University of Turin

PhD Program in Complex Systems for Life Sciences

XXXII Cycle

TITLE

Evaluation of DNA methylation in leukocyte subpopulations associated with tobacco smoking exposure

Candidate

Dr. Giulia Piaggeschi

Tutor

Prof. Francesca Cordero

Co-tutors

Dr. Silvia Polidoro

Dr. Alessio G. Naccarati

Abstract

Cigarette smoking is well-known to have adverse effects on cell-blood compositions. In healthy individuals, several studies showed an increase in the overall number of leukocytes in current smokers compared to former and never smokers. Epigenome-wide association studies (EWAS) suggested a DNA methylation involvement in the regulation of smoking-related pathways and diseases. These studies, mainly conducted on whole blood, have repeatedly reported differentially methylated CpG sites between smokers and never smokers. After smoking cessation, some of these differences in methylation levels persist altered over time while others return to those levels observed in never-smokers. Importantly, differential methylation may partly reflect smoking-related shifts in leukocyte distribution. The majority of EWAS have addressed this issue adjusting their analyses for a leukocyte distribution estimated by the Houseman algorithm. However, this algorithm does not take into account the small immune cell fractions that might still have a role in confounding the results. In this respect, this study aims at clarifying how tobacco smoking impacts both the cell-blood proportions of main leukocyte subpopulations and DNA methylation levels. To investigate these aspects, different molecular epidemiological, statistical and bioinformatics approaches have been used.

We recruited 288 healthy volunteers, aged between 35 and 70 years old for evaluating the association between self-reported smoking habits (current, former and never smokers) and the cell-count distributions of nine leukocyte subpopulations (namely: CD4+T-helper, CD8+T-cytotoxic, CD16/CD56+NK-cells, CD3+T-cells, CD56/CD3+ NKT-cells, CD19+B cells, CD14+monocytes, neutrophils, and eosinophils) as well as of their GPR15 cell receptor as smoking marker quantified by flow cytometry. Current smokers showed a significant lower NK cell count, and an increase of GPR15+cell-type in both T cell (CD3+, CD4+and CD8+) and B cells and a decrease of GPR15+cell-type in monocyte despite the cohort included only light smokers (< 15 cig/day).

We have performed a similar analysis on 358 participants of the Twins-UK Cohort, for whom the cell-frequencies of 41,701 leukocyte-subtypes were available. We have found that active tobacco smoking is associated with increased frequencies of circulating CD8+ T cells expressing the CD25+ activation marker (CD25+). Moreover, we identified novel associations between smoking status and relative abundances of CD8+ CD25+ memory T cells, CD8+ memory T cells expressing the CCR4 chemokine receptor and double-positive CD25+ (DP) T cells. We also observed in current smokers an increase of class-switched memory B cell isotypes IgA, IgG, and IgE relative frequencies and a decrease of circulating CD4+ T cells expressing the CD38 activation marker. Also Finally, using data from 135 former female smokers, we showed that the relative frequency of immune traits associated with active smoking is wholly restored after smoking cessation, with some exceptions for CD8+ T cells (CD8+ CD25+ T cells, CD8+ memory T cells CD25+ and CCR4+) which persist partially altered.

To study smoking-DNA methylation profiles at cell-type level, we have analysed target bisulfite sequencing data (target BS-Seq). However, to the best of our knowledge, specific tools and pipelines for analysing target BS-Seq data are still lacking. Thus, we compared the performance in DNA methylation detection of, BSMAP and Bismark which are the most used aligner and methylation callers tools. To achieve this, we have generated *MethylFastQ*, a new tool to create artificial target bisulfite data. We have tested BSMAP and Bismark tool on synthetic and real datasets showing that BSMAP was more performant during the alignment and methylation recall in datasets with low-quality reads. Furthermore, we applied our developed pipeline to investigate the smoking-related DNA methylation signatures in a pilot study on monocytes and B cells. We observed that these cell-lineages shared a low number of differentially methylated genes, and a high number of these genes were cell-type specifics. Common and cell-type-specific genes were not enriched in particular biological pathways, despite the presence of genes involved in pathways of cancer.

Acknowledgments

I would like to express my sincere gratitude and appreciation to people who helped me during my Ph.D period.

First of all, I want to thank all my supervisors Dott.ssa Francesca Cordero, Dott.ssa Silvia Polidoro e Dott. Alessio G. Naccarati for their supervision, scientific support, and continuous teaching during these years.

I want to thank Prof. Paolo Vineis for welcoming me in own research group and giving me the opportunity to grow in this field.

I want to thank Dr. Marco Beccuti for its support and teaching in the computational analyses.

I want to thank Dr. Mario Falchi and Dott.ssa Alessia Visconti for their supervision and collaboration during my Ph.D visiting in London.

I want to thank Dr. Davide Brusa and Dott.ssa Laura Conti for their support and teaching on flow cytometry analysis.

I want to thank all my colleagues from Vines Unit of IIGM (ex-Hugef), Valentina, Sonia, Chiara, Giovanni, Manuela, Barbara e Antonio, in particular, those that collaborated actively in the present project and all for sharing with me this path.

A special thanks to all my colleagues of q-BioGroup (Giulio, Laura, Greta, Simone and Nicola) for their collaboration, sharing ideas and the desire to grow together.

Finally, I want to thank Association of voluntary Italian blood donors (AVIS) of Turin, in particular, Dott. Roberto Ravera and Dott.ssa Anna Alpe for their precious collaboration in the realization of this project.

Contents

1 Introduction

Tobacco use represents the largest epidemic of the world. Tobacco-related mortality is set to increase to almost 1 billion deaths during the 21st century, most of them in low-income countries. In 2014, 50 years of research on tobacco smoking were celebrated. During these years, the mechanisms behind tobacco smoking were extensively studied increasing our knowledge in this field, but even today some of these mechanisms are not fully understood and predictive markers for lung cancer risk, especially in former smokers, are lacking.

DNA methylation is an epigenetic modification, cell-type specific, widely studied to understand the genetic mechanisms of gene expression whose alterations are at the base of several human diseases, including cancers related to environmental or lifestyle exposures. DNA methylation profiles also represent a powerful biomarker for diagnosing diseases and guiding treatment.

In the last decade, alteration of the methylation profile of DNA from peripheral blood associated with cigarette smoking has been described but limited to whole blood samples. On the other hand, we know that tobacco smoking causes inflammation process inducing leukocyte subpopulation shift in blood and it may confound the results of the association between DNA methylation and smoking exposure. In association studies, array platforms represent the golden standard techniques that interrogate more than 800,000 CpG sites in an elevated number of samples at the same time. However, they present some limitations for estimating the DNA methylation levels. Nowadays, the decrement of the costs of deep sequencing technologies is leading their extensive use to study several aspects of the transcription regulation. Among the others, the DNA methylation can be easily profiled in whole blood and also in different immune cells. This will lead to understand the influence of smoking at a single cell-type level and to identify a novel potential DNA biomarker of smoking exposure.

1.1. Tobacco smoking

1.1.1 Epidemiology

In 2000 the number of smokers in the world was around 1.22 billion, and it is estimating will grow up to 1.9 billion in 2025 (1; 2)

One-in-five (20%) adults in the world smoke tobacco. In the map reported in **Figure 1.1** shows the top five countries where more than 40% of the population smoked in the year 2016.

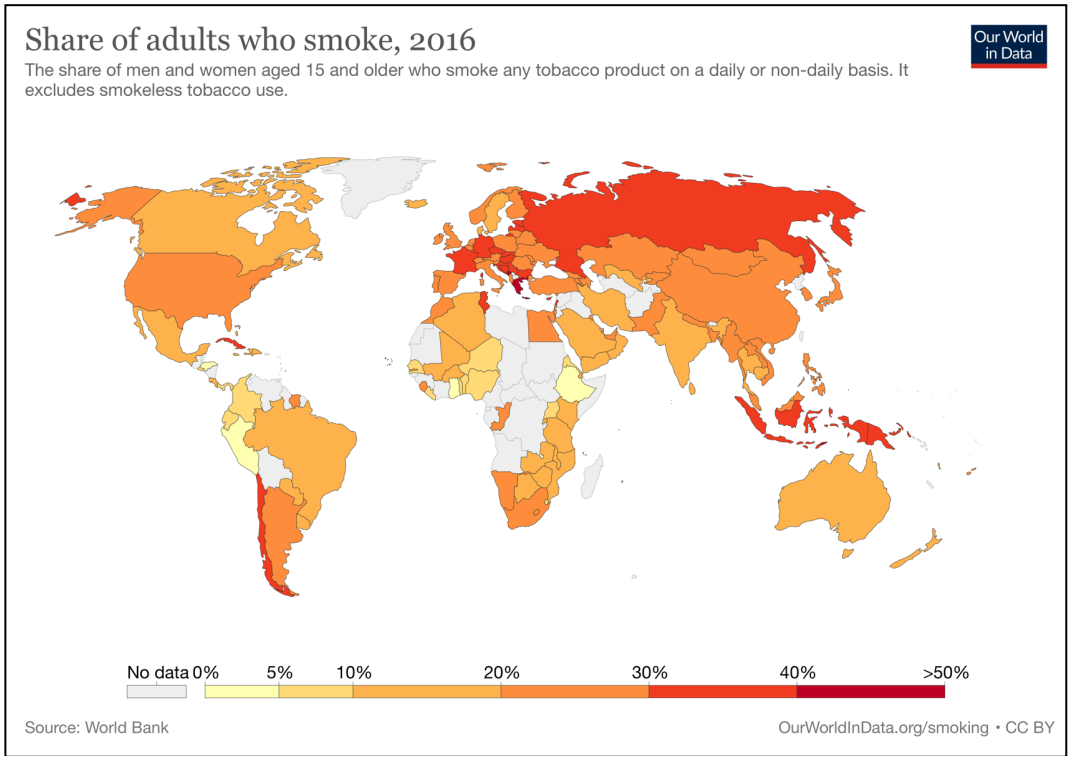


Figure 1.1 Prevalence of smoking across the world in the 2016'. From (1)

Three are Pacific Islands (*i.e.,* Kiribati (47%); Timor (43%); Nauru (40%)) and two are in the Balkans (*i.e.,* Montenegro (46%) and Greece (43%)).

The places where many people smoke are clustered in two regions. South-East Asia and the Pacific islands and Europe – particularly the Balkan region – but also France (33%), Germany (31%), and Austria (30%). In some countries very few people smoke: in Ethiopia, Ghana, Peru and Honduras less than 5% smoke. In Honduras, it's one smoker every 50th person (1).

Tobacco smoking is the major risk factor of mortality and morbidity in the world. It has been associated with at least 17 types of human cancers and several cardiovascular, respiratory and autoimmune diseases. There is wide variety of smoking tobacco products on the world market, including cigarettes, cigars and bidis (*i.e.,* Hand-rolled Indian cigarette) (2).

According to the last data of the World Health Organization (WHO), approximately 8 million people die globally every year for smoking-related diseases, with a high percentage of those deaths occuring in low- and middle-income countries. Moreover,1.2 million deaths annually are caused by second-hand smoking in never smokers. In adults, second-hand smoke causes serious cardiovascular and respiratory diseases, including coronary heart disease and lung cancer. In infants, it raises the risk of sudden infant death syndrome. In pregnant women, it causes pregnancy complications and low birth weight (2; 3).

WHO has estimates that 12% of all deaths among adults could be attributed to tobacco use. Figures for men and women are 16% and 7%, respectively. Tobacco smoking is one of the major problems in public health because it imposes enormous economic costs to society for health-care needs.

In the last decade a declining prevalence of tobacco smoking has been registered worldwide, with an opposite trend for low and middle-income countries (2)(**Figure 1.2**)

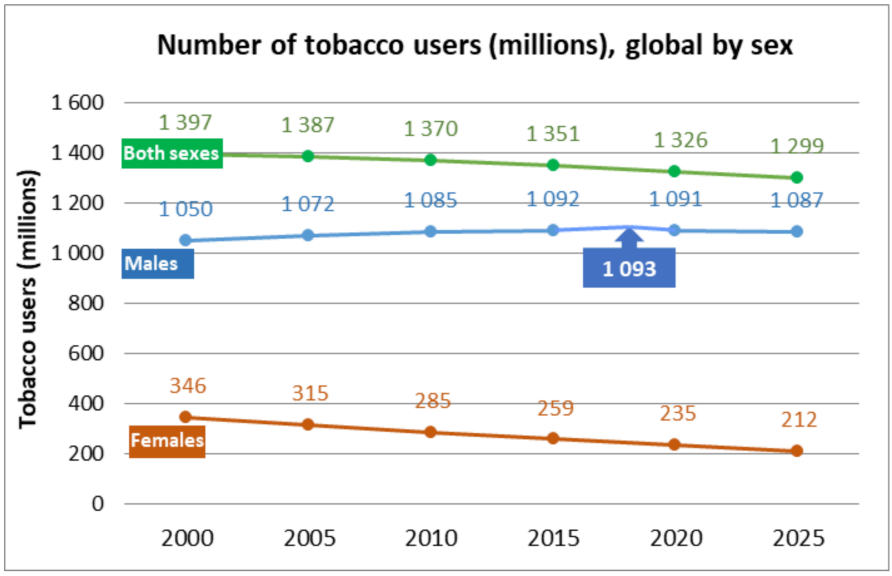


Figure 1.2 Trends in the global number of tobacco users. The total number of tobacco users for both sexes combined has steadily declined world-wide over the period 2000-2015. From 2000 to 2018, the number of male tobacco users in the world was increasing each year, and the peak was in 2018 with 1093 millions of tobacco users. The number of female tobacco users has been declining over the period 2000-2015 and it is expected to continue to 2025. From (2).

Despite these results, the prevention and reduction of tobacco smoking consumption represent an important challenge for all countries in the world. Indeed, since 2005 severe smoking restrictions are currently in force in 181 countries by the treaty of WHO Framework Convention on Tobacco Control, aiming at reducing the damaging health and economic impacts of tobacco consumption. Moreover, from 2013 the reduction of tobacco smoking is one of the key objectives to reduce of 30% the prevalence of premature mortality from noncommunicable diseases (*i.e.,* cardiovascular diseases, cancers, diabetes and respiratory diseases) in persons aged 15 + years by 2025 (2; 3).

The prevalence of tobacco use in the word is largely a male phenomenon (35% of men in the world *vs* 6% of women): in Europe the gap in prevalence between male and female adults is very small (<5%) in countries such as United Kingdom, Denmark, Ireland and Finland. In Europe, the number of cigarettes smoked per day is reduced in current smokers over 15 years of age. Among them, only 6% consume at least 20 cigarettes per day while around 13% consume less than 20 (4).

In Italy, smokers are 11.6 million, representing about the 22% of the population aged >15years. Of these, males are around 7 million. In the range of age between 25 to 44 years the 36.3% of smokers are males, in the range from 45-64 years old the prevalence are women (22.9%), while over 65 the prevalence in both sexes is similar. Although, current smokers smoked in average 11.6 cigarettes/day, more than 21% are heavy smokers that smoked over 20 cigarette/days. Furthermore, quite alarming is the 11.1% of smoking prevalence among children between 14 and 17 years old (4).

Recently, smoke-free policy adopted in Europe had increased the electronic cigarette (*i.e.,* "e-cigarettes", "vape pens", "e-hookahs") consumption among the population. E-cigarettes are used mainly to aid people to quit smoking. The effectiveness of e-cigarettes as a cessation aid is still being researched, but it seems that a proportion of smokers who are trying to quit may be using it as such (2). Moreover, e-cigarette use in smokers is attributable to circumvent smoking bans by using e-cigarettes in places where tobacco smoking is prohibited, thus attenuating the impact of smoking bans (5). However, long-term health effects of e-cigarettes are still unknown, and further research is required. Chemical and toxicological studies togheter with clinical investigationshave led various authors to conclude, with more or fewer caveats, that e-cigarettes are not harmless but are generally less dangerous than cigarettes (2; 5).

Therefore, despite the progress made, the final objective of reducing the tobacco epidemic is still far away. Prevention is an essential weapon in this context, and the research of biomarkers to establish in current smokers the risk to incur tumour and smoking-related diseases represents a good opportunity to overcome this emergency.

1.1.2 Cigarette smoke composition and toxicity

Destructive effect of cigarette smoking derives from cigarette composition**.** Cigarette contains more than 7,000 chemical compounds including direct carcinogens (*e.g.,* methylcholanthrene, benzo[a]pyrenes and acrolein), toxins (*e.g.,*carbon monoxide, acetone, ammonia, nicotine, and hydroquinone), solid with catalytic activity and oxidants (*e.g.,* superoxide and nitrogen oxides). All these

components are contained in the gaseous and particulate phase of the cigarettes (1). Cigarette smoke directly impacts on lungs, by inhalation of fresh smoke. Fresh smoke contains a billion of oxidative moieties that may generate secondary oxidative metabolites and DNA adducts by the activation of an oxidative burst and nitric oxide synthase in the host. Reactive oxygen species (ROS) directly affects the lung cells activating macrophages, epithelial cells, and neutrophils by exerting a pro-inflammatory effect, or through direct damage of lipids, proteins, nucleic acids, and organelles. Induced damage can change the normal function of these critical targets. Increased levels of ROS contribute to apoptosis, inactivation of proteases (such as a1-antitrypsin) and activation of metalloproteases, which immediately contribute to the degradation of lung tissue (6). Moreover, cigarette smoke leads to a significant reduction of glutathione, a major antioxidant present in the lung. Changes in the redox status within the cell initiate the lung inflammatory responses through enhancement of the respiratory burst in phagocytic cells, regulation of intracellular signalling, chromatin remodelling (histone acetylation/deacetylation) and activation of redox-sensitive transcription factors, such as nuclear factor-kB (NF-kB) and activator protein-1 (AP-1). The latter are critical to gene expression of pro-inflammatory mediators such as interleukin (IL)-8, IL-6, and tumour necrosis factor-α (TNF-α) which links cigarette smoke exposure with altered cytokine production. Other pathophysiological mechanisms by which cigarette smoke can alter cytokine gene transcription rely on smoke-induced changes to the epigenome, such as DNA methylation, expression of microRNAs and histone modifications (7).

1.1.3 Cigarette smoke and the immune system

The first putative influence of tobacco smoking on the immune and inflammatory processes was identified in the 1960s (8). This topic has received more attention later on, after the discovery of the association between cigarette smoking and autoimmune diseases. At the moment in literature are present, a large number of studies regarding the molecular and cellular mechanism of the consequences of smoking on the immune system but, some finding across these studies are inconsistent and controversial due to the nature of experiments (*i.e.,*

human *vs.* animal studies, and *in vivo vs in vitro* studies), type and duration of smoking exposure and intrinsic variability related to characteristics of the population studied (*e.g* gender, age, ethnicity). Together, all of these levels of complication made difficult the comprehension of the smoking effects on the immune-inflammatory system, even today (8; 9).

Cigarette smoking alters the development, cytokine production, and effector function of both **innate immune cells** (*i.e.,* macrophages, dendritic and NK cells), and **adaptive immune cells** (*i.e.,* cytotoxic CD8+ T cells, CD4+ helper T cells and B cells). It can be associated with the release but also the inhibition of pro-inflammatory and anti-inflammatory mediators (10; 6; 7)(**Figure1.2**). These effects are triggered by single or combinate actions of tobacco-compounds on immune system components. Cigarette smoke promotes inflammation by inducing the production of pro-inflammatory cytokines, such a TNF-α, IL-1, IL-6, IL-8 and granulocyte-macrophage colony-stimulating factor (GM-CSF), and increasing the accumulation of immune cell in the airways. While, the suppressive proprieties are mainly attributed to nicotine, hydroquinone and carbon monoxide in the smoke. Nicotine shows both pro-inflammatory and immunosuppressing capability. Inhibitory effects of nicotine were associated to its inhibitory effect on alfa7 nicotinic acetylcholine receptor (a7nAChR) found in macrophages, T, and B cells that suppressing cytokines production (IL-6, IL-10, and IL8). Importantly, this activation was shown to suppress Th1 and Th17 responses with reciprocal shift towards the Th2 lineage (7). But nicotine, also favours the increase of circulating neutrophils with a reduction of their functionality. This effect seems to be related to the secretion of catecholamines induced by nicotine. Studies (10; 11). suggested that catecholamines may be the responsible of release of leukocytes into the circulation and stimulation of hemopoietic system related to cigarette exposure Nicotine also exerts a protective action against of free oxygen radicals. Thanks to this anti-inflammatory effect, nicotine and nicotine-metabolites could represent a promising pharmacological strategy for treatment inflammation-related diseases, such as obesity and ulcerative colitis (11).

Moreover, this is further complicated by compounds demonstrating both pro-inflammatory and immunosuppressive properties, such as acrolein, another major

component of tobacco smoke. While inhalation of acrolein promotes airway hypersensitivity responses, it may stimuli neutrophils accumulation in the airway, thereby contributing to immune tolerance (7)

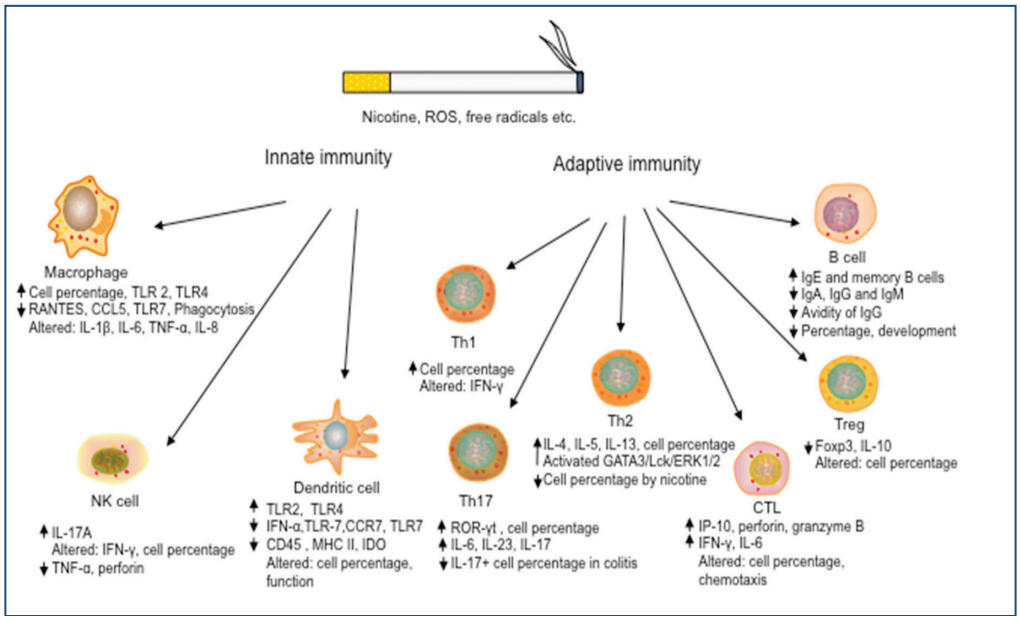


Figure 1.3. Smoking effects on the innate and adaptive immune cells. Cigarette smoking alters the development, cytokine production, and effector function of both innate immune cells (macrophage, NK cells, Dendritic cells) and adaptative immune cells (cytotoxic CD8+ T cells, CD4+ Th cells, regulatory T cells and B cells) leading to pro-inflammatory response and/or dysfunction of immune cells. Adapted by (9).

Tobacco smoke alter several **adaptive immune functions**, including leucocytosis. Peripheral blood of humans exposed to cigarette smoking showed an elevated percentage of total lymphocytes, in particular leads CD8+ /CD4+ T cells ratio. Indeed, in heavy smokers CD8+ T cells increased, whereas the percentage of CD4+ T cells decreased. Also, the numbers of memory T cells and class-switched memory B cells were significantly and positively correlated with smoking habits. Since the process of class-switch recombination in B cells results from repeated antigen recognition, their presence in smokers suggests that cigarette smoke is potentially capable of generating neo-antigens derived from damaged lung tissue or smoke fume components in a chronic manner. Other deleterious effects of cigarette smoke exposure comprise the suppression of immunoglobulin production. Though the secretion of IgA, IgG and IgM appears to be down-regulated in peripheral blood and saliva of smokers, this suppressive effect does

not affect IgE synthesis. Indeed, IgE levels were found increased in smokers (6; 9). Endotoxin (lipopolysaccharides) is one of the most potent inflammatory agents of smoke and is related to elevated levels of IgE in smokers with subsequent development of atopic diseases and asthma (9).

Smoking also affects in a similar way the **innate immunity**. Macrophages are the main lung cell population. They are the first line of cellular defence against exogenous pathogens via phagocytosis and digestion, and recruit/activate lymphocytes via their antigen-presenting ability. Chronic smoke exposure causes an elevated number of macrophages in airways lumen, exhibiting a low maturation level by high expression of CD14 markers (monocytes expression marker), and changes in their morphology. These cells show impaired functions, such as a strong inhibitory effect on lymphocyte proliferation and NK cells, and disability to kill intracellular bacteria (6).

In particular, tobacco smoking exposure has been reported to both suppress and stimulate the activity of NK cells. NK cell activity in peripheral blood was reduced in smokers compared with non-smokers. These alterations appear to be reversible, since a recovery period of six weeks after smoking cessation brought the cytotoxic activity of NK cells back to the levels of never-smokers. NK cells from long-term smokers display a decreased intracellular IL-16 concentration. This depletion of the CD4+-recruiting cytokine strongly suggests that long-term smoking may impact immune responses at the systemic level, and that NK cells are involved. There is strong evidence showing direct negative effect of cigarette smoke on NK cell cytolytic capacity, as well as on their ability to produce inflammatory cytokines in response to microbial agents. Moreover, cigarette smoke exposure caused increased accumulation of primed/activated CD69(+) NK cells in parenchymal and mucosal locations in the airway. The priming and activation of NK cells is believed to result from crosstalk between NK and sentinel cells, such as DCs, and CCR4 appears to be a possible promoter of NK/DC interaction (11; 8).

Finally, cigarette smoke alters the number, distribution and function of dendritic cells (DC) by fostering Th2 response and repressing Th1 cytokine productions. It is traduced with a reduced priming capability of DC, reduced endocytic and

phagocytic activity and reduce secretion of IL-10, IL-12. Cigarette smoking also induced the dramatic increase of Langerhans cells, a subtype of myeloid DCs, in alveolar parenchyma (6; 9).

In summary, tobacco smoking exerts pro-inflammatory and immune-suppressive properties in both innate and adaptative immune system. These mechanisms are mainly attributable to cigarette components such as nicotine and acrolein. Studies in blood reported the smoking effect on primary leukocyte subtypes (*i.e.,* T cells subclass, B cells, NK, Monocytes and Neutrophils) but it influences also all circulating cells including those less frequents (*i.e.,* Th1, Th2, Th17 and Memory cells). In lung tissue, smoking alters the macrophage functions and it suppresses both activity and production of NK cells. While, its effects on dendritic cells alter as a consequence the production of cytotoxic-cells Th1 and Th2.

1.2. DNA methylation

1.2.1 DNA methylation: general overview

DNA methylation is the covalent addition of a methyl group to DNA molecule. In the human genome the most common form of DNA methylation is the addition of methyl group in the position 5' of a cytosine (C) when this nucleotide occur next to guanine (G) forming a **CpG sites**. There are around 28 million CpG sites in the human genome, ~70% of them are generally methylated. DNA methylation is not uniformly distributed over the genome, but it is associated with CpG density. Indeed, has been reported that unmethylated cytosine are usually in CpG-rich regions, called **CpG islands,** and tend to be methylated in CpG-deficient regions.

Non-CpG methylation has a functional role in plants, fungi and in embryonic stem cells, while its function in mammals is currently unknown. DNA methylation occurs in three different contexts: CpG or CG, but also CHG and CHH, where H = {A, C, T}. In human it rarely occurs in CHG and CHH contexts ~3%.

DNA methylation is an important **epigenetic modification**, it is involved in many biological processes including transcriptional activity, genomic imprinting, development, and differentiation in a cell-type-specific manner. Additionally, it changes during life-course and it is reversible.

The function of DNA methylation depends on the position where it occurs on the gene. Studies have emphasized that when it occurs in a transcription start sites is associated with gene repression and silencing, while demethylation is related to gene expression and activation. Whereas, recently, with the advent of high throughput techniques, it has been demonstrated that exons are more methylated than introns, suggesting a role for methylation in regulating splicing (12; 13).

Alterations in DNA methylation levels are widely studied for their association with environmental exposure (*i.e.,*smoking, UV radiations, Bisphenol A), common human diseases, autoimmune disorders and especially with cancer. Given the plasticity, the stability and highly specific nature of DNA methylation, it represents a powerful **molecular biomarker** for risk stratification and disease diagnostics. In fact, the DNA methylation biomarkers are commonly used to support the clinical decisions in various cancer, as well as for early prevention, diagnosis, the prognosis of disease and drug response (14).

1.2.2 DNA methylation, Smoking exposure and Smoking-DNA methylation in blood cell-types

In the last decade, the advent of high-throughput techniques allowed us to study DNA methylation at the genome-wide level in large cohorts of samples (15). For examples, in European Prospective Investigation into Cancer and Nutrition (EPIC)-cohort, a large prospective study conducted across several European countries, was extensively studied the smoking effects on DNA methylation levels (16; 17). Indeed, Epigenome-Wide Association Studies (EWASs) showed that DNA methylation plays a pivotal role in the pathways of smoking and smoking related. These studies, mainly conducted on whole blood, found that three intragenic CpG sites: cg05575921 (***AHRR***), cg03636183 (***F2RL3***), and cg19859270 (***GPR15***) as well as other CpGs within intergenic regions (*2q37.1* and *6p21.33*) are differentially methylated in smokers in comparison to never smokers. These significant smoking-related differences in DNA methylation reflect not only current but also lifetime or long-term exposure to active smoking (16; 18; 17). These genes showing a functional role in response to tobacco smoking. *AHRR* is the repressor of the aryl hydrocarbon receptor, a key regulator of the relationships between the cell and the external environment, including the effect of stressors such as dioxins and polycyclic aromatic hydrocarbons (components of tobacco smoke). Current smokers exhibited a 19% lower methylation level at cg05575921 (*AHRR*) compared to never smokers (19). The second gene *F2RL3* is a relevant gene that encodes the protease-activated receptor-4, involved in both cardiovascular and neoplastic diseases. Finally,

GPR15 is a surface membrane-bound G protein-coupled receptor 15 that regulates cell migration in response to inflammatory insults (20).

In former smokers, some of these differentially methylated sites appear to return to levels of those in never-smokers, whereas other smoking-related CpG sites appear to persist 30 years after smoking quitting (18; 17). Furthermore, these findings were tested in pre-diagnostic blood samples of lung cancer observing the significant association with lung cancer risk for cg05575921 (*AHRR*) and cg03636183 (*F2RL3*), showing a decreasing of lung cancer risk with years after smoking cessation (19; 21) (**Figure 1.4**).

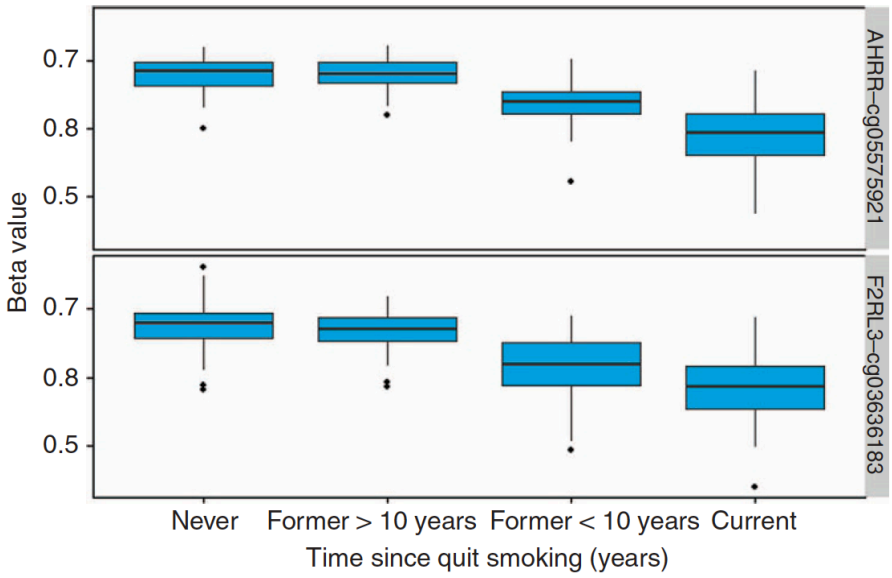


Figure 1.4. Association between smoking cessation and the mean methylation levels of cg05575921 (AHRR) and cg03636183 (F2RL3) in pre-diagnostic blood samples from lung cancer. After smoking cessation, methylation levels increase and after 10 years since quitting appear similar to those of never smokers. The risk of lung cancer decreases substantially after smoking cessation. From (19).

The recent study from Stuve *et al*. (22) tested the consistency of results from association analysis between smoking and DNA methylation measured in normal lung tissue. They found the five CpG loci previously reported as hypomethylated in smokers-blood. It suggests that blood-based biomarkers can reflect changes in the target tissue of these loci and represent a valid diagnostic marker. Together

these findings, supporting the evidence that smoking leads to DNA methylation changes measurable in peripheral blood and useful as predictive markers for lung cancer risk, especially in former smokers.

The main limitation of these studies is the fact that DNA methylation patterns are highly cell type-specific, and analysing purified cells is important to avoid the confounding effects present in a surrogate tissue like whole blood (19; 21). The isolation of specific lymphocyte subpopulations requires additional steps, such as flow cytometric separation using cell-surface receptor antibodies. In epidemiological studies with a large samples size it is not feasible due to several issues: *i)* a large volume of fresh blood difficult to find for each sample; *ii)* elevated costs for the huge volume of antibodies required to sort cells; *iii)* time consuming for the experimental procedures required in cell-sorting.

To overcome this issue, currently, in EWAS the **Houseman algorithm** is frequently adopted (23). This algorithm allows estimating the percentages of cell composition by a deconvolution approach based on DNA methylation signature for each cell-types. However, minor immune leukocyte subsets such as regulatory T cells and NK cells, which are implicated in smoking and in other autoimmune diseases are not taken into account by this method that allows the measurements of main leukocytes subtypes (*i.e.,* B cells, T class, monocyte and granulocyte). Moreover, a recent study has shown that DNA hypomethylation at site cg19859270 within the *GPR15* gene in smokers is caused by the high proportion of CD3+ GPR15+ expressing T cells instead of direct effect of tobacco smoking compounds on DNA methylation (24).

1.2.3 Next Generation Sequencing techniques for methylome profiling

Recent advances in Next-Generation Sequencing (NGS) have allowed to map DNA methylation genome-wide, at single-base resolution and in a large number of samples. The new methods create large opportunities for epigenome research, but they also pose substantial challenges in term of data processing, statistical analysis and biological interpretation of observed differences (15).
Several technologies have been developed to investigate genome-wide DNA methylation changes at single base resolution. The distinguish feature among

these technologies is in treatment of DNA to detect methylation. In particular, there are three main approaches: a) endonuclease digestion, b) affinity enrichment and c) bisulfite conversion (25). The method of choice depends on the biological question.

The bisulfite conversion technologies use **bisulfite treatment** to create an artificial transition which converts unmethylated cytosine (C) to thymines (T) while methylated Cs remains unchanged. Bisulfite-treated DNA can be analysed by both microarray or NGS platforms. (26)

- *Methylation microarray platforms*

Microarray-based methodologies have been and, still today are, widely used in large scale study such as EWAS. They combine bisulfite conversion with specially designed genotyping microarrays for measuring DNA methylation levels at preselected cytosines from genome. Human Methylation450K contains approximately 480k CpG sites, covering 99% RefGen (hg19) and 96% CpG islands. These CpGs cover promoter regions, CpG islands and shores and selected CpGs outside coding regions. The methylation levels where measured using beads labelled with two different dye colours for methylated (green) and for unmethylated (red) cytosines. The main advantages are lower costs compared to whole genome bisulfite sequencing and the experimental procedures are easier and faster for elevate numbers of samples compared to library preparation. Whereas, the microarray platforms are based on florescent signals, which are less sensitive and more prone to technical variation such as dye bias, batch effects and probe design bias (26).

- *Bisulfite DNA methylation sequencing (BS-seq)*

Whole Genome Bisulfite Sequencing (WGBS) represents the gold standard for DNA methylation studies, especially for the identification of differentially methylated regions among multiple samples. After bisulfite conversion the treated reads are then sequenced on the common NGS platform. Sequence alignment on a reference genome enables the detection of the methylated cytosines at single-base resolution.

Basically, for the human genome, with about 28 million of CpGs, at least 1 billion of 100 bp reads with approximately 30x average coverage are needed. Some sites are not covered or present a low coverage (1-10x). For these sites to estimate methylation levels is impossible given that sufficient coverage for downstream analysis is generally 10x.

The main challenge in BS-Seq data analysis arises from its low sequence complexity. Bisulfite conversion reduces most of the genome to a three-nucleotide alphabet, since most of cytosines are not methylated. Thus, sequence alignment becomes a more difficult task and requires specialized tools.

The key advantages of this technology are its comprehensive genome coverage, high qualitative accuracy and reproducibility. While the principal limitations of WGBS are the costs and the difficulties in the analysis of sequenced data. However, since only a small portion of the genome is differentially methylated, often WGBS is not necessary.

Enrichment-based methods offer the opportunity to sequence methylated fractions of the genome in a less expensive way, allowing to increase the sequencing coverage and, therefore, the precision in detecting differentially-methylated regions. The most used Enrichment-based techniques are: *i)* reduced representation bisulfite sequencing (RRBS), it isolates CpG-rich regions through enzymes that recognize CCGG sites and cut the genome in those points; and *ii)* **targeted bisulfite sequencing** (or target enrichment sequencing) works by capturing genomic regions of interest by hybridization to target-specific biotinylated probes, which are then isolated by magnetic pulldown.

Bisulfite-based methods are fairly accurate and reproducible. The major source of bias and measurement error is due to incomplete bisulfite conversion. It is important to measure bisulfite conversion of non-methylated cytosines incorporating controls for bisulfite reactions. Usually, during libraries preparation a bisulfite control represented by unmethylated DNA sequences was included into the samples.. Moreover, bisulfite over-treatment can also cause problem. This process degrades DNA and can lead to methylated cytosines conversion to thymines, which results in methylation underestimation (25; 26).

Basically, the protocol for targeted bisulfite library preparation involves the classical steps used for enrichment based-methods, with the addition of bisulfite conversion step., The steps are: I) genomic DNA fragmentation; II) adapter ligation; III) treatment with sodium bisulfite; IV) PCR amplification; V) target regions selection, VI) targeted amplification and sequencing. **Bisulfite treatment** includes DNA denaturation step and PCR amplification to convert uracils into thymines and to amplify the bisulfite-converted library. Thus, for each double-stranded DNA fragment, bisulfite treatment followed by PCR amplification generates four distinct strands: the bisulfite version of forward and reverse strand and their reverse complements.

Two library protocols have been developed for constructing bisulfite converted libraries: the non-directional protocol and the directional protocol.

In the **non-directional protocol,** all four possible bisulfite DNA strands are sequenced at roughly the same frequency. While, in the **directional protocol**, the sequencing reads will correspond to a bisulfite converted version of either the original forward or reverse strand. Strands complementary to bisulfite forward or reverse strand are generated in the PCR step, but they will not be sequenced as they carry the wrong kind of adapter.

1.2.4 Computational aspect of Bisulfite Sequencing data analysis

Bisulfite sequencing has become the gold standard to quantitatively detect the methylation pattern at single-base resolution. The bisulfite modification makes methylation sequencing analysis a challenging task from computational point of view. The main reasons are:

1. The search space is significantly increased compared to the original reference genome. The bisulfite treatment destroys strand complementarity, because the conversion occurs only on cytosine (Cs). As a result, there will be four different strands after amplification: bisulfite forward and reverse strand, and their respective reverse complements. **Figure 1.5** shows the four strands produced starting from one double-stranded fragment.

2. Cytosine (C) to thymine (T) mapping is asymmetric. Each read can theoretically exist in all possible methylation states. So, a T in the bisulfite read could be mapped either a C or T in reference genome, but not vice versa. Figure 1.4 highlights which Ts map on Cs, but this information is unknown in reality.
3. It can be hard to distinguish a convert cytosine from a sequencing error.

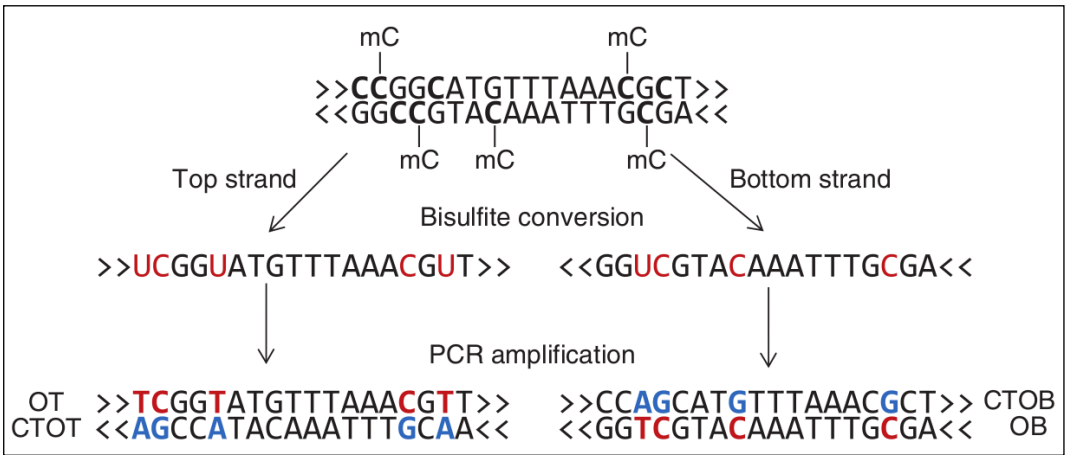


Figure 1.5. Production of bisulfite fragment in the non-directional protocol. Bisulfite treatment converts unmethylated cytosines in uracils. PCR amplification converts uracils in thymines and produces the complement of the treated fragments. As a result, at the end there are four fragments: OT and OB, that are the complementary fragments of OT and OB, respectively. In case of directional protocol, only OT and OB are produced. Adapted from (27)

The first point underlines an additional problem with respect to classic NGS read alignment. A read must be aligned on both forward and reverse strand of the reference; to do that, it is sufficient to calculate its reverse complement.

In case of bisulfite reads, this is not sufficient because strand complementarity does not hold anymore after treatment. As the strand identity of a bisulfite read is a priori unknown, each read must be explicitly aligned to both the forward and the reverse strand of the genome.

The last two points stress the need of an efficient mechanism to allow Cs in the read to be mapped on either a C or T in the genome, and at the same time, to allow the presence of other mismatches.

To overcome these issues, a variety of tools for mapping bisulfite converted reads have been proposed. They can be classified into two major categories in terms of alignment strategies: the **three-letter approach** and the **wild card approach**.

Three-letter aligners first reduce the reference genome into two in-silico variants: in the first one all Cs are converted to Ts, in the second one all Gs are converted into ss (equivalent to a C-to-T conversion on the reverse strand). Reads are converted in a similar manner and then they are mapped to both reference genome variants using a standard read aligner, such as BWA (28) or Bowtie (29). There are many tools that implements this algorithm, such as Bismark (27), BS-Seeker (30) and others. Each of them may use different strategies, as the basic read aligner or the indexing algorithm. Wild card aligners do not convert reference or reads explicitly but treat Cs and Ts in the reads as matches for Cs in the reference genome, enumerating all C-to-T combinations of the read. Alternatively, they used a similarity score matrix that contains a positive match score between a T in a read and a C in the reference genome. This kind of approach is implemented by BSMAP (31), BRAT (32) and other tools.

2 Aim of the thesis

The object of this work reflects the growing interest in clarifying the smoking effects on DNA methylation of the main leukocyte subpopulations. In particular, there is the need to understanding whether the previous findings of altered DNA methylation levels are caused by a direct real effect of smoking or by a casual effect due to a shift in cell-blood composition. In the analysis of DNA methylation data, the computational challenge regards the complexity of data and to overcome the limitations of existing methods used in these analyses.

The aims of this Thesis are: *a)* to evaluate how the leukocytes vary according to different levels of smoking exposure; *b)* to explore the already published epigenetic observations regarding tobacco smoking and DNA methylation by examining in details different leukocyte populations and compare them in smokers, former smokers and non-smokers; *c)* to understand the limitations of available pipelines and to develop an *ad hoc* method for DNA methylation profiling at the cell-type level using bisulfite targeted data.

This Thesis is divided in two main parts:

- The first part is composed of Chapter 3 and Chapter 4 and based on new experimental data and on the analyses of existing available data, respectively. We investigated the smoking effects on leukocytes composition in healthy individuals from two different cohorts.
- The second part is composed of Chapter 5 and Chapter 6, based both on bioinformatics analyses and new experimental approaches, where we performed a comparison of the main pipelines applied to analyse DNA methylation data. We also evaluated the differences in DNA methylation in B cells and Monocytes of current and never smokers as measured by DNA target sequencing.

The present doctoral Thesis stems from the collaboration among the Molecular Epidemiology and Exposomics Unit of the Italian Institute for Genomic Medicine

(IIGM) where I conducted the experimental part, the Department of Computer Science of the University of Turin where I carried out the computational analyses and the Department of Twin Research & Genetic Epidemiology at King's College of London where I performed statistical and computational analyses during my Ph.D visiting abroad.

Smoking effects on leukocyte subpopulations

3. Assessing leukocyte subpopulation proportions in healthy individuals with different smoking exposure

In this chapter, we illustrate the experimental approach and the analyses performed to explore how smoke exposure affects the proportions of leukocyte subpopulations in blood samples of healthy individuals. This characterisation is expected to identify the cell-subtypes involved in cigarette smoking immune response.

3.1 Aim of the work

Cigarette smoking is well-known to have adverse effects on the immune system including, being an immune suppressant in a dose-dependent manner (11; 33). In peripheral blood of healthy individuals, tobacco smoking alters leukocyte cell count and distribution, with several studies showing an increase in the overall number of leukocytes in current smokers compared to former and non-smokers in both sexes (34) This leukocyte variation is influenced by smoking intensity (*i.e.,* smoking pack/years), which induces higher value of leukocytes in heavy smokers, a reduction in moderate, and a decline in mild until to never smoker levels (35). However, smoking cessation restores leukocyte count within one year (36, 37)

The mechanisms underlying this alteration are not fully understood yet, and studies investigating the effect of smoking on the immune system led to conflicting conclusions (33;10). The leading hypothesis suggests that the irritating effect of tobacco smoke on the respiratory tract results in the release of pro-inflammatory cytokines, such as TNF-α, IL-1, IL-6, IL-8, and granulocyte-macrophage colony-stimulating factor, which can, in turn, increase the number of leukocytes (35). Alternatively, it has been suggested that nicotine-induced release of hormones from the adrenal gland (catecholamines) can stimulate cortisol

secretion and thus inhibit antibody responses, T cell proliferation, and neutrophilic phagocytic activity (36;11).

Changes in the distribution of major leukocyte subtypes has also been observed. Studies have shown a significant decreases in circulating NK cells (38), an increase in neutrophils and CD3+ T cells (24), and CD4+ (T-helper) T cells (36;39) in smokers compared to never smokers. Nevertheless, results from these studies frequently report conflicting evidence because of considerable differences in sample size, smoking years, age, gender and ethnicity, among the studies. Despite the clear association between a global leukocyte-shift and cigarette smoking, smoking effect on minor cell-subpopulation remains unexplored. In this respect, the aim of this work was to assess the leukocyte subpopulations proportion in a cohort of healthy individuals with different smoking exposures, in order to identify the cell-types involved in response to the smoking induced alteration.

3.2 Materials and Methods

3.2.1 Study population

For this purpose, 288 healthy volunteers with differential smoking habits were enrolled in collaboration with the Association of voluntary Italian blood donors (AVIS) of Turin, which operates in Italy in the field of blood donation and blood components. Individuals of both sexes with an age range between 35 and 70 years (mean 48.92 ± 7.66 years old) were recruited, between December 2017 and February 2019, after a brief study explanation, the written informed consent form to participate in the study was signed.

Subjects with chronic and autoimmune diseases (*i.e.,* diabetes, celiac disease, rheumatoid arthritis, chronic respiratory diseases), those undergoing to radio and chemotherapy in the last six months before the recruitment and pregnant women were not included in the present study. All subjects also filled an informative questionnaire about self-reported smoking habit and lifestyle information. The inclusion criteria for categories of interest were: *a)* smokers who had smoked for

at least five years and a minimum of 10-15 cigarette/day; *b)* former smokers who had quit smoking at least one year before the date of recruitment and *c)* never smokers who were never directly exposed to passive smoking (*i.e.,* cohabitation with partner who smokes every day in the house). The study was conducted according to the guidelines in the Declaration of Helsinki. The protocol of the study was approved by the Ethics Committee of the University of Turin.

3.2.2 Samples collection

At the enrolment, each volunteer donated around eight ml of peripheral blood sample collected in EDTA (Ethylene Diamine Tetra-Acetic acid) vacutainer tubes. Within two hours after blood collection, 200 µl of fresh blood was used to measure the leukocytes distribution by flow-cytometric analysis, while the remaining part was divided into 500µl aliquots of whole blood, buffy coat and plasma. Buffy coat and plasma were both obtained after centrifuging the blood sample at 4°C, 2500 rpm x 10 minutes. All aliquots were stored at -80° C until further analyses were performed.

3.2.3 Leukocytes count and flow cytometric analysis

A panel of fluorescent-antibodies was designed to quantify nine leukocyte-subpopulations including lymphocytes (*i.e.,* T cells (CD4+ T helper cells, CD8+ cytotoxic T cells), B cells, NK/NKT cells), monocytes and granulocytes such as neutrophils and eosinophils. Cell-subtypes present in the blood were detected by determining the accessory molecules (CD markers) on their surface. In addition,to identify the nine leukocyte-subtypes, the cells were directly stained with mouse monoclonal anti-human antibodies (BD Biosciences): CD3+ T cells, CD4+ T helper cells, CD8+ cytotoxic T cells, CD19+ B cells, CD16+ CD56 NK cells, CD14+ monocytes, CD11b+ CD16+ neutrophils, CD11b+CD16- eosinophils.

For each subject, two FACS tubes containing two different mix of antibodies (**Supplementary, Figure S1**) were prepared.

Briefly, in each FACS tube containing 100µl of whole blood was incubated for 10 minutes with 2.5 µl of human FcR blocking reagent of MACS (Miltenyi Biotec) to avoid unwanted binding of antibodies to human Fc receptor-expressing cells

such as B cells, monocytes, and macrophages. Then, all samples were stained with a mix of antibodies and 3 µl Anti-GPR15+ antibody (RD Systems).
The expression level of this antibody was used as a smoking marker (24). After 10 minutes of incubation, Erylysebuffer lysis solution was added to lysate the erythrocytes and the samples were incubated again 10 minutes, then washed twice with PBS (Phosphate Buffered saline). All incubation steps were carried out at room temperature and in dark conditions. 100 µl of PBS and 100 µl of Erylysebuffer lysis solution were added to the samples before the flow cytometer reading.
All measurements were performed on a BD FACS Verse® flow cytometer and analysed with the BD FACSDiva® Software (version 8.0.1 BD Bioscences).
BD FACS Verse® automatic tube acquisition was used to minimise technical variability.
For each tube ~70,000 events were acquired as threshold of cell-acquisition. The lymphocytes were gated according to cell size (forward scatter) and density (side scatter). The resulting population was used as input for hierarchical gating analyses. The same procedure was applied to analyse monocyte and granulocyte populations (**Supplementary, Figure S2).**

3.2.4 Plasma cotinine concentrations assessment

In the present study, cotinine concentration was measured in plasma samples using Cotinine direct ELISA Kit according to the manufacturer's instruction (DRG Instruments, Germany). All samples were analysed in duplicate at 1:100 dilution in 96-well format. Absorbance (Abs) at 450 nm was detected by GloMax Discover Microplate Reader© (Promega). The cotinine concentration was calculated by extrapolation of the linear portion of the standard curve for each well. To improve the normality of the data, the cotinine measures were log-transformed and corrected for technical sources of variation by ComBat function present in *SVA* package implemented in R statistical software (40). The distribution of cotinine concentration among the smoking categories was tested with ANOVA (one-way; **Supplementary Figure S3**.).

3.2.5 Plasma C-Reactive Protein (CRP) concentration measurements

CRP concentrations were determined through the immunoturbidimetric method with AU Beckman Coulter analyser carried out at Unilabs (*Laboratorio Raffaello*) in Turin. The standard clinical value of CRP concentration ranges from a minimum of 1.6 mg/L until to 5.0 mg/L. Despite this threshold, the instrument detects the values greater than 5.0 mg/L, such as those of the acute inflammation > 10 mg/L.

3.2.6 Statistical Analyses

The distribution of the sample charactheristics (*i.e.*, sex, age, alcohol consumption) among the smoking categories was tested by one-way ANOVA and $\chi 2$ tests, for continuous and categorical variables, respectively.

The outliers were removed from the cell-type percentages (*i.e.*, cell-type measurements deviating more than three standard deviations from the mean of each cell-type) and after the date were normalized.

First, we investigated the association between the variation of cell-type percentages and active smoking, including in the regression model only current and never smoker individuals, and age, sex and alcohol consumption as covariates to adjust the model. The associations with p-value passed a Bonferroni-derived threshold of 0.05/N*eff*, where N*eff* is the effective number of independent tests (*i.e.,* equal to the number of cell-subtypes tested) were considered significant.

Second, we sought whether the cell-type percentages significantly associated with current smokers may persist altered in former smokers after smoking cessation, we compar former *vs* never smokers using the approach described above. Next, in former smokers, we explored if the cell-type variations also were different according to the years of smoking cessation. However, former smokers had quitted smoking at different times (range :1-40 years). Thus, we decided to classify the individuals into three categories: individuals who quitted smoking before to 10 years, individuals between 10-20, and those after 20 years.

Due to a highly skewed distribution, CRP concentrations were categorized following the normal clinical parameters that range from 1.6 mg/L to 5.0 mg/L.

