UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Molecular Pathology Research Group

Alexandra Elsakova

HARNESSING EPIGENETIC CHANGES TO ESTIMATE IMMUNE CELL LEVELS

Master's Thesis

Curriculum Bioengineering

Supervisors:
Research Fellow of Molecular Immunology,
PhD Liina Tserel;
Junior Research Fellow in Bioinformatics,
MSc Ahto Salumets

Tartu 2022

ACKNOWLEDGMENTS

The author expresses sincere gratitude to the workers of the Laboratory of Molecular Pathology of the Institute of Biomedicine and Translational Medicine, supervisors of the work (PhD Liina Tserel and MSc Ahto Salumets) for their help in the experimental part, training in the methodological principles of laboratory methods, computational analysis, and modelling. I am sincerely grateful to PhD Liina Tserel and MSc Ahto Salumets; Professors of Immunology, PhD Pärt Peterson and PhD Kai Kisand for help in discussing the results and their interpretation, valuable advice and comments; for PhD Kai Kisand and PhD Liina Tserel for teaching the methodological principles of using FACS and FSC express; for support and their professional perspective for my parents, Vladimir Elsakov and Yulia Krasheninnikova.

HARNESSING EPIGENETIC CHANGES TO ESTIMATE IMMUNE CELL LEVELS

Abstract:

Epigenetic modifications, in particular DNA methylation, are one of the age-dependent factors in the human body. It accumulates over time and in certain genomic regions, it acts as a biomarker for ageing while modifications in some other regions could be used to estimate immune system status. More specifically, each cell type has its own characteristic epigenetic profile and when a certain region is only methylated/demethylated in this particular region and the reverse is true for the rest of the cell types, it is possible to associate DNA methylation in the whole blood with given cell type levels.

To this end, whole blood from donors of various ages (from 20 to 99 years of age) was used for sequencing of bisulfite-treated DNA sequences of CpGs of interest. In addition, isolated peripheral blood mononuclear cells were used to determine the levels of cell populations selected for the modelling task. After that, the data were combined with previously existing data, and subsequently, a model for predicting the proportions of the cell type of interest was built. This work shows the possibility of using DNA methylation for determining the levels of Temra cells and suggests the possibility to use epigenetic models as alternatives to flow cytometry.

Keywords: ***Ageing, epigenetics, DNA methylation, Temra cells***

CERCS: B500 Immunology, serology, transplantation; B110 Bioinformatics

EPIGENEETILISTE MUUTUSTE KASUTAMINE IMMUUNRAKKUDE TASEMETE HINDAMISEKS

Lühikokkuvõte:

Epigeneetilised modifikatsioonid, iseäranis just DNA metülatsioon, on üks vanuseseoselisi faktoreid. Epigeneetilised modifikatsioonid akumuleeruvad ajas ning teatavates genoomi regioonides saab neid kasutada vanuse ennustamiseks, kuid muutusi osades teistes regioonides saab kasutada ka teistel eesmärkidel, näiteks immuunsüsteemi seisundi hindamiseks. Nimelt iga rakk omab oma unikaalselt epigeneetilist profiili ning kui eksisteerib selline regioon, mis on uuritavas rakutüübis kas metüleeritud või metüleerimata, kuid kõigis teistes rakkudes on epigeneetiline efekt vastupidine, siis on võimalik kasutada DNA metülatsiooni täisveres antud rakutüübi taseme hindamiseks.

Antud töös eraldati täisveri doonoritelt vanusevahemikus 20-99 ning nende bisulfit-töödeldud DNA järjestused huvipakkuvate CpG saitide ümbruses sekveneeriti. Lisaks koguti perifeerse vere mononukleaarsed rakud hindamaks rakupopulatsioonide tasemeid. Saadud andmed kombineeriti varasema andmestikuga ning uuritavaid rakupopulatsioonide tasemeid mudeldati kasutades DNA metülatsiooni väärtusi kindlates CpG saitides. Antud töö näitab, et DNA metülatsiooni saab kasutada Temra rakkude ennustamiseks ning ühtlasi viitab, et epigeneetilisi mudeleid saaks kasutada alternatiivina voolutüstomeetriale.

Võtmesõnad: ***Vananemine, epigeneetika, DNA metülatsioon, Temra rakud***

CERCS: B500 Immunoloogia, seroloogia, transplantoloogia; B110 Bioinformaatika

TABLE OF CONTENT

27k – HumanMethylation27 (Illumina);

450k – HumanMethylation450 (Illumina);

5mC – 5-Methylcytosine;

AAH – Age acceleration calculated with Horvath's clock;

AAHa – Age acceleration calculated with Hannum's clock;

CD – a cluster of differentiation;

CMV – Cytomegalovirus;

CpG – Cytosine-phosphate-Guanine;

DENV – Dengue virus;

DNMT – DNA methyltransferase;

EDTA – Ethylenediaminetetraacetic acid;

FACS – Fluorescence-activated cell sorting or a specialized type of flow cytometry

HLA – human leukocyte antigen;

IL-6 – Interleukin 6;

PTCH1 – Protein patched homolog 1;

RA – Rheumatoid arthritis;

RB – Running buffer;

RBC – Red Blood Cell;

rcf – Relative Centrifugal Force;

RMSE – Root Mean Square Error;

RT – Room temperature;

SLE – Systemic lupus erythematosus;

TE – Tris-EDTA;

Temra – terminally differentiated effector memory T cells;

TET – ten-eleven translocation, are large (∼180- to 230-kDa) multidomain enzymes;

Tris – trisaminomethane;

WBC – whole blood cells.

INTRODUCTION

Nowadays, interest in ageing and age-related issues to improve the quality of human life and its duration is rising. *Ageing* is a process that manifests itself inevitably in life, leading to a decrease in the body's resistance to environmental influences (Gupta et al., 2006; da Costa et al., 2016). For humans being, ageing has always had a special meaning. For centuries, philosophers have discussed the causes of ageing, alchemists have been searching for the elixir of youth, and many religions have attached sacred significance to ageing. However, even today, the biology of the ageing process is still very poorly known, and there are no methods to change the rate of human ageing. Despite intensive research, scientists are still far from overcoming ageing. Today, advances in medicine and rising living standards have made it possible to significantly increase life expectancy (although the change in maximum life expectancy is negligible).

Ageing is not only a case of biological manifestation ("true age" of a person) but is also associated with the accumulation of various changes. Epigenetic processes during life are fundamental factors for determining the state of the body and provide an opportunity to assess the state of the immune system. DNA methylation, like other changes, has a multifaceted effect on the immunological status, influencing the development of cell lines (Makar & Wilson, 2004), differentiation of memory cells, inflammation processes (el Gazzar et al., 2008), cell division process (Robertson et al., 2000). The addition of a methyl residue to DNA affects the activation of the immune system, inflammation processes, and the manifestation of autoimmune, neurogenerative diseases (rheumatoid arthritis, Alzheimer's disease and so on) (Fogel et al., 2017; Jochems et al., 2020). Moreover, all these processes are closely related to the reactions of cellular immunity and its state.

This work shows the possibility of using epigenetic changes to quantify cell populations of effector memory re-expressing CD45RA ($CD4^+$ and $CD8^+$ Temra). In the course of the work, the sequencing results were processed and combined with previously obtained data to create DNA methylation-based models for predicting the levels of Temra cell populations.

1. LITERATURE

1.1. What is ageing?

Ageing is a process that paradoxically combines both immunodeficiency, inflammation (inflammatory ageing), and autoimmunity (Gupta et al., 2006), involving changes in the architecture and functioning of the immune system, often referred to as immunosenescence (Pawelec & Gupta, 2019). In the broadest sense, ageing combines the changes that occur in the body throughout life with high variability over time without a single period (Kirkwood, 2005). They might be both harmful (cardiovascular, metabolic, and neurological diseases, cancer, which leads to disability or death) and harmless (for example, wrinkles and grizzles (da Costa et al., 2016)). Thus, ageing is a gradual decline in bodily functions over time. Normal ageing is associated with a loss of function in various physiological processes and anatomical structures (blood pressure, postural dynamics of respiratory cycles, decreased fertility, and increased mortality risk). Ageing consists of interrelated mechanisms at various biological levels, which can be described at different levels of biological organisation (Guarente, 2014).

Ageing at the organism level is associated with the inability to maintain homeostasis, a manifestation of disorders, and an increased susceptibility to age-related diseases leading to death (Jaul & Barron, 2017; Saul & Kosinsky, 2021). These transformations are often associated with a change in the number of cells and the composition of tissues, a disturbance of intercellular signalling and the activity of response reactions to stress and metabolic changes (Moskalev et al., 2016).

At the tissue level, chronic inflammation occurs, which acts as a driver of many age-related diseases (for example, cardiovascular diseases). Metabolic regulation is impaired due to abnormal signalling mechanisms due to a lack of macronutrient balance (Zhang et al., 2015).

Cellular disruption is another level at which ageing occurs. With age, the number of senescent cells increases, characterised by stable cell cycle arrest. In addition, the formation of reactive oxygen species is enhanced because of a decrease in the activity of the mitochondrial respiratory chain and antioxidant enzymes. And finally, the response to protein denaturation in the endoplasmic reticulum is triggered, and cells are also unable to utilise damaged proteins and maintain proteostasis with the help of lysosomes and proteasomes (Korovila et al., 2017).

At the molecular level, ageing is manifested by the inability of the cell to completely restore damaged macromolecules (Moskalev et al., 2016).

In general, during ageing, mutations in genes accumulate (Yousefzadeh et al., 2021), telomeres get shorter (J. Zhang et al., 2016), and epigenetic modifications of DNA and histones take place (Saul & Kosinsky, 2021), leading to chromatin rearrangements (H. Chen et al., 2014). In this case, heterochromatinisation of chromosome regions important for cell life occurs (Corpet & Stucki, 2014), and deheterochromatinization of repetitive genome sequences occurs, leading to genetic instability (Tsurumi & Li, 2012).

To date, many factors underlying ageing are known. They generally consider this to be programmatic development (Tower, 2015), including molecular crosslinking (Stammers et al., 2020), damage caused by free radicals (Korovila et al., 2017), changes in immunological functions (Weyand & Goronzy, 2016), telomere shortening (Kruk et al., 1995) and the presence of DNA changes (Kruk et al., 1995; da Costa et al., 2016). However, a unified theory covering genes, the capacity of genetic maintenance and repair systems, environment, and chance has been increasingly accepted recently, emphasising the need for a systematic and comprehensive analysis of the ageing process.

Despite the numerous concepts of ageing and their classifications, in 2013, 9 main signs of ageing were identified and are discussed in the following chapter.

1.2. Hallmarks of ageing

The hallmarks of ageing are determined primarily by genetics, but environmental factors also influence or exacerbate these processes. Some of them contribute to the damage that occurs and accumulates over time and is ultimately responsible for age-related pathologies. The distinctive features of ageing determine the difference between chronological and biological age. In 2013 López-Otín et al. described 9 cellular hallmarks of ageing (Figure 1). Subsequently, in 2020 the list was expanded (Fedintsev & Moskalev, 2020) with the 10th hallmark of ageing. It is the accumulation of damage in long-lived macromolecules such as the proteins of the nuclear pore complex and extracellular matrix and histones.

Each of those hallmarks meets the following criteria: 1. occurs in the normal ageing process; 2. ageing is accelerated with experimental aggravation, and 3. ageing process slows down with the reverse process, resulting in life expectancy increase (López-Otín et al., 2013). According to the author's classification, selected signs can be grouped into three categories.

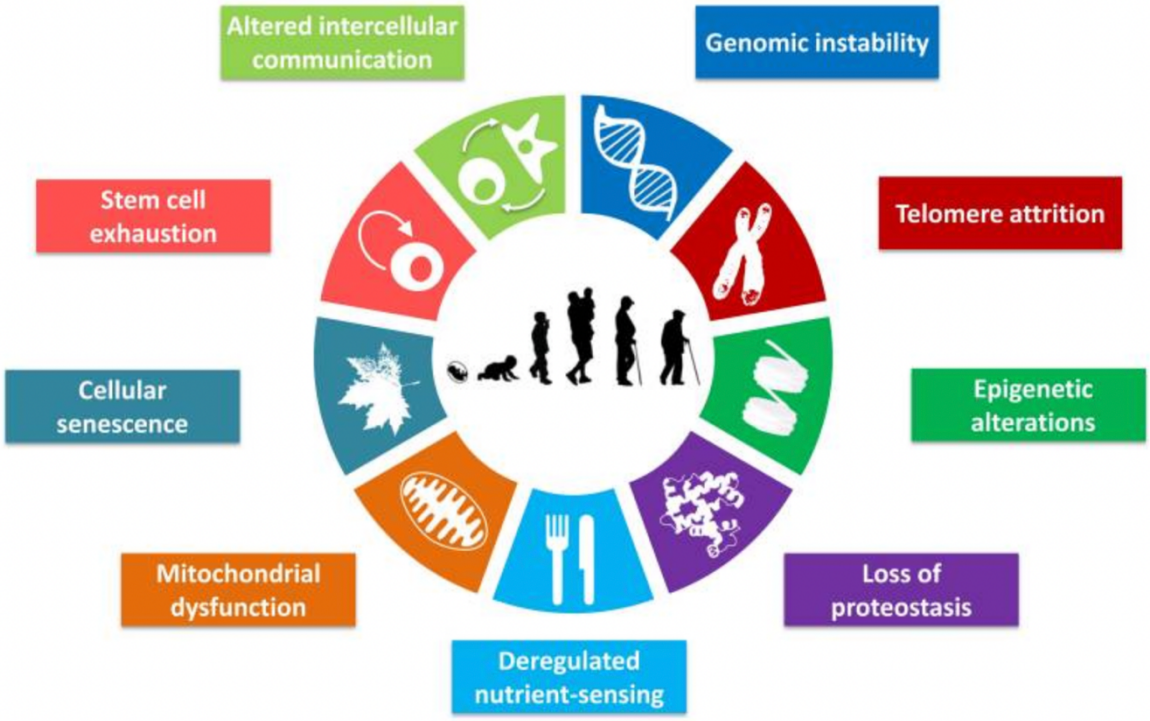


Figure 1. The scheme illustrates the nine hallmarks described in this review: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient-sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication (López-Otín et al., 2013)

Primary hallmarks (DNA damage, including chromosomal aneuploidies, mitochondrial mutations and telomere loss, epigenetic drift, and defective proteostasis) are the leading causes of cellular damage. Next antagonistic signs (mitochondrial dysfunction, cellular ageing) are considered part of the compensatory response to age-related damage. Finally, integrative signs (depletion stem cells, altered intercellular communication), which represent the result of the two previous groups, are responsible for ageing. The selection of these signs of ageing and their further study not only contributes to an increase in life expectancy but is also one of the facts in improving its quality.

1.2.1. Methylation as a prime example of epigenetic modifications

Epigenetic modifications are changes (including inherited ones) in gene expression patterns independent of primary DNA sequence changes, some of which affect the outcome of a locus or chromosome without changing the underlying DNA sequence. Methylation, histone modifications, and non-coding RNAs are examples of epigenetic modification (Sharples et al., 2018).

The process of DNA methylation is one of the earliest discovered epigenetic modifications in vertebrates (Varriale, 2014). DNA methylation usually occurs at CpG sites (cytosine phosphate guanine sites). Cytosine is converted to 5-methylcytosine (5mC). The formation of Me-CpG is catalysed by the enzyme DNA methyltransferase (DNMT) (Gujar et al., 2019). Changes in methylation of CpG islands within the promoter region are the primary regulators of gene expression (Figure 2).

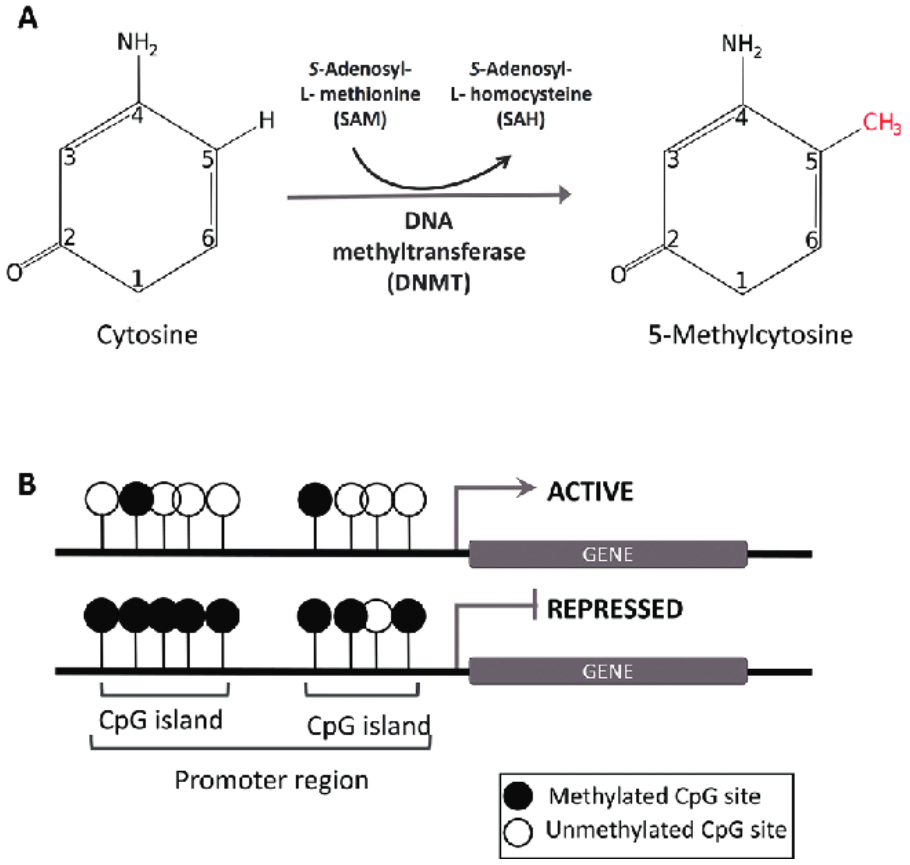


Figure 2. DNA methylation. (A) CpG methylation process DNA methyltransferases by adding the methyl group to the 5th carbon of cytosines that precedes guanine nucleotides. (B) CpG islands are DNA sequences rich in CpG sites. Methylation of CpG islands inside a promoter region may control gene expression (both activate and repress) (Alarcón et al., 2017)

Since DNA methylation in certain regions is strongly associated with age, many scientists have developed epigenetic clocks (Bell et al., 2019) intending to show biological age and the risk of developing age-related pathologies.

1.2.2. Epigenetic clocks

'DNA methylation age' or (DNAmAge) is one of the promising biomarkers of ageing and can also describe the age in various tissues and at different stages of life (Fransquet et al., 2019). There are two types of epigenetic clocks, one that aims to predict the person's age as accurately as possible (e.g. for forensics use) and the other to characterise disease status (e.g. biological age) (Figure 3). Several models for determining biological age have already been made:

1. Hovarth clock – aims to show a person's biological age and is trained on methylation levels of 353 CpG sites from the Illumina 27k array (H. Horvath & Horvath, 2013).

2. Hannum clock – also aims to indicate a person's biological age but is trained on the levels of 71 CpG sites from an Illumina 450k array (Hannum et al., 2013).

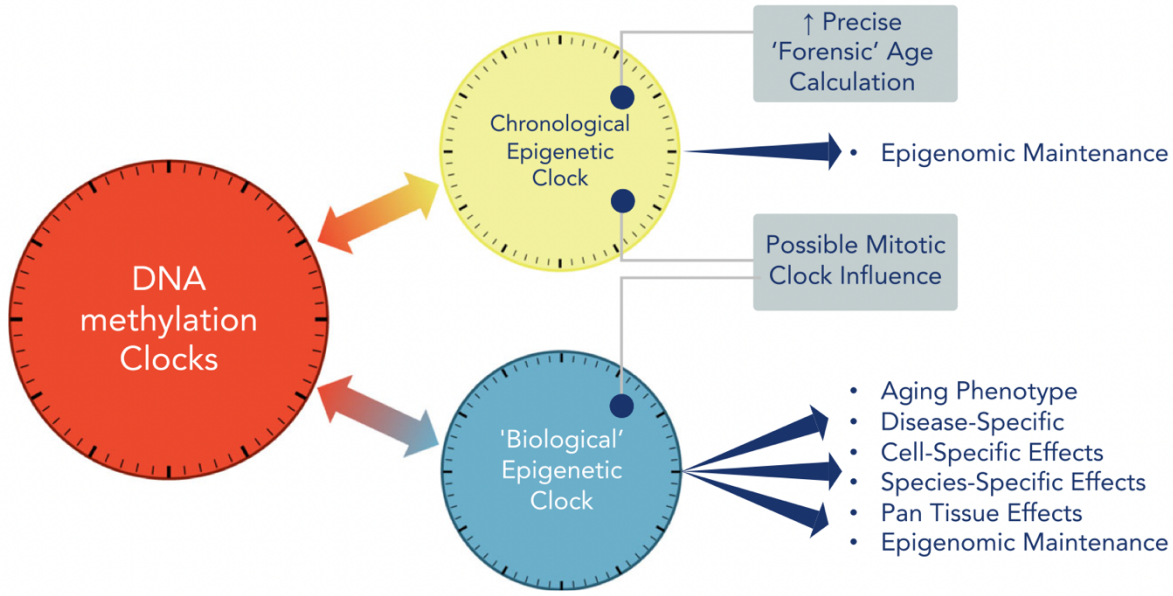


Figure 3. DNA methylation clock structure includes chronological and biological information (adapted from Bell et al., 2019)

The difference between the model's prediction and actual age indicates whether a person is biologically older or younger compared to their actual age. Later versions take into account the composition of blood cells. This ensures that the prediction is not affected by factors influencing blood components, such as the activation of the immune system (S. Horvath et al., 2016; Fransquet et al., 2019).

1.2.3. Methylation and occurrence of various diseases

DNA methylation is very stable compared to other biomarkers, for example, RNA or protein. DNA methylation is often observed in many types of cancer (gastrointestinal, glandular, bladder, Wilms cancer, ovarian, prostate and bone cancer, hepatocellular carcinoma, glioblastoma). In addition, numerous instances of DNMT mutation, varying levels of DNMT expression, or dysregulation of Ten-eleven translocation (TET) are implicated in cancer, all of which suggest a strong link between DNA methylation and cancer (Wajed et al., 2001; Huang & Rao, 2014; Jin & Liu, 2018; Salameh et al., 2020).

Another striking example is immunological pathologies. For example, using DNA methylation analysis, it was reported that altered human leukocyte antigen (HLA) class II DNA methylation could mediate the genetic risk of developing rheumatoid arthritis (RA) (Liu et al., 2013). Experimental confirmation was found using a model of RA in rats. Under expression of the PTCH1 protein in the RA rat model activated the Hedgehog signalling pathway leading to increased secretion of interleukin 6 (IL-6) and tumour necrosis factor-alpha (Jin & Liu, 2018). A similar trend was observed in patients with SLE (S. H. Chen et al., 2017; Chung et al., 2015), multiple sclerosis (Chomyk et al., 2017), as well as in metabolic disorders (Barres & Zierath, 2011; Dayeh et al., 2014) and neurological disorders (Lu et al., 2013; Pellegrini et al., 2021).

1.3. Immune cell population

The immune system is a complex network of organs and cells, consisting of many different cell types such as lymphocytes, macrophages, and macrophage-like cells, including dendritic cells of the spleen and Langerhans cells. The cells of the immune system are organised into tissues and organs (for example, the spleen and lymph nodes). The main cells of the immune system are leukocytes, which act in concert in the fight against bacterial and viral agents and are also involved in the inflammatory response to tissue damage. The cellular differentiation process is characterised by the emergence of certain macromolecules on the surface of the membranes. More precisely, those markers correspond to a particular stage of development (Várady et al., 2013). They are called CD (cluster of differentiation) antigens, 370 of which are currently known, leading to significant heterogeneity of the lymphocyte population (Figure 4).

