Technische Universität München

TUM School of Computation,

Information and Technology

Deciphering Trans-acting Regulatory Molecular Mechanisms with Machine Learning

Toray Sami Akcan

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Julia A. Schnabel

Prüfer*innen der Dissertation:

1. TUM Junior Fellow Dr. Matthias Heinig
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 06.10.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 07.03.2023 angenommen.

DEPARTMENT OF INFORMATICS

Technical University of Munich

Dissertation

Deciphering Trans-acting Regulatory Molecular Mechanisms with Machine Learning

Toray Sami Akcan

June 2022

HelmholtzZentrum münchen
German Research Center for Environmental Health

Acknowledgments

I would like to express my deepest gratitude for the continuous support of all the people during the entire time of my Ph.D.

Special thanks go to my Ph.D. advisor Dr. Matthias Heinig for allowing me to perform my Ph.D. and conduct such exciting research. Continuous support with insightful discussions and directions were essential for the success, and I would like to thank you for all your help, patience, and supervision!

A big thanks go to all the great people in the Heinig Lab and the ICB. It was indeed an extraordinary experience to be part of the ICB family, a work environment that truly stands out, providing the opportunity for creative expression, the security of equality, the competence of highly intellectual people, and the chance to conduct highly innovative state-of-the-art research. Many social activities and interactions were vital for a fun and interactive Ph.D. experience. Special thanks go to Prof. Fabian Theis who makes this possible.

I would also like to thank Prof. Heribert Schunkert from the German Heart Center for collaborative work on fascinating and relevant topics!

Special thanks also go to my thesis advisory committee, including Prof. Dr. Robert Schneider, Prof. Dr. Maria Colomé-Tatché, and Dr. Matthias Heinig. Thank you all for your time, support, and positive feedback, making all this possible!

Finally, I want to shout out a big thanks to all my family members, especially my parents, Imran and Okyanus, my sister Ebru and my brother-in-law Volkan, as well as my aunt Asiye, for accompanying me on this journey and supporting me in any way conceivable! Thanks to all my friends for their continuous and relentless encouragement and for making me smile and feel strong!

You are all awesome. Thank you all !

Toray Sami Akcan, June 2022

Preface

This dissertation results from two research projects that I conducted at the Institute of Computational Biology (ICB) at the Helmholtz Research Center in Munich (HMGU). It entails the entire biological and methodological context to understand the projects better. In the following two paragraphs, I give individual project overviews, highlight my contributions, and acknowledge the important work of my collaborators.

Transcriptional Pausing Project [1]:

In this project, I contributed to improving our understanding of trans-regulatory factors implicated in the transcriptional pausing of the Polymerase II that underlies the transcription of mammalian protein-coding genes. I performed all computational analyses in this project. The project is based on the integration of large-scale genomic data sets of genomic and transcriptomic binding events as well as gene annotation and sequence composition features as context for a machine learning task to extract patterns of trans-regulation modulating the productivity of the Polymerase II. Our results provide the first comprehensive characterization of trans-regulators underlying transcriptional pausing enabling a systematic and targeted investigation thereof by providing specific factors for experimental manipulation. This project is presented in Chapter 3, and the corresponding manuscript is currently (May 2022) under review in *Nucleic Acid Research* and also available on bioRxiv: Akcan and Heinig. 'Predictive Model of Transcriptional Elongation Control Identifies Trans-Regulatory Factors from Chromatin Signatures, Toray Akcan, Matthias Heinig' BioRxiv (2022). The related chapter includes all parts of it, including figures, essentially replicating the manuscript. I want to express my gratitude to Dr. Heinig, who provided me the opportunity to conduct this research which earned me first authorship for this important work, for which I am very grateful.

Coronary Artery Disease Epistasis Project [2]:

In my second project, I contributed to further our knowledge of genetic interactions as additional drivers of Coronary Artery Disease (CAD). This project was a collaborative effort between the Institute of Computational Biology (ICB) at the Helmholtz Zentrum München, the German Heart Center in Munich led by Prof. Heribert Schunkert and Prof. Johan LM Björkegren at the Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, NY, USA, and the Karolinska Institutet, Sweden. The aim was to identify and characterize trans genetic interactions underlying Coronary Artery Disease. I performed all computational analyses in this project. It is based on integrating large-scale genotype, phenotype, and gene expression data for CAD. Developing a filter-based permutation testing approach coupled with linear modeling and subsequent genotype-combination-dependent differential gene expression analyses, enabled us to identify upstream trans-genetic regulatory interactions with downstream regulatory roles on trans-target genes. Our results provide specific interacting genetic loci and allele combinations as well as downstream trans target genes as specific experimental endpoints for a systematic experimental manipulation to further characterize the effects of genetic interaction that contribute to the disease etiology of CAD. I am very grateful to Prof. Schunkert, and Prof. Björkegren for having had access to the underlying data sets to perform respective analyses, without which this work would not have been possible. This project is presented in Chapter 4 with the corresponding manuscript currently (June 2022) in preparation. The related chapter includes all parts of it, including figures, essentially replicating the manuscript. I want to especially thank Dr. Heinig, who initiated me into this collaboration and allowed me to conduct such exciting research, which earned me first authorship for this vital work, for which I am very grateful.

Abstract

Advancements in the digitalization of our societies greatly affect healthcare and science, enabling the investigation of biological phenomena at much greater speed and precision. In fact, contemporaneous developments of high-throughput assays that produce massive amounts of complex biological data, also known as multi-omics data, strictly require data-driven approaches that have the capacity to cope with the quantity, quality, complexity, and interconnectivity of such data types. Artificial intelligence and machine learning have intrinsic properties that harbor these potentials and can overcome these barriers. As such it has become an essential tool in computational biology.

Likewise, we conducted machine learning-driven research of large-scale biological multi-omic data types that aimed to address major challenges in the biological domain of gene regulation, specifically from the perspective of trans-regulatory molecular mechanisms. As opposed to cis-regulatory elements that regulate a limited number of proximal targets that lie in the vicinity of the regulators, trans-regulatory factors act on genes that are distal and lie far away from the regulatory elements. As such, trans-regulators are not constrained in their function by their localization and may act on any distal genomic element thereby potentially occupying roles as key master regulators. This highlights the importance of trans-regulation and indicates that for a holistic understanding of gene transcription and gene regulation beyond gene-proximal cis-regulatory elements, a comprehensive characterization of gene-distal trans-regulatory molecular mechanisms and implicated factors is indispensable. Trans-regulatory mechanisms are characterized by interactions of molecules that can be ascribed to different pools of biological entities with specific characteristics. For instance, interactions between genome-associated elements, like DNA-binding transcription factors, and transcriptome-associated elements, like splicing factors. Due to this multi-modal nature of trans-regulation, elaborate multi-omic data integration and analysis strategies, as it is enabled by machine learning, are required. In this context, we conducted machine learning-driven analyses of various multi-omic data types contributing to this area of research about trans-regulatory molecular mechanisms of gene regulation by focusing specifically on **1)** the identification of novel trans-regulatory elements modulating the promoter-proximal pausing of the Polymerase II (Pol II) during the transcription of mammalian protein-coding genes and **2)** the identification of trans genetic interactions as upstream trans-regulators of downstream trans target genes that underlie Coronary Artery Disease (CAD).

The transcription of genes is characterized by regulated transcriptional arrests of the Pol II, called transcriptional pauses. Transcriptional pause regulatory elements entail cis- and trans-acting factors like DNA sequence motifs and transcription factor bindings that control Pol II pausing during the transcription of mammalian protein-coding genes. A special case of transcriptional pausing is promoter-proximal Pol II pausing. It represents a key rate-limiting step to gene expression and is recognized as a hallmark of protein-coding genes. Despite its important role in gene transcription, we still lack quantitative descriptions of implicated regulatory elements. Predictive models could have the potential to identify previously unknown pause regulatory elements and reveal their relative importance. We addressed this gap with an Extreme Gradient Boosting Tree regression model that accurately predicts (Pearson's rho 0.83, R^2 0.68) the degree of promoter-proximal Pol II pausing and explains almost up to 70% of the observed variance. This was accomplished by engineering features of genome and transcriptome protein binding maps from large-scale Chromatin immunoprecipitation sequencing data (CHIP-seq) and enhanced Cross-linking immunoprecipitation sequencing data (eCLIP-seq) to reveal potential novel trans-acting protein factors. Additional features of gene annotation and sequence compositions served to capture potential novel pause regulatory elements that are intrinsic (cis-acting) to the genes. Pol II productivity, as the target in our models, was quantified using Global-Run-On-sequencing (GRO-seq) data. By validating the obtained model on an independent cell line we demonstrated the generalizability and cell line agnostic character of the model. An array of models based on sets of proteins with specific molecular functional backgrounds further confirmed the strong interconnected nature of the transcriptional pause mechanism with other RNA regulatory processes like for example splicing. By harnessing the model feature contributions we quantified and elucidated the relative importance of individual factors which enabled us to systematically identify previously unknown regulators of pausing with high predictive value. In addition, we identified nine previously unknown

7SK non-coding-RNA interacting RNA-binding proteins predictive of pausing, further strengthening and elucidating the role of the 7SK pause mediator complex and implicated factors in transcriptional pausing. To conclude, our results provide specific proteins predictive of transcriptional pausing for experimental manipulation, for instance with gene knockdown experiments, to evaluate their downstream effects on pausing and gene expression at large.

In the second project, we applied statistical inference techniques to genetic data of patients with Coronary Artery Disease (CAD), a major cause of death worldwide both in developed and developing countries. Despite numerous case-control, epidemiological, quantitative trait loci, and genome-wide association studies, mechanistic details for CAD still remain to be understood. Particularly we lack investigations into genetic interactions (epistasis) which have profound impacts on many quantitative traits not only in humans but also in bacteria and other higher-order model organisms. We addressed this gap with a filter-based permutation testing approach coupled with linear modeling, identifying n=4 interacting SNP pairs through the evaluation of differential SNP correlations between CAD cases and controls in a large cohort of >35k samples (UK Biobank) and validating the SNP interactions by modeling the disease label (case-control status) dependent on individual interacting SNP pairs as multiplicative terms in logistic regression models. For the first time, we were also able to replicate the genetic interactions with analogous models in an independent aggregate cohort of 11 genome-wide case-control studies with >35k samples. Subsequently, we evaluated the downstream effects of these trans epistatic interactions on the transcriptional output of trans target genes. This was achieved by integrating genotype and gene expression data for multiple tissues from the STARNET study and GTEx v8 project and modeling tissue-specific individual gene expressions by each possible genotype-combination of an interacting SNP pair as a multiplicative term in linear regression models. Permutation testing allowed us to further rule out trans differential gene expression results that were most likely due to chance. In total, we identified n=1142 epistatically driven differentially expressed trans target genes in the STARNET cohort that could be replicated (FDR <5%) in the independent GTEx v8 cohort. Few (n=6) of the trans targets were in addition differentially expressed in dependence of the same interacting genetic variants in the same genotype combination at those variants, and showed the same direction of effect. Strikingly, n=2 genes were also differentially expressed in the same tissue type, representing highly confident results of tissue-specific trans-epistatically dysregulated trans target genes. Many of these trans target genes are strongly associated with cardiovascular events. In this context, our results support the hypothesis that combinatorial differential gene regulation could explain the epistatic consequences and provide for the first time specific interacting genetic loci and genotype combinations as well as downstream trans target genes as specific experimental target points for a systematic experimental investigation of the downstream effects of genetic interactions underlying CAD.

Our studies focused on transcriptional regulation, specifically advancing our understanding of trans-regulatory molecular mechanisms. We demonstrated how machine learning coupled with the integration of multi-omic data, like genotype, phenotype, gene expression, or protein binding data, can systematically identify trans-regulatory factors, trans interacting genetic variants and trans epistatically regulated trans target genes. On the one hand, these enabled us to improve our understanding of the transcriptional regulation of protein-coding genes by deciphering the underlying trans-regulatory mechanisms and revealing trans-regulatory factors involved in the critical early steps of transcription. On the other hand, we were able to illustrate the importance of trans genetic interactions in disease, revealing epistatically dysregulated trans-target disease genes. In large, our projects provide a systematic investigation of trans-regulatory molecular mechanisms providing specific experimental endpoints for a targeted investigation of implicated factors and the regulatory processes at large.

Kurzfassung

Fortschritte bei der Digitalisierung unserer Gesellschaften wirken sich stark auf das Gesundheitswesen und die Wissenschaft aus und ermöglichen die Untersuchung biologischer Phänomene mit viel größerer Geschwindigkeit und Präzision. Tatsächlich erfordern Hochdurchsatz-Assays, welche große Mengen komplexer biologischer Daten produzieren (Multi-Omic Daten), datengesteuerte Ansätze, die in der Lage sind, mit der Quantität, Qualität, Komplexität und Interkonnektivität der Daten zurechtzukommen. Künstliche Intelligenz und maschinelles Lernen besitzen Eigenschaften, die diese Potenziale bergen und diese Barrieren überwinden können. Als solches haben sie eine wesentliche Rolle in der Computerbiologie eingenommen.

Auch wir haben maschinellem Lernen genutzt um Forschung an groß angelegten biologischen Multi-Omic-Datentypen durchzuführen. Unser Fokus lag auf dem Bereich der Genregulation, insbesondere transregulatorischer molekularer Mechanismen. Im Gegensatz zu cis-regulatorischen Faktoren, die eine begrenzte Anzahl proximaler Ziele (z.B. Gene) regulieren, die in der Nähe der Regulatoren liegen, wirken transregulatorische Faktoren auf Ziele, die distal und weit entfernt von den regulierten Elementen liegen. Als solche sind Trans-Regulatoren im Gegensatz zu Cis-Regulatoren in ihrer Funktion nicht durch ihre Lokalisierung eingeschränkt und können auf jedes distale genomische Element einwirken, wodurch sie möglicherweise eine Rolle als Hauptregulatoren einnehmen. Dies unterstreicht die Bedeutung der Transregulation und weist darauf hin, dass für ein ganzheitliches Verständnis der Gentranskription und Genregulation über Gen-proximale cis-regulatorische Faktoren hinaus eine umfassende Charakterisierung Gen-distaler transregulatorischer molekularer Mechanismen und beteiligter Faktoren unerlässlich ist. Transregulationsmechanismen sind durch Wechselwirkungen von Molekülen gekennzeichnet, die bestimmten Pools biologischer Einheiten mit spezifischen Eigenschaften zugeordnet werden können. Beispielsweise Wechselwirkungen zwischen Genom-assoziierten Elementen wie DNA-bindenden Transkriptionsfaktoren und Transkriptom-assoziierten Elementen wie Spleißfaktoren. Aufgrund dieser multimodalen Eigenschaften der Transregulierung sind aufwändige Multi-Omic-Datenintegrations- und Analysestrategien erforderlich, wie sie durch maschinelles Lernen ermöglicht werden. In diesem Zusammenhang haben wir auf maschinellem Lernen basierende Analysen verschiedener Multi-Omic-Datentypen durchgeführt, die zu diesem Forschungsgebiet über transregulatorische molekulare Mechanismen der Genregulation beitragen, indem wir uns speziell auf **1)** die Identifizierung neuer transregulatorischer Elemente konzentrieren, die die promotor-proximale Pausierung der Polymerase II (Pol II) während der Transkription von menschlichen protein-kodierenden Genen und **2)** die Identifizierung transgenetischer Interaktionen als vorgeschaltete Transregulatoren von nachgeschalteten Transzielgenen, die der koronaren Herzkrankheit (KHK) zugrunde liegen, fokussierten.

Die Transkription von Genen ist durch regulierte Transkriptionspausen der Polymerase II gekennzeichnet. Regulatorische Elemente der transkriptionellen Pausierung umfassen cis- und trans-regulatorische Faktoren wie DNA Sequenzmotive und Transkriptionsfaktorbindungen, welche die Pausierung der Polymerase II während der Transkription von proteinkodierenden Genen in Säugetieren modulieren. Ein Sonderfall der transkriptionellen Pausierung ist die promotor-proximale Pausierung der Polymerase II, welches den Durchsatz der Genexpression kennzeichnend limitert. Damit stellt die transkriptionelle Pausierung einen essentiellen Mechanismus zur Regulation der Transkription dar, welches jedem proteinkodierenden Gen unterliegt. Jedoch fehlt uns eine umfassende quantitative Beschreibung aller beteiligten regulatorischen Faktoren. Vorhersagemethoden aus dem Bereich des maschinellen Lernens bergen das Potential bisher unbekannte regulatorische Elemente ausfindig zu machen und deren relativen Beiträge zur transkriptionellen Regulation aufzudecken. Wir nutzten dieses Potential aus, indem wir ein Extreme Gradient Boosting Tree Regressionsmodell trainierten, welches den Grad der promotor-proximalen Pausierung der Polymerase II akkurat (pearson's rho 0.83, R^2 0.68) vorhersagen konnte. Dies konnte durch die Erschließung von Genattributen über genomische und transkriptomische Proteinbindungen an Gensequenzen mithilfe von umfangreichen Chromatin Immunoprecipitation Sequencing (CHIP-seq) Daten und enhanced Cross-linking Immunoprecipitation Sequencing (eCLIP-seq) Daten zur Eingabe in das Vorhersagemodell ermöglicht werden. Genannotations- und sequenzattribute, als zusätzliche prädiktive Attribute, dienten dazu potentiel neue intrinsische cis-agierende regulatorische Elemente zu ermitteln.

Die Integration von Global-run-on-sequencing (GRO-seq) Daten ermöglichte die Quantifizierung der transkriptionellen Produktivität der Polymerase II und ermöglichte die Anwendung von Vorhersagemethoden indem es als Zielvariable in die Modelle einging. Durch die erfolgreiche Validierung des Modells auf Daten einer unabhängigen Zelllinie konnten wir die Verallgemeinerbarkeit des Modells und der darunter liegenden Modellattribute demonstrieren. Vorkenntnisse über molekulare Funktionen integrierter Faktoren konnten die starke Verflechtung des Mechanismus der transkriptionellen Pausierung mit anderen RNA-regulatorischen Prozessen wie zum Beispiel dem alternativen Spleißen bestätigen. Mithilfe der Modellstruktur zugrunde liegenden Merkmalsbeiträge konnten wir die relative Wichtigkeit einzelner Faktoren aufklären und quantifizieren, was uns ermöglichte bisher unbekannte Regulatoren mit hohen Vorhersagepotentialen systematisch zu ermitteln. Darüber hinaus identifizierten wir neun zuvor unbekannte 7SK-ncRNA interagierende RNA-bindende Proteine, welche die Rolle des 7SK-Komplexes in der transkriptionellen Pausierung weiter verstärkte. Unsere Ergebnisse ermöglichen eine systematische und gezielte Untersuchung der transkriptionellen Pausierung, indem sie spezifische Faktoren für die experimentelle Manipulation bereitstellen, beispielsweise für Knockdown-Experimente, um ihre Auswirkungen auf das transkiprionelle Pausieren und die Genexpression im Allgemeinen zu bewerten.

Im zweiten Projekt wandten wir statistische Inferenztechniken auf genetische Daten von Patienten mit koronarer Herzkrankheit (KHK) an, einer der häufigsten Todesursachen weltweit sowohl in Industrie- als auch in Entwicklungsländern. Trotz zahlreicher Studien müssen mechanistische Details der KHK noch verstanden werden. Insbesondere fehlt uns die Untersuchung genetischer Wechselwirkungen (Epistasen), welche tiefgreifende Auswirkungen auf viele quantitative Merkmale nicht nur beim Menschen, sondern auch bei Bakterien und anderen Modellorganismen höherer Ordnung haben. Wir konnten diese Lücke mit einem filter-basierten Permutationstestansatz in Verbindung mit linearen Modellen zur Validierung schließen und identifizierten n = 4 interagierende SNP-Paare durch die Evaluation von differentiellen SNP-Korrelationen zwischen KHK-Fällen und Kontrollen in einer großen Kohorte von > 35.000 Proben (UK Biobank). Zudem konnten wir zum ersten Mal die genetischen Interaktionen in einer unabhängigen Kohorte von 11 aggregierten genomweiten Fall-Kontroll-Studien mit ebenfalls >35.000 Proben und analogen Modellen replizieren. Anschließenden führten wir genotypkombinationsabhängige differenzielle trans Genexpressionsanalysen durch, um die nachgeschalteten Effekte der interagierenden SNP-Paaren auf die Transkription von trans Zielgenen zu bewerten. Dafür integrierten wir genotypische und phänotypische Daten für mehrere Gewebe aus der STARNET Studie und dem GTEx v8 Projekt und modellierten gewebespezifische individuelle Genexpressionen in Abhängigkeit von jeder möglichen Genotypkombination eines interagierenden SNP-Paares als multiplikativen Term in linearen Regressionsmodellen. Permutationstests, bei denen die beobachtete Verteilung von Genexpressionen in einer bestimmten Genotypkombination eines interagierenden SNP-Paares mit der erwarteten Verteilung von Genexpressionen in derselben Genotypkombination desselben SNP-Paares auf der Grundlage permutierter Daten verglichen wurde, ermöglichten uns Genexpressionsergebnisse, die mit hoher Wahrscheinlichkeit auf Zufall zurückzuführen sind, auszuschließen. Dabei identifizierten wir in der STARNET-Kohorte n=1142 epistatisch gesteuerte differentiell exprimierte trans-Zielgene, die in der unabhängigen GTEx v8 Kohorte repliziert werden konnten (FDR <5 %). Wenige (n=6) der trans-Zielgene wurden zusätzlich in Abhängigkeit derselben interagierenden genetischen Varianten in derselben Genotypkombination bei diesen Varianten differentiell exprimiert und zeigten dieselbe Wirkungsrichtung. Bemerkenswerterweise wurden auch n=2 Gene im selben Gewebetyp unterschiedlich exprimiert, was sehr zuverlässige Ergebnisse von gewebespezifischen trans-epistatisch fehlregulierten trans-Zielgenen darstellt. Viele dieser trans-Zielgene sind stark mit kardiovaskulären Ereignissen assoziiert. In diesem Zusammenhang unterstützen unsere Ergebnisse die Hypothese, dass kombinatorische differentielle Genregulation die epistatischen Folgen erklären könnte und liefern erstmals spezifisch interagierende genetische Loci und Genotypkombinationen sowie nachgeschaltete trans-Zielgene als spezifische experimentelle Angriffspunkte für eine systematische experimentelle Untersuchung liefern könnten die nachgelagerten Wirkungen genetischer Wechselwirkungen, die KHK zugrunde liegen.

Unsere Studien konzentrierten sich auf die transkriptionelle Regulation und im Speziellen auf transregulatorische molekularen Mechanismen. Wir haben gezeigt, wie maschinelles Lernen in Verbindung mit

Multi-Omic-Datensätzen wie Genotyp, Phänotyp, Genexpression oder Proteinbindungsdaten transregulatorische Faktoren, trans-interagierende genetische Varianten und trans-epistatisch regulierte trans-Zielgene systematisch identifizieren kann. Einerseits ermöglichten uns diese, unser Verständnis der transkriptionellen Regulation proteinkodierender Gene zu verbessern, indem wir die zugrunde liegenden transregulatorischen Mechanismen entschlüsselten und transregulatorische Faktoren aufdeckten, die an den kritischen frühen Schritten der Transkription beteiligt sind. Andererseits konnten wir die Bedeutung transgenetischer Interaktionen bei Krankheiten veranschaulichen, indem wir epistatisch fehlregulierte Krankheitszielgene in trans aufdeckten. Im Großen und Ganzen bieten unsere Projekte eine systematische Untersuchung transregulatorischer molekularer Mechanismen, die spezifische experimentelle Endpunkte für eine gezielte Untersuchung der beteiligten Faktoren und der regulatorischen Prozesse im Allgemeinen liefern.

Contents

1. Introduction

1.1. Thesis Aims

This thesis covers two biological research projects concerned with **1)** the identification of novel trans-regulatory factors modulating the promoter-proximal pausing of the Polymerase II during the transcription of mammalian protein-coding genes and **2)** the identification of trans genetic interactions as upstream trans-regulators of downstream trans target genes that underlie Coronary Artery Disease. In large, both projects aimed to improve our understanding of gene regulatory molecular mechanisms, specifically from the perspective of trans-regulatory factors. Trans-regulatory factors interact with cis-regulatory elements to control for gene expression. However, in contrast to cis-regulatory elements, trans-regulatory factors perform functions that are not constrained by their localization. They may act on any distal genomic element and function as master regulators, as shown for many transcription factors that confer trans-regulatory effects through binding to cis-regulatory transcription factor binding sites (1, 2). This highlights the importance of trans-regulatory mechanisms and implicated factors for a holistic understanding of gene regulation. Moreover, trans-regulation is characterized by interactions of different types of molecules such as proteins, DNA and RNA. This multifactorial complex nature of trans-regulatory mechanisms requires multifarious large-scale multi-omic data sets to accurately capture the genomic context of trans-regulation. Machine learning-driven analysis of such multi-omic data sets enabled us to draw conclusions about the initial biological research questions and enlarge our understanding of trans-regulatory molecular mechanisms underlying gene regulation from two different perspectives. In the following two paragraphs, we provide a brief overview of these two aspects of trans-regulation from the perspective of each project.

In our first project, we aimed to extend our knowledge about the role of trans-regulatory proteins in the context of transcriptional pausing as a key determinant of gene transcription. The transcription of genes is driven by internal and external stimuli that modulate the cell's behavior to continuously adapt itself to these changing environmental conditions in order to sustain cell homeostasis for proper cell functioning (1, 2). This adaptive process is tightly regulated by the coordinated interplay of chromatin and transcription factors (TFs) (3). Initially, a pre-initiation complex (PIC) of transcription factors assembles to enable the synthesis of a short nascent RNA fragment by the polymerase. The polymerase then pauses and requires other regulatory signals to either enter productive elongation or terminate transcription prematurely (6). This is called promoter-proximal pausing of the polymerase (3). The transcriptional pause rates and durations affect RNA burst production (transcription in short bursts) and transcriptional as well as translational noise (for reasons yet unknown), collectively modulating mean RNA and protein levels, representing a key determinant to transcriptional output in large (4). Thus, i(3)represents a critical early regulatory step in the maturation of full-length transcripts as it limits the transcriptional throughput per unit of time (7, 8). Moreover, it is observed across the whole spectrum of gene expression levels, ranging from highly active to largely silenced genes (23) as well as across different life forms ranging from bacteria (244) to mammals (245, 246). It is therefore considered a hallmark of genes, highlighting the ubiquitous importance of the transcriptional pause mechanism. Hence, for a holistic understanding of protein biogenesis and ultimately cell functioning understanding of this regulatory layer poses an important challenge. As a result of higher transcriptional initiation rates compared to productive elongation or premature termination rates (9, 10), paused RNA polymerases accumulate at the promoter site. This accumulation can be seen in assays that capture nascent RNA fragments, such as global run-on sequencing (GRO-seq) (11). Based on GRO-seq data, transcription initiation and productive elongation events can be related to each other with the pausing index (PI), also known as the traveling ratio (TR) (12–14). It is defined as the ratio of GRO-seq reads in a window around the promoter compared to the rest of the gene body, and as such, quantifies the equilibrium between transcription initiation and productive elongation. This, in turn, enables us to contrast these two states of the polymerase and investigate the associated genomic contexts to elucidate the underlying process, in particular, identify novel regulators of transcriptional pausing. Transcriptional pause regulatory factors entail cis- and trans-acting factors that either promote pausing or elongation (16). These have also led to the concepts of 'intrinsic' and 'regulated' pausing. For instance, trans-acting regulatory transcription factors like DSIF, NELF, or P-TEFb and intrinsic cis-acting elements in the DNA/RNA sequences like specific promoter sequence compositions, GC-content, or transcription factor binding motifs and degree of binding motif conservation fine-tune and regulate pausing. Because transcriptional pausing is a convoluted process that emerges in conjunction with other RNA regulatory processes like for instance splicing (75) or transcription termination (77), further layers of complexity arise. However, we still lack quantitative descriptions of associated factors and processes with the potential to reveal their relative importance, identify previously unknown regulators of pausing, and elucidate their roles in other RNA regulatory processes. Hence, we aimed to reduce this gap by building machine learning models based on large-scale genome and transcriptome binding maps, gene annotation, and sequence composition features, systematically identifying previously known and novel cis- and trans-acting regulators of pausing. By integrating data capturing protein-RNA interactions, we further identify novel 7SK pause mediator binding proteins and show their predictive values. We further strengthen the interconnection of the transcriptional pause process with other RNA regulatory events by integrating prior knowledge of implicated factors and quantifying their relative importance. The related project is covered in Chapter 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures.

In our second project, we aimed to enlarge our knowledge about the genetic basis of Coronary Artery Disease (CAD) (5), especially trans-acting genetic interactions. CAD is a cardiovascular, inflammatory disease that arises through occlusions of the coronary arteries with plaque. Research has revealed general environmental risk factors like stress, nutrition, and smoking. Epidemiological investigations (6–9) have led to the discovery of many other risk factors, including diabetes, hypertension, and hyperlipidemia. Similarly, genetic risk factors have been identified through case-control (10–14), quantitative trait loci (15–17) as well as genome-wide association studies (17–28). It has been established to be a complex trait (29, 30) with many disease-associated genetic variants with small effect sizes spread across the genome that can be linked to many genes which do not necessarily have a clear connection to the underlying disease phenotype. Despite a plethora of genetic association studies the genetic basis of CAD still needs to be accurately mapped, as previously identified additive effects and associated variants can not explain all of the heritability of the CAD phenotype ('missing heritability' problem (31, 32)) which is estimated to be around 40-60% (33). Even polygenic risk scores can only explain up to 4% of the variance in the heritability in an independent test population (34). This suggests that the remaining proportion of the heritability could be explained by interactions between genetic or genetic and environmental factors. Genetic interactions also called epistasis (154), occur when a genetic variation's effect depends on the presence or absence of another genetic variation. Such genetic interactions have been successfully identified through systematic screens in yeast, nematodes, and flies affecting fitness and quantitative traits (155), even in humans (35–37). Genetic interactions underlying CAD have also been identified (29–31) but remain scarce and lack replication. The reasons lie in the large sample sizes required for accurate parameter estimates of parametric statistical methods like in logistic regression, the statistical limits that arise when controlling for false positives, the computational limits when considering high order-genetic interactions, and the lack of additional data for proper replication in independent cohorts. Yet, for a holistic understanding of the genetic basis of CAD and complex traits at large, the identification of epistatic interactions is indispensable. Thus we aimed to address this gap with a filter-based permutation testing approach coupled with linear modeling to identify novel trans genetic interactions and reveal their downstream effects on the expression levels of trans target genes. The related project is covered in Chapter 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans.

1.2. Thesis Overview

In the remainder of this introductory chapter we provide a general biological introduction to DNA as a carrier of biological information (Section 1.3.1. DNA - The Blueprint for Life) and focus on the transcriptional cycle of genes (Sections 1.3.2. Promoter Access - 1.3.6. Model of Gene Transcription). We continue with the central mechanism of transcriptional pausing as a rate-limiting regulatory step of the transcription of genes (Sections 1.3.7. Transcriptional Pausing - 1.3.8. Transcriptional Pause Regulatory Elements), laying the foundation for our first project covered in Chapter 3. Chromatin Signatures and their Role in Transcriptional Elongation Control.

We proceed with genetics and genetic variation as a source of phenotypic variation underlying complex diseases (Section 1.4.1. Genetic Variation), introduce studies that systematically identify disease-associated variants (Section 1.4.2. Genome-Wide Association Studies (GWAS)) and follow up with a discussion on how a similar methodology can be applied to any intermediate molecular phenotype to reveal their molecular consequences (Section 1.4.3. Quantitative Trait Loci (eQTL) Studies). Polygenic risk scores as additive proxies quantifying disease risk based on the total number of risk alleles that an individual carries are then introduced (Section 1.4.4. Polygenic Risk Scores (PRSs)), motivating the discussion on epistatic interactions as additional drivers of complex diseases and existing methods for their identification (Section 1.4.5. Genetic Interactions (Epistasis)). Finally, an overview of Coronary Artery Disease as a complex trait (Section 1.4.6. Coronary Artery Disease (CAD)) provides the disease background in which we seek to identify epistatic interaction covered in Chapter 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans.

We conclude Chapter 1. Introduction in section 1.5. Machine Learning & Statistical Inference informing about the global industrial changes resulting from the digitalization and employment of artificial intelligence systems. We then touch upon the general design principles of building such machine learning models, mention the most commonly utilized systems and point to their importance in research and large parts of our digitally driven world.

In Chapter 2. Materials & Methods section 2.1. Omics, we provide brief descriptions of biochemical assays designed for probing biological entities of various types, producing massive amounts of biological data. We then introduce the computational approaches that are employed to analyze such data (Section 2.2. Statistical Inference), starting from basic measures to describe data (Section 2.2.1. Estimates of Location, Variability, and Association) to quantify the uncertainty in obtained measurements (Section 2.2.2. Variance of Estimates) to formulating hypothesis tests (Section 2.2.3. Hypothesis Tests), followed by the most established hypothesis testing methods for group comparisons of discrete and continuous data (Sections 2.2.4. The Fisher's Exact Test & The Chi-Square Test - 2.2.5. The T-test), providing the basis for permutation testing as a means to perform an arbitrary statistical test for multiple group comparisons (Section 2.2.6. Permutation Tests). Lastly, we discuss properly evaluating a series of statistical tests by introducing the multiple hypothesis testing burden and presenting ways to overcome it (Section 2.2.7. The Multiple Testing Burden).

In section Section 2.3. Machine Learning we focus on the technical aspects of supervised machine learning models, starting from linear and logistic regression (Section 2.3.1. Linear Models) followed by tree-based models (Section 2.3.2. Tree Models) comprising simple decision trees (Section 2.3.2.1. The Decision Tree Model), random forests (Section 2.3.2.2. The Random Forest Model) and finally to the most powerful descendant, the Extreme Gradient Boosting Tree model (Section 2.3.2.3. The Extreme Gradient Boosting Regression Tree Model (XGB)). An overview of common procedures and concepts to build and analyze machine learning models (Sections 2.3.3. Feature Scoring - 2.3.7. Regularization) then concludes the machine learning methods section 2.3. Supervised Machine Learning.

We then follow up with the two projects forming the foundation of this dissertation (Chapters 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures & 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans). In our first project (Chapter 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures; 'Transcriptional Pausing'), we investigate the role of DNA and RNA-associated proteins as well as gene annotation and sequence composition features in modulating polymerase II promoter-proximal pausing as a key determinant of transcriptional output of protein-coding genes. We first detail the methods (Section 3.1. Materials & Methods) used to acquire the transcriptional context of gene transcription (Sections 3.1.1. Integration of Transcript Annotations (GENCODE) - 3.1.3. Integration of Transcription Start Site Annotations (CAGE)) and associated factors (Sections 3.1.5. Integration of Genomic Transcription Factor Binding Sites (CHIP-seq) - 3.1.7. Targeting the 7SK non-coding RNA) as well as quantify Polymerase II pause states (Section 3.1.4. Quantification of Promoter-Proximal Pol II Pausing (GRO-seq)) and show how to integrate this information into a machine learning task with gradient boosting tree regressors (Sections 3.1.8. Model Feature Engineering - 3.1.11. Feature Scoring & Interpretation). We then describe the results of our project (Section 3.2. Results), evaluating the model performances (Sections 3.2.1. Predictive Models of Transcriptional Pausing). Finally, we assess the predictive power of functional associations of implicated factors (Sections 3.2.2. Linking Transcriptional Regulatory Steps with Transcriptional Pausing) and provide novel modulators of transcriptional pausing (Section 3.2.3. Modulators of Transcriptional Pausing). In this project, we extend our knowledge of polymerase II transcriptional pausing by providing evidence of its interdependence with other RNA regulatory processes like, for instance, splicing and, more importantly, providing a set of specific highly predictive proteins that act as additional trans-regulatory factors of transcriptional pausing.

In our second project (Chapter 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans; 'CAD Epistasis'), we identify genetic interactions underlying CAD and assess their potential downstream effects on the expression of trans target genes. We first cover the methodological background (Section 4.1. Materials & Methods) starting from the integration of relevant omic data types (Sections 4.1.1. Integration of Genotype and Phenotype Data & Section 4.1.2. Integration of Quantitative Trait Loci for Coronary Artery Disease), to the identification and replication of genetic interactions (Sections 4.1.3. Identification of Candidate Epistatic Interactions & Section 4.1.4. Discovery and Replication of Epistatic Interactions). In a subsequent step, we prepare for differential gene expression analyses (Section 4.1.5. Integration of Gene Expression Data) and provide the methodology to conduct these analyses (Section 4.1.3. Identification of Candidate Epistatic Interactions). We then characterize the identified SNP interactions in more detail (Section 4.2.1. Identification of Trans Epistasis in CAD) and identify SNP interaction-dependent differentially dysregulated trans target genes as potential drivers of CAD (Section 4.2.2. Association of Gene Expression with interacting SNPs). In this project, we improve upon the discovery of epistasis in CAD and provide specific genetic interactions and genotype combinations conferring differential risk for CAD. We also provide specific trans target genes that are differentially regulated in dependence of these interacting SNPs, providing targets for investigating the gene expression effects of genetic variation on the level of specific genotype combinations.

In the last chapter of this thesis (Chapter 5. Summary & Outlook), we review and discuss our results and put our findings into perspective with the current literature and future developments in the field of genomics.

1.3. Gene Transcription & Regulation

In the following subsections, we like to give an overview of DNA as a carrier of biological information (Section 1.3.1. DNA - The Blueprint for Life) and introduce the transcriptional cycle of genes, ranging from promoter recognition to initiation, elongation and transcription termination (Sections 1.3.2. Promoter Access - 1.3.6. Model of Gene Transcription). An outline of the mechanism of transcriptional pausing underlying this transcriptional process (Sections 1.3.7. Transcriptional Pausing - 1.3.8. Transcriptional Pause Regulatory Elements) then lays the foundation for our first project in Chapter 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures

1.3.1. DNA - The Blueprint for Life

DNA (38, 39) (desoxyribonucleic acid), also called the genome, is the building block for life on earth. The DNA is constituted of two polynucleotide chains that coil around each other to form a double-helical three-dimensional structure (double helix) that arises through successive Watson-Crick (38) base pairings of the four nucleotide bases adenine, cytosine, guanine, and thymine, covalently linked to a phosphodiester backbone. These nucleotides are commonly represented in a four-letter code of A (adenine), C (cytosine), T (thymine), and G (guanine). The DNA double-strand forms through complementary base-pairing interactions of these nucleotides, i.e., 'A' pairs with 'T' and 'C' pairs with 'G'. The resulting double helix further winds itself in spools around nucleosomes (40–42) composed of eight protein complexes collectively called histones. This spooling of the DNA is responsible for the efficient packing and condensation of the DNA within the cell. It further fosters the prevention of DNA damage and is a major determinant of DNA replication and the expression of information encoded in the DNA. This three-dimensional structure, called chromatin (43), folds through successively higher-order structures to form so-called chromosomes. Depending on the species, different numbers of chromosomes may be present in their cells, while diploid organisms share the property that they possess two copies of each of their chromosomes. Each healthy human cell contains 23 pairs of chromosomes, 46 in total. These chromosomes harbor regions of DNA stretches of particular interest, called genes (44, 45). Variations of genes, called alleles, are associated with observable and measurable traits called phenotypes (46). For instance, anthropometric characteristics in humans like the eye or hair color or their height, but also disease phenotypes like for instance Coronary Artery Disease (5), Type II Diabetes (47) or cancer (48), just to name a few. Genes can encode proteins and, when expressed (decoded), form the building blocks of diverse biomolecules that sustain cellular productivity. Genes can encode for multiple versions of themselves, that arise from multiple alternative transcription start sites (49).

Strikingly, multicellular organisms carry the same DNA sequence in all of their cells, yet these cells differ in their form and function. Hence, cellular mechanisms that control the use of the DNA, in particular the genes, must exist, i.e., the cell must have a mechanism in place to use only subsets of all available genes for specific cell types to exist (50). So-called transcription factors (TFs) play an integral role in this cell-type-specific selective expression of genes (51–53). These TFs selectively bind to different parts of the regulome to activate or inactivate the expression of a subset of genes depending on internal and external (environmental) factors to change cellular behavior as an adaptation process to changing requirements. However, the presence of these TF binding sites alone is not sufficient to explain cell-type-specific gene expression but rather also depends on the accessibility of the binding sites for a TF to bind at all (54). This accessibility of genomic regions, particularly binding sites of transcription factors, is modulated through the alteration of the histones around which the DNA is wound. Histones have protein ’tails’ that are subject to modification which alters the compaction of the DNA at altered sites (55). Depending on the type of alteration the DNA may be packed densely, called *hetero*chromatin, or loosely, called euchromatin. Euchromatin enables accessibility of gene regulatory genomic DNA to transcription factors that regulate gene expression (56). More importantly, the emergence of different cell types is a consequence of the different accessibility of binding sites for transcription factors generated during the differentiation of cells from common progenitor cells. At specific transition points into specific cell lineages where different cell types are formed from a common progenitor cell, DNA compaction at cell-type specifying sites is altered so that different sets of genes can be expressed that are specific to the cell type to be generated. This differential compaction/decompaction of DNA is itself regulated by specific transcription factors, called chromatin remodelers. They locate inaccessible chromatin and open up these regions to allow the binding of other transcription factors that regulate the expression of cell type-specific genes or vice versa. As mentioned earlier, this differential compaction of DNA and subsequent alteration of the gene expression program not only occurs at the transition points of cell differentiation but also within a specific cell type to adapt to new requirements induced by different stimuli like changes in the temperature, pressure or simply nutrient availability and many more (57).

We can further differentiate between silencing (repressing) and activating transcription factors or regulatory binding sites, referring to their inherent function of either inactivating or activating gene expression (52, 58). Regarding regulatory binding sites, we further distinguish between cis-acting promoter-proximal and gene distal sites. Enhancers and silencers may be in close proximity to each other and may differ only in the transcription factors bindings at those sites. Another regulatory layer represents non-coding RNAs, i.e., RNAs that do not encode for a protein that arises from non-protein-coding regions which make up most of the genome (59, 60). They include for instance well established classes of RNAs such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) with significant roles in the final step of gene expression (translation), small nuclear RNAs involved in alternative processing of protein-coding mRNAs (alternative splicing), and small nucleolar RNAs mainly involved in the modification of small RNAs like ribosomal and transfer RNAs. Many other non-coding RNA biotypes (long non-coding RNAs, piwi-associated RNAs, endogenous short-interfering RNAs, microRNAs, etc.) additionally play key roles as regulators of gene expression and epigenetic control of chromatin, promoter-specific gene regulation, mRNA stability and many more. Their specific roles and function within the cell are yet to be established further, and it remains an active area of research. Still, they undoubtedly represent functional elements with the potential to regulate gene expression in addition to transcription factors.

To conclude and summarize, a simplified model of protein biosynthesis (see Fig. 1.1) through the transcription of a protein-coding gene first involves the alteration of the DNA architecture by chromatin remodeling events (61).

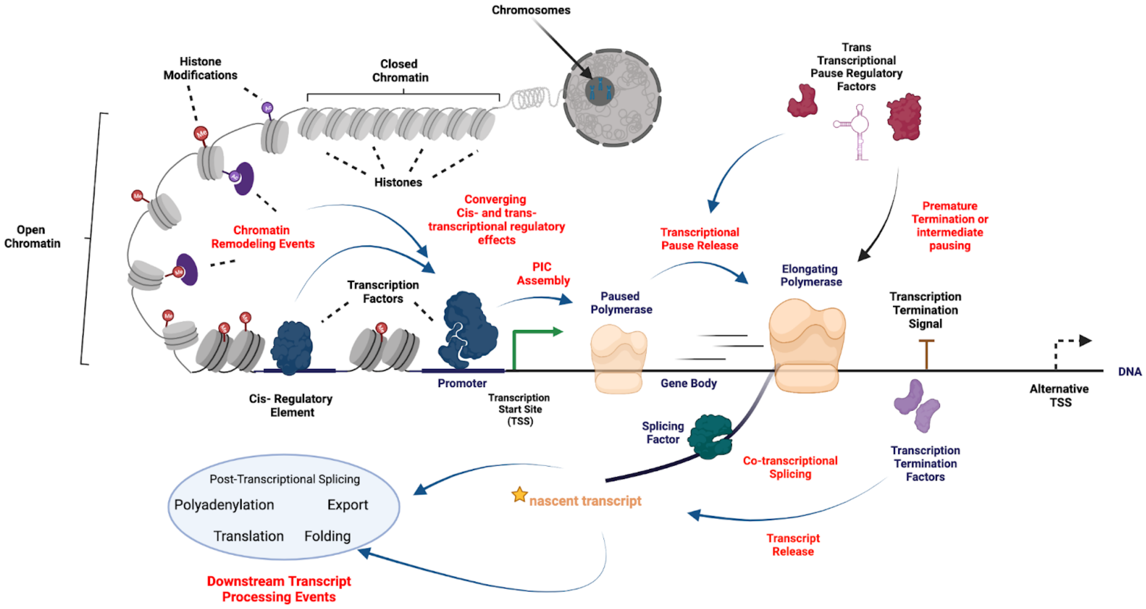

Chromosomes
Histone Modifications
Closed Chromatin
Trans Transcriptional Pause Regulatory Factors
Histones
Open Chromatin
Chromatin Remodeling Events
Converging Cis- and trans-transcriptional regulatory effects
Premature Termination or intermediate pausing
Transcriptional Pause Release
PIC Assembly
Elongating Polymerase
Transcription Factors
Paused Polymerase
Transcription Termination Signal
DNA
Cis- Regulatory Element
Promoter
Transcription Start Site (TSS)
Gene Body
Alternative TSS
Splicing Factor
Co-transcriptional Splicing
Transcription Termination Factors
Post-Transcriptional Splicing
Polyadenylation
Export
Translation
Folding
nascent transcript
Transcript Release
Downstream Transcript Processing Events

Figure 1.1: Conceptual figure of a simplified model of protein biosynthesis. Created with BioRender.com.

The decompaction of densely packed DNA for DNA accessibility is achieved by histone modifications deployed by chromatin remodelers at sites proximal or even distal (e.g. at trans enhancers) to genes. Subsequent binding of transcription factors at regulatory sequences then promotes the expression of the target gene by recruiting additional transcription initiation factors that assemble at the transcription initiation site near the transcription start site of genes (promoters) (62). Once all necessary associations between transcription factors and regulatory regions, as well as protein interactions and protein modifications of implicated factors, have been established, the transcription of the gene's sequence by the so-called Polymerase can start (transcription elongation). Initially, the polymerase pauses at the promoter site (promoter-proximal transcriptional pausing) (3, 63) until the aforementioned associations have been established to start transcription eventually. Transcription starts at the transcription start site (TSS) of the gene, where the polymerase reads the DNA to produce the corresponding nascent (newly synthesized) RNA by successively extending it with nucleotides until it reaches a DNA encoded termination signal upon which it terminates transcription (64, 65). This process is discontinuous like in the case of promoter-proximal pausing in which transcriptional pause events occur in gene body regions which in turn provide opportunities for other factors to associate with the transcribing polymerase complex (elongation complex) for co-transcriptional nascent RNA processing events for example 5' capping or splicing. Once the transcription of the gene has been completed the process is terminated (transcription termination) (66) with the support of specific termination factors and the resulting transcript is then further processed. It then is, for instance, subjected to post-transcriptional splicing events (67) to yield alternative variants of the transcript or polyadenylation (68) to enable the exportation of the transcript to other cellular compartments and to confer transcript stability. Eventually, the transcripts will be translated, i.e. processed to yield proteins (69, 70). Specific translation complexes called ribosomes (71) associate with non-coding tRNAs and rRNAs that are responsible for their processing to yield an amino acid sequence by decoding the RNA sequence according to the amino acid code which maps three successive nucleotides to one of 22 amino acids (translation). The resulting amino acid polymer then collapses into a specific three-dimensional structure (protein folding) (72) with the help of chaperones to yield a functional protein biomolecule that performs its own function within the cell and contributes to cell homogeneity, productivity, and integrity.

We have seen that for the development of a multicellular organism, specific genes have to be expressed in distinct cells to establish different cell types which perform distinct functions in the organism. This requires an elaborate regulatory program to express cell-type-specific genes. This occurs at large during the transcription of genes, therefore understanding the regulation of gene expression requires in-depth knowledge of the mechanisms of gene transcription. In our first project, we sought to extend our knowledge about gene transcription specifically from the perspective of the promoter-proximal pausing of the polymerase II as a key rate-limiting step to gene expression. To continue, we focus on the biological background of the transcriptional cycle of genes and transcriptional pausing specifically to lay the foundation for our first project in Chapter 3. Chromatin Signatures and their Role in Transcriptional Elongation Control.

1.3.2. Promoter Access

The transcription of genes is conducted by RNA polymerases (73). There are three distinct polymerases transcribing three different classes of genes (74). The RNA polymerase I (Pol I) produces ribosomal RNAs, RNA polymerase II (Pol II) produces messenger RNAs and other non-protein-coding RNAs, and RNA polymerase III (Pol III) synthesizes transfer RNAs and the small ribosomal RNAs. These polymerases differ in the mechanisms of transcriptional regulation and associated factors, however, there exists an underlying theme in which the RNA polymerases recognize the promoter region at the beginning of a gene to initiate transcription followed by the opening of the DNA double-strand and polymerase escape from the promoter to start synthesizing RNA. In the following, we will exclusively focus on the Polymerase II, because of its critical role in transcribing protein-coding genes which are the main building workhorses of the cell, but because of existing biochemical protocols that capture its productivity during the transcriptional cycle.

Prior to transcription initiation, the polymerase needs promoter access (75) which at heterochromatic DNA regions harboring genes is inhibited by the 3D chromatin structure that winds and condenses the DNA in those regions. We have seen that DNA condensation is driven by nucleosomes (review Section 1.3.1. DNA - The Blueprint for Life). Therefore specific chromatin remodelers need to remove or shift the nucleosomes at genes to be transcribed so that transcription factors can gain access to the gene’s DNA sequence. Different genes have different promoter characteristics, some of which can impair nucleosome assembly. As an example promoters which contain CpG islands are often found at housekeeping genes and facilitate polymerase access (76, 77). On the other hand, promoters containing TATA-boxes upstream of the transcription start site are often found in genes that are cell-type specific (78). Taken together, chromatin opening is regulated differently for distinct classes of promoters. This promoter opening is performed by transcription factors of which about 1600 are known (51). Most of these factors bind free DNA (79) and only a minority of factors, also known as pioneering factors, can bind nucleosomal DNA and open up chromatin locally to enable transcription (80). These include, for example, histone acetyltransferases or entire complexes of chromatin remodelers (75, 81, 82). Transcription factors also bind distal transcriptional regulatory elements, like enhancers or silencers, to modulate transcription (83, 84). They generally contain multiple binding sites so that entire collections of transcription factors can bind cooperatively, influencing each other (85). Gene distal enhancers influence the transcription of genes by DNA looping structures, i.e., the chromatin architecture, which enable enhancers to communicate with the promoters of genes. These chromatin architectures within which enhancers usually operate are also known as topologically associated domains (TADs) (86). Once promoter access has been cleared and the promoter primed for transcription, the polymerase can initiate transcription.

1.3.3. Transcription Initiation

Because polymerases cannot recognize promoters by themselves, transcription initiation factors are required to recognize and bind conserved DNA sequence elements in promoters, forming a bridge between the polymerase and the promoter. This assembly of transcription factors at the promoter site, and the polymerase is called the pre-initiation complex (PIC). The PIC (87) consists of factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, RNAPII, and Mediator. Pol II and the TFIID make extensive contacts with DNA that extend ~40 bp on either side of the TSS. The TATA-box binding protein TBP binds upstream DNA (from the TSS), recognizes and binds the promoter, and assembles with TFIIB, which recruits the Pol II-TFIIF complex. TFIIB thus acts as a bridge between the promoter and the polymerase and stimulates the transcription of an initial RNA fragment. Though the mechanism of promoter recognition for promoters with conserved DNA sequence elements has been elucidated quite well, we still lack a comprehensive understanding of promoter recognition of promoters without such conserved DNA sequence elements. It is postulated that the initiation factors may recognize the +1 nucleosome or sense physical properties like the bendability of promoter DNA (88). However, further research has to be conducted to establish a comprehensive understanding of promoter recognition. This, in addition, is complicated because PICs themselves can differ between promoters.

The PIC's main function beyond promoter recognition is to open DNA which generally requires a DNA translocase called XBP, a subunit of TFIIH, which binds downstream of Pol II (89). It unwinds DNA in an ATP-dependent manner and propels it into the active center of the polymerase, potentially enabling it to start transcription (51). The polymerase harbors a large subunit, Rpb1, which contains a repetitive amino acid sequence of Y-S-P-T-S-P-S, with 52 repeats in mammals, also known as the polymerase tail or C-terminal domain (90). This tail is subject to post-translational modifications, specifically phosphorylations, which play important roles in controlling and enabling Pol II-mediated transcription (91). In general, dynamic CTD phosphorylation enables stage-appropriate interactions of the polymerase with factors associated with transcription initiation, elongation, termination, and transcript processing. The CTD is largely unphosphorylated during initiation, allowing interactions of the unmodified CTD with the regulatory co-activator complex Mediator, which contacts Pol II and the initiation factors TFIIB and TFIIH. It then modulates the phosphorylation of the CTD of Pol II by the TFIIH kinase subunit CDK7 (91), which enables the polymerase to overcome the tight contacts with promoter DNA and escape from the promoter. Capping of the newly synthesized nascent RNA fragment enables the transition into its elongating phase of nascent RNA synthesis.

1.3.4. Transcription Elongation

Once the RNA has grown to a critical length during the transcription initiation phase, the 5' end of the nascent RNA is capped to protect the transcript from cleavage (92). Subsequently, an elongation complex (EC) forms that processively extends the nascent RNA (93). A nucleotide is added by the closing of the polymerase's active sites and the subsequent catalysis of forming a phosphodiester bond (93). During the nucleotide addition cycle, i.e., the elongation of the nascent RNA, certain DNA sequences can interrupt the cycle and lead to so-called transcriptional pausing (94). Pol II often pauses about 25-50 base pairs downstream of the transcription start site (94), called promoter-proximal pausing. During such pauses, the DNA-RNA hybrid is tilted, preventing nucleotide addition and pause escape (95). These transcriptional pause states can lead to polymerase backtracking, arrest, or even premature transcription termination (94). These paused states are further stabilized by DSIF, which binds around the exit channel of RNA and DNA (96), and NELF, which binds the funnel at the opposite site of Pol II (95). NELF impairs the binding of TFIIS to the funnel (95, 97, 98). TFIIS can rescue paused polymerase by binding to the funnel and aligning the tilted DNA-RNA hybrid with the active site (99, 100) but is inactivated by NELF to suppress pause release. The release of the paused polymerase in these gene bodies requires the kinase CDK9 (101), a subunit of the positive transcription elongation factor b (P-TEFb), which phosphorylates DSIF, NELF, and the CTD of Pol II, triggering the formation of an activated elongation complex.

Transcriptional pausing is a key rate-limiting step to gene expression as it limits the frequency of transcription initiation and the rate of transcription elongation regulating the amount of synthesized RNA per unit of time. Divers' kinds of pausing factors enhance or suppress transcriptional pause states. For instance, the MYC oncogenic transcription factor can promote transcriptional pause release (102), or BRD4 (103) can bind distal enhancers and recruit P-TEFb, promoting pause release. We will look into transcriptional pausing in more detail in a later section (Section 1.3.7. Transcriptional Pausing) and first want to conclude the transcriptional cycle of genes by introducing the mechanism of transcription termination.

1.3.5. Transcription Termination

Termination of transcription (reviewed in (66)) of protein-coding genes involves the dissociation of the nascent RNA from the polymerase and the polymerase from the DNA. It influences the stability and cellular localization and thus determines the functional role of transcribed RNAs. Different pathways target coding mRNA or non-coding RNAs selected by a combination of distinct termination signals on the nascent RNA and the specific phosphorylation patterns of the CTD of Pol II. Transcription termination requires the cleavage and polyadenylation specificity factor CPSF, cleavage stimulatory factor (CstF), and cleavage factor I (CFI) and CFII. It further requires the polyadenylation of the transcript, i.e. the addition of poly-A tail of approximately 200 adenosines to the 3' end, and the terminator sequence (AAUAAA) encoded within the RNA transcript itself (except histone genes). The cleavage of the nascent transcript occurs 18-30 nucleotides downstream of the terminator signal by the CPSF component CPSF73. However, the mechanistic details of transcription termination remain enigmatic and require more in-depth research.

1.3.6. Model of Gene Transcription

To recap, for transcription to happen, different sets of factors are required in each phase of the transcriptional cycle covered in the previous subsections. The initiation phase relies on factors that remodel the chromatin architecture for promotor opening and factors that recognize and bind the promoter and the polymerase. The transitioning into the elongation phase then requires factors that recognize, bind, and phosphorylate the CTD of the polymerase. The elongation phase requires factors that extend and co-transcriptionally process the nascent RNA (e.g., splicing, 3' end processing) and factors that rescue the polymerase from transcriptional pause states. The termination phase relies on factors that process the nascent transcript post-transcriptionally (e.g., poly-adenylation) to invoke transcript release from the polymerase and transcript export from the nucleus for downstream transcript processing events.

A major question arises, how such a large number of factors are coordinated, delivered, and in particular, kept separate between the phases of the transcriptional cycle. Microscopy experiments suggest this process takes place in so-called nuclear hubs or static transcription factories, also known as foci or transcriptional condensates. It has been suggested that these condensates form liquid-liquid phase separation containing the necessary transcription factors to concentrate and localize proteins. A simplified hypothetical model for the organization of Pol II transcription has been suggested in which promoter condensates (initiation phase) contain promoter-associated transcription and initiation factors, co-activators, and unphosphorylated Pol II. In contrast, gene-body condensates (elongation phase) contain phosphorylated Pol II, nascent RNA, elongation factors, RNA processing factors, and elongation-specific co-activators. Each condensate is suggested to support high rates of the underlying phases, i.e., initiation or elongation. Because of the different chemical makeup of each condensate, associated factors could be kept separate from each other. To summarize, the model postulates that transcription factors recruit co-factors and Pol II and promote the formation of a promoter condensate, enabling PIC assembly, transcription initiation, RNA synthesis, and Pol II CTD phosphorylation. The subsequent formation of a gene-body condensate supports elongation and co-transcriptional RNA processing. Once the transcript has been fully synthesized, dephosphorylation of the polymerase CTD leads to the dissociation of Pol II from the gene body condensate to eventually reinitiate transcription by reassociation with the promoter condensate.

After having established an overview of the transcriptional cycle of genes, we want to dive deeper into the intricacies of observed transcriptional pause states of Polymerase II. This will form the basis for the analyses conducted in Chapter 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures.

1.3.7. Transcriptional Pausing

As we have seen in the previous sections, transcription is a discontinuous process marked by transcriptional arrests of the Pol II, called transcriptional pauses. It is considered a universal hallmark of Pol II. It is observed not only at protein-coding genes but also at transcribed enhancers, upstream antisense RNAs, and long non-coding RNAs. It is observed at genes across the whole spectrum of gene expression levels, from highly active genes to those which only show very little activity, while it is infrequently found in genes that are not expressed. It is therefore considered a regulatory layer to tune gene expression levels, rather than acting as an 'on-off' switch, with the potential to reactivate paused genes later. In fact, transcriptional pausing is enriched at genes where small gene expression changes have profound effects, such as on signaling molecules, kinases, receptors, and transcription factors. It is thus suggested that potentiation and plasticity of gene expression by transcriptional pausing is critical during development. Therefore, understanding this regulatory layer is necessary for a holistic understanding of protein biogenesis and non-coding RNA species, and, ultimately, cell functioning.