Four categories were formed: "low" including values <1.6 mg/L, "moderate" comprising values from 1.7 to 3.0 mg/L, "elevate" with values between 3.0 to 5.0 mg/L and “high” referring to values between 5.0 to 10 mg/L. Values greater than 10 mg/L were excluded because they were associated with acute inflammation (*i.e.,* infection and inflammatory disease).

We tested if the proportion of smokers followed a linear trend among the CRP categories by the Cochran-Armitage trend test.

All the analyses were implemented in R statistic software, version 3.5.3

3.3 Results

3.3.1 Characteristics of the study population

The study was carried out in 288 healthy volunteers aged between 35 to 70 years old (average 48.94 ± 7.66), of which 218 were males (76%).

Self-reported smoking information collected by questionnaires at the time of the enrolment was cross-validated using cotinine concentration.

The cotinine values reported by recruited volunteers were consistent with smoking category declared in the questionnaire, showing 89 current, 99 former and 100 never smokers (**Table 3.1**). Age and gender distributions were similar among the smoking categories (p-value=0.948; p-value=0.448, respectively). 89% of current smokers were represented by light smokers (<15 cigarettes/day), and the calculated dose of smoking was an average of 11.80 ± 12.08 pack/years. In former smokers, the categories of cigarettes smoked per day were more homogeneous compared to current smokers. However, the prevalence (52%) was also in this case of light smokers (< 15 cigarettes smoked per day). The time since quitting smoking in former smokers was on average 16.3 ± 9.6 years, with about 40% of the former smokers which had stopped smoking since 10-20 years. There was a statistically significant difference in alcohol consumption among the smoking categories (p-value= 5.4 x10-4).

Table 3.1 Participant characteristics. Descriptive statistics of the study participants stratified by their smoking habits. Mean and standard deviation are reported for continuous variables, absolute numbers and percentages of individuals in each group are reported for categorical variables. P-value* were calculated with Chi-squared test for categorical variables and One-way ANOVA for continuous variables.

Variables	Current smokers	Former smokers	Never smokers	P-value*
Individuals (N)= 288	89	99	100	
Age (range 35-70)	47.97 (8.50)	50.58 (7.50)	48.14 (6.76)	0.948
Gender (%)				0.448
Male	71 (80%)	76 (77%)	72 (72%)	
Female	18 (20%)	23 (23%)	28 (28%)	
Dose of smoking (pack/years)	11.80 (12.08)	10.81 (9.47)		-
Cigarettes smoked/day				
<15	80 (89%)	52 (52%)		-
15-20	6 (6%)	31 (31%)		
> 20	2 (2%)	13 (13%)		
Missing	1 (1%)	3 (3%)		
Smoking cessation years		16.34(9.59)		-
<10		28 (28%)		
10-20		40 (40%)		
> 20		27 (27%)		
Missing		4 (4%)		
Alcohol consumption				$5.4x10^{-4}$
Drinkers	57 (64%)	75 (76%)	66 (66%)	
Occasional	15 (16%)	12 (12%)	5 (5%)	
Never	10 (11%)	8 (8%)	27 (27%)	
Missing	7 (7%)	4 (4%)	1 (1%)	
CRP (mg/L)				$8.6x10^{-4}$
Low (<1.6)	49 (55%)	61(61.6%)	74(74%)	
Moderate (1.7-3.0)	10 (11.2%)	24(24.2%)	11(11%	
Elevate (3.0-5.0)	17(19.1%)	5(5.05%)	7(7%)	
High (5.1-10.)	11(12.3%)	6(6.1%)	8(8%)	
Missing	2(2.4%)	3(3.05%)	-	

3.3.2 Association between smoking habits and leukocyte-subtypes

We investigated the association between active smoking (*i.e.,* comparing smokers *vs* never smoked individuals) and the percentages of the leukocyte subpopulations quantified by flow cytometry: CD4+ T-helper, CD8+ T-cytotoxic, CD3+ T-cells, CD16/CD56+ NK-cells, CD56/CD3+ NKT-cells, CD19+ B cells, CD14+ monocytes, neutrophils, and eosinophils as well as the presence of GRPR15+ receptor in each cell types.

We selected as significant the cell-type percentages passing Bonferroni-derived threshold of p-values $<0.05/18=2.8\text{x}10^{-3}$ (**Table 3.2**).

Table 3.2 Results of association analysis between leukocyte-subtypes in current vs never smokers. P values were obtained by linear regression model adjusted for age, sex and alcohol consumption. Significative p-values passing Bonferroni threshold are labelled in red (p-values $<0.05/18=2.8\text{x}10^{-3}$). Beta= effect size, SE= standard error.

Leukocyte subtypes	Beta	SE	P-value
B	0.06	0.75	0.45
B GPR15	0.35	0.07	$2.2\text{x}10^{-6}$
CD3+ T	0.06	0.07	0.41
CD3+ T GPR15	0.64	0.06	$2.0\text{x}10^{-16}$
CD8+ T	-0.15	0.07	0.03
CD8+ GPR15	0.50	0.07	$7.7\text{x}10^{-13}$
CD4+	0.07	0.07	0.37
CD4+ GPR15	0.65	0.06	$2.0\text{x}10^{-16}$
Monocytes	0.18	0.07	$7.0\text{x}10^{-3}$
Monocytes_GPR15	-0.27	0.08	$6.1\text{x}10^{-4}$
NK T	-0.03	0.07	0.69
NK T GPR15	0.18	0.07	$1.9\text{x}10^{-2}$
NK	-0.33	0.07	$6.2\text{x}10^{-6}$
NK GPR15	-0.18	0.08	$2.7\text{x}10^{-2}$
Eosinophils	-0.05	0.07	0.47
Eosinophils GPR15	0.04	0.07	0.58
Neutrophils	0.20	0.07	$5.3\text{x}10^{-3}$
Neutrophils GPR15	-0.13	0.07	$7.8\text{x}10^{-2}$

In current smokers we observed a significant decrease of NK-cell proportion (p-value= $6.2\text{x}10^{-6}$) and, an increase of GPR15+ expressing in all T cell (CD3+ (p-value= $2.0\text{x}10^{-16}$), CD4+ (p-value= $2.0\text{x}10^{-16}$), CD8+(p-value= $7.7\text{x}10^{-12}$)), and CD19+ B (p-value=$2.2\text{x}10^{-6}$) cells compared to those observed in never smokers.

In contrast, in the same comparison we found a decrease of the proprotion of GPR15+ expressing cells in monocytes (p-value=$6.1x10^{-4}$, **Figure 3.1**).

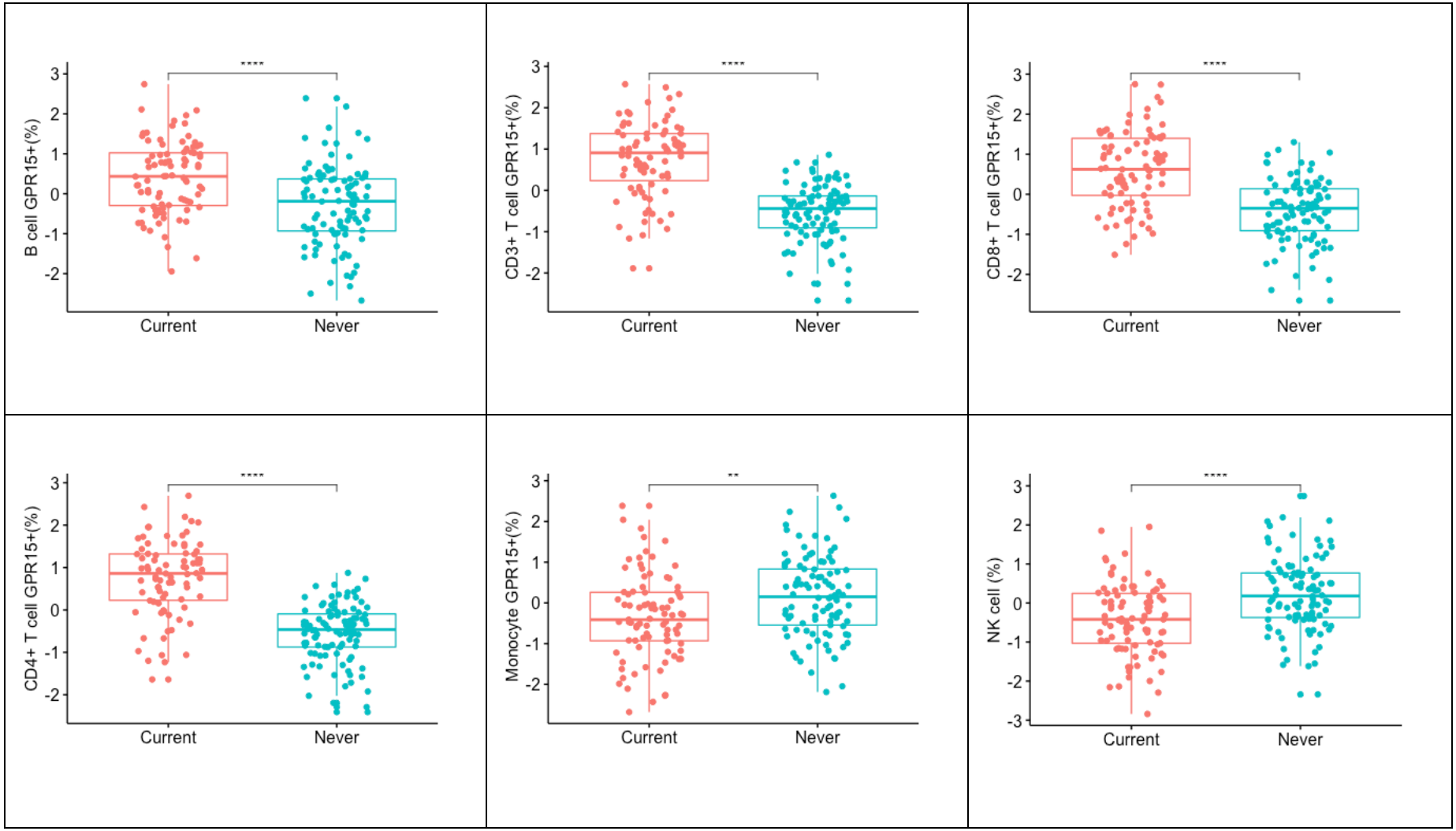

B cell GPR15+(%)
CD3+ T cell GPR15+(%)
CD8+ T cell GPR15+(%)
CD4+ T cell GPR15+(%)
Monocyte GPR15+(%)
NK cell (%)
Current
Never
****
**

Figure 3.1 Leukocyte subpopulations significantly associated with active smoking. Boxplots report the difference of the cell-type percentages between current vs never smokers. Percentage of each cell-subtype was normalized and corrected ~ sex, age, alcohol consumption. **=$p<0.01$; ****= $p<0.0001$

3.3.2 Association between leukocyte-subtypes and former smoking

Next, we explored whether the cell-type percentages significantly associated with active smoking persist altered after smoking cessation. Comparing former vs never smokers we found significantly associated (i.e., associations passing Bonferroni derived threshold p-value < $0.05/6=8.3x10^{-3}$) CD3+ and CD4+ T cells expressing GPR15+ showing an increase in former smokers, and a trend among the smoking categories (Table 3.3).In contrast, the proportion of NK, B-GPR15+ and monocyte GPR15+ cells were not statistically different between former and never group.

We further investigated in former smokers, if the cell-type showed a significant difference in comparison with never smokers change also according to the years since smoking cessation. CD4+ T cells expressing GPR15 marker were significantly decreased with the years of smoking cessation (p-value = 0.02; Figure 3.2)

Table 3.3 Results of cell-type percentages that persist altered after smoking cessation. P values were obtained by linear regression model adjusted for age, sex and alcohol consumption, comparing former vs never smokers. In red are labelled significative p-values passing Bonferroni threshold (p-values <$0.05/18=2.8x10^{-3}$). Beta= effect size, SE= standard error

Leukocyte subtypes	Beta	SE	P-value
B GPR15	0.24	0.14	0.10
CD3+ T GPR15	0.42	0.11	$5.0x10^{-4}$
CD8+ GPR15	0.27	0.13	$3.0x10^{-2}$
CD4+ GPR15	0.45	0.12	$2.1x10^{-4}$
Monocytes_GPR15	-0.01	0.13	0.89
NK	-0.09	0.14	0.50

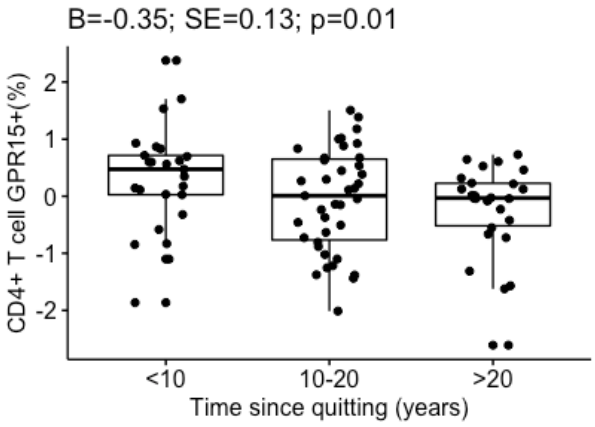


Figure 3.2. Leukocyte subpopulations significantly altered in former smokers. Boxplot shows a decrease of CD4+T cells expressing GPR15+ with the years after smoking cessation, considering only former smokers. It also reports the values of linear regression analysis. ***=p<0.001; ****= p<0.0001; B= effect size; SE= standard error.

3.3.3 C-Reactive Protein levels in smoking categories

C-Reactive Protein (CRP) levels in plasma samples were measured as a marker of systemic inflammation. Smoking habit is known to be strongly associated with a high inflammatory level. As expected, we observed a difference of the CRP values among smoking categories ($P=8.6x10^{-4}$), although more than 50% of individuals in each smoking category showed a low grade of inflammation (**Table 3.1**). To investigate if the number of smokers increased among the CRP categories, we tested the presence of a linear trend in current smokers compared to the whole group. We observed an increase in the proportion of smokers among CRP categories ($p< 0.05$, **Table 3.4**).

Table 3.4. Contingency table of current smokers distribution among the CRP categories. The p-value was computed with Cochran-Armitage test to test the linear trend of smokers among the CRP categories.

	Low (<1.6 mg/L)	Moderate (1.7-3.0 mg/L)	Elevate (3.1-5.0 mg/L)	High (5.1- 10mg/L)	P-value
Current smokers	49	10	17	11	0.004
Total individuals	184	45	29	25	

3.4 Discussion

To date, it is well established that smoking habits increase the overall numbers of leukocytes (9;11;36;38). However, its impact on leukocyte-subpopulations distribution remains object of study.

In the present study, to evaluate the smoking effect on leukocyte subpopulations, we recruited 288 healthy participants, distributed in 89 current, 99 former and 100 never smokers. Lifestyle and smoking habits information and a sample of peripheral venous blood were collected at the enrolment. For each individual, the leukocyte subpopulations: CD3+ T cells, CD4+ T-cells, CD8+ T cells, CD16+CD56+ NK-cells, CD3+CD56+NKT-cells, CD19+ B cells, CD14+ monocytes, CD11b+-CD16+neutrophils, and CD11b+CD16- eosinophils were counted using flow cytometry. For each cell-type the expression of the GPR15 cell receptor as smoking biomarkers was also measured. Moreover, we checked the self-reported smoking habits by the experimental measure of cotinine concentration and the systemic inflammatory level with CRP concentrations in the collected plasma samples.

As a major finding, we observed a significant decrease in circulating NK cells and a significant increase of GPR15+ expressing cells in CD3+, CD4+, CD8+ and B

cells in smokers compared to never smokers. Moreover, we found in current smokers again the significant decrease of GPR15 expressing cells in monocytes. The majority of these results are in according with previous studies. For example, a significantly lower proportion of NK cells in current smoker was found for the first time by Tollerud and colleagues (38). They observed a decrease of NK cells in currents but also among former smokers, including subjects who did not smoke for more than 20 years. Conversely, in our study the NK cell levels in former and never smoker individuals were similar. This can be possibly explained by the different smoking exposure between the two studies, since the difference on average of smoking pack/years between these studies is 20, 34.8 in Tollerud study *vs* 10.81 in our study.

In agreement with what reported in the literature, we also observed in current smokers an increase in the proportions of CD3+ T, CD4+, CD8+, CD19+ B cells expressing GPR15+ compared to never smokers. GPR15+ cell receptor is involved in the regulation of immune response during inflammatory process mediating cell recruitment such as T cell in bowel disease and macrophages in synovial tissue in patients with rheumatoid arthritis (RA) (20). GPR15 gene on cg19859270 site was found differentially methylated in whole blood of smoker *vs* never smoker. These difference in DNA methylation is attributable to a high proportion of CD3+ T cell expressing GPR15 in smoker compared to never rather than by direct impact of tobacco smoke on DNA methylation (24). They speculated that the alteration might be due to the inflammatory effect of smoking on the immune system. Always, in this study, they showed the smoking-related increase of GPR15+ expressing cells in CD19+ B cells. Furthermore, an elevated expression of GPR15 was also found in monocytes of RA patients. On the contrary, in our study, we observed the decreasing of monocytes expressing GRP15+ in smoker compared to never smokers. This result is inconsistent with the previous findings on GPR15+ expression, because if the GPR15+ expression at cell-type level increases as inflammatory status as reported in T and B cells in smokers, it should increase also in monocytes.

Interestingly, in former smokers we showed that the proportion of cells expressing GPR15+ in CD3+ and CD4+ T cells persist altered also after smoking cessation.

In particular, CD4+GPR15+ T cells remain altered after more than 10 years since smoking quitting. Taken together these findings confirm the intermediate levels of GPR15 gene expression observed in former smokers (24) and highlight an inflammatory status not completely restored to never smoker levels also several years after smoking cessation.

To explore the inflammatory levels, in our cohort, we measured the concentration of CRP. Elevated levels of CRP were demonstrated to be associated with cigarette smoking in a dose-dependent manner in male and women with both the duration and intensity of smoking (41). We observed a linear trend between the number of smokers and the CRP categories, suggesting a more elevate inflammatory status in current smokers. This result is intriguing and in line with the obtained results of this study.

Our findings reflect the composition of the cohort that included mostly light smokers (<15 cigarettes/day), while the majority of the published studies were focused on heavy smokers (≥ 20 cigarettes/day). On the other hand, they highlight the smoking effect on leukocyte subpopulation also at low grade of exposition.

We are aware that in the present study we evaluated only the primary leukocyte subpopulations and that the smoking-related leukocyte variations might be attributed to minor differentiate cell-subtypes like regulatory and memory cells. This is a preliminary study where we sought to highlight the main smoking effects on the distribution of primary leukocyte-subpopulations caused to smoking exposure in order to clarify whether a cell-subtypes was more affected than others.

These results need replication in a larger and more variegate cohort where are available the measures of a large number of leukocyte subtypes.

In conclusion, our study shows the significant variation in the proportion of NK and CD3+T, CD4+T, CD8+T, CD19+ B, NKT cells and monocytes expressing GPR15+ caused by smoking exposure. Despite some limitations of our study, we demonstrated that the magnitude effect of smoking on leukocyte subpopulation is clearly visible also in a small and heterogeneous cohort, mainly including light smokers. This is an important and significant finding.

4. Leukocyte shifts in healthy women of the TwinsUK cohort with different smoking habits

This Chapter describes a study performed in collaboration with the Department of Twin Research & Genetic Epidemiology at King's College of London during my Ph.D abroad stay (5 months, from 22/02/2019 to 22/07/2019). We show an association between smoking habits and an elevate number of immune cells in TwinsUK cohort. This analysis is expected to identify the minor cell-subtypes involved in cigarette smoking immune response. Publication of this work is in submission.

4.1 Aim of the work

The rationale of this study is the same of the previous study conducted on healthy individuals from Turin (Chapter 3): to investigate leukocyte shifts in relation to smoking habit. In this case, we want to extend our understanding of the smoking effects on leukocyte subpopulations that are less frequent in blood and, normally, very difficult to analyse. The novelty of this study is the availability of around 42,000 immune cell traits measured by high-resolution deep immunophenotyping flow cytometry analysis in 497 twins from the TwinsUK cohort.

4.2 Methods

4.2.1 Study population

The TwinsUK cohort is a large cohort that includes about 14,000 subjects, with a prevalence of women, extensively studied to understand the genetic and environmental basis of a range of complex diseases (42; 43). The population is not enriched for any particular disease and it is representative of the general UK

population. We selected 358 healthy females within the cohort, aged between 41 and 77. Detailed characteristics of participants are reported in **Table 4.1.**

The individuals were included for: *i*) complete self-reported smoking habits information collected by questionnaire, *ii*) availability of cell-blood subtypes (in this study referred to as **immune traits**, the same nomenclature reported in previous studies on this data (44; 45) measured by flow cytometry, and *iii*) availability of self-reported and/or doctor-diagnosed information on inflammatory and autoimmune diseases.

St. Thomas' Hospital Research Ethics Committee approved the study, and all twins provided written informed consent.

4.2.2 Immunophenotyping

A full description of the immune cells quantification and phenotype is detailed in the papers presented by *Roeaderer et al*., and *Mangino., et al.,* (44; 45). Briefly, leukocytes were characterised in 497 females from the TwinsUK cohort using flow cytometry on seven distinct 14-colour panels. In the first stage, parent lineages were defined within each panel via manual gating based on canonical marker combinations for leukocytes subsets of known functionality. Within each parent lineage, boolean gates were then manually defined for each additional cell surface marker (*i.e.*, markers not used to determine the parent lineage), and information on all combinations of these boolean gates (*i.e.*, whether positive, negative, or ignored) were subsequently used to define subsets of immune cells and to measure their immune **cell subset frequencies** (**CSFs**, evaluated as percentages with respect to the total number of leukocytes in their parent lineage). Using this approach, we captured a grand total of 88,367 immune traits, describing 50 supersets of the parent lineages, and 88,317 CSFs.

In this study, we assumed that measurements equal to zero meant impossibility to detect an immune trait rather than its absence. Therefore, zero values were considered as not available (NA).

To select 41,701 robust immune traits we applied several criteria. First, we removed 41,336 CSFs with median value <0.1 or >99%, as done previously (44;45). Second, we removed 5,246 CSFs with missingness larger than 20%.

Third, we removed 34 redundant CSFs which passed the previous quality check steps and were measured in multiple panels (median Spearman's ρ within pairs = 0.88, range=0.37-0.95).

Immune traits were log-transformed, to improve the normality of their distribution, and then corrected for batch effects using a linear mixed model, as implemented in the R package *lme4* (v1.1.21), with flow cytometry batch number included as random effect. Before carrying out the association analyses, we removed the outliers (*i.e.,* immune trait measurements deviating more than three standard deviations from the mean of each trait).

4.2.3 Self-reported smoking history

Detailed information about smoking history was self-reported via 11 longitudinal questionnaires, collected from 1992 to 2010 in 496 individuals with immunophenotyping available (median number of responses: 7). Consistency of self-reported smoking status was assessed using additional self-reported information, *i.e*., age of start and quitting smoking, and the number of cigarettes and/or packs smoked. For instance, individuals who described themselves as never smokers, but reported, in any questionnaire, age of start and/or quitting smoking, and/or that they had smoked any number of cigarettes were removed from this study. We allowed for smoking relapse after smoking cessation, and considered as current status the latest reported before immunophenotyping. This resulted in the inclusion of 460 individuals, 35 of whom were current smokers, 189 former smokers, and 236 reported never having smoked.

4.2.4 Immune-mediated inflammatory diseases and cancer

History of immune-mediated inflammatory diseases (IMID, *i.e.*, chronic obstructive pulmonary disease, Crohn's disease, systemic lupus erythematosus, multiples sclerosis, polymyalgia rheumatica, psoriatic arthritis, rheumatoid arthritis, and ulcerative colitis) was traced through 15 longitudinal self-administered questionnaires completed between 2004 and 2017 (median number of responses per individual: 3). For each condition, study subjects who reported

being diagnosed by a doctor at least once were treated as IMID cases, and when multiple ages at first diagnosis were provided, the minimum age was considered.
Cancer history was available from the 2019 Office for National Statistics. Non-melanoma skin cancers and carcinomas *in situ* were not taken into account.
Using these pieces of information, 102 individuals were excluded either because having a diagnosis of IMID reported before or within two years from immunophenotyping or being diagnosed for one or more cancers dating five years before or within one year from immunophenotyping.
The final dataset consisted of 358 healthy female individuals, 25 of whom were current smokers, 135 former smokers, and 198 never smokers, and included 28 monozygotic and 52 dizygotic twin pairs, and 198 singletons (**Table 4.1**).

4.2.5 Statistical analysis

First, we aimed at identifying the immune traits involved in the response to active smoking using data from current and never smokers (**Figure 4.1, right panel**). Due to the high variability of time of smoking cessation before immunophenotyping (range: 1-50 years), we excluded former smokers from this analysis to avoid any confounding effects. Associations of immune traits with smoking status were carried out using a linear mixed model, as implementer in the *lmerTest* R package (function *lmer,* v3.1.1), including age at immunophenotyping as a fixed effect, and family as a random effect.
Due to the strong correlation among immune traits, we considered as significant the associations passing a Bonferroni-derived threshold of $0.05/N_{eff}$, where N_{eff} is the effective number of independent tests calculated on the whole set of 497 individuals with immunophenotyping data using the approach proposed by Li & Ji (46).
Due to the unequal sample size between current and never smokers, for each of the N immune traits passing the Bonferroni-derived threshold of $0.05/N_{eff}$ described above, we generated 5,000 random datasets where labels indicating smoking status were randomly permuted between monozygotic/dizygotic twins pairs and among singletons, in order to preserve the family structure and, thus, the underlying genetic correlation. Then, we counted the number of times T the

association p-value in the random datasets was lower than $0.05/N_{eff}$, and used this number to evaluate an empirical p-value as (T+1)/5,001. We confirmed an association as significant when its empirical p-value passed a Bonferroni-derived threshold of $0.05/M_{eff}$, where M_{eff} is the effective number of independent tests evaluated by the approach proposed by Li & Ji on the subset of the N significancy associated immune traits.

Then, we investigated whether the immune traits associated with active smoking and confirmed by permutation testing remained altered after smoking cessation using data from former and never smokers and the statistical model described above (**Figure 4.1, right panel**). Finally, to investigate the presence of a trend in the relative frequencies of the altered immune traits in current *vs* former *vs* never smokers, we performed a further association study including the three smoking categories (*i.e.,* current, former, and never smokers; **Figure 4.1, right panel**), following the design detailed above. Associations passing a Bonferroni-derived threshold of $0.05/M_{eff}$ were considered as significant.

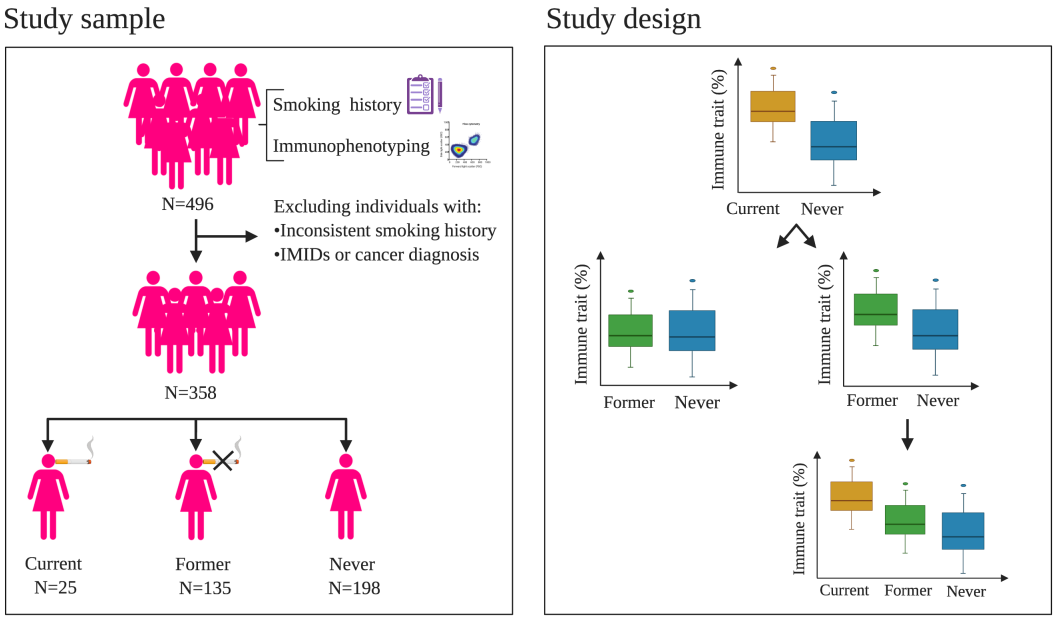


Figure 4.1. **Study sample and design.** Left panel: criteria used to select the 358 women included in this study. Right panel: analysis approach. First, we sought associations between current tobacco smoking and 41,701 immune traits in 25 current and 198 never smokers. Second, we investigated whether the levels of the identified immune traits could be fully or partially restored to never-smoker proportions using data from additional 135 former female smokers. IMIDs: immune-mediated inflammatory diseases.

4.3 Results

4.3.1 Population characteristics

The characteristics of the study participants from the TwinsUK cohort are shown in Table 4.1. The data set included 358 females with an average age of 60.90 ± 8.31 years (range: 41-78). Twenty-five of them (7%) were current smokers; 135 (38%) were former smokers and 198 (55%) never smokers.

Table 4.1 Sample characteristics. The dataset includes 358 females of European ancestry from the TwinsUK cohort. Mean and standard deviation are reported for continuous variables, absolute numbers of individuals in each group are reported for categorical variables. P-values were evaluated using ANOVA for continuous variables and χ^2 test for categorical variables. MZ: monozygotes; DZ: dizygotes

	All	Current smokers	Former smokers	Never smoker	P-value
N	358	25	135	198	-
Age (range 41-78)	-	59.77±8.75	61.71±8.28	60.49±8.28	0.33
Zygosity (MZ/DZ/singletons)	56/104/198	2/0/23	22/26/87	32/78/88	1.61×10^{-6}

4.3.2 Immune traits associated to active smoking

In this study, we first aimed at identifying the immune traits involved in the smoking response using data from current and never smokers.

We identified 848 (2.0%) CSFs associated with active smoking at a Bonferroni-derived threshold of $0.05/2{,}610=1.9\times10^{-5}$, and whose association was further confirmed by permutation testing ($P<0.05/79=6.3\times10^{-4}$, **Methods, Supplementary, Table S1**).

Associated immune traits belonging to the same lineage were highly correlated (**Supplementary, Figures S4-8**), and this correlation was particularly strong among immune traits presenting similar patterns of molecular markers. Therefore, to facilitate the description of the obtained results, we show them as groups of

highly correlated immune traits characterised by at least a common molecular marker. The cell lineages, subsets, and phenotypic markers used here are defined according to the nomenclature reported in Roederer *et al*. (44).

Frequency of circulating T cell, B cells, and monocytes is influenced by active smoking

Significantly associated CSFs included 711 and 133 immune traits belonging to the two major lymphocyte populations of T cells (83.8% of the total associated CSFs) and B cells (15.7%), respectively, and four belonging to the monocyte subset (0.5%).

Among the T cells we found the 264 (37% of total T cells) CSFs of CD8+, 277 (51%) CSFs of CD8+ memory, 162 (23%) CSFs of CD4+, five (0.7%) CSFs of double-positive (DP) CD4+CD8+, and three (0.4%) of γδ T cells lineages. And the relative proportions of B cells included 75 (56 of total B cells %) CSFs of class-switched memory B cells, 44 (33%) CSFs of plasma cells, 10 (7.5%) CSFs of transitional and four (3.1%) CSFs of naïve B cells (**Figure 4.2**). These results indicate that the CSFs are strongly affected in smoker belong to the CD8+ lineage, in particular to CD8+ memory T cells.

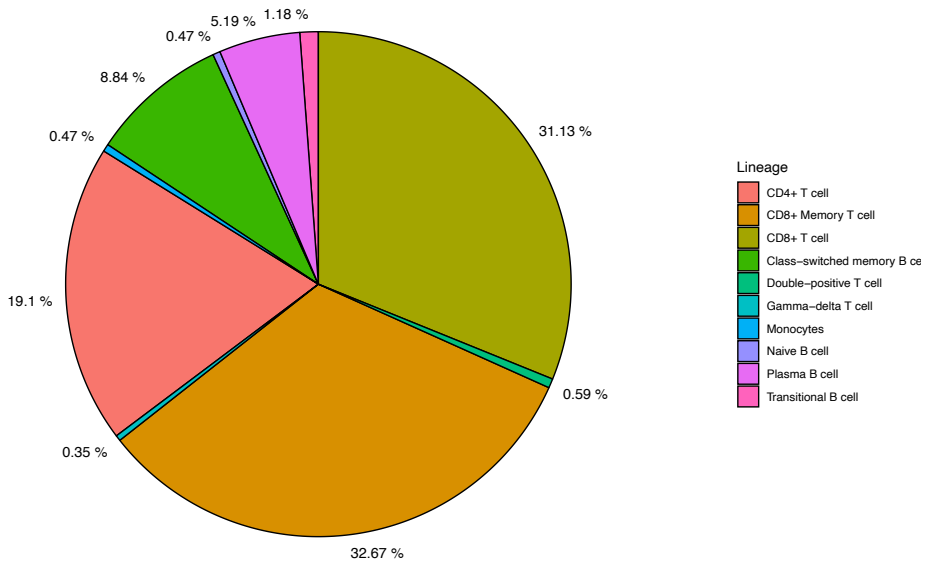


Figure 4.2. **Study Relative proportions of significantly associated CSFs with active smoking.** Plot shows the cell-lineage percentage associated with active smoking calculated on 848 associated CSFs.

T-cells

CD8+ T cells

The largest number of associated CSFs belonged to the CD8+ lineage expressing the CD25 activation marker (184 CFSs**,** 34% of total CD8+ lineage). Sixty-five immune traits expressed exclusively the CD25 activation marker and were strongly correlated with each other (mean Pearson's $|\rho|=0.90$, Supplementary Figure 1, p-value$<8.42\times10^{-6}$), 52 CSFs expressed CD25+ in combination with CD73+ marker with regulatory phenotype (CD8+CD25+CD73+, mean Pearson's $|\rho|=0.94$, p-value$<6.88\times10^{-7}$) and 67 CSFs expressed both the CD25+ and CD127+ markers with proliferation activity (CD8+CD25+CD127+, mean Pearson's $|\rho|=0.90$, p-value$<5.50\times10^{-6}$). The relative proportions of these traits were positively associated with active smoking (*i.e.*, their relative proportions were increased in current smokers; **Figure 4.3**).

Conversely, four CD8+CD25- T cells (n=4, p-value<$5.46x10^{-6}$, mean Pearson's |ρ|=0.94) displayed an opposite direction of effects (*i.e.,* their relative proportions were decreased in current smokers; **Figure 4.3**).

Furthermore*,* in current *vs* never smokers we observed significantly increased the relative proportions of 33 CD8+CD127+T cells and 35 CD8+ expressing the CD95 marker (p-value<$1.85x10^{-5}$, mean Pearson's |ρ|=0.85, p-value<$1.85x10^{-5}$, mean Pearson's |ρ|=0.84, respectively).

CD8+ memory T cells

Within CD8+ memory T cells lineage we found the relative proportion of 34 CSFs expressing CD25+ activation marker (p-value<$1.63x10^{-5}$, mean Pearson's |ρ|=0.98), 32 expressing CD25+CD73+ markers (CD8+CD25+CD73+CD45RO+, p-value<$4.76x10^{-6}$, mean Pearson's |ρ|=0.99) and 30 expressing CD25+CD127+ markers (CD8+CD25+CD127+CD45RO+, p-value<$1.00x10^{-5}$, mean Pearson's |ρ|=0.98) positively associated with active smoking.

CD8+ memory T cells expressing the CCR4 chemokine receptor were positively associated with active smoking (n=88; p-value<$1.87x10^{-5}$, mean Pearson's |ρ|=0.73), whereas CD8+CCR4- memory T cells were negatively associated (n=40; p-value<$1.75x10^{-5}$, mean Pearson's |ρ|=0.54; **Figure 4.3**).

In smokers, we further identified a decreaseof the relative proportions of 12 CD8+ memory cells expressing the CD161 activation and pro-inflammatory marker (CD161+, p-value<$4.01x10^{-6}$, mean Pearson's |ρ|=0.96).

Finally, the relative proportions of 15 central memory T cells (p-value<$1.33x10^{-5}$, mean Pearson's |ρ|=0.78), of 10 long term memory T cells (p-value <$1.51x10^{-5}$, mean Pearson's |ρ|=0.89), and of 16 transitional memory T cells (p-value<$1.59x10^{-5}$, mean Pearson's |ρ|=0.89) were found also increased in smokers.

CD4+ T cells

We identified 107 CD4+ T cells expressing the CD38 activation marker (CD38+, with pro-inflammatory activity; p-values<$1.80x10^{-5}$, mean Pearson's |ρ|=0.60) negatively associated with active smoking. Conversely, the relative proportions of CD4+CD38- were increased in current smokers (n=8; p-value<$1.87x10^{-5}$, mean Pearson's |ρ|=1, **Figure 4.3**).

By contrast, negatively associated with active smoking were observed the relative proportions of a subset of T helper-2 (n=12; p-value<$1.76x10^{-5}$, mean Pearson's |ρ|=0.91), T helper-17 (n=24; p-value<$1.61x10^{-5}$, mean Pearson's |ρ|=0.81), and naïve CD4+ (n=11; p-value<$1.43x10^{-5}$, mean Pearson's |ρ|=0.98).

CD4+CD8+ double positive T cells

The relative proportion of double-positive T cells expressing the CD25+ activation marker (DP, CD4+CD8+CD25+, mean Pearson's |ρ|=0.72) was increased in current *vs* never smokers (n=4, p-value < $1.87x10^{-5}$), while a unique CD4+CD8+CD25- DP T cell (DP CD4+CD8+CD25-) was decreased in current smokers (p-value = $1.71x10^{-6}$).

γδ T cells CD45RA+

The relative proportions of three immune traits belonging to the γδ T cells lineage (Vg9+Vd2- subset) and expressing the CD45RA naïve marker (CD45RA+; p-value < $1.70x10^{-5}$, mean Pearson's |ρ|=1) were decreased in current smokers.

B cells

The relative proportion of class-switched memory B cells isotypes: IgA, IgG and IgE. (n=64, p-value < $1.85x10^{-5}$, mean Pearson's |ρ|=0.66; n=10, p-value < $6.54x10^{-6}$, mean Pearson's |ρ|=0.65 and n=1, p-value = $1.74x10^{-5}$, for IgA , IgG and IgE, respectively) were positively associated with active smoking.

By contrast, B cells expressing the CD38 marker, known as plasma cells (n=44; p-value < $1.86x10^{-5}$, mean Pearson's |ρ|=0.60) decreased in current smokers compared with never smokers.

We further observed ten immune traits in the transitional stage of B cells, co-expressing the CD21 naïve and the CD95 memory markers (CD21+CD95+, p-value < $1.87x10^{-5}$, mean Pearson's |ρ|=0.85) positively associated, and four naïve B cells (CD21+, p-value < $1.60x10^{-5}$, mean Pearson's |ρ|=1) negatively associated with active smoking.

Monocytes

We identified four CSFs belonging to the monocyte lineage which were negatively associated with active smoking (p-value < 3.98×10^{-6}, mean Pearson's $|\rho|=0.95$).

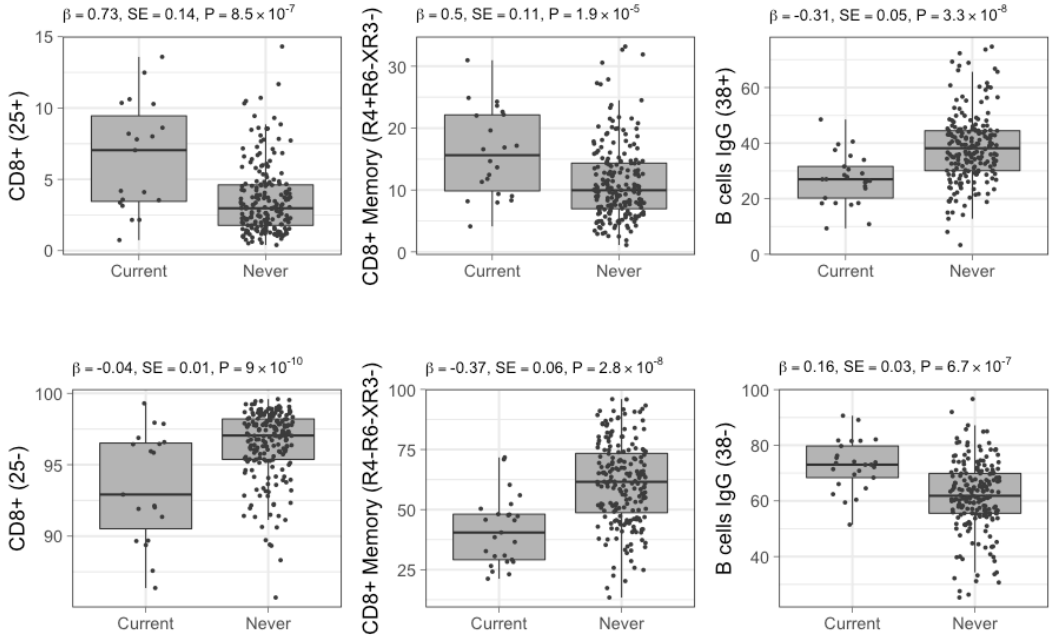


Figure 4.3. **Distribution of selected CSFs in 223 healthy women.** Raw relative proportion (percentage) are plotted, and each boxplot reports effect size (β), standard error (SE), and p-value (P) of the linear regression analysis (current *vs* never smokers). Opposite directions of effects are shown for a selected set of markers: left column CD25 (25+/25-) marker in CD8+ T cells, middle column CCR4 (CCR4+/CCR4-) chemokine receptor marker in CD8+ memory cells, right column CD38 (38+/38-) marker in B cells isotype IgG.

Immune traits are partially restored in former smokers

Next, we investigated whether the 848 CSFs significantly associated in current *vs* never smokers remained significantly different in 135 former *vs* 198 never smokers, that is, whether their relative frequency was restored in individuals who quitted smoking (**Methods, Figure 4.1**). We observed that 390 CSFs (46.0%) including the majority of the B cells (125/133, 94.0%), all the CD4+ (n=162), DP (n=5), and γδ (n=3) T cells, as well as monocytes (n=4) were not significantly different between former and never smokers, suggesting a complete restoration of these immune traits after cessation of smoking (p-value ≥ 0.05**; Methods, Figure**

4.4 ,Supplementary Table S2). In contrast, 254 CSFs (30.0% of the 848 CSFs) belonging to the CD8+ (n=228, 89%) and CD8+ memory T cell (n=26, 11%) showed a significant difference between former and never smokers at a Bonferroni-derived threshold of $0.05/74=6.8x10^{-4}$. Next, we explored whether the 254 CSFs showed a trend among all the smoking categories. All the 254 CSFs displayed a decreasing trend from current, former and never smokers, suggesting a partial restoration in former smokers **(Methods, Figure 4.4, Supplementary Data 2**).

T cells

CD8+ T cells

The relative proportions of CD8+ T cells expressing the CD25 activation marker showed a different behaviour after smoking cessation. Comparing former *vs* never smokers, we observed 17 CSFs of CD8+CD25+ (46% of CSFs CD8+CD25+) that were not fully restored in former smokers (p-value < $1.40x10^{-2}$), and 48 (57%) CSFs CD8+CD25+ that were partially restored in former smokers (i.*e.,* statistically different in the comparison between current and never smokers, **Figure 4.3,** p-value < $1.80x10^{-8}$). A large relative proportion of CD8 T cells expressing the CD25 and CD73 markers were not completely restored (n= 43 (83%), p-value < 0.05). In comparison, the remaining nine (17%) CSFs were completely restored in former smokers (p-value > 0.06). Moreover, former smokers showed 30 CSFs of the CD8+CD25+CD127+ immune traits that were not fully restored (p-value < $3.25x10^{-2}$) and 37 CSFs were partially restored (p-value < $3.03x10^{-8}$).

In former *vs* never smokers we observed partially restored the relative proportions of 28 CSFs of the CD8+CD127+T cells and 15 CFSs of the CD8+CD95+ (p-value < $1.17x10^{-6}$, p-value < $1.15x10^{-6}$, respectively). In contrast, 5 CSFs of the CD8+CD127+ and 20 CFS of the CD8+CD95+ were not completely restored (p-value < $1.57x10^{-3}$, p-value < 0.01, respectively).

CD8 memory T cells

Within the CD8 memory T cells expressing the CD25 activation marker, we observed that the relative proportions of CD8+CD25+CD45RO+ and

CD8+CD25+CD127+CD45RO+ activated immune traits were statistically different in former smokers in comparison with current and never smokers (n=34, p-value < $2.12x10^{-8}$ and n=30, p-value < $9.41x10^{-9}$, respectively). In contrast, within the immune traits co-expressing CD73 marker, only one was partially restored and others were not fully restored (n=31, p-value < 3.81x10-^{3}).

The relative proportions of CD8+CCR4+ memory T cells expressing the CCR4 chemokine receptor (n=32; p-value > 0.28) and those not expressing CCR4+ receptor (n=36, p-value > 0.08) were completely restored in former smokers, whereas other traits showed a p-value < 0.05 (n=30, CD8+ CCR4+, p-value < 0.002 and n=4, CD8+ CCR4-, p-value < 0.03, respectively). Moreover, in last association analysis including all smoking categories, we found significantly associated 26 CSFs of CD8+ CCR4+ that persist partially altered after smoking cessation (p-value < $2.15x10^{-5}$).

All CD8+CD161+ memory T cell were completely restored in former smokers after smoking quitting (n=12, p-value > 0.68).

Finally, we observed that the relative proportions of ten long term memory T cells (n=10, p-value < 4.1 $x10^{-6}$), 16 transitional memory T cells (p-value <$1.8x10^{-4}$) and two of central memory T cells were partially restored after smoking cessation (n=2, p-value < $1.36x10^{-6}$). In contrast, the other 13 central memory T cells were not fully restored in former smokers (p-value < 0.03).

CD4+ T cells

The relative proportions of the identified CD4+ T cells expressing the CD38 activation marker and those not expressing the CD38 marker were restored entirely in former at the levels of never smokers (n=107, p-value > 0.05; n=8, p-value > 0.65, respectively).

Similar trend was observed in former smokers for CD4+ CD38+/- T cell-CSFs and in the relative proportions of the subset of T helper-2 (n=12; p-value > 0.24,), T helper-17 (n=24; p-value > 0.24), and naïve CD4+ T cells (n=11; p-value > 0.27).

Double positive T cells

DP T cells expressed and not the CD25+ activation marker were fully restored after smoking cessation (n=5, p-value > 0.12).

γδ T cells CD45RA+

The relative proportions of three immune traits belonging to the γδ T cells lineage (Vg9+Vd2- subset) and expressing the CD45RA naïve marker (CD45RA+; p-value > 0.10) were wholly restored after smoking quitting.

B cells

Overall, a large proportion of class-switched memory B cells isotype IgA and IgG (n=60, p-value > 0.05; N=9, p-value > 0.13, for IgA and IgG, respectively), including also IgE isotype (n=1, p-value > 0.07), were completely restored in former to never smokers levels. Whereas, four immune traits of IgA and one IgG isotypes were not fully restored (p-value < 0.04).

The relative proportion of plasma cells were fully restored after smoking cessation (n=32, p-value >0.12; and n=9, p-value > 0.26 for IgA and IgG respectively; **Figure 4,4**), except for three CSFs that were not completely restored (p-value < 0.04).

We observed a complete restoration of the relative proportions of the B cells in transitional stage, co-expressing the CD21 naïve and the CD95 memory markers (CD21+CD95+, n=10; p-value > 0.46), and also the four naïve B cells (CD21+, p-value > 0.13).

Monocytes

The four CSFs significantly associated with smoking belonging to the monocyte lineage expressing the HLA-DR activation marker were completely restored after cessation of smoking (p-value > 0.27).

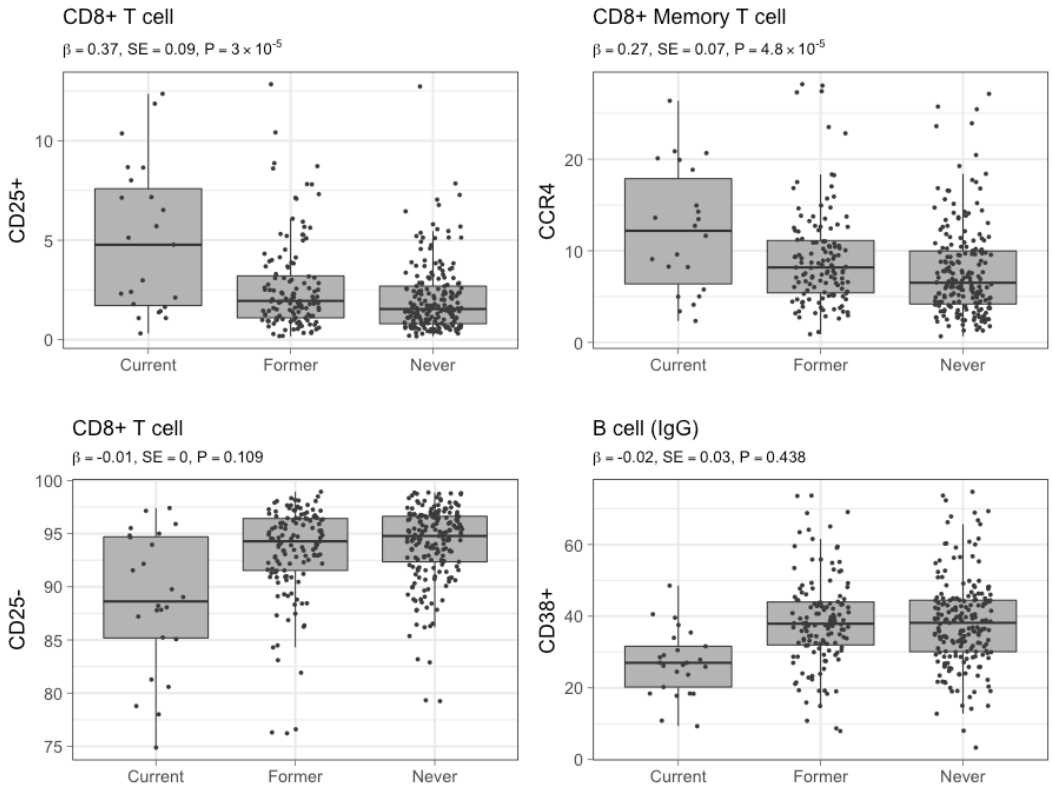


Figure 4.3. **Distribution of selected CSFs in all smoking categories (N=358).** Raw relative proportion (percentage) values are plotted, and each boxplot reports effect size (β), standard error (SE), and p-value (P) of the linear regression analysis (former *vs* never smokers). Top row: immune traits partially restored after smoking cessation, bottom row: immune traits completely restored after smoking cessation.

4.4 Discussion

The effects of tobacco smoking on the innate and adaptive immune system have been extensively investigated showing conflicting results, most likely because of the variability in smoking exposure (*i.e.*, smoking dose and/or whether cigarettes, pipe, or cigars were smoked) as well as the intrinsic differences in study populations (*e.g.*, in age, sex, or ethnicity) and their sample size (8;10). Moreover, the majority of the studies focused only on primary leukocyte subpopulations (*i.e.*, CD3, CD4+, CD8+ T cell, B cell, natural killer cells, monocytes and granulocytes) which are more abundant in blood, and, consequently, easier to measure (36;39).

In the present study, we explored the association between self-reported smoking status and 41,701 immune traits measured by flow cytometry in the peripheral blood of 358 healthy women of European ancestry. To the best of our knowledge, this is the first finely detailed association study aimed at elucidating the relationship between smoking and both the innate and adaptive immune system.

First, we examined the variation related to active smoking by comparing current with never smokers. We observed a change in the relative proportions of circulating leukocytes, mostly in the T and B cell lineages. More in detail, we observed in smokers a global increase in the relative proportions of CD8 T cells, in particular those expressing the CD25 activation marker including DP T cells, and in the CD8+ memory T cells expressing the CCR4 chemokine receptor. In contrast, relative proportions of CD4+ T cells were sensibly decreased in smokers. In B cells, we found an increase in the relative proportion of class-switched memory B cells isotype IgA, IgG, and IgE, and a decrease of plasma cells.

An increase of activated CD8+CD25+ T cells has already been observed in the peripheral blood of smokers (47), in the airway epithelium of smokers with chronic bronchitis (48) , as well as in lung cancer smokers (49). The CD8+CD25+ activated T cells in our results displayed also proliferating (CD127+) and regulatory phenotypes (CD73+) with a predominance of the first. Studies suggest that these latter cells showed an elevate immunosuppressive capacity, able of inhibiting the proliferation of effectors, CD4+ and naïve T cells (50;51), whilst CD8+ with a not-activated phenotype (CD25-) are preferentially naïve or resting T cells (52) .

In the present study, we also showed evidence for the association of CD8+CD4+CD25+ DP and CD8+CD25+ memory T cells with smoking. In the healthy human thymus, activated CD25+ DP T cells displayed suppressive functions similar to regulatory T cells (53). The presence of CD8+CD25+ memory T cells in smokers is in line with their proliferation following antigen recognition due to the systemic inflammation associated with smoking (7;9). Taken together, these results suggest that smoking increases the activation of CD8+ T cells stimulating their proliferation, with concomitant activation of CD8 T reg and DP T cell with immunosuppressive properties.

We have also identified a novel positive association between smoking status and CD8+ memory T cells expressing the CCR4 chemokine receptor. While the role of the CCR4+ receptor in CD8+ memory T cells remains unexplored, studies in animal models showed that cigarette smoking induces the production of CCR4+ ligands in macrophages and dendritic cells, which in turn recruit monocytes in the lung (54), and induce the activation of natural killer cells through the mediation of the CCR4 receptor (55). These findings indicate that the CCR4+ receptor may be involved in smoking response by recruiting CD8+ T cells in inflammatory sites.

