DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
**51**

AHTO SALUMETS

# Bioinformatics analysis of various aspects in immunology

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

**51**

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
**51**

**AHTO SALUMETS**

# Bioinformatics analysis of various aspects in immunology

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on March 6, 2024, by the Council of the Institute of Computer Science, University of Tartu.

*Supervisors*

Prof. Hedi Peterson
Institute of Computer Science,
University of Tartu,
Tartu, Estonia

Prof. Pärt Peterson
Institute of Biomedicine and Translational Medicine,
University of Tartu,
Tartu, Estonia

*Opponents*

Assoc. Prof. Benjamin Fairfax
Department of Oncology,
University of Oxford,
Oxford, United Kingdom

Assist. Prof. Can Kesmir
Department of Biology,
Utrecht University,
Utrecht, Netherlands

The public defense will take place on May 3, 2024, at 12:15 in Narva mnt 18-1021.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (PDF)
ISBN 978-9916-27-495-8 (print) ISBN 978-9916-27-496-5 (PDF)



University of Tartu Press
http://www.tyk.ee/

*"If you thought that science was certain –*
*well, that is just an error on your part."*
*— Richard P. Feynman*

# ABSTRACT

Bioinformatics has become an essential part in modern molecular biology research as advancements in several technologies such as sequencing and microarrays have led to the generation of vast amounts of data that require complex analysis. In my thesis, I present a collection of publications where we applied bioinformatics analysis on molecular biology data to gain insights into immunological processes. Firstly, to provide the necessary domain knowledge, I begin with a brief overview of molecular biology, epigenetic mechanisms, immunology, and aging. Subsequently, I introduce some of the main data science methods we used to gain new knowledge. The research we conducted encompasses a broad spectrum of topics, ranging from the analysis of various T cells to the investigation of COVID-19, which yielded several noteworthy findings. For instance, in one study, we observed a high degree of interindividual variation among the $CD8^+$ terminally differentiated effector memory (TEMRA) population and showed the feasibility to predict the levels of this clinically relevant cell type using DNA methylation. In another study, we found over 16,000 CpG sites that are differentially methylated between regulatory and conventional T cells, with differences in DNA methylation and gene expression in an autoimmune disease linked thyroid-stimulating hormone receptor (TSHR) gene. Yet in another study, we presented evidence supporting the hypothesis that late medullary thymic epithelial cells (mTECs) and Hassal's corpuscles contribute to the creation of a tonic inflammatory environment in the thymus which shapes the overall T cell repertoire. We also found notable similarities between medullary thymic epithelial cell differentiation and keratinocyte differentiation. Next, we found that COVID-19 is associated with apoptotic pathways and some of the apoptotic proteins such as hepatocyte growth factor (HGF) could be used as biomarkers to distinguish between different disease severities. In addition, our analysis indicated that asymptomatic individuals of SARS-CoV2 infection experience a long-term upregulation of inflammatory proteins. Overall, our research showcases the usefulness of bioinformatics in understanding complex biological processes.

# CONTENTS

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 5caC | 5-carboxylcytosine |
| 5fC | 5-formylcytosine |
| 5hmC | 5-hydroxymethylcytosine |
| 5mC | 5-methylcytosine |
| Ac-CoA | Acetyl coenzyme A |
| ACE2 | Angiotensin-converting enzyme 2 |
| AIRE | Autoimmune regulator |
| APC | Antigen-presenting cells |
| ARDS | Acute respiratory distress syndrome |
| CD | Cluster of differentiation |
| CM | Central memory |
| CMV | Cytomegalovirus |
| COVID-19 | Coronavirus disease 2019 |
| CTL | Cytotoxic T cell |
| DAMP | Damage-associated molecular pattern |
| DC | Dendritic cell |
| DMP | Differentially methylated position |
| DMR | Differentially methylated region |
| DNMT | DNA methyltransferase |
| DVP | Differentially variable position |
| EM | Effector memory cell |
| ER | Endoplasmic reticulum |
| FDR | False discovery rate |
| GD | Graves' disease |
| HAT | Histone acetyltransferase |
| HC | Hassall's corpuscles |
| HGF | Hepatocyte growth factor |
| ICU | Intensive care unit |
| IFN | Interferon |
| IL | Interleukin |
| LIPS | Luciferase signal-based immunoprecipitation system |
| MAE | Mean absolute error |
| MHC | Major histocompatibility complex |
| MSE | Mean squared error |
| mTEC | Medullary thymic epithelial cells |

| | |
|---|---|
| nano LC-MS/MS | Nano-scale liquid chromatography coupled to tandem mass spectrometry |
| NK | Natural killer cell |
| NPX | Normalized Protein eXpression |
| PAMP | Pathogen-associated molecular pattern |
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| PEA | Proximity extension assay |
| PRC2 | Polycomb repressive complex 2 |
| RBD | Receptor-binding domain |
| RMSE | Root mean squared error |
| RNApol | RNA polymerase |
| SAM | S-adenosyl methionine |
| SCM | Stem cell memory cells |
| SNP | Single nucleotide polymorphism |
| SNR | Signal-to-noise ratio |
| Tconv | Conventional T cell |
| TCR | T-cell receptor |
| TDG | DNA glycosylase |
| TEMRA | Terminally differentiated effector memory |
| TET | Ten-eleven translocation |
| Tfh | Follicular T helper cell |
| TGF | transforming growth factor |
| Th | T helper cell |
| TLR | Toll-like receptor |
| TMPRSS2 | Transmembrane serine protease 2 |
| TNF | Tumour necrosis factor |
| Treg | Regulatory T cell |
| TSHR | Thyroid-stimulating hormone receptor |
| TSS | Transcription start site |
| UTR | Untranslated region |
| WBC | Whole blood cells |

# LIST OF ORIGINAL PUBLICATIONS

## Publications included in the thesis

I. **Ahto Salumets**[*], Liina Tserel[*], Anna P. Rumm, Lehte Türk, Külli Kingo, Kai Saks, Astrid Oras, Raivo Uibo, Riin Tamm, Hedi Peterson, Kai Kisand, and Pärt Peterson. Epigenetic quantification of immunosenescent $CD8^+$ TEMRA cells in human blood. ***Aging Cell*** 21.5 (2022).

II. **Ahto Salumets**, Liina Tserel, Silva Kasela, Maia Limbach, Lili Milani, Hedi Peterson, Kai Kisand, and Pärt Peterson. Graves' disease-Associated TSHR gene is demethylated and expressed in human regulatory T cells. ***BioRxiv*** (2022).

III. Martti Laan, **Ahto Salumets**, Annabel Klein, Kerli Reintamm, Rudolf Bichele, Hedi Peterson, and Pärt Peterson. Post-Aire medullary thymic epithelial cells and Hassall's corpuscles as inducers of tonic pro-inflammatory microenvironment. ***Frontiers in Immunology*** 12 (2021).

IV. Liis Haljasmägi[*], **Ahto Salumets**[*], Anna Pauliina Rumm[*], Meeri Jürgenson, Ekaterina Krassohhina, Anu Remm, Hanna Sein, Lauri Kareinen, Olli Vapalahti, Tarja Sironen, Hedi Peterson, Lili Milani, Anu Tamm, Adrian Hayday, Kai Kisand, and Pärt Peterson. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. ***Scientific Reports*** 10.1 (2020).

V. Liina Tserel, Piia Jõgi, Paul Naaber, Julia Maslovskaja, Annika Häling, **Ahto Salumets**, Eva Zusinaite, Hiie Soeorg, Freddy Lättekivi, Diana Ingerainen, Mari Soots, Karolin Toompere, Katrin Kaarna, Kai Kisand, Irja Lutsar, and Pärt Peterson. Long-term elevated inflammatory protein levels in asymptomatic SARS-CoV-2 infected individuals. ***Frontiers in Immunology*** 12 (2021).

[*] Authors contributed equally.

## Publications not included in the thesis

VI. Jon Ison, Hervé Ménager, Bryan Brancotte, Erik Jaaniso, **Ahto Salumets**, Tomáš Raček, Anna-Lena Lamprecht, Magnus Palmblad, Matúš Kalaš, Piotr Chmura, John M Hancock, Veit Schwämmle, Hans-Ioan Ienasescu. Community curation of bioinformatics software and data resources. ***Briefings in Bioinformatics*** 21.5 (2020).

VII. Jon Ison, Hans Ienasescu, Piotr Chmura, Emil Rydza, Hervé Ménager, Matúš Kalaš, Veit Schwämmle, Björn Grüning, Niall Beard, Rodrigo Lopez, Severine Duvaud, Heinz Stockinger, Bengt Persson, Radka Svobodová Vařeková, Tomáš Raček, Jiří Vondrášek, Hedi Peterson, **Ahto Salumets**, Inge Jonassen, Rob Hooft, Tommi Nyrönen, Alfonso Valencia, Salvador Capella, Josep

Gelpí, Federico Zambelli, Babis Savakis, Brane Leskošek, Kristoffer Rapacki, Christophe Blanchet, Rafael Jimenez, Arlindo Oliveira, Gert Vriend, Olivier Collin, Jacques van Helden, Peter Løngreen, and Søren Brunak. The bio.tools registry of software tools and data resources for the life sciences. ***Genome Biology*** 20.1 (2019).

# 1. INTRODUCTION

Bioinformatics is an interdisciplinary field that encompasses the development and application of computational tools and statistical methods to understand the biological phenomena underlying the data. With the advancements in sequencing and microarray technologies, mass spectrometry, and proteomics assays, researchers are now able to collect vast amounts of data that require complex analysis. As a result, bioinformatics has become an essential part of molecular biology research, providing crucial insights into underlying biological processes.

In this thesis, I aim to provide a summary of our studies that utilized bioinformatics and data science methods to address various immunological problems. The five papers included in this thesis focus on a range of immunological issues, from age-related changes in the immune system to the characterization of aspects related to Coronavirus disease 2019 (COVID-19) infection. Throughout these articles, we applied a wide range of bioinformatics-related methods to understand the data obtained.

The thesis is structured in a way that the first two chapters provide a general understanding of the work, while the last four chapters offer more detailed information on the study design and findings, as well as the context in which the work was conducted. In Chapter I, I will guide the reader through the fundamental knowledge of genetics, gene regulation, the immune system, and aging to provide a foundation for understanding the work done in the publications. In Chapter II, I will delve into the data generation and preprocessing, and then continue with various analysis methods used in the papers, including hypothesis testing, correlation analysis, and unsupervised and supervised learning techniques. Finally, in later chapters, I will summarise the articles included in my thesis, with an aim to familiarise the reader with the incorporated research.

In Paper I, we focused on the characterization of different $CD4^+$ and $CD8^+$ T cell subpopulations, with a particular interest in terminally differentiated effector memory (TEMRA) cells that are considered to be involved in various age-associated adverse health outcomes. We aimed to harness DNA methylation to study whether modelling $CD8^+$ TEMRA cell levels using DNA methylation levels in particular CpG sites is a feasible solution. In addition, we studied associations of those $CD4^+$ and $CD8^+$ T cells with aging, cytomegalovirus infection, and inflammatory proteins.

In Paper II, we explored DNA methylation differences between regulatory (Tregs) and conventional T cells (Tconvs). We noticed that the thyroid-stimulating hormone receptor (TSHR) gene, a known risk locus for Graves' disease, is differentially methylated between those two cell types and its expression is characteristic of Tregs among other $CD4^+$ T cells. This led us to form a new hypothesis regarding its involvement in Graves' disease (GD) and subsequent methylation profiling of GD patients and healthy individuals.

In Paper III, we studied the key cell type involved in T cell development,

specifically medullary thymic epithelial cells (mTECs) that are crucial for creating central tolerance by making self-antigens available to developing T cells. While those cells are extensively studied, their differentiation, particularly in late stages, is often overlooked. To this end, we analyzed proteomes of late stages of mTECs and draw parallels with keratinocyte differentiation as many other studies have pinpointed similarities between those two cell types concerning the proteins they express.

Lastly, the final two papers in this thesis focus on COVID-19. In the first of these studies, we analyzed longitudinally obtained data on blood inflammation markers, antibodies, and plasma proteins of COVID-19 patients with different clinical manifestations. Our goal was to identify biomarkers that could help distinguish between different classes of disease severity. The second study focused on mostly asymptomatic individuals and reanalyzed them after 7-8 months of infection. We explored the persistence of antibody levels and T cell responses, as well as how the levels of inflammatory proteins changed over time.

# 2. DOMAIN KNOWLEDGE

This thesis is based on a collection of articles that are focused on solving open questions in immunology with the help of computational analysis. More precisely, the work presented here covers bioinformatics analysis of various aspects of molecular biology in the field of immunology. Therefore, to comprehend the work done, one needs to attain a certain level of domain knowledge. Firstly, I will give a brief introduction to molecular biology, molecular genetics, and gene regulation. However, I will put more emphasis on DNA methylation, an epigenetic phenomenon necessary for gene regulation, as this has been studied in two of the articles included in the thesis. After that, I will introduce the immune system and take a more in depth look at T cells as they play a major role in three of the articles but have a minor part also in the other two articles. Finally, I will discuss aging and its connection to the immune system as this forms a crucial part of paper I. But for now, let us dive into molecular biology.

## 2.1. Fundamentals of molecular biology

Cells are generally considered the smallest units of life since they are the simplest functional and structural components that make up all organisms (except viruses and subviral agents). It is speculated that there are somewhere between $10^{13}$ to $10^{14}$ cells in the adult human body [1]. Each of the cells originates from other cells in a process where the mother cell divides usually into two identical cells. We know that the DNA present in their nucleus is identical (again, with a few exceptions). However, when looking at an organism, we observe extraordinary diversity of cells. We see that they differ vastly in their shape, size, lifespan, and function. For example, the average diameter of a eukaryotic cell is around 10-40 $\mu$m but individual human neurons can extend up to 1 meter in length. Regarding the lifespan, it has been found that neutrophils live around 5-6 days while less than 50% of cardiomyocytes are replaced during a person's lifetime. To stay alive, our body needs to perform many complex processes such as food digestion, movement, defense against pathogens, thinking and planning actions, etc. All those functions are carried out by different cells that are grouped together forming tissues and organs. Histologically, the common knowledge has been that there are over 200 different cell types in the human body but single-cell studies have already shown us that it is a considerable understatement [2]. This specialization is acquired by a process called cellular differentiation which essentially means highly controlled modulation of gene expression. Therefore, we reach molecular genetics, a discipline that forms the basis of most molecular biology research including the articles presented in this thesis. It is a study of molecular mechanisms by which genes are expressed and regulated [3].

## 2.2. Molecular genetics

Molecular genetics is essentially a study of an interplay between three classes of biological macromolecules – DNA, RNA, and proteins. Their relationships constitute the central dogma of molecular biology that illustrates the main flow of genetic information. It starts with DNA which is responsible for storing hereditary information. The DNA itself is a deoxyribonucleic acid consisting of 4 types of nucleotides – adenine (A), thymine (T), cytosine (C), and guanine (G). It is usually in the form of a double helix and in canonical Watson-Crick base pairing, A forms a base pair with T using two hydrogen bonds, and C and G are held together by three hydrogen bonds. In eukaryotes, the DNA is stored in the cell's nucleus, and in humans it consists of 23 pairs of linear sequences of DNA called chromosomes that make up our genome. The size of the haploid genome, a term coined to denote the unique 23 chromosomes, consists of 3.1 billion base pairs. The DNA is transmitted from the mother cell to the daughter cells via cell division. Preceding to the division, a process called DNA replication takes place, where DNA is duplicated to ensure that each of the daughter cells ends up with a complete genome. DNA contains functional sections called genes which in humans are sparsely distributed across the DNA. One of the essential processes that are common in all cellular organisms is transcription in which a gene on a DNA strand is used as a template by an RNA polymerase to synthesize RNA. RNA is very similar to DNA but T is replaced by uracil (U) and RNA is mostly single-stranded. There are two major types of RNA – protein coding RNA (mRNA) and non-coding RNA. The estimates of genes vary due to difficulties in gene definitions (e.g., conjoined genes), but stays around 20 000 for protein-coding genes, and there exists at least a similar amount of non-coding RNA genes. The mRNA is subjected to another essential process common among all organisms – translation. Translation takes place at ribosomes where a group of 3 bases of RNA (a codon) is read at a time specifying a particular amino acid – a building block of a protein. During this process, amino acids are linked together ultimately forming a chain of amino acids called proteins. Proteins in turn perform a broad range of different roles in the organism from providing our cells a structure to enabling us to respond to different stimuli. The non-coding RNA also has a variety of roles to play although most of them are involved in regulating their specific target genes [4].

Although the central dogma, which typically refers to the unidirectional flow of genetic information from DNA, to RNA, to proteins, covers the main aspects of biological information flow, there are notable exceptions to this concept. For example, research on retroviruses has demonstrated that cellular reverse transcriptase also exists, enabling a reverse flow of information from RNA to DNA. This phenomenon is not limited to viruses and has been found to function in humans, playing a critical role in replicating DNA sequences located at the end of chromosomes, known as telomeres. Additionally, the discovery of RNA-dependent RNA polymerases has expanded our understanding of genetic information flow,

demonstrating that RNA replication is indeed possible [4].

## 2.3. Overview of transcription

As the central dogma states, genetic information flows from DNA to RNA via transcription. Transcription takes place in the nucleus and in this process, one strand of DNA is used as a template to synthesize RNA via RNA polymerase. Firstly, the DNA double helix needs to unwind locally to allow all the relevant proteins to gain access to this DNA strand. The resulting RNA transcript has the same 5' $\rightarrow$ 3' direction as the non-template DNA strand, which is also referred to as the coding strand. However, the RNA transcript replaces the nucleotide T with U. But, before the RNA polymerase can bind, other proteins need to be bound to specific DNA regions. One of the most important regions is called the promoter which is usually located upstream of the gene but nearby. This is the place where many transcription factors bind that unwind the DNA and guide the RNA polymerase. All in all, those factors form the so-called pre-initiation complex that initiates the transcription of the underlying gene. There are many types of transcription factors, some are general ones, that act in every cell while others are tissue or even cell type specific. In addition to promoter regions, many other DNA regions are involved in transcription, for example, enhancers that as the name implies enhance the transcription. They are usually located far from the gene that they are regulating and serve as a place where many tissue and cell type specific transcription factors bind. Their effect is achieved via looping out of the DNA between the promoter and enhancer and bringing those two DNA regions together. This allows the transcription factors on the enhancer to interact with the proteins on the promoter. This interaction leads to transcription which in turn results in cell-specific gene expression. In addition, there are other types of DNA regions such as silencers that work similarly to enhancers but instead inhibit transcription [4].

## 2.4. The structure of the chromatin

Now coming back to the DNA. As it was said, the DNA of a eukaryotic cell is located in the nucleus and is distributed between multiple linear chromosomes. Each of the chromosomes occupies a distinct territory in the nucleus and is also highly organized to enable sophisticated gene expression. The first layer of DNA packaging involves coiling it around histones. Histones are proteins that package DNA. There are four core histones (H2A, H2B, H3, H4), and DNA is wrapped around eight histone proteins (two of each of the four core histones) forming a nucleosome. Adjacent nucleosomes are connected to one another by a short stretch of linker DNA, where a different type of histone, H1, binds. The entire DNA–protein complex is referred to as chromatin. Generally, chromatin is divided into euchromatin and heterochromatin with the latter split into constitutive and facultative

heterochromatin. The euchromatin is more in an extended state compared to heavily condensed heterochromatin. The constitutive heterochromatin includes permanently "closed" regions on DNA that are usually gene-poor areas that contain highly repetitive sequences. On the other hand, facultative heterochromatin can be reversed and is exemplified by an X chromosome inactivation. In females, one of the two X chromosomes is randomly inactivated in every somatic cell so that it becomes fully heterochromatic, however, in certain stages of meiosis, it is decondensed and fully activated. The euchromatin forms still the majority of DNA, in its very extended form, called "open chromatin", it allows genes to be expressed, however, when the adjacent nucleosomes become more tightly packaged, the RNA polymerases and transcription factors no longer can access the DNA and thus those genes are no longer capable of being expressed. Therefore, the open and condensed euchromatin determines which genes are active and this eventually determines the identity and status of the cell. The (de)condensation of chromatin is regulated by epigenetic mechanisms [3].

## 2.5. Epigenetics

Epigenetics is the study of biological processes and mechanisms that alter gene activity but do not involve changes in the genetic sequence. In addition, it is closely related to a term often used in the scientific literature – epigenomics, which aims to understand the genome-wide distribution of epigenetic changes. The main mechanisms of epigenetics are histone modifications and DNA methylation. Those two do not work in isolation, instead, there is a complex interplay between them. Let us first focus on histone modifications.

### 2.5.1. Histone modifications

Histone modifications play a vital role in regulating the structure of the chromatin. Namely, the N-terminal tails of histones undergo various modifications which in turn affect the condensation level of the chromatin. To be more specific, the tails of the core histones are accessible to other proteins that modify specific amino acids in their tails, for example, by adding methylation or acetylation groups to lysine. For instance, in the case of acetylation, the mechanism of action is dependent on changing the overall charge of the histone tail from positive to neutral. Namely, added negatively charged acetyl group neutralizes the positive charge on lysine in the histone tail. This leads to reduction in affinity that holds histone proteins and negatively charged DNA together, therefore the addition of the acetyl group makes DNA more accessible to transcription factors. This in turn activates gene expression of underlying genes. Deacetylation, the removal of the acetyl group, achieves the opposite effect, as it makes DNA more tightly wrapped around histones and makes a given region much harder to access for transcription factors [5].

Of course, biology is complicated and besides the aforementioned mechanisms, there is actually a vast array of proteins that modify histones. They are grouped into writers, erasers, and readers. Writers add groups (e.g., histone methyltransferases and histone acetyltransferases), erasers remove groups (histone demethylases and deacetylases) and readers bind specifically modified residues and conduct certain actions, e.g., chromodomain proteins bind methylated histones and bromodomain proteins bind acetylated lysines [5]. The effects of histone acetylation and methylation are illustrated in Fig 1.

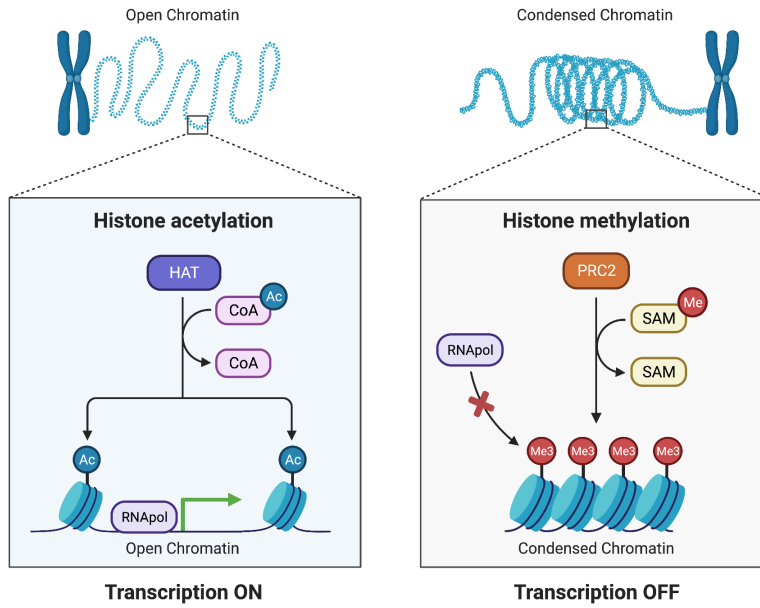


**Figure 1.** Example of histone modifications' effect on gene expression. Histone acetylation, a process carried out by histone acetyl transferases (HATs), leads to gene expression by making DNA accessible to RNA polymerase (RNApol) and its associated transcription factors (not shown). On the contrary, histone methylation, carried out by a protein complexes that have histone methyltransferase activity such as Polycomb Repressive Complex 2 (PRC2) shown here, causes repression of transcription via chromatin closing and consequent prevention of RNA polymerase binding. Acetyl coenzyme A (Ac-CoA) and S-adenosyl methionine (SAM) serve as acetyl and methyl group donors, respectively. This figure was made with Biorender.com by using template "Epigenetics and Gene Expression".

As was said previously, histone modifications do not act in isolation but act in concert with DNA methylation. For example, methylated DNA can recruit different protein complexes that modify histone tails and regulate chromatin structures[6]. Even vice versa is possible, modifications on histone tails can influence DNA methylation [7]. Let us take a closer look at DNA methylation.

### 2.5.2. DNA methylation

DNA methylation is probably the best-known epigenetic mechanism since it has been the easiest to study. DNA methylation refers to a process in which a methyl group is added from S-adenosyl methionine (SAM) to the 5th carbon of the cytosine thus resulting in 5-methylcytosine (5mC) [8]. 5mC is the major chemical modification of DNA among all vertebrates accounting for approximately 1% of genomic DNA in human somatic cells [9].

Studies have shown that methylation of cytosine happens mostly in the so-called CpG context where cytosine is followed by guanine [10]. It has been estimated that there are $\sim$29 million CpGs in the human genome with 60-80% of them being methylated [11]. Although it should be noted, that on a lesser scale, DNA methylation takes place in other contexts as well (C followed by A, C, or T). For example, this has been observed in embryonic stem cells [10, 11]. Nevertheless, as it mostly happens in the CpG context, it produces a nice property. More precisely, as DNA is antiparallel and C binds to G, it means that two methylated cytosines are situated diagonally to each other on opposing strands of DNA – a property that is very important for maintaining DNA methylation over cell divisions, but more on that later. The fact that DNA methylation happens mostly in the CpG context is also the reason why the majority of studies focus only on the methylation levels in CpG sites.

Regarding the function of 5mC, it does not affect base-pairing but it can physically block the binding of transcription factors [12]. However, most importantly, it regulates gene expression by recruiting other proteins that modify chromatin accessibility and gene expression [13]. Mechanistically, the 5mC is located in the major groove of DNA where methyl groups point towards the exterior of the double helix. Since a certain group of proteins called "methyl binding proteins" constantly scan the major groove of DNA, they can be recruited once they interact with 5mC [13]. Those "methyl binding proteins" also cooperate with other chromatin modifying proteins such as histone deacetylases and methylases which after the protein complex has bound, carry out their functions [14]. In general, DNA methylation is usually observed in condensed chromatin and is associated with gene silencing. However, its functions are still somewhat location specific.

There are two main types of CpG location categories – genetic location and relation to CpG islands (Fig 2). CpG islands are regions that show strong enrichment of CpG dinucleotides and are usually a few hundred to several thousand base pairs long. It has been estimated that roughly 7% of all CpGs lie in CpG islands [15, 16]. Regarding the first type of CpG location category, the genetic location shows where the CpG lies with respect to a gene, e.g., in the promoter or in the gene body. In practice, this category is further specified. For example, the promoter area is divided into two parts relative to the place where the mRNA transcription starts denoted as the transcription start site (TSS). The first part corresponds to a distance of 200–1500 bases upstream of the TSS (termed TSS1500) and the sec-

ond 0–200 bases upstream of the TSS (TSS200). In addition, many other regions are distinguished in this category including regions that correspond to untranslated areas in the 5' and 3' end, i.e., sections on mRNA that are not translated by ribosomes (5' UTR and 3' UTR, correspondingly), gene body (sequence between translation start and end site), and areas between genes called intergenic regions. Regarding the relation to CpG islands, 6 regions are usually studied – N shelf, N shore, CpG island, S shore, S shelf and open sea. Shore stands for the region up to ∼2kb and shelf up to ∼4kb from CpG island. The letters N and S denote "north" and "south" that correspond to the 5' and 3' locations relative to the CpG island, and the open sea corresponds to the rest of the genome.

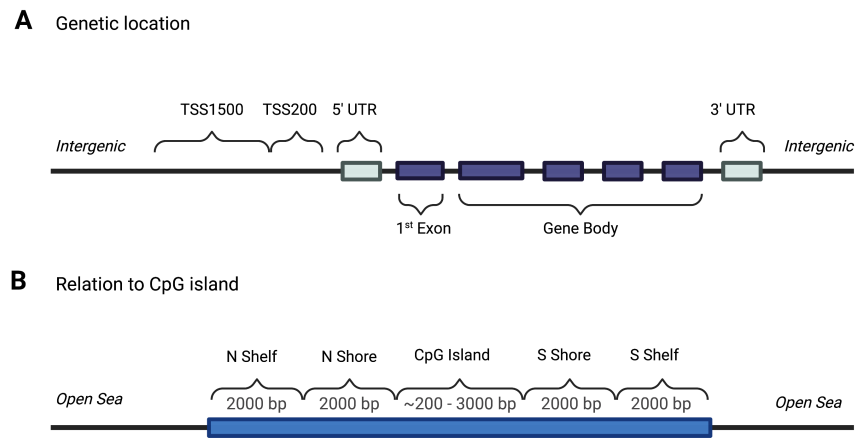


**Figure 2.** Illustration of the main types of CpG location categories – the genetic location (A) and the relation to CpG islands (B). The first type, shown in (A), indicates the location of CpG relative to genetic areas such as untranslated regions (UTRs), gene body, and intergenic regions. The promoter area is further divided into different regions based on the transcription start site (TSS) – 0–200 and 200–1500 bases upstream of the TSS (called TSS200 and TSS1500, respectively). The second way, shown in (B), illustrates CpG location relative to CpG-rich regions in the genome called CpG islands. This includes six regions: N shelf, N shore, CpG island, S shore, S shelf, and open sea. "Shore" refers to regions up to approximately 2kb from the CpG island, and "shelf" refers to regions up to approximately 4kb from the CpG island. "N" (north) and "S" (south) indicate the 5' and 3' locations relative to the CpG island, while "open sea" refers to the remainder of the genome. This figure was made with Biorender.com

Regarding the genetic location, within intergenic regions, the main role of DNA methylation is to repress the expression of potentially harmful genetic elements such as transposable and viral elements. Its role there is hard to overstate

as approximately 45% of the mammalian genome consists of transposable and viral elements that are silenced by bulk methylation [17]. If those elements get expressed, their replication and insertion can lead to gene disruptions and mutations thus bringing along detrimental consequences to the organism. DNA methylation acts as a silencer also in promoters, particularly in those that have high GC content and are associated with CpG islands. In fact, over 70% of promoters are like that [18]. Hence, CpG islands close to transcription start sites in active genes remain unmethylated. Intriguingly, however, DNA methylation in the gene body is associated with a higher level of gene expression [19]. In addition, some studies indicate that DNA methylation in gene bodies and shores have tissue-specific changes [20, 21].

Regarding the molecular machinery by which DNA methylation is established, maintained, and removed, many aspects remain to be elucidated. However, it is known that DNA is methylated by enzymes called DNA methyltransferases (DNMTs) that in humans include DNMT1, DNMT3A and DNMT3B [22]. Additionally, there are two DNMT family proteins that do not have catalytic activity – DNMT2 and DNMT3L [22]. Usually, DNA methylation is divided into two categories – maintenance methylation and *de novo* methylation (Fig 3). The maintenance DNA methylation, also known as inheritance DNA methylation, maintains DNA 5mCs following DNA replication while *de novo* methylation stands for a process by which new methyl groups are introduced to DNA. In the case of inheritance DNA methylation, the DNMT1, also widely regarded as the maintenance methyltransferase, "reads" the methylation marks on the original strands and adds methyl groups to the opposite, newly synthesized strand, therefore harnessing the CpG context property mentioned above. Previously, it was thought that DNMT1 itself can maintain established methylation patterns over cell divisions but given the current evidence, it is very likely that DNMT3A and DNMT3B, usually referred to as *de novo* methyl transferases, also play a role in methylation maintenance [23].

The *de novo* DNA methylation is a more complicated process and is still poorly understood. It has been established that DNMT3A, B and L are involved in this process together with many repressive transcription factors. In addition, in many cases, it seems that certain histone modifications are prerequisites for *de novo* DNA methylation. For example, it has been shown that DNMT3L interacts with unmethylated H3K4 (denotes 4th amino acid in the tail of histone protein H3) and then recruits other DNMT3s to establish *de novo* methylation [7]. Also, studies have shown that DNMT3s can be recruited by other epigenetic repressors such as histone deacetylases, H3K9 methyltransferases, and repressive transcription factors [15, 24]. In addition, many other mechanisms exist, for example, Piwi interacting RNAs, the small non-coding RNA molecules, have been observed to recruit DNMT3A and DNMT3B and increase global DNA methylation in multiple myeloma cells [25]. Therefore, there exist many different mechanisms by which different genomic regions are targeted by *de novo* DNMTs and the exact

mechanisms largely remain elusive.

There are two major ways in how DNA methylation can be erased – passive and active demethylation (Fig 3). The first, passive demethylation, is just a result of dysfunctional maintenance of DNA methylation that causes dilution of 5mCs during DNA replication. The second, active, is mediated by ten-eleven translocation (TET) family enzymes. It should be noted that unlike histone-modifying proteins there are no DNA demethylases per se. Instead, TET enzymes convert 5mC into 5-hydroxymethylcytosine (5hmC) that in turn converts to 5-formylcytosine (5fC) and finally into 5-carboxylcytosine (5caC). Both 5fC and 5caC get excised by thymine DNA glycosylase (TDG) and the given DNA position gets fixed by DNA repair mechanisms called base excision repair [26]. In addition, adding further complexity, there is evidence that 5hmC is not merely an intermediate step in DNA demethylation but is a stable epigenetic mark with regulatory functions [27]. The DNA methylation and demethylation are summed up in Fig 3.

