LUND UNIVERSITY

Statistical Modeling and Learning of the Environmental and Genetic Drivers of Variation in Human Immunity

Bergstedt, Jacob

2018

Document Version:
Publisher's PDF, also known as Version of record

Statistical Modeling and Learning of the Environmental and Genetic drivers of Variation in Human Immunity

Jacob Bergstedt

LUND
UNIVERSITY

Department of Automatic Control

PhD thesis TFRT-1121
ISBN 978-91-7753-908-7 (print)
ISBN 978-91-7753-909-4 (web)
ISSN 0280–5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

Printed in Sweden by Media-Tryck.
Lund 2018

Abstract

During the last decade the variation in the human genome has been mapped in fine detail. Next generation sequencing has made it possible to cheaply and rapidly aquire vast amounts of biomolecular information on large cohorts of people. This have enabled large-scale epidemiological studies to investigate the relationships between environmental and genetic factors and human biomolecular traits. It is now possible to map variation in the genomic blueprint for human biology to variation in levels of epigenomic marks, gene expression levels and protein expression levels. This development has opened up the possibility of a "phenomic science": the data-driven study of the interactions between all levels of the relationship between the genotype, the environment, and the phenotype.

The Milieu Intérieur study of Institut Pasteur, Paris, aims at bringing the technological developments of modern biology to bear on the study of the human immune system in homeostasis. Deep phenotyping has been performed on 1,000 healthy, unrelated people of Western European ancestry. The cohort is evenly stratified across sex, and across five decades of life, between 20 and 70 years of age. In this thesis, we combine the standardised flow cytometry of 173 parameters of innate and adaptive immune cells, genome-wide DNA genotyping, detailed information on life-style and environmental factors and MethylationEPIC array data of the Milieu Intérieur cohort, to identify the genetic and environmental drivers of variation in the human immune system.

The increasing complexity of biological data requires the development of new statistical tools. In this work, we aim to integrate developments in machine learning, convex optimization, causal inference, and statistical methodology, to build robust and reliable tools for analysing the high-dimensional and highly complex biomolecular data of the Milieu Intérieur study. We construct a pipeline to perform genome-wide association studies on phenotypes with heterogenous distributions, while controlling for arbitrarily many environmental factors. The pipeline is applied to study the genetics of human immune system variation in homeostasis and the genetics of the function of the human thymus.

Our pipeline identifies 15 loci that influence immunophenotypes. We show that these loci are enriched in disease-associated variants. We also report a common

genetic variant, situated within the T cell receptor locus, that increases the production of naive T cells within the human thymus. In addition, we find four key non-genetic factors that drive variation in the healthy human immune system: age, sex, latent cytomegalovirus infection and smoking. Age, sex, and smoking have a broad impact on the innate and the adaptive immune subsystems, while cytomegalovirus infection primarily seems to skew the T cell compartment of the adaptive immune subsystem towards inflammatory subsets. We also show that age and sex influence the function of the human thymus.

Immunophenotypes are intimately connected to epigenetic markers in whole-blood. We leverage the >850,000 methylation sites probed in the MethylationEPIC array to build high-dimensional predictive models of 70 immune cell subsets and other traits such as age and smoking status. We employ elastic net regression and stability selection to build sparse, regularized models, and show that they are capable of estimating blood cell composition more accurately and cost-effectively than previous methods. The properties of elastic net regression and stability selection also enable us to investigate the relationship between DNA methylation and immune blood cell composition.

This thesis develops methods for, and performs, the analysis of parts of the rich and multifaceted data of the Milieu Intérieur study. With the construction and analysis of this rich observational data we contribute to the young fields of population immunology and human phenomic science. We discover novel associations that will help in understanding the differences between people in vaccination efficacy and susceptibility to common autoimmune and infectious disesases. Finally, we present predictive models that will facilitate the application of immunological markers in the clinics.

Acknowledgements

This thesis has been ironed out in the crucibles of biomedical research. It has been said that sometimes a chess-board is like a dark jungle, grave danger lurking in every move. On a similar theme, the universe has been described as a dark forest[1] Sometimes PhD work can feel much the same. Luckily, there was always a couple of guiding lights in the darkness. My supervisor Bo Bernhardsson offered invaluable support. I was more or less lost in the forest, trapped within an old willow tree, when Etienne Patin, who would become my co-supervisor and main collaborator for the thesis, offered a path forward. Pontus Giselsson, another co-supervisor, was an inspiration, and he always had time and energy for discussion. Magnus Fontes, a third co-supervisor introduced me to the Pasteur Institute, threw amazing research camps and was, and is, an inspiration. I want to say huge thanks also to my other close collaborator, Cécile Alanio who has taught me a lot about working in research and has always been a major inspiration.

Of course, no quests through murky forests can be completed without a little bit of camaraderie. Therefore I want to thank in particular the other amigos, Rasmus Henningsson and Gabriel Illanes. Thanks also to Björn Olofsson, Fredrik Magnusson, Jerker Nordh, Kerstin Johnsson, Carolina Bergeling and Anders Mannesson. A small but important reason I was able to see my way out of the forest was weekly floorball practice, together with, among others, Mattias Fält and Gautham Nayak Seetanadi.

I want to thank everyone at the Human Evolutionary Genetics laboratory at Institut Pasteur for being wonderful, in particular Lluis for giving me the opportunity to work on Milieu Intérieur. I am also grateful to Matthew Albert for that opportunity.

I also want to thank everyone at the Automatic Control department in Lund for being wonderful. Especially, I want to thank Eva Westin, Mika Nishimura, Monika

[1] "The universe is a dark forest. Every civilization is an armed hunter stalking through the trees like a ghost, gently pushing aside branches that block the path and trying to tread without sound. Even breathing is done with care. The hunter has to be careful, because everywhere in the forest are stealthy hunters like him. If he finds another lifeanother hunter, angel, or a demon, a delicate infant to tottering old man, a fairy or demigodtheres only one thing he can do: open fire and eliminate them." Cixin Liu, *The Dark Forest*.

Rasmusson, Ingrid Nilsson, Cecilia Edelborg, Anders Nilsson, Anders Blomdell, Leif Andersson and Pontus Andersson for keeping the ship afloat.

I want to thank my parents and my sister. I also want to thank my brothers, sister and parents-in-law. Finally, I want to thank Annika Bergstedt, my Galadriel, to whom I dedicate this work.

Contents

Financial support

I am a member of the LCCC Linnaeus Center and the ELLIIT Excellence Center at Lund University and I am supported by the ELLIIT Excellence Center.

Preface

Personal contributions

This thesis presents and analyses data from the Milieu Intérieur project based at Institut Pasteur, Paris. The project investigates drivers of variation in the healthy human immune response. It has three main objectives. First, it aims to produce a rich set of data on the immune system, the genome, the transcriptome, the methylome, and life-style of 1,000 healthy French subjects, evenly stratified across both sex and age. Next, it aims to develop methods capable of analysing such rich and complicated data. Finally, it aims at analysing the data, and present a detailed and precise view of the biomolecular and intrinsic and extrinsic factors that drive the variation in the human immune system.

My own work has dealt mostly with the last two aims of the study. In particular, I have developed a pipeline to perform genome-wide association studies (GWAS) for data of rich complexity. This pipeline is used to study the genetics of the human immune system. I have also designed and conducted a number of analyses to investigate what and how non-genetic factors influence it. To get a broader view of what affects immune system composition, I have developed models to investigate the relative importance of genetic and non-genetic factors on immune system variation. Finally, I have used methods from convex optimization and machine learning to develop predictive models of immune cell variation from high-dimensional DNA methylation data. This work is described in the five chapters of the thesis and in three papers where I am first co-author. Here, I will briefly go through each chapter and paper in turn, and outline my contribution in each.

Chapter 1 - Introduction

The introduction puts the work in perspective, not only from a biological and computational viewpoint, but also in light of the immense technological progress our society is currently experiencing.

Chapter 2 - Statistical methods

The second chapter reviews theory from the intersection of statistics, convex optimization, causal inference and machine learning needed to understand the methods developed and used in the papers. It also uses novel ideas and concepts from the field of causal inference to shed some light on two ubiquitous problems in computational biology: analysing gene expression in whole blood and the problem of population stratification in GWAS.

Chapter 3 - Milieu Intérieur

Chapter 3 very briefly introduces the cells and mechanisms of the human immune system. It then develops some previously unpublished visualizations to showcase the rich data of Milieu Intérieur.

Chapter 4 - GWAS pipeline

Here, I develop the GWAS pipeline that is used in Paper I, to study the genetics of circulating human immune cells and proteins, and in Paper II, to study the genetics of the function of the human thymus. The pipeline leverages the wealth of the non-genetic data of the Milieu Intérieur study to find mutations that impact human immune cells and proteins *independently* of non-genetic factors. To do this, the pipeline selects controls from the database of life-style variables using a technique called stability selection, which has not previously been used in this setting. The design of the selection procedure is guided by new ideas and techniques from the young field of causal inference. The pipeline also includes a novel way to transform the response variable in a data-driven but robust fashion. In summary, a number of individual steps are composed into a pipeline capable of conducting hundreds of genome-wide association studies in a precise and robust manner.

Chapter 5 - Exploring environmental determinants of immune system variation

The fifth chapter outlines the methodology developed to study the non-genetic variation of the immune system. It uses new ideas and concepts from the recently burgeoning field of causal inference to design models that adjust for confounding and batch variables in a principled manner.

Paper I

Patin, E.[*], M. Hasan[*], J. Bergstedt[*], V. Rouilly et al. (2018). "Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors". Nature Immunology 19:3, pp. 302314.
[*]Contributed equally. See Paper I for full author list

In this paper we perform an exhaustive study of the factors that influence the immune system at homeostasis. We look for associations between each of 166 im-

munophenotypes with over 5 million single nucleotide polymorphisms. By developing a novel model selection scheme, leveraging our rich database of intrinsic, demographic and socioeconomic variables, we identify 15 loci that affect immunophenotypes *independently* of environmental context. These loci are showed to be enriched in disease-associated variants. Furthermore, we also investigate non-genetic variables that influence the immune system in healthy humans of Western European ancestry. Using our large sample size of 1,000 observations, and the rich database of non-genetic variables, we are able to conclusively show a broad impact on the immune system of age, sex, latent cytomegalovirus infection and smoking, with unprecedented scope and detail. Finally, we build models for each immunophenotype that investigate the relative importance of the genetic and non-genetic variables that were associated with it.

For this paper, I cleaned and organized the non-genetic data together with EP, VR and BP. I organized and processed the immunophenotype data, together with EP. The study design, in terms of what analyses and datasets to include, was decided by EP and myself, under supervision of LQM and MLA. I developed the GWAS pipeline, under supervision of EP, and conducted the GWAS analyses, together with EP. The analysis of the impact of non-genetic variables on the immunophenotypes was designed and conducted by me. The joint analysis of genetic and non-genetic variables that were associated with an immunophenotype was designed and conducted by myself. I designed and created all, or parts of, Figures 2, 3 and 5. Finally, I contributed in writing several sections of the manuscript.

Paper II

Clave, E.*, I. L. Araujo*, C. Alanio*, E. Patin*, J. Bergstedt*, et al. (2018). "Human thymopoiesis is influenced by a common genetic variant within the TCRA-TCRD locus". Science Translational Medicine 10:457.
*Contributed equally. See Paper I for full author list

Here, we explore the function of the human thymus, measured by T cell receptor excision circles (TRECs), a proxy of T cell thymic production. We evaluate how well this surrogate measure captures thymic function by correlating it with various other measures related to the thymus, such as counts of naive T cells. To study the genetics of thymic function independently of environmental influences, we use the GWAS pipeline developed in Paper I and Chapter 4, together with our dataset of over 5 million SNPs. We find a variant *within the T cell receptor locus* that increases the output of the thymus. To evaluate the variant, we perform experiments that show that the variant influences early steps in T cell development in the mouse thymus, as well as T cell receptor gene segment allocation. Finally, we evaluate the influence of non-genetic variables on the thymus and report a large impact of age and sex.

I designed and conducted most of the statistical analyses of this paper, including the analysis of association between TRECs and immunophenotypes, the GWAS on TRECs, the impact of the variant on T cell development and T cell receptor gene

segment allocation in the mouse thymus and the impact of non-genetic variables on TRECs, under supervision of EC, EP and AT. The GWAS was done together with EP. I designed and created all of, or most of, Figures 1, 2, 5, and 6. Finally, I contributed in writing the manuscript.

Paper III

Bergstedt, J, A. Urrutia, D. Duffy, M. L. Albert, L. Quintana-Murci, and E. Patin (2018). "Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles". bioRxiv.

In this paper, we explore the prediction of immunophenotypes using DNA methylation. In doing so, we are also able to investigate the relationship between DNA methylation and blood cell variation. We develop novel regularized regression models that can predict immunophenotypes with high accuracy. The models we develop are also used to accurately predict other factors from DNA methylation, particularly age, cytomegalovirus infection and smoking.

I came up with the idea of predicting immunophenotypes from DNA methylation using regularized regression. I developed all statistical methodology of this paper, conducted all analyses, and produced all tables and figures and wrote the manuscript, under supervision of EP.

1

Introduction

Out of the many ways the world is currently transforming, the meteoric rise of the use of data might prove to be the most upheaving. The potency of the transformation is felt at all levels of everyday life, from eerily on-point targeted online ads and search query completions, to changes in the political *status quo*. As an example, Cambridge Analytica, involved in the marketing of Donald Trump's election campaign in 2016 and the Brexit campaign, supposedly has collected over 5,000 data points on each of 230 million Americans, for political information campaigns [Cadwalladr, 2018]. The usage of data is also transforming business by creating new markets and drastically changing old ones. The transformative power of data analysis has made the British magazine *The Economist* declare that "the world's most valuable resource is no longer oil, but data" [The Economist, 2017]. As Erik Brynjolfsson and Andrew Mcafee write in the American magazine *The Atlantic*: "Mobile phones, automobiles, factory automation systems and other devices are routinely instrumented to generate streams of data on their activities, making possible an emerging field of "reality mining" to analyse this information. Manufacturers and retailers use radio-frequency identification (RFID) tags to deliver terabits of data on inventories and supplier interactions and then feed this information into analytical models to optimize and reinvent their business processes" [Brynjolfsson and Mcafee, 2011]. To remain competitive, companies have started to base decisions on copious amounts of internally produced data.

Data analysis is also having a profound impact in the policy domain and in public safety. In economics for instance, Liran Einav and Jonathan Levin write in a review in *Science* that the profession has "shifted from a reliance on relatively small-sample government surveys to administrative data with universal or near-universal population coverage" [Einav and Levin, 2014]. In policy, the increasing use of data in decision making is referred to as the "digital transformation", and it is making governments all over the world scramble to form new ministries and action plans to realize its potential [The Swedish Government, 2017]. Meanwhile, law enforcement agencies are collecting and analysing ever vaster amounts of surveillance data, and prediction software is used to find areas at risk for future crime, to set bail, and for sentencing recommendations [Fasman, 2018].

Naturally, data is also transforming the biological and medical sciences. In the 2015 State of the Union address, then US president Barack Obama launched the precision medicine initiative, a project made possible by emerging biomedical data technologies [Collins and Varmus, 2015]. One of the collaborators in the project, Euan A. Ashley, writes in *Nature Reviews Genetics* that precision medicine aims at "understanding disease at a deeper level in order to develop more targeted therapy" [Ashley, 2016]. To do this, vast amounts of data produced by powerful next generation sequencing technologies are analysed to perform deep phenotyping of diseases [Delude, 2015]. Cheap array genotyping platforms have enabled thousands of genome-wide association studies (GWAS) to map common variation in the human genome to various pathologies. The wealth of information created by such studies has started to be used for polygenic risk prediction in clinical application and for personal health management [Torkamani et al., 2018]. The possibility of collecting and analysing ever larger datasets has stimulated the creation of large consortia where collaborative groups of researchers and funding agencies are working together to collect and analyse data to investigate, for example, genetic variation, such as in the 1,000 Genomes Project [The 1000 Genomes Project, 2015] and the Genotype-Tissue Expression (GTEx) project [GTEx Consortium, 2017].

The possibilities engendered by the data revolution in medicine and biology have given rise to the notion that humans should be the ultimate model organisms to study human disease [FitzGerald et al., 2018]. This development is neatly illustrated in the field of immunology. The field has gained tremendous success at elucidating biological mechanisms from experiments on mice. However, the mouse model is not always suitable for illuminating how the immune system functions in a human setting. Many clinical protocols derived from striking results obtained with mice experiments have failed in clinical practice [Davis and Brodin, 2018]. This has led to the birth and vigour of the field of population immunology, where large amounts of biomolecular data and immunophenotyping on cohorts of humans are used to understand and delineate the human immune response *in natura* [Quintana-Murci et al., 2007; Liston et al., 2016]. Two large recent, notable projects with exactly this emphasis is the Human Functional Genomics Project (HFGP) [Netea et al., 2016] and the Milieu Intérieur project [Thomas et al., 2015], the last of which is the topic of this thesis.

A main driver of the rise of data in society, science and technology is the increasing access to ever cheaper and stronger computing power. In parallel, open source software for data collection, management and analysis has become better and easier to use; it is no longer necessary to have a specialized degree to do sophisticated data analysis and prediction modelling. The R and python programming languages and their package management systems has made state-of-the-art machine learning and statistical software readily available to anyone with technical proficiency. There has been tremendous development in this area the last few years, where package universes like the *tidyverse* marks a paradigm shift in the design and theory of data analysis software [RStudio, 2018].

The growth of data collection has naturally been followed by development of algorithms and models suitable for complex and abundant data. These developments are often centered on the concept of regularization. The standard linear model, which has been the workhorse of statistical modelling for the last century, constrains the relationship between the variables in the data it models in a very restrictive way. New models try to constrain variables more flexibly, which allows capturing for instance complex, high-dimensional and nonlinear dependencies. This has led to immense success. For instance, the deep learning model has made it possible for computers to solve tasks that were previously restricted to humans, like object recognition and text and speech processing [Goodfellow et al., 2016].

Deep learning works well for highly nonlinear problems with massive amounts of samples, something that is often incongruous to investigations in biology. Other algorithms more suitable to the relatively less data-rich fields of biology include the random forest model or gradient boosting models. But even these models are often too flexible to work well in biological settings, which are typically characterized by low signal-to-noise ratios and many more variables than samples. Deep learning and its cousins are sometimes referred to as black box models because the parametrization of their regularization can be difficult to interpret. This makes them unsuitable for inference and knowledge generation, for instance in the biological sciences. Instead, the main tools for explicit modelling of complex biological systems are linear models whose parameters are assumed to come from statistical distributions. Such distributions are known as priors in the terminology of Bayesian statistics. These include the mixed effects [Gelman and Hill, 2006] and the elastic net models [Zou and Hastie, 2005]. The priors allow them to model complex relationships between variables while retaining interpretability in terms of the underlying biology.

More observational studies and massive increases in the amount of data, computing power and available, easy to use and powerful statistical software packages have made biological and biomedical research increasingly methodologically challenging. Richer data offer richer opportunities for finding novel relationships. However, with improper methodology it also increases the risk of finding false positives. During the last decade, there has been a lot of focus in the scientific community on the notion of replicability. Scientists at the drug company Amgen in the US tried to reproduce 53 landmark studies in cancer biology and managed to get similar results as the original study in only six cases [Begley and Ellis, 2012]. Out of 1,576 researchers asked the question "Is there a reproducibility crisis" by the journal *Nature*, 52% answered "yes, a significant crisis" and 38% answered "Yes, a slight crisis" [Baker, 2016]. Highlighting the cross-disciplinary nature of modern data-driven research, the number one remedy offered by queried researchers was a "better understanding of statistics". The increasing focus on reproducibility has made the renowned statistican Andrew Gelman talk of a "replication revolution" as a scientific revolution similar in essence, but not in magnitude, to the Darwinian revolution in biology and the quantum revolution in physics [Gelman, 2018].

One of the main culprits in causing false positives is *confounding*: misinterpreta-

tion of regression parameters due to missing variables and incorrect modeling. The study of human systems *in natura* usually requires relying on observational studies. Both HFGP and Milieu Interieur, mentioned previousy, are examples of such studies. Given the rich and complex data collected, there is ample risk for spurious correlational evidence. Luckily, the last decades have seen a breakthrough in the area of causal inference. There have been two parallel developments in this field: the potential outcome framework [Imbens and Rubin, 2015] and the causal structural equations framework [Pearl, 2009]. In Milieu Interieur, we have used the framework of the causal structural equations model [Pearl, 2009]. It puts the task of analysing a complex observational study on a firm theoretical foundation by offering a way to use a conceptual biological model of the system under investigation to understand how to control for spurious correlation. In doing that, the framework clearly and systematically communicates the assumptions put on the variables for the results to be not only correlational, but also causal. Such reasoning is common in epidemiology, and is increasingly getting traction in the social sciences and economics, but is still mostly absent in observational studies in the biological sciences.

Another issue that has been highlighted the last decade is the utility of hypothesis testing and *P* values. Some journals in the social and psychological sciences have even taken the drastic step of banning *P* values altogether [Trafimow and Marks, 2015; Gill, 2018]. Hypothesis testing has been the cornerstone of science throughout the 20th century and has been thought to embody the Semmelweisian ideal of inductive reasoning [Noakes et al., 2008]. However, modern data-rich studies often violate the conditions under which hypothesis testing is appropriate. In particular, such studies often test hypotheses that has been generated by the data itself, so-called selective inference. For valid inference in such settings, the process that generated the hypotheses must be taken into account in the statistical models, something that is often difficult and rarely done. In addition, the null hypothesis is often a "straw man" [Recht, 2018]. It is usually stated in terms of an effect size of a variable under investigation being zero. However, improper control for confounding renders such a hypothesis irrelevant to the underlying research question. Even if there is good control of confounding, it is still impossible to control for all of it, so effect sizes are never exactly zero. That means that the null hypothesis can be rejected as long as enough data is collected. A field that suffers from overreliance on *P* values is genomics where they are the output of standardized genome-wide association studies (GWAS). For instance, the UK Biobank cohort includes over 500,000 individuals. It has been used to study, among other conditions, the genetics of psychiatric diseases [Ward et al., 2017]. The cohort is so large that mutations with small enough effect sizes to put their biological relevance in question could still easily be significant.

In this thesis, we heed the call to study the immune system *in natura* [Quintana-Murci et al., 2007; Davis, 2008]. In doing so, we are adding to the young fields of population immunology [Liston et al., 2016] and human phenomic science [FitzGerald et al., 2018]. Our contribution is the creation and analyses of the rich and multifaceted data of the Milieu Intérieur project, an observational study involving deep

phenotyping of 1,000 healthy individuals. We think that our results will contribute to the development of personalized and precision medicine. Our aim is to integrate the state-of-the-art of study design in population immunology and genomics with new concepts in causal inference and epidemiology, new developments in statistical modelling and machine learning, and new paradigms and ideas in general statistical methodology, while still keeping the biological research question in focus at all-times. Therefore, we have interpretability as the central tenent in our statistical design. All our models are based on causal graphs that clearly communicate our view and understanding of the biological systems we investigate. We use statistical models that capture the correlational structure of our samples, while still being as simple and easy to interpret as possible. We try to use measures that directly relate to a relevant biological phenomenon. We use the effect size as the primary statistical measure. When we use hypothesis testing we aim to put the results in a biological context using effect sizes, confidence intervals, and biological annotations. We use machine learning algorithms mainly to select controls in linear models and for prediction in models where we have a lot more variables than samples. In particular, we use the elastic net model coupled with a stability selection scheme for this purpose, to ensure that we have interpretable and robust models.

The thesis includes two studies of immune function for healthy humans *in natura*. In Paper I: *"Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors"*, we take a broad picture and investigate genetic, intrinsic and environmental determinants of the main actors of the immune system: the immune cell subsets. In Paper II: *"Human thymopoiesis is influenced by a common genetic variant within the TCRA-TCRD locus"*, we zero in on the function of the thymus from a population immunology perspective. One of the main aims of population immunology is to find immunological prognostic markers for clinical use [Davis, 2008]. Our first two studies contribute to the understanding of what markers could be useful. In our final study, Paper III: *"Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles"*, we aim to facilitate the use of immunological markers in the clinics by providing predictive models capable of accurately estimating blood cell composition using only a few stable, cheap and readily acquired methylation probes. Our use of interpretable predictive models also allow us to shed some light on the epigenetic control of immune cell subsets.

2

Statistical methods

In his book "statistical rethinking", Richard McElreath makes a distinction between the small and the large world of the statistical model [McElreath, 2018]. According to him, "The small world is the self-contained logical world of the model", and "The large world is the broader context in which one deploys a model". With some simplification: the small world is the mathematics of the model, while the large world are the assumptions and requirements for it to describe the system it is supposed to represent. In the first part of this chapter, we focus on the "small world" methodology of inference and regression. In the second part, we consider the "large world" methodology of causal inference.

2.1 Notation

First we introduce some notation. Note that this notation is for the introduction, the notation in the papers might deviate from it. Stochastic variables are in upper case, Y, and observations, y, are in lower case. Vectors and matrices are bold. The index i is reserved for indexing individuals. The index j is reserved for indexing *variables* related to individuals. For a predictor matrix $\boldsymbol{x} \in \mathbb{R}^{n \times p}$, $\boldsymbol{x}_j \in \mathbb{R}^n$ is the vector with all values of the jth variable and x_{ij} is the value for the ith individual and the jth variable.

2.2 Inference

We start with observations $\boldsymbol{y} \in \mathbb{R}^{n \times p}$ from a random variable $\boldsymbol{Y}$ which we assume stem from a distribution with the probability density $p_{\boldsymbol{\theta}}$, that depends on the parameters $\boldsymbol{\theta} = \mathbb{R}^d$. We are only considering parametric inference in this thesis. We want to estimate $\boldsymbol{\theta}$ using our observations $\boldsymbol{y}$. This can for example be done by *maximum likelihood* estimation. Usually, the estimation procedure together with the model $p_{\boldsymbol{\theta}}$ gives a way to find a probability density of the estimate $\hat{\boldsymbol{\theta}}$. This is the density of the so-called sampling distribution.

2.2.1 Maximum likelihood

A general way to compute parameter estimates and quantify their uncertainty is the technique of *maximum likelihood*. The *likelihood* function $L(\boldsymbol{\theta})$ is defined by viewing the probability of the observed data as a function of the parameters of the probability density function

$$L(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

We denote the log-likelihood function $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. The *maximum likelihood* estimate $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{2.1}$$

If the model $p_{\boldsymbol{\theta}}$ is correct, it is asymptotically efficient, *i.e.*, no other estimate can have lower variance if the amount of observations is large enough [Wakefield, 2013]. Introduce the observed information $\boldsymbol{I}$ given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} l(\boldsymbol{\theta}).$$

If $\boldsymbol{I}(\boldsymbol{\theta}) > 0$, the sampling distribution of the maximum likelihood estimate is asymptotically given according to

$$I(\boldsymbol{\theta})^{1/2} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \rightarrow_d \mathcal{N}(0, \mathbf{I}_d)$$

where $\rightarrow_d$ indicates that the expression to the left converges *in distribution* to the expression on the right, see for instance [Van der Vaart, 2000].

2.2.2 Confidence intervals

The sampling distribution enables us to construct a confidence interval with the level $1 - \alpha$ for one of the parameters θ_i, *i.e.*, an interval $C_n = (a, b)$, $a < b$ such that $\mathbb{P}(C_n \ni \theta_i) \geq 1 - \alpha$. We use the notation C_n to emphasise that the interval is dependent on our sample. Note that it is the interval C_n that is the stochastic variable, not θ_i, which is a fixed value. A useful interpretation of confidence intervals is the following: imagine that in your life as a scientist you pick $\alpha = 0.05$ and construct such intervals for parameters that you estimate in your experiments. Given that the experiments are unrelated, and that all your models are correct, *i.e*, that $\boldsymbol{Y} \sim p_{\boldsymbol{\theta}}$, then over your lifetime the confidence intervals will trap the true parameter 95% of the time.

Example 1 (Confidence interval for normal distribution)
Suppose that $\hat{\theta} = \mathcal{N}(\theta, \sigma^2)$. Let $Z \sim \mathcal{N}(0, 1)$ and let Φ denote the cumulative distribution function of Z, *i.e.*,

$$\Phi(z) = \mathbb{P}(Z \leq z).$$

Let$\sim z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. The function Φ^{-1} is known as the *quantile* function. A $100(1-\alpha)\%$ confidence interval for θ is then given by

$$C_n = [\hat{\theta} - z_{\alpha/2}\sigma, \hat{\theta} + z_{\alpha/2}\sigma],$$

because

$$\begin{aligned}\mathbb{P}\left(\hat{\theta} - z_{\alpha/2}\sigma < \theta < \hat{\theta} + z_{\alpha/2}\sigma\right) &= \mathbb{P}\left(-z_{\alpha/2} < \frac{\hat{\theta}-\theta}{\sigma} < z_{\alpha/2}\right) \\ &= \Phi\left(z_{\alpha/2}\right) - \Phi\left(-z_{\alpha/2}\right) = 1-\alpha.\end{aligned}$$

A 95% confidence interval for $\hat{\theta}$ is therefore

$$C_n = [\hat{\theta} - 1.96\sigma, \hat{\theta} + 1.96\sigma],$$

since $z_{0.025} = 1.96$. Because the probability of an event is invariant under monotone transformations the procedure also gives a confidence interval for $f(\hat{\theta})$

$$C_n = [f(\hat{\theta} - z_{\alpha/2}\sigma), f(\hat{\theta} + z_{\alpha/2}\sigma)]$$

if f is a monotone transformation, such as for instance $\log(x)$ or 10^x. □

2.2.3 Hypothesis testing

The model $p_{\boldsymbol{\theta}}$ defines a parameter space $\boldsymbol{\theta}$, such that $\boldsymbol{\theta} \in \Theta$. A hypothesis test is done by assuming that the parameters belong to a particular subspace of the parameter space: $H_0 : \boldsymbol{\theta} \in \Theta_0 \subset \Theta$. This hypothesis is referred to as the null hypothesis. For example, it could be that the ith parameter is zero. The assumption is translated to a statement about the distribution of a test-statistic that depends on $\boldsymbol{Y}$: $T(\boldsymbol{Y}) \in \mathbb{R}$. A hypothesis is tested by finding a suitable subset $R \subset \mathbb{R}$ of the range of $T(\boldsymbol{Y})$ with a very low probability mass. The hypothesis is *rejected* if $T(\boldsymbol{y}) \in R$. For many standard hypothesis tests, it holds that

$$R = \{\boldsymbol{y} : T(\boldsymbol{y}) > c\}, \tag{2.2}$$

for some constant c. In that case, the null hypothesis is rejected if the observed test-statistic is a lot larger than what would be likely under the distribution of the test-statistic given by the null hypothesis. A measure for how likely the observed statistic is under a null hypothesis is given by the P value, which is given in the special case of a rejection region like that in (2.2) by

$$P = \mathbb{P}\left(T(\boldsymbol{Y}) > T(\boldsymbol{y}) \mid H_0\right). \tag{2.3}$$

The hypothesis is rejected when P is small. Take a particular rule for when a hypothesis is rejected, say when $P < 0.05$. The frequentist interpretation of probability gives the following interpretation: if all your models are correct, *i.e.* if

$\boldsymbol{Y} \sim p_{\boldsymbol{\theta}}$, and you go by the rule $P < 0.05$ for rejecting the null hypothesis in unrelated experiments, then, in your life as a scientist, you will have falsely rejected the null hypothesis for those experiments at most 5% of the time.

There are different types of errors in hypothesis testing. The type I error is to reject the null hypothesis when it is in fact true, which gives a false positive. The type II error is to fail to reject the null hypothesis when it is in fact false leading to a false negative. If the hypothesis is rejected when $P \leq \alpha$, since $P \sim \mathbb{U}(0,1)$ under the null hypothesis [Wasserman, 2013], the probability of committing a type I error is α. Denote this probability ER. A different characterization of the P value is then

$$P = \inf \{\alpha : H_0 \text{ is rejected at } ER = \alpha\} . \tag{2.4}$$

2.2.4 Multiple testing

In modern data analysis in the biological and medical sciences, there are commonly many more variables to consider than samples ($n \ll p$). Such studies rarely involve only one hypothesis test. In GWAS, it is standard practice to perform several million tests. Using a P value threshold of 0.05 would then give tens of thousands of false positives. Therefore, P values from such analyses must be adjusted to account for the number of tests. The adjustment is designed to control the amount of errors expected over a family of hypothesis tests. Given a chosen error rate over a family of tests, FER, the adjusted P value of a marginal hypothesis test with null hypothesis H_k is given, analogously to (2.4), by

$$\text{adj}P = \inf \{\alpha : H_k \text{ is rejected at } FER = \alpha\} . \tag{2.5}$$

Consider a family of m hypothesis tests. Denote the event that the kth null hypothesis is true by $H_k = 0$. Let B be the total number of type I errors. The *family-wise error rate* (FWER) is then defined by

$$\mathbb{P}(B \geq 1 \mid H_1 = 0, \ldots, H_m = 0). \tag{2.6}$$

Note that this error rate assumes that all hypotheses are in fact null. This makes it conservative.

Example 2 (Bonferroni-corrected P values)
Let B_k be a false positive event for the hypothesis H_k. Given a common level α for each marginal test we have

$$\begin{aligned} \alpha_F &= \mathbb{P}(B \geq 1 \mid H_1 = 0, \ldots, H_m = 0) = \mathbb{P}\left(\cup_{k=1}^{m} B_k \mid H_1 = 0, \ldots, H_m = 0\right) \\ &\leq \sum_{k=1}^{m} \mathbb{P}(B_k \mid H_1 = 0, \ldots, H_m = 0) = m\alpha, \end{aligned} \tag{2.7}$$

which means that the Bonferroni corrected P value $P_{\text{Bon}} = mp$ controls the $FWER$ at level α. □

The FWER is conservative and should be reserved for situations where it is likely that the vast majority of the null hypothesis are true. If, *a priori*, it is expected that many null hypotheses will be rejected correctly, then the *false discovery rate* (FDR) is a more suitable error rate. Let K be the total number of rejections of the null hypothesis. Define the false discovery proportion as the proportion of incorrect rejections

$$\text{FDP} = \begin{cases} \frac{B}{K} & \text{if } K > 0 \\ 0 & \text{if } K = 0. \end{cases}$$

The false discovery rate is the expected proportion of incorrectly rejected null hypotheses

$$\text{FDR} = \mathbb{E}\{FDP\} = \mathbb{E}\left\{ \frac{B}{K} \mid K > 0 \right\} \mathbb{P}\left(K > 0\right). \tag{2.8}$$

The FDR makes a compromise. If there are very few rejected null hypotheses, then controlling the FDR approximately also controls the FWER. If, however, there are plenty of signals, then we accept some few false positives in exchange for potentially plenty of true positives that would have remained hidden under the FWER. Adjusted P values under the FDR can be constructed from a set of marginal P values using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995].

2.2.5 Confidence intervals adjusted for selection

It is widely known that the risk of false positives increases with the number of tests, if no multiple testing correction is done. However, that a similar phenomenon occurs with confidence intervals that are *selected* because they are interesting is less appreciated. We illustrate this with a modification of an example in the 2018 lecture notes of the course "Stats 300C: Theory of Statistics" of Emmanuel Candes, given at Stanford [Candes, 2018].

Example 3 (Coverage of selected confidence intervals)
Sample true effects θ_k from $\mathcal{N}(0, 0.2)$. Assume that we have collected data of these effects and have computed estimates $\hat{\theta}_k$ of them with a sampling distribution given by $\mathcal{N}(\theta_k, 1)$. This setting mimics an experiment where most effect sizes are small, such as for example a GWAS or a study investigating associations between some trait and gene expression data. Confidence intervals on the 95% level are constructed as in Example 1:

$$C_n(i) = [\hat{\theta}_k - 1.96, \hat{\theta}_k + 1.96].$$

We expect the true effect to be captured by the confidence interval 95% of the time. The result of a simulation experiment is shown in Figure 2.1. We simulated 1,000 effects and captured 953 true effects with our confidence intervals, as expected.

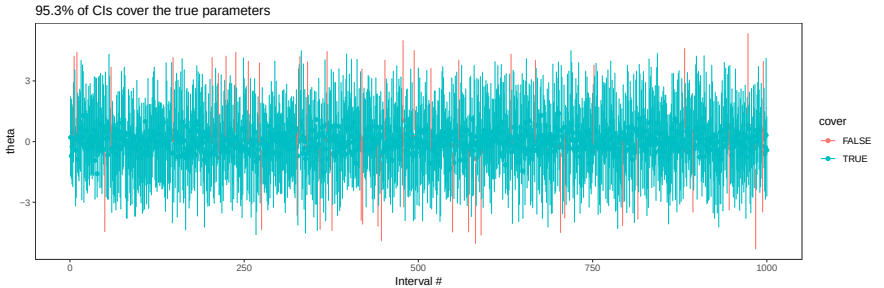

Figure 2.1 Simulation of 1,000 experiments. The true effect is plotted together with the estimated confidence interval. In total, 95.3% of the estimated intervals covered the true effect. Intervals that cover the true effect are shown in blue. Intervals that does not cover the true effect are shown in red.

Now we look only at confidence intervals for interesting parameters. We consider a parameter interesting if it is significant, *i.e.*, if the hypothesis $\theta_k = 0$ has been rejected on the level 0.05. Confidence intervals for such parameters are shown in Figure 2.2. Since the selected parameters are chosen *because they are extreme*, we are more likely to select estimates $\hat{\theta}_k$ that are far out in the tails of the sampling distribution. A confidence interval constructed based on the sampling distribution will therefore not have the correct coverage. Of the 39 selected parameters, 20 are captured by their confidence interval. Running this experiment for a million parameters gives a coverage probability of around 50%. Using a more stringent error rate, like for instance the FWER or the FDR that would typically be used in a real-world study, would lead to even worse coverage. □

Example 3 considers a scenario where a lot of tests are done and where the null hypothesis cannot be rejected for a majority of them. In that case, 95% confidence intervals constructed only for significant parameters will not cover the true parameter 95% of the time. This scenario is extremely common in research in genomics and biomedicine. To find a remedy, we first define the *false coverage rate* (FCR). Let V be the number of confidence intervals not covering their parameter and let R be the number of selected parameters, *i.e.*, the total amount of constructed intervals. The FCR is defined as the expected proportion of confidence intervals not covering their parameter to the number of constructed confidence intervals

$$FCR = \mathbb{E}\left\{\frac{V}{\min(R, 1)}\right\}.$$

Let m be the total number of parameters. Under some technical assumptions not stated here, confidence intervals for selected parameters that control the FCR at level $1 - \alpha$ can then be found by constructing marginal confidence intervals at the level $R\alpha/m$ [Benjamini and Yekutieli, 2005].

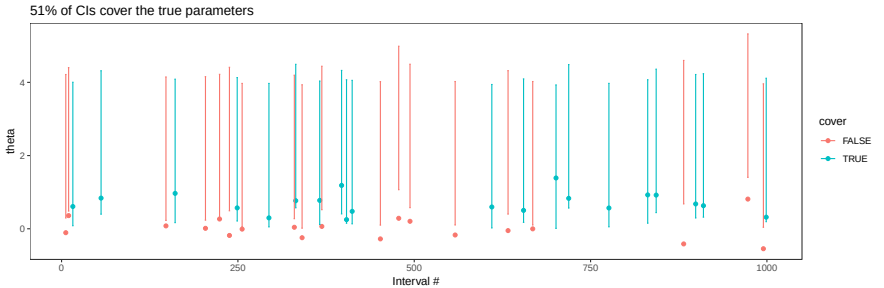

Figure 2.2 The same 1,000 simulated experiments from Figure 2.1. Both true effects and estimates of them were simulated. Here, the experiments that gave rise to extreme estimates were selected, based on a t test of $\hat{\theta}_k = 0$ with $\alpha = 0.05$. Only experiments corresponding to significant estimates are left in the plot. In this case, the estimated confidence intervals only cover their parameter in 51% of the cases. Intervals that cover the true effect are shown in blue. Intervals that does not cover the true effect are shown in red.

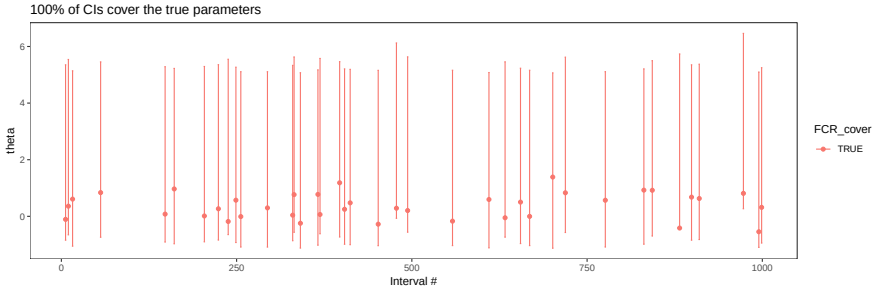

Figure 2.3 The same situation as in Figure 2.2, but in this case the confidence intervals for the extreme estimates are adjusted to control for the false coverage rate. In this case, all estimated intervals cover the true parameter.

Example 4 (Coverage of selected confidence intervals continued)
We repeat the experiment in Example 3, this time constructing FCR-adjusted intervals designed to control the FCR at 0.05. The intervals are therefore constructed on the level $100(1-\alpha)\%$, with $\alpha = 39 \times 0.05/1000 = 0.00195$. The results are shown in Figure 2.3. This time, the intervals cover the true parameter 100% of the time. Running the experiment for a million parameters gives a coverage probability of 97.5%. □

2.3 Regression

Now we consider data from two types of variables: response variables $\boldsymbol{y} \in \mathbb{R}^n$ and predictor variables $\boldsymbol{x} \in \mathbb{R}^{n \times p}$. We think of them as observations from stochastic variables $\boldsymbol{Y}$ and $\boldsymbol{X}$. We want to understand how $\boldsymbol{X}$ relates to $\boldsymbol{Y}$, *i.e.*, we want to find a function f such that $\boldsymbol{Y} = f(\boldsymbol{X})$. This is the problem of *regression*. Given observations $\boldsymbol{y}$ and x, the function f that minimizes the squared error loss is the conditional expectation function $f(\boldsymbol{x}) = \mathbb{E}\{\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}\}$ [Shalizi, 2018]. For simplicity, we will denote the it $\mathbb{E}\{\boldsymbol{Y} \mid \boldsymbol{x}\}$. In parametric regression, this function is assumed to depend on a fixed number of parameters. Typically, the assumed conditional expectation function depends on parameters that illuminate the relationships between the response variable and the predictors. A common simple model is to assume that the conditional expectation function is a linear combination of the predictor variables and that the observations are independent with the same variance. This leads to linear regression, the main workhorse in Science for a century. However, it does not deal well with a situation where there are many predictor variables relative to the number of samples. In statistical learning jargon: it is prone to overfit. It also does not handle dependency between samples.

For more complex data, it is often beneficial to have more general variance assumptions. This can be done, for example, by mimicking linear regression, but assuming that some of its parameters are themselves unobserved stochastic variables with some parameterized distribution. By including the new parameters in the model, it is possible to model dependencies between observations, and also to protect against overfitting. This is a form of regularization, which is the mechanic that most modern machine learning techniques rely on. The main models used in the thesis are mixed models and the elastic net model. Both of these can be seen as special cases of linear models with parameters drawn from particular distributions.

2.3.1 Linear regression

The linear regression model is specified as follows for numerical predictors

$$\begin{aligned} \boldsymbol{Y} &= \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon} \\ \mathbb{E}\{\boldsymbol{\varepsilon}\} = 0, \quad \mathbb{V}\{\boldsymbol{\varepsilon}\} &= \sigma^2 \mathbf{I}_n \,. \end{aligned} \tag{2.9}$$

This means that observations are assumed to be independent with the same variance. The parameter β_p is the additive change in the average response from a unit change in x_p holding all other variables constant:

$$\mathbb{E}\left(Y \mid x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p\right) - \mathbb{E}\left(Y \mid x_1, \ldots, x_p\right) = \beta_j. \tag{2.10}$$

In this thesis, the predictors are either numerical or categorical. Introduce the Dirac delta function

$$\delta(x = a) = \begin{cases} 1, & x = a \\ 0, & x \neq a. \end{cases}$$

For categorical variables we employ corner-point parameterization. For example, if we only have one variable with l possible categories, $x \in \{a_k\}_{k=1}^{l}$, then the assumed conditional expectation function is given by

$$\mathbb{E}\left(Y \mid x\right) = \mu + \beta_2 \delta\left(x = a_2\right) + \ldots + \beta_l \delta\left(x = a_l\right).$$

The parameter $\beta_k, k \in \{2, \ldots, l\}$ is the average difference in response between the group of category a_k and the baseline group of category a_1. We only rarely use interaction terms (where parameters are allowed to depend on other parameters) in the thesis, so we do not explain them or their interpretation here. More information about interaction terms and their interpretation can be found in [Wakefield, 2013].

Example 5 (Log transformation of the response)
The fate of a cell is often determined by binary decisions. For example, a gene that decides the lineage of a cell can be expressed only if one or more transcription factors (proteins that can bind to certain sequences in the DNA string and activate gene expression) are bound. This means that the growth of the population of such cells is determined by multiplicative factors. Measurements of cell numbers from cells that exhibit that type of growth typically exhibits log-normal behaviour. In particular, the standard deviation of the data is proportional to the mean of the data. A good model in this case is

$$\begin{aligned} \log(\boldsymbol{Y}) &= \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon} \\ \mathbb{E}\{\boldsymbol{\varepsilon}\} = 0, \quad \mathbb{V}\{\boldsymbol{\varepsilon}\} &= \sigma^2 \mathbf{I}_n \,. \end{aligned} \tag{2.11}$$

On the scale of the original data, the conditional expectation function becomes [Wakefield, 2013]

$$\mathbb{E}\left\{Y \mid x\right\} = \exp\left\{\beta_1 x_1 + \ldots \beta_P x_P + \frac{\sigma^2}{2}\right\},$$

which gives

$$\frac{\mathbb{E}\left\{Y \mid x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p\right\}}{\mathbb{E}\left\{Y \mid x_1, \ldots, x_p\right\}} = \exp(\beta_j).$$

This means that on the original data scale, $\exp(\beta_j)$ can be interpreted as the ratio of expected responses between populations who differ by one in x_j. Given a sampling distribution for $\hat{\beta}_j$, a confidence interval for $\exp(\beta_j)$ can be constructed from Example 1. □

2.3.2 Mixed models

The model in (2.9) assumes the conditional independence of Y_i, $i = 1, \ldots, n$ given $\boldsymbol{X}$. This assumption does usually not hold, for instance, if data has been collected on the same person across several days, or if the data has been collected in batches. For example, the flow cytometry data that we use in this thesis is collected over 100 days. Each day, samples were taken from around 10 people. It is known that some immunophenotypes show seasonal effects [Carr et al., 2016]. The measurement itself could also be impacted by season. In such cases, observations within processing days would be *dependent*, samples taken during the same day would show a higher degree of similarity with each other than with samples taken during different days. Such dependencies can be modelled by assuming that groups of parameters are drawn from a common distribution. Such models are usually referred to as *mixed models* (along with a multitude of other names). Here, we present the two simple cases of the mixed model that we have used in the thesis. References on the general formulation are [Hodges, 2016; Wakefield, 2013; Ruppert et al., 2003].

Let x_{ij} be the measurement of jth variable for the ith individual. Assume that the there are n observations in total, J groups and introduce a function $r : \{1, \ldots, n\} \to \{1, \ldots, J\}$, that indexes the ith individual into its group. The varying-intercept model for the ith individual is then written

$$\begin{aligned} y_i &= \mu_{r(i)} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \\ \mu_{r(i)} &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\ \varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned} \tag{2.12}$$

Observations from the same group are now correlated with variance σ_μ^2. The varying intercepts, $\mu_{r(i)}$, are commonly referred to as random effects. The variables that are not assumed to be drawn from a distribution are known as fixed effects. Conditional on the random effects, we get back the linear regression model in (2.9). This means that all parameter interpretations for linear regression models still hold for the fixed effects of the varying-intercepts model.

Inference for (2.12) can for example be done by maximum likelihood (see Section 2.2.1), or procedures similar to it, like restricted maximum likelihood (REML), see [Wakefield, 2013; Bates et al., 2015]. In practice, estimates and confidence intervals for mixed models such as (2.12) are usually computed by specialized software. In this thesis we use the *lme4 R* package [Bates et al., 2014] that performs efficient inference for a general class of mixed models using both maximum likelihood, and restricted maximum likelihood estimation.

Example 6 (Pooling)
Random effects models correlation between samples, but it is also beneficial in a more pragmatic way, because it gives rise to *pooling* in the estimates. We illustrate this with a simple example where an analytic formula is available. Consider (2.12),

but with $\beta_j = 0$ for all j so that only the random effect is left. Denote the estimate of the mean in each group by $\bar{y}_r$, and the overall mean across all groups by $\bar{y}$. Further, let n_r be the sample size of group r and n the overall sample size. The maximum likelihood estimate of the random effects can then be approximated by the weighted average of the group means and the overall mean [Gelman and Hill, 2006]

$$\hat{\mu}_r = \frac{\dfrac{n_r \bar{y}_r}{\sigma^2} + \dfrac{\bar{y}}{\sigma_\mu^2}}{\dfrac{n_r}{\sigma^2} + \dfrac{1}{\sigma_\mu^2}}. \tag{2.13}$$

If a group has few samples, the overall mean has a stronger influence on the estimate. This protects against overfitting because estimates of group means that have a higher risk of being inaccurate, because of small sample sizes, are helped by the estimate of the overall mean. The formula in (2.13) does not generalize to more complicated situations, but the concept of pooling does. □

Consider testing the hypothesis $\beta_j = 0$ in the model in (2.12). A first approach would be to use a likelihood-ratio type statistic and use the χ^2 approximation, which is approximately valid for all types of models [Wakefield, 2013]. However, for mixed models it has serious problems for low and medium sample sizes [Faraway, 2016]. Introduce the parameter vector

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

The test is equivalently written $\boldsymbol{L\beta} = 0$, where $\boldsymbol{L}$ is the restriction matrix for the linear hypothesis $\boldsymbol{\beta}_j = 0$, which in this simple case is given by the row vector

$$L_k = \begin{cases} 1 & k = j \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Denote by $\hat{\boldsymbol{V}}$ the covariance matrix of the estimate $\hat{\boldsymbol{\beta}}$. Kenward and Roger have showed that a modification of the Wald-type test statistic

$$\hat{\boldsymbol{\beta}}^T \boldsymbol{L}^T (\boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}^T)^{-1} \boldsymbol{L}\hat{\boldsymbol{\beta}}$$

is approximately distributed according to the F distribution $F_{1,m}$, and they give a means to estimate the degrees of freedom m [Kenward and Roger, 1997]. This test has better small and medium sample size properties than the likelihood ratio test [Kenward and Roger, 1997; Faraway, 2016; Halekoh and Højsgaard, 2014]. For time-constrained analyses in the thesis, we use this test for models of the type in (2.12). Otherwise, we use a parametric bootstrap approach where we fit the model and then simulate a likelihood ratio null distribution. Given a correct model, the parametric bootstrap gives correct inference with enough simulations, see for instance [Faraway, 2016] for more details.

2.3.3 GWAS

The GWAS model for mapping a *single nucleotide polymorphism* (SNP) to a phenotype is another special case of the general mixed effects model. A SNP is a position in the genome that takes two forms in the population, *e.g.*, A and G, usually referred to as *alleles*. As individuals are a combination of their two parents' genomes, they can be either AA, AG, and GG. Assume, for example, that carrying the G allele increases the count of white blood cells. It is usually assumed that white blood cells will be on average β_{SNP} larger in AG individuals, and $2\beta_{SNP}$ in GG individuals, than in AA individuals (in other words, the G form has an additive genetic effect). The SNP predictor is then ordinal and is given for an individual i by counting the number of minor alleles (the allele that is the least frequent in the population).

Suppose that we have the standardized genotypes, $\{z_k\}_{k=1}^m$, $z_k \in \mathbb{R}^n$, of all SNPs that influence a trait. Assuming no interactions among genotypes and between genotypes and the environment, a model for the trait for individual i is

$$\begin{aligned} Y_i &= \mu + g_i + \varepsilon_i \\ g_i &= \sum_{k=1}^{m} z_{ik} u_k \\ \varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right) \end{aligned} \tag{2.15}$$

where u_k is the genetic effect for the kth variant and g_i is the total genetic effect for individual i [Yang et al., 2010]. The term ε_i is a residual term that represents all contributions to the trait from non-heritable factors, which are typically related to the environment of the individual. If some such factors are known, they can be explicitly included as fixed effects. Collect all standardized genotypes z_k in a matrix $\boldsymbol{z} \in \mathbb{R}^{n \times m}$ and all genetic effects u_k in a vector $\boldsymbol{u} \in \mathbb{R}^m$. Define $\mathbb{1}_n$ to be a vector of ones with dimension n. By assuming that the genetic effects are drawn from a zero-mean normal distribution with diagonal covariance matrix we get the following model

$$\begin{aligned} \boldsymbol{Y} &= \mu\, \mathbb{1}_n + \boldsymbol{g} + \boldsymbol{\varepsilon} \\ \boldsymbol{g} &= \boldsymbol{Z}\boldsymbol{u} \\ u &\sim \mathcal{N}\left(0, \sigma_g^2\, \mathbf{I}_m\right) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(0, \sigma^2\, \mathbf{I}_n\right). \end{aligned} \tag{2.16}$$

The variance of the trait is then given by

$$\mathbb{V}\{\boldsymbol{Y}\} = \sigma_g^2 \boldsymbol{z}\boldsymbol{z}^T + \sigma^2\, \mathbf{I}_n\,. \tag{2.17}$$

The matrix

$$\boldsymbol{G} = \boldsymbol{z}\boldsymbol{z}^T \in \mathbb{R}^{n \times n} \tag{2.18}$$

is known as the genetic relationship matrix (GRM) for the trait. The random effect $\boldsymbol{g} \sim \mathcal{N}(0, \sigma_g^2 \boldsymbol{G})$ is an example of very effective regularization: with the $n + 1$ parameters, $\boldsymbol{g}$ and σ_g^2, it models m variables, where m is potentially in the millions.