As an additional finding, we observed an increase of the relative proportion of CD8+ memory T cells, in particular central, transitional and long-term memory subsets, and a decrease of T-helper (Th) cells such as Th2 and Th17 T cell subsets in active smokers. Increase of CD8+ memory T cells and reduction of Th2 in smokers are in line with previous studies indicating a cumulative effect of smoking causing an increase of memory T cells, as well as an immune suppressive effects decreasing Th-2 response (7;9). Conversely, we found a decrease in the relative proportion of Th-17 and CD8+ expressing CD161 marker displayed a phenotype similar to Th17 cells (56). That, is in contrast with what reported in the literature, which shows that smoking increases the number of Th17 in both lung tissue and peripheral blood with pro-inflammatory effect contributing of smoking-induced inflammation (7;9)

A previous study in 20 chronic obstructive pulmonary disease patients and 29 healthy individuals reported an increase of class-switched memory B cells isotype IgA in peripheral blood of current smokers compared to former and never smokers (57). The authors have suggested that the increase of class-switched memory B cells may be the result of chronic inflammation due to continued smoking and might be associated with the formation or release of (neo)antigens, such as smoke particles or damaged lung tissue, also hypothesizing that a continued smoking exposure may cause a secondary immune response increasing the circulating class-switched memory B cells and memory B cells formation in current smokers (57;58). On the other hands, smoking also affects immunoglobulins production showing a decreased level of IgA and IgG in peripheral blood and saliva of smokers (7). Taken together our results support

these findings: we observed, in current *vs* never smokers, an increase of the relative proportion of class-switched memory B cells (IgA, IgG, and IgE isotypes) and a decreased of the relative proportions of plasma cells which releasing in circulation the immunoglobulins in response to antigens recognition (59).

In the present study, we observed a decrease of CD4+ relative proportions expressing the CD38 activation marker. This latter is preferentially expressed by naïve CD4+ T cells and shows a reduced capacity to proliferate and to respond to IL-2 cytokine signalling (60). Interleukin IL-2 homing regulation and proliferation of T cells (61). We can speculate that smoking-inflammation increase IL-2 levels in circulation promote CD8 T cells activation and decreasing number of CD4+CD38+ T cells in current compared to never smokers might be due to an impaired response to IL-2, or the suppressor activity performed by CD8 T reg induced by smoking (51). In contrast with our results, Valiathan *et al*., showed, in smoking HIV patients, an increase of CD4+CD38+ T cells compared to those who have never smoked, indicating an effect of smoking on CD4+ T-cell activation and proliferation (62).

Our results were also consistent with a negative association between smoking status and immune traits belonging to Vγ9Vδ2 T cells and monocytes. The decrease of their relative proportions in smokers could be explained by their preferential recruitment in the lung. A study in mice showed that the γδ T cell frequency in the lung was increased in response to chronic smoke exposure (63), while monocytes are known to be recruited from the blood for generating alveolar macrophages (7)

Interestingly, analysing the former smokers data, we observed that immune trait relative frequencies including the majority of the B cells, all CD4+, DP, γδ T cells and monocytes were restored to non-smokers frequencies after smoking cessation. In contrast, the CD8+CD25+ and CD8+ memory T cells subsets displayed not a complete restoration. In particular, we found activated CD8+CD25+ and CD8+CD25+ memory T cells and those who exhibited the proliferation activity were partially restored, while the T reg phenotype (CD8+CD25+CD73+) was completely restored (**Supplementary, Table S2**). This is in contrast with other studies, one on 174 former smokers who had quit smoking an average of 10.7

years before the study have shown that smoking cessation completely restores the CD8+, CD3, B cell, monocyte counts within one year, while the proportion CD4+ T cells after two years (37). Whereas another two studies displayed a complete restoration of white blood cell count after one year in 231 former smokers, and lymphocytes and monocytes count within 2-5 years after smoking cessation (64).

Our results suggest that the dysregulation of the immune system connected to active smoking may persist, at different degrees, also after smoking cessation, especially in the CD8+CD25+ and CD8+ memory T cells subsets. This could help to explain why the risk of smoking-related pathologies remains also elevated after smoking cessation (65). Despite a more elevated samples size, the studies mentioned above were limited only to primary leukocyte subpopulations or the main groups of blood cells (*i.e.*, lymphocytes, monocytes, neutrophils, eosinophils and basophils) which indicated the smoking effect only on large cell proportions. The difference between our and these results highlight the importance to investigate the smoking effects also on finely detailed leukocyte subpopulations.

We are aware that this study presents some limitations. First, the study cohort included only women of European ancestry, and the effect of smoking on the immune traits may differ in males or in non-European populations. Second, information on smoking status was self-reported and information on second-hand smoking is missing, making it impossible to exclude residual confounding or misclassification bias. Third, our dataset contains a small number of current smokers (n=25) which makes it impossible to investigate the effect of smoking dose (*e.g.*, pack/years) on the immune traits. However, to overpass at least partially these limitations, the selection of subjects for this study was extremely accurate to exclude potential confounders. Indeed, we validated the self-reported smoking status using historical data on smoking habits as well as on the age of starting/quitting smoking, the number of cigarettes and/or packs smoked, and discarded any subject presenting immune-mediated inflammatory diseases (*i.e.*, rheumatoid arthritis, systemic lupus erythematosus, multiples sclerosis, type 1 diabetes), which may affect the immune traits independently from smoking exposure.

In conclusion, in the present study, we detail how tobacco smoking shapes leukocyte cell subset proportions and induces changes in their surface protein expression levels. The shift of immune cells composition in peripheral blood caused by active smoking affects mainly the CD8 T cell lineage that skew towards a chronic inflammatory phenotype. Finally, we observed that changes induced by smoking are not completely reverted even after years since smoking cessations. Further investigations are required to dissect the role of these immune traits in smoking response for a future application and investigation in smoking- and immune-related diseases.

Smoking effects on DNA methylation in leukocyte subpopulations

5. Formalization and definition of a pipeline for DNA methylation analyses

In this Chapter, we explain the approach used to develop a computational pipeline for DNA methylation profiling at cell-type level using target bisulfite sequencing data.

5.1 Introduction and aims of the work

We have previously described the several methods to study DNA methylation with bisulfite treatment followed by next-generation sequencing (Bs-seq) which is the most popular and powerful method to profile genome-wide DNA methylation patterns at single base resolution. In particular, targeted bisulfite sequencing (targeted-BS) allows to sequence specific genomic regions of interest reducing costs and increasing the coverage. Bisulfite treatment converts unmethylated Cs into Ts while the other bases remain unaffected. Bisulfite conversion alters about 90% of cytosines present in the genome. At this point, distinguishing between Cs converted into Ts and a Ts originally present in the DNA molecule is computationally demanding. In addition, the depletion of unmethylated cytosines brings a challenge for aligning bisulfite-converted sequencing reads to a large reference genome and standard short-read aligners are not suitable for Bs-seq reads. On top of that, it is difficult to distinguish a converted C from: *i)* a stochastic sequencing error occurring during all the sequencing steps; *ii)* a Single-Nucleotide Polymorphisms (SNPs). SNP is a DNA sequence variation occurring at nucleotide level (*e.g.* an adenine instead of a guanine) at a frequency >1% in the general population. The presence of SNPs in the samples increases the level of variability of data.

At the moment, in literature a benchmark pipeline for targeted bisulfite data is still lacking. There are many tools available for both aligning reads to reference

genome and extracting methylation information from the reads. However, to select the proper one to use is still difficult, due to the extensive biological and technical variability of the data. Since BS experiments are time-consuming and expensive, the use of synthetic sequencing data (*i.e.,* the creation of a dataset that simulates different biological and technical situations of a BS experiment) has become increasingly popular for assessing and validating bioinformatics tools. To our knowledge, currently there is only one software for generating BS-seq data called *Sherman* (66). However, this tool allows the synthetic dataset generation without reporting any information about the methylation levels of cytosines.

Therefore, the aims of this part of the doctoral work were: *i)* the generation of a simulator to create bisulfite synthetic data; *(ii)* a comparison of the most used tools for DNA methylation data analysis on a synthetic dataset; and *(iii)* a test of the tool performances analysing real datasets.

5.2 State of the art

To achieve the goals of the present studyit is important to know the currently available methods to simulate a synthetic bisulfite dataset and the most used DNA methylation analysis pipelines.

Synthetic NGS data generators

Computational methods can be benchmarked using real and/or synthetic data. A validation with real data is always essential because they represent the biological complexity. However, a validation with real data is still challenging, since the true values are unknown. This complicates their use for performance evaluation of a tool such as accuracy and precision in detecting results.

In this respect, synthetic data generators allow to produce as much data as desired with predefined parameters, for which the true values are known. Artificial datasets permit to generate a big volume of data in a cheap and fast way, compared to costs and time needed to create real datasets.

Synthetic data generators create FASTQ files starting from a given reference genome. They allow to specify a variety of parameters, such as the NGS platform, the read length and the sequencing mode, as well as the coverage or the

sequencing error quantity. A FASTQ file is the standard format to store data sequenced by an NGS system. FASTQ format describes each read through the following three parameters: *i)* a unique identifier; *ii)* a DNA string; and *iii)* a quality score string associated to the sequence. The quality score is measured by Phred. Phred score is a positive value, it represents the estimated probability that the given nucleotide is incorrectly called.

There are several tools to simulate standard NGS data in FASTQ format, like ART (67) and CuReSim (68). However, such tool abundance is not available for bisulfite sequencing data. As mentioned before, there is a unique tool allowing the production of bisulfite sequencing data, called *Sherman* (66). *Sherman* produces only a FASTQ file containing data and it does not report the file related to methylation calling for each sequenced cytosine. It allows to create directional and non-directional libraries with single- and paired-end reads. The user can set number, length and quality of reads, as well as SNPs and sequencing errors can be specified. Bisulfite conversion can be regulated with two parameters, which give the conversion rate for CG and non-CG contexts. However, *Sherman* does not allow the simulation of targeted bisulfite sequencing experiments, it allows only whole-genome ones because it is not possible to select a set of specific fragments from the reference genome. Finally, it is not a parallel tool and runs only in sequential mode.

Bisulfite aligner and Methylation callers

We selected the most used aligner and methylation extractor tools: Bismark (27) and BSMAP (31).

Bismark aligner implements the three-letter algorithm and uses Bowtie (69) or Bowtie 2 (70) as its core read aligner. These two standards read aligners create an index based on Burrows-Wheeler transform (BWT) on the reference genome, which is used to perform an efficient search of the reads. Bowtie is a short-read aligner, which uses a Full-text index in minute space (FM-index) (71) with some improvements to allow the presence of mismatches. The FM-index is a compressed suffix array based on BWT that allows to efficient search all the occurrences of a pattern in a text in sublinear time. Bowtie extends the FM-index implementing a backtracking algorithm that allows a limited number of

mismatches and favors high-quality alignments, combined with a strategy to avoid excessive backtracking. The search proceeds similarly to the exact match. If the matrix range becomes empty, the algorithm greedily selects an already matched position with a minimal quality value and substitutes it with a different base. Then, the exact match search resumes from just after the substituted position. Bowtie 2 (70) is a read aligner which extends the FM-index to permit gapped alignments. For each read, Bowtie 2 extracts some substrings, called seeds, from the read and its reverse complement. These seeds are aligned to the reference genome in an ungapped fashion using the FM-index and matrix ranges are calculated. At each range is assigned a priority based on its size: smaller ranges receive a higher priority. Then, Bowtie 2 chooses rows randomly correspondingly to range with high priority and resolves each selected row's offset into the reference genome. Finally, it performs a parallel dynamic programming alignment algorithm until a sufficient number of alignments are examined.

The first time one is using a certain reference genome, Bismark needs to build two *in silico* versions of it, by applying the C-to-T and G-to-A conversions. From each of the two variants, an index based on BWT is built and it is stored in mass memory for future uses. Before the alignment step, reads are transformed into fully bisulfite-converted versions, applying the same transformations as the genome. Then, each of them is aligned to the two versions of the reference genome using four parallel instances of Bowtie (**Figure 5.1).** This procedure enables Bismark to find the strand of origin of a read. Bismark first tests whether a sequence can be aligned to multiple places in the reference with a minimum number of mismatches. In case the read cannot be uniquely placed it will be discarded. Otherwise, Bismark determines the sequence with the lowest number of mismatches from any of the four alignments. The methylation state of cytosines is inferred by comparing the original read sequence with the corresponding genomic sequence (27).

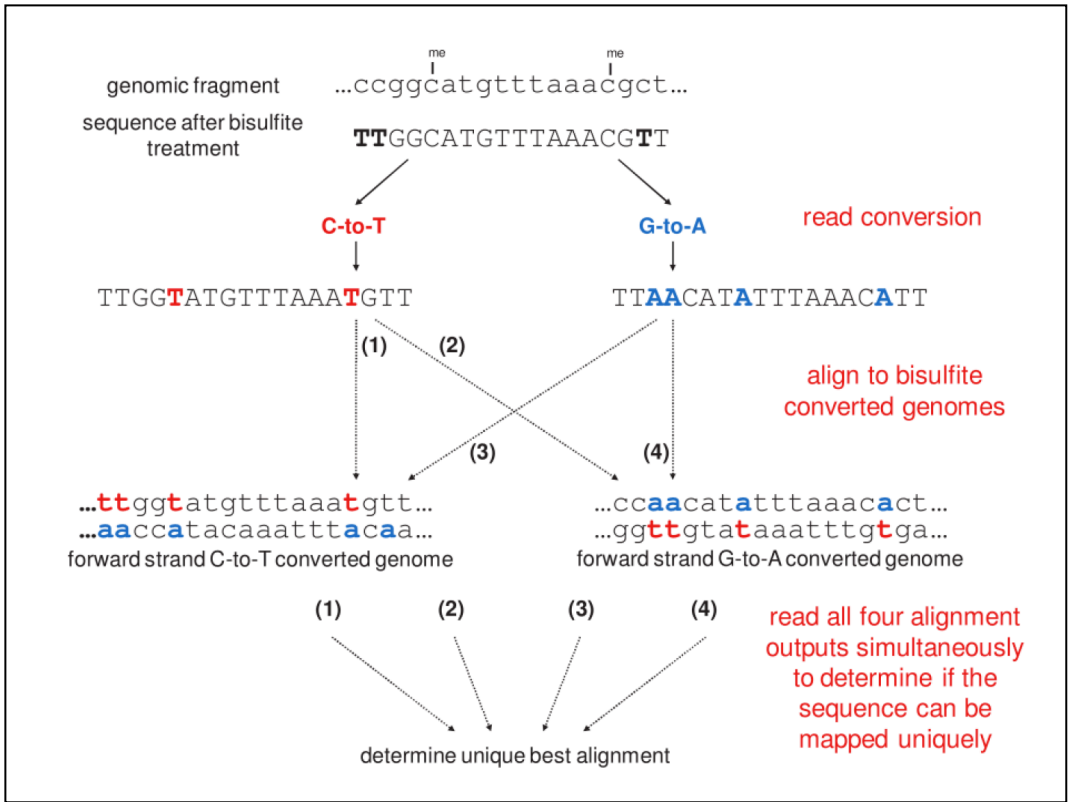


Figure 5.1: Bismark approach to bisulfite mapping. The bisulfite read is converted into a C-to-T and a G-to-A version. These two converted reads are aligned to C-to-T and G-to-A versions of the reference genome. The best unique alignment, if existing, is determined from the four-parallel alignments. It comes from C-to-T read against C-to T reference. Adapted from (27)

BSMAP implements the wild card algorithm: it masks Ts in the read as Cs only at C position in the reference genome, while keeping all the other Ts in the reads unchanged. To do so, BSMAP firstly indexes the reference genome building a hash table which contains all possible k-mers, called seeds. Where k is the length of the substring. The generic entry of the hash table has the seed as a key and the corresponding coordinates as values. To accomplish the C/T mapping issue, the hash table includes all possible bisulfite variants for each seed. To find all possible mapping positions, each read is divided into 4 parts, which are combined two by two to form 6 possible seeds; these seeds are then looked up in the hash table. For each mapping location, the number of mismatches between the read and the reference needs to be counted, allowing a T in the read to map to a C in the reference. DNA sequences are represented as binary strings and DNA nucleotides

are encoded on two bits according to the following encoding: A: 00, C: 01, G: 10, T: 11.

A bitwise masking approach is applied to allow the C/T mapping and to count the number of mismatches. Specifically, a bitwise AND mask (01) is applied to convert a T (11) to C (01) in the read where the base corresponds to a C in the reference, or to keep a C in the read unchanged where the reference is a C. An AND mask (11), which does not change anything, is used where the reference base in not a C (**Figure 3.2**). Mismatch counting is implemented through a bitwise XOR operation between the masked read and the reference. The bitwise XOR of two bits returns zero if they are equal and non- zero if they are different. The number of mismatches between the reference and the read is the number of non-zero two-bit segments (**Figure 5.2**) (31).

In summary, Bismark and BSMAP differ in terms of alignment strategies, Bismark applies a *three-letter* approach whereas BSMAP uses a *wild card* approach.

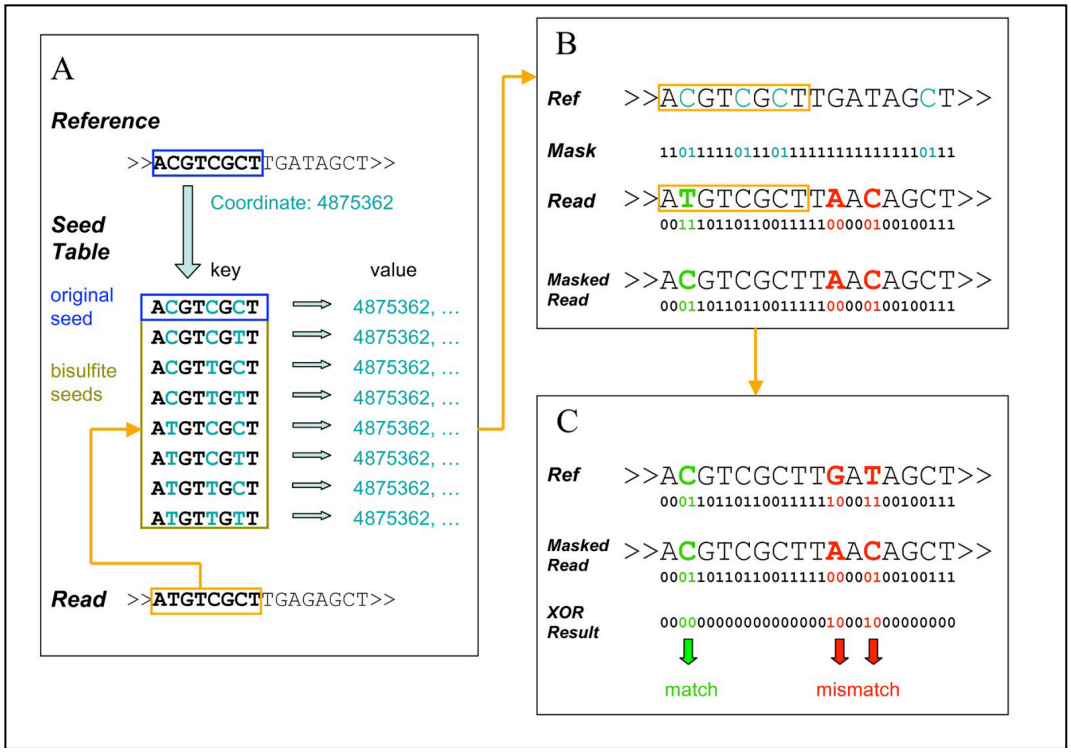


Figure 5.2: BSMAP approach to bisulfite mapping. A) Bisulfite seed table, each read is looked up in the seed table for potential mapping positions; B) Ts of the reads that match over Cs in the reference are masked as Cs through a bitwise AND; C) A bitwise

XOR between the reference and the masked read is calculated. Masked Ts appear as Cs and do not count as mismatches. Adapted from (31).

5.3 Computational implementation

This section is divided into three subsections: In the *first* subsection, we present a formalization of the bisulfite sequencing workflow. It shows how bisulfite data are generated and it was used as input to design a new tool for the generation of bisulfite synthetic datasets. In the *second* subsection we describe the software characteristics and how does works MethylFASTQ. In the *third* subsection, we show the comparison between BSMAP and Bismark performances using synthetic datasets.

5.3.1 Formalization of Bisulfite Sequencing workflow

Here, we described the formalization of the main steps of a typical bisulfite sequencing workflow. The formalization follows the workflow represented in **Figure 5.3**

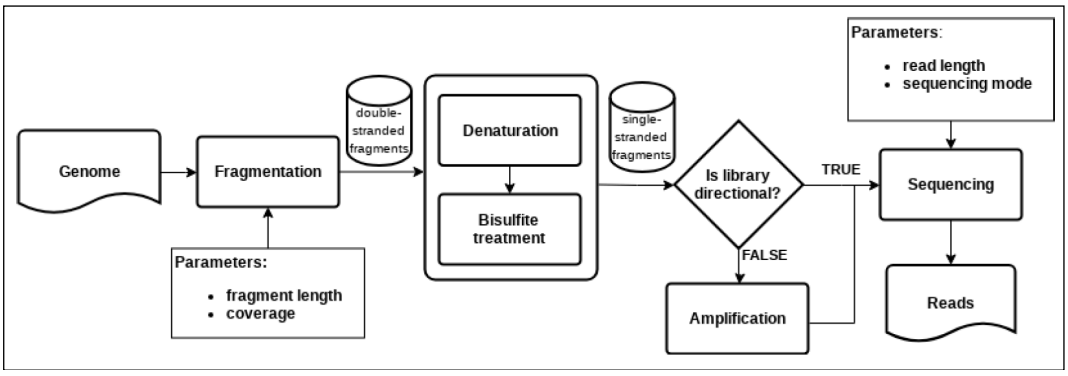


Figure 5.3: Bisulfite Sequencing workflow. The genome of interest is fragmented into a number of double-stranded pieces of known length. Fragment strands are separated through denaturation and then, single-stranded fragments are bisulfite-treated. Amplification produces reverse complement of treated fragments, which are sequenced in the non-directional protocol. Sequencing step processes bisulfite fragments and produces a set of reads.

A *genome* is defined as a pair of strings of length *N*, that represent respectively forward and reverse strand. They are also called plus and minus strand. The two strands are complementary to each other.

Let

$$L = \{A, C, G, T\}$$

be the **nucleotide alphabet**. This alphabet is the classic one that is used to characterize DNA sequences. It has one symbol for each nucleotide: A for adenine, C for cytosine, G for guanine and T for thymine. In the context of DNA methylation, it is necessary to distinguish between a methylated cytosine and an unmethylated one. The nucleotide alphabet does not ensure that: the C symbol represent a generic cytosine, which may be either methylated or non-methylated. So, let us introduce the **methylated nucleotide alphabet,** a new alphabet which allows this distinction:

$$\mathcal{L}_m = \mathcal{L} \cup \{C^m\} = \{A, C, C^m, G, T\}$$

where C and C^m represent respectively non-methylated and methylated cytosine. An asymmetric mapping exists between these two alphabets. A string defined on L_m can be mapped in L, while the opposite is not possible. The mapping from L_m to L has the meaning of nucleotides read from an NGS system. In fact, NGS devices do not reveal the epigenetic mark on DNA unless a suitable pretreatment is done. So, methylated and unmethylated cytosine will be both read as cytosines.

The opposite mapping, from the nucleotide alphabet to the methylated nucleotide alphabet does not exist, because it would be like setting DNA methylation in an arbitrary way.

Let us introduce strands complementarity relation by means of the **reverse complement function**

$$f_{rc} : \mathcal{L}^{\mathbb{N}^+} \rightarrow \mathcal{L}^{\mathbb{N}^+}$$

Given a nucleotide string $s = b_1 b_2 \ldots\ldots b_n$, the reverse complement of s is defined as

$$f_{rc}(s) = s' \quad \text{s.t.} \quad s' = f_c(b_n)\, f_c(b_{n-1}) \ldots f_c(b_2)\, f_c(b_1)$$

where $f_c : L \mapsto L$ is the complementary nucleotide function defined as:

$$f_c(b) = \begin{cases} \text{A} & \text{if} \quad b = \text{T} \\ \text{C} & \text{if} \quad b = \text{G} \\ \text{G} & \text{if} \quad b = \text{C} \\ \text{T} & \text{if} \quad b = \text{A} \end{cases}$$

The reverse complement function can also be defined on the methylated nucleotide alphabet *L*m in a similar way to the previous case. The difference is that the complementary nucleotide of G can be either C or C^m, while both C and C^m have G as a complementary.

Let's consider a genome *g* of length *N*. Formally,

$$g = (s_p, s_m) \in \mathcal{L}_m^N \times \mathcal{L}_m^N \quad \text{s.t.} \quad s_p = f_{rc}(s_m)$$

Here, *N* represents the number of nucleotides which constitute *g*, *sp* is the plus strand and *sm* is the minus strand. Furthermore, let us introduce two functions to select one specific strand of a double-stranded DNA string. Let

$$f_+ : \mathcal{L}_m^{\mathbb{N}^+} \times \mathcal{L}_m^{\mathbb{N}^+} \to \mathcal{L}_m^{\mathbb{N}^+} \quad \text{and} \quad f_- : \mathcal{L}_m^{\mathbb{N}^+} \times \mathcal{L}_m^{\mathbb{N}^+} \to \mathcal{L}_m^{\mathbb{N}^+}$$

be the functions that return respectively the forward and the reverse strand of a given DNA string.

- *Fragmentation*

The genetic material under study is obtained from a group of cells belonging to the same tissue, (*i.e.,* plasma, epithelium, lung, etc.). These different types of DNA are randomly broken into double-stranded fragments of a given length.

Let us call $G = \{g_1, g_2, \ldots\ldots g_p\}$ the sample, namely the pool of starting genomes. Each g_i is randomly fragmented into pieces of length *l*. Let

$$F_i = \{d_1, d_2, \ldots, d_q\} \quad \Big| \quad \forall j \in \{1, 2, \ldots, q\}, \quad d_j \in \mathcal{L}_m^l \times \mathcal{L}_m^l$$

be the set of fragments of i-th genome. Let

$$F = \biguplus_i F_i$$

be the multiset of fragments of sample *G*.

Let us consider a double-stranded fragment to use as an example. Let us denote forward strand as *s*+ and reverse strand as *s*-. They have respectively 5'-3' and 3'-5' direction.

Example 4.1.1.

s_+ 5'–AGCCAGGAACTATTGAGTA$\overset{m}{C}$GAA$\overset{m}{C}$GCCATCGACTTACG–3'

s_- 3'–TCGGT$\overset{m}{C}$CTTGATAACT$\overset{m}{C}$ATGCTTGCGGTAG$\overset{m}{C}$TGAATG$\overset{m}{C}$–5'

- *Bisulfite treatment*

Fragments are denatured to separate the two strands and then, they are treated with sodium bisulfite to highlight cytosine methylation. Specifically, the treatment converts unmethylated cytosines into thymines while it does not affect other nucleotides.

Let

$$f_{bs} : \mathcal{L}_m^n \rightarrow \mathcal{L}^n$$

be the **bisulfite function**. It transforms a string defined on the alphabet L_m in a string defined on the standard nucleotide alphabet *L*.

Given a single-stranded fragment $s = b_1 b_2 \ldots b_n$, the bisulfite function application on s results in a new string $s' = b'_1 b'_2 \ldots b'_n$ such that

$$b'_i = \begin{cases} \mathrm{T} & \text{if} \quad b_i = \mathrm{C} \\ \mathrm{C} & \text{if} \quad b_i = \mathrm{C}^m \\ b_i & \text{otherwise} \end{cases}$$

Let us apply bisulfite function to the fragment of example 4.1.1.

Example 4.1.2.

s'_+ 5'–AGTTAGGAATTATTGAGTACGAACGTTATTGATTTATG–3'
s'_- 3'–TTGGTCTTTGATAATTCATGTTTGTGGTAGCTGAATGC–5'

Let define F_{bs} as the multiset of bisulfite treated fragments. The generic double-stranded fragment is denatured, and the bisulfite function is applied on both its strands.

$$F_{bs} = \biguplus_{s \in F} \{f_{bs}(f_+(s)), f_{bs}(f_-(s))\}$$

- *Amplification*

Amplification step produces many copies of each fragment and of its reverse complementary.

The NGS system needs the multiple copies of each fragment to reinforce the optical signal in order to detect it. The purpose of this formalization is to describe how bisulfite data are generated, not the NGS system's workflow. Thus, the creation of the copies of each fragment is not relevant for the formalization purpose.

For non-directional libraries, it is relevant the creation of the reverse complementary of each fragment.

In case of directional library, let us define the set of fragments that will be sequenced as the set of fragments obtained from the bisulfite treatment step.

$$F_{seq} = F_{bs}$$

In case of non-directional library, let us define the set of fragments that will be sequenced as the set of bisulfite fragments with their reverse complementary fragments.

$$F_{seq} = F_{bs} \uplus \biguplus_{s \in F_{bs}} f_{rc}(s)$$

Let's apply amplification step to the bisulfite treated fragments of example 4.1.2. Let us call the new fragments $s^{rc}+$ and $s^{rc}-$.

Example 4.1.3.

s'_+ 5'–AGTTAGGAATTATTGAGTACGAACGTTATTGATTTATG–3'
s^{rc}_+ 3'–TCAATCCTTAATAACTCATGCTTGCAATAACTAAATAC–5'

s'_- 3'–TTGGTCTTTGATAATTCATGTTTGTGGTAGCTGAATGC–5'
s^{rc}_- 5'–AACCAGAAACTATTAAGTACAAACACCATCGACTTAGC–3'

- *Sequencing*

Fragments may be sequenced either in single-end or in paired-end mode. Let us call $m \in \mathbb{N}+$ the read length.

Single-end sequencing allows us to sequence the 5`-end of the fragments. So, a single-end read may be described as a string $r \in L^m$ that describes the first *m* nucleotides of the fragment. Conversely, paired-end sequencing allows us to sequence both ends of each fragment. So, a generic paired-end read may be described as a pair of strings:

$$(r_1, r_2) \in \mathcal{L}^m \times \mathcal{L}^m$$

The read r_1 contains the first *m* nucleotides of that fragment, while read r_2 contains the first *m* nucleotides of the reverse complement of the same fragment.
Let

$$f_{pref} : \mathcal{L}^{\mathbb{N}+} \times \mathbb{N}^+ \to \mathcal{L}^{\mathbb{N}+}$$

be the prefix function such that it returns the first $m \leq n$ characters of the input string. Given a generic string $s = b_1 b_2 \ldots b_n$ of length *n,* the prefix of length $m \leq n$ of *s* is defined as

$$f_{pref}(s, m) = s' \quad \text{s.t.} \quad s' = b_1 b_2 \ldots b_m$$

Let us define the sequencing functions.
Single-end sequencing function

$$f_{se} : \mathcal{L}^{\mathbb{N}+} \times \mathbb{N}^+ \to \mathcal{L}^{\mathbb{N}+}$$

is defined as

$$f_{se}(s,m) = r \quad \text{s.t.} \quad r = f_{pref}(s,m)$$

Paired-end sequencing function

$$f_{pe} : \mathcal{L}^{\mathbb{N}^+} \times \mathbb{N}^+ \to \mathcal{L}^{\mathbb{N}^+} \times \mathcal{L}^{\mathbb{N}^+}$$

is defined as

$$f_{pe}(s,m) = (r_1, r_2) \quad \text{s.t.}$$
$$r_1 = f_{pref}(s,m) \quad \text{and} \quad r_2 = f_{pref}(f_{rc}(s), m)$$

Let us consider a read length of m = 15 to conclude the example 4.1.3

Let's begin with single-end sequencing.

Example 4.1.4.

$f_{se}(s'_+, m)$	AGTTAGGAATTATTG
$f_{se}(s^{rc}_+, m)$	CATAAATCAATAACG
$f_{se}(s'_-, m)$	CGTAAGTCGATGGTG
$f_{se}(s^{rc}_-, m)$	AACCAGAAACTATTA

As regard to paired-end sequencing, we obtain the following reads.

Example 4.1.5.

	r_1	r_2
$f_{pe}(s'_+, m)$	AGTTAGGAATTATTG	CATAAATCAATAACG
$f_{pe}(s^{rc}_+, m)$	CATAAATCAATAACG	AGTTAGGAATTATTG
$f_{pe}(s'_-, m)$	CGTAAGTCGATGGTG	AACCAGAAACTATTA
$f_{pe}(s^{rc}_-, m)$	AACCAGAAACTATTA	CGTAAGTCGATGGTG

Let f_{seq} be the chosen sequencing function and let m be the read length. So, the set of sequencing reads R is defined as:

$$R = \biguplus_{s \in F_{seq}} f_{seq}(s,m)$$

The cardinality of *R* is the number of reads obtained from sequencing, let us call it *N*. Let's call *L* the read length and *G* the genome length. So, the depth of coverage *C* is defined as $C = NL/G$

5.3.2 *Methyl FASTQ*

Tool overview

MethylFASTQ is a tool written in Python that generates synthetic bisulfite sequencing data in FASTQ format (72)(Detailed information are reported in, **Supplementary**, **Appendix A**). It is both organism and experiment independent. MethylFASTQ is designed to simulates the sequencing process following the bisulfite sequencing experiment workflow.

MethylFASTQ simulates both whole-genome (WGBS) and targeted bisulfite sequencing processes and in directional and non-directional manner. Also, it allows the production of single- end paired-end reads. In the WGBS mode, the user can provide a list containing the chromosome names that have to be sequenced. If no list is provided the entire reference genome will be sequenced. While in targeted mode the user can provide a tabulated file including the genome regions to be sequenced.

The dataset includes the setting of both mutation rate that represent SNPs, insertion and deletion and structural variation and the sequencing errors simulating the errors of NGS system. Generally, they are associated with low quality score of data.

Methylation levels can set in the different contexts (*i.e.,* CG, CHG and CHH).

Finally, MethylFASTQ produces two files: *i)* a FASTQ file containing data, and *ii)* methylation call file. These files are tabulated files reporting the information in the same format as those from real data.

Software architecture

MethylFASTQ is a parallel tool. The parallelization is process-based and utilizes

the built-in module multiprocessing, which supports spawning processes and assigning them a job through a function. The architecture follows producer-consumer software design pattern (**Figure 5.4**). Child processes (producers) produce the data and send them to the parent process (consumer) using a FIFO shared queue. The number of concurrent processes is a tool parameter.

MethylFASTQ is composed by:

- *Load balancer*: it is a parent process that used an heuristic approach, where the chromosome sequence is separated in substring using biological and mathematical knowledges. The load balancing step spilt the extracted substring in order to equally distribute the workload among the number of parallel processes. The number of parallel processes can be set by user.
- *Producer*: its job starts by setting the mutation rate given by user. The cytosines on both strands of the sequence are indexed. And, the cytosines information are stored in hash table. Numerous overlapping fragments are generated equal to the depth of coverage setted. Whenever the number of reads present in the buffer is greater than a certain threshold the producer sent a message to the consumer.
- *Consumer*: it receives the data from the producer by shared queue and permanently store them in a file.

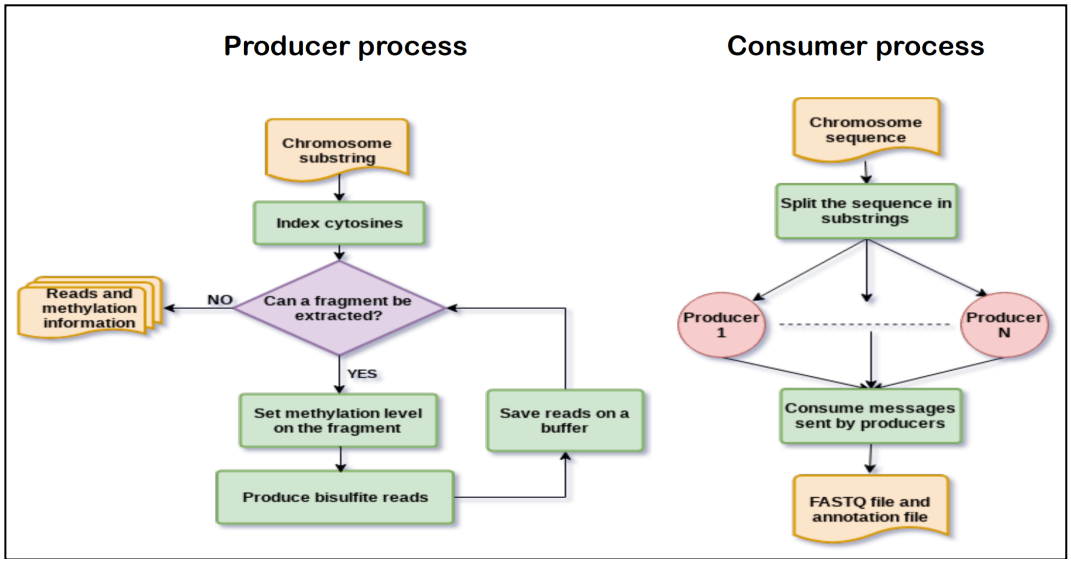


Figure 5.4. MethylFASTQ architecture based on producer-consumer process. Producer process indexed the cytosines present in the chromosome substrings. For each of them methylation is set, and relatives information are stored in the index. Then, the bisulfite fragment is produced, and reads are extracted from it. Reads are stored in a local buffer which is periodically flushed in the queue. When fragments extraction terminates, the consumer pushes in the queue the cytosines information and its execution ends. **Consumer process** starts with the chromosome sequence that is splitted in nonoverlapping substrings, which are further divided by the load balancing algorithm. Obtained substrings are assigned to N producer processes. Then, the consumer waits for items to be available in the queue and elaborate them. When all substrings have been sequenced, the consumer terminates.

Computational tests

Efficiency, together with accuracy and precision**,** is an indicative measure of software performances. Generally, the efficiency is expressed in time needed to complete a task and it depends on a machine workload.

We tested the execution time performances of MethylFASTQ. For each experiment, numerous executions have been performed and the average time has been calculated. Times were expressed in minutes. Runs were performed on 48-core AMD Opteron 6176 CPUs, 2.3 GHz, RAM 503 GB.

We measured MethylFASTQ execution time performance for the generation of: *i)* datasets with different complexity, *ii)* datasets with increasing depth of coverage, and finally *iii)* seven datasets with a different number of parallel processes and, in comparison with *Sherman* tool.

Table 5.1 shows the results on the average creation time on 10 runs to generate four datasets with different complexity. All datasets were generated from human chromosome 21 (*hg19*), in whole genome mode at 10X of coverage using 8 parallel processes. The lower execution time was obtained for creating the dataset with single-end reads of directional library, while the generation of paired-end reads of non-directional library was the most expensive execution. In table 5.2 is reported the average of execution time performances to create the datasets at the same conditions reported above, with different depth of coverage. As expected, the time of execution increases as the depth of coverage increasing. Greater is the depth of coverage and greater will be the number of read to produce.

Table 5.1. MethyFASTQ execution time performances to create the datasets with different complexity. Average time was computed considering 10 runs used to generate the dataset using 8 parallel processes. All datasets were extracted from chromosome 21 (*hg19*), with 10X depth of coverage.

Sequencing	Library	Generation time (min)
single-end	directional	15
single-end	non-directional	24
paired-end	directional	25
paired-end	non-directional	44

Table 5.2. MethyFASTQ execution time performances to create the dataset at the increase of the depth of coverage. Average time was computed considering 5 runs used to generate the dataset using 8 parallel processes. All datasets were extracted from chromosome 21 (hg19), in paired-end reads, in non-directional mode.

Dataset		Average time (min)
Coverage	num. reads	
1	702.720	9
5	3.510.372	25
10	7.014.948	42
15	10.796.256	61
20	14.049.076	76
25	17.549.292	92
30	20.064.748	107

Furthermore, we tested MethylFASTQ in comparison to *Sherman* the already published tool. Both tools generate bisulfite synthetic dataset in a high customizable way. In particular, *Sherman* presents different features compared to MethylFASTQ. For example, *Sherman* produces only a FASTQ file containing data and it does not report the file related to methylation calling for each sequenced cytosine. Moreover, *Sherman* does not allow the simulation of targeted bisulfite sequencing experiment, it allows only whole-genome one because it is not possible to select a set of specific fragments from the reference genome. Finally, *Sherman* is not a parallel tool and run only in sequential mode.

Figure 5.5 reports average execution times to create seven datasets with different sizes. Datasets were extracted to chromosome 21 (*hg19*) and they were generated following non- directional protocol in paired-end reads. Here, we demonstrate the improvement of execution time performance of MethylFASTQ due to the

parallelization of processes in comparison with *Sherman* that runs in sequential mode. With two processes, they obtain a comparable execution time. The execution time performance of MethylFASTQ increases as the number of parallel processes increase.

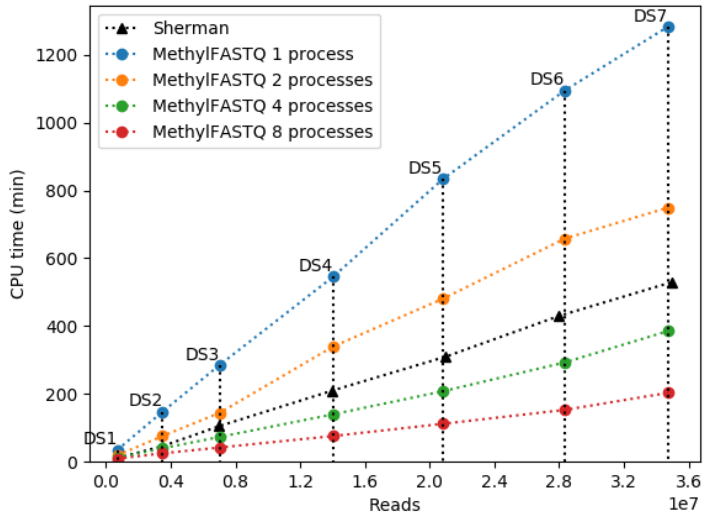


Figure 5.5. MethylFASTQ execution time performances in comparison with Sherman tool. Average times to produce seven different datasets of MethyFASTQ and Sherman. MethylFASTQ has been run with 1, 2, 4, and 8 parallel processes. Datasets were extracted from chromosome 21 (hg19), in paired-end reads, in non-directional mode. Datasets with different size are represented by dots.

5.3.4 Comparison of BSMAP and Bismark

Three metrics were used to assess the performances of Bismark and BSMAP tools on alignment and methylation calling tasks: *i)* **percentage of uniquely mapped reads**, *ii)* **recall of methylation detection** and *iii)* **time consuming in reads alignment.**

Uniquely mapped reads are those reads which map in only one position with a minimum number of mismatches. In case of paired-end reads, the reads are aligned if both the extremities are properly mapped.

Methylation calling is performed by methylation extractors included in both BSMAP and Bismark packages and it was evaluated by the recall. The **Recall** is the fraction of true positive values correctly identified as methylated CG sites. It is

defined as: TP/Pos, where, *TP* is the true positive values identified by the used tools and *Pos* is the total number of cytosines/CpG sites.

Both tools were finally evaluated for **CPU time consumed** for read alignment. Runs were performed on 48-core AMD Opteron 6176 CPUs, 2.3 GHz, RAM 503 GB.

Comparison on synthetic datasets

MethylFASTQ was used to generate 10 datasets from chromosome 21 of the hg19 reference genome. All datasets were generated in non-directional mode, with read length 150bp in paired-end and with 10X of depth of coverage. The read mapping was performed on the entire human genome using *hg19* as reference. Features of the 10 datasets were reported in the **Table 5.3**, they were generated with different percentage of mutation and sequencing errors. Some of them are similar to biological reality while others reported elevated number of mutation and sequencing error to stress the tool performances.

Table 5.3. Characteristics of synthetic datasets. All datasets were extracted from chromosome 21 of hg 19 genome. They are non-directional datasets with paired-end reads with 10x coverage.

ID	num. reads	SNP rate	Error rate	num. CpG sites
SD1	7.024.152	0.1%	0.1%	766.422
SD2	7.023.824	0.1%	1.0%	766.748
SD3	7.018.280	0.1%	2.0%	766.398
SD4	7.019.916	0.1%	5.0%	766.698
SD5	7.021.892	0.3%	0.1%	777.718
SD6	7.016.484	0.3%	0.5%	778.154
SD7	7.017.776	0.5%	0.1%	789.096
SD8	7.017.556	1.0%	1.0%	817.514
SD9	7.021.028	2.0%	5.0%	873.480
SD10	7.022.140	5.0%	2.0%	1.038.142

We evaluated the tools performances as the number of mutations and sequencing errors increase (**Figures 5.6; 5.7**). In the alignment, BSMAP is more stable compared to Bismark at the increase of sequencing errors and SNPs. Bismark performances vary dramatically at the increase of variations and sequencing

errors. An evident example is on synthetic datasets SD9 and SD10 where Bismark align < 20% of reads. Despite these performances for the alignment, in the methylation detection, Bismark performs a little better than BSMAP. Indeed, Bismark on synthetic datasets SD9 and SD10 with a high level of mutation and sequencing errors the recall is almost 50%.

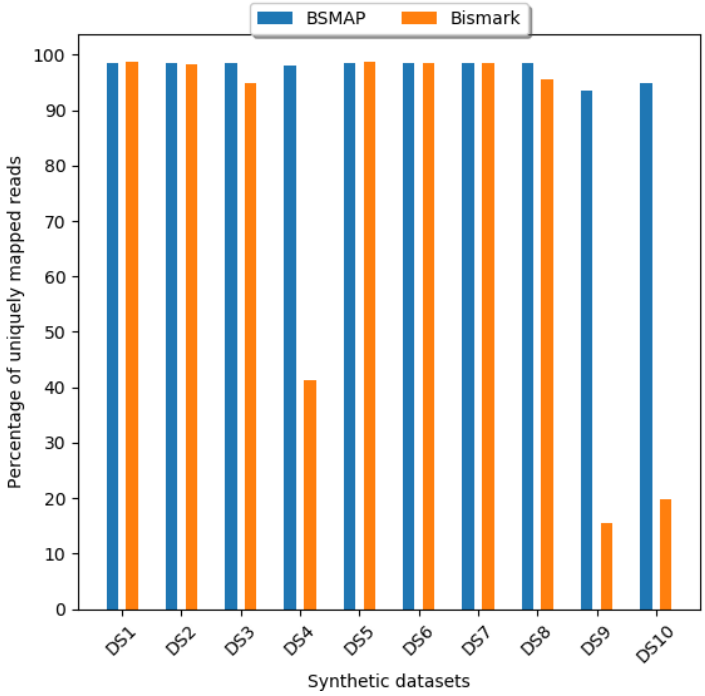


Figure 5.6 BSMAP and Bismark alignment performances on the synthetic datasets.

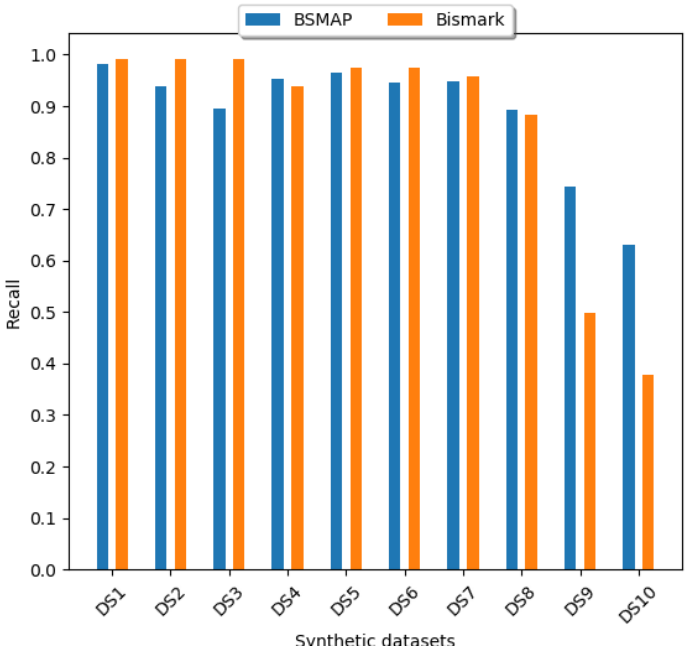


Figure 5.7 BSMAP and Bismark methylation extraction performances on the synthetic datasets.

In the execution time performance of alignment (**Table 5.4**), BSMAP is faster than Bismark: its execution time increases as the number of mutations and sequencing error increase. On the other hand, Bismark shows fast execution time in the datasets with poor quality due to the fact that it maps a short number of reads. These differences are due to different approaches used to reads mapping.
These results of tools performances on synthetic datasets show BSMAP less influenced by low-quality datasets in both alignment and recall and faster than Bismark.

Table 5.4 CPU time used by BSMAP and Bismark to align the synthetic datasets. Alignments were performed with default setting for both tools.

Sample ID	Alignment time (min)	
	BSMAP	Bismark
SD1	11	118
SD2	22	120
SD3	43	114
SD4	191	65
SD5	14	118
SD6	18	119
SD7	16	120
SD8	35	115
SD9	248	49
SD10	226	49
min	11	49
max	248	120
Average	82	99

5.4 Experimental results

We examined BSMAP and Bismark performances in reads alignments, methylation calling and running time on real datasets.

5.4.1 Real targeted bisulfite dataset

Subjects

The dataset is composed by DNA samples of healthy individuals from the EPIC-Italy cohort (73). DNA of 22 samples from buffy coat was analyzed with custom SeqCap Epi Choice S Enrichment Kits proposed by Roche. The custom design of this kit is expected to capture 1054 CpG loci, selected according to their correlation with sex (20 CpGs), epigenetic age (408 CpGs), white blood cells distribution (503 CpG), all-cause mortality (56 CpGs) and smoking habits (67 CpGs).

Methylation levels of these loci were already analyzed using Illumina Infinium 450K BeadChip assay. The procedure is fully described in (17).

Samples and Library preparations

Before the library preparation all DNA samples were thawed and checked for the quality and quantity. The quality control step was assessed by Thermofisher NanoDrop Spectrophotometer observing the ratio of absorbance 260/280 nm. While the samples concentration was determined using BR dsDNA Assay kit and Qbit fluorometer (Invitrogen).

After quality control the samples were ready to be fragmented. The DNA input for all samples was 1 µg to which 5,8 µl of bisulfite conversion control were added. Volume was adjusted for a total of 53µl using buffer EB of QIAGEN and the DNA was fragmented using Covaris M220 set at the following conditions:

- Peak Incident Power (W): 75
- Duty Factor: 10%
- Cycle per Burst: 200
- Temperature: 20°C
- Treatment time (s): 200

High Sensitivity DNA kit on Agilent Bioanalyzer was used to check the correct fragmentation (180-220 bp).

The libraries preparation was performed following steps of SeqCap Epi Enrichment System protocol (Roche). The libraries were prepared in non-directional mode, pooled and sequenced on a Illumina MiSeq platform. Sequencing conditions were read length 150 bp in paired-end mode.

5.4.2 Analysing a real dataset

After read demultiplexing, the quality of sequences was checked with FASTQC (74) tool and Illumina adapters were removed with CutAdapt (75). Trimmed reads were used as input for BSMAP (version 2.90) and Bismark (version 0.19.0) tools. The reads were mapped to the human reference genome (*hg38*) Mapping steps were executed for both tools with the default settings. The number of CpG sites present in the dataset was identified with the methylation extractors present either BSMAP or Bismark packages. CpG sites identified from both tools were filtered for the 1054 known CpG sites.

Comparison on real datasets

A pilot dataset composed of 22 DNA samples of buffy coat from healthy individuals were used to compare the performances of BSMAP and Bismark tools. It derives from a targeted-bisulfite sequencing experiment: the libraries design cover 1054 known CpG methylation sites. Samples were non-directional composed of 150 bp paired-end reads.

Table 5.5 shows the main characteristics of the real samples. Table includes the total number of reads, counting both mates of paired-end reads; the average read quality; and the proportion of read with 0% and more than 5% of nucleotides associated with a low-quality score (*i.e.,* quality score less to 30).

There is a great variability in the number of reads which compose the samples. Especially, between the samples S21 and S7. S21 is the sample with the higher number of reads (1’132’881), while S7 is the one with the lower number of reads (264’571). Half of the reads of each sample has a very good quality, with zero nucleotides with a low-quality score. Whereas, a 25% of each sample is composed of bad quality reads.

Table 5.5 Quality of the real datasets.
All samples showing similar features in reads quality, but they differ in total number of reads for samples.

ID	Sample Quality	num. reads	Read quality 0%	5+%
2460795_S1	36,37	1.089.796	46,56%	26,38%
2462864_S2	36,47	1.132.138	48,64%	25,55%
2468235_S3	36,38	904.106	47,75%	25,52%
2463160_S4	36,43	599.722	48,22%	25,52%
2463313_S5	36,30	1.133.842	44,67%	27,28%
2221350_S6	36,46	1.162.736	48,44%	25,72%
2104434_S7	36,46	529.142	49,83%	24,57%
2101435_S8	36,45	710.768	48,75%	25,47%
2463928_S9	36,4	485.208	47,78%	25,81%
2466541_S10	36,44	1.279.548	48,47%	25,41%
2468401_S11	36,34	1.050.362	46,37%	26,33%
2467013_S12	36,25	1.089.044	45,28%	26,25%
2461447_S13	36,33	1.135.504	46,67%	25,99%
2460308_S14	36,5	883.728	48,99%	25,44%
2229091_S15	36,41	917.690	48,15%	25,47%
2229715_S16	36,51	663.212	50,07%	24,61%
2464281_S17	36,50	1.883.284	49,88%	24,97%
2464163_S18	36,31	1.653.374	46,59%	25,71%
2222529_S19	36,42	1.162.146	48,81%	25,01%
2112435_S20	36,40	2.164.784	47,59%	25,75%
2463312_S21	36,47	2.265.762	48,49%	25,62%
2463056_S22	36,44	2.172.018	47,67%	26,09%
min	36,25	485.208	44,67%	24,57%
max	36,51	2.265.762	50,07%	27,28%
average	36,41	1.184.905	47,90%	25,66%

In the following **Figures 5.8** and **5.9** are reported the performances of tools in **alignment** and **methylation calling** on real dataset. For both tasks both BSMAP and Bismark performs well, even if Bismark overtakes BSMAP. On average, Bismark aligns 81% of reads, while BSMAP only the 78%. Recall values show that Bismark methylation extractor is more performing that BSMAP. On average, their recall values are 98.8% and 96.6%, respectively. Thus, they differ for a 3% in alignments and a 2.2% in recall of methylation.