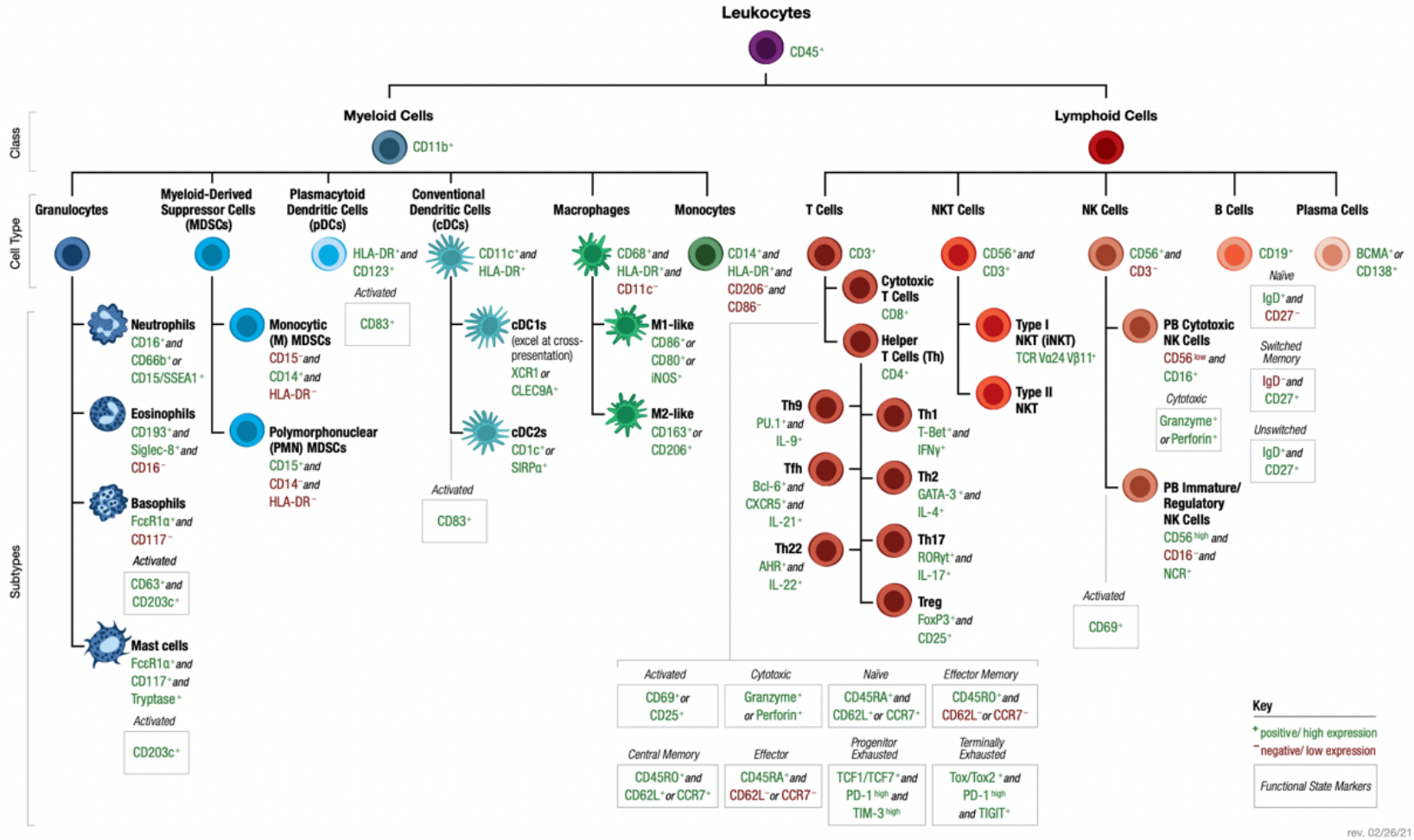


Figure 4. Different types of immune cells. Stages and clusters of leukocyte differentiation are presented (Human Immune Cell Marker Guide, 2022)

With the help of surface antigenic markers (differentiation antigens, CD), it is possible to determine the developmental stage of the cell. Differentiation antigens thus serve as specific markers. For this, surface antigens enable us to differentiate between different subpopulations of lymphocytes and other immunocompetent cells and extract them.

Pre-T cells that have entered the thymus lack the main differentiation markers (CD4 and CD8). They inhabit the upper part of the thymus cortex (subcapsular zone). In the mature thymus, such cells make up 5% of all thymocytes. Their interaction with the stroma of the subcapsular zone leads to the expression of the first specific T-receptor, CD2. Being in close contact with the thymus epithelial cells, they proliferate and complete their path of development in this area with a moderate expression of CD4 and CD8 (Goswami & Awasthi, 2020).

After that, CD8+ T cells enter the circulation as naive T cells co-expressing CD45RA, CCR7, CD27, and CD28 (Figure 5). Next, central memory cells (CM CD45RA-CCR7+) formation occurs after antigen presentation and loss of CD45RA and gaining of CD45RO (Maecker et al., 2012). Upon contact with antigen, CM T cells differentiate into T cells with effector functions (EM), passing through 3 subpopulations: early differentiated (ED) (CD27+CD28+), early-like differentiated (ELD) ($CD27^{-}CD28^{+}$) and intermediate differentiated (ID) ($CD27^{+}CD28^{-}$).

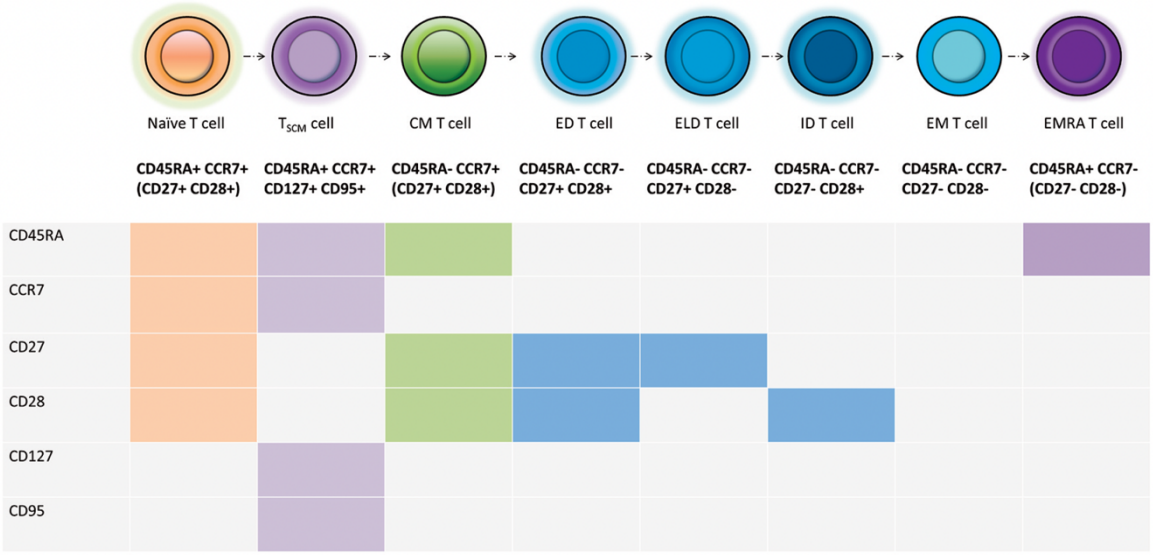


Figure 5. Linear T cell differentiation pathway of $CD4^{+}$ and $CD8^{+}$ T cell subsets; naive T cells, stem cell-like memory T cells, central memory, effector memory, early differentiated, early-like differentiated, intermediately differentiated and effector memory re-expressing CD45RA (Campbell & Narendran, 2019)

At the same time, the level of expression of these differentiation clusters gradually decreases during differentiation in response to antigenic stimulation, leading to an increase in cytotoxicity (Sallusto, 1999).

After losing both CD27 and CD28, T cells become fully differentiated EM T cells (Tomiyama et al., 2002). Further re-expression of CD45RA on fully differentiated effector memory T cells is the final stage of T cell maturation (Colin A. Michie et al., 1992; Campbell & Narendran, 2019). A similar process occurs for CD4+ T lymphocytes, however, they are also determined based on their differentiation into T-helper (Th) (CD45RO+CD127hiCD25low) and T-regulatory (TRegs) (CD4+CD127lowCD25+) subpopulations (Campbell & Narendran, 2019).

Temra cells are generally regarded as terminally differentiated effector cells, with the low secretion of interleukin-2 and high interferon-gamma secretion, high cytotoxicity, low proliferative capacity, and high sensitivity to apoptosis (Fulop et al., 2009). This subpopulation is the final stage of maturation of T-lymphocytes in peripheral blood, expressing markers of ageing, showing a reduced ability to proliferate and the decline in functional activity, having a high pro-inflammatory activity associated with the stage of the disease. Also, these cells are of high importance in antiviral activity: CMV and DENV specific T-cells mainly have the Temra phenotype (Libri et al., 2011; Weiskopf et al., 2015), with the same trend for immunological and oncological pathologies (Yang et al., 2019). At the same time, there exists significant variability in the levels of Temra cells within the population. This subset varies from <0.3% to almost 18% of the total number of T cells (Tian et al., 2017), establishing an opportunity for diagnostic and prognostic possibilities.

1.4. Applications of computational modelling

Mathematical modelling of both normal physiological and pathological processes is currently one of the most relevant areas in scientific research. In the study of biomedical problems, there are processes for the mathematical description of which the apparatus of ordinary differential equations, systems of nonlinear algebraic equations, difference mappings, theories of bifurcations, chaos and order are used. Examples of the successful use of such mathematical tools are known for predicting the development of a disease (Marchuck G., 1985), solving problems of nonlinear dynamics in biology, and chemical kinetics (Riznichenko G., 2010).

Modelling is also common in protein structure prediction, cell cycle dynamics, cancer research, and developmental biology (Barh et al., 2013). Thus, in a recent publication, Combalia et al. (2022) showed the possibility of classifying dermoscopic images using machine learning to

identify and classify skin cancer. Methylation data is also not an exception. The variants of DNA methylation clocks such as the Howarth and Hannum clocks described above are examples of the applications of methylation data in bioinformatics. Models are valuable from the point of view of fundamental science and have applied value, often reducing the experimental work time, and simulating the necessary results *in silico* (Bergsma & Rogaeva, 2020).

In 2018 Bergstedt et al. described the possibility of using methylation and gene expression profiles in determining cell composition, age, smoking activity and serostatus. This work was carried out based on the cellular deconvolution method (Houseman et al., 2016), given a bulk blood gene expression this method estimates underlying blood composition (main cell type levels). More than 850,000 methylation sites were included in the MethylationEPIC array, ordered and robust individual regression models were used to predict the circulating levels of 70 blood cell subsets measured by standardised flow cytometry in 962 healthy donors of Western European origin with high accuracy.

2. AIMS OF THE THESIS

Flow cytometry is one of the leading methods for determining cellular levels. It makes it possible to detect the necessary cell populations by quantifying the intensity of fluorescent markers. At the same time, FACS is time-consuming, labour-intensive and requires qualified practitioners. Indirect determination of cell populations by methylation levels serves as an alternative in this case (Salumets et al., 2022).

This work aimed to create a model for specific cell populations ($CD4^+$ and $CD8^+$ Temra) prediction from the methylation level.

To achieve this, the work was divided into the following goals:

1. DNA isolation from whole blood and its preparation for sequencing;
2. Obtaining methylation levels from selected sites using DNA sequencing;
3. Conducting FACS to get the levels of specific cells subpopulations;
4. Analysis and model construction using R.

3. EXPERIMENTAL PART

3.1. MATERIALS AND METHODS

3.1.1. Study group

The data obtained from 87 healthy donors aged 20 to 64 years (a group named "young individuals") was used for the study. From this group, the results of methylation levels (48 sites) using the whole blood were obtained, and various cell populations were analysed (in particular, $CD4^+$ and $CD8^+$ Temra). Subsequently, the dataset was extended with previously obtained data of patients mostly older than 65 years (a group named "old individuals") (Salumets et al., 2022). Altogether, the dataset consisted of 211 people aged from 20 to 99 years, among which 80 % were women. (Figure 6).

Following equipment was used for all the work: centrifuge and microcentrifuge (Eppendorf), cycler Mastercycler gradient (Eppendorf) and Arktic Thermal Cycler (ThermoFisher Scientific), Horizontal Electrophoresis System and PowerPac Basic Power Supply (Bio-Rad Laboratories). The study was approved by the Research Ethics Committee of the University of Tartu on February 15, 2021 (No 335/T-21). The work was conducted in the Molecular Pathology Research Group of the University of Tartu

3.1.2. Methods: Laboratory work

3.1.2.1. Preparation of white blood cells from EDTA blood

For white blood cell preparation, EDTA blood was used. To each whole blood sample (2 ml), we added up to 5 ml of RBC lysis solution (155 mM NH_4CL, 10 mM $KHCO_3$, 1 mM EDTA-NA_2, pH 8.0), after which the tube was manually mixed and left at +4 °C for 10 min. Next, centrifugation for 10 min at 600 rcf RT was carried out, followed by removal of the supernatant. After that, two washing steps with 2.5 ml of RBC Lysis Solution and centrifugation for 10 min at 500 rcf RT were performed. Next, 200 µl of RBC Lysis Solution was added to the tube and obtained white blood cells were suspended in a homogeneous solution. 2 mL of WBC Lysis Solution (10 mM Tris-HCL, pH 8.0, 25 mM EDTA-NA_2, pH 8.0, 2% SDS) to the mixture was added and vortexed to form a homogeneous solution.

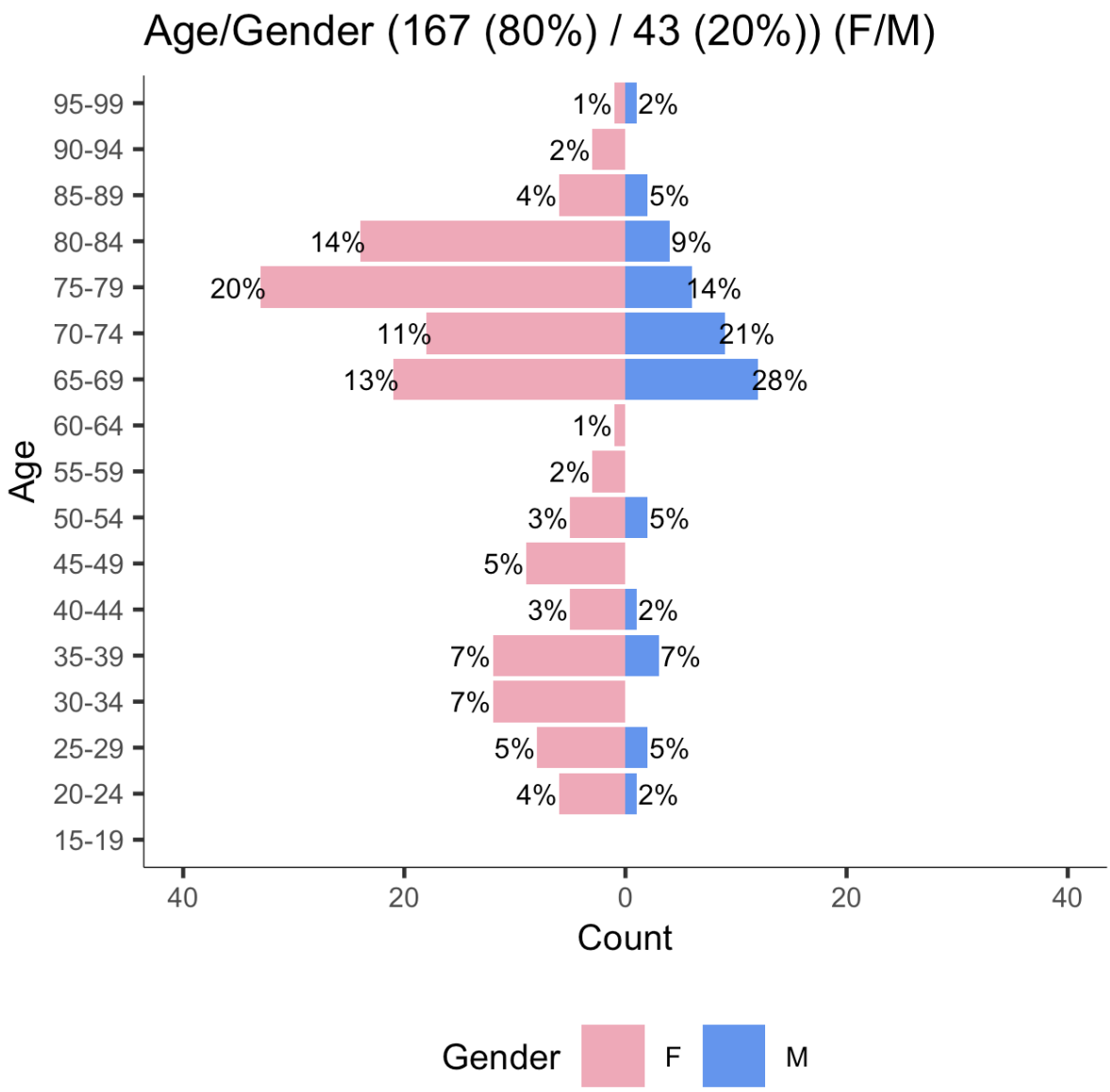


Figure 6. Gender-age distribution of study participants. Younger individuals are represented in the 20-64 age range, while the older group includes individuals from 65 to 99 years old. The y-axis shows the age of the donors, and the x-axis represents the number. Pink and blue colours describe the gender of individuals (female and male respectively)

3.1.2.2. Genomic DNA extraction

White blood cells were incubated for 3 hours at +37 °C. Cell lysate solution was placed at +4 °C for 15 min, and protein-membrane complexes were precipitated with 850 µl of 10 M ammonium acetate, with further centrifugation for 10 min at 2400 rcf RT. Next, 2 ml of isopropanol were added to clean 15 ml tubes, where supernatant was transferred. To precipitate the DNA, tubes were gently shaken for 5 min, after which the strand was caught and transferred to 2 ml of 70 % ethanol for washing for 5 min. The purified DNA was transferred into 1.5 ml

tubes, and the remaining ethanol was vented. 100 µl of 1X TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA-NA_2 pH 8.0) was added to each sample, followed by incubation for 30 min at +56 °C. Finally, tubes were placed on a slow shaker overnight at RT to ensure complete dissolution of the DNA strand in the buffer. After isolation, nucleic acid concentration was measured using NanoDrop ND-1000 spectrophotometry (ThermoFisher Scientific). The mean concentration of genomic DNA was 692.8 ng/µl.

3.1.2.3. Methylation sites detection

According to the manufacturer's instructions, genomic DNA in the amount of 600 ng was treated with sodium bisulfite using the EZ DNA Methylation Kit (Zymo Research Corporation) (EZ DNA MethylationTM Kit…, 2022). Bisulfite treated DNA was amplified in a 10 µl reaction containing 0.72 ng/µl of DNA, 1X Yellow PCR Buffer (Naxo), 1.5 mM MgCl2 (New England BioLabs), 2 mM dNTP mix (Solis BioDyne), 0.2 µM of each primer (Table S1), and 0.06 U/µl HOT FIREPol DNA Polymerase (Solis BioDyne). Cycle conditions were: 1 cycle for 15 min 95 °C; 40 cycles (95 °C for 20 s, 56 °C for 30 s, 72 °C for 1 min); and final elongation at 72 °C 3 min. Primers (Table S1) were designed by a Molecular Pathology (UT) group researcher in a specific way with overhangs that promote the CpGs of interest amplification with the use of UCSC Genome Browser on Human (GRCh37/hg19). To verify the reaction results, electrophoresis was performed in 1.5 % agarose gel with 0.5 mg/mL of Ethidium Bromide (EtBr) for 50 minutes at 120 V for all samples with three randomly selected primer pairs. Subsequently, the nucleic acid concentration was measured using the Qubit Fluorometric Quantification with a Qubit Quantification Starter Kit (ThermoFisher Scientific) using the manufacturer's protocol (Fisher Scientific, 2022). The DNA concentration at this step was 15±4 ng/µl. Next, the combination of amplicons in an equal amount from each individual was carried out. The resulting mixture was purified using Agencourt AMPure XP beads produced by Beckman Coulter (Agencourt ampure ® PCR purification, 2022) and then labelled with Illumina indexes with Nextera XT KIT D (Illumina). Each reaction in the presence of 15 µl contains 5 ng/µl of DNA, Nextera XT index 1 Primer 2.5 µl, Nextera XT index 2 Primer 2.5 µl, 2X KAPA HiFi HotStart PCR mix 7.5 µl (ThermoFisher Scientific). Cycle conditions were as followed 1 cycle for 3 min 95 °C; 7 cycles (95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s); and final elongation for 72 °C 5 min. This process was completed by another round of amplicon purification with Agencourt AMPure XP beads. After this spot check of amplicon income accumulation by the Qubit method, the resulting value was 39±7 ng/µl. Sequencing of bisulfite-treated DNA was performed with Illumina MiSeq at the Core Facility of the Institute of Genomics of the University of Tartu.

3.1.2.4. Flow Cytometry

Cell populations were analysed by the flow cytometry method. As the markers, we used the list of antibodies (Biolegend and BD Biosciences*) (Table 1), while the reaction medium was 1X Running Buffer (25 mM EDTA, 0.5 % BSA, 1X PBS). The list of antibodies specified for T cells was used for most of the work; however, to obtain the exact number of events for T lymphocytes, MDSC markers were used. First, a compensation analysis was carried out using high, medium and low the SPHERO™ Calibration Particles (Spherotech). For thawing the cells, peripheral mononuclear blood cells were added to 10 ml of RPMI media + 10 % FBS + PS, after which centrifugation was performed (10 min at 300 rcf RT), followed by removal of the supernatant and addition of 1 ml of same media. The number of cells was measured using the LUNA-FL™ Dual Fluorescence Cell Counter (Logos Biosystems). To 18 µl of the cell suspension, 2 µl of Acridine Orange Stain (Logos Biosystems) were added, and the cell number was calculated. For further work, we used a maximum of 2 million cells obtained after centrifugation for 10 min at 300 rcf RT followed by adding a mixture of antibodies in a volume of 50 µl. The mixture was incubated for 30 min at 4 °C with 2 subsequent washing steps (1 ml of RB and centrifugation of 10 min 300 rcf RT). Cells were examined with an LSRFortessa™ Cell Analyzer (BD Biosciences), and obtained data were analysed using FCS Express 5 Flow De Novo Software with a gating strategy specified in Figure S1-S2.

Table 1. Markers for flow cytometry analysis for A. – T cell populations and B. – Myeloid-derived suppressor cell populations. The volume of each antibody, as well as the RB, is specified for one sample

(A) T cell			
Catalogue Number	AB name	Marker	Vol, µl
302604	FITC	CD25	5
307630	PerCP-Cy5.5	HLA-DR	0,6
303115	APC	CD31 1:10	3
317426	Alexa Fluor 700	CD4 1:10	3
351310	Brilliant Violet 421	CD127	2
302836	Brilliant Violet 510	CD27	1,25
329924	Brilliant Violet 605	PD1	5
317324	Brilliant Violet 650	CD3	1,2
359612	PE	CD57	0,75
353236	PE-Dazzle	CCR7	2
302910	PE-Cy5	CD28 1:10	3
304126	PE-Cy7	CD45RA 1:10	2
563795	BUV395*	CD8	0,5
		AB total	29,3
		RB	20,7

(B) MDSC			
Catalogue Number	AB name	Marker	Vol, µl
303304	FITC	CD33 1:10	3
317323	BV650	CD3	1,2
302212	APC	CD19	1,25
318332	APC-Cy7	CD56	1,25
301324	BV421	CD11b	1,25
307646	BV510	HLA-DR	1,25
304049	BV711	CD45 1:10	1,5
323006	PE	CD15	2,5
301851	PE-Dazzle	CD14 1:10	1
302026	Alexa Fluor 700	CD16 1:10	1,56
		AB total	15,76
		RB	34,24

For further work, calculations of the relative percentage of the cellular composition of the cell subpopulations of interest (namely, $CD4^+$ and $CD8^+$ Temra) of the proximate composition of whole blood cells were also carried out.

