



**HAL**  
open science

# Defining the boundaries of a healthy immune response using standardized immune monitoring tools

Alejandra Urrutia

► **To cite this version:**

Alejandra Urrutia. Defining the boundaries of a healthy immune response using standardized immune monitoring tools. Immunology. Université Pierre et Marie Curie - Paris VI, 2017. English. NNT : 2017PA066004 . tel-01572191

**HAL Id: tel-01572191**

**<https://theses.hal.science/tel-01572191v1>**

Submitted on 5 Aug 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Université Pierre et Marie Curie

École doctorale Physiologie et Physiopathologie (ED 394)

*Laboratoire d'Immunobiologie des Cellules Dendritiques/Inserm U1223*

**DEFINING THE BOUNDARIES OF A HEALTHY  
IMMUNE RESPONSE USING STANDARDIZED  
IMMUNE MONITORING TOOLS**

*(DÉTERMINATION A L'ÉCHELLE D'UNE POPULATION DE VALEURS DE RÉFÉRENCE  
DE LA RÉPONSE IMMUNITAIRE EN UTILISANT DES OUTILS STANDARDISÉS)*

Par **Alejandra URRUTIA**

Thèse de doctorat d'**IMMUNOLOGIE**

Dirigée par Matthew ALBERT et co-dirigée par Darragh DUFFY

Présentée et soutenue publiquement le **03 Février 2017**

Devant un jury composé de:

Pr. Adrien Six

Pr. Antoine Toubert

Dr. Elisabetta Bianchi

Dr. Eliane Piaggio

Pr. Matthew Albert

Dr. Darragh Duffy

Président de jury

Rapporteur

Rapporteur

Examineur

Directeur de thèse

Co-directeur de thèse



Université Pierre et Marie Curie

École doctorale Physiologie et Physiopathologie (ED 394)

*Laboratoire d'Immunobiologie des Cellules Dendritiques/Inserm U1223*

**DÉTERMINATION A L'ÉCHELLE D'UNE POPULATION DE  
VALEURS DE RÉFÉRENCE DE LA RÉPONSE IMMUNITAIRE EN  
UTILISANT DES OUTILS STANDARDISÉS**

*(DEFINING THE BOUNDARIES OF A HEALTHY IMMUNE RESPONSE USING  
STANDARDIZED IMMUNE MONITORING TOOLS)*

Par **Alejandra URRUTIA**

Thèse de doctorat d'**IMMUNOLOGIE**

Dirigée par Matthew ALBERT et co-dirigée par Darragh DUFFY

Présentée et soutenue publiquement le **03 Février 2017**

Devant un jury composé de:

Pr. Adrien Six

Pr. Antoine Toubert

Dr. Elisabetta Bianchi

Dr. Eliane Piaggio

Pr. Matthew Albert

Dr. Darragh Duffy

Président de jury

Rapporteur

Rapporteur

Examineur

Directeur de thèse

Co-directeur de thèse

## Acknowledgements / Remerciements

*This is a special honor for me to be able to start this manuscript by this acknowledgment section :  
For the people that don't fully know me, I can tell you that this thesis defense has a really special taste for me as this is the result of 15 years of Lab research (but not only) working .  
So, I confess that I'm not as « naive » as can other PhD students be. Before reaching this step, I've been working in different labs meeting different people and all those single « migration » and « presentation » has contributed to my (at least partially) efficient « cross-priming ».  
I want to take advantage of this space to really thanks all the people that participate in my scientific and non-scientific life training.*

*My most sincere thanks,*

*To **Matthew Albert**, my supervisor. It has been a privilege for me to be part of this adventure with you at my side. I will always remember the first time we met when I was applying for the engineer position in your lab by Concours INSERM. You didn't know me but almost immediately after my arrival in the lab (in 2013, 4 years ago) you gave me the opportunity and the freedom to find my space within the LabEx Milieu Intérieur environment. Then you really quickly offered me to start this thesis. I'm not sure if you have realized how much your trust and (I can say here) your « pushy » attitude to make me accept new challenges (the thesis is a minor one ;-)), helped me a lot in term of professional maturation but mostly in term of personal life teaching. And this has reached a new level when you offered me to follow you at San Francisco for my last year of PhD. I'm really grateful for our well balanced interactions (it worked for me) and I've learned a lot thanks to your endless knowledge and your fantastic curiosity and openmind. I know also that this is not the end...*

*To **Darragh Duffy** my co-supervisor. I'm really grateful for our daily interactions covering all the aspects of a lab routine life that I needed. I want to thank you for all the unformal scientific meetings ; kind of personal advisory boards ; always open door in case of complains or satisfaction ; administrative support. I've been really amazed by your capacity to « keep calm » in any situation and your constant professional attitude. This was for me a perfect balance forr this crazy project. I've learned a lot from you during all those years in term of project management and scientifically.*

*A special thanks to the members of my Jury for their time and interest in this project. First, to **Antoine Toubert** and **Elisabetta Bianchi** for their (fast) critical reading of the manuscript. I really want to thank you also for all the nice interactions and chatting that we had during these last years. I'm sure (and hope) that we will have other opportunities to maintain this link.*

***Eliane Piaggio** and **Adrien Six** for accepting so spontaneously to examine my manuscript and to preside at my defense.*

*Thanks also to **Isabelle Cremer** and **Catherine Monnot** (heads of the ED394 UPMC) for having made my PhD project come true. I really want to highlight here your openmindness and reallistic vision about university general aims.*

*To **Darragh** again, as I'm not forgetting the open door for our regular « Friday happy hours ». But for that I need to associate a special thanks to,*

***Molly Ingersoll**. The first time we met was during an Irish 'Christmas 12 pub tour' organized by Darragh (thanks also for making me discover this amazing annual Irish tradition). Molly, I even suspect that you still have a picture of me dancing on a bar table... I really want to thank you for all the conversations that we had (scientific or not), the precious advices that you gave me and your smile and energy (for free too). Your committment for the lab is enormous, you are a source of inspiration for me as a great scientific and group leader woman.*

*To all the LabEx **Milieu Intérieur** (MI) team. During my first year, our daily tempo has been given by the huge sample collection with the related logistic troubles. But this help us to move as a team and to build special links... So thanks,*

To **Milena Hasan** and **Valentina Libri**, you helped me a lot to integrate the MI team and you taught me all about this new worlds of TruCulture tube, huge collection logistic, etc... I admire the work you have accomplished to have made this dream come true.

To **Benoit Beitz** and **Mikael David**, during my two first years we have shared a lot of time during sessions of coffee/lunch, ficolls, flow staining, freezer organization, DNA extraction etc... It was really pleasant to work with you. You allowed me to integrate smoothly this new Institut Pasteur environment.

To **Vincent Rouilly**. Thanks to you, I've discovered this other new world of barcoding/ high-tech traceability and I had the chance to be able to approach and even to talk to a DataAnalyst (or other related names starting with Bio...). You changed my way of doing science and designing my experiments. And mostly thanks for the really friendly and appreciated « happy hour » discussions/sharing (in both continents). You really have contributed a lot in bringing me some lightness during this last though year.

Thanks also to all the past members/ external collaborators of the LabEx MI : **Jonathan Grivel**, **Raouf Djebali**, **Céline Possémé** (good luck for your thesis within ICD now), **Stanislas Mondot**, **Emmanuel Clave**, **Itaua Leston Araujo**, **Benoit Albaud**, **David Gentien** (and all the members of the Genomic Curie Platform), **Pedro Goncalves**, **Friederike Jonsson**. Thanks for so nice collaborations and for having be so patient with me and my demands.

To more recent (and still current) MI members : **Cameron McPherson** and **Bruno Charbit**. Thanks for having taken over some of my tasks in this project. I hope that I didn't leave you too much mess (but you know where to find me if necessary !). I already know that we will keep collaborating.

To **Magnus Fontes**, **Antonio Borderia** & Co. **Magnus**, with your now legendary and unique enthusiasm, you initiated me to this fascinating world of mathematic-applied-to-biology.

Special thanks to **Lluís Quintana-Murci** and the members of his team, **Etienne Patin**, **Hélène Quach** and **Nora Zidane** for your time and nice coffee/terrace sessions. Thanks a lot **Lluís** for all your challenging questions and for having been present whenever it was necessary. I've learned a lot in term of genetic of populations thanks to your wonderful teaching skills (but also from your team!).

To the wonderful **Stéphanie Thomas**. You were my link between MI and ICD. I always admired your energy, positivity and how you were able to talk so easily about science with no scientific background. We miss you, even if I must confess that it was easier for me to write this thesis after your departure ! I wish you all the best for your new life and hope that our paths will cross again.().

To all the members of ICD lab. We didn't shared so much benchwork time but the rest (coffee, labmeetings, retreats, happy hours, events...) was always fun and intense in emotion. So you will always be my main cherished core team (sorry for MI...). To **Alba**, **Armanda**, **Aurélie**, **Biljana**, **Brieuc**, **Cécile**, **Clémence**, **Gabriela**, **Hélène**, **J**, **Marie-Laure**, **Matthieu**, **Nader**, **Pierre-Emmanuel**, **Tracy**, **Vincent**. Thanks to all of you for your friendship and for having made those years so precious to me. I will always cherish the special and intimate shared moments with some of you. I wish you all the best for your next steps.

Thanks to the Pasteur Institute for the great scientific environment and all the members of the Immunology department (merci **Béa** !) and other buildings for your smile and time. Special thanks to **Julie** and **Aleks**. I wish that our friendship will survive across time and continents. Best luck for everything.

To the members of the « other » lab in SF. Special thanks to **Rosa** (ex-ICD member) for welcoming me with all your heart into this new environment. All your efforts have contributed greatly to ease my adaptation. To **Scott** and **Jérémie** (ex-ICD members too) and **Sean** (and not Shaun), **Yoong Wearn** and **Haiyin** for your daily smile, shared commute, enthusiastic chatting and friendship. Thanks to all of you, I quickly felt as being part of a fun group. Thanks for having been part of my new world during this last year and for the support whenever necessary.

To the others members of the CI department and other collaborators (**Cherie, Connie, Magge**) at Genentech for your enthusiasm, efficiency and positivity making everything so easy to perform. This generates an amazing scientific environment that has contributed to my scientific maturation. Special thanks to **Jess, Claudia, Alex, Shannon, Alessandra, Aude-Hélène** and **Cécile** for your spontaneous friendship.

To the members of my first lab at Kremlin-Bicetre and specially to **Christine Lacabaratz** and **Benoit Vingert** for the best training ever in term of cell culture and immunomonitoring skills. I'm glad that our friendship is still there instead of the distance and after all those years.

To all the members of the CIMI at la Pitié-Salp. Special thanks to **Arnaud Moris, Victor Appay** and **Stéphanie Graff-Dubois**. You provided me some of my master pieces during my career's construction. I will always be grateful for all your professional and personal support. Same for **Asier Saez-Cirion** and **Gianfranco Pancino** (here at Pasteur Institute), for contributing to my jump between KB and PS.

We don't know each others (maybe me a little bit more) but I really want to thank you : **Artic Monkeys**. I've listened all your albums, live/acoustic performances during the writing of my thesis. You will forever be associated to this work and probably still inspire me in the future.

I think that I will switch to French and/or Spanish to be really efficient in the next acknowledgment section :

Un énorme merci à la **famille Dubois (S&F)** et à nos soirées « Piscopathes ». Vous m'avez été d'une aide précieuse et inébranlable pendant un moment particulièrement difficile de ma vie. Vous êtes comme une deuxième famille pour moi et j'espère que (tout comme ma gratitude) notre amitié sera éternelle.

Merci à toutes les personnes non citées mais qui m'ont accompagné de près ou de loin tout au long de ces années et qui encore une fois ont contribué à mon évolution (petite pensée pour **Lynda** et nos soirées salsa).

Merci/Gracias à toute ma famille (**Alicia, Tivi...**), pour votre soutien, écoute et votre confiance en mes projets. Muchísimas gracias por su presencia y su oreja cuando lo necesitaba.

Merci à mes frères « Dalton » **Mickaël, Emiliano** et **Emmanuel + Camille** et **Samuel** pour tout, tout, tout et ce qui nous reste à venir.

Merci à mes **parents** sans qui tout ceci ne serait jamais arrivé. Vous m'avez enseigné par des démonstrations quotidiennes que la connaissance est la solution à tout. Ceci est d'autant plus vrai en Sciences. Merci encore pour votre belle leçon de vie que j'essaye d'appliquer au quotidien. J'espère avoir enfin appris à voler de mes propres ailes (como una Urpi)..

Un énorme merci à **Alexis** pour ton soutien et ta confiance. Je ne trouve pas les mots exacts mais c'est grâce à toi si ce projet a pu aboutir. Je n'oublierai jamais ton geste.

Enfin, merci à **Ambre** et **Aymeric**. Vous êtes mes piliers sur tous les aspects de ma vie et vous ne vous en rendez sûrement pas compte, mais vous faite partie intégrante de ce travail.

## ABBREVIATIONS

AIM2	Absent melanoma 2
ASC	Apoptosis-associated speck-like protein containing a CARD
BCG	Bacille Calmette-Guerin
Bcl	B-cell lymphoma
BCR	B cell receptor
CARD	Caspase activation and recruitment domain
CCR	CC-chemokine receptor
CCL	CC-chemokine ligand
CD	Cluster of differentiation
cDCs	Conventional dendritic cells
CLEC	C-type lectin-like receptor
CM	CD4+ or CD8+ central memory T cells
CMV	Cytomegalovirus
CR	Complement receptor
CTL	Cytotoxic CD8+T cell
CTLRs	C-type lectin-like receptors
CXCL	CX-chemokine ligand
CXCR	CX-chemokine receptor
DCs	Dendritic cells
DC-SIGN	DC-specific intercellular adhesion molecule-3-grabbing nonintegrin
Dectin	DC-associated C-type lectin
DNA	Deoxyribonucleic acid
ds or ssRNA/DNA	Double- or single-stranded Ribonucleic acid/Desoxyribonucleic acid
eQTL	Gene expression quantitative trait loci
EM	Effector memory T cells
EMRA	Effector memory CD45RA+ T cells or effector terminal T cells
FADD	Fas-associated protein with a death domain
FcγR or FCGR	Fragment c gamma receptor
FDR	False discovery rate
γδ T cells	Gamma-delta T cells
GM-CSF	Granulocyte-macrophage colony-stimulating factor
GWAS	Genotype wide association study
HKCA	Heat killed <i>Candida albicans</i>
HKEC	Heat killed <i>Escherichia coli</i>
HKHP	Heat killed <i>Helicobacter pilori</i>
HKSA	Heat killed <i>Staphylococcus aureus</i>
HLA	Human leukocyte antigen
HLA-DR	HLA- antigen D related
IAV	Influenza A virus
IFN	Interferon
IFNAR	Interferon alpha receptor
Ig	Immunoglobulin
IL	Interleukin
IL-R	Interleukin-receptor
ILC	Innate lymphoid cell
IRF	Interferon regulatory factor
ISG	Interferon stimulated gene
ITAM	Immunoreceptor tyrosine-based activatory motif

ITIM	Immunoreceptor tyrosine-based inhibitory motif
JAK	Janus kinase
KIR	Killer cell immunoglobulin-like receptor
LBP	Lipopolysaccharide binding protein
LPS	Lipopolysaccharide
MAF	Minor allele frequency
MAPK	Mitogen-activated protein kinase
MAIT	Mucosal-associated invariant T cell
MDA	Melanoma differentiation-associated protein
MFI	Mean fluorescence intensity
MHC-I or-II	Major histocompatibility complex class-I or class-II
MI	<i>Milieu Intérieur</i> study
mRNA	Messenger ribonucleic acid
MyD	Myeloid differentiation primary response gene
NETs	Neutrophil extracellular traps
NF- $\kappa$ B	Nucleofactor-kappaB
NK	Natural killer
NKC	Natural killer gene complex
NLRs	NOD-like receptors
NLRP3	NACHT, LRR and PYD domains-containing protein 3
NOD	Nucleotide-binding oligomerization domain proteins
PAMPs	Pathogen associated molecular patterns
PCA	Principal component analysis
pDCs	Plasmacytoid dendritic cells
poly I:C	Polyinosinic:polycytidylic acid
pQTL	Protein expression quantitative trait loci
PRRs	Pattern recognition receptors
RAG	Recombination-activating gene
RIG-I	Retinoic acid-inducible gene I
RIP	Receptor-interacting protein
RT-qPCR	Reverse-transcriptase-quantitative polymerase chain reaction
RNA seq	Ribonucleotide sequencing
ROS	Reactive oxygen species
S1PR1	Sphingosine-1-phosphate receptor 1
SEB	Streptococcus enterotoxin B
SeV	Sendai virus
SNP	Single nucleotide polymorphism
STAT	signal transducer and activator of transcription
STING	Stimulator of interferon genes
SVM	Support vector machine
Syc	Spleen tyrosine kinase
TCR	T cell receptor
TGF	Tumor growth factor
Th	CD4+ T cell helper
TIR	Toll/IL-1 receptor domain
TLRs	Toll-like receptors
TNF	Tumor necrosis factor
TRAF	TNF receptor-associated factor
TRIF	TIR domain-containing adapter-inducing interferon- $\beta$
Treg	Regulatory T cell



# Table of Contents

<b>CHAPTER I .....</b>	<b>11</b>
<b>Chapter I : Introduction and background .....</b>	<b>12</b>
<b>1. A general introduction to the immune response .....</b>	<b>15</b>
1.1. Major types of immune responses : an equilibrate balance .....	15
1.2. The innate immune system.....	17
1.2.1. Innate immune cells.....	18
1.2.1.1. Mast cells, basophils and eosinophils .....	18
1.2.1.2. Neutrophils.....	19
1.2.1.3. Monocytes/macrophages.....	20
1.2.1.4. Dendritic cells and antigen presentation .....	21
1.2.1.5. Natural Killer cells and Innate lymphoid cells .....	22
1.2.2. Platelets and complement system .....	24
1.3. The adaptive immune system.....	26
1.3.1. T lymphocytes: the cellular response.....	28
1.3.1.1. CD8 <sup>+</sup> T lymphocytes .....	28
1.3.1.2. Helper T lymphocytes.....	28
1.3.2. B lymphocytes: the humoral adaptive response .....	30
<b>2. Inflammation .....</b>	<b>32</b>
<b>3. Inflammatory cytokines .....</b>	<b>34</b>
3.1. Tumor necrosis factor .....	35
3.2. Interleukin-1 $\beta$ .....	36
3.3. Type I Interferons .....	37
3.4. Interferon Gamma.....	38
<b>4. Pattern recognition receptors .....</b>	<b>40</b>
4.1. Toll-like receptors.....	41
4.2. RIG-I-like receptors .....	42
4.3. NOD-like receptors (NLRs) .....	43
4.4. C-type lectin-like receptors (CTLRs) .....	43
<b>5. Microbial recognition by PRRs .....</b>	<b>45</b>
5.1. Bacterial Recognition by PRRs .....	46
5.1.1. Bacterial external components .....	46
5.1.2. Bacterial DNA.....	47
5.2. PAMPs in Mycobacteria .....	48
5.3. Fungal Recognition by PRRs.....	48
5.4. Viral recognition by PRRs.....	49
<b>6. Genetic evolution and phenotypic variation .....</b>	<b>50</b>
<b>7. Immune monitoring and use of whole blood stimulation system.....</b>	<b>51</b>
7.1. Tools for immune monitoring.....	51
7.2. Cell-based approach.....	52
7.3. Animal models .....	52
7.4. Human whole blood model .....	53
<b>CHAPTER II: Standardized whole-blood transcriptional profiling enables the deconvolution of complex induced immune responses .....</b>	<b>56</b>
<b>1. Introduction.....</b>	<b>58</b>
<b>2. Results .....</b>	<b>59</b>
2.1. Distinct transcriptional signatures induced by the IFNB, IFNG, IL1B and TNFA cytokines.....	59
2.2. Variable responses to TLR and microbe stimulation are captured by induced cytokine response.....	71
2.3. Inter-individual variable gene expression supports tracing of cytokine loops.....	78
2.4. Unique microbial gene expression is defined by lymphocyte-derived cytokines.....	81
<b>3. Discussion.....</b>	<b>84</b>

<b>CHAPTER III: Natural variation in innate immune cell parameters is driven by genetic factors</b> .....	<b>90</b>
<b>1. Introduction</b> .....	<b>92</b>
<b>2. Results</b> .....	<b>94</b>
2.1. Variation of immune cell parameters in the general population .....	94
2.2. Impact of age, gender and cytomegalovirus infection on immune cell parameters .....	110
2.3. Tobacco smoking alters extensively immune cell numbers .....	113
2.4. Genome-wide association study of 168 immune cell parameters .....	119
2.5. Most genetic associations act as immune cell-specific protein QTLs .....	126
2.6. Novel <i>trans</i> -acting genetic associations with immune cell parameters .....	133
2.7. Genetic control of mRNA expression explains association with protein levels of immune cell markers.....	135
2.8. The natural variance of innate immune cell parameters is driven by genetic factors. ....	135
2.9. Design and preliminary results from first replicative study .....	138
2.9.1. Donor collection.....	138
2.9.2. Immune cell type assesment.....	139
2.9.2.1. Antibodies.....	140
2.9.2.2. Instrument and cell count assessment.....	141
2.9.2.3. Comparison between <i>Milieu Intérieur</i> and replicative collection.....	143
<b>3. Discussion</b> .....	<b>147</b>
<b>CHAPTER IV: Discussion and Perspectives</b> .....	<b>152</b>
<b>MATERIAL AND METHODS:</b> .....	<b>160</b>
<b>Technical challenges - How to build a reference dataset?</b> .....	<b>160</b>
<b>1. Donor collections</b> .....	<b>161</b>
1.1. The <i>Milieu Intérieur</i> cohort .....	161
1.1.1. Healthy donors' recruitment.....	161
1.1.2. Clinical and epidemiological collection .....	161
1.1.3. Biological collection .....	162
1.2. The Genentech Genotype and Phenotype Registry collection .....	164
<b>2. Focus on the study of induced immune responses</b> .....	<b>164</b>
<b>3. Standardized gene expression quantification from stimulated whole blood</b> .....	<b>167</b>
3.1. Development of a high throughput single-step RNA isolation protocol.....	167
3.1.1. Challenge of Trizol® stabilized whole blood samples.....	167
3.1.2. High-throughput standardized RNA extraction.....	168
3.1.3. RNA quality.....	168
3.2. mRNA quantification: direct hybridization assay .....	169
3.2.1. nCounter codeset by Nanostring Technologies .....	172
3.2.2. Automated sample processing and analysis.....	172
3.3. Technical validation .....	173
3.3.1. Imaging quality control .....	173
3.3.2. Internal positive probes.....	173
3.3.3. Internal negative probes.....	173
3.4. mRNA content normalization .....	174
3.4.1. Criteria for housekeeping gene selection .....	174
3.4.2. Housekeeping gene selection using gNorm R package.....	174
3.5. Final data normalization pipeline using the nSolver® software .....	175
<b>4. Selection criteria for gene expression analysis</b> .....	<b>175</b>
<b>5. Immune cell phenotype analysis and data normalization</b> .....	<b>186</b>
5.1. Human material and staining protocol.....	186
5.2. Reproducibility testing and assay development.....	186
5.3. Cytometric analyses .....	186
5.4. Data transformation.....	198
5.4.1. Outlier removal .....	198
5.4.2. Adjustment for batch effects .....	198

<b>6. Statistical analysis, Data visualization and Softwares</b> .....	<b>199</b>
6.1. Dimension reduction: Principal Component Analysis .....	199
6.2. Dimension reduction: factor analysis.....	199
6.3. Silhouette score analysis.....	200
6.4. Boostrapped hierarchical clustering.....	200
6.5. Support vector machine .....	200
6.6. Non-genetic factors analysis on immune cell phenotypes .....	201
6.6.1. Correlation matrix of cell counts .....	201
6.6.2. Stability selection of analyzed environmental variables .....	201
6.6.3. Analysis of the impact of non-genetic factors on immunophenotypes.....	202
6.7. Genetic-related analysis on circulating cell population parameters .....	203
6.7.1. Genome-wide DNA genotyping.....	203
6.7.2. Genetic relatedness and structure.....	203
6.7.3. Genotype imputation.....	204
6.7.4. Genome-wide association analyses.....	204
6.8. HLA typing and association tests.....	206
6.9. Expression quantitative trait loci analysis .....	206
6.10. Decomposition of the proportion of variance explained.....	207
<b>7. Data mining tool: R Shiny (Interactive Web Application) development</b> .....	<b>207</b>
<b>List of illustrations</b> .....	<b>210</b>
<b>List of tables</b> .....	<b>212</b>
<b>PhD PUBLICATION (Study 1)</b> .....	<b>214</b>
Urrutia et al ., Cell Reports 2016 : .....	214
<b>ADDITIONAL INFORMATIONS (STUDY 1)</b> .....	<b>215</b>
TECAN Script For Automated RNA Purification.....	215
<b>ADDITIONAL PUBLICATIONS</b> .....	<b>216</b>
<b>1. Thomas et al ., CI 2015</b> .....	<b>216</b>
<b>2. Duffy et al ., Immunity 2014</b> .....	<b>216</b>
<b>3. Hasan et al ., CI 2015</b> .....	<b>216</b>
<b>4. Chen et al ., CI 2015</b> .....	<b>216</b>
<b>REFERENCES</b> .....	<b>218</b>
<u>Résumé</u> : .....	238
<u>Abstract</u> : Defining the boundaries of a healthy immune response using standardized immune monitoring tools.....	238

# CHAPTER I

---

## INTRODUCTION AND BACKGROUND

## Chapter I : Introduction and background

### The “Milieu intérieur” : an immune-controlled equilibrium

The immune system plays a key role in maintaining a healthy state by preventing infection and malignant transformation but also by participating in the steady-state tissue homeostasis. As early as 1859, Claude Bernard had conceptualized this constantly preserved equilibrium and named it the “Milieu Intérieur”.

Whenever this equilibrium is lost, this leads to the initiation of an inflammation characterized by the secretion of a myriad of inflammatory mediators released into the circulation in distinct kinetic patterns. However, relatively little is known about the key components that determine the strength and/or the resolution of this inflammation and that can impact the ability to return to a homeostasis status in certain individuals.

Those components can be environmental, genetic or of a pathogenic nature, and can lead to the loss of equilibrium and to disease pathogenesis. To define the factors responsible for these perturbations, an increasing number of studies employ system approaches, thus providing datasets that can be explored for new insights into the workings of the immune system.

These studies aim to capture, at an individual level, high-resolution data across a wide array of parameters, whether molecular (genome, transcriptome, proteome or metabolome), environmental or behavioral. This approach is required to take into account the natural human diversity which results in variability among individuals.

The increased use of “omics” technologies has changed our comprehension of the complex molecular networks involved in cellular signaling pathways, with notable applications for the study of immunity and inflammation. Numerous studies using human or animal models, have described the pathways involved in different pathological models (e.g., infectious diseases, auto-immunity, genetic polymorphism) but few have focused on a deep analysis of healthy immune responses. Knowledge of baseline responses in healthy persons is crucial for the understanding of the pathologic context as it can be used as a reference. Such baseline data is increasingly important as datasets coming from diverse “omics” technologies are integrated. Indeed, the diversity of protocols used between studies makes it difficult to directly compare between datasets. In order to implement accurate comparisons it becomes increasingly important to standardize the different steps of these systemic approaches. Finally, these observed signatures strongly depend on the study design, with most approaches using single

defined primary cell types or cell lines, without taking into account the importance of cellular interactions and/or the microenvironment for a given immune response.

The *Milieu Intérieur* (MI) project aims to determine what are the genetic and environmental factors that drive the human immune response and to provide a large dataset that can be used as reference values by the community. To do so, extended epidemiological and biological data was collected from 1,000 healthy donors with a homogeneous ethnic background, stratified across gender (50% men/women) and age (20 to 69 years). In order to minimize pre-analytical biases, a huge effort has been made in the establishment of standardized and robust procedures. In this context, a suite of whole blood, syringe-based assay systems have been developed, thus permitting reproducible assessment of induced innate and adaptive immune responses. The final goal is to define healthy donor reference values for induced inflammatory genes and propose an analytical strategy for deconvoluting inter-cellular interactions. This approach may help identify new applications for therapeutic inhibition of selected cytokine pathways.

In the context of this collection, my PhD thesis has focused around two main studies.

The first study was driven by three principal aims:

- i) **Development of standardized mRNA expression analysis from whole blood syringe based assay stimulations**
- ii) **Application of specific statistical tools for mRNA analysis to define four major inflammatory cytokine induced signatures**
- iii) **Deconvolution of complex induced immune responses using these restricted cytokine induced gene signatures.**

The second study is a multidisciplinary analysis from the consortium that aims to determine the genetic factors that drive the immune cell parameters assessed by flow cytometry in the MI 1000 healthy donors. Following my contribution to the sample collection and data analysis in the French MI study, I took advantage of a student exchange program to perform a smaller scale replication MI study in the Cancer immunology department at Genentech, Inc (South San Francisco, USA). This replication study will enable a validation of some of the genetic/immune parameter associations that have been identified within the French MI cohort.

As a general introduction to this manuscript, I will present briefly the current consensus about the major types of human responses that can be triggered and the major cell populations participating to those responses highlighting how we have identified them in our studies.

After describing what constitutes an inflammation response, I will present the major inflammatory molecules and the receptor/associated molecules complexes that trigger inflammatory responses. As the number of inflammatory molecules is extensive I will focus on those used in our study. I will finish this introduction by presenting how we can monitor human immune responses and the rationale of using whole blood stimulation systems for human clinical studies.

In the second chapter, I will present my results describing the development and analysis of the first transcriptomic study from the MI collection performed on *in vitro* stimulated whole blood samples from 25 healthy donors.

In the third chapter, I will present a multidisciplinary analysis from the MI consortium associating genetic polymorphisms and flow cytometry parameters on the 1000 healthy individuals. I will then present a replication study on 81 healthy donors of Caucasian ancestry that aims at validating some of those associations in an independent collection of healthy donors. Those results are currently integrated into a manuscript in revision that aims to demonstrate that the natural variation in innate immune cell parameters is driven by genetic factors.

In the fourth chapter, I will discuss these results integrating the future perspectives with other *Milieu Intérieur* datasets, but also related to more general clinical applications.

To be able to build a reference dataset in a healthy population, a highly annotated sample collection and high quality datasets are key factors.

I will finish this manuscript by describing the material and methods used for those studies highlighting the different tools specifically developed at each step of the “ *Milieu Intérieur* ” sample and epidemiological collection, with a special technical focus on the relevance of our syringe-based assay to study induced immune responses and specifically the standardized flow cytometry pipeline developed.

# 1. A general introduction to the immune response

In all complex organisms, the immune response results from a sophisticated system involving several organs, numerous specialized cell types and a multitude of highly regulated ligand/receptor interactions.

The immune cells develop in the bone marrow from pluripotent progenitors and, differently to other cell types, are not tightly associated with a particular organ or tissue. They are able to move freely or in the presence of inflammatory signals can interact with all organs to capture cellular debris (i.e. dying cells) or invading particles/microorganisms and will potentially recirculate through the blood or lymph stream.

In the two studies presented in this manuscript, the focus of our analysis was on the immune response from whole blood samples from healthy individuals.

In this introduction, I will present the current views about the major type of immune responses that can be triggered depending of the pathologic condition and give an overview of the major components that participate in the homeostasis of the human immune response. This will include a brief presentation of the cell populations involved. I will then focus on the inflammation phenomenon first by describing key pro-inflammatory cytokines (interleukin-1 $\beta$  (IL-1 $\beta$ ), Tumor Necrosis Factor (TNF), interferon-Beta (IFN- $\beta$ ) and IFN-Gamma (IFN- $\gamma$ ). Following this I will describe the danger sensors that help the cells to induce this inflammation that are triggered by microbial components (pure or more complex) used in our study leading to the polarization and trafficking of the immune cells. I will finish this chapter by describing how we can monitor human immune responses and the rationale of using whole blood stimulation system for clinical related human studies.

## 1.1. Major types of immune responses : an equilibrate balance

Since the seminal studies of Mossman and Coffman (Mosmann and Coffman, 1989) the immune response has been divided into two types of responses depending on whether T helper (Th) type 1 or Th2 cells are induced to coordinate a response against intracellular or extracellular pathogens respectively.

Nonetheless, since these original concepts were proposed the use of single cell technologies coupled to multiparametric analysis has challenged this reductionist classification. In the adaptive cells (i.e. Th cells) as well as in the innate cell compartment (i.e.



monocytes/macrophages) depending of specific inflammatory conditions, each immune population is highly diverse and can be defined by their production of specific cytokine/chemokine patterns that are tightly regulated by gene modulation and transcription factors.

Therefore it appears more appropriate to use the latest nomenclature describing different types of immune responses and to link them with the innate and adaptive cells involved.

This recent model (inspired by old concepts) describes that in the homeostatic state, there is a balance in the organism between principally **four arms** of the immune system including cells and molecules from the innate and adaptive system. This immune equilibrium is established and educated by a primary microbiota acquired during childhood and then maintained or shaped by environmental factors (i.e. diet, drugs). When an infection or injury occurs, the specific inflammatory environment induced stimulates one specific arm of the immune system. Interestingly, this polarization actively represses the other arms until the infection is cleared. Recently, Gerard Eberl reviewed these responses re-introducing the concept of equilibrium — originally conceptualized by Claude Bernard (1813–1878) and Walter Cannon (1871–1945) as the ‘milieu intérieur’ and homeostasis respectively (Eberl, 2016)

**Type 1 responses:** Intracellular bacteria, viruses and tumor

During a type 1 response, the secretion of cytokines IL-12, IL-15 and IL-18 by myeloid cells such as dendritic cells (DCs) and macrophages and non hematopoietic cells leads to the activation of natural killer (NK) cells and group 1 innate lymphoid cells (ILC1) that will be the major innate sources of IFN- $\gamma$  production. This cytokine will notably increase the expression of MHC class II molecules on antigen-presenting cells, activate myeloid cells and Th1/cytotoxic CD8<sup>+</sup> T cells that will secrete the major type 1 components that have direct or indirect anti-microorganism or anti-tumor properties (i.e., IFN- $\gamma$ , perforin and oxygen radicals). In addition Th1 cells can stimulate the production of antibodies (mostly IgG).

**Type 2 responses:** large organisms (i.e. helminthes).

These responses are promoted by the production of IL-25, IL-33 and thymic stromal lymphopietin (TSLP) by non-haematopoietic cells. This pattern of cytokine production leads to the activation of ILC2s, mast cells, basophils and eosinophils, resulting in the development of Th2 cells, production of IL-4, IL-5 and IL-13 and B cell antibody isotype switching to IgG1<sup>+</sup> or IgE<sup>+</sup> memory phenotype and somatic hypermutations. In parallel, induction of antibodies helps to target the pathogen for direct neutralization, activates the complement system, or mediate local degranulation of mast cells, basophils and eosinophils.

**Type 3 responses:** extracellular microorganisms

During an extracellular infection DCs and macrophages produce IL-1 $\beta$  and IL-23 leading to the activation of ILC3 and Th17 cells. A specific inflammatory environment characterizes the effector phase with the presence of IL-1 $\beta$  accompanied by the production of IL-17 and IL-22 by lymphoid cells, antimicrobial peptides (AMPs) by epithelial cells and the recruitment of neutrophils. In the context of strong inflammation and extensive tissue injury, lymphoid cells can produce both IFN- $\gamma$  and IL-17 due to the release of intracellular motifs leading to a mixed type 1 and type 3 phenotypes.

**Type 4 responses:** barrier response

In the same review, G. Eberl proposes a 4<sup>th</sup> arm as a system to protect hyper-exposed and/or sensitive tissues that can be irreversibly damaged (i.e. eye, mouth, gut). Type 4 immunity aims to prevent infection/injury using secretory systems of neutralizing IgA, mucus and antimicrobial peptides (AMPs) by immune and epithelial cells.

## 1.2. The innate immune system

The **innate immune system** comprises the cells and mechanisms that constitute for the host a first line of defense from infections or general damage. In addition to a physical barrier protection (epithelium), its main functions are to identify/remove foreign substances and to recruit specialized immune cells to sites of infection/injury through the production of chemical factors (cytokines, chemokines, alarmins...) and complement cascade activation that notably promote clearance of antibody complexes or dead cells. Finally the antigen presenting cells from the innate system activates adaptive immunity that will confer long-lasting memory immunity to the host.

A common feature between all innate cells is their extensive patrolling activity due to the expression of a myriad of surface or intracellular Pattern Recognition Receptors (PRRs) that will be described later in this chapter.

As mentioned previously, access to high throughput molecular tools revealed a large diversity of molecular pattern receptors inside each innate population in contrast to what was thought before. This diversity has been initially described in models where due to previous inflammatory signals (infection or tissue damage), the circulating innate immune cells do not present the same inflammatory profile and not express the same molecular pattern at their

surface or at the transcriptional level even in absence of antigens. In addition, due to improved methods to study “intact” innate cell populations, it’s now recognized that these cells can also have a longer half-life than it was previously thought adding another dimension to the importance of a potential long-term “memory” innate signature (Mantovani et al., 2011).

### **1.2.1. Innate immune cells**

#### **1.2.1.1. Mast cells, basophils and eosinophils**

Mast cells, basophils and eosinophils are potent effector cells generally associated with type 2 and 3 immune responses (see section “*Major types of immune responses*” and **Table 1**) and also tissue damage in several allergic inflammation contexts.

Mast cells and basophils are developmentally similar and then express a common set of effector molecules, like mast cell-associated proteases (mMCP), vasodilating substances (such as histamine), various cytokines and pro-inflammatory chemokines. Many of these effector molecules are already stored in cytoplasmic granules allowing a fast release in response to activation of the high-affinity receptor for IgE (FcεRI) or other expressed surface receptors (i.e. Toll-like receptors- TLRs). (Gilfillan and Beaven, 2011; Voehringer, 2013).

Human eosinophils present specific granules that store potent toxic proteins and also numerous cytokines, enzymes and growth factors. Among the main receptors that define the unique biology of the eosinophil are interleukin-5 receptor subunit-α (IL-5Rα) and CC-chemokine receptor 3 (CCR3), as well as sialic acid-binding immunoglobulin-like lectin 8 (SIGLEC-8) in humans. The major eosinophils inducers are the cytokine IL-5 and the eotaxin chemokines (i.e. CC-chemokine ligand 11 or CCL11) that will allow their recruitment into tissues. Other signals like cytokines (i.e. IFNγ or CCL11) and pattern-recognition receptors (PRRs) activation will promote the eosinophil degranulation (Rosenberg et al., 2013).

### 1.2.1.2. Neutrophils

Neutrophils are the main population composing the PMN cell family. These cells are classically characterized by their ability to act as phagocytic cells, to release lytic enzymes from their granules, to produce reactive oxygen species (ROS) and antimicrobial peptides directly killing or limiting the growth of bacteria and fungi (Boregaard, 2010).

They normally represent 50 to 60% of the total circulating leukocytes making them the most abundant cells of the innate immune system and can be identified notably by their high expression of Fc Gamma Receptor IIIA (FcγR3A or CD16) (**Table 1**). After stimulation they induce gene expression of key inflammatory mediators, including complement components, Fc receptors, chemokines (i.e. CCX-chemokine ligand 8, CCL20) and cytokines (i.e. IL-1s, IFNs) (reviewed in Mantovani et al., 2011).

More recently, Neutrophil extracellular traps (NETs) were described as an anti-microbial mechanism by the extravasation from the cytoplasm of a DNA matrix coated by proteins (i.e. neutrophil elastase, matrix metalloproteinase 9) to capture and destroy the microorganism (Brinkmann et al., 2004).

With the recent access of new tools to study “intact” neutrophils, their role have emerged as key components of the effector and regulatory circuits of the innate and adaptive immune systems in increasing number of inflammatory models (Mantovani et al., 2011).

**Innate immunity**  
 (53-86% of leucocytes)

Cell type % of leucocytes	Protein marker <sup>a</sup> ( <i>Hasan et al. 2015</i> )	Gene marker <sup>b</sup>	Type of immune response	Major effector functions and molecules	Major regulators	PRRs expression
<b>Neutrophils</b> 30 – 80%	CD16 <sup>hi</sup>	<i>FCGR3A</i> <i>CSF3R</i> <i>FPR2</i> <i>MME</i>	Type 1, Type 2 and Type 3	<b>Phagocytosis</b> <b>ROS production ++</b> <b>Inflammation:</b> i.e. IL-1 $\alpha$ and -1 $\beta$ , CCL2, CXCL9, CXCL10, CCL20, TNF	CXCR1 or CXCR2/CXCL8	TLRs, NLRs, RLRs, CTLRs
<b>Eosinophils</b> 0 – 7%	CDw125 <sup>+</sup> SSC <sup>hi</sup>	<i>IL5RA</i> <i>SIGLEC-8</i> <i>CCR3</i> <i>PTGDR2</i> <i>SMPD3</i> <i>THBS1</i>	Type 1, Type 2 and Type 3	<b>Degranulation:</b> - Cationic proteins degranulation: i.e. MBP - Enzyme degranulation: i.e. collagenase, histaminase <b>Cytokines:</b> i.e. IL4, IL5	CCR3 /CCL11, IL-5R/IL-5, IL-25, IL-33 SIGLEC-8 TSLP	TLRs* NLRs* RLRs* CTLRs*
<b>Basophils</b> 0 – 2%	Fc $\epsilon$ RI <sup>hi</sup> CD203c <sup>lo/hi</sup> CDw125 <sup>-</sup>	<i>FCERIA</i> <i>ENPP3</i> <i>IL5RA</i>	Type 2 and Type 3	<b>Degranulation:</b> - Proteases: i.e. Mast cells associated proteases (mMCP) - Histamine	Fc $\epsilon$ RI IL-25 IL-33 TSLP	TLRs* NLRs* RLRs* CTLRs*
<b>Monocytes</b> 2 – 12%	CD14 <sup>hi</sup> CD16 <sup>neg</sup> (Conventional) CD14 <sup>lo</sup> CD16 <sup>int</sup> (Intermediaire) CD14 <sup>neg</sup> CD16 <sup>pos</sup> (Inflammatory)	<i>CD14</i> , <i>APOE</i> , <i>CCL7</i> , <i>CD68</i> , <i>CHIT1</i> , <i>CXCL5</i> , <i>MARCO</i> , <i>MSR1</i> , <i>FCGR3A</i>	Type 1, Type 2 and Type 3	<b>Phagocytosis</b> <b>ROS production +++</b> <b>Inflammation:</b> IL-1 $\alpha$ and -1 $\beta$ , TNF, IL-12, CXCL10	CCR2/ CCL2 or CCL7 CX3CR1/CX3CL1	TLRs, NLRs, RLRs, CTLRs
<b>Dendritic Cells</b> 0.3 – 0.9%	HLA-Cl.II <sup>+</sup> cDC1: BDCA1 <sup>+</sup> cDC3: BDCA3 <sup>+</sup> pDC: BDCA2 <sup>+</sup> BDCA4 <sup>+</sup>	<b>cDCs:</b> <i>CCL13</i> , <i>CCL17</i> , <i>CCL22</i> , <i>CD209</i> , <i>HSD11B1</i> , <i>CD1A</i> <b>pDC:</b> <i>IL3RA</i> , <i>THBD</i> , <i>CLE4C</i> , <i>NRP1</i> ,	Type 1, Type 2 and Type 3	<b>Phagocytosis</b> <b>ROS production +</b> <b>MHC-I and -II antigen presentation</b> <b>- Inflammation:</b> IL12, Type I IFNs <b>- Anti-inflammatory:</b> IL10	PAMPs TGF- $\beta$ TSLP, Type I IFNs	TLRs, NLRs, RLRs, CTLRs (i.e. DC-SIGN, CLEC4C)
<b>NK Cells</b> 1 – 6%	NKp46+ CD56 <sup>hi</sup> CD16 <sup>lo/-</sup> CD16 <sup>hi</sup> CD56 <sup>lo</sup>	<i>NCR1</i> <i>NCAMI</i> (CD56) <i>FCGR3A</i>	Type 1	<b>Cytolytic proteins:</b> i.e. Perforin, Granzymes <b>Anti-viral/-tumoral:</b> IFNg, TNF- $\alpha$ , CCL3, CCL4 and CCL5	IL-12,, IL-15, , IL-18, KIRs, MHC , NCRs	TLRs* NLRs* RLRs* CTLRs*

**Table 1 : Characteristic and function of major circulating innate cells**

<sup>a</sup> The surface protein indicated are the markers used to identify each population within the Milieu Interieur general panel (Hasan et al. Cl. 2015)

<sup>b</sup> Gene list commonly used to identify specific populations. The genes indicated in **blue bold** are the ones included in the human Immunology v2 codeset used in our study.

\* Expression detected at mRNA or protein level but the signal pathway resulting from the triggering of these receptors still need to be explored for those cell types.

cDC: conventional dendritic cell; CTLRs: C-type Lectin-like receptors ; MHC: Major histocompatibility complex; NK: natural killer cells ; NLRs: NOD-like receptors; pDC: plasmacytoid DC; PRRs: Pattern rcognition receptors; RLRs: RIG-I-like receptors; ROS: Reactive oxygen species; TLRs: Toll-like receptors.

**1.2.1.3. Monocytes/macrophages**

Monocytes are circulating innate cells that can further differentiate into tissue macrophages and dendritic cells (see next paragraph) (Auffray, et al., 2009). First described by Elie Metchnikoff in the 19<sup>th</sup> century, it's now well established that tissue macrophages are the most efficient phagocytes. Upon activation, monocytes/macrophages produce a myriad of pro-inflammatory molecules (i.e. tumor necrosis factor) and a high production capacity of ROS to destroy engulfed elements.

The subsequent differentiation of bloodstream monocytes is homeostatic or inflammatory environment dependent, generating subpopulations differing by their size, trafficking and innate immune receptor expression.

Human monocytes are divided into subsets on the basis of CD14 and CD16 expression (see **Table 1**) with the CD14<sup>hi</sup>CD16<sup>neg</sup> monocytes, referred as classical monocytes, being the most prevalent subset in human blood and these cells express CCR2 (Ingersoll et al., 2010).

The distinct trafficking ability in the monocyte subsets involves different chemokine-chemokine receptor axes depending on the homeostatic or inflammatory environment. CC-chemokine ligand 2 (CCL2) and CCL7 bind to CCR2 and are secreted by almost all nucleated cells in response to pro-inflammatory cytokines or innate immune receptor triggering (i.e. PRRs) by microbial products. CCL8 and CCL12 are also CCR2 ligands but their role in the monocyte trafficking is minor compared to CCL2 and CCL7.

Subsets of monocytes express the CX3-Chemokine receptor 1 (CX3CR1) that modulates their patrolling function and survival in response to its ligand CX3CL1. The receptors CCR1 and CCR5 and their ligands CCL3 (MIP1- $\alpha$ ) and CCL5 (respectively) play also a major role in monocyte migration (Shi and Pamer, 2011).

#### **1.2.1.4. Dendritic cells and antigen presentation**

The dendritic cells (DCs), described for the first time by Ralph Steinman in 1973, are the most specialized cells in antigen processing thanks to a low rate and controlled degradation of the phagocytosed particles to allow presentation. Dendritic cells have been extensively studied for their large therapeutic potential reflecting their specialized functions in antigen uptake and presentation. In the periphery, DCs are present in an immature state with two major types usually described. The most abundant are the conventional DCs (cDCs) that can be subdivided in three categories depending on surface markers (cDC1, cDC2 and cDC3) and the plasmacytoid DCs (pDCs) (see **Table 1**). The pDCs are the most potent cells in producing type I interferons in response to viruses (Perussia et al., 1985; Chehimi et al., 1989).

One of their unique features is the capacity to efficiently present endogenous and exogenous antigens through both major histocompatibility complex class-I (MHC-I) and -II (MHC-II) pathways. Notably, the unconventional presentation of exogenous antigens on MHC-I by cDCs referred to as cross-presentation (Bevan, 1976; Buckwalter and Albert, 2009) is an important process as they are the only cells able to prime a T cell response.

The triggering of PRRs by their ligands activates DCs and induces their migration to lymphoid organs for antigen presentation to naïve lymphocytes (Albert et al., 2001; Reis e Sousa et al., 2006).

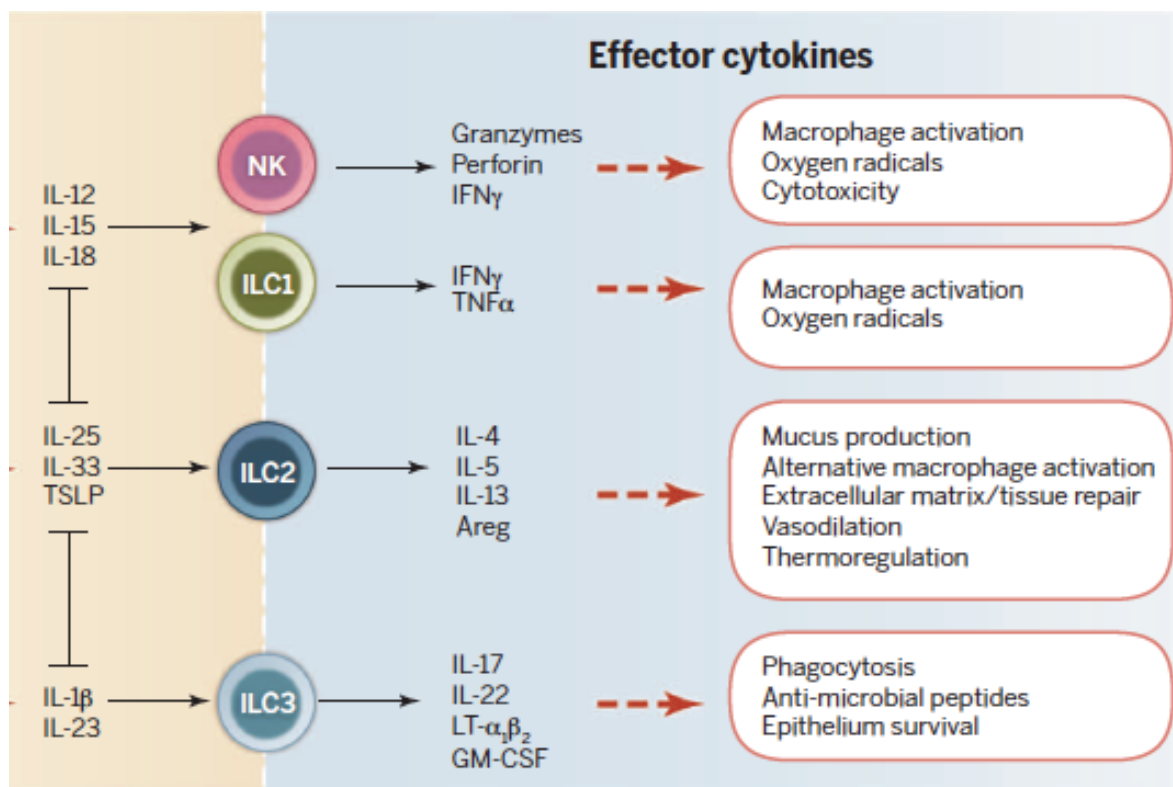
Activated DCs will normally present high levels of MHC molecules bearing pathogen-derived peptides, which can engage T-cell receptors on naïve specific T cells. This delivers the first activating signal ('signal 1') to the T cell. Activated DCs also express a variety of co-stimulatory molecules (i.e. CD40, CD80/CD86; 'signal 2') that will deliver important signals (i.e. proliferation, survival) to T cells, B cells or innate lymphoid cells with the help of CD4<sup>+</sup> T cell (Th). Finally, activated DCs and Th produce mediators (i.e. IL-12, IL-2) to promote T cell differentiation into an effector cell (signal 3). The integration of these three classes of signal by the T cell determines its subsequent fate.

#### **1.2.1.5. Natural Killer cells and Innate lymphoid cells**

NK cells do not have phagocytosis properties to directly target microbes but rather spontaneously destroy compromised host cells presenting a pattern known as "missing self" due to abnormally low levels of self-identifying proteins such as MHC-I. Without prior sensitization, stressed cells can induce NK cell functions like cytotoxicity, cytokine production and proliferation.

In humans, Natural Killer (NK) cells can be divided according to the density of CD56 (**Table 1**). Activated NK cells secrete large amount of cytokines (mostly IFN $\gamma$ ), but they can also secrete interleukins (i.e. IL 10), TNF, growth factors (i.e GM CSF), chemokines (i.e. CCL3, CCL4 and CCL5) and cytotoxic proteins (i.e. perforin, granzymes) (reviewed in Morvan and Lanier, 2016). Their activation and function are tightly regulated by the equilibrium between activating or inhibitory receptors. Among the numerous receptors described, NKG2D (also known as KLRK1), the natural cytotoxicity receptors (NCRs; i.e. NKp46), DNAM1 (or CD226) and CD16 are the best-characterized activating NK cell receptors implicated in immune responses against cancer for example. The well-studied killer cell immunoglobulin-like receptor (KIR) family includes numerous members presenting for mostly (but not all) inhibitory properties (Bashirova, et al., 2006).

It is now well recognized that NK cells are the founding members of the innate lymphoid cell (ILC) family. In the absence of adaptive antigen receptors, NK cells and other ILCs react to the microenvironment through cytokine receptors. Nonetheless NK cells are still considered as the major population presenting specific cytolytic functions (similar to cytotoxic CD8<sup>+</sup>T cells), whereas the ILC1, ILC2 and ILC3 family subtypes are mostly characterized by their signature cytokine secretion profiles (**Figure 1**). Interestingly, ILCs were found to share molecular patterns with their adaptive counterpart, the T helper cells (Eberl et al., 2015).



**Figure 1 : Activator signals and function profile of NK cells, ILC1, ILC2 and ILC3 subtypes**

The effector functions of ILCs mirror the functions of CD8<sup>+</sup> and CD4<sup>+</sup> T cells

(Adapted from Eberl et al. Science, 2015).

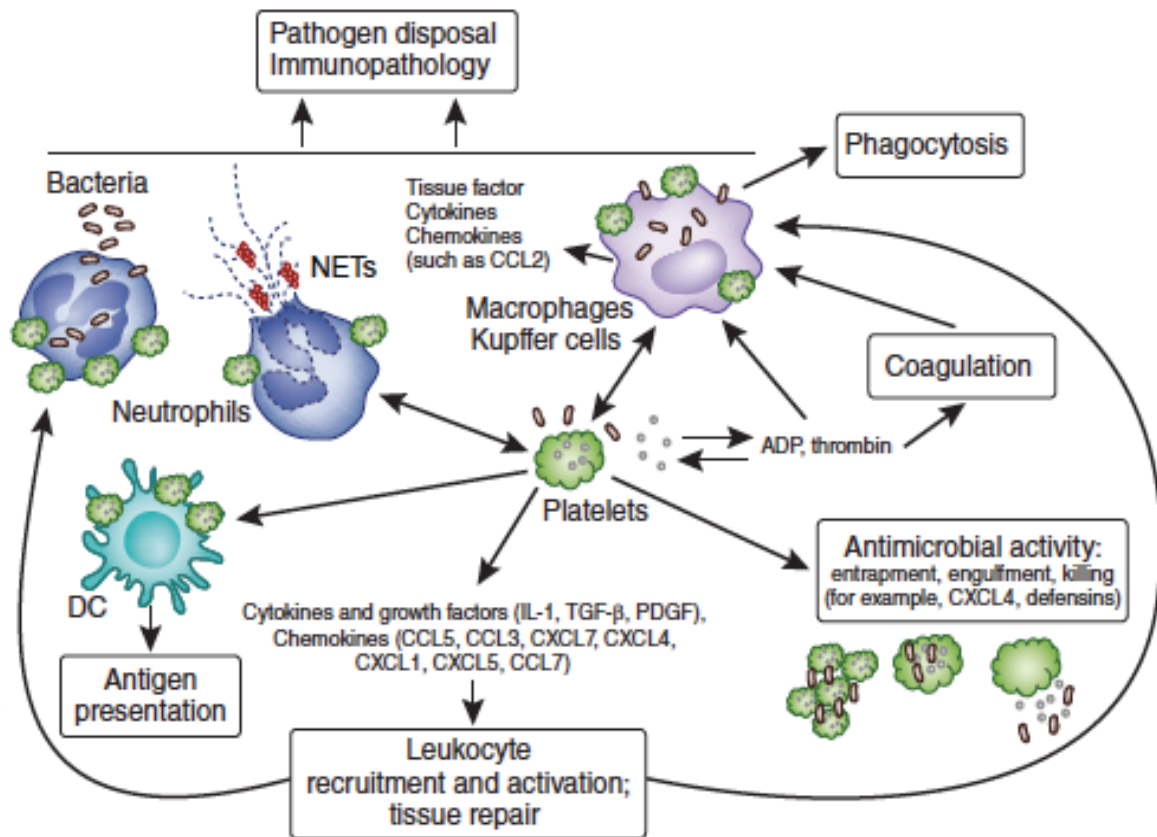


### 1.2.2. Platelets and complement system

The hemostatic process comprises the activation/cleavage molecular cascade of the complement system leads to blood coagulation and platelet activation. The complement cascade in addition to coagulation function results in inflammatory cells recruitment or opsonizes pathogens to facilitate their clearance by Fc $\gamma$ R phagocytic expressing cells.

Commonly overlooked, platelets are also active actors in innate immunity. They express numerous pattern-recognition molecules including TLR4, TLR2 and TLR9 (Cognasse et al., 2015), complement and Fc $\gamma$ R (such as Fc $\gamma$ R2). Platelets act as sentinels in the circulation that potentially result in rapid innate resistance to infection. Platelets have anti-microbial activities and can directly kill microbes by releasing defensins for example or kill intraerythrocytic malaria parasites via platelet factor 4 (PF4). In the context of bacteria contact they become activated and aggregate surrounding the pathogen, leading to engulfment of the microbes by a non-killing phagocytosis-like process (see **Figure 2**).

Platelets contain a high content of  $\alpha$ -granules full of diverse mediators but those can also be newly synthesized upon activation thanks to the presence of stable transcripts acquired during the partitioning of the progenitor hematopoietic cell. These include some of the main inducers of inflammation like IL-1 and CD40L, the ligand for the costimulatory receptor CD40. A huge diversity of chemokines are also stored in platelet granules or newly synthesized such as CCL5 (RANTES), CCL3 (MIP-1 $\alpha$ ), CXCL7 ( $\beta$ -thromboglobulin or NAP-2), CXCL4, CXCL1 (GRO- $\alpha$ ), CXCL5 (ENA-78) and CCL7 (MCP3).. In addition to orchestrating leukocyte function during inflammation, platelets also contribute to tissue repair by releasing growth factors such as PDGF and TGF- $\beta$  (Mantovani & Garlanda, 2013).



**Figure 2 : Platelets and leukocytes as partners in innate immunity.**

Platelets exert direct antimicrobial activity and produce blood coagulation, proinflammatory and chemotactic molecules that participate to the recruitment and activation of leukocytes. Platelets interact with neutrophils, enhancing their activation and the generation of neutrophil extracellular traps (NETs), killing and trapping of pathogens dependent of those traps. Platelets amplifies phagocytosis in macrophages and interact with dendritic cells (DC)

*(Adapted from Mantovani and Garlanda, Nat. Immun., 2013)*

### 1.3. The adaptive immune system

The adaptive immune system is orchestrated by the lymphocytes. The two main classes are the humoral responses mediated by the B cells and T cell mediated responses. They constitute 10–30% of circulating leucocytes. The antigen-specific receptor of B or T cell goes individually through secondary and irreversible somatic rearrangements during the lifetime of the organism whereas in innate immunity, pathogen-specific receptors are already encoded in the germline of each population. The other particular specificity of the adaptive immune system is the lifetime persistence of memory cells able to mount quickly an effective response in case of new encounter with the same antigen.

This process of long lasting acquired immunity constitutes the foundation of the concept of vaccination developed by Louis Pasteur and others in the 19th century.

Nevertheless some T cell subsets are designed as 'unconventional' as they possess an invariant T cell receptor (TCR) as opposed to CD4<sup>+</sup> and CD8<sup>+</sup>  $\alpha\beta$  T cells and share characteristics of helper T cells, cytotoxic T cells and natural killer cells. Natural Killer T cells (CD1d-restricted), Mucosal-associated invariant T (MAIT) and Gamma delta T cells ( $\gamma\delta$  T cells) circulate in a pre-activated status predominantly at epithelium sites. Such properties place them at the border between innate and adaptive immunity. For example, large numbers of V $\gamma$ 9/V $\delta$ 2 T cells respond within hours to common molecules produced by microbes, and highly restricted intraepithelial V $\delta$ 1 T cells will respond to stressed epithelial cells (Vantourout and Hayday, 2013).

Like already mentioned before, all immune cells are coming from bone marrow derived progenitors. Nonetheless one of the unique features of adaptive cells is the requirement of a maturation process implying receptor rearrangement that will confer them their powerful specificity and specific selection/depletion steps to avoid the presence of autoreactive circulating cells (Boehm and Bleul, 2007).

For B cells, the bone marrow is the primary site for the development of an immature form of B cells and the final maturation undergo in lymphoid organs (spleen or lymph nodes). For the T cell compartment, the thymus is the unique organ that support the T cell development from haematopoietic stem cell (HSC)-derived lymphocyte progenitors. Its specific architecture is necessary for T cell lineage commitment and maturation (Zúñiga-Pflücker, 2004; Boehm and Swann, 2013).

In this section, I will focus on  $\alpha\beta$  T cells and B cells biology like depicted in the **Table 2**.

