtypes, further supporting their existence.

A collection of formalin-fixed paraffin-embedded samples from resected primary tumors was used by Puleo. Gene expression was measured with microarray and hierarchical consensus clustering was used to establish the subgroups of PDAC. Puleo and colleagues identified five subtypes named Pure Classical, Immune Classical, Desmoplastic, Stroma Activated and Pure Basal-like. The subtypes are named after measuring the stromal compartment together with the infiltrating immune cells and the TME, assessed by deconvoluting the data into components and using transcriptomic-based techniques. Subtypes rich in stroma are Activated Stroma and Desmoplastic, where the first one is characterized by activated stromal elements while the second one is composed of fibroblast and endothelial cells with a high immune infiltration. Another subtype rich in immune cells is the Immune Classical, contrary to Pure Classical and Pure Basal-like which were found poor in the immune response. In terms of tumor differentiation, Pure Classical was found well differentiated, suggesting lower aggressiveness and slower progression, while Basal-like was found poorly differentiated. Puleo additionally investigated the survival associated with each subtype with a survival analysis which sees the survival curve with the poorest outcome linked to the Pure Basal-like. Activated Stroma, another subtype with little immune cell compartment, was found associated with poor survival. The binary classification of Moffitt is taken as a comparison to highlight the existence of two main PDAC categories which can be further divided into more detailed subcategories. Furthermore, Puleo and colleagues mentioned the potential existence of subtypes which includes both classical and basal characteristics. Finally, a list of 403 signature genes to use as classifiers is proposed.

An overview of the aforementioned studies and the distinctive features of the subtypes is provided in Figure 1.6.

**Figure 1.6** Overview on the PDAC molecular subtypes of Bailey *et al.*, Collisson *et al.*, Moffitt *et al.* and Puleo *et al.*. Table from [113].



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Bailey** | **Pancreatic Progenitor**<br>Transcription factors PDX1, MNX1, HNFGS, FOXAS, HES1 related to early pancreatic development and associated with fatty acid oxidation, steroid hormone biosynthesis, drug metabolism and glycosylation of mucins.<br>*Survival*: 23.7 months | | **Squamous**<br>Overexpression of genes implicated in inflammation, hypoxia, metabolism, activated MYC pathway, TGF-β signaling, autophagy, cell proliferation. Activated α6β1, α6β4 and EGF signaling. Samples hypermethylated with consequent downregulation of endodermal cell fate genes causing loss of endodermal identity. TP53 mutation combined with upregulated TP63 expression linked to tumorigenesis and metastasis development.<br>*Survival*: 13.3 months | | | **ADEX**<br>Transcriptional networks linked to later stages of pancreatic development and differentiation. Upregulated transcription factors (NR5A2, MIST1 and RBPJL) involved in acinar cell differentiation and regeneration after pancreatitis. Activated genes linked to endocrine differentiation and MODY diabetes, Exocrine secretion and regulation of beta cell development.<br>*Survival*: 25.6 months | **Immunogenic**<br>B and T cells infiltration. Expressed genes involved in antigen presentation and in B cell, CD4+ T cell, CD8+ T cell and Toll-like receptor signaling pathways. CTLA4 and PD1 upregulated and linked to immune suppression.<br>*Survival*: 30 months |
| **Collisson** | **Classical**<br>High epithelial and cell adhesion-associated (GATA6) gene expression. KRAS mutation dependent.<br>*Survival*: 23 months | | **Quasi-Mesenchymal**<br>Upregulation of Mesenchyme associated genes.<br>*Survival*: 6.6 months | | | **Exocrine-like**<br>Upregulated Tumor digestive exocrine enzyme genes.<br>*Survival*: 19.7 months | **-** |
| **Moffitt** | **Classical**<br>Overexpressed adhesion-associated, ribosomal and epithelial genes (GATA6).<br>*Survival*: 19 months | | **Basal**<br>Overexpressed mesenchymal genes, also known to be upregulated in the Basal subtype of breast and bladder cancer.<br>*Survival*: 11 months | | | **-** | **-** |
| **Puleo** | **Pure Classical**<br>Low cellular infiltration. Enrichment of Gly12Arg KRAS mutation. Expression of hENT1. Low proteasome/apoptotic signal.<br>*Survival*: 43.1 months | **Immune Classical**<br>Enrichment of Gly12Arg KRAS mutation. Expression of hENT1. High infiltration of immune cells (natural killer, B and T cells). Low proteasome/apoptotic signal.<br>*Survival*: 37.4 months | **Pure Basal-like**<br>Prevalence of CDKN2A or TP53 mutations.<br>*Survival*: 10.3 months | **Desmoplastic**<br>High expression of structural and vascularized stromal components. Low tumor cellularity. High infiltration of immune cells, inflammatory components, fibroblast and endothelial cells.<br>*Survival*: 24.3 months | **Stroma Activated**<br>High levels of stromal components (α-SMA, SPARC, FAP) and myofibroblast-like cancer-associated fibroblast.<br>*Survival*: 20.2 months | - | - |

## 1.4.2 Limitations

The extensive analysis carried out by the four studies has been often taken as a reference whenever PDAC subtyping was discussed. However, despite their contribution to better understanding how PDAC can be classified, numerous studies found inconsistencies and expressed hesitancy [114–119].

The main critics are linked to the real existence of one or more of the proposed subtypes, suspecting that their discovery is influenced by sample preparation.

One evaluation was performed by Janky *et al.* [117] who predicted Collisson classes in a whole tumor samples cohort, using the 62 gene signature. Despite the clusters matching with the labels from the existing classes, the authors observed the survival curves linked to the obtained clusters and found a discrepancy with the prognosis described by the original study. In particular, Exocrine-like was described as a less aggressive subtype linked to good prognosis, while Janky *et al.* linked Exocrine-like to a poor outcome. Rashid *et al.* [116] assessed the clustering robustness of the genes proposed by Collisson, Moffitt and Bailey by using them to cluster nine independent datasets. Survival analysis was additionally performed to assess the prognostic relevance and reproducibility of the subtypes. The binary classification proposed by Moffitt with their tumor-specific subtypes is the one found consistent in identifying patient groups whose survival difference was significant and coherent with what was stated by Moffitt and colleagues. The prognostic power of the Moffitt signature was further validated by Birnbaum *et al.* [114], which employed 15 cohorts of primary tumor samples. The assessment of subtype's prognosis was validated for Moffitt and Bailey signatures but not for the Collisson one.

Besides observing whether the signatures are prognostically relevant or not, other studies focused on investigating the tumor purity and composition as well as the biological relevance of the subtypes proposed. The Cancer Genome Atlas (TCGA) Research Network [115] used their dataset of primary PDAC samples and selected, in turn, one of the signatures. The application of clustering on the reduced dataset returned clusters successively analyzed in their tumor purity to elucidate their biological meaning. Subtypes found low in tumor purity were the Exocrine-like and Immunogenic from Collisson and Bailey, respectively. The presence of normal pancreatic tissue in the samples might be the reason for such poor purity. Skepticism on the Bailey Immunogenic subtype was also raised by Maurer *et al.* [118] who hypothesized the possible influence of a rich stroma compartment in the Immunogenic-predicted samples. Sivakumar *et al.* [119] performed a functional pathway analysis finding the Immunogenic subtype involved in the cell cycle signaling pathway. Despite this promising finding, further investigation on the composition of the sample should be performed to clear the doubt on the possible high Immunogenicity of the samples used as bias sources.

The Immunogenic subtype is not the only subject of criticism. As already mentioned by TCGA, Exocrine-like existence has been in doubt from other studies and even from Collisson himself, who analyzed the subtypes *in vitro* and found no presence of this subtype, suggesting contamination from exocrine cells of normal tissue. The same doubt was expressed for the Bailey ADEX subtype, possibly derived by the presence of acinar cells in the samples employed for its discovery [91, 114, 115]. Same hypothesis was sustained by Puleo who found the ADEX subtypes strongly expressed with genes of acinar cells after computing cell types estimation of resected PDAC samples. Puleo further observed no significant prognostic results associated with ADEX when performing survival analysis. The subtypes proposed by Collisson, Moffitt and Bailey, except the one offered by Puleo given its more recent release, have been extensively assessed and used in literature. These studies estimated and accounted for factors such as immune infiltration, tumor composition, stroma compartment and survival association with their subtypes. However, these studies have been carried out independently and the subtypes show a lack of agreement. Not coordinated studies not only established barely intersecting gene panels used for subtypes classification (Supplementary Figure 5.1), but even the subtypes released have characteristics that hardly overlap. Moreover, the proposed subtypes lack harmonization in their terminology and number, indicating how the real number of PDAC subgroups is still far from being elucidated. As an example, we can see the classical-like subtypes in all four studies (Classical (Moffitt), Classical PDA (Collisson), Pure Classical (Puleo) and Pancreatic Progenitor (Bailey)), while other important tissue compounds such as stroma and immunogenicity have been taken into account by few studies (Puleo, Moffitt for the stroma compound and Bailey for the immune one) while ignored by the others.

Studies started to advance the hypothesis of the existence of subtypes with common characteristics, excluding the rigorous mutually exclusive subtype schema. Taking the example of binary classification with the concept of two antagonistic classes with divergent traits, several studies observed characteristics that belong to both the subtypes [115, 120–122].

Chan-Seng-Yue *et al.* [122] used single-cell RNA-Seq to investigate the issue derived from classifying bulk tumor samples. They started analyzing bulk tumor samples to assign to each of them a label from Moffitt, Collisson, Bailey and Puleo classification schemes. The authors faced the issue of predicting the subtype of a group of samples that would not coincide with any of the existing subtypes. After analyzing data at a single-cell level for each sample, the authors found in 13 out of 15 samples both Basal and Classical cells, results that with bulk data would have not been possible to reveal. Disagreement in binary classification was also experienced by Topham *et al.* [121], who used the tool offered by Moffitt based on logistic regression plus the version reviewed by Rashid *et al.* named PurIST [116]. Twelve percent of the samples were attributed to be neither Basal nor Classical, and further examination revealed the median expression of this group of samples to be intermediated between the two extremes classes Basal and Classical. What was observed introduced the idea of gradient subtyping, where tumors with hybrid features can be explained if we think of binary classification as a continuum subtype showing a progressive shift going from classical-like type to basal-like. Because of the multitude circumstances, PDAC can adopt in terms of tissue composition and cellular environment, the concept of tumor subgroups begins to be seen as a spectrum of features forgetting the idea of distinct and exclusive classes.

To better capture the heterogeneity of PDAC, Nicolle *et al.* [123] took RNA-Seq of patient-derived xenografts (PDX) samples and applied the notion of a continuum spectrum of classes to build a predictor for PDAC patient prognosis. The expression obtained from sequencing the samples was used to decompose the data with independent component analysis (ICA) and, through histology, samples were allocated to five subtypes. The authors computed the correlation between the five subtypes and the ICA components and identified the one with the highest correlation. Such a component takes the name of Pancreatic Adenocarcinoma Molecular Gradient (PAMG), a vector of continuous values, one for each tumor sample. The efficacy of PAMG was assessed by comparing it with a dichotomous classification and the transition observed confirmed the hypothesis of a continuum class more prognostic relevant than separate classes. The same conclusion was reached when Nicolle and colleagues used the tool PurIST to predict the Basal and Classical and compare them with their gradient classifier. The outcome they obtained shows how PAMG detects the heterogeneity of the PDAC accurately, accounting for the tumor diversification more than a binary schema.

The different ways samples were used might influence the subtypes' identification. Puleo employed the samples in their whole, without non-tumor cellularity filtering. On the other hand, Bailey only used samples with tumor cellularity > 40%. Moffitt and Collisson both performed microdissection, removing the stromal compartment from the tumor, the first one *in silico* and the second one *via* laser.

## 1.5 Meningioma

Meningioma is the most frequent central nervous system and intracranial type of cancer [124]. Meningioma arises from the meninges and its prognosis is positive in most cases, as its morphology makes it suitable for tumor resection [125]. About 80% of the meningioma cases are benign while malignant ones are more exceptional with approximately 4-5% of meningioma samples classified as malignant [126, 127]. According to the classification in grades provided by the World Health Organization (WHO), a benign meningioma has a WHO grade I while a malignant meningioma has a WHO grade III, characterized by fast progression and a higher chance of recurrence after surgery with a 5-year survival rate of around 64% [128]. Meningioma is widely known to be age-related and sex-specific, found more frequently in females [125–127]. However, the exact cause responsible for meningioma formation is not fully clear and it seems related to multiple circumstances. Environmental factors such as radiation exposure [129] and head trauma [130] were found highly associated with meningioma while there is no strong relationship with genetic heritage. Indeed, numerous studies performed on patients exposed to X-ray and radiation therapy proved the causation between ionizing radiation and the risk of developing meningioma in elderly life [127, 129, 131–133].

## 1.6 Tumor budding (TB)

Tumor budding (TB) is defined as the detachment of cancer cells from the main bulk, infiltrating the adjacent parenchyma as individual cells or clusters of less than five cells (see Figure 1.7). An agreement was found in literature concerning a poor prognosis for patients with a high count of buds [134–138]. As a matter of fact, TB is already known to be a prognostic indicator for cancer types such as colorectal cancer [139–141], pancreatic cancer [142] and laryngeal cancer [143], plus lung-SCC, largely studied in the last years [135–138].

Numerous studies mirrored TB to the epithelial-mesenchymal transition (EMT). EMT plays a key role in enhancing cell migration by transitioning cell morphology from epithelial to mesenchymal [144]. Mechanisms linked to EMT are the modification of the cytoskeleton and ECM degradation. Moreover, EMT leads to increased motility and invasion of the cells. TGF$\beta$ and WNT are the signaling pathways known to be responsible for EMT [145]. The two signaling, together with microRNA-34 and microRNA-20, regulate the activation of three transcription factors, ZEB1, TWIST and SNAIL, which leads to a decrease in the expression of E-cadherin, an adhesion molecule that determines the epithelial phenotype of a cell. At the same time, cells gain enrichment in mesenchymal markers such as vimentin [146, 147]. Epithelial cells are found in the tumor bulk while cells part of the buds have mesenchymal cell traits. The application of a grading scheme serves as instrument to help clinicians evaluate the magnitude of tumor budding activity and its link to aggressiveness. Two system have been proposed to grade the TB activity, by accounting for the amount of buds and the number of cells within them [148, 149].

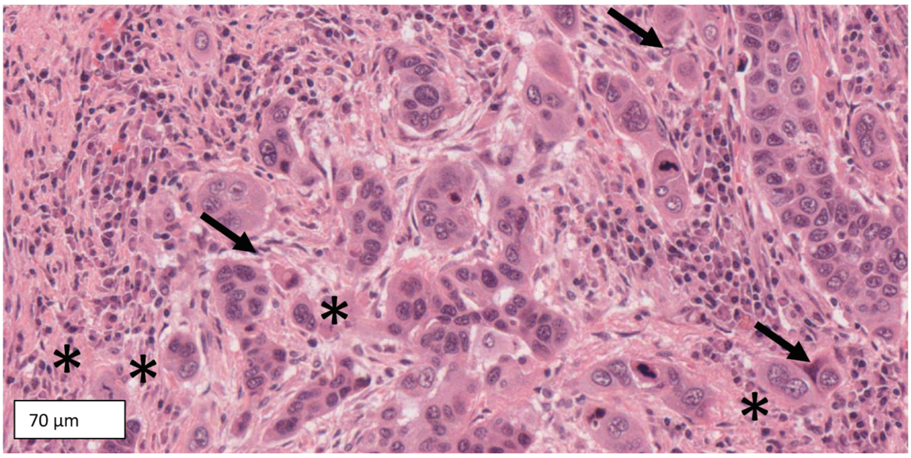


**Figure 1.7** Tissue slide indicating tumor buds (indicated by asterisks) and single-cell invasion (indicated by arrows). Figure from [134], with permission of Elsevier.

## 1.7 Lung Squamous Cell Carcinoma (lung-SCC)

Lung cancer is the deadliest cancer type, registering 18% of deaths from cancer in 2020, and the second most prevalent after breast cancer [1]. Non-small cell lung cancer is the most diffuse histological subtype of lung cancer and can be further divided into subcategories, with lung squamous cell carcinoma (SCC) as the prevailing lung cancer subtype after adenocarcinoma. Lung-SCC is characterized by slow progression and squamous cell morphologic structure, arising in the airways from the trachea to the lungs, in the bronchi proximity. Because of slow tumor growth, it is usually diagnosed at an advanced stage with a five-year survival rate of approximately 16%, a percentage linked to the tumor progression status where stage I can reach a level of five-year survival rate of 50% and above [150]. Lung-SCC is strongly associated with tobacco consumption [151, 152] and with advanced age, and lung-SCC patients present high comorbidity levels. Several studies found this cancer type to be characterized by high mutational compounds, with alteration in TP53, FGFR, PIK3CA, EGFR, and PTEN [153, 154]. Most analyzed prognostic features are tumor-node-metastasis classification, sex, age, and other clinical features, part of the American Joint Committee for Cancer staging system [155, 156]. However, knowledge of histomorphologic prognostic markers is still poor. Research considers TB as a histopathological biomarker for numerous cancers [139–143], and its potential is starting to be considered also for lung-SCC. Indeed, the aggressiveness of lung-SCC has been found related to TB, where patients with high activity were found having poor survival [135–138]. Another morphological trait named STAS is studied in the context of lung cancer. STAS, the abbreviation for spread through air spaces, is the dissociation of tumor cells from the bulk and their spread in the alveolar space [157]. Figure 1.8 gives an example of a tissue slide with STAS foci. STAS has been studied in lung cancer but not so much yet in the SCC subtype [158, 159].

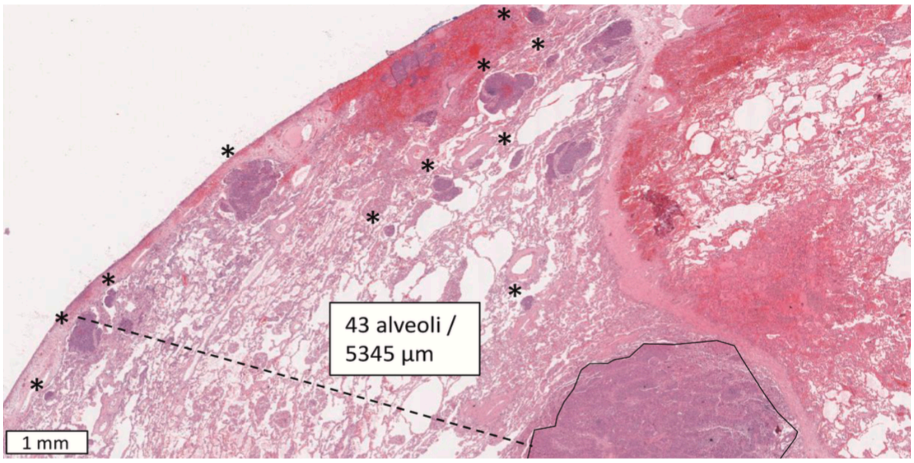


**Figure 1.8** Tissue slide with representation of STAS, indicated by the asterisks. A dashed line indicates how the distance between tumor bulk and the most distant STAS is measured. Distance is then measured counting the alveoli observed in between. Figure from [134], with permission of Elsevier.

## 1.8 Differentially Methylated Regions (DMRs)

DNA methylation plays an important role in regulating tissue-/cell-/condition-specific protein transcription, X chromosome inactivation and embryonic development [61]. Knowing the genomic location where methylation took place can unravel the mechanisms that triggered a disease or condition or that elucidate a phenotypic difference. For instance, the comparison between primary and metastasis or between knockout and mock could be driven by aberrant methylation levels.

Always considered at a single loci or CpG context level (e.g. CpG island/shore or shelf), the focus on DNA methylation has been moved towards the study of aberrant methylation in genomic regions. It has been proven that adjacent CpG sites have proportional methylation levels and that neighboring sites tend to cluster together [160]. Hence, taking CpGs in clusters might give more robust information than taking them individually. Differentially methylated regions (DMRs) are genomic regions in which the methylation status between conditions is different. DMRs carry more biological insight than differentially methylated loci (DML) because of spatial correlation, thus having more statistical power in determining significant results.

### 1.8.1 State-of-the-art tools for DMRs discovery

Numerous tools have been released for the identification of DMRs. Nevertheless, their implementation is not always easy. Basic programming knowledge (most commonly R) is often needed, and many computational tools are applicable only under specific characteristics with a particular data format or from a specific platform, either solely array, WGBS, or both. We summarized the main features of the most popular DMR detection tools in Supplementary Table 5.2.

For some methods, regions are biologically pre-defined such as gene promoters, FANTOM5 enhancers and CpG islands, or tiling regions, meaning sections of the genome of a given size (e.g. 1-5 kb). However, most of the methods allow the identification of *de novo* DMRs, for instance with the use of a genomic window of dynamic size, extended for as long as the DMR is observed. Several methods are considered the gold standard of DNA methylation differential analysis. An example is the Minfi pipeline [161], available in R and built for analyzing array data starting from raw files. Minfi implements the bump hunting algorithm proposed by Jaffe *et al.* [162] and it works first by computing a t-test for each CpG and then collecting CpGs in clusters, as potential DMRs, based on the t-statistics values. In this way, Minfi allows the identification of novel regions without prior information about their size or genomic location. In addition, Minfi included a permutation step to ensure the identification of significant DMRs. Besides the t-test, the identification of DMR is also built on other tests like Fisher's exact test in the case of methylSig [163], a two-dimensional Kolmogorov-Smirnov test used by Metilene [164] and a Wilcoxon or Kruskal-Wallis paired non-parametric tests (depending on the number of conditions to be compared) implemented in methylPipe [165]. In the case of bisulfite sequencing, among the tools for *de novo* regions detection there are BSmooth [166], DMRfinder [167], methylKit [168] and DMRcate [169]. DMRcate is another tool widely implemented in literature and seen as a gold standard choice in DNA methylation analysis. It is designed for both microarray and sequencing data. DMRcate searches for regions by using a Gaussian kernel smoothing approach. Besides novel regions, many existing tools examine methylation levels inside pre-existing regions or tiling windows. Results from comparing pre-established regions of interest are more robust and generally more significant, with their interpretation being more straightforward. Additionally, these methods are particularly suitable when the user wants to observe methylation activity in some specific regions/areas. Examples of methods that analyze predefined regions are edgeR [170], designed for bisulfite sequencing data, and RnBeads [171], suitable for both array and sequencing-based methylation profiles. Nevertheless, the research of indefinite region type or length not only might give more realistic results, given that DMRs are most likely not all of the same size or in the same genomic area, but it also allows the discovery of novel insights related to aberrant methylation activity located in areas of the genome still unknown in literature. Tools like Metilene [164], methylSig [163] and methylPipe [165] give the opportunity to choose between both types of regions and are built for the analysis of bisulfite sequencing data. Another relevant aspect to consider when performing a bioinformatic analysis is that not all users have experience with programming, hence a tool to be used by as many users as possible has to be user-friendly and of easy implementation. Most of the tools, as shown in Supplementary Table 5.2, are available in the Bioconductor library of R, a few are available as a command line tool and rarely available as a graphical user interface (GUI).

## 1.9 Aim of the study

### Aim 1

One of the aims of this study is to assess the reproducibility and consistency of the PDAC subtypes proposed by Collisson *et al.* [89], Moffitt *et al.* [88] and Bailey *et al.* [90] and give a first critical evaluation of Puleo *et al.* [91] subtypes, whose assessment is still poor. The skepticism raised by the literature stimulated us to perform further analysis to verify the generalizability of the PDAC subclasses, elucidate potential biases and, from them, contribute to the complex understanding of PDAC heterogeneity.

### Aim 2

Numerous tools have been proposed for the discovery of DMRs, however, only few of them return interpretable and accurate results as well as easy implementation. A second aim of this study is the presentation of Dimmer 2.0, a tool for *de novo* DMRs search first released in 2016 by Almeida *et al.* [172]. We present an upgraded and extended version of the tool, followed by its application to two study cases:

- PDAC is characterized by its lethality linked to a high metastatic rate which makes the healing process challenging. Metastasis and aggressiveness of the disease are highly related to the rich stromal environment and the ECM compound, which is known to be influencing tumor progression and spread. The first site where PDAC grows metastasis is the liver because of its location concerning the pancreas and because of the abundance of blood vessels. However, whether aberrant methylation plays a role in the metastatic spread is still unknown. To elucidate this aspect, we apply Dimmer 2.0 to PDAC primary and liver-matching metastasis to reveal potential epigenetic markers linked to tumor diffusion.

- Epigenetics might elucidate mechanisms able to explain the aggressiveness that characterizes malignant meningioma, thus we compare its methylation status with the one found in benign tumors.

### Aim 3

High tumor budding activity has been found associated with an advanced stage of cancer and spread metastasis, leading to a poorer survival [173]. Established cut-offs are arbitrarily decided and only successively their survival impact is examined [148, 149]. Third aim of this thesis is to identify a robust and prognostically relevant grading system for SCC of the lung based on tumor budding. Additionally, STAS foci, distance of STAS in alveoli, minimal cell nest size, stromal content and immune cell infiltration are examined for a cut-off associated with survival. All the aforementioned histomorphological traits are analyzed in relationship with death risk, disease progression and relapse to assess their role as potential markers for lung-SCC. Molecular features that drive TB and that can be used as prognostic markers still need to be investigated. Third aim extends to understanding the relationship between TB and diverse molecular profiles, with the final goal of identifying the molecular features that are able to explain the role and origin of this histological phenomenon. A deep understanding of TB can help assess its potential key role as a prognostic biomarker.

# 2 Material and Methods

## 2.1 Data

### 2.1.1 PDAC

PDAC trascriptomic data used was downloaded from public repositories and consists of seven microarray expression matrices and two RNA-Seq expression datasets. Four of these dataset correspond to the ones from Moffitt, Collisson, Bailey and Puleo plus five additional datasets (ICGC-Array, TCGA-PAAD, Badea *et al.* [174], Sandhu *et al.* [175] and Yang *et al.* [176]). From the publications of Moffitt, Collisson, Bailey and Puleo we obtained the subtypes labels assigned to the samples and the lists of genes used as signatures.

Moffitt and Collisson microarray expression data were downloaded from the public repository GEO (accession codes GSE71729 and GSE17891). Data were downloaded already normalized and no further preprocessing was performed. Non-primary PDAC samples are filtered out.

Data from Bailey was obtained by author correspondence in a table of RSEM counts. The data is RNA-Seq gene expression (ICGC PACA-AU) and it was normalized by us with DESeq2 4.2. with the variance stabilizing transformation method. As before, we only kept primary samples and we additionally removed samples of non-ductal histological subtype which were included in the obtained matrix.

Puleo microarray profiles were obtained from ArrayExpress under the accession code E-MTAB-6134.

We downloaded from the International Cancer Genome Consortium (http://www.icgc.org) a dataset of microarray expression data, available under the project PACA-AU. Data was already pre-processed and we further filtered out non-primary samples and non-ductal histological subtype samples.

From the Xena Browser [177] we downloaded the RNA-Seq expression matrix of the TCGA-PAAD project. The matrix, obtained in a log2 counts format, was transformed to counts prior to normalization in DESeq2 4.2. through the variance stabilizing transformation.

From GEO we downloaded the remaining three matrices: Badea *et al.* (GSE15471), whose replicate samples were excluded, Sandhu *et al.* (GSE60980), and Yang *et al.* (GSE62452).

The gene nomenclature used is gene symbols. In the case of translation from probe ids, whenever a gene name is linked to multiple probes the median value is taken.

### 2.1.2 Matched PDAC primary and liver metastasis

DNA methylation data for PDAC and the corresponding liver metastasis was obtained from tissue samples collected from patients by the University of Heidelberg. The samples are 11 pairs of PDAC primary and liver metastasis belonging to the same patient. DNA was quantified with the TaqMan® RNAseP Detection Assay and extracted from FFPE samples using Maxwell® RSC Blood DNA Kit. After bisulfite conversion, FFPE samples were treated with the Infinium FFPE DNA Restore Kit for FFPE restoration to avoid nucleic acid degradation that might derive from FFPE samples. DNA methylation was generated using Infinium MethylationEPIC BeadChip arrays (Illumina) following the original protocol. For the comparison between primary PDAC and liver metastasis idat files were used, with a total of 866 091 CpGs used for the comparison.

### 2.1.3 Meningioma

Samples were collected and prepared for DNA methylation analysis by the University Hospitals of Heidelberg and Mannheim between 2015 and 2018 and published in [178]. Samples were stored as fresh frozen tissues, as well as FFPE and only samples rich in tumor cells were considered. In the case of frozen tissue, Invisorb® Genomic DNA Kit II protocols were implemented for DNA isolation, while Maxwell® 16 FFPE Plus LEV DNA Kit protocols were used for FFPE samples. DNA is then treated with bisulfite before data retrieval, performed with both Illumina Infinium 450k and EPIC array with the use of Illumina iScan for the BeadChips scan which returns .idat files. Additional details about sample preparation and processing are described in the work published by Capper *et al.* [178].

The set is made of 414 samples, of which 355 are classified as benign and 59 as malignant. Only CpGs shared between samples were considered, for a total of 375 197 loci.

### 2.1.4 Lung-SCC

All the data used for discovery (both molecular and clinical) are part of the TCGA-LUSC project, generated by TCGA (https://www.cancer.gov/tcga), with the exception of the validation cohort that was independently generated.

#### 2.1.4.1 Clinical and morphological traits

**Study dataset.** Digitalized H&E tissue slides suitable for histopathologic evaluation were obtained from the GDC Data Portal (https://portal.gdc.cancer.gov) and they were analyzed with the ImageScope x64. Samples with low quality or tumor type different from lung-SCC were excluded. Data used consists of 335 samples.

**Clinical data.** Overall survival (OS) and progression-free interval (PFI) were used for survival analysis, were OS measures the length of the life of a patient after diagnosis, while PFI is considered as the amount of time a patient does not register a progression of the disease, during or after a treatment. AJCC stage IV was merged with stage III because only few samples were available.

**Tumor budding**. Low magnification was used for analyzing the area most enriched in tumor buds. TB was counted in absolute numbers in ten consecutive high-power fields (HPF) in light microscopy (40x magnification in the slide), where one HPF corresponds to a field of diameter 0.38 mm. A tumor bud is defined as a cluster of fewer than five cells (as well as a single cell) located in the stroma adjacent to the tumor bulk.

**STAS and distance in alveoli**. For STAS is considered the count of tumor cells that detached from the tumor bulk and are found in the alveoli [157]. Only cells with undamaged perimeter are considered. Distance of STAS was counted in the number of alveoli between perimeter of bulk and the most distant point of STAS.

**Immune cell and stroma estimation**. Area inside the perimeter of the tumor was analyzed to estimate stroma compound and immune cell infiltration, at low-power magnification. Necrosis or alveolar macrophages observed in the area are excluded. At higher magnification, lymphocytes are searched and their presence in the tumor area is estimated in percentage. Suggestions from Hendry *et al.* [179] were followed for immune cell measurements.

Stromal proportion was also estimated in percentage, dividing the fibrous regions observed in the area by the entire area of the tumor.

**Minimal cell nest size**. Minimal cell nest size (MCNS) represents the number of cells in the smallest buds [180].

**Existing grading system**. A TB grading system previously established was used for assessing prognostic impact [148, 149]. The categories are G1, G2 and G3, based on the a amount of tumor buds combined with MCNS. To samples with amount of buds from 1 to 14 is assigned 1 point while 2 points if more than 14. One point is assigned if MCNS is more than 15, two points is size is from 5 to 15 cells, three points if size is from 2 to 4 and four points to single cells. For each sample, points are summed up. Group G1 is assigned to samples that registered two or three points, G2 to samples with four, five or six points while samples with 7 points are categorized as G3.

**Validation dataset**. Samples for validation are collected by the Department of Thoracic Surgery of the University Hospital of Heidelberg during the years from 2010 to 2017. Samples were obtained as H&E-stained slides from the tissue bank in Heidelberg (Germany) of the National Center for Tumor Diseases (reference number S-2021-315).The cohort consists of 346 lung-SCC samples and survival analyzed are OS and disease-free survival, which measures the amount of time a patient is free of the disease (disease-free survival (DFS)), after a treatment.

All the histopathologic examination were carried out by Fabian Stögbauer and Melanie Boxberg, as deeper described in [134].

#### 2.1.4.2 Omics data

Molecular data was downloaded from the Xena Browser under the project TCGA-LUSC [177].

The molecular data used are gene expression, protein expression, non-silent mutation, and DNA methylation. Only samples with TB information will be considered (333 for gene expression, 123 for protein expression, 322 for non-silent mutation, and 247 for DNA methylation).

The gene expression profile (RNA-Seq) of 333 samples was obtained with Illumina HiSeq 2000 platform and downloaded as $log_2(x+1)$ normalized RSEM counts. Genes with more than 30% of zeros are deleted.

Protein expression was obtained with the reverse phase protein array technique (RPPA) and next quantified with the SuperCurve Fitting method. We downloaded the data already normalized with replicate-base normalization. Data (123 samples) were downloaded and no pre-processing step was applied.

Binary non-silent mutations (SNPs and indels) were downloaded at the gene level for 322 samples. Genes with at least 98% of zeros were removed from the data.

DNA methylation is obtained at the CpG level where methylation was profiled with Illumina Infinium HumanMethylation450 (450K). The data table consists of 247 samples. Methylation values are in the $\beta$-value unit. We further filtered out loci with at least 20% of missing values. K-nearest neighbors was used for CpG with less than 20% of missing values, using Scikit-learn 0.24.2 in Python.

We assessed the association between TB and the gene expression-based lung-SCC subtypes published by Wilkerson *et al.* [181]. The subtype calls for each sample were downloaded from a study conducted by TCGA [153].

## 2.2 Machine learning methods for omics data analysis

Bioinformatics processes biological and clinical data with a combination of computational tools, and statistical and mathematics approaches. Molecular data availability and size keep growing and the advancement in algorithm and programming development enables the bioinformatics field to rapidly expand. An important focus is addressed on methods development and mathematical models that can support decisions and give predictions. The process of extracting information from a given dataset is often performed by machine learning. Defined as an artificial intelligence subfield, machine learning creates automated models able to learn and adapt, returning informative conclusions based on the data pattern. A key task of machine learning is the ability to handle large amounts of data and perform multiple analyses divided into *predictive*, *descriptive*, and *prescriptive*. The *predictive* analysis, in the case of biomedicine, predicts survival, disease relapse, or therapy efficacy; the *descriptive* type is used to observe a phenomenon or condition from a given dataset and explain it; the *prescriptive* type is used to support a decisional action to optimize the outcome. Machine learning has multiple applications in the field of bioinformatics, including biomarker discovery, protein structure prediction, cancer type classification, outcome prediction, drug discovery, and repurposing but also the processing of digital images and text.

Machine learning techniques can be categorized into three groups: supervised, unsupervised, and reinforcement. Supervised techniques are used to make predictions and use datasets where the observations are labeled. The model uses the labels to learn how to classify new observations together with the use of initial training data. Supervised techniques can be further distinguished into classification (e.g. random forest, k-nearest neighbors, support vector machines) and regression (e.g. linear and logistic regression) techniques. Unsupervised techniques, contrarily to supervised, do not use any labeled data and are employed for hidden pattern recognition. Clustering and dimensionality reduction are the two main categories of unsupervised machine learning, where examples of methods are k-means and hierarchical clustering for the first case and principal component analysis for the second one. Reinforcement learning is based on a machine being able to adapt to the environment it is located in by learning from its errors, following a reward/punishment approach. The interaction with the environment is used to teach the machine the best behavior to optimize the rewards.

### Linear regression

Linear regression [182] is a supervised technique employed to model the relationship between one or more predictive features and a response variable. The concept behind linear regression is to build a model capable of explaining such a relationship as well as possible and quantify its strength and statistical reliability. The model searches for the line that best fits the data. The classic linear regression equation is $Y = a + bX$, where $Y$ is the dependent variable and $X$ a matrix whose columns are the independent variables. With linear regression, we can observe how much $X$ can explain $Y$. The regression coefficient $b$ tells how much $Y$ changes, on average, if $X$ changes one unit. The relationship can be positive, where the increase of $X$ corresponds to an increase in $Y$, or negative, where the increase of $X$ corresponds to a decrease in $Y$. Each coefficient is assigned a p-value, indicating whether the relationship expressed by the coefficient is statistically significant. Least-squares is the best known regression method [183]. It looks for the line that minimizes the (average) distance from the regression line to all data points of a given dataset. Such distance takes the name of residual (see Figure 2.1), a measure of error representing the variation of $Y$ that is not explained by $X$. Hence the goal of the least-square method is to minimize the sum of square residuals, here is the name of the method.

Important in linear regression are the key assumptions that have to be respected: no multicollinearity, when, in the case of multivariate regression, variables are highly correlated with one another; no

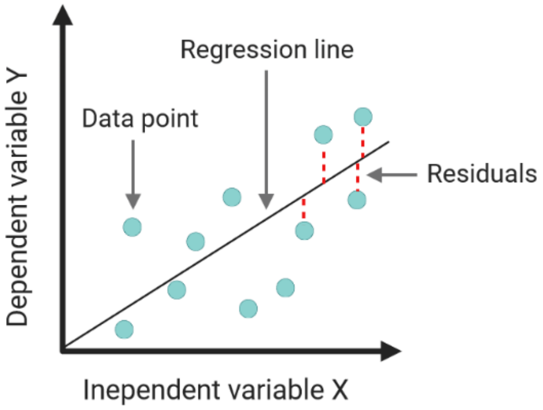


**Figure 2.1** Least-square method works by fitting a line to the data and moving it until the optimal line is found, which is the one with the least sum of squared residuals.

correlation with residuals, when the errors are highly correlated with the target variable; normality and homoscedasticity of residuals, the errors should follow a normal distribution and have equal variance; linearity, which assumes that the dependent variable is a linear combination of the independent variables and the additional coefficients, if any.

### Random forest

Random forest is part of the decision tree learning, the branch of machine learning that performs decision-making based on a tree model [184]. Being a classification method, it follows a supervised approach where classes are known and the algorithm predicts to which class an observation belongs. The model consists of an ensemble of individual decision trees, where a tree is a representation of a chain of decisional processes [184]. Starting from a feature, represented by a node, two branches are generated by classifying the observation in groups according to a condition. Each branch terminates with a node from which new branches will be created. After every split, the model measures the homogeneity, also called purity, of the child nodes. Gini index and entropy are some of the most common purity measures [185]. In this way, a tree is built by progressively categorizing the observation until reaching groups with the highest homogeneity. Hence, a tree performs a classification thus it returns a predicted class. As mentioned, random forest is made of multiple decision trees, where their construction is independent and the class predicted by most of the trees will establish the final prediction. This approach is named majority vote. A random forest model is illustrated in Figure 2.2. The trees in the model are built under a bagging procedure aggregating trees, each of them built based on bootstrapping, hence by random sampling with replacement a subgroup of observation from the initial set of observations. All trees are independent and identically distributed. The characteristics of random forest make it ideal for limiting overfitting, a condition when the predictions fit too well the data it was trained on, limiting its suitability for other independent sets of data. This is because random forest is trained on a set of randomized and independent trees, which maintains low bias while reducing the variance of the model.

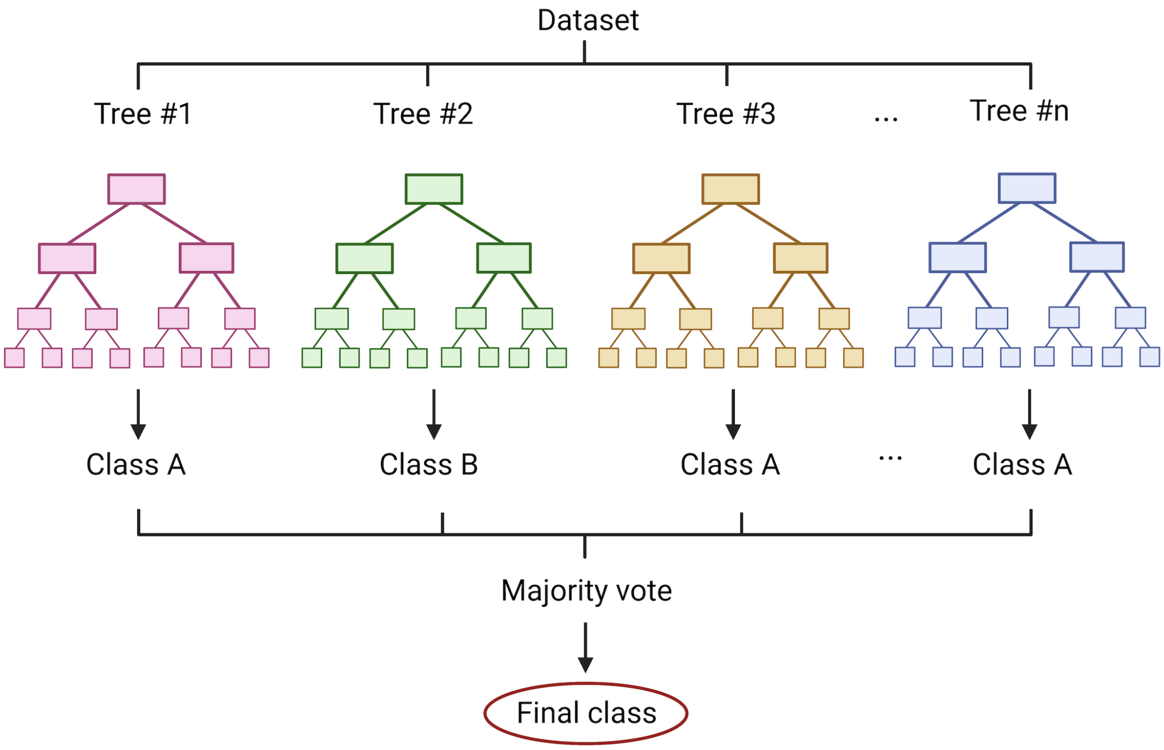


**Figure 2.2** Example of a random forest model. The model is made of an ensemble of independent decisional trees where each of them returns a predicted class. The final prediction is given by the class that was predicted the most, following the rule of majority vote.

### Hierarchical clustering

Hierarchical clustering [186] is a common clustering method that works by organizing observations in homogeneous clusters following a hierarchical structure. There are two ways of performing hierarchical clustering, an agglomerative and a divisive way. Agglomerative hierarchical clustering starts by taking each data point as independent cluster and progressively agglomerating them until only one cluster is left. The aggregation occurs by merging clusters based on their similarity; whether clusters are similar or not is determined by a linkage measure such as single linkage, complete linkage or average linkage [187]. The result of such consequent agglomeration is a dendrogram, a clustering tree that reflects the order of the clusters (see Figure 2.3). Divisive clustering works as the agglomerative but in the opposite direction, starting from one big cluster which contains all the observations and progressively splitting it into subclusters until reaching single data point clusters.

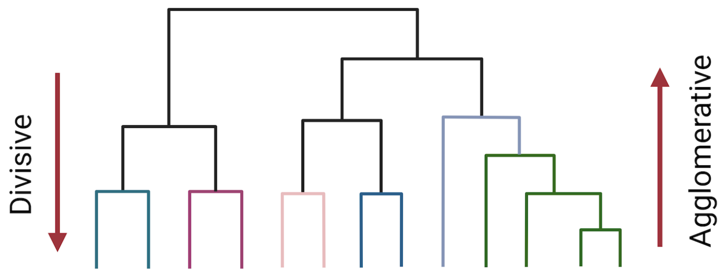


**Figure 2.3** Representation of agglomerative and divisive hierarchical clustering where the final clustering procedure reflects a hierarchical structure.

## 2.3 Computational methods for cancer subtyping

The role of a subtyping signature is to reflect the molecular changes between samples belonging to different subtypes and arrange them in homogeneous categories. Based on diverse molecular data such as genetic variant, gene expression, miRNA, and DNA methylation, potential signatures are computationally identified with the use of algorithms based on pattern recognition in an unsupervised setting, like clustering techniques, or supervised like k-nearest centroids, where subtype labels are known *a priori*. The goal of the analysis is to divide samples into categories that usually reflect cancer aggressiveness, stage, metastatic presence, risk of relapse, and survival expectation as well as different cellular compositions. Starting from a feature-by-sample table, the analysis goal is to identify those features that better differentiate the samples. Hierarchical clustering is often used in the context of tumor subtyping. The 50 genes for breast cancer classification proposed by PAM50 [188] have been obtained by applying hierarchical clustering on gene expression profiles from 189 breast cancer samples. Out of 189, 122 samples show five clear gene expression patterns, indicating five subtypes. A gene signature was established by taking the ten most cluster-specific genes per subtype, here the number 50. The authors offer centroids for the sample classifications, computed with the nearest centroids classifier. Roepman *et al.* [189] identified three intrinsic colorectal cancer subtypes with the implementation of hierarchical clustering on mRNA profiles of a 188 samples cohort. Three subtype-specific signatures are proposed, made of 32, 53, and 102 genes. The signatures were established by computing a pairwise t-test combined with cross-validation. Another popular method is consensus clustering-based non-negative matrix factorization. Non-negative matrix factorization (NMF) [190] is a dimensionality reduction technique used for feature extraction. NMF decomposes a starting matrix $V$ of dimension $m \times n$, where all elements are non-negative, into the product of two matrices, one named feature matrix $W$ ($m \times k$) and a second one named coefficient matrix $H$ ($k \times n$), both non-negative. The values of $W$ and $H$ are iteratively changed until the best approximation of V is reached. The resulting feature matrix is a set of $k$ new latent features deriving from the linear combination of the original features. In biomedicine, the features are usually biological components helpful for interpreting the meaning of the data composition. NMF is commonly used together with consensus clustering and cophenetic coefficient to determine the optimal number of clusters. Consensus clustering-based NMF for subtypes discovery see its implementation in numerous studies [88–90, 191], with different ways of retrieving meaningful subtype-related features. A common approach combined with NMF is the multi-class significance analysis of microarray (SAM). SAM computes t-test and data permutation to verify whether the expression of a gene is significantly different across groups. Such an approach for the identification of cancer subtypes and associated gene signature was implemented in the study of colorectal cancer [191] and allowed the identification of five classes of colorectal cancer from mRNA data and a signature of 786 genes. The same procedure was used for the study of pancreatic ductal adenocarcinoma implemented by Collisson *et al.* [89] and Bailey *et al.* [90], discussed more in detail in the next chapter.

After assigning samples to clusters, these have to be translated into subtypes by recognizing their biological insight. Subtypes are defined as groups with distinct molecular and clinical characteristics and their delineation is crucial for precision medicine. Different subtypes are linked to different survival outcomes, thus analyzing the association with survival information is a fundamental step for the aggressiveness assessment. Subtypes are further distinguished by composition, hence further characterization is carried out by estimating the TME and the cell types populating the samples, such as immune cells. Gene-set enrichment analysis on the genes most enriched in each subtype helps discover the mechanisms and functions they are involved in and that distinguish the subtypes.

## 2.4 Computational methods for PDAC signatures evaluation

This section contains the descriptions of methods used for the computational evaluation of transcriptomic derived PDAC signatures proposed by Moffitt *et al.* [88], Collisson *et al.* [89], Bailey *et al.* [90] and Puleo *et al.* [91].

We first observe the signatures' behavior in clustering independent datasets and as subtypes predictors. Next, we implement additional analyses to understand their prognostic difference, their functional activity, and their cellular content. We finally compare the performance of the real signatures with the one obtained when using a random gene panel or shuffled subtypes labels. A schematic summary of the main steps is shown in Figure 2.4.