3.1.3. Methods: Data analysis

3.1.3.1. Processing of sequencing results and dataset combination

The sequencing results were processed using Bash scripts at the University of Tartu High-Performance Computing (rocket.hpc.ut.ee) (Appendix 3: code). First, the quality of the obtained paired-end reads was checked using FastQC v. 0.11.9 (FastQC Introduction, 2022). Then the adapters and low-quality sequences were removed with TrimGalore v. 0.6.6 (Taking Appropriate QC…, 2022) and Cutadapt v. 3.1 (Martin, 2018) with a Phred score of 35, followed by another quality control (Figure 7).

After that, the reads were aligned in the reference genome GRCh37/hg19 (Karolchik et al., 2012). Site-specific methylation levels were obtained using Bismark v. 0.18.1 Bowtie2 v. 2.3.4.1 (Bismark Bisulfite Mapper-User Guide-v 0.18.1, 2019). Combining the methylation levels and flow cytometry results was done using R (version 3.6.2) and R studio (version 2022.02.0 build 443) with the base R package.

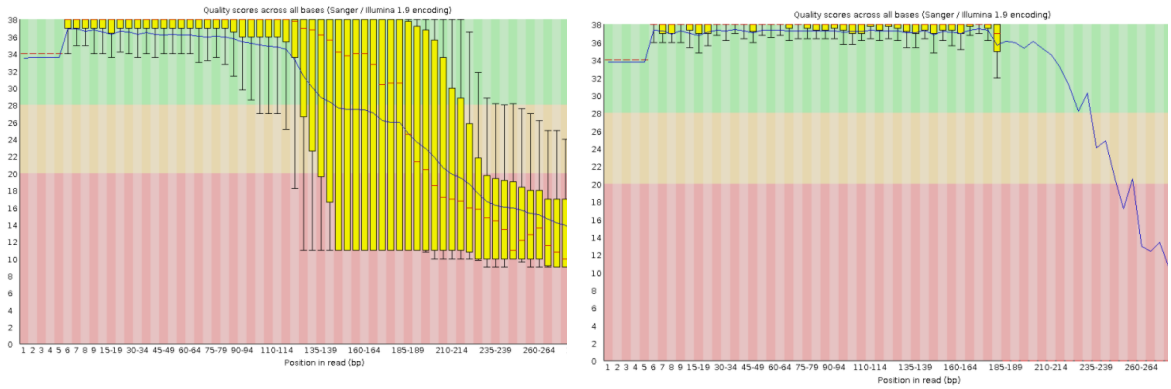


Figure 7. Quality control of the sequencing data with FastQC: before Trimming (left) and after trimming (right). A box plot showing aggregated quality score (Phred score) statistics at each position considering all reads in the file. The y-axis represents the quality indicator, while the x-axis shows the reading position. Colour-coding shows the sequencing quality (considered high (green), medium (orange), and low (red))

3.1.3.2. Dataset cleaning and normalisation

In the dataset, methylation values obtained by less than 300 reads (read depth <300) were filtered out. After that, CpG sites with missing values for many individuals (threshold on >25 % missing) were removed, and imputation of missing methylation values was done using the R package missForest (version 1.4). Next, the distributions of cellular levels were analysed, and 3 transformation methods were applied to better fit the data to the normal distribution: logarithmic with base 10, square root and cube root. After the transformations, the density plots (Figure 8) and the Q-Q plots with the Shapiro-Wilk test (The Shapiro-Wilk…, 2015) were built.

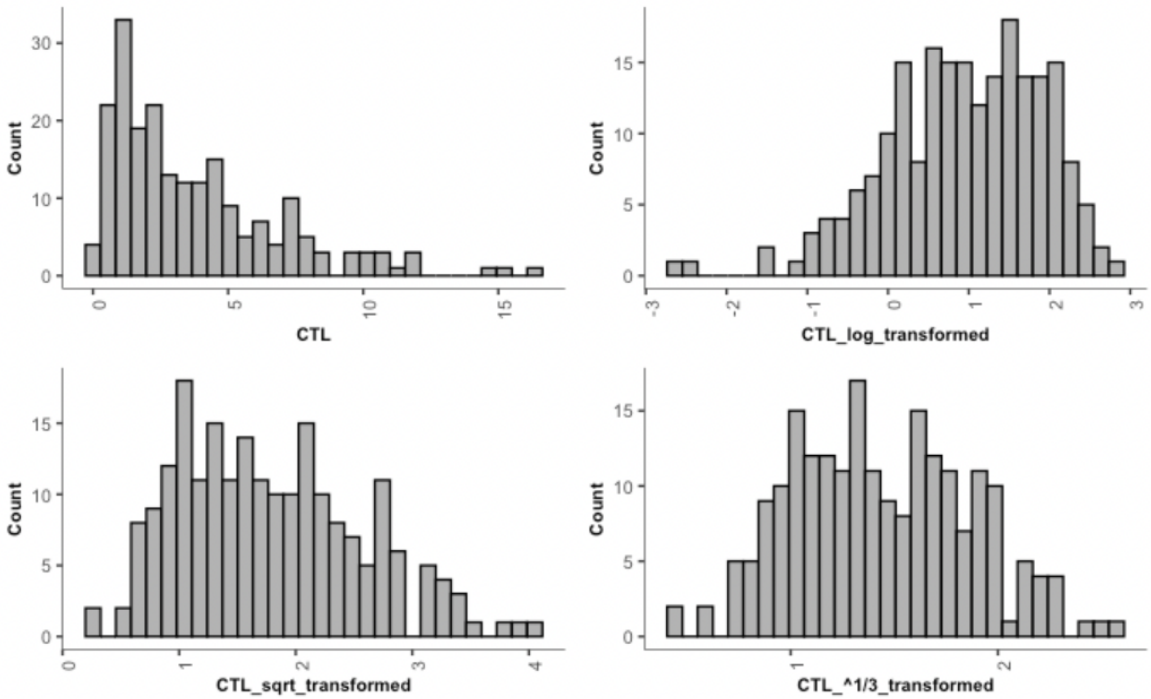


Figure 8. The distribution of $CD8^+$ Temra among whole blood cell content (WBC) in raw (upper left), logarithmic (upper right), square (lower left) and cube (lower right) root transformation

The visual indicators show that the normal distribution appears in the cubic root transformation, which is supported by the quantile-quantile (QQ) plots and the Shapiro-Wilk test (Figure 9). Therefore, for modelling, cube root and original $CD8^+$ and $CD4^+$ Temra were selected.

3.1.3.4. Statistics

Statistical analysis was carried out using R (version 3.6.2) with dplyr (version 1.0.8) and plyr (1.8.7). The main test to perform were the Shapiro p-value, Pearson correlation, p and p adjusted values, and Spearman's Rho criteria with p and p adjusted values (Introduction to Statistics and

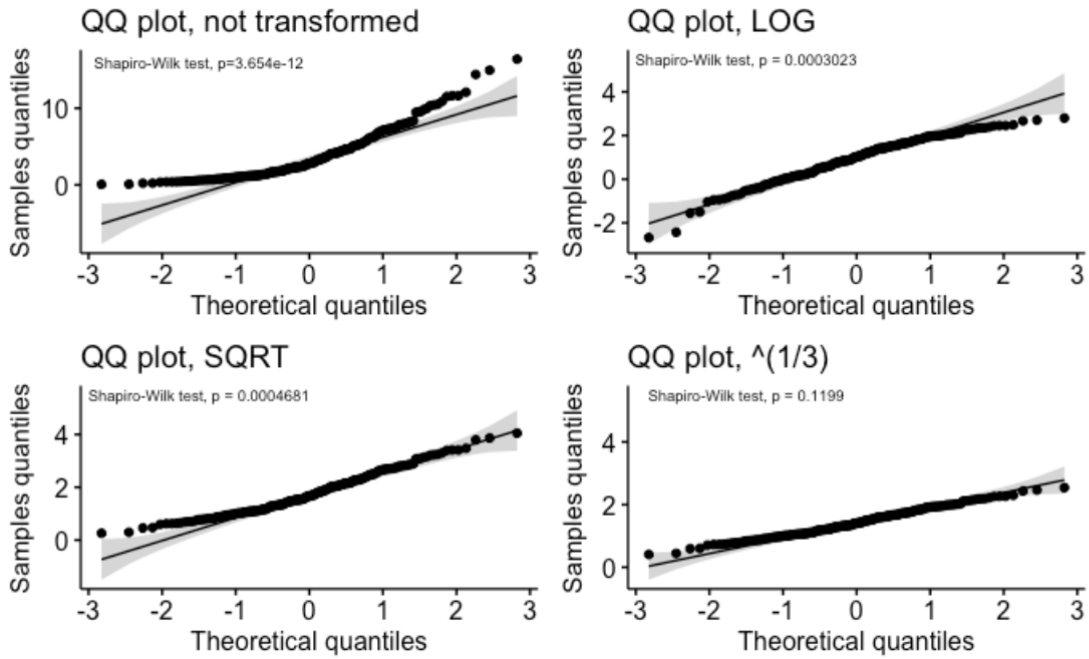


Figure 9. Quantile-quantile plots of CD8+ Temra distribution among whole blood cell content (WBC) in raw (upper left), logarithmic (upper right), square (lower left) and cube (lower right) root transformation. Reported p-values of the Shapiro-Wilk test demonstrate a normal distribution achieved with cubic root transformation: $p > 0.05$ is reached, and the variable can be normally distributed, as shown by the test. Similarly, the distribution of points on QQ plots is as close as possible to the diagonal line or along with it. Judging by these two criteria, the normal distribution was achieved with the cube root transformation

Data Analysis, 2022). These indicators were used to show relationships between the level at a particular methylation site and the relative number of the cell population of interest. Correlational analysis was used to decide which features (CpG sites) should be incorporated to the models.

3.1.3.5. Data analysis and modelling

R studio (version 2022.02.0 build 443) was used for the analysis and modelling. Working with tables was done using the readxl (1.4.0) and writexl (1.4.0) packages. Data pre-processing was performed with dplyr (version 1.0.8) and plyr (1.8.7), visualization with ggplot2 (version 3.3.5), ggpubr (0.4.0) and ComplexHeatmap (2.11.1). Plots were composed using patchwork (version 1.1.1) and tidyverse (version 1.3.1).

As a model for work, a simple linear regression model (1) was chosen:

$$y = \alpha + \beta x \ (1),$$

where y – dependent variable, x – independent variable, $\beta - slope$ and $\alpha -$ intercept (Chaplain & Toland, 2010).

The linear regression model is suitable for prediction in this case because methylation sites (independent variable) and cell proportion (dependent variable) are in a linear relationship since the data was measured from a single tissue type and the CpG site was specific to this particular cell type (demethylated in given cell type and methylated in all others or vice versa). In the first step, data were split into training (0.75) and testing (0.25) sets via the R package Caret (version 6.0-9.1), which was also used for training the models. Next, we built linear regression models for cell populations of interest ($CD8^+$ and $CD4^+$ Temra). Then, the model was evaluated using the root mean squared error (RMSE) and Pearson's correlation coefficient. Finally, all models were tested on the test dataset and visualised their predictions compared to their actual values. The code used in the course of work has been shown in the supplementary section (Appendix 3: code).

3.2. RESULTS

3.2.1. DNA extraction and bisulfite treatment check-up

One of the stages of this work was methylation site detection. For this, whole DNA was isolated from white blood cells with an average concentration of 692.8 ng/µl, and bisulfite treatment of the isolated nucleic acid and its further processing (a detailed description of the work process is presented in section 3.1.2. Methods: Laboratory work) were performed. Agarose gel electrophoresis was used as the central qualitative analysis of the bisulfite converted DNA. To do this, for all of the samples PCR was carried out with 3 randomly selected primer pairs. In the course of the work, it was shown that NA was successfully converted and could be used for methylation analysis (Figure 10).

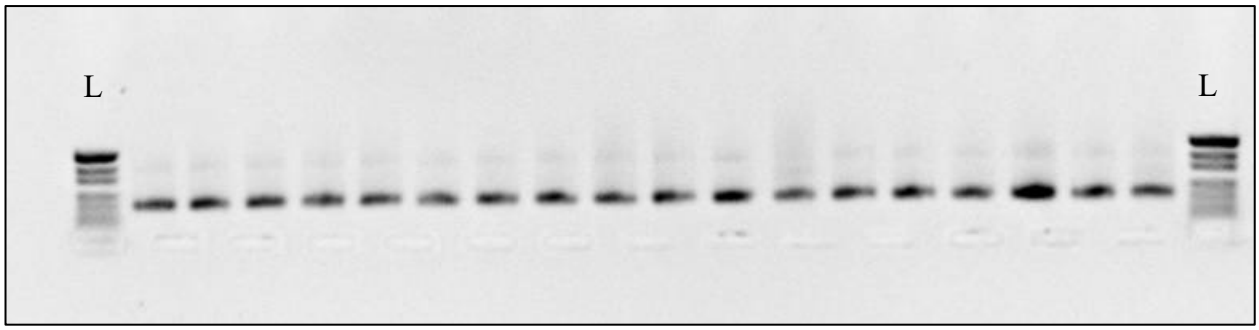


Figure 10. A fragment of control gel electrophoresis in assessing the quality of isolated and bisulfite converted DNA from whole blood cells. Nucleic acid fragments 250 bp are detected

After the test PCR, amplification with 30 primer pairs was performed, followed by purification of the samples with Agencourt AMPure XP beads and their combination in equal amounts. Thereafter, PCR was performed to attach the Nextera XT KIT D indexes for sequencing. Finally, selective quality control of the reaction performed was also carried out at the Core Facility of the Institute of Genomics of the University of Tartu (Figure 11). The fragments after index PCR are longer, so we can assume that amplification was successful and we can combine different samples marked with labelled with indexes to the same sequencing lane.

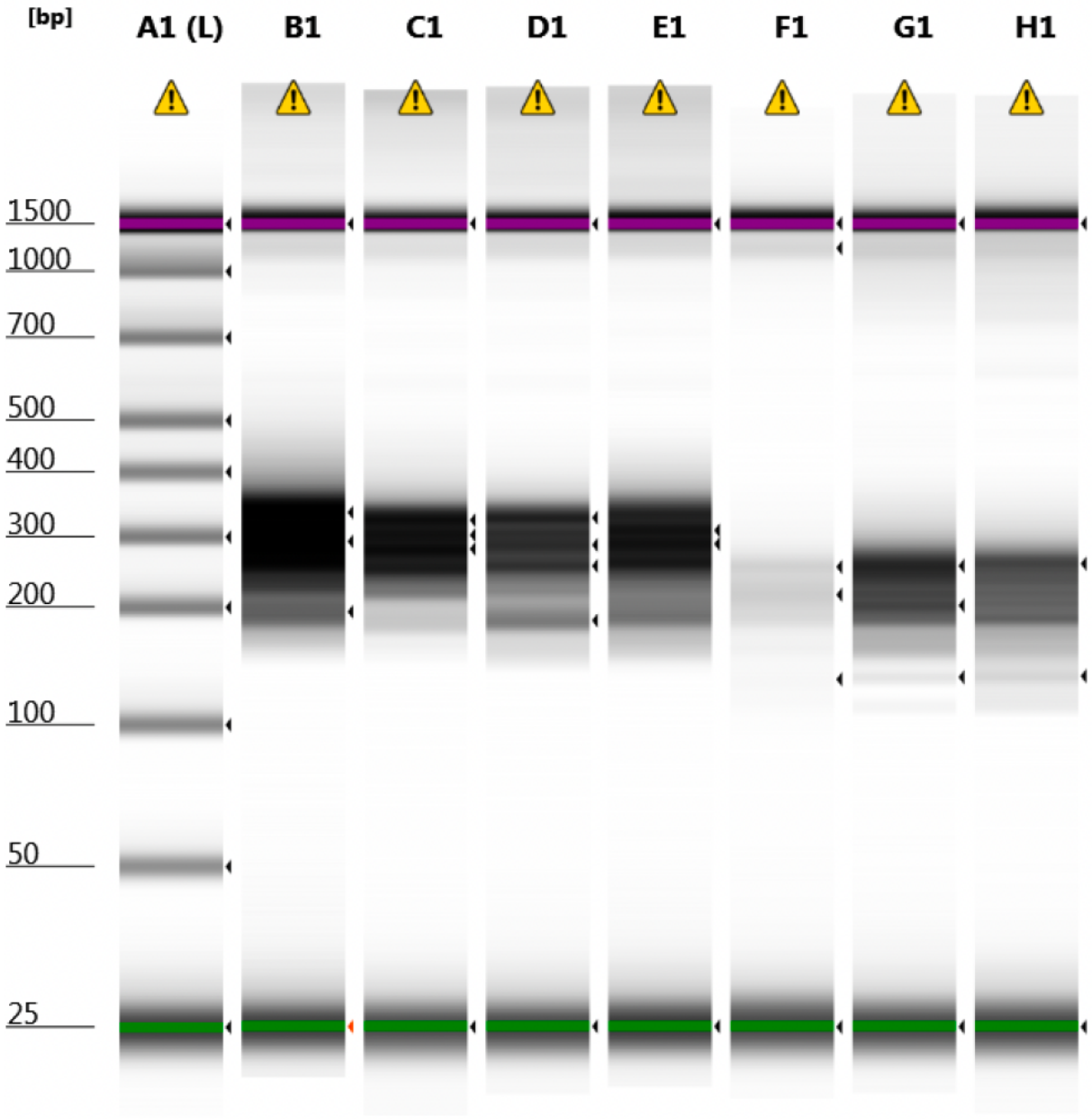


Figure 11. Selective quality control of DNA fragments after index PCR with Illumina indexes. Necessary fragments are visible in the region of 200-300 base pairs. On B1-E1 bands after performing index PCR are shown, while F1-H1 present the same samples before the labelling process

After sequencing, paired reads were obtained for 87 DNA samples. The data was preprocessed using Bash scripts (specified in 3.1.3. Methods: Data analysis) on the University of Tartu High-Performance Computing Server, after which were combined with the previously obtained data of the aged group of individuals.

3.2.2. Temra cells proportions and age

In parallel with obtaining the data on methylation sites, the study of cell populations was carried out with FACS. Flow cytometry allowed us to quantify the levels of cellular populations. During the analysis, the number of cells (events) and their percentage were obtained (Figure S1-S2). Based on these data and blood examination results (the total number of whole blood cells (WBC)), the proportion of cell types of interest in respect to WBC was calculated.

After combining the data produced by the author with data obtained earlier, an age-related increase in Temra cells was evident (Figure 12). At the same time, the distribution of $CD8^+$ Temra cells is more variable among younger individuals, with a tendency to gradually increase with age while the $CD4^+$ Temra proportion increases drastically after 65 years of age (Figure 12).

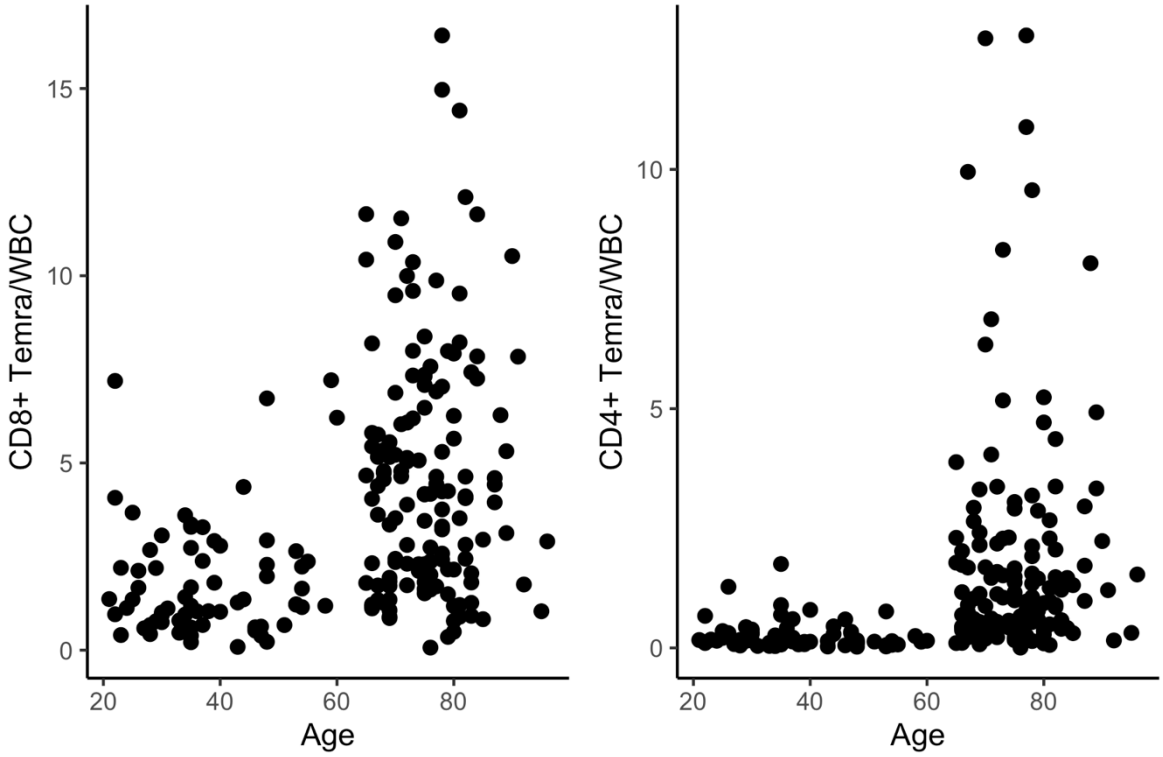


Figure 12. Scatterplots of Temra proportions and age. The y-axis represents the cellular proportions in respect to WBC, and the x-axis corresponds to the donors' age. An age-related increase in the selected cell populations is evident. $CD8^+$ Temra (left) increases gradually with age, while $CD4^+$ Temra (right) number grows rapidly after 65 years of age

The number of $CD8^+$ Temra cells ranges from 0.07 % to 16.4 % with an average of 3.7 % while the corresponding number for $CD4^+$ Temra are 0.07 % – 12.80 % with an average of 1.3 %.

Correlation between cellular proportions and age implies that those proportions could serve as an indicator of health status and immune system ageing.

3.2.3. Site selection for modelling

We matched methylation sites with flow cytometry data to create a model based on epigenetic changes for $CD8^+$ and $CD4^+$ Temra cell populations prediction. Methylation levels (independent variables) and cell levels (dependent variables) were expected to be in a linear relationship. Methylation levels of the selected CpG sites obtained in the course of the work were compared with the proportion of $CD4^+$ and $CD8^+$ Temra among the total number of blood cells. Pearson correlations were calculated between flow cytometry for selected subpopulations and DNA methylation (Figure S3).