The genotype data collected for a GWAS is typically from a genotype array (a technology which enables genotyping, *i.e*, to find out what bases a person has at a particular places in the genome) and consists of around a million SNPs. These markers are chosen to maximize linkage disequilibrium (LD, the correlation between genotypes of SNPs at different genomic locations in a given population) with SNPs not on the array. The SNPs that are not genotyped can then be imputed using the 1,000 Genomes data (The 1,000 Genomes project is a large project organised by the 1,000 Genomes consortium that have sequenced thousands of inviduals from various populations to map the variation in the human genome [The 1000 Genomes Project, 2015]) to give all ~10 million SNPs [Howie et al., 2012]. The final set of variants is determined by filtering based on minor allele frequency (lower minor allele frequencies require higher sample sizes for robust analysis) and various quality control checks. There are typically 1 to 10 million markers in the final set. Assume that the final set includes k markers genotyped from n individuals. Collect the standardized markers in a matrix $\boldsymbol{K} \in \mathbb{R}^{n \times k}$. We want to analyse one of the columns in $\boldsymbol{K}$. Denote it $\boldsymbol{SNP} \in \mathbb{R}^n$. This SNP is from chromosome c. Denote $\boldsymbol{K}_{-c}$ the marker matrix with the cth chromosome removed. If the information exists, we could also include a matrix of environmental predictors $\boldsymbol{x} \in \mathbb{R}^{n \times p}$. The GWAS model for the trait $\boldsymbol{Y} \in \mathbb{R}^n$ is then given by

$$\begin{aligned} \boldsymbol{Y} &= \mu + \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{SNP}\beta_{SNP} + \boldsymbol{g} + \boldsymbol{\varepsilon} \\ \boldsymbol{g} &\sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{K}_{-c}\boldsymbol{K}_{-c}^T\right) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_n\right) \end{aligned} \tag{2.19}$$

The random effect $\boldsymbol{g}$ is supposed to control for all genotypes that are not in LD with $\boldsymbol{SNP}$. According to (2.16), this can be done by the GRM. We do not know the causal variants for the trait, so we use the marker matrix $\boldsymbol{K}$ to approximate the GRM. To make sure that no variants in the approximated GRM are in LD with $\boldsymbol{SNP}$, because this would lower power [Yang et al., 2014] (make it harder to distinguish signals from noise), we remove SNPs that come from the same chromosome. The GRM is therefore approximated in (2.19) by

$$\boldsymbol{K}_{-c}\boldsymbol{K}_{-c}^T. \tag{2.20}$$

In (2.16), all environmental effects were collected in the residual term. If the trait under investigation is well-understood, its environmental predictors can be included explicitly with a predictor matrix $\boldsymbol{x}$, like in (2.19), to reduce the variance of the estimate of the SNP effect and potentially control for confounding.

The model (2.19) is fitted for each SNP in the study, which means that it is usually fitted millions of times. The naive approach would mean that the computation time to

fit all models would scale according to the cube of the number of markers, because the covariance matrix would have to be inverted for each SNP. This is extremely wasteful however, since the covariance matrix is fixed for all SNPs except for the variance components σ_g^2 and σ^2. By using a precomputed spectral decomposition of the approximated GRM matrix $\boldsymbol{K}_{-c}\boldsymbol{K}_{-c}^T$, or a closely related approach, the computation scales linearly with the number of markers [Zhou and Stephens, 2012; Lippert et al., 2011; Yang et al., 2011]. Due to the computational constraints, the hypothesis test used is usually a variant of a likelihood-ratio test.

2.3.4 Elastic net

We saw in (2.16) how a random effect can be used to model the effect of millions of variables using relatively few parameters. However, in that case, we do not get an estimate of the parameter of each variable; we only get an estimate of the effect of their sum. We could instead include all SNPs as explicit predictors in the model. But that would make the linear regression model in (2.9) overdetermined; it would have more variables than samples. It would be possible to use all SNPs as predictors if we assume that the parameters of the SNPs are drawn from a normal distribution. The maximum a posteriori (MAP) estimate of the parameters of this model gives the ridge regression model, which has been popular since the 1960s. Assuming that the parameters are drawn from a common normal distribution and taking the MAP estimate gives rise to *shrinkage* in the estimate: the absolute values of the estimates are smaller than for the unbiased estimates, because the magnitude in the Euclidian norm of the parameters is constrained. This is a form of regularization that is very popular in machine learning. It is for instance used in deep learning models where it is called weight decay. Ridge regression shrinks the parameters, but it does not *select*, *i.e.*, all parameters will still be non-zero. A typical prediction problem involving biomolecular data is sparse: only a few of the potentially millions of predictors will actually be predictive of the phenotype. In such cases ridge regression still performs poorly, because the non-predictive predictors swamp the predictive ones. Using the normal distribution for the parameters is unsatisfactory, because it does not have enough weight centered at zero. Assuming instead that each parameter is drawn from a Laplace distribution, $\beta_k \sim \text{Laplace}(0, 1/\lambda)$, and taking the MAP estimate gives the LASSO problem [Park and Casella, 2008], put forth in its original shape by Robert Tibshirani in 1996 [Tibshirani, 1996]. Given response variables $\boldsymbol{Y} \in \mathbb{R}^n$ and scaled and centered predictors $\boldsymbol{x}$, the LASSO estimator is written as the *convex optimization* problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 : \ \|\boldsymbol{\beta}\|_1 \leq t \right\}. \tag{2.21}$$

The term $\|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2$ is the usual least squares cost used for estimating parameters in linear regression. The term $\|\boldsymbol{\beta}\|_1 \leq t$ constrains the parameters in such a way that, for a given t, the estimator both shrinks and selects. Coefficients that do not rise above

a noise ceiling will be put to exactly zero, and coefficients that are large enough will be shrunk. The way the LASSO does this gives it a lot of favourable properties, such as having the degrees of freedom being exactly the number of non-zero coefficients and being continuous in the observations $\boldsymbol{y}$ of the response, see [Tibshirani, Taylor, et al., 2012] for details. The optimization problem in (2.21) is equivalent to the so-called Lagrangian form of the LASSO problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{2.22}$$

in the sense that for each t there is a corresponding λ that gives the same solution [Hastie et al., 2015].

Typically, the whole solution path in $\boldsymbol{\beta}$ is estimated from (2.22), *i.e.*, estimates $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ are computed for all values of λ that change the number of non-zero estimates of $\boldsymbol{\beta}_k$. This gives an estimate of the dependency between model degrees of freedom, a measure of regularization or model complexity, and prediction accuracy. The algorithm starts with a λ that is large enough to make all coefficients zero. For linear regression, that occurs when

$$\lambda > \max \left(\boldsymbol{x}^T \boldsymbol{y} \right). \tag{2.23}$$

It then gradually increases λ, leading to more coefficients being non-zero. The LASSO problem can be solved efficiently with coordinate descent [Hastie et al., 2015], by using the solution for a λ as a starting guess for the next λ in the sequence.

The solution to (2.21) suffers from a phenomenon called saturation: unlike ridge regression it cannot have more non-zero coefficients than the number of samples. In, admittedly rare, situations the solution is also not unique. These problems can be alleviated by including an additional constraint on the Euclidian norm of the parameters leading to the elastic net problem, which we state in the Lagrangian form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 + \lambda \left((1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right) \right\}, \ \alpha \in [0,1]. \tag{2.24}$$

Another difference between the two optimization problems is that the LASSO tends to select only one of two highly correlated variables, while the elastic net picks both of them while shrinking their coefficients by half. In this thesis we favour elastic net regression.

The parameters λ and α are usually chosen by *cross-validation*: we divide the data up in k blocks, sweep over a grid of λ and α values; for each parameter pair we go through each block in turn, and solve (2.24) using the data in all blocks but the hold-out block, and then we predict the hold-out block using the fitted model and store the prediction error. Usually, the parameter setting that gives the lowest out-of-sample prediction error is then used for prediction. More details on cross-validation

can be found in for instance [Friedman et al., 2001]. The cross-validation scheme that we custom-built for Paper III in the thesis is outlined in Algorithm 1. That scheme repeats the procedure twice, giving twice as many samples the prediction error.

The cross-validation technique estimates a distribution over the regularization path. The regularization path is of interest on its own, because it shows how the prediction error of a response variable is dependent on model complexity.

Algorithm 1 Cross-validation for elastic net linear regression

Stated here using the correlation between out-of-sample predictions and observed values as performance estimate. The case for other performance measures are analogous. Given observed responses $\boldsymbol{y} \in \mathbb{R}^n$ and predictors $\boldsymbol{x} \in \mathbb{R}^{n \times P}$, our cross-validation scheme in Paper III conceptually goes as follows

for $r = \{1, 2\}$ **do**
  Divide data 10 equally sized blocks $\boldsymbol{y}_k$ and $\boldsymbol{x}_k$. Denote data that is not in the kth block with $\boldsymbol{y}_{-k}$ and $\boldsymbol{x}_{-k}$
  for $\alpha \in \{0.05, 0.1, 0.5, 0.95, 1\}$ **do**
    Compute $\lambda_{max} = \frac{1}{\alpha} \max\left(\boldsymbol{x}^T \boldsymbol{y}\right)$
    Let l contain 200 values logarithmically from $10^{-4}\lambda_{max}$ to λ_{max}
    for $k \in \{1, \ldots, 10\}$ **do**
      for $\lambda \in l$ **do**
        Solve (2.24) for $\boldsymbol{x}_{-k}$ and $\boldsymbol{y}_{-k}$
        Find prediction: $\hat{\boldsymbol{Y}}_k = \boldsymbol{x}_k \hat{\boldsymbol{\beta}}$
        Store $\text{corr}(\boldsymbol{y}_k, \hat{\boldsymbol{Y}}_k)$ in $\varepsilon(\alpha, \lambda, k, r)$
      end for
    end for
  end for
end for
The optimal model is found by

$$\hat{\alpha}, \hat{\lambda} = \arg\min_{\alpha, \lambda} \frac{1}{20} \sum_{k,r} \varepsilon(\alpha, \lambda, k, r)$$

2.3.5 Stability selection

Arguably, the LASSO estimator first rose to prominence due to another interpretation: that it is a convex relaxation of the *best subset* problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 : \ |\{\boldsymbol{\beta} \neq 0\}| < c \right\}. \tag{2.25}$$

Here, $|A|$ denotes the *cardinality* of a set A which counts the number of elements in the set. The problem in (2.25) is non-convex and can be NP-hard to solve, and is therefore not useful for large problems. The idea was that it is really this problem that should ideally be solved in a sparse setting, and that (2.21) is an approximation of it. However, it is now acknowledged that the LASSO problem (2.21) has a number of beneficial properties in its own right and that it would not necessarily be better to use (2.25), even if it was possible. Particularly for prediction, recent results seem to favour the LASSO problem for in the low signal-to-noise ratio regimes that are typically encountered with biomolecular data [Hastie et al., 2017].

Unlike the LASSO, (2.25) is a pure selection problem: it does not shrink the coefficients. As a selection algorithm, the LASSO is quite poor, requiring stringent assumptions to recover the true predictors. Best subset selection recovers the true predictors under weaker assumptions [Zhang, Zhang, et al., 2012], but it is untested in practice, it is unknown how it deals with realistic data with unfavourable noise and correlation properties, and even if state-of-the-art optimization techniques are used, it can only be solved approximately and for problems on the scale of a few thousand predictors [Bertsimas et al., 2016].

For selection problems, we instead favour the *stability selection* technique [Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013]. It uses a weak selection algorithm or, in the statistical learning jargon, a weak learner, and constructs a distribution over the probability of a particular variable being selected. It does this by selecting variables on sampled subsets of the original data. The final selection is done by thresholding the estimated probability. It can be seen as a selection model that uses the ensemble-learning framework of, for instance, the random forest or the gradient boosting machine models [Friedman et al., 2001], in the sense that the results from many weak learners are improved by an aggregation scheme. The aggregation over subsamples of the data makes stability selection a very robust selection algorithm. This is important, since selection is inherently discontinuous. We use stability selection, for instance, to select environmental predictors in a GWAS of immunophenotypes. If the selection was not robust, the sampling distribution of the SNP coefficients could have bimodal characteristics, which would invalidate many of the inference procedures for them.

Consider the data model

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbb{V}\{\boldsymbol{\varepsilon}\} &= \sigma^2 \mathbf{I}_n \,. \end{aligned}$$

A selection algorithm estimates the *support* S of the model

$$S(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_j \neq 0\}. \tag{2.26}$$

The learner that we use is weaker than the LASSO with the cross-validation tuning outlined in Algorithm 1. We select q variables by increasing λ in (2.24) until q

predictors has been included. To be more precise, introduce the family of support estimators

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}\|_2^2 + \lambda \left((1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1 \right) \right\}$$
$$\hat{S}(\lambda) = \left\{ j : \hat{\boldsymbol{\beta}}_j \neq 0 \right\}. \tag{2.27}$$

We then use the support estimator $\hat{S}_q = \hat{S}(\lambda^*)$, where λ^* is such that

$$\lambda^* = \min \left\{ \lambda : |\hat{S}(\lambda)| < q \right\}.$$

We keep the ridge regression penalty small by setting α to a value close to 1. Employing this base selection algorithm to estimate $\hat{S}$, the stability selection scheme that we use in the thesis is outlined in Algorithm 2.

Algorithm 2 Stability Selection
Given observed response $\boldsymbol{y} \in \mathbb{R}^n$ and predictors $\boldsymbol{x} \in \mathbb{R}^{n \times p}$ the stability selection scheme goes as follows.

1: Subsample rows of $(\boldsymbol{y}\ \boldsymbol{x})$ in m pairs, where each pair contains half of the rows, giving a total of $2m$ subsets $B_k,\ k = 1, \ldots, 2m$
2: For all subsets, B_k estimate support $\hat{S}_q(B_k)$
3: For all predictors $\boldsymbol{x}_j$, estimate

$$\hat{\mathbb{P}}_j = \frac{1}{2m} \sum_{k=1}^{2m} \mathbb{1}_{j \in \hat{S}_q(B_k)}$$

4: Include $\boldsymbol{x}_j$ as a predictor if $\hat{\mathbb{P}}_j$ is above a certain threshold
5: The threshold is chosen such that, under stringent assumptions, the maximum number of expected false positives is < 2, see [Shah and Samworth, 2013] for more information.

2.4 Causal inference

Traditionally, the area of statistics has dealt mostly with the small world of the model. Textbooks and scientific articles often take a ground truth model for granted. More often than not, this model used to be linear with normally distributed residuals. With more powerful computing and the subsequent coming of simulation-based methods like bootstrap methods and Bayesian modelling, and black-box methods like the random forest and deep learning, the assumptions of linearity and normal residuals have been relaxed.

None of these developments aid in understanding the fundamental challenge of *confounding* in observational studies. Confounding is a phenomenon where estimated associations are faulty because of improper modeling. In science, statistical models are used to gain knowledge. Arguably, the biggest obstacle in this pursuit is not the small world of the model, but the large world, in the sense that when we report new knowledge, new effect sizes and associations, these actually describe the phenomena we are investigating. Confounding can break the link between observations and reality. The problem is that the large world cannot be described in the language of probability. It therefore does not fit in the 20*th* century paradigm of the statistical field, which models observations with probability distributions. If we understand statistics in this narrow sense, the problem of confounding is actually not a statistical problem at all. It is therefore not surprising that it is not represented well in statistical textbooks. Confounding cannot be represented with probability because it is a manifestation of a deeper problem, that of *causality*. The concept of causality is fundamentally asymmetrical. The relation that A *causes* B does not imply that B causes A. However, specifying a joint distribution $\mathbb{P}(A, B)$ and claiming that A is conditionally dependent on B does imply that B is conditionally dependent on A. The concept of causality has to be described in a language that accommodates asymmetric relationships. The approach taken by Judea Pearl in the early 1990s was to describe causal relationships by adding crucial logic to the language of the probabilistic graphical model [Pearl, 2009]. The framework puts the problem of confounding on a firm theoretical basis and gives rise to a calculus of confounding that specifies what statistical model to use for investigating a particular research question. Prior to any data analysis, Pearl's framework requires the specification of a *causal graph* of the system to be investigated. This causal graph formalizes *a priori* knowledge and assumptions of the system and specifies clearly when model estimates map to actual knowledge.

The framework has seen widespread use in the empirical sciences, particularly in epidemiology [Robins et al., 2000], and in the social sciences [Morgan and Winship, 2015], fields that often have to rely on observational studies for understanding the world. The biological sciences, and immunobiology in particular, have traditionally employed more experimental procedures, where the problem of confounding is not as pervasive (there are exceptions to this, for instance batch effects are often confounding). Understanding biological systems *in natura* requires also the biological sciences to confront the challenges of observational studies.

2.4.1 Probabilistic graphical models

A probabilistic graphical model specifies a joint probability distribution over a set of random variables in terms of marginal conditional distributions for each variable. As an example, introduce random variables A, B, C and D. Assume that the joint distribution can be decomposed as

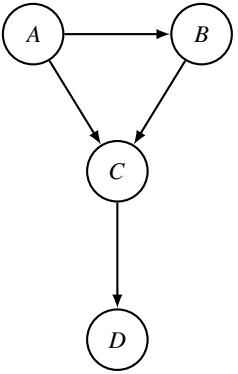

Figure 2.4 Graphical model for the decomposition in (2.28)

$$\mathbb{P}(A, B, C, D) = \mathbb{P}(D \mid C)\,\mathbb{P}(D \mid A, B)\,\mathbb{P}(B \mid A)\,\mathbb{P}(A)\,. \tag{2.28}$$

A decomposition of a joint probability distribution like this can be represented in a graphical model by having a node for each variable and drawing an arrow from one node to another if the variable represented by the first node is conditioned on in the marginal distribution for the variable represented by the second node. A graphical model for the particular decomposition in (2.28) is given in Figure 2.4.

Such graphs are known as directed graphs. A *directed acyclic graph* (DAG) is a directed graph that contains no cycles, *i.e.*, we cannot walk along arrows and end up in a starting node. In the graph in Figure 2.4, A and B are parents to C and A is a parent to B. We say that B is a child of A and that C is a child of A and B. The node D is a descendent of A and B. A node D is a descendent of a node A if it is possible to walk from A to D along the directions of the arrows. For a node, the parents of the parent nodes are also its parents. Denote the parents of the kth node with parents_k. Given a DAG with m nodes, the joint probability distribution is given by

$$\mathbb{P}(x_1, \ldots, x_k) = \prod_{k=1}^{m} \mathbb{P}\left(x_k \mid \text{parents}_k\right).$$

The key reason why the probabilistic graphical model can be used to understand confounding is that it encapsulates conditional independencies between the variables it models. Given random variables A, B and C (not the same as before), we say that A and B are conditionally independent given C, denoted $A \perp\!\!\!\perp B \mid C$, if

$$\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C),$$

or equivalently if

$$\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C).$$

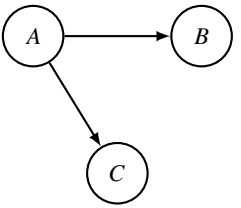

Figure 2.5 Directed acyclical graph describing a fork. The nodes B and C are dependent in this scenario, because of the common ancestor A. Conditioning on A renders them independent.

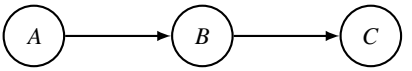

Figure 2.6 Directed acyclical graph describing a chain. The nodes A and C are dependent in this scenario. Conditioning on B makes them independent.

If we condition on a variable C, we say that we observe that variable. By a *path*, we mean an unbroken and nonintersecting walk along the edges in a graph (arrow directions does not have to be followed). If two variables at each end of the path are dependent, then we say that the path is open, otherwise we say that the path is blocked.

Three canonical examples

There are three typical relationships in a graphical model that describe conditional independencies. The first one, the fork or the common cause, is depicted in Figure 2.5. In this case, B and C are dependent because of their common ancestor. However, conditional on A, B and C are conditionally independent since

$$\mathbb{P}(B, C \mid A) = \frac{\mathbb{P}(A, B, C)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A)\,\mathbb{P}(A)}{\mathbb{P}(A)} = \mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A).$$

We say that conditioning on A blocks the path between B and C. Another relationship is the chain, depicted in Figure 2.6. Here, C and A are again dependent, because A is an ancestor of C. In this case, they become independent are conditional on B. The third canonical relationship is that of the *collider*. It is the least intuitive one and it can cause serious biases if not appreciated; a phenomenon known as collider bias. This will be explained in more detail in the causal inference section. The relationship is shown in Figure 2.7. The variable A here is known as a collider. The graph in Figure 2.7 implies that

$$\mathbb{P}(A, B, C) = \mathbb{P}(C)\,\mathbb{P}(B)\,\mathbb{P}(A \mid B, C).$$

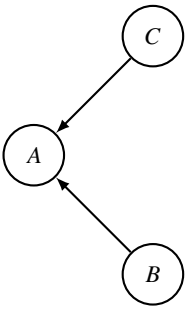

Figure 2.7 Directed acyclical graph describing a collider A. The nodes B and C are independent in this scenario. Conditioning on A makes them dependent.

By marginalizing over A we find that B and C are independent. However, in this case they *become dependent* conditional on A. The conditional distribution of C and B given A is

$$\mathbb{P}(B, C \mid A) = \frac{\mathbb{P}(A, B, C)}{\mathbb{P}(A)} = \frac{\mathbb{P}(C)\,\mathbb{P}(B)\,\mathbb{P}(A \mid B, C)}{\mathbb{P}(A)}$$

which does not in general factorize to $\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A)$. Conditioning on A *opens* the path between B and C. This would occur even if we did not condition on A, but a descendant of it.

The d-separation criterion

Consider sets of variables A, B and C, represented as sets of nodes in a directed acyclic graph G. We now state the d-separation criterion, which explains the conditional independence structure of a graph. We first define d-separation, which we give in terms of the blocking of paths between nodes [Bishop, 2006].

Definition 1 (d-separation or path blocking)
(Mostly taken from Definition 8.2.2 in [Bishop, 2006]. In [Pearl, 2009] it is Definition 1.2.3)
A path p, is said to be d-separated, or *blocked*, if it include nodes i, m and j such that if either

1. m is in a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$, then m is in C.
2. m is a collider $i \to m \leftarrow j$, then the node m, or any descendant of it, are not in C.

A set C is said to d-separate A from B iff C blocks all paths from a node in A to a node in B. □

We can now give a criterion for conditional independence between sets of nodes in a directed acyclic graph.

Theorem 1 (Conditional independency and d-separation)
(Theorem 1.2.4 in [Pearl, 2009])
A and B are conditionally independent given C in any distribution compatible with the graph G if C d-separates A and B. Conversely, if A and B are not d-separated by C then there exists a distribution compatible with G where A and B are dependent conditional on C. □

2.4.2 Causal graphical models

The intuitive leap required to find a calculus of causality is to identify the relation $X \to Y$ in a probabilistic graphical model with the asymmetrical notion that X *causes* Y. In this context, *cause* is taken as a primitive and the veracity of the claim has to be scrutinized for each case. To be more precise, for a causal graphical model, we consider a directed acyclical graph to model deterministic functions f_k such that

$$X_k \leftarrow f_k\left(\text{parents}_k, U_k\right). \tag{2.29}$$

Note that this relation is asymmetrical, going from right to left. There are some restrictions on structure of the graph for it to be a causal model, see Definition 1.3.1 in [Pearl, 2009]. The variables U_k are stochastic variables that represent unobserved and unmodelled variables. The variable X is defined to be a direct cause of Y if Y is a child of X. The variable X is a cause of Y if it is a direct cause of Y or a direct cause of a cause of Y. A simple first requirement for modeling X to be a cause of Y (*i.e.* to draw an arrow between X and Y) is that first, X should be active temporally before Y, and that second, given values of X (and all other causes of Y), the values of Y can be computed.

In a scientific experiment, the impact of a variable X on another variable Y is investigated by taking a random sample from a population and intervening, *i.e.* physically changing X, for some percentage of the population. The difference between the two groups is the average *causal* effect of the treatment.

Inspired by the special case of an interventional experiment, we now define the causal effect for a general causal system. To find the causal effect of X on Y we *intervene* on X, *i.e.* we set the variable to a particular value $X = x$. The causal effect of X on Y, denoted $\mathbb{P}(Y \mid \text{do}(X = x))$, is given by removing all terms in the causal model (2.29) where X occurs on the left, and setting $X = x$ in all equations where X occurs on the right. This corresponds to doing surgery on the graph, removing all links coming in to the node X. As an example, imagine that we are interested in investigating if taking paracetamol gives a person the flue. We take a large random sample of the population. Suppose Y indicates whether someone had the flue the previous week and that X is the number of paracetamol pills the person took during that week. The naive statistical approach to investigate the effect of paracetamol on the risk of getting the flue is to look at an expression involving conditional distributions, such as

$$\mathbb{P}\left(Y \mid X = 1\right) - \mathbb{P}\left(Y \mid X = 0\right). \tag{2.30}$$

But this expression would lead us to conclude that taking paracetamol does give the flue, because it is more likely that people sick with the flue takes paracetamol. If we instead took a random sample of the population and gave half of the people in the sample paracetamol, we would not find any difference in the risk of getting the flue between the two groups. This would correspond to looking at

$$\mathbb{P}\left(Y \mid \mathrm{do}(X = 1)\right) - \mathbb{P}\left(Y \mid \mathrm{do}(X = 0)\right), \tag{2.31}$$

which would be approximately zero.

We now give a simple criterion for how to estimate a causal effect in an observational study of a system modelled by a causal graph. First a definition:

Definition 2 (Back-door)
(Definition 3.3.1 in [Pearl, 2009])
A set of variables Z satisfies the back-door criterion relative to a pair X, Y in a directed acyclical graph G if:

1. No node in Z is a descendant of X
2. Z blocks (by d-separation) every path between X and Y that contains an arrow into X □

The causal effect of X on Y is then given by *adjusting* on the variables in Z:

Theorem 2 (Back-door criterion)
(Theorem 3.3.2 in [Pearl, 2009])
If a set of variables Z satisfies the back-door criterion relative to X and Y, then the causal effect of X on Y is given by

$$\mathbb{P}(Y \mid do(X = x)) = \sum_{Z} \mathbb{P}\left(Y \mid X, Z\right) \mathbb{P}\left(Z\right). \tag{2.32}$$

□

In a linear regression setting, where (2.32) is represented by a linear conditional expectation function, together with some simple assumption on the conditional variance, this amounts to including all the variables of Z in the conditional expectation function, *i.e.*, including them in the regression as covariates. In that case, the causal effect is the regression parameter of the variable X. To make it snappy: if all back-door paths are blocked, then correlation equals causation!

Intuitively, unblocked paths between two variables X and Y in a causal graph represent flows of causation. To estimate the functional form of that causation for one path, all other paths must be blocked. For such control to work well, the functional form of the dependency must be well estimated. If this is not the case, there can still be vestigial confounding. For example, the functional dependency on age is often approximately logarithmic or piece-wise linear. Then, even if age is controlled for linearly, a binary treatment variable affected by it can still pick-up nonlinear age effects.

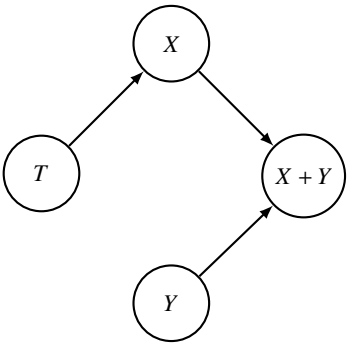

Figure 2.8 Causal graph describing the situation in Example 7. The nodes X and Y are independent. The node T has an effect on X. Conditioning on the collider $X + Y$ would induce spurious correlation between T and Y.

Now we can understand collider bias. If the back-doors paths are given in terms of chains and forks, then they are blocked by conditioning on variables on the paths. However, conditioning on a collider actually opens up the path and gives spurious correlation.

Example 7 (Conditioning on sum)
Consider the simple causal graph in Figure 2.8. Here we have collected data on the variables X, Y, their sum and T, and we are interested in investigating the effect of the variable T on Y. We know from the graph that there is no such effect. But conditioning on the sum $X+Y$ would open up the path $T \to X \to X+Y \leftarrow Y$ which would give a spurious effect. Imagine that you are taking the bus from your home to work. At one point you have to change bus, so there are two bus rides. These two bus rides have little to do with each other, so the duration of each bus ride is independent. The total time it takes for you to get to work is the sum of the duration of each bus ride. In the example, the duration of the first ride is X and the duration of the second ride is Y. The total time it takes to get to work is $X + Y$. Knowing the duration of the first ride does not give any information about the duration of the second ride. But if we somehow know the total time it takes to get from home to work, $X + Y$, then knowing X would lead us to know Y as well. Conditioning on the collider total time to work makes the duration of each individual ride dependent. □

Example 8 (Fisher's smoking controversy)
The famous statistican Ronald Fisher argued in the mid 1900s against the evidence that smoking causes lung cancer [Stolley, 1991]. One of his arguments was that there might exist a genotype that increases both the propensity to smoke and the risk of getting lung cancer. At the time it was impossible to gather data about genotypes, so this variable was unobserved. If such a genotype exists, it is a confounder, and because it is unobserved it is impossible to determine if there is a causal effect of

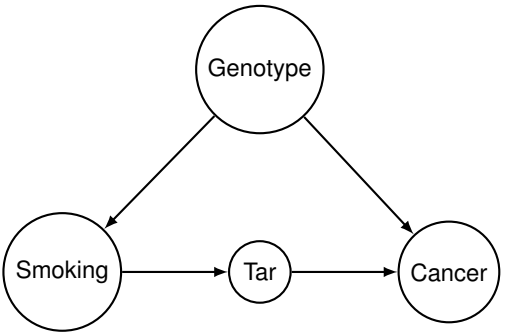

Figure 2.9 Causal graph describing the situation in Example 8. Without information about tar deposits in the lung it is impossible to estimate the causal effect of smoking on cancer risk, because of the unmeasured confounder genotype. But if tar deposits in the lung are observed, the causal effect of smoking on cancer can be estimated even if there is an unmeasured confounder, by summing about the estimable causal effects of smoking on tar deposits, and of tar deposists on cancer risk.

smoking on lung cancer. However, if another variable that lies on a hypothesised chain between smoking and cancer risk can be observed, then the causal effect of smoking on cancer can still be estimated by the so-called front-door adjustment (see Theorem 3.3.4 in [Pearl, 2009]). As an example, we hypothesise that it is the deposit of tar from cigarette smoking that increases the risk of lung cancer. Assume that we can measure this variable. The causal graph of this situation is given in Figure 2.9. The front-door adjustment in this setting would be to estimate the causal effect of smoking on cancer by summing up the causal effects of smoking on tar deposits, and of tar deposits on cancer. A more general calculus of causal effects is developed in Theorem 3.4.1 in [Pearl, 2009]. In the situation depicted in Figure 2.9, there is a direct effect of smoking on tar deposits. There is no back-door effect mediated through genotype, since that path ends in a collider with cancer. The effect of tar deposits on cancer risk can also be readily estimated by adjusting on smoking. Assuming that the probability graph structure in Figure 2.9 is valid, the sum of these effects is the causal effect of smoking on cancer. □

We now use the formalism of causal graphs to illuminate two extremely important situations. First we will revisit the GWAS design from Section 2.3.3. Then we will shed some light on estimating the effect of a variable on gene expression in whole blood.

Example 9 (GWAS)
The process of evolution acts locally on a population. Mutations that increase fitness in one population do not necessarily increase fitness in another distant population. An extreme example of this is the allele that gives rise to sickle cell disease. This allele also renders its carrier immune to malaria. Malaria is a lot more detrimental than

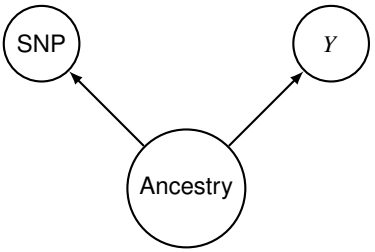

Figure 2.10 Causal graph describing the population stratification problem in GWAS. Ancestry can have an effect on both a SNP and a trait. In that case it induces spurious correlation between the SNP and the trait.

sickle cell disease. The allele is therefore highly beneficial in areas where malaria is prevalent, such as in Africa. However, in Europe, where malaria is currently absent, the allele is virtually missing, because sickle cell disease still has detrimental health effects. Another example is the allele that encodes for lactase, the enzyme that catalyzes the breakdown of lactose, the sugar in milk. This allele is for instance much more frequent in northern Europe than in southern Europe, probably because it has allowed first farmers carrying the allele to digest milk in adulthood [Ségurel and Bon, 2017]. In the whole genome there are dozens of similar examples [Fan et al., 2016]. When we investigate the association of a particular SNP with a trait, we can model this by saying that the ancestry of individuals has an effect on the SNP genotypes. However, ancestry also affects an individual's life, for instance, it will probably to some degree influence the culture that encompasses the individual. We define culture loosely as an individual's way of life. This variable has the potential to have a huge influence on an large number of traits. Assuming that nothing else affects either the trait y or the treatment variable SNP, we can draw a causal graph as in Figure 2.10. Ancestry lies in a fork between SNP and y and is therefore a confounder. This problem is known as population stratification and it caused substantial issues in the beginning of the GWAS era [Tian et al., 2008]. We cannot observe ancestry directly. However, if two people are of similar ancestry that means that they are genetically related. This can be measured by counting the number of alleles they have in common. By doing this for every pair of individuals, we can tabulate how closely related everyone is. This is exactly what the genetic relatedness matrix in (2.18) does. Ancestry can then be estimated by giving similar values to individuals if they are closely related genetically. This is precisely the way the random effects term, $\boldsymbol{g}$, in the GWAS model in (2.19) function. □

Example 10 (Gene expression in whole-blood)
Biomolecular measurements such as gene expression are often measured in whole-blood. Naturally, this is particularly true in systems immunology where blood cells are often the target of interest. An ubiquitous research question is if some variable

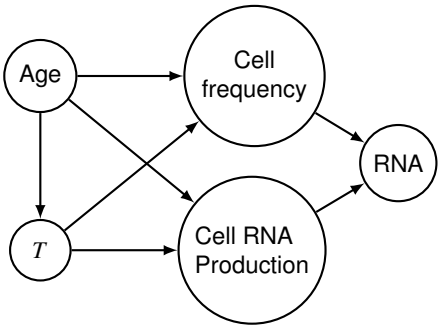

Figure 2.11 Causal graph describing gene expression in whole blood.

T impacts the expression of a gene in whole-blood. Typically, this would involve measurements of the abundance of RNA transcripts from that particular gene. To simplify, we assume here that this gene is only expressed in a particular blood cell subset. As we show in Paper I in the thesis, blood cells are often affected by age: the frequency of naive cells (cell types that are not so far gone in their differentiation chain) diminishes with age, while the frequency of differentiated blood cells rise with age [Patin et al., 2018]. Age is a compound variable that impacts a lot of other factors. We therefore assume that it also affects the variable T. Independently of the frequency of the blood cell, the production of transcripts of each individual cell also has an effect on the frequency of transcripts. This variable is difficult to measure and is therefore not observed. It would be natural to think that this production is also affected by age. We assume that our treatment variable T impacts both cell frequency and cell transcript production. A causal graph of this system is depicted in Figure 2.11. The system is simplified in many aspects, the most obvious one being that a gene is usually expressed by more than one type of blood cell.

In this situation, the consensus in the literature at the moment is to control for immune cell frequency. As a proxy for the actual subset that expresses the gene, the major (commonly six) cell-types in blood are controlled for. The cell-type that expresses the gene would be a subset of these major cell-types, and it is hoped that it is correlated enough with its superset that control for its frequency is achieved. The effect that is estimated by this procedure is described as the effect of T on gene transcript abundance. We will shed some light on this situation using the causal graph in Figure 2.11.

We assume that we have observations of cell frequency, or at least good enough proxies of it. Only conditioning on cell frequency in this situation would open a path to RNA from T through age and cell RNA production. The estimated effect would then be a mix between this path, the fork through age, and the direct path through cell RNA production; information that would be very difficult to interpret and make use of. Note that the regression parameter for T could very well be significant. However,

age would typically also be conditioned on. This would block the paths through age and cell frequency, leaving only the path through RNA production. The estimated effect should then be interpreted as the effect of T on RNA abundance, *mediated* through cell RNA production. Only controlling for age would give the total effect of T on RNA abundance mediated through both cell frequency and RNA production□

3

Milieu Intérieur

The French doctor Claude Bernard coined the term Milieu Intérieur in the mid 1800s to conceptualize his observations of the extraordinary ability of the organs and internal mechanisms of the body to uphold an equilibrium, which is now referred to as homeostasis, in the face of immense and varied external disturbances. An archetypical such system is the immune system, which is capable of neutralizing endless external threats. In honour of Claude Bernard's timeless idea, we use his concept as the name of our study, where we unravel the variation and diversity in the immune system, needed for it to be able to maintain the Milieu Intérieur.

The immune system is a wonderfully complex system developed to protect vertebrate organisms against disease. To achieve that purpose, it must be able to carry out two main functions: it must be able to detect harmful agents in the body, and subsequently neutralize them. Consequently, the human immune system has evolved to become an extremely diverse sensor system, capable of detecting a vast amount of foreign molecules. After detection, it brings a sophisticated weaponry to bear on the carrier of the molecule; be it a tumour cell or a pathogen. The main functions of the immune system are carried out by the immune cells. The immune system is delineated into two subsystems: the innate and the adaptive immune systems. The immune cells of the innate immune system are the early responders to an immune challenge. They consist of a varied repertoire of immune cells, for example macrophages, neutrophils, monocytes, natural killer cells (NK cells) and dendritic cells, whose main purpose is to detect a new immune challenge and, either neutralize the danger immediately, or activate and arm the highly specialized cells of the adaptive immune system. Macrophages are long-lived cells that reside in almost all tissues during homeostasis. It is the main phagocytic cell (it can engulf and neutralize other cells) in the body. It is often the first cell to encounter a pathogen. Neutrophils are the most numerous cells; they reside in the blood and migrate to tissues during infection. They also have phagocytic activity. Monocytes also reside in the blood during homeostasis, and migrate to tissues during infection, where they can differentiate to macrophages. Dendritic cells are the main sensor cells of the innate immune system. Their primary purpose is to communicate with and direct the

cells of the adaptive immune response. The main function carried out by the natural killer cells (NK cells), as their name implies, is to destroy cells infected by viruses.

Immune cells detect foreign molecules through a number of different both generic and highly specialized organic sensors known as receptors. The different parts of the immune system also use these receptors to communicate with each other. The innate immune cells carry generic receptors that recognize molecules that are commonly found on pathogens, but not in the organism itself, such as lipopolysaccharide (LPS), a molecule that is attached to the outer membrane of gram-negative bacteria. Such foreign molecules are known as "pathogen associated molecular patterns" (PAMPs) and are recognized by receptors such as the "toll-like receptors" (TLRs) on innate immune cells.

In contrast to the generic innate receptors, the cells of the *adaptive* immune system carries receptors with an enormous amount of specificities. The adaptive immune response is mediated by T cells and B cells at various stages of differentiation. The T cell and B cell receptors are unique to each cell. They are constructed from DNA sequences that consist of combinations of smaller DNA segments. This *somatic recombination* is a mechanism, unique among all life, to create protein diversity; it is estimated that the total T cell receptor repertoire can consist of 10^{18} different specificities [Murphy and Weaver, 2016]. For T cells, this process is carried out in the *thymus* and we investigate it in the second paper of the thesis.

T cells are alerted to the presence of pathogen-specific molecules, so-called antigens, by communicating with other cells of the body, primarily cells from the innate immune system such as dendritic cells. An antigen from a harmful agent is presented to the T cells by molecules on the surface of cells known as major histocompatibility complex (MHC) type I and type II proteins. Prior to encountering their target antigen, T cells are known as *naive* T cells. After they have been activated by cells from the innate immune system, they differentiate and become effector cells. Some of these effector cells remain in the body even after the threat has been neutralized, conferring a *memory* to the immune system. Differentiated T cells that remain in the body after pathogen clearance are known as memory T cells.

The two main classes of T cells are distinguished by their CD4 or CD8 co-receptors. The T cells that carry CD8 co-receptors are known as *cytotoxic* cells. Their main function is to kill tumour cells, or cells infected by intracellular pathogens. They recognize target cells by the presentation of antigens on MHC class I molecules on the surface of infected cells. T cells that carry the CD4 co-receptor have a range of different functions in the immune system. They can be further classified into different modules that specialize in protection against particular types of pathogens; for instance, the T_H2 cells help control infection by extracellular parasites by directing the responses of eosinophils and mast cells – cells of the innate immune system that specialize in such attacks.

The receptors of B cells also undergo somatic recombination. In contrast to T cells, the B cells are able to secrete their receptors into tissue. Secreted B cell receptors are known as antibodies, and they carry out a wide range of important

functions. For example, they coat extracellular pathogens, which makes it easier for macrophages to destroy the pathogen. B cells differentiate into armed effector cells after encountering their antigen. Such differentiated cells can remain in the body and immediately start producing antibodies once the pathogen is encountered again. Together, this B cell and T cell mediated memory is the mechanism by which vaccines work.

Antibodies are also used in research to detect and count different surface membrane receptors. It is possible to construct antibodies that recognize a certain immune cell receptor. A fluorescent marker is then attached to the antibody. Particular cell types are recognized by the combinations of receptors that they carry on their surface membranes. By shining light of different wavelengths on tissue treated with antibodies of different immune cell receptor specificities, and with different fluorescent markers, it is possible to recognize and count the different immune cells and receptors. This process is known as flow cytometry. Our database of measurements on immunophenotypes was produced by this technique [Hasan et al., 2015].

The aim of the Milieu Intérieur study [Thomas et al., 2015] is to define the parameters that drive the naturally occurring variation in the healthy human immune system. To achieve this, the study has recruited 1,000 healthy volunteers, split evenly between men and women and in each decade between 20 and 70 years of age. Participants are all of at least three generations metropolitan French origin. A wide range of different datasets has been collected, primarily from whole blood, and curated for the study. In this thesis we investigate the environmental and genetic determinants of healthy human immune system variation by analysing a subset of these datasets, with information on:

- Life habits, collected through an extensive questionnaire.
- Genomic variability, collected by genome-wide SNP genotyping.
- Immunophenotypes, including variability in immune cell membrane-bound receptors and levels of circulating immune cell populations, collected by flow cytometry.
- T cell production in the thymus through T cell receptor excision circles (TRECs), which are small pieces of DNA left over after T cell receptor gene assembly.
- DNA methylation, the addition of methyl groups to cytosine bases of DNA that participate in the epigenetic control of gene expression.

Detailed information about the study, such as inclusion criteria and screening processes can be found in [Thomas et al., 2015]. The immunophenotype database is described in the methods section in first paper of the thesis [Patin et al., 2018] and can be downloaded from the *R* package that was built for the article [Bergstedt, 2018]. The development of the flow cytometry protocol is given in [Hasan et al.,

2015]. The SNP genotyping, the genotype quality control filtering and imputation is detailed in the methods section of the first article in the thesis [Patin et al., 2018]. Details about the processing of TRECs can be found in the materials and methods section of the second article of this thesis [Clave et al., 2018]. The DNA methylation data is used in the third article to estimate immune cell counts [Bergstedt et al., 2018]. More information about the DNA methylation dataset can be found in Paper III.

3.1 Immunophenotypes

Using flow cytometry, we enumerated all major subsets of immune cells in blood [Hasan et al., 2015]. Our data resource consists of 173 distinct immunophenotypes, with 83 immune cell levels, 88 cell-surface receptor levels, and 2 ratios between immune cell levels. The immunophenotypes cover all major functions of the immune system. They include, among others, the major subsets of the innate immune response, such as neutrophils, monocytes and natural killer cells (NK cells); subsets of CD4 and CD8 T cells at various differentiation stages, such as naive T cells and memory T cells; and different subsets of B cells.

The proportion of the cells in the blood that is made up by the various major subsets is visualized in Figure 3.1. The plot is a so-called *treemap*. The area of the rectangle for a particular immunophenotype is proportional to the mean number of observed cells in 100 μg of whole blood. There are more innate cells in the blood than adaptive, primarily because of the large amount of circulating neutrophils. T cells are more numerous than B cells, and out of the T cells, CD4 cells are more numerous than CD8 cells. Zooming in on the adaptive immune system in Figure 3.1, we see that the largest proportion of T cells in the blood are memory T cells (CM and EM stands for central memory and effector memory cells). In contrast, the largest subset for B cells are naive cells.

Histograms of innate and adaptive immune cell counts, sorted in descending order according to the mean, are shown in Figures 3.3 and 3.4. The distributions vary in skewness and there seems to be a tendency, particularly among adaptive immune cells, towards more skewed distributions for smaller subsets. Smaller subsets are often more differentiated, and have therefore been subjected to more binary lineage decisions (see Example 5), which have moved the distributions more towards log-normality. The tendency can be seen clearly in the histograms of $CD4^+$ and $CD8^+$ T cells at various differentiation stages, from naive cells to effector memory T cells re-expressing CD45RA (T_{EMRA}), shown in Figure 3.5.

Histograms of surface protein levels are shown for innate and adaptive cells in Figure 3.6 and Figure 3.7 respectively. Distributions for protein levels show less skewness, but, somewhat oddly, some of them have several peaks. These peaks could be an indication that the protein level are strongly influenced by common genetic variants.

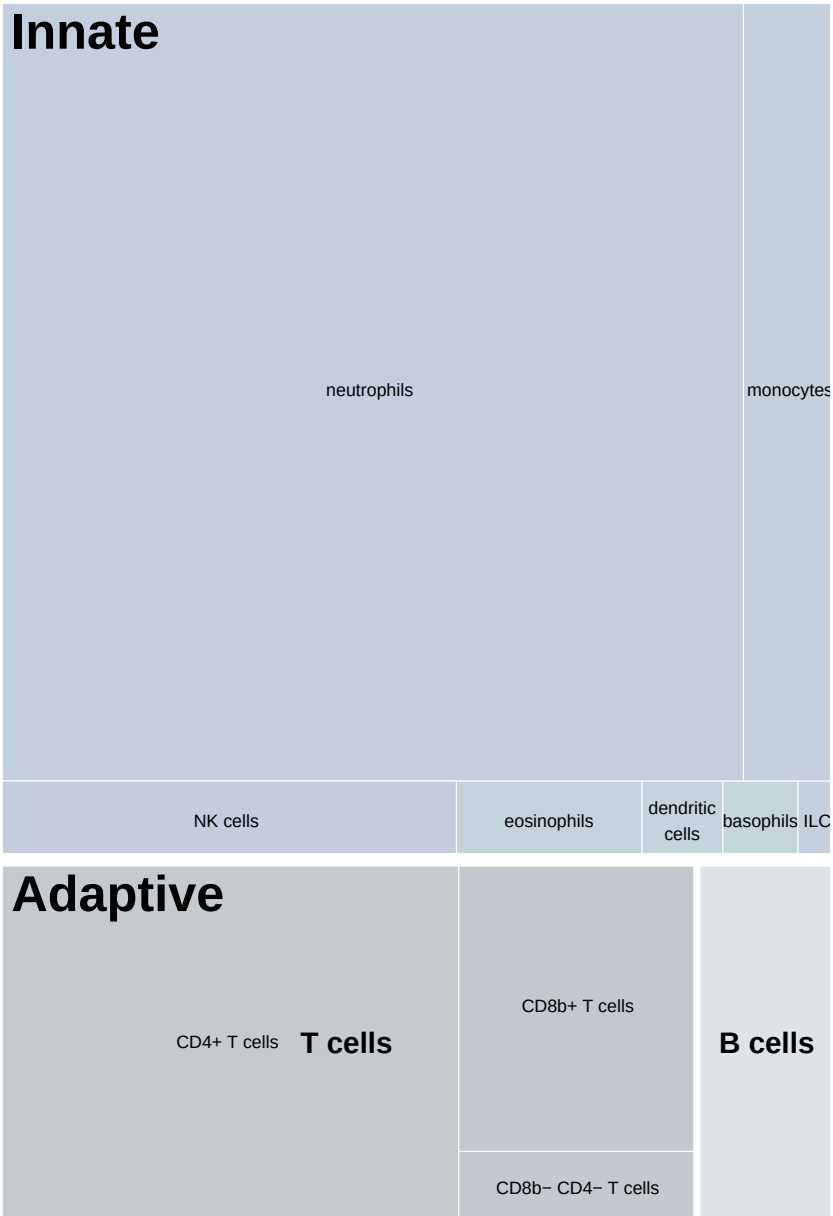

Figure 3.1 Proportion of major cell subsets in whole blood. The area of a rectangle is proportional to the mean number of cells in 100 µg of whole blood for the particular subset.

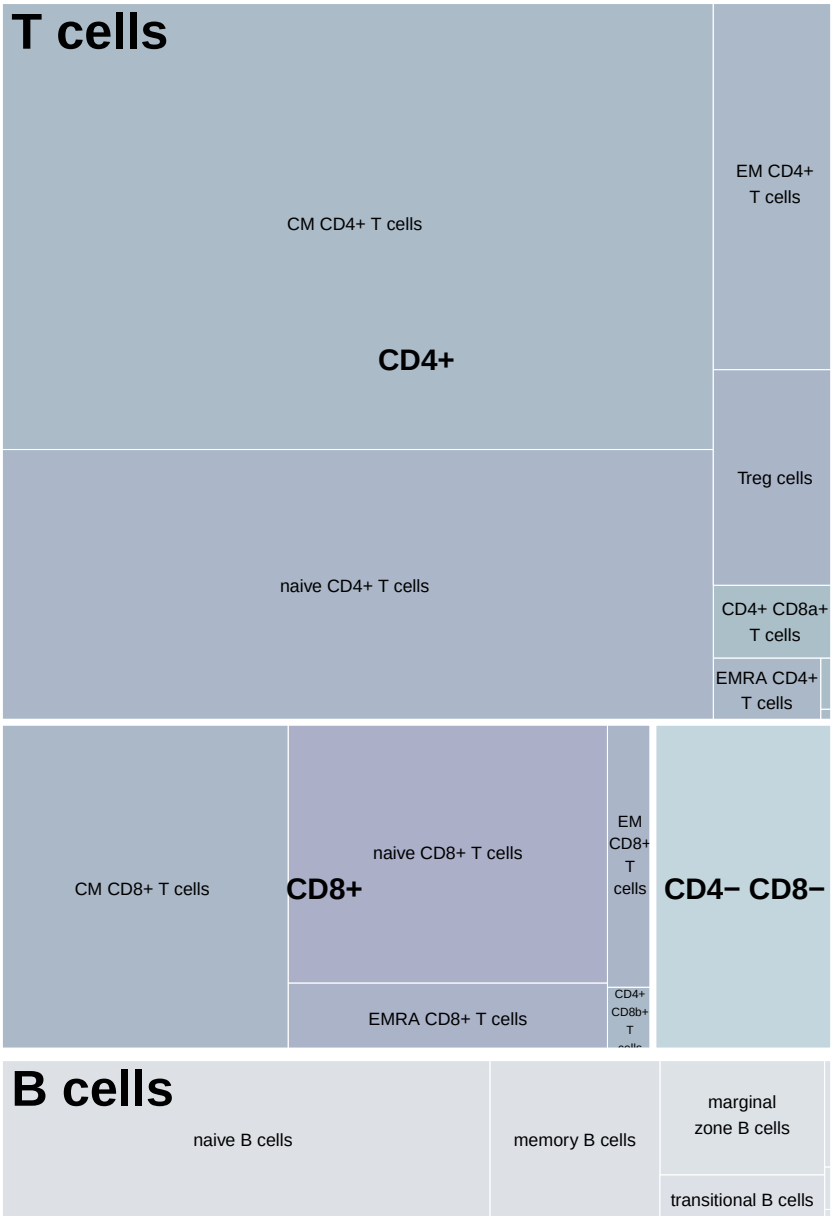

Figure 3.2 Proportion of adaptive immune cells in whole blood. The area of a rectangle is proportional to the mean number of cells in 100 µg of whole blood for the particular subset.

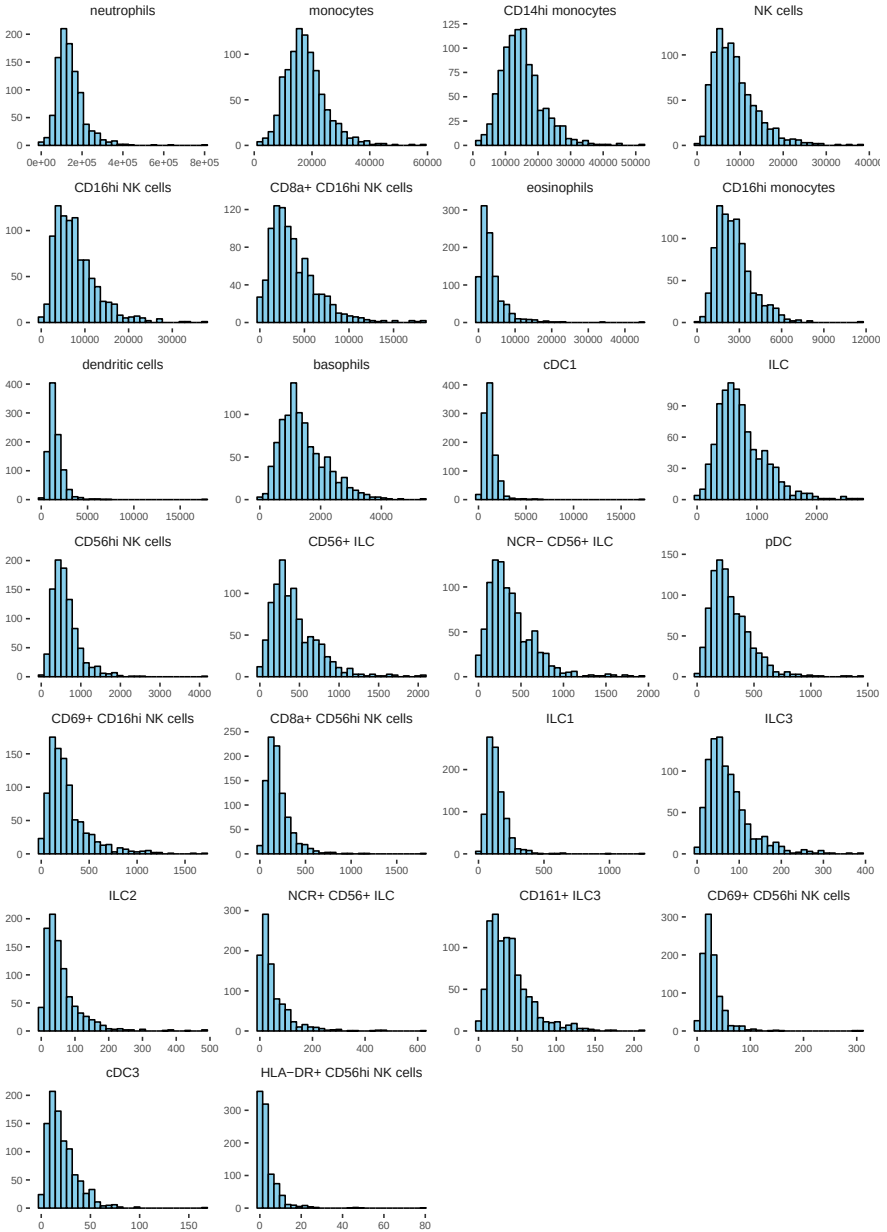

Figure 3.3 Histograms of counts of innate subsets in 100 µg of whole blood. All measurements of counts of all 26 innate cells are shown.