### 2.4.1 Expression distribution

Assuming that a gene part of a signature might be representative of one subtype, hence not active in the others, we expect it to carry a different expression distribution across different subtypes. To evaluate such difference, we use for each gene in the signature the Wilcoxon test and Kruskal-Wallis test (both two-sided) to compare two groups or more. A deeper investigation is dedicated to Puleo,

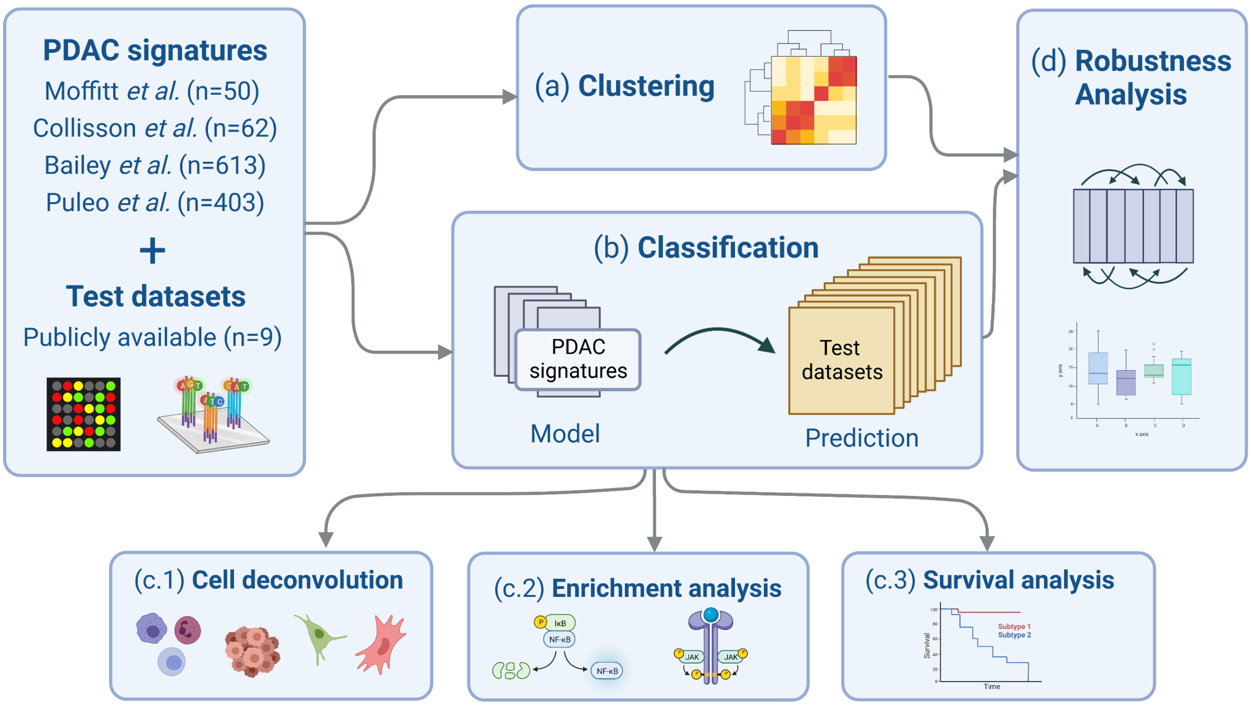


**Figure 2.4** Overview of the main analysis performed for signatures evaluation. Starting from the data collection, established signatures with relative subtype calls plus the download of nine independent validation datasets are obtained. The first analysis is clustering the datasets by using the gene panels (a), followed by building classification models based on the signatures and the subtyping schemas (b). Three additional analyses are performed on the predicted subtypes to assess their cell composition through single-cell cell deconvolution (c.1), their functional pathways with enrichment analysis (c.2), and their survival differences (c.3). Finally, robustness analysis (d) is performed. Figure from [113].

whose five subtypes are compared in a pairwise fashion, with one subtype against the other four. Bonferroni correction (p-value <0.05) is applied to any case. SciPy 1.6.0 in Python is used.

### 2.4.2 Clustering for subtypes reproducibility assessment

We adopted the agglomerative hierarchical clustering technique and truncated the dendrogram at the number of subtypes to reproduce. Before performing the clustering we apply z-score transformation to the data and we filter only the genes from the signatures. This step is performed by taking one dataset and one signature at a time, and submitting the reduced dataset to agglomerative hierarchical clustering with correlation as metric and determining the two (Moffitt), three (Collisson), four (Bailey) or five (Puleo) subtypes. We used two statistical tests, the Wilcoxon rank sum test for two clusters comparison and the Kruskal-Wallis test otherwise, to assess the difference in expression between the obtained clusters. Both tests are used by taking the genes in the signature, one by one, computed two-sided and the returned p-values are Bonferroni corrected (p-value <0.05). We included Rank Index

(RI) and Adjusted Rand Index (ARI) to quantify, in percentage, how much real subtypes match the clusters.

### 2.4.3 Comparison with molecular gradient

The clusters were compared with PAMG scores for each of the nine cohorts. The computation of PAMG is based on four independent datasets, where two are from ICGC (array and RNA-Seq), one from Puleo, and one from PDX. The set of nine datasets collected also contains the first three data just listed, thus we attribute them to the PAMG scores computed on the corresponding data. For the remaining datasets, we used the PAMG scores computed on data from the same platform. Scores were computed by Dr. Nicolas Fraunhoffer and Dr. Nelson Dusetti, which we acknowledge. We additionally perform Cox regression on PAMG, implemented in Python with the Lifelines 0.25.9 package.

### 2.4.4 Use of the signatures to build a subtypes prediction model

The subtype calls for the nine datasets were classified with the random forest technique, implemented with default parameters in Python using the Random Forest Classifier package in Scikit-learn 0.24.2.

Before building the prediction model, we combine the nine datasets considering only the intersecting genes. Each dataset comes from a different source, the samples were prepared under different settings and the data were obtained using different techniques. To remove such confounding factors, we performed a data correction by removing the influence derived from the different origins of the data. We can see how the data adjustment helped remove source influence with a principal component analysis shown in Supplementary Figure 5.2. This step was carried out in R using the function RemoveBatchEffect in the limma library, version 3.48.1.

We applied z-score transformation before using random forest for PDAC subtypes prediction. All four signatures and the four sets of subtypes with related data were used for constructing prediction models, sixteen in total. In this way, each of the nine datasets has a label predicted for the four subtype groups based on the four signatures. The workflow for building a prediction model starts from taking the datasets used in the discovery studies, one by one, together with their subtype labels, and keeping only the genes in the signatures, taking one signature and a time. How the models are built is graphically explained in Figure 2.5.

### 2.4.5 Assessment of signatures robustness

A signature is robust when it performs better than a set of random genes. We want to know how robust the signatures are, both in clustering the samples and predicting their subtype. The robustness in classification is measured by comparing the prediction accuracy of the models when using three different setups: predicting using the signature genes, predicting using a set of random genes of the same size, and predicting using the signature genes but randomizing the class labels. Each of them is

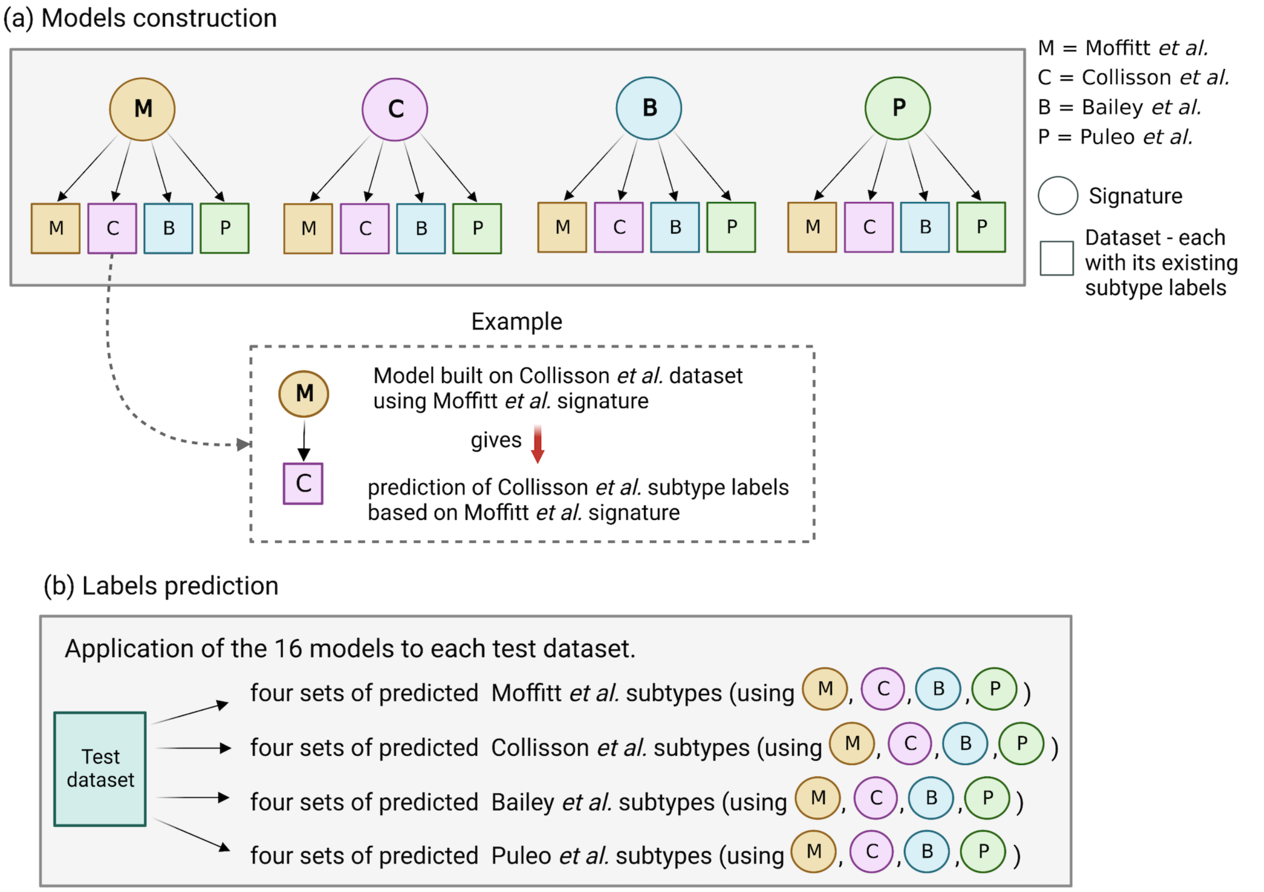

(a) Models construction
M
C
B
P
M
C
B
P
M = Moffitt *et al.*
C = Collisson *et al.*
B = Bailey *et al.*
P = Puleo *et al.*
Signature
Dataset - each with its existing subtype labels
Example
Model built on Collisson *et al.* dataset using Moffitt *et al.* signature
gives
prediction of Collisson *et al.* subtype labels based on Moffitt *et al.* signature
(b) Labels prediction
Application of the 16 models to each test dataset.
Test dataset
four sets of predicted Moffitt *et al.* subtypes (using M, C, B, P)
four sets of predicted Collisson *et al.* subtypes (using M, C, B, P)
four sets of predicted Bailey *et al.* subtypes (using M, C, B, P)
four sets of predicted Puleo *et al.* subtypes (using M, C, B, P)

**Figure 2.5** Schematic overview on the construction of the sixteen prediction models. Figure from [113].

computed 1000 times and computing average 5-fold cross-validation for performance comparison. To evaluate whether the signatures are robust when clustering, we assess the overlap of the clusters when using the true signatures and the ones when using random genes. Such overlap is measured with ARI. ARI was further implemented for pairwise comparison of the random genes-derived clusters.

To summarize the ARI just computed and evaluate the difference between them, the strictly standardized mean difference (SSMD) was used. Taking two independent groups with unequal variance, named 1 and 2, with $\mu$ and $\sigma^2$ as their mean and standard deviation, we calculated SSMD using the formula: $\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ .

### 2.4.6 Cell deconvolution and ssGSEA

To investigate the cell types present in the bulk datasets, we performed cellular deconvolution taking a pancreatic cells signature matrix based on the single-cell RNA-Seq data obtained from normal human pancreas from Tosti *et al.* [192]. Deconvolution was then implemented using CIBERSORTx [193].

To better comprehend the behavior and biological meaning of the genes in every dataset, we compute single sample gene set enrichment analysis (ssGSEA). Such analysis computes enrichment scores aggregated at the gene level, able to represent the functionalities of the genes and determine whether they are upregulated or downregulated. ssGSEA makes possible the comparison of the pathways enriched in different samples. To determine in which pathway lies the difference between the predicted subtypes, we compute t-test (two-sided) and ANOVA, followed by multiple testing correction (Bonferroni p-value < 0.05). Thanks to these two tests, we can filter the enrichment terms which show significantly different activity between subtypes, represented by the enrichment score. After filtering significant pathways and functional terms, we further want to identify in which specific subtype the term is enriched. We do so by taking the average enrichment score registered in every subtype. The term is considered up- or down-regulated based on whether the average value has a positive or negative sign, respectively. This is performed for each term individually while considering that for each dataset we have sixteen predicted labels. We performed the analysis with GSEApy 0.10.5 [194–196] (Python) using two databases for pathway terms, KEGG [197] and Reactome [198], and the Gene Ontology terms (cellular components, biological processes, and molecular functions) [199].

### 2.4.7 Prognostic relevance evaluation

Assessing the prognostic importance of a set of predicted subtypes is crucial for tumor stratification. We use the log-rank test to perform survival analysis and evaluate differences in survival between the predicted subtypes. The test is computed in a pairwise manner using the Lifelines Python package, version 0.25.9 and only on the datasets where survival times are known (seven datasets).

## 2.5 Dimmer 2.0 for the discovery of *de-novo* DMRs

In this section we present Dimmer 2.0, an extended version of the tool Dimmer, released in 2016 by Almeida *et al.* [172]. We first describe the workflow, highlighting novelties and corrections that have applied since its first release, next, we explain and illustrate step by step the algorithm used for DMRs discovery. The pipeline can be described in four main steps, summed up in Figure 2.6.

Dimmer 2.0 offers wide applicability starting from the input data. Data can come from array assay, in either raw format or pre-processed $\beta$-value matrix, or BS-Seq as bismark output files (Figure 2.6, (1.1)). A second input file named sample annotation file is required. This file contains sample IDs and information about the samples which are needed for the upcoming analysis as well as eventual confounding factors to consider in case of regression analysis. The variable of interest is specified in the annotation file which can be dichotomous (e.g. sick/healthy) or continuous (e.g. age). This variable determines whether the DML will be established with a t-test or linear regression (Figure 2.6, (1.2)). In the case of a dichotomous study, Dimmer 2.0 offers the option of designing a paired analysis (e.g. before/after treatment). Once the upload step is done, the data are processed and quality checked. In this step are performed loci filtering and background noise correction are followed by quantile normalization, which estimates and accounts for sex (Figure 2.6, (2)). The cleaning and processing step

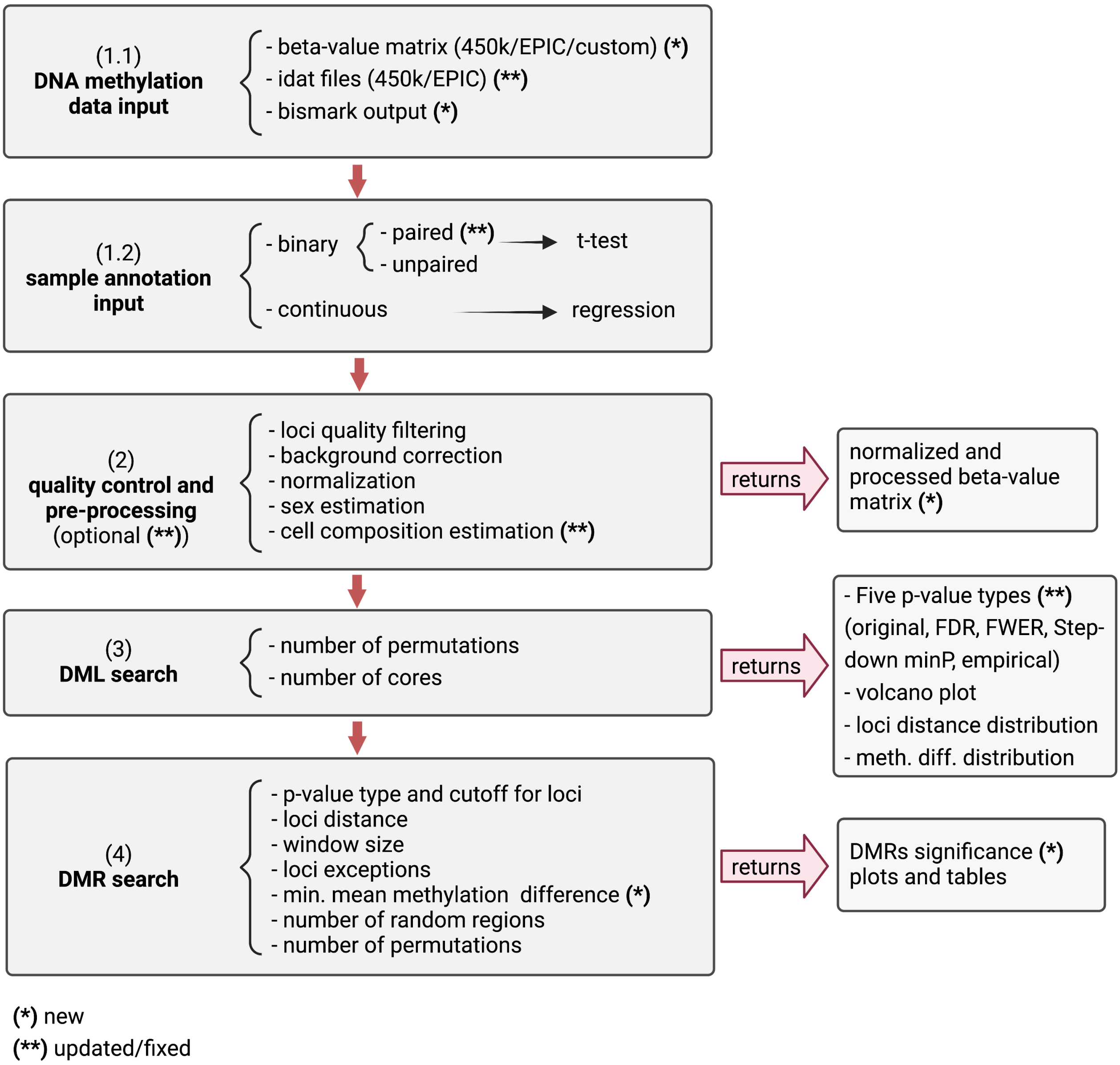

(1.1)
DNA methylation data input
- beta-value matrix (450k/EPIC/custom) (*)
- idat files (450k/EPIC) (**)
- bismark output (*)
(1.2)
sample annotation input
- binary
- paired (**)
- unpaired
t-test
- continuous
regression
(2)
quality control and pre-processing
(optional (**))
- loci quality filtering
- background correction
- normalization
- sex estimation
- cell composition estimation (**)
returns
normalized and processed beta-value matrix (*)
(3)
DML search
- number of permutations
- number of cores
returns
- Five p-value types (**) (original, FDR, FWER, Step-down minP, empirical)
- volcano plot
- loci distance distribution
- meth. diff. distribution
(4)
DMR search
- p-value type and cutoff for loci
- loci distance
- window size
- loci exceptions
- min. mean methylation difference (*)
- number of random regions
- number of permutations
returns
DMRs significance (*) plots and tables
(*) new
(**) updated/fixed

**Figure 2.6** Pipeline of Dimmer 2.0 consists of four steps, starting from the data input (methylation profile (1.1) and samples information (1.2)). Once data is uploaded, quality control, filtering, processing and normalization are executed (2). Step (3) is responsible for differentially methylated loci (DML) identification, returning five types of p-value (original, FDR, FWER, Step-down minP, empirical) which the user can choose from to define whether a locus should be considered significantly differentially methylated or not. The output of step (3) plots the distance between loci and methylation difference for each locus as well as volcano plots. Such figures are meant to support the user in the p-value type decision. The next and final step is the DMR search (4) using the default parameters or the ones passed in the input. A permutation test is used to ensure the DMRs returned are statistically significant. (*) indicates novelties introduced with the update of Dimmer and (**) indicate what was updated or fixed from the first version of the tool.

is available for both 450k and EPIC array data and is based on the adaptation of the minfi R package [161] while cell composition estimation is currently possible only for 450k. Dimmer 2.0 is now ready for identifying DML, computing either a t-test or linear regression analysis for each of the genomic positions of the input data and returning a p-value associated with each locus. The original p-values are corrected following three multiple testing correction techniques, the Benjamini-Hochberg False Discovery Rate (FDR) [200], the family-wise error rate (FWER) and the Step-down minP [201]. An additional p-value type is obtained by shuffling the sample labels before the t-test/regression analysis, returning a distribution of empirical p-values (Figure 2.6, (3)). The user can choose among five p-values types and the decision process is guided by showing figures with their distributions and volcano plots. Additional results such as loci methylation difference and distance are displayed to help the user to choose parameters for the DMRs search step. The final step is the research of DMRs based on the previously identified DML, using the search algorithm which is described in the next section.

### 2.5.1 Novelties

One of the biggest innovations in Dimmer 2.0 is the support of bisulfite sequencing data. The user can upload a Bismark output directly in the tool. Additionally, Dimmer 2.0 can now work with already processed methylation data thanks to the possibility of uploading a $\beta$-value matrix that can be previously customized, from either 450k or EPIC array. Idat files from different array versions do not necessarily have the same set of loci. In these cases, only common loci across samples are analyzed while the additional ones are ignored. In this way, idat files from different EPIC array versions are now supported. Quality control and pre-processing correspond to the one in the first released version of the tool, with a difference in setting the preprocessing step as optional. Furthermore, there is now the chance to export the normalized and processed data table. To further narrow down the DMR discovery, we introduced an additional criterion where the user can decide the minimum mean methylation difference a candidate region must have. We introduce a new approach for computing individual p-values for every DMR. In the old approach, each region had an FDR p-value but individual DMRs composition was not considered, ending up with the same FDR for DMRs of equal size. The new approach works by sampling from the genome a set of random regions (100.000 by default) of the same size and the same number of loci. The p-values of such loci, computed in the DML search step, are combined using Fisher's method [202]. The method returns a score for each of the random regions and the set of scores is used as a background distribution to obtain a p-value for every DMR identified.

### 2.5.2 Corrections

Besides the extensions, we fixed some logical bugs and significantly improved some existing features. We debugged the paired-samples data analysis, where two consecutive samples from the sample annotation file were considered pairs. While the original version of Dimmer provided FDR correction for loci p-values, the correction was not implemented. We fixed this issue and now FDR correction is correctly executed. DMR search speed performance was optimized. The first version of Dimmer used to count the amount of DML and exceptions every time the sliding window was shifted, consuming a significant amount of time. Now the counts are memorized so that exceptions don't have to be counted again, reducing the runtime considerably, especially in cases of abundant DML. Besides these major adjustments, we further fixed minor bugs and applied small backend improvements.

### 2.5.3 The detection algorithm

The method used by Dimmer 2.0, already introduced with the first Dimmer release [172], is based on a sliding genomic window.

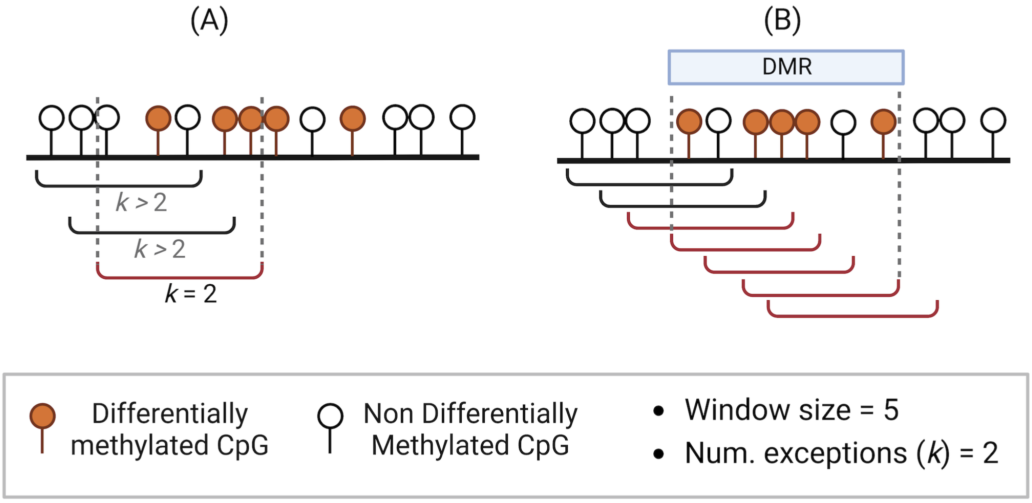

(A)
(B)
DMR
*k* > 2
*k* > 2
*k* = 2
Differentially methylated CpG
Non Differentially Methylated CpG
Window size = 5
Num. exceptions (*k*) = 2

**Figure 2.7** Graphical representation of the DMRs search algorithm in Dimmer 2.0. (A) Once CpGs are classified as differentially/non-differentially methylated, the algorithm places a genomic window of fixed size on the genome and checks whether in the window there are at most *k* non-differentially methylated loci. (B) The window is used to scan the genome as long as criteria are respected. The overlap of the regions that satisfy the criteria will be a potential DMR, trimming the starting and/or ending CpGs in case they are non-differentially methylated.

The algorithm starts by taking a window of fixed size, decided by the user, and moving this window along the genome until a region with at most *k* exceptions is found. An exception is defined as a non-differentially methylated position where a position to be significantly differentially methylated has an associated p-value lower than a significant threshold. This means that if the window size is of five loci and the number of exceptions is two, the candidate region must contain three DML

(Figure 2.7 (A)). Moreover, loci have to be distant at most $n$ base pairs, with $n$ defined by the user. The window is moved forward until requirements are not satisfied anymore. Overlapping windows that satisfy the criteria are considered DMR (Figure 2.7 (B)). Additional criteria to establish whether a position is considered significant in the DMR search is the minimum methylation difference desired. Dimmer 2.0 controls for region width by asking the user the maximum distance allowed between CpGs when scanning the genome (Figure 2.6, (4)). The last step is a permutation test to assess the statistical significance of the candidate DMRs. The labels significantly/non-significantly differentially methylated at the CpG level are randomized and the DMR research is computed again. The goal is to assess the chances of discovering the same DMRs on permuted data. Dimmer 2.0 is available as a command-line tool and graphic user interface (GUI) at https://github.com/baumbachlab/Dimmer/releases/tag/2.2. GUI is only available for array-based data, hence BS-Seq data can be only analyzed using the command-line version of the tool.

### 2.5.4 Implementation on two study cases

Dimmer 2.0 is used two study cases, the first is primary PDAC *versus* metastasis and the second is meningioma benign *versus* malignant samples. In both cases, Dimmer 2.0 is performed with the implementation of a t-test and empirical p-values are considered. All default parameters are used, except for the required minimum mean difference at the loci level, set to 0.1. PDAC *versus* metastasis is analyzed in a paired setting because of the primary samples and the matching metastasis coming from the same patient, differently from the meningioma case where the analysis is unpaired.

## 2.6 Methods for DMRs evaluation

The DMRs obtained are investigated with the implementation of different techniques to retrieve biological insight.

EpiRegio [203] is a repository of regulatory elements (REMs) and it looks for non-coding DNA regions that overlap with a set of genomic regions of interest. Epiregio additionally returns a set of genes associated with those matching REMs. We executed Epiregio using the region query option on the web server application with a 50% REMs overlap percentage. Functional analysis on the regions was performed with the g:GOSt function of the online tool g:Profiler (version $e107_eg54_p17_bf42210$) [204].

The biological interpretation of the regions is additionally carried out with the use of GREAT (Genomic Regions Enrichment of Annotations Tool) [205], a web application for identifying genes in the proximity of a set of imputed regions. GREAT is particularly designed for working on cis-regulatory regions and aims to elucidate the mechanisms involved in the regions by performing enrichment analysis on the neighbor genes. The region-genes association is done by taking the gene regulatory domains that overlap the regions of interest. The regions-genes association is performed under GREAT default settings, meaning a gene is attributed to each regulatory domain that is located at a distance of a minimum of 5 kb upstream and 1 kb downstream of the transcription starting site (TSS). The regulatory domain is further expanded both up- and downstream up to 1000 kb to the closest gene. Only statistically significant terms are shown, where a term is considered significant when both the binomial test performed over the regions and the hypergeometric test performed over the genes have FDR adjusted p-value < 0.05. GREAT version used is 4.0.4.

A gene's promoter or UTR region that overlaps with DMRs, suggests the possibility of an alteration in the gene transcription and expression, as a consequence of the presence of an epigenetic alteration. We additionally focus on the analysis of such genes, identified in R 4.1.0 with the use of the function runseq2gene, part of seq2pathway 1.30.0 [206]. Only coding genes are considered.

Gene enrichment analysis of the genes obtained is performed in Python using the GSEApy 0.10.5 [207] library. Only statistically significant terms with adjusted p-value lower than a significant threshold of 0.05 are shown.

## 2.7 Method for the delineation of a prognostic-relevant TB cut-off

Survival relevant cut-offs for the morphologic traits described in section 1.6 were identified. The approach adopted is based on splitting the samples in two group and assessing the difference of the two survival curves with the implementation of the Log-Rank test. This step is repeated for each value and the one that registered the lowest p-value is taken as cut-off. Whether cut-offs are robust or not was assessed with Kaplan-Meier survival analysis after shuffling time points and events. Multivariate survival analysis was conducted with Cox-regression model including age, sex and AJCC stage as confounding factors.

## 2.8 Computational analysis of the association of TB with molecular data

The identification of features significantly associated with TB was carried out in two ways, distinguishing somatic mutation (binary data type 0/1 where 1=mutation and 0 otherwise) from the other three data types, all continuous data.

The link between gene mutations and TB was assessed by computing Welch's t-test taking one gene at a time. We collect the p-values returned by each test and adjust them for multiple testing (FDR). Genes with corrected p-value < 0.05 are considered significant. The test was implemented in SciPy 1.5.2, in Python.

A univariate linear regression is used to test each gene from the gene expression, protein from protein expression and CpGs from the DNA methylation data table.

The implementation of the regression model is preceded by an additional filtering step. Molecular variables are filtered based on their variability, measured in median absolute deviation (MAD). In the case of gene expression, only the 5000 genes with the highest MAD were kept for analysis. In the case of DNA methylation, only the 10.000 CpGs with the highest MAD were used. The protein expression matrix contains 131 proteins and all of them were kept. As for Welch's t-test on mutational data, we collected the p-values retrieved by the multiple univariate linear regression and applied FDR correction to identify the features with significant p-value (<0.05).

Before performing the regression model, TB was normalized by converting values to a logarithmic scale. The implementation of linear regression was done in Python with the *statsmodel* module (0.12.2).

Significant CpGs were mapped to genes. We considered the genes whose promoter region was mapped with significant CpGs, assuming TSS200 and TSS1500 as gene promoter regions. The mapping was done using the IlluminaHumanMethylation450kanno.ilmn12.hg19 package of R, version 0.6.0.

Correlation between significant molecular features is measured with Pearson correlation and Point Biserial correlation, both implemented using SciPy in Python, version 1.5.2.

### 2.8.1 Pathway enrichment analysis

Enrichment analysis was done using WikiPathway [208], BioPlanet [209], KEGG [197] and Reactome [198] pathways repositories. Results non statistically significant (adjusted p-value < 0.05) were filtered out. Analysis was carried out in Python with the use of the Gseapy 0.10.8 library, while the enrichment dot plot was obtained using the two R packages enrichplot 1.12.3 and multienrichjam 0.0.55.900.

### 2.8.2 Association with published lung-SCC molecular subtypes

TB activity was assessed in relation to different lung-SCC subtypes. We consider the four molecular subtypes proposed by Wilkerson *et al.* [181], obtained via consensus clustering on mRNA expression from a set of 382 squamous cell carcinoma samples. The four subtypes are primitive, classical, secretory, and basal. A study carried out by TCGA classified the samples from the TCGA-LUSC project into the four subtypes [153]. We use these predicted labels to test whether TB is characteristic of different subtypes or not. ANOVA was performed to assess the association between TB and the subtypes, further compared in a one vs all approach with a t-test. Python SciPy 1.5.2 was used for the two tests and a significance threshold of 0.05 was considered.

# 3 Results

## 3.1 Transcriptome subtypes in PDAC are inconsistent

All the methods used in this section are described in section 2.4. The analyses were conducted in Python 3.7.3 and R 4.1.0 and the code is stored in a GitHub repository at https://github.com/biomedbigdata/PDAC-molecular-classifier-validation.

The first hypothesis we want to test is that, taking a gene signature, the expression distribution of each gene appears different between subtypes, where values are higher in the subtype the gene is representing and lower otherwise. Figure 3.1, shows Collisson subtypes as the classes most similar in expression, no matter what genes are used. On the other hand, Puleo subtypes appear to be extremely different for every signature used. However, the test only tells if there is a difference without giving any further information. For this reason, we perform additional analysis on Puleo subtypes, this time comparing them in pairs to see if the high difference was mostly driven by one subtype. The pairwise comparison output shows one subtype standing up respectfully to the others, which is the Pure Classical subtype, followed by Desmoplastic, as shown in Supplementary Figure 5.3.

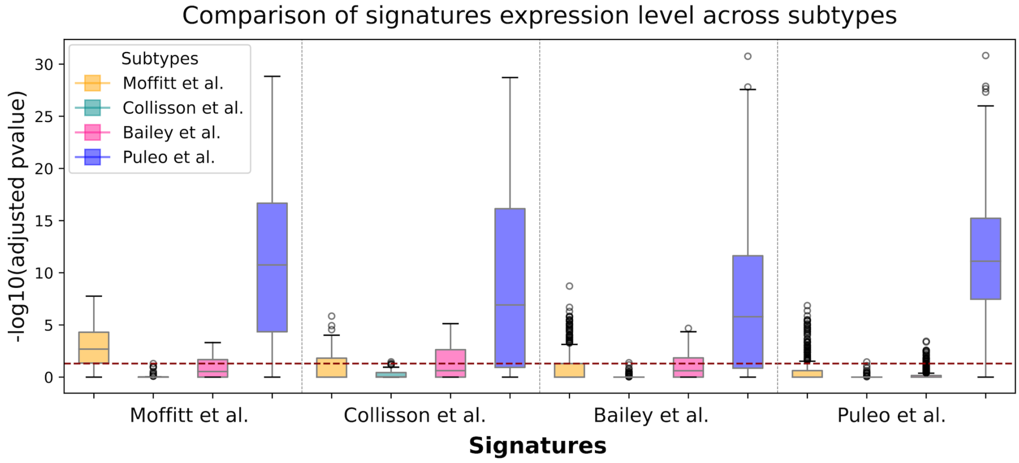


**Figure 3.1** Comparison of the expression distribution of the subtypes, for each signature. Each gene of the signature is used to isolate the expression distribution for every subtype and perform the Wilcoxon rank sum test for two groups comparison and the Kruskal-Wallis test otherwise. The distributions of Bonferroni-corrected $-log_{10}$ of the p-values are shown. The significance threshold for p-value < 0.05 is represented by a horizontal dashed line. Figure from [113].

### 3.1.1 Supervised and unsupervised analysis reveal poor subtypes reproducibility

#### Unsupervised analysis

After observing if the genes in the signatures have different values for different subtypes, we continue testing the signature's ability in clustering samples of nine cohorts. As we did in the previous section, where we compared the expression distribution of subtypes to assess their difference, we do the same here, comparing the clusters taking one gene at a time and showing the adjusted p-value distribution below every heatmap (Figure 3.2).

The clusters obtained on the dataset used for the signature discovery accurately match the true subtypes in the case of Moffitt and Puleo (Supplementary Figure 5.4 A.1 and A.4), as indicated by RI and the distribution of significant p-values. However, this is not always the case. For Collisson and Bailey (Supplementary Figure 5.4 A.2 and A.3), the distribution of p-values hardly reaches the level of significance while the overlap among clusters is high. Similar result variability is observed when doing clustering on independent data, showing either a good classes overlap or a good adjusted p-values distribution. A different case is where there is both a good overlap and significance in the distribution of p-values as we can see, for instance when using the Bailey signature to cluster the Puleo dataset (Figure , D.3). The clusters are compared to the molecular gradient PAMG [123], designed to predict patients' prognosis. The subtypes we are investigating have different prognoses and, in some cases, like the dichotomous Basal and Classical from Moffitt, extreme. With this assumption, even though PAMG is represented as a continuum, we expect its scores to match the expected survival associated with the different subtypes, which are low scores for low survival and high otherwise. We found samples with lower values of PAMG grouped in clusters that correspond to the worst-prognosis subtypes (basal, squamous and quasi-mesenchymal), and the opposite subtypes, the classical-like subtypes, are represented by high PAMG scores.

#### Supervised analysis

Going further with the analysis, we meet the constraint of having subtype calls for a limited number of datasets. To overcome this issue, we classify PDAC subtype labels for each dataset, with prediction models built combining the four subtype schemes with the four signatures. The total number of models is sixteen and their construction is explained in Figure 2.5 under the 2.4.4 method section. The models are applied to the nine datasets and the predicted classes are compared with the real labels, when available, to see if there is an agreement (Supplementary Figure 5.5). Interestingly, we found that the predicted labels are made of a mixture of the real labels. An example is the case of the binary classification proposed by Moffitt. Basal samples are expected to be predicted as basal-like subtypes (Quasi-Mesenchymal, Squamous and Pure Basal/Stroma Activated/Desmoplastic according to Collisson, Bailey and Puleo schemes), but what we observe is that labels predicted for these samples are mixed basal and classical (Supplementary Figure 5.5).

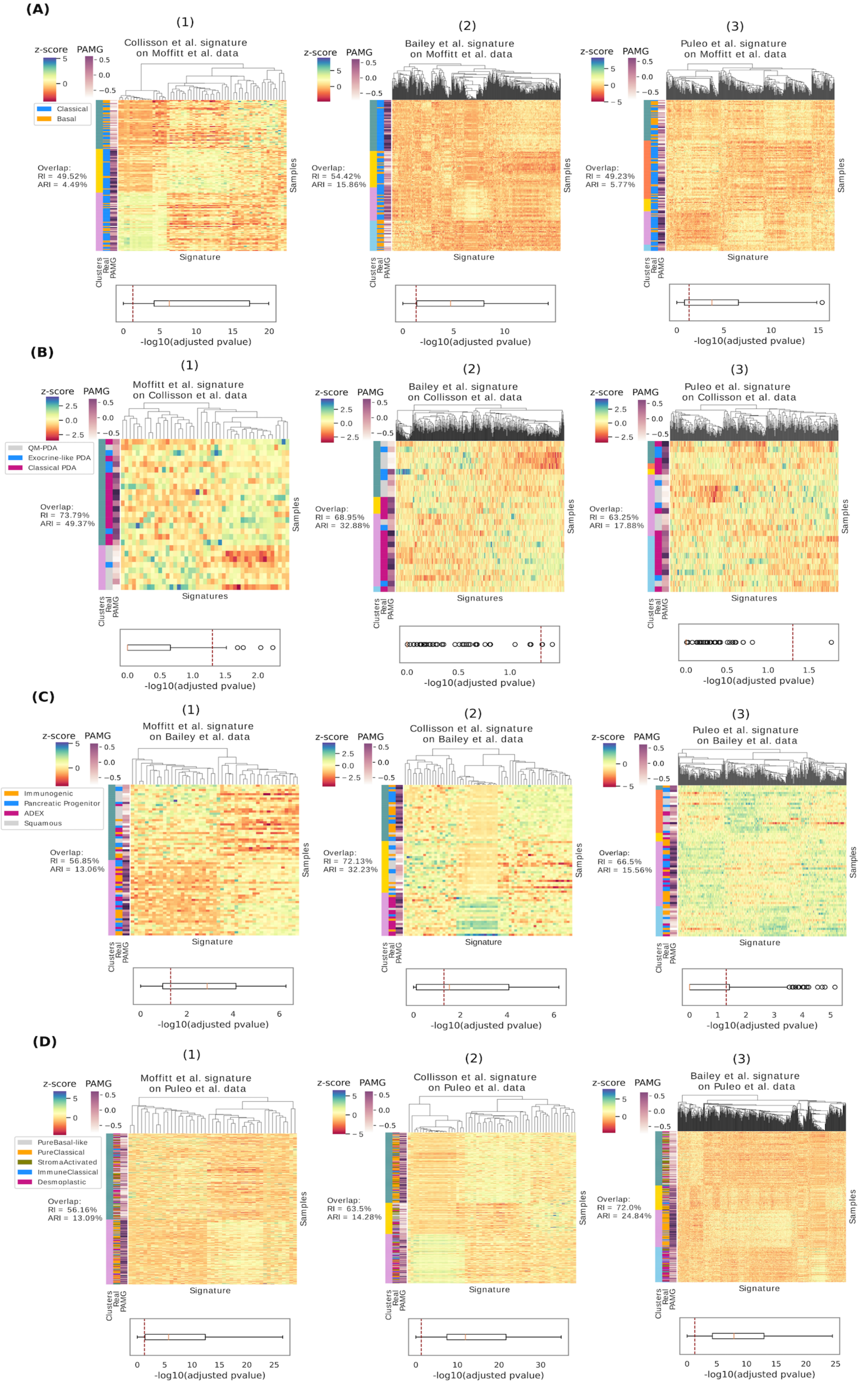

(A)
(1)
Collisson et al. signature
on Moffitt et al. data
z-score
PAMG
Classical
Basal
Overlap:
RI = 49.52%
ARI = 4.49%
Signature
Samples
Clusters
Real
PAMG
-log10(adjusted pvalue)
(2)
Bailey et al. signature
on Moffitt et al. data
Overlap:
RI = 54.42%
ARI = 15.86%
(3)
Puleo et al. signature
on Moffitt et al. data
Overlap:
RI = 49.23%
ARI = 5.77%
(B)
(1)
Moffitt et al. signature
on Collisson et al. data
QM-PDA
Exocrine-like PDA
Classical PDA
Overlap:
RI = 73.79%
ARI = 49.37%
Signatures
(2)
Bailey et al. signature
on Collisson et al. data
Overlap:
RI = 68.95%
ARI = 32.88%
(3)
Puleo et al. signature
on Collisson et al. data
Overlap:
RI = 63.25%
ARI = 17.88%
(C)
(1)
Moffitt et al. signature
on Bailey et al. data
Immunogenic
Pancreatic Progenitor
ADEX
Squamous
Overlap:
RI = 56.85%
ARI = 13.06%
(2)
Collisson et al. signature
on Bailey et al. data
Overlap:
RI = 72.13%
ARI = 32.23%
(3)
Puleo et al. signature
on Bailey et al. data
Overlap:
RI = 66.5%
ARI = 15.56%
(D)
(1)
Moffitt et al. signature
on Puleo et al. data
PureBasal-like
PureClassical
StromaActivated
ImmuneClassical
Desmoplastic
Overlap:
RI = 56.16%
ARI = 13.09%
(2)
Collisson et al. signature
on Puleo et al. data
Overlap:
RI = 63.5%
ARI = 14.28%
(3)
Bailey et al. signature
on Puleo et al. data
Overlap:
RI = 72.0%
ARI = 24.84%

**Figure 3.2** Results from hierarchical clustering. The four datasets from Moffitt (A), Collisson (B), Bailey (C) and Puleo (D) are clustered keeping only the genes in the signatures. Below the heatmaps is placed a boxplot with the $-log_{10}$ adjusted p-values distribution, where the p-values are obtained from the Wilcoxon rank sum and Kruskal-Wallis test, computed to compare two clusters or more, respectively. Overlap between clusters and real subtypes is indicated with RI and ARI and shown in percentage on the left of every heatmap. A color bar with PAMG scores is included. Figure from [113].

The possible explanation is that the models are not suitable to reproduce the subtypes. The clustering results already showed how subtypes lack reliability, and what is observed in Supplementary Figure 5.5 might be the consequence of a poor consensus between the subtypes proposed by the four studies. We then executed the classification models 1000 times and saved their accuracy from 5-fold cross-validation. The average accuracy is shown in Figure 3.3. What appears from the performance result is that the signatures from Moffitt and Puleo perform the worst in predicting Bailey classes. More generally speaking, Moffitt's signature shows the worst accuracy. On the other hand, Moffitt subtypes seem to always be correctly predicted, possibly helped by the fact that a binary classification is less complex to reproduce.

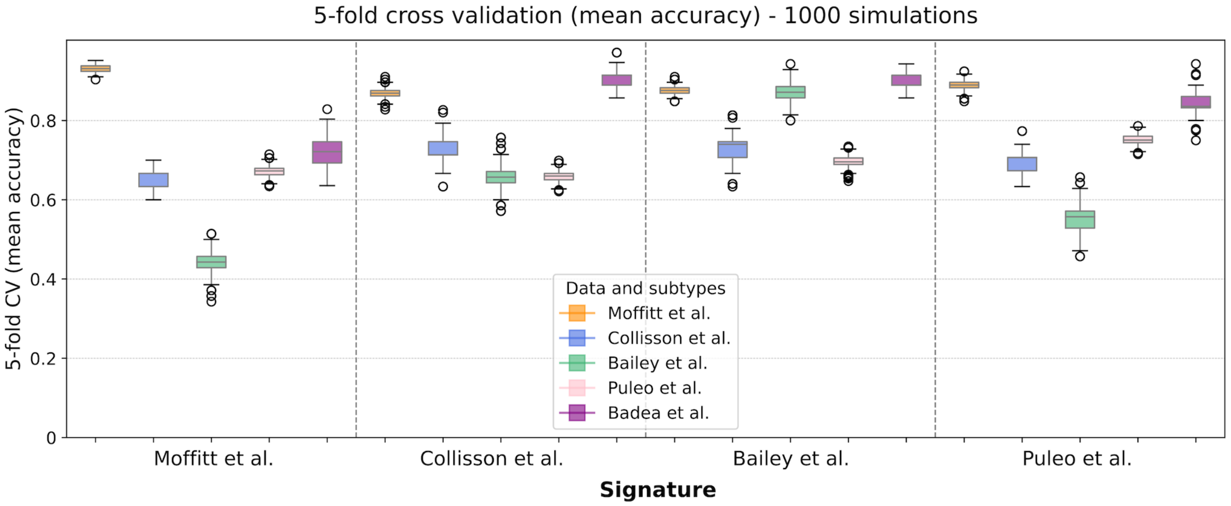


**Figure 3.3** 5-fold cross-validation mean accuracy from 1000 repeated classifications. Results are divided by signatures, where each signature is used to predict the four subtype schemes. Badea *et al.* [82] samples were classified and made available by Collisson, thus included in this analysis. Figure from [113].

### 3.1.2 Signatures show lack of robustness

A signature is reliable for the subtyping of PDAC when not only it can predict the classes it was built for but also when, compared to other sets of genes, is the one that can do it best. This means that between the subtypes and the genes representing them, there is a meaningful connection. Numerous studies have already discovered how signatures can be equal to or even less performant than using random features [106]. We evaluate the signatures robustness by observing how a set of random genes behave when used for the same purpose. We assess the signatures in both clustering and classification settings [82, 210].

We perform the same procedure of using the signatures to cluster the samples of the nine datasets, this time with a set of random genes of the same size as the signature. To assess the similarity between the clusters obtained in the first step and the second one, we compute ARI. We additionally compute pairwise ARI only between random genes-derived clusters. The similarity between real clusters and randomly obtained clusters appears very low, close to zero, as shown in Supplementary Figure 5.6. ARI distributions shown in Supplementary Figure 5.6 are summed up in Figure 3.4 by computing SSMD between the ARI of real vs randomly derived clusters and the pairwise ARI of random vs random. From these results, we found a dataset-related difference. However, comparing the signatures we can see an increase in the SSMD values as increasing the size of the signatures, ending with the Puleo signature whose clusters are the ones diverging the most from the clusters generated by a random signature.

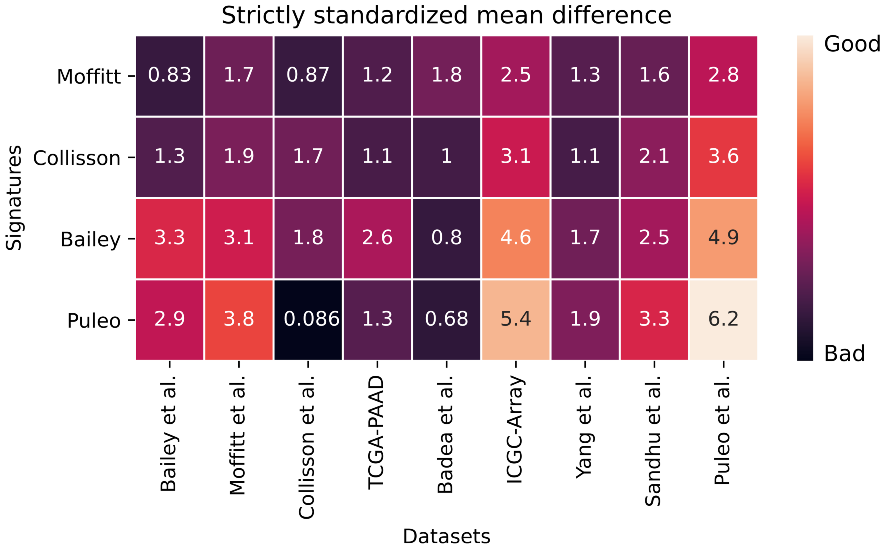


**Figure 3.4** SSMD to compare the ARI computed between the true signatures clusters and the random signatures clusters and the ARI computed taking the random signatures clusters in pairs. The nine datasets are aligned on the x-axis while the four signatures are placed on the rows in ascending order, from the smallest to the biggest. Values, expressed in absolute values, indicated the good and bad robustness of the signatures. Figure from [113].