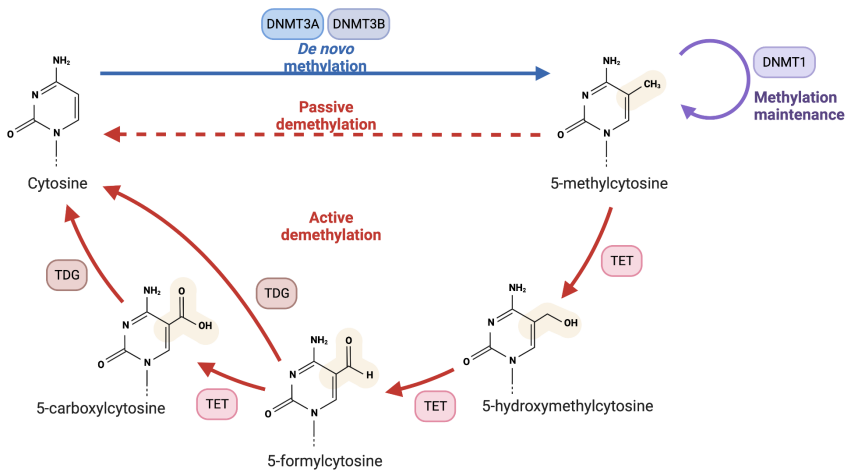


**Figure 3.** Overview of DNA (de)methylation. According to the classical model, 5-methylcytosines (5mCs) are added by DNMT3 enzymes and they are maintained over cell divisions by the enzyme DNMT1. DNA can be demethylated passively or actively using ten-eleven translocation (TET) enzymes that oxidize 5-methylcytosine (5mC) to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Those get removed from DNA by the thymine DNA glycosylase (TDG) and later those positions will be fixed by DNA reparation mechanism that restore a cytosine in its original location. This figure was made with Biorender.com by using template "DNA methylation".

Considering the information above, it is clear that DNA methylation is a rather prevalent modification that plays role in many key physiological processes such as in X-chromosome inactivation, repressing transposable elements, and regulating

gene expression. Therefore, it is no surprise that perturbations in DNA methylation have often been associated with many human diseases, including autoimmune and neurological diseases, cancers as well as aging in general [28]. In addition, it is also evident that a great deal remains unknown regarding how it is established and maintained, as well as its specific functions in different locations of the genome. In the next chapter, I introduce how DNA methylation is measured, how the data looks like, and how to work with such data. However, before that, I also give a brief overview of the immune system as it is the underlying theme of this thesis.

## 2.6. Immune system

Our immune system is a complex network of proteins, cells, tissues, and organs that are distributed across the organism and work together to protect us from pathogens anywhere in the body. At the same time, this system avoids responses against our own healthy cells, beneficial microbes constituting our commensal microflora, and limits responses in such a way that no excessive damage is done to self-tissues.

The immune system is divided into two parts – the innate and adaptive immune system. The innate immune system forms the first line of defense and responds quickly as the mechanisms are already in place before the infection. Typically, it protects us within the first few hours to a few days before the adaptive immune system takes over. Innate immunity relies on a limited set of receptors that recognize parts of pathogens, and it lacks memory in a classical sense. While slower to respond, the adaptive immune system is more specialized. It is named adaptive immunity because it develops and adapts during infection and becomes more capable with each successful exposure to a given pathogen. Hence, it is also characterized by immune memory, meaning that the next time a known pathogen is encountered, it responds faster and protects the organism better. All of this is achieved via a vast array of receptors that can specifically target given pathogens [29]. Both of those systems work closely together, and actually, they are not so distinct as there are cell types having similarities with both of them thus sitting somewhere in between [30]. Now, let us take a closer look at those systems.

The innate immune system deals with infection by causing acute inflammation that is characterized by the accumulation of immune cells, plasma proteins, and fluid into the site of infection. The innate immune system consists of three main components. Firstly, there are physical and chemical barriers such as skin and mucous membranes, that prevent pathogens from entering the body. Secondly, there are immune cells that detect and eliminate pathogens but also remove dying and damaged host cells. Thirdly, there are specialized proteins that mark the pathogens for phagocytic cells, attract other immune cells into the site of infection, and destroy bacterial cell walls and viral envelopes. The innate immune response is activated by the recognition of pathogen-associated molecular patterns

(PAMPs) and damage-associated molecular patterns (DAMPs). The former corresponds to structures and products of bacteria and viruses that are often essential for their survival and the latter denotes molecules that are released by damaged and dying cells. Those molecular patterns are recognized by less than 100 different receptors that are located on the surface as well as inside the cells. It has been estimated that those receptors can recognize roughly 1000 different pathogen and cellular damage-associated molecules. When those receptors bind to their ligand, signalling pathways promoting proinflammatory and antiviral/antimicrobial activity get activated. However, the exact response is cell type specific. In fact, there are many cell types belonging to the innate immune system. For example, there are phagocytic cells such as neutrophils and macrophages that are recruited to the site of infection and devour and destroy microbes. In addition, via secreting cytokines they also communicate with other cells and regulate the overall immune response. Then there are mast cells that once activated release inflammatory mediators, promote acute inflammation and dilate blood vessels. Natural killer (NK) cells induce apoptosis of target cells after activation. Their activation is determined by the integration of signals generated by activating and inhibitory receptors that recognize ligands of infected and healthy cells, correspondingly. There are also dendritic cells (DCs) that form a bridge between innate and adaptive immunity. More precisely, those cells are located in tissues, especially in those that are in close contact with the external environment, and they harbour a diverse set of pattern recognition receptors. They take up pathogens, migrate from the site of infection to the secondary lymphoid organs and present antigens to T cells and thereby induce the adaptive immune response. Collectively, those cells that capture and present antigens to lymphocytes to initiate the adaptive immune response are called antigen-presenting cells (APCs) and the DCs are considered the most efficient in this class [31].

Adaptive immune response is mediated by lymphocytes – B and T cells. Unlike innate immune cells, they have a highly diverse set of receptors potentially capable of recognizing any known and unknown pathogen. Also, differently from innate immunity, its defining feature is memory meaning that the second exposure to the same antigen results in a faster and greater immune response [29].

The maturation of lymphocytes takes place in primary lymphoid organs, B cells develop in the bone marrow and T cells in the thymus. The antigen specificity and receptor diversity are achieved via the rearrangement of variable (V), diversity (D) and joining (J) gene segments in a process called V(D)J recombination. This process involves many stages that include the removal of non-functional receptors as well as elimination of potentially autoreactive lymphocytes (this step in T cells is covered in Paper III) and ends with naïve lymphocytes entering to the secondary lymphoid organs [32].

The adaptive immune response is triggered by antigens, which in the case of T cells are primarily short peptides and in the case of B cells constitute various biomolecules including peptides, folded proteins, nucleic acids, etc. Typically, the

process starts with APCs presenting antigen to naïve lymphocytes, i.e., cells that express antigen recognizing receptors but have not encountered antigen yet. This activation leads to clonal expansion of those antigen-specific lymphocytes. Subsequently, those cells differentiate into effector cells that eliminate or neutralize the pathogen and into memory cells that elicit an immune response when organism is attacked by the same pathogen later. This activation and differentiation take a few days and thus is a reason for the slower response time. Once the pathogen is eliminated, most of the effector cells die but memory cells remain [29].

Adaptive immunity is divided into humoral and cell-mediated immunity. Humoral immunity corresponds to B cell specific antibody-mediated defense mechanism that mainly targets extracellular pathogens but plays a role in clearing viruses as well. More specifically, after B cell activation, most B cells differentiate into plasma cells that secrete large quantities of antibodies but some of the B cells become memory cells instead. Antibodies play many different roles, for example, they neutralize pathogens by attaching directly to their surface and thus inhibiting their functioning. Once they are bound to their target, they also enhance the activity of other immune cells as they can better destroy pathogens marked with antibodies. And finally, antibodies also activate other immune response-associated proteins, such as complement system proteins. On the other hand, cell-mediated immunity is mediated by T cells and plays a major role in eliminating intracellular pathogens such as viruses since they kill infected cells. Given that T cells are the central focus of this thesis, I will delve more extensively into their details. It is important to note that different classes of T cells as well as different differentiation stages can be distinguished by surface molecules on cell membranes that are designated by cluster of differentiation (CD) numbers [29].

There are two major classes of T cells denoted as $CD4^+$ and $CD8^+$ that function very differently. $CD4^+$ T cells, also called T helper (Th) cells, as the name implies, help other immune cells by regulating their activity through the release of cytokines. For example, they activate macrophages and B cells and support the activation of $CD8^+$ T cells. On the other hand, $CD8^+$ T cells, also known as cytotoxic T cells (CTLs), kill infected and tumor cells. Differences in the responses of these two T cell types are determined by the receptors they bind. Specifically, T cells have a specific antigen recognition system, where $CD4^+$ T cells are restricted to major histocompatibility complex (MHC) class II molecules, primarily expressed on professional APCs, such as dendritic cells, macrophages, and B cells. These APCs phagocytize pathogens and present fragments of them on their cell surfaces through MHC II molecules. Then, $CD4^+$ T cells recognize these antigens and coordinate immune responses. In contrast, $CD8^+$ T cells recognize antigens presented by MHC I molecules expressed on virtually all nucleated cells. These antigens typically come from intracellular proteins, including viral or tumor proteins. Therefore, when a cell presents such antigens on MHC I molecules, $CD8^+$ T cells can recognize these antigen-MHC I complexes and eliminate the infected or abnormal host cells [33].

Both $CD4^+$ and $CD8^+$ T cells have various subtypes. For example, naïve $CD4^+$ T cells can differentiate into different types of effector cells, including Th1, Th2, Th9, Th17, Th22, regulatory T cells (Tregs), and follicular helper T cells (Tfh). All of those, except Tregs, are characterized by releasing different sets of cytokines and cause different immune reactions that tackle different types of pathogens. On the contrary, Tregs suppress unwanted immune responses thus helping the organism to avoid autoimmunity. Tregs are more thoroughly covered in Paper II where they are compared to conventional T cells (Tconvs), a term coined to denote the rest of the $CD4^+$ T cells. In addition to differentiation into effector cells, two types of memory cells exist – effector memory (EM) and central memory (CM), with the first denoting memory cells that remain in recently infected tissues while the second corresponds to memory cells in secondary lymphoid organs. Regarding the different subtypes of $CD8^+$ T cells, there are naïve $CD8^+$ T cells that can differentiate into stem cell memory cells (T SCM), T CM, T EM, and T effector cells [34]. Moreover, during aging, many different CD surface molecules appear or disappear thus expanding the number of different subsets even more. This is further addressed in Paper I.

The differentiation of T cells is facilitated by a large and diverse group of small proteins called cytokines. Those proteins are involved in cellular communication and they play a vital role in various biological processes, including regulating cell growth, survival, and differentiation. Their significance extends to other immune cells as well as many other cell types such as endothelial cells and fibroblasts. In fact, all immune cells are known to release at least some cytokines and have receptors that can recognize cytokines secreted by others thus enabling them to respond. In terms of inflammation, there are proinflammatory cytokines such as interleukin (IL)-1, IL-6, and tumour necrosis factor (TNF)-$\alpha$ that cause fever and inflammation but there are also anti-inflammatory cytokines such as IL-10 that have immunosuppressive properties. In addition, there is a special subset of cytokines called chemokines that help to guide the movement of immune cells throughout the body [29, 31]. Due to their involvement in a vast array of different processes and pathways, differences in their levels have been implicated in various infections and diseases. Therefore, many cytokines are considered clinically meaningful as they or a combination of them can be used as biomarkers for different health statuses or to estimate disease severity [35, 36]. In Papers IV and V, we studied the levels of cytokines in coronavirus disease 2019 (COVID-19). However, since Paper I delves into the age-related changes in the immune system, I will also touch upon this topic briefly.

### 2.6.1. Immune system and aging

Aging is a time-dependent functional decline of an organism, and it is the main risk factor for many diseases including cancer, diabetes, cardiovascular and neurodegenerative diseases, and poor vaccine efficacy. A review article published in

2013 by López-Otín et al. [37] proposed nine common denominators of aging, which can be considered the fundamental hallmarks of aging. Those are:

I. genomic instability – accumulation of genetic damage such as point mutations over time
II. telomere attrition – shortening of telomeres,i.e., the terminal ends of linear DNA molecules
III. epigenetic modifications – age-related changes in chromatin structure caused by changes in DNA methylation and activity of histone modifiers
IV. loss of proteostasis – accumulation of misfolded and damaged proteins
V. deregulated nutrient-sensing – dysregulation of metabolic pathways that are regulated by nutrient levels
VI. mitochondrial dysfunction – a decline in respiratory chain function and ATP production caused by many mechanisms such as the accumulation of mutations in mitochondrial DNA (mtDNA), oxidation of mitochondrial proteins, etc
VII. cellular senescence – increase in numbers of cells that are in cell cycle arrest and have acquired senescence-associated secretory phenotype (SASP) – a secretome that is particularly enriched with proinflammatory cytokines
VIII. stem cell exhaustion – the reduced activity of stem cells that results in decreased regenerative potential of tissues
IX. altered intercellular communication – changes/disturbances in signalling between cells that have harmful consequences to the organism

There are many defects in the immune system of elderly people that manifests in poor defense against infections [38]. In particular, a low-grade, chronic, and systemic inflammation, termed inflammaging stands out. In a sense, it falls under the hallmark 9, an altered intercellular communication, as this process makes intercellular communication more inflammatory. It seems to be a significant risk factor for morbidity and mortality in elderly people and is characterized by increased levels of inflammatory cytokines such as IL-6 [39]. This type of inflammation is very different from the normal acute immune response. Instead of being activated by pathogens, it is triggered by the organism's own DAMPs, although, some infections, such as cytomegalovirus (CMV), are also considered to be contributing to it. Of course, those hallmarks are not independent and in fact, there is evidence that all those nine hallmarks of aging are associated with chronic inflammation [40, 41, 42, 43, 44, 45, 46, 47, 48]. Also, the levels of immune cells and inflammation markers have several times been used to derive a score that indicates a person's health status and such scores are shown to correlate with morbidity and mortality [49, 50]. The age-related changes in T cell compartment as well as their associations with CMV are covered in depth in Paper I.

# 3. OVERVIEW OF THE DATA AND METHODS

In this chapter, I cover the main data generation mechanisms and analysis methods that played essential roles in the articles included in the thesis. In the first part, I mainly focus on the DNA methylation data due to its key role in two of the articles and because its preparation was quite substantial. Additionally, I briefly introduce proteomics data but with a lesser degree as it was mostly prepared by the proteomics core facility or by the platform's internal data preprocessing software. In the second part, I will give a short overview of statistical tests, correlation analysis, and unsupervised and supervised machine learning methods that were used throughout our studies.

## 3.1. Data types

In our studies, we mainly analysed DNA methylation and proteomics data. In both cases, we used two different approaches to obtain these data. In order to obtain DNA methylation levels, we measured it either by sequencing bisulfite-treated DNA amplicons (Paper I [51]) or by using DNA methylation arrays (Paper II [52]). Regarding the proteomics data, we mostly measured the levels of plasma proteins and for this purpose utilized the Olink panel based on Proximity Extension Assay (PEA) technology (Papers I [51], IV [53], and V [54]). However, in Paper III [55] we were more interested in the entire proteome and thus used data from nano-scale liquid chromatography coupled to tandem mass spectrometry (nano LC-MS/MS) experiment.

### 3.1.1. DNA methylation data

*Bisuflite sequencing for generating DNA methylation data.* In our work, we used paired-end amplicon sequencing where both ends of the DNA fragment of interest are sequenced. The DNA sequencing stands for a technique that is used to determine the nucleotide sequence of a DNA molecule. It starts with library preparation which means generating a collection of DNA fragments that are ready to be sequenced. Firstly, in the case of amplicon sequencing, preselected genomic regions are amplified using polymerase chain reaction (PCR). Then, short oligonucleotides are ligated to the 5' and 3' ends of DNA amplicons. Those short sequences called adapters play many important roles in sequencing since they form binding sites for primers, contain barcoding sequences, and also immobilize fragments to allow bridge amplification [56]. During paired-end sequencing (Fig 4), both ends of the DNA fragment of interest are sequenced. Those DNA fragments are called inserts because they are "inserted" between adapters. This process yields two sequences called "read 1" and "read 2" (R1 and R2) that are stored in FASTQ files. Besides nucleotide sequences, it also contains quality score for each base call [57]. Base calling denotes a process of inferring bases,

i.e., nucleotides from light intensity signals during sequencing.

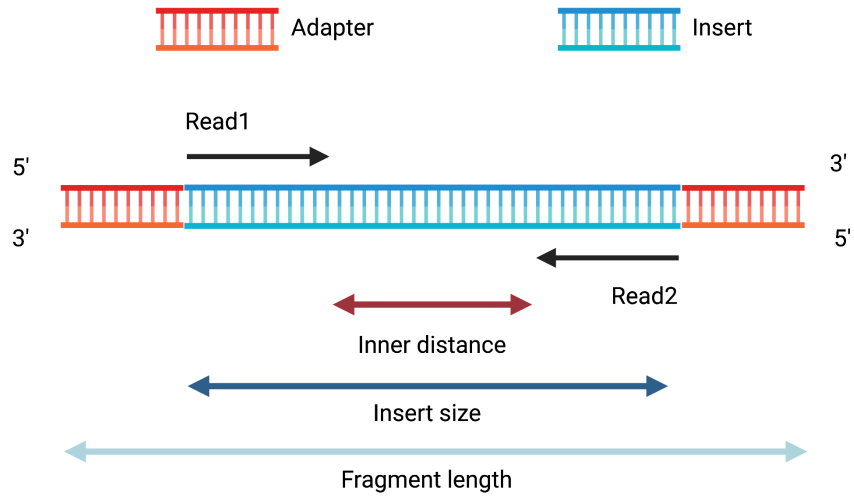


**Figure 4.** Overview of paired-end reads. The DNA fragment of interest denoted as "insert" is situated between short sequences termed adapters. The "fragment" stands for the entire sequence which contains insert and adapters. During sequencing a "read 1" (R1) is obtained in the 5' $\rightarrow$ 3' direction along the forward strand of DNA and "read 2" (R2) in the 5' $\rightarrow$ 3' direction along the reverse DNA strand. The distance between the reads is designated as "inner distance". This figure was made with Biorender.com

The DNA fragment of interest (insert) is often longer than the combined length of R1+R2, resulting in a gap referred to as the "inner distance" or commonly known as the "inner mate distance". As the orientation and inner distance are known, the read mapping tools can harness this property and use this information to improve the mapping quality. The benefits of this are especially conspicuous when difficult areas such as repetitive rich regions of the genome get sequenced. It is also possible that the read lengths are greater than the insert size and thus R1 and R2 are actually overlapping. During read mapping, it is also possible to exploit this property because it means that R1 and R2 can be merged and aligned as a single long read. However, when the read length is greater than the length of the entire insert, it results in an adaptor read-through. This produces reads encompassing adaptor sequences and thus can cause serious problems in downstream analysis. Therefore, those adapter sequences must be identified and trimmed. In addition, it is a common practice to remove low quality regions in reads, since in the case of Illumina, the sequencing quality does not remain constant and drops at the end of the sequence [58]. In the FASTQ file, read quality is encoded in a compact form needing only 1 byte per quality estimate for each base call. More precisely, the quality score (Q) is represented by a character corresponding to ASCII code equal to quality score +33 (ASCII code value = Q + 33). The quality score itself is in the form:

$$Q = -10\log_{10} P,$$

and thus the probability of an incorrect base call (denoted as P) can be calculated as:

$$P = 10^{\frac{-Q}{10}}$$

[57]

As the common trimming cutoffs based on the quality score are 30 and 35, the cutoffs in probability are 0.001 and ∼0.0003, respectively.

There are various tools developed to investigate the quality of FASTQ files and trim the reads. In our study, we used FastQC [59] for studying the quality of sequencing in each sample and aggregating those into a summary report using MultiQC [60]. For removing adapters and filtering low quality bases we used Trim Galore [61].

Since the nucleotide sequences stored in reads are only useful when their genomic location is known, the next step is read mapping. This process is carried out by read mapping tools that align reads on a reference genome. A reference genome is an accepted representation of the genome of a particular species that is assembled from DNA sequences originating from many individual donors. Regarding the read mapping tools, there are many different ones available with some functioning as general-purpose tools and others are specific to certain platform (e.g., Illumina) or source material (e.g., working only with DNA, RNA, miRNA or bisulfite converted DNA). The main challenge of mappers is to find the true location of those reads. However, as the high-throughput sequencing technologies produce large amounts of relatively short reads with a non-zero error rate, the reference genome itself is rather large and there exists considerable genetic variation between samples, then this is a non-trivial problem. To cope with that, the mapping has to be approximate thus allowing mismatches and gaps in alignment, but it also needs to take into account as much information as possible, such as the distance between paired-end reads, quality scores, and some mappers even include platform-specific information. In addition, those tools need to support various read lengths and since the number of reads is usually large, reaching even to hundreds of millions, hence most of those tools can be natively executed in parallel. The majority of those can run in parallel on shared-memory computers but some support parallel execution on distributed-memory computers. Given the high number of demands for this software, those tools usually come with a considerable number of parameters that enhance flexibility but at the same time make finding the best parameters rather difficult. Generally, the output of the read mapping tools is a sequence alignment and map (SAM) or its binary equivalent BAM file that stores alignment information, thus keeping interoperability high [62].

In Paper I, we had to map bisulfite converted DNA, and thus we used a read mapping tool tailored to this purpose called Bismark [63]. In bisulfite treatment,

the DNA is first treated with sodium bisulfite, a compound that specifically deaminates unmethylated cytosines and converts them into uracils, however, at the same time, it leaves methylated cytosines intact. Then, subsequent PCR amplification converts uracils into thymines and thus this process gives rise to four individual strands of DNA [64]. Therefore, as all originally unmethylated cytosines become thymines and methylated cytosine remain cytosines, the methylated cytosines can be distinguished by aligning reads to the original reference genome. As is evident above, the alignment process itself gets considerably more complicated. More specifically, the number of mismatches is significantly greater than before, four DNA strands need to be analyzed, and since many C-s in DNA get replaced by T-s, the read itself gets less informative.

The way Bismark approaches this task is depicted in Fig 5. Prior to the alignment, the reference genome is converted into two genomes resembling full bisulfite conversion – one where all Cs are converted into Ts, and the other where all G-s are converted into A-s. The reads are transformed similarly to the reference genome and are aligned using general-purpose aligner Bowtie 2 [65] running in four parallel instances (Fig 5A). The reads producing the best alignment are then compared to the unmodified reference genome and the methylation states of all cytosines are inferred (Fig 5B [63]). Since the sequencing source material comes from multiple cells, therefore an average methylation percentage in a given position can be calculated using counts of methylated (M) and unmethylated (U) cytosines per site thus giving a formula:

$$\frac{M}{(M+U)},$$

where $(M+U)$ corresponds to read depth indicating the confidence of a given value. It also serves as a good measure for filtering, for example, in paper I [51] we excluded all methylation values where read depth was $<$300.

*Methylation arrays for generating DNA methylation data.* The cost-effective alternative to sequencing for measuring methylation levels in certain DNA positions is a methylation array. To my knowledge, methylation arrays are currently produced only by Illumina with the latest MethylationEPIC array covering $>$850,000 positions in the genome. We also used this array in Paper II. In this platform, each cytosine position is targeted with certain probes that use fluorescence intensities of the methylated and unmethylated signals to estimate DNA methylation levels [66]. In fact, there are two types of assays that enable measuring methylation levels – Infinium I and II (Fig 6). Both of those assays utilize 50 bases long DNA sequences called probes that are attached to microbeads and form a place where bisulfite converted DNA strands from samples can hybridize. Both of those assays harness the C/T conversion of bisulfite treatment meaning methylated cytosine remain cytosines while unmethylated cytosines get converted into thymines [67].

The Infinium I assay design uses two beads to measure cytosine methylation in

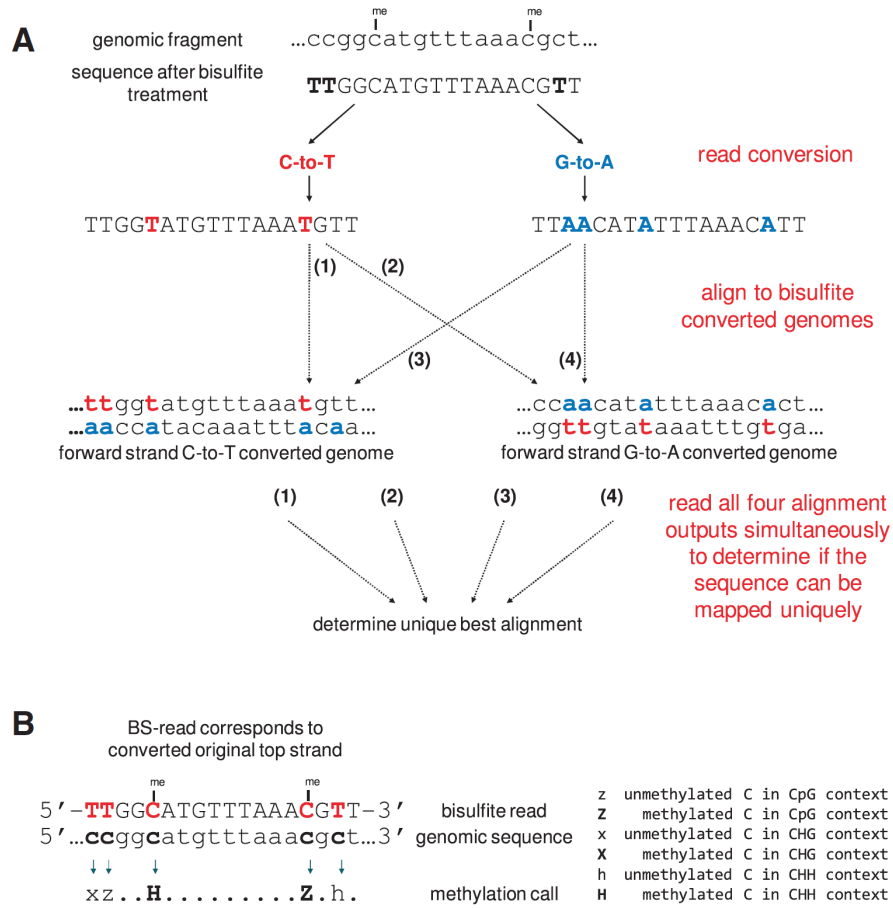


**Figure 5.** The working principle of the Bismark command-line tool. (A) Bisulfite sequencing experiment produces reads that are then subjected to full C-to-T and G-to-A conversions. Then those converted reads are aligned in four parallel processes to equivalently converted versions of the reference genome and the read producing the best alignment gets selected for the next step. (B) The methylation statuses of cytosines in that read are determined by comparing the original read to the unmodified genomic sequence. The CHH context denotes a situation where C is followed by any base except C (termed as H). This figure is reproduced from Krueger and Andrews, 2011 [63].

a given position (Fig 6A). One bead has an unconverted probe attached to it where only methylated DNA fragments can bind since bisulfite conversion keeps methylated cytosine intact. On the contrary, the other contains a converted probe and forms a binding place for unmethylated DNA fragments. After the DNA fragment is bound, it is extended by one fluorescently labelled nucleotide thus producing a fluorescence signal. Therefore, the methylation level, which is denoted with $\beta$ in microarray studies, can be calculated by using intensities from two different probes in the same colour [67]:

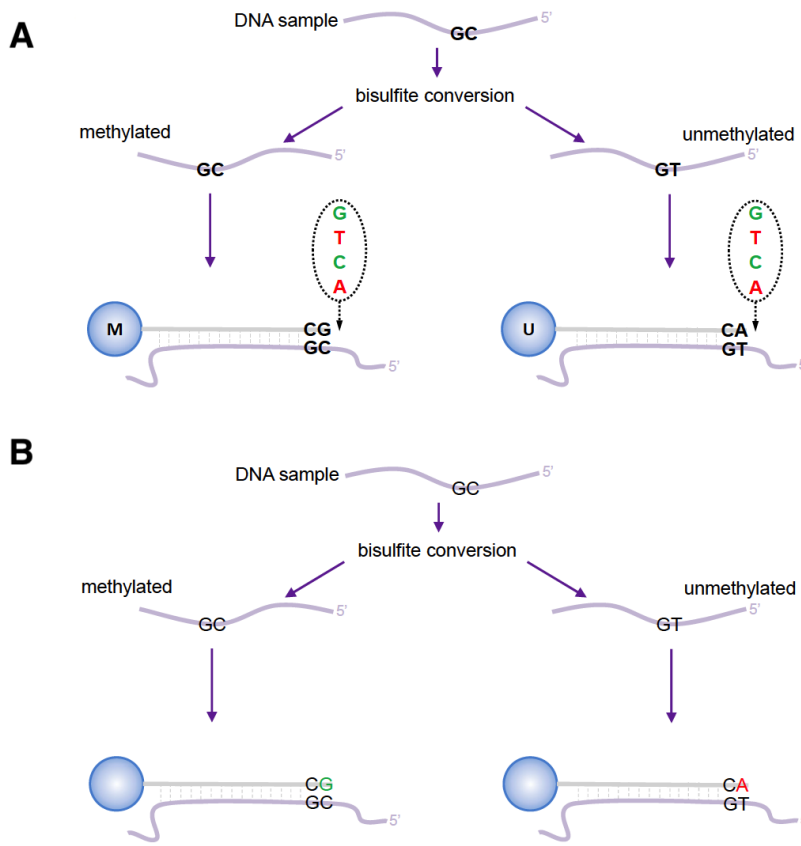


**Figure 6.** Working principle of Illumina Infinium assays. (A) The Infinium I assay measures each cytosine's methylation status using two microbeads. A bead M forms the binding place for methylated DNA fragments as it has unconverted probe while the opposite is true for bead U. After a DNA fragment is hybridized to a probe, a probe gets extended by a fluorescently labelled nucleotide that produces signal that is used in calculation of the methylation level. (B) The Infinium II assay measures each cytosine's methylation status using only one microbead. The probe attached to the bead ends exactly before the cytosine of interest and the addition of fluorescently labelled nucleotide is used in determining the methylation levels since different nucleotides are labelled with different colours. This figure is reproduced from Maksimovic et al. 2012 [67].

$$\beta = M - intensity/(M - intensity + U - intensity),$$

where M- and U-intensities corresponds to intensities from bead M and U, respectively.

In the Infinium II design (Fig 6B), the existence of methylation of each cytosine is determined using only one bead. The probe connected to the bead ends right before the cytosine of interest and the methylation level is determined by single base extension at this position. When the DNA fragment contains methylated

cytosine in this position, the probe sequence gets extended by "G" while in the opposite scenario "A" nucleotide is added. As the nucleotides used to extend the probe sequence have two different fluorescence colours the methylation level at a particular position can be calculated by comparing the intensities of those colours: [67]:

$$\beta = Green(M)/(Red(U) + Green(M))$$

Subsequently, the fluorescence intensities are measured with specific scanners that store the summary intensities for each probe in the Intensity Data (IDAT) files [68]. However, the downside of using fluorescence intensities is that it gives rise to various technical artifacts. For example, the technical variation can be caused by background fluorescence, sample position on a BeadChip as well as intensity differences between different fluorescence colours termed dye bias [69, 70, 71]. In addition, it has also been shown that Infinium I and II probes cause different technical noise as they produce different $\beta$-value distributions [72]. Therefore, there are many methods developed to tackle this problem and they are usually divided into two categories – within array and between array normalization [71]. The former comprises background correction, dye bias adjustment, and Infinium I/II-type bias correction while the latter corresponds to technical variation caused by external factors such as differences in quantities of biological material and aims to adjust measurements at a global level [71]. For example, common methods (but it is by no means an exhaustive list) to reduce technical biases are SWAN normalization [67], subset quantile normalization [73], stratified quantile normalization [74], single sample normal-exponential out-of-band normalization (ssNoob) [75], and functional normalization [76].