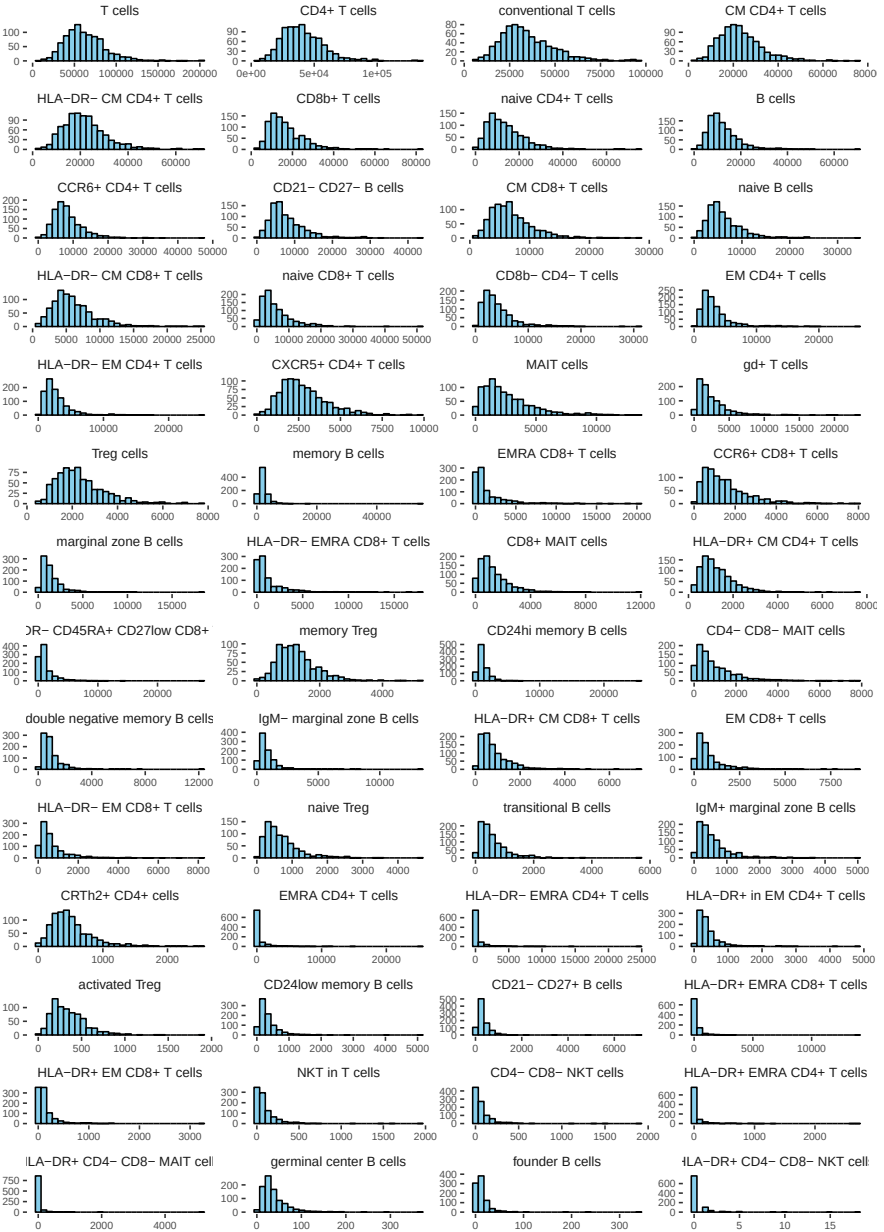

Figure 3.4 Histograms of counts of adaptive subsets in 100 µg of whole blood. All measurements of counts of all 56 adaptive cells are shown.

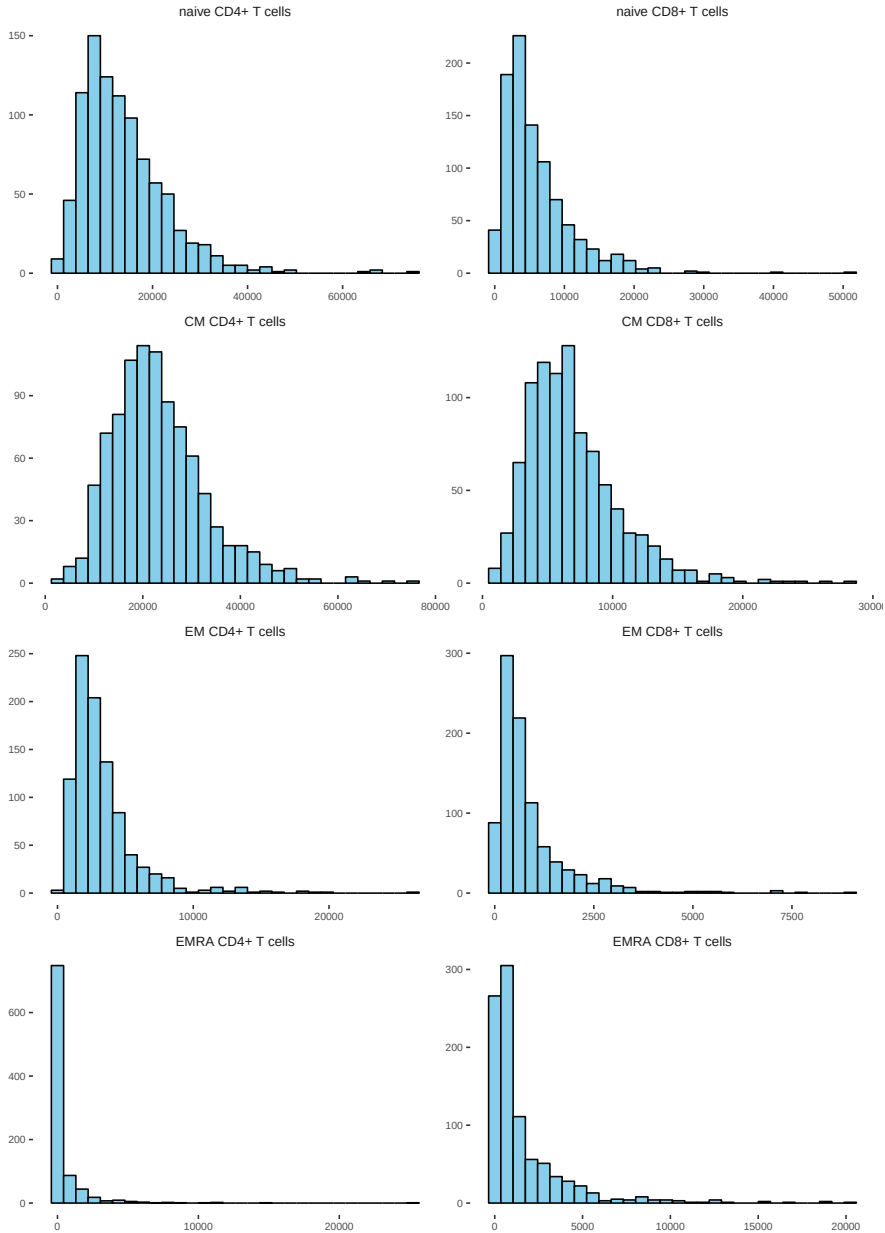

Figure 3.5 Histograms of counts in 100 µg of whole blood of T cells at different differentiation stages.

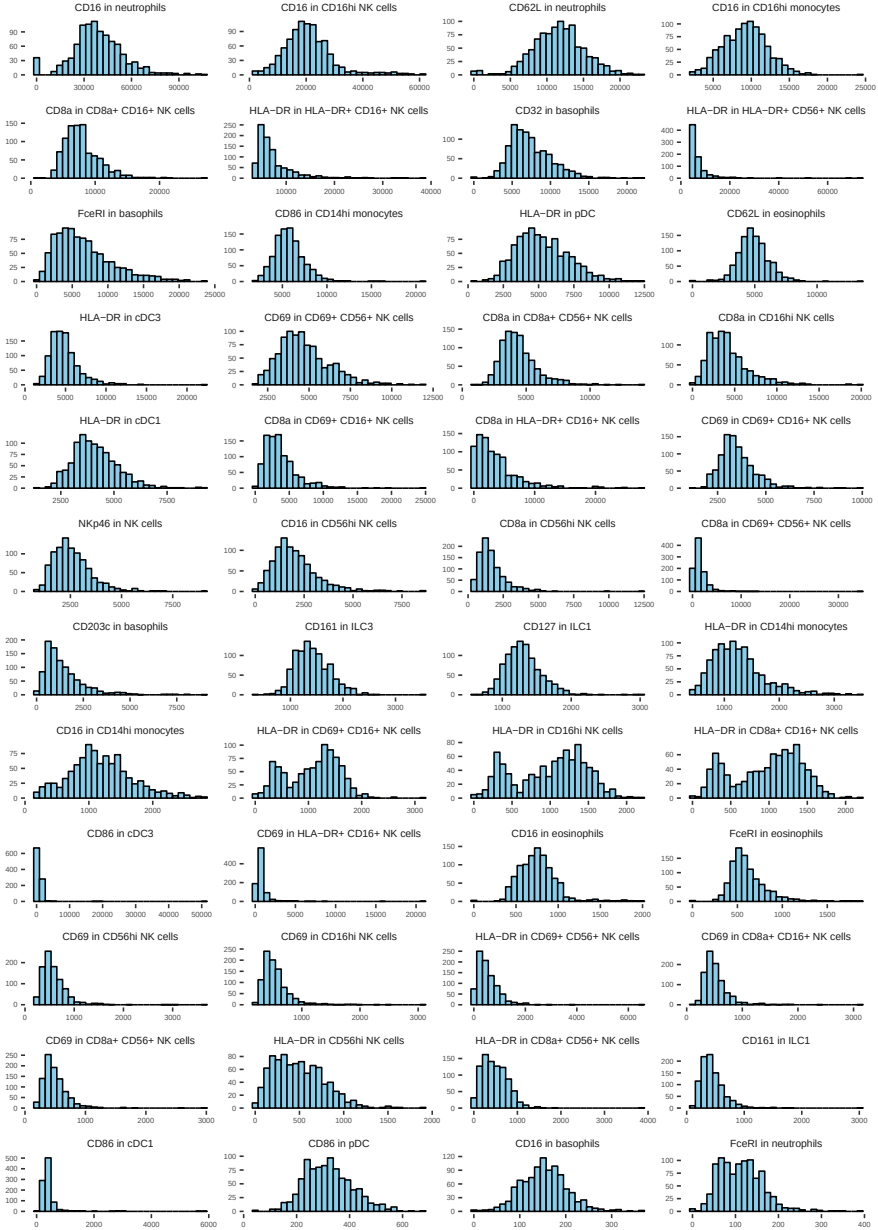

Figure 3.6 Histograms of mean fluorescence intensity [MFI] of surface membrane protein levels for innate cells. All measurements of all 48 MFIs are shown.

Figure 3.7 Histograms of mean fluorescence intensity [MFI] of surface membrane protein levels for adaptive cells. All measurements of all 39 MFIs are shown.

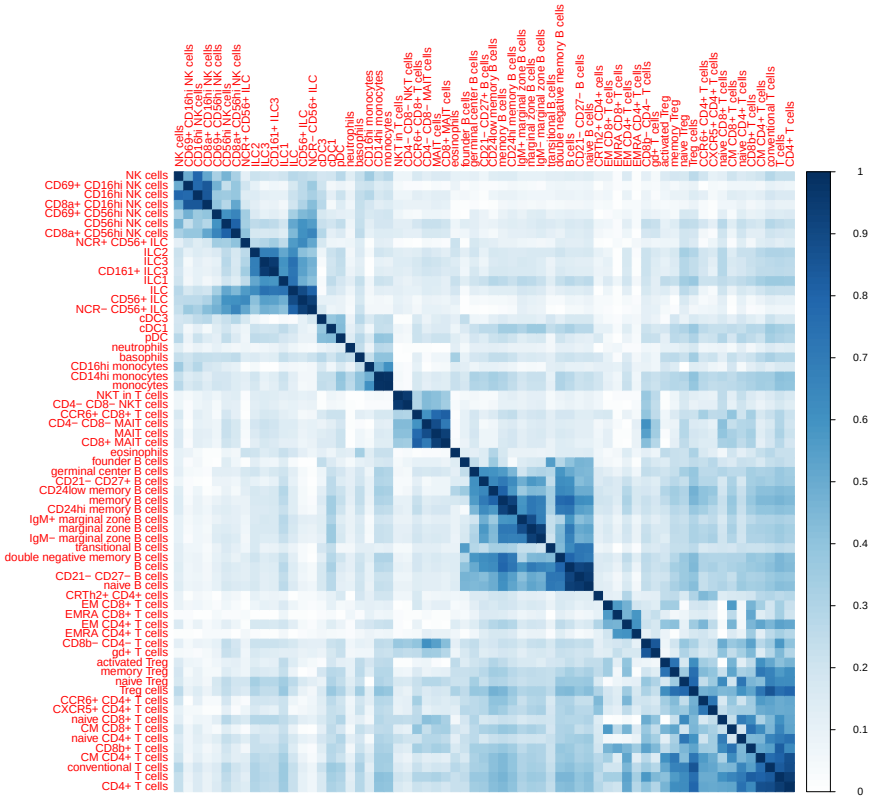

Figure 3.8 The absolute value of the Pearson correlation between log-transformed immune cell counts in 100 µg of whole blood. Cells are ordered to form hierarchical cluster that minimize the minimax linkage [Bien and Tibshirani, 2011].

The absolute value of Pearson correlations, $|r|$, of the log-transformed subset counts, ordered according to hierarchical clustering of $1 - |r|$ with minimax linkage [Bien and Tibshirani, 2011], are shown in Figure 3.8. The clustering reflects the common origin of various subsets. The innate cells of lymphoid lineage such as NK cells and ILCs are clustered together at the top. T cells with innate characteristics have a cluster, that consists of so-called natural killer T (NKT) cells and mucosal associated invariant T (MAIT) cells. The B cells have a cluster of their own. Finally, the conventional adaptive T cells are clustered at the bottom.

3.1.1 Batch effects

The flow cytometry was performed over almost a year between September 2012 and August 2013. Over 106 days of processing, between 1 and 12 samples where analysed per day (mean 9.43). There have been some previous reports of seasonal variation for immunophenotypes [Carr et al., 2016]. Flow cytometry is also a non-trivial technical procedure, relying on judgements of the operator. Producing consistent results over a whole year is difficult, particularly for protein surface MFIs, measurements that are known to be less robust. To investigate the variation of the flow cytometry measurements across days we fitted a linear mixed model, with the day of sampling as a varying-intercept random effect (see Section 2.3.2)

$$\begin{aligned} \log(Y_i) &= \mu + \text{SampleDay}_{d(i)} + \varepsilon_i \\ \text{SampleDay}_{d(i)} &\sim \mathcal{N}(0, \sigma_d^2) \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3.1}$$

where $d(i) : \{1, \ldots, n\} \to \{1, \ldots, 106\}$ is a function that maps from individual to the day that individual was processed. This is an example where the pooling of the mixed effects model (see Example 6) is absolutely necessary. Without it, we would have to estimate 106 means using only between one and twelve samples per mean.

The variance component σ_d^2 is a measure of the variation between days. The *intraclass* coefficient

$$\frac{\sigma_d^2}{\sigma_d^2 + \sigma^2} \tag{3.2}$$

is a measure of how strong the variation is across days relative to the variation across individuals. If the intraclass coefficient is high, then there is a large correlation between samples taken during the same day. Intraclass coefficients are shown for immune cell counts and surface proteins in Figure 3.9 and Figure 3.10 respectively. It is clear that there is a strong batch effect across days of sampling. Almost 90% of all variation in some surface protein MFIs can be attributed to it. The batch effect is particularly strong for MFIs, but some immune cell counts also show strong variation across days. Barplots across days for the four phenotypes with the highest intraclass coefficients are shown for cell counts in Figure 3.11 and for MFIs in Figure 3.12. The variation due to days for immune cell counts seem to stem mostly from outlier values, while the MFIs suffer from batch effects that show both seasonal variation and curious discontinuous jumps.

The batch effect has two negative consequences: 1) it induces correlation between samples taken during the same day, which means that models that does not account for that will have biased variance estimates, and 2) the variation across days is noise, which threaten to swamp the signals.

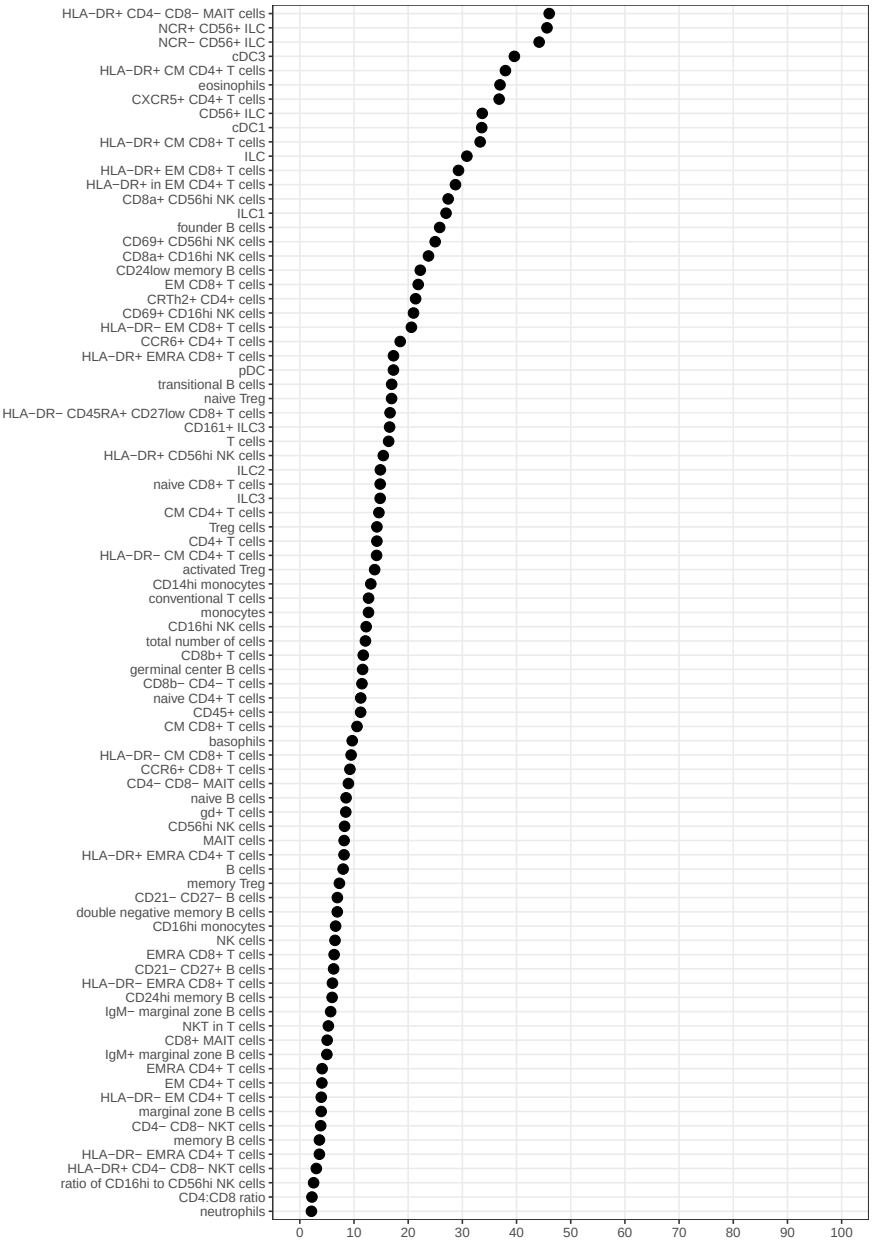

Figure 3.9 Estimated intraclass coefficients (3.2) for immune cell counts

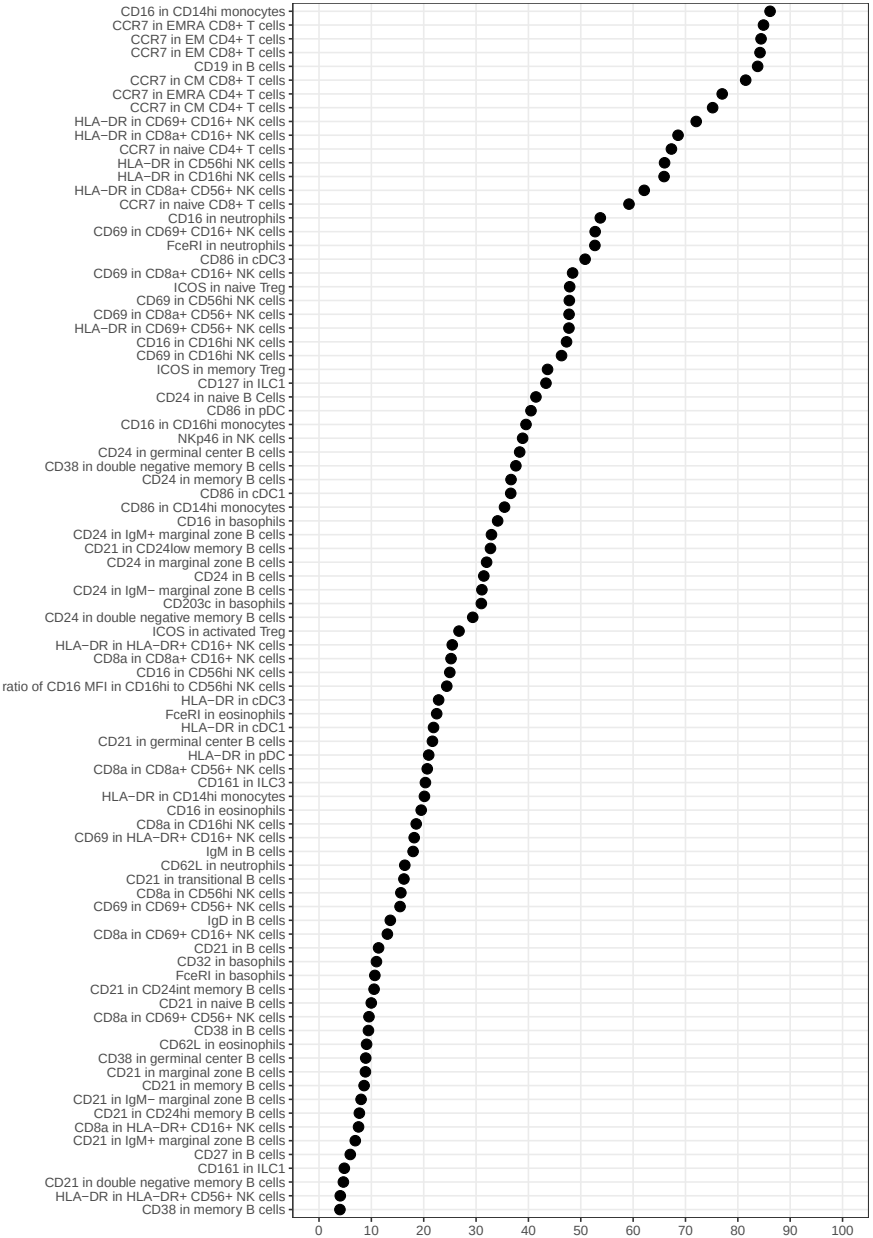

Figure 3.10 Estimated intraclass coefficients (3.2) for surface protein MFIs

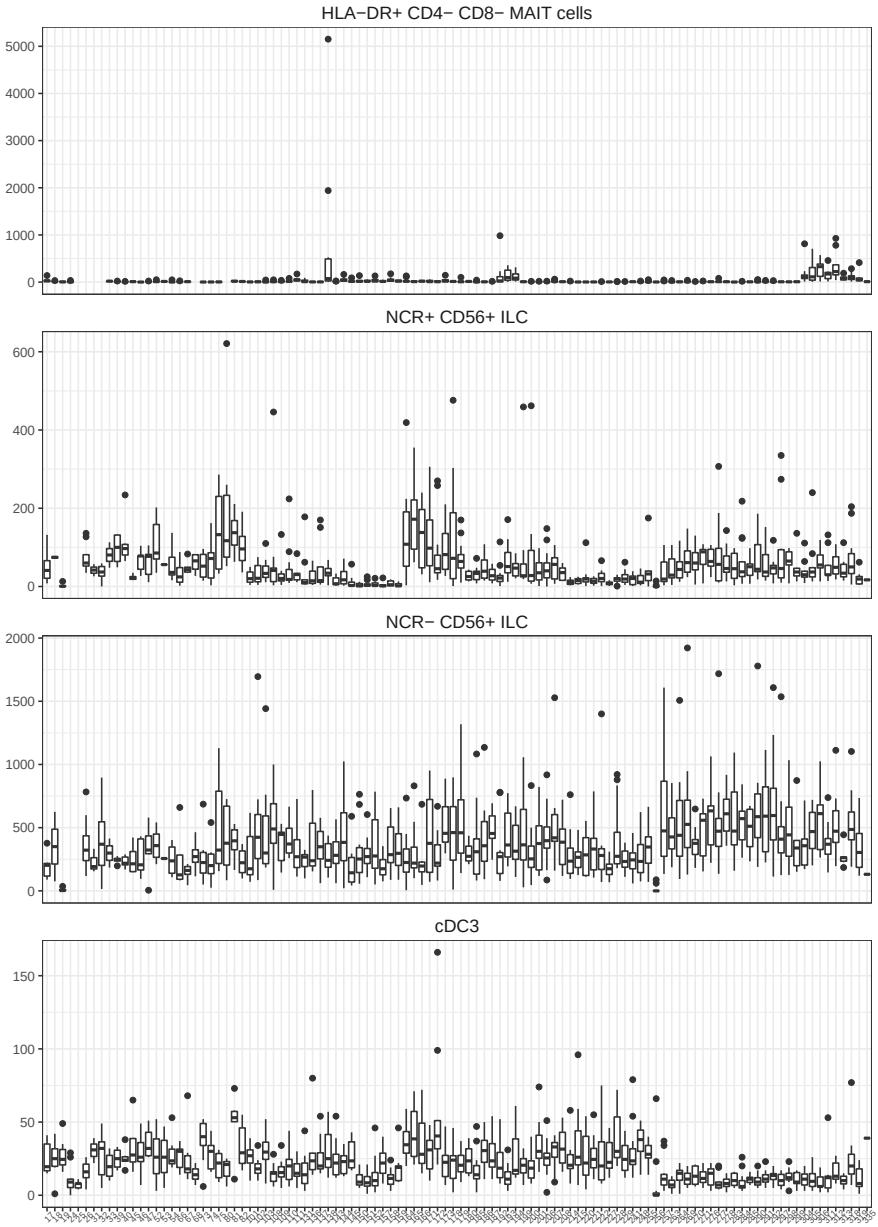

Figure 3.11 Distribution across days of the 4 cell counts with highest intraclass coefficient

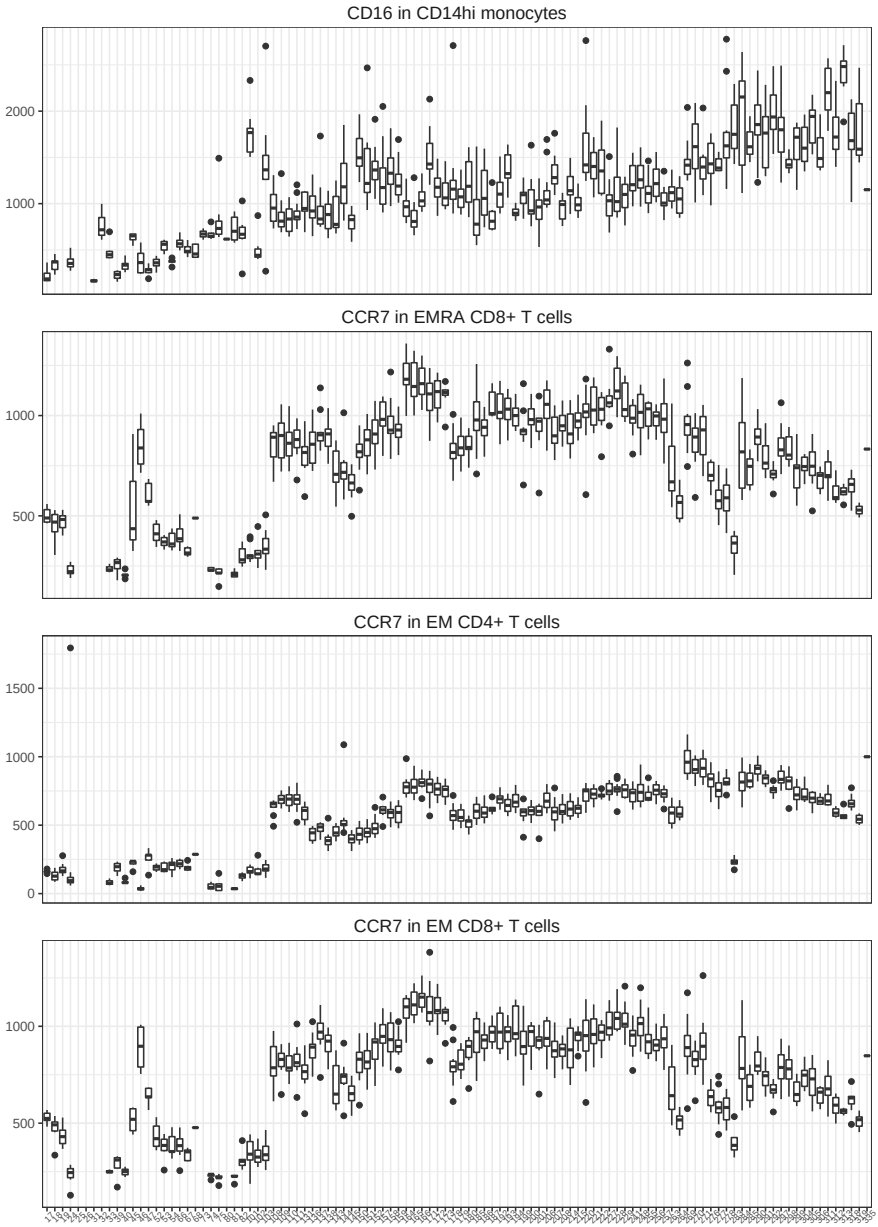

Figure 3.12 Distribution across days of the 4 MFIs with highest intraclass coefficient

3.2 Environmental variables

Subjects of the study answered a detailed questionnaire with questions about life-circumstances and life-habits. It contained detailed questions about, for example, socio-economic background, income and education, diet and eating habits, mental health, medication and drug use, physical activity and exercise. Study participants were also subjected to routine biochemical and hematological tests. The full database of environmental variables include hundreds of variables giving a granular view of the study participants. More details about the database are given in [Thomas et al., 2015].

3.3 Genotypes

The 1000 subjects of the cohort were genotyped at 966,431 SNPs, which after quality control filtering and imputation finally yielded a genotype database of 5,699,237 SNPs. Details can be found in the Supplementary material of Paper I [Patin et al., 2018].

4

GWAS pipeline

One of the main objectives of the Milieu Intérieur project is to map genetic determinants of human immune system variation. In the first article of the thesis [Patin et al., 2018], we mapped the genotypes of the cohort to 166 immunophenotypes (the final 7 immune cells of our database of 173 detailed in Chapter 3 were added after we had finished this work). In the second article [Clave et al., 2018], we mapped the genotypes to T cell receptor excision circles, a phenotype that we used as a proxy for thymic output.

Due to the number of phenotypes, it was not feasible to construct and fine-tune a GWAS model for each trait. However, just using the standard GWAS model (2.19) for each trait was not satisfactory, because the heterogenous character of the distributions of the immunophenotypes, as seen in Chapter 3, threatened to break its assumptions concerning mean-variance characteristics and linearity, and because we wanted to find SNPs that had an effect on the traits *independently* of environmental influences. Our large database of environmental variables offered an unprecedented opportunity to adjust for environmental context. To achieve this, we designed a GWAS pipeline with the objective that for any trait in our immunophenotype database, it should be able to 1) construct a GWAS model fulfilling mean and variance assumptions, and 2) adjust for relevant environmental factors, while still yielding reliable estimates. The pipeline is conceptually outlined in Algorithm 3.

Algorithm 3 GWAS pipeline

1: Detect and remove outliers.
2: Transform traits.
3: Detect and remove outliers on the transformed scale.
4: Impute traits using the missforest R package.
5: Adjust for day of sampling batch effect using the ComBat R package.
6: Select environmental covariates.
7: Run GWAS.

Computational and conceptual constraints do not allow current statistical methods to achieve these objectives while still being able to incorporate the variance of

all steps in the final estimates of the models. These estimates should therefore be interpreted as being conditional on the model selection process. Conceptually, we could get the unconditional estimates by integrating over the probability distribution of possible models. The difference between the conditional and the unconditional estimates is dependent on the variance of that distribution. We did not attempt to estimate the distribution of possible models, so we settled for conditional estimates. However, we strived to design each step in the pipeline to minimize the variance of the distribution.

Some of the phenotypes suffered from severe outliers, which can be seen in Figure 3.11. We therefore first needed to remove such outliers from all traits. To keep our data as pristine as possible, we devised an outlier removal scheme that was very conservative. In total, 200 observations were removed, out of 166×1000 observations. See the methods section of [Patin et al., 2018] for details.

4.1 Transformation of response variables

To be certain of fulfilling mean-variance assumptions, we needed to transform the response variables. For a GWAS, this is often done either by a non-parametric rank-based inverse normal transformation [Beasley et al., 2009], or by a parametric Box-Cox transformation, or its generalization, the Yeo-Johnson transformation. It is our belief that the rank-based inverse normal transformation is not suitable for our study. Since it is non-parametric, and since it does not make any a-priori assumption on the transformation it will estimate, it has high variance and is difficult to interpret. Another popular transformation scheme is the Box-Cox algorithm. Let $\text{gm}(\boldsymbol{y})$ denote the geometric mean of $\boldsymbol{y}$ and introduce the modified power family of transformations

$$\psi(\boldsymbol{y}, \lambda) = \begin{cases} \text{gm}(\boldsymbol{y})^{1-\lambda} \frac{(\boldsymbol{y}^\lambda - 1)}{\lambda}, & \text{if} \quad \lambda \neq 0 \\ \text{gm}(\boldsymbol{y}) \log(\boldsymbol{y}), & \text{if} \quad \lambda = 0. \end{cases} \tag{4.1}$$

Intuitively, it can be interpreted as the transformation

$$f(\boldsymbol{y}, \lambda) = \begin{cases} \boldsymbol{y}^\lambda, & \text{if} \quad \lambda \neq 0 \\ \log(\boldsymbol{y}), & \text{if} \quad \lambda = 0. \end{cases} \tag{4.2}$$

The additional terms in (4.1) are added to make the transformation continuous at zero and to ensure that the Jacobian of the transformation is the identity. The Box-Cox transformation estimates λ to make $\psi(\boldsymbol{y}, \lambda)$ as close to normally distributed as possible. However, since it is estimating a real-valued λ, the Box-Cox transformation can give rise to uncountably many different models. The procedure therefore threatens to severely increase the variance of the model selection distribution. It can also estimate nonsensical transformations with no reasonable interpretations. Both of these options therefore violate central tenets of our statistical design philosophy.

The histograms in Figures 3.3, 3.4, 3.6, and 3.7, show that the characteristics of the immunophenotype distributions approximately range from normal to log-normal. Since both of these distributions have "origin myths", see Example 5 and citelyon2013normal, that are applicable to biomolecular systems, we preferred to simply assume that the immunophenotypes are either normal or log-normal. Because of the transition between these states shown by the histograms, we also included a setting in-between, namely the assumption that $\sqrt{Y} \sim \mathcal{N}(\mu, \sigma^2)$. To further lend credence to these options, the three settings correspond to the natural assumptions that either $\mathbb{V}(Y) \propto 1$, $\mathbb{V}(Y) \propto \mu$ or $\mathbb{V}(Y) \propto \mu^2$ [Wakefield, 2013]. The transformation was chosen by the maximum-likelihood problem

$$\hat{\lambda} = \underset{\lambda \in \{0, 1/2, 1\}, \mu, \sigma}{\arg\max} \mathcal{N}\left(\psi\left(Y, \lambda\right); \mu, \sigma^2\right), \tag{4.3}$$

where $\mathcal{N}\left(y; \mu, \sigma^2\right)$ denotes the density function of the normal distribution.

We also wanted to leverage correlations among the immunophenotypes, shown in Figure 3.8, to impute missing phenotype values. The imputation was done by predicting missing values of a trait using a random forest model regressing non-missing values of the trait on all other traits. This was done using the missForest R package, see [Patin et al., 2018; Stekhoven and Bühlmann, 2011] for details.

4.2 Adjusting for day of processing

From Figure 3.10 and Figure 3.9 it is clear that we needed to adjust for day of sampling batch effects. Ideally, this would be done by having a varying-intercept term for each day, see (3.1), in the GWAS model. Unfortunately, this was not possible, because it would destroy the structure of the GWAS problem that current solvers use to make computations efficient. Given the scope of our GWAS, such structure was mission-critical. Instead, we estimated a mean for each day in a separate model and ran the GWAS software on the residuals of that model. The fitting of a mean for each day was done using the ComBat function in the sva *R* package [Leek et al., 2012]. It uses a multivariate model that assumes that the batch effects estimates can be pooled across days, but also across response variables. Let $\tilde{Y}_{ic}$ denote the transformed and imputed measurements of immunophenotype c for individual i. Furthermore, let $d(i)$ be a function that maps i to the day the ith individual was processed. We estimated the means for each day using the model

$$\begin{aligned} \tilde{Y}_{ic} &= \mu_c + \text{SampleDay}_{cd(i)} + \varepsilon_{ic} \\ \text{SampleDay}_{cd(i)} &\sim \Pi\left(\cdot\right) \\ \varepsilon_{ic} &\sim \mathcal{N}\left(0, \sigma_c^2\right). \end{aligned}$$

The distribution $\Pi\left(\cdot\right)$ is a non-parametric distribution that is estimated from the observations, see the supplementary material of [Johnson et al., 2007] for more

information. Assuming that all day of blood draw effects for all immunophenotypes are distributed according to the same distribution creates pooling in the estimates.

4.3 Covariate selection

Finally, we sought to select environmental covariates for each trait using our database over environmental variables. To make sure that we did not induce spurious collider bias, we excluded variables that had the potential of being downstream of the traits in the causal model of our system. We therefore did not include for instance hematological and some biochemical variables. Some factors that influence immune parameters are already well established. For instance, it is well-known that thymic production of naive T cells declines with age [Goronzy et al., 2015]. Another important factor for the immune system is cytomegalovirus (CMV) infection. During a life-time, a majority of the population becomes chronically infected by CMV. The infection is asymptomatic, but it is known to have a profound impact on memory T-cell parameter variation [Picarda and Benedict, 2018]. There is also previous evidence that sex influences some immunophenotypes, possibly because of hormonal differences [Liston et al., 2016]. Both age, sex and CMV are included in the covariate database.

Example 11 (Causal model of immunophenotype)
We now construct a tentative causal model for an immune cell count that is influenced by both age, sex and CMV. We assume that the other variables in the database potentially impact cell count, but not each other. We assume that m SNPs can have an impact on cell count. We want to find SNPs that does have an impact and quantify their effect. The SNPs are assumed to have no effect on anything but cell count. We further assume that ancestry has a potential effect on all SNPs and that is has the potential to influence the immunophenotypes, see Example 9. Since the screening rules of Milieu Intérieur make sure that all subjects have lived in metropolitan France for more than three generations, we deemed it less likely that ancestry would have a strong impact on the environmental covariates of our population. We therefore removed such links. It is reasonable to assume that age affects all non-genetic variables. We know that it affects CMV, because exposure to CMV increases with time. Sex is also assumed to impact the non-genetic variables, except CMV infection. We include the day of sampling batch effect, but assume that it does not impact any other variable than cell count. A causal model describing such a system is shown in Figure 4.1. □

We selected covariates to the GWAS model among 43 possible variables by the stability selection scheme, detailed in Algorithm 2, with the elastic net, described in Section 2.3.4, as support estimator. We used the two first principal components of the GRM matrix (see Section 2.3.3) as proxies for ancestry. All variables in the causal graph in Figure 4.1 are included, except batch, which had been adjusted for in a previous step. See Supplementary table 1 of the first article for details on

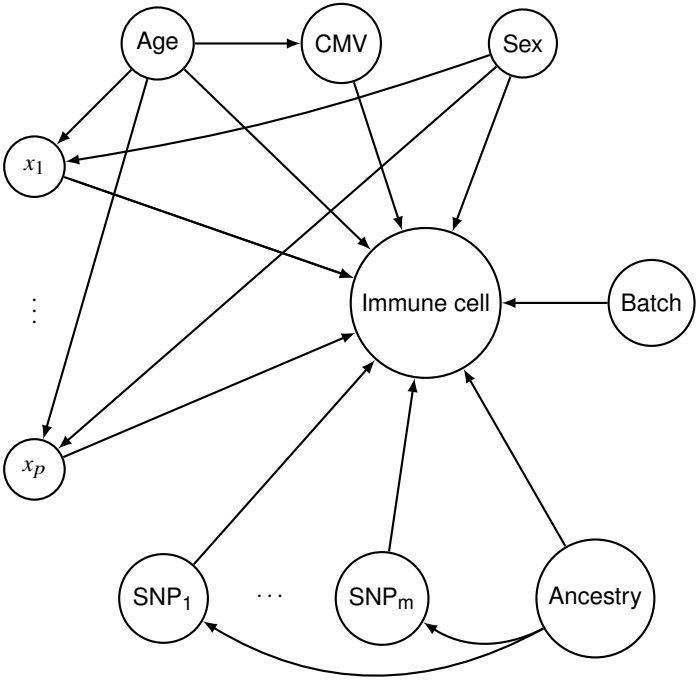

Figure 4.1 Causal model for the system described in Example 11

included covariates. Collect all of these predictors in the matrix X. Let $\tilde{Y}$ be a vector of flow cytometry measurements of an immunophenotype that have undergone the outlier-removal, transformation, imputation, and batch-adjustment steps described previously. The regression model that we used in the elastic net algorithm is given by

$$\mathbb{E}\left(\tilde{\boldsymbol{Y}} \mid \boldsymbol{x}\right) = \boldsymbol{x}\boldsymbol{\beta}. \tag{4.4}$$

Our intuition is that the variables with the strongest direct link to the immunophenotype will be included first in the model. Therefore, in the model depicted in Figure 4.1, the CMV, age and sex predictors should be the first to be included. If ancestry has a strong direct effect, it should also be among the first to be included. This would then break the forks between age, immune cell and $x_1, \ldots, x_p$, and sex, immune cell and $x_1, \ldots, x_p$. The additional variables should then be selected based on their direct link to the immunophenotype.

We show how many times a particular predictor was selected for the adaptive and innate immune cell counts in Figures 4.2 and 4.3, respectively. As expected, both age and CMV were selected for many of the adaptive immune cell counts. Out of the

54 such counts, age was selected 26 times and CMV was selected 25 times. A bit more surprisingly, sex was selected almost as many times, 24 out of 54. Out of the 25 innate cell counts, age was selected 11 times, and CMV was selected 8 times. All of these selections stem from natural killer cells and ILCs, innate cells that develop from the same lymphoid progenitor as T cells. Sex was selected frequently also in the case of innate cells, 10 out of 25 times. Of note is that CMV was not selected as often for innate cells. Strikingly, the predictor that was selected the most for both innate and adaptive cell counts is active smoking. Active smoking was selected 31 out of 54 times for adaptive cell counts, and 18 out of 25 times for innate counts. Although this might make common sense, smoking has not previously been implicated to that degree for impacting circulating immune cell variation. Other predictors that were selected are mostly binary dummy variables related to vaccination and childhood diseases. These are probably selected because they pick up nonlinear age effects that are not well adjusted by the linear age term, see Section 2.4.2 for a brief discussion.

4.4 The final GWAS model

Let for some immunophenotype $\boldsymbol{Y}$ be its flow cytometry measurements, that have undergone the outlier removal, transformation, imputation and adjustment steps described above, and $\boldsymbol{x}$ be the selected covariates. We want to investigate the effect of $\mathbf{SNP}_s$ on the immunophenotype. Instead of denoting the GWAS random effects term $\boldsymbol{g}$ as in Section 2.3.3, we rename it it to **Ancestry** in light of Example 9. Furthermore, let $\boldsymbol{G}_{-c}$ be the GRM approximated by the Milieu Intérieur genotypes using (2.20), excluding the chromosome of $\mathbf{SNP}_s$. The final GWAS model for the sth polymorphism is then

$$\begin{aligned} \boldsymbol{Y} &= \mu \mathbb{1} + \boldsymbol{x}\boldsymbol{\beta} + \mathbf{SNP}_s\beta_s + \mathbf{Ancestry} + \boldsymbol{\varepsilon}, \\ \mathbf{Ancestry} &\sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{G}_{-c}\right) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_n\right), \end{aligned} \tag{4.5}$$

which is of the same form as (2.19). The models for all SNPs were fitted using the *GEMMA* software [Zhou and Stephens, 2012]. We decide that an association between an immunophenotype and a SNP is discovered if the likelihood ratio test for the hypothesis $\beta_s = 0$ has a P value below 10^{-10}.

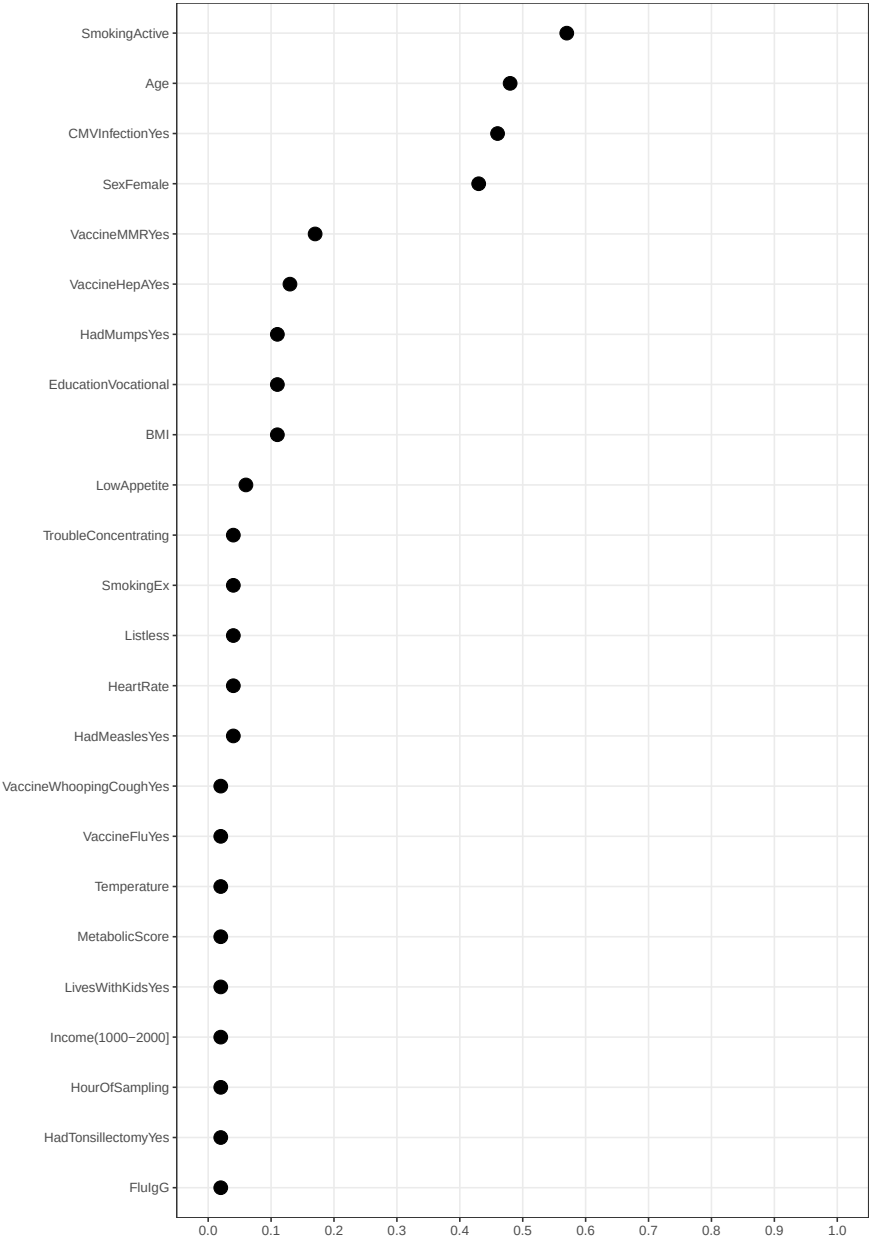

Figure 4.2 Percentage of times predictors were selected for adaptive immune cell counts. Only predictors that were selected at least once are shown.

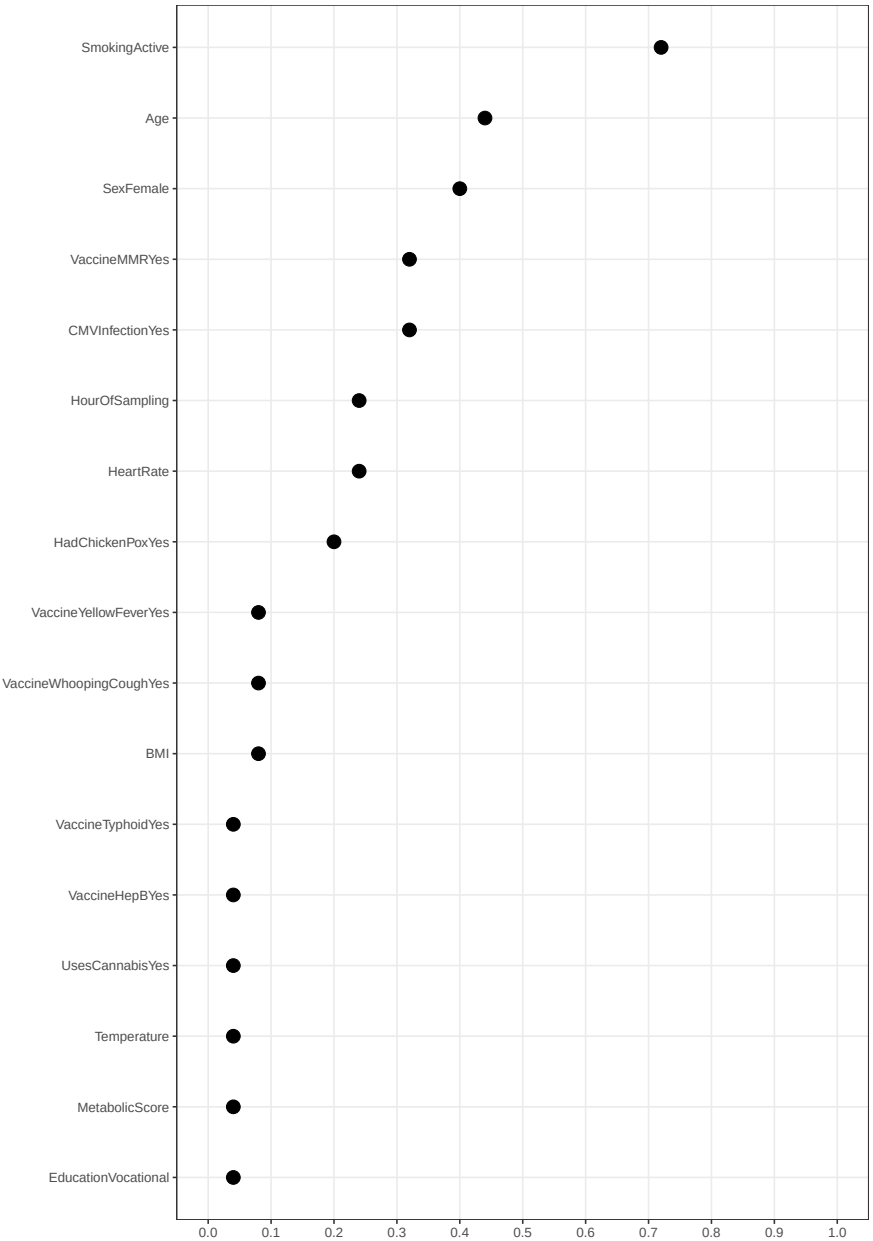

Figure 4.3 Percentage of times predictors were selected for innate immune cell counts. Only predictors that were selected at least once are shown.

5

Exploring environmental determinants of immune system variation

The comprehensive databases of immunophenotypes and non-genetic variables collected for the Milieu Intérieur study, and its large sample size, provided us with an unique opportunity to map the quantitative impact of intrinsic factors, like age and sex, and extrinsic factors like CMV infection, BMI and smoking habits, on the immune system. In contrast to our mapping of the impact of SNPs on the immunophenotypes, we firmly put the effect size in focus of this investigation. Many of the non-genetic variables that we were investigating have been studied in a wide range of biomedical and clinical settings, and their importance is widely appreciated. We therefore wanted the results of the study to be clear quantitative measures that are easy to interpret and apply, both in research and in the clinics. The scope of the study made it possible to compare different areas of the immune system, and different differentiation stages within the same area. To facilitate comparison, we wanted to ensure that the quantitative results of the study were on the same scale. We chose therefore not to transform the immunophenotypes similarly to what was done in Chapter 4. Instead, we log-transformed all immunophenotypes to keep the effect size as a multiplicative proportion, see Example 5. That option was more safe with this analysis, because the fewer amount of models required in comparison to the GWAS made it easier to graphically check and understand the behaviour of each model.

5.1 Causal model

Because the inclusion criteria of Milieu Intérieur included only people who was of more than two generations metropolitan French origin, we deemed it less likely that ancestry would have a strong effect on any of our environmental variables,

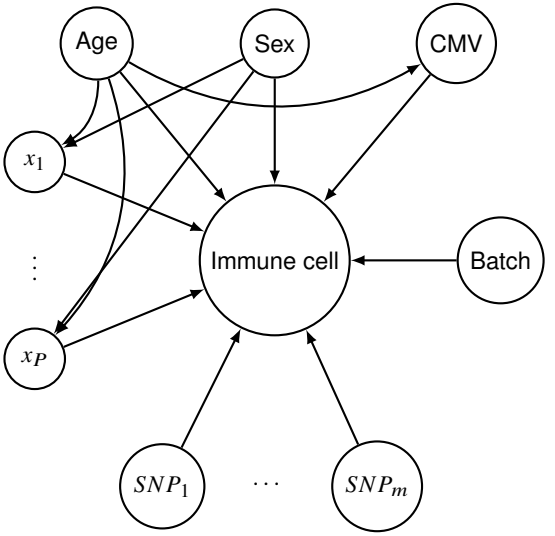

Figure 5.1 Assumptions for the mapping of the impact of environmental variables on immunophenotypes

except possibly for variables like height and BMI. Therefore, we did not take ancestry into account for this analysis. We assumed that m SNPs affect the immune cell. The assumptions for CMV infection, sex, and age mirror those for the causal model in Figure 4.1. Furthermore, we assumed that age and sex have an influence on all environmental variables, but we thought it less likely that the probability of being infected by CMV is strongly influenced by any variable except age, or that being infected by CMV has a strong influence on any variable besides possibly immunophenotypes. For simplicity, we assumed that there are no strong links between all the other environmental variables than CMV, age and sex. This assumption can easily be revisited in detail to judge the validity of a discovered strong link between a non-genetic predictor and an immune cell. The causal model corresponding to these assumptions is shown in Figure 5.1.

5.2 Statistical design

After removing the two proxy variables for ancestry and batch variables, 39 environmental variables were left. We wanted to investigate the causal links between each of these variables and the 163 immunophenotypes, a total of $39 \times 163 = 6357$ causal effects. Whether a causal link exists between predictor x and immunophenotype y was decided by hypothesis testing. All 6357 tests were considered to be one family of tests. There was a strong reason to believe, a priori, that at least some of the en-

vironmental predictors actually has an impact on a number of immunophenotypes. In particular, this is true for CMV infection, sex and age, where there is previous evidence of a causal link. But we suspected that also variables like smoking and BMI have an effect. Therefore, we used the false discovery rate as the error rate for the family (see Section 2.2.4). We decided that a causal link between an environmental predictor and an immunophenotype exists if the test testing their association rejected the null hypothesis with $\alpha = 0.01$. The chosen α is slightly more stringent than the usual 0.05 because of the complexity of the analysis. If we decided that a causal link exists between variable x and immunophenotype y, then we constructed a confidence interval for the strength of the link. This confidence interval was constructed to control the false coverage rate (see Section 2.2.5) at 0.05.

