Genetic determinants of healthy human immune variance: from humoral response to microbiome diversity

THÈSE N^{O} 8955 (2018)

PRÉSENTÉE LE 14 DÉCEMBRE 2018

À LA FACULTÉ DES SCIENCES DE LA VIE

GROUPE FELLAY

PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Petar SCEPANOVIC

acceptée sur proposition du jury:

Prof. B. E. Ferreira De Sousa Correia, président du jury
Prof. J. Fellay, Prof. B. Deplancke, directeurs de thèse
Prof. D. Gfeller, rapporteur
Prof. A. Rausell, rapporteur
Prof. F. Naef, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

Acknowledgements

My utmost gratitude goes to Prof. Jacques Fellay for giving me the opportunity to conduct research in his lab. I would also like to thank him for the freedom he has given me in pursuing my ideas, for being highly encouraging and never putting pressure (even when the thesis writing seemed to be too slow).

I have to thank Jacques, also, for putting together an amazing group of people in his lab. It is because of him that the work atmosphere on everyday basis resembles to the one of a family. I'm extremely lucky to have had the chance to work alongside Chris, Thorball, Nimisha, Alessandro, Samira, Olivier, Flavia, Istvan, Thomas, Paul, Dylan and Sina. It felt so comforting knowing that I can come to you in times I didn't understand R error messages (which was every time I've used it) and that I can rely on you for anything outside of the lab as well. I would also like to thank Marisa for the administrative help.

I would like to acknowledge all the members of *Milieu Intérieur* in Institut Pasteur. In particular, I would like to thank Cécile for the enthusiasm she has put into our collaboration.

EPFL is such a pleasant place for work because of the people that make it. Across SV, GHI, EDBB and Deplancke lab I would like to thank everyone that made my environment all these four years so enjoyable. In particular, would like to thank Nikola, Marjan and Aleksandra for their constant friendship and Özlem for being there with me in these last moments of my PhD.

Finally, I owe all of this to my family. To my parents and my brother. Thank you for your unconditional support and love.

Abstract

Identifying the drivers of the observed interindividual variability of the human immune system is crucial to our understanding of infectious and immune-mediated diseases. The contribution of genetic and non-genetic factors to immunological differences between humans remains largely undefined. The *Milieu Intérieur* Consortium has established a 1000-person healthy population-based cohort (evenly stratified by sex and age), which represents an unparalleled opportunity for assessing the determinants of human immunologic variance.

In this thesis, three population-based studies are presented, all benefiting from the samples and data collected by investigators of the *Milieu Intérieur* Consortium. Human genome-wide genotyping data, more than 100 environmental, lifestyle and physiological variables, and their combination have been tested for their impact on multiple immune phenotypes. Firstly, we identified the respective contributions of age, sex, and genetics to humoral responses to vaccination and persistent viral infection. We observed that specific variants in the human leukocyte antigen (HLA) region are the strongest genetic determinant of antibody response to common antigens. In the second study, investigation of 166 immuno-phenotypes revealed 15 genetic loci associated with variation of immune cell parameters, mainly of innate immune cells. We attributed an important role to genetic variation in the major histocompatibility complex (MHC) region for these phenotypes and narrowed the signals to probable causal associations in HLA genes. In the third work, forces shaping the gut microbiome composition were investigated. We found a strong influence of several non-genetic factors on overall microbiome diversity and on the abundance of specific bacterial species. We showed as well that genetic factors only play a minor role in gut microbiome composition.

Together, these studies quantified the effects of demographic, environment and genetics on the interindividual variability of phenotypes central to the human immune system. Furthermore, they constitute a valuable resource for further explorations of the impact of immune diversity on the individual risk of infections or of immune diseases.

Keywords

Immunity, Infection, Vaccination, GWAS, Serology, Human genomics, Immunoglobulins, Microbiome, Demographics, Environment.

Résumé

L'identification des facteurs responsables de la diversité du système immunitaire humain représente une étape cruciale dans notre compréhension des maladies infectieuses et immunitaires. La contribution de multiples facteurs génétiques et non génétiques aux différences immunologiques reste encore largement indéterminée. Le Consortium *Milieu Intérieur* a mis en place une cohorte de 1000 individus en bonne santé (stratifiée de façon égale selon le sexe et l'âge), créant ainsi une opportunité unique d'évaluer les déterminants de la variance immunologique humaine.

Dans cette thèse sont présentées trois études basées sur les données et échantillons de la cohorte populationnelle recrutée par le Consortium *Milieu Intérieur*. Les données génotypiques de de tous les participants à l'étude ainsi que plus de 100 variables environnementales et physiologiques ont été analysées individuellement et de manière combinée afin de déterminer leur impact sur de nombreux phénotypes immunitaires. Premièrement, nous avons identifié les contributions respectives de l'âge, du sexe et de la génétique sur la réponse humorale aux vaccins et aux infections virales persistantes. Nous avons observé que des variants dans la région de l'antigène d'histocompatibilité humain sont le principal déterminant génétique de la réponse anticorps aux antigènes les plus courants. Deuxièmement, en analysant 166 immuno-phénotypes, nous avons identifié 15 loci génétiques associés à des variations de propriétés des cellules immunitaires, principalement du système immunitaire inné. De nombreux signaux significatifs ont été détectés sur le segment du chromosome 6 où se trouve le complexe majeur d'histocompatibilité ; c'est pourquoi nous avons procédé à une analyse détaillée de la variation génétique de cette région. Nous avons aussi observé que le tabac, l'âge, le sexe et l'infection latente par le cytomégalovirus sont les principaux facteurs non-génétiques affectant les cellules immunitaires. Troisièmement, nous avons examiné les forces influençant la composition du microbiome intestinal. Nous avons observé une forte influence de différents facteurs non-génétiques sur la diversité globale du microbiome, ainsi que sur l'abondance de certaines espèces bactériennes. Nous avons montré également que les facteurs génétiques ne jouent qu'un rôle mineur dans la composition du microbiome intestinal humain.

Cet ensemble d'études a permis de quantifier les effets démographiques, environnementaux et génétiques sur la variabilité interindividuelle de phénotypes immuns. De plus, elles représentent une ressource précieuse pour explorer de manière plus détaillée l'impact de la diversité immunitaire sur les risques d'infections ou de maladies immunitaires.

Mots-clés

Immunité, Infection, vaccination, GWAS, Sérologie, Génomique humaine, Immunoglobulines, Microbiome, Démographie, Environnement.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 The human immune system

The human immune system consists in a complex network of organs, tissues, cells and molecules with a variety of functions that are essential to the maintenance of a healthy state. It ensures beneficial cohabitation with the microbiome and prevents infection. Additionally, it influences much more than host defense against external threats: it notably affects host metabolism and aging, has crucial roles in the early detection and elimination of neoplasms and is able to inflict damage in the context of autoimmune and autoinflammatory diseases [1]. As a result, immune disorders are often associated with increased susceptibility to infection, inflammation, autoimmunity, or even development of cancer [2].

The main components of the immune system are the various types of immune cells. Some are tissue specific, while many circulate in the blood or in the lymphatic system, ready to access injured tissues when recruited. Host immune responses are classically divided into two type of responses: innate and adaptive. Innate responses are particularly important early in the course of an infection as they react rapidly and non-specifically upon encountering a pathogen. Adaptive immune responses are slower to develop but are specific and build up immunological memory [3].

A large number of germline-encoded pattern recognition receptors, expressed on the surface and in intracellular compartments of many cell types, is dedicated to recognizing abnormal signals produced by invading pathogens or damaged cells. One of the best-characterized group of pattern recognition receptors is the Toll-like receptor family (TLRs), which identifies the broad category of infection (e.g. extracellular bacterium versus intracellular DNA virus) by recognizing a few distinctive, highly conserved biochemical structures present in many microorganisms, such as bacterial lipopolysaccharide (LPS) or viral double-stranded RNA (dsRNA) [4].

An early step of the immune response is the activation of the responder cells - macrophages and other cells producing cytokines and chemokines. These, in turn, recruit and activate other immune cells. This cascade of events triggers the adaptive arm of the immune response, composed of multiple subsets of T cells and B cells. Adaptive immunity is characterized by a high degree of specificity against antigens, which is determined by the T and B cell receptor repertoires and by the ability of these cells to generate memory. The cytokine production, triggered by pattern recognition receptors in responder cells, also contribute to a more efficient adaptive immune response, notably by inducing the differentiation of the appropriate type of T cells. This indicates the existence of a tight link between the two arms of the immune system [4, 5].

Constantly confronted with a wide range of stimuli, the immune system works in a highly dynamic fashion. Still, its main features are supposed to remain functionally stable for long periods of time in healthy adults. It has been shown that immune cell frequencies and serum protein levels remain very stable in blood samples taken weeks, months or even years apart [6]. This suggests that each individual has a baseline state of immune system composition, in which cells and proteins are well regulated, and the balances between these are optimal for the healthy condition [7].

The relative stability of immune parameters over time enables investigations of the underlying factors shaping an individual's immune system.

1.2 Variation of the human immune system

Humans react differently to identical immune challenges [8], strongly suggesting a high degree of variation in the composition and regulation of the immune system.

The considerable clinical variability in infectious disease outcomes observed between individuals and between populations was initially attributed to pathogen characteristics, including variable degree of exposure and of pathogenicity. As exposure to a microbial agent is obviously required for infection and disease to occur, infectious diseases were often regarded as examples of purely environmental diseases. That view changed when Charles Nicolle demonstrated that the same pathogen could cause both asymptomatic and symptomatic infections [9].

Immune variation manifests both at the cellular and intracellular levels. Differences can be observed in the relative frequency of different leukocyte populations, variation in the transcriptional and protein profiles within them, as well as in the functional capacity and polarization of effector cells in response to immunological challenges. This variability results markedly different susceptibility to different diseases [10].

The interindividual diversity of the immune system is an important mechanism for limiting the impact specific pathogens can have on morbidity and mortality of a given population. To counterbalance the clear advantage of microorganisms in evolutionary speed, multicellular hosts have developed specialized cells and pathways (our immune system) that ensure pathogen control. The maintenance of immune variability at the population level is therefore a crucial aspect of its function [8].

The fact that healthy individuals display a large degree of variation in specific immune system components offers multiple avenues for studies into the mechanisms that provide robustness and redundancy to the immune system [11].

1.3 Understanding immune variation

The plasticity and resilience of the immune system allows for a large spectrum of functions that are helpful in our dealing with the environment. In parallel, it also increases the probability that, in some circumstances, a subset of individuals will experience a pathologic cascade of immunological events [12].

Individual heterogeneity in immune responses can thus have important medical consequences, such as immunodeficiencies, or meaningful differences in response to anti-infectious therapy or to vaccine administration. It can also influence individual susceptibility to autoimmunity, allergy, cancer, and complex diseases with an inflammatory component like cardiovascular diseases and neurodegenerative disorders [13].

Given the breadth of the effects of immune responses on human pathologies, there is an urgent need to understand immune variability in humans. To develop patient-based individualized treatments, it is necessary to understand these processes at an individual level within a population [14].

Personalized medicine can be defined as the management of a patient's disease or disease predisposition supported by large-scale molecular analyses to achieve optimal medical outcome for the individual [15]. The potential advantages of this approach, both for patients and doctors, include more accurate diagnosis and treatments, safer drug prescription, better disease prevention and even reduction in healthcare costs [16].

A better comprehension of human immune variation in health and disease could pave the way toward concrete applications of personalized medicine. Understanding at which point the extent of immunologic variation becomes pathogenic will be critical in developing primary prevention strategies for the diseases mentioned above. It will help to identify appropriate targets for drug development and identify subsets of individuals that require more active monitoring or need intervention [17].

1.4 Factors influencing immune variance

It is widely accepted that a broad range of factors contributes to human immune system variation. Understanding when and how such influences shape the human immune system is key for defining metrics of immunological health and understanding the risk of immune-mediated and infectious disease [18].

Variability in immune responses can be due to (i) biological parameters (e.g. age, sex), (ii) genetic variants (e.g. single nucleotide polymorphisms, gene methylation marks), (iii) environmental factors (e.g. microbiome, latent or chronic infections, diet, smoking) (Figure 1.1) [19].

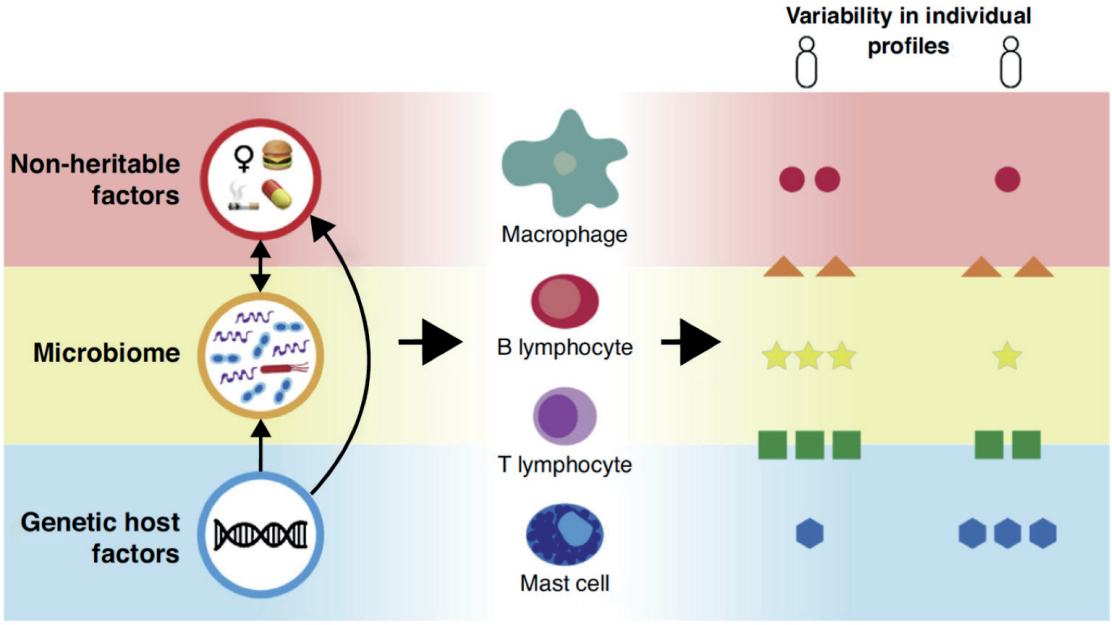

Figure 1.1. The complex interplay between, and the effects of different factors on immune cells and immune system of individuals. Adapted from [18].

It is well known that some immune responses are directly dependent on germline DNA variants (e.g., HLA alleles). On the other hand, the mature T and B cell repertoires are shaped by environmental exposures, which determine which subsets of naïve cells undergo maturation and expansion (e.g., pathogens and commensals, allergens, etc.) [4]. Thus, teasing apart the contributions of environment and genetics to the immune system is a particularly complex task. Yet, a combined understanding of both the heritable and the non-heritable influences on immunity is necessary to fully understand inter-individual variation and its consequences on immunological health and disease [20].

While many studies have examined the impact of individual components on specific immune responses, fewer integrative analyses have been performed. As a consequence, the respective contributions of heritable and non-heritable factors to the composition and function of specific immune responses are often unclear. Our understanding of the genetic, evolutionary, and environmental factors that impact this inter-individual and inter-population immune response heterogeneity is still in its early days [21].

1.5 Non-genetic factors influencing immune variation

Non-heritable influences are typically interpreted as environmental influences, such as infections and vaccinations [22].

The presence of the microorganism is required, but not sufficient, for the development of an infectious disease and Louis Pasteur himself — the father of the microbial theory — emphasized the importance of non-microbial factors in susceptibility to infection [23].

Humans live in a complex environment, and although the influences of pathogens in shaping our immune systems are the most well-described factors, many other non-microbial environmental factors and intrinsic host factors exert a strong influence on human immunity.

It was estimated that intrinsic factors explain 5% of the immunological variation, while up to 60-80% can be explained by environmental factors [24].

Unlike genetic and other intrinsic factors, which are in principle identifiable and measurable in a given individual, dissecting out the meaningful environmental factors represents a more daunting task. This is due to the fact that a limitless list of putative environmental factors could be considered. Collecting data for any of these factors can be especially difficult, and the timeframe over which environmental factors may be acting is mostly unknown, making it hard to dissect their individual contribution to the immune variation.

1.6 Microbiome

Humans are host to trillions of microbes found across multiple body sites such as gut, skin, vagina, oral and nasal cavities. The number of bacterial cells is estimated to be similar to the number of human cells, but together they contain a much vaster genetic repertoire. Until recently, research of the human microbiome used culture-based approaches. It is only through the development of culture-independent techniques over the past decade that it has become possible to examine the full diversity and functionality of the microbiome [25].

The gut microbiome is an essential factor in the education of the immune system. The types of bacteria that colonize the infant gut play a central role in immune education and in the establishment of immune tolerance toward gut commensal microorganisms [26].

The healthy immune system recognizes and interacts with the microbiome with exquisite specificity. Studies have suggested that gut microbes promote the development of intestinal T_h17 cells, which play an important role in infectious disease but are also implicated in the pathogenesis of autoimmune and inflammatory diseases. In fact, aberrant host-microbial interactions at the gut-immune interface are associated with a range of diseases, such as inflammatory bowel disease, rheumatoid arthritis and cancer [27].

The gut microbiome can influence a range of diseases via numerous potential mechanisms. However, the host also influences the gut microbiome: its composition depends on a range of host, environmental and lifestyle factors [28].

The interplay of microbiome, non-heritable and genetic factors, is a major avenue of research and it must be considered when investigating immune variation.

1.7 Genetic factors influencing immune variation

Human genetic variation has been implicated as a central factor in defining individual susceptibility to many diseases though genetic epidemiological studies, complex segregation studies and studies using concordance rates between monozygotic and dizygotic twins. In the field of infectious diseases, severe infections occurring during childhood often represent a monogenic immunodeficiency, while severe symptoms occurring later in life, mostly during secondary infections, might result from more complex genetic predispositions. In fact, increasing layers of evidence show that the human genetic control of common infectious diseases lies in a continuous spectrum from Mendelian susceptibility (rare mutations with

strong effect) to complex polygenic predisposition (polymorphisms with modest effect), with several intermediate situations (for example, a major gene) [29].

Linkage studies of infections and vaccine responses have revealed important roles for several immune-related genes. In particular, human leukocyte antigen (HLA) genes were shown to have a pervasive influence on the activities of the immune system [30]. The HLA class I and II loci are the most polymorphic human genes. They are located in the major histocompatibility complex (MHC), which is an extremely gene-dense region with long-range linkage disequilibrium and hundreds of immunologically active genes. As HLA molecules present microbial antigenic peptides to T cells, the polymorphism of HLA genes is required to mount an efficient immune response against an extremely diverse world of pathogens [31].

HLA polymorphisms have been correlated with low antibody response to measles vaccination in individuals homozygous at one or more HLA loci or carrying specific alleles such as HLA-DRB1*03 and DQA1*0201. Other HLA alleles (HLA-B*44, DRB1*01, DRB1*08, and DQA1*0104) have been observed to associate with very high seropositivity rates following measles-mumps-rubella (MMR) vaccination. It is now clear that HLA genes are indispensable for host defense at the individual and population levels, and that multiple infectious agents are responsible for the selection of their extremely high degree of polymorphism [17].

Major studies into the genetic basis of the variation in the cellular and molecular composition of the human immune system showed that genetic factors (e.g., common and rare variants, variations in copy number) account for 20–40% of total immunological variance [10]. Yet, it's only recently that we have started being able to query human genetic variation at large scale.

The completion of the Human Genome Project at the turn of the new century marked an inflexion point in the history of biology: the post-genomic era began [32, 33]. With the sequence in hand, the next step was to identify the extent of human genetic variation. The HapMap project [34, 35, 36] and 1000 Genomes Project [37] were large-scale studies that followed and successfully managed to identify and catalogue a vast number of nucleotides that vary among individuals. All these efforts, accompanied by the advent of high-throughput genotyping and sequencing technologies, established a new paradigm in the way of approaching biological questions [38].

For the first time, researchers are able to scan the entire genome in a hypothesis-free, agnostic approach. Multiple genome-wide association studies (GWAS) have now identified genetic loci associated with multiple immune system phenotypes, such as immune cell frequencies or the concentration of specific cytokines. The identified alleles revealed many new associated loci, thereby providing leads for more in-depth genetic, etiological and mechanistical studies, as GWAS primarily make use of markers that often represent causal variation indirectly [39].

1.8 Population studies of immunity

The large-scale study of variation among healthy subjects is essential to understand the immune changes that are truly disease-related (and not due, for example, to population stratification or batch effects). With the collection of large research cohorts and the advent of systems analyses in the field of human immunity, it becomes possible to reliably assess

human immune system variation at the population level, to consider interdependencies between immune system components, and to analyze their interindividual variability in health and disease. Combining those advances with the recent progress in DNA analysis allows researchers to correlate complex measures of immune parameters to genetic variation [7].

Several cohort studies have been established to address these questions in recent years. These include the *Milieu intérieur* consortium [40], the Human Functional Genomics Project [41], the Human Immunology Project Consortium (HIPC) [42] and the 10K Immunomes [43].

In summary, several population-based initiatives that focus on healthy individuals and integrate genetic and immunological phenotyping have begun to define the factors behind variable immune responses. Together, these efforts should provide a better definition of immune response variability and a clearer understanding of the key factors that drive it [17].

1.9 Overview of the Thesis

Throughout the thesis, I benefited from the extraordinary resources produced by collaborators of the *Milieu intérieur* Consortium. My particular focus was on the discovery of the genetic underpinnings of immune variance and microbiome diversity in healthy individuals.

Chapter 2 describes GWAS that aimed at identifying the genetic factors responsible for differences in seroprevalence and antibody levels against a range of persistent or recurrent pathogens. We used serological data obtained from the 1'000 healthy individuals about 12 different pathogens and linked them with genetic variation.

In **Chapter 3**, many GWAS of human genetic determinants of immune parameters and cell counts are presented. A total of 166 immune-related phenotypes were tested in the 1,000 healthy people, which identified previously unknown association signals with multiple human genetic variants. Many of these were located in the MHC region and were relevant for innate immune cells. We focused on these signals and fine-mapped the observed associations.

In **Chapter 4**, the main focus is on the observed diversity of the gut microbiome in the MI cohort. After thoroughly investigating the influence of the environment on the gut microbiome diversity, we assessed its association with genetics, while controlling for the identified cofounding effects.

Finally, **Chapter 5** provides a summary with future directions and outlooks.

1.10 References

1. Netea MG, Joosten LA, Li Y, Kumar V, Oosting M, Smeekens S, Jaeger M, Ter Horst R, Schirmer M, Vlamakis H, Notebaart R, Pavelka N, Aguirre-Gamboa RR, Swertz MA, Tunjungputri RN, van de Heijden W, Franzosa EA, Ng A, Graham D, Lassen K, Schraa K, Netea-Maier R, Smit J, de Mast Q, van de Veerdonk F, Kullberg BJ, Tack C, van de Munckhof I, Rutten J, van der Graaf J, Franke L, Hofker M, Jonkers 2, Platteel M, Maatman A, Fu J, Zhernakova A, van der Meer JW, Dinarello CA, van der Ven A, Huttenhouwer C, Koenen H, Joosten I, Xavier RJ, Wijmenga C. Understanding human immune function using the resources from the Human Functional Genomics Project. Nat Med. 2016; 22:831-3.

2. Duffy D, Rouilly V, Libri V, Hasan M, Beitz B, David M, Urrutia A, Bisiaux A, Labrie ST, Dubois A, Boneca IG, Delval C, Thomas S, Rogge L, Schmolz M, Quintana-Murci L, Albert ML; Milieu Intérieur Consortium. Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. Immunity. 2014; 40:436-50.

3. Carmichael A, Wills M. The immunology of infection. Medicine. 2013. 41:611-618.

4. Netea MG, Wijmenga C, O'Neill LA. Genetic variation in Toll-like receptors and disease susceptibility. Nat Immunol. 2012; 13:535-42.

5. Cho J. The heritable immune system. Nat Biotechnol. 2015; 33:608-9.

6. Tsang JS, Schwartzberg PL, Kotliarov Y, Biancotto A, Xie Z, Germain RN, Wang E, Olnes MJ, Narayanan M, Golding H, Moir S, Dickler HB, Perl S, Cheung F; Baylor HIPC Center; CHI Consortium. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. Cell. 2014; 157:499-513.

7. Brodin P, Davis MM. Human immune system variation. Nat Rev Immunol. 2017; 17:21-29.

8. Sanz J, Randolph HE, Barreiro LB. Genetic and evolutionary determinants of human population variation in immune responses. Curr Opin Genet Dev. 2018; 53:28-35.

9. Quintana-Murci L, Alcaïs A, Abel L, Casanova JL. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. Nat Immunol. 2007; 8:1165-71.

10. Liston A, Carr EJ, Linterman MA. Shaping Variation in the Human Immune System. Trends Immunol. 2016; 37:637-646.

11. Kumar V, Wijmenga C, Xavier RJ. Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. Curr Opin Immunol. 2014; 31:51-7.

12. Kaiser J. NIH opens precision medicine study to nation. Science. 2015; 349:1433.

13. Ye CJ, Feng T, Kwon HK, Raj T, Wilson MT, Asinovski N, McCabe C, Lee MH, Frohlich I, Paik HI, Zaitlen N, Hacohen N, Stranger B, De Jager P, Mathis D, Regev A, Benoist C. Intersection of population variation and autoimmunity genetics in human T cell activation. Science. 2014; 345:1254665.

14. De Jager PL, Hacohen N, Mathis D, Regev A, Stranger BE, Benoist C. ImmVar project: Insights and design considerations for future studies of "healthy" immune variation. Semin Immunol. 2015; 27:51-7.

15. Kroemer HK, Meyer zu Schwabedissen HE. A piece in the puzzle of personalized medicine. Clin Pharmacol Ther. 2010; 87:19-20.

16. Netea MG, Joosten LA, Latz E, Mills KH, Natoli G, Stunnenberg HG, O'Neill LA, Xavier RJ. Trained immunity: A program of innate immune memory in health and disease. Science. 2016; 352:aaf1098.

17. Boyd SD, Jackson KJL. Predicting vaccine responsiveness. Cell Host Microbe. 2015; 17:301-307.

18. Schirmer M, Kumar V, Netea MG, Xavier RJ. The causes and consequences of variation in human cytokine production in health. Curr Opin Immunol. 2018; 54:50-58.

19. Duffy D. Milieu intérieur: Defining the boundaries of a healthy immune response for improved vaccination strategies. Hum Vaccin Immunother. 2018; 1:1-5.