<b>Adaptive immunity</b> (14-47% of leucocytes)						
Cell type % of leucocytes	Protein marker <sup>a</sup> ( <i>Hasan et al. 2015</i> )	Gene marker <sup>b</sup>	Type of immune response	Major effector functions and molecules	Major regulators	PRRs expression
<b>T cells</b> 7 – 24%	CD3 <sup>+</sup>	<i>CD2</i> <i>CD3D, CD3E, CD3G</i> <i>CD6</i>				
<b>CD4+ T cells</b> (conventional) 4 – 20%	CD3 <sup>+</sup> CD4 <sup>+</sup>	<b>T helper (Th):</b> <i>ANP32B, BATF, NUP107,</i> <i>CD28, ICOS</i> <b>Th1:</b> <i>CD38, CSF2, IFNG, IL12RB2,</i> <i>LTA, CTLA4, TXB21, STAT4</i> <b>Th2:</b> <i>CXCR6, GATA3, IL26, LAIR2,</i> <i>PMCH, SMAD2, STAT6</i> <b>Th17:</b> <i>IL17A, IL17RA, RORC</i> <b>T reg:</b> <i>FOXP3</i>	Type 1, Type 2 Type 3 and Type 4	<b>. T helper function:</b> i.e. by the presentaion of co- stimulatory molecules <b>. Immunoregulation by cytokine production:</b> IL2 IFN $\gamma$ /TNF IL-4/IL-21 IL-5/IL-33 IL-17/ IL-22 IL-10	MHC-II presentation IL-2 TGF $\beta$ IL-12 IL-6 Eomes T-bet GATA3 ROR $\gamma$ t FOXP3	TLRs* NLRs* RLRs* CTLRs*
<b>CD8+ T cells</b> (conventional) 2 – 11%	CD3 <sup>+</sup> CD8 $\beta$ <sup>+</sup>	<i>CD8A</i> <i>CD8B</i> <i>FLT3LG</i> <i>GZMM</i> <i>PRF1</i>	Type 1	<b>Effector T cells recognizing cognate antigen expressing cells and promoting their apoptosis through:</b> . Cytolytic proteins (i.e. Perforin, Granzymes) - Anti-viral/-tumoral molecules: IFN $\gamma$ , TNF- $\alpha$ , CCL3, CCL4 and CCL5	MHC-I presentation IFN $\gamma$ TNF	TLRs* NLRs* RLRs* CTLRs*
<b>B cells</b> 1 – 7%	CD19 <sup>+</sup>	<i>BLK, CD19,</i> <i>CR2 (CD21),</i> <i>HLA-DOB,</i> <i>MS4A1 (CD20),</i> <i>TNFRSF17</i>	Type 1, Type 2 Type 3 and Type 4	<b>. Antibody production and presentation</b> <b>. Cytokine production</b> <b>. Lymphoid tissue organization</b>	BCR triggering  Fc $\gamma$ R	TLRs NLRs RLRs CTLRs

**Table 2 : Characteristic and function of major circulating adaptive cells**

<sup>a</sup> The surface protein indicated are the markers used to identify each population within the Milieu Interieur general panel (Hasan et al. Clin. Immun. 2015; see Table)

<sup>b</sup> Gene list commonly used to identify specific populations. The genes indicated in **blue bold** are the ones included in the human Immunology v2 codeset used in our study.

\* Expression detected at mRNA or protein level but the signal pathway resulting from the triggering of these receptors still need to be explored for those cell types.

. B cells: lymphocyte B cells ; CTLRs: C-type Lectin-like receptors ; MHC: Major histocompatibility complex; NLRs: NOD-like receptors; PRRs: Pattern rcognition receptors; RLRs: RIG-I-like receptors; T cells: lymphocyte T cells Th: T helper; Treg: regulatory T cells

### **1.3.1. T lymphocytes: the cellular response**

#### **1.3.1.1. CD8<sup>+</sup> T lymphocytes**

CD8<sup>+</sup> T cells or cytotoxic T cells (CTL) induce the death of cells that are dysfunctional (i.e. tumoral) or infected with intracellular pathogens playing then a critical role in the type 1 response (Rocha and Tanchot, 2004).

After thymus selection, the naïve CD8<sup>+</sup> T cells circulate constantly from bloodstream to lymphoid organs making continuous contact with antigen presenting cells (APC) within the lymphoid tissue. As previously described, they get primed (during a pathological condition) when they receive the three signals from the APC and antigen-specific helper CD4<sup>+</sup> T cells in peripheral lymphoid organs to become functionally and phenotypically heterogeneous populations.

Like NK cells, their innate counterpart, effector CTL release newly synthesized cytokines (i.e. IFN- $\gamma$ , TNF- $\alpha$ , MIP-1 $\alpha$ ) and granules containing pre-synthesized cytotoxic proteins (i.e. perforin, granzymes) that can induce the burst or the apoptosis pathway in the target cell (**Table 2**).

Once primed, CD8<sup>+</sup> T cells undergo robust proliferation and differentiate at the same time into effector T cells. The expansion phase is followed by contraction, marked by high levels of cell death, resulting in a small population of memory CD8<sup>+</sup> T cells.

This pool will subset into “effector memory” (TEM) and “central memory” (TCM) depending on their ability to traffic to lymphoid tissues. On a later encounter with the same antigen, these memory cells quickly differentiate into effector cells, shortening the time required to mount an effective response.

During neo-priming or reactivation, CD8<sup>+</sup> T cells undergo characteristic changes in surface molecule expression and intracellular cytokines, chemokines, transcription factors expression that will vary depending on their maturation state (Sallusto et al., 2004; Appay et al. Cytometry, 2008). The protein pattern and/or gene expression analysis allows a precise identification of these subsets and their activation status (De Rosa et al., 2001).

#### **1.3.1.2. Helper T lymphocytes**

As previously mentioned, immature bone marrow T cell progenitors develop in the thymus. During this process they are successively programmed and selected to become naïve CD8<sup>+</sup> T cells (see precedent section), naïve CD4<sup>+</sup> T cells or CD4<sup>+</sup>CD25<sup>+</sup> T-regulatory cells (Tregs).

The common future between naïve CD4<sup>+</sup> and CD8<sup>+</sup> T cells is the expression of a rearranged TCR. However instead of MHC- I recognition, TCRs from CD4<sup>+</sup> T cells recognize antigen bound to MHC-II molecules on APCs that lead to differentiation into distinct types of effector T cells, depending on the ‘signal 3’ they receive in the secondary lymphoid organs (see DCs and antigen presentation chapter). Like for CD8<sup>+</sup> T, naïve CD4<sup>+</sup> T-cells (TN) will progress through central memory (TCM) T-cells and then to effector memory (TEM) T-cells and finally to terminally differentiated effector memory (TEMRA) T-cells.

As mentioned previously, depending on the specific inflammatory conditions there is a huge diversity in the T helper population (Th) defined by their production of specific patterns of cytokines/chemokines that will be associated to the different types of responses (see section “Major type of responses”). Those patterns of expression are tightly regulated by now well-defined genes and transcriptions factors. The major subsets described are the Th1, Th2, Th17, Th22, Th9 and Tregulators (reviewed in Baranovski et al., 2015).

In this section, I will describe briefly those specific T helper subsets and their principal functions and molecular pattern (see **Figure 3** and **Table 2**). Of note, it’s now well established that distinct subsets can regulate each other’s differentiation (Baranovski et al., 2015).

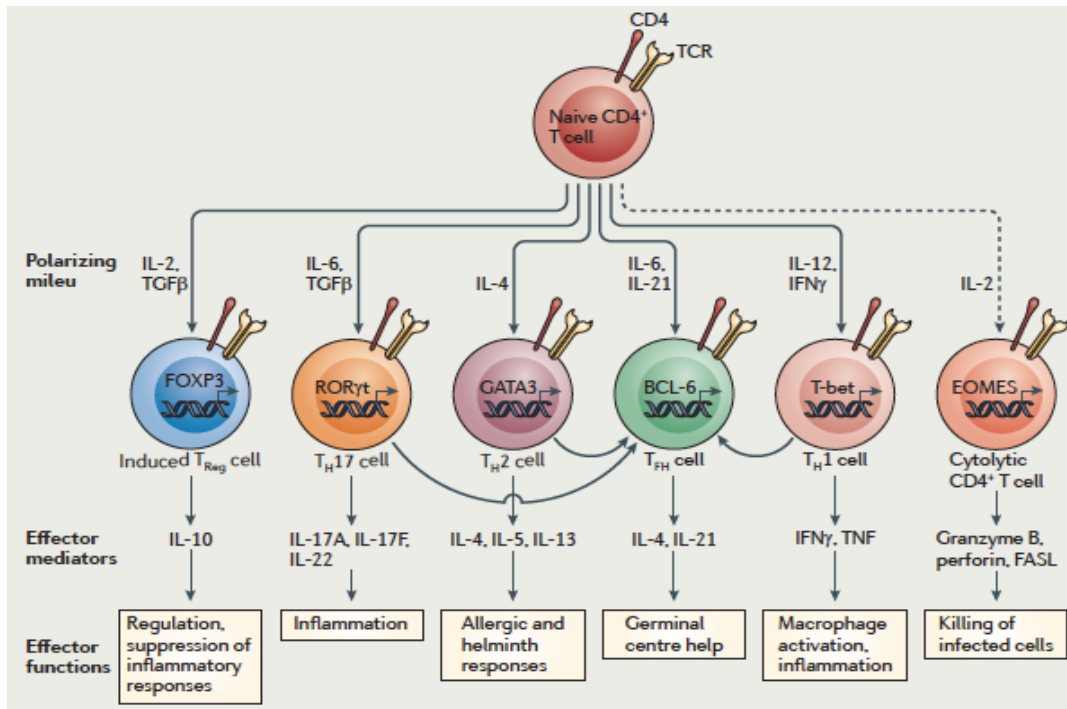
Th1 differentiation: IL12 and interferon- $\gamma$  (IFN- $\gamma$ ) are the critical cytokines initiating the signaling cascade to develop Th1 cells (Szabo et al., 1995). Several transcription factors induce full differentiation of the Th1 cells with the T-box transcription factor (T-bet) as the master regulator, signal transducer and activator of transcription 1 (STAT1), STAT4 and Eomesodermin (Eomes).

Th2 differentiation is favored by IL-4 signaling which will induce IL-5, IL-13 and new IL- 4 release (Sornasse et al., 1996). The master regulator for full differentiation of Th2 cells is GATA3 in coordination with STAT5.

More recently described, the Th17 cells were found to be induced early in response to extracellular bacteria (type 3 response), depending on IL-6 and TGF- $\beta$  for their differentiation and IL-23 for further proliferation, and seem to be involved in stimulating the neutrophil response. The retinoic acid receptor-related orphan receptor gamma-T (ROR $\gamma$ t) (encoded by the gene RORC in human) is the master regulator involved in their differentiation.

Interestingly, exposure to the same TGF- $\beta$  but in the absence of IL-6, IFN- $\gamma$  and IL-12 favors the development of CD4<sup>+</sup>CD25<sup>+</sup>CTLA4<sup>+</sup> Treg cells by the induction of the regulator forkhead box P3 (FoxP3). Those Treg are referred as inducible Treg (iTreg) in contrast to the natural

Treg coming from the thymus. This subset will restrain the immune response by producing inhibitory cytokines such as IL-10 and TGF- $\beta$  (Chen et al., 2003).



**Figure 3 : Subsetting CD4<sup>+</sup> T cell responses based on T<sub>H</sub> cell polarization**

Following recognition of a specific antigen presented by an appropriately activated antigen-presenting cell, naive CD4<sup>+</sup> T cells undergo several rounds of division and can become polarized into distinct effector T helper (T<sub>H</sub>) cell subsets that differentially orchestrate protective immune responses. The differentiation of polarized effector T cells is controlled by unique sets of transcription factors. Their expression is determined by multiple signals but particularly by soluble factors that act on CD4<sup>+</sup> T cells during their activation.

(Adapted from Swain et al. *Nat. Rev. Immunol.*, 2012)

### 1.3.2. B lymphocytes: the humoral adaptive response

Upon activation, B cells differentiate into plasma cells, which proliferate and are able to secrete antibodies, themselves driving protective responses through neutralization, opsonization and complement activation (humoral response).

B lymphocytes display three hallmarks not commonly shared by all immune cells: they express a unique B-cell receptor (BCR) for antigen on their membranes, falling in the adaptive immunity definition; they express CD40 on their surfaces sharing common properties with most APCs, but interestingly, they abundantly express PRRs, which is considered a hallmark of innate immune cells. Indeed, human B-cell subsets express distinct PRRs, including receptors for immunoglobulins (Ig) such as Fc $\gamma$ RIIb/CD32 or Fc $\epsilon$ RII (CD23), complement receptors and TLRs. In contrast to T cells, the triggering of those

constitutively expressed TLRs induce well described functions as class switching recombination or proliferation, differentiation, Ig production depending of the B cell subsets targeted (Garraud et al., 2012).

The capacity of specific B-cell subsets to traffic throughout the body is essential for sampling pathogens and for their APC functions. Accordingly, distinct programs of chemokine receptor expression were also ascribed to the various B-cell subsets (Bowman et al., 2000; Garraud et al., 2012; Nutt et al., 2015).



## 2. Inflammation

After tissue injury or infection a rapid inflammatory response is initiated through the recognition of molecules shed from the outer membrane of invading pathogens as well as danger signals from internal environment (i.e. alarmins from dying cells) by pattern recognition receptors (PRR) expressed on both immune and non-immune cells. Engagement of these receptors triggers the secretion of cytokines, chemokines and other biologically active molecules. The inflammatory cytokines produce changes in the adhesive properties of the endothelial cells, in turn causing circulating inflammatory leukocytes to stick and transmigrate to the site of infection, thereby initiating the process of inflammation.

The hallmark of **acute inflammation** (restricted in time) is the considerable increase (locally or systemically) in acute phase serum proteins such as LPS binding protein (LBP), C reactive protein (CRP), and serum amyloid A (Gabay et al., 1999). This acute phase generally persists only when the stimulus is present. Infected/injured resident cells secrete cytokines that induce rapid and short-lived constriction of blood vessels and chemokines to coordinate the recruitment of neutrophils to the site of injury to remove the invading pathogen. As homeostasis is restored, anti-inflammatory mediators such as IL-4 and lipoxins promote resolution by inhibiting the migration of neutrophils as well as inducing the migration of monocytes to the site of injury. This influx of monocytes helps to resolve inflammation by phagocytosis of apoptotic neutrophils (reviewed by Serhan, 2007). Of note, the time restriction of the acute phase definition depends of the pathologic situation. For example in the case of a slight injury, acute proteins will arise and decrease within minutes or hours, whereas in the case of high antigen concentration during acute phase such as some viral infections characterized by high viral replication during primo-infection (i.e. hepatitis C (HCV) or human immunodeficiency virus (HIV) infections), the acute phase duration occurs during months before leading to the chronic status.

A controlled inflammation resolves with inconsequential damage to the host. However, the loss of control of inflammation can result in chronic disorders commonly observed in clinic such as rheumatoid arthritis, inflammatory bowel disease, and septic shock and can ultimately lead to irreversible damage to tissues and organs and, even death.

**Chronic inflammation** can develop in the case of pathogen persistence (i.e. chronic infection (HCV, HIV)), phagocytic cell inability to degrade deleterious substances, or autoimmune responses (Serhan, 2007) leading to a persistent inflammatory cascade (Kitchens et al., 2003).

In the context of biomarker discoveries and clinical follow-up, many diseases are characterized by persistent cytokine expression (i.e. IL-1 and TNF) and measurement of these molecules can be helpful for medical intervention. However while cytokines play a similar role in many of these diseases, the source and the induced pathogenesis can be extremely variable, adding a layer of difficulty in the design of effective therapies to combat inflammatory disorders (Forrester et al. 2005; Forrester and Libby, 2007).

### 3. Inflammatory cytokines

Cytokines are soluble mediators of cell communication that are critical in immune regulation. These proteins and glycoproteins can be produced by hematopoietic and non-hematopoietic cells and induce specific gene expression programs in responsive cells via autocrine and paracrine mechanisms.

Inflammation results from a complex phenomenon and its understanding requires an analysis of its fundamental mediators. . In our whole blood stimulation model, we have used cytokine mediators to mimic differentially polarized inflammatory environments to focus on specific pathways induced in the whole blood cellular network. In this section I will describe briefly the biology of the key pro-inflammatory cytokines and their receptors used in our whole blood stimulation model (**Table 3**).

TNF, IL-1 $\beta$ , IFN- $\beta$  and IFN- $\gamma$  are pro-inflammatory cytokines rapidly induced within minutes to hours after an inflammatory insult, followed by a rapid clearance (DeForge and Remick, 1991).

Cytokine	Gene name	Main sources	Receptor Gene Name	Immune Target cell	Major function
<b>Interleukins</b>					
IL-1	<i>IL1A, IL1B</i>	Macrophages, Monocytes, Neutrophils, Bcells, DCs, platelets	<i>IL1R1, IL1R2, IL1RAP, IL1RN</i>	B cells, NK cells, T cells	Pyrogenic, pro-inflammatory, proliferation and differentiation, (i.e. Th17)
IL-23	<i>IL23A, IL12B</i>	Macrophages, DCs	<i>IL23R</i>	T cells, Neutrophils	Th17 differentiation
<b>Tumour necrosis factor</b>					
TNF- $\alpha$	<i>TNF</i>	Macrophages, monocytes	<i>TNFRSF1A, TNFRSF1B,</i>	Macrophages, Tumour cells	Phagocyte cell activation, endotoxic shock Tumour cytotoxicity, cachexia
<b>Interferons</b>					
IFN- $\alpha$	<i>IFNA1/13, IFNA2 up to IFN21</i>	Leukocytes, pDCs	<i>IFNAR1, IFNAR2</i>	Various	Anti-viral
IFN- $\beta$	<i>IFNB1</i>	Fibroblasts	<i>IFNAR1, IFNAR2</i>	Various	Anti-viral, anti-proliferative
IFN- $\gamma$	<i>IFNG</i>	T cells	<i>IFNGR1</i>	Various	Anti-viral, macrophage activation, increases neutrophil and monocyte function, MHC-I and -II expression

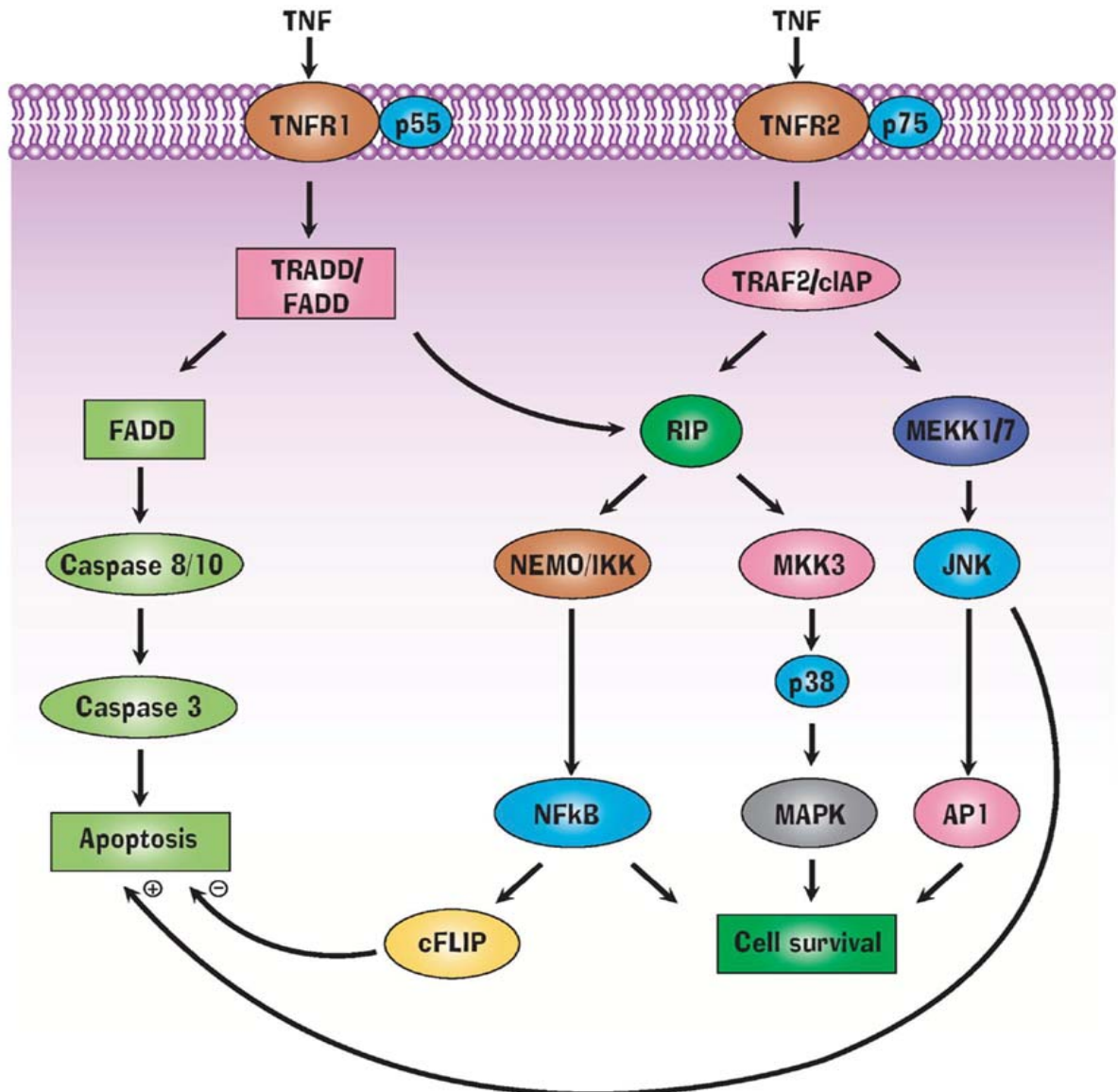
**Table 3 : Expression and function of major inflammatory cytokines**

The genes indicated in *blue bold* are the ones included in the human Immunology v2 codeset used in our study.

### 3.1. Tumor necrosis factor

Tumor necrosis factor (TNF) is a classic proinflammatory cytokine secreted by a number of cells including macrophages/monocytes, mast cells, DCs and tumor cells (Spriggs et al. 1987; Kelsa, 1998). TNF is induced in response to a variety of stimuli including bacterial endotoxins, oxygen radicals, and viruses. (Chensue et al., 1991; Dubravec et al., 1990). TNF induces the anti-microbial function of innate cells but is also a mitogen for T and B cells. This pleiotropic cytokine has been described to participate in a myriad of functional activity including cell growth modulation, inflammation, apoptosis, tumorigenesis, viral replication, septic shock and autoimmunity (Aggarwal et al., 2012). In addition to being produced in response to direct stimulation, TNF production can also induce various other cytokines including IL-1 $\beta$  and IL-8 (Cassatella et al., 1993; DeForge et al., 1992).

TNF can bind TNF receptor 1 (TNFR1 encoded by *TNFRSF1A*) that enhances the respiratory burst in neutrophils, increases IL-6 mRNA, and induces NF $\kappa$ B activation (Tartaglia et al., 1993; Mackay et al., 1993; Kruppa et al., 1992). This receptor is ubiquitously expressed explaining its diverse effects. TNFR1 triggering leads to apoptosis pathway engagement by activation of the adaptor proteins TNFR1-associated death domain (TRADD) and Fas-associated death domain (FADD). The second known receptor, TNFR2 (encoded by *TNFRSF1B*) induces a signaling after TNF binding, that relies on TNF receptor associated factor 2 (TRAF2) and activation and nuclear entry of NF $\kappa$ B. The TNFR2 expression is more limited to certain populations of lymphocytes, including T-regulatory cells (Tregs), endothelial cells, some cells from the central nervous system, cardiac myocytes, thymocytes, islets of Langerhans and human mesenchymal stemcells (reviewed by Faustman and Davis, 2013). Depending on the balance between those two receptors expression and triggering by TNF, in addition to other factors, the cell fate will be determined leading to apoptosis (TNFR1 signaling) or survival (TNFR2 signaling) (Aggarwal et al., 2012; Faustman and Davis, 2013; **Figure 4**). Both receptors can occur in soluble form. Specifically, the membrane protease TACE (TNF alpha converting enzyme) can cleave immature, membrane-bound TNFR resulting in soluble, biologically active TNFR in the circulation (Locksley et al., 2001).



**Figure 4 : TNF Signals through TNFR1 and TNFR2 receptors.**

Triggering of TNFR1 leads to the recruitment of TRADD, FADD, and caspase-8, that activates caspase-3 resulting in apoptosis.

In contrast TNFR2 does not have a cytoplasmic death domain but a TRAF2 binding site. TRAF2, in turn, binds RIP, TAK1, and IKK, leading to cell survival when NFκB is liberated from its inhibitor protein IκBa in the cytoplasm and translocates to the nucleus.

After TNFR1 triggering an alternative pathway relies on TRADD's activation of RIP resulting in cell survival.

Exposure of cells to TNFα in most cases results in the generation of reactive oxygen species, leading to activation of MKK7 and JNK. The activation of ERK and p38MAPK is via TRADD, TRAF2, RIP, TAK1, and MKK3/6.

*(Adapted from Faustman and Davis. Front. Immun., 2013)*

### 3.2. Interleukin-1β

Interleukin 1β is a member of the IL-1 family cytokines. The major members of this family are IL-1α, IL-1β, IL18, IL33 and IL-1 receptor antagonist (IL-1ra) (reviewed by Netea et al.,

2015). The IL-1 genes are induced by a wide variety of stimuli including LPS, viruses, and TNF (Auron et al., 1984). Like IL-6 and TNF, IL-1 is most prominently produced by monocytes/macrophages (Chensue et al., 1991). Upon endotoxin stimulation, for example, IL-1 $\beta$  is rapidly induced in whole blood (Lang et al. Shock, 2003). Although IL-1 $\alpha$  and IL-1 $\beta$  are both induced upon stimulation of the inflammatory response, IL-1 $\alpha$  remains in the cytosol, whereas IL-1 $\beta$  is processed and cleaved into its active form by IL-1 $\beta$  converting enzyme (ICE also referred as caspase-1) (Chin et al. JI, 1993). Infection with bacteria induces activation of caspase-1, which catalyzes the processing of pro-IL-1 $\beta$  to produce the mature cytokine. A complex of proteins responsible for these catalytic processes has been purified and designated as the inflammasome (Martinon et al., 2002 and see section “*NOD-Like receptors*” below).

Direct stimulation with IL-1 $\beta$  can activate the transcription of specific inflammatory genes such as TNF and IL-8 (Akira et al., 1990).

Two IL-1 receptors (IL-1R) can bind IL-1 $\beta$  as well as IL-1 $\alpha$  with the same signaling properties. The pro-inflammatory IL-1RI (Sims et al., 1993) is found predominantly on T cells and fibroblasts and transmits a downstream signaling cascade similar to that observed with TLR binding (see **Figure 6** on section “PRRs” below; Medzhitov, et al., 1997).

Interestingly, the IL-1ra acts as an anti-inflammatory cytokine by binding IL-1RI with a high avidity but without inducing signal. In addition this high avidity blocks future binding of IL-1 $\alpha$  and IL-1 $\beta$  (Antin et al., 1994; Dinarello, 1994).

The second receptor IL-1RII on the other hand, acts as a decoy receptor (Symons et al., 1995) and is expressed on activated T cells, B cells, monocytes, and neutrophils (Giri et al., 1990). IL-1RII occurs as both a membrane-bound and soluble receptor (sIL-1RII) (Sims et al., 1994).

### 3.3. Type I Interferons

Type I interferons (IFNs) are polypeptides secreted by infected cells or upon PRR activation. They induce cell-intrinsic antimicrobial factors in infected and neighbouring cells to limit the spread of infectious agents, particularly viral pathogens. They also modulate innate immune responses by inducing the secretion of IL12 and IL18 to promote antigen presentation and natural killer cell functions. This arm aims also to restrain pro-inflammatory pathways and cytokine production. Finally, they activate the adaptive immune system, thus promoting the development of high-affinity antigen-specific T and B cell responses and immunological memory. Type I IFNs are protective in acute viral infections but can have either protective or deleterious roles in bacterial infections (Trinchieri, 2010, Ivashkiv and Donlin, 2014).

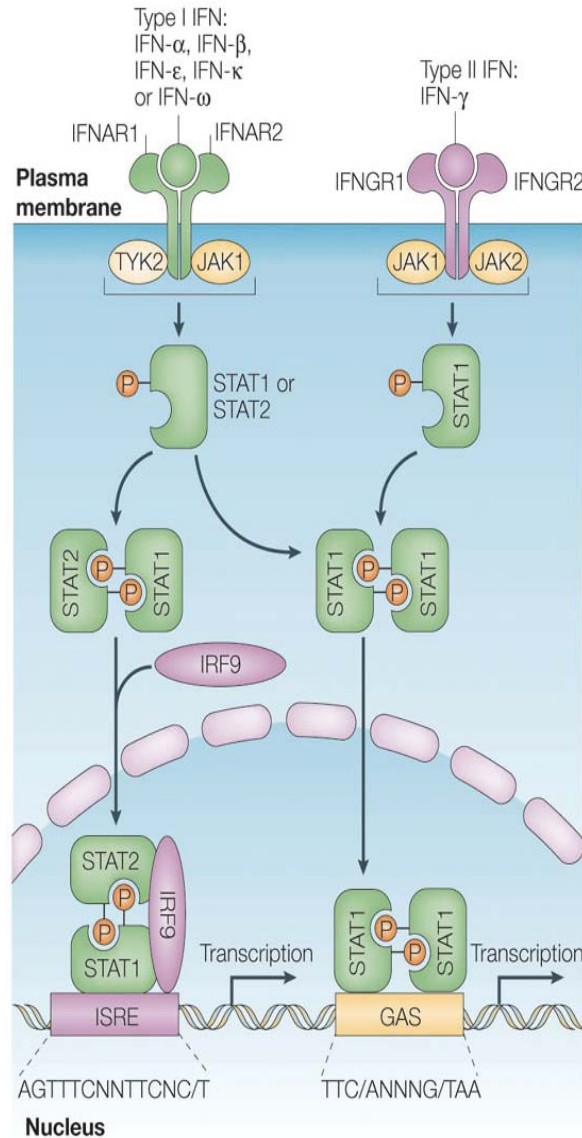
IFN $\alpha$  and IFN $\beta$  bind the ubiquitous IFN- $\alpha$  receptor (IFNAR) a heterodimeric transmembrane receptor composed of IFNAR1 and IFNAR2 subunits. Canonical type I IFN signaling activates the Janus kinase (JAK)–STAT pathway that will complex with IFN-regulatory factor 9 (IRF9) into the nucleus leading to transcription of IFN-stimulated genes (ISGs) (see **Figure 5** in the next section).

### 3.4. Interferon Gamma

Interferon Gamma (IFN- $\gamma$ , also referred to as a Type II IFN) is induced by a unique set of stimuli and is produced by T lymphocytes and natural killer (NK) cells. Interestingly, viral infection of these cells does not directly induce IFN- $\gamma$  production.

In contrast to type I IFNs, IFN- $\gamma$  is primarily an immunomodulator that also can exert to a lesser extent some antiviral activity. It can be induced by antigen presentation through MHC-I or MHC-II for CD8<sup>+</sup> T cells or CD4<sup>+</sup> T cells respectively. In vitro, a direct TCR stimulation by anti-CD3 or mitogen (i.e. Streptococcus Enterotoxin B) can also induce its production. One of the major physiologic roles of IFN- $\gamma$  is its ability to regulate MHC-I and MHC-II protein expression (notably on immune cells). It participates also in the inflammatory response notably by enhancing TNF production and activity.

In whole blood, except for the red blood cells, all the cells (including the platelets) express the single IFN- $\gamma$  receptor (IFNGR) (Farrar and Schreiber, 1993; Plataniias, 2005). The IFNGR triggering by IFN- $\gamma$  induces a JAK1/JAK2-STAT1 pathway that will activate IFN- $\gamma$  binding sites (GAS) in the DNA leading to the transcription of the related IFN stimulated genes (ISG) (see **Figure 5**).



**Figure 5 : Interferon receptors and activation of classical JAK–STAT pathways by type I and type II interferons.**

Type I interferons (IFNs) bind a common receptor at the surface of human cells, the type I IFN receptor composed of two subunits, IFNAR1 and IFNAR2, which are associated with the TYK2 and JAK1, respectively.

The only type II IFN, IFN- $\gamma$ , binds a distinct cell-surface receptor, the type II IFN receptor composed of two subunits, IFNGR1 and IFNGR2, which are associated with JAK1 and JAK2, respectively.

Activation of the JAKs that are associated with the type I IFN receptor results in phosphorylation of STAT2 and STAT1; this leads to the formation of STAT1–STAT2–IRF9 complexes. These complexes translocate to the nucleus and bind ISREs in DNA to initiate gene transcription.

Both type I and type II IFNs also induce the formation of STAT1–STAT1 homodimers that translocate to the nucleus and bind GAS elements that are present in the promoter of certain ISGs, to initiate the transcription of these genes. The consensus GAS element and ISRE sequences are shown (N, any nucleotide).

GAS= IFN- $\gamma$ -activated site ; IFNAR= type I IFN receptor; IFNGR= type II IFN receptor; IRF= IFN-regulatory factor; ISREs= IFN-stimulated response elements; JAK= Janus activated kinases; STAT= signal transducer and activator of transcription; TYK= Tyrosine kinase

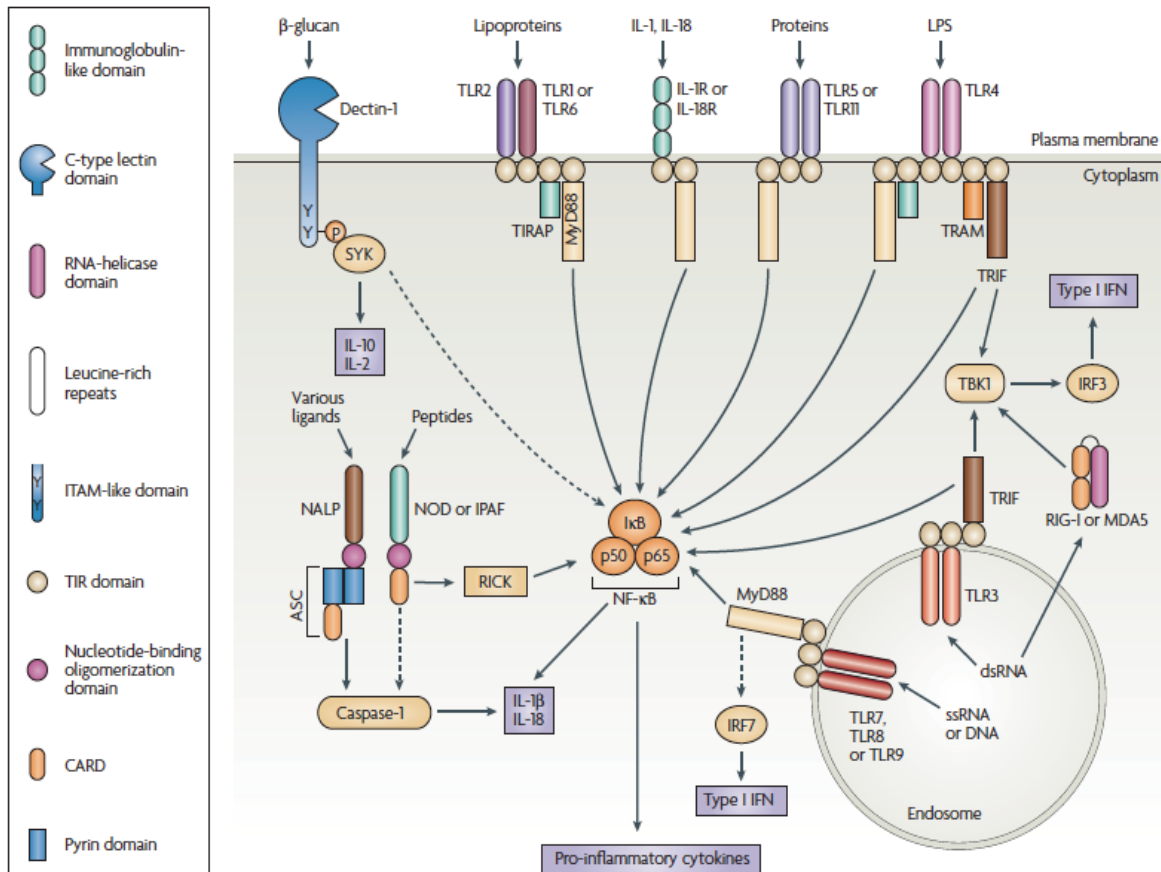
(Adapted from Platanius L. C. *Nat. Rev. Immun.*, 2005)



## 4. Pattern recognition receptors

An effective innate immune response depends largely on its ability to rapidly detect and react against a broad range of foreign pathogens (Janeway, 1989). To distinguish infectious nonself- molecules from self-molecules innate immune cells expressed a large variety of pattern recognition receptors (PRRs) which bind to conserved molecular structures, found on a variety of pathogens, called pathogen-associated molecular patterns (PAMPs) (Medzhitov and Janeway, 1997; Akira et al., 2006). PRRs can also recognize immunostimulatory products that are derived from damaged tissue or necrotic cells, termed damage-associated molecular patterns (DAMPs), and this recognition is crucial for host defence and tissue remodeling (Medzhitov, 2007).

Currently, four major families of PRRs have been described— Toll-like receptors (TLRs), RIG-I-like receptors (RLRs), NOD-like receptors (NLRs) and C-type lectin-like receptors (CTLRRs)—that allow the recognition of a large range of PAMPs, including proteins, nucleic acids, lipids and carbohydrates, derived from foreign microorganisms (see **Figs. 6** and **7**). Interestingly, those receptors can cooperate to act synergistically or to modulate each others against danger signals (Trinchieri and Sher, 2007).



**Figure 6 : Schematic representation of receptor structures and main signaling pathways of the PRR families.**

In this figure are represented the receptor structures with the adaptor molecules and the main signaling pathways that differentiate different classes of pattern recognition receptors (PRRs).

ASC, apoptosis-associated speck-like protein containing a CARD (caspase-recruitment domain) ; ds, double-stranded ; IFN, interferon ;  $\kappa$ B, inhibitor of NucleoFactor $\kappa$ B ; IL, interleukin ; IPAF, ICE-protease-activating factor ; IRF, IFN-regulatory factor ; LPS, lipopolysaccharide ; MDA5, melanoma-differentiation-associated gene 5 ; MyD88, myeloid differentiation primary-response gene 88 ; NALP, NACHT-LRR- and pyrin-domain-containing protein ; NOD, nucleotide-binding oligomerization domain ; RICK, receptor-interacting serine/threonine kinase ; RIG-I, retinoic-acid-inducible gene 1 ; ss, single-stranded ; TBK1, TANK-binding kinase 1 ; TIRAP, Toll/IL-1R (TIR)-domain-containing adaptor protein ; TRAM, TRIF-related adaptor molecule ; TRIF, TIR-domain-containing adaptor protein inducing IFN $\beta$  ; Syk, spleen tyrosine kinase.

(Adapted from Trinchieri and Sher. *Nat. Rev. Immun.*, 2007)

#### 4.1. Toll-like receptors

Toll-like receptors (TLRs) are evolutionarily conserved transmembrane receptors with a leucine-rich extracellular domain involved in ligand recognition and exist as dimeric proteins (heterodimers or homodimers). Furthermore, its cytosolic Toll/IL-1 receptor-like (TIR) domain transmits the intracellular signal in response to ligand binding (Akira and Takeda, 2004). This signaling cascade leads to the activation of nuclear factor  $\kappa$ B (NF- $\kappa$ B), activator

protein 1 (AP1), IRF3, and other transcription factors, driving the production of proinflammatory cytokines, maturation of dendritic cells, and other immunologic responses (see **Figure 6**).

In human, 11 TLRs have been described so far, commonly divided into two subgroups depending on their cellular localization. The first group is composed of TLR1, TLR2, TLR4, TLR5, TLR6 and TLRs 10–13, which are expressed on cell surfaces and recognize mainly microbial surface components. The second group includes TLR3, TLR7, TLR8 and TLR9 that are preferentially expressed in the endoplasmic reticulum in resting cells and rapidly traffic to endolysosomes after stimulation by microbial nucleic acids. Ligation of TLRs by PAMPs induces a tightly controlled cascade of intracellular signaling molecules, leading to the expression of NF- $\kappa$ B-dependent pro-inflammatory cytokines or IRF-dependent type I IFNs (DiDonato et al., 2012).

## 4.2. RIG-I-like receptors

RIG-I like receptors (RLRs) are a family of cytoplasmic RNA helicases, including retinoic acid-inducible protein I (RIG-I) (Pichlmair et al., 2006) and melanoma differentiation-associated protein 5 (MDA5; also known as IFIH1) (Kato et al., 2006) that are essential for innate recognition of viruses and production of type I IFNs through a STING/IRF3 dependent axis to control viral replication and dissemination.

RIG-I is an IFN-inducible protein containing CARDs and a DExD/H box helicase domain and has been identified as a cytoplasmic single (ss) and double (ds) strand RNA detector (**Figure 7**) (Yoneyama et al., 2004).

MDA5, a molecule showing homology to RIG-I, has also been implicated in the recognition of viral dsRNA (Andrejeva et al., 2004; Kang et al., 2002). In addition, these proteins bind the synthetic form, poly I:C.

In addition, other molecules or complexes have been recently discovered to sense cellular RNA or DNA (Sun et al., 2012). For example, the DNA sensor cyclic guanosine monophosphate-AMP synthase (cGAS) induces IFN type I production through the same STING/IRF3 dependent axis (**Figure 7**) (Watson et al., 2015).

### 4.3. NOD-like receptors (NLRs)

NLRs consist of a large group of intracellular PRRs, such as nucleotide-binding oligomerization domain proteins (NODs) and NOD-, LRR- and pyrin domain-containing proteins (NLRPs), that are crucial for host defense against bacterial infection (Philpott et al., 2014).

Among the large number of NOD-LRR family members, NOD1 and NOD2 are the best described and both contain N-terminal CARD domains. NOD1 and NOD2 are cytosolic proteins that respond to intracellular fragments of bacterial peptidoglycan. A missense point mutation in the human NOD2 gene is correlated with susceptibility to Crohn's disease, an inflammatory bowel disease. Ligand binding to NOD1 and NOD2 causes their oligomerization and initiates nuclear factor- $\kappa$ B (NF- $\kappa$ B)-dependent and mitogen-activated protein kinase (MAPK)-dependent gene transcription through the recruitment of RIP2/RICK, a serine/threonine kinase, to the NODs via their respective CARD domains (**Figure 7**).

Some of these sensors are associated to specific complexes called inflammasomes. One of the most studied is the NOD-like receptor family member NOD-, LRR- and pyrin domain-containing 3 (NLRP3). NLRPs form multiprotein inflammasome complexes consisting of an NLRP (or other PRRs), the adaptor ASC and pro-caspase 1. Activation of inflammasomes results in the autocatalytic processing of pro-caspase 1 into its active form, which then cleaves pro-IL-1 $\beta$  and pro-IL-18 into IL-1 $\beta$  and IL-18, respectively, which can then be secreted (see section *IL-1 $\beta$*  above). In addition, inflammasome activation can elicit pyroptosis, a specific form of cell death, of infected cells.

### 4.4. C-type lectin-like receptors (CTLRs)

C-type lectin-like receptors (CTLRs) are a large family of protein receptors (more than 1000 members) crucial in the recognition of self- and nonself-ligands (i.e. carbohydrates). Ligands bind to the conserved C-type lectin-like domains (CTLDs). After triggering, CTLRs can mediate diversified downstream responses like uptake of microorganisms, homeostatic clearance of apoptotic cells, adaptive immune response polarization through cytokine and chemokine expression or cell–cell adhesion (reviewed in Plato et al., 2013).

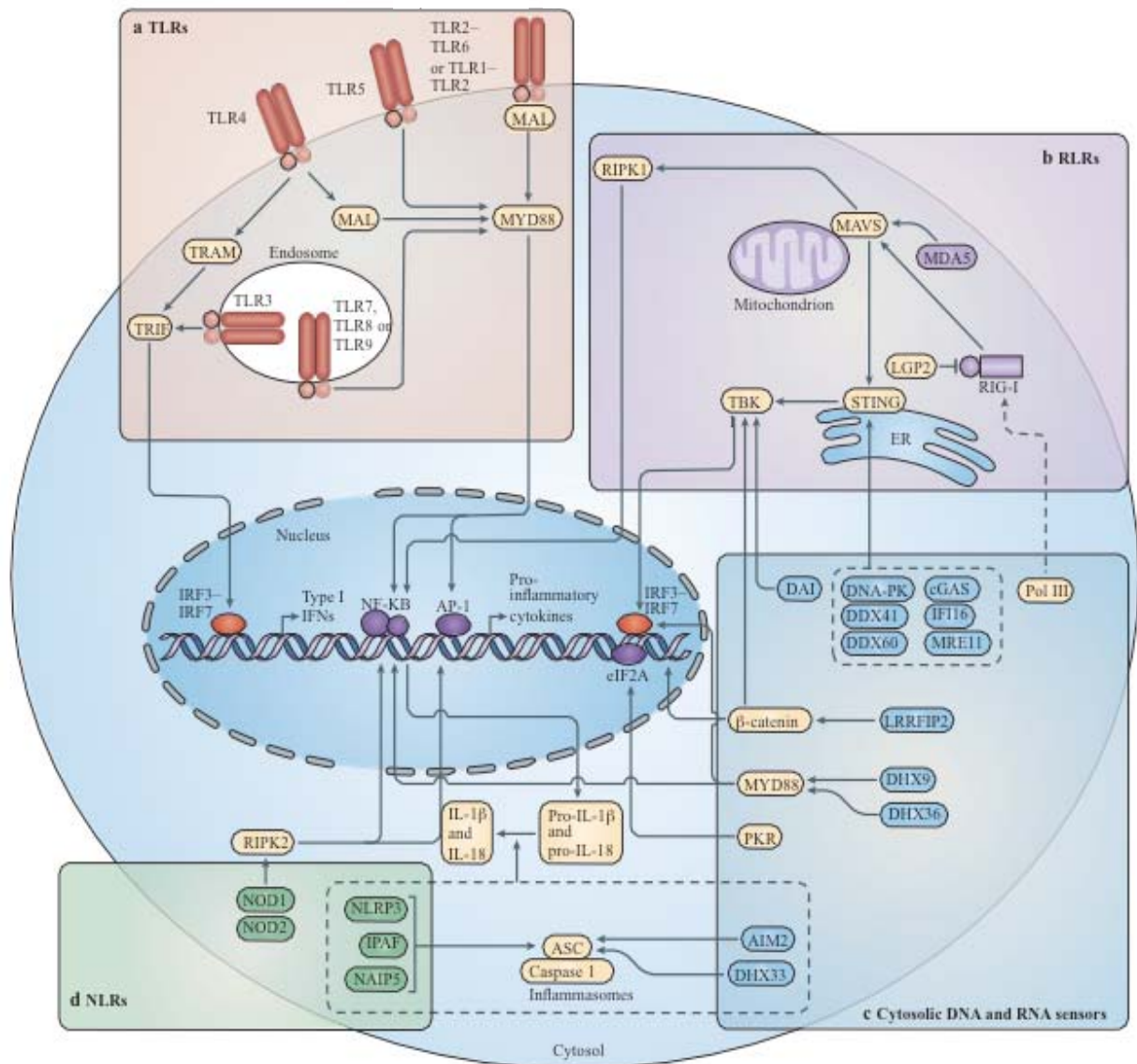
CTLRs can signal directly, through integral motifs in their cytoplasmic tails, or indirectly, through association with adaptor chains. Most activation receptors associate with ITAM-bearing adaptor chains. This is the case for the ones included in the well-studied Dectin-1 cluster (including notably CLEC-1, CLEC-2, Dectin-1 (or CLEC-7A), CLEC-9A) that resides within the natural killer gene complex (NKC). These receptors can be found mostly on

myeloid cells such as DCs, macrophages and neutrophils but also on platelets (Suzuki-Inoue et al. 2007). Multiple pathways downstream of Dectin-1 rely on Syk (**Figure 6**) leading to the control of several axes such as calcineurin/NFAT; ROS/NLRP3 inflammasome; and the phosphorylation of CARD9-Bcl10-Malt1 complex which in turn activates NF- $\kappa$ B. (Tsang et al., 2008; Rawlings et al., 2006).

Others CTLRs members signal through motifs not matching with ITAM, ITAM-like or ITIM sequences and the underlying signaling pathway is poorly understood for most of them with some exceptions such as DC-SIGN (Svajger et al., 2010).

A major general component of a PRR-induced innate immune response is the production of proinflammatory cytokines and interferons (IFN). PRR activation also can initiate phagocytosis, autophagy, cell death, and cytokine processing (Pichlmair et al., 2006; DiDonato et al., 2012).

In addition to inducing different pro-inflammatory molecules pattern, the PRRs can present evolutionary genetic polymorphisms across individuals due to differential pathogen's exposure. These combined factors (among others) can induce heterogeneity across individuals in the intensity and/or the quality of the response upon exposure to the same pathogen.



**Figure 7 : Cellular localization of major PRRs families**

Pattern-recognition receptors (PRRs) share intracellular pathways that lead to the production of pro-inflammatory cytokines and type I interferons (IFNs).

(Adapted from Cao X. *Nat. Rev. Immunol.*, 2016)

## 5. Microbial recognition by PRRs

To evaluate this heterogeneity across healthy individuals, using our whole blood model, we have induced an inflammatory environment by targeting specific PRRs using purified molecular patterns (PRRs ligands) or concomitantly several PRRs by using whole microbes to be able to deconvolute which pathways were participating to the final induced inflammation.

In the coming section, I will describe briefly the different components of microbe 's that can trigger PRRs signaling. The list of the clinically relevant stimuli used in our study with associated triggers is summarized in the **Table 4** below to help the reader to connect with the whole blood model.

	Abbreviation	Concentration	Supplier	Sensor or Receptor	Reference
	Null		Ø	NA	
<b>Microbe</b>					
HK <i>E. coli</i> 0111:B4	HKEC	complex	Invivogen	complex	Takeuchi et al., 1999
HK <i>S. aureus</i>	HKSA	complex	Invivogen	complex	Krishna and Miller, 2012
HK <i>L. rhamnosus</i>	HKLR	complex	Invivogen	complex	Miettinen et al., 2008
BCG (Immucyst)	BCG	complex	Sanofi Pasteur	complex	Means et al., 1999; Godaly and Young, 2005; Randhawa et al., 2011
HK <i>H. pylori</i>	HKHP	complex	Invivogen	complex	Zhao et al., 2007
HK <i>C. albicans</i>	HKCA	complex	Invivogen	complex	Brown et al., 2003; Gantner et al., 2003
Influenza A virus (live)	IAV	complex	Charles Rivers	complex	Diebold et al., 2004; Kato et al., 2006; Ichinohe et al., 2009; Allen et al., 2009
Sendai virus (live)	SeV	Rig-I and Mda/5	Charles Rivers	Rig-I and Mda/5	Yoneyama et al., 2005; Kato et al., 2005
<b>MAMP</b>					
C12-iE-DAP	DAP	NOD1	Invivogen	NOD1	Chamaillard et al., 2003
CPPD	CPPD	NLRP3 and TLR2	Invivogen	NLRP3 and TLR2	Liu-Bryan et al., 2005; Martinon et al., 2006
FSL-1	FSL	TLR2/6	Invivogen	TLR2/6	Shibata et al., 1997; Okusawa et al., 2004
Poly I:C	pIC	TLR3	Invivogen	TLR3	Alexopoulou et al., 2001
LPS-EB (ultrapure)	LPS	TLR4	Invivogen	TLR4	Poltorak et al., 1998; Shimazu et al., 1999
Flagellin-ST	FLA	TLR5	Invivogen	TLR5	Hayashi et al., 2001
Gardiquimod	GARD	TLR7	Invivogen	TLR7	Hemmi et al., 2002
R848	R848	TLR7 and TLR8	Invivogen	TLR7 and TLR8	Jurk et al., 2002
ODN 2216	ODN	TLR9	Invivogen	TLR9	Hemmi et al., 2000; Krieg, 2002
lipoarabinomannan	LAM	Mannose R, CD36	Invivogen	Mannose R, CD36	Jo' zefowski et al., 2011; Sieling et al., 1995
WGP	WPG	Dectin-1	Invivogen	Dectin-1	Goodridge et al, 2011

**Table 4: Complex and purified PRR Stimuli used in the Whole-Blood stimulation systems**  
(Adapted from Duffy et al., Immunity 2014)

## 5.1. Bacterial Recognition by PRRs

### 5.1.1. Bacterial external components

The outer membrane of the double membrane envelope of Gram-negative and Gram-positive bacteria contains numerous components that bind PRRs.

The endotoxin LPS, or lipopolysaccharide, is the main component of the outer membrane of Gram-negative bacteria (i.e. *E.coli*) and contains a Lipid A portion recognized by TLRs (Miller et al., 2005) which induce an intense immune response that can lead to lethal septic shock. Due to this property, LPS is frequently used as an exogenous stimulatory inducer in both *in vivo* and *ex vivo* model systems (see **Table 4**). LPS sheds from the bacteria wall, associates with LPS binding protein (LBP), an acute-phase protein present in the bloodstream, and then binds to CD14, expressed preferentially on the cell surface of monocytes. LPS is

then transferred to MD-2, which associates with the extra-cellular portion of TLR4, followed by oligomerization of TLR4, a key molecule of LPS signaling (Poltorak et 1998; Shimazu et al., 1999).

Lipoproteins and peptidoglycan (PG) are other strong immunogenic components present in both Gram-positive (i.e. *S. aureus*) and Gram-negative bacteria. TLR2 plays a major role in detecting Gram-positive bacteria and is involved in the recognition of a variety of microbial components, including lipoteichoic acid (LTA), lipoproteins, and PG. TLR2 forms heterodimers with TLR1 or TLR6, which allows the discrimination of subtle changes in the lipid portion of lipoproteins (Alexopoulou et al., 2002; Ozinsky et al., 2000; Takeuchi et al., 2001) that can be mimic by the use of synthetic lipoproteins (Pam3CSK4 and FSL-1 respectively). NOD1 detects d-glutamyl-meso-diaminopimelic acid (iE-DAP), which is primarily found in Gram-negative bacteria PG. NOD2 detects muramyl dipeptide (MDP) that is ubiquitously present in bacterial PG (Chamaillard et al., 2003; Girardin et al., 2003 and **Table 4**).

Flagellin is the major protein constituent of bacteria. TLR5 is responsible for the detection of its constant domain D1, is relatively conserved among different species and can be triggered using a purified form (FLA-ST from *S. typhimurium*; **Table 4**) (Hayashi et al., 2001). Some bacteria, such as *Helicobacter pylori* and *Campylobacter jejuni*, produce flagellins that lack proinflammatory properties and therefore escape the flagellin-specific host immune responses (Andersen-Nissen et al., 2005).

### 5.1.2. Bacterial DNA

Bacterial genomic DNA contains unmethylated CpG dinucleotides in a particular base context (referred to as CpG-DNA) that induce strong immunostimulatory activities, such as inflammatory cytokine production and Th1 immune responses (Hemmi et al., 2000; Krieg, 2002). The particular DNA sequences that induce an immune response vary between species. Synthetic oligonucleotides containing the CpG motif are commonly used for *in vitro* experiments (i.e. ODN2216; **Table 4**). Since TLR9 resides in the endosome, bacterial DNA must be delivered to this intracellular compartment, for double-stranded DNA degradation into multiple single-stranded CpG-motif-containing regions that will interact with TLR9 (Ahmad-Nejad et al., 2002; Latz et al., 2004).

The bacterial DNA presents in the cytoplasmic compartment, can bind to cytosolic DNA sensors like AIM2 that will induce inflammasome activation, or such as cGAS leading to a type I IFN production.



## 5.2. PAMPs in Mycobacteria

Mycobacteria are intracellular bacteria that survive in host macrophages by a number of elaborate mechanisms. The mycobacterial cell wall is composed of a thick mixture of lipids and polysaccharides. Purified mycobacterial cell-wall components have been shown to preferentially activate TLR2 and, to a lesser extent, TLR4. Lipomannan (LM) and lipoarabinomannan (LAM) are related powerful immunomodulatory lipoglycans for which the purified forms from *Mycobacterium smegmatis* (LM or LAM-MS) can be used for experimental models (see **Table 4**). Found in slow-growing virulent myco-bacteria (i.e. *M. tuberculosis*, Bacille of Calmette-Guerin or BCG), the mannosyl form of the LAM (manLAM) has been described as a powerful anti-inflammatory molecule (reviewed in Akira et al., 2006).

In addition, TLR2 in association with TLR1 can recognize a secreted antigen of *M. tuberculosis* (Thoma-Uszynski et al., 2001), and mycobacterial DNA, released during endolysosomal degradation or by the bacterial ESX-1 perforating system, can bind to TLR9 or intracytoplasmic DNA sensors (i.e. cGAS) respectively (Watson et al., 2015). It has been shown that Dectin-1 can recognize an unidentified ligand on mycobacteria, which leads to the production of IL-12 (Sancho and Reis e Sousa, 2012).

## 5.3. Fungal Recognition by PRRs

Several fungal PAMPs located in the cell wall or on the cell surface of fungi are recognized by TLR2 or TLR4.

Dectin-1 binds  $\beta$ -glucan and is the primary receptor on macrophages for phagocytosis of various fungi (Brown et al., 2002). The particulate *S. cerevisiae*  $\beta$ -glucan (WGP) is used to specifically trigger Dectin-1 (Li. et al., 2007; see **Table 4**). It has been demonstrated that dectin-1 can collaborate with TLR2 in response to yeast (i.e. *C. albicans*) to elicit a strong inflammatory response via recruitment of the protein tyrosine kinase Syk (Gantner et al., 2003; Rogers et al., 2005; Underhill et al., 2005). In addition to dectin-1, other CTLRs such as the type 3 complement receptor, the mannose receptor, and DC-SIGN are implicated in the recognition and phagocytosis of *Candida* (Takahara et al., 2004). Dectin-1 can orchestrate effective antifungal mechanisms through Th1 and Th17 responses, which are defective in human patients with homozygous non-functional Dectin-1 (Hardison and Brown, 2012).

## 5.4. Viral recognition by PRRs

Various viral structural components, including viral DNA, double-stranded RNA (dsRNA), single-stranded RNA (ssRNA), and surface glycoproteins, are recognized as PAMPs by TLRs and other PRRs. The recognition of viral components by PRRs commonly induces type I IFN production that can activate target cells in both autocrine and paracrine manners.

Like previously described for bacteria, DNA viruses (i.e. herpes simplex virus) contain genomes that are rich in CpG-DNA motifs, which can trigger endosomal TLR9 (Hochrein et al., 2004; Krug et al., 2004a, 2004b; Lund et al., 2003; Tabeta et al., 2004). Cytoplasmic Viral DNA can bind to cytosolic DNA sensors like AIM2 or cGAS.

ssRNA can be recognized by TLR7, TLR8 and RIG-I. Many enveloped viruses such as influenza, traffic into the cytosol through the endosomal compartment. The phagolysosome is a highly acidified environment leading to ssRNA release and recognition by TLR7 or TLR8. The RIG-I sensing targets genomic ssRNA directly released in the cytoplasm (i.e. Sendai virus) or during the replication step (i.e. Influenza virus). The TLR7 and TLR8 genes show high homology to each other. As a consequence, those receptors recognize similar synthetic antiviral imidazoquinoline components (i.e. Gardiquimod for TLR7 and R848 for TLR7 and TLR8) as well as uridine-rich or uridine/guanosine-rich ssRNA of both viral and host origins (Hemmi et al., 2002; Heil et al., 2004; Diebold et al., 2004).

dsRNA such as synthetic analog, polyinosine-deoxycytidylic acid (poly I:C), can be recognized by TLR3 and RLRs members like RIG-I and MDA5 and are potent inducers of type I IFNs (Alexopoulou et al., 2001; **Table 4**). TLR3 is specifically expressed in conventional DCs (cDCs) that phagocytes dying cells, but not in pDCs. Immunization with virus-infected cells or cells containing synthetic dsRNA leads to a striking increase in CTL crosspriming against cell-associated antigens, which is largely dependent on TLR3 expression by antigen-presenting cells (Schulz et al. 2005; Yatim and al., 2015).

Some viral-envelope glycoproteins can be recognized by TLR4 or TLR2 leading then to an inflammation by the production of proinflammatory cytokines rather than specific antiviral responses.

## 6. Genetic evolution and phenotypic variation

Natural selection leaves molecular changes in the genome depending on whether the variant was i) advantageous, increasing its frequency through positive selection; ii) deleterious, leading to its disappearance by negative selection; or iii) favored in combination with the maintenance of allelic diversity, in which case multiple alleles will be maintained at intermediate frequencies by balancing selection (Vitti et al., 2013). Genome-wide analysis (GWAS) for positive selection have detected thousands of loci displaying signatures suggesting that functional polymorphisms in these genes have conferred a selective advantage for host survival. These studies have identified a number of biological functions as being particularly strongly targeted by selection, including pigmentation, metabolic traits and immunity-related genes. Indeed, it seems that immune-related genes have been privileged targets of positive and negative selection in human and non-human populations (reviewed in Barreiro and Quintana-Murci, 2010).

This observation is consistent with the massive re-population imposed by infectious disease waves throughout time (Casanova et Abel, 2005). A current reflection of this past demographic phenomenon is the important observed balancing selection, generally rare in the human genome (Andres et al. 2009), observed for human leukocyte antigen (HLA) genes and KIR. The high level of diversity at these loci is potentially due to acquired advantage for heterozygous individuals to detect a wider variety of microbes (Parham, 2005; Prugnolle et al., 2005; Fumagalli et al., 2009). At the level of innate immunity, several recent studies have characterized the diverse impact of selection on families of innate immune genes providing insights into their relative importance in the human lineage (reviewed in Quintana-Murci and Clark, 2013).

In that context, the aim of the evolutionary genetics approach is to look for the footprints of past natural selection in the genome of present populations.

## **7. Immune monitoring and use of whole blood stimulation system**

Different technological approaches are available to evaluate phenotypic immune variations or to understand the mechanisms underlying an immune response. In addition an increasing number of high-throughput tools are currently available to measure numerous parameters and/or numerous samples helping greatly in designing population-based study. With the help of those technologies and depending of the scientific question several models can be used to characterize immune response. In this section, I will mentioned the tools available for immune monitoring and present the different models that can be used for hypothesis research finishing by the rational of using whole blood stimulation system.