The highest level of correlation (r=|0.6|) was achieved in the analysis of $CD8^+$·Temra cells with CpG site at chromosome 12. At the same time, a lower correlation was observed from the same sites with a population of $CD4^+$ Temra cells, varying in the range of r= |0.2| – |0.4|. Due to a large number of initial sites for sequencing and methylation level assessment, to avoid overfitting of the model, sites were selected. For this, the levels of correlation were calculated for each selected cell population (Figure S3). Methylation sites were evaluated, after which 7 sites with high correlation were selected (chr12.4915855.4915855, chr1.240164755.240164755, chr17.79921715.79921715, chr17.79921776.79921776, chr2:87012808, chr2:87012817, chr2:87020937). 3 out of 7 CpG sites were located near genes expressed in T cells: CD8A genes (chr2:87012808, chr2:87012817, chr2:87020937) and had a high correlation, however, they were excluded since those variables were not statistically significant in multiple regression. The remaining sites were used for further work (Figure 13).

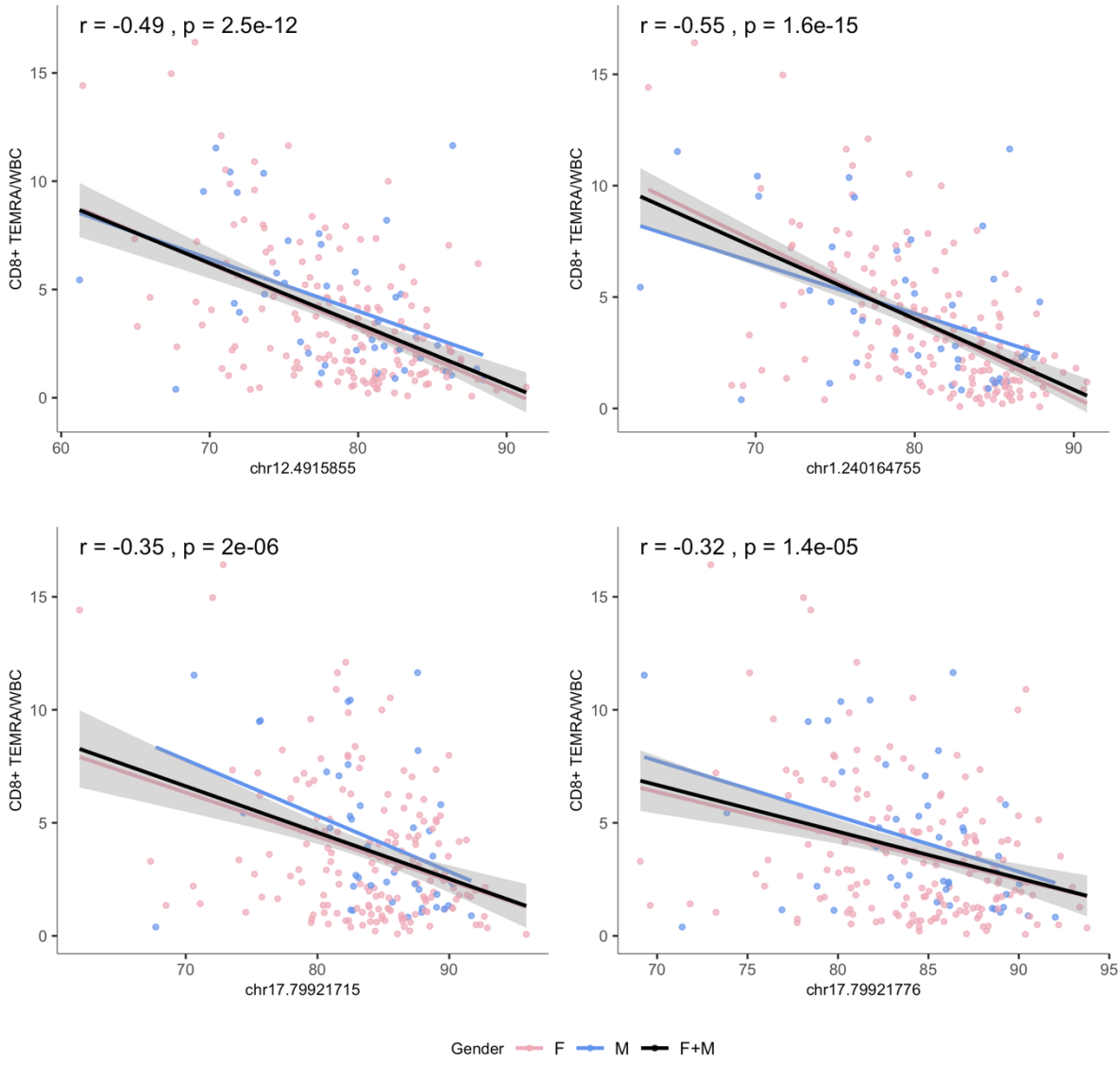


Figure 13. Selected sites for model creation for predicting $CD8^+$ Temra cells. The graph shows the dependence of the cellular proportions on the level of methylation in selected CpG sites. Y-axis represents the cell's proportion in respect to WBC, while the x-axis shows the methylation level. Colouring denotes gender with a pink indicating female and blue male

To create a model for determining the $CD4^+$ Temra cell population levels, the top 5 correlations were selected, 4 of which were located on chromosome 12 (chr1.240164755.240164755, chr12.4915938.4915938, chr12.4915925.4915925, chr12.4915855.4915855, chr12.24689930.24689930). Figure 14 indicates that those CpG sites have a moderate correlation with CD4+ Temra.

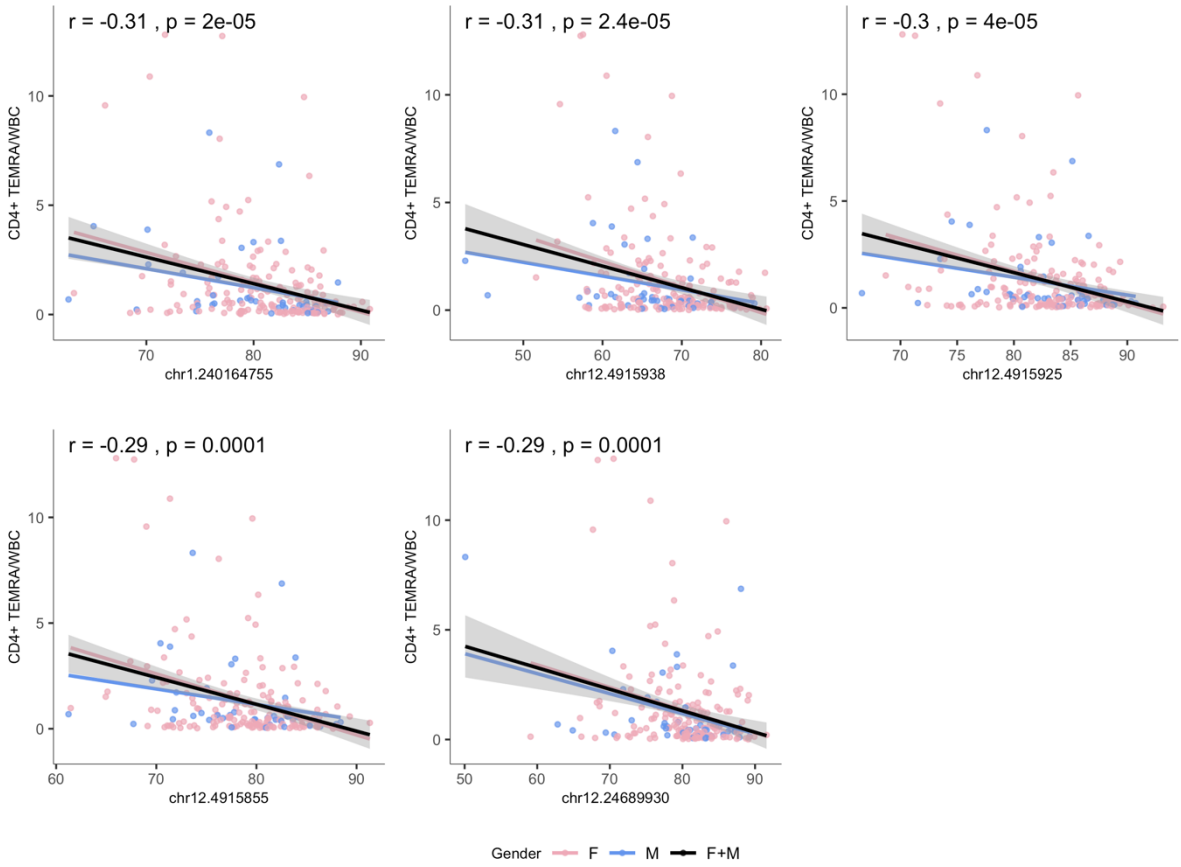


Figure 14. Selected sites for model creation in predicting $CD4^+$ Temra cells. The graph shows the dependence of the cell population number on the level of methylation. The y-axis represents cellular proportion in respect to WBC, while the x-axis shows the methylation level. Gender distribution is denoted with pink (female) and blue (male) colours

3.2.4. Models' performance

Multiple linear regression was used to build the models. All models were trained on a training set (75% of the data) and then tested on a test set (25%). For both cell types, two models were built using either raw values of the dependent variable or transformed values (cube root). The model built to predict $CD8^+$ Temra using raw dependent variable had a correlation and RMSE on test data between actual and predicted values of 0.63 and 2.16 respectively, while the same measurements for cube root were 0.48 and 2.82 (Figure 15). The standard deviation of the residuals (prediction errors) or RMSE parameter indicates an average deviation of ~2 – 3 % between the predicted and actual values, which is quite acceptable for given data. However, there is definitely room to improve the model for example by using different and more sophisticated methods for feature selection and also different types of modelling algorithms.

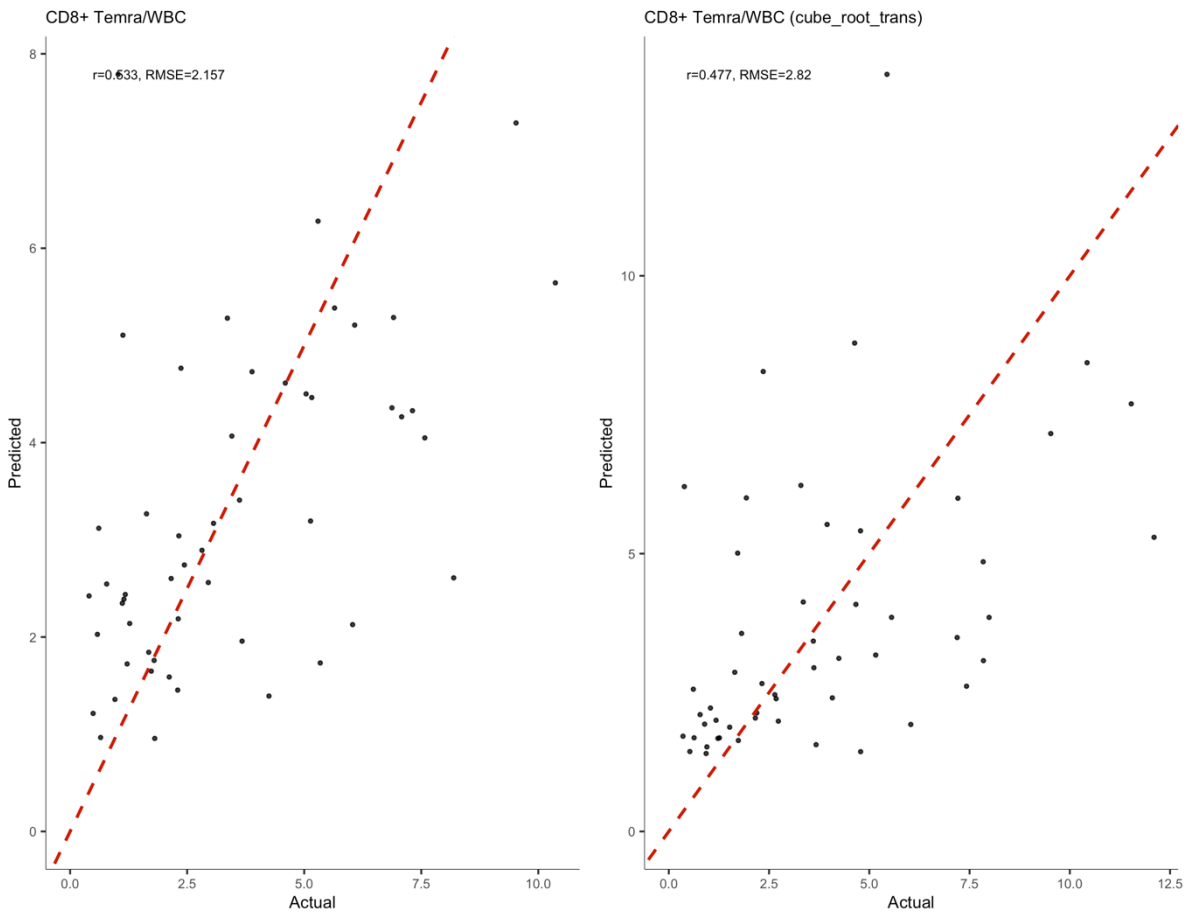


Figure 15. Prediction accuracy of $CD8^+$ Temra/WBC models using raw dependent variable (left) and cube root transformed dependent variable (right). The y-axis shows the predicted values, while the x-axis reflects the actual values. Also, r and RSME measurements are shown together with a diagonal line (red) indicating the ideal model

Similarly, models for $CD4^+$ Temra cells were built, however, they showed lower performance on test data (Figure 16). It implies that our methylation data did not contain suitable CpG sites to predict $CD4^+$ Temra/WBC and a hence different set of CpG sites needs to be studied in this regard. The correlation coefficient (r) and RMSE were 0.38 and 2.44, respectively, for the model built using an untransformed dependent variable and 0.179 and 1.54 for the cube root transformation.

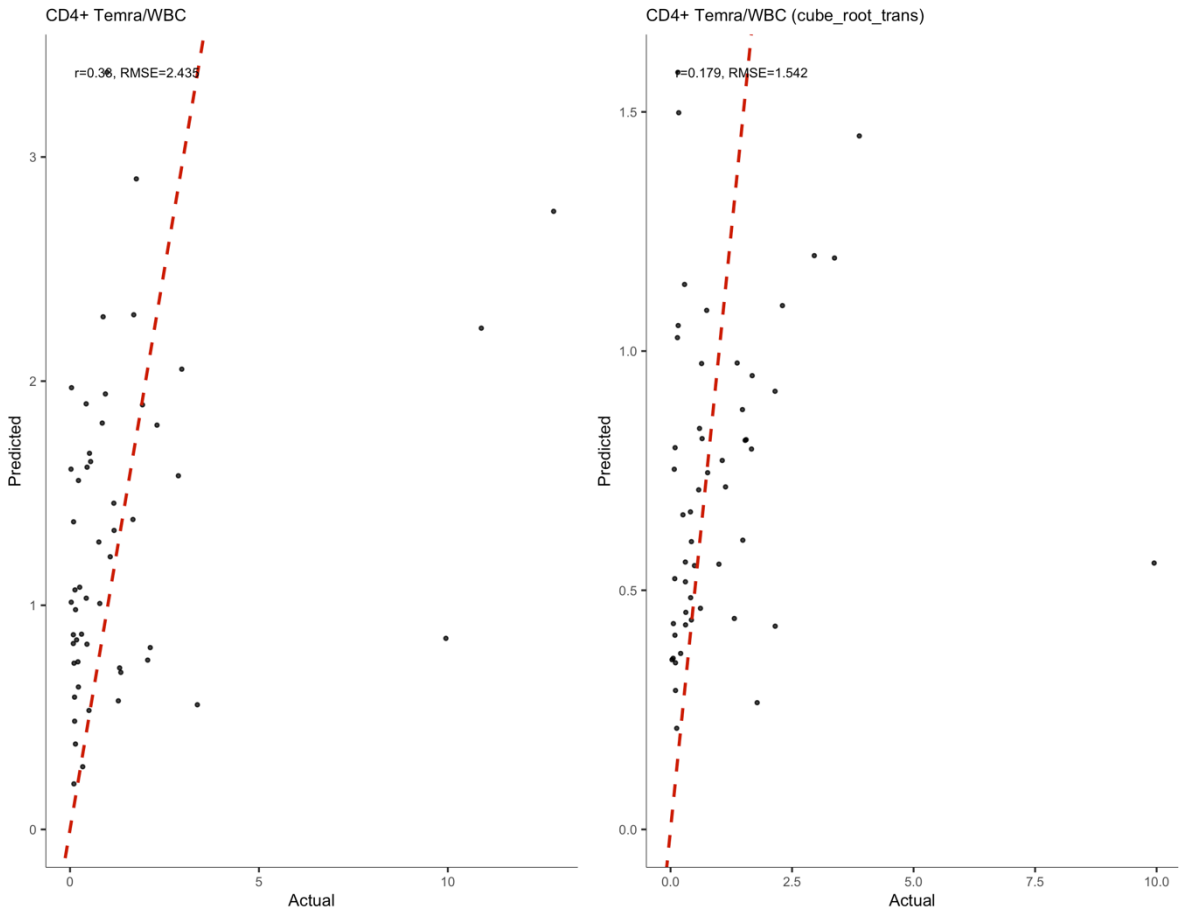


Figure 16. Prediction accuracy of $CD4^+$ Temra models based on r raw dependent variable (left) and cube root transformed dependent variable (right). The y-axis shows the predicted values, while the x-axis reflects the actual values. Also, r and RSME measurements are shown together with a diagonal line (red) indicating ideal model

3.3. DISCUSSION

Since this work includes time-consuming and labour-intensive methods (DNA extraction and flow cytometry), other options for laboratory methods may be considered. Our method of DNA extraction is based on the salting-out process. On its own, it is easy to use and has many advantages compared with, for example, organic isolation methods, where many toxic components (like phenols) are included (Mccormick, 1989). However, it involves several steps of precipitating DNA with alcohol for purification. This option could be replaced using an ion exchange resin such as Chelix 100. This method is in a single reaction tube with different steps and reagents with easy to perform steps and it requires less biomaterial (P. Sean Walsh, 1991). Modifications are also known that include additional purification steps in one column (Seligson D. et al, 1990). Thus, a column containing a resin with positively charged diethylaminoethylcellulose groups is used to bind negatively charged phosphates of the DNA backbone. Contaminants such as protein and RNA can be washed out of the DNA column using medium salt buffers. In the analysis of methylation sites, we used the bisulfite-treated sequence. This method is the gold standard in the detection of epigenetic modifications (Li & Tollefsbol, 2011). The other aspect is replacing flow cytometry in our work. Known techniques aim more on the determination of the levels of main lymphocyte populations in the blood (for example, determination of the levels of $CD3^+$, $CD8^+$, or $CD4^+$) and do not focus on smaller subsets as is the case here (Franke et al., 1994; Lanier, 1981).

In the course of the work, it was confirmed that the accumulation of $CD8^+$ Temra cells associated with high cytotoxicity, low proliferation and sensitivity to apoptosis increases with age (Tomiyama et al., 2002; Yang et al., 2019; Salumets et al., 2022). Results have been validated showing a gradual increase in the levels of $CD8^+$ Temra cells with age, as well as a rapid increase of the $CD4^+$ Temra cell population after 65 years of age. These results are consistent with other studies, however, in some of them, a gradual increase in the size of the cell subpopulation was noticed earlier (after 50 years) (Goronzy & Weyand, 2017; Salumets et al., 2022).

Epigenetic assessment of cell levels using a linear regression model is a viable option for predicting $CD8^+$ populations but requires several improvements. To determine $CD4^+$ Temra levels with DNA methylation, it is necessary to perform a more in-depth study of different CpG sites to see whether suitable sites exist. In the case of $CD8^+$ Temra prediction, it requires more sophisticated methods for feature selection and likely different modelling algorithms should be

tested as well such as Random Forest-based regressor and other types of linear regression algorithms like ridge and lasso. In a similar study, a $CD8^+$ Temra/WBC model was built using ridge regression on older population with model's performance metrics being following: r = 0,887, and RMSE = 2,2 (Salumets et al., 2022). Latter indicates that fine-tuning modelling and feature selection part could result in an improved model.

In contrast to cytometry, the estimation of cellular levels based on epigenetic changes (namely, DNA methylation) is affordable and simpler. Given that many immune cell types such as Temra cells are associated with adverse health outcomes and disturbances in the immune system, such predictions could also potentially give a hint on an individual's immune status. As the broad future perspective of the work analysis to evaluate the usefulness of the model could be performed. This would include finding out whether the predicted values tend to be changed with people inflicted by certain diseases and finding correlations between predicted values and other health-related biological measurements. This work is the first step toward a disease prediction tool that can not only simplify the process of diagnosis but can also be used for prognostic purposes, which will help to intervene earlier in the occurrence of the disease making it less destructive. Therefore, such a model can be used as a potential biomarker.

SUMMARY

Being one of the periods of ontogenetic development, ageing is an inevitable element of age-related development. This process is based on the internal instability of biological molecules, which is the main cause of molecular disturbances. Epigenetic changes, in particular methylation, are no exception and affect all orgasmic systems, including the immune system.
In this work, the possibility of DNA methylation of specific CpG sites, to quantify the level of effector memory re-expressing CD45RA ($CD8^+$ and $CD4^+$ Temra) cells was studied. Thus, DNA was obtained from whole blood samples from healthy donors and was subjected to bisulfite treatment to assess DNA methylation levels. Subsequently, the levels of cellular populations were measured using flow cytometry and then the data were combined. Finally, the models for predicting the proportions of $CD4^+$ and $CD8^+$ Temra cells in respect to whole blood cells were built and their performance was assessed.

