

Josip Juraj Strossmayer University of Osijek

University of Dubrovnik

Ruđer Bošković Institute

Interdisciplinary Doctoral Studies of Molecular Biosciences

Katarina Mandić

**THE ROLE OF LONG-RANGE TRANSCRIPTIONAL REGULATION
ACROSS THE HUMAN DISEASE LANDSCAPE**

PhD thesis

Osijek, 2026.

BASIC DOCUMENTATION CARD

Josip Juraj Strossmayer University of Osijek
University of Dubrovnik
Ruđer Bošković Institute
University Postgraduate Interdisciplinary Doctoral Study of
Molecular Biosciences

PhD thesis

Scientific Area: Interdisciplinary area of science

Scientific Fields: biology, basic medical sciences

The role of long-range transcriptional regulation across the human disease landscape

Katarina Mandić

Thesis performed at: Laboratory for Computational Biology and Translational Medicine, Division of Computing and Data Science, Ruđer Bošković Institute.

Supervisor/s: dr. sc. Anja Barešić and dr. sc. Dubravka Švob Štrac

Short abstract:

One of the standing problems in biomedicine is identifying target genes of non-coding disease-associated variants. Genomic regulatory blocks are regions containing syntenic clusters of highly conserved non-coding elements around key developmental genes. They harbour regulatory elements that may act over long distances, providing a framework for identifying target genes of disease-associated variants.

Number of pages: 135

Number of figures: 38

Number of tables: 14

Number of references: 206

Original in: English

Key words: genomic regulatory blocks, variant interpretation, long-range regulation, multi-omics integration, single-cell transcriptomics, regulatory variants.

Date of the thesis defense:

Reviewers:

- 1.
- 2.
- 3.
4. (substitute)

Thesis deposited in: National and University Library in Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb; City and University Library of Osijek, Europska avenija 24, Osijek; Josip Juraj Strossmayer University of Osijek, Trg sv. Trojstva 3, Osijek.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište Josipa Jurja Strossmayera u Osijeku
Sveučilište u Dubrovniku
Institut Ruđer Bošković
Poslijediplomski interdisciplinarni sveučilišni
studij Molekularne bioznanosti

Doktorska disertacija

Znanstveno područje: Interdisciplinarno područje znanosti
Znanstvena polja: biologija, temeljne medicinske znanosti

Uloga transkripcijske regulacije dalekog dometa diljem krajolika bolesti čovjeka

Katarina Mandić

Disertacija je izrađena u: Laboratorij za računalnu biologiju i translacijsku medicinu, Zavod za računarstvo i podatkovnu znanost, Institut Ruđer Bošković.

Mentor/i: dr. sc. Anja Barešić, dr. sc. Dubravka Švob Štrac

Kratki sažetak doktorske disertacije:

Jedan od nerješениh problema u biomedicini je određivanje ciljnog gena nekodirajućih varijante asociirane s bolestima. Genomski regulatorni blokovi su regije koje sadrže sintenske klasterе evolucijski konzerviranih nekodirajućih elemenata oko ključnog razvojnog gena. Sadrže regulatorne elemente koji mogu djelovati na velikim udaljenostima, pružajući okvir za identifikaciju gena ciljanih varijantama povezanim sa bolestima.

Broj stranica: 135

Broj slika: 38

Broj tablica: 14

Broj literaturnih navoda: 206

Jezik izvornika: Engleski

Ključne riječi: genomski regulatorni blokovi, interpretacija varijanti, regulacija dalekog dometa, multi-omska integracija, transkriptom pojedinačnih stanica

Datum obrane:

Stručno povjerenstvo za obranu:

- 1.
- 2.
- 3.
4. (zamjena)

Disertacija je pohranjena u: Nacionalnoj i sveučilišnoj knjižnici Zagreb, Ul. Hrvatske bratske zajednice 4, Zagreb; Gradskoj i sveučilišnoj knjižnici Osijek, Europska avenija 24, Osijek; Sveučilištu Josipa Jurja Strossmayera u Osijeku, Trg sv. Trojstva 3, Osijek.

Funding

This PhD thesis was performed at the Laboratory for Computational Biology and Translational Medicine, Division of Computing and Data Science, Ruđer Bošković Institute under the supervision of dr. sc. Anja Barešić, as a part of the HRZZ research project Exploring Interactions Between Regulatory Variants in Human Disease Context (UIP-2020-02-1623) and Young Researchers' Career Development Project - Training New Doctoral Students (DOK-2021-02-7831).

Acknowledgements

First and foremost, I would like to thank Anja, who is simply the most dedicated and kind mentor, and from whom I learned so much. Thank you for celebrating all my small victories, for holding my hand in all the worst administrative ordeals, and putting me and all your students in the first place. I loved our long meetings discussing science, and I hope to have many more!

Next, I would like to thank all my IRB colleagues for their support and camaraderie. Soon after I started my PhD I had a strong sense of belonging thanks to you. Miha, the air in our office will still be buzzing from all the mental work we do, also you have to admit it, Linux is just better, and let me know when your Windows machine is done updating. Marina, thank you for all the long walks around the institute, for making me laugh and lending me a listening ear when I needed it the most. Ivan, thanks for letting me steal your fruit occasionally, and for always brightening my day with your unwavering positive energy. Matija, even though I never won the lunch award, I had fun pulling pranks with you (especially to our beloved department head). Filip, thank you for teaching me a lot about military airplanes and how they communicate, you never know when I will need that information. Thank you power rangers for the fun lunches, Damir B. for the authentic café italiano beans, Nenad for seeing wasps. Thanks to everyone from the department Vlatka, Višnja, Dalibor, Elli, Damir K., Ilona, Davor - even though we changed locations and got a new name, we are OG ZEL. Special thanks to Tom for all of the insightful comments and to your dedication to your work - I am extremely honoured to be a part of the department.

I am grateful for all my family and friends. I am lucky to have a tight circle of loved ones on who I can always rely on. Also I am grateful for my mrki softball family, and being on the field with you makes me relax and focus.

To my sisters Ana and Martina, thank you. Martina, your unconditional kindness and warmth have always made me feel safe and confident. Ana, thank you for making me fall in love with science, I miss you.

Finally, to my husband Damir. I think both of us are crazy in our special way, and even though we don't always understand each other's craziness, we deeply admire each other's passion with which we do what we love. Thank you for always pushing me to be a better person and to grow. I am lucky to have you and to be able to build a family with you and Kiara. However, now that Kiara is here I have a feeling that we have become NPCs of our own lives and she is the main character.

List of Abbreviations

ABC	Activity-by-Contact model
BH	Benjamini–Hochberg
Biosamples	Primary cells, cell lines and tissues
bp	Base pair
CAGE	Cap analysis of gene expression
cDC1	Conventional dendritic cell 1
cDC2	Conventional dendritic cell 2
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CGI	CpG island
CNE	Conserved non-coding element
cS2G	Combined SNP-to-gene
CTCF	CCCTC-binding factor
DC	Dendritic cell
DEG	Differentially expressed gene
DEmiR	Differentially expressed microRNA
DEmiR-DEG	Differentially expressed gene associated with a differentially expressed microRNA
DEmiR+DM- DEG	Differentially expressed gene associated with a differentially expressed microRNA and with differentially methylated CpG island
DM	Differentially methylated or differential methylation
DM-CGI	Differentially methylated CpG island

DM-DEG	Differentially expressed gene with a differentially methylated promoter-associated CpG island
DNAm	DNA methylation
empP	Empirical p-value
eRNA	Enhancer RNA
eQTL	Expression quantitative trait locus
FC	Fold change
GO:BP	Gene Ontology: Biological Processes
GRB	Genomic regulatory block
GWAS	Genome-wide association study
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage disequilibrium
methylome	The whole-genome DNA methylation profile
miRNA	microRNA
miRNome	The whole-genome microRNA profile
NK cells	Natural killer cells
noReg-DEG	Differentially expressed gene without associated microRNA or differentially methylated CpG island
O	Oral cavity
OP	Oropharynx
PBMC	Peripheral blood mononuclear cells
PCA	Principal component analysis
PD	Parkinson's disease
pDC	Plasmacytoid dendritic cell
S2G	SNP-to-gene
scRNA-seq	Single-cell RNA sequencing
SCS	Set Counts and Sizes

SNN	Shared nearest neighbour
SNP	Single nucleotide polymorphism
TAD	Topologically associating domain
TCGA	The Cancer Genome Atlas
TCMs	Central memory T-cells
TEMs	Effector memory T-cells
TF	Transcription factor
TPM	Transcripts per million
transcriptome	Whole-genome RNA-seq profile
Tregs	Regulatory T-cells
tSNE	t-distributed stochastic neighbor embedding
TSS	Transcription start site
UMI	Unique molecular identifiers

Contents

1	Introduction	1
1.1	Gene regulation	1
1.1.1	Human genome and regulatory complexity	1
1.1.2	Transcriptional regulation	2
1.1.3	Promoters	3
1.1.4	Enhancers	3
1.2	Long-range transcriptional regulation	6
1.2.1	Distal enhancer-promoter interactions	6
1.2.2	Topologically associated domains	6
1.3	Evolutionarily conserved regulatory landscapes	8
1.3.1	Conserved non-coding elements	8
1.3.2	Genomic regulatory blocks	9
1.3.3	GRB target genes	9
1.3.4	Relationship between GRBs and TADs	11
1.4	Genetic variation and genome-wide association studies	11
1.4.1	Complex traits and disease	11
1.4.2	Genome-wide association studies	12
1.4.3	Non-coding variants affect gene expression	12
1.5	Methods for variant-gene association	14
1.5.1	Chromatin interaction and enhancer-based methods	14
1.5.2	Integrated prediction frameworks	15
1.5.3	Limitations of current approaches	16

1.6	GRB-based approaches for variant-gene association	16
1.6.1	Type 2 diabetes and obesity associated variant involved in long-range regulation	17
1.6.2	Extending GRBs with enhancer-promoter associations	17
1.7	Disease of special interest: head and neck squamous cell carcinoma	20
1.7.1	Disease overview	20
1.7.2	Epigenetic regulation	21
1.7.3	DNA methylation in cancer	21
1.7.4	microRNA in cancer	22
1.8	Disease of special interest: Parkinson's disease	23
1.8.1	Disease overview	23
1.8.2	Immune dysregulation in Parkinson's disease	24
1.8.3	High-throughput sequencing in Parkinson's disease	24
2	Aims and hypotheses	26
3	Methods	27
3.1	Methods for the targPred study	27
3.1.1	Empirical p-value for enhancer-gene links	27
3.1.2	GWAS processing	28
3.1.3	Identification of GWAS loci within genomic regulatory blocks	28
3.1.4	Identification of long-range regulatory traits	29
3.1.5	Locus enrichment in genomic regulatory blocks	30
3.1.6	Validated SNP-gene association datasets	30
3.1.7	Comparison with existing SNP-gene prediction methods	31
3.1.8	Visualization of the childhood obesity locus	31
3.1.9	Statistical testing	32
3.2	Methods for the head and neck squamous cell carcinoma study	32
3.2.1	Croatian HNSCC sample processing	32
3.2.2	Sequencing data analysis	33

3.2.3	Clustering and validation	34
3.2.4	Genome-wide CpG analysis	35
3.3	Methods for the Parkinson's disease study	36
3.3.1	Participant selection and PBMC isolation	36
3.3.2	Single-cell RNA sequencing	37
3.3.3	Alignment and quality control of sequencing data	37
3.3.4	Single-cell data pre-processing	38
3.3.5	Cell type annotation	38
3.3.6	Differential expression and pathway enrichment analysis	39
3.3.7	Validation analyses	39
3.3.8	Exome sequence processing	40
4	Results	41
4.1	Results of the targPred study	41
4.1.1	General properties of targPred	41
4.1.2	Traits under long-range regulation	43
4.1.3	Comparison between targPred and validated SNP-gene links	45
4.1.4	Comparison between targPred and computational SNP-gene prediction methods	46
4.1.5	Long-range regulation in childhood obesity	47
4.2	Results of the head and neck squamous cell carcinoma study	51
4.2.1	Croatian HNSCC Cohort and Multi-Omics Profiling	51
4.2.2	Global Omics Patterns in Croatian HNSCC	52
4.2.3	Differential Expression and Pathway Enrichment in Tumour vs Control	53
4.2.4	Integrative analysis of miRNA and DNAm in HNSCC	55
4.2.5	Epigenetic clusters and biological programs in HNSCC	56
4.2.6	Validation of epigenetic clusters in TCGA-HNSCC	58
4.2.7	Long-range epigenetic regulation	59
4.3	Results of the Parkinson's disease study	62
4.3.1	Generation of the single-cell dataset	63

4.3.2	Data validation	65
4.3.3	Identification of major cell populations	67
4.3.4	Identification of CD4+ T-cell subtypes	68
4.3.5	Identification of dendritic cell subtypes	68
4.3.6	Differential gene expression in PD vs CTRL	70
4.3.7	Pathway enrichment in PD vs CTRL	72
4.3.8	Non-coding variation identification from exome sequencing	73
5	Discussion	76
5.1	Discussion of the targPred study	76
5.2	Discussion of the head and neck squamous cell carcinoma study	78
5.3	Discussion of the Parkinson's disease study	82
6	Conclusion	88
6.1	Development of targPred	88
6.2	Applications of targPred on examples with clinical relevance	89
6.3	Future prospects	90
	References	92
	Summary	116
	Sažetak	117
	Curriculum Vitae	118
A	Supplementary Figures	121
B	Supplementary Tables	130
C	Supplementary Data	134

List of Figures

1.1	Historic overview of genomic discoveries	2
1.2	CAGE signals reflect active promoters and enhancers	4
1.3	Levels of three-dimensional chromatin organization	7
1.4	Features of GRB target and bystander genes	10
1.5	GWAS study design and result interpretation	13
1.6	Type 2 diabetes and obesity risk locus: from long-range regulation to molecular function	18
4.1	Results applying targPred across the GWAS Catalog	42
4.2	Traits under long-range regulation	44
4.3	Comparison of GRB and GWAS based target gene assignments with experimentally validated SNP–gene associations from ClinVar and Alsheik et al.	46
4.4	Comparison of GRB annotation with GeneHancer, ABC, and cS2G	47
4.5	Probability distribution of metabolic syndrome traits associated-loci in GRB	48
4.6	Childhood obesity-associated loci linked to BDNF.	50
4.7	Croatian HNSCC cohort.	52
4.8	Exploratory data analysis of Croatian HNSCC cohort	53
4.9	Gene enrichment analysis of DEGs in tumour versus control samples	54
4.10	Counts of DM CpG probes in cancer versus control groups across genomic elements	56
4.11	Integration and classification of DEGs depending on miRNA and DNAm	57
4.12	Integration of and clustering of transcriptome, miRNome and methylome	58
4.13	Comparison of the Croatian and TCGA-HNSCC cohorts	59

4.14	Distribution of CpG dinucleotides falling into promoters, gene body or intergenic regions in randomly shuffled GRBs	60
4.15	Distribution of ratios of DM CpG probes across promoters, gene body regions, and intergenic regions in randomly shuffled GRBs	61
4.16	Examples of correlation between CpGs and target and bystander genes in two GRBs	62
4.17	Single-cell sequencing data analysis	64
4.18	Biological replicate validation and distribution of PD and CTRL cells	66
4.19	Single-cell analysis of CD4+ T-cell subtypes	69
4.20	Single-cell analysis of DC subtypes	70
4.21	Differential gene expression in CD4+ TEMs and cDC2s	71
4.22	Pathway enrichment analysis of CD4+ TEMs and cDC2s	73
4.23	Non-coding variants from exome sequencing enrichment	75
S1	Higher principal components in Croatian HNSCC and additional sample distance heatmaps	121
S2	Survival curves	122
S3	Genomic neighbourhood of <i>ANK2</i> GRB	123
S4	tSNE donor overlay	124
S5	Exclusion of sample S29 based on quality control metrics	125
S6	Correlation of FC values obtained by single-cell approach and pseudobulk approach	126
S7	Correlation of FC values obtained by single-cell approach and bulk RNA approach	127
S8	Extended cell type annotation	128
S9	Differentiation trajectories for CD4+ T-cells and DCs	129

List of Tables

4.1	Test results across genomic distance intervals	42
4.2	Comparison of prediction scores among targPred categories	48
4.3	Enrichment of GWAS loci within GRBs for selected disease traits	49
4.4	Number of target genes per k-means cluster	57
4.5	Median methylation levels in GRB target and bystander genes	61
4.6	Clinical characteristics of the scRNA-seq cohort	63
4.7	Numbers of DEGs identified in each immune cell subtype	70
4.8	Differential expression of the <i>BCL11B</i> across CD4+ T-cell subtypes	71
4.9	Demographic characteristics of the exome cohort.	74
S1	Mapping statistics for RNA-seq samples	130
S2	Downregulated tumour suppressors with hypermethylated promoters	131
S3	Comparison between proportion of selected cell subtypes between PD and CTRL	132
S4	Variant carrier counts in PD and CTRL samples	132
S5	Enrichment of variants associated with PD-related traits within GRBs	132
S6	GRB target genes that are differentially expressed in CD4+ TCM and cDC2 cells	133

1 Introduction

1.1 Gene regulation

1.1.1 Human genome and regulatory complexity

Before the completion of the Human Genome Project, estimates of the number of human genes were considerably higher than those obtained from the first draft genome sequences. Early projections suggested that the human genome might contain approximately 100,000 genes, whereas the two initial genome publications reported approximately 31,000 and 26,588 protein-coding genes, respectively (Figure 1.1) [1, 2]. Subsequent improvements in genome annotation reduced this estimate further, with current databases recognizing approximately 20,000 protein-coding genes, although the exact number varies among annotation resources [3]. Moreover, protein-coding sequences account for less than 2% of the human genome [4]. These observations demonstrated that the biological complexity of humans and other multicellular organisms cannot be explained solely by genome size or protein-coding gene number [5]. Instead, organismal complexity depends substantially on the precise regulation of gene expression across tissues, cell types and developmental stages. This regulation is mediated by a large repertoire of *cis*-regulatory elements, including promoters, enhancers, silencers and insulators. The ENCODE consortium has catalogued 2.37 million candidate *cis*-regulatory elements in the human genome, illustrating the extensive regulatory potential of non-coding genomic sequences [6]. These elements contribute to the control of when, where and to what extent genes are expressed, thereby enabling cells with the same genomic sequence to acquire distinct molecular identities and specialized functions.

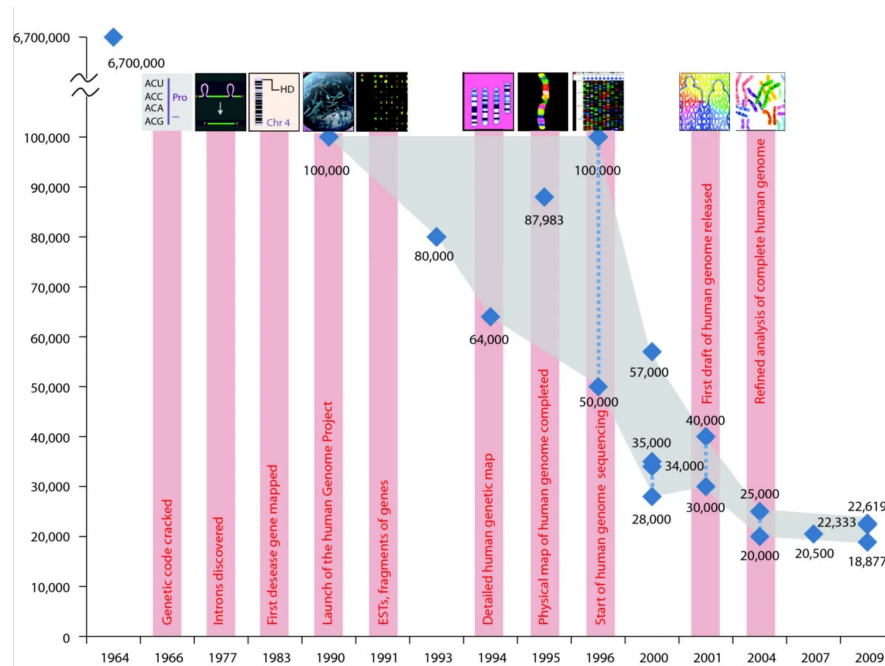


Figure 1.1. A time lent with highlighted articles reporting projected number of genes before the initiation of the Human Genome Project, and after its completion. Adapted from Pertea and Salzberg [7].

1.1.2 Transcriptional regulation

Gene expression can be regulated at multiple stages, including transcription, RNA processing, translation and post-translational protein modification. However, transcription initiation represents a major regulatory checkpoint and is the principal focus of this thesis [8]. Although most cells within an organism contain essentially the same genomic sequence, they differ substantially in their morphology and function because they express distinct sets of genes. The precise spatial and temporal regulation of transcription is therefore essential for biological processes such as embryonic development, cell proliferation, apoptosis and cell type differentiation. Disruption of these regulatory programmes can contribute to disease [9].

Transcriptional *cis*-regulatory elements have historically been divided into four broad categories: promoters, enhancers, silencers and insulators. These elements do not operate independently, but function through interactions with sequence-specific DNA-binding proteins, cofactors and the general transcriptional machinery. Transcription of protein-coding genes is performed by RNA polymerase II, whose recruitment and activity are regulated by general transcription factors, sequence-specific activators and repressors, and transcriptional coactivators

and corepressors [8].

1.1.3 Promoters

Promoter regions can be broadly divided into core and proximal promoter elements. The core promoter surrounds the transcription start site (TSS) and contains sequence features that support the assembly of the pre-initiation complex and the recruitment of RNA polymerase II [10]. The TATA box was among the first core promoter motifs to be characterized and is recognized by the TATA-binding protein, a subunit of the general transcription factor TFIID. However, only a subset of human promoters contains a canonical TATA box, and many promoters use other core promoter elements or lack a single clearly defined motif [8, 11]. Proximal promoter regions extend upstream of the core promoter and contain binding sites for sequence-specific transcription factors that modulate transcriptional output [8]. Approximately 60% of human gene promoters overlap CpG islands (CGIs), regions with a higher-than-expected frequency of CpG dinucleotides, although this proportion and its definition vary among annotation approaches [1].

Cap analysis of gene expression (CAGE) enables the detection of transcription initiation by sequencing the 5' ends of capped RNA molecules at single-nucleotide resolution. At active gene promoters, CAGE typically detects a relatively strong transcription initiation signal associated with the production of a stable transcript, with transcription occurring predominantly in the direction of the gene (Figure 1.2a) [12].

1.1.4 Enhancers

The first human enhancer was identified within the immunoglobulin heavy-chain locus [14]. Enhancers were initially defined by their ability to increase transcription from a target promoter independently of their orientation and, to a large extent, their position relative to the regulated gene [15, 16]. In contrast to core promoters, which direct transcription initiation at a TSS, enhancers were therefore traditionally regarded as distal regulatory sequences that modulate promoter activity. Their function depends on the binding of sequence-specific transcription factors (TFs) and cofactors, in turn recruiting chromatin remodelling complexes and components

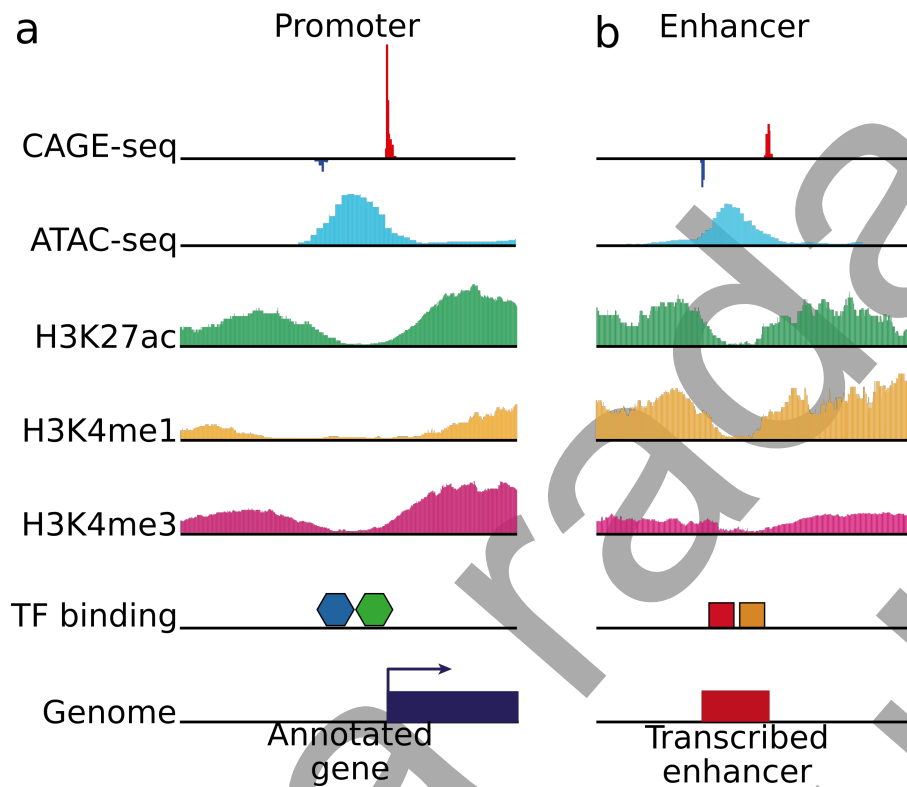


Figure 1.2. Schematic representation of CAGE-seq, ATAC-seq, histone modification ChIP-seq, TF binding, and genome annotation. a Active promoters are characterised by asymmetric CAGE peaks, with the dominant peak corresponding to stable transcription of polyadenylated mRNA. In contrast, active enhancers exhibit bidirectional, symmetrical CAGE peaks with a weaker signal due to the transient nature of eRNAs, which are rapidly degraded. Adapted from Guerrini et al. [13].

of the transcriptional machinery [8].

Regulatory elements can also be characterized by their chromatin state. Active promoters are commonly enriched for trimethylation of histone H3 at lysine 4 (H3K4me3), whereas enhancers are generally associated with monomethylation at the same residue (H3K4me1) [17, 18]. Acetylation of histone H3 at lysine 27 (H3K27ac) is frequently used to distinguish active enhancers from inactive or poised regulatory elements [19, 20]. However, these histone modifications are not exclusive to a single class of regulatory element, and their genomic distributions frequently overlap [21].

In addition to their characteristic chromatin profiles, many active enhancers are transcribed by RNA polymerase II. This transcription generally produces short, low-abundance and unstable transcripts referred to as enhancer RNAs (eRNAs). Enhancer transcription frequently initiates in both directions from a central regulatory region, resulting in pairs of divergent transcripts. These

RNAs are often rapidly degraded by the nuclear exosome and therefore usually are detected at much lower levels than stable gene transcripts [22].

In contrast to promoters, active enhancers generally produce weaker CAGE signals on opposite strands, forming a characteristic bidirectional pattern around the centre of the regulatory element (Figure 1.2b). The FANTOM5 consortium used this property to identify transcribed enhancers across a broad range of human cell types and tissues [12, 22]. The resulting atlas of transcribed enhancers provides a measure of enhancer activity across primary cells, cell lines and tissues (biosamples) and can be used to examine correlations between enhancer and promoter transcription across biological samples. Although promoter transcription is generally stronger and more directionally biased than enhancer transcription, the shared ability of both elements to recruit RNA polymerase II and initiate transcription has contributed to an increasingly blurred distinction between promoters and enhancers [23].

Silencers are regulatory elements that decrease transcription through the recruitment of transcriptional repressors and chromatin-modifying complexes [8]. Compared with promoters and enhancers, silencers remain less well characterized. However, a recent large-scale ENCODE study substantially expanded their annotation and demonstrated that *cis*-regulatory elements may function as either enhancers or silencers depending on the cellular context [6]. This context-dependent regulatory activity suggests that the same regulatory element may contribute to either activation or repression of gene expression, an observation that is considered throughout this thesis when analysing enhancer–promoter associations. Throughout this thesis, we indiscriminately use enhancers to refer to regulatory elements of both enhancing and silencing effect.

Despite their shared molecular properties, enhancers are often separated from their target promoters by considerable genomic distances. Their regulatory effects therefore cannot be inferred solely from linear genomic proximity. Enhancers may regulate promoters located hundreds of kilobases or even megabases away. Consequently, linear genomic distance alone is insufficient to infer regulatory relationships between enhancers and their target genes.

1.2 Long-range transcriptional regulation

1.2.1 Distal enhancer-promoter interactions

Before an enhancer can influence transcription, it must first become accessible to TFs. In inactive chromatin, DNA is wrapped around nucleosomes, limiting access to transcription factor binding sites. Pioneer TFs are capable of recognizing their binding motifs within compact chromatin and recruiting chromatin-remodelling complexes that reposition or evict nucleosomes, thereby increasing DNA accessibility [24]. Accessible chromatin can subsequently be detected experimentally using DNase I hypersensitivity assays or ATAC-seq, both of which have become widely used for identifying active regulatory regions [25, 26].

One of the earliest and best-characterized examples of long-range transcriptional regulation is an enhancer located within an intron of the *LMBR1* gene that regulates the expression of *SHH* over a distance of more than 1 megabase [27]. Studies of this regulatory element demonstrated that point mutations within the enhancer can result in congenital limb malformations despite leaving the coding sequence of *SHH* unaffected [27]. This finding demonstrated that regulatory elements may act over very large genomic distances and that assigning the nearest gene to a regulatory variant may be misleading.

1.2.2 Topologically associated domains

It is widely accepted that long-range chromatin contacts are established through chromatin looping mediated by the interplay of the cohesin complex and the CCCTC-binding factor (CTCF) (Figure 1.3 a) [28]. Cohesin is a ring-shaped protein complex that encircles DNA and extrudes chromatin loops by progressively pulling DNA through the complex. CTCF is a highly conserved DNA-binding protein that acts as a barrier to loop extrusion by binding specific DNA motifs. Stable chromatin loops are preferentially formed when cohesin encounters convergently oriented CTCF-binding sites, in turn halting further loop extrusion and stabilizing the interaction [28, 29].

Chromatin conformation capture methods have demonstrated that the genome is organized

hierarchically, with topologically associating domains (TADs) representing one of the fundamental levels of chromatin organization (Figure 1.3 b) [30–32]. TADs are genomic regions that can span several megabases and are characterized by a high frequency of interactions between loci within the same domain compared with loci located in neighbouring domains [32, 33]. TAD boundaries provide spatial insulation that plays an important role in restricting enhancer-promoter interactions (Figure 1.3 c). By increasing the probability of contacts between regulatory elements and promoters located within the same domain while limiting interactions across domain boundaries, TADs facilitate robust and specific gene regulation [33, 34].

A higher level of chromatin organization is represented by A and B compartments (Figure 1.3 b). These large-scale nuclear compartments broadly correspond to transcriptionally active (A) and transcriptionally inactive (B) chromatin. Unlike TADs, compartment formation appears to

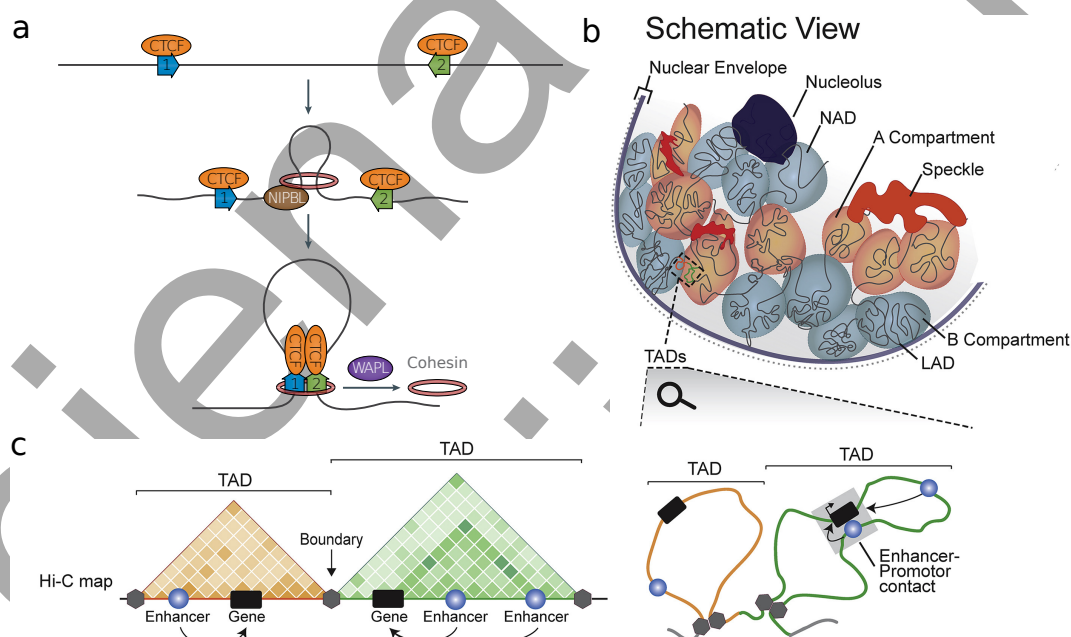


Figure 1.3. a Schematic illustration of the loop extrusion model. Following loading onto chromatin by NIPBL, cohesin extrudes chromatin through its ring-shaped structure until it encounters convergently oriented CTCF binding sites, where loop extrusion is halted. Cohesin is subsequently released by the WAPL complex. b Hierarchical organization of the genome within the nucleus. The upper panel depicts transcriptionally active A (red) and inactive B (blue) chromatin compartments together with nuclear landmarks, including lamina-associated domains (LADs), nucleolus-associated domains (NADs), and nuclear speckles. The lower panel illustrates topologically associating domains (TADs) and enhancer–promoter interactions. c Schematic Hi-C contact map showing two neighbouring TADs separated by an insulating boundary. Chromatin interactions occur more frequently within TADs than across TAD boundaries, facilitating enhancer–promoter contacts within the same domain. Panel a adapted from Schoenfelder and Fraser [33]; panels b and c adapted from Robson, Ringel, and Mundlos [35].

be largely independent of cohesin-mediated loop extrusion, and the mechanisms responsible for their establishment and maintenance remain incompletely understood [33, 35].

1.3 Evolutionarily conserved regulatory landscapes

1.3.1 Conserved non-coding elements

Specific regions of the genome display exceptionally high levels of non-coding sequence conservation across large evolutionary distances and are referred to as conserved non-coding elements (CNEs). Their high degree of conservation indicates strong evolutionary constraint and suggests that these sequences perform important genomic functions. Functional assays have shown that CNEs are enriched for *cis*-regulatory activity and that many act as long-range enhancers of developmental genes [36, 37]. Consistent with their regulatory importance, mutations affecting CNEs have been associated with developmental abnormalities, congenital disease, and complex neurological or behavioural phenotypes [37].

CNEs were initially identified through comparative genomic analyses, often using pairwise sequence comparisons between evolutionarily distant species. Their computational definition is commonly based on two criteria: a minimum percentage of sequence identity and a minimum sequence length over which this identity is maintained [38, 39]. Consequently, the number and genomic distribution of identified CNEs depend on the species being compared and on the selected sequence identity and length thresholds [40]. Alternative approaches have therefore been developed to identify non-coding conservation without relying on a single predefined threshold, including methods based on the statistical properties of whole-genome alignments, particularly pertinent to comparisons of recently diverged genomes [41]. Resources such as ANCORA further enable the visualization and comparison of CNEs identified across multiple pairwise genome comparisons [42].

1.3.2 Genomic regulatory blocks

A striking property of CNEs is that they are not distributed uniformly throughout the genome but frequently form dense clusters around genes with important roles in development [43]. These elements are maintained in conserved syntenic arrays with their target genes, suggesting that evolutionary selection acts not only on individual sequences but also on the organization of entire regulatory landscapes. Large genomic regions with arrays of CNEs, remaining linked in *cis* with their target genes are referred to as genomic regulatory blocks (GRBs) [42, 43]. GRBs may span several megabases and frequently extend across gene-poor regions or gene deserts. They contain complex long-range regulatory landscapes, with distal enhancers regulating target genes while bearing no effect on more proximal genes (Figure 1.4 a). These neighbouring genes, termed bystander genes, are generally unresponsive to the regulatory elements that control the GRB target gene.

The computational boundaries of GRBs are not always unambiguous because their identification depends on the definition and genomic distribution of CNEs. Consequently, different conservation thresholds and segmentation methods may produce somewhat different GRB boundaries. Several computational approaches have therefore been developed, including Hidden Markov Models that identify CNE-enriched genomic regions [44] and threshold-independent approaches based on the statistical properties of sequence conservation [41].

1.3.3 GRB target genes

GRBs typically contain one or a few target genes, many of which encode developmental genes, TFs or other regulators of multicellular development and differentiation [45]. These genes often exhibit complex spatial and temporal expression patterns and consequently require extensive regulatory input to ensure precise expression. Other genes located within the same GRB are referred to as bystander genes and are, by and large, unresponsive to the distal regulatory elements controlling the target gene [40, 45].

There are multiple examples where the disruption of long-range regulatory interactions within GRBs have been associated with a range of human phenotypes. Examples include altered

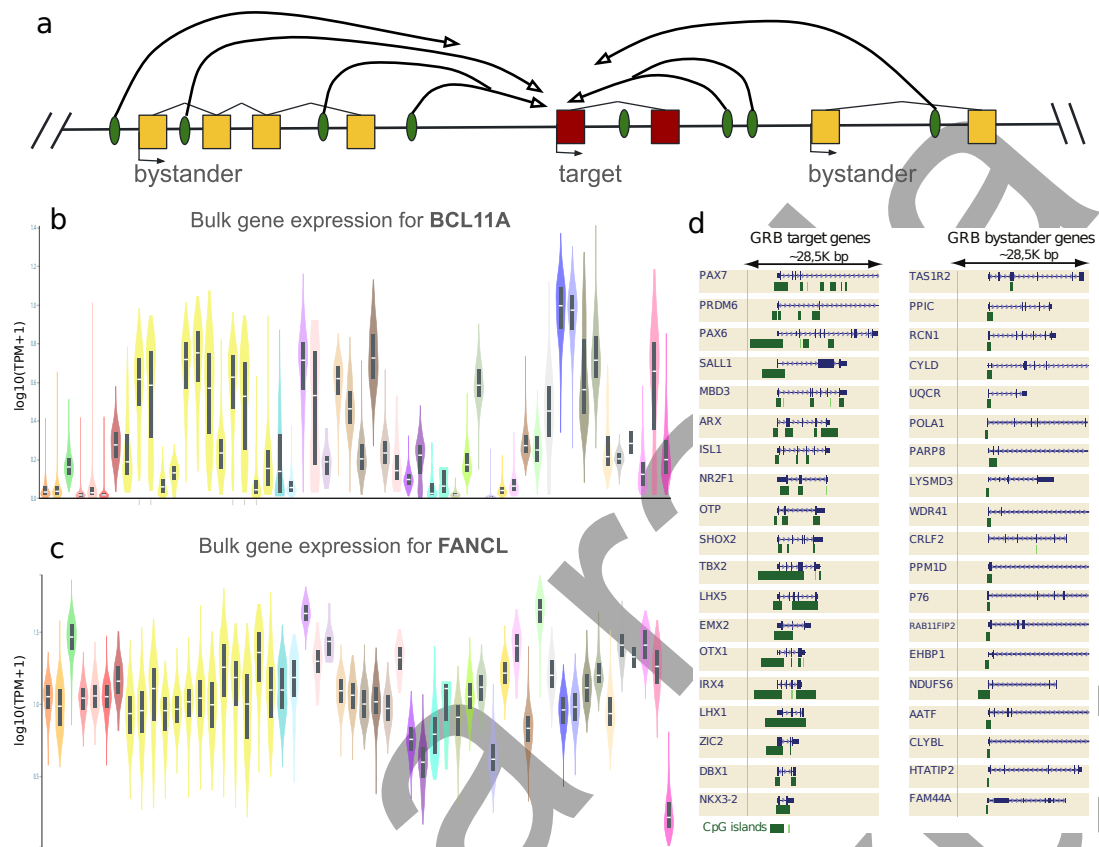


Figure 1.4. a Schematic representation of the GRB model. CNEs (green ovals) regulate the target gene while bypassing neighbouring bystander genes. b and c Gene expression profiles across human tissues and cell types for *BCL11A*, an example of a GRB target gene, and *FANCL*, an example of a bystander gene, illustrating the broader expression pattern of target genes compared with bystander genes. d Schematic comparison of CGIs associated with target and bystander genes, showing that CGIs overlapping target genes are generally more numerous and longer than those associated with bystander genes. Panel a was created using Google Slides; panels b and c were obtained from the GTEx Portal; panel d was adapted from Akalin et al. [45].

regulation of *SHH* in preaxial polydactyly [27], disruption of the *PAX6* regulatory landscape in aniridia [46], and long-range regulation of *IRX3* at loci associated with obesity and type 2 diabetes [47]. These examples illustrate how regulatory variants may influence developmental target genes while bypassing genes located closer to the affected regulatory element.