### **7.1. Tools for immune monitoring**

As depicted in the previous sections, the immune system can be characterized thanks to the expression of a myriad of molecules that will be tightly and specifically regulated whenever an immune response will be triggered.

One first level of analysis is the proteomic approach that is routinely used in clinic to detect and measure the progression of inflammations by the analysis of clinically approved biomarkers (i.e. CRP). To better assess the entire complexity of secreted inflammatory proteins in serum or in culture supernatant, multiplex assay can be used to determine the underlying pathways and cell types involved.

The quality and quantity assessment of specific cellular populations based on intracellular or extracellular markers is commonly analyzed using multiparameter staining by flow cytometry.

In addition, *in vitro* functional cellular assays can measure innate or antigen-specific responsiveness (i.e. migration, cell death) in cells exposed to immunogenic factors.

The current model systems commonly used in research laboratory to analyze the immune response across time, conditions and tissues, are animal models and cell and tissue based systems.

To go further in the cell regulation analysis, the genomic approach using DNA or RNA sequencing can determine potential genetic polymorphisms (i.e. genome-wide association studies) or allow gene identification. RNA sequencing, RT-qPCR or hybridization approaches

measure the abundance of gene expression that can be then compared in different conditions (homeostatic vs inflammatory) or across tissues/cell types.

Since the last centuries those approaches have largely contributed in the understanding of immunological mechanism/regulation and have been the source of the major discoveries in biology and constitute a reference for molecular pathways or general immunological concepts commonly accepted.

## **7.2. Cell-based approach**

Cell-based assays are widely used and contribute strongly to the understanding of mechanisms restricted to singular cell or tissue types and also for drug discovery screening as they provide a quite simple, high-throughput and cost-effective system before use in animal model.

To understand at the molecular level the pathways involved in the regulation of a specific cell type, the cleanest model is the isolation and *in vitro* manipulation of the cell of interest. By using immunological and molecular technologies, this approach allows to decipher the mechanisms implicated in a given population or even at a single cell level. More recently, there has been an increased interest for single cell studies that reveal that even in a highly purified population (upon same environmental conditions) there exists heterogeneity between cells with cyclic gene regulation that modify the level of expression measured at a given timepoint adding a new layer of complexity. The major caveat on these observed signatures is the strong dependence on the study design (i.e. isolation method, growth factors used) and the technology used that caution against general interpretation across studies (Cooper and Shedden, 2013; Gawad et al., 2016). Most approaches use single defined primary cell types or cell lines, without taking into account the importance of cellular interactions and/or the microenvironment for a given immune response. For these reasons the animal model appears more adapted to take into account the impact of a complex environment.

## **7.3. Animal models**

Since a long time, scientists have conducted experiment on animals to observe and understand biological structure and function. Because most of the studies cannot be performed in

humans, the animal model has been accepted as an analogical model to conduct and interpret physio-pathological experiments.

One remarkable historical example from the 19<sup>th</sup> Century was the use by Louis Pasteur and Emile Roux of animal models to produce and test the rabies vaccine before the first human trial (P. Debre, 2000).

Since then, following technological advances, numerous models have been developed or selected to mimic human diseases (i.e. inducible, spontaneous, genetically modified models) contributing in successful human clinical applications.

Nonetheless, recent comparisons revealed the limitations of using animal models (and cell lines) as a way to explore human pathogenic conditions (Editorial, Of men, not mice, 2013). Concerning the specific case of transcriptomic studies, the match between inflammatory responses in human and mice has been reported to be really poor (Seok et al., 2013). These discrepancies can be due to the species differences observed in the innate and adaptive immune cell populations. As examples, we can highlight the absence of CLEC2 (a CTLR member) expression by mouse neutrophils (Mantovani et al., 2011) or expression of TLR4 and CD14 (the two canonical ligands for LPS) in mouse B cells in contrast to human B that are unresponsive to LPS (Garrault et al., 2012). Those findings caution against extrapolation across species and motivate to continue to develop reliable tools for human immune explorations that mimic as much as possible physiological conditions.

#### **7.4. Human whole blood model**

The major limitation in human studies is the access to the relevant tissue(s). In contrast, blood sample collection is relatively non-invasive, easy to perform and containing the cells of the immune system. As the blood stream system circulates through all the tissues, those cells are exposed to stimuli from potential local injury, malignant or infectious sites in addition to the systemic factors (i.e. diet, inflammatory molecules). In the previous chapter, we have already indicated that innate and adaptive immune cells compose blood cells at different stages of differentiation (naïve, memory, activated), recently or anciently educated. For these different reasons, the study of whole blood may reflect systemic changes.

In addition, in contrast to studies examining functional responses of single leukocyte subsets, the use of a whole-blood approach enables the examination of the responses of potentially all leukocytes simultaneously therefore more closely mimicking *in vivo* inflammatory responses generated by the addition of pro-inflammatory molecules or microbes of interest (Crucian and

Sam, 1999; Deenadayalan et al., 2013). This approach allows also the analysis of early events between cells and inflammatory components keeping the complex inter and intra-cellular cross-talk (Morris et al., 2012). Supporting this aspect, using whole blood allows the retaining of any soluble factors present in serum that influence cell activation. As example, the plasma contains glucose, which is an essential source of energy utilized in cellular metabolism, but also the proteins of the complement that we have described previously as active partners of immune cells functions upon inflammation. Same for lipopolysaccharide (LPS) binding protein (LBP), which is necessary for cellular responses to exogenous LPS stimulation (Hamann et al., 2005; Alberts, 2005). In the bench side, utilization of a whole-blood assay avoids potential bias from cellular stimulation associated with leukocyte subset isolation techniques (Nerad et al., 1992; Wilson et al., 1991; Stibenz and Buhrer, 1994)

Another technical critical aspect that can influence the generation of reproducible data is the fact that whole blood assay requires minimal sample manipulation to assess cytokine production when compared to other methods that require labor intensive isolation and culturing of specific cell populations.

Currently, there is an increasing movement of medicine toward the use of individual therapy approach for the cure of different pathological conditions (i.e. autoimmunity, cancer) or to improve vaccine campaign efficacy. This implicates a better understanding of the physiological cells interplay upon different inflammation conditions taking into account the inter-individual variability. In the same manner, there is still a need for accurate analysis of the factors (genetic and/or environmental) that can impact intrinsic immune phenotype, from the same tissue type, that could be linked to potential differences observed across individuals in induced immune responses studies. To that end, the data presented in this thesis aimed to gain insight into simplified deconvolution of acute inflammation in a standardized whole blood model that can help in the mapping of clinical relevant immunogenic components and identify outliers. In order to help in future deeper deciphering of inter-individual variance, the impact of genetic and environmental factors on standardized whole blood immune parameters was determined. Our aims were: i) develop standardized model of acute inflammation and immune phenotype analysis; ii) deconvolute four major inflammation signatures in clinical relevant whole blood model; iii) identify parameters that influence immune response and iii) provide interactive reference datasets from highly annotated healthy individuals.

## CHAPTER II

---

# **STANDARDIZED WHOLE-BLOOD TRANSCRIPTIONAL PROFILING ENABLES THE DECONVOLUTION OF COMPLEX INDUCED IMMUNE RESPONSES**



## CHAPTER II: Standardized whole-blood transcriptional profiling enables the deconvolution of complex induced immune responses

### Context :

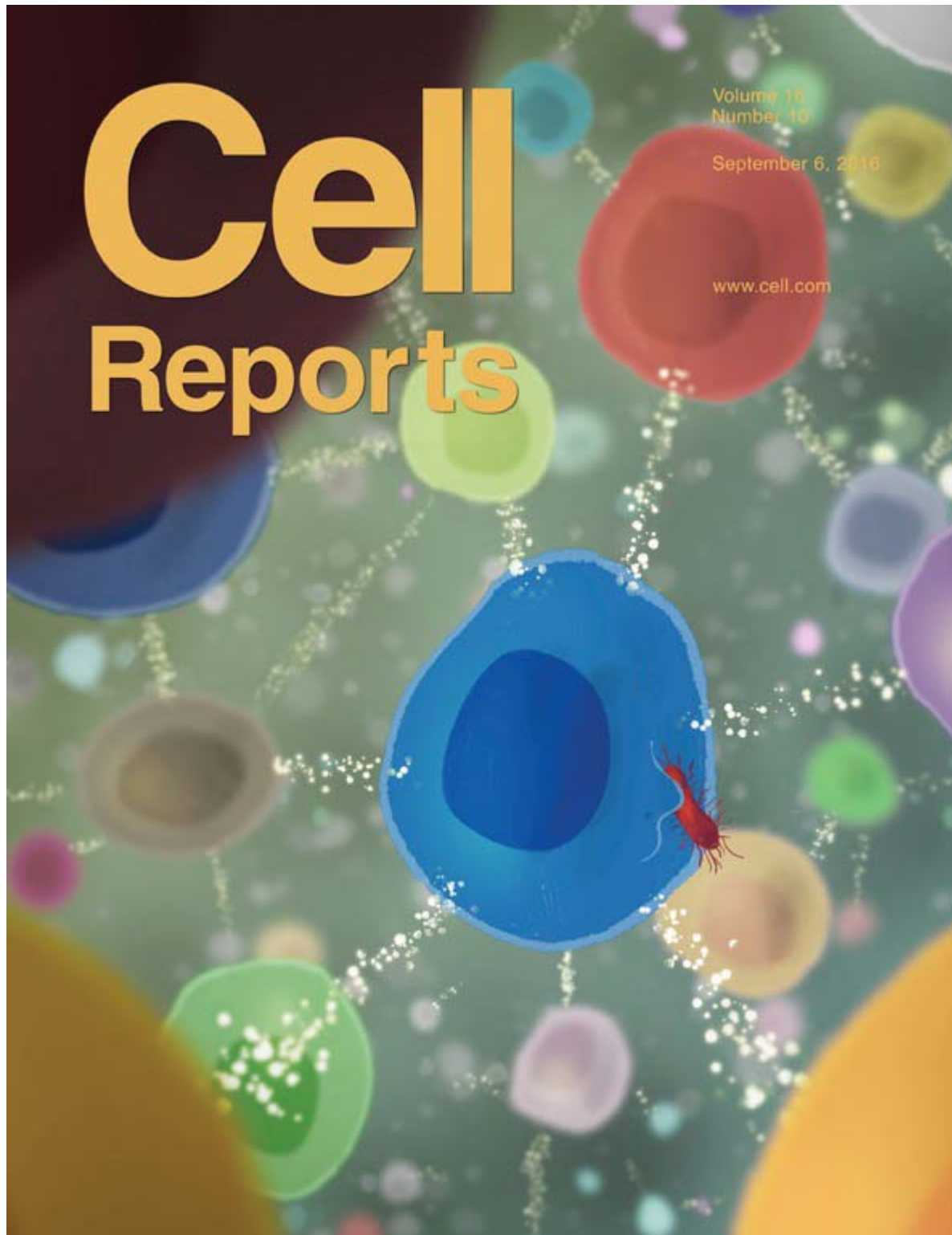
The *Milieu Intérieur* (MI) project aims to determine what are the genetic and environmental factors that drive the human immune response and to provide a large dataset that can be used as reference values by the community.

The use of whole blood as a clinically relevant model for human studies is undeniable but still presents some technical challenges. In the previous section, we have introduced the key elements to consider for the analysis of induced immune responses. The complex cell composition and the fact that each cell presents specific inflammatory response properties requires development of standardized tools to limit introduction of technical bias that could increase external inter-variability.

In a previous study within the consortium, we have shown that the use of standardized devices and protocols able to provide high quality datasets for the analysis of whole blood induced immune response in terms of protein secretion. In terms of transcriptomic analysis, there is still some need in the community for standardized protocols and highly annotated datasets that can be both transposed in different platforms for the induced immune response monitoring in healthy or pathological conditions.

Using the stimulated cell pellets from 25 donors (30-39 years old) from the *Milieu Intérieur* collection, in this collective study my PhD work has been driven by three principal aims i) **Development of standardized mRNA expression analysis from whole blood syringe based assay stimulations** ii) **Use and development of specific statistical tools for mRNA analysis to define four major inflammatory cytokine induced signatures** iii) **deconvolution of complex induced immune responses** using these restricted cytokine induced gene signatures.

The experiments described here were designed to determine if those tools could be useful to the community for future whole blood dataset sharing and leads to a publication in *Cell Reports* (Urrutia et al., 2016).



## 1. Introduction

The initiation of inflammatory responses is typically triggered by a local event engaging sentinel cells, leading to the subsequent recruitment and accumulation of leukocytes. This process can result in the elimination of the initial cause of tissue disruption, the clearance of dying cells and establishes a path towards tissue resolution.

Cytokines mediate cell-to-cell communication, acting to recruit immune cells to inflammatory microenvironment and drive the required effector mechanisms. Despite the inherent complexity of these processes *in natura*, analyses of inflammation have typically focused on the decision-making circuits within cells, and, in most cases, have been restricted to single cell types (Amit et al., 2009; Jovanovic et al., 2015; Lee et al., 2014). Several other studies have assessed *in vivo* responses to vaccination, typically performing sampling over time to assess induced protein, mRNA expression and seroconversion (Banchereau et al., 2014; Li et al., 2014; Tsang et al., 2014).

While informative, these latter approaches permit the testing of only one stimulation condition per individual and are restricted to qualified or experimental vaccines. To properly account for inter-individual variability in the deconvolution of complex immune responses both simple (synthetic or purified ligand) and complex (live or heat killed microbe), stimulations must be performed in the same donor and at the same time, and standardized approaches for all steps from sample collection to analysis must be applied.

To test the hypothesis that responses to Toll-like receptor ligands or whole microbes can be captured by the transcriptional signature of key effector cytokines, we employed a standardized whole-blood stimulation approach with an automated single step RNA extraction, and hybridisation gene array readout. Stimulations were performed at the point-of-care, using syringe-based medical devices (TruCulture tubes), in a pilot study that consisted of 25 well-characterized healthy individuals of European ancestry (Thomas et al., 2015).

Previously, we reported the testing of protein signatures present in the culture supernatant (Duffy et al., 2014). Herein, we used the cell pellets extracted from the TruCulture stimulation systems to define the transcriptional response to clinically relevant cytokines; interferon-alpha 2A (IFNA), interferon-beta 1 (IFNB), interferon-gamma (IFNG), tumor necrosis factor-alpha (TNFA), and interleukin 1-beta (IL1B). By defining unique and distinct gene expression signatures of cytokine-induced transcription, it was possible to test the clustering and classification of responses to Toll-like receptor (TLR) agonists or whole microbes (including

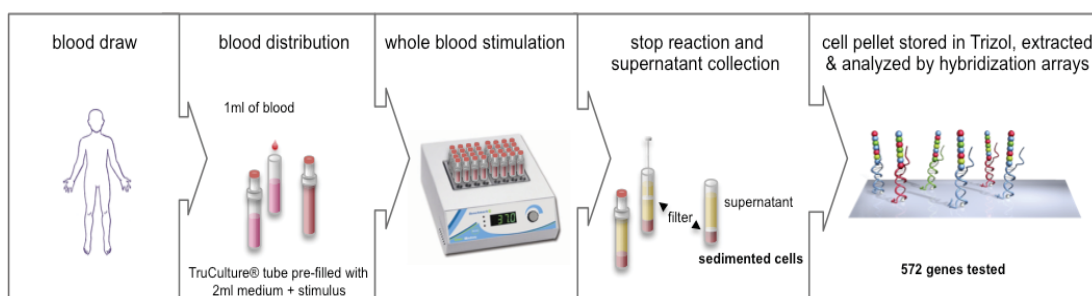
heat killed (HK) gram-negative bacteria, HK gram-positive bacteria, HK fungi, live mycobacteria and viruses).

Our results demonstrate the ability to define complex stimuli in terms of the underlying cytokine loops. Moreover, we provide reference values that reflect the degree of naturally occurring variation of immune responses among healthy individuals originating from a homogeneous European background. These data have been made available as a reference for the community, accessible through an online R-Shiny application that permits data-mining using the analytical methods presented.

## 2. Results

### 2.1. Distinct transcriptional signatures induced by the IFNB, IFNG, IL1B and TNFA cytokines

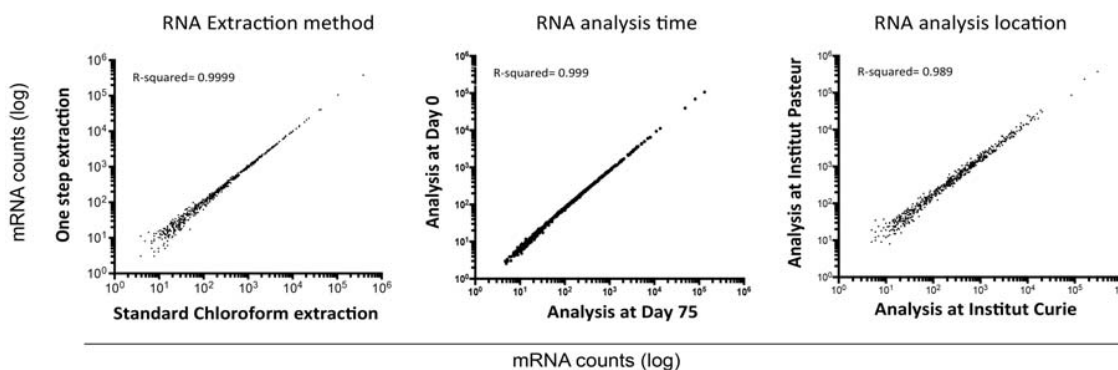
To perform *ex vivo* stimulation, while preserving physiological cellular interactions, we utilized syringe-based medical devices for activating immune cells present in whole-blood. Based on initial dose-finding studies, quality assurance, solubility and stability testing (Duffy et al., 2014), we prioritized stimuli for development in TruCulture whole-blood collection and culture devices (Myriad RBM). After 22-hours stimulation, insertion of a valve separator yielded a cell pellet that was stabilized in Trizol LS, and stored at  $-80^{\circ}\text{C}$  for subsequent mRNA expression analysis utilizing the NanoString nCounter technology (**Figure 8**).



**Figure 8 : Schematic overview of workflow from blood draw to gene expression analysis**

Due to the Trizol content in our samples and to minimize pre-analytical biases, we established an automated mRNA single-step chloroform-free extraction protocol (Tecan script provided

on-line, see [www.milieuinterieur.fr/en](http://www.milieuinterieur.fr/en)). Direct comparison with conventional RNA extraction protocols indicated excellent correlation in gene expression counts between the two extraction methods (Spearman's Rank-Order Correlation,  $r_s > 0.99$ , **Figure 9**).



**Figure 9 : Quality control measures for gene expression analysis**

Comparison between mRNA counts for single step extraction protocol and standard extraction protocol utilizing chloroform step, for nCounter analysis at 2 separate time point (75 days apart), and at 2 different locations (Institut Curie, Paris and Institut Pasteur, Paris) (Representative examples are shown and  $r_s^2$  is reported, based on a Spearman correlation).

Expression data were normalized with nSolver™ Analysis Software (NanoString), using four housekeeping genes: *RPL19*, *TBP*, *POLR2A*, and *HPRT* (**Figure 10A-D**). These 4 housekeeping genes were selected following the application of the geNorm method (Vandesompele et al., 2002), an established algorithm for identifying stable housekeeping genes. The selection of these genes is supported by their strong correlations pre- and post-stimulation ( $r_s > 0.9$ ) across the 25 donors, in contrast with those housekeeping genes that were discarded ( $r_s < 0.7$ ) (**Figure 10B** and data not shown).

The overall rationale for the selection of the NanoString platform, as compared to other transcriptional profiling strategies, is presented in **Table 5**.

This choice was validated by the high reproducibility of the data obtained when experiments were performed at different times or at separate institutional core facilities ( $r_s > 0.98$ , **Figure 9**).

Technology	Method	Gene set	Adaptable for use with whole blood
Fluidigm Biomark	Microfluidic-based PCR	Up to 90 genes	<ul style="list-style-type: none"> <li>• cDNA preparation source of possible pre-analytical bias</li> <li>• PCR reaction inhibited by presence of heparin (and other serum factors) in TruCulture systems</li> </ul>
Nanostring nCounter	RNA Hybridization	Up to 780 genes	<ul style="list-style-type: none"> <li>• Direct quantification of mRNA expression (i.e., no cDNA preparation, no amplification)</li> <li>• suitable for single-step extracted RNA</li> </ul>
Affymetrix microarray	cDNA Hybridization	Whole genome	<ul style="list-style-type: none"> <li>• requirement for high quality RNA</li> <li>• cDNA preparation source of possible pre-analytical bias</li> </ul>
Ion Torrent / Illumina	NGS	Whole genome	<ul style="list-style-type: none"> <li>• requirement for high quality RNA</li> <li>• globin RNA present in whole blood dominate read count (~70% of total RNA)</li> </ul>

Table 5 : Comparison of different gene expression profiling technologies for whole blood analysis

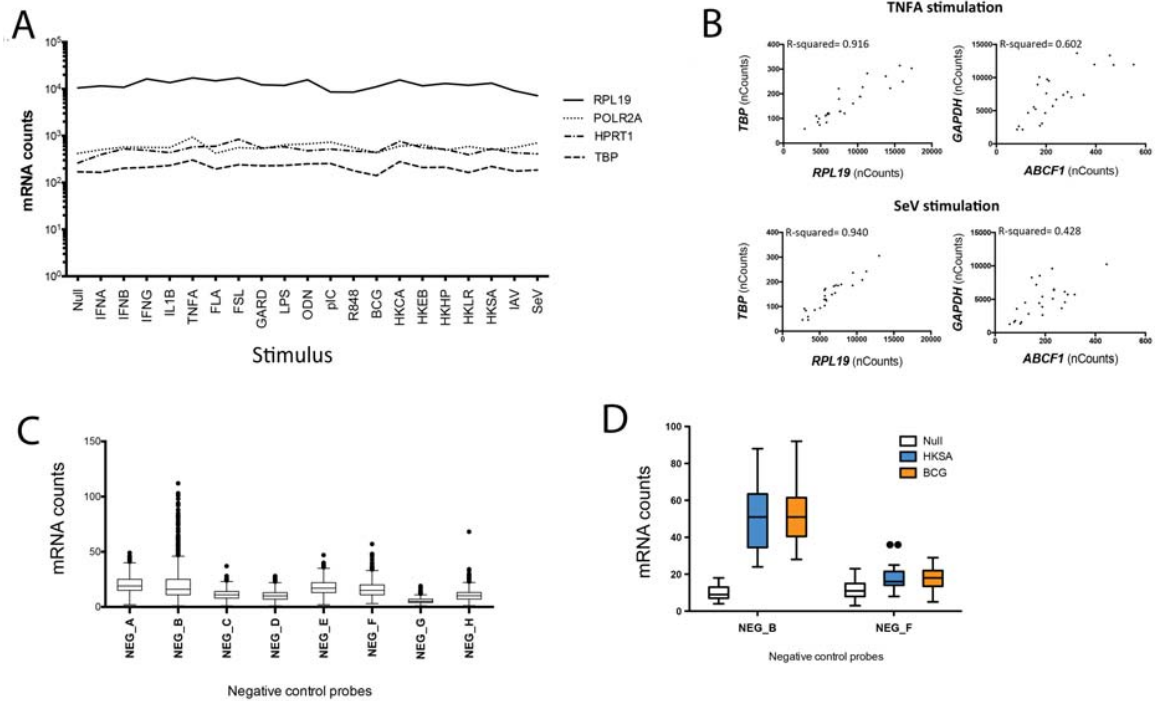


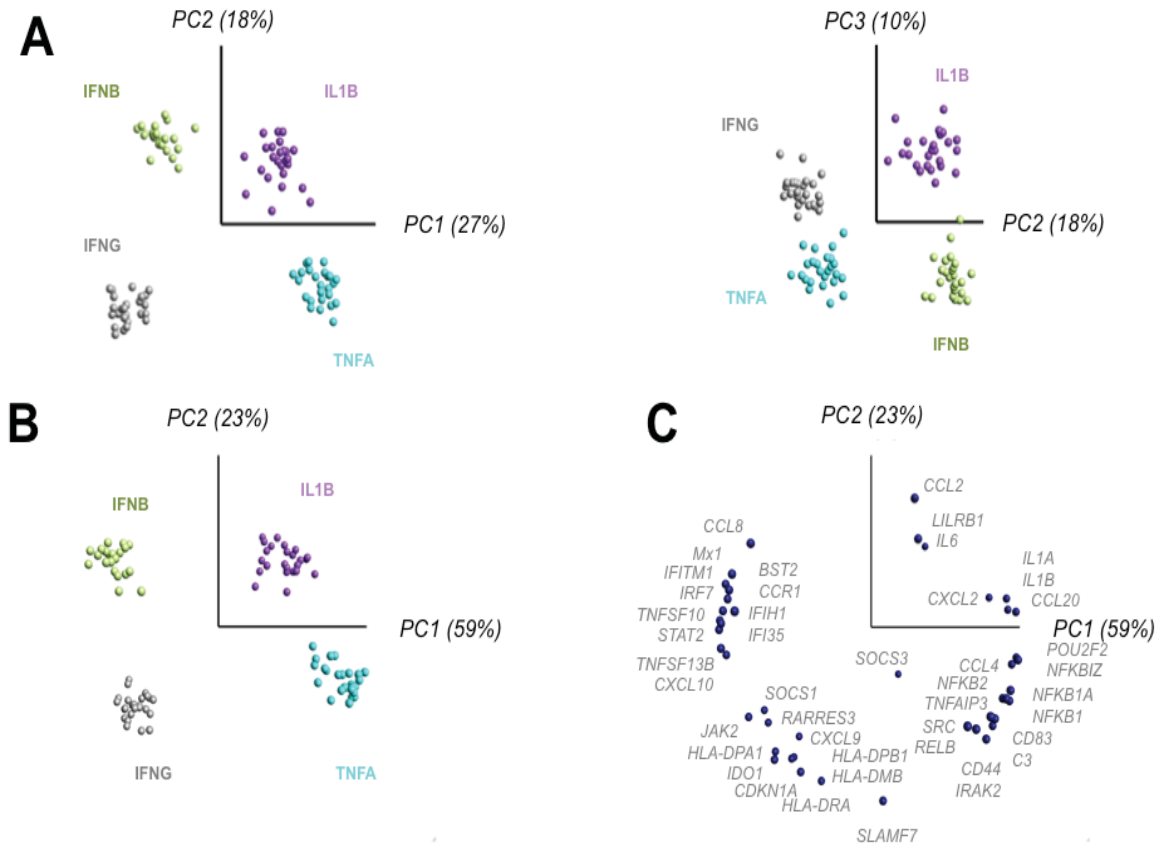
Figure 10 : Internal control evaluation for gene expression analysis

(A) Mean of mRNA counts (log scale) for the 4 selected house keeping genes (HPRT1, POLR2A, RPL19, TBP) across the different stimulation conditions for the 25 donors included in the study. (B) Comparison of mRNA counts (linear scale) for two geNorm selected genes (left plots) versus two candidate house keeping genes (right plots) upon TNFA and SeV stimulation (C) Box-whisker Tukey plots for the negative control probe counts. (D) Example of Neg B and Neg\_F probes from TruCulture stimuli LPS, HKSA, and BCG.

To assess the signatures induced by cytokine stimulation, we analyzed the expression data of a total of 572 genes in the 25 donors, using unsupervised principal component analysis (PCA) (Figure 11A). The PCA revealed strong clustering of stimuli-specific responses, with the first three principal components (PCs) explaining 55% of the total variance; PC1 separated IL-1B and TNFA from IFNB and IFNG; and PC2 distinguished TNFA from IL1B, and IFNB from

IFNG. Of note, the response to IFNA was also tested, and found to be similar to that of the IFNB response (t-test with  $q < 0.05$ , reported no variables as significantly different between the 2 stimuli) (**Figure 12**), and therefore IFNA was excluded from further analyses.

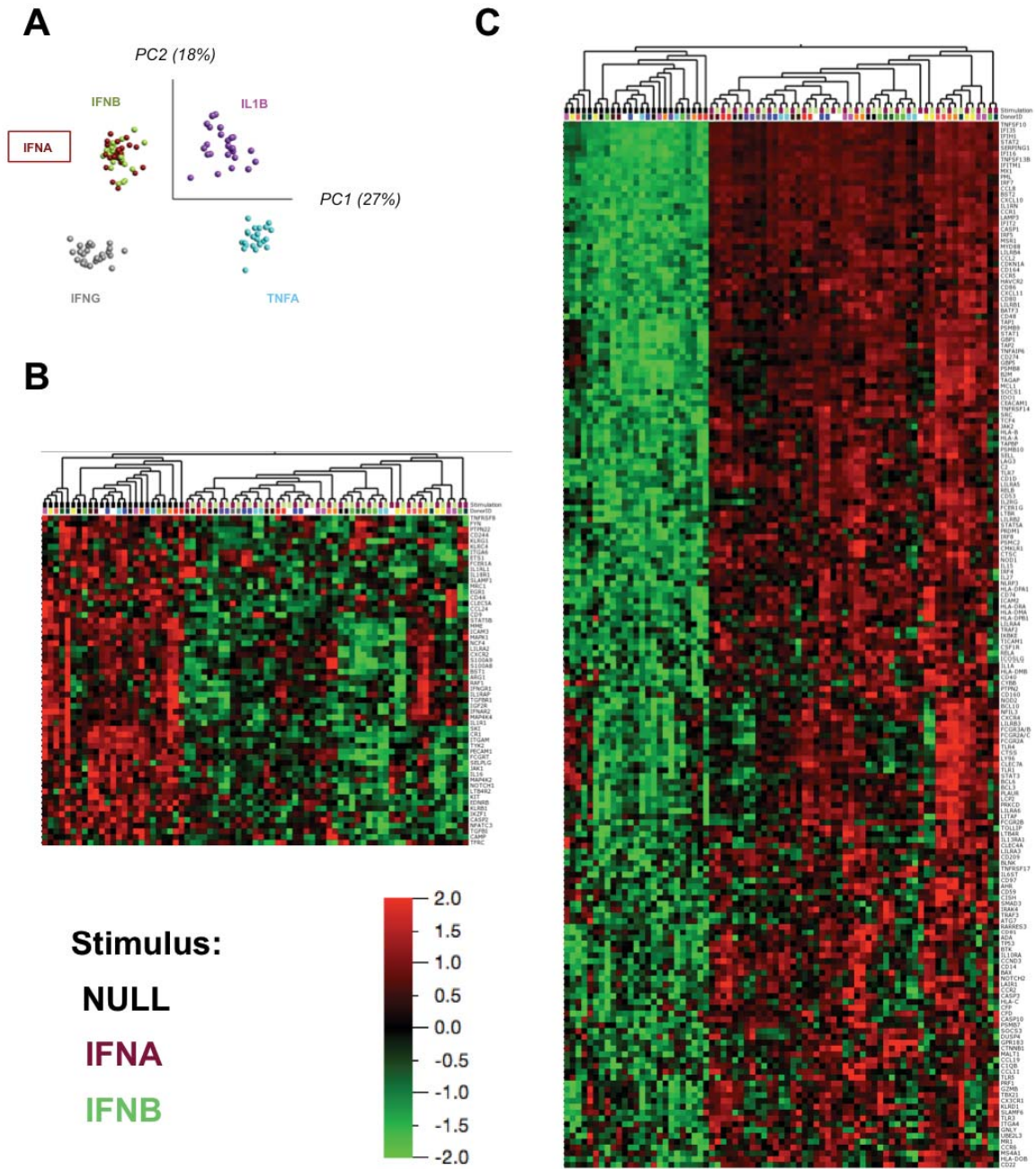
To reduce the dimensionality of the data and exclude genes that did not contribute to unique cytokine induced signatures, we next defined the differential gene expression for each stimulus with respect to the null control using linear support vector machine (SVM) approaches (Burges, 1998). This enabled us the selection of predictive cytokine gene signatures from gene lists ranked according to a paired t-test (individual stimulus vs. null condition). Bootstrapping of data in the SVM training phase ensured robust results (details provided in the materials and methods). The union of the selected cytokine gene signatures yielded a set of 44 genes that separated the four cytokine stimuli (**Table 6**). The resulting PCA projection revealed that the four stimulation conditions could be separated into four clearly distinct clusters based on the expression levels of these 44 genes, with PC1 and PC2 capturing 82% of the total variance (**Figure 11B**). The 44 genes are represented on a biplot – a synchronized dual projection of the variables that drive the loading of the PC vectors (**Figure 11C**).



**Figure 11 : Distinct gene expression signature induced by cytokine stimulation**

(A-C) Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with IFNB (pale green), IFNG (grey), IL1B (purple), and TNFA (turquoise). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes. Prior to applying PCA, values for each of the 572 mRNA were log transformed, centered to a mean value of zero across each donor and scaled to unit variance. The four cytokine stimuli are indicated by the colored circles and the vector position of each of the 25 donors is represented (A) left panel PC1 v PC2, right panel PC2 v PC3, the percentage of variance captured by each PC is indicated). (B) PCA on filtered gene expression data; first for differential gene expression (paired t-test comparing each cytokine with null and a q-value cut-off of  $10^{-3}$ ); followed by the classification of samples using linear support vector machine (SVM) approaches, and genes ranked according to a paired t-test, yielding a union gene set of 44 genes, (C) with a bi-plot of the 44 gene set variable PCA depicted.

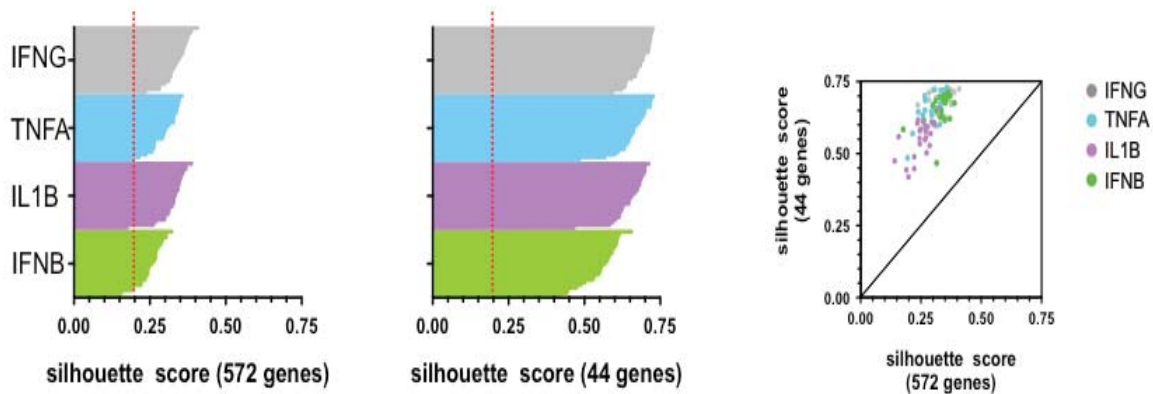




**Figure 12 : IFNA and IFNB show overlapping gene expression profiles**

(A) Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with IFNA (red), IFNB (pale green), IFNG (grey), IL1B (purple), and TNFA (turquoise). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes (the percentage of variance captured by each PC is indicated). (B-C) Hierarchical cluster analysis of the donors and gene expression following stimulation with IFNA, IFNB, and NULL control (black) identified 58 genes commonly down regulated (B) and 212 genes commonly upregulated (C) (ANOVA test,  $q$  value  $< 10^{-3}$ ). Each donor is color-coded revealing that in most instances, individual donors clustered for IFNA / IFNB responses

To quantify the improved clustering provided by this approach, we calculated silhouette scores, i.e., a measure of the distance between the respective k-means clusters, reported for each sample based on the likelihood to fall into one cluster as compared to any of the three other defined clusters. Comparison between the scores that were based on the complete 572 gene set *versus* the selected 44 gene set, revealed a higher score with reduced dimensionality of the feature list and a focus on those most highly discriminating genes (**Figure 13**).

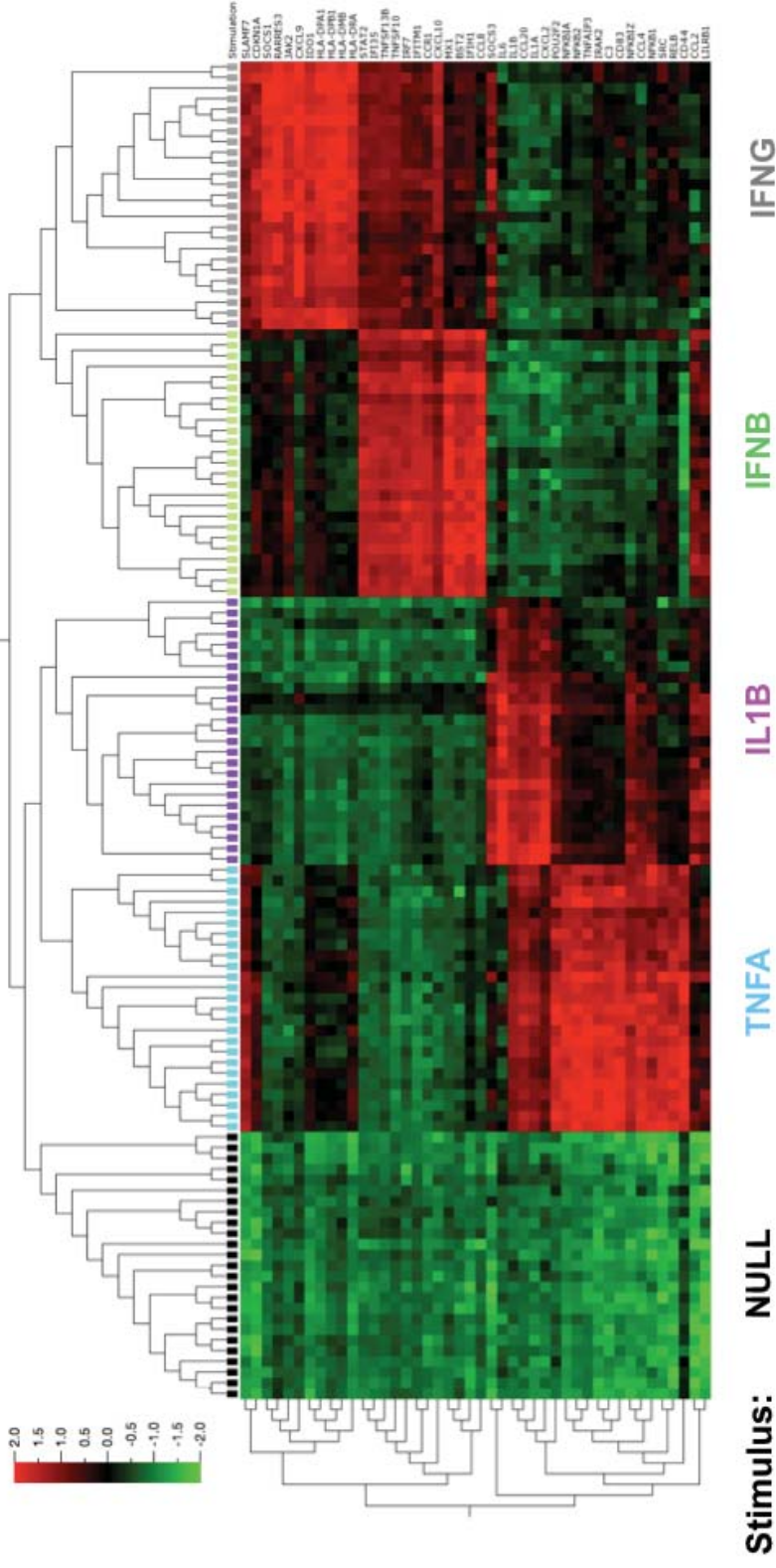


**Figure 13 : Silhouette score comparison from the clusters identifying the cytokine stimulation induced response using 572 gene versus 44 genes**

Silhouette scores for each cytokine IFNB (green), IFNG (grey), IL1B (purple), and TNFA (turquoise) based on the complete 572-gene set and the selected 44-gene set.

While our analyses revealed specific cytokine gene signatures, there was modest overlap in the induced gene lists when the stimulation conditions were compared to the null (**Figure 13**). Hierarchical clustering of the filtered gene list displayed the unique and overlapping gene expression for the four cytokine groups (**Figure 14**).





**Figure 14: Distinct gene expression signature induced by cytokine stimulation**

Hierarchical clustering of the donors based on the filtered gene list and IFNB, IFNG, IL1B and TNFA cytokine stimuli and Null condition showing the unique and overlapping expression

Gene Name	Associated cytokine	q-value (Stim vs. null)*	q-value (Anova on four cytokine stimuli)
<i>BST2</i>	IFN $\beta$	4.3 x 10 <sup>-43</sup>	4.6 x 10 <sup>-43</sup>
<i>C3</i>	TNF $\alpha$	1.6 x 10 <sup>-64</sup>	2.3 x 10 <sup>-64</sup>
<i>CCL2</i>	IL1 $\beta$	3.1 x 10 <sup>-21</sup>	2.7 x 10 <sup>-21</sup>
<i>CCL20</i>	IL1 $\beta$	6.3 x 10 <sup>-62</sup>	1.6 x 10 <sup>-61</sup>
<i>CCL4</i>	TNF $\alpha$	4.1 x 10 <sup>-57</sup>	7.9 x 10 <sup>-57</sup>
<i>CCL8</i>	IFN $\beta$	8.8 x 10 <sup>-53</sup>	9.6 x 10 <sup>-53</sup>
<i>CCR1</i>	IFN $\beta$	8.8 x 10 <sup>-33</sup>	9.0 x 10 <sup>-33</sup>
<i>CD44</i>	TNF $\alpha$	3.2 x 10 <sup>-58</sup>	5.8 x 10 <sup>-58</sup>
<i>CD83</i>	TNF $\alpha$	1.4 x 10 <sup>-59</sup>	2.4 x 10 <sup>-59</sup>
<i>CDKN1A</i>	IFN $\gamma$	1.2 x 10 <sup>-41</sup>	1.3 x 10 <sup>-41</sup>
<i>CXCL10</i>	IFN $\beta$	5.3 x 10 <sup>-51</sup>	5.7 x 10 <sup>-51</sup>
<i>CXCL2</i>	IL1 $\beta$	8.4 x 10 <sup>-39</sup>	7.5 x 10 <sup>-39</sup>
<i>CXCL9</i>	IFN $\gamma$	4.0 x 10 <sup>-43</sup>	4.0 x 10 <sup>-43</sup>
<i>HLA-DMB</i>	IFN $\gamma$	3.8 x 10 <sup>-61</sup>	2.5 x 10 <sup>-61</sup>
<i>HLA-DPA1</i>	IFN $\gamma$	4.2 x 10 <sup>-51</sup>	3.5 x 10 <sup>-51</sup>
<i>HLA-DPB1</i>	IFN $\gamma$	3.5 x 10 <sup>-51</sup>	2.7 x 10 <sup>-51</sup>
<i>HLA-DRA</i>	IFN $\gamma$	4.0 x 10 <sup>-45</sup>	3.9 x 10 <sup>-45</sup>
<i>IDO1</i>	IFN $\gamma$	2.2 x 10 <sup>-61</sup>	1.2 x 10 <sup>-61</sup>
<i>IFI35</i>	IFN $\beta$	3.4 x 10 <sup>-55</sup>	2.7 x 10 <sup>-55</sup>
<i>IFIH1</i>	IFN $\beta$	2.3 x 10 <sup>-54</sup>	2.2 x 10 <sup>-54</sup>
<i>IFITM1</i>	IFN $\beta$	5.7 x 10 <sup>-49</sup>	6.1 x 10 <sup>-49</sup>
<i>IL1A</i>	IL1 $\beta$	1.1 x 10 <sup>-59</sup>	2.0 x 10 <sup>-59</sup>
<i>IL1B</i>	IL1 $\beta$	7.8 x 10 <sup>-83</sup>	2.9 x 10 <sup>-82</sup>
<i>IL6</i>	IL1 $\beta$	1.9 x 10 <sup>-67</sup>	6.9 x 10 <sup>-67</sup>
<i>IRAK2</i>	TNF $\alpha$	4.8 x 10 <sup>-62</sup>	9.8 x 10 <sup>-62</sup>
<i>IRF7</i>	IFN $\beta$	3.4 x 10 <sup>-56</sup>	2.3 x 10 <sup>-56</sup>
<i>JAK2</i>	IFN $\gamma$	5.8 x 10 <sup>-51</sup>	5.0 x 10 <sup>-51</sup>
<i>LILRB1</i>	IL1 $\beta$	1.6 x 10 <sup>-37</sup>	1.5 x 10 <sup>-37</sup>
<i>MX1</i>	IFN $\beta$	1.4 x 10 <sup>-61</sup>	6.3 x 10 <sup>-62</sup>
<i>NFKB1</i>	IL1 $\beta$	8.7 x 10 <sup>-52</sup>	1.0 x 10 <sup>-51</sup>
<i>NFKB2</i>	TNF $\alpha$	2.0 x 10 <sup>-64</sup>	3.6 x 10 <sup>-64</sup>
<i>NFKBIA</i>	TNF $\alpha$	2.6 x 10 <sup>-67</sup>	3.2 x 10 <sup>-67</sup>
<i>NFKBIZ</i>	IL1 $\beta$	2.1 x 10 <sup>-61</sup>	3.5 x 10 <sup>-61</sup>
<i>POU2F2</i>	IL1 $\beta$	1.8 x 10 <sup>-70</sup>	6.6 x 10 <sup>-70</sup>
<i>RARRES3</i>	IFN $\gamma$	2.2 x 10 <sup>-49</sup>	2.1 x 10 <sup>-49</sup>
<i>RELB</i>	TNF $\alpha$	1.8 x 10 <sup>-40</sup>	1.9 x 10 <sup>-40</sup>
<i>SLAMF7</i>	IFN $\gamma$	9.0 x 10 <sup>-62</sup>	2.5 x 10 <sup>-62</sup>
<i>SOCS1</i>	IFN $\gamma$	1.6 x 10 <sup>-42</sup>	1.6 x 10 <sup>-42</sup>
<i>SOCS3</i>	TNF $\alpha$	9.0 x 10 <sup>-62</sup>	3.6 x 10 <sup>-55</sup>
<i>SRC</i>	TNF $\alpha$	2.3 x 10 <sup>-57</sup>	4.3 x 10 <sup>-57</sup>
<i>STAT2</i>	IFN $\beta$	4.3 x 10 <sup>-55</sup>	3.8 x 10 <sup>-55</sup>
<i>TNFAIP3</i>	TNF $\alpha$	2.7 x 10 <sup>-59</sup>	4.9 x 10 <sup>-59</sup>
<i>TNFSF10</i>	IFN $\beta$	2.3 x 10 <sup>-58</sup>	9.9 x 10 <sup>-59</sup>
<i>TNFSF13B</i>	IFN $\beta$	1.7 x 10 <sup>-57</sup>	1.0 x 10 <sup>-57</sup>

**Table 6 : Cytokine gene signature that defines transcriptional response to IFNB, IFNG, IL1B and TNFA**

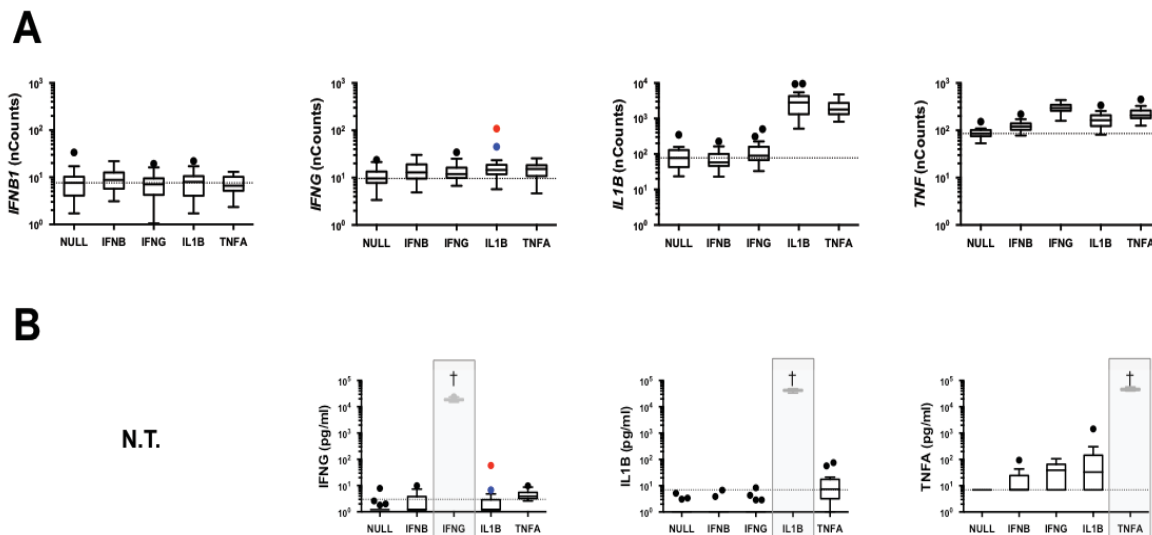
The union set of 44 genes as selected for each cytokine stimulus using linear support vector machine (SVM) approaches and paired T tests with respect to the null control. The q values for each cytokine as compared to the Null (paired T tests) and within the 4 cytokines (multi-group ANOVA) are shown.

To examine the intersection among cytokine-induced genes, we first analyzed the induction of *IFNB*, *IFNG*, *IL1B* and *TNFA* gene expression. While none of the four cytokines triggered high levels of type I or type II IFN expression (**Figure 15A**), IL-1B and TNFA both induced high expression of *IL1B* mRNA, and all four cytokine stimuli induced modest expression of *TNFA* (**Figure 15A**).

These data suggest potential cross-talk among the pathways and highlight a strong feed-forward inter-cellular spread of IL-1B signaling. While this has been previously shown (Dinarello et al., 1987), to our knowledge, there is no mechanistic understanding of how IL-1B activates the inflammasome and triggers caspase-1 activation. Unexpectedly, this analysis revealed two outlier individuals who showed high expression levels of IL-1B-induced *IFNG* (marked by red and blue dots, **Figure 15A**).

To establish if the observed high levels of *IFNG* expression resulted in higher protein secretion, we re-analyzed our previously published protein dataset (Duffy et al., 2014) generated using samples from the same donors and indeed, the two individuals showed the highest levels of IFNG protein in the culture supernatants (**Figure 15B**).

The presence of recombinant protein that was used as the stimulus restricted the interpretation of potential positive feedback loops for the given protein (these data points are masked by a grey box, **Figure 15B**).



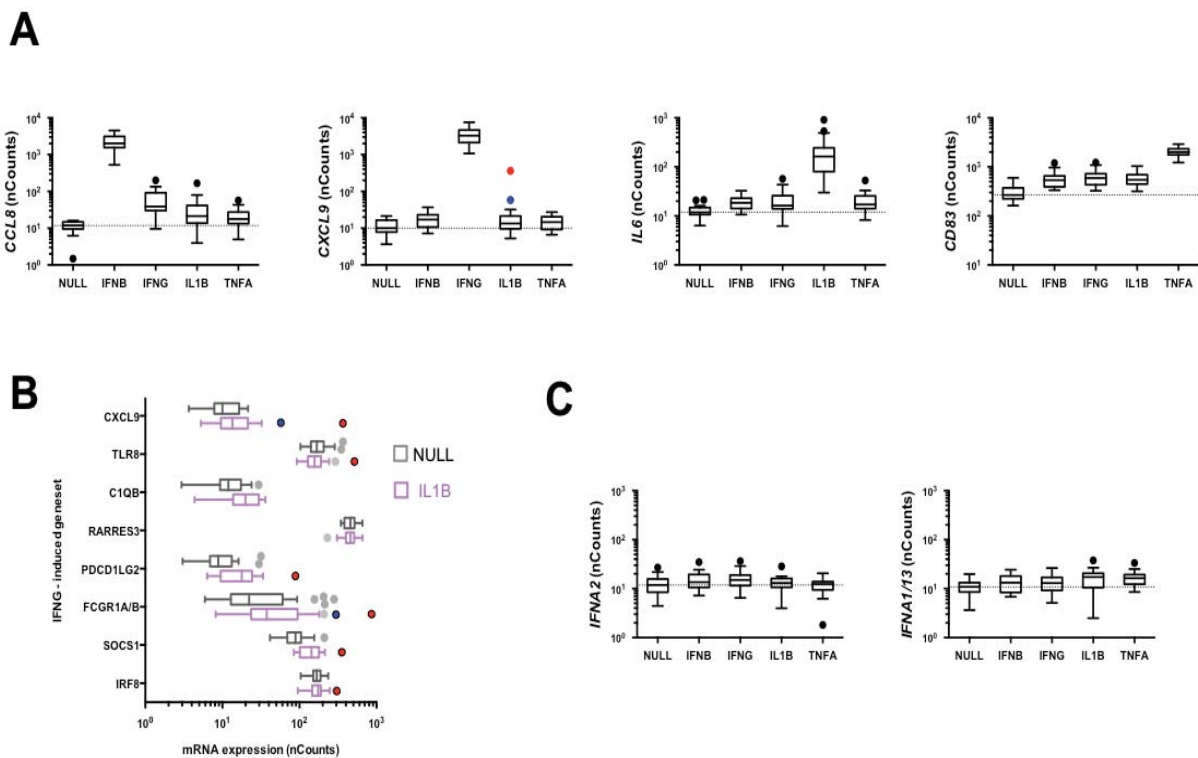
**Figure 15 : Interactions and outlier responses among the cytokine-induced gene expression signatures (1)**

Whole-blood from 25 healthy donors was stimulated using the Null, IFNB, IFNG, IL1B and TNFA stimulation conditions.

- (A) mRNA gene expression absolute nCounts or induced protein expression (B) are plotted for each of the four genes or gene products: *IFNB1*, *IFNG*, *IL1B*, and *TNFA* (N.T. signifies not tested for IFNB protein; grey shaded boxes mask those protein assays that are detecting the input stimulus in the TruCulture tube)

In addition to the induction of IFNG by the two outlier individuals, we also observed higher expression of several IFNG-induced genes, as compared to the other donors studied (**Figure 16A, B, C**).

Together, these data support the concept that the induced innate responses include the spreading of signals through cytokine feedback loops and potential cross-talk among the inter-cellular pathways.



**Figure 16 : Interactions and outlier responses among the cytokine-induced gene expression signatures (2)**

- (A) mRNA expression for the most differentially expressed gene, one per cytokine stimulus as reported in **Table 6** gene list).
- (B) Top IFNG induced gene expression is shown for the Null (grey) and IL1B (purple) stimuli. Data is represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation. Two individual outliers (identified by their induction of *IFNG* expression in response to IL1B stimulation) are indicated using blue and red circles, respectively.
- (C) Cytokine stimulation does not induce expression of IFNA genes. Box-whisker Tukey plots of IFNA2 and IFNA1/13 mRNA expression following stimulation with NULL, IFNB, IFNG, IL1B, and TNFA. Dotted line indicates median null value

## 2.2. Variable responses to TLR and microbe stimulation are captured by induced cytokine response

During vaccination or acute infection, the immune system is exposed to agonists that stimulate Toll-like receptor (TLRs) signaling. In such conditions, small numbers of cells are engaged, triggering in turn the production of cytokines that spread the inflammatory response. To test this concept, we evaluated whether the induced transcriptional responses to the four effector cytokines are capable of capturing the diversity of seven well-defined TLR agonists (Duffy et al., 2014): FSL-1 (FSL, also known as Pam2C) that engages the TLR2-TLR6 heterodimer; poly IC (pIC) that engages TLR3; lipopolysaccharide (LPS) that engages TLR4; flagellin (FLA) that engages TLR5; gardiquimod (GARD) that engages TLR7; R848 that engages both TLR7 and TLR8; and CpG-2216 oligonucleotide (ODN) that engages TLR9. Limiting doses of the respective agonists were selected to more closely reflect *in vivo* responses, and to ensure that we were working within the linear range of physiological responses (please refer to Duffy et al. (2014) or [www.milieuinterieur.fr/en](http://www.milieuinterieur.fr/en) for details on the dose and source of these reagents).

To assess potential similarity in gene expression, we projected the data from each of the seven TLR stimuli onto a fixed PCA coordinate, which was defined by the eigenvectors and eigenvalues of the optimized PCA of the four-cytokine induced mRNA expression data (44 genes defined in **Figure 11C**).

Strikingly, two of the TLR stimuli clustered with a defined cytokine – FLA and FSL vectors both projected onto the IL-1B cluster (**Figure 17A-B**). ODN eigenvectors projected into the IFNB quadrant, with an inter-donor variance in the intensity of gene expression (**Figure 17A**), which was consistent with our previous study of induced proteins.

This analytical approach can be further explored using the online user interface [[www.synapse.org/MilieuInterieur](http://www.synapse.org/MilieuInterieur); doi:10.7303/syn7059574].





**Figure 17 : Projection of TLR stimuli onto PCA analysis as defined by 4 effector cytokines**

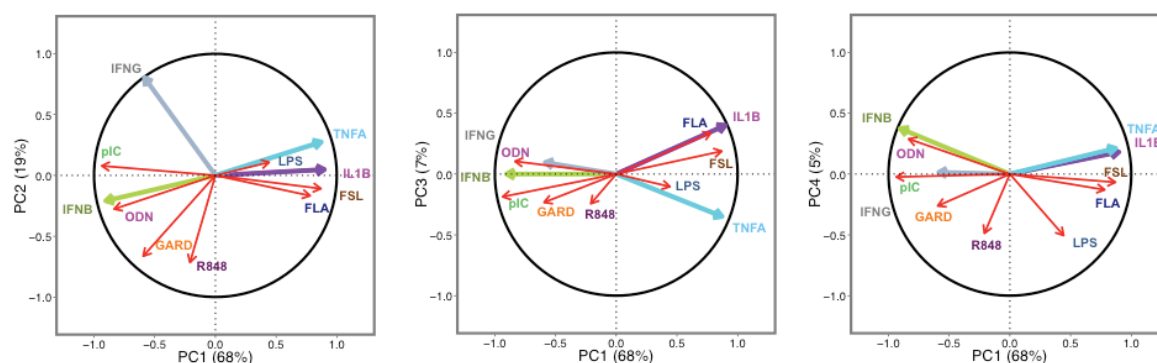
(A) PCA defined by the eigenvectors and eigenvalues as based on the four-cytokine induced mRNA expression data of the 44 genes defined in **Table 6**. Ellipses representing 95% confidence interval (CI) were constructed and replaced the individual samples. Projected sample vectors of TLR stimuli (shown in red) for each of the 25 donors (FSL, pIC, LPS, FLA, GARD, R848, ODN), individually projected onto the first 3 PC vectors, using the 44 selected genes

(B) Projection of different synthetic ligands (WGP, LAM, CPPD) onto the PCA as defined by four-cytokine induced mRNA expression.

We next represented the data on a correlation circle, as an alternative for visualizing the relationships among stimuli (**Figure 18**), allowing us the projection of all TLR stimulation conditions across the four PC axes.

When two stimulation vectors are close to the unit circle, and are proximal to each other, then they are positively correlated (*e.g.*, FLA and FSL). By contrast, if they are orthogonal to each other, they are not correlated (*e.g.*, FLA and R848). Alternatively, when a stimulation vector is close to the center (*e.g.*, LPS in PC1 vs. PC2), it means that information is carried in the other axes (*e.g.*, in the case of LPS almost all variance is carried by PC3 and PC4).

Collectively these data suggest that FLA and FSL induced transcriptional signatures are highly correlated to the IL1B stimulation response; pIC, GARD, R848 and ODN are correlated with type I or type II IFN stimulation; and LPS is intermediate between the two.

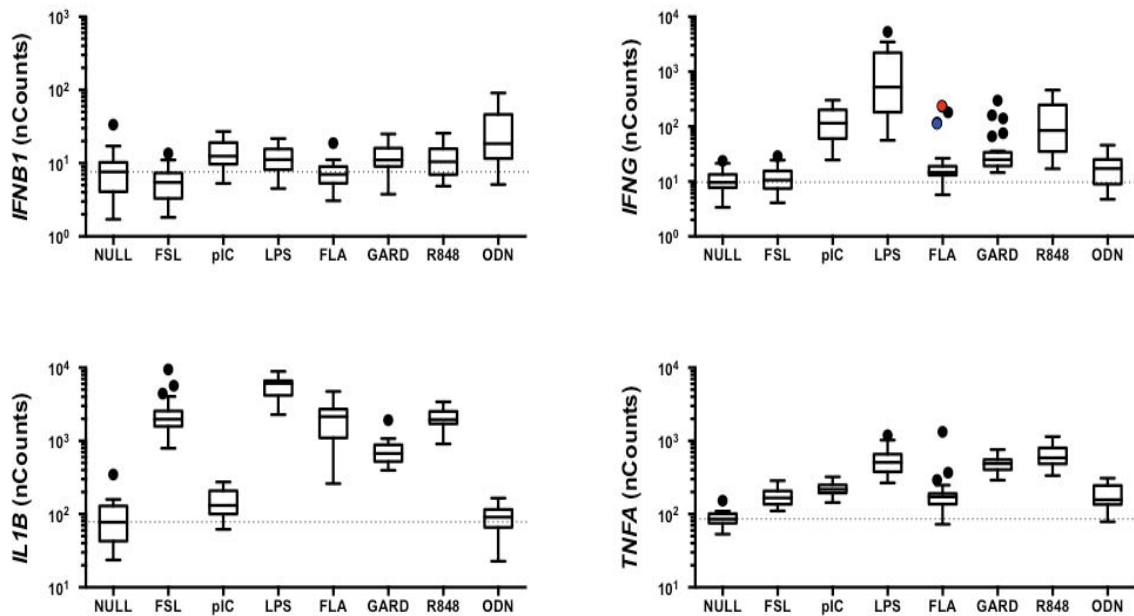


**Figure 18 : TLR induced gene expression can be represented as a function of cytokine-induced gene signatures**

Correlation circles of unit length were constructed using the 44 gene set and PCA loadings were obtained using the gene expression dataset from the four cytokine stimuli (as defined in **Table 6**). The vectors for TLR induced gene expression signatures were generated from the median value for the 25 donors, projected onto the correlation circles across the four PC.

These results were consistent with the TLR induced expression of *IFNB1*, *IFNG*, *IL1B* and *TNFA* (**Figure 19**).