Data normalization is not the only preprocessing step in the ordinary data analysis pipeline. For instance, it is likely that some of those probes may fail, and there are methods aimed to calculate p-values that indicate the quality of the signal. Those p-values are calculated by comparing the total signal for each probe to the background signal estimated using negative control probes. Thus, those p-values provide a meaningful way for filtering out unreliable data before normalization and subsequent downstream analysis [74]. Furthermore, it is also common practice to remove cross-reactive probes and probes where single nucleotide polymorphism (SNP) lies in the cytosine of interest. The rationale for leaving out cross-reactive probes lies in the fact that bisulfite treatment results in the reduction of sequence complexity as most of the C-s get replaced with T-s. Therefore, it is likely that many of the 50 bases long probes will bind to DNA fragments from different regions as was initially planned. Therefore, those probes would produce methylation levels that are actually combinations of methylation levels in multiple regions and thus result in erroneous findings. Similarly, if a position of interest has an SNP, then there is a high chance that the obtained methylation level is false, and it is in fact just marking the genotype of a given person. For example, if a cytosine is normally fully methylated then the analysis of a person

who has both alternative alleles would show 0% of methylation in a particular site while measurements obtained from heterozygous individuals would indicate 50% of methylation [71]. Thus, without knowing the genotypes of the individuals included in the study, leaving out probes that are likely to contain SNPs improves the reliability of the analysis results.

Finally, the DNA methylation levels can be represented either by $\beta$-values that denote the proportion of DNA methylation in a particular position and range from $0 \ldots 1$ or by M-values calculated as $M - value = \log_2(M/U)$, where M is methylated intensity and U unmethylated intensity. According to this formula, when M-value equals 0 then the underlying cytosine is on average half-methylated, therefore, positive values show that more than 50% of Cs are methylated, and vice versa is true for negative M-values. The relationship between $\beta$ and M-values is logistic as is exemplified by following equations [77]:

$$\beta = \frac{2^{M-value}}{2^{M-value}+1};$$

$$M - value = \log_2(\frac{\beta}{1-\beta})$$

The reason why both are used is that $\beta$ values are easily interpretable while due to distributional properties M-values are more appropriate for statistical analysis. More precisely, it has been shown that $\beta$ values have severe heteroscedasticity (i.e., variable variance over different $\beta$ values) outside the middle methylation range and thus its usage violates the assumption of many statistical analysis methods that require data to be homoscedastic. Since M-values are considerably more homoscedastic, their usage is more statistically valid [77].

### 3.1.2. Proteomics data

In our studies, we used proteomics data produced by two different technologies. In Paper III [55], we used data from nano LC-MS/MS experiment and in Papers I [51], IV [53], and V [54], we used data produced by Olink assays. As nano-LC-MS/MS experiment was carried out in the proteomics core facility and we only received analysis ready data, and Olink assay data was preprocessed using its own software, I cover this topic very briefly.

The nano-LC-MS/MS is a method that combines liquid chromatography and mass spectrometry. Liquid chromatography is a laboratory technique used to separate components of a mixture. In this process, a mixture of interest is dissolved in a solvent, called the mobile phase and it is carried through a column containing the stationary phase. Since different components of a mixture have different affinities to the material constituting stationary phase, they take different times to pass through the column and thus become separated. Subsequently, those separated components are subjected to a mass-spectrometry analysis. In

mass-spectrometry, the atoms and molecules of those compounds are ionised, accelerated, deflected by an electromagnet, and finally detected. As the deflection depends on the mass/charge ratio, different atoms and molecules can be identified. The result of mass-spectrometry is an information about the relative abundances of particles with different mass/charge ratios, and this has to be processed further in downstream analysis. In our case, the data was processed with MaxQuant [78], a software that is one of the most frequently used methods in mass-spectrometry based proteomics data analysis. It contains various methods for peptide identification, protein inference, and quantification. The final output in our case was intensities calculated for each protein in normalized LFQ (label-free quantitation) units that can be used as proxies for absolute protein levels [79].

The Olink's PEA technology is based on a pair of antibodies that target each specific protein. Those antibodies are labelled with DNA oligonucleotides that have complementary regions to each other. Upon simultaneous binding to the protein of interest, the DNA oligonucleotides attached to the antibodies hybridize and get extended by a DNA polymerase. This results in the formation of double-stranded DNA "barcode" which is unique for a given protein and is proportional to the initial level of the protein of interest. This DNA "barcode" serves as a substrate for a subsequent quantitative PCR (qPCR) reaction that enables the quantification of DNA amplicons [80]. Then, the readouts from qPCR are converted into relative quantification units called Normalized Protein eXpression (NPX) that measure the level of protein expression in a $\log_2$ scale. As it is a relative measure, it means that proteins with the same NPX values may have different actual concentrations and thus, the NPX values can only be used to compare the same protein across the samples included in one experiment [81].

## 3.2. Statistical analysis methods

In this section, I describe statistical analysis methods that we frequently used in our articles. Hypothesis testing was very prevalent among all the articles and thus I begin with that. Then, I move into correlational analysis and describe the two most common correlation methods used in our studies. After that, I will briefly discuss some unsupervised learning methods that we routinely applied in exploratory analysis, and eventually discuss supervised learning methods that we used for modelling, feature selection, and data imputation.

### 3.2.1. Hypothesis testing

Statistical hypothesis tests provide a quantitative way to draw conclusions about the target population using a representative sample. Statistical tests can be used to determine whether there exists a difference between two or more groups but also to find out whether there is a relationship between the variables of interest. The starting point of hypothesis testing is the formation of null and alternative hypotheses. The null hypothesis, denoted as $H_0$, is a general statement that can be

accepted by default while the alternative hypothesis, named $H_1$, is the negation of the null hypothesis. After that, a suitable test statistic has to be found that would distinguish between the $H_0$ and the $H_1$. The value of the test statistic is calculated from the observed data and is used to evaluate the statistical significance of the result – the probability to observe a test statistic value as extreme or more extreme under the $H_0$. This probability is termed p-value and it is compared to a pre-set criteria for a decision, called significance level denoted as $\alpha$, that generally is set to 0.05. If $p - value \leq \alpha$ then the $H_0$ is rejected and hence the data supports alternative hypothesis while the opposite scenario means failure to reject the $H_0$.

There are numerous statistical tests available with each tailored for a specific situation characterized by different assumptions. Very generally, they are usually divided into two – parametric and non-parametric tests. Parametric tests assume that the data at hand follows a certain distribution such as normal distribution while non-parametric tests are distribution-free tests. There is a trade-off, namely when the data follows a given distribution then non-parametric tests are less powerful, i.e., produce more false negatives. On the other hand, when the distribution requirement is not met and the sample size is small, the non-parametric tests give more reliable results [82].

In our studies, we used many different statistical tests, but we used more often unpaired two-sample t-test and its non-parametric alternative unpaired two-sample Wilcoxon Rank Sum Test also known as Mann-Whitney U Test. Both of those are used for comparing two sets of independent samples. The two-sample t-test compares the means of those two groups and thus the $H_0$ states that the underlying population means of those two groups are the same ($H_0 : \mu_1 = \mu_2$) while the $H_1$ says that they are not equal ($H_1 : \mu_1 \neq \mu_2$). The t-test uses t-statistic as a test statistic and it can be calculated using the following equation:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where $\bar{x}_1$ and $\bar{x}_2$ correspond to the means of studied groups, $s_1^2$ and $s_2^2$, denote the variances of the groups, and $n_1$ and $n_2$ are the group sizes. The t-statistic follows a t-distribution. As there are many t-distributions, each defined by its degrees of freedom, subsequently the degrees of freedom is calculated. The degrees of freedom refers to the number of values that can vary in the calculation of the t-statistic. It is equal to the number of independent values that go into the calculation minus the number of parameters used as intermediate steps in the estimation of the t-statistic itself. For example, in the case of two sample t-test, it is calculated as $(n1 + n2) - 2$, where $n_1$ and $n_2$ are the sample sizes of the two groups being compared and 2 is subtracted because the sample means of two groups are used to calculate the t-statistic. After that, the obtained t-statistic is compared to the t-distribution to calculate the p-value and determine the statistical significance.

The Mann–Whitney U Test works on ranks and is used to test whether the

probability of an observation from population one ($X$) being greater than the observation from population two ($Y$) is different from the probability of an observation from population two being greater than the observation from population one. Therefore $H_0$ states those probabilities are equal: $P(X > Y) = P(X < Y) = 1/2$ and the $H_1$ negates it: $P(X > Y) \neq 1/2$. The first step is to pool the data in ascending order and assign ranks to the data. Then the test statistic U is calculated for groups 1 and 2:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R1;$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R2,$$

where $R1$ and $R2$ are the sum of ranks for groups 1 and 2, respectively, and $n_1$ and $n_2$ denote group sizes. The smallest of the two $(U_1, U_2)$ serves as a test statistic called U, and similarly to the t-test, its distribution under the $H_0$ is known and the p-value is calculated by comparing the U value to the null distribution.

Mistakenly rejecting $H_0$ is called a type I error and it leads to a false positive result while not rejecting $H_0$ when the $H_1$ is true is called a type II error which is known as a false negative result. Since the p-value is uniformly distributed when the $H_0$ is true the $\alpha$ shows the probability of making the type I error. When the number of tests increases, it also means that the amount of false positives increases. For example, when the $H_0$ is correct, then by doing 100 tests the expected number of false positives is 5 and the probability of observing at least one false positive is $1 - (1 - 0.05)^{100} \approx 0.994$. Motivated by this, there are methods called multiple testing correction methods that aim to solve this problem and limit the proportion of false positives among all positives to the pre-set criteria. Those methods work by either adjusting the significance level $\alpha$ or the p-value.

The most well-known method is the Bonferroni correction which ensures that after performing $m$ hypothesis tests, the probability of making type I error is not greater than $\alpha$. To control for this probability, the Bonferroni correction adjusts the significance level by dividing it by the number of tests being performed ($m$). This means that the correct significance level for each test becomes $\alpha/m$, and as this is equivalent to $p \times m \leq \alpha$, the Bonferroni corrected p-value is defined as $min(p \times m, 1)$. If the corrected p-value for a test is less than or equal to $\alpha$, then the null hypothesis can be rejected in favour of the alternative hypothesis.

One of the drawbacks of using the Bonferroni correction is that it can be overly conservative in certain situations, such as studies with a high number of tests and strong correlations among the tests, like in omics studies. In these cases, the Bonferroni correction may lead to a high rate of false negatives, meaning that the null hypothesis is not rejected when it should be. That being said, if the number of independent tests from the total number of tests can be correctly inferred, Bonferroni correction presents itself as a reasonable option. This estimated number is generally referred as the effective number of independent tests ($m_{eff}$) and thus

the Bonferroni corrected p-value becomes $min(p \times m_{eff}, 1)$. One way to estimate $m_{eff}$ is via principal component analysis (PCA), a dimensionality reduction method that is described later in this chapter, but the main idea is to find the number of independent components using PCA that explains most of the variance, e.g., 99% and use this number as $m_{eff}$, thus making Bonferroni correction less conservative [83].

However, there are many other multiple correction methods tailored for such situation. One of the most prevalent methods is the False Discovery Rate (FDR). It is a less conservative approach than Bonferroni and ensures that the proportion of false positives among all significant results does not exceed $\alpha$ but at the same time has shown to have a greater power [84]. The FDR correction according to the Benjamini-Hochberg procedure starts with sorting the p-values in ascending order: $p(1) \leq p(2) \leq \ldots \leq p(m)$, and is followed by finding the largest $k$ where $p(k) \leq \alpha \times \frac{k}{m}$ and reject all null hypothesis for $i = 1 \ldots k$. Alternatively, it is possible to use adjusted p-values where each p-value gets multiplied by the total number of tests $m$ divided by its rank $j$: $p(i) = \min\{\min_{j \geq i}\{\frac{m}{j} \times p(j)\}, 1\}$

In our studies, we utilized all three of the aforementioned methods and selected the appropriate approach based on the characteristics of the data, such as the correlation between measurements, as well as the number of tests being conducted.

### 3.2.2. Correlation analysis

In our studies, we often used correlation analysis. It helps us to investigate the relationship between two variables by estimating the strength as well as the direction of the association. To be more specific, the correlation coefficient characterizes monotonic relationship between 2 variables, and it ranges from $-1 \ldots 1$ with 0 denoting no association, -1 perfect negative association, and 1 perfect positive association. A monotonic relationship means that when the value of one variable increases, also the value of another variable increases, or, when one increases, the other decreases. In our studies, we used two types of correlation coefficients Pearson's $r$ and Spearman's $\rho$. Pearson's correlation coefficient measures the linear association between two variables. While it can be used without any assumptions, a proper inference requires that the data is representative and that both variables being analysed are normally distributed and that their combination follows a normal distribution as well (also known as a bivariate normal distribution) [85].

Pearson's correlation is defined as:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}},$$

where $x_i$ and $y_i$ correspond to values of the x-variable and y-variable, respectively, and $\bar{x}$ and $\bar{y}$ are the means of those variables.

If the relationship between the variables is non-linear (but follows a monotonic pattern) or if their distributions significantly depart from a normal distribution, then Spearman's $\rho$ can be used as a non-parametric alternative to analyse the

relationship. This method is essentially Pearson's $r$ calculated using ranks of the values rather than the actual values, making it more suitable for analysing data that is not normally distributed or has outliers. In addition, Spearman's $\rho$ can be applied to ordinal data and it is defined as:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)},$$

where $d_i$ is the difference between the two ranks of each observation and $n$ is the sample size [85].

Both Pearson's $r$ and Spearman's $\rho$ can be accompanied by t-tests to determine whether the observed correlation is statistically significant. In these tests, the null hypothesis ($H_0$) states that the correlation coefficient is zero, while the alternative hypothesis ($H_1$) claims that it is not. In addition, they are often written together with confidence intervals, which indicate the range where the true coefficient is likely to fall with a certain probability (usually 95%), and are particularly useful when the sample size is small [85].

Since calculating the correlation coefficient may give misleading results if there is another confounding variable, there are also methods to calculate partial and semi-partial correlations that measure the strength of a relationship between two variables while removing the effect of other variables. For example, in Paper I [51] we used partial correlation to control the influence of person's age in the correlational analysis. There is a slight difference between the two, partial correlation removes the effect of other variables from both variables of interest, while semi-partial correlation removes the effect of other variables from only one of the two variables being analysed.

It can be exemplified by a situation where the analysis is controlled for the third potentially confounding variable. If the third variable is independent of the two variables of interest, then both methods (as well as ordinary correlation) will produce the same correlation coefficient. If the confounding variable correlates with only one of the variables of interest, then the partial and semi-partial correlations will be the same and will be smaller than the simple correlation. However, when the confounding variable correlates with both variables of interest, then the semi-partial and partial correlations will be smaller than the ordinary correlation but will also differ slightly from each other [86].

### 3.2.3. Unsupervised machine learning models

Unsupervised machine learning models allow the discovery of patterns in the data. There are two main classes of those methods: clustering and dimensionality reduction methods. The aim of clustering is to partition the data into clusters based on some distance metric so that the cluster would contain more similar observations over their measurement profiles than those residing outside the cluster. In our studies, we relied on hierarchical agglomerative clustering that initially treats

each observation as a separate cluster and then calculates a distance between each of the clusters and merges the two most similar into one. This procedure gets repeated until there is only one cluster left containing the entire data set [87]. There are also a variety of distance metrics available, but we often relied on the most commonly used Euclidean distance defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2},$$

where $x$ and $y$ are observations/clusters and $x_i$ and $y_i$ represent measurement values of $p$ features.

In addition, there are many different methods to define the distance between two clusters. Examples of those include complete-linkage, single-linkage, average-linkage, and centroid clustering. In the case of complete-linkage, the distance is calculated between the furthest data points between the two clusters, which makes this method also quite sensitive to outliers. In single-linkage, the similarity is found between the closest points in each cluster. This approach has a drawback as it may result in joining clusters that are quite dissimilar overall. Alternatively, average-linkage is a method in which the average distance over all pairs of data points from each cluster is calculated. The centroid-linkage corresponds to a situation where distance is measured between the central points of each cluster [87]. The result of clustering is usually depicted as a heatmap, a data matrix where individual values are represented as colours. It is usually shown together with a dendrogram, a tree diagram, that illustrates how the merging was done in hierarchical clustering and thus indicates the relationships between observations. The choice of the distance metric and distance definition can have quite a significant impact on the final result, however, there is no best method for every situation, and hence the choice has to be made based on the nature of the underlying data.

The dimensionality reduction techniques aim to reduce the number of variables in the data. The name comes from the notion that the number of measurements, i.e., variables describing each observation is called the dimensionality of the data. One of the most popular methods for dimensionality reduction is a principal component analysis (PCA) that is often used for descriptive analysis and has proven useful in various situations and disciplines. We also used it throughout our studies to get a glimpse of the relationship between the samples but also for other purposes such as calculating the number of independent components for multiple testing correction. The main idea of this method is to reduce the dimensionality of a dataset while at the same time preserving as much variability as possible. This is achieved by finding new variables called principal components which are linear combinations of the original variables that maximize variance and are uncorrelated with one another. Underneath, PCA solves eigenvalue/eigenvector problem. The eigenvectors and their associated eigenvalues can be calculated from the covariance or correlation matrix of the data. The coefficients of the linear combinations are the elements of the eigenvectors that are called loadings. The eigenvalues

are used to calculate the percentage of variance explained by each component. By definition, the first principal component accounts for the most variance and each subsequent component accounts for the next greatest possible variance. Usually, for visualization purposes, only the first two or three components are used but that does not mean that the last components are completely meaningless. For example, they can be used to detect outliers. In addition, the loadings of the PCA can be used to find variables that contribute most to the separation seen on PCA plot [88]. However, as the principal components are often linear combinations of all variables and when there are lots of features, then the interpretation of the variables based on loadings is a rather difficult task. It should be noted that there are some adaptations of PCA developed such as sparse PCA, that aim to shrink many coefficients to exactly zero, thus enhancing interpretability [89]. Also, as PCA is quite sensitive to outliers, it has prompted the development of more robust variants of PCA [90]. Nevertheless, in our studies, we did not put an emphasis on analysis of PCA loadings and removed outliers beforehand, hence, we did not include other variants of PCA in our analyses.

### 3.2.4. Supervised machine learning models

In Paper I, we used several machine learning models to predict certain cellular proportions and throughout our studies, we used linear regression models to obtain statistically significant results when we had to control for other variables. The models that were used in those studies were linear regression, ridge regression, lasso regression, and random forest. I will briefly introduce all of them below.

Linear models in general are simple, interpretable, and stable methods since they have a straightforward design and relatively few parameters. One of the main benefits of linear models is that they are less likely to overfit. The term overfit marks a situation when a model becomes too complex and starts to pick up on random noise in the training data and does not perform well on unseen data. This is especially important when working with small or scarce datasets. On the other hand, linear models may struggle to accurately capture complex relationships between variables and can result in underfitting, which means the model is not able to accurately predict the output given new input data. Overall, given their simplicity and interpretability, linear models are a good starting point for many machine learning tasks.

Among linear models, linear regression is the most commonly used method. It is used for predicting the values of continuous variable called the dependent variable using one or more predictor variables called independent variables. In addition to the prediction, it can also be used to examine the effect of the independent variables on the dependent variable, such as determining whether their contribution to the model is statistically significant and how the dependent variable changes when the predictor variables are altered.

The multiple linear regression based on $p$ independent variables is in the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i,$$

where $y_i$ is the value of the $i$th observation's dependent variable ($i$ ranges from 1 to $n$ which is the sample size), $x_{1i} \ldots x_{pi}$ are the values of the $p$ independent variables, $\beta_1 \ldots \beta_p$ are coefficients of independent variables, $\beta_0$ is an intercept, and $\varepsilon_i$ is the error term for the $i$th observation. Also, as the regression coefficients are estimates of the true values and subject to sampling uncertainty, they are often accompanied by confidence intervals to indicate the range within which the true value of the coefficient is likely to fall with a certain probability (e.g., 95%).

Linear regression relies on several assumptions, including that there is a linear relationship between the dependent and independent variables, that the errors have conditionally a zero mean ($E(\varepsilon_i | x_{1i} \ldots x_{pi}) = 0$), that the errors have constant variance across the range of the dependent variable called homoscedasticity ($Var(\varepsilon_i x_{1i} \ldots x_{pi}) = \sigma^2$), that the errors are uncorrelated, that the predictors are additive in the case of multiple regression, and that the data itself is representative. Violations of those assumptions may lead to inaccurate conclusions about the significance, coefficient magnitudes, and confidence intervals of the independent variables as well as overall poor predictive performance. One of the most important ways to check whether the assumptions are met is by the examination of the residuals (distances between actual and predicted values) against predicted values. Namely, while many of the assumptions above apply to the error term ($\varepsilon$), residuals are used as a proxy for the error term. Ideally, the residuals are scattered randomly around the zero line with constant variance thus indicating that errors have zero conditional mean, constant variance, and are uncorrelated. In addition, if the residuals follow a normal distribution, it suggests that errors (conditional on x) are also from a normal distribution. This in turn indicates that the given model is suitable for hypothesis testing even if the sample size is small as in this case our coefficient estimates are guaranteed to have a normal distribution [91].

Estimating the values of the coefficients ($\beta_0 \ldots \beta_p$) for a regression model is usually done using ordinary least squares method that minimizes the sum of squared residuals [92]:

$$\hat{\beta} = \arg\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where $y$ represents observed and $\hat{y}$ predicted values of the outcome variable, and $\hat{\beta}$ denotes estimated coefficients of the model.

While linear regression is a simple model, it can still overfit when there are more predictor variables than the sample size can realistically support. One of the ways to handle this is regularization that applies additional constraint, i.e., regularization term to the residual sum of squares [92]:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i))^2 + \lambda R(\beta),$$

where $\lambda$ is a tuning parameter that controls the amount of regularization and $R$ is a regularization function. Regularization reduces model complexity by making coefficients on averages small in magnitude, a process denoted as shrinkage. The outcome is that the model's variance decreases and thus increases its ability to generalize on new data. There are two popular regularization approaches that we also harnessed in our analysis. Those two are lasso and ridge regression [92].

The regularization term in Lasso regression (also known as L1 norm) is defined as:

$$\sum_{i=1}^{p} |\beta_i|,$$

and in ridge regression (also known as L2 norm) is defined as:

$$\sum_{i=1}^{p} \beta_i^2$$

While both of those shrink the coefficients towards 0 there is a significant difference between the two. In ridge regression, coefficients approach towards 0 but will not set exactly to 0 while in lasso, given that the $\lambda$ is large enough, some of the independent variables are exactly set to 0 and thus lasso performs feature selection. It is known that lasso is more likely to remove features in sets of correlated features. Therefore, ridge regression still keeps all of the $p$ variables in the model while lasso favours sparse solutions and ultimately makes the model more interpretable [92].

In addition to the linear models, we also used random forest in our analysis. Random forest is a popular algorithm for both classification and regression problems. In short, it is an ensemble method that grows decision/regression trees on different subsets of training data and then takes a majority vote of their predictions for classification and average in the case of regression [93]. As in our analysis, we used the random forest for a regression task, I will briefly introduce regression trees.

Both the regression and decision trees are flowchart-like tree structures with a root node at the top that is split into internal nodes and finally into leaf nodes. Both the root and internal nodes are decision nodes that contain test condition while the leaf nodes contain labels. In the case of regression trees, the labels are average values of the target variable in a group of data points determined by those test conditions. The trees are learnt recursively by splitting existing nodes into new partitions so that the leaves are maximally pure meaning that they contain maximally similar labels from the training samples. It is important to note that it is achieved greedily by selecting the best split from a set of possible splits. The best split is defined by having the lowest impurity (*Imp*) which in the case of regression trees is often variance (albeit many other measures exist such as the mean squared error and the mean absolute error) [93]:

$$Imp(Y) = \frac{1}{|Y|} \sum_{y \in Y} (y - \bar{y})^2,$$

where $Y$ is the set of target values, and $\bar{y}$ represents the mean of $Y$

Hence, as the test condition splits the tree based on a threshold set on a particular predictor variable, both the predictor variable as well as the cutoff for the threshold are chosen by finding such variable and cutoff point that minimizes variance. Also, the variable that produces the smallest variance among all the predictor variables is included in the root node condition [93].

The first step of random forest is that it draws random samples with replacement from the original dataset that are the same size as the original dataset. This sampling technique is called bootstrapping, and the objective of this is to make the model less sensitive to the original data. Next, each bootstrap sample is used to grow a regression tree. However, unlike a regular regression tree where each node is split using the best split among all variables, in the random forest, the best split is determined based on a random sample of the initial predictor variables. This helps to reduce the correlation between the trees. Lastly, the model can be used to predict the new data by taking the average of the regression tree predictions. In addition, the accuracy of each tree is tested using the data not included in the bootstrap sample called "out-of-bag" data and those errors are then aggregated to give an estimate of the error rate of the entire model.

Random forest algorithm is very versatile, and it is therefore no surprise that it has been incorporated into many methods. In our case, random forest formed the basis of the methods that we used for imputing missing values in the DNA methylation dataset as well as for feature selection in order to find all CpG sites that were useful for modelling certain cellular proportions. The imputation method, named missForest [94], works by training a random forest model for each of the variables present in the dataset using the observed parts (without missing values) of the dataset for training. As it is based on random forest, it is a non-parametric method capable of handling mixed-type data, complex interactions, and non-linear relationships [94].

Regarding feature selection, we used a method called Boruta [95], that is a wrapper around the random forest algorithm that aims to find all relevant features for a particular machine learning task. Feature selection itself has many benefits, such as enabling the discovery of features that are relevant for understanding the mechanisms related to the subject of interest and in the case of modelling, makes models more interpretable, less likely to overfit, and also reduces training time. Generally, feature selection algorithms are divided into two main categories with first consisting of methods that aim to find a minimal set of features for optimal prediction and the second comprising methods that aim to identify all relevant features regardless of whether they are redundant or not. Boruta works by harnessing random forest algorithm property to produce feature importances. Namely, the random forest estimates the importance of a feature by calculating a change in

prediction error on "out-of-bag" data when a given feature is permuted while all others are left unchanged. It is calculated separately for all trees that use given feature for prediction [96]. Boruta calculates Z-scores of those importance values by taking the mean of prediction error differences for each feature and dividing it by its cognate standard deviation. In order to determine whether the variable importance is actually significant, it also introduces "shadow" variables that are obtained by making a copy of each feature and shuffling its values around. Then the Z-scores of best performing "shadow" variables are compared to Z-scores of real features to determine which features are significantly better and are thus deemed important. Also, this procedure is repeated multiple times to make it insensitive to the particular realizations of shadow variables and the random forest model itself [95].

An important aspect of machine learning is model evaluation. In order to get a realistic estimate of the model's performance, the model is usually evaluated using an independent dataset often called validation or test set. However, putting aside substantial part of the data, usually nearly a quarter of it, may not always be a good idea. For example, if the data is scarce, then, on the one hand, the evaluation metrics are affected heavily by the initial partition of the data into test and train set and thus the obtained estimate may be biased due to nonrepresentative test set that just happened due to chance. On the other hand, the model could perform much better when it was trained on a slightly larger sample size. A common way to overcome this is to use k-fold cross-validation that randomly splits data into $k$ equal size parts and uses one fold for testing and the rest for training. As it iterates $k$ times, it ensures that each observation gets to be in the test set exactly once. After this procedure, the evaluation metrics are aggregated and if the results were satisfactory, then the model can be trained on the entire dataset for further use. In addition to the ordinary cross-validation, there is a nested cross-validation where each training set becomes a subject of cross-validation. For example, It is suitable for hyperparameter tuning and is very popular since many machine learning models come with hyperparameters such as $\lambda$ in the case of ridge and lasso regression or the number of trees and predictor variables in the case of random forest.

Regarding the evaluation metrics, in our studies, we focused on regression problems and in this case the performance of the model can be estimated using errors of the predictions on the test set. There are many metrics available, herein I mention the three main ones: mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE). They are defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where $n$ is the number of observations, $y_i$ is the actual value and $\hat{y}_i$ corresponds to the predicted value.

MSE is calculated by taking the mean of the squared differences between the predicted and observed values. As MSE involves squaring the errors, therefore when MSE is used in the loss function, i.e., the function that is minimized during model training, it penalizes large errors more heavily than smaller ones, provided that large errors are $> 1$. However, it also means that the units of MSE are squared, which can make it difficult to interpret. To address this issue, the square root of MSE is often taken, resulting in RMSE, that is on the same scale as the original target variable. Alternatively to MSE and RMSE, MAE is based on taking the absolute value of the difference between predicted and observed values. MAE is often preferred over MSE because it does not penalize large errors as heavily, resulting in a more linear relationship between the error and the metric score. However, MAE is harder to work with mathematically than MSE or RMSE, as absolute values can be more difficult to handle.

# 4. ANALYSIS OF TEMRA CELLS IN HUMAN BLOOD (PUBLICATION I)

This chapter is dedicated to Paper I [51] titled "Epigenetic quantification of immunosenescent $CD8^+$ TEMRA cells in human blood". In this study, we took a closer look at different $CD4^+$ and $CD8^+$ T cell subpopulations with a special interest in TEMRA cells. As was already mentioned in Chapter II, aging is a complex process that affects the immune system and comes with an increased risk of infections and chronic diseases, poor vaccine efficacy, and higher incidence of cancer and autoimmunity [38]. Here we studied the levels of those immune cells and their associations with other immunological measurements as well as aging in general. Finally, we built a predictive model for estimating the proportion of $CD8^+$ TEMRA cells in human blood using only a few CpG sites. This model will serve as a basis for future studies of our research group to find out whether its predictions could be used as a metric for monitoring an individual's immune health status.

## 4.1. Age-related changes in the T cell compartment

Immune system and age-related changes are more thoroughly discussed in Chapter II, however, here, to provide context for this study, I will delve deeper into the changes in T cells during aging.

It has been known for some time that the most crucial changes happening during aging take place in the adaptive immune system and in particular T cell compartment [97]. More precisely, it is evident that homeostatic proliferation is not efficient in maintaining the naïve $CD8^+$ T cell population since aging is often characterized by a reduction in the naïve T cell pool and increased levels of terminally differentiated, exhausted, and senescent $CD8^+$ T cells [97, 98]. Among those, a group known as terminal effector memory $CCR7^-$ $CD45RA^+$ (TEMRA) T cells, which are strongly associated with age, have been the focus of extensive research. They tend to be more sensitive to innate signals, show reduced capacity to T-cell receptor (TCR) dependent activation and are known to have lower TCR clonal diversity [99, 100, 101]). There is an ample amount of articles implicating $CD8^+$ TEMRA cells in various adverse health outcomes such as psoriatic arthritis, age-related chronic inflammation, several comorbidities, and overall reduced immune competence [102, 103, 104, 105, 106, 107, 108]. Albeit it should be noted that some studies have implicated their abundance with positive outcomes, for example, study focusing on octogenarians found that high levels of $CD27^-CD28^+$ $CD8^+$ TEMRA cells were associated with longer life expectancy after adjusting for known risk factors, such as heart failure, frailty, or cancer [109].

## 4.2. Motivation behind the analysis

Motivated by associations of $CD8^+$ TEMRA cells with human health, we first wanted to take a broader scope and analyse its subsets and their relations to other T cells. We also aimed to study less explored $CD4^+$ TEMRAs. In addition to examining changes related to aging and comorbidities, we included information on CMV infection since a cumulative antigenic load caused by viral infections, particularly CMV infection, is believed to contribute to the accumulation of TEMRA cells [110]. More precisely, CMV is a very common, usually asymptomatic virus, infecting around 60% of the global population [111]. It causes a persistent latent infection with episodes of reactivation. Notably, CMV infection stands out due to its ability to trigger T cell responses that do not follow the typical contraction pattern seen after usual infections. This leads to a phenomenon known as "memory T cell inflation," characterized by high levels of CMV-specific T cells, with a significant proportion exhibiting a TEMRA phenotype [112]. Instead of using CMV as a binary indicator like other studies, we utilized an in-house luciferase signal-based immunoprecipitation system (LIPS) assay to quantify CMV-specific antibodies, thus providing a quantitative measure of CMV infection status. The rationale of this is that when CMV infection is in a more active state there will be more CMV-specific antibodies which results in higher LIPS reading. This enabled us to explore for example, whether an increase in CMV-specific antibodies is accompanied by an increase in TEMRA cells. Also, given that many of the cell types included in the analysis may contribute to the inflammaging, a chronic low-grade inflammation that develops in high age, we obtained measurements of immune response linked proteins via using the OLINK inflammation panel (described in Chapter III). Finally, we thought that given TEMRA cells potential effect on an individual's health it would be of great interest to find a way how to measure it cost-efficiently. Namely, flow cytometry is not suitable for large-scale use due to its high cost and thus DNA methylation that relies only on a few CpG sites could serve as an efficient alternative. Prior to this study, a literature review was completed by other lab members to obtain a set of candidate CpGs that would enable us to explore the feasibility of using DNA methylation for predicting $CD8^+$ TEMRA levels.

To meet the goals set in the previous paragraph, we profiled T cells in the elderly cohort of 140 individuals with an age range from 65 to 96 years. It should be noted that we included 25 younger people for comparison, thus extending the age range from 4 to 96 and the number of individuals to 165. Altogether, we studied 26 $CD4^+$ and $CD8^+$ T cell subpopulations, serotyped and measured individuals' CMV levels, obtained the levels of 92 inflammatory markers, and measured the methylation levels of 191 CpGs via deep-amplicon bisulfite sequencing (described in Chapter III).