5.3 Statistical model

As a proxy for the m causal genetic variants in Figure 5.1, we used the genome-wide significant SNPs discovered using the GWAS pipeline developd in Chapter 4. Collect these SNPs in the set Ω. Consider an environmental variable $x \in \{x_1, \dots x_p\}$. To break the paths $x \to \text{Age} \to \text{Immune cell}$, $x \to \text{Age} \to \text{CMV} \to \text{Immune cell}$ and $x \to \text{Sex} \to \text{Immune cell}$, we included age, sex and CMV infection as covariates in the model. Similarly to (3.1), we used a random effects term for the "day of processing" batch variable. We saw that also the hour of processing for the sample had an affect on some immunophenotype measurements. For this analysis, we included a term adjusting for that for all models. Let $\boldsymbol{Y}$ be outlier-removed flow cytometry values for one of the immunophenotypes. As before, let $d(i) : \{1, \dots, n\} \to \{1, \dots, 106\}$ be a function that maps i to the day the ith individual was processed. The final model of the imune cell count for the ith individual is

$$\begin{aligned}
\log Y_i &= \mu + x_i\beta + \text{CMV}_i\beta_{\text{CMV}} + \text{Sex}_i\beta_{\text{Sex}} + \text{Age}_i\beta_{\text{Age}} \\
&\quad + \sum_{m \in \Omega} SNP_{im}\beta_m + \text{SampleHour}_i\beta_h + \text{SampleDay}_{d(i)} + \varepsilon_i \\
\text{SampleDay}_{d(i)} &\sim \mathcal{N}\left(0, \sigma_d^2\right) \\
\varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2\right),
\end{aligned} \tag{5.1}$$

which is of the varying-intercept form given in (2.12). The model was fitted using the *lmer* function from the *lme4* R package. The decision of whether or not a causal link exists between x and Y was done by testing the hypothesis $\beta = 0$ using the Kenward-Rogers test (briefly described in Section 2.3), implemented in the *pbkrtest* R package.

Naturally, the model used for CMV infection, age and sex is the same, but with the x term removed. Since the effect of age was estimated conditional on CMV serostatus, it is the direct effect of age on the immune cell that was estimated, not

the total effect, which would include the contribution mediated by CMV infection. Detailed results and a discussion can be found in the first article of the thesis [Patin et al., 2018].

5.4 Decomposition of variance

Our analysis showed that the environmental variables that primarily influence human immune system variation are CMV infection, age, sex and smoking status. These four factors have a strong and broad effect over the immune system. To compare the relative importance between these factors and the discovered SNP associations, we used linear regression models that for a particular immunophenotype include the four factors, and all genome-wide *suggestive* SNPs, defined to be those SNPs that have a P value below 5×10^{-8} for the test of association with the immunophenotype. Let Y_i be an immunophenotype and Ψ the index set of suggestive SNPs for it. The models are then given by

$$\begin{aligned} Y_i &= \mu + \mathrm{CMV}_i\beta_{\mathrm{CMV}} + \mathrm{Sex}_i\beta_{\mathrm{Sex}} + \mathrm{Age}_i\beta_{\mathrm{Age}} + \mathrm{Smoking}_i\beta_{\mathrm{Smoking}} \\ &\quad + \sum_{m\in\Psi} SNP_{im}\beta_s + \varepsilon_i \\ \varepsilon_i &\sim \left(0, \sigma^2\right), \end{aligned} \tag{5.2}$$

where $\left(0, \sigma^2\right)$ means any symmetrical distribution with mean zero and variance σ^2. We estimated the relative importance by the partial R^2, computed using the *relaimpo* R package [Grömping et al., 2006].

Bibliography

Ashley, E. A. (2016). "Towards precision medicine". *Nature Reviews Genetics* **17**:9, p. 507.

Baker, M. (2016). "1,500 scientists lift the lid on reproducibility". *Nature News* **533**:7604, p. 452.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). "Fitting linear mixed-effects models using lme4". *Journal of Statistical Software, Articles* **67**:1, pp. 1–48.

Bates, D., M. Maechler, B. Bolker, S. Walker, et al. (2014). "Lme4: linear mixed-effects models using eigen and S4". *R package version* **1**:7, pp. 1–23.

Beasley, T. M., S. Erickson, and D. B. Allison (2009). "Rank-based inverse normal transformations are increasingly used, but are they merited?" *Behavior genetics* **39**:5, p. 580.

Begley, C. G. and L. M. Ellis (2012). "Drug development: raise standards for preclinical cancer research". *Nature* **483**:7391, p. 531.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:1, pp. 289–300.

Benjamini, Y. and D. Yekutieli (2005). "False discovery rate-adjusted multiple confidence intervals for selected parameters". *Journal of the American Statistical Association* **100**:469, pp. 71–81.

Bergstedt, J. (2018). *MMI*. URL: `https://github.com/JacobBergstedt/mmi/` (visited on 2018-10-18).

Bergstedt, J., A. Urrutia, D. Duffy, M. L. Albert, L. Quintana-Murci, et al. (2018). "Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles". *bioRxiv*.

Bertsimas, D., A. King, R. Mazumder, et al. (2016). "Best subset selection via a modern optimization lens". *The annals of statistics* **44**:2, pp. 813–852.

Bien, J. and R. Tibshirani (2011). "Hierarchical clustering with prototypes via minimax linkage". *Journal of the American Statistical Association* **106**:495, pp. 1075–1084.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg.

Brynjolfsson, E. and A. Mcafee (2011). "The big data boom is the innovation story of our time". *The Atlantic*.

Cadwalladr, C. (2018). ""I made Steve Bannons psychological warfare tool": meet the data war whistleblower". *The Guardian*.

Candes, E. J. (2018). *Lecture notes in stats 300C: theory of statistics*.

Carr, E. J., J. Dooley, J. E. Garcia-Perez, V. Lagou, J. C. Lee, et al. (2016). "The cellular composition of the human immune system is shaped by age and cohabitation". *Nature immunology* **17**:4, p. 461.

Clave, E., I. L. Araujo, C. Alanio, E. Patin, J. Bergstedt, et al. (2018). "Human thymopoiesis is influenced by a common genetic variant within the TCRA-TCRD locus". *Science Translational Medicine* **10**:457.

Collins, F. S. and H. Varmus (2015). "A new initiative on precision medicine". *New England Journal of Medicine* **372**:9, pp. 793–795.

Davis, M. M. (2008). "A prescription for human immunology". *Immunity* **29**:6, pp. 835–838.

Davis, M. M. and P. Brodin (2018). "Rebooting human immunology". *Annual review of immunology* **36**, pp. 843–864.

Delude, C. M. (2015). "Deep phenotyping: the details of disease". *Nature* **527**, S14–S15.

The Economist (2017). "The worlds most valuable resource is no longer oil, but data". *The Economist*.

Einav, L. and J. Levin (2014). "Economics in the age of big data". *Science* **346**:6210.

Fan, S., M. E. Hansen, Y. Lo, and S. A. Tishkoff (2016). "Going global by adapting local: a review of recent human adaptation". *Science* **354**:6308, pp. 54–59.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Vol. 124. CRC press.

Fasman, J. (2018). "More data and surveillance are transforming justice systems". *The Economist*.

FitzGerald, G., D. Botstein, R. Califf, R. Collins, K. Peters, et al. (2018). "The future of humans as model organisms". *Science* **361**:6402, pp. 552–553.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA:

Gelman, A. (2018). *The competing narratives of scientific revolution*. URL: `https://andrewgelman.com/2018/08/20/competing-narratives-scientific-revolution` (visited on 2018-09-04).

Gelman, A. and J. Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gill, J. (2018). "Comments from the new editor". *Political Analysis* **26**:1, pp. 1–2.

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

Goronzy, J. J., F. Fang, M. M. Cavanagh, Q. Qi, and C. M. Weyand (2015). "Naive T cell maintenance and function in human aging". *The Journal of Immunology* **194**:9, pp. 4073–4080.

Grömping, U. et al. (2006). "Relative importance for linear regression in R: the package relaimpo". *Journal of statistical software* **17**:1, pp. 1–27.

GTEx Consortium (2017). "Genetic effects on gene expression across human tissues". *Nature* **550**, 204 EP -.

Halekoh, U. and S. Højsgaard (2014). "A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest". *Journal of Statistical Software* **59**:9, pp. 1–30.

Hasan, M., B. Beitz, V. Rouilly, V. Libri, A. Urrutia, et al. (2015). "Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping". *Clinical Immunology* **157**:2, pp. 261–276.

Hastie, T., R. Tibshirani, and R. J. Tibshirani (2017). "Extended comparisons of best subset selection, forward stepwise selection, and the lasso". *arXiv preprint arXiv:1707.08692*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hodges, J. S. (2016). *Richly parameterized linear models: additive, time series, and spatial models using random effects*. Chapman and Hall/CRC.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". *Nature genetics* **44**:8, p. 955.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Johnson, W. E., C. Li, and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* **8**:1, pp. 118–127.

Kenward, M. G. and J. H. Roger (1997). "Small sample inference for fixed effects from restricted maximum likelihood". *Biometrics*, pp. 983–997.

Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". *Bioinformatics* **28**:6, pp. 882–883.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, et al. (2011). "FaST linear mixed models for genome-wide association studies". *Nature methods* **8**:10, p. 833.

Liston, A., E. J. Carr, and M. A. Linterman (2016). "Shaping variation in the human immune system". *Trends in immunology* **37**:10, pp. 637–646.

McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**:4, pp. 417–473.

Morgan, S. L. and C. Winship (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Murphy, K. and C. Weaver (2016). *Janeway's immunobiology*. Garland Science.

Netea, M. G., L. A. B. Joosten, Y. Li, V. Kumar, M. Oosting, et al. (2016). "Understanding human immune function using the resources from the human functional genomics project". *Nature Medicine* **22**, 831 EP -.

Noakes, T., J. Borresen, T. Hew-Butler, M. Lambert, and E. Jordaan (2008). "Semmelweis and the aetiology of puerperal sepsis 160 years on: an historical review". *Epidemiology & Infection* **136**:1, pp. 1–9.

Park, T. and G. Casella (2008). "The Bayesian lasso". *Journal of the American Statistical Association* **103**:482, pp. 681–686.

Patin, E., M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, et al. (2018). "Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors". *Nature Immunology* **19**:3, pp. 302–314.

Pearl, J. (2009). *Causality*. Cambridge university press.

Picarda, G. and C. A. Benedict (2018). "Cytomegalovirus: shape-shifting the immune system". *The Journal of Immunology* **200**:12, pp. 3881–3889.

Quintana-Murci, L., A. Alcaïs, L. Abel, and J.-L. Casanova (2007). "Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases". *Nature immunology* **8**:11, p. 1165.

Recht, B. (2018). personal communication.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). "Marginal structural models and causal inference in epidemiology". *Epidemiology* **11**:5.

RStudio (2018). *Tidyverse*. URL: `https://www.tidyverse.org/` (visited on 2018-09-04).

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). "Semiparametric regression".

Ségurel, L. and C. Bon (2017). "On the evolution of lactase persistence in humans". *Annual review of genomics and human genetics* **18**, pp. 297–319.

Shah, R. D. and R. J. Samworth (2013). "Variable selection with error control: another look at stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**:1, pp. 55–80.

Shalizi, C. (2018). *Advanced data analysis from an elementary point of view*.

Stekhoven, D. J. and P. Bühlmann (2011). "MissForest – non-parametric missing value imputation for mixed-type data". *Bioinformatics* **28**:1, pp. 112–118.

Stolley, P. D. (1991). "When genius errs: RA Fisher and the lung cancer controversy". *American Journal of Epidemiology* **133**:5, pp. 416–425.

The Swedish Government (2017). *Action on digital transformation*. URL: `https://www.government.se/press-releases/2017/06/action-on-digital-transformation` (visited on 2018-09-04).

The 1000 Genomes Project (2015). "A global reference for human genetic variation". *Nature* **526**, pp. 68–74.

Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, et al. (2015). "The milieu intérieur study – an integrative approach for study of human immunological variance". *Clinical Immunology* **157**:2, pp. 277–293.

Tian, C., P. K. Gregersen, and M. F. Seldin (2008). "Accounting for ancestry: population substructure and genome-wide association studies". *Human molecular genetics* **17**:R2, R143–R150.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tibshirani, R. J., J. Taylor, et al. (2012). "Degrees of freedom in lasso problems". *The Annals of Statistics* **40**:2, pp. 1198–1232.

Torkamani, A., N. E. Wineinger, and E. J. Topol (2018). "The personal and clinical utility of polygenic risk scores". *Nature Reviews Genetics* **19**:9, pp. 581–590.

Trafimow, D. and M. Marks (2015). "Editorial". *Basic and Applied Social Psychology* **37**:1, pp. 1–2.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*. Springer Science & Business Media.

Ward, J., R. J. Strawbridge, M. E. Bailey, N. Graham, A. Ferguson, et al. (2017). "Genome-wide analysis in UK biobank identifies four loci associated with mood instability and genetic correlation with major depressive disorder, anxiety disorder and schizophrenia". *Translational psychiatry* **7**:11, p. 1264.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, et al. (2010). "Common snps explain a large proportion of the heritability for human height". *Nature genetics* **42**:7, p. 565.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011). "GCTA: a tool for genome-wide complex trait analysis". *The American Journal of Human Genetics* **88**:1, pp. 76–82.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price (2014). "Advantages and pitfalls in the application of mixed-model association methods". *Nature genetics* **46**:2, p. 100.

Zhang, C.-H., T. Zhang, et al. (2012). "A general theory of concave regularization for high-dimensional sparse estimation problems". *Statistical Science* **27**:4, pp. 576–593.

Zhou, X. and M. Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies". *Nature genetics* **44**:7, p. 821.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**:2, pp. 301–320.

Paper I

Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors

Etienne Patin* **Milena Hasan*** **Jacob Bergstedt***
Vincent Rouilly **Valentina Libri** **Alejandra Urrutia**
Cécile Alanio **Petar Scepanovic** **Christian Hammer**
Friederike Jönsson **Benoît Beitz** **Hélène Quach**
Yoong Wearn Lim **Julie Hunkapiller** **Magge Zepeda**
Cherie Green **Barbara Piasecka** **Claire Leloup** **Lars Rogge**
François Huetz **Isabelle Peguillet** **Olivier Lantz** **Magnus Fontes**
James P. Di Santo **Stéphanie Thomas** **Jacques Fellay**
Darragh Duffy **Lluís Quintana-Murci†** **Matthew L. Albert†**
The Milieu Intérieur Consortium

* These authors contributed equally † These authors jointly directed the work

Abstract

The quantification and characterization of circulating immune cells provide key indicators of human health and disease. To identify the relative effects of environmental and genetic factors on variation in the parameters of innate and adaptive immune cells in homeostatic conditions, we combined standardized flow cytometry of blood leukocytes and genome-wide DNA genotyping of 1,000 healthy, unrelated people of Western European ancestry. We found that smoking, together with age, sex and latent infection with cytomegalovirus, were the main non-genetic factors that affected variation in parameters of human immune cells. Genome-wide association studies of 166 immunophenotypes identified 15 loci that showed enrichment for disease-associated variants. Finally, we demonstrated that the parameters of innate cells were more strongly controlled by genetic variation than were those of adaptive cells, which were driven by mainly environmental exposure. Our data establish a resource that will generate new hypotheses in immunology and highlight the role of innate immunity in susceptibility to common autoimmune diseases.

Originally published in Nature immunology, March 2018. Reprinted with permission.

1. Main

The immune system has an essential role in maintaining homeostasis in people challenged by microbial infection, a physiological mechanism conceptualized by the French physician Claude Bernard in 1865, when he defined the notion of milieu intérieur [Bernard, 1865]. Hostpathogen interactions trigger immune responses through the activation of specialized immune cell populations, which can eventually result in pathogen clearance. The study of immune cell populations circulating in the blood provides a view into innate cells that are transiting between the bone marrow and tissues, and into adaptive cells that are recirculating through the lymphoid organs. Clinical studies of patients with past or chronic latent infection have reported profound perturbations in subsets of circulating immune cells due to altered trafficking, selective population expansion or attrition [Altfeld and Gale Jr, 2015; Orme et al., 2015]. However, several studies have suggested that extensive differences also exist among healthy people in the composition of their white blood cells [Tollerud et al., 1989; Reichert et al., 1991]. Evaluation of the naturally occurring variation in parameters of immune cells, together with environmental and genetic determinants of such variation, could accelerate the generation of hypotheses in basic immunology and ultimately improve the characterization of pathological states.

Population-immunology approaches, which compare immunological status across a large number of healthy people, have highlighted the predominant effect of intrinsic factors such as age and sex on the composition of human blood cells [Liston et al., 2016]. Several subpopulations of activated and memory T cells increase with age [Goronzy and Weyand, 2017], which might result in part from diminished thymic activity [Sauce and Appay, 2011] and might explain reduced vaccination efficacy in the elderly [Furman et al., 2014b]. Seasonal fluctuations in B cells, regulatory T cells (T_{reg} cells) and monocytes [Aguirre-Gamboa et al., 2016] and a strong effect of cohabitation on human immunological profiles [Carr et al., 2016] have been observed, which suggests that environmental exposure also drives variation in the immune system. For example, latent infection with cytomegalovirus (CMV), which is detected in 40% to >90% of the general population [Boeckh and Geballe, 2011], has been associated with an increased number of effector memory T cells [Wertheimer et al., 2014], which could in turn alter immune responses to heterologous infection [Furman et al., 2015]. However, the respective effects of age, sex and CMV infection on both innate cells and adaptive cells, as well as the precise nature of the environmental factors that affect variation in the immune system, are largely unknown.

Technological advances in flow cytometry, combined with genome-wide DNA genotyping, now allow delineation of the genetic basis of inter-person variation in the parameters of immune cells. A seminal genome-wide association study (GWAS) has identified 13 genetic loci strongly associated with the proportion of various leukocyte subpopulations in a cohort of 249 Sardinian families [Orrù et al., 2013]. Another study has reported deep immunophenotyping of ~1,800 independent traits in 245

healthy twin pairs, which has identified 11 independent genetic loci that account for up to 36% of the variation of 19 different traits [Roederer et al., 2015]. A third study has estimated the genetic heritability in the frequency of 95 different immune cells in 105 healthy twin pairs and has suggested that variation in immune cells is explained largely by non-heritable factors [Brodin et al., 2015]. Finally, four novel loci have been associated with B cell and T cell traits in a cohort of 442 healthy human donors in a study that delineated both non-genetic factors and genetic factors that affect immune cell traits that mediate adaptive immunity [Aguirre-Gamboa et al., 2016]. Together such studies have provided valuable insights into the contribution of genetic factors to inter-person differences in populations of adaptive immune cells, but they have largely neglected several major types of innate cells in the circulation. An integrated evaluation of the nature and respective effects of intrinsic, environmental and genetic factors that drive human variation in both innate immunity and adaptive immunity is thus lacking.

Here we report the use of standardized flow cytometry to comprehensively establish the composition of white blood cells from 1,000 healthy, unrelated people of Western European ancestry that compose the Milieu Intérieur cohort. We confirmed with this broad resource that age, sex, CMV seropositivity and smoking had major, independent effects on the parameters of innate and adaptive immune cells. We identified, through a GWAS, 15 loci associated with parameters of circulating leukocyte subpopulations, 12 of which were previously unknown. Finally, we found that cellular mediators of innate and adaptive immunity were affected differentially by non-genetic factors and genetic factors under homeostatic conditions.

2. Results

2.1 Variation in immune cell parameters in the general population.

The Milieu Intérieur cohort includes 500 men and 500 women stratified across five decades from 20 years of age to 69 years of age. Subjects were surveyed for various demographic variables, including past infections, vaccination and surgical histories, and health-related habits (Supplementary Table 1). Detailed inclusion and exclusion criteria used to define healthy subjects recruited into the cohort have been previously reported [Thomas et al., 2015].

To describe natural variation of both innate immune cells and adaptive immune cells in the 1,000 subjects, we used ten eight-color immunophenotyping flow-cytometry panels (Supplementary Figs. 1–10 and Supplementary Table 2), which allowed us to report a total of 166 distinct immunophenotypes (Supplementary Table 3). Our resource included 75 immunophenotypes obtained in innate immune cells (46%) and 91 immunophenotypes obtained in adaptive immune cells (54%). Innate cells were defined as those lacking somatic recombination of the genome [Vivier et al., 2011] and included granulocytes (neutrophils, basophils and eosinophils),

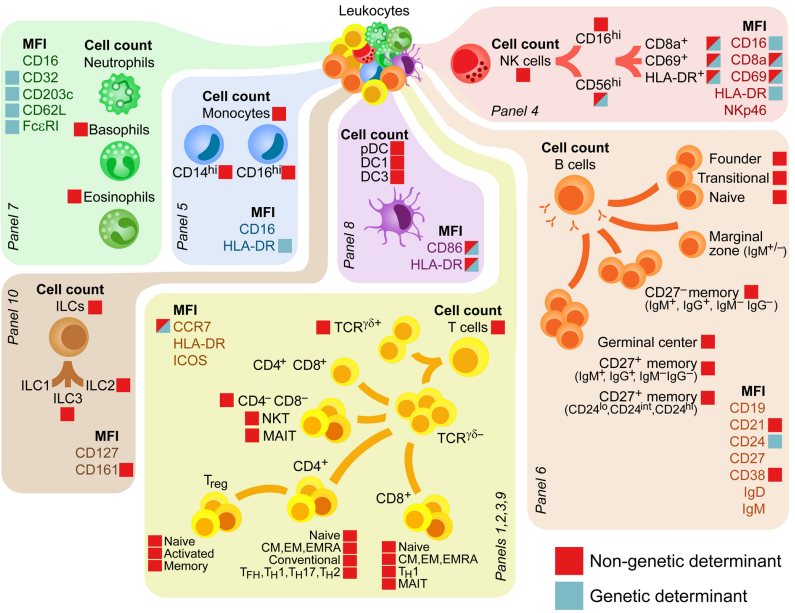

Figure 1. Quantification of immune cells and cell-surface markers measured in the Milieu Intérieur cohort. Strategy: flow cytometry was used to quantify (as MFI) the expression of phenotypic markers of differentiation or activation in cells of various lineages or differentiation states (interconnecting lines), as well as to quantify the cells themselves, for the identification of immunophenotypes significantly associated with non-genetic or genetic factors (key); numbers in parentheses (bottom left corners) indicate eight-color panels performed, grouped on the basis of cellular lineage (Supplementary Figs. 1–10 and Supplementary Tables 2 and 3). ILC1, ILC2 and ILC3, subsets of ILCs; T_{CM} cells, central memory T cells; T_{EM} cells, effector memory T cells; T_{FH}, T_H1, T_H17 and T_H2, subsets of helper T cells; NKp46, activating receptor; ICOS, costimulatory receptor.

monocytes, natural killer (NK) cells, dendritic cells (DCs) and innate lymphoid cells (ILCs) (Fig. 1). Adaptive cells were defined by their dependence on activity of the RAG1–RAG2 recombinase and included T cells ($\gamma\delta$ T cells, mucosa-associated invariant T cells (MAIT cells), NKT cells, T_{reg} cells and helper T cells) and B cells. The immunophenotypes of both innate immune cells and adaptive immune cells included 76 absolute counts of circulating cells, 87 expression levels of cell-surface markers (quantified as mean fluorescence intensity (MFI)), and 3 ratios of cell counts or MFI values (Supplementary Fig. 11 and Supplementary Table 3).

To reduce technical variation introduced by sample-temperature fluctuations and pre-analytical procedures, we strictly followed a standardized protocol for tracking and processing samples [Hasan et al., 2015]. Through the use of technical replicates, we verified that the immunophenotypes measured were highly reproducible (Supplementary Figs. 12 and 13 and Supplementary Table 3), which demonstrated the high precision of the data. We nevertheless identified two technical batch effects that affected flow-cytometry analyses. One effect corresponded to the hour at which the blood sample was obtained from fasting subjects (Supplementary Fig. 14a), which might possibly be explained by the spike in cortisol at the time of waking [Patterson et al., 2013]. The second effect corresponded to temporal variation of immunophenotypes over the 1-year sampling period, which did not follow the periodic distribution observed for cellular traits under seasonal fluctuations [Carr et al., 2016], and affected mainly measures of MFI (Supplementary Fig. 14b). We corrected for these batch effects in all subsequent analyses (Supplementary Fig. 15) and provide the distribution, ranges and statistics of all batch-corrected counts of immune cells (Supplementary Table 3), which should facilitate comparisons with cytometry data collected as part of routine clinical practice. This resource can be accessed through an online application (http://milieu-interieur.cytogwas.pasteur.fr/), which can be queried by personal characteristics such as age or sex.

Owing to the hierarchical structure of the differentiation of immune cells (i.e., cellular lineages emerge from common progenitor cells), a substantial portion of the counts of immune cells obtained in this study were highly correlated (Supplementary Fig. 16). These correlations were not directly attributable to the influence of factors such as age or sex, which were regressed out in this analysis. We observed correlations between the number of circulating ILC populations and that of NK cell populations, reflective of their common developmental pathway and dependence on γ_c cytokines [Serafini et al., 2015]. Likewise, MAIT cells and $CCR6^+CD8^+$ T cells were also correlated, owing to the formers being the major subset of $CCR6^+$ T cells in the circulation [Dusseaux et al., 2011]. Finally, we identified a strong correlation between the number of T_{reg} cells and that of conventional $CD4^+$ T cells, in confirmation of experimental work that defined a self-regulatory circuit driven by the cytokine IL-2 that integrates the homeostasis of these cell populations [Amado et al., 2013].

2.2 Effects of age, sex and CMV infection on parameters of innate and adaptive cells.

Published studies have shown that two intrinsic factors, age and sex, are responsible for inter-person variation in the composition of white blood cells [Liston et al., 2016; Goronzy and Weyand, 2017; Aguirre-Gamboa et al., 2016; Furman et al., 2015; Pennell et al., 2012; Furman et al., 2014a; Astle et al., 2016]. We used linear mixed models to quantify the respective effect of each of these intrinsic factors on variation in the composition of innate and adaptive cells. We observed a significant effect of age on 35% of the parameters of immune cells (adjusted P value, < 0.01; Fig. 2a and Supplementary Fig. 17a), among which only 29% were measured for innate cells. We detected a general decrease in the number of ILCs and plasmacytoid DCs (pDCs) and an increase in the number of $CD16^{hi}$ monocytes with increasing age (Fig. 2a), which might contribute to the altered immune response to viral infection in elderly people and age-associated inflammation [Furman et al., 2015; Della Bella et al., 2007; Puchta et al., 2016]. We found a modest increase in the number of memory T cells with age, in support of the view that the observed expansion of these cell populations in elderly subjects is not due to aging itself but to CMV seropositivity [Wertheimer et al., 2014], which we accounted for in the model. Our analyses also revealed that the number of naive $CD8^+$ T cells decreased more than twice as rapidly with age as the number of naive $CD4^+$ T cells did, at a rate of 3.6% per year (99% false-coverage rate (FCR)-adjusted confidence interval (99% CI): [3.0%, 4.1%]) and 1.6% per year (99% CI: [1.1%, 2.1%]), respectively (Fig. 2a–c), in support of the view that $CD8^+$ T cells are more susceptible to concentrations of homeostatic cytokines and/or that the production of $CD4^+$ T cells is 'preferentially' enhanced in the human thymus [Vrisekoop et al., 2008].

Although sex differences have been previously reported for various immune responses and diseases [Pennell et al., 2012], studies examining parameters of circulating cells have reported inconsistent results, owing to both differences in flow-cytometry procedures and relatively small, underpowered or poorly stratified study cohorts. We found a significant effect of sex on 16% of the immunophenotypes measured (adjusted P value, < 0.01; Fig. 2d and Supplementary Fig. 17b), of which 38% were measured in innate cells. We found a larger number of activated NK cells in men than in women. In contrast, MAIT cells were systematically greater in number in women, across all age decades (Fig. 2e–f), collectively suggestive of a lasting effect of early hormonal differences on the development and biology of immune cells.

Environmental exposures are also known to drive variation in the immune system, among which persistent infection with CMV is one of the strongest candidates [Liston et al., 2016; Wertheimer et al., 2014; Furman et al., 2015; Brodin et al., 2015]. We observed a significant effect of latent infection with CMV on 13% of the parameters of immune cells (Fig. 2g and Supplementary Fig. 17c), of which more than 75% were measured in adaptive cells. We confirmed that CMV triggered a major change

in the number of memory T cells, which was independent of age effects [Wertheimer et al., 2014; Brodin et al., 2015]. In particular, CMV seropositivity was associated with a 12.5-fold greater number of $CD4^+$ effector memory T cells that re-express the naive-cell marker CD45RA (T_{EMRA} cells) (99% CI: [8.8, 17.6]), and a 4.6-fold greater number of $CD8^+$ T_{EMRA} cells (99% CI: [3.5, 6.0]) (Fig. 2gi). However, we did not find evidence that CMV infection affected the number of cells in the naive T cell compartment or central memory T cell compartment. In support of that observation, the total number of $CD8^+$ T cells and $CD4^+$ T cells increased in parallel with the expanded number of memory T cells, suggestive of independent regulation of the naive T cell pool and the effector memory T cell and/or T_{EMRA} cell pool(s). CMV-seropositive donors also had lower numbers of circulating NKT cells and MAIT cells (Fig. 2g). Together our broad resource provided comprehensive quantification of the respective effects of age, sex and CMV infection on parameters of immune cells. Moreover, our results suggested a stronger effect of these factors on adaptive cells than on innate cells.

2.3 Tobacco smoking extensively alters the number of innate and adaptive cells.

Capitalizing on the detailed lifestyle and demographic data obtained for the Milieu Intérieur cohort, we evaluated the influence of additional environmental factors on parameters of immune cells with linear mixed models, controlling for the defined effects of age, sex, CMV serological status and batch effects. A total of 39 variables were chosen for analysis and tested for association with each immunophenotype. These included socio-economic characteristics, past infections, health-related habits, and surgery and vaccination history (Supplementary Fig. 18 and Supplementary Table 1). We identified a unique environmental factor that significantly altered the number of circulating immune cells: active smoking of tobacco cigarettes. This affected 36% of the immunophenotypes measured (Fig. 3a and Supplementary Fig. 19), of which 36% were measured in innate cells.

We observed a 23% greater number of circulating $CD45^+$ cells (99% CI: [11%, 37%]) and a 26% greater number of conventional lymphocytes (99% CI: [10%, 45%]) in smokers than in non-smokers (Fig. 3b). Published studies have suggested that smokers have alterations in circulating cell populations due to diminished adherence of leukocytes to blood-vessel walls, possibly as a result of lower antioxidant concentrations [Tsuchiya et al., 2002]. Furthermore, we found in active smokers a significant increase of 43% in activated T_{reg} cells (99% CI: [17%, 76%]) and 41% in memory T_{reg} cells (99% CI: [15%, 71%]), a pattern that was also observed, to a lesser extent, in past smokers (Fig. 3bd). Active smokers also showed a decreased number of NK cells, ILCs, $\gamma\delta$ T cells and various subsets of MAIT cells (Fig. 3b). These findings were consistent with a study showing that smoking triggers local release of IL-33 by the lung epithelium [Kearley et al., 2015], which in turn engages the IL-33 receptor ST2 on both innate lymphocytes and non-classical lymphocytes

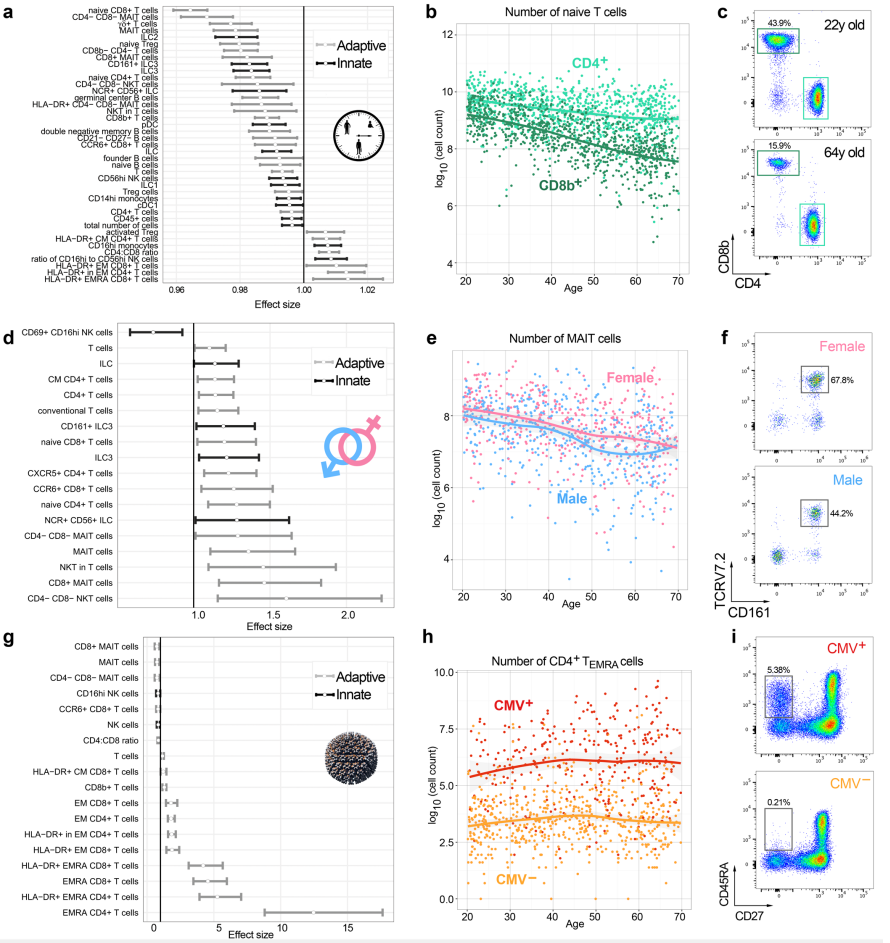

Figure 2. Effects of age, sex and CMV infection on the number of innate and adaptive cells in healthy people. a,d,g, Quantification of the effect of age **(a)**, sex **(d)** and CMV serostatus **(g)** on the abundance of circulating adaptive or innate immune cells (key; left margin) obtained from healthy donors ($n = 1,000$), estimated in a linear mixed model with a log-transformed immunophenotype as the response, controlled for batch effects and genome-wide significant SNPs, then transformed to the original scale (with 99% CIs adjusted for false coverage). **b**, Quantification of naive $CD8b^+$ or $CD4^+$ T cells (above plot) obtained from healthy donors (as in **a**,**d**,**g**) of various ages (horizontal axis), presented with regression lines fitted by local polynomial regression. **e**, Quantification of MAIT cells obtained from male or female (above plot) healthy donors (as in **a**,**d**,**g**) of various ages (horizontal axis), presented as in **b**.
Caption continues on the next page

Figure 2. *Caption continued:* **h**, Quantification of $CD4^+$ T_{EMRA} cells obtained from CMV^+ or CMV^- (above plot) healthy donors (as in **a**,**d**,**g**) of various ages (horizontal axis), presented as in **b**. **c**, Flow cytometry of naive T cells obtained from a donor 22 years of age or a donor 64 years of age (left margin). Numbers adjacent to outlined areas indicate percent $CD8b^+CD4^-$ T cells. **f**, Flow cytometry of naive T cells obtained from a female donor and a male donor (left margin). Numbers adjacent to outlined areas indicate percent $TCRV7.2^+CD161^+$ T cells (T_{EMRA} cells). **i**, Flow cytometry of naive T cells obtained from a CMV^+ donor and a CMV^- donor (left margin). Numbers adjacent to outlined areas indicate percent $CD45RA^+CD27^-$ T cells (MAIT cells). Effects on MFI, Supplementary Fig. 17.

[Monticelli et al., 2011]. Collectively, these findings revealed that active smoking had a profound effect on parameters of immune cells that was similar in magnitude to that of age, and that it affected both innate cells and adaptive cells.

2.4 GWAS of 166 parameters of immune cells.

To identify common genetic variants that affect inter-person variation in parameters of immune cells, we genotyped the Milieu Intérieur cohort at 945,213 single-nucleotide polymorphisms (SNPs) enriched for exonic SNPs. After quality control (Supplementary Fig. 20), genotype imputation was performed, which yielded a total of 5,699,237 highly accurate SNPs, which were tested for association with the 166 immunophenotypes by linear mixed models. The models were adjusted for the genetic relatedness among subjects and any non-genetic variable identified as being predictive of each specific immunophenotype by stability selection based on elastic net regression (Supplementary Table 3). We confirmed that we had the power to identify medium effect genotypephenotype associations by simulations and by empirically replicating well-known genetic associations with non immunological traits, such as eye and hair color or levels of uric acid and cholesterol.

In the context of immunological traits, we found 14 independent genetic loci associated with 42 of 166 immunophenotypes (25%), at a conservative genome-wide significant threshold of $P < 1.0 \times 10^{-10}$ (Fig. 4a, Table 1, Supplementary Fig. 21 and Supplementary Tables 4 and 5). We then conducted conditional GWAS by adjusting those 42 immunophenotypes on the 14 leading associated variants (Table 1) and found an additional independent locus that reached genome-wide significance (Supplementary Fig. 22 and Supplementary Table 6). Genome-wide significant associations were replicated in an independent cohort of 75 donors of European descent for all immunological traits measured in this replication cohort ($P < 0.05$; Table 1). Also, we confirmed that our measurements of immune cells were stable, as all genome-wide significant associations were confirmed for immunophenotypes measured in a sample of blood newly obtained from 500 of the 1,000 subjects of the Milieu Intérieur cohort, at 7–44 d after the initial visit ($P < 10^{-3}$; Table 1). We also provide a list of 26 suggestive association signals ($P < 5.0 \times 10^{-8}$), including various candidate genes encoding biologically relevant molecules (Supplementary Table 6).

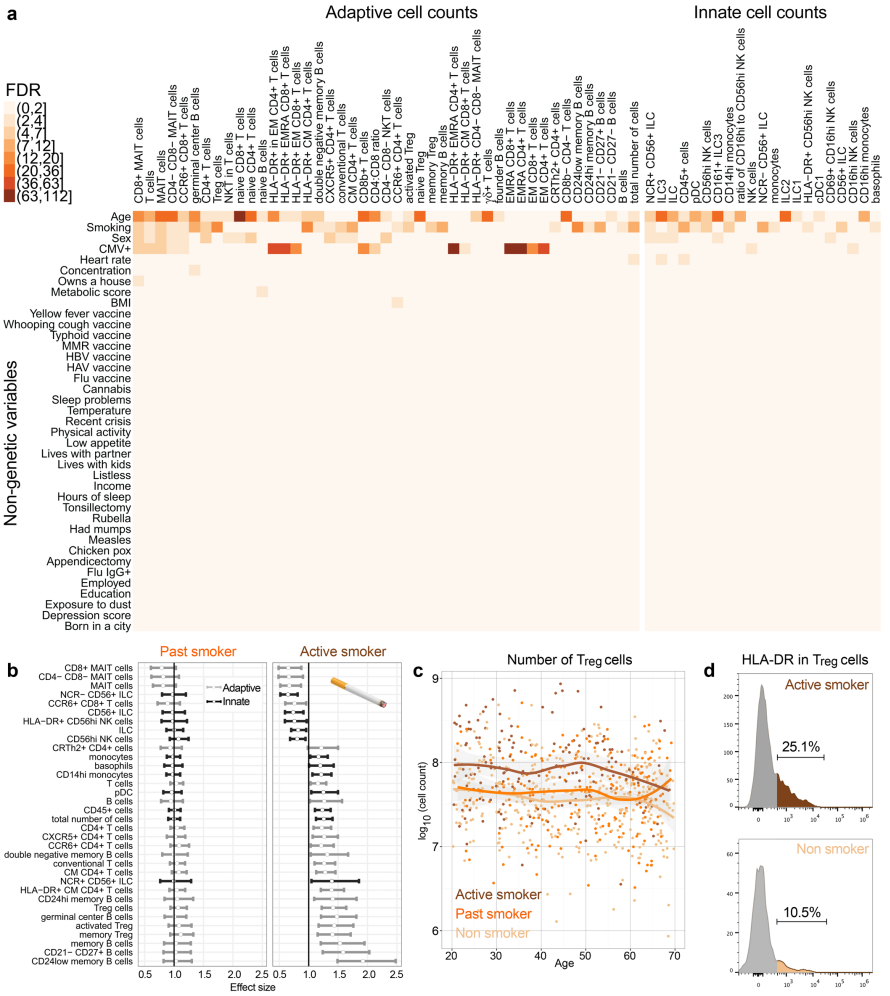

Figure 3. Effects of smoking on the number of innate and adaptive immune cells in healthy people. a, Association between 39 non-genetic factors (left margin) and the number of adaptive and innate cells (above plot) in healthy donors (n = 1,000), presented as $-\log_{10}$ of adjusted *P* values with a false-discovery rate (FDR) of <1%. We controlled for age, sex, and CMV status, except when their effects were specifically estimated, together with batch effects and genome-wide significant SNPs (Table 1). **b**, Quantification of the effect of past or active smoking (above plots) on the abundance of circulating adaptive and innate immune cells (key; left margin), with multiplicative effect sizes estimated in a linear mixed model with a log-transformed immunophenotype as response, controlled for age, sex, CMV serostatus, batch effects and genome-wide significant SNPs, then transformed to the original data scale (99% CIs adjusted for false coverage). *Caption continues on the next page*

Figure 3. *Caption continued:* **c**, Quantification of circulating T_{reg} cells in donors of various ages (horizontal axis) with a status of active smoker, past smoker or non-smoker (above plot), presented with regression lines fitted by local polynomial regression. **d**, Flow-cytometry analysis of HLA-DR in T_{reg} cells from an active smoker and a non-smoker (left margin). Gray shaded curves, HLA-DR^- T_{reg} cells. Numbers above bracketed lines indicate percent HLA-DR^+ T_{reg} cells (red or tan curve; effect of smoking on MFI, Supplementary Fig. 19).

The associated genetic loci showed enrichment for SNPs associated by GWAS with diseases (31% observed versus 5% expected; resampling *P* value, 0.0032), most of which were autoimmune diseases, including rheumatoid arthritis, Vogt-Koyanagi-Harada syndrome and atopic dermatitis (Supplementary Table 4). These findings highlighted the importance of the alteration of immune cell populations by genetic loci in the context of ultimate organismal traits that affect human health.

2.5 Genetic associations identify mainly immune cell-specific protein quantitative *trait* loci.

Of the 42 immunophenotypes for which a significant genetic association was detected, 36 (86%) were MFI measurements, which quantifies the cell-specific expression of protein markers conventionally used to determine the differentiation or activation state of leukocytes. For 28 of these 36 MFI measurements (78%), the genetic association was observed between the protein MFI and SNPs located in the vicinity of the gene encoding the corresponding protein (Table 1 and Supplementary Fig. 21); i.e., local protein quantitative trait loci (local-pQTLs). For example, genetic variation near *ENPP3* (which encodes the phosphodiesterase CD203c) was associated with the MFI of CD203c in basophils (rs2270089; $P = 2.1 \times 10^{-28}$); genetic variation near *CD24* (which encodes the B cell-differentiation marker CD24) was associated with the MFI of CD24 in marginal zone B cells (rs12529793; $P = 3.8 \times 10^{-21}$); and genetic variation near *CD8A* (which encodes the co-receptor CD8a) was associated with the MFI of CD8a in $CD69^+CD16^{hi}$ NK cells (rs71411868; $P = 5.9 \times 10^{-58}$).

We identified two independent local-pQTLs in the *FCGR* cluster (Table 1), which encodes the most important Fc receptors for inducing the phagocytosis of opsonized microbes. Genetic variation near *FCGR3A* was associated here with the MFI of the NK cell receptor CD16 (FcγRIII) in $CD16^{hi}$ NK cells (rs3845548; $P = 3.0 \times 10^{-87}$). The same variants were also shown to affect the number of $CD62L^-$ myeloid cDCs in a published study [Orrù et al., 2013]. The second signal-associated variation in *FCGR2B* was associated with the MFI of the NK cell receptor CD32 (FcγRII) in basophils (rs61804205; $P = 1.7 \times 10^{-36}$) but not in eosinophils or neutrophils. Consistent with that, it is known that basophils express both CD32a and CD32b, while eosinophils and neutrophils express mainly CD32a [Cassard et al., 2012]. Conversely, a local pQTL was identified at *SELL* (which encodes the adhesion molecule CD62L) that was associated with the MFI of CD62L in eosinophils and

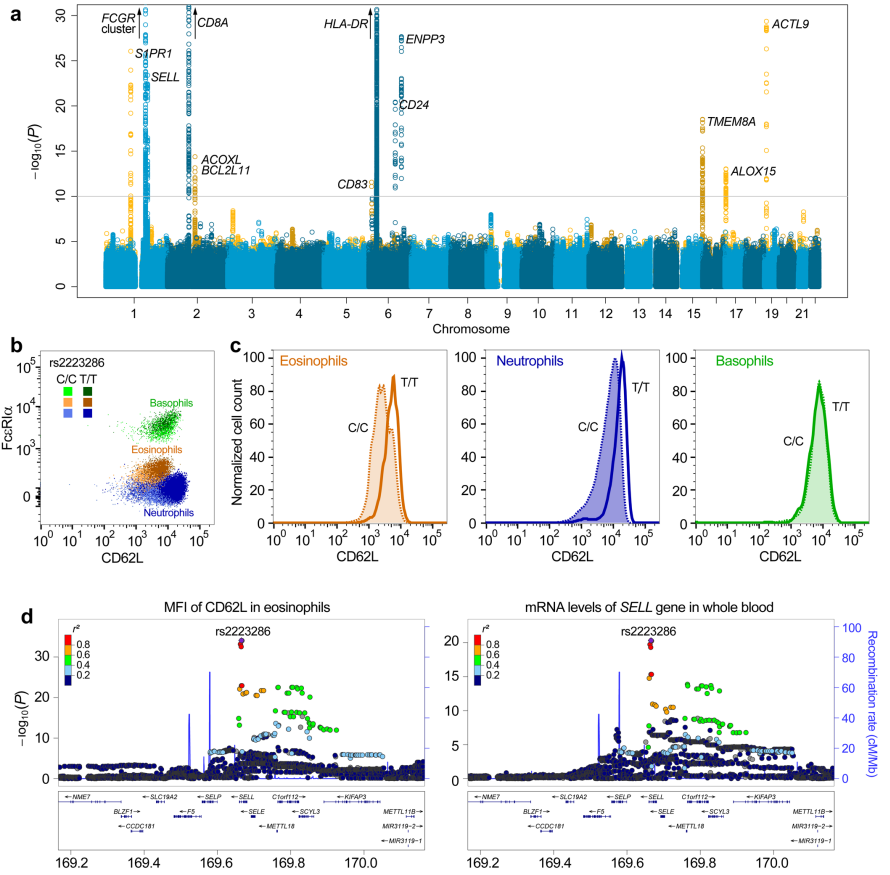

Figure 4. Genome-wide significant associations with 166 immunophenotypes measured in healthy people. a, Genome-wide significant associations with variants acting locally (local-pQTLs (blue)) or not (cell count QTLs or trans-pQTLs (yellow)) on immunophenotypes in healthy subjects ($n = 1{,}000$), presented as Manhattan plots (gray line, genome-wide significance threshold ($P < 1 \times 10^{-10}$); 'zoomed' Manhattan plots for all hits, Supplementary Fig. 21). **b**, Flow-cytometry analysis of the expression of FcεRIα and CD62L by various granulocytes (colors; key) of donors homozygous for the major rs2223286 allele (T/T) or minor rs2223286 allele (C/C) (color intensity; key). *Caption continues on the next page*

Figure 4. *Caption continued:* **c**, CD62L expression by eosinophils, neutrophils and basophils (above plots) from age-matched donors homozygous for the major rs2223286 allele (solid line, open curve) or minor rs2223286 allele (dotted line, shaded curve) (key). **d**, Genetic associations between SNPs in the *SELL* genomic region and cell-surface expression of CD62L by eosinophils (left) or level of *SELL* mRNA in whole blood (right), presented as 'zoomed' Manhattan plots: each symbol is an SNP; color indicates linkage disequilibrium (r^2), with the best hit (rs2223286) in purple; blue lines indicate local recombination rates.

neutrophils (rs2223286; $P = 1.6 \times 10^{-35}$ and $P = 8.8 \times 10^{-13}$, respectively) but not in basophils (Fig. 4b,c).

Various other local-pQTLs were found to be cell specific; three different association signals in the *HLA-DR* gene region were found to be associated with the MFI of HLA-DR in pDCs and $CD14^{hi}$ monocytes (rs114973966; $P = 2.2 \times 10^{-56}$), in conventional DCs (cDC1 cells, as defined by the expression of the transmembrane glycoprotein BDCA1) (rs2760994; $P = 6.1 \times 10^{-38}$) and in cDC3 cells (rs143655145; $P = 2.6 \times 10^{-11}$). To determine if these signals were independent of each other, we conducted omnibus association tests on imputed HLA alleles. We found that the association signals in $CD14^{hi}$ monocytes, pDCs and cDC1 cells actually resulted from different amino acidaltering variants at the same codon in position 13 of the HLA-DR β1 protein (omnibus test $P = 2.0 \times 10^{-47}$, $P = 7.0 \times 10^{-90}$ and $P = 5.3 \times 10^{-41}$ in $CD14^{hi}$ monocytes, pDC and cDC1 cells, respectively; Supplementary Tables 7 and 8) that has been shown to explain a large part of the association signal in the HLA locus for type 1 diabetes [Hu et al., 2015]. A different amino acid variant, at position 67 of HLA-DR β1, was identified in cDC3 cells (on the basis of their expression of the integral membrane protein BDCA3; $P = 3.9 \times 10^{-13}$). Conditional analyses also revealed independent associations of the cell-surface expression of HLA-DR with two residues in the class I *HLA-B* gene (position 97 ($P = 3.8 \times 10^{-17}$) and position 194 ($P = 1.3 \times 10^{-18}$); Supplementary Tables 7 and 8). Collectively, these results showed that the protein expression of markers of immune cell differentiation and activation was affected by common genetic variants, of which some are known to be linked to human pathogenesis.

2.6 Immune cell local-pQTLs control mRNA levels of nearby genes.

Although four of the nine local-pQTLs identified by our analyses could probably be explained by amino acidaltering variants in surrounding genes (Supplementary Tables 4 and 7), the remaining signals did not present obvious candidate causal variants. To delineate the functional basis of these associations, we investigated whether the corresponding SNPs were also associated with mRNA levels of nearby genes (i.e., expression quantitative trait loci (eQTLs)) using gene-expression data obtained from the same donors [Piasecka et al., 2018] and results from the Geno-

typeTissue Expression Project [GTEx Consortium, 2015]. Five of the local-pQTLs were strongly associated with the transcript levels of a surrounding gene (linear regression model adjusting on major cell proportions; $P < 1.0 \times 10^{-5}$; Fig. 4d). The SNPs that controlled the MFI of CD16 in $CD16^{hi}$ NK cells and that of CD32 in basophils, CD62L in eosinophils, CD8a in $CD69^{+}$ $CD16^{hi}$ NK cells and CD203c in basophils were associated with the mRNA levels of their genes (FCGR2B, SELL, CD8A and ENPP3, respectively) (Supplementary Table 4). These analyses indicated that genetic variants associated with immunophenotypes were able to directly affect the expression of genes encoding markers of immune cells in whole blood. This suggested that eQTL mapping in various immune cell compartments might greatly improve knowledge of the genetic factors that control inter-person variation in parameters of flow cytometry.

2.7 Novel trans-acting genetic associations with parameters of immune cells.

We detected six loci that did not exclusively act as local-pQTLs on immunophenotypes (Table 1 and Supplementary Fig. 21). These included variants associated with immune cell counts or genetically independent of the genes encoding immune cell markers with which they are associated (i.e., 'trans-pQTLs'). A variant in the vicinity of *S1PR1* (which encodes the sphingosine 1-phosphate receptor $S1P_1$ (CD363)) was associated with the MFI of CD69 in $CD16^{hi}$ NK cells (rs6693121; $P = 4.8 \times 10^{-37}$). CD69 is known to downregulate cell-surface expression of $S1P_1$ on lymphocytes, a mechanism that elicits egress from the thymus and secondary lymphoid organs [Garris et al., 2014]. Genetic variation in an intron of *ACOXL* (which encodes an acyl-coenzyme A oxidase-like protein) near *BCL2L11* (which encodes the apoptosis-related protein BCL2L11) was associated with the absolute number of $CD8a^{+}CD56^{hi}$ NK cells (rs12986962; $P = 9.1 \times 10^{-19}$). BCL2L11 (BIM) is an important regulator of lymphocyte apoptosis [Pellegrini et al., 2004] and is associated with chronic lymphocytic leukemia and the total number of blood cells [Van Der Harst et al., 2012]. A third association involved genetic variants near *ACTL9* (which encodes an actin-like protein) and the ratio of the MFI of CD16 in $CD16^{hi}$ NK cells to that in $CD56^{hi}$ NK cells (rs114412914, $P = 4.3 \times 10^{-30}$). The same variants have been also found to be associated with $CD56^{++}CD16^{-}$ NK cells in another study [Aguirre-Gamboa et al., 2016].

Although they were identified here for their trans effects on markers of the differentiation or activation of immune cells, three trans-acting genetic associations were also local-eQTLs for nearby genes encoding proteins related to the immune system [GTEx Consortium, 2015] (Supplementary Tables 4 and 6). The MFI of the chemokine receptor CCR7 in $CD4^{+}$ or $CD8b^{+}$ naive T cells was associated with a variant in *TMEM8A* (which encodes a transmembrane protein) (rs11648403; $P = 3.0 \times 10^{-19}$) that also controlled the level of *TMEM8A* mRNA ($P = 2.5 \times 10^{-27}$). TMEM8A is expressed on the surface of resting T cells and is downregulated after

cell activation [Motohashi et al., 2000], suggestive of a possible functional association and/or co-regulation with CCR7. Variants in the vicinity of *ALOX15* (which encodes arachidonate 15-lipoxygenase) were associated with increased protein levels of the high-affinity immunoglobulin E (IgE) receptor in eosinophils (rs56170457; $P = 9.2 \times 10^{-14}$) and increased levels of *ALOX15* mRNA ($P = 2.7 \times 10^{-13}$). These results, together with the high expression of ALOX15 protein and its proinflammatory effect on circulating eosinophils [Feltenmark et al., 2008], suggested an important role for this lipoxygenase in IgE-dependent allergic reactions. Finally, conditional GWAS identified an additional trans-acting association between a variant near *CD83* (which encodes the co-stimulatory molecule CD83) and the MFI of HLA-DR in cDC1 cells (rs72836542; $P = 2.8 \times 10^{-12}$; Supplementary Fig. 22); the same variant was also identified as a local-eQTL of CD83 expression ($P = 5.4 \times 10^{-21}$). These results suggested that CD83, an early activation marker of human DCs, upregulates HLA-DR expression in activated DCs.

2.8 Natural variation in the parameters of innate immune cells is 'preferentially' driven by genetic factors.

A large proportion of both MFI immunophenotypes and cell-number immunophenotypes that presented a genome-wide association were detected in innate immune cells (35 of 44 (80%)), including granulocytes, monocytes, NK cells and DCs (Table 1), while 47% of all immunophenotypes were measured in innate cells (Supplementary Table 3). Furthermore, of the adaptive-cell immunophenotypes that showed genetic associations, three of the nine measurements (33%) were related to naive T cells or B cells, while parameters of naive adaptive cells represented <10% of all measurements of adaptive cells. These observations suggested a stronger effect of genetic variants on innate and naive adaptive cell subpopulations than on differentiated or experienced adaptive immune cells.