The development of a plethora of assays, for instance, global-run-on sequencing (GRO-seq) (104) or precision-run-on-sequencing (PRO-seq) (105) monitoring Pol II distribution and dynamics, have greatly improved our understanding of polymerase II pausing. Early investigation revealed high densities of Pol II near many promoters at about 25-50bp downstream of the TSS, referred to as promoter-proximal pausing, indicating that it goes through a rate-limiting step before being released into the gene body for productive transcript elongation. Therefore promoter-proximal pause release is a key determinant of gene expression with an underlying mechanism of great interest.

1.3.8. Transcriptional Pause Regulatory Elements

Because trans-acting factors like DSIF, NELF, or P-TEFb can fine-tune and regulate promoter-proximal pausing, this step has been called 'regulated pausing'. On the other hand, pausing due to cis-acting DNA or RNA sequence features or universal barriers like nucleosomes has been called 'intrinsic pausing'. In the following, we first want to dive more into the details of the cis-acting factor's roles in intrinsic pausing.

Cis-acting factors of transcriptional pausing

The interaction of transcription factors with DNA or RNA is primarily driven by DNA/RNA encoded sequence properties. Many factors bind specific DNA sequences with high affinity. These sequence elements, therefore, modulate the activity of the corresponding factors. Likewise, transcriptional pausing is also a function of sequence elements directing the binding of trans-acting factors which has an effect on the transcriptional activity of the polymerase. For instance, specific binding motifs at promoter sites that influence promoter opening and PIC formation positively correlate with the transcriptional pause states of the Pol II. Because of DNA sequence variation at these sequence elements, the binding affinities of associated factors differ from gene to gene. This has been shown for promoter sites with motifs that strongly agree with the consensus sequences for specific binding sites, which also show a positive correlation with pausing levels. Moreover, strong promoter motifs can retain PIC components even after promoter escape of the polymerase, which enables the PIC to recruit new polymerases maintaining high occupancy of the pause site at the promoter. Besides motif structures, the relative positions of these sequence elements from the TSS also affect pausing.

Regarding specific sequence properties, it has been observed that a high G/C content of the RNA/DNA hybrid during transcript elongation presents an obstacle to the forward movement of the Polymerase. This, in turn, renders the Polymerase susceptible to backtracking, which dislodges the 3′ end of the nascent RNA from the catalytic site which in turn can lead to a transcriptional arrest or premature transcription termination. In this context, it has been shown that CpG island promoters, which have a GC skew, strongly correlate with pausing.

Another observation has been made that the nontemplate DNA strand upstream of paused Pol II has the potential to form structures called G quadruplexes, a stable secondary structure held together by G-G base pairs.

These structures pose an obstacle to reannealing the non-template strand with the template strand after transcription, which has been suggested to potentially enable the formation of R-loops between nascent RNA and the DNA template. Such R-loop formations are observed at promoter sites and impact genome stability.

Although DNA/RNA sequence elements of promoters and, in general, gene bodies or even cis-regulatory sites present opportunities to regulate transcriptional pausing, they cannot explain the longevity of paused Pol II. Therefore sequence properties play important roles in transcriptional pause states, yet trans-acting transcription factors ultimately regulate the productivity of the polymerase.

In contrast to intrinsic pausing, regulated pausing is governed by trans-acting factors like, for example, NELF or PTEFb. In the following, we want to cover such instances further and provide insights into trans-regulatory factors implicated in transcriptional elongation control.

Trans-acting factors of transcriptional pausing

We have seen that NELF occupies the binding site of TFIIS, thereby maintaining transcriptional pause states as it inhibits binding of TFIIS, which would otherwise induce pause-release. In this setting, pause-release requires the kinase activity of P-TEFb. P-TEFb phosphorylates many factors, including DSIF and NELF, which have been shown to be necessary to overcome DSIF/NELF-mediated inhibition of early elongation. In fact, biochemical experiments have shown that the inhibition of P-TEFb activity leads to Pol II pausing at nearly all mRNA promoters. P-TEFb is typically (>75%) found as part of larger complexes like the SEC or 7SK non-coding RNA complex. Components of the SEC interact with Mediator, which recruits P-TEFb for pause release at sites occupied by Meditator. 7SK complex bound P-TEFb is inactive and requires its dissociation from the inhibitory 7SK complex. The 7SK complex includes the most abundantly expressed non-coding RNA 7SK and is bound by MeCPE, LARP7, and HEXIM proteins. Many cellular stresses and signaling pathways can liberate P-TEFb from 7SK, enabling large-scale activation of gene expression. A well-known factor that liberates P-TEFb from 7SK is the bromodomain-containing protein Brd4, whose CTD can bind P-TEFb to stimulate PTEFb kinase activity for targeted gene activation. Other transcriptional pause regulatory factors like SUPT6H, SUPT16H, MYC, TAF1, TBP, or PAF1 exist, with their specific functions and roles in transcriptional pausing yet to be illuminated through different experimental techniques. Briefly, SUPT6H is a transcriptional enhancer (106). SUPT16H is a component of the FACT complex, which is a histone chaperone that both destabilizes and restores nucleosomal structures and facilitates the passage of Pol II during transcription (107, 108). MYC regulates transcriptional pause release (102), TAF1 and TBP associate with each other and enable promoter-proximal pausing of the Pol II (87), and PAF1 acts as an additional regulator of transcriptional pausing (109).

Beyond cis- and trans-acting pause regulatory elements, nascent RNA regulatory events like, for instance, co-transcriptional splicing or polyadenylation during transcription termination have substantial effects on the pausing of the polymerase. In the following, we briefly cover such interconnected events with the transcriptional pausing mechanism.

Transcriptional regulatory processes connected to transcriptional pausing

Beside cis- and trans-acting pause regulatory elements, transcriptional pausing is further modulated by the interconnection with other pre-, co-, or post-transcriptional regulatory processes like chromatin remodeling, splicing, or RNA processing.

The chromatin architecture is defined by the positioning of nucleosomes, the posttranslational modification of their histones, and ultimately, the DNA wrapped around them. It is modulated by chromatin remodelers and tightly linked to transcription initiation, elongation as well as co-transcriptional splicing (81, 110–112). The regulation of Pol II pausing at promoter-proximal nucleosomes by chromatin remodelers like, for instance, Chd1 (113) has also been established.

Co-transcriptional splicing is strongly dependent on the availability of nascent RNA, which in turn is a function of Pol II pausing. Therefore splicing is intricately connected to transcriptional pausing of the polymerase. To this end, active spliceosomes are complexed to the Pol II S5P CTD during elongation and co-transcriptional splicing (114), and transcription kinetics strongly impact splicing decisions. Slow Pol II elongation rates allow more spliceosome assembly time and favor splicing. Moreover, it has been shown that the recruitment of P-TEFb and Pol II elongation is impaired by the inhibition of the spliceosomal U2 snRNP function (115). This indicates that the release of paused Pol II requires the formation of functional spliceosomes. This also suggested the presence of positive feedback from the splicing machinery to the transcription machinery.

Transcriptional pausing is also connected to transcription termination (116, 117). It has been suggested that it results from the simultaneous interaction of the CPSF complex with the polyadenylation signal and the body of the polymerase. It has also been correlated to the formation of RNA-DNA hybrids (R-loops) when the nascent transcript anneals to the template strand leading to transcriptional pause states. However, the mechanistic details remain to be understood.

1.4. Genetic Variation & Complex Disease

In the following subsections, we lay the foundation for genetic variation analyses conducted in the context of Coronary Artery Disease (CAD) as a complex trait, covered later in Chapter 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans. Complex trait outcomes (CT) can be seen as a function of five components, namely additive genetic effects (A), allele dominance effects (D), genetic interaction effects (I), environmental effects (E), and the effects resulting from interactions between genetic and environmental factors (EI):

$$CT = A + D + I + E + EI \tag{1}$$

Examples of additive genetic effects (A) are the consequences of single genetic variations (Single Nucleotide Polymorphisms (118)) or the additive effects of an entire collection of SNPs as quantified by Polygenic Risk Scores (PRS (119)). Allele dominance effects (D) refer to masked effects of an allele by the presence of another allele of the same gene. Genetic interaction effects (I) refer to the observed joint (multiplicative) effects of variants that the additive genetic effects of the individual variants do not account for. Interactions may also arise with the environment (EI), with the potential to inform about disease-relevant environmental exposures. Lastly, the phenotype is also influenced by environmental effects (E) only, for instance, harmful environments with exposure to radioactive radiation. Therefore, the characterization of complex traits requires accurate quantification of these sources of potentially causal factors underlying a complex trait.

In the following, we first introduce genetic variation (Section 1.4.1. Genetic Variation) and present the so-called Genome-wide Association Studies (GWAS) that examine the additive genetic effects (component A) of SNPs on disease across the genome (Section 1.4.2. Genome-Wide Association Studies (GWAS)). We then show how these genetic variants can also be linked to intermediate molecular traits with Quantitative Trait Loci (QTL) studies (Section 1.4.3. Quantitative Trait Loci (QTL) Studies). We proceed with an introduction to Polygenic Risk Scores (PRSs) (Section 1.4.4. Polygenic Risk Scores (PRSs)) as an additional approach to explain complex trait outcomes with the additive genetic effects (component A) of entire collections of genetic variants. PRSs seek to overcome the limitations of GWA studies that try to explain complex diseases with only single independent associations of genetic variations. Beyond the additive effects of genetic variants collectively captured by Polygenic Risk Scores, genetic interactions (Epistasis) represent another source of genetic variability with multiplicative effects (component I) that PRSs do not capture. Many computational approaches have been developed for their identification, discussed in Section 1.4.5. Genetic Interactions (Epistasis). Finally, we introduce Coronary Artery Disease as a complex trait and briefly discuss previous efforts to identify epistatic interactions underlying CAD (Section 1.4.6. Coronary Artery Disease (CAD)).

1.4.1. Genetic Variation

As we have seen earlier (review Section 1.3.1. DNA - The Blueprint for Life), diploid (120) organisms have duplicated chromosomes and thus possess duplicates of each gene. Alternative forms of a gene arise through changes in the DNA sequence of a particular gene, referred to as alleles. We call an organism homozygous with respect to a gene if both of its alleles are identical in their DNA sequence, else heterozygous. Differences in genes and corresponding alleles may result in observable phenotypic differences through the alteration of cellular programs driven by altered protein structures and interactions resulting from the changes in the corresponding genes. Phenotypic differences also include diseases (30, 121–124). These differences occur through changes in the DNA sequence (125, 126), either at a single nucleotide base, called single nucleotide polymorphisms (SNPs) or in longer stretches of deletions or insertions of the sequence. These changes arise from three main sources, namely, genetic mutation, recombination, and gene flow events (127). Mutations may naturally occur through environmental influences, e.g., exposition to radioactive materials (128). Genetic variation is due to genetic recombination events that arise when homologous regions of paired chromosomes recombine (cross-over) during the Prophase I and Metaphase I during meiosis division (129). Gene flow occurs when genetic material is transferred from one population to another through migration events (130). Alterations of the genetic makeup of genes due to these sources of variation may significantly alter an organism's phenotypic traits leading to differential fitness of that organism. Somatic mutations occur in non-reproductive cells (germ line cells) and can cause phenotypic traits (131, 132), however, in contrast to germline mutations they are not hereditary. These DNA alterations are not limited to genes and may occur at gene regulatory sites (133–135) that influence the expression of genes (136–138), collectively called the regulome.

We have seen that the genome is an organism's DNA sequence that encodes the rules that make up the organism (review Section 1.3.1. DNA - The Blueprint for Life). About 99% of the genomes of unrelated individuals are identical, and the remaining 1% difference is, beyond environmental factors, the reason why people in this world look significantly different. The 1% difference is called ‘genetic variation’ and forms the foundation of evolution (125). Some variations may not show noticeable effects (neutral evolution), some may lead to clearly observable differences like skin and hair colors, and others may cause rare Mendelian or common complex diseases (139). Genetic variations are of different types ranging from single nucleotide variations to changes in the number of entire chromosome sets. These differences are relative to the reference genome sequenced in the Human Genome Project (140, 141). They can occur anywhere in the genome and may affect the structure or the expression of genes and ultimately their function if they occur at the sequences encoding genes or factors that regulate genes or simply regions that have gene regulatory functions like, for example, promoters or enhancers. These differences in the expression of genes or their structural properties resulting from genetic variations may have a profound impact on the organism since they disrupt the sequences that encode proteins that perform all cellular functions. Thus understanding genetic variation is necessary to understand phenotypic differences and disease. Although experimental studies can provide reliable insights into their effects and consequences on the organism, the huge number of variants and variant types does not allow extensive and exhaustive experimental investigations. The sequencing of human genomes has revealed, on average, about 4 million genetic variants in any given individual (139). Genetic disease association tests are conducted to narrow down a subset of variants to focus on. These models focus on the most common (90%) type of genetic variation known as single nucleotide polymorphism (SNPs) (118), i.e., the difference in a single nucleotide. In the following section, we will introduce the methodology behind identifying and quantifying the effects of single nucleotide polymorphisms using quantitative trait loci (QTL) studies.

1.4.2. Genome-Wide Association Studies (GWAS)

In the previous section, we have seen that single nucleotide polymorphisms can lead to monogenic Mendelian or polygenic complex diseases. Understanding the link between genetic variation and disease is thus a central goal of genetics. As we have seen (review Section 1.4. Genetic Variation & Complex Disease), complex trait phenotypes are a function of multiple components, ranging from additive genetic effects, allele dominance effects, genetic interaction effects, environmental effects, and the effects resulting from interactions between genetic and environmental factors. Genome-Wide-Association Studies (GWAS) (142) are a prime example of a systematic assessment of the individual population-level additive genetic effects of millions of SNPs in multiple diseases.

A genome-wide association study (GWAS) (142) tests potential relationships between diverse kinds of phenotypes (continuous, discrete, binary) and genotypes (SNPs). These include disease phenotypes like, for example, coronary artery disease, breast cancer, depression, multiple sclerosis, or phenotypes for traits like blood pressure or BMI. Over 5.700 GWA studies have been conducted for more than 3.300 traits. Such studies are conducted through univariate linear modeling (see Section 2.3.1. Linear Models) of the trait of interest in dependence of the genotypes along with potential confounding variables such as population structure, age, or sex. In the case of quantitative phenotypes, linear regression (see Section 2.3.1.1. Linear Regression) is applied. In contrast, with categorical phenotypes, for instance, breast cancer status, logistic regression models (see Section 2.3.1.2. Logistic Regression) are applied, which generalize linear regression models to binary traits. The genotype is treated as the independent (predictor) and the phenotype as the dependent variable (target). Genotypes are typically encoded as $\{0,1,2\}$ where 0 and 2 represent homozygous genotypes for the reference or alternative alleles and 1 represents the heterozygous genotypes. Thus, it quantifies how often an allele of interest at a genetic variant is observed in a sample. This allows encoding of genetic variation and contrasting groups of samples based on their allele carrier status for estimating the genetic additive effects of individual alleles on the phenotype of interest. The resulting test statistics, i.e., odds ratios from the logistic models or beta regression coefficients from the linear regression models along with p-values, are used to quantify the magnitude of the genetic associations with the disease outcome. Currently (Sept. 2021), there are 4321 studies in which millions of genotypes are independently tested for association with the trait of interest.

However, genome-wide significance (GWS) (143) is a specific threshold ($\alpha = 5e10^{-8}$) to determine the statistical significance of reported associations. So only if a *p*-value for the genotype from a logistic or linear regression model is found to be lower than this threshold will the null hypothesis of no association be rejected, and the SNP will be reported as a genome-wide significant GWAS hit. The GWS is based on performing a Bonferroni correction (see Section 2.2.7. The Multiple Testing Burden) for all the independent common SNPs ($\sim 10^{6}$) across the human genome. At an alpha level of $\alpha = 0.05$ a Bonferroni correction would yield this new significance threshold of $\alpha = \frac{0.05}{10^{6}} = 5e10^{-8}$. Controlling for false positives through a more stringent p-value threshold, as is the case with the GWS, is necessary due to the high number of genetic associations tested simultaneously (multiple testing burden). Otherwise, the probability of falsely rejecting the null hypothesis of no genetic trait association increases substantially. Replication analysis with independent cohorts is an established routine to limit the discovery of false positives as well.

GWAS hits that pass genome-wide association can then be investigated downstream. For instance, hits in protein-coding regions of the genome can highlight potential drug targets. Generally, the biological interpretation of these genetic loci and their functional impact and meaning in disease is difficult. Especially the interpretation of the association of multiple SNPs underlying a specific phenotype, in contrast to Mendelian diseases, is an arduous process. In addition to single-locus trait associations, associations of entire blocks of SNPs with many highly correlated SNPs can be linked with traits of interest, which further complicate pinning down specific disease-causing variants (144) due to their high intercorrelation. More strikingly, a significant proportion (approx. 90%) of GWAS hits are found in non-coding parts of the genome and far away from any protein-coding region (145), substantially complicating the interpretation of the specific mechanistic consequences of such disease-associated variants. Therefore, after having identified millions of potentially causal variants underlying a wide range of disease phenotypes, the focus now has shifted to obtaining mechanistic explanations like the molecular and cellular consequences of these (146). These are called Quantitative Trait Locus (QTL) studies, covered in the following section

1.4.3. Quantitative Trait Loci (eQTL) Studies

As mentioned earlier (review Section 1.3.1. DNA - The Blueprint for Life & 1.4.1. Genetic Variation), genetic variation can cause clear, measurable phenotypic changes (called quantitative and qualitative traits) as is the case with, e.g., monogenic Mendelian diseases like sickle-cell anemia or complex polygenic traits like for instance in CAD. Because approximately 90% of the GWAS hits are found in non-coding parts of the genome (145) the focus now has shifted to obtaining mechanistic explanations like the molecular and cellular consequences of these (146). Variants in non-coding parts of the genome, that do not perturb protein-coding sequences, may, for instance, affect gene regulatory elements that, in turn, can propagate the effects to the transcriptional landscape of genes through differential binding (affinities) of gene regulatory factors that bind to the altered regulatory elements. This is crucial in developing therapeutic strategies to combat the disease (121). To this end, Quantitative Trait Loci mapping (QTL mapping) (147) is conducted where, beyond population-level phenotypes of common complex diseases, intermediate molecular phenotypes like differential gene expression, DNA methylation, or protein abundances are investigated in dependence of genetic variation. The loci that are associated with a phenotype are then called QTLs. QTLs are usually identified by linear modeling of phenotypes in dependence of genotype structures and, if available additional covariates like, for example, sex or age. Many QTL analyses have been conducted ranging from molecular readouts such as gene readouts (eQTLs) (148), protein abundance (pQTLs) (149), and DNA methylations (meQTLs) (150) or histone modifications (hQTLs) (151). The standard genotype encoding is in the allele dosage format (factor variable that specifies the number of the allele of interest observed at each gene copy) or dominant factor encoding (two indicator variables that each specify the number of observed dominant alleles). We further distinguish between cis- and trans-acting QTLs, a distinction between loci that are proximal (typically ≤ 1Mbp) to the target molecular readout (cis-QTL) and trans-QTLs which reside far away from the molecular readout (either > 1Mbp or even on a different chromosome). Among these QTL analysis types, the eQTL analyses gained much attention, as genes that are linked to genetic variation represent attractive therapeutic targets that can be easier studied with experimental approaches. Cis-eQTLs are hypothesized to alter the chromatin structure or the transcription factor binding sites of gene regulatory elements to mediate the gene expression locally (152). In contrast, trans-eQTLs are thought to mediate gene expression by altering intermediate regulatory elements or factors (153). Disease-associated genes could be identified by overlapping GWAS hits with eQTLs that might reveal the gene a variant is linked to.

1.4.4. Polygenic Risk Scores (PRSs)

A major drawback of GWA studies is that only single loci are probed for their association with complex diseases, which in turn are multifactorial with interactions of multiple genes and environmental factors as opposed to monogenic mendelian diseases where only the disruption of one gene is responsible for the disease phenotype (154). So-called Polygenic Risk Scores (PRSs) (119) have been developed to quantify an individual's genetic risk conferred by an accumulation of many genetic variants (SNPs) of small effects summarizing it into a single variable to overcome this limitation. The aggregation serves as a proxy for overall disease risk, some genetic variants conferring high risk and being rare and others conferring small risk but being common. Thereby polygenic risk scores enabled the identification of larger fractions of the population, as opposed to rare monogenic mutations, at comparable or greater disease risk (34). A major advantage of PRSs is that they can be calculated for many diseases simultaneously based on data from a single genotyping array as opposed to monogenic diseases where the sequencing of specific genes and subsequent careful interpretation of the functional effects of mutations found is necessary.

Determining a person’s susceptibility to diseases with PRSs allows physicians to monitor and treat affected people much earlier with preventive approaches that could improve outcomes and overcome disease predispositions early on. It will also enable the characterization of diseases more holistically if data is collected over several years when the disease has not fully developed in the person to acquire information about the individual developmental stages of the disease. Besides benefits for scientific and clinical use, PRSs can encourage people to adopt a different, more healthy lifestyle or to change their life goals as time becomes scarce with lethal predisposition where cure or treatment is unavailable.

Therefore PRSs have gained great attention over the last years focusing on a wide range of phenotypes (155). Typically, polygenic risk scores are calculated (156) as the (weighted) sum of such trait-associated risk alleles, giving an individual's estimated genetic predisposition for a given trait. Weights for the individual alleles can be directly obtained from the genetic association studies (157, 158), e.g., the beta-coefficient or odds ratios of the linear models in a GWAS for a trait or weights from custom multivariate models. Alternatively, non-weighted approaches also exist, in which only the number of trait-associated alleles is summed up. Many disease conditions involve environmental, and lifestyle factors, and combining PRSs with other known risk factors can further improve risk prediction. A major advantage of polygenic risk scores is that they inherently factor in the effect of co-occurring SNPs, e.g., the polygenic risk score will reflect the additive effect of multiple SNPs in carriers. If, for example, the co-occurrence of two or more alleles in a carrier increases the disease risk, then it will be reflected in the distribution of the disease's polygenic risk score-dependent prevalence. However, polygenic risk scores are unable to account for non-additive genetic interactions. Such SNP interactions are difficult to find and complicate the establishment of the genetic basis of many complex traits since the effects of many QTLs might be obscured by interactions with other loci. Therefore, although polygenic risk scores are instrumental to quantifying disease risk in dependence of the number of disease risk alleles, they do not allow identifying specific disease risk allele interactions. This minor distinction is important to accurately elucidate the genetic disease basis and pinpoint the causal disease factors. Such SNP interactions are identified in so-called epistasis discovery analyses (159).

1.4.5. Genetic Interactions (Epistasis)

Although polygenic risk scores capture the cumulative effect of individual genetic variants, they do not account for genetic interaction effects. Genetic interactions, called epistasis (160), occur if the effect of a variant that is affecting a complex trait depends on the genotype of another trait affecting variant, i.e. when the effect of a genetic variation depends on the presence or absence of another genetic variation. Systematic screens for genetic interactions in yeast, nematodes, and flies affecting fitness and quantitative traits have revealed the importance of epistasis (161). Pervasive epistasis has also been shown by transferring whole genomic fragments like entire chromosomes or smaller sequence intervals between two inbred strains (162). Due to these observations, it is reasonable to assume the existence of epistasis in humans too. The scientific community largely dismissed this as most genetic variation in complex traits is additive (163). However, these additive effects and associated variants can not explain all of the heritability of phenotypes (31, 32). This has been termed the ‘missing heritability problem’. Thus for a holistic understanding of the genetic basis of complex traits, the identification of epistatic interactions is indispensable. Genetic interactions in humans with large effect sizes in complex traits have been identified, but remain relatively scarce (35–37, 164–168). For instance, a recent study shows that genetic interactions or interactions between genetic and environmental factors modify the effect sizes of causal variants in human complex traits (169). This again highlights the importance of genetic interaction and the limitations of single-locus-trait association analyses. However, detecting epistasis in humans is generally difficult because large sample sizes are required for accurate parameter estimates of parametric statistical methods like in logistic regressions. Frequentist approaches based on the statistical assessments of the significance of genetic interactions require the investigators to balance the false positive and false negative rates, ultimately reducing the power of the methods as stringent significance thresholds are applied. Lastly, computational limits are reached due to the exponential growth of the search space of possible SNP interactions.

There are different forms of epistasis. For instance, the effect of a disease-associated variant might be concealed by another co-occurring variant in healthy samples (positive epistasis). On the other hand, the effect of a variant could also be enhanced by other variants (negative epistasis). Other reasons for differential disease risk of individuals with the same mutational backgrounds are environmental factors and risk factors like, for example, diet or smoking. Beyond positive and negative epistatic interactions, epistasis can be classified depending on the strength of the variant effect, called magnitude epistasis, or the direction of effect, called sign epistasis. A distinction can also be made on whether the interaction is measured in a specific genetic context compared to the average genetic context across a set of different genetic contexts, also called background-relative epistasis and background-averaged epistasis, respectively.

Lastly, we distinguish between interactions that involve two (pairwise epistasis) or more (higher-order epistasis) mutations. The identification of higher-order epistasis is an analytically challenging problem. An exhaustive screening for higher-order epistasis is impractical at a genome-wide scale at present, simply due to the sheer unmanageable amount of potential SNP interactions to consider, i.e., already n^2 possible SNP interactions when only considering pairwise interactions. In general, the problem grows exponentially (n^x) with the interaction order x, i.e., with n = 1 million SNPs, which is approximately the number of independent SNPs (170), we have to test $1e^{12}$ pairwise, $1e^{18}$ 3-way, $1e^{24}$ 4-way, $1e^{30}$ 5-way interactions, and so on. Not only computational limits but also statistical limits are reached by this huge search space. Although epistasis detection for 2-way or 3-way interactions can now be calculated with GPU accelerated computational approaches (171), the effort to balance the false-positive and the false-negative rate with multiple testing correction procedures limit the discovery of interactions simply as a result of applying too stringent significance thresholds. Moreover, even though the number of SNPs may be huge, they might have very low minor allele frequencies (MAFs), additionally leading to sparse data sets, i.e., to be able to observe rare genotype combinations requires a big sample size.

Beyond experimental approaches, (e.g. recessive epistasis-driven coat color variation in mice (159)) major computational efforts have been undertaken to identify epistatic interactions (172), ranging from exhaustive methods like multifactor dimensionality reduction (173), likelihood ratio-based tests (174), receiver operating characteristic curve analysis (175) or non-exhaustive methods like random forests (176, 177) or Bayesian networks (178) and combinatorial optimization approaches like ant colony optimization (179) and computational evolution systems (180) but also simply linear models (181).

As an example, Boolean operation-based testing and screening (BOOST) (174) runs an exhaustive analysis of all potential pairwise SNP interactions by building an additive logistic regression model of the individual SNPs (main effects model) and a full model additionally incorporating an interaction effect (multiplicative term in the model) of the SNPs. The test statistic for the interaction effects is then defined as the difference of the maximum log-likelihoods. Another approach is called genome-wide interaction search (GWIS) (175) to detect pairwise SNP interactions based on building classification models and evaluating their respective ROC curves, i.e., an additive model for each of the two SNPs taken individually and one multiplicative model for the SNP pair. The SNP pair is said to have better prediction power over the individual SNPs if the ROC curve of the SNP pair associated model lies over the two ROC curves of the individual models. To test the significance of the prediction power, authors developed a model-free hypothesis test, called the difference in sensitivity and specificity (DSS), in which they quantify the gain in sensitivity and specificity of a ROC curve over another. The BOOST and GWIS approaches are restricted to detecting pairwise interacting SNP. Methods exist to overcome this limitation, for instance, with an exhaustive search for higher-order epistasis. One such method is Multifactor Dimensionality Reduction (MDR) (173) which takes a different approach in that it is model-free and non-parametric. It is based on comparing the observed case-control counts in certain genotype combinations with contingency tables and ranking genotype combinations based on the observed degree of sample count differences. However, although exhaustive methods allow for identifying higher-order SNP interactions, they can not be scaled to a genome-wide analysis and have to be constrained to smaller SNP sets (several hundred). This can be done through a series of preselections that reduce the entire SNP set to a manageable size for exhaustive higher-order genetic interaction analysis.

An example is to conduct a single SNP analysis to keep only SNPs with significant marginal effects (e.g. in a logistic regression model) and then test the SNP combination effects for the remaining marker subset (182). This approach is biased since SNP interactions with no individual marginal effects are not considered, yet it allows for an exhaustive screening. Another filtering strategy is called Regressional ReliefF (Relief) (183). The Relief algorithm essentially calculates the proximity between individuals based on genome-wide genetic similarity and subsequently weights the genetic variants based on how well individuals proximal to each other are separated based on their genotypes. The Relief method is prevalent, and many other Relief-based derivatives have been implemented, for example, Relieved-F (184), TuRF (185), Evaporating Cooling Relief (186), ReliefMSS (187), and many more.

Another approach is to filter based on data integration techniques, i.e. selecting SNP sets that are relevant to the phenotype of interest (188–190) e.g. GWAS trait-associated SNPs, or to narrow down to a reduced list of SNPs located in genes that encode for proteins involved in relevant interactions that one is interested in, e.g. by querying protein-protein interaction databases like IntAct (191), BioGRID (192), STRING (193) or ChEMBL (194). On the other hand, one could also follow a pathway-driven approach by selecting pathways of interest and mapping SNP to the genes involved in selected pathways. Public databases like KEGG Pathway (195), Reactome(196), or BioCarta (196, 197) can be mined for these pathways. However, such data-integration techniques are biased since they are incomplete, i.e., some pathways or gene interactions are more studied than others, so SNPs in well-studied branches will be given more weight.

Identifying epistatic interactions is an ongoing effort, and as time progresses, we will most likely see further algorithmic and computational advancements. For discussions and reviews of other established methods, we encourage the reader to investigate the following resources (172, 196–200). In the following section, we additionally want to give a brief overview of Coronary Artery Disease (CAD) to introduce the complex trait of interest for identifying epistatic interactions presented in our second project in Chapter 4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans. It is based on the evaluation of LD differences of SNP pairs between CAD cases and controls. This has recently been suggested as a means to rank pairs of loci for epistasis testing (20). However, it was not yet applied to study CAD.

1.4.6. Coronary Artery Disease (CAD)

Coronary Artery Disease (CAD) or Coronary Heart Disease (CHD) (reviewed in (5)) is a complex trait and a leading cause of death in both developed and developing countries. It is a cardiovascular atherosclerotic inflammatory disease that is caused by occlusions of the coronary arteries. The disruption of the endothelial function of the arterial walls as a result of the accumulation of lipoprotein droplets in the intima of the coronary vessels leads to atherosclerosis. These lipoproteins are bound by water-insoluble lipids that foster their circulation in the bloodstream. In high concentrations, low-density lipoproteins (LDL) may permeate the disrupted endothelium and undergo oxidation which in turn attracts leukocytes that, as a result, lead to the formation of foamy cells, visual as the earliest forms of atherosclerotic lesions. These lesions then attract smooth muscle cells (SMCs), which in turn trigger the proliferation and production of a large volume of extracellular matrix with collagen and proteoglycans that lead to the formation of fibrous plaque, i.e., atherosclerotic plaque. This plaque encroaches the lumen of the coronary vessel and is calcified by new small blood vessels. The resulting final plaque with an enriched lipid-core and necrotic material is highly thrombogenic and poses a risk to the host. In addition, the proteoglycans prolong the existence of lipoproteins in the intima by binding them, and modifications of the lipoproteins further propagate inflammatory responses. In response to these inflammatory signals, matrix metalloproteinases are then secreted, modulating various vascular cell functions. Among these are, for instance, proliferation, cell death, new vessel formation, and destruction of the extracellular matrix of arteries or myocardium. The formation of atherosclerotic plaque due to this cellular reaction chain leads to the obstruction of the blood flow and an unmet oxygen supply. This, in turn, leads to the well-known symptoms of CAD such as substernal discomfort, heaviness, or a pressure-like feeling which may radiate to other bodily places.

Beyond environmental factors like geographical locations, ethnicity, gender, and insights from epidemiological investigations that have led to the discovery of many risk factors, including smoking, diabetes, hypertension, and hyperlipidemia, genetic risk factors have been identified that drive the CAD phenotype. Several case-control (10–14), epidemiological (6–9), quantitative trait loci (15–17) as well as genome-wide case-control association studies (17–28), and investigations into epistatic interactions (201–203) through computational approaches (reviewed in (204)) have been conducted. Genome-wide association studies (GWAS) uncovered 321

CAD-associated variants since 2007 (18, 205). In particular, it has been found that the heritability of CAD risk increases with the number of affected relatives and onset at a young age. Mendelian disorders like familial hypercholesterolemia, a single gene disorder caused by mutations in the LDL receptor genes (LDLR), are also associated with the pathogenesis of CAD. Understanding the LDL cholesterol metabolism greatly improved the understanding of the molecular basis of CAD. GWAS studies have revealed specific markers that robustly associate with CAD, some of which are contained in coding sequences of two cyclin-dependent kinases (CDKN2A, CDKN2B) associated with the regulation of the cell cycle and suggested to have a role in transforming growth factor β (TGF-β)-induced growth inhibition which itself is involved in the pathogenesis of CAD. By investigating tissues affected by atherosclerosis, an antisense noncoding RNA in the INK4 locus (ANRIL) of unknown function has been identified, and alterations of expressions of these CAD-associated factors (CDKN2A, CDKN2B, ANRIL) have led to further insights into their roles in the CAD. Yet, a comprehensive understanding of these and other factors in the advancement of CAD is still unknown. Case-control association studies have been carried out to identify differentially expressed genes, which led to the discovery of many differentially expressed genes that have been categorized with positional cloning (206) as disease-causing, susceptibility, or disease-linked genes. Disease-causing genes have high predictive power and are directly responsible for the development of CAD, and can be readily used for genetic testing. Susceptibility genes are associated with increased or decreased risk for CAD and show genetic variation in CAD individuals. Disease-linked genes, as identified by genomic and proteomic approaches, serve as biomarkers and show differential expression patterns linked to CAD and myocardial infarction.

These efforts have led to the successful development of various therapeutic approaches that have greatly improved the health of affected individuals. For example, during a heart attack in percutaneous coronary intervention, a blocked artery is enlarged with a tiny balloon to reduce the damage to the heart, and a small wire mesh tube (stent) is inserted permanently to keep the artery open, greatly decreasing the chance of artery contraction. Another approach is to treat individuals using recombinant fibroblast growth factor 2 (FGF2), which can stimulate the growth and migration of cell types promoting vascular tree branching and augmenting the coronary flow. A common strategy is treating with antiplatelet agents that decrease platelet aggregation and prevent thrombus formation. Commonly used antiplatelet agents include aspirin, sulfinpyrazone, and nonsteroidal anti-inflammatory agents. Several other therapeutic agents, such as β-blockers, nitrates in the form of sublingual nitroglycerin, or calcium antagonists, have also been successfully used.

1.5. Machine Learning & Statistical Inference

Sophisticated Artificial Intelligence (AI) or Machine Learning (ML) systems (207–209) augment human intelligence and perception and automate tasks, transforming entire industries (209). AI and ML refer to the field of study of algorithms with the ability to learn to perform tasks without being explicitly programmed to do so, i.e., they are trained to learn patterns in historical data to make predictions without being given the underlying patterns on how to do so. Data is the essential ingredient to AI/ML systems that contains a picture of an aspect of the real world. AI/ML systems aim to learn this picture that approximates real-world events and ultimately learn a representation of that world. They can be applied in any sector that produces sufficient amounts of high-quality data.

The underlying principle to then create such systems (207) is always the same and consists of **1)** an object of interest that we are trying to predict **2)** historical data that relates to the object of interest **3)** an AI/ML model that learns the relationship between the historical data with the object of interest underlying the data. An object of interest might be an individual's disease status (e.g. binary indicator variable for a disease). Historical data might be their genetic background. A model then maps the genetic background to the cancer status by learning the underlying patterns in the historical data that determine the disease status. For a model to learn such a relationship, the data must contain a systematic pattern that allows for this, i.e., the object of interest (also called the target variable) needs to be correlated with the factors (also called input variables, predictors, or features) that influence it by a systematic pattern (Fig. 1.2 A). Selecting the best combination of features the model should pay attention to and with which it should try to predict the target is called feature engineering and selection (Fig.

1.2 B). Learning the systematic patterns through a mathematical mapping of the pre-selected features to the targets is also called model training (Fig. 1.2 C). Ideally, a model is provided with features strongly correlated with the target. The process of evaluating the model performance in how well it makes predictions is referred to as validation and testing (Fig. 1.2 D). Lastly, investigation of the prediction contributions of individual model features then serves to understand the model behavior and its predictions in terms of its predictors. This is called feature interpretation (Fig.1.2 E).

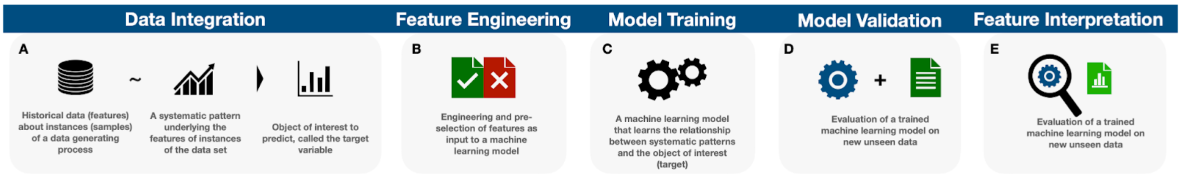

Data Integration
Feature Engineering
Model Training
Model Validation
Feature Interpretation
A
~
Historical data (features) about instances (samples) of a data generating process
A systematic pattern underlying the features of instances of the data set
Object of interest to predict, called the target variable
B
Engineering and pre-selection of features as input to a machine learning model
C
A machine learning model that learns the relationship between systematic patterns and the object of interest (target)
D
+
Evaluation of a trained machine learning model on new unseen data
E
Evaluation of a trained machine learning model on new unseen data

Figure 1.2: Conceptual figure illustrating the general principle of training and evaluating a machine learning algorithm.

Ideally, a model trained on some data should be able to make accurate predictions on completely unseen new data, i.e., it should generalize well, which is a sign that it truly has learned some generally applicable rules by which the data follows and real-world scenarios can be modeled with.

We mainly distinguish between *supervised*, *unsupervised,* and *reinforcement* learning models (207, 210). Supervised machine learning models operate on datasets that include features and known outputs as opposed to unsupervised models, which refer to models that learn patterns in the data where an outcome of interest is unknown. So, in contrast to supervised learning, where a target is predicted from a set of predictor variables, unsupervised methods also build models of the data but do not distinguish between targets and predictors but rather capture general global patterns that underlie the data (e.g., cluster analysis (211)). Supervised learning models can be further subdivided into *classification* and *regression* models in which either a class membership or a real-valued number is predicted, respectively. For instance, with a classification model, we could predict the disease status (e.g., a binary indicator variable for a disease) of individuals (classification) or predict the quantity of a biomarker that is a proxy for a disease (regression). Reinforcement learning takes an entirely different approach than supervised and unsupervised models in that it does not consider labeled or unlabeled data. Not even data itself is necessary but rather an *environment* in which the model can evolve through a series of simulations. Given a reward, a reinforcement model tries to maximize its reward through trial and error. This is very much like raising a child in which feedback from the parents guides the child's behavior for it to behave optimally in its environment.

Many model classes (207) differ in how they approach learning patterns in an unsupervised, supervised, or reinforced way, and it depends on the specific use case to decide what type of model should or could be applied. Factors influencing this decision are whether we are faced with labeled, unlabeled data or an entire environment that determines if supervised, unsupervised, or reinforcement models are needed. The type of data, whether it is categorical or continuous, especially the target in supervised models, determines whether we have to build classification or regression models. The type of relationship between predictors and targets is also important. If the target is likely to be a linear combination of its predictors, then linear models are appropriate. However, if we suspect many non-linearities and interactions among predictors, more complex models like tree models are needed. The number of features and samples (data points) is yet another point to consider. With large data sets (hundreds of thousands), especially samples, neural network architectures are suitable as these can handle large volumes of data and complex patterns. On the other hand, if the number of observations is very low then complex models are not suitable. Another example is prior knowledge about coefficients and their characteristics (value range, non-zero, etc.) which may guide the investigator to Bayesian approaches, for instance, Bayesian Regression Analysis (212). For instance, prior knowledge about the general sparsity of a model or about specific effect sizes of certain independent variables could be accounted for by incorporating this knowledge into the model-building process in terms of additional hyperparameters.

By far, the most widely used are linear models (213, 214) describing a response variable in terms of a linear combination of (multiple) predictor variables, i.e., the weighted sum of the predictors. The choice of the linear model type depends on the response type. In terms of a continuous response variable, linear and logistic regression models (214) are applied in the case of categorical responses. Though the response is modeled by the linear additive combination of the predictors, the predictors can be power terms or other nonlinear transformations of the original predictors. Linear models are fast to implement, intuitive and easy to interpret, applicable to many problems, and computationally inexpensive. On the other hand, they perform poorly in complex processes with non-linear relationships of the predictor variables.

Tree-based (215) models are a very popular way to account for interaction effects between predictor variables and non-linear dependencies of the response. They are based on decision or regression trees, hierarchically ordered if-then-else rules. They consist of many splitting points (called nodes) which evaluate a condition on specific predictors by which individual observations can be classified or grouped into similar groups. A decision tree is rather weak and tends to overfit, which is why it is usually used in ensemble learning, where the predictions of multiple models are aggregated for a final prediction outcome. Random forests (216) are an ensemble extension of decision trees. It consists of many decision trees (sub-models), called a random forest, over which the predictions of individual trees are aggregated (bagging). It is a strong and widely used model, very intuitive, and able to model non-linear relationships. However, if not tuned well, they tend to overfit. Extensions of random forests also exist, incorporating the idea of gradient boosting where a series of models are fit, in which each successive model seeks to minimize the error of the previous model. Extreme gradient boosting trees (217) is a prime example of this approach. These are highly complex models and must be tuned well, as overfitting might occur easily. Even though tree models perform very well in settings with non-linear relationships between predictors, it is sometimes more beneficial to use even more complex models. Especially with massive data with hundreds of thousands or millions of predictors or with certain data types like images or sound, neural networks are more applicable.

Neural networks (218–220) are models that loosely mimic the biological brain's network function. It consists of a series of interconnected neurons (nodes) arranged in layers. Usually, neural networks have at least three layers, namely the input layer where the data is fed into, the output layer where the result or the prediction of the network is propagated, and at least one layer in between the two, i.e., the hidden layer in which the inputs are transformed and processed. A neural network with multiple hidden layers is called a deep neural network, and the training procedure is called deep learning. The network transmits signals (activations) through these layers of neurons from neuron to neuron as a result of a mathematical operation that connects them. The successive activation of each neuron then results in a prediction at the final output layer of the network. Neural networks are extremely versatile and can be used for large data sets with complex non-linear patterns for regression and (multiclass) classification.

Artificial intelligence systems can be used to create value across many industries (207). AI has also transformed the health industry and science in general (221), opening up new avenues for drug development, patient monitoring, automated diagnosis, personalized medicine, and the analysis of DNA. Likewise, we have conducted machine learning-driven analysis of large-scale biological data sets and built Extreme Gradient Boosting Tree Regressors to predict transcriptional pause states of genes (see Section 3.1.10. Model Training), logistic regression models to predict the binary disease status of Coronary Artery Disease patients from large-scale genotypic data (see Section 4.1.3. Identification of Candidate Epistatic Interactions) as well as linear regression models to predict gene expressions from genotype data (see Section 4.1.6. Identification of Epistatic Effects on Gene Expression in Trans). With this section, we conclude the introductory chapter that lays the foundation for gene transcription and transcriptional pausing, genetic variation, interactions, complex diseases, and machine learning systems to be used in those contexts. In the following chapter, we want to continue with the necessary data types and methods underlying our projects.

2. Materials & Methods

This chapter covers the necessary data types and sets utilized in our projects (Section 2.1. Omics) and introduces the computational methodologies of harnessing these data sets to be able to answer our projects' biological questions. First statistical methods are covered (Section 2.2. Statistical Inference), including basic measures to describe data (Section 2.2.1. Estimates of Location, Variability, and Association) and quantify the uncertainty in obtained measurements (Section 2.2.2. Variance of Estimates). An introduction to hypothesis testing follows (Section 2.2.3. Hypothesis Tests), along with specific established methods for group comparisons for discrete and continuous data (Sections 2.2.4. The Fisher's Exact Test & The Chi-Square Test - 2.2.5. The T-test). These lay the foundation for permutation testing as a means to conduct group comparisons with arbitrary test statistics (Section 2.2.6. Permutation Tests). A discussion on properly evaluating a series of statistical tests by introducing the multiple testing burden and how to overcome it (Section 2.2.7. The Multiple Testing Burden) then concludes the statistical section. What follows are machine learning models of supervised nature (Section 2.3. Supervised Machine Learning), focusing on linear and logistic regression models (Section 2.3.1. Linear Models) at first, followed by tree-based models (Section 2.3.2. Tree Models). Lastly, concepts to build and analyze such models are presented (Sections 2.3.3. Feature Scoring - 2.3.7. Regularization).

2.1. Omics

Various biochemical techniques and protocols enable the measurement and quantification of different kinds of biological entities in the cell. These biological data sets form the basis for conducting analyses and studying biological processes of interest and are referred to as omics (222). An entire collection of specific entities of the same type, such as transcripts and proteins, is referred to with the "-ome" suffix. For example, the entire set of transcripts is referred to as the transcriptome, the entire set of metabolites as the metabolome, the set of proteins as the proteome, and so forth. Identifying, quantifying, and characterizing all biological entities involved in cellular processes lie at the heart of omics sciences. High throughput sequencing (HTS) technologies (223) were a key driver of omic sciences as they enabled the sequencing of hundreds of millions of DNA molecules in parallel, providing large data sets with the potential to obtain more comprehensive insights about the cellular realm. For instance, HTS technologies enabled whole-exome sequencing (WES) which can be harnessed to identify novel variants that may underlie cardiovascular disorders, RNA sequencing (RNA-seq) to compare the transcriptome between patient groups, Chromatin immunoprecipitation sequencing (ChIP-seq) to identify protein-DNA interaction or ribosome sequencing (Ribo-seq) to capture actively translated mRNA transcripts. Many more HTS coupled biochemical protocols exist and are further developed, providing more insights into cellular entities and mechanisms. The next section briefly introduces the sequencing methodology as an integral part of omic data sets thereafter.

2.1.1. Next-Generation Sequencing

Next-generation sequencing (NGS) (reviewed in (224, 225)) is a DNA sequencing technology that is used to determine the DNA sequence of a DNA biosample. It is an HTS technology that works massively parallel and has revolutionized genomic research since an entire human genome can be sequenced within a single day with scalable costs. Available NGS platforms (reviewed in (226)), for example, from companies Illumina, Roche, or Thermo Fischer, apply different approaches to achieve high-throughput sequencing with differences in sequence quality, quantity, and choice of application. However, the general approach is to extract genomic DNA from bio samples, fragment DNA for library preparation, ligate adapters to the DNA fragments, and amplify and sequence the fragments, to yield nucleotide base call intensities. After sequencing, a series of computational steps are performed, ranging from the removal of sequencing adapters (e.g., with Trimmomatic (227)), and poor quality reads (e.g., with FastQC (228)) to the alignment to the reference genome (e.g., with BWA (229), Bowtie2 (230) or the STAR aligner (231)). In the following, we briefly describe some of the HTS-driven omic data types used in the analyses throughout the dissertation underlying the data sets.

2.1.2. RNA-Sequencing

RNA-sequencing (RNA-seq) (232) is a transcriptome-wide technique to quantify the cellular content of RNAs. A major application of RNA-seq data is differential gene expression analysis, i.e., to discover quantitative changes in expression levels between experimental groups. Moreover, it allows for identifying alternative transcripts or post-transcriptional modifications that occur during mRNA processing, such as polyadenylation or 5' capping. The RNA-seq protocol begins with RNA extraction and depletion of ribosomal RNA or enrichment of mRNA, followed by copy DNA (cDNA) synthesis through reverse transcription and subsequent high throughput sequencing. Obtained reads are then aligned and quantified for downstream analyses. Gene annotations help us to associate the read fragments with gene regions of interest, quantify the mapped reads and use it as a proxy for the expression levels of the gene they mapped to. Differences in gene length and in total sequencing output per experiment are expected to lead to systematic differences in read counts. A common normalization procedure is to calculate the fragments of reads that map per kilobase of gene sequence per million sequenced reads (FPKM values) (233) in order to make gene expression levels comparable across different genes and experiments.

2.1.3. Cap Analysis Gene Expression (CAGE)

Similar to RNA-seq, Cap Analysis of Gene Expression (CAGE) (234, 235) is a transcriptome profiling technique that, in contrast to RNA-seq, produces snapshots of the 5′ ends of messenger RNAs (sets of short nucleotide sequences called 'tags') and its counts. As opposed to RNA-seq, it identifies transcriptional start sites (TSSs) and the corresponding promoter regions of genes. It enables the investigation of gene regulation on a TSS level as it accurately distinguishes between multiple alternative promoters with the respective alternative transcripts (234). It also allows the investigation of the transcription initiation frequency of specific transcription start sites at single base-pair resolution across the genome (234, 236).CAGE data is also very valuable in validating RNA-seq-based gene expression levels to increase the confidence in expressed transcripts as it also informs about individual alternative transcript expression levels. Briefly, CAGE targets the cap of Pol II transcripts to pull down the 5'-complete cDNAs, which are reversely transcribed from the captured transcripts. Massive parallel sequencing of these 5' ends of cDNAs and analysis of the sequenced tags provides counts of transcription start sites and transcript quantifications on a genome-wide scale, thereby providing an effective genome-wide transcriptional profiling technique as an alternative to microarray or RNA-seq data. To make tag counts comparable among experiments they are normalized to the number of raw sequences that were read (Tags per Million (TPM)) which gives the expected count for a particular tag if we had sequenced one million raw CAGE tags.

2.1.4. Global-Run-On-Sequencing (GRO-seq)

As opposed to RNA-seq, which quantifies the total cellular content of RNAs, *Global-run-on-sequencing* (GRO-seq) (237) quantifies all initiated (nascent) transcripts that are being newly synthesized by the engaged polymerases. This, in particular, is useful for annotating and quantifying short-lived RNA molecules or assessing the polymerase's productivity. This is accomplished by arresting ongoing transcription, e.g., through cold temperatures, introducing brominated nucleotides (BrdU), and preventing *de novo* assembly of the pre-initiation complex with an anionic detergent like sarkosyl and thereby avoiding re-initiation of new polymerase. Transcription is then resumed, and already initiated polymerases will integrate the brominated nucleotides, and the nascent RNA molecules can be affinity purified with antibodies against bromodeoxyuridine (anti-BrdU). Extracted RNA fragments are then subjected to high throughput sequencing and subsequent mapping and quantification with computational pipelines. The output is similar to those from RNA-seq experiments with the exception that it captures all nascent RNA. Quantification of these nascent transcripts is analogous to the approaches conducted on RNA-seq experiment outputs. However, the raw nascent RNA read counts can also be used as they inform about the productivity of the polymerase and about sites of active transcription. An advantage of GRO-seq over RNA-seq is the possibility to contrast the transcriptional events of initiation to elongation by measuring the relative ratio of GRO-seq reads that map near the promoter site to the reads that map into the gene body. In fact, this forms the basis of our machine learning regression task in our first project in which we predict this ratio, also called the Pausing Index or Traveling Ratio, from large-scale genomic and transcriptomics protein binding maps as well as gene sequence composition feature (see Section 3.2.1. Predictive Models of Transcriptional Pausing).

2.1.5. Chromatin Immunoprecipitation-Sequencing (CHIP-seq)

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (238) is a genome-wide in vivo assay to identify and selectively enrich DNA sequences bound by proteins such as transcription factors and in general chromatin-associated factors. During CHIP-seq, DNA is cross-linked to a protein complex of interest using formaldehyde, and the bound DNA sequences are then fragmented, and protein-specific antibodies are used to immunoprecipitate the protein-DNA complex. Subsequent sequencing of the bound DNA fragment allows for identifying and quantifying binding sites of the immunoprecipitated protein. Sequenced fragments are then mapped to the genome in a computational pipeline. Aligned reads are asymmetrically centered around the binding site, with read densities on the positive and negative strand. These represent candidate binding sites, however, because reads from a ChIP-seq experiment are a mixture of enriched signal reads but also a large number of background noise reads from non-specific (background) regions throughout the genome (239) regions with high read densities may not correspond to the enriched signal reads. To account for these and other unknown biases, a matching control sample with non-specific antibodies can be generated against which the original samples can be statistically compared to determine sites of enriched signal reads (240). This procedure is called peak calling (214) and is usually conducted with statistical methods (e.g., MACS2 (241)). ENCODE has a uniformly processing pipeline following strict quality standards (242) that produces versioned data in high quality, in many instances available in multiple formats with extensive documentation and a user-friendly interface.

2.1.6. Enhanced Cross-linking Immunoprecipitation-Sequencing (eCLIP-seq)

In contrast to CHIP-seq, *enhanced Cross-linking immunoprecipitation followed by high-throughput sequencing* (eCLIP-seq) (243) is used for the identification of transcriptome-wide protein-RNA interactions. eCLIP provides another important view on how proteins interact with RNA to regulate gene expression. As opposed to crosslinking with formaldehyde, as with CHIP-seq, protein-RNA interactions are covalently linked using ultraviolet light (UV). After cell lysation, the bound RNA fragments are fragmented with an RNAse, and protein-specific antibodies are then used to immunoprecipitate the protein-RNA complex. Sequencing of the bound RNA fragment then allows for identifying binding sites of the immunoprecipitated protein. The sequenced fragments are then mapped to the genome in a computational pipeline, and specific protein binding sites are identified with peak calling procedures (e.g. PureCLIP (244)) and statistical methods, similar to ChIP-seq experiments (see the previous section). An adapted background control normalization procedure is applied to account for unknown, but also known biases like the dependence of the background on gene expression levels or varying protein sizes that have selective RNA targets (245).

2.1.7. Genotyping

The genomes of different individuals are not identical. They vary on average by 1% (125). The most common type of genetic variation is called a single nucleotide polymorphism (SNPs) (118) or point mutation, a single base pair variation in the genome. In order to be able to associate variants with phenotypes, it is crucial first to localize these variants. Genotyping (246) describes the processes to capture known variants identified through a systematic comparison of the reference genome against the genomes of additional sequenced samples and measure them. A prime example is the HapMap project which first identified all variants with a minor allele frequency (MAF) > 0.05 and genotyped a common SNP every five kilobases along the genome (247). Genotyping methods differ in the number of identifiable SNPs, technical specifications, and the analysis cost (reviewed in (226, 248, 249)). For instance, in hybridization-based SNP arrays, hundreds of thousands to millions of allele-specific short DNA molecules (oligonucleotide probes) are designed for all SNPs of interest, for instance, SNPs a priori known to be associated with diseases or specific traits. These oligos contain the reference and the alternative allele and are hybridized with fluorescently labeled DNA fragments extracted from a specific biosample of interest, e.g., a specific tissue. Once hybridization of the DNA fragments with an oligonucleotide with a specific allele occurs, fluorescent light is emitted, which signifies the presence of the allele in the DNA fragment. This, in turn, enables the determination of the genotype of the sample of interest by measuring the light intensities of hybridization events for each allele, which is called genotype calling. Although all variants could be identified, including deletions and insertions (INDELs), structural variants (SVs), or copy number variants (CNVs), as is the case with whole-genome sequencing (WGS), pre-selection is usually applied to reduce the financial burden. A great reduction in financial costs is achieved in whole-exome sequencing (WES) (250), where the focus lies only on protein-coding regions (exome) of the genome (about 3%). However, although most Mendelian diseases are caused by mutations found within the exome (251), it misses genetic variation in non-coding regions, especially at gene regulatory sites with potential significant genetic implications. Alternatively, the HapMap project has demonstrated that it is sufficient to genotype a common SNP every five kilobases as one in five SNPs has 20 or more perfect proxies, and three in five have five or more (International HapMap Consortium 2005), such that almost all genetic variation can be captured with proxies in high linkage disequilibrium reducing the financial burden substantially and effectively enabling cost-efficient genome-wide association studies. However, with an ongoing reduction in sequencing costs, we can expect this picture to change with an increased rate of sequencing whole genomes opening up new avenues to investigate genetic variation in greater depth and breadth. Still, because usually only a fraction of all possible SNPs are captured through the pre-selection of loci of interest or genomic regions, additional genotype imputation methods are applied to increase the power of genotyping studies, essentially estimating the most probable haplotypes (252) and thereby the genotypes for untyped variants based on prior knowledge of genome structures and recombination events (reviewed in (253–255)).

2.2. Statistical Inference

In the following, we want to provide a few fundamental concepts to obtain descriptions of data (Section 2.2.1. Estimates of Location, Variability, and Association & 2.2.2. Variance of Estimates) and the methodology for hypothesis tests, along with a few established statistical tests for group comparison (Sections 2.2.4. The Fisher's Exact Test & The Chi-Square Test - 2.2.5. The T-test). These lay the necessary background for the statistical procedure for group comparison with arbitrary test statistics (Sections 2.2.6. Permutation Tests) covered thereafter. A discussion on how to perform a large number of statistical tests (Section 2.2.7. The Multiple Testing Burden) then concludes the statistical section.

2.2.1. Estimates of Location, Variability, and Association

At the heart of statistical analyses is the goal of designing experiments around some subject and analyzing the observational data, trying to infer properties of a population or process beyond the existing data. This is achieved by testing hypotheses and deriving estimates from a subsample of the population to derive such properties. This includes basic descriptions of the data in terms of statistical measures that summarize the behavior of the data and, ultimately, the underlying subject matter

Random Variable

In our digitally-driven world data comes in different types and forms like sensor measurements, texts, and images, questionnaires, and many more. In the previous section (review Section 2.1. Omics), we have specifically seen examples of experimental assays that generate massive amounts of biological data. This data has a generating process (a real-world phenomenon) attached to it in which the data points follow a specific probability distribution that, in turn, is governed by the natural laws underlying the phenomenon. Investigating these distributions allows us to derive conclusions about the underlying data-generating process and the biological phenomenon at large. For instance, an experiment might aim to assess the prevalence of CAD patients in a certain subpopulation and thus randomly sample a subset of people from that subpopulation and annotate whether the individual has CAD or not for each sample. A 'random variable' is the conceptual entity that holds the numeric outcomes resulting from random processes (256), in our example, the annotations whether a randomly drawn sample has CAD or not. The associated probability distribution (and density) function(s) assign a specific probability to each observed value of the random variable (256). Established knowledge about specific characteristics of the random variable's underlying probability distributions then enables us to understand the associated phenomena in great detail. For instance, one might draw conclusions about the population-level prevalence of CAD based on the observed prevalence in a representative cohort.

Generally speaking, a random variable is a numerical collection of the outcomes of a statistical experiment or the data points obtained by some measuring instrument. A random variable may hold numeric or categorical results depending on the experiment. In the case of numerical data points, we further distinguish between discrete and continuous data, which refer to data of only integer values such as count data (e.g., number of CAD risk-alleles carried) and to values in a continuous numeric scale (e.g., gene expression levels), respectively. On the other hand, categorical data takes on only a specific set of values and is categorized into nominal and ordinal data. Nominal data is descriptive and non-numeric (e.g., a list of gene biotypes). Ordinal data is like nominal data but with an intrinsic order to the elements (e.g., a list of gene biotypes ordered by degree of sequence conservation across species).