The same idea was applied when evaluating how robust the signatures are when used as predictors. We took the mean 5-fold cross-validation accuracy obtained from the prediction models executed 1000 times on the data with known subtypes, and repeated the same but with two modifications. A first

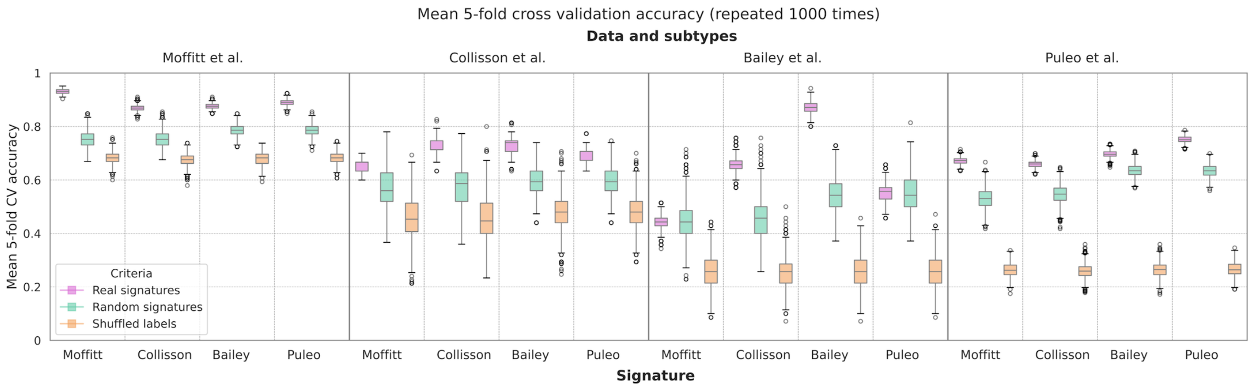


**Figure 3.5** Distribution of mean 5-fold cross-validation accuracy obtained by executing 1000 times the three types of prediction models: the first trained using true signatures, the second trained using random genes, and a third trained on randomized subtype labels keeping the true signatures. Figure from [113].

**Table 3.1** The maximum accuracy reached by the model trained on true signature (a), random genes (b), and shuffled labels (c). The highest values among the three models are expressed in bold. Cases, where random genes outperformed signatures or returned the same average accuracy, are indicated with (*), (**), (***), (****). Table from [113].

| Data and subtypes: | Moffitt | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Signature: | Moffitt | | | Collisson | | | Bailey | | | Puleo | | |
| | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) |
| | 0.95 | 0.84 | 0.75 | 0.91 | 0.85 | 0.73 | 0.91 | 0.84 | 0.73 | 0.92 | 0.85 | 0.74 |
| Data and subtypes: | Collisson | | | | | | | | | | | |
| Signature: | Moffitt (***) | | | Collisson | | | Bailey | | | Puleo | | |
| | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) |
| | 0.7 | 0.78 | 0.69 | 0.82 | 0.77 | 0.8 | 0.81 | 0.74 | 0.7 | 0.77 | 0.74 | 0.7 |
| Data and subtypes: | Bailey | | | | | | | | | | | |
| Signature: | Moffitt (*) | | | Collisson (****) | | | Bailey | | | Puleo (**) | | |
| | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) |
| | 0.51 | 0.71 | 0.44 | 0.75 | 0.75 | 0.5 | 0.94 | 0.72 | 0.45 | 0.65 | 0.81 | 0.47 |
| Data and subtypes: | Puleo | | | | | | | | | | | |
| Signature: | Moffitt | | | Collisson | | | Bailey | | | Puleo | | |
| | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) | (a) | (b) | (c) |
| | 0.71 | 0.66 | 0.33 | 0.69 | 0.64 | 0.35 | 0.73 | 0.7 | 0.35 | 0.78 | 0.69 | 0.34 |

modification trains the model on a set of random genes and a second one is using the real signatures but shuffling the class labels. Results are shown in Figure 3.5 where the distribution of mean accuracy is compared between the three types of models. Accuracy from the random genes and the model with shuffled labels reaches values that are comparable to the ones returned by the signature-based models, and in some cases even higher.

More specifically, we collected the highest values registered in the 1000 runs, available in Table 3.1. The prediction of Bailey subtypes was the most critical one, indeed, three out of the four examined signatures performed worse (Table 3.1 (*), (**)) or equal (Table 3.1 (****)) to random genes. The same outcome is observed when using the Moffitt signature for the prediction of Collisson subtypes (Table 3.1 (***)).

### 3.1.3 Cell deconvolution confirms normal tissue contamination

After focusing on the evaluation of the signatures' performance and robustness, we now move the focus to the meaning of the subtypes predicted by the sixteen models. Even though the subtypes predicted do not correspond to the clinically established ones, they might still carry biological meaning. The isolation of pancreas-specific cells will help investigate their role in the different PDAC subgroups. For each of the nine datasets, single-cell deconvolution estimated the enrichment of each cell in every sample, as shown in Supplementary Figure 5.7. A first overview suggests that results are conditional to the dataset used. We found the ADEX and Exocrine-like subtypes from Bailey and Collisson to be enriched in acinar cells, corroborating the hypothesis on the presence of healthy tissue in the samples used by the authors (results from one of the nine cohort in Figure 3.6). Immune cell infiltration, macrophages in particular, was observed in the Immunogenic predicted samples in Badea *et al.*, Yang *et al.* and ICGC-Array, more specifically when Bailey signature is used for prediction. Additionally, macrophages are abundant in the Desmoplastic subtype together with a high presence of acinar cells. Ductal cell types were found particularly expressed in classical-like subtypes. After examining the cellular composition we want to investigate which are the active pathways and the functionalities linked to the subtypes. Results are collected in the Supplementary Figure 5.8 while one representative result is Figure 3.7. For sake of simplicity, we only plot the first 30 terms found significantly associated with most of the datasets. Mostly enriched activities were found related to fatty acid and lipid metabolism and metabolic pathways in general. For instance, metabolic activity is downregulated in the Basal subtype, visible in 3.7. Samples predicted to be Exocrine-like were expressed by cytochrome P450, where the characteristic of being drug-resistant was already discovered by Noll *et al.* [211].

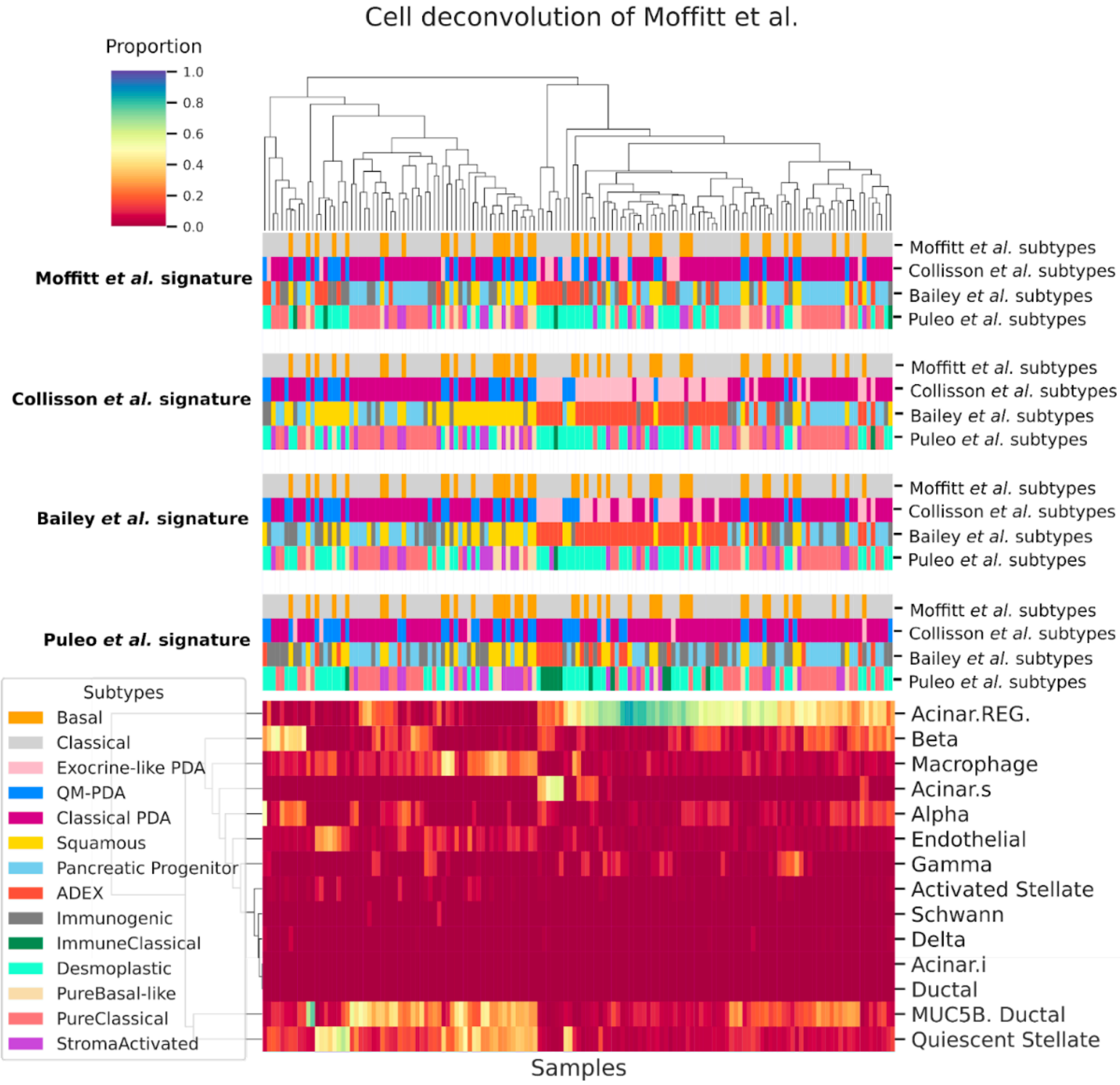


**Figure 3.6** Cell deconvolution of the dataset used by Moffitt *et al*. Samples on the column and cell types on the rows, where the values in the heatmap represent the proportion of each cell type in the samples. The top color bars indicate the labels predicted by the 16 models. Figure from [113].

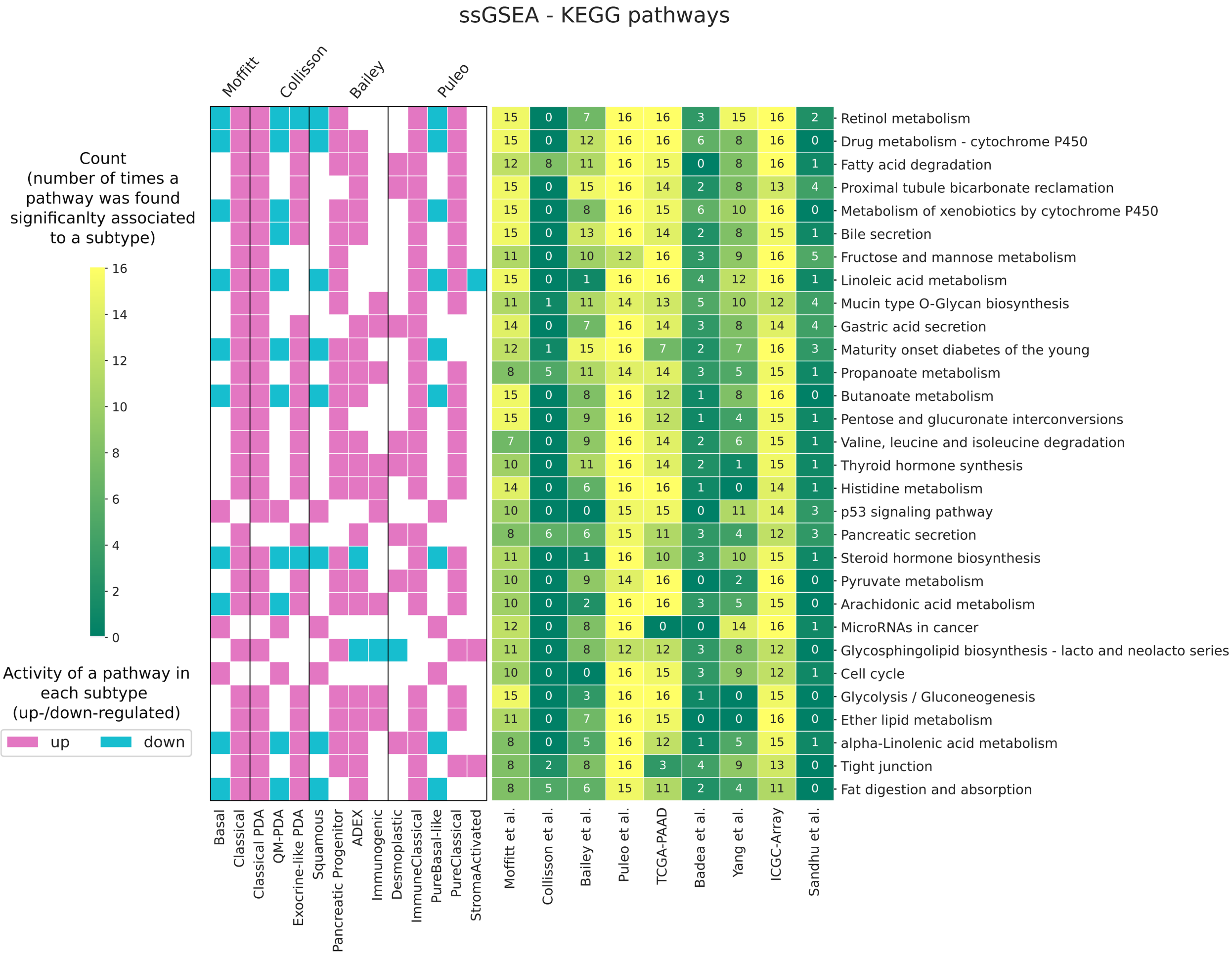


**Figure 3.7** KEGG pathways that are mostly associated with the predicted subtypes. Terms are sorted in descending order, starting from the one found significantly enriched in the highest number of predictions. Only the top 30 terms are shown. Table cells are colored according to the number of times a term was significantly associated with the subtypes, going from zero to sixteen given the sixteen predictive models. The vertical color bars indicate in which subtype the term is mostly enriched and if it's up-/down-regulated. Figure from [113].

### 3.1.4 Basal/classical as the most prognostically-relevant subtypes

We want to evaluate whether the classes we predicted with the use of the 16 models not only differ in survival but also if their difference reflects what the authors observed with their analysis. We took the seven datasets with available survival information and performed survival analysis. A pairwise log-rank test was used on the predicted subtypes and $-log_{10}$ of the p-values are shown in Figure 3.8. A first observation is that Classical-Basal survival was found significantly different (a), as claimed by the own authors. The main observation from the other subtypes is the predominant difference between basal-like and classical-like subtypes, particularly visible in Figure 3.8 (b) and (d.2).

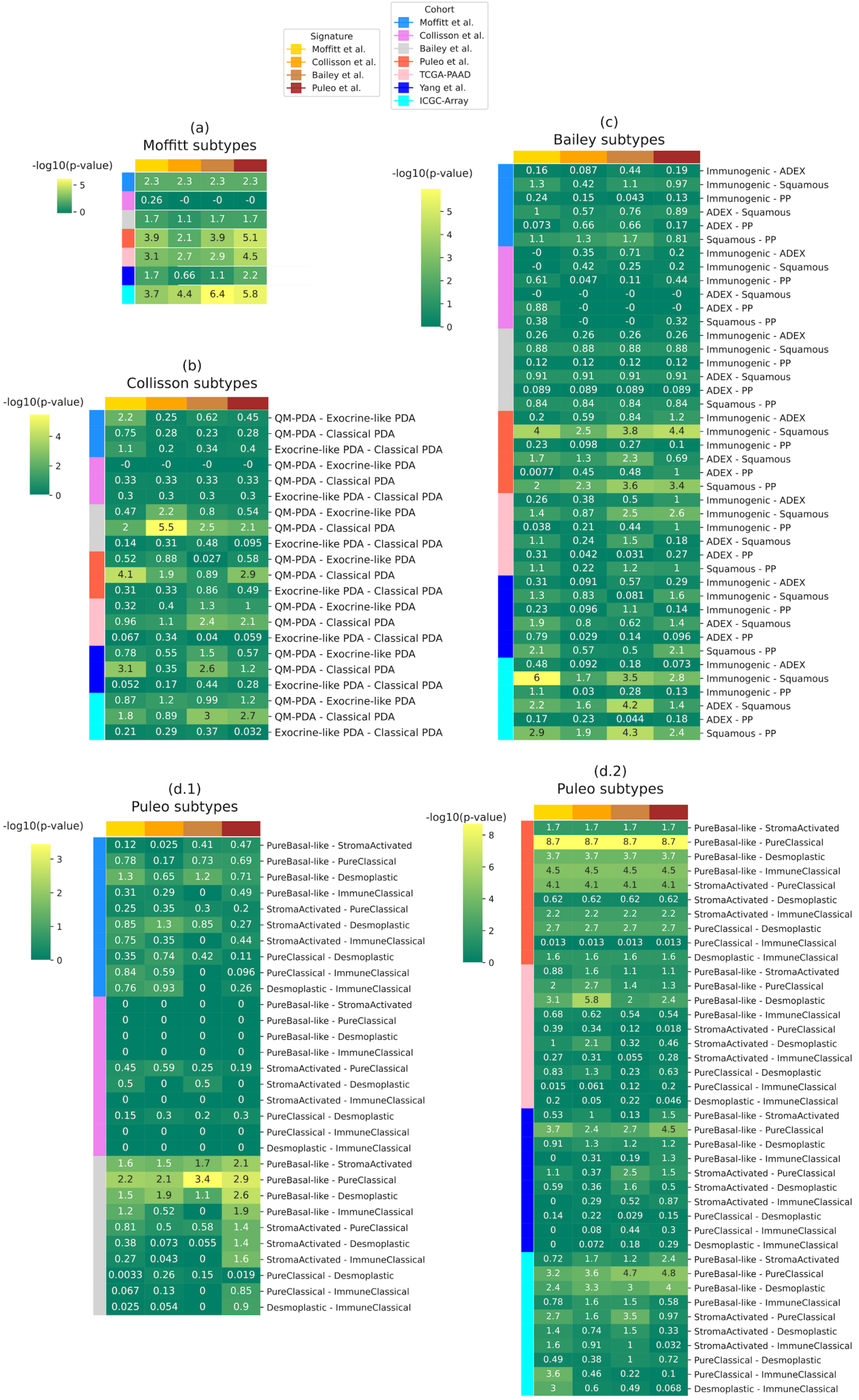


**Figure 3.8** Survival analysis to compare survival curves of predicted subtypes. Analysis conducted with a pairwise log-rank test and $-log_{10}$ of the p-value is stored in heatmaps divided by subtype schemes: (a) Moffitt, (b) Collisson, (c) Bailey and (d.1) and (d.2) Puleo. Results are divided by datasets, placed on the rows, and signatures, on the columns. Figure from [113].

## 3.2 De novo discovery of differentially methylated regions in PDAC and meningioma

### 3.2.1 Results on primary PDAC and matched liver metastasis

We used Dimmer 2.0 to investigate the molecular conditions that trigger PDAC progression and aggressivity. We compared the methylation status of 11 PDAC samples with their corresponding liver metastasis and identified 708 statistically significant DMRs. The regions found significant are long, on average, 573 base pairs (Figure 3.9, left) and contain approximately 11 significant CpGs, with an average CpG density of 0.02 (Figure 3.9, right). One of the regions detected is located in chr16:56 641 998-56 642 130, shown in Figure 3.10. Enrichment analysis revealed a small cluster of proteins linked to zinc homeostasis, as shown in Table 3.2. Zinc plays a fundamental role in cellular functions and zinc dysregulation may affect the immune reaction, the organism's cellular life balance, and endocrine functions. Zinc deficiency is also linked to gastrointestinal diseases and particularly in the pancreas [212]. To assign biological meaning to the DMRs returned, we implement functional enrichment analysis by following three approaches useful for genomic regions interpretation. We first implemented EpiRegio and found 146 REMs with 144 associated genes. Genes found significantly enriched in the zinc regulation pathways are MT1A, MT1G and MT1H, which belong to the metallothioneins (MT) protein family, and SLC39A7, a protein coding gene with zinc transport function. Zinc in the pancreas plays multiple tasks such as insulin regulation and engagement in glucagon secretion. The role of MT in the context of cancer metastasis is known to be linked to drug resistance and tumor progression

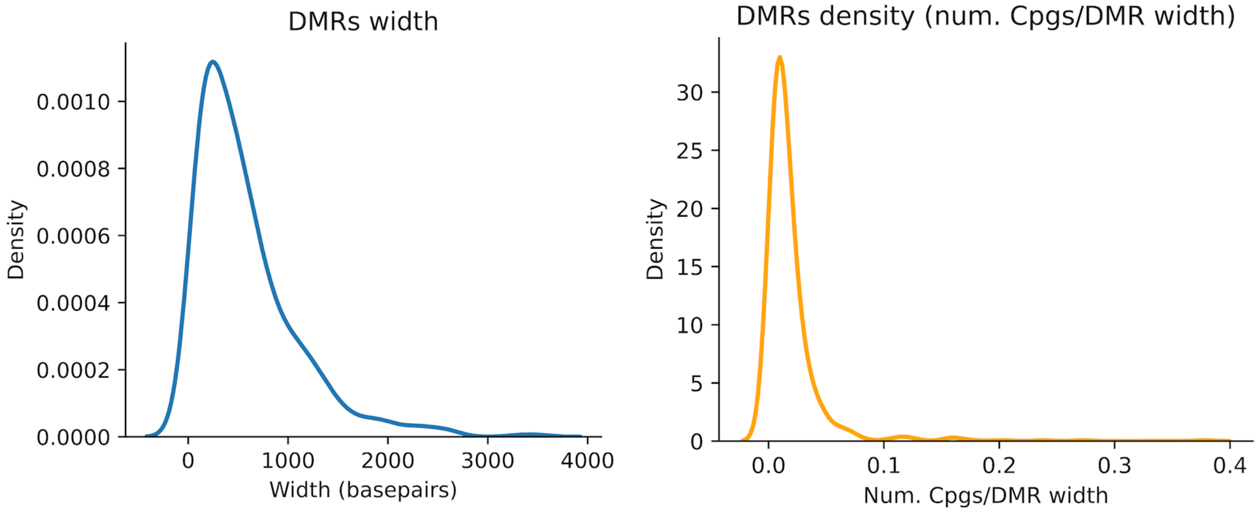


**Figure 3.9** Distribution of the width of the 708 DMRs returned by Dimmer 2.0 (left) measured in base pairs and distribution of the region's density (right) measured as the amount of CpGs found significant inside a region.

[213]. An example is the case of MT1G, known to have a tumor suppressor and metastasis prevention role.

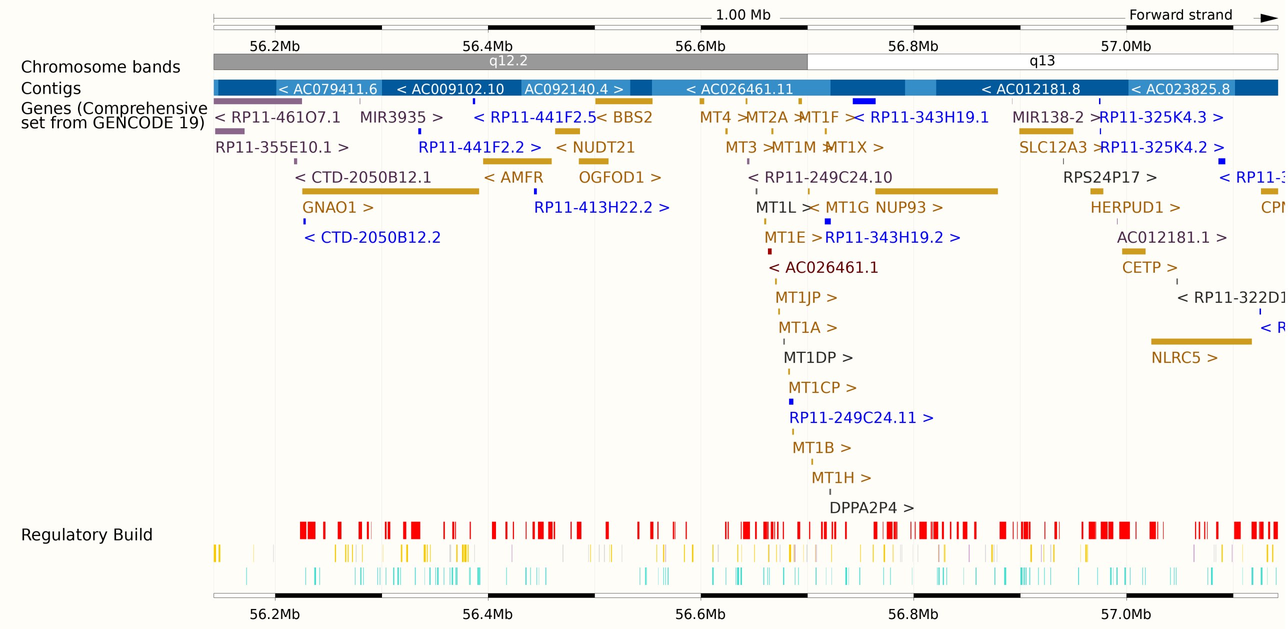


**Figure 3.10** Genomic view of one of the DMR identified by Dimmer 2.0 from comparing primary PDAC *versus* their matching liver metastasis. Region located on chromosome 16, starting at 56 641 998 and ending at 56 642 130. The view was obtained through the Ensemble Genome Browser [53].

**Table 3.2** Enrichment analysis of the 144 genes associated with the REMs identified by EpiRegio. Only significant terms with adjusted p-value < 0.05 are shown. Gene set repositories that registered significant terms are WikiPathway, Reactome and Biological Processes (BP) from Gene Ontology terms.

| **Repository** | **Term** | **Overlap** | **Adjusted p-value** | **Genes** |
|---|---|---|---|---|
| WikiPathway | Zinc homeostasis WP3529 | 4/36 | 0.011 | MT1A, MT1G, MT1H, SLC39A7 |
| Reactome | Response to metal ions R-HSA-5660526 | 3/11 | 0.008 | MT1A, MT1G, MT1H |
| Reactome | Metallothioneins bind metals R-HSA-5661231 | 3/11 | 0.008 | MT1A, MT1G, MT1H |
| GO BP | cellular zinc ion homeostasis (GO:0006882) | 4/28 | 0.021 | MT1A, MT1G, MT1H, SLC39A7 |
| GO BP | zinc ion homeostasis (GO:0055069) | 4/30 | 0.021 | MT1A, MT1G, MT1H, SLC39A7 |

When the gene promoter is hypermethylated, the expression of MT1G decreases, resulting in PDAC cells resistant to chemotherapy [214]. The relationship between the downregulation of MT1G and cancer progression was also observed in hepatocellular carcinoma and it was linked to the activation of cancer pathways like nuclear factor $\kappa$B (NF-$\kappa$B) pathway, a family of transcription factors involved in immune response regulation and B cells apoptosis prevention [215]. The NF-$\kappa$B transcription factor is also proved to be linked with the transcription of genes that regulate cell proliferation promotion [216]. High expression levels of SLC39A7 were found linked to poor prognosis and metastasis spread in the case of colorectal and breast cancer [217, 218]. We continue investigating the biological meaning of the DMRs by looking at the genes whose TSS is associated with them. The GREAT tool identified 928 associated genes and returned significant molecular functions and biological processes performed by these genes, thanks to GO enrichment analysis, reported in Table 3.3.

**Table 3.3** GREAT output showing GO enrichment analysis on the genes with TSS nearby or matching the DMRs retrieved comparing PDAC and liver metastasis. Statistically significant Biological Processes (BP) and Molecular Function (MF) terms with FDR adjusted p-value < 0.05 in both binomial and hypergeometric tests.

| **Repository** | **Term** | **Binomial FDR p-value** | **Hypergeometric FDR p-value** |
|---|---|---|---|
| GO BP | Regulation of toll-like receptor signaling pathway | 0.041 | 0.008 |
| GO BP | Regulation of transcription involved in cell fate commitment | 0.043 | 7.759e-05 |
| GO MF | Transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding | 1.972e-06 | 1.268e-10 |
| GO MF | Transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding | 9.628e-07 | 1.71e-09 |
| GO MF | Transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding | 1.122e-06 | 2.233e-11 |

Interesting is the enriched regulation of toll-like receptor (TLR) signaling pathways, responsible for the innate immune response when a pathogen is detected. The activation of TLRs leads to the activation of the NF-$\kappa$B pathway [219]. While GREAT enrichment analysis exclusively focuses on GO terms, we additionally performed functional pathway analysis on the associated genes returned by GREAT. Results are shown in Table 3.4. We found enrichment in the pathway involved in the regulation of gene expression in early pancreatic precursor cells. Engaged in the pathway is NR5A2, a gene whose overexpression has been found linked to PDAC cell mobility and infiltration [220]. Because of the negative correlation between gene activity and DNA methylation status, where altered methylation in gene promoters can influence the binding of the transcription factor, we want to observe whether the DMRs are located in proximity or coincide with gene promoters.

**Table 3.4** Functional pathway analysis of the 928 genes found associated with the DMRs. Analysis executed using Reactome and KEGG pathway databases. Only significant terms with FDR-adjusted p-value < 0.05 are reported.

| **Repository** | **Term** | **Overlap** | **Adjusted p-value** | **Genes** |
|---|---|---|---|---|
| Reactome | TNFs bind their physiological receptors R-HSA-5669034 | 7/24 | 0.034 | EDARADD, TNFSF18, TNFRSF6B, TNFSF4, CD27, TNFRSF25, TNFRSF1A |
| Reactome | Regulation of beta-cell development R-HSA-186712 | 8/32 | 0.034 | NEUROD1, NR5A2, HNF4A, ONECUT1, PDX1, SLC2A2, PAX6, NKX6-1 |
| Reactome | Nuclear Receptor transcription pathway R-HSA-383280 | 10/51 | 0.034 | THRB, NR5A2, HNF4A, NR1I2, RORC, ESRRG, NR2E1, NR3C1, NR0B2, ESR1 |
| Reactome | Circadian Clock R-HSA-400253 | 11/63 | 0.035 | SREBF1, F7, RAI1, CPT1A, MEF2C, CRTC1, CRY2, CUL1, SIK1, NR3C1, HIF1A |
| Reactome | Regulation of gene expression in beta cells R-HSA-210745 | 6/20 | 0.045 | NEUROD1, HNF4A, PDX1, SLC2A2, PAX6, NKX6-1 |
| Reactome | Regulation of gene expression in early pancreatic precursor cells R-HSA-210747 | 4/8 | 0.048 | NR5A2, ONECUT1, PDX1, NKX6-1 |
| KEGG | Maturity onset diabetes of the young | 9/26 | 0.0004 | NEUROD1, NR5A2, HNF4A, ONECUT1, PDX1, SLC2A2, PAX6, NKX6-1, MNX1 |

We identified the genes whose promoter overlaps the DMRs and performed functional enrichment analysis. The genes found are 350 and the enrichment analysis results are reported in Table 3.5. Table 3.5 shows enrichment in glycosphingolipids (GSLs), biomolecules located in the cell plasma membrane, known for their regulator function of different cells. Changes in GSLs expression have been found linked to cancer growth and are used as biomarkers for immune-based cancer treatment [221].

**Table 3.5** Enrichment analysis on the genes with promoter regions overlapping the DMRs. Analysis performed on the gene ontology Biological Processes (BP) and Molecular Function (MF). Only statistically significant terms with FDR-adjusted p-value < 0.05 are shown.

| Repository | Term | Overlap | Adjusted P-value |
|---|---|---|---|
| GO MF | sialyltransferase activity (GO:0008373) | 5/21 | 0.008 |
| GO MF | double-stranded DNA binding (GO:0003690) | 26/651 | 0.011 |
| GO MF | alpha-N-acetylneuraminate alpha-2 8-sialyltransferase activity (GO:0003828) | 3/6 | 0.011 |
| GO MF | sequence-specific DNA binding (GO:0043565) | 26/707 | 0.024 |
| GO MF | sequence-specific double-stranded DNA binding (GO:1990837) | 26/712 | 0.024 |
| GO MF | RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978) | 36/1149 | 0.03 |
| GO MF | DNA-binding transcription activator activity RNA polymerase II-specific (GO:0001228) | 15/333 | 0.038 |
| GO MF | RNA polymerase II transcription regulatory region sequence-specific DNA binding (GO:0000977) | 40/1359 | 0.038 |
| GO MF | cis-regulatory region sequence-specific DNA binding (GO:0000987) | 35/1149 | 0.039 |
| GO MF | glutathione transferase activity (GO:0004364) | 4/27 | 0.04 |
| GO BP | glycosphingolipid biosynthetic process (GO:0006688) | 5/16 | 0.012 |

Dysregulation of the GSLs is common and occurs after translation, used as therapeutic and diagnostic markers for multiple cancer types such as breast and colorectal cancer, glioblastoma and recently explored in the context of PDAC progression [222–225]. Literature shows a growing demonstration of the link between aberrant GLSs expression and cancer signaling pathways such as the epithelial-to-mesenchymal transition (EMT), an event associated with the metastatic promotion and malignant cells diffusion [226]. Genes found enriched in GSLs biosynthetic process are ST8SIA1, B3GALT4, ST8SIA4 ST8SIA5, and ST6GALNAC5. ST8SIA1 and ST8SIA4 are known to be responsible for metastasis outgrowth in breast cancer [227, 228], ST8SIA5 was observed hypermethylated in PDAC [212] and ST6GALNAC5 was found responsible for promoting PDAC resistance to the chemotherapy gemcitabine [229].

### 3.2.2 Results on benign and malignant meningioma

We investigate the role of DNA methylation in the progression of meningioma from benign to malignant. We applied Dimmer 2.0 intending to find potential epigenetic markers able to explain the differentiation between the two diagnostic stages. Dimmer 2.0 identified 2100 statistically significant DMRs, with an average width of 985 base pairs and an average density (number of CpGs/DMR width) of 0.01, with 10 CpGs per region, on average. Density and DMRs size distribution is illustrated in Figure 3.11). Visualization of one of the DMRs is proposed in Figure 3.12 and has as coordinates chr6:30 711 073-30 712 559.

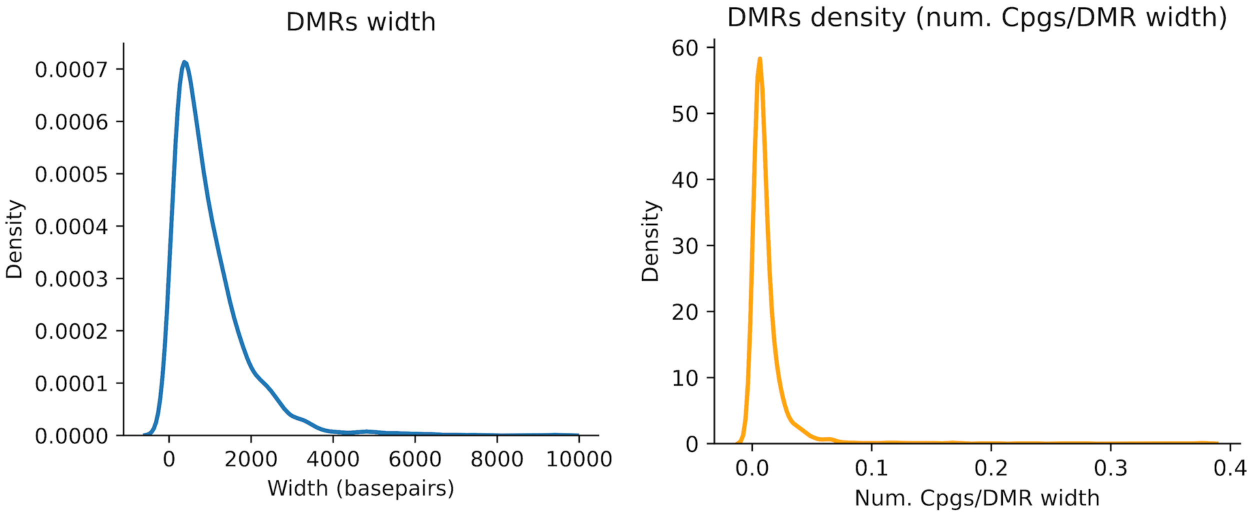


**Figure 3.11** Distribution of the width of the 2100 DMRs returned by Dimmer 2.0 (left) measured in basepairs and distribution of the region's density (right) measured as the amount of CpGs found significant inside a region.

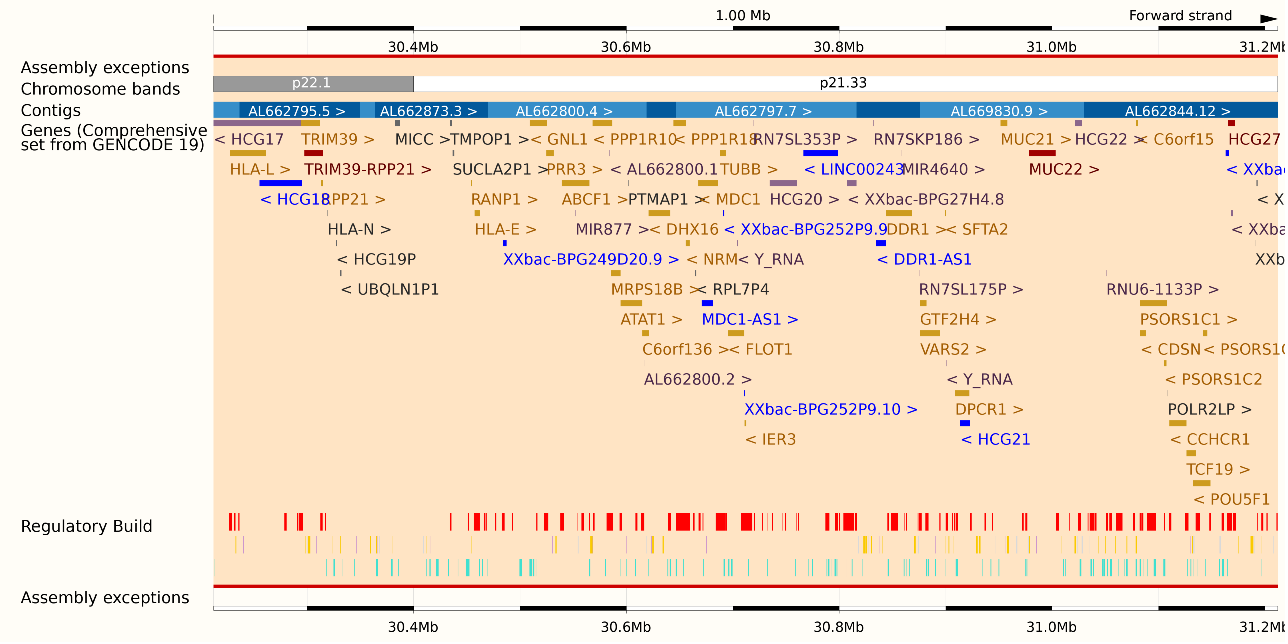


**Figure 3.12** Genomic view of the DMR identified by Dimmer 2.0 from comparing benign meningioma samples *versus* malignant ones. Region located on chromosome 6, starting at 30 711 073 and ending at 30 712 559. The view is obtained through the Ensemble Genome Browser [53].

The implementation of Epiregio found 383 significant REMS with 369 associated genes. However, enrichment analysis on the genes did not identify any significant term. With GREAT we identified

2264 genes overlapping the DMRs by their TSS. GREAT performs enrichment analysis on the genes using GO terms, where genes were found engaged in embryonic organ morphogenesis and development, protein binding, and regulation of transcriptional activity. Results are reported in Supplementary Table 5.3. An independent pathway enrichment analysis revealed enrichment in cancer-related pathways and genes involved in beta cell regulation (Table 3.6). Interesting is the enrichment of PI3K/AKT, a signaling pathway with a fundamental function in cell cycle regulation, known to be dysregulated in numerous cancer types [230]. Moreover, it has been proved a connection between the activation of PI3K/AKT and later stages of meningioma, explaining the progression to more aggressive conditions, thus proposed as a therapeutic target [231]. By overlapping DMRs with gene promoters, we identified 845 genes, particularly the enrichment of HOX genes, potential prognostic markers [232]. Pathway analysis showed enrichment in several cell types, such as neurons, stem cells, and epithelial cells. Results are available in Supplementary Table 5.4.

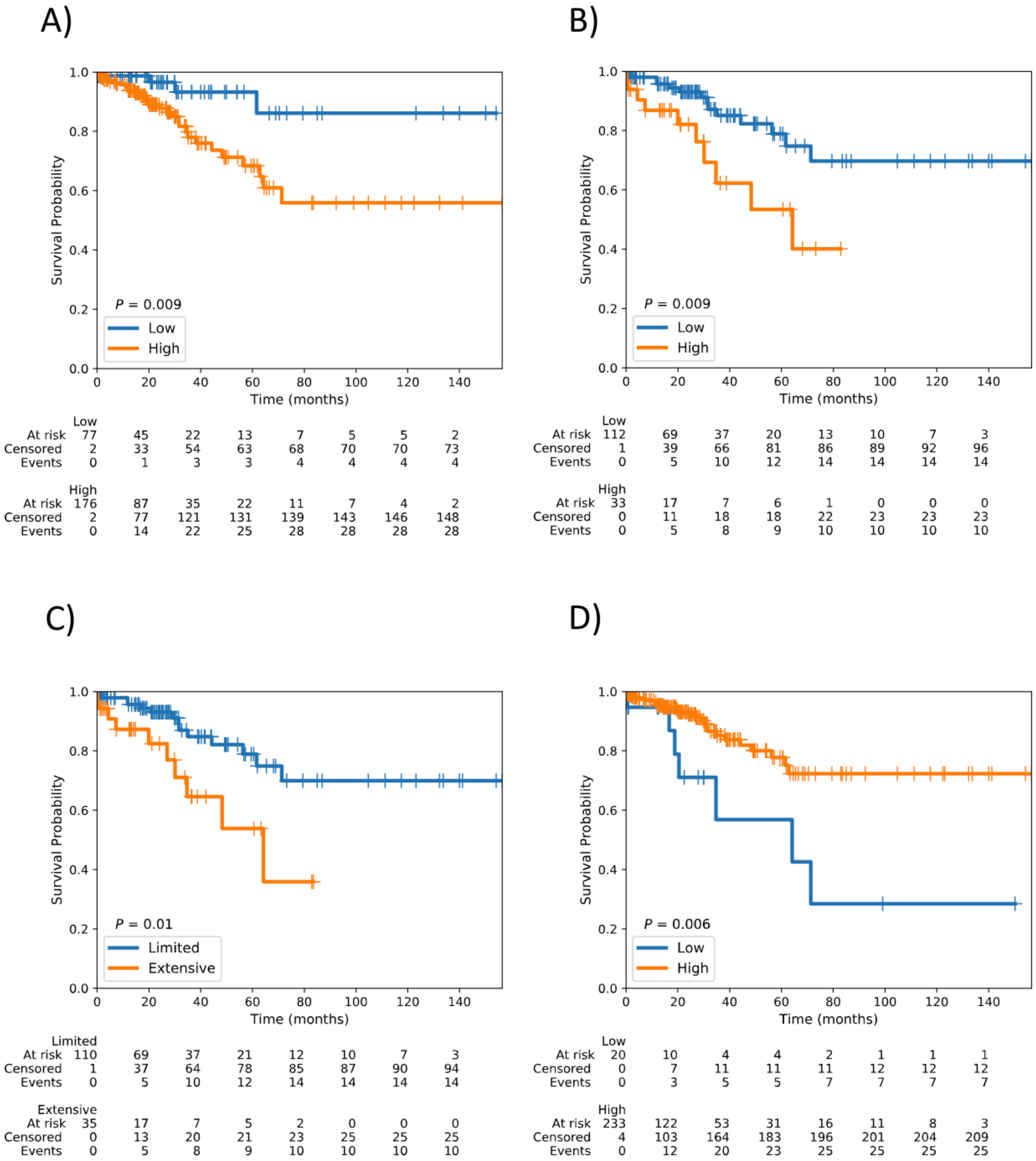
**Table 3.6** Functional pathway analysis of the 2264 genes associated with the DMRs. Analysis executed using Reactome and KEGG pathway databases. Only significant terms with FDR-adjusted p-value < 0.05 are reported.

| Repository | Term | Overlap | Adjusted p-value |
|---|---|---|---|
| Reactome | Developmental Biology Homo sapiens R-HSA-1266738 | 139/786 | 4.63E-05 |
| Reactome | Regulation of beta-cell development Homo sapiens R-HSA-186712 | 16/32 | 4.63E-05 |
| Reactome | Constitutive Signaling by Aberrant PI3K in Cancer Homo sapiens R-HSA-2219530 | 20/61 | 0.003 |
| Reactome | Transcriptional regulation of pluripotent stem cells Homo sapiens R-HSA-452723 | 15/42 | 0.0069 |
| Reactome | PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling Homo sapiens R-HSA-6811558 | 23/83 | 0.0069 |
| Reactome | Activation of anterior HOX genes in hindbrain development during early embryogenesis Homo sapiens R-HSA-5617472 | 24/89 | 0.0069 |
| Reactome | Activation of HOX genes during differentiation Homo sapiens R-HSA-5619507 | 24/89 | 0.0069 |
| Reactome | Negative regulation of the PI3K/AKT network Homo sapiens R-HSA-199418 | 24/90 | 0.007 |
| Reactome | Regulation of gene expression in early pancreatic precursor cells Homo sapiens R-HSA-210747 | 6/8 | 0.007 |
| Reactome | Regulation of gene expression in beta cells Homo sapiens R-HSA-210745 | 9/20 | 0.0189 |
| KEGG | Maturity onset diabetes of the young | 16/26 | 3.77E-07 |
| KEGG | Signaling pathways regulating pluripotency of stem cells | 37/143 | 0.0002 |
| KEGG | Gastric cancer | 34/149 | 0.0037 |
| KEGG | Neuroactive ligand-receptor interaction | 61/341 | 0.0103 |
| KEGG | Parathyroid hormone synthesis, secretion and action | 25/106 | 0.0103 |
| KEGG | Transcriptional misregulation in cancer | 38/192 | 0.0135 |
| KEGG | Basal cell carcinoma | 16/63 | 0.0365 |

## 3.3 The amount of tumor buds serves as a prognostic marker in histomorphology

We used the established grading system [148, 149] to grade the samples of the study cohort according to their TB activity. This grading appeared not survival relevant for neither OS nor PFI (p-value for OS=0.28, for PFI=0.45). Hence, we identified a cut-off that is prognostically significant for four morphological traits of lung-SCC: TB, STAS, the distance of STAS in the alveoli, and immune cell infiltration. The amount of stroma was also tested for its prognostic role but no significant cut-off was found. For TB counts the cut-off identified is of two tumor buds. Hence, patients are divided into two groups according to the number of buds found in 10 HPF, where one group represents low budding activity (zero or one bud) and the second group represents the high activity (two buds or more) with a p-value of 0.009 (Figure 3.13 A). The same cut-off was identified for the amount of STAS observed (low STAS if zero or one, high STAS otherwise), with a p-value of 0.009 (Figure 3.13 B). In the case of the distance of STAS in alveoli, a cut-off of three was determined and samples are considered as *limited* in distance in case the number of alveoli is between zero and two, and *extensive* in case of three or more alveoli, with p-value=0.01 (Figure 3.13 C). In the case of immune cell infiltration, the survival of patients differs the most when divided into low and high immune cell content with a cut-off of 2% (p-value=0.006) (Figure 3.13 D). Minimum cell nest size was also considered and a cut-off of three cells was identified to determine whether a cell nest is to be considered small (less than three cells) or large (three or more cells). Cut-off for the amount of stroma was found at 35%, grouping patients into low and high presence of stroma. The cut-offs identified for stroma content and minimum cell nest size did not appear to be useful for prognostic stratification either for OS or for PFI.

Cut-offs for TB, STAS, and immune cells content failed to predict PFI (p-value = 0.8, 0.14, 0.97, respectively) contrary to distance of STAS (p-value=0.03). To further assess the prognostic impact of the cut-offs, additional survival analysis was conducted with Cox-regression including age, sex, AJCC stage as risk factors for both OS and PFI. Higher risk of death for patients with high TB (HR=3.57, p-value=0.02), high STAS (HR=3.13, p-value=0.01), an extensive distance of STAS in alveoli (HR=2.63, p-value=0.03) and rich immune cell infiltration (HR=3.08, p-value=0.01) as shown in Figure 3.14, left. High risk of disease progression was observed for patients with a high number of STAS (HR=1.96, p-value=0.04), and an extensive distance of STAS (HR=2, p-value=0.03), while high immune cell content is predictive for a reduced disease progression (HR=0.38, p-value=0.05). The amount of buds was not a determinant for PFI (HR=0.96, p-value=0.87) while stage III/IV, considered in the interrelation with the age, sex AJCC stage and TB, is predictive of PFI. The same result was observed when considering richness in immune cell infiltration, where stage III/IV is linked to an increased risk of disease progression. Models with age, sex, AJCC stages and STAS or distance of STAS were not significant. Summary of the results is in Figure 3.14.