Target and bystander genes can also be distinguished by several genomic and transcriptional characteristics. GRB target genes are frequently involved in development, differentiation, cell adhesion, cell-cell signalling and axon guidance and often show intermediate gene-expression entropy (Figure 1.4 b), reflecting expression patterns that are neither ubiquitous nor restricted to a single cell type [40, 45]. In contrast, housekeeping and other ubiquitously expressed genes commonly occur as bystanders, and tend to show more uniform expression across tissues

(Figure 1.4 c). Target genes are also associated with distinctive promoter architecture, including longer or multiple CGIs that may extend from the promoter into the gene body, compared with neighbouring bystander genes (Figure 1.4 d) [45].

1.3.4 Relationship between GRBs and TADs

Harmston et al. demonstrated that GRBs are closely associated with the three-dimensional organization of the genome, with GRB boundaries frequently coinciding with TAD boundaries in both vertebrate and *Drosophila* genomes [44]. TADs overlapping GRBs are generally larger, contain fewer genes and exhibit stronger boundary insulation than other TADs, reflecting the extensive regulatory landscapes surrounding developmental genes.

These findings suggest that GRBs and TADs represent complementary views of the same regulatory architecture. Whereas GRBs describe the evolutionary conservation of long-range regulatory landscapes, TADs describe their three-dimensional organization within the nucleus. Together, they provide a structural and evolutionary framework for understanding long-range gene regulation [44].

1.4 Genetic variation and genome-wide association studies

1.4.1 Complex traits and disease

According to their genetic architecture, human diseases with a genetic component can be broadly classified into monogenic and complex disorders. Monogenic disorders are caused by pathogenic variants in a single gene leading to disruption of that gene's function, typically exhibiting large genetic effects and often resulting in rare, highly penetrant and early-onset diseases [48, 49]. In contrast, complex disorders are influenced by numerous genetic variants, each contributing a relatively small effect to the overall phenotype, together with environmental factors. Consequently, these are generally associated with common, later-onset diseases and traits with multifactorial aetiologies [50]. The small individual effect of each contributing variant, combined with environmental influences and interactions between genetic loci, makes

the genetic basis of complex traits substantially more difficult to dissect than that of monogenic disorders [49].

1.4.2 Genome-wide association studies

Advances in high-throughput genotyping technologies and the establishment of large population biobanks have enabled genome-wide interrogation of common genetic variation. Genome-wide association studies (GWAS) identify statistical associations between genetic variants, most commonly single nucleotide polymorphisms (SNPs), and phenotypic traits or diseases by comparing allele frequencies between individuals with different phenotypes (Figure 1.5 a) [51, 52]. During the past decade, the increasing availability of large, deeply phenotyped cohorts has substantially increased the statistical power of GWAS and led to a rapid growth in the number of reported associations [53]. National biobanks, particularly resources such as the UK Biobank and FinnGen, have been among the major drivers of population-scale genetic studies [54, 55]. The GWAS Catalog has become the principal public repository of published GWAS results (<https://www.ebi.ac.uk/gwas/>) and, as of the beginning of 2026, contains more than 7,000 studies and over one million reported associations [52].

It is not always straightforward to identify the causal variant underlying a GWAS association. GWAS typically report the lead (or tagging) SNP, defined as the variant showing the strongest statistical association with the studied phenotype. However, neighbouring variants are often highly correlated due to linkage disequilibrium (LD), the non-random association of alleles at different loci that are inherited together more frequently than expected by chance [57]. Consequently, a lead SNP usually represents a genomic locus rather than a single causal variant (Figure 1.5 b). A GWAS locus may therefore contain multiple correlated variants and several nearby genes, making it difficult to determine both the causal variant and its target gene.

1.4.3 Non-coding variants affect gene expression

Interpretation of coding variants is often relatively straightforward, as variants located within protein-coding exons may directly alter the amino acid sequence of the encoded protein, potentially modifying its function or resulting in complete loss of function. In contrast, more

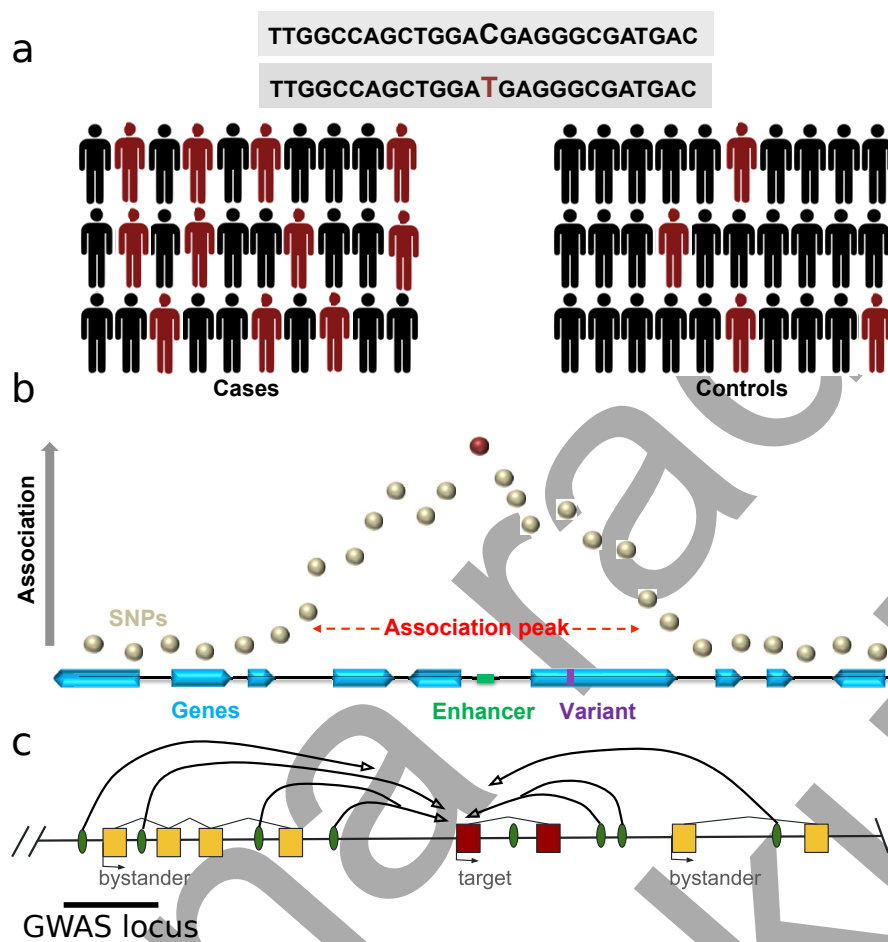


Figure 1.5. a Individuals affected by a disease or trait of interest are genotyped and compared with unaffected controls to identify SNPs associated with the phenotype. b Schematic representation of a GWAS association plot. The SNP showing the strongest association with the trait is highlighted in red, while neighbouring SNPs within the association peak are in LD with the lead variant. c Schematic representation of a GWAS locus within a GRB, illustrating how the nearest gene is often assigned as the candidate gene based on genomic proximity, whereas the GRB model predicts the distal target gene as the functional target of the associated regulatory variant. Panels a and b were adapted from Wangler, Hu, and Shulman [56]; panel c was created using Google Slides.

than 90% of variants identified by GWAS are located outside protein-coding regions, making it considerably more difficult to determine their functional consequences [51, 58]. Non-coding variants are enriched within regulatory regions [59], and a substantial proportion are thought to influence phenotype by altering gene expression rather than protein sequence. These variants may alter TF binding sites, chromatin accessibility or enhancer activity, ultimately leading to changes in gene expression.

As discussed previously in the Section 1.2, the three-dimensional organization of chromatin brings regulatory elements into close physical proximity with their target promoters despite

large genomic distances separating them in the linear genome. Consequently, disease-associated variants located within regulatory elements may influence genes situated hundreds of kilobases or even several megabases away. Nevertheless, disease-associated variants are still frequently assigned to the nearest gene during biological interpretation (Figure 1.5 c)[60]. This assumption is particularly problematic within GRBs, where long-range enhancer-promoter interactions are a defining feature.

Consequently, assigning non-coding variants to their target genes has become one of the central challenges in the interpretation of GWAS loci. Accurate variant-gene associations are essential for understanding the molecular mechanisms underlying disease, identifying affected biological pathways and prioritizing candidate therapeutic targets. As a result, considerable effort has been devoted to developing computational and experimental approaches for associating regulatory variants with their target genes [61], surveyed hereafter.

1.5 Methods for variant-gene association

To improve the interpretation of non-coding variants identified by GWAS, numerous computational methods have been developed to associate regulatory variants with their target genes. These approaches integrate complementary sources of evidence, including chromatin interaction data, chromatin accessibility, histone modifications, molecular QTLs and additional genomic annotations, to infer regulatory interactions and prioritize candidate target genes. Existing methods differ primarily in the type of biological evidence they integrate.

1.5.1 Chromatin interaction and enhancer-based methods

A prominent enhancer-based framework is the Activity-by-Contact (ABC) model, which predicts enhancer–gene interactions based on the simple notion that the regulatory effect of an enhancer depends both on its biochemical activity and on how frequently it comes into physical contact with the promoter of its target gene. The ABC score estimates the regulatory contribution of an enhancer as the product of its biochemical activity, quantified using chromatin accessibility (DNase-seq or ATAC-seq) together with H3K27ac chromatin immunoprecipitation followed by

sequencing (ChIP-seq) data, and its contact frequency with a gene promoter measured by Hi-C. The model was trained and experimentally validated using CRISPRi-FlowFISH perturbation experiments, which enable parallel perturbation of hundreds of candidate regulatory elements and quantification of their effects on target-gene expression [62]. The ABC framework has subsequently been widely applied to diverse human cell types and disease-associated loci for prioritizing high-confidence enhancer–gene and variant–gene associations [63].

GeneHancer is an integrated database of human enhancers and their predicted target genes. To link enhancers to genes, GeneHancer integrates multiple independent sources of regulatory evidence, including enhancer RNA co-expression, expression quantitative trait loci (eQTLs) and promoter-capture Hi-C data, to assign confidence scores to enhancer–gene associations and identify high-confidence (“elite”) regulatory interactions [64]. GeneHancer is integrated into the GeneCards and UCSC Genome Browser resources, facilitating the interpretation of regulatory variants and GWAS loci [64]. In contrast to the quantitative modelling performed by ABC, GeneHancer integrates multiple complementary evidence sources into a unified resource for annotating enhancer–gene associations.

1.5.2 Integrated prediction frameworks

Another widely used framework is the Open Targets Genetics locus-to-gene (L2G) model, which employs a gradient-boosted machine-learning approach to prioritize candidate causal genes at GWAS loci. Rather than explicitly modelling enhancer–promoter interactions, L2G integrates multiple complementary sources of evidence, including fine-mapped GWAS credible sets, molecular QTL colocalization, chromatin interaction data, genomic distance and variant pathogenicity, to generate a probability score for each candidate gene within a locus [65]. The resulting predictions are integrated into the Open Targets Platform, enabling users to prioritize candidate disease genes and evaluate their potential as therapeutic targets.

More recently, Gazal et al. developed the combined SNP-to-gene (cS2G) framework, by systematically evaluating numerous strategies for associating variants with genes and combining the highest-performing approaches into a single predictive model. By integrating information from coding regions, promoters, fine-mapped cis-eQTLs, enhancer–gene linking methods

including ABC and EpiMap, and chromatin accessibility-derived annotations, cS2G substantially improved the recall of disease genes while maintaining high precision, outperforming several previously available approaches [66].

Recent advances in artificial intelligence have also led to the development of genomic foundation models, large deep-learning models trained on DNA sequence together with large-scale functional genomic datasets. Models such as Enformer and the Nucleotide Transformer have demonstrated impressive performance across a wide range of sequence-based prediction tasks and can be adapted to numerous downstream applications [67–69]. Nevertheless, these approaches currently infer regulatory function primarily from sequence and associated genomic features and do not explicitly incorporate the dynamic three-dimensional chromatin organization or evolutionary conservation of regulatory landscapes.

1.5.3 Limitations of current approaches

Despite their considerable success, most existing methods rely predominantly on experimentally derived functional genomic datasets generated in specific tissues, cell types or developmental stages. Consequently, their predictions depend on the availability and quality of these datasets and may not fully capture regulatory interactions outside the assayed biological context. Consequently, long-range regulatory architecture and the evolutionary conservation of enhancer–promoter relationships are often under-represented.

1.6 GRB-based approaches for variant-gene association

The utility of GRBs for interpreting disease-associated variants has been demonstrated in several smaller studies spanning developmental disorders, metabolic disease and neuropsychiatric traits. Two representative examples illustrate how incorporating long-range regulatory architecture can reveal causal target genes that would not be identified using genomic proximity alone.

1.6.1 Type 2 diabetes and obesity associated variant involved in long-range regulation

In 2010, Ragvin et al. have applied the GRB model to identify the affected gene of non-coding variants highly associated with type 2 diabetes and obesity [47]. They re-examined SNPs falling into the intronic region of the *FTO* gene, initially assumed to represent the causal gene because the associated variants were located within its introns. However, this regions falls into a GRB and, furthermore, the LD block of these variants span several deeply conserved CNEs (Figure 1.6). The target gene for this GRB is *IRX3*, and using a transgenic assay, the authors demonstrated that a regulatory region of the CNEs drives the expression of the reporter gene in a similar temporal-spatial manner as the expression of *IRX3*, while the *FTO* expression was not detected [47].

Two later studies showed further evidence and mechanistically explained the involvement of the long-range enhancer in regulating *IRX3* rather than *FTO*. Smemo et al. [71] performed a chromatin conformation capture experiment in developing mouse embryos and shows that indeed the disease-associated locus forms frequent contacts with the promoter region of *IRX3* (Figure 1.6) [71]. Claussnitzer et al. [72] next demonstrated, using CRISPR-Cas9 editing that the obesity-associated SNP rs1412085 disrupts a conserved motif of the ARID5B repressor which leads to a high expression of *IRX3* during early adipocyte differentiation (Figure 1.6). Consequentially, this leads to a developmental shift from energy-dissipating adipocytes to energy-storing white adipocyte, as well as an increase in lipid storage (Figure 1.6) [72].

This example illustrates the importance of considering GRBs in interpreting GWAS SNPs since the role of long-range regulation is often overlooked. Interestingly, the original GRB-based interpretation proposed by Ragvin et al. [47] preceded the mechanistic studies by several years but was not cited in these subsequent investigations.

1.6.2 Extending GRBs with enhancer-promoter associations

Although the GRB framework successfully identified candidate target genes through evolutionary conservation, it did not directly prioritize individual enhancer–gene interactions. Barešić et al.

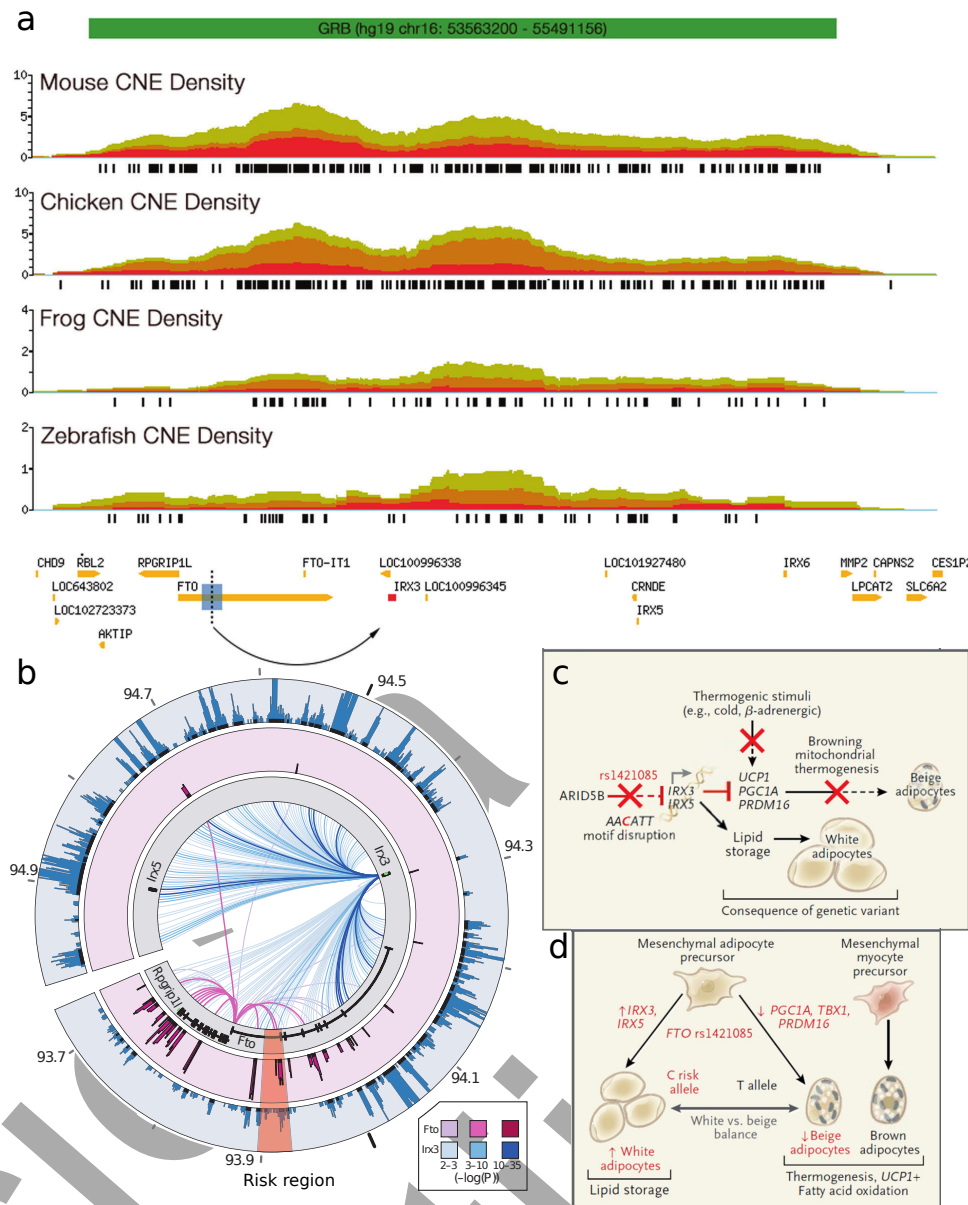


Figure 1.6. a Genomic view of the *IRX3* GRB with four tracks showing CNE densities based on pairwise comparisons between the human genome and progressively more evolutionarily distant vertebrates: mouse, chicken, frog, and zebrafish. Genes within the region are shown below, with the obesity-associated locus overlapping *FTO* highlighted in blue and the target gene *IRX3* highlighted in red. b Chromatin contact map spanning the *FTO*, *IRX3*, and *IRX5* locus. The highlighted obesity-associated region forms long-range chromatin contacts with the promoter of *IRX3*. c The rs1421085 risk variant disrupts the binding motif of the ARID5B repressor, resulting in increased expression of *IRX3* and *IRX5*. This inhibits adipocyte browning and mitochondrial thermogenesis while promoting lipid storage in white adipocytes. d Proposed effect of the rs1421085 risk allele on adipocyte precursor differentiation, favouring white adipocyte development at the expense of thermogenic beige and brown adipocytes. Panel a was adapted from Barešić et al. [70]; panel b from Smemo et al. [71]; panels c and d from Claussnitzer et al. [72].

[70] extended this concept by integrating GRBs with enhancer-promoter association scores derived from FANTOM5 transcriptional activity profiles. They applied this framework to re-evaluate schizophrenia-associated GWAS loci and identify candidate genes regulated through long-range transcriptional interactions [70].

Using this combined approach, Barešić et al. [70] identified a novel pathway enriched for neuronal transcription factors that was not apparent in the original GWAS interpretation. One illustrative example involved a schizophrenia-associated SNP located within an intron of the *VRK2* gene, originally assigned as the likely target based on genomic proximity. In contrast, the GRB-based framework associated the locus with the developmental transcription factor *BCL11A* and identified a long-range enhancer predicted to regulate *BCL11A* despite being located more than 1 megabase away [70].

Together, these studies demonstrated that combining evolutionary conservation with functional evidence of enhancer activity substantially improves the biological interpretation of non-coding GWAS loci. Building on this concept, the targPred framework was developed to extend the approach from individual diseases to the complete GWAS Catalog, enabling systematic prioritization of candidate target genes across a broad range of human phenotypes [73].

To showcase utility of such tool, two pertinent biomedical examples were chosen. Head and neck squamous cell carcinoma (HNSCC) example demonstrates how multiomics datasets benefit from the inclusion of long-range regulatory effects on gene transcriptional activity and epigenetics, while Parkinson's disease (PD) single-cell transcriptome dataset is used to identify pathways affected by dysregulation of GRB target genes and a set of newly detected genomic variants is used to exemplify SNP–gene assignment problem on a single disease.

1.7 Disease of special interest: head and neck squamous cell carcinoma

1.7.1 Disease overview

Head and neck cancer ranks as the sixth most common malignancy globally, with HNSCC representing over 90% of all cases. Unfortunately, HNSCC is frequently diagnosed at an advanced stage, when treatment is more difficult, contributing to a suboptimal 5-year survival rate of approximately 66% [74]. Although this cancer originates from a single cell type, the squamous epithelial cell, HNSCC exhibits remarkable heterogeneity [75]. Tumour heterogeneity is influenced by the primary site of origin, with the oral cavity, oropharynx, larynx, and sinonasal tract being the most common locations [76]. In addition to anatomical site, infection with high-risk human papillomavirus (HPV) types, particularly in the oropharynx, substantially shapes HNSCC biology and patient survival [77]. HPV-positive and HPV-negative HNSCC differ markedly in their aetiology, genetic alterations, and epigenetic profiles [74, 78]. HPV-positive HNSCC tends to affect younger individuals with little or no history of smoking or alcohol consumption [79]. HPV-negative tumours arise at various sites, HPV-positive HNSCC predominantly originates in the oropharynx and has been associated with improved overall survival and better therapeutic response in many Western countries [80]. The incidence of HPV-positive oropharyngeal squamous cell carcinoma has risen substantially in Western populations, with HPV implicated in nearly 70% of cases, in turn likely contributing to the observed improvement in 5-year survival across all HNSCC cases [74, 81].

Although numerous studies have reported more favourable outcomes for patients with HPV-positive HNSCC, there remains no consensus on specific biomarkers that could enhance diagnostic, prognostic, or therapeutic strategies [82–84]. This lack of agreement may stem from differences in study populations, inconsistent documentation of tumour site, and insufficient stratification by HPV status. Accordingly, the Croatian cohort analysed in this thesis represents a distinct population characterised by higher rates of smoking and alcohol consumption [85, 86].

1.7.2 Epigenetic regulation

Epigenetic regulation refers to heritable and stable changes in gene expression that occur without alterations in the underlying DNA sequence, mediated instead by chemical modifications to DNA, histones, and chromatin architecture, as well as by non-coding RNAs [87]. The main epigenetic mechanisms include DNA methylation (DNAm), histone modifications (such as acetylation, methylation and phosphorylation), chromatin remodelling, and regulation by microRNAs (miRNAs) and long non-coding RNAs: these all act in combination to shape higher-order chromatin structure and determine the accessibility of genes to the transcriptional machinery [88–90]. These modifications are written, read and erased by dedicated enzyme systems — DNA methyltransferases, histone acetyltransferases and deacetylases, histone methyltransferases and demethylases, and chromatin-remodeling complexes — that dynamically integrate developmental cues and environmental signals to fine-tune gene expression patterns across cell types and over time [89, 91, 92]. Because epigenetic marks can be mitotically and sometimes meiotically inherited [93, 94], they provide an additional layer of information beyond the genome sequence, and their disruption contributes to diverse pathological states, including cancer, cardiovascular and metabolic diseases, by aberrantly silencing tumour-suppressor genes, activating oncogenes, or reprogramming signalling and stress-response pathways [87, 88, 92].

1.7.3 DNA methylation in cancer

In mammalian cells, maintenance of DNAm is largely carried out by DNMT1, de novo methylation by DNMT3A and DNMT3B, and active DNA demethylation by TET family dioxygenases that oxidize 5-methylcytosine to successive intermediates, while methyl-CpG-binding proteins and extensive crosstalk with histone modifications and chromatin architecture integrate DNAm into broader epigenetic networks [95]. CpG dinucleotides often cluster in promoter-associated CpG islands whose methylation status is tightly linked to transcriptional activity and plays central roles in genomic imprinting, X-chromosome inactivation, transposon silencing, and tissue-specific gene expression [96].

In cancer, as well as HNSCC, DNAm patterns become profoundly disrupted, with two

prominent patterns emerging in various cancers: genome-wide hypomethylation and hypermethylation of promoter-associated CGIs. Hypomethylation of gene deserts, repeat sequences, transposons or CpGs located in introns are connected to genome instability [97]. For example, normal mammalian cells maintain high methylation levels in retrotransposon elements, and hypomethylation may lead to activation and subsequent destabilization of the genome facilitating carcinogenesis. On the other hand, hypermethylation of promoter-associated CGIs leads to transcriptional silencing of tumour-suppressor genes and regulatory non-coding RNAs, in turn promoting transcriptional transformation [98].

1.7.4 microRNA in cancer

miRNAs are short, single-stranded non-coding RNAs of about 22 nucleotides that regulate gene expression post-transcriptionally by base-pairing to messenger RNAs and promoting their degradation or translational repression [99, 100]. They are generated through a conserved biogenesis pathway where primary miRNA transcripts are processed by the Drosha microprocessor complex into precursor miRNAs and then by Dicer into mature miRNAs. These are then loaded into Argonaute-containing RNA-induced silencing complexes to fine-tune networks controlling proliferation, differentiation and apoptosis [99, 101].

In cancer, a widespread downregulation of tumour-suppressor miRNAs and overexpression of oncogenic miRNAs lead to altered global miRNA expression profiles. These changes that contribute to many hallmarks of malignancy and can stratify tumours by tissue of origin, molecular subtype and clinical outcome [99, 100]. In HNSCC, DNAm and miRNA-mediated post-transcriptional control jointly shape the tumour gene expression landscape and contribute to the emergence of distinct molecular subgroups and clinical phenotypes [98, 102, 103]. Promoter CGI methylation modulates transcription of oncogenes and tumour-suppressor genes, and integrative multi-omics studies have shown that changes in methylation at specific loci are tightly correlated with altered mRNA expression profiles along tumour progression and metastasis [98]. In parallel, deregulated miRNAs fine-tune many of the same pathways at the post-transcriptional level, with recurrent oncogenic and tumour-suppressor miRNAs controlling networks involved in cell-cycle regulation, DNA repair, epithelial–mesenchymal transition and

therapy resistance [104, 105]. DNAm can also directly silence miRNA genes, while subsets of miRNAs target DNA methyltransferases, histone modifiers and chromatin regulators, creating feedback circuits that couple methylation states to miRNA expression and downstream gene regulation [106, 107].

Taken together, these findings underscore that a comprehensive investigation of HNSCC must include epigenetic regulation — particularly DNAm and miRNA networks — when interpreting gene-expression changes and their impact on tumour biology [98, 103].

1.8 Disease of special interest: Parkinson's disease

1.8.1 Disease overview

Parkinson's disease (PD) is the second most common neurodegenerative disorder, affecting more than six million people worldwide and representing a major public health challenge [108]. It is a progressive disorder that predominantly affects individuals over 60 years of age and is characterised by a combination of motor symptoms, including bradykinesia, rigidity, tremor, and postural instability, together with a range of non-motor symptoms such as impaired olfaction, dementia, insomnia, constipation, and depression [109].

Only 5-10% of PD cases are familial, while the majority are idiopathic, with disease onset thought to result from a complex interplay between genetic and environmental factors [110]. Previous studies have implicated oxidative stress, mitochondrial dysfunction, and immune activation in disease development and progression, however, the precise molecular mechanisms underlying PD remain incompletely understood [111, 112].

Neuropathologically, PD is characterised by the accumulation of α -synuclein aggregates, known as Lewy bodies, and the progressive degeneration of dopaminergic neurons in the substantia nigra pars compacta. Increasing evidence suggests that α -synuclein pathology triggers neuroinflammatory responses, leading to microglial activation and disruption of neuronal homeostasis. Sustained neuroinflammation is thought to contribute to the progressive loss of dopaminergic neurons and has emerged as a key mechanism in PD pathogenesis [113, 114].

1.8.2 Immune dysregulation in Parkinson's disease

In addition to resident microglia, increasing evidence suggests that peripheral immune cells contribute to neuroinflammation in PD. Dendritic cells (DCs) function as professional antigen-presenting cells, linking the innate and adaptive immune systems through antigen presentation to CD4⁺ T-cells via major histocompatibility complex class II molecules [114]. Consequently, DCs regulate T-cell activation and differentiation, thereby shaping downstream immune responses.

Studies have demonstrated that both DCs and CD4⁺ T-cells infiltrate the central nervous system of patients with PD, where they influence microglial activation and the progression of neurodegeneration [115, 116]. This interaction promotes a pro-inflammatory environment characterised by increased cytokine production and sustained immune activation, which may further exacerbate neuronal loss. The growing evidence for peripheral immune involvement in PD highlights the importance of understanding the functional states and transcriptional programmes of these immune cell populations.

1.8.3 High-throughput sequencing in Parkinson's disease

Recent advances in genomics and transcriptomics have provided valuable insight into the molecular mechanisms underlying complex diseases, including PD [117]. In particular, the development of single-cell RNA sequencing (scRNA-seq) has enabled transcriptomic profiling at cellular resolution, allowing the identification of distinct cell populations and characterisation of their transcriptional states [118]. Compared to bulk RNA sequencing, scRNA-seq resolves cellular heterogeneity and enables the detection of cell type-specific transcriptional changes that may otherwise remain masked.

Single-cell studies have increasingly implicated peripheral immune dysregulation in PD. T-cells from patients with PD have been shown to recognise α -synuclein-derived peptides, suggesting antigen-specific immune activation [119]. Furthermore, single-cell transcriptomic analyses have identified clonal expansion of cytotoxic CD4⁺ and CD8⁺ T-cells, alterations in natural killer cell populations associated with disease severity, and distinct transcriptional profiles of α -synuclein-reactive T-cells [120, 121]. Similar approaches have revealed expan-

sion of effector T-cell populations in Alzheimer's disease, suggesting that peripheral immune dysregulation may represent a common feature of neurodegenerative disorders [122, 123].

While transcriptomic profiling provides insight into disease-associated cellular states, genomic analyses offer complementary information by identifying genetic variants associated with disease susceptibility. Together, single-cell transcriptomics and exome sequencing enable the investigation of both transcriptional alterations and their potential genetic determinants, providing a comprehensive framework for studying the molecular mechanisms underlying PD.

2 Aims and hypotheses

GRBs have an important role in human complex disease and utilising high-throughput molecular data will aid in understanding the mechanisms of long-range transcriptional regulation.

- H1. There is a set of phenotypic traits, for which disease mechanisms are influenced and/or dominated by long-range regulation of transcription of the GRB target genes.
- H2. The GRB model and long-range regulation can give new mechanistic insight into the role of epigenetics in transcriptional regulation in HNSCC.
- H3. Target gene expression regulated through long-range transcriptional regulation has an important role in driving immunological response in immune cell types in PD.

The main aim of this work is to better understand the role of GRBs in complex human disease by applying statistical methods on high-throughput data. This is achieved through identification of phenotypic traits that have disease-associated loci enriched in GRBs, in turn pinpointing traits for which the assignment of the correct causal gene is missed (H1). Furthermore, two diseases will be showcased: HNSCC and PD. The HNSCC example will explore different epigenetic mechanisms (methylation and microRNA regulation) of transcriptional activity (H2). The PD example will exemplify the role of long-range transcription in immunological aspects of a complex disease (H3).

3 Methods

3.1 Methods for the targPred study

3.1.1 Empirical p-value for enhancer–gene links

enhancer–gene associations were quantified based on the CAGE data from the FANTOM5 project [22]. To account for both activating and repressive regulatory effects, we extended the empirical p-value (empP) method previously described by Barešić et al. [70]. The original FANTOM5 dataset comprised 1,046 biosamples, including cell lines, primary cells, and tissues, with time-series samples excluded. Technical and donor replicates were merged by calculating the median transcripts-per-million (TPM) value, resulting in a final dataset of 309 biosamples (full list available on the Github repository). Replicate samples were identified by removing the “donor” and “pool” identifiers from the sample names.

For each enhancer detected in the FANTOM5 dataset, all promoter genes located within ± 6.4 megabase (corresponding to the maximum GRB size) were considered as candidate targets. For each enhancer–gene pair, biosamples were divided into two groups according to enhancer activity: active enhancers (“E+”, enhancer CAGE signal > 0) and inactive enhancers (“E0”, enhancer CAGE signal $= 0$). The observed test statistic was defined as the signed difference between the median promoter expression in the two groups,

$$T_{\text{obs}} = \text{median}(\text{promoter in E+}) - \text{median}(\text{promoter in E0}). \quad (3.1)$$

Null distributions were generated by randomly permuting biosample group labels (“E+”/“E0”) 10,000 times, recomputing T_{perm} each iteration. The two-tailed p-value tested bidirectional

effects - activation and repression ($|T_{\text{perm}}| \geq |T_{\text{obs}}|$):

$$\text{empP} = \frac{r + 1}{\text{iter} + 1}, \quad (3.2)$$

where r denotes the number of permutations producing a test statistic at least as extreme as the observed value and $\text{iter} = 10,000$. The addition of a pseudocount ensures that empP are always greater than zero and reduces bias caused by the discrete nature of permutation testing [124].

3.1.2 GWAS processing

GWAS summary statistics were downloaded from the GWAS Catalog file repository using the data freeze available on 17.02.2026, including all studies present in the catalogue at that time [51, 52]. All downstream analyses were performed in R version 4.3.3. Only studies conducted in participants of European ancestry (as defined by the broad ancestral category) were included in subsequent analyses. SNPs involved in SNP-SNP interactions, multi-SNP haplotypes, or lacking an rsID were excluded. Studies reporting multiple traits were split so that each trait was analysed independently. In addition, reported and mapped genes were combined into a single set, hereafter referred to as GWAS genes.

Linkage disequilibrium (LD) blocks were calculated using the European population from the 1000 Genomes Project Phase 3 (GRCh37) dataset [125]. Canonical autosomes and X chromosome were selected, and the genotype data were converted to BIM/BED/FAM format. LD blocks were computed for each GWAS SNP using PLINK v2.0 and was performed using a 500 kb window and an r^2 threshold of 0.6. For each GWAS SNP, the corresponding locus was defined by its associated LD block.

3.1.3 Identification of GWAS loci within genomic regulatory blocks

GRB coordinates and their corresponding target genes were obtained from Tan and Lenhard [40]. Briefly, human-dog pairwise comparison GRBs were used, defined from CNEs sharing at least 98% sequence identity across a 50-nucleotide window. GWAS loci were considered to reside within a GRB if they overlapped the GRB by at least one nucleotide [70]. GRB target genes are

distinguished from bystander genes by several characteristic features, including longer CGIs, a greater number of alternative transcription start sites with wider spacing, and a characteristic combination of transcription factor binding sites within their promoter regions. The list of GRB target genes was obtained from Tan and Lenhard [40]. All remaining genes located within a GRB were classified as bystander genes.

3.1.4 Identification of long-range regulatory traits

Identification of traits exhibiting long-range regulation was based on comparing locus-associated enhancer–gene pairs located within GRBs with those outside GRBs. First, for each trait, locus-associated enhancers were identified and classified according to whether they overlapped a GRB. Second, significant enhancer–gene pairs were identified for each enhancer using an significance threshold of $empP < 0.001$. Raw $empP$ values were used without multiple-testing correction, together with a more stringent significance threshold. This approach was necessary because multiple-testing correction could not be applied consistently to enhancer–gene pairs located outside GRBs on the basis of GRB regions.

For each trait, the median genomic distance of enhancer–gene pairs was calculated separately for loci inside and outside GRBs. Differences between the two groups were assessed using the Wilcoxon rank-sum test, and resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure to control the false discovery rate. Finally, disease traits were assigned to higher-level categories using the Experimental Factor Ontology (EFO) and MONDO ontologies.

Within the EFO ontology, the terms *overnutrition* (MONDO:0003916) and *diabetic eye disease* (EFO:0009486) are placed within the aetiologic and body system subtrees respectively, rather than being classified according to developmental or physiological processes. For the purposes of this analysis, these traits were therefore included in the metabolic phenotype category. They are presented in italics throughout the Results (Chapter 4) to distinguish them from true descendants of the metabolic disease (MONDO:0005066) ontology term.

3.1.5 Locus enrichment in genomic regulatory blocks

Enrichment of trait-associated loci within GRBs was assessed using genome shuffling implemented in `bedtools` v2.30. GRB regions were randomly shuffled 10,000 times while preventing overlap between shuffled regions and both the original GRBs and one another. To evaluate enrichment, only GWAS studies with at least 30 associated SNPs were included. For each permutation, the number of loci overlapping the shuffled GRB regions was counted to generate a null distribution. Empirical p-values were then calculated using equation (3.2) with r defined as the number of permutations with the number of overlaps at least as extreme as that observed for the original GRB regions.

3.1.6 Validated SNP–gene association datasets

Two reference datasets of validated SNP–gene associations were used for performance evaluation. The first dataset was constructed from disease-associated coding variants deposited in the ClinVar database [126]. Variant call format files were downloaded from the NCBI public repository using the data freeze of 09.03.2026 [126] and processed to remove entries lacking a specified disease annotation. Disease ontologies were then used to map ClinVar disease annotations to EFO which is the ontology used for GWAS traits. While a limited number of direct mappings were available, most disease terms were matched through synonym mapping using the MONDO ontology [127]. Since the ClinVar and GWAS Catalog resources use different disease annotation systems, the MONDO ontology was used as an intermediary framework to harmonize disease concepts and enable consistent cross-mapping to EFO. The resulting dataset consisted of trait–gene associations successfully mapped to corresponding GWAS traits.

The second validation dataset comprised experimentally supported SNP–gene associations obtained from Alsheikh et al. [128]. Only SNP–trait pairs present in the GWAS dataset and overlapping GRBs were retained for subsequent analyses.

3.1.7 Comparison with existing SNP–gene prediction methods

To compare existing enhancer–gene prediction methods with targPred, each enhancer–gene pair was assigned to one of three targPred categories:

- **Category 0:** the enhancer–gene pair was not significant in our dataset and did not include a GRB target gene.
- **Category 1:** the enhancer–gene pair was significant in our dataset or included a GRB target gene.
- **Category 2:** the enhancer–gene pair was significant in our dataset and included a GRB target gene.

Genome-wide ABC enhancer–gene scores, calculated across 131 biosamples, were downloaded from the official repository (<https://mitra.stanford.edu/engreitz/oak/public/Nasser2021>). The ABC dataset contains predicted distal enhancer–gene connections with scores > 0.015 within a 2 megabase window [62, 63]. GeneHancer enhancer–gene scores generated for the original publication were obtained directly from the authors upon request [64]. GeneHancer scores show a discontinuous distribution, with most of the scores ranging from 0 to 50, and a distinct second group of enhancer–gene scores with values > 250 (Figure 4.4c). For comparison with targPred, only enhancer–gene pairs from the second group were filtered. For both ABC and GeneHancer datasets, only enhancers overlapping GRBs were retained for further analysis.

Genome-wide cS2G SNP-trait-gene scores were obtained from the published dataset (<https://zenodo.org/records/7754032>) [66]. For each SNP located within a GRB, the corresponding GWAS locus was assigned, and locus-associated enhancer–gene pairs were classified into targPred categories using the same procedure described above.

3.1.8 Visualization of the childhood obesity locus

The chromatin contact map for the childhood obesity locus was generated using Hi-C data from fully differentiated human fibroblasts at a resolution of 10 kb [129]. Lift-over from hg38 human

genome assembly to hg19 version of the region was performed using the rtracklayer v1.54. The genomic region was visualized with the plotgardener v1.0, while two-sided gene expression distribution plots were adapted from the targPred resource (<http://targpred.irb.hr/>).

3.1.9 Statistical testing

The significance of enhancer–gene associations was assessed using the empirical permutation-based approach described above, and $\text{emp}P$ were adjusted for multiple testing at the GRB level using the BH procedure. Two-sided Wilcoxon rank-sum tests were used for pairwise comparisons between groups, and the resulting p -values were adjusted for multiple testing using the BH procedure where applicable. Unless otherwise stated, statistical significance was defined as an adjusted $p < 0.05$. For the identification of traits exhibiting long-range regulation, a more stringent significance threshold of $\text{emp}P < 0.001$ was applied when selecting significant enhancer–gene pairs. Location shift when comparing cS2G scores between targPred classes was calculated using Hodges-Lehmann estimate.