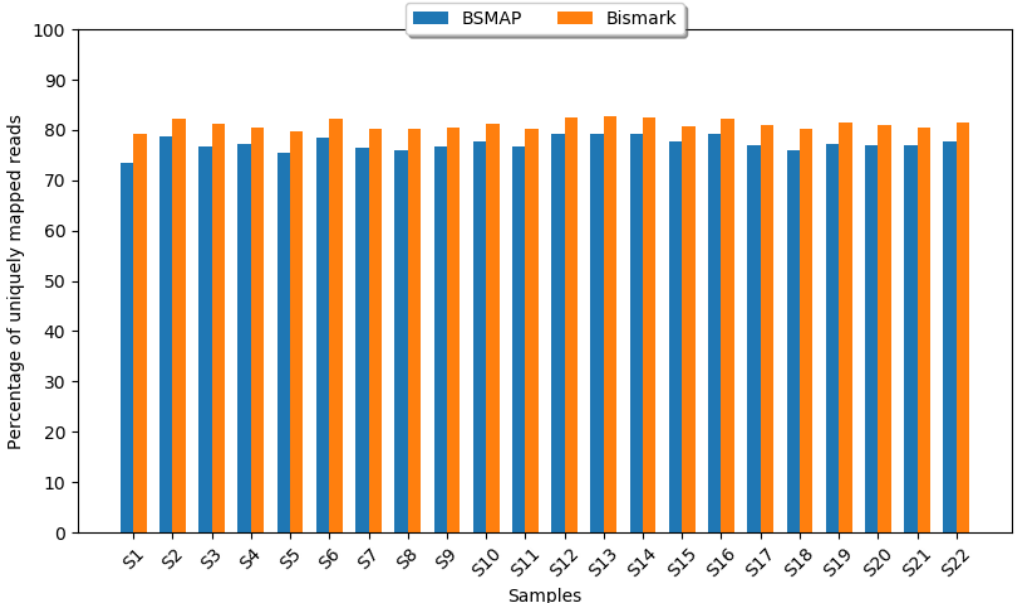


Figure 5.8 BSMAP and Bismark alignment performances on the real datasets.

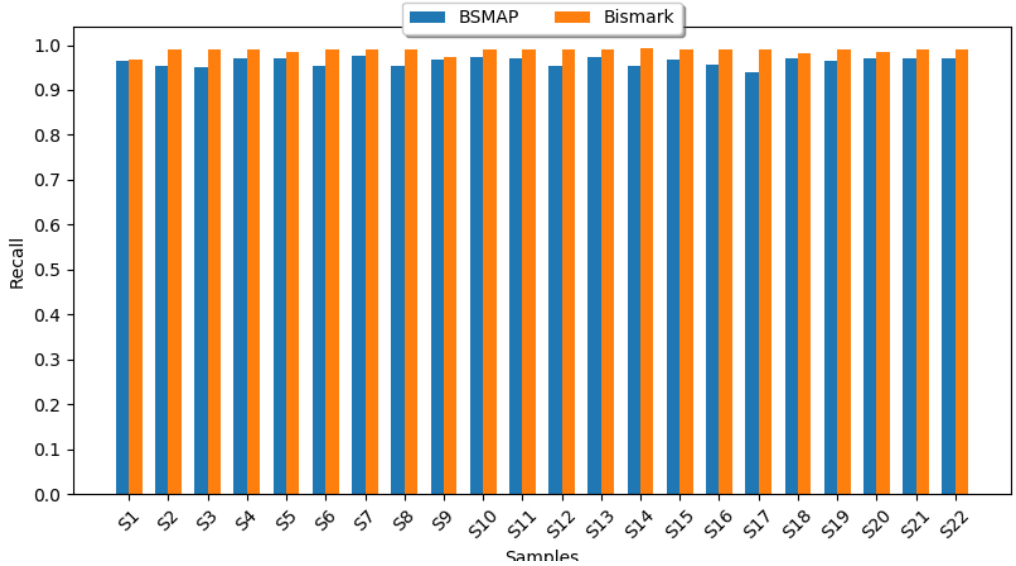


Figure 5.9 BSMAP and Bismark methylation extraction performances on the real datasets.

Table 5.6 shows that both tools identify a common large subset of CpG sites: 96% on average. Although each tool detects an exclusive subset of CpG sites, Bismark spots a relatively large number of them (2.7% on average), while BSMAP just a few (0.22%). Alignment times on the targeted-bisulfite dataset are very similar (**Table 5.7**). There are no big differences for small size samples. However, on average BSMAP is slightly faster than Bismark.

Table 5.6 CpG detected by BSMAP and Bismark in targeted-bisulfite sequencing experiment. Methylation detection of 1054 CpG sites. The second column identifies the proportion of CpG sites detected by both tools over each sample. The last two columns show the proportion of CpG sites detected exclusively by BSMAP or by Bismark, respectively.

Sample	CpG detected by both tools	BSMAP exclusive	Bismark exclusive
2460795_S1	96.49%	0.09%	2.66%
2462864_S2	95.45%	0.00%	3.61%
2468235_S3	96.87%	0.09%	2.28%
2463160_S4	96.58%	0.57%	1.99%
2463313_S5	95.54%	0.09%	3.51%
2221350_S6	95.07%	0.09%	3.98%
2104434_S7	96.87%	0.28%	1.52%
2101435_S8	97.44%	0.19%	1.52%
2463928_S9	95.35%	1.14%	1.42%
2466541_S10	96.96%	0.09%	2.18%
2468401_S11	96.87%	0.09%	2.18%
2467013_S12	96.68%	0.09%	2.47%
2461447_S13	97.06%	0.09%	2.09%
2460308_S14	97.15%	0.19%	1.80%
2229091_S15	97.15%	0.09%	1.99%
2229715_S16	96.30%	0.57%	1.14%
2464281_S17	95.16%	0.19%	3.89%
2464163_S18	95.35%	0.09%	3.80%
2222529_S19	96.58%	0.57%	1.61%
2112435_S20	93.83%	0.00%	5.31%
2463312_S21	95.26%	0.09%	3.98%
2463056_S22	95.16%	0.09%	3.89%
min	93.83%	0	1.14%
max	97.44%	1.14%	5.31%
Average	96.14%	0.22%	2.67%

Table 5.7 CPU time used by BSMAP and Bismark to align the samples of real dataset.
Alignments were performed with default setting for both tools.

Sample		CPU time (min)	
ID	num. reads	BSMAP	Bismark
2460795_S1	544.898	19	23
2462864_S2	566.069	20	24
2468235_S3	452.053	27	18
2463160_S4	299.861	14	14
2463313_S5	566.921	19	25
2221350_S6	581.368	22	25
2104434_S7	26.4571	15	12
2101435_S8	355.384	14	15
2463928_S9	242.604	16	14
2466541_S10	639.774	21	29
2468401_S11	525.181	23	28
2467013_S12	544.522	25	25
2461447_S13	567.752	27	22
2460308_S14	441.864	26	22
2229091_S15	458.845	22	24
2229715_S16	331.606	18	15
2464281_S17	941.642	28	38
2464163_S18	826.687	25	33
2222529_S19	581.073	25	33
2112435_S20	1.082.392	30	45
2463312_S21	1.132.881	31	46
2463056_S22	1.086.009	30	45
min	242.604	14	12
max	1.132.881	31	46
Average	592.453	23	26

5.4.3 Comparison of methylation levels detected by BSMAP with Illumina Infinium (HM450K) BeadChip assay.

Based on the results obtained through tools performances on synthetic and real datasets, we decided to explore the methylation levels estimated by *methratio* packages of BSMAP. Since BSMAP demonstrated a good compromise between alignment percentage and time needed to complete the mapping process.

On the 22 samples, we compared the methylation levels of CpG sites detected by BSMAP with their levels analysed by Illumina Infinium 450K BeadChip data.

Results of the comparison on the 1054 CpG methylation levels are shown in **Figure 5.10.** For 19 out of 22 samples the Pearson correlation coefficients were on average higher than 90% (R=0.95), while for 3 samples thy were resulted

slightly lower (R=0.88) due to a higher dispersion. These results showed that BSMAP obtained a good performance also in the estimation of methylation levels.

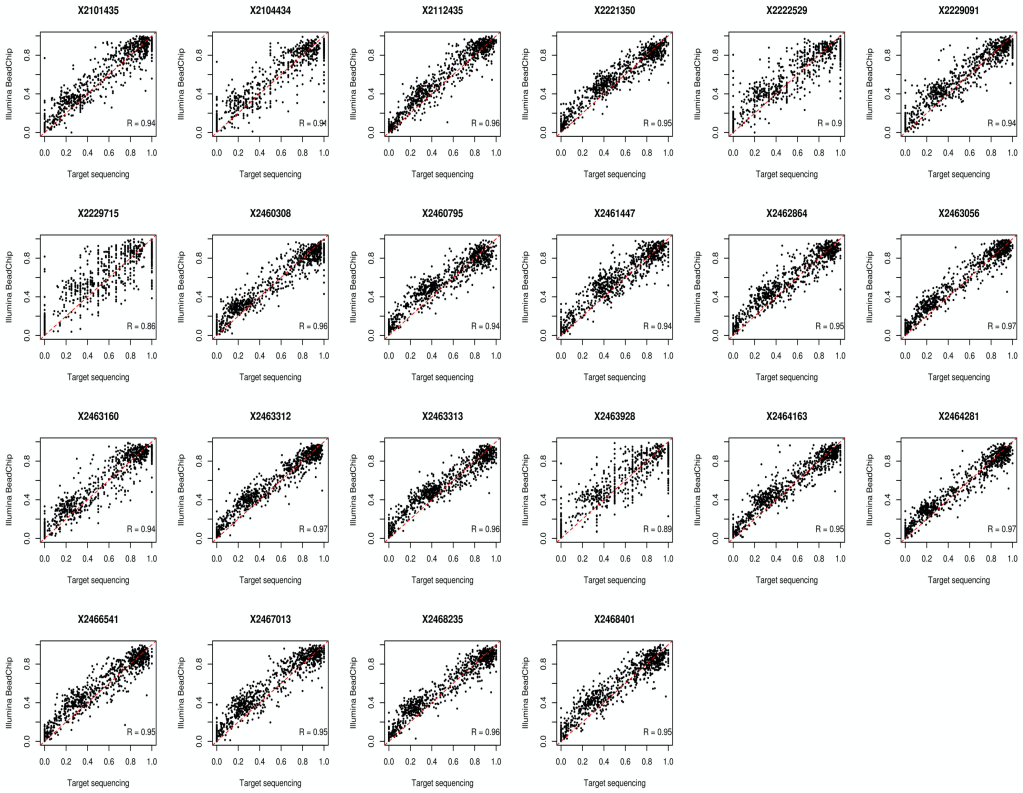


Figure 5.10 Comparison of methylation levels between target bisulfite sequencing and Illumina Infinium 450K BeadChip estimated on 22 samples. In correlation plots are reported on y-axes the methylation levels of Illumina Infinium 450K BeadChip and on x-axes the methylation levels detected in target data.

5.5 Discussion

Analysis of bisulfite sequencing data allows us to observe cytosine methylation at a single-base level. Bisulfite treatment converts unmethylated cytosines in thymines, specialized tools are required to align bisulfite reads over a genome (15). Target bisulfite sequencing is a good compromise to study DNA methylation of particular regions of interest associated with specific phenotype with the elevated levels of accuracy and reproducibility of WGBS at reduced costs.

To analyse bisulfite sequencing data several tools have been published and as many comparison studies have been made to evaluate the tool performances on WGBS and/or reduced representation bisulfite sequencing data (RRBS) (76; 77; 78). However, a comparison of these tool on target bisulfite sequencing data was still lacking.

In computational field software performances are evaluated using simulate and real data. The synthetic dataset allows us to understand where tool fails because true positive values are known. At the moment in literature a simulator for target bisulfite sequencing data that allows both the production of data and information on sequenced cytosines is missing. For this purpose, we developed *MethylFASTQ* a new user-friendly tool to generate bisulfite sequencing data in a fast and high customizable way. We used the synthetic datasets created by *MethylFASTQ* to compare the performance of the most used aligner and methylation caller tools, BSMAP (31) and Bismark (27). These tools were evaluated by the performance in mapping efficiency (*i.e.,* % of reads correctly mapped), recall in methylation detection (*i.e.,* % of cytosines correctly identified) and computational time consuming (*i.e.,* CPU time consumed during the task of alignment).

Datasets with different levels of mutations and sequencing errors were created to stress the tool performances. During the alignment, BSMAP is less sensible to the high level of mutations and sequencing error compared to Bismark that performs wrongly. Even if Bismark shows a very good performance in the recall of methylation. This result is in agreement with previous studies reporting that Bismark is more performant than BSMAP in methylation detection task (77; 78).

Regarding the execution time performances assessed during the alignment on synthetic datasets, BSMAP shows that computational time grows at the increase number of SNPs and sequencing error presence. In contrast, Bismark execution time is shorter for low-quality datasets.

Moreover, we evaluated the tool performances on a real dataset composed of 22 blood samples with 1054 known CpG sites. Tool performances were evaluated by the same parameters used with synthetic data. In this case, both tools perform in a similar manner during mapping and recall steps, through with a slight difference

Bismark seems to be more performant. While in execution time performances BSMAP is better than Bismark in all samples of real datasets.

The differences between these tools highlights in the present study can be most likely ascribed to the diverse approaches used during reads mapping. In fact, for each read BSMAP tests every matching position and keeps the alignment with the lowest number of mismatches. With an increase in the number of SNPs and sequencing errors, the possible number of matching positions increases. As a consequence, the computational time needed to perform the overall alignment increases. While Bismark is based on Bowtie 2, which performs the mismatches identification using the FM-index. In the first time, various short seeds are extracted from the read and they are searched using the index, with a policy that allows few mismatches. Then, low-quality reads are filtered during this phase. Successively, seeds with a lower number of matching positions are selected for the second phase of the alignment.

Moreover, we showed a good correlation between the methylation levels estimated by BSMAP and those previously analysed with Illumina Infinium 450K BeadChip.

Our study underlines that the quality of the reads is an important parameter to consider because may introduce bias during the methylation calling. However, our results suggest that reads with poor quality are discarded during the step of mapping and they do not affect the methylation detection. This represents an important finding from the experimental point of view, since it is not always easy to obtain data of good quality.

A limitation of the present study may be that we performed the recall in methylation detection on a real dataset with a unique depth of coverage. To improve our understanding of tools performance in methylation profiling, we need further studies on CpG sites with different depth of coverage. Indeed, a study on RRBS data suggested that sequencing depth, methylation levels and positions of CpG sites have a significant impact on tool performances. Authors showed that CpG sites in the CpG islands and promoters with higher sequencing depth and lower methylation levels are more likely to be efficiently identified and estimated without any bias compared to CpG sites on CGI shore or gene body (9).

The possibility to have a tool able to correctly detect DNA methylation data with a low coverage may improve the analysis of target bisulfite sequencing. In fact, these data are affected by high duplicate levels when they are sequenced to high coverage. Thus, for bisulfite target data to obtain a good compromise between a good coverage of the genome and a low level of duplicate reads is very difficult; in the majority of cases, it is desirable a low coverage for maintaining low the duplicate levels (1).

Thanks to the present study, we understood how BSMAP and Bismark work and we highlighted their weaknesses in target bisulfite data analysis. We built our pipeline including also all standard steps for duplicate removing and filtering of reads in on-target regions common to all targeted sequencing pipelines. To guarantee the software reproducibility, our pipeline was encapsulated in a Docker container.

6. Smoking DNA methylation in purified monocytes and B cells

This Chapter describes a pilot study to investigate smoking-related DNA methylation signatures in monocytes and B cells.

6.1 Introduction and Aims of the work

DNA methylation is a common epigenetic modification widely studied to understand the genetic mechanisms behind complex diseases and various exposures including smoking. DNA methylation is cell-type specific and it changes among individuals.

DNA methylation is usually measured in whole blood samples, because this represents a biospecimen easy to collect in a minimally-invasive way and analyses can be repeated over the time for the same subject (79). However, blood consists of many functionally and developmentally distinct cell populations whose distribution is variable. Inflammatory status, such as the one derived from smoking exposure, are associated with an abnormally increase in the number of whole blood cells. This cell-number variation might affect the interpretation of methylation results based on whole blood DNA, which is the biospecimen analysed by, the majority of association studies on DNA methylation and smoking exposures. These studies have revealed important smoking-related DNA methylation biomarkers such as the well-known: cg05575921 (***AHRR***), cg03636183 (***F2RL3***) and cg19859270 (***GPR15***).

In these studies, the variation of leukocyte cells in blood was adjusted with the *Houseman* algorithm (23). This algorithm allows the distribution correction for the primary leukocyte subtypes (*i.e.,* B cells, T class, monocyte and granulocyte), but the minor leukocyte fractions are non-included in this method. As demonstrated in our study on the TwinsUK cohort (Chapter 3), these subtypes are

largely affected by smoking exposure. Moreover, this method has been developed for Illumina Bead Chip array technology and it may present some limitations with NGS data, like bisulfite sequencing data.
Therefore, in this study we aimed at understanding if DNA methylation measured in whole blood reflects the smoking effect on leukocyte subtypes and, wheter the smoking-related DNA patterns are similar or specific across the analysed cell-types.

6.2 Materials and methods

6.2.1 Study participants

Six age-matched healthy females (*i.e.,* three current and three never smokers) were enrolled in collaboration with the Association of voluntary Italian blood donors (AVIS) of Turin, following the same criteria and procedures described in the previous study (Chapter 3).
Blood samples and informative questionnaires about lifestyle information, including self-reported smoking habit, were collected for each participant.
The study was conducted according to the guidelines in the Declaration of Helsinki. The protocol of the study was approved by the University of Turin Ethics Committee. All participants signed a written informed consent to participate in the study.

6.2.2 Blood collection

For each subject, 30 mL of peripheral blood in EDTA (Ethylene Diamine Tetra-Acetic acid) vacutainer tubes were collected to obtain the six distinct cell-blood populations. All blood samples were processed within two hours since the collection. Three aliquots of whole blood were taken from one Vacutainer: 1ml for further DNA extraction, 400 µl were added to RNA *later* solution (Thermo Fisher Scientific) for RNA extraction and 2.6 ml were centrifuged at 4°C, 2500

rpm x 10 minutes to obtain the plasma fraction. The remaining 26 ml were used for cell-blood separation.

All aliquots were stored at -80° C until further analyses were performed.

6.2.3 Sample preparation and Cells sorting

Purified blood cell populations were obtained from 26 ml of the remaining fresh blood for each sample. The blood was split into two parts: 20 ml to isolate peripheral blood mononuclear cell (PBMC) consisting of lymphocytes and monocytes, and 6 mL for granulocytes and neutrophils.

Twenty ml of whole blood were diluted in 20 ml of PBS and distributed in sterile Falcon tubes containing 5 mL of Ficoll-Paque Plus™ (GE Healthcare, Sweden). PBMCs were separated by density centrifugation at 2200 rpm x 20' at room temperature. Then, PBMCs were twice washed in PBS, eluted in 10 ml of PBS and counted.

Six ml of whole blood were splitted in six Falcon tubes filled with 1 ml of whole blood and 14 ml of RBC lysing solution 10X.and, then, inverted for 10 minutes at 4°C until liquid became clear red. At this point, the tubes were centrifuged at 250g 10' at 4°C. The lysis procedure was repeated two times. Afterward, the pellet of cells was twice washed with PBS, combined in a unique tube, and counted.

A panel of fluorescent-antibodies was designed to sort six leukocyte subpopulations (*i.e.*, CD4+, CD8+, B-, NK-cells, monocytes and neutrophils). Two diverse mixes of florescent cell-surface marker antibodies were prepared: one for CD4+, CD8+, B- and NK-cells and the second one for monocytes and granulocytes (**Supplementary, Figure S5; Antibodies Sorting Panel**).

In separated FACS Sorting tubes, $20x10^{-6}$ of PBMC and $20x10^{-6}$ of granulocytes were incubated for 10 minutes at room temperature with 3 µl of FcR blocking reagent of MACS (Miltenyi Biotec), then the antibodies staining mix was added, incubated for 20' and twice washed with PBS. The obtained cellular pellet was re-suspended in PBS sorting buffer (0.1% BSA in PBS), filtered with sorting filter (70 µm) and immediately sorted. Cells were sorted using BD FACS Aria III® (BD Biosciences, USA) at low speed, 4-way, 70 µm of the nozzle, and analysed

by BD DiVa Software (version 8.0.2). The purity of all FACS-sorted cell populations was analysed by flow cytometry using BD FACS Aria III® (**Supplementary,Table S3**).

6.2.4 DNA extraction and bisulfite sequencing

Genomic DNA was isolated from the sorted cells using ReliaPrep™ Blood gDNA kit (Promega) according to the manufacturer's instructions. DNA concentration was measured by BR dsDNA Assay kit and Qbit fluorometer (Invitrogen).
500 ng DNA for sample were processed for target bisulfite sequencing using Illumina TruSeq®Methyl Capture EPIC library Prep Kit and paired-end reads of 85 bp were sequenced by the Illumina HiSeq 4000, in collaboration with the Genecore Facility at EMBL, (Heidelberg, Germany).

6.2.5 Bioinformatics analysis

Bisulfite sequencing data were analysed using our pipeline encapsulated in Docker container (the pipeline is described in **Supplementary Material**).
FASTQ files and BED files containing target regions were used as input of pipeline. Briefly, reads quality was checked with FASTQC tool (74) and subsequently, adaptors and low-quality reads were removed with CutAdapt tool (75). Reads were aligned to the human reference genome GRCh38 (*hg38*)(http://www.ncbi.nlm.nih.gov/bioproject/31257), using BSMAP (version 2.90). Before the mapping, pseudo-autosomal regions on Chr Y were masked to N, so that reads may be mapped to equivalent regions on Chr X. In addition, the lambda genome (NC_001416) was added to the reference *hg38* genome to estimate the bisulfite conversion efficiency, as quality control of experimental procedure. Afterwards, mapped reads were filtered to remove duplicates and in on-target regions with Picard and Samtools tools. Moreover, Picard builds several mapping metrics reporting count on-target reads, depth of coverage per base and the estimated insert size distributions. Percentage of methylated cytosines (methylation calling step) and bisulfite conversion efficiency were estimated with *methratio* function implemented in BSMAP tool. Methylation level for each

cytosine is defined by a ratio between the number of reads which the cytosine appears methylated/number of reads that mapped in that position.

Methylation calling files were imported into R statistical software and analysed by *methylKit* package (80). The data were filtered discarding the cytosines with extreme coverage (*i.e.* cytosines covered < 10 reads and with high methylation levels more than 99.9th percentile of coverage in each sample) to eliminate the bases affected by PCR bias. Then, the read coverage distributions between samples were normalized. These two functions will help reduce the bias in the statistical tests that might occur due to systematic over-sampling of reads in certain samples.

The different methylation proportion across the samples was analysed at the level of the genomic regions. Differentially methylated regions (DMRs) were defined by tile the genome in windows of 1000bp length and 1000bp step-size. We sought DMRs with 25% difference between the samples using the logistic regression model, including age and batch effect as adjustment variables in the model. DMRs with adjusted p-value <0.001 were considered to be significant (81). DMRs significantly different from the comparisons (see Results, the analysis approach is summarized in **Figure 6.1**,), were annotated to the genomic regions and gene regions of human reference *hg38* genome using *annotar* R package (82). Functional enrichment analysis of the gene lists cointaining the significantly DMRs was performed using EnrichR (83) and Gorilla (84) web tools.

6.3 Results

6.3.1 Description of samples and cell-population

Three current smokers and three never smokers of similar age (44.6 ±3.2 years, range: 40-48) were included in the study. Current smokers declared to smoke 4-5 cigarettes/day and were considered as light smokers. For each individual, six cell populations (B cells, CD4+ and CD8+ T cells, NK cells, monocytes and neutrophils) were sorted. The purities of sorted cells based on cell surface markers

ranged from 96.22% for NK cells o a 99% for monocytes, neutrophils, CD4+ and CD8+ T cells in all samples (**Supplementary Table S3).**

6.3.2 Results of target bisulfite sequencing

In this study, we investigated the smoking-related difference in the DNA methylation levels in monocytes, and B cells.
For each subject, the DNA extracted from purified monocytes and B cells was sequenced. For one of the never smoker subjects (Non-smoker 3), only DNA from monocytes was sequenced because the amount of DNA extracted from B cells was very low (less than 400 ng, which is the minimum amount of DNA required for sequencing by the TruSeq®Methyl Capture EPIC library Perp Kit).
In **Tables 6.1** and **6.2** the individual outcomes from target bisulfite sequencing of monocytes and B cells are reported. The samples were evaluated for: *i)* the number of sequenced reads; *ii)* the percentage of on-target reads; *iii)* the depth of coverage; and *iiii)* the number of CpG sites covered.
The number of total input reads among the samples was similar, with an average of 57,407,725 and 56,825,863 reads for monocytes and B cells, respectively. The lowest number of total input reads was 34,284,430 and 30 x10^6 for monocytes and B cells respectively, whereas the highest was 76,312,293 and 75,125,805 in monocytes and B cells, respectively.
Similarly, the percentage of aligned reads (*i.e.,* the reads uniquely mapped in paired-end) was 86.69% in monocytes and 86.03% in B cells. The percentage of filtered reads (*i.e.,* reads mapped with correct orientation and the distance consistent with the library insert size) was higher in monocytes than B cells (80.32 vs 77.82, for monocytes and B cells, respectively).
A crucial parameter to evaluate in the experiments of target sequencing is the percentage of on-target reads. This represents the number of mapped reads overlapping a set of target regions by at least one base. Our samples showed a high percentage of on-target reads (range: 85-92% for both cell-lineages), where the maximum value was 92.18% of Smoker 3 sample in monocytes and the average value was 87% in both cell-lineages. Median of depth of coverage varied among the samples, the maximum value was 26x (fold of coverage) observed in

monocytes of Non-smoker 2 sample and in B cells of Nonsmoker 1 sample. The minimum value of 12x was observed in B cells of sample Non-smokers 2 sample, in line with the lower number of total input reads reported in the table. All samples showed a 99.9% of bisulfite conversion efficiency, suggesting a total conversion of unmethylated citosines to thymines.

The average of the number of CpG sites detected in both cell-lineages were roughly the same, 8,075,285 for monocytes and 8,010,382 for B cells. Within each cell-lineage, we found that smokers showed a lower average of CpG sites in monocytes compared to never smokers (7,690,306 and 8,460,264, for current and never smokers, respectively). In contrast, in B cells we observed the opposite situation, with a higher average number of CpG sites in smokers than in never smokers (8,125,033 and 7,838,407 for smokers and never smokers, respectively).

Table 6.1 Summary statistics for target bisulfite sequencing of monocytes.

Samples	Total input raw reads	Reads aligners pair	% alignment	Total reads after filtering	% input reads after filtering	Reads on-target (primary)	% reads on-target (primary)	Median target coverage	Lambda conversion efficiency	Number of CpG sites
Smoker 1	55230090	46804376	84,74	40094212	72,6	34262730	85,46	18	99.9%	7849197
Smoker 2	65135564	55261125	84,84	46992590	72,15	40172906	85,49	21	99.9%	8319438
Smoker 3	34284430	30970684	90,33	28905800	93,33	26646149	92,18	14	99.9%	6902284
Non-smk 1	66115257	56165953	84,95	50052320	75,71	42638463	85,19	23	99.9%	8574653
Non-smk 2	76312293	64671192	84,74	57063790	74,78	48778167	85,48	26	99.9%	8946090
Non-smk 3	47368713	42874336	90,51	40018807	93,33	36672289	91,64	19	99.9%	7860049
min	34284430	30970684	84,74	28905800	72,15	26646149	85,19	14	99.9%	6902284
max	76312293	64671192	90,51	57063790	93,33	48778167	92,18	26	99.9%	8946090
average	57407724	49457944	86,69	43854586	80,32	38195117	87,57	20,17	99.9%	8075285

Table 6.2 Summary statistics for target bisulfite sequencing of B cells.

Samples	Total input raw reads	Reads aligners pair	% alignment	Total reads after filtering	% input reads after filtering	Reads on-target (primary)	% reads on-target (primary)	Median target coverage	Lambda conversion efficiency	Number of CpG sites
Smoker 1	63853274	54156666	84,81	48139158	75,39	41031746	85,24	22	99.9%	8431270
Smoker 2	64546796	54758329	84,83	46912342	72,68	40041970	85,36	21	99.9%	8350795
Smoker 3	51322717	43404977	84,57	37083578	72,26	31781405	85,7	17	99.9%	7593033
Non-smk 1	75125805	63700151	84,79	56972360	75,84	48108117	84,44	26	99.9%	9102037
Non-smk 2	29280721	26684918	91,13	24794158	92,91	22838685	92,11	12	99.9%	6574777
min	29280721	26684918	84,57	24794158	72,26	22838685	84,44	12	99.9%	6574777
max	75125805	63700151	91,13	56972360	92,91	48108117	92,11	26	99.9%	9102037
average	56825863	48541008	86,03	42780319	77,82	36760385	86,57	19,60	99.9%	8010382

6.3.3 Smoking-related DNA methylation signatures: in monocytes and B cells

To investigate the smoking-related DNA methylation signatures in monocytes and B cells, we compared current *vs* never smokers within of each cell-lineage, separately (**Figure 6.1).**

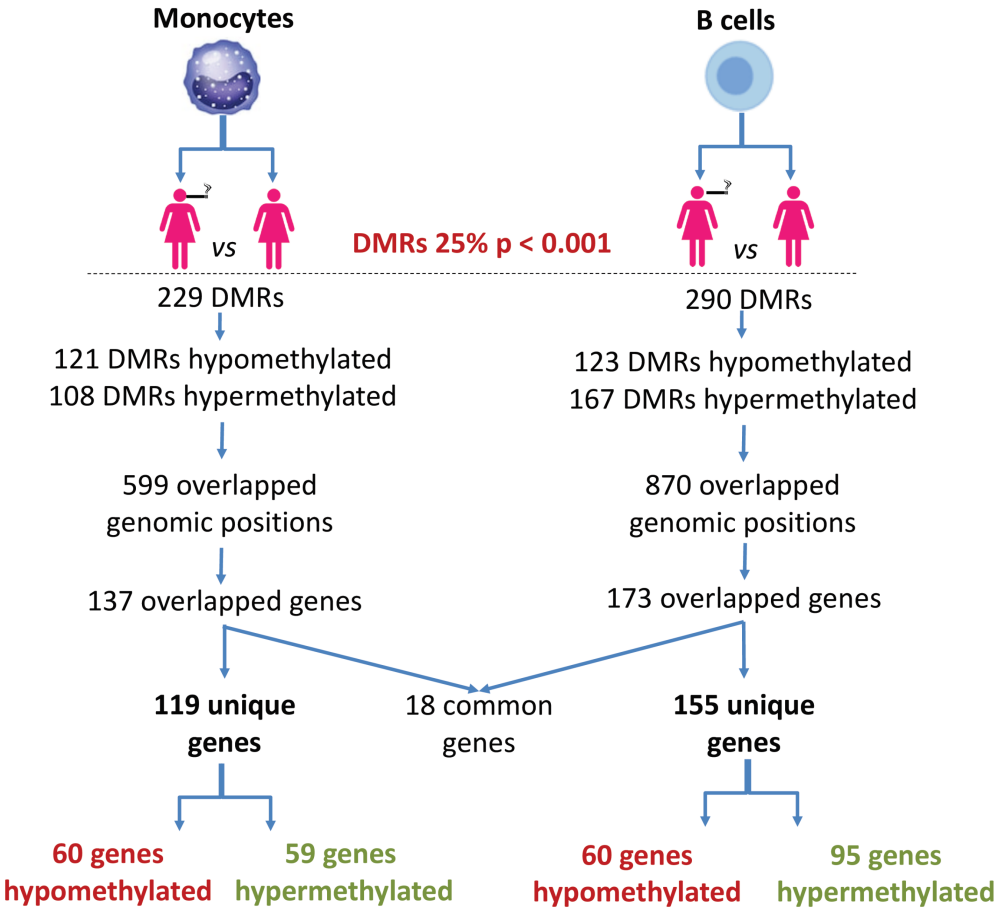


Figure 6.1. Analysis approach and number of differentially methylated genes of monocytes and B cells involved in smoking exposure. The Figure reports the number of DMRs resulted from the comparison between current *vs* never smokers in monocyte and B cells, separately. DMRs were annotated on the genome and the resulted genes from these two comparisons were merged in order to identify the differentially methylated genes in common between the cell-lineages and exclusively present in monocytes and B cells.

We analysed the difference in DNA methylation levels comparing smokers vs never smokers in monocytes and B cells. In monocytes, we found 229 DMRs at adjusted p-value < 0.001 and methylation difference more than 25%: 121 were

hypomethylated (*i.e.,* a lower methylation level in smokers vs never) and 108 were hypermethylated (*i.e.,* a higher methylation level in smokers vs never) in smokers compared to never smokers. We conducted the same comparison in B cells, and we observed 290 DMRs, of which 123 were hypomethylated and 167 were hypermethylated in smokers vs never.

To gain insight on these differences, we annotated the DMRs according to genomic and gene localizations. In the genomic positions were included the CpG islands, shores (*i.e.,* regions up to 2 kb from CpG island), shelves (*i.e.,* regions from 2 to 4 kb from CpG island), and open sea (*i.e.,* the rest of the genome). Moreover, we annotated DMRs also to gene localization including 1-5Kb upstream of the TSS, the promoter (< 1Kb upstream of the TSS), 5'UTR, first exons, exons, introns, 3'UTR, and intergenic regions (the intergenic regions exclude the previous list of annotations). The genome localization and CpGs distribution of significantly different DMRs in monocytes and B cells is reported in **Figure 6.2**. We reported that hypomethylated and hypermethylated regions in both cell-lineages follow a similar distribution, with a high percentage (~39%) of DMRs located in introns (gene bodies) and a low percentage (~3%) of DMRs located in 3' UTR (gene non-coding region) regions. Similarly, the distribution of DMRs in the genomic regions showed a high percentage (>70%) of DMRs located in inter-CGI or open-sea regions. In contrast, a lower percentage (<6%) of DMRs was observed to be placed in CpG islands.

We further explored the gene-containing DMRs in both the cell-lineages to identify those smoking-related genes that were in common or specific between monocytes- and B cells. We found 18 smoking-methylated genes in common between monocytes and B cells, and, 154 and 119 genes exclusively methylated in monocytes and B cells, respectively. The unique genes of monocytes were 60 hypomethylated and 59 hypermethylated, whereas in B cells were divided in 60 hypomethylated and 95 hypermethylated comparing smokers vs never smokers.

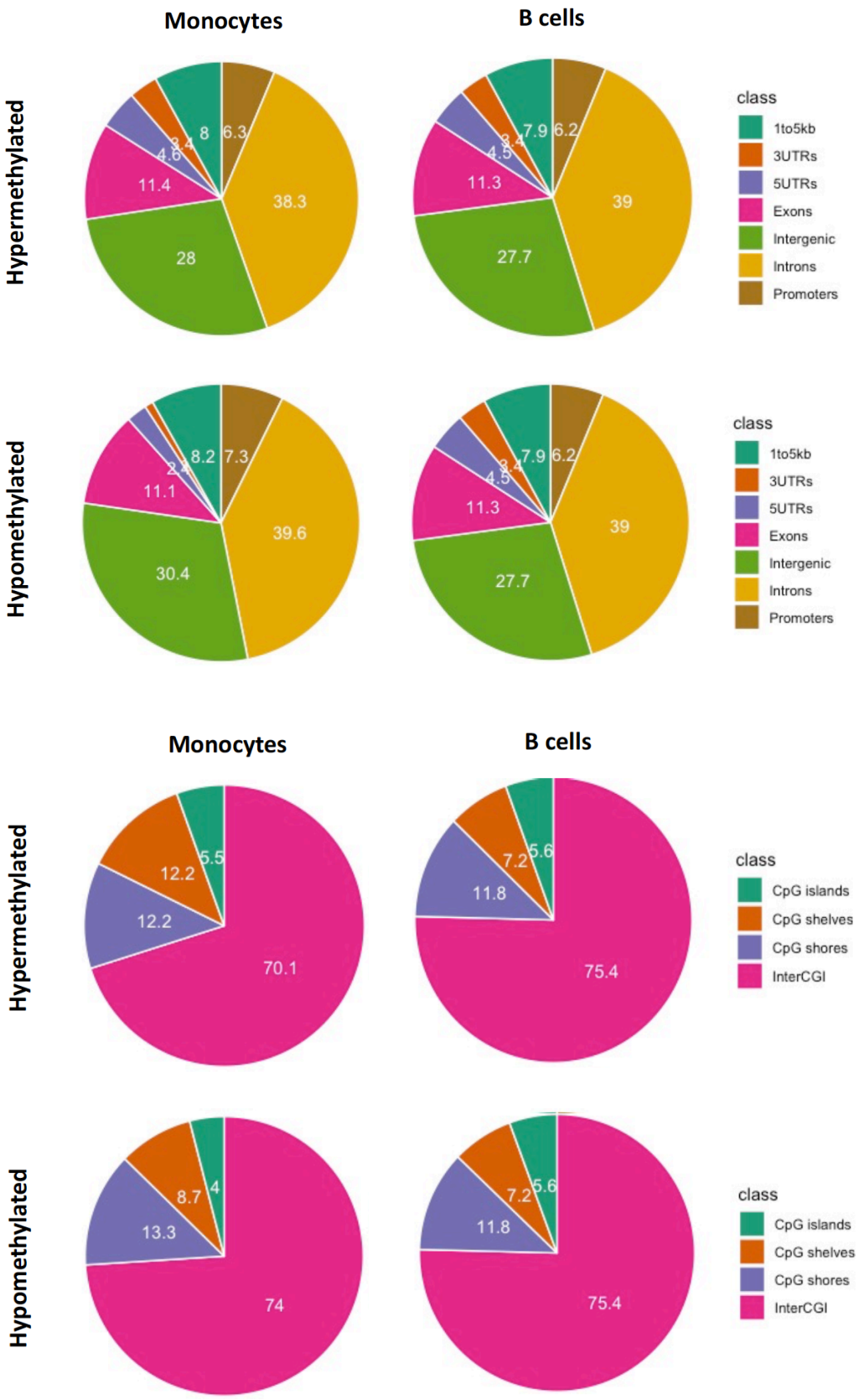


Figure 6.2. Annotation of DMRs of monocytes and B cells. Pie charts show the percentage of DMRs overlapping gene and genomic regions.

6.3.4. Pathways and functional enrichment analysis of differential methylated genes of monocytes and B cells.

We performed a pathways enrichment analysis to explore the biological function and the molecular pathways in which were involved the identified differentially methylated genes. The analysis was conducted on gene sets showing a hypomethylation and hypermethylation only in monocytes and B cells, and on all common genes between the cell lineages. We compared the gene lists with human KEGG, Reactome, Wiki Pathways databases using Enrichr (**Supplementary, Table S4**)

We observed that the hypomethylated genes of B cells were enriched in the lysosome pathway (adj.p-value < 0.05) which also includes the *LAPTM4B* gene involved in several cancers (85). We also reported the top ten pathways of each tested database that were not statistically significant (adj.p-value> 0.05) but with nominal p-value <0.05. Among these, we observed in hypomethylated gene set of B cells, *GALC* and *GLB1* genes overlapped sphingolipid metabolism pathways in all databases. *GALC* gene is associated with pulmonary artery enlargement in COPD (86) and DNA hypomethylation of *GLB1* has been associated with obesity (87). In contrast, the list of hypermethylated genes in B cells was not enriched in any particular pathways (adj.p-value=1) and, it also showed inconsistent results among the three databases. Indeed, we observed in Wiki pathways *KCNU1*, *KCNMB2* and *PDE5A* genes overlapped cGMP-PKG signalling pathway; in KEGG *SETBP1*and *SMYD3* genes overlapped histone modifications pathway and in Reactome database *SEMA5A*, *ADAMTS17*, and *THSD7A* overlapped pathways associated with glycosylation.

Hypomethylated genes of monocytes showed consistent results among the top terms from the different databases. At nominal p-value <0.05, we observed *RHEB*, *ADCY3*, *PRDM16*, *PPARG* genes involved in thermogenesis and *SLC22A4P*, *PDGFRA, RHEB* in choline metabolism in cancer. Some of these genes (*PPARG*, *ADCY3* and *RHEB*) also overlapped pathways involved in ageing. Hypermethylated genes of monocytes showed inconsistent results among the databases, even if we observed the recursive presence of *ITGA1* gene overlapping

miRNA targets in ECM and membrane receptors, regulation of actin cytoskeleton, *NOTCH1* regulation of human endothelial cell calcification and L1CAM interaction pathways. This gene coding integrin α1β1 has been reported to be involved in various cancers, including colorectal cancer (88).

For the list of genes shared between monocytes and B cells, we did not find any strongly significant enrichment. The results showed pathways involved in cellular signalling and interleukins production.

To examine in depth the biological function of gene sets, we also performed the gene ontology enrichment analysis with the GOrilla tool. All submitted gene lists did not show enriched terms. We also tried to submit the entire list (*i.e.*, all hypomethylated and hypermethylated genes together) of B-cell and monocytes specific genes and, we found in B cells an enrichment in GO:0051271 term associated with negative regulation of cellular component movement (**Supplementary Table S4**).

6.4 Discussion

In the present study, we sequenced monocytes and B cells from six healthy females, three current smoker and three never smokers, for a total of 11 samples, of which six were from monocytes, and five to B cells.

Overall, the sequencing results showed a difference in the number of total reads in input among samples, but that seems not to have a great impact on the percentage of on-target reads and the number of CpG sites covered by the sample. Basically, all samples displayed a good quality such as an elevate percentage of on-target reads (> 80%), bisulfite conversion efficiency (99%), and the number of CpG sites covered. However, depth of coverage was low in all samples around 20x, while for target bisulfite sequencing is recommended a deeper coverage, around 30x.

Next, we identified the DMRs in monocytes and B cells, separately, comparing smokers vs never smokers individuals. We observed a more elevated number of DMRs in B cells compared to monocytes (290 vs 229) suggesting a different degree of methylation, which might be due to their different cell-lineage and to a

different role in the immune system. In line with that, we also observed a difference in the methylation levels of significant DMRs. B cells reported, in fact, a high number of hypermethylated and a low number of hypomethylated regions, conversely, monocytes reported the opposite proportions. The genomic and gene localization of both hypomethylated and hypermethylated DMRs were similar in both cell lineages, with a high proportion of DMRs located in introns and intergenic regions. This result is consistent with what reported in previous studies: CpG sites methylated were located in introns, 3'UTRs and intergenic regions. In contrast, CpG sites unmethylated tend to be in CpG islands and 5'UTR (89).

We also identified a group of smoking-related genes in common between the cell-lineages and uniquely methylated in monocytes and not in B cells and vice versa. As expected, we observed only 18 genes shared between monocytes and B cells, involved in cellular function but not enriched in any particular biological pathways. This result supports a clear difference in the methylation patterns between lymphocytes (B cells) and the myeloid cells (monocytes), as demonstrated in previous studies (89; 90). And it is imputable to the difference in the cellular lineage. Among the B cells, hypomethylated genes showed enrichment in the lysosome pathway, including *LAPTM4B* gene. This gene is associated with overexpression and poor prognosis in various malignancies including breast cancer, bladder cancer, ovarian cancer, HCC, gastric cancer and cervical cancer. Besides, its over-expression has also been identified in non-small cell lung cancer (85). In contrast, the gene sets only methylated in monocytes did not show any significant enrichment, despite the presence of genes involved in choline metabolism in cancer. These findings are interesting, and we can speculate on a possible link between smoking-related methylation signatures and cancer implications.

However, we are aware that this pilot study has several limitations, including limited sample size and low statistical power. We chose to test the smoking-related DNA methylation differences in monocytes and B cells for their different cell lineage (myeloid and lymphoid, respectively), and we expected to find an elevated DNA methylation difference in smoking response. On the other hand, in peripheral blood, these cells are < 10% of total white blood cells in circulation,

and they could be contributing marginally to methylation value observed in the blood of known smoking-related genes such as *AHRR*, *F2RL3* and *GPR15*. In contrast, in T cells, which are about 20-30% of the total lymphocytes in circulation, we could expect to see the same difference observed in whole blood. Moreover, our studies of Chapter 3 and 4 and other studies in literature (24;36; 38) on the leukocyte-shift caused by active smoking, have shown that the T cells are the main cell-population affected by smoking.

In conclusion, this pilot study reports a significative difference in smoking-related DNA methylation between monocyte and B cells using also light smoker samples. The resulted genes did not show any particular enrichmentin biological pathways. To confirm these results, further studies in a more variegated cohorts with larger sample size and with a high prevalence of heavy smokers are needed. However, these preliminary findings suggest that to study purified monocytes and B cells instead of whole blood samples is foundamental to understand if the smoking-related signatures are at least in part the same.

Conclusions

The present Ph.D work was focused on the analyses of smoking effects on leukocyte subpopulations to explore specific DNA methylation changes caused by this environmental exposure in blood samples. In this respect, two aspects were evaluated: first, the blood cells variation caused by smoking and, subsequently, the differences in DNA methylation levels in each cell-types due to smoking exposure. Both of these aspects were investigated following three approaches: experimental, epidemiological and computational.

Smoking effects on the main leukocyte subpopulations were studied in 288 healthy volunteers covered all smoking categories (*i.e.,* 88 currents, 99 former and 100 never smokers) recruited in collaboration with the Association of voluntary Italian blood donors (AVIS) of Turin. As reported above, this cohort was mainly characterized by light smokers (<15 cigarette/day). For each subject the primary leukocyte subpopulations were evaluated, including their expression of GPR15 receptor as a smoking marker, by flow cytometry. Results showed a significant decrease of NK cells, and the increase expression of GPR15+ in CD4, CD8 and B cells in current smokers compared to never smokers. These results are in agreement with those reported in the literature about heavy smokers, but our study underlines that smoking affects significantly the leukocyte distributions also at low dose of exposition, like in light smokers.

A similar investigation was conducted taking advantage of the available data in the TwinsUK cohort. This cohort is composed of healthy females characterized by around 42,000 immune cell traits measured by high-resolution deep immunophenotyping. Out of the whole cohort, we selected 358 individuals for the complete information of their self-reported smoking habit and without any reported diagnosis of autoimmune disease. In this study, we found a statistically significant increase of several CD8 T cell-subpopulations and class-switched memory B cells isotype IgA, IgG and IgE in smokers vs never smokers. After

smoking cessation the relative proportions of the majority of these cells return to never smokers levels, whereas activated CD8 T cells and CCR4 + CD8 Memory T cells persist partially altered in former smokers.

Results obtained from TwinsUK-study suggest that smoking not only globally changes the distribution of the major leukocyte subpopulations, but also it affects cells less frequent in circulation, such as CD8 T cell-subtypes, DP T cells and class-switched memory B cells. Moreover, high light that active smoking affects mainly CD8 T cells that skew towards a chronic inflammatory phenotype.

The second aspect of the present thesis concerned an argument widely debated in the molecular epidemiological studies: does the DNA methylation levels measured in whole blood reflect the real association with a particular phenotype such as smoking exposures? Or the obtained results are confounded by the variation of specific cell-types present in whole blood? To study this issue, we decided to study the smoking DNA methylation profiles in each cell-types belonging to primary leukocytes. Thanks to decreasing costs deep sequencing techniques, for this purpose we used targeted bisulfite sequencing. This technique represents a valid compromise between the elevated precision and reproducibility of the whole genome sequencing and the possibility to study large study populations.

First of all, we realized that currently in the literature, a benchmark pipeline to analyze this type of data is still lacking. Therefore, we compared the most used tools to analyze bisulfite data. Since the software performance is evaluated on a synthetic dataset, and a simulator specific for targeted bisulfite was missing, we developed MethylFASTQ, a simulator of synthetic bisulfite sequencing data. We tested BSMAP and Bismark tool on synthetic and real datasets showing BSMAP more performant during alignment and methylation recall in datasets with low-quality reads.

We used our developed pipeline to investigate the smoking-related DNA methylation signatures in monocytes and B cells, in a pilot study. Comparing the DNA methylation levels of smokers and never smokers in each cell-lineage, we found several differentially methylated regions overlapping annotated genes. A low number of these genes were shared between monocytes and B cells, whereas

an elevated number of these genes were cell-type specifics. The common and cell-specific gene sets were involved in several biological pathways, including genes overlapped with pathways of cancer, without showing a particular enrichment in one of those. Further investigations are needed to validate these smoking-related DNA methylation signatures.

In conclusion, the present thesis showed that active smoking affects significantly the leukocyte composition, including the fraction less frequent in blood. This is an important finding since in epidemiological studies the cell-blood composition is usually corrected by the Houseman algorithm. As mentioned, this algorithm presents some limitations that must be taken into account. So, in the near future, we need to test the existing methods of cell-type deconvolution for understating their limits or to develop new computational approaches for dissecting cell-types composition from whole blood without sorting cells, considering both the variation of minor blood-cells and the complexity of the bisulfite sequencing data.

References

1. Ritchie, H. and Roser, M. Smoking. *OurWorldInData.org.* [Online] 2020. https://ourworldindata.org/smoking'.
2. Organization, World Health. WHO global report on trends in prevalence of tobacco use 2000-2015, thrid edition. 2019.
3. Organization, World Health. WHO European Tobacco Use, Trend report 2019.
4. Pacifici, R. *et al*., Indagine ISS-DOXA 2019. Tobacco smoking in Italy.s.l. : Tabaccologia , 2019, Vol. 3.
5. Filippidis, F. T. *et al.* Two-year trends and predictors of e-cigarette use in 27 European Union member states. 26, *Tob Control*, 2017.
6. Genera, Surgeon. *The health Consequences of Smoking-50 years of progress.* 2014.
7. Strzelak A., *et al.* Tobacco Smoke Induces and Alters Immune Responses in the Lung Triggering Inflammation, Allergy, Asthma and Other Lung Diseases: A Mechanistic Review. *Int. J. Environ. Res. Public Health*, 2018, Vol. 13.
8. Goncalves, R.B. *et al.* Impact of smoking on inflammation: overview of molecular mechanisms. *Inflammation research*, 2011.
9. Qiu, F. *et al*. Impacts of cigarette smoking on immune responsiveness: Up and down or upside down? *Oncotarget*, 2017, Vol. 8.
10. Stampfli, M. R. and Anderson, G.P. How cigarette smoke skews immune responses to promote infection, lung disease and cancer. *Nat. Rev. Immunology*, 2009, Vol. 9, pp. 377-382.
11. Arnson Y, *et al***.** Effects of tobacco smoke on immunity, inflammation and autoimmunity. Journal of autoimmunity, 2010, Vol. 34, pp. 258-265.
12. Jones, P.A., Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Review Genetics*, 2012, Vol. 13, pp. 484-492.
13. Bird, A., DNA methylation patterns and epigenetic memory. *Genes Development*, 2002, Vol. 16, pp. 6-21.
14. Bock, C. *et al.,* Quantitative comparison of DNA methylation assays for biomarker development and clinical applicantions. *Nature Biotechnology*, 2016, Vol. 34, pp. 726-737.
15. Bock, C. *et al.* Analysing and interpreting DNA methylation data.*Nature Genetics Reviews*, 2012, Vol. 13.
16. Shenker, N. S. *et al.* Epigenome-wide association study in the Epuropean Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking.*Human Molecular Genetics* , 2013, Vol. 22.

17. Ambatipudi, S. *et al.,* Tobacco smoking-associated genome-wide DNA methylation change in the EPIC study. *Epigenomics,* 2016 , Vol. 8.
18. Guida, F. *et al.,* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation.*Hum Mol Genet*, 2015, Vol. 24, pp. 2349-2359.
19. Fasanelli F. *et al.,* Hypomethylation of smoking-related genes is associated with future lung cancer in fuor prospective cohort.*Nature communication*, 2015, Vol. 6.
20. Koks, Gea *et al*., Smoking-Induced Expression of the GPR15 Gene Indicates Its Potential Role in Chronic In fl ammatory Pathologies. *Am. J. Pathol*., 2015, Vol. 185.
21. Baglietto, L., *et al.,* DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International Journal of Cancer,* 2017, Vol. 140, pp. 50-61.
22. Stueve, T. R. *et al.,* Epigenome-wide analysis DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Human Molecular Genetics*, 2017, Vol. 26.
23. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics,* 2012, Vol. 13.
24. Bauer, M. *et al.,* A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. *Clin Epigenetics,* 2015, Vol. 7.
25. Laird P.W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 2010, Vol. 11, pp. 191-203.
26. Sun, Z. *et al.,* Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics*, 2015, Vol. 7.
27. Krueger,F *et al.,* Bismark: a flexible aligner and methylation caller for Bisulfite-Seq apllications. *Bioinformatics*, 2011, Vol. 11.
28. Li,H., and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009, Vol. 25.
29. Langmead, B. *et al*., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 2009, Vol 10
30. Chen, P.Y. *et al BS Seeker: precise mapping for bisul te sequencing*. *BMC Bioinformatics*, 2010, Vol. 11.
31. Xi, Y. and Li, W. *BSMAP: whole genome bisul te sequence MAPping program. BMC Bioinformatics*, 2009, Vol. 10.
32. Harris, E. Y., *et al.* BRAT: bisul te-treated reads analysis tool. *Bioinformatics*, 2009, Vol. 24.
33. Sopori, M. Effects of cigarette smoke on the immune system**.** *Nat Rev Immunol,* 2002, Vol. 2, pp. 372-377.