A)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Low | | | | | | | | |
| At risk | 77 | 45 | 22 | 13 | 7 | 5 | 5 | 2 |
| Censored | 2 | 33 | 54 | 63 | 68 | 70 | 70 | 73 |
| Events | 0 | 1 | 3 | 3 | 4 | 4 | 4 | 4 |
| High | | | | | | | | |
| At risk | 176 | 87 | 35 | 22 | 11 | 7 | 4 | 2 |
| Censored | 2 | 77 | 121 | 131 | 139 | 143 | 146 | 148 |
| Events | 0 | 14 | 22 | 25 | 28 | 28 | 28 | 28 |

B)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Low | | | | | | | | |
| At risk | 112 | 69 | 37 | 20 | 13 | 10 | 7 | 3 |
| Censored | 1 | 39 | 66 | 81 | 86 | 89 | 92 | 96 |
| Events | 0 | 5 | 10 | 12 | 14 | 14 | 14 | 14 |
| High | | | | | | | | |
| At risk | 33 | 17 | 7 | 6 | 1 | 0 | 0 | 0 |
| Censored | 0 | 11 | 18 | 18 | 22 | 23 | 23 | 23 |
| Events | 0 | 5 | 8 | 9 | 10 | 10 | 10 | 10 |

C)

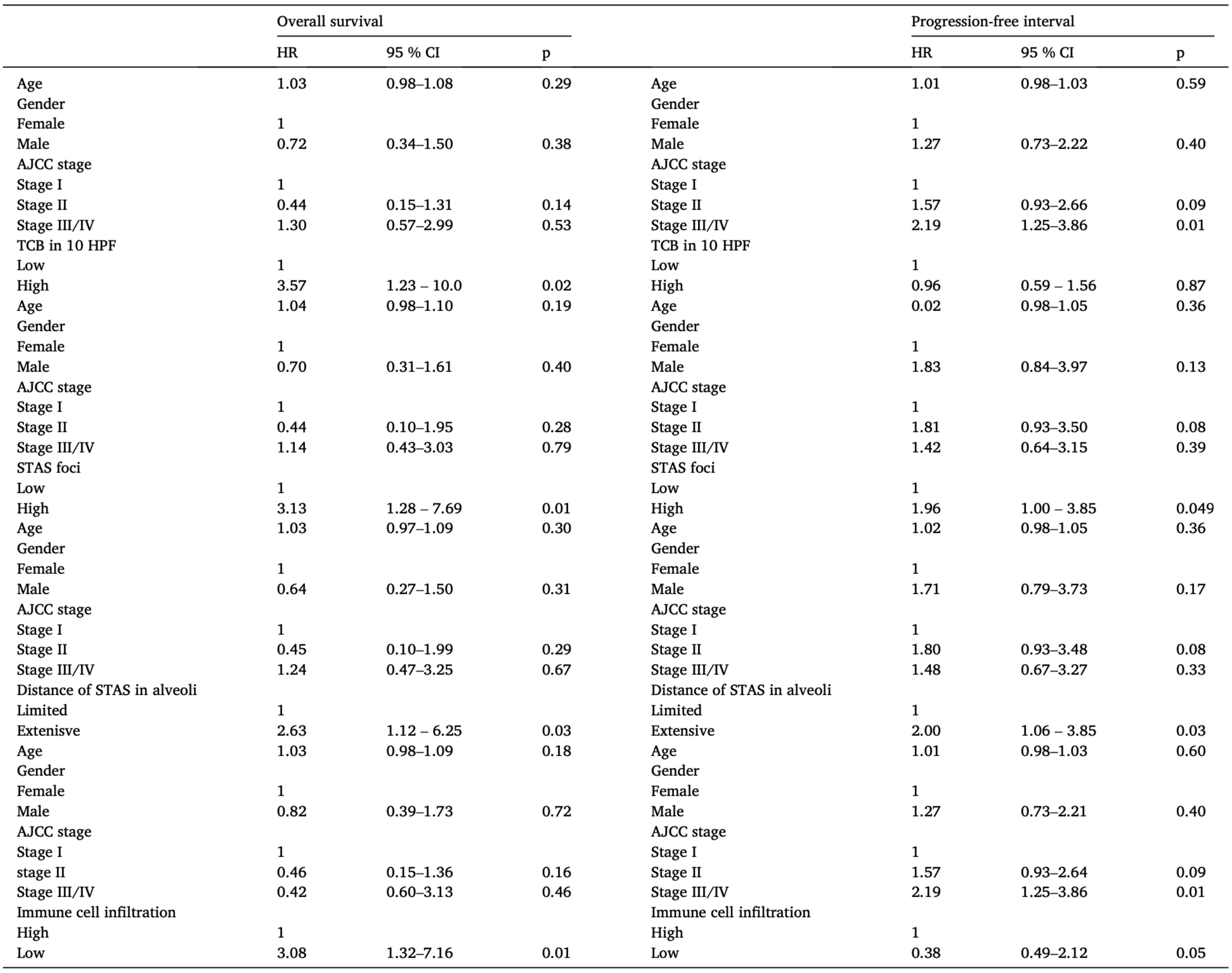
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Limited | | | | | | | | |
| At risk | 110 | 69 | 37 | 21 | 12 | 10 | 7 | 3 |
| Censored | 1 | 37 | 64 | 78 | 85 | 87 | 90 | 94 |
| Events | 0 | 5 | 10 | 12 | 14 | 14 | 14 | 14 |
| Extensive | | | | | | | | |
| At risk | 35 | 17 | 7 | 5 | 2 | 0 | 0 | 0 |
| Censored | 0 | 13 | 20 | 21 | 23 | 25 | 25 | 25 |
| Events | 0 | 5 | 8 | 9 | 10 | 10 | 10 | 10 |

D)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Low | | | | | | | | |
| At risk | 20 | 10 | 4 | 4 | 2 | 1 | 1 | 1 |
| Censored | 0 | 7 | 11 | 11 | 11 | 12 | 12 | 12 |
| Events | 0 | 3 | 5 | 5 | 7 | 7 | 7 | 7 |
| High | | | | | | | | |
| At risk | 233 | 122 | 53 | 31 | 16 | 11 | 8 | 3 |
| Censored | 4 | 103 | 164 | 183 | 196 | 201 | 204 | 209 |
| Events | 0 | 12 | 20 | 23 | 25 | 25 | 25 | 25 |

**Figure 3.13** Kaplan-Meier plot of the TCGA dataset. Survival anlalysis conducted on OS. The two survival curves are obtained using the cut-off identified for TB (A), STAS (B), distance of STAS in alveoli (C) and immune cell proliferation (D). Figure from [134], with permission of Elsevier.

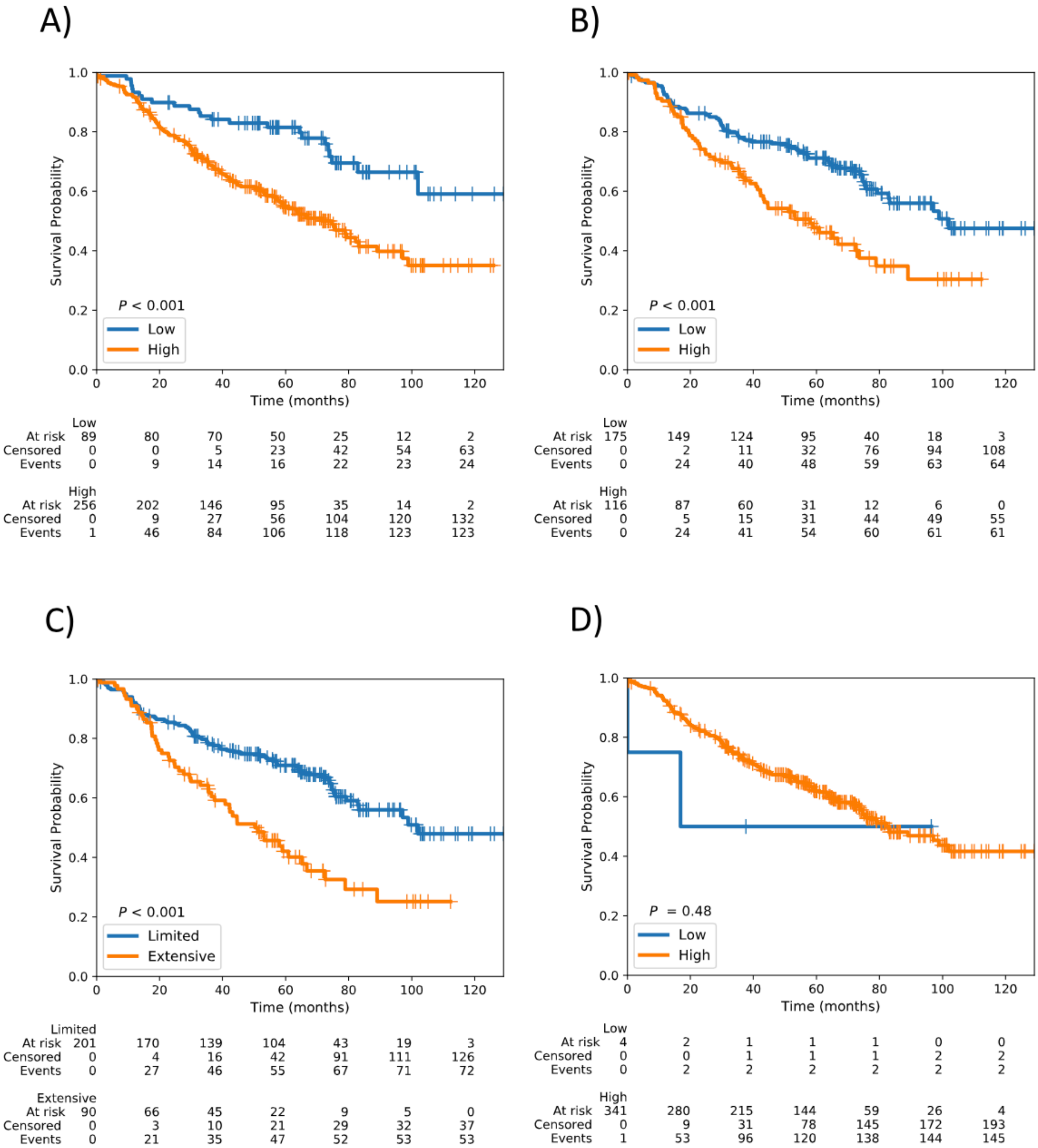
**Figure 3.14** Cox regression of the TCGA cohort. Analysis performed on TB, STAS, distance of STAS in alveoli and immune cell proliferation presented one after the other an divided by OS (left) and PFI (right). For each variable hazard-ratio (HR), confidence interval (CI) and p-value are shown. Table from [134], with permission of Elsevier.



| | Overall survival | | | | Progression-free interval | | |
|---|---|---|---|---|---|---|---|
| | HR | 95 % CI | p | | HR | 95 % CI | p |
| Age | 1.03 | 0.98–1.08 | 0.29 | Age | 1.01 | 0.98–1.03 | 0.59 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.72 | 0.34–1.50 | 0.38 | Male | 1.27 | 0.73–2.22 | 0.40 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 0.44 | 0.15–1.31 | 0.14 | Stage II | 1.57 | 0.93–2.66 | 0.09 |
| Stage III/IV | 1.30 | 0.57–2.99 | 0.53 | Stage III/IV | 2.19 | 1.25–3.86 | 0.01 |
| TCB in 10 HPF | | | | TCB in 10 HPF | | | |
| Low | 1 | | | Low | 1 | | |
| High | 3.57 | 1.23 – 10.0 | 0.02 | High | 0.96 | 0.59 – 1.56 | 0.87 |
| Age | 1.04 | 0.98–1.10 | 0.19 | Age | 0.02 | 0.98–1.05 | 0.36 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.70 | 0.31–1.61 | 0.40 | Male | 1.83 | 0.84–3.97 | 0.13 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 0.44 | 0.10–1.95 | 0.28 | Stage II | 1.81 | 0.93–3.50 | 0.08 |
| Stage III/IV | 1.14 | 0.43–3.03 | 0.79 | Stage III/IV | 1.42 | 0.64–3.15 | 0.39 |
| STAS foci | | | | STAS foci | | | |
| Low | 1 | | | Low | 1 | | |
| High | 3.13 | 1.28 – 7.69 | 0.01 | High | 1.96 | 1.00 – 3.85 | 0.049 |
| Age | 1.03 | 0.97–1.09 | 0.30 | Age | 1.02 | 0.98–1.05 | 0.36 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.64 | 0.27–1.50 | 0.31 | Male | 1.71 | 0.79–3.73 | 0.17 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 0.45 | 0.10–1.99 | 0.29 | Stage II | 1.80 | 0.93–3.48 | 0.08 |
| Stage III/IV | 1.24 | 0.47–3.25 | 0.67 | Stage III/IV | 1.48 | 0.67–3.27 | 0.33 |
| Distance of STAS in alveoli | | | | Distance of STAS in alveoli | | | |
| Limited | 1 | | | Limited | 1 | | |
| Extenisve | 2.63 | 1.12 – 6.25 | 0.03 | Extensive | 2.00 | 1.06 – 3.85 | 0.03 |
| Age | 1.03 | 0.98–1.09 | 0.18 | Age | 1.01 | 0.98–1.03 | 0.60 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.82 | 0.39–1.73 | 0.72 | Male | 1.27 | 0.73–2.21 | 0.40 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| stage II | 0.46 | 0.15–1.36 | 0.16 | Stage II | 1.57 | 0.93–2.64 | 0.09 |
| Stage III/IV | 0.42 | 0.60–3.13 | 0.46 | Stage III/IV | 2.19 | 1.25–3.86 | 0.01 |
| Immune cell infiltration | | | | Immune cell infiltration | | | |
| High | 1 | | | High | 1 | | |
| Low | 3.08 | 1.32–7.16 | 0.01 | Low | 0.38 | 0.49–2.12 | 0.05 |

An independent cohort is used to assess the validity of the cut-offs in dividing patients by prognosis. We performed a univariate survival analysis where a log-rank test was used to assess the difference in OS between the two groups of samples identified by the cut-offs of four variables: TB, STAS, distance of STAS and immune cells infiltration. Results show survival curves significantly different based on TB, STAS, distance of STAS but not on immune cells content, as illustrated in Figure 3.15. This time, besides OS, also the time a patient is disease-free after treatment was investigate. Analysis on DFS found a p-value lower than 0.001 for TB, distance of STAS and STAS, while p-value=0.81 for immune cells infiltration.

Predictive models were built for the four variables independently, with age, sex and AJCC stage as additional risk factors. Cox regression shows comparable results between OS and DFS. If the patient is at stage III/IV and presents a high TB activity, the death risk is increased, as well as the disease relapse. The same conclusion can be taken when considering STAS and distance of STAS. Results are summarized in Figure 3.16.



A)

| | 0 | 20 | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|---|
| Low | | | | | | | |
| At risk | 89 | 80 | 70 | 50 | 25 | 12 | 2 |
| Censored | 0 | 0 | 5 | 23 | 42 | 54 | 63 |
| Events | 0 | 9 | 14 | 16 | 22 | 23 | 24 |
| High | | | | | | | |
| At risk | 256 | 202 | 146 | 95 | 35 | 14 | 2 |
| Censored | 0 | 9 | 27 | 56 | 104 | 120 | 132 |
| Events | 1 | 46 | 84 | 106 | 118 | 123 | 123 |

B)

| | 0 | 20 | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|---|
| Low | | | | | | | |
| At risk | 175 | 149 | 124 | 95 | 40 | 18 | 3 |
| Censored | 0 | 2 | 11 | 32 | 76 | 94 | 108 |
| Events | 0 | 24 | 40 | 48 | 59 | 63 | 64 |
| High | | | | | | | |
| At risk | 116 | 87 | 60 | 31 | 12 | 6 | 0 |
| Censored | 0 | 5 | 15 | 31 | 44 | 49 | 55 |
| Events | 0 | 24 | 41 | 54 | 60 | 61 | 61 |

C)

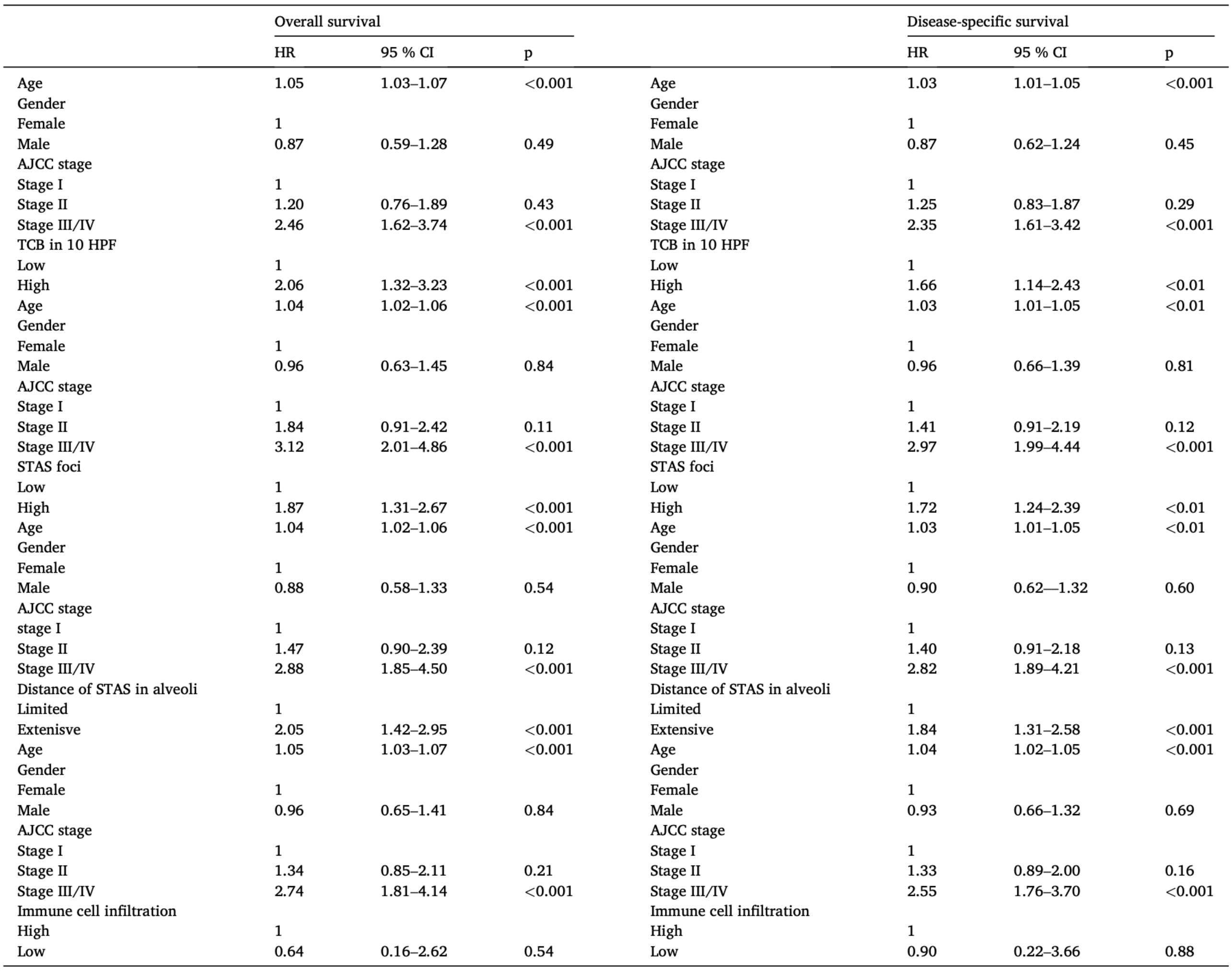
| | 0 | 20 | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|---|
| Limited | | | | | | | |
| At risk | 201 | 170 | 139 | 104 | 43 | 19 | 3 |
| Censored | 0 | 4 | 16 | 42 | 91 | 111 | 126 |
| Events | 0 | 27 | 46 | 55 | 67 | 71 | 72 |
| Extensive | | | | | | | |
| At risk | 90 | 66 | 45 | 22 | 9 | 5 | 0 |
| Censored | 0 | 3 | 10 | 21 | 29 | 32 | 37 |
| Events | 0 | 21 | 35 | 47 | 52 | 53 | 53 |

D)

| | 0 | 20 | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|---|
| Low | | | | | | | |
| At risk | 4 | 2 | 1 | 1 | 1 | 0 | 0 |
| Censored | 0 | 0 | 1 | 1 | 1 | 2 | 2 |
| Events | 0 | 2 | 2 | 2 | 2 | 2 | 2 |
| High | | | | | | | |
| At risk | 341 | 280 | 215 | 144 | 59 | 26 | 4 |
| Censored | 0 | 9 | 31 | 78 | 145 | 172 | 193 |
| Events | 1 | 53 | 96 | 120 | 138 | 144 | 145 |

**Figure 3.15** Kaplan-Meier plot of the validation dataset. Survival anlalysis conducted on OS. The two survival curves are obtained using the cut-off identified for TB (A), STAS (B), distance of STAS in alveoli (C) and immune cell proliferation (D). Figure from [134], with permission of Elsevier.

**Figure 3.16** Cox regression of the validation cohort. Analysis performed on TB, STAS, distance of STAS in alveoli and immune cell proliferation presented one after the other an divided by OS (left) and DFS (right). For each variable hazard-ratio (HR), confidence interval (CI) and p-value are shown. Table from [134], with permission of Elsevier.



| | Overall survival | | | | Disease-specific survival | | |
|---|---|---|---|---|---|---|---|
| | HR | 95 % CI | p | | HR | 95 % CI | p |
| Age | 1.05 | 1.03–1.07 | <0.001 | Age | 1.03 | 1.01–1.05 | <0.001 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.87 | 0.59–1.28 | 0.49 | Male | 0.87 | 0.62–1.24 | 0.45 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 1.20 | 0.76–1.89 | 0.43 | Stage II | 1.25 | 0.83–1.87 | 0.29 |
| Stage III/IV | 2.46 | 1.62–3.74 | <0.001 | Stage III/IV | 2.35 | 1.61–3.42 | <0.001 |
| TCB in 10 HPF | | | | TCB in 10 HPF | | | |
| Low | 1 | | | Low | 1 | | |
| High | 2.06 | 1.32–3.23 | <0.001 | High | 1.66 | 1.14–2.43 | <0.01 |
| Age | 1.04 | 1.02–1.06 | <0.001 | Age | 1.03 | 1.01–1.05 | <0.01 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.96 | 0.63–1.45 | 0.84 | Male | 0.96 | 0.66–1.39 | 0.81 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 1.84 | 0.91–2.42 | 0.11 | Stage II | 1.41 | 0.91–2.19 | 0.12 |
| Stage III/IV | 3.12 | 2.01–4.86 | <0.001 | Stage III/IV | 2.97 | 1.99–4.44 | <0.001 |
| STAS foci | | | | STAS foci | | | |
| Low | 1 | | | Low | 1 | | |
| High | 1.87 | 1.31–2.67 | <0.001 | High | 1.72 | 1.24–2.39 | <0.01 |
| Age | 1.04 | 1.02–1.06 | <0.001 | Age | 1.03 | 1.01–1.05 | <0.01 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.88 | 0.58–1.33 | 0.54 | Male | 0.90 | 0.62—1.32 | 0.60 |
| AJCC stage | | | | AJCC stage | | | |
| stage I | 1 | | | Stage I | 1 | | |
| Stage II | 1.47 | 0.90–2.39 | 0.12 | Stage II | 1.40 | 0.91–2.18 | 0.13 |
| Stage III/IV | 2.88 | 1.85–4.50 | <0.001 | Stage III/IV | 2.82 | 1.89–4.21 | <0.001 |
| Distance of STAS in alveoli | | | | Distance of STAS in alveoli | | | |
| Limited | 1 | | | Limited | 1 | | |
| Extenisve | 2.05 | 1.42–2.95 | <0.001 | Extensive | 1.84 | 1.31–2.58 | <0.001 |
| Age | 1.05 | 1.03–1.07 | <0.001 | Age | 1.04 | 1.02–1.05 | <0.001 |
| Gender | | | | Gender | | | |
| Female | 1 | | | Female | 1 | | |
| Male | 0.96 | 0.65–1.41 | 0.84 | Male | 0.93 | 0.66–1.32 | 0.69 |
| AJCC stage | | | | AJCC stage | | | |
| Stage I | 1 | | | Stage I | 1 | | |
| Stage II | 1.34 | 0.85–2.11 | 0.21 | Stage II | 1.33 | 0.89–2.00 | 0.16 |
| Stage III/IV | 2.74 | 1.81–4.14 | <0.001 | Stage III/IV | 2.55 | 1.76–3.70 | <0.001 |
| Immune cell infiltration | | | | Immune cell infiltration | | | |
| High | 1 | | | High | 1 | | |
| Low | 0.64 | 0.16–2.62 | 0.54 | Low | 0.90 | 0.22–3.66 | 0.88 |

## 3.4 Gene expression analysis reveals EMT as potential TB driver

### 3.4.1 Single-omics analysis

Analysis of the relationship between expression and TB returned 595 statistically significant genes. The regression coefficients linked to each gene used as predictors in the analysis suggest how intense their relationship with TB is. In Table 3.7 we display the five statistically significant genes with the highest coefficients.

**Table 3.7** Five of the 595 statistically significant genes identified by linear regression using univariate gene expression, top 5 with the highest regression coefficients.

| Gene name | Adjusted p-value | Coefficient |
|---|---|---|
| TNFRSF12A | 3.93E-08 | 0.478 |
| KCNN4 | 4.06E-08 | 0.471 |
| DAPL1 | 8.63E-07 | -0.437 |
| TMPRSS11A | 6.19E-06 | -0.412 |
| IL11 | 6.37E-06 | 0.41 |

The first two genes, TNFRSF12A and KCNN4, are known for their contribution to metastasis invasion. TNFRSF12A (tumor necrosis factor receptor superfamily, member 12A) is a protein-coding gene responsible for the production of the TWEAK cytokine, a protein part of the TNF superfamily of receptors involved in the activation of the AKT signaling pathway. AKT pathway plays a fundamental role in cancer promotion, stimulating cell migration and preventing cell apoptosis [233]. TNFRSF12A is known to be up-regulated and overexpressed in cancer, it is additionally linked to metastasis formation and reduced patient survival, making it a prognostic biomarker and drug target for glioma [234], breast cancer [235], gastric cancer [236] and pancreatic cancer [237]. Numerous studies demonstrated the oncogenic role of KCNN4 in promoting tumor development and resistance to chemotherapy [219, 238, 239]. In lung adenocarcinoma, the silencing of KCNN4 was found responsible for malignant cell death induction and the suppression of cell invasion mechanisms [240]. To gain insight into the 595 genes we performed pathway enrichment analysis. Results see the genes strongly associated with cell invasion mechanisms given the enrichment in the pathway responsible for EMT, corroborating the hypothesis of EMT as a driver for TB. Enrichment analysis results performed on WikiPathway are illustrated in Figure 3.17. It is possible to see which gene is involved in which pathway in Supplementary Figure 5.9. An additional enrichment analysis carried out using the pathway repository BioPlanet found an enrichment in WNT signaling pathway (p-value = 4.619E-13), further supporting the involvement of EMT in the TB development. Genes found involved in the WNT signaling belong to the Collagen type 1 family and ADAMS gene family.

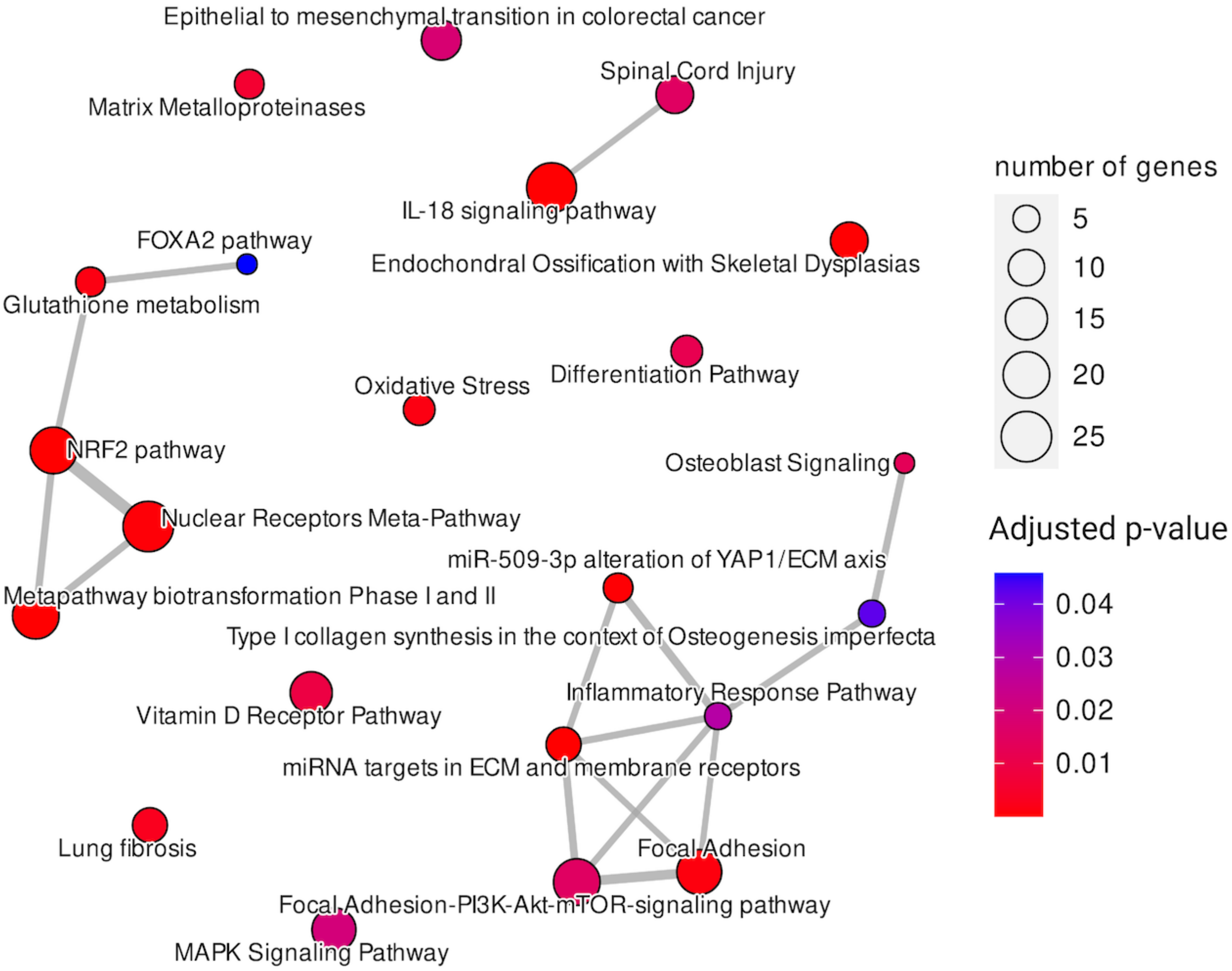


**Figure 3.17** Pathways significantly enriched in the 595 significant genes. Enrichment analysis was performed using WikiPathways as a pathways repository. The figure shows the interaction between pathways based on the intersecting genes. The width of the edges is proportional to the number of genes the two terms have in common while the size of the nodes is determined by the size of the gene set. The nodes' color reflects the adjusted p-value of each term.

DNA methylation levels of 543 CpGs were found significantly associated with TB. Aberrant methylation in the gene promoter region is known to be responsible for altered gene activity such as gene silencing or activation. We retrieved those genes whose promoter region includes at least one of the 543 CpGs, collecting a list of 76 genes. As for gene expression, we use enrichment analysis to identify the biological functionalities of the genes. High enrichment was found in the glucuronidation pathway (Table 3.8), a phase II metabolic reaction responsible for the metabolism of chemical substances and known for its involvement in drug inactivation. Genes found involved in this mechanism belongs to the UDP-glucuronosyltransferases (UGT) gene family, known as drug targets. Indeed, high expression levels of UGT genes have been found responsible for the inefficacy of a wide variety of drugs [241, 242].

**Table 3.8** Pathway enrichment analysis performed on the 76 genes whose promoter regions were found enriched with methylated CpGs associated with TB. Repositories shown are WikiPathway and Reactome and only significant terms (adjusted p-value < 0.05) are reported.

| Repository | Term | Overlap | Genes | Adjusted p-value |
|---|---|---|---|---|
| WikiPathway | Glucuronidation | 5/26 | UGT1A10, UGT1A9, UGT1A8, UGT1A7, UGT1A6 | 3.98E-06 |
| WikiPathway | Codeine and Morphine Metabolism | 4/15 | UGT1A10, UGT1A9, UGT1A8, UGT1A6 | 1.18E-05 |
| WikiPathway | Metapathway biotransformation Phase I and II | 5/183 | UGT1A10, GPX2, UGT1A9, UGT1A7, UGT1A6 | 0.016 |
| WikiPathway | Aryl Hydrocarbon Receptor Pathway | 3/46 | UGT1A9, UGT1A7, UGT1A6 | 0.016 |
| WikiPathway | Irinotecan pathway | 2/13 | UGT1A10, UGT1A9 | 0.02 |
| WikiPathway | Estrogen metabolism | 2/18 | UGT1A9, UGT1A6 | 0.03 |
| WikiPathway | NRF2 pathway | 4/146 | GPX2, UGT1A9, UGT1A7, UGT1A6 | 0.03 |
| WikiPathway | Tamoxifen metabolism | 2/21 | UGT1A10, UGT1A8 | 0.033 |
| Reactome | Glucuronidation | 4/20 | UGT1A9, UGT1A8, UGT1A7, UGT1A6 | 0.0001 |

Only one protein, MIG6, was found significantly associated with TB. MIG6 (also called ERRFI1) is well known for its tumor suppressor key role [243], found down-regulated in multiple cancer types like breast [244] and lung cancer [245]. MIG6 negatively regulates EGFR, a protein involved in the activation of signaling pathways like PI3K-AKT, responsible for cell migration promotion, by inhibiting the catalytic activity [246]. The development of chemotherapy drugs targeting EGFR inhibition showed promising results in improving lung cancer patients' survival [247]. However, the identification of drug-resistant cases addressed researchers in further analysis to elucidate the role of EGFR as a therapeutic target [248] and focus on its interaction with MIG6 taking this last one as a potential target in lung cancer [249]. Moreover, the activity of MIG6 was found to be dependent on EGFR expression level in cancer cells that, when low, affects MIG6 activity in regulating AKT and PHLPP, linked to tumorigenesis promotion [250]. Nevertheless, In this analysis MIG6 was found positively associated with TB with a regression coefficient of 3.97, suggesting that its role in the TB activity of lung cancer requires further study. TB was found associated with mutations in 281 genes. We identified frequently mutated genes that are involved in phosphoinositides metabolism, with seven out of the 281 mutated genes involved in it (PLCB4, MTMR3, MTMR12, INPP5D, PLCE1, PLCH1, PIK3C2B). Interesting is the case of PIK3C2B, a protein-coding gene part of the PI3K family, found linked to TB (Welch's t-test = 3.78). It is involved in signaling pathways and cell migration regulation and its mutation might affect cancer progression [251, 252]. The presence of a somatic mutation in PIK3C2B was found in non-squamous cell lung cancer [253]. Hence, a mutation in PIK3C2B might be linked to high activity TB hence to a worst prognosis.

### 3.4.2 Multi-omics integration suggests a link to epigenetic activity

Among the genes obtained from the analysis on methylation, there is glutathione peroxidase 2 (GPX2). The role of GPX2 is to maintain the cell status by protecting it against oxidative damage caused by lipid peroxidation. Literature determined that overexpression of GPX2 is found in different cancer types, including non-small cell lung cancer, associated with poorer survival of the patients [254]. Results found one CpG (cg26155983) in the promoter region of GPX2. This CpG is significantly associated with TB (regression coefficient = 0.29) it has a methylation level of $\beta$ value = 0.6. Moreover, it's negatively correlated with the gene expression level(Pearson correlation coefficient = -0.79, see Figure 3.18), suggesting that gene transcription is potentially inhibited by epigenetic activity. However, the GPX2 expression level was also found negatively linked to TB in lung-SCC (regression coefficient of -0.27). Other genes with methylated promoters significantly associated with TB are UGT1A6/7/8/9, involved in the metabolic pathway as shown in the previous section. As for GPX2, expression of UGT1A6/7/9 was also found negatively linked to TB. Methylation intensity in the promoters of UGT1A6 and UGT1A9 is similar ($\beta$-value = 0.7) and both show a high negative correlation with expression (respectively -0.74 and -0.76 Pearson correlation), as illustrated in Figure 3.18. The gene expression of these two genes was also found significantly linked to TB.

Expression of protein MIG6 was found positively associated with TB. However, the negative expression levels of the protein (average of -0.14) together with the methylated promoter of the associated gene (average $\beta$-value=0.38) suggests a possible epigenetic silencing of the protein transcription. This hypothesis was further supported by the negative correlation (-0.29) observed between methylation values and gene expression, while no correlation was registered between methylation and protein expression. However, it is not clear what possible functionality MIG6 has in the context of tumor budding located in lung-SCC cancer patients. Further experimental analysis should be carried out to better understand the role of this gene. The seven mutated genes associated with TB and involved in phosphoinositides (PLCB4, MTMR3, MTMR12, INPP5D, PLCE1, PLCH1, PIK3C2B) were found highly expressed though expression levels were found not correlated with the presence or absence of a somatic mutation.

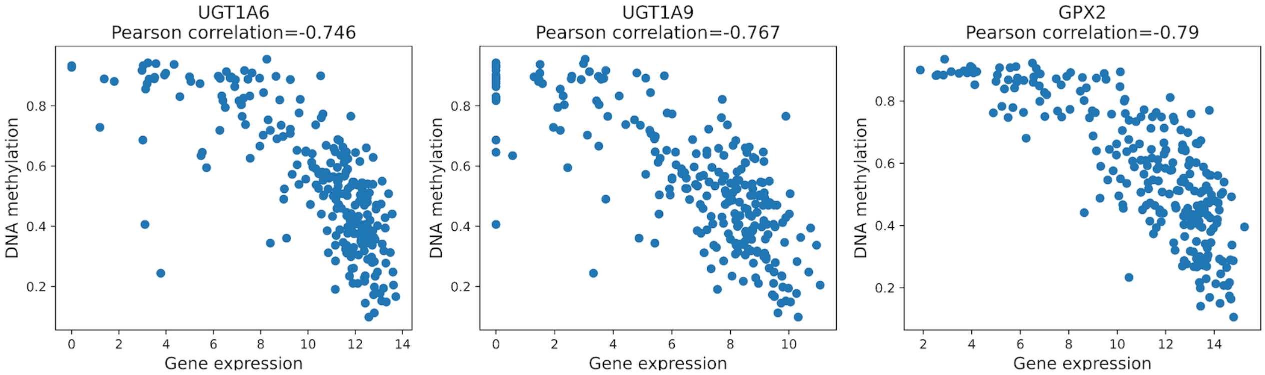


**Figure 3.18** Pearson correlation computed between gene expression profile and average promoter methylation for UGT1A6, UGT1A9 and GPX2. The expression and methylation level of the three genes was found significantly associated with TB activity.

### 3.4.3 TB is significantly associated with published molecular lung-SCC subtypes

Primitive, classical, secretory and basal are four molecular subtypes identified for lung-SCC and published in 2010 by Wilkerson and colleagues [181]. We investigated whether TB activity differs across the four subtypes and we found the difference to be statistically significant (p-value = 0.01) suggesting that the subtypes are characterized by diverse amount of buds. We further tested the subtypes' differences by comparing them with each other in a one vs others comparison. The subtype which

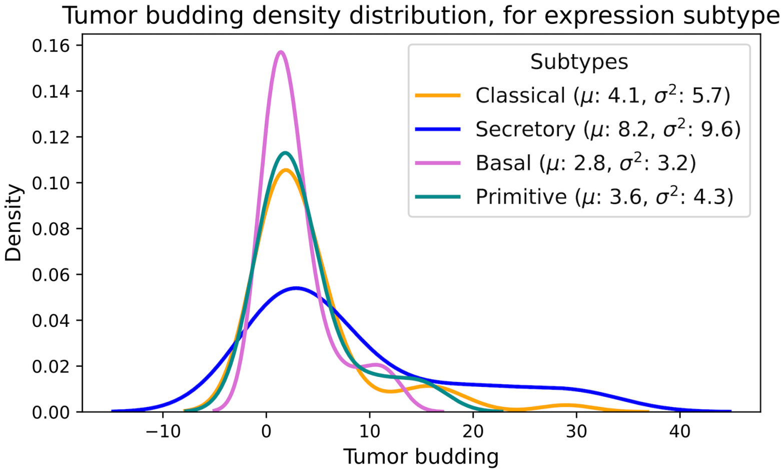


**Figure 3.19** Tumor budding distribution of the four molecular subtypes proposed by Wilkerson [181]. The density curve of the Secretory subtype differs from the others.

distinguishes itself from the others is the secretory subtype (p-value = 0.001). The basal subtype registered a p-value of 0.08 while the two remaining subtypes, primitive (p-value = 0.51) and classical (p-value = 0.39), are both not significant. Furthermore, comparing the distribution of tumor buds count we notice a different behavior in the secretory subtype, characterized by higher mean and larger standard deviation, as illustrated in Figure 3.19.

# 4 Discussion

### Computational evaluation of published molecular PDAC signatures

In oncology, the delineation of which tumor type affects a patient is crucial to designing a precise treatment approach, which coincides with the aim of the first project we presented. The complexity of PDAC is well known, with high heterogeneity of the tumor mass and the tumor microenvironment, and a high metastatic spread that makes this cancer type one of the most deadly. Numerous studies focused their research on identifying the subtypes PDAC differentiates in and the genetic markers that explain this differentiation. We performed an extensive analysis to understand the limits of four of the main established signatures for PDAC subtyping. Despite the promising results shown by the four studies ([88–91]), some main points were fully investigated yet, such as the reproducibility of the subtypes on independent datasets or the signatures' robustness. Additionally, besides the reliability of the subtype schemes in their existence, their composition and quantity across different studies show disagreement, where even just the number of subtypes still remains an open issue. A common conclusion we reached after the computational analyses is the fact that the subtypes are built fitting the dataset used for discovery, reflecting its specific composition and data preparation. Indeed, not all the subtypes can be reproduced on external datasets. Instability on the number of subtypes was firstly detected from clustering analysis where the clusters obtained not only do not always match the real labels but some of the subtypes are made of only one sample, meaning that a lower number of clusters would be preferred. An example is the clusters aimed to reproduce Bailey subtypes, where three subtypes rather than the proposed four seem more suitable, insinuating that a normal cell contamination was believed to be a distinct subtype. This was confirmed by cell deconvolution where healthy acinar cells were found enriched in the samples predicted as ADEX (Bailey) and as Exocrine-like (Collisson). We believe that in the data used by these two studies there is a cluster of healthy cells particularly prevalent in some samples, whichi influenced the classes identification. In samples predicted as Immunogenic (Bailey) and Desmoplastic (Puleo) we found the presence of macrophages, denoting a possible immuno-contamination. Further investigation should be addressed in understanding the immunogenicity of PDAC when contemplating a subtypes identification. Samples rich in immune cells such as T-cells and macrophages have been linked to poor prognosis and cancer aggressiveness, however, whether the immune reaction should be considered or not a subtype should be independently validated [255–258]. Bailey and Puleo's signatures reported better performance in classification when compared to the other two studies, Moffitt and Collisson. Furthermore, the Bailey signature was found as the most robust classifier when compared with random genes or the use of a shuffled target labels in the prediction model. Other signatures registered an overall performance better than randomized models, however, there have been observed cases of random signatures reaching equal or even higher accuracy. Subtypes like Immunogenic and ADEX (from Bailey), Exocrine-like (from Collisson), and Desmoplastic (from Puleo) require analysis at a more detailed level. With our analysis, we confirm the essential classification in the two classes basal-like and classical-like as the most solid and stable stratification, also demonstrated by survival analysis which sees the two diverging subtypes as the most prognostically different. The binary classification was additionally proven to be a successful determinant for patient survival and therapeutic outcome by Rashid and colleagues [116]. The authors propose a classifier named PurIST, able to predict whether a PDAC sample belongs to a basal or classical configuration, additionally showing promising prediction for chemotherapy reaction and prognosis. To be clinically applicable, clinical trials should be used to evaluate PurIST. Even though a two-tumor class appears to be robust, it also showed limitations and discrepancies [121, 150, 259]. The studies proposed by Moffitt, Collisson, Bailey, and Puleo certainly contributed to getting closer to elucidating the possible existing PDAC subgroup types. With our study, we demonstrated that the combination of machine learning techniques with high-throughput data is not enough for the purpose: we found inconsistencies and weak points that we can use as the foundation from which to continue the challenge of PDAC stratification. We believe that the different sample composition and data preparation are the first issues to be responsible for the subtype disagreement across studies. Only Moffitt and Collisson performed tumor microdissection and only Bailey and Puleo accounted for immune cell infiltration; on the other hand, only Bailey and Collisson considered the exocrine features as one independent subtype. Moreover, all these studies implemented an unsupervised approach, and the use of such analysis for subtype identification should be followed by an intensive evaluation on independent cohorts to confirm the veracity of the clusters found, validation that is either missing or performed superficially. Given the results achieved, we think that PDAC subtype identification might benefit from a different approach. In the literature, it has been raising the concept of a continuous tumor stratification, as it might better resemble the biological and clinical traits of PDAC [123]. The comparison between predicted and real subtype classes revealed a mismatch. However, such a mismatch might reinforce the hypothesis of a spectrum representing PDAC progression and change of morphology from a classical-like state to a more basal-like state. We strongly believe that future studies should focus on the concept of PDAC stratification as a continuum instead of mutually exclusive classes, given that one patient can present characteristics belonging to more than one class which can explain a sample misclassification or non classification. We observed the impact that cell heterogeneity, stroma, and TME have on PDAC stratification, therefore we consider that the use of single-cell profiling can provide crucial insights to understand the real composition of the tumor [122, 260–264]. Together with the use of single-cell data, we also believe that an omics data integration can give a broader picture of PDAC configuration and it can help reveal the mechanisms that drive its differentiation [265, 266]. All these aspects need consideration to avoid creating models that are only suitable for the data used for their construction and not for external data, and we stress the importance of the validation of the results obtained as one of the take-home messages. Indeed, although bioinformatics research often shows promising biomarkers, these should be cautiously validated in their prediction accuracy, robustness, and reproducibility as well as their clinical utility.

### De-novo DMRs discovery with Dimmer 2.0

With the second project, we offer a full pipeline for epigenome-wide differential study, which is an extended version of Dimmer ([172]). We have carefully surveyed the available methods for the discovery of differentially methylated regions (DMRs) and decided to upgrade Dimmer and propose Dimmer 2.0, a tool of easy implementation and results interpretation while ensuring reliability and robustness. It is common knowledge that the study of epigenetic marks is challenging yet of high relevance, and that epigenetic changes are often responsible for the initiation of new diseases or for creating conditions that enhance their progression. Methylation is measured for hundreds of thousands of genomic locations and the manipulation of such a big dataset can be demanding. Moreover, even after allocating where the alteration took place, linking that change to the phenotype might be challenging. Such a complex scenario stimulated us to provide a tool that can help scientists for a faster understanding of how methylation contributes to a phenotype. Methylation data can be generated from multiple platforms and presented in different formats. While numerous tools focus on a single platform, either array or sequencing-based and exclusively raw files, Dimmer 2.0 has the advantage of being suitable for both raw and preprocessed array platform data and sequencing technology data. Dimmer 2.0 is designed for the identification of *de novo* DMRs, for which the discovery of new biological insight is made possible. In Dimmer 2.0 a broad range of parameters can be modified to fit personal requirements and give the user the power to analyze different case studies (e.g. continuous phenotype or paired studies); moreover, it can work with both CpG and non-CpG contexts, not always possible in other DMR discovery tools. To test the significance of DMRs and minimize the amount of false positive regions (type 1 error), the best approach is to perform a permutation test on the final set of returned regions. Tools available in the literature rarely offer the chance to perform permutation analysis. The only ones including permutation are dmrseq [267] and the tools implementing the bump hunting algorithm (ChAMP [268] and Minfi [161]). A strength of Dimmer 2.0 is the easy implementation thanks to the dual availability of GUI and command line tool which makes it user-friendly and does not require any coding ability or previous computation. We further compared the characteristics of Dimmer 2.0 with the main tools published and highlighted its capability in identifying *de novo* DMRs and ensuring statistically significant results. A table containing the main existing tools and their characteristics can be found in Supplementary Table 5.2. We demonstrated the potential of Dimmer 2.0 using two distinct study cases and assessed the output in multiple ways to gain biological insight. We first compare paired samples of PDAC primary tumor to liver metastasis to unravel the epigenetic mechanisms linked to metastasis formation. The DMRs obtained were found to intersect regulatory elements involved in maintaining the stability of intracellular zinc level. Indeed, cells with a decreased zinc amount are linked to malignancies and cell growth. From the DMRs, we further obtained the genes whose TSS region matches the DMRs. Such genes were found involved in the Toll-like receptor signaling pathway which plays a key role in the innate immune response of an organism and in the regulation of the NF-$\kappa$B pathway, whose activation triggers cell proliferation, angiogenesis, and EMT. The second study case sees the comparison of methylation status of benign meningioma patients *versus* malignant ones. The DMRs we identified were found involved in the dysregulation of the PI3K/AKT pathway, known to affect meningioma progression. Additionally, the identified DMRs were found overlapping with the promoter regions of the HOX gene family. Methylation and expression levels of the HOX genes were found higher in the more advanced meningioma stages by some studies [269, 270]. Overall, with this study we showed the strength of Dimmer 2.0 presenting its novelties and advantages over other existing tools which makes it a solid choice for the analysis of methylation. Moreover, we showed the ability of Dimmer 2.0 to improve the biological understanding by identifying the mechanisms regulated by DNA methylation that are responsible for differentiating conditions.