20. Liston A, Goris A. The origins of diversity in human immunity. Nat Immunol. 2018; 19:209-210.

21. Davis MM, Brodin P. Rebooting Human Immunology. Annu Rev Immunol. 2018; 36:843-864.

22. Carr EJ, Dooley J, Garcia-Perez JE, Lagou V, Lee JC, Wouters C, Meyts I, Goris A, Boeckxstaens G, Linterman MA, Liston A. The cellular composition of the human immune system is shaped by age and cohabitation. Nat Immunol. 2016; 17:461-468.

23. Casanova JL, Abel L. The genetic theory of infectious diseases: a brief history and selected illustrations. Annu Rev Genomics Hum Genet. 2013; 14:215-43.

24. Wu X, Chen H, Xu H. The genomic landscape of human immune-mediated diseases. J Hum Genet. 2015; 60:675-81.

25. Honda K, Littman DR. The microbiome in infectious disease and inflammation. Annu Rev Immunol. 2012; 30:759-95.

26. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, Sogin ML. The Microbiome and Human Biology. Annu Rev Genomics Hum Genet. 2017; 18:65-86.

27. Sekirov I, Russell SL, Antunes LC, Finlay BB. Gut microbiota in health and disease. Physiol Rev. 2010; 90:859-904.

28. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016; 8:51.

29. Alcaïs A, Abel L, Casanova JL. Human genetics of infectious diseases: between proof of principle and paradigm. J Clin Invest. 2009; 119:2506-14.

30. Casanova JL, Abel . The human model: a genetic dissection of immunity to infection in natural conditions. Nat Rev Immunol. 2004; 4:55-66.

31. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 2013; 14:661-73

32. Lander E. S. et al. Initial sequencing and analysis of the human genome. Nature. 2001: 409: 860-921.

33. Venter J. C. et al. The sequence of the human genome. Science. 2001; 291: 1304-1351.

34. The International HapMap Consortium. The haplotype map of the human genome. Nature. 2005; 437: 1299-1320.

35. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449: 851-861.

36. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467: 52-58.

37. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061-1073.

38. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S. Sequencing studies in human genetics: design and interpretation. Nat Rev Genet. 2013; 14:460-70.

39. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53.

40. Thomas S, Rouilly V, Patin E, Alanio C, Dubois A, Delval C, Marquier LG, Fauchoux N, Sayegrih S, Vray M, Duffy D, Quintana-Murci L, Albert ML. Milieu Intérieur Consortium. The *Milieu Intérieur* study—an integrative approach for study of human immunological variance. Clin Immunol. 2015;157:277–93.

41. Pappalardo JL, Hafler DA. The Human Functional Genomics Project: Understanding Generation of Diversity. Cell. 2016; 167:894–896.

42. Brusic V, Gottardo R, Kleinstein SH, Davis MM & HIPC steering committee. Computational resources for high-dimensional immune analysis from the Human Immunology Project Consortium. Nat. Biotechnol. 2014; 32:146–148.

43. Zalocusky KA et al. The 10,000 Immunomes Project: A resource for human immunology. bioRxiv. 2017; doi:10.1101/180489

Chapter 2: Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines

This work has been published in ***Genome Medicine*** (2018; 10:59)

Petar Scepanovic[1,2†], Cécile Alanio[3,4,5†], Christian Hammer[1,2,6], Flavia Hodel[1,2], Jacob Bergstedt[7], Etienne Patin[8,9,10], Christian W. Thorball[1,2], Nimisha Chaturvedi[1,2], Bruno Charbit[4], Laurent Abel[11,12,13], Lluis Quintana-Murci[8,9,10], Darragh Duffy[3,4,5], Matthew L. Albert[6*], Jacques Fellay[1,2,14*] for The *Milieu Intérieur* Consortium.

Author Affiliations: 1 School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; 2 Swiss Institute of Bioinformatics, Lausanne, Switzerland; 3 Immunobiology of Dendritic Cell Unit, Institut Pasteur, Paris, France; 4 Center for Translational Research, Institut Pasteur, Paris, France; 5 Inserm U1223, Institut Pasteur, Paris, France ; 6 Department of Cancer Immunology, Genentech, South San Francisco, CA, USA; 7 Department of Automatic Control, Lund University, Lund, Sweden; 8 Unit of Human Evolutionary Genetics, Department of Genomes and Genetics, Institut Pasteur, Paris, France; 9 Centre National de la Recherche Scientifique, URA 3012, Paris, France; 10 Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris 75015, France; 11 Laboratory of Human Genetics of Infectious Diseases, Necker branch, Inserm U1163, Paris, France; 12 Paris Descartes University, Imagine Institute, Paris, France; 13 St Giles laboratory of Human Genetics of Infectious Diseases, Rockefeller branch, The Rockefeller University, New York, NY, USA; 14 Precision Medicine Unit, Lausanne University Hospital, Lausanne, Switzerland; † These authors contributed equally to the study; * Corresponding Authors.

Contribution to the study: I was involved in the study design and preformed the genetic analyses – ran a genome-wide association study with serology data, investigated specific associations and conducted rare variant burden testing. Together with Cécile Alanio, I wrote the manuscript.

2.1 Abstract

Background**.** Humoral immune responses to infectious agents or vaccination vary substantially among individuals, and many of the factors responsible for this variability remain to be defined. Current evidence suggests that human genetic variation influences (i) serum immunoglobulin levels, (ii) seroconversion rates, and (iii) intensity of antigen-specific immune responses. Here, we evaluate the impact of intrinsic (age and sex), environmental and genetic factors on the variability of humoral response to common pathogens and vaccines.

Methods**.** We characterized the serological response to 15 antigens from common human pathogens or vaccines, in an age- and sex-stratified cohort of 1,000 healthy individuals (*Milieu Intérieur* cohort). Using clinical-grade serological assays, we measured total IgA, IgE, IgG and IgM levels, as well as qualitative (serostatus) and quantitative IgG responses to cytomegalovirus, Epstein-Barr virus, herpes simplex virus 1 & 2, varicella zoster virus, *Helicobacter pylori*, *Toxoplasma gondii*, influenza A virus, measles, mumps, rubella, and hepatitis B virus. Following genome-wide genotyping of single nucleotide polymorphisms and imputation, we examined associations between ~5 million genetic variants and antibody responses using single marker and gene burden tests.

Results**.** We identified age and sex as important determinants of humoral immunity, with older individuals and women having higher rates of seropositivity for most antigens. Genome-wide association studies revealed significant associations between variants in the human leucocyte antigen (HLA) class II region on chromosome 6 and anti-EBV and anti-rubella IgG levels. We used HLA imputation to fine map these associations to amino acid variants in the peptide-binding groove of HLA-DRβ1 and HLA-DPβ1, respectively. We also observed significant associations for total IgA levels with two loci on chromosome 2 and with specific KIR-HLA combinations.

Conclusions. Using extensive serological testing and genome-wide association analyses in a well-characterized cohort of healthy individuals, we demonstrate that age, sex and specific human genetic variants contribute to inter-individual variability in humoral immunity. By highlighting genes and pathways implicated in the normal antibody response to frequently encountered antigens, these findings provide a basis to better understand disease pathogenesis.

2.2 Background

Humans are regularly exposed to infectious agents, including common viruses such as cytomegalovirus (CMV), Epstein-Barr virus (EBV) or herpes simplex virus-1 (HSV-1), that have the ability to persist as latent infections throughout life – with possible reactivation events depending on extrinsic and intrinsic factors [1]. Humans also receive multiple vaccinations, which in many cases are expected to achieve lifelong immunity in the form of neutralizing antibodies. In response to each of these stimulations, the immune system

mounts a humoral response, triggering the production of specific antibodies that play an essential role in limiting infection and providing long-term protection. Although the intensity of the humoral response to a given stimulation has been shown to be highly variable [2, 3, 4], the genetic and non-genetic determinants of this variability are still largely unknown. The identification of such factors may lead to improved vaccination strategies by optimizing vaccine-induced immunoglobulin G (IgG) protection, or to new understanding of autoimmune diseases, where immunoglobulin levels can correlate with disease severity [5].

Several genetic variants have been identified that account for inter-individual differences in susceptibility to pathogens [6, 7, 8, 9], and in infectious [10] or therapeutic [11] phenotypes. By contrast, relatively few studies have investigated the variability of humoral responses in healthy humans [12, 13, 14]. In particular, Hammer C., *et al.* examined the contribution of genetics to variability in human antibody responses to common viral antigens, and fine-mapped variants at the HLA class II locus that associated with IgG responses. To replicate and extend these findings, we measured IgG responses to 15 antigens from common infectious agents or vaccines as well as total IgG, IgM, IgE and IgA levels in 1,000 well-characterized healthy donors. We used an integrative approach to study the impact of age, sex, non-genetic and genetic factors on humoral immunity in healthy humans.

2.3 Methods

2.3.1 Study participants

The *Milieu Intérieur* cohort consists of 1,000 healthy individuals that were recruited by BioTrial (Rennes, France). The cohort is stratified by sex (500 men, 500 women) and age (200 individuals from each decade of life, between 20 and 70 years of age). Donors were selected based on stringent inclusion and exclusion criteria, previously described [15]. Briefly, recruited individuals had no evidence of any severe/chronic/recurrent medical conditions. The main exclusion criteria were: seropositivity for human immunodeficiency virus (HIV) or hepatitis C virus (HCV); ongoing infection with the hepatitis B virus (HBV) – as evidenced by detectable HBs antigen levels; travel to (sub-)tropical countries within the previous 6 months; recent vaccine administration; and alcohol abuse. To avoid the influence of hormonal fluctuations in women during the peri-menopausal phase, only pre- or post-menopausal women were included. To minimize the importance of population substructure on genomic analyses, the study was restricted to self-reported Metropolitan French origin for three generations (*i.e.*, with parents and grandparents born in continental France). Whole blood samples were collected from the 1,000 fasting healthy donors on lithium heparin tubes, from September 2012 to August 2013. The clinical study was approved by the Comité de Protection des Personnes - Ouest 6 on June 13th, 2012, and by the French Agence Nationale de Sécurité du Médicament on June 22nd, 2012. The study is sponsored by 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).

2.3.2 Serologies

Total IgG, IgM, IgE, and IgA levels were measured using clinical grade turbidimetric test on AU 400 Olympus at the BioTrial (Rennes, France). Antigen-specific serological tests were performed using clinical-grade assays measuring IgG levels, according to the manufacturer's instructions. A list and description of the assays is provided in Additional File 1: Table S1. Briefly, anti-HBs and anti-HBc IgGs were measured on the Architect automate (CMIA assay, Abbott). Anti-CMV IgGs were measured by CMIA using the CMV IgG kit from Beckman Coulter on the Unicel Dxl 800 Access automate (Beckman Coulter). Anti-Measles, anti-Mumps and anti-Rubella IgGs were measured using the BioPlex 2200 MMRV IgG kit on the BioPlex 2200 analyzer (Bio-Rad). Anti-*Toxoplasma gondi*, and anti-CMV IgGs were measured using the BioPlex 2200 ToRC IgG kit on the BioPlex 2200 analyzer (Bio-Rad). Anti-HSV1 and anti-HSV2 IgGs were measured using the BioPlex 2200 HSV-1 & HSV-2 IgG kit on the BioPlex 2200 analyzer (Bio-Rad). IgGs against *Helicobacter Pylori* were measured by EIA using the PLATELIA *H. Pylori* IgG kit (BioRad) on the VIDAS automate (Biomérieux). Anti-influenza A IgGs were measured by ELISA using the NovaLisa IgG kit from NovaTec (Biomérieux) that explores responses to grade 2 H3N2 Texas 1/77 strain. In all cases, the criteria for serostatus definition (positive, negative or indeterminate) were established by the manufacturer, and are indicated in Additional File 1: Table S2. Donors with an unclear result were retested, and assigned a negative result if borderline levels were confirmed with repeat testing.

2.3.3 Non-genetic variables

A large number of demographical and clinical variables are available in the Milieu Intérieur cohort as a description of the environment of the healthy donors [15]. These include infection and vaccination history, childhood diseases, health-related habits, and socio-demographical variables. Of these, 53 where chosen for subsequent analysis of their impact on serostatus. This selection is based on the one done in [16], with a few variables added, such as measures of lipids and C-reactive protein (CRP).

2.3.4 Testing of non-genetic variables

Using serostatus variables as the response, and non-genetic variables as treatment variables, we fitted a logistic regression model for each response and treatment variable pair. A total of 14 * 52 = 742 models where therefore fitted. Age and sex where included as controls for all models, except if that variable was the treatment variable. We tested the impact of the clinical and demographical variables using a likelihood ratio test. All 742 tests where considered a multiple testing family with the false discovery rate (FDR) as error rate.

2.3.5 Age and sex testing

To examine the impact of age and sex we performed logistic and linear regression analyses for serostatus and IgG levels, respectively. For logistic regression, we included both scaled linear and quadratic terms for the age variable (model = glm(y~Age+I(Age^2)+Sex, family=binomial)). Scaling was achieved by centering age variable at the mean age. When indicated, we used a second model that includes age, sex as well as an interaction term for age and sex (model = glm(y~Age+Sex+Age*Sex, family=binomial)). All continuous traits (i.e.

quantitative measurements of antibody levels) were log10-transformed in donors assigned as positive using the clinical cutoff suggested by the manufacturer. We used false discovery rate (FDR) correction for the number of serologies tested (associations with P < 0.05 were considered significant).

2.3.6 DNA genotyping

Blood was collected in 5mL sodium EDTA tubes and was kept at room temperature (18–25°) until processing. DNA was extracted from human whole blood and genotyped at 719,665 single nucleotide polymorphisms (SNPs) using the HumanOmniExpress-24 BeadChip (Illumina). The 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 variants using the HumanExome-12 BeadChip. The HumanExome variant call rate was lower than 97% in 11 donors, which were thus removed from this dataset. We filtered out from both datasets genetic variants 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 < 99%, (v) were monomorphic, (vi) were on sex chromosomes, or (vii) diverged significantly from Hardy-Weinberg equilibrium (HWE $P < 10^{-7}$). These quality-control filters yielded a total of 661,332 and 87,960 variants for the HumanOmniExpress and HumanExome BeadChips, respectively. Average concordance rate for the 16,753 SNPs shared between the two genotyping platforms was 99.9925%, and individual concordance rates ranged from 99.8% to 100%.

2.3.7 Genetic relatedness and structure

As detailed elsewhere [16], relatedness was detected using KING [17]. Six pairs of related participants (parent-child, first and second-degree siblings) were detected and one individual from each pair, randomly selected, was removed from the genetic analyses. The genetic structure of the study population was estimated using principal component analysis (PCA), implemented in EIGENSTRAT (v6.1.3) [18]. The PCA plot of the study population is shown in Additional File 2: Figure S1.

2.3.8 Genotype imputation

We used Positional Burrows-Wheeler Transform for genotype imputation, starting with the 661,332 quality-controlled SNPs genotyped on the HumanOmniExpress array. Phasing was performed using EAGLE2 (v2.0.5) [19]. As reference panel, we used the haplotypes from the Haplotype Reference Consortium (release 1.1) [20]. After removing SNPs that had an imputation info score < 0.8 we obtained 22,235,661 variants. We then merged the imputed dataset with 87,960 variants directly genotyped on the HumanExome BeadChips array and removed variants that were monomorphic or diverged significantly from Hardy-Weinberg equilibrium ($P < 10^{-7}$). We obtained a total of 12,058,650 genetic variants to be used in association analyses.

We used SNP2HLA (v1.03) [21] to impute 104 4-digit HLA alleles and 738 amino acid residues (at 315 variable amino acid positions of the HLA class I and II proteins) with a minor allele frequency (MAF) of >1%.

We used KIR*IMP [22] to impute KIR alleles, after haplotype inference on chromosome 19 with SHAPEIT2 (v2.r790) [23]. A total of 19 KIR types were imputed: 17 loci plus two extended haplotype classifications (A vs. B and KIR haplotype). A MAF threshold of 1% was applied, leaving 16 KIR alleles for association analysis.

2.3.9 Genetic association analyses

For single variant association analyses, we only considered SNPs with a MAF of >5% (N=5,699,237). We used PLINK (v1.9) [24] to perform logistic regression for binary phenotypes (serostatus: antibody positive versus negative) and linear regression for continuous traits (log10-transformed quantitative measurements of antibody levels in seropositive donors). The first two principal components of a PCA based on genetic data, age and sex were used as covariates in all tests. In order to correct for baseline difference in IgG production in individuals, total IgG levels were included as covariates when examining associations with antigen-specific antibody levels, total IgM, IgE and IgA levels**.** From a total of 53 additional variables additional co-variates, selected by using elastic net [25] and stability selection [26] as detailed elsewhere [16], were included in some analyses (Additional File 1: Table S3). For all genome-wide association studies, we used a genome-wide significant threshold ($P_{threshold} < 2.6 \times 10^{-9}$) corrected for the number of antigens and immunoglobulin classes tested (N=19). For specific HLA analyses, we used PLINK (v1.07) [27] to perform conditional haplotype-based association tests and multivariate omnibus tests at multi-allelic amino acid positions.

2.3.10 Variant annotation and gene burden testing

We used SnpEff (v4.3g) [28] to annotate all 12,058,650 variants. A total of 84,748 variants were annotated as having (potentially) moderate (e.g. missense variant, inframe deletion, etc.) or high impact (e.g. stop gained, frameshift variant, etc.) and were included in the analysis. We used bedtools v2.26.0 [29] to intersect variant genomic location with gene boundaries, thus obtaining sets of variants per gene. By performing kernel-regression-based association tests with SKAT_CommonRare (testing the combined effect of common and rare variants) and SKATBinary implemented in the SKAT v1.2.1 [30], we tested 16,628 gene sets for association with continuous and binary phenotypes, respectively. By SKAT default parameters, variants with MAF ≤ $\frac{1}{\sqrt{2n}}$ are considered rare, whereas variants with MAF ≥ $\frac{1}{\sqrt{2n}}$ were considered common, where N is the sample size. We used genome-wide Bonferroni correction for multiple testing, accounting for the number of phenotypes tested ($P_{threshold} < 2.6 \times 10^{-9}$).

2.4 Results

2.4.1 Characterization of humoral immune responses in the 1,000 study participants

To characterize the variability in humoral immune responses between healthy individuals, we measured total IgG, IgM, IgA and IgE levels in the plasma of the 1,000 donors of the *Milieu Interieur* (MI) cohort. After log10 transformation, total IgG, IgM, IgA and IgE levels showed normal distributions, with a median ± sd of 1.02 ±0.08 g/l, 0.01 ±0.2 g/l, 0.31 ±0.18 g/l and 1.51 ±0.62 UI/ml, respectively (Additional File 2: Figure S2A).

We then evaluated specific IgG responses to multiple antigens from the following infections and vaccines: (i) 7 common persistent pathogens, including five viruses: CMV, EBV (EA, EBNA, and VCA antigens), herpes simplex virus 1 & 2 (HSV-1 & 2), varicella zoster virus (VZV), one bacterium: *Helicobacter pylori* (H. Pylori), and one parasite: *Toxoplasma gondii* (T. Gondii); (ii) one recurrent virus: influenza A virus (IAV); and (iii) four viruses for which most donors received vaccination: measles, mumps, rubella, and HBV (HBs and HBc antigens). The distributions of $\log_{10}$-transformed antigen-specific IgG levels in the 1,000 donors for the 15 serologies are shown in Additional File 2: Figure S2B. Donors were classified as seropositive or seronegative using the thresholds recommended by the manufacturer (Additional File 1: Table S2).

The vast majority of the 1,000 healthy donors were chronically infected with EBV (seropositivity rates of 96% for EBV VCA, 91% for EBV EBNA and 9% for EBV EA) and VZV (93%). Many also showed high-titer antibodies specific for IAV (77%), HSV-1 (65%), and T. Gondii (56%). By contrast, fewer individuals were seropositive for CMV (35%), HSV-2 (21%), and H. Pylori (18%) (Additional File 2: Figure S3A). The majority of healthy donors carried antibodies against 5 or more persistent/recurrent infections of the 8 infectious agents tested (Additional File 2: Figure S3B). 51% of MI donors were positive for anti-HBs IgG - a large majority of them as a result of vaccination, as only 15 study participants (3% of the anti-HBs positive group) were positive for anti-HBc IgG, indicative of previous HBV infection (spontaneously cured, as all donors were negative for HBs antigen, criteria for inclusion in the study). For rubella, measles, and mumps, seropositivity rates were 94%, 91%, and 89% respectively. For the majority of the donors, this likely reflects vaccination with a trivalent vaccine, which was integrated in 1984 as part of national recommendations in France, but for some – in particular the >40 year-old individuals of the cohort, it may reflect acquired immunity due to natural infection.

2.4.2 Associations of age, sex, and non-genetic variables with serostatus

Subjects included in the *Milieu Intérieur* cohort were surveyed for a large number of variables related to infection and vaccination history, childhood diseases, health-related habits, and socio-demographical variables (http://www.milieuinterieur.fr/en/research-activities/cohort/crf-data). Of these, 53 where chosen for subsequent analysis of their impact on serostatus. This selection is based on the one done in [16], with a few variables added, such as measures of lipids and CRP. Applying a mixed model analysis that controls for potential confounders and batch effects, we found expected associations of HBs seropositivity with previous administration of HBV vaccine, as well as of Influenza seropositivity with previous administration of Flu vaccine. We also found associations of HBs seropositivity with previous administration of Typhoid and Hepatitis A vaccines - which likely reflects co-immunization, as well as with Income, Employment, and Owning a house – which likely reflects confounding epidemiological factors (Additional File 2: Figure S4). Full results of the association of non-genetic variables with serostatus are available in Additional File 1: Table S4.

We observed a significant impact of age on the probability of being seropositive for antigens from persistent or recurrent infectious agents and/or vaccines. For 14 out of the 15 examined serologies, older people (> 45 years old) were more likely to have detectable specific IgG, with

a mean beta estimate of 0.04 for linear associations (Figure 2.1A). Additionally, we found a significant quadratic term for five out of the 15 serologies, highlighting that the rate of change in probability of seropositivity with respect to age is higher for rubella and lower for HSV-1, HP, HBs and EBV EBNA in older people as compared to younger donors (Additional File 2: Figure S5A). We identified four different profiles of age-dependent evolution of seropositivity rates (Figure 2.1B). Profile 1 is typical of childhood-acquired infection, *i.e.* microbes that most donors had encountered by age 20 (EBV, VZV, and influenza). We observed in this case either (i) a limited increase in seropositivity rate after age 20 for EBV; (ii) stability for VZV; or (iii) a small decrease in seropositivity rate with age for IAV (Additional File 2: Figure S5B-F). Profile 2 concerns prevalent infectious agents that are acquired throughout life, with steadily increasing prevalence (observed for CMV, HSV-1, and *T. gondii*). We observed in this case either (i) a linear increase in seropositivity rates over the 5 decades of age for CMV (seropositivity rate: 24% in 20-29 years-old; 44% in 60-69 years-old; slope=0.02) and T. Gondii (seropositivity rate: 21% in 20-29 years-old; 88% in 60-69; slope=0.08); or (ii) a non-linear increase in seropositivity rates for HSV-1, with a steeper slope before age 40 (seropositivity rate: 36% in 20-29 years-old; 85% in 60-69; slope=0.05) (Additional File 2: Figure S5G-I). Profile 3 showed microbial agents with limited seroprevalence - in our cohort, HSV-2, HBV (anti-HBs and anti-HBc positive individuals, indicating prior infection rather than vaccination), and *H. Pylori*. We observed a modest increase of seropositivity rates throughout life, likely reflecting continuous low-grade exposure (Additional File 2: Figure S5J-L). Profile 4 is negatively correlated with increasing age and is unique to HBV anti-HBs serology (Additional File 2: Figure S5M). This reflects the introduction of the HBV vaccine in 1982 and the higher vaccination coverage of younger populations. Profiles for Measles, Mumps and Rubella are provided in Additional File 2: Figure S5N-P.

We also observed a significant association between sex and serostatus for 8 of the 15 antigens, with a mean beta estimate of 0.07 (Figure 2.1C). For six serological phenotypes, women had a higher rate of positivity, IAV being the notable exception. These associations were confirmed when considering “Sharing house with partner”, and “Sharing house with children” as covariates. Full results of associations of age and sex with serostatus are present in Additional File 1: Table S5. Finally, we found a significant interaction of age and sex for odds of being seropositive for EBV EBNA, reflecting a decrease in seropositivity rate in older women (beta -0.0414814; P=0.02, Additional File 2: Figure S5Q).