In support of that hypothesis, the presence of HLA-DR molecules, assessed at the surface of both innate immune cells and adaptive immune cells, was strongly associated with *HLA-DR* genetic variation in monocytes, NK cells and DCs (Table 1) but not in $CD4^+$ or $CD8^+$ central memory T cells, effector memory T cells or T_{EMRA} cells ($P > 1.0 \times 10^{-6}$; Supplementary Table 5). Because we observed substantial correlations among the number of HLA-DR^+ memory T cells (linear model $R^2 \approx 0.3$; $P < 0.05$; Supplementary Fig. 16), we hypothesized that they were controlled at least in part by the same genetic factors, which we further assessed by multivariate GWAS. This refined approach detected a suggestive genetic association near *HLA-DRB1* with a variant (rs35743245; multivariate mixed model $P = 1.0 \times 10^{-8}$) in strong linkage disequilibrium with that detected in pDCs, monocytes and NK cells ($r^2 = 0.92$; Supplementary Fig. 23). This finding provided proof of the concept that the immunophenotypes of both innate cells and adaptive cells can be controlled by the same genetic factors but their effects are stronger in innate cells than in experienced adaptive cells.

Table 1. Genome-wide signals of associatoin with immunophenotypes in the Milieu Intérieur cohort. *P* values of the linear mixed model used for GWAS. [a]Other immunophenotypes correspond to any measured immunophenotype in the Milieu Intérieur cohort that was also significantly associated with the candidate variant, but to a lesser extent than the main immunophenotype. [b]Replication was performed in an independent cohort of 75 European-descent Americans. Only panels 4 and 7 could be used, due to sample limitations; effects were in the same direction as the primary cohort. [c]*P* values for biological replicates were estimated on the basis of immunophenotypes measured from blood newly obtained ~17 d after the initial visit, in 500 subjects of the Milieu Intérieur cohort [d]Previous identification noted by reference number; - indicates no previous identification. [e]EAF is the frequency of the effect allele, which was defined as the allele with a positive effect on the immunophenotype

Locus	Flow-cytometry panel	Immunopheno type	Other immunophenotypes[a]	*P* value	Replication *P* value[b]	*P* value for biological replicates[c]	Previous identification[d]	Effect size (SE)	Chr	Position	Candidate variant	Effect allele[e]	Other allele	EAF[e]	Candidate gene	Distance to TSS (kb)
1	4	CD69 in CD16^{hi} NK cells	CD69^{+}CD16^{hi} NK cells; CD69 in CD8a^{+} and CD69^{+}CD16^{+} NK cells	4.8×10^{-37}	6.3×10^{-4}	2.0×10^{-16}	-	0.14 (0.01)	1	101744633	rs6693121	A	C	0.40	*SIPR1*	41.0
2	4	CD16 in CD16^{hi} NK cells	CD16 in CD56^{hi} NK cells; HLA-DR in CD16^{hi}, CD8a^{+}CD16^{+} and CD69^{+}CD16^{+} NK cells	3.0×10^{-87}	7.1×10^{-7}	2.6×10^{-41}	15	22.77 (1.04)	1	161507448	rs3845548	C	T	0.87	*FCGR3A*	12.4
3	7	CD32 in basophils	-	1.7×10^{-36}	3.6×10^{-7}	1.6×10^{-18}	-	11.23 (0.86)	1	161653737	rs61804205	C	T	0.10	*FCGR2B*	20.8
4	7	CD62L in eosinophils	CD62L in neutrophils	1.6×10^{-35}	3.7×10^{-2}	1.4×10^{-8}	-	542.78 (42.08)	1	169665632	rs2223286	C	T	0.33	*SELL*	0.0
5	4	CD8a in CD69^{+}CD16^{hi} NK cells	CD8a in CD16^{hi}, CD56^{hi}, CD69^{+} CD56^{hi}, CD8 CD56^{hi}, CD8a^{+}CD16^{hi} and HLA-DR^{+}CD16^{hi} NK cells	5.9×10^{-58}	5.9×10^{-2}	3.4×10^{-24}	15	0.44 (0.03)	2	87026807	rs71411868	A	G	0.76	*CD8A*	0.0
6	4	Number of CD8a^{+}CD56^{hi} NK cells	CD56^{hi} NK cells; CD69^{+}CD56^{hi} NK cells; CD56^{+} ILCs	9.1×10^{-19}	2.7×10^{-2}	2.5×10^{-9}	-	1.57 (0.18)	2	111808558	rs12986962	A	G	0.62	*ACOXL / BCL2L11*	0.0
7	8	HLA-DR in cDC3 cells	-	2.6×10^{-11}	-	3.1×10^{-10}	-	0.11 (0.02)	6	32340176	rs143655145	T	C	0.19	*HLA-DRA*	67.4
8	8	HLA-DR in cDC1 cells	-	6.1×10^{-38}	-	1.3×10^{-17}	-	0.12 (0.01)	6	32574308	rs2760994	T	C	0.63	*HLA-DRB1*	16.7
9	8	HLA-DR in pDCs	CD86 in pDCs; HLA-DR^{+}CD56^{hi} NK cells; HLA-DR in CD14^{hi} monocytes	2.2×10^{-56}	-	2.7×10^{-26}	-	9.06 (0.54)	6	32599163	rs114973966	T	C	0.18	*HLA-DRB1*	41.5
10	6	CD24 in IgM^{+} marginal zone B cells	CD24 in B cells, and in naive, memory, CD4^{-}CD8^{-} (double-negative) memory, IgM^{-} marginal zone and marginal zone B cells	3.8×10^{-21}	-	5.5×10^{-10}	-	0.20 (0.02)	6	107168676	rs12529793	C	T	0.92	*CD24*	254.7
11	7	CD203c in basophils	-	2.1×10^{-28}	3.2×10^{-2}	3.9×10^{-14}	-	8.83 (0.77)	6	132043056	rs2270089	G	A	0.09	*ENPP3*	0.0
12	1	CCR7 in CD4^{+} naive T cells	CCR7 in CD8b^{+} naive T cells	3.0×10^{-19}	-	2.0×10^{-7}	-	0.07 (0.01)	16	429129	rs11648403	C	T	0.57	*TMEM8A*	0.0
13	7	FCεRI in eosinophils	-	9.2×10^{-14}	5.1×10^{-5}	1.9×10^{-7}	-	0.96 (0.13)	17	4560141	rs56170457	G	T	0.75	*ALOX15*	25.9
14	4	Ratio of CD16 MFI in CD16^{hi} and CD56^{hi} NK cells	CD16 in CD56^{hi} NK cells	4.3×10^{-30}	2.4×10^{-2}	8.9×10^{-13}	10	0.39 (0.03)	19	8788184	rs114412914	G	A	0.85	*ACTL9*	21.0

We next systematically quantified the effects of genetic and nongenetic factors on innate and adaptive cells. We established, for each immunophenotype, a linear-regression model that included the four non-genetic variables with the greatest effect (Figs. 2 and 3) and all genome-wide significant and suggestive variants (Table 1 and Supplementary Table 6) and estimated their respective contributions to the total variance. We found that a larger proportion of the variance of the immunophenotypes of innate cells (Fig. 5b,d) than that of the immunophenotypes of adaptive cells (Fig. 5a,c) was explained by genetic factors. Inversely, the variance in the number of adaptive cells was dominated by non-genetic factors such as age and CMV serostatus (Fig. 5a). To determine if these differences were significant, we used a mixed model that accounted for correlations among immunophenotypes. Conclusively, we estimated that the variance explained by genetics was 66% larger for measurements of innate cells than for those of adaptive cells (95% CI, 13–143%; $P = 0.012$ (bootstrap); $P = 0.032$ (Mann-Whitney U-test)), while the variance explained by nongenetic factors was 46% smaller for measurements of innate cells than for those of adaptive cells (95% CI, 22–63%; $P = 1.8 \times 10^{-3}$ (bootstrap); $P = 8.1 \times 10^{-3}$ (Mann-Whitney U-test)). When we considered non-genetic factors separately, the ratio of explained variance for measurements of innate cells to that of adaptive cells was the smallest for smoking (0.46, 95% CI: 0.17–1.25), followed by age (0.63; 95% CI, 0.42–0.95), CMV infection (0.71; 95% CI, 0.51–0.99) and sex (0.95; 95% CI, 0.60–1.51). Together our results indicated that genetic factors accounted for a substantial fraction of human variation in parameters of immune cells, with their influence being stronger on innate immune cells than on the phenotypes of adaptive immune cells.

3. Discussion

Over the past two decades, research into human immunology has employed multi-parameter cytometry to count and assess the activation state of immune cells in healthy and disease conditions. Although the parameters of immune cells do vary in the general population, the extent to which intrinsic, environmental and genetic factors explain this variability has remained elusive. To tackle these questions, we generated a broad resource by combining standardized flow cytometry with genome-wide DNA genotyping in a demographically well-defined cohort of 1,000 healthy people. We confirmed the strong and independent effects of age and CMV infection on naive T cell populations and memory T cell populations, respectively, and provided robust evidence for sex differences in the number of innate cells and adaptive cells. We showed that homeostasis of the immune system was altered after chronic exposure to cigarette smoke, which elicited both a decrease in the abundance of MAIT cells, possibly due to their increased migration to sites of inflammation, and an increase in the number of activated and memory T_{reg} cells, suggestive of a role for these immunosuppressive populations in the increased susceptibility of smokers

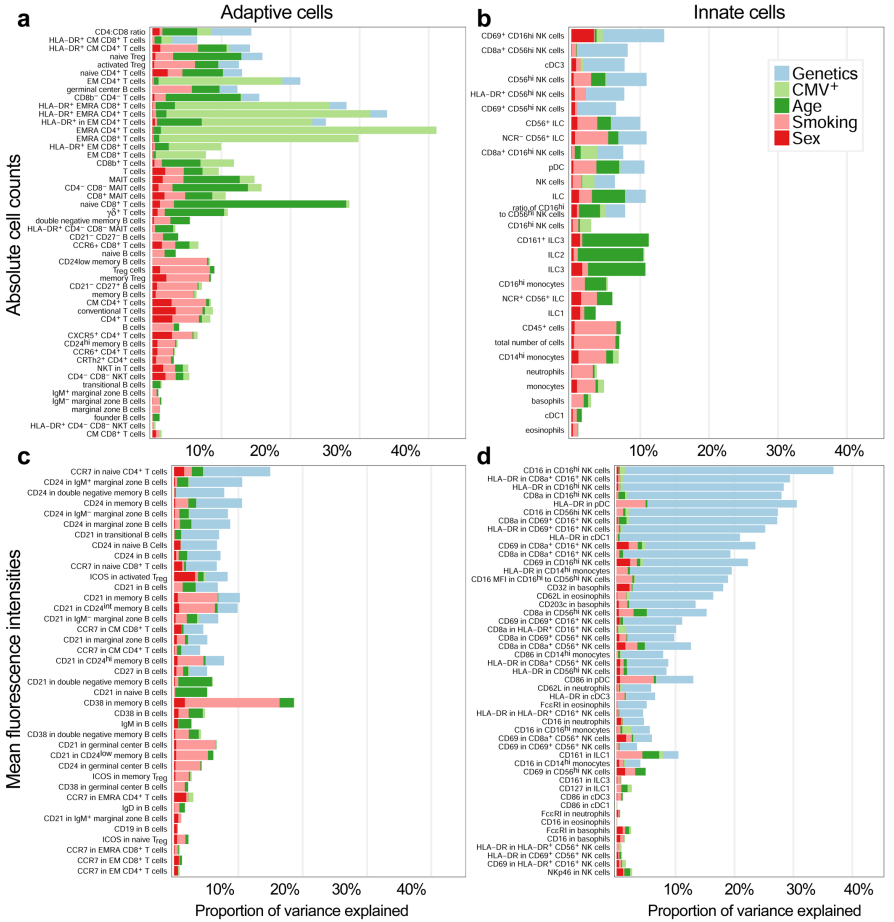

Figure 5. Proportion of variance of the parameters of innate and adaptive cells explained by non-genetic and genetic factors. Analysis of the effect of various factors (key) on variance of cellular abundance (76 absolute cell counts and two count ratios, assessed by flow cytometry; **a**,**b**) and MFI of various cell-surface markers (left margin; 87 MFI values and a ratio of MFI values, assessed by flow cytometry; **c**,**d**), presented as variance of the 91 parameters of adaptive cells (**a**,**c**) and 75 parameters of innate cells (**b**,**d**) decomposed into proportions explained (R^2) by intrinsic factors (key: age and sex (Fig. 2) or by environmental exposure (CMV infection and smoking; Figs. 2 and 3) and genetic factors (independent significant and suggestive GWAS hits, Table 1 and Supplementary Table 6).

to infection [Stämpfli and Anderson, 2009]. Furthermore, we found that human genetic variation substantially affected parameters of immune cells, particularly the cell-surface expression of markers conventionally used to identify leukocyte differentiation or activation. These results highlight the need to consider non-genetic and genetic features when interpreting parameters such as the circulating white blood cells of patients, a critical aspect in clinical monitoring. For example, expression of HLA-DR on monocytes is routinely measured by flow cytometry to predict the clinical course of septic shock and identify patients who might benefit from immunoadjuvant therapies [Venet et al., 2013]. We identified a strong effect of HLA-DRβ1 coding variation on the expression of HLA-DR by CD14^{hi} monocytes, which would suggest that tools used to predict fatal outcome in sepsis should be tailored to the patient's genetic makeup.

The most prominent result of our study was the lower number of genetic associations detected in memory T cells and B cells, relative to that in innate cells, an observation that could be explained by their strong dependence on the varying individual history of past infections. Adaptive immune cells are known to have a much longer half-life than that of myeloid innate cells, in mice and humans [Kolaczkowska and Kubes, 2013; Farber et al., 2014]. Stimulus-induced differentiation and population expansion might also result in the possible masking of genetic associations for adaptive cell types. Consistent with that, genetic associations in adaptive immune cells were observed mainly for immunophenotypes of naive adaptive cells. Our observations are further supported by a GWAS of 36 blood traits in 173,480 people, which found that the genetic heritability of monocyte and eosinophil counts was larger than that of lymphocyte counts [Astle et al., 2016]. However, that is at odds with another published study that concluded that adaptive immune traits are affected more by genetics, whereas innate immune traits are affected more by environment, on the basis of the estimated genetic heritability of 23,394 immunophenotypes in 497 adult female twins [Mangino et al., 2017]. We suggest that such deep immunophenotyping in large-scale cohorts, combined with statistical tests for differences in heritability that account for inherent correlations among phenotypes, might reveal a more balanced contribution of genetics to the natural variation in the traits of innate and adaptive immune cells.

Our findings that genetic factors preferentially controlled variation in innate immune cells have other important consequences. A published study of 105 healthy twin pairs concluded that variation in cell population frequencies is driven largely by non-heritable influences [Brodin et al., 2015]. We found instead that genetic variation explained a large part of the variance in the parameters of immune cells, particularly MFI measurements (i.e., cell-surface expression of protein markers) assessed in innate cells. This discrepancy might stem from the fact that the previously published study considered only a fraction of innate myeloid and lymphoid populations [Casanova and Abel, 2015], and possibly because of its limited power due to a moderate sample size. Also, our results suggested that the genetic control of cell-surface expression of immune cell markers was stronger than that of cell counts, and

the former were not assessed in most previously published population-immunology studies [Aguirre-Gamboa et al., 2016; Orrù et al., 2013; Brodin et al., 2015].

Finally, the mapping of genetic loci encoding proteins that control parameters of immune cells identified cell-specific pQTLs that showed enrichment for genetic variants associated with human diseases and traits. For example, we identified position 13 of the HLA-DRβ1 protein as a predictor of HLA-DR expression at the surface of pDCs and monocytes, which in turn is strongly associated with type 1 diabetes [Hu et al., 2015]; this would suggest an association of innate immunity with the disease [Astle et al., 2016]. Furthermore, the expression of CD56 and CD16 in NK cells was controlled by genetic variants near *ACTL9* that have been shown to be associated with atopic dermatitis [Paternoster et al., 2015], suggestive of the possible involvement of NK cells in this pathology [Bubnoff et al., 2010]. More generally, genetic variants found to modulate parameters of innate immune cells, in our study here and in published studies [Aguirre-Gamboa et al., 2016; Orrù et al., 2013; Roederer et al., 2015], have been directly linked to the etiology of several autoimmune disorders, such as inflammatory bowel disease, ulcerative colitis and atopic dermatitis. Together these findings illustrate the value of our approach, which mapped previously unknown genetic associations to specific cell populations and cellular states, providing new insights into the mechanisms underlying disease pathogenesis. Further evaluation of the natural variability in cellular mediators of immunity, together with the elucidation of their environmental and genetic determinants, will facilitate detailed delineation of the involvement of the immune system in human health and disease.

4. Methods

The Milieu Intérieur cohort. The 1,000 healthy donors of the Milieu Intérieur cohort were recruited by BioTrial (Rennes, France), and included 500 women and 500 men, and 200 people from each decade of life, between 20 and 69 years of age. Donors were selected based on stringent inclusion and exclusion criteria, detailed elsewhere [Thomas et al., 2015]. The clinical study was approved by the Comité de Protection des Personnes – Ouest 6 (Committee for the protection of persons) on 13 June 2012 and by the French Agence Nationale de Sécurité du Médicament (ANSM) on 22 June 2012. The study is sponsored by the Institut Pasteur (Pasteur ID-RCB Number: 2012-A00238-35) and was conducted as a single center study without any investigational product. The protocol is registered under ClinicalTrials.gov (study# NCT01699893).

Human material and staining protocol. Whole blood samples were collected from the 1,000 healthy, fasting donors on Li-heparin, every working day from 8 AM to 11 AM, from September 2012 to August 2013, in Rennes, France. Tracking procedures were established in order to ensure delivery to Institut Pasteur, Paris,

within 6 h of blood draw, at a temperature between 18°C and 25°C. To check the stability of our flow cytometry measures through time, a second blood sample was drawn for half of the cohort during a second visit, ~17 d on average after the first visit, ranging from 7 d to 44 d. After receipt, samples were kept at room temperature before sample staining. Details on staining protocols can be found elsewhere [Hasan et al., 2015].

Reproducibility testing and assay development. For optimization studies and panel development, whole blood samples were collected from healthy volunteers enrolled at the Institut Pasteur Platform for Clinical Investigation and Access to Research Bioresources (ICAReB) within the Diagmicoll cohort. The biobank activity of ICAReB platform is NF S96-900 certified. The Diagmicoll protocol was approved by the French Ethical Committee (CPP) Ile-de-France I, and the related biospecimen collection was declared to the Research Ministry under the code N° DC 2008-68. The reproducibility tests were performed as detailed elsewhere [Hasan et al., 2015].

Flow cytometry. Ten eight-color flow-cytometry panels were developed. Details on staining antibodies are in Supplementary Table 2. A unique lot of each antibody was used for the entire study. Each antibody was selected and titrated as described earlier [Hasan et al., 2015]. Gating strategies are described in Supplementary Figs. 1–10. The acquisition of cells was performed using two MACSQuant analyzers (Serial numbers 2420 & 2416), each fit with identical three lasers and ten detector optical racks (FSC, SSC and eight fluorochrome channels). Calibration of instruments was performed using MacsQuant calibration beads (Miltenyi, ref. 130-093-607). Flow cytometry data were generated using MACSQuantify software version 2.4.1229.1 and were saved as.mqd files (Miltenyi). The files were converted to FCS compatible format and analyzed by FlowJo software version 9.5.3. A total of 313 immunophenotypes were exported from FlowJo. These included 110 cell proportions, 106 cell counts, 89 MFI values and 8 ratios. We excluded from subsequent analyses all cell proportions, 35 immunophenotypes that were measured several times on different panels and were exported for quality controls, and two MFI values that were measured with a problematic clone (Supplementary Table 3). A total of 166 flow-cytometry measurements were thus analyzed, including 76 cell counts, 87 MFI values and 3 ratios (Supplementary Table 3). Problems in flow cytometry processing, such as abnormal lysis or staining, were systematically flagged by trained experimenters, which resulted in 8.70% missing data among the 166,000 measured values.

Outlier removal. Despite the exclusion of flagged problematic values, a limited number of outlier values were observed. As the goal of this study was to identify common non-genetic and genetic factors that control immune cell levels, we removed these outlier values. Outliers were detected using a distance-based algorithm instead

of a parametric method (for example, removal based on a number of s.d. from the mean), because of the substantial and highly variable skewness of the distributions of flow cytometry measurements. A value in the higher tail was considered an outlier if the distance to the closest point in the direction of the mean of the distribution was more than 60% of the total range of the sample, while a value in the lower tail was considered an outlier if that distance was more than 15% of the total range of the sample. To choose these threshold values, we simulated 10,000 log-normal distributions with a skewness similar to that of the flow cytometry measurements. We then searched for threshold values so that simulated values outside of these ranges were observed in less than 5% of the distributions. Outliers were only looked for in the 50 highest and lowest values. This threshold was chosen to make sure that we did not miss any effect on immunophenotypes of common genetic variants (minor allele frequency > 5%) or that of one of 39 continuous or common categorical non-genetic factors studied here. All values more extreme than the points labeled as outliers were also labeled outliers. A total of 24 values were removed at this stage.

Batch effects on flow-cytometry measurements. Two batch effects on flow cytometry measurements were considered: the hour at which blood samples were drawn (from 8 AM to 11 AM) and the day at which samples were processed (8–12 samples per day, from September 2012 to August 2013). The effect of the hour of blood draw was evaluated with linear regression on all immunophenotypes. We observed that the hour of blood draw affected a limited number of cell counts, mainly $CD16^{hi}$ NK cells (Supplementary Fig. 14a). The sampling-day effect was evaluated by estimating its variance component on all immunophenotypes. Visual inspection was used to determine whether temporal fluctuations (observed for those immunophenotypes with a large variance explained) were seasonal or not. We observed that sample processing day had a substantial effect on MFI. Fluctuations in MFI across time were strongly discontinuous, suggestive of technical issues possibly related to the compensation matrix, rather than seasonal effects (Supplementary Fig. 14b).

Inclusion and imputation of candidate non-genetic factors. A large number of demographic variables were available for the Milieu Intérieur cohort [Thomas et al., 2015]. These included infection and vaccination history, childhood diseases, health-related life habits, and socio-demographic variables. Of these, 39 variables were chosen for subsequent analyses (Supplementary Table 1) based on the fact that they were intrinsic factors (i.e., age, sex) or measured the exposure of people to exogenous factors and thus might not be affected by the immunophenotypes themselves. These variables were filtered based on their distribution (i.e., categorical variables with only rare levels, such as infrequent vaccines, were excluded) and on their levels of dependence on other variables (for example, height and BMI). The dependency matrix among the 39 non-genetic variables, together with batch variables, was obtained based on the generalized R^2 measures for pairwise fitted

generalized linear models. If the response was a continuous variable, we used a Gaussian linear model. If the response was binary, we used logistic regression. Categorical variables were used only as predictors. Missing values were imputed using the random forest-based R package missForest.

Effect of candidate non-genetic factors on immunophenotypes. To analyze the effect of non-genetic factors on immunophenotypes, we fitted a linear mixed model for each of the 166 immunophenotypes and each of the 39 non-genetic treatment variables. A total of 6,474 models were therefore fitted using the lme4 R package [Bates et al., 2014]. All models were fitted to complete cases. Due to lack of a priori knowledge on how the non-genetic variables affected the immunophenotypes, we did not attempt to make a full causal structural equation model for all variables. Instead, we chose to keep the amount of controls in the models small to increase interpretability of the results, and to make the study easier to reproduce. We included age, sex and CMV seropositivity as fixed-effect controls for all models (Fig. 3 and Supplementary Fig. 19), except when they were the treatment variable to be tested (Fig. 2 and Supplementary Fig. 17). The intrinsic factors (i.e., age and sex) were included as covariates because they are known to have an effect on immunophenotypes [Liston et al., 2016; Goronzy and Weyand, 2017; Aguirre-Gamboa et al., 2016; Furman et al., 2015; Pennell et al., 2012; Furman et al., 2014a; Astle et al., 2016], as well as on many of the other environmental exposures, and were therefore possible confounders. CMV seropositivity was included because it has been shown to strongly affect some immunophenotypes [Liston et al., 2016; Wertheimer et al., 2014; Furman et al., 2015; Brodin et al., 2015]. We also controlled for genome-wide significant SNPs for corresponding immunophenotypes (Table 1). Genetic variants were included to reduce the residual variance of the models and to make the inferences more robust. To correct for the batch effect related to the day of sample processing, we included it as a random effect for all models; we included a constant for each day and assumed that all constants were drawn from the same normal distribution. This procedure models correlation among subjects processed during the same day. We also included the hour of blood draw as a fixed-effect control for all models.

The distributions of the immunophenotypes have variable skewness. We considered normal, lognormal and negative binomial response distributions, and chose to model all immunophenotypes as lognormal based on diagnostic plots, AIC measures and our aim to have comparable results across immunophenotypes and facilitate the interpretation of effect sizes. A total of 46 immunophenotypes had zero values. A unit value was added to those before log-transformation.

For each model, we tested the hypothesis that the regression parameter for the treatment variable was zero by an F-test with the Kenward-Roger approximation. This test has better small- and medium-sample properties than the traditional chi-square-based likelihood ratio test for mixed models [Kenward and Roger, 1997] and

can readily be applied using the pbkrtest R package [Halekoh and Højsgaard, 2014]. We assumed that our sample size was large enough for this test to be appropriate and chose therefore not to do parametric bootstrapping. We considered all 6,474 tests as one multiple-testing family, and we used the false-discovery rate (FDR) as error rate. An effect was considered significant if the adjusted P value was smaller than 0.01. If a test was significant, confidence intervals were constructed using the profile likelihood method in such a way that the false-coverage rate (FCR) was controlled at a level of 0.01. The FCR measures the rate of confidence intervals that do not cover the true parameter and is needed if confidence intervals are selected based on a criterion that makes these intervals especially interesting, such as significant hypothesis tests [Benjamini and Yekutieli, 2005]. FCR-adjusted confidence intervals are always wider than regular intervals. All these analyses were done, and can be reproduced, with the mmi R package (`http://github.com/JacobBergstedt/mmi`).

Genome-wide DNA genotyping. The 1,000 subjects of the Milieu Intérieur cohort were genotyped at 719,665 SNPs by the HumanOmniExpress-24 BeadChip (Illumina, California). SNP call rate was higher than 97% in all donors. To increase coverage of rare and potentially functional variation, 966 of the 1,000 donors were also genotyped at 245,766 exonic SNPs by the HumanExome-12 BeadChip (Illumina, California). The HumanExome SNP call rate was lower than 97% in 11 donors, which were thus removed from this data set. We filtered out from both data sets SNPs that: (i) were unmapped on dbSNP138, (ii) were duplicated, (iii) had a low genotype clustering quality (GenTrain score < 0.35), (iv) had a call rate of < 99%, (v) were monomorphic, (vi) were on sex chromosomes and (vii) were in Hardy-Weinberg disequilibrium (HWE) ($P < 10^{-7}$). These SNP quality-control filters yielded a total of 661,332 SNPs and 87,960 SNPs for the HumanOmniExpress and HumanExome BeadChips, respectively. The two data sets were then merged, after excluding triallelic SNPs, SNPs with discordant alleles between arrays (even after allele flipping), SNPs with discordant chromosomal position, and SNPs shared between arrays that presented a genotype concordance rate of < 99%. Average concordance rate for the 16,753 SNPs shared between the two genotyping platforms was 99.9925%, and individual concordance rates ranged from 99.80% to 100%, which confirmed that no problem occurred during DNA sample processing. The final data set included 732,341 QC-filtered genotyped SNPs.

Genetic relatedness and structure. Possible pairs of genetically related subjects were detected using an estimate of the kinship coefficient and the proportion of SNPs that were not identical by state between all possible pairs of subjects, obtained with KING [Manichaikul et al., 2010]. Genetic structure was visualized with the Principal Component Analysis (PCA) implemented in EIGENSTRAT [Patterson et al., 2006]. For comparison purposes, the analysis was performed on 261,827 independent SNPs and 1,723 people, which include the 1,000 Milieu Intérieur subjects together with

a selection of 723 people from 36 populations of North Africa, the Near East, and Western and Northern Europe [Behar et al., 2010].

Genotype imputation. Prior to imputation, we phased the final SNP data set with SHAPEIT2 [Delaneau et al., 2012] using 500 conditioning haplotypes, 50 MCMC iterations, and 10 burn-in and 10 pruning iterations. SNPs and allelic states were then aligned to the 1,000 Genomes Project imputation reference panel (Phase1 v3.2010/11/23). We removed SNPs that have the same position in our data and in the reference panel but incompatible alleles, even after allele flipping, and ambiguous SNPs with C/G or A/T alleles. Genotype imputation was performed by IMPUTE v.259, considering 1-Mb windows and a buffer region of 1 Mb. Out of the 37,895,612 SNPs obtained after imputation, 37,164,442 were imputed. We removed 26,005,463 imputed SNPs with information metric ≤ 0.8, 43,737 duplicated SNPs, 955 monomorphic SNPs, and 449,903 SNPs with missingness of $> 5\%$ (individual genotype probabilities < 0.8 were considered as missing data). After quality-control filters, a total of 11,395,554 high-quality SNPs were further filtered for minor allele frequencies $> 5\%$, yielding a final set of 5,699,237 SNPs for association analyses.

Genome-wide association analysis. Prior to the GWAS, we transformed immunophenotypes using a procedure different from that used for the analysis of non-genetic factors. This is because we tested for association between immunophenotypes and millions of genetic variants, among which some have an unbalanced genotypic distribution (i.e., SNPs with a low minor allele frequency), which makes this analysis more sensitive to deviations from distributional assumptions. Our primary aim was therefore to use transformations that make the GWAS as robust as possible against such deviations. Also, we map loci associated with immunophenotypes based on *P* values, so it was less important to keep effect sizes on the same scale, in contrast with the analysis of non-genetic factors, for which we favored the interpretability of effect sizes. A unit value was first added to all phenotypes with zero values. The transformations were then chosen based on an AIC measure using the Jacobian-adjusted Gaussian likelihood, among three possible choices of increasing skewness: identity transformation, squareroot-tranformation and log-transformation. We kept the amount of possible transformations low to minimize the amount of added unmodelled stochasticity. The added unit value was kept only for immunophenotypes for which the log-transformation was chosen.

After transformation, a second round of outlier removal was done, to remove extreme values on the new scale. The thresholds for the lower and higher tail were 20%, obtained as for the first step of outlier removal (in the description of the distance-based outlier removal algorithm above), but on the Gaussian scale. The immunophenotypes were then imputed using the missForest *R* package, as missing data was not allowed by the subsequent analyses. We finally adjusted all immunophenotypes for the batch effect of processing days. We used the ComBat non-parametric empirical-

Bayes framework [Johnson et al., 2007], instead of the mixed model described above (in the subsection Effect of candidate non-genetic factors on immunophenotypes), because the GEMMA mixed model used to conduct GWAS (discussed below) includes only the random effect capturing genetic relatedness. ComBat adjusts for batch effects by leveraging multivariate correlations among response variables. We did not include variables of interest in the ComBat model (none of the non-genetic variables were significantly different across sample processing days, with the exception of smoking (regression P value = 0.002)).

To reduce the residual variance of GWAS models and make the inferences more robust [Mefford and Witte, 2012], we sought to adjust models for covariates selected among 42 variables. These included the 39 non-genetic variables (Supplementary Table 1), the hour-of-blood-draw variable, and the two first principal components of a PCA based on genetic data (Supplementary Fig. 20b). Covariates were selected by stability selection [Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013], with elastic net regression as the selection algorithm. A selection algorithm uses a cost function that drives regression parameters of non-predictive variables to zero, unlike least-square regressions. The elastic net method was used in particular because it has lower variance than stepwise methods and overcomes limitations of the LASSO method related to correlated variables [Friedman et al., 2001]. To perform stability selection, we estimated, for each of the $i \in \{1, \ldots, 42\}$ variables, the probability $p_i = \mathbb{P}(\beta_i = 0)$, *i.e.*, that the elastic net regression parameter β_i of variable i equals zero. Specifically, we first took 50 subsamples of half of the data, performed variable selection on each subsample, and estimated p_i as the number of subsamples in which $\beta_i > 0$, divided by the total number of subsamples. The variables were then chosen to be controls in the GWAS models by thresholding the probability $\hat{p}_i$. It has been shown that this procedure, with the right threshold and under certain assumptions, controls the FDR of selected variables [Shah and Samworth, 2013]. The procedure is more stable than selecting variables by, for instance, stepwise regression or elastic net without stability selection, and thus adds less unmodelled variability to the estimates. Still, because this approach does select predictive variables for each individual response variable, it adds more variance to the model selection, relative to that of models in which only age, sex, CMV infection and smoking would be systematically included. However, controlling for the selected variables would be expected to generate more parsimonious models (i.e., the inclusion of unnecessary covariates could reduce power [Wakefield, 2013]) and to decrease the risk of type 1 errors (for example, some of the many rare genetic variants that are tested could associate, by chance, with an immunophenotype when the model does not fulfil inference assumptions due to a specific, unmodelled covariate).

The univariate GWAS was conducted for each imputed, transformed and batch-effect corrected immunophenotype using the linear mixed model implemented in GEMMA [Zhou and Stephens, 2014], adjusting on selected covariates. GEMMA is an efficient mixed model that controls for genetic relatedness among donors and allows for multivariate analyses. Genetic relatedness matrices (GRMs) were

estimated for each chromosome separately, using the 21 other chromosomes, to exclude from the GRM estimation potentially associated SNPs (i.e., 'leave-one-chromosome' approach [Yang et al., 2014]). A conditional GWA analysis was also carried out for each of the 14 immunophenotypes that showed the strongest genome-wide significant signals ('main immunophenotypes' in Table 1), by including as a covariate in GEMMA the genotypes of the most strongly associated variant. A multivariate GWAS was conducted on a set of six candidate immunophenotypes (i.e., number of HLA-DR$^+$ memory T cells), using GEMMA linear mixed model adjusted on covariates that were selected for at least one of the six traits. For all genome-wide association analyses, a conservative genome-wide significant threshold of $P < 1 \times 10^{-10}$ was used, to account for testing multiple SNPs and immunophenotypes.

Power estimation. We used simulations to estimate the minimum effect of a variant that we could detect with 95% power by our GWAS. Specifically, we sampled 100,000 times a SNP in our data, and simulated an immunophenotype by adding to a randomly sampled immunophenotype the effect k of that SNP, k being drawn from a uniform distribution of bounds 0 and 1 (k is expressed in unit of phenotype s.d., as in 'scheme 1' of ref. [Zhang et al., 2010]). We then ran the GEMMA mixed model on the simulated data, and estimated the probability that the variant was detected, assuming our genome-wide significant threshold of $P < 1 \times 10^{-10}$. We found that we have 95% power to detect a SNP with a medium effect of 0.6 phenotype s.d. We also confirmed empirically the power to identify medium-effect genotype-phenotype associations in the Milieu Intérieur cohort by replicating well-known genetic associations with non-immune traits, including the association of *OCA2* and *HERC2* with eye and hair color (rs12913832; $P = 6.7 \times 10^{-138}$ and $P = 8.5 \times 10^{-18}$, respectively), the association of *SLC45A2* with hair color (rs16891982; $P = 3.2 \times 10^{-9}$), the association of the *UGT1A* gene cluster with bilirubin levels (rs6742078; $P = 2.6 \times 10^{-75}$), the association of *SLC2A9* with uric acid levels (rs6832439; $P = 4.3 \times 10^{-14}$), and the association of *CETP* with HDL levels (rs711752; $P = 4.5 \times 10^{-8}$).

Enrichment for variants associated with diseases. We explored the implication of our 15 genome-wide significant variants in human diseases and traits using previously published hits of genome-wide association studies (GWASs), obtained from the 31/08/2017 version of the EBI-NHGRI GWAS Catalog. A candidate variant was considered as implicated in a disease/trait if it was previously associated with such a disease or trait with a P value of $< 5 \times 10^{-8}$ or if it was in linkage disequilibrium (LD) with a variant associated with such a disease or trait ($r^2 > 0.6$). We tested if our 15 genome-wide significant variants showed enrichment for known associations with diseases or traits by resampling. We sampled 100,000 times 15 random SNPs with minor allele frequencies matched to those observed, and we calculated for each resampled set the proportion of variants known to be, or in LD with a variant known to be, associated with a disease. The enrichment P value was estimated as

the proportion of resamples for which this proportion was larger than that observed in our set. LD was precomputed for all 5,699,237 SNPs with PLINK 1.9 (options '–show-tags all–tag-kb 500–tag-r2 0.6') [Chang et al., 2015].

HLA typing and association tests. Four-digit classical alleles and variable amino acid positions in the HLA class I and II proteins were imputed with SNP2HLA v 1.03 [Jia et al., 2013]. 104 HLA alleles and 738 amino acid residues (at 315 positions) with MAF of > 1% were included in the analysis. Conditional haplotype-based association tests were performed using PLINK v. 1.07 [Purcell et al., 2007], as well as multivariate omnibus tests used to test for association at multi-allelic amino acid positions.

Replication cohort. We recruited 75 donors through the Genentech Genotype and Phenotype (gGAP) Registry. This sample size provides 95% power to replicate SNPs with an effect of > 0.9 phenotype s.d. Ethical agreement was obtained for all gGAP donors. Samples were received at room temperature and were processed 1 h after blood draw. Prior to staining, the blood was washed with PBS 1x. Except for the antibodies to CD32, the antibodies for population identification were titrated using the same clones and providers as in the primary study (Supplementary Table 2). Cell labeling were performed manually in deep-well plates. Data acquisition was performed within one hour using a calibrated FacsCantoII (Becton Dickinson). We selected panels 4 and 7 for the replication study, because 10 of the 16 GWAS hits were identified with these panels, and because of sample limitations. Immunophenotypes were transformed based on models chosen in the primary cohort. The GEMMA linear mixed model was used to test for replication, with age and sex as covariates and a GRM estimated from 1,960,432 autosomal SNPs obtained by the Illumina HumanOmni1-Quad v1.0 array.

Gene-expression assays. NanoString nCounter, a hybridization-based multiplex assay, was used to measure gene expression in unstimulated whole blood of the 1,000 Milieu Intérieur subjects, with the Human Immunology v2 Gene Expression CodeSet. These data are described in detail elsewhere [Piasecka et al., 2018]. Expression probes that bind to cDNAs in which at least three known common SNPs segregate in humans were removed from the analyses (i.e., *HLA-DQB1*, *HLA-DQA1*, *HLA-DRB1*, *HLA-B* and *C8G*). After quality-control filters, mRNA levels were available for 986 people at 90 candidate genes; i.e., immunity-related genes in a 1-Mb window around the genome-wide significant and suggestive associations identified in this study. For each sample, probe counts were $\log_2$ transformed, normalized and adjusted for batch effects. eQTL mapping was performed in a 1-Mb window around corresponding association signals, using the linear mixed model implemented in GenABEL [Aulchenko et al., 2007]. All models were adjusted on the proportion of eight major cell populations, including neutrophils, $CD19^+$ B cells, $CD4^+$ T cells,

$CD8^+$ T cells, $CD4^+CD8^+$ T cells, $CD4^-CD8^-$ T cells, NK cells and $CD14^+$ monocytes, to account for the effect of heterogeneous blood cell composition on gene expression.

Decomposition of the proportion of variance explained. We analyzed each of the 166 batch-corrected and transformed immunophenotypes (described in the subsection 'Genome-wide association analysis' above) with a linear regression model including the four non-genetic factors with the greatest effect (Fig. 2) (i.e., age, sex, CMV seropositivity status and smoking) and genome-wide genetic factors that were either significant ($P < 1 \times 10^{-10}$) or suggestive ($P < 5 \times 10^{-8}$). The contribution of each of these variables to the variance of each immunophenotype was calculated by averaging over the sums of squares in all orderings of the variables in the linear model, using the lmg metric in the *relaimpo R* package [Grömping et al., 2006].The averaging over orderings was done to avoid bias due to correlations among predictors.

The difference in contribution to explained variance between innate and adaptive immunophenotypes was tested using linear mixed models, where we used the log-transformed proportions of variance of each immunophenotype explained by age, sex, CMV serostatus, smoking or genetics as different response variables, and indicator variables for the immunophenotype being innate or adaptive, and being a count or an MFI value. The sum of the individual contributions of associated genetic variants was used to estimate the overall contribution of genetics. Since some of the immunophenotypes were correlated, their proportion of variance explained were also correlated. To account for this, we included a random effect term whose covariance matrix was modeled as a variance component multiplied by the sample correlation matrix among the immunophenotypes. Due to the small sample size, hypothesis testing was done by building a null distribution of likelihood ratios using the parametric bootstrap. The models were fitted using the R package *lme4qtl* (`http://github.com/variani/lme4qtl`). Because the distribution of variance explained by genetics was zero-inflated, we also tested for differences in the proportion of variance explained by non-genetic and genetic factors between innate and adaptive cell measurements with a non-parametric Mann-Whitney U-test. Because the Mann-Whitney U-test cannot account for correlations among immune cell measurements, we conducted this test on a subset of immunophenotypes that were selected to be uncorrelated ($h \leq 0.6$ with the *protoclust R* package). 50 immunophenotypes were kept, including 19 adaptive and 31 innate cell measures, among which the median Pearsons r value was 0.039.

References

Aguirre-Gamboa, R., I. Joosten, P. C. Urbano, R. G. van der Molen, E. van Rijssen, et al. (2016). "Differential effects of environmental and genetic factors on T and B cell immune traits". *Cell reports* **17**:9, pp. 2474–2487.

Altfeld, M. and M. Gale Jr (2015). "Innate immunity against HIV-1 infection". *Nature immunology* **16**:6, p. 554.

Amado, I. F., J. Berges, R. J. Luther, M.-P. Mailhé, S. Garcia, et al. (2013). "IL-2 coordinates IL-2–producing and regulatory T cell interplay". *Journal of Experimental Medicine* **210**:12, pp. 2707–2720.

Astle, W. J., H. Elding, T. Jiang, D. Allen, D. Ruklisa, et al. (2016). "The allelic landscape of human blood cell trait variation and links to common complex disease". *Cell* **167**:5, pp. 1415–1429.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. Van Duijn (2007). "GenABEL: an R library for genome-wide association analysis". *Bioinformatics* **23**:10, pp. 1294–1296.

Bates, D., M. Maechler, B. Bolker, S. Walker, et al. (2014). "Lme4: linear mixed-effects models using eigen and S4". *R package version* **1**:7, pp. 1–23.

Behar, D. M., B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, et al. (2010). "The genome-wide structure of the jewish people". *Nature* **466**:7303, p. 238.

Benjamini, Y. and D. Yekutieli (2005). "False discovery rate-adjusted multiple confidence intervals for selected parameters". *Journal of the American Statistical Association* **100**:469, pp. 71–81.

Bernard, C. (1865). *Introduction à l'étude de la médecine expérimentale par m. Claude Bernard*. Baillière.

Boeckh, M. and A. P. Geballe (2011). "Cytomegalovirus: pathogen, paradigm, and puzzle". *The Journal of clinical investigation* **121**:5, pp. 1673–1680.

Brodin, P., V. Jojic, T. Gao, S. Bhattacharya, C. J. L. Angel, et al. (2015). "Variation in the human immune system is largely driven by non-heritable influences". *Cell* **160**:1-2, pp. 37–47.

Bubnoff, D. von, E. Andrès, F. Hentges, T. Bieber, T. Michel, et al. (2010). "Natural killer cells in atopic and autoimmune diseases of the skin". *Journal of Allergy and Clinical Immunology* **125**:1, pp. 60–68.

Carr, E. J., J. Dooley, J. E. Garcia-Perez, V. Lagou, J. C. Lee, et al. (2016). "The cellular composition of the human immune system is shaped by age and cohabitation". *Nature immunology* **17**:4, p. 461.

Casanova, J.-L. and L. Abel (2015). "Disentangling inborn and acquired immunity in human twins". *Cell* **160**:1-2, pp. 13–15.

Cassard, L., F. Jönsson, S. Arnaud, and M. Daëron (2012). "Fcγ receptors inhibit mouse and human basophil activation". *The Journal of Immunology*, p. 1200968.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* **4**:1, p. 7.

Delaneau, O., J.-F. Zagury, and J. Marchini (2012). "Improved whole-chromosome phasing for disease and population genetic studies". *Nature methods* **10**:1, p. 5.

Della Bella, S., L. Bierti, P. Presicce, R. Arienti, M. Valenti, et al. (2007). "Peripheral blood dendritic cells and monocytes are differently regulated in the elderly". *Clinical immunology* **122**:2, pp. 220–228.

Dusseaux, M., E. Martin, N. Serriari, I. Péguillet, V. Premel, et al. (2011). "Human MAIT cells are xenobiotic-resistant, tissue-targeted, CD161hi IL-17–secreting T cells". *Blood* **117**:4, pp. 1250–1259.

Farber, D. L., N. A. Yudanin, and N. P. Restifo (2014). "Human memory T cells: generation, compartmentalization and homeostasis". *Nature Reviews Immunology* **14**:1, p. 24.

Feltenmark, S., N. Gautam, Å. Brunnström, W. Griffiths, L. Backman, et al. (2008). "Eoxins are proinflammatory arachidonic acid metabolites produced via the 15-lipoxygenase-1 pathway in human eosinophils and mast cells". *Proceedings of the National Academy of Sciences* **105**:2, pp. 680–685.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA:

Furman, D., B. P. Hejblum, N. Simon, V. Jojic, C. L. Dekker, et al. (2014a). "Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination". *Proceedings of the National Academy of Sciences* **111**:2, pp. 869–874.

Furman, D., V. Jojic, B. Kidd, S. Shen-Orr, J. Price, et al. (2014b). "Apoptosis and other immune biomarkers predict influenza vaccine responsiveness". *Molecular systems biology* **10**:9, p. 750.

Furman, D., V. Jojic, S. Sharma, S. S. Shen-Orr, C. J. Angel, et al. (2015). "Cytomegalovirus infection enhances the immune response to influenza". *Science translational medicine* **7**:281, 281ra43–281ra43.

Garris, C. S., V. A. Blaho, T. Hla, and M. H. Han (2014). "Sphingosine-1-phosphate receptor 1 signalling in T cells: trafficking and beyond". *Immunology* **142**:3, pp. 347–353.

Goronzy, J. J. and C. M. Weyand (2017). "Successful and maladaptive T cell aging". *Immunity* **46**:3, pp. 364–378.

Grömping, U. et al. (2006). "Relative importance for linear regression in R: the package relaimpo". *Journal of statistical software* **17**:1, pp. 1–27.

GTEx Consortium (2015). "The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans". *Science* **348**:6235, pp. 648–660.

Halekoh, U. and S. Højsgaard (2014). "A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest". *Journal of Statistical Software* **59**:9, pp. 1–30.

Hasan, M., B. Beitz, V. Rouilly, V. Libri, A. Urrutia, et al. (2015). "Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping". *Clinical Immunology* **157**:2, pp. 261–276.

Hu, X., A. J. Deutsch, T. L. Lenz, S. Onengut-Gumuscu, B. Han, et al. (2015). "Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk". *Nature genetics* **47**:8, p. 898.

Jia, X., B. Han, S. Onengut-Gumuscu, W.-M. Chen, P. J. Concannon, et al. (2013). "Imputing amino acid polymorphisms in human leukocyte antigens". *PloS one* **8**:6, e64683.

Johnson, W. E., C. Li, and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* **8**:1, pp. 118–127.

Kearley, J., J. S. Silver, C. Sanden, Z. Liu, A. A. Berlin, et al. (2015). "Cigarette smoke silences innate lymphoid cell function and facilitates an exacerbated type I interleukin-33-dependent response to infection". *Immunity* **42**:3, pp. 566–579.

Kenward, M. G. and J. H. Roger (1997). "Small sample inference for fixed effects from restricted maximum likelihood". *Biometrics*, pp. 983–997.

Kolaczkowska, E. and P. Kubes (2013). "Neutrophil recruitment and function in health and inflammation". *Nature Reviews Immunology* **13**:3, p. 159.

Liston, A., E. J. Carr, and M. A. Linterman (2016). "Shaping variation in the human immune system". *Trends in immunology* **37**:10, pp. 637–646.

Mangino, M., M. Roederer, M. H. Beddall, F. O. Nestle, and T. D. Spector (2017). "Innate and adaptive immune traits are differentially affected by genetic and environmental factors". *Nature communications* **8**, p. 13850.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, et al. (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* **26**:22, pp. 2867–2873.

Mefford, J. and J. S. Witte (2012). "The covariate's dilemma". *PLoS genetics* **8**:11, e1003096.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**:4, pp. 417–473.

Monticelli, L. A., G. F. Sonnenberg, M. C. Abt, T. Alenghat, C. G. Ziegler, et al. (2011). "Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus". *Nature immunology* **12**:11, p. 1045.

Motohashi, T., S. Miyoshi, M. Osawa, H. J. Eyre, G. R. Sutherland, et al. (2000). "Molecular cloning and chromosomal mapping of a novel five-span transmembrane protein gene, M83". *Biochemical and biophysical research communications* **276**:1, pp. 244–250.

Orme, I. M., R. T. Robinson, and A. M. Cooper (2015). "The balance between protective and pathogenic immune responses in the TB-infected lung". *Nature immunology* **16**:1, p. 57.

Orrù, V., M. Steri, G. Sole, C. Sidore, F. Virdis, et al. (2013). "Genetic variants regulating immune cell levels in health and disease". *Cell* **155**:1, pp. 242–256.

Paternoster, L., M. Standl, J. Waage, H. Baurecht, M. Hotze, et al. (2015). "Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis". *Nature genetics* **47**:12, p. 1449.

Patterson, N., A. L. Price, and D. Reich (2006). "Population structure and eigenanalysis". *PLoS genetics* **2**:12, e190.

Patterson, S., P. Moran, E. Epel, E. Sinclair, M. E. Kemeny, et al. (2013). "Cortisol patterns are associated with T cell activation in HIV". *PloS one* **8**:7, e63429.

Pellegrini, M., P. Bouillet, M. Robati, G. T. Belz, G. M. Davey, et al. (2004). "Loss of Bim increases T cell production and function in interleukin 7 receptor–deficient mice". *Journal of Experimental Medicine* **200**:9, pp. 1189–1195.

Pennell, L. M., C. L. Galligan, and E. N. Fish (2012). "Sex affects immunity". *Journal of autoimmunity* **38**:2-3, J282–J291.

Piasecka, B., D. Duffy, A. Urrutia, H. Quach, E. Patin, et al. (2018). "Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges". *Proceedings of the National Academy of Sciences* **115**:3, E488–E497.

Puchta, A., A. Naidoo, C. P. Verschoor, D. Loukov, N. Thevaranjan, et al. (2016). "TNF drives monocyte dysfunction with age and results in impaired anti-pneumococcal immunity". *PLoS pathogens* **12**:1, e1005368.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". *The American Journal of Human Genetics* **81**:3, pp. 559–575.

Reichert, T., M. DeBruyère, V. Deneys, T. Tötterman, P. Lydyard, et al. (1991). "Lymphocyte subset reference ranges in adult caucasians". *Clinical immunology and immunopathology* **60**:2, pp. 190–208.

Roederer, M., L. Quaye, M. Mangino, M. H. Beddall, Y. Mahnke, et al. (2015). "The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis". *Cell* **161**:2, pp. 387–403.

Sauce, D. and V. Appay (2011). "Altered thymic activity in early life: how does it affect the immune system in young adults?" *Current opinion in immunology* **23**:4, pp. 543–548.

Serafini, N., C. A. Vosshenrich, and J. P. Di Santo (2015). "Transcriptional regulation of innate lymphoid cell fate". *Nature Reviews Immunology* **15**:7, p. 415.

Shah, R. D. and R. J. Samworth (2013). "Variable selection with error control: another look at stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**:1, pp. 55–80.

Stämpfli, M. R. and G. P. Anderson (2009). "How cigarette smoke skews immune responses to promote infection, lung disease and cancer". *Nature Reviews Immunology* **9**:5, p. 377.

Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, et al. (2015). "The milieu intérieur study – an integrative approach for study of human immunological variance". *Clinical Immunology* **157**:2, pp. 277–293.

Tollerud, D., J. Clark, L. M. Brown, C. Neuland, L. Pankiw-Trost, et al. (1989). "The influence of age, race, and gender on peripheral blood mononuclear-cell subsets in healthy nonsmokers". *Journal of clinical immunology* **9**:3, pp. 214–222.

Tsuchiya, M., A. Asada, E. Kasahara, E. F. Sato, M. Shindo, et al. (2002). "Smoking a single cigarette rapidly reduces combined concentrations of nitrate and nitrite and concentrations of antioxidants in plasma". *Circulation* **105**:10, pp. 1155–1157.

Van Der Harst, P., W. Zhang, I. M. Leach, A. Rendon, N. Verweij, et al. (2012). "Seventy-five genetic loci influencing the human red blood cell". *Nature* **492**:7429, p. 369.

Venet, F., A.-C. Lukaszewicz, D. Payen, R. Hotchkiss, and G. Monneret (2013). "Monitoring the immune response in sepsis: a rational approach to administration of immunoadjuvant therapies". *Current opinion in immunology* **25**:4, pp. 477–483.

Vivier, E., D. H. Raulet, A. Moretta, M. A. Caligiuri, L. Zitvogel, et al. (2011). "Innate or adaptive immunity? the example of natural killer cells". *Science* **331**:6013, pp. 44–49.

Vrisekoop, N., I. den Braber, A. B. de Boer, A. F. Ruiter, M. T. Ackermans, et al. (2008). "Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool". *Proceedings of the National Academy of Sciences* **105**:16, pp. 6115–6120.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*. Springer Science & Business Media.

Wertheimer, A. M., M. S. Bennett, B. Park, J. L. Uhrlaub, C. Martinez, et al. (2014). "Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans". *The Journal of Immunology*, p. 1301721.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price (2014). "Advantages and pitfalls in the application of mixed-model association methods". *Nature genetics* **46**:2, p. 100.

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, et al. (2010). "Mixed linear model approach adapted for genome-wide association studies". *Nature genetics* **42**:4, p. 355.

Zhou, X. and M. Stephens (2014). "Efficient multivariate linear mixed model algorithms for genome-wide association studies". *Nature methods* **11**:4, p. 407.

Acknowledgments

We thank the Center for Translation Research, Institut Pasteur, and the OMNI Biomarker Development-Flow Cytometry Biomarker group, Genentech, for support. This work benefited from support of the French governments program Investissement dAvenir, managed by the Agence Nationale de la Recherche (reference 10-LABX-69-01). J.B. is a member of the LCCC Linnaeus Center and the ELLIIT Excellence Center at Lund University and is supported by the ELLIIT Excellence Center.