## 4.3. T cell differentiation markers and stages relevant to this study

Firstly, to put things in context, usually, $CD8^+$ TEMRA cells have been studied based on markers CD45RA and CCR7, but our study covered a much broader array of markers. More precisely, the $CD3^+$ T cells were divided into two – $CD4^+$ T helper and $CD8^+$ cytotoxic T cells. Those were in turn divided based on their expression of CD45RA and CCR7 into naïve ($CD45RA^+$ $CCR7^+$), central memory (CM; $CD45RA^-$ $CCR7^+$), effector memory (EM; $CD45RA^-$ $CCR7^-$ ) and terminally differentiated effector memory (TEMRA; $CD45RA^+$ $CCR7^-$) cells. The EM and TEMRA populations were divided further into 4 $CD4^+$ and 5 $CD8^+$ subsets. It should be noted that the $CD4^+$ T cells downregulate first CD27 and later CD28 marker while vice versa is true for $CD8^+$ T cells. Also, downregulation of CD127 and upregulation of PD1 are independent events of CD27 and CD28 downregulation. Thus, the studied $CD4^+$ T cells were $CD27^-$, $CD27^-$ $CD28^-$, $CD27^-$ $CD28^-$ $CD57^+$ and $PD1^+$ while $CD8^+$ EM/TEMRA compartment contained $CD28^-$, $CD28^-$ $CD27^-$, $CD28^-$ $CD27^-$ $CD57^+$, $CD127^-$ and $PD1^+$ subsets. The relationship of those cell populations is illustrated in Fig 7.

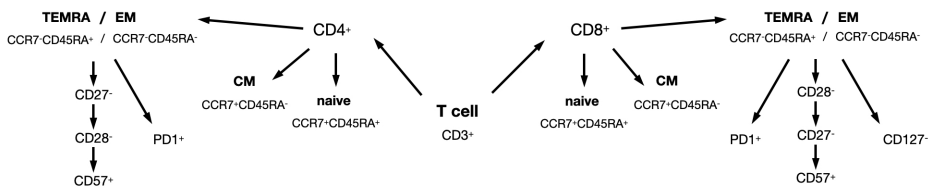


**Figure 7.** Schematic representation of studied $CD4^+$ and $CD8^+$ T cell populations in various differentiation stages. $CD3^+$ marker identifies all T cells and from there each subset is defined by the presence of certain surface marker(s).

However, it should be emphasized that there are some debates about how different T cell subsets should be defined and whether or not only surface markers are sufficient to define a certain cellular population [113]. Despite disagreements, some general characteristics have been established [114]. For example, TEMRA cells are characterized by the expression of CD45RA and the existence of certain senescence characteristics such as short telomeres, cell cycle arrest, and senescence-like secretome. Although, unlike senescent cells (in the T cell case $CD57^+$ cells), they have shown strong effector capabilities and their cell cycle arrest is reversible, e.g., it has been shown that many chronic viral infections such as CMV and Epstein-Barr virus are responsible for the proliferation of TEMRA cells. Exhausted cells, usually denoted as $PD1^+$ cells, have been described as effector T cells with reduced functionality. Senescent cells comprise a small cell population that has irreversible cell cycle arrest and possesses senescence-associated secretome that is responsible for low-grade inflammation that in turn harms surrounding tissues [114].

## 4.4. Main findings

During this study, we first confirmed some previously known trends such as the reduction in $CD4^+$ T cells and the increase in $CD8^+$ TEMRA cells with age. Using a moving average, we found a steady increase of $CD8^+$ TEMRA cells over the years with gradient rise after 50 years of age and emergence of $CD4^+$ TEMRA cells in the same period. Such stark changes are very intriguing and need further studies on larger samples to validate this result, especially given that our <65-year-olds group was very small.

Among elderly individuals' $CD8^+$ T cell compartment, TEMRA cells were the most prevalent, reaching on average nearly 60%. In particular, we saw that majority of $CD8^+$ TEMRA cells were age-associated $CD28^-$ or $CD27^-$ and $CD28^-$ subpopulations, of which many were positive for CD57, a senescence marker for T cells. In order to spot the most variable populations we used a signal-to-noise ratio (SNR) by dividing each cell population mean with its corresponding standard deviation. We used that because variation in the cellular compartments was in correlation with their overall proportion. Therefore, in addition to their high average levels, the $CD8^+$ TEMRA populations had the highest interindividual variability of all T cells in older individuals. Those findings are illustrated in Fig 8.

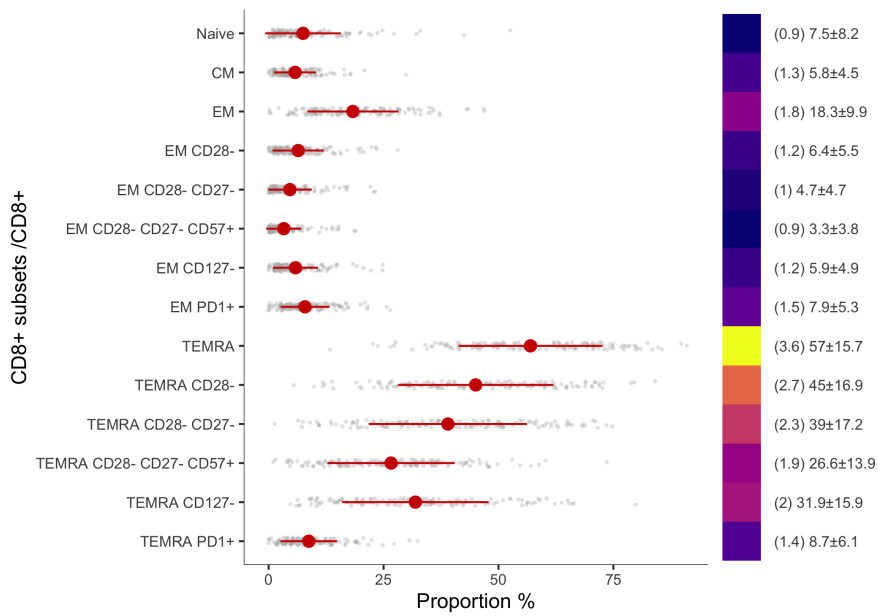


**Figure 8.** Proportions of $CD8^+$ T-cell subsets among $CD8^+$ compartment. Mean is shown by a red dot and the adjacent line denotes standard deviation, given information is also written next to each T cell subset. The value in brackets and colour bar show the level of signal-to-noise ratio (SNR). Brighter colour corresponding to higher value.

Subsequently, we calculated pairwise Pearson's correlations between cellular

population levels in the $CD8^+$ compartment. After clustering the obtained correlation matrix (Fig 9), we noticed that EM, CM, and TEMRA cells correlated well within their families, except $PD1^+$ $CD8^+$ TEMRA which did not correlate with other TEMRAs. It is also somewhat expected as the acquisition of the PD1 marker is not related to the loss of $CD27^-$ and $CD28^-$ as shown in Fig 7 and CMV-specific T cells are less likely to express PD1 [115].

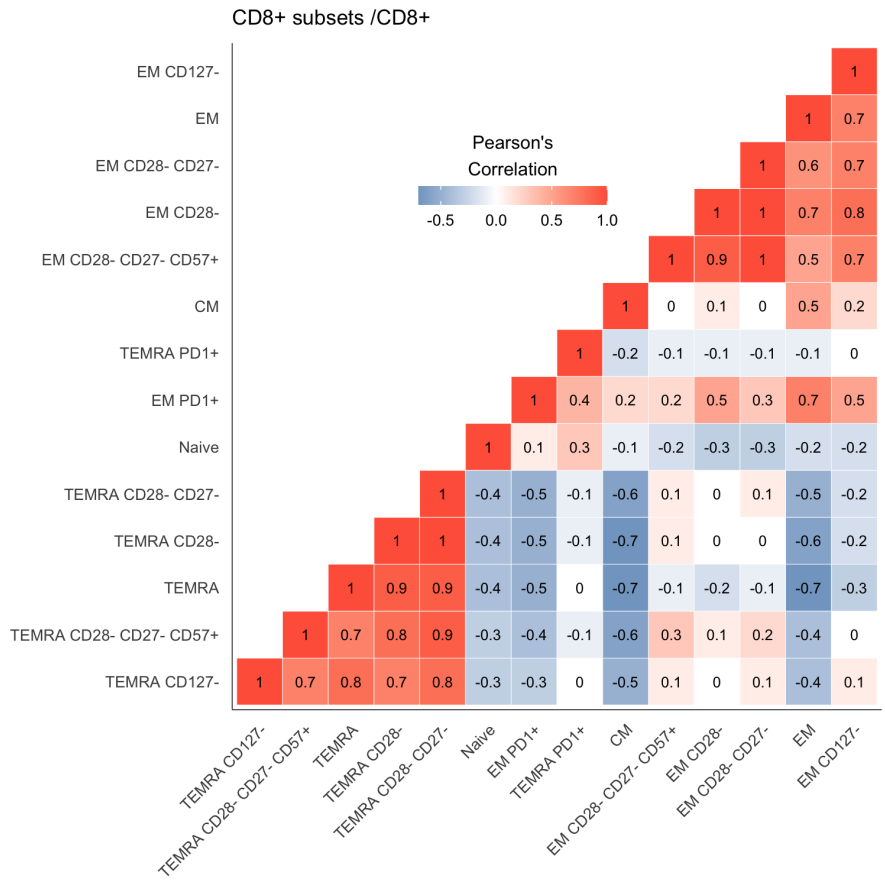


**Figure 9.** Clustered correlation matrix of pairwise Pearson's correlation coefficients calculated using the levels of $CD8^+$ T cell subsets.

Interestingly, as both $CD4^+$ and $CD8^+$ TEMRA cells are considered as age-related cell types, their levels did not correlate with age in people older than 65 years, therefore, implying that other factors are at play. Firstly, we studied whether this could be explained by the disease status of older individuals. As many age-related chronic diseases are inflammatory and elderly individuals are often diagnosed with several concurrent comorbidities (e.g., hypertension, cardiovascular disease, type 2 diabetes (T2D), autoimmunity, and chronic kidney disease (CKD)) [116], our study cohort exhibited a very diverse set of diseases. In order to

find out whether a particular disease status could explain the observed variability in $CD4^+$ and $CD8^+$ T cells, we divided our individuals into four groups. Firstly, we formed a control group that only contained individuals who either were healthy or had hypertension but did not manifest any other age-related disease. Altogether, this control group contained 33 individuals. Next, we formed the following groups – individuals who had diabetes (35), kidney disease (20), and autoimmune disease (28). As the three aforementioned groups were overlapping to some extent, we did not focus on finding the differences between diseases but instead only focused on comparisons with controls using linear models for determining statistical significance and also included sex and age as covariates. However, we did not find any significant association between cellular levels and those groups most likely due to the lack of better-suited control group.

CMV infection is very common and the majority (91.2%) of our elderly cohort was positive for CMV antibodies. Thus, we developed a LIPS method specific to two CMV tegument pp150 protein [117] fragments: p150d1 and p150d2, that enabled us to measure the antibody levels specific to CMV. Our measurements distinguish CMV positive and negative individuals very well as can be seen from boxplots and ROC analysis in Fig 10A-B.

Regarding cellular levels, we found p150d1 fragment specific antibodies to be positively correlated with $CD8^+$ TEMRA cells (Pearson's r = 0.45), and its subpopulations ($CD28^-$, $CD27^-$ $CD28^-$, $CD57^+$, $CD127^-$), thus giving additional support to the notion that CMV infection increases the number of TEMRA cells as seen in Fig 10C. However, it is not certain whether increased numbers of TEMRA cells actually restrict the spread of the infection or are they spreading to a such extent because they are inefficient to do so. A similar trend was also evident with p150d2 fragment specific antibodies, but as its range was higher, lower correlation could be explained by noisier reading due to light pollution from adjacent wells.

Regarding the inflammatory plasma proteins, the most notable association came with TRANCE (also known as RANKL), a TNF family cytokine that is well known for its functions in osteoclast differentiation but which also plays a role in the immune system [118]. We found that it was negatively correlating with $CD8^+$ TEMRA and its subpopulations. This prompts several new hypotheses, for example, TRANCE/RANKL has been shown to suppress proinflammatory cytokines in mice [119] and thus may be avoiding the generation of highly differentiated/senescent cells in the first place. However, in humans, such an effect has not been reported yet. It is also possible that TRANCE/RANKL affects the production of naïve T cells due to involvement in the differentiation of medullary thymic epithelial cells [120].

## 4.5. Modelling of $CD8^+$ TEMRA levels using epigenetics data

In the second part of our analysis we aimed to explore the feasibility of using DNA methylation to predict $CD8^+$ TEMRA levels. We chose to obtain epigenetic

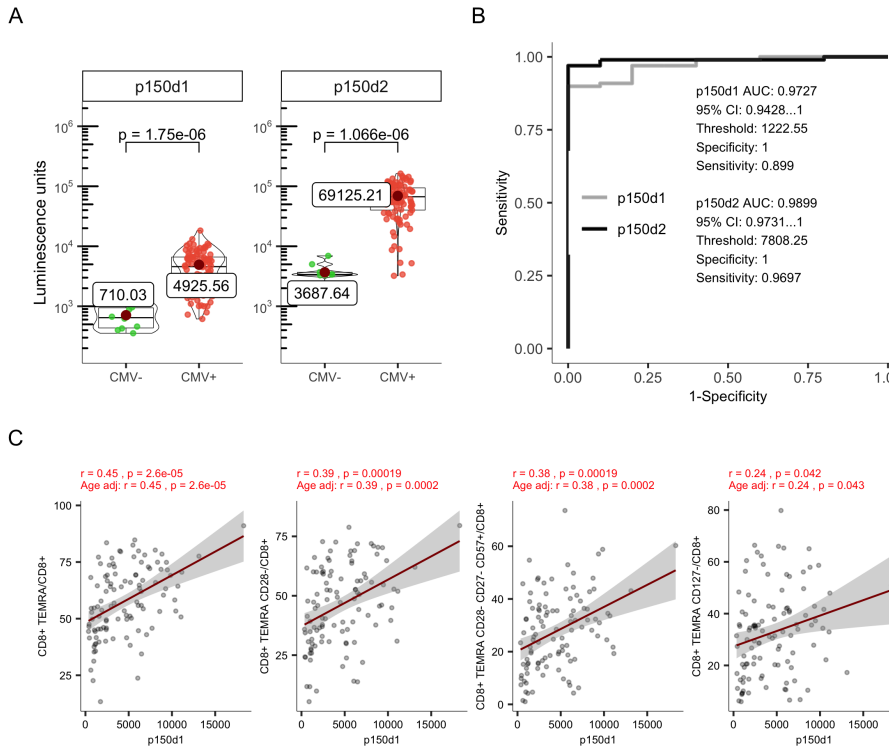


**Figure 10.** CMV-specific antibody measurements and their associations with CD8$^+$ TEMRA cells. (A) Boxplots showing the levels of anti-p150d1 and p50d2 antibodies in luminescence units in CMV positive and negative individuals. (B) ROC curve indicating classification performance of p150d1 and p150d2 LIPS measurements. (C) Correlations and age adjusted partial correlations between p150d1 antibody measurement and TEMRA subsets levels.

data on the whole blood cells (WBC) level since the peripheral blood is a common source of human genomic DNA and thus the results would be more relevant to the wider audience. As mentioned above, the CpG sites were first selected based on the extensive literature review with the most important source being an article by Tserel et al. that focused on age-related changes in DNA methylation and gene expression in CD4$^+$ and CD8$^+$ T cell populations [121]. All in all, we selected 191 CpG sites for the analysis of which many were close to the genes expressed in differentiated T cells. The steps needed to obtain the methylation data as well as the preprocessing of it are further described in Chapter III. To add more confidence to the results, after alignment we filtered out methylation values that were calculated using less than 300 reads. In addition, the data was then reduced (rows and columns removed) in a such way that overall it contained less than 5% of missing values and thus would be suitable for imputation. We chose to impute missing values using a random forest based imputation method called missForest [122] that is also discussed in Chapter III.

In order to model dependent variables ("CD8$^+$ TEMRA/WBC", "CD8$^+$/WBC" and "CD3$^+$/WBC"), we mostly relied on linear models (and analysed their Pearson's correlations) as a linear relationship between DNA methylation and cellular levels could be assumed. More precisely, we selected the CpGs in such a way that they would be specific to a certain population, i.e., unmethylated in the cell type of interest and methylated in the other white blood cells or vice versa. Thus, ideally, studied sites' methylation levels would be proportional to the proportion of the cell type of interest. As we were relying mostly on linear models, we investigated the distributions of dependent variables. We saw that distributions of "CD8$^+$ T cells/WBC" and "CD8$^+$ TEMRA cells/WBC" were following a skewed shape and thus were deviating from a preferrable normal distribution (based on visual inspection on quantile-quantile plots and Shapiro-Wilk test p-value $< 0.05$). Therefore, we applied transformation functions on those measurements – square root for "CD8$^+$ T cells/WBC" and cube root for "CD8$^+$ TEMRA cells/WBC".

Since the number of CpG sites (features) exceeded considerably the common statistics criterion of sample size / 10, a feature selection method had to be included. But before that, we split the data into training and test set (75% / 25%). Then, on training data and using transformed target variables, we first used a all-relevant feature selection method called Boruta [123] (described in the Chapter III) and on top of it we used lasso regression (also explained in Chapter III) to leave only one feature from each set of strongly correlated features. To train the models we used a nested 5-fold cross-validation scheme. We used outer folds to test different modelling algorithms (ridge regression and random forest) and inner folds for hyperparameter tuning (in the case of the random forest the hyperparameter was the number of trees and in the case of ridge regression it was $\lambda$). We used RMSE to evaluate the models' performance and to select final models. Finally, we tested all models on the test dataset and visualized their predictions in comparison with their actual values. It should be noted, that for testing the model's performance on test data we inverse-transformed the predictions of those models. The final CD8$^+$ TEMRA/WBC model was trained using ridge regression and contained 7 CpGs. Many of the 7 CpGs used to predicting CD8$^+$ TEMRA cells were located close to CD8A and GALNT8/KCNA6 that are expressed in T cells. Information about CD8$^+$ TEMRA levels captured by those 7 CpGs is illustrated by PCA plot (Fig 11A). Models' performance is shown on a test set containing 28 samples together with correlation and RMSE between actual and predicted values (r = 0.887, RMSE = 2.2) (Fig 11B). Due to the limited sample size and high variability of target measurements, we confirmed the usefulness of our selected features by additionally sampling a new training and test set from the entire dataset multiple times and automatically built linear regression models using the same features. This further supported the suitability of the selected CpG sites in the prediction of the cellular compartments since the observed correlation between predicted and actual values remained strong ($r > 0.6$).

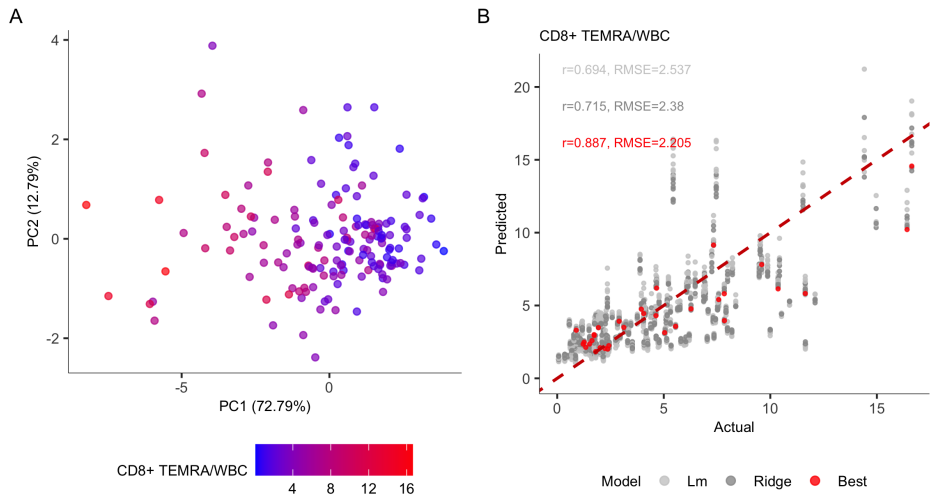


**Figure 11.** CD8$^+$ TEMRA associations with CpG sites included in the model. (A) PCA based on the methylation levels of 7 CpG sites included in the model and coloured according to the level of CD8$^+$ TEMRA/WBC. (B) The accuracy of the final model on training set shown in red together with predictions of models trained on resampled training dataset using linear (light gray) and ridge (dark gray) regression.

## 4.6. Summary and impact

To our knowledge, at the time of publishing this article, this was the most comprehensive analysis of effector memory and TEMRA subpopulations among elderly individuals. While many studies have addressed the differences between young and old individuals, the changes that occur after the age of 65 years have been less studied. We also found that the levels of CD8$^+$ TEMRA cells exhibit high interindividual variation, and among elderly people, their levels are not determined by age. In addition, the results of this analysis further validate many findings relating to CD8$^+$ TEMRA cells such as association with CMV infection, and also demonstrate the feasibility of using DNA methylation to predict CD8$^+$ TEMRA proportions from white blood cells. This provides a groundwork for later studies which will show whether epigenetics-based CD8$^+$ TEMRA model could serve as a tool to track immunosenescence.

## 4.7. Contribution

My contribution to this publication is the following, I performed the preprocessing and analysis of the data, interpreted the results, made figures, and co-wrote the manuscript. Also, I participated in the critical review of the paper.

# 5. ANALYSIS OF REGULATORY AND CONVENTIONAL T CELLS (PUBLICATION II)

This chapter is based on Paper II [52] titled "Graves' disease-associated TSHR gene is demethylated and expressed in human regulatory T cells". While the previous chapter focused on how age affects the immune system, herein we investigated immunological self-tolerance. More specifically, we aimed to investigate DNA methylation – one of the most studied epigenetic modifications, in the context of regulatory (Tregs) and conventional T cells (Tconvs). While Tconvs are crucial for shaping the immune response to clear pathogens, Tregs have an opposite role to play – they suppress unwanted immune reactions. We started with DNA methylation profiling of those cell types to see which genes are differentially methylated and therefore potentially involved in immune suppression. During this study the thyroid-stimulating hormone receptor (TSHR) gene, a known risk locus for Graves' disease (GD) [124], piqued our interest due to its differential methylation and expression in Tregs and enabled us to form a new hypothesis regarding its involvement in GD. Subsequently, it incited us to explore methylation differences in Tregs between GD patients and healthy individuals.

## 5.1. Overview of Treg function

Our immune system is a very complex system consisting of a vast array of different cell types and signaling pathways. On the one hand, it needs to protect us from a broad range of pathogenic microorganisms and agents while at the same time it needs to avoid launching misguided reactions against the organism's own healthy cells and tissues. $CD4^+$ T cells play a significant role in both of those tasks – Tconvs ($CD4^+$ $CD25^-$) are involved in immune system activation while $CD4^+$ $CD25^+$ T cells called Tregs confer immunological self-tolerance via their immunosuppressive functions. Tregs form a minority among $CD4^+$ T cells, comprising around 2-5% of the total population in peripheral blood [125]. They can be divided into two sub-populations based on whether they originate from the embryonic/neonatal thymus or are induced in the periphery [126]. They can also be divided in terms of their differentiation status – naive cells (nTregs), central memory cells (cmTregs), effector memory cells (emTregs), and effector Tregs (eTreg) [127].

Tregs have remained enigmatic regarding the mechanisms by which they suppress unwanted immune reactions. In fact, while many different mechanisms have been described, there is no universal consensus on this matter. This controversy could at least partly be attributed to the reason that different mechanisms may depend on the type and state of the target cell as well as the location of the immune reaction. For example, studies have shown that Tregs exert their immunosuppressive mechanisms on different Tconvs – T helper (Th) cells type 1 (Th1) and Type

2 (Th2) [128], but also have an effect on dendritic cells (DCs) [129] and $CD8^+$ T cells [130].

One of the proposed mechanisms of Treg action includes out-competition of Th cells via aggregating around DCs and physically hindering Th cell access to DCs [131]. In this process, Tregs downregulate CD80 and CD86 on DCs – receptors that normally bind to CD28 on the Th cells and prime their activation [129]. This downregulation is achieved through CTLA-4 receptors on the surface of Tregs that have a greater affinity to CD80 and CD86 than CD28 receptor on Th cells [129]. In addition, it has been shown that Tregs stimulate the expression of enzyme indoleamine 2,3-dioxygenase that produces kynurenine, a compound that is toxic to the adjacent Th cells [132]. Also, activated Tregs highly express the high affinity IL-2 receptor (IL-2R) that binds to the IL-2, a growth factor of T cells, that leads to lower levels of free IL-2 and consequently inhibits the activation of Th cells [133]. This also has a bystander effect on $CD8^+$ T cells that also need IL-2 for proliferation, but are not producing it in needed quantities on their own and need to rely mostly on IL-2 produced by Th cells [130, 133]. Besides those effects, Tregs produce many anti-inflammatory cytokines such as IL-10, transforming growth factor (TGF)-$\beta$ and IL-35 that help to produce and maintain an immune-suppressive environment [134, 135, 136].

Nevertheless, there are still many unanswered questions regarding to pathways that trigger Treg activation, control their suppressive function and enable Tregs to behave in an antigen-specific manner. Be that as it may, as a result of its functions, Tregs prevent autoimmune diseases, suppress allergies and induce tolerance to dietary antigens [137, 138, 139]. Their significance is further supported by studies that show that depletion of Tregs or their key genes leads to the emergence of autoimmune diseases [140, 141, 142, 143].

## 5.2. Importance of epigenetics in Treg's phenotype

While the mechanisms by which Tregs mediate immune tolerance are only partly known, there exists more knowledge on how Treg's phenotype is induced at the molecular level. The main transcription factor needed for the development of Tregs is the forkhead box protein P3 (FOXP3) [144, 128]. It acts as both – a transcriptional repressor and activator depending on its interaction partners [145, 146, 147, 148]. It is a key player in the induction and maintenance of immunosuppressive properties of Treg cells. It has been shown that mutations in that gene region lead to autoimmune diseases [140]. The expression of genes that are needed for the development of Tregs are governed by epigenetic changes that control chromatin structure and hence the accessibility of regulatory regions on DNA [149]. Previous methylation analyses have shown many differences between the methylation profiles of Tregs and Tconvs [150, 151, 152]. One of the most prominent examples is FOXP3 demethylation, particularly in the first intron, called "conserved non-coding sequence 2" [150, 151]. Therefore, it is evident that epigenetic

factors play an essential role in Treg's development.

## 5.3. Study design and methods

In our study, we obtained DNA methylation profiles for two groups of individuals using Illumina EPIC platform. After quality control and preprocessing, the first group comprised of Tregs and Tconvs extracted from 5 female and 1 male samples. The second group contained Tregs from 8 female and 3 male Graves' patients, and 7 female and 3 male healthy controls. We preprocessed the data with R package minfi [74] using the IDAT files from HiScanSQ scanner.

Microarray data analysis involves many steps of quality control and data filtering (see Chapter III) which have been tailored to be quite straightforward. For example, detection p-values can be obtained for each single CpG probe for each sample separately, and hence, we excluded samples whose average detection p-value was >0.05. After that, we performed data normalization using "noob" (normal-exponential out-of-band probes) [75] method for background correction and dye-bias normalization. Secondly, we removed unwanted variation by regressing out variability explained by the control probes and preserved the global methylation differences using the between-array normalization method called "functional normalization" [76].

After normalization, we removed CpGs based on the following criteria:

I. detection p-value $>0.01$ in at least one sample;
II. mapped to sex chromosomes, except for separate analysis on X chromosome;
III. polymorphic CpGs;
IV. cross-reactive probes [66]

After data preprocessing, the total number of CpG sites for downstream analyses was 766,126 for Treg vs Tconv comparison and 726,882 for Graves' patients vs healthy control comparison. In order to add biological information to the data we annotated CpG sites using an annotation package IlluminaHumanMethylationEPICanno.ilm10b4.hg19.

As an additional step of QC, we used a multidimensional scaling algorithm on 10,000 CpGs that had the largest standard deviations across the samples and calculated principal component 1 that captures most of the variance in the data. Based on that, we excluded samples further than 3 SD away from the group mean. Since outliers affect also normalization, we did the normalization and filtering steps described above again using slightly fewer samples. Given that those samples were within the limits of allowed p-values in the standard quality control steps, this implies that the standard threshold criteria might be a bit too loose sometimes.

Subsequently, we performed statistical analysis on M-values. As explained in Chapter III, the proportion of methylation, denoted as $\beta$, is useful for visualization and interpretation, however, it is more appropriate to use M-values in the

case of statistical calculations [77]. In order to find differentially methylated positions (DMPs), we relied on R package limma [153] and used it to fit a linear model to each CpG. Since Tregs and Tconvs were extracted from the same individual, we included two variables to each model – individual and cell type, thus enabling us to control for individuals' differences. After obtaining moderated t-statistics and associated p-values for each CpG, we adjusted those using the FDR method. Similarly, we used limma to find differentially variable positions (DVPs). Variability was calculated for each CpG in each sample by subtracting the group mean and taking the absolute deviation. After that, a linear model was then fitted to the absolute deviations and statistical significance was determined. In addition, we identified differentially methylated regions (DMRs) using R package DMRcate [154] that also relies on limma to estimate differential methylation at individual CpG site level. However, this method passes t-statistics obtained using limma funcitons to Gaussian smoothing function and uses smoothed statistics to compute p-values which is followed by multiple testing correction. Lastly, DMRs are identified by grouping statistically significant CpGs in such a way that the next consecutive significant CpG is less than prespecified number of nucleotides away, which our case was set to 500.

## 5.4. Main findings

Firstly, when we studied the distributions of genome-wide methylation levels in Tregs and Tconvs, we noticed that there is a rather large difference between those two. Both of them follow a bimodal distribution characteristic to cellular DNA methylation, however, Tconvs had a higher number of extremely methylated positions while Tregs had more CpG sites with intermediate methylation levels (Fig 12A). Moreover, we observed differences in the global variability of methylation levels between Tregs and Tconvs (Fig 12B), likely implying more diverse subpopulations in Tregs. Albeit, it is also possible, that in Tregs, there is an active regulation that causes the methylation level distribution to move from opposite modes into the middle area and thus a global epigenetic modifier could induce the methylation change in Tregs. For example, the enrichment analysis of transcription factors using TRANSFAC data on gene sets containing hypo- and hypermethylated DMPs showed enrichment of several DNA methylation associated transcription factors of which Kaiso was the most interesting. More precisely, Kaiso is expressed in Tregs, is known to bind methylated DNA [155], and has been suggested to regulate DNA methylation homeostasis [156]. Therefore, in light of this, Kaiso might play an important role in the formation of Treg-specific demethylation pattern.

Methylation differences between Tregs and Tconvs have been studied previously several times [152, 157, 158], however, our analysis yielded considerably more DMPs. Of course, those differences can be explained by our slightly larger sample size as well as platform differences (our platform of choice had more CpG

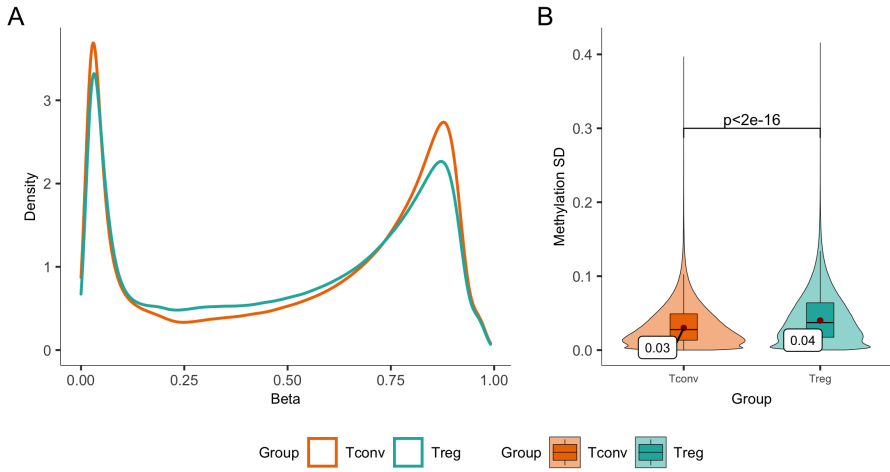


**Figure 12.** (A) The density plot of DNA methylation levels across all studied sites in Tregs (green) and Tconvs (orange). (B) Boxplots indicating higher variability of DNA methylation in Tregs in terms of invididual CpG's standard deviation.

sites to begin with). In total, we found 16,624 DMPs on autosomes, and in agreement with previous studies, the majority of those CpG sites were hypomethylated in Tregs and preferentially located in gene bodies and not in promoter regions or CpG islands. In addition, we also found over 3000 DMRs with nearly 60% of them being hypomethylated in Tregs (Fig 13). DMRs are genomic regions that contain closely located consecutive CpGs with similar methylation patterns. Concerning DVPs, our analysis resulted in a much modest number – only found 27 with 24 of them having higher variance in Tregs.