34. Shipa, Shahena Aktar *et al,* Effect of intensity of cigarette smoking on leukocytes among adult men and women smokers in Bangladesh**.** *Asia Pacific Journal of Medical Toxicology*, 2017, Vol. 6, pp. 12-17.
35. Malenica *et al.* Effect of Cigarette Smoking on Haematological Parameters in Healthy Population*.* 132-136, *Medical archives*, 2017, Vol. 71.
36. Tollerud. *Et al.,* The effects of cigarette smoking on T cell subsets. A population-based survey of healthy caucasians*. The American review of respiratory disease*, 1989.
37. Higuchi T. *et al.,* Current cigarette smoking is a reversible cause of elevated white blood cell count: Cross-sectional and longitudinal studies.417, *Preventive medicine reports*, 2016, Vol. 4.
38. Tollerud, *et al*., Association of Cigarette Smoking with Decreased Numbers of Circulating Natural Killer Cells. *Am Rev Respir Dis*., 1989, Vol. 139, pp. 194-198.
39. Freedman, D.S. *et al.,* Cigarette smoking and leukocyte subpopulations in men. *Annals of epidemiology, 1996.*
40. Leek, Jeffrey T, *et al.,* sva:Surrogate Variable Analysis. R package version 3.30.1, 2019.
41. Gallus, S. *et al*., Effect of Tobacco Smoking Cessation on C-Reactive Protein Levels in A Cohort of Low-Dose Computed Tomography Screening Participants. *Scientific Report*, 2018, Vol. 8.
42. Verdi, S. *et al.,* TwinsUK: The UK Adult Twin Registry Update*.Twins Research and Human Genetics*, 2019.
43. Andrew,T., *et al*., Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women*. Twin Research*, 2001, Vol. 4.
44. Roeaderer M., *et al.,* The Genetic Architecture of the Human Immune System: A bioresource of Autoimmunity and Disease Pathogenesis*. Cell*, 2015, Vol. 161.
45. Mangino,M., *et al.* Innate and adaptive immune traits are differentially affected by genetic and environmental factors*. Nature Communications*, 2017, Vol. 8.
46. Li, J., and Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 2005, Vol. 95.
47. Chen, G., *et al.,* Cigarette Smoke Disturbs the Survival of CD8+ Tc/Tregs Partially through Muscarinic Receptors-Dependent Mechanisms in Chronic Obstructive Pulmonary Disease*. PlosOne,* 2016.
48. Saetta, M., *et al.,* Activated T-lymphocytes and macrophages in bronchial mucosa of subjects with chronic bronchitis*. The American review of respiratory disease*, 1993.
49. Ginns, *et al.,* Elevated concentration of soluble interleukin-2 receptors in serum of smokers and patients with lung cancer. Correlation with clinical activity. *The American review of respiratory disease*, 1990.

50. Churlaud, G. *et al.,* Human and Mouse CD8(+)CD25(+)FOXP3(+) Regulatory T Cells at Steady State and during Interleukin-2 Therapy. *Frontiers in immunology*, 2015, Vol. 6 .
51. Vieyra-Lobato, M.R. *et al.,* Description of CD8(+) Regulatory T Lymphocytes and Their Specific Intervention in Graft-versus-Host and Infectious Diseases, Autoimmunity, and Cancer. *Journal of immunology research* , 2018.
52. Caruso, A. e*t al.,* Flow cytometric analysis of activation markers on stimulated T cells and their correlation with cell proliferation. *Cytometry*, 1997, Vol. 27.
53. Nunes-Cabaco, *et al.,* Differentiation of human thymic regulatory T cells at the double positive stage. *European journal of* immunology , 2011, Vol. 41.
54. Ritter, M. *et al.,* Elevated expression of TARC (CCL17) and MDC (CCL22) in models of cigarette smoke-induced pulmonary inflammation. *Biochemical and biophysical research communications*, 2005, Vol. 334.
55. Stolberg, V.R. *et al.,* Role of CC chemokine receptor 4 in natural killer cell activation during acute cigarette smoke exposure. *The American journal of pathology*, 2014, Vol. 184.
56. Fergusson, J.R. *et al.,* CD161-expressing human T cells. *Frontiers in Immunology,* 2011.
57. Brandsma, C. A., *et al.,* Increased level of (class switched) memory B cells in peripheral blood of current smokers. *Respir. Res*, 2009, Vol. 10.
58. Brandsma, C. A., *et al.,* Differential switching to IgG and IgA in active smoking COPD patients and healthy controls. *Eur. Resp. J,* 2012, Vol. 40.
59. Jackson D.A. & Elsawa S.I., Factors Regulating Immunoglobulin Production by Normal and Disease-Associated Plasma Cells. *Biomolecules*, 2015.
60. Scalzo-Inguanti and Plebanski, CD38 identifies a hypo-proliferative 3-secreting CD4+ T-cell subset that does not fit into existing naive and memory phenotype paradigms. *European journal of immunology* , 2011.
61. Kalia, V. & Sarkar, S. Regulation of Effector and Memory CD8 T Cell Differentiation by IL-2. A Balancing Act.*Frontiers in Immunology*, 2018.
62. Valiathan, R. *et al.,*. Tobacco smoking increases immune activation and impairs T-cell function in HIV infected patients on antiretrovirals: a cross-sectional pilot study. *Plos One*, 2014, Vol. 9.
63. Hong, M. J., *et al.,* Protective role of gammadelta T cells in cigarette smoke and influenza infection. *Mucosal immunology* , 2018, Vol. 11.
64. Elfrieke D Van Tiel, *et al.,*. Quitting Smoking May Restore Hematological Characteristics within Five Years. *Annal of epidemiology*, 2002.
65. Ebbert, J.O. *et al.,*. Lung Cancer Risk Reduction After Smoking Cessation: Observations From a Prospective Cohort of Women. *Journal of Clinical Oncology*, 2003.
66. Krueger,F., Sherman - bisulfite treated Read FastQ Simulator. https://www.bioinformatics.babraham.ac.uk/projects/sherman/.

67. Huang, *et al.,* ART: a next-generation sequencing read simulator. *Bioinformatics*, 2011, Vol. 28.
68. Caboche, *et al.,* Comparison of mapping algorithms used in highthroughput sequencing: application to Ion Torrent data. *BMC Genomics*, 2014, Vol. 15.
69. Langmead, B., *et al. Genome Biology, 2009, Vol. 10.*
70. Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods, 2012, Vol. 9.*
71. P. Ferragina and G. Manzini. Opportunistic data structures with applications. *Proceedings of the 41st Annual Symposium on Foundations of Computer Science., 2000.*
72. Piaggeschi,G.and Licheri,N., *et al.,* MethylFASTQ: a tool simulating bisulfite sequencing data.
73. Palli, D. *et al.,* A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *89, Tumori, 2003, Vol. 6.*
74. Andrews, FASTQC. A quality control tool for high throughput sequence data. *2014*
75. Martin *et al. https://doi.org/10.14806/ej.17.1.200.* 2011.
76. Chatterjee, *et al.,* Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research,* 2012, Vol. 40.
77. Sun, *et al.,* A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinformatics,* 2018, Vol. 34.
78. Tsuji, Junko and Weng, Zhiping. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Briefings in Bioinformatics,* 2016, Vol. 17.
79. Terry, B. M. *et al.,* DNA methylation in white blood cells: association with risk factors in epidemiologic studies. *Epigenetics,* 2011, Vol. 6.
80. Akalin, A. *et al., methylKit:* a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology, 2012.*
81. Storey, J. D, and Tibshirani, R. Statistical significance for genomewide studies.. *PNAS,* 2003.
82. Cavalcante, R.G. and Sartor, M.A *annotatr:* genomic regions in context. *Bioinformatics, 2017.*
83. Kuleshov, M. *et al., Enrichr:* a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research,* 2016.
84. Eden, E. *et al., GOrilla:* a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics, 2009.*
85. Wang, L., *et al.,.* LAPTM4B is a novel diagnostic and prognostic marker for lung adenocarcinoma and associated with mutant EGFR. *BMC Cancer,* 2019
86. Lee, J.H., *et al.,* IREB2 and GALC are associated with pulmonary artery enlargement in chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol,* 2015.

87. Benton, M.C. *et al.,.* An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss.*Genome Biology,* 2015.
88. Boudjaid, S., *et al.,.* Integrin α1β1 expression is controlled by c-MYC in colorectal cancer cells. *Oncogene*, 2016.
89.Reinius, *et al.,* Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility.*Plos One,* 2012, Vol. 7.
90. Hachiya, T., *et al.,.* Genome-wide identification of inter-individually variable DNA methyation sites improves the efficacy of epigenetic association studies. *npj Genomic Medicine,* 2017, Vol. 2.

Publication

MethylFASTQ: a tool simulating bisulfite sequencing data

Giulia Piaggeschi*
Dept. of Computer Science
University of Turin
Turin, Italy
giulia.piaggeschi@unito.it

Nicola Licheri*
Dept. of Computer Science
University of Turin
Turin, Italy
licheri@di.unito.it

Greta Romano
Dept. of Computer Science
University of Turin
Turin, Italy
grromano@unito.it

Simone Pernice
Dept. of Computer Science
University of Turin
Turin, Italy
pernice@di.unito.it

Laura Follia
Dept. of Computer Science
University of Turin
Turin, Italy
laura.follia@unito.it

Giulio Ferrero
Dept. of Computer Science
University of Turin
Turin, Italy
giulio.ferrero@unito.it

***Abstract*—DNA methylation is a DNA modification playing an important role in several diseases, including cancer. The gold-standard technique for measuring DNA methylation is Bisulfite Sequencing (BS). The treatment with bisulfite alters the sequence of DNA making the analysis of BS data computationally difficult. There are many tools for analysing BS data but the choice of which to use is difficult due to the extensive biological and technical variability of the data. Synthetic and real datasets can be exploited to evaluate the tool performance and to obtain an accurate data analysis. Today, Sherman is the only available tool to generate BS synthetic datasets. However, this tool does not report any information about the methylated cytosines.**
For this purpose, in this paper we present MethylFASTQ, an easy-to-use bioinformatics tool that generates synthetic bisulfite datasets in FASTQ format. MethylFASTQ works in parallel manner using producer-consumer approach. It returns:
i) a complete dataset in FASTQ format simulating the results of a BS experiment
ii) a report file storing the information about the methylation level of the dataset (i.e. methylated cytosines).
First, we test MethylFASTQ performances with an increasing number of concurrent processes and we report the comparison of MethylFASTQ with respect to Sherman tool. Then, we also describe an application of synthetic datasets generated with our tool and we use them as input for two bisulfite mapping and methylation calling tools.
Finally, we propose MethylFASTQ as a tool to generate synthetic bisulfite sequencing data.

***Index Terms*—DNA methylation, Next Generation Sequencing (NGS), synthetic dataset, parallel computing**

I. Introduction

DNA methylation (DNAm) is the addiction of a methyl group to a DNA molecule. The DNA sequence is composed by four bases: adenine (A), thymine (T), cytosine (C) and guanine (G). The most common form of DNA methylation is the methylation of cytosine which form the 5-methylcytosine (5mC) and it affects a high number of cytosines present in the genome [1]. Methylation changes the activity of DNA without changing its base sequence.

The changes in patterns and levels of DNA methylation are associated with several diseases as cancer and genetic disorders [2]. The gold-standard technique used to study DNAm is the Whole Genome Bisulfite Sequencing (WGBS) that allows to measure methylation in the whole human genome. Conversely, targeted bisulfite sequencing (targeted-BS) allows to sequence the specific genomic regions. Both approaches belong to Next Generation Sequencing (NGS) techniques, a set of advanced technologies that allow the identification of a DNA sequence. The bisulfite treatment converts unmethylated Cs into Ts, while the other bases remain unaffected. Bisulfite conversion alters about 90% of cytosines present in the genome. At this point, distinguishing between Cs converted into Ts and a Ts originally present in the DNA molecule is computationally demanding [1]. On top of that, it is difficult to distinguish a converted C from: i) a stochastic sequencing error occurring during all the sequencing steps; ii) a Single Nucleotide Polimorfisms (SNPs). SNPs are base mutations of the genome that differ among individuals. The presence of SNPs in the samples increases the level of variability of the above data.

Since BS experiments are time and money consuming, the use of synthetic sequencing data (i.e. the creation of a dataset that simulates different biological and technical situations of a BS experiment) has become increasingly popular for assessing and validating bioinformatics tools. Simulations can also be used to evaluate software performances, for debugging purposes and to develop new computational tools [3].

II. Related works

The bioinformatics tools can be benchmarked using real and/or synthetic sequencing data. However, tools validation with real data is essential. Unfortunately, this is a difficult task because the true positive values are unknown and they

* These authors contributed equally

are masked by the extensive biological noise and by the variability of the data. These limitations complicate the use of real data for assessing the accuracy of tools and other performance measures [3]. Synthetic data generator tools allow the production of data with predefined parameters by defining the true positive values.
Furthermore, synthetic datasets allow the generation of a high volume of data in an inexpensive and fast way compared to costs and time needed to create real datasets in laboratory.
Synthetic data generators create FASTQ files starting from a given reference genome. FASTQ file is the *de facto* standard format to store biological data that are sequenced by NGS techniques. FASTQ format describes each read (i.e. substring of DNA) through three fields: the **sequence id** that specifies the unique identifier of the read; the **base sequence** that is the ordered sequence of bases; and the **quality score** that is a measure of quality associated to each base of the sequence.
Synthetic data generators allow to specify a variety of parameters, such as the NGS technique, the read length, the sequencing mode, the coverage and quantity of sequencing errors. The coverage parameter represents the number of times that a single base is sequenced or the number of reads aligned over a single base.
In literature there are several tools that simulate NGS data in FASTQ format, such as ART [4] and CuReSim [5]. However, tools for BS data are still lacking. At the best of our knowledge, *Sherman* is the only one tool that allows to simulate bisulfite sequencing [6]. *Sherman* is a Perl script that generates bisulfite sequencing data in FASTQ format.
Sherman allows the creation of single- and paired-end reads. The number of reads, their length and read quality can be set as tool parameters. SNPs and sequencing errors can also be set and specified. Bisulfite conversion can be regulated with two parameters, which provide the conversion rate in specific DNA contexts (i.e CG and non-CG contexts).

III. METHYLFASTQ

A. Tool overview

MethylFASTQ is a tool written in Python that generate synthetic bisulfite sequencing data in FASTQ format. It is highly customizable because MethylFASTQ is organism-independent and experiment-independent. MethylFASTQ is designed to simulates the sequencing process, following the bisulfite sequencing experiment work-flow (Figure 1).

Given a reference genome sequence as input, the user can create single-end or paired-end reads of directional and non-directional NGS libraries. The single-end mode consists in the production of one read in one direction (i.e. Forward read) for each DNA fragment. Otherwise, the paired-end mode consists in the production of two reads in two directions (i.e. Forward and Reverse reads) for each DNA fragment.
In the non-directional protocol, all four possible bisulfite DNA fragments are sequenced at the same frequency. In the directional protocol, the sequencing reads will correspond to a bisulfite converted version of either the original forward or reverse DNA fragments.

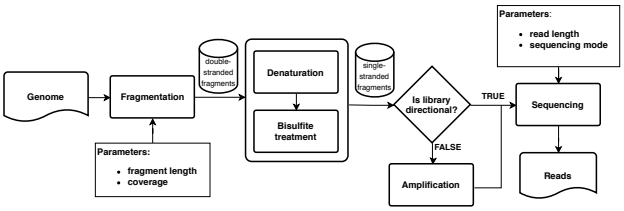


Fig. 1. **Bisulfite Sequencing workflow**. The genome of interest is fragmented in a number of double-stranded pieces of known length. Fragment strands are separated through denaturation and then, single-stranded fragments are bisulfite-treated. Amplification produces reverse complement of treated fragments, which are sequenced in the non-directional protocol. Sequencing step processes bisulfite fragments and produces a set of reads of known length.

MethylFASTQ also allows to simulate both WGBS experiment and targeted-BS data. Two files are returned: a FASTQ file(s) and a methylation call file. In case of single-end sequencing, a single FASTQ file is produced. Differently, in case of paired-end sequencing two FASTQ files are produced which contain respectively the forward and reverse reads. The methylation call file contains the information about the sequenced cytosines.
Two experimental modes are implemented: 1) in the WGBS mode the user can optionally provide a list containing the chromosome names that have to be sequenced. If no list is provided the entire reference genome will be sequenced; 2) in targeted-BS mode the user must provide a tabulated file containing the genome regions to be sequenced. This file will contain the chromosome number and the indexes of first and last base for each region that will be sequenced. Moreover, the user may define the fragment size (i.e the reads length) and the depth of coverage. Methylation can be set through three context-based probabilities: CG, CHG and CHH (where H= A,T or C). The user can also specify probabilities about SNPs and sequencing errors. All the reads which cover a specific base will report the mutated base with a quality is not discernible from a non-mutated base.
Each read in the FASTQ file has an unique record "id" which provides information about its true mapping position in the reference genome. Specifically, the record "id" of a generic read has the form **chr:pos:strand**, where:

- **chr** is the chromosome from which the fragment has been extracted;
- **pos** is the position of the first base in the chromosome;
- **strand** identifies the DNA strand. It can be either forward (+) or reverse (-);

Regarding the methylation call file, it is a file which presents a line for each covered cytosine. Each line has the form **chr pos strand ctx nmeth ntot beta**, where:

- **chr** is the chromosome in which the cytosine is located;
- **pos** is the index of the cytosine in the chromosome (starting from 0);
- **strand** is the strand, it can be either forward (+) or reverse (-);
- **ctx** is the cytosine context, it can be either CG, CHG or CHH;

- **nmeth** represents how many times the cytosine appears as methylated;
- **ntot** represents how many times the base was sequenced;
- **beta** is the beta value of cytosines, defined as the ratio $nmeth/ntot$.

B. Software architecture

MethylFASTQ is modularized in three different modules.

1) `methylfastq` module contains the list of command line arguments and the main class `MethylFASTQ`. This class checks the input parameters and reads the input reference genome file, starting sequencing either in WGBS mode or targeted-BS mode.
2) `sequencing` module implements the sequencing procedures by means of two classes. The first class, called `ChromosomeSequencer`, splits an entire chromosome record in subsequences. These are independently sequenced by the second class, called `FragmentSequencer`.
3) `dna` module contains auxiliary classes that implement different types of DNA sequences, such as double- and single-stranded fragment or single- and paired-end reads.

MethylFASTQ architecture follows the well-known **producer-consumer** software design pattern. The *producer*'s job (Figure 2) is to generate the data and to send it to the consumer. Conversely, the *consumer* (Figure 3) has to consume the received data one at time. Parallelization is process-based and utilizes the built-in module `multiprocessing`, which supports spawning processes and assigning them a job through a function. Inter-process communication is performed using a FIFO queue implemented in `multiprocessing` module, which is process-safe and thread-safe. A process attempting to get an element from an empty queue is blocked until an element is available. In a similar way, a process attempting to put an element in a full queue is blocked until a free slot is available.

The parent process acts as the consumer, whereas the producers are represented by the child processes.
MethylFASTQ works with a chromosome sequence at a time. Chromosome substrings separated by unspecified bases, represented by 'N' characters, are located and extracted. Extracted substrings are split in order to equally distribute the workload among a number of parallel processes.

The load balancing step starts by calculating the total size of the extracted substrings and their average length ($\bar{m}$) that should be assigned to each process. Sequences length $\hat{m} \geq \bar{m}$, longer than the average value, are splitted into M substrings of length $\bar{m}$ and one of length r, where M, r are chosen such that $\hat{m} = \bar{m} \cdot M + r$ with $0 \leq r < M$.

The resulting substrings are sorted with respect to their length in descending order, so that shorter substrings will be processed after the longer ones.
Finally, the user can define a set of processes (workers) that will elaborate the substrings. Sequences with their offsets are distributed among the workers and sequenced in a parallel manner.

Data generated by the workers can be of three types:

1) a list of single-end reads in FASTQ format;
2) a list of paired-end reads in FASTQ format, where the generic paired-end read is a pair;
3) a list storing the methylation information about covered cytosines of the sequenced substring;

so that each kind of data can be stored in a different file.

Workers instantiate a `FragmentSequencer` object using as input parameters the chromosome substring and its initial and final offsets. Random SNPs are set on the string, using the SNP rate parameter given by the user. Then, cytosines on both strands of the sequence are indexed. Cytosines information are stored in a hash table, where the cytosine position into the fragment acts as a key and a `Cytosine` object is the corresponding value. This object contains the strand and context information, as well as two values that take into account how many times that base is covered by a read, and how many times it appears methylated.

Numerous overlapping fragments are extracted from the sequence, so that each base is covered (on average) by a number of reads equal to the chosen depth of coverage. A methylation is generated w.r.t. a probability based on the context (CG, CHG, CHH). Single- or paired-end reads, depending on the chosen sequencing mode, are then extracted from bisulfite strands and stored into a buffer. If the non-directional library has been chosen, reads are also extracted from reverse complement of the bisulfite fragment strands. Whenever the number of reads in the buffer is greater than a certain threshold, it is flushed in the shared queue, so that the parent process can permanently store them in a file. Reads generation involves sequencing error set up and the creation of the relative FASTQ record. Setting up the sequencing errors changes each base with a probability given as input. Quality score associated to changed bases is drastically lowered.
FASTA file scanning and FASTQ record creation are accomplished using BioPython package [7].

IV. Results

In this section are described the results from: (1) the application of MethylFASTQ to generate different synthetic datasets with associated execution times; (2) the comparison between MethylFASTQ and *Sherman* tools performances; (3) the application of MethylFASTQ synthetic datasets in the BS analysis pipeline performed using two BS data mapping and methylation caller tools (BSMAP [8] and Bismark [9]). The experiments were performed on a 48-core AMD Opteron 6176 CPUs at 2.3 GHz with 503 GB of RAM.

A. MethylFASTQ performances

The measure of the execution time is an indicative quantification of software performance. Indeed, the time needed to complete a task is dependent on the machine workload.

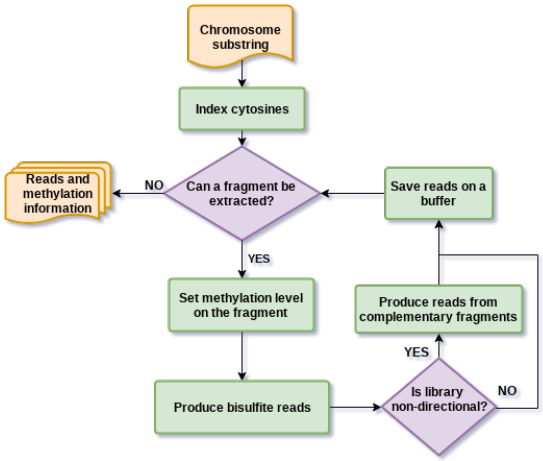


Fig. 2. **Producer process**. Cytosines of the chromosome substring are indexed. Several overlapping substrings are extracted from the chromosome substrings. For each of them, methylation is set and relatives information are stored in the index. Then, the bisulfite fragment is produced and reads are extracted from it. Reads are stored in a local buffer which is periodically flushed in the queue. When fragments extraction terminates, the consumer pushes in the queue the cytosines information and its execution ends.

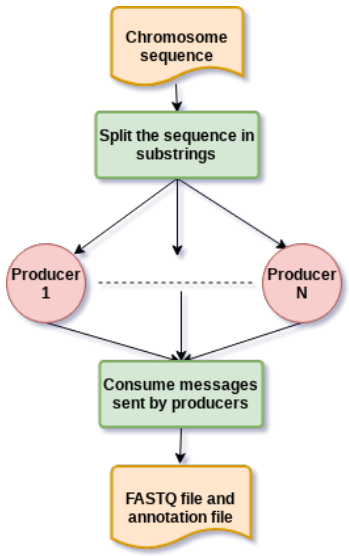


Fig. 3. **Consumer process**. The chromosome sequence is splitted in non-overlapping substrings, which are further divided by the load balancing algorithm. Obtained substrings are assigned to N producer processes. Then, the consumer waits for items to be available in the queue and elaborate them. When all substrings have been sequenced, the consumer terminates.

As reported in Table I the average execution time for the generation of each dataset increases in proportion with the features complexity. Indeed, the lower execution time was obtained for creating the dataset with single-end reads of directional library while the generation of paired-end reads of non-directional library was the most expensive execution. As reported in Figure 4 the MethylFASTQ execution time rapidly drops as the number of parallel processes increases. The execution time using one process was longer than ten hours, while with two processes the execution time was halved, and finally dropped to minutes with seven and eight processes.

Sequencing	Library	Generation time (min)
single-end	directional	15
single-end	non-directional	24
paired-end	directional	25
paired-end	non-directional	44

TABLE I
AVERAGE TIME COMPUTED CONSIDERING 10 RUNS USED TO CREATE THE DATASETS USING EIGHT PARALLEL PROCESSES. ALL THE DATASETS ARE EXTRACTED FROM CHROMOSOME 21 OF HG19 REFERENCE AND HAS 10X COVERAGE. FOR EACH EXPERIMENT, 10 METHYLFASTQ EXECUTIONS HAVE BEEN PERFORMED AND THE AVERAGE TIME HAS BEEN CALCULATED. TIMES ARE EXPRESSED IN MINUTES.

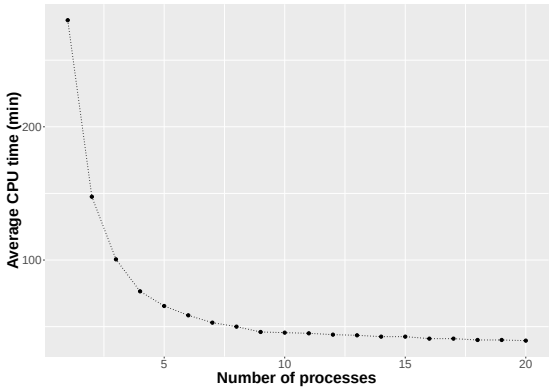


Fig. 4. **MethylFASTQ execution times performances**. Average time of 10 runs to create a dataset as the number of parallel processes increases. The dataset is extracted from human chromosome 21. It is a non-directional library with paired-end reads with 10x coverage.

B. Comparison between MethylFASTQ and Sherman tools

We compared the performance of MethylFASTQ and the already published *Sherman* tool [6] (Figure 5). Both tools generate bisulfite synthetic data in high customizable way and they allow the setting of the reads length, the single-end/paired-end mode and the directionality of the libraries. In addition, they allow the setting of the bisulfite conversion rate for all the cytosines and the simulation of different reads quality scores as well as the number of random SNPs in each read. The final output of both these tools is a FASTQ file, however, Sherman does not produce a report file related to methylation calling for each sequenced cytosine. *Sherman* also does not allow the simulation of a targeted-BS experiment but only a WGBS, because it is not possible to select a set of specific fragments from the reference genome.
The results of the tools comparison show that when both tools run with one process *Sherman* performs better in terms of execution time than MethylFASTQ (Figure 5). This is probably due to the double step of MethylFASTQ that is:
(i) apply the methylation function on genome substrings and save them
(ii) produce a report file storing the information of data methylation profile.
Since *Sherman* is not a parallel tool, the below comparison of

execution times will show the performances of MethylFASTQ using up to eight processes, while *Sherman* runs in sequential mode. The results are different when MethylFASTQ runs with an increasing number of processes. Indeed, the run of MethylFASTQ with two processes obtains comparable execution time with respect to *Sherman*. Instead, with a further increase of the processes number, MethylFASTQ performs better than *Sherman*, due to the parallelization.

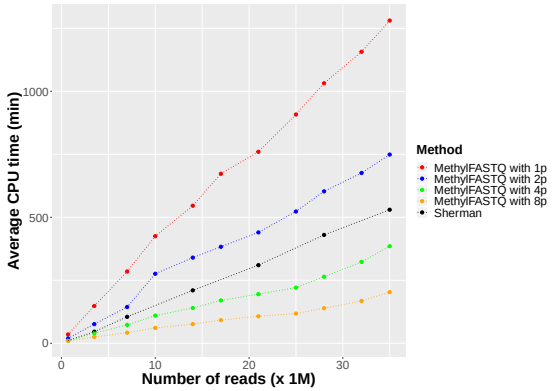


Fig. 5. **Comparison between *Sherman* and MethylFASTQ tools.** Average times to produce datasets of seven different sizes by *Sherman* and MethylFASTQ. MethylFASTQ has been run with 1, 2, 4 and 8 producer processes. Datasets were extracted from human chromosome 21 of human genome hg19. They are non-directional libraries with paired-end reads.

C. *MethylFASTQ helps on the comparison of bisulfite aligners and methylation callers*

The synthetic datasets generated with MethylFASTQ were used as input for a comparative analysis between BSMAP [8] and Bismark [9] performances on the alignment and the methylation calling tasks. These tools follow two different approaches for BS reads mapping: BSMAP applies an approach based on the hashing technique; it masks cytosines in the reference genome to allow bisulfite mismatches. Conversely, Bismark converts both reads and reference in 3-letter sequences and then it applies an algorithm based on the Burrows-Wheeler transform [10]. Methylation calling is performed by methylation extractors included in BSMAP and Bismark packages. All the tools have been tested using their default settings.

The alignment percentage and the recall on identified CG sites were used as performance measurements. The **alignment percentage** considers only the uniquely mapped reads (i.e those reads that are mapped in only one position with a minimum number of mismatches). In case of paired-end reads the reads are aligned if both the extremities are properly mapped. The **recall** is the fraction of true positive values correctly identified as methylated CG sites. It is defined as: TP/Pos, where, TP is the number of CG sites identified by the tool and Pos is the total number of CG sites.

Ten synthetic datasets with different combinations of parameters have been generated to evaluate the tools performances as the library settings and the reads quality level change (Table II).

	Sample	Aligned reads		Recall	
ID	num. reads	BSMAP	Bismark	BSMAP	Bismark
SD1	7.024.152	98.45%	98.68%	98.19%	99.13%
SD2	7.023.824	98.53%	98.37%	93.88%	99.10%
SD3	7.018.280	98.58%	94.95%	89.41%	99.13%
SD4	7.019.916	98.04%	41.37%	95.36%	93.89%
SD5	7.021.892	98.46%	98.65%	96.46%	97.40%
SD6	7.016.484	98.50%	98.55%	94.51%	97.34%
SD7	7.017.776	98.47%	98.58%	94.80%	95.74%
SD8	7.017.556	98.55%	95.55%	89.18%	88.40%
SD9	7.021.028	93.52%	15.46%	74.32%	49.82%
SD10	7.022.140	94.86%	19.77%	63.11%	37.91%
min	7.016.484	93.51%	15.46%	63.11%	37.91%
max	7.024.152	98.58%	98.68%	98.19%	99.13%
avg	7.020.305	97.6%	76%	88.92	85.79%

TABLE III
ALIGNMENT AND METHYLATION EXTRACTION PERFORMANCES ON THE SYNTHETIC DATASETS. MAPPING AND METHYLATION CALLING RESULTS ON SYNTHETIC DATASETS OF BSMAP AND BISMARK TOOLS.

The comparison between alignment performances using these synthetic datasets show that BSMAP is stable as the sequencing error rate or the presence of SNPs increases (Table III). The alignment percentages have little variability, even for low quality datasets. Conversely, Bismark alignment performances vary dramatically with the increase of sequencing errors/SNPs rate. However, the alignment performances have not a great impact on the methylation extraction. Indeed, using low quality datasets with associated low alignment percentages, the methylation extraction works properly. An example is the synthetic dataset 9 (SD9) for which Bismark aligns only 15% reads obtaining a recall of 50% (Table III).

ID	num. reads	SNP rate	Error rate	num. CG sites
SD1	7.024.152	0.1%	0.1%	766.422
SD2	7.023.824	0.1%	1.0%	766.748
SD3	7.018.280	0.1%	2.0%	766.398
SD4	7.019.916	0.1%	5.0%	766.698
SD5	7.021.892	0.3%	0.1%	777.718
SD6	7.016.484	0.3%	0.5%	778.154
SD7	7.017.776	0.5%	0.1%	789.096
SD8	7.017.556	1.0%	1.0%	817.514
SD9	7.021.028	2.0%	5.0%	873.480
SD10	7.022.140	5.0%	2.0%	1.038.142

TABLE II
CONSTRUCTION PARAMETERS OF THE USED SYNTHETIC DATASETS. ALL THE DATASETS ARE EXTRACTED FROM CHROMOSOME 21 OF HG19 REFERENCE. THEY ARE NON-DIRECTIONAL DATASETS WITH PAIRED-END READS OF LENGTH 150 BASES USING A 10X COVERAGE. DATASETS WERE GENERATED FROM HUMAN CHROMOSOME 21 OF HUMAN GENOME HG19.

V. CONCLUSION

In this paper we present MethylFASTQ a new parallel tool to generate bisulfite synthetic datasets. MethylFASTQ allows us to generate both reads and a report file of methylation call, which contains information about methylated cytosines. We showed that our tool helps to find the weaknesses of two mapping and bisulfite caller tools, Bismark and BSMAP. In the future, we will implement MethylFASTQ in C/C++ language

in order to switch from multiprocessing to multithreading, enhancing software performances.

Availability and implementation

MethylFASTQ is released under the GNU GPLv3 license. It is freely available at https://github.com/qBioTurin/MethylFASTQ.

References

[1] Katarzyna Wreczycka, Alexander Gosdschan, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, and Altuna Akalin. Strategies for analyzing bisulfite sequencing data. *Journal of biotechnology*, 261:105–115, 2017.

[2] Christoph Bock. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705, 2012.

[3] Merly Escalona, Sara Rocha, and David Posada. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature Reviews Genetics*, 17(8):459, 2016.

[4] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2011.

[5] Ségolène Caboche, Christophe Audebert, Yves Lemoine, and David Hot. Comparison of mapping algorithms used in high-throughput sequencing: application to ion torrent data. *BMC genomics*, 15(1):264, 2014.

[6] F. Krueger. Sherman - bisulfite-treated Read FastQ Simulator. https://www.bioinformatics.babraham.ac.uk/projects/sherman/. Accessed: 2018-09-20.

[7] P. J. A. Cock et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, mar 2009.

[8] Yuanxin Xi and Wei Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232, 2009.

[9] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572, 2011.

[10] Micheal Burrows and David Wheeler. A Block-Sorting Lossless Data Compression Algorithm. Technical report, DIGITAL SRC RESEARCH REPORT, 1994.

Contents

Appendices

Appendix A

MethylFASTQ

A.1 Synthetic dataset

MethylFASTQ is a Python tool to simulate bisulfite sequencing data in a highly customizable way.

It is organismal independent, because it receives as input a FASTA file containing a reference genome. It is also experiment-independent, since it allows to simulate both whole-genome and targeted bisulfite sequencing.

MethylFASTQ simulates both directional and non-directional libraries, with single- and paired-end reads. Its outputs are two kind of files: FASTQ file(s) and a methylation call file.

In case of single-end sequencing, the tool produces a single FASTQ file, which contains the 5'-end of each sequenced fragment. In case of paired-end sequencing, it produces two FASTQ files, which contain respectively the 5'-end and the reverse complement of 3'-end of the sequenced fragment.
The methylation call file is a tabulated file containing the true methylation call of the sequenced cytosines.

In whole-genome mode, the user can optionally provide a list containing the chromosomes names that have to be sequenced. If no list is provided, the entire reference genome will be sequenced.
In targeted mode, the user has to provide a tabulated file containing the genome regions to be sequenced. A generic line of this file has three fields, in the following order:

1. `chr`: the FASTA record identifier associated to the chromosome to be sequenced;
2. `begin`: the index of the first nucleotide of the region;
3. `end`: the index of the last nucleotide of the region.

The tool allows the creation of directional and non-directional sequencing libraries, with single- or paired-end reads. The tool simulates the entire sequencing process, according to the formalization of the previous section. The user may specify fragment size, read length and the depth of coverage. Methylation can be set through three context-based probabilities: CpG, CHG and CHH.

The user can also specify probabilities about SNPs and sequencing errors. SNPs are nucleotide mutations of the reference genome. All the reads which cover that specific nucleotide have to report the mutated base. Its quality is not discernible from that of a non-mutated nucleotide. A sequencing error simulates an error during PCR or sequencing step. Hence, it is read specific and it is associated with a low quality value.

Each read in the FASTQ file has a record id which gives a number of information about its true mapping position. Specifically, the record id of a generic read has the form `chr:pos:strand`, where:

- `chr` is the chromosome from which the fragment has been extracted. The chromosome name is extracted from the input FASTA file;
- `pos` is the position of the first nucleotide into the chromosome;
- `strand` identifies the bisulfite strand. For directional libraries, it can be either forward (`f`) or reverse (`r`). For non-directional libraries, it can also be the reverse complement of either the forward (`f:rc`) or reverse (`r:rc`) strand.

As regard to the methylation call file, it is a tabulated file which presents a line for each covered cytosine. Each line has the form `chr pos strand ctx nmeth ntot beta`, where:

- `chr` is the chromosome in which the cytosine is located;
- `pos` is the index of the cytosine in the chromosome (it begins from 0);
- `strand` is the , it can be either forward (+) or reverse (−);
- `ctx` is the cytosine context, it can be either CG, CHG or CHH;
- `nmeth` represents how many times the cytosine appears as methylated;
- `ntot` represents how many times the base was sequenced;
- `beta` is the beta value of that cytosine, defined as the ratio `nmeth`/`ntot`.

A.1.1 Software description

MethylFASTQ has a chromosome-based approach. It works with a FASTA record at a time, that usually corresponds to a chromosome. The chromosome sequence is splitted is various large subsequences, which are then sequenced in a parallel manner.

Parallelization is not thread-based, because global interpreter lock (GIL) does not allow CPU usage to multiple threads simultaneously. The GIL is the Python interpreter mutex, that must be held by the current thread in order to safely access objects [**?**]. So, parallelization is process-based and utilizes the built-in module `multiprocessing`, which supports spawning processes and assigning them a job through a function. Child processes produce the data and send them to the parent process using a shared queue. The number of concurrent processes is a tool parameter. It sets the upper bound in the number of workers that can execute simultaneously.

Extracted sequences are assigned to a worker. From each sequence numerous overlapping fragments are extracted, in order to follow the formalization and to obtain the selected coverage. Methylation of a certain fragment is set independently of that of the others and bisulfite reads are extracted from it.

The source code of MethylFASTQ is modularized in different classes located in three modules.

- `methylfastq` module contains MethylFASTQ entry point, the main class `methylFASTQ` and the list of command line arguments accepted by the tool;
- `sequencing` module contains the sequencing procedures by means of two classes. The first class splits an entire FASTA record in subsequences, which are independently sequenced by the other class of the module.
- `dna` module contains various auxiliary classes that implement different types of DNA sequences, such as double- and single-stranded fragment or single- and paired-end reads.

In the following of this section, an in-depth view of each module previously mentioned will be provided.

Methylfastq module

The `methylfastq` module contains the main function, that is the software entry point. In the main function there are listed the command line argument accepted by the script, which are parsed by the built-in module `argparse`. The complete list of parameters is presented in the following:

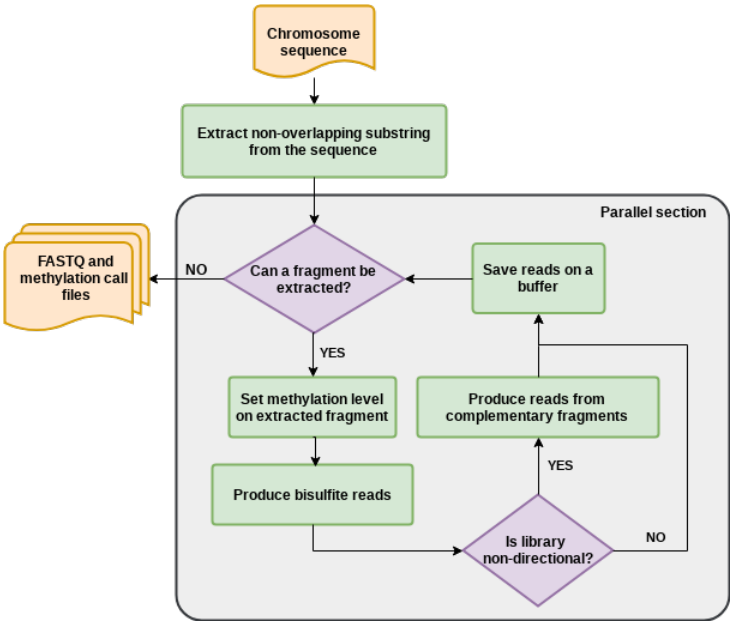


Figure A.1: **Sequencing workflow**. Distincts non-overlapping substrings are extracted from the chromosome sequence and assigned to different processes. Each process extracts numerous fragments from its own sequence. For each of them, random methylation is set and the corresponding bisulfite fragment is produced. Reads are extracted from the fragment, and from its reverse complement if the library has to be non-directional. Produced reads are saved in a buffer and the process is repeated until no fragments remain.

- `-i` or `--in`: path of the FASTA file containing the genome to be sequenced.
- `-o` or `--out`: path where to save the output files.
- `--seq`: the selected sequencing mode. Accepted values are `single_end` and `paired_end`.
- `--lib`: the selected library mode. Accepted values are `directional` and `non_directional`.
- `--chr`: a list of one or more elements that reports the FASTA id of the record to be sequenced.
- `--regions`: a tab-separated file containing the genome regions to be sequenced.
- `--coverage`: the selected depth of coverage value.
- `--fragment`: size of the genome fragments during fragmentation step.
- `--read`: length of the reads to be produced.
- `--maxq`: maximum quality associable to sequenced nucleotides.
- `--minq`: minimum quality associable to sequenced nucleotides.
- `--processes`: processes number to be used during sequencing procedure.
- `--cg`: methylation probability for CG context.
- `--chg`: methylation probability for CHG context.
- `--chh`: methylation probability for CHH context.
- `--snp`: rate to apply mutation in the genome before sequencing phase.
- `--error`: rate to apply sequencing error during sequencing phase.

After argument parsing, an object `MethylFASTQ` is instantiated using the given parameters. It checks the input parameters and creates the directory for the output files. If the tool is executed in targeted mode, it also loads the interval regions to be sequenced from the apposite file.

Execution starts scanning the input FASTA file and selecting only the records whose name appears in the dedicated input parameter. In WGBS

mode, the used parameter is the list of chromosome names, while in targeted mode the record name is searched in the interval regions list. Sequencing procedures are implemented in `sequencing` module.

Sequencing module

`Sequencing` module is the core of MethylFASTQ. It contains two classes. The first one, called `ChromosomeSequencer`, receives an entire FASTA record as input. Its task is to split the FASTA record in multiple substrings, which are then individually sequenced by the other class, called `FragmentSequencer`.

If the chosen sequencing mode is the targeted one, the splitting task is straightforward: it is sufficient to extract the substrings correspondents to the interval regions from the FASTA record. Otherwise, if the chosen sequencing mode is WGBS, it locates and extracts all the substrings separated by unspecified nucleotides, represented by 'N' characters.

Then, a load balancing algorithm splits the extracted substrings in order to equally distribute the workloads across multiple processes. This procedure aims to optimize resource use, maximize throughput and avoid overload of any single process. Algorithm 1 describes the load balancing procedure. It receives as input num_p, the number of concurrent processes that have to be used and *sequences*, a list containing the extracted substrings. A generic element of the list is a triple $(seq, begin, end)$, where:

- *seq* is a chromosome substring;
- *begin* is the position of the first nucleotide into the overall chromosome sequence;
- *end* is the position of the last nucleotide into the overall chromosome sequence.

It starts calculating the total size of the extracted substrings and an average number of bases that should be assigned to each process. Then, sequences longer than the average value are splitted into substrings of length equal to the average. Finally, obtained substrings are sorted with respect to their size in descending order, so that smaller substrings will be processed last. Then, substrings are sequenced in a parallel fashion.

A number of `FragmentSequencer` objects are instantiated with a subsequence and its relative begin and end indexes. Each of these objects sets random mutations on its sequence using the SNP rate parameter.

Then, cytosines on both strands of its sequence are indexed, in order to take into account how many times each of them is covered by a read, and how many times each of them appears as methylated.

Algorithm 1 Load balancing algorithm

Require: *substrings*: an array of triples $(seq, begin, end)$, *num_p*: the number of concurrent processes

Ensure: *new_substrings*: an array of triples $(seq, begin, end)$

```
1:  function LOAD_BALANCING(sequences, num_p)
2:      new_substrings ← newlist()
3:      totsize ← SUMFRAGMENTLENGTHS(fragments)
4:      avg = totsize/num_p
5:      for all (seq, begin, end) ∈ substrings do
6:          size ← end − begin
7:          if size > avg then
8:              num_pieces ← ⌈size/avg⌉
9:              prev ← 0;   curr ← avg
10:             for n ← 0 to num_pieces do
11:                 subseq ← (seq[prev : curr], begin + prev, begin + curr)
12:                 new_substrings.append(subseq)
13:                 prev ← curr;   curr ← curr + avg
14:                 if curr > size then
15:                     curr ← size
16:         else
17:             new_substrings.append((seq, begin, end))
18:     return new_substrings
```

Numerous overlapping fragments are extracted from the sequence, in such a way that each base is covered, on average, by a number of reads equal to the chosen depth of coverage. For each fragment, methylation is randomly put to both strand using the methylation context probabilities and the bisulfite conversion is applied.

Single- or paired-end reads, depending on the chosen sequencing mode, are then extracted from bisulfite strands and stored on a buffer. If the non-directional library has been chosen, reads are also extracted from reverse complement of the bisulfite fragment strands.

DNA module

The `dna` module contains the implementations of various DNA sequence classes, such as single- and double-stranded DNA fragments, and single- and paired-end reads.

The most general one represent a generic DNA sequence, and hence this class is called simply `dna`. A generic DNA sequence is represented by a nucleotide sequence and its begin and end indexes, that allow to know the original location of that sequence in the genome. This class implements all the common operations, such as reverse complement computation, complement sequence calculation, application of bisulfite treatment and so on.

The `fragment` class represents a **double-stranded fragment**. It is used by the `FragmentSequencer` class during fragment extraction. Besides the operation implemented by the `dna` class, a very important new one is to apply methylation to both its strands and to save these informations. Furthermore, there are two methods to extract its forward and reverse strand.

Single-stranded fragment derived from a `fragment` are implemented by the `single_stranded_fragment` class. This class maintains the information about the strand which generates the single-stranded fragment. Possible strands are enumerated by the enum class `strand`. Methods of this class override those of `dna` class, in order to take into account strand diversity.

The last two classes implement respectively single- and paired-end reads. **Single-end reads** are described by `read` class, which is composed by a dna object and a list of integer values. The first component describes the nucleotide sequence of the read, its mapping position and the strand information. The integer list describes the Phred quality scores of the nucleotides.

To represent a realistic read quality, Phred score values of a read follow a gaussian function. In this way, a nucleotide in a certain position has, on average, a better quality of next nucleotides and a worse quality of the previous ones.

Single-end read class has two essential method. The first one allows to set

sequencing errors over the read with a given rate. Quality values associated to mutated nucleotides is lowered. The second one generates the FASTQ record of the read.

Paired-end reads are described by the `paired_end_read` class and are characterized as a pair of single-end reads. Methods of this class simply apply single-end read methods to both the reads of the pair.

A.1.2 Implementation

Let us view how MethylFASTQ actually works. In the following of this section, it will be explained how methylation is set on reads, how fragments are extracted from the reference in order to satisfy the selected depth of coverage and how output data are stored in mass memory.

Apart from FASTQ files, the tool produces another file that contains detailed information about each sequenced cytosine. These information comprehend the number of times a specific cytosine has been sequenced and how many times it appears as methylated.

Cytosine indexing

In order to store these informations, as soon as a new subsequence is extracted from a chromosome, its cytosines of both strands are indexed. This procedure is accomplished by the `__initialize_cytosines` method of `FragmentSequencer` class.
Cytosine informations are stored in the attribute `__cytosines`, that is an hash table. An hash table allows to store key/value pairs. Cytosine position in the fragment is used as key, while a `Cytosine` object is the corresponding value. The `Cytosine` class stores four valuable informations.

- `context`, that represents the cytosine context;
- `strand`, that identifies the strand on which the cytosine is;
- `ntot`, that is the number of times that the cytosine is covered by a read;
- `nmeth`, that is the number of times that the cytosine appears as methylated in the reads.

Context and strand informations are initialized by the class constructor. Context value is used to set cytosine methylation according to the proper methylation probability.

Cytosine indexing procedure in provided in Algorithm 2. To index cytosines of both strands, the fragment is scanned. For each 'C' or 'G' character

found, a new cytosine on forward or reverse strand, respectively, is initialized. The cytosine context has to be identified. Three different contexts are currently recognized: CpG, CHG and CHH, where H = {A, C, T}.

Algorithm 2 Cytosine indexing

Require: *dna_seq*: a DNA string
Ensure: *C_index*: an hash table having cytosine positions as keys and Cytosine objects as values

```
1:  function INDEX_CYTOSINES(dna_seq)
2:      C_index ← new HashTable()
3:      for i ← 0 to dna_seq.length() do
4:          context ← CHH
5:          if dna_seq[i] = C then
6:              if dna_seq[i + 1] = G then
7:                  context ← CG
8:              else if dna_seq[i + 2] = G then
9:                  context ← CHG
10:             C_index[i] ← new Cytosine(context, strand.forward)
11:         else if dna_seq[i] = G then
12:             if dna_seq[i − 1] = C then
13:                 context ← CG
14:             else if dna_seq[i − 2] = C then
15:                 context ← CHG
16:             C_index[i] ← new Cytosine(context, strand.reverse)
17:     return C_index
```

Fragmentation and sequencing

Read production involves several steps. shown below in the respective order:

1. fragment extraction from the reference subsequence;
2. random methylation of the fragment according through the methylation probabilities given by the user;
3. single- or paired-end read production from the fragment;
4. storage of the produced reads.

Fragment extraction happens in the `fragmentation` method of `FragmentSequencer` class. The entire chromosome subsequence is scanned using an index, that is

increased by a little value at each step. The incrementation value is calibrated in such a way that each nucleotide is covered, on average, a number of times equal to the depth of coverage specified by the user.

The index points to the first nucleotide of the next fragment to be extracted. Its length is given by the appropriate parameter given by the user. A double-stranded fragment object (`fragment` class) is instantiated using the nucleotide sequence and its begin and end position in the whole sequence.

Methylation is then set over the extracted fragment. Pseudocode of this procedure is displayed in Algorithm 3. The nucleotide sequence is scanned by one character at a time. If the current base is a C or a G, the cytosine of the appropriate strand is set as methylated with the proper context probability and its counters are updated.

Algorithm 3 Fragment methylation

Require: $fragment$: a double-stranded fragment, C_index: the cytosine index

Ensure: $meth\_fragment$: the methylated double-stranded fragment

```
 1: function METHYLATE_FRAGMENT(dna_fragment)
 2:     meth_fragment ← FRAGMENT(fragment)
 3:     for i ← fragment.begin to fragment.end do
 4:         if meth_fragment[i] = C or meth_fragment[i] = G then
 5:             cytosine ← C_index[i]
 6:             cytosine.ntot ← cytosine.ntot + 1
 7:             meth_prob ← methylation_probability[cytosine.context]
 8:             if RANDOM(0, 1) < meth_prob then
 9:                 cytosine.nmeth ← cytosine.nmeth + 1
10:                 meth_fragment[i].set_as_methylated ()
11:     return meth_fragment
```

Fragmentation method is invoked in both single- and paired-end sequencing methods. The procedure is then straightforward. For each methylated double-stranded fragment:

1. forward and reverse strand are extracted and the bisulfite method is invoked on them;

2. if the dataset under construction is non-directional, also the reverse complement of the two previous bisulfite sequences is computed;

3. from each of the two (or four, in case of non-directionality) sequences,

reads of the appropriate type are generated and stored in a buffer, which is periodically flushed on the shared queue.

Read generation from a single-stranded fragment involves sequencing error set up and the creation of the relative FASTQ record. Sequencing error set up method scans the sequence and mutates each nucleotide with a probability equal to the proper input parameter. Quality score associated to mutated nucleotide is drastically lowered. FASTQ record creation is accomplished using BioPython [**?**] package.

Algorithm 4 shows the generation of a paired-end read. Two substrings, $r1$ and $r2$, are extracted from the single-stranded fragment. They are the prefix and the suffix of the fragment, respectively. Sequencing errors are then set on both the reads, which are then paired in the resulting paired-end read object.

Algorithm 4 Generation of a paired-end read

Require: $fragment$: a single-stranded fragment object, $rlength$: the read length
Ensure: $read$: a paired-end read object

1: **function** PAIREDEND_SEQUENCING($fragment$)
2: $prefix \leftarrow fragment[: rlength]$
3: $suffix \leftarrow fragment[fragment.length() - rlength :]$
4: $r1 \leftarrow$ SINGLEEND_READ($prefix$) $.set\_errors()$
5: $r2 \leftarrow$ SINGLEEND_READ($suffix.reverse\_complement\,()$) $.set\_errors()$
6: **return** PAIREDEND_READ($r1, r2$)

Data persistence

MethylFASTQ architecture follows the well-known **producer-consumer** software design pattern. The producer and the consumer are two processes, who communicate using a shared queue. The producer's job is to generate the data, sent it to the consumer using the queue and repeat. The consumer's job is to consume the received data one at a time, removing it from the queue and make use of. The parent process acts as the consumer, whereas the child processes are the producers.

The parent process instantiates a FIFO queue to allow inter-process communication with their child processes. Queue data structure is implemented in `multiprocessing` module: it is process-safe and thread-safe. A process attempting to get an element from an empty queue is blocked until an element

is available. In a similar way, a process attempting to put an element in a full queue is blocked until a free slot is available.

The parent process spawns a thread that acts as the consumer. This thread waits that an element became available in the queue in order to process it. Then, the parent instantiates a worker pool of a number of processes equal to the value specified by the user. Workers are the producers. The parent shares the data about a substring to be sequenced and the queue with them.

Producers generate the data, namely the FASTQ reads and the true methylation call data. These data are sent to the parent using the shared queue. Different messages are used to communicate these types of data.

A message of type `tuple` indicates that a child process sent data that have to be stored. The tuple is a pair (`datatype`, `data`). The first element is a string that indicates what kind of data has been received. It is used by the parent process to select the correct file in which store the data, that is the second element of the pair. Three data types are supported.

1. `fastq_se` indicates that a list of single-end reads in FASTQ format has been received.

2. `fastq_pe` indicates that a list of paired-end reads in FASTQ format has been received. A generic element of that list is a pair (r1, r2) in FASTQ format.

3. `ch3` indicates that a list containing the methylation information about covered cytosines of a certain fragment has been received.