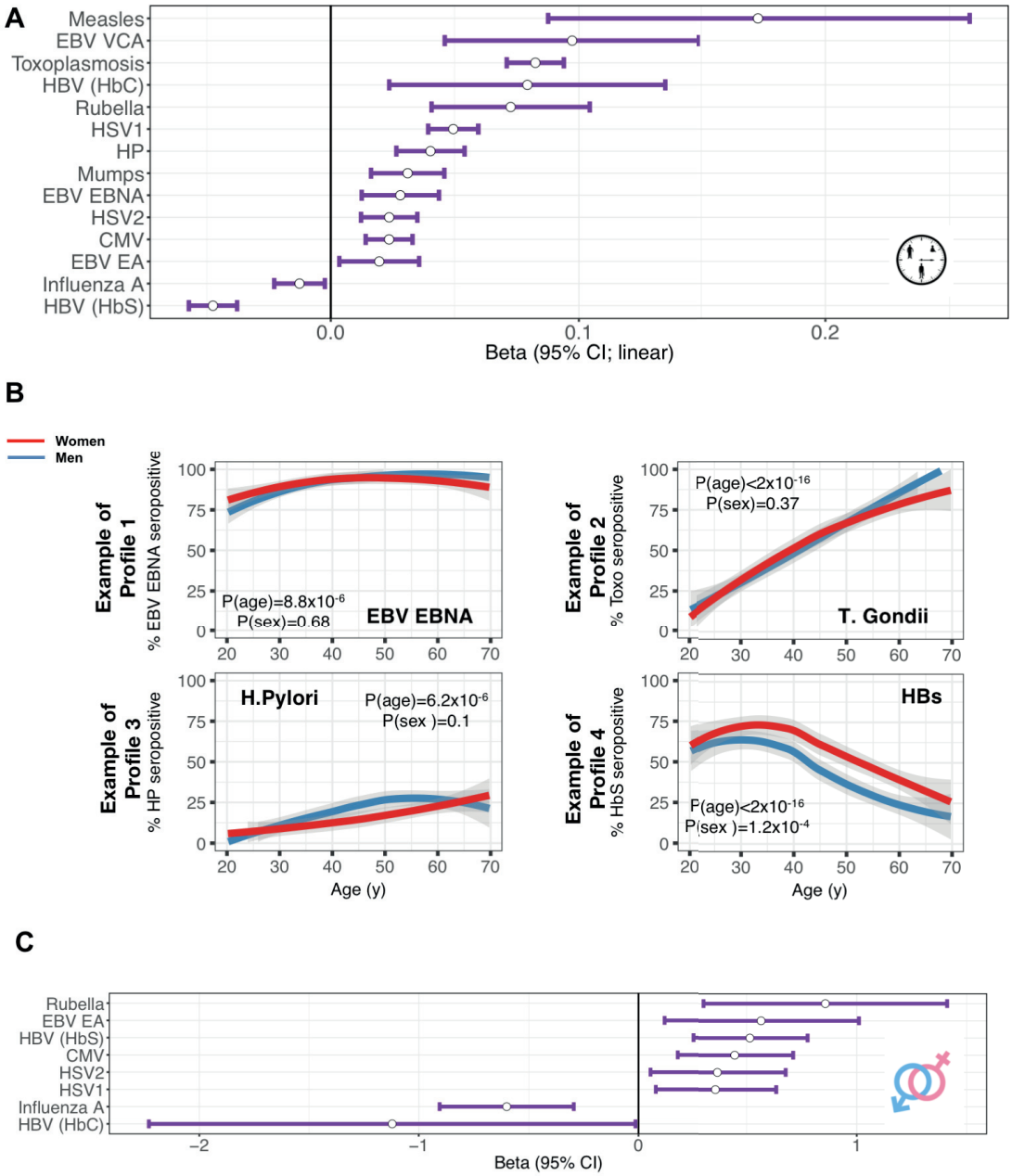

Figure 2.1. Age and sex impact on serostatus. (A) Effect sizes of significant linear associations (adjusted P-values (adj. P<0.05)) between age and serostatus as determined based on clinical-grade serologies in the 1,000 healthy individuals from the Milieu Intérieur cohort. Effect sizes were estimated in a generalized linear mixed model, with serostatus as response variable, and age and sex as treatment variables. This model includes both scaled linear and quadratic terms for the age variable. Scaling was achieved by centering age variable at the mean age. All results from this analysis are provided in Additional File 1: Table S5. Dots represent the mean of the beta. Lines represent the 95% confidence intervals. (B) Odds of being seropositive towards EBV EBNA (Profile 1; upper left), Toxoplasma gondii (Profile 2; upper right), Helicobacter Pylori (Profile 3; bottom left), and HBs antigen of HBV (Profile 4; bottom right), as a function of age in men (blue) and women (red) in the 1,000 healthy donors. Indicated P-values were obtained using a logistic regression with Wald test, with serostatus binary variables (seropositive versus seronegative) as the response, and age and sex as treatments.

Similar plots from all examined serologies are provided in Additional File 2: Figure S5. (C) Effect sizes of significant associations (adjusted P-values (adj. P<0.05) between sex (Men=reference, vs. Women) and serostatus. Effect sizes were estimated in a generalized linear mixed model, with serostatus as response variable, and age and sex as treatment variables. All results from this analysis are provided in Additional File 1: Table S5. Dots represent the mean of the beta. Lines represent the 95% confidence intervals.

2.4.3 Impact of age and sex on total and antigen-specific antibody levels

We further examined the impact of age and sex on the levels of total IgG, IgM, IgA and IgE detected in the serum of the patients, as well as on the levels of antigen-specific IgGs in seropositive individuals. We observed a low impact of age and sex with total immunoglobulin levels (Figure 2.2A). Age also had a strong impact on specific IgG levels in seropositive individuals, affecting 9 out of the 15 examined serologies (Figure 2.2B). Correlations between age and pathogen-specific IgG levels were mostly positive, *i.e.* older donors had more specific IgG than younger donors, as for example in the case of Rubella (Additional File 2: Figure S6A). The notable exception was *T. gondii*, where we observed lower amounts of specific IgG in older individuals (b=-0.013(-0.019, -0.007), $P=3.7x10^{-6}$, Additional File 2: Figure S6B). On the other hand, sex was significantly correlated with IgG levels specific to Mumps and VZV (Figure 2.2C). Full results of associations of age and sex with total immunoglobulin and antigen-specific antibody levels are presented in Additional File 1: Table S5.

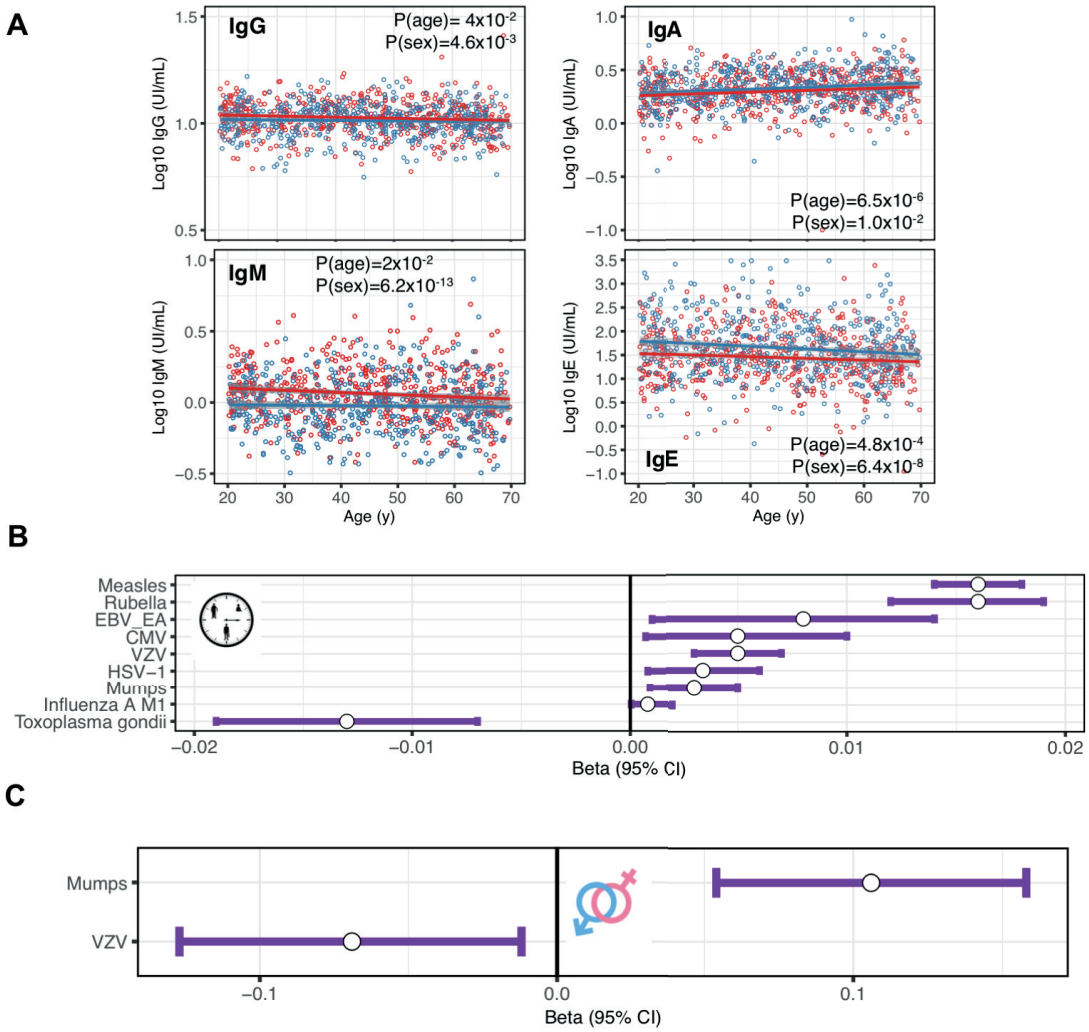

Figure 2.2. Age and sex impact on total and antigen-specific antibody levels. (A) Relationships between Log10-transformed IgG (upper left), IgA (upper right), IgM (bottom left) and IgE (bottom right) levels and age. Regression lines were fitted using linear regression, with Log10-transformed total antibody levels as response variable, and age and sex as treatment variables. Indicated adj. P were obtained using the mixed model, and corrected for multiple testing using the FDR method. (B-C) Effect sizes of significant associations (adjusted P-values (adj. P<0.05) between age (B) and sex (C) on Log10-transformed antigen-specific IgG levels in the 1,000 healthy individuals from the Milieu Intérieur cohort. Because of low number of seropositive donors (n=15), HBc serology was removed from this analysis. Effect sizes were estimated in a linear mixed model, with Log10-transformed antigen-specific IgG levels as response variables, and age and sex as treatment variables. All results from this analysis are provided in Additional File 1: Table S5. Dots represent the mean of the beta. Lines represent the 95% confidence intervals.

2.4.4 Genome-wide association study of serostatus

To test if human genetic factors influence the rate of seroconversion upon exposure, we performed genome-wide association studies. Specifically, we searched for associations

between 5.7 million common polymorphisms (MAF > 5%) and the 15 serostatus in the 1,000 healthy donors. Based on our results regarding age and sex, we included both as covariates in all models. After correcting for the number of antibodies considered, the threshold for genome-wide significance was $P_{threshold} = 2.6 \times 10^{-9}$, for which we did not observe any significant association. In particular, we did not replicate the previously reported associations with *H. Pylori* serostatus on chromosome 1 (rs368433, P = 0.56, OR = 1.08) and 4 (rs10004195, P = 0.83, OD = 0.97) [31]. We verified this result by performing an additional analysis that matched the design of the previous study, i.e. a case-control association study comparing individuals in the upper quartile of the anti-*H. Pylori* antibody distribution to the rest of the study population: no association was found (P=0.42 and P=0.48 for rs368433 and rs10004195, respectively). The quantile-quantile (QQ) plots and lambda values of all genome-wide logistic regressions are available in Additional File 2: Figure S7.

We then focused on the HLA region and confirmed the previously published association of influenza A serostatus with specific amino acid variants of HLA class II molecules [12]. The strongest association in the *MI* cohort was found with residues at position 31 of the HLA-DRβ1 subunit (omnibus P = 0.009, Additional File 1: Table S6). Residues found at that position, isoleucine (P = 0.2, OD (95% CI) = 0.8 (0.56, 1.13)) and phenylalanine (P = 0.2, OR (95% CI) = 0.81 (0.56, 1.13)), are consistent in direction and in almost perfect linkage disequilibrium (LD) with the glutamic acid residue at position 96 in HLA-DRβ1 that was identified in the previous study (Additional File 1: Table S7). As such, our result independently validates the previous observation.

2.4.5 Genome-wide association study of total and antigen-specific antibody levels

To test whether human genetic factors also influence the intensity of antigen-specific immune response, we performed genome-wide association studies of total IgG, IgM, IgA and IgE levels, as well as antigen-specific IgG levels.

We found no SNPs associated with total IgG, IgM, IgE and IgA levels. Additional File 2: Figure S8 shows QQ plots and lambda values of these studies. However, we observed nominal significance and the same direction of the effect for 3 out of 11 loci previously published for total IgA [13, 32, 33, 34, 35], 1 out of 6 loci for total IgG [13, 32, 36] and 4 out of 11 loci for total IgM [13, 37] (Additional File 1: Table S8). Finally, we also report a suggestive association (genome-wide significant, $P < 5.0 \times 10^{-8}$, but not significant when correcting for the number of antibody levels tested in the study) of a SNP rs11186609 on chromosome 10 with total IgA levels ($P = 2.0 \times 10^{-8}$, beta = -0.07 for the C allele). The closest gene for this signal is *SH2D4B*.

We next explored associations between human genetic variants and antigen-specific IgG levels in seropositive donors. Information on possible inflation of false positive rates of these linear regressions are available in Additional File 2: Figure S9. We detected significant associations for anti-EBV (EBNA antigen) and anti-rubella IgGs. Associated variants were in both cases located in the HLA region on chromosome 6. For EBV, the top SNP was rs74951723 ($P = 3 \times 10^{-14}$, beta = 0.29 for the A allele) (Figure 2.3A). For rubella, the top SNP was rs115118356 ($P = 7.7 \times 10^{-10}$, beta = -0.11 for the G allele) (Figure 2.3B). rs115118356 is in LD with rs2064479, which has been previously reported as associated with titers of anti-rubella IgGs ($r^2 = 0.53$ and D' = 0.76) [38].

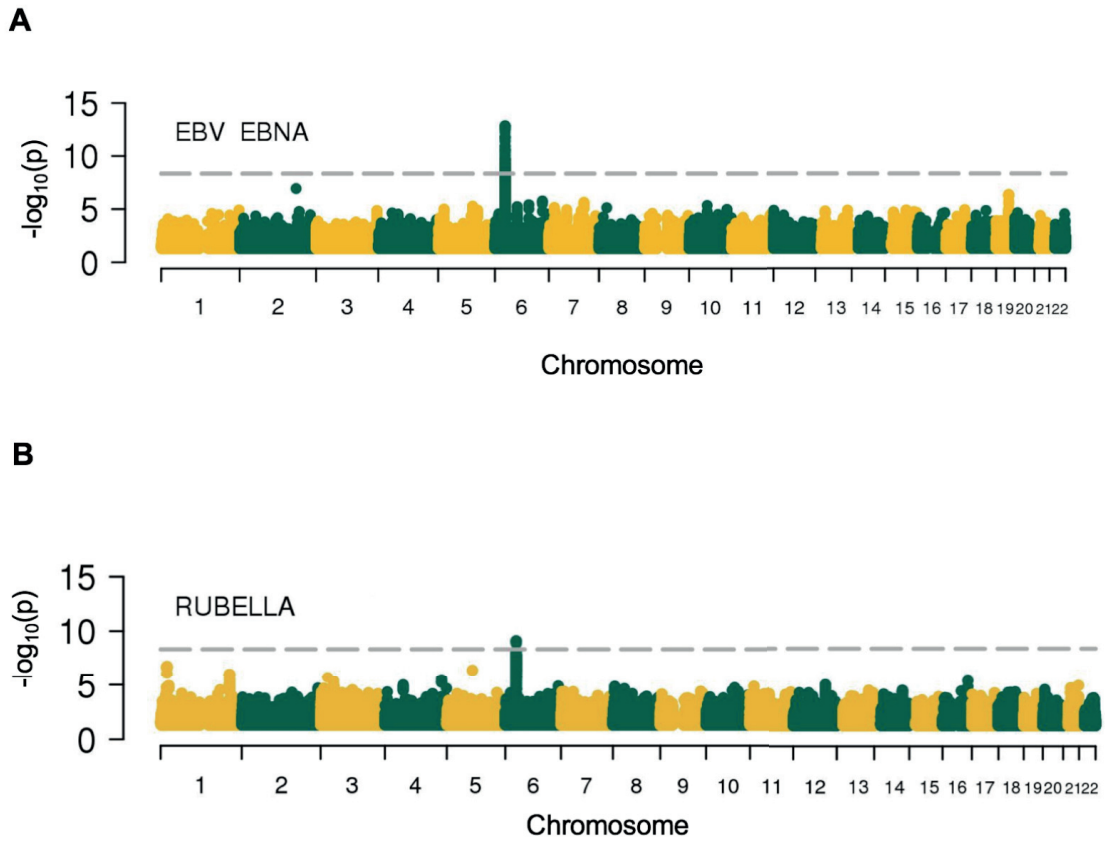

Figure 2.3. Association between host genetic variants and serological phenotypes. Manhattan plots of association results for (A) EBV anti-EBNA IgG, (B) Rubella IgG levels. The dashed horizontal line denotes genome-wide significance ($P = 2.6 \times 10^{-9}$).

To fine map the associations observed in the HLA region, we tested 4-digit HLA alleles and variable amino positions in HLA proteins. At the level of HLA alleles, *HLA-DQB1*03:01* showed the lowest P-value for association with EBV EBNA ($P = 1.3 \times 10^{-7}$), and *HLA-DPB1*03:01* was the top signal for rubella ($P = 3.8 \times 10^{-6}$). At the level of amino acid positions, position 58 of the HLA-DRβ1 protein associated with anti-EBV (EBNA antigen) IgG levels ($P = 2.5 \times 10^{-11}$). This is consistent with results of previous studies linking genetic variations in HLA-DRβ1 with levels of anti-EBV EBNA-specific IgGs [12, 39, 40] (Additional File 1: Table S9). In addition, position 8 of the HLA-DPβ1 protein associated with anti-rubella IgG levels ($P = 1.1 \times 10^{-9}$, Table 2.1). Conditional analyses on these amino-acid positions did not reveal any additional independent signals.

Table 2.1. Associations of EBV EBNA and Rubella antigens with HLA (SNP, allele and amino acid position)

		Phenotype	
		EBV EBNA IgG levels	Rubella IgG levels
SNP	ID (Allele)	rs74951723 (A)	rs115118356 (G)
	P-value	3×10^{-14}	7.68×10^{-10}
	Beta (95% CI)	0.29 (0.21, 0.36)	-0.11 (-0.15, -0.08)
Classical HLA allele	Allele	HLA-DQB1*03:01	HLA-DPB1*03:01
	P-value	1.26×10^{-7}	3.8×10^{-6}
	Beta (95% CI)	0.17 (0.11, 0.23)	-0.12 (-0.18, -0.07)
Amino acid	Protein (position)	HLA-DRβ1 (56)	HLA-DPβ1 (8)
	Omnibus P-value	2.53×10^{-11}	1.12×10^{-9}

2.4.6 KIR associations

To test whether specific KIR genotypes, and their interaction with HLA molecules, are associated with humoral immune responses, we imputed KIR alleles from SNP genotypes using KIR*IMP [22]. First, we searched for potential associations with serostatus or IgG levels for 16 KIR alleles that had a MAF > 1%. We did not find any significant association after Bonferroni correction for multiple testing. Second, we tested specific KIR-HLA combinations. We filtered out rare combinations by removing pairs that were observed less then 4 times in the cohort. After correction for the number of tests performed and phenotypes considered ($P_{threshold} < 5.4 \times 10^{-7}$), we observed significant associations between total IgA levels and the two following HLA-KIR combinations: HLA-B*14:02 / KIR3DL1 and HLA-C*08:02 / KIR2DS4 ($P = 3.9 \times 10^{-9}$ and $P = 4.9 \times 10^{-9}$ respectively, Table 2.2).

Table 2.2. Association testing between KIR-HLA interactions and serology phenotypes

Phenotype	KIR	HLA	Estimate	Std. Error	P-value
IgA levels	KIR3DL1	HLA-B*14:02	0.456	0.077	3.9×10^{-09}
IgA levels	KIR2DS4	HLA-B*14:02	0.454	0.077	4.5×10^{-09}
IgA levels	KIR3DL1	HLA-C*08:02	0.449	0.076	4.9×10^{-09}
IgA levels	KIR2DS4	HLA-C*08:02	0.448	0.076	5.7×10^{-09}

2.4.7 Burden testing for rare variants

Finally, to search for potential associations between the burden of low frequency variants and the serological phenotypes, we conducted a rare variant association study. This analysis only included variants annotated as missense or putative loss-of-function (nonsense, essential splice-site and frame-shift, N=84,748), which we collapsed by gene and tested together using the kernel-regression-based association test SKAT [30]. We restricted our analysis to genes that contained at least 5 variants. Two genes were identified as significantly associated with total IgA levels using this approach: *ACADL* ($P = 3.4 \times 10^{-11}$) and *TMEM131* ($P=7.8 \times 10^{-11}$) (Table 2.3). By contrast, we did not observe any significant associations between rare variant burden and antigen-specific IgG levels or serostatus. All the QQ plots and lambda values of analysis of binary, total Ig levels and pathogen-specific quantitative phenotypes are shown in Additional File 2: Figure S10, S11 and S12.

Table 2.3. Significant associations of rare variants collapsed per gene set with IgA levels.

Phenotype	Chromosome	Gene	P-value	Q	Nº of Rare Markers	Nº of Common Markers
IgA levels	2	ACADL	$3.42x10^{-11}$	18.09	5	2
	2	TMEM131	$7.83x10^{-11}$	17.89	13	2

2.5 Discussion

We performed genome-wide association studies for a number of serological phenotypes in a well-characterized age- and sex-stratified cohort and included a unique examination of genetic variation at HLA and KIR loci, as well as KIR-HLA associations. As such, our study provides a broad resource for exploring the variability in humoral immune responses across different isotypes and different antigens in humans.

Using a fine-mapping approach, we replicated the previously reported associations of variation in the HLA-DRβ1 protein with influenza A serostatus and anti-EBV IgG titers [4, 12], implicating amino acid residues in strong LD with the ones previously reported by Hammer et al. In accordance with the same study, we did not observe any significant association with another measure of EBV serostatus, the presence of anti-EBNA antibodies, suggesting that a larger sample size will be required to uncover potentially associated variants. We replicated an association between HLA class II variation and anti-Rubella IgG titers [38], and further fine-mapped it to position 8 of the HLA-DPβ1 protein. Interestingly, position 8 of HLA-DPβ1, as well as positions 58 and 31 of HLA-DRβ1, are all part of the extracellular domain of the respective proteins. Our findings confirm these proteins as critical elements for the presentation of processed peptide to $CD4^+$ T cells, and as such may reveal important clues in the fine regulation of class II antigen presentation. We also identified specific HLA/KIR combinations, namely HLA-B*14:02/KIR3DL1 and HLA-C*08:02/KIR2DS4, which associate with higher levels of circulating IgA. Combinations of HLA and killer cell immunoglobulin-like receptor (KIR) genes have been associated with diseases as diverse as autoimmunity, viral infections, reproductive failure, and cancer [41]. To date, the molecular basis for these associations are mostly unknown. One could speculate that the association identified between IgA levels and specific KIR-HLA combinations may reflect different levels of tolerance to commensal microbes. However, formal testing of this hypothesis will require additional studies. Also, given the novelty of KIR imputation method and the lack of possibility of benchmarking its reliability in the *MI* cohort, further replication of these results will be needed. Yet these findings support the concept that variations in the sequence of HLA Class II molecules, or specific KIRs/HLA class I interactions play a critical role in shaping humoral immune responses in humans. In particular, our findings confirm that small differences in the capacity of HLA class II molecules to bind specific viral peptides can have a measurable impact on downstream antibody production. As such, our study emphasizes the importance of considering HLA diversity in disease association studies where associations between IgG levels and autoimmune diseases are being explored.

We identified nominal significance for some but not all of the previously reported associations with levels of total IgG, IgM and IgA, as well as a suggestive association of total IgA levels with

an intergenic region on chromosome 10 - closest gene being *SH2D4B.* By collapsing the rare variants present in our dataset into gene sets and testing them for association with the immunoglobulin phenotypes, we identified two additional loci that participate to natural variation in IgA levels. These associations mapped to the genes *ACADL* and *TMEM131*. *ACADL* encodes an enzyme with long-chain acyl-CoA dehydrogenase activity, and polymorphisms have been associated with pulmonary surfactant dysfunction [42]. As the same gene is associated with levels of circulating IgA in our cohort, we speculate that *ACADL* could play a role in regulating the balance between mucosal and circulating IgA. Further studies will be needed to test this hypothesis, as well as the potential impact of our findings in other IgA-related diseases.

We were not able to replicate previous associations of *TLR1* and *FCGR2A* locus with serostatus for *H. Pylori* [31]. We believe this may be a result of (i) different analytical methods; or (ii) notable differences in previous exposure among the different cohorts as illustrated by the different levels of seropositivity - 17% in the *Milieu Interieur* cohort, versus 56% in the previous ones, reducing the likelihood of replication due to decreased statistical power.

In addition to genetics findings, our study re-examined the impact of age and sex, as well as non-genetic variables, on humoral immune responses. Although this question has been previously addressed, our well-stratified cohort brings interesting additional insights. One interesting finding is the high rate of seroconversion for CMV, HSV-1, and T. Gondii during adulthood. In our cohort, the likelihood of being seropositive for one of these infections is comparable at age 20 and 40. This observation raises interesting questions about the factors that could prevent some individuals from becoming seropositive upon late life exposure, considering the high likelihood of being in contact with the pathogens because of their high prevalence in humans (CMV and HSV-1) or because of frequent interactions with an animal reservoir (toxoplasmosis). Second, both age and sex have a strong correlation with serostatus, *i.e.* older and female donors were more likely to be seropositive. Although increased seropositivity with age probably reflects continuous exposure, the sex effect is intriguing. Indeed, our study considered humoral immunity to microbial agents that differ significantly in terms of physiopathology and that do not necessarily have a childhood reservoir. Also, our analysis shows that associations persist after removal of potential confounding factors such as marital status, and/or number of kids. As such, we believe that our results may highlight a general impact of sex on humoral immune response variability, *i.e.* a tendency for women to be more likely to seroconvert after exposure, as compared to men of same age. Gender-specific differences in humoral responses have been previously observed for a large number of viral and bacterial vaccines including influenza, hepatitis A and B, rubella, measles, rabies, yellow fever, meningococcus, pneumococcus, diphtheria, tetanus and Brucella [43, 44]. Along the same line, women often respond to lower vaccine doses then men [43, 45], and higher levels of antibodies have been found in female schoolchildren after rubella and mumps vaccination [46] as well as in adult women after smallpox vaccination [47]. This could be explained, at least partially, by a shift towards Th2 immunity in women as compared to men [48]. Finally, we observed an age-related increase in antigen-specific IgG levels in seropositive individuals for most serologies, with the notable exception of toxoplasmosis. This may indicate that aging plays a general role in IgG production. An alternative explanation that requires further study is that this could be the consequence of reactivation or recurrent exposure.

2.6 Conclusions

In sum, our study provides evidence that age, sex and host genetics contribute to natural variation in humoral immunity in humans. The identified associations have the potential to help improve vaccination strategies, and/or dissect pathogenic mechanisms implicated in human diseases related to immunoglobulin production such as autoimmunity.

2.7 List of abbreviations

HLA: Human Leucocyte Antigen (HLA); CMV: Cytomegalovirus; EBV: Epstein-Barr virus; HSV1: Herpes simplex virus 1; HSV2: Herpes simplex virus 2; Ig: Immunoglobulin; HCV: Hepatitis C virus; HBV: Hepatitis B virus; H. Pylori: Helicobacter Pylori; CRP: C-reactive protein; FDR: False discovery rate; SNP: Single nucleotide polymorphism; MAF: Minor allele frequency; MI: Milieu Interieur; VZV: Varicella zoster virus; T. Gondii: Toxoplasma gondii; IAV: influenza A virus; QQ: quantile-quantile; LD: Linkage disequilibrium.

2.8 Declarations

2.8.1 Ethics approval and consent to participate

The clinical study was approved by the Comité de Protection des Personnes - Ouest 6 on June 13th, 2012, and by the French Agence Nationale de Sécurité du Médicament on June 22nd, 2012, and have been performed in accordance with the Declaration of Helsinki. 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). Informed consent was obtained from participants after the nature and possible consequences of the studies were explained.

2.8.2 Availability of data and material

Genotype data supporting the conclusions of this article are available in the European Genome-Phenome Archive under the accession code EGAS00001002460. Full summary association results are available for download from Zenodo (http://doi.org/10.5281/zenodo.1217136).