REFERENCES

1. *Agencourt ® ampure ® pcr purification*. (2022). www.fishersci.com
2. Alarcón, A., Figueroa, U., Espinoza, B., Sandoval, A., Carrasco-Aviño, G., Aguayo, F. R., & Corvalan, A. H. (2017). Epstein-Barr Virus–Associated Gastric Carcinoma: The Americas' Perspective. In *Gastric Cancer*. InTech. https://doi.org/10.5772/intechopen.70201
3. Barh, D., Chaitankar, V., Yiannakopoulou, E. C., Salawu, E. O., Chowbina, S., Ghosh, P., & Azevedo, V. (2013). In Silico Models: From Simple Networks to Complex Diseases. In *Animal Biotechnology: Models in Discovery and Translation* (pp. 385–404). Elsevier Inc. https://doi.org/10.1016/B978-0-12-416002-6.00021-3
4. Barres, R., & Zierath, J. R. (2011). DNA methylation in metabolic disorders. *American Journal of Clinical Nutrition*, *93*(4). https://doi.org/10.3945/ajcn.110.001933
5. Bell, C. G., Lowe, R., Adams, P. D., Baccarelli, A. A., Beck, S., Bell, J. T., Christensen, B. C., Gladyshev, V. N., Heijmans, B. T., Horvath, S., Ideker, T., Issa, J. P. J., Kelsey, K. T., Marioni, R. E., Reik, W., Relton, C. L., Schalkwyk, L. C., Teschendorff, A. E., Wagner, W., … Rakyan, V. K. (2019). DNA methylation aging clocks: Challenges and recommendations. In *Genome Biology* (Vol. 20, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-019-1824-y
6. Bergsma, T., & Rogaeva, E. (2020). DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. In *Neuroscience Insights* (Vol. 15). SAGE Publications Ltd. https://doi.org/10.1177/2633105520942221
7. Bergstedt, J., Urrutia, A., Duffy, D., Albert, M. L., Quintana-Murci, L., & Patin, E. (2018). *Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles*. https://doi.org/10.1101/456996
8. *Bismark Bisulfite Mapper-User Guide-v0.15.0*. (2016). http://bowtie-bio.sourceforge.net/bowtie2
9. Campbell, J. P., & Narendran, P. (2019). *Type 1 diabetes impairs the mobilisation of highly-differentiated CD8+T cells during a single bout of acute exercise*. https://www.researchgate.net/publication/331981166
10. Chaplain, M. A. J., & Toland, J. F. (2010). *Regression: Linear Models in Statistics (Springer Undergraduate Mathematics Series)*. www.springer.com/series/3423

11. Chen, H., Zheng, X., & Zheng, Y. (2014). Age-associated loss of lamin-b leads to systemic inflammation and gut hyperplasia. *Cell*, *159*(4), 829–843. https://doi.org/10.1016/j.cell.2014.10.028
12. Chen, S. H., Lv, Q. L., Hu, L., Peng, M. J., Wang, G. H., & Sun, B. (2017). DNA methylation alterations in the pathogenesis of lupus. In *Clinical and Experimental Immunology* (Vol. 187, Issue 2, pp. 185–192). Blackwell Publishing Ltd. https://doi.org/10.1111/cei.12877
13. Chomyk, A. M., Volsko, C., Tripathi, A., Deckard, S. A., Trapp, B. D., Fox, R. J., & Dutta, R. (2017). DNA methylation in demyelinated multiple sclerosis hippocampus. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-08623-5
14. Chung, S. A., Nititham, J., Elboudwarej, E., Quach, H. L., Taylor, K. E., Barcellos, L. F., & Criswell, L. A. (2015). Genome-wide assessment of differential DNA methylation associated with autoantibody production in systemic lupus erythematosus. *PLoS ONE*, *10*(7). https://doi.org/10.1371/journal.pone.0129813
15. Colin A. Michie, A. M. C. A. P. C. L. B. (1992). Lifespan of human Lymphocyte subsets defined by CD45 isoforms. *Nature*.
16. Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N. R., Liopyris, K., Marchetti, M. A., Podlipnik, S., Puig, S., Rinner, C., Tschandl, P., Weber, J., Halpern, A., & Malvehy, J. (2022). Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*, *4*(5), e330–e339. https://doi.org/10.1016/S2589-7500(22)00021-8
17. Corpet, A., & Stucki, M. (2014). Chromatin maintenance and dynamics in senescence: a spotlight on SAHF formation and the epigenome of senescent cells. In *Chromosoma* (Vol. 123, Issue 5, pp. 423–436). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00412-014-0469-6
18. da Costa, J. P., Vitorino, R., Silva, G. M., Vogel, C., Duarte, A. C., & Rocha-Santos, T. (2016). A synopsis on aging—Theories, mechanisms and future prospects. In *Ageing Research Reviews* (Vol. 29, pp. 90–112). Elsevier Ireland Ltd. https://doi.org/10.1016/j.arr.2016.06.005

19. Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A. H., Kirkpatrick, C. L., Wollheim, C. B., Eliasson, L., Rönn, T., Bacos, K., & Ling, C. (2014). Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-Diabetic Donors Identifies Candidate Genes That Influence Insulin Secretion. *PLoS Genetics*, *10*(3). https://doi.org/10.1371/journal.pgen.1004160
20. el Gazzar, M., Yoza, B. K., Chen, X., Hu, J., Hawkins, G. A., & McCall, C. E. (2008). G9a and HP1 couple histone and DNA methylation to TNFα transcription silencing during endotoxin tolerance. *Journal of Biological Chemistry*, *283*(47), 32198–32208. https://doi.org/10.1074/jbc.M803446200
21. *EZ DNA Methylation™ Kit Streamlined bisulfite conversion of DNA Highlights*. (2022). www.zymoresearch.com
22. *FastQC introduction*. (2022). http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/
23. Federica Sallusto, D. L. R. F. M. L. A. L. (1999). *Two subsets of memoryT lymphocytes withdistinct homing potentialsand effector functions*.
24. Fedintsev, A., & Moskalev, A. (2020). Stochastic non-enzymatic modification of long-lived macromolecules - A missing hallmark of aging. In *Ageing Research Reviews* (Vol. 62). Elsevier Ireland Ltd. https://doi.org/10.1016/j.arr.2020.101097
25. Fedintsev, A., & Moskalev, A. (2020). Stochastic non-enzymatic modification of long-lived macromolecules - A missing hallmark of aging. In *Ageing Research Reviews* (Vol. 62). Elsevier Ireland Ltd. https://doi.org/10.1016/j.arr.2020.101097
26. Fisher Scientific, T. (n.d.). *User Guide: Qubit dsDNA HS Assay Kits*.
27. Fogel, O., Richard-Miceli, C., & Tost, J. (2017). Epigenetic Changes in Chronic Inflammatory Diseases. In *Advances in Protein Chemistry and Structural Biology* (Vol. 106, pp. 139–189). Academic Press Inc. https://doi.org/10.1016/bs.apcsb.2016.09.003
28. Franke, L., Nugel, E., Docke, W.-D., & Porstmann', T. (1994). *Quantitative Determination of CD4/CD8 Molecules by a Cell Marker ELISA* (Vol. 40, Issue 1).
29. Fransquet, P. D., Wrigglesworth, J., Woods, R. L., Ernst, M. E., & Ryan, J. (2019). The epigenetic clock as a predictor of disease and mortality risk: A systematic review and meta-analysis. In *Clinical Epigenetics* (Vol. 11, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13148-019-0656-7
30. Fulop, T., Franceschi, C., Hirokawa, K., & Pawelec, G. (2009). *Handbook on Immunosenescence Basic Understanding and Clinical Applications*.

31. Goronzy, J. J., & Weyand, C. M. (2017). Successful and Maladaptive T Cell Aging. In *Immunity* (Vol. 46, Issue 3, pp. 364–378). Cell Press. https://doi.org/10.1016/j.immuni.2017.03.010
32. Goswami, R., & Awasthi, A. (2020). Editorial: T Cell Differentiation and Function in Tissue Inflammation. In *Frontiers in Immunology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2020.00289
33. Guarente, L. (2014). Aging research - Where do we stand and where are we going? In *Cell* (Vol. 159, Issue 1, pp. 15–19). Cell Press. https://doi.org/10.1016/j.cell.2014.08.041
34. Gujar, H., Weisenberger, D. J., & Liang, G. (2019). The roles of human DNA methyltransferases and their isoforms in shaping the epigenome. In *Genes* (Vol. 10, Issue 2). MDPI AG. https://doi.org/10.3390/genes10020172
35. Gupta, S., Agrawal, A., Agrawal, S., Su, H., & Gollapudi, S. (2006). A paradox of immunodeficiency and inflammation in human aging: Lessons learned from apoptosis. In *Immunity and Ageing* (Vol. 3). https://doi.org/10.1186/1742-4933-3-5
36. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S. V., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, *49*(2), 359–367. https://doi.org/10.1016/j.molcel.2012.10.016
37. Horvath, H., & Horvath, S. (2013). DNA methylation age of human tissues and cell types. In *Genome Biology* (Vol. 14). http://genomebiology.com//14/10/R115
38. Horvath, S., Gurven, M., Levine, M. E., Trumble, B. C., Kaplan, H., Allayee, H., Ritz, B. R., Chen, B., Lu, A. T., Rickabaugh, T. M., Jamieson, B. D., Sun, D., Li, S., Chen, W., Quintana-Murci, L., Fagny, M., Kobor, M. S., Tsao, P. S., Reiner, A. P., … Assimes, T. L. (2016). An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1030-0
39. Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., & Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-1140-4
40. Huang, Y., & Rao, A. (2014). Connections between TET proteins and aberrant DNA modification in cancer. In *Trends in Genetics* (Vol. 30, Issue 10, pp. 464–474). Elsevier Ltd. https://doi.org/10.1016/j.tig.2014.07.005
41. Introduction to Statistics and Data Analysis. (2022).

42. Jaul, E., & Barron, J. (2017). Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population. In *Frontiers in Public Health* (Vol. 5). Frontiers Media S.A. https://doi.org/10.3389/fpubh.2017.00335
43. Jin, Z., & Liu, Y. (2018). DNA methylation in human diseases. In *Genes and Diseases* (Vol. 5, Issue 1, pp. 1–8). Chongqing yi ke da xue, di 2 lin chuang xue yuan Bing du xing gan yan yan jiu suo. https://doi.org/10.1016/j.gendis.2018.01.002
44. Jochems, S. P., Jacquelin, B., Tchitchek, N., Busato, F., Pichon, F., Huot, N., Liu, Y., Ploquin, M. J., Roché, E., Cheynier, R., Dereuddre-Bosquet, N., Stahl-Henning, C., le Grand, R., Tost, J., & Müller-Trutwin, M. (2020). DNA methylation changes in metabolic and immune-regulatory pathways in blood and lymph node CD4 + T cells in response to SIV infections. *Clinical Epigenetics*, *12*(1). https://doi.org/10.1186/s13148-020-00971-w
45. Karolchik, D., Hinrichs, A. S., & James Kent, W. (2012). The UCSC genome browser. *Current Protocols in Bioinformatics*, *SUPPL.40*. https://doi.org/10.1002/0471250953.bi0104s40
46. Kirkwood, T. B. L. (2005). Understanding the odd science of aging. In *Cell* (Vol. 120, Issue 4, pp. 437–447). Elsevier B.V. https://doi.org/10.1016/j.cell.2005.01.027
47. Korovila, I., Hugo, M., Castro, J. P., Weber, D., Höhn, A., Grune, T., & Jung, T. (2017). Proteostasis, oxidative stress and aging. In *Redox Biology* (Vol. 13, pp. 550–567). Elsevier B.V. https://doi.org/10.1016/j.redox.2017.07.008
48. Kruk, P. A., Rampino, N. J., & Bohr, V. A. (1995). DNA damage and repair in telomeres: Relation to aging (human telomeres). In *Biochemistry* (Vol. 92). https://www.pnas.org
49. Lanier l.l., w. N. L. (1981). Paraformaldehyde fixation of hematopoietic cells for quantitative flow cytometry (FACS) analysis. *Journal Oflmmunological Methods*, *47*, 25–30.
50. Li, Y., & Tollefsbol, T. O. (2011). DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in Molecular Biology*, *791*, 11–21. https://doi.org/10.1007/978-1-61779-316-5_2
51. Libri, V., Azevedo, R. I., Jackson, S. E., di Mitri, D., Lachmann, R., Fuhrmann, S., Vukmanovic-Stejic, M., Yong, K., Battistini, L., Kern, F., Soares, M. v.d., & Akbar, A. N. (2011). Cytomegalovirus infection induces the accumulation of short-lived, multifunctional CD4+CD45RA+CD27- T cells: The potential involvement of interleukin-7 in this process. *Immunology*, *132*(3), 326–339. https://doi.org/10.1111/j.1365-2567.2010.03386.x

52. Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T. J., & Feinberg, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, *31*(2), 142–147. https://doi.org/10.1038/nbt.2487
53. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. In *Cell* (Vol. 153, Issue 6, p. 1194). Elsevier B.V. https://doi.org/10.1016/j.cell.2013.05.039
54. Lu, H., Liu, X., Deng, Y., & Qing, H. (2013). DNA methylation, a hand behind neurodegenerative diseases. In *Frontiers in Aging Neuroscience* (Vol. 5, Issue DEC). Frontiers Media SA. https://doi.org/10.3389/fnagi.2013.00085
55. Maecker, H. T., McCoy, J. P., & Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. In *Nature Reviews Immunology* (Vol. 12, Issue 3, pp. 191–200). https://doi.org/10.1038/nri3158
56. Makar, K. W., & Wilson, C. B. (2004). DNA Methylation Is a Nonredundant Repressor of the Th2 Effector Program. *The Journal of Immunology*, *173*(7), 4402–4406. https://doi.org/10.4049/jimmunol.173.7.4402
57. Marchuck G. (1985). *Modelling in Immunology*. http://library.keldysh.ru/mvk.asp?id=1989-5
58. Martin, M. (2018). *cutadapt Documentation Release 1.16*.
59. Mccormick, R. M. (1989). A So-lid-Phase Extraction Procedure for DNA Purification. In *ANALYTICALBIOCHEMISTRY* (Vol. 181).
60. Moskalev, A. A., Proshkina, E. N., Belyi, A. A., & Solovyev, I. A. (2016). Genetics of aging and longevity. *Vavilov Journal of Genetics and Breeding*, *20*(4), 426–440. https://doi.org/10.18699/vj16.171
61. P. Sean Walsh, D. A. M. and R. H. (19911). Chelex 100 as a Medium for Simple Extraction of DNA for PCR-Based Typing from Forensic Material. *BioTechniques* , *10*(4), 506–513.
62. Pawelec, G., & Gupta, S. (2019). Editorial: Immunology of aging. In *Frontiers in Immunology* (Vol. 10, Issue JULY). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2019.01614

63. Pellegrini, C., Pirazzini, C., Sala, C., Sambati, L., Yusipov, I., Kalyakulina, A., Ravaioli, F., Kwiatkowska, K. M., Durso, D. F., Ivanchenko, M., Monti, D., Lodi, R., Franceschi, C., Cortelli, P., Garagnani, P., & Bacalini, M. G. (2021). A Meta-Analysis of Brain DNA Methylation Across Sex, Age, and Alzheimer's Disease Points for Accelerated Epigenetic Aging in Neurodegeneration. In *Frontiers in Aging Neuroscience* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fnagi.2021.639428
64. Riznichenko G. (2010). Mathematical Modelling in Immunology.
65. Robertson, K. D., Ait-Si-Ali, S., Yokochi, T., Wade, P. A., Jones, P. L., & Wolffe, A. P. (2000). *338 nature genetics • volume 25 • july 2000 DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters*. http://genetics.nature.com
66. Salameh, Y., Bejaoui, Y., & el Hajj, N. (2020). DNA Methylation Biomarkers in Aging and Age-Related Diseases. In *Frontiers in Genetics* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fgene.2020.00171
67. Salumets, A., Tserel, L., Rumm, A. P., Türk, L., Kingo, K., Saks, K., Oras, A., Uibo, R., Tamm, R., Peterson, H., Kisand, K., & Peterson, P. (2022a). Epigenetic quantification of immunosenescent CD8 $^{+}$ TEMRA cells in human blood. *Aging Cell*. https://doi.org/10.1111/acel.13607
68. Saul, D., & Kosinsky, R. L. (2021). Epigenetics of aging and aging-associated diseases. In *International Journal of Molecular Sciences* (Vol. 22, Issue 1, pp. 1–25). MDPI AG. https://doi.org/10.3390/ijms22010401
69. Seligson D. et al. (1990). *Method of isolating and purifying nuclec acds from biological samples*.
70. Sharples, A. P., Seaborne, R. A., & Stewart, C. E. (2018). Epigenetics of Skeletal Muscle Aging. In *Epigenetics of Aging and Longevity* (pp. 389–416). Elsevier. https://doi.org/10.1016/b978-0-12-811060-7.00019-x
71. Stammers, M., Ivanova, I. M., Niewczas, I. S., Segonds-Pichon, A., Streeter, M., Spiegel, D. A., & Clark, J. (2020). Age-related changes in the physical properties, crosslinking, and glycation of collagen from mouse tail tendon. *Journal of Biological Chemistry*, *295*(31), 10562–10571. https://doi.org/10.1074/jbc.RA119.011031
72. Taking appropriate QC measures for RRBS-type or other-Seq applications with Trim Galore! (2022).
73. *The shapiro-wilk and related tests for normality*. (2015).

74. Tian, Y., Babor, M., Lane, J., Schulten, V., Patil, V. S., Seumois, G., Rosales, S. L., Fu, Z., Picarda, G., Burel, J., Zapardiel-Gonzalo, J., Tennekoon, R. N., de Silva, A. D., Premawansa, S., Premawansa, G., Wijewickrama, A., Greenbaum, J. A., Vijayanand, P., Weiskopf, D., … Peters, B. (2017). Unique phenotypes and clonal expansions of human CD4 effector memory T cells re-expressing CD45RA. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01728-5
75. Tomiyama, H., Matsuda, T., & Takiguchi, M. (2002). Differentiation of Human CD8 + T Cells from a Memory to Memory/Effector Phenotype. *The Journal of Immunology*, *168*(11), 5538–5550. https://doi.org/10.4049/jimmunol.168.11.5538
76. Tomiyama, H., Matsuda, T., & Takiguchi, M. (2002). Differentiation of Human CD8 + T Cells from a Memory to Memory/Effector Phenotype .*The Journal of Immunology*, *168*(11), 5538–5550. https://doi.org/10.4049/jimmunol.168.11.5538
77. Tower, J. (2015). Programmed cell death in aging. In *Ageing Research Reviews* (Vol. 23, Issue PA, pp. 90–100). Elsevier Ireland Ltd. https://doi.org/10.1016/j.arr.2015.04.002
78. Tsurumi, A., & Li, W. X. (2012). Global heterochromatin loss: A unifying theory of aging? *Epigenetics*, *7*(7), 680–688. https://doi.org/10.4161/epi.20540
79. Várady, G., Cserepes, J., Németh, A., Szabó, E., & Sarkadi, B. (2013). Cell surface membrane proteins as personalized biomarkers: Where we stand and where we are headed. In *Biomarkers in Medicine* (Vol. 7, Issue 5, pp. 803–819). https://doi.org/10.2217/bmm.13.90
80. Varriale, A. (2014). DNA Methylation, Epigenetics, and Evolution in Vertebrates: Facts and Challenges. International Journal of Evolutionary Biology, 2014, 1–7. https://doi.org/10.1155/2014/475981
81. Wajed, S. A., Laird, P. W., & Demeester, T. R. (2001). *DNA Methylation: An Alternative Pathway to Cancer*.
82. Weiskopf, D., Bangs, D. J., Sidney, J., Kolla, R. v., de Silva, A. D., de Silva, A. M., Crotty, S., Peters, B., & Sette, A. (2015). Dengue virus infection elicits highly polarized CX3CR1+ cytotoxic CD4+ T cells associated with protective immunity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(31), E4256–E4263. https://doi.org/10.1073/pnas.1505956112
83. Weyand, C. M., & Goronzy, J. J. (2016). Aging of the immune system: Mechanisms and therapeutic targets. *Annals of the American Thoracic Society*, *13*, S422–S428. https://doi.org/10.1513/AnnalsATS.201602-095AW

84. Yang, D.-H., Ahn, J.-S., Ahn, S.-Y., Jung, S.-H., Lee, J.-J., Kim, H.-J., Do, Y. R., Song, S. Y., Song, G.-Y., Kim, H. J., & Kim, T. (2019). Genomic Profiles and Subset Characterization of CD8+terminally Differentiated Effector Memory (TEMRA) Cells from Cancer Patients. *Blood*, *134*(Supplement_1), 2329–2329. https://doi.org/10.1182/blood-2019-126533
85. Yang, D.-H., Ahn, J.-S., Ahn, S.-Y., Jung, S.-H., Lee, J.-J., Kim, H.-J., Do, Y. R., Song, S. Y., Song, G.-Y., Kim, H. J., & Kim, T. (2019). Genomic Profiles and Subset Characterization of CD8+terminally Differentiated Effector Memory (TEMRA) Cells from Cancer Patients. *Blood*, *134*(Supplement_1), 2329–2329. https://doi.org/10.1182/blood-2019-126533
86. Yousefzadeh, M., Henpita, C., Vyas, R., Soto-Palma, C., Robbins, P., & Niedernhofer, L. (2021). Dna damage—how and why we age? *ELife*, *10*, 1–17. https://doi.org/10.7554/eLife.62852
87. Zhang, J., Rane, G., Dai, X., Shanmugam, M. K., Arfuso, F., Samy, R. P., Lai, M. K. P., Kappei, D., Kumar, A. P., & Sethi, G. (2016). Ageing and the telomere connection: An intimate relationship with inflammation. In *Ageing Research Reviews* (Vol. 25, pp. 55–69). Elsevier Ireland Ltd. https://doi.org/10.1016/j.arr.2015.11.006
88. Zhang, R., Chen, H. Z., & Liu, D. P. (2015). The Four Layers of Aging. In *Cell Systems* (Vol. 1, Issue 3, pp. 180–186). Cell Press. https://doi.org/10.1016/j.cels.2015.09.00

APPENDIX 1: Tables

Table S1. List of primers used for the work. The chromosomal position and sequence of the oligonucleotide are presented

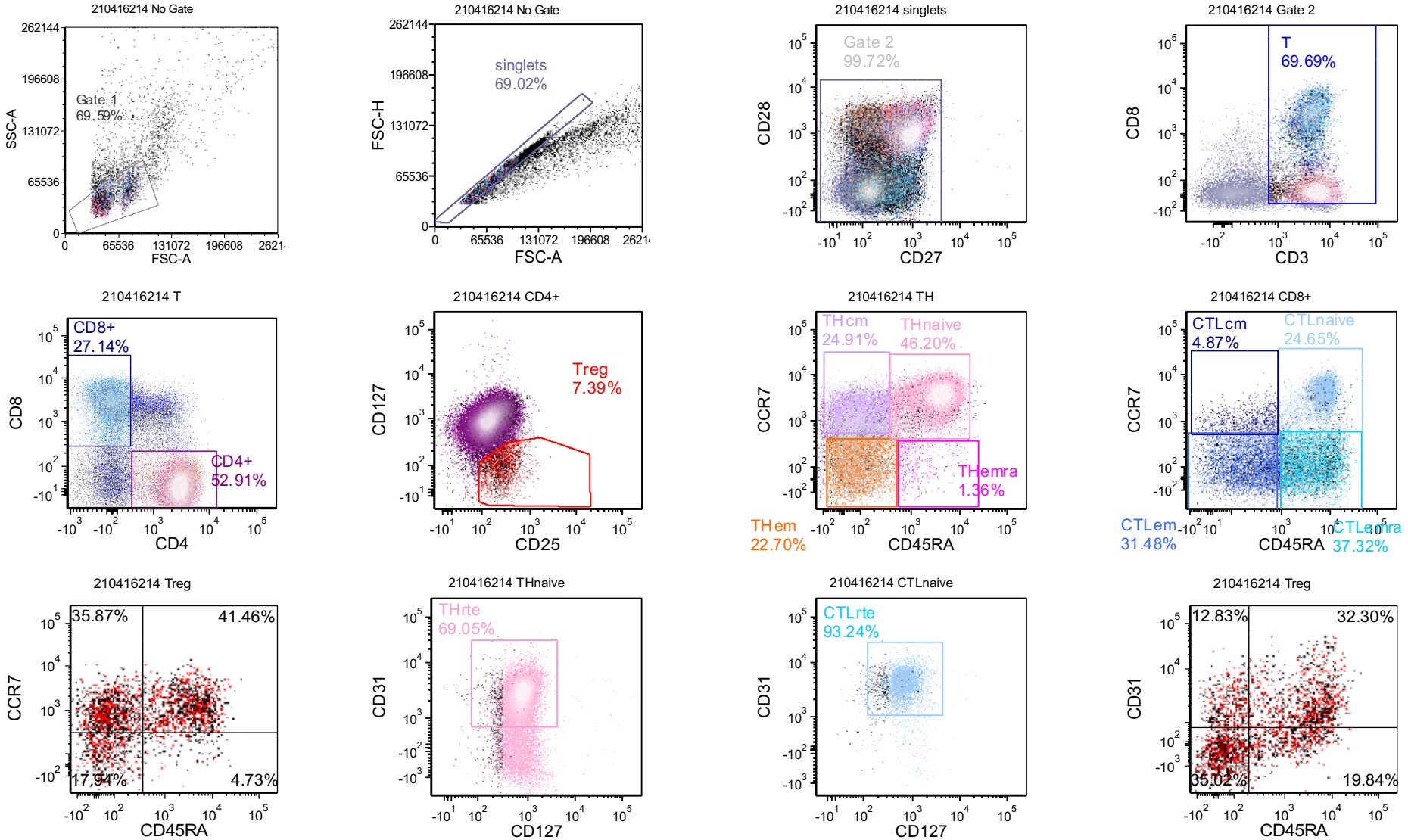
Primer Name	Sequence
cg00219921 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGATGGTATGGGTGTTTTTAG
cg00219921 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCACAAAAATCACAATACTATTATA
cg25939861 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGATGATGGTTAGATTTGGGG
cg25939861 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCACCCCACTAACTAACACT
cg02150910 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGGGTTAGAGTATAAGATGGT
cg02150910 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTACTAAAACAAATCCAACCTC
cg23663547 F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGAGTTTAGAGTGTGGTTTTGA
cg23663547 R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACACACTCATCCCCTCACTT
cg26215982 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTTGTTTGAAGGGTGGT
cg26215982 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCAAAACCAATCCTAACAAAACA
cg01940810 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATAAATATCAAATTTTTCAATAACAAC
cg01940810 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTTATTTTTTGGAGAATAAATGTTTTG
cg04467549 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGTGATAATAGTGTTTTTGTTAGA
cg04467549 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTCAACATACAAACTCCAT
cg06567722 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGAGGTTTTTGTAGGTGAT
cg06567722 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAAACAACCTACTACTACTACA
cg13669740 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAAATATAACCTCTTCTACCACT
cg13669740 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTAGTTTGTGTATTTGTTGAT
cg02051545 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGATTTAGATATTAAGTGGTTGT
cg02051545 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAATATCTTAAAACACTAATATCTCC
cg19884600 F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGTTTGGGTAAAAGATAAGATAATGA
cg19884600 R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATATAAATTCAATATTTAACTTCCACT

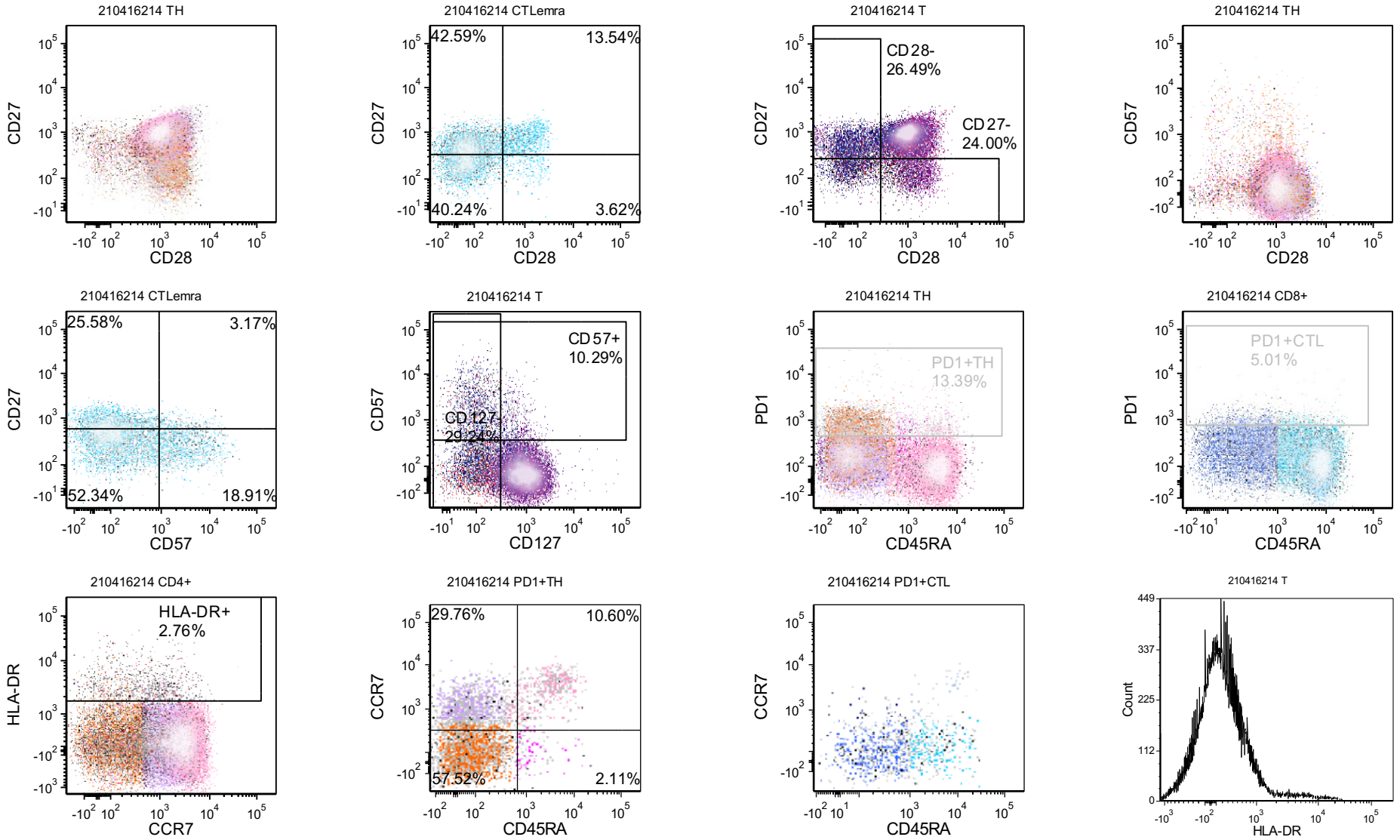
Primer Name	Sequence
cg00112929 F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGAAGAGTTAAGGAGAATAGGGT
cg00112929 R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTTCCCACACCCACCACTA
cg23364656 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTAGGGTTAGAGGTTTATTTGT
cg23364656 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCTACTAACCTATTCATCAACACC
cg24612198 F	AGAAGTAGTAAGTTTGTTGGT
cg24612198 R	ACTCCATCCTACTCACCTAAT
cg24841244 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGGGTAGTTAGTATTAGGTTAG
cg24841244 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACTCTCACCCAAACTAATAA
cg07545925 F	GAGTTTTTGAGTGGGAATTTAGTA
cg07545925 R	TCCACCCTCTACTACAAATAT
cg15880738 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTGGTTGGTTGGTTGGTTGT
cg15880738 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACAACCAAACCCTTCCCCT
cg06147361 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTAGAAGAGATAGTTATTATGTTA
cg06147361 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCCAAAAATATAAATTCAATCATATC
cg10315334 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAATACCCCTCAACTAACCCT
cg10315334 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGTATTTGTGGGAGAGGTTGT
cg02867514 F	GGTATTTGTGGGAGAGGTTGTG
cg02867514 R	AATACCCCTCAACTAACCCT
cg22235901 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTTATGGTGGAGGGTTTGT
cg22235901 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCCAACCACTAAAACCTAA
cg03318654 F	TGTGAGTAGGATAGGATTTAGGAGG
cg03318654 R	ACCTAATTTCCTTCCAACCA
cg13486641 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGGAGTAATGGTATTTTGGGA
cg13486641 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATTTCTAACCTCAAACTACTATAA
cg20832020 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGGTAGAAGAGGTTATATTTGT
cg20832020 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCTCAACCCCTAAACCCAA
cg22496559 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTTCAATACAAACCTATAAAC
cg22496559 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTTAGGGTTGGATTTGAA

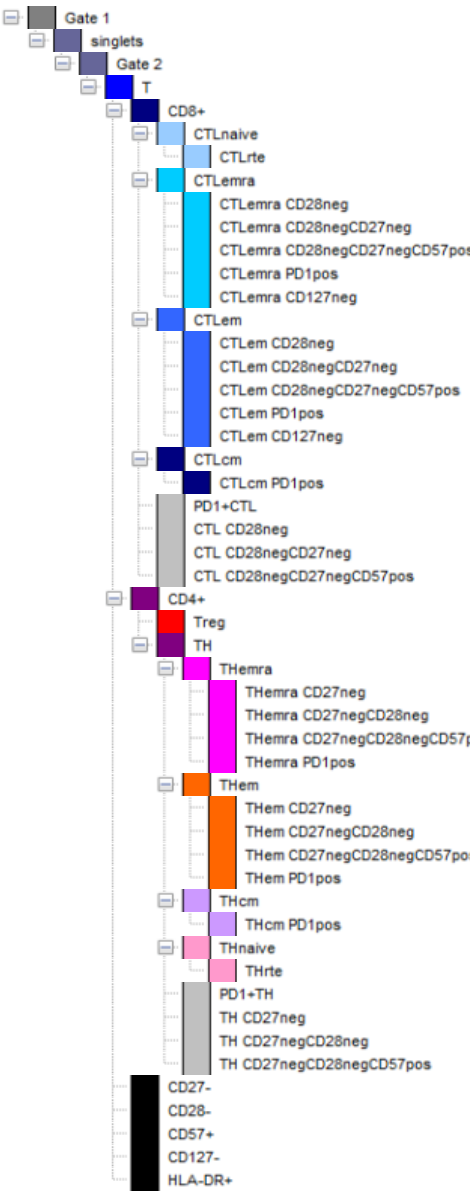
Primer Name	Sequence
cg08455089 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGGTAGAGGAAGAGGTGGT
cg08455089 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCACAATCCTTATAAAATATCCAA
cg05221370 F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGTTTAGAGAAGAAAAGTAATTTGT
cg05221370 R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AAAACAATATTACCTCATACACAAAA
cg03069731 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGAGTGGGAATTGTGTTTGA
cg03069731 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTATCCTAAACTCAACCCAAAA
cg00087425 F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATAGGATTGGAGATGTTTGAGG
cg00087425 R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCTCCCTCACACAAAACT
cg20063728 Illumina F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGTTTAGGATTAGTAGTTTAAAGT
cg20063728 Illumina R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTTACCTAACACCTTCTATAAAATT

APPENDIX 2: Figures

Figure S1. Identification of cell populations using flow cytometry. Removing dead cells and doublets. T cell populations are listed in the FACS table description

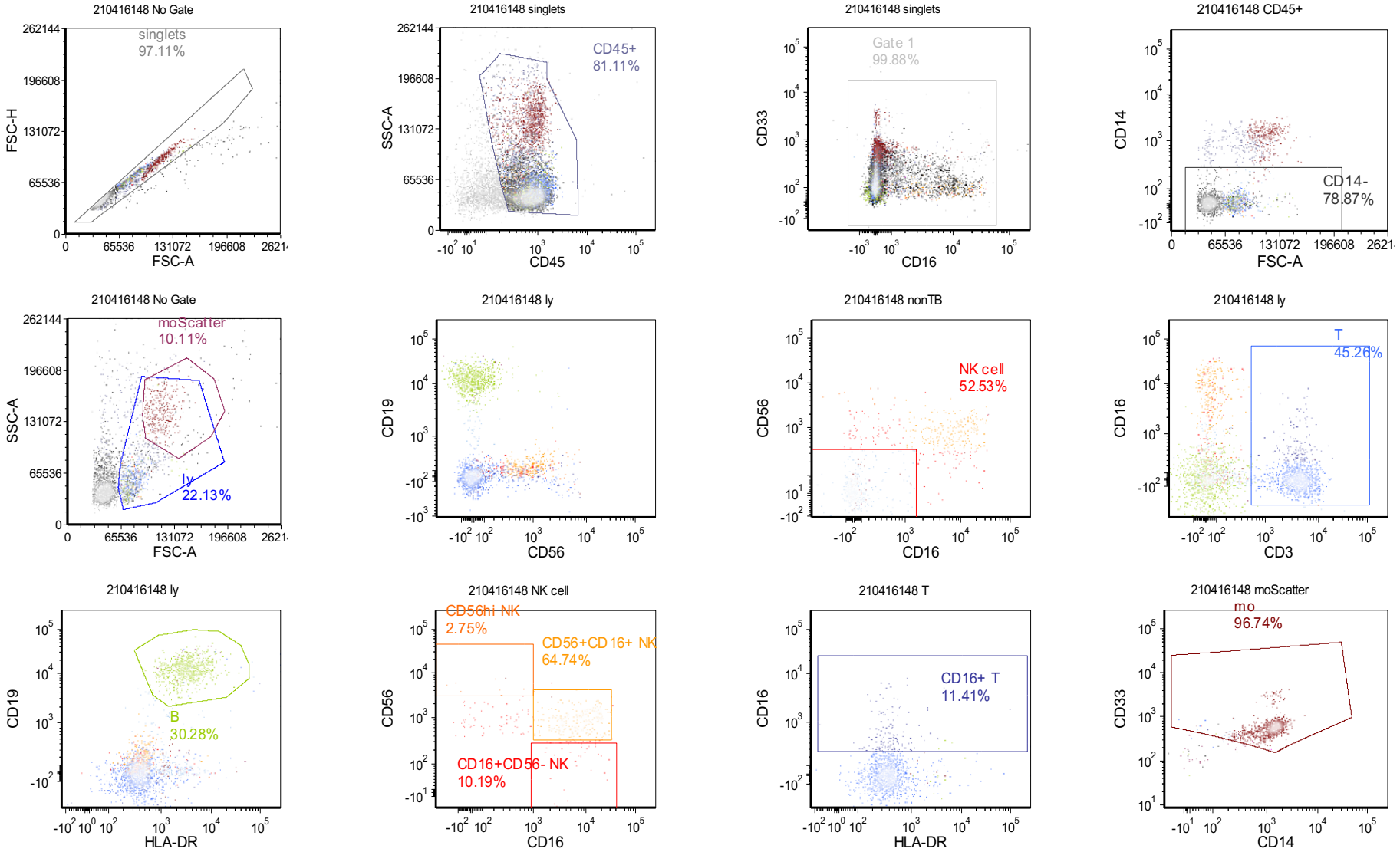


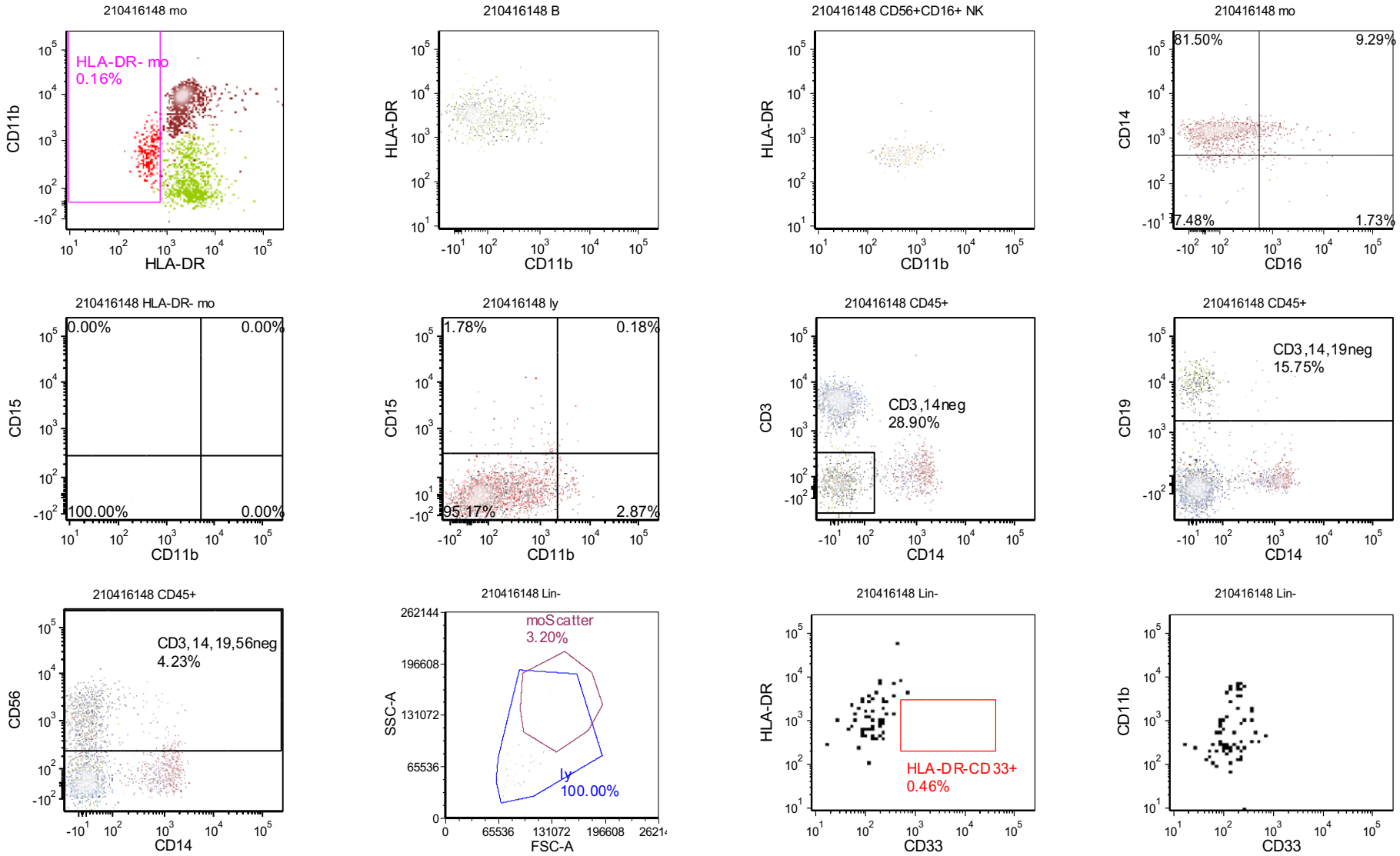

210416214 TH
CD27
CD28
210416214 CTLemra
42.59%
13.54%
40.24%
3.62%
CD27
CD28
210416214 T
CD28-
26.49%
CD27-
24.00%
CD27
CD28
210416214 TH
CD57
CD28
210416214 CTLemra
25.58%
3.17%
52.34%
18.91%
CD27
CD57
210416214 T
CD57+
10.29%
CD127-
29.24%
CD57
CD127
210416214 TH
PD1+TH
13.39%
PD1
CD45RA
210416214 CD8+
PD1+CTL
5.01%
PD1
CD45RA
210416214 CD4+
HLA-DR+
2.76%
HLA-DR
CCR7
210416214 PD1+TH
29.76%
10.60%
57.52%
2.11%
CCR7
CD45RA
210416214 PD1+CTL
CCR7
CD45RA
210416214 T
449
337
225
112
0
Count
HLA-DR

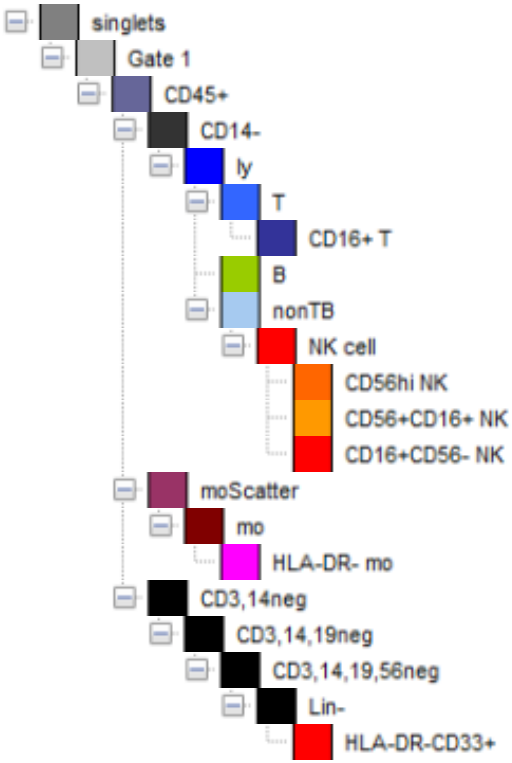


210416214	events	%	HLA-DR MFI
T cells	75161	69.69	
CD4+	39765	52.91	
Treg	2938	7.39	727.26
Treg n	1218	41.46	
Treg rte	949	32.30	
Treg cm	1054	35.87	
Treg em	527	17.94	
TH	36827	92.61	
THnaive	17014	46.20	204.82
THrte / TH	11749	29.55	
THcm	9173	24.91	220.52
THem	8360	22.70	309.21
THemra	502	1.36	735.69
THem CD27neg /TH	3300	8.96	
THem CD27,28neg /TH	244	0.66	
THem CD27,28negCD57pos /TH	45	0.12	
THem PD1pos /TH	2940	7.98	
THemra CD27neg /TH	227	0.62	
THemra CD27,28neg /TH	106	0.29	
THemra CD27,28negCD57pos /TH	21	0.06	
THemra PD1pos /TH	129	0.35	
THcm PD1pos /TH	949	2.58	
CTL	75161	27.14	
CTLnaive	5029	24.65	354.59
CTLrte / TH	4689	22.98	
CTLcm	994	4.87	299.93
CTLem	6422	31.48	402.09
CTLemra	7614	37.32	237.31
CTLem CD28neg /CTL	2461	12.06	
CTLem CD28,CD27neg /CTL	643	3.15	
CTLem CD28,CD27negCD57pos /CTL	240	1.18	
CTLem PD1pos /CTL	591	2.90	
CTLem CD127neg	2719	13.33	
CTLemra CD28neg /CTL	6170	30.24	
CTLemra CD28,CD27neg /CTL	2544	12.47	
CTLemra CD28,CD27negCD57pos /CTL	1250	6.13	
CTLemra PD1pos /CTL	309	1.51	
CTLemra CD127neg	5240	25.68	
CTLcm PD1pos	42	0.21	

Figure S2. Identification of cell populations using flow cytometry. Removing dead cells and doublets. MDSC populations are listed in the FACS table description




210416148 mo
HLA-DR- mo
0.16%
CD11b
HLA-DR
210416148 B
HLA-DR
CD11b
210416148 CD56+CD16+ NK
HLA-DR
CD11b
210416148 mo
81.50%
9.29%
7.48%
1.73%
CD14
CD16
210416148 HLA-DR- mo
0.00%
0.00%
100.00%
0.00%
CD15
CD11b
210416148 ly
1.78%
0.18%
95.17%
2.87%
CD15
CD11b
210416148 CD45+
CD3,14neg
28.90%
CD3
CD14
210416148 CD45+
CD3,14,19neg
15.75%
CD19
CD14
210416148 CD45+
CD3,14,19,56neg
4.23%
CD56
CD14
210416148 Lin-
moScatter
3.20%
ly
100.00%
SSC-A
FSC-A
210416148 Lin-
HLA-DR-CD33+
0.46%
HLA-DR
CD33
210416148 Lin-
CD11b
CD33



210416148	
T /ly	45.26
CD16+ T /T	11.41
B /ly	30.28
NK /ly	13.20
CD56+CD16+ /NK	64.74
CD56hiNK /NK	2.75
CD16+NK /NK	10.19
CD14+CD16- mo	81.50
CD14+CD16+ mo	9.29
CD14-CD16+ mo	1.73
HLA-DR- mo /mo	0.16
CD11b+CD15+ MDSC /MDSC	0.00
CD11b+CD15- MDSC /MDSC	0.00
CD11b-CD15- MDSC /MDSC	100.00
e-MDSC /CD45+	0.01