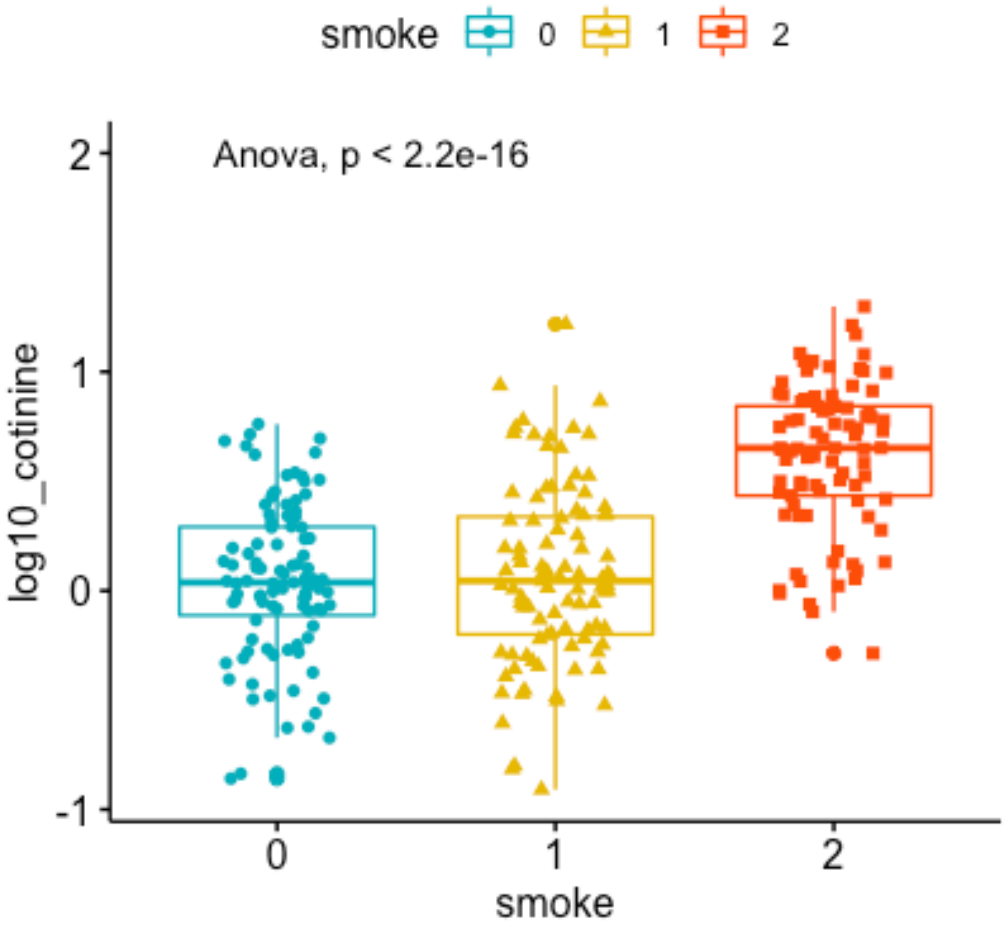
A message of type `int` is the signal that a producer has terminated his job. The consumer knows how many jobs have to be accomplished. This signal is used to decrease the number of remaining jobs, so that when this value became zero, the consumer can terminate.

Supplementary Material

Figure S1: Antibodies Panel used to measure leucocyte percentages. In first row are reported the seven lasers to measure the fluorescence express by cell-types.

Tube	Cell-types	FITC	PE	PrCP-Cy5.5	PE-Cy7	APC	Bv421	Bv510
1	Lymphocytes (NK,B,CD3, CD4, CD8) + Monocytes	CD16+CD56 (NK cells)	GPR15	CD19 (B cells)	CD14 (Monocytes)	CD3 (T cells)	**CD4+**	**CD8+**
2	Granulocytes (Neutrophils and Eosinophils)	CD16	GPR15		CD14	CD11b		

Figure S2: Log transformed cotinine levels in smoking categories. The p-values were calculated with Anova One-way test among groups. Smoking categories were codified with 0= never smokers (blue), 1= former smokers(yellow) and
2= current smokers (red).



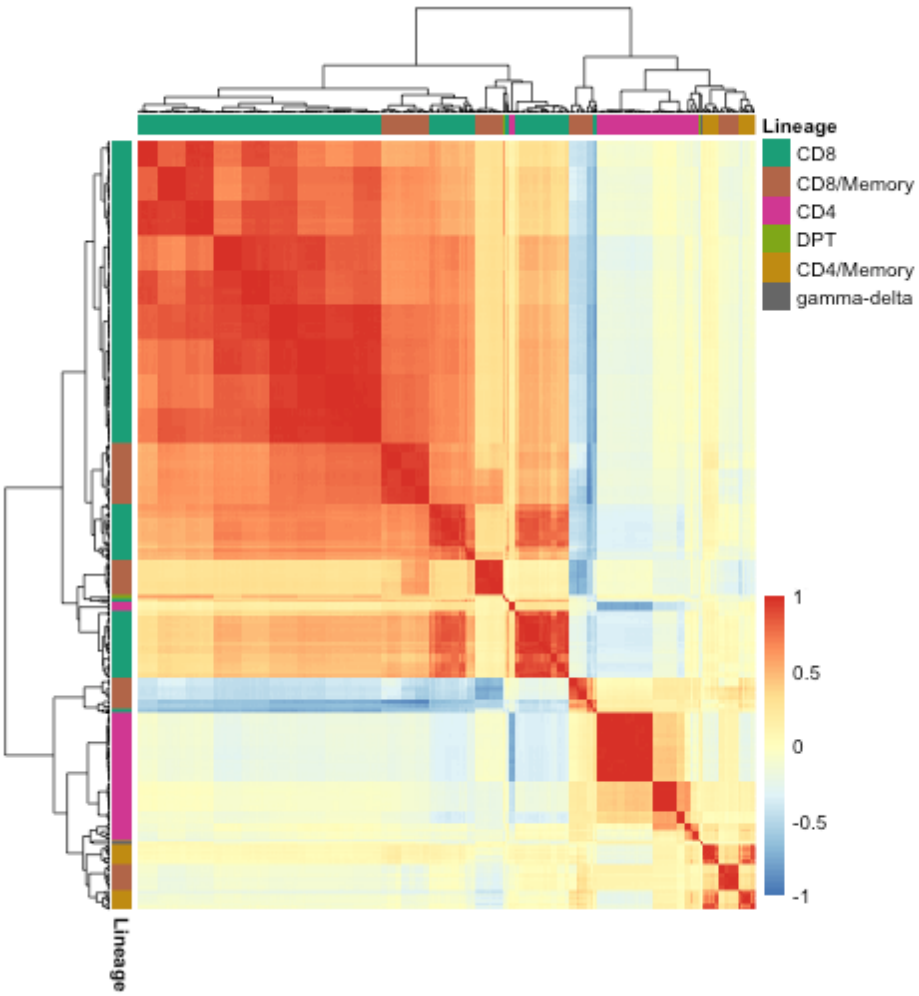


Figure S4. Correlations among T cells associated with smoking status (current *vs* never smokers). The heatmap shows Pearson's correlation coefficients evaluated in the dataset of 497 individuals with immunophenotyping. DPT: Double Positive T cells.

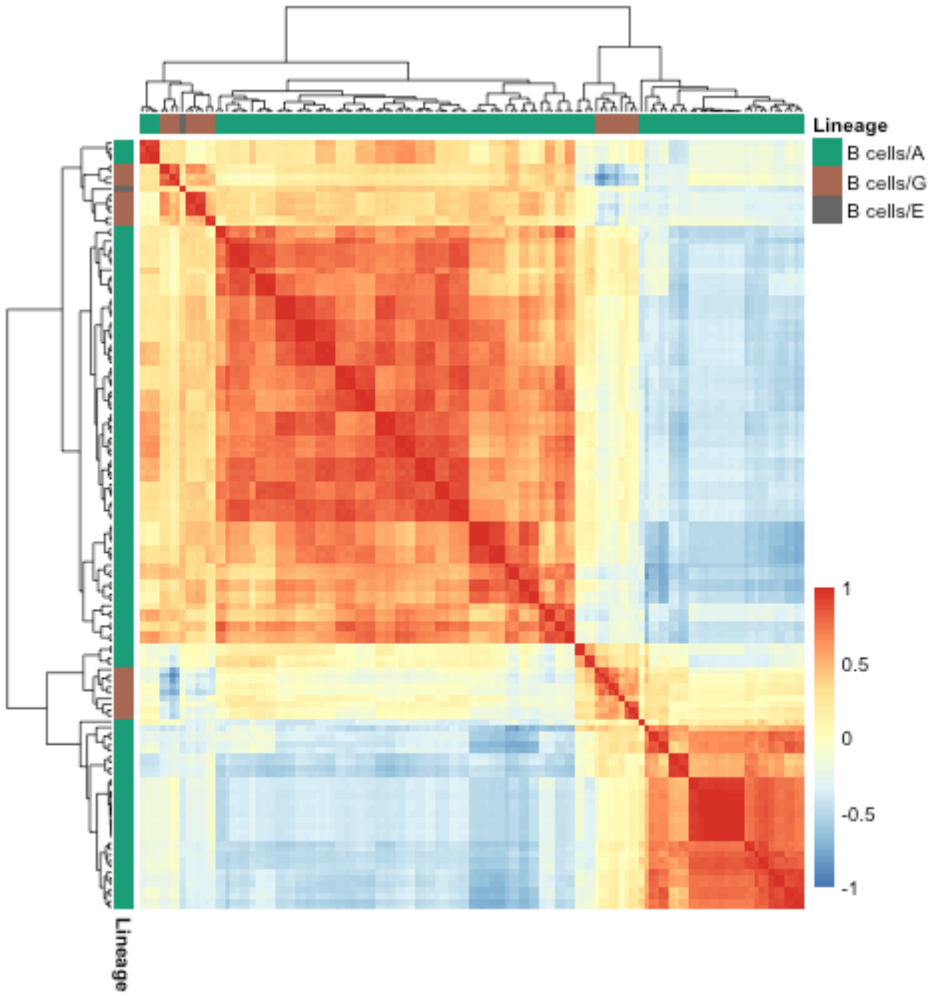


Figure S5. Correlations among B cells associated with smoking status (current *vs* never smokers). The heatmap shows Pearson's correlation coefficients evaluated in the dataset of 497 individuals with immunophenotyping. B cell/A: B cell isotype IgA, B cell/G: B cell isotype IgG, B cell/E: B cell isotype IgE.

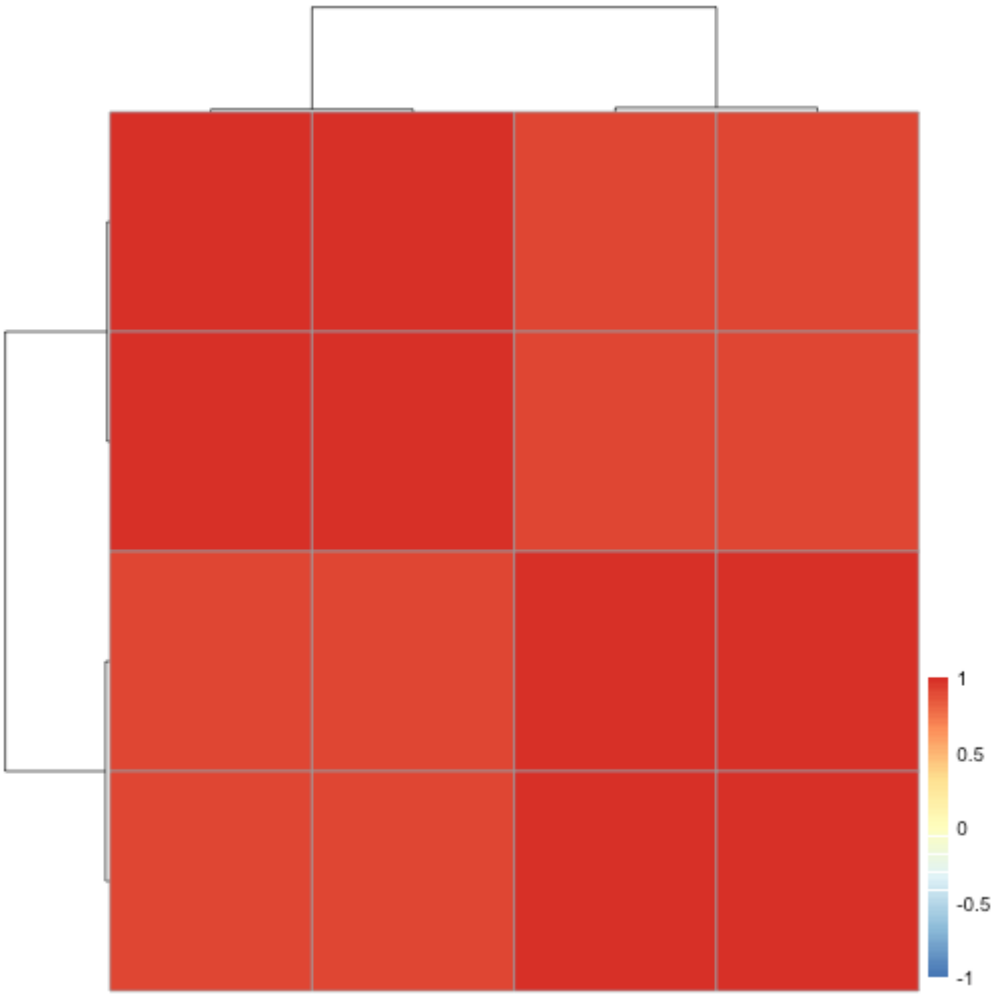


Figure S6. Correlations among monocytes associated with smoking status (current *vs* never smokers). The heatmap shows Pearson's correlation coefficients evaluated in the dataset of 497 individuals with immunophenotyping.

Supplementary Data

Supplementary Table S1. Summary statistics for the associations between immune traits and smoking status (current *vs* never smokers). For each immune trait (N=41,701), the table reports the immune cell population lineage and subset it belongs to, the number of individuals used in the association study (N), and the association study results, as effect size (Beta), standard error (SE), and p-value (P).

Supplementary Table 2. Summary statistics for the associations between immune traits and smoking status (former *vs* never smokers). For each immune trait which was associated with smoking status (current *vs* never smokers, N=848), the table reports the immune cell population lineage and subset it belongs to, the number of individuals used in the association study (N), and the association study results, as effect size (Beta), standard error (SE), and p-value (P). For immune traits significantly different between former and never smokers (N=254) also the results of the association study with the three smoking categories (current *vs* former *vs* never smokers) are reported, as the number of individuals included in the study (N), effect size (Beta.all), standard error (SE.all), and p-value (P.all).

Table S1: Results association between smokers and never smokers.

Type	population	Lineage	Subset.name	Beta	SE	P
CSF	BCell	B cells/A	B cells/A/10-38+	-0.394501786086839	0.0472379165188713	9,20E-01
CSF	BCell	B cells/A	B cells/A/5-24-38-95+	0.699728704926875	0.0972256833368048	1,07E+01
CSF	BCell	B cells/A	B cells/A/5-10-24-38-95+	0.70225593420892	0.097644893028051	1,11E+03
CSF	BCell	B cells/A	B cells/A/24-38-95+	0.657465728080215	0.091766140088583	1,31E+02
CSF	BCell	B cells/A	B cells/A/10-24-38-95+	0.661178293815361	0.0936291930749029	2,39E+03
CSF	BCell	B cells/A	B cells/A/10-38-	0.324887881027117	0.0461614932510667	2,77E+03
CSF	BCell	B cells/A	B cells/A/38-	0.321978990594117	0.0458746458096623	3,10E+03
CSF	BCell	B cells/A	B cells/A/24-27+38-	0.614703039789068	0.0878263257490955	3,56E+03
CSF	BCell	B cells/A	B cells/A/5-10-24-27+38-	0.665069775632493	0.0964486863789021	6,29E+03
CSF	BCell	B cells/A	B cells/A/5-10-38-95+	0.564722638337581	0.0822357258255765	7,17E+03
CSF	BCell	B cells/A	B cells/A/10-38-95+	0.538918322555802	0.0787111658588745	8,06E+03
CSF	BCell	B cells/A	B cells/A/5-10-24-38-	0.532185795937791	0.0780307401920595	9,48E+03
CSF	BCell	B cells/A	B cells/A/10-24-27+38-	0.633594171934642	0.0932696201627639	1,13E+04
CSF	BCell	B cells/A	B cells/A/5-38-95+	0.545281367648957	0.0803990046910234	1,17E+04
CSF	BCell	B cells/A	B cells/A/38-95+	0.521530181697715	0.0773254093558505	1,44E+04
CSF	BCell	B cells/A	B cells/A/5-24-38-	0.526164714128224	0.0781217558392406	1,54E+03
CSF	BCell	B cells/A	B cells/A/5-38-	0.386381360835395	0.0574602558948245	1,62E+04
CSF	BCell	B cells/A	B cells/A/5-24-27+38-	0.666337845741511	0.0993633053541826	1,83E+04
CSF	BCell	B cells/A	B cells/A/5-10-21+24-38-	0.588166746945653	0.0883574568826379	2,42E+04
CSF	BCell	B cells/A	B cells/A/5-21+24-38-	0.579366944261582	0.0875428824497858	3,02E+04
CSF	BCell	B cells/A	B cells/A/5-24-27+38-95+	0.747997086066456	0.114877246679564	5,49E+04
CSF	BCell	B cells/A	B cells/A/10-27+38-95+	0.561033466451792	0.0863121535301561	5,81E+04
CSF	BCell	B cells/A	B cells/A/5-10-27+38-95+	0.585043482891488	0.0901528118612549	6,12E+04
CSF	BCell	B cells/A	B cells/A/27+38-	0.426770612930585	0.0659361601891492	6,83E+04

CSF	BCell	B cells/A	B cells/A/27+38-95+	0.540778625197362	0.083686119000654	7,14E+04
CSF	BCell	B cells/A	B cells/A/10-27+38-	0.434335333418635	0.0675142494701461	8,46E+04
CSF	BCell	B cells/A	B cells/A/5-10-24-27+38-95+	0.76022648658215	0.118207677233449	8,50E+04
CSF	BCell	B cells/A	B cells/A/5-27+38-95+	0.564223572486889	0.0877954831574345	8,66E+04
CSF	BCell	B cells/A	B cells/A/24-27+38-95+	0.702671274756632	0.109701540342242	9,89E+04
CSF	BCell	B cells/A	B cells/A/5-10-27+38-	0.441440127132397	0.0692683304472682	1,19E+05
CSF	BCell	B cells/A	B cells/A/10-24-27+38-95+	0.702217313867128	0.11076527693634	1,42E+05
CSF	BCell	B cells/A	B cells/A/21+24-27+38-	0.690080046236891	0.109106918991262	1,52E+05
CSF	BCell	B cells/A	B cells/A/10-21+24-27+38-	0.68681427747845	0.108704442989912	1,60E+05
CSF	BCell	B cells/A	B cells/A/5-10-21+24-27+38-	0.700055452519357	0.111734776404373	2,13E+05
CSF	BCell	B cells/A	B cells/A/5-21+24-27+38-	0.698018692144139	0.11163751332544	2,26E+05
CSF	BCell	B cells/A	B cells/A/21+24-38-95+	0.705819837526868	0.113103899954595	2,40E+05
CSF	BCell	B cells/A	B cells/A/10-24-38-	0.426504226946917	0.0686103778252315	2,70E+05
CSF	BCell	B cells/A	B cells/A/10-21+24-38-95+	0.713507329160193	0.116579029049873	4,56E+05
CSF	BCell	B cells/A	B cells/A/5-10-21+24-38-95+	0.720740854947372	0.117966175249114	4,82E+05
CSF	BCell	B cells/A	B cells/A/24-38-	0.422894541566451	0.0692860058564971	4,94E+05
CSF	BCell	B cells/A	B cells/A/24-38+	-0.428780508340263	0.0713684815086477	8,04E+04
CSF	BCell	B cells/A	B cells/A/10-21+38-95+	0.557704269873734	0.092814977869413	8,10E+05
CSF	BCell	B cells/A	B cells/A/5-10-21+38-95+	0.558096051908351	0.093664533636699	1,05E+04
CSF	BCell	B cells/A	B cells/A/21+24-27+38-95+	0.774643231338106	0.130096435937591	1,11E+06
CSF	BCell	B cells/A	B cells/A/5-21+38-	0.404685449153408	0.0680288839849893	1,12E+06
CSF	BCell	B cells/A	B cells/A/5-21+24-27+38-95+	0.781105351006614	0.131476192492577	1,19E+05
CSF	BCell	B cells/A	B cells/A/21+38-95+	0.544384961890096	0.0917320980408893	1,20E+06
CSF	BCell	B cells/A	B cells/A/5-10-21+38-	0.414883226400301	0.0701307020710642	1,33E+06
CSF	BCell	B cells/A	B cells/A/5-21+38-95+	0.548367270309202	0.0927765652850067	1,36E+06
CSF	BCell	B cells/A	B cells/A/5-10-21+24-27+38-95+	0.793380264378847	0.134379010612125	1,45E+06
CSF	BCell	B cells/A	B cells/A/38+95-	-0.446995504847011	0.0762244467106822	1,71E+06

CSF	BCell	B cells/A	B cells/A/10-21+24-27+38-95+	0.781524538174046	0.133494144152792	1,87E+06
CSF	BCell	B cells/A	B cells/A/10-21+27+38-95+	0.602693159838195	0.104993150384042	3,28E+06
CSF	BCell	B cells/G	B cells/G/38+	-0.312527405538825	0.0544620655834366	3,33E+06
CSF	BCell	B cells/A	B cells/A/21+38-	0.336044456650204	0.0586339765522412	3,48E+06
CSF	BCell	B cells/A	B cells/A/21+27+38-95+	0.587950735302525	0.102630293809925	3,49E+06
CSF	BCell	B cells/A	B cells/A/5-10-21+27+38-95+	0.606501582303471	0.106006740844492	3,61E+06
CSF	BCell	B cells/A	B cells/A/10-21+38-	0.338150322210311	0.059178203804918	3,78E+06
CSF	BCell	B cells/A	B cells/A/5-21+27+38-95+	0.590598633504179	0.104238929650143	4,80E+03
CSF	BCell	B cells/A	B cells/A/10-24-38+	-0.421275791473889	0.0745530305768921	5,09E+06
CSF	BCell	B cells/A	B cells/A/10-38+95-	-0.453288653532252	0.0804335132943237	5,51E+06
CSF	BCell	B cells/A	B cells/A/21+24-38-	0.446080727458997	0.0791423459697091	5,61E+06
CSF	BCell	B cells/G	B cells/G/5-38-	0.201656476475953	0.0368760904828069	1,29E+06
CSF	BCell	B cells/A	B cells/A/21+27+38-	0.457025363997234	0.0841103708224585	1,53E+07
CSF	BCell	B cells/A	B cells/A/5-10-21+27+38-	0.459339953433159	0.0845264001987107	1,54E+07
CSF	BCell	B cells/A	B cells/A/27-38+	-0.48529803358798	0.0893564951276492	1,56E+07
CSF	BCell	B cells/A	B cells/A/10-21+27+38-	0.465761020022493	0.0858500320624475	1,60E+06
CSF	BCell	B cells/A	B cells/A/10-27-38+	-0.49265981823639	0.0911684954952338	1,79E+07
CSF	BCell	B cells/G	B cells/G/5-21+38-	0.248612946633804	0.046132678060314	1,92E+07
CSF	BCell	B cells/G	B cells/G/21+38-	0.249121768789619	0.0465230857079639	2,25E+06
CSF	BCell	B cells/A	B cells/A/5-21+27+38-	0.457096450769596	0.0853690984462483	2,26E+07
CSF	BCell	B cells/G	B cells/G/10-38+	-0.314008099992953	0.0589206207450133	2,57E+07
CSF	BCell	B cells/A	B cells/A/5-38+	-0.298068427437551	0.0560473390334938	2,78E+07
CSF	BCell	B cells/A	B cells/A/10-21+24-38+	-0.419938703841086	0.0796847379597187	3,32E+07
CSF	BCell	B cells/G	B cells/G/10-21+38-	0.257463400524626	0.0489818443452106	3,61E+07
CSF	BCell	B cells/G	B cells/G/5-10-21+38-	0.26311309100754	0.0508804423875002	5,41E+07
CSF	BCell	B cells/G	B cells/G/5-10-38-	0.198359676729829	0.0383873582284178	5,55E+07
CSF	BCell	B cells/G	B cells/G/38-	0.156796855390897	0.0305908523339081	6,68E+07

CSF	BCell	B cells/A	B cells/A/21+24-38+	-0.3937628586604	0.0772195431861279	7,49E+06
CSF	BCell	B cells/G	B cells/G/24-38+95+	-0.479897935481996	0.0942517700650826	7,82E+07
CSF	BCell	B cells/G	B cells/G/10-38-	0.166619290100548	0.0328556964620894	8,68E+07
CSF	BCell	B cells/A	B cells/A/10-21+38+95-	-0.431174618532106	0.0863667317817414	1,24E+08
CSF	BCell	B cells/A	B cells/A/10-21+27-38+	-0.493766808277057	0.0990696750517664	1,31E+08
CSF	BCell	B cells/A	B cells/A/24-27-38+	-0.535040323854429	0.108693344386122	1,74E+08
CSF	BCell	B cells/A	B cells/A/5-10-21+24-	0.324865339532555	0.0666531889521752	2,16E+08
CSF	BCell	B cells/G	B cells/G/5-38+	-0.288588607298889	0.0593171495935577	2,27E+08
CSF	BCell	B cells/G	B cells/G/21+24-38+95+	-0.480783579001518	0.0987553534342835	2,29E+08
CSF	BCell	B cells/G	B cells/G/38+95+	-0.422907430681068	0.087509382580603	2,57E+07
CSF	BCell	B cells/A	B cells/A/5-21+24-	0.315609091764101	0.0654218106589573	2,72E+08
CSF	BCell	B cells/A	B cells/A/10-24-27-38+	-0.522115033519646	0.109150300501013	3,26E+08
CSF	BCell	B cells/A	B cells/A/21+38+95-	-0.398993512224548	0.0845060487833317	4,25E+08
CSF	BCell	B cells/A	B cells/A/21+27-38+	-0.461414878826826	0.0977479424171729	4,32E+08
CSF	BCell	B cells/A	B cells/A/5-10-21+24-95+	0.426485279908179	0.09040938869921	4,37E+08
CSF	BCell	B cells/A	B cells/A/10-21+24-95+	0.420966859616765	0.0895820215447671	4,73E+08
CSF	BCell	B cells/A	B cells/A/5-10-38+	-0.292550522516742	0.0623734079184542	5,04E+08
CSF	BCell	B cells/A	B cells/A/24-27-38+95-	-0.648944768016674	0.138671248482593	5,16E+08
CSF	BCell	B cells/A	B cells/A/24-38+95-	-0.512388106597754	0.109978656919857	5,60E+08
CSF	BCell	B cells/G	B cells/G/10-24-38+	-0.322402141027604	0.0692160067469545	5,70E+08
CSF	BCell	B cells/G	B cells/G/10-27+38-	0.256174184601792	0.0550458784598882	5,78E+08
CSF	BCell	B cells/G	B cells/G/24-38+	-0.299824255252177	0.0645648087419268	6,05E+08
CSF	BCell	B cells/A	B cells/A/24+38+95+	-0.557843807203931	0.120572475231689	6,51E+08
CSF	BCell	B cells/G	B cells/G/27+38-	0.23118137895197	0.0499888251586187	6,54E+08
CSF	BCell	B cells/A	B cells/A/21+24-95+	0.40505436729387	0.0876813475948723	6,73E+08
CSF	BCell	B cells/A	B cells/A/5-21+24-95+	0.413412721839335	0.08952124739123	6,78E+08
CSF	BCell	B cells/A	B cells/A/5-24+38+95+	-0.560973969680706	0.123245810680059	9,02E+08

CSF	BCell	B cells/A	B cells/A/5-10-21+24-27+95+	0.488244726952176	0.107532857026385	9,51E+08
CSF	BCell	B cells/A	B cells/A/5+27-38+	-0.902342473438917	0.19954424077026	1,03E+09
CSF	BCell	B cells/A	B cells/A/5+10-21+27-38+	-0.906668822441461	0.200588533818411	1,05E+09
CSF	BCell	B cells/A	B cells/A/10-21+27-38+95-	-0.57707873232888	0.12794273629013	1,07E+09
CSF	BCell	B cells/A	B cells/A/5+27-38+95-	-0.913346905827282	0.202453823728515	1,09E+09
CSF	BCell	B cells/A	B cells/A/10-24-38+95-	-0.513959879977124	0.114189436928719	1,12E+09
CSF	BCell	B cells/A	B cells/A/5-10-21-24-38-95+	0.576276298906406	0.127972626893902	1,14E+09
CSF	BCell	B cells/A	B cells/A/5+10-21+27-38+95-	-0.916705539560277	0.204355944569485	1,22E+09
CSF	BCell	B cells/A	B cells/A/5-21-24-38-95+	0.575656485203109	0.128322459280054	1,22E+09
CSF	BCell	B cells/A	B cells/A/5+21+27-38+	-0.897141722329893	0.200195072464176	1,23E+09
CSF	BCell	B cells/A	B cells/A/5+10-27-38+95-	-0.915269239422849	0.204197862748204	1,23E+09
CSF	BCell	B cells/A	B cells/A/10-21+24-27+95+	0.480412561163211	0.107418054555155	1,27E+09
CSF	BCell	B cells/A	B cells/A/5+10-27-38+	-0.900996587545838	0.201592262256503	1,30E+09
CSF	BCell	B cells/A	B cells/A/5-21-38-95+	0.529098627225324	0.118496073183249	1,31E+09
CSF	BCell	B cells/A	B cells/A/5+10-21+27-	-0.803870967608825	0.180409766299958	1,37E+09
CSF	BCell	B cells/A	B cells/A/5+21+27-38+95-	-0.903321101110698	0.202700672705385	1,38E+09
CSF	BCell	B cells/A	B cells/A/5-21+24-27+95+	0.468167864639344	0.10539612994073	1,45E+09
CSF	BCell	B cells/A	B cells/A/5+24-27-38+95-	-0.928450409654407	0.209386660867166	1,51E+09
CSF	BCell	B cells/A	B cells/A/10-24-27-38+95-	-0.622664997322969	0.140854910278639	1,58E+09
CSF	BCell	B cells/A	B cells/A/5+10-21+27-95-	-0.825407713374716	0.186753031012618	1,60E+09
CSF	BCell	B cells/G	B cells/G/24-27+38+	-0.359919085207681	0.0815550582628815	1,64E+09
CSF	BCell	B cells/A	B cells/A/10-21+24-27-38+	-0.511942433531603	0.116017105949115	1,65E+09
CSF	BCell	B cells/A	B cells/A/10-21+24-27+	0.39651270211453	0.0899394035717269	1,67E+09
CSF	BCell	B cells/A	B cells/A/5-10-24-27+95+	0.4222788784654	0.0958868493653467	1,70E+08
CSF	BCell	B cells/E	B cells/E/5-10-24-27+38-	0.754925884739329	0.17166865864343	1,74E+09
CSF	BCell	B cells/A	B cells/A/10-21-24-38-95+	0.538042594288341	0.122689659456167	1,85E+09
CSF	BCell	B cells/A	B cells/A/5-21-27-38+95-	-0.778444517375815	0.176559306287734	1,86E+09

CSF	BCell	B cells/A	B cells/A/21+24-27+95+	0.468196577847108	0.106881558399905	1,87E+08
CSF	Monocyte	Monocytes	Monocytes/16-64+141-DR+	-0.912694789702719	0.180271972490504	9,00E+07
CSF	Monocyte	Monocytes	Monocytes/16-64+DR+	-0.899396790989666	0.179410990091352	1,13E+08
CSF	Monocyte	Monocytes	Monocytes/16-DR+	-0.797917013015882	0.16792430711564	3,70E+07
CSF	Monocyte	Monocytes	Monocytes/16-141-DR+	-0.79147859767956	0.167124408718432	3,98E+08
CSF	TCell	CD8	CD8/25-	-0.0401205955995759	0.00624624397105357	8,98E+04
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-XR3-	-0.373718804990211	0.064757452500993	2,80E+06
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-XR3-XR5-	-0.373523674276885	0.0648135939999875	2,92E+06
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-R10-XR3-	-0.372462898741606	0.0646536882616686	2,95E+06
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-R10-XR3-XR5-	-0.372320415783658	0.0647126742570866	3,07E+06
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-XR3-	-0.402834788529514	0.0705249801204665	3,83E+06
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-XR3-XR5-	-0.402792367840453	0.0705772707759151	3,93E+06
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R10-XR3-	-0.401734515133089	0.0704508099999396	4,03E+05
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R10-XR3-XR5-	-0.401774006709406	0.0705043942101045	4,10E+06
CSF	TCell	CD8	CD8/25+38-73+127+PD1-RO-	0.895658679170521	0.158481983356093	5,18E+06
CSF	TCell	CD8	CD8/25+38-73+127+RO-	0.887158455256977	0.157095299781435	5,29E+06
CSF	TCell	CD8	CD8/25+38-73+127+DR-RO-	0.897049832724384	0.159057032596309	5,50E+06
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-PD1-RO-	0.9222001638295	0.164109391667912	6,07E+05
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-RO-	0.91308421244039	0.162856687648771	6,47E+06
CSF	TCell	CD8	CD8/25+38-73+127+DR-PD1-RO-	0.917204152099258	0.163764619059025	6,67E+06
CSF	TCell	CD8	CD8/25+39-73+127+DR-PD1-RO-	0.884133238520194	0.158119530293901	7,01E+06
CSF	TCell	CD8	CD8/25+73+127+PD1-RO-	0.869377783862024	0.155664877215303	7,25E+05
CSF	TCell	CD8	CD8/25+38-39-73+127+RO-	0.900391992825333	0.161278016515219	7,29E+06
CSF	TCell	CD8	CD8/25+39-73+127+DR-RO-	0.874456040392071	0.156735690154907	7,45E+06
CSF	TCell	CD8	CD8/25+39-73+127+PD1-RO-	0.875133094083868	0.156903914176365	7,52E+05
CSF	TCell	CD8	CD8/25+73+127+RO-	0.859950562496538	0.154265066367619	7,63E+06

CSF	TCell	CD8	CD8/25+39-73+127+RO-	0.865746506010754	0.155448512771685	7,82E+06
CSF	TCell	CD8	CD8/25+73+127+DR-PD1-RO-	0.875194531684365	0.157342097319207	8,11E+06
CSF	TCell	CD8	CD8/25+73+127+DR-RO-	0.865204470638127	0.156042621172555	8,85E+06
CSF	TCell	CD8	CD8/25+38-39-73+DR-PD1-RO-	0.871692408624342	0.15866240981859	1,14E+07
CSF	TCell	CD8	CD8/25+38-DR-RO-	0.86044750314707	0.156752108785439	1,17E+07
CSF	TCell	CD8	CD8/25+38-39-73+DR-RO-	0.863366024517634	0.157714275193745	1,26E+07
CSF	TCell	CD8	CD8/25+38-39-73+PD1-RO-	0.855297039672264	0.156324606570155	1,27E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-	0.824406797617867	0.150975562558143	1,33E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-PD1-	0.827920564824763	0.151751854763783	1,36E+07
CSF	TCell	CD8	CD8/25+38-39-73+RO-	0.847706384912267	0.155383432572199	1,38E+07
CSF	TCell	CD8	CD8/25+38-DR-PD1-RO-	0.863909032139694	0.158592613643382	1,44E+07
CSF	TCell	CD8	CD8/25+38-73+DR-PD1-	0.786229900184668	0.144579549964044	1,49E+07
CSF	TCell	CD8	CD8/25+38-73+DR-	0.780587213812916	0.143847039585714	1,58E+07
CSF	TCell	CD8	CD8/25+DR-RO-	0.837486917568089	0.154363932085586	1,60E+07
CSF	TCell	CD8	CD8/25+39-73+DR-PD1-RO-	0.846074009501128	0.155995727874952	1,61E+07
CSF	TCell	CD8	CD8/25+38-39-127+DR-RO-	0.883927294165827	0.163458079014254	1,74E+07
CSF	TCell	CD8	CD8/25+39-73+PD1-RO-	0.832406028203075	0.154005891426298	1,77E+07
CSF	TCell	CD8	CD8/25+39-73+DR-RO-	0.837112875559749	0.154986621531781	1,80E+07
CSF	TCell	CD8	CD8/25+38-RO-	0.836523240013603	0.154927298202466	1,82E+05
CSF	TCell	CD8	CD8/25+38-127+DR-RO-	0.879141662538317	0.1628467351221	1,82E+07
CSF	TCell	CD8	CD8/25+38-73+127+DR-	0.806158095537646	0.149447573655003	1,84E+07
CSF	TCell	CD8	CD8/25+38-73+127+DR-PD1-	0.810828803200447	0.150313282908552	1,85E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+	0.814070866538389	0.150941553086821	1,85E+07
CSF	TCell	CD8	CD8/27-28+RA-R5-R7+	0.925222386642296	0.171579307782059	1,87E+07
CSF	TCell	CD8	CD8/25+38-39-73+DR-	0.79442305314557	0.147351773942617	1,88E+07
CSF	TCell	CD8	CD8/25+38-39-73+DR-PD1-	0.798181454652136	0.148077546530094	1,89E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+PD1-	0.817617003783051	0.151753961531504	1,90E+07

CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R6-XR3-	-0.397496869900654	0.073760317435007	1,92E+07
CSF	TCell	CD8	CD8/25+38-39-127+DR-PD1-RO-	0.887326687882323	0.164701122666064	1,92E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-	0.597404404791602	0.110928326508178	1,92E+07
CSF	TCell	CD8	CD8/25+DR-PD1-RO-	0.842367476952512	0.15635066156953	1,92E+07
CSF	TCell	CD8	CD8/25+39-73+RO-	0.824166444647228	0.153008470871523	1,93E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR5-	0.600135699900775	0.111514277892245	1,96E+07
CSF	TCell	CD8	CD8/25+38-127+RO-	0.870498695957801	0.161697728032611	1,96E+07
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R6-XR3-XR5-	-0.397387998983119	0.0738107731358945	1,97E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-	0.598008721962152	0.111185472151558	1,99E+07
CSF	TCell	CD8	CD8/25+38-39-DR-RO-	0.845849877830491	0.157202998958879	1,99E+07
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R6-R10-XR3-	-0.396407978233415	0.0736688439352082	1,99E+07
CSF	TCell	CD8	CD8/25+38-127+DR-PD1-RO-	0.882319488590486	0.164061104792101	2,01E+07
CSF	TCell	CD8	CD8/25+38-39-127+RO-	0.875012955943157	0.16272231061823	2,02E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR5-	0.600943827038547	0.111815603718799	2,03E+07
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4-R6-R10-XR3-XR5-	-0.396369838611447	0.0737219155722318	2,04E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+	0.594409206508284	0.110656760781507	2,06E+07
CSF	TCell	CD8	CD8/25+RO-	0.816580905141918	0.152082784311237	2,11E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR5-	0.596837793794769	0.111241387515974	2,12E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-	0.594997655591279	0.110937714151124	2,14E+07
CSF	TCell	CD8	CD8/25+38-73+PD1-	0.766916519134394	0.143022824832078	2,17E+07
CSF	TCell	CD8	CD8/25+38-127+PD1-RO-	0.87359797715972	0.162933243555863	2,18E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR5-	0.597689603566827	0.111568842596081	2,21E+07
CSF	TCell	CD8	CD8/25+38-39-127+PD1-RO-	0.878593538945153	0.16399494629992	2,22E+07
CSF	TCell	CD8	CD8/25+38-PD1-RO-	0.839494382188042	0.156715361230734	2,24E+07
CSF	TCell	CD8	CD8/25+39-73+127+DR-	0.797946607320552	0.149043486378368	2,25E+07
CSF	TCell	CD8	CD8/25+39-73+127+DR-PD1-	0.802181835019198	0.14995299463477	2,29E+07
CSF	TCell	CD8	CD8/25+38-73+	0.76176652377512	0.142382203501579	2,30E+07

CSF	TCell	CD8	CD8/25+38-39-DR-PD1-RO-	0.849464565785246	0.159027723562196	2,40E+07
CSF	TCell	CD8	CD8/25+38-73+127+	0.795627982358133	0.149107158573825	2,45E+07
CSF	TCell	CD8	CD8/25+38-73+127+PD1-	0.80007925573024	0.149976167998968	2,47E+07
CSF	TCell	CD8	CD8/25+39-127+DR-RO-	0.862934181734907	0.161787004060835	2,49E+07
CSF	TCell	CD8	CD8/25+127+DR-RO-	0.859140270614244	0.161120840924908	2,51E+07
CSF	TCell	CD8	CD8/25+PD1-RO-	0.821245624241783	0.1539957699289	2,52E+07
CSF	TCell	CD8	CD8/25+38-39-73+	0.782342911047117	0.146783103047584	2,53E+07
CSF	TCell	CD8	CD8/25+127+RO-	0.85031811892131	0.159514690772029	2,53E+07
CSF	TCell	CD8	CD8/25+38-39-73+PD1-	0.786083570708594	0.147495392880044	2,54E+07
CSF	TCell	CD8	CD8/25+73+DR-PD1-	0.758986195036239	0.142567494235977	2,61E+07
CSF	TCell	CD8	CD8/25+39-127+DR-PD1-RO-	0.866568569186108	0.16305343133525	2,73E+07
CSF	TCell	CD8	CD8/25+127+PD1-RO-	0.854737360480533	0.160826657282608	2,74E+07
CSF	TCell	CD8	CD8/25+127+DR-PD1-RO-	0.863082813579618	0.16241070682259	2,74E+07
CSF	TCell	CD8	CD8/25+39-127+RO-	0.854311131785275	0.160847033645532	2,78E+07
CSF	TCell	CD8	CD8/25+73+DR-	0.751682804279102	0.141688755905113	2,86E+07
CSF	TCell	CD8	CD8/25+38-39-127+DR-	0.815768287363735	0.153912576591429	2,91E+07
CSF	TCell	CD8	CD8/25+38-39-RO-	0.829315999667875	0.156466362822825	2,94E+07
CSF	TCell	CD8	CD8/25+39-DR-RO-	0.820491478814529	0.154867537302286	2,97E+06
CSF	TCell	CD8	CD8/25+39-73+DR-PD1-	0.774775463871747	0.14630976786313	2,99E+07
CSF	TCell	CD8	CD8/25+39-73+127+	0.789357034733652	0.149118799827034	3,01E+07
CSF	TCell	CD8	CD8/25+39-73+DR-	0.769878302864277	0.145475966455181	3,04E+07
CSF	TCell	CD8	CD8/25+39-127+PD1-RO-	0.858273098505679	0.162161715588574	3,04E+07
CSF	TCell	CD8	CD8/25+39-73+127+PD1-	0.793756461440163	0.150056013470959	3,07E+07
CSF	TCell	CD8	CD8/25+38-39-DR-	0.798117468434644	0.150981013372361	3,12E+07
CSF	TCell	CD8	CD8/RA-R5-R7+	0.669208477274053	0.126675743246773	3,13E+07
CSF	TCell	CD8	CD8/25+38-39-127+DR-PD1-	0.821988638395934	0.155751880617582	3,25E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-	0.560596908923904	0.106211539972373	3,27E+07

CSF	TCell	CD8	CD8/25+38-39-73-127+DR-RO-	106.188.581.329.332	0.201187127664762	3,27E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR5-	0.562880492834448	0.106816328066694	3,41E+06
CSF	TCell	CD8	CD8/25+38-39-127+	0.80909905658404	0.153601537202093	3,41E+07
CSF	TCell	CD8	CD8/25-39-	-0.0444599806061857	0.00843843975832863	3,41E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-	0.557499145537725	0.10580598930031	3,42E+07
CSF	TCell	CD8	CD8/25+38-39-DR-PD1-	0.804128135901068	0.152763729232617	3,48E+07
CSF	TCell	CD8	CD8/25+73+PD1-	0.740150826157069	0.140664361061763	3,53E+07
CSF	TCell	CD8	CD8/25+38-39-PD1-RO-	0.832880934777779	0.158285156197885	3,53E+07
CSF	TCell	CD8	CD8/25+39-DR-PD1-RO-	0.824805714818305	0.156810597866212	3,57E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-XR5-	0.559602979500635	0.106421571145106	3,60E+07
CSF	TCell	CD8	CD8/25+38-39-127+PD1-	0.815691826419006	0.155495925331863	3,78E+06
CSF	TCell	CD8	CD8/25+73+127+DR-PD1-	0.779429238597012	0.148600899207489	3,80E+07
CSF	TCell	CD8	CD8/25+73+127+DR-	0.77375803701471	0.147593473364133	3,85E+07
CSF	TCell	CD8	CD8/25+39-73+PD1-	0.764102133060326	0.145796196710912	3,89E+07
CSF	TCell	CD8	CD8/25+73-127+DR-RO-	103.651.278.750.971	0.197774180730637	3,90E+07
CSF	TCell	CD8	CD8/25+73+	0.733004589788374	0.139895622241349	3,92E+06
CSF	TCell	CD8	CD8/25+38-39-	0.787474228946778	0.15034153131524	3,94E+07
CSF	TCell	CD8	CD8/25+39-73+	0.759122203351649	0.144976736942969	3,98E+07
CSF	TCell	CD8	CD8/25+39-127+DR-	0.798204056206945	0.152522191666416	4,01E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+XR5-	0.55091941836414	0.105331628391099	4,12E+07
CSF	TCell	CD8	CD8/25+39-RO-	0.805190459867173	0.153962707612486	4,12E+07
CSF	TCell	CD8	CD8/28+RA-R5-R7+	0.697512217317135	0.133771286877674	4,35E+07
CSF	TCell	CD8	CD8/25+38-39-PD1-	0.793756685762331	0.152149968532986	4,36E+07
CSF	TCell	CD8	CD8/25+39-DR-	0.778862295803255	0.149321852527071	4,37E+07
CSF	TCell	CD8	CD8/25+39-127+DR-PD1-	0.804171348962367	0.154276752542893	4,43E+07
CSF	TCell	CD8	CD8/25+38-39-73-DR-	0.90197901239045	0.173086208152416	4,46E+07

CSF	TCell	CD8	CD8/25+39-127+	0.790848130691659	0.151998088718627	4,64E+07
CSF	TCell	CD8	CD8/25+73+127+PD1-	0.769982844516792	0.148013952135927	4,67E+07
CSF	TCell	CD8	CD8/25+39-73-127+DR-RO-	103.812.841.860.981	0.199521818021783	4,68E+07
CSF	TCell	CD8	CD8/25+73+127+	0.764345520042837	0.147031443717753	4,75E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-	0.554320483147561	0.106636234018013	4,75E+07
CSF	TCell	CD8	CD8/25+39-DR-PD1-	0.784928199590338	0.151057336720197	4,80E+07
CSF	TCell	CD8	CD8/25+38-DR-	0.769637886465857	0.148132266172238	4,82E+07
CSF	TCell	CD8	CD8/25+39-PD1-RO-	0.809644027845178	0.155911904604573	4,92E+07
CSF	TCell	CD8	CD8/25+38-73-127+DR-PD1-RO-	105.925.743.465.549	0.204034275237458	4,94E+07
CSF	TCell	CD8	CD8/25+38-39-73-127+PD1-RO-	106.295.297.299.783	0.204799885409079	4,95E+06
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR5-	0.556410956056095	0.107240693877473	4,98E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+	0.551001031674688	0.106209716706917	4,99E+07
CSF	TCell	CD8	CD8/25+39-127+PD1-	0.797326901324424	0.153855981134153	5,12E+07
CSF	TCell	CD8	CD8/25+38-DR-PD1-	0.776996006396638	0.14994373237666	5,15E+07
CSF	TCell	CD4	CD4/38+73-127-RO-	-0.592020669002826	0.11433556991957	5,21E+07
CSF	TCell	CD8	CD8/25+38-39-73-	0.891795287418237	0.17229278345837	5,27E+07
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-	0.911965770725708	0.176283890514746	5,33E+07
CSF	TCell	CD8	CD8/25+39-	0.76818025109301	0.14853310066041	5,40E+07
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-	0.91095208048995	0.176175907362998	5,41E+06
CSF	TCell	CD8	CD8/25+38-127+DR-	0.791730714261348	0.153133054957398	5,42E+07
CSF	TCell	CD8	CD8/25+38-39-73-127+	0.907317011100994	0.175653826033944	5,54E+07
CSF	TCell	CD8	CD8/28+RA-R5-	0.595036031865201	0.115254926013453	5,57E+07
CSF	TCell	CD4	CD4/38+39-73-127-RO-	-0.59470595774001	0.115189205741545	5,60E+07
CSF	TCell	CD8	CD8/25+38-	0.755385661847385	0.146476689698423	5,80E+06
CSF	TCell	CD8	CD8/25+38-127+DR-PD1-	0.79924933514561	0.155064005887829	5,85E+07
CSF	TCell	CD8	CD8/25+39-PD1-	0.774788930145129	0.15032966763157	5,88E+07
CSF	TCell	CD8	CD8/25+39-73-DR-	0.885883466982923	0.171967661174729	5,92E+07

CSF	TCell	CD8	CD8/25+38-39-73-127+DR-PD1-RO-	106.096.197.497.397	0.206053907792513	6,05E+07
CSF	TCell	CD8	CD8/25+38-PD1-	0.762717842199934	0.148263047190036	6,16E+07
CSF	TCell	CD8	CD8/25+38-39-73-PD1-	0.901559871582111	0.175418109380378	6,27E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-	0.644961381200558	0.125511300947715	6,31E+07
CSF	TCell	CD8	CD8/25+38-127+	0.783766666065136	0.152567059480945	6,34E+07
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-PD1-	0.922073879644001	0.179572953065411	6,40E+07
CSF	TCell	CD8	CD8/25+73-127+DR-PD1-RO-	103.157.397.711.495	0.200827049616979	6,41E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR5-	0.649799972178198	0.126531941054336	6,41E+07
CSF	TCell	CD4	CD4/25-38+73-	-0.296020549235085	0.0577014987834502	6,42E+06
CSF	TCell	CD8	CD8/25+38-39-73-127+PD1-	0.918356789260997	0.178994009288223	6,53E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-	0.646917816145617	0.126095480067021	6,56E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR5-	0.651996500162206	0.127170709956562	6,67E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+	0.640384789874185	0.12498404843723	6,78E+06
CSF	TCell	CD8	CD8/25+38-127+PD1-	0.791536404586851	0.154524020330457	6,80E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+PD1-RO-	0.840027097218004	0.164027673177487	6,88E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR5-	0.645059740069042	0.125985365283675	6,89E+07
CSF	TCell	CD8	CD8/25+39-73-127+DR-	0.900203684823723	0.175858394493939	6,90E+07
CSF	TCell	CD8	CD8/25+39-73-127+PD1-RO-	102.789.761.277.149	0.200765723922018	6,91E+07
CSF	TCell	CD8	CD8/25+39-73-	0.873602962267694	0.170675612607669	6,93E+07
CSF	TCell	CD8	CD8/25+38-39-73-PD1-RO-	100.702.284.970.668	0.196756402160137	6,96E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-	0.642421248894774	0.125609896067456	7,08E+07
CSF	TCell	CD8	CD8/25+DR-	0.749354839579432	0.146550573341644	7,12E+07
CSF	TCell	CD8	CD8/25+39-73-127+	0.892387126017858	0.174563725119681	7,12E+06
CSF	TCell	CD4	CD4/27+31+127-RA+	-0.609281886890086	0.119214362236241	7,19E+07
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-	0.893935857815204	0.17492975869135	7,20E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR5-	0.647322258853654	0.126671415363367	7,22E+07
CSF	TCell	CD8	CD8/25+DR-PD1-	0.757187883241064	0.148390527135573	7,49E+07

CSF	TCell	CD8	CD8/25+39-73-127+DR-PD1-RO-	103.396.521.109.229	0.202606985014835	7,50E+07
CSF	TCell	CD8	CD8/25+127+DR-	0.773718473391884	0.15184541768424	7,72E+07
CSF	TCell	CD4	CD4/25-38+39-73-127-RO-	-0.584537294743484	0.114933305885073	8,09E+07
CSF	TCell	CD8	CD8/25+39-73-127+DR-PD1-	0.909484722950847	0.178962104985858	8,20E+07
CSF	TCell	CD8	CD8/25+39-73-PD1-	0.882564564801285	0.173715863486468	8,28E+07
CSF	TCell	CD8	CD8/25+127+DR-PD1-	0.781047366474429	0.153733269325661	8,28E+07
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-RO+	0.862239547894159	0.169833931008488	8,42E+07
CSF	TCell	CD8	CD8/25+39-73-127+PD1-	0.902489920494273	0.177775664770421	8,42E+07
CSF	TCell	CD8	CD8/25+	0.734176917927622	0.144625233308176	8,48E+07
CSF	TCell	CD8	CD8/25+38-39-73+DR-RO+	0.856727906338546	0.169017452798133	8,74E+07
CSF	TCell	CD8	CD8/25+PD1-	0.742340729986619	0.14645991705868	8,79E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3+XR5-	102.192.320.410.268	0.201721964403727	8,90E+05
CSF	TCell	CD8	CD8/25+127+	0.763977130747828	0.150947929454479	9,07E+07
CSF	TCell	CD8	CD8/25+127+DR-RO+	0.838538059710216	0.165766445551728	9,13E+07
CSF	TCell	CD8	CD8/25+38-73-127+DR-	0.88296265223338	0.174743921602819	9,40E+07
CSF	TCell	CD8	CD8/25+127+PD1-	0.772006523841541	0.152889106696221	9,58E+07
CSF	TCell	CD8	CD8/25+38-73-127+	0.876863248673461	0.174099165492744	1,01E+08
CSF	TCell	CD4	CD4/25-38+39-73-DR+	-0.747793204706806	0.148388771423797	1,02E+08
CSF	TCell	CD8	CD8/25+38-39-73+127+DR-PD1-RO+	0.863126330016295	0.171433404422427	1,03E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3+XR5-	0.995533391853959	0.197698623368859	1,03E+08
CSF	TCell	CD8	CD8/25+38-39-73+127+RO+	0.856370587653628	0.170416422544955	1,07E+08
CSF	TCell	CD4	CD4/25-38+73-127-RO-	-0.58578155098465	0.116587929280112	1,07E+08
CSF	TCell	CD8	CD8/27+RA-R5-R7+	0.669927401071901	0.133388082219652	1,08E+08
CSF	TCell	CD8	CD8/25+38-39-73+DR-PD1-RO+	0.859667636915083	0.171144919481654	1,08E+08
CSF	TCell	CD8	CD8/25+38-73-127+DR-PD1-	0.895082115100362	0.178216553332057	1,08E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-XR3+XR5-	0.933609889647887	0.185824617337566	1,09E+08
CSF	TCell	CD8	CD8/25+38-39-73+RO+	0.850795943526872	0.16947042971432	1,10E+08