For data analysis and predictive modeling, these data types help determine the methods to visually display the data, analyze the data, determine the statistical or predictive models that could be applied, and, especially, characterize the underlying data generating process by investigating the probability distributions underlying these processes. To conclude, random variables are the essential logical entities that enable the investigation of data at all. In the following, we want to cover the essential descriptive estimates of such random variables briefly.

Central Tendency

Let X be a random variable with continuous or count data with a finite number of distinct values, e.g. gene expression levels of a biomarker for a disease, then a basic approach to get an overview of the variable is to get a typical value that the variable represents, i.e., an estimate of where most of the entries in X are located, the central tendency or location. The most basic estimate of this location is the *mean* or the average value of the random variable X. It is defined as

$$Mean\ (X)\ =\ \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2}$$

where n is the total number of data points in X. So the mean of X is essentially the sum of all values divided by the number of values in X. In our previous example, the mean would represent the average expression of the biomarker that informs about the disease. This biomarker's mean expression levels can then be systematically compared against the same marker's expression levels in groups of samples with different therapeutic treatments to evaluate the average change in biomarker expression levels in dependence of different therapeutic treatments that enable the assessment of the treatment effects of the therapeutics administered to better combat the disease.

The mean is sensitive to extreme values (outliers) as the mean is shifted towards these values. A more robust estimate of the central tendency of a random variable is to calculate its *median*. It is the middle number of the sorted values of the variable X. In the case of an even number of values, the middle value is the average of the two values that divide the sorted data into its upper and lower halves. The median is formally defined as

$$Median\ (X)\ =\ X[\tfrac{n}{2}], \qquad \text{if n is even} \tag{3.1}$$

$$Median\ (X)\ =\ \frac{X[\frac{n-1}{2}] + X[\frac{n+1}{2}]}{2}, \text{ if n is odd,} \tag{3.2}$$

where n is the total number of values in X.

Variance

Beside the central tendency of values in X one is often interested in whether the values are tightly clustered or spread out, known as the *variability* of X. Estimates of variation are based on the differences in the observed data from the estimate of their central value. The most widely used estimate for the variability of a random variable X is known as the variance, given by

$$Var(X)\ =\ s^2\ = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n-1} \tag{4}$$

Because the variance is not on the same scale as the original data due to the square term, the standard deviation overcomes this limitation and is used more often. It is the square root of the variance:

$$SD(X)\ =\ s\ \ = \sqrt{Variance} \tag{5}$$

To continue our example, in addition to comparing mean expression levels, the variance of expression levels of genes of interest between groups of samples with different backgrounds (e.g., diseased and healthy; groups of samples treated with different medications) could be compared. Jointly considering the mean and the variance of gene expression levels allows us to better assess differences in group means, as these might just arise out of natural variance that underlies the measurements. The T-test is an example of a statistical test that performs a comparison of group means taking into account the variability of the means (see Section 2.2.5. The T-test).

Another approach to investigating the spread of data is based on the *percentiles* of the data. In contrast to the variance and standard deviation, which give a point estimate of the dispersion, percentiles report the fraction of values that lie within a specific range, i.e., the p-th percentile is a value such where at least p percent of the values in X take on this value or less and at least $100 - p$ percent of the values take on this value or more. For instance, to retrieve the 50-th percentile, the data is first sorted, and starting from the smallest value, we proceed 50% of the way to the largest value and report that value as the 50-th percentile below which 50% of data points lie. The 50th percentile is also the median, as discussed earlier.

Correlation & Covariance

Given two random variables X and Y, one is often interested in the relationship between the two variables. If large values of X correspond to large values of Y, or vice versa, or if even, large values of X correspond to low values of Y and vice versa, then we say that the variables X and Y are linearly correlated. Covariance and correlation are two concepts that both assess this relationship between variables (256). The covariance is a quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean, i.e. the joint variability of two random variables:

$$Cov\,(X, Y) \; = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}\,, \tag{6}$$

The overall magnitude of the covariance can be quantified with the correlation between two random variables. The widely used correlation coefficient (rho) developed by Pearson (257) can be employed. It is the product of the deviations from the mean of each variable divided by the product of the standard deviations of each variable. It is a generalization of the variance to two random variables and is linked to the covariance:

$$\rho\,(X, Y) \; = \frac{Cov(X, Y)}{s_x s_y}\,, \tag{7}$$

where s_x and s_y are the standard deviations of variables X and Y, respectively. The correlation coefficient always lies within the range [-1;1] with negative coefficients describing inverse relationships. There are other correlation coefficients, for instance, Spearman's p (258) or Kendall's tau (259), and are based on the ranks of the data values. In contrast to Pearson's correlation coefficient, which quantifies the linear relationship of the variables with constant value changes, rank-based correlation coefficients quantify monotonic relationships where the values increase or decrease at a non-constant rate.

Correlation is an essential concept to identify variables that co-vary. For instance, entire gene-correlation networks can be built to identify and link genes that show similar gene expression profiles to built gene-co-expression networks (260–262). Based on this, entire modules of associated genes with shared or similar functions could be inferred, or networks of diseased and healthy samples could be compared to identify disease network modules with associated genes (263, 264).

2.2.2. Variance of Estimates

At the heart of statistical inference lies the goal of understanding some property of a whole population. However, usually, a statistic (estimate), such as the mean, is calculated on a sample (subset) of the population (superset) rather than the whole population itself because of reasons ranging from financial limitations to infeasibility concerns or time constraints. Then conclusions are made about the whole population regarding the property that was estimated based on the sample. Because the sample cannot fully represent the whole population, *uncertainty* exists in the estimate and the conclusions, i.e., it might be in error and could be different if we draw a different sample from the population. Therefore we are always interested in how much different it might be from the true population parameter, also called the *sampling variability*. The *standard error (SE)* can be used as a metric to quantify this variability in the sampling distribution for a statistic, for instance, the mean:

$$SE \; = \; \frac{s}{\sqrt{n}}\,, \tag{8}$$

where s is the standard deviation of the sample values. As the sample size increases the estimate of the error decreases, as we become more and more accurate since we consider a larger portion of the population. However, as mentioned earlier, taking resamples from the population is not always possible, which is why bootstrap resampling (265) is introduced to quantify the variability of *any* test statistic. One simply draws additional samples with replacement from the initial sample (bootstrap resampling) instead of the larger population and recalculates the statistic for each resulting subsample. Conceptually, we are replicating the original sample many (thousands or millions) of times which embodies the underlying patterns of the original sample. Given an initial sample of a population, the procedure is as follows:

1. Draw n samples with replacement from the initial sample (called a resample)
2. Calculate any test statistic for that resample and record it
3. Repeat steps 1-2 many times
4. Calculate the standard deviation of recorded test statistics and, based on that, the standard error

The number of bootstrap iterations in step 3 is set arbitrarily and usually lies in the range of thousands or millions. The more iterations we make the more accurate will be the estimate of the variability of our statistics.

Instead of giving a point estimate of the variability of the test statistics, as is done with the standard error, we can give a range in which the test statistics will fall, called the *confidence interval*. So instead of performing the calculations in step 4 we trim $x\%$ of the lower and upper part of the distribution of the test statistics and report the trim points between which the test statistics should lie $100 - x\%$ of the time if a similar bootstrapping procedure is followed. For instance, the 90% confidence interval for the mean gene expression levels of m resamples from a bootstrap resampling procedure can be obtained by trimming $[(100 - 90)/2]\% = 5\%$ from either end of the test statistic distribution. This could be done by calculating the percentiles of the test statistics distribution and reporting the 5th- and 95th-percentile points as the confidence interval boundaries and removing test statistics that lie below and above those boundaries, respectively.

Bootstrap resampling is also applied in permutation testing covered later in Section 2.2.6. Permutation Tests as well as in tree-based machine learning models in Section 2.3.2. Tree Models, and are generally used to obtain confidence intervals for diverse kinds of statistical tests, for example, the tests covered in Sections 2.2.4. The Fisher's Exact Test & The Chi-Square Test - 2.2.5. The T-test.

2.2.3. Hypothesis Tests

A major goal in statistics is to design experiments around some subject to confirm or reject a hypothesis about the subject. This is done with so-called *hypothesis tests*. To provide an intuition for hypothesis testing, we will consider a commonly observed scenario when analyzing data and comparing *groups of data.* This differs from investigating single variables with descriptive metrics as shown in Section 2.2.1. Estimates of Location, Variability, and Association. Most often, the goal is to compare two (or multiple groups) with different characteristics against each other. In Section 2.2.1. Estimates of Location, Variability and Association we have seen how groups of samples could be compared based on the mean of some property of the samples, e.g. comparing the expression levels of a biomarker for a disease between healthy and diseased samples. However, how can we know that an observed difference in the expression levels of that biomarker between the groups did not arise out of pure chance? What if the biomarker has no effect, and we observe this difference simply resulting from our sampling procedure? What if we had an infinite number of samples? Would the trend still be visible? To answer these questions, we need a negative control group and formally assess the hypothesis that there is no difference between groups of samples with different backgrounds.

To begin, we will construct a simplified two-group example of this instance to derive a general algorithmic procedure to formally perform multiple group comparisons to assess the likelihood of observed group differences with respect to some property thereof. Let an A/B test describe an experiment where two groups (A and B) are compared to each other in which one of the groups is exposed to some specific treatment (treatment group) and the other not (control group). The goal is to establish that the treatment truly changes some properties of the exposed group compared to the non-exposed group. For instance, we are interested in the effects of the treatment of Coronary Artery Disease patients with acetylsalicylic acid. We want to test whether medication with acetylsalicylic acid lowers the rate of heart attacks. For that purpose, we could form two groups

(A and B) of randomly selected CAD patients, treat group B with the new drug, and leave group A untreated as the control. We then count how many heart attacks occurred for each group. This can be summarized in a $2x2$ contingency table as follows:

	Control (Group A)	**Treatment** (Group B)
No-Heart Attack	a	b
Heart Attack	c	d

Table 2.1: Toy example of a 2x2 contingency table for hypothesis testing.

Variables a, b, c and d represent the table counts of observations for each combination of the groups and heart attack status. We are interested in whether we observe fewer heart attacks under treatment as compared to the control, which is the same as asking whether the ratio of $\frac{a}{b}$ (the odds of no heart attack to heart attack under no treatment) is significantly different from the ratio $\frac{b}{d}$ (the odds of no heart attack to heart attack under treatment). The ratio should be more or less equal if the treatment has no effect. However, if the treatment is effective, it should lead to fewer heart attacks in the treatment group such that the ratios should differ substantially. Any differences between the groups (control and treatment or A and B) are either due to **1)** the effect of the treatment or **2)** simply random chance and the luck of the draw in which subjects are assigned to each of the group's A and B which lead to such a difference. Hypothesis tests help to learn to make this distinction whether random chance might be responsible for an observed effect.

We distinguish between the *null hypothesis*, i.e., the hypothesis that chance is to blame and in reality, there is no difference between the odds in the two groups, and the *alternative hypothesis*, i.e., the counterpoint to the null hypothesis that chance is likely not to blame and in reality there is a difference between the two groups. We further distinguish between the two cases of *one-way-tests* which count chance results in only one direction, and *two-way-tests* which count chance results in two directions. To continue the above example, we assume that the heart attack rate is low in treatment group B and high in the control group A. Therefore, we assume that the distribution of these measurements (heart attack or not) is different between the groups treated differently. As mentioned earlier, this difference is either true and has a biological explanation (here the molecular effects of acetylsalicylic acid) or just arose due to chance because of the assignment of the patients into those groups. A hypothesis test tests exactly this, whether random chance is a reasonable explanation for the observed difference in the measurements between the groups or that the difference is more than what chance might produce. Our hypothesis test could be formulated as follows:

Null Hypothesis (H0)	**:** The odds of a **heart attack** in the treated group is ***equal*** as in the control group
Alternative Hypothesis (H1)	**:** The odds of a **heart attack** in the treated group is ***higher*** than in the control group

So the hypothesis test consists of a *null hypothesis (H0)*, which makes a statement that there is no difference to be observed, and the *alternative hypothesis (H1)* that there is a difference between the groups. In this specific example, we state that the treated group will have fewer heart attack rates, which corresponds to a one-way test. If we had only stated that the heart attack rates would differ or not without giving a direction, it would correspond to a two-way-test because we did not care about the direction, whether it is higher or lower.

In the following two sections, we want to briefly cover three popular statistical tests to assess the probability of the observed data under the null hypothesis, called the Fisher's Exact Test, and an analogous test called the Chi-Square test, both used to investigate contingency tables as illustrated earlier as well the T-Test to compare group means. In general, we will accept the null hypothesis, if the probability of the observed data (or the derived test statistic) under the null hypothesis is large enough. In other words, if the observed data is explained well enough by the model specified under the null hypothesis. It is common practice to use a significance level of 5% to decide whether to accept or reject the null hypothesis.

2.2.4. The Fisher's Exact Test & The Chi-Square Test

For the special case of comparing count data between two groups with respect to some variable, as shown earlier with the 2x2 contingency table of treatment against heart attack rate, statistician Fisher has developed a procedure where all the possible permutations of two groups along with respective frequencies of cell counts are enumerated to determine *exactly* how extreme the original observed result is. It is based on the observation that the table counts are hypergeometrically distributed. Because *all* permutations are considered, it is called an exact test, i.e., Fisher's exact test (for details, see (266)). As compared to the Fishers' Exact test, which is usually applied to 2x2 contingency tables and contingency tables with low cell counts (n<5), the Chi-Square test (for details, see (267)) is a computationally feasible generalization to test the independence among variables in contingency tables with arbitrary numbers of columns and rows, i.e., r x c contingency tables.

These tests additionally serve to conveniently test for enrichments of specific biological entities in groups of data, for example, done in Section 3.2.2. Linking Transcriptional Regulatory Steps with Transcriptional Pausing where we test whether splicing factors are enriched in genomic and transcriptomic intron regions associated with transcriptional pausing as compared to a background set of all observed intronic regions.

2.2.5. The T-test

In the case of continuous data for comparing groups, a computationally feasible approach exists, called the Student's T-test (268). Variations of the t-test exist in which we either compare **1)** the means for two groups **2)** compare the means of the same groups at different time points, and **3)** compare the mean of a group to a known mean prior. These correspond to the null hypothesis (H0) that the differences between the means of the groups (or a group mean and a prior mean in case **3)**) are zero. As discussed earlier in hypothesis tests, we can further distinguish between a one-way and a two-way t-test, i.e., consider the direction in which the means differ between the groups, which is also true for the t-test. The t-test assumes that **1)** the groups are independent **2)** they are approximately normally distributed, and **3)** they have a similar variance. The t-test then calculates the so-called t-statistic, the number of standard deviations from the mean in a t-distribution. The t-distribution is the reference distribution of the null hypothesis to which the observed t-statistics are thus compared. The t-distribution is a type of normal distribution used for smaller sample sizes, where the variance in the data is unknown. It is a good approximation to the permutation distribution of the null hypothesis and thus saves computational resources and time. The t-statistics t is given by

$$t = \frac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}, \tag{9}$$

where $\bar{X}$ and $\bar{Y}$ are the group means, s_x and s_y the standard deviations of each group and n_x, n_y the group sizes of groups X and Y, respectively. Given the t-score we can lookup in a t-distribution table to decide whether we reject the null hypothesis of no significant difference in means $\bar{X}$ and $\bar{Y}$. For that, we identify the degrees of freedom, i.e., the number of variables that are free to vary, set the α threshold commonly $\alpha = 0.05$, and find the critical value in the t-distribution table with these parameters and reject the null hypothesis if our t value is greater than the critical value.

Several analogous tests exist for comparing multiple groups with continuous measurements. For instance, Analysis of Variance (ANOVA) (269) is used to compare more than two groups with respect to one variable, Multiple Analysis of Variances (MANOVA) (270) is used to compare more than two groups with respect to multiple variables, Analysis of Covariances (ANCOVA) (271) is used to compare more than two groups with respect to one variable while controlling for other variables (covariates) and Multiple Analysis of Covariances (MANCOVA) (272) is used to compare more than two groups with respect to multiple variables while controlling for covariates.

2.2.6. Permutation Tests

Permutation testing is an algorithmic procedure to formally perform multiple group comparisons to assess the likelihood of observed group differences concerning some property. In contrast to the asymptotic tests introduced in the previous two sections, permutation tests can be applied to multiple groups, any test-.statistic, and fine-tuned to adjust the accuracy of the estimates. In permutation tests, all possible values of a test statistic under all possible rearrangements of observed data points are calculated to obtain a test statistic distribution under the null hypothesis H0. The permutation distribution of the test statistic can then be used to evaluate a hypothesis test and determine whether an observed effect is likely due to chance.

To formally generalize, in a permutation testing procedure, we essentially shuffle the results from original groups, generate data of the same size as the original groups from the shuffled data, and observe how often we get a difference in the test statistic that is at least as extreme as the observed test statistic from the original non-shuffled data. The shuffling of the groups and subsequent resampling from it essentially represents the null-hypothesis H0 of both groups being equivalent. The general permutation procedure for a set of groups (A, B, C, D ...) to compare is as follows:

1. Aggregate and shuffle the data points from all groups into a single data set
2. Draw bootstrap resamples from the combined data with the same sizes as the original groups
3. Calculate the test statistic of interest for the bootstrap resample groups and record the test statistic
4. Repeat steps 1-3 multiple times to yield a permutation distribution of the test statistic
5. Calculate the proportion of permutation test statistics being more extreme than the observed test statistics of the unshuffled data

Intuitively, if the observed test statistics are not much different from the permuted test statistics, then the observed test statistics are in the range of what chance might produce because we have randomly shuffled the data prior to calculating the permutation test statistics. Calculating how often the permutation statistics are greater than the observed statistics provides us with a probability of randomly observing a result as extreme as the observed because we have randomly shuffled the data. This probability is also called the *p-value*. Given the p-values, we could filter for observations whose results are unlikely to have resulted due to chance (probability equals p) by setting a threshold for the p-value, usually 0.05. So for test statistics with p-values lower than this threshold, also called the *alpha level*, we conclude that chance is unlikely to have produced the statistics and call these results statistically significant and the individual results *significant hits*.

In step 4 the number of iterations is set arbitrarily but usually lies within the range of thousands or millions, depending on the computational resources and time available. The more iterations we conduct, the more precise the distribution of the permuted test statistic will be, and the more accurate will be the p-values. Ideally, we should perform an *exhaustive permutation test* in which we build the distribution of the permuted test statistics for all possible ways of shuffling and dividing the data. However, in step 2 we cannot perform bootstrap resampling as it is resampling with replacement leading to an infinite number of samples that can be drawn. Instead, one could simply resample without replacement. This would result in a so-called *exact test* as it guarantees that the null model will not yield significant test results more than the alpha level of the test.

This procedure lays the methodological foundation for identifying genetic interactions in Section 4.1.3. Identification of Candidate Epistatic Interactions where we apply a permutation testing approach to investigate whether the observed difference of SNP interactions between cases and controls is likely due to chance.

2.2.7. The Multiple Testing Burden

Multiple testing refers to the simultaneous testing of multiple hypotheses (273). The problem with multiple testing is that with an increasing number of tests, the probability that we obtain statistically significant hits increases, although chance is to blame. So we reject the true null hypotheses although the effect is due to chance, yet we declare it otherwise. This is also called the Type I error, mistakenly concluding that a random effect is statistically significant. On the other hand, mistakenly concluding that a true effect is due to chance is called a Type II error.

For instance, suppose we conduct $n = 100$ tests at the alpha level $\alpha = 0.05$ then the probability that we have at least one significant result is given by

$$P(at\ least\ one\ significant) = 1 - P(none\ significant) \tag{10.1}$$

$$= 1 - (1 - \alpha)^{n} \tag{10.2}$$

$$= 1 - (1 - 0.05)^{100} \tag{10.3}$$

$$= 0.994 \tag{10.4}$$

Therefore it is almost guaranteed that we observe at least one significant result although in reality none of the tests are significant. In order to control for the overall Type I error rate, also known as the family-wise error rate (FWER) (265), different p-value adjustment methods have been developed. One conservative way to adjust for the number of tests is by simply dividing the alpha level by the number of tests. So we would reject the null hypothesis if the p-value is less than $\frac{\alpha}{n} = 0.0005$. This is also called the *Bonferroni Correction* method (274). *To* repeat our example with p-value adjustment we get

$$P(at\ least\ one\ significant) = 1 - P(none\ significant) \tag{11.1}$$

$$= 1 - (1 - \frac{\alpha}{n})^{n} \tag{11.2}$$

$$= 1 - (1 - 0.0005)^{100} \tag{11.3}$$

$$= 0.048 \tag{11.4}$$

So the portability of observing at least one significant hit by chance has dropped to only nearly 0.05.

In contrast to the Bonferroni correction method, which controls the probability of obtaining a single false-positive result, the most commonly applied method to control for Type I errors is to control the false discovery rate (FDR) (275, 276). The FDR controls the fraction of false-positive findings over the total number of positive test results. The so-called q-value is an estimate of the false discovery rate from the p-values from the multiple tests and defined as

$$q_i = \frac{p_i N}{i} \tag{12}$$

where p_i is the i-th p-value from the increasingly ordered list of p-values, i the index of the i-th p-value and N the number of statistical tests conducted. Because p_i is the probability of accepting a false result by chance and N the number of tests, their product is the expected number of false tests. The denominator is the number of results we actually accept at the i-th p-value threshold. Therefore the q-value is the number of expected false positives based on the p-values divided by the total number of positives accepted at the same p-values. However, because the q-values are not a monotonic function of the p-values, lower p-values could result in higher q-values. A small adjustment to enforce monotonicity is performed, in which the q-values are simply replaced with the lowest value among all lower-rank q-values.

2.3. Supervised Machine Learning

Supervised machine learning models (review Section 1.5. Machine Learning & Statistical Inference) operate on datasets that include inputs, called *features*, and known outputs, called *targets*, which allow the models to learn patterns and relationships in the data to predict the targets from the features. Supervised learning models can be subdivided into classification and regression models, in which either a class membership or a real-valued number is predicted, respectively. For instance, with a classification model, we could predict patients' disease state (yes or no) or predict the quantity of a biomarker as a proxy for a disease (regression). In the following sections, some of the most widely used classification and regression models are introduced that are also employed in our projects covered later. We cover linear regression and classification models (Section 2.3.1. Linear Models) as well as three variations of tree-based models (Section 2.3.2. Tree Models). An overview of common procedures and concepts to build and analyze such models (Sections 2.3.3. Feature Scoring & 2.3.7. Regularization) concludes the section on supervised machine learning.

2.3.1. Linear Models

2.3.1.1. Linear Regression

A common goal in machine learning, as well as in statistics, is the characterization of the relationship between a random variable X and another continuous random variable Y, and even predict Y given X . *Linear regression (214)* is a supervised machine learning method that assumes a linear relationship between a variable X and Y, essentially fitting a straight line to the data, given by

$$Y = b_o + b_1 X + e \tag{13}$$

The variable X is referred to as the predictor (independent) variable, Y as the predicted target (response; dependent) variable, e as the error term, b_o as the intercept (or constant) and b_1 as the slope for X , both referred to as the coefficients of the model. Unless there is a deterministic relationship, not all the points will fall on the line, so there is an inherent error when approximating the underlying trend, called the prediction (residual) error, i.e. the difference of each point to the regression line, hence the error term e . The fitted values of a model are given by

$$\widehat{Y}_i = \widehat{b}_o + \widehat{b}_1 X_i \tag{14}$$

Throughout this section, variables with a hat-symbol denote predicted or estimated variables as opposed to known variables without the hat-symbol. The residual errors $\widehat{e}_i$ of the model are given by the difference of observed (Y) and predicted ($\widehat{Y}$) values, i.e.

$$\widehat{e}_i = Y_i - \widehat{Y}_i \tag{15}$$

The regression model tries to find the best coefficient $\widehat{b}_0$ and $\widehat{b}_1$ to predict Y_i from X_i which corresponds to fitting a line to the data by minimizing the distance of the line to each point, i.e. minimizing the total error. This is accomplished by minimizing a cost function given by the sum of squared residual errors $\widehat{e}_i$, also known as the residual sum of squares (RSS), defined as

$$RSS = \sum_{i=1}^{n} (Y_i - \widehat{b}_0 - \widehat{b}_1 X_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} \widehat{e}_i^{\,2} \tag{16}$$

The coefficients $\widehat{b}_0$ and $\widehat{b}_i$ are the values that minimize the RSS. Minimizing the RSS is also termed least squares regression, ordinary least squares (OLS) regression, or simple linear regression. In this model, for each unit increase in X the response Y will increase by $\widehat{b_1}$. The ordinary least square estimate is the same as the Maximum Likelihood Estimate (MLE) under a Gaussian model (277).

Usually, multiple predictors are available, and the regression equation is simply extended to accommodate these additional variables where the relationship between each coefficient $\widehat{b_j}$ and variables X_j is linear. This is called multiple linear regression (MLR). The fitted values are thus given by

$$\widehat{Y}_i = \widehat{b}_o + \widehat{b}_1 X_{1,i} + \widehat{b}_2 X_{2,i} + \dots + \widehat{b}_j X_{j,i} \ , \tag{17}$$

where j indexes the j-th coefficient and variable and i the i-th observation. The interpretation of the coefficients is analogous to the simple linear regression, so $\widehat{Y}$ changes by the coefficients $\widehat{b_j}$ for each unit change in X_j assuming all other variables X_k for $k \neq j$ remain the same. As opposed to the simple linear regression where we minimize the RSS, in multiple linear regression, we can, among other metrics, minimize the root mean squared error (RMSE) which measures the overall accuracy of the model (278).

A convenient property of linear regression is that an analytical solution exists to minimize the cost function. It is equivalent to the maximum likelihood estimate of Gaussian Linear Models (277) and given by

$$b = (X^T X)^{-1} X^T y, \tag{18}$$

where b and y represent the vectors of coefficients and targets, respectively.

A widely used and very intuitive metric to assess the model performance of regression models in general, thus also linear regression, is the coefficient of determination, also known as the *R-squared statistic* or R^2. It quantifies the proportion of variance explained by the model. The formula is given by

$$R^2 = 1 - \frac{\sum_i^n (Y_i - \widehat{Y}_i)^2}{\sum_i^n (Y_i - \overline{Y})^2}, \tag{19}$$

where $\overline{y}$ denotes the mean of the predicted values. The closer the R^2 is to 1, the better the prediction performance.

Linear regression is useful since a linear combination of variables can model many phenomena. Moreover, it is fast, scalable, easy to interpret, and, more importantly, has great explanatory power despite its simplicity. In the following section, we want to introduce another linear model called logistic regression in which, in contrast to linear models with continuous outcomes, binary outcomes are predicted.

2.3.1.2. Logistic Regression

Logistic regression (214) predicts a *binary outcome* (e.g. 0/1; yes/no; diseased/healthy) as opposed to linear regression, where the outcome is continuous. This is also called a classification problem. Usually, the class of interest is encoded ‘1’ and the common class as ‘0’. Logistic regression is very similar to MLR covered in the last section except for the difference in the response and the interpretation of the coefficients. The response can first be thought of as the probability that the class label will be 1. This would yield the following equation

$$p \ = \ b_o + b_1X_1 + \ b_2 \ X_2 + \ \dots + b_jX_j \tag{20}$$

However, because p is not guaranteed to stay in the range [0;1], which is a necessity for probabilities, it is modeled by applying the logistic function to the predictors to ensure this property. The logistic function is a sigmoid function that takes any real input and outputs a value between zero and one and is defined as

$$S(x) \ = \ \frac{1}{1+e^{-x}} \ = \ \frac{e^{x}}{1+e^{x}} \tag{21}$$

Hence the probability that the class label will be 1 is given by

$$p \ = \frac{e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots + b_jX_j \ )}}{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}} \tag{22}$$

Then the probability that the class label will be 0 is given by

$$1 \ - \ p \ = \frac{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )} - e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}}{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}} \tag{23.1}$$

$$= \frac{1}{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}} \tag{23.2}$$

To get the exponential expression out of the denominator, odds instead of probabilities are considered. Odds are simply the ratio of ‘successes’ to ‘non-successes’, both denoted as 1s and 0s, respectively. Therefore the odds of observing the class label of interest ($Y \ = \ 1$) is given by

$$Odds(Y \ = \ 1) \ = \frac{p}{1-p} \ , \tag{24}$$

Therefore we get

$$\frac{p}{1-p} \ = \frac{e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}}{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}} \ * \ \frac{1+e^{(b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j \ )}}{1} \tag{25.1}$$

$$= e^{b_o+b_1X_1+ b_2 \ X_2+ \dots +b_jX_j} \tag{25.2}$$

To remove the exponential expression we apply the log-function to both sides of the equation and get

$$log(\frac{p}{1-p}) \ = \ b_o + b_1X_1 + \ b_2 \ X_2 + \ \dots + b_jX_j \tag{26}$$

We thus have expressed the log of the odds of observing the class label of 1 as a linear combination of our predictors. To obtain probabilities from the log odds we apply the logistic function given by

$$P(Y \ = \ 1) \ = \frac{1}{1+e^{-log(Odds(Y=1)))}} \tag{28}$$

We thereby obtain a linear model to predict the probability of observing the class label 1. To actually classify samples as 1s and 0s we apply a cutoff rule. Usually, a probability cutoff of 0.5 is chosen, so samples with a predicted probability $\geq \ 0.5$ are classified as 1s and $< \ 0.5$ as 0s.

The coefficients $\widehat{b}_i$ in the logistic regression model give the log of the odds ratio for variable X_i. Consider a single binary explanatory variable X_1, then the associated coefficient b_1 is defined as:

$$log(Odds\ Ratio) \ = \ log(\frac{odds(Y=1\,|\,X=1)}{odds(Y=1|\,X=0)}) \tag{29.1}$$

$$= log(odds(Y \ = \ 1\,|\,X \ = \ 1)) \ - \ log(odds(Y \ = \ 1\,|\,X \ = \ 0)) \tag{29.2}$$

$$= b_0 \ + \ b_1 * 1 + \ ... \ + \ b_n X_n - (b_0 + b_1 * 0 + \ ... \ + \ b_n X_n) \tag{29.3}$$

$$= b_1 \tag{29.4}$$

So it gives the log of the odds ratio that $Y \ = \ 1$ when $X \ = \ 1$ versus the odds that $Y \ = \ 1$ when $X \ = \ 0$, when all other variables are held constant. For instance, if the odds ratio equals 2, then the odds that Y = 1 are twice as high when X = 1 as opposed to X = 0. Because the coefficients $\widehat{b}_i$ are the log of the odds ratios we can exponentiate the coefficients to get the odds ratios, which are more readily interpretable. For continuous variables, the interpretation is analogous and gives the change in the odds ratio for a unit change in X.

In simple linear regression (review Section 2.3.1.1. Linear Regression) we considered the RSS as our cost function to penalize model fits. In logistic regression we consider the logistic loss cost (LLC) function given by

$$LLC \ = \frac{1}{m}\sum_{i=1}^{m} - Y_i log(\widehat{Y_i}) \ + \ (1 \ - \ Y_i)log(1 \ - \ \widehat{Y}_i), \tag{30}$$

where m is the number of samples, Y_i the true class label, and $\widehat{Y}_i$ the predicted class label. Intuitively, the logistic loss penalizes more predicting 1 when the actual class label is 0 or when predicting 0 when the class label is actually 1. An analytical solution does not exist for the maximum likelihood estimator of the logistic model (277) which is why numerical optimization methods like, for instance, gradient descent (279) or Newton's method (280) have to be used to train the model.

2.3.2. Tree Models

Tree models, also called Classification and Regression Trees (CART) (215), or simply decision trees, are a highly popular, effective, and efficient class of machine learning models for regression and classification tasks. Especially their powerful descendants, random forests and boosted trees are among the most widely used models for predictive modeling. They are straightforward to interpret, as they are based on building a series of if-then-else rules, fast and highly scalable as well as competitive relative to the most advanced methods like neural networks. We used a variation of decision trees, called Extreme Gradient boosting trees (XGB), in our second project (Chapter 3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures) for predicting the transcriptional pause state of genes. We thus want to introduce in the following sections the underlying methodology of tree-based models. We first cover the basic idea of building decision trees (Section 2.3.2.1. The Decision Tree Model) and introduce a powerful extension of these, called random forests, in Section 2.3.2.2. The Random Forest Model and conclude this section with one of the most advanced tree-based models, the Extreme Gradient boosting Trees, in Section 2.3.2.3. The Extreme Gradient Boosting Regression Tree Model (XGB).

2.3.2.1. The Decision Tree Model

Tree models are very powerful, popular, easy to interpret, scalable and fast. In contrast to linear models, they have the power to identify hidden patterns that correspond to variable interactions without explicitly modeling these interactions during the modeling phase. Trees are a series of if-then-else rules imposed by the predictors that partition the data set into smaller subsets of homogenous samples. A recursive partitioning algorithm constructs the trees.

Given a target variable Y and a set of P predictor variables X_i for $i = 1, ..., P$, a partition A of samples, the recursive partitioning algorithm will find the best way to partition A into two subpartitions A_1 and A_2, according to the following procedure:

1. For each predictor variable X_i
    a. For each value v_i of X_i
        i. Split the samples in A with X_i values <= v_i as one partition (A_1) and the remaining samples where $X_i > v_i$ as the second partition (A_2)
        ii. Measure the homogeneity of the samples in each of the subpartitions
    b. Select the value v_i that produces maximum within-partition homogeneity of samples
2. Select the variable X_i and the split value v_i that produces maximum within-partition homogeneity of samples

The recursive nature arises when we first initialize partition A with the entire dataset at hand, apply the partitioning algorithm to split A into subpartitions A_1 and $A_2$and repeat partitioning on each of the subpartitions A_1 and A_2 and all resulting subpartitions until no further partition can be made that sufficiently improves the homogeneities of the partitions.

Different metrics exist to measure partition homogeneity and their use cases differ in dependence of the type of the target variable. We will focus on the metrics of a 2-class classification problem where the response is binary, but generalization to multi-class problems also exists. Usually, the Gini Impurity (GI) or Entropy of Information (EI) are used (281). The GI of a partition A is defined as

$$GI(A) = p(1 - p), \tag{31}$$

where p gives the proportion of misclassified samples in the partition.

The EI of a partition is given by

$$EI(A) = \ - p\, log_2(p) - (1 - p)\, log_2(1 - p) \tag{32}$$

During the partitioning of the data any such metric is used and a weighted (by the number of samples in the partition) average is calculated and whichever partition yields the lowest impurity score is then selected along with the split variable X_j and split value v_j which correspond to a learned rule.

Predicting a continuous variable is analogous except that the scoring metric is the root mean squared error within the partition, given by

$$RMSE = \sqrt{\frac{\sum_i^n (Y_i - \widehat{Y_i})^2}{n}}, \tag{33}$$

where n is the number of samples, Y_i the true outcome and $\widehat{Y}_i$ the predicted outcome. The predicted value $\widehat{Y}_i$ for a sample i is the average true outcome $\overline{Y}_X$ of a partition X that contains sample i .

Due to the inherent nature of the recursive partitioning, the algorithm terminates when there is no partitioning possible anymore. This is when each sample forms its own partition (fully grown tree), which corresponds to an overfitted model. In this setting, there exists a definite path from the root of the tree to reach each sample, also called leaves, that results in perfect prediction of the target values. This results in an 100% accuracy of the model as the model shifted from learning general rules that identify reliable relationships in the data to learning specific rules that only apply to specific observations. These specific rules thus represent noise in the data. To overcome this problem, the tree growth process, i.e., the recursive partitioning, is terminated at an earlier stage when certain tree growth conditions are met. In fact, the goal is to stop the tree from growing at a stage that will generalize well to new data. A common way to accomplish this is to avoid splitting a partition if the resulting subpartitions are too small, i.e., specifying a minimum partition leaf size. Another way is to stop partitioning if the resulting subpartitions do not significantly increase the homogeneity or decrease the impurity. Finding the optimal parameter settings is exploratory work and an instance of a bias-variance tradeoff (discussed later in Section 2.3.4. The Bias-Variance Tradeoff). Usually, it is done in combination with cross-validation (discussed later in Section 2.3.5. Cross Validation) and regularization (discussed later in Section 2.3.7. Regularization) and is called hyperparameter tuning (discussed later in Section 2.3.6. Hyperparameter Tuning).

Regarding prediction, harnessing the prediction power of multiple trees is often superior to only a single decision tree. Random forests are extensions of the simple decision tree algorithm, which tries to harness the power of multiple trees based on a random subset of data and variables. In the following, we want to introduce random forest to provide an intuition for gradient boosting trees covered in the section thereafter (Section 2.3.2.3. The Extreme Gradient Boosting Regression Tree Model (XGB)).

2.3.2.2. The Random Forest Model

Random forests (216) are based on the idea of ensemble learning in which the average or the majority vote of multiple models trained on the same data (Bagging (282)), often outperform a single model. The random forest model uses bagging, i.e. bootstrap aggregation, and bootstrap sampling (review Section 2.2.2. Variance of Estimates) the predictors in each step. A random forest is built according to the following procedure:

1. Take a bootstrap subsample from the sample space (called the “bag”)
2. Take a bootstrap subsample p from the predictor space P
3. Apply recursive partitioning as discussed earlier (review Section 2.3.2.1. The Decision Tree Model) with the randomly selected subset of predictors p until tree growth conditions are met
4. Repeat steps 1-3 multiple times, resulting in multiple trees (collectively called a random forest)

A rule of thumb to set p is to choose $\sqrt{P}$ predictors where P is the total number of available predictors. The individual tree performances are given by the out-of-bag (OOB) prediction performances, i.e. when predicting the samples that were left out (not contained in the ‘bag’) during bootstrap sampling (Step 1) with the grown tree. Once training is complete, predictions for unseen samples can be made by averaging the predictions from all the individual regression trees or by taking the majority vote in the case of a classification task.

Predictors are scored according to variable importance measures. One way to judge the importance of a predictor is by randomly permuting the predictor, thus removing any predictive power of that variable, and then quantifying the decrease in model accuracy on the OOB data, which is effectively a cross-validated estimate. On the other hand, one can measure the mean decrease in the Gini Impurity score for all of the nodes that were split on a specific variable which measures how much the purity of nodes is improved by including the variable. However, this way of quantifying variable importance is inferior to the former, as it is based on the training data compared to the accuracy based on the OOB data.

Random forest models also require hyperparameter optimization. Among others, important parameters to avoid overfitting are the minimum number of samples for terminal nodes, the maximum number of nodes in the tree, the depth of the trees, and the number of trees to grow (Step 4).

2.3.2.3. The Extreme Gradient Boosting Regression Tree Model (XGB)

Extreme Gradient Boosting (XGB) Tree models (217) apply gradient boosting to decision trees. Gradient Boosting (283) is based on the idea that an ensemble of multiple weak learners can generate a single strong learner while allowing the optimization of an arbitrary differentiable loss function. In the context of tree models, the weak learners correspond to individual decision trees. Specifically, in gradient boosting trees, a series of decision tree models (like in random forests (216)) are trained sequentially (in contrast to parallelly in random forests), where each successive decision tree seeks to minimize the error of the previous decision tree (boosting in contrast to bagging (282) as in random forests). The combination of each previous and successive model is expected to perform better than either model alone since each successive model overcomes the shortcomings of the combined boosted ensemble of all previous models. To formalize, the model can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F, \tag{34}$$

where K is the number of trees, f_k is a function in the functional space F , and F is the set of all possible trees and x_i the i-th sample. The objective function to assess the goodness of model fits is given by

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} w(f_t), \tag{35}$$

where n is the number of samples, T is the number of trees, y_i the true target values, $\hat{y}_i$ the predicted target values, l a loss function (e.g. logistic loss, squared error loss) and $w(f_t)$ the complexity of the tree f_t defined in detail later. In order to optimize the objective function, we have to build optimal trees f_t that produces minimal prediction error. However, constructing all possible trees is computationally too expensive and a simplified additive approach is followed to select locally optimal trees, i.e. new trees are added sequentially. This is in contrast to the random forest model where trees are trained parallelly. Predictions of the t-th iteration ($\hat{y}_i^{(t)}$) can be formulated as:

$$\hat{y}_i^{(0)} = 0 \tag{36.1}$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \tag{36.2}$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \tag{36.3}$$

$$\dots$$

$$\hat{y}_i^{(t)} = \sum_{t=1}^{T} f_t(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{36.4}$$

A tree structure f_t is added if it optimizes the objective function. However, the cost functions in XGB are approximated by a Taylor series (284) of order two to improve upon the computational complexity. This approximation is necessary to scale XGB to a large dataset with many features. The approximation of order two is sufficient as we seek new parameters in the neighborhood of the starting points. In the general case, for an arbitrary loss function our optimization goal for the new tree at step t is defined as:

$$L = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^{\,2}(x_i)] + w(f_t), \tag{37}$$

where g_i and h_i are defined as

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{38}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{39}$$

Because the objective function only depends on g_i and h_i XGBoost can support custom loss functions. The complexity of a tree $w(f)$ is defined as

$$w(f) = \gamma L + \frac{1}{2}\lambda \sum_{j=1}^{L} w_j^2 \tag{40}$$

where w is the vector of scores on leaves, q a mapping function that assigns each data point to the corresponding leaf, L is the number of leaves and γ , λ regularization parameters subject to hyperparameter tuning. Considering this the objective value of the t-th tree is defined as:

$$obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{L} \frac{G_j^2}{H_j+\lambda} + \gamma L \tag{41}$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ while $I_j = |i|q(x_i) = j\}$ which is the set of indices of data points assigned to the j-th leaf. Given the objective function, we can learn trees, however, enumerating all possible trees is computationally infeasible. Thus instead of evaluating all possible trees, we can build trees in a stepwise fashion, and optimize one level of the tree at a time. Specifically, we split a leaf into two leaves and calculate it's Gain, defined as

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}] - \gamma \tag{42}$$

where subscripts L and R index the left and right leaves, respectively. This basically represents the score on the new left leaf, the score on the new right leaf, the score on the original leaf, and the regularization on the additional leaf.

Extreme Gradient boosting trees are very powerful, able to compete with neural networks, fast, scalable as well as flexible, and applicable to regression and (multi-class) classification tasks (217, 285, 286). However, they are also subject to overfitting very easily but have many hyperparameters to adjust their behavior and overcome this issue. For instance, to avoid overfitting, in each tree, the difference between the tree gain (after adding a split node) and a predefined hyperparameter gamma can be calculated and if the difference is negative the branch can be removed. A root node will never be removed if its branches are not removed. Other key determinants of overfitting are the tree depth, the minimum number of observations in terminal leaf nodes, or the number of successive trees to be grown.

In our first project about the transcriptional pausing of the Polymerase II we have harnessed the predictive power of XGB models to predict the degree to which a gene is paused based on features extracted from large-scale genomic data sets (Section 3.1.10. Model Training). With this section, we want to conclude the algorithmic introduction of supervised machine learning models used in our research projects and proceed to briefly cover basic concepts to consider when building supervised machine learning models in general. We will start off by discussing the process of interpreting a model and its predictors, called feature interpretation.

2.3.3. Feature Scoring

A key step after having built a machine learning model of any type is to investigate the contributions (feature importances) of predictors (features) to the model predictions. This process is called *feature interpretation (287)* since we are trying to understand and interpret the features the model has chosen to make its predictions. Feature importances measure the degree to which a model relies on a particular feature to make its predictions. The general idea is that the more a model's prediction performance depends on a specific feature, the more predictions will change as a function of perturbing that feature. Depending on the model type, the feature interpretation approach will vary substantially.

The best way to interpret a model is to use models which are interpretable out of the box (in contrast to black-box models) like, for instance, linear or logistic regression models, simple tree-based models, or Naive Bayes, just to name a few. For example, in simple linear regression, we investigate the beta coefficients of the features to assess their direction of effect and effect sizes or the odds ratios of features in logistic regression models. In a random forest model, there are no beta coefficients, and we can use the Gini Impurity score for all of the nodes that were split on a specific variable to assess the variable's split performance and thereby its predictive power. Simple models that are interpretable out of the box are often preferred over complex ones for their ease of interpretation despite lower accuracy. However, the growing availability of big data increasingly requires complex models. This, in turn, necessitates a trade-off between accuracy and interpretability of a model's output or elaborate feature interpretation methods. Approaches exist like building surrogate models in which interpretable models are trained on the same data as the complex black box model but trained to predict the predictions of the black-box model. This essentially provides an interpretable model that can explain the predictions of a more complex model. Analogous to this idea, local surrogate models can be trained as well. For instance, Local Interpretable Model-agnostic Explanations (LIME) (288) builds local surrogate models to explain why single predictions were made in contrast to a global surrogate model in which a collective set of predictions are explained.

A relatively new framework for interpreting model predictions used in our first project with tree-based models is SHapley Additive exPlanations (SHAP) (289). In contrast to other approaches like LIME, SHAP has been justified as the only consistent additive feature attribution approach with several unique properties which agree with human intuition. It is based on game theory and combines ideas from several established methods, and can explain the output of any machine learning model. SHAP transfers game-theoretic concepts to machine learning and asks the following question: Let M be a set of arbitrary entities, C the coalition (the set) of all entities with $m \in M$ and a (coalition) value V cooperatively produced by these members $m \in C$, then, how much does each individual member m of the coalition C, that collaboratively produces a value V, contribute to that final value V? In the context of machine learning this translates to asking how much each feature value of each individual sample contributed to the prediction V of the sample target value compared to the average of the target. For instance, suppose we have trained a machine learning model that predicts gene expressions based on three gene sequence composition features, the gene length, the number of exons, and the number of transcription start sites for a gene. For a gene of interest with the following feature values, gene length = 1000, number of exons = 5, and number of transcription start sites = 2. Given a gene expression level of 10 FPKM, the question is, how much does each of these feature values of gene x contribute to the final prediction of the expression level of 10 FPKM of gene x, compared to the average expression level that we observe across all genes?

The most naive solution is to attribute a value of $\frac{1}{|C|}$ to each member in C, assuming equal contributions of each individual member (feature value). However, in the case of coalition member interactions (feature interactions), this does not lead to fair attributions, as certain coalition permutations can cause coalition members to contribute more than the sum of their parts. The so-called Shapley value can be computed for each coalition member to overcome this limitation. It quantifies the marginal contribution of a member over all possible permutations of the coalition, i.e., all possible (membership) subsets of the coalition C. This can be accomplished by enumerating all possible coalitions (a subset of feature values for a specific sample) with and without a specific member (feature value) and calculating the difference in the produced value V (target prediction for the sample) between these sets, i.e., a coalition containing a member against a coalition that does not contain that specific member. This yields the marginal contribution of that specific member in that specific coalition where it was contained. The overall marginal contribution of that specific member can be obtained by averaging all the differences we get by comparing all coalitions with and without that specific member. This is repeated for each coalition member to arrive at a vector of SHAP values, representing all marginal contributions of all members over all possible permutations of coalitions. This can be formulated as

$$\phi_i = \frac{1}{n} \sum_{\forall S \subseteq C \setminus \{i\}}^{N} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S)), \tag{43}$$

where ϕ_i is the SHAP value for member i (a feature value), n the total number of members, N the total number of possible subsets $S \subseteq C \setminus \{i\}$, $v(x)$ a function that maps a set of members (binary indicator vector where 1 denotes 'present' and 0 denotes 'missing') to their respective cumulative contribution V to the target prediction which in turn is weighted by the number of possible coalitions of size $|S| \setminus \{i\}$ to adjust for differential coalition contributions due to differential coalition sizes. The formula can be interpreted as the mean marginal contribution of member i to each possible coalitions S excluding member i, weighted by the individual coalition sizes and normalized by the total number of members. These Shapley values enable the calculation of the total contributions of specific coalition sets. Let s' be a binary indicator vector that specifies the members (features values) to be considered. The total value of that coalition is given by

$$g(s') = \phi_o \sum_{i=1}^{|s'|} \phi_i s'_i \tag{44}$$

However, calculating the SHAP values for all possible coalitions by reevaluating the model is NP-hard since the permutation space increases exponentially such that this approach becomes computationally infeasible with large data sets with hundreds of thousands of features. Therefore, the authors propose Kernel SHAP as an approximation to the calculation of the SHAP values (289). The idea is to pass samples with omitted features through the trained model instead of retraining the model with omitted features. This way, the model does not need to be retrained. However, the trained model cannot make predictions on samples with omitted features that were originally included in the training phase of the model. A solution to circumvent this problem is to replace the values of the features that will be omitted with the mean of those features over all samples, random numbers or sampling values from the data at random while fixing the original values of features that are not omitted. These synthetic samples simulate the missingness of features and can approximate the Shapley values (contributions) of features that were omitted by simply taking the average over all model outputs of the synthetic samples. Through a specific weighting of these model outputs of a specific feature permutation (synthetic sample) based on the total number of features in the model, the number of coalitions with the same number of features as this particular sample, and the number of features included and excluded in this sample's permutation, it is guaranteed that the resulting value is equivalent to the Shapley value ϕ (see (289) for proof). This can be formulated as the following loss function that needs to be minimized

$$L(f, g, \pi) = \sum_{s \in S} (f(h(s)) - g(s))^2 * \pi_s \tag{45}$$

where s is a binary indicator vector specifying the features to include in a specific coalition (feature subset), f the machine learning model, $h(x)$ a mapping function that returns a vector of feature values where features that are not indicated by x are replaced, for instance, by the mean of these features, g a linear model containing the coefficients (shapley values ϕ_i) and finally the kernel π defined as

$$\pi_s = \frac{n-1}{\binom{n}{|s|}} |s| * (n - |s|) \tag{46}$$

where n is the total number of features, thereby essentially weighting by the total number of features in the model, the number of coalitions with the same number of features as this particular sample and the number of features included and excluded in this sample's permutation. This problem corresponds to a weighted multiple linear regression problem and can be solved in polynomial time complexity, enabling the application of Shapley Explanations to large-scale data sets with many features. Fitting the kernel weighted linear model across all data points then allows to attribute the marginal feature contributions.

There are other forms of SHAP like TreeShap (286) for tree-based models or DeepShap (290) for neural network architectures that make model-specific optimizations to optimize the time complexity further. However, Kernel Shap is universal and can be applied to any machine learning model, as we have done in our tree models.

Shapley values have several desirable properties. For instance, as compared to LIME (introduced at the beginning of the section), Shapley values fairly distribute model contributions among the feature values of the samples (called Efficiency Property). Because SHAP values are additive and are calculated per sample and feature value providing high prediction transparency, they also provide a full explanation, i.e., the sum of the Shapley values of a sample corresponds to the predicted target of that sample. This is in contrast to many common feature importance metrics that average the contributions of the features over the samples, which reduce the interpretability of the models. With the additive property of SHAP, for instance, samples could be clustered by their Shapley values to obtain sets of samples that are similarly influenced by the same features and feature values. Alternatively, one could cluster the features by their Shapley values to obtain sets of features that contribute similarly.

2.3.4. The Bias-Variance Tradeoff

The bias-variance tradeoff (291) refers to finding the model parameter settings at which a model generalizes well beyond the training data. The bias is the difference between the expected value of an estimator $\widehat{f}$ and the true function f that we want to estimate. The variance is the difference between the expected value of the squared estimator minus the squared expectation of the estimator. The bias component quantifies the average accuracy of the model across different possible training sets, while the variance component quantifies the model's sensitivity to small changes in the training set. Let $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ be the training data set for the machine learning task, where x_i are the training instances (samples) and y_i real valued targets associated with each x_i via $y = f(x) + \epsilon$, where ϵ is noise with zero mean and variance σ^2. A machine learning model then finds a function $\widehat{f}(x)$ that approximates the true function $f(x)$ by learning patterns underlying the training instances (x_i, y_i) of the training data D by minimizing the loss function, defined as the squared difference between predicted and actual outcomes, i.e. $(f(x) - \widehat{f}(x))^2$. The expected value of the loss function with respect to unseen samples (different instances of training data sets from the population) can be decomposed into its bias-variance components (265) to understand better and adjust the performance of the underlying learning algorithm. For notational convenience, let $\widehat{f}(x) = \widehat{f}$ and $y = f(x) + \epsilon = f + \epsilon$, then the bias-variance decomposition for the squared error loss is as follows:

$$E[(y - \widehat{f})^2] = (f - E[\widehat{f}])^2 + Var[\epsilon] + Var[\widehat{f}] \tag{47}$$

$$= Bias[\widehat{f}]^2 + \sigma^2 + Var[\widehat{f}] \tag{48}$$

This shows that the squared error loss is composed of three main components, the bias, the variance, and the noise. The error term (noise) is intrinsic to the measurements $y_{i,,}$ and thus irreducible. Similar to the bias-variance decomposition of the squared error loss of regression models, the 0/1 loss of classification models can be decomposed into its bias-variance components (292) as well.

The bias and variance components allow us to assess the generalizability of a model, i.e., whether it is overfitting or underfitting. Overfitting refers to the phenomenon where a model has, in addition to real and general relationships and patterns in the data, learned the noise in the data (high variance). Underfitting is the opposite case, where a model learns too little, i.e., it has not learned any useful pattern underlying the data that could be used to explain new unseen data (high bias). For instance, a fully grown (maximal tree depth) decision tree in which the model has learned each individual sample instead of the patterns underlying the samples is said to have high variance because it can make accurate predictions across all samples in the training data set. However, it is highly biased towards the training dataset and may not generalize well to new unseen data instances. Thus, the model has to be trained so that it generalizes well to unseen data by properly balancing the model's bias and variance components. For instance, simplifying a model by training on a reduced feature set obtained, for example, through feature selection procedures, can decrease the variance of the model. Analogously, adding features introduces variance but decreases the bias. Alternatively, increasing the number of training instances can decrease variance. In general, adjusting the bias-variance tradeoff can be achieved in terms of how the model is trained with respect to the data (see the following section) and the model parameter settings (see Section 2.3.6. Hyperparameter Tuning).

2.3.5. Cross Validation

To increase the confidence of trained models and the patterns learned from the data, any machine learning model should be tested (validated) on how they perform on independent data they never were exposed to. For model validation, some portion of the original full data is usually set aside (hold-out test data set), then a model is trained on the other part and subsequently validated on the hold-out test data set. However, there is still some uncertainty about the hold-out test data set, e.g., one could have randomly chosen the samples that are easier to predict simply by chance. Thereby the model would perform quite well on this hold-out set, yet if confronted with another data set, it might perform much worse. The idea of cross-validation (293) mitigates this problem by introducing a series of hold-out data sets, called folds, to validate on (called k-fold-cross-validation). Let k be the number of folds, then the procedure for k-cross-validation is as follow:

1. Split the original data into k equal parts
2. Set one part $(1/k)$ of the original full data set aside, called the holdout test dataset or fold
3. Train a model on the remaining part of the original dataset, called the training set
4. Measure the performance of the model on the holdout test data set and keep track of model performance metrics
5. Replace the $1/k$ holdout test dataset and set aside another $1/k$ holdout test dataset that does not contain any datapoint from the previous holdout test datasets
6. Repeat steps 2-4 either k times or until every sample has been used in a holdout data portion
7. Finally average the model performances over all holdout-data sets

Because the optimal value for k is unknown, it can be included in the hyperparameter estimation procedure (discussed in the next section). However, tuning the fold parameter (k) is computationally expensive. Thus, usually 5- or 10-fold cross-validation is performed, in which either 20% or 10% of the original full data is used to test the model.

2.3.6. Hyperparameter Tuning

Many machine learning models have an array of parameters to be set that are not learned during the model's training (hyperparameters) and instead have to be inferred by evaluating the models with different parameter settings. Finding the optimal parameters is called *hyperparameter tuning or optimization (294)*. Hyperparameters control the model's behavior and allow the investigator to balance the bias-variance tradeoff (review Section 2.3.4. The Bias-Variance Tradeoff). However, the exact values to be used are not known. The idea is simple, for a set of different hyperparameter combinations, one takes the one that performs the best on the data. Instead of taking the hyperparameter combination that performs best on the full data set, one takes the parameter combination in which the model yields the lowest overall error as computed by averaging the errors from each cross-validation fold (review Section 2.3.5. Cross Validation). The model is cross-validated multiple times with different hyperparameter settings, and the settings that yield the best overall cross-validation runs are retained to obtain globally optimal hyperparameters.

However, although the model is evaluated on holdout test data sets (the k folds), its hyperparameters are adjusted according to the fold performances, i.e. the hyperparameters might fit the k folds better than a completely unseen dataset that was not even included during the k-fold-cross-validation simply because it's parameters were tuned according to the distribution the k folds. Hence it is advisable to either set aside yet another test data set before conducting hyperparameter tuning with k-fold cross-validation or just perform k-fold cross validation but validate the final model with the best k-fold cross-validation hyperparameter settings on a completely independent dataset from another source not used during cross-validation. Usually, the latter is unavailable, so one could, for instance, set aside 30% of the original data set, perform hyperparameter tuning with k-fold cross-validation on the remaining data set, and validate the model with the best parameters on the 30% holdout test data set. This ensures an unbiased estimate of the model performance on the unseen data set as well as its parameters.

2.3.7. Regularization

Regularization (295) during model training is an additional means to adjust the bias-variance tradeoff. It allows us to avoid overfitting, obtain sparse models to alleviate model interpretation, and to overcome multicollinearity of input features (independent variables). Multicollinearity describes the situation in which one or more independent variables can be expressed as a linear combination of other independent variables leading to a large variance of estimated coefficients and non-interpretable models. Non-linear models like tree models are not sensitive to multicollinearity but still suffer from overfitting or large feature spaces such that regularization can be a powerful tool to overcome these problems.

Regularization works by adding a penalty term to the cost function of a machine learning model. Therefore, depending on the model and thus the cost function used, the regularization procedure changes accordingly. The most common regularization techniques (295) for linear models are L1 regularization, L2 regularization, and the combination of both L1 and L2 regularization (296), referred to as lasso regression, ridge regression and elastic net regularized regression, respectively. Ridge regression penalizes the sum of squared coefficients and shrinks the coefficients (reducing bias). In contrast, lasso regression enforces the sum of the absolute coefficients to be less than a constant value, forcing some of the coefficients to be zero (reducing variance). Lasso is suitable when there are small numbers of significant coefficients with others close to zero, i.e., when only a small subset of coefficients influences the response. Hence, lasso regularization leads to sparse significant coefficients alleviating feature selection and interpretation. On the other hand, ridge regularization is suitable when there are many coefficients, especially correlated predictors, with about the same effect size, i.e., when most predictors influence the response. However, an elastic net provides the benefits of both L1 and L2 regularization, essentially performing feature elimination from Lasso and feature coefficient reduction via Ridge to improve model predictions.

3. Predictive model of transcriptional elongation control identifies trans-regulatory factors from chromatin signatures

This chapter covers our first project that is concerned with the identification of novel trans-regulatory factors modulating the promoter-proximal pausing of the Polymerase II during the transcription of mammalian protein-coding genes (review Section 1.1. Thesis Aims and Section 1.3. Gene Transcription & Regulation). The contents are entirely based on my manuscript currently (May 2022) under review in *Nucleic Acid Research* and also available on bioRxiv: Akcan and Heinig. ‘Predictive Model of Transcriptional Elongation Control Identifies Trans-Regulatory Factors from Chromatin Signatures, Toray Akcan, Matthias Heinig’ BioRxiv (2022). The referenced supplementary tables are not included in this dissertation and can be found in the supplementary materials section of the publication.