One unanticipated result was the similarity between FLA and FSL, and the IL1B gene expression signature. In the case of FLA, we suggest this may be occurring due to the engagement of the intracellular sensor NLRC4, in turn activating caspase-1 (Gay et al., 2014); however the mechanisms underlying FSL activation of the inflammasome also remains uncharacterized. Notably, these analyses also identified the two outlier individuals discussed above, who showed high expression levels of FLA induced *IFNG* (blue and red dots, **Figure 19**).



**Figure 19 Four cytokine gene induced expression upon TLR stimulation**

*IFNB1*, *IFNG*, *TNFA*, and *IL1B* gene expression nCounts are shown for the Null and TLR stimulation conditions. Data is represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation. Two individual outliers (identified by their induction of *IFNG* expression in response to IL1B stimulation, [Figure 15A](#)) are indicated using blue and red circles, respectively

We applied the same approach to characterize several less well-studied agonists.

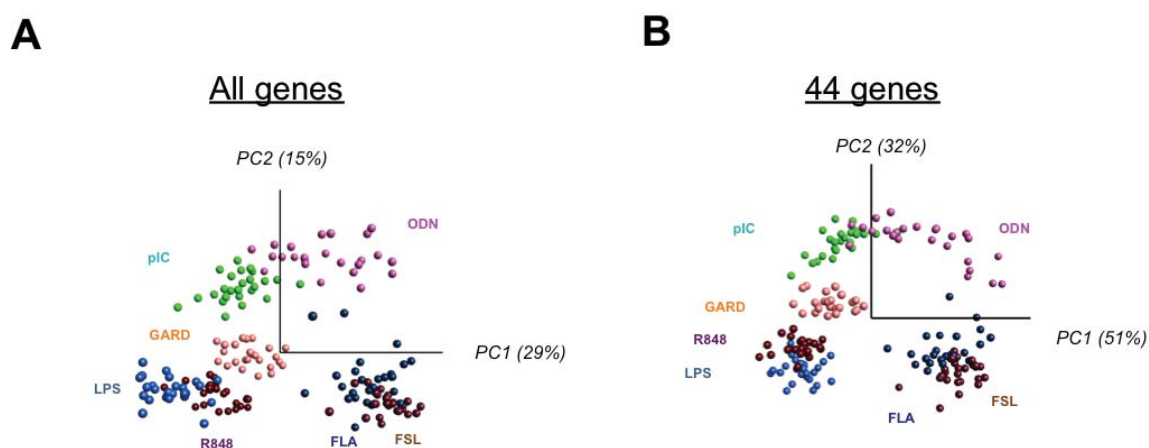
These included whole  $\beta$ -glucan particles (WGP) derived from *Saccharomyces cerevisiae*, known to engage Dectin-1 and lacking TLR-stimulating activity (Li et al., 2007); lipoarabamanin (LAM), a component of mycobacterial cell walls and an inducer of TLR2; and calcium pyrophosphate dihydrate crystals (CPPD), the etiological agent of pseudogout (Martinon et al., 2006), and a stimulator of NLRP3.

Consistent with inflammasome activation, CPPD mapped to the *IL1B* cluster, and similar to FSL1, we demonstrate that the LAM induced gene expression overlaid the *IL1B* gene set ([Figure 17B](#)). By contrast, WGP induced an mRNA expression signature that projected between *IL1B* and *TNFA*. Extension of this method may support the classification of unknown adjuvants or innate stimuli.

Next, we performed unsupervised PCA on the TLR stimulated gene expression data using the entire 572-gene set ([Figure 20A](#)).

The first two PCs, capturing 44% of the total variance, segregated all TLR stimuli with the exception of FLA and FSL (shown to have similar gene expression patterns), and to a lesser extent LPS and R848.

The clustering achieved with the entire dataset was then compared to a PCA plot built using the 44 gene signature, selected for the four effector cytokines (**Table 6**). Strikingly, the vectors built from the cytokine-gene set fully captured the diversity of responses among the TLR stimuli (**Figure 20B**).

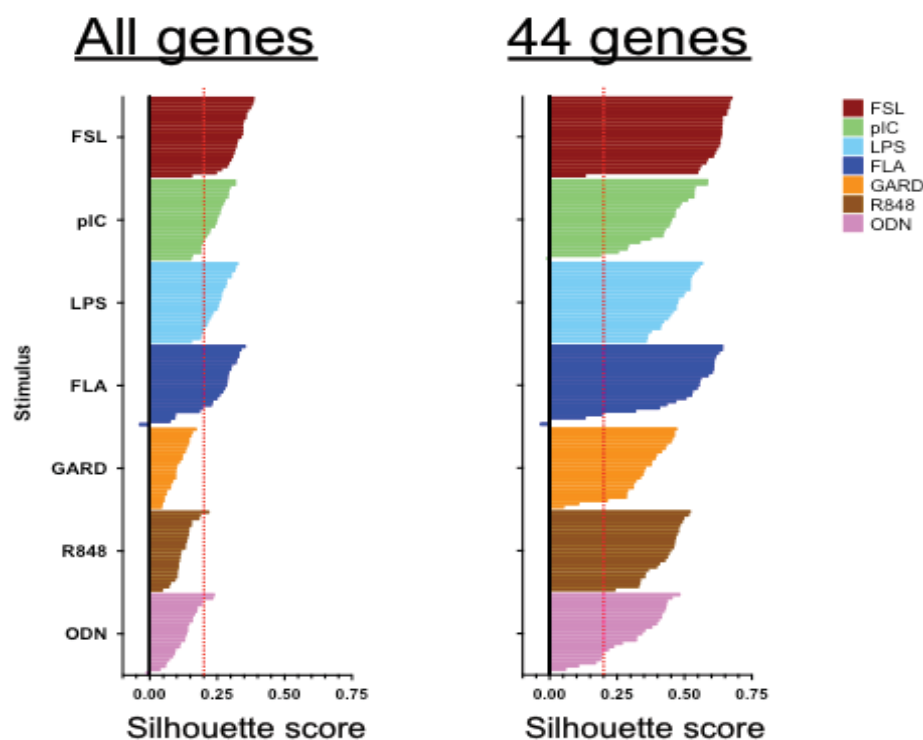


**Figure 20 : Distinct and variable response to TLR agonist stimulation can be captured using the cytokine induced 44 gene signature**

- (A, B) Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with FSL (maroon), pIC (green), LPS (light blue) FLA (dark blue), GARD (orange), R848 (brown) and ODN (pink). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes
- (A) PC1 v PC2, the percentage of variance captured by each PC is indicated). A parallel PCA was constructed using the mRNA expression data from the filtered set of 44 cytokine induced genes (from gene lists reported in **Table 5**)
- (B) PC1 v PC2, the percentage of variance captured by each PC is indicated.

Moreover, the cytokine-optimized gene set provided improved definition of the clusters, as indicated by a higher silhouette scores (**Figure 21**). This is most evident for the improved discrimination of LPS from R848 (**Figure 20B**, see PC2; and an increase in the median silhouette score from 0.26 to 0.46 for LPS, and from 0.11 to 0.35 for R848 samples, **Figure 21**).

These observations support the hypothesis that, in situations of limited agonist concentration and heterogeneous cell types, the characteristic TLR gene signatures can be identified by a limited set of cytokine-induced genes. From the perspective of population-based studies, this introduces the concept that a handful of highly discriminatory gene expression responses are sufficient to distinguish the transcriptional landscape activated by TLR pathways.



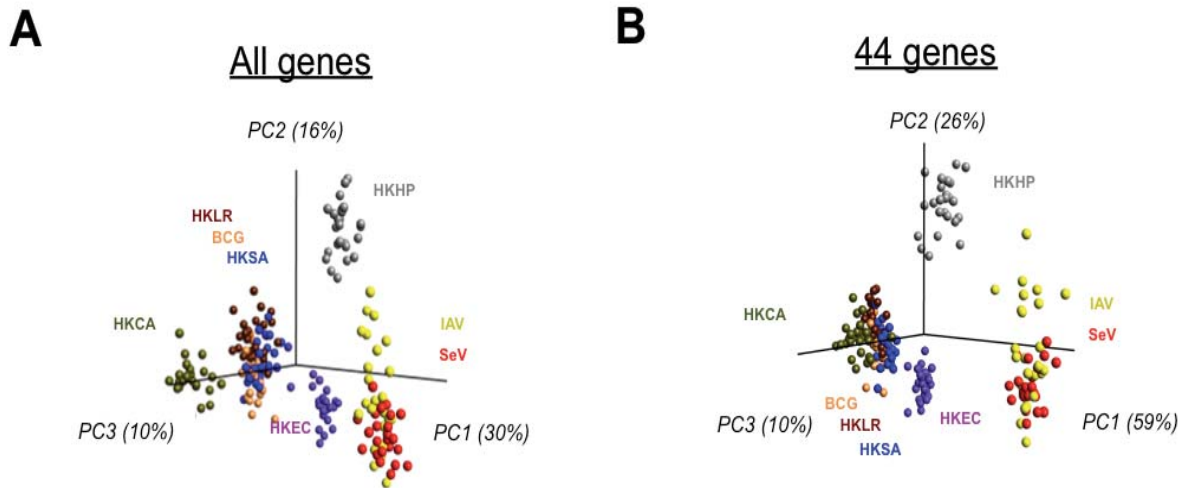
**Figure 21 : Comparison of the silhouette score from the clusters identifying the response to TLR agonists using 572 gene versus 44 gene signature**

Silhouette scores were determined for each sample based on k-means clustering ( $k=7$ ). Samples are plotted according to TLR stimulus. The red-line indicates a silhouette score of 0.2 (considered a strong fit). The median silhouette score for 572-gene set was 0.19; and for the 44-gene set it was 0.45

To test the robustness of this prediction, we subsequently evaluated the gene expression patterns induced by whole microbes, first using the entire 572-gene set (**Figure 22A**).

The microbes included heat-killed *Escherichia coli* O111:B4 (HKEC), *Staphylococcus aureus* (HKSA), *Lactobacillus rhamnosus* (HKLR), *Helicobacter pylori* (HKHP), *Candida albicans* (HKCA), a clinical preparation of live bacillus Calmette-Guerin (BCG), H1N1 attenuated influenza A/PR8 (IAV) and Sendai virus (SeV). The first three principal components, capturing 56% of the total variance, segregated samples from the viral stimuli and HKEC from the other microbes in PC1; HKHP was separated by PC2; and the remaining microbes falling along PC3 with HKCA being distinguishable from HKLR, HKSA and BCG.

Again, we demonstrated improved clustering when using the 44-gene set, as defined by the response to the four effector cytokines (**Fig 11B,C**). Strikingly, when using the 44-gene set, the variance captured by the first three principle components reached 95% (**Figure 22B**).



**Figure 22 : Distinct and variable response to microbial stimulation can be captured using the cytokine induced 44 gene signature**

Whole-blood stimulation was also preformed using HKHP (grey), HKLR (brown), HKSA (blue), HKEC (purple), HKCA (grey-green), BCG (orange), IAV (yellow) and SeV (red).

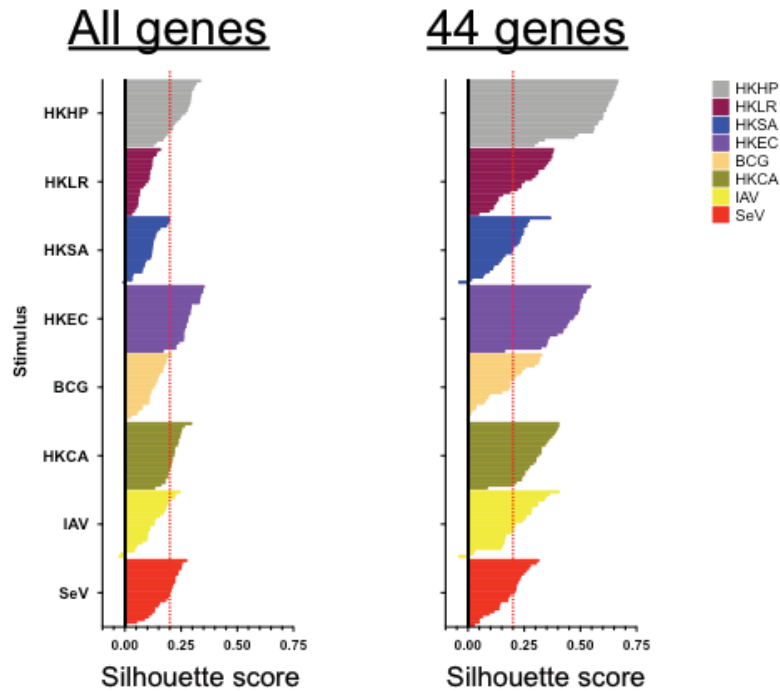
(A) PCA was used to project mRNA expression data from 572 genes.

(B) The parallel PCA was constructed using the mRNA expression data from 44 genes.

Indeed, even with whole microbe stimulation – representing a higher level of biological complexity due to the activation of multiple signaling pathways – we obtained an improved silhouette scores for k-means clustering across all stimuli when the PCA was based on the 44-gene set (**Figure 23**).

For example, the clustering of HKHP samples improved from a median silhouette score of 0.27 to 0.52, when applying the selected 44-gene set in place of the complete 572 genes. Notably, HKLR, HKSA and BCG were less distinguishable, likely a result of common agonist activity and similar levels of induced cytokines. IAV and SeV also co-segregated for similar reasons. Nonetheless, a doubling of the median silhouette score indicated that here too a focused feature list improved clustering of the data.

In light of these results, we conclude that a standardized sample collection combined with precise measurement of induced gene expression allows to massively reduce the dimensionality of the data space, while preserving the ability to discriminate the inflammatory trigger as well as the variability among human donors.



**Figure 23 : Comparison of the silhouette score from the clusters identifying the response to microbial stimulation using 572 gene versus 44 gene signature.**

Silhouette scores were determined for each sample based on k-means clustering ( $k=8$ ). Samples are plotted according to microbial stimulus. Note, IAV and SeV were mixed among two clusters (not depicted); and two samples were misclustered using the 572-gene set *versus* five samples misclustered using the 44-gene set (not depicted). The red-line indicates a silhouette score of 0.2 (considered a strong fit). The median silhouette score for 572-gene set was 0.18; and for the 44-gene set it was 0.26

### 2.3. Inter-individual variable gene expression supports tracing of cytokine loops

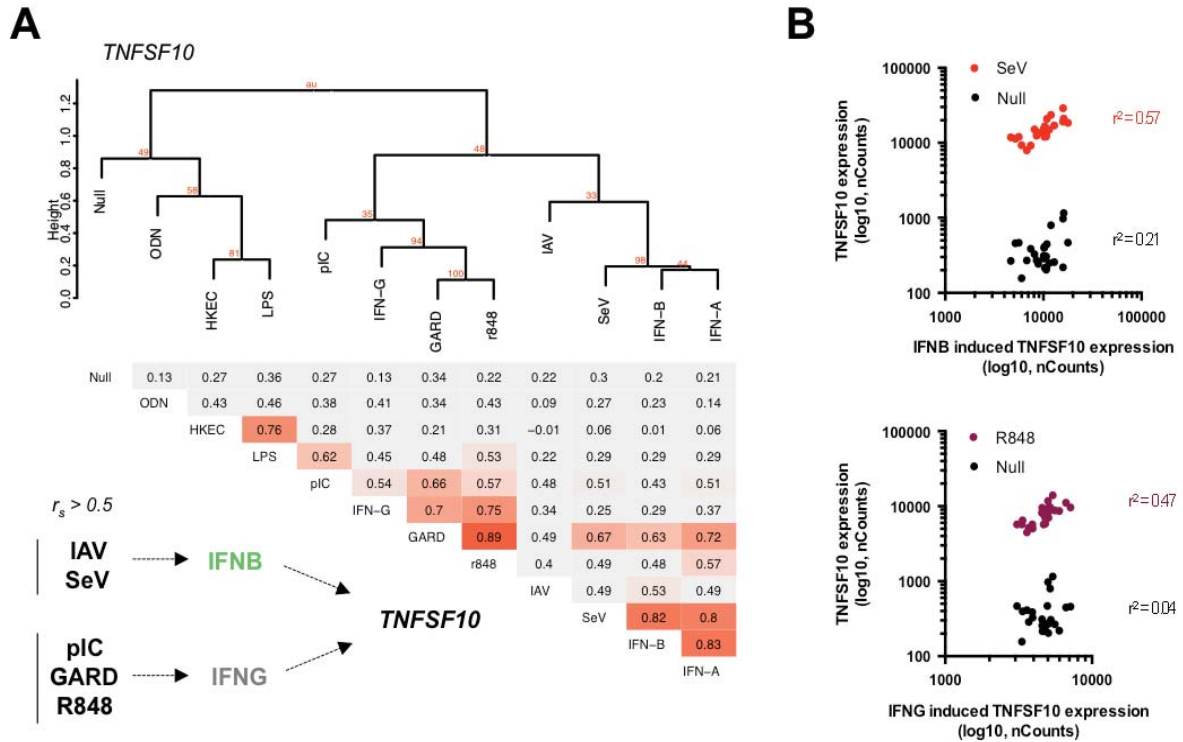
We next extended the concept of correlation among the stimulation conditions to shed lights onto possible cytokine loops involved in individual gene expression. This approach provides an exploratory analysis of possible cell-to-cell interactions that can be tested in future experimental studies.

Spearman correlation matrices and hierarchical clustering, based on a connected correlation dissimilarity metric, were performed for each gene, and results were bootstrapped to ensure the identified correlations were robust. Using these outputs, we identified cases where the variable responses to TLR or microbe stimulations could be explained by the inter-individual gene expression variance observed when using one of the four-cytokine stimuli.

To illustrate this observation, the dendrogram depicting the clusters of Spearman correlations and a table indicating the respective  $r_s$  coefficients are shown for *TNFSF10* (Figure 24A).

A cut off value of 2-fold expression change greater than the null condition was utilized for inclusion of stimuli in the cluster.

Interestingly, the viral stimuli clearly clustered with type I IFN stimulation with SeV showing a high correlation with IFNB induced *TNFSF10* ( $r_s = 0.82$ ); whereas GARD and R848 clustered with IFNG ( $r_s = 0.7$  and  $0.75$ , respectively) (Figure 24A,B).



**Figure 24 : Correlation among *TNFSF10* stimulus-induced gene expression helps to trace cytokine loops**

(A-B) Gene expression data from all twenty-three stimulation conditions were used to generate Spearman correlation matrices and hierarchical cluster analysis followed by bootstrapping.

(A) The dendrograms shown depict clustering of stimuli based on Spearman correlations for *TNFSF10* and the associated triangular matrix indicates the respective pairwise  $r_s$  coefficients.

(B) Scatter plots for indicated stimulation pairs are shown. Each dot represents the absolute nCount for a single individual of the 25 healthy donors tested for *TNFSF10*.

Red numbers at the intersection of the dendrogram branches indicate approximately unbiased (au) p-values, reported as percentage for 1000 sampled dendrograms. Color scale on table indicates strength of correlation. Proposed schematics for stimulus-driven cytokine induced gene expression is proposed using indicated cut-off for  $r_s$

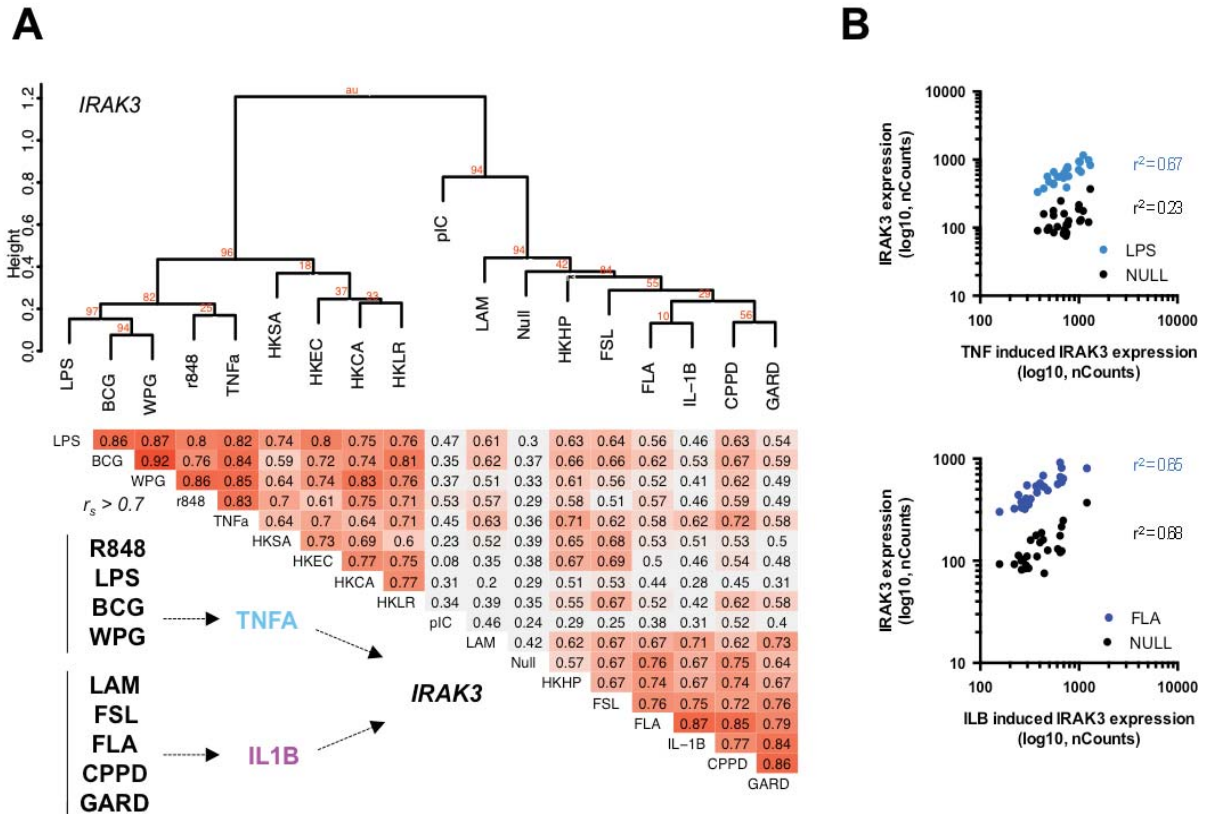
As a second example *IRAK3* is shown, illustrating distinct clustering of bacterial / TLR stimuli with TNF or IL1B (Figure 25A,B).

Schematic depictions of the putative stimulus-induced cytokine-mediated expression of *TNFSF10* or *IRAK3* are shown with dotted line arrows provided for illustrative purposes. This



analytical approach allows us to predict the distinct cytokine loops that drive common gene expression following stimulation by TLR agonists or microbes.

While this modeling approach to population-based data must be experimentally validated, we highlight the possibility that inter-individual variance can be utilized as a means to identify causal pathways driving gene expression, which will support future experimental inquiry.



**Figure 25 : Correlation among IRAK3 stimulus-induced gene expression helps to trace cytokine loops**

(A-B) Gene expression data from all twenty-three stimulation conditions were used to generate Spearman correlation matrices and hierarchical cluster analysis followed by bootstrapping.

(A) The dendrograms shown depict clustering of stimuli based on Spearman correlations for *IRAK3* and the associated triangular matrix indicates the respective pairwise  $r_s$  coefficients.

(B) Scatter plots for indicated stimulation pairs are shown. Each dot represents the absolute nCount for a single individual of the 25 healthy donors tested for *IRAK3*.

Red numbers at the intersection of the dendrogram branches indicate approximately unbiased (au) p-values, reported as percentage for 1000 sampled dendrograms. Color scale on table indicates strength of correlation. Proposed schematics for stimulus-driven cytokine induced gene expression is proposed using indicated cut-off for  $r_s$

## 2.4. Unique microbial gene expression is defined by lymphocyte-derived cytokines

While the four cytokines studied herein represent major effector pathways in host response and disease pathogenesis, we were cognizant of additional upstream factors that help to specify the inflammatory reaction.

To identify other potential effector cytokines, we generated a list of genes upregulated by each stimulus as compared to the null condition (stimulus > null, paired t-test  $q < 10^{-3}$ ), and then merged the resulting gene lists for the four cytokines, the seven TLR and the eight microbial stimuli.

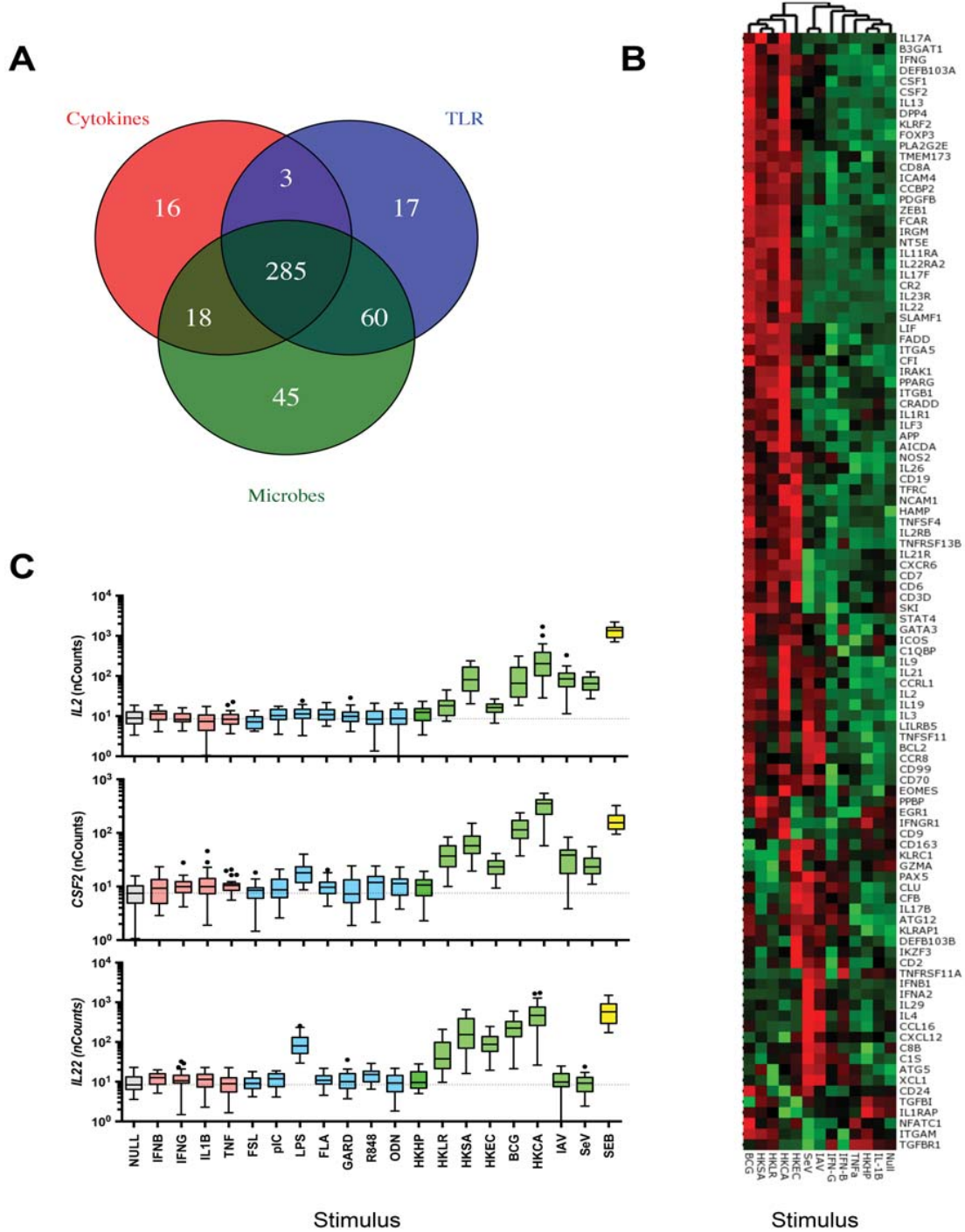
A Venn diagram depicts the overlap and intersections in gene expression for these three groups, respectively (**Figure 26A**). Additionally, we calculated the median gene expression for each stimulus and generated heat maps, clustering by both genes and samples, using either the set of genes that were expressed after microbial but not cytokine stimulation (**Figure 26B**); TLR but not cytokine stimulation (**Figure 27A**); and microbial but not TLR stimulation (**Figure 27B**).

Strikingly, the complex stimuli induced a subset of genes indicative of lymphocyte activation. This subset of genes included: (i) transcription factors such as *FoxP3* (highly induced after bacterial stimulation), *EOMES* (induced by HKCA) and *GATA3* (induced by BCG); (ii) cytolytic effectors such as *GZMA* (highly induced by HKEC); (iii) anti-microbial genes such as *NOS2* (induced after bacterial stimulation), *DEFB103A* (induced by BCG) and *HAMP* (highly induced by HKEC) (**Figure 27B**).

Additionally, we detected the differential induction of 18 cytokines, which included *IL2* (induced by HKSA, BCG, HKCA, IAV and SeV), *CSF2* (highly induced by HKCA) and *IL22* (induced after bacterial and HKCA stimulation) (**Figure 26C**).

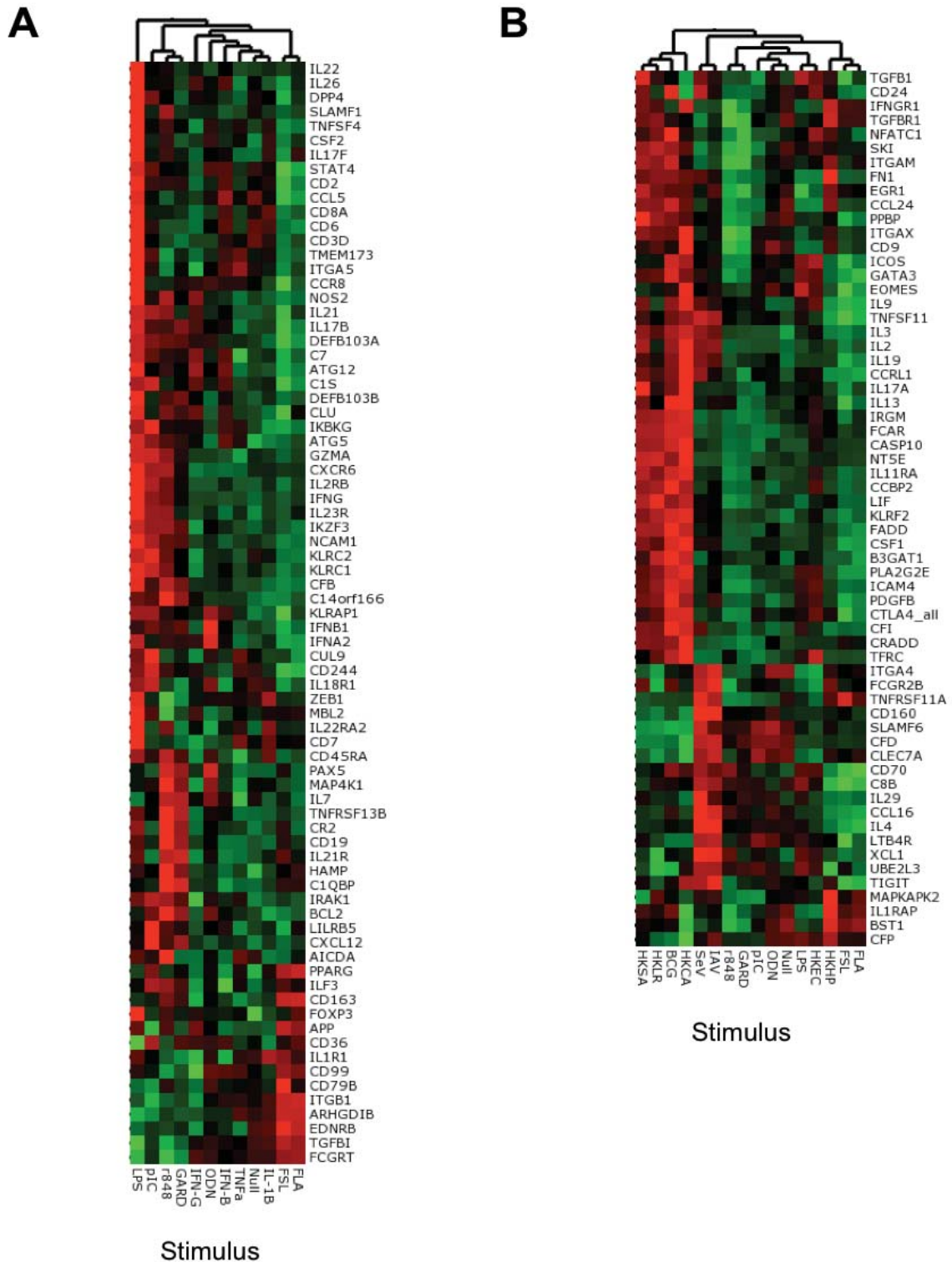
As indicated by the comparison with Staphylococcal enterotoxin B (SEB) stimulation and consistent with the presence of microbial antigen-specific T cells within the repertoire of healthy donors (Becattini et al., 2015; Geiger et al., 2009), these cytokine genes likely reflect the activation of lymphocyte subsets (**Figure 26C**).

The characterization of these lymphocyte-derived cytokines may further establish the role of feed-forward cytokine loops in the deconvolution of microbial-induced gene signatures.



**Figure 26 : Microbial induced lymphokines are absent from TLR and cytokine gene expression signatures**

- (A) Gene expression data from all twenty-three stimulation conditions were used to generate stimulus-induced signatures (stimulus > null, paired t-test with q-value cut-off of  $10^{-3}$ ). The union sets of cytokine (IFNB, IFNG, IL1B, TNFA); TLR (FSL, pIC, LPS, FLA, GARD, R848, ODN); and microbes (HKHP, HKLR, HKSA, HKEC, HKCA, BCG, IAV, SeV) were generated. The Venn diagram indicates the number of shared and unique genes among the three groups of stimuli.
- (B) Hierarchical clustering of the donors and genes based on the 105 genes present in the union set of microbes but not cytokines was performed. The median gene expression value was used for each stimulus, with variables log transformed, mean centered per donor and scaled to unit variance. NB, the dendrogram for clustering of genes not shown.
- (C) Representative gene expression data is shown for *IL2*, *CSF2* and *IL22* for each stimuli, as well as Staphylococcal enterotoxin B (SEB) stimulation for reference. Data is represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation.



**Figure 27 : Gene expression patterns not captured by four effector cytokine induced changes**  
 Hierarchical cluster analysis of donors and gene expression based on genes expressed after microbial stimulation but not cytokine stimulation (A), and TLR but not cytokine stimulation (B) as defined by firstly by a paired T test for all stimuli versus null ( $q < 10^{-3}$ ) and merging the gene lists.

### 3. Discussion

In this study, we aimed at testing if standardized whole-blood stimulation systems can support the identification of a handful set of genes that are capable of deconvoluting complex responses to immune stimulation.

We utilized medically relevant stimuli to determine their inflammatory signatures and, in doing so, established the degree of naturally-occurring variation present in a population of well-defined healthy donors of European descent. The definition of host immune responses to adjuvants and microbial agents, and subsequent characterization of inter-individual variability in the human population, is of major fundamental interest and provides the necessary foundation for understanding human health and disease pathogenesis.

Although functional tests are routinely used in laboratory investigation (Folds and Schmitz, 2003), the standardization of such assays has been challenging. While whole blood assays are more biologically relevant and introduce less experimental bias than, for example, PBMC stimulation, they are not without technical challenges in particular due to the high levels of globin RNA and enzyme-inhibiting compounds (e.g., heparin interference of reverse transcriptase) (Chaussabel et al., 2010).

Previous efforts have focused on removing the globin RNA before downstream analysis, however these processes can introduce, in turn, higher levels of technical variance as compared to what was achieved with our data generation pipeline (Shin et al., 2014).

Specifically, the innovation brought forward in this study is an automated single step RNA extraction method from whole blood, which minimized pre-analytic bias and generated highly reproducible results when using a gene hybridization read-out. These solutions are essential for multicenter population-based studies, as well as for assays with ambitions for clinical deployment.

Using the reference data presented herein, we tested the hypothesis that responses to TLR ligands or whole microbes can be captured by the transcriptional signature of key effector cytokines.

We tested a total of 23 stimulation systems, all built into whole-blood syringes for point of care sampling.

Using linear SVM learning, it was possible to identify a 44-gene set, selected based on their ability to differentially cluster cytokine-induced genes. Strikingly, these same genes, when

applied to the stratification of response to TLR or microbes, resulted in improved discrimination among the stimuli as indicated by a marked improvement in silhouette scores. In the era of an increased use of whole-genome transcriptional profiles, our results suggest that limiting the pre-analytical bias introduced by cell separation and non-standardized stimulation protocols may be more important than obtaining greater numbers of measured genes.

In addition to sample collection and data analysis standardization, we minimized intrinsic variability by the recruitment of donors of Western European ancestry (3rd generation born in Metropolitan France). Furthermore, we minimized pre-analytic or environmental sources of variability, by applying highly precise inclusion and exclusion criteria (Thomas et al., 2015). To restrict other sources of variability, in addition to the standardization of the assay systems, all donors were sampled at the same time of day (09.00-11.00), during the same week, and in the same location.

Such a reliable monitoring of induced immune gene expression responses permitted the classification of inflammatory and host immune responses based on the variance observed in healthy donors.

In addition to defining detailed healthy reference ranges to be considered in future clinical studies, this work permitted the identification of a number of outlier responses. This included the identification of two individuals that responded to FLA or IL1B by producing IFNG and in turn expressing IFNG-stimulated genes.

Following from this observation, we extended the approach of tracing cytokine loops and gene expression pathways, using inter-individual variance and correlation among the stimulation signatures as a means to deconvolute complex transcriptional responses. This approach may also support the future classification of unknown adjuvants, innate stimuli, new pathogenic agents or the stratification of disease and treatment response.

If extended to the study of disease states, it may be possible to classify, for example, subsets of rheumatoid arthritis patients that are responsive to IL1B *versus* TNFA blockade (Gibbons and Hyrich, 2009; McInnes and Schett, 2007).

This reference data set and the applied analytical approach offers an important, useful resource to the community, nevertheless, several specific limitations should be highlighted.

First, some of the employed TLR stimuli may engage secondary receptors in addition to their commonly ascribed receptors. Notably, the observation that FLA is highly correlated with the

IL1B-induced gene signature suggests that it may also trigger NLRC4 within the whole-blood stimulation systems. This may occur within neutrophils, which express high levels of the NLRC4 inflammasome and release IL1B (Chen et al., 2014). If correct, it would also help to explain why, despite the high prevalence of dominant negative forms of TLR5 in Europeans (Barreiro et al., 2009; Hawn et al., 2003), all twenty-five donors showed an induced response after FLA stimulation (Barreiro et al., 2009).

Alternatively, TLR sensor pathways on platelets and neutrophils may be unique in their ability to engage caspase-1 (Hayashi et al., 2003).

We also observed that IAV and SeV were highly correlated with pIC, suggesting that the latter is engaging RIG-I like receptors (RLRs) in addition to TLR3.

We also acknowledge that, in the natural setting, human immune responses typically occur in mucosal tissues and, as such, stromal cells and tissue resident immune populations such as macrophages and ILCs may need to be considered to fully apply our dataset to physiologic and pathologic responses.

Lastly, our analyses consider a single analytical time point only, thus capturing a snapshot of the complexity inherent in dynamic immune responses.

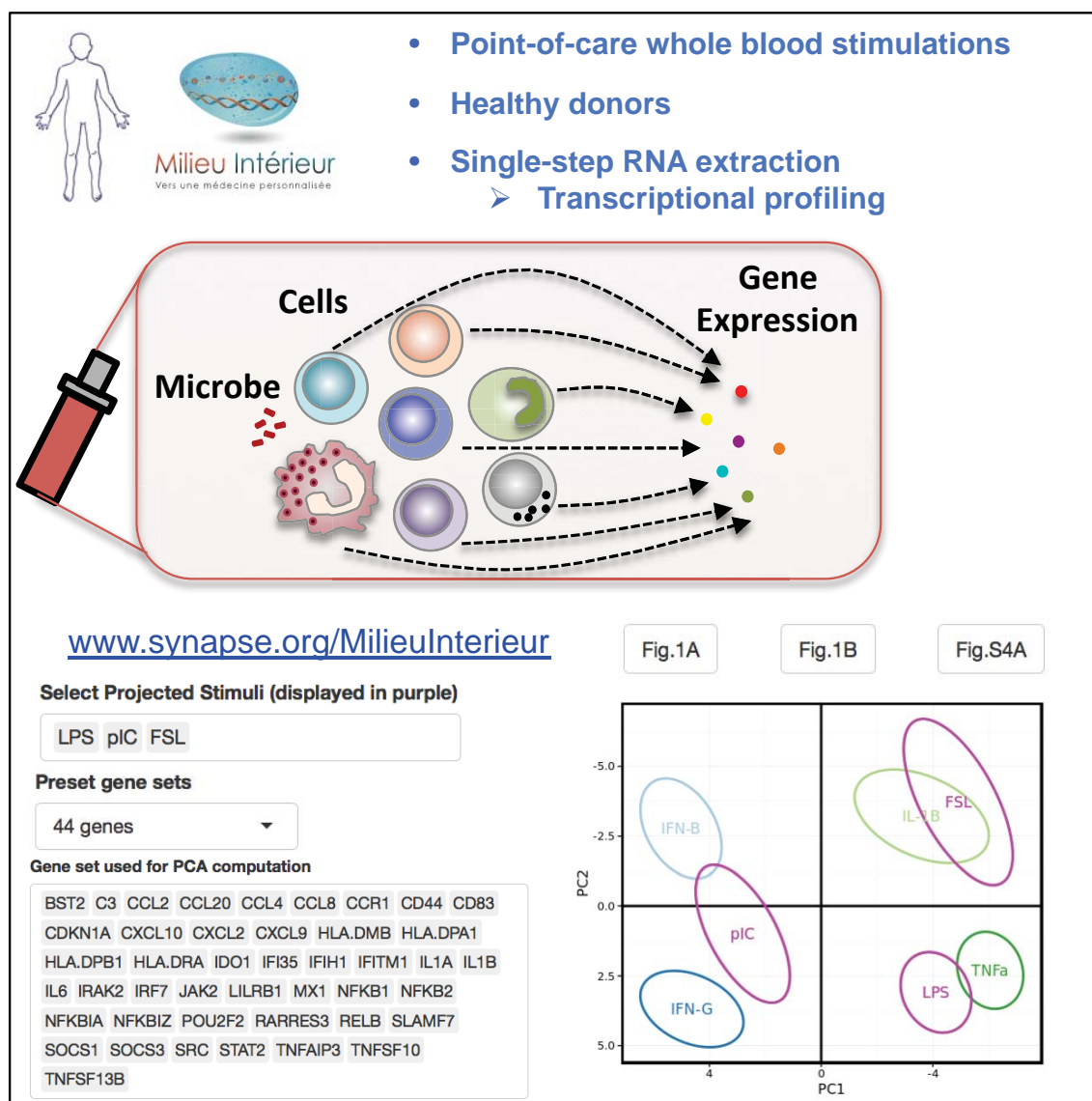
Finally, it is our aim with this resource paper to highlight the growing need to make data more accessible and easier to explore.

In line with recent efforts (Gorenshteyn et al., 2015; Speake et al., 2015), we have thus developed an online R-Shiny application software that will allow readers to fully query the data set based on their specific questions. This application software was built as a direct companion to the presented analyses with publically available R-scripts and downloading options for gene expression data. In sum, the data resource presented here, and the available online tools, provide a foundation for association studies, kinetic analyses and *in vivo* mechanistic experimentation.

For example, it remains to be established how the inter-individual variation in gene expression that we identified here is accounted for by host genetic variants (*i.e.*, expression quantitative trait loci, eQTLs), specifically in cases where gene expression variation is altered upon activation with certain immune stimuli (*i.e.*, response/interaction eQTLs).

Conceptually, the strategy to trace inter-cellular cytokine driven gene expression may support such future eQTL association studies, especially in cases where inter-cellular trans-eQTL are identified.

From a practical viewpoint, the tools will support a path towards more targeted immune monitoring from whole-blood, enabling the use of standardized approaches that capture the common variation within the human population as summarized in **Figure 28**.



**Figure 28 : Graphical abstract summarizing the study 1**

‘Standardized whole-blood transcriptional profiling enables the deconvolution of complex induced immune responses’





**NATURAL VARIATION IN INNATE IMMUNE CELL  
PARAMETERS IS DRIVEN BY GENETIC FACTORS**

## CHAPTER III: Natural variation in innate immune cell parameters is driven by genetic factors

### General presentation:

As already mentioned, the heart of the *Milieu Intérieur* study is to identify what are the natural genetic and environmental factors that drive the human immune response. To address this question 1,000 healthy individuals were genotyped and annotated with extended epidemiological data. The collection was designed to generate equilibrated groups with 50% men/women and age stratification (20 to 69 years). The circulating immune cells parameters have been measured on whole blood for the entire collection. A standardized flow cytometry analysis pipeline has been specifically developed to limit technical noise and variability across the collection timeline (almost 1 year).

This multidisciplinary work was primarily lead by E. Patin (geneticist at Institut Pasteur) for the dataset accuracy and genetic analysis, and M. Hasan (Immunologist at Institut Pasteur) for the flow cytometry dataset analysis coordination. During my PhD work, my specific contribution in the *MI* work was i) contribution in the daily automated 8-colors flow staining on whole blood (10 panels); ii) analysis of the flow cytometry data files for two panels (DCs and B cells populations); iii) leadership and execution of the replication study and iv) interpretation of results.

This analysis aims to demonstrate that the application of standardized procedures associated with a statistically significant size collection improves the confidence in the associations observed. Nonetheless, the best way to demonstrate this is to be able to replicate those findings in an independent collection. Taking advantage of a partnership with the Cancer immunology department at Genentech, Inc. (South San Francisco, USA) and their Genotype and Phenotype (gGAP) Registry, I have organized a replicative collection in the context of a student exchange program in collaboration with different facilities from the company.

Due to time, logistic and budget considerations the collection size and the parameters measured have been restricted. Interestingly, in this registry the volunteers were already genotyped. In order to keep the statistical power to confirm some associations found in the *MI* analysis, we have focused the recruitment on volunteers with European origin ancestry, from 20 to 50 years old to fit with *MI* range (but avoiding peri-menopausal women), and

homogeneously stratified for the most frequent single nucleotide polymorphisms (SNPs) that we've found associated with immune parameters. This possibility of “artificial SNPs enrichment” allowed also the restriction of the flow cytometry analysis to the immune cell populations found associated to those genetic polymorphisms. Based on the *MI* analysis, three 8-colour panels were selected to determine the quantitative and qualitative characteristics of the main circulative populations (“General” panel), natural killer cells (“NK” panel) and polymorphonuclear cell populations (“PMN” panel). Following a power calculation, we discovered that a sample size of 150 individuals was sufficient to replicate our top GWAS findings.

Then using the standardized technical pipeline from the *MI* collection as a template, some modifications have been optimized and validated due to local logistic differences.

In this chapter of the thesis, I will first present the *MI* analysis that aims to determine genetic and environmental factors that impact the most the natural variance of circulating immune cell parameters measured by flow cytometry.

This will provide a context for my presentation of the preliminary analyses of this replicative study.

## 1. Introduction

The immune system plays an essential role in maintaining homeostasis in individuals challenged by microbial infections, a physiological mechanism conceptualized by the French physician Claude Bernard in 1859, under the term *milieu intérieur*. Host-pathogen molecular interactions trigger immune responses through the activation of specialized immune cell populations, which may 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 adaptive cells that are recirculating through lymphoid organs. Past or chronic latent infections have been reported to profoundly perturb subsets of circulating immune cells due to altered trafficking, selective expansions or attrition (Park and Rehmann, 2014; Altfeld and Gale, 2015; Orme et al., 2015). Several studies suggested that extensive differences in white blood cell composition are also present among healthy individuals (Tollerud et al., 1989; Reichert et al., 1991). However, the degree of naturally occurring variation of immune cell parameters, together with the environmental and genetic determinants of such variation, remain to be fully characterized.

Standardized flow cytometry in relatively small studies of healthy individuals has highlighted the predominant effect of age on several T cell subpopulations (Tollerud et al., 1989; Sridharan et al., 2011), which may be due in part to diminished thymic activity (Sauce and Appay, 2011) and explain reduced vaccination efficacy in the elderly (Buchholz et al., 2011). Latent cytomegalovirus (CMV) infection, detected in 40% to >90% of the general population (Boeckh and Geballe, 2011), has been associated with an increased number of effector memory T cells (Sylwester et al., 2005; Libri et al., 2011; Wertheimer et al., 2014), which could in turn alter immune responses to heterologous infection (Furman et al., 2015).

These findings illustrate how critical it is to define reference ranges of major white blood cell subsets, based on large cohorts of healthy subjects from different age and CMV serological status groups. This will support the characterization of pathological states, and ultimately the development of personalized strategies for clinical management of patients. Furthermore, studies on such cohorts could help to establish how lifestyle habits, such as smoking and diet, modify white blood cell subset distribution and provide insights into the way these parameters influence immune functions.

Recent technological advances in flow cytometry and genome-wide DNA genotyping now allow the dissection of the genetic basis of population variation in immune cell parameters.

A seminal genome-wide association study reported 13 genetic loci strongly associated with the proportion of different leukocyte subpopulations in a cohort of 249 Sardinian families, including CD39<sup>+</sup> activated CD4<sup>+</sup> Treg and T cells expressing high levels of CD25 or CD8 (Orrù et al., 2013).

More recently, a study reported the deep immunophenotyping of ~1,800 independent traits in 245 healthy twin pairs, identifying 11 independent genetic loci that accounted for up to 36% of the variation of 19 different traits (Roederer et al., 2015).

Finally, a third study estimated the genetic heritability of 95 different immune cell frequencies in 105 healthy twin pairs, and suggested that variation in human immune cells is largely explained by non-heritable factors (Brodin et al., 2015).

Together, these family- and twin-based studies were designed to provide valuable insights into the genetic basis of inter-individual differences in adaptive immune cell populations, omitting evaluation of several major innate cell types in circulation. As such, an integrated evaluation of both genetic and non-genetic factors driving human variation in innate and adaptive immunity remains lacking.

Here, we report the use of semi-automated flow cytometry to comprehensively profile the white blood cell composition of 1,000 healthy, unrelated individuals of western European ancestry, which compose the *Milieu Intérieur* cohort.

We used ten 8-color immunophenotyping panels to quantify the absolute numbers – as well as the expression levels of relevant immune cell markers – of a wide range of circulating immune cells, yielding a total of 168 confidently measured immunophenotypes. Notably, the *Milieu Intérieur* cohort was designed to identify both non-genetic and genetic factors that contribute to the inter-individual variation in human immune cell parameters.

We confirm that age, gender and CMV seropositivity have major and independent effects on white blood cell composition, and characterize the profound impact of smoking on circulating numbers of Treg and MAIT cells.

We identified by genome-wide association study 14 loci associated with parameters of circulating leukocyte subpopulations, 11 of which have not been previously described. Finally, we established models that quantify the respective contributions of genetic and non-genetic factors in controlling the characteristics and subset distribution of white blood cells of healthy individuals.

## 2. Results

### 2.1. Variation of immune cell parameters in the general population

Defining standardized ranges of circulating immune cells in a large cohort of healthy individuals is a critical step for tailoring clinical strategies to individual patients. The *Milieu Intérieur* cohort includes 500 men and 500 women, stratified across five decades of age from 20 to 69 years. Subjects were surveyed for a number of demographic and lifestyle variables, including nutrition, sleep, smoking, vaccination and medical histories (**Table 7**). Detailed inclusion and exclusion criteria used to define "healthy" subjects recruited into the cohort have been previously reported (Thomas et al., 2015).

Variable	Category	Description	Coded values and units
Age	Demographics	Age	years
Owens a house	Demographics	Does the subject own his/her housing?	1=landlord; 2=lesser
Physical activity	Demographics	Hours per week of physical activity during leisure	hours per week
Sex	Demographics	Clinical sex	1=male; 2=female
Lives with partner	Demographics	Subject shares or not housing with partner	0=Alone; 1=With a partner
Lives with kids	Demographics	Subject shares or not housing with children	0=Without children; 1=With children
Born in a city	Geographic origin	Born in a city with a number of inhabitants larger than 20,000	0=no; 1=yes
BMI	Basic physiological measurements	BMI	kg/m <sup>2</sup>
Heart rate	Basic physiological measurements	Heart rate	bpm
Temperature	Basic physiological measurements	Ear temperature	°C
Dietary habits	Food and nutrition	Nutrinet nutritional profiles	1,2,3,4
CMV	Laboratory measure	CMV serology	0=negative; 1=positive
Flu IgG+	Laboratory measure	Log10-transformed ratio of anti-IAV IgG levels and the assay threshold value	NA
Metabolic score	Laboratory measure	Metabolic score, estimated as described in Thomas et al., Clin Immunol 2015	Number of risk factors for the Metabolic Syndrome, ranging from 0 to 5
Smoking	Smoking habits	Tobacco smoking	Non-smoker = 0, Ex-Smoker = 1, Smoker = 2
Cannabis	Sleep habits, drug habits, and psychological problems	Haschich smoking	0 = never, 1 = rarely, 2 = regularly
Low appetite	Sleep habits, drug habits, and psychological problems	Little or too much appetite in last 2 weeks	0, 3, 8, 14 days the last two weeks
Concentration	Sleep habits, drug habits, and psychological problems	Difficulty concentrating on things like reading newspapers or watching the television, in last 2 weeks	0, 3, 8, 14 days the last two weeks
Sleep problems	Sleep habits, drug habits, and psychological problems	Does the subject often find it difficult to fall asleep or to remain asleep?	0,1,2,3 = Never, Sometimes, occasionally, most of the time, all the time
Hours of sleep	Sleep habits, drug habits, and psychological problems	Hours of sleep	hours/day
Listless	Sleep habits, drug habits, and psychological problems	Feeling tired or having little energy, last 2 weeks	0, 3, 8, 14 days the last two weeks
Depression score	Sleep habits, drug habits, and psychological problems	Poor self-image, or lack of interest or pleasure in doing things, or feeling sad, depressed or despairing, during last 2 weeks	0, 3, 8, 14 days the last two weeks
Recent crisis	Sleep habits, drug habits, and psychological problems	Major negative life event (e.g., death of a close relative) in last 12 months	0 = no, 1 = yes
Employed	Socio-professional information	Steady job	0=No, 1=Yes; NA=missing data
Education	Socio-professional information	Level of education	1=No diploma; 2=Primary school certificate; 3=CAP, BEP, Brevet de colleges; 4=Baccalauréat; 5=Higher education, cycle 1 (DUT, BTS, DEUG, L2); 6=Higher education, cycle 2 and 3 (L3, M1, M2, PhD)
Exposure to dust	Socio-professional information	Exposure to dust	0=No exposure; 1=Past exposure; 2=Current exposure; NA=Unknown
Income	Socio-professional information	Net monthly income of the household	1=0-1000€ 2=1001-2000€ 3=2001-3000€ 4=3001-4000€ 5=4001-5000€ 6=5001€ and more; NA=missing data
	Batch effect	Hour at which 25ml blood sample was taken	hours
	Batch effect	Date at which V1 was done	days since 09-01-2012
Had measles	Medical history	Childhood disease: Measles	0=No; 1=Yes
Had rubella	Medical history	Childhood disease: Rubella	0=No; 1=Yes
Had chicken pox	Medical history	Childhood disease: Chicken pox	0=No; 1=Yes
Had mumps	Medical history	Childhood disease: Mumps	0=No; 1=Yes
Tonsillectomy	Medical history	Tonsillectomy	0=No; 1=Yes
Appendectomy	Medical history	Appendectomy	0=No; 1=Yes
MMR vaccine	Vaccination history	Vaccination against mumps, rubella and measles	0=No; 1=Yes
Typhoid vaccine	Vaccination history	Vaccination against typhoid	0=No; 1=Yes
Whooping cough vac.	Vaccination history	Vaccination against whooping cough	0=No; 1=Yes
Yellow fever vaccine	Vaccination history	Vaccination against yellow fever	0=No; 1=Yes
HAV vaccine	Vaccination history	Vaccination against Hepatitis A	0=No; 1=Yes
HBV vaccine	Vaccination history	Vaccination against hepatitis B	0=No; 1=Yes
Flu vaccine	Vaccination history	Vaccination against flu	0=No; 1=Yes

**Table 7 : Demographic variables included in the *MI* study**

Panel	Dye	Specificity	Clone (Company)
<b>Panel 1 (T cells)</b>			
	eF450	CD3	SK7 (eBio)
	V500	HLA-DR	L243/G46-6 (BD)
	FITC	CD45RA	L48 (BD)
	PE	CD8 $\alpha$	BW135/80 (Miltenyi)
	PerCP-eF710	CD27	O323 (eBio)
	PE-Cy7	CD8 $\beta$	SID18BEE (eBio)
	APC	CCR7	FR11-11E8 (Miltenyi)
	APC-H7	CD4	SK3 (BD)
<b>Panel 2 (Treg)</b>			
	eF450	ICOS	ISA-3 (eBio)
	eF506	FVD eF 506	NA (eBio)
	FITC	CD45RA	L48 (BD)
	PE	CD25	NA (Miltenyi)
	PerCP-Cy5.5	HLA-DR	L243/G46-6 (BD)
	PE-Cy7	CD8 $\beta$	SID18BEE (eBio)
	APC	CD127	MB15-18C9 (Miltenyi)
	APC-H7	CD4	SK3 (BD)
<b>Panel 3 (MAIT/NKT cells)</b>			
	eF450	CD3	SK7 (eBio)
	V500	HLA-DR	L243 G46-6 (BD)
	FITC	V $\alpha$ 7.2	3C10 (BioLegend)
	PE	$\gamma$ 8TCR	11F2 (BD)
	PerCP-Cy5.5	CD161	HP-3G10 (eBio)
	PE-Cy7	CD8b	SID18BEE (eBio)
	APC	V $\alpha$ 24	6B11 (eBio)
	APC-H7	CD4	SK3 (BD)
<b>Panel 4 (NK cells)</b>			
	eF450	CD3	SK7 (eBio)
	VioBlue	CD14	TUK4 (Miltenyi)
	525/50BP	FVD eF 506	NA (eBio)
	FITC	CD69	L78 (BD)
	PE	CD335 (NKp46)	9E2/NKp46 (BD)
	PerCP-Cy5.5	HLA-DR	L243/G46-6 (BD)
	PE-Cy7	CD8 $\alpha$	SK1 (BD)
	APC	CD16	B73.1 (BD)
	APC-eF780	CD56	CMSSB (eBio)
<b>Panel 5 (Lineage)</b>			
	eF450	CD3	SK7 (eBio)
	V500	CD14	M5E2 (BD)
	FITC	CD56	NCAM16.2 (BD)
	PE	CD45	H130 (BD)
	PerCP-Cy5.5	CD16	3G8 (BD)
	PE-Cy7	CD8b	SID18BEE (eBio)
	APC	CD19	SJ25C1 (BD)
	APC-H7	CD4	SK3 (BD)
<b>Panel 6 (B cells)</b>			
	V450	IgM	G20-127 (BD)
	V500	IgD	IA6-2 (BD)
	FITC	IgG	IS11 3B2.2.3 (Miltenyi)
	PE	CD38	1B6 (Miltenyi)
	PerCP-eFluor710	CD27	O323 (eBio)
	PE-Cy7	CD21	B-ly4 (BD)
	APC	CD19	SJ25C1 (BD)
	APC-eF780	CD24	SN3 (eBio)
<b>Panel 7 (PMN cells)</b>			
	eF450	CD62L	Dreg 56 (BD)
	eF506	FVD eF 506	NA (eBio)
	FITC	FC $\epsilon$ R1 $\alpha$	AER-37 CRA1 (eBio)
	PE	CDw125	A14 (BD)
	PerCP-Cy5.5	CD16	3G8 (BD)
	PE-Cy7	CD32	FL18.26 (BD)
	APC	CD203c	FR3-16A11 (Miltenyi)
	APC-eF780		
<b>Panel 8 (Dendritic cells)</b>			
	VioBlue	CD14	TUK4 (Miltenyi)
	eF506	FVD eF 506	NA (eBio)
	VioGreen	CD19	LT19 (Miltenyi)
	VioGreen	CD3	BW264/56 (Miltenyi)
	FITC	BDCA1	AD5-8E7 (Miltenyi)
	PE	BDCA2	AC144 (Miltenyi)
	PerCP	HLA-DR	AC122 (Miltenyi)
	PE-Vio770	CD86	FM95 (Miltenyi)
	APC	BDCA4	AD5-17F6 (Miltenyi)
	APC-Vio770	BDCA3	AD5-14H12 (Miltenyi)
<b>Panel 9 (T helper cells)</b>			
	V450	CCR4 (CD194)	1G1 (BD)
	VioGreen	CRTH2 (CD294)	BM16 (Miltenyi)
	AF488	CXCR5 (CD185)	RF8B2 (BD)
	PE	$\gamma$ 8TCR	11F2 (BD)
	PerCP-Cy5.5	CCR6 (CD196)	11A9 (BD)
	PE-Cy7	CD8 $\beta$	SID18BEE (eBio)
	APC	CXCR3 (CD183)	1C6/CXCR3 (BD)
	APC-H7	CD4	SK3 (BD)
<b>Panel 10 (ILC)</b>			
	eF450	CD3	SK7 (eBio)
	VioBlue	CD14	TUK4 (Miltenyi)
	eF506	FVD eF506	NA (eBio)
	FITC	CD127	eBioRDR5 (eBio)
	PE	CD294	BM16 (Miltenyi)
	PerCP-Cy5.5	CD161	HP-3G10 (eBio)
	PE-Cy7	CD117	104D2 (eBio)
	AF647	NKp44	P44-8.1 (BD)
	APC-eF780	CD56	CMSSB (eBio)

**Table 8 : Staining antibodies used for the ten 8-color flow cytometry panels.**

To describe natural variation of both innate and adaptive immune cells in the 1,000 subjects of the *Milieu Intérieur* cohort, we used ten 8-color immunophenotyping flow cytometry panels (Figures 56-65 in « Materials & Methods » and Table 8), which allowed us to report a total of 168 distinct immunophenotypes (Table 9)