3.2 Methods for the head and neck squamous cell carcinoma study

3.2.1 Croatian HNSCC sample processing

In this thesis we used a Croatian cohort of HNSCC patients treated at Clinical Hospital Dubrava between 2015 and 2018. A total of 61 primary tumours from the oral cavity (O) and oropharynx (OP) were collected, together with non-malignant tissue from individuals without HNSCC. The whole genome microRNA profile (miRNome) and methylation array (methylome) datasets used here were generated in previous studies on this cohort and are only briefly summarised below. Detailed descriptions of sample collection, wet-lab processing and initial analyses are provided elsewhere [85, 86].

For miRNome profiling, 22 tumour samples and three tonsil controls were sequenced, two tumours were subsequently excluded (one recurrent tumour and one outlier) as reported in the

original study. From the corresponding methylome study, 12 of the miRNA-profiled tumours plus one additional tumour with methylation data only, and eight healthy oral mucosa swabs as controls were included in methylation profiling using Infinium MethylationEPIC BeadChip array (Illumina). For the whole genome RNA-seq (transcriptome) analysis, 10 tumours (six O and four OP) with existing miRNome and methylome data were selected, together with three tonsil control tissues from the miRNome dataset; HPV status for all samples was determined by polymerase chain reaction as described previously [85]. In the present work, we re-used these three omics layers as the basis for integrative bioinformatic analysis.

3.2.2 Sequencing data analysis

Transcriptome profiling of the 13 selected samples was performed using TruSeq Stranded mRNA libraries (Illumina) sequenced on a NextSeq 500 instrument with NextSeq High 150-cycle reagents in paired-end mode on a high-output flow cell, debarcoding was carried out on the Illumina BaseSpace platform as part of the original experimental workflow. FastQC v0.11.9 was used for quality control of sequencing data. Subsequently, the alignment was performed using the GRCh37 human assembly (downloaded from GENCODE, released in 2013). Gene counts were quantified using salmon v1.5.0 with default parameters; an average of 25.2 ± 2.2 million fragments per sample (Supplementary Table S1).

Analyses were conducted in R v4.2.2 using Bioconductor packages. Gene-level transcript counts were derived from salmon v1.5.0 quantifications, miRNA counts were reprocessed for the subset of samples described by [85], and differentially methylated (DM) CpG sites from [86] were re-analysed and extended to CGI-level summaries. Count normalisation and group-wise differential expression for mRNA and miRNA data were performed with DESeq2 v1.36, after pre-filtering genes and miRNAs with fewer than 10 total counts across all samples. Variance-stabilised counts were obtained using the DESeq2 regularised logarithm (rlog) transformation and used as input for principal component analysis (PCA) and Euclidean sample distance calculations. Differential expression was computed for tumour versus control samples, and differentially expressed genes (DEGs) and miRNAs (DEmiRs) were defined with the significance threshold of BH-adjusted *p-values* < 0.05 .

Gene set enrichment analysis on DEGs was performed and visualised using gProfiler2 v0.2.1, with all annotated human genes as the background for the hypergeometric test. We applied the tailor-made multiple-testing correction method Set Counts and Sizes (SCS) [130], applied the significance threshold of $p < 0.05$, and assessed enrichment across Gene Ontology: Biological Processes (GO:BP) [131], Kyoto Encyclopedia of Genes and Genomes (KEGG) [132], Reactome [133], miRTarBase [134] and TRANSFAC [135], and the same parameters were used for all subsequent gene set enrichment analyses. Tumour suppressor genes were searched against the TSGene 2.0 database [136]. Identification of miRNA targets was carried out by querying significant DEGs against DE miRs with multiMiR v1.18.0 and the miRTarBase database, retaining only experimentally validated miRNA–gene interactions. Previously published DNAm data [86] were imported with RnBeads v2.14.0, and differential methylation (DM) between tumour and control groups was evaluated at CpG sites and CGIs. CpG sites were grouped into CGIs using RnBeads hg19 v1.28.0 annotation, and results are reported at the CGI level unless otherwise stated. For each CGI, DM was defined as the quotient of mean methylation between groups across all constituent CpG sites, with significance assessed at BH-adjusted $p < 0.05$. Promoter regions were defined as 1,500 bases upstream and 500 bases downstream of transcription start sites, and CGIs overlapping these promoter intervals were designated as promoter CGI targets.

3.2.3 Clustering and validation

To integrate the three omics layers and identify epigenetic mechanisms in HNSCC, we constructed a unified analysis pipeline centred on differentially expressed genes and their associated DE miR and DM-CGIs. We centred the analysis on the set of DEGs (Supplementary Data 5) and linked each gene to DE miRs and DM promoter-associated CGI using Spearman correlation between gene expression counts, miRNA expression counts and DNAm beta values across samples. When multiple miRNAs were associated with a given DEG, we retained only the miRNA showing the largest absolute Spearman correlation with gene expression. For genes with more than one associated CGI, we selected the CGI with the lowest DM p-value. This yielded a unified table containing DEGs and their corresponding DE miRs and/or DM-CGIs. We then

assessed the interplay between DEmiRs, DM-CGIs and DEGs in two complementary ways: (i) a simple overlap analysis that recorded whether each DEG was associated with DEmiR activity and/or CGI methylation changes and (ii) unsupervised k-means clustering. For clustering, we used log-transformed, centred and scaled values of DEG and DEmiR fold changes together with DNAm quotients, and ran k-means from the `stats` v4.2.2 package with 8–12 clusters and $nstart = 100$, selecting the best solution with 11 clusters. Survival analysis for all tumour patients was conducted with the `survival` v3.5 package, generating Kaplan–Meier curves under default settings and testing differences between oral versus oropharyngeal tumours and HPV-positive versus HPV-negative cases using the default G-rho test.

For external validation, we obtained the The Cancer Genome Atlas (TCGA) dataset of HNSCC samples from the Genomics Data Commons portal (<https://portal.gdc.cancer.gov/>), restricting the tumour site to tonsil, floor of mouth, tongue and oral cavity, hereafter termed the TCGA-HNSCC dataset. TCGA transcriptome and miRNome data were imported as raw gene and miRNA counts and processed with DESeq2, while methylome data were imported as beta values and analysed using `limma` v3.46. Subsequent validation focused on the 1,834 genes that were significantly differentially expressed both in our cohort and in TCGA-HNSCC, forming the basis for comparative expression and integrative clustering analyses (Figure 4.13).

Code used for this analysis, as well as TCGA data is deposited in Github repository (https://github.com/mlkr-rbi/HNSCC_Integrative_analysis).

3.2.4 Genome-wide CpG analysis

Analyses were performed on the human reference genome GRCh37, restricted to canonical autosomes and the X chromosome, using Illumina 450k CpG probe annotation and GRB coordinates with target/bystander information. Overlaps between probes and GRBs were identified with `GenomicRanges`, and probes were classified into promoter, gene-body and intergenic contexts using the UCSC (University of California, Santa Cruz) reference field (promoter defined as 1500 base pairs upstream and 500 downstream of the transcription start site; gene body as both exons and introns; remaining probes as intergenic). Genome-wide CpG dinucleotides were counted with `Biostrings`, as well as counts in and out of GRBs. Densities

were calculated per base pair for promoters, gene bodies and intergenic regions. The same counting procedure was applied to significantly DM CpGs, defined as probes with $p < 0.05$ and direction calculated using limma. For permutation-based significance testing, a set of 1,000 shuffled GRB ranges were used that were generated by preserving GRB lengths, disallowing overlap among shuffled regions and allowed up to 10% of region length to overlap with observed GRBs, using a fixed random seed to ensure reproducibility. For each permutation, we recomputed overlaps and class-specific counts (total, significant, hypermethylated and hypomethylated). Two-sided empP were then calculated using Equation (3.2), where r is the number of permutation observations more extreme than the observed value and i is the total number of permutations. Observed ratios (for example, the proportion of overlaps that are significantly hypermethylated) were evaluated against the corresponding permutation ratios using the same formula. CpG count and DM ratio differences between target and bystander genes were assessed using the Wilcoxon rank-sum test. For DM ratio comparisons, effect size was quantified using Cliff's δ , as the Wilcoxon test alone does not indicate the magnitude of the observed differences. Pearson's correlation analysis was performed using 392 samples from the TCGA cohort with both DNAm and expression data provided.

3.3 Methods for the Parkinson's disease study

3.3.1 Participant selection and PBMC isolation

The study included 17 individuals with Parkinson's disease (PD) and 10 neurologically healthy control participants, aged 48–85 years. Patients with PD and healthy controls were recruited at the University Hospital Centre Zagreb. PD diagnosis was established according to the United Kingdom Brain Bank diagnostic criteria. Peripheral blood was collected into EDTA tubes from all participants, and one PD participant was sampled twice, one week apart, to generate a biological replicate.

The study was conducted in accordance with the Declaration of Helsinki and received approval from the Ethics Committee of the University Hospital Centre Zagreb, the University of Zagreb School of Medicine and the University Hospital for Infectious Diseases “Dr. Fran

Mihaljević". Each participant provided written informed consent before being included in the study.

Peripheral blood mononuclear cells (PBMCs) were isolated within 2–3 h of blood collection by Ficoll-Paque density gradient centrifugation following the manufacturer's protocol. Isolated PBMCs were used for magnetic enrichment of CD4⁺ T-cells and dendritic cells. Aliquots were cryopreserved for flow cytometry validation, where sufficient remaining material was available.

CD4⁺ T-cells were enriched by positive magnetic selection using CD4 MicroBeads (Miltenyi Biotec), whereas dendritic cells were enriched from the remaining PBMC fraction by negative selection using the Human Pan-DC Enrichment Kit (Miltenyi Biotec). Cell numbers were determined using a haemocytometer, adjusted to equal concentrations, and combined in equal proportions before scRNA-seq library preparation.

3.3.2 Single-cell RNA sequencing

Gene expression libraries were prepared using the Chromium Next GEM Single Cell 5' Reagent Kit v2 (10x Genomics) following the manufacturer's protocol. An initial target recovery of 10,000 cells per sample was performed, followed by a reduction to 5,000 cells in subsequent batches, after ensuring this number also achieved the desired sequencing depth. Libraries were quality-controlled using an Agilent Bioanalyzer 2100, pooled, and sequenced on an Illumina NextSeq 550 platform using High Output 150-cycle chemistry, targeting approximately 20,000 reads per cell.

3.3.3 Alignment and quality control of sequencing data

Sequencing data were processed using the standard Cell Ranger v7.2 (10x Genomics) pipeline. FASTQ files were first demultiplexed and concatenated for each sample. The count function was then used to perform read alignment, gene expression quantification, and cell calling. Alignment and feature counting were carried out against the prebuilt human reference genome (GRCh38, 2020) downloaded from the 10x Genomics website. During cell calling, droplets without cells were identified and removed from the raw count matrix. Quality control metrics generated during this step are summarised in Supplementary Data 1.

3.3.4 Single-cell data pre-processing

Most downstream analyses were performed in R v4.2 using the Seurat v5.0 package within a Singularity container to ensure computational reproducibility. The container image is publicly available on Docker Hub (<https://hub.docker.com/r/prskalok/seurat>).

Quality control consisted of several filtering steps: (i) removal of cells containing more than 5% mitochondrial genes, (ii) exclusion of cells expressing fewer than 200 genes or exceeding an upper threshold defined as three median absolute deviations from the sample median, and (iii) removal of cells containing more than 20,000 unique molecular identifiers (UMIs). Gene expression data were normalised using `LogNormalize`, after which the 2,000 most variable features were identified using the `FindVariableFeatures` function with the `vst` selection method. The data were then scaled using `ScaleData` and PCA was performed using `RunPCA` for downstream analyses. Cell embeddings were visualised using `RunTSNE`. Doublet identification was performed using `DoubletFinder` v2.0.4. A parameter sweep was performed, and for most samples, the parameters identified as optimal by the package's algorithm were used. For other samples, upon inspecting the results of the parameter sweep (data not shown), the neighbourhood size parameter (`pK`) was selected according to the protocol recommended by the developers of the package [137]. Sample integration was performed using reciprocal PCA, followed by shared nearest neighbour (SNN) clustering.

3.3.5 Cell type annotation

Cell type annotation was performed using a reference-based approach with a PBMC reference dataset [138]. This method uses anchors to harmonise reference and query datasets, followed by projection of reference labels onto the query dataset. Cell type labels in the reference dataset are organised into two annotation levels: major cell types (e.g. DCs, CD4⁺ T-cells, and CD8⁺ T-cells) and cell subtypes (e.g. pDCs, cDC1s, and cDC2s). Final cell populations were defined by combining SNN clustering with reference-based annotation.

Trajectory inference and pseudotime analyses were performed using `Monocle` v2.3 on CD4⁺ T-cells and DCs [139]. Owing to the large number of annotated CD4⁺ T-cells, a random subset of

5,000 cells was selected for trajectory inference, whereas all DCs were included in the analysis. The 500 most highly variable genes were used as the ordering filter. Cell ordering was performed using the `orderCells` function.

3.3.6 Differential expression and pathway enrichment analysis

Differential expression analysis was performed using the `FindMarkers` function from the Seurat package with the Wilcoxon rank-sum test. Raw p-values were adjusted for multiple testing using the Bonferroni correction and are hereafter reported as adjusted p-values (p_{adj}). Differentially expressed genes were defined using a significance threshold of $p_{\text{adj}} < 0.05$ and an absolute $\log_2\text{FC}$ greater than 0.5.

Pathway enrichment analysis was performed in R v4.2 using the `gProfiler2` v0.2 package. Annotated human genes were used as the background set for the hypergeometric test, and multiple testing correction was performed using the Set Counts and Sizes method. Visualisations were generated using the `ggplot2` v3.4 and `G0plot` v1.0 packages. Databases used for querying: GO:BP [131], KEGG [132], and Reactome [133].

3.3.7 Validation analyses

A pseudobulk approach was implemented by aggregating gene counts for each donor and cell subtype, generating a gene-by-sample count matrix equivalent to a bulk RNA-seq dataset. Differential expression analysis was then performed using the standard DESeq2 pipeline [140]. $\log_2\text{FC}$ values from the DESeq2 analysis were extracted for genes identified as significantly differentially expressed using the single-cell `FindMarkers` analysis.

External validation was performed by comparing the top-expressed genes with bulk RNA-seq data obtained from the BLUEPRINT project (<https://projects.ensembl.org/blueprint/>) and a reference PBMC single-cell dataset [141]. Six CD4⁺ alpha-beta T-cell samples from venous blood and three conventional dendritic cell samples from cord blood were selected from the BLUEPRINT project, and the mean TPM values were calculated across samples. For the reference PBMC dataset and the generated single-cell dataset, gene expression was calculated as the mean normalised expression across annotated CD4⁺ T-cells and DCs. Validation was

performed by comparing the top 500 expressed genes using Spearman's correlation.

3.3.8 Exome sequence processing

The variant calling format files were generated using GATK variant caller and were obtained for 61 PD patients and 16 control (CTRL) participants. The EXON field and functional annotation field provided from the variant calling format files were used to select only non-coding variants and filter out coding and splicing variants. Next, the non-coding variants were overlapped with GWAS identified loci associated with a set of manually curated PD-related traits: age of onset of Parkinson disease, cognitive decline measurement, cognitive domain measurement, cognitive function measurement, cognitive impairment, cognitive impairment measurement, cognitive inhibition measurement, dementia, frontotemporal dementia, insomnia, insomnia measurement, Lewy body dementia, neuroticism measurement, Parkinson disease, Parkinson's disease symptom measurement [142–144]. For each unique variant, we investigated over-representation in PD patients versus CTRL using one-sided Fisher's exact test based on variant carrier counts. Additionally, for each PD-related trait, the proportion of loci located within GRBs was compared with an expected proportion of 33.33% using a one-sided binomial test. The baseline expectation roughly corresponds to the genomic coverage of GRBs. BH procedure was used in both cases to correct for multiple testing.

Code that was used for these analysis is available on Github repository (https://github.com/PKatarina/PD_SingleCell)

4 Results

4.1 Results of the targPred study

The central contribution of this work is development of a targPred web-based publicly available resource, revealing transcriptomic evidence of long-range regulatory effects for all genomic loci harbouring common human non-coding variants.

We applied targPred to predict candidate target genes for 280,711 variants associated with 11,921 phenotypic traits from the GWAS Catalog, comprising 757,253 autosomal and X-chromosomal SNP–trait associations. These predictions were further supported using a statistical framework for enhancer–gene associations. Hereafter, a GWAS locus is defined as the genomic region encompassing variants that are in LD with the lead GWAS SNP. For loci residing within GRBs, targPred assigns the GRB target gene as a plausible alternative to the GWAS-proposed target gene in 38% of cases (Figure 4.1a, middle ring). In addition, the resource identifies novel candidate target genes, distinct from both the reported and mapped GWAS genes, for 158,611 locus–trait pairs, corresponding to 56% of all locus–trait pairs located within GRBs (Figure 4.1a, highlighted in yellow).

4.1.1 General properties of targPred

For a subset of GRB loci where the locus overlaps a known enhancer (67% of loci; Figure 4.1a, inner ring), we additionally provide an experimentally supported measure of the association between eRNA activity and gene expression across a broad collection of biosamples, including cell lines, primary cells, and tissues. Genome-wide enhancer–gene empP were calculated using CAGE data from the FANTOM5 project [22]. These empP values were derived from a

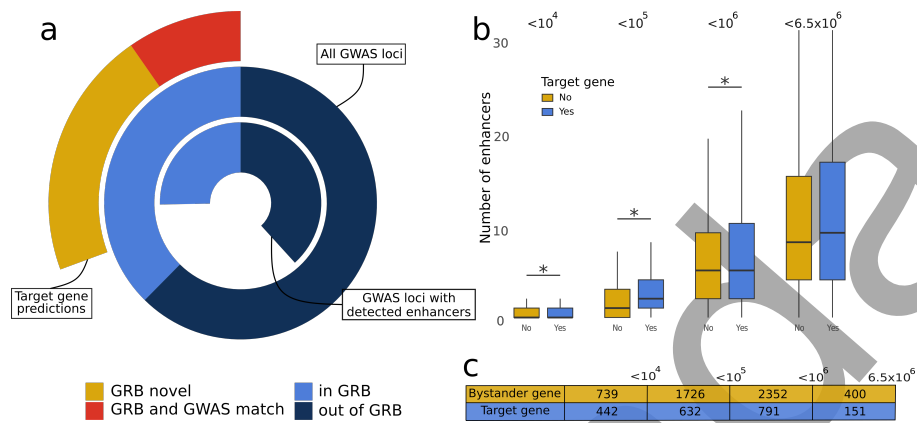


Figure 4.1. a Middle ring: all GWAS loci. Inner ring: loci overlapping enhancers. Outer ring: SNP–gene assignments predicted by targPred, where red indicates loci that targPred assigns the same target gene as the GWAS Catalog, and yellow indicates loci where targPred predicts a novel target gene. b Cumulative number of enhancers significantly associated with bystander and target genes (BH-adjusted $empP < 0.05$), grouped by genomic distance. c Number of bystander and target genes with significant enhancer associations within each genomic distance bin. Adapted from Mandić et al. [73].

permutation test of enhancer and promoter CAGE signals across 309 biosamples, capturing both activating (increased gene expression) and repressive (decreased gene expression) regulatory effects (see Methods for details) and can be obtained by querying the targPred resource for individual SNPs or genomic loci. Significant $empP$ values (BH-adjusted $empP < 0.05$) indicate enhancer–promoter activity consistent with a functional regulatory interaction.

On average, GRB target genes were associated with a greater number of locus-overlapping enhancers than bystander genes across multiple genomic distance bins (Figure 4.1b). This difference was statistically significant for the $< 10^4$, $< 10^5$, and $< 10^6$ distance bins ($p < 0.05$), suggesting that GRB target genes are embedded within more complex regulatory landscapes spanning larger genomic distances (Table 4.1).

Table 4.1. Median number of associated enhancers for GRB target and bystander genes across genomic distance intervals, with corresponding Wilcoxon rank-sum p-values.

Genomic distance	Target median	Bystander median	p-value
1– 10^3	1	1	0.898
10^3 – 10^4	1	1	0.0467
10^4 – 10^5	3	2	3.18×10^{-11}
10^5 – 10^6	6	6	0.0103
$\geq 10^6$	10	9	0.135

4.1.2 Traits under long-range regulation

We next assessed long-range regulation across traits represented in the GWAS, focusing specifically on loci harbouring enhancers. Trait-associated enhancer–gene pairs ($empP < 0.001$) located within GRBs were compared with those outside GRBs, and differences in genomic distance were evaluated using the Wilcoxon rank-sum test. In total, 1607 disease traits contained at least one GRB locus harbouring an enhancer–gene pair. Among these, 779 traits exhibited significantly greater enhancer–gene distances within GRBs than outside GRBs (BH-adjusted $p < 0.05$). The 30 most significant disease traits, ranked by p-value and subsequently ordered according to the median distance difference, were selected for illustration (Figure 4.2).

We identified six broad disease categories: cancer (EFO term “cancer”, MONDO:0004992), immune disorders (EFO term “immune system disorder”, MONDO:0005046), asthma (EFO term “asthma”, MONDO:0004979), metabolic diseases (EFO term “metabolic disease”, MONDO:0005066), psychiatric disorders (EFO term “psychiatric disorder”, MONDO:0002025), and an “other” category. Using this classification, *Clostridium difficile* infection and oral lichen planus were assigned to the “other” category. Nevertheless, both traits are associated with immune- and cancer-related processes, as *Clostridium difficile* infection is strongly linked to inflammatory bowel disease [145], while oral lichen planus is a T-cell-mediated inflammatory disease associated with both autoimmune disease and oral cancer [146].

Among the significant traits, we also observed a distinct group of psychiatric disorders exhibiting long-range regulation (Figure 4.2b), including substance-related, cognitive, and mental disorders. Furthermore, genome permutation testing demonstrated an enrichment of GWAS loci within GRBs, indicating that loci associated with several psychiatric disorders occur within GRBs more frequently than expected by chance (Figure 4.2c). This finding is consistent with previous observations for schizophrenia and autism spectrum disorder [70]. Similar results were obtained using independent GWAS datasets for these traits (Figure 4.2c). In addition, we observed enrichment of GWAS loci from a major depressive disorder study and from a study combining eight psychiatric disorders (Figure 4.2c).

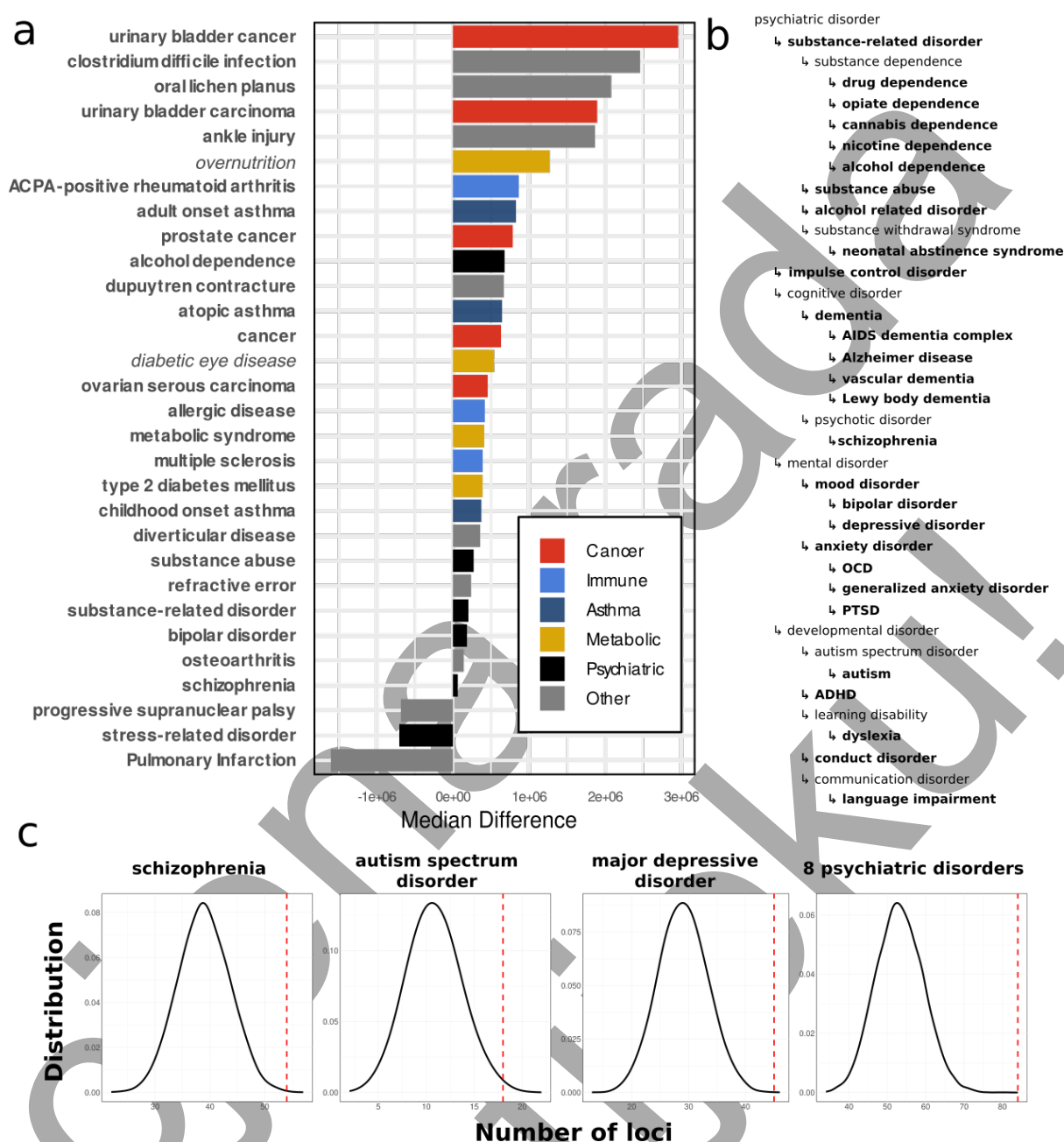


Figure 4.2. Traits under long-range regulation. a Top 30 disease traits exhibiting long-range regulation within GRBs, selected based on the BH-adjusted p-value and ordered according to the median enhancer–gene distance within versus outside GRBs. The classification of italicized traits into the metabolic category is described in the Methods. b Overview of the psychiatric disorder ontology highlighting (bold) traits exhibiting significant long-range regulation within GRBs compared with outside GRBs. Abbreviations: OCD, obsessive-compulsive disorder; PTSD, post-traumatic stress disorder; ADHD, attention deficit/hyperactivity disorder. c Probability distributions of loci overlapping 10,000 permuted GRB regions for four representative psychiatric disorders. Empirical p-values were calculated based on the observed number of loci overlapping GRBs. All four traits showed significant enrichment within GRBs ($p < 0.05$, Table 4.3). Adapted from Mandić et al. [73].

4.1.3 Comparison between targPred and validated SNP–gene links

To evaluate the accuracy of targPred predictions, we used two ground-truth datasets of validated SNP–gene associations. First, disease-associated coding variants deposited in the ClinVar database were used, as coding variants generally provide precise and unambiguous gene assignments. Second, we used a collection of experimentally validated SNP–gene associations compiled by Alsheikh et al. [128]. It should be noted that both datasets are relatively limited in size and therefore do not fully capture the complexity of disease phenotypes.

To validate targPred predictions against the ClinVar dataset, we searched for genes that harboured both a coding variant reported in ClinVar and a nearby non-coding variant predicted by targPred or the GWAS Catalog. The two variants had to be associated with the same disease trait, while the non-coding variant was required to reside within the same GRB as the coding variant. This approach identified 432 variants (Figure 4.3a, Supplementary Data 2). targPred provided target gene assignments for 201 of these variants, of which 102 were not reported in the GWAS Catalog. As an example, two variants (rs55754224 and rs6829664) associated with atrial fibrillation are located within an intron of *CAMK2D*, a gene assigned as the target gene in the GWAS Catalog (Supplementary Figure S3)[147, 148]. In contrast, targPred predicts *ANK2* as the target gene. *ANK2* is associated with several cardiovascular phenotypes, including atrial fibrillation, long QT syndrome, and cardiac arrhythmia, according to ClinVar. This prediction therefore identifies *ANK2* as a promising candidate target gene for these non-coding variants in atrial fibrillation, as well as in two additional cardiovascular traits.

The second validation dataset consisted of manually curated publications providing experimental evidence for regulatory variant–target gene associations [128]. For the subset of targPred associations included in the dataset of Alsheikh et al., most experimentally validated SNP–gene links were supported by the same gene assignment in both the GRB and GWAS based target gene lists (Figure 4.3b, Supplementary Data 2), while a smaller number of associations were unique to targPred. For example, variant rs1859962, associated with prostate carcinoma, is assigned to the long non-coding RNA gene *CASC17* in the GWAS Catalog [149, 150]. In contrast, both targPred and the Alsheikh dataset assign this variant to *SOX9*, located more than 1

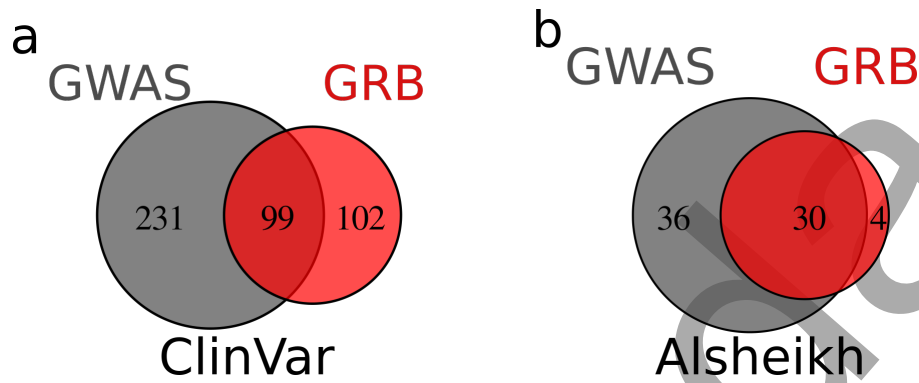


Figure 4.3. Comparison of GRB and GWAS based target gene assignments with experimentally validated SNP-gene associations from ClinVar [126] (a) and Alsheikh et al. [128] (b). A prediction was considered a true overlap if at least one of the predicted target genes matched a target gene in the corresponding validation dataset. Adapted from Mandić et al. [73].

megabase away.

4.1.4 Comparison between targPred and computational SNP-gene prediction methods

Finally, we compared targPred enhancer-gene assignments with enhancer-gene predictions from ABC and GeneHancer, as well as SNP-gene scores from the cS2G model [62, 64, 66]. All comparisons with other prediction methods were restricted to enhancers located within GRB regions, and were classified into three categories. Category 0 comprised pairs that neither represented a significant enhancer-gene association in targPred nor included a GRB target gene. Category 1 included pairs satisfying either of these criteria, whereas category 2 comprised pairs that satisfied both.

For comparison with the ABC dataset, enhancer-gene pairs with scores > 0.015 were selected, following the threshold used previously (Figure 4.4b)[63]. For GeneHancer, pairs with scores > 250 were selected owing to the discontinuous score distribution in the original dataset (Figure 4.4a and c). Across all three prediction methods, enhancer-gene pairs belonging to category 2 received significantly higher scores than pairs in the other categories (Table 4.2). Likewise, ABC, GeneHancer, and cS2G all assigned the highest scores to category 2 compared with categories 1 and 0, although this increase reached statistical significance only for ABC and cS2G (Table 4.2). Despite this overall agreement, targPred identified numerous significant long-

range regulation enhancer–gene pairs that were not detected by the other prediction methods Supplementary Data 3. A representative example involving the *BDNF* target gene is presented below.

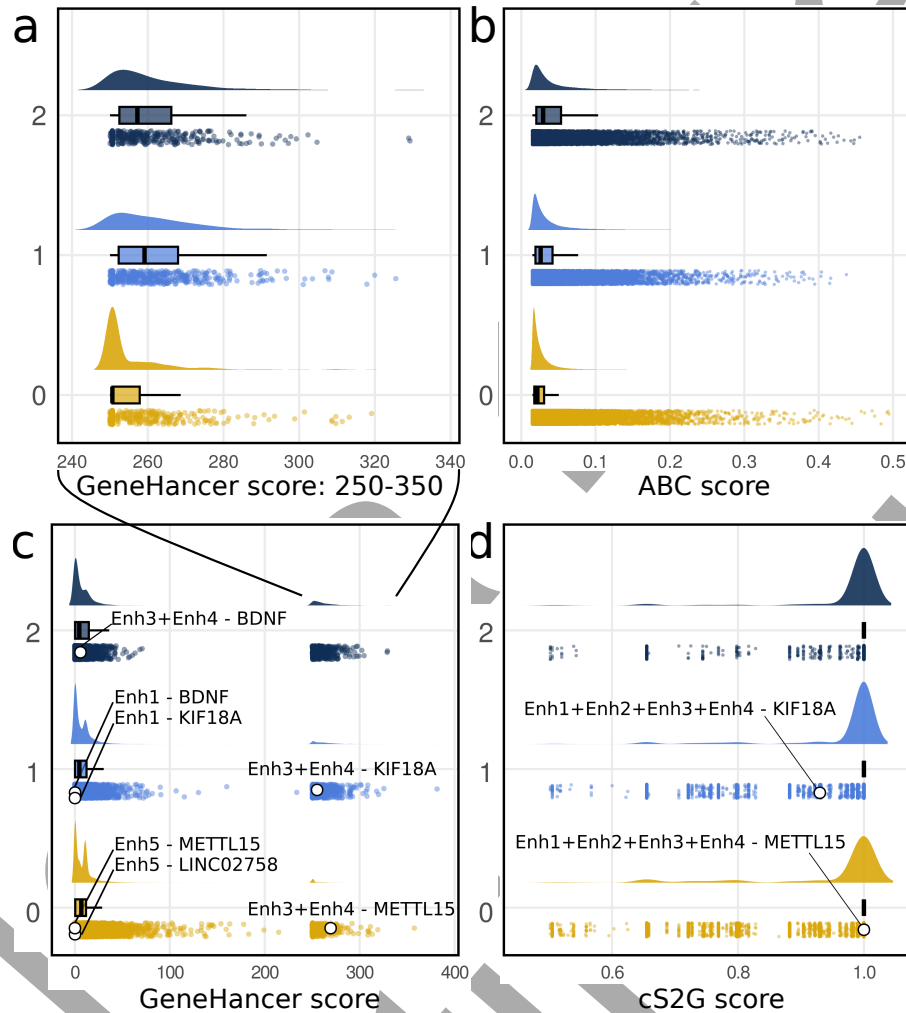


Figure 4.4. Comparison of GRB annotation with three SNP–gene prediction methods: GeneHancer (a, c), ABC (b), and cS2G (d) [62, 64, 66]. Panel c shows the distribution of GeneHancer scores across the full score range, whereas panel a presents only high-confidence enhancer–gene pairs with GeneHancer scores > 250. White points in panels c and d indicate the scores of enhancers associated with the childhood obesity loci (rs7940756 and rs10835310). No ABC enhancer overlapped either of these loci. Adapted from Mandić et al. [73].

4.1.5 Long-range regulation in childhood obesity

To illustrate locus–gene assignment by targPred, we present two childhood obesity-associated loci supported by chromatin contact data. We previously showed that SNPs associated with obesity and other metabolic syndrome-related traits frequently participate in long-range regulatory

Table 4.2. Comparison of prediction scores among targPred categories for GeneHancer, ABC, and cS2G.

Method	Cat. comparison	p-value	HL shift estimate	Median	<i>n</i> pairs
GeneHancer	0 vs. 1	3.72×10^{-31}	—	250.65	1,036
	0 vs. 2	7.03×10^{-26}	—	259.20	911
	1 vs. 2	0.211	—	257.26	835
ABC	0 vs. 1	$< 2.2 \times 10^{-308}$	—	0.021	124,212
	0 vs. 2	$< 2.2 \times 10^{-308}$	—	0.026	106,900
	1 vs. 2	2.22×10^{-70}	—	0.029	41,128
cS2G	0 vs. 1	1.23×10^{-35}	-2.78×10^{-5}	1	9,978
	0 vs. 2	3.68×10^{-44}	-8.67×10^{-7}	1	7,778
	1 vs. 2	3.46×10^{-11}	-6.76×10^{-6}	1	2,407

Cat., targPred category; HL shift, Hodges–Lehmann estimate of the location shift between the first and second categories in each comparison; *n* pairs, number of enhancer–gene pairs.

associations (Figure 4.2a) and we identified four such traits that are enriched within GRBs (Figure 4.5 and table 4.3).

Figure 4.6a shows two childhood obesity-associated loci (rs7940756 and rs10835310) overlapping a GRB (chr11:27,604,430–29,063,089). For both loci, the nearest genes, *METTL15* and *LINC02758*, were reported as the target genes in the original GWAS study [151]. In contrast, targPred identifies *BDNF* as a more plausible target gene, supported by multiple lines of functional evidence. *BDNF* has previously been implicated in adult obesity, type 2 diabetes mellitus, and other metabolic syndrome-related traits, whereas *METTL15* encodes a methyltransferase protein and *LINC02758* is a long non-coding RNA with no clearly established function.

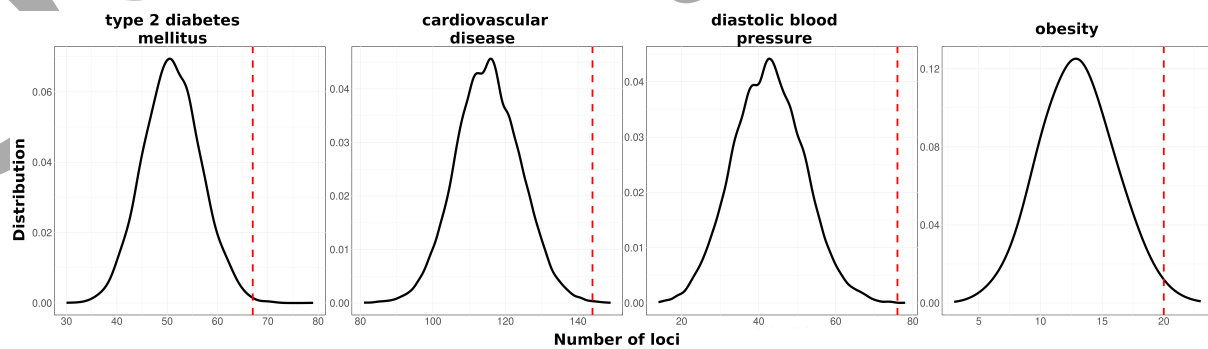


Figure 4.5. Probability distribution of loci overlapping N = 10,000 permuted GRB regions was estimated for four traits associated with the metabolic syndrome. P-values were calculated empirically (Table 4.3). Adapted from Mandić et al. [73].

Table 4.3. Enrichment of GWAS loci within GRBs for selected disease traits.

Trait	Study acc.	Ontology ID	<i>n</i> SNPs	p-value
Schizophrenia	GCST006803	MONDO:0005090	98	0.0006
Autism spectrum disorder	GCST007556	EFO:0003756	31	0.0033
Major depressive disorder	GCST012090	MONDO:0002009	73	1.00×10^{-4}
Eight psychiatric disorders	GCST009600	EFO:0004895	134	0
Type 2 diabetes mellitus	GCST006867	MONDO:0005148	138	0.0020
Cardiovascular disease	GCST007072	EFO:0000319	316	0.0008
Diastolic blood pressure	GCST007704	EFO:0006336	118	1.00×10^{-4}
Obesity	GCST009382	EFO:0001073	35	0.0065

Study acc., GWAS Catalog study accession; *n* SNPs, number of SNPs.

The *METTL15* locus spans the entire *METTL15* and *KIF18A* genomic regions, together with four enhancers (Enh1-4; Figure 4.6a). Based on available transcriptional activity profiles, both *BDNF* and *KIF18A* respond to these enhancer elements, whereas *METTL15* exhibits the expression profile of a ubiquitously expressed gene that is largely unresponsive to enhancer activity (Figure 4.6b). Furthermore, chromatin contact maps demonstrate that the region containing these four enhancers interacts with the promoter of *BDNF*, effectively forming a regulatory hub [129]. The *LINC02758* locus contains a single enhancer (Enh5) without nearby genes. targPred predicts that this locus regulates the distal *BDNF* gene across a genomic distance of approximately 1 megabase. Consistent with this prediction, *BDNF* showed a significant response to Enh5, while Hi-C data confirmed chromatin contacts between these two regions (arrow, Figure 4.6a). This example further illustrates how targPred can identify novel SNP–gene associations that are not detected by recently published prediction methods (Figure 4.4b-d).