2.8.3 Competing interests

C.H. and M.L.A. are employees of Genentech Inc., a member of The Roche Group. The remaining authors declare that they have no competing interests.

2.8.4 Funding

This work benefited from support of the French government's Invest in the Future Program, managed by the Agence Nationale de la Recherche (ANR, reference 10-LABX-69-01).

It was also supported by a grant from the Swiss National Science Foundation (31003A_175603, to JF). C.A. received a PostDoctoral Fellowship from Institut National de la Recherche Médicale.

2.8.5 Acknowledgements

We would like to thank to all the donors for their contribution to the study. We also thank the members of the The *Milieu Intérieur* Consortium for their insightful comments. The *Milieu Intérieur* Consortium is composed of the following team leaders: Laurent Abel (Hôpital Necker, Paris, France), Andres Alcover (Institut Pasteur, Paris, France), Hugues Aschard (Institut Pasteur, Paris, France), Kalla Astrom (Lund University, Lund, Sweden), Philippe Bousso (Institut Pasteur, Paris, France), Pierre Bruhns (Institut Pasteur, Paris, France), Ana Cumano (Institut Pasteur, Paris, France), Caroline Demangel (Institut Pasteur, Paris, France), Ludovic Deriano (Institut Pasteur, Paris, France), James Di Santo (Institut Pasteur, Paris, France), Françoise Dromer (Institut Pasteur, Paris, France), Darragh Duffy (Institut Pasteur, Paris, France), Gérard Eberl (Institut Pasteur, Paris, France), Jost Enninga (Institut Pasteur, Paris, France), Jacques Fellay (EPFL, Lausanne, Switzerland) Odile Gelpi (Institut Pasteur, Paris, France), Ivo Gomperts-Boneca (Institut Pasteur, Paris, France), Milena Hasan (Institut Pasteur, Paris, France), Serge Hercberg (Université Paris 13, Paris, France), Olivier Lantz (Institut Curie, Paris, France), Claude Leclerc (Institut Pasteur, Paris, France), Hugo Mouquet (Institut Pasteur, Paris, France), Sandra Pellegrini (Institut Pasteur, Paris, France), Stanislas Pol (Hôpital Côchin, Paris, France), Antonio Rausell (INSERM UMR 1163 – Institut Imagine, Paris, France), Lars Rogge (Institut Pasteur, Paris, France), Anavaj Sakuntabhai (Institut Pasteur, Paris, France), Olivier Schwartz (Institut Pasteur, Paris, France), Benno Schwikowski (Institut Pasteur, Paris, France), Spencer Shorte (Institut Pasteur, Paris, France), Vassili Soumelis (Institut Curie, Paris, France), Frédéric Tangy (Institut Pasteur, Paris, France), Eric Tartour (Hôpital Européen George Pompidou, Paris, France), Antoine Toubert (Hôpital Saint-Louis, Paris, France), Mathilde Touvier (Université Paris 13, Paris, France), Marie-Noëlle Ungeheuer (Institut Pasteur, Paris, France), Matthew L. Albert (Roche Genentech, South San Francisco, CA, USA), Lluis Quintana-Murci (Institut Pasteur, Paris, France).

2.9 Additional Files

All the additional files are available at the online version of the article: https://doi.org/10.1186/s13073-018-0568-8.

Additional File 1: Table S1. Assay details for serologies. Table S2. Cutoffs and seroprevalence for serologies. Table S3. List of covariates used for each phenotype. Table S4. Associations of environmental variables with serostatus. Table S5. Association of serologies with age and sex. Table S6. Associations of amino acid positions in HLA proteins with Influenza A serology. Table S7. LD between residues in HLA-DRβ1 at position 13 and 96. Table S8. Replication of SNPs associated with levels of total IgM, IgA and IgG. Table S9. LD between residues in HLA-DRβ1 at position 15 and 11. (XLSX 70 KB)

Additional File 2: Fig.S1 Principal Component Analysis. Fig.S2 Distribution of serological variables, and clinical thresholds. Fig.S3 Seroprevalence data. Fig.S4 Impact of non-genetic factors on serostatus. Fig.S5 Evolution of serostatus with age and sex. Fig.S6 Correlations

between age and IgG specific to Rubella and T. Gondii. Fig.S7 QQ plots for logistic regressions preformed in the study. Fig.S8 QQ plots for linear regressions preformed on total Ig levels. Fig.S9 QQ plots for linear regressions preformed for pathogen-specific IgG levels. Fig.S10 QQ plots for burden testing analyses preformed for all binary phenotypes. Fig.S11 QQ plots for burden testing analyses preformed for total Ig levels. Fig.S12 QQ plots for burden testing analyses preformed for pathogen-specific IgG levels. (DOCX 92.2 MB)

2.10 References

1. Traylen CM, Patel HR, Fondaw W, Mahatme S, Williams JF, Walker LR, Dyson OF, Arce S, Akula SM. Virus reactivation: a panoramic view in human infections. Future Virol. 2011;6:451-63.

2. Grundbacher FJ. Heritability estimates and genetic and environmental correlations for the human immunoglobulins G, M, and A. Am J Hum Genet. 1974;26:1-12.

3. Tsang JS, Schwartzberg PL, Kotliarov Y, Biancotto A, Xie Z, Germain RN, Wang E, Olnes MJ, Narayanan M, Golding H, Moir S, Dickler HB, Perl S, Cheung F; Baylor HIPC Center; CHI Consortium. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. Cell. 2014;157:499-513.

4. Rubicz R, Leach CT, Kraig E, Dhurandhar NV, Duggirala R, Blangero J, Yolken R, Göring HH. Genetic factors influence serological measures of common infections. Hum Hered. 2011;72:133-41.

5. Almohmeed YH, Avenell A, Aucott L, Vickers MA. Systematic review and meta-analysis of the sero-epidemiological association between Epstein Barr virus and multiple sclerosis. PLoS One. 2013;8:e61110.

6. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, Sievertsen J, Muntau B, Ruge G, Loag W, Ansong D, Antwi S, Asafo-Adjei E, Nguah SB, Kwakye KO, Akoto AO, Sylverken J, Brendel M, Schuldt K, Loley C, Franke A, Meyer CG, Agbenyega T, Ziegler A, Horstmann RD. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012;489:443-6.

7. McLaren PJ, Coulonges C, Ripke S, van den Berg L, Buchbinder S, Carrington M, Cossarizza A, Dalmau J, Deeks SG, Delaneau O, De Luca A, Goedert JJ, Haas D, Herbeck JT, Kathiresan S, Kirk GD, Lambotte O, Luo M, Mallal S, van Manen D, Martinez-Picado J, Meyer L, Miro JM, Mullins JI, Obel N, O'Brien SJ, Pereyra F, Plummer FA, Poli G, Qi Y, Rucart P, Sandhu MS, Shea PR, Schuitemaker H, Theodorou I, Vannberg F, Veldink J, Walker BD, Weintrob A, Winkler CA, Wolinsky S, Telenti A, Goldstein DB, de Bakker PI, Zagury JF, Fellay J. Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. PLoS Pathog. 2013;9:e1003515.

8. Casanova JL, Abel L. The genetic theory of infectious diseases: a brief history and selected illustrations. Annu Rev Genomics Hum Genet. 2013; 14:215-43.

9. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, Hinds DA. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun. 2017;8:599.

10. McLaren PJ, Coulonges C, Bartha I, Lenz TL, Deutsch AJ, Bashirova A, Buchbinder S, Carrington MN, Cossarizza A, Dalmau J, De Luca A, Goedert JJ, Gurdasani D, Haas DW, Herbeck JT, Johnson EO, Kirk GD, Lambotte O, Luo M, Mallal S, van Manen D, Martinez-Picado J, Meyer L, Miro JM, Mullins JI, Obel N, Poli G, Sandhu MS, Schuitemaker H, Shea PR, Theodorou I, Walker BD, Weintrob AC, Winkler CA, Wolinsky SM, Raychaudhuri S, Goldstein DB, Telenti A, de Bakker PI, Zagury JF, Fellay J. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. Proc Natl Acad Sci U S A. 2015;112:14658-63.

11. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, McHutchison JG, Goldstein DB. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature. 2009;461:399-401.

12. Hammer C, Begemann M, McLaren PJ, Bartha I, Michel A, Klose B, Schmitt C, Waterboer T, Pawlita M, Schulz TF, Ehrenreich H, Fellay J. Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. Am J Hum Genet. 2015;97:738-43.

13. Jonsson S, Sveinbjornsson G, de Lapuente Portilla AL, Swaminathan B, Plomp R, Dekkers G, Ajore R, Ali M, Bentlage AEH, Elmér E, Eyjolfsson GI, Gudjonsson SA, Gullberg U, Gylfason A, Halldorsson BV, Hansson M, Holm H, Johansson Å, Johnsson E, Jonasdottir A, Ludviksson BR, Oddsson A, Olafsson I, Olafsson S, Sigurdardottir O, Sigurdsson A, Stefansdottir L, Masson G, Sulem P, Wuhrer M, Wihlborg AK, Thorleifsson G, Gudbjartsson DF, Thorsteinsdottir U, Vidarsson G, Jonsdottir I, Nilsson B, Stefansson K. Identification of sequence variants influencing immunoglobulin levels. Nat Genet. 2017;49:1182-91.

14. Rubicz R, Yolken R, Drigalenko E, Carless MA, Dyer TD, Kent J Jr, Curran JE, Johnson MP, Cole SA, Fowler SP, Arya R, Puppala S, Almasy L, Moses EK, Kraig E, Duggirala R, Blangero J, Leach CT, Göring HH. Genome-wide genetic investigation of serological measures of common infections. Eur J Hum Genet. 2015;23:1544-8.

15. Thomas S, Rouilly V, Patin E, Alanio C, Dubois A, Delval C, Marquier LG, Fauchoux N, Sayegrih S, Vray M, Duffy D, Quintana-Murci L, Albert ML; Milieu Intérieur Consortium. The Milieu Intérieur study - an integrative approach for study of human immunological variance. Clin Immunol. 2015;157:277-93.

16. Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, Alanio C, Scepanovic P, Hammer C, Jönsson F, Beitz B, Quach H, Lim YW, Hunkapiller J, Zepeda M, Green C, Piasecka B, Leloup L, Rogge L, Huetz F, Peguillet I, Lantz O, Fontes M, Di Santo JP, Thomas S, Fellay J, Duffy D, Quintana-Murci L, Albert ML, for The Milieu Intérieur Consortium. Natural variation in innate immune cell parameters is preferentially driven by genetic factors. Nat Immunol. 2018;19:302-314.

17. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867-73.

18. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.

19. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, L Price A. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443-48.

20. McCarthy S et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279-83.

21. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, Raychaudhuri S, de Bakker PI. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One. 2013;8:e64683.

22. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, Lathrop M, Karlsen TH, Franke A, Moffatt M, Cookson W, Trowsdale J, McVean G, Sawcer S, Leslie S. Imputation of KIR Types from SNP Variation Data. Am J Hum Genet. 2015;97:593-607.

23. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF, Vitart V, Navarro P, Zagury JF, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10:e1004234.

24. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

25. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11:407-9.

26. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2010;72:417-73.

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.

28. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80-92.

29. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2.

30. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92:841-53.

31. Mayerle J, den Hoed CM, Schurmann C, Stolk L, Homuth G, Peters MJ, Capelle LG, Zimmermann K, Rivadeneira F, Gruska S, Völzke H, de Vries AC, Völker U, Teumer A, van Meurs JB, Steinmetz I, Nauck M, Ernst F, Weiss FU, Hofman A, Zenker M, Kroemer HK, Prokisch H, Uitterlinden AG, Lerch MM, Kuipers EJ. Identification of genetic loci associated with Helicobacter pylori serologic status. JAMA. 2013;309:1912-20.

32. Swaminathan B, Thorleifsson G, Jöud M, Ali M, Johnsson E, Ajore R, Sulem P, Halvarsson BM, Eyjolfsson G, Haraldsdottir V, Hultman C, Ingelsson E, Kristinsson SY, Kähler AK, Lenhoff S, Masson G, Mellqvist UH, Månsson R, Nelander S, Olafsson I, Sigurðardottir O, Steingrimsdóttir H, Vangsted A, Vogel U, Waage A, Nahi H, Gudbjartsson DF, Rafnar T, Turesson I, Gullberg U, Stefánsson K, Hansson M, Thorsteinsdóttir U, Nilsson B. Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. Nat Commun. 2015;6:7213.

33. Viktorin A, Frankowiack M, Padyukov L, Chang Z, Melén E, Sääf A, Kull I, Klareskog L, Hammarström L, Magnusson PK. IgA measurements in over 12 000 Swedish twins reveal sex differential heritability and regulatory locus near CD30L. Hum Mol Genet. 2014;23:4177-84.

34. Frankowiack M, Kovanen RM, Repasky GA, Lim CK, Song C, Pedersen NL, Hammarström L. The higher frequency of IgA deficiency among Swedish twins is not explained by HLA haplotypes. Genes Immun. 2015;16:199-205.

35. Yang C, Jie W, Yanlong Y, Xuefeng G, Aihua T, Yong G, Zheng L, Youjie Z, Haiying Z, Xue Q, Min Q, Linjian M, Xiaobo Y, Yanling H, Zengnan M. Genome-wide association study identifies TNFSF13 as a susceptibility gene for IgA in a South Chinese population in smokers. Immunogenetics. 2012;64:747-53.

36. Liao M, Ye F, Zhang B, Huang L, Xiao Q, Qin M, Mo L, Tan A, Gao Y, Lu Z, Wu C, Zhang Y, Zhang H, Qin X, Hu Y, Yang X, Mo Z. Genome-wide association study identifies common variants at TNFRSF13B associated with IgG level in a healthy Chinese male population. Genes Immun. 2012;13:509-13.

37. Yang M, Wu Y, Lu Y, Liu C, Sun J, Liao M, Qin M, Mo L, Gao Y, Lu Z, Wu C, Zhang Y, Zhang H, Qin X, Hu Y, Zhang S, Li J, Dong M, Zheng SL, Xu J, Yang X, Tan A, Mo Z. Genome-wide scan identifies variant in TNFSF13 associated with serum IgM in a healthy Chinese male population. PLoS One. 2012;7:e47990.

38. Lambert ND, Haralambieva IH, Kennedy RB, Ovsyannikova IG, Pankratz VS, Poland GA. Polymorphisms in HLA-DPB1 are associated with differences in rubella virus-specific humoral immunity after vaccination. J Infect Dis. 2015;211:898-905.

39. Rubicz R, Yolken R, Drigalenko E, Carless MA, Dyer TD, Bauman L, Melton PE, Kent JW Jr, Harley JB, Curran JE, Johnson MP, Cole SA, Almasy L, Moses EK, Dhurandhar NV, Kraig E, Blangero J, Leach CT, Göring HH. A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). PLoS Genet. 2013;9:e1003147.

40. Pedergnana V, Syx L, Cobat A, Guergnon J, Brice P, Fermé C, Carde P, Hermine O, Le-Pendeven C, Amiel C, Taoufik Y, Alcaïs A, Theodorou I, Besson C, Abel L. Combined linkage and association studies show that HLA class II variants control levels of antibodies against Epstein-Barr virus antigens. PLoS One. 2014;9:e102501.

41. Rajagopalan S, Long EO. Understanding how combinations of HLA and KIR genes influence disease. J Exp Med. 2005;201:1025-9.

42. Goetzman ES, Alcorn JF, Bharathi SS, Uppala R, McHugh KJ, Kosmider B, Chen R, Zuo YY, Beck ME, McKinney RW, Skilling H, Suhrie KR, Karunanidhi A, Yeasted R, Otsubo C, Ellis B, Tyurina YY, Kagan VE, Mallampalli RK, Vockley J. Long-chain acyl-CoA dehydrogenase deficiency as a cause of pulmonary surfactant dysfunction. J Biol Chem. 2014;289:10668-79.

43. Giefing-Kröll C, Berger P, Lepperdinger G, Grubeck-Loebenstein B. How sex and age affect immune responses, susceptibility to infections, and response to vaccination. Aging Cell. 2015;14:309-21.

44. Cook IF. Sexual dimorphism of humoral immunity with human vaccines. Vaccine. 2008;26:3551-5.

45. Klein SL, Jedlicka A, Pekosz A. The Xs and Y of immune responses to viral vaccines. Lancet Infect Dis. 2010;10:338-49.

46. Ovsyannikova IG, Jacobson RM, Dhiman N, Vierkant RA, Pankratz VS, Poland GA. Human leukocyte antigen and cytokine receptor gene polymorphisms associated with heterogeneous immune responses to mumps viral vaccine. Pediatrics. 2008;121:e1091-9.

47. Kennedy RB, Ovsyannikova IG, Pankratz VS, Vierkant RA, Jacobson RM, Ryan MA, Poland GA. Gender effects on humoral immune responses to smallpox vaccine. Vaccine. 2009;27:3319-23.

48. Girón-González JA, Moral FJ, Elvira J, García-Gil D, Guerrero F, Gavilán I, Escobar L. Consistent production of a higher TH1:TH2 cytokine ratio by stimulated T cells in men compared with women. Eur J Endocrinol. 2000;143:31-6.

Chapter 3: HLA variants play a major role in determining the natural variation in innate immune cell parameters

This work was part of a study published in ***Nature Immunology*** (2018; 19:302–314). The full study is provided at the end of the chapter.

3.1 Introduction

The immune system has an essential role in maintaining homeostasis and several studies have suggested that extensive differences exist among healthy people in the composition of the immune cell repertoire [1,2,3].

Technological advances in flow cytometry, based on improved instrument design and the increased availability of a reagents targeting specific molecules, now permit low-cost and deep phenotyping of immune cell populations in a large cohort of individuals [4]. Combined with genome-wide DNA genotyping and assessment of environmental factors, the genetic and non-genetic basis of inter-person variation in the parameters of immune cells can thus be more easily interrogated [5].

Using a systems immunology approach, the Milieu Intérieur (MI) consortium comprehensively measured the composition of white blood cells from 1,000 healthy, unrelated people of Western European ancestry. A total of 166 distinct immunophenotypes were obtained, including 75 in innate immune cells (46%) and 91 in adaptive immune cells (54%). The immunophenotypes of both innate 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 [6].

This broad resource confirmed that age, sex, CMV seropositivity and smoking had major, independent effects on many parameters of innate and adaptive immune cells. Genome-wide association studies were also conducted, revealing 15 loci associated with parameters of circulating leukocyte subpopulations. The most prominent result was that genetic associations were primarily detected with innate cell parameters. Within these, associations with variation in the human leukocyte antigen (HLA) region were observed for 6 innate immunophenotypes [6].

The HLA region presents extreme sequence diversity, substantial linkage disequilibrium (LD) and high gene density. Sequence and structural variations differ between populations and further complicate haplotype inference. Identifying causal and independent loci from association signals in HLA is thus challenging, as often they cannot be fine-mapped to a single variant at a single locus but often comprise independent effects from multiple loci. [7]

In order to help the interpretation of the GWAS signals, i.e. to provide an additional biological context, the variation at classical HLA alleles and in HLA proteins can be considered. The complex relationship between these levels of variation and SNP variation can be inferred by HLA typing [8]. Unfortunately, direct typing of classical HLA alleles is costly and often prohibitive for many large-scale studies.

Here we leveraged a reference panel with SNP and HLA data available, together with individual genotypes of MI participant to impute their HLA alleles and variable amino acid positions in HLA proteins. This approach allowed us to obtain additional level of information without further cost involved. We then conducted a fine-mapping study of the signals observed within the HLA region with the relevant immunophenotypes.

3.2 Methods

3.2.1 The Milieu Intérieur cohort

Detailed information about recruitment of donors and immunoprofiling are available in the full study provided at the end of the chapter.

3.2.1 Genotyping, genome-wide imputation and genome-wide association analyses

The Milieu Intérieur cohort was genotyped at 945,213 single-nucleotide polymorphisms (SNPs) enriched for exonic SNPs. After quality control, genotype imputation was performed, which yielded a total of 5,699,237 highly accurate SNPs. The models of genome-wide associations were adjusted for 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. Each immunophenotype, on which the tests were performed, was imputed, transformed and batch-effect corrected. Additional information is available in the method section of the paper at the end of the chapter.

3.2.2 HLA imputation

HLA imputation was performed by using SNP2HLA v 1.03. We used as a reference panel the data from Type 1 Diabetes Genetics Consortium (T1DGC) study, which contains SNP genotyping and classic HLA serotyping information for 5225 unrelated individuals [9]. We obtained 104 four-digit classical alleles HLA alleles and 738 amino acid residues (at 315 positions) in the HLA class I and II proteins with MAF of > 1%.

3.2.3 HLA association testing, conditional and omnibus tests

Similarly to genome-wide association analyses, we conducted linear regression with HLA alleles on 6 immunophenotypes by using PLINK v.1.9 [10]. The first two principal components, calculated on the genetic matrix, were included in all tests to correct for residual population stratification. Additional correlated non-genetic variables identified for each immunophenotype were also included in the association tests. We considered associations to be significant if they passed a genome-wide significance threshold corrected for the number of phenotypes tested in the entire study (i.e. 166).

Conditional linear regression tests were run by including the identified significant associations in the model as additional covariates.

Multivariate omnibus tests were used to test for association at multi-allelic amino acid positions. If there are R amino acid residues at a tested HLA protein position, then the omnibus test performs an R-1 degrees of freedom test, comparing the alternate (each amino acid residue having a unique effect) versus the null hypothesis (no amino acid residue having any different effect). The tests were performed by using PLINK v. 1.07 [11].

3.2.4 Proportion of explained variance calculations

We calculated the explained variance of each 6 immunophenotypes with a linear regression model including the four non-genetic factors with the greatest effect (i.e., age, sex, CMV seropositivity status and smoking) and genome-wide genetic factors that were significant ($P < 1 \times 10^{-10}$). 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 R software.

3.3 Results

3.3.1 Fine-mapping of HLA association results

We conducted association tests of HLA alleles and variable residues at amino acid positions in HLA proteins with 6 immunophenotypes that had significant SNP associations in the HLA region. We observed significant associations at all levels, with the strongest association signals at the level of amino acids (Table 3.1).

Table 3.1. Summary of genome-wide significant association results in the HLA locus.

Immunophenotype	SNP		Classical HLA allele		Amino acid	
	ID	P-value	Allele	P-value	Protein (position)	Omnibus P-value
HLA-DR in cDC1	rs2760994	5.12×10^{-39}	HLA-DQA1*05:01	1.45×10^{-24}	HLA-DRβ1 (13)	5.29×10^{-41}
HLA-DR in pDC	rs114973966	2.82×10^{-58}	HLA-DQA1*03:01	1.61×10^{-56}	HLA-DRβ1 (13)	7.02×10^{-90}
CD86 in pDC	rs140872668	2.11×10^{-19}	HLA-DQA1*03:01	4.06×10^{-18}	HLA-DRβ1 (13)	4.24×10^{-22}
HLA-DR in $CD14^{hi}$ monocytes	rs116018922	4.46×10^{-40}	HLA-DQA1*03:01	2.90×10^{-36}	HLA-DRβ1 (13)	1.97×10^{-47}
HLA-DR in cDC3	rs114176373	2.77×10^{-12}	HLA-DQB1*04:02	5.61×10^{-8}	HLA-DRβ1 (74)	3.86×10^{-13}
$HLA\text{-}DR^{+}$ $CD56^{hi}$ NK cells	rs28383322	5.37×10^{-14}	HLA-DQB1*02:02	1.27×10^{-7}	HLA-DRβ1 (67)	5.38×10^{-14}

We then ran conditional tests at the level of HLA alleles and amino acids by including in the linear regression model as a covariate the genotypes of the most strongly associated HLA allele or amino acid position, respectively.
We first ran the test at the level of HLA alleles and observed one additional independent significant signal for HLA-DR in pDC phenotype (Table 3.2).

Table 3.2. Associations of HLA classical alleles and conditional tests with candidate immunophenotypes.

Immunophenotype	Risk HLA allele	Conditioning on…	P-value	Beta (95% CI)
HLA-DR in cDC1	HLA-DQA1*05:01	-	$1.4x10^{-24}$	0.11 (0.09 - 0.13)
	HLA-DQB1*05:01	HLA-DQA1*05:01	$2.6x10^{-9}$	-0.08 (-0.11 - -0.06)
HLA-DR in pDC	HLA-DQA1*03:01	-	$1.6x10^{-56}$	336.6 (297.6 - 375.6)
	HLA-DQA1*02:01	HLA-DQA1*03:01	$2.9x10^{-28}$	-253.8 (-297.3 - -210.2)
	HLA-DQA1*01:01	HLA-DQA1*03:01, HLA-DQA1*02:01	$4.4x10^{-10}$	-130.9 (-171.7 - -90.2)
CD86 in pDC	HLA-DQA1*03:01	-	$4.1x10^{-18}$	11.0 (8.6 - 13.4)
	HLA-DQA1*02:01	HLA-DQA1*03:01	$2.2x10^{-5}$	-6.2 (-9.1 - -3.4)
HLA-DR in $CD14^{hi}$ monocytes	HLA-DQA1*03:01	-	$2.9x10^{-36}$	2.3 (2.0 - 2.6)
	HLA-DQA1*02:01	HLA-DQA1*03:01	$5.8x10^{-10}$	-1.3 (-1.7 - -0.9)
HLA-DR in cDC3	HLA-DQB1*02:02	-	$5.6x10^{-8}$	15.65 (10.04 - 21.25)
$HLA\text{-}DR^{+}$ $CD56^{hi}$ NK cells	HLA-DQB1*04:02	-	$2.7x10^{-7}$	-0.26 (-0.36 - -0.17)

On the other hand, controlling for the top associated residue at amino acid position 13 in HLA-DRβ1 protein revealed two independent effects in the same phenotype - MFI of HLA-DR in pDC (Table 3.3).

Table 3.3. Significant associations upon conditioning on top residue in HLA amino-acid positions with candidate immunophenotypes.

HLA amino acid position	Omnibus test P-value	Amino-acid substitutions (frequency)	Beta (95% CI)
HLA-B position 97, conditioning on HLA-DRβ1 position 13	3.8×10^{-17}	Asn (f=3.35%)	-17.4 (-92.9 - 58.2)
		Arg (f=53.5%)	31.5 (3.1 - 59.9)
		Ser (f=24.3%)	-47.6 (-82.1 - -13.1)
		Thr (f=12.7%)	22.4 (-19.2 - 63.9)
		Val (f=2.0%)	-71.1 (-172.7 - 30.5)
		Trp (f=4.1%)	-11.7 (-83.0 - 59.5)
HLA-B position 194, conditioning on HLA-DRβ1 position 13 and HLA-B position 97	1.3×10^{-18}	Ile (f=83.2%)	-30.9 (-73.2 - 11.3)
		Val (f=16.9%)	30.9 (-11.3 - 73.2)
		Indel (f=0.1%)	-167.4 (-786.5 - 451.7)

In order to test if the top GWAS SNP association can be explained by the significant signals observed at the HLA allele and amino acid levels, we included them as covariates in the models and reran regression tests. We observed that no residual signal remained by using amino acid positions and thus we concluded that they explained the entirety of the SNP association signals (Table 3.4).