To briefly recap our aims, the cell continuously adapts to changing environmental conditions to sustain cell homeostasis for proper cell functioning (1, 2). It achieves this through the modulation of the transcription of genes dependent on internal and external stimuli. At first, the polymerase synthesizes a short nascent RNA fragment once the pre-initiation complex (PIC) assembles at the promoter site. This is followed by the pausing of the polymerase and requires other regulatory signals for it to either enter productive elongation or terminate transcription prematurely (6). This represents a rate-limiting and regulatory step that renders gene transcription a discontinuous process (297). The limitation on the transcriptional throughput per unit of time represents a critical early regulatory step in the maturation of full-length transcripts (7, 8). Hence, understanding this regulatory layer is indispensable for a holistic understanding of protein biogenesis. Transcriptional pause regulatory factors entail cis- and trans-acting factors, either promoting transcriptional pause or elongation states (16). However, we still lack quantitative descriptions of the relative importance of associated factors and processes. Likewise, we lack systematic approaches to identify previously unknown regulators of pausing and elucidate their roles in other RNA regulatory processes. Hence, we aimed to reduce these gaps by identifying novel cis- and especially trans-regulatory factors and elucidating their functional backgrounds and relative importance.

This project is based on integrating large-scale genomic data sets for feature engineering purposes as inputs for a machine learning task to learn genomic features to distinguish between the Polymerase II transcriptional pause and elongation states (Fig. 3.1).

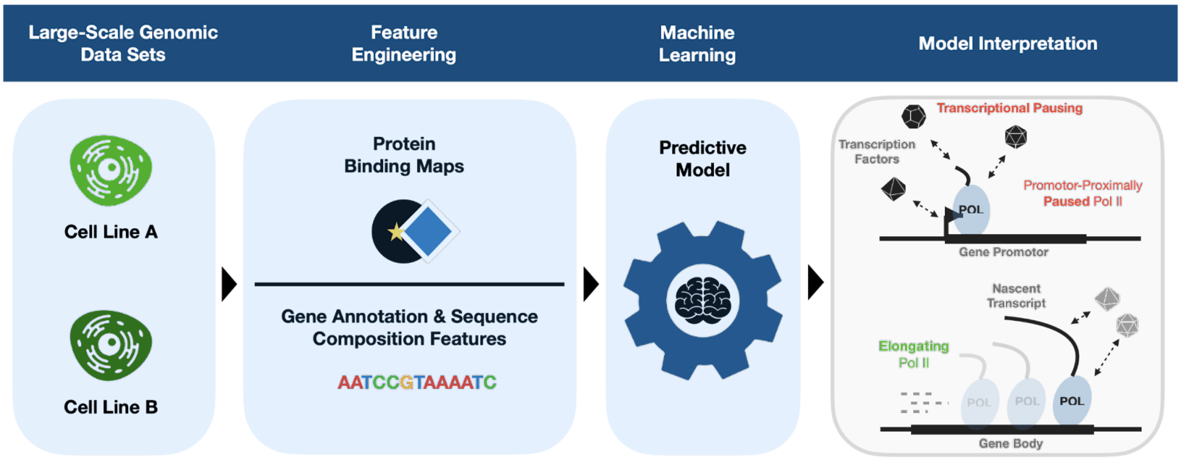


Figure 3.1: **High-level overview of the transcriptional pausing workflow pipeline.**

For this purpose, we captured the genomic and transcriptomic context of gene transcription by integrating large-scale data sets of CHIP-seq and eCLIP-seq experiments from ENCODE, providing elaborate protein binding maps that enable the investigation of genomic binding events with the potential to reveal novel trans-regulatory factors. On the other hand, the extraction of DNA sequence features from gene annotations from GENCODE served to identify potential novel intrinsic cis-regulatory elements of pausing. The integration of GRO-seq data enabled us to capture the productivity of the polymerase and contrast states of promoter-proximal pausing and transcriptional elongation. By harnessing the power of predictive models by integrating the transcriptional context in the form of machine learning features into an Extreme Gradient Boosting Tree regression model to learn this contrast, we demonstrated the predictive value of obtained features and underlying cis- and trans-acting regulatory elements. The investigation of the underlying model structure in terms of feature contributions then allowed us to discern factors with high model impacts and propose novel regulators of transcriptional pausing. The integration of prior knowledge about molecular functions of incorporated factors further enabled us to confirm the strong interconnection of the transcriptional pause mechanism with other RNA regulatory processes, especially splicing. In addition, the identification of previously unknown 7SK ncRNA interacting RNA-binding proteins and demonstrated predictive values in obtained models of pausing further strengthened the role of the 7SK pause mediator complex in the transcriptional pause mechanism.

In the following section, we will first provide the methodological background of described analyses (Section 3.1. Materials & Methods) and then focus on the results in more detail (Section 3.2. Results) and conclude the chapter with a discussion on the obtained results (Section 3.3. Discussion).

3.1. Materials & Methods

This project heavily draws upon publicly available datasets provided by the *Encyclopedia of DNA Elements (ENCODE)* Consortium (298, 299) and the GENCODE (300) project, generating large-scale genomic data sets or annotations based on biochemical protocols.

The *ENCODE* Consortium is an international collaborative effort to build a comprehensive list of functional elements in the human genome. ENCODE employs a broad array of different kinds of assays and methods to identify such functional elements. This is accomplished by sequencing a diverse range of RNA sources by developing bioinformatics methods and human curation. ENCODE applies high-quality standards, both when generating data and processing data. Invaluable data resources are provided, ranging from 3D chromatin structure to chromatin accessibility, chromatin interactions and methylation, RNA quantifications, and transcription factor bindings. More importantly, these data sets are generated for a broad range of diverse cell types, enabling us to investigate inter-cell type differences. We made use of resources from ENCODE of large-scale transcription factor binding maps and RNA quantifications for two different cell lines to build features for predictive models applied in this first project about transcriptional pausing.

The GENCODE project (300) was formed as part of the pilot phase of the ENCODE project to identify all protein-coding genes within the ENCODE regions and aim to build an encyclopedia of genes and gene variants including protein-coding and non-coding genes, pseudogenes, small and long non-coding RNA genes, and many others. Its current release (Release 36, December 2020) includes 60660 genes and 232117 transcripts. This release also utilizes the latest GRCh38 human reference genome assembly. GENCODE provides an invaluable resource for investigating gene and gene variants on a sequence and expression levels. We used the GENCODE gene annotation data sets to obtain annotations of protein-coding and non-coding genes for downstream analyses in our projects.

As invaluable resources for computational biologists, we harnessed multiple data sets from both sources in our project, as introduced in the following subsections.

3.1.1. Integration of Transcript Annotations (GENCODE)

For feature engineering purposes as predictors in our machine learning models, we crafted gene-centric features of protein binding events and annotations for protein-coding and non-coding genes with gene annotations obtained from the GENCODE (300) database for the hg19 (GrCH37) genome build. To increase the confidence of obtained transcripts (N=81745 transcripts of 20167 genes), only those which were also supported by RefSeq

(301) annotations were considered (N=30186 of 18889 genes). To capture interpretable binding sites derived from CHIP-seq and eCLIP-seq datasets (Sections 3.1.5. Integration of Genomic Transcription Factor Binding Sites (CHIP-seq) & 3.1.6. Integration of Transcriptomic Transcription Factor Binding Sites (eCLIP-seq)) we obtained 5-prime, intronic, coding exonic, and 3-prime genomic regions for each of those transcripts. The transcripts were further annotated with their respective gene symbols from HUGO gene nomenclatures (HGNC) (302) from GENCODE.

A set of non-coding transcripts annotated as one of *miscRNA*, *miRNA*, *snoRNA*, *snRNA and lincRNA* representing miscellaneous, micro, small nucleolar, small nuclear and long intervening RNA biotypes, respectively, was obtained through the appropriate filtering of the GENCODE annotation set. Analogous to the protein-coding transcripts, the non-coding transcripts and their 5-prime, intronic, exonic, and 3-prime genomic regions served to capture interpretable binding sites derived from CHIP-seq and eCLIP-seq datasets. In addition, the non-coding transcripts were used in downstream analyses in the context of the 7SK non-coding RNA (see 7SK non-coding RNA).

3.1.2. Integration of Transcript Quantifications (RNA-seq)

To only consider expressed transcripts, pre-processed transcript quantifications from total RNA-seq (review Section 2.1.2. RNA-Sequencing) experiments with two replicates each were obtained from the ENCODE project (303, 304) for the K562 and HepG2 cell lines for the hg19 (GrCH37) genome build. The transcript expressions were filtered for valid ENSEMBLE (305) IDs, annotations in the aforementioned GENCODE and RefSeq transcript annotation set, and to be expressed (fragments per kilobase million (FPKM) > 0) in both of the replicates. The FPKMs were log10-transformed. These filtering steps lead to the consideration of 16403 (K562) and 16670 (HepG2) protein-coding and 2655 and 1950 non-coding transcripts for the K562 and HepG2 cell lines, respectively.

3.1.3. Integration of Transcription Start Site Annotations (CAGE)

We further integrated Cap-analysis Gene Expression Data (CAGE) (306) (review Section 2.1.3. Cap Analysis Gene Expression (CAGE)) transcription start sites (TSS) to increase the confidence of previously captured transcripts. The read counts of the most correlating replicates were aggregated per cell fraction and normalized to transcripts per million reads (TPMs). The resulting TSS were then aggregated to CAGE transcription start site clusters (CTSS cluster) with a parametric clustering (307) approach with a minimum TPM threshold of 0.1 per cluster while excluding singletons with TPM less than 0.1. Previously captured transcripts were further filtered, retaining only transcripts whose transcription start site was also the dominant CAGE transcription start site (CTSS) in a cell-type-specific CTSS cluster. This led to considering 16194 and 16412 protein-coding transcripts in the K562 and HepG2 cell lines, respectively.

3.1.4. Quantification of Promoter-Proximal Pol II Pausing (GRO-seq)

To enable the quantification of transcriptional pausing at protein-coding genes we integrated Global-Run-On-sequencing (GRO-seq) (308) (review Section 2.1.4. Global-Run-On-Sequencing (GRO-seq)). GRO-seq allows assessing Pol II productivity based on the nascent RNA fragment output during the transcriptional cycle. The commonly used pausing index (PI), also known as the traveling ratio (102, 309)), which is the log2 ratio of GRO-seq read signals at the transcription start site (TSS) to the GRO-seq read signals in the gene body, was used as a measure of pausing. By maximizing the inverse correlation of the PI (Fig. 3.2) with the corresponding transcript expressions (Pearson's ρ=-0.68 (K562) and ρ=-0.66 (HepG2)) with varying TSS window sizes, we were able to improve upon the definition of the PI.

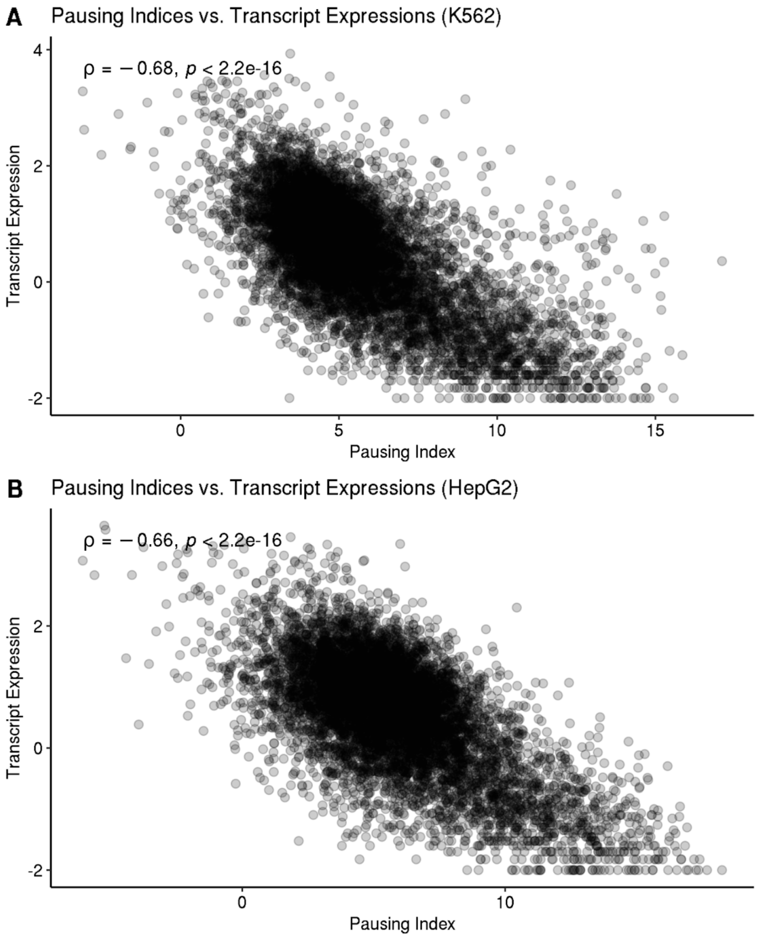


Figure 3.2: Association of the transcriptional pausing index with transcript expression levels. Pausing indices (x-axis) are inversely correlated with transcript expression levels (FPKMs, y-axis) in the K562 (A) and HepG2 (B) cell line. Pearson's correlation coefficient rho (ρ) with the associated p-value is depicted in the upper left.

This approach is motivated by the fact that high PIs, representative of transcriptional pausing, should result in low gene expression profiles and vice versa. This approach led to a sharp TSS window size of 3bp ranging 1bp up-and downstream of the TSS, rendering the remainder of the transcripts as the gene body windows. Read lengths of 30bp (K562, GSM1480325) and at least 25bp (HepG2, GSM2428726) ensure that the most frequent Pol II pause site and associated components (310) are covered by this approach. The GRO-seq read counts within the windows were then normalized by the respective window size and a pseudo count of 1 read was added to each resulting window to enable the log2 transformation when building the ratio. For each of the 16194 (K562) and 16412 (HepG2) expressed protein-coding transcripts we calculated the PI in a strand-specific manner. Transcripts that only contained the DNA base letters (A,T,C,G) along the whole transcript were considered, which further led to the exclusion of 16 and 9 protein-coding transcripts in the K562 and HepG2 cell lines, respectively. This filtering was necessary to ensure that we exclude erroneously mapped reads to capture the full GRO-seq read signals along with the remaining transcripts and enable comparable signal counts. Because GRO-seq signals can not be uniquely ascribed with overlapping transcripts, resulting in convoluted PI signals, we only considered non-overlapping transcripts and thus were left with 8555 and 8456 protein-coding transcripts in the K562 and HepG2 cell line, respectively (Fig. 3.3).

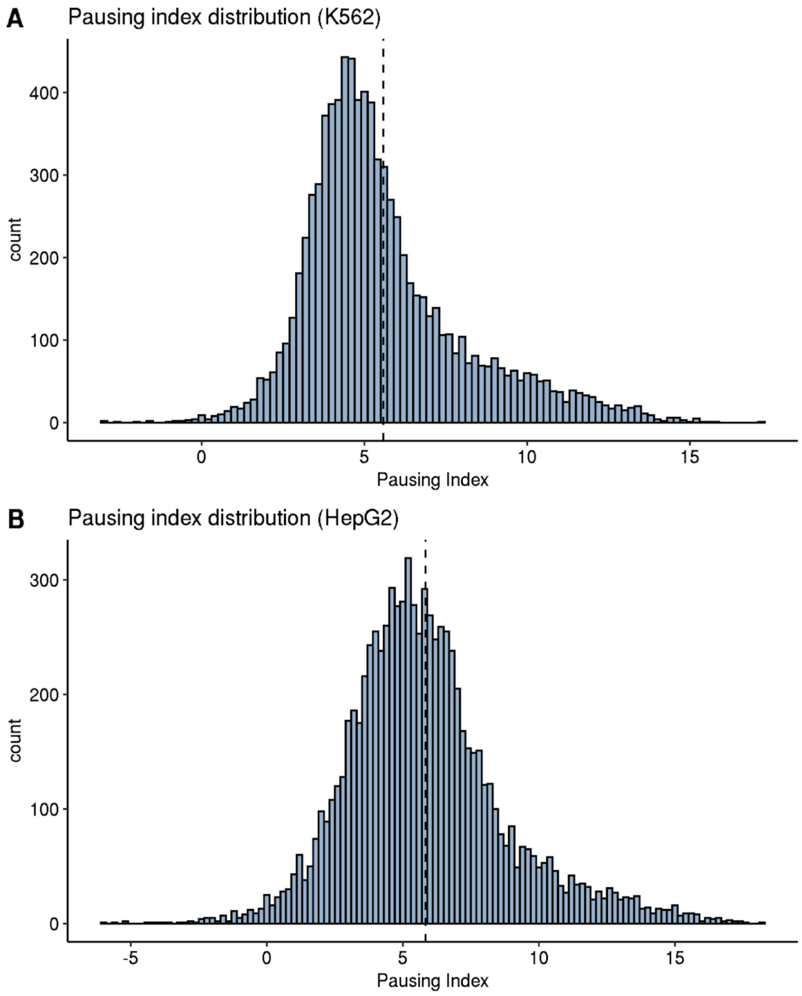


Figure 3.3: Distribution of pausing indices. Histograms of the distribution of pausing indices (PIs) in the K562 (A) and HepG2 (B) cell line. Dashed lines indicate the mean pausing indices, the x-axes the PIs and the y-axes the PI counts.

3.1.5. Integration of Genomic Transcription Factor Binding Sites (CHIP-seq)

Gene-centric genomic protein binding events identified by chromatin immunoprecipitation sequencing (CHIP-seq) (311) (review Section 2.1.5. Chromatin Immunoprecipitation-Sequencing (CHIP-seq)) data served to engineer features of as inputs for the machine learning models. Binding sites for DNA binding proteins (DBPs) were obtained from the ENCODE project from all available CHIP-seq experiments for the K562 and HepG2 cell lines for the hg19 (GrCH37) genome build. Available peak-called data sets (bed-files) were harvested for these binding sites while perturbation experiments were excluded, and to increase the confidence in the obtained binding sites, only optimal irreproducible discovery rate (IDR) (308, 312) thresholded replicated peaks were considered for downstream analyses. Additionally, untagged and newer versioned experiments were prioritized over tagged and older versioned experiments. This yielded 5041190 (K562) and 4138805 (HepG2) genomic binding signals for 309 (K562) and 211 (HepG2) factors (see **Supplementary Tables S1 & S2** for CHIP-seq factors per cell line) and served for feature engineering purposes (see Feature Engineering).

3.1.6. Integration of Transcriptomic Transcription Factor Binding Sites (eCLIP-seq)

Gene-centric transcriptomic protein binding events identified by enhanced cross-linking immunoprecipitation sequencing (CHIP-seq) (311) (review Section 2.1.6. Enhanced Cross-linking Immunoprecipitation-Sequencing (eCLIP-seq)) data served as additional features as inputs for the machine learning models. All available processed eCLIP-seq experiments from the ENCODE project for the K562 and HepG2 cell lines for the hg19

(GrCH37) genome were harvested to obtain binding sites of RNA-binding proteins (RBPs). Perturbation experiments were excluded and newer versioned experiments were prioritized over older ones. Only optimal IDR thresholded replicated peaks were considered. We thereby obtained 409839 (K562) and 435015 (HepG2) transcriptomic binding signals for 120 (K562) and 103 (HepG2) factors (see **Supplementary Table S2** of eCLIP-seq factors per cell line).

3.1.7. Targeting the 7SK non-coding RNA

For identifying known and novel 7SK non-coding RNA (7SK) binding proteins, we first filtered the GENCODE transcript annotation data set for all 7SK annotated transcripts. Pseudo 7SK transcripts were included if they were expressed at least at the median expression level of all expressed non-coding transcripts. This ensures that we can capture factors binding pseudo 7SK transcripts, which may, in turn, compete (313) for respective binding sites with factors that bind the non-pseudo version. Factors with at least one eCLIP binding site on any of the 7SK transcripts (see **Supplementary Tables S9 & S10)** corresponded to the set of 7SK binding factors. This set of 7SK binding factors was identified in a cell-type-specific manner. Beyond the identification of 7SK binding proteins, this set assessed their predictive value in the context of transcriptional pausing.

3.1.8. Model Feature Engineering

As predictors for the machine learning models in predicting the gene-wise pausing index of protein-coding genes, we have engineered features of DNA- and RNA-binding events at protein-coding transcripts and the closest non-coding transcripts up- or downstream of the TSS of each protein-coding transcript. We further engineered DNA sequence and annotation features of protein-coding transcripts to capture DNA sequence effects that might modulate transcriptional pausing.

Specifically the following features were created:

- transcript length
- strand specification
- chromosome specification
- location on the linear genome
- number of annotated exons
- average exon width
- exon density (ratio of the length of the transcript including introns to the number of exons)
- fraction of exonic sequence (ratio of the length of all exonic sequences to the transcript length)
- GC content of the whole transcript including introns
- width of CAGE transcription start site cluster (CTSS)
- GC content of CTSS
- distance to most proximal CpG island along with information about the CpG island (length, and features of the sequence: number of CpGs, number of C and G, percentage of CpG, percentage C or G, and ratio of observed to expected CpG)
- binary encoding whether the transcript is a housekeeping gene
- binary encoding of RBP binding events separately for 5’/3’-UTR, introns and coding exons
- binary encoding of DBP binding events separately for 5’/3’-UTR, introns and coding exons excluding Pol II bindings as these are expected to be naturally correlated with the prediction target
- binary encoding of RBP/DBP binding events separately for 5’/3’-UTR, introns and coding exons of the two most TSS proximal non-coding RNAs excluding Polymerase II bindings as these are expected to be naturally correlated with the prediction target

CpG island annotations were obtained for the hg19 genome build from the UCSC golden path (cpgIslandExt.txt.gz). Housekeeping gene annotations were taken from (314). To avoid having more features than samples (genes) which would lead to overfitting, we only considered the two most proximal ncRNAs because in combination with CHIP-seq and eCLIP-seq signals on more than two proximal ncRNAs, the feature space would overgrow the sample space. To achieve faster and more accurate model convergences, we rescaled numeric features not in the range [0:1] to that range. DNA- and RNA-binding signal features were binary encoded (binding (1) or non-binding (0)) (see **Supplementary Tables S11 & S12** for the number of binding events per factor on individual genomic or transcriptomic regions for each cell line). See Fig. 3.4 for the distributions of DNA sequence features in the K562 cell line (see **Supplementary Figures S3.1** for Fig. 3.4 analog for the HepG2 cell line).

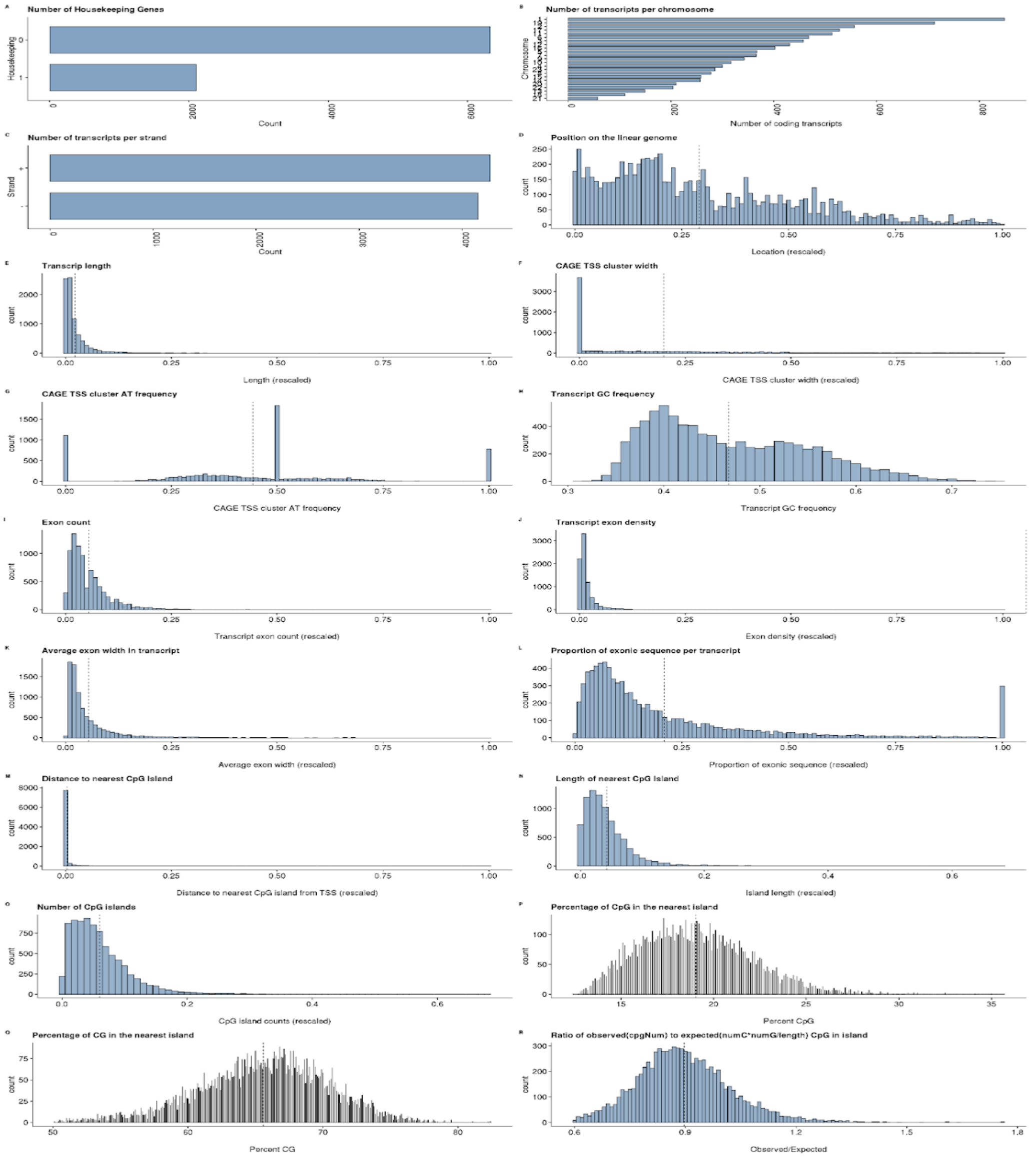


Figure 3.4: Distribution of gene annotation and sequence composition features in the K562 cell line. Numeric features were rescaled to the range [0;1]. In sub-figures A-C the x-axes show the counts of features, and the y-axes the feature values. In sub-figues D-R the x-axes show the feature values, and the y-axes the counts of features.

As discussed in the next section, various data matrices for a series of regression tasks based on different feature sub-spaces based on prior domain knowledge were built with these feature vectors.

3.1.9. Model Feature Engineering based on Prior Knowledge

To characterize and quantify the relevance and importance of pre- co- or post-transcriptional events in the context of transcriptional pausing, we stratified the feature space of protein binding events into functionally related sets of proteins. This was achieved through integrating Gene Ontology (GO) (315, 316) annotations which provide sets of proteins implicated in specific biological processes. We aimed at functional sets of proteins annotated in biological process (BP) ontology terms in the context of transcription ranging from Chromatin, Initiation, Elongation, Termination, and Splicing. Specifically, we considered the following GO terms: **Chromatin (**chromosome organization, GO:0051276**;** chromatin organization, GO:0006325; chromatin remodeling, GO:0006338), **Initiation (**RNA polymerase II preinitiation complex assembly, GO:0051123; transcription initiation from RNA polymerase II promoter, GO:0006367), **Elongation (**transcription elongation from RNA polymerase II promoter, GO:0006368), **Termination (**termination of RNA polymerase II transcription, GO:0006369), **Splicing (**mRNA splicing via spliceosome GO:0045292; regulation of alternative mRNA splicing via spliceosome, GO:0000381) and **Processing (**mRNA export from nucleus, GO:0006406; mRNA 3'-end processing, GO:0031124). We extended the Elongation factors by pause regulatory factors from the literature (3, 87, 317) if not already included. This pause regulatory factor set was comprised of the super elongation complex (SEC) factors CCNT1, CCNT2, ELL, ELL2, ELL3, AFF1, AFF4, MLLT1, MLLT3, established pausing factors NELFA, NELFB, NELFCD, NELFE, SUPT4H1, SUPT5H, SUPT6H, SUPT16H, BRD4, MYC, TAF1, TBP, PAF1, and CDK9 (P-TEFB), as well as 7SK ncRNA pause mediator complex binding factors LARP7, HEXIM1, HEXIM2, and MEPCE (see also **Supplementary Table S13**). However, only n=19 of all established pausing factors could be considered since not all were assayed in the CHIP-seq and eCLIP-seq experiments. Therefore the final **Elongation** factor set consisted of POLR2A, POLR2B, POLR2G, POLR2H, MLLT1, SUPT5H, GTF2F1, BRD4, WDR43, NCBP2, HNRNPU, LARP7, MYC, TAF1, TBP, AFF1, EZH2, PAF1, and SSRP1. However, Polymerase associated factors (POLR2A, POLR2B, POLR2G, POLR2H) were excluded since these are expected to affect pausing by definition. To quantitatively assess the importance of unknown or less well-established 7SK associated factors (review Section 3.1.7. Targeting the 7SK non-coding RNA or **Supplementary Table S4** of 7SK binding factors per cell line), a set of 7SK binding proteins was generated as well. To capture general pausing associated factors, we further formed the union of the Elongation and 7SK associated factor set (**Elongation+7SK**). See **Supplementary Tables S14 & S15** for a list of factors in each resulting functional factor set per cell line**.**

Subsequently, each factor set was further stratified into sequence-specific and non-sequence-specific binders based on annotations from The Molecular Signatures Database (MSigDB) (318, 319), a collection of annotated gene sets, the Catalog of Inferred Sequence Binding Preferences (CIS-BP) (320), a library of transcription factors and their binding motifs and the Homo sapiens comprehensive model collection (HOCOMOCO) (321), a collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis based on binding motifs (see **Supplementary Tables S16 & S17**).

To arrive at different feature matrices based upon prior domain knowledge, the feature vector space of binding events was then accordingly grouped by these factor sets (see **Supplementary Table S18** of factor presence in feature subspaces). These feature matrices always included DNA sequence and annotation features of protein-coding genes. We thereby obtained feature matrices to build an array of predictive models based on features with a defined biological function.

We further built 100 random models randomizing over the number of factors, the factors itself and their binding patterns to enable a baseline comparison of model performances. The binding patterns were randomized according to the observed binding proportions.

3.1.10. Model Training

Extreme Gradient Boosting Tree (XGB) regressors (review Section 2.3.2.3. The Extreme Gradient Boosting Regression Tree Model (XGB)) were trained to predict the pausing index with each of the subsets mentioned previously (review Section 3.1.9. Model Feature Engineering based on Prior Knowledge). We validated our models with a 5-fold cross validation (review Section 2.3.5. Cross Validation) procedure as well as applied trained models to completely independent test data sets from the cross cell line, which provided us with an unbiased estimate of the model performances as trained models have neither seen the genes target distribution nor the specific feature distributions of the cross cell type. This was accomplished by reducing each feature matrix to features that are common to both cell lines. We refer to these as the *synchronized* models. In contrast to these models, the *individual* models incorporated all available features specific to a cell type. To enable a proper validation for individual models 50% of the available data points were held out prior to training as an independent test data set. Although this hold out test data set is not from an independent cell line as is the case with the synchronized models, it still provides an unbiased model performance estimate as trained models have also not seen any of the data points. Regression with squared loss was chosen for the learning objective and the coefficient of determination (R-squared, R^2) was used to evaluate trained models. See **Supplementary Table S19** for hyperparameter (review Section 2.3.6. Hyperparameter Tuning) specification and the zenodo repository for R-Data structures with all model matrices.

3.1.11. Feature Scoring & Interpretation

As a scoring metric for model feature contributions we have used Shapley additive explanations (SHAP) (286, 289) (review Section 2.3.3. Feature Scoring). SHAP has the potential to explain the output of any machine learning model and in contrast to the well known variable importance metric in tree-based machine learning models it is able to show the positive or negative relationship for each feature with the target. Additionally, as opposed to most feature importance metrics which average over all genes, SHAP assigns each gene its own set of SHAP values which greatly enhances the prediction transparency. Moreover, SHAP values are additive and enable to aggregate over feature contributions of subsets of features. This enabled us to capture contributions of protein binding per protein and group these proteins into sets of positive and negative regulatory factors. As an example we obtain model contribution scores for genomic and transcriptomic transcription factor bindings on the 5’UTR, exons, introns and 3’UTR as identified by CHIP-seq and eCLIP-seq, respectively. We can then calculate total contributions of a specific factor by aggregating the SHAP scores per such factor over each gene region (5’UTR, exons, introns and 3’UTR) on the genome or transcriptome which in turn enables us to select factors with high effect sizes to pinpoint specific pause regulatory factors.

3.2. Results

3.2.1. Predictive Models of Transcriptional Pausing

The turning point of promoter-proximally paused Pol II (Fig. 3.5A, promoter-proximally paused Pol II) into its productive form of nascent RNA synthesis (Fig. 3.5A, elongating Pol II) is a function of trans-regulatory protein co-factors as well as cis-regulatory DNA and RNA sequence features (3, 322). We refer to these cis- and trans-regulatory factors as chromatin signatures.

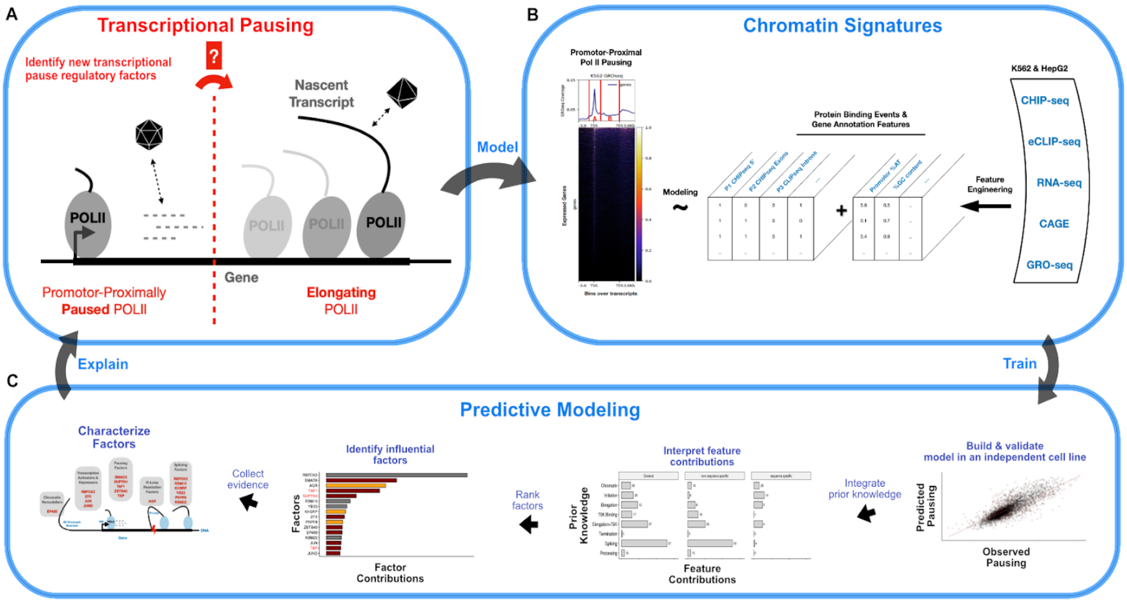


Figure 3.5: Low-level overview of the transcriptional pausing workflow pipeline. (A) Identifying transcriptional pause regulatory factors by contrasting promoter-proximally paused Polymerase II against the Polymerase's productive elongation phase of nascent RNA synthesis. **(B)** Feature engineering (middle, right) with large-scale genomic data sets for two different cell lines to build the chromatin context of transcriptional pausing (A) based on chromatin signatures consisting of gene-wise protein binding patterns and gene annotation and sequence composition features. Construction of the prediction target (left), the pausing index, by relating GRO-seq read densities at the TSS to GRO-seq read densities in the gene body. **(C)** Predictive modeling to forecast promoter-proximally paused Pol II with chromatin signatures (B), followed by evaluating prior knowledge in the context of transcriptional pausing and selection of factors as novel regulators of pausing.

To identify such pause regulatory cis- and trans-acting factors we compiled large-scale genomic and transcriptomic binding maps based on ENCODE data sets. Potential cis-regulatory elements were captured with gene annotation and sequence composition features. These chromatin signatures then served to follow a systematic machine learning approach with Extreme Gradient Boosting tree (XGB) regressors to predict the degree of transcriptional pausing at protein-coding genes (Fig. 3.5B) and subsequently reveal explanatory factors (Fig. 3.5C).

To validate our models, we have integrated relevant data sets of two different cell lines (K562 and HepG2). To facilitate the validation in independent cell lines, we obtained relevant data sets for two different cell lines (K562 and HepG2). The gene-wise pausing index served as the prediction target (review Section 3.1.4. Quantification of Promoter-Proximal Pol II Pausing (GRO-seq)); see **Figure 3.3** for pausing index distributions) It quantifies the degree to which a gene is paused (high pausing index) or elongated (low pausing index). Model features were systematically compiled through the integration of genome-wide CHIP-seq (review Section 3.1.5. Integration of Genomic Transcription Factor Binding Sites (CHIP-seq)) and eCLIP-seq (review Section 3.1.6. Integration of Transcriptomic Transcription Factor Binding Sites (eCLIP-seq)) data from the ENCODE project which provide DNA and RNA binding sites on the genome and transcriptome respectively (see **Supplementary Tables S11 & S12**). Gene annotation and sequence composition features were engineered based on GENCODE transcript annotations (review Section 3.1.8. Model Feature Engineering, see **Figure 3.4** (K562) and **Supplementary Figure S3.1** (HepG2)). To validate the expressions of gene transcripts and increase the confidence in transcription start sites, we further integrated CAGE transcription start sites (review Section 3.1.3. Integration of Transcription Start Site Annotations (CAGE)). In total, we obtained 2503 features of 2485 DNA & RNA binding and 18 gene annotation features in the K562 cell line and 1832 features of 1814 DNA & RNA binding and 18 gene annotation features in the HepG2 cell line. Subsequently, we trained an Extreme Gradient

Boosting Tree regressor (review Section 3.1.10. Model Training and **Supplementary Table S20**) to predict the pausing index of protein-coding genes (n=8426 in K562), which achieved high accuracy and was able to explain up to 68% of the observed variance (R^2~0.68 on 50% hold-out test data set, K562) in the pausing index (Fig. 3.6A).

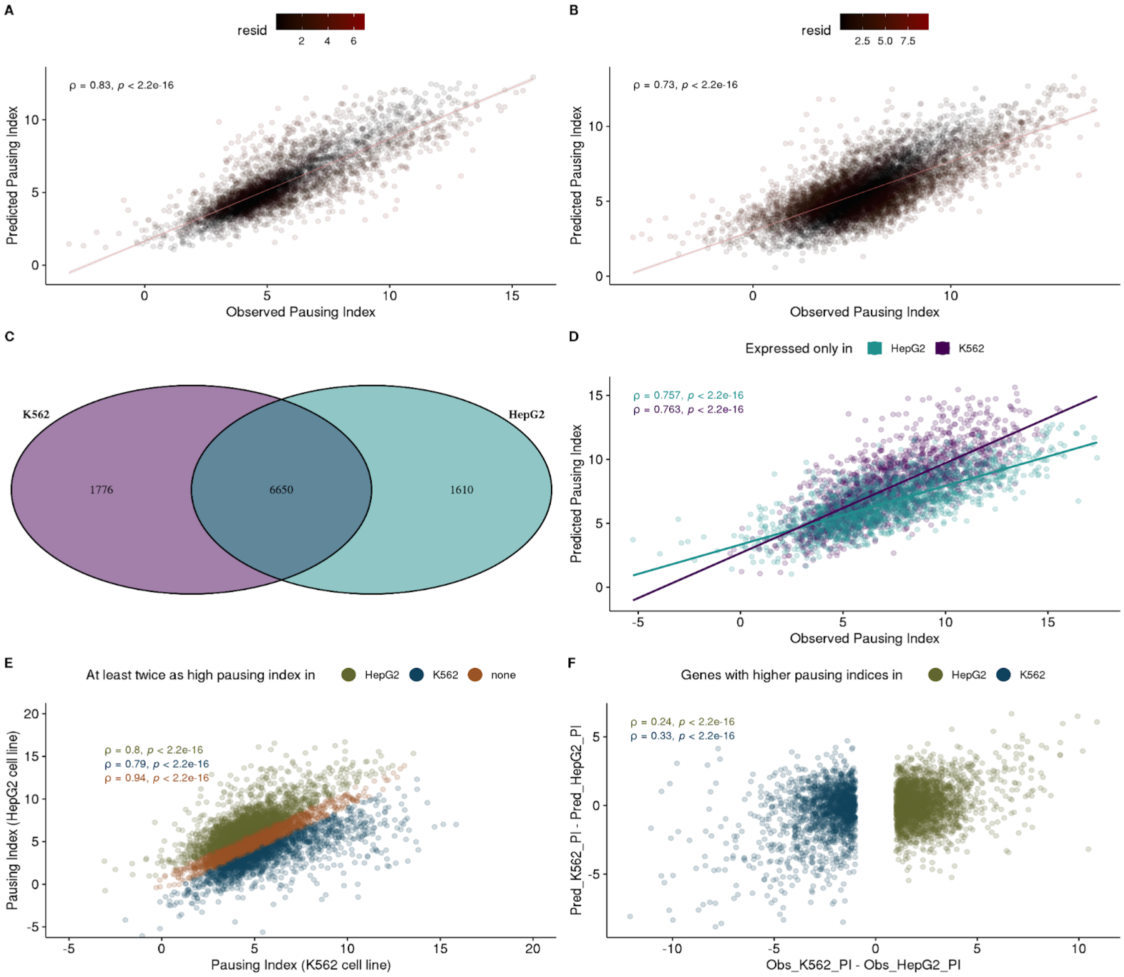


Figure 3.6: Transcriptional pausing model prediction performances. (A) Scatterplot of observed and predicted pausing indices (log2-scale, K562). Predictions stem from a 5-fold cross-validated and regularized XGB regression model applied to an independent 50% hold-out test dataset from the same cell line taken prior to model training. The upper left shows Pearson's correlation coefficient rho (ρ) with the associated p-value. The residual regression error is colored in red (see legend *resid*). **(B)** Scatterplot of observed and predicted pausing indices (log2-scale, K562). Predictions stem from a 5-fold cross-validated and regularized XGB regression model applied to the independent test dataset of the cross-cell line (HepG2). The model was trained with features available in both cell lines. The upper left shows the Pearson's correlation coefficient rho (ρ) with the associated p-value. The residual regression error is colored in red (see legend *resid*). **(C)** Venn diagram of expressed transcripts between the two cell lines (K562, HepG2). **(D)** Scatterplot of observed and predicted pausing indices. Predictions stem from a 5-fold cross-validated and regularized XGB regression model trained on each cell line and applied to data of genes exclusively expressed in the cross-cell line. The upper left shows Pearson's correlation coefficient rho (ρ) with the associated p-value. The residual regression error is colored in red (see legend *resid*). **(E)** Scatterplot of observed pausing indices of each cell line. HepG2 specific transcripts with at least a 2-fold higher pausing index than in K562 are colored green, K562 specific transcripts with at least a 2-fold higher pausing index than in HepG2 are colored blue, and transcripts with similar pausing indices in both cell lines (less than a 2-fold change), thus not specific to any of the cell lines, are colored in orange. The upper left shows the Pearson correlation coefficients rho (ρ) for each of the subgroups with the associated p-value. **(F)** Scatterplot of differences of observed pausing indices between the cell lines against the prediction differences obtained from models trained in each cell line and

applied to data from the cross cell line. Differences are shown for gene transcripts identified in E), which showed a 2-fold change between cell lines. Each model was trained on features available in both cell lines.

The model performance can be evaluated in different validation settings, specifically with **1)** a model trained on one cell line and applied to the full data of the other cell line (Fig. 3.6B) **2)** a model trained on one cell line and applied to genes exclusively expressed in the other cell line (Fig. 3.6D) as well as **3)** with a model trained on one cell line and applied to genes that are present in both of the cell lines that show significantly different pausing indices representing extreme observation specific to the other cell line (Fig. 3.6F). For model performances of a model trained on the HepG2 cell line and validated on the K562 cell line, see **Supplementary Figure S3.2** for figure 3.6 analog for the HepG2 cell line.

High prediction performance on the independent cross-cell type test data set (Fig. 3.6B, performance on HepG2 data of K562 model) in which the model was able to explain up to 53% of the variance in the pausing signal, demonstrates the predictive power and generalizability of the underlying model and features. The drop in the model performance from an R^2 of 0.68 (Fig. 3.6A) to an R^2 of 0.53 and is likely caused by the reduced amount of available features in the HepG2 cell line (39% of all features (n=987) of n=2503 features available in the K562 cell line).

The observed performances in the cross-cell type prediction task (Fig. 3.6B) may result from **1)** the signal of ubiquitously expressed genes that are similar between cell types, as might be the case with housekeeping genes, or **2)** from general learned rules that allow predicting cell type-specific pausing indices from cell type-specific chromatin signatures. To distinguish between the two cases, we identified the sets of cell-type genes (Fig. 3.6C) and evaluated the performances of models trained on one of the cell lines to predict the pausing index of genes exclusively expressed in the other cell line (Fig. 3.6D). The K562 model explained up to 57% (HepG2 model up to 58%) of the observed variance in the pausing indices in the HepG2 (K562) cell line, respectively.

We additionally evaluated the model's capability to predict cell type-specific distributions of differential (fold change >= 2) pausing indices (Fig. 3.6E, blue, green). Specifically, we evaluated the agreement of the differences of observed pausing indices between the cell lines against the differences of predictions of the pausing indices using models trained in one of the cell lines and applied to data in the other cell line (Fig. 3.6F). There is a substantial decrease in the model performances from a correlation coefficient of 0.73 for the prediction on the entire HepG2 cell type data (Fig 3.6B) or 0.76 on HepG2 cell type-specific genes (Fig 3.6D) to 0.24 (Fig. 3.6F, HepG2 specific pausing indices; green). Nonetheless, the model maintains some predictive power for extreme observation of pausing indices specific to the cross cell type. This further underlines the model's ability to generalize to other cell lines.

Taken together, the model’s predictive power on the intra-cell type holdout test data sets (Fig. 3.6A), the inter-cell type test data set (Fig. 3.6B) as well as its ability to predict pausing indices of cell type-specific genes (Fig. 3.6D) and cell type-specific differential pausing indices (Fig. 3.6F), demonstrate that the model has sufficient discriminatory power to explain a large fraction of the observed variance in the pausing index and that it captured general cell-type independent rules of pausing regulation. We thus proceeded with feature interpretation and selection approaches to be able to extract potentially novel regulators of transcriptional pausing. These downstream analyses were performed on data from the K562 cell line due to the increased amount of available data points (features).

3.2.2. Linking Transcriptional Regulatory Steps with Transcriptional Pausing

We next aimed at a mechanistic explanation of the underlying predictive features. For that purpose, we have used Shapley Additive Explanations (SHAP) (286, 323) as a feature scoring metric (see Materials & Methods) that captures the directional contribution of each feature specifically for each gene on the target variable. In dependence of the factor's relevance for pausing and their interaction with other model features, a feature may positively (increase) or negatively (decrease) affect the pausing index or exert no effect at all (Fig. 3.7A). The sum of the individual effects then converge in predicted pausing indices representing the average output whether a gene is paused or not.

Due to the intricate connection of transcriptional pausing with other steps of gene expression ranging from chromatin organization (324–326), transcription initiation (297, 310, 327), to splicing (328–330) and post-transcriptional transcript processing (331–333), we next evaluated factors associated with these pre-, co- or post-transcriptional events according to their importances in predicting transcriptional pausing. Based on Gene Ontology (GO) annotations of protein membership terms in biological processes we first generated sets of of regulators (see Methods and **Supplementary Tables S14 & S15**) representative of major RNA processing events ranging from *Chromatin*, *Initiation*, *Elongation*, *Splicing*, *Termination* to *Processing*. We extended the *Elongation* factor by established pausing factors identified from literature. Due to the pause regulatory role of the 7SK non-coding RNA complex (7SK) (334–339) we built a set of factors that bind the 7SK in the eCLIP-seq datasets (see Methods and **Supplementary Tables S9 & S10** for 7SK binding factors per cell line) to be able to evaluate the role of of these binders in transcriptional pausing. This set consisted of the very well known 7SK binder LARP7, the pausing associated factor AQR previously not associated with the 7SK as well as the following factors not previously associated with pausing: SSB (LARP3), HNRNPK, DGCR8, PCBP1, ATF, ZNF800, XRCC6, NCBP2, SBDS, YWHAG, GRWD1, ZNF622, SRSF7, TARDBP and BUD13. A general set of pause regulatory factors was further generated by forming the union of the Elongation and 7SK associated factor sets (*Elongation+7SK*). All resulting sets of regulators were then grouped into known sequence-specific and non-sequence-specific binders (see **Supplementary Tables S16 & S17**) to be able to assess the relevance of sequence specific binding events. Subsequently we aggregated model feature contributions (Fig. 3.7A) per functional process (Fig. 3.7B) over all factors from a specific process.

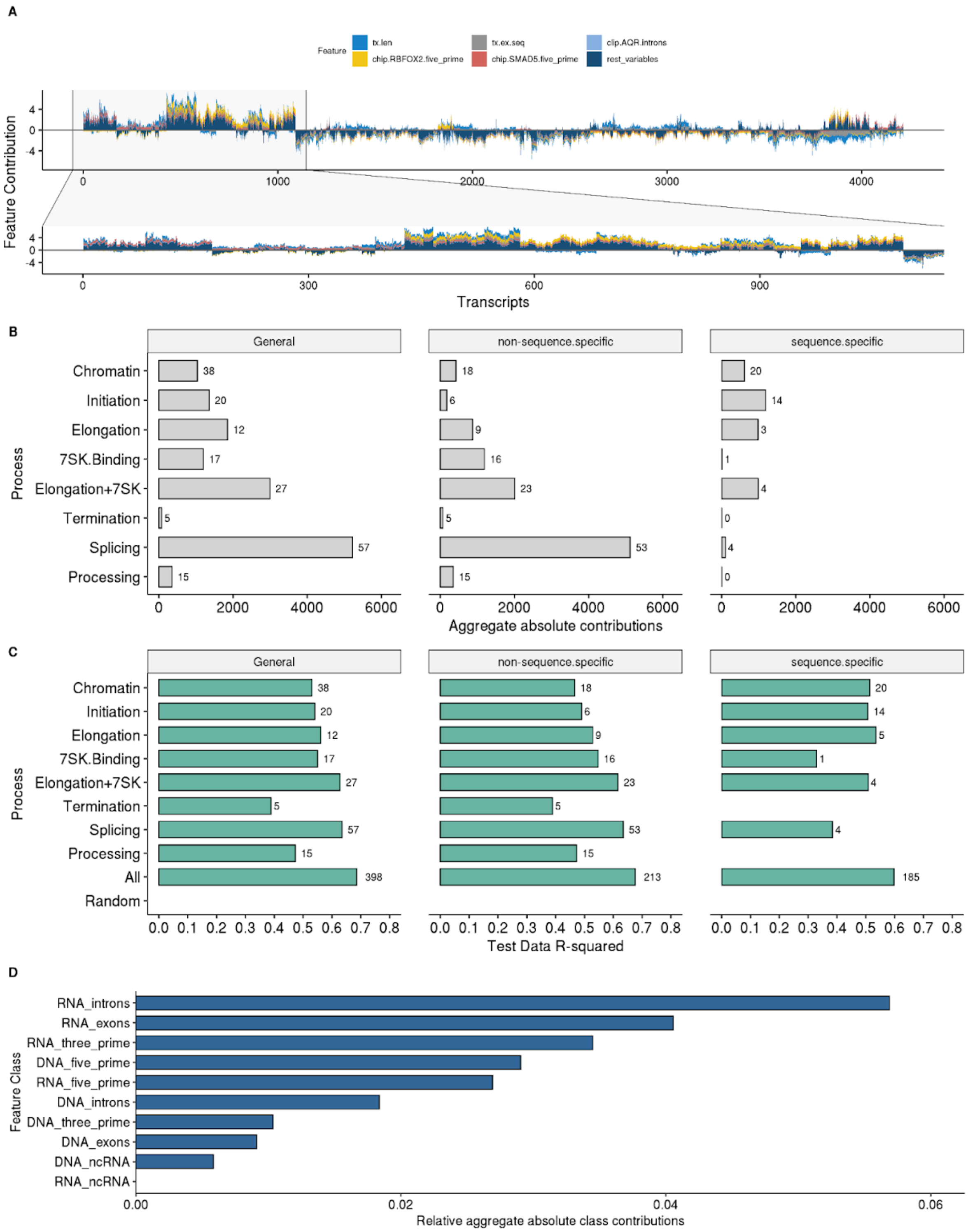


Figure 3.7: Interpretation of the transcriptional pausing model. (A) Distribution of SHAP feature contributions (y-axis; only top 5 individual features colored and remaining features aggregated in *rest_variables*) on each gene transcript (x-axis) with a zoom-in on a subset of transcripts for better visual investigation **(B)** Absolute total factor class contributions based on prior knowledge, subdivided by sequence and non-sequence specific binding factors. Class *Processing* refers to mRNA polyadenylation and export from the nucleus. **(C)** R^2 model performances of individual models of functional factor classes based on prior knowledge evaluated on a 50% holdout test data set from the same cell line (K562) **(D)** Absolute total factor

contributions based on their binding modes.

This ranking (Fig. 3.7B) revealed the high importance of splicing factors in transcriptional pausing, followed by elongation and 7SK binding proteins. These findings strongly support the interconnection of co-transcriptional splicing events (115, 328, 340) and provide quantitative measures about the relevance of the newly identified 7SK binding proteins as transcriptional pause regulatory factors. On the other hand, the *Elongation* factor set of established pausing factors validated our approach.

After having quantified the relative importance of major RNA processing events in the context of transcriptional pausing, we asked how models exclusively trained on features defined by the different sets of regulators would perform individually. As a baseline comparison of model performances, we have also built randomized models with randomized input data (see Materials & Methods). The model performances (R^2 values) of cross-validated models in the K562 cell line on the independent 50% holdout test data sets (see also **Supplementary Table S20** for all model results) for each feature space is given in figure 3.7C. Each feature space performs reasonably well relative to the number of factors they incorporate. For instance, the splicing factor-based model (*Splicing*) considers only n=57 of all available factors (n=398) yet performs almost equally well as the full model (*All*) with all n=398 available factors. Similarly, compared to the chromatin-associated model *Chromatin with n = 38* factors, the *Initiation* model incorporates only about half the number of factors (n=20) yet performs slightly better (R^2 of 0.54 vs. 0.53).

Despite the low numbers of factors considered in the 7SK (*7SK.Binding*) and established pausing factor (*Elongation*) models, these sets perform very well, which underlines the predictive value of implicated factors. Their predictive power is further demonstrated by the set of the union of 7SK and established elongation factors (*Elongation+7SK*) which not only outperform (R^2 0.62) each individual factor set alone (*7SK.Binding*: R^2 0.55, *Elongation*: R^2 0.56) but also perform almost equally well as the full model (R^2 0.62 vs. 0.68). This finding further highlights the importance of identified novel 7SK binders as potential pause regulators. To summarize, each factor set shows high predictive power relative to the number of factors they consider. However, due to the different amount of factors considered in each model, their performances should not be compared to each other. On the other hand, we can conclude that their high predictive value demonstrates the intricate interconnection of underlying processes with the transcriptional pausing outcome. Moreover, the role of the 7SK ncRNA with associated factors as a transcriptional pause mediator complex is supported and strengthened, which allowed us to suggest the factors contained in the 7SK factor set (*7SK.Binding*) (see **Supplementary Tables S9 & S10**) as additional 7SK ncRNA binding proteins to be implicated in transcriptional pause regulation based on their predictive value. The results for the HepG2 cell line are highly similar and support the aforementioned conclusions (see **Supplementary Figure S3.3)**.

In the next step, we asked whether genomic or transcriptomic binding events are primarily responsible for the observed predictive power of the models. Upon investigation, we could establish that RNA binding events had higher contributions than DNA binding events (Fig. 3.7D). Interestingly, splicing factors are enriched in RNA intron binding sites (Fisher's exact test, one-sided (greater), p = 0.034, odds ratio 4.45, confidence interval [1.11;Inf] in K562 and p=0.032, odds ratio 7.1 [1.15;Inf] in HepG2), which is the highest-ranked functional class (see Fig. 3.8 for K562 and **Supplementary Figure S3.4** for HepG2 data).

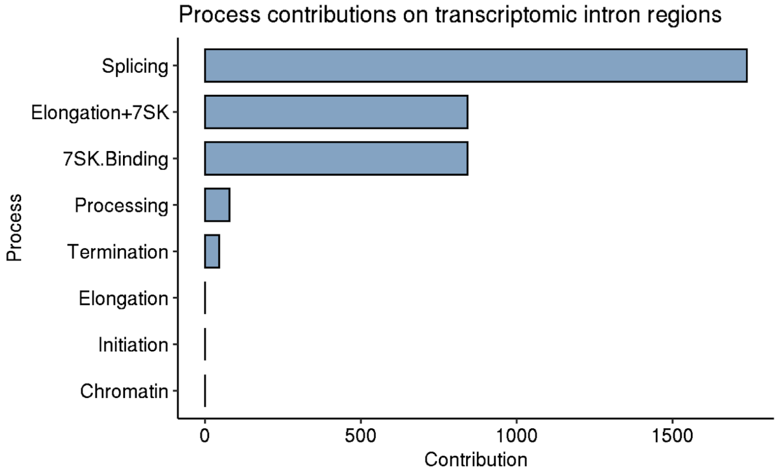


Figure 3.8: Model contributions of regulatory processes. Aggregate feature contributions (x-axis) of RNA intron binding factors by functional classes (y-axis) in the K562 cell line.

Lastly, the high contribution score of genomic binding events on the 5' region of transcripts (Fig 3.7D, DNA_five_prime) is in accordance with observed genomic five prime modulated transcriptional pause states from the literature (341).

Gene annotation and sequence composition features account for 26% of all feature contributions (see **Supplementary Figures S3.5-3.8**). However, due to their static nature, they cannot explain the variation of transcriptional pausing between cell lines, which is why we have focused the discussion on the individual proteins and their dynamic binding events.

3.2.3. Modulators of Transcriptional Pausing

We next aimed to identify specific pause regulatory factors based on the model feature contribution scores of protein binding events. We thus ranked individual DNA- and RNA-binding factors by aggregating SHAP feature contributions per factor over all genes into a single contribution score (see Fig. 3.9 for K562 and **Supplementary Figure S3.9** for HepG2 data)

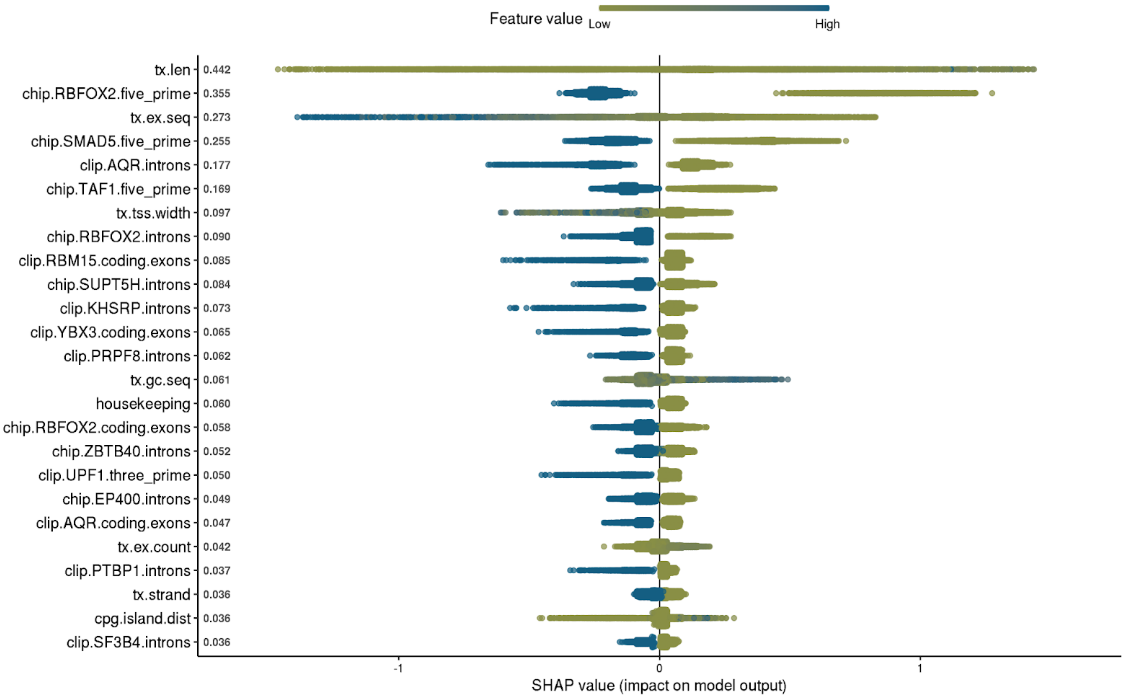


Figure 3.9: Individual model feature contributions. Feature contribution (x-axis) of the top 25 features (y-axis) from the full (*All)* individual K562 model.