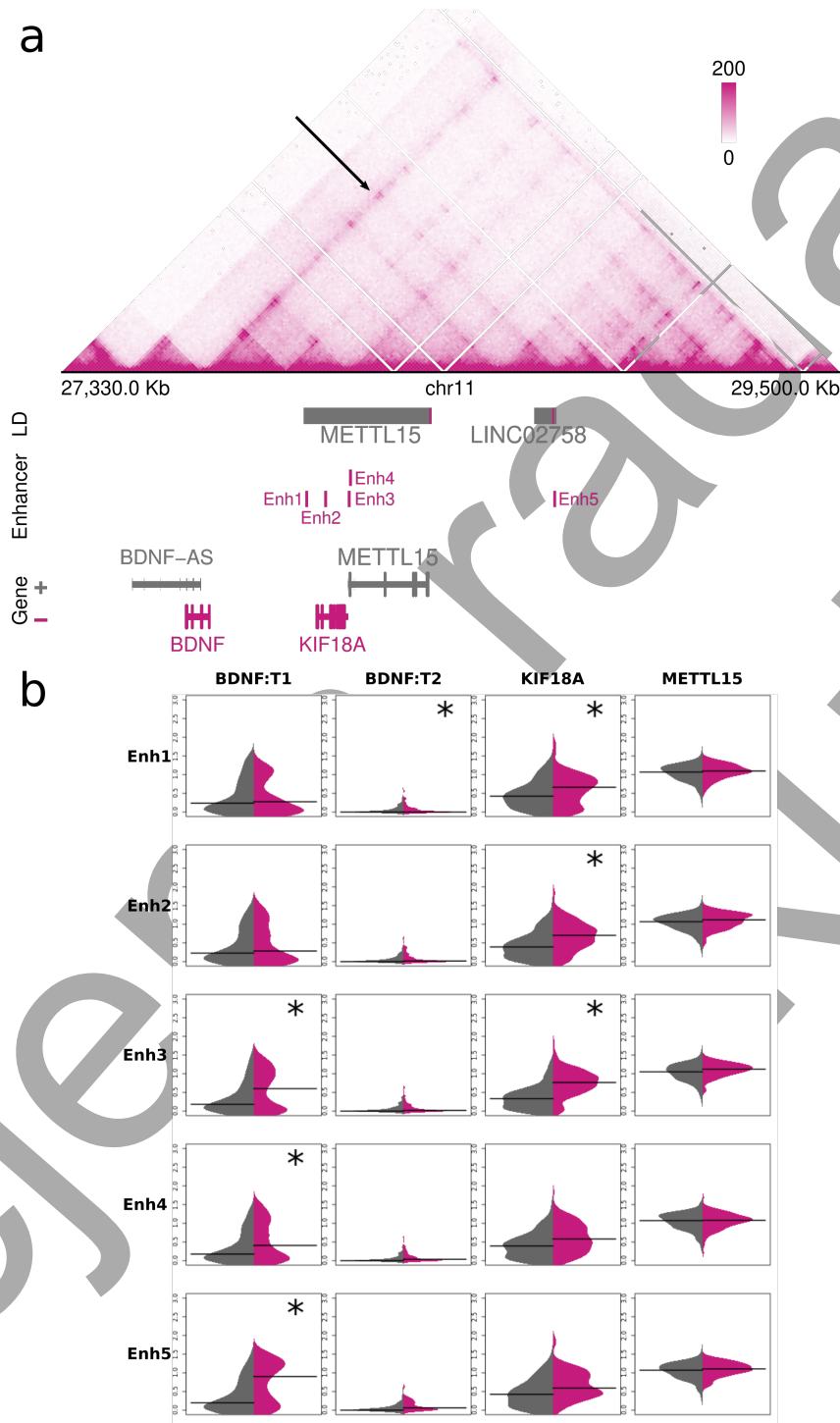


Figure 4.6. Childhood obesity-associated loci linked to BDNF. a Hi-C interaction map of the GRB containing the *BDNF* target gene, with annotated childhood obesity-associated loci. A long-range chromatin interaction is indicated by the arrow. The two loci are highlighted in grey, while overlapping SNPs are shown in pink in the LD track and labelled according to the GWAS target gene assignment for each locus. b Two-sided distribution matrix of log-TPM expression values for all enhancer-gene pairs within this GRB. Gene expression distributions in tissues where the enhancer is inactive and active are shown in grey and pink, respectively, with median values indicated by black lines. Significant enhancer-gene associations (Bonferroni adjusted $empP < 0.05$) are marked with an asterisk (*). Adapted from Mandić et al. [73].

4.2 Results of the head and neck squamous cell carcinoma study

The first disease of interest due to its complex aetiological links to epigenetic regulation is the HNSCC. It has a poor overall survival rate and mounting evidence suggests that epigenetic changes, rather than genomic changes alone, are major drivers of HNSCC development and progression. Despite originating from squamous cell epithelium, HNSCC is highly heterogeneous, reflecting differences in anatomical site, exposure to tobacco and alcohol, infection with HPV, and broader lifestyle factors. In this thesis an integrative approach will be utilised, combining transcriptome, miRNome and methylome data to characterise epigenetic programmes in HNSCC and to relate them to long-range regulatory mechanisms.

This thesis presents a detailed integrative analysis of a Croatian HNSCC cohort and a complementary TCGA-HNSCC cohort. The Croatian dataset provides a well-defined, lifestyle-specific patient group aimed for in depth investigation of multi-omics patterns, whereas TCGA-HNSCC serves as a broader, heterogeneous reference set. In later sections, TCGA-HNSCC is also used as an independent resource to explore GRBs, their target and bystander genes, and long-range epigenetic regulation in HNSCC, thereby linking disease-associated epigenetic changes to the GRB framework developed in the rest of the thesis.

4.2.1 Croatian HNSCC Cohort and Multi-Omics Profiling

We investigated a Croatian cohort of 61 primary HNSCC tumours of oral and oropharyngeal origin collected at Clinical Hospital Dubrava between 2015 and 2018, representing patients with high smoking and alcohol intake. For comparison, non-malignant tissues were obtained from individuals without HNSCC: fresh tonsil specimens from patients undergoing tonsillectomy and healthy oral mucosa swabs. These controls were hence independent control samples, rather than patient-matched pairs. Multi-omics profiling was performed on overlapping subsets of this cohort (Figure 4.7). For the miRNome, we analysed 22 HNSCC tumours together with three non-malignant tonsil controls. Genome-wide DNAm array was assessed in 12 tumours from

this set plus one additional tumour, alongside eight healthy oral mucosa swabs as methylation controls. Transcriptome profiling was carried out for 10 HNSCC tumours (six oral and four oropharyngeal) and three tonsil controls drawn from the miRNome dataset. The integrative analyses in this thesis focus on this overlapping multi-omics subset, providing a coherent view of miRNA, DNAm and gene expression changes in Croatian HNSCC relative to non-malignant oral and tonsillar tissues.

4.2.2 Global Omics Patterns in Croatian HNSCC

Across the Croatian multi-omics subset, global variation in gene expression, miRNA profiles and DNAm was dominated by differences between tumour and non-malignant samples. PCA of the transcriptome revealed a clear separation of HNSCC tumours from tonsil controls along the first two components (Figure 4.8), explaining more than half of the variance in the dataset. Additional eight components reinforce this tumour–control distinction amounting to a total of 84.4% variance explained (Supplementary Figure S1). Comparable patterns were observed for the miRNome and methylome, where sample-distance heatmaps showed clustering primarily

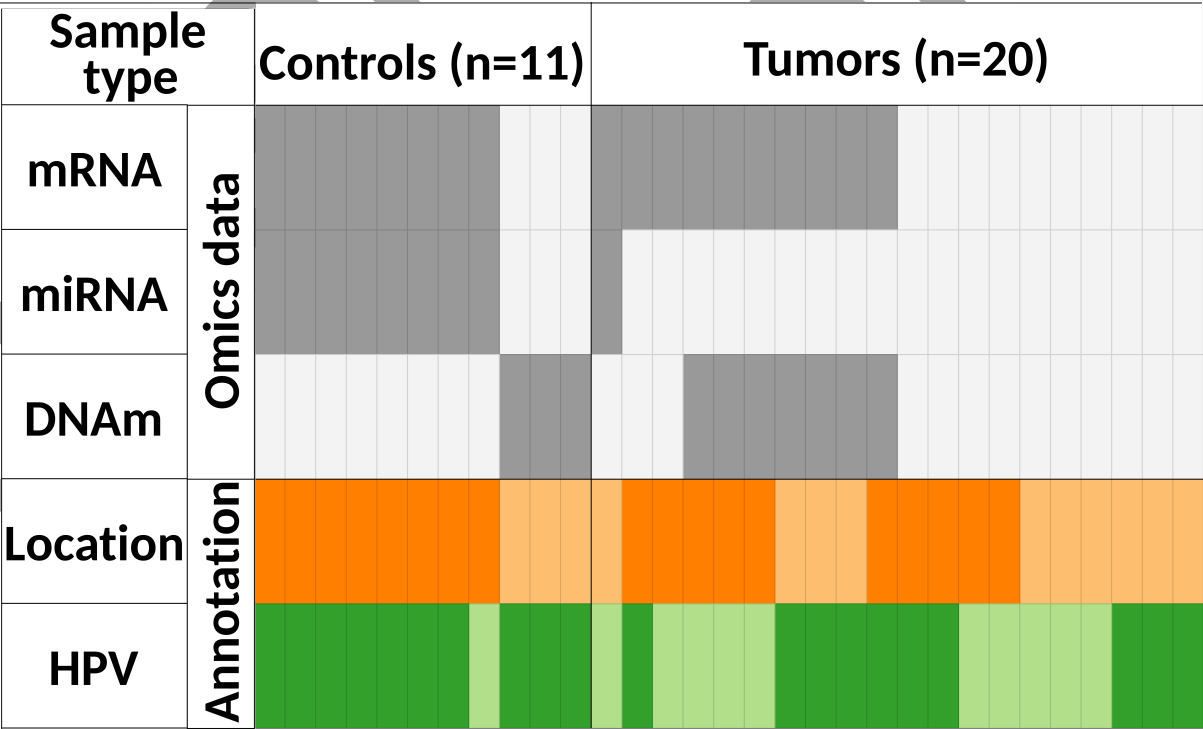


Figure 4.7. Croatian HNSCC cohort. Sequenced samples are shown in white fields; O, oral; OP, oropharyngeal; P, HPV-positive; N, HPV-negative. Adapted from Mandić et al. [103].

by tumour status rather than by anatomical site (Figure 4.8 b and c). Additionally, the patient survival rates did not show statistically significant differences in HPV positive versus negative nor tumour location (Supplementary Figure S2). In line with these global patterns, neither HPV status nor oral versus oropharyngeal location introduced major structure into the leading components of variation, indicating that, at the level of whole-transcriptome, miRNA and methylation profiles, tumour development itself is the principal driver of molecular divergence in this cohort. This consistent separation between tumour and control across all omics layers provides a robust framework for subsequent differential expression, integrative epigenetic analyses and the investigation of long-range regulatory mechanisms.

4.2.3 Differential Expression and Pathway Enrichment in Tumour vs Control

Among 18,104 genes with sufficient read support, we identified 1,248 (6.9%) genes with increased expression and 1,533 (8.5%) with decreased expression in tumours compared with

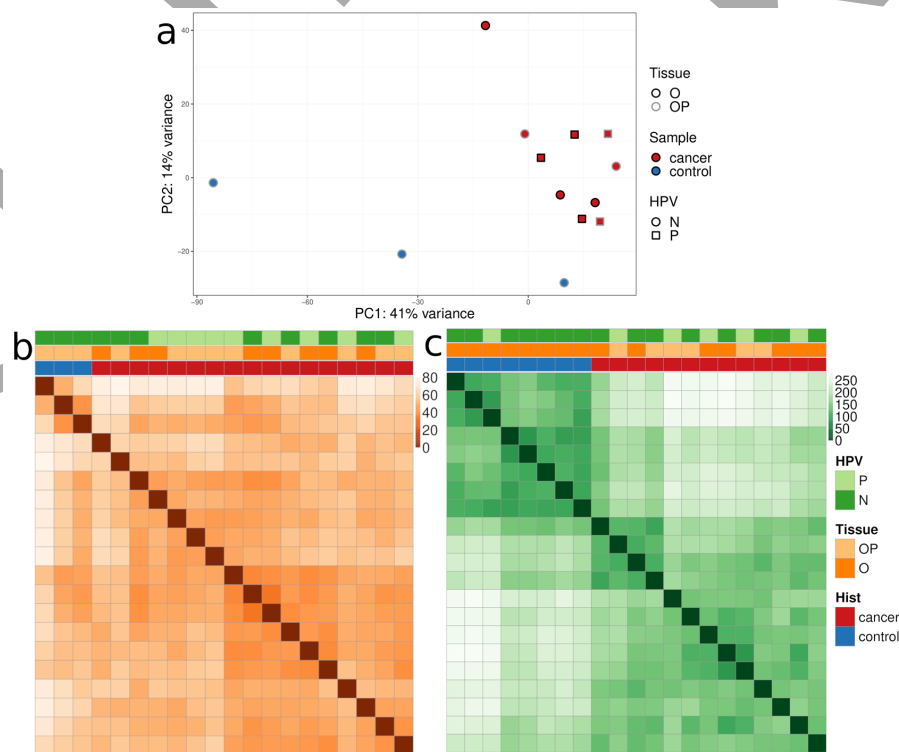


Figure 4.8. a PCA plot for transcriptome data. b sample distance heatmap for miRNA data. c sample distance heatmap for methylome. N, HPV-negative; P, HPV-positive; O, oral tumours; OP, oropharyngeal tumours. Adapted from Mandić et al. [103].

control transcriptomes (BH-adjusted $p < 0.05$). Gene set enrichment analysis of these DEGs highlighted signalling pathways linked to tumour progression, including multiple cell-cycle regulators and oncogenic pathways such as NF- κ B (Figure 4.9; Supplementary Data 4). There are 183 (16%) GRB target genes found to be deregulated, and some are members of key developmental signalling pathways such as WNT/ β -catenin pathway and FGF signalling, which are morphogen and growth-factor pathways involved in cell fate and proliferation coordination (Supplementary Data 5).

We next examined post-transcriptional and epigenetic regulation. Integration of 514 significantly deregulated miRNAs (DEmiRs) from the miRNome with experimentally supported miRNA–target interactions yielded 3,042 miRNA–gene pairs spanning 90 miRNAs and 1,227 target genes (DEmiR-DEGs). Most targets were associated with a single miRNA, but some genes central to cell proliferation and apoptosis, notably *BCL2* and *MYC*, were linked to 18 and 16 DEmiRs, respectively. Conversely, certain miRNAs showed broad targeting, for example,

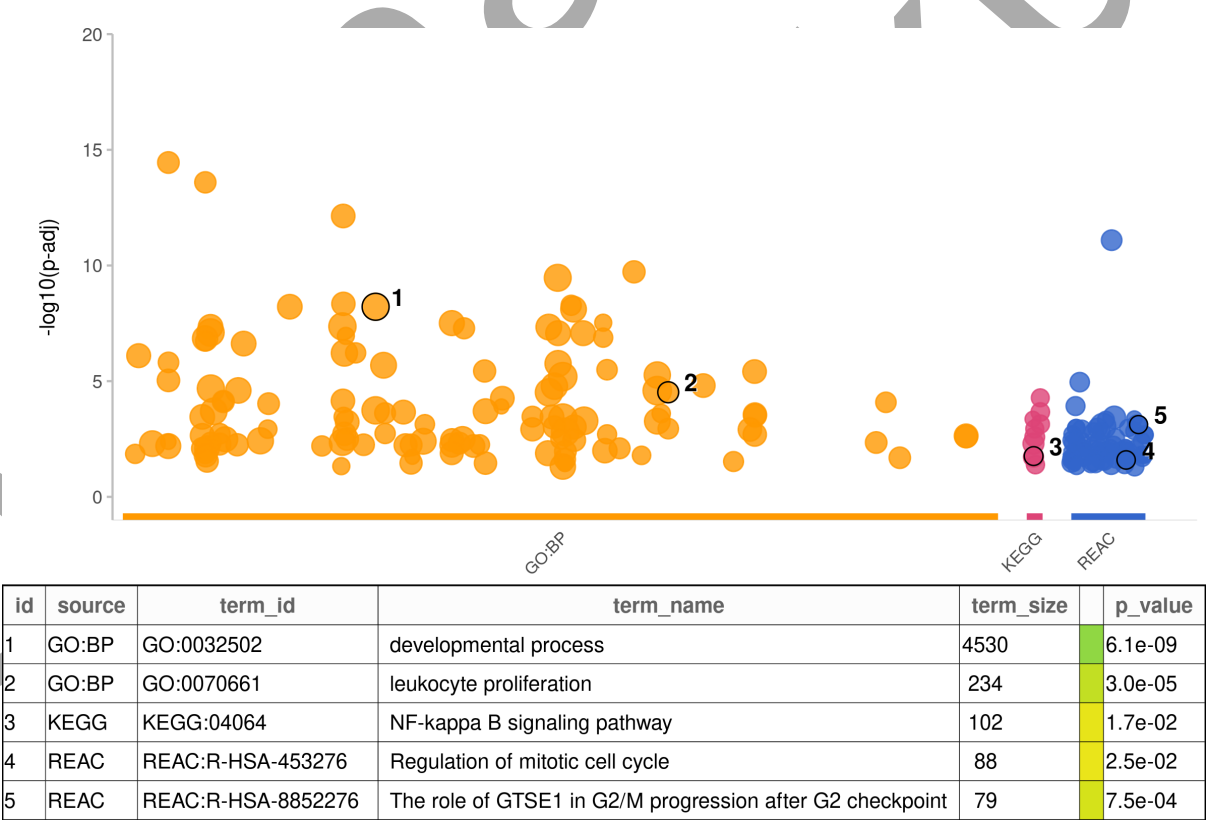


Figure 4.9. Gene enrichment analysis of DEGs in tumour versus control samples. Databases used for enrichment: GO:BP (orange), KEGG (pink), and Reactome (REAC, blue). The highlighted terms point to key concepts associated with cancer progression and regulation. Adapted from Mandić et al. [103].

the tumour-suppressive hsa-miR-26b-5p was predicted to interact with 255 distinct DEGs and displayed strong negative correlations with known oncogenes such as *MMP10*, *IGSF3* and *ARNTL2* in our cohort.

Global DNAm patterns revealed extensive epigenetic reprogramming, with more than 300,000 CpG probes hypomethylated and approximately 72,000 hypermethylated in tumours relative to controls (BH-adjusted $p < 0.05$, Supplementary Data 6). DM probes were predominantly located in gene bodies and, to a lesser extent, intergenic regions enriched for repetitive elements, the latter consistently showing tendency towards hypomethylation (Figure 4.10). At the CGI level, we observed 6,557 hypomethylated and 5,457 hypermethylated islands (Supplementary Data 7). To pinpoint deregulated tumour suppressors, we intersected our data with a curated set of tumour suppressors and identified 10 genes that combined reduced expression with promoter CGI hypermethylation, consistent with epigenetic silencing (Supplementary Table S2) [136]. Focusing on differentially methylated DEGs, enrichment analysis using TRANSFAC indicated significant over-representation of transcription factor targets from the *E2F*, *SP/KLF* and *AP-2* families (Supplementary Data 8). In particular, genes regulated by the *E2F* family showed very strong enrichment (SCS adjusted $p < 10^{-43}$) and were further supported by Reactome pathways related to cell-cycle control, including G1/S transition and G2/M checkpoints (SCS-adjusted $p < 10^{-4}$), reinforcing the central role of promoter methylation in modulating key cell-cycle regulators in HNSCC.

4.2.4 Integrative analysis of miRNA and DNAm in HNSCC

We stratified each DEG according to whether it was linked to a deregulated miRNA (DEmiR) and/or a differentially methylated CpG island (DM CGI), resulting in four categories: genes associated only with DEmiRs (DEmiR-DEGs), only with DM CGIs (DM-DEGs), with both regulatory features (DEmiR+DM-DEGs), or with neither (noReg-DEGs; Figure 4.11). These groups showed distinct functional enrichments across GO:BP, KEGG, miRTarBase, Reactome and TRANSFAC terms (Supplementary Data 9): DM-DEGs were primarily enriched for transcription factor targets without strong over-representation of specific biological pathways, DEmiR-DEGs were linked to signalling and immune-related pathways, noReg-DEGs were

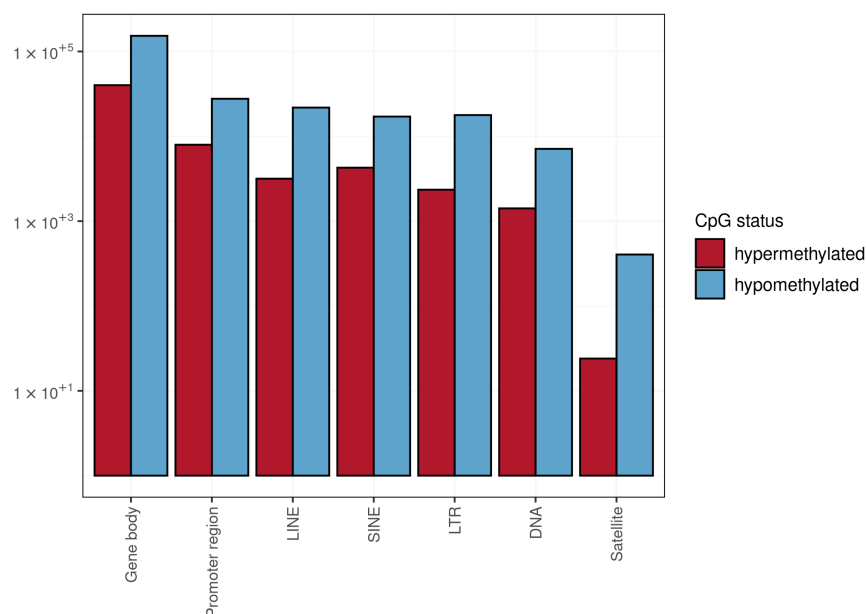


Figure 4.10. Counts of DM CpG probes in cancer versus control groups across genomic elements. Gene body – across exons and introns; Promoter region – from 500 to 2000 distance from canonical transcription start site; LINE – Long interspersed nuclear elements; SINE – Short interspersed nuclear elements; LTR – Long terminal repeat elements including retroposons; DNA – DNA repeat elements; Satellite – Satellite repeats. Annotation for genomic elements were used from Jin et al. [152]. Adapted from Mandić et al. [103].

enriched for immune and muscle system processes, and DE*miR*+DM-DEGs were associated with broad, high-level regulatory processes encompassing biological, metabolic and cellular functions, consistent with wide-ranging regulatory roles.

4.2.5 Epigenetic clusters and biological programs in HNSCC

K-means clustering of DEGs together with their corresponding DE*miR* and methylation changes revealed that, by and large, clusters were dominated either by miRNA regulation or by DNAm, with few clusters showing strong contributions from both mechanisms (Figure 4.12). This pattern indicates that, in HNSCC, gene expression tends to be driven predominantly by one of the two epigenetic modalities rather than by their combined action. In line with the conventional view that miRNAs and DNAm usually repress gene expression, several clusters (3, 4, 7 and 10) displayed negative relationships between the epigenetic feature and gene expression, whereas other clusters (1, 5, 6 and 9) showed non-canonical patterns. Because many miRNAs have multiple targets, there was substantial sharing of deregulated miRNAs across clusters, with 73%

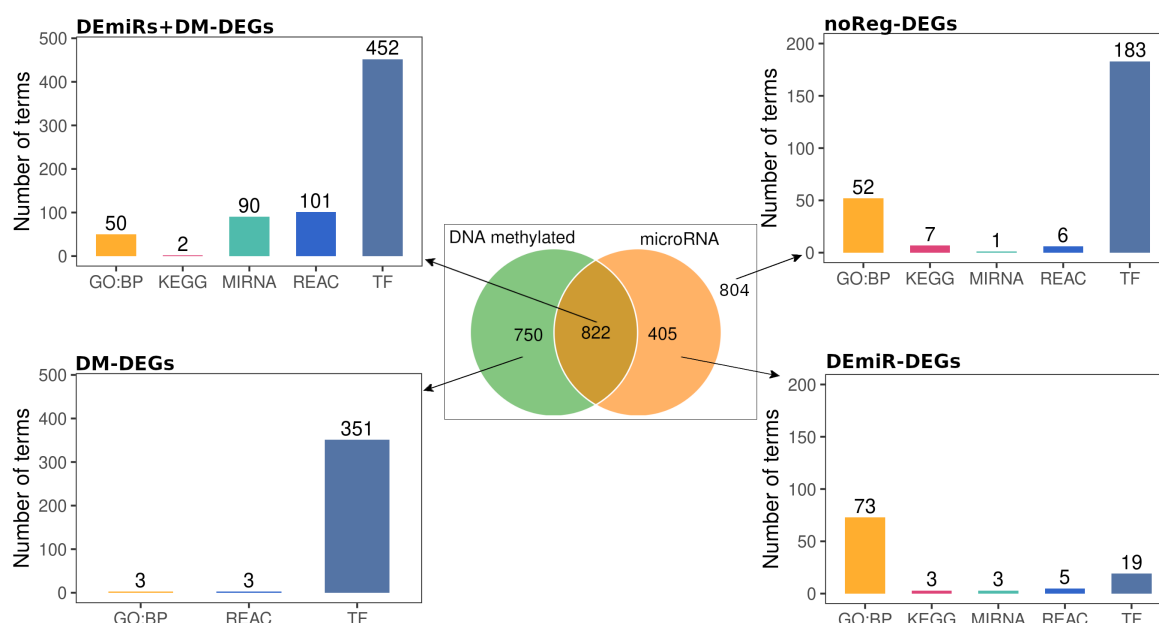


Figure 4.11. Intersection of DEGs linked with DEmiRs or DM-CGIs, in the centre. The outer panels show the number of gene set enrichment terms per database for each group of DEGs. Adapted from Mandić et al. [103].

of under-expressed miRNAs appearing in both clusters 3 and 9 and 77% of over-expressed miRNAs shared between clusters 5 and 7. Deregulated GRB target genes are distributed across multiple k-means clusters, with the larger clusters 2 and 8 containing more target genes, suggesting that target genes employ a wide range of epigenetic modules (Table 4.4).

Gene set enrichment analysis for each cluster further connected these regulatory patterns to biological programmes (Figure 4.12, Supplementary Data 10). DM-DEGs in clusters 1, 4, 6 and 10 were strongly enriched for targets of *E2F* transcription factors, even though most *E2F* genes themselves were not differentially expressed, consistent with previous reports that CpG methylation at *E2F* binding motifs modulates transcription factor binding [153]. Cluster 6 additionally comprised under-expressed genes with hypomethylated CGI and was enriched for *ZF5*-regulated genes, suggesting a potential repressive role for *ZF5* in this context [154], whereas clusters 4 and 5 were both related to cell-cycle processes but differed in their pattern of miRNA versus DNAm.

Table 4.4. Number of target genes per k-means cluster.

K-means cluster	1	2	3	4	5	6	7	8	9	10	11	Total
Target gene count	10	30	11	11	18	4	14	36	11	18	20	183

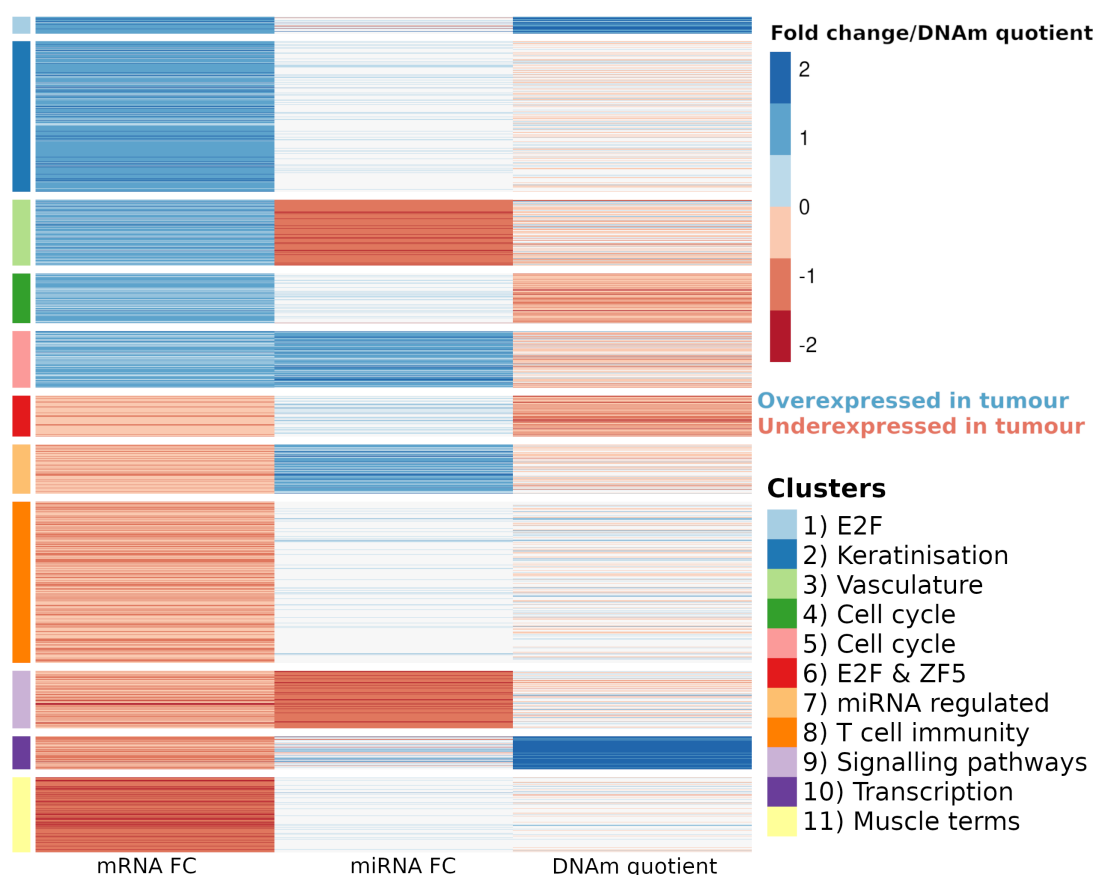


Figure 4.12. Heatmap of integrated transcriptome, miRNome and methylome data. Fold change values for transcriptome (mRNAFC column), miRNome (miRNA FC) and DNAm quotient were scaled and normalised before k-means clustering. Gene set enrichment shown for all eleven clusters. Adapted from Mandić et al. [103].

4.2.6 Validation of epigenetic clusters in TCGA-HNSCC

To test whether these regulatory patterns are specific to the Croatian cohort or can be further generalisable, we repeated the integrative analysis on TCGA-HNSCC transcriptome, methylome and miRNome data using analogous processing steps (Supplementary Data 11 and 12). FC estimates for the 1,834 overlapping DEGs showed a strong concordance between our cohort and TCGA (Pearson's $r = 0.76$, $p < 2.2 \times 10^{-16}$; Figure 4.13 a), and the TCGA dataset also reproduced the tendency for DEGs to be regulated predominantly either by DNAm or by miRNAs (Figure 4.13 b), supporting the robustness of the identified epigenetic programmes. Having established that these epigenetic clusters and transcriptional changes are reproducible across cohorts, we next asked whether they intersect with GRBs and potentially form long-range regulatory networks that are central to the broader focus of this thesis.

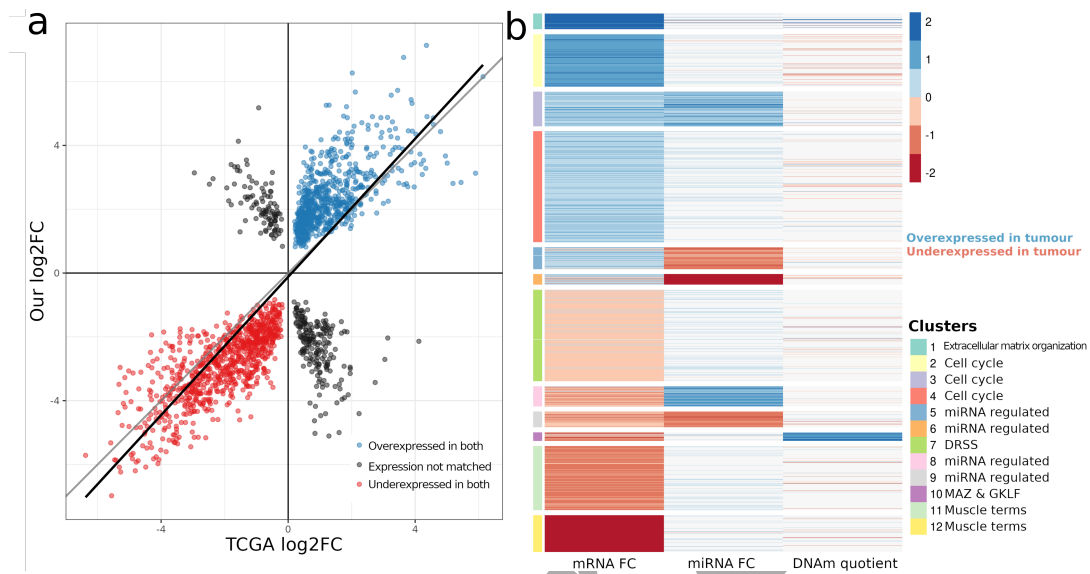


Figure 4.13. Comparison of the Croatian and TCGA-HNSCC cohorts. a correlation of fold change values of 1834 DEGs present in both datasets, with the $y = x$ identity line shown in grey, and while the best fit line through the data shown in black. b k-means clustering on the integrated transcriptome, miRNome and methylome data from the TCGA on the same set of DEGs; with biological programmes assigned to each cluster using gene set enrichment. Adapted from Mandić et al. [103].

4.2.7 Long-range epigenetic regulation

The role of GRBs and long-range regulation in HNSCC has not yet been systematically investigated. Therefore, in the context of this thesis, we used HNSCC as a disease model to explore how GRBs shape CpG architecture, CpG probe methylation and gene regulation. GRB target genes tend to harbour long, and often multiple CGI in promoters that extend into the gene body, a feature that distinguishes them from neighbouring bystander genes [40, 45]. However, CpG distributions across entire GRBs have not yet been systematically examined beyond the promoter region. We compared CpG distributions within GRBs to those observed in randomly shuffled genomic regions of equivalent length. This analysis revealed that GRBs have a markedly different CpG density relative to non-GRB regions (Figure 4.14 a-c) across genomic regions of different classes. GRBs show depletion of CpGs in promoter and gene body regions (Figure 4.14 a-b), while intergenic regions are enriched compared to the shuffled non-GRB regions (Figure 4.14 c). In line with previous research, target genes have a higher number of CpGs in promoters than bystander genes (Figure 4.14 d).

A hallmark feature of cancer is the alteration of the methylation status across the whole

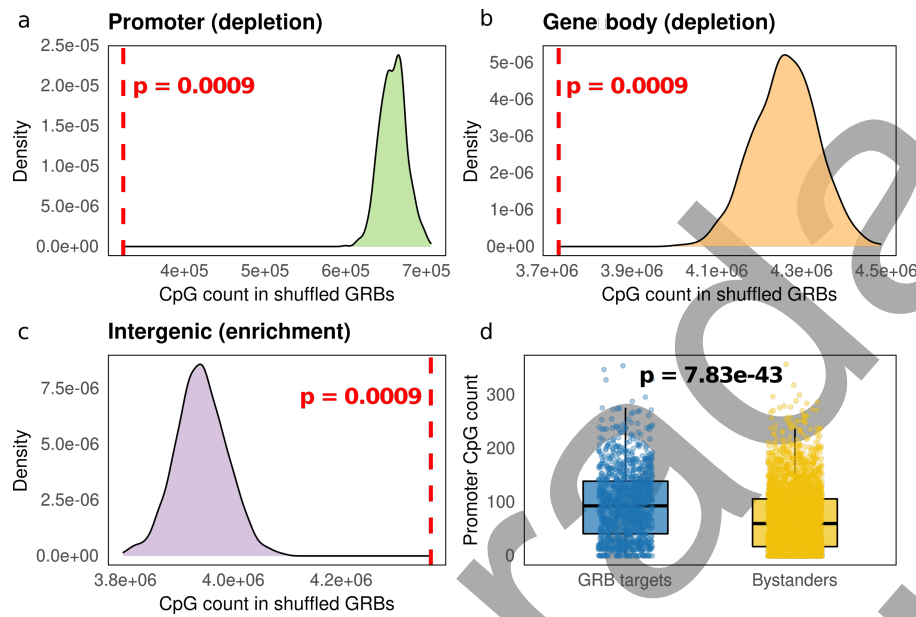


Figure 4.14. Count distribution of CpG dinucleotides falling into a promoters; b gene body regions; c intergenic regions of randomly shuffled GRB regions with the vertical line denoting the observed GRBs with corresponding empirical p-value. d CpG dinucleotide count in promoters of target and bystander genes with Wilcoxon rank sum p-value.

genome [98]. To assess whether GRBs and target genes exhibit similar changes in the methylation status in HNSCC, we applied a similar approach to the one used for the CpG architecture. We compared the ratio of differentially methylated CpG probes detected from 391 TCGA samples in observed GRBs to those in randomly shuffled genomic regions. GRBs show a markedly higher rate of DM probes (empirical $p < 0.05$) across promoters, gene bodies and intergenic regions (Figure 4.15 a-c), with the exception of the rate of hypomethylation in promoters. In promoters, hypermethylation events are more frequent than hypomethylation for both GRBs and shuffled ranges, whereas gene bodies and intergenic regions display the opposite trend, with hypomethylation predominating. A higher rate of hypermethylation is observed in the promoters of target genes when compared to bystander genes ($p < 0.05$, Figure 4.15 d), while this difference is significant in other cases, the effect size is small or negligible (Table 4.5), suggesting a potential silencing of target genes with tumour suppressor role in this disease.

Next, we showcase two examples where, in tumour samples, CpG probes located at a significant distance from the gene promoter show strong correlations with the expression of *WNT2B* and *PITX1*, indicating a possible contribution of distal methylation to long-range epigenetic regulation (Figure 4.16). *WNT2B* participates in Wnt signalling, a pathway central to the control

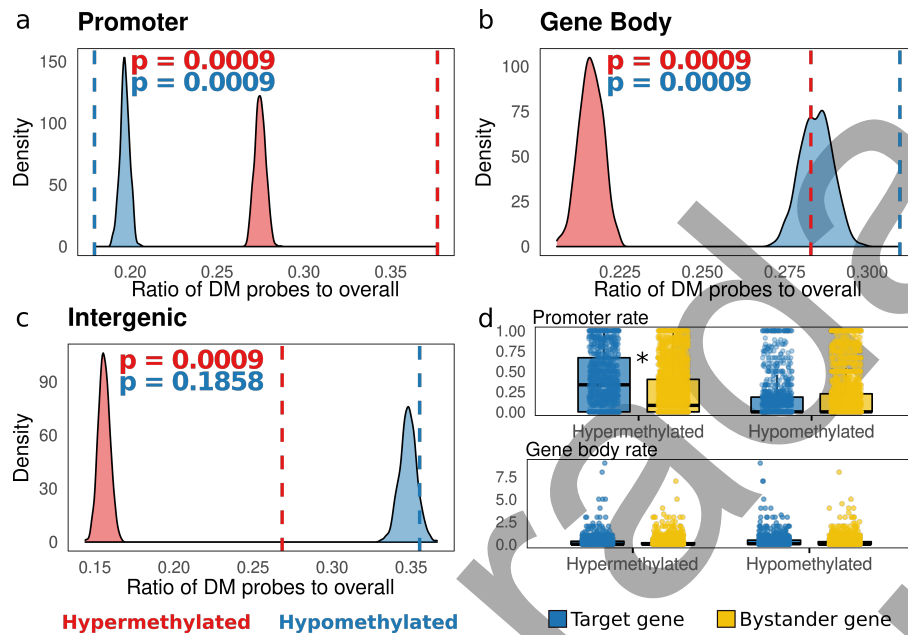


Figure 4.15. Distribution of ratios of hypermethylated (in red) and hypomethylated (in blue) CpG probes across a promoters; b gene body regions; c intergenic regions in randomly shuffled GRBs. Vertical lines in corresponding colours represent the ratios for observed GRBs. d target and bystander gene DM ratio for promoters and gene body regions. Adapted from Mandić et al. [103].

Table 4.5. Median methylation levels in GRB target and bystander genes for promoter and gene body regions, stratified by hyper- and hypomethylation, with corresponding Wilcoxon rank-sum p-values and Cliff's δ effect sizes.

Region	Methylation change	p-value	Target median	Bystander median	Cliff's δ (magnitude)
Promoter	Hypermethylation	1.30×10^{-46}	0.3333	0.0801	0.27 (medium)
	Hypomethylation	0.334	0	0	-0.02 (negligible)
Gene body	Hypermethylation	9.73×10^{-18}	0	0	0.14 (small)
	Hypomethylation	1.08×10^{-15}	0.0556	0	0.14 (small)

of cell proliferation and differentiation, whereas *PITX1* is a developmental transcription factor implicated in tumour-related transcriptional regulation. Although the functional impact of intergenic CpGs on long-range regulation in HNSCC remains incompletely understood, these results highlight the importance of considering long-range epigenetic regulation.