Like previous studies, we found methylation changes at so-called Treg signature genes, such as IKZF2, TIGIT, IKZF4, CTLA4, BCL11B, IL2RA and TNFRSF9 whose functions are well described in Tregs. A separate analysis on the X chromosome using samples from female participants revealed hypomethylation in FOXP3 gene intron 1, close to the CNS1 and CNS2 regions (Fig 14). Both – top DMPs and top DMR mapped to this region. FOXP3 encodes the well-known master regulator that is required for Treg development and maintenance of its suppressive functions [159, 160]. The hypomethylation in FOXP3 intron 1 maintains its expression in Tregs [161]. In addition, we found hypermethylation in an upstream region of THEMIS, a gene that is strongly downregulated by FOXP3 in response to TCR stimulation, thus suggesting that its suppression may be a part of the Treg developmental program [145, 162].

As the protein encoded by TSHR is a well-established autoantigen in GD, it was interesting to find hypomethylated CpGs within CEP128/TSHR region in Tregs (Fig 15). More precisely, GD causes hyperthyroidism which is a condition in which TSHR specific autoantibodies lead to excess thyroid hormone synthe-

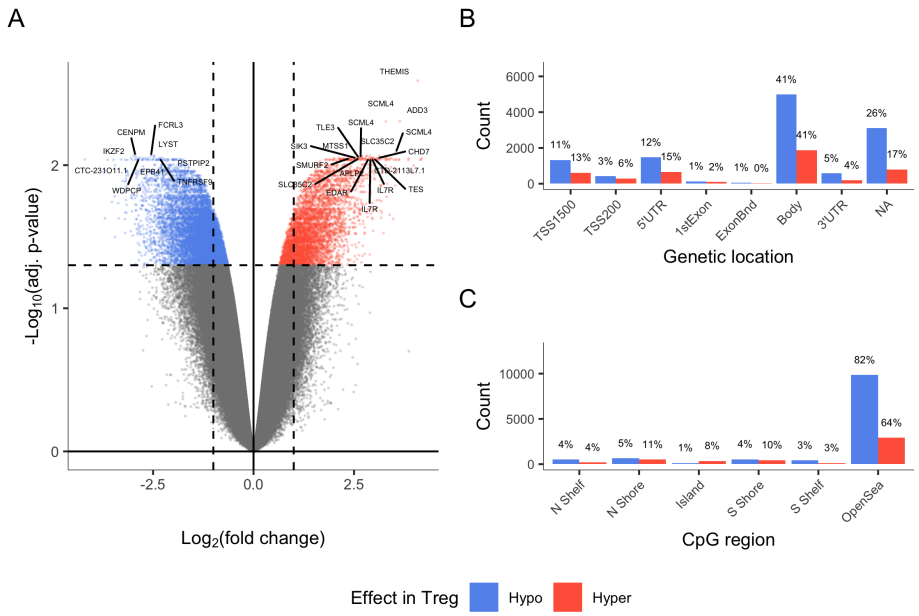


**Figure 13.** (A) Results of differential analysis summarised on a volcano plot. DMPs hypomethylated in Tregs are coloured red and hypermethylated blue. The y-axis shows $-\log_{10}$ of FDR adjusted p-values from differential analysis and x-axis shows $\log_2$ of fold change. (B) Genetic location and (C) CpG island relation of hyper- and hypomethylated DMPs illustrated using bar plots. The y-axis shows total count of CpGs in that particular region while text on bar plots indicates proportions.

sis and secretion [163]. In addition, there have been several articles that report reduced levels of Tregs in GD [164, 165, 166]. That incited us to explore those genes more closely. Those two aforementioned neighbouring genes are located head-to-head, and the CEP128 region has been identified as a crucial risk locus for autoimmune thyroid diseases [167]. As those two genes are in close proximity, their 5' regulatory regions are 14 kb apart, it is possible, that polymorphisms in the CEP128 gene regulatory region may affect TSHR's expression. In addition, several studies have found associations between GD and SNPs in the TSHR gene [124]. It is likely that the TSHR locus hypomethylation is specific to Tregs among other T cell populations, especially considering that our analysis also showed that the TSHR gene is expressed in Tregs, particularly its memory subset, but not in Tconvs. Moreover, our earlier analysis of bulk DNA methylation in $CD4^+$ T cells showed hypermethylation of TSHR intron 1, which was increased in GD patients [168].

Currently, there are at least two different (albeit not mutually exclusive) scenarios to explain the TSHR gene's associations with GD [169, 170, 171, 172]. Firstly, it has been proposed that GD-associated SNPs regulate mRNA splicing of TSHR that results in increased expression of more autoantigenic TSHR subunit

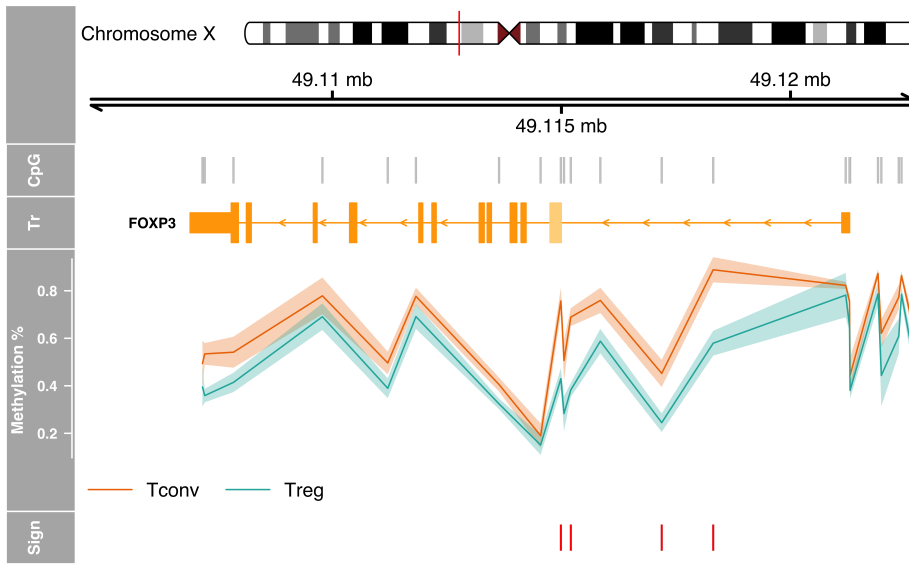


**Figure 14.** Methylation levels of CpGs in FOXP3 region indicate hypomethylation in intron 1. It is shown together with chromosomal information and information about exons and introns in the primary transcript (track denoted as "Tr"). The location of DMPs is shown with red stripes in "Sign" track and all CpGs in that region are present in the "CpG" track.

in the thyroid. Secondly, it has been suggested that the disease-associated SNPs lead to an increased risk of autoimmunity by reduced expression of TSHR in the thymus that results in the escape of TSHR-reactive T cells from central tolerance. We, however, thought that TSHR locus demethylation and upregulation in Tregs might have a role in immune tolerance to TSHR. Therefore, we carried out additional analysis of Treg DNA methylation profiles in healthy individuals and GD patients. Unfortunately, that resulted only in 19 differentially methylated CpG sites with no difference in TSHR locus. It is still possible that DNA methylation in the TSHR locus could be different in certain Treg subtypes, therefore, this remains an interesting topic for further studies.

## 5.5. Summary and impact

In this article, we studied DNA methylation profiles in $CD4^+$ $CD25^+$ Tregs and $CD4^+$ $CD25^-$ conventional T cells from healthy individuals and found >16k CpG sites to be differentially methylated between those two. To our knowledge, this is considerably more than previously reported. In this analysis, we show demethylation in many Treg signature genes as well as show that DMPs are enriched with transcriptional repressor Kaiso binding motifs. Additionally, we highlight that CpGs in Tregs are less defined in their methylation status at the cell population level. While this observation has been evident from other studies it has never been

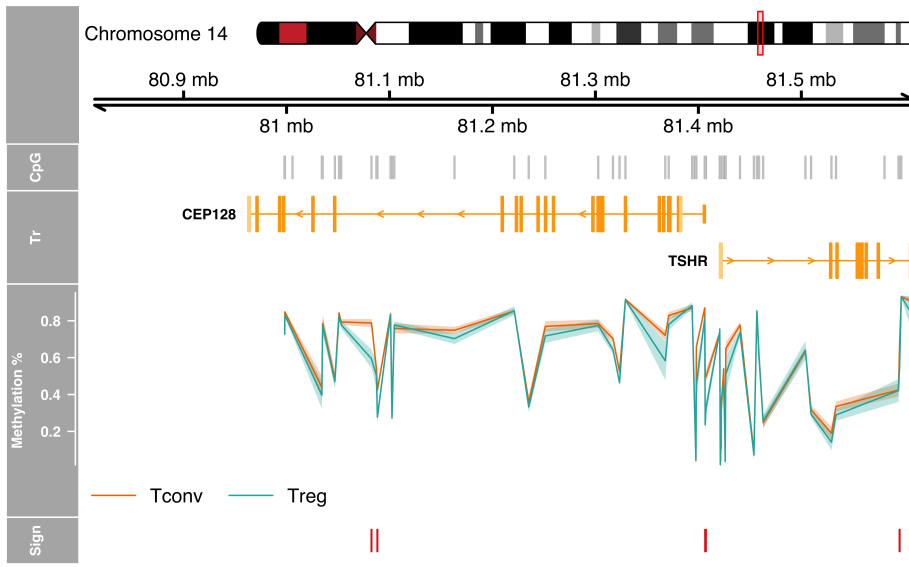


**Figure 15.** Hypomethylated CpGs are present in CEP128/TSHR region which is a risk locus for autoimmune thyroid diseases. It is shown together with chromosomal information and information about exons and introns in the primary transcript (track denoted as "Tr"). The location of DMPs is shown with red stripes in "Sign" track and all CpGs in that region are present in the "CpG" track.

addressed further. Here we characterized it a bit more. Perhaps most interestingly, we found hypomethylation and increased expression of the TSHR gene in Tregs compared to Tconvs. TSHR is a known risk gene for GD and thus prompted us to do subsequent DNA methylation profiling in healthy individuals and GD patients. Although the DNA methylation differences between Tregs from healthy individuals and GD patients were small, revealing only 19 DMPs, we hope that our study will encourage further research in this area that could explain the genetic link and role of anti-TSHR autoantibodies in GD besides the two prevailing views.

## 5.6. Contribution

My contribution to this publication is the following: I performed the preprocessing and analysis of the data, interpreted the results, made figures, and co-wrote the manuscript. Also, I participated in the critical review of the paper.

# 6. ANALYSIS OF THYMIC EPITHELIAL CELLS' PROTEOME AT VARIOUS DIFFERENTIATION STAGES (PUBLICATION III)

While the previous papers discussed $CD4^+$ and $CD8^+$ T cells, the third paper titled "Post-Aire Medullary Thymic Epithelial Cells and Hassall's Corpuscles as Inducers of Tonic Pro-Inflammatory Microenvironment" focuses on the environment where T cells develop. As was mentioned in Chapter II, the development of T cells takes place in the thymus. The development of T cells itself is a very complex process with many discoveries yet to be made. This process starts with T lymphocyte precursors entering the thymus from the bloodstream and involves rearrangement and expression of TCR genes, expression of CD4 and CD8 receptors as well as positive and negative selection. In this maturation process, many different cell types are involved but one of the most crucial roles is played by medullary thymic epithelial cells (mTECs). Those cells express a protein called AIRE (autoimmune regulator) that makes self-antigens, normally expressed only in certain peripheral organs, available to developing T cells. This process is crucial for creating central tolerance and thus is extensively studied. However, less attention has been paid to mTECs after they lose AIRE expression (post-AIRE cells), and here we aimed to characterize their role further. In addition, as there is evidence from previous studies showing similar protein expression during mTEC and keratinocyte differentiation we wanted to see whether those similarities still hold when the entire proteome is analysed.

## 6.1. Background and motivation of the study

The T-cell receptor (TCR), a protein complex capable of recognizing antigen peptides bound to the major histocompatibility complex (MHC), is responsible for the activation of the T cells [173]. T cells can recognize a vast array of different peptides and this is achieved through the random joining of gene segments in the TCR gene loci via process called V(D)J recombination [174]. It is the utmost importance, that T cells can distinguish between peptides originating from healthy endogenous cells and peptides from foreign sources such as pathogens. Therefore, as V(D)J recombination can also produce harmful T cells causing autoimmune reactions, those T cells need to be either removed or altered functionally before they exit the thymus and go into circulation. Those two processes comprise negative selection and the thymic regulatory T cell (Treg) induction. Both of those processes take place in thymic medulla and are collectively known as mechanisms of central tolerance [175].

The mechanisms of central tolerance depend on complex interactions between thymic cells. The key role in central tolerance induction is played by mTECs that a have unique property – they express a vast variety of different proteins of which

many are normally expressed only in certain tissues [176]. This process, controlled by a transcriptional regulator AIRE, makes possible for mTECs to present those peptides on MHCI and II complexes to immature T cells. Those immature T cells that strongly bind to self-peptides, and thus are potentially dangerous, are eliminated either by apoptosis that is triggered by this high affinity binding or directed towards Treg lineage [175].

Since AIRE expressing mTEC are crucial for central tolerance, they have been extensively studied [177]. In addition, as those cells are functionally mature cell populations, they were considered the endpoint of this lineage. However, recent single-cell RNA sequencing analyses in mice [178] and humans [179] have highlighted that mTECs can be divided into several subpopulations. Those studies (as well as many others that are more extensively covered in the article [55]) collectively suggest that the most studied AIRE-expressing mTECs differentiate into post-AIRE mTECs and then further to Hassall's corpuscles (HCs). Post-AIRE mTECs are characterized by the downregulation of AIRE and its related proteins, and thus lose the ability to express a broad range of self-peptides. In addition, they also obtain corneocyte-like (terminally differentiated keratinocytes) phenotype by expressing proteins associated with end-stage keratinocytes [180, 181, 182]. HCs are structures formed by concentrically arranged unnucleated cells with a typical diameter of 20-100 $\mu$m [183]. The cells in those structures also show expression of corneocyte-associated proteins and no AIRE expression [184, 185]. Currently, the role of post-AIRE cells is not clear, but there is some evidence that they could be involved in creating the tonic pro-inflammatory microenvironment in the thymus. This thymic microenvironment is characterized by constitutive low-grade expression of proinflammatory cytokines [186, 187, 188] and is believed to play a role in shaping the repertoire of T cells [189, 190].

Motivated by previous results described above, we wanted to study how similar are mTEC and keratinocyte end-stage differentiation and to see whether we could find hints about the role of post-AIRE mTECs and HCs.

## 6.2. Study design and methods

We obtained proteomics data of human thymi from three cardiac surgery patients with an age range of 1-2 years. More precisely, we microdissected three distinct morphological thymic areas to study three consecutive differentiation stages of mTECs (Fig 16):

I. thymic medulla (labeled as "mTECs" in the analysis)

II. epithelial layer immediately surrounding the HCs and characterized by flattened nuclei (labeled as "late mTECs")

III. HCs

Additionally, as we wanted to compare mTEC differentiation to keratinocyte differentiation, we obtained three consecutive differentiation stages of human epi-

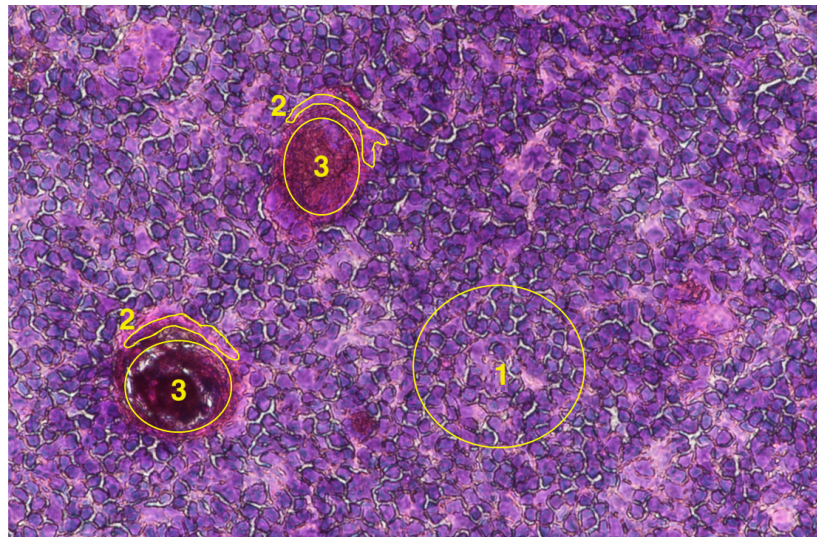


**Figure 16.** Microdissected areas from the thymic samples. Example of morphological areas where microdissection was performed: 1 – mTEC, 2 – late mTEC, 3 – HC

.

dermal keratinocytes. We extracted the following samples from healthy grown-ups:

I. stratum basale (labeled as "Basale")

II. stratum spinosum (labeled as "Spinosum")

III. stratum granulosum + stratum corneum (labeled as "Corneum")

We obtained proteomics data using nano-flow LC-MS/MS (described in Chapter III). This experiment as well as subsequent processing with MaxQuant [78] was done by the proteomics core facility of University of Tartu. After receiving the data containing normalized LFQ intensities, we filtered out data using the following criteria to limit the number of false positives:

I. Q-value $>0.01$

II. Peptides $>2$

Before data analysis, we $\log_2$-transformed the LFQ intensities to reduce the effect of outliers. After that, we used limma [153] for differential analysis and since limma can work with missing values, we did not incorporate separate imputation step into the analysis pipeline. Altogether, we carried out 6 differential analyses by studying differences between the first and second, first and third, and second and third differentiation stage in both thymic and skin samples. We adjusted obtained p-values for multiple testing correction with FDR, however, due to the small sample size (3 in each group), we made the analysis less stringent and set significance level to 0.1 instead of common 0.05. Our downstream analysis involved gene enrichment analysis using g:profiler R package [191] but also hierarchical clustering and PCA to draw parallels between keratinocyte and mTEC

differentiation. In addition, in order to find similarly changing proteins in the thymus and skin, we calculated a mean of protein expression level in each differentiation stage for each protein. Then calculated fold changes from the first to second and from the second to third differentiation stage. Subsequently, we clustered those values using Euclidean distance to find the most similarly changing proteins in those two tissues. Moreover, to further add confidence that keratinocyte and mTEC differentiation resemble, we used permutation tests and calculated p-values for the observed overlaps between differential analyses of those two sources. More specifically, we sampled 10,000 times the same number of proteins as were significant in the thymic and skin differential analyses (using $\alpha$ = 0.1). Then, we calculated overlaps between the resulting two sets and used the overlap as a test statistic to form a probability distribution.

## 6.3. Main findings

Firstly, our results showed several similarities between mTEC and keratinocyte differentiation. For example, in both cases, we observed upregulation of proteins involved in exocytosis ($p = 5.9 * 10^{-6}$ and $3.5 * 10^{-4}$ in the thymus and skin, correspondingly) and extracellular exosome generation ($p = 1.8 * 10^{-16}$ in the thymus and skin $9.2 * 10^{-5}$). Our clustering analysis showed that samples from the corneum (stratum granulosum and stratum corneum) clustered together with HCs instead of other keratinocyte stages. Moreover, by using permutation tests, we saw that the overlap of significant results from thymic and skin differential analyses was statistically significant in all cases – among the upregulated, downregulated and all significant proteins combined. It was also interesting to see that the late stages of differentiation in both thymus and skin were marked by increased expression of two autoimmune skin blistering associated auto-antigens – EPPK1 and A2ML1 [192, 193]. In addition, by studying similarly changing proteins in the skin and thymus, we found similar expression patterns among many epidermis and nucleus-related proteins as well as among collagens as is illustrated in Fig 17.

As described above, the thymus is characterized by a tonic proinflammatory microenvironment. In light of that, it was interesting to see the upregulation of proteins involved in inflammatory processes during mTEC differentiation. More precisely, our gene enrichment analysis showed enrichment of immune system related terms such as leukocyte mediated immunity ($p = 2.6 * 10^{-10}$) and innate immunity ($p = 1.4 * 10^{-5}$), but also upregulation of S100 family proteins was evident. This protein family contains 25 members that are involved in various roles ranging from energy metabolism to $Ca^{2+}$ homeostasis but also play a role in inflammation [194]. In fact, many of the S100 family proteins are involved in the production of proinflammatory cytokines [194]. Moreover, in our analysis, we saw that HCs and late mTECs were expressing more S100A8 and S100A9 proteins compared to mTECs (Fig 18) that induce the expression of several proinflammatory cytokines via binding to Toll-like receptor (TLR) 4 [195]. We also

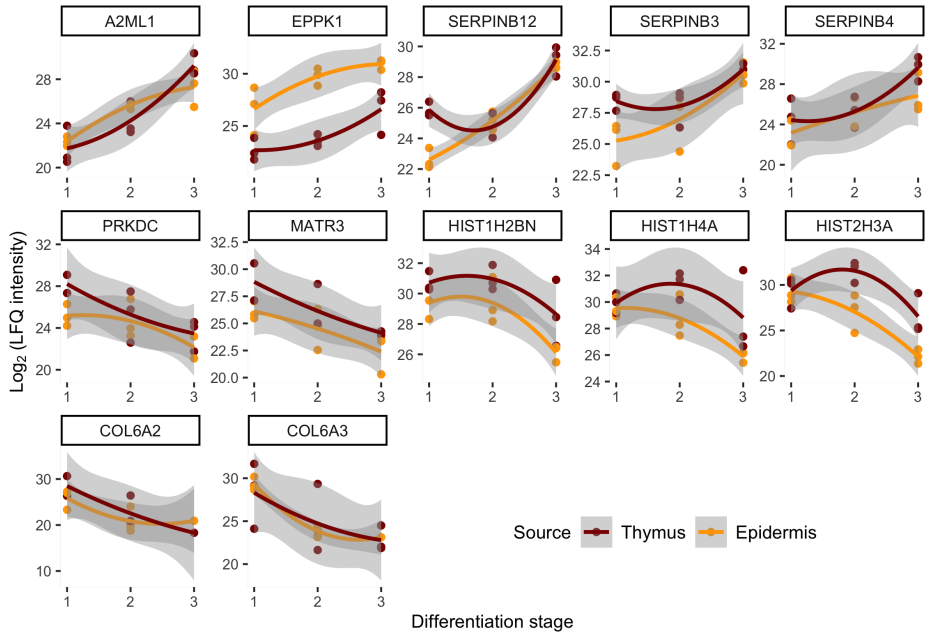


**Figure 17.** Expression levels of selected proteins during various differentiation stages in epidermis (yellow) and thymus (dark red). y-axis indicates $\log_2$ intensity levels (LFQ) and x-axis shows differentiation stage. The differentiation stages 1, 2, 3 denote mTEC, late mTEC, HC in thymus and in epidermis correspond to stratum basale, stratum spinosum and stratum granulosum + stratum corneum. The loess regression lines connect the means in those stages thus representing the average change in protein levels.

noticed the upregulation of TLR4 binding related processes at the late stages of mTEC differentiation. It is important to note that TLR4 signalling pathway leads to the production of type I interferons [196] and the expression of type I interferons has been observed in normal human thymi [187]. Type I interferons themselves are pleiotropic cytokines with immunoregulatory functions, and their function, either as proinflammatory or anti-inflammatory, depends on the biological context [197]. In the thymus they have been shown to be relevant for the survival of thymocytes [198] as well as for promoting Treg development [199]. Therefore, our findings support the notion that post-AIRE mTECS and HCs are involved in establishing the tonic inflammatory environment and contributing to the development of thymocytes.

## 6.4. Summary and impact

Our analysis is the first proteome analysis of HCs and we confirmed that many findings previously noticed in mice are also present in human tissues. We also found two clinically relevant autoantigens, that are expressed only in the latest stage of mTEC differentiation, and they may have a functional role in HCs. In

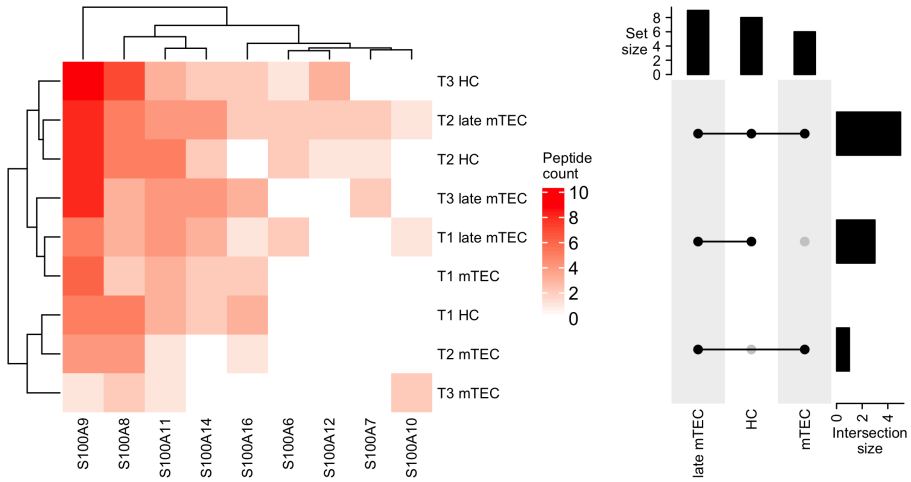


**Figure 18.** Clustered heatmap of detected peptide counts of S100A family proteins in all studied samples. Upset plots represent the overlaps of detected proteins (count $\geq 1$ included) between three differentation stages.

addition, our study supports the view that late mTEC and HC are involved in creating the tonic inflammatory environment in the thymus and proposes that they mediate their effect via S100A8 and S100A9 proteins. In agreement with previous studies that suggest similarities between mTEC maturation and keratinocyte differentiation, we showed that many common proteins are behaving similarly in those two tissues. The fact that they appear quite similar also brings along an interesting question from the developmental biology side that, however, remained unexplored in this study. Namely, while keratinocytes and mTECs emerge from two different embryonic layers – ectoderm and mesoderm, how do they still reach a similar phenotype?

## 6.5. Contribution

I contributed to this article by doing the data analyses after maxQuant normalization and by preparing the figures. In addition, I presented the results to co-authors and participated in the critical review of the paper.

# 7. ANALYSIS OF SARS-COV-2 IMMUNE RESPONSES AND ASSOCIATED INFLAMMATION (PUBLICATIONS IV-V)

While the previous chapters were mostly focusing on either specific immune cells or cells crucial for immune system development, the current chapter is based on two articles that studied severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In Paper IV "Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19" we longitudinally studied blood inflammation markers, antibodies, and plasma proteins of COVID-19 patients with different clinical manifestations. In this study, we aimed to characterize the inflammation happening during COVID-19 over the course of the disease as well as to find biomarkers distinguishing between various degrees of severity of this infection. In Paper V titled "Long-term elevated inflammatory protein levels in asymptomatic SARS-CoV-2 infected individuals", we aimed to gain an understating of the persistence of immune responses after 7-8 months of infection in asymptomatic individuals. Given that this virus has caused a global pandemic and received a lot of media coverage for several years, it hardly needs any general introduction. However, as I am referencing specific proteins of SARS-CoV-2 and aspects of its infection while giving an overview of the two aforementioned articles, basics of this virus structure, life cycle, and associated clinical features should be covered.

## 7.1. Background and motivation of the study

SARS-CoV-2 is a highly contagious positive-sense single-stranded RNA (+ssRNA) virus that causes COVID-19. SARS-CoV-2 belongs to the family of betacoronaviruses, its genome size is roughly 30kb and being a +ssRNA virus, the genome itself is an mRNA that can be translated directly into proteins [200]. The genome of SARS-CoV-2 is known to encode 29 proteins which include 4 structural proteins that make up the viral particle, 16 non-structural proteins that play a variety of roles related to the regulation of viral RNA replication and transcription, and finally 9 accessory proteins that interact with host cells and help the virus to evade host defenses [201].

The structural proteins comprise nucleocapsid (N) protein that packages RNA genome [202], the M protein that promotes virus assembly and defines the shape of the viral envelope [203], the E protein, that functions by interacting with other structural proteins and is involved in various aspects of viral replication cycle [204], and the spike (S) protein, that binds to host cell surface receptors and facilitates the entry of the virus into the host cell [205]. Due to its indispensable role, the S protein is also the most extensively studied and is the main target for vaccines. It contains two subunits – S1 and S2. The former contains the receptor-binding do-

main (RBD), which binds to angiotensin-converting enzyme 2 (ACE2) on target cells and the latter, the S2 subunit, is responsible for anchoring the S protein to the membrane and mediates membrane fusion thus enabling the virus to enter the cell [205].

Briefly, the life cycle of SARS-CoV-2 (Fig 19) is following, firstly, a viral particle binds to the host ACE2 receptor, a receptor that is abundantly present on the respiratory epithelium, via the S protein. Then, the transmembrane serine protease 2 (TMPRSS2) on the host cell membrane primes the S protein which in turn leads to the fusion of viral and cellular membranes via the S2 subunit. This process results in the release of viral genomic RNA into the host cell. Once the viral RNA is in the cell, host ribosomes get hijacked and viral RNA is translated into polyproteins. Those polyproteins get cleaved by proteases and produce components of the viral polymerase (RNA-dependent RNA polymerase). Then viral RNA replication and transcription are initiated producing more copies of genomic RNA-s and subgenomic mRNAs. Genomic RNA-s serve as genomes for new viral particles while subgenomic RNAs are translated into structural and accessory proteins at the membrane of the endoplasmic reticulum (ER). Then, the genomic RNA gets bound by N proteins and combines with an ER-derived compartment that mediates trafficking between the ER and the Golgi complex. After that, a mature virion is formed and exported from the host cell by exocytosis. Following this, the released virion becomes ready for another round of infection [206, 207].

The infection caused by SARS-CoV-2 ranges from asymptomatic to critical illness with acute respiratory distress syndrome [200]. Studies have shown that asymptomatic cases comprise approximately 18-33% of infected individuals [208, 209]. In symptomatic individuals, the pneumonia caused by SARS-CoV-2 is characterized by an early and late stage. The early stages consist of virus replication and tissue damage caused directly by the virus. The late stage corresponds to host immune response that involves the recruitment of immune cells such as monocytes, neutrophils, and T cells but also elevated levels of cytokines such as IL-6, (TNF-$\alpha$), interferon (IFN)-$\gamma$, and many others [200]. Severe infection is often caused by overactivation of the immune system that manifests itself in cytokine storm that is characterized by high levels of cytokines such as IL-6 and TNF-$\alpha$ [210, 211]. Although, it should be noted that the cytokine storm is a poorly defined concept and its role in severe COVID-19 has been criticised since often the levels of cytokines do not reach the levels commonly associated with cytokine storm [212].

While all people are at risk of being infected and getting a severe disease, it is the elderly people with pre-existing medical conditions such as obesity, chronic kidney disease, diabetes, etc, that have been considered as the main risk group [213]. In Paper IV, we used longitudinal analysis to identify whether differences in plasma protein or antibody levels could also explain why some people have a worse course of infection. In this article, we aimed to characterize the COVID-19 immune response by studying the levels of plasma proteins in patients who were admitted to

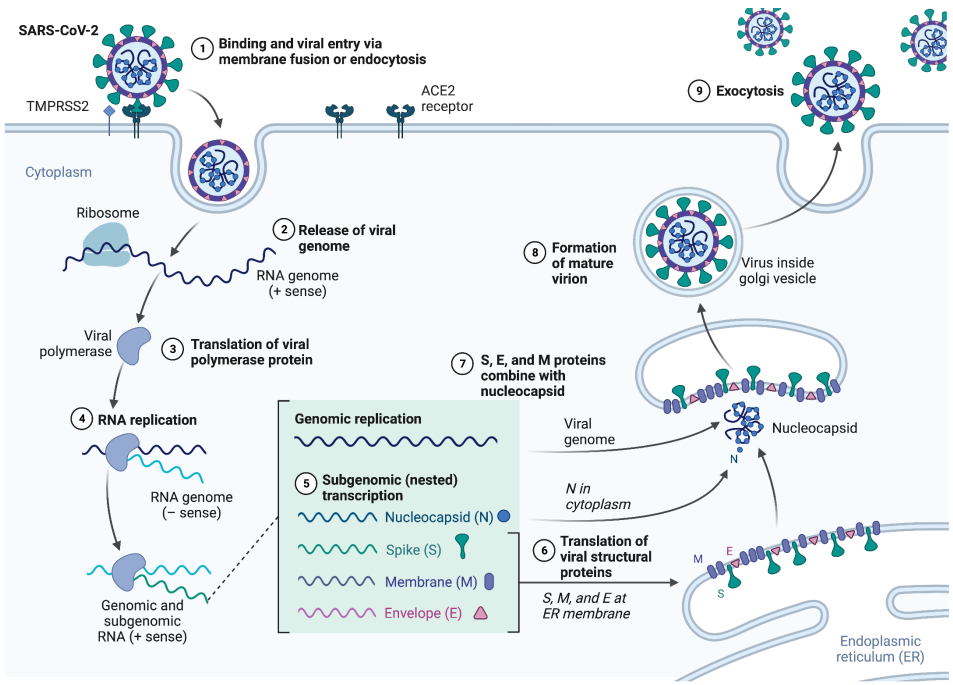


**Figure 19.** SARS-CoV-2 life cycle. Key steps in SARS-CoV-2 life cycle are numbered. This figure was made with Biorender.com by using template "Life Cycle of Coronavirus"

.

the intensive care unit (ICU), those who were treated in ordinary COVID-19 ward, and those who had mild disease and were not hospitalized. In Paper V, we studied mostly asymptomatic individuals 7-8 months after the infection. We explored how long the antibody levels sustain, whether there are any differences regarding the antibody's target, and how well they neutralize the virus. Besides that, we also wanted to see whether those individuals show any ongoing inflammation.