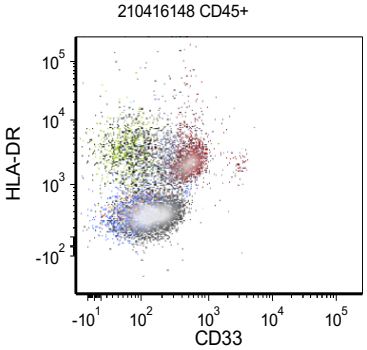
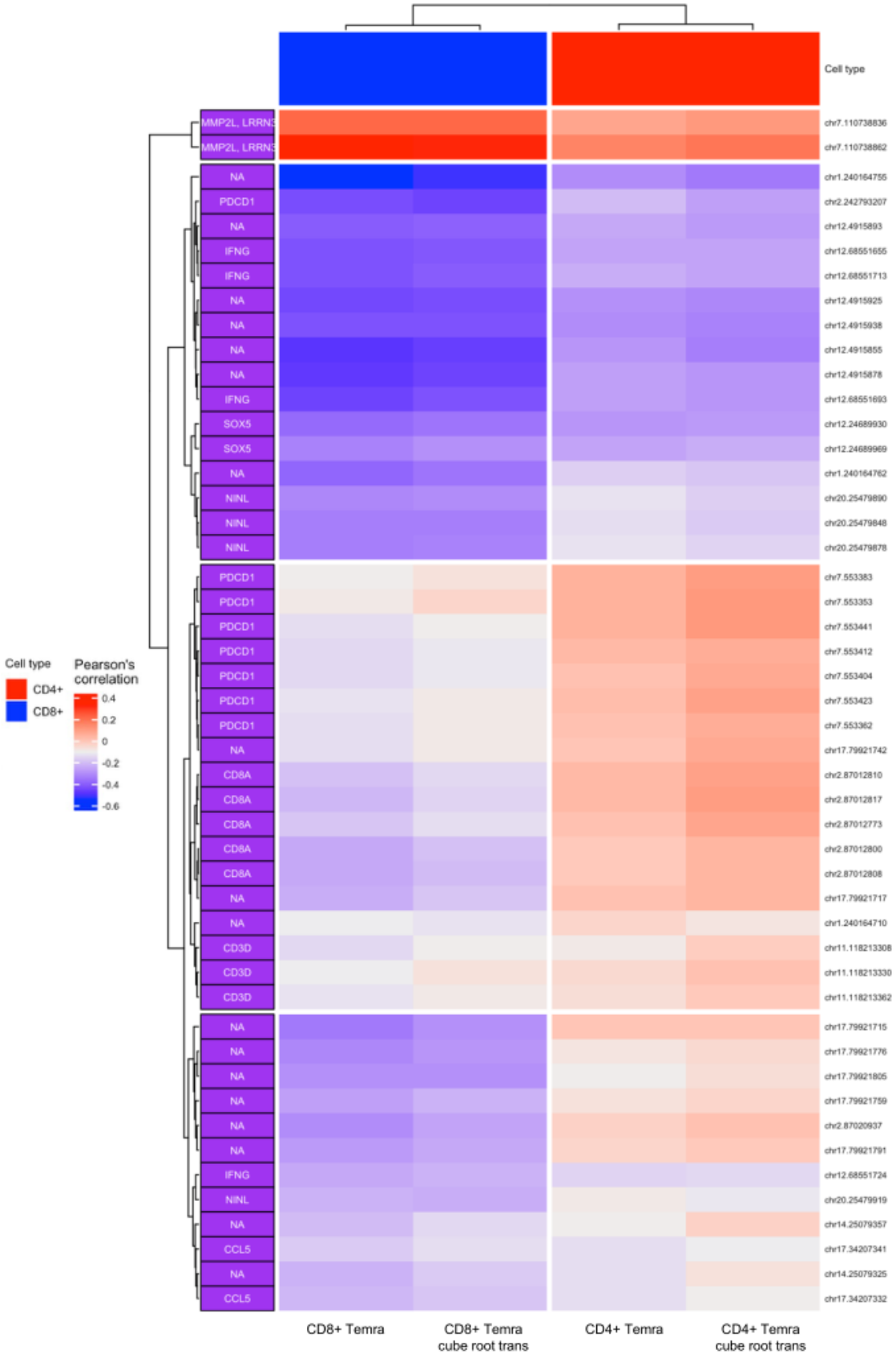


Figure S3. Cluster analysis of selected cell populations. Correlations between $CD8^+$ and $CD4^+$ Temra and their transformations with methylation sites are presented. Also, the genes in which CpG is located shown are. NA - the site is located outside the gene



APPENDIX 3: CODE

1. HPC scrips

1. trim_galore

module load any-cutadapt/3.1
module load trim_galore

```
sbatch trim_galore_script.sh #to use the file created with editor/
#!/bin/bash
#SBATCH --job-name=trim_galore
#SBATCH --nodes=1
#SBATCH --time=4:00:00
#SBATCH --mem=40GB

quality=35
dir=/gpfs/space/home/elsakova/seq/

file_parts=$( ls $dir | grep -Po ".*(?=R1_001.fastq)" )
for file_part in ${file_parts[@]}
        do
                trim_galore --paired --quality $quality "$dir${file_part}R1_001.fastq"
                "$dir${file_part}R2_001.fastq"
        done
```

2. Fastqc
module load fastqc

```
fastqc *.fq
```

3. bismark
module load bismark

```
# Bismark II
 sbatch bismark_step2_2.sh
#!/bin/bash
#SBATCH --job-name=bisulfit_2
#SBATCH --nodes=1
#SBATCH --time=1-05:00:00
#SBATCH --mem=80GB
#SBATCH --workdir=/gpfs/rocket/home/elsakova/seq/

file_parts=$( ls $dir | grep -Po ".*(?=R1_001_val_1\.fq)" )

for file_part in ${file_parts[@]}
        do
                gpfs/rocket/home/b15184/BSAS/bismark_v0.18.1/bismark -n 1 -X 1000 --bowtie2
        /gpfs/rocket/home/elsakova/genome/ -1 "${file_part}R1_001_val_1.fq" -2
        "${file_part}R2_001_val_2.fq"

        done
```

```
## Bismark III
sbatch bisulfit_3.sh
#!/bin/bash
#SBATCH --job-name=bisulfit_3
#SBATCH --nodes=1
#SBATCH --time=0-10:00:00
#SBATCH --mem=40GB
gunzip *gz
file_arr=(*.sam)
len=${#file_arr[@]}
for ((i=0;i<len;i++)); do
        /bismark_methylation_extractor -p --no_overlap --bedGraph --counts --buffer_size 10G --
genome_folder /gpfs/rocket/home/elsakova/seq/????/ ${file_arr[i]}
        done
```

2. Temra Cells prediction model

libraries

```
library(readxl)
library(writexl)
library(ggpubr)
library(dplyr)
library(tidyverse)
library(hrbrthemes)
library(viridis)
library(ggplot2)
library(patchwork)

exceldata1 <- as.data.frame(read_excel("SEQ for R step 1.xlsx"))
BSAS_I_VI_combined <- exceldata1
exceldata2 <- as.data.frame(read_excel("SEQ 2.xlsx"))
BSAS_I_VI_combined_count <- exceldata2
COMB_PREPRO_IMP <- as.data.frame(read_excel("BSAS_I_VI_combined_prepro_imp.
xlsx"))
CD4_CD8 <- as.data.frame(read_excel("CD4_8.xlsx"))
CD4_CD8
S_A <- as.data.frame(read_excel("Koond 24.11.21 isikuandmetetaKK.xlsx"))
STAT <- as.data.frame(read_excel ("correlations_all_M_F_correlations_result
s_FACS_CpG.xlsx"))
AHTOS_TABLE <- as.data.frame(read_excel("BSASIII_combined_300_cov_mes_075_i
mp_diseases2.xlsx"))
AHTOS_TABLE
COMB_CD4CD8 <- as.data.frame(read_excel("Data+S&A.xlsx"))
```

low values and missing values

```
BSAS_I_VI_combined_count[BSAS_I_VI_combined_count < 300] <-  NA
BSAS_I_VI_combined <- as.matrix(BSAS_I_VI_combined)
BSAS_I_VI_combined[which(is.na(BSAS_I_VI_combined_count))] <- NA
BSAS_I_VI_combined <- as.data.frame(BSAS_I_VI_combined)
BSAS_I_VI_combined
```

The code to prepare the data for imputation (make sure that there are not too many missing values in row/column)

```
preprocess_imp <- function(data, cutoff = 0.1) {
  rownames(data) <- data$Code
  sites <- grep('chr', colnames(data), value = TRUE)
  facs_info <- colnames(data)[!(colnames(data) %in% sites)]
  data_sub <- data[, sites]
  data_sub <- data_sub[, colMeans(is.na(data_sub)) <= cutoff]
  data_sub <- data_sub[rowMeans(is.na(data_sub)) <= cutoff, ]
  print(paste(c("Removed rows:",
                rownames(data)[!rownames(data) %in% rownames(data_sub)]), c
ollapse = ";  "))
  print(paste(c("Removed columns:",
                sites[!sites %in% colnames(data_sub)]), collapse = ";  "))
  print(paste("Missing data before:", round(mean(is.na(data[, sites])), dig
its = 6)))
  print(paste(c("Dimensions:", dim(data[, sites])), collapse = " "))
  print(paste("Missing data after:", round(mean(is.na(data_sub)), digits =
```

```
6)))
  print(paste(c("Dimensions:", dim(data_sub)), collapse = " "))
  c_names <- c(colnames(data_sub), facs_info)
  data_final <- data[rownames(data) %in% rownames(data_sub), colnames(data)
%in% c_names]
  return(data_final)
}

BSAS_I_VI_combined_prepro <- preprocess_imp(BSAS_I_VI_combined, cutoff = 0.
25)
BSAS_I_VI_combined_prepro
#summary(BSAS_I_VI_combined_prepro)
BSAS_I_VI_combined_prepro_conv <- type.convert(BSAS_I_VI_combined_prepro)
summary(BSAS_I_VI_combined_prepro_conv)
```

Imputation code

```
impute_missForest <- function(data, include_status = FALSE) {
  require(missForest)
  sites <- grep('chr', colnames(data), value = TRUE)
  if(include_status == TRUE) {
    data$Status <- as.factor(data$Status)
    sites <- c('Status', sites)
  }
  facs_info <- colnames(data)[!(colnames(data) %in% sites)]
  data_sub <- data[, sites]
  data_sub_imp <- missForest(data_sub)
  print(paste('OOBerror:', data_sub_imp$OOBerror))
  data[, sites] <- data_sub_imp$ximp
  return(data)
}
```

MissForest and extraction to xlsx

```
BSAS_I_VI_combined_prepro_imp <- impute_missForest(BSAS_I_VI_combined_prepr
o_conv)
BSAS_I_VI_combined_prepro_imp
write_xlsx(BSAS_I_VI_combined_prepro_imp, "D:\\Imuno\\R\\BSAS_I_VI_combined
_prepro_imp.xlsx")
```

Combining 2 tables for further work

```
COMB_CD4_CD8 <- merge(COMB_PREPRO_IMP, CD4_CD8,by=c("Kood"))
COMB_CD4_CD8 <- merge(COMB_CD4_CD8, S_A, by=c("Number Cells"))
COMB_CD4_CD8 <- within(COMB_CD4_CD8, rm ("email", "märkus"))
write_xlsx(COMB_CD4_CD8,"D:\\Imuno\\R\\Data+S&A.xlsx")
COMB_CD4_CD8
```

Combining 2 tables

```
data_CD4_8 <- as.data.frame(read_excel("COMB_ALL_AHTOS.xlsx"))
data_CD4_8
summary(data_CD4_8$Themra_WBC)
test <- data_CD4_8[data_CD4_8$Themra_WBC > 5, c('Kood','Themra_WBC')]

colnames(test)
```

CORRELATION and STATs

```
cor_test_fun <- function(data, input_list, filename, save = T, use_sub_data
sets = T, arrange_data = T, include_partial = F, covariate = NULL) {
  require(writexl)
  require(plyr)
  require(dplyr)
  require(ppcor)
  cor_test_fun_helper <- function(data, feature1, feature2) {
    if(include_partial) {
      data_sub <- data[, c(feature1, feature2, covariate)]
      data_sub <- data_sub[complete.cases(data_sub), ]
    } else {
      data_sub <- data[, c(feature1, feature2)]
      data_sub <- data_sub[complete.cases(data_sub), ]
    }
    # test whether normality holds
    normality_test_pval_f1 <- tryCatch(shapiro.test(data_sub[, feature1])$p
.value, error = function(x) return(NA))
    normality_test_pval_f2 <- tryCatch(shapiro.test(data_sub[, feature2])$p
.value, error = function(x) return(NA))
    # calculate sample size i.e. pairwise complete observations
    n_obs <- nrow(data_sub)
    # extract all relevant info for correlation
    res_pearson <- cor.test(data_sub[, feature1], data_sub[, feature2], met
hod = 'pearson')
    res_spearman <- cor.test(data_sub[, feature1], data_sub[, feature2], me
thod = 'spearman')
    if(include_partial) {
      partial_pearson <- pcor.test(data_sub[, feature1],
                                   data_sub[, feature2],
                                   data_sub[, covariate], method = 'pearson
')
      partial_spearman <- pcor.test(data_sub[, feature1],
                                    data_sub[, feature2],
                                    data_sub[, covariate], method = 'spearm
an')
      # partial correlation estimate
      partial_pearson_cor <- partial_pearson$estimate
      partial_spearman_cor <- partial_spearman$estimate
      # partial correlation p-value
      partial_pearson_pval <- partial_pearson$p.value
      partial_spearman_pval <- partial_spearman$p.value
    }
    # combine all results into named vector
    res <- c(feature1, feature2, n_obs, normality_test_pval_f1, normality_t
est_pval_f2,
             res_pearson$estimate, res_pearson$p.value, paste(round(res_pea
rson$conf.int, 3), collapse = '...'),
             res_pearson$parameter, res_spearman$estimate, res_spearman$p.v
alue)
    if(include_partial){
      res <- c(res, c(partial_pearson_cor, partial_pearson_pval, partial_sp
earman_cor, partial_spearman_pval))
      names(res) <- c('feature1', 'feature2', 'n_obs', 'shapiro_pval_featur
```

```
e1', 'shapiro_pval_feature2',
                      'pearson_cor', 'pearson_pval', 'pearson_conf_int', 'p
earson_df',
                      'spearman_rho', 'spearman_pval', 'partial_pearson_cor
', 'partial_pearson_pval',
                      'partial_spearman_cor' , 'partial_spearman_pval')
    } else {
      names(res) <- c('feature1', 'feature2', 'n_obs', 'shapiro_pval_featur
e1', 'shapiro_pval_feature2',
                      'pearson_cor', 'pearson_pval', 'pearson_conf_int', 'p
earson_df',
                      'spearman_rho', 'spearman_pval')
    }
    # make df from named vector
    res_df_sub <- data.frame(as.list(res))
    return(res_df_sub)
  }
  data <- as.data.frame(data)
  # input is list that contains features1 and features2 or elements
  # that contain features1 and features2
  if(is.data.frame(input_list)) {
    feature_pair_df <- input_list
  } else {
    if(length(input_list) == 2 & all(names(input_list) %in% c('features1',
'features2'))) {
      feature_pair_df <- expand.grid(input_list[['features1']], input_list[
['features2']])
    } else {
      # if it contains many pairs of features1 and 2
      for(item in input_list) {
        features1 <- item[['features1']]
        features2 <- item[['features2']]
        feature_pair_df_sub <- expand.grid(features1, features2)
        if(exists('feature_pair_df')) {
          feature_pair_df <- base::rbind(feature_pair_df, feature_pair_df_s
ub)
        } else {
          feature_pair_df <- feature_pair_df_sub
        }
      }
    }
  }
  # let's specify Gender specific datasets
  if(use_sub_datasets) {
    data_M <- data[data$Gender == 'M', ]
    data_F <- data[data$Gender == 'F', ]
    data_list <- list('all' = data,
                      'M' = data_M,
                      'F' = data_F)
  } else {
    data_list <- list('all' = data)
  }
  for(i in 1:nrow(feature_pair_df)) {
    # extract feature pair
```

```
    feature1 <- as.character(feature_pair_df[i, 'Var1'])
    feature2 <- as.character(feature_pair_df[i, 'Var2'])
    # with following for loop we build a dataframe with 1 row that contians
relevant correlations and p-values
    for(name_df in names(data_list)) {
      tryCatch(
        expr = {
          cor_df_sub <- cor_test_fun_helper(data_list[[name_df]], feature1,
feature2)
          cor_df_sub$dataset <- name_df
          if(exists('cor_df')) {
            cor_df <- bind_rows(cor_df, cor_df_sub)
          } else {
            cor_df <- cor_df_sub
          }
        },
        error = function(e){
          print(e)
          print(paste('name_df: ', name_df, ', feature1: ', feature1, ', fe
ature2: ', feature2))
        }
      )
    }
    if(exists('cor_df_final')) {
      cor_df_final <- bind_rows(cor_df_final, cor_df)
    } else {
      cor_df_final <- cor_df
    }
    rm(cor_df)
  }
  # convert columns to correct type
  cor_df_final <- type.convert(cor_df_final)
  # now we need to adjust p-values (but we have p-values for different data
sets (TD, nTD, HC etc))
  cor_df_final$adj_pearson_pval <- NA
  cor_df_final$adj_spearman_pval <- NA
  if(include_partial) {
    # 'partial_pearson_cor', 'partial_pearson_pval',
    #'partial_spearman_cor' , 'partial_spearman_pval'
    cor_df_final$adj_partial_pearson_pval <- NA
    cor_df_final$adj_partial_spearman_pval <- NA
  }
  for(name_df in names(data_list)) {
    cor_df_final[cor_df_final$dataset == name_df, ]$adj_pearson_pval <- p.a
djust(cor_df_final[cor_df_final$dataset == name_df, ]$pearson_pval,

method = 'fdr')
    cor_df_final[cor_df_final$dataset == name_df, ]$adj_spearman_pval <- p.
adjust(cor_df_final[cor_df_final$dataset == name_df, ]$spearman_pval,

method = 'fdr')
    if(include_partial) {
      cor_df_final[cor_df_final$dataset == name_df, ]$adj_partial_pearson_p
val <- p.adjust(cor_df_final[cor_df_final$dataset == name_df, ]$partial_pea
```

```
rson_pval,

method = 'fdr')
      cor_df_final[cor_df_final$dataset == name_df, ]$adj_partial_spearman_
pval <- p.adjust(cor_df_final[cor_df_final$dataset == name_df, ]$partial_sp
earman_pval,

method = 'fdr')
    }
  }
  # sort the columns
  if(include_partial) {

    reorder_cols <- c('dataset', 'feature1', 'feature2', 'n_obs', 'shapiro_
pval_feature1', 'shapiro_pval_feature2',
                      'pearson_cor', 'pearson_conf_int','pearson_pval', 'ad
j_pearson_pval', 'pearson_df',
                      'spearman_rho', 'spearman_pval', 'adj_spearman_pval',
'partial_pearson_cor', 'partial_pearson_pval',
                      'adj_partial_pearson_pval',
                      'partial_spearman_cor' , 'partial_spearman_pval', 'ad
j_partial_spearman_pval')
  } else {
    reorder_cols <- c('dataset', 'feature1', 'feature2', 'n_obs', 'shapiro_
pval_feature1', 'shapiro_pval_feature2',
                      'pearson_cor', 'pearson_conf_int','pearson_pval', 'ad
j_pearson_pval', 'pearson_df',
                      'spearman_rho', 'spearman_pval', 'adj_spearman_pval')
  }
  cor_df_final <- cor_df_final[, reorder_cols]
  # sort the rows
  if(arrange_data) {
    cor_df_final <- dplyr::arrange(cor_df_final, adj_pearson_pval)
  }
  # divide results to two if possible (CpG)
  #cpg_bool <- grepl('chr', cor_df_final$feature1) | grepl('chr', cor_df_fi
nal$feature2)
  # write xlsx file
  if(save) {
    write_xlsx(cor_df_final,
               path = paste('correlations_all_M_F_', filename, '.xlsx', sep
= ''))
  }
  return(cor_df_final)
}

features_for_correlation <- list(
  'Features' = list(features1 = grep('chr', colnames(data_CD4_8), value = T
),
                    features2 = c('CTLemra_WBC', 'Themra_WBC'))
)
```

Checking correlations CD8+/CD4+

```
cor_res <- cor_test_fun(data = data_CD4_8,
                        input_list = features_for_correlation,
                        filename = 'correlations_results_FACS_CpG_ALL',
                        save = T, use_sub_datasets = F)

newdata <- cor_res[order (cor_res$ pearson_pval),]
newdata
```

Transformation

```
data_CD4_8$CTLemra_WBC_log_trans <- log(data_CD4_8$CTLemra_WBC)
data_CD4_8$CTLemra_WBC_sqrt_trans <- sqrt(data_CD4_8$CTLemra_WBC)
data_CD4_8$CTLemra_WBC_cube_root_trans <- (data_CD4_8$CTLemra_WBC)^(1/3)

data_CD4_8$Themra_WBC_log_trans <- log(data_CD4_8$Themra_WBC)
data_CD4_8$Themra_WBC_sqrt_trans <- sqrt(data_CD4_8$Themra_WBC)
data_CD4_8$Themra_WBC_cube_root_trans <- (data_CD4_8$Themra_WBC)^(1/3)

data_CD4_8
write_xlsx(data_CD4_8,"D:\\Imuno\\R\\TRANSFORMATIONS.xlsx")
```

1. For age/sex visualisation study following code:

```
agebreaks <- c(0,1,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85, 90, 9
5, 100, 105, 150)
agelabels <- c("0-1","1-4","5-9","10-14","15-19","20-24","25-29","30-34",
               "35-39","40-44","45-49","50-54","55-59","60-64","65-69",
               "70-74","75-79","80-84","85-89", "90-94", "95-99", "100-104"
, "105-")
data_CD4_8$Age_group <- cut(data_CD4_8$Age, breaks = agebreaks, right = FAL
SE, labels =agelabels)

age_gender_pyramid <- function(data, filename, title = '', age_groups = NUL
L, text_nudge = 4, width_parameter = 1.1,
                               keep_above_65 = F, use_all_levels = F, label
_size = 2.5) {


  data <- data[data$Age >= 15, ]
  data <- data[, c('Age_group', 'Gender')]
  data <- data[!is.na(data$Age_group), ]
  data$Age_group <- factor(data$Age_group, levels = c("15-19","20-24","25-2
9","30-34",
                                                      "35-39","40-44","45-4
9","50-54","55-59","60-64","65-69",
                                                      "70-74","75-79","80-8
4","85-89", "90-94", "95-99"))


  data_sub <- data %>%
    dplyr::group_by(Age_group, Gender) %>%
    dplyr::summarise(Count = n())

  # data_sub$Count[is.na(data_sub$Gender)] <- 0
  # data_sub$Gender[data_sub$Count == 0] <- 'F'
```

```
  # lisame proportsioonid
  data_sub$Proportion <- NA
  data_sub$Proportion[data_sub$Gender == 'M'] <- data_sub$Count[data_sub$Ge
nder == 'M'] /
    sum(data_sub$Count[data_sub$Gender == 'M']) * 100
  data_sub$Proportion[data_sub$Gender == 'F'] <- data_sub$Count[data_sub$Ge
nder == 'F'] /
    sum(data_sub$Count[data_sub$Gender == 'F']) * 100

  data_sub$Proportion <- round(data_sub$Proportion)
  data_sub$Proportion <- paste(data_sub$Proportion, '%', sep = '')
  data_sub$Proportion[data_sub$Proportion == '0%'] <- NA

  n_F <- sum(data$Gender == 'F')
  n_M <- sum(data$Gender == 'M')
  p_F <- paste(round(mean(data$Gender == 'F') * 100), '%', sep = '')
  p_M <- paste(round(mean(data$Gender == 'M') * 100), '%', sep = '')

  p <- ggplot(data_sub, aes(x = Age_group, fill = Gender, label = Proportio
n,
                            y = ifelse(test = Gender == 'F',
                                       yes = -Count, no = Count))) +
    geom_bar(stat = 'identity') +
    geom_text(nudge_y = ifelse(test = data_sub$Gender == 'F',
                               yes = -text_nudge, no = text_nudge), size =
label_size) +
    scale_y_continuous(labels = abs, limits = width_parameter * max(data_su
b$Count) * c(-1,1)) +
    scale_x_discrete(drop = !use_all_levels) +
    coord_flip() +
    theme_classic() +
    labs(x = 'Age', y = 'Count', title = paste(title, ' (', n_F, ' (', p_F,
')', ' / ', n_M, ' (', p_M, ')', ') ', '(F/M)', sep = '')) +
    scale_fill_manual(values=c('F' = 'pink2', 'M' = 'cornflowerblue')) +
    theme(axis.text=element_text(size=7),
          axis.title=element_text(size=9),
          title = element_text(size=9),
          legend.text=element_text(size=7),
          legend.position = 'bottom',
          axis.line = element_line(size = 0.15))

  #ggsave(filename = paste('age_gender_pyramid_', filename, '.png', sep = '
'), plot = p, dpi = 320, width = 9, height = 7)
  return(p)

}

# call a funciton to generate a plot
p_pyramid_BSASIII <- age_gender_pyramid(data_CD4_8,
                                        title = 'Age/Gender',
                                        filename = 'BSASIII_test',
                                        text_nudge = 2, width_parameter = 1
.2, keep_above_65 = T, use_all_levels = T)
```

```
# save a plot
ggsave(p_pyramid_BSASIII, filename = 'Fig_S2.png', dpi = 320, width = 4, he
ight = 4)


p_pyramid_BSASIII
```