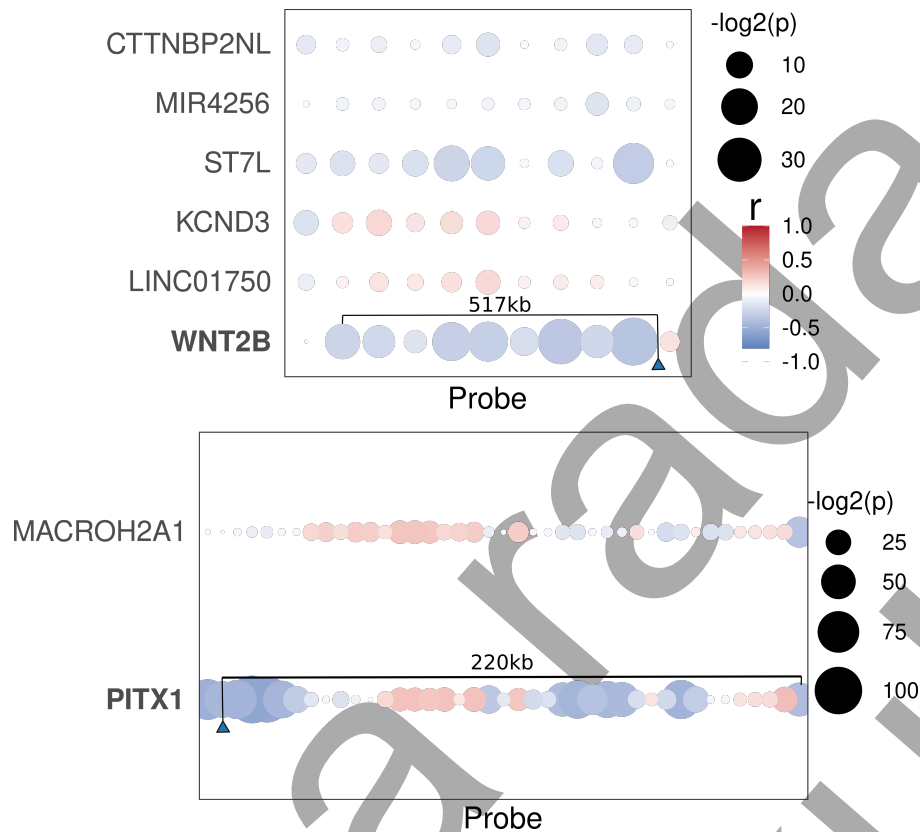


Figure 4.16. Pearson's correlation of CpG probes and genes within two GRBs with the target gene shown with bold letters (other genes are bystanders). For the target genes the blue triangle represents the relative position of the transcription start site in relation to CpG probes and the distance to the furthest significant probe is outlined. Adapted from Mandić et al. [103].

4.3 Results of the Parkinson's disease study

PD is the second disease investigated in this thesis. Although genetic factors contribute substantially to disease susceptibility, the molecular mechanisms linking genetic variation to disease pathogenesis remain incompletely understood [110]. Increasing evidence suggests that immune activation and chronic neuroinflammation play important roles in PD progression [113, 114]. To investigate these mechanisms, peripheral immune cell populations were profiled using scRNA-seq transcriptomics. DCs and CD4⁺ T-cells, two key regulators of innate and adaptive immune responses, were isolated from PBMCs of patients with PD and healthy controls and analysed at single-cell resolution. The resulting transcriptomic data were used to identify disease-associated immune cell subsets, characterize their transcriptional programs, and determine biological pathways altered in PD. Finally, additional analyses of exome sequencing data were performed to investigate disease-associated genetic variants, with a particular focus on identifying non-

coding variants that may contribute to PD susceptibility. Together, these analyses provide a comprehensive overview of peripheral immune dysregulation and genetic variation in PD, and establish the basis for the results presented in this chapter.

4.3.1 Generation of the single-cell dataset

To gain a better understanding of the role of immune cells in PD pathology, scRNA-seq was performed on CD4⁺ T-cells and pre-enriched DC populations isolated from PBMCs. Cells were obtained from whole blood samples of 17 patients with PD and 10 healthy controls. A summary of the cohort characteristics is provided in Table 4.6.

Initially, scRNA-seq libraries were prepared targeting a recovery of 10,000 cells per sample to maximise cell yield. Following suboptimal sequencing saturation in the first sequencing batch (Supplementary Data 1), the protocol was adjusted to target a recovery of 5,000 cells per sample to achieve the required sequencing depth. Owing to sequencer capacity, samples were processed in batches of four. For each sample, 2,500 cells from each enriched fraction (CD4⁺ T-cells and DCs) were combined for gene expression library preparation. During cell separation, PD sample S13 yielded only the CD4⁺ T-cell fraction; therefore, a total of 5,000 CD4⁺ T-cells were included for sequencing. To assess transcriptional variability over time, a biological replicate was collected from one patient one week apart (samples S28 and S30). In addition, sample S29 was excluded from downstream analyses due to poor sample quality and low annotation confidence across all cell type clusters (more detail in Section 4.3.2).

Table 4.6. Clinical characteristics of participants included in the scRNA-seq cohort. Adapted from Meglaj Bakrač et al. [155].

Feature	CTRL	PD
N	10	17
<i>Demographic data</i>		
Female, <i>n</i> (%)	4 (40)	10 (59)
Median age at inclusion (range)	60 (48–81)	65.5 (50–86)
Median age at onset (range)	n/a	61 (39–69)
Early-onset PD (<50 years), <i>n</i>		4
Late-onset PD (≥50 years), <i>n</i>		11
Onset unknown, <i>n</i>		2

CTRL, control; PD, Parkinson's disease; n/a, not applicable.

Following single-cell 5' gene expression sequencing, Cell Ranger processing, quality control, and integration of all samples, a total of 45,345 high-quality cells were retained, comprising 24,660 cells from PD samples and 20,685 cells from CTRL samples. The distribution of detected genes and UMIs per cell after quality control is shown in Figure 4.17a. Cells were visualised using tSNE for dimensionality reduction, demonstrating that cells containing a higher number of UMIs clustered together (Figure 4.17b). Sequencing metrics including the total number of reads and cells detected per sample, are summarised in Supplementary Data 1. Integration of all samples using reciprocal PCA followed by SNN clustering identified 27 distinct cell clusters (Figure 4.17c).

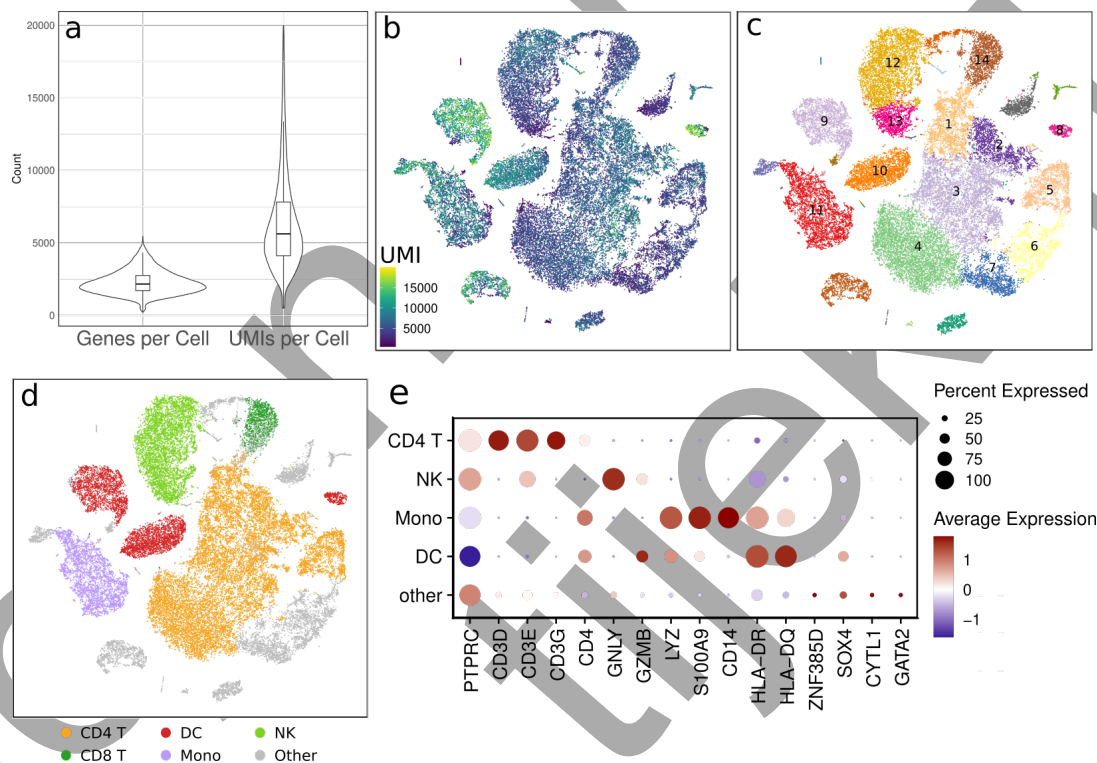


Figure 4.17. Single-cell sequencing data analysis. a Distribution of the number of detected genes and UMIs per cell following quality control. b tSNE projection of all cells coloured by the number of UMIs per cell. c tSNE projection showing unsupervised SNN clustering, identifying 27 cell clusters. Cluster numbers correspond to the major cell populations identified in the dataset: clusters 1–5, CD4+ T-cells; clusters 6–7, low-confidence CD4+ T-cells; clusters 8–10, dendritic cells; cluster 11, monocytes; clusters 12–13, natural killer cells; and cluster 14, low-confidence CD8+ T-cells. d Final cell type annotation based on reference-guided annotation and SNN clustering. e The expression of selected marker genes across the major cell populations [156–159]. Dot size represents the proportion of cells expressing each marker, while colour indicates the average expression level. Adapted from Meglaj Bakrač et al. [155].

4.3.2 Data validation

This section describes the internal and external validation performed on the generated dataset. Internal validation included assessment of sample quality to identify outliers, donor-specific cell distribution in the tSNE embedding, biological replicates collected one week apart, the distribution of PD and CTRL cells across the tSNE embedding, and differential expression analysis using a pseudobulk approach. External validation was performed by comparing the gene expression profiles of CD4+ T-cells and DCs from this study with bulk RNA-seq data from the BLUEPRINT project (<https://projects.ensembl.org/blueprint/>) and a reference PBMC single-cell dataset [141].

Quality control metrics following cell calling were assessed across all samples (Supplementary Data 1). Sample S29 was identified as a low-quality outlier due to a fraction of reads assigned to cells below 0.6 (recommended > 0.7), together with a high estimated cell count (Figure S5a). In addition, it had the lowest mean number of reads and detected genes per cell (Figure S5b). The poor cell quality most likely contributed to the low mean annotation confidence observed across all major cell populations (Figure S5c). This sample was therefore excluded from subsequent analyses.

Cells from each donor were mapped onto the integrated tSNE embedding to assess their distribution across the identified clusters and cell types (Figure S4). Cell distributions were consistent across donors and showed no obvious association with disease status or sex.

Biological replicates from one PD patient were collected one week apart using the same experimental protocol (samples S28 and S30). Differential expression analysis between these samples identified only a small number of DEGs, indicating minimal variation between the two samples (Figure 4.18a). Consistent with this finding, no obvious separation between the biological replicates was observed in the tSNE embedding (Figure 4.18b).

The distribution of PD and CTRL cells was further examined across CD4+ T-cells and DCs. Cell type clustering represented the primary source of variation within the dataset, with no clear separation between PD and CTRL cells (Figure 4.18c). A slight rightward shift of PD cells along the first tSNE dimension was observed in both CD4+ T-cell and DC populations.

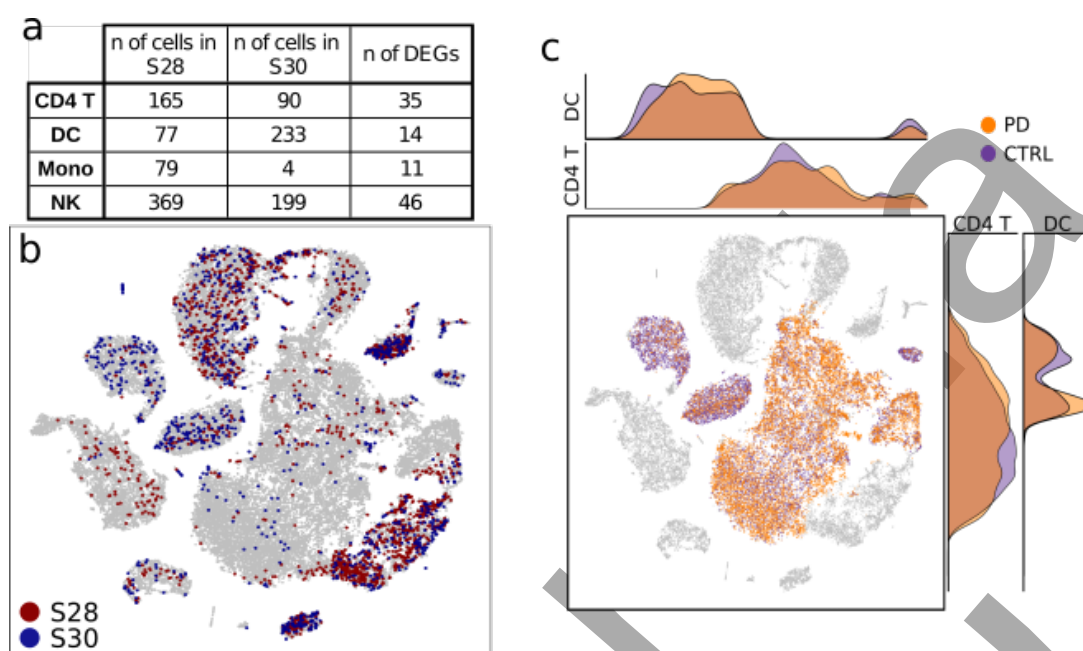


Figure 4.18. Biological replicate validation and distribution of PD and CTRL cells. a Number of cells recovered from the biological replicates collected one week apart (samples S28 and S30), together with the number of differentially expressed genes (DEGs) identified between them. b tSNE projection showing the distribution of cells from biological replicates S28 and S30. c tSNE projection of PD and CTRL cells with the corresponding density distributions across the CD4⁺ T-cell and DC populations. Adapted from Meglaj Bakrač et al. [155].

Pseudobulking was utilised to address the pseudoreplication problem of single-cell statistical methods. Many single-cell differential expression methods work under the assumption that individual cells are independent and identically distributed, which is often not the case in single-cell analyses, since multiple cells can come from one donor. The pseudobulk approach was applied by aggregating cells from each donor and performing differential expression analysis for each cell subtype using a standard bulk RNA-seq pipeline. Although no genes reached the significance threshold, likely owing to reduced statistical power following aggregation, FC values strongly correlated with those obtained from the single-cell analysis, with Pearson's correlation coefficients exceeding 0.8 for all analysed cell subtypes (Figure S6).

External validation was performed by comparing the top-expressed genes from the single-cell dataset with bulk RNA-seq data from the BLUEPRINT project and a reference PBMC single-cell dataset [141]. Average gene expression in CD4⁺ T-cells was compared with bulk RNA-seq data from CD4-positive alpha-beta T-cells isolated from venous blood, while DCs were compared with conventional DCs isolated from cord blood. Gene expression profiles from

both CD4+ T-cells and DCs showed a high correlation with the corresponding bulk RNA-seq datasets (Figure S7a,b). Similarly, comparison with the PBMC single-cell reference dataset demonstrated a high correlation for both cell types (Figure S7c,d).

4.3.3 Identification of major cell populations

Upon successful validation of samples, cell type annotation was performed using a human PBMC reference dataset [138]. Cell clusters were annotated using a combination of SNN clustering and reference-based annotation. Initially, all cells were assigned to seven clusters: CD4+ T-cells, CD8+ T-cells, DCs, monocytes, NK cells, other T-cells, and “other” (Figure S8a). The “other T-cells” and “other” clusters were subsequently merged into a single dispersed cluster labelled “other”. In addition, a subset of CD4+ T-cells (clusters 6 and 7) identified by the reference-based annotation was reassigned to the “other” cluster due to low prediction scores (Figure S8b) and the absence of CD3 complex gene expression (Figure 4.17c,d). The CD8+ T-cell cluster was also merged into the “other” group owing to its low annotation confidence (Figure S8b). The resulting five major cell populations expressed canonical marker genes characteristic of their respective cell types (Figure 4.17e). Some cells within the “other” cluster expressed marker genes associated with granulocyte-monocyte progenitors (*ZNF385D*, *SOX4*, *CYTLL1*, and *GATA2*) [120], which likely prevented their assignment to one of the major immune cell populations.

The scRNA-seq data also highlighted the limitations of magnetic cell separation. In addition to the expected CD4+ T-cells and DCs, monocytes, CD8+ T-cells, and NK cells were detected. Sample S13, which contained only the CD4+ T-cell fraction, demonstrated that a proportion of monocytes and CD8+ T-cells remained within the CD4+ T-cell-enriched fraction, suggesting that the NK cells originated from the DC-enriched fraction (Figure S8c). Monocyte and NK-cell identities were confirmed by the expression of established marker genes, including *CD14*, *S100A9*, and *CD68* for monocytes, and *NKG7*, *GNLY*, and *NCAM1* for NK cells (Figure S8d) [156–159]. As the analysed samples were enriched for the target cell populations (CD4+ T-cells and DCs), these non-target cell populations were not considered representative of the peripheral immune compartment and were therefore excluded from subsequent differential expression and downstream analyses.

4.3.4 Identification of CD4+ T-cell subtypes

The tSNE embedding identified CD4+ T-cells as a single large cluster (Figure 4.19a). All cells expressed the canonical marker genes encoding the CD3 complex subunits (*CD3D*, *CD3E*, and *CD3G*) together with *CD4*. In addition, the majority of CD4+ T-cells expressed *CAMK4* and *BCL11B*, both not commonly used as markers of CD4+ T-cells (Figure S8e).

Reference-based annotation at level 2 identified four distinct CD4+ T-cell subtypes: naïve T-cells, central memory T-cells (TCMs), effector memory T-cells (TEMs), and regulatory T-cells (Tregs) (Figure 4.19a). Cell identities were further confirmed by the expression of established marker genes shown in Figure 4.19b. Specifically, *GZMK*, *GZMA*, and *CCL5* were associated with TEMs, *ITGB1* with TCMs, *CCR7* and *SELL* with naïve T-cells, and *FOXP3*, *IL2RA*, and *CTLA4* with regulatory T-cells. Marker gene selection was guided by the reference dataset and supported by previous studies [160–162]. To further characterise the identified CD4+ T-cell subtypes, pseudotime analysis was performed by ordering cells along a trajectory inferred from their gene expression profiles. As expected, the resulting trajectory reflected the differentiation continuum from naïve T-cells through TCMs to TEMs (Figure 4.19c; Figure S9a).

To investigate potential differences in CD4+ T-cell subtype composition, subtype proportions within the total CD4+ T-cell population were compared between PD and CTRL samples (Figure 4.19c). No significant differences were observed for any of the identified CD4+ T-cell subtypes (Supplementary Table S3).

4.3.5 Identification of dendritic cell subtypes

SNN clustering identified three distinct DC clusters, which corresponded well with the DC subtypes assigned by reference-based annotation (Figure 4.20a). These clusters represented the three major DC subtypes: conventional dendritic cell 1 (cDC1), conventional dendritic cell 2 (cDC2), and plasmacytoid dendritic cells (pDCs). Genes encoding the HLA-DR and HLA-DQ complexes were highly expressed across all three subtypes, whereas *CD86* expression was more prominent in the conventional DC subsets. Cell identities were further confirmed by the expression of established marker genes (Figure 4.20b). *ITGAX* was used as a general marker of

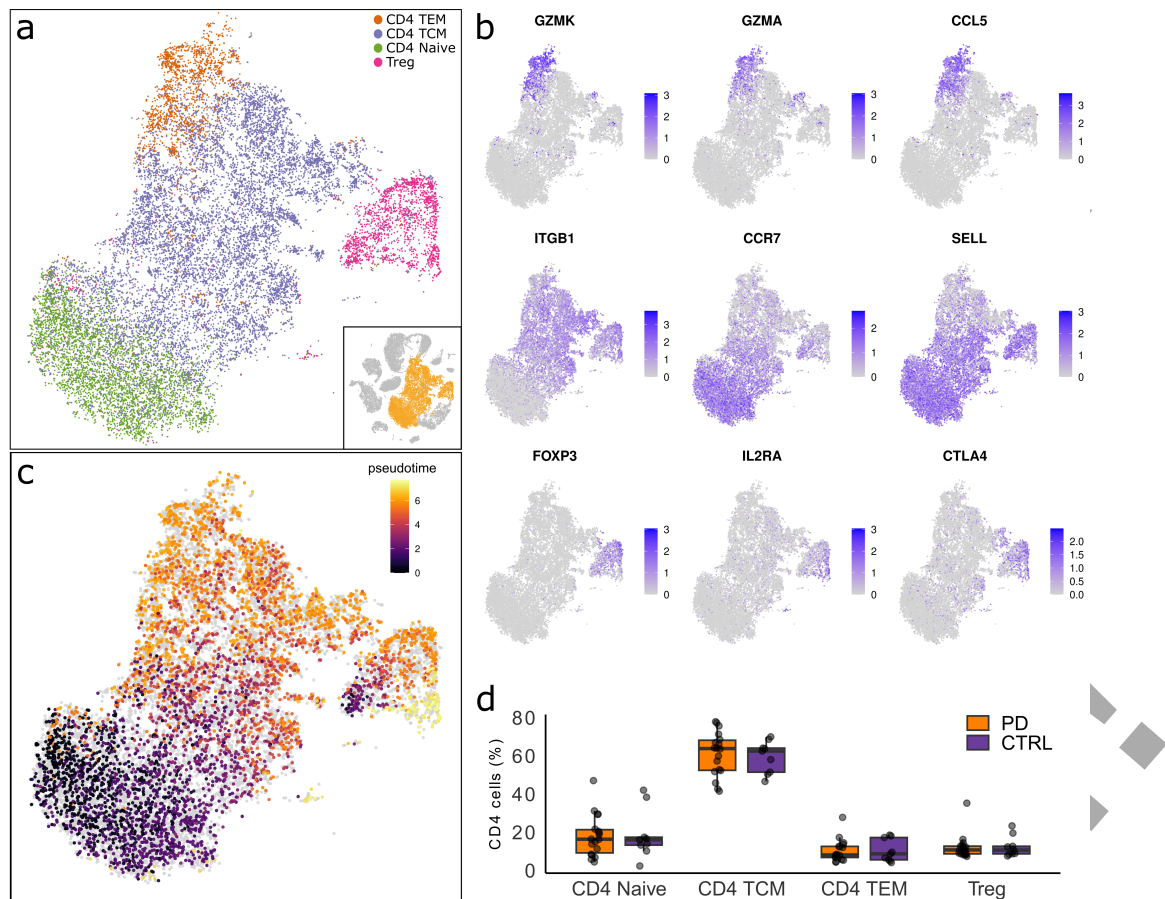


Figure 4.19. Single-cell analysis of CD4+ T-cell subtypes. a tSNE projection showing the identified CD4+ T-cell subtypes. b Feature plots showing the expression of marker genes used for CD4+ T-cell subtype annotation, with higher expression levels indicated in purple. c Relative proportions of CD4+ T-cell subtypes within the total CD4+ T-cell population in PD and CTRL samples. Adapted from Meglaj Bakrač et al. [155].

conventional DCs, while subtype-specific markers included *CADM1* and *CLEC9A* for cDC1s, *CD1C* and *CLEC10A* for cDC2s, and *CLEC4C* and *IL3RA* for pDCs [120, 156].

A small subset of cells expressed *CD163*, a marker associated with the DC3 subtype of cDC2s. As DC3 cells are also characterised by expression of the monocyte marker *CD14*, the expression profile of *CD14* is shown alongside *CD163* in Figure 4.20b to provide additional context [163]. The differentiation trajectory analysis did not show branching for DC subtypes (Supplementary Figure S9), consistent with their distinct developmental origins: cDCs originate from the myeloid lineage, whereas pDCs arise from a separate lymphoid progenitor.

To investigate potential differences in DC subtype composition, subtype proportions within the total DC population were compared between PD and CTRL samples (Figure 4.20c). No significant differences were observed for any of the identified DC subtypes (Supplementary

Table S3).

4.3.6 Differential gene expression in PD vs CTRL

Differential gene expression analysis between the PD and CTRL groups was performed using Seurat's FindMarkers function, focusing on the four CD4⁺ T-cell subtypes and three DC subtypes identified in the dataset. The total number of differentially expressed genes (DEGs) identified in each cell subtype is summarised in Table 4.7, while all the DEGs can be found in Figure 4.21 and Supplementary Data 13. Across all cell subtypes, the number of upregulated genes was substantially higher than the number of downregulated genes.

Of note, we have identified that CD4⁺ cells in our dataset broadly express *BCL11B* gene

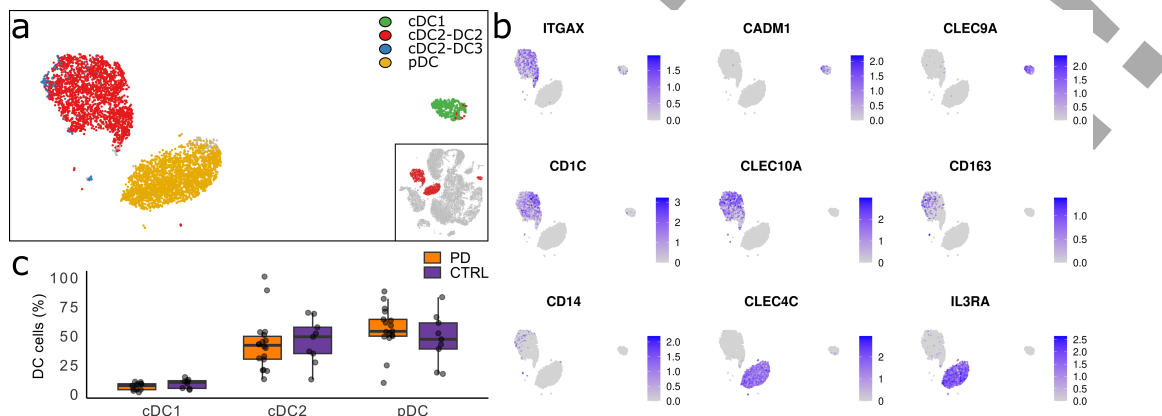


Figure 4.20. Single-cell analysis of DC subtypes. a tSNE projection showing the identified DC subtypes. b Feature plots showing the expression of marker genes used for DC subtype annotation, with higher expression levels indicated in purple. c Relative proportions of DC subtypes within the total DC population in PD and CTRL samples. Adapted from Meglaj Bakrač et al. [155].

Table 4.7. Numbers of DEGs identified in each immune cell subtype. Adapted from Meglaj Bakrač et al. [155].

Cell subtype	Total DEGs	Upregulated	Downregulated
CD4 naïve T-cells	328	320	8
CD4 TCM	1604	1593	11
CD4 TEM	457	450	7
CD4 regulatory T-cells	210	207	3
cDC1	18	16	2
cDC2	230	190	40
pDC	447	433	14

DEG, differentially expressed gene; TCM, central memory T-cell; TEM, effector memory T-cell; cDC, conventional dendritic cell; pDC, plasmacytoid dendritic cell.

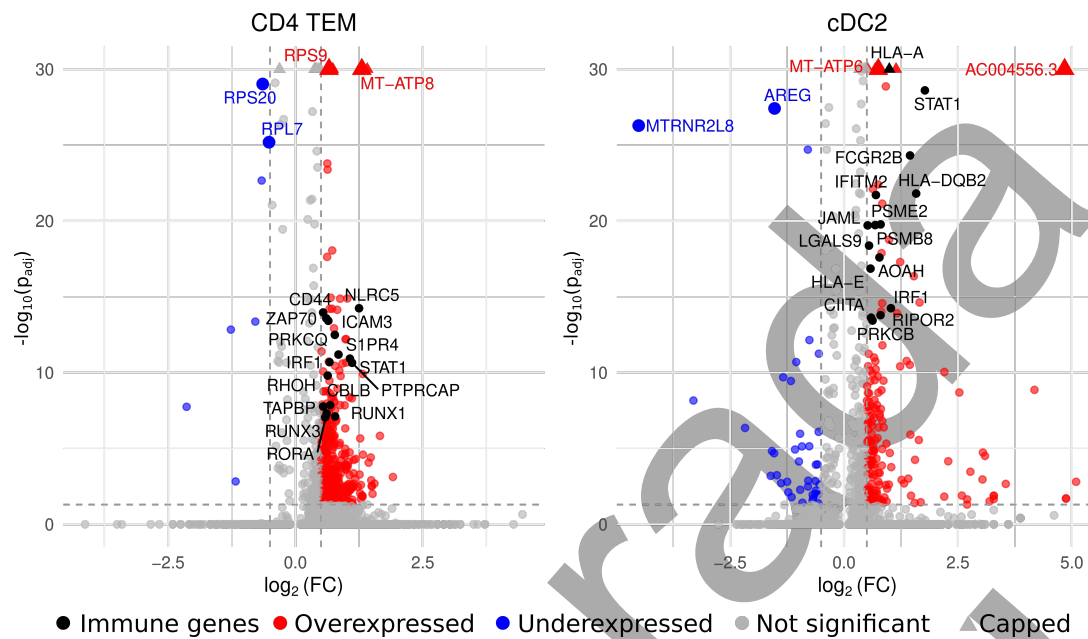


Figure 4.21. Differential gene expression in CD4+ TEMs and cDC2s. Volcano plots showing DEGs in CD4+ TEMs (left) and cDC2s (right). Immune-related genes with an adjusted $p < 10^{-10}$ are labelled in black. Genes exceeding the upper limit of the y-axis are predominantly mitochondrial and ribosomal protein genes. The complete lists of DEGs are provided in Supplementary Data 13. Adapted from Meglaj Bakrač et al. [155].

(Supplementary Figure S8e): a GRB target gene found to be under long-range regulation during T-cell development [164]. On closer inspection of deregulation of expression of *BCL11B* we identified that PD patient exhibit significant overexpression across all CD4+ cells as well as Naïve and TCM subtypes, albeit with a small FC value (Table 4.8). Full list of GRB target genes deregulated in PD versus CTRL are reported in Table S6.

Table 4.8. Differential expression statistics of *BCL11B* across CD4+ T-cell subtypes, including differential expression significance, $\log_2FC$, and the proportion of cells expressing the gene in Parkinson's disease (PD) and control (CTRL) samples.

Cell type	p-value	$\log_2FC$	pct. PD	pct. CTRL	Adj. p-value
Naïve	1.29×10^{-9}	0.352	0.792	0.711	3.34×10^{-5}
TCM	2.39×10^{-28}	0.294	0.775	0.663	6.19×10^{-24}
TEM	3.87×10^{-6}	0.393	0.681	0.572	0.100
Treg	—	—	—	—	—
CD4+ total	4.14×10^{-41}	0.269	0.767	0.664	1.07×10^{-36}

FC, fold change; PD, Parkinson's disease; CTRL, control; TCM, central memory T-cell; TEM, effector memory T-cell; Treg, regulatory T-cell.

4.3.7 Pathway enrichment in PD vs CTRL

To better understand the functional implications of the identified transcriptional changes in PD, gene set enrichment analysis was performed using the GO:BP, KEGG, and Reactome databases for the DEGs identified in each cell subtype. As KEGG and Reactome terms provided limited biological insight, subsequent analyses focused on GO:BP.

Among the CD4⁺ T-cell subtypes, DEGs from naïve T-cells were enriched for only three GO terms. In contrast, CD4⁺ TCMs, TEMs, and Tregs were enriched for 83, 180, and 46 terms, respectively. A complete list of enriched terms is provided in Supplementary Data 14. Enriched terms in CD4⁺ TCMs and Tregs were predominantly associated with metabolic and biosynthetic processes, translation, and gene expression, with relatively few terms related to immune function. We therefore focused further analyses on CD4⁺ TEMs, a subtype known to initiate effector T-cell responses following activation. Compared with the other CD4⁺ T-cell subtypes, DEGs in CD4⁺ TEMs were enriched for a greater number of immune-related processes, including lymphocyte activation and differentiation (Supplementary Data 14). Additional enriched terms included RNA metabolism and ribosomal subunit organisation, likely reflecting the contribution of ribosomal protein genes; a common yet biologically irrelevant trend often found in such analyses [165].

Within the DC populations, cDC1 DEGs were enriched for 11 GO terms, several of which were associated with antigen processing and presentation (Supplementary Data 14). In comparison, cDC2 DEGs were enriched for 129 GO terms, primarily related to immune functions, including antigen processing and presentation, lymphocyte activation, cytokine production, and the regulation of immune responses. Consequently, subsequent analyses focused on cDC2s and pathways representing their principal biological functions: antigen processing and presentation, T-cell activation, and cytokine production. In contrast, only a single enriched term, “gene expression (transcription)”, was identified for pDCs.

The top 30 GO terms for CD4⁺ TEMs and cDC2s were ranked by recall value (Figure 4.22a,b), and representative terms reflecting the principal functions of each cell type were selected for further visualisation (Figure 4.22c,d).

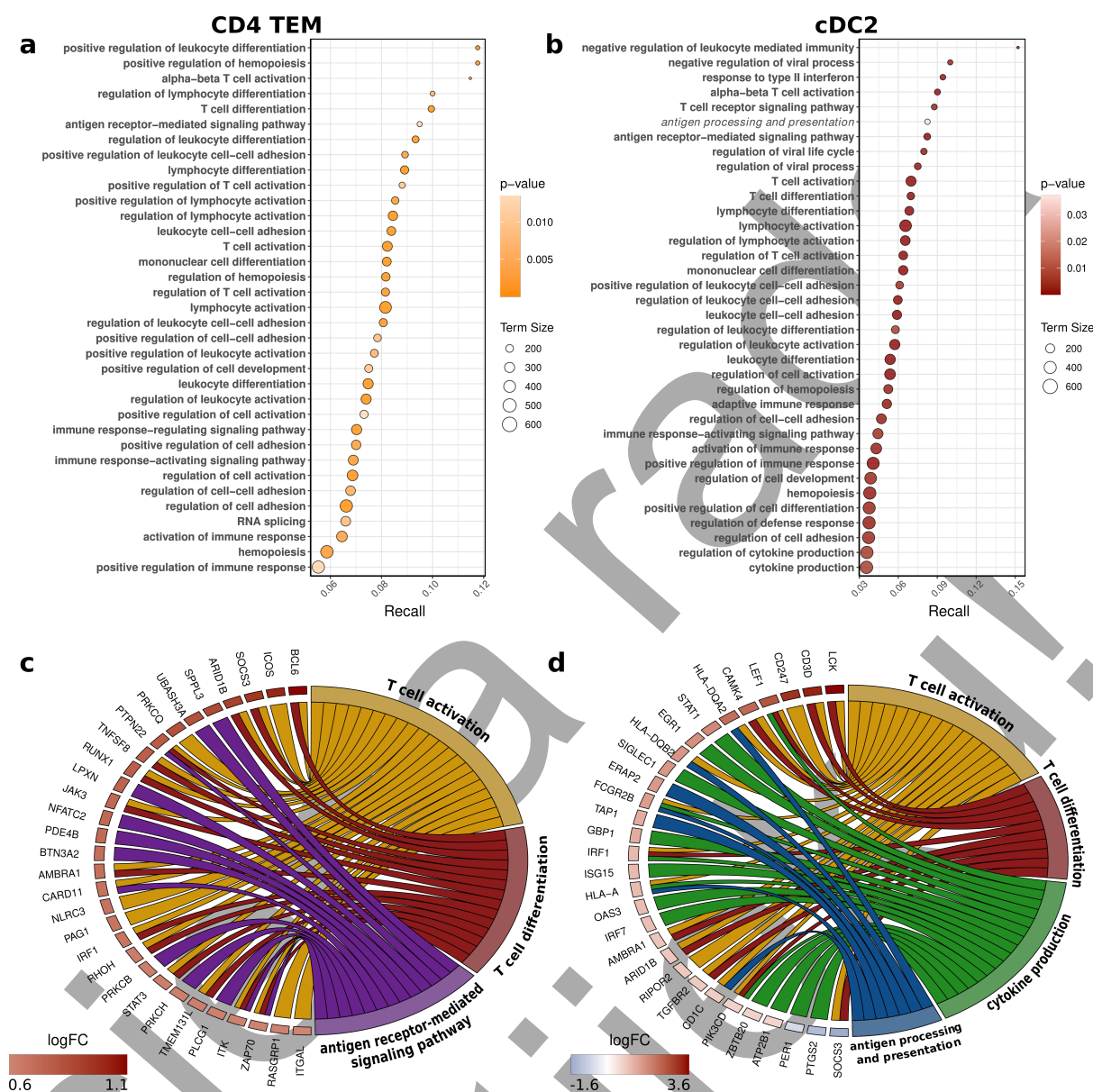


Figure 4.22. Pathway enrichment analysis of CD4+ TEMs and cDC2s. a,b GO:BP process terms enriched among DEGs in PD compared with CTRL samples, ranked by recall. Circle size represents the number of genes associated with each term, while colour indicates the enrichment p-value. c,d Circos plots showing selected enriched pathways and the top 30 DEGs ranked by log₂FC. Panels a and c correspond to CD4+ TEMs, whereas panels b and d correspond to cDC2s. Adapted from Meglaj Bakrač et al. [155].

4.3.8 Non-coding variation identification from exome sequencing

Moving beyond the single-cell analysis prepared for Meglaj Bakrač et al. [155], we proceed to investigate genetic burden, as identified by exome sequencing from an overlapping, but larger, PD cohort (Table 4.9). The sequencing of 61 patients did not reveal coding variants associated with PD (Meglaj-Bakrač, in preparation). To further utilise the exome dataset by available

targPred data, we expanded the search to the non-coding part which were captured during whole exome sequencing (Figure 4.23a). Available annotations from variant calling format files were utilise to select a set of strictly non-coding variants - variant that overlap different transcripts and those with conflicting annotations were excluded (Figure 4.23a).

Table 4.9. Demographic characteristics of the exome cohort.

Feature	CTRL	PD
N	16	64
<i>Demographic data</i>		
Female, <i>n</i> (%)	5 (31)	27 (42)
Median age at inclusion (range)	64.0 (48–81)	66.0 (48–86)

CTRL, control; PD, Parkinson's disease.

The non-coding variants were further overlapped with loci associated with PD-related traits obtained in the targPred workflow (Section 4.1.1). To identify over-representation of the variants in PD vs CTRL we used Fisher's exact test and identified only one variant which passed the nominal p-value threshold (Supplementary Table S4). Additionally, we examined whether the identified loci are enriched in GRBs by applying a binomial test, with the expected probability 33% - which is roughly the GRB coverage of the genome. This resulted in three traits that were found over-represented in GRBs (Figure 4.23, Supplementary Table S5).

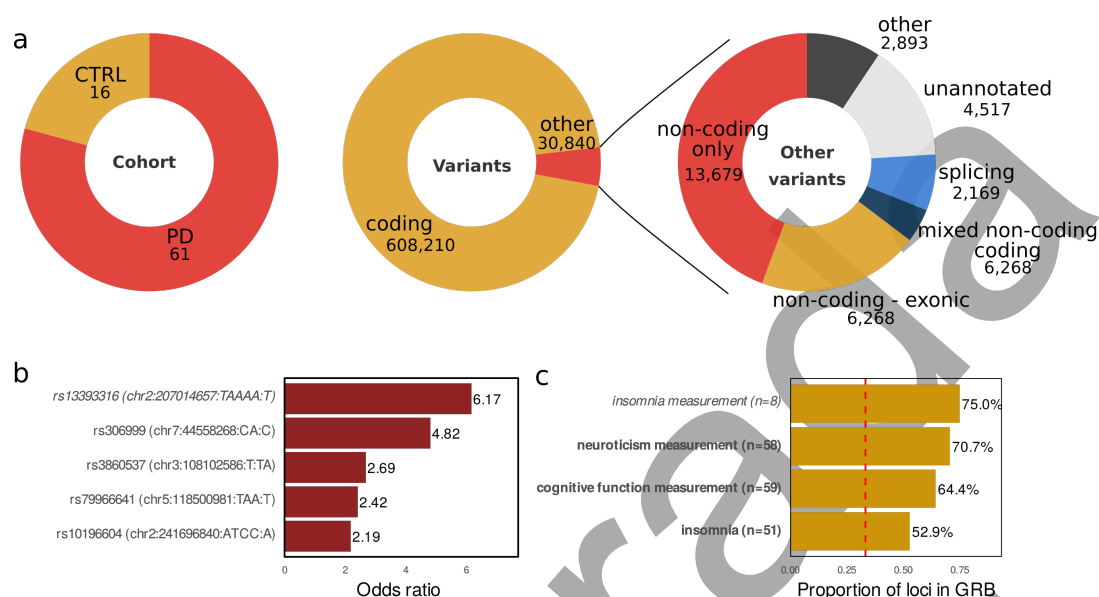


Figure 4.23. Non-coding variants from exome sequencing enrichment. **a** Non-coding variants identified from whole exome sequencing. Left ring shows the number of individuals in the cohort. Middle ring is the total count of variants denoted as exons and coding. Right ring categorises the “rest” category from the middle ring as follows: non-coding only - strictly non-coding variants; non-coding exonic - contain a functional non-coding flag and exon flag; mixed non-coding coding - coding depending on different transcripts; splicing, unannotated and other - annotation from functional flag. **b** Top 5 variants enriched in PD, selected by adjusted p -value and ordered by odds ratio, Fisher’s exact test. **c** Top four traits showing proportion of loci overlapping GRBs, binomial test. Number of loci containing a non-coding variant in the exome sequencing data is shown in parentheses. Expected frequency of locus overlap with GRB is $1/3$ (denoted with red vertical line). Labels in *italic* denote variants and traits passing the nominal, while **bolded** passed the BH-adjusted threshold of p – *value* < 0.05 .