## 7.2. Design of the studies and methods

In article IV, we studied longitudinal plasma samples, collected almost on a daily basis, from 40 hospitalized COVID-19 patients at the Tartu University Hospital. 15 of those patients were treated in ICU and thus formed an ICU group while disease in the other 25 people was more modest and they were treated in the regular hospital ward, hence comprised the non-ICU cohort. The age of the patients ranged from 21 to 92 with a mean of 66. We compared those to 119 SARS-CoV-2 negative controls (age range 23-87) and in some instances, where possible, to patients with mild illness. We classified individuals as mild if they attended the emergency medical department but were not subsequently hospitalized.

In Paper V, we studied samples from 56 antibody positive mostly asymptomatic individuals at two time points. Those individuals were identified by a seroepidemiological study carried out between May 8 and July 31, 2020. Then in November 2020, samples were collected again from those people for subsequent analysis, corresponding roughly 7-8 months after their initial infection in March 2020 (outbreak in a region where samples were collected happened in March 2020). As a control group, we used 115 age and sex-matched uninfected individuals.

In both of those studies, we measured levels of inflammation-associated plasma proteins with Proximity Extension Assay produced by Olink Proteomics (described in Chapter III). The resulting data in NPX units is on $\log_2$ scale where a higher value indicates higher protein expression. As a preprocessing step, we excluded values below the assay's limit of detection (LOD). In both analyses, we were mostly interested in group-wise differences, and since the normality of data distribution could not be assumed, we relied on nonparametric tests. In Paper IV, we used a two-sample Wilcoxon test in order to determine statistical significance and additionally made use of the PCA to adjust the p-values. More precisely, since many of those proteins were highly correlated with one another, we calculated the number of independent components that would explain $>99\%$ of variance using PCA. As a next step, we used this number of components in Bonferroni correction to adjust p-values and thus reduce the probability of type I error. However, in Paper V, the statistical significance between groups was determined using the Kruskal-Wallis test in combination with Dunn's Test. More precisely, we used the Kruskal-Wallis test to find whether there was a difference between the medians of three analysis groups and then used Dunn's test to pinpoint which specific

medians were significantly different from others.

Regarding the exploratory analysis, we also used PCA for data visualization in order to understand which sets of proteins better explain the group-wise differences. Similarly, we used hierarchical clustering based on Euclidean distance to find individuals with similar protein expression profiles but we also used it to cluster correlation matrices of proteins expression to find similarly (or oppositely) behaving proteins. In our analysis, we mostly used nonparametric Spearman's rank correlation as a linear relationship often could not be assumed. However, we used Pearson's correlation to find proteins acting in a similar/opposite way since linearity, especially in a smaller dataset, enhances biological interpretability. As the data in Paper IV was longitudinally obtained and contained many timepoints, we used local regression, a nonparametric smoother, to get a polynomial trend line from otherwise noisy data to understand how protein levels changed over the course of the disease.

## 7.3. Main findings

In Paper IV, we studied longitudinal plasma samples from 15 ICU patients and 25 patients with moderate disease. We compared their measurements to healthy controls and when possible, also to individuals with mild disease, i.e., not hospitalized people. As expected, we saw that two main biomarkers of inflammation – C-reactive protein and procalcitonin, were highly elevated among ICU and non-ICU patients in comparison to controls (p-values $\leq 0.0001$ and $\leq 0.001$, correspondingly). Also, patients were characterized by increased neutrophil levels (p-value $\leq 0.001$ in ICU patients and $p \leq 0.05$ in the non-ICU cohort) and ICU patients also showed decreased lymphocyte levels ($p \leq 0.01$).

Our antibody analysis with the LIPS method showed that all patients developed IgG antibodies against S1, S2, N, and RBD during their hospitalization with an average of 13 days from the start of the disease. Interestingly, there was no significant difference in seroconversion time between ICU and non-ICU patients, however, there was a trend indicating that non-ICU patients reach peak levels earlier while the levels in ICU patients tended to reach higher numbers.

Further, using Olink assay, we found higher levels of many well-established inflammation markers associated with activated myeloid and T cells among ICU and non-ICU patients in comparison to controls. In most cases, we saw a trend indicating the highest average levels in ICU patients, followed by non-ICU patients and slightly lower levels among individuals with mild disease, and the lowest levels in controls. A triad on cytokines, IL-6—CXCL10—IL-10, which have been associated with disease severity [214], were also following this pattern and possibly indicating a strong inflammatory response. Cytokines of this triad achieved their maximal levels within 24–72 h after hospitalization which is approximately 10 days after the appearance of the first symptoms.

Besides differences in proinflammatory cytokine levels, we also noticed higher

levels in many apoptosis related proteins, such as CASP8, HGF, TNFSF14, and TGFB1, among COVID-19 patients. This phenomenon could be explained by the lymphopenia and T cell apoptosis reported in COVID-19 [215]. In addition, the most striking finding from Paper IV was related to the antiapoptotic protein called hepatocyte growth factor (HGF). It was significantly higher among COVID-19 patients, but it also separated ICU and non-ICU patients (Fig 20). HGF mostly functions as an antiapoptotic protein [216], and thus possibly reflecting a feedback mechanism. More precisely, as the severe COVID-19 is characterized by stronger lymphopenia, this apoptosis could drive the levels of antiapoptotic proteins such as HGF to inhibit excess apoptosis. In addition, an analysis published after our study showed that HGF together with CXCL13 were the best predictors for discriminating between ICU and non-ICU patients. In this analysis, HGF was slightly better than CXCL13 and showed 88.6% sensitivity and 81.5% specificity for classifying ICU-s and non-ICUs [36].

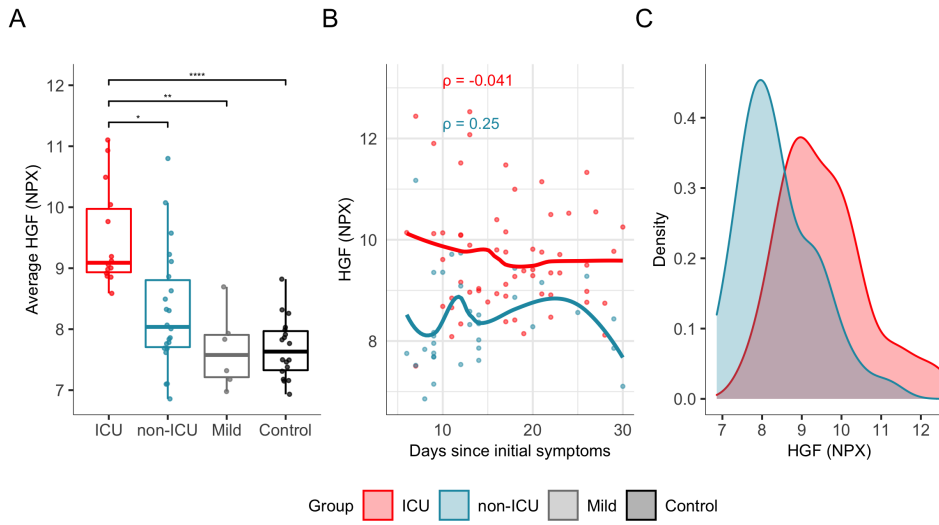


**Figure 20.** HGF levels in our study cohort. (A) Average levels of HGF in studied groups shown with boxplots. (B) Scatterplots of HGF levels over time with x-axis denoting the time in days since initial symptoms and y-axis showing the expression level of HGF. Group specific trends are shown with loess regresson lines. (C) Density plots illustrating the distributions of all measured HGF levels in ICU and non-ICU patients.

Finally, to find out whether our selected proteins (top 10 from early-stage inflammation markers and top 10 from apoptosis-related proteins, except IL7 due to low overall levels) could help to segregate different disease groups we clustered those proteins using patients' measurements at their peak levels. We identified 4 clusters of individuals (Fig 21) that indicated that severe disease was associated with the upregulation of inflammatory and apoptotic markers. Starting from the top, the first cluster of 4 individuals, characterized by the highest values overall, contained 3 ICU patients with acute respiratory distress syndrome (ARDS) and 1 ICU patient who died before receiving an ARDS diagnosis. The second clus-

ter contained 15 patients with 8 of them having severe disease. The third and fourth clusters contained mostly non-ICU patents and healthy controls, respectively. This heatmap also shows that patients with ARDS tend to have higher levels of HGF, IL6 and CCL7.

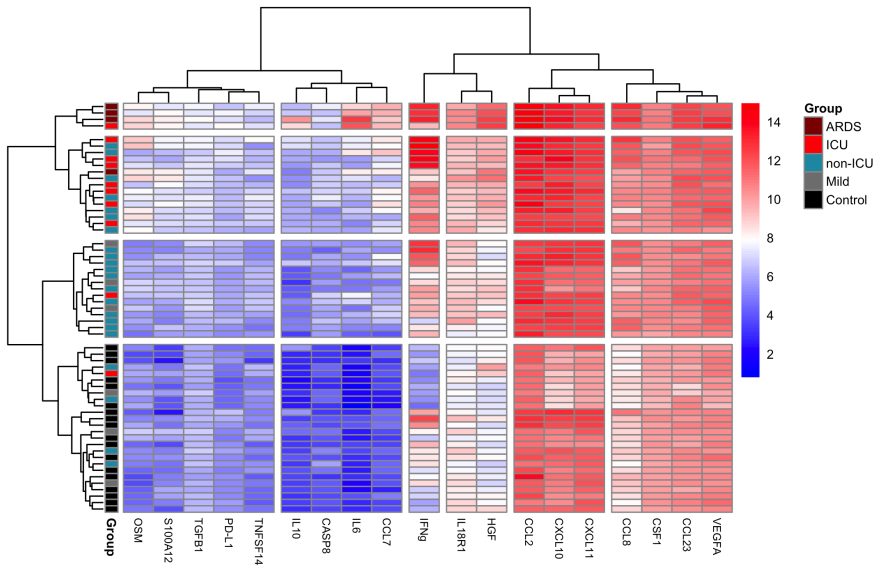


**Figure 21.** Clustered heatmap of 19 selected proteins constructed using maximal expression levels in 60 individuals. Annotation track indicates the analysis group where each of the studied individual belongs.

In the subsequent Paper V, we studied 56 mostly asymptomatic SARS-CoV-2 antibody positive individuals and conducted a reanalysis of them 7-8 months after their infection together with 115 age-matched seronegative controls. In this study, we focused on antibodies, T cell responses, and the levels of inflammatory proteins. We observed a decline in antibodies specific to the NC over time whereas antibodies specific to RBD remained constant. In addition, the levels of RBD specific antibodies also correlated with virus neutralization measured by $CD4^+$ T cell responses ($\rho = 0.54$, $p = 0.00014$). Regarding the analysis of inflammatory markers, we made an unexpected observation. Namely, asymptomatic individuals had elevated serum levels of S100A12, TGF-$\alpha$, IL18, and OSM proteins. Those plasma proteins are markers of activated macrophages and monocytes and thus their higher levels in months after infection suggest a persistent pulmonary inflammation.

## 7.4. Summary and impact

COVID-19 pandemic caused a massive influx of scientists into SARS-CoV-2 research. It resulted in an unprecedented growth of publications devoted to COVID-

19 which at the time of writing the thesis exceeded over one million according to the Dimensions database [217]. In the beginning, many knowledge gaps needed to be filled. Our analyses, albeit on smaller sample sizes, also enabled us to fill in some missing pieces and validate many others. In Paper IV, we showed an upregulation of many inflammatory cytokines that agreed with previous studies but also highlighted that COVID-19 is associated with the apoptotic pathway. Moreover, we showed that the antiapoptotic protein HGF is differently expressed in ICU and non-ICU patients, and subsequent study further highlighted its classification power between ICU and non-ICU patients. In Paper V we showed that while in general antibodies decline, the RBD-specific antibodies stayed on a higher level. In addition, our analysis indicated that even in asymptomatic individuals, there is a long-term upregulation of inflammatory proteins pointing out the need for post-infection clinical monitoring of SARS-CoV-2 infected asymptomatic individuals.

## 7.5. Contribution

I contributed to Paper IV by analysing the data, making figures, and co-wrote the article's results and methods sections. In Paper V, I made a preliminary analysis of a larger dataset that served as a starting point for the given article. In addition, I presented the results to co-authors and participated in the critical review of the papers.

# 8. CONCLUSION

In this thesis, I have presented a collection of papers that are a mix of immunology, molecular biology, statistics, and computer science. These publications, while focusing on different aspects of immunology, demonstrate how molecular biology research has become increasingly intertwined with data science, even in small research groups like ours. It is clear that this trend is continually growing and possessing the skills to effectively analyze, understand, and argue with data is becoming increasingly crucial for researchers in the field of biology. I hope that this thesis inspires those with a background in biology to continue to develop their skills in data science, and to explore the potential for new discoveries through this interdisciplinary approach.

The main findings from the articles are as follows:

- In Paper I, we conducted a comprehensive analysis of different types of T cells and found a high degree of interindividual variation among the $CD8^+$ TEMRA population. Our investigation revealed that while $CD8^+$ TEMRA levels increase with age, among the elderly, their levels do not correlate with age and are determined by other factors, such as CMV infection. We also found that this cell type can be predicted using DNA methylation levels of a few CpG sites, and hope that this model, or its subsequent adaptations, could serve as a tool for monitoring age-associated immune dysfunction.
- In Paper II, we studied DNA methylation profiles in regulatory (Treg) and conventional (Tconvs) T cells and found over 16,000 CpG sites to be differentially methylated. We also found that Tregs exhibited less defined global DNA methylation status, with more CpG sites having intermediate methylation levels. The most intriguing finding was the observation of hypomethylation in and nearby the thyroid-stimulating hormone receptor (TSHR) gene in Tregs together with its Treg-specific expression. Namely, this genetic location is a known risk locus for Graves' disease (GD), and although our subsequent profiling of Tregs from healthy individuals and GD patients did not reveal any difference in that genetic area, we hope that our research incites further studies to elucidate whether changes in this genetic locus affect Tregs' phenotype that in turn predispose the development of GD.
- In Paper III, we studied medullary thymic epithelial cells' differentiation and highlighted its similarities with keratinocyte differentiation. Besides that, our study supports the view that late mTECs and Hassal's corpuscles contribute to the creation of a tonic inflammatory environment in the thymus that is relevant for shaping the repertoire of T cells. Additionally, our analysis proposes that their involvement in this process is mediated by S100A8 and S100A9 proteins. Furthermore, we found two clinically relevant autoantigens that are expressed in the late stages of mTEC differentiation.

- In Paper IV and V, we contributed to the SARS-CoV-2 research. For instance, we showed that COVID-19 is associated with apoptotic pathways, especially in the case of severe disease. We showed that the antiapoptotic protein HGF has higher levels in patients that are subjected to the intensive care unit (ICU). It is noteworthy that other subsequent studies have further illustrated HGF's usefulness as a biomarker for distinguishing ICU and non-ICU patients. Additionally, we confirmed many previous findings indicating higher levels of proinflammatory cytokines among more severe disease states. Our analysis also indicated that there is a long-term upregulation of inflammatory proteins in asymptomatic individuals.

# BIBLIOGRAPHY


[1] Eva Bianconi et al. “An estimation of the number of cells in the human body”. In: *Annals of Human Biology* 40.6 (2013), pp. 463–471. DOI: 10.3109/03014460.2013.807878. URL: https://doi.org/10.3109/03014460.2013.807878.

[2] The Tabula Sapiens Consortium. “The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans”. In: *Science* 376.6594 (2022). DOI: 10.1126/science.abl4896. URL: https://doi.org/10.1126/science.abl4896.

[3] Tom Strachan and Andrew Read. “Human Molecular Genetics 5th Edition, Chapter 2”. In: (2018), pp. 41–67.

[4] Tom Strachan and Andrew Read. “Human Molecular Genetics 5th Edition, Chapter 1”. In: (2018), pp. 3–40.

[5] Tom Strachan and Andrew Read. “Human Molecular Genetics 5th Edition, Chapter 10”. In: (2018), pp. 325–356.

[6] Jaime L. Miller and Patrick A. Grant. “The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans”. In: (2012), pp. 289–317. DOI: 10.1007/978-94-007-4525-4_13. URL: https://doi.org/10.1007/978-94-007-4525-4_13.

[7] Steen K. T. Ooi et al. “DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA”. In: *Nature* 448.7154 (2007), pp. 714–717. DOI: 10.1038/nature05987. URL: https://doi.org/10.1038/nature05987.

[8] Jiaojiao Li et al. “Insights into S-adenosyl-l-methionine (SAM)-dependent methyltransferase related diseases and genetic polymorphisms”. In: *Mutation Research/Reviews in Mutation Research* 788 (2021), p. 108396. DOI: 10.1016/j.mrrev.2021.108396. URL: https://doi.org/10.1016/j.mrrev.2021.108396.

[9] Melanie Ehrlich et al. “Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells”. In: *Nucleic Acids Research* 10.8 (1982), pp. 2709–2721. DOI: 10.1093/nar/10.8.2709. URL: https://doi.org/10.1093/nar/10.8.2709.

[10] Michael J. Ziller et al. “Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types”. In: *PLoS Genetics* 7.12 (2011). Ed. by Dirk Schübeler, e1002389. DOI: 10.1371/journal.pgen.1002389. URL: https://doi.org/10.1371/journal.pgen.1002389.

[11] Ryan Lister et al. “Human DNA methylomes at base resolution show widespread epigenomic differences”. In: *Nature* 462.7271 (2009), pp. 315–322. DOI: 10.1038/nature08514. URL: https://doi.org/10.1038/nature08514.

[12] Yulia A Medvedeva et al. “Effects of cytosine methylation on transcription factor binding sites”. In: *BMC Genomics* 15.1 (2014). DOI: 10.1186/1471-2164-15-119. URL: https://doi.org/10.1186/1471-2164-15-119.

[13] Thomas Clouaire and Irina Stancheva. “Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin?” In: *Cellular and Molecular Life Sciences* 65.10 (2008), pp. 1509–1522. DOI: 10.1007/s00018-008-7324-y. URL: https://doi.org/10.1007/s00018-008-7324-y.

[14] Adrian P Bird and Alan P Wolffe. “Methylation-Induced Repression—Belts, Braces, and Chromatin”. In: *Cell* 99.5 (1999), pp. 451–454. DOI: 10.1016/s0092-8674(00)81532-9. URL: https://doi.org/10.1016/s0092-8674(00)81532-9.

[15] Zachary D. Smith and Alexander Meissner. “DNA methylation: roles in mammalian development”. In: *Nature Reviews Genetics* 14.3 (2013), pp. 204–220. DOI: 10.1038/nrg3354. URL: https://doi.org/10.1038/nrg3354.

[16] Aimée M. Deaton and Adrian Bird. “CpG islands and the regulation of transcription”. In: *Genes & Development* 25.10 (2011), pp. 1010–1022. DOI: 10.1101/gad.2037511. URL: https://doi.org/10.1101/gad.2037511.

[17] Richard Cordaux and Mark A. Batzer. “The impact of retrotransposons on human genome evolution”. In: *Nature Reviews Genetics* 10.10 (2009), pp. 691–703. DOI: 10.1038/nrg2640. URL: https://doi.org/10.1038/nrg2640.

[18] Serge Saxonov, Paul Berg, and Douglas L. Brutlag. “A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters”. In: *Proceedings of the National Academy of Sciences* 103.5 (2006), pp. 1412–1417. DOI: 10.1073/pnas.0510310103. URL: https://doi.org/10.1073/pnas.0510310103.

[19] Peter A. Jones. “The DNA methylation paradox”. In: *Trends in Genetics* 15.1 (1999), pp. 34–37. DOI: 10.1016/s0168-9525(98)01636-9. URL: https://doi.org/10.1016/s0168-9525(98)01636-9.

[20] Rafael A Irizarry et al. “The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores”. In: *Nature Genetics* 41.2 (2009), pp. 178–186. DOI: 10.1038/ng.298. URL: https://doi.org/10.1038/ng.298.

[21] Peter A. Jones. “Functions of DNA methylation: islands, start sites, gene bodies and beyond”. In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492. DOI: 10.1038/nrg3230. URL: https://doi.org/10.1038/nrg3230.

[22] Frank Lyko. “The DNA methyltransferase family: a versatile toolkit for epigenetic regulation”. In: *Nature Reviews Genetics* 19.2 (2017), pp. 81–92. DOI: 10.1038/nrg.2017.80. URL: https://doi.org/10.1038/nrg.2017.80.

[23] Peter A. Jones and Gangning Liang. “Rethinking how DNA methylation patterns are maintained”. In: *Nature Reviews Genetics* 10.11 (2009), pp. 805–811. DOI: 10.1038/nrg2651. URL: https://doi.org/10.1038/nrg2651.

[24] Silvina Epsztejn-Litman et al. “De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes”. In: *Nature Structural & Molecular Biology* 15.11 (2008), pp. 1176–1183. DOI: 10.1038/nsmb.1476. URL: https://doi.org/10.1038/nsmb.1476.

[25] H Yan et al. “piRNA-823 contributes to tumorigenesis by regulating de novo DNA methylation and angiogenesis in multiple myeloma”. In: *Leukemia* 29.1 (2014), pp. 196–206. DOI: 10.1038/leu.2014.135. URL: https://doi.org/10.1038/leu.2014.135.

[26] Xiaoji Wu and Yi Zhang. “TET-mediated active DNA demethylation: mechanism, function and beyond”. In: *Nature Reviews Genetics* 18.9 (2017), pp. 517–534. DOI: 10.1038/nrg.2017.33. URL: https://doi.org/10.1038/nrg.2017.33.

[27] Martin Bachman et al. “5-Hydroxymethylcytosine is a predominantly stable DNA modification”. In: *Nature Chemistry* 6.12 (2014), pp. 1049–1055. DOI: 10.1038/nchem.2064. URL: https://doi.org/10.1038/nchem.2064.

[28] Zelin Jin and Yun Liu. “DNA methylation in human diseases”. In: *Genes & Diseases* 5.1 (2018), pp. 1–8. DOI: 10.1016/j.gendis.2018.01.002. URL: https://doi.org/10.1016/j.gendis.2018.01.002.

[29] Abul K. Abbas, Andrew H. Lichtman, and Shiv Pillai. “Cellular and Molecular Immunology 10th Edition, Chapter 1”. In: (2021), pp. 26–59.

[30] Leonardo M. R. Ferreira. “Gammadelta T Cells: Innately Adaptive Immune Cells?” In: *International Reviews of Immunology* 32.3 (2013), pp. 223–248. DOI: 10.3109/08830185.2013.783831. URL: https://doi.org/10.3109/08830185.2013.783831.

[31] Abul K. Abbas, Andrew H. Lichtman, and Shiv Pillai. “Cellular and Molecular Immunology 10th Edition, Chapter 4”. In: (2021), pp. 220–351.

[32] Abul K. Abbas, Andrew H. Lichtman, and Shiv Pillai. “Cellular and Molecular Immunology 10th Edition, Chapter 8”. In: (2021), pp. 636–722.

[33] Abul K. Abbas, Andrew H. Lichtman, and Shiv Pillai. “Cellular and Molecular Immunology 10th Edition, Chapter 6”. In: (2021), pp. 429–526.

[34] Vita Golubovskaya and Lijun Wu. “Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy”. In: *Cancers* 8.3 (2016),

p. 36. DOI: 10.3390/cancers8030036. URL: https://doi.org/10.3390/cancers8030036.

[35] Rebecca N. Monastero and Srinivas Pentyala. “Cytokines as Biomarkers and Their Respective Clinical Cutoff Levels”. In: *International Journal of Inflammation* 2017 (2017), pp. 1–11. DOI: 10.1155/2017/4309485. URL: https://doi.org/10.1155/2017/4309485.

[36] Matthieu Perreau et al. “The cytokines HGF and CXCL13 predict the severity and the mortality in COVID-19 patients”. In: *Nature Communications* 12.1 (2021). DOI: 10.1038/s41467-021-25191-5. URL: https://doi.org/10.1038/s41467-021-25191-5.

[37] Carlos López-Otıén et al. “The Hallmarks of Aging”. In: *Cell* 153.6 (2013), pp. 1194–1217. DOI: 10.1016/j.cell.2013.05.039. URL: https://doi.org/10.1016/j.cell.2013.05.039.

[38] Janko Nikolich-Žugich. “The twilight of immunity: emerging concepts in aging of the immune system”. en. In: *Nature Immunology* 19.1 (2017), pp. 10–19. DOI: 10.1038/s41590-017-0006-x. URL: http://dx.doi.org/10.1038/s41590-017-0006-x.

[39] Claudio Franceschi and Judith Campisi. “Chronic Inflammation (Inflammaging) and Its Potential Contribution to Age-Associated Diseases”. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 69.Suppl 1 (2014), S4–S9. DOI: 10.1093/gerona/glu057. URL: https://doi.org/10.1093/gerona/glu057.

[40] Cláudia Cavadas et al. “The pathophysiology of defective proteostasis in the hypothalamus — from obesity to ageing”. In: *Nature Reviews Endocrinology* 12.12 (2016), pp. 723–733. DOI: 10.1038/nrendo.2016.107. URL: https://doi.org/10.1038/nrendo.2016.107.

[41] Alejo Efeyan, William C. Comb, and David M. Sabatini. “Nutrient-sensing mechanisms and pathways”. In: *Nature* 517.7534 (2015), pp. 302–310. DOI: 10.1038/nature14190. URL: https://doi.org/10.1038/nature14190.

[42] Randy L. Hunter et al. “Inflammation induces mitochondrial dysfunction and dopaminergic neurodegeneration in the nigrostriatal system”. In: *Journal of Neurochemistry* 100.5 (2006), pp. 1375–1386. DOI: 10.1111/j.1471-4159.2006.04327.x. URL: https://doi.org/10.1111/j.1471-4159.2006.04327.x.

[43] Diana Jurk et al. “Chronic inflammation induces telomere dysfunction and accelerates ageing in mice”. In: *Nature Communications* 5.1 (2014). DOI: 10.1038/ncomms5172. URL: https://doi.org/10.1038/ncomms5172.

[44] Audrey Lasry and Yinon Ben-Neriah. “Senescence-associated inflammatory responses: aging and cancer perspectives”. In: *Trends in Immunology*

36.4 (2015), pp. 217–228. DOI: 10.1016/j.it.2015.02.009. URL: https://doi.org/10.1016/j.it.2015.02.009.

[45] Carl Nathan and Amy Cunningham-Bussel. “Beyond oxidative stress: an immunologist’s guide to reactive oxygen species”. In: *Nature Reviews Immunology* 13.5 (2013), pp. 349–361. DOI: 10.1038/nri3423. URL: https://doi.org/10.1038/nri3423.

[46] Juhyun Oh, Yang David Lee, and Amy J Wagers. “Stem cell aging: mechanisms, regulators and therapeutic opportunities”. In: *Nature Medicine* 20.8 (2014), pp. 870–880. DOI: 10.1038/nm.3651. URL: https://doi.org/10.1038/nm.3651.

[47] Christine Nardini et al. “The epigenetics of inflammaging: The contribution of age-related heterochromatin loss and locus-specific remodelling and the modulation by environmental stimuli”. In: *Seminars in Immunology* 40 (2018), pp. 49–60. DOI: 10.1016/j.smim.2018.10.009. URL: https://doi.org/10.1016/j.smim.2018.10.009.

[48] Jian-Hua Chen, C. Nicholes Hales, and Susan E. Ozanne. “DNA damage, cellular senescence and organismal ageing: causal or correlative?” In: *Nucleic Acids Research* 35.22 (2007), pp. 7417–7428. DOI: 10.1093/nar/gkm681. URL: https://doi.org/10.1093/nar/gkm681.

[49] Ayelet Alpert et al. “A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring”. In: *Nature Medicine* 25.3 (2019), pp. 487–495. DOI: 10.1038/s41591-019-0381-y. URL: https://doi.org/10.1038/s41591-019-0381-y.

[50] Nazish Sayed et al. “An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging”. In: *Nature Aging* 1.7 (2021), pp. 598–615. DOI: 10.1038/s43587-021-00082-y. URL: https://doi.org/10.1038/s43587-021-00082-y.

[51] Ahto Salumets et al. “Epigenetic quantification of immunosenescent CD8+ TEMRA cells in human blood”. In: *Aging Cell* 21.5 (2022), e13607. DOI: 10.1111/acel.13607.

[52] Ahto Salumets et al. “Graves’ disease-associated TSHR gene is demethylated and expressed in human regulatory T cells”. In: *bioRxiv* (2022).

[53] Liis Haljasmägi et al. “Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19”. In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-77525-w.

[54] Liina Tserel et al. “Long-Term Elevated Inflammatory Protein Levels in Asymptomatic SARS-CoV-2 Infected Individuals”. In: *Frontiers in Immunology* 12 (2021). DOI: 10.3389/fimmu.2021.709759.

[55] Martti Laan et al. “Post-Aire Medullary Thymic Epithelial Cells and Hassall’s Corpuscles as Inducers of Tonic Pro-Inflammatory Microenviron-

ment". In: *Frontiers in Immunology* 12 (2021). DOI: 10.3389/fimmu.2021.635569.

[56] Henk P. J. Buermans and Johan den Dunnen. "Next generation sequencing technology: Advances and applications". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842.10 (2014), pp. 1932–1941. DOI: 10.1016/j.bbadis.2014.06.015. URL: https://doi.org/10.1016/j.bbadis.2014.06.015.

[57] Peter J. A. Cock et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic Acids Research* 38.6 (2009), pp. 1767–1771. DOI: 10.1093/nar/gkp1137. URL: 10.1093/nar/gkp1137.

[58] Juliane C. Dohm et al. "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing". In: *Nucleic Acids Research* 36.16 (2008). DOI: 10.1093/nar/gkn425. URL: https://doi.org/10.1093/nar/gkn425.

[59] Babraham Bioinformatics. *FastQC*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (accessed: 01.01.2013).

[60] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (2016), pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354. URL: https://doi.org/10.1093/bioinformatics/btw354.

[61] Babraham Bioinformatics. *Trim Galore*. URL: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (accessed: 01.01.2023).

[62] Nuno A. Fonseca et al. "Tools for mapping high-throughput sequencing data". In: *Bioinformatics* 28.24 (2012), pp. 3169–3177. DOI: 10.1093/bioinformatics/bts605. URL: https://doi.org/10.1093/bioinformatics/bts605.

[63] Felix Krueger and Simon R. Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *Bioinformatics* 27.11 (2011), pp. 1571–1572. DOI: 10.1093/bioinformatics/btr167. URL: https://doi.org/110.1093/bioinformatics/btr167.

[64] Yuanyuan Li and Trygve O. Tollefsbol. "DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis". In: *Methods in Molecular Biology*. Humana Press, 2011, pp. 11–21. DOI: 10.1007/978-1-61779-316-5_2. URL: https://doi.org/10.1007/978-1-61779-316-5_2.

[65] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923. URL: https://doi.org/10.1038/nmeth.1923.

[66] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". en. In: *Genome Biology* 17.1 (2016). DOI: 10.1186/s13059-016-1066-1. URL: http://dx.doi.org/10.1186/s13059-016-1066-1.

[67] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. “SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips”. In: *Genome Biology* 13.6 (2012), R44. DOI: 10.1186/gb-2012-13-6-r44. URL: https://doi.org/10.1186/gb-2012-13-6-r44.

[68] Mike L Smith et al. “illuminaio: An open source IDAT parsing tool for Illumina microarrays”. In: *F1000Research* 2 (2013), p. 264. DOI: 10.12688/f1000research.2-264.v1. URL: https://doi.org/10.12688/f1000research.2-264.v1.

[69] Yang Xie, Xinlei Wang, and Michael Story. “Statistical methods of background correction for Illumina BeadArray data”. In: *Bioinformatics* 25.6 (2009), pp. 751–757. DOI: 10.1093/bioinformatics/btp040. URL: https://doi.org/10.1093/bioinformatics/btp040.

[70] Ricardo A. Verdugo et al. “Importance of randomization in microarray experimental designs with Illumina platforms”. In: *Nucleic Acids Research* 37.17 (2009), pp. 5610–5618. DOI: 10.1093/nar/gkp573. URL: https://doi.org/10.1093/nar/gkp573.

[71] Sarah Dedeurwaerder et al. “A comprehensive overview of Infinium HumanMethylation450 data processing”. In: *Briefings in Bioinformatics* 15.6 (2013), pp. 929–941. DOI: 10.1093/bib/bbt054. URL: https://doi.org/10.1093/bib/bbt054.