Author contributions

Conceptualization, E.P., L.Q.-M. and M.L.A.; methodology, M.H., V.L., A.U., F.J., B.B., C.L., F.H., L.R., I.P., O.L. and J.P.D.; software, E.P., J.B., V.R., P.S., C.H., B.P. and J.F.; validation, M.H., V.R., F.J., Y.W.L. and M.L.A.; formal analysis, E.P., J.B., V.R., P.S., C.H., C.G., B.P. and J.F.; investigation, E.P., M.H., J.B., V.L., A.U., C.A., F.J., H.Q., M.Z., B.P., C.L., L.R., F.H., O.L., J.P.D. and M.L.A.; data curation, E.P., M.H., J.B., V.R., V.L., A.U., B.P., C.L., L.R., I.P., O.L. and J.P.D.; writing (original draft), E.P., J.B., C.A., D.D. and M.L.A.; writing (review and editing), E.P., M.H., J.B., C.A., L.R., O.L., M.F., J.P.D., J.F., L.Q.-M. and M.L.A.; supervision, E.P., M.H., M.F., J.F., D.D. and M.L.A.; project administration, J.H., S.T. and D.D.; and funding acquisition, M.F., J.F., L.Q.-M. and M.L.A. L.Q.M. and M.L.A. are co-coordinators of the Milieu Intérieur Consortium (more information available at `http://www.milieuinterieur.fr/en`).

Data and materials availability

The SNP array data that support the findings of this study have been deposited in the European Genome-Phenome Archive (EGA) with the accession code EGAS00001002460. The flow cytometric data can be downloaded as an R package (`http://github.com/JacobBergstedt/mmi`) and explored with the online Shiny application (available at `http://milieu-interieur.cytogwas.pasteur.fr/`). The code developed to identify non-genetic factors that affect immunophenotypes and quantify their effects has been made available online (`http://github.com/JacobBergstedt/mmi`).

Supplementary material

The supplementary material of the article can be found at https://www.nature.com/articles/s41590-018-0049-7. A PDF with the supplementary figures can be downloaded at http://lup.lub.lu.se/record/8506d40c-14f9-45d3-965b-eb49ff8ff918. The supplementary tables in xlsx format can be downloaded at http://lup.lub.lu.se/record/8506d40c-14f9-45d3-965b-eb49ff8ff918.

Paper II

Human thymopoiesis is influenced by a common genetic variant within the *TCRA-TCRD* locus

Emmanuel Clave* **Itauá Leston Araujo*** **Cécile Alanio***
Etienne Patin* **Jacob Bergstedt*** **Alejandra Urrutia**
Silvia Lopez-Lastra **Yan Li** **Bruno Charbit**
Cameron Ross MacPherson **Milena Hasan**
Breno Luiz Melo-Lima **Corinne Douay** **Noémie Saut**
Marine Germain **David-Alexandre Trégouët**
Pierre-Emmanuel Morange **Magnus Fontes** **Darragh Duffy**
James P. Di Santo, **Lluis Quintana-Murci** **Matthew L. Albert†**
Antoine Toubert† **The Milieu Intérieur Consortium**

* These authors contributed equally † These authors jointly directed the work

Abstract

The thymus is the primary lymphoid organ where naïve T cells are generated; however, with the exception of age, the parameters that govern its function in healthy humans remain unknown. We characterized the variability of thymic function among 1000 age- and sex-stratified healthy adults of the Milieu Intérieur cohort, using quantification of T cell receptor excision circles (TRECs) in peripheral blood T cells as a surrogate marker of thymopoiesis. Age and sex were the only nonheritable factors identified that affect thymic function. TREC amounts decreased with age and were higher in women compared to men. In addition, a genome-wide association study revealed a common variant (rs2204985) within the T cell receptor *TCRA-TCRD* locus, between the *DD2* and *DD3* gene segments, which associated with TREC amounts. Strikingly, transplantation of human hematopoietic stem cells with the rs2204985 GG genotype into immunodeficient mice led to thymopoiesis with higher TRECs, increased thymocyte counts, and a higher TCR repertoire diversity. Our population immunology approach revealed a genetic locus that influences thymopoiesis in healthy adults, with potentially broad implications in precision medicine.

Originally published in Science translational medicine, September 2018. Reprinted with permission.

1. Introduction

In healthy individuals, continuous production of naïve self-tolerant T cells by the thymus ensures potent immune responses toward newly encountered antigens and contributes to maintenance of the naïve T cell repertoire [Miller, 2011; Mathis and Benoist, 2009]. Thymic function has been extensively studied for its capacity to shape the adaptive immune repertoire through positive and negative selection [Stritesky et al., 2012]. However, little is known about the environmental or genetic determinants of thymopoiesis in healthy individuals. Such insights would be relevant for optimizing regenerative strategies [Boehm and Swann, 2013], especially in conditions where thymic function is altered, such as aging [Palmer et al., 2018; Ferrando-Martínez et al., 2013], HIV infection [Dion et al., 2004], or allogeneic hematopoietic stem cell transplantation (allo-HSCT) [Krenger et al., 2011].

A bilateral cross-talk between thymocytes and thymic stromal cells directs sequential intrathymic T cell development and helps maintain activity of thymic stromal niches [Abramson and Anderson, 2017; Kurd and Robey, 2016]. Thymocyte progenitors receive signals from cortical thymic epithelial cells (TECs) for their commitment to the T cell lineage via the engagement of the NOTCH1 receptor with Delta-like 4 ligand, a major Forkhead box protein (FOX)N1 target in the thymic epithelium [Zuklys et al., 2016]. The medulla, via medullary TECs (mTECs) and dendritic cells, has a critical role in establishing self-tolerance by negative selection and induction of regulatory T cells (T_{regs}), especially but not exclusively via mTECs expressing the autoimmune regulator gene (*AIRE*) [Abramson and Anderson, 2017]. Naïve T cells are heterogeneous and include so-called recent thymic emigrants (RTEs), a subset that undergoes further post-thymic maturation [Fink and Hendricks, 2011]. Some phenotypic markers have been proposed to identify RTEs, such as CD31 (PECAM-1) in $CD4^+$ T cells. However, CD31 expression can be maintained during cytokine-driven proliferation of $CD4^+$ T cells, limiting its use as a specific marker of thymopoiesis. RTEs are enriched in T cell receptor excision circles (TRECs), which are produced during thymic *TCR* somatic recombination [Kohler and Thiel, 2009]. TRECs persist within mature T cells as episomal DNA [Villartay et al., 1988], cannot replicate, and are diluted out by peripheral cell divisions. Their quantification in peripheral blood provides a noninvasive surrogate marker of thymopoiesis, especially relevant in steady-state homeostatic conditions of the T cell compartment.

Signal joint TRECs (sjTRECs) are generated during the recombination of the TCRα chain, in double-positive (DP) $CD4^+CD8^+$ thymocytes, before positive and negative selection and lineage commitment [Villartay et al., 1988]. Polymerase chain reaction (PCR)-based quantification of sjTRECs is used in clinical laboratories as a diagnostic test for the recovery of the naïve T cell repertoire during HIV treatment, after allo-HSCT, and in the screening of severe combined immunodeficiencies in newborns [Douek et al., 1998; Clave et al., 2009; Puck, 2012]. Similarly, assays are available to measure βTRECs generated during the TCRβ chain recombination

at the $CD4^-CD8^-$ double-negative (DN) 3 stage. Because the β chain recombines before the α chain, βTRECs are much less abundant than sjTRECs in the periphery and frequently fall below the detection threshold in quantitative PCR. Given the dilution of βTRECs at each cell division between βTREC and sjTREC generation, the $\log_2$ transformation of the sjTREC/βTREC ratio gives an estimate of the number of intrathymic divisions occurring between DN and DP stages [Dion et al., 2004].

Here, we quantified TRECs from the peripheral blood of 1000 healthy individuals of western European ancestry [the Milieu Intérieur (MI) cohort] at immunological steady state, stratified by sex and age across five decades of life from 20 to 69 years [Thomas et al., 2015]. This population immunology approach revealed determinants of heterogeneity in human thymic function and identified a common genetic variation within the *TCRA-TCRD* locus directly affecting thymopoiesis.

2. Results

2.1 Validation of TRECs as surrogate markers of thymic function in the MI cohort.

We first standardized and validated sjTREC and βTREC high throughput assays (Fig. 1A and fig. S1) and then applied them to DNA from the 1000 MI donors. sjTREC counts normalized per 150,000 whole blood cells were used in subsequent analyses and correlated ($r^2 = 0.99$, $P < 10^{-16}$) with sjTRECs calculated as absolute numbers per microliter of blood (fig. S1C), the latter being not affected by T cell peripheral divisions. Log_{10}-transformed values of sjTRECs (Log_{10} sjTRECs) showed a normal distribution (kurtosis test, $P = 0.25$), with a mean of 2.4 ± 0.03 (minimum to maximum range, 0.2 to 4.1; fig. S1D). By contrast, Log_{10}-transformed values of βTRECs (Log_{10} βTRECs) showed a bimodal distribution, with 368 donors having samples below the limit of assay detection. In donors with detectable βTRECs in whole blood, Log_{10} βTRECs followed a normal distribution (kurtosis test, $P = 0.70$), with a mean of 1.75 ± 0.06 (minimum to maximum range, 0 to 3.1). Finally, the number of intrathymic divisions was also normally distributed across the healthy donors with detectable βTRECs (kurtosis test, $P = 0.72$), with a mean of 3.0 ± 0.21 (minimum to maximum range, -4.3 to 10.6; fig. S1D).

We evaluated whether TRECs are associated with any of 173 immune cell variables, defined through 10 eight-color immunophenotyping flow cytometry panels [Patin et al., 2018]. Using a linear mixed model approach controlling for potential confounders and batch effects, sjTRECs were found to be strongly associated with naïve $CD8^+$ and $CD4^+$ T cell counts or other cell types that are known to develop within the thymus, including naïve T_{regs} and invariant natural killer T (iNKT) cells (Fig. 1B and fig. S2). Naïve $CD8^+$ T cell counts doubled with a 10-fold increase in sjTRECs [confidence interval (CI), 78 to 136%; Kenward-Rogers (K-R) approximate F test, adjusted $P = 3 \times 10^{-47}$], whereas naïve $CD4^+$ T cell, NKT cell, and naïve T_{reg} counts showed 63% (CI, 40 to 88%; adjusted $P = 3 \times 10^{-21}$), 40% (CI, 6

to 84%; adjusted $P = 7 \times 10^{-3}$), and 44% (CI, 25 to 65%; adjusted $P = 5 \times 10^{-13}$) increases, respectively, per 10-fold increase in sjTRECs (Fig. 1C). We also found significant associations of T cell compartments with βTRECs (fig. S3). Naïve $CD8^+$ T cell, naïve $CD4^+$ T cell, and naïve T_{reg} counts showed 14% (CI, 6 to 22%, adjusted $P = 9 \times 10^{-6}$), 11% (CI, 3 to 19%; adjusted $P = 6 \times 10^{-4}$), and 8% (CI, 1 to 16%; adjusted $P = 0.008$) increases, respectively, per 10-fold increase in βTRECs (fig. S3).

2.2 Nonheritable factors associated with TREC amounts in the MI cohort.

Not only several nonheritable factors have been previously identified as affecting thymic function, in particular aging [Mitchell et al., 2010], but also endocrine factors such as sex steroid and growth hormones, body mass index, and metabolic syndrome [Taub and Longo, 2005; Yang et al., 2009; Youm et al., 2016]. Subjects included in the MI cohort were surveyed for a large number of variables related to nutrition, sleep, smoking, vaccination, and medical history [Thomas et al., 2015]. From these, we selected 56 candidate variables that potentially affect thymic function (table S1 and Supplementary Materials and Methods) and applied linear mixed models, controlling for potential confounders and batch effects, to identify factors that contribute to variance in thymic function. We estimated with power simulations that our false-negative rate was <5% if the candidate variable explained >2.5% of the variance. We found no significant effects of cytomegalovirus (CMV) seropositivity, influenza A serostatus, metabolic score index, or C-reactive protein on sjTRECs and βTRECs (figs. S4 to S7). By contrast, age had a strong effect on sjTRECs and, to a lesser extent, on βTRECs and the number of intrathymic divisions (Fig. 2, A to C, and figs. S4 to S6). sjTRECs showed a decrease of 4.9% per year (CI, 4.3 to 5.6%; K-R F test, adjusted $P = 4 \times 10^{-104}$; Fig. 2A) yet remained detectable in >95% of 60- to 69-year-old donors. Among donors with detectable βTRECs, we detected a 2% decrease per year (CI, 0.7 to 3.4%; adjusted $P = 8 \times 10^{-5}$; Fig. 2B). We also observed fewer donors with detectable amounts of βTRECs as a function of age [odds of having detectable amounts decreased by 3.36% per year; CI, 1.76 to 4.99%; likelihood ratio test (LRT), adjusted $P = 6 \times 10^{-10}$; fig. S5D]. Finally, we estimated a decrease of 0.26 intrathymic divisions every 10 years of age (CI, 0.03 to 0.5; K-R F test, adjusted $P = 9 \times 10^{-3}$). Strikingly, sex also showed a strong effect on sjTREC amounts with 67% (CI, 38 to 102%; adjusted $P = 2 \times 10^{-15}$) higher sjTREC amounts in women of all ages, relative to men (Fig. 2D). In contrast, no associations were found between sex and either the probability of having detectable βTRECs, βTREC amounts for donors with detectable βTRECs, or the number of intrathymic divisions (adjusted $P > 0.05$; figs. S5 and S6).

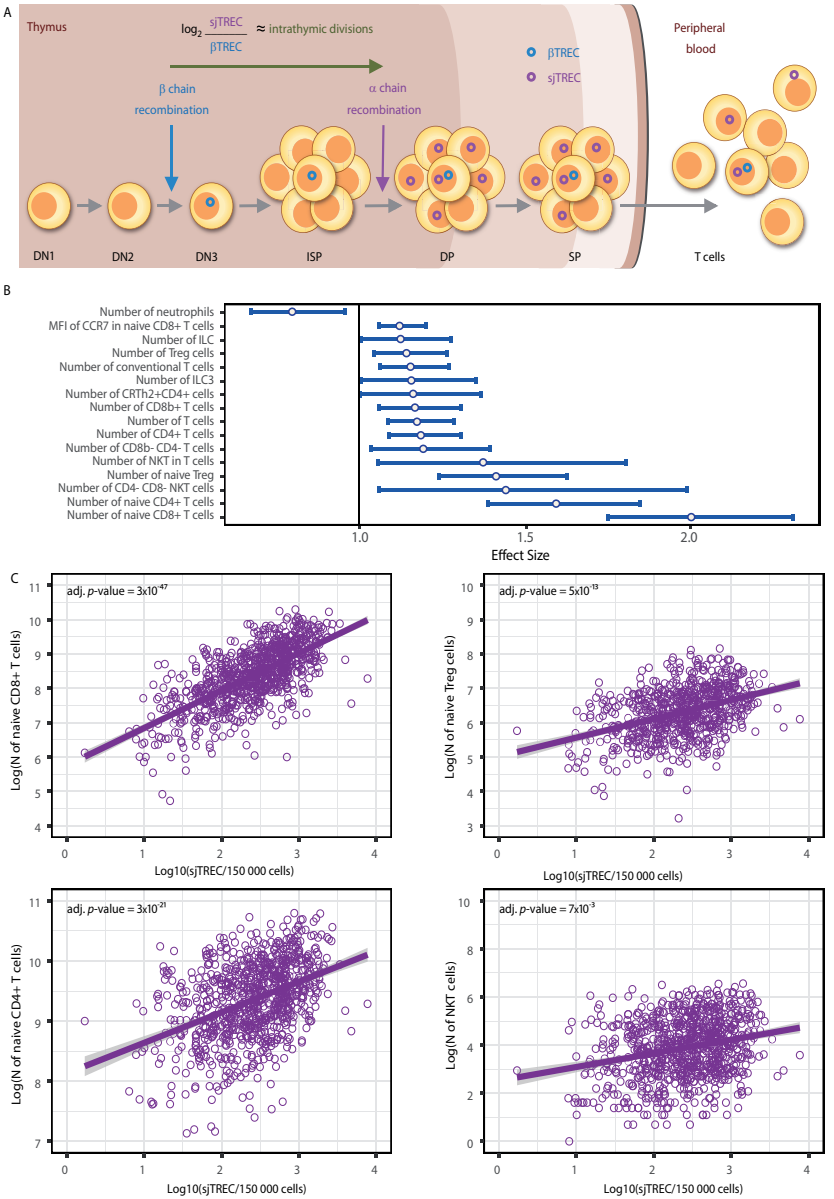

Figure 1. Thymic function associates with naïve T cell immune phenotypes. (**A**) βTRECs (blue) are episomal DNA generated during the TCRB recombination. *Caption continues on the next page*

Figure 1. *Caption continued:* sjTRECs (purple) derive from the deletion of the *TCRD* locus during *TCRA* locus recombination (shown in fig. S1B). DN, double negative; ISP, immature single positive; DP, double positive; SP, single positive. (**B**) Effect sizes of significant associations [adjusted *P* values (adj. *P*) < 0.05] between sjTRECs and immune cells and parameters measured by flow cytometry in 969 healthy individuals from the MI cohort. Effect sizes were estimated in a mixed model (see Supplementary Materials and Methods). MFI, mean fluorescence intensity. (**C**) Relationships between sjTRECs and the $\log_{10}$-transformed number of naïve $CD4^+$ and $CD8^+$ T cells, naïve T_{reg}, and iNKT cells. Regression lines were fitted using linear regression. Adjusted *P* values were obtained using the mixed model and based on the Kenward-Rogers F test.

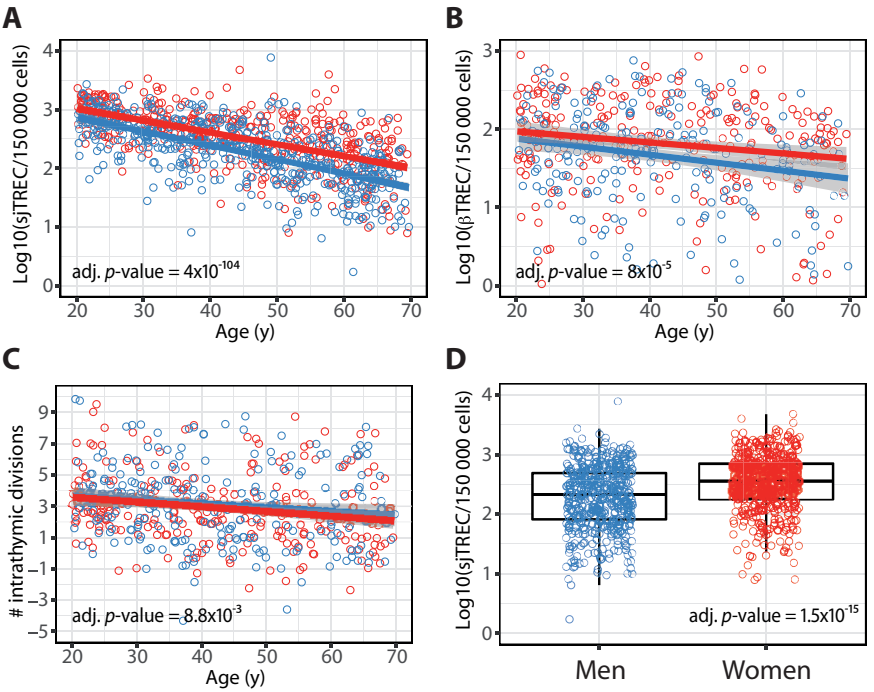

Figure 2. Age and sex strongly affect thymic function in healthy donors. (**A**) sjTRECs as a function of age in 487 women (red) and 492 men (blue). (**B**) βTRECs as a function of age, in 264 women and 242 men donors with detectable amounts. (**C**) Number of intrathymic divisions as a function of age, in donors with detectable βTRECs. (**D**) sjTRECs as a function of sex in 487 women and 492 men. Regression lines were fitted using linear regression. *P* values were adjusted to control the false discovery rate at 5% and estimated on the basis of the Kenward-Rogers approximate F tests.

2.3 Association of a genetic variation at the *TCRA-TCRD* locus with sjTRECs.

We next conducted a genome-wide association study of Log_{10}-transformed sjTREC numbers on 5,699,237 common single-nucleotide polymorphisms (SNPs) with a linear mixed model adjusted for age, sex, genetic relatedness, and other covariates selected using a data-driven variable selection scheme [Zhou and Stephens, 2014]. No association was detected at genome-wide significance (LRT, $P < 5.0 \times 10^{-8}$). Nevertheless, seven independent genomic regions on chromosomes 2, 4, 5, 10, 11, 14, and 17 showed suggestive evidence for association (LRT, $P < 1.0 \times 10^{-5}$; Fig. 3A). To test for replication of these suggestive associations, we measured sjTRECs in an independent cohort, the Marseille Thrombosis Association study (MARTHA) cohort, which includes 612 unrelated patients of European descent affected with venous thromboembolism [Morange et al., 2011]. We validated in this cohort the association of decreased sjTRECs with increasing age (4.05% per year; CI, 3.55 to 4.56%; K-R F test, $P = 5 \times 10^{-45}$; fig. S8A), and their higher abundance in women, relative to men (86%; CI, 60 to 116%; $P = 1.6 \times 10^{-15}$; fig. S8B). Among 14 SNPs tagging the seven suggestively associated loci, only variants on chromosome 14 showed statistical evidence for replication in the MARTHA cohort (table S2). These variants all mapped within a 25-kb region included in the *TCRA-TCRD* locus (Fig. 3B).

To fine-map the signal, we genotyped the eight most informative imputed SNPs within this region in the MI cohort and combined these data with array-based or imputed genotype data from the MARTHA cohort. This led us to identify four SNPs in linkage disequilibrium (rs8013419, rs10873018, rs12147006, and rs2204985) with genome-wide statistical significance (DerSimonian and Laird meta-analysis, $P < 2 \times 10^{-8}$; table S2) located in the intergenic *DD2* and *DD3* segments (Fig. 3C). Among them, rs2204985 (located 472 bases upstream of *DD3*) was considered the most likely candidate variant (effect allele frequency of 0.49; meta-analysis, $P = 1.9 \times 10^{-8}$; table S2) because it is located in an open-chromatin region targeted by the transcription factors Runt-related transcription factor 3 (RUNX3), E74-like factor 1 (ELF1), FOXM1, and RNA polymerase II according to the Encyclopedia of DNA Elements (ENCODE) consortium reference data set (Fig. 3D) [ENCODE Project Consortium et al., 2012].

2.4 Influence of the *TCRA-TCRD* genetic polymorphism on T cell development in immunodeficient mice.

Immunodeficient mice engrafted with human hematopoietic stem cells (HSCs) are able to develop a diverse repertoire of thymus-dependent human T cells [Halkias et al., 2015]. To directly evaluate in vivo the impact of the rs2204985 polymorphism on thymopoiesis, we reconstituted immunodeficient Balb/c *Rag2*$^{-/-}$*Il2rg*$^{-/-}$*Sirpa*^{NOD}(BRGS) mice [Lopez-Lastra et al., 2017] with human CD34^{+} hematopoietic progenitors harvested from fetal livers having different genotypes

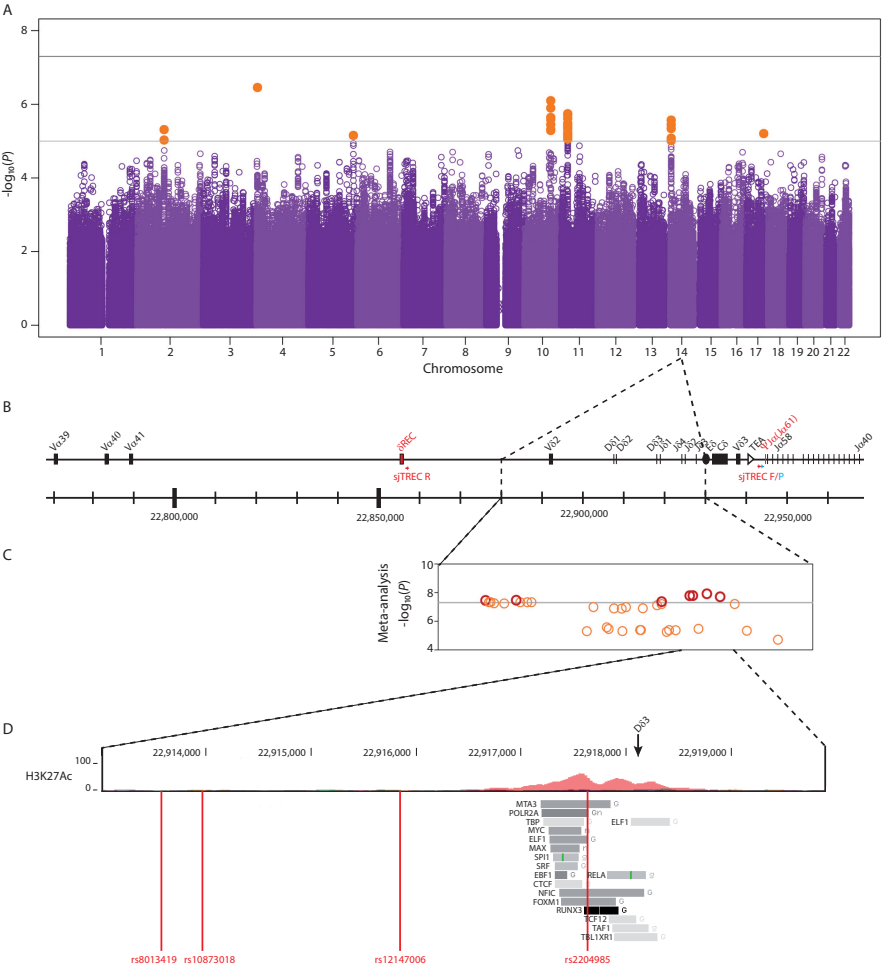

Figure 3. Genome-wide association study reveals an impact of *TCRA-TCRD* genetic variation on thymic function. (**A**) Manhattan plot for genetic association with sjTRECs in the 969 donors of the MI cohort. Light and dark gray lines indicate the threshold for suggestive association ($P = 1.0\times10^{-5}$) and genome-wide significant association ($P = 5.0\times10^{-8}$), respectively. (**B**) Detailed view of the *TCRA-TCRD* locus. Primers (sjTREC-F/R) and probe (sjTREC-P) used to quantify sjTRECs are shown in red and cyan, respectively. (**C**) Fine mapping of the genetic association between the *TCRA-TCRD* locus and sjTRECs. Meta-analysis *P* values were obtained by combining array-based, probe-based, and imputed genotypes of the MI and MARTHA cohorts (table S2). Variants that are significantly associated at the genome-wide level are indicated in red. (**D**) Physical position of the four most strongly associated variants, relative to active transcription activity [ENCODE Project Consortium et al., 2012]. The position of Dδ3 is indicated.

for the rs2204985 variant (Fig. 4A). Controlling for mouse recipient sex in a linear model, we observed significantly higher sjTRECs (multiplicative effect size CI, 1.24 to 2.16; t test, $P = 7 \times 10^{-4}$; Fig. 4B) and total $CD3^+$ thymocyte numbers (multiplicative effect size CI, 1.63 to 3.92; $P = 9 \times 10^{-5}$; Fig. 4C) in thymi of mice reconstituted with $CD34^+$ progenitors of the rs2204985 GG genotype, as compared to mice reconstituted with AA or GA genotypes. Significant results were also obtained when controlling for the origin of the human fetal liver sample (1.6 times increase in sjTRECs: CI, 1.1 to 2.5; K-R F test, $P = 0.047$; 2.5 times increase in thymocytes: CI, 1.54 to 4.25; $P = 6.6 \times 10^{-3}$). We next studied thymocyte developmental stages on human $CD45^+$ cells by flow cytometry (fig. S9). We observed that mice grafted with cells from rs2204985 genotype GG donors had larger thymocyte counts at all stages, starting as early as the $CD3^-CD4^-CD8^-$ DN population (Fig. 4D). These data support the hypothesis of a T cell-intrinsic effect of the identified genetic variant, which associates with thymocyte counts.

As shown in fig. S1, sjTRECs are produced by the δ*Rec*-ψ*Jα(Jα61)* recombination leading to the prominent *TCRD* locus deletion. However, there are alternative rearrangements including the one between δ*Rec* and *Jα58* gene segments that represents 23% of total δ*Rec* rearrangement (fig. S1B) [Verschuren et al., 1997]. We found a similar effect of the rs2204985 genotype on the alternative δ*Rec*-*Jα58* rearrangement as on sjTRECs (fig. S10), excluding an effect of the genetic variant on the *Jα* segment usage during primary *TCRA* rearrangements. Evaluating the *TCRA-TCRD* repertoire diversity according to rs2204985 genotypes (table S3), we found that the numbers of total and productive rearrangements did not differ (Mann-Whitney U test, $P > 0.05$). We found no specific overlap of *TCRA-TCRD* clonotypes, as calculated by the Morisita index, between mice grafted with the same fetal liver $CD34^+$ cells or even with the same rs2204985 genotype (fig. S11). Conversely, repertoire diversity, as quantified by productive clonality or Shannon equitability indexes, was significantly greater in mice grafted with cells of the GG genotype (Mann-Whitney U test, $P = 0.016$ and $P = 0.003$, respectively; table S3). Whereas no differences in *TCRAV* and *TCRAJ* gene segment usage were observed among mice grafted with cells of the AA or GG genotypes (Mann-Whitney U test, $P > 0.05$; Fig. 5, A and B), large differences were found in *TCRDV* and *TCRDJ* usage, with a preferential usage of gene segments close to the variant region (*DJ*, *DV2*, and *DV3*) in rs2204985 AA individuals (adjusted $P < 0.05$; Fig. 5B). Accordingly, the calculated frequency of T cells carrying a productive *TCRD* rearrangement was higher in AA individuals (Mann-Whitney U test, $P = 0.012$; Fig. 5C). A more detailed analysis of *TCRDV* and *TCRDJ* usage restricted to productive *TCRD* rearrangements showed that *DV1*, *DD2*, and *DJ1* segments were used preferentially in GG, whereas *DV2*, *DD3*, and *DJ3* were used preferentially in AA individuals (Fig. 5D), confirming that the rs2204985 variant locally affects *TCRD* rearrangements.

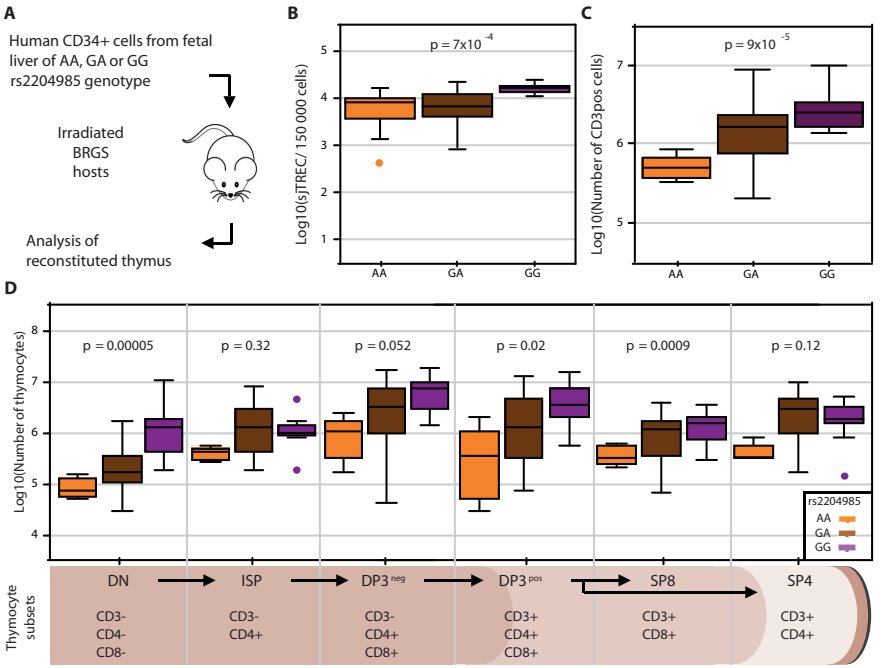

Figure 4. Effect of *TCRA-TCRD* human genetic variation on thymic function in humanized immunodeficient mice. (**A**) Immunodeficient Balb/c $Rag2^{-/-}Il2rg^{-/-}Sirpa^{NOD}$(BRGS) mice were reconstituted with human $CD34^{+}$ hematopoietic progenitors harvested from fetal livers with rs2204985 genotype AA (orange), GA (brown), or GG (purple). (**B**) Effects of rs2204985 genotypes on sjTRECs in all mice (AA, n = 19; GA, n = 58; GG, n = 15). (**C**) Effects of rs2204985 genotypes in immunophenotyped mice (AA, n = 5; GA, n = 31; GG, n = 13) on number of $CD3^{+}$ thymocytes and (**D**) on thymocyte subsets at different developmental stages. Indicated *P* values correspond to the genotype effect in a linear model including genotype and mouse recipient sex.

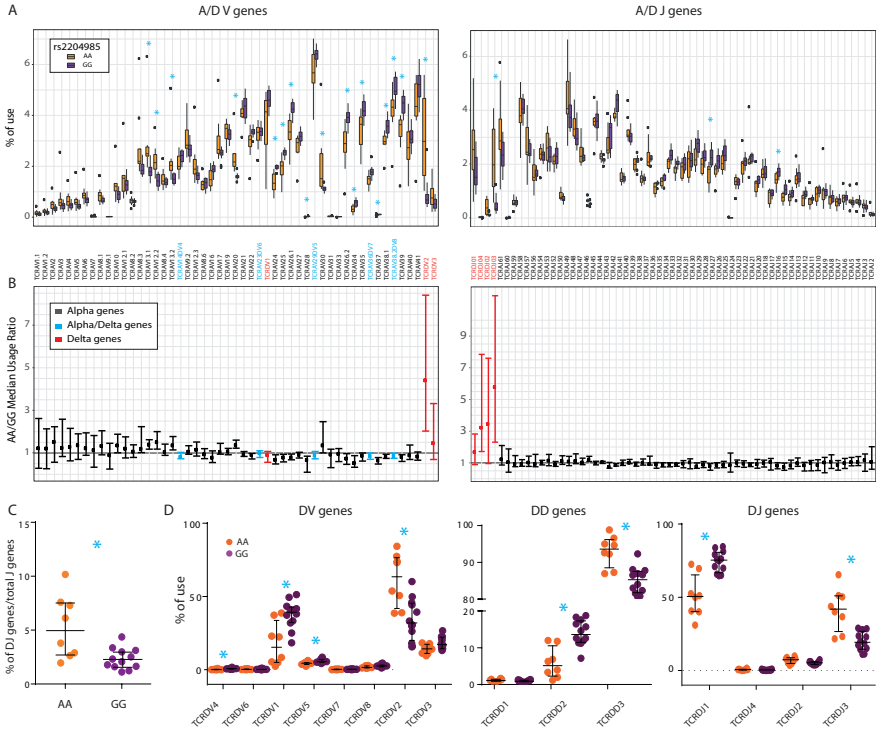

Figure 5. Effects of *TCRA-TCRD* human genetic variation on thymic TCR repertoire in humanized immunodeficient mice. The human *TCRA-TCRD* locus was sequenced using genomic DNA from 8 (3 males and 5 females) and 12 (4 males and 8 females) immunodeficient mice thymi grafted with rs2204985 AA (orange) and GG (purple) human fetal livers, respectively (table S3). (**A**) Effects of the donor genotype on V (left) and J (right) gene usage, among TCRα and TCRδ productive rearrangements. (**B**) Ratio of median percentage of V (left) or J (right) gene usage in GG-grafted mice, over that in AA-grafted mice. Gene segments used specifically by TCRδ are indicated in red, by TCRα in gray, and shared by both in cyan. Whiskers indicate bias-corrected and accelerated bootstrap 95% CIs. (**C**) Effect of genotypes on the percentage of *TCRD* specific J genes (*TCRDJ01* to *04*) among total *TCRD* and *TCRA J* genes used in productive rearrangements. (**D**) Effect of genotypes on the percentages of *DV* (left), *DD* (center), and *DJ* (right) genes usages among *TCRD* productive rearrangements. Genes are ordered according to their genomic location (see Fig. 3B). Blue asterisks indicate $P < 0.05$ obtained using nonparametric Mann-Whitney U test, adjusted for multiple testing using the false discovery rate as error rate.

2.5 Modeling the variance of thymic function in healthy adults.

Finally, we developed a model that estimates TREC content in healthy adults as a function of the rs2204985 genotype, age, and sex. We combined data of the MI and MARTHA cohorts in a mixed model, controlling for population stratification and batch variables. We found a 43% increase of sjTRECs in rs2204985 GG homozygotes, relative to AA homozygotes in the MI cohort (marginal CI, 22 to 69%; Fig. 6A). Similarly, in the MARTHA cohort, we found a 44% increase of sjTRECs in rs2204985 GG homozygotes, relative to AA homozygotes (marginal CI, 21 to 71%; Fig. 6B). The relative contribution of age, sex, and the rs2204985 variant to the variance of $\log_{10}$ sjTRECs was estimated to be 37.8, 4.78, and 1.32% in the MI cohort and 25.6, 8.5, and 1.3% in the MARTHA cohort, respectively (Fig. 6C). There was no indication that the effect of age on sjTRECs was dependent on rs2204985 genotypes (CI: 0.94 to 0.95, 0.94 to 0.96, and 0.94 to 0.96 for AA, GA, and GG, respectively, in MI; CI: 0.95 to 0.97, 0.95 to 0.96, and 0.95 to 0.98 for AA, GA, and GG, respectively, in MARTHA). We next sought to express the effect of the *TCRA-TCRD* genetic variation as a function of "thymic age," defined as the age of a male carrying the AA genotype with sjTRECs equal to those predicted by a linear model fitted on age, sex, and the rs2204985 genotype, using combined data of the MI and MARTHA cohorts. We then estimated the difference between actual age and thymic age for women and men carrying the GG genotype of 18.5 years (CI, 15 to 22.2) and 7.3 years (CI, 4.57 to 10.1), respectively (Fig. 6D). To support the application of rs2204985 genotyping in future clinical studies, we have developed the Shiny application allowing interactive visualization of the MI data (`https://mithymus.pasteur.fr`) (fig. S12).

3. Discussion

The thymus is the primary lymphoid organ where T lymphocytes are generated in the adaptive immune system of all vertebrates, through spatiotemporal interactions between thymocytes and specialized microenvironments [Abramson and Anderson, 2017]. The thymus is sensitive to insults received throughout life upon inflammation and infections, reflected in its functional decline with age [Palmer et al., 2018; Ferrando-Martínez et al., 2013; Douek et al., 1998]. It is, however, an extremely plastic tissue endowed with endogenous regenerative capacities after an acute damage during chemotherapies or irradiation [Boehm and Swann, 2013; Lopes et al., 2017; Wertheimer et al., 2018]. However, the parameters that control the levels of thymic function in homeostatic conditions remain largely unknown, an unmet need to develop precision and regenerative medicine. Here, by combining TREC quantification and a population immunology approach, we report the assessment of nongenetic and genetic determinants of thymic function in healthy adults.

sjTRECs are produced by the thymus and diluted out during T cell divisions [Douek et al., 1998]. Taking into account the dynamics of TRECs and peripheral

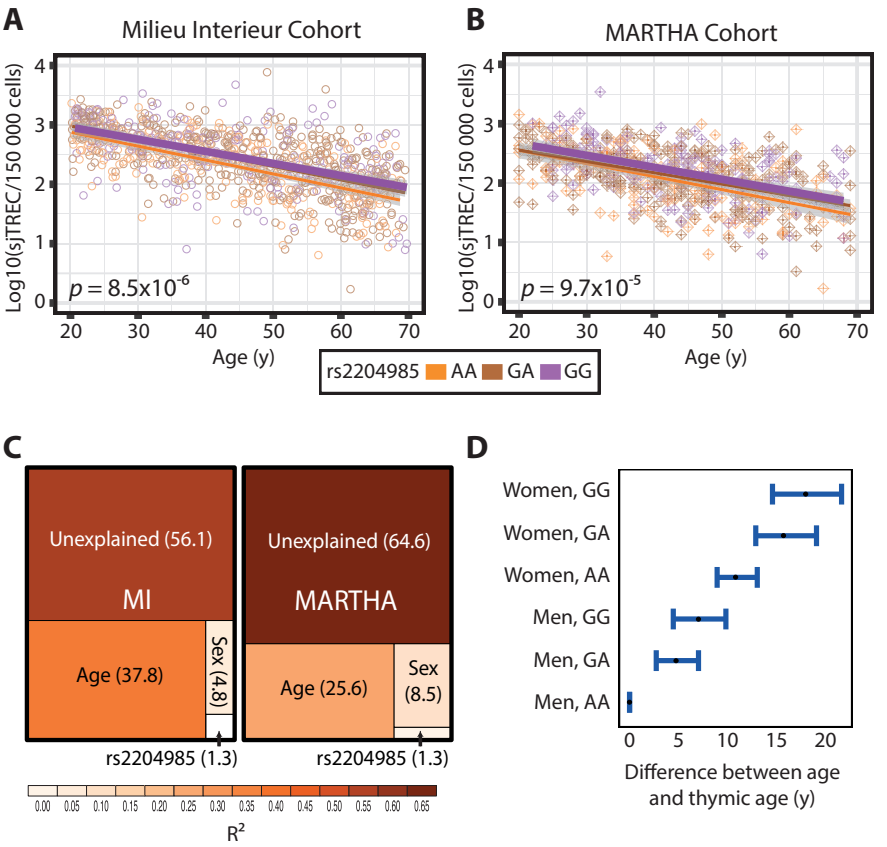

Figure 6. Combined effects of sex, age, and *TCRA-TCRD* genetic variation on human thymic function. sjTRECs as a function of age and rs2204985 genotypes in (**A**) the MI cohort (n = 969) and (**B**) the replication MARTHA cohort (n = 612). Regression lines were fitted using linear regression. *P* values were obtained with a mixed model of $\log_{10}$(sjTRECs), including rs2204985 genotypes as predictor; covariates were selected using a data-driven variable selection scheme; and correcting for population stratification was performed using the genetic relatedness matrix (GRM) as a random effect. Orange, brown, and purple indicate AA, GA, and GG genotype, respectively. (**C**) Proportions of variance of sjTRECs explained by age, sex, and *TCRA-TCRD* genetic variation in MI (left) and MARTHA (right) cohorts. The surface area and color of subrectangles indicate proportions attributed to specific predictors, as measured by the R^2 of the regression model. (**D**) Difference between actual age and thymic age as a function of sex and rs2204985 variant. Thymic age is predicted from a regression model, where AA men are assumed as the baseline.

cell division in young healthy individuals, an average of 4% per year involution in thymic output was previously estimated on the basis of the dynamics of TRECs and peripheral cell division in young healthy individuals [Bains et al., 2009], which is in line with our values of a 4.9 and 4% decrease per year in sjTREC amounts in the MI and MARTHA cohorts, respectively. In addition, our study in immunodeficient mice reconstituted with human $CD34^+$ HSC allowed a direct investigation of the impact of *TCRA-TCRD* genetic variation on the developing thymocytes independent of any peripheral dilution of TRECs.

The only nonheritable factors that we found strongly affecting thymopoiesis in the healthy population were age and sex, with a higher thymic function in women relative to men. Previous studies reported higher thymic mass, as measured by computed tomography in young (20 to 30 years old) women relative to young men [Ackman et al., 2013]. We demonstrate that the impact of sex on sjTRECs is observed during all of adulthood. In mice, androgens have a direct detrimental effect on stromal TECs [Rode and Boehm, 2012], and male cortical TECs express low levels of genes implicated in thymocyte expansion and positive selection [Dumont-Lagacé et al., 2015]. We suggest that the sex differences observed in our study could similarly reflect sex differences in TEC function, resulting in a more efficient bilateral cross-talk between thymocytes and thymic stroma and higher thymopoiesis in women [Abramson and Anderson, 2017]. Overall, the strong and replicated effect of sex on TREC content reinforces the need of stratifying immunological studies by sex [Markle and Fish, 2014].

Twin studies reported that RTE numbers are highly heritable, although no genetic associations have been found so far [Roederer et al., 2015]. In addition, naïve $CD27^+$ CD4 T cell counts have a high estimated heritability in healthy twins [Brodin et al., 2015] and in the MI cohort [Patin et al., 2018]. Collectively, these studies estimated a higher heritability of naïve rather than differentiated T cells in the adaptive compartment and suggest that T cell generation could be under genetic control. We found that TRECs, used as the closest readout of TCR rearrangements, are influenced by genetic variation at the *TCRA-TCRD* locus, which offers insights into the TCR locus function. The *TCRA-TCRD* locus is organized in a single genetic locus contributing to two different TCR specificities, TCR$\gamma\delta$ and TCR$\alpha\beta$. It therefore requires a complex program to regulate chromatin accessibility of *TCRA* and *TCRD* gene segments to the recombination machinery at two different developmental stages [Carico and Krangel, 2015]. The four most associated variants are located in a short segment spanning 4 kb within the *DD2* and *DD3* intergenic region, in a close 5' position to the TCRδ enhancer (Eδ). Our best candidate variant, rs2204985, is located in an open-chromatin region [ENCODE Project Consortium et al., 2012] close to a CCCTC-binding factor binding site, a critical element mediating chromatin looping and the access of the recombination machinery to the chromatin [Carico and Krangel, 2015; Chen et al., 2015].

Although the precise molecular mechanisms underlying the observed association of sjTRECs with genetic polymorphism will require further studies, the data col-

lected in immunodeficient mice experiments allow generation of some hypotheses. The TCRδ rearrangement is the first to occur at the earliest $CD34^{+}CD38^{-}CD1a^{-}$ DN stage [Dik et al., 2005] and is tightly ordered in humans, *DD2-DD3* rearrangements occurring before *DD2-DJ1* rearrangements [Cieslak et al., 2014]. δ*REC*-ψ*J*α rearrangements measured with sjTRECs are first detected in immature single-positive cells and reach peak levels in single-positive thymocytes [Dik et al., 2005]. We show in our study the effect of the rs2204985 variant, or a close genetic element in linkage disequilibrium, on DN thymocyte numbers before sjTREC generation and on *DV* and *DJ* usages. This supports a direct role of the genetic variation at an early stage of thymocyte differentiation. Our interpretation is that the higher sjTRECs and TCR diversity in mice engrafted with the rs2204985 GG genotype could relate to a higher rate of T cell generation starting at the early DN stage. The higher usage of *DV1*, as well as sjTRECs, in GG as compared to AA genotypes excluded a possible effect of the genetic variation on the reciprocal usage of *DV1* and δ*REC*. The Eδ element is a major regulator of *TCRD* accessibility in DN thymocytes [Monroe et al., 1999], functioning over a limited chromosomal distance [Bassing et al., 2003]. It has been suggested that Eδ may require additional upstream elements to promote *TCRD* accessibility [Bassing et al., 2003]. Hence, we hypothesize that the rs2204985 variant, or a genetic element nearby, could influence chromatin conformation and *TCRD* accessibility directly or through the binding of transcription factors and participate in the regulation of the *TCRD* recombination center [Zhao et al., 2016]. It will be interesting to investigate whether this polymorphism affects the generation of the different TCRγδ T cell subsets [Dik et al., 2005; Cieslak et al., 2014]. It remains also to be explained how the TCR genetic polymorphism could be linked to thymocyte survival or thymocyte proliferation at the DN stage. Notably, physiological DNA double-strand breaks generated in developing lymphocytes activate a broad transcriptional program [Bredemeyer et al., 2008], some of them promoting lymphocyte survival via, for instance, the activation of p38MAPK in DN thymocytes [Pedraza-Alva et al., 2006]. In addition, transcription factors binding the rs2204985 genomic region might affect DN survival/proliferation, such as FOXM1, which is required for cellular proliferation in normal cells [Bella et al., 2014]. It is intriguing to find evidence for genetic control of T cell generation in loci deleted in all mature peripheral T cells, TCRγδ through *TCRD* rearrangements and TCRαβ T cells through sjTREC generation. This suggests selective pressure at a critical step in T cell development, which might be otherwise unnecessary or possibly harmful if functional in the periphery. In support of the pathogenic potential of this genomic region is its proposed involvement during oncogene activation in T cell acute lymphoblastic leukemia [Le Noir et al., 2012].

By providing reference values of thymic function in a large healthy population via a key genetic control, our data provide a resource that may be useful in the context of precision medicine and regenerative strategies for diverse diseases. This study contributes to a better understanding of aging of the immune system, a major public health concern [World Health Organization, 2015]. We showed that the decrease

in thymic output with age was different in men and women and was independent of several environmental factors, including latent CMV infection, previously shown to associate with exhaustion of differentiated T cells [Patin et al., 2018; Nikolich-Zugich et al., 2017]. About 50% of the variance in sjTREC numbers remained unexplained, suggesting a role for still unknown environmental or genetic factors. Nonetheless, we showed differences in healthy thymic function depending on the *TCRA-TCRD* genetic variation in two independent cohorts of western European origin. It is important to estimate this impact in other ethnic groups, especially given the differences in frequency of the rs2204985 G allele across populations, ranging from 25% in East Asia to >80% in South America [The 1000 Genomes Project, 2015].

Considering the clinical implications of our findings, we anticipate that there may be settings where it would be beneficial to achieve a higher potential for T cell production. This would be the case, for instance, in an uncomplicated allo-HSCT setting or in the recovery of lymphopenic conditions in young patients. In contrast, it would be detrimental to fuel the system if the thymic environment is damaged as, for instance, in older individuals, in graft versus host disease in allo-HSCT [Krenger et al., 2011; Clave et al., 2009] or in autoimmune conditions where women are known to have an overall higher susceptibility [Dragin et al., 2016]. Such cases could result in the generation of T cells defective in their selection process with an autoreactive potential which could be pathogenic.

4. Materials and methods

Study design.

MI cohort.

The 1000 healthy donors of the MI cohort were recruited from September 2012 to August 2013 by Biotrial, stratified by sex (500 men and 500 women) and age (200 individuals from each decade between 20 and 69 years of age). Donors were selected on the basis of inclusion and exclusion criteria detailed elsewhere [Thomas et al., 2015]. To avoid the influence of hormonal fluctuations in women during the perimenopausal phase, only pre- or postmenopausal women were included. To avoid issues related to population stratification, the study was restricted to French citizens with Metropolitan French origin for three generations. The clinical study was approved by the Comité de Protection des Personnes-Ouest 6 on 13 June 2012 and by the French Agence Nationale de Sécurité du Médicament on 22 June 2012. The study is sponsored by the Institut Pasteur (Pasteur ID-RCB number: 2012-A00238-35) and was conducted as a single center study without any investigational product. The protocol is registered under ClinicalTrials.gov (study number NCT01699893). Primary data for the humanized mouse experiments are shown in table S5.

Replication cohort.

Our replication cohort included 612 patients from the MARTHA cohort [Morange et

al., 2011]. Donors were all of European descent and were examined between January 1994 and October 2005 for having suffered a single venous thrombosis event, without detectable cause. The study was approved by the Institutional Ethics Committee ("Département Santé de la Direction Générale de la Recherche et de lInnovation"; Projects DC: 2008880 & 09.576), and written informed consent was obtained from each subject. MARTHA biobank is hosted by the HEMOVASC bioresource center (CRB APHM). sjTRECs of all donors were quantified in DNA extracted from blood. Genotypes for candidate variants were obtained from the Illumina Human610-Quad SNP array [Morange et al., 2011] or probe-based genotyping.

Statistical analysis. We tested for association between TRECs and immunophenotypes, and TRECs and nonheritable factors by fitting linear mixed models, using the *mmi R* package (`https://github.com/jacobbergstedt/mmi`). The CIs were false coverage-adjusted intervals designed to keep the rate of false coverage at 5%. Hypothesis tests were done using K-R F tests with the false discovery rate as error rate. Impact of nonheritable factors on βTREC detection status was analyzed using logistic regression and LRTs. Genome-wide association studies were conducted using linear mixed models controlling for nonheritable variables and using the GRM as one of the correlation matrices. A similar model was used to compute effect sizes and 95% CIs for the rs2204985 polymorphism, age, and sex, with respect to sjTRECs in the MI cohort. For the MARTHA cohort, the four principal components of the genotype matrix that explained most variance were used instead of the GRM, and the hypothesis test was conducted using the K-R F test. The DerSimonian and Laird method was used to compute the meta-analysis *P* values. Both linear regression models and linear mixed models were used to compute 95% CIs and *P* values for the effect of the rs2204985 polymorphism on sjTREC numbers and thymus T cell progenitors in humanized mice. For gene segment usage, nonparametric 95% CIs were estimated by a bootstrap procedure. Thymic age and proportion of variance was estimated from a linear regression model with $\log_{10}$-transformed sjTRECs as response and age, sex, and the rs2204985 polymorphism as predictors.

DNA extraction from human whole blood. Blood was collected in 5 ml EDTA tube and was kept at room temperature (18–25°C) until processing. DNA extraction was performed using the Nucleon BACC3 kit (GE-Healthcare). Upon arrival at the processing site, blood was transferred into a 50 ml polypropylene tube. 20 ml of sterile Reagent A 1x (lysis buffer) was added to the blood sample in aseptic conditions and mixed by rotation for 4 minutes at room temperature. After red blood cell lysis, the tube was centrifuged 1300g for 5 minutes and the supernatant was discarded. The cell pellet was resuspended with 40 µl of PBS 1X, transferred to a 0.5 ml 2D-cap tube and stored at -80°C before processing. After thawing, 1 ml of sodium Reagent B was added directly to the cell pellet for resuspension before transfer to a 15ml screw capped propropylene centrifuge tube. 250 µl of sodium

perchlorate solution were then added for deproteinisation and the tube was mixed by inverting the tube at least 7 times. 1 ml of chloroform was then added to the tube, which was mixed by inversion. Without remixing the phases, 150 µl of Nucleon resin were added and the tube, which was then centrifuged at 1300g for 3 minutes. Without disturbing the Nucleon resin layer, the upper phase was transferred to a clean 15 ml tube and 2 volumes of cold absolute ethanol were added. The tube was then mixed by inversion until the DNA precipitate appeared. Using a heat-sealed Pasteur pipette, the precipitated DNA was hooked out and placed into a clean 1.5 ml microcentrifugation tube. 1 ml of cold 70% ethanol was added, the DNA was washed and the supernatant was discarded after centrifugation at 4000g for 5 minutes. After DNA pellet air dry for 10 minutes, 400 µl of deionized water were added and the tube kept overnight at 4°C to complete the resuspension before DNA quantification.

T-cell receptor excision circle (TREC) assays. sjTRECs and βTRECs are episomal circular DNAs generated during TCRα and TCRβ chain recombination, respectively (Fig. 1a and fig. S1). We implemented the protocol from Clave et al. [Clave et al., 2009], to quantify sjTRECs and βTRECs. The protocol is based on a quantitative PCR of genomic DNA extracted from whole blood, using the Biomark HD system (Fluidigm France, Paris, France). 1 to 2 µg of genomic DNA was preamplified for 3 minutes at 95°C and then 18 cycles of 95°C for 15 seconds, 60°C for 30 seconds and 68°C for 30 seconds, in a 50 µl reaction that contained the primers listed in table S4, 200 µM of each dNTP, 2.5 mM MgSO4 and 1.25 unit of Platinum Taq DNA pol High Fidelity (ThermoFisher Scientific, Courtaboeuf, France) in 1x buffer. Columns of 48.48 Dynamic array IFCs were loaded with 5 µl containing 2.25 µl of a 1/400th dilution of preamplified DNA, 2.5 µl of 2x Takyon Low Rox Probe MM (Eurogentec) and 0.25 µl of sample Loading Reagent and raws with an equal mixture of 2x Assay loading Reagent and 2x Assay Biomark that contains only the 2 primers and the probe specific for each assay. The sum of the 10 βTRECs was multiplied by 1.3 in order to take into account the 3 *Dβ1Jβ* that were not quantified (*Dβ2–Jβ2.5, 2.6* and *2.7*). sjTRECs and βTRECs were normalized to 150,000 cells (around 1 µg of DNA) using the Albumin gene quantification. As $\log_{10}$-transformed values of βTRECs ($\log_{10}$ βTRECs) showed a bimodal distribution, we analyzed it as either a binary variable (0 if undetectable; 1 if detectable), or a quantitative variable in donors with detectable βTRECs only ($\log_{10}$ βTRECs>0). When comparing βTRECs and flow cytometry measures, we included donors with levels below detection limit, and put their value at the detection limit. The number of intrathymic division was obtained using the $\log_2$ of sjTRECs/βTRECs ratio, and could thus be calculated in donors with detectable βTRECs only. When indicated, a shorter probe (sjTREC-LNA) was used for sjTREC quantification, which contains only the 26 first nucleotides of the standard probe and 4 Locked Nucleic Acids (LNA) in order to keep the same Tm as the original probe (table S4). From the 1000 MI donors, 979 DNA were available for TREC analysis and 10 were further excluded because of low Albumin quantification

(less than 50,000 copies).