### Unraveling the prognostic and molecular role of tumor budding in lung-SCC

Lung cancer is one of the leading cancer types with lung squamous cell carcinoma (lung-SCC) as one of the prevalent histological subtypes. Late diagnosis and high morbidity are a few of the main issues leading to a low therapy response rate. While prevention is still the main solution for a successful recovery, the progression in the personalised medicine research field represents a prominent aspect of treating patients in the most targetable way. To do so, it is fundamental to understand what differentiates tumor samples of the same kind. Different patients have different survival and reaction to treatments and with the analysis of their morphological and molecular aspects, it was possible to delineate prognostic biomarkers. Besides traditional platinum-based chemotherapy, progress has been made in the field of molecular-based therapies [271]. Identified molecular alterations are already used for the development of therapies targeted to patients affected by squamous cell lung cancer [272], however not yet for lung-SCC. Histomorphological traits have been proposed to categorize lung-SCC samples by prognosis, including tumor-stroma ratio and tumor-infiltrating lymphocytes. These aspects are prognostic biomarkers with the potential to improve the disease's early detection, but not yet fully investigated [273]. With the third project, we assessed the prognostic role of tumor budding and other histomorphological traits that characterize lung-SCC. We identified cut-offs for tumor budding, STAS, distance of STAS in alveoli and immune cell infiltration that efficiently worked as a stratification system for patients with different prognosis. Results on the study cohort showed that a high budding activity increases the risk of death, as well as STAS foci and an extensive distance between the farthest STAS and the tumor bulk. On the other hand, we observed a richer immune cell infiltration linked to a more positive prognosis. Looking at the progression-free interval, we found that the progression of the disease appeared not to be affected by these traits, except for the distance of STAS in alveoli, where a longer distance might promote disease progression. We further assessed the validity of the four traits using a validation cohort, which corroborates our finding of TB, STAS, and the distance of STAS. Interesting is the case of immune cell infiltration where the cut-off identified was not significant in the external cohort. Studies identified the role of the immune compartment as ambiguous since it has been linked to both tumor suppressor and promoting role [274]. The immune compound might not be a determinant for prognosis, hence we invite a deeper investigation of the immune compartment, known to be rich and heterogeneous in cancer [275]. In the second part of the project, we studied the potential biological processes and pathways that drive TB. Despite the prognostic impact of TB being already well studied [276–278], the mechanisms that explain why TB worsens survival are still unknown. We performed an association study based on multiple omics profiles including gene expression, protein expression, DNA methylation, and somatic mutation to identify the molecular features that help understanding the biological role of TB in tumorigenesis and that can potentially be targeted for therapeutic purposes. The association with gene expression data revealed an enrichment in EMT. This mechanisms might explain the presence of TB, hence it can be the potential TB driver which ultimately leads to the initiation of cancer metastasis [279]. We found two genes, TNFR12A and KCNN4 highly expressed and with a positive relationship with TB. TNFR12A and KCNN4 are both known as oncogenes and responsible for activating metastasis spread and known for their involvement in the EMT process [280]. TNFR12A and KCNN4 constitute possible therapeutic targets for lung-SCC because of their high association with TB, albeit experimental validation is needed to fully confirm their validity. GPX2 is normally found overexpressed in lung cancer [254]. Our results found methylated the promoter region of GPX2 and the RNA expression negatively associated with TB. Indeed, methylation in the GPX2 promoter was found positively linked to TB and negatively correlated with gene expression, suggesting gene silencing as a result of hypermethylation. However, the epigenetic role of this gene in lung-SCC is not yet fully investigated. We believe that studying the presence of GPX2 in patients with tumor budding with a focus on epigenetic mechanisms might help disclose the mechanisms behind TB. UGT1A6 and UGT1A9 are known to be involved in drug inactivation and we found a positive link between TB and their methylation status. The tumor suppressor role of MIG6 is widely known in the literature [281, 282]. However, our analysis returned contrasting results showing a positive association with the protein expression level of MIG6 with TB. A closer look at a possible link with methylation activity was done and revealed no correlation with protein expression but a negative one with the expression of the gene. Patients with mutated PIK3C2B *versus* not mutated show significantly different TB amounts, indicating that the rise of TB might be caused by a somatic mutation in this gene. We additionally investigated the presence of different TB quantities in lung-SCC molecular-based subtypes, finding significant results showing how different subtypes are characterized by a different amount of TB. We found molecular features whose altered methylated activity represents a promising starting point for further analysis and we believe that the study of TB from an epigenetic point of view might give relevant clarification on how and why TB begins and how it affects prognosis. The prognostic role of the aforementioned potential biomarkers should be investigated with an analysis of survival to assess whether they are also predictive of survival outcomes. However, the availability of independent cohorts with TB counts per sample is scarce, limiting the computational evaluation of our findings. Moreover, there is one open question which is whether the cells located in the buds are molecularly different from the ones in the bulk is still unknown. The advancement of spatial transcriptomics offers a powerful tool at an extremely detailed level thanks to the combination of spatial data with single-cell expression, and we believe that such an approach could help reveal if there is a change in molecular profile when a cell disassociates from the bulk to become bud [283].

# 5 Conclusion and Outlook

Thanks to the advancement in the field of biotechnologies, the amount of data available increased exponentially allowing the analysis of complex diseases at a very detailed genomic level. However, this huge amount of data cannot be investigated without the use of proper tools designed for information mining. Bioinformatics plays a crucial role in combining large data with mathematical and statistical methods, offering completely new perspectives. Such an approach supports research that is progressively moving towards making clinical and therapeutic decisions as personalized as possible. Patients affected by complex diseases, especially in the case of cancer, are characterized by personal molecular hallmarks and even a same disease can be triggered by different causes and it can develop following different directions addressed by personal genetic heritage or environment.

With these three projects, we focus our attention on the validation and identification of biomarkers. The field of biomarker discovery is intensely working to assist early disease discovery with the final goal of timely and targeted intervention. Research studies in the field of translational oncology are more and more focused on omics and especially on multi-omics with the need to assist traditional approaches like histopathological examinations. Moreover, behind the combination of molecular data from multiple levels, there is the intention of a broader comprehension of the diseases. The identification and understanding of biomarkers are possible with a holistic approach including the observation of multiple molecular levels together with clinical data for an accurate overview of a patient. Moreover, molecules from different omics data might not be relevant singularly but their combination unravels unseen mechanisms. The study of genomics is the widest and the one with the highest application in clinical practice. Genes mutation and CNV have been intensively investigated and proposed as prognostic and diagnostic biomarkers and with the development of NGS techniques, transcriptomics and epigenomics are increasingly studied, with transcriptome as the most analyzed. Indeed, the analysis of altered expression levels can be informative to determine the mechanisms that triggered a disease. What initiated the disorder can be the consequence of numerous interplaying factors like the study of gene expression together with DNA methylation, proven to be a regulator of genes' activity. However, there is is the urge for a novel analytical approach for multi-omics investigation which allows the combination of more than only two or three omics data types, which is currently done for sake of interpretability and computational implementation. A possible approach could be to reduce the study field with the use of what is known and use it as a starting point to investigate novel information. Nevertheless, the analysis of multiple omic levels can still be considered an emerging field since the literature on the topic is still scarce. However, promising findings have been shown and intensive research on machine learning techniques will assist scientists in overcoming this challenge. Large-scale biological data gives the advantage of highly detailed information, however, it also brings the necessity to know how to gain meaningful information from it. Moreover, data is affected by noise and technical biases and it can be computationally expensive given the size. Machine learning contributes to quickly examining the big amount of data, while still considering important aspects such as a check of data quality, cleaning, and standardization, to allow the discovery of statistically robust and reliable findings.

One of the main challenges in the field of biomarker discovery is the clinical applicability of computationally discovered biomarkers, which is still low. Analyses are often not extensively validated and their reproducibility is critical. Important signs of progress are underway in deep learning and the analysis of extensive data in a multivariate setting is key to identifying biomarkers.

The motivation behind the development of these Ph.D. projects is the contribution that bioinformatics can give to precision oncology by understanding cancer biology thanks to the combination of machine learning techniques with omics data. The implementation of existing and new approaches helped us to make a step forward in the direction of personalized medicine. Despite all the progress made in biomedical research, cancer remains part of a complex system of internal and external factors and it should be studied in its totality. Biologists, clinicians, pathologists, and bioinformaticians work for the same purpose and we encourage them to closely collaborate, hoping to reach a moment when all these scientific figures actively work together to defeat one of the biggest medical challenges.

# Declaration of contribution

### Contribution to project 1 - An extensive computational evaluation of published gene signatures for PDAC subtyping

This project was started and executed by me as part of the Big Data in Biomedicine group at the Technical University of Munich, under the direct supervision of Dr. Markus List, Prof. Dr. Tim Kacprowski, Dr. med. vet. Katja Steiger and Nicole Pfarr. I designed and developed the analysis pipelines and implemented them. This work resulted in a publication, published in NAR Cancer on October 17th, 2022, under the Cancer Computational Biology section [113].

### Contribution to project 2 - Dimmer 2.0

I contributed to the development of Dimmer 2.0, an improved version of Dimmer [172] released in 2016 by Almeida and colleagues. This project was carried out in collaboration with Johannes Kersting and Alex Dietrich from the group of Big Data in Biomedicine under the direct supervision of Dr. Markus List. I contributed to testing the tool on novel generated data whilst collaborating on the tool improvement and expansion and testing the new functionalities.

### Contribution to project 3 - An analysis on the prognostic role and molecular profile of tumor budding in lung-SCC

This work was performed in collaboration with Dr. med. Melanie Boxberg and Dr. med. Fabian Stögbauer under the direct supervision of Prof. Dr. Tim Kacprowski. The work was divided into two projects, where in the first one I gave my contribution to identify a prognosis-relevant cut-off for the tumor budding trait and spread through air spaces variable in lung-SCC and validating it with survival analysis. This work was published on May 2nd, 2022 [134]. The second project aims to investigate the relationship between tumor budding and molecular features. In the second project, I designed and performed all the analysis and a manuscript is in preparation.

# List of Publications

1. **Lautizi, M.**, Baumbach, J., Weichert, W., Steiger, K., List, M., Pfarr, N., & Kacprowski, T. (2022). The limits of molecular signatures for pancreatic ductal adenocarcinoma subtyping. NAR cancer, 4(4), zcac030. https://doi.org/10.1093/narcan/zcac030

2. Lazareva, O.*, **Lautizi M.***, Fenn, A.*, List, M., Kacprowski, T., Baumbach, J. Multi-Omics Analysis in a Network Context. Systems Medicine. Elsevier; 2021. pp. 224-33. https://doi:10.1016/B978-0-12-801238-3.11647-2

3. Stalidzans, E., Zanin, M., Tieri, P., Castiglione, F., Polster, A., Scheiner, S., Pahle, J., Stres, B., List, M., Baumbach, J., **Lautizi, M.**, Schmidt, Harald H.H.W., 2020. Mechanistic modeling and multiscale applications for precision medicine: theory and practice. Network and Systems Medicine, 3(1), pp.36-56. https://doi:10.1089/nsm.2020.0002.

4. Stögbauer, F., **Lautizi, M.**, Kriegsmann, M., Winter, H., Muley, T., Kriegsmann, K., Jesinghaus, M., Baumbach, J., Schüffler, P., Weichert, W., Kacprowski, T., & Boxberg, M. (2022). Tumour cell budding and spread through air spaces in squamous cell carcinoma of the lung - Determination and validation of optimal prognostic cut-offs. Lung cancer (Amsterdam, Netherlands), 169, 1–12. https://doi.org/10.1016/j.lungcan.2022.04.012

[*] Shared first authorship.

# Abbreviations

| | |
|---|---|
| **DNA** | Deoxyribonucleic acid |
| **RNA** | Ribonucleic acid |
| **RSEM** | RNA-Seq by Expectation-Maximization |
| **ROC** | Receiver operating characteristic |
| **AUC** | Area under the ROC curve |
| **FFPE** | Formalin-Fixed Paraffin-Embedded |
| **H&E** | Hematoxylin and Eosin |
| **AJCC** | American Joint Committee on Cancer |
| **BRCA** | BReast CAncer gene |
| **TP53** | Tumor Protein P53 |
| **KRAS** | Kirsten rat sarcoma virus |
| **CDKN2A** | Cyclin Dependent Kinase Inhibitor 2A |
| **SMAD4** | Mothers against decapentaplegic homolog 4 |
| **FGFR** | Fibroblast growth factor receptors |
| **PIK3CA** | Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha |
| **EGFR** | Estimated glomerular filtration rate |
| **PTEN** | Phosphatase and tensin homolog |
| **APC** | Adenomatous polyposis coli |
| **BRAF** | v-raf murine sarcoma viral oncogene homolog B1 |
| **MAPK** | Mitogen-activated protein kinase |
| **PI3K** | Phosphoinositide 3-kinases |
| **mTOR** | Mammalian target of rapamycin |
| **MHC** | Major histocompatibility complex |
| **IL-1** | Interleukin-1 |
| **PD-L1** | Programmed death-ligand 1 |
| **PD-1** | Programmed cell death protein 1 |
| **CTLA-4** | Cytotoxic T-lymphocyte-associated protein 4 |
| **MT1A/G/H** | Metallothionein 1A/1G/1H |
| **SLC39A7** | Solute carrier family 39 member 7 |
| **NR5A2** | Nuclear Receptor Subfamily 5 Group A Member 2 |

| | |
|---|---|
| **B3GALT4** | Beta-1,3-galactosyltransferase 4 |
| **ST8SIA4/5** | ST8 Alpha-N-Acetyl-Neuraminide Alpha-2,8-Sialyltransferase 4/5 |
| **ST6GALNAC5** | ST6 N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 5 |
| **HOX** | Homeotic genes |
| **KCNN4** | Potassium Calcium-Activated Channel Subfamily N Member 4 |
| **MIG6** | Alias of ERRFI1 |
| **ERRFI1** | ERRFI1 |
| **EGFR** | Epidermal Growth Factor Receptor |
| **PHLPP** | PH Domain And Leucine Rich Repeat Protein Phosphatase 1 |
| **LCB4** | Sphingoid long chain base kinase 4 |
| **MTMR3/12** | Myotubularin Related Protein 3/12 |
| **INPP5D** | Inositol Polyphosphate-5-Phosphatase D |
| **PLCE1** | Phospholipase C Epsilon 1 |
| **PLCH1** | Phospholipase C Eta 1 |
| **PIK3C2B** | Phosphatidylinositol-4-Phosphate 3-Kinase Catalytic Subunit Type 2 Beta |
| **UGT1A6/7/8/9/10** | UDP Glucuronosyltransferase Family 1 Member A6/A7/A8/A9/A10 |
| **ADAMTS** | A disintegrin and metalloproteinase with thrombospondin motifs |

# Bibliography

[1] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome". en. In: *Nature* 458.7239 (Apr. 2009), p. 719.

[3] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". In: *cell* 144.5 (2011), pp. 646–674.

[4] D Hanahan and R A Weinberg. "The hallmarks of cancer". en. In: *Cell* 100.1 (Jan. 2000), pp. 57–70.

[5] Primo Schär. "Spontaneous DNA Damage, Genome Instability, and Cancer—When DNA Replication Escapes Control". en. In: *Cell* 104.3 (Feb. 2001), pp. 329–332.

[6] Prescott Deininger. "Genetic Instability in Cancer: Caretaker and Gatekeeper Genes". en. In: *Ochsner J.* 1.4 (Oct. 1999), p. 206.

[7] Li Wang et al. "Functional regulations between genetic alteration-driven genes and drug target genes acting as prognostic biomarkers in breast cancer". en. In: *Sci. Rep.* 12.1 (June 2022), pp. 1–14.

[8] "Key signal transduction pathways and crosstalk in cancer: Biological and therapeutic opportunities". In: *Transl. Oncol.* 26 (Dec. 2022), p. 101510.

[9] David A Frank. *Signal Transduction in Cancer*. Springer US.

[10] M A Dunham et al. "Telomere maintenance by recombination in human cells". In: *Nat. Genet.* 26.4 (Dec. 2000).

[11] Claire M Pfeffer and Amareshwar T K Singh. "Apoptosis: A Target for Anticancer Therapy". en. In: *Int. J. Mol. Sci.* 19.2 (Feb. 2018).

[12] Rebecca S Y Wong. "Apoptosis in cancer: from pathogenesis to treatment". en. In: *J. Exp. Clin. Cancer Res.* 30.1 (2011), p. 87.

[13] Benedito A Carneiro and Wafik S El-Deiry. "Targeting apoptosis in cancer therapy". en. In: *Nat. Rev. Clin. Oncol.* 17.7 (July 2020), p. 395.

[14] Naoyo Nishida et al. “Angiogenesis in Cancer”. en. In: *Vasc. Health Risk Manag.* 2.3 (Sept. 2006), p. 213.

[15] A W Lambert, D R Pattabiraman, and R A Weinberg. “Emerging Biological Principles of Metastasis”. In: *Cell* 168.4 (Feb. 2017).

[16] Federica Sotgia et al. “Understanding the Warburg effect and the prognostic value of stromal caveolin-1 as a marker of a lethal tumor microenvironment”. en. In: *Breast Cancer Res.* 13.4 (July 2011), pp. 1–13.

[17] Yasumasa Kato et al. “Acidic extracellular microenvironment and cancer”. en. In: *Cancer Cell Int.* 13.1 (Sept. 2013), pp. 1–8.

[18] T Soussi and K G Wiman. “TP53: an oncogene in disguise”. en. In: *Cell Death Differ.* 22.8 (May 2015), pp. 1239–1249.

[19] Victor D Li, Karen H Li, and John T Li. “TP53 mutations as potential prognostic markers for specific cancers: analysis of data from The Cancer Genome Atlas and the International Agency for Research on Cancer TP53 Database”. en. In: *J. Cancer Res. Clin. Oncol.* 145.3 (Mar. 2019), pp. 625–636.

[20] Ignacio A Rodriguez-Brenes, Dominik Wodarz, and Natalia L Komarova. “Quantifying replicative senescence as a tumor suppressor pathway and a target for cancer therapy”. en. In: *Sci. Rep.* 5.1 (Dec. 2015), pp. 1–13.

[21] M Burnet. “Cancer–A biological approach: I. the processes of control. II. The significance of somatic mutation”. en. In: *BMJ* 1.5022 (Apr. 1957), pp. 779–786.

[22] Daniel S Chen and Ira Mellman. “Oncology meets immunology: the cancer-immunity cycle”. en. In: *Immunity* 39.1 (July 2013), pp. 1–10.

[23] Hans Raskov et al. “Cytotoxic CD8+ T cells in cancer and cancer immunotherapy”. en. In: *Br. J. Cancer* 124.2 (Jan. 2021), pp. 359–367.

[24] Anoop Kallingal et al. “Cancer immune escape: the role of antigen presentation machinery”. en. In: *J. Cancer Res. Clin. Oncol.* 149.10 (Aug. 2023), pp. 8131–8141.

[25] Karthik Dhatchinamoorthy, Jeff D Colbert, and Kenneth L Rock. “Cancer immune evasion through loss of MHC class I antigen presentation”. en. In: *Front. Immunol.* 12 (Mar. 2021), p. 636568.

[26] Gulce Sari and Kenneth L Rock. “Tumor immune evasion through loss of MHC class-I antigen presentation”. en. In: *Curr. Opin. Immunol.* 83.102329 (Aug. 2023), p. 102329.

[27] Padmanee Sharma et al. “Primary, adaptive, and acquired resistance to cancer immunotherapy”. en. In: *Cell* 168.4 (Feb. 2017), pp. 707–723.

[28] Xianjie Jiang et al. “Role of the tumor microenvironment in PD-L1/PD-1-mediated tumor immune escape”. en. In: *Mol. Cancer* 18.1 (Dec. 2019).

[29] Alireza Labani-Motlagh, Mehrnoush Ashja-Mahdavi, and Angelica Loskog. “The tumor microenvironment: A milieu hindering and obstructing antitumor immune responses”. en. In: *Front. Immunol.* 11 (May 2020), p. 940.

[30] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. “Multi-omics approaches to disease”. en. In: *Genome Biol.* 18.1 (May 2017), pp. 1–15.

[31] Ana Conesa and Stephan Beck. “Making multi-omics data accessible to researchers”. en. In: *Scientific Data* 6.1 (Oct. 2019), pp. 1–4.

[32] Emil Uffelmann et al. “Genome-wide association studies”. en. In: *Nature Reviews Methods Primers* 1.1 (Aug. 2021), pp. 1–21.

[33] M S Kris A. Wetterstrand. *The Cost of Sequencing a Human Genome*. en. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost. Accessed: 2023-1-6. Mar. 2019.

[34] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4 et al. “Initial sequencing and analysis of the human genome”. In: *nature* 409.6822 (2001), pp. 860–921.

[35] Andrea Sboner et al. “The real cost of sequencing: higher than you think!” en. In: *Genome Biol.* 12.8 (Aug. 2011), pp. 1–10.

[36] Bob Weinhold. “Epigenetics: The Science of Change”. en. In: *Environ. Health Perspect.* (2006).

[37] Martin C Frith, Michael Pheasant, and John S Mattick. “Genomics: The amazing complexity of the human transcriptome”. en. In: *Eur. J. Hum. Genet.* 13.8 (June 2005), pp. 894–897.

[38] Philip D Glaves and Jonathan D Tugwood. “Generation and Analysis of Transcriptomics Data”. en. In: *Drug Safety Evaluation* (2011), pp. 167–185.

[39] Stanislaw Supplitt et al. “Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine”. en. In: *Int. J. Mol. Sci.* 22.3 (Jan. 2021), p. 1422.

[40] Akhilesh Pandey and Matthias Mann. “Proteomics to study genes and genomes”. en. In: *Nature* 405.6788 (June 2000), pp. 837–846.

[41] Sarah Franklin and Thomas M Vondriska. “Genomes, Proteomes and the Central Dogma”. en. In: *Circ. Cardiovasc. Genet.* 4.5 (Oct. 2011), p. 576.

[42] Nancy Wiebelhaus et al. “Protein folding stability changes across the proteome reveal targets of Cu toxicity in E. coli”. en. In: *ACS Chem. Biol.* 16.1 (Jan. 2021), p. 214.

[43] Fabian Birzele, Gergely Csaba, and Ralf Zimmer. “Alternative splicing and protein structure evolution”. en. In: *Nucleic Acids Res.* 36.2 (Feb. 2008), p. 550.

[44] Penghao Wang and Susan R Wilson. “Mass spectrometry-based protein identification by integrating de novo sequencing with database searching”. en. In: *BMC Bioinformatics* 14.Suppl 2 (2013), S24.

[45] Clary B Clish. “Metabolomics: an emerging but powerful tool for precision medicine”. en. In: *Cold Spring Harbor Molecular Case Studies* 1.1 (Oct. 2015).

[46] Shira Shaham-Niv, Sigal Rencus-Lazar, and Ehud Gazit. “Metabolite medicine offers a path beyond lists of metabolites”. en. In: *Communications Chemistry* 4.1 (Aug. 2021), pp. 1–5.

[47] Luke K Ursell et al. “Defining the human microbiome”. en. In: *Nutr. Rev.* 70.suppl_1 (Aug. 2012), S38–S44.

[48] Willem M de Vos et al. “Gut microbiome and health: mechanistic insights”. en. In: *Gut* 71.5 (May 2022), pp. 1020–1032.

[49] Juliana Durack and Susan V Lynch. “The gut microbiome: Relationships with disease and opportunities for therapy”. en. In: *J. Exp. Med.* 216.1 (Jan. 2019), p. 20.

[50] Djawad Radjabzadeh et al. “Gut microbiome-wide association study of depressive symptoms”. en. In: *Nat. Commun.* 13.1 (Dec. 2022), pp. 1–10.

[51] “The role of the gut microbiome in the development of schizophrenia”. In: *Schizophr. Res.* 234 (Aug. 2021), pp. 4–23.

[52] Rajeshwar Govindarajan et al. “Microarray and its applications”. en. In: *J. Pharm. Bioallied Sci.* 4.Suppl 2 (Aug. 2012), S310.

[53] Fiona Cunningham et al. “Ensembl 2022”. en. In: *Nucleic Acids Res.* 50.D1 (Nov. 2021), pp. D988–D995.

[54] Olivier Harismendy et al. “Evaluation of next generation sequencing platforms for population targeted sequencing studies”. en. In: *Genome Biol.* 10.3 (Mar. 2009), pp. 1–13.

[55] Yu Cao et al. “A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies”. en. In: *Front. Microbiol.* 8 (Sept. 2017).

[56] Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. “Overview of Next Generation Sequencing Technologies”. en. In: *Curr. Protoc. Mol. Biol.* 122.1 (Apr. 2018), e59.

[57] Milan Fedurco et al. “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies”. en. In: *Nucleic Acids Res.* 34.3 (Jan. 2006), e22–e22.

[58] Gerardo Turcatti et al. “A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis”. en. In: *Nucleic Acids Res.* 36.4 (Feb. 2008), e25–e25.

[59] Céline Adessi et al. “Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms”. en. In: *Nucleic Acids Res.* 28.20 (Oct. 2000), e87–e87.

[60] Lisa D Moore, Thuc Le, and Guoping Fan. “DNA Methylation and Its Basic Function”. en. In: *Neuropsychopharmacology* 38.1 (July 2012), pp. 23–38.

[61] Bilian Jin, Yajun Li, and Keith D Robertson. “DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?” en. In: *Genes Cancer* 2.6 (June 2011), p. 607.

[62] Gaurab Aditya Dhar et al. “DNA methylation and regulation of gene expression: Guardian of our health”. en. In: *Nucleus* 64.3 (Aug. 2021), pp. 259–270.

[63] Manel Esteller. “CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future”. en. In: *Oncogene* 21.35 (Aug. 2002), pp. 5427–5440.

[64] Jae-Won Cho et al. “The importance of enhancer methylation for epigenetic regulation of tumorigenesis in squamous lung cancer”. en. In: *Exp. Mol. Med.* 54.1 (Jan. 2022), pp. 12–22.

[65] Madeleine P Ball et al. “Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells”. en. In: *Nat. Biotechnol.* 27.4 (Mar. 2009), pp. 361–368.

[66] H Cui et al. “Loss of imprinting in colorectal cancer linked to hypomethylation of H19 and IGF2”. In: *Cancer Res.* 62.22 (Nov. 2002).

[67] G A Ulaner et al. “Loss of imprinting of IGF2 and H19 in osteosarcoma is accompanied by reciprocal methylation changes of a CTCF-binding site”. In: *Hum. Mol. Genet.* 12.5 (Mar. 2003).

[68] S K Murphy et al. “Frequent IGF2/H19 domain epigenetic alterations and elevated IGF2 expression in epithelial ovarian cancer”. In: *Mol. Cancer Res.* 4.4 (Apr. 2006).

[69] W Timp and A P Feinberg. “Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host”. In: *Nat. Rev. Cancer* 13.7 (July 2013).

[70] Jean-Pierre J Issa et al. “Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon”. en. In: *Nat. Genet.* 7.4 (Aug. 1994), pp. 536–540.

[71] Daniel J Weisenberger et al. “CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer”. en. In: *Nat. Genet.* 38.7 (June 2006), pp. 787–793.

[72] Lixin Dong et al. “Genome-wide Analysis Reveals DNA Methylation Alterations in Obesity Associated with High Risk of Colorectal Cancer”. en. In: *Sci. Rep.* 9.1 (Mar. 2019), pp. 1–11.

[73] Li Zhou et al. “Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing”. en. In: *Sci. Rep.* 9 (2019).

[74] Cecilia Lövkvist et al. "DNA methylation in human epigenomes depends on local topology of CpG sites". en. In: *Nucleic Acids Res.* 44.11 (June 2016), pp. 5123–5132.

[75] Sebastian Moran, Carles Arribas, and Manel Esteller. "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences". en. In: *Epigenomics* 8.3 (Mar. 2016), pp. 389–399.

[76] Nan Lin et al. "Genome-wide DNA methylation profiling in human breast tissue by Illumina TruSeq methyl capture EPIC sequencing and infinium methylationEPIC beadchip microarray". en. In: *Epigenetics* 16.7 (2021), pp. 754–769.

[77] Ai Ling Teh et al. "Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples". en. In: *Epigenetics* 11.1 (Jan. 2016), p. 36.

[78] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". en. In: *Genome Biol.* 17 (2016).

[79] Eoghan R Malone et al. "Molecular profiling for precision cancer therapies". en. In: *Genome Med.* 12.1 (Jan. 2020), pp. 1–19.

[80] *National Cancer Institute, http://www.cancer.gov/dictionary?CdrID=45618*.

[81] Abigail Shaw et al. "Tumour biomarkers: diagnostic, prognostic, and predictive". In: *BMJ* 351 (2015).

[82] David Venet, Jacques E Dumont, and Vincent Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". en. In: *PLoS Comput. Biol.* 7.10 (Oct. 2011), e1002240.

[83] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[84] David R Cox. "Regression Models and Life-Tables". en. In: *Breakthroughs in Statistics* (1992), pp. 527–541.

[85] T Sørlie et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 98.19 (Sept. 2001), pp. 10869–10874.

[86] Laetitia Marisa et al. "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value". en. In: *PLoS Med.* 10.5 (May 2013), e1001453.

[87] Felipe De Sousa E Melo et al. "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions". In: *Nature medicine* 19.5 (2013), pp. 614–618.

[88] Richard A Moffitt et al. “Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma”. en. In: *Nat. Genet.* 47.10 (Oct. 2015), pp. 1168–1178.

[89] Eric A Collisson et al. “Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy”. en. In: *Nat. Med.* 17.4 (Apr. 2011), pp. 500–503.

[90] Peter Bailey et al. “Genomic analyses identify molecular subtypes of pancreatic cancer”. en. In: *Nature* 531.7592 (Mar. 2016), pp. 47–52.

[91] Francesco Puleo et al. “Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features”. en. In: *Gastroenterology* 155.6 (Dec. 2018), 1999–2013.e3.

[92] C M Perou et al. “Molecular portraits of human breast tumours”. en. In: *Nature* 406.6797 (Aug. 2000), pp. 747–752.

[93] Christina Curtis et al. “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups”. en. In: *Nature* 486.7403 (Apr. 2012), pp. 346–352.

[94] Jeyapradha Duraiyan et al. “Applications of immunohistochemistry”. en. In: *J. Pharm. Bioallied Sci.* 4.Suppl 2 (Aug. 2012), S307.

[95] Fieke E M Froeling et al. “Molecular Subtyping and Precision Medicine for Pancreatic Cancer”. en. In: *J. Clin. Med. Res.* 10.1 (Jan. 2021).

[96] Jiangang Liu et al. “Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data”. en. In: *BMC Med. Genomics* 1.1 (Sept. 2008), pp. 1–12.

[97] Laura Cantini et al. “Classification of gene signatures for their information value and functional redundancy”. en. In: *npj Systems Biology and Applications* 4.1 (Dec. 2017), pp. 1–11.

[98] Francisco Azuaje, Yvan Devaux, and Daniel Wagner. “Challenges and standards in reporting diagnostic and prognostic biomarker studies”. en. In: *Clin. Transl. Sci.* 2.2 (Apr. 2009), pp. 156–161.

[99] Sandra Ng et al. “The benefits and pitfalls of machine learning for biomarker discovery”. en. In: *Cell Tissue Res.* 394.1 (Oct. 2023), pp. 17–31.

[100] Michael Olivier et al. “The Need for Multi-Omics Biomarker Signatures in Precision Medicine”. In: *International Journal of Molecular Sciences* 20.19 (2019). ISSN: 1422-0067. DOI: 10.3390/ijms20194781. URL: https://www.mdpi.com/1422-0067/20/19/4781.

[101] Milan Picard et al. “Integration strategies of multi-omics data for machine learning analysis”. en. In: *Comput. Struct. Biotechnol. J.* 19 (June 2021), pp. 3735–3746.

[102] Ramon Diaz-Uriarte et al. “Ten quick tips for biomarker discovery and validation analyses using machine learning”. en. In: *PLoS Comput. Biol.* 18.8 (Aug. 2022), e1010357.

[103] Vivien Marx. “The big challenges of big data”. In: *Nature* 498.7453 (2013), pp. 255–260.

[104] Jason E McDermott et al. “Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data”. en. In: *Expert Opin. Med. Diagn.* 7.1 (Jan. 2013), pp. 37–51.

[105] Lin Shi et al. “Variable selection and validation in multivariate modelling”. en. In: *Bioinformatics* 35.6 (Mar. 2019), pp. 972–980.

[106] David Venet, Jacques E Dumont, and Vincent Detours. “Most random gene expression signatures are significantly associated with breast cancer outcome”. en. In: *PLoS Comput. Biol.* 7.10 (Oct. 2011), e1002240.

[107] Panagiotis Sarantis et al. “Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy”. en. In: *World J. Gastrointest. Oncol.* 12.2 (Feb. 2020), pp. 173–181.

[108] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. “Cancer statistics, 2018”. In: *CA: a cancer journal for clinicians* 68.1 (2018), pp. 7–30.

[109] Axel Bengtsson, Roland Andersson, and Daniel Ansari. “The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data”. en. In: *Sci. Rep.* 10.1 (Oct. 2020), p. 16425.

[110] Andrew Cannon et al. “Desmoplasia in pancreatic ductal adenocarcinoma: insight into pathological function and therapeutic potential”. en. In: *Genes Cancer* 9.3-4 (Mar. 2018), p. 78.

[111] “Translating complexity and heterogeneity of pancreatic tumor: 3D in vitro to in vivo models”. In: *Adv. Drug Deliv. Rev.* 174 (July 2021), pp. 265–293.

[112] María Laura Gutiérrez, Luis Muñoz-Bellvís, and Alberto Orfao. “Genomic Heterogeneity of Pancreatic Ductal Adenocarcinoma and Its Clinical Impact”. en. In: *Cancers* 13.17 (Sept. 2021), p. 4451.

[113] Manuela Lautizi et al. “The limits of molecular signatures for pancreatic ductal adenocarcinoma subtyping”. en. In: *NAR Cancer* 4.4 (Oct. 2022), zcac030.

[114] David J Birnbaum et al. “Validation and comparison of the molecular classifications of pancreatic carcinomas”. en. In: *Mol. Cancer* 16.1 (Nov. 2017), p. 168.

[115] Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu and Cancer Genome Atlas Research Network. “Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma”. en. In: *Cancer Cell* 32.2 (Aug. 2017), 185–203.e13.

[116] Naim U Rashid et al. “Purity Independent Subtyping of Tumors (PurIST), A Clinically Robust, Single-sample Classifier for Tumor Subtyping in Pancreatic Cancer”. en. In: *Clin. Cancer Res.* 26.1 (Jan. 2020), pp. 82–92.

[117] Rekin's Janky et al. "Prognostic relevance of molecular subtypes and master regulators in pancreatic ductal adenocarcinoma". en. In: *BMC Cancer* 16 (Aug. 2016), p. 632.

[118] Carlo Maurer et al. "Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes". en. In: *Gut* 68.6 (June 2019), pp. 1034–1043.

[119] Shivan Sivakumar et al. "Master Regulators of Oncogenic KRAS Response in Pancreatic Cancer: An Integrative Network Biology Analysis". en. In: *PLoS Med.* 14.1 (Jan. 2017), e1002223.

[120] Henry C-H Law et al. "The Proteomic Landscape of Pancreatic Ductal Adenocarcinoma Liver Metastases Identifies Molecular Subtypes and Associations with Clinical ResponseProteomics of Pancreatic Ductal Adenocarcinoma Metastases". In: *Clinical Cancer Research* 26.5 (2020), pp. 1065–1076.

[121] James T Topham et al. "Subtype-Discordant Pancreatic Ductal Adenocarcinoma Tumors Show Intermediate Clinical and Molecular Characteristics". en. In: *Clin. Cancer Res.* 27.1 (Jan. 2021), pp. 150–157.

[122] Michelle Chan-Seng-Yue et al. "Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution". en. In: *Nat. Genet.* 52.2 (Feb. 2020), pp. 231–240.

[123] Rémy Nicolle et al. "Establishment of a pancreatic adenocarcinoma molecular gradient (PAMG) that predicts the clinical outcome of pancreatic cancer". en. In: *EBioMedicine* 57 (July 2020), p. 102858.

[124] Thara Tunthanathip et al. "Multiple, Primary Brain Tumors with Diverse Origins and Different Localizations: Case Series and Review of the Literature". en. In: *J. Neurosci. Rural Pract.* 9.4 (2018), p. 593.

[125] Zhiwei Shao et al. "Molecular Mechanism and Approach in Progression of Meningioma". en. In: *Front. Oncol.* 10 (2020).

[126] Patrick N Harter, Yannick Braun, and Karl H Plate. "Classification of meningiomas—advances and controversies". en. In: *Chinese Clinical Oncology* 6.Suppl 1 (July 2017), S2–S2.

[127] Joseph Wiemels, Margaret Wrensch, and Elizabeth B Claus. "Epidemiology and etiology of meningioma". en. In: *J. Neurooncol.* 99.3 (Sept. 2010), pp. 307–314.

[128] Gabrielle Truitt et al. "Partnership for defining the impact of 12 selected rare CNS tumors: a report from the CBTRUS and the NCI-CONNECT". en. In: *J. Neurooncol.* 144.1 (June 2019), pp. 53–63.

[129] W T Longstreth Jr et al. "Dental X-rays and the risk of intracranial meningioma: a population-based case-control study". en. In: *Cancer* 100.5 (Mar. 2004), pp. 1026–1034.

[130] L E Phillips et al. "History of head trauma and risk of intracranial meningioma: Population-based case-control study". en. In: *Neurology* 58.12 (June 2002), pp. 1849–1852.

[131] Nobuko Hijiya et al. "Cumulative incidence of secondary neoplasms as a first event after childhood acute lymphoblastic leukemia". en. In: *JAMA* 297.11 (Mar. 2007), pp. 1207–1215.

[132] Philip Ryan et al. "Amalgam fillings, diagnostic dental x-rays and tumours of the brain and meninges". In: *European Journal of Cancer Part B: Oral Oncology* 28.2 (1992), pp. 91–95.

[133] Siegal Sadetzki et al. "Radiation-induced meningioma: a descriptive study of 253 cases". en. In: *J. Neurosurg.* 97.5 (Nov. 2002), pp. 1078–1082.

[134] Fabian Stögbauer et al. "Tumour cell budding and spread through air spaces in squamous cell carcinoma of the lung–Determination and validation of optimal prognostic cut-offs". In: *Lung Cancer* 169 (2022), pp. 1–12.

[135] Christina Neppl et al. "Validation of the International Tumor Budding Consensus Conference (ITBCC) 2016 recommendation in squamous cell carcinoma of the lung—a single-center analysis of 354 cases". en. In: *Mod. Pathol.* 33.5 (Dec. 2019), pp. 802–811.

[136] Ryota Masuda et al. "Tumor budding is a significant indicator of a poor prognosis in lung squamous cell carcinoma patients". en. In: *Mol. Med. Rep.* 6.5 (Nov. 2012), p. 937.

[137] Li Qian et al. "Potential key roles of tumour budding: a representative malignant pathological feature of non-small cell lung cancer and a sensitive indicator of prognosis". en. In: *BMJ Open* 12.3 (Mar. 2022), e054009.

[138] Justine Fan, Samuel M DeFina, and He Wang. "Prognostic Value of Selected Histologic Features for Lung Squamous Cell Carcinoma". en. In: *Exploratory Research and Hypothesis in Medicine* 0.000 (Mar. 2022), pp. 0–0.

[139] Inti Zlobec and Alessandro Lugli. "Tumour budding in colorectal cancer: molecular rationale for clinical translation". en. In: *Nat. Rev. Cancer* 18.4 (Apr. 2018), pp. 203–204.

[140] AC Rogers et al. "Systematic review and meta-analysis of the impact of tumour budding in colorectal cancer". In: *British journal of cancer* 115.7 (2016), pp. 831–840.

[141] Linde De Smedt et al. "Expression profiling of budding cells in colorectal cancer reveals an EMT-like phenotype and molecular subtype switching". en. In: *Br. J. Cancer* 116.1 (Nov. 2016), pp. 58–65.

[142] Kate O'Connor et al. "Tumor budding is an independent adverse prognostic factor in pancreatic ductal adenocarcinoma". In: *The American journal of surgical pathology* 39.4 (2015), pp. 472–478.

[143] Sulen Sarioglu et al. "Tumor budding as a prognostic marker in laryngeal carcinoma". In: *Pathology-Research and Practice* 206.2 (2010), pp. 88–92.

[144] Xiaowei Lai et al. "Epithelial-Mesenchymal Transition and Metabolic Switching in Cancer: Lessons From Somatic Cell Reprogramming". en. In: *Front. Cell Dev. Biol.* 8 (Aug. 2020).

[145] Linde De Smedt et al. “Expression profiling of budding cells in colorectal cancer reveals an EMT-like phenotype and molecular subtype switching”. In: *British journal of cancer* 116.1 (2017), pp. 58–65.

[146] Alessandro Lugli et al. “Tumour budding in solid cancers”. In: *Nature Reviews Clinical Oncology* 18.2 (2021), pp. 101–115.

[147] Alexandru Dan Grigore et al. “Tumor budding: the name is EMT. Partial EMT.” In: *Journal of clinical medicine* 5.5 (2016), p. 51.

[148] Melanie Boxberg et al. “Tumour budding activity and cell nest size determine patient outcome in oral squamous cell carcinoma: proposal for an adjusted grading system”. In: *Histopathology* 70.7 (2017), pp. 1125–1137.

[149] Melanie Boxberg et al. “Tumor budding and cell nest size are highly prognostic in laryngeal and hypopharyngeal squamous cell carcinoma”. In: *The American journal of surgical pathology* 43.3 (2019), pp. 303–313.

[150] Zhao Chen et al. “Non-small-cell lung cancers: a heterogeneous set of diseases”. en. In: *Nat. Rev. Cancer* 14.8 (July 2014), pp. 535–546.

[151] “Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis”. In: *Lung Cancer* 31.2-3 (Mar. 2001), pp. 139–148.

[152] “Clinicopathologic Features of Advanced Squamous NSCLC”. In: *J. Thorac. Oncol.* 11.9 (Sept. 2016), pp. 1411–1422.

[153] Cancer Genome Atlas Research Network et al. “Comprehensive genomic characterization of squamous cell lung cancers”. In: *Nature* 489.7417 (2012), p. 519.

[154] Pablo Perez-Moreno et al. “Squamous cell carcinoma of the lung: molecular subtypes and therapeutic opportunities”. en. In: *Clin. Cancer Res.* 18.9 (May 2012), pp. 2443–2451.

[155] Brian O’Sullivan et al. “The TNM classification of malignant tumours—towards common understanding and reasonable expectations”. In: *The Lancet Oncology* 18.7 (2017), pp. 849–851.

[156] Mahul B Amin et al. “The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging”. In: *CA: a cancer journal for clinicians* 67.2 (2017), pp. 93–99.

[157] Kyuichi Kadota et al. “Tumor spread through air spaces is an important pattern of invasion and impacts the frequency and location of recurrences after limited resection for small stage I lung adenocarcinomas”. In: *Journal of Thoracic Oncology* 10.5 (2015), pp. 806–814.

[158] Huining Liu et al. “Prognostic impact of tumor spread through air spaces in non-small cell lung cancers: a meta-analysis including 3564 patients”. In: *Pathology & Oncology Research* 25.4 (2019), pp. 1303–1310.

[159] Hironori Uruga et al. “Semiquantitative assessment of tumor spread through air spaces (STAS) in early-stage lung adenocarcinomas”. In: *Journal of Thoracic Oncology* 12.7 (2017), pp. 1046–1051.

[160] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. “Detection of significantly differentially methylated regions in targeted bisulfite sequencing data”. en. In: *Bioinformatics* 29.13 (May 2013), pp. 1647–1653.

[161] Martin J Aryee et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. en. In: *Bioinformatics* 30.10 (May 2014), pp. 1363–1369.

[162] A E Jaffe et al. “Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies”. In: *Int. J. Epidemiol.* 41.1 (Feb. 2012).

[163] Yongseok Park et al. “MethylSig: a whole genome DNA methylation analysis pipeline”. en. In: *Bioinformatics* 30.17 (May 2014), pp. 2414–2422.

[164] Frank Jühling et al. “metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data”. en. In: *Genome Res.* 26.2 (Feb. 2016), p. 256.

[165] Kamal Kishore et al. “methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data”. en. In: *BMC Bioinformatics* 16.1 (Sept. 2015), pp. 1–11.

[166] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. “BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions”. en. In: *Genome Biol.* 13.10 (Oct. 2012), pp. 1–10.

[167] John M Gaspar and Ronald P Hart. “DMRfinder: efficiently identifying differentially methylated regions from MethylC-seq data”. en. In: *BMC Bioinformatics* 18.1 (Nov. 2017), pp. 1–8.

[168] Altuna Akalin et al. “methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles”. en. In: *Genome Biol.* 13.10 (Oct. 2012), pp. 1–9.

[169] Timothy J Peters et al. “Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate”. en. In: *Nucleic Acids Res.* 49.19 (Nov. 2021), e109.

[170] Yunshun Chen et al. “Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR”. en. In: *F1000Res.* 6 (2017).

[171] Fabian Müller et al. “RnBeads 2.0: comprehensive analysis of DNA methylation data”. en. In: *Genome Biol.* 20.1 (Mar. 2019), pp. 1–12.

[172] Diogo Almeida et al. “Efficient detection of differentially methylated regions using DiMmeR”. en. In: *Bioinformatics* 33.4 (Nov. 2016), pp. 549–551.

[173] Tetsuhiko Taira et al. “Characterization of the immunophenotype of the tumor budding and its prognostic implications in squamous cell carcinoma of the lung”. In: *Lung Cancer* 76.3 (2012), pp. 423–430.

[174] Liviu Badea et al. “Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia”. en. In: *Hepatogastroenterology* 55.88 (Nov. 2008), pp. 2016–2027.

[175] V Sandhu et al. “Molecular signatures of mRNAs and miRNAs as prognostic biomarkers in pancreatobiliary and intestinal types of periampullary adenocarcinomas”. en. In: *Mol. Oncol.* 9.4 (Apr. 2015), pp. 758–771.

[176] Shouhui Yang et al. “A Novel MIF Signaling Pathway Drives the Malignant Character of Pancreatic Cancer by Targeting NR3C2”. en. In: *Cancer Res.* 76.13 (July 2016), pp. 3838–3850.

[177] Mary J Goldman et al. “Visualizing and interpreting cancer genomics data via the Xena platform”. en. In: *Nat. Biotechnol.* 38.6 (June 2020), pp. 675–678.

[178] David Capper et al. “Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience”. en. In: *Acta Neuropathol.* 136.2 (July 2018), pp. 181–210.

[179] Shona Hendry et al. “Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors”. In: *Advances in anatomic pathology* 24.6 (2017), p. 311.

[180] Wilko Weichert et al. “Proposal of a prognostically relevant grading scheme for pulmonary squamous cell carcinoma”. In: *European Respiratory Journal* 47.3 (2016), pp. 938–946.

[181] Matthew D Wilkerson et al. “Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types”. en. In: *Clin. Cancer Res.* 16.19 (Sept. 2010), pp. 4864–4875.

[182] Astrid Schneider, Gerhard Hommel, and Maria Blettner. “Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications”. en. In: *Deutsches Ärzteblatt International* 107.44 (Nov. 2010), p. 776.

[183] Stacey J Shaefer and Louis Theodore. *REG. Regression Analysis: Method of Least Squares*. 2007.

[184] Leo Breiman. “Random Forests”. en. In: *Mach. Learn.* 45.1 (Oct. 2001), pp. 5–32.

[185] Ioana Popa-Burke et al. “The Effect of Initial Purity on the Stability of Solutions in Storage”. en. In: *SLAS Discovery* 19.2 (Feb. 2014), pp. 308–316.

[186] Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio. "Hierarchical Agglomerative Clustering". en. In: *Encyclopedia of Systems Biology* (2013), pp. 886–887.

[187] Adji Achmad Rinaldo Fernandes and Solimun Solimun. "Comparison of the Use of Linkage in Cluster Integration With Path Analysis Approach". en. In: *Front. Appl. Math. Stat.* 8 (Aug. 2022).

[188] Joel S Parker et al. "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes". en. In: *J. Clin. Oncol.* 27.8 (Mar. 2009), p. 1160.

[189] Paul Roepman et al. "Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition". en. In: *Int. J. Cancer* 134.3 (Feb. 2014), pp. 552–562.

[190] D D Lee and H S Seung. "Learning the parts of objects by non-negative matrix factorization". en. In: *Nature* 401.6755 (Oct. 1999), pp. 788–791.

[191] Anguraj Sadanandam et al. "A colorectal cancer classification system that associates cellular phenotype and responses to therapy". en. In: *Nat. Med.* 19.5 (May 2013), p. 619.

[192] Luca Tosti et al. "Single-Nucleus and In Situ RNA-Sequencing Reveal Cell Topographies in the Human Pancreas". en. In: *Gastroenterology* 160.4 (Mar. 2021), 1330–1344.e11.

[193] Aaron M Newman et al. "Robust enumeration of cell subsets from tissue expression profiles". en. In: *Nat. Methods* 12.5 (May 2015), pp. 453–457.