*(Remark for the reader : To facilitate reading, the original Table 9 has been split into 6 parts. The last section presents the entire list of immunophenotypes used as exclusion criterion reformatted to fit on the same page as two lists aligned and separated by a grey column).*



Immunophenotype	Exclusion criterion <sup>a</sup>	Measure type <sup>b</sup>	FACS panel	Inmate <sup>c</sup>	Adaptive <sup>c</sup>	Average <sup>d</sup>	SD <sup>d</sup>	2.5th centile <sup>d</sup>	97.5th centile <sup>d</sup>	V1/V2 correlation <sup>e</sup>	Significance	Associated non-genetic covariates <sup>f</sup>
Number of CD16hi NK cells	-	Absolute count	4	1	0	7858	4730	1917	20332	0.573	2.54E-45	CMV
Number of CD56hi NK cells	-	Absolute count	4	1	0	611	327	185	1465	0.685	7.35E-71	age, smoking
Number of CD69+ CD16hi NK cells	-	Absolute count	4	1	0	272	215	37	869	0.606	8.02E-52	sex
Number of CD69+ CD56hi NK cells	-	Absolute count	4	1	0	28	19	5	80	0.462	4.06E-28	sex
Number of CD8a+ CD16hi NK cells	-	Absolute count	4	1	0	3825	2649	221	10485	0.622	3.94E-55	CMV
Number of CD8a+ CD56hi NK cells	-	Absolute count	4	1	0	198	134	39	549	0.688	9.51E-72	
Number of HLA-DR+ CD56hi NK cells	-	Absolute count	4	1	0	4	5	0	16	0.355	1.47E-16	
Number of NK cells	-	Absolute count	4	1	0	8808	4949	2554	21720	0.567	4.29E-44	CMV
MFI of CD16 in CD16hi NK cells	-	MFI	4	1	0	20910	5269	9559	31625	0.677	1.67E-68	
MFI of CD16 in CD56hi NK cells	-	MFI	4	1	0	2029	921	570	4185	0.740	3.48E-88	
MFI of CD69 in CD16hi NK cells	-	MFI	4	1	0	544	187	320	1003	0.571	7.24E-45	sex
MFI of CD69 in CD56hi NK cells	-	MFI	4	1	0	587	223	352	1055	0.435	7.65E-25	
MFI of CD69 in CD69+ CD16+ NK cells	-	MFI	4	1	0	3493	634	2576	5031	0.451	1.03E-26	
MFI of CD69 in CD69+ CD56+ NK cells	-	MFI	4	1	0	4696	1290	2420	7589	0.254	4.61E-09	
MFI of CD69 in CD8a+ CD16+ NK cells	-	MFI	4	1	0	520	172	305	933	0.539	2.35E-39	sex
MFI of CD69 in CD8a+ CD56+ NK cells	-	MFI	4	1	0	490	143	317	840	0.252	5.24E-09	sex
MFI of CD69 in HLA-DR+ CD16+ NK cells	-	MFI	4	1	0	753	629	26	2399	0.119	4.00E-03	
MFI of CD8a in CD16hi NK cells	-	MFI	4	1	0	4176	2063	1021	9305	0.830	6.17E-129	
MFI of CD8a in CD56hi NK cells	-	MFI	4	1	0	1654	789	540	3739	0.757	3.21E-94	age, smoking
MFI of CD8a in CD69+ CD16+ NK cells	-	MFI	4	1	0	3684	2170	728	8816	0.752	0.00E+00	
MFI of CD8a in CD69+ CD56+ NK cells	-	MFI	4	1	0	1516	1212	-36	4473	0.368	8.87E-18	sex
MFI of CD8a in CD8a+ CD16+ NK cells	-	MFI	4	1	0	7899	2321	4402	13116	0.795	0.00E+00	
MFI of CD8a in CD8a+ CD56+ NK cells	-	MFI	4	1	0	4411	1239	2569	7458	0.658	0.00E+00	sex, smoking
MFI of CD8a in HLA-DR+ CD16+ NK cells	-	MFI	4	1	0	3624	3269	-781	12027	0.348	6.95E-16	
MFI of HLA-DR in CD16hi NK cells	-	MFI	4	1	0	980	249	439	1462	0.651	5.75E-62	
MFI of HLA-DR in CD56hi NK cells	-	MFI	4	1	0	483	153	245	827	0.450	1.36E-26	
MFI of HLA-DR in CD69+ CD16+ NK cells	-	MFI	4	1	0	1066	269	486	1606	0.600	1.40E-50	
MFI of HLA-DR in CD69+ CD56+ NK cells	-	MFI	4	1	0	495	292	112	1262	0.140	8.50E-04	
MFI of HLA-DR in CD8a+ CD16+ NK cells	-	MFI	4	1	0	949	236	433	1421	0.666	1.19E-65	
MFI of HLA-DR in CD8a+ CD56+ NK cells	-	MFI	4	1	0	468	150	222	840	0.453	5.25E-27	
MFI of HLA-DR in HLA-DR+ CD16+ NK cells	-	MFI	4	1	0	7603	3514	3186	17560	0.235	5.63E-08	
MFI of HLA-DR in HLA-DR+ CD56+ NK cells	-	MFI	4	1	0	6923	3746	2704	15556	0.165	1.03E-04	
MFI of NKp46 in NK cells	-	MFI	4	1	0	2550	660	1470	3942	0.752	1.38E-92	
Ratio of CD16 MFI in CD16hi to CD56hi NK cells	-	Ratio	4	1	0	11.0	4.0	5.0	21.0	0.717	2.57E-80	smoking
Ratio of CD16hi to CD56hi NK cells	-	Ratio	4	1	0	15.0	10.0	3.0	40.0	0.745	0.00E+00	age

Immunophenotype	Exclusion criterion <sup>a</sup>	Measure type <sup>b</sup>	FACS panel	Inmate <sup>c</sup>	Adaptive <sup>c</sup>	Average <sup>d</sup>	SD <sup>d</sup>	2.5th centile <sup>d</sup>	97.5th centile <sup>d</sup>	V1/V2 correlation <sup>e</sup>	Significance	Associated non-genetic covariates <sup>f</sup>
Number of CD14hi monocytes	-	Absolute count	5	1	0	15083	6260	5308	29825	0.428	6.06E-24	age, sex, smoking
Number of CD16hi monocytes	-	Absolute count	5	1	0	2643	1221	854	5663	0.410	4.95E-22	age, smoking
Number of monocytes	-	Absolute count	5	1	0	17764	6939	6555	34750	0.399	7.72E-21	smoking
MFI of CD16 in CD14hi monocytes	-	MFI	5	1	0	1182	257	803	1811	0.457	2.00E-27	CMV
MFI of CD16 in CD16hi monocytes	-	MFI	5	1	0	9208	2239	4921	13929	0.653	0.00E+00	
Number of basophils	-	Absolute count	7	1	0	1440	750	381	3210	0.581	9.18E-47	Flu IgG+, temperature
Number of eosinophils	-	Absolute count	7	1	0	3468	3228	245	12913	0.585	1.26E-47	
Number of neutrophils	-	Absolute count	7	1	0	141317	65799	46151	303200	0.459	1.08E-27	
MFI of CD16 in basophils	-	MFI	7	1	0	154	34	94	231	0.420	4.59E-23	
MFI of CD16 in eosinophils	-	MFI	7	1	0	754	164	522	1176	0.500	2.71E-33	
MFI of CD16 in neutrophils	-	MFI	7	1	0	38546	10746	17378	62169	0.672	3.80E-67	
MFI of CD203c in basophils	-	MFI	7	1	0	1428	895	123	3643	0.694	1.66E-73	temperature
MFI of CD32 in basophils	-	MFI	7	1	0	7385	2208	4008	12254	0.762	0.00E+00	sex
MFI of CD32 in eosinophils	-	MFI	7	1	0	1918	1054	72	3906	0.830	6.04E-129	
MFI of CD32 in neutrophils	-	MFI	7	1	0	5370	2574	568	10142	0.767	0.00E+00	
MFI of CD62L in eosinophils	-	MFI	7	1	0	4926	942	3305	6673	0.594	2.59E-49	heart rate, smoking
MFI of CD62L in neutrophils	-	MFI	7	1	0	11598	2567	6861	16646	0.541	1.09E-39	had tonsillectomy
MFI of FcεRI in basophils	-	MFI	7	1	0	6592	3421	638	14490	0.875	0.00E+00	
MFI of FcεRI in eosinophils	-	MFI	7	1	0	611	149	406	1001	0.397	1.41E-20	
MFI of FcεRI in neutrophils	-	MFI	7	1	0	115	27	78	183	0.363	2.50E-17	
Number of Conventional DC1	-	Absolute count	8	1	0	1221	683	346	2671	0.448	2.45E-26	age
Number of Conventional DC3	-	Absolute count	8	1	0	22	15	3	56	0.325	4.22E-14	
Number of pDC	-	Absolute count	8	1	0	292	174	58	703	0.479	2.42E-30	age, smoking
MFI of CD86 in CD14hi monocytes	-	MFI	8	1	0	5785	1338	3687	8461	0.621	4.97E-55	
MFI of CD86 in cDC1	-	MFI	8	1	0	404	230	152	881	0.589	2.31E-48	
MFI of CD86 in cDC3	-	MFI	8	1	0	813	392	347	1624	0.214	6.49E-07	
MFI of CD86 in pDC	-	MFI	8	1	0	319	66	200	450	0.613	2.54E-53	smoking
MFI of HLA-DR in CD14hi monocytes	-	MFI	8	1	0	1253	402	592	2284	0.719	0.00E+00	metabolic score, smoking
MFI of HLA-DR in cDC1	-	MFI	8	1	0	4114	878	2643	6013	0.762	2.47E-96	
MFI of HLA-DR in cDC3	-	MFI	8	1	0	4837	1485	2322	8398	0.386	1.72E-19	
MFI of HLA-DR in pDC	-	MFI	8	1	0	5274	1523	2587	8313	0.793	2.49E-109	smoking

Immunophenotype	Exclusion criterion <sup>a</sup>	Measure type <sup>b</sup>	FACS panel	Inmate <sup>c</sup>	Adaptive <sup>c</sup>	Average <sup>d</sup>	SD <sup>d</sup>	2.5th centile <sup>d</sup>	97.5th centile <sup>d</sup>	V1/V2 correlation <sup>e</sup>	Significance	Associated non-genetic covariates <sup>f</sup>
Number of CD161+ ILC3	-	Absolute count	10	1	0	40	27	6	114	0.632	1.96E-57	age, sex
Number of CD56+ ILC	-	Absolute count	10	1	0	433	289	71	1150	0.552	1.78E-41	smoking
Number of ILC	-	Absolute count	10	1	0	740	393	192	1700	0.520	2.51E-36	age, smoking
Number of ILC1	-	Absolute count	10	1	0	135	73	38	321	0.496	8.91E-33	age
Number of ILC2	-	Absolute count	10	1	0	60	54	6	191	0.696	5.03E-74	age
Number of ILC3	-	Absolute count	10	1	0	74	51	13	204	0.671	4.97E-67	age, sex
Number of NCR- CD56+ ILC	-	Absolute count	10	1	0	378	263	43	1025	0.582	6.45E-47	age, sex, smoking
Number of NCR+ CD56+ ILC	-	Absolute count	10	1	0	52	60	0	217	0.440	2.00E-25	age
MFI of CD127 in ILC1	-	MFI	10	1	0	1303	187	963	1691	0.390	7.31E-20	age, owns a house, smoking
MFI of CD161 in ILC3	-	MFI	10	1	0	444	187	179	889	0.586	9.91E-48	age, sex, smoking
Number of CD4+ T cells	-	MFI	10	1	0	1416	266	917	1955	0.400	6.22E-21	age, sex, smoking
Number of CD8b- CD4- T cells	-	Absolute count	1	0	1	40455	15707	15595	77116	0.544	4.00E-40	age, sex, smoking
Number of CD8b+ T cells	-	Absolute count	1	0	1	4193	3017	826	13413	0.718	7.83E-81	age
Number of CM CD4+ T cells	-	Absolute count	1	0	1	16910	8613	6212	37342	0.599	2.41E-50	age, CMV
Number of CM CD8+ T cells	-	Absolute count	1	0	1	23024	9505	8918	46158	0.560	5.84E-43	sex, smoking
Number of EM CD4+ T cells	-	Absolute count	1	0	1	6962	3484	2129	15091	0.574	1.71E-45	CMV
Number of EM CD8+ T cells	-	Absolute count	1	0	1	3293	2420	800	10322	0.687	2.54E-71	CMV
Number of EMRA CD4+ T cells	-	Absolute count	1	0	1	883	956	69	3189	0.652	4.24E-62	CMV
Number of EMRA CD8+ T cells	-	Absolute count	1	0	1	474	1212	5	3884	0.791	9.70E-109	CMV
Number of HLA-DR+ CM CD4+ T cells	-	Absolute count	1	0	1	1680	2382	67	8404	0.780	1.31E-103	CMV
Number of HLA-DR+ CM CD8+ T cells	-	Absolute count	1	0	1	1305	791	343	3219	0.415	1.48E-22	age, smoking
Number of HLA-DR+ EM CD8+ T cells	-	Absolute count	1	0	1	882	745	139	2885	0.583	3.88E-47	CMV
Number of HLA-DR+ EMRA CD8+ T cells	-	Absolute count	1	0	1	142	185	5	656	0.600	2.00E-50	CMV
Number of HLA-DR+ EMRA CD4+ T cells	-	Absolute count	1	0	1	52	128	0	396	0.660	4.07E-64	CMV
Number of HLA-DR+ EMRA CD8+ T cells	-	Absolute count	1	0	1	226	394	3	1407	0.680	2.70E-69	CMV
Number of HLA-DR+ EMRA CD4+ T cells	-	Absolute count	1	0	1	420	371	80	1412	0.618	2.19E-54	age, CMV
Number of HLA-DR+ EMRA CD8+ T cells	-	Absolute count	1	0	1	13674	8080	2372	33365	0.745	6.61E-90	age, sex
Number of naive CD4+ T cells	-	Absolute count	1	0	1	5866	4527	677	18286	0.785	4.94E-106	age, sex
Number of naive CD8+ T cells	-	Absolute count	1	0	1	62555	24219	25989	117050	0.486	2.67E-31	age, CMV, lives with kids, smoking
MFI of CCR7 in CM CD4+ T cells	-	MFI	1	0	1	842	205	506	1254	0.571	5.85E-45	age, sex
MFI of CCR7 in CM CD8+ T cells	-	MFI	1	0	1	596	98	396	802	0.323	6.62E-14	sex
MFI of CCR7 in EM CD4+ T cells	-	MFI	1	0	1	579	83	408	733	0.407	1.26E-21	sex
MFI of CCR7 in EM CD8+ T cells	-	MFI	1	0	1	778	94	583	961	0.219	3.77E-07	sex
MFI of CCR7 in EMRA CD4+ T cells	-	MFI	1	0	1	638	147	400	954	0.257	2.59E-09	age, smoking
MFI of CCR7 in EMRA CD8+ T cells	-	MFI	1	0	1	810	102	608	1014	0.340	2.44E-15	Flu IgG+
MFI of CCR7 in naive CD4+ T cells	-	MFI	1	0	1	2035	361	1399	2771	0.728	5.14E-84	age, CMV
MFI of CCR7 in naive CD8+ T cells	-	MFI	1	0	1	2445	466	1495	3453	0.695	1.33E-73	age, sex, smoking
CD4+ to CD8+ T cell ratio	-	Ratio	1	0	1	3	1	1	6	0.910	5.42E-193	age, smoking
Number of activated Treg	-	Absolute count	2	0	1	354	181	91	782	0.665	1.68E-65	age, CMV
Number of conventional T cells	-	Absolute count	2	0	1	34899	13737	13972	66169	0.613	3.39E-53	age, sex, smoking
Number of memory Treg	-	Absolute count	2	0	1	1297	604	386	2635	0.587	6.93E-48	smoking
Number of naive Treg	-	Absolute count	2	0	1	707	464	144	1987	0.701	2.21E-75	age
Number of Treg cells	-	Absolute count	2	0	1	2374	1026	938	4773	0.587	5.06E-48	age, smoking
MFI of ICOS in activated Treg	-	MFI	2	0	1	233	44	152	332	0.480	1.56E-30	sex
MFI of ICOS in memory Treg	-	MFI	2	0	1	111	18	77	150	0.449	1.90E-26	age, CMV, sex, smoking
MFI of ICOS in naive Treg	-	MFI	2	0	1	62	18	35	96	0.214	6.72E-07	age, CMV, sex, smoking
Number of CD4- CD8- MAIT cells	-	Absolute count	3	0	1	1037	1004	48	3840	0.812	8.17E-119	age, CMV, sex, smoking
Number of CD4- CD8- NKT cells	-	Absolute count	3	0	1	69	89	1	341	0.850	6.60E-141	age, CMV, sex
Number of CD8+ MAIT cells	-	Absolute count	3	0	1	1352	1145	89	4315	0.811	1.60E-118	age, CMV, owns a house, sex, smoking

Immunophenotype	Exclusion criterion <sup>a</sup>	Measure type <sup>b</sup>	FACS panel	Innate <sup>c</sup>	Adaptive <sup>c</sup>	Average <sup>d</sup>	SD <sup>d</sup>	2.5th centile <sup>d</sup>	97.5th centile <sup>d</sup>	V1/V2 correlation <sup>e</sup>	Significance	Associated non-genetic covariates <sup>f</sup>
Number of HLA-DR+ CD4- CD8- MAIT cells	-	Absolute count	3	0	1	34	103	0	233	0.416	1.28E-22	
Number of HLA-DR+ CD4- CD8- NKT cells	-	Absolute count	3	0	1	0	1	0	3	0.224	2.02E-07	
Number of MAIT cells	-	Absolute count	3	0	1	2681	2107	222	8860	0.799	2.73E-112	age, CMV, sex, smoking
Number of NKT in T cells	-	Absolute count	3	0	1	92	104	4	424	0.808	5.22E-117	age, sex
Number of $\gamma\delta$ + T cells	-	Absolute count	3	0	1	2601	2505	308	9136	0.779	2.16E-103	age
Number of B cells	-	Absolute count	6	0	1	12435	6827	3701	30585	0.633	9.59E-58	smoking
Number of CD21- CD27- B cells	-	Absolute count	6	0	1	8024	4630	1942	19803	0.705	1.21E-76	age
Number of CD21- CD27+ B cells	-	Absolute count	6	0	1	334	299	73	974	0.717	1.88E-80	smoking
Number of CD24hi memory B cells	-	Absolute count	6	0	1	1172	830	213	3137	0.655	8.41E-63	smoking
Number of CD24low memory B cells	-	Absolute count	6	0	1	341	298	49	1203	0.700	4.68E-75	age, smoking
Number of double negative memory B cells	-	Absolute count	6	0	1	982	772	224	2806	0.614	2.41E-53	age, smoking
Number of founder B cells	-	Absolute count	6	0	1	13	13	0	50	0.513	3.24E-35	age, smoking
Number of germinal center B cells	-	Absolute count	6	0	1	39	29	7	111	0.581	1.04E-46	
Number of IgM- marginal zone B cells	-	Absolute count	6	0	1	877	865	118	2955	0.684	1.35E-70	
Number of IgM+ marginal zone B cells	-	Absolute count	6	0	1	594	508	78	2097	0.663	7.46E-65	
Number of marginal zone B cells	-	Absolute count	6	0	1	1474	1232	233	4549	0.736	1.51E-86	
Number of memory B cells	-	Absolute count	6	0	1	2116	1443	462	5411	0.687	2.92E-71	smoking
Number of naive B cells	-	Absolute count	6	0	1	6379	3660	1521	15059	0.705	1.46E-76	age
Number of transitional B cells	-	Absolute count	6	0	1	619	482	86	1956	0.790	3.86E-108	age
MFI of CD19 in B cells	-	MFI	6	0	1	5320	605	4241	6550	0.607	5.40E-52	had MMR vaccine
MFI of CD21 in B cells	-	MFI	6	0	1	5242	797	3724	6730	0.734	4.96E-86	age
MFI of CD21 in CD24hi memory B cells	-	MFI	6	0	1	5413	823	3935	6997	0.745	9.14E-90	smoking
MFI of CD21 in CD24int memory B cells	-	MFI	6	0	1	4990	690	3600	6303	0.722	6.40E-82	smoking
MFI of CD21 in CD24low memory B cells	-	MFI	6	0	1	3969	636	2785	5274	0.642	1.08E-59	smoking
MFI of CD21 in double negative memory B cells	-	MFI	6	0	1	3688	660	2386	5029	0.675	0.00E+00	age
MFI of CD21 in germinal center B cells	-	MFI	6	0	1	2133	817	836	4203	0.295	8.39E-12	smoking
MFI of CD21 in IgM- marginal zone B cells	-	MFI	6	0	1	5926	935	4006	7661	0.694	1.93E-73	age, smoking
MFI of CD21 in IgM+ marginal zone B cells	-	MFI	6	0	1	5930	1035	3874	8010	0.730	1.55E-84	age, smoking
MFI of CD21 in marginal zone B cells	-	MFI	6	0	1	5967	996	4060	7784	0.754	2.33E-93	age, smoking
MFI of CD21 in memory B cells	-	MFI	6	0	1	5092	747	3633	6549	0.770	1.27E-99	smoking
MFI of CD21 in naive B cells	-	MFI	6	0	1	5385	915	3526	7224	0.753	6.42E-93	age
MFI of CD21 in transitional B cells	-	MFI	6	0	1	5025	795	3598	6784	0.667	5.59E-66	age
MFI of CD24 in B cells	-	MFI	6	0	1	1136	300	669	1876	0.656	2.87E-63	age
MFI of CD24 in double negative memory B cells	-	MFI	6	0	1	931	249	517	1478	0.648	2.87E-61	smoking
MFI of CD24 in germinal center B cells	-	MFI	6	0	1	162	72	39	332	0.261	1.55E-09	age
MFI of CD24 in IgM- marginal zone B cells	-	MFI	6	0	1	1597	413	859	2587	0.650	1.04E-61	age
MFI of CD24 in IgM+ marginal zone B cells	-	MFI	6	0	1	2196	580	1229	3520	0.644	3.95E-60	age
MFI of CD24 in marginal zone B cells	-	MFI	6	0	1	2028	560	1123	3356	0.515	1.40E-35	age
MFI of CD24 in memory B cells	-	MFI	6	0	1	1537	396	855	2503	0.656	3.72E-63	age
MFI of CD24 in naive B cells	-	MFI	6	0	1	771	182	472	1215	0.663	7.88E-65	age
MFI of CD27 in B cells	-	MFI	6	0	1	1287	514	462	2458	0.779	0.00E+00	
MFI of CD38 in B cells	-	MFI	6	0	1	4632	2350	1634	10632	0.623	2.18E-55	
MFI of CD38 in double negative memory B cells	-	MFI	6	0	1	6903	5234	161	20892	0.511	0.00E+00	
MFI of CD38 in germinal center B cells	-	MFI	6	0	1	19977	1751	16514	23195	0.250	8.30E-09	sex, smoking
MFI of CD38 in memory B cells	-	MFI	6	0	1	1156	278	602	1707	0.838	2.41E-133	
MFI of IgD in B cells	-	MFI	6	0	1	2346	649	1144	3816	0.714	1.74E-79	age
MFI of IgM in B cells	-	MFI	6	0	1	520	210	225	1058	0.714	2.36E-79	age
Number of CCR6+ CD4+ T cells	-	Absolute count	9	0	1	8358	4201	2548	19105	0.530	8.41E-38	age, sex, smoking
Number of CCR6+ CD8+ T cells	-	Absolute count	9	0	1	1540	1141	222	4463	0.750	1.35E-91	
Number of CRTh2+ CD4+ T cells	-	Absolute count	9	0	1	513	326	89	1389	0.610	1.48E-52	
Number of CXCR5+ CD4+ T cells	-	Absolute count	9	0	1	2787	1408	744	6209	0.506	3.79E-34	sex
Number of CD45+ cells	-	Absolute count	5	1	1	312995	1E+05	139925	580025	0.428	6.06E-24	age, smoking
Total number of cells	-	Absolute count	5	1	1	333076	1E+05	145000	662325	0.408	9.50E-22	age, smoking

Immunophenotype	Measure type <sup>a</sup>	FACS panel	Exclusion criterion	Measure type <sup>b</sup>	FACS panel	Immunophenotype	Exclusion criterion	Measure type	FACS panel
Number of CD4+ CD8+ T cells	Absolute count	1	Control	Proportion	1	Proportion of MAIT in CD4+ T cells	Proportion	Proportion	3
Number of CD4+ T cells	Absolute count	2	Control	Proportion	2	Proportion of MAIT in CD8+ T cells	Proportion	Proportion	3
Number of CD8+ T cells	Absolute count	2	Control	Proportion	2	Proportion of NKT in CD4+ CD8+ T cells	Proportion	Proportion	3
Number of CD4+ T cell ratio	Ratio	2	Control	Proportion	2	Proportion of NKT in CD8+ T cells	Proportion	Proportion	3
Number of CD8+ T cell ratio	Ratio	2	Control	Proportion	2	Proportion of CD56hi in NK cells	Proportion	Proportion	4
Number of CD8+ CD4+ T cells	Absolute count	3	Control	Proportion	3	Proportion of CD69+ in CD56hi NK cells	Proportion	Proportion	4
Number of CD8+ T cells	Absolute count	3	Control	Proportion	3	Proportion of CD16hi NK cells	Proportion	Proportion	4
Total number of cells	Absolute count	3	Control	Proportion	3	Proportion of HLA-DR+ in CD16hi NK cells	Proportion	Proportion	4
CD4+ to CD8+ T cell ratio	Ratio	3	Control	Proportion	3	Proportion of B cells in CD45+ cells	Proportion	Proportion	5
Number of B cells	Absolute count	5	Control	Proportion	5	Proportion of CD14hi monocytes in CD45+ cells	Proportion	Proportion	5
Number of CD16hi NK cells	Absolute count	5	Control	Proportion	5	Proportion of CD14hi monocytes in CD45+ cells	Proportion	Proportion	5
Number of CD56hi NK cells	Absolute count	5	Control	Proportion	5	Proportion of CD8+ CD4+ T cells	Proportion	Proportion	5
Number of CD8+ CD4+ T cells	Absolute count	5	Control	Proportion	5	Proportion of CD8+ CD4+ T cells	Proportion	Proportion	5
Number of CD4+ T cells	Absolute count	5	Control	Proportion	5	Proportion of NK cells in CD45+ cells	Proportion	Proportion	5
Number of CD8+ T cells	Absolute count	5	Control	Proportion	5	Proportion of CD21+ CD27+ in B cells	Proportion	Proportion	6
Number of CD8+ T cell ratio	Ratio	5	Control	Proportion	5	Proportion of CD24hi in memory B cells	Proportion	Proportion	6
Number of CD4+ T cells	Absolute count	9	Control	Proportion	9	Proportion of Founder in B cells	Proportion	Proportion	6
Number of CD8+ T cells	Absolute count	9	Control	Proportion	9	Proportion of IgA+ CD27+ B cells	Proportion	Proportion	6
Number of CD4+ T cell ratio	Ratio	9	Control	Proportion	9	Proportion of IgA+ CD27+ in B cells	Proportion	Proportion	6
Number of CD8+ T cell ratio	Ratio	9	Control	Proportion	9	Proportion of IgM+ in double negative memory B cells	Proportion	Proportion	6
Number of CD8+ T cells	Absolute count	10	Control	Proportion	10	Proportion of IgM+ in marginal zone B cells	Proportion	Proportion	6
Number of CD4+ T cells	Absolute count	10	Control	Proportion	10	Proportion of IgM+ in marginal zone B cells	Proportion	Proportion	6
Number of CD8+ T cells	Absolute count	10	Control	Proportion	10	Proportion of memory B cells in IgD+ CD27+ B cells	Proportion	Proportion	6
Number of CD4+ T cell ratio	Ratio	10	Control	Proportion	10	Proportion of naive in B cells	Proportion	Proportion	6
Number of CD8+ T cell ratio	Ratio	10	Control	Proportion	10	Proportion of Plasmacytes in B cells	Proportion	Proportion	6
Number of CD4+ T cells	Absolute count	1	Control	Proportion	1	Proportion of Plasmacytes in IgD+ CD27+ B cells	Proportion	Proportion	6
Number of CD8+ T cells	Absolute count	1	Control	Proportion	1	Proportion of transitional in IgD+ CD27+ B cells	Proportion	Proportion	6
Number of CD4+ T cells	Absolute count	1	Control	Proportion	1	Proportion of CD32+ in basophils	Proportion	Proportion	7
Number of CD8+ T cells	Absolute count	1	Control	Proportion	1	Proportion of CD32+ in eosinophils	Proportion	Proportion	7
Number of CD4+ T cells	Absolute count	1	Control	Proportion	1	Proportion of CD32+ in neutrophils	Proportion	Proportion	7
Number of CD8+ T cells	Absolute count	1	Control	Proportion	1	Proportion of CD32+ in neutrophils	Proportion	Proportion	7
Number of CD4+ T cell ratio	Ratio	1	Control	Proportion	1	Proportion of cDC1 in HLA-DR+ CD14+ cells	Proportion	Proportion	8
Number of CD8+ T cell ratio	Ratio	1	Control	Proportion	1	Proportion of cDC2 in HLA-DR+ CD14+ cells	Proportion	Proportion	8
Number of CD4+ T cells	Absolute count	2	Control	Proportion	2	Proportion of pDC in HLA-DR+ CD14+ cells	Proportion	Proportion	8
Number of CD8+ T cells	Absolute count	2	Control	Proportion	2	Proportion of CCR6+ in CD4+ T cells	Proportion	Proportion	9
Number of CD4+ T cell ratio	Ratio	2	Control	Proportion	2	Proportion of CCR6+ in CD8+ T cells	Proportion	Proportion	9
Number of CD8+ T cell ratio	Ratio	2	Control	Proportion	2	Proportion of CXCR5+ in CD4+ T cells	Proportion	Proportion	9
Number of CD4+ T cells	Absolute count	3	Control	Proportion	3	Proportion of CD161+ ILCS in LLC	Proportion	Proportion	10
Number of CD8+ T cells	Absolute count	3	Control	Proportion	3	Proportion of CD56+ cells in LLC	Proportion	Proportion	10
Number of CD4+ T cell ratio	Ratio	3	Control	Proportion	3	Proportion of ILCS in LLC	Proportion	Proportion	10
Number of CD8+ T cell ratio	Ratio	3	Control	Proportion	3	Proportion of ILCS in LLC	Proportion	Proportion	10
Number of CD4+ T cells	Absolute count	3	Control	Proportion	3	Proportion of HLA-DR+ NKT in CD4+ cells	Proportion	Proportion	10
Number of CD8+ T cells	Absolute count	3	Control	Proportion	3	Proportion of HLA-DR+ NKT in CD8+ cells	Proportion	Proportion	10
Number of CD4+ T cell ratio	Ratio	3	Control	Proportion	3	Proportion of HLA-DR+ NKT in CD4+ cells	Proportion	Proportion	10
Number of CD8+ T cell ratio	Ratio	3	Control	Proportion	3	Proportion of HLA-DR+ NKT in CD8+ cells	Proportion	Proportion	10

**Table 9 : List of the immunophenotypes measured in this study.**

<sup>a</sup> A total of 313 measured immunophenotypes were measured. We excluded from subsequent analyses all cell proportions ("Proportion", n=110) and immunophenotypes that were measured several times on different panels ("Control", n=35).

<sup>b</sup> Absolute count corresponds to the number of cell events measured in a given gate of flow cytometry. MFI corresponds to the mean fluorescence intensity of a given marker at the cell-surface of a gated population of cells.

<sup>c</sup> Innate cells were defined as those populations lacking somatic recombination of the genome. Adaptive cells, defined by their dependence on RAG1/2 activity.

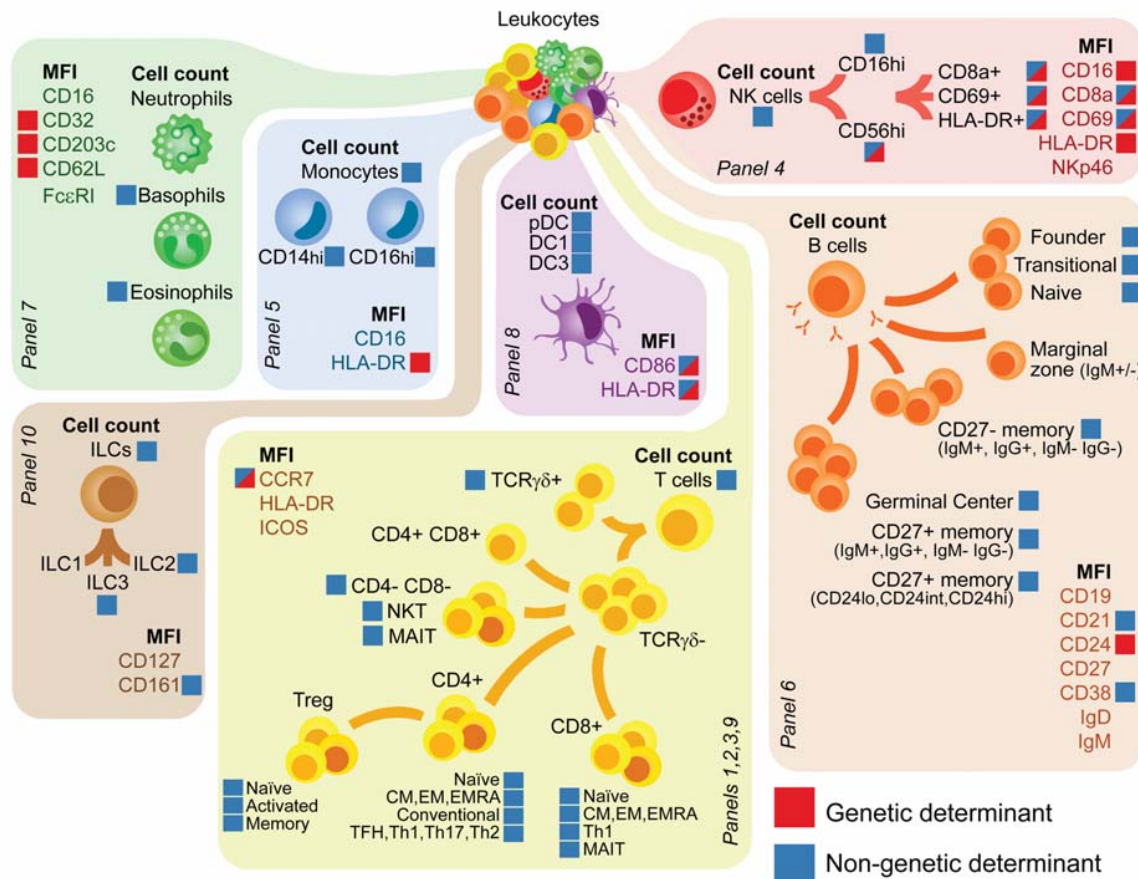
<sup>d</sup> Average, standard deviation (SD) and centiles were estimated on immunophenotypes after outlier removal, imputation and batch correction.

<sup>e</sup> Spearman correlation coefficient between immunophenotypes measured at visit 1 and visit 2 (~17 days after visit 1), on the same 500 individuals who went to both visits.

<sup>f</sup> Non-genetic factors associated with immunophenotypes, identified by multiple regression (Figure 40A).

Remark : To facilitate reading, the original table has been split into 6 parts. This last section on the left contains two last parts with the entire list of immunophenotypes used as exclusion criterion (the two continuous lists have been aligned and separated by a grey column)

These include 77 and 91 immunophenotypes obtained in innate and adaptive immune cells, respectively. Innate cells were defined as those lacking somatic recombination of the genome (Vivier et al., 2011), and included granulocytes (neutrophils, basophils and eosinophils), monocytes, natural killer (NK) cells, dendritic cells and innate lymphoid cells (ILCs) (**Figure 29**). Adaptive cells were defined by their dependence on RAG1/2 activity and included T cells ( $\gamma\delta$  T, MAIT, NKT, Treg and Th cells) and B cells. The immunophenotypes in both innate and adaptive immune cells included 76 absolute counts of circulating cell numbers, 89 expression levels of cell-surface protein markers (quantified by the mean fluorescence intensity, or MFI), and 3 ratios of cell counts or MFI (**Figures 29 and 30 and Table 9**).



**Figure 29 : Immune cells and cell-surface markers measured in the *MI* cohort**

Panel numbers refer to the cytometric analyses performed, grouped based on cellular lineage (**Figures 56-65 and Tables 8 and 9**). The expression of phenotypic markers of differentiation or activation was quantified based on their mean fluorescent intensity (MFI), indicated per panel. Interconnecting lines illustrate cellular lineages or differentiation states. Red and blue squares indicate immunophenotypes significantly associated in this study with genetic or non-genetic factors, respectively

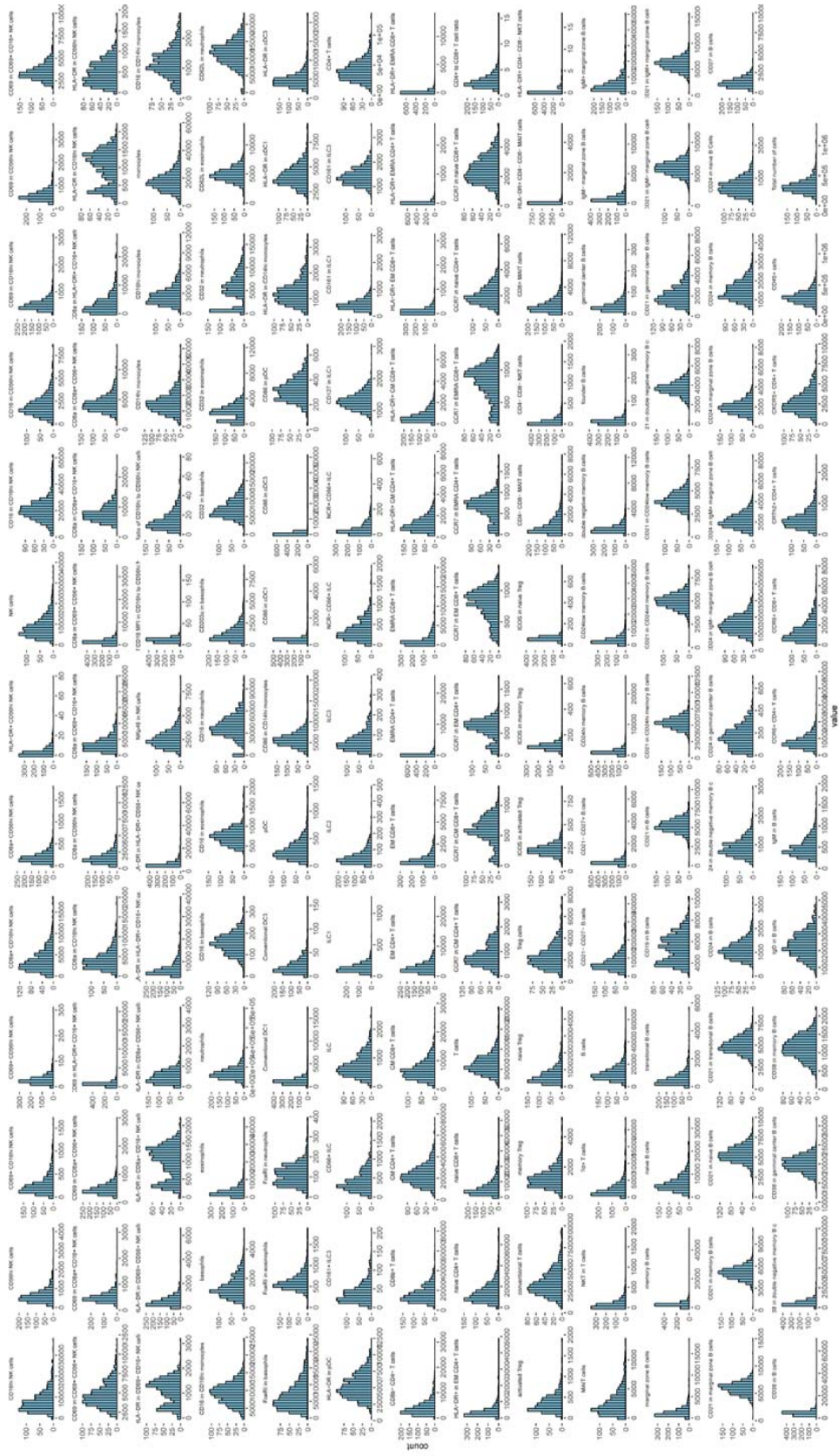
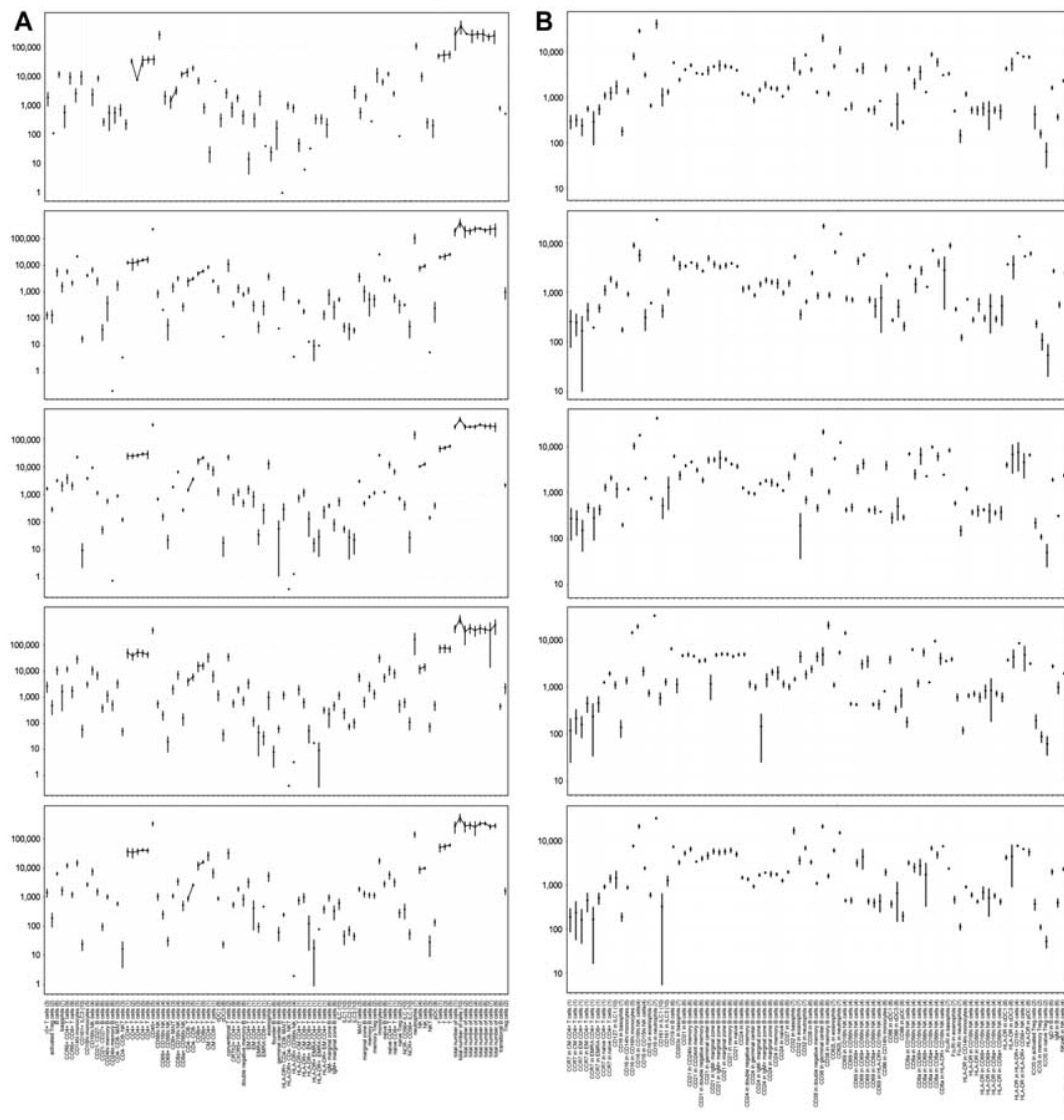


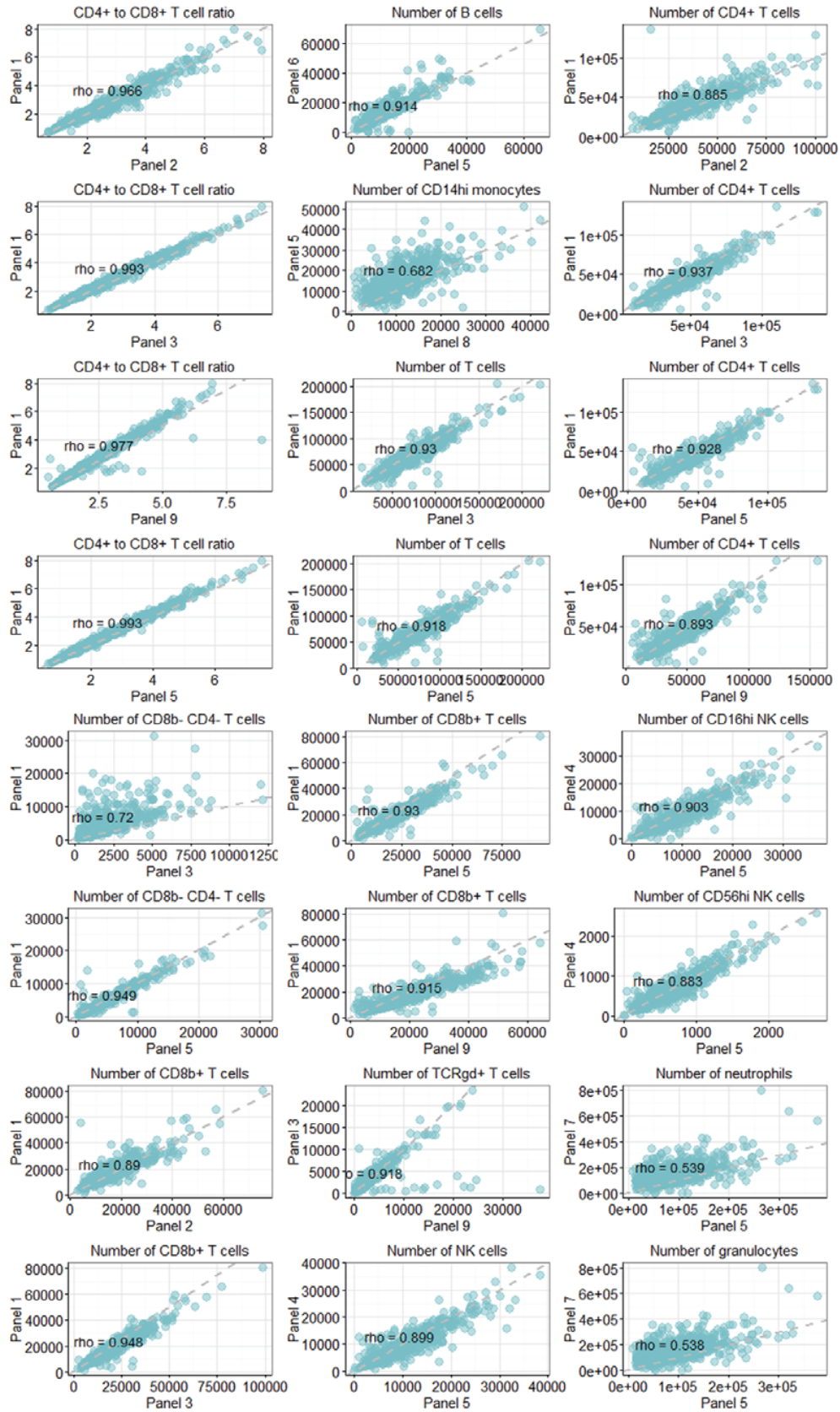
Figure 40 : Distributions of 169 raw immunophenotypes measured in the M1 cohort

To reduce technical variation introduced by sample temperature fluctuations and pre-analytical procedures, we strictly followed a standardized protocol for tracking and processing samples (Hasan et al., 2015). We verified that measured immunophenotypes were highly reproducible using technical replicates (**Figures 31 and 32** and **Table 9**), demonstrating the high accuracy of the data.



**Figure 31 : Repeatability and reproducibility of semi-automated flow cytometry measurements**  
 Repeatability of (A) absolute cell counts and (B) MFI estimated based on measurements performed at five different time points over five months on the same five individuals. Points represent the average of measurements across replicates of the same individual, and bars 3 times their standard deviation.



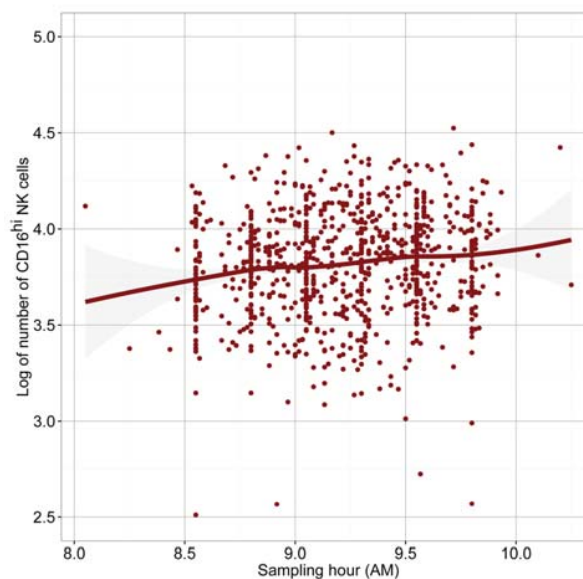


**Figure 32 : Reproducibility of semi-automated flow cytometry measurements.**

Comparison of cell counts measured in different flow cytometry panels across the 1,000 *Milieu Intérieur* subjects. Spearman's  $\rho$  correlation coefficients are shown.

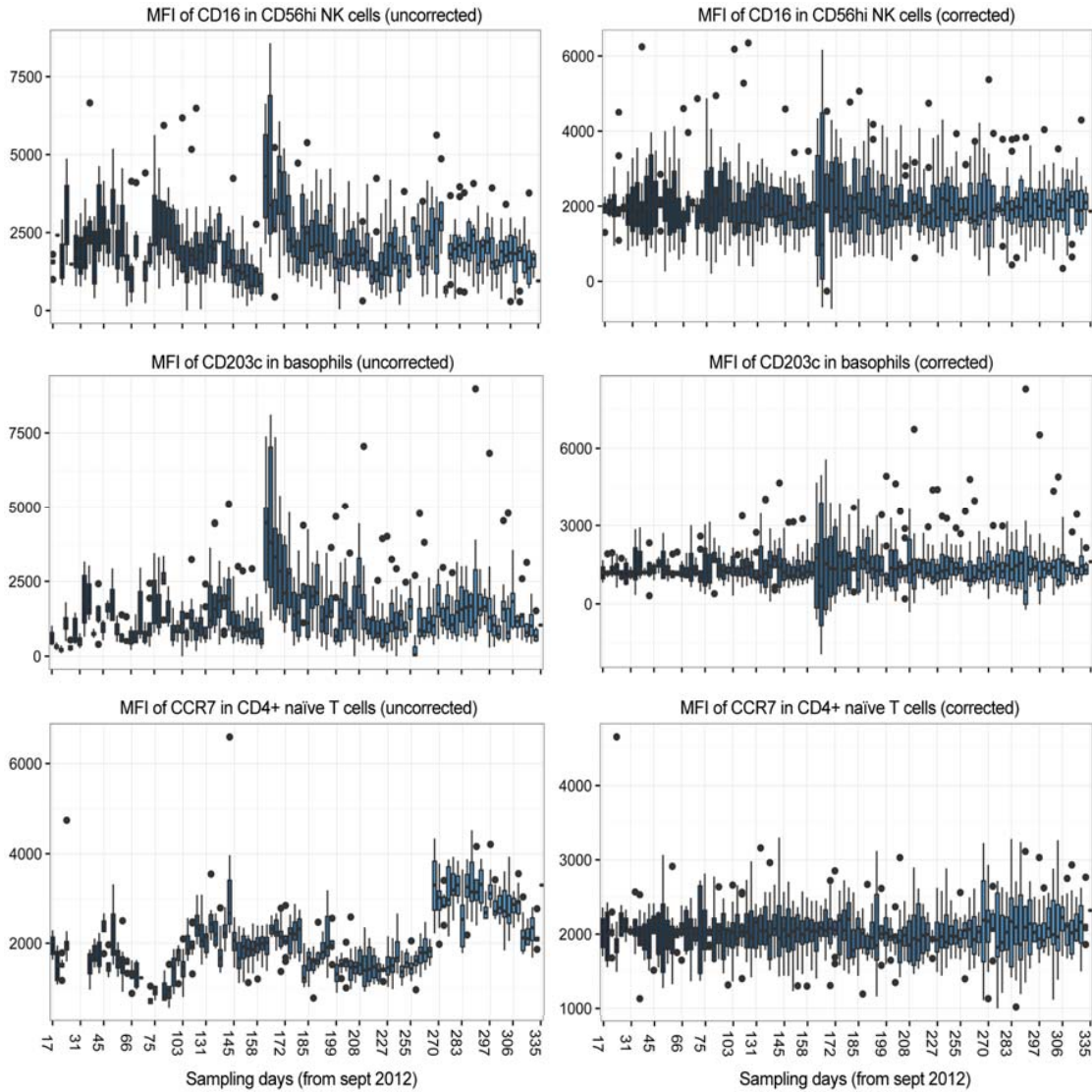
We nevertheless identified two technical batch effects that impacted flow cytometric analyses. One effect corresponded to the hour at which the blood sample was drawn from fasting subjects (**Figure 33**), which may possibly be explained by the spike in cortisol at the time of waking (Patterson et al., 2013). The second batch effect corresponded to variation of immunophenotypes over the one-year sampling period, which primary affected MFI measures (**Figure 34**). We corrected for these batch effects in all subsequent analyses (**Figure 35**; Materials & Methods). We provide the distribution, ranges and statistics of all batch-corrected immune cell counts (**Table 9**), thereby facilitating comparisons with cytometry data collected as part of routine clinical practice. Values can be accessed through a user-friendly web application

([http://104.236.137.56:3838/LabExMICytometryBrowser\\_ShinyApp/](http://104.236.137.56:3838/LabExMICytometryBrowser_ShinyApp/), draft Shiny application available for review), which can be queried based on personal characteristics such as age or gender.



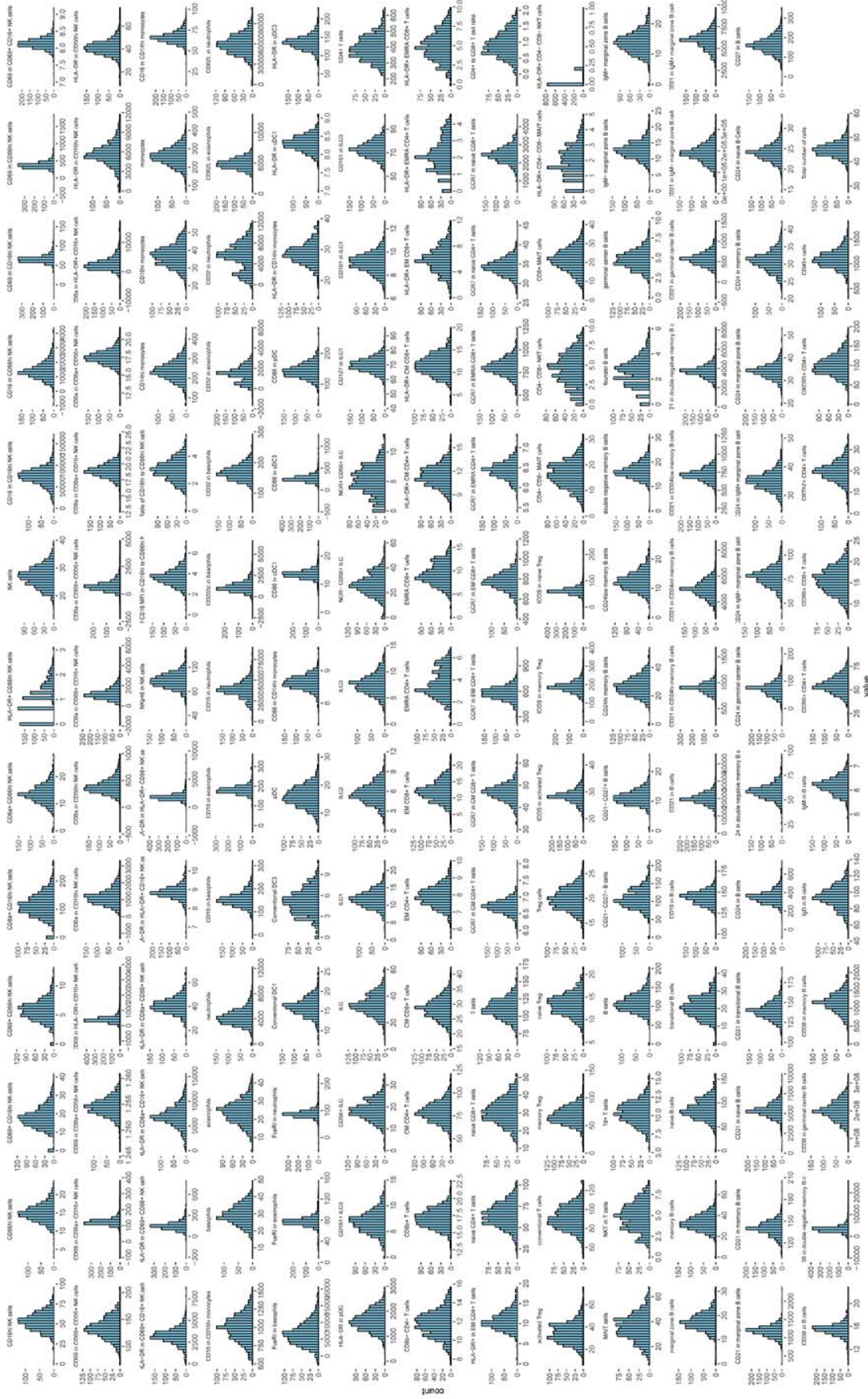
**Figure 33 : Sample hour effect on measured immunophenotypes**

Effect of the sampling hour on the absolute number of CD16<sup>hi</sup> NK cells. Whole blood samples were collected from the 1,000 *Milieu Intérieur* healthy subjects every working day from 8 to 11AM. The sampling hour has a significant effect of the absolute number of CD16<sup>hi</sup> NK cells, as well as other immune cell counts, which were adjusted for this batch effect.



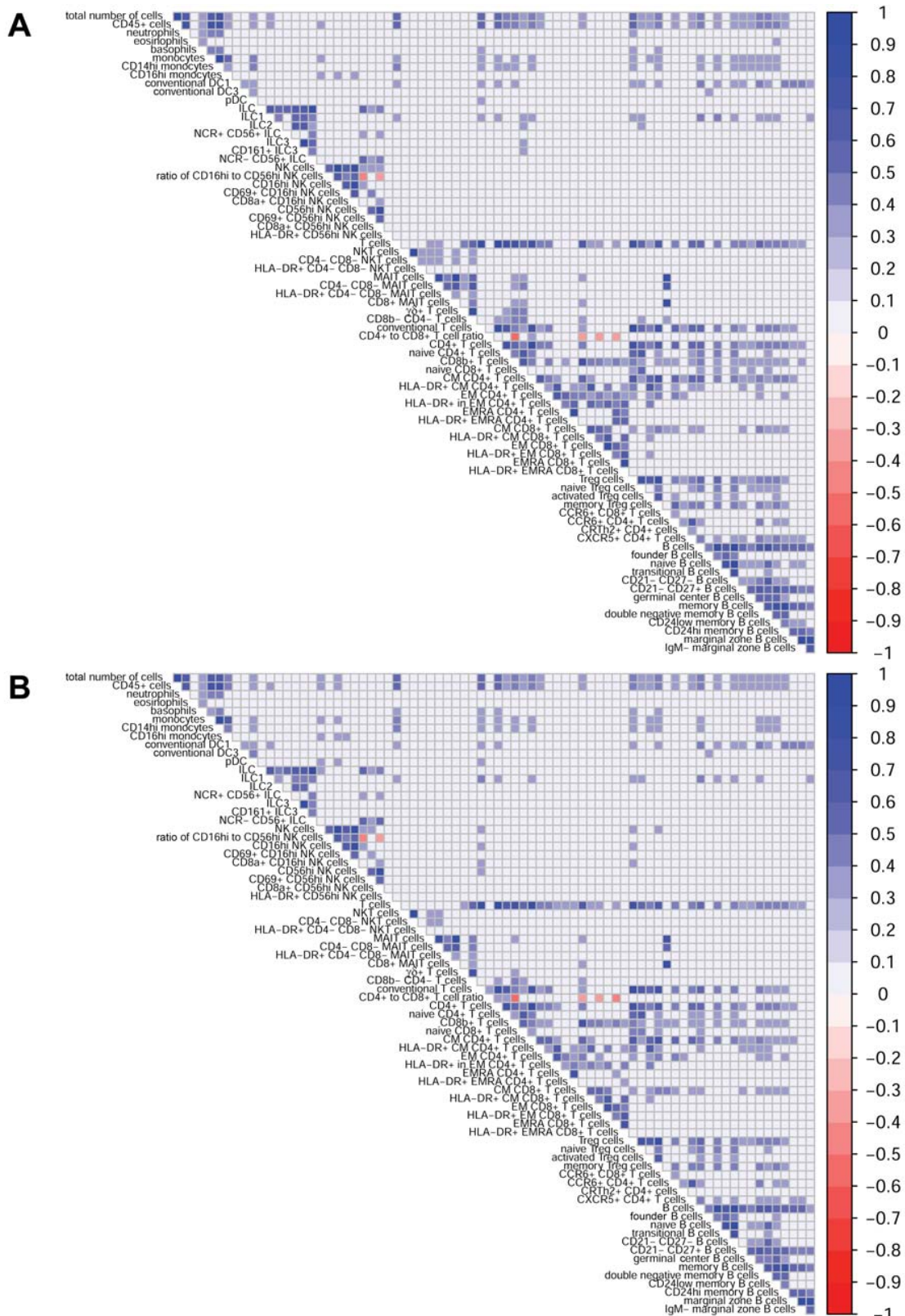
**Figure 34 : Effect of sampling over time on measured immunophenotypes**

Effect of the sampling day on immunophenotypes measured in the *Milieu Intérieur* cohort. Whole blood samples were collected from the 1,000 *Milieu Intérieur* healthy subjects every working day from September 2012 to August 2013. Variation in the mean fluorescence intensity (MFI) of several protein markers was observed along the sampling period. All MFIs were corrected for this batch effect using the ComBat non-parametric Bayesian framework (Johnson et al., 2007). Three examples of uncorrected (left column) and corrected (right columns) values are represented.