CSF	TCell	CD8	CD8/25+127+RO+	0.828421297519912	0.165149340802426	1,11E+08
CSF	TCell	CD8	CD8/25+38-73-DR-	0.849455450750928	0.169480390168287	1,14E+08
CSF	TCell	CD8	CD8/25+38-73-127+PD1-	0.889836987245174	0.177579389398003	1,14E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-XR3+XR5-	0.925421206200019	0.184677778094791	1,15E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3+XR5-	103.701.145.233.215	0.206982180357562	1,15E+07
CSF	TCell	CD8	CD8/27+28+RA-R5-R7+	0.678817347912871	0.13562975533613	1,17E+08
CSF	TCell	CD4	CD4/38+73-127-DR-RO-	-0.578972300861669	0.115734875998524	1,19E+08
CSF	TCell	CD8	CD8/27-57-127+244-RA-	0.775467846981258	0.155063800683782	1,20E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R6-XR3-	-0.950784101803191	0.190215423608864	1,21E+07
CSF	TCell	CD4	CD4/25-38+73-DR-	-0.292285935539984	0.0585256142792199	1,21E+08
CSF	TCell	CD4	CD4/38+73-DR-	-0.291341668957826	0.0583387285795782	1,22E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R6-XR3-XR5-	-0.950397965769078	0.190256651994638	1,23E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R6-R10-XR3-	-0.950519450828205	0.190356428886769	1,24E+08
CSF	TCell	CD8	CD8/27-57-95+127+244-RA-	0.775763920606306	0.155359391100233	1,24E+07
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R6-R10-XR3-XR5-	-0.950235716545105	0.190384245480621	1,25E+08
CSF	TCell	CD8	CD8/27-127+244-RA-	0.773714766565097	0.155048914918524	1,26E+08
CSF	TCell	CD8	CD8/25+73-127+DR-	0.870009730947577	0.174479447651548	1,28E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3+XR5-	101.197.101.157.543	0.203046643398192	1,30E+08
CSF	TCell	CD8	CD8/25+38-39-73+127+PD1-RO+	0.857595607984258	0.172107481513734	1,30E+08
CSF	TCell	CD8	CD8/27-95+127+244-RA-	0.773935223819175	0.155354157735631	1,31E+08
CSF	TCell	CD8	CD8/27-28+57-127+244-RA-	0.775884854253569	0.155802572432737	1,32E+07
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-	0.860057524730818	0.172712841086303	1,32E+08
CSF	TCell	CD8	CD8/27-28+57-95+127+244-RA-	0.776315912888953	0.155930896391764	1,33E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-	0.609120649283286	0.122370637392061	1,33E+08
CSF	TCell	CD8	CD8/25+38-73+127+DR-RO+	0.834234046932266	0.167627329565778	1,34E+08
CSF	TCell	CD8	CD8/25+38-73-	0.83802037613311	0.168431630106449	1,35E+08

CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-XR5-	0.614241943076456	0.123469385325541	1,35E+08
CSF	TCell	CD8	CD8/25+38-39-73+PD1-RO+	0.854070902021497	0.171680311961212	1,36E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R6-XR3-	-0.97031821408253	0.195122005410639	1,36E+08
CSF	TCell	CD4	CD4/38+127-RO-	-0.575838960980693	0.115796408795736	1,36E+08
CSF	TCell	CD8	CD8/27-28+127+244-RA-	0.775086201266979	0.155913840625363	1,38E+07
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R6-XR3-XR5-	-0.969987562552939	0.195165704149835	1,38E+08
CSF	TCell	CD8	CD8/27-28+95+127+244-RA-	0.775486284931094	0.156049461873173	1,39E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3+	100.835.507.630.744	0.20289801286482	1,39E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-	0.60477060971568	0.121821907880603	1,42E+08
CSF	TCell	CD8	CD8/25+73-127+	0.859394824191448	0.173180537367915	1,43E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R6-R10-XR3-	-0.968417182108096	0.195166975977587	1,43E+08
CSF	TCell	CD4	CD4/38+39-73-	-0.297633663052406	0.0600256995424213	1,43E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-	0.598718072957056	0.120672852628757	1,44E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-XR5-	0.609712278176525	0.122907323087383	1,44E+08
CSF	TCell	CD4	CD4/38+39-127-RO-	-0.579224182347182	0.116787212474376	1,45E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-XR5-	0.603438742206411	0.121710834562235	1,46E+08
CSF	TCell	CD4	CD4/38+39-73-DR+	-0.736716497538433	0.148506710510706	1,47E+07
CSF	TCell	CD8/Memory	CD8/Memory/161+R6-XR3-XR5-	-0.894158575649381	0.180377761936256	1,48E+08
CSF	TCell	CD8	CD8/25+73-127+DR-PD1-	0.880846593217173	0.177801741571583	1,49E+08
CSF	TCell	CD8	CD8/27-31-95+127+244-	0.767988446169645	0.155023095922587	1,49E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R6-R10-XR3-XR5-	-0.967329945056132	0.195460986912225	1,52E+08
CSF	TCell	CD4	CD4/25-38+39-73-	-0.297062830466197	0.0600653770032293	1,52E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R6-R10-XR3-XR5-	-0.893381696868991	0.180479264394882	1,53E+08
CSF	TCell	CD8	CD8/27-31-127+244-RA-	0.791029372125551	0.159888796985322	1,54E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+	0.5939169664904	0.120062156729328	1,54E+08
CSF	TCell	CD8	CD8/25+38-73-PD1-	0.849165264430683	0.171673552730769	1,54E+07

CSF	TCell	CD8	CD8/25+38-73+DR-RO+	0.80046793265701	0.161866443997693	1,55E+08
CSF	TCell	CD8	CD8/27-31-57-127+244-RA-	0.791455290365092	0.160037349295472	1,55E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3+	0.983600068582276	0.198945258008203	1,57E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+XR5-	0.598466436781359	0.121080745235989	1,57E+08
CSF	TCell	CD8	CD8/25+38-73-DR-RO-	0.945763037058885	0.191319133061879	1,57E+08
CSF	TCell	CD8	CD8/25+39-73+127+DR-RO+	0.832938584126571	0.168562175220228	1,58E+08
CSF	TCell	CD8	CD8/27-95+127+244-	0.732702055643052	0.148277288033366	1,58E+08
CSF	TCell	CD8	CD8/25+39-73+DR-RO+	0.827426564411335	0.167540173353609	1,60E+08
CSF	TCell	CD8	CD8/27-31-95+127+244-RA-	0.791185553947494	0.160243249627004	1,61E+08
CSF	TCell	CD8	CD8/27-31-57-95+127+244-RA-	0.791746826192192	0.16039349366831	1,62E+08
CSF	TCell	CD8	CD8/25+73-127+PD1-	0.87113881692346	0.176533906655618	1,63E+08
CSF	TCell	CD8	CD8/27-31-57-95+127+244-	0.766793133277445	0.155385572630337	1,63E+08
CSF	TCell	CD8	CD8/25+38-73+127+DR-PD1-RO+	0.83605376932471	0.16943294340567	1,63E+08
CSF	TCell	CD8	CD8/25+73-DR-	0.831289146888192	0.168475117417158	1,63E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-XR3+	0.92083358523535	0.186692678606516	1,66E+08
CSF	TCell	CD4	CD4/27+31+57-127-RA+	-0.59477020180336	0.120651306661806	1,67E+08
CSF	TCell	CD8	CD8/27-28+31-127+244-RA-	0.794004474317577	0.161147498357381	1,69E+08
CSF	TCell	CD8	CD8/27-28+31-57-127+244-RA-	0.793678885383684	0.161149674553717	1,70E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3+	102.648.670.935.026	0.208400909292848	1,70E+07
CSF	TCell	CD4	CD4/38+73-	-0.288991690834849	0.0587496961664584	1,71E+08
CSF	TCell	CD8	CD8/27-28+31-95+127+244-RA-	0.794048214765004	0.161259729629989	1,71E+08
CSF	TCell	DPT	DPT/25-39-DR+	-106.852.494.601.682	0.214984172226102	1,71E+08
CSF	TCell	CD4	CD4/38+39-73-DR-	-0.295957907978454	0.0601464762093715	1,72E+07
CSF	TCell	CD8	CD8/27-28+31-57-95+127+244-RA-	0.793811104946928	0.161262235722198	1,72E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-XR3+	0.913265225203085	0.18552096309441	1,73E+08
CSF	TCell	CD8	CD8/27-57-95+127+244-	0.731025936862354	0.148591028590826	1,75E+08

CSF	TCell	CD8	CD8/25+38-73+127+RO+	0.82639854226632	0.168006731610345	1,75E+08
CSF	TCell	CD8	CD8/25+38-73+DR-PD1-RO+	0.803149045700472	0.163408065393483	1,78E+08
CSF	TCell	CD4	CD4/38+73-127-PD1-RO-	-0.567424244567535	0.115702313052548	1,87E+08
CSF	TCell	CD4	CD4/27+28+31+57-127-RA+	-0.609511621533267	0.124349168019342	1,90E+08
CSF	TCell	CD8	CD8/25+DR-RO+	0.793874659905452	0.16200291594116	1,90E+08
CSF	TCell	CD8	CD8/25+73-DR-PD1-	0.841189475209678	0.171658316027553	1,91E+08
CSF	TCell	CD8	CD8/25+39-73+DR-PD1-RO+	0.831248337290158	0.169666297377428	1,92E+08
CSF	TCell	CD8	CD8/27-28+31-57-95+127+244-	0.767494703637734	0.156662941570231	1,92E+08
CSF	TCell	CD8	CD8/25+39-73+127+DR-PD1-RO+	0.834402530631449	0.17035311311578	1,93E+08
CSF	TCell	CD8	CD8/27-28+31-95+127+244-	0.767220280628002	0.156652032308926	1,93E+08
CSF	TCell	CD8	CD8/25+38-39-73-DR-RO-	0.946975022563097	0.193406414176077	1,95E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+XR3+XR5-	0.8653569893116	0.176727755417188	1,95E+08
CSF	TCell	CD8	CD8/25+38-73+RO+	0.793320789313046	0.162122953936789	1,97E+08
CSF	TCell	CD8	CD8/25+73-	0.816516827214282	0.166892887657457	1,98E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3+	0.999738644584699	0.204372229408406	1,99E+08
CSF	TCell	CD8	CD8/27-28+57-95+127+244-	0.733827443304689	0.150148362247379	2,03E+08
CSF	TCell	CD8	CD8/25+39-73+RO+	0.820267985056145	0.167934200106866	2,05E+07
CSF	TCell	CD8	CD8/25+39-73+127+RO+	0.826368804952199	0.169202860470211	2,05E+08
CSF	TCell	CD8	CD8/25+38-73-127+RO-	0.969958281783265	0.198603979262936	2,06E+08
CSF	TCell	CD8	CD8/27-28+95+127+244-	0.733084985819374	0.150237751189452	2,10E+08
CSF	TCell	CD4	CD4/25-38+39-DR+	-0.700469183827762	0.143503091231553	2,11E+08
CSF	TCell	CD4	CD4/27+28+31+127-RA+	-0.612824443834532	0.125674450919909	2,12E+07
CSF	TCell	CD8	CD8/25+38-73+127+PD1-RO+	0.828378259244389	0.169884922560686	2,13E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-XR3+XR5-	0.880975175109124	0.18078161330641	2,17E+08
CSF	TCell	CD8	CD8/25+38-73+PD1-RO+	0.796620986422484	0.163735621657944	2,24E+08
CSF	TCell	CD8	CD8/25+73-PD1-	0.827018705277322	0.170092032129622	2,27E+08
CSF	TCell	CD4	CD4/25-38+39-127-RO-	-0.575536137803467	0.118509928526154	2,32E+08

CSF	TCell	CD8	CD8/25+38-39-127+DR-RO+	0.819619051913999	0.168829522301742	2,34E+08
CSF	TCell	CD8	CD8/25+RO+	0.782636350060212	0.161213821261747	2,34E+08
CSF	TCell	CD8	CD8/25+38-73-RO-	0.91649270102084	0.188786246597752	2,36E+08
CSF	TCell	CD8	CD8/25+38-73-127+DR-RO-	0.973712897965955	0.200624057583903	2,38E+08
CSF	TCell	CD4	CD4/38+73-127+	-0.295364055873402	0.0609183622891517	2,38E+08
CSF	TCell	CD8	CD8/25+39-73+PD1-RO+	0.824444838617056	0.170138355993002	2,45E+08
CSF	TCell	CD8	CD8/25+PD1-RO+	0.793337850407505	0.163784347089412	2,46E+08
CSF	TCell	CD4	CD4/25-38+73-127-DR-RO-	-0.572464736311224	0.118212359079276	2,47E+08
CSF	TCell	CD4	CD4/38+73-PD1-	-0.293286401937499	0.0605910331504833	2,47E+08
CSF	TCell	CD4	CD4/25-38+73-PD1-	-0.294452583820471	0.0608390280299047	2,48E+08
CSF	TCell	CD4	CD4/25-38+73-127+	-0.295018734198996	0.0609632148805547	2,48E+08
CSF	TCell	CD8	CD8/25+39-73+127+PD1-RO+	0.828177355890604	0.171073749095735	2,50E+08
CSF	TCell	CD8	CD8/25+38-39-73-127+RO-	0.973642788970427	0.201170864381891	2,52E+08
CSF	TCell	CD4	CD4/25-38+39-73-DR-	-0.292938431821091	0.0605976850807829	2,53E+08
CSF	TCell	CD8	CD8/25+73-DR-RO-	0.910190732507963	0.188175498062413	2,55E+08
CSF	TCell	CD4	CD4/25-38+127-RO-	-0.570701427652571	0.118027876810991	2,56E+08
CSF	TCell	CD8	CD8/25+38-39-127+DR-PD1-RO+	0.827346251941764	0.171142539809142	2,57E+08
CSF	TCell	CD8	CD8/25+73+127+DR-RO+	0.803973247515928	0.166330008989791	2,58E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR3+XR5-	0.948594962830946	0.19625041789842	2,60E+08
CSF	TCell	CD8	CD8/25+38-39-127+RO+	0.813647480151884	0.168423711437438	2,60E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3+XR5-	0.895639073441476	0.185326535748196	2,61E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-XR3-	-0.867835488508845	0.179613611935643	2,61E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-XR3-XR5-	-0.867520349593784	0.179663422081009	2,64E+07
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-XR3-	-0.614897540024935	0.12741323016215	2,66E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-R10-XR3-	-0.614620286757921	0.127371590292827	2,66E+08
CSF	TCell	CD8	CD8/25+38-39-DR-RO+	0.815065713397053	0.168924578083662	2,68E+08
CSF	TCell	CD4	CD4/38+127-DR-RO-	-0.566628359034826	0.117501136285951	2,71E+08

CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R10-XR3-	-0.866904263895364	0.179750660223639	2,71E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R10-XR3-XR5-	-0.866676978934144	0.179787744152144	2,74E+07
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+XR3+	0.855361576744128	0.17749963460935	2,77E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R6-XR3-XR5-	-0.766396999234538	0.159195222991119	2,82E+08
CSF	TCell	CD8	CD8/25+38-39-127+PD1-RO+	0.821850372413154	0.170766887175277	2,83E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3+XR5-	0.931657826359585	0.19352357005914	2,85E+08
CSF	TCell	CD4	CD4/25-38+73-127-PD1-RO-	-0.557022881780256	0.115771642340977	2,86E+08
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-RO+	0.897327638681147	0.186626777318054	2,88E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-XR3+XR5-	0.886968164438144	0.184445932198348	2,90E+08
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-PD1-RO+	0.911126352922938	0.189635755380661	2,92E+08
CSF	TCell	CD4	CD4/25-38+39-73-DR+RO+	-0.753911089966943	0.156894062312869	2,97E+07
CSF	TCell	CD8	CD8/25+38-39-RO+	0.808688717894007	0.168453474649013	2,99E+08
CSF	TCell	CD8	CD8/25+38-39-73-127+RO+	0.891928846389925	0.18585133367033	3,00E+08
CSF	TCell	CD8	CD8/25+38-39-DR-PD1-RO+	0.823017788965108	0.171476872276664	3,00E+08
CSF	TCell	CD8	CD8/25+38-39-73-127+PD1-RO+	0.906309869562748	0.188876355839074	3,01E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R6-R10-XR3-XR5-	-0.76466079207615	0.159351141766936	3,03E+08
CSF	TCell	CD8	CD8/25+38-39-73-RO-	0.921232341970071	0.192039010504189	3,06E+08
CSF	TCell	CD4	CD4/38+39-DR+	-0.689497471544949	0.143743227740751	3,09E+08
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-XR3+	0.870935533750622	0.181722007505583	3,12E+08
CSF	TCell	CD8	CD8/25+73+127+DR-PD1-RO+	0.806711697244945	0.168366074874405	3,12E+08
CSF	TCell	CD8	CD8/25+39-73-DR-RO-	0.909753918575962	0.189861473430292	3,13E+08
CSF	TCell	CD8	CD8/25+38-39-73-DR-RO+	0.892326490232378	0.186477793831473	3,19E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR3+XR5-	0.939369812567759	0.19629106371021	3,21E+07
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-RO-	0.943831212268446	0.197308495011857	3,24E+07
CSF	TCell	CD4	CD4/38+73-DR-PD1-	-0.29218859233987	0.0611425570548298	3,27E+08
CSF	TCell	CD4	CD4/25-38+73-DR-PD1-	-0.293131534966344	0.0613406821971455	3,27E+08
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-RO+	0.906038636133021	0.189549304454945	3,27E+08

CSF	TCell	CD4	CD4/27+28+31+127-244-RA+	-0.61313111008181	0.128262259877634	3,27E+08
CSF	TCell	CD4	CD4/38+73-127+DR-	-0.292602781609675	0.0612391172753005	3,27E+07
CSF	TCell	CD8	CD8/25+73+DR-RO+	0.764117029367035	0.159846195574731	3,28E+08
CSF	TCell	CD8	CD8/25+38-39-PD1-RO+	0.817223963389791	0.171002596088612	3,29E+08
CSF	TCell	CD8	CD8/25+38-39-73-RO+	0.886068514932227	0.185594418236084	3,36E+08
CSF	TCell	CD4	CD4/27+28+31+57-127-244-RA+	-0.608372227014028	0.12739977894979	3,36E+07
CSF	TCell	CD8	CD8/25+38-39-73-PD1-RO+	0.900538792237508	0.188650846789729	3,36E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-XR3-	-0.883050868055305	0.184944924589111	3,37E+08
CSF	TCell	CD4	CD4/38+39-127-DR-RO-	-0.566076856631195	0.118602221159286	3,38E+08
CSF	TCell	CD8	CD8/25+39-127+DR-RO+	0.802198712429052	0.168123527613775	3,40E+07
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-XR3-XR5-	-0.882810988371753	0.184987856019262	3,41E+08
CSF	TCell	CD4	CD4/25-38+73-127+DR-	-0.292227782998949	0.0612831815649459	3,42E+08
CSF	TCell	CD8	CD8/25-DR-	-0.0364913869399958	0.00764890626872658	3,46E+08
CSF	TCell	CD8	CD8/28+RA-R7+	0.549884612360861	0.115359383677272	3,48E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R6-XR3-	-0.883373198198046	0.185301633397836	3,48E+08
CSF	TCell	CD8	CD8/RA-R7+	0.501080830121048	0.105113654876166	3,48E+08
CSF	TCell	CD8	CD8/27+28+RA-R5-	0.592331416404662	0.124341431750712	3,51E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R10-XR3-	-0.881500250407289	0.184988408284205	3,52E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+PD1-R4-R10-XR3-XR5-	-0.881338231735941	0.185010919215009	3,54E+08
CSF	TCell	CD8	CD8/25+73-127+RO-	0.936939107920783	0.196714433085972	3,56E+08
CSF	TCell	CD8	CD8/25+73+127+RO+	0.792334728747962	0.166447726673389	3,59E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R6-R10-XR3-	-0.882548963054134	0.185408755355887	3,60E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-R10-XR3-	-0.627246234342965	0.131850020326532	3,61E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-XR3-	-0.626602967000621	0.131817557497264	3,67E+08
CSF	TCell	CD8	CD8/25+73+DR-PD1-RO+	0.768093911030074	0.161540145984708	3,68E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-XR3-XR5-	-0.634751888939393	0.133538625633541	3,68E+08

CSF	TCell	CD8	CD8/25+38-73-127+PD1-RO-	0.968346139182108	0.203654597931485	3,69E+08
CSF	TCell	CD8	CD8/25+39-73-127+DR-RO+	0.886030879061105	0.186444658948965	3,70E+08
CSF	TCell	CD8	CD8/25+73-RO-	0.880136867356196	0.185156137267238	3,71E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-R10-XR3-XR5-	-0.634094522549264	0.133479501265693	3,73E+08
CSF	TCell	CD8	CD8/25+39-127+DR-PD1-RO+	0.809286119093832	0.17039737735316	3,76E+08
CSF	TCell	CD8	CD8/25+39-73-127+DR-PD1-RO+	0.899058002981471	0.18941061628503	3,79E+08
CSF	TCell	CD8	CD8/25+39-127+RO+	0.795482860453574	0.167585782338113	3,80E+08
CSF	TCell	CD8	CD8/25+39-73-127+RO+	0.879844216038351	0.185448400638884	3,83E+08
CSF	TCell	CD8	CD8/25+39-DR-RO+	0.796447990922899	0.167872299599951	3,84E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R4-R6-XR3-	-0.902985666809013	0.190309030082895	3,84E+08
CSF	TCell	CD4	CD4/38+73-127-DR-PD1-RO-	-0.556937358666414	0.117397172081115	3,85E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R4-R6-XR3-XR5-	-0.903046065326009	0.190347165977066	3,86E+08
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-RO-	0.945279677420507	0.199333805562978	3,90E+08
CSF	TCell	CD4	CD4/25-38+39-DR+RO+	-0.737706905455055	0.155520032183173	3,92E+08
CSF	TCell	CD4	CD4/38+39-73-PD1-	-0.29522916583328	0.0623146017573184	3,93E+08
CSF	TCell	CD4	CD4/38+39-73-127+	-0.294446316496592	0.0621578257188325	3,94E+08
CSF	TCell	CD8	CD8/25+39-73-127+RO-	0.943144917317258	0.198973500996253	3,94E+07
CSF	TCell	CD8	CD8/25+39-73-127+PD1-RO+	0.89372272515136	0.188665887772804	3,96E+08
CSF	TCell	CD4	CD4/38+39-	-0.284773655457708	0.0601286753284384	3,96E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R4-R6-R10-XR3-	-0.901290868573862	0.190334579150402	4,01E+08
CSF	TCell	CD8/Memory	CD8/Memory/161+R4-R6-R10-XR3-XR5-	-0.901403046141497	0.190360058437755	4,01E+08
CSF	TCell	CD8	CD8/25+39-73-DR-RO+	0.879178686562951	0.185764170838111	4,04E+07
CSF	TCell	CD4/Memory	CD4/Memory/161+R6-XR3-	-0.623599734539479	0.131783300374691	4,07E+08
CSF	TCell	CD4	CD4/38+39-127-PD1-RO-	-0.559249643235919	0.118275723654005	4,13E+08
CSF	TCell	CD4	CD4/25-38+39-73-127+	-0.293755230999343	0.0621597177404371	4,14E+08
CSF	TCell	CD4	CD4/25-38+39-73-PD1-	-0.294765117461132	0.0623878448768904	4,17E+08
CSF	TCell	CD4	CD4/25-38+39-	-0.284317779977831	0.0601783158255386	4,17E+08

CSF	TCell	CD4/Memory	CD4/Memory/161+R6-R10-XR3-	-0.622968822683622	0.131839401946959	4,20E+08
CSF	TCell	CD8	CD8/25+39-127+PD1-RO+	0.80328607657231	0.170062522425322	4,22E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3+	0.88477293507831	0.187267816375689	4,23E+08
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-RO+	0.892040950769695	0.188908356916755	4,24E+08
CSF	TCell	CD8	CD8/28+31-57-95+127+244-	0.641712992303291	0.135876931850373	4,24E+08
CSF	TCell	CD8	CD8/25+73+127+PD1-RO+	0.795221243750958	0.168444802783736	4,28E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+XR3+XR5-	0.852862391038929	0.180658282564758	4,29E+08
CSF	TCell	CD8	CD8/25+39-DR-PD1-RO+	0.804075031292645	0.170367073311343	4,29E+08
CSF	TCell	CD8	CD8/25+39-73-RO+	0.8720823078001	0.184796545136947	4,29E+08
CSF	TCell	CD8	CD8/28+31-95+127+244-	0.641126488233354	0.135856632798423	4,31E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3+XR5-	0.951150909127391	0.201539709281352	4,32E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+XR3-XR5-	-0.675818914794435	0.143221976390331	4,35E+08
CSF	TCell	CD8	CD8/25+39-RO+	0.789153256954674	0.167304132273268	4,35E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR3+	0.93547779652782	0.198308216302156	4,37E+08
CSF	TCell	CD8	CD8/25+73+RO+	0.752131789985639	0.159487132738982	4,38E+08
CSF	TCell	CD8	CD8/27-28+31-127+244-	0.732468187965203	0.155404797052904	4,44E+08
CSF	TCell	CD4	CD4/38+127-PD1-RO-	-0.552208215839938	0.11720393768314	4,45E+08
CSF	TCell	CD8	CD8/25+39-73-PD1-RO+	0.885953812015832	0.188075188547273	4,46E+08
CSF	TCell	CD8	CD8/27-28+31-57-127+244-	0.732357656401983	0.155457762763339	4,48E+08
CSF	TCell	CD8	CD8/25+73-DR-PD1-RO-	0.904936318534896	0.192110128160142	4,49E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3+XR5-	0.873132841869018	0.185423631561375	4,52E+07
CSF	TCell	CD4	CD4/25-38+DR-	-0.277233259060166	0.0589311602078586	4,54E+08
CSF	TCell	CD4	CD4/38+DR-	-0.276295860360152	0.0587423478409735	4,56E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-XR3+	0.876877708863957	0.186283224295734	4,56E+08
CSF	TCell	CD8	CD8/27-31-127+244-	0.726600278879728	0.154418590352137	4,59E+08
CSF	TCell	CD8	CD8/31-57-95+127+244-	0.635683950325213	0.135259907719948	4,70E+08
CSF	TCell	CD4	CD4/38+127+PD1-	-0.289403279439817	0.0616025274815656	4,71E+08

CSF	TCell	CD8	CD8/31-95+127+244-	0.634861610413229	0.135117222896161	4,72E+08
CSF	TCell	CD8	CD8/25+73+PD1-RO+	0.757146140016746	0.161185779701555	4,76E+08
CSF	TCell	CD8	CD8/25+38-73-PD1-RO-	0.915280294505155	0.194874701710755	4,78E+08
CSF	TCell	CD8	CD8/25+39-PD1-RO+	0.797614586033831	0.169924924149563	4,81E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R10-XR3-XR5-	-0.674421115442577	0.143639834111662	4,82E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3+	0.915991316376584	0.195049122909115	4,83E+08
CSF	TCell	CD4	CD4/25-38+127+PD1-	-0.289259630583628	0.0616591257278109	4,86E+08
CSF	TCell	CD8	CD8/25+39-73-RO-	0.881944715899092	0.188118148483791	4,96E+08
CSF	TCell	CD4	CD4/38+39-73-DR-PD1-	-0.293289809883539	0.0626309260514654	5,02E+08
CSF	TCell	CD4	CD4/25-38+127-DR-RO-	-0.561544534221784	0.119965215926963	5,09E+08
CSF	TCell	CD8	CD8/27-31-57-127+244-	0.724451705432356	0.154769151609908	5,12E+08
CSF	TCell	CD8	CD8/25+38-127+DR-RO+	0.784196564732202	0.167599228183774	5,14E+08
CSF	TCell	CD8	CD8/27-28+31-57-95+244-	0.696226152839714	0.148793150325515	5,17E+08
CSF	TCell	CD8	CD8/27-28+57-95+244-	0.669193184423518	0.143036367426627	5,18E+08
CSF	TCell	CD4	CD4/38+39-127+DR-	-0.285650716767411	0.0610963305289936	5,21E+08
CSF	TCell	CD4	CD4/38+39-73-127+DR-	-0.291495020417576	0.0623910117866556	5,25E+08
CSF	TCell	CD4	CD4/25-38+39-73-DR-PD1-	-0.29283228601928	0.0626999097913538	5,30E+08
CSF	TCell	CD4	CD4/38+73-127+PD1-	-0.294176897470004	0.0630383469473071	5,38E+08
CSF	TCell	CD4	CD4/25-38+39-127-DR-RO-	-0.56259180543733	0.120521657891025	5,39E+08
CSF	TCell	CD8	CD8/28+31-57-95+244-	0.609765314299467	0.13061940357011	5,40E+08
CSF	TCell	CD8	CD8/27-28+31-95+244-	0.694760792877089	0.148835022207261	5,43E+08
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-RO-	0.904730408806536	0.193845376736622	5,44E+08
CSF	TCell	CD8	CD8/25+38-127+DR-PD1-RO+	0.793545578181148	0.170075506148424	5,45E+08
CSF	TCell	CD4	CD4/25-38+39-127+DR-	-0.285105402107795	0.0611132307536126	5,45E+08
CSF	TCell	CD8	CD8/25-39-DR-	-0.0422701944332904	0.00905664299131233	5,46E+08
CSF	TCell	CD8	CD8/25+73-127+PD1-RO-	0.931966957091435	0.199784697105646	5,50E+08

CSF	TCell	CD4	CD4/25-38+39-73-127+DR-	-0.290838472351589	0.062401696052342	5,52E+08
CSF	TCell	CD4	CD4/38+127+DR-PD1-	-0.28834816157121	0.0618686661856744	5,56E+08
CSF	TCell	CD4	CD4/25-38+73-127+PD1-	-0.293894392276911	0.0630978093591444	5,60E+08
CSF	TCell	CD4	CD4/28+31+127-RA+	-0.561788829607158	0.120577553153726	5,61E+08
CSF	TCell	CD8	CD8/27-28+95+244-	0.667456733318517	0.143219469420783	5,61E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR3+	0.905261798994361	0.194294718709483	5,64E+08
CSF	TCell	CD4	CD4/38+39-73-DR+RO+	-0.73405644263527	0.157522558966425	5,67E+08
CSF	TCell	CD8	CD8/28+31-95+244-	0.608503752379835	0.130660027822403	5,67E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3-	0.545353059553982	0.117093170372907	5,67E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3-XR5-	0.546063798684205	0.11726325492707	5,69E+08
CSF	TCell	CD4	CD4/25-38+127+DR-PD1-	-0.28815907647983	0.0619210453152593	5,73E+08
CSF	TCell	CD8	CD8/27-28+RA-R5-	0.610785771091514	0.131308790890748	5,80E+08
CSF	TCell	CD8	CD8/25+38-73-127+DR-PD1-RO+	0.874321566253698	0.188140451425762	5,90E+08
CSF	TCell	CD8	CD8/25+38-127+RO+	0.776383431468484	0.167153998644484	5,98E+07
CSF	TCell	CD4	CD4/38+39-DR-	-0.28115438801627	0.0605724256676385	6,02E+08
CSF	TCell	CD4	CD4/27+31+57-127-244-RA+	-0.58511389213198	0.125994610483944	6,03E+08
CSF	TCell	CD4	CD4/38+39-73-127-	-0.410280387570961	0.0883498103007178	6,03E+08
CSF	TCell	CD4	CD4/27+31+127-244-RA+	-0.589367198666695	0.126956576624465	6,04E+08
CSF	TCell	CD8	CD8/25+38-73-127+DR-RO+	0.858113714521018	0.184877705461832	6,05E+08
CSF	TCell	CD4	CD4/25-38+73-127-DR-PD1-RO-	-0.546176684173668	0.117686417264439	6,09E+08
CSF	TCell	CD8	CD8/25+38-127+PD1-RO+	0.786105176845228	0.169621796714691	6,25E+08
CSF	TCell	CD8	CD8/25+38-73-127+PD1-RO+	0.868184375237007	0.187417015966573	6,30E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+XR3+	0.846037307538251	0.18258790448396	6,30E+08
CSF	TCell	CD4	CD4/25-38+39-DR-	-0.280715485685758	0.0606279718447527	6,33E+08
CSF	TCell	CD4	CD4/25-38+39-127-PD1-RO-	-0.556124537331034	0.120089618913934	6,35E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3-	0.543820713076585	0.117421499164091	6,36E+08
CSF	TCell	CD8	CD8/25+73-PD1-RO-	0.876221350134994	0.189190510273152	6,37E+08

CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3-XR5-	0.544515518820323	0.117593480660072	6,39E+07
CSF	TCell	CD4	CD4/38+73-127+DR-PD1-	-0.292862074748053	0.0632942375683073	6,41E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3+	0.927100170923094	0.200255935239235	6,44E+08
CSF	TCell	CD8	CD8/27-28+57-95+244-RA-	0.684502441461922	0.147907006172723	6,46E+08
CSF	TCell	CD8	CD8/27-95+244-RA-	0.676224221587029	0.146169048356785	6,49E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4-R10-XR5-	-0.0913358587249732	0.0197366753936218	6,51E+08
CSF	TCell	CD8	CD8/25+38-73-127+RO+	0.851435722039905	0.184182736297324	6,57E+08
CSF	TCell	CD4	CD4/25-38+39-73-127-	-0.412316404476866	0.0891638392376851	6,57E+08
CSF	TCell	CD4	CD4/38+	-0.273026421136211	0.0590984167790019	6,57E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-R10-XR5-	-0.577848381051656	0.125078291804724	6,66E+08
CSF	TCell	CD8	CD8/27-28+57-244-RA-	0.682829108285686	0.147771398522863	6,66E+08
CSF	TCell	CD4	CD4/25-38+73-127+DR-PD1-	-0.292552960740644	0.0633493255382639	6,66E+08
CSF	TCell	CD4	CD4/25-38+	-0.274057026109493	0.0593818249807595	6,71E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4-R10-	-0.0906597095272056	0.0196226893636595	6,73E+08
CSF	TCell	CD8	CD8/27+RA-R5-	0.538329324054449	0.116590966547642	6,75E+07
CSF	TCell	CD4/Memory	CD4/Memory/161+R4+R6-XR5-	-0.574091633173397	0.124350116023992	6,75E+08
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3+	0.865112724197223	0.187377010695812	6,77E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR3-	0.539888975250002	0.1169972081899	6,85E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR3-XR5-	0.540573398393609	0.117160050576731	6,87E+08
CSF	TCell	CD8	CD8/31-57-127+244-RA-	0.676359517313552	0.146666764533356	6,91E+08
CSF	TCell	CD8	CD8/27-244-RA-	0.672699574759585	0.145934772126037	6,98E+07
CSF	TCell	CD8	CD8/28+31-57-127+244-RA-	0.67704957642283	0.146909307092386	7,00E+08
CSF	TCell	CD8	CD8/27-28+95+244-RA-	0.682704843662899	0.148102390983754	7,00E+08
CSF	TCell	CD8	CD8/31-127+244-RA-	0.67576054629294	0.146663393730481	7,03E+08
CSF	TCell	CD8	CD8/28+31-127+244-RA-	0.676793252344695	0.146955819761875	7,09E+08

CSF	TCell	CD4	CD4/38+39-127+PD1-	-0.288199954804456	0.0626157284253357	7,17E+08
CSF	TCell	CD4	CD4/38+127+	-0.28173390462468	0.0612403052596682	7,20E+08
CSF	TCell	CD8	CD8/27-28+244-RA-	0.681035213653397	0.14796020928491	7,21E+08
CSF	TCell	CD4	CD4/25-38+127-PD1-RO-	-0.548465213733419	0.119223836197629	7,26E+08
CSF	TCell	CD4	CD4/38+39-DR+RO+	-0.718849248937054	0.156163566384861	7,27E+07
CSF	TCell	CD8	CD8/27+RA-R7+	0.567469894816719	0.123468868462353	7,38E+08
CSF	TCell	CD4	CD4/25-38+127+	-0.281532103880451	0.0612941266654703	7,44E+08
CSF	TCell	CD4	CD4/25-38+39-127+PD1-	-0.28771439860048	0.0626458945705748	7,49E+08
CSF	TCell	CD8	CD8/27+28+RA-R7+	0.57988131976817	0.126274452880936	7,50E+08
CSF	TCell	CD8	CD8/27-28+31-57-95+244-RA-	0.701791374246654	0.152786907506662	7,52E+06
CSF	TCell	CD4	CD4/25-38+PD1-	-0.28077841787586	0.0611834024362179	7,58E+08
CSF	TCell	CD4	CD4/38+39-127-	-0.404534086344553	0.0881101192209411	7,59E+08
CSF	TCell	CD4	CD4/38+PD1-	-0.279602189355525	0.0609326408371084	7,59E+08
CSF	TCell	CD8	CD8/27-57-95+244-RA-	0.669695352993405	0.145897620608079	7,61E+08
CSF	TCell	CD8	CD8/27-28+31-95+244-RA-	0.701176140057931	0.152801419314178	7,67E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR3-	0.538530713227174	0.117370045898668	7,68E+08
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR3-XR5-	0.539186374540311	0.117537512005173	7,71E+08
CSF	TCell	CD4	CD4/38+39-73-127+PD1-	-0.293131506716074	0.0639951218293209	7,86E+08
CSF	TCell	CD8	CD8/27-28+31-57-244-RA-	0.699869255442155	0.152800872062602	7,96E+08
CSF	TCell	CD8	CD8/27-57-244-RA-	0.666845349032662	0.145666680063276	8,03E+08
CSF	TCell	CD8	CD8/25+73-127+DR-RO+	0.845655760178044	0.184849963465097	8,09E+08
CSF	TCell	CD8	CD8/27-28+31-244-RA-	0.699316569405787	0.152818100684031	8,10E+08
CSF	TCell	CD8	CD8/25+73-127+DR-PD1-RO+	0.860388789950543	0.188119162138475	8,13E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-R10-XR3-XR5-	-0.62972880168866	0.137776026608917	8,23E+08
CSF	TCell	CD4	CD4/25-38+39-73-127+PD1-	-0.292528173755748	0.0640194780340042	8,25E+08
CSF	TCell	CD4	CD4/25-38+39-127-	-0.406568490591058	0.0889300696651832	8,25E+08

CSF	TCell	CD4	CD4/38+39-127+DR-PD1-	-0.286837784061682	0.0627976764801974	8,35E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+XR3-	-0.627126497973648	0.137292436450261	8,38E+08
CSF	TCell	CD8	CD8/25+127+DR-PD1-RO+	0.773772627986985	0.16943794067501	8,40E+08
CSF	TCell	CD8	CD8/25+39-73-PD1-RO-	0.878132225216337	0.192265590337297	8,42E+08
CSF	TCell	CD4	CD4/38+127-DR-PD1-RO-	-0.544069391004914	0.119213794998804	8,50E+08
CSF	TCell	CD4	CD4/25-38+39-127+DR-PD1-	-0.286365349848248	0.0628307397280998	8,72E+08
CSF	TCell	CD8	CD8/31-57-95+127+244-RA-	0.672123490201375	0.14749837276144	8,76E+08
CSF	TCell	CD8	CD8/28+31-57-95+127+244-RA-	0.672496128165532	0.147609474747362	8,80E+08
CSF	TCell	CD8	CD8/25+38-DR-RO+	0.749061574562424	0.164406261268988	8,80E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-XR3-XR5-	-0.626981087043417	0.137696049003046	8,87E+08
CSF	TCell	CD8	CD8/31-95+127+244-RA-	0.6715034044025	0.147498951659853	8,93E+08
CSF	TCell	CD8	CD8/28+31-95+127+244-RA-	0.672215052027648	0.147657123773291	8,93E+08
CSF	TCell	CD8	CD8/25+73-127+RO+	0.837194124342105	0.18396876524473	8,99E+08
CSF	TCell	CD8	CD8/25+73-127+PD1-RO+	0.852824118871792	0.187408418298395	8,99E+08
CSF	TCell	CD4	CD4/38+39-73-127+DR-PD1-	-0.29158169402454	0.0641996193401817	9,29E+08
CSF	TCell	CD8	CD8/25+38-DR-PD1-RO+	0.757410116106156	0.166733075692879	9,33E+08
CSF	TCell	CD8	CD8/27-31-95+244-RA-	0.693460020938906	0.152765926804372	9,48E+08
CSF	TCell	CD4	CD4/38+127+DR-	-0.279405890825915	0.0615873273996459	9,50E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R10-XR3-	-0.62457368515521	0.137669225607547	9,59E+08
CSF	TCell	CD4	CD4/25-38+DR-PD1-	-0.279715242263356	0.061695396897226	9,64E+08
CSF	TCell	CD4	CD4/38+DR-PD1-	-0.278755784292863	0.0614910691475648	9,66E+08
CSF	TCell	CD4	CD4/25-38+39-73-127+DR-PD1-	-0.291003275457607	0.0642265240959468	9,74E+06
CSF	TCell	CD4	CD4/25-38+127+DR-	-0.279153006751967	0.0616384567266858	9,83E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+R6-XR3-XR5-	-0.641672260189832	0.141625147043621	9,84E+07
CSF	TCell	CD8	CD8/31-57-95+244-	0.59061423565307	0.130390566430338	9,88E+08
CSF	TCell	CD8	CD8/25+127+PD1-RO+	0.76467809375932	0.168945034605288	1,00E+09
CSF	TCell	Vg9+/Vd2Low	Vg9+/Vd2Low/8-27-RA+R7-	-0.585748586983405	0.12941763526401	1,00E+09

CSF	TCell	CD4	CD4/38+39-73-127-DR-RO-	-0.523713078732317	0.115740627149374	1,01E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3-	0.605313196601615	0.133819245052623	1,02E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4-R6-XR3-	-0.93974313929646	0.207776879618821	1,02E+09
CSF	TCell	CD8	CD8/25+38-RO+	0.740846465636086	0.163837632750834	1,02E+09
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-RO+	0.834463506457848	0.184608778867602	1,03E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+R6-R10-XR3-XR5-	-0.640445227927305	0.141690894825076	1,03E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3-XR5-	0.605582465798238	0.133985373066496	1,03E+09
CSF	TCell	CD8	CD8/25+38-73-DR-RO+	0.819882483333005	0.181455972879962	1,03E+09
CSF	TCell	CD8	CD8/27-31-57-95+244-RA-	0.688963767430114	0.152461039446485	1,04E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-XR5-	-0.111286958878413	0.0246323004276665	1,04E+09
CSF	TCell	CD8	CD8/27-31-244-RA-	0.689892497277646	0.152757247831767	1,05E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-	-0.110688744129336	0.0245086604810046	1,05E+09
CSF	TCell	CD8	CD8/25+38-PD1-RO+	0.749644389996092	0.166161316454214	1,07E+08
CSF	TCell	CD8	CD8/27-95+244-	0.652021403466806	0.144680163874664	1,09E+09
CSF	TCell	CD8	CD8/27-57-95+244-	0.649743104459437	0.144205973419836	1,10E+09
CSF	TCell	CD8	CD8/57-127+244-RA-	0.616768928682627	0.136934654556212	1,10E+09
CSF	TCell	CD8	CD8/27-31-57-244-RA-	0.685824633368384	0.152395356210529	1,12E+09
CSF	TCell	CD8	CD8/25+38-73-PD1-RO+	0.826382333300675	0.18367629179616	1,12E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3-	0.600200498684927	0.133385386803046	1,13E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3-	0.603970276121216	0.134261163398369	1,13E+09
CSF	TCell	CD8	CD8/31-95+244-	0.586562208821718	0.130417119382603	1,13E+08
CSF	TCell	CD4	CD4/28+31+57-127-RA+	-0.540780616283128	0.120263336196307	1,14E+08
CSF	TCell	CD8	CD8/28+57-127+244-RA-	0.617087246746414	0.137290591980858	1,14E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3-XR5-	0.600437409988637	0.13354173443495	1,14E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4+R6-XR3-XR5-	-0.647813717366098	0.144099940526313	1,15E+09
CSF	TCell	CD8	CD8/25+38-73-RO+	0.811444201869014	0.180541220449085	1,15E+09
CSF	TCell	CD4	CD4/38+39-PD1-	-0.282391947870868	0.0628773070151767	1,16E+09

CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4+R6-R10-XR3-XR5-	-0.647299211136794	0.144119661148584	1,17E+09
CSF	TCell	CD4	CD4/38-127+DR-PD1-	0.165353328219173	0.0368047461597156	1,17E+09
CSF	TCell	CD8	CD8/127+244-RA-	0.614427436969097	0.136887392625491	1,17E+09
CSF	TCell	CD4	CD4/38+39-127+	-0.280785360849799	0.062575152031167	1,18E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4-R6-R10-XR3-	-0.934853388233079	0.208334872745968	1,19E+09
CSF	TCell	CD4	CD4/25-38+39-PD1-	-0.282049254764548	0.0629535065612398	1,21E+09
CSF	TCell	CD8	CD8/27-31+57-127+244-RA-	0.782901177780289	0.174716385328962	1,21E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4-XR5-	-0.0914934212961188	0.0204069458789989	1,22E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3-XR5-	0.602385803828649	0.134424091681927	1,22E+09
CSF	TCell	CD8	CD8/28+127+244-RA-	0.615009990343984	0.137291710301374	1,22E+09
CSF	TCell	CD8	CD8/28+31-57-244-RA-	0.626242986725274	0.139785178636143	1,22E+09
CSF	TCell	CD8	CD8/27-31+57-95+127+244-RA-	0.789529748196633	0.176349866712966	1,23E+09
CSF	TCell	CD8	CD8/27-28+31+57-127+244-RA-	0.7775018035795	0.173704304555932	1,24E+09
CSF	TCell	CD4	CD4/38-39-127+DR-PD1-	0.16886296439605	0.0376971183127298	1,24E+09
CSF	TCell	CD4	CD4/38+39-73-127-PD1-RO-	-0.516649731093333	0.115426336085855	1,25E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3-	0.599148773862084	0.133901672492524	1,25E+09
CSF	TCell	CD4	CD4/38+39-73-PD1+	-0.444735152037835	0.0994072887482714	1,26E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-R10-	-0.56537095511732	0.126462313966209	1,27E+09
CSF	TCell	CD8	CD8/27-28+31+57-95+127+244-RA-	0.782979914744237	0.175158677616279	1,27E+09
CSF	TCell	CD8	CD8/28+31-244-RA-	0.625331398000338	0.139892087696028	1,27E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3-XR5-	0.599353777828311	0.134059517494144	1,28E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-	-0.560440538715859	0.125471066524833	1,29E+09
CSF	TCell	CD4	CD4/38-127+PD1-	0.16362726073002	0.0366024950799024	1,29E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-R10-XR5-	-0.11011923448653	0.024662784838975	1,31E+09

CSF	TCell	CD8/Memory	CD8/Memory/R4-R6-R10-	-0.109569548656713	0.0245399105340952	1,31E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-XR5-	-0.569963244622657	0.127686344646023	1,31E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R4+R6-R10-XR5-	-0.573813290113883	0.128558530637891	1,31E+09
CSF	TCell	CD4	CD4/25-38+39-127+	-0.278841595766825	0.062492426783795	1,31E+09
CSF	TCell	CD4	CD4/38-39-127+PD1-	0.167559892544562	0.0375377496561539	1,33E+09
CSF	TCell	CD8	CD8/27-RA-R5-R7+	0.686711880371661	0.15399499027081	1,33E+09
CSF	TCell	CD8	CD8/27-28+31-57-244-	0.667170756974593	0.149722642702678	1,36E+09
CSF	TCell	CD8	CD8/27-28+31+127+244-RA-	0.772497135563769	0.173452718340774	1,36E+09
CSF	TCell	CD8	CD8/27-31+127+244-RA-	0.776915988660833	0.17445323433681	1,37E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4-	-0.0903049039714309	0.0202685838047164	1,37E+09
CSF	TCell	CD8	CD8/27-31+95+127+244-RA-	0.783651350103075	0.176128109015147	1,39E+09
CSF	TCell	CD4	CD4/25-38+127-DR-PD1-RO-	-0.540369983444431	0.12145155223686	1,39E+09
CSF	TCell	CD8	CD8/25+73-DR-RO+	0.803277107540805	0.180554632644532	1,39E+09
CSF	TCell	CD8	CD8/27-28+31-244-	0.66633328149708	0.149731184490861	1,39E+09
CSF	TCell	CD8	CD8/27-28+31+95+127+244-RA-	0.778127048847168	0.174939091687065	1,40E+09
CSF	TCell	CD8	CD8/57-95+127+244-RA-	0.613216249442422	0.137954960258033	1,41E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R6-XR3-	-0.664677047408425	0.149488282726194	1,41E+09
CSF	TCell	CD4	CD4/38+39-DR-PD1-	-0.28067966838682	0.0631606895521344	1,42E+09
CSF	TCell	CD4	CD4/38+39-73-127-DR-	-0.382501937332189	0.0860561846960728	1,42E+08
CSF	TCell	CD4	CD4/28+31+127-244-RA+	-0.549877680943661	0.123766190672331	1,43E+09
CSF	TCell	CD8	CD8/25+73-DR-PD1-RO+	0.817189368008149	0.183976775358443	1,44E+09
CSF	TCell	CD8	CD8/28+57-95+127+244-RA-	0.613344994351257	0.138127918542905	1,44E+09
CSF	TCell	CD8	CD8/27+28+31-57-95+127+244-	0.612475630844836	0.137975380576096	1,46E+09
CSF	TCell	Vg9+/Vd2Low	Vg9+/Vd2Low/8-RA+R7-	-0.55680107649554	0.12546326073076	1,46E+08
CSF	TCell	CD8	CD8/27+28+31-95+127+244-	0.61186785984242	0.137974168346238	1,48E+09
CSF	TCell	CD4	CD4/25-38+39-DR-PD1-	-0.280333075633445	0.0632355938801867	1,48E+09

CSF	TCell	CD8/Memory	CD8/Memory/161-R4-XR3-	-0.315628236778307	0.0711911657417138	1,49E+09
CSF	TCell	CD8	CD8/95+127+244-RA-	0.610825483488897	0.137914955713268	1,51E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+R4-R6-XR3-	-0.711416831665422	0.160595607239467	1,52E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R6-R10-XR3-	-0.597549484882502	0.134933580126497	1,52E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-XR3-XR5-	-0.31556561131967	0.0712500532102474	1,52E+09
CSF	TCell	CD8	CD8/28+31-57-95+244-RA-	0.621015469496889	0.140259204400274	1,53E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1-R6-XR3-	-0.597390382238878	0.13493564192522	1,53E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R6-R10-XR3-	-0.662420789093629	0.149643802047682	1,54E+09
CSF	TCell	CD8	CD8/28+95+127+244-RA-	0.611220697244796	0.138132460482151	1,54E+09
CSF	TCell	CD8	CD8/27-31-57-95+244-	0.676590121053524	0.152898568707554	1,55E+09
CSF	TCell	DPT	DPT/25+38-39-127+PD1-RO-	0.777900489262663	0.17568899644696	1,56E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3-	0.504072974295536	0.113980999341868	1,57E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3-	0.508070630803031	0.114898466144587	1,57E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3-XR5-	0.50484070340405	0.114181357029934	1,57E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R10-XR3-	-0.314437715100931	0.0711200622421525	1,57E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3-XR5-	0.508822976726997	0.115099805816721	1,58E+09
CSF	TCell	CD8	CD8/25+73-RO+	0.793020848200053	0.179511534768787	1,59E+09
CSF	TCell	CD4	CD4/25-38-39-127+DR-PD1-	0.169602966721333	0.0383644461080093	1,59E+09
CSF	TCell	CD8	CD8/28+31-95+244-RA-	0.620080229609445	0.140364862381649	1,59E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R10-XR3-XR5-	-0.314428458972907	0.071180675879129	1,60E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+R4-R6-R10-XR3-	-0.709883789769698	0.160786736989883	1,61E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R6-XR3-	-0.317088546789733	0.0718385162416111	1,62E+09
CSF	TCell	CD8	CD8/25+73-PD1-RO+	0.807639156649613	0.183035285694445	1,63E+09
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4+R6-XR3-	-0.597284173226932	0.135389162402244	1,63E+09
CSF	TCell	CD8	CD8/27+28+31-57-95+244-	0.586778617900604	0.133057813820088	1,65E+09
CSF	TCell	CD8	CD8/RA-R5-	0.440397516654055	0.0998907207012934	1,66E+09
CSF	TCell	CD8	CD8/25+DR-PD1-RO+	0.729959058904225	0.165595225257383	1,66E+08

CSF	TCell	CD8	CD8/27+31-57-95+127+244-	0.607094811158041	0.137721202618216	1,66E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R6-XR3-XR5-	-0.316950260400218	0.0718961532457046	1,66E+09
CSF	TCell	CD4	CD4/25-38+39-73-127-DR-	-0.383163902956959	0.0869319020170723	1,67E+09
CSF	TCell	CD8	CD8/31-57-244-RA-	0.61147623649194	0.13886545329698	1,69E+09
CSF	TCell	CD4	CD4/25-38-39-127+PD1-	0.168342375714678	0.0382124209861189	1,70E+09
CSF	TCell	Vg9+/Vd2Low	Vg9+/Vd2Low/8-27-RA+	-0.565473367530477	0.128449354986949	1,70E+09
CSF	TCell	CD4	CD4/25-38-127+DR-PD1-	0.165553747208892	0.0375862883403475	1,70E+09
CSF	TCell	CD4	CD4/25-38+39-73-127-DR-RO-	-0.517579998499224	0.117584315045729	1,71E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R6-R10-XR3-	-0.315922792579957	0.0717731460368195	1,71E+09
CSF	TCell	CD8	CD8/27+28+31-95+244-	0.585733136436065	0.133096685831551	1,71E+09
CSF	TCell	CD8	CD8/31-244-RA-	0.610325969967811	0.138720062439016	1,72E+09
CSF	TCell	CD8	CD8/27+31-95+127+244-	0.605806331147103	0.137769445789803	1,74E+09
CSF	TCell	CD8	CD8/28+57-244-RA-	0.563382983998127	0.128174949618689	1,75E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4-R6-R10-XR3-XR5-	-0.315827415972556	0.0718340485518769	1,75E+09
CSF	TCell	CD8	CD8/57-244-RA-	0.543864624816224	0.123758186446187	1,76E+09
CSF	TCell	CD8	CD8/27-31-95+244-	0.67489115632869	0.153576687299033	1,76E+08
CSF	TCell	CD4/Memory	CD4/Memory/161+PD1+R4+R6-R10-XR3-	-0.593989515374895	0.135205503817636	1,76E+09
CSF	TCell	DPT	DPT/25+38-39-127+DR-PD1-RO-	0.778588533717716	0.176985848561972	1,77E+09
CSF	TCell	CD4	CD4/38+39-127-DR-	-0.377101964535693	0.085911587596748	1,80E+09
CSF	TCell	CD8	CD8/57-95+127+244-	0.484546071056682	0.110553505962608	1,85E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+XR3-	0.499416713407172	0.11394420817602	1,85E+08
CSF	TCell	CD8	CD8/244-RA-	0.540772115610992	0.123406295270837	1,85E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+XR3-XR5-	0.500158353170333	0.114136321538171	1,86E+09
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-XR3-	0.503480660208687	0.11490058810425	1,86E+09
CSF	TCell	DPT	DPT/25+38-39-127+RO-	0.752897026360886	0.171647747313721	1,86E+09
CSF	TCell	CD4	CD4/25-38-127+PD1-	0.163921937821929	0.0373981959136074	1,87E+07
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-XR3-XR5-	0.504194740833888	0.115096021376847	1,87E+09