[194] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (Oct. 2005), pp. 15545–15550.

[195] Vamsi K Mootha et al. "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes". en. In: *Nat. Genet.* 34.3 (July 2003), pp. 267–273.

[196] Maxim V Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". en. In: *Nucleic Acids Res.* 44.W1 (July 2016), W90–7.

[197] Minoru Kanehisa et al. "New approach for understanding genome variations in KEGG". en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D590–D595.

[198] Bijay Jassal et al. "The reactome pathway knowledgebase". en. In: *Nucleic Acids Res.* 48.D1 (Jan. 2020), pp. D498–D503.

[199] M Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". en. In: *Nat. Genet.* 25.1 (May 2000), pp. 25–29.

[200] Manish Bhattacharjee, Sunil K Dhar, and Sundarraman Subramanian. *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*. en. World Scientific, Mar. 2011.

[201] Peter H Westfall and S Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. en. John Wiley & Sons, Jan. 1993.

[202] E Rödel. *Fisher, RA: Statistical methods for research workers, 14. Aufl., Oliver & Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 s*. 1971.

[203] Nina Baumgarten et al. "EpiRegio: analysis and retrieval of regulatory elements linked to genes". en. In: *Nucleic Acids Res.* 48.W1 (May 2020), W193–W199.

[204] Uku Raudvere et al. "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". en. In: *Nucleic Acids Res.* 47.W1 (July 2019), W191–W198.

[205] Y Tanigawa, E S Dyer, and G Bejerano. "WhichTF is functionally important in your open chromatin data?" In: *PLoS Comput. Biol.* 18.8 (Aug. 2022).

[206] Bin Wang, John M Cunningham, and Xinan (holly) Yang. "Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data". en. In: *Bioinformatics* 31.18 (Sept. 2015), p. 3043.

[207] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python". en. In: *Bioinformatics* (Nov. 2022), btac757.

[208] Marvin Martens et al. "WikiPathways: connecting communities". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D613–D621. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1024. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D613/35364599/gkaa1024.pdf. URL: https://doi.org/10.1093/nar/gkaa1024.

[209] Ruili Huang et al. "The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics". In: *Frontiers in Pharmacology* 10 (2019). ISSN: 1663-9812. DOI: 10.3389/fphar.2019.00445. URL: https://www.frontiersin.org/articles/10.3389/fphar.2019.00445.

[210] Levi Waldron et al. "Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer". en. In: *J. Natl. Cancer Inst.* 106.5 (Apr. 2014).

[211] Elisa M Noll et al. "CYP3A5 mediates basal and acquired therapy resistance in different subtypes of pancreatic ductal adenocarcinoma". en. In: *Nat. Med.* 22.3 (Mar. 2016), pp. 278–287.

[212] Melinda Wang et al. "Zinc: Roles in pancreatic physiology and disease". en. In: *Pancreatology* 20.7 (Oct. 2020), p. 1413.

[213] Manfei Si and Jinghe Lang. "The roles of metallothioneins in carcinogenesis". en. In: *J. Hematol. Oncol.* 11 (2018).

[214] Kai Li et al. “Metallothionein-1G suppresses pancreatic cancer cell stemness by limiting activin A secretion via NF-$\kappa$B inhibition”. en. In: *Theranostics* 11.7 (2021), p. 3196.

[215] Feng Zhang et al. “Integrative Analysis of Metallothioneins Identifies MT1H as Candidate Prognostic Biomarker in Hepatocellular Carcinoma”. en. In: *Front. Mol. Biosci.* 0 (2021).

[216] Koji Taniguchi and Michael Karin. “NF-$\kappa$B, inflammation, immunity and cancer: coming of age”. en. In: *Nat. Rev. Immunol.* 18.5 (Jan. 2018), pp. 309–324.

[217] Nengquan Sheng et al. “Knockdown of SLC39A7 inhibits cell growth and induces apoptosis in human colorectal cancer cells”. en. In: *Acta Biochim. Biophys. Sin.* 49.10 (Sept. 2017), pp. 926–934.

[218] Limei Liu, Jiaomin Yang, and Chao Wang. “Analysis of the prognostic significance of solute carrier (SLC) family 39 genes in breast cancer”. en. In: *Biosci. Rep.* 40.8 (Aug. 2020).

[219] Hui Yu et al. “Targeting NF-$\kappa$B pathway for the therapy of diseases: mechanism and clinical study”. en. In: *Signal Transduction and Targeted Therapy* 5.1 (Sept. 2020), pp. 1–23.

[220] Feng Guo et al. “NR5A2 transcriptional activation by BRD4 promotes pancreatic cancer progression by upregulating GDF15”. en. In: *Cell Death Discovery* 7.1 (Apr. 2021), pp. 1–15.

[221] Sophie Groux-Degroote and Philippe Delannoy. “Cancer-Associated Glycosphingolipids as Tumor Markers and Targets for Cancer Immunotherapy”. en. In: *Int. J. Mol. Sci.* 22.11 (June 2021), p. 6145.

[222] Norihiko Sasaki et al. “Ganglioside GM2, highly expressed in the MIA PaCa-2 pancreatic ductal adenocarcinoma cell line, is correlated with growth, invasion, and advanced stage”. en. In: *Sci. Rep.* 9.1 (Dec. 2019), pp. 1–13.

[223] T R Sarkar et al. “GD3 synthase regulates epithelial-mesenchymal transition and metastasis in breast cancer”. en. In: *Oncogene* 34.23 (June 2015), pp. 2958–2967.

[224] Dong Hoon Kwak et al. “Relationship between ganglioside expression and anti-cancer effects of the monoclonal antibody against epithelial cell adhesion molecule in colon cancer”. en. In: *Exp. Mol. Med.* 43.12 (Dec. 2011), pp. 693–701.

[225] Shih-Chi Yeh et al. “Glycolipid GD3 and GD3 synthase are key drivers for glioblastoma stem cells and tumorigenicity”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.20 (May 2016), pp. 5592–5597.

[226] “Glycosphingolipids are mediators of cancer plasticity through independent signaling pathways”. In: *Cell Rep.* 40.7 (Aug. 2022), p. 111181.

[227] Khoa Nguyen et al. “ST8SIA1 regulates tumor growth and metastasis in TNBC by activating the FAK-AKT-mTOR signaling pathway”. en. In: *Mol. Cancer Ther.* 17.12 (Dec. 2018), p. 2689.

[228] Xiaolu Ma et al. "Functional roles of sialylation in breast cancer progression through miR-26a/26b targeting ST8SIA4". en. In: *Cell Death Dis.* 7.12 (Dec. 2016), e2561.

[229] Jinsheng Ding et al. "OXCT1 Enhances Gemcitabine Resistance Through NF-$\kappa$B Pathway in Pancreatic Ductal Adenocarcinoma". en. In: *Front. Oncol.* 11 (2021).

[230] Yan He et al. "Targeting PI3K/Akt signal transduction for cancer therapy". en. In: *Signal Transduction and Targeted Therapy* 6.1 (Dec. 2021), pp. 1–17.

[231] Gregoire Mondielli et al. "Co-Targeting MAP Kinase and Pi3K-Akt-mTOR Pathways in Meningioma: Preclinical Study of Alpelisib and Trametinib". en. In: *Cancers* 14.18 (Sept. 2022), p. 4448.

[232] A Di Vinci et al. "HOXA7, 9, and 10 are methylation targets associated with aggressive behavior in meningiomas". In: *Transl. Res.* 160.5 (Nov. 2012).

[233] S R Wiley and J A Winkles. "TWEAK, a member of the TNF superfamily, is a multifunctional cytokine that binds the TweakR/Fn14 receptor". In: *Cytokine Growth Factor Rev.* 14.3-4 (June 2003).

[234] Yu Zhang et al. "A Novel Immune-Related Prognostic Biomarker and Target Associated With Malignant Progression of Glioma". en. In: *Front. Oncol.* 11 (Apr. 2021).

[235] Jungho Yang et al. "High TNFRSF12A level associated with MMP-9 overexpression is linked to poor prognosis in breast cancer: Gene set enrichment analysis and validation in large-scale cohorts". In: *PLoS One* 13.8 (Aug. 2018), e0202113.

[236] Jun-Yan Liu et al. "NETO2 promotes invasion and metastasis of gastric cancer cells via activation of PI3K/Akt/NF-$\kappa$B/Snail axis and predicts outcome of the patients". en. In: *Cell Death Dis.* 10.3 (Feb. 2019), pp. 1–14.

[237] Sidsel C Lindgaard et al. "Circulating Protein Biomarkers for Prognostic Use in Patients with Advanced Pancreatic Ductal Adenocarcinoma Undergoing Chemotherapy". en. In: *Cancers* 14.13 (July 2022), p. 3250.

[238] Shaohua Chen et al. "KCNN4 is a potential prognostic marker and critical factor affecting the immune status of the tumor microenvironment in kidney renal clear cell carcinoma". en. In: *Transl. Androl. Urol.* 10.6 (June 2021), p. 2454.

[239] Ping Xu et al. "KCNN4 promotes the progression of lung adenocarcinoma by activating the AKT and ERK signaling pathways". In: *Cancer Biomarkers* 31.2 (2021), pp. 187–201.

[240] Ping Xu et al. "KCNN4 promotes the progression of lung adenocarcinoma by activating the AKT and ERK signaling pathways". In: *Cancer Biomark.* 31.2 (Jan. 2021), pp. 187–201.

[241] Sarit Assouline et al. "A phase I trial of ribavirin and low-dose cytarabine for the treatment of relapsed and refractory acute myeloid leukemia with elevated eIF4E". en. In: *Haematologica* 100.1 (Jan. 2015), e7.

[242] Hiba Ahmad Zahreddine et al. “The sonic hedgehog factor GLI1 imparts drug resistance through inducible glucuronidation”. en. In: *Nature* 511.7507 (May 2014), pp. 90–93.

[243] Y-W Zhang et al. “Evidence that MIG-6 is a tumor-suppressor gene”. en. In: *Oncogene* 26.2 (July 2006), pp. 269–276.

[244] Michael K Wendt et al. “The antitumorigenic function of EGFR in metastatic breast cancer is regulated by expression of Mig6”. en. In: *Neoplasia* 17.1 (Jan. 2015), pp. 124–133.

[245] Zixuan Li et al. “Downregulation of Mig-6 in nonsmall-cell lung cancer is associated with EGFR signaling”. en. In: *Mol. Carcinog.* 51.7 (July 2012), pp. 522–534.

[246] Yuri Frosi et al. “A two-tiered mechanism of EGFR inhibition by RALT/MIG6 via kinase suppression and receptor degradation”. In: *Journal of Cell Biology* 189.3 (2010), pp. 557–571.

[247] James Chih-Hsin Yang. “A selective ALK inhibitor in ALK-rearranged patients”. en. In: *Lancet Oncol.* 14.7 (June 2013), pp. 564–565.

[248] Byung Soo Lee et al. “Hippo effector YAP directly regulates the expression of PD-L1 transcripts in EGFR-TKI-resistant lung adenocarcinoma”. en. In: *Biochem. Biophys. Res. Commun.* 491.2 (Sept. 2017), pp. 493–499.

[249] Da Hyun Kang et al. “Suppression of Mig-6 overcomes the acquired EGFR-TKI resistance of lung adenocarcinoma”. en. In: *BMC Cancer* 20.1 (June 2020), pp. 1–13.

[250] Junmei Cairns et al. “Differential roles of ERRFI1 in EGFR and AKT pathway regulation affect cancer proliferation”. en. In: *EMBO Rep.* 19.3 (Mar. 2018).

[251] Nesrin Sabha et al. “PIK3C2B inhibition improves function and prolongs survival in myotubular myopathy animal models”. en. In: *J. Clin. Invest.* 126.9 (Sept. 2016), pp. 3613–3625.

[252] Yiqun Zhang et al. “A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations”. In: *Cancer cell* 31.6 (2017), pp. 820–832.

[253] Pengyuan Liu et al. “Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing”. en. In: *Carcinogenesis* 33.7 (Apr. 2012), pp. 1270–1276.

[254] Mei Wang et al. “Glutathione peroxidase 2 overexpression promotes malignant progression and cisplatin resistance of KRAS-mutated lung cancer cells”. In: *Oncol. Rep.* 48.6 (Dec. 2022), pp. 1–14.

[255] Y Ino et al. “Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer”. en. In: *Br. J. Cancer* 108.4 (Mar. 2013), pp. 914–923.

[256] Erik S Knudsen et al. “Stratification of Pancreatic Ductal Adenocarcinoma: Combinatorial Genetic, Stromal, and Immunologic MarkersImmunologic Subtypes of Pancreatic Cancer”. In: *Clinical Cancer Research* 23.15 (2017), pp. 4429–4440.

[257] Robert H Vonderheide and Lauren J Bayne. "Inflammatory networks and immune surveillance of pancreatic carcinoma". en. In: *Curr. Opin. Immunol.* 25.2 (Apr. 2013), pp. 200–205.

[258] Christine Feig et al. "The pancreas cancer microenvironment". en. In: *Clin. Cancer Res.* 18.16 (Aug. 2012), pp. 4266–4276.

[259] Yousra Mohamed Abd-El-Halim et al. "A glycosyltransferase gene signature to detect pancreatic ductal adenocarcinoma patients with poor prognosis". en. In: *EBioMedicine* 71 (Sept. 2021), p. 103541.

[260] Teresa G Krieger et al. "Single-cell analysis of patient-derived PDAC organoids reveals cell state heterogeneity and a conserved developmental hierarchy". In: *Nature Communications* 12.1 (2021), p. 5826.

[261] Yu Wang et al. "Single-cell analysis of pancreatic ductal adenocarcinoma identifies a novel fibroblast subtype associated with poor prognosis but better immunotherapy response". en. In: *Cell Discov* 7.1 (May 2021), p. 36.

[262] Barbara T Grünwald et al. "Spatially confined sub-tumor microenvironments in pancreatic cancer". en. In: *Cell* 184.22 (Oct. 2021), 5577–5592.e18.

[263] Daniel Schreyer et al. "Deconstructing Pancreatic Cancer Using Next Generation-Omic Technologies-From Discovery to Knowledge-Guided Platforms for Better Patient Management". en. In: *Front Cell Dev Biol* 9 (2021), p. 795735.

[264] William L Hwang et al. "Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment". en. In: *Nat. Genet.* 54.8 (Aug. 2022), pp. 1178–1191.

[265] Rémy Nicolle et al. "Pancreatic Adenocarcinoma Therapeutic Targets Revealed by Tumor-Stroma Cross-Talk Analyses in Patient-Derived Xenografts". en. In: *Cell Rep.* 21.9 (Nov. 2017), pp. 2458–2470.

[266] Liwei Cao et al. "Proteogenomic characterization of pancreatic ductal adenocarcinoma". en. In: *Cell* 184.19 (Sept. 2021), 5031–5052.e26.

[267] Keegan Korthauer et al. "Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing". en. In: *Biostatistics* 20.3 (Feb. 2018), pp. 367–383.

[268] Yuan Tian et al. "ChAMP: updated methylation analysis pipeline for Illumina BeadChips". en. In: *Bioinformatics* 33.24 (Aug. 2017), pp. 3982–3984.

[269] Angela Di Vinci et al. "HOXA7, 9, and 10 are methylation targets associated with aggressive behavior in meningiomas". en. In: *Transl. Res.* 160.5 (Nov. 2012), pp. 355–362.

[270] Grace Collord et al. “An integrated genomic analysis of anaplastic meningioma identifies prognostic molecular signatures”. en. In: *Sci. Rep.* 8.1 (Sept. 2018), p. 13537.

[271] Yuan Zhuang et al. “Identification of an individualized immune-related prognostic risk score in lung squamous cell cancer”. In: *Frontiers in oncology* 11 (2021), p. 546455.

[272] Jun Qian and Pierre P Massion. “Next-generation molecular therapy in lung cancer”. In: *Translational lung cancer research* 7.Suppl 1 (2018), S31.

[273] GG Van den Eynden et al. “A fibrotic focus is a prognostic factor and a surrogate marker for hypoxia and (lymph) angiogenesis in breast cancer: review of the literature and proposal on the criteria of evaluation”. In: *Histopathology* 51.4 (2007), pp. 440–451.

[274] Xiaodan Zheng, Yuhai Hu, and Chengfang Yao. “The paradoxical role of tumor-infiltrating immune cells in lung cancer”. In: *Intractable & rare diseases research* 6.4 (2017), pp. 234–241.

[275] Shuguang Zuo et al. “Pan-cancer analysis of immune cell infiltration identifies a prognostic immune-cell characteristic score (ICCS) in lung adenocarcinoma”. In: *Frontiers in immunology* 11 (2020), p. 1218.

[276] N Thakur et al. “Tumor Budding as a Marker for Poor Prognosis and Epithelial-Mesenchymal Transition in Lung Cancer: A Systematic Review and Meta-Analysis”. In: *Front. Oncol.* 12 (June 2022).

[277] Ki-Jae Park et al. “Intensity of tumor budding and its prognostic implications in invasive colon carcinoma”. en. In: *Dis. Colon Rectum* 48.8 (Aug. 2005), pp. 1597–1602.

[278] Kenichi Ohtsuki et al. “Prognostic value of immunohistochemical analysis of tumor budding in colorectal carcinoma”. en. In: *Anticancer Res.* 28.3B (2008), pp. 1831–1836.

[279] Hui Li et al. “The tumor microenvironment: An irreplaceable element of tumor budding and epithelial-mesenchymal transition-mediated cancer metastasis”. en. In: *Cell Adh. Migr.* 10.4 (July 2016), pp. 1–13.

[280] I Zlobec et al. “Role of APAF-1, E-cadherin and peritumoral lymphocytic infiltration in tumour budding in colorectal cancer”. en. In: *J. Pathol.* 212.3 (July 2007), pp. 260–268.

[281] Y-W Zhang et al. “Evidence that MIG-6 is a tumor-suppressor gene”. en. In: *Oncogene* 26.2 (Jan. 2007), pp. 269–276.

[282] Mari Sasaki et al. “The tumor suppressor MIG6 controls mitotic progression and the G2/M DNA damage checkpoint by stabilizing the WEE1 kinase”. en. In: *Cell Rep.* 24.5 (July 2018), pp. 1278–1289.

[283] Yuki Ozato et al. “Spatial and single-cell transcriptomics decipher the cellular environment containing HLA-G+ cancer cells and SPP1+ macrophages in colorectal cancer”. en. In: *Cell Rep.* 42.1 (Jan. 2023), p. 111929.

[284] Charles D Warden et al. “COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis”. en. In: *Nucleic Acids Res.* 41.11 (Apr. 2013), e117–e117.

[285] Peter A Stockwell et al. “DMAP: differential methylation analysis package for RRBS and WGBS data”. en. In: *Bioinformatics* 30.13 (Mar. 2014), pp. 1814–1822.

[286] Marco Catoni et al. “DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts”. en. In: *Nucleic Acids Res.* 46.19 (July 2018), e114–e114.

[287] Yongseok Park and Hao Wu. “Differential methylation analysis for BS-seq data under general experimental design”. en. In: *Bioinformatics* 32.10 (Jan. 2016), pp. 1446–1453.

[288] Sheng Li et al. “An optimized algorithm for detecting and annotating regional differential methylation”. en. In: *BMC Bioinformatics* 14.5 (Apr. 2013), pp. 1–9.

[289] Yalu Wen et al. “Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics”. en. In: *Bioinformatics* 32.22 (Aug. 2016), pp. 3396–3404.

[290] Akanksha Srivastava et al. “HOME: a histogram based machine learning approach for effective identification of differentially methylated regions”. en. In: *BMC Bioinformatics* 20.1 (May 2019), pp. 1–15.

[291] Egor Dolzhenko and Andrew D Smith. “Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments”. en. In: *BMC Bioinformatics* 15.1 (June 2014), pp. 1–8.

# Appendix

## 5.1 Supplementary Tables

**Table 5.2** List of tools for the DMRs search.

| Tool | Num. of groups compared (2, 2+, continuous) | Data protocol (Array, BS-seq) | Input format | Preprocessing | Batch correction, confounding factor correction | Context (CpG, CpG and non-CpG) | DMRs type (predefined, *de novo*, both) |
|---|---|---|---|---|---|---|---|
| BiSeq | 2 | BS-seq | Bismark output | smoothing | - | CpG | Both |
| Bsmooth (bsseq) | 2 | BS-seq | Table with methylation read counts | smoothing | - | CpG and non-CpG | De novo |
| ChAMP pipeline (bumphunter) | 2+ | Array | Idat | filter by coverage, normalization, smoothing | Cell composition estimation, batch effect correction | CpG | De novo |
| ChAMP pipeline (ProbeLasso) | 2 | Array | Idat | filter by coverage,normalization | Cell composition estimation,batch effect correction | CpG | De novo |
| COHCAP | 2+ | Array, BS-seq | $\beta$-values / methylation % / bismark output | - | - | CpG | Predefined |
| **Dimmer 2.0** | 2, continuous | Array, BS-seq | Idat / $\beta$-values / bismark output / customed $\beta$-values matrix | Filter by coverage, background correction | Cell composition estimation, adjust for covariates, sex prediction | CpG and non-CpG | De novo |
| DMAP | 2 | BS-seq | Bismark output | filter by coverage | - | CpG and non-CpG | De novo |
| DMRcaller | 2 | BS-seq | Bismark output | filter by coverage,smoothing | - | CpG and non-CpG | Both |
| DMRcate | 2 | Array, BS-seq | M-values / bismark output | filter by position | - | CpG | De novo |
| DMRfinder | 2+ | BS-seq | Bismark output | filter by coverage | - | CpG | De novo |
| dmrseq | 2+ | BS-seq | Bismark output / table with methylation read counts | smoothing,filter by coverage | Adjust for covariates | CpG | De novo |
| DSS-general | 2+ | BS-seq | Bismark output / table with methylation read counts | - | Adjust for covariates | CpG | De novo |
| edgeR | 2+ | BS-seq | Bismark output | filter by coverage, normalization (dispersion and size) | Adjust for covariates | CpG | Predefined |
| eDMR | 2 | BS-seq | Output of methylkit or DML results in bed format | - | - | CpG | De novo |
| GetisDMR | 2 | BS-seq | Bismark output | - | Adjust for covariates | CpG | De novo |
| HOME | 2+, time series | BS-seq | Table with methylation read counts | - | - | CpG and non-CpG | De novo |
| methylKit | 2+ | BS-seq | Table with methylation % / bismark output | filter by coverage, correction for overdispersion | Adjust for covariates, batch effect correction | CpG and non-CpG | De novo |
| methylPipe | 2+ | BS-seq | Bismark output | smoothing | - | CpG and non-CpG | Both |
| methylSig | 2, continuous | BS-seq | Bismark output | filter by coverage, filter by position | Adjust for covariates | CpG and non-CpG | Both |
| metilene | 2 | BS-seq | Bed file with methylation ratio | - | - | CpG | Both |
| minfi pipeline (bumphunter) | 2 | Array | Idat | Filter by position, background correction, normalization | Cell composition estimation, sex prediction | CpG | De novo |
| RADMeth | 2 | BS-seq | Table with case and control methylation proportion | - | Adjust for covariates, batch effect correction | CpG and non-CpG | De novo |
| RnBeads | 2 | Array, BS-seq | Idat / $\beta$-values / table with methylation read counts | filter by coverage, normalization | Adjust for covariates, cell composition estimation, batch effect correction, sex and age prediction | CpG | Predefined |

| Tool | P-values for each DMR | Permutation | Graphic output | Language | Implementation | Reference |
|---|---|---|---|---|---|---|
| BiSeq | - | - | Plots | R | R (Bioconductor) | [160] |
| Bsmooth (bsseq) | - | - | Plots | R | R (Bioconductor) | [166] |
| ChAMP pipeline (bumphunter) | Yes | Yes | Plots | R | R (Bioconductor), R shiny app | [268] |
| ChAMP pipeline (ProbeLasso) | Yes | - | Plots | R | R (Bioconductor), R shiny app | same [268] |
| COHCAP | - | - | Plots | R/Java | R (Bioconductor), GUI | [284] |
| **Dimmer 2.0** | Yes | Yes | Plots | Java | GUI, command line tool | current study |
| DMAP | Yes | - | - | C | Command line tool | [285] |
| DMRcaller | Yes | - | Plots | R | R (Bioconductor) | [286] |
| DMRcate | Yes | - | Plots | R | R (Bioconductor) | [169] |
| DMRfinder | - | - | - | R and Python | Command line tool | [167] |
| dmrseq | Yes | Yes | Plots | R | R (Bioconductor) | [267] |
| DSS-general | - | - | Plots | R | R (Bioconductor) | [287] |
| edgeR | Yes | - | Plots | R | R (Bioconductor) | [170] |
| eDMR | Yes | - | Plots | R | R package | [288] |
| GetisDMR | - | - | - | C++, R | Command line tool | [289] |
| HOME | - | - | - | Python | python package | [290] |
| methylKit | - | - | Plots | R | R (Bioconductor) | [168] |
| methylPipe | Yes | - | Plots | R | R (Bioconductor) | [165] |
| methylSig | - | - | Plots | R | R (Bioconductor) | [163] |
| metilene | Yes | - | Plots | C | Command line tool | [164] |
| minfi pipeline (bumphunter) | Yes | Yes | - | R | R (Bioconductor) | [161] |
| RADMeth | - | - | - | C++ | Command line tool | [291] |
| RnBeads | Yes | - | Plots | R | R (Bioconductor) | [171] |

**Table 5.3** GREAT output showing gene ontology enrichment analysis on the genes with TSS nearby or matching the regions found with significant different methylation level between benign and malignant meningioma samples. Statistically significant Biological Processes (BP) and Molecular Function (MF) terms with FDR adjusted p-value < 0.05 in both binomial and hypergeometric tests. Only top 20 most significant terms are shown.

| Repository | Term | Binomial FDR p-value | Hypergeometric FDR p-value |
|---|---|---|---|
| GO MF | sequence-specific DNA binding | 1.36E-119 | 3.93E-63 |
| GO MF | transcription factor activity, sequence-specific DNA binding | 1.86E-90 | 1.11E-46 |
| GO MF | RNA polymerase II transcription factor activity, sequence-specific DNA binding | 1.40E-86 | 2.83E-49 |
| GO MF | regulatory region DNA binding | 2.52E-72 | 2.35E-42 |
| GO MF | regulatory region nucleic acid binding | 3.51E-72 | 2.38E-42 |
| GO MF | RNA polymerase II regulatory region sequence-specific DNA binding | 1.26E-71 | 3.26E-39 |
| GO MF | transcription regulatory region DNA binding | 2.57E-71 | 1.09E-41 |
| GO MF | RNA polymerase II regulatory region DNA binding | 3.24E-71 | 4.94E-39 |
| GO MF | transcription regulatory region sequence-specific DNA binding | 6.84E-68 | 8.42E-39 |
| GO MF | sequence-specific double-stranded DNA binding | 1.60E-66 | 7.65E-37 |
| GO MF | double-stranded DNA binding | 1.59E-63 | 2.11E-33 |
| GO MF | transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding | 2.02E-56 | 3.09E-32 |
| GO MF | transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding | 2.10E-46 | 1.25E-26 |
| GO MF | RNA polymerase II core promoter proximal region sequence-specific DNA binding | 1.80E-38 | 3.70E-21 |
| GO MF | core promoter proximal region DNA binding | 7.57E-38 | 4.90E-22 |
| GO MF | core promoter proximal region sequence-specific DNA binding | 1.21E-37 | 2.62E-21 |
| GO MF | transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding | 2.03E-35 | 6.55E-21 |
| GO MF | transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding | 4.57E-21 | 3.50E-11 |
| GO MF | HMG box domain binding | 1.86E-15 | 2.01E-07 |
| GO MF | transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding | 2.92E-14 | 1.51E-05 |

**Table 5.4** Enrichment analysis on the genes with promoter regions overlapping the DMRs. Analysis performed on the gene ontology Biological Processes (BP) and Molecular Function (MF), and the pathway repositories Reactome and KEGG. Only top 20 statistically significant terms with FDR adjusted p-value < 0.05 are shown.

| Repository | Term | Overlap | Adjusted p-value |
|---|---|---|---|
| Reactome | Activation of anterior HOX genes in hindbrain development during early embryogenesis Homo sapiens R-HSA-5617472 | 14/89 | 0.0085 |
| Reactome | Activation of HOX genes during differentiation Homo sapiens R-HSA-5619507 | 14/89 | 0.0085 |
| Reactome | Regulation of beta-cell development Homo sapiens R-HSA-186712 | 8/32 | 0.0114 |
| Reactome | GPCR ligand binding Homo sapiens R-HSA-500792 | 36/447 | 0.0315 |
| Reactome | Regulation of gene expression in early pancreatic precursor cells Homo sapiens R-HSA-210747 | 4/8 | 0.0315 |
| Reactome | Amino acid transport across the plasma membrane Homo sapiens R-HSA-352230 | 7/31 | 0.0344 |
| Reactome | Transcriptional regulation of pluripotent stem cells Homo sapiens R-HSA-452723 | 8/42 | 0.0376 |
| Reactome | Ras activation uopn Ca2+ infux through NMDA receptor Homo sapiens R-HSA-442982 | 5/17 | 0.049 |
| Reactome | POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation Homo sapiens R-HSA-2892245 | 4/10 | 0.049 |
| KEGG | Maturity onset diabetes of the young | 10/26 | 1.31E-05 |
| KEGG | Neuroactive ligand-receptor interaction | 33/341 | 0.0012 |
| GO CC | integral component of plasma membrane (GO:0005887) | 109/1454 | 6.39E-07 |
| GO CC | nucleus (GO:0005634) | 242/4484 | 0.0012 |
| GO CC | intracellular membrane-bounded organelle (GO:0043231) | 262/5192 | 0.0362 |
| GO MF | sequence-specific DNA binding (GO:0043565) | 137/707 | 1.31E-50 |
| GO MF | sequence-specific double-stranded DNA binding (GO:1990837) | 133/712 | 2.27E-47 |
| GO MF | double-stranded DNA binding (GO:0003690) | 125/651 | 7.83E-46 |
| GO MF | RNA polymerase II transcription regulatory region sequence-specific DNA binding (GO:0000977) | 182/1359 | 2.23E-44 |
| GO MF | RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978) | 163/1149 | 1.27E-42 |
| GO MF | cis-regulatory region sequence-specific DNA binding (GO:0000987) | 162/1149 | 4.67E-42 |

## 5.2 Supplementary Figures

**Figure 5.1** Amount of genes in common across signatures from the four studies.

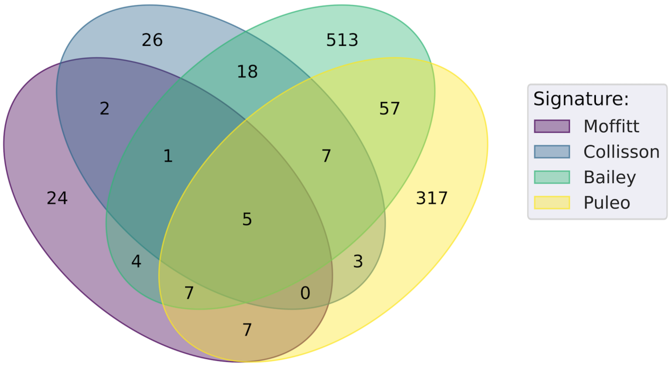


**Figure 5.2** Visualization in principal component analysis (PCA) of the nine validation datasets, before (left) and after (right) batch effect correction.

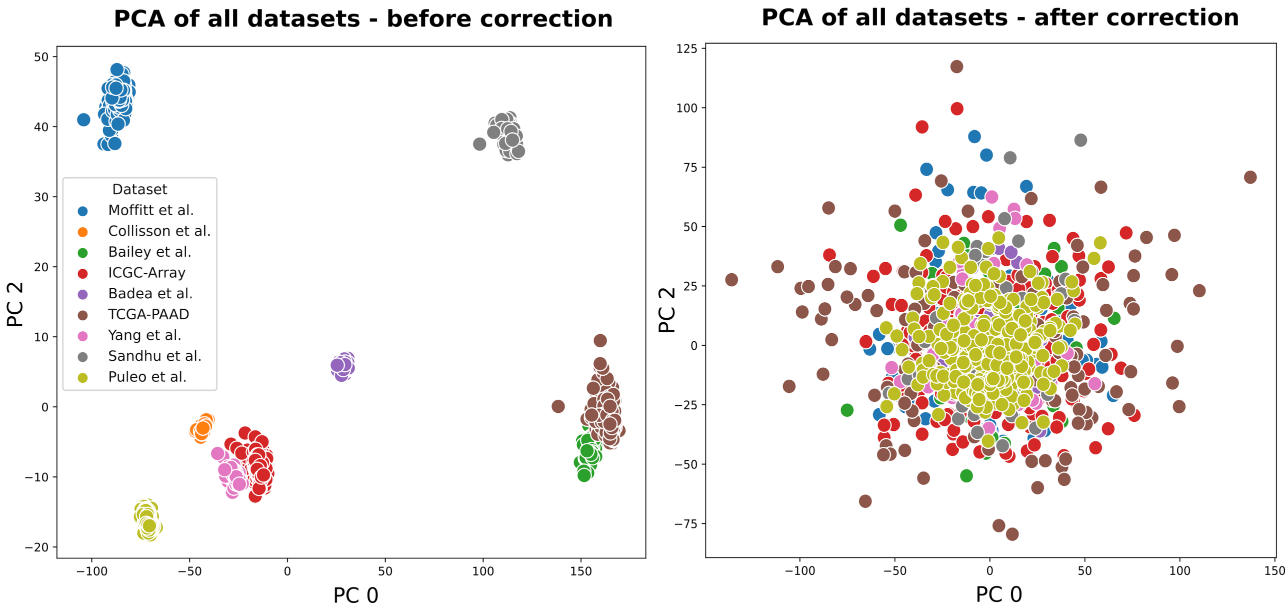


**Figure 5.3** Comparison of Puleo *et al.* subtypes carried out with Wilcoxon test in a pairwise setting, taking one gene at a time for each of the four signatures. P-values are subject to multiple testing correction and $-log_{10}$ of the p-value is shown

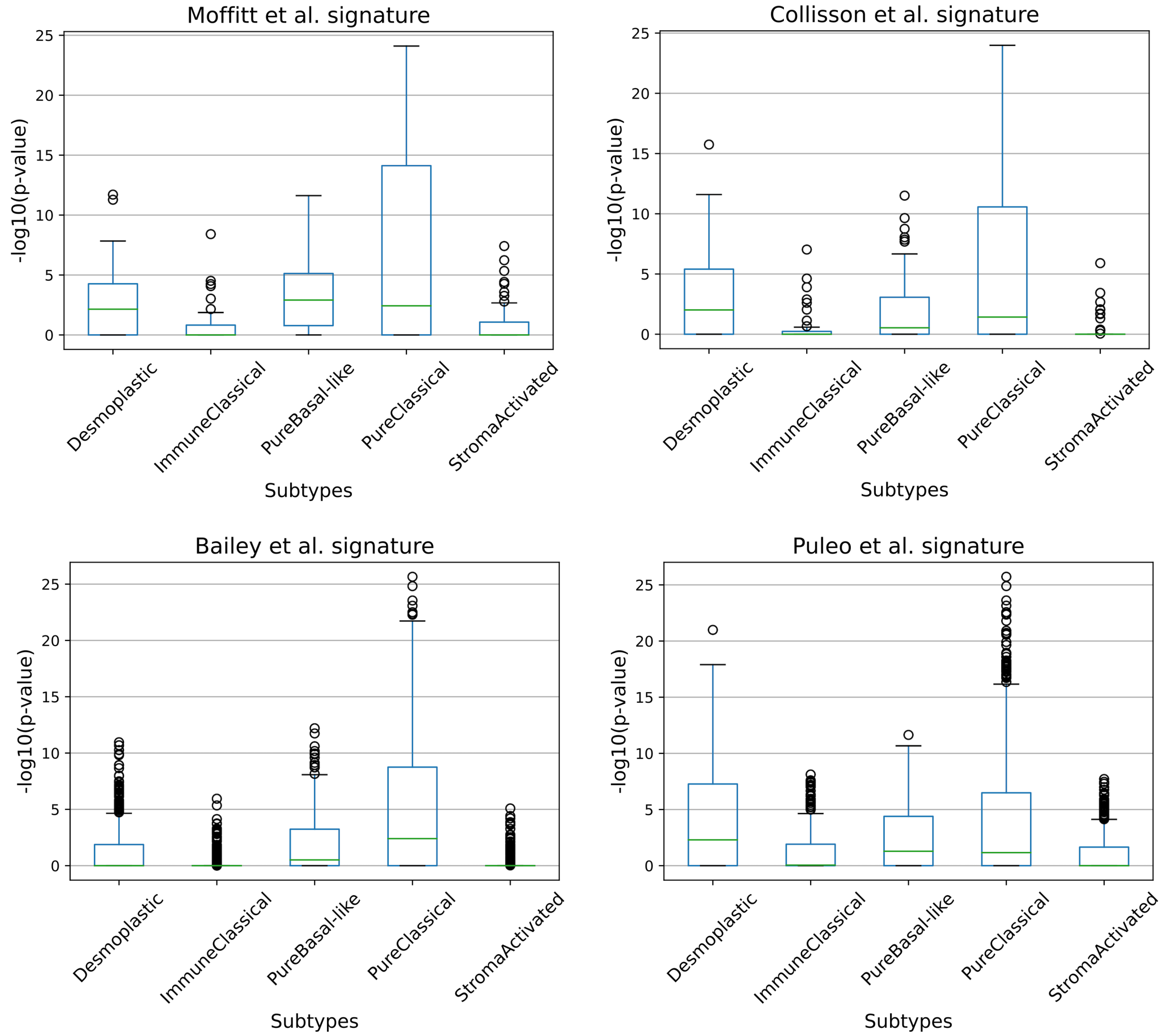
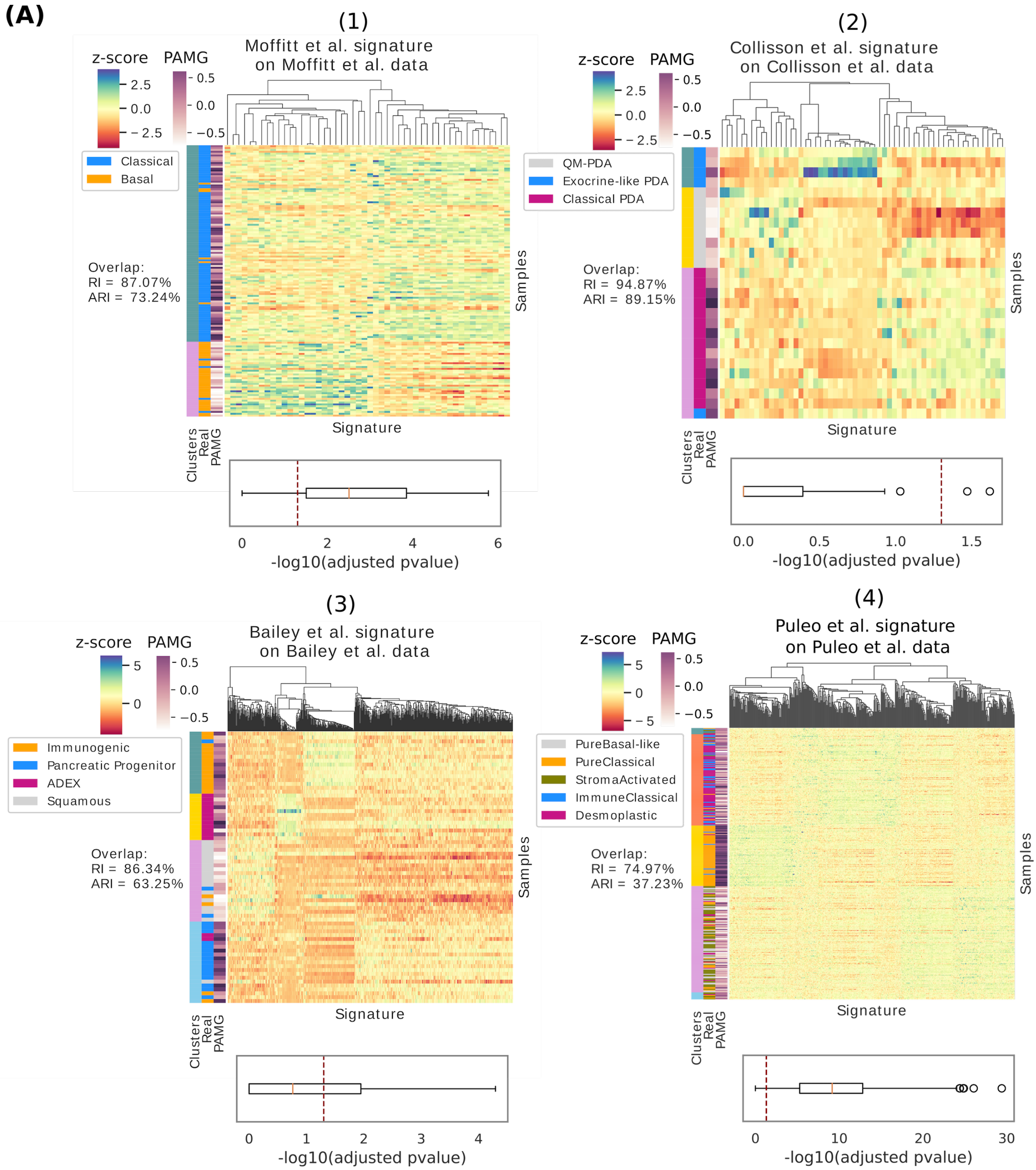
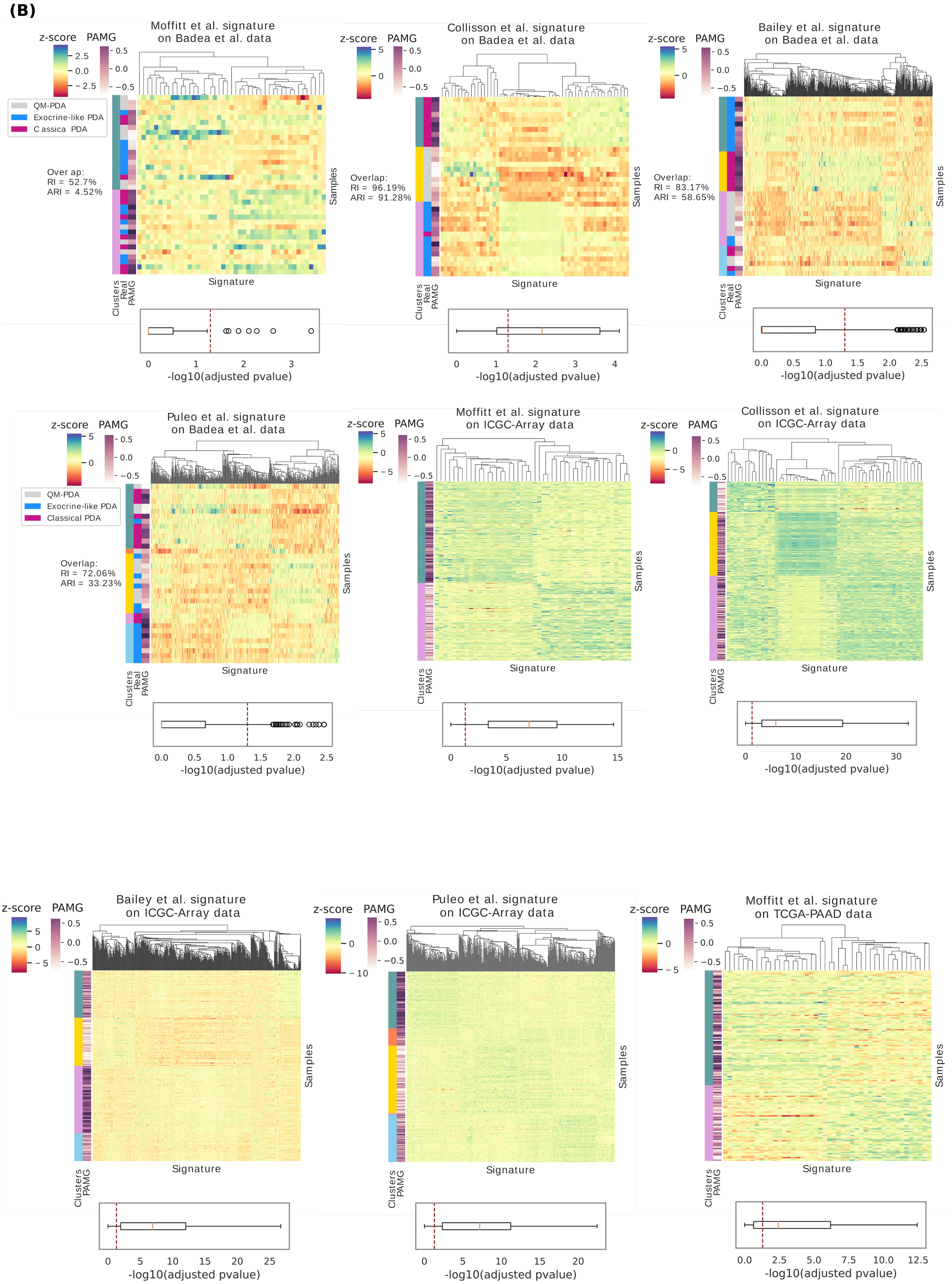
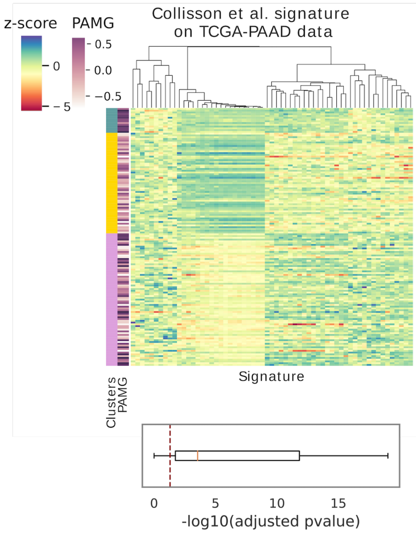
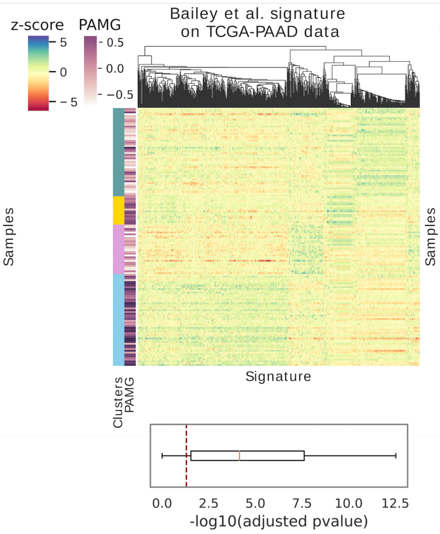


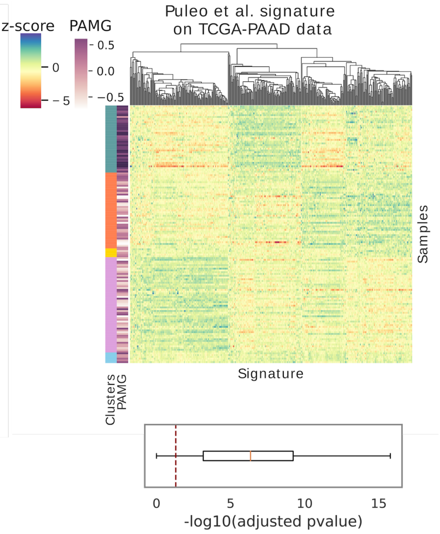
**Figure 5.4** Clustering of the nine datasets using the genes from the four signatures Moffitt *et al.*, Collisson *et al.*, Bailey *et al.* and Puleo *et al.*. (A) When the datasets and the signatures come from the same study; (B) clustering of the remaining five datasets. Each gene in a signature is assessed comparing the expression distribution of the obtained clusters, using Wilcoxon rank sum test for two clusters and Kruskal-Wallis test for three, four and five. The p-values obtained are $-log_{10}$ transformed and shown in a boxplot with a red dashed line representing the significance threshold (p-value=0.05). (*) Badea *et al.* dataset has real labels assigned by Collisson *et al.*, which are compared with the obtained clusters.