TRANSFORMATION

```
plot_histogram <- function(data_vect, x_name) {
  data <- data.frame(var = data_vect)
  print(data)
  p <- ggplot(data, aes(x = var)) +
    geom_histogram(color='black', fill='gray70') +
    theme_classic() +
    theme(axis.text=element_text(size=8),
          axis.title=element_text(size=8,face='bold'),
          legend.text=element_text(size=8),
          legend.title=element_text(size=8, face = 'bold'),
          axis.text.x = element_text(angle = 90, hjust = 1, size = 8),
          axis.line = element_line(size = 0.2),
          plot.title = element_text(size = 8, face = 'bold'),
          legend.position = 'bottom') +
    labs(y = 'Count', x = x_name)
  return(p)
}

p_hist <- plot_histogram(data_vect = data_CD4_8$CTLemra_WBC, x_name = "CD8+
Temra")
p_hist


p_hist1 <- plot_histogram(data_vect = log(data_CD4_8$CTLemra_WBC), x_name =
"CD8+ Temra log transformed")
p_hist1


p_hist2 <- plot_histogram(data_vect = sqrt(data_CD4_8$CTLemra_WBC), x_name
= expression(sqrt('CD8+ Temra')))
p_hist2


p_hist3 <- plot_histogram(data_vect = (data_CD4_8$CTLemra_WBC)^(1/3), x_nam
e = expression(sqrt('CD8+ Temra', 3)))
p_hist3

shapiro.test(data_CD4_8$CTLemra_WBC)
shapiro.test(log(data_CD4_8$CTLemra_WBC))
shapiro.test(sqrt(data_CD4_8$CTLemra_WBC))
shapiro.test((data_CD4_8$CTLemra_WBC)^(1/3))

transformations<-wrap_plots(p_hist, p_hist1, p_hist2, p_hist3)

ggsave(transformations, filename = 'Fig_S3.png', dpi = 320, width = 4, heig
ht = 4)
```

Q-Q- normality test (Shapiro-wilk test)

```
NORM <- ggqqplot(data_CD4_8$CTLemra_WBC)+
  labs(
    title = "QQ plot, not transformed",
    x = "Theoretical quantiles",
    y = "Samples quantiles"
  )

a <- NORM + annotate("text", x = -1.5, y = 16,
                     label = "Shapiro-Wilk test, p=3.654e-12", size = 2.5)


LOG <- ggqqplot(log(data_CD4_8$CTLemra_WBC))+
  labs(
    title = "QQ plot, LOG",
    x = "Theoretical quantiles",
    y = "Samples quantiles"
  )

b <- LOG + annotate("text", x = -1.5, y = 5.5,
                    label = "Shapiro-Wilk test, p = 0.0003023", size = 2.5)



SQRT <- ggqqplot(sqrt(data_CD4_8$CTLemra_WBC))+
  labs(
    title = "QQ plot, SQRT",
    x = "Theoretical quantiles",
    y = "Samples quantiles"
  )

c <- SQRT + annotate("text", x = -1.5, y = 5.5,
                     label = "Shapiro-Wilk test, p = 0.0004681", size = 2.5
)



TRANS_1 <- ggqqplot((data_CD4_8$CTLemra_WBC)^(1/3))+
  labs(
    title = "QQ plot, ^(1/3)",
    x = "Theoretical quantiles",
    y = "Samples quantiles"
  )

d <- TRANS_1 + annotate("text", x = -1.5, y = 5.5,
                        label = "Shapiro-Wilk test, p = 0.1199", size = 2.5
)

QQPLOTS<-wrap_plots(a, b, c, d)
QQPLOTS

ggsave(QQPLOTS, filename = 'QQ.png', dpi = 320, width = 4, height = 4)
```

2. FACS/AGE

```
CTL<-ggplot(data = data_CD4_8, aes_string(x = "Age", y =  "CTLemra_WBC")) +
  labs(y= "CD8+ Temra/WBC")+
  geom_point(size = 2) +
  theme_classic()

TH<-ggplot(data = data_CD4_8, aes_string(x = "Age", y =  'Themra_WBC')) +
   labs(y= "CD4+ Temra/WBC")+
  geom_point(size = 2) +
  theme_classic()

TEMRAS<-wrap_plots (CTL+TH)

ggsave(TEMRAS, filename = 'Temras.png', dpi = 320, width = 6, height = 4)
```

3. Met VS T cells, Scatterplot

```
Chr <- c('chr12.4915855.4915855', 'chr1.240164755.240164755', 'chr17.799217
15.79921715', 'chr17.79921776.79921776')

AHL <- data.frame(Chr)


Chr1 <- c("chr1.240164755.240164755", "chr12.4915938.4915938",  "chr12.4915
925.4915925",    "chr12.4915855.4915855", "chr12.24689930.24689930")

AHL1 <- data.frame(Chr1)


# function for one scatterplot #parameters
library(patchwork)
plot_scatterplot <- function(data, pval_data, feature_x, feature_y,
                             title = '', text_size = 6, title_size = 6, poi
nt_size = 0.75,
                             label_size = 2) {

  require(ggplot2) #libraries
  require(cowplot)
  require(ggpubr)
  require(Hmisc)
  require(scales)
  require(patchwork)

  data$Sugu <- factor(data$Gender, levels = c('F', 'M', 'F+M')) #have the f
actor to present

  x_coord <- min(data[, feature_x], na.rm = T)
  y_coord_max <- max(data[, feature_y], na.rm = T)
  y_coord_min <- min(data[, feature_y], na.rm = T)

  pval_data_sub <- pval_data %>%
    dplyr::filter(feature1 %in% c(feature_x, feature_y) & feature2 %in% c(f
eature_x, feature_y))


  pval <- pval_data_sub$adj_pearson_pval[pval_data_sub$dataset == "all"]
  cor_r <- pval_data_sub$pearson_cor[pval_data_sub$dataset == "all" ]
```

```
  p <- ggplot(data = data, aes_string(x = feature_x, y = feature_y, colour
= 'Gender')) +
    geom_point(size = point_size, alpha = 0.7) +
    geom_smooth(method = 'lm', na.rm = TRUE, se = FALSE, lwd = 0.9) +
    geom_smooth(method = 'lm', na.rm = TRUE, lwd = 0.9, colour = 'black') +
    theme_classic() +
    theme(axis.text = element_text(size = text_size),
          axis.title = element_text(size = title_size),
          title  = element_text(size=title_size),
          axis.line = element_line(size = 0.15),
          legend.position = 'bottom')  +
    labs(x = make_correct_name(feature_x), y = make_correct_name(feature_y)
, title = title) +
    ggplot2::annotate('text', x = x_coord, y = (1 + (0.05 *((y_coord_max-y_
coord_min)/y_coord_max))) * y_coord_max,
                      label = paste('r =', round(cor_r, 2), ', p =', format
C(pval, format = 'g', digits = 2)),
                      color ='black', hjust = 0,
                      size = label_size) +
    scale_colour_manual(values = c('pink2', 'cornflowerblue', 'black'),
                        labels = c('F', 'M', 'F+M'), limits = c('F', 'M', '
F+M'), drop = F)

  return(p)
}


# make several scatterplots
plot_selected_scatterplots <- function(data, pval_data, features_x, feature
s_y,
                                       title = '', text_size = 6, title_siz
e = 6, point_size = 1,
                                       label_size = 2, ncol = 4) {
  require(ggplot2)
  require(ggpubr)

  plot_list <- list()

  for(i in 1:length(features_x)){

    p <- plot_scatterplot(data = data, pval_data = pval_data,
                          feature_x = features_x[i], feature_y = features_y
[i],
                          title = title, text_size = text_size,

                          title_size = title_size, point_size = point_size,
                          label_size = label_size)

    plot_list[[i]] <- p
  }

  p_final <- wrap_plots(plot_list, ncol = ncol) +
```

```
    plot_layout(guides = 'collect') &
    theme(legend.position = 'bottom')

  return(p_final)

}

# plot selected correlations (features_x and features_y are given as vector
s)
plots_cor <- plot_selected_scatterplots(data = data_CD4_8, pval_data = cor_
res,
                                        features_x = AHL$Chr,
                                        features_y = rep('CTLemra_WBC', len
gth(AHL$Chr)),
                                        title = ' ',

                                        text_size = 10, title_size = 10, po
int_size = 2,
                                        label_size = 6, ncol = 2)
# save plot
ggsave(plots_cor, filename = 'CTL.png', width = 15, height = 15)

plots_cor


# plot selected correlations (features_x and features_y are given as vector
s)
plots_cor1 <- plot_selected_scatterplots(data = data_CD4_8, pval_data = cor
_res,
                                         features_x = AHL1$Chr1,
                                         features_y = rep('Themra_WBC', len
gth(AHL1$Chr1)),
                                         title = '',
                                         text_size = 10, title_size = 10, p
oint_size = 2,
                                         label_size = 6, ncol = 3)


# save plot
ggsave(plots_cor1, filename = 'Themra.png', width = 20, height = 15)

plots_cor1

ggplot(data = data_CD4_8, aes_string(x = "chr17.79921715.79921715", y =  "C
TLemra_WBC", colour = "Age")) +
  geom_point(size = 2) +
  theme_classic() +
  scale_color_gradient(low = 'blue', high = 'red')
```

4. A heatmap for CpGs

```
HeatMap_data <- as.data.frame(read_excel("Genes for the HeatMap.xlsx"))
```

Function to plot heatmap:

```
plot_heatmap_FACS_CpG <- function(data, row_features, col_features,
                                  filename, col_k = 5, row_k = 5, cutoff =
0) {
  require(ggplot2)
  require(reshape2)
  require(ComplexHeatmap)
  require(ggplot2)
  require(RColorBrewer)
  require(viridis)
  require(ggplotify)
  require(stringr)

  reorder_cormat <- function(cormat){
    # Use correlation between variables as distance
    dd <- as.dist((1-cormat)/2)
    hc <- hclust(dd)
    cormat <- cormat[hc$order, hc$order]
  }
  data <- data[, c(row_features, col_features)]
  #data <- data[complete.cases(data), ]
  cormat <- round(cor(data, method = 'pearson', use = 'pairwise.complete.ob
s'), 2)
  #cormat <- reorder_cormat(cormat)
  cormat <- cormat[(rownames(cormat) %in% row_features), (colnames(cormat)
%in% col_features)]
  keep_features_row <- c()
  keep_features_col <- c()
  for(r_name in rownames(cormat)) {
    if(!all(abs(cormat[r_name, ]) < cutoff)) {
      keep_features_row <- c(keep_features_row, r_name)
    }
  }
  for(c_name in colnames(cormat)) {
    if(!all(abs(cormat[, c_name]) < cutoff)) {
      keep_features_col <- c(keep_features_col, c_name)
    }
  }
  print(dim(cormat))
  cormat <- cormat[keep_features_row, keep_features_col]
  print(dim(cormat))
  #colnames(cormat) <- sapply(colnames(cormat), function(x) make_correct_na
me(x))
  rownames(cormat) <- sapply(rownames(cormat), function(x) str_extract(patt
ern = 'chr\\d{1,2}\\.\\d{5,}', string = x))
  # FACS annotatsioonid
  CD4_CD8_annotaton <- ifelse(grepl(pattern = 'Th|CD4', keep_features_col),
'CD4+',
                              ifelse(grepl(pattern = 'CTL|CD8', keep_featur
es_col), 'CD8+', NA))

  # FACS annotations
  column_anno <- HeatmapAnnotation('Cell type' = CD4_CD8_annotaton,
                                   col = list('Cell type' = c('CD4+' = 'red
',
```

```
                                                                'CD8+' = 'blu
e')),
                                        annotation_legend_param = list(labels_gp
= gpar(fontsize = 6),
                                                                       title_gp
= gpar(fontsize = 6)),
                                        show_legend = T,
                                        simple_anno_size_adjust = T,
                                        annotation_name_gp = gpar(fontsize = 6),
                                        height = unit(1.5, units = 'cm'))
  # CpG annotations
  cells <- sapply(str_split(keep_features_row, pattern = '\\.'), function(x
) x[5])
  genes <- sapply(str_split(keep_features_row, pattern = '\\.'), function(x
) x[4])
  row_anno = rowAnnotation('Gene' = anno_text(genes, location = 0.5, just =
'center',
                                              gp = gpar(fill = 'purple', co
l = 'white', border = 'black', fontsize = 6),
                                              width = max_text_width(genes)
*1.2),
                           annotation_legend_param = list(labels_gp = gpar(
fontsize = 6), title_gp = gpar(fontsize = 7, fontface = 'plain')),
                           show_legend = T,
                           simple_anno_size_adjust = T,
                           annotation_name_gp = gpar(fontsize = 5),
                           width = unit(1.5, units = 'cm'))

  p <- Heatmap(cormat, heatmap_legend_param = list(title = "Pearson's\ncorr
elation", labels_gp = gpar(fontsize = 5),
                                                   title_gp = gpar(fontsize
= 7, fontface = 'plain'),
                                                   direction = 'vertical',
title_position = 'topleft'),
               cluster_rows = T, row_names_gp = gpar(fontsize = 5),
               column_names_gp = gpar(fontsize = 6), row_split = row_k, col
umn_split = col_k, top_annotation = column_anno,
               left_annotation = row_anno, show_heatmap_legend = T, row_tit
le = NULL, column_title = NULL, column_dend_height = unit(0.5, units = 'cm'
),
               row_dend_width = unit(1, units = 'cm'))
  p <- grid::grid.grabExpr(draw(p, heatmap_legend_side = 'left', annotation
_legend_side = 'left'))
  ggsave(paste('heatmap_CpG_FACS',filename, '.png', sep = ''), plot = p,
         dpi = 320, width = 7.2, height = 12)
  return(p)
}
cpg_FACS_plot <- plot_heatmap_FACS_CpG(data = HeatMap_data,
                                       col_features = c('CTLemra_WBC', 'The
mra_WBC', 'CTLemra_WBC_cube_root_trans', 'Themra_WBC_cube_root_trans'),
                                       row_features = grep('chr', colnames(
HeatMap_data), value = T), filename = 'filename',
                                       cutoff = 0.1,
                                       col_k = 2, row_k = 4)
```

```
cpg_FACS_plot
```

MODELLING

CTL

```
library(caret)
train_indx_0 <- caret::createDataPartition(data_CD4_8$CTLemra_WBC, p = 0.75
,
                                           list = FALSE,
                                           times = 1)

train_CTL <- data_CD4_8[train_indx_0, ]
test_CTL <- data_CD4_8[-train_indx_0, ]

lm0 <- train(CTLemra_WBC~chr12.4915855.4915855+chr1.240164755.240164755+chr
17.79921715.79921715+chr17.79921776.79921776, data = train_CTL, method = "l
m")
summary(lm0)


lm0
res_0 <- resid(lm0)

train_indx_1 <- caret::createDataPartition(data_CD4_8$CTLemra_WBC_cube_root
_trans, p = 0.75,
                                           list = FALSE,
                                           times = 1)

train_CTL_norm <- data_CD4_8[train_indx_1, ]
test_CTL_norm <- data_CD4_8[-train_indx_1, ]

lm1 <- train(CTLemra_WBC_cube_root_trans~chr12.4915855.4915855+chr1.2401647
55.240164755+chr17.79921715.79921715+chr17.79921776.79921776, data = train_
CTL_norm, method = "lm")
summary(lm1)

lm1

res_1 <- resid(lm1)
```

Themra

```
train_indx_2 <- caret::createDataPartition(data_CD4_8$Themra_WBC, p = 0.75,
                                           list = FALSE,
                                           times = 1)

train_TH <- data_CD4_8[train_indx_2, ]
test_TH <- data_CD4_8[-train_indx_2, ]


lm2 <- train(Themra_WBC~chr1.240164755.240164755+chr12.4915938.4915938+chr1
2.4915925.4915925+chr12.4915855.4915855+chr12.24689930.24689930, data = tra
in_TH, method = "lm")
```

```
summary(lm2)

lm2

res_2 <- resid(lm2)

train_indx_3 <- caret::createDataPartition(data_CD4_8$Themra_WBC_cube_root_
trans, p = 0.75,
                                           list = FALSE,
                                           times = 1)

train_TH_norm <- data_CD4_8[train_indx_3, ]
test_TH_norm <- data_CD4_8[-train_indx_3, ]

lm3 <- train(Themra_WBC_cube_root_trans~chr1.240164755.240164755+chr12.4915
938.4915938+chr12.4915925.4915925+chr12.4915855.4915855+chr12.24689930.2468
9930, data = train_TH_norm, method = "lm")
summary(lm3)

lm3

res_3 <- resid(lm3)
```

Test plots

```
plot_pred_vs_actual <- function(test_data, model, dependent_variable, depen
dent_variable_name, label_size = 2.5) {

  require(ggpmisc)
  require(ggplot2)

  # predict dependent feature values for test data
  predicted_values <- predict(model, test_data)
  # take actual values
  actual_values <- test_data[, dependent_variable]
  predicted_actual_df <- data.frame(predicted = predicted_values,
                                    actual = actual_values)

  # calculate RMSE and cor
  cor_pred_actual <- round(cor(predicted_actual_df$actual, predicted_actual
_df$predicted), 3)
  rmse_pred_actual <- round(sqrt(mean((predicted_actual_df$actual - predict
ed_actual_df$predicted)**2)), 3)
  annot_df <- data.frame('cor' = cor_pred_actual, 'rmse' = rmse_pred_actual
)

  # calculate coordinates for labels
  x_coord <- min(predicted_actual_df$actual, na.rm = T)
  y_coord <- max(predicted_actual_df$predicted, na.rm = T)

  # scatterplot
  p <- ggplot(predicted_actual_df, aes(x = actual, y = predicted)) +
    geom_point(size = 0.7, alpha = 0.8) +
    theme_classic() + geom_abline(intercept = 0, slope = 1, color = 'red3',
linetype='dashed', size=0.8) +
```

```
    theme(axis.title=element_text(size=9),
          axis.text=element_text(size=7),
          plot.title = element_text(size=9),
          legend.title = element_text(size = 9),
          legend.text = element_text(size=7),
          legend.position = 'bottom',
          axis.line = element_line(size = 0.15)) +
    labs(x ='Actual', y='Predicted', title = dependent_variable_name) +
    expand_limits(x = 0, y = 0) +
    ggplot2::annotate('text', x = x_coord, y = y_coord,
                      label = paste(' r=', annot_df$cor, ', RMSE=', annot_d
f$rmse, sep = ''),
                      hjust = 0,
                      size = label_size) +
    guides(colour = guide_legend(override.aes = list(size=2)))

  return(p)
}


plot_pred_vs_actual_cube <- function(test_data, model, dependent_variable,
dependent_variable_name, label_size = 2.5) {

  require(ggpmisc)
  require(ggplot2)

  # predict dependent feature values for test data
  predicted_values <- predict(model, test_data)
  predicted_values <- predicted_values^3
  # take actual values
  actual_values <- test_data[, dependent_variable]
  actual_values <- actual_values^3
  predicted_actual_df <- data.frame(predicted = predicted_values,
                                    actual = actual_values)

  # calculate RMSE and cor
  cor_pred_actual <- round(cor(predicted_actual_df$actual, predicted_actual
_df$predicted), 3)
  rmse_pred_actual <- round(sqrt(mean((predicted_actual_df$actual - predict
ed_actual_df$predicted)**2)), 3)
  annot_df <- data.frame('cor' = cor_pred_actual, 'rmse' = rmse_pred_actual
)

  # calculate coordinates for labels
  x_coord <- min(predicted_actual_df$actual, na.rm = T)
  y_coord <- max(predicted_actual_df$predicted, na.rm = T)

  # scatterplot
  p <- ggplot(predicted_actual_df, aes(x = actual, y = predicted)) +
    geom_point(size = 0.7, alpha = 0.8) +
    theme_classic() + geom_abline(intercept = 0, slope = 1, color = 'red3',
linetype='dashed', size=0.8) +
    theme(axis.title=element_text(size=9),
          axis.text=element_text(size=7),
```

```
          plot.title = element_text(size=9),
          legend.title = element_text(size = 9),
          legend.text = element_text(size=7),
          legend.position = 'bottom',
          axis.line = element_line(size = 0.15)) +
    labs(x ='Actual', y='Predicted', title = dependent_variable_name) +
    expand_limits(x = 0, y = 0) +
    ggplot2::annotate('text', x = x_coord, y = y_coord,
                      label = paste(' r=', annot_df$cor, ', RMSE=', annot_d
f$rmse, sep = ''),
                      hjust = 0,
                      size = label_size) +
    guides(colour = guide_legend(override.aes = list(size=2)))

  return(p)
}

# make a plot

CTL<-plot_pred_vs_actual(test_CTL, lm0,
                         "CTLemra_WBC",
                         'CTLemra/WBC', label_size = 2.5)

CTL_norm<-plot_pred_vs_actual_cube(test_CTL_norm, lm1,
                                   "CTLemra_WBC_cube_root_trans",
                                   'CTLemra/WBC (cube_root_trans)', label_s
ize = 2.5)
CTL+CTL_norm

TH<-plot_pred_vs_actual(test_TH, lm2,
                        "Themra_WBC",
                        'Themra/WBC', label_size = 2.5)

TH_norm<-plot_pred_vs_actual_cube(test_TH_norm, lm3,
                                  "Themra_WBC_cube_root_trans",
                                  'Themra/WBC (cube_root_trans)', label_siz
e = 2.5)

TH+TH_norm
```

Residuals

```
CTL_res<-plot(fitted(lm0), res_0)+abline(0,0, col="red")
CTL_res_qq<-qqnorm(res_0)
CTL_res_den<-plot(density(res_0))
CTL_norm_res<-plot(fitted(lm1), res_1)+abline(0,0, col="red")
CTL_norm_res_qq<-qqnorm(res_1)
CTL_norm_res_den<-plot(density(res_1))

TH_res<-plot(fitted(lm2), res_2)+abline(0,0, col="red")
TH_res_qq<-qqnorm(res_2)
TH_res_den<-plot(density(res_2))
TH_norm_res<-plot(fitted(lm3), res_3)+abline(0,0, col="red")
TH_norm_res_qq<-qqnorm(res_3)
TH_norm_res_den<-plot(density(res_3))
```

NON-EXCLUSIVE LICENCE TO REPRODUCE THE THESIS AND MAKE THE THESIS PUBLIC

I, Alexandra Elsakova,

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

HARNESSING EPIGENETIC CHANGES TO ESTIMATE IMMUNE CELL LEVELS

supervised by Liina Tserel and Ahto Salumets,

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Alexandra Elsakova

27/05/2022