CSF	TCell	DPT	DPT/25+38-39-127+DR-RO-	0.765877305512029	0.174655784986083	1,87E+09

Table S2: Results association between former and never smokers.

type	population	Lineage	Subset.name	Beta	SE	P
CSF	TCell	CD8	CD8/28+31-57-95+127+244-RA-	0.278745318198675	0.076089582890598	0.000289990173956613
CSF	TCell	CD8	CD8/25+DR-	0.293132786481285	0.0790001355581487	0.000242929097271729
CSF	TCell	CD8	CD8/25+38-73-DR-RO-	0.387533766068999	0.0991080108917319	0.000112337309258628
CSF	TCell	CD8	CD8/25+73-DR-	0.377898402361257	0.0885907636751163	2,62E+07
CSF	TCell	CD8	CD8/25+	0.28555966884388	0.0780599854464338	0.000295810921550245
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-RO-	0.380443521629626	0.10235666503027	0.000237689531716643
CSF	TCell	CD8	CD8/27-57-127+244-RA-	0.335671579147545	0.0821732152532263	5,55E+09
CSF	TCell	CD8	CD8/25+127+DR-RO+	0.353952581307035	0.0876459533379949	6,73E+09
CSF	TCell	CD8	CD8/25+38-39-73-DR-	0.411335015342049	0.0919209141949218	1,06E+09
CSF	TCell	CD8	CD8/25+38-39-DR-PD1-	0.320978686696053	0.0830395903472369	0.000133809016790725
CSF	TCell	CD8	CD8/25+39-73-DR-	0.397380312854939	0.0907844274077137	1,63E+09
CSF	TCell	CD8	CD8/25+38-73-127+DR-PD1-	0.382179210571245	0.0946979820671037	6,79E+09
CSF	TCell	CD8	CD8/25+38-39-DR-PD1-RO+	0.390943889228919	0.0894857377657414	1,69E+08
CSF	TCell	CD8	CD8/25+38-39-73-	0.393587015307172	0.0910090719995073	2,04E+09
CSF	TCell	CD8	CD8/25+38-127+DR-PD1-	0.314365160433831	0.0827108580964809	0.000172267070080438
CSF	TCell	CD8	CD8/25+38-39-PD1-RO+	0.387536477599097	0.0892738111629006	1,90E+09
CSF	TCell	CD8	CD8/25+38-39-DR-RO+	0.382112691878776	0.0887672947351997	2,22E+09
CSF	TCell	CD8	CD8/25+73-127+DR-PD1-	0.380102854582068	0.0944168259050092	7,08E+09
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-	0.399618874032802	0.0927296923205947	2,17E+09
CSF	TCell	CD8	CD8/25+39-73-	0.39196786016355	0.0902604846958396	1,89E+09
CSF	TCell	CD8	CD8/25+39-127+DR-PD1-	0.323337750373039	0.082662121264543	0.000111798887692789
CSF	TCell	CD8	CD8/25+38-39-RO+	0.378404669520828	0.0885325858053007	2,53E+09
CSF	TCell	CD8	CD8/25+RO+	0.340339365356262	0.0844699935634209	6,98E+09
CSF	TCell	CD8	CD8/25+73-127+PD1-RO+	0.407817775310901	0.096211711253237	2,95E+09

CSF	TCell	CD8	CD8/25+39-DR-PD1-RO+	0.386292631425019	0.0890665168990061	1,94E+09
CSF	TCell	CD8	CD8/25+38-39-127+DR-PD1-RO+	0.383491327359853	0.0899955782363511	2,67E+09
CSF	TCell	CD8	CD8/25+38-39-73-PD1-	0.394742069005174	0.0924865033809105	2,59E+09
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-	0.396488074633422	0.0922356022838241	2,28E+09
CSF	TCell	CD8	CD8/25+39-DR-RO+	0.378835998795242	0.0883102984766438	2,37E+09
CSF	TCell	CD8	CD8/25+39-127+PD1-	0.319357026135737	0.0824987239792841	0.000131119635469215
CSF	TCell	CD8	CD8/25+38-39-127+DR-RO+	0.375755797940035	0.0892992708720245	3,35E+09
CSF	TCell	CD8	CD8/25+38-39-127+PD1-RO+	0.38033428003461	0.0898291944900718	3,00E+09
CSF	TCell	CD8	CD8/25+39-PD1-RO+	0.382636562519108	0.0888823392454076	2,22E+08
CSF	TCell	CD8	CD8/25+39-73-PD1-	0.391969594719632	0.0917171938699335	2,53E+09
CSF	TCell	CD8	CD8/25+39-RO+	0.374707619706361	0.0880787864602152	2,75E+09
CSF	TCell	CD8	CD8/25+38-39-127+RO+	0.37226467928078	0.0891089659506998	3,80E+09
CSF	TCell	CD8	CD8/28+31-127+244-RA-	0.278562252386548	0.0758133036553949	0.000278293463199169
CSF	TCell	CD8	CD8/25+39-127+DR-PD1-RO+	0.37894009411526	0.0895864653269187	3,05E+08
CSF	TCell	CD8	CD8/25+38-73-PD1-RO+	0.397256372630806	0.0938751049215837	3,03E+09
CSF	TCell	CD8	CD8/25+39-127+DR-RO+	0.373006403795313	0.0888791226811998	3,51E+09
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-	0.399269680524034	0.0941882910444766	2,93E+09
CSF	TCell	CD8	CD8/25+39-127+PD1-RO+	0.375824940617	0.0894406402364969	3,43E+09
CSF	TCell	CD8	CD8/25+38-39-73-DR-RO+	0.433232379179402	0.0964399844863142	9,87E+08
CSF	TCell	CD8	CD8/25+39-127+RO+	0.369251763914585	0.088673244273675	4,02E+09
CSF	TCell	CD8	CD8/25+39-73-127+DR-	0.399217892483648	0.093934531774138	2,80E+09
CSF	TCell	CD8	CD8/25+38-39-73-127+	0.395483042099605	0.0939963068845586	3,35E+09
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-RO+	0.43695390597139	0.0976098467849021	1,06E+09
CSF	TCell	CD8	CD8/25+38-39-73-RO+	0.429456406016891	0.0961370034706978	1,10E+09
CSF	**TCell**	**CD8/Memory**	**CD8/Memory/161-PD1-R4+R10-XR5-**	**0.237784090552813**	**0.0623671819749235**	**0.000165762737267396**
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-RO+	0.430089036720409	0.0968686604456405	1,24E+09
CSF	TCell	CD8	CD8/25+39-73-127+	0.394873872903655	0.0934818752902178	3,12E+09

CSF	TCell	CD8	CD8/25+38-73-DR-	0.380233556952276	0.0892616930303901	2,68E+09
CSF	TCell	CD8	CD8/25+38-39-73-PD1-RO+	0.433766096323371	0.0973279123066759	1,15E+09
CSF	TCell	CD8	CD8/25+39-73-DR-RO+	0.430712484525477	0.0962283462359698	1,06E+09
CSF	TCell	CD8	CD8/25+39-73-127+DR-RO+	0.429744101304089	0.0967110454882319	1,22E+07
CSF	TCell	CD8	CD8/25+39-127+DR-	0.306214335704515	0.0832814652317717	0.000276027782963577
CSF	TCell	CD8	CD8/25+38-39-73-127+RO+	0.426653229011587	0.0966417819532551	1,39E+09
CSF	TCell	CD8	CD8/25+38-127+	0.290900881049089	0.0829917235530565	0.00051994341215947
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-PD1-RO+	0.43291467814435	0.0980779577740258	1,39E+09
CSF	TCell	CD8	CD8/25+38-127+DR-PD1-RO+	0.36565816950219	0.088947981693765	5,01E+09
CSF	TCell	CD8	CD8/25+39-73-RO+	0.426572327549677	0.0959315141697953	1,21E+09
CSF	TCell	CD8	CD8/25+38-39-73-127+DR-PD1-	0.399419500978298	0.0959405298027492	4,03E+09
CSF	TCell	CD8	CD8/25+39-73-127+RO+	0.426069985456882	0.0964512230129704	1,37E+09
CSF	TCell	CD8	CD8/25+38-39-73-127+PD1-RO+	0.430030882617347	0.097871329645451	1,52E+09
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-RO+	0.433116302007592	0.0974716341221707	1,22E+09
CSF	TCell	CD8	CD8/25+38-127+DR-RO+	0.356794276241391	0.0881131731317515	6,44E+09
CSF	TCell	CD8	CD8/25+38-DR-PD1-RO+	0.360806533616091	0.0863601358215841	3,79E+09
CSF	TCell	CD8	CD8/25+38-73-	0.37462392176063	0.088804770386212	3,19E+09
CSF	TCell	CD8	CD8/25+38-39-127+DR-PD1-	0.327717872577294	0.0834495153847349	0.000105077443784683
CSF	TCell	CD8	CD8/25+38-127+PD1-RO+	0.362025038929913	0.088713396383238	5,66E+09
CSF	TCell	CD8	CD8/25+39-127+	0.301835246263803	0.0830968456072177	0.000326086017962754
CSF	TCell	CD8	CD8/25+39-73-127+DR-PD1-RO+	0.430691018908653	0.0979044675107892	1,49E+09
CSF	TCell	CD8	CD8/25+38-39-73-127+PD1-	0.396247727178242	0.0957954483570984	4,50E+09
CSF	TCell	CD8	CD8/25+38-73-127+DR-	0.38105468856522	0.0928219380377049	5,12E+09
CSF	TCell	CD8	CD8/25+39-73-PD1-RO+	0.429878857323661	0.0972178262148367	1,34E+09
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-	0.38179373694898	0.0908925985494678	3,44E+09
CSF	TCell	CD8	CD8/25+39-73-127+DR-PD1-	0.397709263189741	0.0956453194915853	4,11E+08
CSF	TCell	CD8	CD8/25+39-73-127+PD1-RO+	0.428098095320008	0.0977138199800286	1,61E+09

CSF	TCell	CD8	CD8/25+73-DR-RO-	0.369676549080042	0.0961121936515163	0.000144428652337011
CSF	TCell	CD8	CD8/25+38-39-73-DR-RO-	0.371721895216154	0.101820120162104	0.0003047206398027
CSF	TCell	CD8	CD8/25+38-127+RO+	0.353139893871354	0.0878727986865095	7,29E+09
CSF	TCell	CD8	CD8/25+38-DR-RO+	0.351840092696469	0.0855849695206722	5,00E+09
CSF	TCell	CD8	CD8/25+73-	0.371811024013046	0.0878790295833503	3,03E+09
CSF	TCell	CD8	CD8/25+38-PD1-RO+	0.356513481241714	0.0860678618332165	4,40E+09
CSF	TCell	CD8	CD8/25+39-73-127+PD1-	0.394289455084928	0.0952322428270552	4,43E+09
CSF	TCell	CD8	CD8/25+38-73-127+	0.377704091908899	0.0925395228204267	5,64E+09
CSF	TCell	CD8	CD8/25+127+DR-PD1-RO+	0.36112687648041	0.0885052451480954	5,68E+09
CSF	TCell	CD8	CD8/25+38-39-127+PD1-	0.323749230114777	0.0833525529928043	0.000124541008349804
CSF	TCell	CD8	CD8/25+73-127+DR-	0.380473830734469	0.0925707022760324	5,02E+09
CSF	TCell	CD8	CD8/25+38-39-DR-	0.317162118596791	0.0822009266608846	0.000137505858885694
CSF	TCell	CD8	CD8/25+38-73-PD1-	0.376723923806776	0.0904479932535616	3,99E+08
CSF	TCell	CD8	CD8/25+39-73-127+RO-	0.36325500065033	0.102307969738814	0.000441417519411982
CSF	TCell	CD8	CD8/25+73-DR-PD1-	0.378739499221872	0.0902622368749794	3,51E+09
CSF	TCell	CD8	CD8/25+73-127+	0.376284202848312	0.0919888156424484	5,44E+09
CSF	TCell	CD8	CD8/25+39-73-DR-RO-	0.374089612539858	0.097380818556014	0.00014715019630843
CSF	TCell	CD8	CD8/25+38-RO+	0.347302451536898	0.0852813410091295	5,85E+09
CSF	TCell	CD8	CD8/25+127+DR-	0.292666409502048	0.082310582095543	0.000432814384935759
CSF	TCell	CD8	CD8/25+38-39-73-DR-PD1-RO-	0.383653255443218	0.103710280050087	0.000254007268989036
CSF	TCell	CD8	CD8/25+73-DR-PD1-RO-	0.377790491179939	0.0975516876883613	0.000130274831120629
CSF	TCell	CD8	CD8/25+127+PD1-RO+	0.356626540687739	0.0882375136713618	6,65E+09
CSF	TCell	CD8	CD8/25+38-73-127+DR-RO+	0.408619878083737	0.0952803166466924	2,39E+09
CSF	TCell	CD8	CD8/25+73-127+RO-	0.352463984966323	0.101433974477395	0.00058091713151385
CSF	TCell	CD8	CD8/25+39-DR-	0.313482172193645	0.0813834156683301	0.000141046875109545
CSF	TCell	CD8	CD8/25+127+RO+	0.349041424060388	0.0873534267179702	8,00E+09
CSF	TCell	CD8	CD8/25+127+	0.288210591202715	0.08173800696742	0.000482186679147151

CSF	TCell	CD8	CD8/25+DR-PD1-RO+	0.354025681726921	0.0857104171922518	4,62E+09
CSF	TCell	CD8	CD8/25+73-PD1-	0.373318352826824	0.0894758797954103	3,87E+09
CSF	TCell	CD8	CD8/25+39-73-DR-PD1-RO-	0.362099704352422	0.100275777870154	0.00035334269182065
CSF	TCell	CD8	CD8/25+127+DR-PD1-	0.30996596110816	0.0819366156062956	0.00018454134425059
CSF	TCell	CD8	CD8/25+38-73-127+DR-PD1-RO+	0.413003576179162	0.0966493850568855	2,55E+09
CSF	TCell	CD8	CD8/25+38-73-127+RO+	0.405612918844637	0.0950232643558187	2,60E+09
CSF	TCell	CD8	CD8/25+38-73-127+PD1-	0.379244792115111	0.0944325835748134	7,36E+09
CSF	TCell	CD8	CD8/25+38-39-	0.311042992608747	0.0819901905704917	0.000176798975524345
CSF	TCell	CD8	CD8/25+39-DR-PD1-	0.31668217252481	0.0822517462545009	0.000142030136583834
CSF	TCell	CD8	CD8/25+DR-RO+	0.345562748733783	0.0848305833352855	5,83E+07
CSF	TCell	CD8	CD8/25+73-127+DR-RO+	0.407962764734424	0.0951568962841779	2,40E+08
CSF	TCell	CD8	CD8/25+73-RO-	0.356938615295561	0.0944713543789883	0.000188041517879223
CSF	TCell	CD8	CD8/25+38-73-127+PD1-RO-	0.358494896887828	0.104321918855381	0.000666527627348518
CSF	TCell	CD8	CD8/25+39-73-127+DR-RO-	0.359654163664836	0.103423739385005	0.000575334337842343
CSF	TCell	CD8	CD8/25+38-73-127+PD1-RO+	0.410161566733254	0.096385724793796	2,75E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+XR3-XR5-	0.227850938768587	0.059123024615889	0.000141043265461865
CSF	TCell	CD8	CD8/25+38-39-PD1-	0.314873458783043	0.0828552584806657	0.000172349536582159
CSF	TCell	CD8	CD8/25+73-127+PD1-RO-	0.359221412537111	0.102445130920041	0.000518449766805805
CSF	TCell	CD8	CD8/25+38-73-PD1-RO-	0.367300260367971	0.100957583754831	0.000319678538288909
CSF	TCell	CD8	CD8/25+PD1-RO+	0.348973187347142	0.0853444656488893	5,47E+09
CSF	TCell	CD8	CD8/25+73-127+RO+	0.404622007832638	0.0948186545611047	2,61E+09
CSF	TCell	CD8	CD8/25+38-73-RO-	0.373637814978406	0.0976389541513579	0.000155885174369965
CSF	TCell	CD8	CD8/25+39-	0.30794895596511	0.0810546323796052	0.000173039046485413
CSF	TCell	CD8	CD8/25+73-127+PD1-	0.376831562787865	0.0938202613063381	7,35E+09
CSF	TCell	CD8	CD8/25+38-127+PD1-	0.309925156625948	0.0824582225313194	0.000202611523857685
CSF	TCell	CD8	CD8/25+73-127+DR-PD1-RO+	0.410337331041714	0.0965302624264365	2,80E+09
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-XR5-	0.217908799185048	0.0603122326769658	0.000352416408202456

CSF	TCell	CD8	CD8/25+73-PD1-RO-	0.366684189406291	0.0959476566417325	0.000159031463298766
CSF	TCell	CD8	CD8/25+38-39-73-RO-	0.357805020650375	0.101052021878988	0.000457602148672824
CSF	TCell	CD8	CD8/25+39-PD1-	0.311469194452206	0.0819011725657455	0.000170585718645625
CSF	TCell	CD8	CD8/25+38-39-127+DR-	0.309932168193327	0.0839906254400073	0.000262573234310434
CSF	TCell	CD8	CD8/25+39-73-RO-	0.362043259290916	0.0964278144919525	0.00020596658404044
CSF	TCell	CD8	CD8/25+38-DR-	0.295423045737862	0.080194037212891	0.00026863702935438
CSF	TCell	CD8	CD8/25+38-39-73-PD1-RO-	0.370780161175273	0.1030225640544	0.000369634686751909
CSF	TCell	CD8	CD8/25+38-DR-PD1-	0.299666831759494	0.0811238422948874	0.000258627249754622
CSF	TCell	CD8	CD8/25+DR-PD1-	0.298200945837427	0.0798437931249313	0.00022155302326712
CSF	TCell	CD8	CD8/25+39-73-PD1-RO-	0.351339036001714	0.099424331627591	0.000469639972784133
CSF	TCell	CD8	CD8/25+38-39-127+	0.305762493175662	0.0838748532755584	0.000310456195485029
CSF	TCell	CD8	CD8/25+38-73-DR-RO+	0.396093154347649	0.0929309718746254	2,66E+09
CSF	TCell	CD8	CD8/25+38-73-DR-PD1-RO+	0.400622509416898	0.0942727878161757	2,81E+09
CSF	TCell	CD8	CD8/28+RA-R5-	0.23519788161347	0.0601238669581785	0.000111229790010046
CSF	TCell	CD8	CD8/25+38-73-RO+	0.392221176853298	0.0925388481766542	2,94E+09
CSF	TCell	CD8	CD8/25+73-DR-RO+	0.392801716211748	0.0924270712044186	2,81E+09
CSF	TCell	CD8	CD8/25+PD1-	0.290651404334756	0.0788522798725753	0.000266474210029417
CSF	TCell	CD8	CD8/25+38-	0.286907028484823	0.0795257591039303	0.000356969259670449
CSF	TCell	CD8	CD8/25+38-PD1-	0.290895620499674	0.0804668381584889	0.000347358330120463
CSF	TCell	CD8	CD8/25+73-DR-PD1-RO+	0.396754722517125	0.0938972929313765	3,11E+09
CSF	TCell	CD8	CD8/25+73-RO+	0.388648976044978	0.0920280569233537	3,14E+09
CSF	TCell	CD8	CD8/25+38-127+DR-	0.295505711906742	0.0832339005745218	0.000441237492087381
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-	0.236685237586389	0.0619509581567708	0.000160785434865312
CSF	TCell	CD8	CD8/25+73-PD1-RO+	0.392892597615874	0.0934702151894288	3,41E+07
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-	0.237679673812421	0.0622500666438236	0.000162317855032488
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR5-	0.238881529848315	0.0626895908017968	0.000167157178067579

CSF	TCell	CD8	CD8/27-28+95+127+244-	0.2930285804675	0.0807642366641372	0.00033088479698702
CSF	TCell	CD8	CD8/25+127+PD1-	0.292153278459253	0.0827246497649223	0.000472467470177301
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3-XR5-	0.212076950813267	0.0584928068986644	0.000335973627207591
CSF	TCell	CD8	CD8/27-28+31-57-95+127+244-	0.302673668771309	0.0850092998039438	0.000424799828576031
CSF	TCell	CD8	CD8/28+31-95+244-RA-	0.253562488293705	0.0716354220122403	0.000458508701740687
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3-XR5-	0.251617070005681	0.0671270266059413	0.000211872145078373
CSF	TCell	CD8	CD8/27-28+57-127+244-RA-	0.339305467152066	0.0828849064683119	5,35E+09
CSF	TCell	CD8	CD8/27-28+127+244-RA-	0.339147413245567	0.0828937900003072	5,40E+09
CSF	TCell	CD8	CD8/27-28+57-95+127+244-RA-	0.339736747164748	0.0829962446973802	5,36E+09
CSF	TCell	CD8	CD8/27-57-95+127+244-RA-	0.337136888292945	0.0823843189171225	5,38E+09
CSF	TCell	CD8	CD8/27-28+95+127+244-RA-	0.339614721304884	0.0830075742664519	5,40E+09
CSF	TCell	CD8	CD8/27-127+244-RA-	0.334325683197326	0.0821343735348872	5,89E+09
CSF	TCell	CD8	CD8/27+28+RA-R5-	0.238095477602917	0.0629211157763659	0.000183330757605724
CSF	TCell	CD8	CD8/27-95+127+244-RA-	0.335755340842675	0.0823487592778147	5,72E+09
CSF	TCell	CD8	CD8/27-57-244-RA-	0.295449864700574	0.0780408171312687	0.000182140746841996
CSF	TCell	CD8	CD8/25+38-39-73+DR-PD1-RO+	0.315743670142306	0.0910959579656413	0.000598787193609205
CSF	TCell	CD8	CD8/31-57-244-RA-	0.249568621357787	0.0713840810068003	0.000536833912458826
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3-	0.228058525741234	0.0589762552623699	0.000134021410105103
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R10-XR3-XR5-	0.228388256013285	0.0590533841234034	0.000133742282803404
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3-	0.272342461236349	0.0660148276785125	4,75E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R10-XR3-XR5-	0.272581823344974	0.0660895715810247	4,77E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3-	0.228081324795589	0.0592024572783923	0.000141786157163993
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-R10-XR3-XR5-	0.22841298508658	0.0592808092809632	0.000141508429102688
CSF	TCell	CD8	CD8/27-28+31-127+244-RA-	0.347136020380909	0.0864797275183238	7,40E+09

CSF	TCell	CD8	CD8/27-28+31-57-127+244-RA-	0.347280564066345	0.0865070615361225	7,39E+09
CSF	TCell	CD8	CD8/27-31-57-127+244-RA-	0.343923172280449	0.0856049170220356	7,30E+09
CSF	TCell	CD8	CD8/27-28+31-95+127+244-RA-	0.34744130243568	0.0866066620113579	7,47E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3-	0.272658478626267	0.0662772309802011	4,99E+09
CSF	TCell	CD8	CD8/27-28+31-57-95+127+244-RA-	0.347580241142917	0.0866339914135202	7,46E+09
CSF	TCell	CD8	CD8/27-31-127+244-RA-	0.342470363669051	0.0855165146421426	7,68E+09
CSF	TCell	CD8	CD8/27-31-57-95+127+244-RA-	0.345461180779119	0.0858772802542323	7,14E+09
CSF	TCell	CD8	CD8/27-31-95+127+244-RA-	0.343929780931566	0.08578675270207	7,55E+09
CSF	TCell	CD8/Memory	CD8/Memory/161-R4+R6-XR3-XR5-	0.227941402034391	0.0593604112032933	0.000148900210798042
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-	0.218861718981512	0.0607025898155526	0.0003625917863001
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R10-	0.216842129451697	0.0598956572821485	0.000342972731277596
CSF	TCell	CD8/Memory	CD8/Memory/PD1-R4+R6-R10-XR5-	0.22013317148599	0.0611500526240417	0.000370159918631619
CSF	TCell	CD8	CD8/27-31-57-244-RA-	0.299956000690311	0.0820461133661361	0.000298340505566998
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-R10-XR3-XR5-	0.266707355711362	0.0660344064353267	6,77E+09
CSF	TCell	CD8	CD8/57-127+244-RA-	0.254171905592357	0.0690586965169997	0.000272022245437664
CSF	TCell	CD8	CD8/28+57-127+244-RA-	0.255380221782188	0.0693829700786588	0.000271780107334116
CSF	TCell	CD8	CD8/127+244-RA-	0.253356137051372	0.0690200738738917	0.000282199248774462
CSF	TCell	CD8	CD8/31-57-127+244-RA-	0.277726754323089	0.0754984398134608	0.000273887398520211
CSF	TCell	CD8	CD8/28+127+244-RA-	0.254827043781106	0.0693600468445181	0.000278788982835651
CSF	TCell	CD8	CD8/28+31-57-127+244-RA-	0.278665237683621	0.0757959832764595	0.000276001641352761
CSF	TCell	CD8	CD8/31-127+244-RA-	0.277222822203209	0.075497547887184	0.000280803078076166
CSF	TCell	CD8	CD8/57-95+127+244-RA-	0.25443373511037	0.0695204572155357	0.000294036621249573
CSF	TCell	CD8	CD8/28+57-95+127+244-RA-	0.255289520097331	0.0697579258211721	0.000294229190084854
CSF	TCell	CD8	CD8/31-57-95+127+244-RA-	0.278237910715729	0.07587558593237	0.000286058666726883
CSF	TCell	CD8	CD8/28+31-95+127+244-RA-	0.278650924647768	0.0761075554436472	0.000292289086028161

CSF	TCell	CD8	CD8/31-95+127+244-RA-	0.277749455268155	0.0758757683166848	0.00029305951824171
CSF	TCell	CD8	CD8/28+95+127+244-RA-	0.254753378579883	0.0697352482739328	0.000301487434668192
CSF	TCell	CD8	CD8/95+127+244-RA-	0.2536397211415	0.0694832643857418	0.000304654287940729
CSF	TCell	CD8	CD8/27-28+57-95+244-RA-	0.303465995984382	0.0774431215734546	0.000108551914928104
CSF	TCell	CD8	CD8/27-28+57-244-RA-	0.301918410401332	0.0773236642319564	0.000114712868744883
CSF	TCell	CD8	CD8/27-28+57-95+127+244-	0.293472704222359	0.0807889886241182	0.000325518892613148
CSF	TCell	CD8	CD8/27-28+95+244-RA-	0.302579002924615	0.0774844175225359	0.00011452465434
CSF	TCell	CD8	CD8/27-28+244-RA-	0.300994189314693	0.0773621861561523	0.000121175161697351
CSF	TCell	CD8	CD8/27-95+127+244-	0.281661558932	0.0794942154323901	0.000452611883401293
CSF	TCell	CD8	CD8/28+57-244-RA-	0.221919325558451	0.0626359985529317	0.000453823221663268
CSF	TCell	CD8	CD8/27-57-95+127+244-	0.283180561759044	0.0798296544253044	0.000445834015623252
CSF	TCell	CD8	CD8/28+31-57-244-RA-	0.253648638713774	0.0713754637300969	0.000435453173861295
CSF	TCell	CD8	CD8/28+31-244-RA-	0.253477003441149	0.0714144478384348	0.000442427750653502
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3-	0.25145962421783	0.0667099988931211	0.000195465633568647
CSF	TCell	CD8	CD8/27-28+31+57-95+127+244-RA-	0.332239628384667	0.0892710026161369	0.000232805675720702
CSF	TCell	CD8	CD8/27-31+57-95+127+244-RA-	0.331314628097994	0.089428191911737	0.000248190086582769
CSF	TCell	CD8/Memory	CD8/Memory/R4+R6-R10-XR3-	0.21166901317377	0.0584019973630943	0.000337659551687414
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+XR3-XR5-	0.251263140456531	0.0668285194423926	0.000202784327508795
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3-XR5-	0.209688359124308	0.0579560782848201	0.000345529311731505
CSF	TCell	CD8/Memory	CD8/Memory/R4+R10-XR3-	0.209272511862065	0.0578682605732275	0.000347707407060729
CSF	TCell	CD8	CD8/28+31-57-95+244-RA-	0.253747351735464	0.0715969525154651	0.000451065137777552
CSF	TCell	CD8	CD8/27-28+31-95+127+244-	0.302000632587414	0.0849496868050112	0.000433329302732511
CSF	TCell	CD8/Memory	CD8/Memory/161-PD1-R4+R6-XR3-	0.251797754764545	0.0670103091584696	0.000204555946115537
CSF	TCell	CD8	CD8/27-28+31+95+127+244-RA-	0.330379505992015	0.0891494735753591	0.000247194729968091
CSF	TCell	CD8	CD8/27-31-95+127+244-	0.291777458832829	0.0837676404721652	0.000562529115937762
CSF	TCell	CD8	CD8/27-28+31+57-127+244-RA-	0.326879249472751	0.0889125257600592	0.000276255067263027
CSF	TCell	CD8	CD8/27-31+95+127+244-RA-	0.329027829125281	0.0893055324508725	0.000268202667827306

CSF	TCell	CD8	CD8/57-244-RA-	0.224052810927006	0.0632148920991587	0.000450884593389549
CSF	TCell	CD8	CD8/27-31+57-127+244-RA-	0.325101349371343	0.0889953196600443	0.000301593921507969
CSF	TCell	CD8	CD8/27-31-57-95+127+244-	0.293427867994343	0.0842007892779265	0.000559137969508776
CSF	TCell	CD8	CD8/27-28+31-57-95+244-RA-	0.305118866692978	0.0805025915058724	0.000179261956810158
CSF	TCell	CD8	CD8/27-28+31+127+244-RA-	0.324910732387852	0.0887768054045281	0.000293992830914167
CSF	TCell	CD8	CD8/27-28+31-95+244-RA-	0.30432668855018	0.080494912866605	0.000185929233460572
CSF	TCell	CD8	CD8/27-28+31-57-244-RA-	0.303958667767241	0.080423042387741	0.000186787227081857
CSF	TCell	CD8	CD8/27-28+31-244-RA-	0.303178335536875	0.080415885364018	0.000193634820801385
CSF	TCell	CD8	CD8/27-31+127+244-RA-	0.322708287459471	0.0888501675331401	0.000326170468123883
CSF	TCell	CD8	CD8/244-RA-	0.22049800039032	0.0632111222332629	0.000552462244788044
CSF	TCell	CD8	CD8/27-57-95+244-RA-	0.296564698894205	0.0783083432405379	0.000181225438968555
CSF	TCell	CD8	CD8/27-95+244-RA-	0.290355615621107	0.0786848526333962	0.000262298439038791
CSF	TCell	CD8	CD8/27-244-RA-	0.289040920124657	0.0784172713371291	0.000266463279705433
CSF	TCell	CD8	CD8/27-28+57-95+244-	0.270933266113995	0.0762658430737978	0.000437795641755694
CSF	TCell	CD8	CD8/31-244-RA-	0.245600578802634	0.0715182215884788	0.000670530173510451
CSF	TCell	CD8	CD8/27-28+95+244-	0.269642756156568	0.0762762520951059	0.000466520673127083
CSF	TCell	CD8	CD8/27-31-57-95+244-RA-	0.300404073819999	0.0823161374228102	0.00030573266945151
CSF	TCell	CD8	CD8/27-28+31-57-95+244-	0.273930888411814	0.0797101878089142	0.000665121893818441
CSF	TCell	CD8	CD8/27-31-95+244-RA-	0.293344928911927	0.0827258573357139	0.00044795557471446
CSF	TCell	CD8	CD8/27-31-244-RA-	0.292676627232421	0.082483801952582	0.000444208119334088

Table S3: Purities of sorted cells.

Sample	N°monocytes	Purity	N° Neutrophils	Purity	N° NK cell	Purity	N° B cell	Purity	N°CD4+	Purity	N° CD8+	Purity
Non_smoker1	366715	99,70%	6,0*106	99,40%	1190410	95,20%	335127	97,10%	3,0*106	99,40%	2,0*106	99,60%
Non_smoke2	805663	99,50%	3836878	99,30%	1603518	91,10%	552704	96,70%	3027294	99%	1088507	99%
Smoker 1	939934	99,90%	6857618	99,30%	1238652	97,25%	1198578	97,10%	5413880	99,30%	1766318	98,90%
Smoker 2	773145	95,80%	8315211	99,20%	2756138	97,60%	501156	96,25%	1905684	99,70%	2280389	99,50%
Smoker 3	1046481	99,60%	7877924	98,20%	2345683	96,70%	756103	99,40%	2917351	99,80%	1262729	99,80%
Non_smoker 3	482894	99,30%	7619338	99,60%	1056676	99,45%	767856	97,45%	4277608	99,60%	1711703	97,80%
mean	735805,3333	98,97%	6901393,8	99,17%	1698512,833	96,22%	685254	97,33%	3508363,4	99,47%	1621929,2	99,10%
min	366715	95,80%	3836878	98,20%	1056676	91,10%	335127	96,25%	1905684	99,00%	1088507	97,80%
max	1046481	99,90%	8315211	99,60%	2756138	99,45%	1198578	99,40%	5413880	99,80%	2280389	99,80%

Table S4: Results of functional and pathways enrichment

HYPOMETHYLATED GENES IN B CELLS	p-value	Adj.p-value	Combined Score	Genes
WIKY PATHWAYS				
Degradation pathway of sphingolipids, including diseases WP4153	0.03	1.0	131.41	GLB1
Pyrimidine metabolism and related diseases WP4225	0.03	1.0	131.41	DPYS
T-Cell antigen Receptor (TCR) Signaling Pathway WP69	0.03	1.0	25.34	VAV3;SKAP1
REACTOME				
Glycosphingolipid metabolism Homo sapiens R-HSA-1660662	0.01	1.0	79.47	GALC;GLB1
EPH-ephrin mediated repulsion of cells Homo sapiens R-HSA-3928665	0.01	1.0	63.71	VAV3;EPHA3
Neurofascin interactions Homo sapiens R-HSA-447043	0.02	1.0	18.06	CNTN1
Sphingolipid metabolism Homo sapiens R-HSA-428157	0.02	1.0	34.04	GALC;GLB1
Type I hemidesmosome assembly Homo sapiens R-HSA-446107	0.03	1.0	13.14	CD151
IRAK2 mediated activation of TAK1 complex upon TLR7/8 or 9 stimulation	0.03	1.0	11.49	IRAK2
IRAK2 mediated activation of TAK1 complex Homo sapiens R-HSA-937042	0.03	1.0	11.50	IRAK2
Reactions specific to the complex N-glycan synthesis pathway Homo sapiens R-HSA-975578	0.03	1.0	11.51	MAN2A1
Urea cycle Homo sapiens R-HSA-70635	0.03	1.0	11.52	CPS1
Transport of organic anions Homo sapiens R-HSA-879518	0.03	1.0	10.16	SLCO2A1

Table S4: Results of functional and pathways enrichment

KEGG				
Lysosome	3.66 x10-6	0.01	136.14	MFSD8;GALC;LAPTM4B;GLB1;AP3B1
Sphingolipid metabolism	9.06x10-3	1.0	65.63	GALC;GLB1
HYPERMETHYLATED GENES IN B CELLS				
WIKI PATHWAYS				
Histone Modifications WP2369	0.04	1.0	20.16	SETBP1;SMYD3
REACTOME				
Defective B3GALTL causes Peters-plus syndrome (PpS) Homo sapiens R-HSA-5083635	7.18x10-4	1.0	123.58	SEMA5A;ADAMTS17;THSD7A
O-glycosylation of TSR domain-containing proteins Homo sapiens R-HSA-5173214	7.76x10-4	0.6	119.02	SEMA5A;ADAMTS17;THSD7A
Diseases associated with O-glycosylation of proteins Homo sapiens R-HSA-3906995	3.21x10-3	1.0	58.50	SEMA5A;ADAMTS17;THSD7A
cGMP effects Homo sapiens R-HSA-418457	3.25x10-3	1.0	13.40	KCNMB2;PDE5A
Nitric oxide stimulates guanylate cyclase Homo sapiens R-HSA-392154	6.24x10-3	1.0	85.51	KCNMB2;PDE5A
Gap junction trafficking Homo sapiens R-HSA-190828	7.79x10-3	1.0	73.01	DAB2;GJB3
Diseases of glycosylation Homo sapiens R-HSA-3781865	8.52x10-3	1.0	34.21	SEMA5A;ADAMTS17;THSD7A
Gap junction trafficking and regulation Homo sapiens R-HSA-157858	8.91x10-3	1.0	66.26	DAB2;GJB3
Neuronal System Homo sapiens R-HSA-112316	0.01	1.0	14.83	GNAL;RPS6KA2;KCNMB2;KCNK17;CACNG3

Table S4: Results of functional and pathways enrichment

O-linked glycosylation Homo sapiens R-HSA-5173105	0.01	1.0	23.91	SEMA5A;ADAMTS17;THSD7A
KEGG				
cGMP-PKG signaling pathway	0.04	1.0	11.83	KCNU1;KCNMB2;PDE5A
Adherens junction	0.05	1.0	18.00	FER;PTPRM
HYPOMETHYLATED GENES IN MONOCYTES				
WIKI PATHWAYS				
Thermogenesis WP4321	3.31x10-4	0.16	97.30	RHEB;PRDM16;ADCY3;PPARG
Differentiation of white and brown adipocyte WP2895	2.63x10-3	0.62	155.89	PRDM16;PPARG
Regulation of Actin Cytoskeleton WP51	0.01	1.0	29.67	VIL1;PDGFRA;MYLK
HIF1A and PPARG regulation of glycolysis WP2456	0.02	1.0	152.61	PPARG
FTO Obesity Variant Mechanism WP3407	0.02	1.0	152.61	PRDM16
Pathways in clear cell renal cell carcinoma WP4018	0.03	1.0	27.64	PDGFRA;RHEB
NAD metabolism, sirtuins and aging WP3630	0.03	1.0	101.63	PPARG
Transcriptional cascade regulating adipogenesis WP4211	0.04	1.0	81.86	PPARG
Osteoblast Signaling WP322	0.04	1.0	74.31	PDGFRA
REACTOME				
Netrin-1 signaling Homo sapiens R-HSA-373752	7.28x10-3	1.0	76.85	ABLIM3;TRPC4
Choline catabolism Homo sapiens R-HSA-6798163	0.02	1.0	219.04	DMGDH
CREB phosphorylation through the activation of Adenylate Cyclase Homo sapiens R-HSA-442720	0.02	1.0	180.59	ADCY3
Organic cation transport Homo sapiens R-HSA-549127	0.03	1.0	131.42	SLC22A4
Role of second messengers in netrin-1 signaling Homo sapiens R-HSA-418890	0.03	1.0	114.87	TRPC4

Table S4: Results of functional and pathways enrichment

Adenylate cyclase activating pathway Homo sapiens R-HSA-170660	0.03	1.0	114.87	ADCY3
Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	0.04	1.0	21.21	SLC22A4;SLC14A2
DCC mediated attractive signaling Homo sapiens R-HSA-418885	0.04	1.0	74.31	ABLIM3
Organic cation/anion/zwitterion transport Homo sapiens R-HSA-549132	0.04	1.0	74.31	SLC22A4
SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion Homo sapiens R-HSA-399955	0.04	1.0	74.31	FARP2
KEGG				
Choline metabolism in cancer	3.44x10-3	1.0	56.36	SLC22A4;PDGFRA;RHEB
Longevity regulating pathway	3.74x10-3	0.58	53.89	RHEB;ADCY3;PPARG
Thermogenesis	5.41x10-3	0.56	29.63	RHEB;PRDM16;ADCY3;PPARG
Phospholipase D signaling pathway	0.01	0.80	30.31	PDGFRA;RHEB;ADCY3
Calcium signaling pathway	0.02	1.0	20.53	PDGFRA;ADCY3;MYLK
Gastric acid secretion	0.02	1.0	33.36	ADCY3;MYLK
Rap1 signaling pathway	0.03	1.0	17.61	FARP2;PDGFRA;ADCY3
Gap junction	0.03	1.0	26.23	PDGFRA;ADCY3
Human cytomegalovirus infection	0.03	1.0	15.13	PDGFRA;RHEB;ADCY3
Dilated cardiomyopathy (DCM)	0.03	1.0	24.92	SGCD;ADCY3
HYPERMETHYLATED GENES IN MONOCYTES				
WIKI PATHWAYS				
miRNA targets in ECM and membrane receptors WP2911	1.97x10-3	0.93	188.83	COL5A1;ITGA1
Regulation of Actin Cytoskeleton WP51	0.01	1.0	30.47	ITGA1;PIK3C3;FGFR2

Table S4: Results of functional and pathways enrichment

AMP-activated Protein Kinase (AMPK) Signaling WP1403	0.02	1.0	38.68	LEPR;PIK3C3
Hair Follicle Development: Cytodifferentiation (Part 3 of 3) WP2840	0.03	1.0	27.36	CUX1;GLI2
Leptin and adiponectin WP3934	0.03	1.0	117.33	LEPR
Hedgehog Signaling Pathway WP47	0.05	1.0	63.72	GLI2
NOTCH1 regulation of human endothelial cell calcification WP3413	0.05	1.0	58.81	ITGA1
miR-509-3p alteration of YAP1/ECM axis WP3967	0.05	1.0	58.81	COL5A1
Leptin Insulin Overlap WP3935	0.05	1.0	58.81	LEPR
REACTOME				
L1CAM interactions Homo sapiens R-HSA-373760	3.01x10-3	1.0	60.48	ITGA1;NRCAM;SCN1A
Interaction between L1 and Ankyrins Homo sapiens R-HSA-445095	3.41x10-3	1.0	13.06	NRCAM;SCN1A
Axon guidance Homo sapiens R-HSA-422475	4.37x10-3	1.0	21.10	ARHGEF28;ITGA1;NRCAM;SLIT3;FGFR2;SCN1A
Phospholipid metabolism Homo sapiens R-HSA-1483257	0.01	1.0	27.20	LPCAT2;PIK3C3;SYNJ2
PI Metabolism Homo sapiens R-HSA-1483255	0.01	1.0	47.42	PIK3C3;SYNJ2
Collagen biosynthesis and modifying enzymes Homo sapiens R-HSA-1650814	0.02	1.0	44.18	COL17A1;COL5A1
Signaling by FGFR in disease Homo sapiens R-HSA-1226099	0.02	1.0	44.18	CUX1;FGFR2
Intra-Golgi and retrograde Golgi-to-ER traffic Homo sapiens R-HSA-6811442	0.02	1.0	22.89	SCOC;CUX1;KLC1
GLI proteins bind promoters of Hh responsive genes to promote transcription	0.02	1.0	18.44	GLI2
Neurofascin interactions Homo sapiens R-HSA-447043	0.02	1.0	18.44	NRCAM

Table S4: Results of functional and pathways enrichment

KEGG				
Inositol phosphate metabolism	0.02	1.0	34.88	PIK3C3;SYNJ2
Protein digestion and absorption	0.03	1.0	25.99	COL17A1;COL5A1
Phosphatidylinositol signaling system	0.04	1.0	22.45	PIK3C3;SYNJ2
Endocytosis	0.04	1.0	13.52	WIPF1;NEDD4L;FGFR2
GENES IN COMMON BETWEEN MONOCYTES AND B CELLS				
WIKI PATHWAY				
IL-5 Signaling Pathway WP127	5.85x10-4	0.28	413.57	SPRED1;JAK2
Pyrimidine metabolism and related diseases WP4225	8.07x10-3	1.0	594.98	DPYD
ncRNAs involved in STAT3 signaling in hepatocellular carcinoma WP4337	0.01	1.0	380.62	JAK2
Leptin Insulin Overlap WP3935	0.02	1.0	273.64	JAK2
The human immune response to tuberculosis WP4197	0.02	1.0	187.78	JAK2
EPO Receptor Signaling WP581	0.02	1.0	160.92	JAK2
PDGFR-beta pathway WP3972	0.03	1.0	140.14	JAK2
IL17 signaling pathway WP2112	0.03	1.0	128.74	JAK2
Mammary gland development pathway - Pregnancy and lactation (Stage 3 of 4) WP2817	0.03	1.0	123.63	JAK2
Fluoropyrimidine Activity WP1601	0.03	1.0	118.86	DPYD
REACTOME				
Interferon Signaling Homo sapiens R-HSA-913531	6.79x10-4	1.0	124.07	TRIM62;JAK2;IFNA21
VEGFR2 mediated cell proliferation Homo sapiens R-HSA-5218921	1.34x10-3	1.0	88.92	SPRED1;ITPR2;JAK2

Table S4: Results of functional and pathways enrichment

Immune System Homo sapiens R-HSA-168256	1.74x10-3	0.89	27.36	TRIM62;SPRED1;ITPR2;TPP2;JAK2;IFNA21
Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215	1.98x10-3	0.76	44.63	TRIM62;SPRED1;JAK2;IFNA21
VEGFA-VEGFR2 Pathway Homo sapiens R-HSA-4420097	2.77x10-3	0.85	61.34	SPRED1;ITPR2;JAK2
Signaling by VEGF Homo sapiens R-HSA-194138	2.97x10-3	0.76	59.14	SPRED1;ITPR2;JAK2
Downstream signaling of activated FGFR2 Homo sapiens R-HSA-5654696	3.0x10-3	0.65	58.87	SPRED1;ITPR2;JAK2
Downstream signaling of activated FGFR4 Homo sapiens R-HSA-5654716	3.0x10-3	0.57	58.87	SPRED1;ITPR2;JAK2
Downstream signaling of activated FGFR3 Homo sapiens R-HSA-5654708	3.0x10-3	0.51	58.87	SPRED1;ITPR2;JAK2
Signaling by FGFR4 Homo sapiens R-HSA-5654743	3.0x10-3	0.47	58.07	SPRED1;ITPR2;JAK2
KEGG				
Kaposi sarcoma-associated herpesvirus infection	5.83x10-4	0.18	133.47	ITPR2;JAK2;IFNA21
Cholinergic synapse	4.48x10-3	0.69	107.28	ITPR2;JAK2
JAK-STAT signaling pathway	9.16x10-3	0.94	64.37	JAK2;IFNA21
Necroptosis	9.16x10-3	0.71	64.37	JAK2;IFNA21
Hepatitis B	9.27x10-3	0.57	63.82	JAK2;IFNA21
Influenza A	0.01	0.52	59.64	JAK2;IFNA21
NOD-like receptor signaling pathway	0.01	0.48	56.33	ITPR2;IFNA21
Tuberculosis	0.01	0.43	55.88	JAK2;IFNA21
Human immunodeficiency virus 1 infection	0.02	0.52	43.81	ITPR2;IFNA21
Pantothenate and CoA biosynthesis	0.02	0.52	238.38	DPYD
GOrilla results: Methylated genes in B cells	P-value	FDR q-value	Enrichment	Genes

Table S4: Results of functional and pathways enrichment

GO:0051271 negative regulation of cellular component movement	3.24E-4	7.12E-1	4.12	SEMA5A; C16orf45;SEMA3A;MCC;NAV3;PTPRM

Figure S5: Antibodies Panel used for cell-sorting. In first row are reported the seven lasers to separate the cell-type by expressed fluorescence.

Tube	Cell-types	FITC	PE	PrCP-Cy5.5	PE-Cy7	APC	A700	PE-CF594
1	LYMPHOCITE (B,NK,CD4,CD8) 4 WAY	CD16+CD56 (NK cells)		CD19 (B cells)	CD14 (Monocytes)	CD3 (T cells)	CD4+	CD8+
2	MONO/GRANULO (Mono, NK, B, Neu) 4 WAY	CD16+CD56 (NK cells)	CD3 (T cells)	CD19 (B cells)	CD14+ (Monocytes)	CD11b+ (myeloid)		

Target bisulfite sequencing pipeline

The pipeline includes the common four steps of target sequencing analysis: *a)* pre-processing NGS reads, *b)* aligning reads to a reference genome, *c)* filtering duplicated reads, *d)* methylation calling.

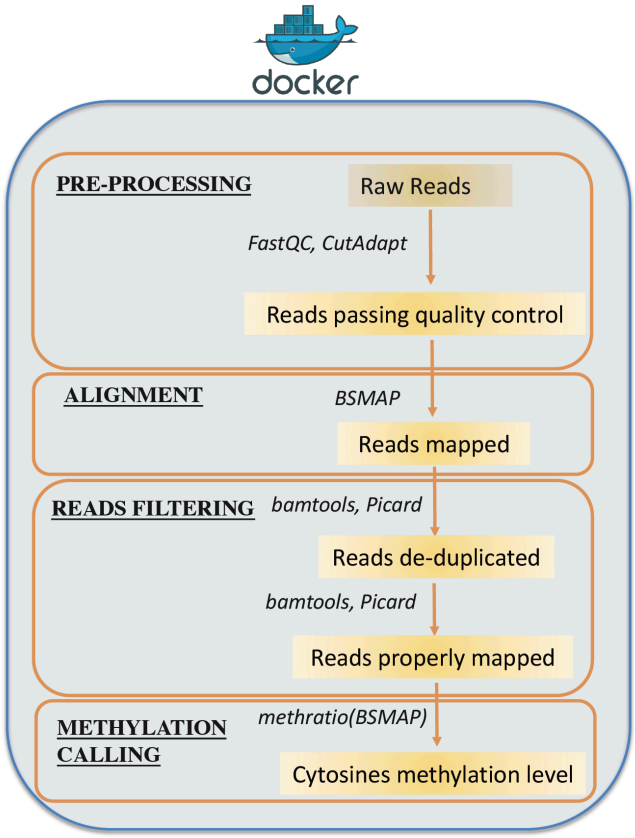


Figure: Workflow of analysis pipeline.
In each block are detailed the algorithms applied to analyse data

Pre-processing step

Sequencing reads are quality controlled using FastQC(1). It allows to evaluate the overall library quality, presence of sequencing adapters and presence of bad

quality bases in the 3'-end of the read. Low quality bases and adapter were removed by CutAdapt (2).

Alignment

Reads passed quality control step, are mapped to reference genome using BSMAP (version 2.90)(3).

Filtering steps

This step includes splitting reads to remove duplicates and filtering the properly mapped paired-end reads. During splitting step, four BAM files are produced. Each of them includes aligned reads on a specific bisulfite strand (bisulfite top and bottom strand and their reverse complements). BAM file are merged in two files, that contain reads aligning on top and bottom strand, respectively. These files are then sorted, duplicates are removed and then, they are merged again, in order to obtain a single BAM file. These steps are performed using bamtools (4) and Picard (5) packages. Filtering for keep only mapped and properly paired reads and to clip overlapping reads are performed using bamtools. Furthermore, Picard and bamtools packages on filtered and clipped reads allow to estimate several metrics for evaluating the quality of experiments and analysis. A brief summary of these metrics is provided as follow.

- *Mapping metrics* provide details about the quality of the read alignments and the proportion of the reads that passed machine signal-to-noise threshold quality filters.
- *Hybrid selection analysis metrics* provide a number of metrics assessing the quality of the targeted bisulfite sequencing experiment.
- *The number of on-target reads*, namely the number of reads that overlap one of the target regions by at least one base.
- *The depth of coverage* over the target regions.
- The *insert size distribution* of the reads. DNA is randomly fragmented, and later size selected. So, it is normal to observe a range of fragment sizes. If this range is too large or too small, the number of on-target reads can be adversely affected.

Methylation calling analysis

The percent methylation at each cytosine in the samples is determinate using *methratio* package supplied by BSMAP tools. It creates a tab-delimited file that contains a line for each detected cytosine. Fields of the line specify the cytosine position, the number of times it appears as methylated and unmethylated and its ß-score. *Methratio* is also used to calculate the bisulfite conversion efficiency.

Docker

Docker (6) is an open source project that allows operating system level virtualization, portable deployment of containers across platforms, and git-like versioning, among others. Docker technology has recently become very popular throughout the scientific community and not just. It allows to run applications in an isolated environment and to efficiently distribute the package, in the form of Docker images, in a portable manner across different platforms. In a similar way to virtual machines Docker images provides all the required software is already installed, configured and tested.

Virtual machines and Docker containers are similar in their goals. They provide analysis portability, isolating an application into a self-contained unit that can run anywhere. They provide analysis reproducibility, freezing the version of tools and library used.

References

1. Andrews, S., FASTQC. A quality control tool for high throughput sequence data. *2014.*

2. Martin et al. *https://doi.org/10.14806/ej.17.1.200.* 2011.

3. Xi, Y. and Li, W. BSMAP: whole genome bisul te sequence MAPping program. *BMC Bioinformatics,* 2009, Vol. 10.

4. Barnett, D. W., et al. BamTools: a C++ API and toolkit for analyzing and managing BAM files, *Bioinformatics*, Volume 27, Issue 12, 15 June 2011, Pages 1691–1692, https://doi.org/10.1093/bioinformatics/btr174

5.Picard toolkit. http://broadinstitute.github.io/picard/

6.Boettiger, C. An introduction to Docker for reproducible research. ACM SIGOPS *Operating Systems Review*, 49(1):71,2015.