DNA genotyping and imputation. The 1,000 subjects were genotyped at 719,665 SNPs by the HumanOmniExpress-24 BeadChip (Illumina, California) [Patin et al., 2018]. To increase coverage of rare and potentially functional variation, 966 of the 1,000 donors were also genotyped at 245,766 exonic SNPs by the HumanExome-12 BeadChip (Illumina, California). A total of 945,213 unique SNPs were thus genotyped. SNP quality-control filters yielded a total of 661,332 and 87,960 SNPs for the HumanOmniExpress and HumanExome BeadChips, respectively. The two datasets were then merged. Average concordance rate for the 16,753 SNPs shared between the two genotyping platforms was 99.9925%. The final dataset included 732,341 QC-filtered genotyped SNPs. Genotype imputation was performed by IMPUTE v.2 (54), considering 1-Mb windows and a buffer region of 1 Mb. After quality-control filters, a total of 11,395,554 high-quality SNPs were obtained, which were further filtered for minor allele frequencies >5%, yielding a final set of 5,699,237 SNPs for association analyses.

Probe-based genotyping of candidate variants. Eight SNPs (table S2) were selected for confirmation and fine mapping, on the basis of the availability of accurate TaqMan©assays and their strength of association. Genotyping was done on the Biomark HD (Fluidigm, San Francisco). Briefly, 60 to 120 ng of genomic DNA were preamplified with 0.2X of the 8 TaqMan Genotyping Assays (ThermoFisher Scientific, Massachusetts) and 1x Preamp Master Mix (Fluidigm) for 14 cycles (95°C for 15 sec, 60°C for 60 sec). The FLEXsix Genotyping IFCs (Fluidigm) were loaded with 1/50th dilution of the preamplified product, 2X Takyon Low Rox Probe MM (Eurogentec) and 40X TaqMan Genotyping Assay (ThermoFisher Scientific) according to manufacturer's instructions

Reconstitution of Human Immune System (HIS) mice.

Mouse strain.

HIS mice were generated in Balb/c *Rag2*$^{-/-}$*Il2rg*$^{-/-}$*Sirpa*^{NOD}(BRGS) recipients using human fetal liver hematopoietic stem cells as previously described [Lopez-Lastra et al., 2017]. Briefly, newborn mice (3 to 5 days of age) received sublethal irradiation (3 Gy) and were injected intrahepatically with the equivalent of 2.0×10^5 $CD34^+CD38^-$ human fetal liver cells. A total of 92 HIS mice in 15 independent experiments (4–10 mice per experiment) were analyzed at 8–29 weeks of age. Thymocytes and splenocytes were mechanically dissociated using a Cell Strainer (100µm nylon Falcon®). Cells (5.0×10^5) were frozen as dry pellet. DNA was prepared using the Proteinase K method (54°C for 1 min, 95°C for 10 min).

Human and mouse sex determination.

Sexing of human donors was made by single amplification of the *ZFX/ZFY* genes

in 25 µl PCR using of the hSex2 primer pair (table S4), 200 µM each dNTPs, 1.5 mM MgSO4 and 1 unit of HiFi Taq Platinum (Thermofisher). Cycling conditions were 94°C for 5 min and 40 cycles of 94°C for 30 sec, 56°C for 30 sec and 72°C for 2 min. PCR product was subsequently loaded on a 0.8% agarose gel giving a 1329 pb band for *ZFX* and a 906 pb band for *ZFY*. Sexing of mouse recipient was made single amplification of the *SRY/IL3* genes in 25 µl PCR using mSRY and mIL3 primer pairs, 200 µM each dNTPs, 1.5 mM MgSO4 and 1 unit of HiFi Taq Platinum (Thermofisher). Cycling conditions were 94°C for 3 min and 35 cycles of: 94°C for 15 sec, 57°C for 15 sec and 72°C for 30 sec, and, 72°C for 5 min. PCR product was subsequently loaded on a 1.5% agarose gel giving a 402 pb band for *SRY* and a 544 pb band for *IL3*.

Quantification of T-cell Receptor (TCR) Excision Circles (TRECs) in HIS mice.
TREC quantification adapted from Clave et al. [Clave et al., 2009] was performed in all 92 HIS mice samples. Real time quantification was made using ViiA7 (Applied Biosystems by Life Technologies, Austin, TX, USA) in 384-well plates loaded with 20 µl containing 5 µl of DNA (0.5 to 1 µg of genomic DNA), 10 µl of 2x Takyon Low Rox Probe MM (Eurogentec) and 5 µl of specific primer-probe mix (table S4). In addition, the quantification of the alternative rearrangement in the *TCR*δ locus (called δ*Rec-Jα58*) of TRECs in HIS mice was made using ViiA7 in 384-well plates loaded with 20 µl containing 5 µl of DNA (0.5 to 1 µg of genomic DNA), 10 µl of 2x Takyon Low Rox Probe MM (Eurogentec) and 5 µl of specific primer-probe mix (table S4). sjTRECs were normalized to 150 000 cells using the Albumin gene quantification.

DNA genotyping in HIS mice.
The samples were genotyped to the SNPs rs2204985 and rs10873018 with 5 µl containing 2 µl of DNA (10 to 20ng of genomic DNA), 0.25 µl of 2x Takyon Low Rox Probe MM (Eurogentec) and 2.5 µl of 40X TaqMan Genotyping Assay (ThermoFisher Scientific) according to manufacturer's instructions.

Immunophenotyping of human thymocytes in HIS mice.
Flow cytometry was performed in 49 of the 99 mice (22 males, 27 females), reconstituted from 8 human liver donors (3 males, 5 females) (table S5). Total thymus cell number was determined by trypan blue cell staining. Then, 10^6 cells were labelled with human specific CD45 PERCP-Cy5.5, CD8 PE, CD4 APC H7, CD3 V500 AmCyan and CD1a FITC monoclonal antibodies (All from BD Biosciences), then read on a FACSCanto II flow-cytometer. Data were analyzed using FACSDiva software (BD Biosciences).

Next Generation sequencing of the TCRAD locus recombination.
Sequencing of *TCRAD* locus in HIS mouse thymi was performed on genomic DNA at Adaptive Biotechnologies (Seattle, USA) using the immunoSEQ assay at the

survey level. Numbers of sequenced rearrangements are listed in table S3. Analysis of diversity was made using the ImmunoSEQ software tool and Morisitas index (fig S11) using the *R tcR* package.

Association between TREC measurements and immunophenotypes. Immunophenotyping was conducted on whole blood from all donors. Details on technical procedures and complete results are available in Patin et al. [Patin et al., 2018]. Ten 8-color flow cytometry panels were developed, allowing for the measurement of 168 traits, including 76 cell counts, 89 MFI and 3 ratios. Five additional measurements were obtained from flow cytometry raw data, to test the association of TRECs with HLA-DR memory T-cells. The minimum, median and maximum proportion of missing observations for the 173 variables were respectively, 0%, 7% and 25%. Missing data was imputed using the *missForest R* package [Stekhoven and Bühlmann, 2011].

We investigated the relationship between these immunophenotypes and sjTRECs and βTRECs as well as and the number of intrathymic divisions, by fitting linear mixed models. Substantial batch effects were observed in the flow cytometry data. To control for those batch effects, we included the day of blood draw as a random effect in the model. We also controlled for age, sex, cytomegalovirus (CMV) serostatus and smoking, the four factors having previously been shown to impact immunophenotypes [Patin et al., 2018]. Specifically, let y be one of the 173 immunophenotypes, x be either $\log_{10}$(sjTRECs), $\log_{10}$(βTRECs) or the number of intrathymic divisions. Then we fitted models that assume the following linear relationship for the ith individual:

$$\begin{aligned}\log y_i = \mu + x_i\beta + \mathrm{CMV}_i\beta_{\mathrm{CMV}} + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{Age}_i\beta_{\mathrm{Age}} \\ + \mathrm{ExSmoker}_i\beta_{\mathrm{ExSmoker}} + \mathrm{Smoker}_i\beta_{\mathrm{Smoker}} + \mathrm{SampleDay}_{j(i)} + \varepsilon_i, \qquad (5.1)\end{aligned}$$

with $\mathrm{SampleDay}_{j(i)} \sim \mathcal{N}(0, \sigma_b^2)$, $J(i) \in \{1, \ldots, 106\}$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and where CMV_i, Sex_i, $\mathrm{ExSmoker}_i$, and Smoker_i are indicator functions (*e.g.*, $\mathrm{CMV}_i = 1$ if individual i is CMV^+). βTRECs below the limit of detection were set to the detection limit.

Hypothesis tests on the parameters were performed using Kenward-Rogers approximated F-tests, implemented in the *pbkrtest R* package [Halekoh and Højsgaard, 2014]. The 173 x 3 = 519 tests conducted at this stage were considered a multiple testing family, and corresponding P values were corrected for multiple testing accordingly, using the false discovery rate (FDR) as error rate (figs. S2A and S3A). Confidence intervals of significantly-associated parameters (adj. $P < 0.05$) were constructed using the profile likelihood as implemented in the *lme4* package (Fig. 1B). The confidence intervals were adjusted for this selection to control the rate of false coverage over the selected intervals at 0.05 [Benjamini and Yekutieli, 2005]. All the flow cytometric data, together with source code used to perform analyses can be found in the *mmi R* package (`github.com/jacobbergstedt/mmi`).

Impact of non-genetic factors on TREC measurements. 56 physiological and demographic variables (table S1) that potentially affect thymic function as suggested in the literature and that showed a sufficient frequency for analysis in our cohort were selected from the detailed dataset available in the MI cohort [Thomas et al., 2015] (accessible on `http://www.synapse.org/MilieuInterieur`). Association between the 56 demographic variables and $\log_{10}$-transformed sjTRECs, $\log_{10}$-transformed βTRECs (in donors with detectable amounts), and the number of intrathymic divisions, was tested with linear mixed models. We estimated the false negative rate (1 − power) of these tests with the *pwr.f2.test* function of the *pwr R* package, assuming a sample size of $n = 969$, a significance level of $P = 0.05/56 = 9 \times 10^{-4}$ and proportions of variance explained by a demographic variable ranging from 0 to 1.

Because of their strong impact on TRECs and other candidate demographic variables, age and sex were included in all models. Specifically, let y be either $\log_{10}$(sjTRECs), $\log_{10}$(βTRECs) or the number of intrathymic divisions, and x one of the candidate demographic variables, except age and sex. Then we fitted mixed models that assume the following relationship for the ith individual:

$$y_i = \mu + f_x(x_i) + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{Age}_i\beta_{\mathrm{Age}} + \mathrm{TRECDay}_{j(i)} + \varepsilon_i,$$

with $\mathrm{TRECDay}_{j(i)} \sim \mathcal{N}(0, \sigma_T^2)$, $j(i) \in \{1, \ldots, 7\}$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, and where $f(x_i)$ is either a linear term when x_i takes ordered values, *i.e*, $f(x_i) = x_i\beta$, or a standard corner point parametrization with $K-1$ parameters, when x_i is a categorical variable with K levels. Removing the term related to x_i gives the model used for age and sex. For βTREC levels, only donors with detectable amounts were included, leaving 464 observations. A total of 56 x 3 = 168 hypothesis tests were performed, using the Kenward-Rogers F-test approximation implemented in the *pbkrtest R* package.

We also investigated the impact of non-genetic variables on the βTREC detection status, defined as a binary variable that equals 0 if βTRECs were not detected, or 1 if βTRECs were detected. Specifically, let y be the βTREC detection status, and x one of the candidate demographic variables, except age and sex. Then we fitted logistic regression models:

$$\mathrm{logit}(\mathbb{P}(y_i = 1)) = \mu + f_x(x_i) + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{Age}_i\beta_{\mathrm{Age}},$$

where $\mathrm{logit}(x) = \log(x/(1-x))$, using the *glm* function in *R*. Removing the term related to x_i gives the model used for age and sex. A total of 56 chi-square likelihood ratio tests were conducted using these models. In total 56 + 168 = 224 hypothesis tests were done. All these tests were considered as one multiple testing family, and corresponding P values were corrected for multiple testing accordingly, using the FDR as error rate.

Genotype-wide association studies. Univariate genome-wide association study of $\log_{10}$-transformed sjTRECs was conducted using the linear mixed model im-

plemented in GEMMA [Zhou and Stephens, 2014]. Genetic relatedness matrix was estimated with GEMMA, using a leave-one-chromosome approach. Because GEMMA cannot fit more random effect terms than the GRM, batch effects were not considered as random effects, as above, but as indicator terms. Covariates were selected using a variable selection scheme based around the elastic net and stability selection [Meinshausen and Bühlmann, 2010]. Among all demographic and batch effect variables, only age, sex and two TREC processing plates were selected and included as covariates in the model. The effect of a SNP on $\log_{10}$(sjTRECs) was tested with a likelihood ratio statistic, and considered to be genome-wide significant if $P < 5 \times 10^{-8}$.

Impact of the rs2204985 polymorphism on sjTRECs and thymocytes in humanized mice. To investigate the relationship between the rs2204985 polymorphism and sjTREC and thymocyte levels in mice we fitted the following linear models controlling for mouse sex:

$$\log \text{sjTREC}_i = \mu + \text{SNP}_i\beta_{\text{SNP}} + \text{Sex}_i\beta_{\text{Female}} + \varepsilon_i,$$

and

$$\log\left(\text{CD3}^+\text{CD45}^+_i\right) = \mu + \text{SNP}_i\beta_{\text{SNP}} + \text{Sex}_i\beta_{\text{Female}} + \varepsilon_i,$$

where for both models ε_i are normally distributed residuals. In this section we assume an additive genotype model, (i.e. SNP_i counts the number of G alleles. Inference for these models was done using Wald tests and confidence intervals.

Liver donor determines the genotype of the mice. It could also potentially affect phenotype values. The liver series origin could thus be a confounder, especially considering that we observed variability in sjTRECs and thymocyte counts between samples of different liver sample origin. We therefore also tested the impact of the rs2204985 genotype on mice sjTREC and thymocyte values when controlling for liver series origin. We did this by estimating a mean for each human liver donor using a random effect. This random effect incorporates possible differences in age and sex of the human liver samples. Since mouse sex has no significant effect on sjTRECs in mice ($P = 0.288$) we did not include it in this model. We therefore fitted the model:

$$\log \text{sjTREC}_i = \mu + \text{SNP}_i\beta_{\text{SNP}} + \text{Liver}_{j(i)} + \varepsilon_i,$$

where $\text{Liver}_{j(i)} \sim \mathcal{N}(0, \sigma_s^2)$ is the mean for the human liver sample grafted in mouse i, $j(i) \in \{1, \ldots, 16\}$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In contrast to sjTRECs, mouse sex showed an impact on thymocyte counts. This association remained even after controlling for human liver sample origin. To investigate the impact of the polymorphism on $\text{CD3}^+\text{CD45}^+$ values we therefore fitted a similar mixed model as above but included mouse sex as a covariate. All models were fitted using the *lme4 R* package. Confidence intervals were constructed using

the profile likelihood based method implemented in that package. For hypothesis tests on β_{SNP}, because of the low number of degrees of freedom for these data and models, we chose not to rely on large sample approximations, and used instead parametric bootstrapping of likelihood ratios.

Effect of the rs2204985 SNP on sjTRECs in the MI and MARTHA cohorts. To estimate the effect of the rs2204985 polymorphism on sjTRECs, we chose to compute its effect size and its confidence interval by fitting a new model, including both ancestry and batch variables as random effects. This was done in *R* using the *lme4qtl* package [Ziyatdinov et al., 2018]. The computations in this package is performed by the *lme4* package, which does not factorize the GRM prior to fitting models and is therefore not suitable for GWAS. Because our initial analyses showed that age and sex are the main non-genetic factors impacting sjTRECs, it was natural to fit the following model to compute the effect size of the SNP in the MI cohort:

$$\begin{aligned}\log_{10}(\text{sjTREC}_i) = \mu + \text{SNP}_i\beta_{\text{SNP}} + \text{Age}_i\beta_{\text{Age}} + \text{Sex}_i\beta_{\text{Female}} \\ + \text{Ancestry}_i + \text{TRECDay}_{j(i)} + \varepsilon_i,\end{aligned}$$

with $\text{Ancestry}_i \sim \mathcal{N}(0, \sigma^2_G\text{GRM})$, $\text{TRECDay}_{j(i)} \sim \mathcal{N}(0, \sigma^2_T)$, $j(1, \ldots, 7)$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $i = 1, \ldots, 969$. The *lme4qtl* package makes it possible to include a random effect term correlated according to a constant matrix like the ancestry term in the model above in a usual *lme4* model object. However, possibly due to this added functionality, we were not able to compute the confidence interval of β_{SNP} using the usual profile likelihood methods implemented in *lme4*. Instead we used a large sample Wald approximation.

We could not access the GRM matrix for the MARTHA cohort, and used instead the four principal components that explain the most variance of the SNP array data. To compute the effect size of the SNP in the MARTHA cohort, we therefore used the following model:

$$\begin{aligned}\log_{10}(\text{sjTREC}_i) = \mu + \text{SNP}_i\beta_{\text{SNP}} + \text{Age}_i\beta_{\text{Age}} + \text{Sex}_i\beta_{\text{Female}} + PC1_i\beta_{PC1} \\ + PC2_i\beta_{PC2} + PC3_i\beta_{PC3} + PC4_i\beta_{PC4} + \text{TRECPlate}_{j(i)} + \varepsilon_i,\end{aligned}$$

where $\text{TRECPlate}_{j(i)} \sim \mathcal{N}(0, \sigma^2_p)$ is the batch variable for TREC processing (what plate was used for sample i), $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $i = 1, \ldots, 590$. The model was fitted, and profile likelihood based confidence intervals was constructed, using the *lme4* package. Testing was done using Kenward-Rogers approximated F-tests implemented in the *pbkrtest* package in *R*.

Meta-analysis of the MI and MARTHA cohorts was conducted with the *rna.mi* function in the *metafor R* package [Viechtbauer, 2010]), with the random effects approach based on the DerSimonian and Laird estimator.

Thymic age estimation. To illustrate how thymic function is impacted by non-genetic and genetic factors, we sought to estimate the effect of an individual's genotype at rs2204985 and sex in terms of a quantity we call "thymic age", which we define for an individual as the age of the average male with the AA rs2204985 genotype with the same number of predicted sjTRECs as this individual. Genotypes in this context are considered a categorical variable that take values in $\{, \mathrm{AG}, \mathrm{GG}\}$. Consider the regression model:

$$\begin{aligned}&\mathbb{E}\left\{\log_{10}(\mathrm{sjTREC}_i) \mid \mathrm{Age}_i, \mathrm{Sex}_i, \mathrm{AG}_i, \mathrm{GG}_i\right\}\\ &\quad = \mu + \mathrm{Age}_i\beta_{\mathrm{Age}} + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{AG}_i\beta_{\mathrm{AG}} + \mathrm{GG}_i\beta_{\mathrm{GG}} + \mathrm{MARTHA}_i\beta_{MARTHA},\end{aligned}$$

where AG_i and GG_i are indicators that individual i is of genotype AA or GG. Because we pooled data of the MI and MARTHA cohort, we adjusted for cohort differences by including the MARTHA term, which equals 1 if individual i is part of the MARTHA cohort. The previous equation gives the following relationship for AA males:

$$\mathbb{E}\left\{\log_{10}(\mathrm{sjTREC}_j) \mid \mathrm{Age}_j\right\} - \mathrm{MARTHA}_j\beta_{MARTHA} = \mu + \mathrm{Age}_j\beta_{\mathrm{Age}}$$

The "thymic age" of an individual i is defined as the age of the AA male, *i.e*, Age_j, when

$$\begin{aligned}&\mathbb{E}\left\{\log_{10}(\mathrm{sjTREC}_i) \mid \mathrm{Age}_i, \mathrm{Sex}_i, \mathrm{AG}_i, \mathrm{GG}_i\right\}\\ &\quad = \mathbb{E}\left\{\log_{10}(\mathrm{sjTREC}_j) \mid \mathrm{Age}_j\right\} - \mathrm{MARTHA}_j\beta_{MARTHA}. \qquad (5.2)\end{aligned}$$

By the first two equations of this section this is equivalent to

$$\begin{aligned}&\mu + \mathrm{Age}_j\beta_{\mathrm{Age}} = \mu + \mathrm{Age}_i\beta_{\mathrm{Age}} + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{AG}_i\beta_{\mathrm{AG}} + \mathrm{GG}_i\beta_{\mathrm{GG}} \iff\\ &\mathrm{Age}_i - \mathrm{ThymicAge}_i = \frac{1}{\beta_{\mathrm{Age}}}\left(\mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{AG}_i\beta_{\mathrm{AG}} + \mathrm{GG}_i\beta_{\mathrm{GG}}\right).\end{aligned}$$

which gives an expression for the difference between actual age and the thymic age. The 95% CI for this quantity was computed by parametric bootstrapping.

Proportion of sjTRECs variance estimation. The contribution of rs2204985 genotypes, age and sex to the explained variance of $\log_{10}$ sjTREC values was estimated by the fitting linear regression model

$$\mathbb{E}\left\{\log_{10}(\mathrm{sjTREC}_i) \mid \mathrm{Age}_i, \mathrm{Sex}_i, \mathrm{SNP}_i\right\} = \mu + \mathrm{Age}_i\beta_{\mathrm{Age}} + \mathrm{Sex}_i\beta_{\mathrm{Female}} + \mathrm{SNP}_i\beta_{\mathrm{SNP}}$$

The proportion of variance explained by a particular predictor was estimated by averaging the sum of squares for that particular variable over different orderings in the regression model. The estimation was done using the *relaimpo R* package.

R Shiny interactive Web application. The interactive web application was developed using the Shiny package for the R statistical environment. The purpose of the webapp is to allow exploration of study results around the SNP rs2204985. The models developed in the study were exposed in the app to give users the ability to make their own predictions using either real or synthetic data. All predictions are stored temporarily during the user's session and may be downloaded at any time during the session. The app is hosted by the Institute Pasteur, Paris, and access is provided via the Milieu Intérieur Project hosted on Synapse (`http://www.synapse.org/MilieuInterieur`). A permanent, citable, link is provided (`https://mithymus.pasteur.fr/`).

References

Abramson, J. and G. Anderson (2017). "Thymic epithelial cells". *Annual review of immunology* **35**, pp. 85–118.

Ackman, J. B., B. Kovacina, B. W. Carter, C. C. Wu, A. Sharma, et al. (2013). "Sex difference in normal thymic appearance in adults 20–30 years of age". *Radiology* **268**:1, pp. 245–253.

Bains, I., R. Thiébaut, A. J. Yates, and R. Callard (2009). "Quantifying thymic export: combining models of naive T cell proliferation and TCR excision circle dynamics gives an explicit measure of thymic output". *The Journal of Immunology*, jimmunol–0900743.

Bassing, C. H., R. E. Tillman, B. B. Woodman, D. Canty, R. J. Monroe, et al. (2003). "T cell receptor (TCR) α/δ locus enhancer identity and position are critical for the assembly of TCR δ and α variable region genes". *Proceedings of the National Academy of Sciences* **100**:5, pp. 2598–2603.

Bella, L., S. Zona, G. N. de Moraes, and E. W.-F. Lam (2014). "FOXM1: a key oncofoetal transcription factor in health and disease". In: *Seminars in cancer biology*. Vol. 29. Elsevier, pp. 32–39.

Benjamini, Y. and D. Yekutieli (2005). "False discovery rate-adjusted multiple confidence intervals for selected parameters". *Journal of the American Statistical Association* **100**:469, pp. 71–81.

Boehm, T. and J. B. Swann (2013). "Thymus involution and regeneration: two sides of the same coin?" *Nature Reviews Immunology* **13**:11, p. 831.

Bredemeyer, A. L., B. A. Helmink, C. L. Innes, B. Calderon, L. M. McGinnis, et al. (2008). "DNA double-strand breaks activate a multi-functional genetic program in developing lymphocytes". *Nature* **456**:7223, p. 819.

Brodin, P., V. Jojic, T. Gao, S. Bhattacharya, C. J. L. Angel, et al. (2015). "Variation in the human immune system is largely driven by non-heritable influences". *Cell* **160**:1-2, pp. 37–47.

Carico, Z. and M. S. Krangel (2015). "Chromatin dynamics and the development of the TCRα and TCRδ repertoires". In: *Advances in immunology*. Vol. 128. Elsevier, pp. 307–361.

Chen, L., Z. Carico, H.-Y. Shih, and M. S. Krangel (2015). "A discrete chromatin loop in the mouse Tcra-Tcrd locus shapes the TCRδ and TCRα repertoires". *Nature immunology* **16**:10, p. 1085.

Cieslak, A., S. Le Noir, A. Trinquand, L. Lhermitte, D.-M. Franchini, et al. (2014). "RUNX1-dependent RAG1 deposition instigates human TCR-δ locus rearrangement". *Journal of Experimental Medicine* **211**:9, pp. 1821–1832.

Clave, E., M. Busson, C. Douay, R. Peffault de Latour, J. Berrou, et al. (2009). "Acute graft-versus-host disease transiently impairs thymic output in young patients after allogeneic hematopoietic stem cell transplantation". *Blood* **113**:25, pp. 6477–6484.

Dik, W. A., K. Pike-Overzet, F. Weerkamp, D. de Ridder, E. F. de Haas, et al. (2005). "New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling". *Journal of Experimental Medicine* **201**:11, pp. 1715–1723.

Dion, M.-L., J.-F. Poulin, R. Bordi, M. Sylvestre, R. Corsini, et al. (2004). "HIV infection rapidly induces and maintains a substantial suppression of thymocyte proliferation". *Immunity* **21**:6, pp. 757–768.

Douek, D. C., R. D. McFarland, P. H. Keiser, E. A. Gage, J. M. Massey, et al. (1998). "Changes in thymic function with age and during the treatment of HIV infection". *Nature* **396**:6712, p. 690.

Dragin, N., J. Bismuth, G. Cizeron-Clairac, M. G. Biferi, C. Berthault, et al. (2016). "Estrogen-mediated downregulation of AIRE influences sexual dimorphism in autoimmune diseases". *The Journal of clinical investigation* **126**:4, pp. 1525–1537.

Dumont-Lagacé, M., C. St-Pierre, and C. Perreault (2015). "Sex hormones have pervasive effects on thymic epithelial cells". *Scientific reports* **5**, p. 12895.

ENCODE Project Consortium et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* **489**:7414, p. 57.

Ferrando-Martínez, S., M. C. Romero-Sánchez, R. Solana, J. Delgado, R. de la Rosa, et al. (2013). "Thymic function failure and c-reactive protein levels are independent predictors of all-cause mortality in healthy elderly humans". *AGE* **35**:1, pp. 251–259.

Fink, P. J. and D. W. Hendricks (2011). "Post-thymic maturation: young T cells assert their individuality". *Nature Reviews Immunology* **11**:8, p. 544.

Halekoh, U. and S. Højsgaard (2014). "A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest". *Journal of Statistical Software* **59**:9, pp. 1–30.

Halkias, J., B. Yen, K. T. Taylor, O. Reinhartz, A. Winoto, et al. (2015). "Conserved and divergent aspects of human T-cell development and migration in humanized mice". *Immunology and cell biology* **93**:8, p. 716.

Kohler, S. and A. Thiel (2009). "Life after the thymus: CD31+ and CD31- human naive CD4+ T-cell subsets". *Blood* **113**:4, pp. 769–774.

Krenger, W., B. R. Blazar, and G. A. Holländer (2011). "Thymic T-cell development in allogeneic stem cell transplantation". *Blood*, blood–2011.

Kurd, N. and E. A. Robey (2016). "T-cell selection in the thymus: a spatial and temporal perspective". *Immunological reviews* **271**:1, pp. 114–126.

Le Noir, S., R. B. Abdelali, M. Lelorch, J. Bergeron, S. Sungalee, et al. (2012). "Extensive molecular mapping of TCRα/δ and TCRβ involved chromosomal translocations reveals distinct mechanisms of oncogenes activation in t-all". *Blood*, blood–2012.

Lopes, N., H. Vachon, J. Marie, and M. Irla (2017). "Administration of RANKL boosts thymic regeneration upon bone marrow transplantation". *EMBO molecular medicine* **9**:6, pp. 835–851.

Lopez-Lastra, S., G. Masse-Ranson, O. Fiquet, S. Darche, N. Serafini, et al. (2017). "A functional DC cross talk promotes human ILC homeostasis in humanized mice". *Blood advances* **1**:10, pp. 601–614.

Markle, J. and E. N. Fish (2014). "SeXX matters in immunity". *Trends in immunology* **35**:3, pp. 97–104.

Mathis, D. and C. Benoist (2009). "Aire". *Annual review of immunology* **27**, pp. 287–312.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**:4, pp. 417–473.

Miller, J. F. (2011). "The golden anniversary of the thymus". *Nature Reviews Immunology* **11**:7, p. 489.

Mitchell, W., P. O. Lang, and R. Aspinall (2010). "Tracing thymic output in older individuals". *Clinical & Experimental Immunology* **161**:3, pp. 497–503.

Monroe, R. J., B. P. Sleckman, B. C. Monroe, B. Khor, S. Claypool, et al. (1999). "Developmental regulation of TCRδ locus accessibility and expression by the TCRδ enhancer". *Immunity* **10**:5, pp. 503–513.

Morange, P.-E., T. Oudot-Mellakh, W. Cohen, M. Germain, N. Saut, et al. (2011). "KNG1 Ile581Thr and susceptibility to venous thrombosis". *Blood*, blood–2010.

Nikolich-Zugich, J., F. Goodrum, K. Knox, and M. J. Smithey (2017). "Known unknowns: how might the persistent herpesvirome shape immunity and aging?" *Current opinion in immunology* **48**, pp. 23–30.

Palmer, S., L. Albergante, C. C. Blackburn, and T. Newman (2018). "Thymic involution and rising disease incidence with age". *Proceedings of the National Academy of Sciences* **115**:8, pp. 1883–1888.

Patin, E., M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, et al. (2018). "Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors". *Nature Immunology* **19**:3, pp. 302–314.

Pedraza-Alva, G., M. Koulnis, C. Charland, T. Thornton, J. L. Clements, et al. (2006). "Activation of p38 MAP kinase by DNA double-strand breaks in V(D)J recombination induces a G2/M cell cycle checkpoint". *The EMBO Journal* **25**:4, pp. 763–773.

Puck, J. M. (2012). "Laboratory technology for population-based screening for severe combined immunodeficiency in neonates: the winner is T-cell receptor excision circles". *Journal of Allergy and Clinical Immunology* **129**:3, pp. 607–616.

Rode, I. and T. Boehm (2012). "Regenerative capacity of adult cortical thymic epithelial cells". *Proceedings of the National Academy of Sciences* **109**:9, pp. 3463–3468.

Roederer, M., L. Quaye, M. Mangino, M. H. Beddall, Y. Mahnke, et al. (2015). "The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis". *Cell* **161**:2, pp. 387–403.

Stekhoven, D. J. and P. Bühlmann (2011). "MissForest – non-parametric missing value imputation for mixed-type data". *Bioinformatics* **28**:1, pp. 112–118.

Stritesky, G. L., S. C. Jameson, and K. A. Hogquist (2012). "Selection of self-reactive T cells in the thymus". *Annual review of immunology* **30**, pp. 95–114.

Taub, D. D. and D. L. Longo (2005). "Insights into thymic aging and regeneration". *Immunological reviews* **205**:1, pp. 72–93.

The 1000 Genomes Project (2015). "A global reference for human genetic variation". *Nature* **526**, pp. 68–74.

Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, et al. (2015). "The milieu intérieur study – an integrative approach for study of human immunological variance". *Clinical Immunology* **157**:2, pp. 277–293.

Verschuren, M., I. Wolvers-Tettero, T. M. Breit, J. Noordzij, E. Van Wering, et al. (1997). "Preferential rearrangements of the T cell receptor-delta-deleting elements in human T cells." *The Journal of Immunology* **158**:3, pp. 1208–1216.

Viechtbauer, W. (2010). "Conducting meta-analyses in R with the metafor package". *Journal of statistical software* **36**:3.

Villartay, J.-P. de, R. D. Hockett, D. Coran, S. J. Korsmeyer, and D. I. Cohen (1988). "Deletion of the human T-cell receptor δ-gene by a site-specific recombination". *Nature* **335**:6186, p. 170.

Wertheimer, T., E. Velardi, J. Tsai, K. Cooper, S. Xiao, et al. (2018). "Production of BMP4 by endothelial cells is crucial for endogenous thymic regeneration". *Science immunology* **3**:19.

World Health Organization (2015). *World report on ageing and health*. World Health Organization.

Yang, H., Y.-H. Youm, B. Vandanmagsar, J. Rood, K. G. Kumar, et al. (2009). "Obesity accelerates thymic aging". *Blood* **114**:18, pp. 3803–3812.

Youm, Y.-H., T. L. Horvath, D. J. Mangelsdorf, S. A. Kliewer, and V. D. Dixit (2016). "Prolongevity hormone FGF21 protects against immune senescence by delaying age-related thymic involution". *Proceedings of the National Academy of Sciences* **113**:4, pp. 1026–1031.

Zhao, L., R. L. Frock, Z. Du, J. Hu, L. Chen, et al. (2016). "Orientation-specific RAG activity in chromosomal loop domains contributes to Tcrd V(D)J recombination during T cell development". *Journal of Experimental Medicine* **213**:9, pp. 1921–1936.

Zhou, X. and M. Stephens (2014). "Efficient multivariate linear mixed model algorithms for genome-wide association studies". *Nature methods* **11**:4, p. 407.

Ziyatdinov, A., M. Vázquez-Santiago, H. Brunel, A. Martinez-Perez, H. Aschard, et al. (2018). "Lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals". *BMC bioinformatics* **19**:1, p. 68.

Zuklys, S., A. Handel, S. Zhanybekova, F. Govani, M. Keller, et al. (2016). "Foxn1 regulates key target genes essential for T cell development in postnatal thymic epithelial cells". *Nature immunology* **17**:10, p. 1206.

Acknowledgments

We thank J. Fellay, P. Scepanovic, and C. A. W. Thorball for support with genetic analysis and J.-M. Doisne, G. M. Ranson, and H. Strick-Marchand for support with humanized mouse experiments. We thank V. Asnafi, E. Macintyre, A. Cieslak, and I. André-Schmutz for helpful discussions. We also thank the Centre dImmunologie Humaine (Institut Pasteur, Paris, France) for support.

Funding

This work was supported by the French governments Invest in the Future Program, managed by the Agence Nationale de la Recherche (ANR; 10-LABX-69-01). I.L.A. was a recipient of the Science without Borders PhD program from Brazil Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). J.B. is a member of the Lund Center for Control of Complex Engineering Systems (LCCC) Linnaeus Center and the Excellence Center at Linköping-Lund in Information Technology (ELLIT) Excellence Center at Lund University and is supported by the ELLIIT Excellence Center. J.P.D.S., Y.L., and S.L.-L. received funding from Institut Pasteur,

INSERM, the Laboratoire dExcellence REVIVE. A.T. and E.C. received funding from ANR grant PriCelAge (ANR-14-CE14-0030-01). Genetic investigations in the MARTHA study were supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013).

Author contributions

A.T. and M.L.A. conceived the project. E.C., I.L.A., and C.A. contributed to the design of experiments, supervised and conducted the experiments, generated the manuscript figures, and wrote the manuscript. E.P. and J.B. supervised and conducted statistical analyses, generated the manuscript figures, and wrote the manuscript. A.U., S.L.-L., Y.L., B.C., C.R.M., M.H., B.L.M.-L., and C.D. assisted in conducting experiments. N.S., M.G., D.-A.T., and P.-E.M. provided materials and data. M.F., D.D., and J.P.D.S. contributed to the supervision of the project and data analyses. A.T. and M.L.A. supervised the project, contributed to the design of the experiments and their analyses, and wrote the manuscript. M.L.A., L.Q.-M., and A.T. secured the funding. All authors reviewed and accepted the manuscript.

Data and materials availability

All data associated with this study are present in the paper or the Supplementary Materials. TCR repertoire data are accessible through ImmunACCESS database under the Genetic of Thymic function in hMice name project (`https://clients.adaptivebiotech.com/immuneaccess`). The SNP array data have been deposited in the European Genome-Phenome Archive with the accession code EGAS00001002460. The flow cytometric data and the code implementing statistical analyses can be downloaded as an R package (`http://github.com/JacobBergstedt/mmi`) and explored with the online Shiny application (`http://milieu-interieur.cytogwas.pasteur.fr/`).

Supplementary material

In the original article the materials and methods in Section 4 was put in the supplementary material. A version of the supplementary material without the materials and methods can be found in `http://lup.lub.lu.se/record/8506d40c-14f9-45d3-965b-eb49ff8ff918`. The original supplementary material can be found at `www.sciencetranslationalmedicine.org/cgi/content/full/10/457/eaao2966/DC1`.

There has been minor typographical changes in the supplementary material from the original published article

Paper III

Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles

Jacob Bergstedt Alejandra Urrutia Darragh Duffy
Matthew L. Albert Lluís Quintana-Murci Etienne Patin

Abstract

DNA methylation is a stable epigenetic alteration that plays a key role in cellular differentiation and gene regulation, and that has been proposed to mediate environmental effects on disease risk. Epigenome-wide association studies have identified and replicated associations between methylation sites and several disease conditions, which could serve as biomarkers in predictive medicine and forensics. Nevertheless, heterogeneity in cellular proportions between the compared groups could complicate interpretation. Reference-based cell-type deconvolution methods have proven useful in correcting epigenomic studies for cellular heterogeneity, but they rely on reference libraries of sorted cells and only predict a limited number of cell populations. Here we leverage >850,000 methylation sites included in the MethylationEPIC array and use elastic net regularized and stability selected regression models to predict the circulating levels of 70 blood cell subsets, measured by standardized flow cytometry in 962 healthy donors of western European descent. We show that our predictions, based on a hundred of methylation sites or lower, are less error-prone than other existing methods, and extend the number of cell types that can be accurately predicted. Application of the same methods to age, smoking consumption and several serological responses to pathogen antigens also provide accurate estimations. Together, our study substantially improves predictions of blood cell composition based on methylation profiles, which will be critical in the emerging field of medical epigenomics.

bioRxiv, October 2018.

1. Introduction

Cellular subtypes that compose organisms derive from various differentiation lineages during development. As stem cells differentiate into more specialized cells, their genome accumulates epigenetic modifications, *i.e.*, stable chemical additions to the DNA that can affect gene expression but do not change the DNA sequence, resulting in cell-specific gene expression. DNA methylation (DNAm), a stable epigenetic mark that refers to the attachment of a methyl group to DNA cytosine, plays a key role in cellular differentiation and gene regulation. Epigenome-wide association studies (EWAS) have searched for DNAm sites that covary with disease conditions or disease-related traits, as these DNAm changes could mediate the effects of environmental perturbations on the transcriptional reprogramming of differentiated cells and, in turn, organismal phenotypes [Jirtle and Skinner, 2007; Feil and Fraga, 2012]. However, interpretation of the results can be problematic, because statistical associations between DNAm and a condition of interest could be due to either a perturbation of the epigenetic properties of a cell subtype that causes the condition, or heterogenetiy in the proportions of differentiated cells caused by the condition [Rakyan et al., 2011; Lappalainen and Greally, 2017]. For example, because rheumatoid arthritis triggers a change in the granulocyte-tolymphocyte ratio, an EWAS of this disease identified thousands of associated DNAm sites that became non-significant upon correction for cellular heterogeneity [Liu et al., 2013]. Thus, there is a clear need in the epigenomics field for methods that reliably enumerate cell sub-populations from heterogeneous tissues [Teschendorff and Relton, 2018].

Currently, the gold standard approach for cell counting is flow cytometry, a laser-based technology that simultaneously detects several fluorescent-labelled protein markers at a single-cell resolution. However, this approach is labour-intensive and costly, requires skilled practitioners, and its performance is affected by sample degradation. Alternatively, cell composition can be indirectly estimated from gene expression profiles, which are known to be cell-specific [Newman et al., 2015; Shen-Orr and Gaujoux, 2013]. These methods, referred to as cellular deconvolution, rely on transcriptional profiles of reference cell populations to predict the cellular composition of sampled cell mixtures, which are also strongly affected by degradation and are difficult to standardize. In a seminal study, Houseman and colleagues used projection methods similar to the ones used for gene expression to estimate blood cell mixture proportions from DNAm profiles [Houseman et al., 2012], a more stable molecular measure. Because DNAm changes are thought to be involved directly in the lineage decision of hematopoietic cells [Álvarez-Errico et al., 2015; Deaton et al., 2011], they provide a direct link with blood cell identity. This method, referred to as the 'Houseman method' or 'Houseman model', uses DNAm profiles from six sorted cell subtypes as a reference, and assumes that the heterogeneous sample of interest is a mixture of these cells, whose proportions are estimated by projecting the sample matrix on to the reference matrix. The method can estimate the proportion of six major immune cells in blood, using a reference library of 600 CpG sites. Koestler *et*

al. proposed a refined reference library, called IDOL [Koestler et al., 2016], achieving better estimation of the six subsets with only 300 CpG sites. Although these methods have been extensively used, they only estimate six major cell subsets, and need at least 300 probes, limiting their usefulness as a tool for adequately controlling confounding in EWAS, and for applications in clinical research.

Here, we build novel parsimonious models for predicting the circulating levels of 70 blood cell subsets measured by flow cytometry in 962 healthy donors of the Milieu Intérieur study [Thomas et al., 2015; Patin et al., 2018], based on blood DNAm levels at >850,000 sites (Illumina Methylation EPIC beadchip; [Ait Kaci Azzou et al., n.d.]). The models are based on two key assumptions: 1) methylation at some sites marks differentiation events that can identify a particular blood cell lineage, and 2) only few methylation probes on the EPIC array mark such differentiation events. The first assumption implies a linear relation between the cell proportion in whole blood and methylation levels at the sites that mark it. We therefore use linear regression models to predict blood cell composition from DNAm levels. The second assumption means that only a small fraction of the probes will actually be predictive. We must therefore look for *sparse* models, which discard many of the included predictors in a data-driven fashion.

We use two approaches to build predictive models of immune cell proportions. The two assumptions mentioned above lead naturally to regularized linear regression models. Therefore, to infer optimal models in terms of prediction accuracy, and to investigate how prediction accuracy depends on the number of predictors, we use the elastic net method [Tibshirani, 1996; Zou and Hastie, 2005]. Similar models have previously been used for the prediction of age, smoking status, alcohol consumption and educational attainment based on DNAm [Zhang et al., 2018; McCartney et al., 2018]. We believe that elastic net regression will be able to find both the predictors that mark differentiation events for the lineage of the cell, but also the probes that are correlated with such predictors. In addition, we use the more stringent selection technique, *stability selection* [Shah and Samworth, 2013; Meinshausen and Bühlmann, 2010], to find a minimal stable set of predictors for each proportion. Stability selection selects predictors of each immune cell proportion that are consistently predictive in 100 subsamples of the dataset. We then build predictive models from the stability selected set of predictors using ordinary least squares. Compared with the elastic net, stability selection is more demanding of the predictors it selects. Consequently, it targets probes that mark differentiation events that are the most important for the cell. We therefore explore the biological functions associated with the stability selected probes, to improve knowledge of the epigenetic changes that characterize differentiated immune cells. A similar two-pronged approach is used to predict other conditions and traits collected within the Milieu Intérieur study, including age, smoking, height, BMI, routine chemical and hematological laboratory tests, and the serological responses to antigens of 13 common pathogens [Scepanovic et al., 2018]. Several of these traits have not previously been modelled using all DNAm probes jointly. Our study substantially improves predictions of blood cell composi-

tion based on DNA methylation profiles, which will be critical for applications in medical epigenomics, forensics and disease prognosis.

2. Results

2.1 Optimization of predictive models.

To predict immune cell proportions with optimal accuracy, given our assumption of sparsity and linearity, we use elastic net regularization. It is controlled by two regularization parameters: λ, which controls the $\mathbb{L}^1$ regularization that enforces sparsity on the coefficients, and α, which controls $\mathbb{L}^2$ regularization that restricts the magnitude of the coefficients. We use 5 different values for α and 200 different values for λ. Each possible pair of α and λ parameter values give a different amount of predictors and regularization, and is a step in the so-called *regularization path*. We measure the prediction accuracy along the regularization path by the mean absolute error (MAE) and the correlation (R) between the hold-out sample values and the out-of-sample predictions in 10-fold, twice repeated two-dimensional cross-validation, described in Algorithm 4. The procedure gives 20 samples from the distribution of out-of-sample prediction accuracy along the regularization path. We use those samples to estimate the mean accuracy and its 95% confidence intervals.

The performance of models, together with the number of predictors that is optimal in terms of prediction accuracy, is shown in Table 1 for each cell proportion, as well as 23 other continuous traits, including age and morphometric and physiological measures. DNA methylation levels can accurately predict age and sex [Zhang et al., 2018], intrinsic factors that are predictive of many traits, including immune cell counts in whole blood [Patin et al., 2018]. It is therefore important to discern when predictors based on methylation probes give additional information to these two commonly available factors. For comparison, we therefore include in Table 1 the prediction accuracy of a linear model that only includes age and sex as predictors. We also build predictive models for binary phenotypes, including smoking status and serostatus for 13 different common infections, using elastic net regularization together with the cost function of the binomial likelihood with a logit link function. Similarly to the approach we use for the continuous traits, we estimate prediction accuracy in terms of model complexity using cross-validation. For binary traits we measure prediction accuracy by the classification rate, *i.e.*, the proportion of correct class predictions (probability threshold is taken at 0.5). Prediction accuracy for models with optimally many predictors for the binary traits are shown in Table 2.

2.2 Blood cell deconvolution.

Accurate estimations are obtained with elastic net regularized models for 35 immune cell proportions (estimated correlation between predicted and observed out-of-sample values R>0.6; Table 1). The four immune cells that we predict with the highest accuracy are $CD8^+$ naive T cells, with a correlation between predicted and

Table 1. Mean absolute errors (MAE) and correlation (*R*) between out-of-samples predictions and observed values from three different predictive models for each trait. The first results are from elastic net models that have been tuned by our cross-validation scheme, detailed in Algorithm 4. Our scheme gives 20 samples from the distribution of accuracy estimates, which are used to construct confidence intervals. Results are also shown for stability selected linear models. Such models include only predictors that are robustly predictive of the trait. For comparison, the predictive accuracy of each trait is shown also for a simple linear model that only includes age and sex as predictors. Predictive accuracy estimates for the simple model for traits where the difference in *R* between the elastic net model and the simple model is less than 0.10 is shown in bold face. The character '#' stands for 'number of probes'. In the case of elastic net models, this is the mean number of probes for the 20 repetitions.
Table continued on next page.

	Elastic net					Stability selected linear model			Linear model with age and sex	
Trait	R	R CI	MAE	MAE CI	#	R	MAE	#	R	MAE
Age	0.99	[0.98, 0.99]	1.67	[1.5, 1.88]	701	0.98	2.00	38	**1**	**0**
%CD8$^+$ naive T cells	0.92	[0.87, 0.96]	0.38	[0.32, 0.48]	312	0.89	0.48	13	0.59	0.62
%B cells in CD45$^+$ cells	0.90	[0.8, 0.96]	0.41	[0.3, 0.53]	606	0.90	0.45	15	0.05	1.06
%CD8$^+$ T cells in CD45$^+$ cells	0.90	[0.84, 0.94]	0.84	[0.7, 0.97]	555	0.84	1.03	13	0.17	1.5
%NK cells in CD45$^+$ cells	0.88	[0.71, 0.95]	0.49	[0.41, 0.59]	1072	0.89	0.60	14	0.11	0.95
%T cells in CD45$^+$ cells	0.85	[0.75, 0.93]	2.24	[1.93, 2.7]	560	0.77	2.62	6	0.18	3.63
%CD4$^+$ naive T cells	0.85	[0.76, 0.92]	0.92	[0.72, 1.04]	662	0.86	1.07	13	0.18	1.71
%CD4$^+$ T cells in CD45$^+$ cells	0.85	[0.74, 0.93]	1.62	[1.29, 1.95]	338	0.77	1.93	12	0.21	2.59
%MAIT in T cells	0.84	[0.74, 0.91]	1.32	[1.12, 1.55]	973	0.84	1.48	15	0.32	2.44
CD4 CD8 ratio	0.82	[0.7, 0.88]	0.47	[0.38, 0.6]	102	0.91	0.15	18	0.22	0.28
%CD8a$^+$CD4$^-$ T cells	0.81	[0.59, 0.9]	0.41	[0.32, 0.49]	273	0.73	0.45	12	0.42	0.46
%monocytes in CD45$^+$ cells	0.81	[0.7, 0.85]	0.71	[0.64, 0.82]	480	0.70	0.89	13	0.03	1.03
%CD8b$^+$ T cells	0.80	[0.72, 0.88]	1.21	[0.98, 1.44]	57	0.70	1.57	5	0.14	1.59
%CD8$^+$ EMRA T cells	0.80	[0.7, 0.87]	0.29	[0.23, 0.38]	604	0.79	0.34	8	0.09	0.33
%naive in B cells	0.78	[0.7, 0.86]	6.10	[5.14, 7.03]	113	0.69	6.33	13	-0.03	7.95
%CD8a$^+$CD16^{hi} NK cells	0.77	[0.6, 0.87]	0.45	[0.36, 0.53]	218	0.74	0.43	13	0.04	0.56
%TCR$\gamma\delta^+$ in T cells	0.77	[0.64, 0.85]	1.49	[1.25, 1.76]	901	0.51	2.01	8	0.21	1.72
%CD14^{hi} monocytes in CD45$^+$ cells	0.77	[0.65, 0.83]	0.68	[0.6, 0.79]	690	0.64	0.86	12	0.08	0.93
%CD16^{hi} monocytes in CD45$^+$ cells	0.76	[0.69, 0.84]	0.22	[0.19, 0.25]	1447	0.68	0.24	8	0.29	0.2
%CD4$^+$ CM T cells	0.76	[0.65, 0.86]	1.31	[1.08, 1.65]	936	0.67	1.44	8	0.26	1.5
%conventional T cells	0.75	[0.64, 0.86]	2.26	[1.91, 2.97]	193	0.63	2.57	7	0.23	2.68
%CD8$^+$ MAIT cells	0.74	[0.6, 0.86]	0.18	[0.14, 0.23]	1402	0.55	0.24	10	0.14	0.25
%memory B cells in B cells	0.74	[0.65, 0.81]	4.34	[3.51, 4.99]	423	0.68	4.49	13	-0.03	5.81
Height	0.73	[0.64, 0.79]	5.00	[4.42, 5.57]	507	0.76	4.55	3	**0.78**	**3.83**
%eosinophils	0.72	[0.53, 0.89]	0.54	[0.45, 0.64]	372	0.64	0.58	6	0.06	0.64
%CD4$^+$ EM T cells	0.72	[0.61, 0.84]	0.36	[0.3, 0.42]	361	0.67	0.38	6	0.2	0.41
%marginal zone in B cells	0.71	[0.58, 0.83]	4.05	[3.38, 4.96]	100	0.52	4.42	13	0.03	4.63
%neutrophils	0.70	[0.51, 0.8]	5.47	[4.68, 6.28]	46	0.64	5.87	3	-0.01	6.89
%CD4$^-$CD8$^-$ MAIT cells	0.69	[0.52, 0.85]	0.15	[0.13, 0.18]	657	0.67	0.17	15	0.41	0.19
%basophils	0.69	[0.53, 0.81]	0.13	[0.11, 0.15]	599	0.63	0.15	6	0.25	0.14
%CD8$^+$ CM T cells	0.68	[0.56, 0.77]	0.58	[0.49, 0.69]	140	0.72	0.59	9	0.15	0.78
Red blood cells (hematology)	0.68	[0.61, 0.75]	0.25	[0.21, 0.29]	596	NA	NA	NA	**0.66**	**0.23**
%naive T_{reg}	0.67	[0.51, 0.77]	0.09	[0.08, 0.11]	443	0.33	0.13	3	0.34	0.09
%T_{reg}	0.67	[0.56, 0.76]	0.20	[0.17, 0.23]	554	0.30	0.24	3	0.18	0.19
Eosinophils (hematology)	0.66	[0.48, 0.86]	0.06	[0.05, 0.08]	893	0.59	0.07	5	0.04	0.06
%TCR$\gamma\delta^+$ cells	0.65	[0.41, 0.81]	0.41	[0.3, 0.55]	1498	0.62	0.36	5	0.32	0.37
%CCR6$^+$ cells	0.64	[0.54, 0.74]	0.67	[0.57, 0.8]	174	0.56	0.75	5	0.05	0.79
%transitional in B cells	0.61	[0.51, 0.72]	1.79	[1.42, 2.17]	76	0.50	1.84	8	-0.06	1.98
%memory T_{reg}	0.61	[0.41, 0.78]	0.13	[0.11, 0.15]	122	0.47	0.13	5	0.04	0.11
%CD8b$^+$CD45RA$^+$CD27$_{int}$ T cells	0.59	[0.44, 0.69]	0.34	[0.28, 0.39]	26	0.53	0.30	2	0.17	0.27

Table 1. Continued

	Elastic net					Stability selected linear model			Linear model with age and sex	
Trait	R	R CI	MAE	MAE CI	#	R	MAE	#	R	MAE
%CD8$^+$ EM T cells	0.58	[0.44, 0.67]	0.16	[0.12, 0.23]	25	0.65	0.18	4	0.08	0.17
Weight	0.57	[0.47, 0.65]	7.91	[6.78, 9.03]	299	0.51	7.61	1	**0.61**	**5.8**
Lymphocytes (hematology)	0.55	[0.41, 0.67]	0.33	[0.27, 0.38]	56	0.39	0.39	3	0.23	0.34
%CD8b$^+$CD4$^+$ T cells in CD45$^+$ cells	0.53	[0.4, 0.66]	0.04	[0.03, 0.04]	406	0.56	0.03	3	0.3	0.03
CD16^{hi} CD56^{hi} NK cell ratio	0.52	[0.35, 0.65]	6.26	[5.46, 7.25]	676	0.53	0.40	4	0.21	0.4
Cholesterol	0.45	[0.32, 0.56]	0.88	[0.78, 0.99]	50	0.38	0.79	2	**0.4**	**0.65**
Diastolic pressure	0.44	[0.26, 0.55]	7.03	[6.36, 8.05]	238	0.24	8.09	2	**0.37**	**5.39**
%live CD8b$^+$CD4$^+$ T cells	0.43	[0.28, 0.56]	0.05	[0.04, 0.06]	896	NA	NA	NA	0.1	0.04
Abdominal circumference	0.42	[0.27, 0.57]	7.21	[6.34, 7.94]	207	0.31	7.50	3	**0.4**	**6.53**
MCHC	0.42	[0.31, 0.55]	0.61	[0.53, 0.7]	422	0.43	0.61	2	0.08	0.71
%ILC2	0.42	[0.26, 0.64]	0.01	[0, 0.01]	212	NA	NA	NA	**0.37**	**0.01**
LDL	0.41	[0.28, 0.54]	0.80	[0.7, 0.92]	23	0.35	0.74	2	**0.4**	**0.6**
%activated T_{reg}	0.41	[0.3, 0.57]	0.06	[0.05, 0.07]	46	0.30	0.05	2	0.25	0.05
Systolic pressure	0.40	[0.25, 0.51]	10.76	[9.8, 12.2]	470	NA	NA	NA	**0.43**	**9.01**
%ILC	0.39	[0.21, 0.54]	0.05	[0.05, 0.06]	1028	0.27	0.05	2	0.16	0.04
%CXCR3$^+$ cells	0.35	[0.11, 0.54]	0.82	[0.71, 0.95]	50	NA	NA	NA	0.11	0.86
Basophils (hematology)	0.35	[0.22, 0.45]	0.01	[0.01, 0.01]	334	0.41	0.01	3	-0.02	0.01
%HLA$^-$DR$^+$ in CM CD8$^+$ T cells	0.35	[0.16, 0.54]	5.71	[4.91, 6.98]	1810	0.23	5.71	1	0.16	4.53
BMI	0.34	[0.21, 0.53]	2.48	[2.13, 2.79]	59	0.07	2.49	1	**0.29**	**2.15**
CRP	0.34	[0.11, 0.5]	1.44	[1.09, 2]	2143	0.19	1.28	2	0.05	1.27
%CD4$^+$ EMRA T cells	0.52	[0.26, 0.71]	0.15	[0.11, 0.22]	820	0.40	0.19	3	0.04	0.12
Neutrophils (hematology)	0.51	[0.33, 0.67]	0.93	[0.81, 1.1]	427	0.32	0.95	3	0.1	0.77
Monocytes (hematology)	0.51	[0.35, 0.64]	0.10	[0.09, 0.12]	105	0.49	0.09	4	0	0.09
HDL	0.50	[0.32, 0.6]	0.24	[0.22, 0.27]	229	NA	NA	NA	**0.47**	**0.19**
%cDC1	0.49	[0.3, 0.67]	0.13	[0.11, 0.17]	561	0.56	0.12	1	0.02	0.12
Leukocytes (hematology)	0.49	[0.27, 0.71]	1.17	[0.96, 1.39]	115	0.28	1.20	2	0.17	1.01
%CD14^{hi} monocytes	0.49	[0.33, 0.62]	0.89	[0.77, 1.03]	179	0.34	0.95	4	0.06	0.89
%CD69$^+$CD16^{hi} NK cells	0.49	[0.31, 0.64]	0.05	[0.04, 0.06]	792	0.22	0.07	4	0.15	0.04
%HLA$^-$DR$^+$ in EM CD4$^+$ T cells	0.49	[0.35, 0.62]	4.54	[3.81, 5.67]	1629	NA	NA	NA	0.33	3.97
%CD8a$^+$CD4$^+$ T cells	0.48	[0.3, 0.67]	0.13	[0.09, 0.2]	245	-0.06	0.13	1	0.02	0.11

observed out-of-sample values of R=0.92 (95% CI: [0.87, 0.96]), using 312 predictors (95% CI: [295, 338]); B cells (R=0.90, 95% CI: [0.8, 0.96]) using 606 predictors (95% CI: [582, 635]); CD8$^+$ T cells (R=0.90, 95% CI: [0.84, 0.94]) using 555 predictors (95% CI: [526, 591]); and natural killer (NK) cells (R=0.88, 95% CI: [0.71, 0.95]) using 1072 predictors (95% CI: [1036, 1126]). For most immune cell proportions, methylation levels clearly provide additional information in comparison to just age and sex.