Table 3.4. Association test for top associated GWAS SNPs including significant HLA alleles and variable amino acids as covariates.

Phenotype	Associated classical HLA allele	P-value (SNP) conditional on significant HLA alleles	Associated amino acid position in HLA protein	P-value (SNP) conditional on significant amino acids
MFI of CD86 in pDC	HLA-DQA1*03:01	4.38×10^{-5}	HLA-DRβ1 (13)	3.16×10^{-4}
MFI of HLA-DR in $CD14^{hi}$ monocytes	HLA-DQA1*03:01	1.62×10^{-3}	HLA-DRβ1 (13)	1
MFI of HLA-DR in cDC1	HLA-DQA1*05:01	3.24×10^{-22}	HLA-DRβ1 (13)	0.05
MFI of HLA-DR in pDC	HLA-DQA1*03:01, HLA-DQA1*02:01	2.48×10^{-5}	HLA-DRβ1 (13), HLA-B (97), HLA-B (194)	1
$HLA\text{-}DR^{+}$ $CD56^{hi}$ NK cells			HLA-DRβ1 (67)	2×10^{-3}
MFI of HLA-DR in cDC3			HLA-DRβ1 (74)	3.18×10^{-4}

3.3.2 Variance explained by HLA amino acids

We then sought to calculate the additional variance explained by the strongest signals we observed in the HLA region – the variable amino acid positions in the HLA proteins. We observed a significant increase of explained variance for each of the phenotype (Table 3.5).

Table 3.5. Variance explained by the associated variable HLA amino acids.

Phenotype	r^2 (covariates)	Associated amino acid position in HLA protein	r^2 (covariates, amino acids)	Δr^2
MFI of CD86 in pDC	22.50%	HLA-DRβ1 (13)	28.50%	6.00%
MFI of HLA-DR in $CD14^{hi}$ monocytes	3.10%	HLA-DRβ1 (13)	20.20%	17.10%
MFI of HLA-DR in cDC1	3.70%	HLA-DRβ1 (13)	17.50%	13.80%
MFI of HLA-DR in pDC	3.10%	HLA-DRβ1 (13), HLA-B (97), HLA-B (194)	31.60%	28.50%
$HLA\text{-}DR^{+}$ $CD56^{hi}$ NK cells	3.80%	HLA-DRβ1 (67)	9.50%	5.70%
MFI of HLA-DR in cDC3	1.80%	HLA-DRβ1 (74)	8.00%	6.20%

3.4 Discussion

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, in conventional DCs (cDC1 cells, as defined by the expression of the transmembrane glycoprotein BDCA1) and in cDC3 cells (as defined on the basis of their expression of the integral membrane protein BDCA3). We here looked to determine if these signals were independent of each other and 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 acid–altering variants at the same codon in position 13 of the HLA-DRβ1 protein. A different amino acid variant, at position 67 of HLA-DRβ1, was identified as associated with cDC3 cells. 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 and position 194). We additionally fine-mapped the initial signal at the level of SNP to HLA amino acid variation for two other innate immunophenotypes – MFI of CD86 in pDC cells and the number of $HLA\text{-}DR^{+}$ $CD56^{hi}$ NK cells. Collectively, these results showed that the protein expression of markers of innate immune cell differentiation and activation were strongly affected by common genetic variants in the HLA region.

These results could have a broader significance for personalized diagnosis of immune-related diseases. For example, expression of HLA-DR on monocytes can be measured by flow cytometry to predict the clinical course of septic shock and identify patients who might benefit from immunoadjuvant therapies [12]. 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. Additionally, the position 13 of the HLA-DRβ1 protein that we identified as a predictor of HLA-DR expression at the surface of pDCs and monocytes, has been shown to explain a large part of the association signal in the HLA locus for type 1 diabetes [13] and this would suggest an association of innate immunity with the disease [14].

Together these findings provide a new insight into the mechanisms underlying disease pathogenesis and further evaluation of the natural variability in cellular mediators of immunity will improve our understanding of the involvement of the immune system in human health and disease.

3.5 References

1. Bernard C. Introduction à l'étude de la médecine expérimentale. Libraires de l'Académie Impériale de Médecine. 1865.

2. Tollerud DJ, Clark JW, Brown LM, Neuland CY, Pankiw-Trost LK, Blattner WA, Hoover RN. The influence of age, race, and gender on peripheral blood mononuclear-cell subsets in healthy nonsmokers. J Clin Immunol. 1989;9:214-22.

3. Reichert T, DeBruyère M, Deneys V, Tötterman T, Lydyard P, Yuksel F, Chapel H, Jewell D, Van Hove L, Linden J. Lymphocyte subset reference ranges in adult Caucasians. Clin Immunol Immunopathol. 1991;60:190-208.

4. Hasan M, Beitz B, Rouilly V, Libri V, Urrutia A, Duffy D, Cassard L, Di Santo JP, Mottez E, Quintana-Murci L, Albert ML, Rogge L; Milieu Intérieur Consortium. Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping. Clin Immunol. 2015;157:261-76.

5. Liston A, Carr EJ, Linterman MA. Shaping Variation in the Human Immune System. Trends Immunol. 2016; 37:637-646.

6. Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, Alanio C, Scepanovic P, Hammer C, Jönsson F, Beitz B, Quach H, Lim YW, Hunkapiller J, Zepeda M, Green C, Piasecka B, Leloup L, Rogge L, Huetz F, Peguillet I, Lantz O, Fontes M, Di Santo JP, Thomas S, Fellay J, Duffy D, Quintana-Murci L, Albert ML, for The Milieu Intérieur Consortium. Natural variation in innate immune cell parameters is preferentially driven by genetic factors. Nat Immunol. 2018;19:302-314.

7. Matzaraki V, Kumar V, Wijmenga C, Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. Genome Biol. 2017;18:76.

8. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.

9. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, Raychaudhuri S, de Bakker PI. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One. 2013;8:e64683.

10. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.

12. Venet F, Lukaszewicz AC, Payen D, Hotchkiss R, Monneret G. Monitoring the immune response in sepsis: a rational approach to administration of immunoadjuvant therapies. Curr Opin Immunol. 2013;25:477-83.

13. Hu X, Deutsch AJ, Lenz TL, Onengut-Gumuscu S, Han B, Chen WM, Howson JM, Todd JA, de Bakker PI, Rich SS, Raychaudhuri S. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. Nat Genet. 2015;47:898-905.

14. Astle WJ, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167:1415-1429.e19.

3.6 Full study (published in *Nature Immunology*, 2018)

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

Authors: Etienne Patin[1,2,3†*], Milena Hasan[4†], Jacob Bergstedt[5,6†], Vincent Rouilly[3,4], Valentina Libri[4], Alejandra Urrutia[4,7,8,9], Cécile Alanio[4,7,8], **Petar Scepanovic**[10,11], Christian Hammer[10,11], Friederike Jönsson[12,13], Benoît Beitz[4], Hélène Quach[1,2,3], Yoong Wearn Lim[9], Julie Hunkapiller[14], Magge Zepeda[15], Cherie Green[16], Barbara Piasecka[1,2,3,4], Claire Leloup[14], Lars Rogge[4,17], François Huetz[18,19], Isabelle Peguillet[20,21], Olivier Lantz[20,21,22,23], Magnus Fontes[6,24], James P. Di Santo[4,8,25], Stéphanie Thomas[4,7,8], Jacques Fellay[9,10], Darragh Duffy[4,7,8], Lluís Quintana-Murci[1,2,3†], Matthew L. Albert[4,7,8,9†*] and The Milieu Intérieur Consortium.

Author Affiliations: 1 Unit of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France; 2 CNRS UMR 2000, Paris, France; 3 Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France; 4 Center for Translation Research, Institut Pasteur, Paris, France; 5 Department of Automatic Control, Lund University, Lund, Sweden; 6 International Group for Data Analysis, Institut Pasteur, Paris, France; 7 Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris, France; 8 INSERM U1223, Paris, France; 9 Department of Cancer Immunology, Genentech, South San Francisco, CA, USA; 10 School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; 11 Swiss Institute of Bioinformatics, Lausanne, Switzerland; 12 Antibodies in Therapy and Pathology, Department of Immunology, Institut Pasteur, Paris, France; 13 INSERM U760, Paris, France; 14 Department of Human Genetics, Genentech, South San Francisco, CA, USA; 15 Employee Donation Program, Genentech, South San Francisco, CA, USA; 16 Department of Development Sciences, Genentech, South San Francisco, CA, USA; 17 Immunoregulation Unit, Department of Immunology, Institut Pasteur, Paris, France; 18 INSERM U783, Faculté de Médecine, Site Necker-Enfants Malades, Université Paris Descartes, Paris, France; 19 Lymphocyte Population Biology, CNRS URA 1961, Institut Pasteur, Paris, France; 20 Center of Clinical Investigations CIC-BT1428 IGR/Curie, Paris, France; 21 Department of Biopathology, Institut Curie, Paris, France; 22 Equipe Labellisée de la Ligue de Lutte Contre le Cancer, Institut Curie, Paris, France; 23 INSERM/ Institut Curie U932, Paris, France; 24 Centre for Mathematical Sciences, Lund University, Lund, Sweden; 25 Innate Immunity Unit, Institut Pasteur, Paris, France; † These authors contributed equally to the study; * Corresponding Authors.

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.

Introduction: 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*". [1] Host–pathogen 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 [2,3]. However, several studies have suggested that extensive differences also exist among healthy people in the composition of their white blood cells [4,5]. 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 [6]. Several subpopulations of activated and memory T cells increase with age [7], which might result in part from diminished thymic activity [8] and might explain reduced vaccination efficacy in the elderly [9]. Seasonal fluctuations in B cells, regulatory T cells (T_{reg} cells) and monocytes [10] and a strong effect of cohabitation on human immunological profiles [11] 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 [12], has been associated with an increased number of effector memory T cells [13], which could in turn alter immune responses to heterologous infection [14]. 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 [15]. 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 [16]. 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 [17]. 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 [10]. 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.

Results:

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 [18].

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 [19] and included granulocytes (neutrophils, basophils and eosinophils), monocytes, natural killer (NK) cells, dendritic cells (DCs) and innate lymphoid cells (ILCs) (Fig. 3.1). Adaptive cells were defined by their dependence on activity of the RAG1–RAG2 recombinase and included T cells (γδ 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).

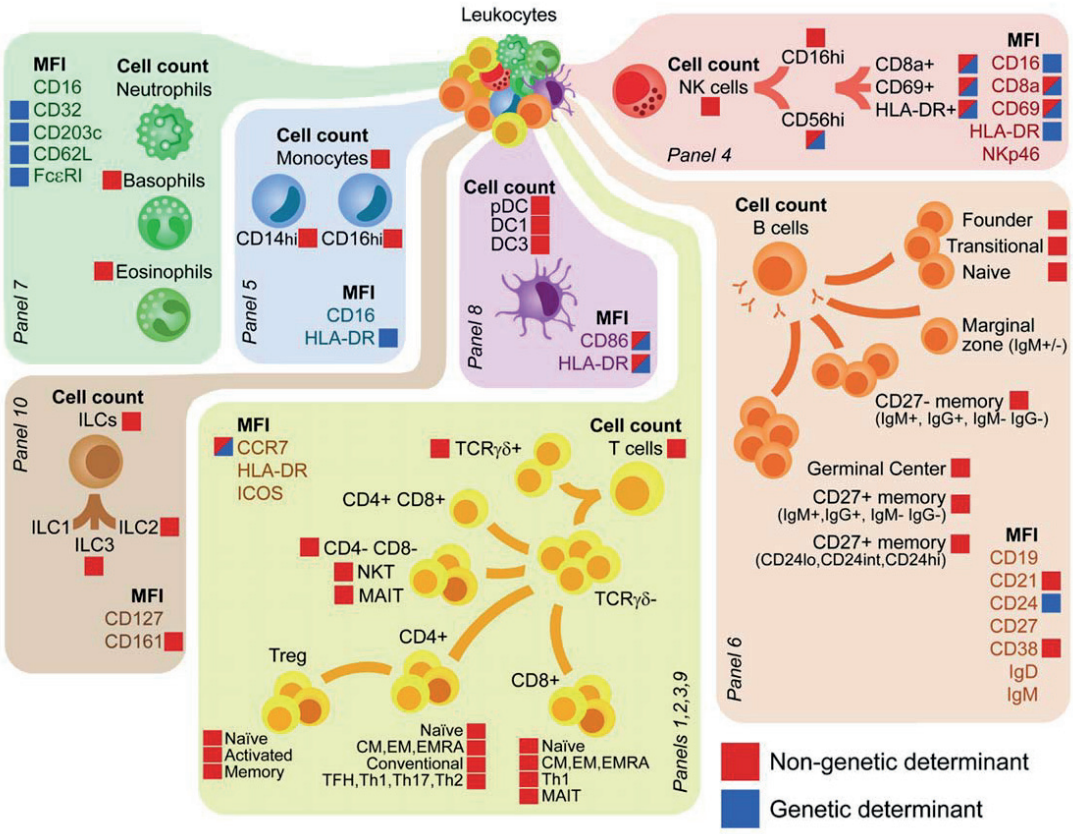

Fig. 3.1. Quantification of immune cells and cell-surface markers measured in the Milieu Intérieur cohort. Strategy: flow cytometry was used to quantify (as MFI) the expression of phenotypic markers of differentiation or activation in cells of various lineages or differentiation states (interconnecting 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.

To reduce technical variation introduced by sample-temperature fluctuations and pre-analytical procedures, we strictly followed a standardized protocol for tracking and processing samples [20]. 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 [21]. 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 [11], 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 γc cytokines [22]. Likewise, MAIT cells and $CCR6^+CD8^+$ T cells were also correlated, owing to the former's being the major subset of $CCR6^+$ T cells in the circulation [23]. 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 [24].

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 [6,7,10,14,25–27]. 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. 3.2a 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 CD16hi monocytes with increasing age (Fig. 3.2a), which might contribute to the altered immune response to viral infection in elderly people and age-associated inflammation [14,28,29]. 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 [13], 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. 3.2a–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 [30].

Although sex differences have been previously reported for various immune responses and diseases [25], 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. 3.2d 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. 3.2e–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 [6,13,14,17]. We observed a significant effect of latent infection with CMV on 13% of the parameters of immune cells (Fig. 3.2g 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 effects13,17. 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. 3.2g–i). 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. 3.2g). 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.

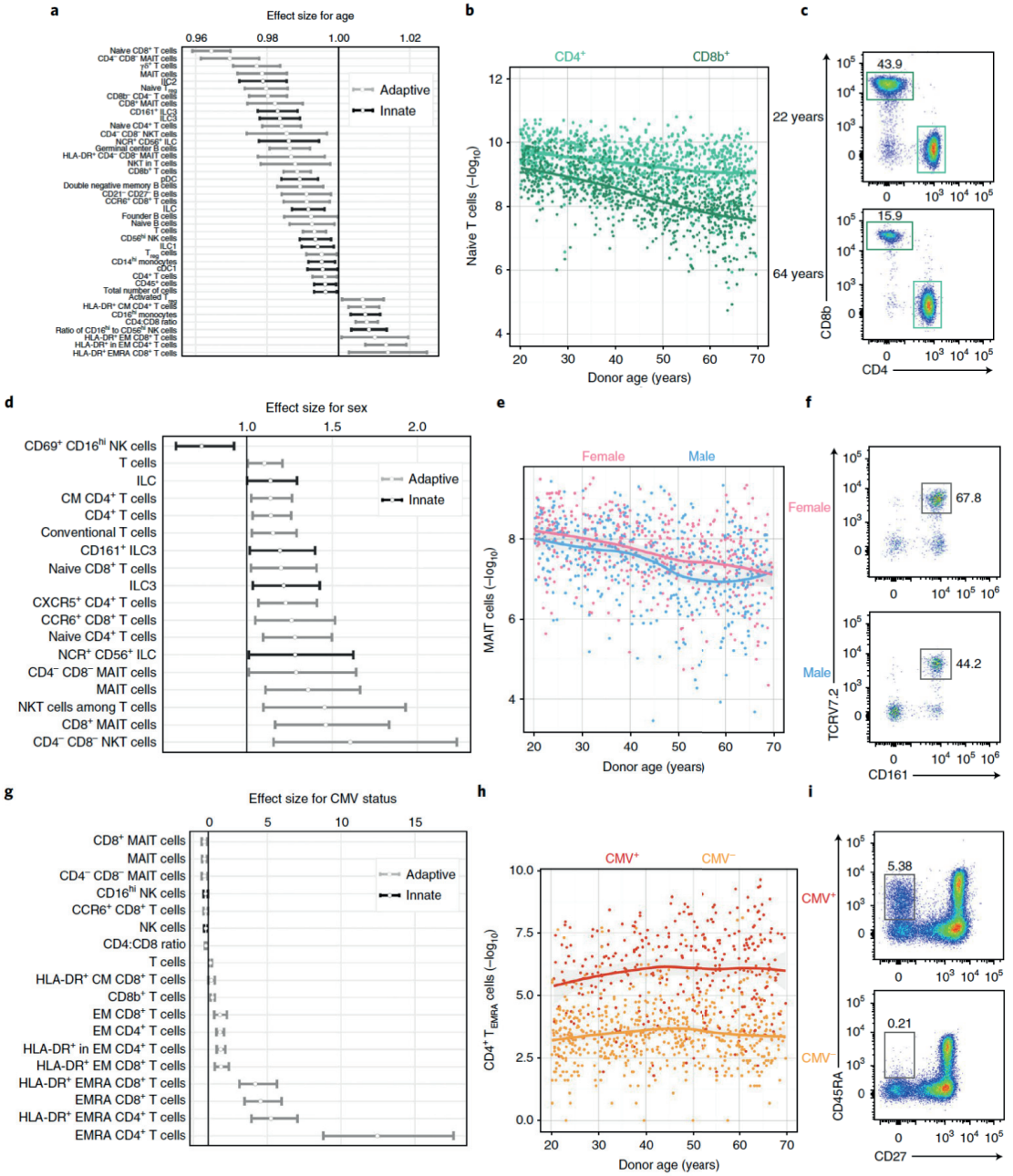

Fig. 3.2. Effects of age, sex and CMV infection on the number of innate and adaptive cells in healthy people. a,d,g, Quantification of the effect of age (a), sex (d) and CMV serostatus (g) on the abundance of circulating adaptive or innate immune cells (key; left margin) obtained from healthy donors (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. 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.

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. 3.3a 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. 3.3b). 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 [31]. 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. 3.3b–d). Active smokers also showed a decreased number of NK cells, ILCs, γδ T cells and various subsets of MAIT cells (Fig. 3.3b). These findings were consistent with a study showing that smoking triggers local release of IL-33 by the lung epithelium [32], which in turn engages the IL-33 receptor ST2 on both innate lymphocytes and non-classical lymphocytes [33]. 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.

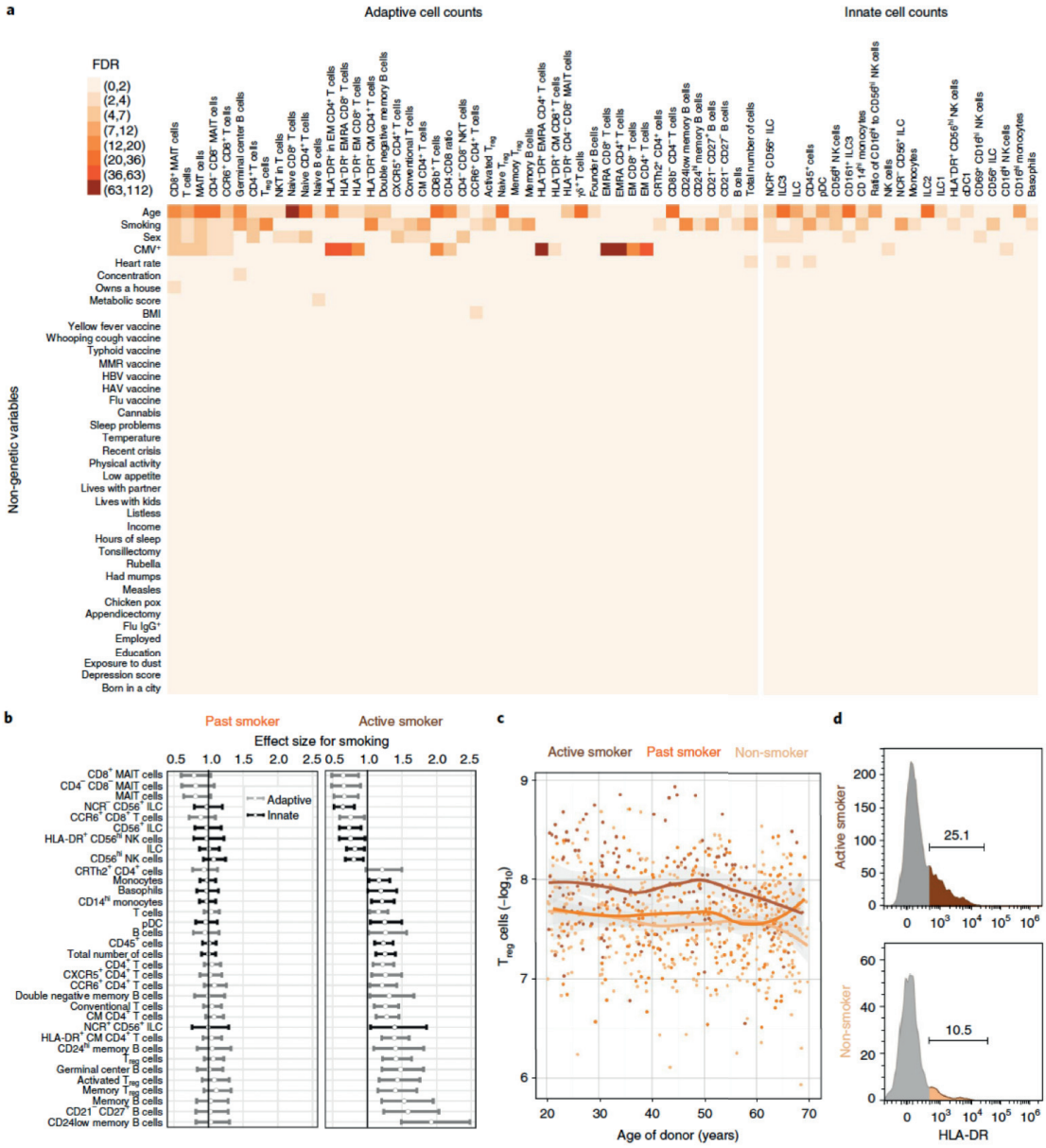

Fig. 3.3. Effects of smoking on the number of innate and adaptive immune cells in healthy people. a, Association between 39 non-genetic factors (left margin) and the number of adaptive and innate cells (above plot) in healthy donors (n = 1,000), presented as $-\log_{10}$ of adjusted P values with a false-discovery 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). 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– Treg cells. Numbers above bracketed lines indicate percent HLA-DR$^+$ T_{reg} cells (red or tan curve; effect of smoking on MFI, Supplementary Fig. 19).

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 genotype–phenotype associations by simulations and by empirically replicating well-known genetic associations with nonimmunological 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. 3.4a, Table 3.1, 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 3.1) 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 3.1). 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 3.1). 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). 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.

Table 3.6. Genome-wide signals of association 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 in 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	Immunophenotype	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	*S1PR1*	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 $HLA\text{-}DR^{+}CD16^{hi}$ 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; $HLA\text{-}DR^{+}CD56^{hi}$ 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

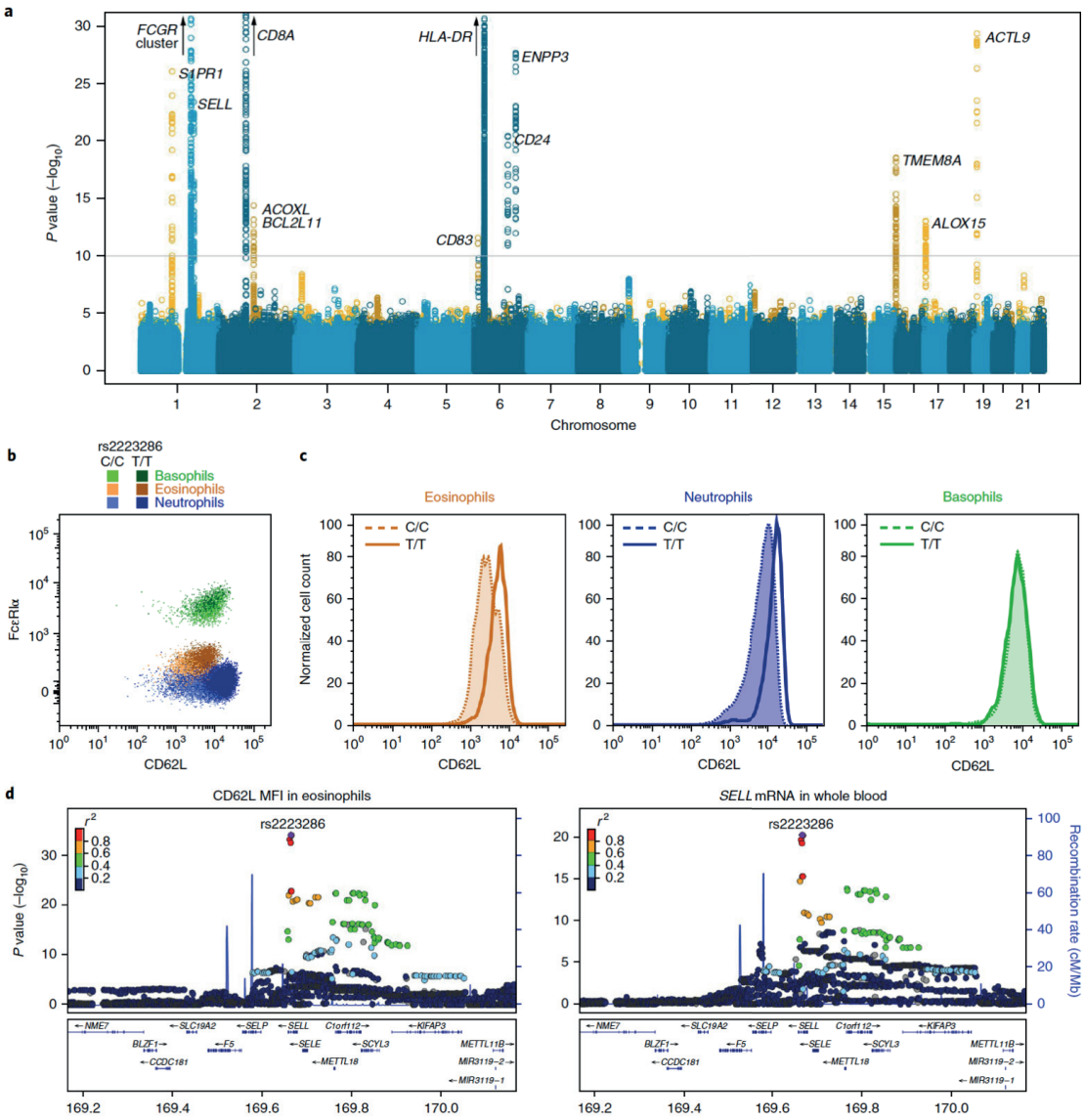

Fig. 3.4. Genome-wide significant associations with 166 immunophenotypes measured in healthy people. a, Genome-wide significant associations with variants acting locally (local-pQTLs (blue)) or not (cell count QTLs or trans-pQTLs (yellow)) on immunophenotypes in healthy subjects (n = 1,000), presented 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). 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.