[72] Sarah Dedeurwaerder et al. “Evaluation of the Infinium Methylation 450K technology”. In: *Epigenomics* 3.6 (2011), pp. 771–784. DOI: 10.2217/epi.11.105. URL: https://doi.org/10.2217/epi.11.105.

[73] Nizar Touleimat and Jörg Tost. “Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation”. In: *Epigenomics* 4.3 (2012), pp. 325–341. DOI: 10.2217/epi.12.21. URL: https://doi.org/10.2217/epi.12.21.

[74] Martin J. Aryee et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays”. en. In: *Bioinformatics* 30.10 (2014), pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049. URL: http://dx.doi.org/10.1093/bioinformatics/btu049.

[75] Timothy J. Triche et al. “Low-level processing of Illumina Infinium DNA Methylation BeadArrays”. In: *Nucleic Acids Research* 41.7 (2013), e90–e90. DOI: 10.1093/nar/gkt090. URL: https://doi.org/10.1093/nar/gkt090.

[76] Jean-Philippe Fortin et al. “Functional normalization of 450k methylation array data improves replication in large cancer studies”. en. In: *Genome Biology* 15.11 (2014). DOI: 10.1186/s13059-014-0503-2. URL: http://dx.doi.org/10.1186/s13059-014-0503-2.

[77] Pan Du et al. “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis”. en. In: *BMC Bioinformatics* 11.1 (2010). DOI: 10.1186/1471-2105-11-587. URL: http://dx.doi.org/10.1186/1471-2105-11-587.

[78] Stefka Tyanova, Tikira Temu, and Juergen Cox. “The MaxQuant computational platform for mass spectrometry-based shotgun proteomics”. In: *Nature Protocols* 11.12 (2016), pp. 2301–2319. DOI: 10.1038/nprot.2016.136. URL: https://doi.org/10.1038/nprot.2016.136.

[79] Jürgen Cox et al. “Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ”. en. In: *Molecular & Cellular Proteomics* 13.9 (2014), pp. 2513–2526. DOI: 10.1074/mcp.m113.031591. URL: http://dx.doi.org/10.1074/mcp.M113.031591.

[80] Olink Proteomics. *PEA – A high-multiplex immunoassay technology with qPCR or NGS readout*. https://www.olink.com/content/uploads/2021/09/olink-white-paper-pea-a-high-multiplex-immunoassay-technology-with-qpcr-or-ngs-readout-v1.0.pdf.

[81] Olink Proteomics. *Data normalization and standardization*. https://www.olink.com/content/uploads/2022/04/white-paper-data-normalization-v2.1.pdf.

[82] Zulfiqar Ali and SBala Bhaskar. “Basic statistical tools in research and data analysis”. In: *Indian Journal of Anaesthesia* 60.9 (2016), p. 662. DOI: 10.4103/0019-5049.190623. URL: https://doi.org/10.4103/0019-5049.190623.

[83] Xiaoyi Gao, Joshua Starmer, and Eden R. Martin. “A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms”. In: *Genetic Epidemiology* 32.4 (2008), pp. 361–369. DOI: 10.1002/gepi.20310. URL: https://doi.org/10.1002/gepi.20310.

[84] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x. URL: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[85] Patrick Schober, Christa Boer, and Lothar A. Schwarte. “Correlation Coefficients”. In: *Anesthesia & Analgesia* 126.5 (2018), pp. 1763–1768. DOI: 10.1213/ane.0000000000002864. URL: https://doi.org/10.1213/ane.0000000000002864.

[86] Seongho Kim. “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients”. In: *Communications for Statistical Applications and Methods* 22.6 (2015), pp. 665–674. DOI: 10.5351/csam.

2015.22.6.665. URL: https://doi.org/10.5351/csam.2015.22.6.665.

[87] Zhongheng Zhang et al. “Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R”. In: *Annals of Translational Medicine* 5.4 (2017), pp. 75–75. DOI: 10.21037/atm.2017.02.05. URL: https://doi.org/10.21037/atm.2017.02.05.

[88] Ian T. Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202. URL: https://doi.org/10.1098/rsta.2015.0202.

[89] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. “A Modified Principal Component Technique Based on the LASSO”. In: *Journal of Computational and Graphical Statistics* 12.3 (2003), pp. 531–547. DOI: 10.1198/1061860032148. URL: https://doi.org/10.1198/1061860032148.

[90] Marc Hallin, Davy Paindaveine, and Thomas Verdebout. “Efficient R-Estimation of Principal and Common Principal Components”. In: *Journal of the American Statistical Association* 109.507 (2014), pp. 1071–1083. DOI: 10.1080/01621459.2014.880057. URL: https://doi.org/10.1080/01621459.2014.880057.

[91] Robert J Casson and Lachlan DM Farmer. “Understanding and checking the assumptions of linear regression: a primer for medical researchers”. In: *Clinical & Experimental Ophthalmology* 42.6 (2014), pp. 590–596. DOI: 10.1111/ceo.12358. URL: https://doi.org/10.1111/ceo.12358.

[92] Peter Flach. In: *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012. Chap. 7, pp. 194–230.

[93] Peter Flach. In: *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012. Chap. 5, pp. 129–156.

[94] D. J. Stekhoven and P. Buhlmann. “MissForest–non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1 (2011), pp. 112–118. DOI: 10.1093/bioinformatics/btr597. URL: https://doi.org/10.1093/bioinformatics/btr597.

[95] Miron B. Kursa and Witold R. Rudnicki. “Feature Selection with the Boruta Package”. In: *Journal of Statistical Software* 36.11 (2010). DOI: 10.18637/jss.v036.i11. URL: https://doi.org/10.18637/jss.v036.i11.

[96] Andy Liaw and Matthew Wiener. “Classification and Regression by randomForest”. In: *R News* 2.3 (2002), pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

[97] Jeff E. Mold et al. “Cell generation dynamics underlying naive T-cell homeostasis in adult humans”. en. In: *PLOS Biology* 17.10 (2019). Ed. by Avinash Bhandoola, e3000383. DOI: 10.1371/journal.pbio.3000383. URL: http://dx.doi.org/10.1371/journal.pbio.3000383.

[98] Chan C. Whiting et al. “Large-Scale and Comprehensive Immune Profiling and Functional Analysis of Normal Human Aging”. en. In: *PLOS ONE* 10.7 (2015). Ed. by Derya Unutmaz, e0133627. DOI: 10.1371/journal.pone.0133627. URL: http://dx.doi.org/10.1371/journal.pone.0133627.

[99] Sian M Henson, Natalie E Riddell, and Arne N Akbar. “Properties of end-stage human T cells defined by CD45RA re-expression”. en. In: *Current Opinion in Immunology* 24.4 (2012), pp. 476–481. DOI: 10.1016/j.coi.2012.04.001. URL: http://dx.doi.org/10.1016/j.coi.2012.04.001.

[100] Luca Pangrazzi et al. “Increased IL-15 Production and Accumulation of Highly Differentiated CD8+ Effector/Memory T Cells in the Bone Marrow of Persons with Cytomegalovirus”. In: *Frontiers in Immunology* 8 (2017). DOI: 10.3389/fimmu.2017.00715. URL: http://dx.doi.org/10.3389/fimmu.2017.00715.

[101] Paul Klenerman. “The (gradual) rise of memory inflation”. en. In: *Immunological Reviews* 283.1 (2018), pp. 99–112. DOI: 10.1111/imr.12653. URL: http://dx.doi.org/10.1111/imr.12653.

[102] Stefan Brunner et al. “Persistent viral infections and immune aging”. en. In: *Ageing Research Reviews* 10.3 (2011), pp. 362–369. DOI: 10.1016/j.arr.2010.08.003. URL: http://dx.doi.org/10.1016/j.arr.2010.08.003.

[103] Ioakim Spyridopoulos et al. “CMV seropositivity and T-cell senescence predict increased cardiovascular mortality in octogenarians: results from the Newcastle 85+ study”. en. In: *Aging Cell* 15.2 (2015), pp. 389–392. DOI: 10.1111/acel.12430. URL: http://dx.doi.org/10.1111/acel.12430.

[104] Yen-Ling Chiu et al. “A comprehensive characterization of aggravated aging-related changes in T lymphocytes and monocytes in end-stage renal disease: the iESRD study”. en. In: *Immunity & Ageing* 15.1 (2018). DOI: 10.1186/s12979-018-0131-x. URL: http://dx.doi.org/10.1186/s12979-018-0131-x.

[105] Marco Diani et al. “Increased frequency of activated CD8+ T cell effectors in patients with psoriatic arthritis”. en. In: *Scientific Reports* 9.1 (2019).

DOI: 10.1038/s41598-019-47310-5. URL: http://dx.doi.org/10.1038/s41598-019-47310-5.

[106] Yen-Ling Chiu et al. “Emergence of T cell immunosenescence in diabetic chronic kidney disease”. en. In: *Immunity & Ageing* 17.1 (2020). DOI: 10.1186/s12979-020-00200-1. URL: http://dx.doi.org/10.1186/s12979-020-00200-1.

[107] Lola Jacquemont et al. “Terminally Differentiated Effector Memory CD8+ T Cells Identify Kidney Transplant Recipients at High Risk of Graft Failure”. en. In: *Journal of the American Society of Nephrology* 31.4 (2020), pp. 876–891. DOI: 10.1681/asn.2019080847. URL: http://dx.doi.org/10.1681/ASN.2019080847.

[108] Tim K. Boßlau et al. “Abdominal Obesity-Related Disturbance of Insulin Sensitivity Is Associated with CD8+ EMRA Cells in the Elderly”. en. In: *Cells* 10.5 (2021), p. 998. DOI: 10.3390/cells10050998. URL: http://dx.doi.org/10.3390/cells10050998.

[109] Carmen Martin-Ruiz et al. “CMV-independent increase in CD27-CD28+ CD8+ EMRA T cells is inversely related to mortality in octogenarians”. en. In: *npj Aging and Mechanisms of Disease* 6.1 (2020). DOI: 10.1038/s41514-019-0041-y. URL: http://dx.doi.org/10.1038/s41514-019-0041-y.

[110] Evelyna Derhovanessian et al. “Infection with cytomegalovirus but not herpes simplex virus induces the accumulation of late-differentiated CD4+ and CD8+ T-cells in humans”. en. In: *Journal of General Virology* 92.12 (2011), pp. 2746–2756. DOI: 10.1099/vir.0.036004-0. URL: http://dx.doi.org/10.1099/vir.0.036004-0.

[111] Michael J. Cannon, D. Scott Schmid, and Terri B. Hyde. “Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection”. In: *Reviews in Medical Virology* 20.4 (2010), pp. 202–213. DOI: 10.1002/rmv.655. URL: https://doi.org/10.1002/rmv.655.

[112] Geraldine A. O’Hara et al. “Memory T cell inflation: understanding cause and effect”. In: *Trends in Immunology* 33.2 (2012), pp. 84–90. DOI: 10.1016/j.it.2011.11.005. URL: https://doi.org/10.1016/j.it.2011.11.005.

[113] Christian U. Blank et al. “Defining "T cell exhaustion"”. en. In: *Nature Reviews Immunology* 19.11 (2019), pp. 665–674. DOI: 10.1038/s41577-019-0221-9. URL: http://dx.doi.org/10.1038/s41577-019-0221-9.

[114] Jörg J. Goronzy and Cornelia M. Weyand. “Mechanisms underlying T cell ageing”. en. In: *Nature Reviews Immunology* 19.9 (2019), pp. 573–583. DOI: 10.1038/s41577-019-0180-1. URL: http://dx.doi.org/10.1038/s41577-019-0180-1.

[115] Sara P. H. van den Berg et al. “The hallmarks of CMV-specific CD8 T-cell differentiation”. In: *Medical Microbiology and Immunology* 208.3-4 (Apr. 2019), pp. 365–373. DOI: 10.1007/s00430-019-00608-7. URL: https://doi.org/10.1007/s00430-019-00608-7.

[116] Tamas Fulop et al. “he integration of inflammaging in age-related diseases”. en. In: *Journal of General Virology* 40.12 (2018), pp. 17–35. DOI: 10.1016/j.smim.2018.09.003.

[117] John Paul Tomtishen III. “Human cytomegalovirus tegument proteins (pp65, pp71, pp150, pp28)”. en. In: *Virology Journal* 9.1 (2012). DOI: 10.1186/1743-422x-9-22. URL: http://dx.doi.org/10.1186/1743-422X-9-22.

[118] Reiko Hanada et al. “RANKL/RANK—beyond bones”. en. In: *Journal of Molecular Medicine* 89.7 (2011), pp. 647–656. DOI: 10.1007/s00109-011-0749-z. URL: http://dx.doi.org/10.1007/s00109-011-0749-z.

[119] Kenta Maruyama et al. “Receptor Activator of NF-κB Ligand and Osteoprotegerin Regulate Proinflammatory Cytokine Production in Mice”. en. In: *The Journal of Immunology* 177.6 (2006), pp. 3799–3805. DOI: 10.4049/jimmunol.177.6.3799. URL: http://dx.doi.org/10.4049/jimmunol.177.6.3799.

[120] Simona W. Rossi et al. “RANK signals from CD4+3- inducer cells regulate development of Aire-expressing epithelial cells in the thymic medulla”. en. In: *Journal of Experimental Medicine* 204.6 (2007), pp. 1267–1272. DOI: 10.1084/jem.20062497. URL: http://dx.doi.org/10.1084/jem.20062497.

[121] Liina Tserel et al. “Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes”. en. In: *Scientific Reports* 5.1 (2015). DOI: 10.1038/srep13107. URL: http://dx.doi.org/10.1038/srep13107.

[122] D. J. Stekhoven and P. Buhlmann. “MissForest–non-parametric missing value imputation for mixed-type data”. en. In: *Bioinformatics* 28.1 (2011), pp. 112–118. DOI: 10.1093/bioinformatics/btr597. URL: http://dx.doi.org/10.1093/bioinformatics/btr597.

[123] Miron B. Kursa and Witold R. Rudnicki. “Feature Selection with the-BorutaPackage”. en. In: *Journal of Statistical Software* 36.11 (2010). DOI: 10.18637/jss.v036.i11. URL: http://dx.doi.org/10.18637/jss.v036.i11.

[124] Wei Qian et al. “Association between TSHR gene polymorphism and the risk of Graves disease a meta-analysis”. en. In: *The Journal of Biomedical Research* 30.6 (2016), pp. 466–475. DOI: 10.7555/JBR.30.20140144. URL: http://www.jbr-pub.org.cn//article/id/jbr160603.

[125] Clare Baecher-Allan and David A. Hafler. “Human regulatory T cells and their role in autoimmune disease”. en. In: *Immunological Reviews* 212.1 (2006), pp. 203–216. DOI: 10.1111/j.0105-2896.2006.00417.x. URL: http://dx.doi.org/10.1111/j.0105-2896.2006.00417.x.

[126] Giovanni Antonio Maria Povoleri et al. “Thymic Versus Induced Regulatory T Cells – Who Regulates the Regulators?” In: *Frontiers in Immunology* 4 (2013). DOI: 10.3389/fimmu.2013.00169. URL: http://dx.doi.org/10.3389/fimmu.2013.00169.

[127] Daniil Shevyrev and Valeriy Tereshchenko. “Treg Heterogeneity, Function, and Homeostasis”. In: *Frontiers in Immunology* 10 (2020). DOI: 10.3389/fimmu.2019.03100. URL: http://dx.doi.org/10.3389/fimmu.2019.03100.

[128] Shimon Sakaguchi et al. “Regulatory T Cells and Immune Tolerance”. en. In: *Cell* 133.5 (2008), pp. 775–787. DOI: 10.1016/j.cell.2008.05.009. URL: http://dx.doi.org/10.1016/j.cell.2008.05.009.

[129] Kajsa Wing et al. “CTLA-4 Control over Foxp3 + Regulatory T Cell Function”. en. In: *Science* 322.5899 (2008), pp. 271–275. DOI: 10.1126/science.1160062. URL: http://dx.doi.org/10.1126/science.1160062.

[130] Alice McNally et al. “CD4 + CD25 + regulatory T cells control CD8 + T-cell effector differentiation by modulating IL-2 homeostasis”. en. In: *Proceedings of the National Academy of Sciences* 108.18 (2011), pp. 7529–7534. DOI: 10.1073/pnas.1103782108. URL: http://dx.doi.org/10.1073/pnas.1103782108.

[131] Yasushi Onishi et al. “Foxp3 + natural regulatory T cells preferentially form aggregates on dendritic cells in vitro and actively inhibit their maturation”. en. In: *Proceedings of the National Academy of Sciences* 105.29 (2008), pp. 10113–10118. DOI: 10.1073/pnas.0711106105. URL: http://dx.doi.org/10.1073/pnas.0711106105.

[132] Ursula Grohmann et al. “CTLA-4–Ig regulates tryptophan catabolism in vivo”. en. In: *Nature Immunology* 3.11 (2002), pp. 1097–1101. DOI: 10.1038/ni846. URL: http://dx.doi.org/10.1038/ni846.

[133] Takatoshi Chinen et al. “An essential role for the IL-2 receptor in Treg cell function”. en. In: *Nature Immunology* 17.11 (2016), pp. 1322–1333. DOI: 10.1038/ni.3540. URL: http://dx.doi.org/10.1038/ni.3540.

[134] Lauren W. Collison et al. “The inhibitory cytokine IL-35 contributes to regulatory T-cell function”. en. In: *Nature* 450.7169 (2007), pp. 566–569. DOI: 10.1038/nature06306. URL: http://dx.doi.org/10.1038/nature06306.

[135] Yisong Y. Wan and Richard A. Flavell. “TGF-$\beta$ and Regulatory T Cell in Immunity and Autoimmunity”. en. In: *Journal of Clinical Immunology*

28.6 (2008), pp. 647–659. DOI: 10.1007/s10875-008-9251-y. URL: http://dx.doi.org/10.1007/s10875-008-9251-y.

[136] Keishi Fujio, Kazuhiko Yamamoto, and Tomohisa Okamura. *Overview of LAG-3-Expressing, IL-10-Producing Regulatory T Cells*. 2017. DOI: 10.1007/82_2017_59. URL: http://dx.doi.org/10.1007/82_2017_59.

[137] Youhai Chen et al. "Regulatory T Cell Clones Induced by Oral Tolerance: Suppression of Autoimmune Encephalomyelitis". en. In: *Science* 265.5176 (1994), pp. 1237–1240. DOI: 10.1126/science.7520605. URL: http://dx.doi.org/10.1126/science.7520605.

[138] Omid Akbari et al. "Antigen-specific regulatory T cells develop via the ICOS–ICOS-ligand pathway and inhibit allergen-induced airway hyperreactivity". en. In: *Nature Medicine* 8.9 (2002), pp. 1024–1032. DOI: 10.1038/nm745. URL: http://dx.doi.org/10.1038/nm745.

[139] R. Wang et al. "Mechanisms of Regulatory T-cell Induction by Antigen-IgG-transduced Splenocytes". en. In: *Scandinavian Journal of Immunology* 66.5 (2007), pp. 515–522. DOI: 10.1111/j.1365-3083.2007.02004.x. URL: http://dx.doi.org/10.1111/j.1365-3083.2007.02004.x.

[140] Craig L. Bennett et al. "The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3". en. In: *Nature Genetics* 27.1 (2001), pp. 20–21. DOI: 10.1038/83713. URL: http://dx.doi.org/10.1038/83713.

[141] Ruka Setoguchi et al. "Homeostatic maintenance of natural Foxp3+ CD25+ CD4+ regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization". en. In: *Journal of Experimental Medicine* 201.5 (2005), pp. 723–735. DOI: 10.1084/jem.20041982. URL: http://dx.doi.org/10.1084/jem.20041982.

[142] Dat Q. Tran et al. "Analysis of Adhesion Molecules, Target Cells, and Role of IL-2 in Human FOXP3+Regulatory T Cell Suppressor Function". en. In: *The Journal of Immunology* 182.5 (2009), pp. 2929–2938. DOI: 10.4049/jimmunol.0803827. URL: http://dx.doi.org/10.4049/jimmunol.0803827.

[143] Dorothy K. Sojka, Angela Hughson, and Deborah J. Fowell. "CTLA-4 is required by CD4+CD25+ Treg to control CD4+ T-cell lymphopenia-induced proliferation". en. In: *European Journal of Immunology* 39.6 (2009), pp. 1544–1551. DOI: 10.1002/eji.200838603. URL: http://dx.doi.org/10.1002/eji.200838603.

[144] Shimon Sakaguchi. "Naturally Arising CD4+ Regulatory T Cells for Immunologic Self-Tolerance and Negative Control of Immune Responses". en. In: *Annual Review of Immunology* 22.1 (2004), pp. 531–562. DOI: 10.1146/annurev.immunol.21.120601.141122. URL: http://dx.doi.org/10.1146/annurev.immunol.21.120601.141122.

[145] Alexander Marson et al. “Foxp3 occupancy and regulation of key target genes during T-cell stimulation”. en. In: *Nature* 445.7130 (2007), pp. 931–935. DOI: 10.1038/nature05478. URL: http://dx.doi.org/10.1038/nature05478.

[146] Ye Zheng et al. “Genome-wide analysis of Foxp3 target genes in developing and mature regulatory T cells”. en. In: *Nature* 445.7130 (2007), pp. 936–940. DOI: 10.1038/nature05563. URL: http://dx.doi.org/10.1038/nature05563.

[147] Marc A. Gavin et al. “Foxp3-dependent programme of regulatory T-cell differentiation”. en. In: *Nature* 445.7129 (2007), pp. 771–775. DOI: 10.1038/nature05543. URL: http://dx.doi.org/10.1038/nature05543.

[148] Wen Lin et al. “Regulatory T cell development in the absence of functional Foxp3”. en. In: *Nature Immunology* 8.4 (2007), pp. 359–368. DOI: 10.1038/ni1445. URL: http://dx.doi.org/10.1038/ni1445.

[149] Julia K. Polansky et al. “DNA methylation controls Foxp3 gene expression”. en. In: *European Journal of Immunology* 38.6 (2008), pp. 1654–1663. DOI: 10.1002/eji.200838105. URL: http://dx.doi.org/10.1002/eji.200838105.

[150] Stefan Floess et al. “Epigenetic Control of the foxp3 Locus in Regulatory T Cells”. en. In: *PLoS Biology* 5.2 (2007). Ed. by Philippa Marrack, e38. DOI: 10.1371/journal.pbio.0050038. URL: http://dx.doi.org/10.1371/journal.pbio.0050038.

[151] Steven Z. Josefowicz, Li-Fan Lu, and Alexander Y. Rudensky. “Regulatory T Cells: Mechanisms of Differentiation and Function”. en. In: *Annual Review of Immunology* 30.1 (2012), pp. 531–564. DOI: 10.1146/annurev.immunol.25.022106.141623. URL: http://dx.doi.org/10.1146/annurev.immunol.25.022106.141623.

[152] Yuxia Zhang et al. “Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells”. en. In: *Blood* 122.16 (2013), pp. 2823–2836. DOI: 10.1182/blood-2013-02-481788. URL: http://dx.doi.org/10.1182/blood-2013-02-481788.

[153] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7 (2015), e47–e47. DOI: 10.1093/nar/gkv007. URL: https://doi.org/10.1093/nar/gkv007.

[154] Timothy J Peters et al. “De novo identification of differentially methylated regions in the human genome”. In: *Epigenetics & Chromatin* 8.1 (2015). DOI: 10.1186/1756-8935-8-6. URL: https://doi.org/10.1186/1756-8935-8-6.

[155] Bethany A. Buck-Koehntop et al. “Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso”.

en. In: *Proceedings of the National Academy of Sciences* 109.38 (2012), pp. 15229–15234. DOI: 10.1073/pnas.1213726109. URL: http://dx.doi.org/10.1073/pnas.1213726109.

[156] Darya Kaplun et al. "Kaiso Regulates DNA Methylation Homeostasis". en. In: *International Journal of Molecular Sciences* 22.14 (2021), p. 7587. DOI: 10.3390/ijms22147587. URL: http://dx.doi.org/10.3390/ijms22147587.

[157] Christian Schmidl et al. "The enhancer and promoter landscape of human regulatory and conventional T-cell subpopulations". en. In: *Blood* 123.17 (2014), e68–e78. DOI: 10.1182/blood-2013-02-486944. URL: http://dx.doi.org/10.1182/blood-2013-02-486944.

[158] Naganari Ohkura et al. "Regulatory T Cell-Specific Epigenomic Region Variants Are a Key Determinant of Susceptibility to Common Autoimmune Diseases". en. In: *Immunity* 52.6 (2020), 1119–1132.e4. DOI: 10.1016/j.immuni.2020.04.006. URL: http://dx.doi.org/10.1016/j.immuni.2020.04.006.

[159] Shimon Sakaguchi et al. "Regulatory T Cells and Human Disease". en. In: *Annual Review of Immunology* 38.1 (2020), pp. 541–566. DOI: 10.1146/annurev-immunol-042718-041717. URL: http://dx.doi.org/10.1146/annurev-immunol-042718-041717.

[160] Naganari Ohkura et al. "T Cell Receptor Stimulation-Induced Epigenetic Changes and Foxp3 Expression Are Independent and Complementary Events Required for Treg Cell Development". en. In: *Immunity* 37.5 (2012), pp. 785–799. DOI: 10.1016/j.immuni.2012.09.010. URL: http://dx.doi.org/10.1016/j.immuni.2012.09.010.

[161] Ling Lu, Joseph Barbi, and Fan Pan. "The regulation of immune tolerance by FOXP3". en. In: *Nature Reviews Immunology* 17.11 (2017), pp. 703–717. DOI: 10.1038/nri.2017.75. URL: http://dx.doi.org/10.1038/nri.2017.75.

[162] Marianne Chabod et al. "A Spontaneous Mutation of the Rat Themis Gene Leads to Impaired Function of Regulatory T Cells Linked to Inflammatory Bowel Disease". en. In: *PLoS Genetics* 8.1 (2012). Ed. by Derry C. Roopenian, e1002461. DOI: 10.1371/journal.pgen.1002461. URL: http://dx.doi.org/10.1371/journal.pgen.1002461.

[163] Terry F. Davies et al. "Graves' disease". en. In: *Nature Reviews Disease Primers* 6.1 (2020). DOI: 10.1038/s41572-020-0184-y. URL: http://dx.doi.org/10.1038/s41572-020-0184-y.

[164] Mónica Marazuela et al. "Regulatory T Cells in Human Autoimmune Thyroid Disease". en. In: *The Journal of Clinical Endocrinology & Metabolism* 91.9 (2006), pp. 3639–3646. DOI: 10.1210/jc.2005-2337. URL: http://dx.doi.org/10.1210/jc.2005-2337.

[165] Chaoming Mao et al. “Impairment of Regulatory Capacity of CD4+CD25+ Regulatory T Cells Mediated by Dendritic Cell Polarization and Hyperthyroidism in Graves’ Disease”. en. In: *The Journal of Immunology* 186.8 (2011), pp. 4734–4743. DOI: 10.4049/jimmunol.0904135. URL: http://dx.doi.org/10.4049/jimmunol.0904135.

[166] Ziyi Chen et al. “Decreased Treg Cell and TCR Expansion Are Involved in Long-Lasting Graves’ Disease”. In: *Frontiers in Endocrinology* 12 (2021). DOI: 10.3389/fendo.2021.632492. URL: http://dx.doi.org/10.3389/fendo.2021.632492.

[167] Bin Wang et al. “CEP128 is a crucial risk locus for autoimmune thyroid diseases”. en. In: *Molecular and Cellular Endocrinology* 480 (2019), pp. 97–106. DOI: 10.1016/j.mce.2018.10.017. URL: http://dx.doi.org/10.1016/j.mce.2018.10.017.

[168] Maia Limbach et al. “Epigenetic profiling in CD4+ and CD8+ T cells from Graves’ disease patients reveals changes in genes associated with T cell receptor signaling”. en. In: *Journal of Autoimmunity* 67 (2016), pp. 46–56. DOI: 10.1016/j.jaut.2015.09.006. URL: http://dx.doi.org/10.1016/j.jaut.2015.09.006.

[169] O. J. Brand et al. “Association of the thyroid stimulating hormone receptor gene (TSHR) with Graves’ disease”. en. In: *Human Molecular Genetics* 18.9 (2009), pp. 1704–1713. DOI: 10.1093/hmg/ddp087. URL: http://dx.doi.org/10.1093/hmg/ddp087.

[170] Roger Colobran et al. “Association of an SNP with intrathymic transcription of TSHR and Graves’ disease: a role for defective thymic tolerance”. en. In: *Human Molecular Genetics* 20.17 (2011), pp. 3415–3423. DOI: 10.1093/hmg/ddr247. URL: http://dx.doi.org/10.1093/hmg/ddr247.

[171] Mihaela Stefan et al. “Genetic–epigenetic dysregulation of thymic TSH receptor gene expression triggers thyroid autoimmunity”. en. In: *Proceedings of the National Academy of Sciences* 111.34 (2014), pp. 12562–12567. DOI: 10.1073/pnas.1408821111. URL: http://dx.doi.org/10.1073/pnas.1408821111.

[172] Ricardo Pujol-Borrell et al. “Central Tolerance Mechanisms to TSHR in Graves’ Disease: Contributions to Understand the Genetic Association”. en. In: *Hormone and Metabolic Research* 50.12 (2018), pp. 863–870. DOI: 10.1055/a-0755-7927. URL: http://dx.doi.org/10.1055/a-0755-7927.

[173] Jean S. Marshall et al. “An introduction to immunology and immunopathology”. en. In: *Allergy, Asthma & Clinical Immunology* 14.S2 (2018). DOI: 10.1186/s13223-018-0278-1. URL: http://dx.doi.org/10.1186/s13223-018-0278-1.

[174] Michael S Krangel. “Mechanics of T cell receptor gene rearrangement”. en. In: *Current Opinion in Immunology* 21.2 (2009), pp. 133–139. DOI: 10.1016/j.coi.2009.03.009. URL: http://dx.doi.org/10.1016/j.coi.2009.03.009.

[175] Kristin A. Hogquist, Troy A. Baldwin, and Stephen C. Jameson. “Central tolerance: learning self-control in the thymus”. en. In: *Nature Reviews Immunology* 5.10 (2005), pp. 772–782. DOI: 10.1038/nri1707. URL: http://dx.doi.org/10.1038/nri1707.

[176] Jens Derbinski et al. “Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self”. en. In: *Nature Immunology* 2.11 (2001), pp. 1032–1039. DOI: 10.1038/ni723. URL: http://dx.doi.org/10.1038/ni723.

[177] Noam Kadouri et al. “Thymic epithelial cell heterogeneity: TEC by TEC”. en. In: *Nature Reviews Immunology* 20.4 (2019), pp. 239–253. DOI: 10.1038/s41577-019-0238-0. URL: http://dx.doi.org/10.1038/s41577-019-0238-0.

[178] Kristen L. Wells et al. “Combined transient ablation and single cell RNA sequencing reveals the development of medullary thymic epithelial cells”. In: (2020). DOI: 10.1101/2020.06.19.160424. URL: https://doi.org/10.1101/2020.06.19.160424.

[179] Jong-Eun Park et al. “A cell atlas of human thymic development defines T cell repertoire formation”. en. In: *Science* 367.6480 (2020). DOI: 10.1126/science.aay3224. URL: http://dx.doi.org/10.1126/science.aay3224.

[180] Stephen N. Sansom et al. “Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia”. en. In: *Genome Research* 24.12 (2014), pp. 1918–1931. DOI: 10.1101/gr.171645.113. URL: http://dx.doi.org/10.1101/gr.171645.113.

[181] Ann E. Walts et al. “Involucrin, a marker of squamous and urothelial differentiation. An immunohistochemical study on its distribution in normal and neoplastic tissues”. en. In: *The Journal of Pathology* 145.4 (1985), pp. 329–340. DOI: 10.1002/path.1711450406. URL: http://dx.doi.org/10.1002/path.1711450406.

[182] E. Bitoun. “LEKTI proteolytic processing in human primary keratinocytes, tissue distribution and defective expression in Netherton syndrome”. en. In: *Human Molecular Genetics* 12.19 (2003), pp. 2417–2430. DOI: 10.1093/hmg/ddg247. URL: http://dx.doi.org/10.1093/hmg/ddg247.

[183] Adil Asghar, MSyed Yunus, and ANafis Faruqi. “Polymorphism of Hassall’s corpuscles in thymus of human fetuses”. en. In: *International Journal of Applied and Basic Medical Research* 2.1 (2012), p. 7. DOI: 10.

4103/2229-516x.96791. URL: http://dx.doi.org/10.4103/2229-516x.96791.

[184] Masashi Yano et al. “Aire controls the differentiation program of thymic epithelial cells in the medulla for the establishment of self-tolerance”. en. In: *Journal of Experimental Medicine* 205.12 (2008), pp. 2827–2838. DOI: 10.1084/jem.20080046. URL: http://dx.doi.org/10.1084/jem.20080046.

[185] Xiaoping Wang et al. “Post-Aire Maturation of Thymic Medullary Epithelial Cells Involves Selective Expression of Keratinocyte-Specific Autoantigens”. In: *Frontiers in Immunology* 3 (2012). DOI: 10.3389/fimmu.2012.00019. URL: http://dx.doi.org/10.3389/fimmu.2012.00019.