A comparison of the performance of our elastic net models and the Houseman model, using either the standard or IDOL reference libraries [Houseman et al., 2012; Koestler et al., 2016], is given in Table 3. Our models outperform the two models for the six major cell-types that they are currently able to estimate. The correlations between predicted and observed out-of-sample values are systematically higher for our models, relative to the Houseman model with either the default or IDOL reference library (Table 3). Furthermore, our models are less error-prone (Table 3). These findings suggest that elastic net regression models, trained on whole blood standardized cytometry data, can outperform constrained projection techniques based on reference values obtained in a limited number of isolated blood

Table 2. Results from the logistic regression elastic net models and stability selected logistic regression models for binary traits. Logistic regression elastic net models were fitted using the logistic regression cost function together with elastic net regularization on the regression parameters. The regularization was tuned using our cross-validation scheme detailed in Algorithm 4. The column '#' gives the number of probes included in the model. The column CR gives the classification rate: how many of the out-of-sample classes that were correctly predicted. The naive prediction is to always guess the most prevalent condition. The percentage of people in the whole sample that belongs to the most prevalent class is given in the column "Prev". Methylation predictors only add something if they can improve on the naive prediction. This measure is given in the "Diff" column which is computed as "CR" - "prevalence".

		Elastic net				Stability selected linear model		
Trait	Prev	CR	CR CI	Diff	#	CR	Diff	#
Smoking	0.52	0.89	[0.82, 0.94]	0.38	192.90	0.64	0.12	3
CMV	0.65	0.87	[0.81, 0.94]	0.22	255.95	0.90	0.25	13
Toxoplasmosis	0.56	0.72	[0.65, 0.8]	0.16	1057.65	0.69	0.13	1
Hepatitis B	0.52	0.65	[0.59, 0.75]	0.13	337.65	0.69	0.17	1
Herpes Simplex 1	0.65	0.68	[0.59, 0.76]	0.03	5312.05	0.70	0.05	1
Helicobacter pylori	0.82	0.82	[0.77, 0.88]	0.00	144.25	0.84	0.02	1

Table 3. Mean absolute error (MAE) and correlation (R) between predicted and observed out-of-sample values compared between our elastic net models and the Houseman model with either the standard reference library or IDOL

	R			MAE		
Trait	Houseman	IDOL	Elastic net	Houseman	IDOL	Elastic net
B cells	0.875	0.853	0.898	3.043	2.752	0.408
$CD4^+$ T cells	0.783	0.824	0.847	2.207	3.522	1.619
$CD8^+$ T cells	0.824	0.852	0.896	6.368	3.768	0.840
Monocytes	0.786	0.784	0.806	3.994	2.304	0.713
NK cells	0.788	0.813	0.884	3.512	2.232	0.494
Neutrophils	0.634	0.634	0.703	11.377	10.751	5.471

cell sub-types.

2.3 Linear models selected by stability selection.

We next evaluate how prediction accuracy varies with the number of predictors in our models. The regularization paths for the nine best predicted traits are shown in Figure 1. Interestingly, out-of-sample prediction error decreases rapidly with the number of predictors, and plateaus at around 50 predictors (Figure 1). This

Table 4. Mean absolute error (MAE) and correlation (R) between predicted and observed out-of-sample values compared between our stability selected linear models and the Houseman model with either the standard reference library or IDOL

	R			MAE		
Trait	Houseman	IDOL	Stab. sel.	Houseman	IDOL	Stab. sel.
B cells	0.899	0.889	0.905	3.107	2.846	0.454
$CD4^+$ T cells	0.757	0.788	0.774	2.622	3.645	1.933
$CD8^+$ T cells	0.829	0.825	0.844	6.381	4.024	1.032
Monocytes	0.708	0.719	0.704	3.973	2.334	0.891
NK cells	0.799	0.802	0.891	3.591	2.259	0.600
Neutrophils	0.609	0.609	0.640	11.333	10.666	5.871

indicates that accurate predictions can be achieved with much fewer predictors than the hundreds of DNAm probes used by current prediction models of cell composition [Houseman et al., 2012; Koestler et al., 2016] and age [Horvath, 2013; Zhang et al., 2018]. These results suggest that blood cell composition can be predicted well using only a few number of probes that are markers for differentiation events. To find such probes, we estimate a minimal robust predictor set using stability selection. We select and build the models on a subsample of 866 randomly selected individuals, and then evaluate on a hold-out sample of 96 randomly selected individuals.

The predictive accuracy of the stability-selected predictive models is high (Table 1) and comparable to that of elastic net regression models, while using considerably fewer predictors. Prediction performance is also apparent when predicted out-of-sample values are plotted against the observed values for the 16 most accurate models (Figure 2). For instance, using only six methylation probes, the correlation between estimated and observed values for T cells is R=0.77 and the MAE is lower than 3%. We verify that our stability selected models are competitive by comparing their prediction accuracy to that of the Houseman model using either the standard or IDOL reference panels. Although our models use only 15, 12, 13, 13, 14 and 3 predictors for B cells, $CD4^+$ T cells, $CD8^+$ T cells, monocytes, NK cells and neutrophils, respectively, they yield comparable out-of-sample correlations and lower MAE (Table 4), relative to current methods. Together, these results demonstrate that prediction models that use a dozen or fewer methylation probes selected by stability selection can achieve prediction accuracy comparable to that of gold-standard, reference-based cell deconvolution techniques that use hundreds of probes.

2.4 Biological relevance of the stability selected methylation probes.

Because blood cell proportions could be accurately predicted with just a dozen of DNAm probes, we next investigate the relevance of the stability-selected probes to cell biology. We find several, methylome-wide significant DNAm probes that are

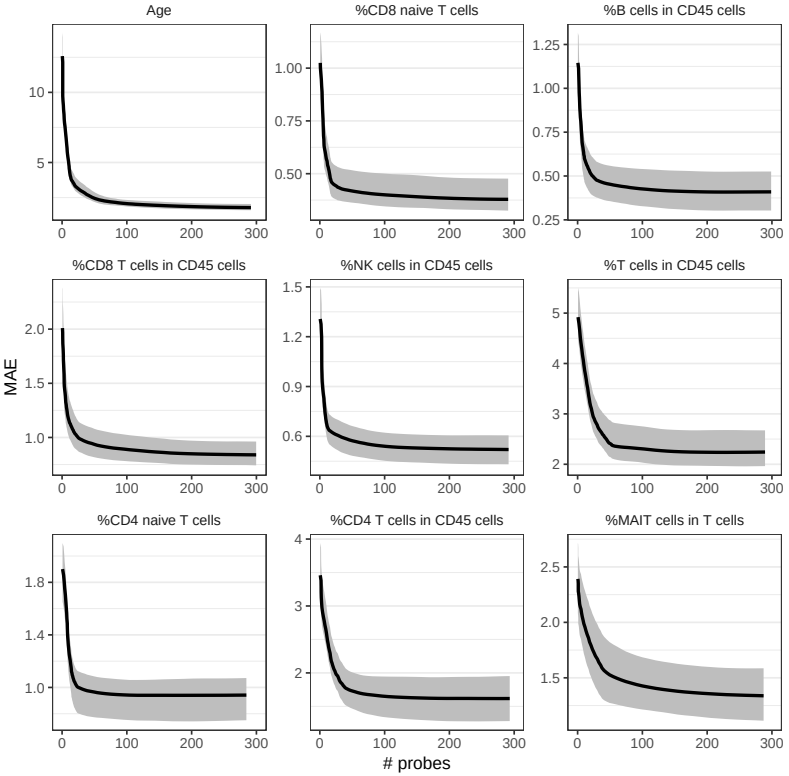

Figure 1. Mean absolute prediction errors (MAE) as a function of the number of predictors included in elastic net models predicting quantitative traits. The regularization parameter α is here set to 0.95. Confidence bands are estimated non-parametrically from the 20 samples of the prediction error given by our cross-validation scheme detailed in Algorithm 4.

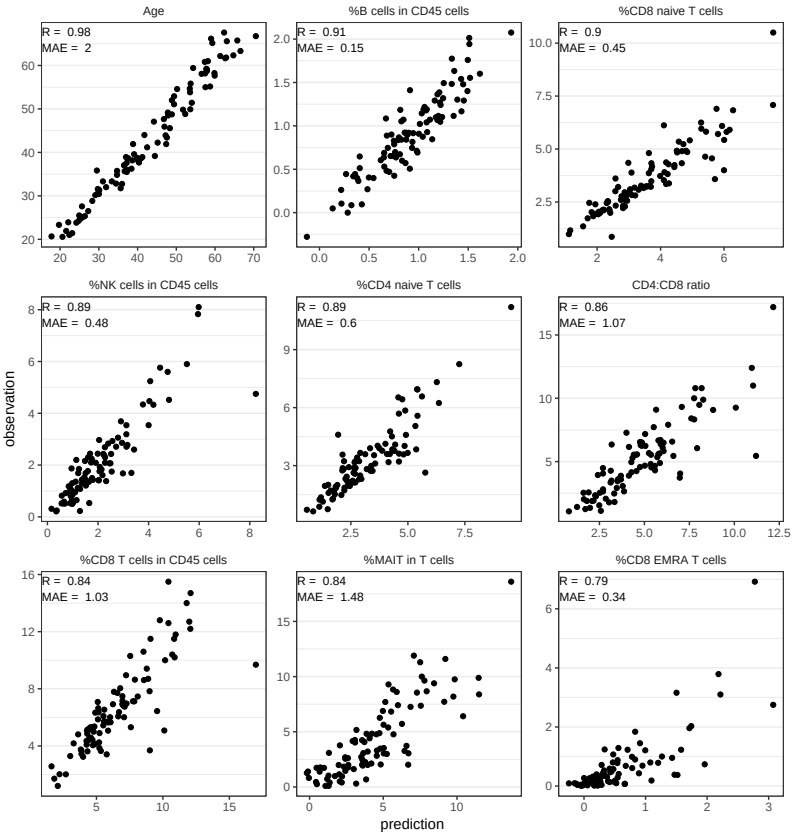

Figure 2. Out-of-sample predictions from stability selected linear models plotted against observed hold-out values. Plots are shown for the 9 best predicted traits. Observed and predicted values are obtained from the hold-out sample of 96 individuals.

found close to, or within, genes with well-known functions in immune cell differentiation (Table 5). For instance, DNAm levels within *CD4*, *CD8A* and *CD8B* genes are associated with the CD4:CD8 ratio (P=3.9x10^{-11}), the proportion of CD8a^{+} NK cells (P=4.6x10^{-17}) and the proportion of CD8b^{+} T cells (P=1.6x10^{-9}), respectively. The proportion of neutrophils are associated with DNAm levels in the *PDE4B* gene body (P=1.8x10^{-8}), which plays a key role in neutrophil function [Ariga et al., 2004]. Similarly, the proportion of MAIT cells are associated with DNAm levels in the 5'UTR of *IL21R* (P=8.3x10^{-21}), which is known to regulate MAIT cell numbers [Wilson et al., 2015]. Several cell sub-types, including leukocytes, lymphocytes, monocytes and ILC, are associated with DNAm sites within *AHRR*, *F2RL3* and *GATA3* genes, which are known to be strongly affected by cigarette consumption [Joehanes et al., 2016; Bojesen et al., 2017; Chatziioannou et al., 2017]. We consistently showed recently that circulating levels of these different blood cell subsets are significantly impacted by smoking status [Patin et al., 2018]. Finally, a number of the selected DNAm probes have previously been associated with disease (Table 5). For instance, DNAm within the *ACSF3* gene is associated with the proportion of naive B cells (P=4.1x10^{-9}) and has been shown to be differentially methylated in B cells of patients with rheumatoid arthritis [Julià et al., 2017], suggesting that B cell sub-type fractions are altered in these patients. Together, these findings support stability selection as a robust tool to select relevant associated variables, and illustrate the biological relevance of DNAm probes selected as predictors of immune cell proportions.

2.5 Prediction of other factors.

Among the other quantitative factors assessed in the Milieu Intérieur cohort, prediction by the elastic net method is the most accurate for age (Table 1). Using 701 predictors (95% CI: [673, 749]), we estimate age with an MAE of 1.67 years (95% CI: [1.5, 1.88]), confirming that it can be estimated from DNAm with high accuracy [Horvath, 2013; Zhang et al., 2018]. From Table 1 it appears that our elastic net models are also able to estimate red blood cell counts, height and weight with high accuracy. However, a comparison with the model that only uses age and sex reveals that the predictive power of methylation levels for these two traits probably mostly stem from their ability to predict age and sex.

We next evaluate the accuracy of elastic net models to predict, based on DNAm data, smoking status and the serostatus for 13 common infections, including infections by *Toxoplasma gondii*, *Helicobacter pylori*, cytomegalovirus (CMV), Epstein-Barr virus (EBV), hepatitis B virus (HBV), Herpes Simplex virus (HSV), Varicella Zoster virus (VZV), mumps virus and measles virus [Scepanovic et al., 2018]. Binary traits for which the prediction of the most prevalent condition outperforms the naive prediction are shown in Table 2. We obtain good prediction results for smoking consumption and CMV serostatus, which is natural considering that both factors have been shown to broadly affect immune cell variation [Patin et al., 2018]. The out-of-sample classifications for both of these traits are correct almost 90% of the

Table 5. Methylome-wide significant ($P < 3 \times 10^{-8}$) DNAm probes selected by stability selection for all continuous traits. A model was fitted for each trait with all predictors selected by stability selection. *P* values were then computed for each predictor. The predictors who had *P* values smaller than 3×10^{-8} are included in the table. *Table continued on next page.*

Trait	Methylation probe	Coefficient	Standard error	P value	Chr.	Position	Closest gene	Genic region	Published associations
Age	cg08097417	44.342	4.6133	8.33e-21	chr7	130419133	KLF14	TSS1500	Age (Florath et al, Hum Mol Genet 2014; Hannum et al, Mol Cell 2013)
Age	cg10501210	-10.188	1.196	7.62e-17	chr1	207997020	miR-29b-2		Age (Tserel et al, Immun Ageing 2014)
Age	cg22083892	-8.645	1.244	7.44e-12	chr12	21928661	KCNJ8	TSS1500	
Height	cg26020914	-22.118	2.6003	7.90e-17	chrX	18444359	CDKL5	5'UTR	
Weight	cg24447042	78.969	4.653	5.67e-56	chrX	128657893	SMARCA1	TSS1500	
Abdominal circumference	cg01243823	-43.54	5.1116	7.17e-17	chr16	50732212	NOD2	Body	BMI (Mendelson et al., PLoS Medicine 2017); Memory T cell differentiation (Komori et al., J Immunol 2015)
BMI	cg16740586	20.548	2.8828	2.15e-12	chr21	43655919	ABCG1	Body	BMI (Demerath et al., Hum Mol Genet 2015)
Diastolic pressure	cg09761247	-41.541	4.5998	1.09e-18	chrX	148585951	IDS	Body	
Diastolic pressure	cg19996355	189.541	25.531	2.73e-13	chr19	19729375	PBX4	1stExon	Age (Johansson et al., PloS One 2013)
HDL	cg23581718	-4.098	0.3211	2.74e-34	chrX	41173768			
LDL	cg22454769	2.719	0.4601	4.95e-09	chr2	106015767	FHL2	TSS200	Age (Li et al., Sci Rep 2017)
Cholesterol	cg16867657	5.29	0.3668	2.10e-42	chr6	11044877	ELOVL2	TSS1500	Age (Li et al., Sci Rep 2017)
CRP	cg12992827	-10.626	1.4894	2.06e-12	chr3	101901234			BMI (Wahl et al., Nature 2017)
CRP	cg00444883	15.865	2.5926	1.43e-09	chr8	92060568			
Leukocytes (hematology)	cg05575921	-4.585	0.7691	3.65e-09	chr5	373378	AHRR	Body	Smoking (Bojesen et al., Thorax 2017; Chatziioannou et al., Sci Rep 2017)
Lymphocytes (hematology)	cg04551776	-3.881	0.3825	6.18e-23	chr5	393366	AHRR	Body	Smoking (Chatziioannou et al., Sci Rep 2017)
Lymphocytes (CD3$^+$)	cg09736846	-29.648	4.175	2.70e-12	chr8	101443681			
Neutrophils (hematology)	cg18377866	-23.853	2.4474	2.31e-21	chr3	193965288	LOC101929337	TSS200	
Neutrophils	cg11674865	-96.313	7.0731	2.76e-38	chr1	161591488			
Neutrophils	cg17781418	-40.562	4.6305	1.07e-17	chr3	71305262	FOXP1	5'UTR	CD4$^+$ T cell differentiation (Garaud et al., Eur J Immunol. 2017)
Neutrophils	cg14973204	-3.129	0.4149	1.21e-13	chr12	133052753			
Neutrophils	cg10236264	-34.085	5.9929	1.78e-08	chr1	66793339	PDE4B	Body	Neutrophil function (Ariga et al., J Immunol 2004) ; Maternal glycemic response (Cardenas et al., Diabetes 2018)
Basophils (hematology)	cg14973204	-0.178	0.0258	1.04e-11	chr12	133052753			
Monocytes (hematology)	cg05575921	-0.38	0.0363	2.94e-24	chr5	373378	AHRR	Body	Smoking (Bojesen et al., Thorax 2017; Chatziioannou et al., Sci Rep 2017)
CD16^{bi} monocytes	cg23213217	2.015	0.2896	7.16e-12	chr1	224370155	DEGS1	TSS1500	Monocyte count (Houseman et al., BMC Bioinformatics 2012)
CD16^{bi} monocytes	cg23417673	-6.397	0.9542	3.81e-11	chr16	85096433	KIAA0513	TSS1500	
CD16^{bi} monocytes	cg05575921	0.576	0.0937	1.21e-09	chr5	373378	AHRR	Body	Smoking (Bojesen et al., Thorax 2017; Chatziioannou et al., Sci Rep 2017)
CD8a$^+$ NK cells	cg03196485	-2.217	0.2581	4.58e-17	chr2	87021117	CD8A	5'UTR	
ILC	cg23617037	0.62	0.0808	4.68e-14	chr6	149804659	ZC3H12D	5'UTR	
ILC	cg03636183	0.23	0.038	2.10e-09	chr19	17000585	F2RL3	Body	Smoking (Dogan et al., Am J Med Genet 2017)
cDC1	cg22697239	-4.095	0.2892	5.49e-41	chr11	44626708	CD82	Body	
CCR6$^+$ cells	cg09222732	-15.187	1.473	1.71e-23	chr6	466893			
CXCR3$^+$ CCR6$^+$ cells	cg22858308	-6.945	0.7813	3.93e-18	chr6	143095613	HIVEP2	Body	
CD4$^-$ CD8b$^-$ T cells	cg00994629	12.089	1.3009	1.43e-19	chr14	22694547			Gestational age (Bohlin et al., Genome Biol 2016)
$\gamma\delta$ TCR$^+$ cells	cg00994629	12.982	1.4714	6.82e-18	chr14	22694547			Gestational age (Bohlin et al., Genome Biol 2016)
$\gamma\delta$ TCR$^+$ cells in lymphocytes	cg00994629	48.705	5.1129	1.91e-20	chr14	22694547			Gestational age (Bohlin et al., Genome Biol 2016)
CD4:CD8 ratio	cg03855955	-7.399	1.1036	3.88e-11	chr12	6900351	CD4	5'UTR	
CD8b$^+$ T cells	cg04329870	-12.479	2.0443	1.61e-09	chr2	87048747	CD8B	Body	
Live CD8$^+$ T cells	cg01029623	15.712	2.3006	1.93e-11	chr12	122016779	KDM2B	Body	CD8$^+$ T cells (Kennedy et al., Front Immunol 2016)
Live CD8$^+$ T cells	cg04329870	-17.201	2.6626	2.03e-10	chr2	87048747	CD8B	Body	
CD8$^+$ CD45$^+$ CD27^{int} cells	cg08641278	8.138	0.72	1.37e-27	chr10	73848764	SPOCK2	1stExon	
CD8$^+$ naive T cells	cg17458390	12.345	1.8609	6.07e-11	chr10	63752709	ARID5B	Body	Atherosclerosis (Liu et al, Nat Commun 2017)
HLA-DR$^+$ of CD8$^+$ CM T cells	cg25242306	67.027	9.8538	2.02e-11	chr13	74667131	KLF12	5'UTR	
HLA-DR$^+$ of CD8$^+$ EMRA T cells	cg02097498	-52.524	8.79	3.45e-09	chr16	10965851			
CD4$^+$ EM T cells	cg26144437	5.433	0.9428	1.18e-08	chr1	145474469	ANKRD34A	Body	Abestos exposure in lung cancer (Kettunen et al., Int J Cancer 2017)
CD4$^+$ EMRA T cells	cg09841874	-9.742	1.3872	4.65e-12	chr20	46251037	NCOA3	Body	
HLA-DR$^+$ of CD4$^+$ EM T cells	cg10921592	-11.554	1.3589	9.01e-17	chr6	33039414	HLA-DPA1	Body	
HLA-DR$^+$ of CD4$^+$ EMRA T cells	cg08151292	-233.484	29.3931	6.65e-15	chr20	3758189	SPEF1	3'UTR	
CD4$^+$ CD8a$^+$ T cells	cg19660239	17.474	1.3247	4.25e-36	chr19	53400545			
CD4$^+$ CD8b$^+$ T cells	cg24148817	-0.676	0.096	4.14e-12	chr6	37461033	C6orf129	Body	
CD4$^+$ CD8b$^+$ T cells	cg11679455	-1.213	0.1805	3.41e-11	chr10	8100761	GATA3	Body	Smoking (Joehanes et al., Circ Cardiovasc Genet 2016)
Live CD4$^+$ CD8b$^+$ T cells	cg11679455	-1.81	0.2267	6.20e-15	chr10	8100761	GATA3	Body	Smoking (Joehanes et al., Circ Cardiovasc Genet 2016)
Naive T$_{reg}$	cg26836479	1.753	0.2441	1.89e-12	chr19	42706353	DEDD2	Body	Gestational age (Bohlin et al., Genome Biol 2016)
Naive T$_{reg}$	cg14395620	-1.454	0.2062	4.46e-12	chr4	40285234			
Naive T$_{reg}$	cg03354487	0.73	0.125	8.11e-09	chr6	20039211			
T$_{reg}$	cg02255107	-3.489	0.4399	9.17e-15	chr3	16347334	OXNAD1	3'UTR	
T$_{reg}$	cg26714968	2.213	0.3012	5.95e-13	chr2	234267824	DGKD	Body	
T$_{reg}$	cg13788583	3.999	0.6658	3.13e-09	chr20	8132217	PLCB1	Body	
Activated T$_{reg}$	cg24683414	2.12	0.2272	1.54e-19	chr1	92952581	GFI1	TSS1500	

Table 5. Continued

Trait	Methylation probe	Coefficient	Standard error	P value	Chr.	Position	Closest gene	Genic region	Published associations
MAIT cells in lymphocytes	cg20732539	19.44	2.02	8.32e-21	chr16	27416077	IL21R	5'UTR	IL21R regulates MAIT cell numbers (Wilson et al., J Exp Med 2015)
MAIT cells in lymphocytes	cg10827488	-18.236	2.5987	4.87e-12	chr11	113953838	ZBTB16	Body	Multiple sclerosis (Souren et al., bioRxiv) ; NK and NKT cell differentiation (Schlums et al., Immunity 2015; Mao et al., PNAS 2016)
MAIT cells in lymphocytes	cg09088625	-34.172	6.0198	1.93e-08	chr3	46246578	CCR1	5'UTR	Monocyte/DC differentiation (Rodríguez-Ubreva et al., Cell Reports 2017)
$CD8^+$ MAIT cells	cg04116545	8.27	1.4394	1.31e-08	chr6	125684679			
Naive B in all B cells	cg06800849	80.426	13.519	4.13e-09	chr16	89180587	ACSF3	Body	Rheumatoid arthritis in B cells (Julià et al., Hum Mol Genet 2017)
Marginal zone B in B cells	cg13651690	-236.308	41.3711	1.61e-08	chr14	106320748			
Plasmocytes in B cells	cg25780496	-12.362	2.0867	4.75e-09	chr15	101137253	LINS	5'UTR	
Transitional B in B cells	cg25385366	-12.187	2.104	1.02e-08	chr21	43809360	TMPRSS3	Body	

time. The estimated regularization paths for the different binary traits are shown in Figure 3. which indicate that near optimal prediction can be achieved with less than 50 DNAm probes. The optimal classification rate for CMV serostatus is *CR*=87% (95% CI: [81%, 94%]) using 256 predictors, while for smoking, *CR*=89% (CI: [82%, 94%]) using 193 predictors.

We also select a robust minimal set of predictors using stability selection for the binary phenotypes. Models are selected and fitted using the same training set of 866 samples as for continuous traits, and then evaluated on the 96 hold-out samples. The prediction accuracy of the models is shown in Table 2. Interestingly, the stability selected model for CMV performs slightly better than the elastic net model, using only 13 probes, while the selected model for smoking performs notably worse. This indicates that the relationship between DNAm and smoking is less sparse than that for DNAm and CMV serostatus. Methylome-wide significant probes selected for smoking are well known DNAm sites predictive of cigarette consumption (Table 6). We find that HBV, *T. gondii* and HVS1 infections associate with DNAm sites close to *EVOLV2* and *KLF14* genes, known to be strongly associated with age. This suggests no effects of these infections on DNAm besides that of age, with which they are themselves strongly correlated [Scepanovic et al., 2018]. More interestingly, DNAm associated with *H. pylori* seropositivity is found within the poliovirus receptor-like 3 gene (P=4.6x10^{-12}), an intestinal epithelium receptor for bacterial toxins [LaFrance et al., 2015], suggesting a role of this protein in *H. pylori* infection.

3. Discussion

Our study reports novel, accurate models to predict blood cell composition from whole blood DNAm profiles. Models were built using a unique dataset that comprises both the quantification of 70 blood cell proportions by standardized flow cytometry [Patin et al., 2018] and blood methylomes established with the MethylationEPIC array [Ait Kaci Azzou et al., n.d.], assessed in 962 healthy donors of western European ancestry. Predictive models are built using the elastic net method [Zou and Hastie, 2005], a regularized linear regression model that has been recently

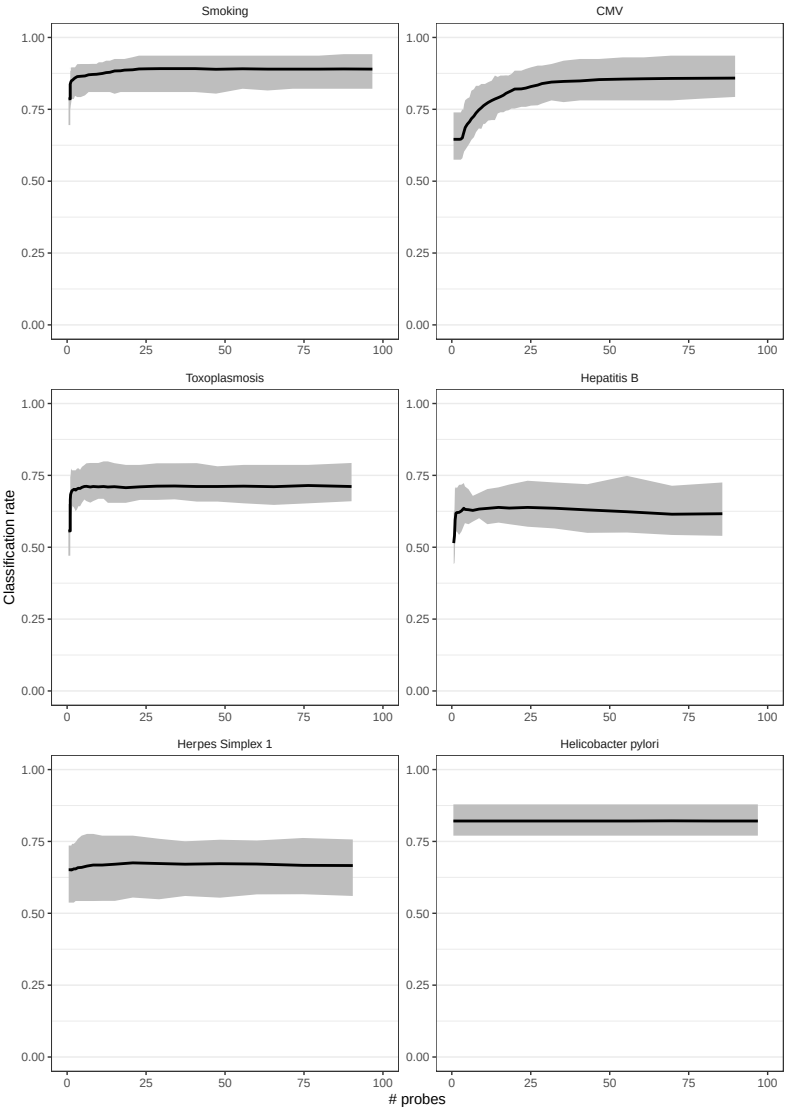

Figure 3. Classification rate as a function of the number of predictors included in logistic regression elastic net models predicting binary traits. The regularization parameter α is here set to 0.95. The 6 binary traits best predicted by the models are shown. Confidence bands are estimated non-parametrically using the 20 samples from the distribution of predictive accuracy given by our cross-validation scheme detailed in Algorithm 4

Table 6. Methylome-wide significant ($P < 3 \times 10^{-8}$) DNAm probes selected by stability selection for all binary traits. A logistic regression model with all predictors chosen by stability selection was fitted for each trait. The predictors who had p values smaller than 3×10^{-8} are included in the table.

Trait	Methylation probe	Coefficient	Standard error	P value	Chr.	Position	Closest gene	Genic region	Published association
Hepatitis B	cg08097417	-17.65	2.000	1.10e-18	chr7	130419133	KLF14	TSS1500	Age (Li et al., Sci Rep 2017)
Toxoplasmosis	cg16867657	11.96	0.888	2.32e-41	chr6	11044877	ELOVL2	TSS1500	Age (Li et al., Sci Rep 2017)
Smoking	cg05575921	-15.98	2.080	1.59e-14	chr5	373378	AHRR	Body	Smoking (Bojesen et al., Thorax 2017; Chatziioannou et al., Sci Rep 2017)
Smoking	cg21566642	-9.62	1.626	3.33e-09	chr2	233284661			Smoking (Joehanes et al., Circ Cardiovasc Genet 2016; Chatziioannou et al., Sci Rep 2017)
Herpes Simplex 1	cg16867657	8.25	0.804	1.14e-24	chr6	11044877	ELOVL2	TSS1500	Age (Li et al., Sci Rep 2017)
Helicobacter pylori	cg21306573	-10.48	1.516	4.65e-12	chr3	110788276	PVRL3-AS1	Body	PVRL3 is a receptor for *Clostridium difficile* toxins (LaFrance et al., PNAS 2015)
Measles	cg09472506	-18.27	1.966	1.45e-20	chr11	100803740	ARHGAP42	Body	ARHGAP42 is associated with risk for hypertension (Bai et al., J Clin Invest 2017)

used to predict age from MethylationEPIC array data [Zhang et al., 2018]. The prediction accuracy, measured as the correlation between predicted and observed out-of-sample values and the MAE, is improved for our models, compared to the widely-used Houseman model, based on either the standard or improved IDOL reference libraries [Houseman et al., 2012; Koestler et al., 2016]. We are also able to accurately predict 35 subset frequencies, in contrast to the six that are currently possible to estimate by the Houseman model using either reference panel. These results suggest that our models should better prevent false positives in EWAS due to cellular heterogeneity, relative to existing gold-standard methods. Nevertheless, it must be noted that we assessed prediction accuracy based on cellular fractions estimated with the same flow cytometry technique, panel design and standardization steps as those used for the training dataset, which may disfavor the other methods trained on other types of cell enumeration techniques.

We also show that it is possible to find predictive models of immune cell proportions that are comparable in terms of accuracy to elastic net models, and to the Houseman models with either reference library, using considerably fewer predictors. This is done by employing the stability selection technique [Shah and Samworth, 2013; Meinshausen and Bühlmann, 2010]. Because of their much smaller size, such models can more robustly, flexibly and cost-effectively predict blood cell composition, age, and smoking consumption than previous models.

Thanks to the exhaustive immunophenotyping performed in our training dataset, we can extend the number of blood cell subsets that can be accurately predicted from blood DNAm data. Notably, our models can accurately predict the blood frequencies of MAIT cells, eosinophils, basophils and T_{reg} cells (R>0.6; Table 1). Importantly, all these leukocyte subsets have previously been reported to vary with various disease conditions, and are thus expected to confound interpretation of EWAS. For instance, circulating levels of MAIT cells are known to be strongly altered during infection [Le Bourhis et al., 2010] and in systemic lupus erythematosus and rheumatoid arthritis patients [Cho et al., 2014]. Eosinophil numbers change with exposure to allergens and in asthmatic patients [Kita, 2011]. Similarly, T_{reg} populations and subpopulations show altered frequencies in several autoimmune and allergic diseases

[Dominguez-Villar and Hafler, 2018]. Therefore, adjusting for these newly-predicted cell populations may improve correction for cellular heterogeneity in epigenomic studies of immune-related disorders. More generally, we envisage that prediction models of blood cellular composition could also be employed to better understand disease pathophysiology *per se*. While EWAS assume that disease-associated DNAm sites affect the transcriptional reprogramming of already differentiated cells, there is increasing evidence that diseases can also be caused by stable alterations of cellular repertoires, a phenomenon recently referred to as polycreodism [Lappalainen and Greally, 2017]. We suggest that model-based estimation of blood cell composition in large longitudinal cohorts, for which methylomes but no flow cytometric measurements exist, will represent a powerful new approach to evaluate whether perturbations in cell proportions can predict disease outcome.

4. Methods

4.1 DNA methylation data.

The *Milieu Intérieur* cohort includes 1,000 healthy donors who were recruited by BioTrial (Rennes, France) and were stratified by gender (*i.e.*, 500 women and 500 men) and age (*i.e.*, 200 individuals from each decade of life, between 20 and 70 years of age). Donors were selected based on stringent inclusion and exclusion criteria, detailed elsewhere [Thomas et al., 2015]. DNAm data was retrieved for all donors from a previous study [Ait Kaci Azzou et al., n.d.], where detailed methods are provided. In brief, the DNA methylome was profiled with the Infinium MethylationEPIC BeadChip on whole blood-derived samples. Raw fluorescence intensities of 866,895 methylation sites across the human genome were processed with the *R* (version 3.5) *Bioconductor* package *minfi*. Values were corrected for probe color bias and differences in type-I and type-II probe distributions, using the single sample NOOB (ssNOOB) method implemented in minfi. Because we wanted to use the methylation data primarily for prediction, which can easily be evaluated on out-of-sample observations and in validation cohorts, we wanted to exclude as few probes as possible. Therefore, we did not exclude probes from the X and Y chromosomes. We did neither exclude possibly cross-reactive probes. From the 866,895 initial probes, we only excluded probes that had a *detection* $P \geq 0.01$ for *more* than 3 samples. A total of 858,923 probes were kept for the analyses. We suppose in this study that DNAm levels are linearly related to cell proportions. We therefore use β methylation values instead of m values.

4.2 Flow cytometry data.

Flow cytometry data was retrieved for all *Milieu Intérieur* donors from a previous study [Patin et al., 2018], where detailed methods are provided. Briefly, whole blood samples were collected from the 1,000 healthy, fasting donors on Li-heparin. Sample staining was performed within 6h of blood draw.Ten 8-color flow cytometry panels

were developed. The acquisition of cells was performed using two MACSQuant analyzers, which were calibrated using MacsQuant calibration beads. Flow cytometry data were generated using MACSQuantify™ software. Among the 313 exported immunophenotypes, we only kept 70 cell proportions and 2 ratios as candidate measures for prediction.

4.3 Houseman model using standard and IDOL reference libraries.

We used the implementation of the Houseman model in the *EstimateCellCounts2* function of the *Bioconductor R* package *FlowSorted.Blood.EPIC* to predict immune cell proportions for all our 962 samples with both the default and IDOL reference panels.

4.4 Statistical modeling

We suppose that there are DNAm CpG sites in the genome of a cell that mark a particular cellular lineage, in the sense that the methylation state of these sites are specific to the cells belonging to that lineage. Therefore, we expect the state of methylation at a number of CpG sites to mark the identity of a particular blood cell. In whole blood, the percentage of cells that are methylated at such DNAm sites should be linearly related to the proportion of the cell in the blood. We further suppose that it is primarily such DNAm sites, and sites related to them, that are predictive of blood cell proportions. We therefore use a linear model to predict blood cell proportions from DNA methylation levels in whole blood. Let $S^P = \{x_p\}_{p=1}^P, P = 858923$ denote our observations of the percentage of methylation at all measured DNAm sites. Let $C \ll P$ be the number of sites that are related to a differentiation event that offers information on the identity of a particular blood cell. This could be a primary event that directly determines cell identity, or it could be an event that gives information on the identity of the cell because of the correlation structure with other cells or genetic and environmental factors. We expect that only few sites correspond to primary events and we further expect the average methylation at such sites to be highly predictive of the immune cell proportion whose lineage it marks. We expect more events that offer correlational information on the proportion of immune cells. Typically, such sites are distributed according to a long tail of decreasing predictive power. Let $D \ll C$ be the number of sites that correspond to primary differentiation events for a particular blood cell. To summarize: for a particular blood cell, we are targeting two sets of probes, S^C and S^D such that

$$S^D = \{x_p\}_{p=1}^D \subset S^C = \{x_p\}_{p=1}^C \subset \{x_p\}_{p=1}^P,$$

and we suppose a predominantly linear relationship between these variables and the cell proportion. We are therefore looking for sparse linear models, where the coefficients of the predictors in $S^P \setminus S^C$ is set to zero. We employ two different strategies to target the predictors in S_C and S_D. Let $n = 962$ be our sample size.

We expect $D \ll n$, but do not necessarily suppose that $C < n$. For S^C we therefore need to *select* predictors, but the linear regression equation system could still be overdetermined, so we also need to *regularize* the coefficients of the fitted linear model. To do this, we employ elastic net regularization [Zou and Hastie, 2005]. In the case of S^D, we only want to *select* predictors and then fit an unbiased least squares regression model. We achieve this by using the stability selection technique [Shah and Samworth, 2013; Meinshausen and Bühlmann, 2010].

Elastic net regression. Let now $X \in \mathbb{R}^{962\times 858923}$ be the matrix corresponding to S^P, with the methylation percentages as columns. Furthermore, let $y \in \mathbb{R}^{962}$ be observations of a cell proportion. The elastic net cost function combines the least squares term with two regularization terms on the magnitude of the coefficients for the columns in X

$$\hat{\beta} = \arg\min_{\beta} \Big\{ \|y - X\beta\|_2^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \Big\}, \ \alpha \in [0,1]. \tag{5.1}$$

The parameter α chooses between a pure Euclidian norm squared penalty, $\|\beta\|_2^2$, corresponding to ridge regression at $\alpha = 0$ and a pure $\mathbb{L}_1$ norm, $\|\beta\|_1$, penalty corresponding to the LASSO [Tibshirani, 1996] penalty at $\alpha = 1$. If $\alpha \neq 0$ then the estimator in (5.1) will do a selection: coefficients that do not rise above a noise floor will be put to exactly zero. The pure LASSO penalty has a *saturation* property: it cannot select more predictors than the number of samples [Hastie et al., 2015]. Note that for the pure LASSO penalty, all coefficients will be zero if

$$\lambda > \max\left(X^T y\right). \tag{5.2}$$

To target X^C, we suppose that an α between zero and one will be optimal. To find this parameter, we employ our own cross-validation scheme, detailed in Algorithm 4. We fit the optimization problem (5.1) by the *glmnet* package in *R*.

Stability selected linear regression. Elastic net regression with regularization parameters tuned by cross-validation will typically include predictors of weak predictive power as well as some false positives [Su et al., 2017]. To target S_D, we therefore use a more stringent selection scheme. As mentioned above, we suppose that $D \ll n$. Therefore, we are now only aiming to select predictors to use in a linear model; we do not want to regularize the parameters. Define the *support* S of a linear model by

$$S(\beta) = \{p : \beta_p \neq 0\}. \tag{5.3}$$

Algorithm 4 Cross-validation for elastic net linear regression
Stated here using the correlation between out-of-sample predictions and observed values as performance estimate. The case for the MAE is analogous. Given observed responses $y \in \mathbb{R}^n$ and predictors $X \in \mathbb{R}^{n \times P}$, our cross-validation scheme conceptually goes as follows

for $r = \{1, 2\}$ **do**
    Divide data 10 equally sized blocks y_k and X_k. Denote data that is not in the kth block with y_{-k} and X_{-k}
    for $\alpha \in \{0.05, 0.1, 0.5, 0.95, 1\}$ **do**
        Compute $\lambda_{max} = \frac{1}{\alpha} \max\left(X^T y\right)$
        for $k \in \{1, \ldots, 10\}$ **do**
            Let l contain 200 values logarithmically from $10^{-4}\lambda_{max}$ to λ_{max}
            for $\lambda \in l$ **do**
                Solve (5.1) for X_{-k} and y_{-k}
                Find prediction: $\hat{y}_k = X_k \hat{\beta}$
                Store $\text{corr}(y_k, \hat{y}_k)$ in $\varepsilon(\alpha, \lambda, k, r)$
            end for
        end for
    end for
end for
The optimal model is found by

$$\hat{\alpha}, \hat{\lambda} = \arg\min_{\alpha,\lambda} \frac{1}{20} \sum_{k,r} \varepsilon(\alpha, \lambda, k, r)$$

First we introduce a weak support estimator. This estimator uses the cost function in (5.1) with α fixed at 0.8, while keeping λ large enough so that it never includes more than q variables. Given this constraint, the support is then estimated to be the included variables. To be more precise, introduce the family of support estimators

$$\begin{aligned} \hat{\beta} = \arg\min_{\beta} \Bigg\{ & \|y - X\beta\|_2^2 + \\ & \lambda \left(0.2\|\beta\|_2^2 + 0.8\|\beta\|_1 \right) \Bigg\}, \\ \hat{S}(\lambda) = & \left\{ p : \hat{\beta}_p \neq 0 \right\}. \end{aligned} \tag{5.4}$$

We then use the support estimator $\hat{S}_q = \hat{S}(\lambda^*)$, where λ^* is such that

$$\lambda^* = \min \left\{ \lambda : |\hat{S}(\lambda)| < q \right\}.$$

To find S^D, we wrap this weak support estimator in a subsampling scheme known as stability selection. The full scheme is outlined in Algorithm 5. Let X_{ss} be the columns of X corresponding to predictors selected by the stability selection scheme. The coefficient estimates for the final linear regression model of the immune cell proportion with measurements in y is then

$$\hat{\beta} = \left(X_{ss}^T X_{ss} \right)^{-1} X_{ss}^T y. \tag{5.5}$$

We use the implementation of stability selection in the *stabs R* package [Hofner et al., 2015].

Algorithm 5 Stability Selection

Given observed response $y \in \mathbb{R}^n$ and predictors $X \in \mathbb{R}^{n \times p}$ the stability selection scheme goes as follows. Let $q = 50$.

1: Subsample rows of $(\, y \; X \,)$ in M pairs, where each pair contains half of the rows, giving a total of $2M$ subsets $B_m,\ m = 1, \ldots, 2M$
2: For all subsets, B_m estimate support $\hat{S}_q(B_m)$
3: For all probes x_p, estimate $\hat{\mathbb{P}}_p = \frac{1}{2M} \sum_{m=1}^{2M} \mathbb{1}_{p \in \hat{S}_q(B_m)}$
4: Include x_p as a predictor if $\hat{\mathbb{P}}_p$ is above a certain threshold
5: The threshold is chosen such that, under stringent assumptions, the maximum number of expected false positives is less than 2, see [Shah and Samworth, 2013] for more information.

Other traits. The models above were developed primarily for immune cell proportions, but we use them also for the other traits. We suppose that most of the predictive power of whole blood DNA methylation for any trait comes from its intimate link with immune cell proportions. Therefore, we anticipate that prediction models of a form suitable for immune cell frequencies should work well also for traits related to them.

For binary traits, code the classes as either 0 or 1. The procedure we use for binary traits follows the algorithms above verbatim, except that the least squares term $\|y - X\beta\|_2^2$ in the cost function in (5.1) is replaced by the negative log-likelihood of the binomial distribution given a logit link function

$$-\frac{1}{n} \sum_{i=1}^{n} \left(y_i \left(\mu + \beta^T x_i \right) - \log \left(1 + e^{\mu + \beta^T x_i} \right) \right). \tag{5.6}$$

Logistic regression with elastic net regularization is implemented in *glmnet*. For stability selection, we use the *stabs R* package with a custom built selection function based on *glmnet*.

References

Ait Kaci Azzou, S., E. Patin, A. Urrutia, H. Quach, J. Bergstedt, et al. (n.d.). "Limited impact of environmental exposures on the human blood methylome in adulthood". *in preparation* ().

Álvarez-Errico, D., R. Vento-Tormo, M. Sieweke, and E. Ballestar (2015). "Epigenetic control of myeloid cell differentiation, identity and function". *Nature Reviews Immunology* **15**:1, p. 7.

Ariga, M., B. Neitzert, S. Nakae, G. Mottin, C. Bertrand, et al. (2004). "Nonredundant function of phosphodiesterases 4D and 4B in neutrophil recruitment to the site of inflammation". *The Journal of Immunology* **173**:12, pp. 7531–7538.

Bojesen, S. E., N. Timpson, C. Relton, G. D. Smith, and B. G. Nordestgaard (2017). "AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality". *Thorax* **72**:7, pp. 646–653.

Chatziioannou, A., P. Georgiadis, D. G. Hebels, I. Liampa, I. Valavanis, et al. (2017). "Blood-based omic profiling supports female susceptibility to tobacco smoke-induced cardiovascular diseases". *Scientific Reports* **7**, p. 42870.

Cho, Y.-N., S.-J. Kee, T.-J. Kim, H. M. Jin, M.-J. Kim, et al. (2014). "Mucosal-associated invariant T cell deficiency in systemic lupus erythematosus". *The Journal of Immunology*, p. 1302701.

Deaton, A. M., S. Webb, A. R. Kerr, R. S. Illingworth, J. Guy, et al. (2011). "Cell type-specific DNA methylation at intragenic CpG islands in the immune system". *Genome research*, pp. 1074–1086.

Dominguez-Villar, M. and D. A. Hafler (2018). "Regulatory T cells in autoimmune disease". *Nature Immunology*, p. 1.

Feil, R. and M. F. Fraga (2012). "Epigenetics and the environment: emerging patterns and implications". *Nature Reviews Genetics* **13**:2, p. 97.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Hofner, B., L. Boccuto, and M. Göker (2015). "Controlling false discoveries in high-dimensional situations: boosting with stability selection". *BMC bioinformatics* **16**:1, p. 144.

Horvath, S. (2013). "DNA methylation age of human tissues and cell types". *Genome biology* **14**:10, p. 3156.

Houseman, E. A., W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, et al. (2012). "DNA methylation arrays as surrogate measures of cell mixture distribution". *BMC Bioinformatics* **13**:1, p. 86.

Jirtle, R. L. and M. K. Skinner (2007). "Environmental epigenomics and disease susceptibility". *Nature Reviews Genetics* **8**:4, p. 253.

Joehanes, R., A. C. Just, R. E. Marioni, L. C. Pilling, L. M. Reynolds, et al. (2016). "Epigenetic signatures of cigarette smoking". *Circulation: Genomic and Precision Medicine* **9**:5, pp. 436–447.

Julià, A., D. Absher, M. López-Lasanta, N. Palau, A. Pluma, et al. (2017). "Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells". *Human Molecular Genetics* **26**:14, pp. 2803–2811.

Kita, H. (2011). "Eosinophils: multifaceted biological properties and roles in health and disease". *Immunological Reviews* **242**:1, pp. 161–177.

Koestler, D. C., M. J. Jones, J. Usset, B. C. Christensen, R. A. Butler, et al. (2016). "Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL)". *BMC Bioinformatics* **17**:1, p. 120.

LaFrance, M. E., M. A. Farrow, R. Chandrasekaran, J. Sheng, D. H. Rubin, et al. (2015). "Identification of an epithelial cell receptor responsible for Clostridium difficile TcdB-induced cytotoxicity". *Proceedings of the National Academy of Sciences*, p. 201500791.

Lappalainen, T. and J. M. Greally (2017). "Associating cellular epigenetic models with human phenotypes". *Nature Reviews Genetics* **18**:7, p. 441.

Le Bourhis, L., E. Martin, I. Péguillet, A. Guihot, N. Froux, et al. (2010). "Antimicrobial activity of mucosal-associated invariant T cells". *Nature Immunology* **11**:8, p. 701.

Liu, Y., M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, et al. (2013). "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis". *Nature biotechnology* **31**:2, p. 142.

McCartney, D. L., A. J. Stevenson, S. J. Ritchie, R. M. Walker, Q. Zhang, et al. (2018). "Epigenetic prediction of complex traits and death". *bioRxiv*.

Meinshausen, N. and P. Bühlmann (2010). "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**:4, pp. 417–473.

Newman, A. M., C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, et al. (2015). "Robust enumeration of cell subsets from tissue expression profiles". *Nature methods* **12**:5, pp. 453–457.

Patin, E., M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, et al. (2018). "Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors". *Nature Immunology* **19**:3, pp. 302–314.

Rakyan, V. K., T. A. Down, D. J. Balding, and S. Beck (2011). "Epigenome-wide association studies for common human diseases". *Nature Reviews Genetics* **12**:8, p. 529.

Scepanovic, P., C. Alanio, C. Hammer, F. Hodel, J. Bergstedt, et al. (2018). "Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines". *Genome Medicine* **10**:1, p. 59.

Shah, R. D. and R. J. Samworth (2013). "Variable selection with error control: another look at stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**:1, pp. 55–80.

Shen-Orr, S. S. and R. Gaujoux (2013). "Computational deconvolution: extracting cell type-specific information from heterogeneous samples". *Current opinion in immunology* **25**:5, pp. 571–578.

Su, W., M. Bogdan, E. Candes, et al. (2017). "False discoveries occur early on the lasso path". *The Annals of Statistics* **45**:5, pp. 2133–2150.

Teschendorff, A. E. and C. L. Relton (2018). "Statistical and integrative system-level analysis of DNA methylation data". *Nature Reviews Genetics* **19**:3, p. 129.

Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, et al. (2015). "The milieu intérieur study – an integrative approach for study of human immunological variance". *Clinical Immunology* **157**:2, pp. 277–293.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Wilson, R. P., M. L. Ives, G. Rao, A. Lau, K. Payne, et al. (2015). "STAT3 is a critical cell-intrinsic regulator of human unconventional T cell numbers and function". *Journal of Experimental Medicine* **212**:6, pp. 855–864.

Zhang, Q., C. Vallerga, R. Walker, T. Lin, A. Henders, et al. (2018). "Improved prediction of chronological age from DNA methylation limits it as a biomarker of ageing". *bioRxiv*.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**:2, pp. 301–320.

Acknowledgements

This work benefited from support of the French governments Program *Investissement dAvenir*, managed by the Agence Nationale de la Recherche (ANR, reference 10-LABX-69-01). J.B. is a member of the LCCC Linnaeus Center and the ELLIIT Excellence Center at Lund University and is supported by the ELLIIT Excellence Center.