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 3.1 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+CD16hi NK cells (rs71411868; $P = 5.9 \times 10^{-58}$).

We identified two independent local-pQTLs in the *FCGR* cluster (Table 3.1), 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 CD16hi 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 [15]. 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 [34]. 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 neutrophils (rs2223286; $P = 1.6 \times 10^{-35}$ and $P = 8.8 \times 10^{-13}$, respectively) but not in basophils (Fig. 3.4b,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 CD14hi 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 CD14hi monocytes, pDCs and cDC1 cells actually resulted from different amino acid–altering variants at the same codon in position 13 of the HLA-DRβ 1 protein (omnibus test $P = 2.0 \times 10^{-47}$, $P = 7.0 \times 10^{-90}$ and $P = 5.3 \times 10^{-41}$ in CD14hi 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 [35]. A different amino acid variant, at position 67 of HLA-DRβ1, 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.

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 acid–altering 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 [36] and results from the Genotype–Tissue Expression Project [37]. 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. 3.4d). The SNPs that controlled the MFI of CD16 in CD16hi NK cells and that of CD32 in basophils, CD62L in eosinophils, CD8a in CD69+ CD16hi 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.

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 3.1 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 CD16hi 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 [38]. 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 [39] and is associated with chronic lymphocytic leukemia and the total number of blood cells [40]. 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 [10].

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 [37] (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 [41], 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 [42], suggested an important role for this lipoxygenase in IgE-dependent allergic reactions. Finally, conditional GWAS identified an additional transacting 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.

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 3.1), 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 3.1) 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.

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. 3.2 and 3.3) and all genome-wide significant and suggestive variants (Table 3.1 and Supplementary Table 6) and estimated their respective contribution(s) to the total variance. We found that a larger proportion of the variance of the immunophenotypes of innate cells (Fig. 3.5b,d) than that of the immunophenotypes of adaptive cells (Fig. 3.5a,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. 3.5a). 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 × 10^{-3} (bootstrap); P = 8.1 × 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.

Fig. 3.5. 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. 3.2.) or by environmental exposure

(CMV infection and smoking; Figs. 3.2 and 3.3) and genetic factors (independent significant and suggestive GWAS hits, Table 3.1 and Supplementary Table 6).

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 to infection [43]. 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 [44]. 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 [45,46]. 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 [27]. 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 [47]. 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 [17]. 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 [48], 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 [10,15,17].

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 [35]; this would suggest an association of innate immunity with the disease [27]. 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 [49], suggestive of the possible involvement of NK cells in this pathology [50]. More generally, genetic variants found to modulate parameters of innate immune cells, in our study here and in published studies [10,15,16], 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.

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 [18]. 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 [20].

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 [20].

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 [20]. 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 labelled as outliers were also labelled 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 [18]. 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 package51. 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. 3.3 and Supplementary Fig. 19), except when they were the treatment variable to be tested (Fig. 3.2 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

[6,7,10,14,25–27], 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 [6,13,14,17]. We also controlled for genome-wide significant SNPs for corresponding immunophenotypes (Table 3.1). 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 [52] and can readily be applied using the pbkrtest R package [53]. 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 [54]. 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 [55]. Genetic structure was visualized with the Principal Component Analysis (PCA) implemented in EIGENSTRAT [56]. 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 [57].

Genotype imputation.

Prior to imputation, we phased the final SNP data set with SHAPEIT2 [58] 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.2 [59], 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 [60], 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 [61], 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 [62,63], with elastic net regression as the selection algorithm. A selection algorithm uses a cost function that drives regression parameters of nonpredictive 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 [64]. To perform stability selection, we estimated, for each of the $i \in \{1, \ldots, 42\}$ variables, the probability $pi = P(\beta i = 0)$ that the elastic net regression parameter βi of variable i equals zero. Specifically, we first took 50 subsamples of half of the data, performed variable selection on each subsample, and estimated pi 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 pi. It has been shown that this procedure, with the right threshold and under certain assumptions, controls the FDR of selected variables [63]. 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 power65) 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 [66], 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 [67]). 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 3.1), 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. [68]). 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') [69].

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 [70]. 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 [71], 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 1×. 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 [36]. 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 log2 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 [72]. 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. 3.2) (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 [73]. 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 Pearson's r value was 0.039.

Data 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).

Additional information.

Supplementary information is available at the online version of the paper at https://doi.org/10.1038/ s41590-018-0049-7.

References:

1. Bernard, C. Introduction à l'étude de la médecine expérimentale. (Libraires de l'Académie Impériale de Médecine, 1865).
2. Altfeld, M. & Gale, M. Jr. Innate immunity against HIV-1 infection. Nat. Immunol. 16, 554–562 (2015).
3. Orme, I. M., Robinson, R. T. & Cooper, A. M. The balance between protective and pathogenic immune responses in the TB-infected lung. Nat. Immunol. 16, 57–63 (2015).
4. Tollerud, D. J. et al. The influence of age, race, and gender on peripheral blood mononuclear-cell subsets in healthy nonsmokers. J. Clin. Immunol. 9, 214–222 (1989).
5. Reichert, T. et al. Lymphocyte subset reference ranges in adult Caucasians. Clin. Immunol. Immunopathol. 60, 190–208 (1991).
6. Liston, A., Carr, E. J. & Linterman, M. A. Shaping Variation in the Human Immune System. Trends Immunol. 37, 637–646 (2016).
7. Goronzy, J. J. & Weyand, C. M. Successful and maladaptive T cell aging. Immunity 46, 364–378 (2017).
8. Sauce, D. & Appay, V. Altered thymic activity in early life: how does it affect the immune system in young adults? Curr. Opin. Immunol. 23, 543–548 (2011).
9. Furman, D. et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol. Syst. Biol. 9, 659 (2013).
10. Aguirre-Gamboa, R. et al. Differential effects of environmental and genetic factors on T and B cell immune traits. Cell Rep 17, 2474–2487 (2016).
11. Carr, E. J. et al. The cellular composition of the human immune system is shaped by age and cohabitation. Nat. Immunol. 17, 461–468 (2016).

12. Boeckh, M. & Geballe, A. P. Cytomegalovirus: pathogen, paradigm, and puzzle. J. Clin. Invest. 121, 1673–1680 (2011).
13. Wertheimer, A. M. et al. Aging and cytomegalovirus infection differentially and jointly affect distinct circulating T cell subsets in humans. J. Immunol. 192, 2143–2155 (2014).
14. Furman, D. et al. Cytomegalovirus infection enhances the immune response to influenza. Sci. Transl. Med. 7, 281ra43 (2015).
15. Orrù, V. et al. Genetic variants regulating immune cell levels in health and disease. Cell 155, 242–256 (2013).
16. Roederer, M. et al. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. Cell 161, 387–403 (2015).
17. Brodin, P. et al. Variation in the human immune system is largely driven by non-heritable influences. Cell 160, 37–47 (2015).
18. Thomas, S. et al. The Milieu Intérieur study - an integrative approach for study of human immunological variance. Clin. Immunol. 157, 277–293 (2015).
19. Vivier, E. et al. Innate or adaptive immunity? the example of natural killer cells. Science 331, 44–49 (2011).
20. Hasan, M. et al. Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping. Clin. Immunol. 157, 261–276 (2015).
21. Patterson, S. et al. Cortisol patterns are associated with T cell activation in HIV. PLoS ONE 8, e63429 (2013).
22. Serafini, N., Vosshenrich, C. A. J. & Di Santo, J. P. Transcriptional regulation of innate lymphoid cell fate. Nat. Rev. Immunol. 15, 415–428 (2015).
23. Dusseaux, M. et al. Human MAIT cells are xenobiotic-resistant, tissuetargeted, CD161hi IL-17-secreting T cells. Blood 117, 1250–1259 (2011).
24. Amado, I. F. et al. IL-2 coordinates IL-2-producing and regulatory T cell interplay. J. Exp. Med. 210, 2707–2720 (2013).
25. Pennell, L. M., Galligan, C. L. & Fish, E. N. Sex affects immunity. J. Autoimmun. 38, J282–J291 (2012).
26. Furman, D., Hejblum, B. P., Simon, N., Jojic, V., Dekker, C. L., Thiébaut, R., Tibshirani, R. J. & Davis, M. M. Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. Proc Natl Acad Sci USA (2), 869–874 (2014).
27. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell 167, 1415–1429.e19 (2016).
28. Della Bella, S. et al. Peripheral blood dendritic cells and monocytes are differently regulated in the elderly. Clin. Immunol. 122, 220–228 (2007).
29. Puchta, A. et al. TNF drives monocyte dysfunction with age and results in impaired anti-pneumococcal immunity. PLoS. Pathog. 12, e1005368 (2016).
30. Vrisekoop, N. et al. Sparse production but preferential incorporation of recently produced naive T cells in the human peripheral pool. Proc. Natl Acad. Sci. USA 105, 6115–6120 (2008).
31. Tsuchiya, M. et al. Smoking a single cigarette rapidly reduces combined concentrations of nitrate and nitrite and concentrations of antioxidants in plasma. Circulation 105, 1155–1157 (2002).
32. Kearley, J. et al. Cigarette smoke silences innate lymphoid cell function and facilitates an exacerbated type I interleukin-33-dependent response to infection. Immunity 42, 566–579 (2015).

33. Monticelli, L. A. et al. Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. Nat. Immunol. 12, 1045–1054 (2011).
34. Cassard, L., Jönsson, F., Arnaud, S. & Daëron, M. Fcγ receptors inhibit mouse and human basophil activation. J. Immunol. 189, 2995–3006 (2012).
35. Hu, X. et al. Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. Nat. Genet. 47, 898–905 (2015).
36. Piasecka, B. et al. Distinctive roles of age, sex and genetics in shaping transcriptional variation of human immune responses to microbial challenges. Proc. Natl. Acad. Sci. USA 115, E488–E497 (2017).
37. GTEx Consortium. The Genotype–Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660 (2015).
38. Garris, C. S., Blaho, V. A., Hla, T. & Han, M. H. Sphingosine-1-phosphate receptor 1 signalling in T cells: trafficking and beyond. Immunology 142, 347–353 (2014).
39. Pellegrini, M. et al. Loss of Bim increases T cell production and function in interleukin 7 receptor-deficient mice. J. Exp. Med. 200, 1189–1195 (2004).
40. van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369–375 (2012).
41. Motohashi, T. et al. Molecular cloning and chromosomal mapping of a novel protein gene, M83. Biochem. Biophys. Res. Commun. 250, 244–250 (2000).
42. Feltenmark, S. et al. Eoxins are proinflammatory arachidonic acid metabolites produced via the 15-lipoxygenase-1 pathway in human eosinophils and mast cells. Proc. Natl Acad. Sci. 105, 680–685 (2008).
43. Stämpfli, M. R. & Anderson, G. P. How cigarette smoke skews immune responses to promote infection, lung disease and cancer. Nat. Rev. Immunol. 9, 377–384 (2009).
44. Venet, F., Lukaszewicz, A.-C., Payen, D., Hotchkiss, R. & Monneret, G. Monitoring the immune response in sepsis: a rational approach to administration of immunoadjuvant therapies. Curr. Opin. Immunol. 25, 477–483 (2013).
45. Kolaczkowska, E. & Kubes, P. Neutrophil recruitment and function in health and inflammation. Nat. Rev. Immunol. 13, 159–175 (2013).
46. Farber, D. L., Yudanin, N. A. & Restifo, N. P. Human memory T cells: generation, compartmentalization and homeostasis. Nat. Rev. Immunol. 14, 24–35 (2014).
47. Mangino, M., Roederer, M., Beddall, M. H., Nestle, F. O. & Spector, T. D. Innate and adaptive immune traits are differentially affected by genetic and environmental factors. Nat. Commun. 8, 13850 (2017).
48. Casanova, J. L. & Abel, L. Disentangling inborn and acquired immunity in human twins. Cell 160, 13–15 (2015).
49. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nat. Genet. 47, 1449–1456 (2015).
50. von Bubnoff, D. et al. Natural killer cells in atopic and autoimmune diseases of the skin. J. Allergy Clin. Immunol. 125, 60–68 (2010).
51. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using. J. Stat. Softw. 67, 41–48 (2015).
52. Kenward, M. G. & Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983–997 (1997).
53. Halekoh, U. & Højsgaard, S. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models - The R package pbkrtest. J. Stat. Softw. 59, 1–30 (2014).

54. Benjamini, Y. & Yekutieli, D. False discovery rate–adjusted multiple confidence intervals for selected parameters. J. Am. Stat. Assoc. 1000, 71–93 (2005).
55. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873 (2010).
56. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. PLoS. Genet. 2, e190 (2006).
57. Behar, D. M. et al. The genome-wide structure of the Jewish people. Nature 466, 238–242 (2010).
58. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10, 5–6 (2013).
59. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
60. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007).
61. Mefford, J. & Witte, J. S. The Covariate's Dilemma. PLoS Genet. 8, e1003096 (2012).
62. Meinshausen, N. & Bühlmann, P. Stability selection. J. R. Stat. Soc. B 72, 417–473 (2010).
63. Shah, R. D. & Samworth, R. J. Variable selection with error control: another look at stability selection. J. R. Stat. Soc. B 75, 55–80 (2013).
64. Hastie, T., Tibshirani, R. & Friedman, J. Elements of Statistical Learning (Springer, 2009).
65. Wakefield, J. Bayesian and Frequentist Regression Methods (Springer, 2013).
66. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11, 407–409 (2014).
67. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46, 100–106 (2014).
68. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360 (2010).
69. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
70. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE 8, e64683 (2013).
71. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
72. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294–1296 (2007).
73. Grömping, U. Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. 17, 1–27 (2006).

Chapter 4: A comprehensive assessment of demographic, environmental and host genetic associations with gut microbiome diversity in healthy individuals

Petar Scepanovic[1,2], Flavia Hodel[1,2], Stanislas Mondot[3], Valentin Partula[4,5], Christian Hammer[6], Etienne Patin[7,8], Mathilde Touvier[4], Olivier Lantz[9,10], Matthew L. Albert[6], Darragh Duffy[11], Lluis Quintana-Murci[7,8], Jacques Fellay[1,2,12*] and The *Milieu Intérieur* Consortium.

Author Affiliations: 1 School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; 2 Swiss Institute of Bioinformatics, Lausanne, Switzerland; 3 MICALIS Institute (INRA/AgroParisTech), Jouy-en-Josas, France; 4 Sorbonne-Paris-Cité Research Center for Epidemiology and Statistics CRESS, Nutritional Epidemiology Research Team EREN (INSERM U1153/INRA U1125/CNAM/Université Paris-XIII Nord), Bobigny, France; 5 University of Paris-VII Denis Diderot, Sorbonne-Paris-Cité University, Paris, France; 6 Department of Cancer Immunology, Genentech Inc., San Fransisco, CA 94080, USA; 7 Unit of Human Evolutionary Genetics, Department of Genomes and Genetics, Institut Pasteur, Paris, France; 8 Centre National de la Recherche Scientifique, UMR2000, Paris, France; 9 Institut Curie, PSL Research University, Inserm U932, 75005, Paris, France; 10 Center of Clinical Investigations, CICBT1428 IGR/Curie, 75005, Paris, France; 11 Immunobiology of Dendritic Cells laboratory (INSERM U1223/Institut Pasteur), Paris, France; 12 Precision Medicine Unit, Lausanne University Hospital, Lausanne, Switzerland; * corresponding author.

Contribution to the study: I design the study and conducted the analyses of genetic influences on the gut microbiome. Together with Flavia Hodel, I analysed the associations of the environmental variables. I wrote the original draft of the manuscript.

4.1 Abstract

Introduction. The gut microbiome is an important determinant of human health. It consists in a complex mix of microbial species with large compositional variation across individuals. This diversity is influenced by multiple environmental factors but also by human genetic variation. In the framework of the *Milieu Intérieur* Consortium, a population-based study aimed at deciphering immune response variance in healthy individuals, we assessed the commensal intestinal microbiome of 1000 individuals by 16S rRNA gene sequencing. We

identified demographic, environmental and clinical variables associated with gut microbiome diversity and used those as covariates in genome-wide association studies.

Results. A total of 1'000 healthy individuals of Western European ancestry, with a 1:1 sex ratio and stratified across five-decades of life (age 20 – 69), were recruited in France and genotyped using the Illumina OmniExpress and HumanExome arrays. 16S rRNA profiles were obtained from stool samples in 858 non-related individuals. We collected detailed demographic and environmental information through a questionnaire, as well as multiple results from standardized blood analyses. We used Spearman correlations, permutational analysis of variance and multivariate association with linear models to identify variables correlated with α-diversity, ß-diversity, or microbial community abundances. We then used linear and logistic regression to search for associations between >5 million single nucleotide polymorphisms (SNPs) and the same indicators of gut microbiome diversity, including the significant non-genetic factors as covariates. No genome-wide significant associations were identified after correction for multiple testing. A small fraction of previously reported associations between specific taxa and host genetic variants could be replicated in our cohort, while no replication was observed for gut microbiome diversity metrics.

Conclusion. In a well-characterized cohort of healthy individuals, age was the only variable that consistently influenced the gut microbiome composition. Upon careful adjustment for demographics, environmental and clinical factors, we did not observe any convincing association between specific human polymorphisms and diversity metrics or taxonomic content of the gut microbiome.

4.2 Background

A wide diversity of microbial species colonizes the human body. [1] Through a range of functions these microbes provide considerable benefits to the host. They notably generate metabolites that can act as energy sources for cell metabolism, promote the development and the functionality of the immune system and prevent colonization by pathogenic microorganisms. [2]

The human intestine harbors a particularly rich microbial community. Multiple 16s rRNA gene sequencing and metagenomic studies established that each individual gut microbiome, defined as the collection of genomes and their products of resident microorganisms, harbors a unique combination of microbial life. [3, 4] An estimated 150 to 400 species reside in each person's gut. [5]

Typically, the human gut microbiome is dominated by four phyla: *Bacteroidetes*, *Firmicutes*, *Actinobacteria* and *Proteobacteria*. [6, 7] They contain almost all of the bacterial species found in the human gastrointestinal tract, which can also be classified in higher-level taxonomic groups such as genera, families, orders and classes. [8] The relative proportions of microbial species vary extensively between individuals [9] and is age-dependent. [10, 11] The microbiome composition evolves rather quickly during the first three years of life, followed by a more gradual maturation. [12] After that, it remains relatively constant throughout adult life. [13]

A variety of environmental factors such as diet, lifestyle, diseases and medications can induce substantial shifts in the microbiome composition. [14, 15] Multiple studies have shown that diet is the main force influencing gut microbial diversity. [16, 17, 18, 19, 20, 21, 22, 23] Yet, diet only explains a small percentage of the microbiome variation observed in human populations. Host genetics is also supposed to play a role in determining the relative abundance of specific gut microbes. [24, 25] Several groups searched for associations between human genetic variation and gut microbiome diversity [21, 22, 23, 26, 27, 28, 29], but only few genetic loci replicated among all these studies. Thus, most of the interindividual variability in gut microbiome composition is still unexplained.

Leveraging on the in-depth phenotypic information available in the *Milieu intérieur* (MI) cohort, a population-based cohort of 1000 healthy individuals of Western European ancestry, we investigate the role of demographic and environmental factors in inter-individual gut microbiome variation. We also evaluate the potential impact of human genetic variants using a genome-wide association study (GWAS) framework, including as covariates the demographic and environmental factors that were found to be correlated with various measures of gut microbiome diversity.

4.3 Results

4.3.1 Gut microbiome diversity

To characterize the bacterial diversity of the gut flora of the 1'000 healthy donors, we used 16s rRNA gene sequencing. We obtained profiles for 862 individuals, with a sequencing depth ranging from 5,064 to 240,472 reads per sample (mean ± SD: 21,363 ± 19,087 reads). A total of 8,422 operational taxonomy units (OTUs) were detected, which correspond to 11 phyla, 24 classes, 43 orders, 103 families, 328 genera and 698 species. On average we detected 193 species per microbiome (standard error 1.9, standard deviation 55.1) with minimum of 58 and a maximum of 346 species. Inter-individual variability was pronounced already at the phylum level. Fig. 1A presents the relative abundances of the 8 phyla observed in more than 10% of study participants. *Firmicutes* and *Proteobacteria* were detected in all individuals, and *Bacteroidetes* in all but one individual. *Firmicutes* was the dominant phylum in the vast majority of individuals (91.8%).

Starting from the OTU counts, we calculated α and β microbiome diversity metrics (see Methods):

- α-diversity measures diversity within each sample. We calculated observed richness (number of distinct species present in the given sample), Chao1 richness estimate (estimate of the number of unobserved species), ACE (Abundance-based Coverage Estimator) and Simpson's diversity index (corresponding to the probability that two randomly picked sequences belong to the same species). The histograms of their distributions are shown in Additional File 1: Figure S1A. Throughout the study we used Simpson's diversity index as a representative metric of α-diversity. The results for other metrics are given in the supplementary material.
- β-diversity measures the difference in taxonomic composition between samples. We calculated compositional Jaccard (unweighted), Bray-Curtis (weighed) and Unifrac (weighted) dissimilarity matrices. We used Bray-Curtis dissimilarity matrix as a

representative metric of β -diversity. The results for other indexes are given in the supplementary material. Fig. 1B presents multidimensional scaling (MDS) plots of Bray-Curtis dissimilarity matrices by age (left panel) and sex (right panel), indicating an absence of stratification. Similar homogeneous distribution of other dissimilarity metrics on the MDS plot is available in Additional File 1: Figure S2.

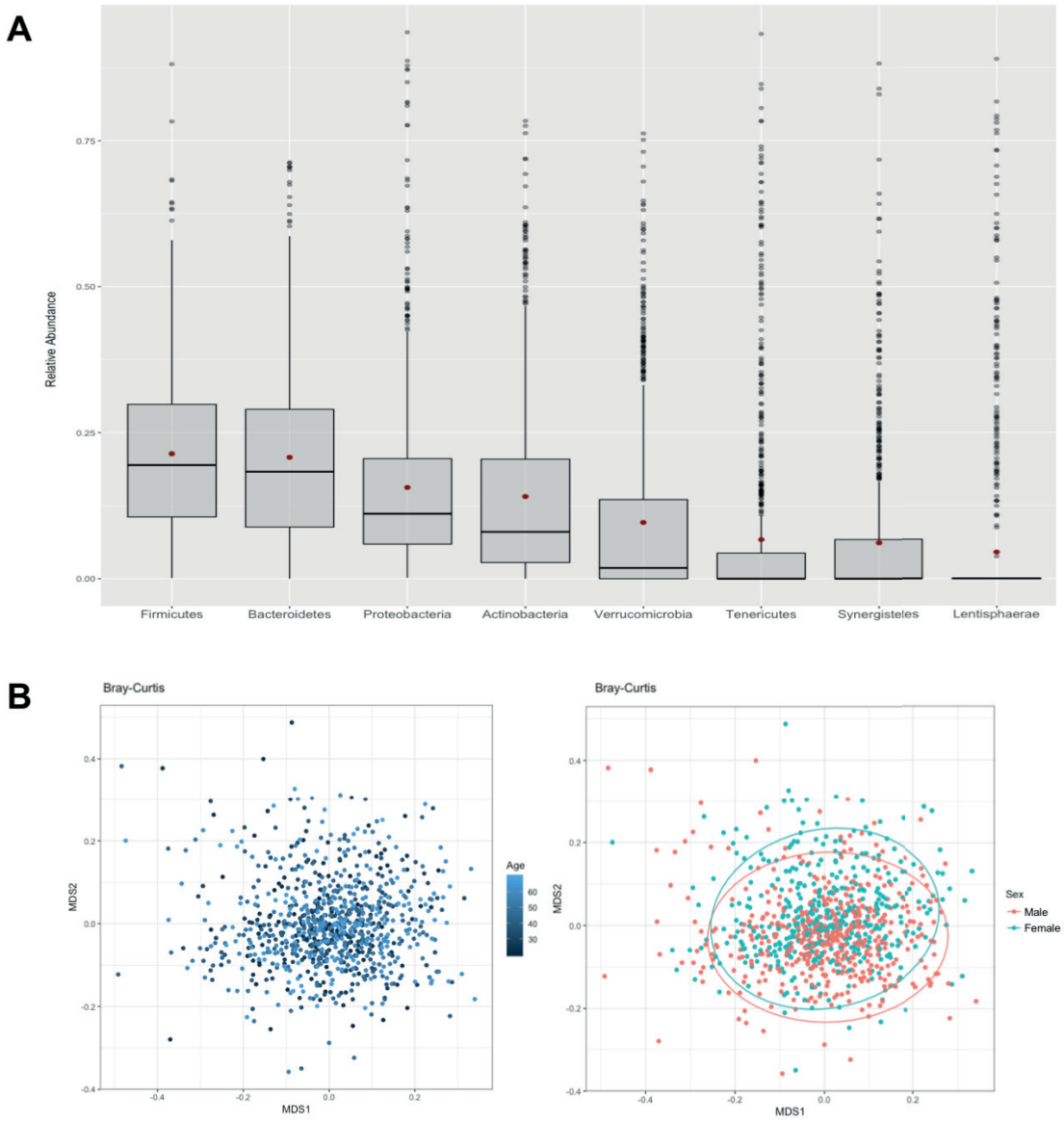

Figure 4.1. Gut microbiome diversity. (A) Box-plots of relative abundances of 8 phyla that were observed in more than 10% of the donors. Dots represent the means and outliers are also represented. (B) Multidimensional scaling plot of Bray-Curtis dissimilarity matrix with donors colored according to age (left panel) and sex (right panel). The curves present 95% normal confidence ellipses.