5 Discussion

5.1 Discussion of the targPred study

In this thesis, we present a method for improved utilisation of the accumulating GWAS data to facilitate the mechanistic interpretation of disease-associated non-coding genomic variants, named targPred. We developed a comparative genomics approach based on the conserved synteny of large genomic regions, termed genomic regulatory blocks, across vertebrate evolution, thereby providing an optimal set of distal regulatory variants for each gene. This substantially expands the number of SNP–gene associations by incorporating variants acting through long-range genomic interactions, while simultaneously avoiding the multiple testing burden and the resulting p-value inflation. This balance of tested enhancer–gene pairs is achieved by focusing exclusively on genomic regions with well-established evidence of harbouring distal regulatory elements. More than 37,000 unique GWAS loci (158,611 locus–trait pairs) have novel target genes predicted by targPred at present, with the GWAS Catalog continuously expanding in both the number of deposited studies and their biological complexity, with no indication of reaching redundancy [166].

By comparing targPred to similar publicly available tools, we demonstrated that different approaches frequently produce non-overlapping SNP–gene associations. We first evaluated how well the target gene assignments from the original GWAS Catalog and targPred agreed with two reference datasets of gene–trait associations: a set of annotated coding variants linked to individual traits (Figure 4.3a)[126], and a set of experimentally validated regulatory variants associated with genes and traits through extensive literature curation (Figure 4.3b)[128]. These comparisons indicate that both the original GWAS assignments and the GRB-based predictions

recover valid SNP–gene associations, although the Alsheikh dataset displays high fraction of overlapping assignments between GWAS and our method, implicating modest novelty from the addition of our predictions. This observation suggests that long-range variant–gene associations that have already been described in the literature are often correctly represented in the GWAS Catalog. However, when no prior evidence exists for a non-coding variant, other sources of knowledge, such as coding variant datasets including ClinVar, appear to be underutilised. Coding variants affecting the same gene can provide independent evidence supporting its involvement in the trait and, consequently, the target gene predictions made by targPred. Furthermore, although targPred provides only a modest enrichment of only four novel validated associations in the Alsheikh dataset, the prostate carcinoma locus associated with the *SOX9* gene and missed by the original GWAS assignment, illustrates that the validity and utility of novel SNP–gene assignments should ultimately be assessed on a locus-by-locus basis.

We further compared the SNP–gene predictions produced by targPred with those generated by several state-of-the-art prediction methods, including ABC, GeneHancer and cS2G (Figure 4.3). Although these comparisons are primarily based on quantitative prediction scores, they demonstrate that, for SNP–gene pairs passing the empirical significance threshold in targPred, high prediction scores from these methods are significantly enriched. This concordance provides independent support for the long-range regulatory interactions identified by targPred. The agreement between these independent methods is particularly encouraging given their distinct underlying assumptions. While ABC, GeneHancer and cS2G rely on functional genomic and epigenomic evidence, targPred exploits the evolutionary conservation of GRBs and CNEs together with an empirical statistical framework, to identify candidate long-range regulatory interactions.

The *BDNF* locus presented in figures 4.3 and 4.6 illustrates both the interpretation of targPred predictions and the potential biological utility of alternative SNP–gene associations intended to guide downstream experimental studies. Within this GRB, *BDNF* is identified as the target gene regulated through long-range interactions, with Enh5 showing a significant empirical association with increased *BDNF* transcriptional activity. The regulatory variant overlapping Enh5 was originally assigned to the *LINC02758* gene in the GWAS Catalog [151]. In contrast,

the ABC method does not predict enhancer–gene interactions for this locus, while cS2G links other enhancers within the same GRB to *METLL15* and *KIF18A*. GeneHancer, which includes the broadest collection of enhancer annotations, including those from FANTOM5, associates Enh5 with *METLL15* and *KIF18A*, although with low-confidence scores.

Finally, a comprehensive analysis of GWAS non-coding variants across hundreds of traits demonstrated that certain diseases and phenotypic traits are more frequently associated with regulatory elements and their corresponding long-range regulatory interactions. As expected, a broad range of psychiatric disorders showed significant enrichment, consistent with previous observations that genes under extreme long-range regulation are over-represented in neuronal functions (Figure 4.2c). Additional enrichments were also observed for cancer, immune disorders, asthma and metabolic disorders. Collectively, the evidence from previous proof-of-concept studies in obesity [47] and schizophrenia [70], together with the examples of prostate cancer and childhood obesity presented in this study, the propensity for long-range regulation in three of the eight asthma ontology terms (Figure 4.2a), and the over-representation of GWAS loci under long-range regulation across the psychiatric disease lineage (Figure 4.2B), supports the use of targPred as an important step in the interpretation of GWAS SNP-trait-gene associations.

5.2 Discussion of the head and neck squamous cell carcinoma study

HNSCC is one of the diseases of special interest in this thesis, as its development and progression appear to be strongly influenced by epigenetic regulation, including DNAm and miRNA-mediated control. Within this contribution, we use the Croatian HNSCC cohort to investigate these regulatory changes in detail in relation to gene expression, as well as validating the results with a larger and publicly available TCGA cohort. Furthermore, epigenetic regulation was investigated in relation to GRBs, their target genes, and broader long-range epigenetic effects.

In this study, we integrated miRNome, methylome, and transcriptome data to investigate regulatory mechanisms involved in HNSCC development and progression. Unsupervised clustering identified distinct patterns across the three omics datasets, suggesting that miRNA- and

DNAm-mediated regulation of gene expression largely occurs in a mutually exclusive manner. The clustering analysis also revealed key tumour progression-related pathways, including altered cell cycle regulation, transcription, immune response, and vasculature-associated genes. In addition, our results underscore the relevance of promoter DNAm in genes targeted by the *E2F* transcription factor family.

The exploratory analysis of the transcriptome data separated samples into two principal clusters corresponding to tumour and control tissues, with no clear segregation by HPV status or anatomical tumour site (Figure 4.8), also consistent in higher principal components (Figure S1). Comparable tumour–control distinctions were observed in the overlapping miRNA expression and DNAm analyses reported previously [85, 86]. These results imply that the tumours in the Croatian HNSCC cohort may derive from related cells of origin or activate common oncogenic programmes during tumourigenesis. Moreover, the set of differentially expressed genes ($n = 2,781$, Figure 4.9) is enriched for pathways that mirror transcriptomic signatures previously associated with radioresistant HNSCC, including hypoxia response, the p53 pathway, NF- κ B signalling and inflammatory responses [74].

The unsupervised clustering procedure shown in Figure 4.12 indicates that specific regulatory mechanisms shape gene expression patterns in tumours compared with control samples, forming distinct biological programmes. Two clusters of overexpressed genes involved in key cell cycle regulatory pathways showed predominantly hypomethylated promoters. In addition, among the underexpressed genes in tumours, we identified a cluster of DM-DEGs associated with regulation of transcription (Figure 4.12, cluster 10). Clusters 4 and 10 support a shared influence of DNAm on gene expression, whereby promoter methylation suppresses expression and reduced methylation promotes gene expression [167–169]. However, the opposite pattern of overexpression with hypermethylation and underexpression with hypomethylation was observed in clusters 1 and 6, respectively, both of which were enriched for *E2F* and other transcription factor targets. Similarly, an inhibitory effect of miRNAs on target genes was evident in cluster 3 enriched for vasculature-related genes, and in cluster 7 not showing any specific biological processes apart from grouping miRNA targets.

A positive correlation between miRNAs and their target genes was observed in clusters 5

and 7, consistent with previous studies [170, 171]. These results may reflect either direct or indirect effects of miRNAs on DE miR-DEGs, potentially through interference with the action of repressive miRNA–protein complexes [170]. However, these findings indicate correlation rather than causation, and additional studies are needed to establish the underlying mechanisms. It should also be noted that miRNAs often regulate multiple targets, while some interactions in existing databases are based on high-throughput methods that may not fully capture the complexity of in vivo biological interactions. Finally, clusters 2, 8 and 11 did not show strong evidence of regulation by either miRNA or DNAm. Notably, cluster 11 comprised of underexpressed genes with a strong association with muscle system processes; an unusual finding in the context of this type of cancer. However, this pattern was also reproduced in the TCGA dataset (4.13, clusters 11 and 12).

We found a cluster of genes under *E2F* control that was significantly underexpressed in tumour samples relative to controls (Figure 4.12). Previous studies have shown that the *E2F* family is negatively regulated by miRNAs in HNSCC and in other cancers, including lung cancer, paediatric brain tumours and chronic myeloid leukaemia [[102, 172–174]. In addition, the TCGA analysis reported increased copy number of *E2F1* in HPV-positive HNSCC [102]. Although *E2F* is often highly expressed and is known to play a central role in tumour growth, proliferation, metastasis and drug resistance [175], our data did not show overexpression of the *E2F* family in cancer compared with controls. Instead, clustering identified several dysregulated gene clusters corresponding to *E2F* targets, namely clusters 1, 4, 5, 6 and 10. The most prominent of these, clusters 4, 5 and 10, were involved in cell cycle and transcriptional regulation and also showed differential methylation in promoter regions (Supplementary Data 8 and Figure 4.12). Taken together, our findings extend the previously reported link between *E2F1* activity and methylation-dependent regulation of *HOX* genes in oral cancer to multiple pathways involved in cell cycle control [175].

The role of immune modulation in HNSCC progression is highlighted by cluster 8 enriched for genes related to T cell immunity and showed underexpression in tumour samples while being targeted by overexpressed miRNAs. This dysregulation of T cell immunity genes was to a lesser extent associated with hypermethylation, reflecting the complex interplay between

immune regulation and the tumour microenvironment and underscoring immune evasion as an important feature of HNSCC. In addition, T cell exhaustion is a conserved feature in cancer and has previously been linked to epigenetic regulation of tumour progression [176]. However, because the control samples from the Croatian cohort were derived from tonsil tissue bearing inherent immunological activity, the findings related to cluster 8 should be interpreted with caution.

The results showed that genes involved in tumour vascularisation were DEMiR + DM-DEGs (Figure 4.12, cluster 3), suggesting that both epigenetic drivers contribute to their regulation, with a more pronounced influence from DEMiRs. This is consistent with previous work linking DNAm to vascular and cardiovascular disease directly, or proposing that both epigenetic changes act together, with hypomethylation contributing to activation of the most prominent gene clusters. Downregulation of miRNAs has also been associated with increased expression of angiogenesis-related genes. In our cluster 3, miR-150 was underexpressed and targeted multiple genes, supporting its potential role as a tumour suppressor. In addition, the extracellular matrix is known to play an important role in angiogenesis, and we observed overexpression of several matrix metalloproteinase genes, including *MMP2*, *MMP10*, *MMP11* and *MMP13* that have also been reported in other cancers [177]. Because matrix metalloproteinase genes are common miRNA targets, their deregulation is frequently observed in cancer. Overall, these findings support a role for both epigenetic modalities in HNSCC through regulation of vasculature-related processes, with DEMiRs acting as the dominant epigenetic silencers.

In addition to the findings presented in Mandić et al. [103], in this thesis, we found that GRBs have a distinct CpG distribution compared with the rest of the genome, with depletion in promoter and gene body regions and enrichment in intergenic regions (Figure 4.14). Furthermore, DM patterns generally followed the genome-wide trend: promoters were predominantly hypermethylated, whereas gene body and intergenic regions were predominantly hypomethylated. However, across all three regions, GRBs showed an approximately 10 percentage-point higher proportion of hypermethylated CpGs compared with randomly sampled genomic regions (Figure 4.15). Taken together, these observations indicate that GRBs follow the general HNSCC-associated methylation pattern, but with a more pronounced shift towards hypermethylation,

suggesting that long-range regulatory domains may be particularly susceptible to epigenetic remodelling in this disease.

5.3 Discussion of the Parkinson's disease study

PD is the second disease of special interest in this thesis, as it has a strong genetic component which is yet not fully understood [178]. Furthermore, emerging studies suggest the role of immune system in the progression and development of the disease [114]. Within this work, we identified certain cell types that may be drivers of the immune response in PDs compared to CTRLs. Additionally, a set of non-coding variants were identified from non-coding sequences captured during whole exome sequencing and linked to potential long range regulation.

Single-cell transcriptomic analysis provided a detailed characterisation of the CD4+ T-cell- and DC-enriched fractions isolated from PBMCs of PD and CTRL participants. In addition to the expected CD4+ T-cell and DC populations, several other immune cell types were identified, demonstrating the limitations of the magnetic separation procedure (Figure 4.17d,e). Due to an experimental issue during magnetic separation, one sample contained only the CD4+ T-cell fraction, providing an opportunity to assess the origin of the non-target cell populations. In this sample, cells assigned to the monocyte cluster were detected alongside the CD4+ T-cell fraction (Supplementary Figure S8c), indicating that monocytes were retained within this fraction during separation. In contrast, the absence of NK cells from this sample suggests that the NK cell population observed in the complete dataset was predominantly retained within the DC-enriched fraction.

Single-cell analysis identified four CD4+ T-cell subtypes in the dataset: naïve T-cells, TCMs, TEMs, and Tregs, with their identities supported by the expression of subtype-specific marker genes (Figure 4.19 b). Among these populations, DEGs identified in CD4+ TEMs between PD and CTRL samples showed greater enrichment of immune-related processes compared with the other CD4+ T-cell subtypes, including terms associated with lymphocyte activation and differentiation (Figure 4.22). Previous studies have implicated CD4+ T-cells in autoimmune mechanisms related to PD through their recognition of α -synuclein, with stronger α -synuclein-

specific T-cell responses observed during the early stages of the disease [179]. Although the specific contribution of CD4⁺ TEMs to PD remains poorly understood, their capacity to migrate to sites of inflammation suggests a potential involvement in neuroinflammatory processes [180]. Increased levels of CD4⁺ TEMs have also been associated with cognitive decline, suggesting a broader involvement of this cell population in cognitive ageing [181]. While their role in PD progression remains to be established, the observed enrichment of immune-related processes supports the potential involvement of CD4⁺ TEMs in the peripheral immune response associated with PD and warrants further investigation.

DCs play an important role in CD4⁺ T-cell activation and the subsequent differentiation of effector T-cell populations. In the present single-cell dataset, no significant differences in the proportions of the major DC subtypes were observed between PD and CTRL samples (Figure 4.20 c). Previous studies have reported reduced numbers of circulating DCs in PD, particularly among conventional DCs, with decreased DC abundance also associated with greater motor symptom severity [113]. These observations have led to the suggestion that peripheral DC depletion may reflect their migration to affected regions of the brain. However, a recent single-cell study based on a similarly sized cohort also found no evidence of DC depletion, indicating that this observation remains to be validated in larger PD cohorts [182]. Similarly, no significant differences in the proportions of CD4⁺ T-cells or their identified subtypes were detected between PD and CTRL samples in the present analysis (Figure 4.19 d). This contrasts with previous reports of reduced circulating CD4⁺ T-cell numbers and altered proportions of CD4⁺ T-cell subsets in PD [183–185]. Differences between studies may reflect variation in cohort characteristics and methodological approaches, as well as the greater resolution provided by single-cell analysis.

Three distinct DC subtypes were identified based on the expression of characteristic marker genes: cDC1s, cDC2s, and pDCs (Figure 4.20 a). These populations have distinct roles in coordinating immune responses. cDC1s primarily promote CD8⁺ T-cell activation and Th1 polarisation, whereas cDC2s predominantly interact with CD4⁺ T-cells and contribute to their differentiation into different effector phenotypes. In contrast, pDCs are major producers of type I interferons and have an important role in antiviral immune responses [186–188].

The cDC2 population is increasingly recognised as heterogeneous, with several studies identifying distinct subpopulations within this group [163, 188]. Villani et al. [163] initially described two cDC2 subsets, DC2 and DC3, characterised by distinct transcriptional profiles, with DC3s displaying features overlapping those of both monocytes and cDC2s [187]. Both populations are involved in CD4⁺ naïve T-cell polarisation and inflammatory responses, while the dual phenotype of DC3s distinguishes them from other conventional DC populations [163, 188–190].

Within the cDC2 cluster, however, a small population expressing *CD163* and *CD14* was observed and likely represents the DC3 subtype (Figure 4.20 a and b). The expression of these markers is consistent with the previously described dual cDC2-monocyte phenotype of DC3s. Although DC3s did not form a distinct cluster, potentially due to limitations of the clustering approach, their characteristic marker expression suggests that this population was captured within the broader cDC2 cluster. However, an orthogonal approach to SC-seq is required to further confirm the identity of these cells. Nevertheless, this observation illustrates the importance of refined cell classification for resolving immune cell heterogeneity and the utility of SC-seq for detecting rare immune populations. As DCs comprise only approximately 1–2% of PBMCs [191], enrichment of the DC fraction increased the representation of these cells and facilitated the detection of the rare population presumed to represent DC3s.

DEG analysis highlighted cDC2s and CD4⁺ TEMs as cell populations exhibiting prominent immune-related transcriptional changes in PD (Figure 4.22). The DEG profiles of both populations were enriched for immune-related processes, with gene set enrichment analysis identifying shared terms such as “T-cell activation” and “positive regulation of immune response”, suggesting a potential functional interplay between these cell types. As antigen presenting cells, cDC2s play an important role in CD4⁺ T-cell activation and polarisation and produce T-cell-attracting chemokines, including *CCL5* and *CCL22*, as well as pro-inflammatory cytokines that promote T-cell activation [187, 188]. Although cDC2s have traditionally been associated with Th2 and Th17 polarisation, they can also contribute to Th1 differentiation depending on the DC subset involved [156]. In particular, cDC2s have been shown to promote CD4⁺ TEM responses through cytokines such as IL-1 β and IL-12p70 [190]. The presence of CD4⁺ helper T-cells in the sub-

stantia nigra of patients with PD further supports their potential involvement in PD pathogenesis [192]. Upon reactivation, CD4⁺ TEMs can produce both pro- and anti-inflammatory cytokines and may therefore contribute to the sustained inflammatory environment associated with PD [193–195].

We examined the transcriptional profiles of CD4⁺ TEMs and cDC2s in the context of their potential coordinated involvement in the immune response associated with PD. Among the top 30 DEGs contributing to the selected pathways in each cell type, several genes showed pronounced differences in expression between PD and CTRL samples (Figure 4.22 c and d). In CD4⁺ TEMs, *BCL6*, *ICOS*, and *SOCS3* were among the prominently upregulated genes, whereas cDC2s showed increased expression of *LCK*, *CD3D*, and *CD247*. In contrast, *PTGS2* and *SOCS3* were downregulated in cDC2s.

While the involvement of *ICOS*, *CD3D*, and *CD247* in PD remains relatively unexplored, *BCL6*, a transcriptional repressor involved in the regulation of immune and inflammatory processes, has previously been associated with neurodegenerative diseases [196]. Its increased expression in the present dataset is also consistent with previous observations in PD [197]. *LCK* has similarly been implicated in PD-associated immune processes, including the promotion of Th17 cell differentiation [198, 199]. In addition, *SOCS3* has been associated with neuroinflammation and PD pathology [200, 201]. Interestingly, *SOCS3* showed opposing expression patterns between the two cell populations, being upregulated in CD4⁺ TEMs and downregulated in cDC2s, suggesting that its regulation may be dependent on the cellular context.

Interestingly, the majority of CD4⁺ T-cells expressed *CAMK4* and *BCL11B* (Figure S8e), which are not commonly used as markers of this cell population. However, *BCL11B* is a crucial commitment factor during T-cell development from haematopoietic precursors, where it contributes to the repression of alternative lineage programmes following initial T-cell specification by other transcription factors [202]. At the same time, *BCL11B* participates in a network of interacting transcription factors involved in neuronal development and cortical layer specification [70, 203]. This pleiotropic role of *BCL11B* illustrates an important characteristic of GRB target genes: their expression across different developmental and cellular contexts requires precise spatio-temporal regulation mediated by extensive regulatory landscapes. A prominent example

is an enhancer located in a gene desert approximately 850 kb downstream of *BCL11B*, which was shown to interact with *BCL11B*, but not with other nearby genes. Functional reporter assays further demonstrated that this enhancer directs *BCL11B* expression in T-cells, but not in non-T-lineage haematopoietic cells [164].

Additionally, *BCL11B* was significantly upregulated in PD compared with CTRL samples, although the magnitude of the expression difference was small (Table 4.8). Several other GRB target genes were also differentially expressed in PD (Supplementary Table S6), suggesting that altered expression of genes under extensive long-range regulatory control may be relevant in this cellular context. Further investigation of these genes and their regulatory landscapes in specific immune cell populations could therefore provide additional insight into the contribution of long-range regulation to immune dysregulation in PD.

This study provides insight into specific immune cell subtypes associated with PD, however, several limitations should be considered. Although enrichment facilitated the detection of rare cell populations, analysis of selected immune fractions does not capture the complete peripheral immune profile and limits the investigation of interactions between targeted and non-targeted cell populations. Complementary analysis of the total PBMC population could therefore provide a broader representation of immune alterations associated with PD.

Magnetic separation also resulted in incomplete sample purity, as demonstrated by the presence of several non-target immune cell populations within the CD4⁺ T-cell- and DC-enriched fractions. Future studies could employ fluorescence-activated cell sorting to obtain populations of greater purity. Nevertheless, the use of SC-seq allowed individual cell populations to be distinguished and analysed separately despite these limitations, avoiding the averaging of gene expression across mixed cell populations inherent to bulk RNA sequencing.

The annotation of CD4⁺ T-cell subtypes was additionally limited by the specificity of the marker genes available in the reference dataset, resulting in several CD4⁺ T-cell populations clustering together. Inclusion of more subtype-specific markers, including the transcription factors *TBX21*, *GATA3*, and *RORC*, or cytokines such as *IFNG*, *IL4*, and *IL17A*, could improve the resolution of CD4⁺ T-cell heterogeneity [204]. Finally, although the cohort size was comparable to other single-cell studies in the field, larger cohorts would increase statistical

power, facilitate the detection of rare cell populations, and provide a better representation of inter-individual variability [120, 157, 205].

Additional limitations relate to the availability of detailed pharmacological and clinical information, both not always systematically recorded in routine clinical practice. Although all patients received comparable anti-Parkinsonian treatment and were screened to exclude recent infections, vaccinations, or biological therapies, subtle effects of unreported medications on peripheral immune profiles cannot be excluded. Nevertheless, the relatively consistent clinical background and careful inclusion criteria likely reduced the influence of these factors. Future studies incorporating detailed pharmacological records, standardised cognitive assessments, and α -synuclein biomarkers could help identify additional associations and provide a more comprehensive understanding of peripheral immune alterations in PD.

Using exome sequencing data, we extended the analysis beyond coding variants and identified additional non-coding variants which may aid interpretation of PD. These non-coding variants overlap GWAS associated loci, and for some traits, these loci were more likely to fall into GRBs (Figure 4.23c). This result suggest a potential long-range involvement of these non-coding variant, demonstrating that targPred may be utilised in this case for generating additional hypothesis.

6 Conclusion

The work presented in this thesis was fully incepted and funded by the Croatian Science Foundation project “Exploring Interactions Between Regulatory Variants in Human Disease Context (IntRegVar)”, with main objective to improve prediction of enhancer-promoter activity associations, achieved by development of the targPred method and accompanying web-based resource.

6.1 Development of targPred

Starting from a proof of concept on a single disease phenotype and a basic association score presented in Barešić et al. [70], in this work we developed the targPred method by combining the genomic framework of GRBs with enhancer–promoter associations derived from transcriptional activity and applied it systematically to the GWAS Catalog. In terms of methodology, original association calculation was improved to both enhancing and repressive scenarios, and a web-based resource was created enabling browsing the results and querying in a user-friendly way, aimed at users with little bioinformatic experience.

Using this framework, we identified disease-associated variants with a potential role in long-range regulation and linked them to distal target genes that may be overlooked by conventional nearest-gene assignment. Application of targPred across a broad range of phenotypic traits revealed that certain traits are enriched with long-range regulation. In particular, a broad range of psychiatric disorders showed a high propensity for long-range regulation, consistent with previous findings that genes involved in neuronal development and function are frequently under extreme long-range regulatory control. However, this pattern was not restricted to neuronal phenotypes, we found evidence for long-range regulation also across immune, metabolic,

cardiovascular and cancer-related traits. These findings demonstrate that consideration of the wider regulatory landscape can provide an additional layer of biological interpretation for non-coding GWAS associations.

6.2 Applications of targPred on examples with clinical relevance

The utility of the GRB framework was first explored in HNSCC using a multi-omics approach. We found that the density and distribution of CpGs differ significantly between GRBs and non-GRB regions and that aberrant DNA methylation in cancer is associated with altered expression of GRB target genes. Since GRB target genes frequently encode developmental and differentiation-related transcription factors, disruption of their tightly controlled regulatory network may have important consequences for tumour development and progression. These results demonstrate that the GRB concept and the regulatory architecture surrounding GRB target genes are also relevant for interpreting epigenetic alterations in disease. Furthermore, investigating the epigenetic patterns of enhancers involved in long-range regulation represents a potential avenue for future research. Extending this approach to other cancer types may provide further insight into the contribution of disrupted long-range regulation to their molecular aetiology.

Additionally, the role of GRB target genes was investigated in immune cells of patients with PD. Multiple GRB target genes were differentially expressed in immune cell populations in PD patients, indicating that genes under extensive long-range regulatory control may also contribute to the altered peripheral immune landscape associated with the disease. Of particular interest was *BCL11B*, which was differentially expressed in CD4⁺ T-cells and is known to be regulated by enhancers located at extreme genomic distances from its promoter. *BCL11B* is an important regulator of T-cell development, while also having essential functions in neuronal development. The single-cell approach enabled the expression of GRB target genes to be examined in specific immune cell populations, providing a more detailed view of their potential involvement in PD-associated immune dysregulation. In the future, integration with single-cell chromatin

accessibility data, such as scATAC-seq, could complement these findings by characterising the regulatory landscape of GRBs in individual cell types and providing further insight into the regulatory elements controlling GRB target gene expression.

6.3 Future prospects

With the targPred resource built and its utility demonstrated through aforementioned examples, several directions of further work are planned.

From the methodological side, the association method, while improved and producing promising results as demonstrated in Mandić et al. [73], relies on several other computational methods, each with own pitfalls and compromises. The GRBs used here are based on human-dog comparison, and while these were considered to be of high quality, other pairwise comparisons to human could be considered, or a combined method using several comparisons at the same time. Methods for the definition of genome-wide regulatory elements are still under active development, and while CAGE-based enhancerome [22] is considered to be technically robust and biologically relevant, due to the wide complexity of the dataset, sets created from other genomic experiments could be used (ATAC-seq, ChIP-seq, various chromatin capture approaches, etc.). It is worth noting here that various approaches of AI-based predictions of regulatory sequences are under active development by the author's group and elsewhere, and this field is expected to contribute some interesting predictors utilising various datasets in the near future.

The genomic prediction methods can be as good as the data provided in the process of learning. Indeed, at the inception stage of targPred in 2022, it was built on the then available version of the GWAS Catalog. All the results presented here originate from the recent update of the tool, recalculated on the February 2026 data freeze. The latest data doubled in amount compared to the 2022 version, and this growing rate is only expected to accelerate [51, 52]. With that in mind, data maintenance is crucial to ensure that targPred keeps up with the biological complexity available for analysis, and an automation of data acquisition from the GWAS Catalog is planned, to facilitate regular updates. However, it is worth noting here that the true utility of this tool will not be possible unless a functionality to process individual variant files is

implemented, as showcased in the PD example. To that end, targPred is built in mind to have an on-the-fly enhancer-promoter association calculation, for any variant previously unreported in common GWAS sets.

Finally, as previously stated, targPred is built for use by the clinical genomics community, as well as research purposes. Due to this intended breadth of use, the tool needs to be presented to a broad audience and motivated to provide feedback on intended use and issues. An example of such successful collaboration with clinician that spans beyond this thesis can be found in our recent work [206] driving expansion of targPred in visualisations, documentation, and access options. Several future directions of improvement are planned, including exploration of an AI agent to aid in accessing the associations for lay users.

Taken together, this thesis provides several lines of evidence supporting the use of GRBs and their target genes as a framework for interpreting gene regulation in human disease. Across genetic variation, cancer-associated epigenetic alterations and disease-related transcriptional changes, the results consistently demonstrate the importance of considering regulatory relationships beyond genomic proximity. Importantly, the GRB framework can be integrated with already available genomic, epigenomic and transcriptomic datasets without the need for additional experimental data, while providing biologically informed hypotheses about enhancer–promoter relationships and distal target genes.

References

- [1] J. Craig Venter et al. “The sequence of the human genome”. In: *Science* 291.5507 (2001), pp. 1304–1351. DOI: 10.1126/SCIENCE.1058040.
- [2] Eric S. Lander et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (2001), pp. 860–921. DOI: 10.1038/35057062.
- [3] Steven L. Salzberg. “Open questions: How many genes do we have?” In: *BMC Biology* 16.1 (2018), p. 94. DOI: 10.1186/S12915-018-0564-X.
- [4] P. Carninci et al. “Molecular biology: The transcriptional landscape of the mammalian genome”. In: *Science* 309.5740 (2005), pp. 1559–1563. DOI: 10.1126/SCIENCE.1112014/SUPPL\ _FILE/P1-2CAGE\ _MOUSE\ _ACCESSION.LIST.ZIP.
- [5] Tyler A. Elliott and T. Ryan Gregory. “What’s in a genome? The C-value enigma and the evolution of eukaryotic genome content”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (2015). DOI: 10.1098/RSTB.2014.0331/30148.
- [6] Jill E. Moore et al. “An expanded registry of candidate cis-regulatory elements”. In: *Nature* 2026 6 (2026), pp. 1–10. DOI: 10.1038/s41586-025-09909-9.
- [7] Mihaela Pertea and Steven L. Salzberg. “Between a chicken and a grape: estimating the number of human genes”. In: *Genome biology* 11.5 (2010). DOI: 10.1186/GB-2010-11-5-206.
- [8] Glenn A Maston, Sara K Evans, and Michael R Green. “Transcriptional Regulatory Elements in the Human Genome”. In: (2006). DOI: 10.1146/annurev.genom.7.080505.115623.

- [9] Tong Ihn Lee and Richard A. Young. “Transcriptional Regulation and its Misregulation in Disease”. In: *Cell* 152.6 (2013), p. 1237. DOI: 10.1016/J.CELL.2013.02.014.
- [10] Stephen T. Smale and James T. Kadonaga. “The RNA polymerase II core promoter”. In: *Annual review of biochemistry* 72 (2003), pp. 449–479. DOI: 10.1146/ANNUREV.BIOCHEM.72.121801.161520.
- [11] Boris Lenhard, Albin Sandelin, and Piero Carninci. “Metazoan promoters: emerging characteristics and insights into transcriptional regulation”. In: *Nature reviews. Genetics* 13.4 (2012), pp. 233–245. DOI: 10.1038/NRG3163.
- [12] Alistair R.R. Forrest et al. “A promoter-level mammalian expression atlas”. In: *Nature* 507.7493 (2014), pp. 462–470. DOI: 10.1038/NATURE13182.
- [13] Matteo Maurizio Guerrini, Akiko Oguchi, Akari Suzuki, and Yasuhiro Murakawa. “Cap analysis of gene expression (CAGE) and noncoding regulatory elements”. In: *Seminars in Immunopathology* 44.1 (2022), pp. 127–136. DOI: 10.1007/S00281-021-00886-5.
- [14] Julian Banerji, Laura Olson, and Walter Schaffner. “A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes”. In: *Cell* 33.3 (1983), pp. 729–740. DOI: 10.1016/0092-8674(83)90015-6.
- [15] P. Moreau, R. Hen, B. Wasylyk, R. Everett, M. P. Gaub, and P. Chambon. “The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants”. In: *Nucleic acids research* 9.22 (1981), pp. 6047–6068. DOI: 10.1093/NAR/9.22.6047.
- [16] Julian Banerji, Sandro Rusconi, and Walter Schaffner. “Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences”. In: *Cell* 27.2 Pt 1 (1981), pp. 299–308. DOI: 10.1016/0092-8674(81)90413-X.
- [17] Dustin E. Schones et al. “Dynamic Regulation of Nucleosome Positioning in the Human Genome”. In: *Cell* 132.5 (2008), pp. 887–898. DOI: 10.1016/j.cell.2008.02.022.
- [18] Bradley E. Bernstein et al. “Genomic maps and comparative analysis of histone modifications in human and mouse”. In: *Cell* 120.2 (2005), pp. 169–181. DOI: 10.1016/j.cell.2005.01.001.

- [19] A. Gordon Robertson et al. “Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding”. In: *Genome research* 18.12 (2008), pp. 1906–1917. DOI: 10.1101/GR.078519.108.
- [20] Nathaniel D. Heintzman et al. “Histone modifications at human enhancers reflect global cell-type-specific gene expression”. In: *Nature* 459.7243 (2009), pp. 108–112. DOI: 10.1038/NATURE07829.
- [21] Robin Andersson, Albin Sandelin, and Charles G. Danko. “A unified architecture of transcriptional regulatory elements”. In: *Trends in genetics : TIG* 31.8 (2015), pp. 426–433. DOI: 10.1016/J.TIG.2015.05.007.
- [22] Robin Andersson et al. “An atlas of active enhancers across human cell types and tissues”. In: *Nature* 507.7493 (2014), pp. 455–461. DOI: 10.1038/nature12787.
- [23] Robin Andersson and Albin Sandelin. “Determinants of enhancer and promoter activities of regulatory elements”. In: *Nature Reviews Genetics* 2019 21:2 21.2 (2019), pp. 71–87. DOI: 10.1038/s41576-019-0173-8.
- [24] Kenneth S. Zaret and Jason S. Carroll. “Pioneer transcription factors: establishing competence for gene expression”. In: *Genes & Development* 25.21 (2011), pp. 2227–2241. DOI: 10.1101/GAD.176826.111.
- [25] Gregory E. Crawford et al. “Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)”. In: *Genome Research* 16.1 (2006), pp. 123–131. DOI: 10.1101/GR.4074106.
- [26] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature Methods* 2013 10:12 10.12 (2013), pp. 1213–1218. DOI: 10.1038/nmeth.2688.
- [27] Laura A. Lettice et al. “Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.11 (2002), pp. 7548–7553. DOI: 10.1073/PNAS.112212199.

- [28] Adrian L. Sanborn et al. “Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.47 (2015), E6456–E6465. DOI: 10.1073/PNAS.1518552112/SUPPL\FILE/PNAS.1518552112.ST01.XLSX.
- [29] Gilad Barshad and Charles G. Danko. “Revisiting models of enhancer–promoter communication in gene regulation”. In: *Genome Research* 35.6 (2025), p. 1277. DOI: 10.1101/GR.278389.123.
- [30] Erez Lieberman-Aiden et al. “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”. In: *Science* 326.5950 (2009), pp. 289–293. DOI: 10.1126/SCIENCE.1181369.
- [31] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. “Capturing chromosome conformation”. In: *Science* 295.5558 (2002), pp. 1306–1311. DOI: 10.1126/SCIENCE.1067799/SUPPL\FILE/INDEX.HTM.
- [32] Jesse R. Dixon et al. “Topological domains in mammalian genomes identified by analysis of chromatin interactions”. In: *Nature* 2012 485:7398 485.7398 (2012), pp. 376–380. DOI: 10.1038/nature11082.
- [33] Stefan Schoenfelder and Peter Fraser. “Long-range enhancer–promoter contacts in gene expression control”. In: *Nature Reviews Genetics* 2019 20:8 20.8 (2019), pp. 437–455. DOI: 10.1038/s41576-019-0128-0.
- [34] Elizabeth Ing-Simmons et al. “Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin”. In: *Genome Research* 25.4 (2015), pp. 504–513. DOI: 10.1101/GR.184986.114.
- [35] Michael I. Robson, Alessa R. Ringel, and Stefan Mundlos. “Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D”. In: *Molecular Cell* 74.6 (2019), pp. 1110–1122. DOI: 10.1016/J.MOLCEL.2019.05.032/ASSET/00160DC1-D34D-43EE-8D6E-66B46634D54C/MAIN.ASSETS/GR6.JPG.

- [36] Adam Woolfe et al. “Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development”. In: *PLoS Biology* 3.1 (2004), e7. DOI: 10.1371/JOURNAL.PBIO.0030007.
- [37] Dimitris Polychronopoulos, James W.D. King, Alexander J. Nash, Ge Tan, and Boris Lenhard. *Conserved non-coding elements: Developmental gene regulation meets genome organization*. 2017. DOI: 10.1093/nar/gkx1074.
- [38] Albin Sandelin et al. “Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes”. In: *BMC Genomics* 5 (2004), p. 99. DOI: 10.1186/1471-2164-5-99.
- [39] Gill Bejerano et al. “Ultraconserved elements in the human genome”. In: *Science* 304.5675 (2004), pp. 1321–1325. DOI: 10.1126/SCIENCE.1098119.
- [40] Ge Tan and Boris Lenhard. “Computational Genomics of Regulatory Elements and Regulatory Territories”. In: November (2017).
- [41] Alexander J. Nash and Boris Lenhard. “A novel measure of non-coding genome conservation identifies genomic regulatory blocks within primates”. In: *Bioinformatics (Oxford, England)* 35.14 (2019), pp. 2354–2361. DOI: 10.1093/BIOINFORMATICS/BTY1014.
- [42] Pär G. Engström, Shannan J. Ho Sui, Øyvind Drivenes, Thomas S. Becker, and Boris Lenhard. “Genomic regulatory blocks underlie extensive microsynteny conservation in insects”. In: *Genome Research* 17.12 (2007), pp. 1898–1908. DOI: 10.1101/GR.6669607.
- [43] Hiroshi Kikuta et al. “Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates”. In: *Genome Research* 17.5 (2007), pp. 545–555. DOI: 10.1101/GR.6086307.
- [44] Nathan Harmston, Elizabeth Ing-Simmons, Ge Tan, Malcolm Perry, Matthias Merken-schlager, and Boris Lenhard. “Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation”. In: *Nature Communications* 8.1 (2017). DOI: 10.1038/S41467-017-00524-5.

- [45] Altuna Akalin et al. “Transcriptional features of genomic regulatory blocks”. In: *Genome biology* 10.4 (2009). DOI: 10.1186/GB-2009-10-4-R38.
- [46] Pavla Navratilova and Thomas S. Becker. “Genomic regulatory blocks in vertebrates and implications in human disease”. In: *Briefings in Functional Genomics & Proteomics* 8.4 (2009), pp. 333–342. DOI: 10.1093/BFGP/ELP019.
- [47] Anja Ragvin et al. “Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.2 (2010), pp. 775–780. DOI: 10.1073/PNAS.0911591107.
- [48] Eric Vallabh Minikel et al. “Quantifying prion disease penetrance using large population control cohorts”. In: *Science Translational Medicine* 8.322 (2016). DOI: 10.1126/SCITRANSLMED.AAD5169/SUPPL_FILE/8-322RA9_SM.PDF.
- [49] Aravinda Chakravarti. “Magnitude of Mendelian versus complex inheritance of rare disorders”. In: July (2021), pp. 1–7. DOI: 10.1002/ajmg.a.62463.
- [50] Melina Claussnitzer et al. “A brief history of human disease genetics”. In: *Nature* 2020 577:7789–7789 (2020), pp. 179–189. DOI: 10.1038/s41586-019-1879-7.
- [51] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 2021 1:1 1.1 (2021), pp. 1–21. DOI: 10.1038/s43586-021-00056-9.
- [52] Maria Cerezo et al. “The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity”. In: *Nucleic Acids Research* 53.D1 (2025), pp. D998–D1005. DOI: 10.1093/NAR/GKAE1070.
- [53] Abdel Abdellaoui, Loic Yengo, Karin J.H. Verweij, and Peter M. Visscher. “15 years of GWAS discovery: Realizing the promise”. In: *American Journal of Human Genetics* 110.2 (2023), p. 179. DOI: 10.1016/J.AJHG.2022.12.011.
- [54] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 2018 562:7726–7726 (2018), pp. 203–209. DOI: 10.1038/s41586-018-0579-z.