Subsequently, selected the minimal set of most influential factors (16 out of 398) that together account for up to 50% of all feature contributions (Fig. 3.10A). In this ranking established pausing factors from the literature (Fig. 3.10A, highlighted in red) ranked high among these top influential factors and thus served as a validation of our factor ranking approach. Interestingly, three factors that are not primarily related to pausing were ranked higher than the established pausing factors. These might represent novel pause regulatory factors with at least the effect size of established pausing factors. However, all other lower-ranking factors can be considered almost equally relevant since they have similarly high contributions.

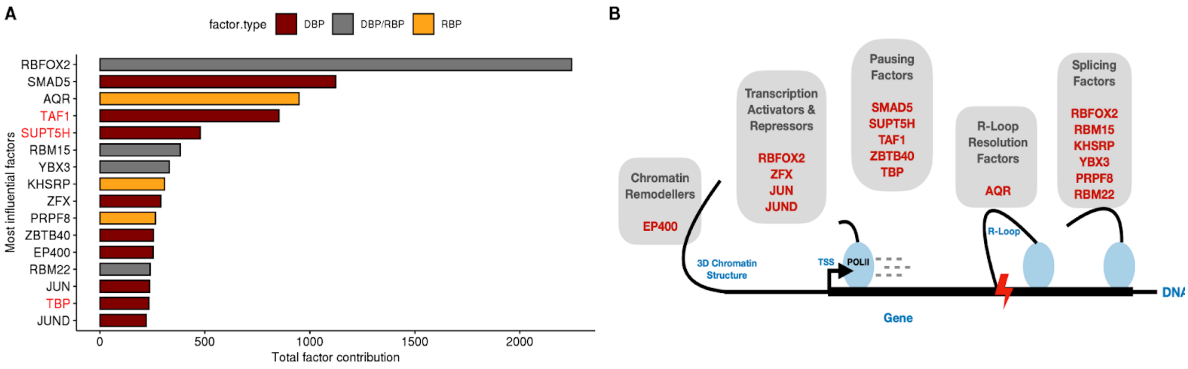


Figure 3.10: Feature contributions of the minimal model of transcriptional pausing. (A) Factor contributions of factors that together make up at least 50% of total model contributions (increasingly ordered). Established pausing/elongation factors are colored in red. The bar fill colors indicate DNA-binding (DBP; dark red), RNA-binding (RBP; orange), or DNA- and RNA-binding (DBP/RBP; grey) factors. **(B)** Functional associations of identified factors.

Strikingly, a model based on binding features of only these 16 most influential factors (including gene annotation and sequence composition features and only the five known pausing or 7SK-related factors AQR, BRD4, SUPT5H, TAF1, TBP) achieves an R^2 of 0.65 (on 50% holdout test data set; see Figure 3.11 for K562 and **Supplementary Figure S3.10** for HepG2 data performances of minimal models per cell line) and compete with the full model which incorporates all 398 factors and achieves an R^2 of 0.68.

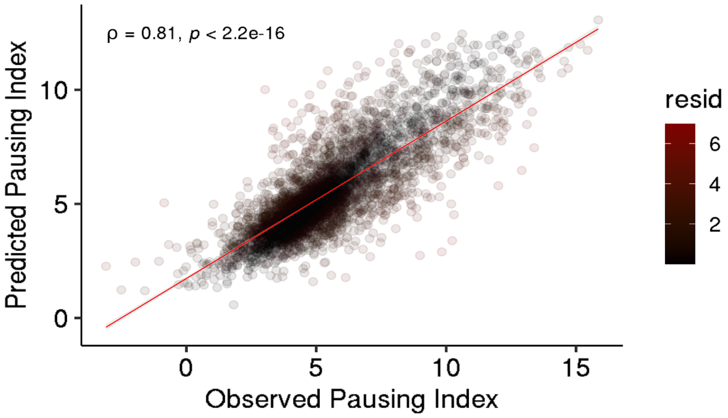


Figure 3.11: Minimal model performance in the K562 cell line. Observed (x-axis) vs predicted (y-axis) pausing index of the 16 most influential factor model.

This minimal model with n=16 factors also outperforms the *Elongation+7SK* model (Fig. 3.7C) with n=27 factors (R^2 of 0.65 as compared to an R^2 of 0.61). This indicated that the *Elongation+7SK* was incomplete and lacked pausing-related factors. Lastly, The HepG2-based minimal model with n=9 factors consisting of the factors RBFOX2, AQR, TAF1, TBP, RBM15, RBM22 KHSRP, PRPF8 and YBX3, are all included in the minimal model identified in the K562 cell line.

Upon consideration of the functional background of the most influential factor set (n=16, K562) (Fig. 3.10B) the interconnected nature of transcriptional pausing with other RNA-processing events becomes further evident.

3.2.4. Pausing factors

A few well-established pausing factors like TAF1, TBP, and SUPT5H ranked high in our models. Especially TAF1 and TBP validate our approach as these are components of the pre-initiation complex (PIC) whose formation inherently leads to pausing (87). This, in turn, can be regulated by other pausing factors like NELF and DSIF (SUPT5H), which increase pausing, whereas the P-TEFb associates with pause release and decreases pausing.

3.2.5. Chromatin remodelers

EP400 is a chromatin remodeler that greatly impacted our model. It modulates chromatin state by nucleosome positioning and posttranslational modifications of histones. The chromatin state is regulated by chromatin remodelers and intricately linked to transcription initiation, elongation, and co-transcriptional splicing (81, 110–112). Specifically, EP400 acts as a histone acetyltransferase and deposits H3.3/H2.AZ into promoters and enhancers after PIC assembly, thereby exerting gene activating functions (342). Moreover, it associates and interacts with the well-known pausing factor MYC (102, 342, 343). This association might be a direct link to transcriptional pausing. In fact, regulation of Pol II pausing at promoter-proximal nucleosomes by chromatin remodelers like for instance, Chd1 (113), are already established.

3.2.6. Transcriptional repressors and activators

Among the established pausing and chromatin remodeling factors, we can find transcriptional repressors and activators like ZFX, JUN, JUND, or RBFOX2.

ZFX family members are active in multiple types of human tumors, bind downstream from the TSS at the majority of CpG island promoters, regulate essential housekeeping genes and exert a transcription activating function in general. They act similarly to the MYC family of transcription factors characterized by their pervasive binding at promoter sites and profound proliferation defects upon knockdown (344, 345). MYC can recruit P-TEFb (102, 346) and thus plays an essential role in transcriptional pause release. A similar connection with pausing could exist for ZFX given their similar functional behavior. A transcriptional pause regulatory role is further supported by the observation that ZFX binds slightly downstream from the most frequent Pol II pause site and slightly upstream of the downstream peak of H3K4me3 signal (344, 345)

JUN and JUND are subcomponents of the activating protein 1 (AP-1) (347, 348). AP-1 is responsible for cell proliferation, neoplastic transformation, apoptosis as well as the expression of immune mediators. It is targeted and suppressed by the negative elongation factor NELF (349). However, so far no transcriptional pause regulatory role has been reported yet.

RBFOX2 can act as a regulator of alternative splicing as discussed later, as well as a transcriptional repressor through the recruitment of the repressive polycomb-complex 2 (PRC2) to its site of action (344, 350, 351). It primarily targets chromatin-associated RNA, especially promoter-proximal nascent RNA and might be more intricately linked to transcriptional pausing. In fact, *RBFOX2* knockout cardiomyocytes show decreased pausing indices and coordinated transcriptional pause enhancing roles of RBFOX2 and PRC2 at gene promoters (351).

3.2.7. Co-Transcriptional splicing and mRNA regulatory factors

The intricate connection of transcriptional pausing to co-transcriptional splicing events (329, 330, 352, 353) is supported by the presence of several splicing-associated factors (RBFOX2, PRPF8, RBM15, RBM22, KHSRP, YBX3, AQR) among the top regulators. Co-transcriptional splicing of pre-mRNAs is directly coupled to the nascent RNA that forms during the transcriptional cycle, which, in turn, is subject to transcriptional pausing. Thus transcriptional pausing is a rate-limiting step to co-transcriptional splicing. Indeed it has been shown that active spliceosomes are bound and complexed to the Pol II S5P C-terminal-domain during transcript elongation as well as co-transcriptional splicing (114) and that transcription kinetics strongly impact splicing decisions. For instance, slow Pol II elongation rates buy time for the spliceosome to assemble and favor splicing. More strikingly, it has been shown that inhibiting the spliceosomal U2 snRNP function would lead to Pol II pausing in promoter-proximal regions, impairing the recruitment of P-TEFb, reducing the Pol II elongation velocity at the beginning of genes (115). These observations point towards the existence of positive feedback from the splicing to the transcription machinery since the release of paused Pol II requires the formation of functional spliceosomes. To that end, RBFOX2 is a well-established regulator of alternative splicing (354–356) with an essential role in transcriptional pausing (351). Similarly, the pre-mRNA splicing factors or spliceosome components RBM15 (357), RBM22 (358, 359), PRPF8 (360), KHSRP (361) and YBX3 (362) are likely linked to pausing, as is the case for RBFOX2 and splicing in general.

AQR is an R-loop resolution factor (363). When nascent RNA anneals back to template DNA (364–367), R-loops are formed, forming an RNA/DNA hybrid structure. They have been suggested to likely be part of the mechanism for Pol II pausing (366), with the goal to hold back Pol II elongation (368) and the DNA replisome (369). However, R-loops may also form as a result of splicing defects which further highlights the importance of splicing events during transcriptional pausing since the lack of splicing-dependent nascent RNA processing leads to an increased formation of R-loops which would have otherwise be prevented through timely splicing events.

3.2.8. Novel pausing factors

ZBTB40 and SMAD5 have not been previously associated with the regulation of transcriptional pausing. Thus, we suggest a novel link. ZBTB40 is a regulator of osteoblast activity and bone mass (370). SMAD5 from the SMAD family of proteins acts as a signal transducer. It is activated in the cytoplasm and accumulated in the nucleus where it modulates transcription and potentially exerts a pause regulatory via chromatin remodeling events through the recruitment of a diverse set of coactivators and corepressors (344, 350).

3.3. Discussion

Promoter-proximal Pol II pause regulatory elements play an essential role in transcriptional regulation. Their identification and characterization are crucial to deciphering gene regulatory mechanisms that maintain cell homeostasis and enable cell plasticity. We improved our understanding of pause regulatory elements with machine learning models that can predict the extent of proximal promoter pausing from large-scale features of genomic and transcriptomic protein binding maps and gene annotation and sequence composition features. These models provide novel insights into cis- and trans-regulatory elements underlying transcriptional pausing. In particular, our model achieved a high prediction performance (R^2 ~ 0.68 with n=389, factors; R^2 ~ 0.65 with only n=16 factors), which is indicative of features that can explain a large part of the variance observed in transcriptional pausing. The generalizability of the model is demonstrated by its high prediction performance on cross cell type-specific binding data (R^2 ~ 0.52), indicative of rules learned that are general and cell type unspecific, which is in accordance with the observation that pausing of genes is prevalent across a large fraction of cell types (309). Additional models trained on subsets of proteins implicated in either chromatin remodeling, transcription initiation, elongation, splicing, and further downstream transcript processing showed high predictive power and strongly supported the intricate connection of these processes (3, 297, 310, 324, 325, 328, 371, 372) with transcriptional pausing. Strikingly, splicing factors have the highest predictive power for pausing, which is in agreement with many studies that propose and show dual roles for individual proteins as is the case, for instance, with RBFOX2 (354–356), SRSF2 (330), U2AF65 (115) or MAGOH (115) providing a direct causal link between the two processes of splicing and transcriptional pausing. A major goal of our analysis was the identification of novel regulators of transcriptional pausing. To that end, we used two approaches, in which we first identified novel 7SK binding RBPs and subsequently showed their high predictive power for pausing. Secondly, we investigated the feature contributions in our model to extract protein factors with higher contribution scores than established pausing factors. Many of these factors like RBOFX2 (354–356), AQR (363), JUN, and JUND (347) have been shown to affect transcriptional pausing or to be implicated in certain processes associated with pausing. Thus these factors and their functional background provide some initial mechanistic hypotheses and represent attractive targets for experimental validation.

We have chosen to obtain and analyze data for the HepG2 and K562 cell lines from the ENCODE project. They have been extensively characterized with multiple types of assays. In particular, it provides an unparalleled number of genomic and transcriptomic binding maps and thereby enables the identification of novel regulators of promoter-proximal pausing. However, a limitation is that not all previously characterized regulators of pausing are available. Moreover, to quantify proximal promoter pausing, only GRO-seq data or similar variations are available for these two cell lines. On the other hand, multi-omics approaches like TT-seq (373) or mNET-seq (297, 374, 375) estimate the kinetic rates of initiation and pause duration more accurately yet are only available for the K562 and Raji B cell lines. Nonetheless, these provide at least ground for future studies of transcriptional pausing with enhanced precision. Once they are available across cell lines, elaborate validation procedures will also be available. Unfortunately, these data types are not available for a second ENCODE cell line which would not make cross-validation possible. Taken together, we provide a framework to foster our understanding of transcriptional regulation from the perspective of the critical early steps in transcriptional elongation. At the same time, we expect improvements with more accurate kinetic profiling of the polymerase, broader availability of protein binding maps, and improved binding site prediction from DNA sequence alone.

4. Trans-epistasis underlying Coronary Artery Disease confers differential disease risk and perturbs gene expressions in trans

This chapter covers the contents of our second research project concerned the identification of trans genetic interactions as upstream trans-regulators of downstream trans target genes that underlie Coronary Artery Disease (review Section 1.1. Thesis Aims). The contents are entirely based on my manuscript in preparation (June 2022) and essentially replicate it, including the figures.

To briefly recap our aims, Coronary Artery Disease (CAD) is a cardiovascular atherosclerotic inflammatory disease characterized by occlusions of the coronary arteries. It is one of the leading causes of death worldwide (5), with environmental and genetic risk factors contributing to the disease etiology of CAD (33, 376–378). Genome-wide-association studies (GWAS) (17–28) identified numerous genetic risk factors with an overall heritability of CAD estimated as 40-60% (33), of which up to 38% is collectively explained by loci identified through GWAS (379), yet, only 4% of the variance is explained in an independent test population by additive polygenic risk scores that incorporate such loci (34). This indicates that multiplicative effects between different genetic factors (epistasis) or genetic and environmental factors might explain the remaining proportion of heritability (380). Indeed, numerous epistatic interactions have been already identified in model organisms (161, 162) and even in humans to modulate gene expression levels (381, 382) which support the hypothesis that epistatic interactions may also contribute to complex diseases such as CAD. Because epistasis changes the linkage disequilibrium (LD) between pairs of loci (383), differences in LD between case and control groups can be used as a means to prioritize SNP pairs for epistasis testing (384), thereby reducing the set of possible SNP pairs and overcoming the multiple testing burden. This approach is favored over a previous approach on a Type 2 Diabetes data set (385) but has not yet been applied to study CAD.

Therefore, in this project, we aimed to identify epistatic interactions between previously identified GWAS loci (379) for CAD by applying an LD filter to avoid the multiple testing burden. Our approach integrates two independent cohorts to separate discovery and replication analysis. We preselected pairs of loci based on the LD difference between cases and controls in the UK Biobank (UKBB) cohort and subsequently subjected these pairs of loci to statistical testing for epistasis with a permutation testing procedure to nominate SNP pairs for replication analysis. Replication analysis included >35.000 case-control samples, and we were able to identify and replicate n=4 interacting pairs of SNPs for which we further characterized their effects on gene expression through genotype-combination dependent differential gene expression analysis, which revealed n=2 differentially regulated trans target genes that could be independently replicated with regards to the SNP pair genotype combination, the direction of effects, and more importantly, in the same or a highly related tissue type.

4.1. Materials & Methods

This project used multiple large-scale data sets to study Coronary Artery Disease. Genotype data from the UK Biobank data served to discover SNP interactions. The UK Biobank (UKBB) (386) is a large-scale database of genetic and health information from half a million volunteers from the United Kingdom aged between 40 and 69 years. The UKBB is a large and rich data set providing data on blood, urine, and saliva samples, health-related records, as well as detailed information about people's lifestyles. This allows linking lifestyle factors with disease states, enabling a deeper understanding of how individuals experience diseases and their underlying causes. Moreover, volunteers will be followed for at least 30 years after enrollment, enabling disease progression analysis with time-resolved health records.

For validation purposes, we have integrated and aggregated multiple data sets of genome-wide associations studies for CAD namely from the German Myocardial Infarction Family Studies (GerMIFS) I (387), II (388), III (389), IV (379), V (390), VI (391), VII (392), the LUdwigshafen RIsk and Cardiovascular Health Study (LURIC) (393), Cardiogenics (CG) (394), Wellcome Trust Case Control Consortium (WTCCC) (395), and Myocardial Infarction Genetics Consortium (MIGEN) (396).

To enable downstream differential gene expression analyses, we have integrated data from the *Stockholm-Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET) study* (397, 398), which is concerned with the investigation of living patients with cardiovascular diseases. Samples of multiple disease-relevant tissues have been taken during open thorax surgery of 600 coronary artery disease (CAD) patients, including from blood, atherosclerotic-lesion-free internal mammary artery (MAM), atherosclerotic aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM) and liver (LIV). An inclusion criterion for the study was the patient's eligibility for coronary artery by-pass graft (CABG) surgery, while exclusion criteria were other severe systemic diseases, such as active systemic inflammatory diseases or cancer. The study includes patients of Caucasian origins, mainly Northern European (Finnish) descent, with 31% being female. In addition, 32% had diabetes, 75% had hypertension, 67% had hyperlipidemia, and 33% had myocardial infarction before age 60. The study has 566 genotypes and 3577 RNA-seq profiles from previously mentioned tissues for 600 patients. Genotypes were imputed to a total of 14,098,063 DNA variant calls (6,245,505 with minor allele frequency (MAF) >5%). RNA sequencing was performed with poly-A (LIV, SKLM, VAF, SF, and blood) and ribo-zero (AOR, MAM) protocols with 50-100 bp read lengths, with single-end sequencing and a read depth of 15-30 million reads. The STARNET study served to identify genotype-combination-dependent differential gene expression profiles in CAD patients.

To validate differential gene expression analyses, we have further integrated data from the *Genotype-Tissue Expression (GTEx) project (399)*. The *GTEx project* provides resources to study human tissue-specific gene expression and its relationship to genetic variation. Genetic variations can be linked to differential gene expression patterns and identified as expression quantitative trait loci (eQTLs). GTEx enables this by collecting multiple human tissues and quantifying RNA types along with dense genotyping of the individuals to assess genetic variation within their genomes and ultimately link them to gene expression patterns. This fosters a comprehensive understanding of the mechanism of gene regulation in dependence on genetic variation, which then can be used to understand diseases. In addition to the publicly available RNA-seq and genotype data sets, their database allows investigators to view and obtain precalculated cis- and trans-eQTL for all tissues, eQTL associated with diseases, allele-specific expressions as well as tissue-specific alternative splicing information. The latest version (v8, 2021) of GTEx provides gene expression data of 17382 samples from 948 donors with age groups ranging from 20-70 years across 54 human tissues and genotype data of 15253 samples from 838 donors also across 54 tissues. The GTEx cohort served to validate genotype-combination-dependent differential gene expression profiles identified in the STARNET cohort.

In the following subsections from 4.1.1. Integration of Genotype and Phenotype Data to 4.1.6. Identification of Epistatic Effects on Gene Expression in Trans we will cover the technical aspects in analyzing these data sets for identifying epistatic interactions underlying Coronary Artery Disease.

4.1.1. Integration of Genotype and Phenotype Data

For identifying potential SNP interactions underlying CAD, we obtained genotype data from the UK Biobank project (UKBB; Project ID 25214; n=~500,000 samples) (386). Samples were then filtered for phenotypes of the circulatory system of ischemic heart diseases with ICD-10 codes I20, I21, I22, I23, I24, and I25, corresponding to phenotypes of unstable angina, acute myocardial infarction, subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction, certain current complications following ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction (within the 28 day period), other acute ischemic heart diseases and chronic ischemic heart disease, respectively. We thereby obtained 36191 diseased and 451218 healthy samples. This cohort served as the discovery data set for potential genetic interactions.

As a replication cohort of potential SNP interaction, we aggregated individual-level genotypes of 11 CAD case-control studies, namely the German Myocardial Infarction Family Studies (GerMIFS) I (387), II (388), III (389), IV (379), V (390), VI (391), VII (392), the LUdwigshafen RIsk and Cardiovascular Health Study (LURIC) (393), Cardiogenics (CG) (394), Wellcome Trust Case Control Consortium (WTCCC) (395), and Myocardial Infarction Genetics Consortium (MIGEN) (396). We refer to this aggregate data set as the GWAS cohort. It consists of 17584 diseased and 18157 healthy samples and served as a replication/validation cohort of potential SNP interactions identified in the UK Biobank cohort. Expected genotype frequencies calculated as two times the allele frequency based on genotype data from the 1000 genomes project (400) served to replace missing genotype data. Missing genotypes were imputed based on haplotypes obtained from the 1000 Genomes (400) data.

4.1.2. Integration of Quantitative Trait Loci for Coronary Artery Disease

We have integrated CAD-associated quantitative trait loci (SNPs) identified in a 1000 genomes-based genome-wide meta-analysis of coronary artery disease (379) to enable epistasis discovery for CAD. This set consisted of n=202 variants, of which n=157 had a valid rs-identifier and were measured in the aforementioned GWAS cohort. This set of n=157 CAD-associated SNPs (see **Supplementary Table 4.1** for SNP meta-data) served to discover SNP-SNP interactions in the UK Biobank cohort.

4.1.3. Identification of Candidate Epistatic Interactions

We identified SNP interactions with a randomization test based on differential SNP correlations between cases and controls. To ensure that the sample size-dependent calculation of SNP correlations is not inflated in any of the subgroups of either cases or controls, we downsampled the controls in the UKBB data set to match the ratio of cases and controls observed in the GWAS data set. This enables the comparison of SNP correlation between cohorts. We downsampled the majority class within a cohort to obtain comparable correlation coefficients between cases and controls. Because correlation coefficients might be affected by differential CAD risk allele counts, we stratified the samples into groups with the same number of risk alleles. A group size constraint of a minimum of n=15 samples was chosen, and samples at the lower and higher end of the distribution of the risk allele counts were successively merged until the group size constraint was met. Therefore samples at the extremes of the risk allele count distribution had to be assigned to groups with differential amounts of risk alleles. Subsequently, we calculated the Spearman SNP correlations within cases and controls separately. The SNP correlations were then aggregated per SNP pair, per subpopulation of cases and controls, and over the risk allele groups weighted by the size of the groups. This ensures that the calculation of SNP correlations is not biased by differential risk allele counts. The following specifies the computations made:

$$C_{i,j,k} = \sum_{l=1}^{n} \frac{\rho_{i,j,k,l} \, |m_l|}{N} \text{, where}$$

C is the correlation of SNPs i, j with $i \neq j$ in the k-th sub-population with $k \in \{cases;\ controls\}$ obtained through the summation of Spearman SNP correlations $\rho_{i,j,k,l}$ of $l \in \{1 .. n\}$ groups of samples with the same number of risk alleles, weighted by the individual group sizes $|m_l|$ and normalized by the total size of all subgroups n as N.

A randomization test was then performed in which the absolute differences of squared SNP correlations between cases and controls served as a test statistic, i.e

$$S_{i,j} = abs(C^2_{i,j,cases} \quad - C^2_{i,j,controls} \quad) \text{, where}$$

$S_{i,j}$ gives the test statistic for SNPs i, j with $i \neq j$. The disease label (case-control status) was permuted 1e4 times, and the empirical null distribution of the test statistic was obtained by calculating the statistic S on the permuted data. To obtain empirical p-values of observing a test statistic on permuted data at least as extreme as the observed test statistics on non-permuted data, we compared the observed test statistic S based on non-permuted data against the null distribution of the statistic S obtained by the permutation testing on the permuted data. Haplotype effects and cis-epistasis were avoided by only considering inter-chromosomal (trans-acting) SNP pairs (n=11447; see **Supplementary Table 4.2** for the list of SNP pairs). For some SNP pairs, the correlation statistic could not be calculated due to insufficient variance in an SNP genotype and thus was excluded (n=2252). Following the permutation test, we selected the top 1% SNP pairs (n=92: see **Supplementary Table 4.3**) with the lowest empirical p-values and subjected them to validation with linear models (see Section 4.1.4. Discovery and Replication of Epistatic Interactions)

4.1.4. Discovery and Replication of Epistatic Interactions

To validate the SNP interactions (n=92) identified in the permutation testing procedure (see Section 4.1.3. Identification of Candidate Epistatic Interactions), we performed logistic regression analysis on the UKBB data set. The CAD case-control status as the target variable was modeled in dependence of a multiplicative term for SNP-SNP genotype interactions, the terms for the genotypes of each of the individual interacting SNPs and sex as an additive covariate:

$$logit(P(CAD|data)) = \beta_0 + \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_3 SNP_1 * SNP_2 + \beta_4 SEX.$$

Based on FDR adjusted p-values obtained from the models ($H_0 : \beta_3 = 0$; FDR < 5%) we selected the set of validated discovery SNP-SNP interactions. However, we further subjected this validated discovery set of SNP-SNP interactions identified in UKBB to replication in the GWAS cohort. Replication was performed by testing the identical model on the GWAS replication cohort and selecting SNP pairs with nominal $p < 0.05$ and the same direction of effects and similar effect sizes as observed in the UKBB cohort.

4.1.5. Integration of Gene Expression Data

We integrated gene expression and genotype data from the Stockholm-Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET) study (397, 398) as well as the Genotype-Tissue Expression project (GTEx v8)(399) to enable the investigation of the cis- and trans-regulatory effects of the identified SNP interaction on the expression of genes. The STARNET data served as the discovery cohort and the GTEx data as the replication cohort of potential cis- or trans-regulated genes. Gene expression profiles were obtained for multiple tissues (see **Supplementary Table 4.4** for the number of genes and samples per tissue per cohort), and non-coding RNA biotypes were included. Gene read counts were normalized per fragment per kilobase per million (FPKMs), and genes not expressed in more than 90% of the samples were excluded. The effect of SNPs identified in the interaction analysis on the closest gene at the SNP locus (cis-eQTL effects) was evaluated with linear regression models, in which the mapped genes normalized expression levels served as the targets and the individual SNPs as predictors.

4.1.6. Identification of Epistatic Effects on Gene Expression in Trans

To unveil potential downstream epistatic effects on the transcriptional landscape of genes, we evaluated genotype-combination dependent differential gene expressions for all genes localized in trans of each interacting SNP pair. Specifically, for each possible genotype combination at the two SNPs of an interacting SNP pair, we assessed the significance of the differences in observed expression levels between samples that carry a specific genotype combination against samples that do not carry the same genotype-combination. Thus, a binary carrier status variable served to distinguish between these cases, assigning a value of one for each sample carrying a specific genotype-combinations or zero otherwise. We then conducted linear regression analysis in each tissue of the STARNET cohort, predicting individual gene expressions from the carrier status:

$$Trans\ Gene_{log10(FPKM)} = \beta_0 + \beta_1 IND_{i,j,k},$$

where $IND_{i,j,k}$ represents the binary carrier status variable that specifies which individuals carry a specific genotype-combination k between interacting SNPs i and j with $i \neq j$. Differentially expressed genes were selected based on whether the carrier status indicator variable showed a significant effect ($H_0 : \beta_1 = 0$, FDR < 5%) for a given SNP pair, genotype combination, and tissue. Validation of differential gene expressions was conducted in the GTEx v8 cohort, performing the same methodology except limiting the analysis on the STARNET discovery set of differentially expressed genes. The replication criterion was FDR < 5%. Certain genotype combinations of interacting SNP pairs have relatively small allele frequencies and lead to small sample counts, which can inflate the test statistics of the linear models. To bypass this risk imposed by these rare events, we determined the significance of the regression coefficients for the carrier indicator variable. This was accomplished with a permutation testing procedure comparing the absolute t-statistic (T) from a two-sided t-test between samples with and without a specific genotype combination against the expected distribution of the test statistic T. The expected (null) distribution of the test statistics was obtained by permuting the assignment of samples to genotypes $b = 1e^4$ times, inherently removing any relationship between gene expression levels and genotypes, and calculating the T-statistic b times on the permuted data. Evaluating how often we observe a permutation T-statistic that is at least as extreme as the observed T-statistic of the non-permuted data yields an empirical p-value that allows us to rule out individual differential trans gene expression results that are likely due to chance (FDR < 5%).

4.2. Results

4.2.1. Identification of Trans Epistasis in CAD

The integration of large-scale genotype and phenotype data (review Section 4.1.1. Integration of Genotype and Phenotype Data) from the UK Biobank as a discovery cohort as well as 11 additional genome-wide CAD case-control studies as a replication cohort together with n=157 quantitative trait loci for CAD from a large meta-analysis (379) (review Section 4.1.2. Integration of Quantitative Trait Loci for Coronary Artery Disease) enabled the identification of genetic interactions underlying CAD through a filter-based permutation testing approach coupled with linear modeling (Fig. 4.1).

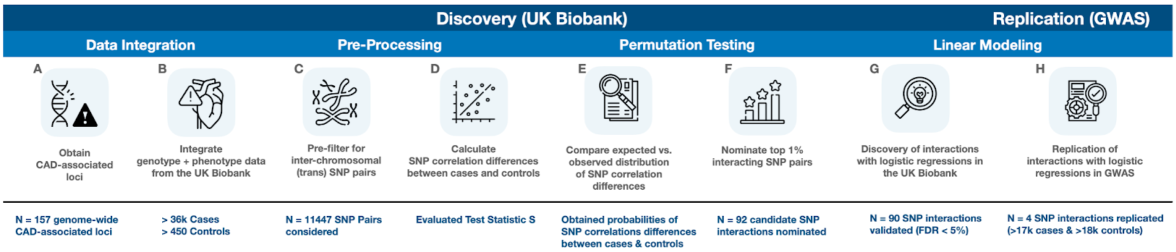

Discovery (UK Biobank)
Replication (GWAS)
Data Integration
Pre-Processing
Permutation Testing
Linear Modeling
A
Obtain CAD-associated loci
B
Integrate genotype + phenotype data from the UK Biobank
C
Pre-filter for inter-chromosomal (trans) SNP pairs
D
Calculate SNP correlation differences between cases and controls
E
Compare expected vs. observed distribution of SNP correlation differences
F
Nominate top 1% interacting SNP pairs
G
Discovery of interactions with logistic regressions in the UK Biobank
H
Replication of interactions with logistic regressions in GWAS
N = 157 genome-wide CAD-associated loci
> 36k Cases
> 450 Controls
N = 11447 SNP Pairs considered
Evaluated Test Statistic S
Obtained probabilities of SNP correlations differences between cases & controls
N = 92 candidate SNP interactions nominated
N = 90 SNP interactions validated (FDR < 5%)
N = 4 SNP interactions replicated (>17k cases & >18k controls)

Figure 4.1: Discovery and replication analysis workflow for the identification of trans genetic interactions underlying Coronary Artery Disease.

The CAD prevalence in the percentiles of a polygenic risk score based on the number of risk alleles per sample (Fig. 4.2 A) successfully separates cases and controls (Fig. 4.2 B) and revealed a non-linear increase in the disease prevalence (Fig. 4.2 C), indicative of multiplicative effects in addition to the additive genetic effects, supporting the idea of potential genetic interactions underlying CAD.

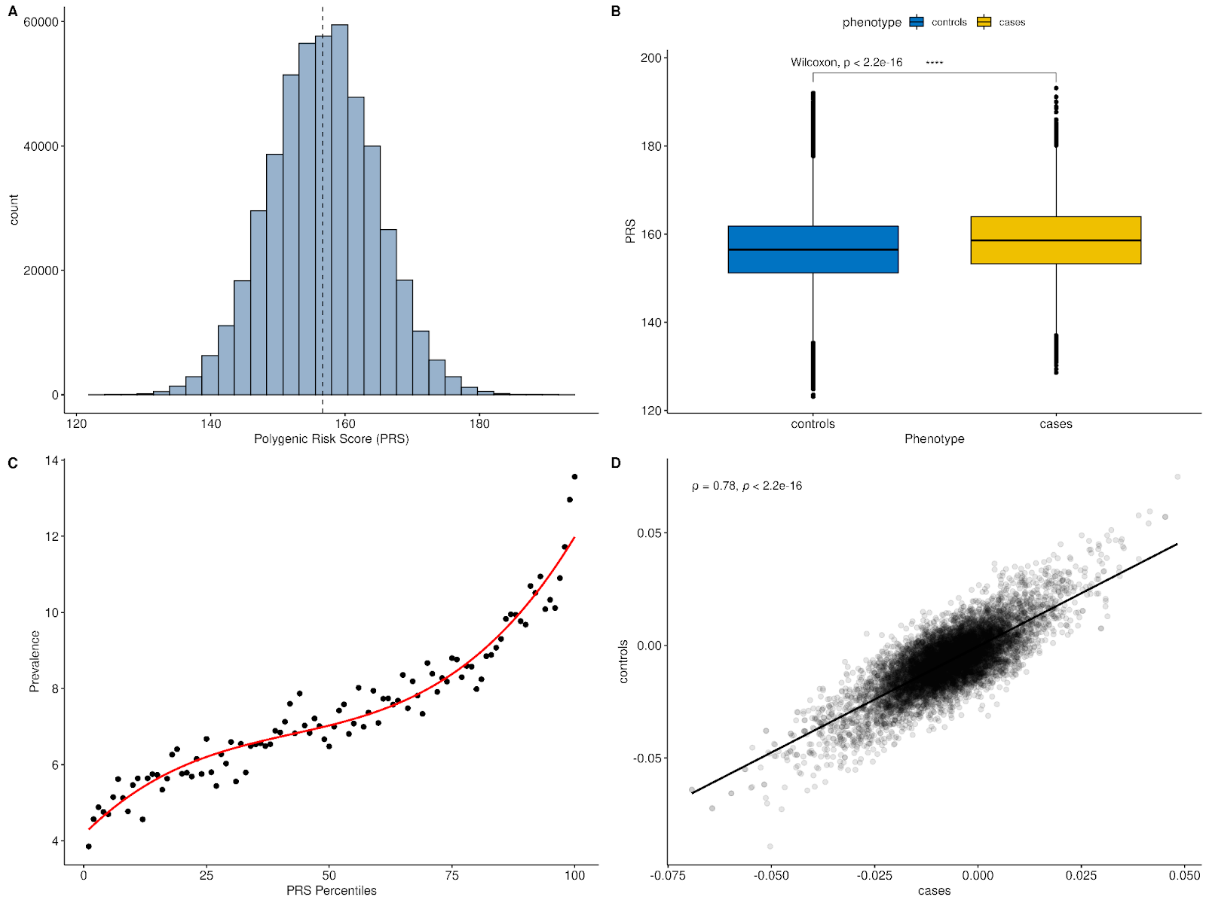


Figure 4.2: The polygenic risk score for CAD (UKBB cohort). (A) Distribution of the polygenic risk score distribution for CAD, quantifying disease risk with the number of CAD risk alleles carried. **(B)** Distribution of the polygenic risk score in cases and controls. **(C)** Prevalence of CAD cases in the percentiles of the polygenic risk score. **(D)** Interchromosomal SNP - SNP pair correlations between cases and controls.

An investigation into the differences of linkage disequilibria (LD) of interchromosomal SNP pairs between cases and controls revealed differential SNP correlation structures in which the SNP-SNP correlations are globally highly similar (Pearson correlation = 0.77), yet with differences observable for individual SNP pairs (Fig. 4.2 D) which have the potential to contribute to the disease etiology and further separate healthy from unhealthy samples. These trends can also be observed in the GWAS cohort (see **Supplementary Figure 4.1 A-D**).

SNP pairs were ranked and prioritized based on the statistical significance of these differences of LD correlations between cases and controls in the UK Biobank as the discovery cohort. The statistical significance was assessed by a randomization procedure (review Section 4.1.3. Identification of Candidate Epistatic Interactions; Fig. 1 D-F). The observed distribution of SNP correlation differences between cases and controls was compared against the expected distribution of SNP correlation differences between cases and controls based on permuted data. Subsequently, we selected the top 1% (n=92; see **Supplementary Table 4.3**) SNP pairs with the lowest empirical p-values obtained from the permutation test to prioritize SNP pairs. We then evaluated the interaction effects of these pre-filtered SNP pairs in multiplicative terms in logistic regression models on the CAD case-control status (review Section 4.1.4. Discovery and Replication of Epistatic Interactions). Among these potential interactors the majority (98%; n=90) showed significant interaction effects (FDR < 5%). These n=90 SNPs were further subjected to replication in the GWAS cohort (same direction of effect and similar effect size) with analogous models, which identified the SNP pairs *rs72685791 - rs12202017*, *rs10841443 - rs12899265*, *rs73222236 - rs11911017* and *rs4719608 - rs2487928* with replicated interaction effects on CAD (see **Table 4.1** for meta-information from logistic regressions). Due to the high agreement (100%) of effect directions and effect sizes (mean absolute difference of interaction beta-coefficients between the cohorts ~ 0.01) between the cohorts, as well as the initial validation, approaches with permutation testing procedures and the validation with FDR adjusted logistic regression results, we chose to not additionally control for false positives of interactions identified in the GWAS cohort.

SNP Pairs	**SNP1** β_1	**SNP1** se_1	**SNP1** p_1	**SNP2** β_2	**SNP2** se_2	**SNP2** p_2	**Int.** β_3	**Int.** se_3	**Int.** p_3	**Int.** p_3'
rs72685791 rs12202017	0.143 (0.147)	0.033 (0.046)	1.4e-5 (1.4e-3)	0.175 (0.170)	0.036 (0.048)	1.2e-6 (4.6e-4)	-0.073 (-0.056)	0.021 (0.028)	5.9e-4 (4.8e-2)	2.8e-3 -
rs73222236 rs11911017	0.071 (0.087)	0.013 (0.017)	1.2e-7 (1e-6)	0.045 (0.112)	0.020 (0.029)	2.3e-3 (1.7e-4)	-0.068 (-0.063)	0.020 (0.030)	8.9e-4 (3.3e-2)	4.5e-3 -
rs4719608 rs2487928	-0.020 (-0.027)	0.019 (0.026)	0.299 (0.301)	0.026 (0.016)	0.014 (0.019)	5.6e-2 (0.384)	0.049 (0.053)	0.017 (0.023)	4.7e-3 (2e-2)	7.4e-3 -
rs10841443 rs12899265	0.016 (-7.4e-5)	0.013 (0.018)	0.198 (0.996)	-0.105 (-0.085)	0.033 (0.057)	1.8e-3 (0.137)	0.071 (0.101)	0.023 (0.037)	2.1e-3 (7.3e-3)	2.9e-3 -

Table 4.1: Meta-information of SNP interactions from logistic regression models for the UKBB data set. Meta-information on the GWAS cohort from analogous models is given in brackets. Column “*SNP Pairs*” lists all validated and replicated interacting SNP pairs, “β_i” the beta-coefficients of each independent variable, “se_i” the standard error of the i-th β-coefficient estimate and “p_i” the nominal p-value of the i-th independent variable. *“SNP1”* and *“SNP2”* refer to the first and second mentioned SNP of a SNP pair given in the *“SNP Pair”* column , respectively. Column “Int.” refers to the interaction term of the SNPs given in the *“SNP Pair”* column. Column “Int. p_3' ” gives the FDR adjusted p-values (FDR < 5%) of the interactions in the UKBB data set. Meta-information about the intercept and the “Sex” covariate is not shown.

Literature research about the individual interacting SNPs (see **Table 4.2**), specifically the associations of the SNP-associated genes with CAD revealed multiple functional implications. For instance, variations in GUCY1A1 affect platelet aggregation and confer an increased risk for CAD (401). Strikingly, it is also associated with ischemic events after coronary intervention (402), large artery atherosclerotic stroke risk (403) as well as myocardial infarction (404). TARID is an antisense RNA of TCF21 and can induce the expression of TCF21 by inducing the demethylation of the TCF21 promotor (405). Interestingly, it is proposed that TCF21 acts as a master regulator of CAD-associated genes (406). Thus TARID influences expression levels of CAD target genes through the epigenetic regulation of TCF21. TARID is also involved in cell cycle pathways associated with coronary artery disease and has been shown to induce cell proliferation (358). IGF-1R is a receptor gene and binds IGF-1 with high affinity, which in turn is strongly associated with cardiovascular events (407, 408) and diseases (409), conferring protective effects (410–412). JCAD (Junctional protein associated with coronary artery disease or KIAA1462) is a novel CAD disease gene that advances atherosclerotic plaque formation(413). ITGB8 has been shown to be strongly associated with CAD severity in epicardial adipose tissue (414) and functions to activate TGF-β. TGF-β is strongly associated with advanced atherosclerosis and CAD or CVDs in large (415–418). No evidence for a link between MSL2, LINC00189, LINC02398, MACC1, and CAD or CVDs exists yet.

SNP	EA	NEA	EAF	p	Loc	Gene	Conseq.
rs72685791	G	A	0.176 (UKBB) 0.179 (GWAS)	7.7e-3 (UKBB) 1.2e-3 (GWAS)	4:156620217	GUCY1A1	Intron Variant
rs12202017	A	G	0.249 (UKBB) 0.228 (GWAS)	1.3e-7 (UKBB) 2.4e-6(GWAS)	6:134173151	LINC01312; TARID	Non Coding Transcript Variant; Intron Variant
rs73222236	G	A	0.442 (UKBB) 0.427 (GWAS)	9.6e-5 (UKBB) 5.6e-6 (GWAS)	3:135888642	MSL2	Intron Variant
rs11911017	T	G	0.482 (UKBB) 0.487 (GWAS)	9.2e-1 (UKBB) 4.2e-4 (GWAS)	21:30567941	LINC00189 BACH1	Intron Variant
rs10841443	G	C	0.275 (UKBB) 0.258 (GWAS)	1.4e-4 (UKBB) 1.4e-1(GWAS)	12:20220033	LINC02398	Intron Variant
rs12899265	C	T	0.490 (UKBB) 0.493 (GWAS)	3.6e-1 (UKBB) 1.6e-2(GWAS)	15:99219598	IGF1R	Intron Variant
rs4719608	A	G	0.467 (UKBB) 0.455 (GWAS)	1.6e-1 (UKBB) 1.0e-1(GWAS)	7:20292134	ITGB8; MACC1	Upstream Gene Variant
rs2487928	A	G	0.400 (UKBB) 0.392 (GWAS)	1.2e-6 (UKBB) 2.0e-3(GWAS)	10:30323892	JCAD	Intron Variant

Table 4.2: Meta-information about individual SNPs of interacting SNP pairs. The column "*EA*" gives the effect allele, "*NEA*" the non-effect allele, "*EAF*" the allele frequency of the effect allele, "*p*" the nominal p-value from a logistic regression model of the SNPs additive CAD association, "*Loc*" the position on the genome, "*Gene*" the associated gene(s) and "*Conseq*" the consequences of the variants.

The interaction effects of specific genotype combinations of identified SNP pairs can be seen by contrasting the observed frequencies of specific genotype combinations against expected genotype combination frequency estimates from the logistic regression models for CAD cases (Fig. 4.3).

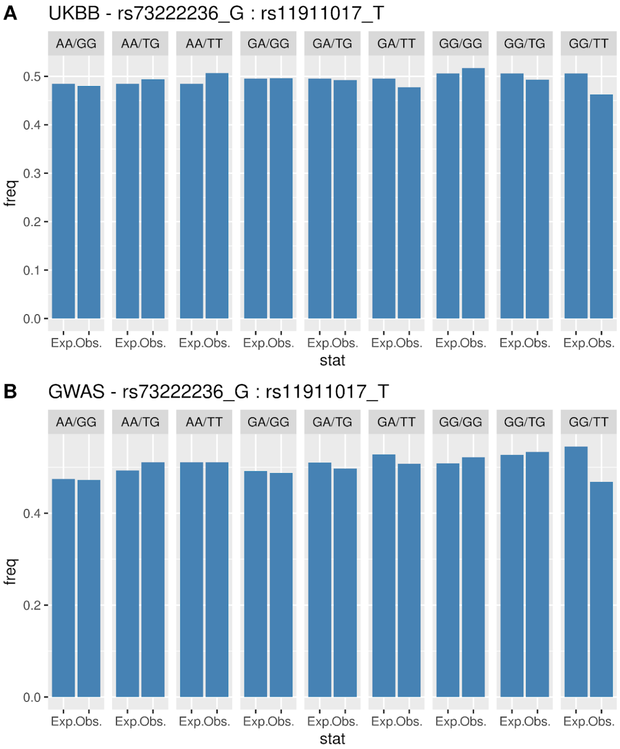


Figure 4.3: Proportions of CAD cases in each genotype combination as expected from an additive linear model (no interaction term) against the observed proportion of CAD cases.

For instance, the GG/TT combination of SNP pair rs73222236 - rs11911017 shows a clear difference between observed and expected case proportions in both cohorts and represents an SNP interaction as it leads to differential case proportions. Similar effects can be observed for the remaining SNP pairs (see **Supplementary Figures 4.2-4.4**).

4.2.2. Association of Gene Expression with interacting SNPs

Due to the localization of the majority (87,5%) of SNPs in non-protein-coding intronic sequences with the potential to alter gene regulatory elements (GREs) and the expression of genes in cis of the SNPs, we hypothesized that certain genotype combinations of interacting loci as identified earlier (Table 4.1) might additionally influence the expression of trans target genes through this differential regulation of SNP associated cis genes. To investigate genotype-combination-dependent differential trans gene expression levels, we integrated large-scale gene expression and genotype data for multiple tissues from the STARNET study and the GTEx project (review Section 4.1.1. Integration of Genotype and Phenotype Data).

The differential trans gene (>1Mbp) expression analysis was conducted within each tissue of the STARNET cohort as it consists of CAD cases only. We constructed linear regression models predicting gene expression levels from individual genotype combinations of interacting SNPs (review **Table 4.1**) (see Section 4.1. Materials & Methods). To overcome the risk of inflating test statistics resulting from low sample counts in specific genotype combinations, we evaluated the statistical significance of the observed differential expression patterns in these rare combinations. This was achieved by comparing the observed distribution of gene expressions in a certain genotype combination against the expected distribution of gene expressions in the same genotype combination based on permuted data obtained from a randomization test. This allowed us to further exclude differential trans gene expression results that are likely due to chance (FDR < 5%), account for outliers, and identify n=5766 unique differentially expressed genes across all SNP pairs (n=4), genotype combinations (n=36), and tissues (n=7) in the STARNET study. Following the same methodology, we were able to replicate (FDR < 5%) n=1142 unique differential gene expressions in the GTEx cohort. Among these, n=299 unique genes are differentially expressed in dependence of the same SNP pair genotype combination in the GTEx cohort (see **Supplementary Table 4.5** for all detailed results). Moreover, n=6 genes, including two novel uncharacterized genes, are significantly differentially expressed in the same SNP pair genotype combination with the same direction of effects between the cohorts (see **Supplementary Table 4.6** for this gene subset with detailed results). Strikingly, n=2 genes (SLC25A4 and IGF2) were, in addition, differentially expressed in the same or a highly related tissue type between the cohorts, representing the most consistent and confident result subset.

For instance, the interaction between the SNPs rs73222236 and rs11911017, with cis-genes MSL2 and LINC000189, leads to differential expected and observed case frequencies in the GA/TT genotype combination (Fig. 4.3). The data suggest a protective effect as fewer cases are observed under the interaction (“Obs”) as compared to the additive model (“Exp.”) which does not account for the interaction effect. In this genotype combination we can observe the differential expression of trans target gene SLC25A43, in both, mammary artery (MAM) tissue in STARNET (Fig. 4.4B; beta = 0.03, nominal p = 8.5e-5 and FDR adjusted p = 1.4e-2; nominal permutation p-value for GA/TT genotype = 3.6e-2 and FDR adjusted permutation p = 3.7e-2) as well as aortic tissue in GTEx (Fig. 4.4 C, beta = 0.01, nominal p = 1.2e-4 and adjusted p = 2.2e-2; nominal permutation p-value for GA/TT genotype = 5e-4 and FDR adjusted permutation p = 1e-3). Both tissue types are highly related. The GTEx tissue ‘Artery - Aorta’ refers to tissue samples from the ascending aorta or other thoracic regions, while the ‘Mammary Artery’ refers to samples from the internal thoracic artery.

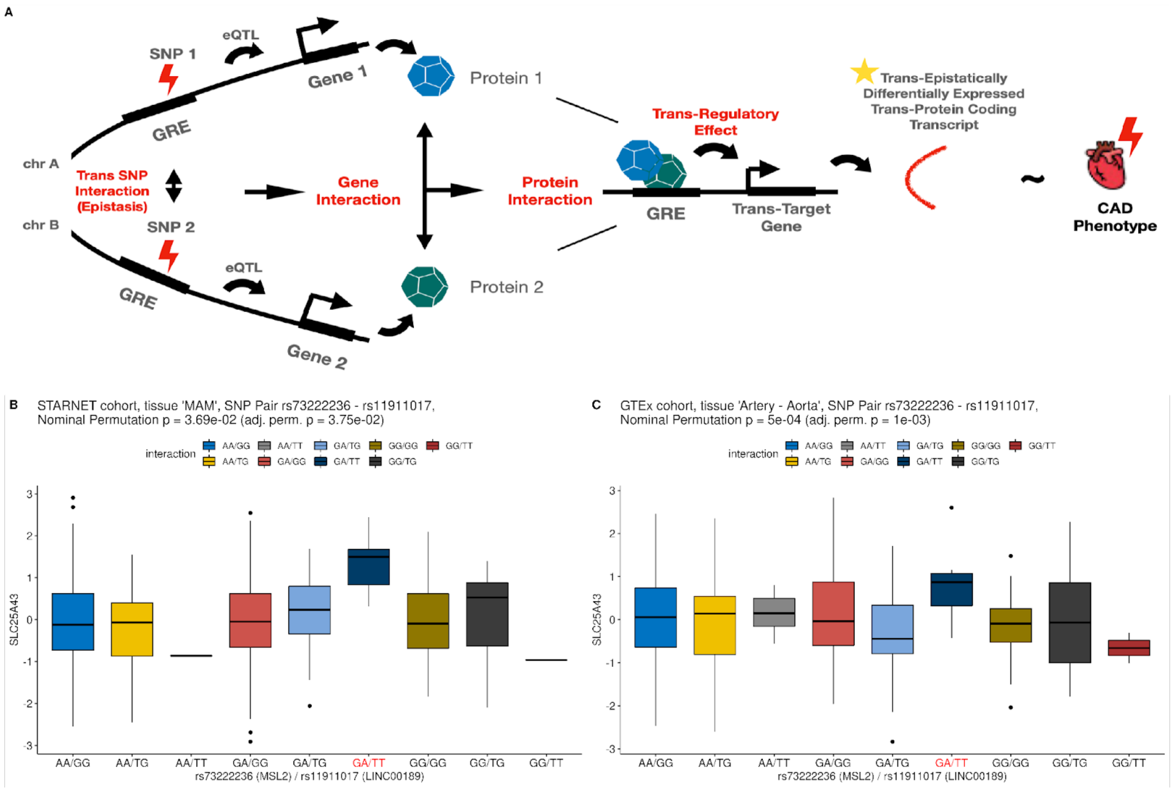


Figure 4.4: (A) Conceptual figure of the genotype-combination dependent differential trans-gene expression analyses. Identified SNP pairs (SNP 1 and SNP 2) might affect the expression of interacting genes in cis (eQTL effects on Gene 1 and Gene 2) which in turn might co-dysregulate trans-target genes and propagate the genetic interaction effect into the gene expression space. **(B)** Genotype combination (GA/TT) dependent differential expression of trans target gene SLC25A43 of MSL2 / LINC000189 associated interacting SNPs (rs73222236 / rs11911017) in 'MAM' tissue in the STARNET cohort. **(C)** Genotype combination (GA/TT) dependent differential expression of trans target gene SLC25A43 of MSL2 / LINC000189 associated interacting SNPs (rs73222236 / rs11911017) in 'Artery - Aorta' tissue in the GTEx cohort.

The effects of individual interacting SNPs on genes in cis (<1MB) were estimated with an expression quantitative trait locus (eQTL) analysis, in which cis gene expression levels were modeled in dependence of the individual SNPs genotypes as predictors (see **Supplementary Table 4.7** for cis-eQTL results). The genetic loci rs73222236 and rs11911017 showed a significant association (alpha = 0.05; p = 1.7e-3 for MSL2 and p = 6e-8 for LINC000189; **Supplementary Figure 4.5 A-B**) with respective genes localized in cis in the 'Artery - Aorta' tissue in GTEx, which firmly support the hypothesis (Fig. 4.3 A) of the differential regulation of trans target genes, in this case, SLC25A43, in dependence of genotype-dependent differentially regulated upstream factors, here MSL2 and LINC000189, as regulatory factors modulating trans target genes and propagating the genetic interaction effect in trans into the transcriptional landscape. No expression data for LINC00189 was available in the STARNET cohort, and no statistically significant cis effect of rs11911017 on MSL2 could be observed (see **Supplementary Figure 4.5 C**). Therefore, the mechanistic details of the trans-regulatory mechanism in the STARNET cohort remain to be understood.

The SNP interaction between rs73222236 and rs11911017 in the GG/TT genotype-combination distinctly leads to differential expected and observed case frequencies (Fig. 4.2) with protective effects as fewer cases are observed under the interaction ("Obs") as compared to the additive model ("Exp."). It also leads to the differential expression of trans target genes IGF2, COL27A1, and novel gene ENSG00000247679 with consistent direction of effects (see **Supplementary Figures 4.6-4.8**). For example, IGF2 is differentially regulated in the GG/TT genotype-combination in skeletal muscle tissue in STARNET (see **Supplementary Figure 4.6 A**; beta = 0.73, nominal p = 2e-4 and FDR adjusted p = 4.4e-2; nominal permutation p-value for

GG/TT genotype = 1.5e-3 and FDR adjusted permutation p = 9.2e-3) as well as skeletal muscle tissue in GTEx (see **Supplementary Figure 4.6 B**, beta = 0.29, nominal p = 3.2e-5 and adjusted p = 1.4e-3; nominal permutation p-value for GG/TT genotype = 6.3e-3 and FDR adjusted permutation p = 3.1e-2).

Considering the SNP pair rs72685791 and rs12202017, the AA/GG genotype-combination leads to differential expected and observed case frequencies (see **Supplementary Figure 4.2**) as well as the differential expression with consistent direction of effects of GH1 and novel gene ENSG00000250613 (see **Supplementary Figures 4.7-4.8**).

Differentially expressed trans target genes could not be observed for the SNP pair rs4719608 - rs2487928 and SNP pair rs10841443 - rs12899265.

4.3. Discussion

Coronary Artery Disease (CAD) is one of the major causes of death worldwide (5). Many genetic association studies (376, 419, 420) have been conducted, yet, research into genetic interactions (epistasis) remains scarce, mainly due to statistical and computational limits imposed by the problem complexity. Hence, we improved our knowledge of epistatic interactions underlying CAD with a filter-based randomization testing approach, which identified n=4 SNP-SNP interactions (review Table 4.1) in the UK Biobank cohort between the genetic loci rs72685791 & rs12202017, rs73222236 & rs11911017, rs4719608 & rs2487928, and rs10841443 & rs12899265. We validated the interactions in logistic regression models by modeling the disease label in dependence of an interaction term between the two loci of an interacting SNP pair. For the first time, we successfully replicated the interactions in an independent aggregate data set of 11 large-scale genome-wide case-control studies (379, 387–396), which we refer to as the GWAS cohort.

We integrated gene expression and genotype data for multiple tissues from the STARNET (397, 398) study to evaluate the potential downstream effects of interacting loci on the differential expressions of genes in cis and trans. Differentially expressed trans genes were identified through linear regression analysis in which the expression of individual genes localized in trans of the SNP pairs was modeled in dependence of individual multiplicative terms of the individual loci of the interacting SNP. Permutation testing of gene expression patterns in each genotype combination allowed us to exclude results that are likely due to chance. By following an analogous approach in the GTEx v8 (399) cohort, except limiting analysis on the STARNET discovery set of genes, we successfully replicated the differential expressions of n=1142 unique genes. These genes passed the FDR threshold of <5% in the linear regression and permutation testing procedure in both cohorts. The differential expressions of n=6 genes, including two novel uncharacterized genes, are observable in the same SNP pair genotype combination and show consistent direction of effects between the cohorts. This set of genes comprises SLC25A43, IGF2, COL27A1, GH1, as well as novel genes ENSG00000247679 and ENSG00000213269. Strikingly, SLC25A4 and IGF2 were additionally differentially expressed in the same or a highly related tissue type between the cohorts, representing the most confident results, as they meet all replication criteria.

Literature research about the association of identified trans target genes with CAD revealed strong implications in cardiovascular events and diseases (CVD) in general. We focused the literature research on n=6 genes (SLC25A43, IGF2, COL27A1, GH1, ENSG00000247679, ENSG00000213269) as these are differentially expressed in dependence of the same SNP pair genotype-combination, have consistent directions of effects and, in two cases, are observed in related tissue types. To begin with, SLC25A43 is a nuclear phosphate carrier, localized in the inner membrane of the mitochondrion and is a member of the solute carrier family 25, the largest of all transporter families (SLC families) (421, 422). SLC family members (n > 400) mediate solute (e.g. ions, nucleotides, and sugars) influx and efflux across plasma and intracellular membranes, conducting essential roles in the localization of molecular substances required for cell homeostasis (423). Given their essential role in mediating substrate concentrations across cellular compartments, they have been recognized as important drug targets (424, 425). In fact, SLC transporters have been shown to contribute to many Mendelian and complex multifactorial diseases if perturbed (423). For instance, genetic variants at solute carrier SLC22A3 are associated with coronary heart disease (CHD), hindering the inflammatory response and reducing the risk for CHD (426, 427). Similarly, the members of the solute carrier family 25 are responsible for a series of diseases (428). Nonetheless, a link to CAD does not exist yet. In this context, our analysis proposes that SLC25A43, as a trans target of rs73222236 and rs11911017 in the GA/TT genotype combination, might exhibit protective effects in CAD as fewer CAD cases are observed under genetic interaction (review Fig. 2). Additional CVD-associated solute carriers are among our differential expression results, namely SLC22A20 (429), SLC23A3 (430), and SLC24A3 (431), and further, underline the importance of SLC in CAD and CVDs in large.