**Figure 35 : Transformed distributions of the 168 immunophenotypes studied in the MI cohort.** Distributions are shown after outlier removal, imputation, batch correction and Yeo-Johnson or Box-Cox transformations. Raw distributions are shown in Figure 30

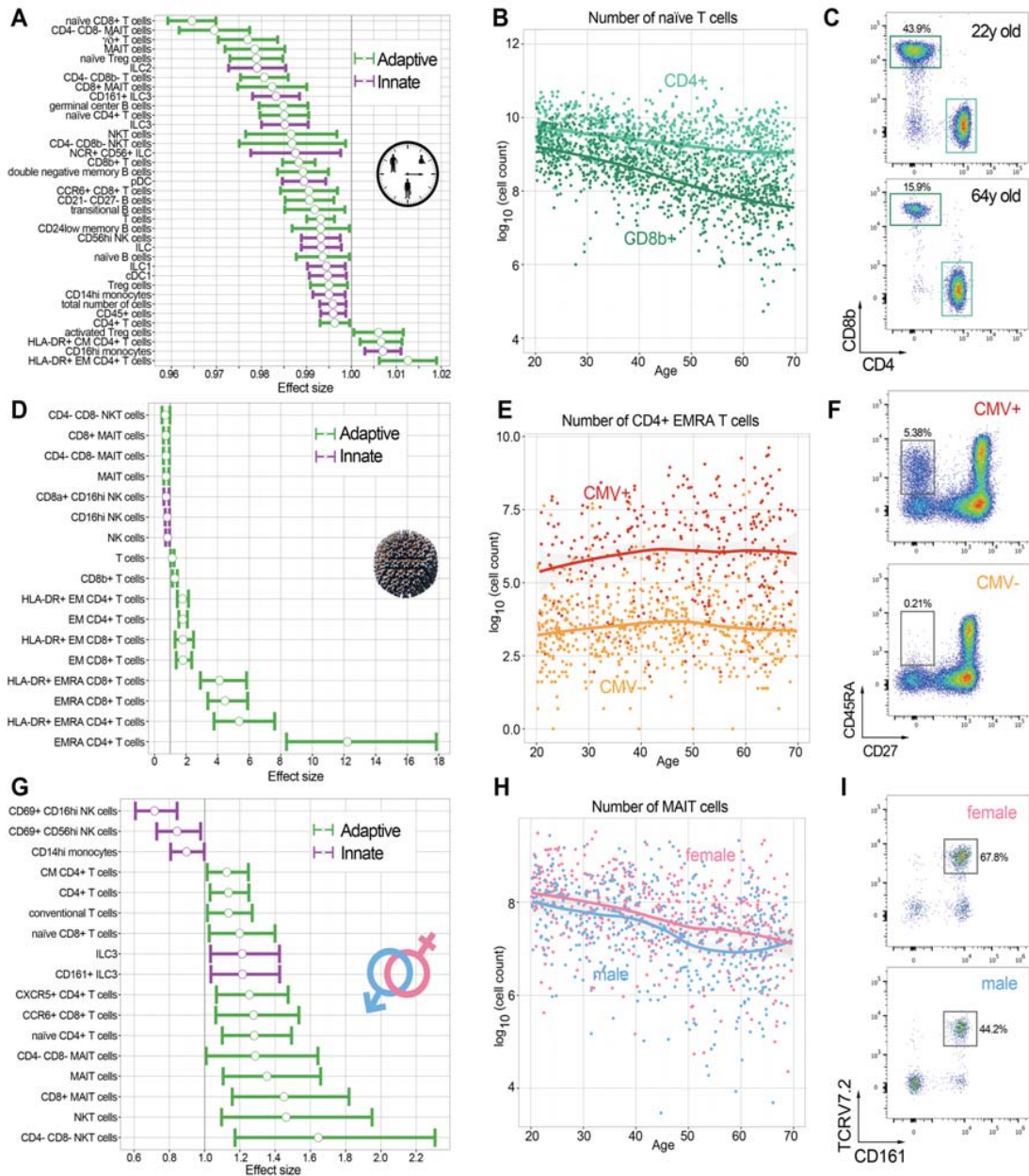
Owing to the hierarchical structure of immune cell differentiation (*i.e.*, cellular lineages emerge from common progenitor cells), a substantial portion of the immune cell counts measured in this study are highly correlated (**Figure 36**). These correlations were not directly attributable to the influence of non-genetic factors such as age or gender, which were regressed out in this analysis. Interestingly, we observed correlations between circulating levels of ILC and NK populations, reflecting their common developmental pathway and dependence on  $\gamma_c$  cytokines (Serafini et al., 2015). We also observed a correlation between MAIT cells and CCR6<sup>+</sup> CD8<sup>+</sup> T cells, a result of the former being the major subset of CCR6<sup>+</sup> T cells in circulation (Dusseaux et al., 2011). An appreciable correlation was detected among T and B lymphocytes, highlighting their origin from a common lymphoid progenitor and an interdependence in response to antigenic challenge. Finally, we identified a strong correlation between the number of Treg and conventional CD4<sup>+</sup> T cells, validating previous experimental work that defined an interleukin-2 (IL-2) driven self-regulatory circuit that integrates the homeostasis of these cell populations (Amado et al., 2013).



**Figure 36 : Correlations among circulating levels of immune cell populations in the *MI* cohort.** Correlations were estimated between (A) non-regressed immunophenotypes or (B) immunophenotypes regressed on non-genetic covariates, which were selected using a stability selection algorithm among 40 demographic variables (Table 7; Materials & Methods). Correlations lower than 0.3 and larger than -0.3 were set to 0

## 2.2. Impact of age, gender and cytomegalovirus infection on immune cell parameters

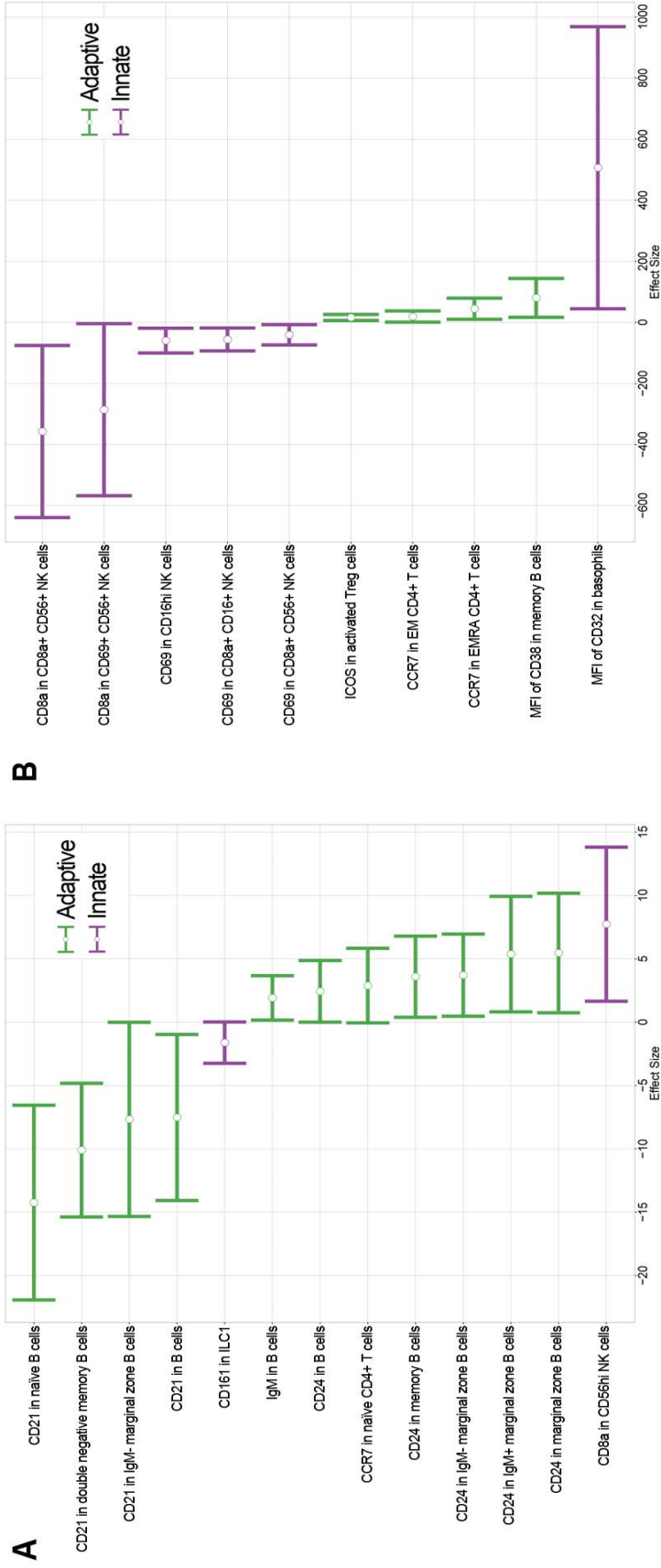
Prior studies have suggested that environmental exposures are responsible for inter-individual differences in white blood cell composition, in particular for T cell subpopulations, which are impacted by factors such as age, gender and CMV seropositivity (Tollerud et al., 1989; Sridharan et al., 2011; Pennell et al., 2012; Griffiths et al., 2013). We used multiple linear regression to quantify the impact of each of these non-genetic factors on human variation in white blood cell composition. We observed a substantial effect of age on both innate and adaptive immune cells, affecting 30% of immunophenotypes (**Figures 37A** and **38**). We detected a general decline in the number of innate lymphoid and dendritic cells as a function of increasing age (**Figure 37A**). Contrary to previous studies, we find a modest increase in the number of memory T cells in elderly subjects, which may stem from the reliance of these studies on frequency analyses, as compared to our implementation of absolute cell count measurements. Our analyses also show that naïve CD8<sup>+</sup> T cells decrease more than twice as rapidly with age as compared to naïve CD4<sup>+</sup> T cells (3.5% and 1.5% per year, respectively; **Figure 37A-C**), which supports the view that CD8<sup>+</sup> T cells are more susceptible to concentrations of homeostatic cytokines and/or that the production of CD4<sup>+</sup> T cells is preferentially enhanced in the human thymus (Vrisekoop et al., 2008).



**Figure 37 : Age, sex and CMV infection impact immune cell counts in healthy individuals.**

Significant effects of (A-C) increasing age (D-F) CMV seropositivity and (G-I) female sex on circulating levels of immune cells in 1,000 healthy individuals. (A, D, G) Effect sizes were estimated in a multiple regression that includes the three non-genetic factors, together with batch effects and genome-wide significant SNPs. The confidence intervals are false coverage-adjusted and calculated using sandwich estimated standard errors. Adaptive and innate immune cells are represented in green and purple, respectively. (B, E, H) Regression lines were fitted using local multinomial regression. (B) Impact of age on naïve CD8b<sup>+</sup> (in dark green) and CD4<sup>+</sup> (in light green) T cells. (E) Impact of age and CMV serostatus on CD4<sup>+</sup> EMRA T cells. CMV+ individuals are represented in red and CMV- in orange. (H) Impact of age and sex on the absolute count of MAIT cells. Females are represented in pink and men in blue. (C) Flow cytometry plots of naïve CD8b<sup>+</sup> and CD4<sup>+</sup> T cells for representative persons in their 20s and their 60s. (F) Flow cytometry plots of EMRA CD4<sup>+</sup> T cells in representative CMV- and CMV+ subjects. (I) Flow cytometry plots of MAIT cells in representative woman and man.





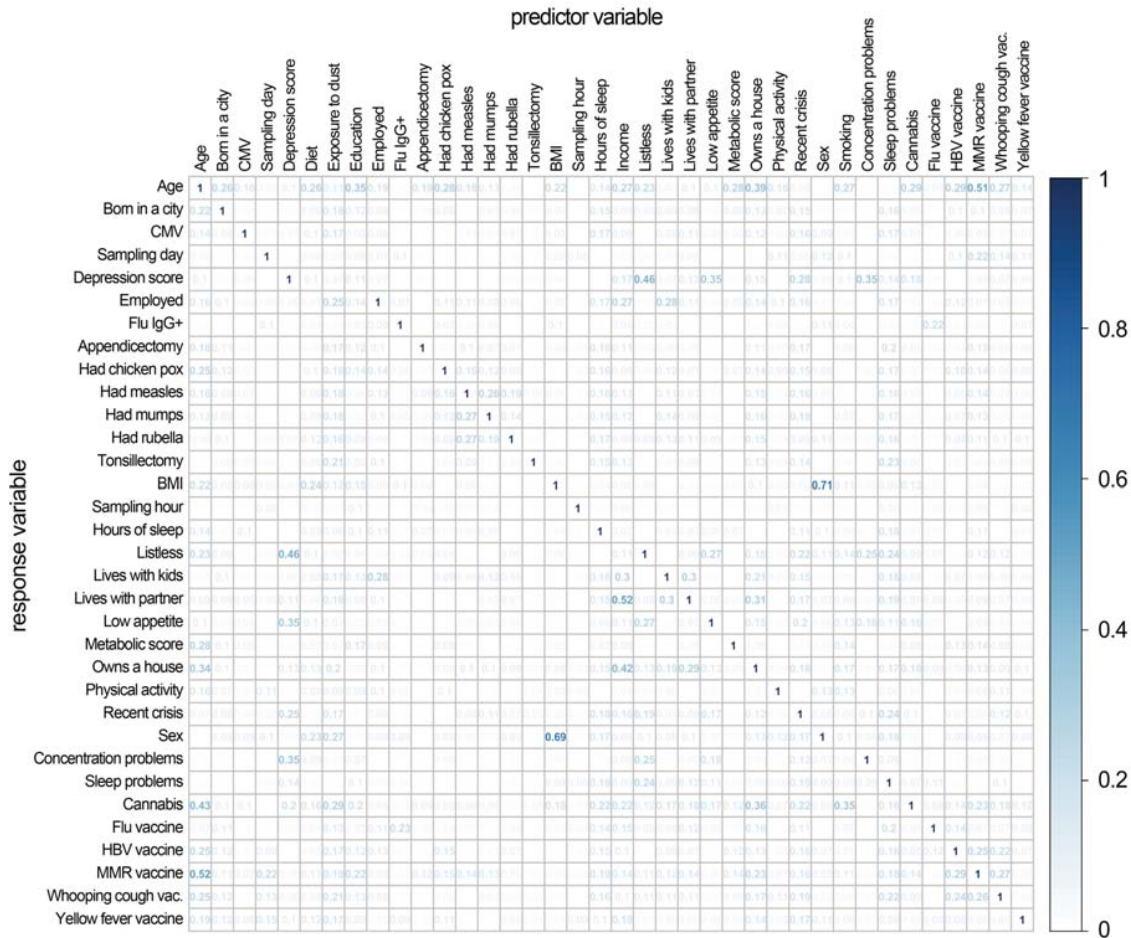
**Figure 38 : Age and sex impact protein levels of immune cell markers in healthy individuals.**

Significant multiplicative effects of (A) increasing age and (B) female sex on the mean fluorescence intensity (MFI) of protein markers of immune cells of 1,000 healthy individuals, controlling for the other factor, batch effects and genomewide significant SNPs (Table 10). CMV serostatus was not included because it has no significant effect on measured MFIs (data not shown).

We observed a profound effect of latent CMV infection on several immune cell parameters (**Figures 37D and 38**). We confirm that CMV triggers a major change in the number of memory T cells, which is independent from age effects (Griffiths et al., 2013). In particular, CMV seropositivity associated with a 12- and 4.5-times higher number of CD4<sup>+</sup> and CD8<sup>+</sup> T<sub>EM</sub> and T<sub>EMRA</sub> cells (**Figure 37D-F**). However, we did not observe evidence that CMV infection impacts the number of naïve T cells or the T<sub>CM</sub> compartments. Supporting this observation, the total number of CD8<sup>+</sup> and CD4<sup>+</sup> T cells increased in parallel with the expanded number of memory T cells, thus suggesting independent regulation of the naïve and EM/EMRA T cell pools. Finally, CMV seropositive donors presented lower numbers of circulating NKT and MAIT cells (**Figure 37D**), suggesting that latent CMV infection may result in competition for cytokine growth factors with the expansion of conventional effector memory T cells and the subsequent reduced numbers of ILC and invariant T cell populations. Although sex differences have been previously reported for various immune responses and diseases (Whitacre, 2001; Pennell et al., 2012), previous studies examining circulating cellular parameters have reported inconsistent results, owing to differences between flow cytometry procedures and small, underpowered or poorly-stratified study cohorts. We report a significant impact of sex on 16% of measured immunophenotypes (**Figures 37G and 38**). Most notably, we find a higher number of CD14<sup>hi</sup> monocytes and activated NK cells in men, as compared to women. By contrast, MAIT cells were systematically increased in women across all age decades (**Figure 37H-I**). These results suggest a lasting effect of early hormonal differences on immune cell development and biology.

### 2.3. Tobacco smoking alters extensively immune cell numbers

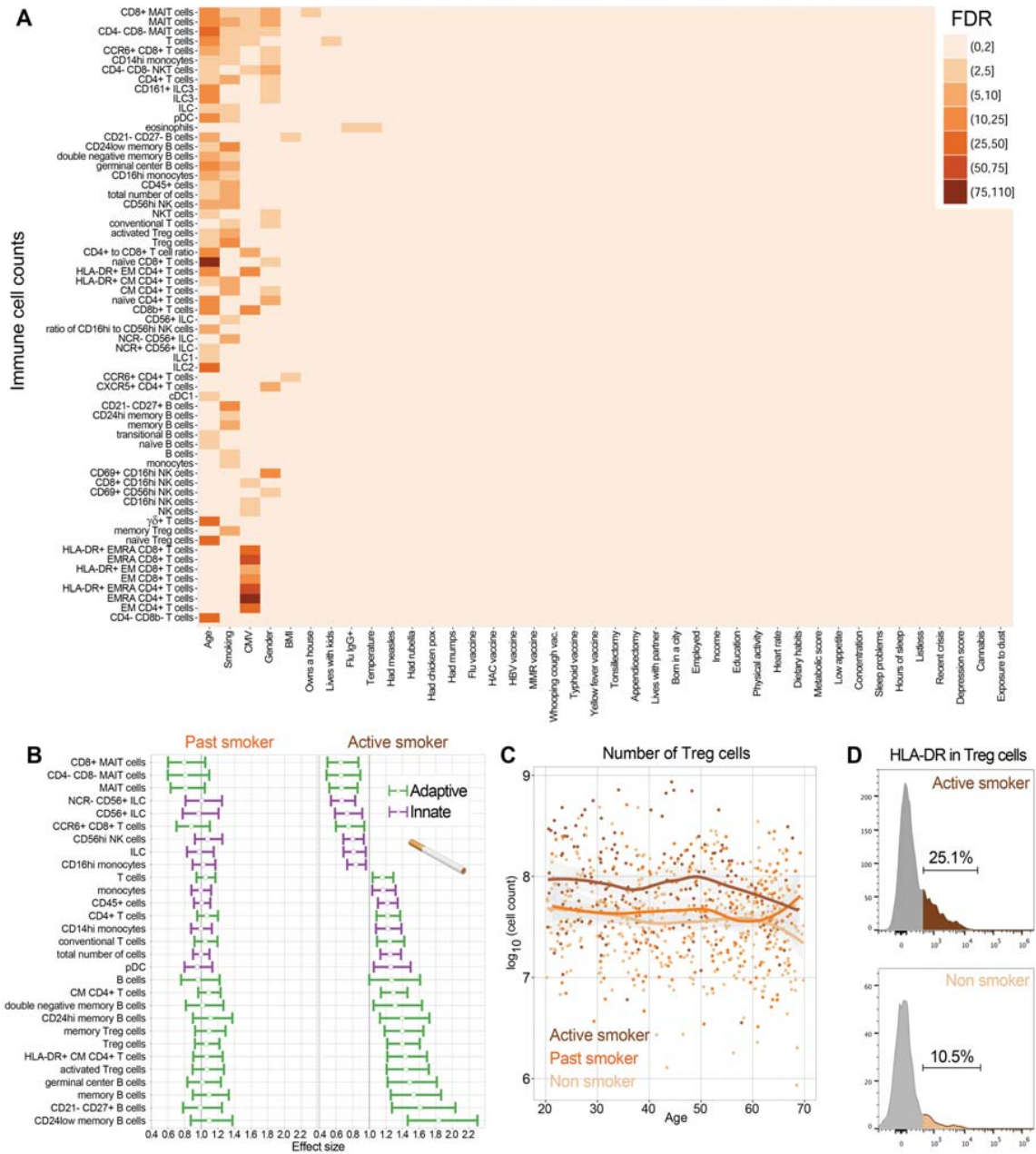
Capitalizing on the lifestyle and demographic data available for the *Milieu Intérieur* cohort, we evaluated the influence of additional non-genetic factors on immune cell parameters, controlling for the defined effects of age, gender and CMV serological status. A total of 40 variables were chosen for analysis and tested for each immunophenotype. These include socio-economic characteristics, dietary habits, past infections, and surgery and vaccination history (**Figure 39 and Table 7**).



**Figure 39 : Correlation matrix among 40 demographic variables analyzed in this study**

Details on the demographic variables can be found in [Table 7](#). The estimation of correlation coefficients was based on the generalized  $R^2$  measures for pairwise fitted generalized linear models (Methods). Variables listed in rows are response variables and those listed in columns are predictor variables in the models. Categorical variables (i.e., diet, exposure to dust, level of education, income and smoking) were used only as predictors

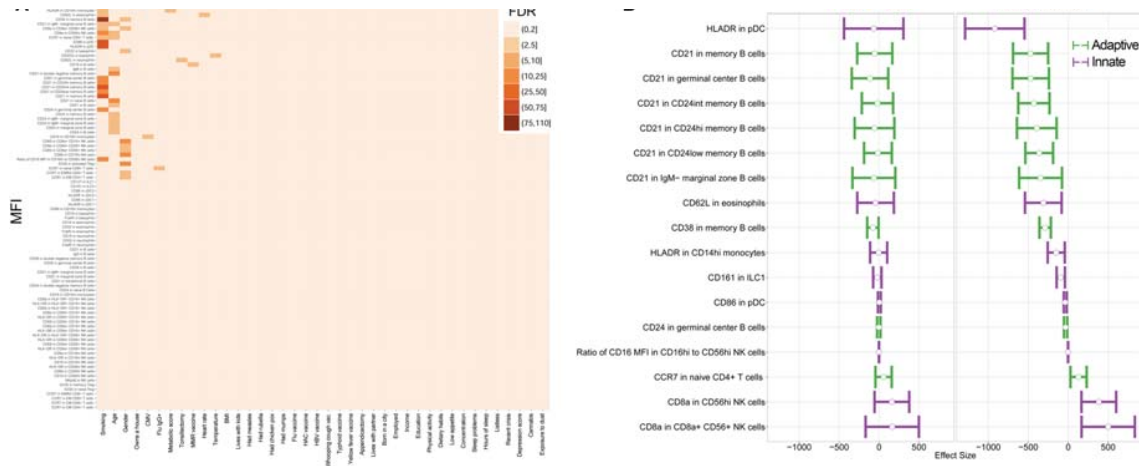
We identified a unique environmental factor that significantly alters circulating levels of human immune cells: active tobacco cigarette smoking, which affects 27% of measured immunophenotypes ([Figures 40](#) and [41](#)).



**Figure 40 : Tobacco smoking strongly impacts immune cell counts in healthy individuals.**

(A) Significant multiplicative effects of 40 non-genetic factors on immune cell counts in healthy individuals. Colors represent levels of association (*i.e.*,  $-\log_{10}$  (P-values)) between the 40 non-genetic factors and immune cell counts, at a false discovery rate (FDR) < 1%. Unless when specifically measured, immunophenotypes were regressed on age, gender, CMV status, batch effects and genome-wide significant SNPs (Table 10). (B) Effects of active and past smoking on circulating levels of immune cells in 1,000 healthy individuals, controlling for age, gender, CMV status, batch effects and genome-wide significant SNPs. The confidence intervals are false coverage-adjusted and calculated using sandwich estimated standard errors. (C) Impact of age and smoking on the number of circulating Treg cells. Brown indicates active smokers, orange indicates past smokers and yellow indicates non-smokers. (D) Flow cytometry plots of HLA-DR expression in Treg cells of representative non-smoker and active smoker.

We observed a 21% increase in the number of circulating CD45<sup>+</sup> cells, and a 25% increase in the number of conventional lymphocytes in smokers as compared to non-smokers (Figure 40B).



**Figure 41 : Smoking strongly impacts protein levels of immune cell markers in healthy individuals.** (A) Significant multiplicative effects of 40 non-genetic factors on protein levels of immune cell markers (i.e., MFIs) in healthy individuals. Colors represent levels of association (i.e.,  $-\log_{10}$  (q-values)) between the 40 non-genetic factors and protein levels of immune cell markers, at a false discovery rate (FDR) < 1%. Unless when specifically measured, immunophenotypes were regressed for age, gender, CMV status, batch effects and genome-wide significant SNPs (Table 10). (B) Significant effect sizes of active smoking on protein levels of immune cell markers in 1,000 healthy individuals while controlling for age, gender, CMV status, batch effects and genome-wide significant SNPs. The confidence intervals are false coverage-adjusted and calculated using sandwich estimated standard errors. Effect sizes in past smokers are shown, for comparison purposes

Previous studies suggested that smokers have alterations in circulating cell populations due to diminished adherence of leukocytes to blood vessel walls, possibly as a result of lower antioxidant concentrations (Tsuchiya et al., 2002).

In addition, we found in active smokers a significant increase of 61%, 65% and 72% of naïve, memory and activated Treg cells, respectively, which was also visible to a lesser extent in past smokers (Figure 40B-D). This increase may reflect the need to counter-balance inflammation and immune activation triggered by cigarette smoke-induced damage to the lung epithelium. Increased numbers of Treg cells may also contribute to the diminished Th1 immunity observed in smokers (Vassallo et al., 2005), and the increased susceptibility to bacterial infection (Bagaitkar et al., 2008).

Active smokers also showed decreased numbers of NK cells, ILCs,  $\gamma\delta$  T cells and all subsets of MAIT cells (Figure 40B). These findings are consistent with a recent study showing that smoking triggers local release of interleukin (IL)-33 by the lung epithelium (Kearley et al., 2015), in turn engaging the IL-33 receptor, ST2, on both innate and non-classical lymphocytes (Monticelli et al., 2011). This likely accounts for the activation and retention of ST2 positive cells in the lung, with an associated decrease in the number of circulating cells. Alternatively, smoking may increase the number of sub-acute bacterial infections in the lung, leading to an attrition of circulating MAIT and ILC subsets via their preferential homing to the lung.

Locus	FACS panel	Immunophenotype	Other immunophenotypes <sup>a</sup>	P-value	Effect size (SE)	Validation P-value <sup>b</sup>	Candidate SNP	Chr	Position	MA <sup>c</sup>	MAF <sup>c</sup>	Candidate gene	Distance to TSS (kb)	Published in previous study
1	4	MFI of CD69 in CD16 <sup>hi</sup> NK cells	Number of CD69 <sup>+</sup> CD16 <sup>hi</sup> NK cells; MFI of CD69 in CD8a <sup>+</sup> and CD69 <sup>+</sup> CD16 <sup>+</sup> NK cells	3.8 x 10 <sup>-30</sup>	7.01 (0.60)	1.0 x 10 <sup>-6</sup>	rs6693121	1	101744633	C	0.40	<i>SIPRI</i>	41.0	-
2	4	MFI of CD16 in CD16 <sup>hi</sup> NK cells	MFI of CD16 in CD56 <sup>hi</sup> NK cells; MFI of HLA-DR in CD16 <sup>hi</sup> , CD8a <sup>+</sup> CD16 <sup>+</sup> and CD69 <sup>+</sup> CD16 <sup>+</sup> NK cells	2.6 x 10 <sup>-66</sup>	21646.41 (1169.55)	7.4 x 10 <sup>-32</sup>	rs3845548	1	161507448	C	0.13	<i>FCGR3A</i>	12.4	Orru et al., <i>Cell</i> 2013
3	7	MFI of CD32 in eosinophils	MFI of CD32 in neutrophils	9.8 x 10 <sup>-233</sup>	868.71 (20.04)	6.0 x 10 <sup>-135</sup>	rs1801274	1	161479745	A	0.49	<i>FCGR2A</i>	0.0	Roederer et al., <i>Cell</i> 2015
4	7	MFI of CD32 in basophils	-	1.6 x 10 <sup>-34</sup>	22.76 (1.79)	4.6 x 10 <sup>-15</sup>	rs61804205	1	161653737	T	0.10	<i>FCGR2B</i>	20.8	-
5	7	MFI of CD62L in eosinophils	MFI of CD62L in neutrophils	9.2 x 10 <sup>-35</sup>	-536.96 (42.13)	1.5 x 10 <sup>-9</sup>	rs2223286	1	169665632	T	0.33	<i>SELL</i>	0.0	-
6	4	MFI of CD8a in CD16 <sup>hi</sup> NK cells	Number of CD8a <sup>+</sup> CD16 <sup>hi</sup> NK cells; MFI of CD8a in CD56 <sup>hi</sup> , CD69 <sup>+</sup> CD56 <sup>hi</sup> , CD8 <sup>+</sup> CD56 <sup>hi</sup> , CD69 <sup>+</sup> CD16 <sup>hi</sup> , CD8a <sup>+</sup> CD16 <sup>hi</sup> and HLA-DR <sup>+</sup> CD16 <sup>hi</sup> NK cells	1.6 x 10 <sup>-50</sup>	-363.27 (23.01)	4.3 x 10 <sup>-22</sup>	rs71411868	2	87026807	A	0.24	<i>CD8A</i>	0.0	Orru et al., <i>Cell</i> 2013
7	4	Number of CD56 <sup>hi</sup> NK cells	Number of CD69 <sup>+</sup> CD56 <sup>hi</sup> NK cells; Number of CD8a <sup>+</sup> CD56 <sup>hi</sup> NK cells	9.1 x 10 <sup>-20</sup>	-0.95 (0.10)	2.9 x 10 <sup>-8</sup>	rs875063	2	111813085	A	0.30	<i>ACOXL/ BCL2L11</i>	0.0	-
8	8	MFI of HLA-DR in cDC3	-	2.6 x 10 <sup>-11</sup>	8.06 (1.16)	8.2 x 10 <sup>-9</sup>	rs143655145	6	32340176	C	0.19	<i>HLA-DRA</i>	67.4	-

9	8	MFI of HLA-DR in cDC1	-	1.8 x 10 <sup>-39</sup>	-0.12 (0.01)	5.9 x 10 <sup>-16</sup>	rs2760994	6	32574308	T	0.38	HLA-DRB1	16.7	-
10	4	MFI of HLA-DR in pDC	MFI of CD86 in pDC; Number of HLA-DR <sup>+</sup> CD56 <sup>hi</sup> NK cells; MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	2.5 x 10 <sup>-39</sup>	348.86 (20.12)	5.5 x 10 <sup>-29</sup>	rs114973966	6	32599163	C	0.18	HLA-DRB1	41.5	-
11	6	MFI of CD24 in IgM <sup>+</sup> marginal zone B cells	MFI of CD24 in naïve, memory, double negative memory, IgM <sup>+</sup> marginal zone and marginal zone B cells	5.9 x 10 <sup>-22</sup>	-2.35 (0.24)	9.2 x 10 <sup>-9</sup>	rs12529793	6	107168676	C	0.08	CD24	254.7	-
12	7	MFI of CD203c in basophils	-	3.5 x 10 <sup>-29</sup>	632.87 (54.60)	5.3 x 10 <sup>-11</sup>	rs138925115	6	132035965	G	0.08	ENPP3	0.0	-
13	1	MFI of CCR7 in CD4 <sup>+</sup> naïve T cells	MFI of CCR7 in CD8b <sup>+</sup> naïve T cells	2.4 x 10 <sup>-20</sup>	-3.22 (0.34)	6.4 x 10 <sup>-9</sup>	rs11648905	16	425298	G	0.43	TMEM8A	0.0	-
14	4	Ratio of CD16 MFI in CD16 <sup>hi</sup> and CD56 <sup>hi</sup> NK cells	MFI of CD16 in CD56 <sup>hi</sup> NK cells	2.0 x 10 <sup>-24</sup>	-0.49 (0.05)	3.4 x 10 <sup>-10</sup>	rs114412914	19	8788184	G	0.15	ACTL9	21.0	-

**Table 10 : Genome-wide signals of association with immunophenotypes in the MI cohort.**

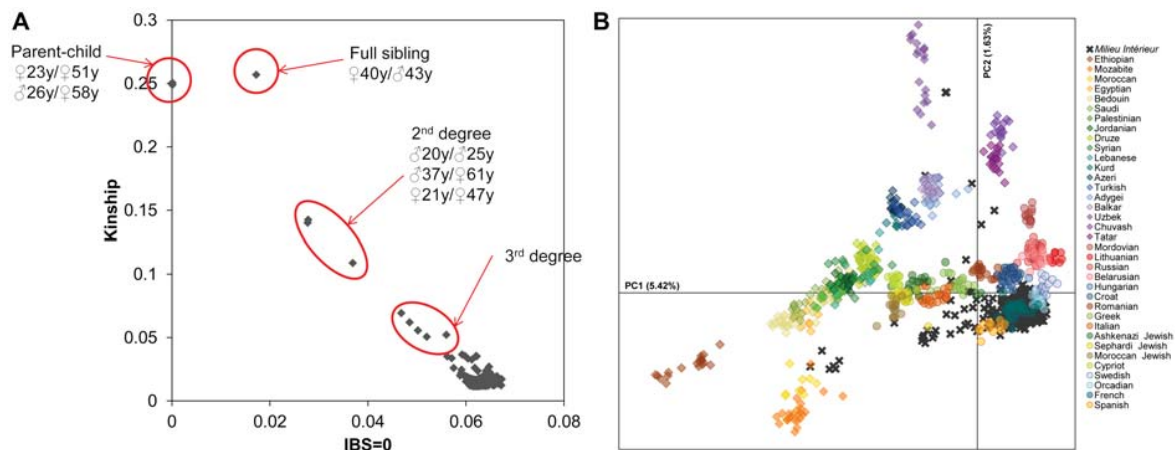
<sup>a</sup>Other immunophenotypes correspond to any measured immunophenotype in the *Milieu Intérieur* cohort that was also significantly associated with the candidate SNP, but to a lesser extent than the main immunophenotype.

<sup>b</sup>Validation P-values were estimated based on immunophenotypes measured from a new blood draw taken ~17 days after the initial visit, in 500 subjects of the *Milieu Intérieur* cohort.

<sup>c</sup>MAF is the frequency of the minor allele (MA) in the *Milieu Intérieur* cohort.

## 2.4. Genome-wide association study of 168 immune cell parameters

To identify common genetic variants affecting immunophenotypes, the *Milieu Intérieur* cohort was genotyped at 945,213 SNPs, enriched in exonic SNPs (Materials & Methods). After quality control filters (**Figure 42**), genotype imputation was performed and yielded a total of 5,699,237 highly accurate SNPs, which were tested for association with all immune cell measurements.



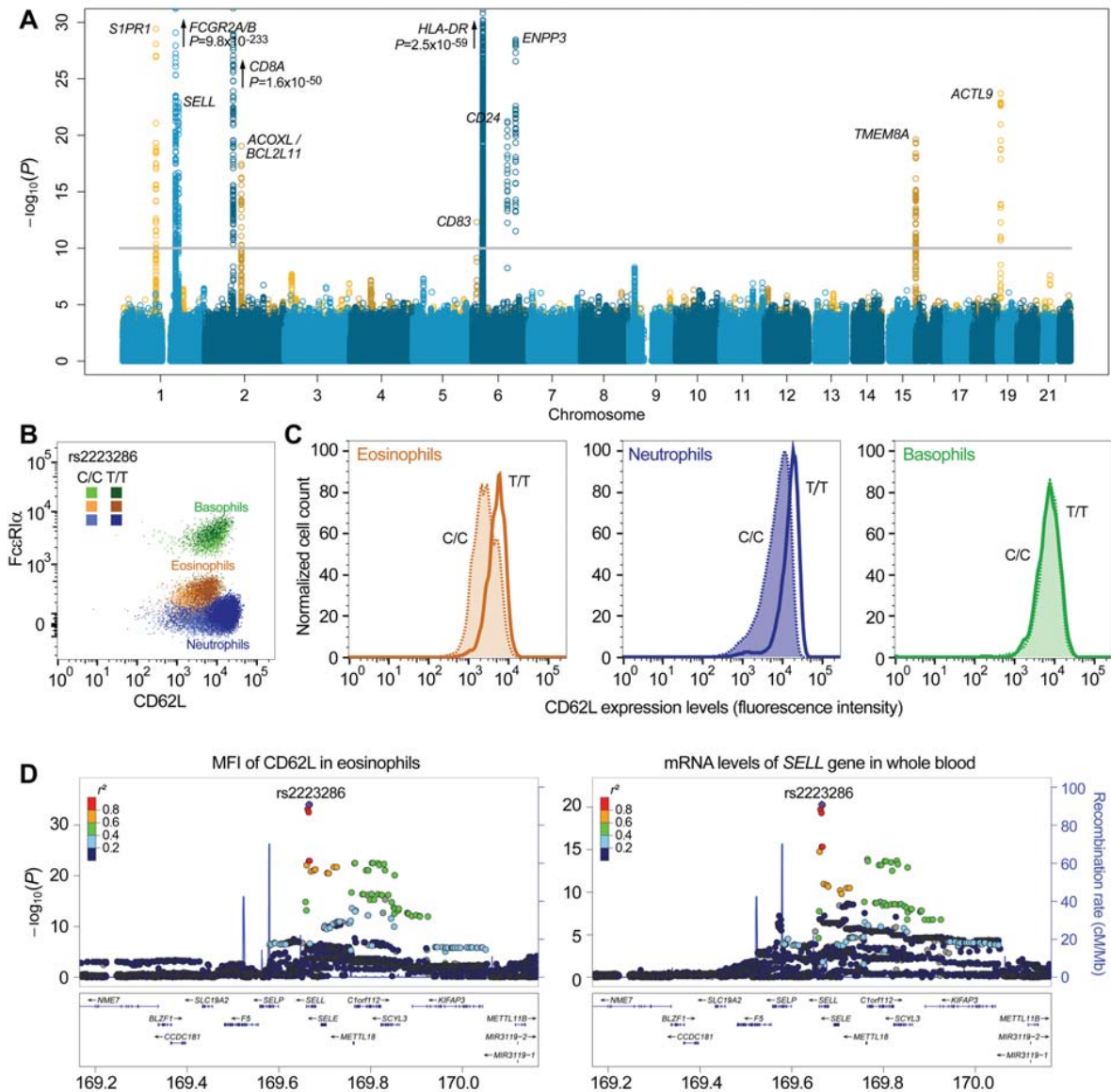
**Figure 42 : Genetic relatedness and structure in the *MI* cohort.**

- (A)** Genetic relatedness in the *Milieu Intérieur* cohort. Pairs of related subjects were detected using an estimate of the kinship coefficient and the proportion of SNPs that are not identical-by-state between all possible pairs of subjects, using KING (Manichaikul et al., 2010). **(B)** Genetic structure of the *MI* cohort. Genetic structure was estimated with the Principal Component Analysis (PCA) implemented in EIGENSTRAT (Patterson et al., 2006). For comparison purposes, the analysis was performed on 261,827 independent SNPs and 1,723 individuals, which include the 1,000 *MI* subjects together with 723 individuals from a selection of 36 populations of North Africa, the Near East, western and northern Europe (Behar et al., 2010)

We first confirmed our power to identify large-effect genotype-phenotype associations by replicating in the cohort well-known genetic associations with non-immune traits, such as eye and hair color or uric acid and cholesterol levels (Materials & Methods).

Genome-wide association analyses were then conducted on the 168 immunophenotypes, using linear mixed models (Zhou and Stephens, 2014). The models were adjusted for the genetic relatedness among subjects and any non-genetic variable identified as associated with each specific immunophenotype by stability selection based on elastic net regression (**Table 9**; Materials & Methods).





**Figure 43 : Genome-wide significant associations with immunophenotypes measured in the MI cohort.**

(A) Manhattan plots of genome-wide significant associations with variants acting locally (in blue) or in *trans* (in yellow) on immunophenotypes. The gray line indicates the genome-wide significance threshold ( $P < 10^{-10}$ ). Zoomed Manhattan plots for all hits are shown in **Figures 44** and **45**. (B) Differential expression of the CD62L protein marker in granulocytes of representative individuals homozygous for the major (T/T, in dark colors) and minor (C/C, in light colors) rs2223286 alleles. (C) Cell-specific CD62L expression is shown for age-matched individuals homozygous for the major (open distribution with solid line) or minor (shaded distribution with dotted line) rs2223286 allele. (D) Zoomed Manhattan plots of genetic associations between SNP rs2223286 in the *SELL* gene and cell-surface expression on CD62L in eosinophils or *SELL* mRNA levels in whole blood. Each point is a SNP, whose color represents its level of linkage disequilibrium ( $r^2$ ) with the best hit (in purple). Blue lines indicate local recombination rates.

We found 14 independent genetic loci associated with 42 out of 168 immunophenotypes (25%), at a conservative genome-wide significant threshold of  $P < 1.0 \times 10^{-10}$  (**Figure 43A**, **Tables 10** and **11**).

We then conducted conditional GWAS for these 42 immunophenotypes by adjusting on the genotypes of genetic variants associated in the main analyses (**Table 10**) and found seven additional signals, of which six were located in close vicinity of the main signal (**Figure 45** and **Table 12**).

We validated all genome-wide significant associations (validation  $P$ -values  $< 10^{-3}$ ; **Table 10**), by measuring corresponding immunophenotypes in a new blood draw taken in 500 of the 1,000 subjects of the *Milieu Intérieur* cohort, sampled 7 to 44 days after the initial visit. We also provide a list of 26 suggestive association signals ( $P < 5.0 \times 10^{-8}$ ), including a number of biologically relevant candidate genes (**Table 12**).

Locus	FACS panel	Immunophenotype	Other immunophenotypes <sup>a</sup>	P-value	Effect size (SE)	Validation P-value <sup>b</sup>	Candidate SNP	Chr	Position	MA <sup>c</sup>	MAF <sup>c</sup>	Candidate gene	Distance to TSS (kb)	Published in previous study
1	4	MFI of CD69 in CD16 <sup>hi</sup> NK cells	Number of CD69 <sup>+</sup> CD16 <sup>hi</sup> NK cells; MFI of CD69 in CD8a <sup>+</sup> and CD69 <sup>+</sup> CD16 <sup>+</sup> NK cells	3.8 x 10 <sup>-30</sup>	7.01 (0.60)	1.0 x 10 <sup>-6</sup>	rs6693121	1	101744633	C	0.40	<i>SIPRI</i>	41.0	-
2	7	MFI of CD32 in eosinophils	MFI of CD32 in neutrophils	9.8 x 10 <sup>-233</sup>	868.71 (20.04)	6.0 x 10 <sup>-135</sup>	rs1801274	1	161479745	A	0.49	<i>FCGR2A</i>	0.0	Roederer et al., <i>Cell</i> 2015
3	4	MFI of CD16 in CD16 <sup>hi</sup> NK cells	MFI of CD16 in CD56 <sup>hi</sup> NK cells; MFI of HLA-DR in CD16 <sup>hi</sup> , CD8a <sup>+</sup> CD16 <sup>+</sup> and CD69 <sup>+</sup> CD16 <sup>+</sup> NK cells	2.6 x 10 <sup>-66</sup>	-21.646, 41 (1169.55)	7.4 x 10 <sup>-32</sup>	rs3845548	1	161507448	C	0.13	<i>FCGR3A</i>	12.4	Ortu et al., <i>Cell</i> 2013
4	7	MFI of CD32 in basophils	-	1.6 x 10 <sup>-34</sup>	22.76 (1.79)	4.6 x 10 <sup>-15</sup>	rs61804205	1	161653737	T	0.10	<i>FCGR2B</i>	20.8	-
5	7	MFI of CD62L in eosinophils	MFI of CD62L in neutrophils	9.2 x 10 <sup>-35</sup>	-536.96 (42.13)	1.5 x 10 <sup>-9</sup>	rs2223286	1	169665632	T	0.33	<i>SELL</i>	0.0	-
6	4	MFI of CD8a in CD16 <sup>hi</sup> NK cells	Number of CD8a <sup>+</sup> CD16 <sup>hi</sup> NK cells; MFI of CD8a in CD56 <sup>hi</sup> , CD69 <sup>+</sup> CD56 <sup>hi</sup> , CD8 <sup>+</sup> CD56 <sup>hi</sup> , CD69 <sup>+</sup> CD16 <sup>hi</sup> , CD8a <sup>+</sup> CD16 <sup>hi</sup> and HLA-DR <sup>+</sup> CD16 <sup>hi</sup> NK cells	1.6 x 10 <sup>-30</sup>	-363.27 (23.01)	4.3 x 10 <sup>-22</sup>	rs71411868	2	87026807	A	0.24	<i>CD8A</i>	0.0	Ortu et al., <i>Cell</i> 2013
7	4	Number of CD56 <sup>hi</sup> NK cells	Number of CD69 <sup>+</sup> CD56 <sup>hi</sup> NK cells; Number of CD8a <sup>+</sup> CD56 <sup>hi</sup> NK cells	9.1 x 10 <sup>-20</sup>	-0.95 (0.10)	2.9 x 10 <sup>-8</sup>	rs875063	2	111813085	A	0.30	<i>ACOXL/ BCL2L1I</i>	0.0	-
8	8	MFI of HLA-DR in cDC3	-	2.6 x 10 <sup>-11</sup>	8.06 (1.16)	8.2 x 10 <sup>-9</sup>	rs143655145	6	32340176	C	0.19	<i>HLA-DRA</i>	67.4	-
9	8	MFI of HLA-DR in cDC1	-	1.8 x 10 <sup>-39</sup>	-0.12 (0.01)	5.9 x 10 <sup>-16</sup>	rs2760994	6	32574308	T	0.38	<i>HLA-DRBI</i>	16.7	-
10	4	MFI of HLA-DR in pDC	MFI of CD86 in pDC; Number of HLA-DR <sup>+</sup> CD56 <sup>hi</sup> NK cells; MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	2.5 x 10 <sup>-59</sup>	348.86 (20.12)	5.5 x 10 <sup>-29</sup>	rs114973966	6	32599163	C	0.18	<i>HLA-DRBI</i>	41.5	-
11	6	MFI of CD24 in IgM <sup>+</sup> marginal zone B cells	MFI of CD24 in naive, memory, double negative memory, IgM <sup>+</sup> marginal zone and marginal zone B cells	5.9 x 10 <sup>-22</sup>	-2.35 (0.24)	9.2 x 10 <sup>-9</sup>	rs12529793	6	107168676	C	0.08	<i>CD24</i>	254.7	-
12	7	MFI of CD203c in basophils	-	3.5 x 10 <sup>-29</sup>	632.87 (54.60)	5.3 x 10 <sup>-11</sup>	rs138925115	6	132035965	G	0.08	<i>ENPP3</i>	0.0	-
13	1	MFI of CCR7 in CD4 <sup>+</sup> naive T cells	MFI of CCR7 in CD8b <sup>+</sup> naive T cells	2.4 x 10 <sup>-20</sup>	-3.22 (0.34)	6.4 x 10 <sup>-9</sup>	rs11648905	16	425298	G	0.43	<i>TMEM8A</i>	0.0	-
14	4	Ratio of CD16 MFI in CD16 <sup>hi</sup> and CD56 <sup>hi</sup> NK cells	MFI of CD16 in CD56 <sup>hi</sup> NK cells	2.0 x 10 <sup>-24</sup>	-0.49 (0.05)	3.4 x 10 <sup>-10</sup>	rs114412914	19	8788184	G	0.15	<i>ACTL9</i>	21.0	-

<sup>a</sup>Other immunophenotypes correspond to any measured immunophenotype in the *M/I* cohort that was also significantly associated with the candidate SNP, but to a lesser extent than the main immunophenotype.

<sup>b</sup>Validation P-values were estimated based on immunophenotypes measured from a new blood draw taken ~17 days after the initial visit, in 500 subjects of the *M/I* cohort.

<sup>c</sup>MAF is the frequency of the minor allele (MA) in the *M/I* cohort.

Immunophenotype	Candidate gene	NanoString genes <sup>d</sup>	Candidate SNP eQTL P-value <sup>e</sup>	Associated gene expression <sup>f</sup>	Amino-acid altering mutation	Disease/trait-associated SNP	Disease/trait	Reference
MFI of CD69 in CD16 <sup>hi</sup> NK cells	<i>SIPRI</i>	<i>SIPRI</i>	0.915	<i>SIPRI</i>	-	-	-	-
MFI of CD32 in eosinophils	<i>FCGR2A</i>	<i>ITLN2, FCER1G, FCGR2A, FCGR2B, FCGR3A, FCGR3B, FCGR2C</i>	$1.9 \times 10^{-5}$	<i>FCGR2B</i>	rs1801274	rs1801274	Systemic lupus erythematosus; Inflammatory bowel disease; Kawasaki disease; Ulcerative colitis	Jostins et al., <i>Nature</i> 2012; Khor et al., <i>Nat Genet</i> 2011; Andersson et al., <i>Nat Genet</i> 2011; Asano et al., <i>Nat Genet</i> 2009
MFI of CD16 in CD16 <sup>hi</sup> NK cells	<i>FCGR3A</i>	<i>ITLN2, FCER1G, FCGR2A, FCGR2B, FCGR3A, FCGR3B, FCGR2C</i>	$6.3 \times 10^{-21}$	<i>FCGR2B</i>	rs9427398 ( $r^2=0.72$ )	rs72717009 ( $r^2=0.62$ ); rs67418890 ( $r^2=0.74$ )	Rheumatoid arthritis; Cholesterol ester content of extra large HDL	Okada et al., <i>Nature</i> 2013; Ketunen et al., <i>Nat Genet</i> 2012
MFI of CD32 in basophils	<i>FCGR2B</i>	<i>ITLN2, FCER1G, FCGR2A, FCGR2B, FCGR3A, FCGR3B, FCGR2C</i>	$4.1 \times 10^{-22}$	<i>FCGR2B</i>	-	-	-	-
MFI of CD62L in eosinophils	<i>SELL</i>	<i>SELL, SELE</i>	$4.9 \times 10^{-21}$	<i>SELL</i>	-	-	-	-
MFI of CD8a in CD16 <sup>hi</sup> NK cells	<i>CD8A</i>	<i>CD8A, CD8B</i>	0.011	<i>CD8A</i>	-	-	-	-
Number of CD56 <sup>hi</sup> NK cells	<i>ACOXL/ BCL2L11</i>	<i>BCL2L11</i>	0.254	<i>BCL2L11</i>	-	rs10207392 ( $r^2=0.60$ )	Mean Corpuscular Volume	van der Harst et al., <i>Nature</i> 2012
MFI of HLA-DR in cDC3	<i>HLA-DRA</i>	<i>C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPA1, HLA-DPBI</i>	-	-	-	-	-	-
MFI of HLA-DR in cDC1	<i>HLA-DRB1</i>	<i>C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPA1, HLA-DPBI</i>	-	-	-	-	-	-
MFI of HLA-DR in pDC	<i>HLA-DRB1</i>	<i>C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPA1, HLA-DPBI</i>	-	-	rs9272785 ( $r^2=0.91$ )	rs660895 ( $r^2=0.87$ ); rs114800139 ( $r^2=0.60$ )	IgA nephropathy; Rheumatoid arthritis; Vogt-Koyanagi-Harada syndrome	Yu et al., <i>Nat Genet</i> 2011; Plenge et al., <i>N Engl J Med</i> 2007; Hou et al., <i>Nat Genet</i> 2014
MFI of CD24 in IgM <sup>+</sup> marginal zone B cells	<i>CD24</i>	<i>ATG5, CD24</i>	$1.9 \times 10^{-5}$	<i>CD24</i>	-	-	-	-
MFI of CD203c in basophils	<i>ENPP3</i>	<i>ARG1</i>	0.030	<i>ARG1</i>	-	-	-	-
MFI of CCR7 in CD4 <sup>+</sup> naive T cells	<i>TMEM8A</i>	-	-	-	rs2071915 ( $r^2=0.61$ )	-	-	-
Ratio of CD16 MFI in CD16 <sup>hi</sup> and CD56 <sup>hi</sup> NK cells	<i>ACTL9</i>	-	-	-	-	rs2164983 ( $r^2=0.96$ )	Atopic dermatitis	Puttemoster et al., <i>Nat Genet</i> 2011

**Table 11 : Genome-wide signals of association with immunophenotypes in the MI cohort, with their consequences on gene expression, protein sequence and human diseases and traits**

<sup>d</sup>mRNA levels of immunity-related genes that are close to the 14 genome-wide hits were measured in whole blood of the 1,000 MI subjects.

<sup>e</sup>eQTL mapping was performed between the candidate SNP and mRNA levels of surrounding immunity-related genes. The strongest P-value among all tested genes is reported.

<sup>f</sup>The gene whose expression is the most strongly associated with the candidate SNP is reported.

**Remark :** To facilitate reading, the original table has been split into 2 parts. To be able to follow, the « Immunophenotypes » (green) and « Candidate genes » (orange) columns were duplicated

Locus	GWAS hit	FACS panel	Immunophenotype	Other immunophenotypes <sup>a</sup>	P-value	Effect size (SE)	Validation P-value <sup>b</sup>	Candidate SNP	Chr	Position	MA <sup>c</sup>	MAF <sup>c</sup>	Candidate gene	Distance to TSS (kb)	LD (r <sup>2</sup> ) with main signal
1	Genome-wide significant when conditioning on rs3845548	4	MFI of HLA-DR in CD8a <sup>+</sup> CD16 <sup>+</sup> NK cells	MFI of HLA-DR in CD16 <sup>+</sup> NK cells	2.3 x 10 <sup>-12</sup>	1746.93 (247.14)	3.5 x 10 <sup>-4</sup>	rs140970775	1	161503013	A	0.08	FCGR3A	16.8	0.62
2	Genome-wide significant when conditioning on rs61804205	7	MFI of CD32 in eosinophils	-	2.7 x 10 <sup>-20</sup>	-9.66 (1.03)	4.8 x 10 <sup>-9</sup>	rs17413333	1	161675580	G	0.404	FCGR2B	42.7	0.01
3	Genome-wide significant when conditioning on rs34988194	4	MFI of CD8a in CD8a <sup>+</sup> CD16 <sup>+</sup> NK cells	MFI of CD8a in CD16 <sup>+</sup> NK cells; MFI of CD8a in CD69 <sup>+</sup> CD16 <sup>+</sup> NK cells	8.7 x 10 <sup>-30</sup>	0.61 (0.05)	2.4 x 10 <sup>-16</sup>	rs4832315	2	87021035	A	0.44	CD8A	0.0	0.39
4	Genome-wide significant when conditioning on rs2760994	8	MFI of HLA-DR in cDC1	-	1.0 x 10 <sup>-11</sup>	-0.07 (0.01)	3.6 x 10 <sup>-8</sup>	rs72836542	6	14091682	C	0.284	CD83	26.2	0.00
5	Genome-wide significant when conditioning on rs2760994	8	MFI of HLA-DR in cDC1	-	7.5 x 10 <sup>-13</sup>	0.07 (0.01)	1.9 x 10 <sup>-8</sup>	rs116072659	6	32414290	G	0.284	HLA-DRA	6.7	0.01
6	Genome-wide significant when conditioning on rs114973966	8	MFI of HLA-DR in pDC	-	2.8 x 10 <sup>-23</sup>	-193.10 (19.02)	1.6 x 10 <sup>-11</sup>	rs114748268	6	32445691	C	0.422	HLA-DRA	38.1	0.29
7	Genome-wide significant when conditioning on rs114973966	8	MFI of HLA-DR in CD14 <sup>+</sup> monocytes	-	1.4 x 10 <sup>-12</sup>	1.02 (0.14)	4.7 x 10 <sup>-12</sup>	rs139299944	6	32602665	C	0.437	HLA-DRB1	45.1	0.16
15	Suggestive in main GWAS	4	MFI of HLA-DR in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	MFI of HLA-DR in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	3.9 x 10 <sup>-8</sup>	0.14 (0.03)	0.148	rs12402741	1	161404254	G	0.13	FCGR3A	115.6	-
16	Suggestive in main GWAS	5	MFI of CD16 in CD16 <sup>+</sup> monocytes	-	2.7 x 10 <sup>-10</sup>	-46.03 (7.29)	4.8 x 10 <sup>-4</sup>	rs10919544	1	161508763	T	0.35	FCGR3A	11.1	-
17	Suggestive in main GWAS	7	MFI of CD16 in neutrophils	-	2.6 x 10 <sup>-9</sup>	4065.48 (686.78)	1.5 x 10 <sup>-4</sup>	rs17411858	1	161617516	A	0.13	FCGR3A	97.7	-
18	Suggestive in main GWAS	2	Number of memory Treg	Number of activated Treg	8.8 x 10 <sup>-9</sup>	-5.76 (1.04)	6.0 x 10 <sup>-4</sup>	rs74746864	1	199023240	A	0.11	LINC01221	0.0	-
19	Suggestive in main GWAS	6	MFI of CD21 in CD24 <sup>+</sup> memory B cells	MFI of CD21 in memory and transitional B cells	1.7 x 10 <sup>-9</sup>	-351.99 (58.50)	5.9 x 10 <sup>-5</sup>	rs61821132	1	207651059	T	0.07	CR2	0.0	-
20	Suggestive in main GWAS	4	Ratio of CD16 MFI in CD16 <sup>+</sup> to CD56 <sup>+</sup> NK cells	Ratio of CD16 <sup>+</sup> to CD56 <sup>+</sup> NK cells; Number of NKp44 <sup>+</sup> CD56 <sup>+</sup> ILC3	5.4 x 10 <sup>-9</sup>	0.23 (0.04)	0.073	rs2009581	2	111807677	G	0.27	ACOXL / BCL2L11	0.0	-
21	Suggestive in main GWAS	5	Number of CD14 <sup>+</sup> monocytes	-	6.4 x 10 <sup>-9</sup>	-9.35 (1.60)	2.2 x 10 <sup>-4</sup>	rs10932325	2	211191424	A	0.30	LANCL1	0.0	-
22	Suggestive in main GWAS	1	MFI of CCR7 in naive CD4 <sup>+</sup> T cells	-	2.0 x 10 <sup>-8</sup>	-2.44 (0.43)	3.9 x 10 <sup>-4</sup>	rs2630797	3	21881706	C	0.22	ZNF385D	88.9	-
23	Suggestive in main GWAS	6	MFI of CD27 in B cells	-	2.0 x 10 <sup>-8</sup>	17.08 (3.14)	5.5 x 10 <sup>-5</sup>	rs17014824	3	24603532	A	0.10	THRB	67.2	-
24	Suggestive in main GWAS	4	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	-	1.4 x 10 <sup>-8</sup>	2.02 (0.36)	0.307	rs72872226	3	54732073	T	0.22	CACNA2D3	0.0	-
25	Suggestive in main GWAS	4	MFI of CD8a in CD16 <sup>+</sup> NK cells	-	5.0 x 10 <sup>-8</sup>	174.63 (31.56)	1.6 x 10 <sup>-4</sup>	rs22920226	5	32738891	C	0.13	NPR3	0.0	-
26	Suggestive in main GWAS	8	MFI of HLA-DR in cDC1	-	7.3 x 10 <sup>-10</sup>	-0.06 (0.01)	1.4 x 10 <sup>-4</sup>	rs72836542	6	14091682	C	0.28	CD83	25.8	-
27	Suggestive in main GWAS	1	MFI of CCR7 in CM CD4 <sup>+</sup> T cells	-	3.8 x 10 <sup>-8</sup>	-0.05 (0.01)	0.080	rs115246869	6	29836741	G	0.45	HLA-G	42.0	-
28	Suggestive in main GWAS	1	CD4 <sup>+</sup> to CD8 <sup>+</sup> T cell ratio	-	9.0 x 10 <sup>-9</sup>	-0.12 (0.02)	3.6 x 10 <sup>-4</sup>	rs116217042	6	31164595	G	0.15	HLA-C	75.3	-
29	Suggestive in main GWAS	1	Number of naive CD8 <sup>+</sup> T cells	-	1.5 x 10 <sup>-8</sup>	1.36 (0.25)	7.6 x 10 <sup>-4</sup>	rs143626981	6	31267207	C	0.30	HLA-C	27.3	-
30	Suggestive in main GWAS	4	MFI of HLA-DR in CD56 <sup>+</sup> NK cells	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	5.6 x 10 <sup>-9</sup>	1.75 (0.30)	1.2 x 10 <sup>-3</sup>	rs28383187	6	32572113	A	0.49	HLA-DRB1	14.5	-
31	Suggestive in main GWAS	2	MFI of ICOS in activated Treg	-	4.7 x 10 <sup>-9</sup>	-0.79 (0.13)	5.7 x 10 <sup>-6</sup>	rs116687380	6	32680640	T	0.33	HLA-DRB1	123.0	-
32	Suggestive in main GWAS	4	MFI of CD8a in CD16 <sup>+</sup> NK cells	-	4.9 x 10 <sup>-9</sup>	127.80 (21.73)	1.9 x 10 <sup>-5</sup>	rs10738466	9	16903366	A	0.42	BNC2	32.6	-
33	Suggestive in main GWAS	4	MFI of CD8a in CD69 <sup>+</sup> CD56 <sup>+</sup> NK cells	-	1.6 x 10 <sup>-8</sup>	181.02 (31.79)	0.644	rs10886566	10	121770668	G	0.17	SEC23IP	118.6	-
34	Suggestive in main GWAS	4	MFI of CD69 in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	-	4.4 x 10 <sup>-8</sup>	146.85 (26.70)	0.201	rs61145574	11	81663931	TA	0.08	miR4300HG	0.0	-
35	Suggestive in main GWAS	4	MFI of CD69 in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	-	2.9 x 10 <sup>-8</sup>	142.35 (25.53)	0.305	rs6483300	11	88355868	C	0.09	GRM5	0.0	-
36	Suggestive in main GWAS	1	MFI of CCR7 in CM CD8 <sup>+</sup> T cells	-	1.7 x 10 <sup>-8</sup>	1.19 (0.21)	0.050	rs25680	12	6554628	G	0.25	CD27	0.0	-
37	Suggestive in main GWAS	6	Number of IgM <sup>+</sup> marginal zone B cells	-	2.3 x 10 <sup>-8</sup>	-0.91 (0.16)	5.5 x 10 <sup>-4</sup>	rs9902832	17	3534833	T	0.25	SHIPK	0.0	-
38	Suggestive in main GWAS	7	MFI of PiceA1 in eosinophils	-	1.9 x 10 <sup>-10</sup>	-3.69 (0.58)	1.4 x 10 <sup>-8</sup>	rs7223840	17	4562150	T	0.25	ALOX15	45.5	-
39	Suggestive in main GWAS	4	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	4.2 x 10 <sup>-9</sup>	2.38 (0.41)	5.6 x 10 <sup>-6</sup>	rs35358447	19	8791581	AG	0.16	ACTL9	17.6	-
40	Suggestive in main GWAS	4	MFI of CD69 in CD16 <sup>+</sup> NK cells	-	2.6 x 10 <sup>-8</sup>	5.75 (1.03)	0.970	rs73192987	21	36690844	G	0.10	LOC100506403	262.2	-

<sup>a</sup>Other immunophenotypes correspond to any measured immunophenotype in the *MJ* cohort that was also significantly associated with the candidate SNP, but to a lesser extent than the main immunophenotype.