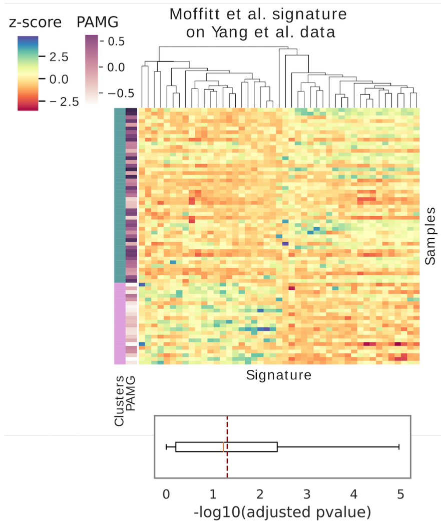


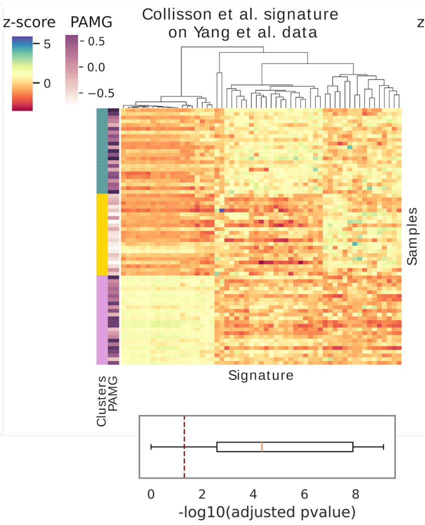

(B)
Moffitt et al. signature on Badea et al. data
z-score
PAMG
QM-PDA
Exocrine-like PDA
Classical PDA
Overlap:
RI = 52.7%
ARI = 4.52%
Samples
Signature
Clusters
Real
PAMG
-log10(adjusted pvalue)
Collisson et al. signature on Badea et al. data
Overlap:
RI = 96.19%
ARI = 91.28%
Bailey et al. signature on Badea et al. data
Overlap:
RI = 83.17%
ARI = 58.65%
Puleo et al. signature on Badea et al. data
Overlap:
RI = 72.06%
ARI = 33.23%
Moffitt et al. signature on ICGC-Array data
Collisson et al. signature on ICGC-Array data
Bailey et al. signature on ICGC-Array data
Puleo et al. signature on ICGC-Array data
Moffitt et al. signature on TCGA-PAAD data

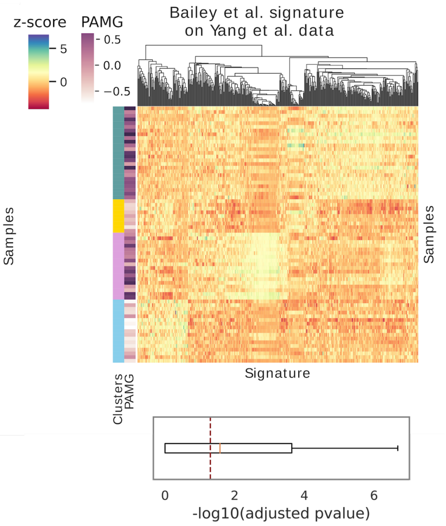

Collisson et al. signature
on TCGA-PAAD data
z-score
PAMG
0
− 5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
5
10
15
-log10(adjusted pvalue)

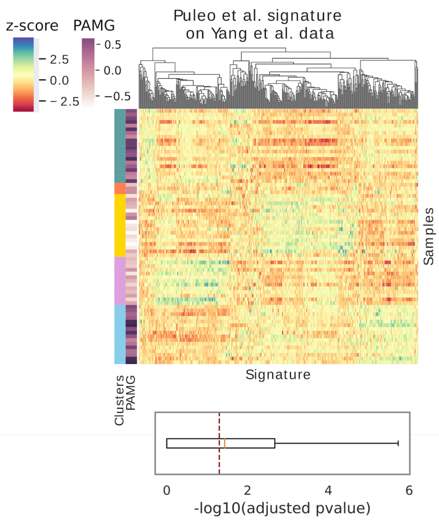

Bailey et al. signature
on TCGA-PAAD data
z-score
PAMG
5
0
− 5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0.0
2.5
5.0
7.5
10.0
12.5
-log10(adjusted pvalue)

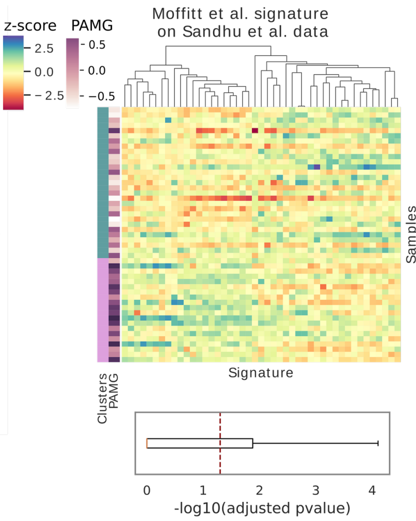

Puleo et al. signature
on TCGA-PAAD data
z-score
PAMG
0
− 5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
5
10
15
-log10(adjusted pvalue)

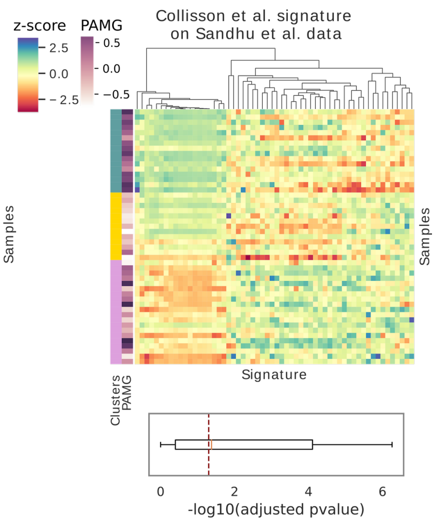

Moffitt et al. signature
on Yang et al. data
z-score
PAMG
2.5
0.0
− 2.5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
1
2
3
4
5
-log10(adjusted pvalue)


Collisson et al. signature
on Yang et al. data
z-score
PAMG
5
0
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
2
4
6
8
-log10(adjusted pvalue)


Bailey et al. signature
on Yang et al. data
z-score
PAMG
5
0
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
2
4
6
-log10(adjusted pvalue)


Puleo et al. signature
on Yang et al. data
z-score
PAMG
2.5
0.0
− 2.5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
2
4
6
-log10(adjusted pvalue)


Moffitt et al. signature
on Sandhu et al. data
z-score
PAMG
2.5
0.0
− 2.5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
1
2
3
4
-log10(adjusted pvalue)


Collisson et al. signature
on Sandhu et al. data
z-score
PAMG
2.5
0.0
− 2.5
0.5
0.0
−0.5
Samples
Signature
Clusters
PAMG
0
2
4
6
-log10(adjusted pvalue)

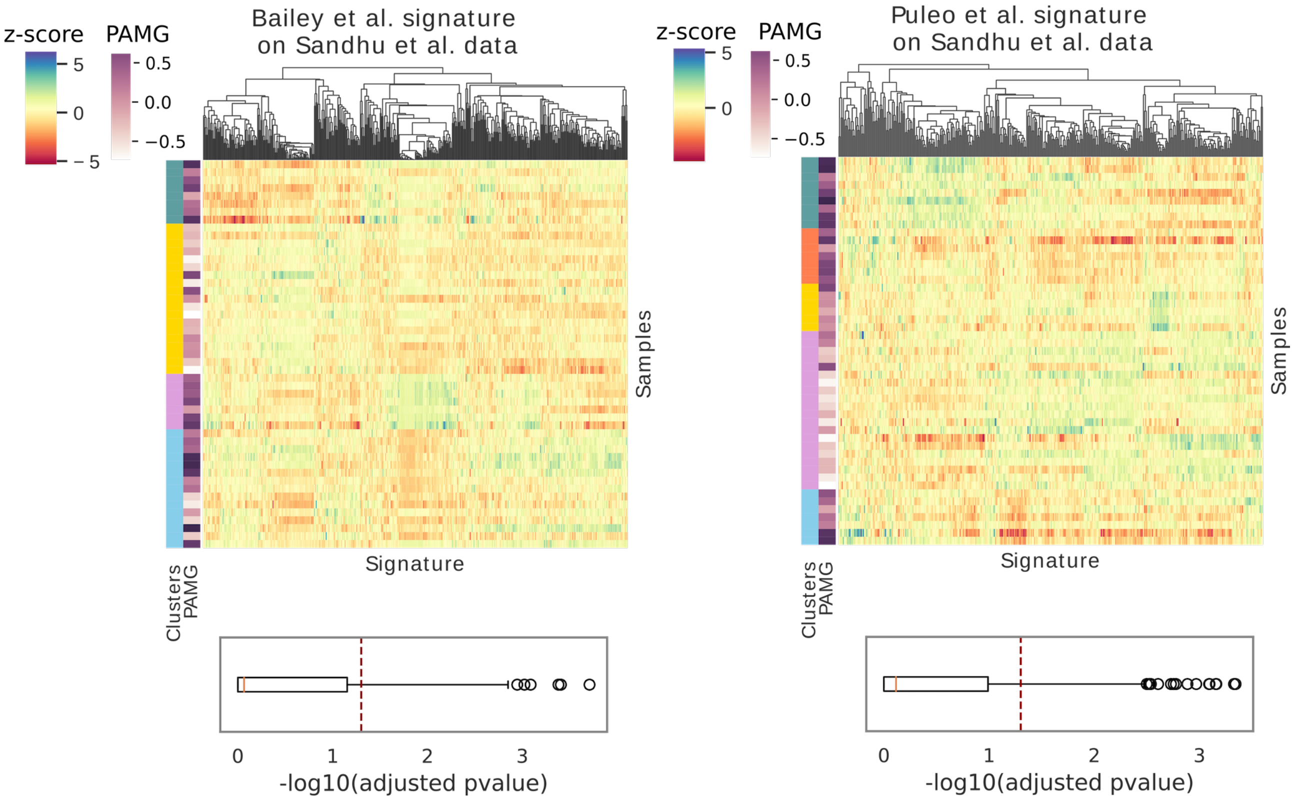
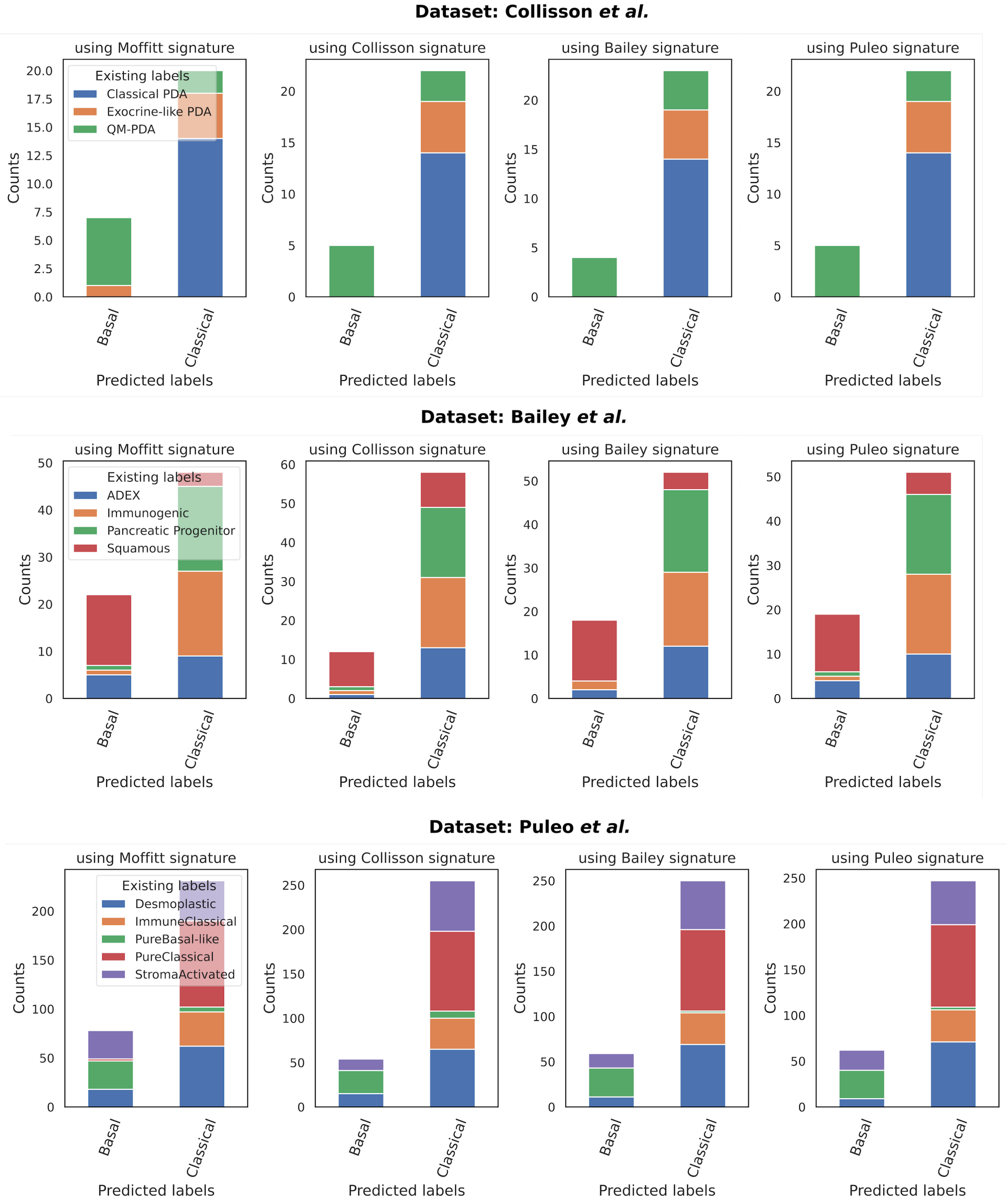

z-score
5
0
−5
PAMG
0.5
0.0
−0.5
Bailey et al. signature
on Sandhu et al. data
Samples
Clusters
PAMG
Signature
0
1
2
3
-log10(adjusted pvalue)
z-score
5
0
PAMG
0.5
0.0
−0.5
Puleo et al. signature
on Sandhu et al. data
Samples
Clusters
PAMG
Signature
0
1
2
3
-log10(adjusted pvalue)

**Figure 5.5** Predicted classes are compared with the real classes, for each dataset where labels are available. Train datasets are (a) Moffitt *et al.*, (b) Collisson *et al.*, (c) Bailey *et al.* and (d) Puleo *et al.*, used for building prediction models using, one by one, the four signatures. Each model is used on the nine datasets to classify their samples. Only datasets with real labels available are collected in this figure to check the agreement between predicted and pre-assigned subtypes. For samples in Collisson *et al.* and Badea *et al.* no prediction of Immune Classical subtype was found.

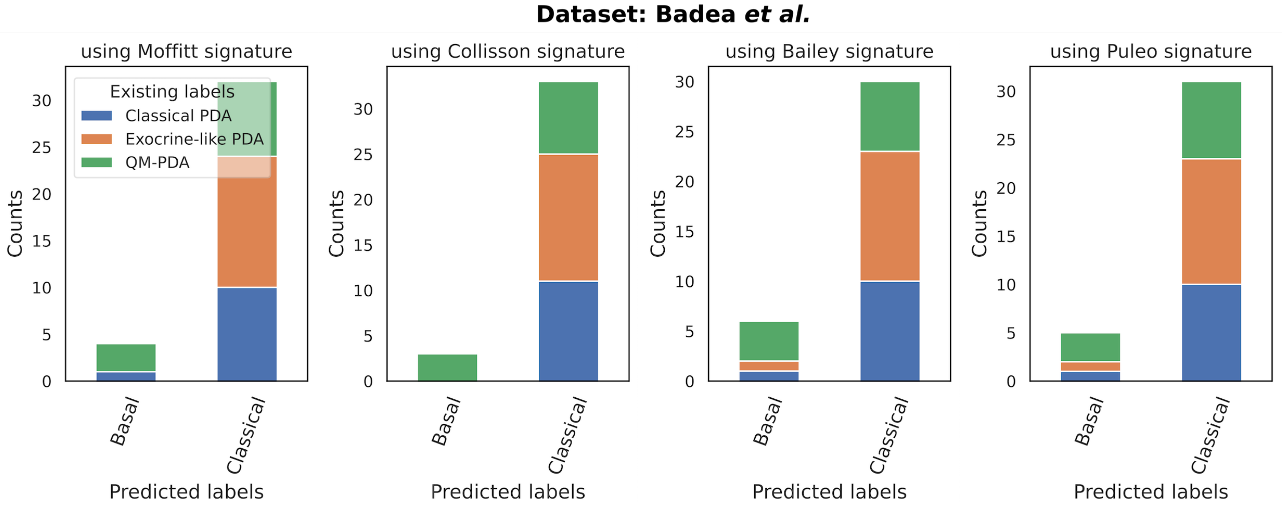


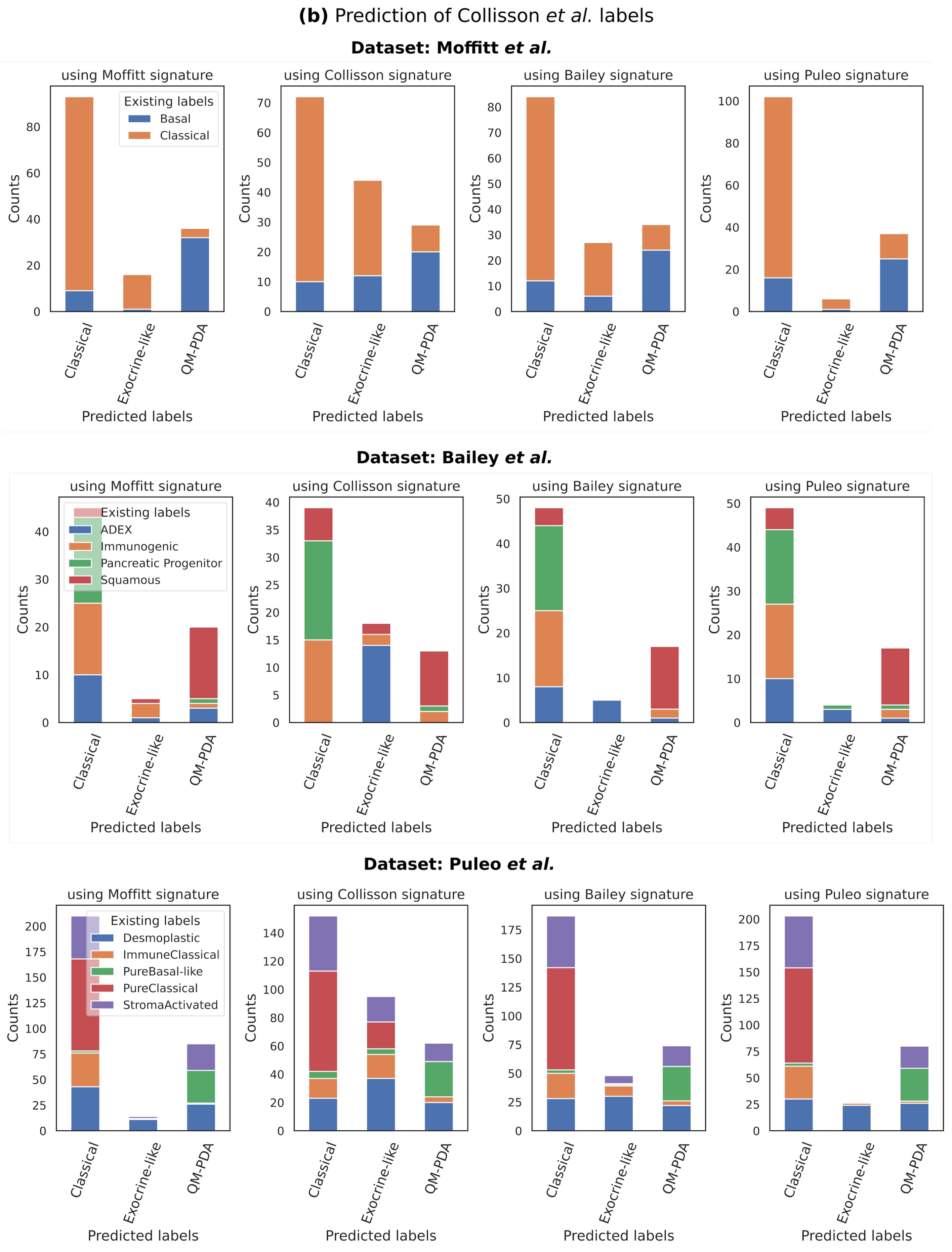
**(a)** Prediction of Moffitt *et al.* labels

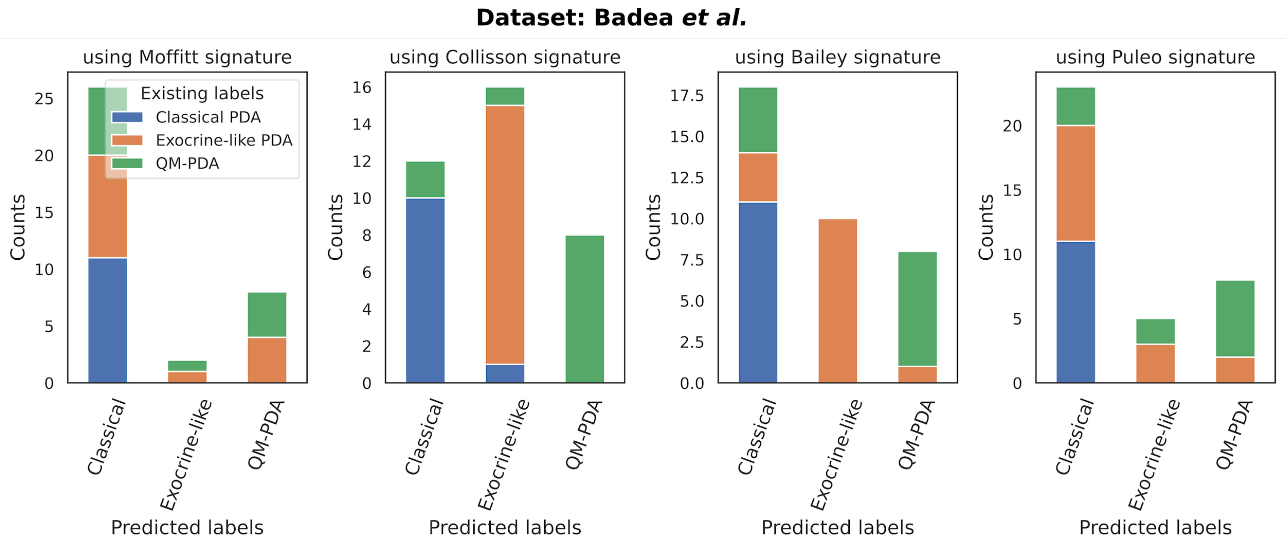
**Dataset: Collisson *et al.***

**Dataset: Bailey *et al.***

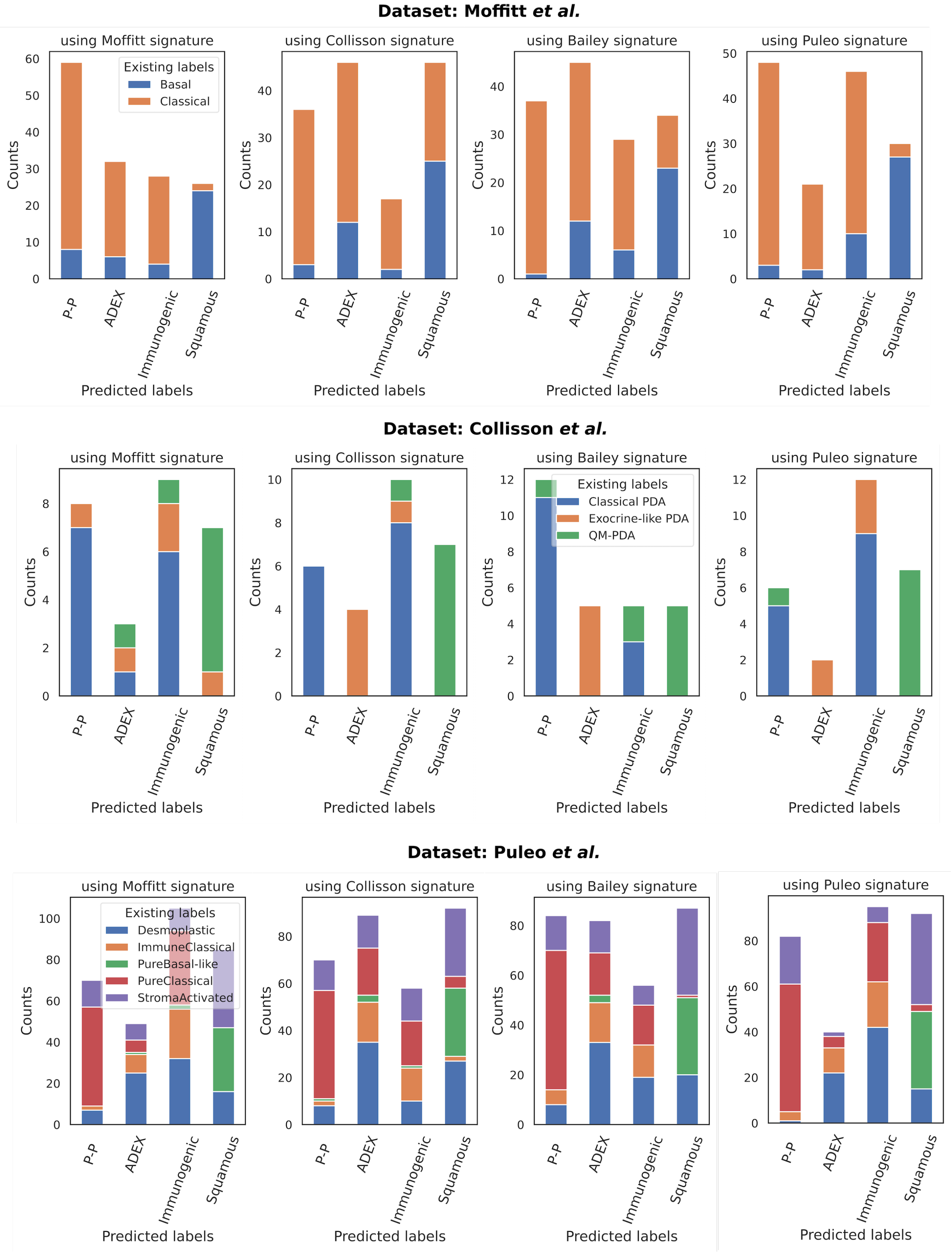
**Dataset: Puleo *et al.***


Dataset: Badea et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Classical PDA
Exocrine-like PDA
QM-PDA
Counts
0
5
10
15
20
25
30
Basal
Classical
Predicted labels


(b) Prediction of Collisson et al. labels
Dataset: Moffitt et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Basal
Classical
Counts
0
20
40
60
80
0
10
20
30
40
50
60
70
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
Classical
Exocrine-like
QM-PDA
Predicted labels
Dataset: Bailey et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
ADEX
Immunogenic
Pancreatic Progenitor
Squamous
Counts
0
10
20
30
40
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
0
10
20
30
40
50
Classical
Exocrine-like
QM-PDA
Predicted labels
Dataset: Puleo et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
Counts
0
25
50
75
100
125
150
175
200
0
20
40
60
80
100
120
140
0
25
50
75
100
125
150
175
0
25
50
75
100
125
150
175
200
Classical
Exocrine-like
QM-PDA
Predicted labels


Dataset: Badea *et al.*
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Classical PDA
Exocrine-like PDA
QM-PDA
Counts
Predicted labels
Classical
Exocrine-like
QM-PDA
0
5
10
15
20
25
2
4
6
8
12
14
16
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

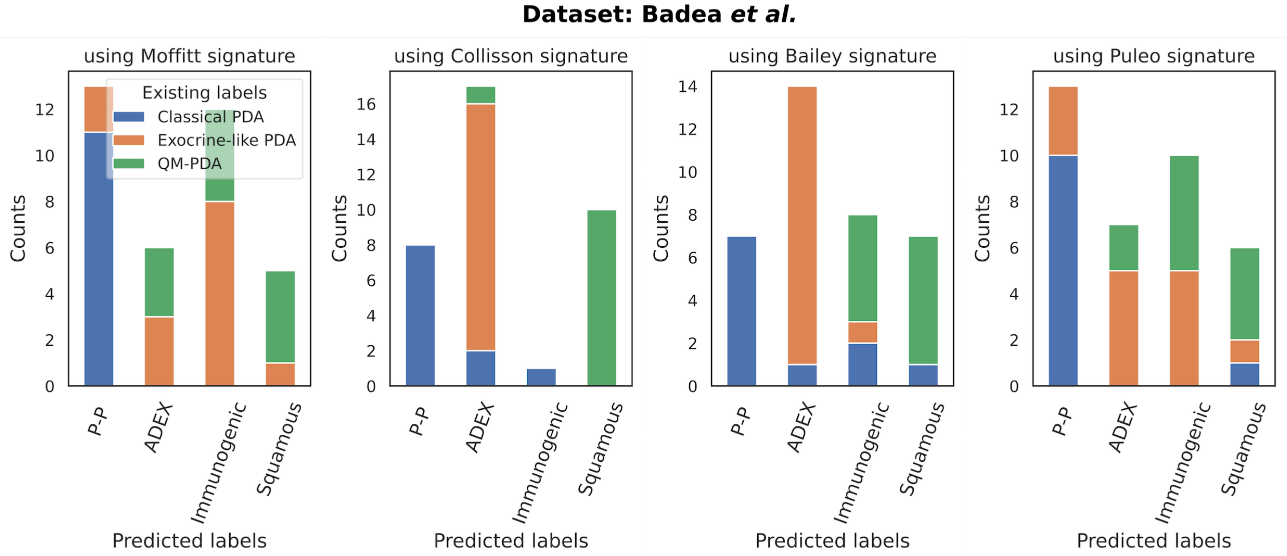
**(c)** Prediction of Bailey *et al.* labels

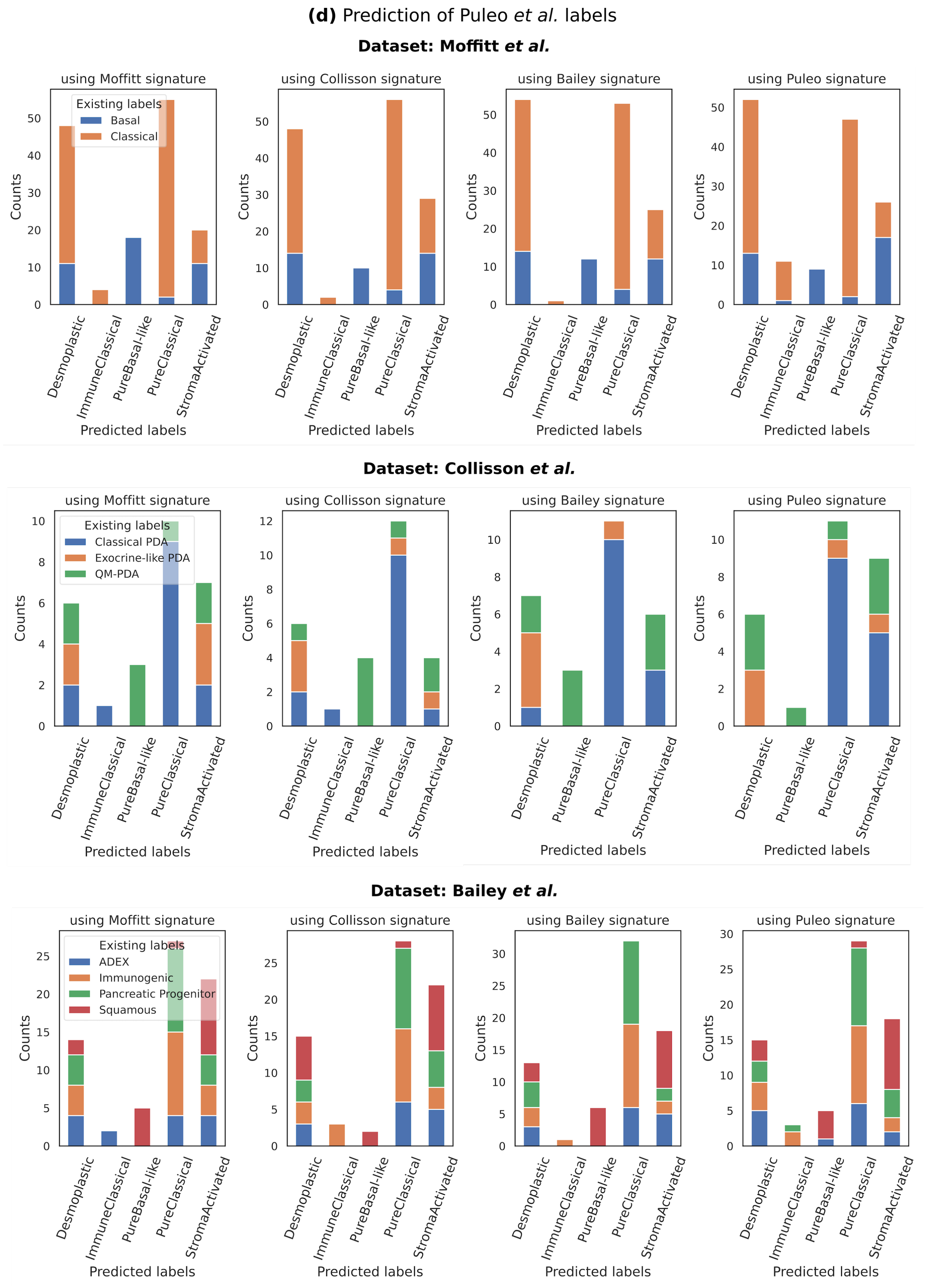


**Dataset: Moffitt *et al.***

**Dataset: Collisson *et al.***

**Dataset: Puleo *et al.***


Dataset: Badea *et al.*
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Classical PDA
Exocrine-like PDA
QM-PDA
Counts
Predicted labels
P-P
ADEX
Immunogenic
Squamous


(d) Prediction of Puleo et al. labels
Dataset: Moffitt et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Basal
Classical
Counts
Predicted labels
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
0
10
20
30
40
50
Dataset: Collisson et al.
Existing labels
Classical PDA
Exocrine-like PDA
QM-PDA
0
2
4
6
8
10
12
Dataset: Bailey et al.
Existing labels
ADEX
Immunogenic
Pancreatic Progenitor
Squamous
0
5
10
15
20
25
30

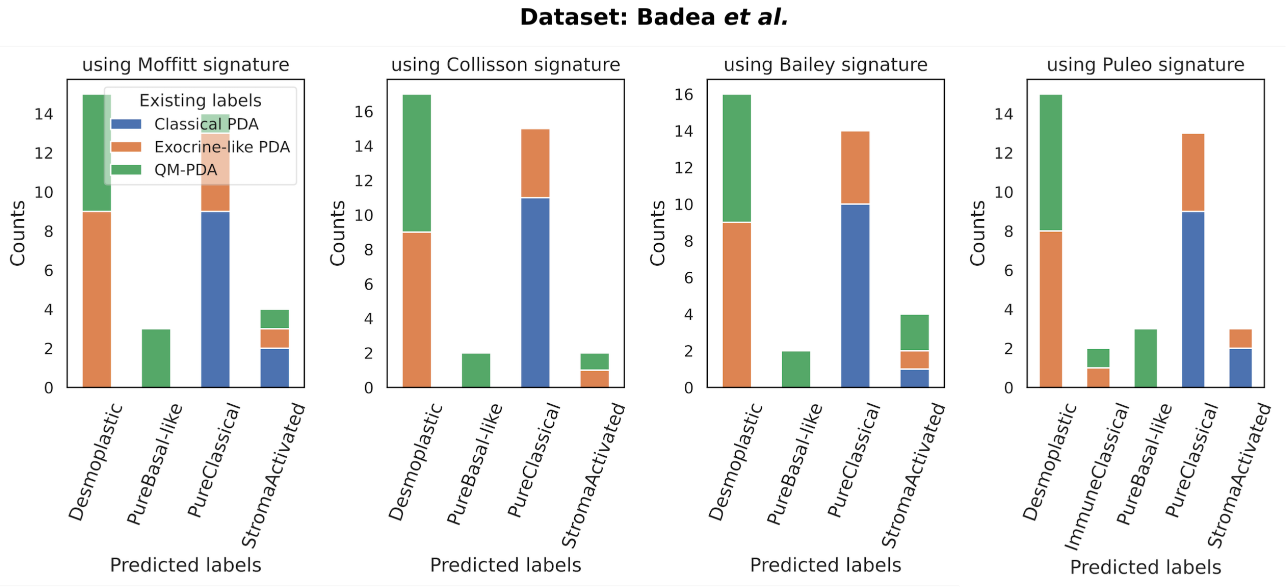
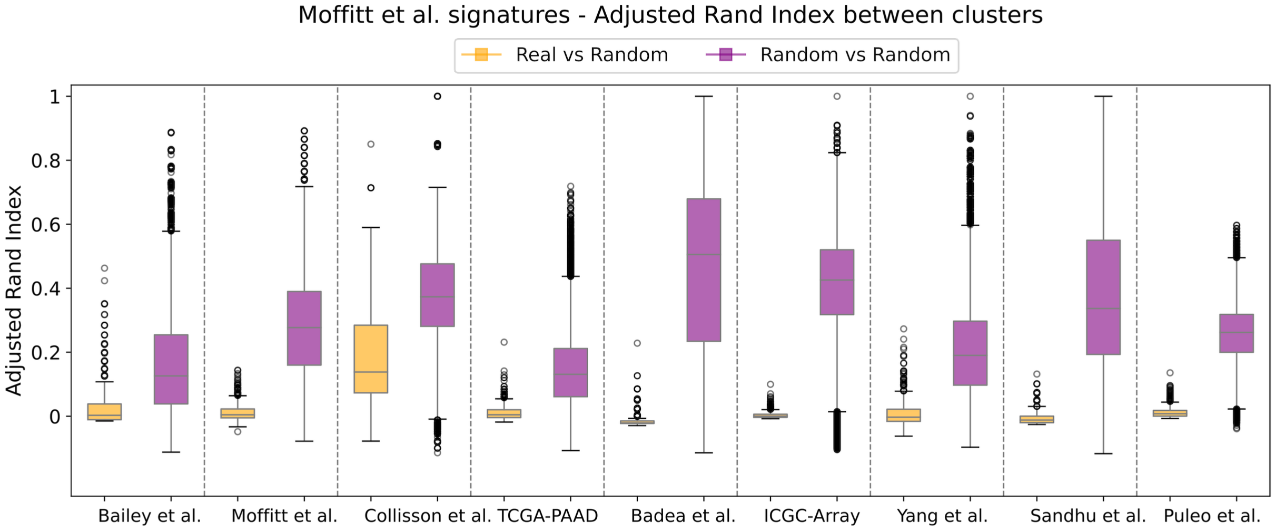
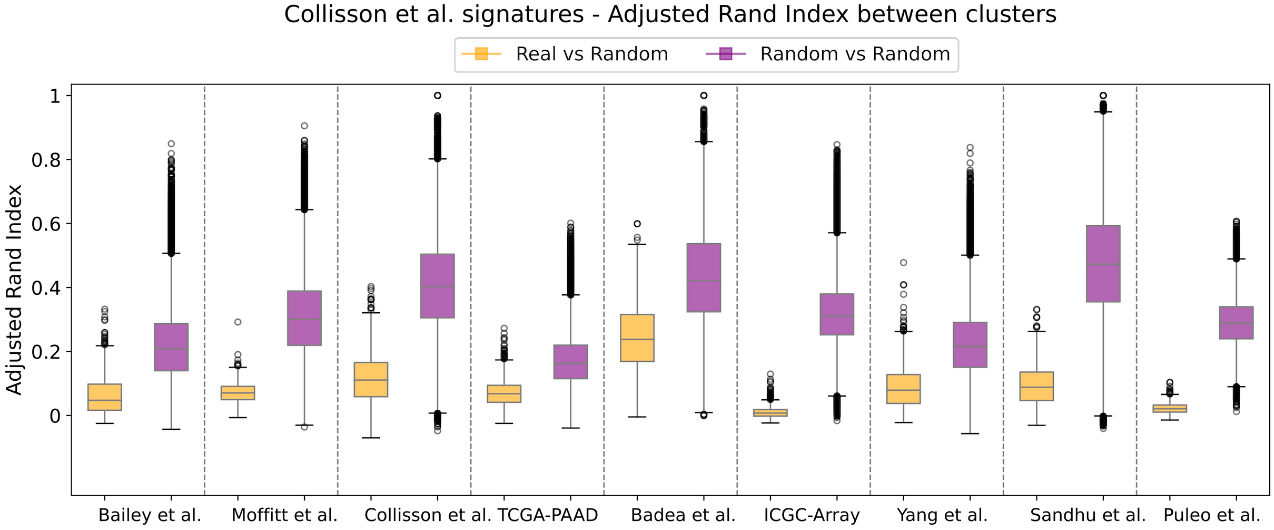
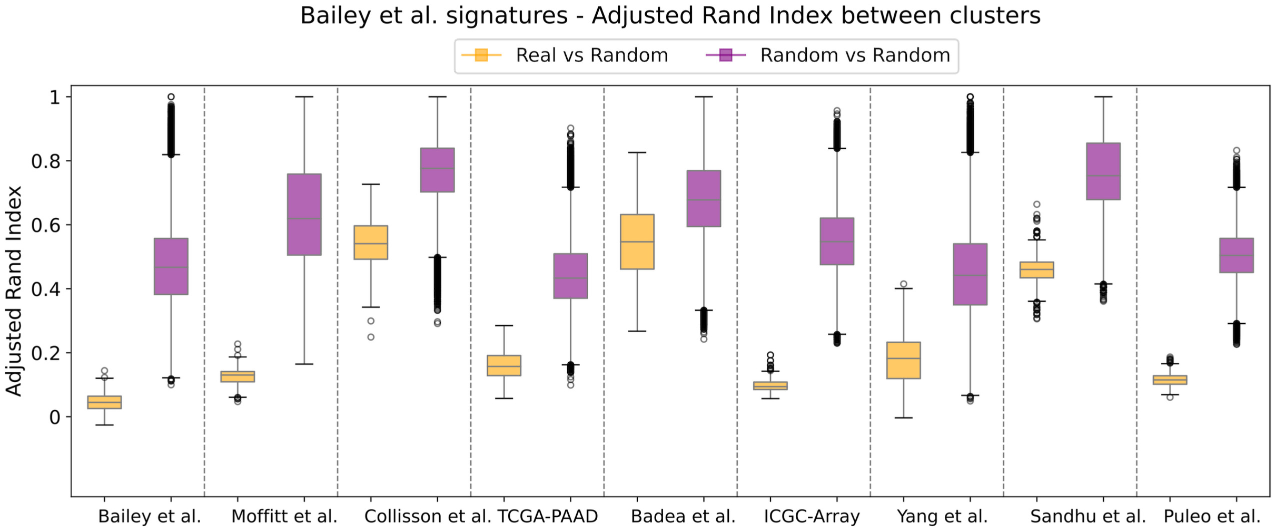

Dataset: Badea et al.
using Moffitt signature
using Collisson signature
using Bailey signature
using Puleo signature
Existing labels
Classical PDA
Exocrine-like PDA
QM-PDA
Counts
0
2
4
6
8
10
12
14
16
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
Predicted labels

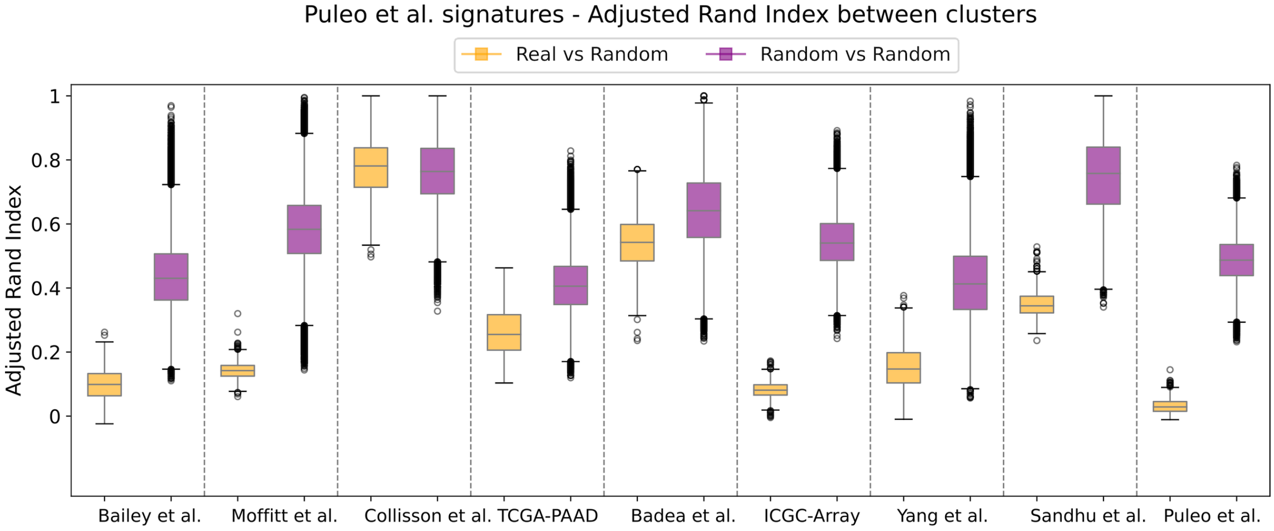
**Figure 5.6** Adjusted Rand Index (ARI) is used for evaluating the clustering robustness of the signatures. From top to bottom: clusters derived using Moffitt *et al.*, Collisson *et al.*, Bailey *et al.* and Puleo *et al.* signatures are compared with the ones identified when employing random genes of the same size of the signature (Real vs Random). ARI is also used for a pair-wise comparison between clusters deriving from the use of random gene sets (Random vs Random). Comparison between clusters obtained with the use of real signatures and clusters obtained with the use of the same number of genes, but random (Real vs Random). To assess the comparison, Adjusted Rand Index (ARI) is computed. The clusters obtained from random genes are further evaluated with a pairwise ARI (Random vs Random).