IGF2 (Insulin-like growth factor 2) occupies important roles in multiple processes ranging from cell proliferation and growth to migration and differentiation as well as survival in general (432), but more importantly, together with insulin-like-growth factors−1, it is strongly associated with CVDs like heart failure, cardiac hypertrophy and diabetes (433, 434). They even emerged as epigenetic mediators thereof (435) and, as demonstrated in animal models (436), can delay infarction and improve post-infarction healing. Specifically for IGF2, it has been shown to influence the size of atherosclerotic lesions (432, 437). Beyond cardiac phenotypes, IGF2 also occupies critical roles in developing various cancers, including breast, colon, and lung cancer (432), which further underlines the significance of IGF2 in disease.

COL27A1, among several other collagen genes, is strongly associated with spontaneous coronary artery dissection (SCAD). The SCAD phenotype is characterized by a tear inside a coronary artery that leads to the disruption of blood flow resulting in an oxygen deficiency in the heart muscle, which dies as a result thereof, eventually leading to a heart attack (438). A heart attack resulting from SCAD, due to the tearing of the arteries, differs from a heart attack caused by stiffening of arteries (CAD). However, both phenotypes are strongly related, and our results suggest that COL27A1 might be implicated in the disease etiology of both phenotypes.

GH1, a pituitary growth hormone, is differentially regulated in growth hormone signaling pathways in a study that examined the relationship between a genetically determined decrease in height and an increased risk of CAD (439). Investigations of an Asian Indian population at high risk of CVD also established the association between genetic variation in the promoter region of GH1 and its receptor with CAD and stature (440). Likewise, our results support the observation about the association of GH1 with CAD.

Evidence for the association of novel genes ENSG00000247679 and ENSG00000213269 with CAD was not available. Their specific functions remain to be understood. However, our results provide the first indication of potential novel epistatically trans-regulated factors underlying CAD.

Our analysis of trans-epistasis underlying Coronary Artery Disease has certain limitations we would like to address. For instance, the epistatic interactions have relatively small effect sizes (mean absolute beta-coefficient of 0.065 (UKBB) and 0.068 (GWAS)). Still, we successfully replicated the SNP interactions in an aggregate cohort of 11 independent case-control studies with over 35k samples and a heterogeneous European population background. The filter-based permutation testing approach limits the discovery of other epistatic interactions, as it only selects the top 1% of trans SNP interactions. Consequently, we might miss other causal pairs that are lower ranked or that interact in cis. However, we reach statistical limits without these prefilters. Control of false discoveries (false positives) through p-value correction procedures greatly restricts the discovery of true positives. Regarding the genotype combination dependent differential trans gene expression analysis, a major limitation lies in the smaller numbers of samples with certain genotype combinations, particularly the homozygous minor allele genotypes. This limitation could be overcome to some extent by applying a secondary validation procedure based on a t-test permutation testing approach of gene expressions in these genotype combinations. Replicating identified differential expressions in an independent cohort also greatly increased the confidence in the obtained results. Nevertheless, once broadly available for CAD, further validation of larger sample sets with specific genotype combinations would further increase the confidence in the obtained results. Lastly, compared to the STARNET cohort, which consisted solely of CAD cases, the GTEx cohort comprised healthy controls only. Thus, replication in another independent cohort other than the STARNET study with CAD cases only would be beneficial.

Despite these limitations, we were successful in identifying, and for the first time also independently replicating, trans epistatic interactions for Coronary Artery Disease. Subsequent genotype combination-dependent differential trans gene expression analyses provided strong evidence that combinatorial gene regulation could explain the epistatic consequence. Nevertheless, additional experimental follow-up studies are required to confirm this regulatory hypothesis and discern the specific functions of the epistatically dysregulated trans target genes. To conclude, our analyses equip us with specific experimental targets (interacting loci, allele combinations, and trans gene targets) for a systematic and targeted experimental manipulation to evaluate downstream effects of genetic variation underlying CAD in much greater detail.

5. Summary & Outlook

Healthcare and science see great benefit from the exponential increase in the availability of big data resulting from the global digital transformation (207, 209). Developments in machine learning enable the examination and characterization of large-scale genomic data sets in a systematic and targeted way with high precision (221, 223, 441–447). As such, machine learning has become an essential instrument of computational biology that enables precision research on a large scale (448–450). We harnessed these potentials by conducting machine learning-driven analysis of large-scale biological data sets for **1)** the identification of novel trans-regulatory transcription factors that modulate transcriptional pausing of the polymerase II and **2)** the identification of trans epistatic interactions underlying Coronary Artery Disease (CAD) and the characterization of their potential downstream transcriptional effects. Thus our contributions specifically advance our knowledge about transcription and transcriptional regulation (3, 66, 87, 317, 451) as well as genetic interactions (160) underlying disease, which is, in large, achieved by modeling the respective biological contexts as inputs to machine learning models to discern and interpret the biological patterns that constitute each biological phenomenon under investigation as summarized in the following two paragraphs.

5.1. Towards a comprehensive understanding of transcriptional regulation

In our first project, we specifically sought to identify trans-acting promoter-proximal Polymerase II (Pol II) pause regulatory elements. They play an essential role in transcriptional gene regulatory programs and maintain cell homeostasis and plasticity by controlling the productivity of Polymerase II. Their identification and characterization are thus crucial for a holistic understanding of gene transcription. We built the first comprehensive interpretable complex machine learning model (Extreme Gradient Boosting Tree) of transcriptional pausing and enabled the identification of such novel trans-acting factors. Our model accurately predicts the extent of promoter-proximal Pol II pausing from large-scale features of genomic and transcriptomic protein binding maps and gene annotation and sequence composition features. The model's generalizability was demonstrated through its application to data from an independent cell line. Harnessing the underlying model contributions, we were able to pinpoint and rank specific transcriptional pause regulatory factors, many of which have already been shown to affect transcriptional pausing or to be implicated in certain processes associated with pausing. Additional models trained on subsets of proteins implicated in transcription and specific transcription regulatory processes ranging from chromatin remodeling, transcription initiation, elongation, splicing, and further downstream transcript processing enabled us to strengthen the connection of these processes with transcriptional pausing additionally. The importance of the established 7SK pause mediator complex and associated factors in transcriptional pausing was further demonstrated by the predictive power of existing and novel 7SK binding factors as identified from large-scale RNA-Protein interaction data. Taken together, these trans-acting pause regulatory factors and their functional backgrounds provide initial mechanistic hypotheses and represent interesting targets for experimental validation.

We have chosen to analyze data for the HepG2 and K562 cell lines from the ENCODE project, as they have been extensively characterized with multiple types of essays and provide a large number of genomic and transcriptomic binding maps thereby enabling the identification of novel regulators of promoter-proximal pausing, however, not all previously characterized, and most likely also several unknown regulators of pausing, are available. Moreover, only GRO-seq data with similar variations quantifying promoter-proximal pausing are available for these two cell lines. Multi-omics approaches like TT-seq (373) or mNET-seq (297, 374, 375) capture Pol II pausing more accurately yet are only available for a limited number of cell lines for which protein binding data is not available, which prevents the identification of novel regulators of promoter-proximal pausing across cell types. However, they provide ground for future studies of transcriptional pausing with greater precision and will enable elaborate validation procedures once they are available across multiple cell types, which is currently not the case with the ENCODE cell line, which would make cross-validation impossible. Taken together, we established a framework to extend our knowledge of transcriptional regulation of protein-coding genes from the perspective of promoter-proximal Polymerase II pausing through the identification of novel trans-acting transcriptional pause regulatory factors.

We expect great improvements with accurate kinetic profiling of the polymerase's productivity. ChIP-seq data, once available for a broader range of proteins, will likely reveal yet other novel regulators of transcriptional pausing. Once broadly available across different cell types, protein binding data can reveal potential cell-type-specific transcriptional pause regulators. Alternatively, potential transcriptional pause master regulators could be identified by integrating protein binding data from multiple cell types. However, probing about 1600 transcription factors in all cell lines, ideally in different conditions, poses a challenge with the available technology as of now. An alternative feasible solution is to predict transcription factor binding sites from ChIP-seq which substantially mitigates this problem (452, 453). Coupled with open chromatin data like ATAC-seq (454) candidate active transcription factor binding sites can be inferred based on conserved transcription factor binding sequences that might be located in those open chromatin regions (455, 456). Further improvements in model predictions can be expected with improved protein binding site predictions. With ongoing advancements in genomics at large, we envision a model that predicts transcriptional pausing outcomes from DNA sequence alone.

5.2. Enriching our understanding of genetic interactions and their role in complex disease

In our second project, we aimed to identify and characterize genetic interactions, called epistatic interactions, underlying Coronary Artery Disease (CAD), one of the major causes of death worldwide. Despite numerous studies, ranging from genetic association to case-control or epidemiological studies, CAD remains one of the leading causes of death worldwide in both developed and developing countries. Particularly we lack studies focusing on epistasis discovery mainly due to statistical and computational limits imposed by the intrinsic problem complexity given by the exponentially growing search space in dependence of the order of genetic interaction considered. However, we able to extend our knowledge of the effects of pairwise genetic interactions underlying CAD with a filter-based permutation testing coupled with linear modeling, which enabled us to discern n=4 genetic interactions that confer differential disease risk in a large-cohort (UK Biobank) of >35k CAD cases. For the first time, we could replicate all interactions in an independent aggregate data set of 11 large-scale genome-wide case-control studies (>35k CAD cases). Genotype-combination dependent linear regression analysis coupled with statistical hypothesis testing further enabled us to elucidate the downstream effects of interacting loci on the transcriptional output of genes in trans (>1 Mbp), identifying many differentially expressed unique genes in a large-cohort of CAD cases (STARNET) that could be replicated in an independent cohort (GTEx v8). Among this list of genotype-combination dependent differentially regulated trans target genes, we found two genes, namely SLC25A43 and IGF2, that could be replicated across all replication criteria, i.e., the same tissue type, genetic interaction, genotype combination of the genetic interaction, and direction of the effects in the specific genotype combination, representing highly confident and consistent results. Literature research of identified genes revealed strong implications in the disease etiology of CAD. To summarize, for the first time, we were able to identify and replicate specific genetic interactions and trans target genes thereof as entry points for a systematic and targeted experimental investigation of epistasis underlying CAD.

Nonetheless, our analysis comes with a few limitations. For instance, the effect sizes of the identified genetic interactions are relatively small, with mean absolute beta-coefficients of 0.065 in UKBB and 0.068 in GWAS. In spite of this, for the first time, we were successful in replicating the interactions in an independent large-scale (>35k samples) aggregate cohort of 11 case-control studies and also, for the first time, in identifying and replicating their downstream effects on the transcriptional output of trans target genes in two additional independent cohorts. Another limitation lies in the filter-based permutation testing approach, which **1)** only considers inter-chromosomal SNP pairs (neglecting intra-chromosomal SNP pairs) and **2)** only selects the top 1% of trans SNP interactions. Consequently, we might miss interacting SNP that lie on the same chromosome (cis-interacting loci) and other causal pairs that are missed by the thresholding approach. However, these prefilters allow us to circumvent confusing haplotype effects and overcome statistical limits in which the control of false discoveries (false positives) through p-value correction procedures like the FDR correction method limits the discovery of true positives. The differential trans gene expression analysis shows a limitation in the low number of individuals with certain genotype combinations, particularly the homozygous minor allele genotypes, which require additional validation procedures. This limitation was alleviated with a randomization test of gene expressions in these rare genotype combinations and the replication analysis in an independent cohort. Lastly, because the STARNET cohort consisted solely of CAD cases as opposed to the GTEx cohort, which consisted solely of healthy controls replication analysis in yet another independent cohort other than the STARNET study with CAD cases only as well would even further increase the confidence in obtained results. To conclude, despite these limitations, we were able to identify and independently replicate trans epistatic interactions in Coronary Artery Disease for the first time. The results of our downstream differential trans gene expression analysis, dependent on the identified genetic interactions, support the hypothesis that the combinatorial gene regulation could explain the epistatic effect. Nevertheless, further experimental studies are required to confirm this regulatory hypothesis and elucidate the specific cellular functions of identified trans target genes. In this context, our analyses provide the first entry points of specific genetic loci and allele combinations and trans-regulated gene targets for such a targeted investigation of the downstream effects of genetic variation underlying CAD.

We expect better profiling of epistatic interactions once bigger sample sets are broadly available (34). Additional cohorts for validation purposes with a broad range of ethnic backgrounds would greatly increase the confidence in genetic interaction analyses as it would decrease the selection bias resulting from the overrepresentation of certain subpopulations. This will establish a generalized knowledge of the genetic basis of complex diseases underlying the whole human population. Alternatively, CRISPR-Cas based genetic perturbation experiments screens in human cell lines harbor great potential to reveal or validate genetic interactions (457, 458), yet need to be employed in a more systematic manner.

5.3. What the future holds

Our work is one of the many demonstrations of how machine learning models coupled with large-scale multi-omic data integration can enable the researcher to answer complex biological questions previously hidden behind the mass and complexity of the data points at hand. As time progresses, new experimental techniques, especially assays that measure different omic entities in parallel within the same sample (coupled assays), will result in an ever more exponential increase in the volume, variety, veracity, and complexity of biological datasets. A prime example is high-throughput single-cell sequencing (459) and its derivatives like single-cell ChIP-seq (460), single-cell ATAC-seq (461) or single-cell HI-C (462), just to name a few. These and similar future developments represent an overwhelmingly big and important source of biological information harboring the potential to significantly advance and revolutionize healthcare and science (415). For instance, genetic interactions could be investigated in single heart muscle cells, once available, as opposed to bulk measurements from populations of cells (463, 464). Single-cell ChIP-seq enables the investigation of dynamic protein binding events at single-cell resolution with the potential to identify differential protein binding patterns between single cells which might give further clues about the binding characteristics of transcriptional regulators that were previously masked in bulk measurements. More importantly, single-cell ChIP-seq coupled with GRO-seq or another similar polymerase profiling technique will allow us to study transcriptional pausing in single cells and identify regulators of pausing by jointly comparing the protein binding patterns and the polymerase's productivity between single cells (465). Experimental techniques that assay DNA-binding of multiple proteins in single cells in parallel harbor the potential to reveal specific combinatorial protein binding profiles to shed light on regulatory networks of proteins and study the dynamics of these networks between single cells (460).

Not only experimental but also ongoing algorithmic advancements, especially in the field of artificial intelligence and machine learning (220, 466, 467), a decline in computational costs, and increased computational resource availability will provide the grounds to efficiently harness such data to draw an ever more precise picture of our biological world (468–470). Thus machine learning is an indispensable part of computational biology and is applied in many research contexts. For instance, machine learning models can be used to improve the annotation of the genome, i.e., they can be trained to learn specific sequence elements of a given type. The sequence composition features we have engineered for our model of transcriptional pausing are an example of this. Similar approaches have been conducted across many sequence types ranging from promoters (471), and splice sites (472) to enhancers (473), just to name a few. These models can be combined to obtain a comprehensive system to annotate genomes (424) accurately. Machine learning has also enabled the prediction of gene expression levels from DNA sequence alone (474–476). This is particularly encouraging, providing the first indications that a model of transcriptional pausing based on pure DNA sequence alone is promising. Many more exciting developments and applications of machine learning models in genomics will continue to greatly improve our understanding of biological phenomena at large (468).

Nevertheless, our understanding will not only be improved by the quantity of data but also by its quality. Particularly, improved sequencing techniques will provide broader and deeper coverage of the (epi-) genome with higher resolution. Especially machine learning models will greatly benefit from the increase in the signal-to-noise ratio reflected in improved prediction performances. Similarly, larger sample sizes will allow us to estimate the true effects of many genetic variants with small effect sizes with much greater accuracy.

Further developments in the field of multi-omic data integration will allow us to elucidate the relationships between different omic entities and open up new avenues for holistic analyses, which will greatly improve our systemic understanding of biology (462, 477, 478). For instance, single-cell assays jointly probing DNA- and RNA-binding proteins will greatly improve our understanding of combinatorial protein binding profiles of transcriptome and genome binding interacting proteins (460). Coupled with gene-expression data, these combinatorial networks of genomic and transcriptomic protein binding profiles will further highlight the importance of interactions between different omic entities necessary for proper gene regulation and maintenance of cellular homeostasis (479).

Coupled with precision medicine, improved sequencing technologies can revolutionize our healthcare system as it will be possible to collect and integrate high-quality multi-omic data on an individual basis and analyze it with high velocity and accuracy by harnessing the power of artificial intelligence systems. This is already common practice for a range of cancer types in which the genetic makeup of genes is considered to treat cancer (480–484). For instance, during breast cancer treatment, a recurrence score is calculated based on the expression levels of 16 breast cancer-related genes providing an estimate of the 9-year distant recurrence risk and the likelihood of chemotherapy benefit on an individual basis (485). A similar approach has also been applied to CAD, in which a polygenic risk score quantifies the genetic variation of diseased patients to inform about their genetic risk (486). However, further improvements that are likely to be made are required for its application in a clinical setting. In general, personalized precision medicine enables treatment decisions to be made based on individual objective characteristics as opposed to trends that are observed across a population of individuals. As such, developments of this and similar kind will greatly foster personalized medicine and improve patient treatment and survival rates, as highly individualized treatment will be available, as opposed to bulk solutions that cannot address individual concerns.

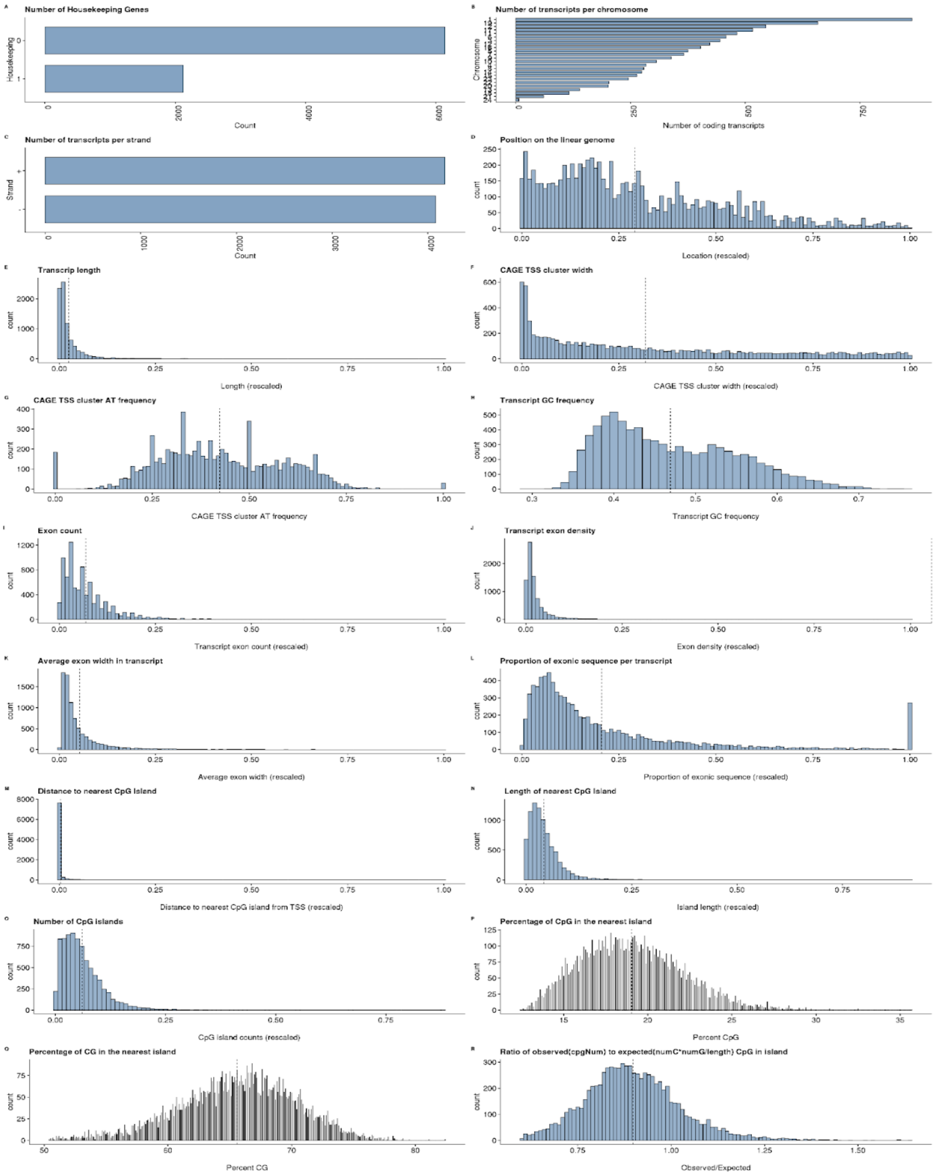
6. Appendix

6.1. List of Figures

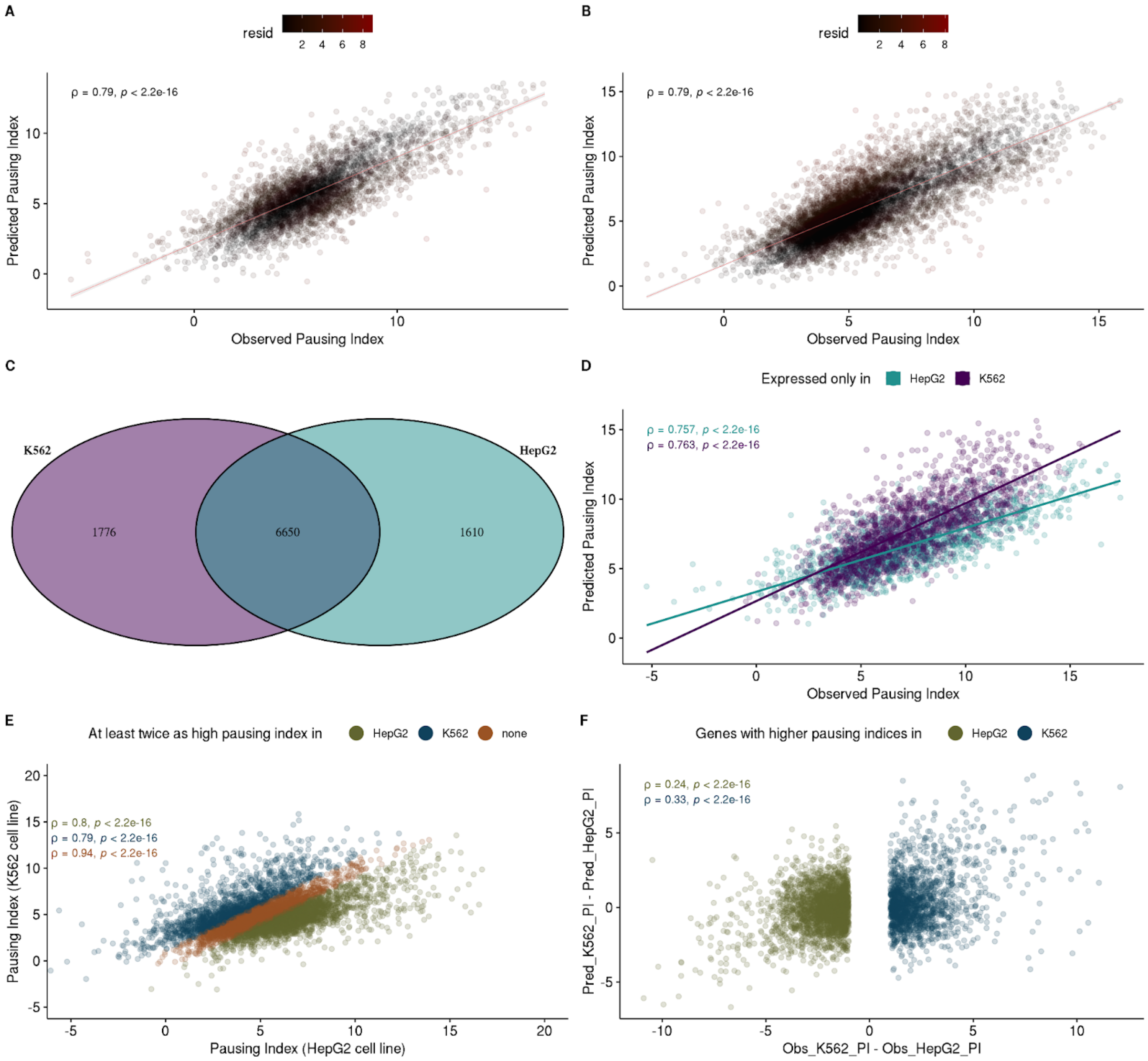
6.2. List of Tables

6.3. Chromatin Signatures and their Role in Transcriptional Elongation Control (Project 1)

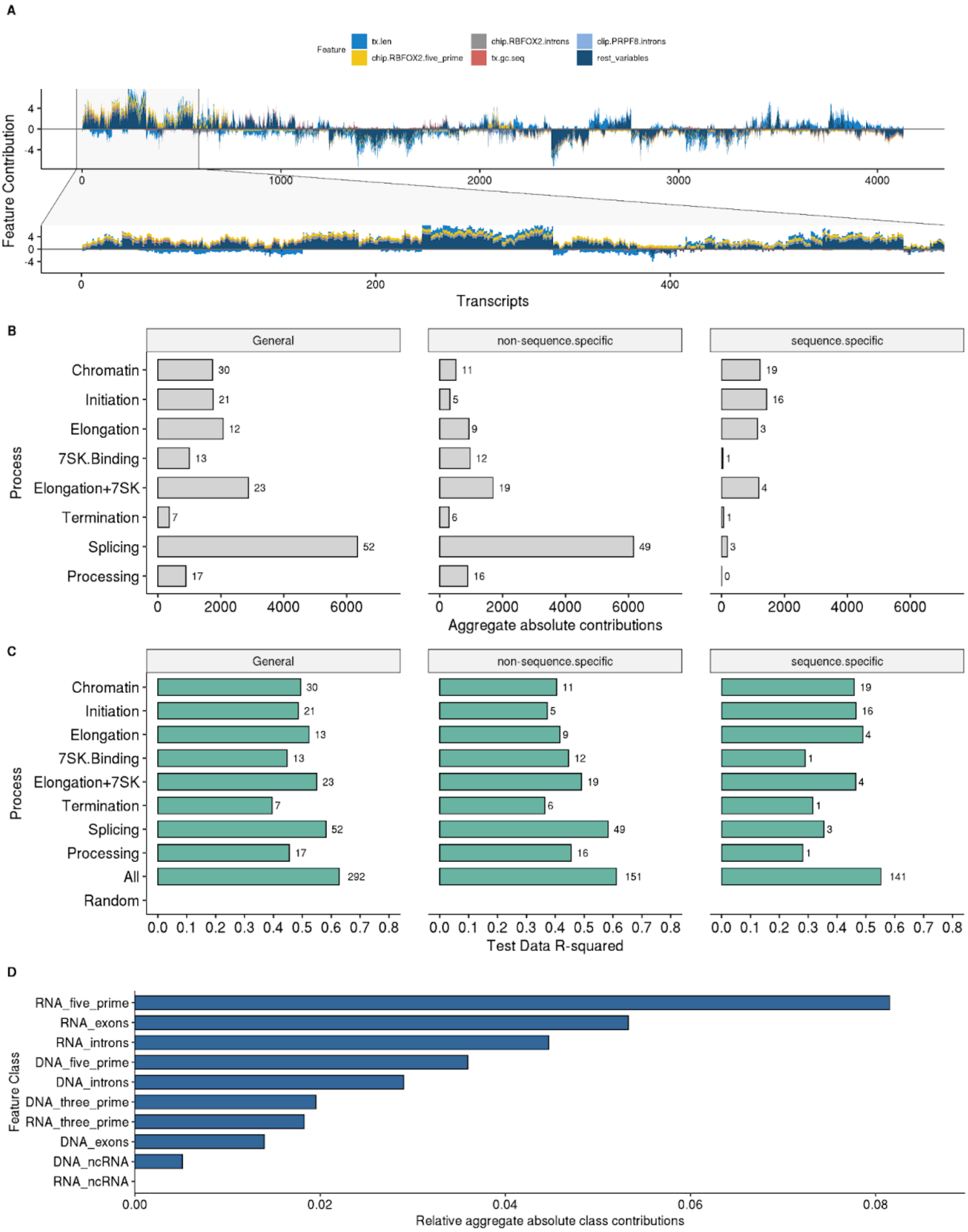
6.3.1. Supplementary Figures



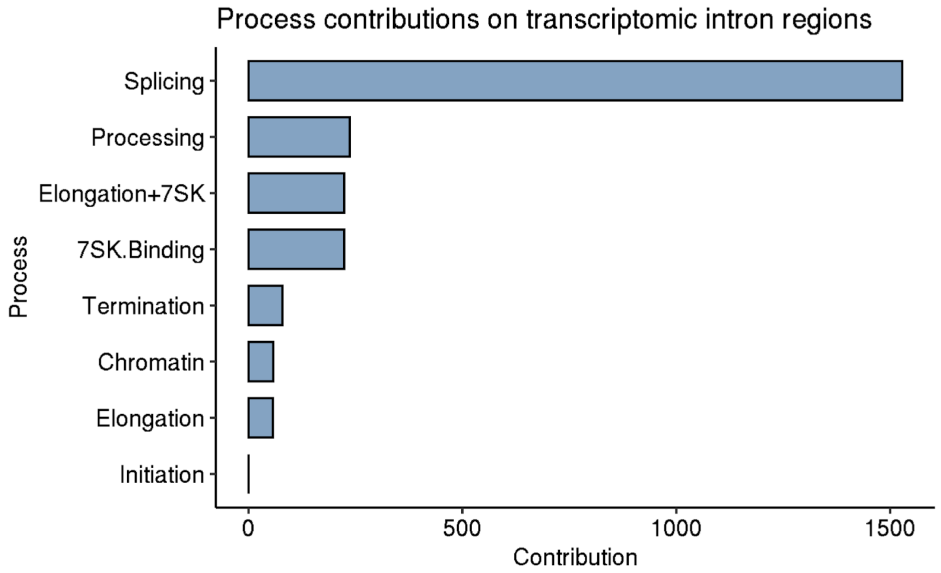
Supplementary Figure 3.1: Gene annotation and sequence composition features (HepG2). See caption of main figure 3.4 for more details.



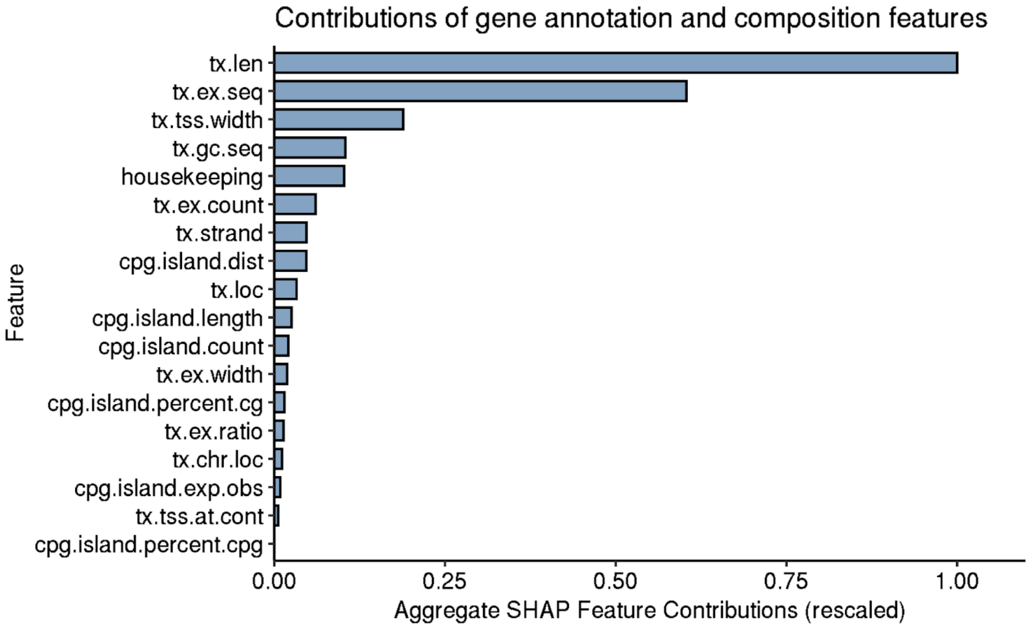
Supplementary Figure 3.2: Figure 3.6 analog for the HepG2 cell line. See caption of main figure 3.2 for more details.



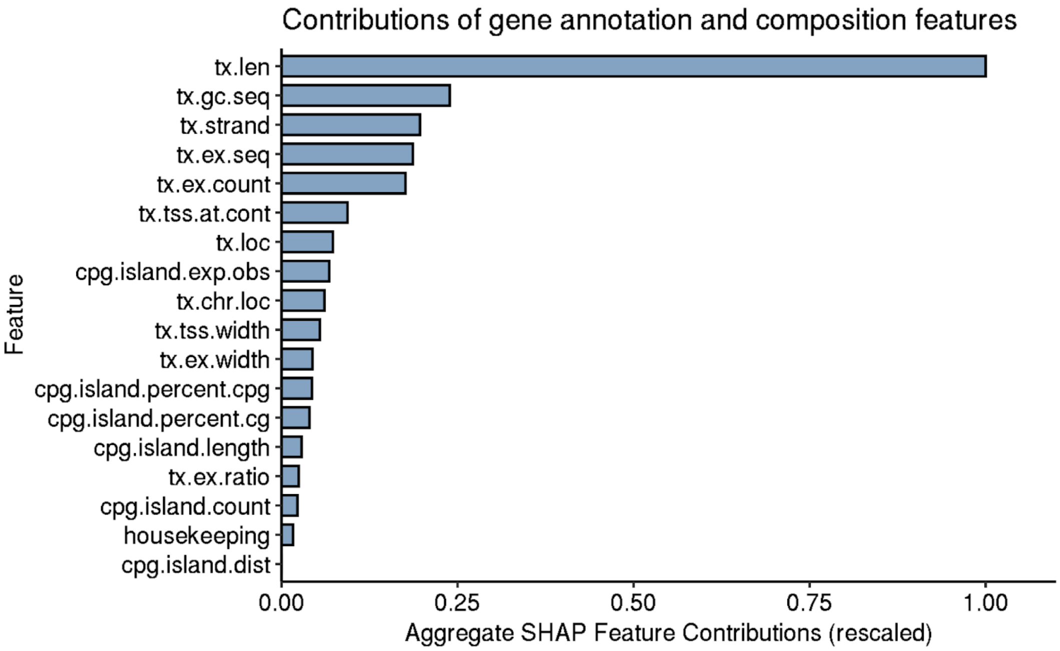
Supplementary Figure 3.3: Figure 3.7 analog for the HepG2 cell line. See caption of main figure 3.3 for more details.



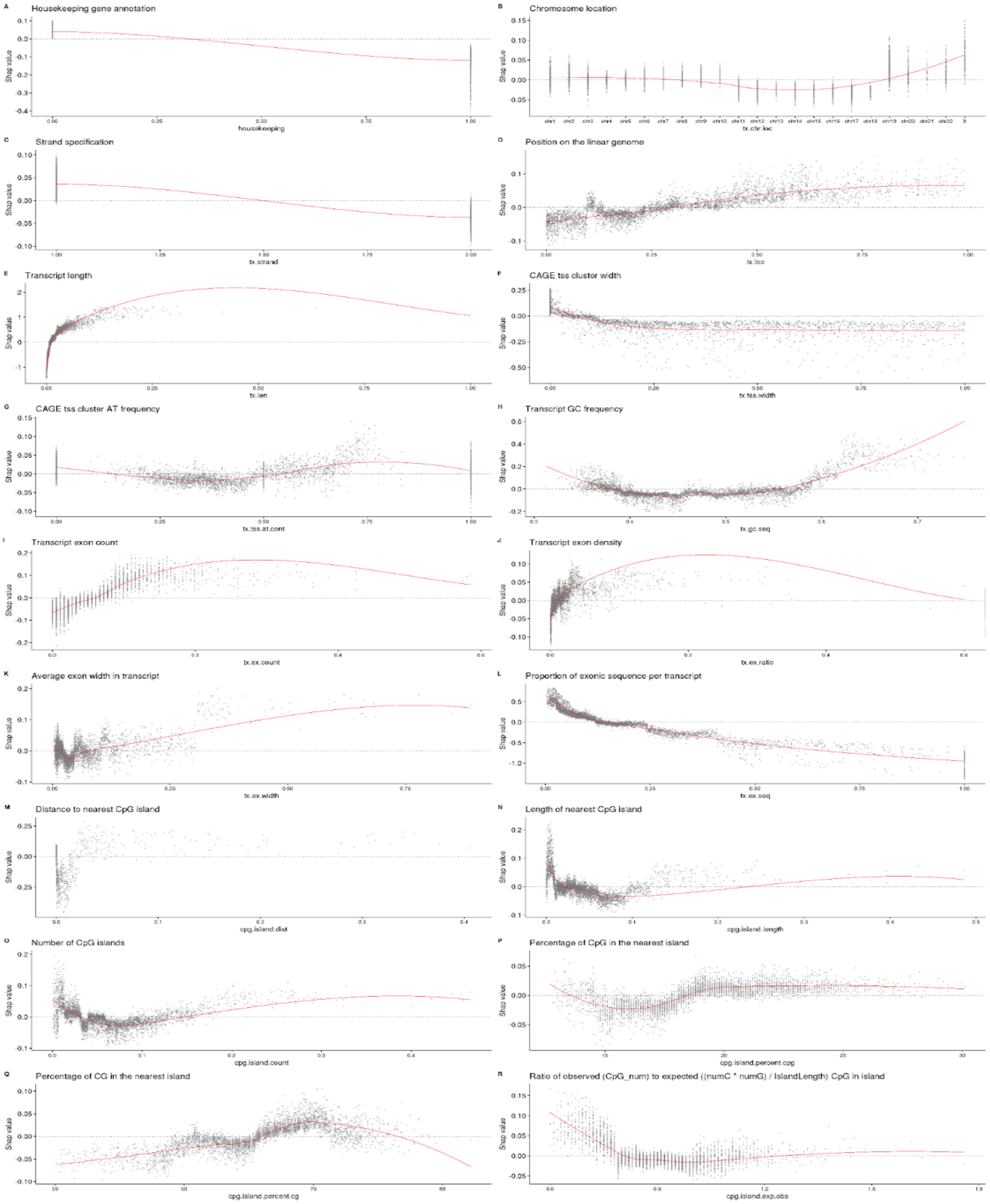
Supplementary Figure 3.4: Feature contributions on RNA introns (HepG2 cell line). See caption of figure 3.8 for more details.



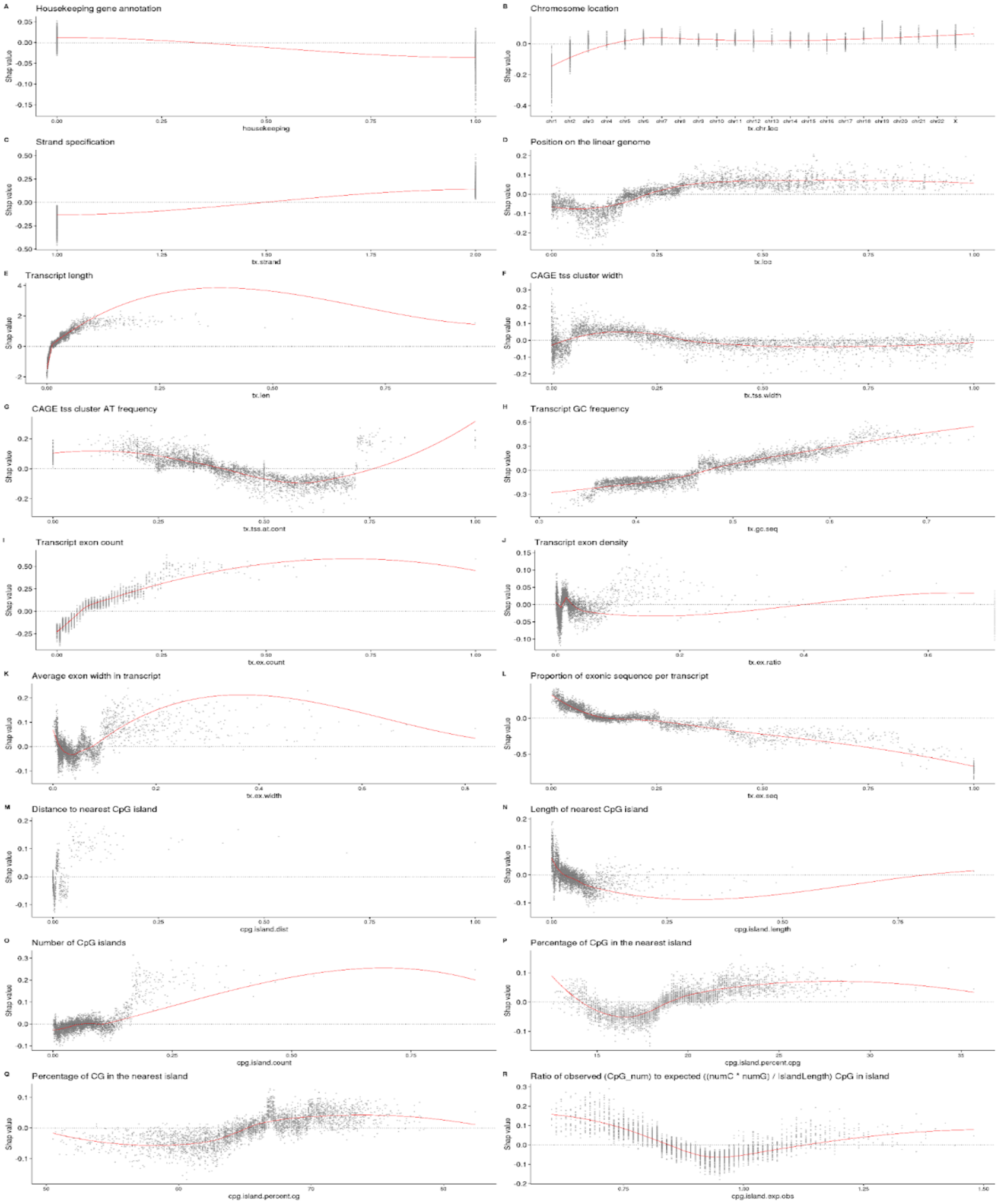
Supplementary Figure 3.5: Aggregate feature contributions of gene annotation and sequence composition features (K562). Distribution of aggregate feature contributions (x-axis) of gene annotation and sequence composition features (y-axis) in the K562 cell line.



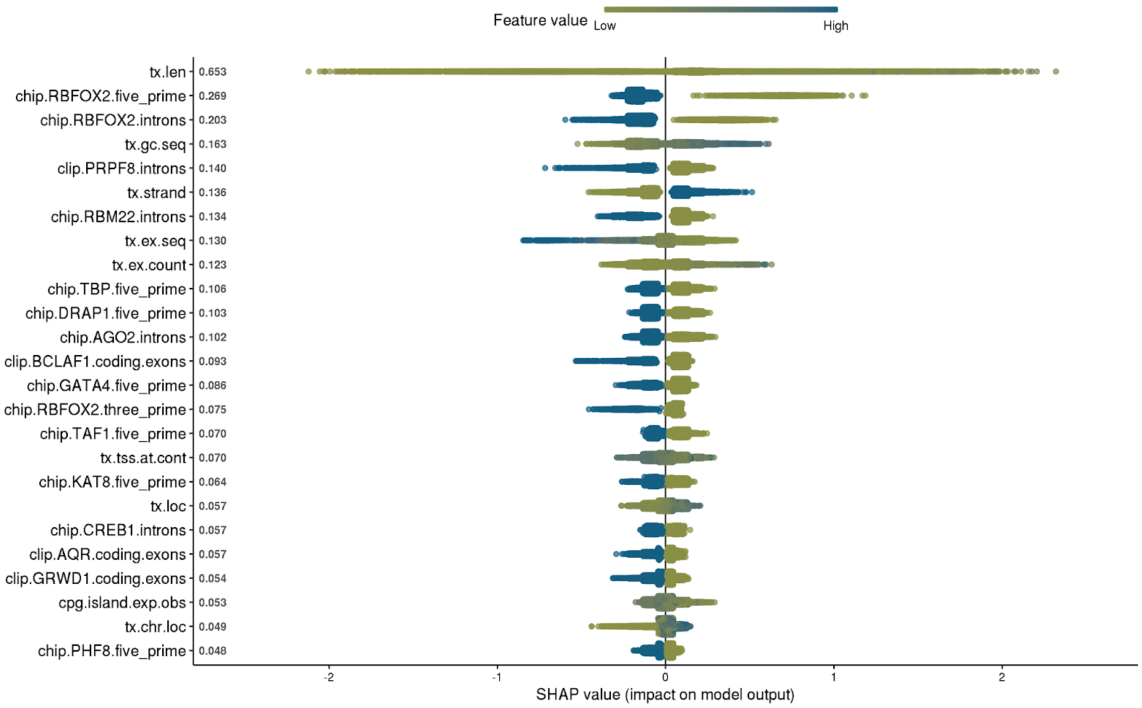
Supplementary Figure 3.6: Aggregate feature contributions of gene annotation and composition features (HepG2). See caption of supplementary figure 3.9 for more details.



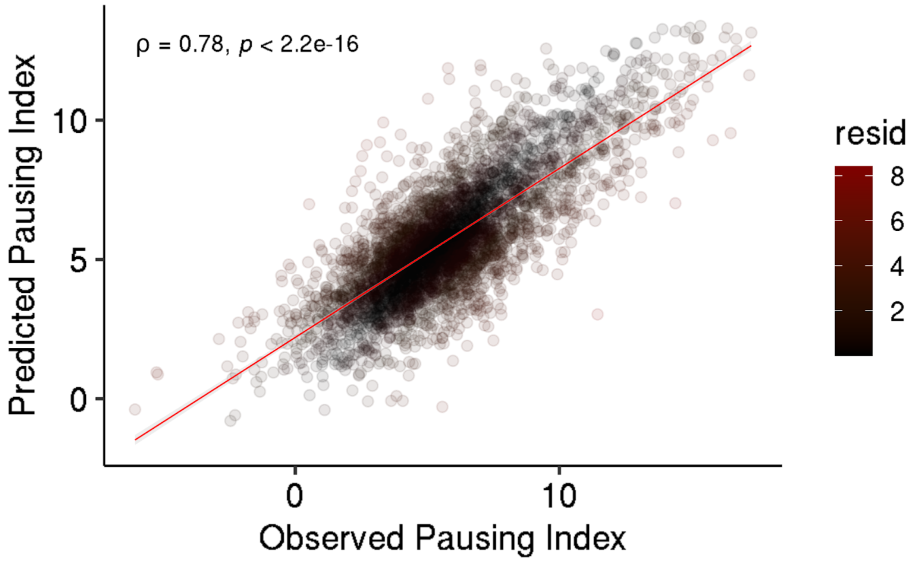
Supplementary Figure 3.7: Model feature contribution distributions of gene annotation and composition features (K562). Distribution of feature contributions (SHAP values, y-axes) in the K562 cell line of gene annotation and sequence composition features values (x-axes). In the “tx.strand” feature, “1” denotes “+” (forward) strand and “2” denotes “-” (reverse strand).



Supplementary Figure 3.8: Model feature contribution distributions of gene annotation and composition features (HepG2). See caption of supplementary figure 3.11 for more details.



Supplementary Figure 3.9: Feature Contributions (HepG2). See caption of figure 3.9 for more details.



Supplementary Figure 3.10: Minimal model performance. Observed (x-axis) vs predicted (y-axis) pausing index of the most influential factor model (n=9 factors) in the HepG2 cell line.

6.3.2. Supplementary Tables

Please refer to the supplementary materials in my corresponding manuscript currently (May 2022) under review in *Nucleic Acid Research* and also available on bioRxiv: Akcan and Heinig. *'Predictive Model of Transcriptional Elongation Control Identifies Trans-Regulatory Factors from Chromatin Signatures, Toray Akcan, Matthias Heinig'* BioRxiv (2022) to access the supplementary tables mentioned in the following.

Supplementary Table S3.1 (see xls file sheet "S1 K562 CHIPseq Factors") : List of DNA binding proteins in the K562 cell line derived from the ENCODE CHIP-seq experiments.

Supplementary Table S3.2 (see xls file sheet "S2 HepG2 CHIPseq Factors") : List of DNA binding proteins in the HepG2 cell line derived from the ENCODE CHIP-seq experiments.

Supplementary Table S3.3 (see xls file sheet "S3 K562 CHIPseq Accessions") : List of accession numbers of ENCODE CHIP-seq experiment for the K562 cell line.

Supplementary Table S3.4 (see xls file sheet "S4 HepG2 CHIPseq Accessions"): List of accession numbers of ENCODE CHIP-seq experiment for the HepG2 cell line.

Supplementary Table S3.5 (see xls file sheet "S5 K562 eCLIPseq Factors"): List of RNA binding proteins in the K562 cell line derived from the ENCODE eCLIP-seq experiments.

Supplementary Table S3.6 (see xls file sheet "S6 HepG2 eCLIPseq Factors"): List of RNA binding proteins in the HepG2 cell line derived from the ENCODE eCLIP-seq experiments.

Supplementary Table S3.7 (see xls file sheet "S7 K562 eCLIPseq Accessions") : List of accession numbers of ENCODE eCLIP-seq experiment for the K562 cell line.

Supplementary Table S3.8 (see xls file sheet "S8 HepG2 eCLIPseq Accessions"): List of accession numbers of ENCODE eCLIP-seq experiment for the HepG2 cell line.

Supplementary Table S3.9 (see xls file sheet "S9 K562 7SK Binding Factors"): List of RNA binding proteins that bind the 7SK ncRNA in the K562 cell line as evidenced by eCLIP-seq binding signals. Binding signals on pseudo 7SK ncRNA transcript variants that are expressed above median ncRNA expression levels were included in the analyses. This was motivated by the transcripts' high mean pairwise sequence similarity (487) of 0.74 and high mean conservation score of 923.58 (PAM250 scoring matrix) based on a multiple sequence alignment (ClustalW alignment) of corresponding 7SK transcripts.

Supplementary Table S3.10 (see xls file sheet "S10 HepG2 7SK Binding Factors"): List of RNA binding proteins that bind the 7SK ncRNA in the HepG2 cell line as evidenced by eCLIP-seq binding signals. Binding signals of pseudo 7SK ncRNA transcript variants that are expressed above median ncRNA expression levels were included. This was motivated by the transcripts' high mean pairwise sequence similarity (487) of 0.81 and high mean conservation score of 302.29 (PAM250 scoring matrix) based on a multiple sequence alignment (ClustalW alignment) of corresponding 7SK transcripts.

Supplementary Table S3.11 (see xls file sheet "S11 K562 Factor Bindings"): Number of genomic and transcriptomic binding events in the K562 cell line on gene transcript regions per protein.

Supplementary Table S3.12 (see xls file sheet "S12 HepG2 Factor Bindings"): Number of genomic and transcriptomic binding events in the HepG2 cell line on gene transcript regions per protein.

Supplementary Table S3.13 (see xls file sheet "S13 Known Pausing Factors"): List of established pausing factors derived from literature research.

Supplementary Table S3.14 (see xls file sheet "S14 K562 Factors per Process"): List of proteins in the K562 cell line associated with specific functional processes.

Supplementary Table S3.15 (see xls file sheet "S15 HepG2 Factors per Process"): List of proteins in the HepG2 cell line associated with specific functional processes.

Supplementary Table S3.16 (see xls file sheet “S16 K562 Sequence Specificity”): An indicator matrix (1=’yes’; 0=’no’) for the K562 cell line specifying protein’s sequence specificity (column *SS*), non-sequence specificity (column *NSS*), RNA- (column *RBP*) or DNA-binding factor (column DBP).

Supplementary Table S3.17 (see xls file sheet “S17 HepG2 Sequence Specificity”): An indicator matrix (1=’yes’; 0=’no’) for the HepG2 cell line specifying protein’s sequence specificity (column SS), non-sequence specificity (column NSS), RNA- (column RBP) or DNA-binding factor (column DBP).

Supplementary Table S3.18 (see xls file sheet "S18 Subspace Factors Presence"): An indicator matrix that specifies if a factor is contained in any of the feature subspaces, where “1” denotes present, “0” denotes” not present.

Supplementary Table S3.19 (see xls file sheet “S19 Hyperparameters”): Hyperparameters specification of the Extreme Gradient Boosting Tree regression model.

Supplementary Table S3.20 (see xls file sheet “S20 All Model Results”): Detailed model prediction results for each cell line and each feature subspace. The column ‘*subspace’* specifies the feature subspace the model was trained on. The appendix “*ss*” in feature subspace names indicates a model that was trained on binding features of sequence specific proteins, whereas “*nss”* indicates a model that was trained on binding data of non-sequence specific proteins. The model type ‘*synchronised.model.matrices’* refers to a model that was trained on features observed in both of the cell lines, as opposed to ‘*individual.model.matrices’* which refers to models that also incorporated features that are exclusive to a cell line. The column ‘*train.rsqrd’* gives the 5-fold cross-validation performance (R^2), while the column ‘*test.rsqrd’* gives the performance on a 50% hold out test data set taken prior to training. The column ‘*mean.shap’* gives the average feature contributions over all proteins considered in a model.

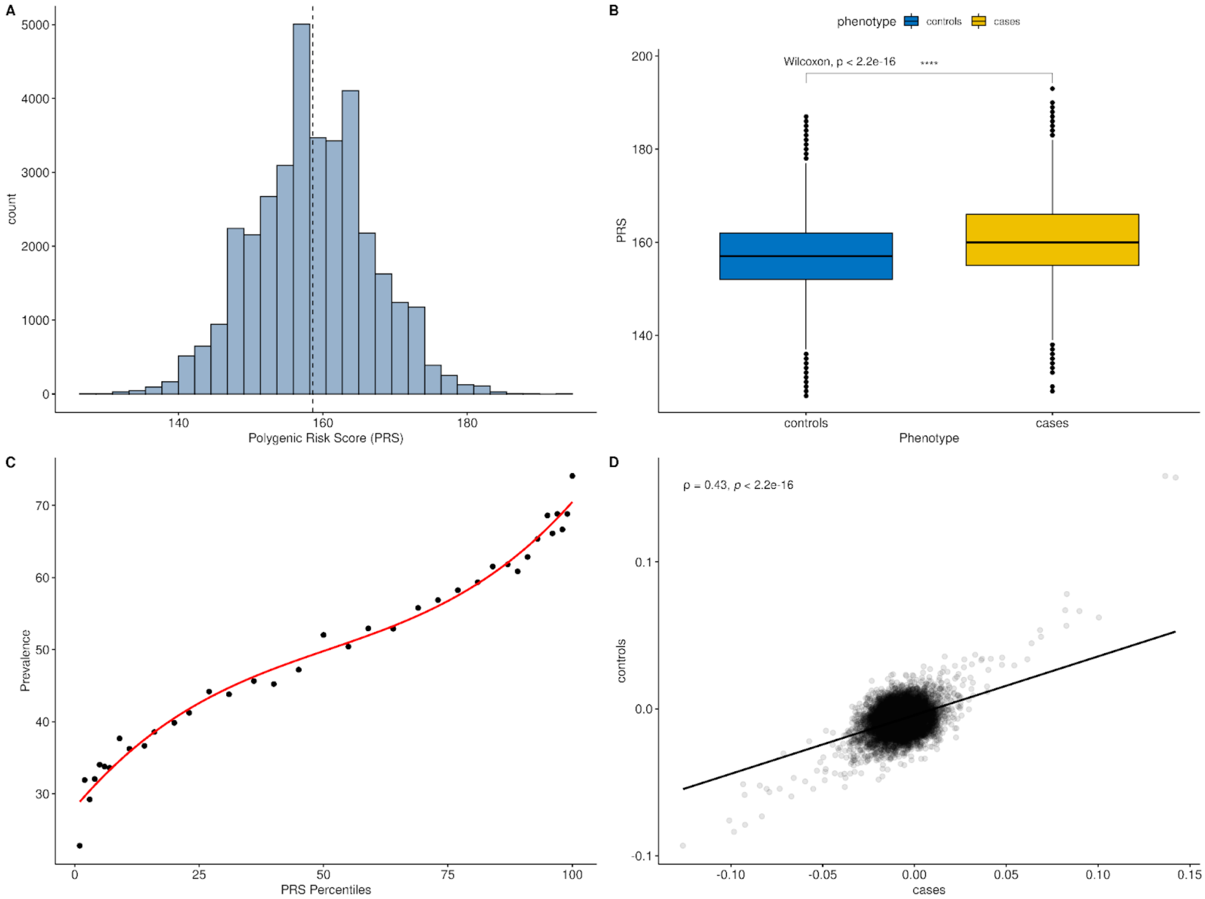
Supplementary Table S3.21 (see xls file sheet “S21 Data Accessions”): List of accession numbers of all data sets.

6.3.3. Data & Code Availability

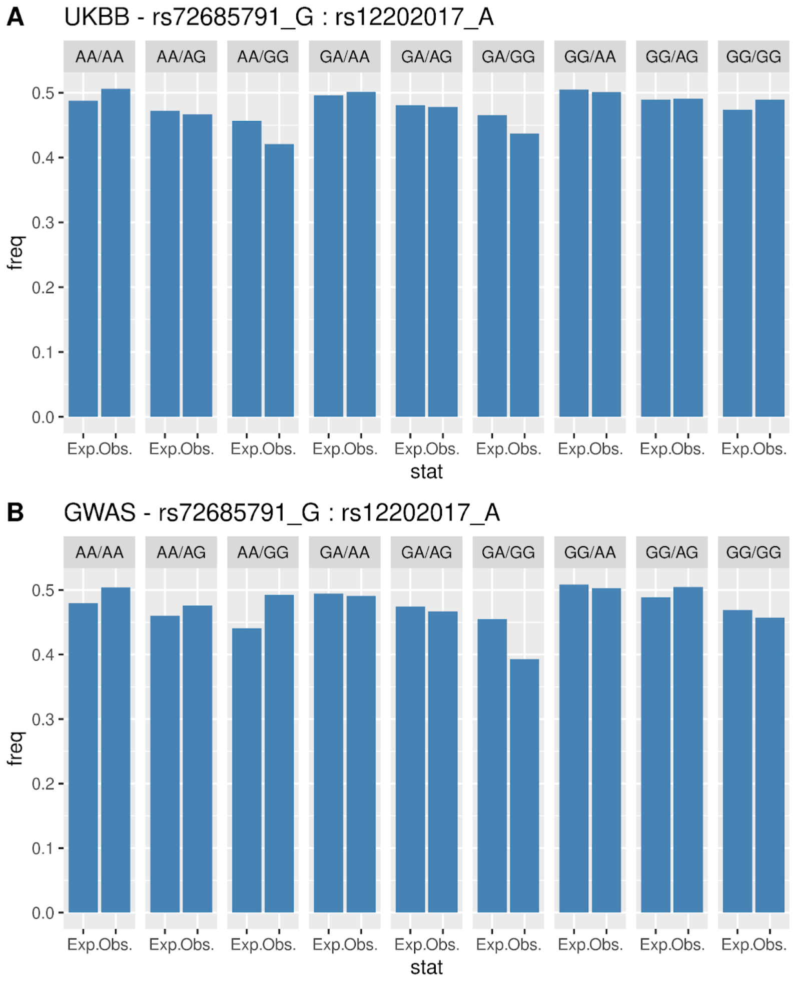
The code is available at https://github.com/heiniglab/POLII_pausing and all data as well as results are available at 10.5281/zenodo.5236311. GRO-seq data was available under the GEO accessions ***GSM1480325*** and ***GSM2428726*** for the K562 and HepG2 cell lines, respectively. Transcript quantifications from RNA-seq experiments (tsv-files) were obtained from ENCODE with the experiment number **ENCSR885DVH** and accession numbers of replicated experiments **ENCFF424CXV** and **ENCFF073NHK** for the K562 cell line, as well as the experiment number **ENCSR181ZGR** with accession numbers of replicated experiments **ENCFF205WUQ**, **ENCFF915JUZ** for the HepG2 cell line. CHIP-seq and eCLIP-seq ENCODE accession numbers are listed in supplementary tables S3 & S4 and S7 & S8, respectively. Housekeeping gene annotations were taken from (314) (see **Supplementary Table S21;** housekeeping.RDS in zenodo repository). CpG island annotations were obtained from the UCSC golden path for the hg19 genome build (cpgIslandExt.txt.gz) (see **Supplementary Table S21;** cpg.islands.RDS in zenodo repository). The GENCODE project (see **Supplementary Table S21**) served to obtain gene annotations, HGNC gene symbols mappings and RefSeq metadata files. Preprocessed CAGE transcription start sites based on ENCODE data are provided in the zenodo repository as an R-data structure (CTSS.RDS).