<sup>b</sup>Validation P-values were estimated based on immunophenotypes measured from a new blood draw taken ~17 days after the initial visit, in 500 subjects of the *MJ* cohort.

<sup>c</sup>MAF is the frequency of the minor allele (MA) in the *MJ* cohort.

GWAS hit	Immunophenotype	Candidate gene	NanoString genes <sup>d</sup>	Candidate SNP eQTL P-value <sup>e</sup>	Associated gene expression <sup>f</sup>	Amino-acid altering mutation	Disease-associated SNP	Disease	Reference
Genome-wide significant when conditioning on rs3845548	MFI of HLA-DR in CD8a <sup>+</sup> CD16 <sup>+</sup> NK cells	FCGR3A	ITLN2, FCER1G, FCGR2A, FCGR2B, FCGR3A, FCGR3B, FCGR2C	3.3 x 10 <sup>-15</sup>	FCGR2B	-	-	-	-
Genome-wide significant when conditioning on rs61804205	MFI of CD33 in basophils	FCGR2B	ITLN2, FCER1G, FCGR2A, FCGR2B, FCGR3A, FCGR3B, FCGR2C	0.025	FCGR2A	-	-	-	-
Genome-wide significant when conditioning on rs34988194	MFI of CD8a in CD8a <sup>+</sup> CD16 <sup>+</sup> NK cells	CD8A	CD8A, CD8B	0.243	CD8A	-	-	-	-
Genome-wide significant when conditioning on rs2760994	MFI of HLA-DR in cDC1	CD83	CD83	3.2 x 10 <sup>-22</sup>	CD83	-	-	-	-
Genome-wide significant when conditioning on rs2760994	MFI of HLA-DR in cDC1	HLA-DRA	C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPAI, HLA-DPBI	-	-	rs116072659 (r <sup>2</sup> =0.72)	rs9274407 (r <sup>2</sup> =0.64)	Amoxicillin-induced liver injury	Lucena et al., <i>Gastroenterology</i> 2011
Genome-wide significant when conditioning on rs114973966	MFI of HLA-DR in pDC	HLA-DRA	C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPAI, HLA-DPBI	-	-	rs114748268 (r <sup>2</sup> =0.62)	-	-	-
Genome-wide significant when conditioning on rs114973966	MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	HLA-DRB1	C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPAI, HLA-DPBI	-	-	-	-	-	-
Suggestive in main GWAS	MFI of HLA-DR in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	FCGR3A	ITLN2, FCER1G, FCGR2A, FCGR3A, FCGR2C, FCGR3B, FCGR2B	3.4 x 10 <sup>-9</sup>	FCGR2B	rs9427398 (r <sup>2</sup> =0.62)	rs72717009 (r <sup>2</sup> =1.0)	Rheumatoid arthritis	Okada et al., <i>Nature</i> 2013
Suggestive in main GWAS	MFI of CD16 in CD16 <sup>hi</sup> monocytes	FCGR3A	ITLN2, FCER1G, FCGR2A, FCGR3A, FCGR2C, FCGR3B, FCGR2B	2.3 x 10 <sup>-3</sup>	FCGR2A	-	-	-	-
Suggestive in main GWAS	MFI of CD16 in neutrophils	FCGR3A	ITLN2, FCER1G, FCGR2A, FCGR3A, FCGR2C, FCGR3B, FCGR2B	3.8 x 10 <sup>-22</sup>	FCGR2B	-	-	-	-
Suggestive in main GWAS	Number of memory Treg	LINC01221	C4BPA, CD55, CR2, CRI, CD46, CD54	0.024	CD46	-	-	-	-
Suggestive in main GWAS	Ratio of CD16 MFI in CD16 <sup>hi</sup> to CD56 <sup>hi</sup> NK cells	CR2	BCL2L11	0.451	BCL2L11	-	rs7578982 (r <sup>2</sup> =0.72)	Monocyte count	Okada et al., <i>PLoS Genet</i> 2011
Suggestive in main GWAS	Number of CD14 <sup>hi</sup> monocytes	LANCL1	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CCR7 in native CD4 <sup>+</sup> T cells	ZNF385D	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD27 in B cells	THRB	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	CACMA2D3	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD8a in CD16 <sup>hi</sup> NK cells	NP63	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of HLA-DR in cDC1	CD83	CD83	3.2 x 10 <sup>-22</sup>	CD83	-	-	-	-
Suggestive in main GWAS	MFI of CCR7 in CM CD4 <sup>+</sup> T cells	HLA-G	HLA-A	0.445	HLA-A	-	-	-	-
Suggestive in main GWAS	CD4 <sup>+</sup> to CD8 <sup>+</sup> T cell ratio	HLA-C	TUBB, HLA-C, HLA-B, LTA, TNF	4.6 x 10 <sup>-3</sup>	HLA-C	-	rs3130559 (r <sup>2</sup> =0.73)	Chronic obstructive pulmonary disease-related biomarkers	Kim et al., <i>Am J Respir Crit Care Med</i> 2012
Suggestive in main GWAS	Number of native CD8 <sup>+</sup> T cells	HLA-C	TUBB, HLA-C, HLA-B, LTA, TNF	2.4 x 10 <sup>-4</sup>	HLA-C	-	-	-	-
Suggestive in main GWAS	MFI of HLA-DR in CD56 <sup>hi</sup> NK cells	HLA-DRB1	C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPAI, HLA-DPBI	-	-	rs481139 (r <sup>2</sup> =0.61)	-	-	-
Suggestive in main GWAS	MFI of ICOS in activated Treg	HLA-DRB1	C2, CFB, HLA-DRA, HLA-DOB, TAP2, PSMB8, TAP1, PSMB9, HLA-DMB, HLA-DMA, HLA-DPAI, HLA-DPBI	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD8a in CD16 <sup>hi</sup> NK cells	BNC2	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD8a in CD69 <sup>+</sup> CD56 <sup>+</sup> NK cells	SEC23IP	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD69 in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	miR4300HG	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of CD69 in HLA-DR <sup>+</sup> CD16 <sup>+</sup> NK cells	GRM5	CTSC	7.8 x 10 <sup>-3</sup>	CTSC	-	-	-	-
Suggestive in main GWAS	MFI of CCR7 in CM CD8 <sup>+</sup> T cells	CD27	CD9, LITBR, CD27, GAPDH, LAG3	1.7 x 10 <sup>-24</sup>	CD27	rs25680	-	-	-
Suggestive in main GWAS	Number of IgM <sup>+</sup> marginal zone B cells	SHPK	ITGAE	-	-	rs9436 (r <sup>2</sup> =0.95)	-	-	-
Suggestive in main GWAS	MFI of Feer1 in eosinophils	ALOX15	-	-	-	-	-	-	-
Suggestive in main GWAS	MFI of HLA-DR in CD8a <sup>+</sup> CD56 <sup>+</sup> NK cells	ACTL9	-	-	-	-	rs2164983 (r <sup>2</sup> =1.0)	Atopic dermatitis	Patemoster et al., <i>Nat Genet</i> 2011
Suggestive in main GWAS	MFI of CD69 in CD16 <sup>hi</sup> NK cells	LOC100506403	RUNX1	0.340	RUNX1	-	-	-	-

**Table 12 : Other genome-wide significant (P < 10-10) or suggestive (P < 5x10-8) signals of association with immunophenotypes in the MI cohort**

<sup>d</sup>mRNA levels of immunity-related genes that are close to the 14 genome-wide hits were measured in whole blood of the 1,000 MI subjects.  
<sup>e</sup>eQTL mapping was performed between the candidate SNP and mRNA levels of surrounding immunity-related genes. The strongest P-value among all tested genes is reported.  
<sup>f</sup>The gene whose expression is the most strongly associated with the candidate SNP is reported.

*Remark : To facilitate reading, the original table has been split into 2 parts. To be able to follow the « GWAS hits » (blue), « Immunophenotypes » (green) and « Candidate genes » (orange) columns were duplicated.*

## 2.5. Most genetic associations act as immune cell-specific protein QTLs

Of the 42 immunophenotypes for which a significant genetic association was detected, 36 (86%) were MFI, which measures the cell-specific expression of protein markers conventionally used to determine the differentiation or activation state of leukocytes. Strikingly, for 29 of these 36 MFI measurements (81%), genetic association was observed between the protein MFI and SNPs located in the vicinity of the gene encoding the corresponding protein (**Figure 44**), *i.e.*, local protein QTLs (herein referred to as local-pQTLs). For instance, genetic variation close to the *ENPP3* gene was associated with CD203c MFI in basophils (rs138925115,  $P = 3.5 \times 10^{-29}$ ), *CD24* with CD24 MFI in marginal zone B cells (rs12529793,  $P = 5.9 \times 10^{-22}$ ) and *CD8A* with the MFI of CD8a in CD69<sup>+</sup> CD16<sup>hi</sup> NK cells (rs71411868,  $P = 1.6 \times 10^{-50}$ ).

We identified three independent local-pQTLs in the *FCGR* gene cluster (**Table 10**), which encodes the most important Fc receptors for inducing phagocytosis of opsonized microbes. Two of them were previously reported; genetic variation close to *FCGR3A* (Orri et al., 2013) and in *FCGR2A* (Roederer et al., 2015) were associated here with the MFI of CD16 in CD16<sup>hi</sup> NK cells (rs3845548,  $P = 2.6 \times 10^{-66}$ ) and the MFI of CD32 in eosinophils (rs1801274,  $P = 9.8 \times 10^{-233}$ ), respectively.

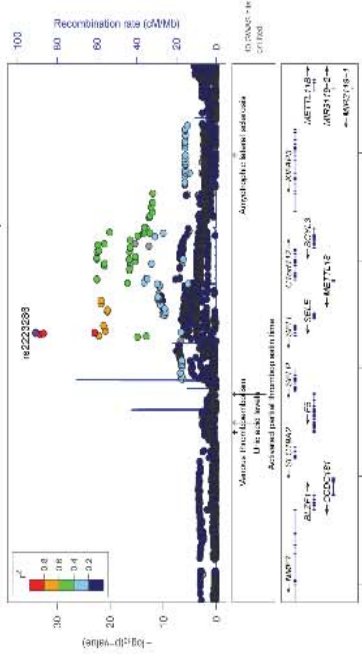
The latter variant corresponds to the CD32a R/H amino acid-altering variant, which has been strongly associated with different auto-immune disorders, such as systemic lupus erythematosus, inflammatory bowel disease, Kawasaki disease and ulcerative colitis (**Table 11**).

The third signal associated *FCGR2B* variation with the MFI of CD32 in basophils (rs61804205,  $P = 1.6 \times 10^{-34}$ ). Consistently, it is known that the anti-CD32 antibody used in this study can recognize both *FCGR2A* and *FCGR2B* gene products (CD32a and CD32b, respectively), and that basophils express both CD32a and CD32b proteins, while eosinophils and neutrophils predominantly express CD32a (Cassard et al., 2012).

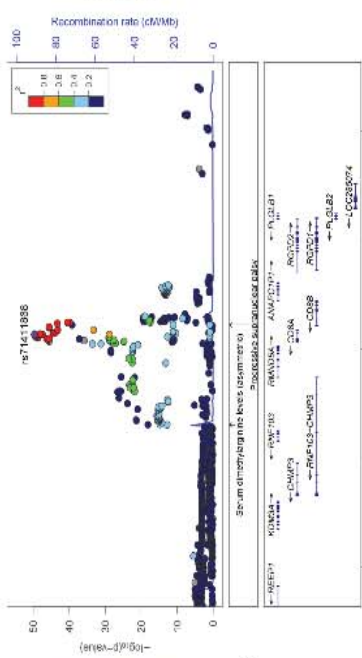




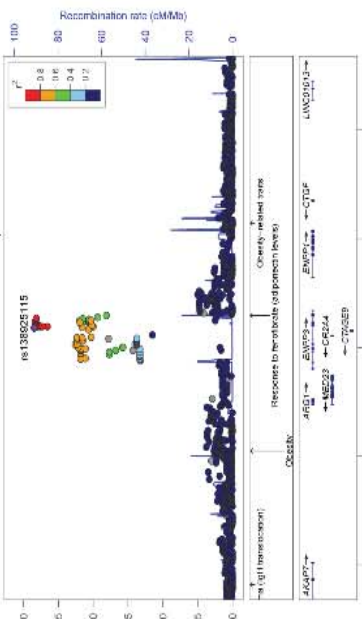
MFI of CD62L in eosinophils



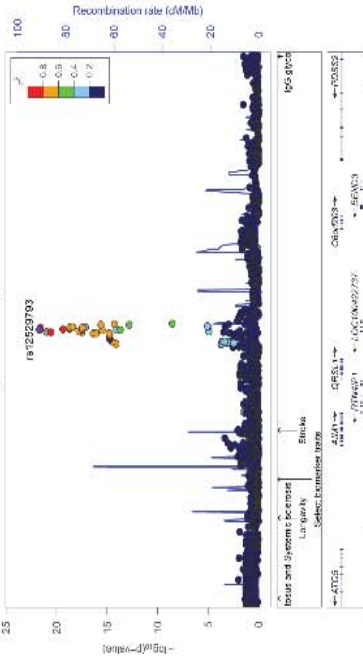
MFI of CD8a in CD16hi NK cells



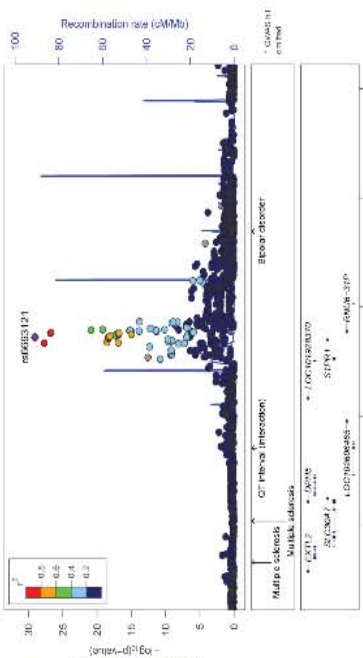
MFI of CD203c in basophils



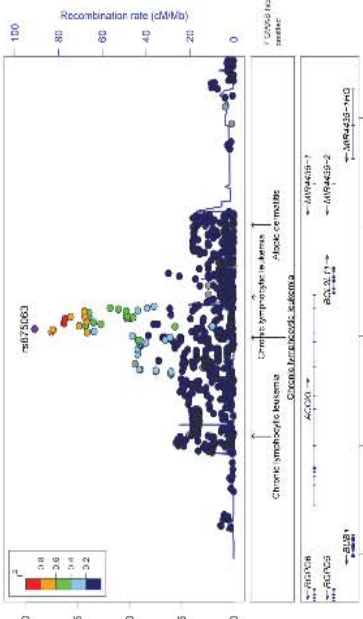
MFI of CD24 in IgM+ marginal zone B cells

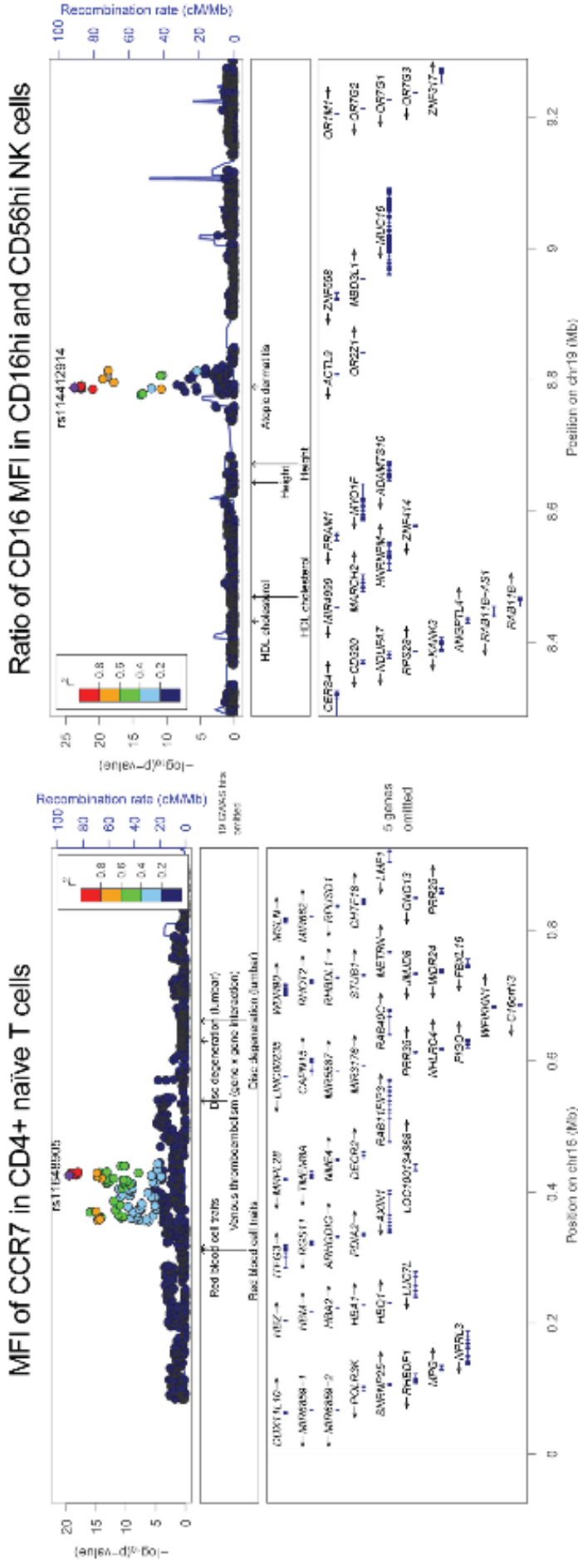


MFI of CD69 in CD16hi NK cells



Number of CD56hi NK cells





**Figure 44 : Local association signals for the 14 genome-wide significant hits associated with immunophenotypes measured in the *MI* cohort.** Each point is a SNP, whose color represents its level of linkage disequilibrium ( $r^2$ ) with the best hit (in purple). Blue lines indicate local recombination rates.

A number of other local-pQTLs were cell-specific; we identified a local-pQTL at the *SELL* gene associated with CD62L MFI in eosinophils and neutrophils (rs2223286,  $P = 9.2 \times 10^{-35}$  and  $4.6 \times 10^{-14}$ , respectively), but not in basophils (**Figure 43B-C**).

Similarly, three different association signals were found in the *HLA-DR* gene region, with the MFI of HLA-DR in pDCs and CD14<sup>hi</sup> monocytes (rs114973966,  $P = 2.5 \times 10^{-59}$ ), in cDC1 (rs2760994,  $P = 1.8 \times 10^{-39}$ ) and in cDC3 cells (rs143655145,  $P = 2.6 \times 10^{-11}$ ).

To verify if these signals are independent from each other, we conducted omnibus association tests on imputed HLA alleles (Jia et al., 2013). We found that the association signals in CD14<sup>hi</sup> monocytes, pDCs and cDC1 actually resulted from different amino acid-altering variants at the same multi-allelic position 13 of the HLA-DR $\beta$ 1 protein ( $P = 2.0 \times 10^{-47}$ ,  $7.0 \times 10^{-90}$  and  $5.3 \times 10^{-41}$  in CD14<sup>hi</sup> monocytes, pDC and cDC1, respectively; **Tables 13 and 14**), recently shown to explain a large part of the association signal in the *HLA* locus for type 2 diabetes (Hu et al., 2015).

A different amino-acid variant, at position 67 of HLA-DR $\beta$ 1, was identified in cDC3 ( $P = 3.9 \times 10^{-13}$ ).

Conditional analyses also revealed independent associations of HLA-DR cell-surface expression with two residues in class I *HLA-B* gene (position 97 and 194;  $P = 3.8 \times 10^{-17}$  and  $1.3 \times 10^{-18}$ ; **Tables 13 and 14**).

Collectively, these results show that cell-specific protein levels of markers that play an important role in immune cell differentiation and activation can be affected by common genetic variants, of which some are known to be implicated in human pathogenesis.

Immunophenotype	HLA-DRβ1 position 13													
	Omnibus test		Beta (95% CI) of amino-acid substitutions								Best SNP P-value <sup>a</sup>	Best SNP P-value <sup>a</sup> conditioning on associated HLA amino-acid positions	Proportion of variance explained by HLA amino acid positions	
	P-value	Val. P-val	Phe (F=12.0%)	Gly (F=4.3%)	His (F=17.8%)	Arg (F=14.8%)	Ser (F=39.2%)	Tyr (F=12.1%)						
MFI of HLA-DR in cDC1	5.3x10 <sup>-41</sup>	1.8x10 <sup>-16</sup>	-0.11 (-0.14, -0.08)	0.09 (0.05, 0.14)	-0.01 (-0.01, 0.04)	-0.06 (-0.09, -0.04)	0.10 (0.08, 0.11)	-0.09 (-0.11, -0.06)	rs2760994	5.1 x 10 <sup>-39</sup>	0.049	13.80%		
MFI of HLA-DR in pDC	7.0x10 <sup>-90</sup>	1.0x10 <sup>-39</sup>	-167.8 (-220.2, -115.5)	-79.2 (-163.9, -5.4)	343.6 (303.8, 383.5)	35.2 (-12.3, 82.6)	-1.7 (-36.5, 33.0)	-304.3 (-352.8, -255.8)	rs114973966	2.8 x 10 <sup>-88</sup>	1.0	28.50%		
MFI of CD86 in pDC	4.2x10 <sup>-22</sup>	6.9x10 <sup>-8</sup>	-6.63 (-9.65, -3.62)	-2.98 (-7.81, 1.85)	11.20 (8.71, 13.68)	-0.79 (-3.50, 1.92)	0.70 (-1.28, 2.68)	-8.02 (-10.95, -5.09)	rs140872668 <sup>b</sup>	2.1x10 <sup>-16</sup>	3.2x10 <sup>-4</sup>	6.00%		
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	2.0x10 <sup>-17</sup>	2.9x10 <sup>-30</sup>	0.44 (-0.01, -0.89)	-0.18 (-0.89, 0.53)	2.41 (2.06, 2.76)	-0.88 (-1.27, -0.48)	-0.39 (-0.68, -0.10)	-1.63 (-2.06, -1.20)	rs116018922 <sup>b</sup>	4.5x10 <sup>-60</sup>	1.0	17.10%		
MFI of HLA-DR in cDC3	-	-	-	-	-	-	-	-	rs114176373 <sup>c</sup>	2.8 x 10 <sup>-12</sup>	3.2x10 <sup>-3</sup>	6.20%		
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	-	-	-	-	-	-	-	-	rs28383322	5.4 x 10 <sup>-14</sup>	2.0x10 <sup>-3</sup>	5.70%		
HLA-DRβ1 position 67														
Omnibus test		Beta (95% CI) of amino-acid substitutions												
P-value	Val. P-val	Phe (F=14.7%)	Ile (F=42.1%)	Leu (F=43.2%)										
MFI of HLA-DR in cDC1	-	-	-	-	-	-	-	-	rs2760994	5.1 x 10 <sup>-39</sup>	0.049	13.80%		
MFI of HLA-DR in pDC	-	-	-	-	-	-	-	-	rs114973966	2.8 x 10 <sup>-88</sup>	1.0	28.50%		
MFI of CD86 in pDC	-	-	-	-	-	-	-	-	rs140872668 <sup>b</sup>	2.1x10 <sup>-16</sup>	3.2x10 <sup>-4</sup>	6.00%		
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	-	-	-	-	-	-	-	-	rs116018922 <sup>b</sup>	4.5x10 <sup>-60</sup>	1.0	17.10%		
MFI of HLA-DR in cDC3	-	-	-	-	-	-	-	-	rs114176373 <sup>c</sup>	2.8 x 10 <sup>-12</sup>	3.2x10 <sup>-3</sup>	6.20%		
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	5.4x10 <sup>-14</sup>	8.7x10 <sup>-6</sup>	0.18 (0.10, 0.27)	-0.23 (-0.28, -0.17)	0.14 (0.08, 0.19)				rs28383322	5.4 x 10 <sup>-14</sup>	2.0x10 <sup>-3</sup>	5.70%		
HLA-DRβ1 position 74														
Omnibus test		Beta (95% CI) of amino-acid substitutions												
P-value	Val. P-val	Arg (F=11.2%)	Ala (F=67.7%)	Gln (F=12.1%)	Leu (F=2.7%)	Glu (F=6.4%)								
MFI of HLA-DR in cDC1	-	-	-	-	-	-	-	-	rs2760994	5.1 x 10 <sup>-39</sup>	0.049	13.80%		
MFI of HLA-DR in pDC	-	-	-	-	-	-	-	-	rs114973966	2.8 x 10 <sup>-88</sup>	1.0	28.50%		
MFI of CD86 in pDC	-	-	-	-	-	-	-	-	rs140872668 <sup>b</sup>	2.1x10 <sup>-16</sup>	3.2x10 <sup>-4</sup>	6.00%		
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	-	-	-	-	-	-	-	-	rs116018922 <sup>b</sup>	4.5x10 <sup>-60</sup>	1.0	17.10%		
MFI of HLA-DR in cDC3	3.9x10 <sup>-13</sup>	1.1x10 <sup>-7</sup>	5.29 (2.48, 8.10)	-1.06 (-3.00, 0.89)	-7.40 (-10.13, -4.68)	15.56 (10.06, 21.07)	1.25 (-2.53, 5.03)				rs114176373 <sup>c</sup>	2.8 x 10 <sup>-12</sup>	3.2x10 <sup>-3</sup>	6.20%
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	-	-	-	-	-	-	-	-	rs28383322	5.4 x 10 <sup>-14</sup>	2.0x10 <sup>-3</sup>	5.70%		
HLA-B position 97, conditioning on HLA-DRβ1 position 13														
Omnibus test		Beta (95% CI) of amino-acid substitutions												
P-value	Val. P-val	Asn (F=3.35%)	Arg (F=53.5%)	Ser (F=24.3%)	Thr (F=12.7%)	Val (F=2.0%)	Trp (F=4.1%)							
MFI of HLA-DR in cDC1	-	-	-	-	-	-	-	-	rs2760994	5.1 x 10 <sup>-39</sup>	0.049	13.80%		
MFI of HLA-DR in pDC	3.8x10 <sup>-17</sup>	1.7x10 <sup>-12</sup>	-17.4 (-92.9, 58.2)	31.5 (3.1, 59.9)	-47.6 (-82.1, -13.1)	22.4 (-19.2, 63.9)	-71.1 (-172.7, 30.5)	-11.7 (-83.0, 59.5)	rs114973966	2.8 x 10 <sup>-88</sup>	1.0	28.50%		
MFI of CD86 in pDC	-	-	-	-	-	-	-	-	rs140872668 <sup>b</sup>	2.1x10 <sup>-16</sup>	3.2x10 <sup>-4</sup>	6.00%		
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	-	-	-	-	-	-	-	-	rs116018922 <sup>b</sup>	4.5x10 <sup>-60</sup>	1.0	17.10%		
MFI of HLA-DR in cDC3	-	-	-	-	-	-	-	-	rs114176373 <sup>c</sup>	2.8 x 10 <sup>-12</sup>	3.2x10 <sup>-3</sup>	6.20%		
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	-	-	-	-	-	-	-	-	rs28383322	5.4 x 10 <sup>-14</sup>	2.0x10 <sup>-3</sup>	5.70%		
HLA-B pos 194, conditioning on HLA-DRβ1 pos 13 and HLA-B pos 97														
Omnibus test		Beta (95% CI) of amino-acid substitutions												
P-value	Val. P-val	Ile (F=83.2%)	Val (F=6.9%)	Indel (F=0.1%)										
MFI of HLA-DR in cDC1	-	-	-	-	-	-	-	-	rs2760994	5.1 x 10 <sup>-39</sup>	0.049	13.80%		
MFI of HLA-DR in pDC	1.3x10 <sup>-18</sup>	1.2x10 <sup>-9</sup>	-30.9 (-73.2, 11.3)	30.9 (-11.3, 73.2)	-167.4 (-786.5, 451.7)				rs114973966	2.8 x 10 <sup>-88</sup>	1.0	28.50%		
MFI of CD86 in pDC	-	-	-	-	-	-	-	-	rs140872668 <sup>b</sup>	2.1x10 <sup>-16</sup>	3.2x10 <sup>-4</sup>	6.00%		
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	-	-	-	-	-	-	-	-	rs116018922 <sup>b</sup>	4.5x10 <sup>-60</sup>	1.0	17.10%		
MFI of HLA-DR in cDC3	-	-	-	-	-	-	-	-	rs114176373 <sup>c</sup>	2.8 x 10 <sup>-12</sup>	3.2x10 <sup>-3</sup>	6.20%		
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	-	-	-	-	-	-	-	-	rs28383322	5.4 x 10 <sup>-14</sup>	2.0x10 <sup>-3</sup>	5.70%		

**Table 13 :** Significant associations of HLA amino-acid positions with immunophenotypes in the MI cohort

<sup>a</sup> Best SNP and P-value are obtained with a linear model adjusting for PC1 and PC2, for comparison purposes: the linear mixed model used in main analyses is not supported for the omnibus test. They may thus be slightly different from best SNP and P-values in Table 10.

<sup>b</sup>This SNP is in LD with rs114973966

<sup>c</sup>This SNP is in LD with rs143655145

Immunophenotype	Risk HLA allele	Conditioning on...	P-value	Validation P-value	Beta (95% CI)	Best SNP <sup>a</sup>	Best SNP P-value <sup>a</sup>	Best SNP P-value conditioning on HLA allele
MFI of HLA-DR in cDC1	HLA-DQA1*05:01	-	1.4x10 <sup>-24</sup>	1.2x10 <sup>-12</sup>	0.11 (0.09 - 0.13)	rs2760994	5.1 x 10 <sup>-39</sup>	3.2x10 <sup>-22</sup>
	HLA-DQBI*05:01	HLA-DQA1*05:01	2.6x10 <sup>-9</sup>	2.7x10 <sup>-3</sup>	-0.08 (-0.11 - -0.06)			
	HLA-DQA1*03:01	-	1.6x10 <sup>-56</sup>	1.9x10 <sup>-28</sup>	336.6 (297.6 - 375.6)	rs114973966	2.8 x 10 <sup>-58</sup>	2.5x10 <sup>-5</sup>
MFI of HLA-DR in pDC	HLA-DQA1*02:01 / HLA-DRBI*07:01	HLA-DQA1*03:01	2.9x10 <sup>-28</sup>	2.4x10 <sup>-12</sup>	-253.8 (-297.3 - -210.2)			
	HLA-DQA1*01:01	HLA-DQA1*03:01, HLA-DQA1*02:01	4.4x10 <sup>-10</sup>	3.1x10 <sup>-4</sup>	-130.9 (-171.7 - -90.2)			
	HLA-DQA1*03:01	-	4.1x10 <sup>-18</sup>	8.2x10 <sup>-8</sup>	11.0 (8.6 - 13.4)	rs140872668 <sup>b</sup>	2.1x10 <sup>-19</sup>	4.4x10 <sup>-5</sup>
MFI of CD86 in pDC	HLA-DQA1*02:01 / HLA-DRBI*07:01	HLA-DQA1*03:01	2.2x10 <sup>-5</sup>	9.4x10 <sup>-3</sup>	-6.2 (-9.1 - -3.4)			
	HLA-DQA1*03:01	-	2.9x10 <sup>-36</sup>	5.0x10 <sup>-24</sup>	2.3 (2.0 - 2.6)	rs116018922 <sup>b</sup>	4.5x10 <sup>-40</sup>	1.6x10 <sup>-3</sup>
MFI of HLA-DR in CD14 <sup>hi</sup> monocytes	HLA-DQA1*02:01 / HLA-DRBI*07:01	HLA-DQA1*03:01	5.8x10 <sup>-10</sup>	1.4x10 <sup>-7</sup>	-1.3 (-1.7 - -0.9)			
	HLA-DQBI*02:02	-	5.6x10 <sup>-8</sup>	1.8x10 <sup>-4</sup>	15.65 (10.04 - 21.25)	rs114176373 <sup>c</sup>	5.4 x 10 <sup>-14</sup>	1.8x10 <sup>-13</sup>
Number of HLA-DR+ CD56 <sup>hi</sup> NK cells	HLA-DQBI*04:02	-	2.7x10 <sup>-7</sup>	1.6x10 <sup>-1</sup>	-0.26 (-0.36 - -0.17)	rs28383322	2.8 x 10 <sup>-12</sup>	1.8x10 <sup>-10</sup>

**Table 14 : Associations of HLA classical alleles with immunophenotypes in the MI cohort**

<sup>a</sup>Best SNP and P-value are obtained with a linear model adjusting for PC1 and PC2, for comparison purposes: the linear mixed model used in main analyses is not supported for the omnibus test. They may thus be slightly different from best SNP and P-values in **Table 10**.

<sup>b</sup>This SNP is in LD with rs114973966

<sup>c</sup>This SNP is in LD with rs143655145

## 2.6. Novel *trans*-acting genetic associations with immune cell parameters

We detected five novel associations that involve SNPs acting in *trans* on immunophenotypes (**Figure 44**). These SNPs include variants that are genetically independent from the genes encoding immune cell markers with which they are associated, or that are associated with immune cell counts.

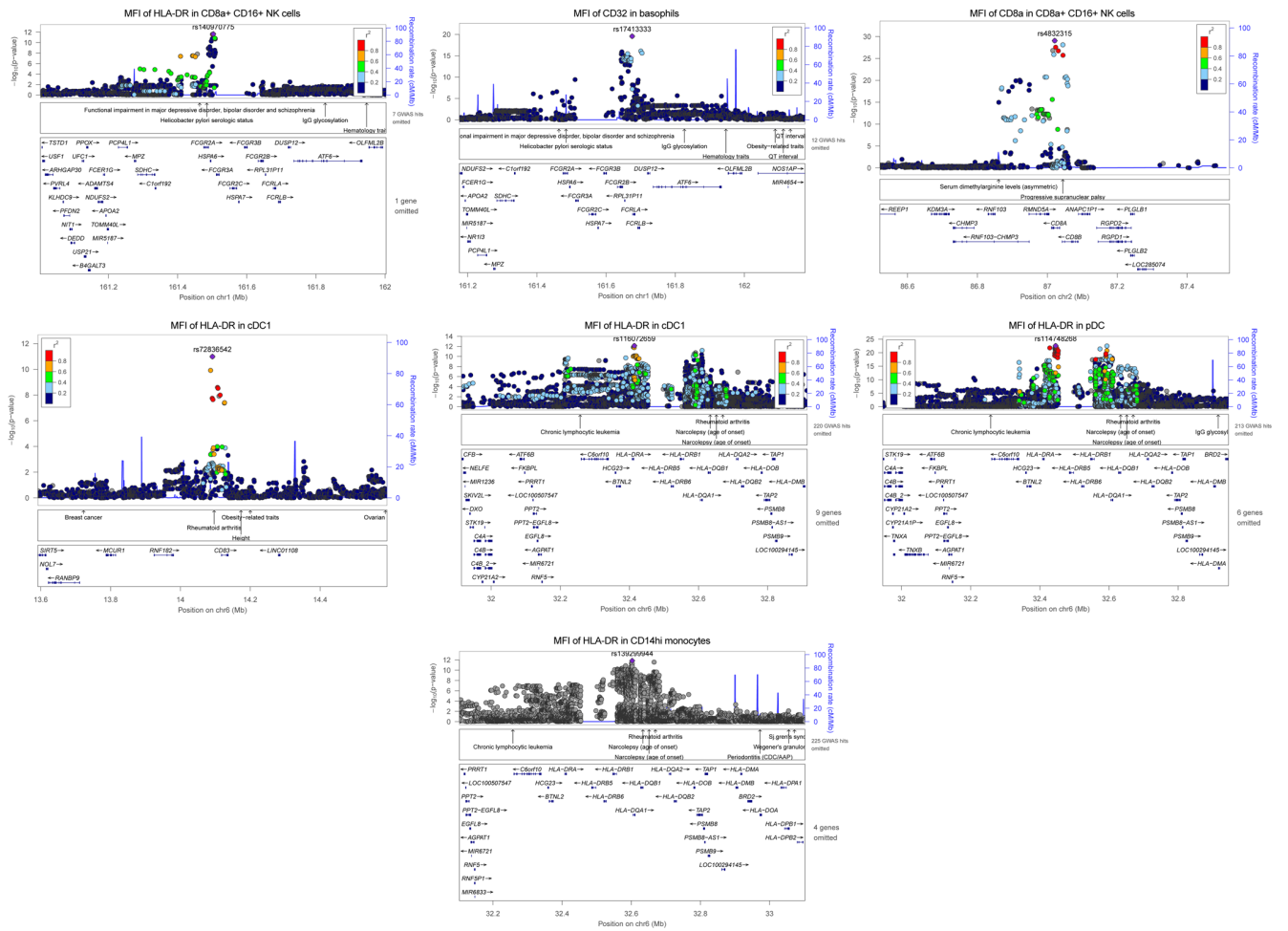
Variants close to the *SIPRI* gene were associated with the MFI of CD69 in CD16<sup>hi</sup> NK cells ( $P = 3.8 \times 10^{-30}$ ). CD69 is known to downregulate cell-surface expression of the sphingosine-1-phosphate receptor-1 (S1P1) on lymphocytes, a mechanism that elicits egress from the thymus and secondary lymphoid organs (Garris et al., 2014).

The second association signal lies in an intron of the *ACOXL* gene, close to *BCL2L11*, and is associated with the absolute count of CD56<sup>hi</sup> NK cells ( $P = 9.1 \times 10^{-20}$ ). *BCL2L11* (also known as BIM) is an important regulator of lymphocyte apoptosis (Pellegrini et al., 2004) and is associated with chronic lymphocytic leukemia and total blood cell number (van der Harst et al., 2012).

The third *trans*-acting association correlates genetic variants close to the *ACTL9* gene with the ratio of CD16 MFI in CD16<sup>hi</sup> and CD56<sup>hi</sup> NK cells ( $P = 2.0 \times 10^{-24}$ ). While little is known about the function of the *ACTL9* gene, the most strongly associated SNPs at the locus were recently shown to be associated with atopic dermatitis (**Table 11**) (Paternoster et al., 2015), thereby suggesting a possible involvement of NK cells in this pathology (von Bubnoff et al., 2010).

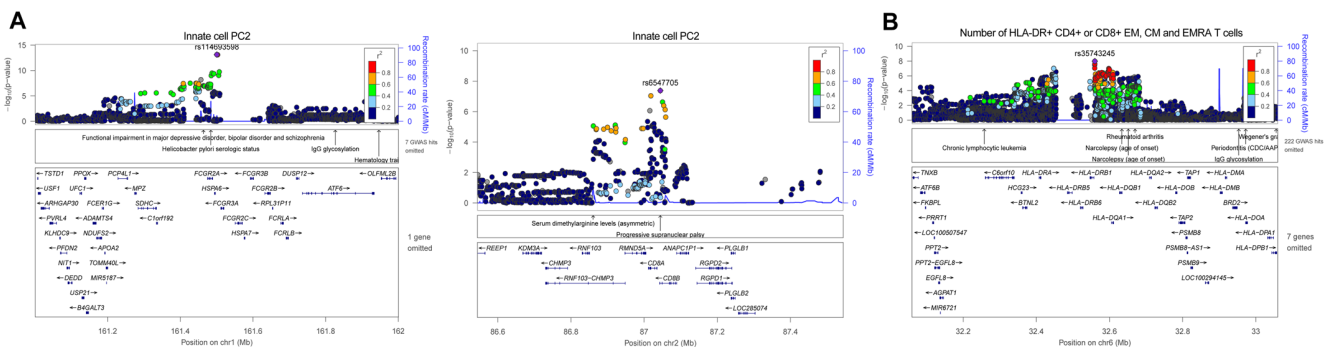
The fourth signal locates in the *TMEM8A* gene and is associated with the MFI of CCR7 in CD4<sup>+</sup> and CD8b<sup>+</sup> naïve T cells ( $P = 2.4 \times 10^{-20}$ ). *TMEM8A* is expressed on the surface of resting T cells and is down-regulated after cell activation (Motohashi et al., 2000), suggesting a possible functional association and/or co-regulation with CCR7.

Finally, conditional GWAS identified a fifth *trans*-acting association, between HLA-DR MFI in cDC1 and the *CD83* gene ( $P = 1.0 \times 10^{-11}$ , **Figure 45** and **Table 12**). These results suggest that CD83, an early activation marker of human DCs, upregulates HLA-DR expression in activated dendritic cells.



**Figure 45 : Local association signals for the 7 genome-wide significant hits identified by conditional GWAS of the 14 immunophenotypes showing strong genetic association in the *MI* cohort.**

Each point is a SNP, whose color represents its level of linkage disequilibrium ( $r^2$ ) with the best hit (in purple). Blue lines indicate local recombination rates. These association signals were identified when conditioning on main signals (Materials & Methods).



**Figure 46 : Local association signals detected by multi-trait GWAS of immunophenotypes measured in the *MI* cohort.**

(A) Local association signals at loci influencing principal components of a PCA of all innate cell immunophenotypes. No suggestive signal ( $P < 5 \times 10^{-8}$ ) was observed for PCs of a PCA of adaptive cell immunophenotypes.

(B) Local association signal with the absolute numbers of HLA-DR<sup>+</sup> CM, EM and EMRA T cells, either CD4<sup>+</sup> or CD8<sup>+</sup>. The six local corresponding immunophenotypes were analyzed altogether using a multivariate GWAS (Materials and Methods)

## 2.7. Genetic control of mRNA expression explains association with protein levels of immune cell markers

While three of the 14 genome-wide significant associations are probably explained by amino acid-altering variants in surrounding genes (H/R polymorphism in *FCGR2A*, position 13 in *HLA-DR*; **Tables 11** and **13**), the remaining signals do not present obvious candidate causal variants. To better understand these associations, we tested if corresponding SNPs were associated with mRNA levels of nearby genes (*i.e.*, gene expression QTL, eQTL).

We quantified the expression of immunity-related genes located close to each associated locus, using mRNA extracted from whole blood of all 1,000 *Milieu Intérieur* subjects (Materials & Methods).

Three of the pQTLs that we identified were strongly associated with the mRNA levels of a surrounding immunity-related gene (**Tables 11** and **12**). The SNPs controlling CD62L MFI in eosinophils, CD32 MFI in basophils and HLA-DR MFI in cDC1 were strongly associated with *SELL* mRNA levels ( $P = 1.9 \times 10^{-21}$ ; **Figure 43D**), *FCGR2B* mRNA levels ( $P = 4.1 \times 10^{-22}$ ; **Table 11**) and *CD83* mRNA levels ( $P = 3.2 \times 10^{-22}$ ; **Table 12**), respectively.

Together, our analyses indicate that several genetic variants associated with immunophenotypes affect the expression of genes encoding markers of immune cells in whole blood.

This suggests that the intersection of eQTL and pQTL mapping in different immune cell compartments can greatly improve our knowledge of the genetic factors controlling human inter-individual variation in immune cell parameters.

## 2.8. The natural variance of innate immune cell parameters is driven by genetic factors

A large proportion of both MFI and cell number immunophenotypes that presented a genome-wide association (34/42, 81%) were measured in innate immune cells, including granulocytes, monocytes, NK and dendritic cells (**Table 10**), while 47% of all immunophenotypes were measured in innate cells (**Table 9**). Also, of the adaptive cell immunophenotypes showing genetic associations, 3 of the 8 hits (38%) were related to naïve T or B cells, while naïve adaptive cell parameters represented <10% of all adaptive cell measurements. These



observations suggested a stronger effect of genetic variants on innate and naïve adaptive cell subpopulations, as compared to differentiated or experienced adaptive immune cells.

In accordance with these observations, the presence of HLA-DR molecules, which was assessed at the surface of both innate and adaptive immune cells, was strongly associated with *HLA-DR* genetic variation in monocytes, NK and dendritic cells (**Table 10**), while it showed no significant association in memory CD4<sup>+</sup> or CD8<sup>+</sup> T cells (including CM, EM and EMRA T cells;  $P > 1.0 \times 10^{-6}$ ). Because HLA-DR<sup>+</sup> memory T cell numbers were modestly correlated ( $r^2 \sim 0.3-0.4$ , **Figure 36**), we assumed they were at least partly controlled by the same genetic factors, which we further examined using a multivariate GWAS (Materials & Methods). This refined approach indeed detected a suggestive genetic association close the *HLA-DRB1* gene (rs35743245,  $P = 1.0 \times 10^{-8}$ ;  $r^2 = 0.92$  with rs114973966) in strong linkage disequilibrium with that detected in pDCs, monocytes and NK cells ( $r^2 = 0.92$ ; **Figure 46**).

This finding provides a proof-of-concept that immunophenotypes in both innate and adaptive cells can be controlled by the same genetic factors, but their weaker effects in experienced adaptive cells may require greater power to be reliably detected.

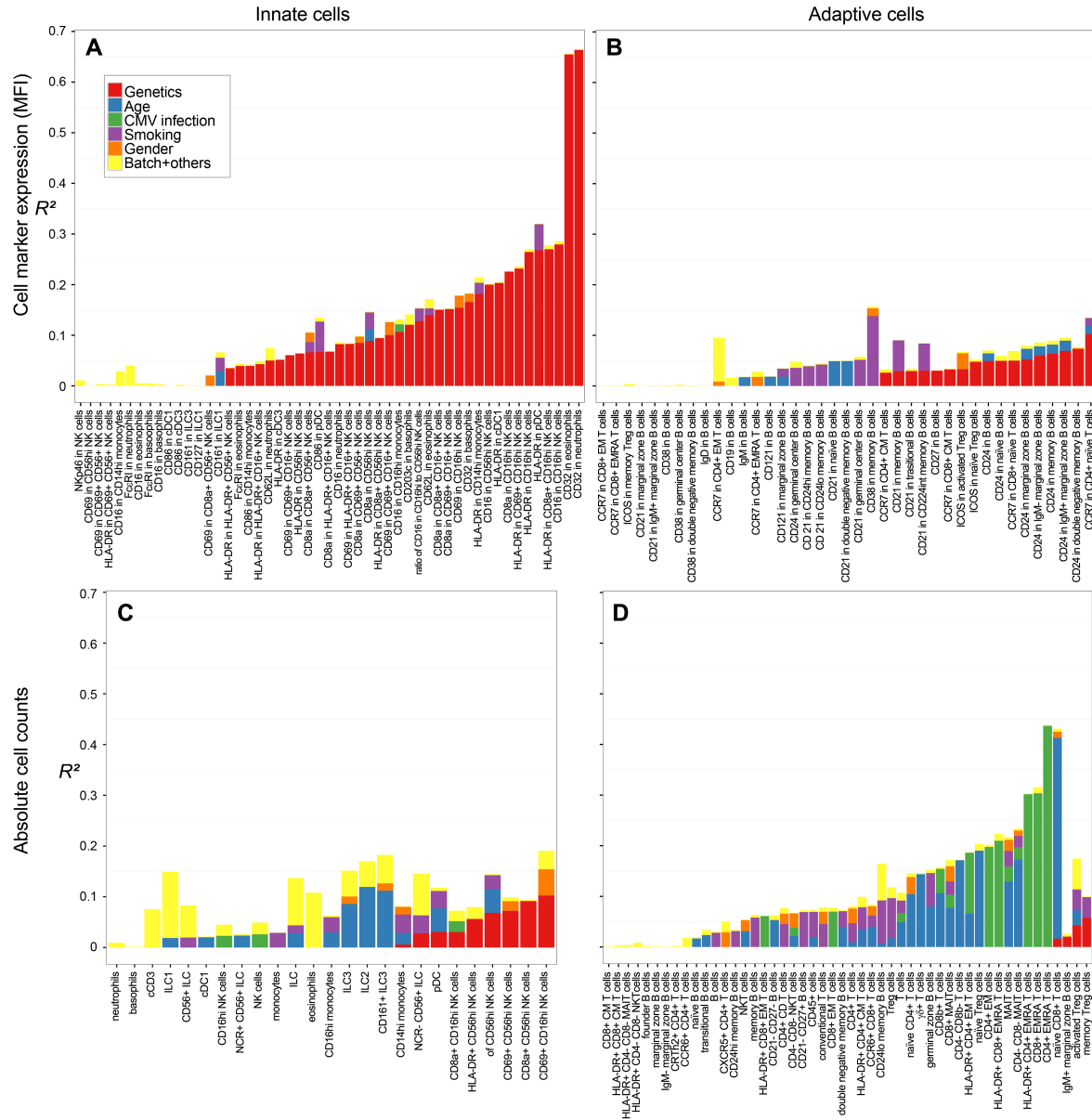
Following these observations, we next systematically quantified the impact of genetic and non-genetic factors on innate and adaptive cells, by establishing for each immunophenotype a linear regression model that included all significant variables (**Figures 29, 37, 40 and 43**), and by estimating their respective contribution to the total variance (Materials & Methods).

We indeed observed that a larger proportion of the variance of innate cell immunophenotypes was explained by genetic factors (**Figures 47A and 47C**), as compared to adaptive cell immunophenotypes (**Figures 47B and 47D**). Inversely, the variance in adaptive cell numbers was dominated by non-genetic factors such as age and CMV serostatus (**Figure 47D**).

To formally test if these differences were significant, we used a generalized least squares model that accounts for correlations among immunophenotypes (Materials & Methods). Conclusively, the increase in the proportion of variance explained by genetics for innate cell measurements was strongly significant ( $9.4\% \pm 0.2\%$ ;  $P = 1.2 \times 10^{-12}$ ), a result that was further supported by a non-parametric test (Wilcoxon signed rank test  $P = 2.4 \times 10^{-5}$ ).

Because CD32a is a receptor for the Fc-domain of IgG antibodies, we were concerned that the strong association detected between the CD32a R/H amino acid-altering variant (rs1801274, **Table 10**) and CD32 MFI in eosinophils and neutrophils was due to genotype-dependent non-specific interactions with the antibody used for detecting its expression. For this reason, we

repeated our analyses after excluding the two associations observed between the CD32a R/H variant and CD32 expression.



**Figure 47 : Proportion of variance of innate and adaptive cell parameters explained by non-genetic and genetic factors.**

Flow cytometry measurements were separated into (A, B) expression levels of cell surface markers (mean fluorescence intensity, MFI) and (C, D) absolute counts of circulating immune cells. The total variance  $R^2$  of each innate (A, C) or adaptive (B, D) cell immunophenotype was decomposed into proportions explained by non-genetic factors (*i.e.*, age, gender, CMV infection and smoking; **Figure 40A**) and genetic factors (*i.e.*, GWAS hits, **Tables 10 and 12**)

Conclusively, the larger proportion of variance explained by genetics for innate cell measurements was still strongly significant ( $P = 1.8 \times 10^{-12}$ ;  $P_{Wilcoxon} = 5.7 \times 10^{-5}$ ). In sum, we

find that genetic factors account for a substantial fraction of human variation in immune cell parameters, and their influence is stronger in innate immune cells, as compared to adaptive immune cell phenotypes.

## 2.9. Design and preliminary results from first replicative study

In order to validate some of the findings from the *Milieu Intérieur* collection, an independent collection has been organized in partnership with Genentech company (Cancer Immunology department), where I spent a portion of my PhD training. Due to differences between the two sites in term of donor recruitment, instrument facilities and reagent availability, some modifications to the protocol were performed. While requiring some additional methods development, I saw this as a true opportunity to pressure-test the robustness of our findings.

### 2.9.1. Donor collection

Using the Genentech genotyped and phenotyped (gGAP) registry, a first evaluation was performed by a geneticist of our group to determine the minor allele frequency (MAF) distribution for the genes for which the most significant associations were found in the *MI* analysis. The donors presenting those SNPs (i.e. homozygote for the minor allele) were prioritized as well as donors presenting the corresponding haplotypes (heterozygotes and homozygotes for the major allele). This selection aims to assure a statistically sufficient number of donors presenting a specific MAF to be able to perform correlations with phenotypic immune parameters.

Thus far 81 donors have been recruited with 65% of men and a median of age of 39 years old. The range was between 25 to 50 years old with only one active and 4 former smokers (**Table 15**).

	Total number	Ancestry	Age (years old)		Smoking Status
			Median	Range	
Female	31	Caucasian	38.6	24.9 - 46	1 former
Male	50		39.1	29 - 50	1 active + 3 formers

**Table 15 : Demographic parameters of the replicative collection**

We have indicated that a primary donor selection has been performed in term of genetic background. Nevertheless, the analysis of the allele distribution frequency for the genes of interest in this preliminary collection was concordant with the ones observed in the *MI* collection. This indicates the absence of a massive genetic bias within the donors included which allows a broad genotype/phenotype comparison (**Table 16** and data not shown).

Example of selected SNPs	Homo minor (Donor nb and allele frequency)	Heterozygote (Donor nb and allele frequency)	Homo major (Donor nb and allele frequency)
Rs875063 (BCL2L11)	N=4 (6,1%)	N=25 (38,5%)	N=36 (55,4%)
Rs6693121 (S1PR1)	N=15 (23,1%)	N=25 (38,5%)	N=25 (38,5%)

**Table 16:** Examples of allele frequency distribution within the replicative collection.

### 2.9.2. Immune cell type assesment

This « artificial » SNPs enrichment allowed also focusing on the parameters measured by flow cytometry to the ones found associated to those genetic polymorphisms.

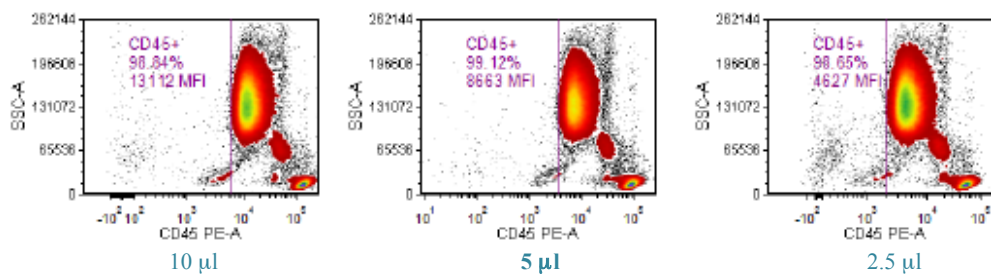
As described in the previous section, using the *MI* cohort dataset, the majority of genetic associations were found within the innate cells compartment whereas the variance observed within the adaptive cells compartment was preferentially associated to environmental factors. Based on this *MI* analysis, three « innate » 8-colour panels (out of the 10 used for the *MI* collection; **Figures 56-65** in « Materials & Methods » and **Table 8**) were selected to assess quantitative and qualitative immune phenotypes (**Figure 49**). Those three selected panels aimed to measure the (i) general immune cell populations or « lineage » panel, for the assessment of the major innate and adaptive immune cells ; (ii) the natural killer cell phenotypes or « NK » panel ; (iii) the polymorphonuclear cell population phenotypes or « PMN » panel, allowing the analysis of neutrophils, eosinophils and basophils populations.

In order to validate the standardized flow cytometry analysis pipeline developed for the *MI* collection we applied the same strategy except for some modifications due to local logistic differences. I will indicate those changes whenever necessary in the next sections.

### 2.9.2.1. Antibodies

At the pre-collection stage of the *MI* study, a huge effort in term of antibody selection (clone and supplier), staining protocol and titration has been done to guaranty the dataset quality (Hasan et al., 2015).

In collaboration with the clinical cytometry platform from Genentech, we applied the same antibody mix and the same protocol. Nonetheless, a new antibody titration was necessary due to change of individual antibody batch and cytometer instrument (**Figure 48**).



**Figure 48 : Antibody titration for flow cytometry analysis of the replicative study**

Titration of the CD45 PE antibody (Ab) before inclusion in the lineage panel using 10µl, 5µl or 2.5µl per 100µl of blood (from left to right). The facs plots represent the CD45 staining (x axis) against the granularity parameter (SSC) on the y axis allowing the detection of the three main immune circulating populations : PMN ( $CD45^{lo}SSC^{hi}$ ), monocytes ( $CD45^{int}SSC^{int}$ ) and lymphocytes ( $CD45^{hi}SSC^{lo}$ ). The CD45+ frequency and mean fluorescence intensity (MFI) are indicated for each condition (purple).Whereas the overall frequency do not change, the MFI decrease in parallel to Ab quantity. The condition 5µl has been finally selected as the CD45 staining within PMN population was sufficiently bright (half distance between  $10^3$  to  $10^4$  log scale) and not too saturated within lymphocyte population (MFI at  $10^5$  log scale maximum). In the *MI* collection 10µl per mix of this antibody was used.

In addition, due to a customized coupling of the anti-CD32 (FcγRIIa) to the PE-Cy<sup>7</sup> fluorochrome (specially performed for the *MI* study by the company Becton Dickinson), this antibody (included in the PMN panel) was not anymore available for this study. As a consequence, we replaced the original reagent by an anti-CD32 from the same clone but coupled to allophycocyanin (APC) from the same supplier and commercially available. In parallel, the anti-CD203c APC (Miltenyi Biotech) originally used in this panel was replaced by the corresponding anti-CD203c PE-Vio770 (same clone and supplier).

We have titrated those antibodies and validated this new mix in term of population identification and frequency (**Figure 49** and data not shown).

Lineage panel				NK cells panel				PMN cells panel			
Laser	Fluorochrome	Supplier	Marker	Laser	Fluorochrome	Supplier	Marker	Laser	Fluorochrome	Supplier	Marker
Violet	eF450	eBiosc.	CD3	Violet	eF450 + Vioblue	eBioSc.+ Miltenyi	CD3 + CD14	Violet	eF450	BD	CD62L
	525/50 BP	BD	CD14		525/50 BP	eBiosc	FVD eF506		525/50 BP	eBiosc.	FVD eF506
Blue	FITC	BD	CD56	Blue	FITC	BD	CD69	Blue	FITC	eBiosc.	FCeRIa
	PE	BD	CD45		PE	BD	NKp46		PE	BD	CDw125
	PerCP-Cy5.5	BD	CD16		PerCP-Cy5.5	BD	HLA-DR		PerCP-Cy5.5	BD	CD16
	PE-Cy7	eBiosc.	CD8b		PE-Cy7	BD	CD8a		PE-Vio770	Miltenyi	CD203c
Red	APC	BD	CD19	Red	APC	BD	CD16	Red	APC	BD	CD32
	APC-H7	BD	CD4		APC-eF780	eBiosc.	CD56				

**Figure 49: Modified flow cytometry panels used for the replicative study.**

From the left to the right the lineage, Natural killer (NK) and polymorphonuclear (PMN) cell panels are described. The tables indicate for each panel the markers used with the fluorochrome and supplier associated. The two human antibodies (anti-CD32 and anti-CD203c) modified in the PMN panel are highlighted in green.