- [55] Mitja I. Kurki et al. “FinnGen provides genetic insights from a well-phenotyped isolated population”. In: *Nature* 2023 613:7944 613.7944 (2023), pp. 508–518. DOI: 10.1038/s41586-022-05473-8.
- [56] Michael F. Wangler, Yanhui Hu, and Joshua M. Shulman. “Drosophila and genome-wide association studies: a review and resource for the functional dissection of human complex traits”. In: *Disease Models & Mechanisms* 10.2 (2017), p. 77. DOI: 10.1242/DMM.027680.
- [57] Montgomery Slatkin. “Linkage disequilibrium — understanding the evolutionary past and mapping the medical future”. In: *Nature Reviews Genetics* 2008 9:6 9.6 (2008), pp. 477–485. DOI: 10.1038/nrg2361.
- [58] Kyoko Watanabe et al. “A global overview of pleiotropy and genetic architecture in complex traits”. In: *Nature Genetics* 2019 51:9 51.9 (2019), pp. 1339–1348. DOI: 10.1038/s41588-019-0481-0.
- [59] Matthew T. Maurano et al. “Systematic localization of common disease-associated variation in regulatory DNA”. In: *Science* 337.6099 (2012), pp. 1190–1195. DOI: 10.1126/SCIENCE.1222794.
- [60] Ellora Hui Zhen Chua, Samen Yasar, and Nathan Harmston. “The importance of considering regulatory domains in genome-wide analyses - the nearest gene is often wrong!” In: *Biology open* 11.4 (2022). DOI: 10.1242/BI0.059091.
- [61] Ella K. Samer et al. “Deciphering the impact of genomic variation on function”. In: *Nature* 633.8028 (2024), pp. 47–57. DOI: 10.1038/s41586-024-07510-0.
- [62] Charles P. Fulco et al. “Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations”. In: *Nature Genetics* 51.12 (2019), pp. 1664–1669. DOI: 10.1038/S41588-019-0538-0.
- [63] Joseph Nasser et al. “Genome-wide enhancer maps link risk variants to disease genes”. In: *Nature* 593.7858 (2021), pp. 238–243. DOI: 10.1038/S41586-021-03446-X.

- [64] Simon Fishilevich et al. “GeneHancer: Genome-wide integration of enhancers and target genes in GeneCards”. In: *Database* 2017 (2017), pp. 1–17. DOI: 10.1093/database/bax028.
- [65] Edward Mountjoy et al. “An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci”. In: *Nature genetics* 53.11 (2021), p. 1527. DOI: 10.1038/S41588-021-00945-5.
- [66] Steven Gazal et al. “Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity”. In: *Nature Genetics* 54.6 (2022), pp. 827–836. DOI: 10.1038/S41588-022-01087-Y.
- [67] Žiga Avsec et al. “Effective gene expression prediction from sequence by integrating long-range interactions”. In: *Nature Methods* 18.10 (2021), pp. 1196–1203. DOI: 10.1038/s41592-021-01252-x.
- [68] Hugo Dalla-Torre et al. “Nucleotide Transformer: building and evaluating robust foundation models for human genomics”. In: *Nature Methods* 2024 22:2 22.2 (2024), pp. 287–297. DOI: 10.1038/s41592-024-02523-z.
- [69] Žiga Avsec et al. “Advancing regulatory variant effect prediction with AlphaGenome”. In: *Nature* 2026 649:8099 649.8099 (2026), pp. 1206–1218. DOI: 10.1038/s41586-025-10014-0.
- [70] Anja Barešić, Alexander Jolyon Nash, Tarik Dahoun, Oliver Howes, and Boris Lenhard. “Understanding the genetics of neuropsychiatric disorders: the potential role of genomic regulatory blocks”. In: *Molecular Psychiatry* 25.1 (2020), pp. 6–18. DOI: 10.1038/s41380-019-0518-x.
- [71] Scott Smemo et al. “Obesity-associated variants within FTO form long-range functional connections with IRX3”. In: *Nature* 507.7492 (2014), p. 371. DOI: 10.1038/NATURE13138.
- [72] Melina Claussnitzer et al. “FTO Obesity Variant Circuitry and Adipocyte Browning in Humans”. In: *The New England journal of medicine* 373.10 (2015), p. 895. DOI: 10.1056/NEJMOA1502214.

- [73] Katarina Mandić, Dalibor Hršak, Filip Uljanić, Boris Lenhard, and Anja Barešić. “The role of long-range transcriptional regulation in interpretation of non-coding variants associated with human disease”. In: *bioRxiv : the preprint server for biology* (2026). DOI: 10.64898/2026.06.15.731051.
- [74] Daniel E. Johnson, Barbara Burtneš, C. René Leemans, Vivian Wai Yan Lui, Julie E. Bauman, and Jennifer R. Grandis. *Head and neck squamous cell carcinoma*. 2020. DOI: 10.1038/s41572-020-00224-3.
- [75] Madison Canning et al. *Heterogeneity of the head and neck squamous cell carcinoma immune landscape and its impact on immunotherapy*. 2019. DOI: 10.3389/fcell.2019.00052.
- [76] Wusheng Yan, Ignacio I Wistuba, Michael R Emmert-Buck, and Heidi S Erickson. “Squamous cell carcinoma – similarities and differences among anatomical sites”. In: *American Journal of Cancer Research* 1.3 (2011), p. 275.
- [77] Kazuhiro Kobayashi, Kenji Hisamatsu, Natsuko Suzui, Akira Hara, Hiroyuki Tomita, and Tatsuhiko Miyazaki. “A Review of HPV-Related Head and Neck Cancer”. In: *Journal of Clinical Medicine* 2018, Vol. 7, Page 241 7.9 (2018), p. 241. DOI: 10.3390/JCM7090241.
- [78] Andrew W. Joseph and Gypsyamber D’Souza. “Epidemiology of Human Papillomavirus-Related Head and Neck Cancer”. In: *Otolaryngologic Clinics of North America* 45.4 (2012), pp. 739–764. DOI: 10.1016/J.OTC.2012.04.003.
- [79] Maura L. Gillison, Anil K. Chaturvedi, William F. Anderson, and Carole Fakhry. “Epidemiology of Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma”. In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33.29 (2015), pp. 3235–3242. DOI: 10.1200/JCO.2015.61.6995.
- [80] Shanthi Marur, Gypsyamber D’Souza, William H. Westra, and Arlene A. Forastiere. “HPV-associated head and neck cancer: a virus-related cancer epidemic”. In: *The Lancet. Oncology* 11.8 (2010), pp. 781–789. DOI: 10.1016/S1470-2045(10)70017-6.

- [81] Shanthi Marur and Arlene A. Forastiere. “Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment”. In: *Mayo Clinic Proceedings* 91.3 (2016), pp. 386–396. DOI: 10.1016/J.MAYOCP.2015.12.017.
- [82] B. Kofler, S. Laban, C. J. Busch, B. Lörincz, and R. Knecht. “New treatment strategies for HPV-positive head and neck cancer”. In: *European Archives of Oto-Rhino-Laryngology* 271.7 (2014), pp. 1861–1867. DOI: 10.1007/S00405-013-2603-0/FIGURES/4.
- [83] Francesco Perri et al. “Management of HPV-Related Squamous Cell Carcinoma of the Head and Neck: Pitfalls and Caveat”. In: *Cancers* 2020, Vol. 12, Page 975 12.4 (2020), p. 975. DOI: 10.3390/CANCERS12040975.
- [84] Yingming Sun, Zhe Wang, Sufang Qiu, and Ruoyu Wang. “Therapeutic strategies of different hpv status in head and neck squamous cell carcinoma”. In: *International Journal of Biological Sciences* 17.4 (2021), pp. 1104–1118. DOI: 10.7150/IJBS.58077.
- [85] Ksenija Božinović et al. “Genome-wide miRNA profiling reinforces the importance of miR-9 in human papillomavirus associated oral and oropharyngeal head and neck cancer”. In: *Scientific Reports* 9.1 (2019), pp. 1–13. DOI: 10.1038/s41598-019-38797-z.
- [86] Nina Milutin Gašperov et al. “DNA Methylome Distinguishes Head and Neck Cancer from Potentially Malignant Oral Lesions and Healthy Oral Mucosa”. In: *International Journal of Molecular Sciences* 2020, Vol. 21, Page 6853 21.18 (2020), p. 6853. DOI: 10.3390/IJMS21186853.
- [87] Kimberly E. Stephens, Christine A. Miaskowski, Jon D. Levine, Clive R. Pullinger, and Bradley E. Aouizerat. “Epigenetic Regulation and Measurement of Epigenetic Changes”. In: *Biological research for nursing* 15.4 (2012), p. 373. DOI: 10.1177/1099800412444785.
- [88] Yutaka Kondo. “Epigenetic Cross-Talk between DNA Methylation and Histone Modifications in Human Cancers”. In: *Yonsei Medical Journal* 50.4 (2009), p. 455. DOI: 10.3349/YMJ.2009.50.4.455.

- [89] Guoqiang Zhang and Sriharsa Pradhan. “Mammalian epigenetic mechanisms”. In: *IUBMB life* 66.4 (2014), pp. 240–256. DOI: 10.1002/IUB.1264.
- [90] Irina V. Bure, Marina V. Nemtsova, and Ekaterina B. Kuznetsova. “Histone Modifications and Non-Coding RNAs: Mutual Epigenetic Regulation and Role in Pathogenesis”. In: *International Journal of Molecular Sciences* 23.10 (2022), p. 5801. DOI: 10.3390/IJMS23105801.
- [91] Diane E. Handy, Rita Castro, and Joseph Loscalzo. “Epigenetic Modifications: Basic Mechanisms and Role in Cardiovascular Disease”. In: *Circulation* 123.19 (2011), p. 2145. DOI: 10.1161/CIRCULATIONAHA.110.956839.
- [92] Yan Lin Wu et al. “Epigenetic regulation in metabolic diseases: mechanisms and advances in clinical study”. In: *Signal Transduction and Targeted Therapy* 2023 8:1 8.1 (2023), pp. 98–. DOI: 10.1038/s41392-023-01333-7.
- [93] Ksenia Skvortsova, Nicola Iovino, and Ozren Bogdanović. “Functions and mechanisms of epigenetic inheritance in animals”. In: *Nature Reviews Molecular Cell Biology* 2018 19:12 19.12 (2018), pp. 774–790. DOI: 10.1038/s41580-018-0074-2.
- [94] Marcos Francisco Perez and Ben Lehner. “Intergenerational and transgenerational epigenetic inheritance in animals”. In: *Nature Cell Biology* 2019 21:2 21.2 (2019), pp. 143–151. DOI: 10.1038/s41556-018-0242-9.
- [95] Lisa D. Moore, Thuc Le, and Guoping Fan. “DNA Methylation and Its Basic Function”. In: *Neuropsychopharmacology* 2013 38:1 38.1 (2012), pp. 23–38. DOI: 10.1038/npp.2012.112.
- [96] Mirang Kim and Joseph Costello. “DNA methylation: an epigenetic mark of cellular memory”. In: *Experimental & Molecular Medicine* 2017 49:4 49.4 (2017), e322–e322. DOI: 10.1038/emm.2017.10.
- [97] Melanie Ehrlich. *DNA hypomethylation in cancer cells*. 2009. DOI: 10.2217/EPI.09.33.

- [98] Jadwiga Gazdzicka, Karolina Golabek, Joanna Katarzyna Strzelczyk, and Zofia Ostrowska. “Epigenetic Modifications in Head and Neck Cancer”. In: *Biochemical genetics* 58.2 (2020), pp. 213–244. DOI: 10.1007/S10528-019-09941-1.
- [99] Marilena V. Iorio and Carlo M. Croce. “MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review”. In: *EMBO Molecular Medicine* 4.3 (2012), p. 143. DOI: 10.1002/EMMM.201100209.
- [100] Yong Peng and Carlo M. Croce. “The role of MicroRNAs in human cancer”. In: *Signal Transduction and Targeted Therapy* 2016 1:1 1.1 (2016), pp. 15004–. DOI: 10.1038/sigtrans.2015.4.
- [101] Veronica Davalos and Manel Esteller. “MicroRNAs and cancer epigenetics: A macrorevolution”. In: *Current Opinion in Oncology* 22.1 (2010), pp. 35–45. DOI: 10.1097/CCO.0B013E328333DCBB.
- [102] Michael S. Lawrence et al. “Comprehensive genomic characterization of head and neck squamous cell carcinomas”. In: *Nature* 517.7536 (2015), pp. 576–582. DOI: 10.1038/nature14129.
- [103] Katarina Mandić et al. “Integrative analysis in head and neck cancer reveals distinct role of miRNome and methylome as tumour epigenetic drivers”. In: *Scientific Reports* 2024 14:1 14.1 (2024), pp. 1–12. DOI: 10.1038/s41598-024-59312-z.
- [104] Masashi Shiiba, Katsuhiko Uzawa, and Hideki Tanzawa. “MicroRNAs in Head and Neck Squamous Cell Carcinoma (HNSCC) and Oral Squamous Cell Carcinoma (OSCC)”. In: *Cancers* 2.2 (2010), pp. 653–669. DOI: 10.3390/CANCERS2020653.
- [105] Joshua Lubov et al. “Meta-analysis of microRNAs expression in head and neck cancer: uncovering association with outcome and mechanisms”. In: *Oncotarget* 8.33 (2017), pp. 55511–55524. DOI: 10.18632/ONCOTARGET.19224.
- [106] Alice Ramassone, Sara Pagotto, Angelo Veronese, and Rosa Visone. “Epigenetics and MicroRNAs in Cancer”. In: *International Journal of Molecular Sciences* 2018, Vol. 19, Page 459 19.2 (2018), p. 459. DOI: 10.3390/IJMS19020459.

- [107] María J. Pajares, Ester Alemany-cosme, Saioa Goñi, Eva Bandres, Cora Palanca-Ballester, and Juan Sandoval. “Epigenetic Regulation of microRNAs in Cancer: Shortening the Distance from Bench to Bedside”. In: *International Journal of Molecular Sciences* 22.14 (2021), p. 7350. DOI: 10.3390/IJMS22147350.
- [108] Zejin Ou et al. “Global trends in the incidence, prevalence, and years lived with disability of Parkinson’s disease in 204 countries/territories from 1990 to 2019”. In: *Front. Public Health* 9 (2021). DOI: 10.3389/fpubh.2021.776847.
- [109] Yoav Ben-Shlomo, Sirwan Darweesh, Jorge Llibre-Guerra, Connie Marras, Marta San Luciano, and Caroline Tanner. “The epidemiology of Parkinson’s disease”. In: *Lancet* 403.10423 (2024), pp. 283–292. DOI: 10.1016/s0140-6736(23)01419-8.
- [110] Christina M. Lill. “Genetics of Parkinson’s disease”. In: *Molecular and Cellular Probes* 30.6 (2016), pp. 386–396. DOI: 10.1016/J.MCP.2016.11.001.
- [111] Md Shahidul Islam et al. “Pesticides and Parkinson’s disease: current and future perspective”. In: *J. Chem. Neuroanat.* 115 (2021), p. 101966. DOI: 10.1016/j.jchemneu.2021.101966.
- [112] Junjun Zhao, Shuqing Yu, Yan Zheng, Hui Yang, and Jianliang Zhang. “Oxidative modification and its implications for the neurodegeneration of Parkinson’s disease”. In: *Mol. Neurobiol.* 54.2 (2017), pp. 1404–1418. DOI: 10.1007/s12035-016-9743-3.
- [113] Antonio Ciaramella et al. “Blood Dendritic Cell Frequency Declines in Idiopathic Parkinson’s Disease and Is Associated with Motor Symptom Severity”. In: *PLoS ONE* 8.6 (2013). DOI: 10.1371/JOURNAL.PONE.0065352.
- [114] Gregory P. Williams, Aubrey M. Schonhoff, Asta Jurkuvenaite, Nicole J. Gallups, David G. Standaert, and Ashley S. Harms. “CD4 T cells mediate brain inflammation and neurodegeneration in a mouse model of Parkinson’s disease”. In: *Brain* 144.7 (2021), pp. 2047–2059. DOI: 10.1093/brain/awab103.
- [115] A. Mula, Xingxing Yuan, and Jinrong Lu. “Dendritic cells in Parkinson’s disease: regulatory role and therapeutic potential”. In: *Eur. J. Pharmacol.* 976 (2024), p. 176690. DOI: 10.1016/j.ejphar.2024.176690.

- [116] V Brochard. “Infiltration of CD4+ lymphocytes into the brain contributes to neurodegeneration in a mouse model of Parkinson disease”. In: *J. Clin. Investig.* 119 (2009), pp. 182–192.
- [117] Amirhossein Ahmadi, Juan D. Gispert, Arcadi Navarro, Natalia Vilor-Tejedor, and Iman Sadeghi. “Single-cell transcriptional changes in neurodegenerative diseases”. In: *Neuroscience* 479 (2021), pp. 192–205. DOI: 10.1016/j.neuroscience.2021.10.025.
- [118] Jelena Pozojevic and Malte Spielmann. “Single-cell sequencing in neurodegenerative disorders”. In: *Mol. Diagn. Ther.* 27.5 (2023), pp. 553–561. DOI: 10.1007/s40291-023-00668-9.
- [119] David Sulzer et al. “T cells from patients with Parkinson’s disease recognize α -synuclein peptides”. In: *Nature* 546.7660 (2017), pp. 656–661. DOI: 10.1038/nature22815.
- [120] Liu Lin Xiong et al. “Single-cell RNA sequencing reveals peripheral immunological features in Parkinson’s disease”. In: *Npj Park. Dis.* 10.1 (2024), p. 185. DOI: 10.1038/s41531-024-00790-3.
- [121] Antoine Freuchet et al. “Differential memory enrichment of cytotoxic CD4 T cells in Parkinson’s disease patients reactive to α -synuclein”. In: *Npj Park. Dis.* 11.1 (2025), p. 127. DOI: 10.1038/s41531-025-00981-6.
- [122] Chiara Rickenbach et al. “Altered T-cell reactivity in the early stages of Alzheimer’s disease”. In: *Brain* 148.9 (2025), pp. 3364–3378. DOI: 10.1093/brain/awaf167.
- [123] Y Wang. “Characteristics of T cells in single-cell datasets of peripheral blood and cerebrospinal fluid in Alzheimer’s disease patients”. In: *J. Alzheimer’s. Dis.* 99 (2024), S265–S280.
- [124] B. V. North, D. Curtis, and P. C. Sham. “A note on the calculation of empirical P values from Monte Carlo procedures [1]”. In: *American Journal of Human Genetics* 71.2 (2002), pp. 439–441. DOI: 10.1086/341527.
- [125] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74. DOI: 10.1038/NATURE15393.

- [126] Melissa J. Landrum et al. “ClinVar: improving access to variant interpretations and supporting evidence”. In: *Nucleic acids research* 46.D1 (2018), pp. D1062–D1067. DOI: 10.1093/NAR/GKX1153.
- [127] Nicole A Vasilevsky et al. “Mondo: integrating disease terminology across communities”. In: *Genetics* 232.4 (2026). DOI: 10.1093/GENETICS/IYAF215.
- [128] Ammar J. Alsheikh et al. “The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases”. In: *BMC Medical Genomics* 2022 15:1 15.1 (2022), pp. 74–. DOI: 10.1186/S12920-022-01216-W.
- [129] Betul Akgol Oksuz et al. “Systematic evaluation of chromosome conformation capture assays”. In: *Nature methods* 18.9 (2021), pp. 1046–1055. DOI: 10.1038/S41592-021-01248-7.
- [130] Uku Raudvere et al. “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic Acids Research* 47 (2019), pp. 191–198. DOI: 10.1093/nar/gkz369.
- [131] Michael Ashburner et al. “Gene ontology: Tool for the unification of biology”. In: *Nature Genetics* 25.1 (2000), pp. 25–29. DOI: 10.1038/75556.
- [132] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. “New approach for understanding genome variations in KEGG”. In: *Nucleic Acids Research* 47.D1 (2019), pp. D590–D595. DOI: 10.1093/nar/gky962.
- [133] Antonio Fabregat et al. “The Reactome Pathway Knowledgebase”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D649–D655. DOI: 10.1093/nar/gkx1132.
- [134] Chih Hung Chou et al. “MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions”. In: *Nucleic Acids Research* 46.D1 (2018), pp. D296–D302. DOI: 10.1093/nar/gkx1067.
- [135] V. Matys et al. “TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.” In: *Nucleic acids research* 34.Database issue (2006), pp. 108–110. DOI: 10.1093/nar/gkj143.

- [136] Min Zhao, Pora Kim, Ramkrishna Mitra, Junfei Zhao, and Zhongming Zhao. “TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes”. In: *Nucleic acids research* 44.D1 (2016), pp. D1023–D1031. DOI: 10.1093/NAR/GKV1268.
- [137] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. “DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors”. In: *Cell Syst.* 8.4 (2019), 329–337.e4. DOI: 10.1016/j.cels.2019.03.003.
- [138] Yuhao Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), 3573–3587.e29. DOI: 10.1016/j.cell.2021.04.048.
- [139] Xiaojie Qiu et al. “Reversed graph embedding resolves complex single-cell trajectories”. In: *Nat. Methods* 14.10 (2017), pp. 979–982. DOI: 10.1038/nmeth.4402.
- [140] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biol.* 15.12 (2014). DOI: 10.1186/s13059-014-0550-8.
- [141] Tim Stuart et al. “Comprehensive integration of single-cell data”. In: *Cell* 177.7 (2019), 1888–1902.e21. DOI: 10.1016/j.cell.2019.05.031.
- [142] R. Savica et al. “Medical records documentation of constipation preceding Parkinson disease: a case-control study”. In: *Neurology* 73.21 (2009), pp. 1752–1758. DOI: 10.1212/wnl.0b013e3181c34af5.
- [143] G. Stennis Watson and James B. Leverenz. “Profile of Cognitive Impairment in Parkinson’s Disease”. In: *Brain Pathology* 20.3 (2010), p. 640. DOI: 10.1111/J.1750-3639.2010.00373.X.
- [144] Antonio Terracciano et al. “Neuroticism and risk of Parkinson’s Disease: A meta-analysis”. In: *Movement disorders : official journal of the Movement Disorder Society* 36.8 (2021), p. 1863. DOI: 10.1002/MDS.28575.
- [145] Kelly C. Cushing-Damm et al. “Genetic insight into the relationship between inflammatory bowel disease and *Clostridioides difficile* infection”. In: *mSphere* 9.11 (2024). DOI: 10.1128/MSPHERE.00567-24.

- [146] Mary Pat Reeve et al. “Oral and non-oral lichen planus show genetic heterogeneity and differential risk for autoimmune disease and oral cancer.” In: *American Journal of Human Genetics* 111.6 (2024), pp. 1047–1060. DOI: 10.1016/J.AJHG.2024.04.020.
- [147] Carolina Roselli et al. “Multi-ethnic genome-wide association study for atrial fibrillation”. In: *Nature Genetics* 50.9 (2018), pp. 1225–1233. DOI: 10.1038/S41588-018-0133-9.
- [148] Jonas B. Nielsen et al. “Biobank-driven genomic discovery yields new insight into atrial fibrillation biology”. In: *Nature Genetics* 50.9 (2018), pp. 1234–1239. DOI: 10.1038/S41588-018-0171-3.
- [149] Fredrick R. Schumacher et al. “Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci”. In: *Nature Genetics* 50.7 (2018), pp. 928–936. DOI: 10.1038/S41588-018-0142-8.
- [150] Rosalind A. Eeles et al. “Identification of seven new prostate cancer susceptibility loci through a genome-wide association study”. In: *Nature Genetics* 41.10 (2009), pp. 1116–1121. DOI: 10.1038/NG.450.
- [151] Jonathan P. Bradfield et al. “A trans-ancestral meta-analysis of genome-wide association studies reveals loci associated with childhood obesity”. In: *Human Molecular Genetics* 28.19 (2019), pp. 3327–3338. DOI: 10.1093/HMG/DDZ161.
- [152] Ying Jin, Oliver H. Tam, Eric Paniagua, and Molly Hammell. “TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets”. In: *Bioinformatics* 31.22 (2015), p. 3593. DOI: 10.1093/BIOINFORMATICS/BTV422.
- [153] Yimeng Yin et al. “Impact of cytosine methylation on DNA binding specificities of human transcription factors”. In: *Science (New York, N.Y.)* 356.6337 (2017). DOI: 10.1126/SCIENCE.AAJ2239.
- [154] Michitaka Numoto, Ken Yokoro, and Junko Koshi. “ZF5, which is a Kruppel-type transcriptional repressor, requires the zinc finger domain for self-association”. In: *Bio-*

- chemical and biophysical research communications* 256.3 (1999), pp. 573–578. DOI: 10.1006/BBRC.1999.0375.
- [155] Sarah Meglaj Bakrač et al. “Single-cell analysis of the peripheral immune landscape in Parkinson’s disease: insights into dendritic cell and CD4+ T-cell transcriptomics”. In: *npj Parkinson’s Disease* 2026 12:1 12.1 (2026), pp. 73–. DOI: 10.1038/s41531-026-01283-1.
- [156] Charles Antoine Dutertre et al. “Single-cell analysis of human mononuclear phagocytes reveals subset-defining markers and identifies circulating inflammatory dendritic cells”. In: *Immunity* 51.3 (2019), 573–589.e8. DOI: 10.1016/j.immuni.2019.08.008.
- [157] Hui Xu and Jianping Jia. “Single-cell RNA sequencing of peripheral blood reveals immune cell signatures in Alzheimer’s disease”. In: *Front. Immunol.* 12 (2021), p. 645666. DOI: 10.3389/fimmu.2021.645666.
- [158] Florian Mair and Thomas Liechti. “Comprehensive phenotyping of human dendritic cells and monocytes”. In: *Cytometry A* 99.3 (2021), pp. 231–242. DOI: 10.1002/cyto.a.24269.
- [159] Aaron J. Wilk et al. “A single-cell atlas of the peripheral immune response in patients with severe COVID-19”. In: *Nat. Med.* 26.7 (2020), pp. 1070–1076. DOI: 10.1038/s41591-020-0944-y.
- [160] Tobias Poch et al. “Single-cell atlas of hepatic T cells reveals expansion of liver-resident naive-like CD4+ T cells in primary sclerosing cholangitis”. In: *J. Hepatol.* 75.2 (2021), pp. 414–423. DOI: 10.1016/j.jhep.2021.03.016.
- [161] Peter A. Szabo et al. “Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease”. In: *Nat. Commun.* 10.1 (2019). DOI: 10.1038/s41467-019-12464-3.
- [162] Xuanqi Liu, Zhenhua Zhu, and Xiangdong Wang. “Specificity and function of T cell subset identities using single-cell sequencing”. In: *Clin. Transl. Discov.* 3.3 (2023), e199. DOI: 10.1002/ctd2.199.

- [163] Alexandra Chloé Villani et al. “Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors”. In: *Science* 356.6335 (2017). DOI: 10.1126/science.aah4573.
- [164] Long Li et al. “A far downstream enhancer for murine Bcl11b controls its T-cell specific expression”. In: *Blood* 122.6 (2013), pp. 902–911. DOI: 10.1182/BLOOD-2012-08-447839.
- [165] Jordan W. Squair et al. “Confronting false discoveries in single-cell differential expression”. In: *Nature Communications* 2021 12:1 12.1 (2021), pp. 5692–. DOI: 10.1038/s41467-021-25960-2.
- [166] Yakov A. Tsepilov et al. “The Human Pleiotropic Map of GWAS Associations and Therapeutic Implications”. In: *bioRxiv : the preprint server for biology* (2026). DOI: 10.64898/2026.04.28.721048.
- [167] Sharon Eden and Howard Cedar. “Role of DNA methylation in the regulation of transcription”. In: *Current Opinion in Genetics & Development* 4.2 (1994), pp. 255–259. DOI: 10.1016/S0959-437X(05)80052-8.
- [168] Michael J. Pikaart, Félix Recillas-Targa, and Gary Felsenfeld. “Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators”. In: *Genes & development* 12.18 (1998), pp. 2852–2862. DOI: 10.1101/GAD.12.18.2852.
- [169] A. Bird. “DNA methylation patterns and epigenetic memory”. In: *Genes & Development* 16.1 (2002), pp. 6–21. DOI: 10.1101/GAD.947102.
- [170] Hua Tan, Shan Huang, Zhigang Zhang, Xiaohua Qian, Peiqing Sun, and Xiaobo Zhou. “Pan-cancer analysis on microRNA-associated gene activation”. In: *EBioMedicine* 43 (2019), pp. 82–97. DOI: 10.1016/j.ebiom.2019.03.082.
- [171] Peng Xu et al. “A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers”. In: *Frontiers in Genetics* 0 (2020), p. 278. DOI: 10.3389/FGENE.2020.00278.

- [172] Yaming Zhang and Haili Hu. “Long non-coding RNA CCAT1/miR-218/ZFX axis modulates the progression of laryngeal squamous cell cancer”. In: *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 39.6 (2017). DOI: 10.1177/1010428317699417.
- [173] Renata Gruszka, Krzysztof Zakrzewski, Paweł Piotr Liberski, and Magdalena Zakrzewska. “mRNA and miRNA Expression Analyses of the MYC/E2F/miR-17-92 Network in the Most Common Pediatric Brain Tumors”. In: *International Journal of Molecular Sciences* 2021, Vol. 22, Page 543 22.2 (2021), p. 543. DOI: 10.3390/IJMS22020543.
- [174] Letizia Venturini et al. “Expression of the miR-17-92 polycistron in chronic myeloid leukemia (CML) CD34+ cells”. In: *Blood* 109.10 (2007), pp. 4399–4405. DOI: 10.1182/BLOOD-2006-09-045104.
- [175] Amal Kassab, Ishita Gupta, and Ala Eddin Al Moustafa. “Role of E2F transcription factor in oral cancer: Recent insight and advancements”. In: *Seminars in Cancer Biology* 92 (2023), pp. 28–41. DOI: 10.1016/J.SEMCANCER.2023.03.004.
- [176] Julia A. Belk, Bence Daniel, and Ansuman T. Satpathy. “Epigenetic regulation of T cell exhaustion”. In: *Nature Immunology* 2022 23:6 23.6 (2022), pp. 848–860. DOI: 10.1038/s41590-022-01224-z.
- [177] Saray Quintero-Fabián et al. “Role of Matrix Metalloproteinases in Angiogenesis and Cancer”. In: *Frontiers in Oncology* 9 (2019), p. 480558. DOI: 10.3389/FONC.2019.01370/XML.
- [178] Mike A. Nalls et al. “Identification of novel risk loci, causal insights, and heritable risk for Parkinson’s disease: a meta-genome wide association study”. In: *The Lancet. Neurology* 18.12 (2019), p. 1091. DOI: 10.1016/S1474-4422(19)30320-5.
- [179] Cecilia S. Lindestam Arlehamn et al. “ α -Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson’s disease”. In: *Nat. Commun.* 11.1 (2020). DOI: 10.1038/s41467-020-15626-w.

- [180] Jens Geginat, Federica Sallusto, and Antonio Lanzavecchia. “Cytokine-driven proliferation and differentiation of human naive, central memory, and effector memory CD4+ T Cells”. In: *J. Exp. Med.* 194.12 (2001), pp. 1711–1719. DOI: 10.1084/jem.194.12.1711.
- [181] Cláudia Serre-Miranda et al. “Effector memory CD4+ T cells are associated with cognitive performance in a senior population”. In: *Neurol. Neuroimmunol. Neuroinflammation* 2.1 (2015). DOI: 10.1212/nxi.0000000000000054.
- [182] Gael Moquin-Beaudry et al. “Mapping the peripheral immune landscape of Parkinson’s disease patients with single-cell sequencing”. In: *Brain* 148.8 (2025), pp. 2847–2860. DOI: 10.1093/brain/awaf066.
- [183] Natasa Kustrimovic et al. “Parkinson’s disease patients have a complex phenotypic and functional Th1 bias: cross-sectional studies of CD4+ Th1/Th2/T17 and Treg in drug-naïve and drug-treated patients”. In: *J. Neuroinflammation* 15.1 (2018). DOI: 10.1186/s12974-018-1248-8.
- [184] Sen Jiang, Hua Gao, Qin Luo, Pengfei Wang, and Xinling Yang. “The correlation of lymphocyte subsets, natural killer cell, and Parkinson’s disease: a meta-analysis”. In: *Neurol. Sci.* 38.8 (2017), pp. 1373–1380. DOI: 10.1007/s10072-017-2988-4.
- [185] Claire H. Stevens et al. “Reduced T helper and B lymphocytes in Parkinson’s disease”. In: *J. Neuroimmunol.* 252.1-2 (2012), pp. 95–99. DOI: 10.1016/j.jneuroim.2012.07.015.
- [186] Juan Liu, Xiaomin Zhang, Yujie Cheng, and Xuetao Cao. “Dendritic cell migration in inflammation and immunity”. In: *Cell. Mol. Immunol.* 18.11 (2021), pp. 2461–2471. DOI: 10.1038/s41423-021-00726-4.
- [187] Yasuyuki Saito, Satomi Komori, Takenori Kotani, Yoji Murata, and Takashi Matozaki. “The Role of type-2 conventional dendritic cells in the regulation of tumor immunity”. In: *Cancers* 14.8 (2022), p. 1976. DOI: 10.3390/cancers14081976.

- [188] Xiangyun Yin et al. “Human blood CD1c+ dendritic cells encompass CD5high and CD5low subsets that differ significantly in phenotype, gene expression, and functions”. In: *J. Immunol.* 198.4 (2017), pp. 1553–1564. DOI: 10.4049/jimmunol.1600193.
- [189] Pierre Bourdely et al. “Transcriptional and functional analysis of CD1c+ human dendritic cells identifies a CD163+ subset priming CD8+CD103+ T cells”. In: *Immunity* 53.2 (2020), 335–352.e8. DOI: 10.1016/j.immuni.2020.06.002.
- [190] Ingrid M. Leal Rojas et al. “Human blood CD1c+ dendritic cells promote Th1 and Th17 effector function in memory CD4+ T cells”. In: *Front. Immunol.* 8.AUG (2017), p. 971. DOI: 10.3389/fimmu.2017.00971.
- [191] Andrew J. Kassianos, Sarah L. Jongbloed, Derek N.J. Hart, and Kristen J. Radford. “Isolation of human blood DC subtypes”. In: *Methods in molecular biology (Clifton, N.J.)* 595 (2010), pp. 45–54. DOI: 10.1007/978-1-60761-421-0_3.
- [192] Yuhua Chen et al. “Clinical correlation of peripheral CD4+-cell sub-sets, their imbalance and Parkinson’s disease”. In: *Mol. Med. Rep.* 12.4 (2015), pp. 6105–6111. DOI: 10.3892/mmr.2015.4136.
- [193] Xiaoming Zhang et al. “CD4 T cells with effector memory phenotype and function develop in the sterile environment of the fetus”. In: *Sci. Transl. Med.* 6.238 (2014), 238ra72. DOI: 10.1126/scitranslmed.3008748.
- [194] Nicolas Dzamk. “Cytokine activity in Parkinson’s disease”. In: *Neuronal Signal* 7.4 (2023), NS20220063. DOI: 10.1042/ns20220063.
- [195] Giulia Di Lazzaro et al. “Differential profiles of serum cytokines in Parkinson’s disease according to disease duration”. In: *Neurobiol. Dis.* 190 (2024), p. 106371. DOI: 10.1016/j.nbd.2023.106371.
- [196] Beverly W. Baron and Peter Pytel. “Expression pattern of the BCL6 and ITM2B proteins in normal human brains and in Alzheimer disease”. In: *Appl. Immunohistochem. Mol. Morphol.* 25.7 (2017), pp. 489–496. DOI: 10.1097/pai.0000000000000329.

- [197] Longping Yao et al. “Bioinformatic analysis of genetic factors from human blood samples and postmortem brains in Parkinson’s disease”. In: *Oxid. Med. Cell. Longev.* 2022 (2022), p. 9235358. DOI: 10.1155/2022/9235358.
- [198] A. L. Shaffer, Xin Yu, Yunsheng He, Jennifer Boldrick, Erick P. Chan, and Louis M. Staudt. “BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control”. In: *Immunity* 13.2 (2000), pp. 199–212. DOI: 10.1016/s1074-7613(00)00020-0.
- [199] Jingwei Zhao et al. “CCL5 promotes LFA-1 expression in Th17 cells and induces LCK and ZAP70 activation in a mouse model of Parkinson’s disease”. In: *Front. Aging Neurosci.* 15 (2023), p. 1250685. DOI: 10.3389/fnagi.2023.1250685.
- [200] Dan Feng et al. “LncRNA SOX21-AS1 promotes activation of BV2 cells via epigenetical silencing of SOCS3 and aggravates Parkinson’s disease”. In: *Gerontology* 70.10 (2024), pp. 1063–1073. DOI: 10.1159/000539784.
- [201] Y Liu. “LncRNA MALAT1 facilitates Parkinson’s disease progression by increasing SOCS3 promoter methylation”. In: *Gerontology* 70 (2024), pp. 1294–1304.
- [202] Ellen V. Rothenberg. “Transcriptional drivers of the T-cell lineage program”. In: *Current Opinion in Immunology* 24.2 (2012), pp. 132–138. DOI: 10.1016/J.COI.2011.12.012.
- [203] Matthew J. Lennon, Simon P. Jones, Michael D. Lovelace, Gilles J. Guillemin, and Bruce J. Brew. “Bcl11b-A Critical Neurodevelopmental Transcription Factor-Roles in Health and Disease”. In: *Frontiers in cellular neuroscience* 11 (2017). DOI: 10.3389/FNCEL.2017.00089.
- [204] Francesco Annunziato, Lorenzo Cosmi, Francesco Liotta, Enrico Maggi, and Sergio Romagnani. “Defining the human T helper 17 cell phenotype”. In: *Trends Immunol.* 33.10 (2012), pp. 505–512. DOI: 10.1016/j.it.2012.05.004.
- [205] Melissa Saichi et al. “Single-cell RNA sequencing of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity”. In: *Nat. Cell Biol.* 23.5 (2021), pp. 538–551. DOI: 10.1038/s41556-021-00681-2.

- [206] Dorian Dulčić, Katarina Mandić, Dalibor Hršak, and Anja Barešić. “Long-range regulatory target prediction reveals shared genetic background across ulcerative colitis, Crohn’s disease, primary sclerosing cholangitis and ankylosing spondylitis”. In: *bioRxiv* (2026).

Summary

More than 95% of disease-associated variants identified in genome-wide association studies (GWASs) fall into non-coding regions of the genome, making it difficult to discern the affected gene or the molecular mechanism. Often, the gene nearest to the variant is implicated as the candidate gene; however, this view is simplistic and does not take into account the three-dimensional organisation of the genome. The variant and gene may be distal in the linear genome and come into close physical proximity, making it even more complicated to pinpoint the exact affected gene. To tackle this problem, this thesis utilises findings from comparative genomics that revealed non-coding regions that show an extreme level of evolutionary conservation and have a regulatory role. These conserved non-coding regions are clustered together in arrays of synteny named genomic regulatory blocks (GRBs). GRBs form a regulatory landscape for one important developmental target gene while other genes in the region are typically not responsive to the regulatory elements. GRBs may span several megabases in genomic length and facilitate long-range regulation, and can thereby be utilised to interpret disease-associated variants found through GWAS. Furthermore, GRBs have a role in complex disease and this work systematically applies the GRB approach on the whole GWAS Catalog to identify phenotypic traits which have a propensity for long-range regulation of transcription activity, and genomic localization of loci within GRBs. Moreover, additional high-throughput molecular data is analysed in greater detail for two examples: head and neck squamous cell carcinoma and Parkinson's disease.

Sažetak

Više od 95% varijanti povezani s bolestima utvrđeni u cijelogenomskim asocijacijskim studijama (GWAS) nalazi se u nekodirajućim regijama genoma, što otežava određivanje ciljnog gena ili molekularnog mehanizma djelovanja. Često se gen koji je najbliži varijanti implicira kao gen kandidat, ali, ovaj jednostavan pristup ne uzima u obzir trodimenzionalnu organizaciju genoma. Varijanta i gen mogu biti udaljeni u linearnom genomu i fizički doći u blizinu, što otežava točno određivanje ciljnog gena. Kako bi riješila taj problem, ova teza koristi nalaze iz komparativne genomike koje su otkrile postojanje nekodirajućih regija koje imaju ekstremnu evolucijsku očuvanost te imaju regulatornu ulogu tijekom transkripcije gena. Ove očuvane nekodirajuće regije su grupirane u sintenske nizove koji se nazivaju genomski regulatorni blokovi (GRB-ovi). GRB-ovi formiraju regulatorni krajolik za jedan važan razvojni gen dok drugi geni iz te regije su uglavnom neosjetljivi na regulatorne elemente. GRB-ovi mogu imati raspon od nekoliko megabaza genomske duljine i omogućavati regulaciju dalekog dometa. Nadalje, GRB-ovi imaju ulogu u kompleksnim bolestima te se u ovom radu sustavno primjenjuje GRB pristup na cijeli GWAS katalog kako bi se utvrdila fenotipska svojstva koja imaju sklonost regulaciji transkripcijske aktivnosti dalekog dometa te genomske lokalizacije lokusa unutar GRB-a. Također, dodatni molekularni podaci generirani visoko-propusnim metodama su detaljnije analizirani za dvije reprezentativne bolesti: karcinom pločastih stanica glave i vrata i Parkinsonovu bolest.