[186] Arnaud D. Colantonio et al. “IFN-$\alpha$ Is Constitutively Expressed in the Human Thymus, but Not in Peripheral Lymphoid Organs”. en. In: *PLoS ONE* 6.8 (2011). Ed. by Derya Unutmaz, e24252. DOI: 10.1371/journal.pone.0024252. URL: http://dx.doi.org/10.1371/journal.pone.0024252.

[187] Anthony Meager et al. “Anti-Interferon Autoantibodies in Autoimmune Polyendocrinopathy Syndrome Type 1”. In: *PLoS Medicine* 3.7 (2006). Ed. by Ludvig Sollid, e289. DOI: 10.1371/journal.pmed.0030289. URL: https://doi.org/10.1371/journal.pmed.0030289.

[188] Stefan Lienenklaus et al. “Novel Reporter Mouse Reveals Constitutive and Inflammatory Expression of IFN-$\beta$ In Vivo”. In: *The Journal of Immunology* 183.5 (2009), pp. 3229–3236. DOI: 10.4049/jimmunol.0804277. URL: https://doi.org/10.4049/jimmunol.0804277.

[189] Jianwei Wang et al. “Hassall’s corpuscles with cellular-senescence features maintain IFN$\alpha$ production through neutrophils and pDC activation in the thymus”. In: *International Immunology* 31.3 (2018), pp. 127–139. DOI: 10.1093/intimm/dxy073. URL: https://doi.org/10.1093/intimm/dxy073.

[190] Matouš Vobořil et al. “Toll-like receptor signaling in thymic epithelium controls monocyte-derived dendritic cell recruitment and Treg generation”. In: *Nature Communications* 11.1 (2020). DOI: 10.1038/s41467-020-16081-3. URL: https://doi.org/10.1038/s41467-020-16081-3.

[191] 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.W1 (2019), W191–W198. DOI: 10.1093/nar/gkz369. URL: https://doi.org/10.1093/nar/gkz369.

[192] Sakuhei Fujiwara et al. “Epiplakin, a Novel Member of the Plakin Family Originally Identified as a 450-kDa Human Epidermal Autoantigen”. In: *Journal of Biological Chemistry* 276.16 (2001), pp. 13340–13347. DOI:

10.1074/jbc.m011386200. URL: https://doi.org/10.1074/jbc.m011386200.

[193] Isabelle Schepens et al. "The Protease Inhibitor Alpha-2-Macroglobuline-Like-1 Is the p170 Antigen Recognized by Paraneoplastic Pemphigus Autoantibodies in Human". In: *PLoS ONE* 5.8 (2010). Ed. by H. Peter Soyer, e12250. DOI: 10.1371/journal.pone.0012250. URL: https://doi.org/10.1371/journal.pone.0012250.

[194] Laura L. Gonzalez, Karin Garrie, and Mark D. Turner. "Role of S100 proteins in health and disease". In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1867.6 (2020), p. 118677. DOI: 10.1016/j.bbamcr.2020.118677. URL: https://doi.org/10.1016/j.bbamcr.2020.118677.

[195] Siwen Wang et al. "S100A8/A9 in Inflammation". In: *Frontiers in Immunology* 9 (2018). DOI: 10.3389/fimmu.2018.01298. URL: https://doi.org/10.3389/fimmu.2018.01298.

[196] Shizuo Akira and Kiyoshi Takeda. "Toll-like receptor signalling". In: *Nature Reviews Immunology* 4.7 (2004), pp. 499–511. DOI: 10.1038/nri1391. URL: https://doi.org/10.1038/nri1391.

[197] George D Kalliolias and Lionel B Ivashkiv. "Overview of the biology of type I interferons". In: *Arthritis Research & Therapy* 12.Suppl 1 (2010), S1. DOI: 10.1186/ar2881. URL: https://doi.org/10.1186/ar2881.

[198] Hiroshi Moro et al. "T Cell-Intrinsic and -Extrinsic Contributions of the IFNAR/STAT1-Axis to Thymocyte Survival". In: *PLoS ONE* 6.9 (2011). Ed. by Jose Alberola-Ila, e24972. DOI: 10.1371/journal.pone.0024972. URL: https://doi.org/10.1371/journal.pone.0024972.

[199] Amina Metidji et al. "IFN-$\alpha/\beta$ Receptor Signaling Promotes Regulatory T Cell Development and Function under Stress Conditions". In: *The Journal of Immunology* 194.9 (2015), pp. 4265–4276. DOI: 10.4049/jimmunol.1500036. URL: https://doi.org/10.4049/jimmunol.1500036.

[200] Marco Cascella et al. "Features, Evaluation, and Treatment of Coronavirus (COVID-19) [Updated 2022 Jun 30]". In: (2022). URL: https://www.ncbi.nlm.nih.gov/books/NBK554776/.

[201] Chongzhi Bai, Qiming Zhong, and George Fu Gao. "Overview of SARS-CoV-2 genome-encoded proteins". In: *Science China Life Sciences* 65.2 (2021), pp. 280–294. DOI: 10.1007/s11427-021-1964-4. URL: https://doi.org/10.1007/s11427-021-1964-4.

[202] Chung-ke Chang et al. "Recent insights into the development of therapeutics against coronavirus diseases by targeting N protein". In: *Drug Discovery Today* 21.4 (2016), pp. 562–572. DOI: 10.1016/j.drudis.2015.11.015. URL: https://doi.org/10.1016/j.drudis.2015.11.015.

[203] Benjamin W. Neuman et al. "A structural analysis of M protein in coronavirus assembly and morphology". In: *Journal of Structural Biology* 174.1

(2011), pp. 11–22. DOI: 10.1016/j.jsb.2010.11.021. URL: https://doi.org/10.1016/j.jsb.2010.11.021.

[204] Yipeng Cao et al. “Characterization of the SARS-CoV-2 E Protein: Sequence, Structure, Viroporin, and Inhibitors”. In: *Protein Science* 30.6 (2021), pp. 1114–1130. DOI: 10.1002/pro.4075. URL: https://doi.org/10.1002/pro.4075.

[205] Cody B. Jackson et al. “Mechanisms of SARS-CoV-2 entry into cells”. In: *Nature Reviews Molecular Cell Biology* 23.1 (2021), pp. 3–20. DOI: 10.1038/s41580-021-00418-x. URL: https://doi.org/10.1038/s41580-021-00418-x.

[206] Ella Hartenian et al. “The molecular virology of coronaviruses”. In: *Journal of Biological Chemistry* 295.37 (2020), pp. 12910–12934. DOI: 10.1074/jbc.rev120.013930. URL: https://doi.org/10.1074/jbc.rev120.013930.

[207] Brandon Malone et al. “Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design”. In: *Nature Reviews Molecular Cell Biology* 23.1 (2021), pp. 21–39. DOI: 10.1038/s41580-021-00432-z. URL: https://doi.org/10.1038/s41580-021-00432-z.

[208] Kenji Mizumoto et al. “Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020”. In: *Eurosurveillance* 25.10 (2020). DOI: 10.2807/1560-7917.es.2020.25.10.2000180. URL: https://doi.org/10.2807/1560-7917.es.2020.25.10.2000180.

[209] Hiroshi Nishiura et al. “Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19)”. In: *International Journal of Infectious Diseases* 94 (2020), pp. 154–155. DOI: 10.1016/j.ijid.2020.03.020. URL: https://doi.org/10.1016/j.ijid.2020.03.020.

[210] Jin Wang et al. “Cytokine storm and leukocyte changes in mild versus severe SARS-CoV-2 infection: Review of 3939 COVID-19 patients in China and emerging pathogenesis and therapy concepts”. In: *Journal of Leukocyte Biology* 108.1 (2020), pp. 17–41. DOI: 10.1002/jlb.3covr0520-272r. URL: https://doi.org/10.1002/jlb.3covr0520-272r.

[211] Ahmet Kursat Azkur et al. “Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19”. In: *Allergy* 75.7 (2020), pp. 1564–1581. DOI: 10.1111/all.14364. URL: https://doi.org/10.1111/all.14364.

[212] Puja Mehta and David C Fajgenbaum. “Is severe COVID-19 a cytokine storm syndrome: a hyperinflammatory debate”. en. In: *Curr. Opin. Rheumatol.* 33.5 (2021), pp. 419–430.

[213] Erin K Stokes et al. “Coronavirus disease 2019 case surveillance - United States, January 22-May 30, 2020”. en. In: *MMWR Morb. Mortal. Wkly. Rep.* 69.24 (2020), pp. 759–765.

[214] Adam G. Laing et al. “A dynamic COVID-19 immune signature includes associations with poor prognosis”. In: *Nature Medicine* 26.10 (2020), pp. 1623–1635. DOI: 10.1038/s41591-020-1038-6. URL: https://doi.org/10.1038/s41591-020-1038-6.

[215] Linnan Zhu et al. “Single-Cell Sequencing of Peripheral Mononuclear Cells Reveals Distinct Immune Response Landscapes of COVID-19 and Influenza Patients”. In: *Immunity* 53.3 (2020), 685–696.e3. DOI: 10.1016/j.immuni.2020.07.009. URL: https://doi.org/10.1016/j.immuni.2020.07.009.

[216] Guang-Hui Xiao et al. “Anti-apoptotic signaling by hepatocyte growth factor/Met via the phosphatidylinositol 3-kinase/Akt and mitogen-activated protein kinase pathways”. In: *Proceedings of the National Academy of Sciences* 98.1 (2000), pp. 247–252. DOI: 10.1073/pnas.98.1.247. URL: https://doi.org/10.1073/pnas.98.1.247.

[217] *Dimensions*. URL: https://app.dimensions.ai. (accessed: 01.01.2023).

# ACKNOWLEDGEMENTS

I would like to thank all those who have been part of my PhD journey. Your support, encouragement, and friendship have made this journey so much more enjoyable and memorable.

First and foremost, I would like to give a special thanks to my supervisor Hedi. I am very grateful for the opportunity to join the BIIT Research group, as it has allowed me to convert myself from a biologist to a bioinformatician and it has opened the door to the data science for me. Additionally, your kindness and generosity have created a very positive working environment for all of us. Thank you for providing good working conditions and encouraging us to travel.

I would also like thank to my co-supervisor Pärt. Your help with writing scientific papers and our discussions on immunology, science in general and topics outside of science have been invaluable. Your guidance and advice have helped me to grow both professionally and personally. I am also thankful for other members of Molecular Pathology Research Group, particularly to Kai, Liina, and Martti. Your expertise, collaboration, and hard work have been instrumental for me to complete my PhD, and I appreciate all that you have done.

I would like to extend special thanks to my thesis opponents, Assoc. Prof. Benjamin Fairfax and Assist. Prof. Can Kesmir. Benjamin, I appreciate all your comments and suggestions; they have been instrumental in shaping my thesis. Can, thank you for agreeing to be my opponent at such short notice. I am sincerely grateful for your willingness to participate. I would also like to express my gratitude to my thesis reviewers, my colleague Dima, and Assoc. Prof. Harri Lähdesmäki.

The BIIT research group has become a second family to me, and I am grateful for the support, laughter, and friendship I have experienced here. I would like to thank Elena, Erik, Uku, Ivan, Lemps, Kaur, Nurlan, Mari-Liis, Sulev, Tõnis, and many others for the great (often philosophical) conversations and overall good memories. You have created an incredibly friendly and productive work environment. I would also like thank Jaak for his leadership of the BIIT research group. Your guidance and hard work have been essential for creating such a nice workplace.

I would like to give a special thank you to Liis and Kaido. Liis, your strong principles, getting-things-done attitude and ability to say no when needed, have been a real inspiration. Kaido, your knowledge and spectacular logical reasoning skills are something that I truly admire. I also appreciate our discussions on various topics and the fun activities we have shared. Our friendship is truly invaluable to me.

My family has given me a constant support throughout my PhD journey. I am very grateful for my parents, Helve and Lembit, being there for me, supporting me when needed (and much more), and always encouraging and believing in me. Also, I am very fortunate to have such a good and caring sister like you have been,

Liina. I am also thankful for my fiancée, Kristel, for her love and support and for bringing immense joy to my life every day but also motivating me to be the best version of myself. I am grateful for my son, Romet, who is the highlight of my day. I cherish every moment I spend with you!

Lastly, I would like to express my gratitude to my dear friends. Freddy, you have been my closest comrade since our university days, and I cannot thank you enough for all the fun memories. You have been a true inspiration, demonstrating through your actions what it means to be a genuinely good person. I would also like to thank a "gang" that now goes by the name "Homaarid" – Kreete, Mona, Marili, Anette, and your significant others, thank you for making university time so much more eventful. I am also lucky to have some amazing childhood friends in my life – Matti, Kristian, Rene, Marko, Taavi, Silver, Morten, Mikk, Alar, and Jonas. You have brought a lot of joy to my life, and I thank you for the fun memories, entertaining conversations, and also being motivating examples to achieve more in life.

Thank you all for your support and contributions to my PhD journey. I am deeply grateful for everything you have done for me, and I will always cherish these memories.

# SISUKOKKUVÕTE

## Immunoloogia erinevate tahkude bioinformaatiline analüüs

Bioinformaatika on väljakutseid pakkuv ja interdistsiplinaarne valdkond, mis ühendab endas nii bioloogia, statistika kui ka arvutiteaduse. Selles valdkonnas luuakse nii uusi rakendusi kui ka viiakse läbi bioloogiliste andmete eeltöötlust ja analüüsi. Üks mis kindel, selle eesmärgiks on jõuda lähemale bioloogiliste protsesside mõistmisele, mis peituvad andmete "taga". Kuigi bioinformaatika kui teadusharu on eksisteerinud juba üle 50 aasta, on selle tähtsus molekulaarbioloogias kiiresti kasvanud, eriti viimase kümnendi jooksul. See on suures osas tänu sekveneerimise tehnoloogiate ja mikrokiipide kiirele arengule, mis on võimaldanud toota suures koguses andmeid. Seetõttu on bioinformaatika muutunud igapäevaseks osaks molekulaarbioloogia uurimistööst.

Oma doktoritöös andsin ülevaate viiest artiklist, milles kasutasime bioinformaatilist analüüsi erinevate immuunsüsteemi aspektide kirjeldamiseks ja mõistmiseks. Enne artiklite juurde asumist andsin lugejale ülevaate molekulaarbioloogiast ning keskendusin detailsemalt epigeneetikale, eriti DNA metülatsioonile, sest see mängis märkimisväärset rolli kahes töösse kaasatud artiklis. Lisaks tutvustasin lühidalt immunoloogia maailma ning põgusalt mainisin ka vananemisega seotud muutusi. Teises peatükis tutvustasin meetodeid, mida me kasutasime kõige sagedamini antud artiklites. See ülevaade hõlmas nii statistilisi teste, korrelatsioone, dimensionaalsuse vähendamise meetodeid kui ka klasteranalüüsi. Lõpuks andsin ka lühikese sissejuhatuse masinõppesse, tutvustades töös kasutatud mudeleid ning mudeli hindamise meetodeid. Järgnevas neljas peatükis andsin ülevaate töös kasutatud artiklitest.

Esimeses artiklis keskendusime erinevate T-raku alatüüpide kirjeldamisele. Me valisime selleks just eelkõige eakamaid inimesi hõlmava kohordi, sest mitmeid T-raku alatüüpe on seostatud madalatasemelise põletiku, krooniliste haiguste ja üldise immuunkompetentsi langusega, mis esinevad just vanemaealistel inimestel. Üheks selliseks alatüübiks on näiteks terminaalselt diferentseerunud $CD8^+$ TEMRA rakud, mis olid ka meil täiendava tähelepanu all. Me leidsime, et ehkki üldiselt võiks öelda, et $CD8^+$ TEMRA rakkude arv suureneb vanusega, siis see seos kaob enam kui 65-aastaste inimeste hulgas ning selle rakupopulatsiooni tasemete varieeruvust saab seletada muude teguritega, näiteks tsütomegaloviiruse infektsiooniga. Lisaks viisime läbi nende ja paljude teiste T-rakkude sait-spetsiifilise DNA metülatsiooni analüüsi, mille tulemusena leidsime väikese hulga CpG saite, mille metülatsioonitasemed võimaldasid meil treenida mudeli $CD8^+$ TEMRA rakkude osakaalu ennustamiseks. Me usume ja loodame, et see mudel või selle järgnevad iteratsioonid võiksid olla kasulikud inimese tervisliku seisundi hindamisel. Täpsemalt, me loodame, et see võimaldab tuvastada inimesi, kellel on küll antud rakud kuhjunud, kuid kellel ei ole veel terviseprobleeme ilmnenud. Seega aitaks ühest küljest selline tööriist neil inimestel tulevastele haigustele varem jaole saada ning

ühtlasi motiveeriks neid ka haigusi ennetama.

Teises publikatsioonis uurisime DNA metülatsiooni erinevusi regulatoorsete ja konventsionaalsete $CD4^+$ T-rakkude vahel. Leidsime, et regulatoorsetel T-rakkudel on vähem defineeritud globaalne metülatsioonijaotus ning ühtlasi olid neil diferentsiaalselt metüleeritud positsioonid pigem alametüleeritud võrreldes konventsionaalsete T-rakkudega. Eriti huvitavaks leiuks osutus kilpnääret stimuleeriva hormooni retseptori (TSHR) geeni lähedal paiknevate CpG saitide alametüleeritus ning antud geeni Tregi-spetsiifilise ekspressioon $CD4^+$ T-rakkude hulgas. Nimelt on teada, et see geneetiline regioon on võtmetähtsusega Graves'i haiguse puhul. Kuigi sellele järgnev Treg-rakkude analüüs tervete inimeste ja Graves'i haigete vahel ei näidanud antud lookuses erinevusi DNA metülatsiooni osas, siis siiski loodame, et meie uurimus loob aluse tulevastele uuringutele. Need uuringud võiksid selgitada, kas selles regioonis toimuvad muutused mõjutavad Treg-rakkude fenotüüpi ning kui nad seda teevad, siis kas need muutused võivad kaasa aidata autoimmuunhaiguste tekkele.

Kolmandas artiklis keskendusime tüümuse epiteeli rakkude diferentseerumise uurimisele. Tüümuse säsi epiteelrakud (mTEC-id) on võimelised ekspresseerima koespetsiifilisi antigeene tänu transkriptsioonifaktor AIRE-le. Nende antigeenide eskpressioon võimaldab neil läbi viia negatiivset selektsiooni, suunates autoreaktiivsed T-rakud apoptoosi. Kuigi mTEC-id on hästi uuritud, on nende kohta teada vähe pärast AIRE-ekspressiooni kadumist. Antud töös näitasime, et mTEC-ide edasine diferentseerumine sarnaneb keratinotsüütide diferentseerumisega. Lisaks toetas meie analüüs hüpoteesi, et mTEC-i hilisemad staadiumid ja Hassali kehad aitavad tüümuses luua püsiva põletikulise keskkonna, mis on oluline T-rakkude repertuaari kujunemisel. Meie analüüsi põhjal võiksid selles protsessis osaleda valgud S100A8 ja S100A9. Lisaks leidsime ka kaks kliiniliselt olulist autoantigeeni, mis olid esindatud mTEC-i diferentseerumise hilises staadiumis.

Antud doktoritöö kahe viimase artikli keskseks teemaks oli SARS-CoV-2 ning sellest tingitud COVID-19. Meie analüüs näitas, et COVID-19 on seotud apoptootiliste signaaliradadega ning seda eriti raske haigestumise puhul. Näiteks leidsime, et antiapoptootiline valk HGF esines kõrgemal tasemel patsientidel, kes olid suunatud intensiivravi osakonda (ICU). Tähelepanuväärne on ka asjaolu, et hilisemad uuringud on näidanud, et HGF-i tasemeid veres saab kasutada biomarkerina ICU ja mitte-ICU patsientide eristamiseks. Lisaks kinnitasime mitmeid varasemaid leide, nagu näiteks raskema haiguskulu seotust kõrgemate proinflammatoorsete tsütokiinide tasemetega. Meie analüüs näitas ka seda, et isegi asümptomaatilistel inimestel võib SARS-CoV-2 infektsioon kaasa tuua pikaajalise proinflammatoorsete tsütokiinide taseme tõusu.

Lõpetuseks, loodan et antud töö pakkus lugejale omajagu huvitavaid avastusi ja andis arusaadava ülevaate antud valdkonnast ning kasutatud metoodikast. Samuti soovin, et minu töö innustab biolooge omandama oskusi andmeteaduse valdkonnas.

# PUBLICATIONS

# CURRICULUM VITAE

## Personal data


| | |
|---|---|
| Name: | Ahto Salumets |
| Date of birth: | January 29th 1992 |
| Nationality: | Estonian |
| E-mail: | ahto.salumets@ut.ee |


## Education

| | |
|---|---|
| 2016–... | PhD in computer science, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia |
| 2014–2016 | MSc in gene technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia |
| 2011–2014 | BSc in gene technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia |
| 1999–2011 | Saue Gymnasium, Estonia |

## Employment

| | |
|---|---|
| 2023–... | Data Scientist, Transporeon GmbH |
| 2022–2023 | Junior Data Scientist, Transporeon GmbH |
| 2019–2022 | Junior Research Fellow of Bioinformatics, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia |
| 2016–2019 | Research Project Specialist, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia |

## Honours & awards

| | |
|---|---|
| 2021 | Co-supervised BSc student Liisa Pomerants to 2nd prize in the Estonian National Contest for University Students with her BSc thesis "Analysis of the B cell receptor repertoire with two methods" |
| 2019 | 1st place in the Estonian Bioinnovation Days 2019 hackaton with a DNA methylation based application that aims to monitor individuals' health |

## Teaching

| | |
|---|---|
| 2018/2019 | Spring semester - Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Teaching Assistant in Bioinformatics Seminar |
| 2017/2018 | Spring semester - Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Teaching Assistant in Bioinformatics Seminar |
| 2017/2018 | Autumn semester - Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Teaching Assistant in Bioinformatics Seminar |

## Supervised theses

| | |
|---|---|
| 2022 | Anella Salmistu, MSc thesis "DNA methylation pattern-based analysis of $CD4^+$ regulatory T cells" |
| 2022 | Alexandra Elsakova, MSc thesis "Harnessing epigenetic changes to estimate immune cell levels" |
| 2022 | Simo Pähk, BSc thesis "Prediction of Cell Counts from DNA Methylation" |
| 2022 | Maksym Zarodniuk, BSc thesis "Profiling antibody responses and their contribution to disease in APS1 rat model using high-throughput sequencing approaches" |
| 2021 | Liisa Pomerants, BSc thesis "Analysis of the B cell receptor repertoire with two methods" |
| 2020 | Katrin Ruisu, MSc thesis "Analysis and Web Interface Generation for the Qumanize Application" |
| 2019 | Anella Salmistu, BSc thesis "Evaluation of $CD8^+$ TEMRA cells in psoriasis" |

## Scientific work

Main fields of interest:

- bioinformatics
- data science
- immunology
- epigenetics
- genomics

## Publications and preprints

2022 **Ahto Salumets**, Liina Tserel, Silva Kasela, Maia Limbach, Lili Milani, Hedi Peterson, Kai Kisand, and Pärt Peterson. Graves' disease-Associated TSHR gene is demethylated and expressed in human regulatory T cells. ***BioRxiv*** (2022).

2022 **Ahto Salumets***, Liina Tserel*, Anna P. Rumm, Lehte Türk, Külli Kingo, Kai Saks, Astrid Oras, Raivo Uibo, Riin Tamm, Hedi Peterson, Kai Kisand, and Pärt Peterson. Epigenetic quantification of immunosenescent $CD8^+$ TEMRA cells in human blood. ***Aging Cell*** 21.5 (2022).

2021 Martti Laan, **Ahto Salumets**, Annabel Klein, Kerli Reintamm, Rudolf Bichele, Hedi Peterson, and Pärt Peterson. Post-Aire medullary thymic epithelial cells and Hassall's corpuscles as inducers of tonic pro-inflammatory microenvironment. ***Frontiers in Immunology*** 12 (2021).

2021 Liina Tserel, Piia Jõgi, Paul Naaber, Julia Maslovskaja, Annika Häling, **Ahto Salumets**, Eva Zusinaite, Hiie Soeorg, Freddy Lättekivi, Diana Ingerainen, Mari Soots, Karolin Toompere, Katrin Kaarna, Kai Kisand, Irja Lutsar, and Pärt Peterson. Long-term elevated inflammatory protein levels in asymptomatic SARS-CoV-2 infected individuals. ***Frontiers in Immunology*** 12 (2021).

2020 Liis Haljasmägi*, **Ahto Salumets***, Anna Pauliina Rumm*, Meeri Jürgenson, Ekaterina Krassohhina, Anu Remm, Hanna Sein, Lauri Kareinen, Olli Vapalahti, Tarja Sironen, Hedi Peterson, Lili Milani, Anu Tamm, Adrian Hayday, Kai Kisand, and Pärt Peterson. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. ***Scientific Reports*** 10.1 (2020).

2020 Jon Ison, Hervé Ménager, Bryan Brancotte, Erik Jaaniso, **Ahto Salumets**, Tomáš Raček, Anna-Lena Lamprecht, Magnus Palmblad, Matúš Kalaš, Piotr Chmura, John M Hancock, Veit Schwämmle, Hans-Ioan Ienasescu. Community curation of bioinformatics software and data resources. ***Briefings in Bioinformatics*** 21.5 (2020).

2019 Jon Ison, Hans Ienasescu, Piotr Chmura, Emil Rydza, Hervé Ménager, Matúš Kalaš, Veit Schwämmle, Björn Grüning, Niall Beard, Rodrigo Lopez, Severine Duvaud, Heinz Stockinger, Bengt Persson, Radka Svobodová Vařeková, Tomáš Raček, Jiří Vondrášek, Hedi Peterson, **Ahto Salumets**, Inge Jonassen, Rob Hooft, Tommi Nyrönen, Alfonso Valencia, Salvador Capella, Josep Gelpí, Federico Zambelli, Babis Savakis, Brane Leskošek, Kristoffer Rapacki, Christophe Blanchet, Rafael Jimenez, Arlindo Oliveira, Gert Vriend, Olivier Collin, Jacques van Helden, Peter Løngreen, and Søren Brunak. The bio.tools registry of software tools and data resources for the life sciences. ***Genome Biology*** 20.1 (2019).

# ELULOOKIRJELDUS

## Isikuandmed


| | |
|---|---|
| Nimi: | Ahto Salumetes |
| Sünniaeg: | 29.01 1992 |
| Rahvus: | Eestlane |
| E-post: | ahto.salumets@ut.ee |


## Haridus

| | |
|---|---|
| 2016–... | Doktoriõpe informaatika erialal, arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti |
| 2014–2016 | Magistriõpe geenitehnoloogia erialal, molekulaar- ja rakubioloogia instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool, Eesti |
| 2011–2014 | Bakalaureuseõpe geenitehnoloogia erialal, molekulaar- ja rakubioloogia instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool, Eesti |
| 1999–2011 | Saue Gümnaasium, Eesti |

## Teenistuskäik

| | |
|---|---|
| 2023–... | Andmeteadlane, Transporeon GmbH |
| 2022–2023 | Nooremandmeteadlane, Transporeon GmbH |
| 2019–2022 | Bioinformaatika nooremteadur, arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool |
| 2016–2019 | Teadusprojekti spetsialist , arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool |

## Teaduspreemiad ja tunnustused

| | |
|---|---|
| 2021 | Juhendatav Liisa Pomerants saavutas II preemia üliõpilaste teadustööde riiklikul konkursil bakalaureusetööga "B-raku retseptorite repertuaari uurimine kahel meetodil" |
| 2019 | I koht Eesti Bioinnovatsiooni päevad 2019 häkatonil rakenduse eest, mis hindab inimese tervislikku seisundit DNA metülatsiooni põhjal |

## Õppetöö

| | |
|---|---|
| 2018/2019 | Kevadsemester - arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool, õppeassistent aines bioinformaatika seminar |
| 2017/2018 | Kevadsemester - arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool, õppeassistent aines bioinformaatika seminar |
| 2017/2018 | Sügissemester - arvutiteaduse instituut, loodus- ja täppisteaduste valdkond, Tartu ülikool, õppeassistent aines bioinformaatika seminar |

## Juhendatud väitekirjad

| | |
|---|---|
| 2022 | Anella Salmistu, magistritöö "DNA metülatsioonipõhine $CD4^+$ regulatoorsete T-rakkude analüüs" |
| 2022 | Alexandra Elsakova, magistritöö "Epigeneetiliste muutuste kasutamine immuunrakkude tasemete hindamiseks" |
| 2022 | Simo Pähk, bakalaureusetöö "Rakkude arvukuse mudeldamine DNA metülatsiooni põhjal" |
| 2022 | Maksym Zarodniuk, bakalaureusetöö “Antikeha vastuste profileerimine ning nende haigusseoselise rolli välja selgitamine APS1 roti mudelil kasutades kõrge läbilaskevõimega sekveneerimismeetodeid" |
| 2021 | Liisa Pomerants, bakalaureusetöö "B-raku retseptorite repertuaari uurimine kahel meetodil" |
| 2020 | Katrin Ruisu, magistritöö "Qumanize rakenduse analüüs ja sellele veebiliidese loomine" |
| 2019 | Anella Salmistu, bakalaureusetöö "$CD8^+$ TEMRA rakkude hindamine psoriaasi korral" |

## Teadustegevus

Peamised uurimisvaldkonnad:

- bioinformaatika
- andmeteadus
- immunoloogia
- epigeneetika
- genoomika

## Publikatsioonid ja eeltrükid

2022 **Ahto Salumets**, Liina Tserel, Silva Kasela, Maia Limbach, Lili Milani, Hedi Peterson, Kai Kisand, and Pärt Peterson. Graves' disease-Associated TSHR gene is demethylated and expressed in human regulatory T cells. ***BioRxiv*** (2022).

2022 **Ahto Salumets***, Liina Tserel*, Anna P. Rumm, Lehte Türk, Külli Kingo, Kai Saks, Astrid Oras, Raivo Uibo, Riin Tamm, Hedi Peterson, Kai Kisand, and Pärt Peterson. Epigenetic quantification of immunosenescent $CD8^+$ TEMRA cells in human blood. ***Aging Cell*** 21.5 (2022).

2021 Martti Laan, **Ahto Salumets**, Annabel Klein, Kerli Reintamm, Rudolf Bichele, Hedi Peterson, and Pärt Peterson. Post-Aire medullary thymic epithelial cells and Hassall's corpuscles as inducers of tonic pro-inflammatory microenvironment. ***Frontiers in Immunology*** 12 (2021).

2021 Liina Tserel, Piia Jõgi, Paul Naaber, Julia Maslovskaja, Annika Häling, **Ahto Salumets**, Eva Zusinaite, Hiie Soeorg, Freddy Lättekivi, Diana Ingerainen, Mari Soots, Karolin Toompere, Katrin Kaarna, Kai Kisand, Irja Lutsar, and Pärt Peterson. Long-term elevated inflammatory protein levels in asymptomatic SARS-CoV-2 infected individuals. ***Frontiers in Immunology*** 12 (2021).

2020 Liis Haljasmägi*, **Ahto Salumets***, Anna Pauliina Rumm*, Meeri Jürgenson, Ekaterina Krassohhina, Anu Remm, Hanna Sein, Lauri Kareinen, Olli Vapalahti, Tarja Sironen, Hedi Peterson, Lili Milani, Anu Tamm, Adrian Hayday, Kai Kisand, and Pärt Peterson. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. ***Scientific Reports*** 10.1 (2020).

2020 Jon Ison, Hervé Ménager, Bryan Brancotte, Erik Jaaniso, **Ahto Salumets**, Tomáš Raček, Anna-Lena Lamprecht, Magnus Palmblad, Matúš Kalaš, Piotr Chmura, John M Hancock, Veit Schwämmle, Hans-Ioan Ienasescu. Community curation of bioinformatics software and data resources. ***Briefings in Bioinformatics*** 21.5 (2020).

2019 Jon Ison, Hans Ienasescu, Piotr Chmura, Emil Rydza, Hervé Ménager, Matúš Kalaš, Veit Schwämmle, Björn Grüning, Niall Beard, Rodrigo Lopez, Severine Duvaud, Heinz Stockinger, Bengt Persson, Radka Svobodová Vařeková, Tomáš Raček, Jiří Vondrášek, Hedi Peterson, **Ahto Salumets**, Inge Jonassen, Rob Hooft, Tommi Nyrönen, Alfonso Valencia, Salvador Capella, Josep Gelpí, Federico Zambelli, Babis Savakis, Brane Leskošek, Kristoffer Rapacki, Christophe Blanchet, Rafael Jimenez, Arlindo Oliveira, Gert Vriend, Olivier Collin, Jacques van Helden, Peter Løngreen, and Søren Brunak. The bio.tools registry of software tools and data resources for the life sciences. ***Genome Biology*** 20.1 (2019).

# DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

# DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.