4.3.2 Selection of demographic, environmental and clinical variables

Demographic, social, behavioral and nutritional information was collected via a detailed questionnaire, while multiple biochemical parameters were measured in blood samples.

Correlations between dietary consumption parameters and gut microbiome have already been investigated in the MI cohort [30]. We considered an additional 274 variables and filtered them based on prevalence, missingness and collinearity, resulting in a final number of 110 variables to be included in association analyses (see Methods). They are described in Additional File 2: Table S1.

4.3.3 Association of non-genetic variables with gut microbiome

To investigate the potential impact of demographic and environmental factors on the gut microbiome, we looked for associations of diversity metrics and individual taxa with the 110 non-genetic variables selected above.

We used Spearman rank correlation testing with four different α-diversity metrics (Additional File 2: Table S2). Five covariates were significant (FDR < 0.05) in univariable tests for all α-diversity metrics and are presented in Table 1. We then tested them in multivariable models also including four dietary variables (which were previously found to be significantly associated with α-diversity [30]) and ran ANOVAs. Only age and levels of alanine aminotransferase remained significant in the multivariable model for all four α-diversity metrics (Table 1 and Additional File 2: Table S3).

Table 4.1. Significant association of non-genetic variables with Simpson's diversity index.

	Univariable Model (Spearman p-value)	Multivarible Model (ANOVA p-value)
Age	1.5×10^{-6}	2×10^{-7}
Level of ALAT	1.2×10^{-3}	4.7×10^{-2}
Glomerular filtration rate	8.5×10^{-3}	1.9×10^{-1}
Raw fruits	1×10^{-3}	2.6×10^{-1}
Fish	4×10^{-3}	3×10^{-1}
Fatty sweet products	1×10^{-3}	7.6×10^{-1}
Sodas / Sugary drinks	1×10^{-3}	8.4×10^{-1}
Having breakfast	1.3×10^{-2}	9.5×10^{-1}
Eating in fast-food restaurants	6.6×10^{-3}	9.8×10^{-1}

We then investigated the impact of non-genetic variables on the β-diversity indexes, running PERMANOVA tests for 110 variables. The results of the tests are present in Additional File 2: Table S4. 15 covariates were significant (FDR < 0.05) by univariable testing for all three indexes. We then tested them in multivariable models, also including raw fruit consumption (which was previously found to be significantly associated with β -diversity [30]) and reran

PERMANOVAs. A total of 10 variables were significant in the final models (Table 2 and Additional File 2: Table S5).

Table 4.2. Significant association of non-genetic variables with Bray-Curtis diversity metrics.

	Univariable Model	Multivarible Model
Age	1.95×10^{-2}	9.99×10^{-4}
Level of ALAT	4.81×10^{-2}	9.99×10^{-4}
Sex	1.95×10^{-2}	9.99×10^{-4}
Chicken pox vaccination	1.95×10^{-2}	2×10^{-3}
Having breakfast	1.95×10^{-2}	3×10^{-3}
Eats lunch	4.92×10^{-2}	1.6×10^{-2}
Diastolic blood pressure	2.6×10^{-2}	1.7×10^{-2}
Little or too much appetite	2.7×10^{-2}	2.1×10^{-2}
Raw fruits	5×10^{-3}	2.2×10^{-2}
Teeth extraction	4.92×10^{-2}	4.4×10^{-2}
Level of HDL	2.7×10^{-2}	5.79×10^{-2}
Cannabis use	2.7×10^{-2}	1.35×10^{-1}
Feeling tired	2.92×10^{-2}	2.58×10^{-1}
House owner	2.92×10^{-2}	2.71×10^{-1}
Eating in fast-food restaurants	2.6×10^{-2}	2.86×10^{-1}
Temperature	2.6×10^{-2}	6.23×10^{-1}

We then calculated the cumulative explained variance of Bray-Curtis dissimilarity by using all the non-genetic variables available. We observed that 16.4% of the variance can be explained by non-genetic factors (Additional File 2: Table S6).

Next, we searched for associations between demographic and environmental variables and individual taxa. We used multivariate association with linear models to test 110 variables and 475 taxa that were observed in more than 10% of donors (see Methods). The full list of tested taxa is available in Additional File 2: Table S7. Table 3 shows the only three significant associations with FDR threshold of 0.05.

Table 4.3. Significant associations of non-genetic variables with individual taxa.

Covariate	**Taxa**	**Coefficient**	**Sample size**	**P-value**	**Q-value**
Age	*Comamonadaceae*	3.99×10^{-4}	317	3.09×10^{-9}	5.89×10^{-5}
Age	*Schlegelella*	3.32×10^{-4}	255	5.48×10^{-6}	3×10^{-2}
Consumption of minerals	*Clostridium papyrosolvens*	2.44×10^{-2}	119	8.32×10^{-7}	4.72×10^{-3}

4.3.4 Association of host genetics with gut microbiome

We used a GWAS framework to search for potential associations between human genetic polymorphisms and gut microbiome diversity. We included in the regression models all the statistically significant covariates for each phenotype. The full list of all the covariates used, including the first two principal components of the genotyping matrix, is available in Additional File 2: Table S8.

We ran GWAS with four α-diversity metrics and three β-diversity indexes. We did not observe any statistically significant signal (Figure 2A, Additional File 1: Figure S3, and Figure 2B and Additional File 1: Figure S4). The quantile-quantile plots and lambda values for all genome-wide analyses are shown in Additional File 1: Figure S5 and Figure S6. We then focused on SNPs that were previously reported to be associated with β-diversity [20, 21, 22]. Upon correction for the 66 SNPs considered (0.05/66), none was significant (Additional File 2: Table S9).

We also used a GWAS approach to test individual taxa for association with host genetic variation. We used both a quantitative phenotype (non-zero log transformed relative abundance) and a binary phenotype (presence vs. absence) for all taxa (see Methods). After correction for the number of phenotypes tested, we did not observe any statistically significant signal. We also imputed HLA and KIR alleles and tested them for association with the phenotypes, observing no significant associations. Using a less stringent threshold for multiple testing correction ($P_{threshold} < 5 \times 10^{-8}$), a total of 170 SNPs were associated with the quantitative phenotype of 53 taxa, and 65 SNPs were associated with the binary phenotype of 23 taxa. The lists of these SNPs and their association p-values are available in Additional File 2: Table S10 and Additional File 2: Table S11, respectively.

We then considered all the SNPs previously reported to be associated with individual taxa (Additional File 2: Table S14) [20, 21, 22, 23, 26, 28]. Only 13 out of 336 SNPs passed the corrected nominal significance threshold (0.05/336) for association with a quantitative phenotype. Of these, 9 were concordant at the phylum level with the original report. For binary phenotypes, 10 SNPs passed the corrected nominal significance threshold, including 2 that were concordant at the phylum level.

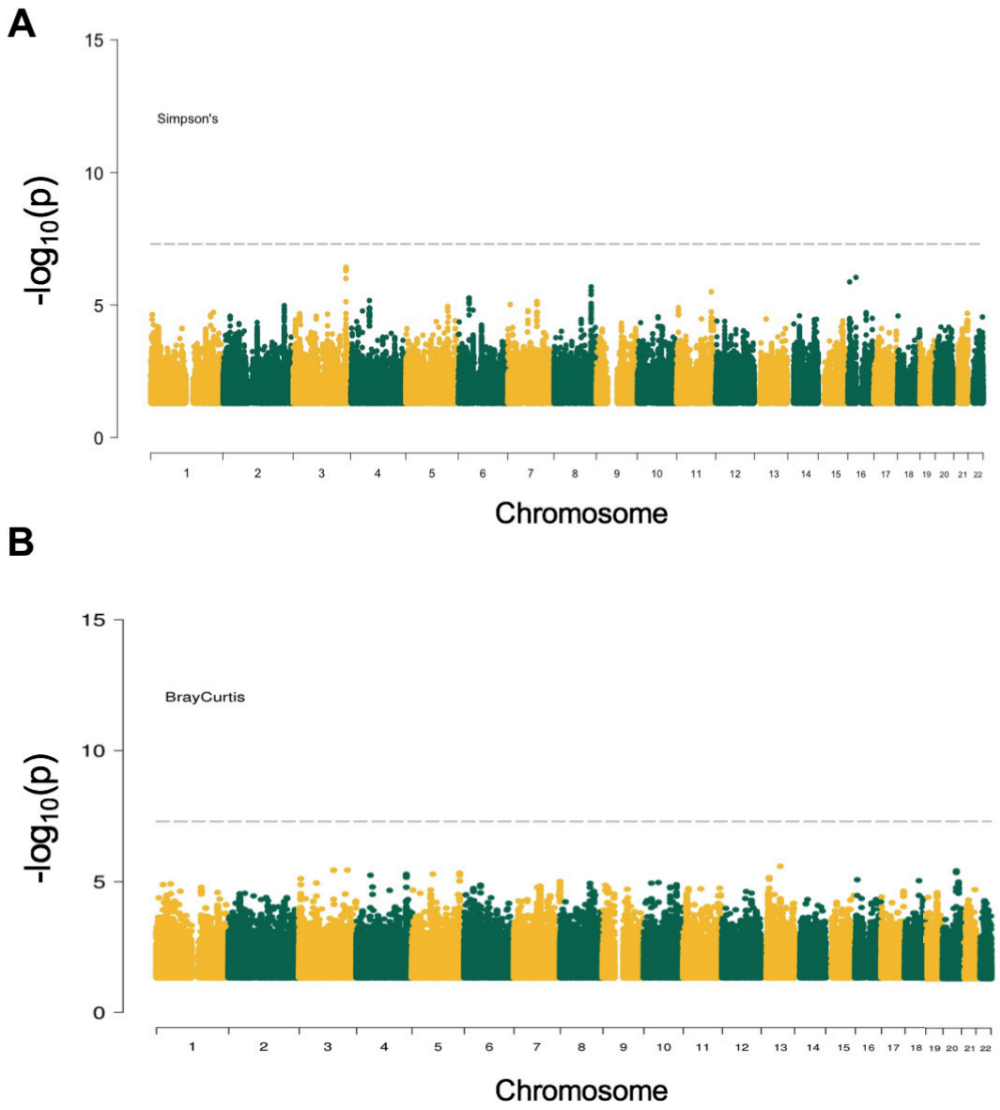

Figure 4.2. Results of genome-wide association study between host genetic variants and microbiome diversity metrics. (A) Manhattan plot for Simpson's diversity metric (representative α-diversity metric). The dashed horizontal line denotes the genome-wide significance threshold ($P_{\alpha\text{-threshold}} < 1.25 \times 10^{-8}$). (B) Manhattan plot for Bray-Curtis dissimilarity matrix (representative ß-diversity index). The dashed horizontal line denotes the genome-wide significance threshold ($P_{\theta\text{-threshold}} < 1.67 \times 10^{-8}$).

4.4 Discussion

We investigated the potential influence of demographic, environmental, clinical and genetic factors on the gut microbiome composition in 858 unrelated healthy individuals of French descent. The *Milieu intérieur* cohort is particularly well suited for such a comprehensive assessment. The study participants have a homogeneous genetic background and are evenly stratified by sex and age, which provides an excellent opportunity to search for unique determinants of gut microbiome diversity.

First, we used the rich data collected through questionnaires that collected detailed medical histories as well as lifestyle and socio-demographic information. We also considered laboratory results that could indicate underlying physiological differences (e.g. levels of hemoglobin, glucose, hepatic transaminases, etc.). We searched for potential association of these variables with several α- and β- diversity metrics of the gut microbiome, as well as with quantitative and binary phenotypes derived from the detected abundance of individual microbial taxa.

As the MI cohort was designed to better understand healthy immunity, strict criteria were used during enrolment to exclude individuals with chronic medical conditions. Therefore, the use of prescription medication was very limited among study participants. In fact, out of the final 110 non-genetic variables that were analyzed, only one concerned drug treatment. Surprisingly, even the use of over-the-counter drugs, such as proton pump inhibitors, was observed in less than 10% of the individuals and could thus not be evaluated in our study. As a consequence, we made no attempt at replicating the previously reported associations between drug intake and gut microbiome composition [17, 18, 19].

Since the influence of dietary variables on the gut microbiome has already been evaluated in the MI cohort [30], we focused our attention on other environmental influences, lifestyle variables and biochemical measurements. Age strongly and consistently associated with α-diversity metrics in all models, whereas sex and BMI did not show any significant association. A more surprising finding was the correlation between higher levels of alanine aminotransferase and lower gut microbiome diversity. The directionality of the observed correlation is unclear. Indeed, much work is still needed to get a better understanding of the interplay between the microbiome and liver disease [31]. In the analysis of β-diversity indexes, we identified six additional variables that were significant in the multivariable PERMANOVA models. An estimation of the explained variance in β-diversity metrics by all associated variables demonstrated a small individual effect for each variable (Additional File 2: Table S4), which together explained 16.4% of the variance. This is concordant with previous reports [17, 18, 19, 20, 22].

We then studied the impact of the same set of non-genetic variables on individual taxa. We observed a strong correlation of age with the *Comamonadaceae* family and with the genus *Schlegelella*, which is part of the same family. We also found an association between *Clostridium papyrosolvens*, belonging to the *Clostridia* class and *Firmicutes* phylum, and the oral intake of mineral supplements. *Clostridium papyrosolvens* is an anaerobic bacterium that is involved in the degradation of diverse carbohydrates (such as cellulose, arabinose and glucose) [32] and could thus play a role in modulating the individual glycemic response.

Our in-depth investigation of demographic, environmental and clinical variables allowed us to identify subsets of factors that are highly associated with various measures of gut microbiome composition. Including them as covariates in genome-wide association studies had the potential to increase our power to detect true genetic effects. However, after necessary correction for multiple testing, we did not observe any statistically significant association. This was the case for a total of 7 different α- and β- diversity metrics and for 475 individual taxa, tested either as quantitative and, for 375 of them as well as, a binary phenotype.

Lastly, we checked for replication in the MI cohort of the SNPs previously reported to be associated with the gut microbiome composition at the β-diversity or the taxonomic levels [20, 21, 22, 23, 26, 28]. None of the variants associated with β-diversity metrics replicated. For individual taxa, replication at the phylum level occurred for 2 SNPs for binary phenotypes (presence vs. absence of the phylum) and 9 SNPs for quantitative phenotypes (abundance). Of these, only one SNP (rs7856187) replicated at the family level – *Lachnospiraceae* [28]. Of note, the only SNP that was significant in a recent meta-analysis [20], rs4988235, did not show any association in our study (Additional File 2: Table S12).

4.5 Conclusions

Our study provides a detailed investigation of potential demographic, environmental, clinical and genetic influences on the diversity of the gut microbiome in healthy individuals. We identified variables associated with overall microbiome composition and with a small number of individual taxa. The absence of any significant results in the genome-wide association analyses, on the other hand, indicates that common human genetic variants are unlikely to play a major role in shaping the gut microbiome diversity observed in healthy populations. Future studies should include larger sample sizes and a more comprehensive evaluation of human genetic variation (also including rare and structural variants not captured by genotyping arrays). Data should also be pooled across cohorts, as recently proposed [33], to accelerate discovery in the field of host-microbiome interactions.

4.6 Methods

4.6.1 The *Milieu Intérieur* cohort

The 1,000 healthy donors of the *Milieu Intérieur* cohort were recruited by BioTrial (Rennes, France). The cohort is stratified by sex (500 men, 500 women) and age (200 individuals from each decade of life, between 20 and 70 years of age). Participants were selected based on stringent inclusion and exclusion criteria, detailed elsewhere [34]. Briefly, they had no evidence of any severe/chronic/recurrent medical conditions. The main exclusion criteria were seropositivity for human immunodeficiency virus or hepatitis C virus; travel to (sub-) tropical countries within the previous 6 months; recent vaccine administration; and alcohol abuse. Subjects were excluded if they were at the time on, or were treated in the three months preceding enrolment with, nasal, intestinal or respiratory antibiotics or antiseptics. Volunteers following a specific diet prescribed by a doctor or dietician for medical reasons (calorie-controlled diet or diet favouring weight loss in very overweight patients, diets to decrease cholesterol levels) and volunteers with food intolerance or allergy were also excluded. To avoid the influence of hormonal fluctuations in women during the peri-menopausal phase, only pre- or post-menopausal women were included. To minimize the importance of population substructure on genomic analyses, the study was restricted to individuals of self-reported Metropolitan French origin for three generations (i.e., with parents and grandparents born in continental France). Fasting whole blood samples were collected from the 1000 participants on lithium heparin tubes between September 2012 and August 2013.

4.6.2 Fecal DNA extraction and amplicon sequencing

Human stool samples were produced at home no more than 24 hours before the scheduled medical visit and collected in a double-lined sealable bag maintaining strict anaerobic conditions. Upon reception at the clinical site, the fresh stool samples were aliquoted and stored immediately at -80°C. DNA was extracted from stool as previously published [35, 36]. DNA quantity was measured with Qubit using broad range assay. Barcoding polymerase chain reaction (PCR) was carried out using indexed primers targeting the V3-V5 region of the 16S rRNA gene as described in [37]. AccuPrime™ Pfx SuperMix (Invitrogen - 12344-040) was used to perform the PCR. PCR mix was made up of 18 µL of AccuPrime™ Pfx SuperMix, 0.5 µL of both V3-340F and V5-926R primers (0.2 µM) and 1 µL of DNA (10 ng). PCR was carried out as follow: 95°C for 2 min, 30 cycles of 95°C for 20 sec, 55°C for 15 sec, 72°C for 5 min and a final step at 72°C for 10 min. Amplicon concentration was then normalized to 25 ng per PCR reaction using SequalPrep™ Normalization Plate Kit, 96-well (Thermo Fisher Scientific). Equal volumes of normalized PCR reaction were pooled and thoroughly mixed. The amplicon libraries were sequenced at the Institut Curie NGS platform on Illumina MiSeq using the 2*300 base pair V3 kit.

4.6.3 16s sequencing data processing and identification of microbial taxa

Raw reads were trimmed using sickle [38], then error corrected using SPAdes [39] and merged using PEAR [40]. Reads were clustered into operational taxonomy units (OTUs) at 97% of identity using vsearch pipeline [41]. Chimeric OTUs were identified using UCHIME [42] and discarded from downstream analysis. Taxonomy of representative OTU sequences was determined using RDP classifier [43]. OTU sequences were aligned using ssu-align [44]. The phylogenetic tree was inferred from the OTUs multiple alignments using Fastree2 [45]. For 138 individuals, the gut microbiome composition could not be established because of technical issues in exploitability of sequencing results. These were excluded from further analysis.

4.6.4 Gut microbiome diversity estimates

We calculated two types of microbial diversity indicators: α- and β-diversity indices. As estimates of α-diversity, we used Simpson's diversity index, observed richness, Chao1 richness estimate and ACE (Abundance-based Coverage Estimator). We applied Yeo-Johnson transformation with R package VGAM [46] to normalize these phenotypes. The histograms of raw and transformed distributions are shown in Additional File 1: Figure S2. As estimates of β-diversity, we used Bray-Curtis (weighed), compositional Jaccard (unweighted) and Unifrac (weighted) dissimilarity matrices. All diversity indicators were generated on non-rarefied data using the R package vegan [47].

4.6.5 Demographic, environmental and clinical variables

A large number of demographical, environmental and clinical variables are available in the *Milieu Intérieur* cohort [34]. These notably include infection and vaccination history, childhood diseases, health- and diet- related habits, socio-demographical variables, and laboratory measurements. After manual curation, we considered 274 variables as potentially interesting for our analyses. Of those, we removed 130 that: (i) were only variable in less than

5% of participants; or (ii) were missing in more than 10% of participants. We tested for collinearity among the remaining 144 variables using Spearman rank correlation. All pairwise correlations with a Spearman's ρ > 0.6 or < -0.6 and a false discovery rate (FDR) < 5% were considered colinear; one variable from each pair was randomly removed from further analysis, resulting in a final set of 110 variables (described in Additional File 2: Table S1). Of these, 39 had some missing values (<1% in 25, 1-5% in 10, 5-10% in 4 individuals), which were imputed using random forest method in the R package mice [48].

4.6.6 Testing of demographic, environmental and clinical variables

We searched for associations between the 110 demographic, environmental and clinical variables selected above and the various gut microbiome phenotypes. For α-diversity indexes (Simpson's index, observed richness, Chao1 richness estimate and ACE), we used non-parametric Spearman correlations. For β-diversity dissimilarities (Jaccard, Bray-Curtis and Unifrac matrices), we used permutational analysis of variance (PERMANOVA) with 1000 permutations. PERMANOVAs identify variables that are significantly associated with β-diversity and measure the fraction of variance explained by the factors tested. The variables that were significantly associated (FDR < 0.05) with the diversity estimates were included in the respective multivariable models: we used multivariable ANOVAs for α-diversity and PERMANOVAs for β-diversity. After eliminating the variables that were not significant in the first multivariable model, we reran the tests iteratively until all included predictors were significant. Spearman correlations, ANOVA and PERMANOVAs tests were performed in R v3.5.1. Finally, to search for associations with individual taxa, we implemented multivariate association with linear models by using MaAsLin [49] with default parameters.

4.6.7 Human DNA genotyping

As previously described [50], blood was collected in 5mL sodium EDTA tubes and kept at room temperature (18–25°) until processing. After extraction, DNA was genotyped at 719,665 single nucleotide polymorphisms (SNPs) using the HumanOmniExpress-24 BeadChip (Illumina). The SNP call rate was > 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 variants using the HumanExome-12 BeadChip. The variant call rate was < 97% in 11 donors, which were thus removed from this dataset. We filtered out from both datasets genetic variants based on a set of criteria detailed in [51]. These quality-control filters yielded a total of 661,332 and 87,960 variants for the HumanOmniExpress and HumanExome BeadChips, respectively. Average concordance rate for the 16,753 SNPs shared between the two genotyping platforms was 99.99%, and individual concordance rates ranged from 99.8% to 100%.

4.6.8 Genetic relatedness and structure

Relatedness was detected using KING [52]. Six pairs of related participants (parent-child, first and second-degree siblings) were identifed. Of those, four pairs had both genotyping and microbiome datasets and one individual from each pair, randomly selected, was removed from the genetic analyses, leaving in total 858 individuals with both genotyping and 16s rRNA gene sequencing data. The genetic structure of the study population was

estimated using principal component analysis (PCA), implemented in EIGENSTRAT (v6.1.3) [53]. The PCA plot of the study population is shown in Additional File 1: Figure S7.

4.6.9 Genotype imputation

As described previously [51], we used Positional Burrows-Wheeler Transform for genotype imputation, starting with the 661,332 quality-controlled SNPs genotyped on the HumanOmniExpress array. Phasing was performed using EAGLE2 (v2.0.5) [54]. As reference panel, we used the haplotypes from the Haplotype Reference Consortium (release 1.1) [55]. After removing SNPs that had an imputation info score < 0.8, we obtained 22,235,661 variants. We then merged the imputed dataset with 87,960 variants directly genotyped on the HumanExome BeadChips array and removed variants that were monomorphic or diverged significantly from Hardy-Weinberg equilibrium ($P < 10^{-7}$). We obtained a total of 12,058,650 genetic variants to be used in association analyses.

We used SNP2HLA (v1.03) [56] to impute 104 4-digit human leukocyte antigen (HLA) alleles and 738 amino acid residues (at 315 variable amino acid positions of the HLA class I and II proteins) with a minor allele frequency (MAF) of >1%.

We used KIR*IMP [57] to impute killer cell immunoglobulin like receptors (KIR) alleles, after haplotype inference on chromosome 19 with SHAPEIT2 (v2.r790) [58]. A total of 19 KIR types were imputed: 17 loci plus two extended haplotype classifications (A vs. B and KIR haplotype). A MAF threshold of 1% was applied, leaving 16 KIR alleles for association analysis.

4.6.10 Genetic association analyses

For single variant association analyses, we only considered SNPs with a MAF of >5% (N=5,293,637). Unless otherwise stated, we used PLINK (v1.9) [59] for association testing. In all test, we included the first two first principal components of the genotyping matrix as covariates to correct for residual population stratification. The demographic, environmental and clinical variables that were identified as significantly associated were also included as covariates in the respective analyses. A full list of covariates for each phenotype is available in Additional File 2: Table S8.

We used linear regression (within PLINK) and microbiomeGWAS [60] to test for SNP associations with α-diversity indexes and β-diversity dissimilarities, respectively. Linear regression was also used to search for associations with relative abundance of specific taxa. Only taxa present in at least 10% of individuals were tested (N=475), i.e. 8/11 (remaining/total) phyla, 16/24 classes, 20/43 orders, 50/103 families, 135/328 genera and 246/698 species. The list of all tested taxa is presented in Additional File 2: Table S7. We used logistic regression to test binary phenotypes (presence/absence of specific taxa). Here, we excluded taxa that were present in >90% of individuals, resulting in a total of 374 phenotypes (4 phyla, 8 classes, 15 orders, 38 families, 104 genera and 205 species). For all GWAS, we used a significance threshold corrected for the number of tests performed. For α-diversity (N=4): $P_{\alpha\text{-threshold}} < 1.25 \times 10^{-8}$, for β-diversity (N=3): $P_{\beta\text{-threshold}} < 1.67 \times 10^{-8}$, for taxa abundance (N=475): $P_{taxa\text{-}linear} < 1.05 \times 10^{-10}$ and for presence or absence of taxa (N=374): $P_{taxa\text{-}logistic} < 1.33 \times 10^{-10}$.

4.7 List of Abbreviations

SNP: single nucleotide polymorphism; MAF: minor allele frequency; MI: *Milieu Intérieur*; QQ: quantile-quantile; LD: linkage disequilibrium; PCR: polymerase chain reaction; ANOVA: analysis of variance; PERMANOVA: permutational analysis of variance; FDR: false discovery rate; OTU: operational taxonomy unit; HIV: human immunodeficiency virus; HCV: hepatitis C virus; ACE: Abundance-based coverage estimator; GWAS: genome-wide association study; HLA: human leukocyte antigen; KIR: killer cell immunoglobulin like receptors; PCA: principal component analysis; MDS: Multidimensional scaling.

4.8 Declarations

4.8.1 Ethics approval and consent to participate

The clinical study was approved by the Comité de Protection des Personnes - Ouest 6 on June 13th, 2012, and by the French Agence Nationale de Sécurité du Médicament on June 22nd, 2012 and has been performed in accordance with the Declaration of Helsinki. 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). Informed consent was obtained from participants after the nature and possible consequences of the studies were explained.

4.8.2 Availability of data and material

Genotype data supporting the conclusions of this article are available in the European Genome-Phenome Archive under the accession code EGAS00001002460. Full summary association results are available for download from Zenodo (http://doi.org/10.5281/zenodo.1438474).