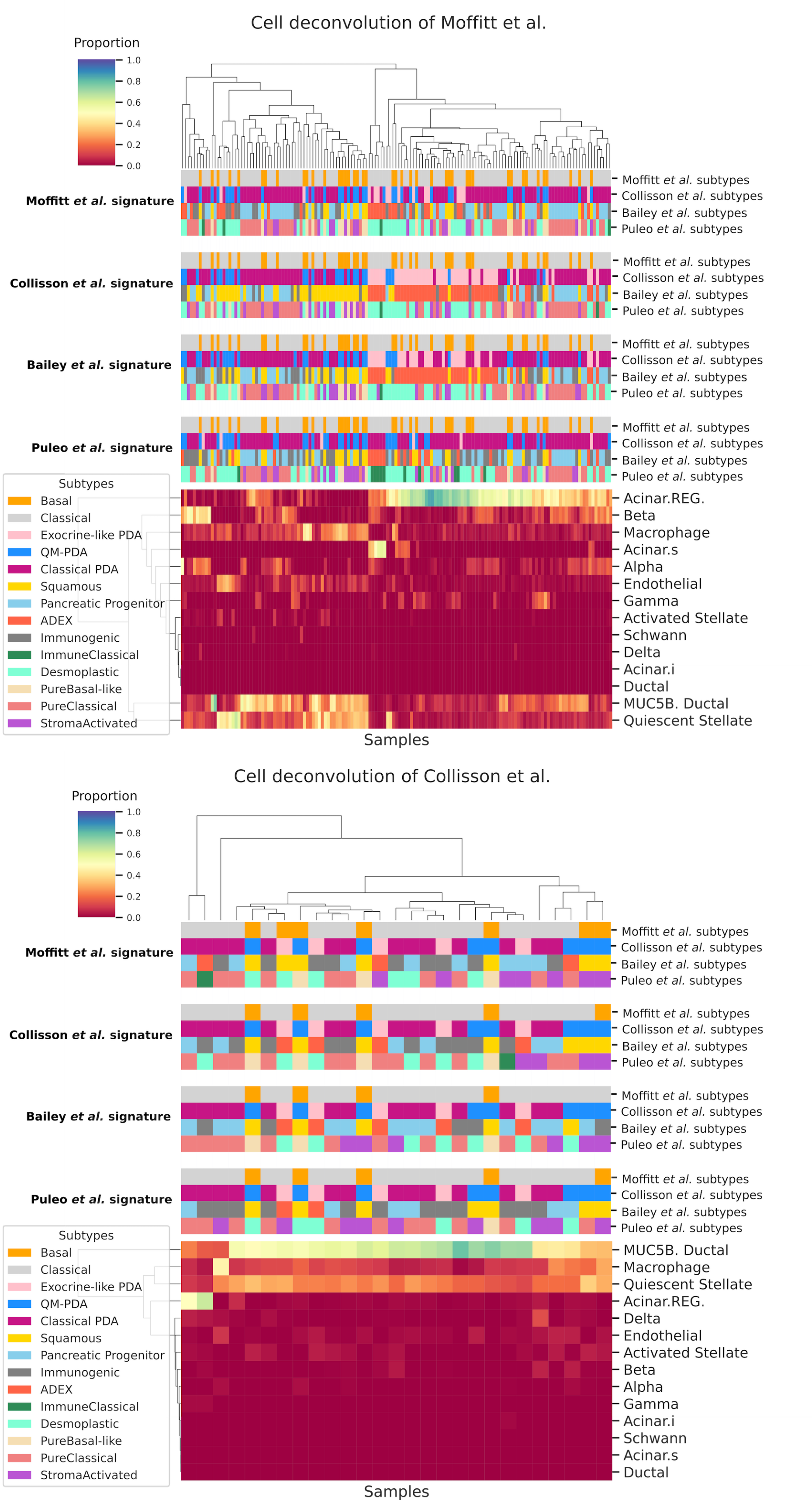


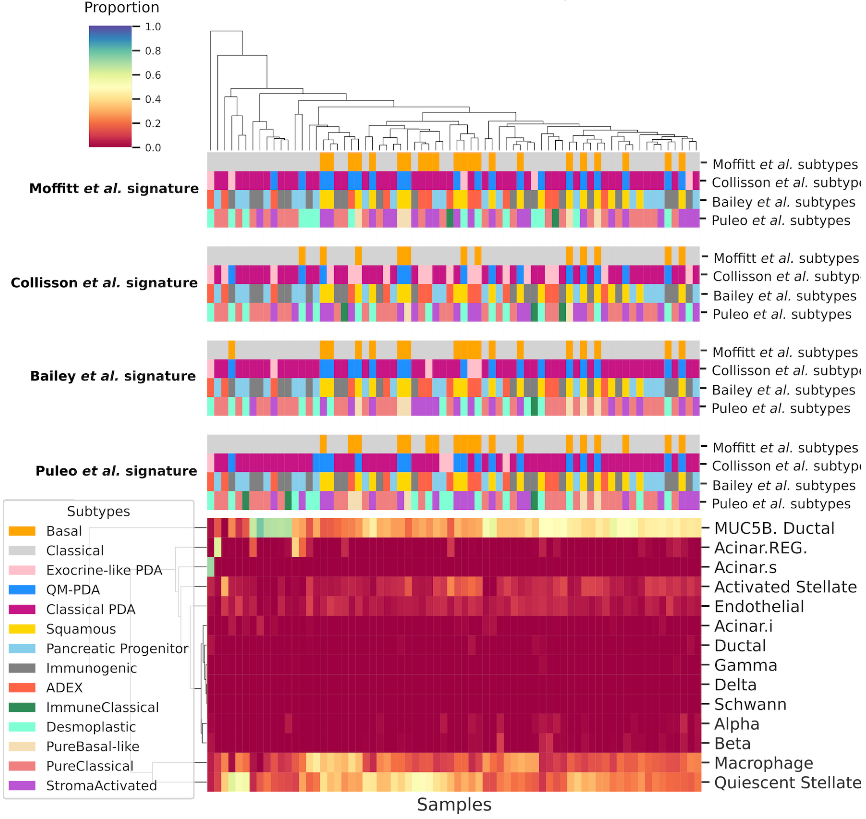
Bailey et al. signatures - Adjusted Rand Index between clusters



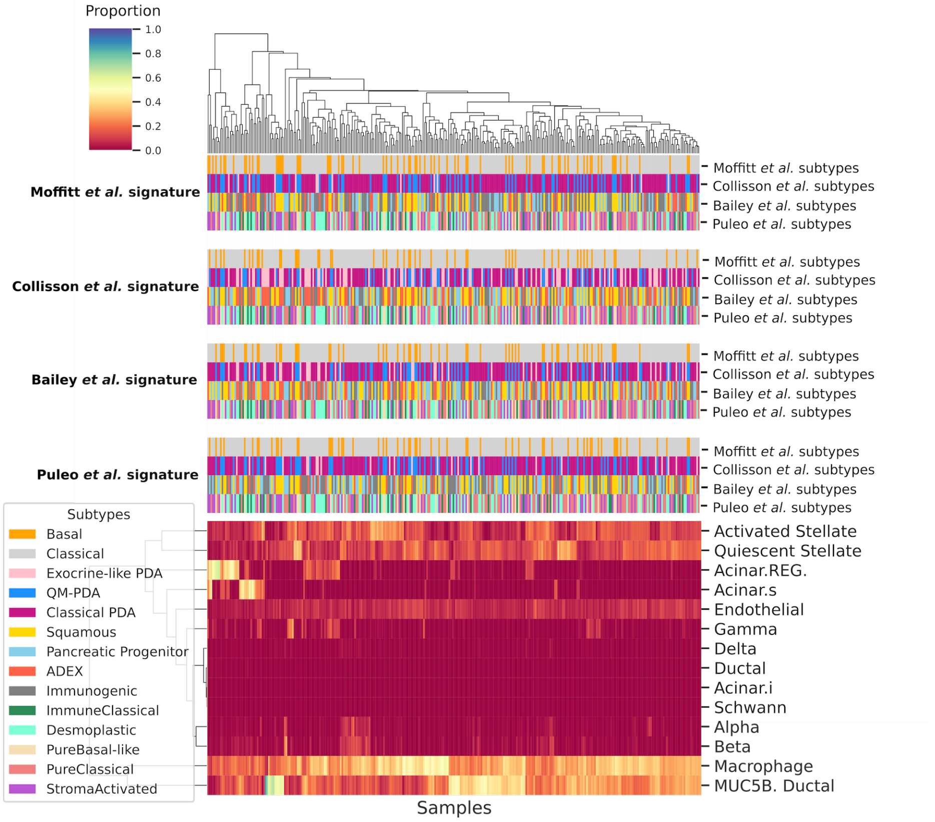
Puleo et al. signatures - Adjusted Rand Index between clusters

**Figure 5.7** Cell deconvolution estimated the healthy pancreas cell types proportion in each of the nine dataset. Clustering analysis is used to identify enrichment of particular cells in the predicted subtypes, shown on top of each heatmap.

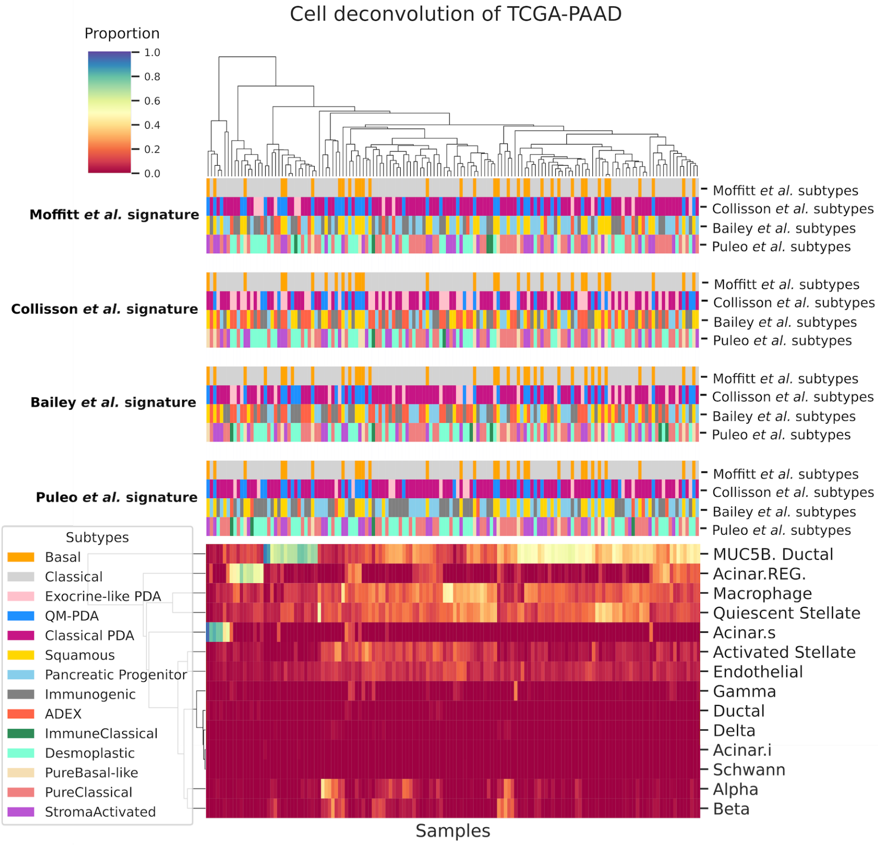


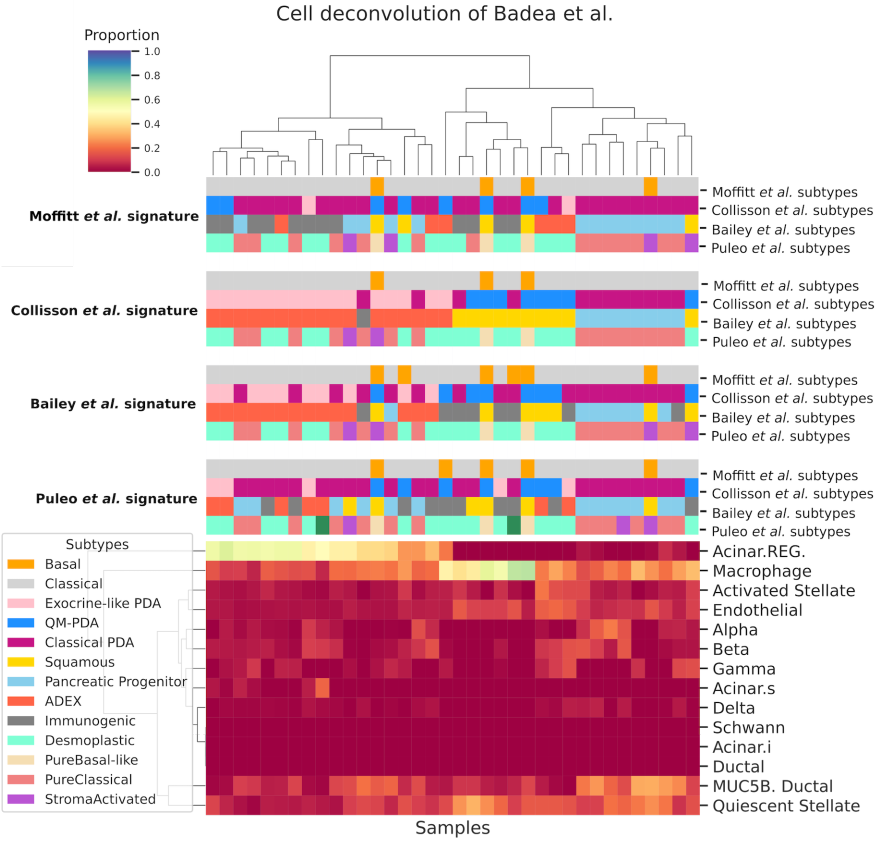


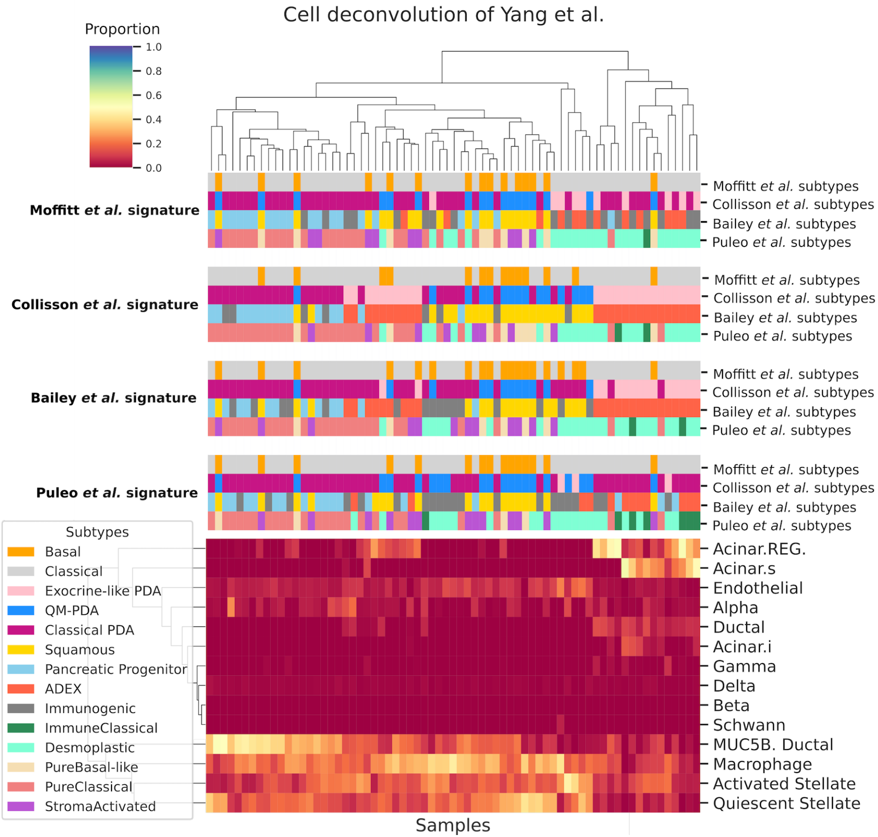
Cell deconvolution of Bailey et al.

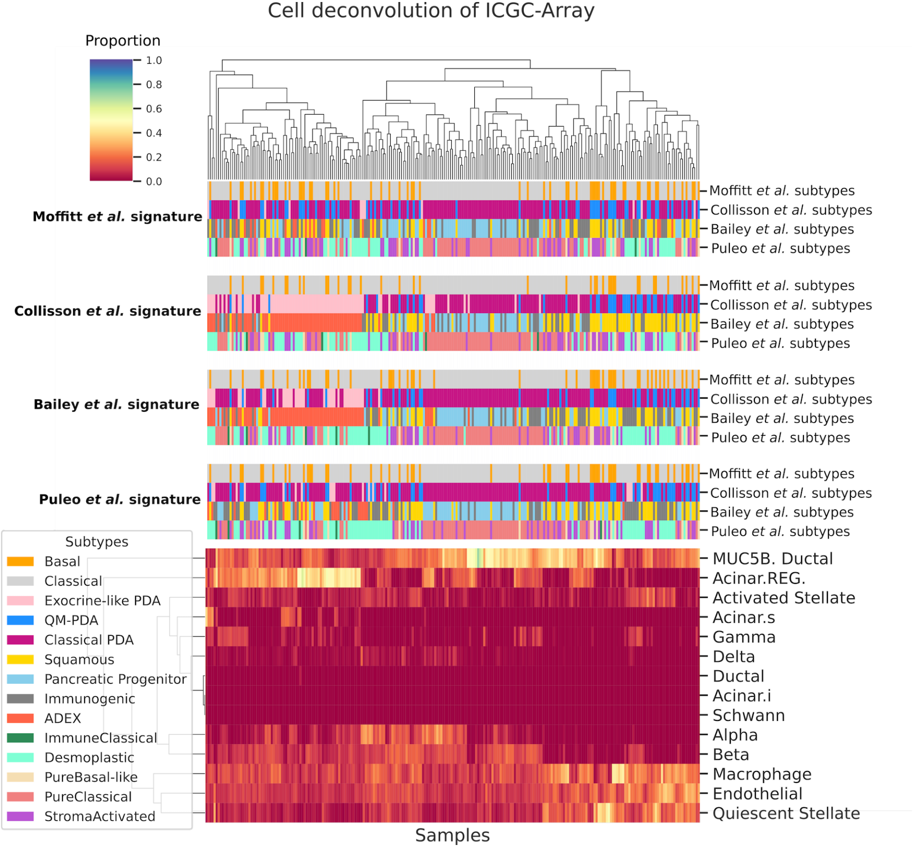


Cell deconvolution of Puleo et al.


Cell deconvolution of TCGA-PAAD
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Moffitt et al. signature
Collisson et al. signature
Bailey et al. signature
Puleo et al. signature
Moffitt et al. subtypes
Collisson et al. subtypes
Bailey et al. subtypes
Puleo et al. subtypes
Subtypes
Basal
Classical
Exocrine-like PDA
QM-PDA
Classical PDA
Squamous
Pancreatic Progenitor
Immunogenic
ADEX
ImmuneClassical
Desmoplastic
PureBasal-like
PureClassical
StromaActivated
MUC5B. Ductal
Acinar.REG.
Macrophage
Quiescent Stellate
Acinar.s
Activated Stellate
Endothelial
Gamma
Ductal
Delta
Acinar.i
Schwann
Alpha
Beta
Samples


Cell deconvolution of Badea et al.
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Moffitt et al. signature
Collisson et al. signature
Bailey et al. signature
Puleo et al. signature
Moffitt et al. subtypes
Collisson et al. subtypes
Bailey et al. subtypes
Puleo et al. subtypes
Subtypes
Basal
Classical
Exocrine-like PDA
QM-PDA
Classical PDA
Squamous
Pancreatic Progenitor
ADEX
Immunogenic
Desmoplastic
PureBasal-like
PureClassical
StromaActivated
Acinar.REG.
Macrophage
Activated Stellate
Endothelial
Alpha
Beta
Gamma
Acinar.s
Delta
Schwann
Acinar.i
Ductal
MUC5B. Ductal
Quiescent Stellate
Samples


Cell deconvolution of Yang et al.
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Moffitt et al. signature
Collisson et al. signature
Bailey et al. signature
Puleo et al. signature
Moffitt et al. subtypes
Collisson et al. subtypes
Bailey et al. subtypes
Puleo et al. subtypes
Subtypes
Basal
Classical
Exocrine-like PDA
QM-PDA
Classical PDA
Squamous
Pancreatic Progenitor
ADEX
Immunogenic
ImmuneClassical
Desmoplastic
PureBasal-like
PureClassical
StromaActivated
Acinar.REG.
Acinar.s
Endothelial
Alpha
Ductal
Acinar.i
Gamma
Delta
Beta
Schwann
MUC5B. Ductal
Macrophage
Activated Stellate
Quiescent Stellate
Samples


Cell deconvolution of ICGC-Array
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Moffitt et al. signature
Collisson et al. signature
Bailey et al. signature
Puleo et al. signature
Moffitt et al. subtypes
Collisson et al. subtypes
Bailey et al. subtypes
Puleo et al. subtypes
Subtypes
Basal
Classical
Exocrine-like PDA
QM-PDA
Classical PDA
Squamous
Pancreatic Progenitor
Immunogenic
ADEX
ImmuneClassical
Desmoplastic
PureBasal-like
PureClassical
StromaActivated
MUC5B. Ductal
Acinar.REG.
Activated Stellate
Acinar.s
Gamma
Delta
Ductal
Acinar.i
Schwann
Alpha
Beta
Macrophage
Endothelial
Quiescent Stellate
Samples

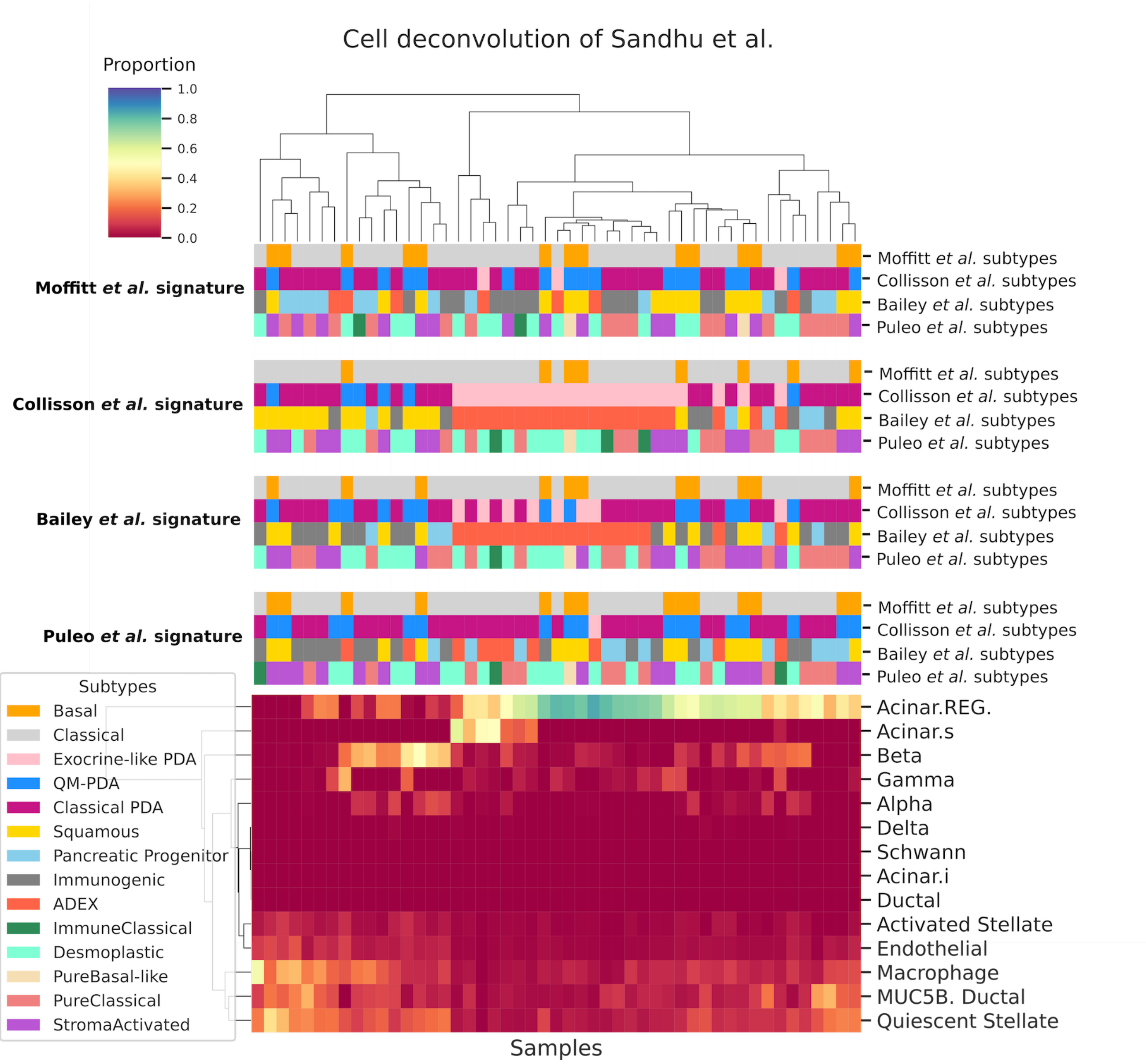
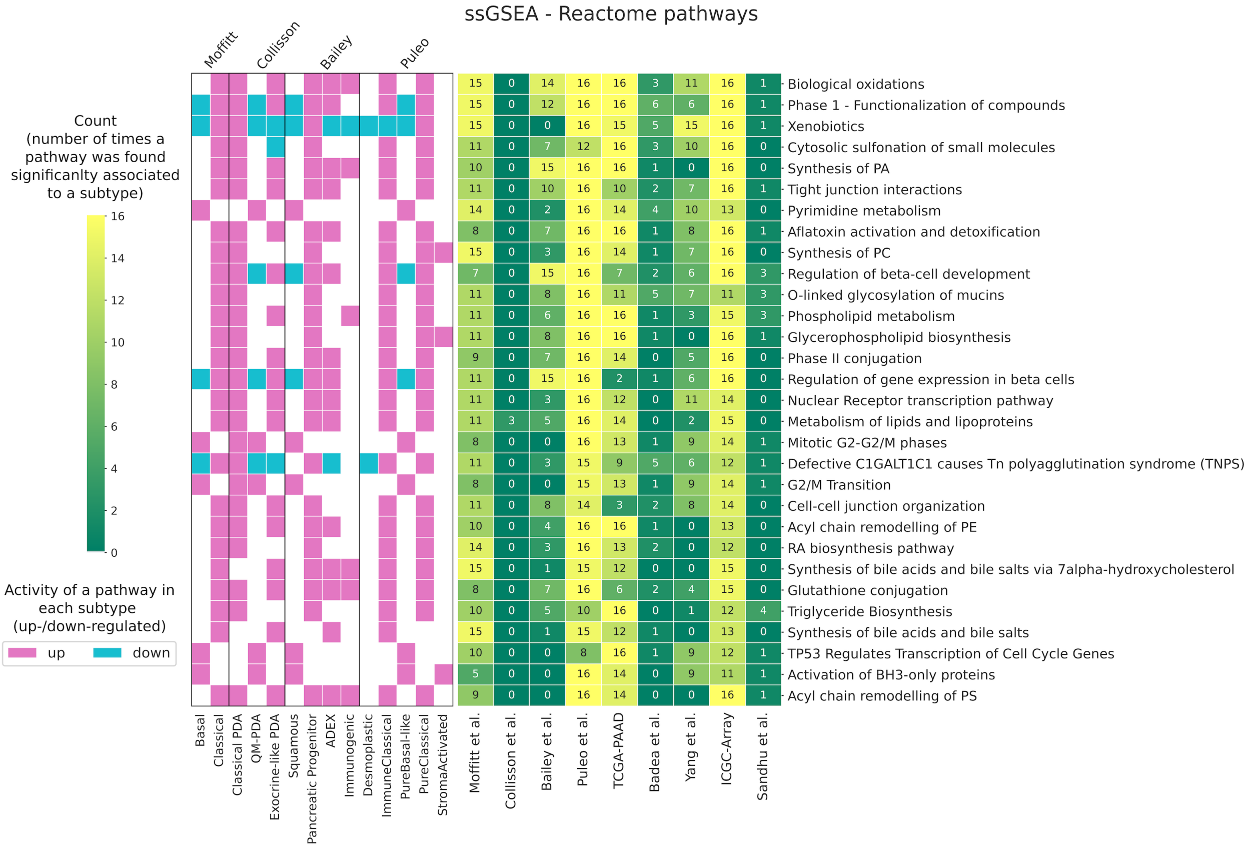
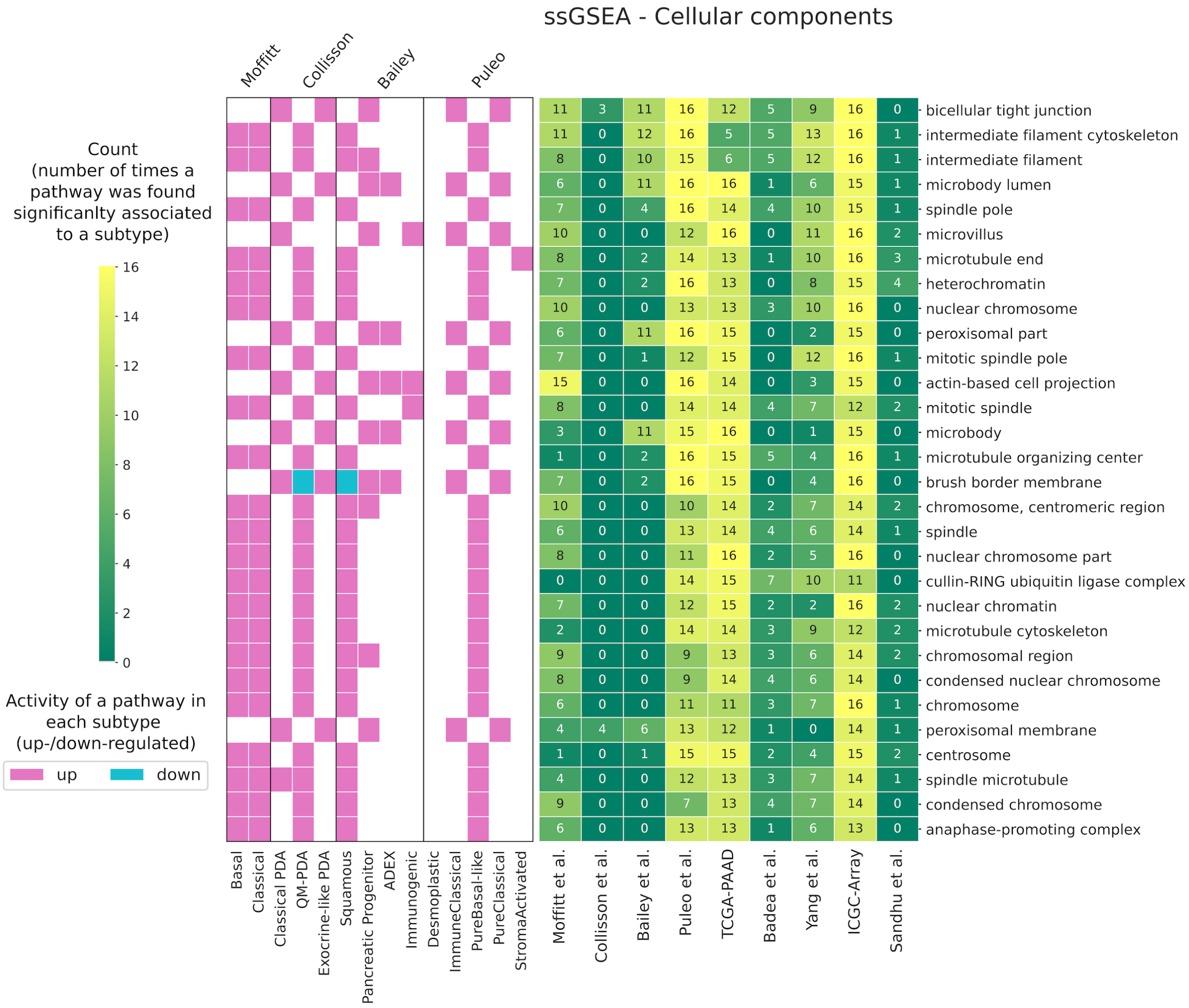
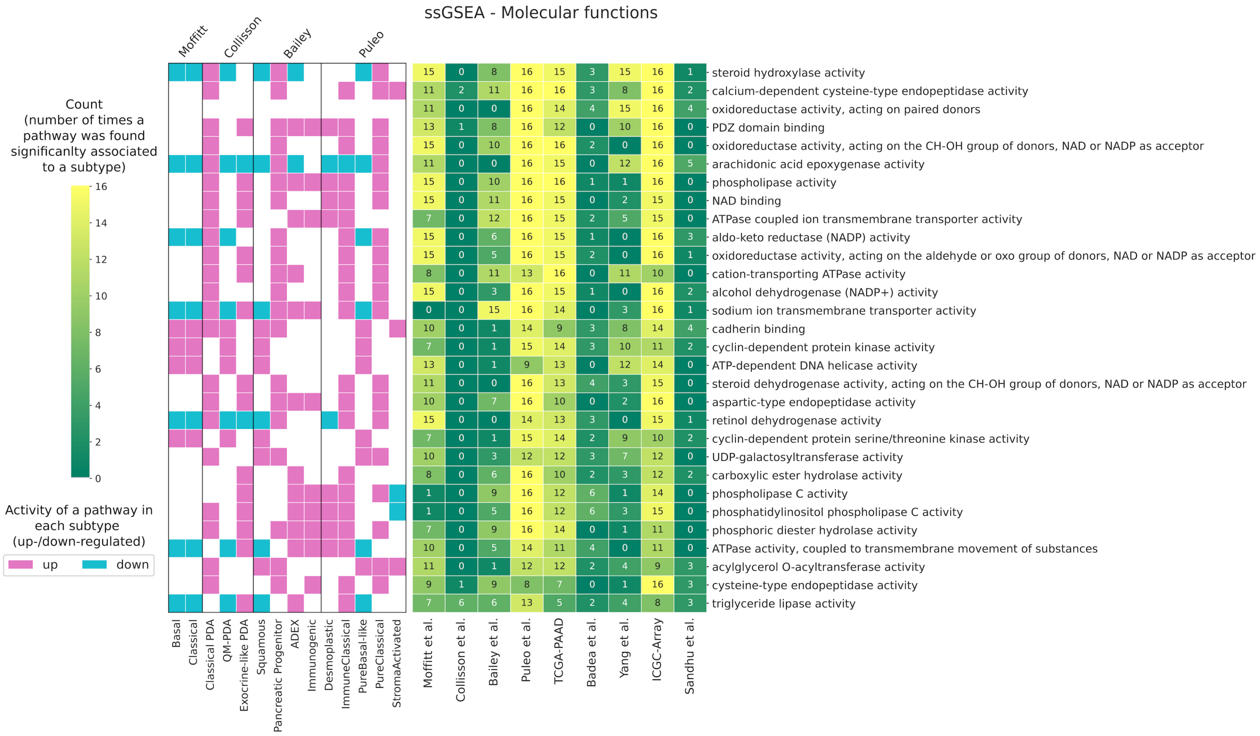

Cell deconvolution of Sandhu et al.
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Moffitt et al. signature
Collisson et al. signature
Bailey et al. signature
Puleo et al. signature
Moffitt et al. subtypes
Collisson et al. subtypes
Bailey et al. subtypes
Puleo et al. subtypes
Subtypes
Basal
Classical
Exocrine-like PDA
QM-PDA
Classical PDA
Squamous
Pancreatic Progenitor
Immunogenic
ADEX
ImmuneClassical
Desmoplastic
PureBasal-like
PureClassical
StromaActivated
Acinar.REG.
Acinar.s
Beta
Gamma
Alpha
Delta
Schwann
Acinar.i
Ductal
Activated Stellate
Endothelial
Macrophage
MUC5B. Ductal
Quiescent Stellate
Samples

**Figure 5.8** ssGSEA results. Each of the nine validation dataset was subject to ssGSEA, results from Reactome pathways repository and the GO terms are shown below where only the 30 terms found most associated across the nine cohorts are reported. Additionally, whether a term is up-/down-regulated is shown on the left of the heatmap.




ssGSEA - Cellular components
Moffitt
Collisson
Bailey
Puleo
Count
(number of times a
pathway was found
significanlty associated
to a subtype)
16
14
12
10
8
6
4
2
0
Activity of a pathway in
each subtype
(up-/down-regulated)
up
down
11 3 11 16 12 5 9 16 0 bicellular tight junction
11 0 12 16 5 5 13 16 1 intermediate filament cytoskeleton
8 0 10 15 6 5 12 16 1 intermediate filament
6 0 11 16 16 1 6 15 1 microbody lumen
7 0 4 16 14 4 10 15 1 spindle pole
10 0 0 12 16 0 11 16 2 microvillus
8 0 2 14 13 1 10 16 3 microtubule end
7 0 2 16 13 0 8 15 4 heterochromatin
10 0 0 13 13 3 10 16 0 nuclear chromosome
6 0 11 16 15 0 2 15 0 peroxisomal part
7 0 1 12 15 0 12 16 1 mitotic spindle pole
15 0 0 16 14 0 3 15 0 actin-based cell projection
8 0 0 14 14 4 7 12 2 mitotic spindle
3 0 11 15 16 0 1 15 0 microbody
1 0 2 16 15 5 4 16 1 microtubule organizing center
7 0 2 16 15 0 4 16 0 brush border membrane
10 0 0 10 14 2 7 14 2 chromosome, centromeric region
6 0 0 13 14 4 6 14 1 spindle
8 0 0 11 16 2 5 16 0 nuclear chromosome part
0 0 0 14 15 7 10 11 0 cullin-RING ubiquitin ligase complex
7 0 0 12 15 2 2 16 2 nuclear chromatin
2 0 0 14 14 3 9 12 2 microtubule cytoskeleton
9 0 0 9 13 3 6 14 2 chromosomal region
8 0 0 9 14 4 6 14 0 condensed nuclear chromosome
6 0 0 11 11 3 7 16 1 chromosome
4 4 6 13 12 1 0 14 1 peroxisomal membrane
1 0 1 15 15 2 4 15 2 centrosome
4 0 0 12 13 3 7 14 1 spindle microtubule
9 0 0 7 13 4 7 14 0 condensed chromosome
6 0 0 13 13 1 6 13 0 anaphase-promoting complex
Basal
Classical
Classical PDA
QM-PDA
Exocrine-like PDA
Squamous
Pancreatic Progenitor
ADEX
Immunogenic
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
Moffitt et al.
Collisson et al.
Bailey et al.
Puleo et al.
TCGA-PAAD
Badea et al.
Yang et al.
ICGC-Array
Sandhu et al.


ssGSEA - Molecular functions
Moffitt
Collisson
Bailey
Puleo
Count
(number of times a
pathway was found
significanlty associated
to a subtype)
16
14
12
10
8
6
4
2
0
Activity of a pathway in
each subtype
(up-/down-regulated)
up
down
15 0 8 16 15 3 15 16 1 steroid hydroxylase activity
11 2 11 16 16 3 8 16 2 calcium-dependent cysteine-type endopeptidase activity
11 0 0 16 14 4 15 16 4 oxidoreductase activity, acting on paired donors
13 1 8 16 12 0 10 16 0 PDZ domain binding
15 0 10 16 16 2 0 16 0 oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
11 0 0 16 15 0 12 16 5 arachidonic acid epoxygenase activity
15 0 10 16 16 1 1 16 0 phospholipase activity
15 0 11 16 15 0 2 15 0 NAD binding
7 0 12 16 15 2 5 15 0 ATPase coupled ion transmembrane transporter activity
15 0 6 16 15 1 0 16 3 aldo-keto reductase (NADP) activity
15 0 5 16 15 2 0 16 1 oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
8 0 11 13 16 0 11 10 0 cation-transporting ATPase activity
15 0 3 16 15 1 0 16 2 alcohol dehydrogenase (NADP+) activity
0 0 15 16 14 0 3 16 1 sodium ion transmembrane transporter activity
10 0 1 14 9 3 8 14 4 cadherin binding
7 0 1 15 14 3 10 11 2 cyclin-dependent protein kinase activity
13 0 1 9 13 0 12 14 0 ATP-dependent DNA helicase activity
11 0 0 16 13 4 3 15 0 steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
10 0 7 16 10 0 2 16 0 aspartic-type endopeptidase activity
15 0 0 14 13 3 0 15 1 retinol dehydrogenase activity
7 0 1 15 14 2 9 10 2 cyclin-dependent protein serine/threonine kinase activity
10 0 3 12 12 3 7 12 0 UDP-galactosyltransferase activity
8 0 6 16 10 2 3 12 2 carboxylic ester hydrolase activity
1 0 9 16 12 6 1 14 0 phospholipase C activity
1 0 5 16 12 6 3 15 0 phosphatidylinositol phospholipase C activity
7 0 9 16 14 0 1 11 0 phosphoric diester hydrolase activity
10 0 5 14 11 4 0 11 0 ATPase activity, coupled to transmembrane movement of substances
11 0 1 12 12 2 4 9 3 acylglycerol O-acyltransferase activity
9 1 9 8 7 0 1 16 3 cysteine-type endopeptidase activity
7 6 6 13 5 2 4 8 3 triglyceride lipase activity
Basal
Classical
Classical PDA
QM-PDA
Exocrine-like PDA
Squamous
Pancreatic Progenitor
ADEX
Immunogenic
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
Moffitt et al.
Collisson et al.
Bailey et al.
Puleo et al.
TCGA-PAAD
Badea et al.
Yang et al.
ICGC-Array
Sandhu et al.

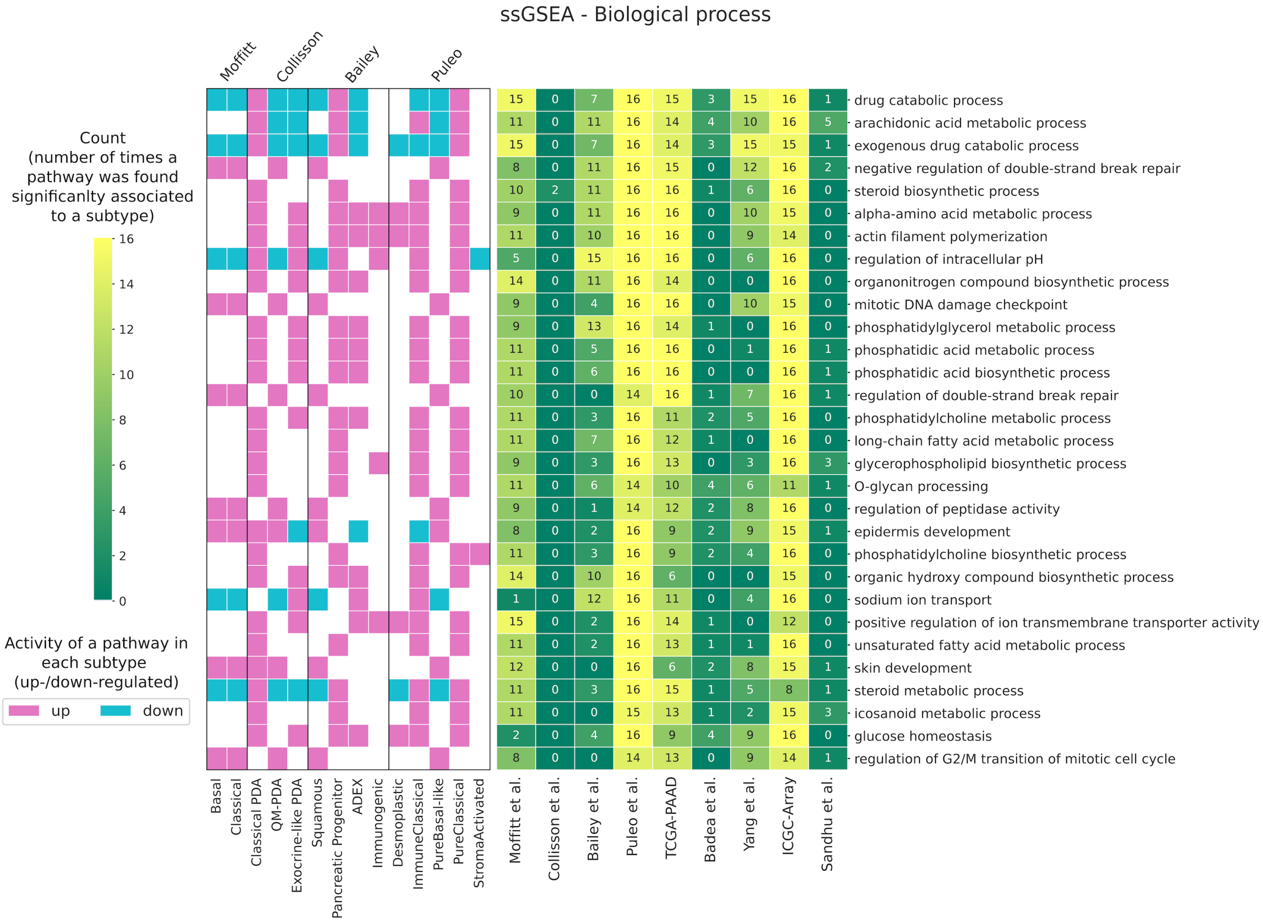
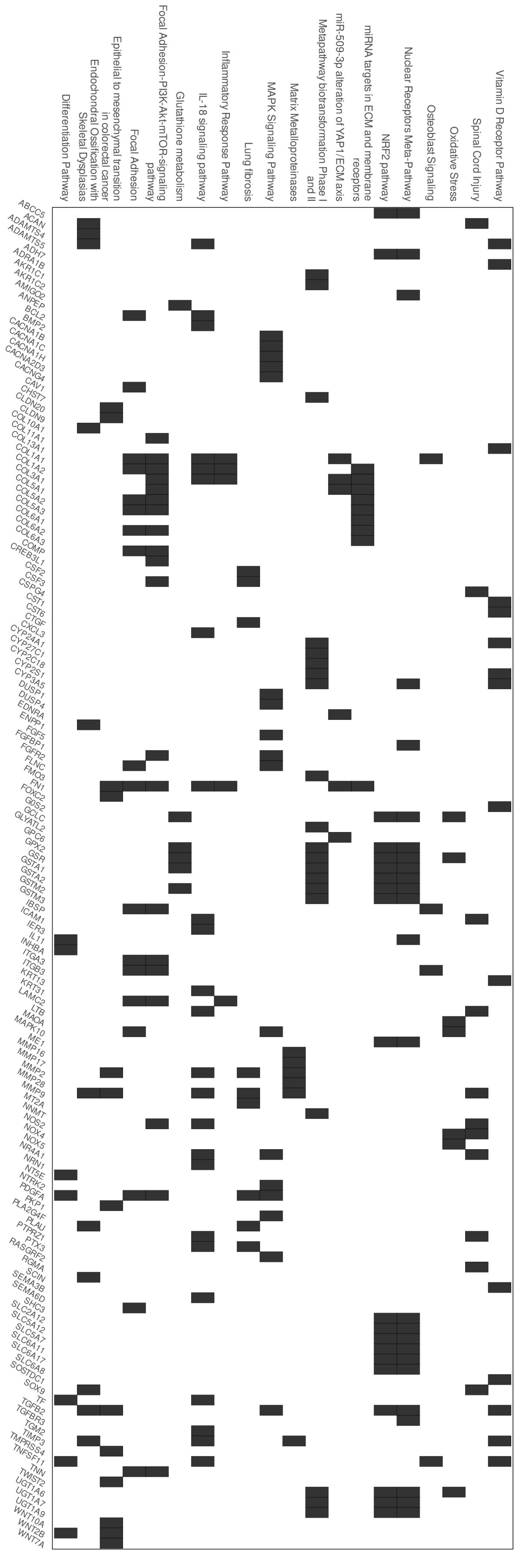

ssGSEA - Biological process
Moffitt
Collisson
Bailey
Puleo
Count
(number of times a
pathway was found
significanlty associated
to a subtype)
16
14
12
10
8
6
4
2
0
Activity of a pathway in
each subtype
(up-/down-regulated)
up
down
Basal
Classical
Classical PDA
QM-PDA
Exocrine-like PDA
Squamous
Pancreatic Progenitor
ADEX
Immunogenic
Desmoplastic
ImmuneClassical
PureBasal-like
PureClassical
StromaActivated
Moffitt et al.
Collisson et al.
Bailey et al.
Puleo et al.
TCGA-PAAD
Badea et al.
Yang et al.
ICGC-Array
Sandhu et al.
15 0 7 16 15 3 15 16 1 drug catabolic process
11 0 11 16 14 4 10 16 5 arachidonic acid metabolic process
15 0 7 16 14 3 15 15 1 exogenous drug catabolic process
8 0 11 16 15 0 12 16 2 negative regulation of double-strand break repair
10 2 11 16 16 1 6 16 0 steroid biosynthetic process
9 0 11 16 16 0 10 15 0 alpha-amino acid metabolic process
11 0 10 16 16 0 9 14 0 actin filament polymerization
5 0 15 16 16 0 6 16 0 regulation of intracellular pH
14 0 11 16 14 0 0 16 0 organonitrogen compound biosynthetic process
9 0 4 16 16 0 10 15 0 mitotic DNA damage checkpoint
9 0 13 16 14 1 0 16 0 phosphatidylglycerol metabolic process
11 0 5 16 16 0 1 16 1 phosphatidic acid metabolic process
11 0 6 16 16 0 0 16 1 phosphatidic acid biosynthetic process
10 0 0 14 16 1 7 16 1 regulation of double-strand break repair
11 0 3 16 11 2 5 16 0 phosphatidylcholine metabolic process
11 0 7 16 12 1 0 16 0 long-chain fatty acid metabolic process
9 0 3 16 13 0 3 16 3 glycerophospholipid biosynthetic process
11 0 6 14 10 4 6 11 1 O-glycan processing
9 0 1 14 12 2 8 16 0 regulation of peptidase activity
8 0 2 16 9 2 9 15 1 epidermis development
11 0 3 16 9 2 4 16 0 phosphatidylcholine biosynthetic process
14 0 10 16 6 0 0 15 0 organic hydroxy compound biosynthetic process
1 0 12 16 11 0 4 16 0 sodium ion transport
15 0 2 16 14 1 0 12 0 positive regulation of ion transmembrane transporter activity
11 0 2 16 13 1 1 16 0 unsaturated fatty acid metabolic process
12 0 0 16 6 2 8 15 1 skin development
11 0 3 16 15 1 5 8 1 steroid metabolic process
11 0 0 15 13 1 2 15 3 icosanoid metabolic process
2 0 4 16 9 4 9 16 0 glucose homeostasis
8 0 0 14 13 0 9 14 1 regulation of G2/M transition of mitotic cell cycle

**Figure 5.9** Results from the enrichment analysis performed on the genes whose expression was found significantly associated with TB. For each term, part of the WikiPathway repository, is possible to see the genes enriched. Only significant terms (p-value < 0.05) are shown.



# Acknowledgement

Approaching the end of this long journey I do not know where to start to thank all the amazing people that made this possible. All the challenges I went through, including changing country and facing the COVID wave spreading, were supported by all the friends and colleagues I have met since day one. I will start by thanking Markus List, who supported me so much that a normal acknowledgment will never be enough. I am truly grateful for his help and the extreme kindness and with no doubt, I can call myself lucky for being able to work with him. I thank Jan Baumbach for giving me the opportunity to take part in his group and giving me all the support for initiating this journey. I thank Tim Kacprowski for his support in all my projects and for always being open to listening to my overthinking. He has been such a mentor and I have learned so much from him. I wanna thank Dietmar Zehn for his availability in supporting me in the final stage of my Ph.D. I would like to thank all the people from the Klinikum Rechts der Isar that welcomed me to their team. I thank Wilko Weichert, who gave me the chance to work in the pathology group, and Katja Steiger and Nicole Pfarr for being so supportive and for guiding me since the beginning of my Ph.D. There are so many more people I would like to thank that made my life easier and who I see as friends more than co-workers. I can not thank Martina enough, she always has been a reference point for all of us and she helped me in the most difficult moments. I have to thank Drazen, who had to deal with all my questions and weird issues and I thank him for being so patient and always available in helping others. I wanna thank my colleagues, that I would rather call friends. Among them, Marisol, Gihanna, Amit, Chit, Zakaria and so many more that I can not list or I would need another whole page. We always supported and helped each other and had so much fun that I am going to miss all of you so much (I already do since we already are spread all over Europe). I am so grateful for having worked in a stimulating and friendly environment and all the work I have done and everything I have learned it would not have been possible without these special people I have met. If I think of myself from the beginning I notice such a big difference and I am so proud of how I grew, both personally and especially professionally. I feel sad knowing that this chapter is closing, but extremely excited to jump into the next one and see what life brings next!