Curriculum Vitae

Personal information

Name: Katarina Mandić

Email: katarina.mandić@irb.hr

ORCID: 0009-0008-2968-369X

Affiliation: Laboratory for Computational Biology and Translational Medicine, Division of Computing and Data Science, Ruđer Bošković Institute, Zagreb, Croatia

Education

2022–

Josip Juraj Strossmayer University of Osijek, Osijek, Croatia.

PhD candidate in the interdisciplinary field of biology and basic medical sciences

University Postgraduate Interdisciplinary Doctoral Study, Molecular Biosciences Thesis: *The role of long-range transcriptional regulation across human disease landscape*

2017–2020

MSc in the field of molecular biotechnology

University of Zagreb, Faculty of Food Technology and Biotechnology

2013–2017

BSc in biotechnology

University of Zagreb, Faculty of Food Technology and Biotechnology

Professional experience

2021–2024 - Research assistant at the Laboratory for machine learning and knowledge representation, Division of electronics, Ruđer Bošković Institute

2024– - Research assistant at the Laboratory for Computational Biology and Translational Medicine, Division of Computing and Data Science, Ruđer Bošković Institute, Zagreb, Croatia

Publications

1. Dulčić, D., Mandić, K., Hršak, D., Barešić, A. **Long-range regulatory target prediction reveals shared genetic background across ulcerative colitis, Crohn's disease, primary sclerosing cholangitis and ankylosing spondylitis.**
Preprint, 2026.
DOI: 10.64898/2026.06.29.735270
2. Mandić, K., Hršak, D., Uljanić, F., Lenhard, B., Barešić, A. **The role of long-range transcriptional regulation in interpretation of non-coding variants associated with human disease.**
Preprint, 2026.
DOI: 10.64898/2026.06.15.731051
3. Meglaj Bakrač, S., Mandić, K., Cvetko Krajinović, L. *et al.* **Single-cell analysis of the peripheral immune landscape in Parkinson's disease: insights into dendritic cell and CD4+ T-cell transcriptomics.**
npj Parkinson's Disease, 12, 73 (2026).
DOI: 10.1038/s41531-026-01283-1
4. Mandić, K., Milutin Gašperov, N., Božinović, K. *et al.* **Integrative analysis in head and neck cancer reveals distinct role of miRNome and methylome as tumour epigenetic drivers.**
Scientific Reports, 14, 9062 (2024).
DOI: 10.1038/s41598-024-59312-z

Conference presentations

- Božinović, K., Mandić, K., Milutin Gašperov, N., Sabol, I., Barešić, A., Grce, M. **Integrative analysis of multi-omics data identifies microRNA or DNA methylation as independent epigenetic driving forces in HNSCC.**

Poster presentation at the *EMBO Workshop: Epigenome Inheritance and Reprogramming in Health and Disease*, Mediterranean Institute for Life Sciences, Split, Croatia, September 22–26, 2022.

- Mandić, K., Božinović, K., Milutin Gašperov, N., Sabol, I., Barešić, A., Dediol, E., Bennet Saidu, N. E., Krasić, J., Sinčić, N., Grce, M. **Multi-omics analysis of HNSCC reveal non-overlapping epigenomic regulation of gene expression.**

Poster presentation at the *Intelligent Systems for Molecular Biology/ European Conference on Computational Biology (ISMB/ECCB) 2023*, Lyon, France, July 23–27, 2023.

- Uljanić, F., Mandić, K., Hršak, D., Barešić, A. **Genomic regulatory blocks are enriched in variants associated with mental and behavioural disorders and show evidence of long-range regulation.**

Poster presentation at the *International Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2024)*, Gran Canaria, Spain, July 15–17, 2024.

Awards and grants

- **Financial support for Incentive Projects for the Best Scientific Work of Scientists (ZI6-26)**, Ruđer Bošković Institute, 2026.

- **COST Action CA17138 Short-Term Scientific Mission (STSM).**

Two-week research stay at the Laboratory of Transplantation Genomics, Research and Development, Finnish Red Cross Blood Service, Helsinki, Finland, 2022.

A Supplementary Figures

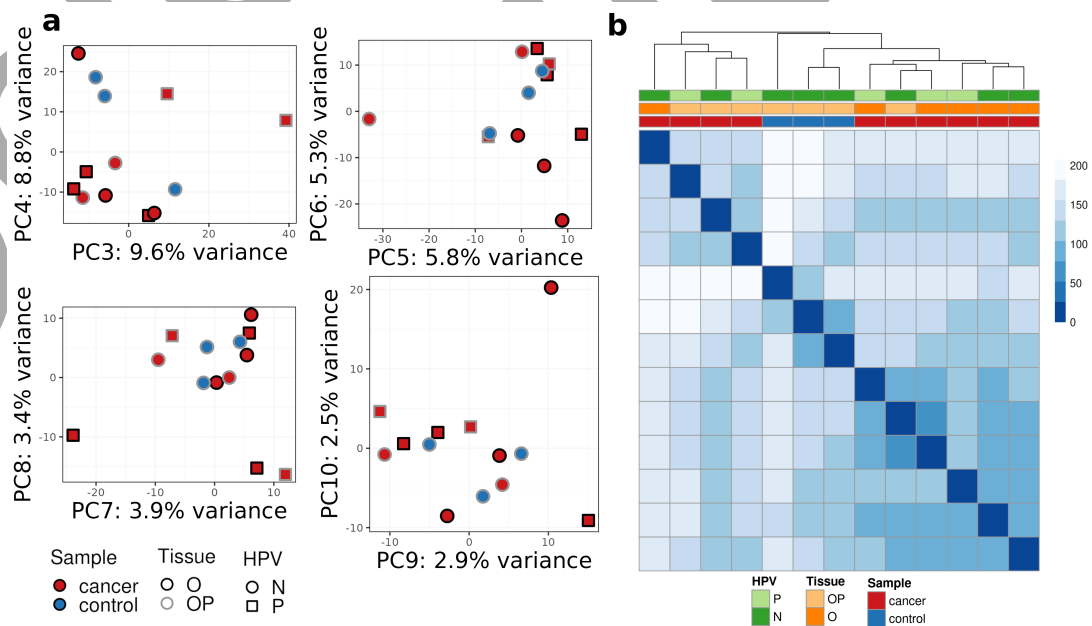


Figure S1. a) Principal components 3-10. b) Sample distance heatmap using Euclidean metric. Adapted from Mandić et al. [103].

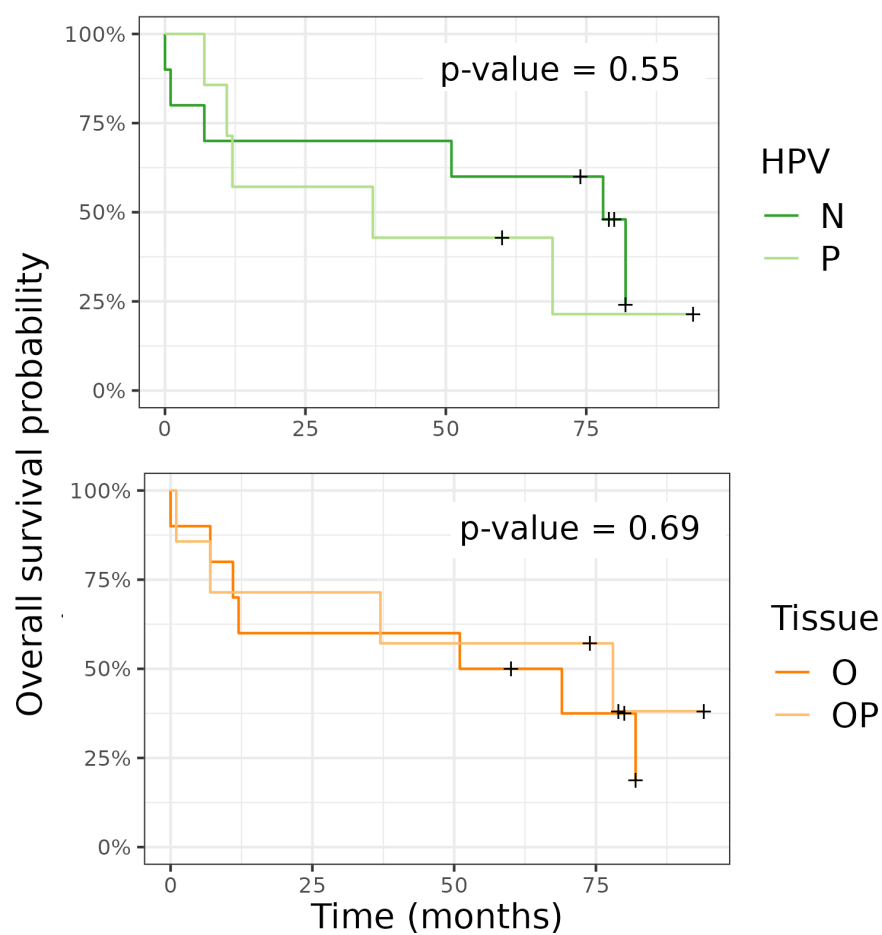
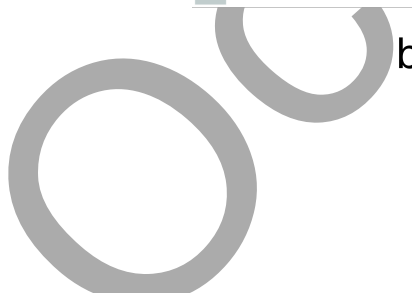


Figure S2. Kaplan–Meier plots comparing overall survival between HPV-positive and HPV-negative patients and between oral and oropharyngeal tumour locations. Adapted from Mandić et al. [103].



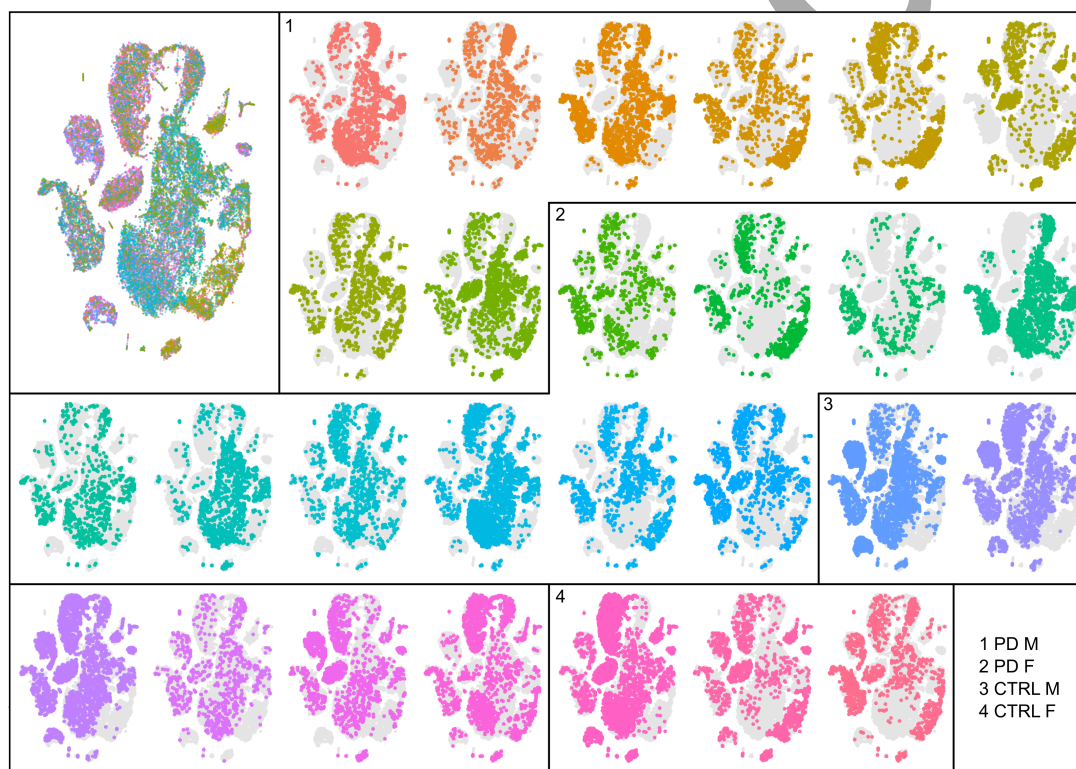


Figure S4. tSNE donor overlay. Each tSNE projection represents a single sample. Samples are grouped according to disease status and sex as follows: 1, male PD; 2, female PD; 3, male CTRL; and 4, female CTRL. PD, Parkinson's disease; CTRL, control; M, male; F, female. Adapted from Meglaj Bakrač et al. [155].

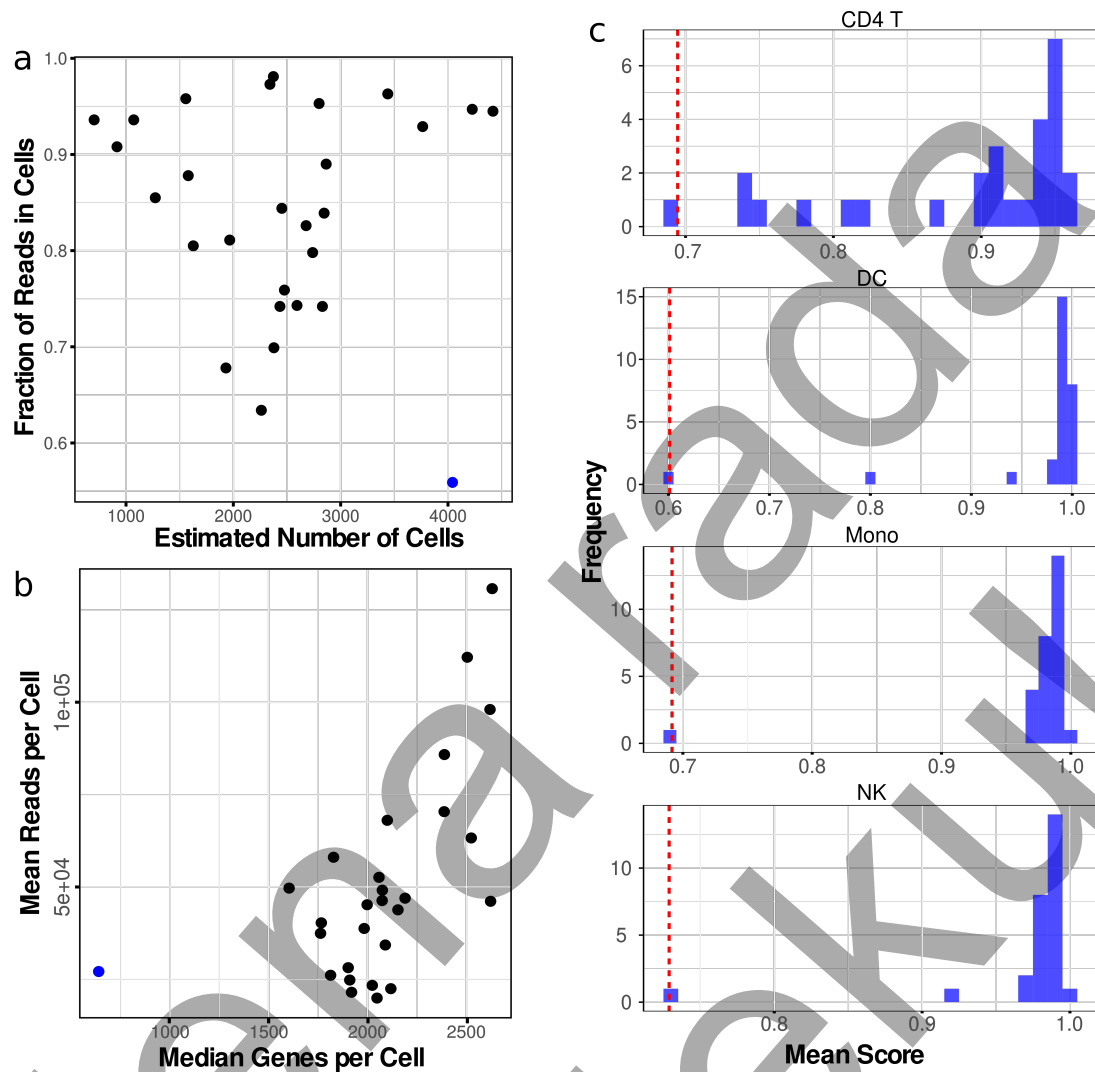


Figure S5. Exclusion of sample S29 based on quality control metrics. a Scatter plot showing sample S29 (blue) with the lowest fraction of reads assigned to cells and the highest estimated cell count. b Sample S29 exhibited the lowest mean number of reads per cell together with a low median number of detected genes per cell. c Mean prediction confidence scores calculated from reference-based annotation for the major cell populations (CD4+ T-cells, DCs, monocytes, and NK cells). Sample S29 (red line) consistently showed the lowest confidence scores across all cell populations. Adapted from Meglaj Bakrač et al. [155].

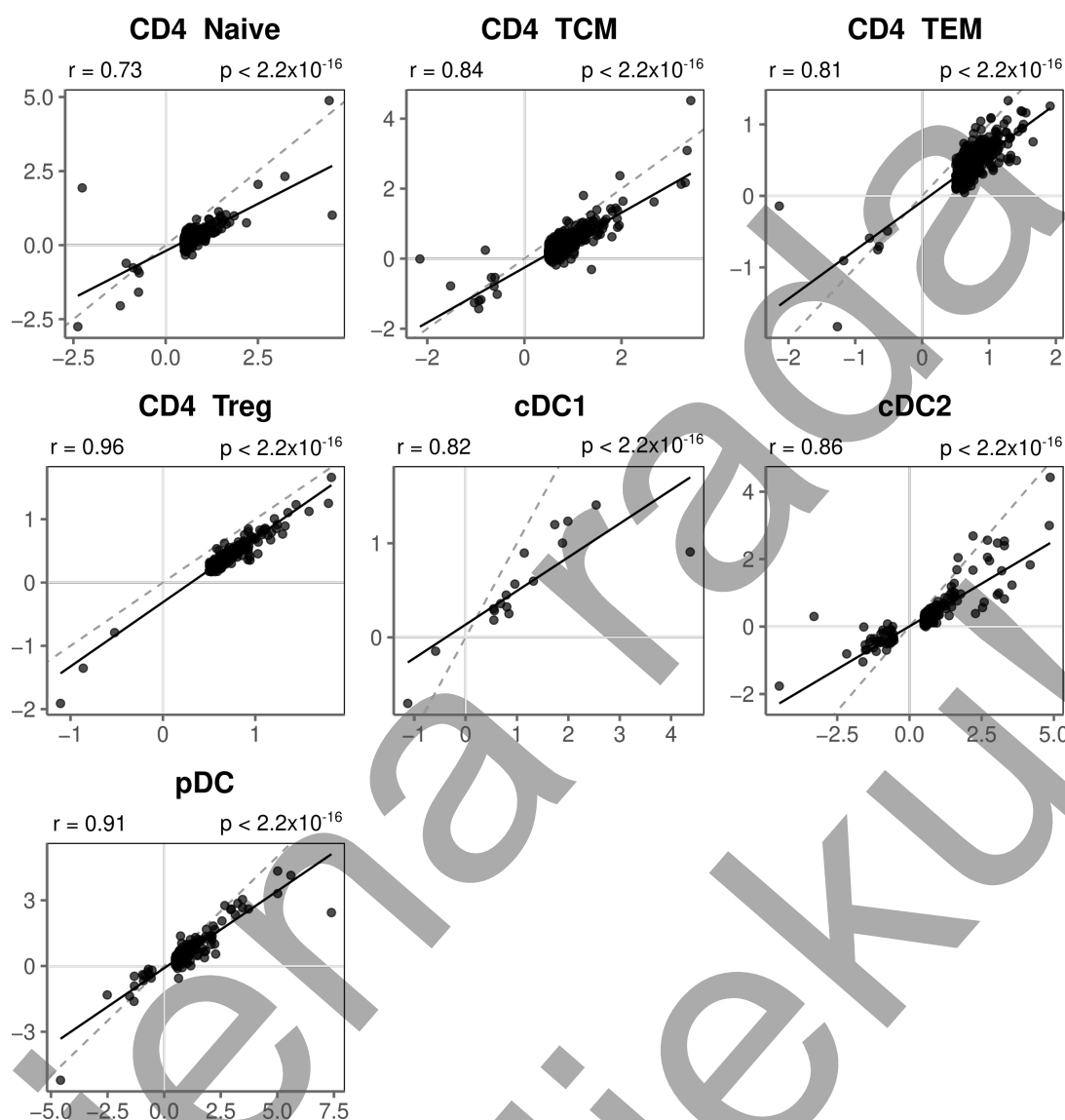


Figure S6. Scatterplots comparing $\log_2FC$ values obtained using the single-cell approach (x-axis) and the pseudobulk approach (y-axis). Only genes reaching the significance threshold in the PD versus CTRL comparison using the single-cell approach (Bonferroni-adjusted $p < 0.05$ and $|\log_2FC| > 0.5$, FindMarkers) were included. $\log_2FC$ values from the pseudobulk analysis were calculated using a standard DESeq2 pipeline. Adapted from Meglaj Bakrač et al. [155].

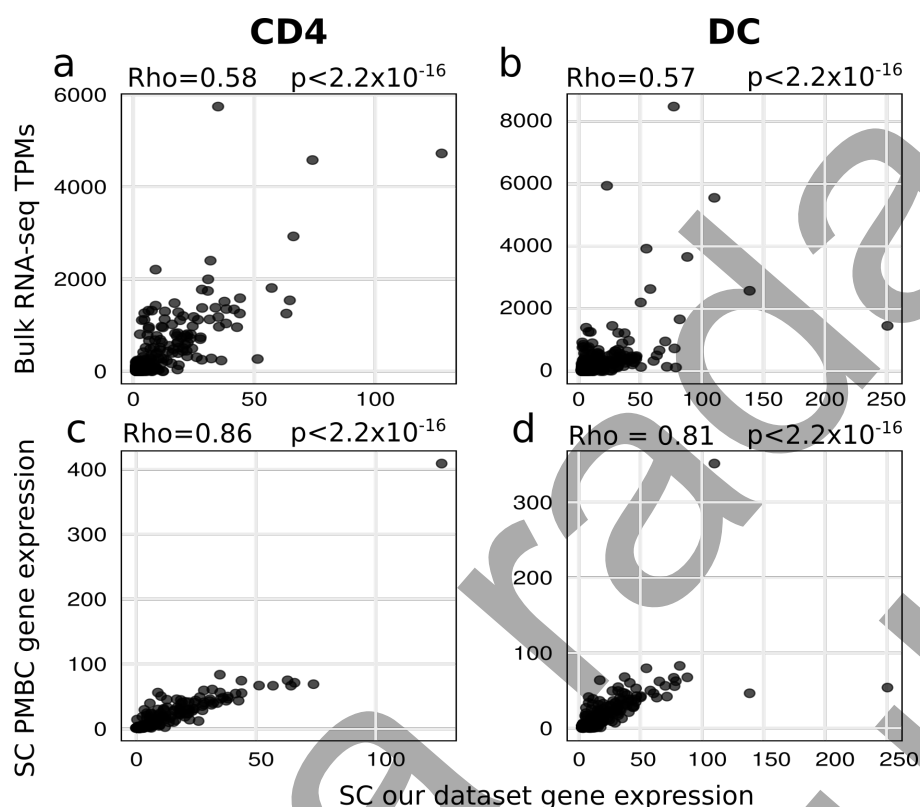


Figure S7. Comparison of gene expression profiles between the generated single-cell dataset and publicly available bulk RNA-seq and PBMC single-cell reference datasets. Scatter plots showing the relationship between gene expression levels, with Spearman's correlation coefficient (ρ) calculated for each comparison. **a** Bulk RNA-seq comparison for CD4+ T-cells. **b** Bulk RNA-seq comparison for DCs. **c** PBMC single-cell comparison for CD4+ T-cells. **d** PBMC single-cell comparison for DCs. Statistical significance was assessed at $p < 0.05$. Adapted from Meglaj Bakrač et al. [155].

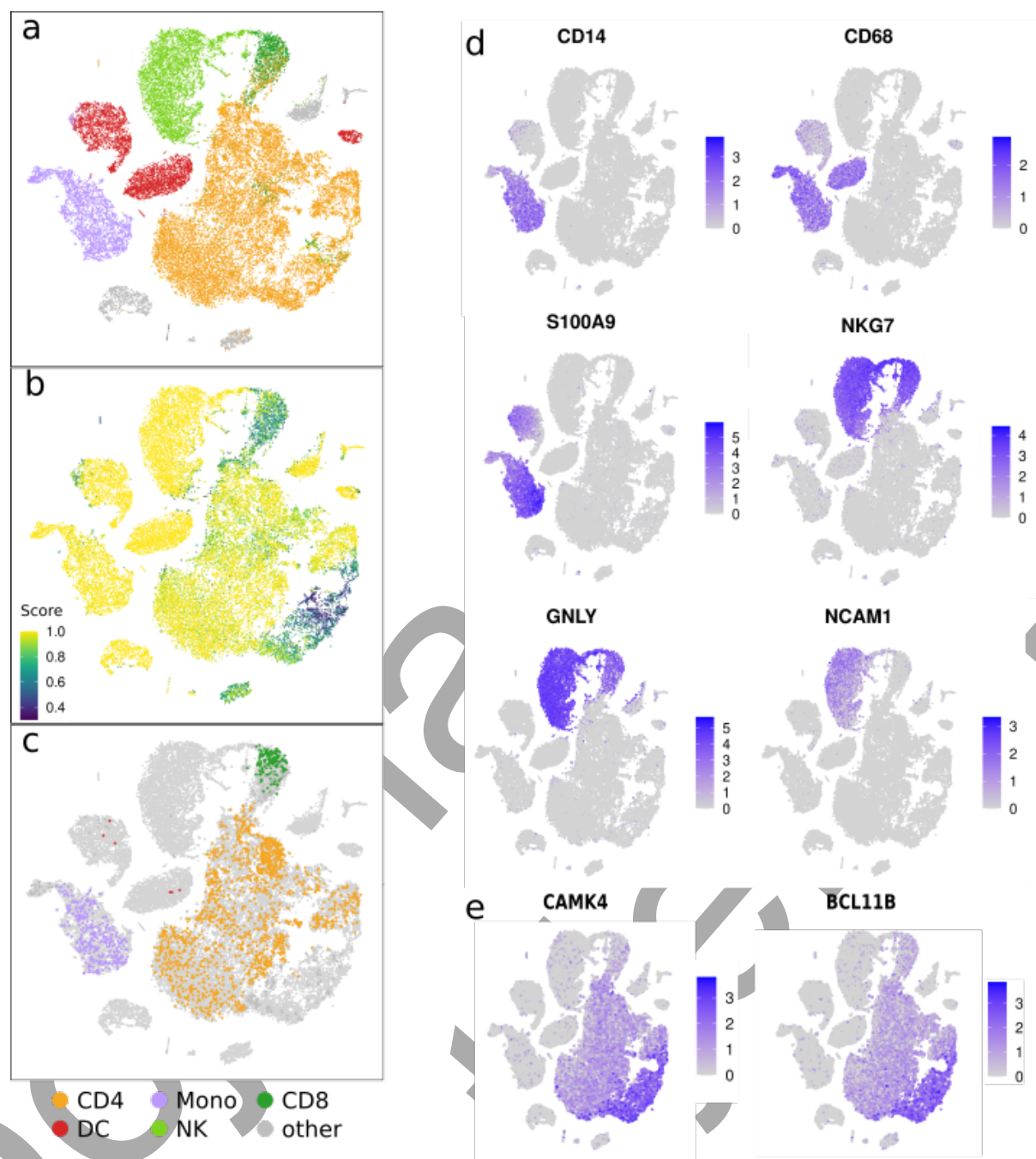


Figure S8. Extended cell type annotation. **a** Initial cell type annotation obtained using the reference-based approach. **b** Annotation confidence scores highlighting a subset of low-confidence CD4+ T cells that were excluded from downstream analyses. **c** tSNE projection highlighting cells originating from sample S13, which contained only the CD4+ T-cell fraction. **d** Feature plots showing the expression of marker genes used to identify monocytes (*CD14*, *CD68*, *S100A9*) and NK cells (*NKG7*, *GNLY*, *NCAM1*). **e** Feature plots showing the expression of *CAMK4* and *BCL11B* across the majority of CD4+ T cells. Higher expression levels are indicated in purple. Adapted from Meglaj Bakrač et al. [155].

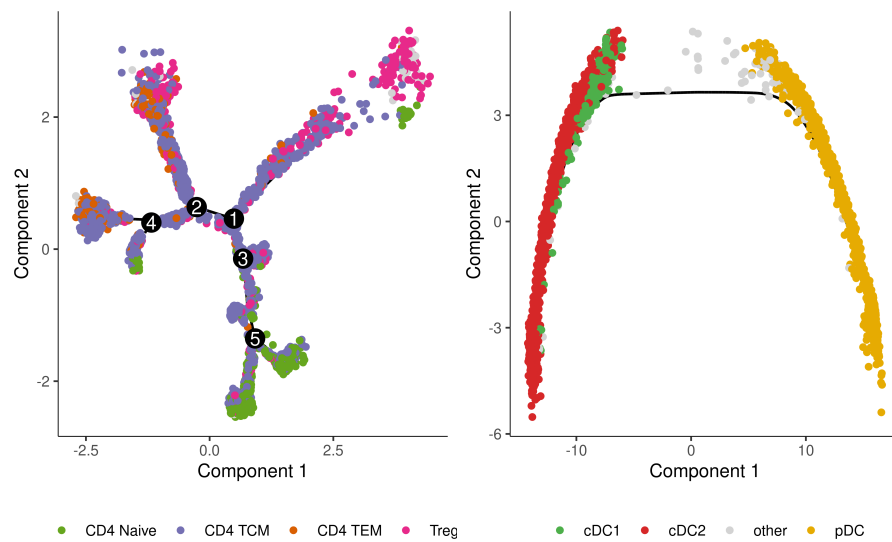


Figure S9. Differentiation trajectory analysis of a CD4⁺ T-cells and b DCs. As cDCs and pDCs do not originate from a shared progenitor, they do not form a common branching trajectory. Adapted from Meglaj Bakrač et al. [155].

B Supplementary Tables

Table S1. Mapping rates and observed fragment counts for RNA-seq samples.

Sample	Mapping rate (%)	Observed fragments
KBD10_quant	88	20,987,077
KBD11_quant	89	26,268,283
KBD13_quant	81	26,501,942
KBD15_quant	91	26,506,063
KBD16_quant	91	21,451,086
KBD17_quant	92	25,220,897
KBD18_quant	91	26,211,943
KBD21_quant	89	25,593,932
KBD2_quant	89	22,734,666
KBD3_quant	92	24,431,441
RNA10T_quant	84	27,502,903
RNA16T_quant	89	26,865,842
RNA3T_quant	82	27,068,651

Table S2. Downregulated tumour suppressors with hypermethylated promoters.

Ensembl ID	Gene	CGI coordinates	Mean expression	FC	Gene p	Gene adj. p	DNAm quot.	DNAm p	DNAm adj. p
ENSG00000117425	<i>PTCH2</i>	chr1:45308544-45309159	28.08	-3.16	1.18×10^{-6}	8.47×10^{-5}	0.482	1.91×10^{-3}	8.68×10^{-3}
ENSG00000162409	<i>PRKAA2</i>	chr1:57110664-57111337	132.97	-4.08	2.94×10^{-6}	1.75×10^{-4}	1.073	2.06×10^{-5}	3.34×10^{-4}
ENSG00000162706	<i>CADM3</i>	chr1:159141203-159141718	120.04	-3.90	9.70×10^{-9}	2.13×10^{-6}	1.119	1.59×10^{-6}	6.09×10^{-5}
ENSG00000178568	<i>ERBB4</i>	chr2:213402181-213403867	19.66	-2.90	5.93×10^{-3}	0.0412	0.119	3.65×10^{-4}	2.57×10^{-3}
ENSG00000175161	<i>CADM2</i>	chr3:85008285-85008837	7.75	-3.61	7.28×10^{-4}	9.68×10^{-3}	1.963	6.06×10^{-4}	3.69×10^{-3}
ENSG00000107562	<i>CXCL12</i>	chr10:44879715-44882391	1196.07	-3.45	3.27×10^{-6}	1.87×10^{-4}	1.672	1.81×10^{-3}	8.37×10^{-3}
ENSG00000140538	<i>NTRK3</i>	chr15:88799523-88801017	34.22	-5.15	8.82×10^{-15}	3.92×10^{-11}	1.313	2.75×10^{-4}	2.09×10^{-3}
ENSG00000181007	<i>ZFP82</i>	chr19:36909282-36909854	28.31	-2.25	2.48×10^{-5}	8.29×10^{-4}	3.291	2.42×10^{-5}	3.73×10^{-4}
ENSG00000161298	<i>ZNF382</i>	chr19:37095681-37096589	17.79	-2.19	3.53×10^{-3}	0.0289	2.106	3.16×10^{-5}	4.47×10^{-4}
ENSG00000198300	<i>PEG3</i>	chr19:57351284-57351995	128.52	-4.17	4.09×10^{-6}	2.20×10^{-4}	1.698	8.66×10^{-4}	4.83×10^{-3}
ENSG00000198300	<i>PEG3</i>	chr19:57352051-57352284	128.52	-4.17	4.09×10^{-6}	2.20×10^{-4}	1.013	1.92×10^{-3}	8.73×10^{-3}

CGI, CpG island; FC, fold change; DNAm, DNA methylation; p, p-value; adj. p, multiple-testing-adjusted p-value.

Table S3. Differences in the proportions of CD4+ T-cell and DC subtypes between PD and CTRL groups tested using Wilcoxon Rank Sum. No significant differences ($p < 0.05$) were detected for any subtype.

Cell type	Cell subtype	Median PD	Median CTRL	p-value
CD4 T	CD4 naive	0.158	0.154	0.980
CD4 T	CD4 TCM	0.629	0.619	0.596
CD4 T	CD4 TEM	0.076	0.080	1.000
CD4 T	Treg	0.102	0.100	0.900
DC	cDC1	0.064	0.092	0.091
DC	cDC2	0.402	0.484	0.561
DC	pDC	0.556	0.462	0.241

PD, Parkinson's disease; CTRL, control; TCM, central memory T-cell; TEM, effector memory T-cell; Treg, regulatory T-cell; DC, dendritic cell; cDC, conventional dendritic cell; pDC, plasmacytoid dendritic cell.

Table S4. Variant carrier counts in Parkinson's disease (PD) and control (CTRL) samples, with corresponding Fisher's exact test p-values and BH-adjusted p-values.

Variant ID	PD	CTRL	p-value	Adj. p-value
chr2:207014657:TAAAA:T	18	1	0.0472	1.000
chr3:108102586:T:TA	29	4	0.0889	1.000
chr2:241696840:ATCC:A	42	8	0.134	1.000
chr7:44558268:CA:C	15	1	0.0974	1.000
chr5:118500981:TAA:T	22	3	0.155	1.000

PD, Parkinson's disease; CTRL, control; Adj. p-value, Benjamini-Hochberg-adjusted p-value.

Table S5. Enrichment of loci associated Parkinson's disease-related traits within GRBs, with corresponding binomial test p-values and BH-adjusted p-values. Loci included in this test overlap non-coding variants found in the exome sequence analysis.

Mapped trait	p-value	Adj. p-value
Insomnia measurement	0.0197	0.0737
Neuroticism measurement	6.86×10^{-9}	1.03×10^{-7}
Cognitive function measurement	1.04×10^{-6}	7.84×10^{-6}
Insomnia	0.00302	0.0151

Adj. p-value, Benjamini-Hochberg-adjusted p-value; GRB, genomic regulatory block.

Table S6. GRB target genes that are differentially expressed in CD4+ TCM and cDC2 cells. Complete differential expression statistics are provided in Supplementary Data 13.

Cell subtype	Genes
CD4+ TCM	<i>RPL7, ARHGAP15, RORA, PAG1, SORL1, DPYD, RUNX1, CITED2, S1PR1, BHLHE40, CXCR4, ZFP36L1, RBMS1, RUNX3, ZBTB20, ETV6, PPP3CA, BCL11B, MAP4K1, SOCS3, FOXN3, ARRDC3, SMARCA2, CELF2, MAF, ARHGEF1, SAMSN1, CD28, NR3C2, MBNL1, FYN, LRIG1, ARL15, RASGRP2, RUNX2, IRF2BP2, MID1IP1, FLII, EXOC4, CLTC, ARNTL, MBNL2, CALM2, NR1D1, FBXW7, CD40LG, CUX1, DNAJC1, MTIF3, PURA, RFTN1, GREM2, CCND2, CTNNB1, GATA3, XPO1, NFIA, ZBTB38, ZSWIM6, CDK4, CD200R1, MDGA1, RPA3, NFKB2, IFT57, RNF220, ACVR1, NOL4L, ALCAM, PPM1A, GCH1, TMEM167A, ERF, KLHDC3, LMO4, SAR1B, PBX3, CD63, TRIM8, PJA1, FAM53B, TM2D1, GLRX3, TLE4, SHQ1, MLLT3, AFTPH, PLXNA4, HS3ST3B1, DDX1, HMG3, ARID3B, OGFR1, KIF16B, RARG, VEGFB, HNRNPR, ARHGAP35, SLC44A1, C5orf24, ANKUB1, MEA1, LY75, CD83, CRIP1, RASGEF1B, SLC38A2, VMA21, JAZF1, ATP2A2, ZNF574, GBE1, ELL2, CDC42EP3, CHCHD7, POLR3A, PLXNC1, JADE1, SMIM20, EIF2B2, ZBTB16, SPRY1, CXCR5</i>
cDC2	<i>DPYD, FOS, ARHGAP15, MBNL1, RPL7, BCL11B, ZFP36L1, EXOC4, SAMSN1, ZBTB20, SOCS3, ANXA1, ETV6, FOXN3, NR4A2, MEF2C, FZD2, RBMS1, LEF1, CUX1, BHLHE40, FYN</i>

C Supplementary Data

The following supplementary datasets are available externally due to their size and/or because they have previously been published.

Supplementary Data 1. Quality control metrics and mapping results for SC-seq. The dataset is available as Supplementary Data 1 in Meglaj Bakrač et al. [155].

Supplementary Data 2. Complete list of validated SNP–gene associations. The dataset is available as Supplementary Table 3 in Mandić et al. [73].

Supplementary Data 3. Complete list of predicted enhancer–gene associations for all three computational methods. The dataset is available as Supplementary Table 5 in Mandić et al. [73].

Supplementary Data 4. Gene set enrichment analysis of all DEGs in HNSCC tumour versus controls. The dataset is available as Supplementary Table 3 in Mandić et al. [103].

Supplementary Data 5. Complete list of DEGs in HNSCC tumour versus controls. The dataset is available as Supplementary Table 2 in Mandić et al. [103].

Supplementary Data 6. Result of DM analysis on CpG sites. The dataset is available on Figshare repository <https://doi.org/10.6084/m9.figshare.24722268>.

Supplementary Data 7. Result of DM analysis on CGIs. The dataset is available as Supplementary Table 5 in Mandić et al. [103].

Supplementary Data 8. Gene set enrichment analysis of genes with DM promoters. The dataset is available as Supplementary Table 7 in Mandić et al. [103].

Supplementary Data 9. Gene set enrichment analysis of DEGs categorized based on their involvement with DM promoters, association with DEmiRs, or both. The dataset is available as Supplementary Table 8 in Mandić et al. [103].

Supplementary Data 10. Gene set enrichment analysis result on 11 clusters of integrated

DEGs. The dataset is available as Supplementary Table 9 in Mandić et al. [103].

Supplementary Data 11. Table of integrated DEGs DEmiRs and DM CGIs in TCGA dataset. The dataset is available as Supplementary Table 10 in Mandić et al. [103].

Supplementary Data 12. TCGA gene set enrichment analysis result on 12 clusters of integrated DEGs. The dataset is available as Supplementary Table 11 in Mandić et al. [103].

Supplementary Data 13. Complete list of differentially expressed genes between PD and CTRL in analysed cell subsets. The dataset is available as Supplementary Data 2 in Meglaj Bakrač et al. [155].

Supplementary Data 14. Complete list of gene set enrichment results in analysed cell types. The dataset is available as Supplementary Data 3 in Meglaj Bakrač et al. [155].