4.8.3 Competing interests

C.H. and M.L.A. are employees of Genentech Inc., a member of The Roche Group. The remaining authors declare that they have no competing interests.

4.8.4 Funding

This work benefited from support of the French government's Invest in the Future Program, managed by the Agence Nationale de la Recherche (ANR, reference 10-LABX-69-01). It was also supported by a grant from the Swiss National Science Foundation (31003A_175603, to JF).

4.8.5 Acknowledgements

We would like to thank to all the donors for their contribution to the study.

The *Milieu Intérieur* Consortium is composed of the following team leaders: Laurent Abel (Hôpital Necker, Paris, France), Andres Alcover (Institut Pasteur, Paris, France), Hugues Aschard (Institut Pasteur, Paris, France), Kalla Astrom (Lund University, Lund, Sweden), Philippe Bousso (Institut Pasteur, Paris, France), Pierre Bruhns (Institut Pasteur, Paris, France), Ana Cumano (Institut Pasteur, Paris, France), Caroline Demangel (Institut Pasteur, Paris, France), Ludovic Deriano (Institut Pasteur, Paris, France), James Di Santo (Institut Pasteur, Paris, France), Françoise Dromer (Institut Pasteur, Paris, France), Darragh Duffy (Institut Pasteur, Paris, France), Gérard Eberl (Institut Pasteur, Paris, France), Jost Enninga (Institut Pasteur, Paris, France), Jacques Fellay (EPFL, Lausanne, Switzerland) Odile Gelpi (Institut Pasteur, Paris, France), Ivo Gomperts-Boneca (Institut Pasteur, Paris, France), Milena Hasan (Institut Pasteur, Paris, France), Serge Hercberg (Université Paris 13, Paris, France), Olivier Lantz (Institut Curie, Paris, France), Claude Leclerc (Institut Pasteur, Paris, France), Hugo Mouquet (Institut Pasteur, Paris, France), Sandra Pellegrini (Institut Pasteur, Paris, France), Stanislas Pol (Hôpital Côchin, Paris, France), Antonio Rausell (INSERM UMR 1163 – Institut Imagine, Paris, France), Lars Rogge (Institut Pasteur, Paris, France), Anavaj Sakuntabhai (Institut Pasteur, Paris, France), Olivier Schwartz (Institut Pasteur, Paris, France), Benno Schwikowski (Institut Pasteur, Paris, France), Spencer Shorte (Institut Pasteur, Paris, France), Vassili Soumelis (Institut Curie, Paris, France), Frédéric Tangy (Institut Pasteur, Paris, France), Eric Tartour (Hôpital Européen George Pompidou, Paris, France), Antoine Toubert (Hôpital Saint-Louis, Paris, France), Mathilde Touvier (Université Paris 13, Paris, France), Marie-Noëlle Ungeheuer (Institut Pasteur, Paris, France), Matthew L. Albert (Roche Genentech, South San Francisco, CA, USA), Lluis Quintana-Murci (Institut Pasteur, Paris, France). Matthew L. Albert and Lluis Quintana-Murci are co-coordinators of the consortium. Additional information can be found at: http://www.milieuinterieur.fr/en.

4.8.6 Additional Files

Additional files are available at the following link: http://doi.org/10.5281/zenodo.1443336.

Additional File 1: (DOCX 3.4 MB)
Figure S1. Raw and transformed distributions of α-diversity phenotypes.
Figure S2. Multidimensional scaling plots of Jaccard and Unifrac distance matrices.
Figure S3. Manhattan plots for α-diversity metrics: richness, Chao1 and ACE.
Figure S4. Manhattan plots for ß-diversity matrices: Jaccard and Unifrac.
Figure S5. QQ plots and lambda values of GWAS of α-diversity phenotypes.
Figure S6. QQ plots and lambda values of GWAS of β-diversity indices.
Figure S7. PCA plot of the genetic matrix data of MI donors.

Additional File 2: (XLSX 631 KB)
Table S1. Description of all the covariates used in the study.
Table S2. Spearman correlations of all the covariates with four α-diversity metrics.
Table S3. Results of multivariable ANOVAs with three other α-diversity metrics.
Table S4. PERMANOVA results for all of the covariates with three β-diversity indices.
Table S5. Results of multivariable PERMANOVAs of two others β-diversity indices.
Table S6. Explained cumulative variance of Bray-Curtis dissimilarity metric by all non-genetic covariates (dietary and 110 tested in this study).
Table S7. List of taxa tested for association with genetic variants.

Table S8. List of identified covariates that were used for each phenotype in addition to the first two principal components of the genotyping matrix.
Table S9. Replication of the SNPs previously reported to be associated with β-diversity.
Table S10. Nominal associations of SNPs with relative abundances of taxa.
Table S11. Nominal associations of SNPs with dichotomized taxa in the MI cohort.
Table S12. Replication of the SNPs previously reported to be associated with individual taxa.

4.9 References

1. Sekirov I, Russell SL, Antunes LC, Finlay BB. Gut microbiota in health and disease. Physiol Rev. 2010; 90:859-904.

2. Honda K, Littman DR. The microbiome in infectious disease and inflammation. Annu Rev Immunol. 2012; 30:759-95.

3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J; MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59-65.

4. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, Sogin ML. The Microbiome and Human Biology. Annu Rev Genomics Hum Genet. 2017; 18:65-86.

5. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016; 8:51.

6. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207-14.

7. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science. 2006; 312:1355-9.

8. Wexler AG, Goodman AL. An insider's perspective: Bacteroides as a window into the microbiome. Nat Microbiol. 2017; 2:17026.

9. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017; 550:61-66.

10. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao JZ, Abe F, Osawa R. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. BMC Microbiol. 2016; 16:90.

11. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vázquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N. Temporal variability is a personalized feature of the human microbiome. Genome Biol. 2014; 15:531.

12. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT, Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Wang J. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe. 2015; 17:690-703.

13. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, Rosenbaum M, Gordon JI. The long-term stability of the human gut microbiota. Science. 2013; 341:1237439.

14. Schroeder BO, Bäckhed F. Signals from the gut microbiota to distant organs in physiology and disease. Nat Med. 2016; 22:1079-1089.

15. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. Human gut microbiome viewed across age and geography. Nature. 2012; 486:222-7.

16. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, Suez J, Mahdi JA, Matot E, Malka G, Kosower N, Rein M, Zilberman-Schapira G, Dohnalová L, Pevsner-Fischer M, Bikovsky R, Halpern Z, Elinav E, Segal E. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015; 163:1079-1094.

17. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De SL, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. Population-level analysis of gut microbiome variation. Science 2016; 352:560-4.

18. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, Wang J, Imhann F, Brandsma E, Jankipersadsing SA, Joossens M, Cenit MC, Deelen P, Swertz MA, Weersma RK, Feskens EJ, Netea MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 2016; 352:565-9.

19. Jackson MA, Verdi S, Maxan ME, Shin CM, Zierer J, Bowyer RCE, Martin T, Williams FMK, Menni C, Bell JT, Spector TD, Steves CJ. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. Nat Commun. 2018; 9:2655.

20. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N, Shilo S, Lador D, Vila AV, Zmora N, Pevsner-Fischer M, Israeli D, Kosower N, Malka G, Wolf BC, Avnit-Sagi T, Lotan-Pompan M, Weinberger A, Halpern Z, Carmi S, Fu J, Wijmenga C, Zhernakova A, Elinav E, Segal E. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018; 555:210-215.

21. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe. 2016; 19:731-43.

22. Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen FA, Rühlemann MC, Szymczak S, Holm K, Esko T, Sun J, Pricop-Jeckstadt M, Al-Dury S, Bohov P, Bethune J, Sommer F, Ellinghaus D, Berge RK, Hübenthal M, Koch M, Schwarz K, Rimbach G, Hübbe P, Pan WH, Sheibani-Tezerji R, Häsler R, Rosenstiel P, D'Amato M, Cloppenborg-Schmidt K, Künzel S, Laudes M, Marschall HU, Lieb W, Nöthlings U, Karlsen TH, Baines JF, Franke A. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet. 2016; 48:1396-1406.

23. Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, Guttman DS, Griffiths A, Panaccione R, Otley A, Xu L, Shestopaloff K, Moreno-Hagelsieb G; GEM Project Research Consortium, Paterson AD, Croitoru K. Association of host genome with intestinal microbial composition in a large healthy cohort. Nat Genet. 2016; 48:1413-1417.

24. Khachatryan ZA, Ktsoyan ZA, Manukyan GP, Kelly D, Ghazaryan KA, Aminov RI. Predominant Role of Host Genetics in Controlling the Composition of Gut Microbiota. Fraser JA, ed. PLoS ONE. 2008; 3:e3064.

25. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. Human genetics shape the gut microbiome. Cell. 2014; 159:789-99.

26. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 2015; 16:191.

27. Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. Genome-Wide Association Studies of the Human Gut Microbiota. PLoS One. 2015; 10:e0140301.

28. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, Deelen P, Vatanen T, Schirmer M, Smeekens SP, Zhernakova DV, Jankipersadsing SA, Jaeger M,

Oosting M, Cenit MC, Masclee AA, Swertz MA, Li Y, Kumar V, Joosten L, Harmsen H, Weersma RK, Franke L, Hofker MH, Xavier RJ, Jonkers D, Netea MG, Wijmenga C, Fu J, Zhernakova A. The effect of host genetics on the gut microbiome. Nat Genet. 2016; 48:1407-1412.

29. Kolde R, Franzosa EA, Rahnavard G, Hall AB, Vlamakis H, Stevens C, Daly MJ, Xavier RJ, Huttenhower C. Host genetic variation and its microbiome interactions within the Human Microbiome Project. Genome Med. 2018; 10:6.

30. Partula V, Mondot M, Torres MJ, Kesse-Guyot E, Deschasaux M, Assmann K, Latino-Martel P, Buscail C, Julia C, Galan P, Hercberg S, Quintana-Murci L, Albert ML, Duffy D, Lantz O, Touvier M, The Milieu Intérieur Consortium. Associations between usual dietary consumptions and gut microbiota composition in healthy French adults: results from the 1000-individuals *Milieu Intérieur* study. (submitted)

31. Tripathi A, Debelius J, Brenner DA, Karin M, Loomba R, Schnabl B, Knight R. The gut-liver axis and the intersection with the microbiome. Nat Rev Gastroenterol Hepatol. 2018; 15:397-411.

32. Madden RH, Bryder MJ, Poole NJ. Isolation and characterization of an anaerobic, cellulolytic bacterium, Clostridium papyrosolvens sp-nov. Int J Syst Bacteriol. 1982; 32:87–91.

33. Wang J, Kurilshikov A, Radjabzadeh D, Turpin W, Croitoru K, Bonder MJ, Jackson MA, Medina-Gomez C, Frost F, Homuth G, Rühlemann M, Hughes D, Kim HN; MiBioGen Consortium Initiative, Spector TD, Bell JT, Steves CJ, Timpson N, Franke A, Wijmenga C, Meyer K, Kacprowski T, Franke L, Paterson AD, Raes J, Kraaij R, Zhernakova A. Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. Microbiome. 2018; 6:101.

34. Thomas S, Rouilly V, Patin E, Alanio C, Dubois A, Delval C, Marquier LG, Fauchoux N, Sayegrih S, Vray M, Duffy D, Quintana-Murci L, Albert ML. Milieu Intérieur Consortium. The *Milieu Intérieur* study—an integrative approach for study of human immunological variance. Clin Immunol. 2015;157:277–93.

35. Ó Cuív P, Aguirre de Cárcer D, Jones M, Klaassens ES, Worthley DL, Whitehall VL, Kang S, McSweeney CS, Leggett BA, Morrison M. The effects from DNA extraction methods on the evaluation of microbial diversity associated with human colonic tissue. Microb Ecol. 2011; 61:353-62.

36. Yu Z, Morrison M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. Biotechniques. 2004; 36:808-12.

37. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013; 79:5112-20.

38. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. 2011 [citeulike:13260426].

39. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19:455-77.

40. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014; 30:614-20.

41. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016; 4:e2584.

42. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27:2194-200.

43. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009; 37:D141-5.

44. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009; 25:1335-7.

45. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010; 5:e9490.

46. Yee TW. The VGAM Package for Categorical Data Analysis. J Stat Softw. 2010; 32:1-34.

47. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Peter Solymos P, Stevens MHH, Szoecs E, Wagner H. vegan: community ecology package. R package version 2.5-2. 2018. https://cran.r-project.org/web/packages/vegan/index.html.

48. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011; 45:1-67.

49. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012; 13:R79.

50. Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, Alanio C, Scepanovic P, Hammer C, Jönsson F, Beitz B, Quach H, Lim YW, Hunkapiller J, Zepeda M, Green C, Piasecka B, Leloup L, Rogge L, Huetz F, Peguillet I, Lantz O, Fontes M, Di Santo JP, Thomas S, Fellay J, Duffy D, Quintana-Murci L, Albert ML, for The Milieu Intérieur

Consortium. Natural variation in innate immune cell parameters is preferentially driven by genetic factors. Nat Immunol 2018; 19:302–314.

51. Scepanovic P, Alanio C, Hammer C, Hodel F, Bergstedt J, Patin E, Thorball CW, Chaturvedi N, Charbit B, Abel L, Quintana-Murci L, Duffy D, Albert ML, Fellay J; Milieu Intérieur Consortium. Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. Genome Med. 2018; 10:59.

52. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.

53. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.

54. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, L Price A. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016; 48:1443–8.

55. McCarthy S, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48:1279–83.

56. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, Raychaudhuri S, de Bakker PI. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One. 2013;8:e64683.

57. Vukcevic D, Traherne JA, Næss S, Ellinghaus E, Kamatani Y, Dilthey A, Lathrop M, Karlsen TH, Franke A, Moffatt M, Cookson W, Trowsdale J, McVean G, Sawcer S, Leslie S. Imputation of KIR types from SNP variation data. Am J Hum Genet. 2015; 97:593–607.

58. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF, Vitart V, Navarro P, Zagury JF, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014; 10:e1004234.

59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7.

60. Hua X, Song L, Yu G, Goedert JJ, Abnet CC, Landi MT, Shi J. MicrobiomeGWAS: a tool for identifying host genetic variants associated with microbiome composition. bioRxiv. 2015. https://doi.org/10.1101/03118

Chapter 5: Conclusions

Today, it is not possible for a patient to walk into a doctor's clinic and receive a test that faithfully reflects the health of their immune system. This is not for lack of parameters to measure, but because of a more fundamental problem: our very poor understanding of the immense heterogeneity and variability in the immune system. To improve our ability to provide an answer to this rather simple question, novel insights into the factors governing this variability are needed.

The works presented in this thesis focused on discovering the drivers of the inter-individual variance of the healthy immune system in humans. In particular, the potential effects of human genome variation and environmental factors were investigated using various phenotypic outcomes.

In the second chapter of the thesis, the study of factors influencing humoral immunity is described. We examined the humoral responses of 1,000 healthy people to common infections and vaccines. We measured antibody responses to fifteen antigens from twelve infectious agents: cytomegalovirus, Epstein-Barr virus, herpes simplex virus 1 and 2, varicella zoster virus, influenza A virus, measles, mumps, rubella, and hepatitis B virus, *Helicobacter pylori*, and *Toxoplasma gondii*. In order to assess the importance of non-genetic factors, attention was given to numerous demographic variables. Age and sex were identified as the most important determinants of humoral response, with older individuals and women showing stronger antibody responses against most antigens. With regard to genetic factors, we performed genome-wide association studies. The results showed that differences in response to Epstein-Barr virus and rubella associate with variation in the human leucocyte antigen (HLA) gene region. We fine-mapped these signals to amino acids in the extracellular domains of HLA proteins, confirming their essentiality for the presentation of processed peptides to $CD4^+$ T cells and thus reveling important clues in the fine regulation of class II antigen presentation. We also identified novel specific HLA/KIR combinations that support the idea that these interactions can play a critical role in shaping humoral immune responses in humans.

In the third chapter, a large-scale genetic and immunological profiling study is presented. The detailed cellular composition of white blood cells was established through the generation of 166 immune-system phenotypes. The effects of environment on these phenotypes were assessed, and a particularly strong impact of age, sex and cytomegalovirus (CMV) infection on many cell subpopulations was observed. Using a genome-wide association study approach, 15 loci were identified that were associated with immunological diversity within the healthy human population. Innate cells were more strongly controlled by genetic variation than cells involved in the adaptive arm of the immune response, which were more likely to show

variability because of environmental influences. We investigated in detail the associations arising from the HLA gene region and we managed to pinpoint them to coding variations and amino acids within HLA proteins that were the more probable cause of the signal. Many of these signals were reported previously to be associated with human diseases and traits and were directly linked with etiology of several autoimmune disorders. Our findings thus provide new insight into the mechanisms underlying disease pathogenesis.

In the fourth chapter, the investigation of the factors influencing gut microbiome diversity is presented. An unprecedented number of demographic, environmental, clinical and lifestyle factors were evaluated for their potential association with gut microbiome composition. The total variance explained by these factors was estimated to be ~16%. We also conducted genome-wide association studies, in which the significant non-genetic factors were included as covariates to reduce noise. No human genetic polymorphism could be identified that significantly associated with gut microbiome diversity of individual microbial taxa. This allowed us to conclude that non-genetic variables are the driving forces determining microbiome composition in heathy individuals, and that the direct influence of human genetics on the overall gut composition has a minor role.

The various studies presented in this thesis contribute to the generation of a more detailed map of the genetic architecture of variation in the immune function. In particular, the pivotal role of HLA genes is highlighted by several lines of evidence.

Population studies carried out over the last several decades have identified a long list of human diseases that are significantly more common among individuals that carry particular HLA alleles. We here confirmed the role of HLA genes in the determination of the healthy immune variance. Our works further accredited that small differences in the capacity of HLA class II molecules to bind specific viral peptides can have a measurable impact on the downstream antibody production. In a similar fashion, the modulation of parameters of innate immune cells is strongly governed by HLA variation.

We concentrated on pinpointing specific amino acids within HLA molecules that could be responsible for these effects. This, in its turn, can help in revealing the causal genes which can lead to explaining the disease heritability and to a better understanding of the molecular pathways involved in disease pathogenesis.

An example can be given by our investigation of association we observed between the levels of antibodies directed towards the EBNA epitope of EBV with a SNP in the HLA region.
Understanding the biology underlying this genetic association is critical, in particular for increasing our knowledge on susceptibility to multiple sclerosis (MS), an autoimmune demyelinating disease of the central nervous system. Indeed, in the clinical setting, the serum titers of IgG against EBV EBNA have been identified as a strong and robust predictive marker of MS occurrence, with an 8-fold higher relative risk for individuals with high levels of IgG anti-EBNA, as compared with those with low levels (Almohmeed Y.H., *et. al*. *PLoS One*. 2013).
We found that associations between EBV EBNA titers and HLA are due to variations in amino acid composition of HLA-DRβ1. Specifically, the level of IgG mounted against EBV EBNA associated with HLA-DRB1*07:01 haplotype, and with amino acid positions 98 and 104 of the HLA-DRβ1 molecule. These amino acid residues are implicated in the peptide binding groove conformation. Statistically, this association can also be attributed to the presence of HLA-

DRB1*07:01 haplotype, which was observed as a protective allele for MS by previous genome-wide association studies that have mapped MS susceptibility to the HLA-DR locus.
We can now conclude based on these findings that the conformation of the peptide binding groove of the HLA-DRβ1 molecule (encoded by HLA-DRB1*0701) plays a crucial role in determining the repertoire of self and non-self peptides that can be efficiently presented to immune cells; and that the same amino acid variants can confer both higher capacity to mount an anti-EBNA IgG response and higher risk of developing MS.

Our works help to pinpoint potential therapeutic targets and can lead to a better understanding of the structure and the nature of potential antigens for autoimmune or inflammatory diseases, and these can then be tested through binding assays and molecular modeling. Therefore, we emphasize the importance of considering HLA diversity in disease association studies, as the associations we have identified have also the potential to help improve vaccination strategies, and dissect pathogenic mechanisms implicated in human diseases.

Yet, this is still just the first step. The results obtained are the founding bricks on which such further investigation should be conducted to more comprehensively understand the many factors that play a regulating role in human immunity.

Future explorations in the field will need to consider additional types of genetic variants, which were for the most part not interrogated in our genotyping-based studies. Indeed, both rare and common SNPs, as well as structural variants, are very likely to play a role in the variability of human immune responses. In the second chapter of the thesis, we used burden testing to interrogate rare variants. The results presented will need to be replicated, yet they suggest that with a bigger sample size and a more detailed characterization of genetic polymorphisms (*e.g.* from whole genome sequencing), many more genetic influences on immunity could be discovered. This is expected, as genetic diversity in immune-related genes is well tolerated at the population level, because the downstream effects are context-dependent (e.g. depending on the exposure to a pathogen). Immune genetic diversity can even be maintained by evolutionary pressures, in case of balancing selection, which leads to the persistence, at the population level, of a relatively high level of variation in immune function that may become useful in future challenges.

A clear limitation of our investigations is their exclusive scope on individual of European ancestry. Efforts should be intensified to include diverse human populations in biomedical research, but especially in genetic studies. Because of their longest genetic history, individuals of African ancestry are more likely to carry variants that are not observed in other populations. Also, human populations have experienced vastly different evolutionary forces depending on their geographical and genealogical history, resulting in additional levels of variability, which could be highly informative about human health.

A fundamental question in biology is how the nature (human genome) and the environment (nurture) jointly contribute to the observed variability between individuals of the same species. Obviously, genomic and environmental influences are often hard to disentangle, and novel conceptual and statistical approaches will be needed to better understand their respective and combined impact on immune traits. The studies presented in this thesis are

only a start, but they are an integral part of a fascinating journey: the patient scientific unravelling of what makes us humans.

Curriculum Vitae

Petar Scepanovic

Email: scepanovic.petar90@gmail.com
Date and Place of Birth: 11th July 1990, Cetinje, Montenegro
Citizenship: Montenegrin

EXPERIENCE

October 2014 – December 2018	PhD Student in Human Genomics Laboratory Supervisor: Prof. Jacques Fellay, École polytechnique fédérale de Lausanne, Switzerland Co-supervisor: Prof. Bart Deplancke, École polytechnique fédérale de Lausanne, Switzerland
July 2013 – September 2014	Research Assistant in Bioinformatics Laboratory Supervisor: Prof. Paolo Provero, University of Turin, Italy
March 2012 – September 2014	Research Assistant in Molecular Biology Laboratory Supervisor: Prof. Carola Ponzetto, University of Turin, Italy

EDUCATION

October 2014 – December 2018	PhD Biotechnology and Bioengineering École polytechnique fédérale de Lausanne, Switzerland
June 2017 – July 2017	Summer Program in Entrepreneurship HEC Paris, France
October 2012 – July 2014	MSc Molecular Biotechnology University of Turin, Italy GPA: 110/110 cum laude
October 2009 – July 2012	BSc Biotechnology University of Turin, Italy GPA: 102/110

LANGUAGES

Montenegrin	Native
English	Fluent (C2)
Italian	Fluent (C2)
French	Intermediate (B2)

PUBLICATIONS

Morena D, Maestro N, Bersani F, Forni PE, Lingua MF, Foglizzo V, **Šćepanović P**, Miretti S, Morotti A, Shern JF, Khan J, Ala U, Provero P, Sala V, Crepaldi T, Gasparini P, Casanova M, Ferrari A, Sozzi G, Chiarle R, Ponzetto C, Taulli R. Hepatocyte Growth Factor-mediated satellite cells niche perturbation promotes development of distinct sarcoma subtypes. *Elife*. 2016 Mar 17; 5.

SIB Swiss Institute of Bioinformatics Members. * The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res*. 2016 Jan 4; 44: D27-37. (* As collaborator)

Patin E, Hasan M, Bergstedt J, Rouilly V, Libri V, Urrutia A, Alanio C, **Scepanovic P**, Hammer C, Jönsson F, Beitz B, Quach H, Lim YW, Hunkapiller J, Zepeda M, Green C, Piasecka B, Leloup L, Rogge L, Huetz F, Peguillet I, Lantz O, Fontes M, Di Santo JP, Thomas S, Fellay J, Duffy D, Quintana-Murci L, Albert ML, for The Milieu Intérieur Consortium. Natural variation in innate immune cell parameters is preferentially driven by genetic factors. *Nat Immunol*. 2018 Mar; 19: 302-314.

Scepanovic P, Alanio C, Hammer C, Hodel F, Bergstedt J, Patin E, Thorball CW, Chaturvedi N, Charbit B, Abel L, Quintana-Murci L, Duffy D, Albert ML, Fellay J, The Milieu Intérieur Consortium. Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Med*. 2018 Jul 27; 10: 59.

Asgari S, Chaturvedi N, **Scepanovic P**, Hammer C, Semmo N, Giostra E, Müllhaupt B, Angus P, Thompson AJ, Moradpour D, Fellay J. Human genomics of acute liver failure due to hepatitis B virus infection: an exome sequencing study in liver transplant recipients. *J Viral Hepat.* 2018; 1–7.

SKILLS

Computational	Python, R, bash, Unix, Microsoft Office, bioinformatics tools for analysis of genotyping, exome, genome, transcriptome, microbiome, viral and microarray data.
Cell Biology and Biochemistry	Cell cultures, establishment of primary cell cultures (murine satellite cells), FACS analysis (apoptosis, cell cycle, membrane and cytoplasmic staining), Western blotting, immunofluorescence, immunohistochemistry
Molecular Biology	Genomic DNA extraction, PCR, agarose gel electrophoresis, RNA extraction, RT-PCR, quantitative real time PCR, cDNA cloning, generation and use of lentiviral vectors, stable and inducible expression of shRNAs and miRNAs
Animal models	Work with transgenic (screening and crossing) and immunocompromised mice (tumor xenograft experiments)

TEACHING EXPERIENCE

February 2015 – February 2016	'Probability and Statistics', Teaching assistant for second year Bachelor's course at School of Life Sciences, EPFL, Switzerland
February 2015 – February 2016	'Integrated laboratory in Life sciences', Teaching assistant for second year Bachelor's course at School of Life Sciences, EPFL, Switzerland

AWARDS & FELLOWSHIPS

Gap Summit	16th – 18th April 2018	Selected to attend as Future Leader of Tomorrow GapSummit, Cambridge, UK
Innosuisse	19th December 2017	Team leader of MyBion - Best Start-up award at Business Concept course of Innosuisse, Switzerland
GYSS	15th – 21st January 2017	Selected to represent EPFL at Global Young Scientists Summit (GYSS) in Singapore
AIRC	03/2013 – 09/2014	Fellowship by Italian Association for Cancer Research (AIRC)
EDISU	10/2009 – 07/2014	Scholarship from Regional Agency for the Right to University Education (EDISU) Piedmont