6.4. Trans Epistasis in Coronary Artery Disease (Project 2)

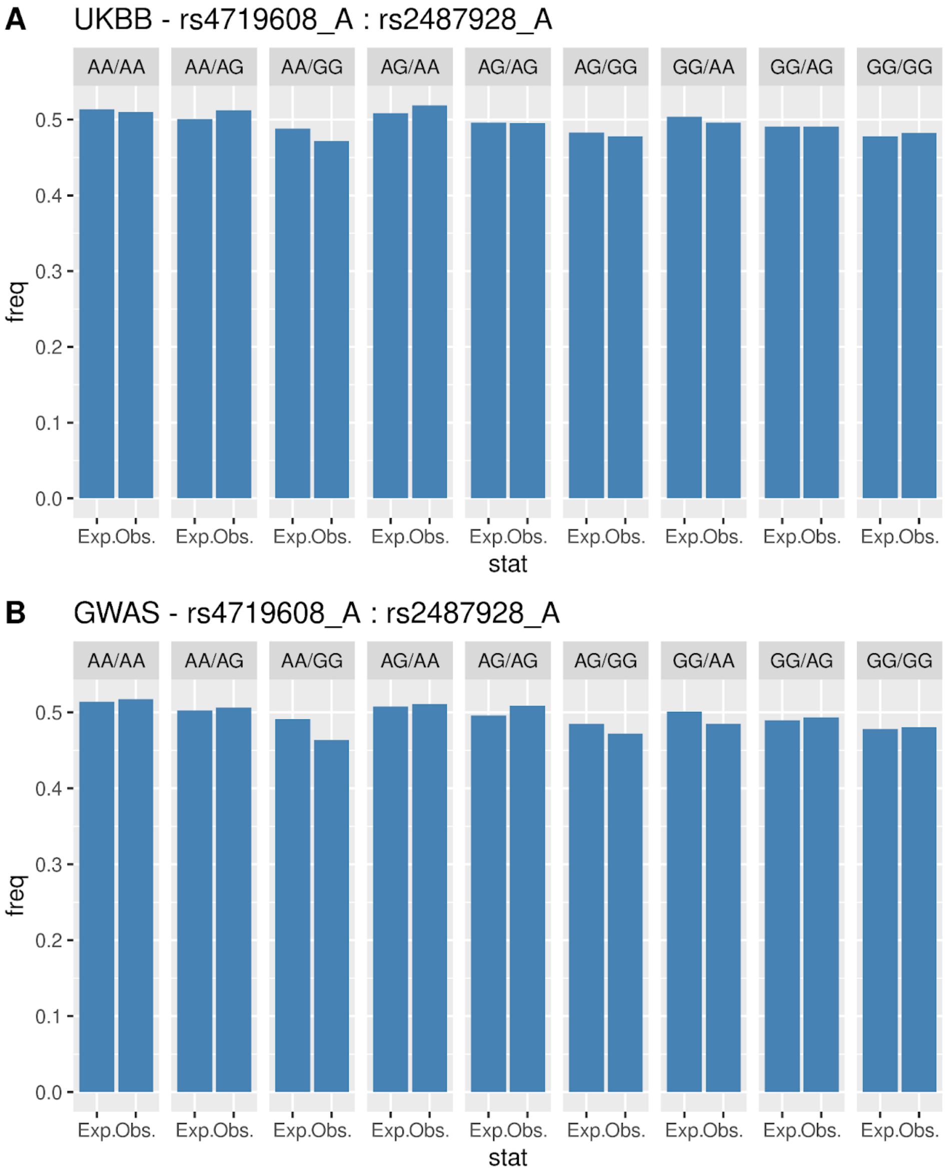
6.4.1. Supplementary Figures



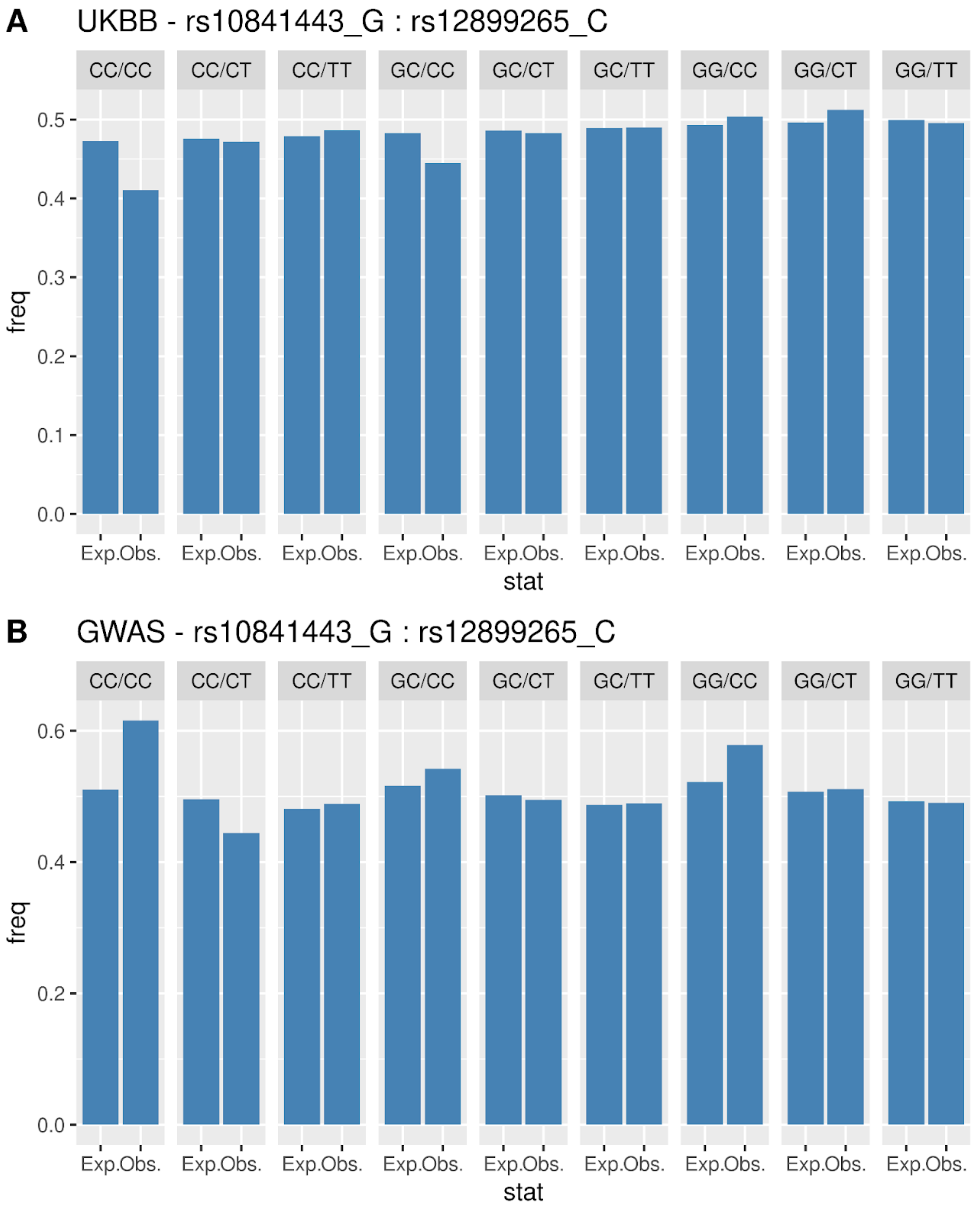
Supplementary Figure 4.1: The polygenic risk score (GWAS cohort). Please see the caption of main figure 2 for more details.



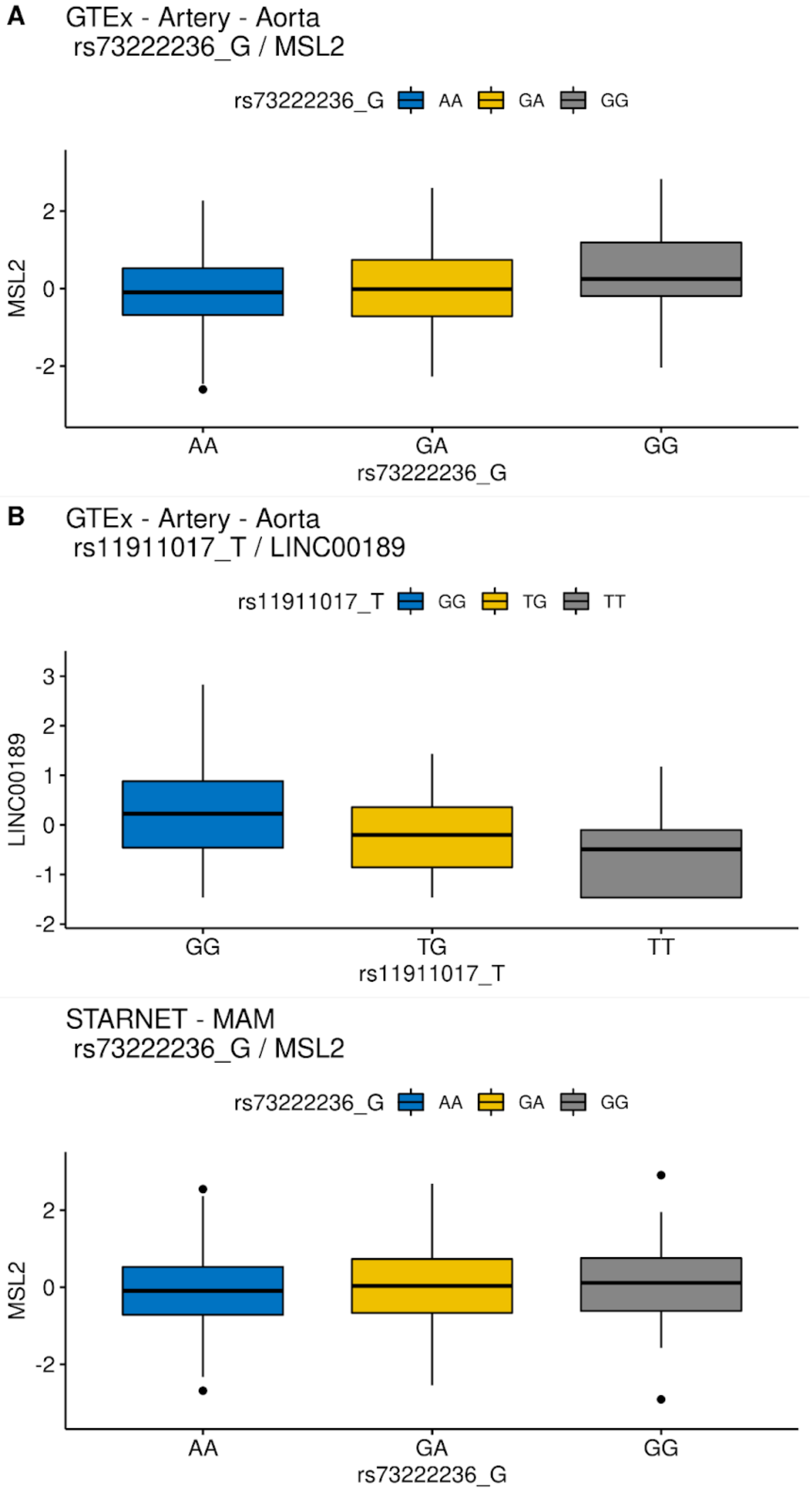
Supplementary Figure 4.2: Genotype-combination dependent case frequencies. Frequencies (y-axes) of observed cases ("Obs") in specific genotype combinations (columns) of SNP pair rs72685791 - rs12202017 against expected frequencies of case estimates ("Exp") based on logistic regression models. Subplot A shows the results for UKBB and subplot B for the GWAS cohort.



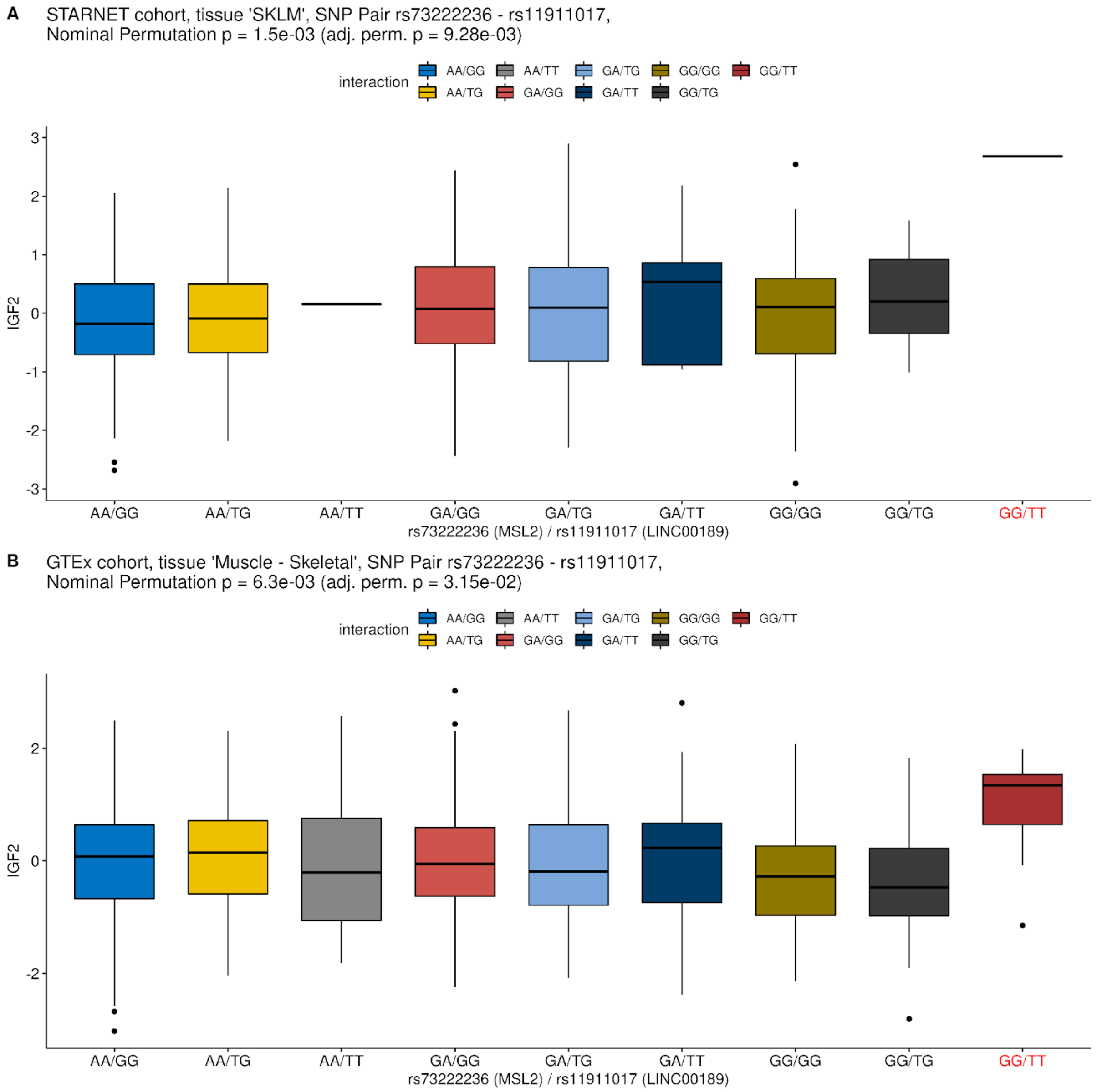
Supplementary Figure 4.3: Genotype-combination dependent case frequencies. Frequencies (y-axes) of observed cases (“Obs”) in specific genotype combinations (columns) of SNP pair rs4719608 - rs2487928 against expected frequencies of case estimates (“Exp”) based on logistic regression models. Subplot A shows the results for UKBB and subplot B for the GWAS cohort.



Supplementary Figure 4.4: Genotype-combination dependent case frequencies. Frequencies (y-axes) of observed cases ("Obs") in specific genotype combinations (columns) of SNP pair rs10841443 - rs12899265 against expected frequencies of case estimates ("Exp") based on logistic regression models. Subplot A shows the results for UKBB and subplot B for the GWAS cohort.



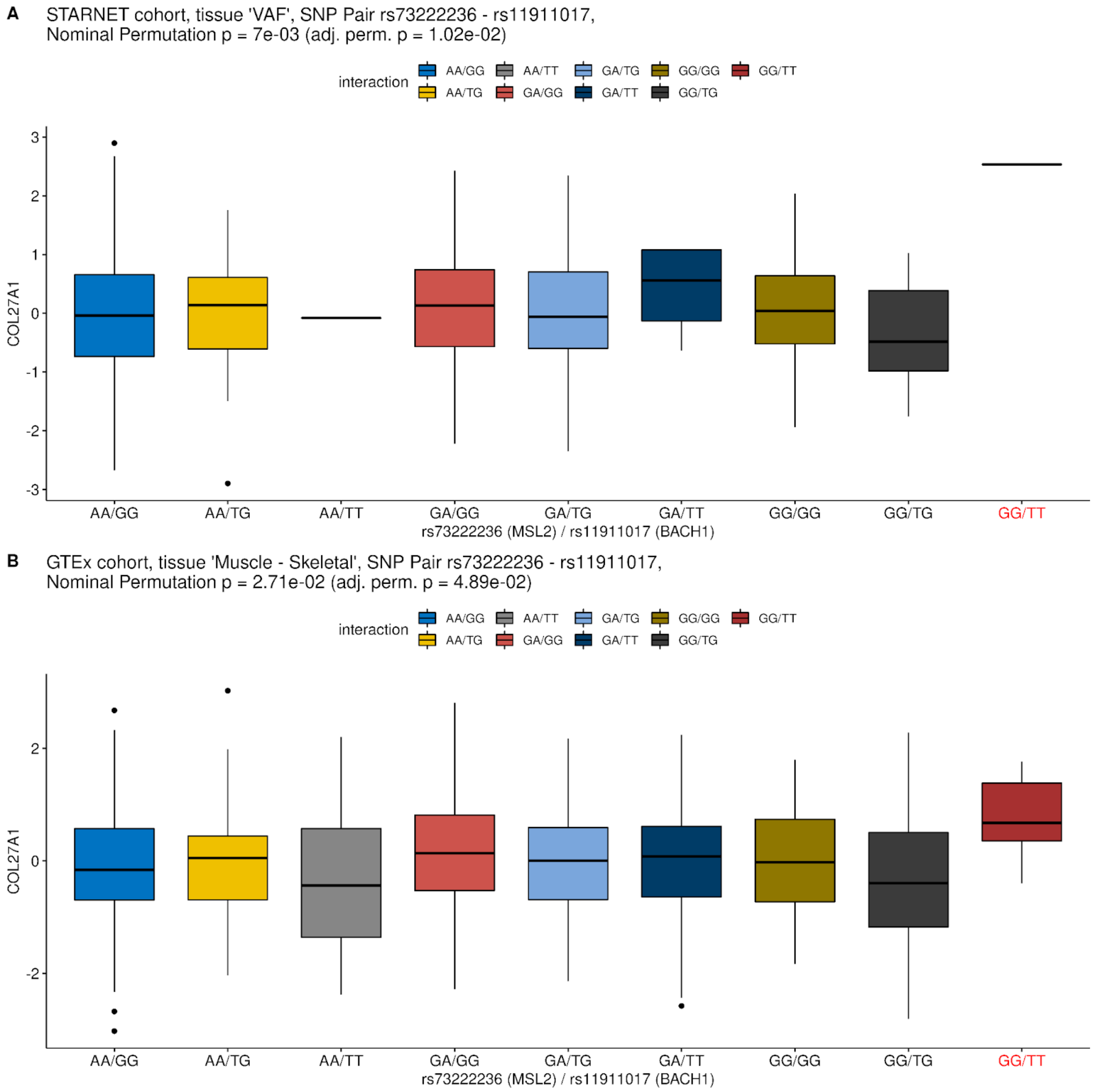
Supplementary Figure 4.5: Cis-QTL analysis results for rs73222236 and rs11911017. Expression profiles (y-axes) of cis (<1Mbp) genes MSL2 (A) and LINC00189 (B) in each cohort dependent on rs73222236 and rs11911017 genotypes (x-axes). LINC00189 is not expressed in the STARNET cohort, thus missing in subplot A.



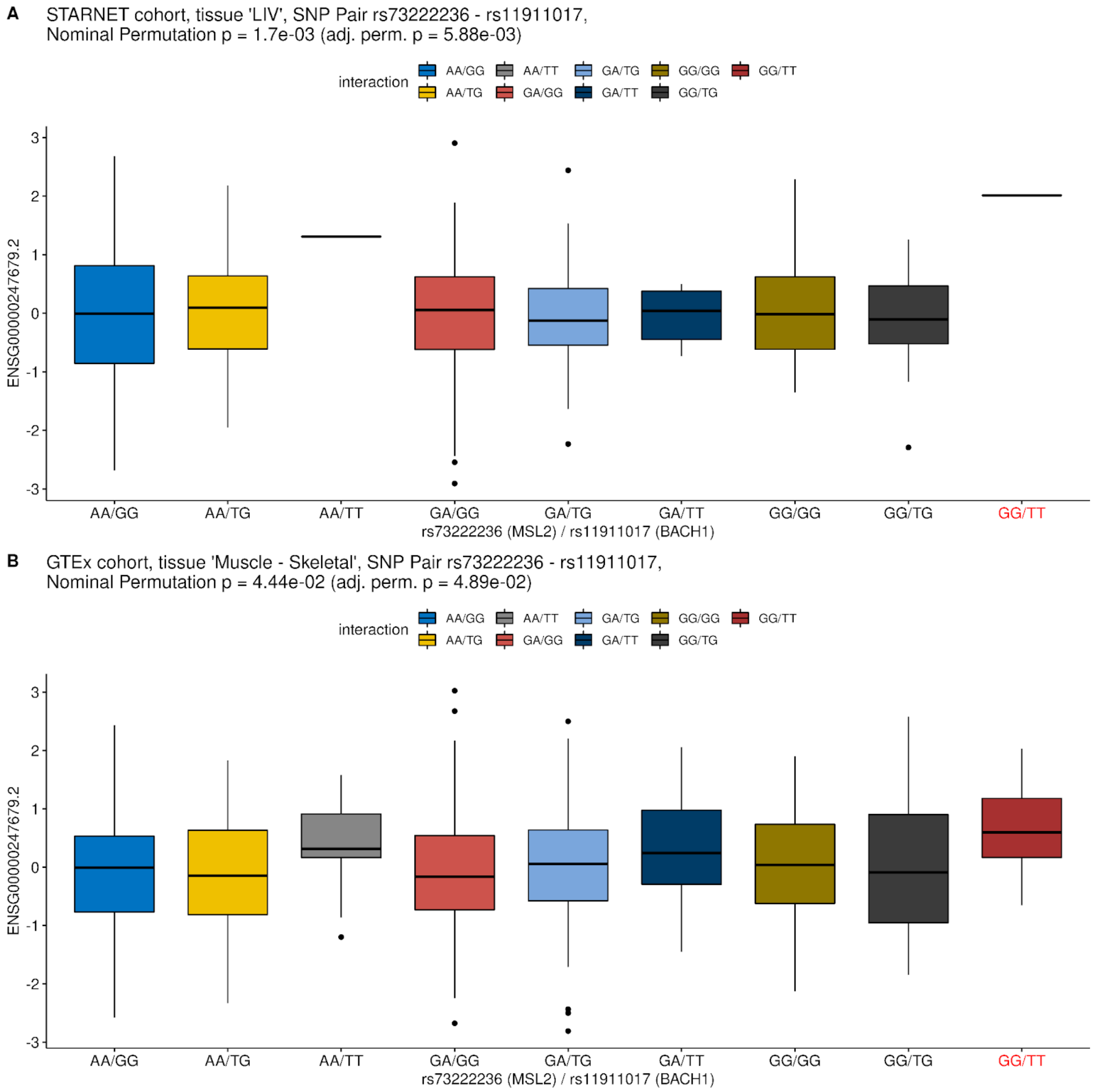
A STARNET cohort, tissue 'SKLM', SNP Pair rs73222236 - rs11911017, Nominal Permutation p = 1.5e-03 (adj. perm. p = 9.28e-03)

B GTEx cohort, tissue 'Muscle - Skeletal', SNP Pair rs73222236 - rs11911017, Nominal Permutation p = 6.3e-03 (adj. perm. p = 3.15e-02)

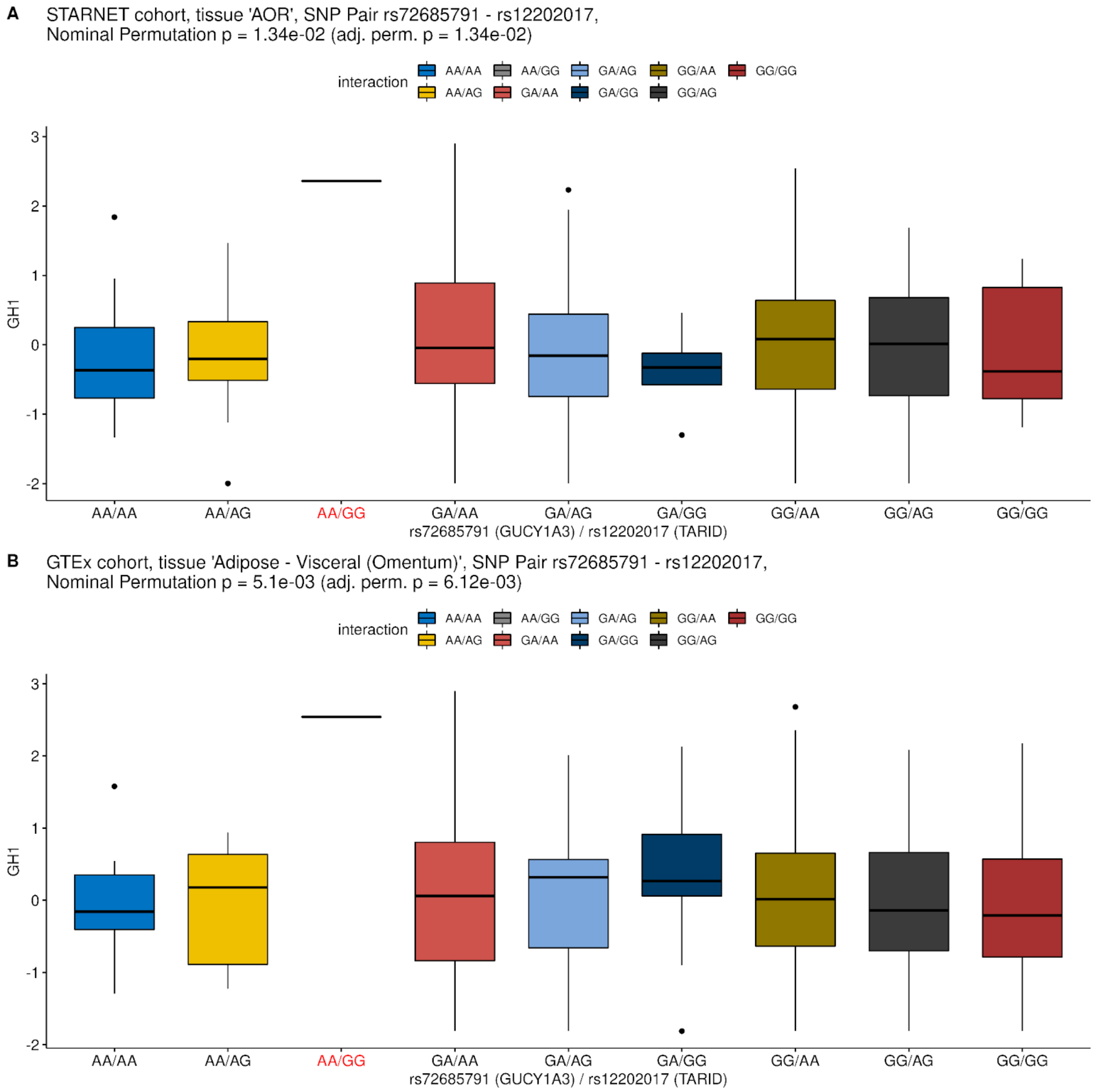
Supplementary Figure 4.6: Genotype combination dependent differential expression of IGF2. The GG/TT genotype combination of SNP pair rs73222236 and rs11911017 (x-axis, highlighted in red) leads to the significant (alpha=0.05) differential expression (y-axis) of trans-target gene IGF2 in skeletal muscle (SKLM) tissue in STARNET (A) and skeletal muscle tissue in GTEx (B).



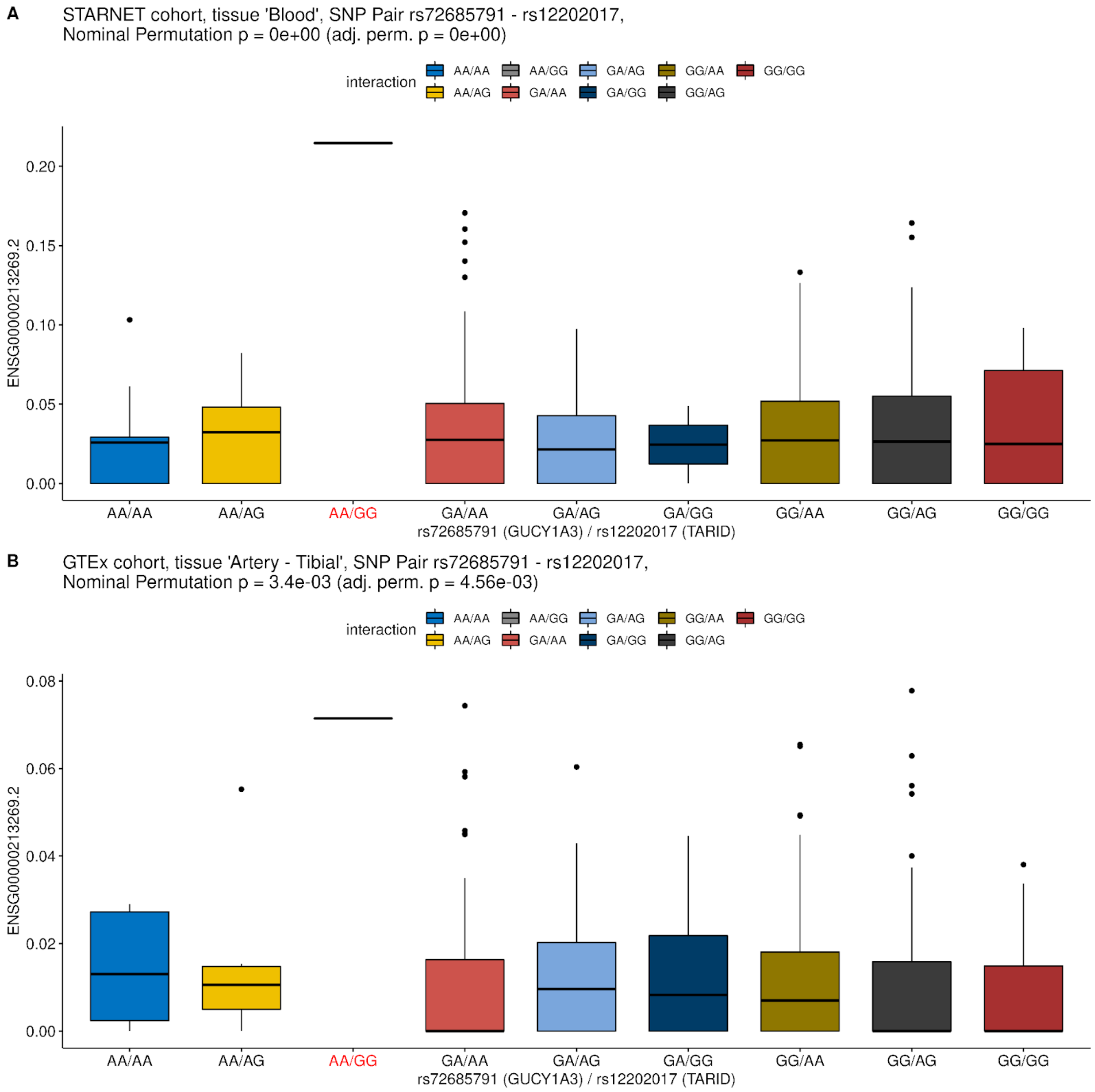
Supplementary Figure 4.7: Genotype combination dependent differential expression of COL27A43. The GG/TT genotype combination of SNP pair rs73222236 and rs11911017 (x-axis, highlighted in red) leads to the significant (alpha=0.05) differential expression (y-axis) of trans-target gene COL27A43 in visceral abdominal fat (VAF) tissue in STARNET (A) and skeletal muscle tissue in GTEx (B).



Supplementary Figure 4.8: Genotype combination dependent differential expression of ENSG00000247679. The GG/TT genotype combination of SNP pair rs73222236 and rs11911017 (x-axis, highlighted in red) leads to the significant (alpha=0.05) differential expression (y-axis) of trans-target gene ENSG00000247679 in liver (LIV) tissue in STARNET (A) and skeletal muscle tissue in GTEx (B).



Supplementary Figure 4.9: Genotype combination dependent differential expression of GH1. The AA/GG genotype combination of SNP pair rs72685791 and rs12202017 (x-axis, highlighted in red) leads to the significant (alpha=0.05) differential expression (y-axis) of trans-target gene GH1 in aortic artery (AOR) tissue in STARNET (A) and adipose visceral (omentum) tissue in GTEx (B).



Supplementary Figure 4.10: Genotype combination dependent differential expression of ENSG00000213269. The AA/GG genotype combination of SNP pair rs72685791 and rs12202017 (x-axis, highlighted in red) leads to the significant (alpha=0.05) differential expression (y-axis) of trans-target gene ENSG00000213269 in aortic artery (AOR) tissue in STARNET (A) and liver tissue in GTEx (B).

6.4.2. Supplementary Tables

Please access the Zenodo repository with DOI:10.5281/zenodo.6687913 to access the supplementary tables mentioned in the following.

Supplementary Table 4.1 (see xls file sheet "S1 CAD Risk Loci"): Metadata of CAD risk loci.

Supplementary Table 4.2 (see xls file sheet "S2 Interchromosomal SNP Pairs"): List of all interchromosomal SNP pairs.

Supplementary Table 4.3 (see xls file sheet "S3 Top 1% SNP Pairs"): Top 1% candidate SNP interactions derived from the permutation testing of SNP correlation differences between cases and controls.

Supplementary Table 4.4 (see xls file sheet "S4 Gene Expression Metadata"): Table of the number of genes and samples per tissue in gene expression data sets of the STARNET and GTEx v8 cohort. The STARNET cohort consisted of tissues from blood (Blood), atherosclerotic-lesion-free internal mammary artery (MAM), atherosclerotic aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM) and liver (LIV). The GTEx cohort consisted of tissues 'Adipose-Subcutaneous', 'Adipose-Visceral-Omentum', 'Artery-Tibial', 'Artery-Aorta', 'Artery-Coronary', 'Heart_Atrial-Appendage', 'Heart-Left-Ventricle', 'Liver', 'Muscle-Skeletal' and 'Whole-Blood'.

Supplementary Table S4.5 (see xls file sheet "S5 Trans Analysis Results"): Detailed results of the interacting SNP-pair associated genotype-combination dependent differential trans gene expression analyses. Coulmn 'cohort' gives the cohort, 'snp.pair' the rs-identifiers of the interacting SNP pairs, 'genotype.combination' the specific genotype-combination of the interacting SNP pairs under which a trans gene is differentially expressed, 'tissue' the tissue in which the gene is differentially expressed, 'gene' the differentially expressed trans target, 'beta' the effect size from the linear regression model of the genotype-combination term, 'nominal.p' the nominal p-value of that term and lastly, 'adj.p' the adjusted p-value of the same term.

Supplementary Table S4.6 (see xls file sheet "S6 Confident Result Set"): See description of supplementary table 4.5, with the exception that the table only shows the detailed results for the n=6 most confident trans-differential expression results.

Supplementary Table S4.7 (see xls file sheet "S7 Cis-eQTL Results"): Cis-eQTL results for each cohort (column "cohort"), tissue (column "tissue") and individuals SNP (column "snp") of interacting SNP pairs. Column "glm.p" gives the p-value of the cis-effect of a loci on the expression of a cis-gene (<1Mbp; column "gene"), whereas column "t.test.p" gives the p-value of a t-test between homozygous carriers, essentially comparing the extremes of the genotype-dependent gene expression distribution.

6.4.3. Data & Code Availability

The code is available at https://github.com/heiniglab/cad_epistasis. The data from the German Myocardial Infarction Family Studies (GerMIFS) I (387), II (388), III (389), IV (379), V (390), VI (391), VII (392) cannot be shared publicly due to ethical and confidentiality reasons but can be shared on reasonable requests addressed to the corresponding authors. Data from the LUdwigshafen RIsk and Cardiovascular Health Study (LURIC) (393), Cardiogenics (CG) (394) (Dataset ID: EGAC00001000088), Wellcome Trust Case Control Consortium (WTCCC) (395), Myocardial Infarction Genetics Consortium (MIGEN) (396) (dbGaP accession phs000902.v1.p1) and Stockholm-Tartu Reverse Network Engineering Task (STARNET) studies (397, 398) (dbGaP accession phs001203.v1.p1) were provided by third parties by permission and will be shared on request to the corresponding authors with permission of the third party. Access to the UK Biobank [(386)] can be requested on their website and was granted under the project ID 25214. Gene expression and genotype data data from the Genotype-Tissue Expression (GTEx) project [(399)] were obtained from the GTEx Portal with the project ID 20848.

7. Bibliography

1. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G., *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

2. Wang,G., Wang,F., Huang,Q., Li,Y., Liu,Y. and Wang,Y. (2015) Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. *Biomed Res. Int.*, **2015**, 757530.

3. Core,L. and Adelman,K. (2019) Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev.*, **33**, 960–982.

4. Rajala,T., Häkkinen,A., Healy,S., Yli-Harja,O. and Ribeiro,A.S. (2010) Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput. Biol.*, **6**, e1000704.

5. Malakar,A.K., Choudhury,D., Halder,B., Paul,P., Uddin,A. and Chakraborty,S. (2019) A review on coronary artery disease, its risk factors, and therapeutics. *J. Cell. Physiol.*, **234**, 16812–16823.

6. Singh,R.B., Niaz,M.A., Ghosh,S., Beegom,R., Chibo,H., Agarwal,P., Singh,R., Srivastav,S., Rastogi,S.S. and Postiglione,A. (1995) Epidemiological study of coronary artery disease and its risk factors in an elderly urban population of north India. *J. Am. Coll. Nutr.*, **14**, 628–634.

7. Ralapanawa,U. and Sivakanesan,R. (2021) Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review. *J. Epidemiol. Glob. Health*, **11**, 169–177.

8. Jamee Shahwan,A., Abed,Y., Desormais,I., Magne,J., Preux,P.M., Aboyans,V. and Lacroix,P. (2019) Epidemiology of coronary artery disease and stroke and associated risk factors in Gaza community -Palestine. *PLoS One*, **14**, e0211131.

9. Aday,A.W. and Matsushita,K. (2021) Epidemiology of Peripheral Artery Disease and Polyvascular Disease. *Circ. Res.*, **128**, 1818–1832.

10. Applegate,W.B., Hughes,J.P. and Vander Zwaag,R. (1991) Case-control study of coronary heart disease risk factors in the elderly. *J. Clin. Epidemiol.*, **44**, 409–415.

11. Jain,P., Jain,P., Bhandari,S. and Siddhu,A. (2008) A case-control study of risk factors for coronary heart disease in urban Indian middle-aged males. *Indian Heart J.*, **60**, 233–240.

12. Aggarwal,A., Aggarwal,S., Goel,A., Sharma,V. and Dwivedi,S. (2012) A retrospective case-control study of modifiable risk factors and cutaneous markers in Indian patients with young coronary artery disease. *JRSM Cardiovasc Dis*, **1**.

13. Nifina,N. and Krishnan,A. (2021) Association of bowel and tear suppression with coronary artery disease: A case control study. *J. Ayurveda Integr. Med.*, **12**, 80–86.

14. Li,C., Ma,R., Zhang,X., Ma,J., Wang,X., He,J., Zhang,J., Wang,K., Hu,Y., Pang,H., *et al.* (2020) Risk of coronary heart disease in the rural population in Xinjiang: A nested case-control study in China. *PLoS One*, **15**, e0229598.

15. Broadbent,H.M., Peden,J.F., Lorkowski,S., Goel,A., Ongen,H., Green,F., Clarke,R., Collins,R., Franzosi,M.G., Tognoni,G., *et al.* (2008) Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum. Mol. Genet.*, **17**, 806–814.

16. Yang,R., Li,L., Seidelmann,S.B., Shen,G.-Q., Sharma,S., Rao,S., Abdullah,K.G., Mackinlay,K.G., Elston,R.C., Chen,Q., *et al.* (2010) A genome-wide linkage scan identifies multiple quantitative trait loci for HDL-cholesterol levels in families with premature CAD and MI. *J. Lipid Res.*, **51**, 1442–1451.

17. Zhao,Q., Dacre,M., Nguyen,T., Pjanic,M., Liu,B., Iyer,D., Cheng,P., Wirka,R., Kim,J.B., Fraser,H.B., *et al.* (2020) Molecular mechanisms of coronary disease revealed using quantitative trait loci for TCF21 binding, chromatin accessibility, and chromosomal looping. *Genome Biol.*, **21**, 135.

18. Erdmann,J., Kessler,T., Munoz Venegas,L. and Schunkert,H. (2018) A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc. Res.*, **114**, 1241–1257.

19. van der Harst,P. and Verweij,N. (2018) Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.*, **122**, 433–443.

20. Verweij,N., Eppinga,R.N., Hagemeijer,Y. and van der Harst,P. (2017) Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci. Rep.*, **7**, 2761.

21. Wu,X., Lin,X., Li,Q., Wang,Z., Zhang,N., Tian,M., Wang,X., Deng,H. and Tan,H. (2020) Identification of novel SNPs associated with coronary artery disease and birth weight using a pleiotropic cFDR method. *Aging* , **13**, 3618–3644.

22. Huan,T., Joehanes,R., Song,C., Peng,F., Guo,Y., Mendelson,M., Yao,C., Liu,C., Ma,J., Richard,M., *et al.* (2019) Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat. Commun.*, **10**, 4267.

23. Yao,C., Chen,G., Song,C., Keefe,J., Mendelson,M., Huan,T., Sun,B.B., Laser,A., Maranville,J.C., Wu,H., *et al.* (2018) Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.*, **9**, 3268.

24. Peden,J.F. and Farrall,M. (2011) Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum. Mol. Genet.*, **20**, R198–205.

25. Farrall,M., Green,F.R., Peden,J.F., Olsson,P.G., Clarke,R., Hellenius,M.-L., Rust,S., Lagercrantz,J., Franzosi,M.G., Schulte,H., *et al.* (2006) Genome-wide mapping of susceptibility to coronary artery disease identifies a novel replicated locus on chromosome 17. *PLoS Genet.*, **2**, e72.

26. Krause,M.D., Huang,R.-T., Wu,D., Shentu,T.-P., Harrison,D.L., Whalen,M.B., Stolze,L.K., Di Rienzo,A., Moskowitz,I.P., Civelek,M., *et al.* (2018) Genetic variant at coronary artery disease and ischemic stroke locus 1p32.2 regulates endothelial responses to hemodynamics. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, E11349–E11358.

27. Jones,M.B., An,A., Shi,L.J. and Shi,W. (2020) Regional Variation in Genetic Control of Atherosclerosis in Hyperlipidemic Mice. *G3* , **10**, 4679–4689.

28. Findley,A.S., Richards,A.L., Petrini,C., Alazizi,A., Doman,E., Shanku,A.G., Davis,G.O., Hauff,N., Sorokin,Y., Wen,X., *et al.* (2019) Interpreting Coronary Artery Disease Risk Through Gene-Environment Interactions in Gene Regulation. *Genetics*, **213**, 651–663.

29. Mitchell,K.J. (2012) What is complex about complex disorders? *Genome Biol.*, **13**, 237.

30. Schork,N.J. (1997) Genetics of complex disease: approaches, problems, and solutions. *Am. J. Respir. Crit. Care Med.*, **156**, S103–9.

31. Maher,B. (2008) Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21.

32. Eichler,E.E., Flint,J., Gibson,G., Kong,A., Leal,S.M., Moore,J.H. and Nadeau,J.H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

33. McPherson,R. and Tybjaerg-Hansen,A. (2016) Genetics of Coronary Artery Disease. *Circ. Res.*, **118**, 564–578.

34. Khera,A.V., Chaffin,M., Aragam,K.G., Haas,M.E., Roselli,C., Choi,S.H., Natarajan,P., Lander,E.S., Lubitz,S.A., Ellinor,P.T., *et al.* (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, **50**, 1219–1224.

35. Clayton,D.G. (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.*, **5**, e1000540.

36. Evans,D.M., Spencer,C.C.A., Pointon,J.J., Su,Z., Harvey,D., Kochan,G., Oppermann,U., Dilthey,A., Pirinen,M., Stone,M.A., *et al.* (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.*, **43**, 761–767.

37. Moutsianas,L., Jostins,L., Beecham,A.H., Dilthey,A.T., Xifara,D.K., Ban,M., Shah,T.S., Patsopoulos,N.A., Alfredsson,L., Anderson,C.A., *et al.* (2015) Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat. Genet.*, **47**, 1107–1113.

38. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

39. Travers,A. and Muskhelishvili,G. (2015) DNA structure and function. *FEBS J.*, **282**, 2279–2295.

40. McGinty,R.K. and Tan,S. (2015) Nucleosome structure and function. *Chem. Rev.*, **115**, 2255–2273.

41. Cutter,A.R. and Hayes,J.J. (2015) A brief review of nucleosome structure. *FEBS Lett.*, **589**, 2914–2922.

42. Khorasanizadeh,S. (2004) The nucleosome: from genomic organization to genomic regulation. *Cell*, **116**, 259–272.

43. Chen,P., Li,W. and Li,G. (2021) Structures and Functions of Chromatin Fibers. *Annu. Rev. Biophys.*, **50**, 95–116.

44. Portin,P. and Wilkins,A. (2017) The Evolving Definition of the Term ‘Gene’. *Genetics*, **205**, 1353–1364.

45. Alphey,L.S., Crisanti,A., Randazzo,F.F. and Akbari,O.S. (2020) Opinion: Standardizing the definition of gene drive. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 30864–30867.

46. Deans,A.R., Lewis,S.E., Huala,E., Anzaldo,S.S., Ashburner,M., Balhoff,J.P., Blackburn,D.C., Blake,J.A., Burleigh,J.G., Chanet,B., *et al.* (2015) Finding our way through phenotypes. *PLoS Biol.*, **13**, e1002033.

47. Olokoba,A.B., Obateru,O.A. and Olokoba,L.B. (2012) Type 2 diabetes mellitus: a review of

current trends. *Oman Med. J.*, **27**, 269–273.

48. Hassanpour,S.H. and Dehghani,M. (2017) Review of cancer from perspective of molecular. *J. Cancer Surviv.*, **4**, 127–129.

49. Xu,C., Park,J.-K. and Zhang,J. (2019) Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS Biol.*, **17**, e3000197.

50. Yuan,F., Pan,X., Zeng,T., Zhang,Y.-H., Chen,L., Gan,Z., Huang,T. and Cai,Y.-D. (2020) Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data. *Front Bioeng Biotechnol*, **8**, 350.

51. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

52. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

53. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.

54. Klemm,S.L., Shipony,Z. and Greenleaf,W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, **20**, 207–220.

55. Ramazi,S., Allahverdi,A. and Zahiri,J. (2020) Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *J. Biosci.*, **45**.

56. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.

57. López-Maury,L., Marguerat,S. and Bähler,J. (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.*, **9**, 583–593.

58. Courey,A.J. and Jia,S. (2001) Transcriptional repression: the long and the short of it. *Genes Dev.*, **15**, 2786–2796.

59. Palazzo,A.F. and Lee,E.S. (2015) Non-coding RNA: what is functional and what is junk? *Front. Genet.*, **6**, 2.

60. Statello,L., Guo,C.-J., Chen,L.-L. and Huarte,M. (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, **22**, 96–118.

61. Tyagi,M., Imam,N., Verma,K. and Patel,A.K. (2016) Chromatin remodelers: We are the drivers!! *Nucleus*, **7**, 388–404.

62. Greber,B.J. and Nogales,E. (2019) The Structures of Eukaryotic Transcription Pre-initiation Complexes and Their Functional Implications. *Subcell. Biochem.*, **93**, 143–192.

63. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, **13**, 720–731.

64. Chen,F.X., Smith,E.R. and Shilatifard,A. (2018) Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **19**, 464–478.

65. Jonkers,I. and Lis,J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, **16**, 167–177.

66. Porrua,O. and Libri,D. (2015) Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.*, **16**, 190–202.

67. Wilkinson,M.E., Charenton,C. and Nagai,K. (2020) RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.*, **89**, 359–388.

68. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.

69. Kapp,L.D. and Lorsch,J.R. (2004) The molecular mechanics of eukaryotic translation. *Annu. Rev. Biochem.*, **73**, 657–704.

70. Das,S., Vera,M., Gandin,V., Singer,R.H. and Tutucci,E. (2021) Intracellular mRNA transport and localized translation. *Nat. Rev. Mol. Cell Biol.*, **22**, 483–504.

71. Baßler,J. and Hurt,E. (2019) Eukaryotic Ribosome Assembly. *Annu. Rev. Biochem.*, **88**, 281–306.

72. Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.

73. Roeder,R.G. and Rutter,W.J. (1969) Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, **224**, 234–237.

74. Sentenac,A. (1985) Eukaryotic RNA polymerases. *CRC Crit. Rev. Biochem.*, **18**, 31–90.

75. Fuda,N.J., Ardehali,M.B. and Lis,J.T. (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, **461**, 186–192.

76. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.

77. Schübeler,D. (2015) Function and information content of DNA methylation. *Nature*, **517**, 321–326.

78. Müller,F. and Tora,L. (2014) Chromatin and DNA sequences in defining promoters for transcription initiation. *Biochim. Biophys. Acta*, **1839**, 118–128.

79. Zhu,F., Farnung,L., Kaasinen,E., Sahu,B., Yin,Y., Wei,B., Dodonova,S.O., Nitta,K.R., Morgunova,E., Taipale,M., *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature*, **562**, 76–81.

80. Iwafuchi-Doi,M. and Zaret,K.S. (2016) Cell fate control by pioneer transcription factors. *Development*, **143**, 1833–1837.

81. Lorch,Y. and Kornberg,R.D. (2017) Chromatin-remodeling for transcription. *Q. Rev. Biophys.*, **50**, e5.

82. An,W., Palhan,V.B., Karymov,M.A., Leuba,S.H. and Roeder,R.G. (2002) Selective requirements for histone H3 and H4 N termini in p300-dependent transcriptional activation from chromatin. *Mol. Cell*, **9**, 811–821.

83. Banerji,J., Rusconi,S. and Schaffner,W. (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**, 299–308.

84. Kolovos,P., Knoch,T.A., Grosveld,F.G., Cook,P.R. and Papantonis,A. (2012) Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin*, **5**, 1.

85. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.

86. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J., *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.

87. Fant,C.B., Levandowski,C.B., Gupta,K., Maas,Z.L., Moir,J., Rubin,J.D., Sawyer,A., Esbin,M.N., Rimel,J.K., Luyties,O., *et al.* (2020) TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol. Cell*, **78**, 785–793.e8.

88. Vermeulen,M., Mulder,K.W., Denissov,S., Pijnappel,W.W.M.P., van Schaik,F.M.A., Varier,R.A., Baltissen,M.P.A., Stunnenberg,H.G., Mann,M. and Timmers,H.T.M. (2007) Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, **131**, 58–69.

89. Egly,J.-M. and Coin,F. (2011) A history of TFIIH: two decades of molecular biology on a pivotal transcription/repair factor. *DNA Repair* , **10**, 714–721.

90. Eick,D. and Geyer,M. (2013) The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem. Rev.*, **113**, 8456–8490.

91. Kornberg,R.D. (2005) Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.*, **30**, 235–239.

92. Bentley,D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.*, **15**, 163–175.

93. Gnatt,A.L., Cramer,P., Fu,J., Bushnell,D.A. and Kornberg,R.D. (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science*, **292**, 1876–1882.

94. Landick,R. (2006) The regulatory roles and mechanism of transcriptional pausing. *Biochem. Soc. Trans.*, **34**, 1062–1066.

95. Vos,S.M., Farnung,L., Urlaub,H. and Cramer,P. (2018) Structure of paused transcription complex Pol II-DSIF-NELF. *Nature*, **560**, 601–606.

96. Bernecky,C., Plitzko,J.M. and Cramer,P. (2017) Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp. *Nat. Struct. Mol. Biol.*, **24**, 809–815.

97. Palangat,M., Renner,D.B., Price,D.H. and Landick,R. (2005) A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15036–15041.

98. Kettenberger,H., Armache,K.-J. and Cramer,P. (2003) Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell*, **114**, 347–357.

99. Conaway,J.W., Shilatifard,A., Dvir,A. and Conaway,R.C. (2000) Control of elongation by RNA polymerase II. *Trends Biochem. Sci.*, **25**, 375–380.

100. Cheung,A.C.M. and Cramer,P. (2011) Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature*, **471**, 249–253.

101. Marshall,N.F. and Price,D.H. (1995) Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.*, **270**, 12335–12338.

102. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R.A., McCuine,S., Burge,C.B., Sharp,P.A. and Young,R.A.

(2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–445.

103. Li,Y., Liu,M., Chen,L.-F. and Chen,R. (2018) P-TEFb: Finding its ways to release promoter-proximally paused RNA polymerase II. *Transcription*, **9**, 88–94.

104. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

105. Kwak,H., Fuda,N.J., Core,L.J. and Lis,J.T. (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, **339**, 950–953.

106. Endoh,M., Zhu,W., Hasegawa,J., Watanabe,H., Kim,D.-K., Aida,M., Inukai,N., Narita,T., Yamada,T., Furuya,A., *et al.* (2004) Human Spt6 stimulates transcription elongation by RNA polymerase II in vitro. *Mol. Cell. Biol.*, **24**, 3324–3336.

107. Belotserkovskaya,R., Oh,S., Bondarenko,V.A., Orphanides,G., Studitsky,V.M. and Reinberg,D. (2003) FACT facilitates transcription-dependent nucleosome alteration. *Science*, **301**, 1090–1093.

108. Orphanides,G., LeRoy,G., Chang,C.H., Luse,D.S. and Reinberg,D. (1998) FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell*, **92**, 105–116.

109. Chen,F.X., Woodfin,A.R., Gardini,A., Rickels,R.A., Marshall,S.A., Smith,E.R., Shiekhattar,R. and Shilatifard,A. (2015) PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. *Cell*, **162**, 1003–1015.

110. Smolle,M., Workman,J.L. and Venkatesh,S. (2013) reSETting chromatin during transcription elongation. *Epigenetics*, **8**, 10–15.

111. Zraly,C.B. and Dingwall,A.K. (2012) The chromatin remodeling and mRNA splicing functions of the Brahma (SWI/SNF) complex are mediated by the SNR1/SNF5 regulatory subunit. *Nucleic Acids Res.*, **40**, 5975–5987.

112. Schwartz,S. and Ast,G. (2010) Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO J.*, **29**, 1629–1636.

113. Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP (2018) *Mol. Cell*, **69**, 648–663.e7.

114. RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing (2018) *Mol. Cell*, **72**, 369–379.e4.

115. Caizzi,L., Monteiro-Martins,S., Schwalb,B., Lysakovskaia,K., Schmitzova,J., Sawicka,A., Chen,Y., Lidschreiber,M. and Cramer,P. (2021) Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell*, **81**, 1920–1934.e9.

116. Rosonina,E., Kaneko,S. and Manley,J.L. (2006) Terminating the transcript: breaking up is hard to do. *Genes Dev.*, **20**, 1050–1056.

117. Vo,T.V., Dhakshnamoorthy,J., Larkin,M., Zofall,M., Thillainadesan,G., Balachandran,V., Holla,S., Wheeler,D. and Grewal,S.I.S. (2019) CPF Recruitment to Non-canonical Transcription Termination Sites Triggers Heterochromatin Assembly and Gene Silencing. *Cell Rep.*, **28**, 267–281.e5.

118. Varela,M.A. and Amos,W. (2010) Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. *Genomics*, **95**, 151–159.

119. Lewis,C.M. and Vassos,E. (2020) Polygenic risk scores: from research tools to clinical instruments. *Genome Med.*, **12**, 1–11.

120. Van de Peer,Y., Mizrachi,E. and Marchal,K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.*, **18**, 411–424.

121. Cano-Gamez,E. and Trynka,G. (2020) From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.*, **11**, 424.

122. Fu,Y., Tessneer,K.L., Li,C. and Gaffney,P.M. (2018) From association to mechanism in complex disease genetics: the role of the 3D genome. *Arthritis Res. Ther.*, **20**, 216.

123. Hunter,D.J. (2005) Gene-environment interactions in human diseases. *Nat. Rev. Genet.*, **6**, 287–298.

124. Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.*, **3**, 779–789.

125. Vihinen,M. (2018) Systematics for types and effects of DNA variations. *BMC Genomics*, **19**, 974.

126. Nakamura,Y. (2009) DNA variations in human and medical genetics: 25 years of my experience. *J. Hum. Genet.*, **54**, 1–8.

127. Trindade Maia,R. and de Araújo Campos,M. (2021) Introductory chapter: Genetic variation - the source of biological diversity. In *Genetic Variation*. IntechOpen.

128. Saclier,N., Chardon,P., Malard,F., Konecny-Dupré,L., Eme,D., Bellec,A., Breton,V., Duret,L., Lefebure,T. and Douady,C.J. (2020) Bedrock radioactivity influences the rate and spectrum of mutation. *Elife*, **9**.

129. Schwarzacher,T. (2003) Meiosis, recombination and chromosomes: a review of gene isolation and fluorescent in situ hybridization data in plants. *J. Exp. Bot.*, **54**, 11–23.

130. Slatkin,M. (1981) Estimating levels of gene flow in natural populations. *Genetics*, **99**, 323–335.

131. Sherman,M.A., Yaari,A.U., Priebe,O., Dietlein,F., Loh,P.-R. and Berger,B. (2022) Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat. Biotechnol.*, 10.1038/s41587-022-01353-8.

132. Cagan,A., Baez-Ortega,A., Brzozowska,N., Abascal,F., Coorens,T.H.H., Sanders,M.A., Lawson,A.R.J., Harvey,L.M.R., Bhosle,S., Jones,D., *et al.* (2022) Somatic mutation rates scale with lifespan across mammals. *Nature*, **604**, 517–524.

133. Joshi,M., Kapopoulou,A. and Laurent,S. (2021) Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective. *Front. Genet.*, **12**, 660899.

134. Perdomo-Sabogal,Á. and Nowick,K. (2019) Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease. *Genome Biol. Evol.*, **11**, 2178–2193.

135. Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.

136. Storey,J.D., Madeoy,J., Strout,J.L., Wurfel,M., Ronald,J. and Akey,J.M. (2007) Gene-expression variation within and among human populations. *Am. J. Hum. Genet.*, **80**, 502–509.

137. Williams,R.B.H., Chan,E.K.F., Cowley,M.J. and Little,P.F.R. (2007) The influence of genetic variation on gene expression. *Genome Res.*, **17**, 1707–1716.

138. Cheung,V.G. and Spielman,R.S. (2002) The genetics of variation in gene expression. *Nat. Genet.*, **32 Suppl**, 522–525.

139. Frazer,K.A., Murray,S.S., Schork,N.J. and Topol,E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.

140. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

141. Hood,L. and Rowen,L. (2013) The Human Genome Project: big science transforms biology and medicine. *Genome Med.*, **5**, 79.

142. Uffelmann,E., Huang,Q.Q., Munung,N.S., de Vries,J., Okada,Y., Martin,A.R., Martin,H.C., Lappalainen,T. and Posthuma,D. (2021) Genome-wide association studies. *Nature Reviews Methods Primers*, **1**, 1–21.

143. Xu,C., Tachmazidou,I., Walter,K., Ciampi,A., Zeggini,E., Greenwood,C.M.T. and UK10K Consortium (2014) Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.*, **38**, 281–290.

144. Jorde,L.B. (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res.*, **10**, 1435–1444.

145. Giral,H., Landmesser,U. and Kratzer,A. (2018) Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med*, **5**, 181.

146. Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.*, **102**, 717–730.

147. Devoto,M. and Falchi,M. (2012) Genetic mapping of quantitative trait loci for disease-related phenotypes. *Methods Mol. Biol.*, **871**, 281–311.

148. Nica,A.C. and Dermitzakis,E.T. (2013) Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **368**, 20120362.

149. He,B., Shi,J., Wang,X., Jiang,H. and Zhu,H.-J. (2020) Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. *BMC Biol.*, **18**, 97.

150. Hawe,J.S., Wilson,R., Schmid,K.T., Zhou,L., Lakshmanan,L.N., Lehne,B.C., Kühnel,B., Scott,W.R., Wielscher,M., Yew,Y.W., *et al.* (2022) Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat. Genet.*, **54**, 18–29.

151. Pelikan,R.C., Kelly,J.A., Fu,Y., Lareau,C.A., Tessneer,K.L., Wiley,G.B., Wiley,M.M., Glenn,S.B., Harley,J.B., Guthridge,J.M., *et al.* (2018) Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. *Nat. Commun.*, **9**, 2905.

152. Tehranchi,A., Hie,B., Dacre,M., Kaplow,I., Pettie,K., Combs,P. and Fraser,H.B. (2019) Fine-mapping cis-regulatory variants in diverse human populations. *Elife*, **8**.

153. Albert,F.W., Bloom,J.S., Siegel,J., Day,L. and Kruglyak,L. (2018) Genetics of trans-regulatory variation in gene expression. *Elife*, **7**.

154. Golan,D., Lander,E.S. and Rosset,S. (2014) Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, E5272–81.

155. Crouch,D.J.M. and Bodmer,W.F. (2020) Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 18924–18933.

156. Choi,S.W., Mak,T.S.H. and O'Reilly,P.F. (2018) A guide to performing Polygenic Risk Score analyses. *bioRxiv*, 10.1101/416545.

157. Konuma,T. and Okada,Y. (2021) Statistical genetics and polygenic risk score for precision medicine. *Inflamm. Regen.*, **41**, 18.

158. Wray,N.R., Goddard,M.E. and Visscher,P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, **17**, 1520–1528.

159. Phillips,P.C. (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.

160. Domingo,J., Baeza-Centurion,P. and Lehner,B. (2019) The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.*, **20**, 433–460.

161. Mackay,T.F.C. (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **15**, 22–33.

162. Spiezio,S.H., Takada,T., Shiroishi,T. and Nadeau,J.H. (2012) Genetic divergence and the genetic architecture of complex traits in chromosome substitution strains of mice. *BMC Genet.*, **13**, 38.

163. Hill,W.G., Goddard,M.E. and Visscher,P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.*, **4**, e1000008.

164. Randolph,H.E., Mu,Z., Fiege,J.K., Thielen,B.K., Grenier,J.-C., Cobb,M.S., Hussin,J.G., Li,Y.I., Langlois,R.A. and Barreiro,L.B. (2020) Single-cell RNA-sequencing reveals pervasive but highly cell type-specific genetic ancestry effects on the response to viral infection. *bioRxiv*, 10.1101/2020.12.21.423830.

165. Nédélec,Y., Sanz,J., Baharian,G., Szpiech,Z.A., Pacis,A., Dumaine,A., Grenier,J.-C., Freiman,A., Sams,A.J., Hebert,S., *et al.* (2016) Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, **167**, 657–669.e21.

166. Young,A.I., Wauthier,F. and Donnelly,P. (2016) Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nat. Commun.*, **7**, 12724.

167. Takeshita,T., Mao,X.Q. and Morimoto,K. (1996) The contribution of polymorphism in the alcohol dehydrogenase beta subunit to alcohol sensitivity in a Japanese population. *Hum. Genet.*, **97**, 409–413.

168. Kilpeläinen,T.O., Qi,L., Brage,S., Sharp,S.J., Sonestedt,E., Demerath,E., Ahmad,T., Mora,S., Kaakinen,M., Sandholt,C.H., *et al.* (2011) Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.*, **8**, e1001116.

169. Patel,R.A., Musharoff,S.A., Spence,J.P., Pimentel,H., Tcheandjieu,C., Mostafavi,H., Sinnott-Armstrong,N., Clarke,S.L., Smith,C.J., Durda,P.P., *et al.* (2022) Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet.*, 10.1016/j.ajhg.2022.05.014.

170. Kanai,M., Tanaka,T. and Okada,Y. (2016) Empirical estimation of genome-wide significance

thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.*, **61**, 861–866.

171. Nobre,R., Ilic,A., Santander-Jimenez,S. and Sousa,L. (2021) Retargeting tensor accelerators for epistasis detection. *IEEE Trans. Parallel Distrib. Syst.*, **32**, 2160–2174.

172. Niel,C., Sinoquet,C., Dina,C. and Rocheleau,G. (2015) A survey about methods dedicated to epistasis detection. *Front. Genet.*, **0**.

173. Moore,J.H. and Andrews,P.C. (2015) Epistasis analysis using multifactor dimensionality reduction. *Methods Mol. Biol.*, **1253**, 301–314.

174. Wan,X., Yang,C., Yang,Q., Xue,H., Fan,X., Tang,N.L.S. and Yu,W. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.

175. Goudey,B., Rawlinson,D., Wang,Q., Shi,F., Ferra,H., Campbell,R.M., Stern,L., Inouye,M.T., Ong,C.S. and Kowalczyk,A. (2013) GWIS--model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics*, **14 Suppl 3**, S10.

176. Jiang,R., Tang,W., Wu,X. and Fu,W. (2009) A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, **10 Suppl 1**, S65.

177. Schwarz,D.F., König,I.R. and Ziegler,A. (2010) On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, **26**, 1752–1758.

178. Liu Yanlan and Luo Jiawei (2012) An improved Markov blanket approach to detect SNPs-disease associations in case-control studies. *Int. j. digit. content technol. appl.*, **6**, 278–286.

179. Wang,Y., Liu,X., Robbins,K. and Rekaya,R. (2010) AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Res. Notes*, **3**, 117.

180. Moore,J.H. and Hill,D.P. (2015) Epistasis analysis using artificial intelligence. *Methods Mol. Biol.*, **1253**, 327–346.

181. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

182. Marchini,J., Donnelly,P. and Cardon,L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

183. Moore,J.H. and Williams,S.M. (2009) Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, **85**, 309–320.

184. Wrappers for feature subset selection (1997) *Artif. Intell.*, **97**, 273–324.

185. Moore,J.H. and White,B.C. (2007) Tuning ReliefF for genome-wide genetic analysis. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 166–175.

186. McKinney,B.A., Reif,D.M., White,B.C., Crowe,J.E. and Moore,J.H. (2007) Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics*, **23**, 2113–2120.

187. Chikhi,S. and Benhammada,S. (2009) ReliefMSS: a variation on a feature ranking ReliefF algorithm. *Int. j. bus. intell. data min.*, **4**, 375.

188. Grady,B.J., Torstenson,E.S., McLaren,P.J., DE Bakker,P.I.W., Haas,D.W., Robbins,G.K.,

Gulick,R.M., Haubrich,R., Ribaudo,H. and Ritchie,M.D. (2011) Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in ART-naïve ACTG clinical trials participants. *Pac. Symp. Biocomput.*

189. Ritchie,M.D. (2015) Finding the epistasis needles in the genome-wide haystack. *Methods Mol. Biol.*, **1253**, 19–33.

190. Pattin,K.A. and Moore,J.H. (2008) Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.*, **124**, 19–29.

191. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N., *et al.* (2014) The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–63.

192. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B.-J., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F., *et al.* (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.

193. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

194. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrián-Uhalte,E., *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.

195. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

196. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R., *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

197. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Ma'ayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* , **2016**.

198. Chatelain,C., Durand,G., Thuillier,V. and Augé,F. (2018) Performance of epistasis detection methods in semi-simulated GWAS. *BMC Bioinformatics*, **19**, 1–17.

199. Upton,A., Trelles,O., Cornejo-García,J.A. and Perkins,J.R. (2015) Review: High-performance computing to detect epistasis in genome scale data sets. *Brief. Bioinform.*, **17**, 368–379.

200. Ponte-Fernandez,C., Gonzalez-Dominguez,J., Carvajal-Rodriguez,A. and Martin,M.J. (2020) Evaluation of Existing Methods for High-Order Epistasis Detection. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **PP**.

201. Zeng,L., Moser,S., Mirza-Schreiber,N., Lamina,C., Coassin,S., Nelson,C.P., Annilo,T., Franzén,O., Kleber,M.E., Mack,S., *et al.* (2021) Cis-epistasis at the LPA locus and risk of cardiovascular diseases. *Cardiovasc. Res.*, 10.1093/cvr/cvab136.

202. Li,Y., Cho,H., Wang,F., Canela-Xandri,O., Luo,C., Rawlik,K., Archacki,S., Xu,C., Tenesa,A., Chen,Q., *et al.* (2020) Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *J. Am. Heart Assoc.*, **9**, e014146.

203. Lu,W.-H., Zhang,W.-Q., Zhao,Y.-J., Gao,Y.-T., Tao,N., Ma,Y.-T., Liu,J.-W. and Wulasihan,M. (2020) Case-Control Study on the Interaction Effects of rs10757278 Polymorphisms at 9p21 Locus and Traditional Risk Factors on Coronary Heart Disease in Xinjiang, China. *J. Cardiovasc. Pharmacol.*, **75**, 439–445.

204. Priya,S. and Manavalan,R.K. (2020) Genetic interactions effects of cardiovascular disorder using computational models: A review. *Curr. Biotechnol.*, **9**, 177–191.

205. Koyama,S., Ito,K., Terao,C., Akiyama,M., Horikoshi,M., Momozawa,Y., Matsunaga,H., Ieki,H., Ozaki,K., Onouchi,Y., *et al.* (2020) Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.*, **52**, 1169–1177.

206. Wang,Q. (2005) Molecular genetics of coronary artery disease. *Curr. Opin. Cardiol.*, **20**, 182–188.

207. Sarker,I.H. (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput Sci*, **2**, 160.

208. Serre,T. (2019) Deep Learning: The Good, the Bad, and the Ugly. *Annu Rev Vis Sci*, **5**, 399–426.

209. Nadkarni,S. and Prügl,R. (2021) Digital transformation: a review, synthesis and opportunities for future research. *Management Review Quarterly*, **71**, 233–341.

210. Sah,S. (2020) Machine learning: A review of learning types. *Preprints*, 10.20944/preprints202007.0230.v1.

211. Bridges,C.C.,Jr (1966) Hierarchical cluster analysis. *Psychol. Rep.*, **18**, 851–854.

212. Fornalski,K.W. (2015) Applications of the robust Bayesian regression analysis. *Int. J. Soc. Syst. Sci.*, **7**, 314.

213. Seal,H.L. (1967) Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika*, **54**, 1–24.

214. Schneider,A., Hommel,G. and Blettner,M. (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch. Arztebl. Int.*, **107**, 776–782.

215. Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (2017) Classification and Regression Trees Chapman & Hall/CRC.

216. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

217. Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16. Association for Computing Machinery, New York, NY, USA, pp. 785–794.

218. Palm,G. (1986) Warren McCulloch and Walter Pitts: A Logical Calculus of the Ideas Immanent in Nervous Activity. In *Brain Theory*. Springer Berlin Heidelberg, pp. 229–230.

219. Schmidhuber,J. (2015) Deep learning in neural networks: an overview. *Neural Netw.*, **61**, 85–117.

220. Eraslan,G., Avsec,Ž., Gagneur,J. and Theis,F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, **20**, 389–403.

221. Li,Y., Wu,F.-X. and Ngom,A. (2018) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, **19**, 325–340.

222. Subramanian,I., Verma,S., Kumar,S., Jere,A. and Anamika,K. (2020) Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights*, **14**, 1177932219899051.

223. Lightbody,G., Haberland,V., Browne,F., Taggart,L., Zheng,H., Parkes,E. and Blayney,J.K. (2019) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.*, **20**, 1795–1811.

224. Schloss,J.A., Gibbs,R.A., Makhijani,V.B. and Marziali,A. (2020) Cultivating DNA Sequencing Technology After the Human Genome Project. *Annu. Rev. Genomics Hum. Genet.*, **21**, 117–138.

225. Levy,S.E. and Myers,R.M. (2016) Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.*, **17**, 95–115.

226. Kanzi,A.M., San,J.E., Chimukangara,B., Wilkinson,E., Fish,M., Ramsuran,V. and de Oliveira,T. (2020) Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. *Front. Genet.*, **11**, 544162.

227. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

228. Bioinformatics,B. (2011) FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*.

229. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

230. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

231. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

232. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

233. Zhao,Y., Li,M.-C., Konaté,M.M., Chen,L., Das,B., Karlovich,C., Williams,P.M., Evrard,Y.A., Doroshow,J.H. and McShane,L.M. (2021) TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J. Transl. Med.*, **19**, 269.

234. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T., *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15776–15781.

235. Kawaji,H., Lizio,M., Itoh,M., Kanamori-Katayama,M., Kaiho,A., Nishiyori-Sueki,H., Shin,J.W., Kojima-Ishiyama,M., Kawano,M., Murata,M., *et al.* (2014) Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.*, **24**, 708–717.

236. Nepal,C., Hadzhiev,Y., Previti,C., Haberle,V., Li,N., Takahashi,H., Suzuki,A.M.M., Sheng,Y., Abdelhamid,R.F., Anand,S., *et al.* (2013) Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.*, **23**, 1938–1950.

237. Gardini,A. (2017) Global Run-On Sequencing (GRO-Seq). *Methods Mol. Biol.*, **1468**, 111–120.

238. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

239. Xu,H., Handoko,L., Wei,X., Ye,C., Sheng,J., Wei,C.-L., Lin,F. and Sung,W.-K. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, **26**, 1199–1204.

240. Liang,K. and Keleş,S. (2012) Normalization of ChIP-seq data with control. *BMC Bioinformatics*, **13**, 199.

241. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

242. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P., *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

243. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K., *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.

244. Krakau,S., Richard,H. and Marsico,A. (2017) PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol.*, **18**, 240.

245. Lovci,M.T., Ghanem,D., Marr,H., Arnold,J., Gee,S., Parra,M., Liang,T.Y., Stark,T.J., Gehman,L.T., Hoon,S., *et al.* (2013) Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.*, **20**, 1434–1442.

246. Davey,J.W., Hohenlohe,P.A., Etter,P.D., Boone,J.Q., Catchen,J.M. and Blaxter,M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, **12**, 499–510.

247. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

248. Kim,S. and Misra,A. (2007) SNP genotyping: technologies and biomedical applications. *Annu. Rev. Biomed. Eng.*, **9**, 289–320.

249. Kwok,P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum. Genet.*, **2**, 235–258.

250. Meienberg,J., Zerjavic,K., Keller,I., Okoniewski,M., Patrignani,A., Ludin,K., Xu,Z., Steinmann,B., Carrel,T., Röthlisberger,B., *et al.* (2015) New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res.*, **43**, e76.

251. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

252. Garg,S. (2021) Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.*, **22**, 101.

253. Das,S., Abecasis,G.R. and Browning,B.L. (2018) Genotype Imputation from Large Reference

Panels. *Annu. Rev. Genomics Hum. Genet.*, **19**, 73–96.

254. Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.

255. Li,Y., Willer,C., Sanna,S. and Abecasis,G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.

256. Irizarry,R.A. (2019) Introduction to Data Science: Data Analysis and Prediction Algorithms with R CRC Press.

257. Wang,J. (2013) Pearson Correlation Coefficient. In Dubitzky,W., Wolkenhauer,O., Cho,K.-H., Yokota,H. (eds), *Encyclopedia of Systems Biology*. Springer New York, New York, NY, pp. 1671–1671.

258. Spearman Rank Correlation Coefficient (2008) In *The Concise Encyclopedia of Statistics*. Springer New York, New York, NY, pp. 502–505.

259. Kendall,M.G. (1938) A NEW MEASURE OF RANK CORRELATION. *Biometrika*, **30**, 81–93.

260. van Dam,S., Võsa,U., van der Graaf,A., Franke,L. and de Magalhães,J.P. (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, **19**, 575–592.

261. Zhang,Z., Chen,L., Xu,P., Xing,L., Hong,Y. and Chen,P. (2020) Gene correlation network analysis to identify regulatory factors in sepsis. *J. Transl. Med.*, **18**, 381.

262. Niu,X., Zhang,J., Zhang,L., Hou,Y., Pu,S., Chu,A., Bai,M. and Zhang,Z. (2019) Weighted Gene Co-Expression Network Analysis Identifies Critical Genes in the Development of Heart Failure After Acute Myocardial Infarction. *Front. Genet.*, **10**, 1214.

263. Paci,P., Fiscon,G., Conte,F., Wang,R.-S., Farina,L. and Loscalzo,J. (2021) Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *NPJ Syst Biol Appl*, **7**, 3.

264. He,B., Xu,J., Tian,Y., Liao,B., Lang,J., Lin,H., Mo,X., Lu,Q., Tian,G. and Bing,P. (2020) Gene Coexpression Network and Module Analysis across 52 Human Tissues. *Biomed Res. Int.*, **2020**, 6782046.

265. Bishop,C.M. (2006) Pattern Recognition and Machine Learning Springer.

266. Freeman,G.H. and Halton,J.H. (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, **38**, 141–149.

267. Pearson,K. (1900) X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**, 157–175.

268. STUDENT (1908) The probable error of a mean. *Biometrika*, **6**, 1–25.

269. Kaufmann,J. and Schering,A.G. (2014) Analysis of Variance ANOVA. *Wiley StatsRef: Statistics Reference Online*, 10.1002/9781118445112.stat06938.

270. S. Pillai,K.C. (2014) Multivariate analysis of variance (MANOVA). *Wiley StatsRef: Statistics Reference Online*, 10.1002/9781118445112.stat02476.

271. Philippas,D. (2014) Analysis of Covariance (ANCOVA). In Michalos,A.C. (ed), *Encyclopedia of Quality of Life and Well-Being Research*. Springer Netherlands, Dordrecht, pp. 157–161.

272. Huberty,C.J. and Petoskey,M.D. (2000) 7 - Multivariate Analysis of Variance and Covariance. In Tinsley,H.E.A., Brown,S.D. (eds), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press, San Diego, pp. 183–208.

273. Chen,S.-Y., Feng,Z. and Yi,X. (2017) A general introduction to adjustment for multiple comparisons. *J. Thorac. Dis.*, **9**, 1725–1729.

274. Etymologia: Bonferroni correction (2015) *Emerg. Infect. Dis.*, **21**, 289.

275. Bartroff,J. and Song,J. (2020) Sequential Tests of Multiple Hypotheses Controlling False Discovery and Nondiscovery Rates. *Seq. Anal.*, **39**, 65–91.

276. Benjamini,Y., Bretz,F. and Sarkar,S.K. (2004) Recent Developments in Multiple Comparison Procedures IMS.

277. Murphy,K.P. (2012) Machine Learning: A Probabilistic Perspective MIT Press.

278. Chai,T. and Draxler,R.R. (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.*, **7**, 1247–1250.

279. Ruder,S. (2016) An overview of gradient descent optimization algorithms. *arXiv [cs.LG]*.

280. Polyak,B.T. (2007) Newton's method and its use in optimization. *Eur. J. Oper. Res.*, **181**, 1086–1096.

281. Raileanu,L.E. and Stoffel,K. (2004) Theoretical Comparison between the Gini Index and Information Gain Criteria. *Ann. Math. Artif. Intell.*, **41**, 77–93.

282. Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

283. Natekin,A. and Knoll,A. (2013) Gradient boosting machines, a tutorial. *Front. Neurorobot.*, **7**, 21.

284. Rahim,K. and Hachaïchi,Y. (2019) Taylor approximation : From history to teaching. 10.13140/RG.2.2.26154.77761.

285. Friedman,J.H. (2001) Greedy function approximation: A gradient boosting machine. *aos*, **29**, 1189–1232.

286. Lundberg,S.M., Erion,G., Chen,H., DeGrave,A., Prutkin,J.M., Nair,B., Katz,R., Himmelfarb,J., Bansal,N. and Lee,S.-I. (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, **2**, 56–67.

287. Belle,V. and Papantonis,I. (2021) Principles and Practice of Explainable Machine Learning. *Front Big Data*, **4**, 688969.

288. Ribeiro,M.T., Singh,S. and Guestrin,C. (2016) 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *arXiv [cs.LG]*.

289. Lundberg,S. and Lee,S.-I. (2017) A Unified Approach to Interpreting Model Predictions.

290. Shrikumar,A., Greenside,P. and Kundaje,A. (2017) Learning Important Features Through Propagating Activation Differences. *arXiv [cs.CV]*.

291. Doroudi,S. (2020) The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates. *AERA Open*, **6**, 2332858420977208.

292. James,G.M. (2003) Variance and Bias for General Loss Functions. *Mach. Learn.*, **51**, 115–135.

293. Kohavi,R. and Others (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*. Montreal, Canada, Vol. 14, pp. 1137–1145.

294. Feurer,M. and Hutter,F. (2019) Hyperparameter Optimization. In Hutter,F., Kotthoff,L., Vanschoren,J. (eds), *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing, Cham, pp. 3–33.

295. Ng,A.Y. (2004) Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04. Association for Computing Machinery, New York, NY, USA, p. 78.

296. Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **67**, 301–320.

297. Gressel,S., Schwalb,B. and Cramer,P. (2019) The pause-initiation limit restricts transcription activation in human cells. *Nat. Commun.*, **10**, 1–12.

298. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

299. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K., *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

300. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J., *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.

301. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–45.

302. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.

303. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

304. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K., *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

305. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J., *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.

306. Noguchi,S., Arakawa,T., Fukuda,S., Furuno,M., Hasegawa,A., Hori,F., Ishikawa-Kato,S., Kaida,K., Kaiho,A., Kanamori-Katayama,M., *et al.* (2017) FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data*, **4**, 1–10.

307. Frith,M.C., Valen,E., Krogh,A., Hayashizaki,Y., Carninci,P. and Sandelin,A. (2008) A code for transcription initiation in mammalian genomes. *Genome Res.*, **18**, 1–12.

308. Lopes,R., Agami,R. and Korkmaz,G. (2017) GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. *Methods Mol. Biol.*, **1543**, 45–55.

309. Day,D.S., Zhang,B., Stevens,S.M., Ferrari,F., Larschan,E.N., Park,P.J. and Pu,W.T. (2016) Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biol.*, **17**, 120.

310. Shao,W. and Zeitlinger,J. (2017) Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.*, **49**, 1045–1051.

311. Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

312. Li,Q., Brown,J.B., Huang,H. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

313. Salmena,L., Poliseno,L., Tay,Y., Kats,L. and Pandolfi,P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.

314. Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.

315. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

316. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

317. Luo,Z., Lin,C. and Shilatifard,A. (2012) The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell Biol.*, **13**, 543–547.

318. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

319. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

320. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

321. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A., *et al.* (2017) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.

322. Gajos,M., Jasnovidova,O., van Bömmel,A., Freier,S., Vingron,M. and Mayer,A. (2021) Conserved DNA sequence features underlie pervasive RNA polymerase pausing. *Nucleic Acids Res.*, **49**, 4402–4420.

323. Lundberg,S.M., Nair,B., Vavilala,M.S., Horibe,M., Eisses,M.J., Adams,T., Liston,D.E., Low,D.K.-W., Newman,S.-F., Kim,J., *et al.* (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, **2**, 749–760.

324. Gilchrist,D.A. and Adelman,K. (2012) Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochim. Biophys. Acta*, **1819**, 700–706.

325. Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation (2010) *Cell*, **143**, 540–551.

326. Vaid,R., Wen,J. and Mannervik,M. (2020) Release of promoter-proximal paused Pol II in response to histone deacetylase inhibition. *Nucleic Acids Res.*, **48**, 4877–4890.

327. Lerner,E., Ingargiola,A., Lee,J.J., Borukhov,S., Michalet,X. and Weiss,S. (2017) Different types of pausing modes during transcription initiation. *Transcription*, **8**, 242–253.

328. Saldi,T., Cortazar,M.A., Sheridan,R.M. and Bentley,D.L. (2016) Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.*, **428**, 2623–2635.

329. Carrillo Oesterreich,F., Bieberstein,N. and Neugebauer,K.M. (2011) Pause locally, splice globally. *Trends Cell Biol.*, **21**, 328–335.

330. Akhtar,J., Kreim,N., Marini,F., Mohana,G., Brüne,D., Binder,H. and Roignant,J.-Y. (2019) Promoter-proximal pausing mediated by the exon junction complex regulates splicing. *Nat. Commun.*, **10**, 1–17.

331. Fusby,B., Kim,S., Erickson,B., Kim,H., Peterson,M.L. and Bentley,D.L. (2016) Coordination of RNA Polymerase II Pausing and 3' End Processing Factor Recruitment with Alternative Polyadenylation. *Mol. Cell. Biol.*, **36**, 295–303.

332. Ishov,A.M., Gurumurthy,A. and Bungert,J. (2020) Coordination of transcription, processing, and export of highly expressed RNAs by distinct biomolecular condensates. *Emerg Top Life Sci*, **4**, 281–291.

333. Yonaha,M. and Proudfoot,N.J. (1999) Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol. Cell*, **3**, 593–600.

334. McNamara,R.P., Bacon,C.W. and D'Orso,I. (2016) Transcription elongation control by the 7SK snRNP complex: Releasing the pause. *Cell Cycle*, **15**, 2115–2123.

335. Studniarek,C., Tellier,M., Martin,P.G.P., Murphy,S., Kiss,T. and Egloff,S. (2021) The 7SK/P-TEFb snRNP controls ultraviolet radiation-induced transcriptional reprogramming. *Cell Rep.*, **35**, 108965.

336. C Quaresma,A.J., Bugai,A. and Barboric,M. (2016) Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. *Nucleic Acids Res.*, **44**, 7527–7539.

337. Ji,X., Zhou,Y., Pandit,S., Huang,J., Li,H., Lin,C.Y., Xiao,R., Burge,C.B. and Fu,X.-D. (2013) SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell*, **153**, 855–868.

338. Barboric,M., Lenasi,T., Chen,H., Johansen,E.B., Guo,S. and Peterlin,B.M. (2009) 7SK snRNP/P-TEFb couples transcription elongation with alternative splicing and is essential for vertebrate development. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 7798–7803.

339. Egloff,S., Vitali,P., Tellier,M., Raffel,R., Murphy,S. and Kiss,T. (2017) The 7SK snRNP associates with the little elongation complex to promote snRNA gene expression. *EMBO J.*, **36**, 934–948.

340. Guo,Y.E., Manteiga,J.C., Henninger,J.E., Sabari,B.R., Dall'Agnese,A., Hannett,N.M., Spille,J.-H., Afeyan,L.K., Zamudio,A.V., Shrinivas,K., *et al.* (2019) Pol II phosphorylation

regulates a switch between transcriptional and splicing condensates. *Nature*, **572**, 543–548.

341. Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5′ Pause Release, Termination, and Transcription Elongation Rate (2019) *Mol. Cell*, **73**, 107–118.e4.

342. EP400 Deposits H3.3 into Promoters and Enhancers during Gene Activation (2016) *Mol. Cell*, **61**, 27–38.

343. Fuchs,M., Gerber,J., Drapkin,R., Sif,S., Ikura,T., Ogryzko,V., Lane,W.S., Nakatani,Y. and Livingston,D.M. (2001) The p400 complex is an essential E1A transformation target. *Cell*, **106**, 297–307.

344. Rhie,S.K., Yao,L., Luo,Z., Witt,H., Schreiner,S., Guo,Y., Perez,A.A. and Farnham,P.J. (2018) ZFX acts as a transcriptional activator in multiple types of human tumors by binding downstream of transcription start sites at the majority of CpG island promoters. *Genome Res.*, 10.1101/gr.228809.117.

345. Ni,W., Perez,A.A., Schreiner,S., Nicolet,C.M. and Farnham,P.J. (2020) Characterization of the ZFX family of transcription factors that bind downstream of the start site of CpG island promoters. *Nucleic Acids Res.*, **48**, 5986–6000.

346. Rahl,P.B. and Young,R.A. (2014) MYC and transcription elongation. *Cold Spring Harb. Perspect. Med.*, **4**, a020990.

347. Shaulian,E. and Karin,M. (2001) AP-1 in cell proliferation and survival. *Oncogene*, **20**, 2390–2400.

348. Gazon,H., Barbeau,B., Mesnard,J.-M. and Peloponese,J.-M.,Jr (2017) Hijacking of the AP-1 Signaling Pathway during Development of ATL. *Front. Microbiol.*, **8**, 2686.

349. Yu,L., Zhang,B., Deochand,D., Sacta,M.A., Coppo,M., Shang,Y., Guo,Z., Zeng,X., Rollins,D.A., Tharmalingam,B., *et al.* (2020) Negative elongation factor complex enables macrophage inflammatory responses by controlling anti-inflammatory gene expression. *Nat. Commun.*, **11**, 2286.

350. Hill,C.S. (2016) Transcriptional Control by the SMADs. *Cold Spring Harb. Perspect. Biol.*, **8**.

351. Wei,C., Xiao,R., Chen,L., Cui,H., Zhou,Y., Xue,Y., Hu,J., Zhou,B., Tsutsui,T., Qiu,J., *et al.* (2016) RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. *Mol. Cell*, **62**, 875–889.

352. Alexander,R.D., Innocente,S.A., Barrass,J.D. and Beggs,J.D. (2010) Splicing-dependent RNA polymerase pausing in yeast. *Mol. Cell*, **40**, 582–593.

353. Andersen,P.K. and Jensen,T.H. (2010) A pause to splice. *Mol. Cell*, **40**, 503–505.

354. Braeutigam,C., Rago,L., Rolke,A., Waldmeier,L., Christofori,G. and Winter,J. (2013) The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. *Oncogene*, **33**, 1082–1092.

355. Splicing Activation by Rbfox Requires Self-Aggregation through Its Tyrosine-Rich Domain (2017) *Cell*, **170**, 312–323.e10.

356. Quentmeier,H., Pommerenke,C., Bernhart,S.H., Dirks,W.G., Hauer,V., Hoffmann,S., Nagel,S., Siebert,R., Uphoff,C.C., Zaborski,M., *et al.* (2018) RBFOX2 and alternative splicing in B-cell lymphoma. *Blood Cancer J.*, **8**, 1–4.

357. Zhang,L., Tran,N.-T., Su,H., Wang,R., Lu,Y., Tang,H., Aoyagi,S., Guo,A., Khodadadi-Jamayran,A., Zhou,D., *et al.* (2015) Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *Elife*, **4**.

358. Xiao,R., Chen,J.-Y., Liang,Z., Luo,D., Chen,G., Lu,Z.J., Chen,Y., Zhou,B., Li,H., Du,X., *et al.* (2019) Pervasive Chromatin-RNA Binding Protein Interactions Enable RNA-Based Regulation of Transcription. *Cell*, **178**, 107–121.e18.

359. Rasche,N., Dybkov,O., Schmitzová,J., Akyildiz,B., Fabrizio,P. and Lührmann,R. (2012) Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome catalytic centre. *EMBO J.*, **31**, 1591–1604.

360. Wickramasinghe,V.O., Gonzàlez-Porta,M., Perera,D., Bartolozzi,A.R., Sibley,C.R., Hallegger,M., Ule,J., Marioni,J.C. and Venkitaraman,A.R. (2015) Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength. *Genome Biol.*, **16**, 201.

361. Briata,P., Bordo,D., Puppo,M., Gorlero,F., Rossi,M., Perrone-Bizzozero,N. and Gherzi,R. (2016) Diverse roles of the nucleic acid-binding protein KHSRP in cell differentiation and disease. *Wiley Interdiscip. Rev. RNA*, **7**, 227–240.

362. Rambout,X., Dequiedt,F. and Maquat,L.E. (2018) Beyond Transcription: Roles of Transcription Factors in Pre-mRNA Splicing. *Chem. Rev.*, **118**, 4339–4364.

363. Sollier,J., Stork,C.T., García-Rubio,M.L., Paulsen,R.D., Aguilera,A. and Cimprich,K.A. (2014) Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol. Cell*, **56**, 777–785.

364. Vidal,M. and Starowicz,K. (2017) Polycomb complexes PRC1 and their function in hematopoiesis. *Exp. Hematol.*, **48**, 12–31.

365. Pherson,M., Misulovin,Z., Gause,M., Mihindukulasuriya,K., Swain,A. and Dorsett,D. (2017) Polycomb repressive complex 1 modifies transcription of active genes. *Sci Adv*, **3**, e1700944.

366. R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters (2017) *Mol. Cell*, **68**, 745–757.e5.

367. Chédin,F. (2016) Nascent Connections: R-Loops and Chromatin Patterning. *Trends Genet.*, **32**, 828–838.

368. Huertas,P. and Aguilera,A. (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell*, **12**, 711–721.

369. Tuduri,S., Crabbé,L., Conti,C., Tourrière,H., Holtgreve-Grez,H., Jauch,A., Pantesco,V., De Vos,J., Thomas,A., Theillet,C., *et al.* (2009) Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat. Cell Biol.*, **11**, 1315–1324.

370. Doolittle,M.L., Calabrese,G.M., Mesner,L.D., Godfrey,D.A., Maynard,R.D., Ackert-Bicknell,C.L. and Farber,C.R. (2020) Genetic analysis of osteoblast activity identifies Zbtb40 as a regulator of osteoblast activity and bone mass. *PLoS Genet.*, **16**, e1008805.

371. Gromak,N., West,S. and Proudfoot,N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol. Cell. Biol.*, **26**, 3986–3996.

372. Nag,A., Narsinh,K. and Martinson,H.G. (2007) The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat. Struct. Mol. Biol.*, **14**, 662–669.

373. Schwalb,B., Michel,M., Zacher,B., Frühauf,K., Demel,C., Tresch,A., Gagneur,J. and Cramer,P. (2016) TT-seq maps the human transient transcriptome. *Science*, **352**, 1225–1228.

374. Prudêncio,P., Rebelo,K., Grosso,A.R., Martinho,R.G. and Carmo-Fonseca,M. (2020) Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data. *Methods*, **178**, 89–95.

375. Gressel,S., Schwalb,B., Decker,T.M., Qin,W., Leonhardt,H., Eick,D. and Cramer,P. (2017) CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife*, **6**.

376. Kessler,T. and Schunkert,H. (2021) Coronary Artery Disease Genetics Enlightened by Genome-Wide Association Studies. *JACC Basic Transl Sci*, **6**, 610–623.

377. Khera,A.V. and Kathiresan,S. (2017) Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.*, **18**, 331–344.

378. Aragam,K.G., Jiang,T., Goel,A., Kanoni,S., Wolford,B.N., Weeks,E.M., Wang,M., Hindy,G., Zhou,W., Grace,C., *et al.* (2021) Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *medRxiv*.

379. Nikpay,M., Goel,A., Won,H.-H., Hall,L.M., Willenborg,C., Kanoni,S., Saleheen,D., Kyriakou,T., Nelson,C.P., Hopewell,J.C., *et al.* (2015) A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.*, **47**, 1121–1130.

380. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A., *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

381. Brown,A.A., Buil,A., Viñuela,A., Lappalainen,T., Zheng,H.-F., Richards,J.B., Small,K.S., Spector,T.D., Dermitzakis,E.T. and Durbin,R. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. *Elife*, **3**, e01381.

382. Hemani,G., Shakhbazov,K., Westra,H.-J., Esko,T., Henders,A.K., McRae,A.F., Yang,J., Gibson,G., Martin,N.G., Metspalu,A., *et al.* (2014) Detection and replication of epistasis influencing transcription in humans. *Nature*, **508**, 249–253.

383. Pedruzzi,G. and Rouzine,I.M. (2019) Epistasis detectably alters correlations between genomic sites in a narrow parameter window. *PLoS One*, **14**, e0214036.

384. Caylak,G., Tastan,O. and Cicek,A.E. (2021) Potpourri: An Epistasis Test Prioritization Algorithm via Diverse SNP Selection. *J. Comput. Biol.*, **28**, 365–377.

385. Wellcome Trust Case Control Consortium, Craddock,N., Hurles,M.E., Cardin,N., Pearson,R.D., Plagnol,V., Robson,S., Vukcevic,D., Barnes,C., Conrad,D.F., *et al.* (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, **464**, 713–720.

386. Sudlow,C., Gallacher,J., Allen,N., Beral,V., Burton,P., Danesh,J., Downey,P., Elliott,P., Green,J., Landray,M., *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.

387. Samani,N.J., Erdmann,J., Hall,A.S., Hengstenberg,C., Mangino,M., Mayer,B., Dixon,R.J., Meitinger,T., Braund,P., Wichmann,H.-E., *et al.* (2007) Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.*, **357**, 443–453.

388. Erdmann,J., Grosshennig,A., Braund,P.S., König,I.R., Hengstenberg,C., Hall,A.S., Linsel-Nitschke,P., Kathiresan,S., Wright,B., Trégouët,D.-A., *et al.* (2009) New susceptibility

locus for coronary artery disease on chromosome 3q22.3. *Nat. Genet.*, **41**, 280–282.

389. Erdmann,J., Willenborg,C., Nahrstaedt,J., Preuss,M., König,I.R., Baumert,J., Linsel-Nitschke,P., Gieger,C., Tennstedt,S., Belcredi,P., *et al.* (2011) Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. *Eur. Heart J.*, **32**, 158–168.

390. Myocardial Infarction Genetics Consortium Investigators, Stitziel,N.O., Won,H.-H., Morrison,A.C., Peloso,G.M., Do,R., Lange,L.A., Fontanillas,P., Gupta,N., Duga,S., *et al.* (2014) Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N. Engl. J. Med.*, **371**, 2072–2082.

391. Nelson,C.P., Goel,A., Butterworth,A.S., Kanoni,S., Webb,T.R., Marouli,E., Zeng,L., Ntalla,I., Lai,F.Y., Hopewell,J.C., *et al.* (2017) Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.*, **49**, 1385–1391.

392. Li,L., Pang,S., Zeng,L., Güldener,U. and Schunkert,H. (2021) Genetically determined intelligence and coronary artery disease risk. *Clin. Res. Cardiol.*, **110**, 211–219.

393. Winkelmann,B.R., März,W., Boehm,B.O., Zotz,R., Hager,J., Hellstern,P., Senges,J. and LURIC Study Group (LUdwigshafen RIsk and Cardiovascular Health) (2001) Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics*, **2**, S1–73.

394. CARDIoGRAMplusC4D Consortium, Deloukas,P., Kanoni,S., Willenborg,C., Farrall,M., Assimes,T.L., Thompson,J.R., Ingelsson,E., Saleheen,D., Erdmann,J., *et al.* (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, **45**, 25–33.

395. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

396. Myocardial Infarction Genetics Consortium, Kathiresan,S., Voight,B.F., Purcell,S., Musunuru,K., Ardissino,D., Mannucci,P.M., Anand,S., Engert,J.C., Samani,N.J., *et al.* (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.*, **41**, 334–341.

397. Franzén,O., Ermel,R., Cohain,A., Akers,N.K., Di Narzo,A., Talukdar,H.A., Foroughi-Asl,H., Giambartolomei,C., Fullard,J.F., Sukhavasi,K., *et al.* (2016) Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*, **353**, 827–830.

398. Björkegren,J.L.M., Kovacic,J.C., Dudley,J.T. and Schadt,E.E. (2015) Genome-wide significant loci: how important are they? Systems genetics to understand heritability of coronary artery disease and other common complex disorders. *J. Am. Coll. Cardiol.*, **65**, 830–845.

399. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.

400. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

401. Lambrecht,L., Wobst,J., Wolf,B., Schunkert,H. and Kessler,T. (2019) Platelet Inhibition In Carriers Of The Gucy1A3 Coronary Artery Disease Risk Allele. *Atherosclerosis*, **287**, e140.

402. Kessler,T., Wolf,B., Eriksson,N., Kofink,D., Mahmoodi,B.K., Rai,H., Tragante,V., Åkerblom,A., Becker,R.C., Bernlochner,I., *et al.* (2019) Association of the coronary artery disease risk gene GUCY1A3 with ischaemic events after coronary intervention. *Cardiovasc. Res.*, **115**,

1512–1518.

403. Li,J.-L., Liu,L.-Y., Jiang,D.-D., Jiang,Y.-Y., Zhou,G.-Q., Mo,D.-C. and Luo,M. (2019) Associations between GUCY1A3 genetic polymorphisms and large artery atherosclerotic stroke risk in Chinese Han population: a case-control study. *Lipids Health Dis.*, **18**, 233.

404. Wobst,J., Dang,T.A., Kessler,T., Ameln,S. von, Tennstedt,S., Hengstenberg,C., Erdmann,J. and Schunkert,H. (2015) Functional evaluation of GUCY1A3 mutations associated with myocardial infarction risk. *BMC Pharmacol. Toxicol.*, **16**, 1–2.

405. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.

406. Sazonova,O., Zhao,Y., Nürnberg,S., Miller,C., Pjanic,M., Castano,V.G., Kim,J.B., Salfati,E.L., Kundaje,A.B., Bejerano,G., *et al.* (2015) Characterization of TCF21 Downstream Target Regions Identifies a Transcriptional Network Linking Multiple Independent Coronary Artery Disease Loci. *PLoS Genet.*, **11**, e1005202.

407. Lee,W.-S. and Kim,J. (2018) Insulin-like growth factor-1 signaling in cardiac aging. *Biochim. Biophys. Acta Mol. Basis Dis.*, **1864**, 1931–1938.

408. González-Guerra,J.L., Castilla-Cortazar,I., Aguirre,G.A., Muñoz,Ú., Martín-Estal,I., Ávila-Gallego,E., Granado,M., Puche,J.E. and García-Villalón,Á.L. (2017) Partial IGF-1 deficiency is sufficient to reduce heart contractibility, angiotensin II sensibility, and alter gene expression of structural and functional cardiac proteins. *PLoS One*, **12**, e0181760.

409. Frystyk,J., Ledet,T., Møller,N., Flyvbjerg,A. and Orskov,H. (2002) Cardiovascular disease and insulin-like growth factor I. *Circulation*, **106**, 893–895.

410. Ock,S., Ham,W., Kang,C.W., Kang,H., Lee,W.S. and Kim,J. (2021) IGF-1 protects against angiotensin II-induced cardiac fibrosis by targeting αSMA. *Cell Death Dis.*, **12**, 688.

411. Higashi,Y., Gautam,S., Delafontaine,P. and Sukhanov,S. (2019) IGF-1 and cardiovascular disease. *Growth Horm. IGF Res.*, **45**, 6–16.

412. Heinen,A., Nederlof,R., Panjwani,P., Spychala,A., Tschaidse,T., Reffelt,H., Boy,J., Raupach,A., Gödecke,S., Petzsch,P., *et al.* (2019) IGF1 Treatment Improves Cardiac Remodeling after Infarction by Targeting Myeloid Cells. *Mol. Ther.*, **27**, 46–58.

413. Douglas,G., Mehta,V., Al Haj Zen,A., Akoumianakis,I., Goel,A., Rashbrook,V.S., Trelfa,L., Donovan,L., Drydale,E., Chuaiphichai,S., *et al.* (2020) A key role for the novel coronary artery disease gene JCAD in atherosclerosis via shear stress mechanotransduction. *Cardiovasc. Res.*, **116**, 1863–1874.

414. Lee,N., Nicoloro,S.M., Straubhaar,J.R., Darrigo,M., Tam,S., Czech,M.P. and Fitzgibbons,T.P. (2013) Expression of ITGB8 in epicardial adipose tissue is highly and directly correlated with the severity of coronary atherosclerosis. 10.13028/96HA-S822.

415. Low,E.L., Baker,A.H. and Bradshaw,A.C. (2019) TGFβ, smooth muscle cells and coronary artery disease: a review. *Cell. Signal.*, **53**, 90–101.

416. Hanna,A. and Frangogiannis,N.G. (2019) The Role of the TGF-β Superfamily in Myocardial Infarction. *Front Cardiovasc Med*, **6**, 140.

417. Toma,I. and McCaffrey,T.A. (2012) Transforming growth factor-β and atherosclerosis: interwoven atherogenic and atheroprotective aspects. *Cell Tissue Res.*, **347**, 155–175.

418. Grainger,D.J. (2007) TGF-beta and atherosclerosis in man. *Cardiovasc. Res.*, **74**, 213–222.

419. Wang,Y. and Wang,J.-G. (2019) Genome-Wide Association Studies of Hypertension and Several Other Cardiovascular Diseases. *Pulse*, **6**, 169–186.

420. Prins,B.P., Lagou,V., Asselbergs,F.W., Snieder,H. and Fu,J. (2012) Genetics of coronary artery disease: genome-wide association studies and beyond. *Atherosclerosis*, **225**, 1–10.

421. Palmieri,F. (2013) The mitochondrial transporter family SLC25: identification, properties and physiopathology. *Mol. Aspects Med.*, **34**, 465–484.

422. Gutiérrez-Aguilar,M. and Baines,C.P. (2013) Physiological and pathological roles of mitochondrial SLC25 carriers. *Biochem. J*, **454**, 371–386.

423. Schumann,T., König,J., Henke,C., Willmes,D.M., Bornstein,S.R., Jordan,J., Fromm,M.F. and Birkenfeld,A.L. (2020) Solute Carrier Transporters as Potential Targets for the Treatment of Metabolic Disease. *Pharmacol. Rev.*, **72**, 343–379.

424. Zhang,Y., Zhang,Y., Sun,K., Meng,Z. and Chen,L. (2019) The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. *J. Mol. Cell Biol.*, **11**, 1–13.

425. Lin,L., Yee,S.W., Kim,R.B. and Giacomini,K.M. (2015) SLC transporters as therapeutic targets: emerging opportunities. *Nat. Rev. Drug Discov.*, **14**, 543–560.

426. Li,L., He,M., Zhou,L., Miao,X., Wu,F., Huang,S., Dai,X., Wang,T. and Wu,T. (2015) A solute carrier family 22 member 3 variant rs3088442 G→A associated with coronary heart disease inhibits lipopolysaccharide-induced inflammatory response. *J. Biol. Chem.*, **290**, 5328–5340.

427. Zhao,Q., Wei,H., Liu,D., Shi,B., Li,L., Yan,M., Zhang,X., Wang,F. and Ouyang,Y. (2017) PHACTR1 and SLC22A3 gene polymorphisms are associated with reduced coronary artery disease risk in the male Chinese Han population. *Oncotarget*, **8**, 658–663.

428. Palmieri,F., Scarcia,P. and Monné,M. (2020) Diseases Caused by Mutations in Mitochondrial Carrier Genes SLC25: A Review. *Biomolecules*, **10**.

429. Shendre,A., Irvin,M.R., Wiener,H., Zhi,D., Limdi,N.A., Overton,E.T. and Shrestha,S. (2017) Local Ancestry and Clinical Cardiovascular Events Among African Americans From the Atherosclerosis Risk in Communities Study. *J. Am. Heart Assoc.*, **6**.

430. Yuan,S., Zheng,J.-S., Mason,A.M., Burgess,S. and Larsson,S.C. (2022) Genetically predicted circulating vitamin C in relation to cardiovascular disease. *Eur. J. Prev. Cardiol.*, **28**, 1829–1837.

431. Georges,A., Yang,M.-L., Berrandou,T.-E., Bakker,M., Dikilitas,O., Kiando,S.R., Ma,L., Satterfield,B.A., Sengupta,S., Yu,M., *et al.* (2020) Genetic association studies of fibromuscular dysplasia identify new risk loci and shared genetic basis with more common vascular diseases. *bioRxiv*, 10.1101/2020.09.16.20195701.

432. Bergman,D., Halje,M., Nordin,M. and Engström,W. (2013) Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology*, **59**, 240–249.

433. Colao,A. (2008) The GH-IGF-I axis and the cardiovascular system: clinical implications. *Clin. Endocrinol.* , **69**, 347–358.

434. Rodríguez,S., Gaunt,T.R., O'Dell,S.D., Chen,X.-H., Gu,D., Hawe,E., Miller,G.J., Humphries,S.E. and Day,I.N.M. (2004) Haplotypic analyses of the IGF2-INS-TH gene cluster in relation to cardiovascular risk traits. *Hum. Mol. Genet.*, **13**, 715–725.

435. Iosef Husted,C. and Valencik,M. (2016) Insulin-like growth factors and their potential role in cardiac epigenetics. *J. Cell. Mol. Med.*, **20**, 1589–1602.

436. Zaina,S., Pettersson,L., Thomsen,A.B., Chai,C.-M., Qi,Z., Thyberg,J. and Nilsson,J. (2003) Shortened life span, bradycardia, and hypotension in mice with targeted expression of an Igf2 transgene in smooth muscle cells. *Endocrinology*, **144**, 2695–2703.

437. Zaina,S., Pettersson,L., Ahrén,B., Brånén,L., Hassan,A.B., Lindholm,M., Mattsson,R., Thyberg,J. and Nilsson,J. (2002) Insulin-like growth factor II plays a central role in atherosclerosis in a mouse model. *J. Biol. Chem.*, **277**, 4505–4511.

438. Zekavat,S.M., Chou,E.L., Zekavat,M., Pampana,A., Paruchuri,K., Lino Cardenas,C.L., Koyama,S., Ghazzawi,Y., Kii,E., Uddin,M.M., *et al.* (2022) Fibrillar Collagen Variants in Spontaneous Coronary Artery Dissection. *JAMA Cardiol*, **7**, 396–406.

439. Nelson,C.P., Hamby,S.E., Saleheen,D., Hopewell,J.C., Zeng,L., Assimes,T.L., Kanoni,S., Willenborg,C., Burgess,S., Amouyel,P., *et al.* (2015) Genetically determined height and coronary artery disease. *N. Engl. J. Med.*, **372**, 1608–1618.

440. Maitra,A., Shanker,J., Dash,D., Sannappa,P.R., John,S., Siwach,P., Rao,V.S., Sridhara,H. and Kakkar,V.V. (2010) Polymorphisms in the pituitary growth hormone gene and its receptor associated with coronary artery disease in a predisposed cohort from India. *J. Genet.*, **89**, 437–447.

441. Filipp,F.V. (2019) Opportunities for Artificial Intelligence in Advancing Precision Medicine. *Curr. Genet. Med. Rep.*, **7**, 208–213.

442. Martínez-García,M. and Hernández-Lemus,E. (2021) Data Integration Challenges for Machine Learning in Precision Medicine. *Front. Med.*, **8**, 784455.

443. Daniels,H., Jones,K.H., Heys,S. and Ford,D.V. (2021) Exploring the Use of Genomic and Routinely Collected Data: Narrative Literature Review and Interview Study. *J. Med. Internet Res.*, **23**, e15739.

444. MacEachern,S.J. and Forkert,N.D. (2021) Machine learning for precision medicine. *Genome*, **64**, 416–425.

445. Plant,D. and Barton,A. (2021) Machine learning in precision medicine: lessons to learn. *Nat. Rev. Rheumatol.*, **17**, 5–6.

446. Zhang,S., Bamakan,S.M.H., Qu,Q. and Li,S. (2019) Learning for Personalized Medicine: A Comprehensive Review From a Deep Learning Perspective. *IEEE Rev. Biomed. Eng.*, **12**, 194–208.

447. Johnson,K.B., Wei,W.-Q., Weeraratne,D., Frisse,M.E., Misulis,K., Rhee,K., Zhao,J. and Snowdon,J.L. (2021) Precision Medicine, AI, and the Future of Personalized Health Care. *Clin. Transl. Sci.*, **14**, 86–93.

448. Piccialli,F., Calabrò,F., Crisci,D., Cuomo,S., Prezioso,E., Mandile,R., Troncone,R., Greco,L. and Auricchio,R. (2021) Precision medicine and machine learning towards the prediction of the outcome of potential celiac disease. *Sci. Rep.*, **11**, 5683.

449. Peng,J., Jury,E.C., Dönnes,P. and Ciurtin,C. (2021) Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges. *Front. Pharmacol.*, **12**, 720694.

450. Uddin,M., Wang,Y. and Woodbury-Smith,M. (2019) Artificial intelligence for precision medicine

in neurodevelopmental disorders. *NPJ Digit Med*, **2**, 112.

451. Cramer,P. (2019) Organization and regulation of gene transcription. *Nature*, **573**, 45–54.

452. Bardet,A.F., Steinmann,J., Bafna,S., Knoblich,J.A., Zeitlinger,J. and Stark,A. (2013) Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**, 2705–2713.

453. Behjati Ardakani,F., Schmidt,F. and Schulz,M.H. (2018) Predicting transcription factor binding using ensemble random forest models. *F1000Res.*, **7**, 1603.

454. Buenrostro,J.D., Wu,B., Chang,H.Y. and Greenleaf,W.J. (2015) ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, **109**, 21.29.1–21.29.9.

455. Bentsen,M., Goymann,P., Schultheis,H., Klee,K., Petrova,A., Wiegandt,R., Fust,A., Preussner,J., Kuenne,C., Braun,T., *et al.* (2020) ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.*, **11**, 4267.

456. Funk,C.C., Casella,A.M., Jung,S., Richards,M.A., Rodriguez,A., Shannon,P., Donovan-Maiye,R., Heavner,B., Chard,K., Xiao,Y., *et al.* (2020) Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data across 27 Tissue Types. *Cell Rep.*, **32**, 108029.

457. Ramkumar,P. and Kampmann,M. (2018) CRISPR-based genetic interaction maps inform therapeutic strategies in cancer. *Transl. Cancer Res.*, **7**, S61–S67.

458. Diehl,V., Wegner,M., Grumati,P., Husnjak,K., Schaubeck,S., Gubas,A., Shah,V.J., Polat,I.H., Langschied,F., Prieto-Garcia,C., *et al.* (2021) Minimized combinatorial CRISPR screens identify genetic interactions in autophagy. *Nucleic Acids Res.*, **49**, 5684–5704.

459. Tang,X., Huang,Y., Lei,J., Luo,H. and Zhu,X. (2019) The single-cell sequencing: new developments and medical applications. *Cell Biosci.*, **9**, 53.

460. Rotem,A., Ram,O., Shoresh,N., Sperling,R.A., Goren,A., Weitz,D.A. and Bernstein,B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.

461. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.

462. Swanson,E., Lord,C., Reading,J., Heubeck,A.T., Genge,P.C., Thomson,Z., Weiss,M.D.A., Li,X.-J., Savage,A.K., Green,R.R., *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*, **10**.

463. van Dijk,D., Sharma,R., Nainys,J., Yim,K., Kathail,P., Carr,A.J., Burdziak,C., Moon,K.R., Chaffer,C.L., Pattabiraman,D., *et al.* (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, **174**, 716–729.e27.

464. Haliburton,J.R., Shao,W., Deutschbauer,A., Arkin,A. and Abate,A.R. (2017) Genetic interaction mapping with microfluidic-based single cell sequencing. *PLoS One*, **12**, e0171302.

465. Behjati Ardakani,F., Kattler,K., Nordström,K., Gasparoni,N., Gasparoni,G., Fuchs,S., Sinha,A., Barann,M., Ebert,P., Fischer,J., *et al.* (2018) Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters. *Epigenetics Chromatin*, **11**, 66.

466. Greenwald,H.S. and Oertel,C.K. (2017) Future Directions in Machine Learning. *Front. Robot.*

AI, **3**.

467. Deep learning for genomics (2019) *Nat. Genet.*, **51**, 1.

468. Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat. Rev. Genet.*, **16**, 321–332.

469. Huang,K., Xiao,C., Glass,L.M., Critchlow,C.W., Gibson,G. and Sun,J. (2021) Machine learning applications for therapeutic tasks with genomics data. *Patterns (N Y)*, **2**, 100328.

470. Whalen,S., Schreiber,J., Noble,W.S. and Pollard,K.S. (2022) Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.*, **23**, 169–181.

471. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

472. Degroeve,S., De Baets,B., Van de Peer,Y. and Rouzé,P. (2002) Feature subset selection for splice site prediction. *Bioinformatics*, **18 Suppl 2**, S75–83.

473. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A., *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

474. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.

475. Zhang,Y., Zhou,X. and Cai,X. (2020) Predicting Gene Expression from DNA Sequence using Residual Neural Network. *bioRxiv*, 10.1101/2020.06.21.163956.

476. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

477. Wang,C., Sun,D., Huang,X., Wan,C., Li,Z., Han,Y., Qin,Q., Fan,J., Qiu,X., Xie,Y., *et al.* (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.*, **21**, 198.

478. Jiang,S. and Mortazavi,A. (2018) Integrating ChIP-seq with other functional genomics data. *Brief. Funct. Genomics*, **17**, 104–115.

479. Ouyang,Z., Zhou,Q. and Wong,W.H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 21521–21526.

480. Politi,K. and Herbst,R.S. (2015) Lung cancer in the era of precision medicine. *Clin. Cancer Res.*, **21**, 2213–2220.

481. Naito,Y. and Urasaki,T. (2018) Precision medicine in breast cancer. *Chin Clin Oncol*, **7**, 29.

482. Tran,N.H., Cavalcante,L.L., Lubner,S.J., Mulkerin,D.L., LoConte,N.K., Clipson,L., Matkowskyj,K.A. and Deming,D.A. (2015) Precision medicine in colorectal cancer: the molecular profile alters treatment strategies. *Ther. Adv. Med. Oncol.*, **7**, 252–262.

483. Bonelli,P., Borrelli,A., Tuccillo,F.M., Silvestro,L., Palaia,R. and Buonaguro,F.M. (2019) Precision medicine in gastric cancer. *World J. Gastrointest. Oncol.*, **11**, 804–829.

484. Santangelo,G., Caruso,G., Palaia,I., Tomao,F., Perniola,G., Di Donato,V., Fischetti,M., Muzii,L.

and Benedetti Panici,P. (2020) The emerging role of precision medicine in the treatment of ovarian cancer. *Expert Review of Precision Medicine and Drug Development*, **5**, 283–297.

485. Syed,Y.Y. (2020) Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. *Mol. Diagn. Ther.*, **24**, 621–632.

486. Howe,L.J., Dudbridge,F., Schmidt,A.F., Finan,C., Denaxas,S., Asselbergs,F.W., Hingorani,A.D. and Patel,R.S. (2020) Polygenic risk scores for coronary artery disease and subsequent event risk amongst established cases. *Hum. Mol. Genet.*, **29**, 1388–1395.

487. An improved method of testing for evolutionary homology (1966) *J. Mol. Biol.*, **16**, 9–16.