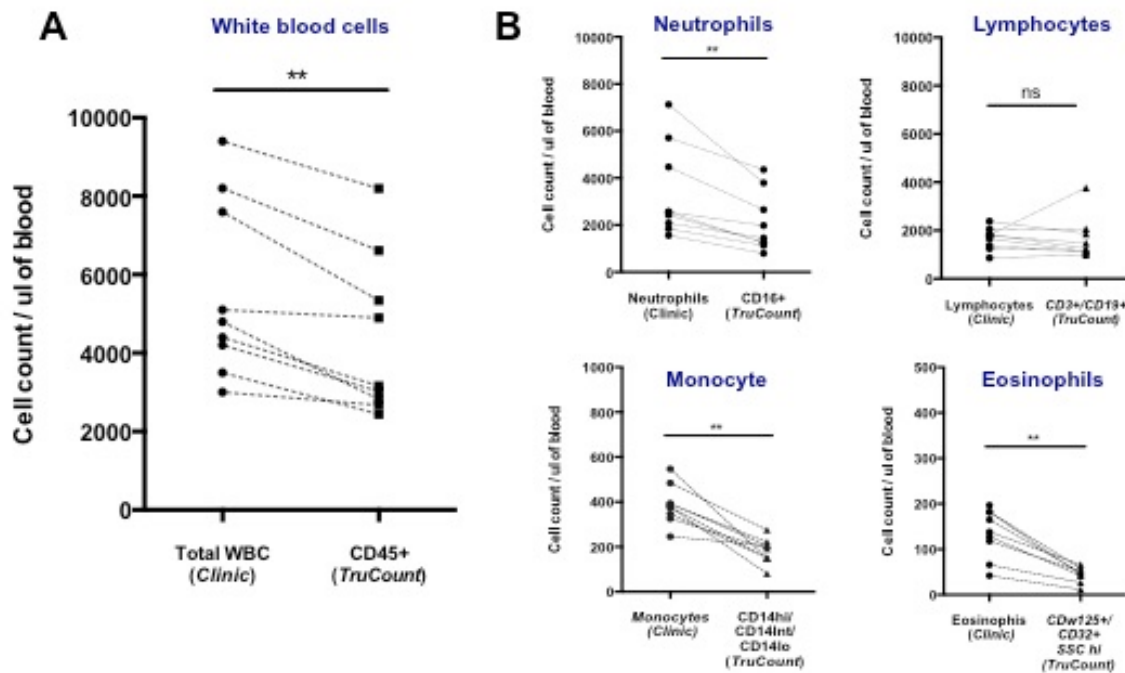
### 2.9.2.2. Instrument and cell count assessment

The assessment of immunophenotypes for the *Milieu Intérieur* cohort has been performed on two *MacsQuant* instruments (Miltenyi Biotech) specifically selected, calibrated and dedicated to the collection (Hasan et al., 2015; Materials & Methods).

In order to replicate our previous findings, we organized this replicative study in collaboration with the clinical flow cytometry facility at Genentech for the immune cell phenotyping. Our interest for this platform was driven by their capability to offer daily-calibrated instruments. Despite an impressive armaterium of flow cytometers proposed by this platform, the *MacsQuant* instrument was not one of them. Nonetheless, I considered this as a real opportunity to test the robustness of our previous phenotype/genotype associations.

The cytometer *BDFacsCantoII* (Becton Dickinson) has been selected as the most similar to the *MacsQuant* in term of fluorescent parameter detection and for its clinical grade fluidic system. However, the laser and optic system remain different. Other major difference between those two instruments is the absence of a direct cell count measurement for the *BDFacsCantoII*.

In order to optimize the immune cell count, we compared for 9 donors the *TruCount* bead method (Becton Dickinson) associated to the Lineage panel with a clinical grade *white blood cells count (WBC) and differential* performed in parallel by the clinical grade Quest Diagnostics laboratory (Figure 50).



**Figure 50 : Method comparison for cell count assessment.**

The cell count measurement (expressed per microliter of blood) has been assessed for 9 donors using a clinical grade *white blood cells count (WBC) and differential* and the *TruCount beads* included in the Lineage panel tube and acquired in the *BDFacsCantoll*.

- (A) The leukocyte counts given by the *WBC and differential* (left) or after calculation using the frequency of CD45<sup>+</sup> normalized with the *TruCount beads* (CD45<sup>+</sup> *TruCount*) included in the same tube (right) are plotted.
- (B) Same analysis for different population subsets: neutrophil counts (*WBC and differential*) versus CD16<sup>hi</sup> frequency normalized with the *TruCount beads* (top left); lymphocyte counts versus CD3<sup>+</sup> plus CD19<sup>+</sup> *TruCount* (top right); monocyte counts versus of CD14<sup>hi</sup> (conventional) plus CD14<sup>lo</sup> (intermediate) plus CD14<sup>hi</sup> (inflammatory) *TruCount* (bottom left); eosinophil counts versus of CDw125<sup>+</sup>CD132<sup>+</sup>SSC<sup>hi</sup> *TruCount* (bottom right). Non-parametric Wilcoxon test \*p<0.05.

The CD45<sup>+</sup> *TruCount* calculation was performed as follow:

$(N_{CD45^+}/N_{Measured\ Beads}) \times (48650 \text{ (TruCount Beads Lot 15341= 48650 target beads)} / 50\mu\text{l of blood})$ .

Using those two methods simultaneously, we constantly observed a significant lower hematopoietic CD45<sup>+</sup> *TruCount* compared to the clinical leucocyte count (*WBC*) with 4 to 40% decrease observed across donors (average of 22.7%) (**Figure 50A**). This difference highlight probably technical differences between the two methods where one relies on light and electronic captors measuring the size and granularity to identify the populations (*WBC and differential*) whereas the other technology detects fluorescent parameters (*TruCount*).

To validate this hypothesis, we next extended this analysis by the comparison of the cell count by those two approaches on population subsets presenting different abundances. We performed a comparison on the populations provided by the *WBC and differential*

(neutrophils, eosinophils, basophils, monocytes and lymphocytes) to the assigned equivalent populations in our gating strategies for this panel (**Figure 60**), analyzed after fluorescent immunostaining coupled with the *TruCount* beads.

The calculated population counts using the *TruCount* beads were obtained using the following formula: (% population of interest gated in CD45<sup>+</sup>) X (CD45<sup>+</sup> *TruCount*).

We confirmed significant higher counts from the clinical-grade assessment compared to our flow approach for neutrophils (p=0.0039), monocytes (p=0.0039), eosinophils (p=0.0078), but not for lymphocytes for which the counts measured in both methods were very similar (n.s., p= 0.3008). (**Figure 50B**). These results indicated that the lower count observed using the *TruCount* was independent of the population frequency as the neutrophils (the most abundant) and the eosinophils (the less abundant) presented the same pattern whereas the lymphocytes (the second most frequent population) were not impacted by the method used. In addition, those results suggested that the technical differences between both methods did not explain entirely these discrepancies.

Due to those observations and in order to reduce pre-analytical bias potentially introduced by the addition of *TruCount* beads within our antibody mix, we decided to request a systematic *WBC and differential* in parallel to our sample process by flow cytometry. Those clinical data will be used as hematopoietic cell reference count to normalize the counts of all the subsets extracted from the three panels for the entire collection.

### 2.9.2.3. Comparison between *Milieu Intérieur* and replicative collection

The same gating strategies as previously described were applied to analyze the NK cells, lineage and PMN cell panels (**Figures 59, 60 and 62** in « Materials & Methods » section) for the 81 donors included. Those analysis provided the frequencies and mean fluorescence intensity (MFI) for the same immunophenotypes than in the *MI* cohort analysis (see **Table 9**). As mentioned before, this is an on going analysis and the dataset still need to be curated notably by controlling and correcting for the potential batch effects.

In the coming section, I will briefly present some preliminary comparative analysis between both studies to give a first overview of what can be expected from the future dataset.

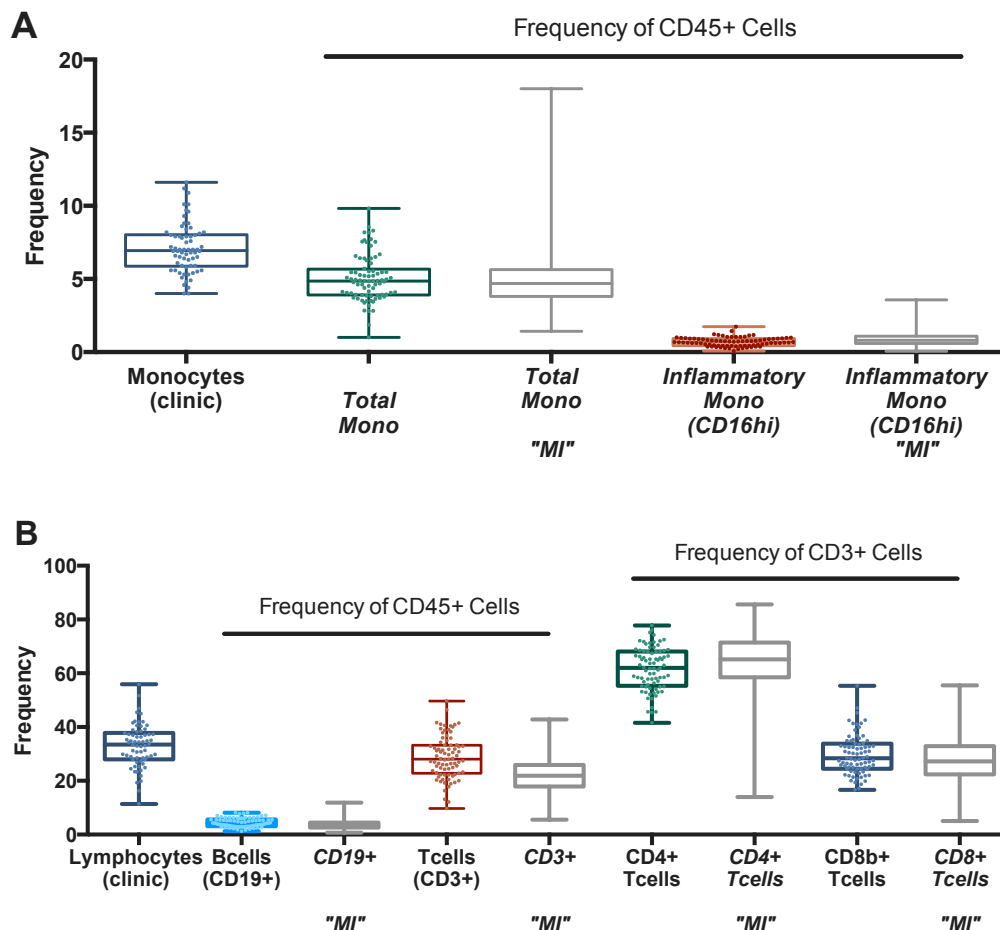
The comparison of the major circulating population frequencies showed similar ranges between the values measures in both collections. As an example, for the circulating monocyte populations, the median of total monocytes (CD14<sup>hi</sup>, CD14<sup>lo</sup> and CD16<sup>hi</sup>) frequency of CD45<sup>+</sup>



was 4.9% (from 1.0 to 9.8 %) for the replicative study compared to a median of 4.7% (range of 1.43 to 18%) for the *MI* study (**Figure 51A**).

This was also true for the less frequent inflammatory monocyte population ( $CD16^{hi}$ ) with a median of 0.69% and 0.81% for the replicative and the *MI* study respectively (**Figure 51A**).

We observed the same pattern for the PMN and NK cells (data not shown).



**Figure 51: Comparison of monocyte and lymphoid population frequency across studies.**

The box plots show the comparison between the frequencies measured within the 81 donors included in the replicative study using the *WBC and differential* (dark blue) or the flow panels (light blue, green and brown) and compared to the frequencies measured for the 1,000 donors from the *Milieu Intérieur* cohort ("*MI*"; grey box plots).

- (A)** Total monocyte frequency (including the  $CD14^{hi}$ ,  $CD14^{int}CD16^{int}$  and  $CD14^{lo}CD16^{hi}$ ) within the  $CD45^{+}$  gate using the Lineage panel. The inflammatory monocyte frequencies ( $CD14^{lo}CD16^{hi}$ ) were indicated the right for both studies.
- (B)** Lymphoid population frequencies frequency (including the  $CD14^{hi}$ ,  $CD14^{int}CD16^{int}$  and  $CD14^{lo}CD16^{hi}$ ) within the  $CD45^{+}$  gate using the Lineage panel. The inflammatory monocyte frequencies ( $CD14^{lo}CD16^{hi}$ ) were indicated the right for both studies.

The same analysis was performed within the lymphoid compartment. We observed similar median frequencies for the B cells ( $CD19^{+}$ ) with 4.2% for the replicative collection vs 3.54%

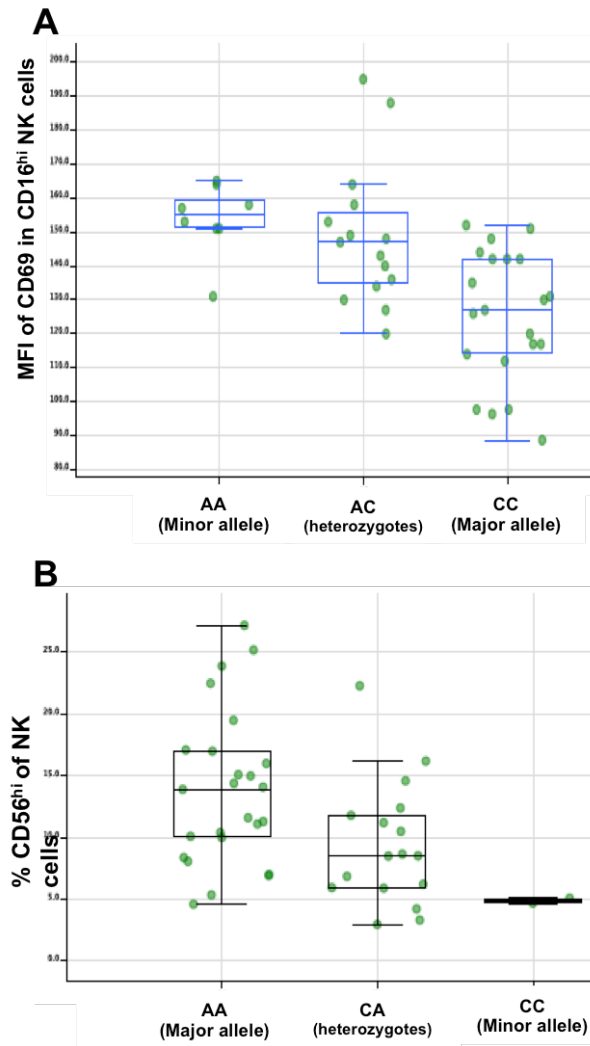
in the *MI* study. For the T cells, the median frequency was slightly increased within the replicative collection than within the *MI* cohort (28.1% versus 21.9% respectively). This was accompanied by a preferential decrease of the ratio  $CD4^+/CD8^+$  within the  $CD3^+$  T cell subset in the replicative study (**Figure 51B**).

As observed for the counts in the precedent section, the frequencies of monocytes were constantly lower if measured by flow cytometry rather than by *WBC and differential* (median of 6.95% compared to 4.9% for the flow approach). Again, this could be probably explained by different technical properties. The flow cytometry approach requires a fluorophore excitation with fluorescence detection and restricts the population measurement to the ones identified by the expression of our selected marker(s). By contrast, the clinical *WBC and differential* measure their size (front scatter) and light refraction and/or electrical impedance properties (side scatter) without staining and wash step.

Nonetheless, the lymphoid cell frequencies, as observed previously for the counts, showed similar results between both methods when we compared the sum of B cells and T cells frequencies to the clinical lymphocytes frequency (32.3% versus 33.5%; **Figure 51B**). This suggest potential biological differences

In order to check the accuracy of our immune parameter measurements and taking advantage of the genetic/phenotype associations found in the *MI* analysis, we compared within the same populations the MFI of specific surface proteins with genetic variants loci.

The preliminary analysis was encouraging with an overall increase of the activation marker CD69 MFI within the CD16hi NK cells subset when the genetic variation of the gene *S1PR1* was present (rs6693121) (**Figure 52A**) or a decrease of the CD56hi NK cell subset frequency alleles in presence of the *BCL2L11* gene variant (rs875063) (**Figure 52B**). Those observations were in concordance with the associations observed within the *MI* collection for rs6693121 (p-value =  $3.8 \times 10^{-30}$ , SE= 7.01) and rs875063 (p-value =  $9.1 \times 10^{-20}$ , SE= -0.95).



**Figure 52 : SNPs controlling circulating immune cell parameters.**

Box plots with the individual donors indicated (green points) representing the (A) MFI of CD69 in the CD16<sup>hi</sup> NK cell subset (measured with the NK cell panel), across the different variants of the gene S1PR1 (minor allele rs6693121); (B) frequency of CD56<sup>hi</sup> NK cells in the NKp46<sup>+</sup> NK cell population (measured with the NK cell panel), across the different variants of the gene BCL2L11 (minor allele rs875063).

### 3. Discussion

Over the last two decades, research in immunology has employed multi-parametric cytometry to enumerate and assess the activation state of human immune cells in healthy and disease conditions. Although such immune cell parameters vary in the general population, the genetic and non-genetic factors that explain this variance remain largely unexplored. Powerful genome-wide association studies have identified genetic factors that contribute to the risk of developing chronic infectious or auto-immune diseases, but little is known about the specific tissues or immune cell compartments where susceptibility alleles are phenotypically expressed. Such precise information is critical in the context of translational research and clinical immune monitoring.

Here we combined standardized flow cytometry with genome-wide DNA genotyping in a demographically-well defined cohort of 1,000 healthy unrelated subjects, to identify non-genetic and genetic factors that modulate human immune cell parameters. We confirm the strong and independent impacts of age and CMV infection on naïve and memory T cell populations, respectively, and provide robust evidence for sex differences in innate and adaptive cell numbers. We also show that active cigarette smoking, and to a lesser extent past smoking, elicit a decline in innate immune cells and MAIT cells and an increase in the levels of activated and memory Treg cells, which may predispose healthy individuals to infection. Furthermore, human genetic variation considerably impacts immune cell parameters in healthy individuals, particularly the cell-surface expression of markers conventionally used to identify leukocyte differentiation or activation. Overall, our analyses strongly suggest that individual non-genetic and genetic features should be taken into consideration when interpreting parameters of circulating white blood cells of patients. For instance, HLA-DR expression on monocytes is routinely measured by flow cytometry in clinics to predict mortality in septic shock and identify patients who should benefit from immunoadjuvant therapies (Venet et al., 2013). We identified a strong effect of HLA-DR $\beta$ 1 coding variation on HLA-DR expression in CD14<sup>hi</sup> monocytes, which suggests that prognostic tools of fatal outcome in sepsis should be tailored to patient's genetic makeup.

The most striking result of our study is the disproportionate number of genetic associations found in innate immune cells. The virtual absence of genetic associations detected in memory T and B cells, even after adjusting for age and CMV infection, could be attributed to their strong dependence on individual past infections. Notably, adaptive immune cells are known to

possess a much longer half-life as compared to myeloid innate cells, in mice and humans (Kolaczowska and Kubes, 2013; Macallan et al., 2005; Vrisekoop et al., 2008; Wen et al., 2013). Additionally, stimulus-induced differentiation and expansions may result in the possible masking of genetic associations for adaptive cell types. In agreement with this conclusion, strong genetic associations in adaptive immune cells were primarily observed for immunophenotypes of naïve adaptive cells.

A previous study of the immune system of 105 pairs of healthy twins concluded that less than 15% of immune cell parameters are controlled by genetic factors (Brodin et al., 2015). At odds with this estimation, we find that more than 36% of measured immunophenotypes are dominated by genetics (**Figure 47**). This discrepancy stems mainly from the fact that this previous study considered only a fraction of innate myeloid and lymphoid populations (Casanova and Abel, 2015), and possibly because of a limited power due to restricted sample size (Martin et al., 1978).

This hypothesis is strongly supported by the preliminary results observed from our replicative study that suggests similar genetic association with innate cells (NK) phenotype changes. Innate immune cells represent more than 70% of circulating white blood cells and are involved in a number of disease conditions (Bot, 2014). Importantly, genetic variants found to modulate innate immune cell parameters, in our study and in previous studies (Orrù et al., 2013; Roederer et al., 2015), have been directly implicated in the aetiology of several autoimmune disorders, such as inflammatory bowel disease, ulcerative colitis and atopic dermatitis. These findings illustrate the value of our approach, which mapped novel genetic associations to specific cell populations and cellular states, providing new insights into disease pathogenesis.

Further evaluations of normal ranges of leukocytes and their key environmental and genetic determinants will improve the utility of cellular immunophenotyping in the development of personalized clinical management strategies, and will facilitate a detailed dissection of immune cell populations implicated in disease.

The replication of those measurements by mounting an independent donor collection constitutes a first step in that direction. In order to assess the immune cell parameters we used standardized flow cytometry procedures as previously applied (Hasan et al., 2015) with some modifications to fit with the local site configuration. The key points that we have focused on to limit pre-analytic variation were the sample processing, the reagent selection and the instrument parameters. Due to the absence of an automated platform, the potential variations that can be introduced at the step of the sample processing were limited by restricting the

number of trained users (two) for the entire collection and the use of an antibody mix per day for each panel to limit inter-donor variability due to multiple pipetting.

This technical variability has been measured by repeatability testing (three donors processed three times within the same day) that showed a coefficient of variation <10% (data not shown). We selected the same reagents as for the *MI* collection and used the same protocol/volume except for the antibodies that were newly titrated. This is of particular importance as we purchased new batches what can impact greatly the ratio of antibody-fluorochrome tandem and then the fluorescence intensity (one of the key parameters associated with genetic polymorphisms). In addition, each flow cytometer presents specific optical configuration (laser power, filter band pass...), which requires instrument-dependent optimization. Of note, for some of the antibodies we reduced the antibody volume by 10 times in this replicative study compared to the *MI* study (data not shown). One important configuration distinguishing both studies was the instrument used. We have seen that this change was impacting the method for the cell counts assessment as the MacsQuant (Miltenyi; used in the French collection) allows a direct cell count by aspirating a determined volume of blood whereas the FacCanto II (Becton Dickinson) used in the replicative study, not. We have shown that the use of a clinical grade white blood cell count from the same day was able to compensate this lack.

Nonetheless, we also observed a constant higher frequency and cell count for different innate cell types (neutrophils, monocytes and eosinophils) whereas the lymphoid cells assessment was not affected by the method used. Even if a technical dissimilarity could easily explain these differences, we cannot exclude that the markers included in our panel do not cover the potential high heterogeneity of those populations. An alternative (or cumulative) explanation could be that those populations are more sensitive to the staining procedure than the lymphocytes even if we limit as much as possible pre-analytical bias by using whole blood, with a processing within the 4 hours post-sampling using standardized protocol.

Importantly, despite those differences between cell count assessments, the hierarchical proportion between populations was perfectly preserved in all donors (**Figures 50 and 51**).

This aspect will be further explored to determine if some cell specific corrections are required. As for the *MI* study, the cytometer stability has been controlled across time, prior and during the study. Those tests included the use of Rainbow beads (Becton Dickinson) to check instrument fluorescence detection stability and to determine the application settings to apply daily for a given instrument to allow mean fluorescence intensity comparison across samples during the duration of the entire collection (around 6 months). Those settings have been

validating by reproducibility testing (three repeated draws for 3 donors across time) that showed a CV <25%. The preliminary analysis of the range of the cell frequency revealed similar ranges that the one observed within the *MI* collection. Interestingly, some outliers within specific populations have been also detected (data not shown).

Indeed, even if deeper statistical analysis have to be performed, we observed encouraging trends in term of genotype/phenotype associations within the innate compartment that replicate those observed in the *MI* collection confirming the value of our standardized approach.

In summary, we find that, in a thorough examination of numerous innate and adaptive cell parameters of 1,000 healthy subjects, inter-individual variation of innate immune cell parameters is dominated by genetic factors, while that of adaptive cell parameters is strongly influenced by non-genetic factors, in which genetic effects are probably masked by adaptation and reaction to environmental exposures.

## **CHAPTER IV**

---

### **DISCUSSION AND PERSPECTIVES**



## CHAPTER IV: Discussion and Perspectives

The overall aims of this PhD work was to provide to the community reference datasets and analytical strategies to decipher the natural variance in the immune response at the level of population-based studies. This work has been performed in the context of the *Milieu Intérieur* (MI) project that thanks to a highly annotated 1000 donor collection, aims to determine what are the genetic and environmental factors that drive the healthy human immune response.

The findings presented in the first study of this manuscript include i) a robust and reproducible transcriptomic analysis pipeline for in vitro stimulated whole blood samples and ii) a robust identification of a restricted set of genes that are capable of deconvoluting complex responses to immune stimulation. Those tools and analytical approaches can be used to map inflammatory signature from new adjuvant/drug candidates or to determine the natural variance of induced immune responses in population-based studies (healthy or pathogenic condition) that can predict responses to treatment or vaccination for example.

The latest is of particular interest for clinical decision and can be even more powerful if coupled with other phenotypic and/or genetic analysis like the one presented in the second study of this manuscript. The major findings from our study are the confirmation of environmental factors associations with adaptive immune phenotype (i.e. CMV infection and increase of highly differentiated T cells) whereas the majority of genetic associations were found within innate immune cells compartment.

Those approaches are particularly powerful when performed on the same time on the same individual (system biology) with reproducible tools to lead to the field of the personalized medicine.

At the end of the second chapter, we have already discussed the advantages and specific limitations of our induced immune response exploration. But some other considerations can be discussed in the context of the entire collection.

## Deconvolution of blood tissue induced signatures

In this manuscript we have shown that the standardized characterization of immune response in whole blood generate a powerful dataset that reflect the immune profile of a single individual in steady state condition as well as on pathological conditions (that we can mimic by *in vitro* stimulations). Blood is a complex tissue (as described in the first chapter) and differences across individuals in transcript abundance can be attributed to either genetic transcriptional regulation and/or relative changes in composition or expression patterns of leukocyte populations. The deciphering of these two factors can be next addressed by extensive isolation and profiling of individual blood cell populations (with potential associated technical bias) or by computational deconvolution analysis to deduce cellular composition or cell-specific levels of gene expression using statistical methodologies (Abbas et al., 2009). These whole blood “deconvolution” is now commonly used in studies with RNA sequencing datasets (Chaussabel and Baldwin., 2014).

In our study however, the use of the Human Immunology v2 codeset from Nanostring technology allows the analysis of a limited number of genes that restrict robust creation of those modules as numerous “lineage” genes are shared by multiple cell types depending of stimulation condition. Moreover, some populations are Nonetheless, a cell deconvolution approach could be validated in the future by combination with the cell counts obtain by flow cytometry.

It is interesting to note that despite the numerous deep transcriptional profiling studies, there is still no general consensus for data mining analysis. This is in part due to the use of different technological approaches at each step of the sample preparation/processing. Those differences impact greatly the analysis pipeline required for final validation. In addition, the increase use of high throughput approaches allows the generation of big datasets without prior precise scientific hypothesis (also referred as discovery research). As a consequence, there is more and more public datasets that can be used and compared for our own hypothesis research. The primary caveat of this approach is the bioinformatic skill required or an efficient bridge to build between bioinformaticians and immunologists. In addition, again due to technical differences or not equivalent sample annotation, those datasets are not always easy to reproduce or to use for multi-study comparison.

At each step of the *MI* collection, the aim was to limit the technical bias to provide reproducible datasets and web-based analytical tools to help the scientists to perform data

mining with their own biological hypotheses. In this manuscript, we have demonstrated that by using those specific tools, we were able to detect some extreme phenotypes that are interesting for further mechanistic explorations.

One example is the IL-1 $\alpha$  non-secretor phenotype observed upon whole blood bacterial stimulation (LPS or *E. coli*). First identified in 4 to 8% of a set of 25 healthy donors (Duffy et al., 2014), we have confirmed this phenotype in a set of 400 *MI* donors with a frequency increasing up to 35% in the oldest group (unpublished data) and also in 15% of the healthy individuals included in the gGAP replicative study (data not shown). Interestingly, the *IL1A* mRNA expression analysis (Nanotrinq codeset) performed on the 1000 *MI* donors for the null and whole bacteria conditions (on going analysis in the consortium), reveals normal level of induced transcripts on these donors. This indicates that the absence of detectable protein in the supernatant following stimulation is potentially impacted by differences in post-translational modifications and/or cell death pathways (implicated in the IL-1 $\alpha$  release) rather than by protein synthesis. This latest hypothesis is supported by the observation of the same level of intracellular IL-1  $\alpha$  in the monocyte population since 3 hours post-stimulation by LPS between non-secretor and secretor donors from the replicative study (data not shown). This is an on going exploration in our laboratory with potential clinical interest concerning inter-individual variability in term of cell death pathway-related inflammatory mechanisms.

### **Characterization of new adjuvants**

The analytical tools developed in this study allow defining a restricted set of genes specific for the signature of four clinical relevant pro-inflammatory cytokines that reflect distinct inflammatory origins. This list of genes if considered as biomarkers could be of particular interest for clinical profiling of unknown diseases or follow-up without going trough RNA sequencing that can be more expensive and requires bioinformatic resources. For example those approaches could be of particular interest for medical decision concerning autoimmune inflammatory diseases as spondyloarthritis or Lupus where additional factors (genetic, environmental) can impact the disease outcome or the response to the treatment (Banchereau et al., 2016). Other interesting application could be the classification of new adjuvants/drugs in term of immune perturbation but also the classification of donors/patients depending of their response to those unknown components that would help to predict potential non-responders before starting expensive and/or toxic treatments.

## Phenotype/Genotype associations

From our genotype/phenotype analysis, we have been able first to confirm some already described associations (i.e. eye color/) demonstrating the robustness of our dataset and allow us to find new hits associated within the innate compartment. Those findings are the first step that will need then to be validated by functional explorations. A first approach would be the analysis of RNA expression associated with gene polymorphism.

At the end of the chapter two we already evoked that it remains to be established how the inter-individual variation in gene expression that we identified is accounted for by host genetic variants (i.e., expression quantitative trait loci, eQTLs), specifically in cases where gene expression variation is altered upon activation with certain immune stimuli (i.e., response/interaction eQTLs). This analysis can be enlarged to other phenotypes already (or soon) available from *MI* collection like the secreted proteins from the same stimulated samples (pQTLs). This will give other layer of understanding in individual immune profiling. The transcriptomic analysis indeed is the first step of immune regulation but do not take into account the post-transcriptional regulation neither potential post-traductional events that can impact protein expression and then impact immune effect. In addition, those events can be genetically or epigenetically regulated.

In this manuscript I have described the assessment of immunophenotypes in genetically annotated donors in a replicative study. This dataset has shown encouraging results and will be important in order to validate our findings but also to describe a potential variability due to a different genetic and/or environmental background. One of the major parameters impacting the immune phenotypes in the *MI* cohort was the tobacco. However, we will not be able to address this question in this replicative “healthy” study as only one active and 4 former smokers were included. Some other explorations still need to be extended like serologies for latent infections (i.e. CMV).

In parallel of immune phenotype assessment, for each donor included in the replicative study we performed TruCulture stimulation during 24 hours for 5 conditions: Null, LPS, SEB, HKEC and HKSA. The stimulated supernatants were frozen and the stimulated pellets were stabilized in Trizol LS before storage at -80C. This collection will give the opportunity to performed eQTLs and pQTLs analysis like in the *MI* study in order to validate or complete our findings. Using this new reference dataset, we will be able to apply the same analytical tool pipeline providing a new step to the community toward the personalized medicine approach.

## Personalized medicine

We have to keep in mind that even if it is a good approach, the use of blood tissue does not give the entire picture of an individual immune profile. Characterization of other compartments can complete the system biology profiling and help to understand other factors that can influence the immune response. One first interesting compartment (not so difficult to reach) is the stroma. In our collection, we have addressed this by banking the fibroblasts generated after skin biopsy. Interestingly they represent cells from the connective tissue (and the comparison may be enlarged to many cells), are genetically pre-defined and constitute a first line in contact with local inflammatory inducers. Like all the other cells, fibroblasts express PRRs to be able to sense the danger and trigger immune response and/or participate in its amplification. This innate defense can be explored and associated to others phenotypes and genotypes. Differences in circulating immune response observed between donors/patients could be due to differences in the first steps of inflammation triggered by those cells.

An other important compartment that has now been demonstrated to impact the immune response behavior and composition, is the commensal microbiota (Pasparakis et al., 2014). This flora is present in all the epitheliums (skin and mucosal epitheliums like gut, airways and uro-genital tract) and is composed by different bacteria (bacteriome), viruses (virrome), fungi (mycobiome) species that can vary depending of the individuals and are localization specific. The composition of this microbiota shapes the immune response “education” since the childhood and understanding the relation between its composition and immune response profiles in healthy individuals could greatly help in associating a lost of this equilibrium with pathogenesis (Honda and Littman., 2016). In the context of the *Milieu Intérieur* collection, the gut bacteriome profiling has been recently achieved on the 1,000 donors and a preliminary analysis indicates three clusters of donors depending of their bacteria composition. An ongoing similar bacteriome based on 16s sequencing from the nasal swabs samples is performed. The future association of those datasets with the immune profiling will be a great value to define key parameters that impact natural variation of immune response. Those datasets would be then enlarged to others microbiome analysis i.e. the mycobiome that Hoarau et al. have recently associated to a microbial dysbiosis in familial Crohn’s disease (Hoarau et al., 2016). Those microbial analyses are of particular interest in the cancer field (but not restricted to) as it has been shown that the efficacy of checkpoint inhibitors are

correlated to gut microbiota in mice and could explain the discrepancy between patients to respond to immunotherapies (Vetizou et al., 2015; Sivan et al.,2015).

For all those reasons, using those standardized integrative approaches and render public the related annotated dataset (accompanied by visualization tools) would be highly valuable to the scientific community to move faster to personalized medicine decision.

## **MATERIALS AND METHODS**

---

### **HOW TO BUILD A REFERENCE DATASET?**





## MATERIAL AND METHODS:

### Technical challenges - How to build a reference dataset?

In the context of a French Laboratoire d'Excellence (LabEx) research program, the *Milieu Intérieur* Consortium was developed with the objective to define the determinants (genetic or environmental) of human immune variance in healthy individuals.

With this objective in mind, the environmental factors were assessed thanks to an extensive questionnaire allowing the collection of the every-day life habits, a serological status assessment for a myriad of microbe infections/vaccines and a sample collection for the characterization of the metagenomic diversity based on sequence analysis of bacterial, fungal and viral populations in fecal and nasal samples whereas the genomic variability has been assessed using genome-wide SNP genotyping. To determine the immune variance, we've measured the levels and quality of circulating immune cell populations based on standardized flow cytometry procedures and the induced transcriptional and protein signatures in response to medically relevant immune inducers.

In parallel, the establishment of a cell bank from those genetically annotated donors, including EBV-transformed B cell lines and fibroblasts will allow mechanistic studies.

In the context of this collection, one of the goals of this thesis was to build a reference dataset for induced gene expression from human whole blood samples.

In this section, I'll detail how the *Milieu Intérieur* collection aims to serve as a reference highlighting the efforts put to standardized each aspect/tools used for the data/sample collection. I'll then focused on the development of the induced gene expression analysis and the analytical tools used to provide robust biological analysis.

Additional information can be found at: <http://www.pasteur.fr/labex/milieu-interieur>

## 1. Donor collections

### 1.1. The *Milieu Intérieur* cohort

#### 1.1.1. Healthy donors' recruitment

All subjects provided informed consent prior to enrollment in the study. Taking advantage of a pre-existing donor database used for pre-screening, final 1,000 donors stratified by gender (i.e., 500 women and 500 men) and age (i.e., 200 individuals from each decade of life, between 20 and 70 years of age) were recruited in accordance with the study criteria (extensively depicted in the « Additional publication » : Thomas et al., 2015).

To maximize our ability to associate genetic and epigenetic variation with individual phenotypes, a detailed list of inclusion and exclusion criteria were defined mainly to avoid major perturbed immune system. In example, donors could not have evidence/report of neurological, psychiatric disorders or severe/chronic/recurrent pathological conditions. Same for history or evidence of alcohol abuse, recent use of illicit drugs, vaccine administration and use of immune modulatory agents. Only pre- or post-menopausal women were included to avoid the influence of hormonal fluctuations. After initial evaluation for recruitment criteria, additional physical examination and clinical laboratory testing were performed at visit V0 (**Figure 53**).

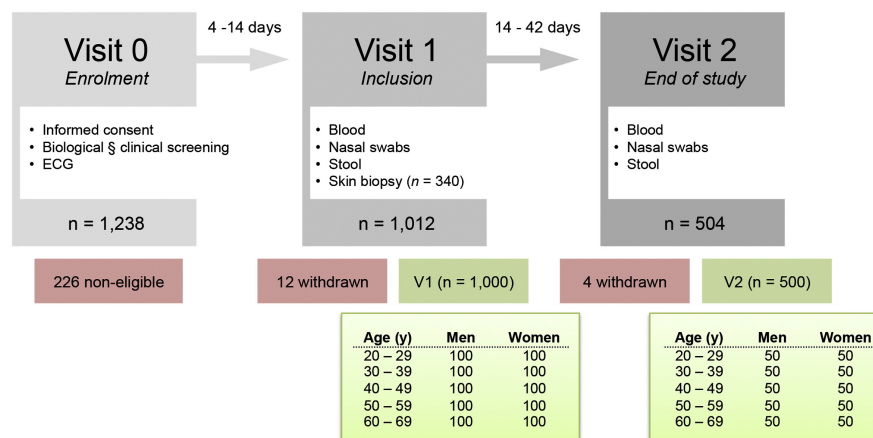
After this visit the others considered criteria were a BMI restricted to  $\geq 18.5$  and  $\leq 32$  kg/m<sup>2</sup>, an absence of urinary human chorionic gonadotropin (hCG) test for female donors, of urinary toxicology screens for cannabinoid use, proteinuria and glycosuria on all donors.

To avoid highly variable genetic backgrounds due to different ancestry, the study was restricted to French citizens (European-descent), from Metropolitan French origin for three generations.

#### 1.1.2. Clinical and epidemiological collection

An independent sampling of 20 mL of blood has been collected at V0 and V2 (for repeat) and used for clinical chemistry, hematologic and serologic assessments. In order to obtain clinical grade results, all the assessments were performed in a certified Laboratoire de biologie médicale, Centre Eugene Marquis (Rennes, France) (**Figure 53**).

During this first pre-visit (V0), an epidemiological collection has been performed with the record of detailed medical histories and questionnaires. Those questionnaires included general information about socio-demographic, lifestyle and family health history that were recorded in an electronic case report form. For example, the donors were asked to provide their family status, income, occupational status and educational level to integrate these information as social-demographic variables; smoking habits, alcohol intake, sleeping habits, depressive symptoms, nutritional behavior and habits were recorded as lifestyle variables.



**Figure 53 : Schematic representation of donor recruitment for the *Milieu Intérieur* study.**

To include 1000 healthy persons stratified according to sex (500 men, 500 women) and age (200 donors per decade of life, 20–69 years of age), we enrolled a total of 1238 individuals at visit 0 (Enrollment). Of those screened, 226 donors were considered non-eligible for reasons of consent withdrawal ( $n = 54$ ), past medical history ( $n = 67$ ), identification of an exclusion criteria during the onsite physical examination ( $n = 54$ ), or during laboratory testing ( $n = 51$ ). An additional 16 donors withdrew consent in the course of the study. During visit 1, whole blood, fecal samples and nasal swabs were collected. Punch biopsies of the skin were obtained from 340 of these donors. Half of the subjects were randomly selected (stratified by age and sex) to return for a visit 2, when repeat sampling of whole blood, fecal samples and nasal swabs was performed. Detailed medical histories and questionnaires were completed from all donors, recorded by medical personnel using an electronic case report form.

### 1.1.3. Biological collection

During the first visit after inclusion (V1), blood, nasal swabs and stool samples were collected from all donors (1000) according to established protocols. For 500 individuals, the same samples were collected during a second visit (V2) – 14 to 42 days later – in order to provide biological replicates that can be used to validate final phenotypic studies (**Figure 53**).

For 340 randomized donors, a skin biopsy has been performed at V1 in order to generate fibroblast primary cell lines.

The blood sampling (up to 12 donors per day) included 20 mL of blood into Na Heparin tubes for Flow Cytometry study on whole blood, plasma and PBMC banking. In addition, a 5mL EDTA tube was collected for DNA banking. The tubes were maintained at 18–25 °C during transport and processed within 6 h of collection. A standardized whole blood staining protocol has been developed for 10 panels (8 colors each) and adapted to an automated platform (EVOWare, Tecan). The samples were then acquired in a Macs Quant instrument (Hasan et al., 2015).

For the induced immune response study, 60 mL of blood were collected. Within 15 min of collection, 1mL of blood was distributed into 40 different TruCulture® tubes at the collection site (Rennes). Those tubes were specifically developed and adapted to this study to induce reproducible innate or adaptive immune responses (more details about the development will be described in the next paragraph). After thawing to room temperature, the collection syringes were filled with 1ml whole blood at the collection site within 15 minutes maximum after drawing, and incubated for 22 hours ( $\pm 10$  min) in room air at 37 °C ( $\pm 1$  °C), utilizing a bench-top heating block (VLMH GmbH). The specific media included allows tube's incubation at 37°C directly in a bench-top heating block without need of CO<sub>2</sub> exchange incubator. After incubation, the immune cell activation was stopped by insertion of a valve separator (provided with the syringe device) until reaching 1 mm above the pellet and the culture supernatant was collected, aliquoted and stored at -80°C. At the end of the incubation period, the supernatants were aliquoted, and the nucleic acid content from the cell pellet stabilized in Trizol. Both samples were stored immediately at -80 °C at the collection site.

For the gut microbiota analysis, the volunteers were asked to produce fecal specimen at their home using the GENbag Anaer atmosphere generator (Aerocult, Biomerieux). This system maintains anaerobic conditions (validated by an anaerobic indicator strip) to preserve anaerobic bacteria species. The stools samples were aliquoted into cryotubes and stored at -80 °C upon arrival at the clinical site.

To assess the airways microbiota, nostrils were sampled for each individual using dry flocced swabs (FLOQSwab™) which were stored in stabilization media provided in the kit and frozen at -80 °C.

Finally, in order to generate a randomized collection of human fibroblast cell lines, skin punch biopsies were performed on 340 individuals at the clinical collection site (Rennes) under local anesthesia. The same day, the sample was shipped at 4 °C to Genethon (Evry, Ile de France, France) where the fibroblast cell lines were generated and aliquots stored in liquid Nitrogen.

All the samples stored at  $-80^{\circ}\text{C}$  were weekly transported in dry ice to Institut Pasteur (Paris) and all the fibroblast cell lines were transported in dry shipper at the end of the collection.

### **1.2. The Genentech Genotype and Phenotype Registry collection**

In order to validate the genome-wide association analysis observed within the Milieu Interieur cohort, we have put in place a replicative study using volunteers from the Genotype and Phenotype (gGAP) Registry (Genentech, Inc., South San Francisco). So far 81 healthy individuals were recruited and integrated into a preliminary analysis. The individuals are regular blood donors tested seronegative for most non-cured chronic infections (i.e. HIV, HCV, HBV). Donors with Caucasian ancestry were recruited (validated by genotyping analysis) with a median of 39 years of age (range: 25 to 50 years old) and 62% are men. The donors were asked not to have intense physical exercise 12h before the sampling and were clinically examined to determine their healthy status. The day of the visit a clinically approved white blood cells (WBC) differential count was performed to confirm the absence of inflammation.

## **2. Focus on the study of induced immune responses**

As mentioned before, the study of whole blood allows the preservation of physiological cellular interactions and is an interesting snapshot of an individual whole immune status, as the tissue specific immune cell subtypes tend to recirculate through blood stream. Nevertheless, even the study of an “untouched” tissue requires the establishment of standardized and reproducible in vitro assay systems to be able to measure biological variance instead of technical variance.

In that aim, we have used the TruCulture tubes to reproducibly stimulate 1mL of whole blood. Series of repeatability and reproducibility tests were performed showing coefficients of variance (CVs) ranging from 5 to 14%, demonstrating a low technical and biological variance (Duffy et al., 2014).

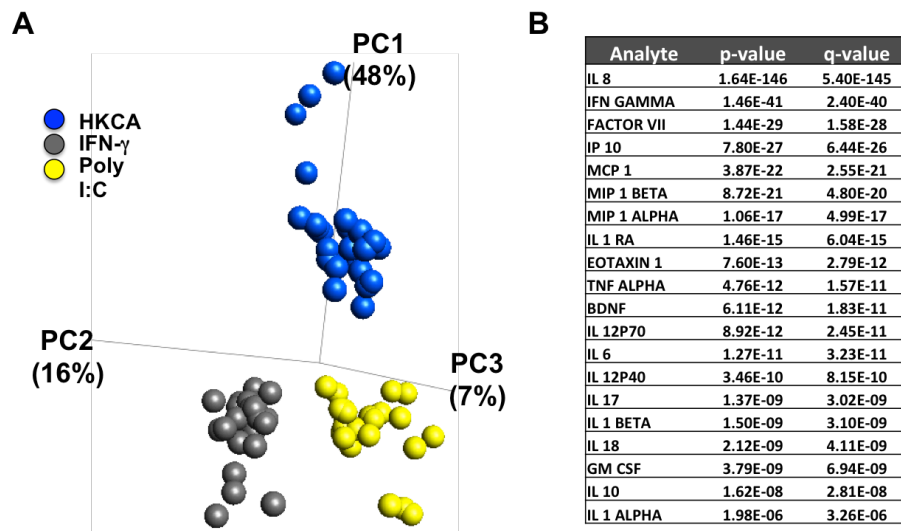
In close partnership with the company Myriad RBM, a first list of fifty-four stimuli including microbial, viral, fungi components (purified or complex), cytokines, superantigens were considered for study and evaluated. To validate the clinical grade of the stimulation system,

each component was evaluated in term of sterility, solubility, dose response, short and long-term stability, and reproducibility. All those evaluations and the description of the exclusion criteria are fully described in the first manuscript published by the consortium (« Additional publication » : Duffy et al., 2014). Dose concentrations were selected for the stimuli that maximized the ability to detect low-expressed proteins, without exceeding the upper limit of the biologic range for highly expressed proteins.

To summarize, final 27 stimuli were selected, dissolved in 2 ml of proprietary TruCulture medium and included in TruCulture syringe devices (Myriad RBM) (Table 17). The devices were then frozen and stored at -20°C until use.

In order to validate the robustness and the sensitivity of the TruCulture whole blood stimulation system at the level of a population-based analysis, the protein signature upon 28 conditions (including Null) was performed in a first study including 25 donors from the group aged 30–39 years old, stratified by gender (13 women, 12 men).

A standardized multiplex protein immunoassay platform (Myriad RBM) was used for protein detection and the panel design included low and highly induced protein analytes. This study has demonstrated that using this syringe-based system, we are able to capture unique induced signature at the protein level and also to detect some extreme phenotypes interesting to follow for further exploration. Representative data for one microbe, TLR agonist, and cytokine stimulus is shown (Figure 54), and the selected dose for all assays can be found in Table 17.



**Figure 54 : Distinct protein inflammatory signatures for whole microbe, TLR and cytokine stimulation**

(A) Principal component analysis (PCA) was performed on the data set obtained from 25 healthy donors. Blue circle represents samples stimulated with heat killed *Candida albicans* (blue circles), IFN- $\gamma$  (grey circles) and Poly I:C (yellow circles). The PCA was run with data obtained from the analysis of 33 proteins. The PCA plot shown captures 71% of the total variance within the selected data set (PCA1, 48%; PCA2, 16%; PCA3, 7%).

(B) The induced responses to whole-blood stimulations with HKCA, Poly I:C and IFN- $\gamma$  were compared and the 20 first differentially expressed proteins were identified (ANOVA q value < 0.05).

Stimulus	Abbreviation	Concentration	Supplier	Sensor or Receptor	Reference
Nul	Ø		NA		
<b>Microbe</b>					
HK <i>E. coli</i> 0111:B4	HKEC	10 <sup>7</sup> bacteria	Invivogen	complex	Takeuchi et al., 1999
HK <i>S. aureus</i>	HKSA	10 <sup>7</sup> bacteria	Invivogen	complex	Krishna and Miller, 2012
HK <i>L. rhamnosus</i>	HKLR	10 <sup>7</sup> bacteria	Invivogen	complex	Miettinen et al., 2008
BCG (Immucyst)	BCG	3 × 10 <sup>8</sup> bacteria	Sanofi Pasteur	complex	Means et al., 1999; Godaly and Young, 2005; Randhawa et al., 2011
HK <i>H. pylori</i>	HKHP	10 <sup>7</sup> bacteria	Invivogen	complex	Zhao et al., 2007
HK <i>C. albicans</i>	HKCA	10 <sup>7</sup> bacteria	Invivogen	complex	Brown et al., 2003; Gantner et al., 2003
Influenza A virus (live)	IAV	100 HAU	Charles Rivers	complex	Diebold et al., 2004; Kato et al., 2006; Ichinohe et al., 2009; Allen et al., 2009
Sendai virus (live)	SeV	10 HAU	Charles Rivers	Rig-I and Mda/5	Yoneyama et al., 2005; Kato et al., 2005
<b>MAMP</b>					
C12-IE-DAP	DAP	4 µg/ml	Invivogen	NOD1	Chamaillard et al., 2003
CPPD	CPPD	100 µg/ml	Invivogen	NLRP3 and TLR2	Liu-Bryan et al., 2005; Martinon et al., 2006
FSL-1	FSL	2 µg/ml	Invivogen	TLR2/6	Shibata et al., 1997; Okusawa et al., 2004
Poly I:C	pIC	20 µg/ml	Invivogen	TLR3	Alexopoulou et al., 2001
LPS-EB (ultrapure)	LPS	10 ng/ml	Invivogen	TLR4	Poltorak et al., 1998; Shimazu et al., 1999
Flagellin-ST	FLA	0.25 µg/ml	Invivogen	TLR5	Hayashi et al., 2001
Gardiquimod	GARD	3 µM	Invivogen	TLR7	Hemmi et al., 2002
R848	R848	1 µM	Invivogen	TLR7 and TLR8	Jurk et al., 2002
ODN 2216	ODN	25 µg/ml	Invivogen	TLR9	Hemmi et al., 2000; Krieg, 2002
lipoarabinomannan	LAM	10 µg/ml	Invivogen	Mannose R, CD36	Józefowski et al., 2011; Sieling et al., 1995
WGP	WPG	40 µg/ml	Invivogen	Dectin-1	Goodridge et al., 2011
<b>Cytokines</b>					
IFN-α <sub>25</sub> (Intron A)	IFN-A	1,000 IU/ml	Merck	IFNAR	González-Navajas et al., 2012
IFN-β (Betaseron)	IFN-B	1,000 IU/ml	Bayer	IFNAR	González-Navajas et al., 2012
IFN-γ (Imukin)	IFN-G	1,000 IU/ml	Boehringer Ingelheim	IFNγR	Platanias, 2005
TNF-α	TNF-A	10 ng/ml	Miltenyi Biotec	TNFR	Kolb and Granger, 1968
IL-1β	IL-1B	25 ng/ml	Peptotec	IL1R	March et al., 1985
IL-23	IL-23	50 ng/ml	Miltenyi Biotec	IL23R	Oppmann et al., 2000
<b>T Cells</b>					
α-CD3 + α-CD28	CD3+CD28	0.4 µg/ml + 0.33 µg/ml	RND Systems + Beckman Coulter	TCR	Smith-Garvin et al., 2009
Enterotoxin SEB	SEB	0.4 µg/ml	Bernhard Nocht Institute	TCR and MHC II	Fleischer and Schrezenmeier, 1988

**Table 17 : Innate and adaptive immune stimuli used for development of whole-blood stimulation system.** Abbreviations are as follows : HK, heat Killed ; HAU, hemagglutinin units ; IU, international units. The 28 stimulation conditions used for the preparation of TruCulture tubes are listed, with the indicated dose and commercial supplier. Stimuli are ordered based on four categories : whole microbes, MAMP, cytokine and T cell agonist.

### **3. Standardized gene expression quantification from stimulated whole blood**

Having demonstrated that our *in vitro* whole blood stimulation system was able to capture unique stimulation protein signature, the next step was to measure this signature at the gene expression level.

As mentioned before, after a given stimulation for a given whole blood sample, in parallel to the supernatant collection for the proteomic study, the stimulated cell pellet was stabilized and stored with Trizol LS® directly in the TruCulture tube.

In the introduction section of this manuscript, I have already briefly mentioned what are the different technologies available for gene expression analysis. Nonetheless, depending of the biological question and most importantly depending of the sample source (i.e. single population versus whole blood; heparin vs other anticoagulant tube) there is no general consensus across studies in the pipeline to apply from the RNA purification to the gene expression analysis.

In this next methodological section, I will describe first the development of our single-step mRNA purification, the gene expression measurement platform used and finally described the statistical and visualization tools applied for the transcriptomic analysis.

#### **3.1. Development of a high throughput single-step RNA isolation protocol**

##### **3.1.1. Challenge of Trizol® stabilized whole blood samples**

The reference protocol to isolate RNA from Trizol® lyzed samples requires a first step of chloroform extraction. The addition of chloroform followed by centrifugation separates the solution into an upper aqueous phase containing RNA and a lower organic phase. This aqueous phase need to be collected and transferred before purification. This step is really challenging to automatize into a high throughput platform (due to the centrifugation's step) and is highly susceptible to user variability (due to the requirement of a precise pipetting step). To be able to process a large number of samples (60000 tubes in our entire collection)



in addition to minimize pre-analytical biases, we established an mRNA single-step chloroform-free extraction protocol in collaboration with the Macherey-Nagel Company (detailed below).

To validate the protocol, direct comparison with conventional RNA extraction protocols (following manufacturer's instruction) indicated an extremely high correlation in gene expression counts (Nanostring Technologies hybridization) between the two extraction methods (Spearman's Rank-Order Correlation,  $r_s > 0.99$ , **Figure 9** « section Study 1 »).

### **3.1.2. High-throughput standardized RNA extraction**

Samples were randomized and extracted in groups of 95. Cell pellets were thawed on ice 30 - 60 min prior to initiating processing. To complete the thawing and the RNA release, the tubes were vortexed twice for 5 min at 2000 rpm. Before processing, a centrifugation (3000 g for 5 min at 4°C) of the thawed samples was performed to pellet the cellular debris generated during the Trizol® lysis. The barcoded tubes were loaded in the rack module of the Freedom EVO® platform (TECAN, Switzerland) and scanned for sample traceability. For extraction, the modified protocol of the NucleoSpin® 96 RNA tissue kit (Macherey-Nagel, Germany) was adapted to the Freedom EVO® integrated vacuum system. Briefly, 600µl of clarified phase of the Trizol® lysate was transferred to a deep well plate preloaded with 900µl of 100% Ethanol. The binding mixture was transferred into the silica membrane plate. The columns were washed with Buffers MW1 and MW2 (x2), and RNA eluted into 0.5ml 2D barcoded tubes (ThermoScientific) using 60µl RNase-free water. As an internal control of the extraction process a tube containing a defined quantity of spiked RNA was included in each run.

To avoid unnecessary freeze and thaw of the RNA, distinct aliquots for quality control and gene expression analysis were prepared, and all aliquots frozen at -80°C until the time of use.

### **3.1.3. RNA quality**

RNA concentration was estimated using Qubit® RNA HS Assay Kit (Life Technologies, USA) according to the protocol provided by the manufacturer.

An automated and high throughput RNA integrity assessment was performed using the Standard RNA Reagent Kit on a LabChip®GX (Perkin Elmer, USA). The RNA Quality Score

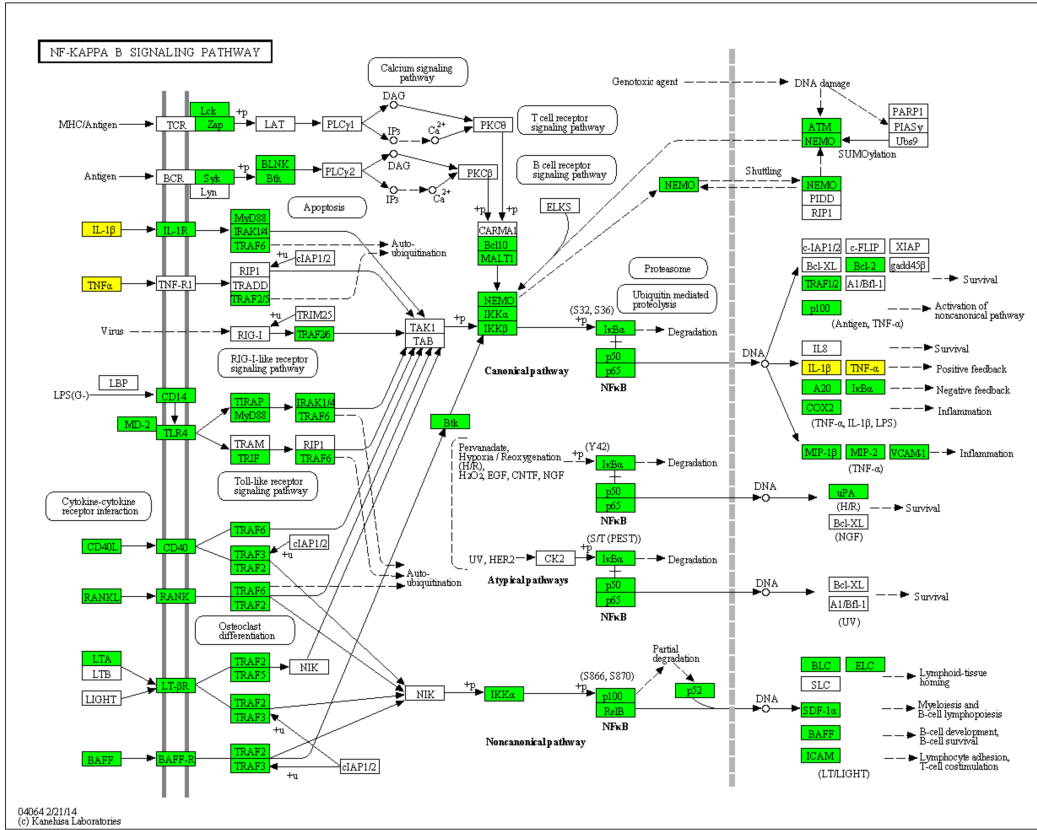
(RQS) was calculated using the LabChip® System software, and all samples with a RQS > 4 were processed for gene expression analysis.

### **3.2. mRNA quantification: direct hybridization assay**

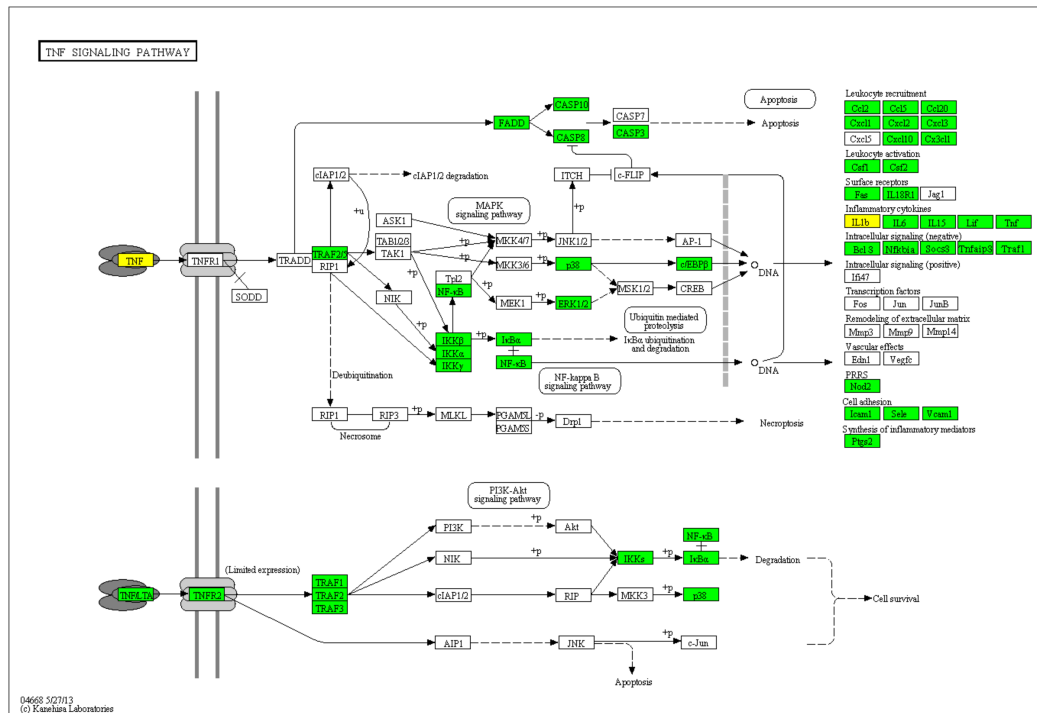
After comparison with other transcriptional profiling strategies, we have selected the NanoString platform (**Table 5** « section Study 1 »). This choice was validated by the high reproducibility of the data obtained when experiments were performed at different times or at separate institutional core facilities ( $r_s > 0.98$ , **Figure 9** « section Study 1 »).

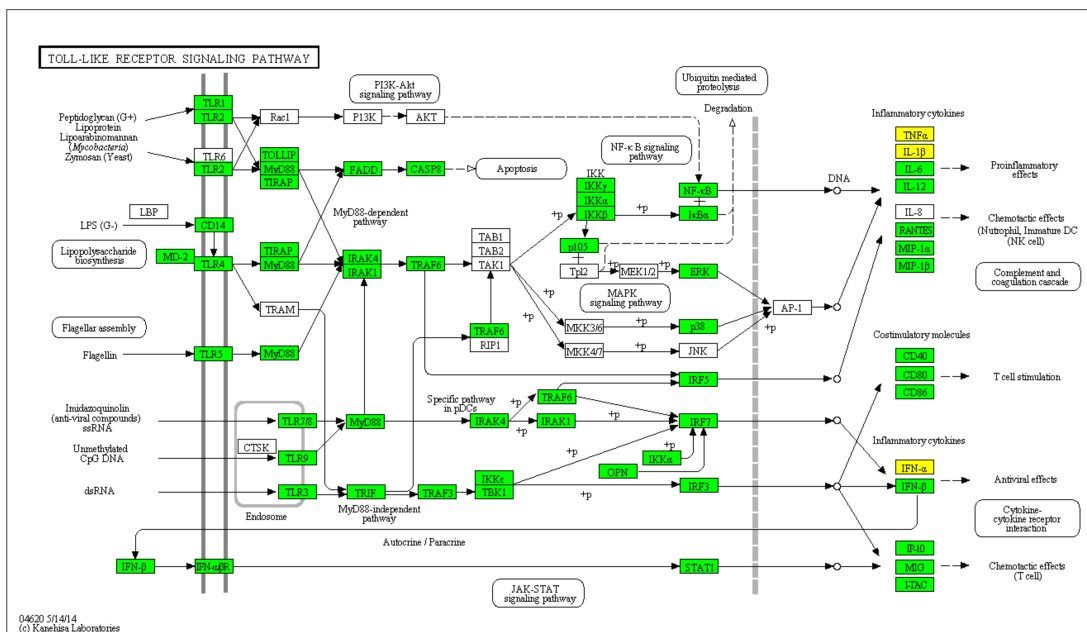
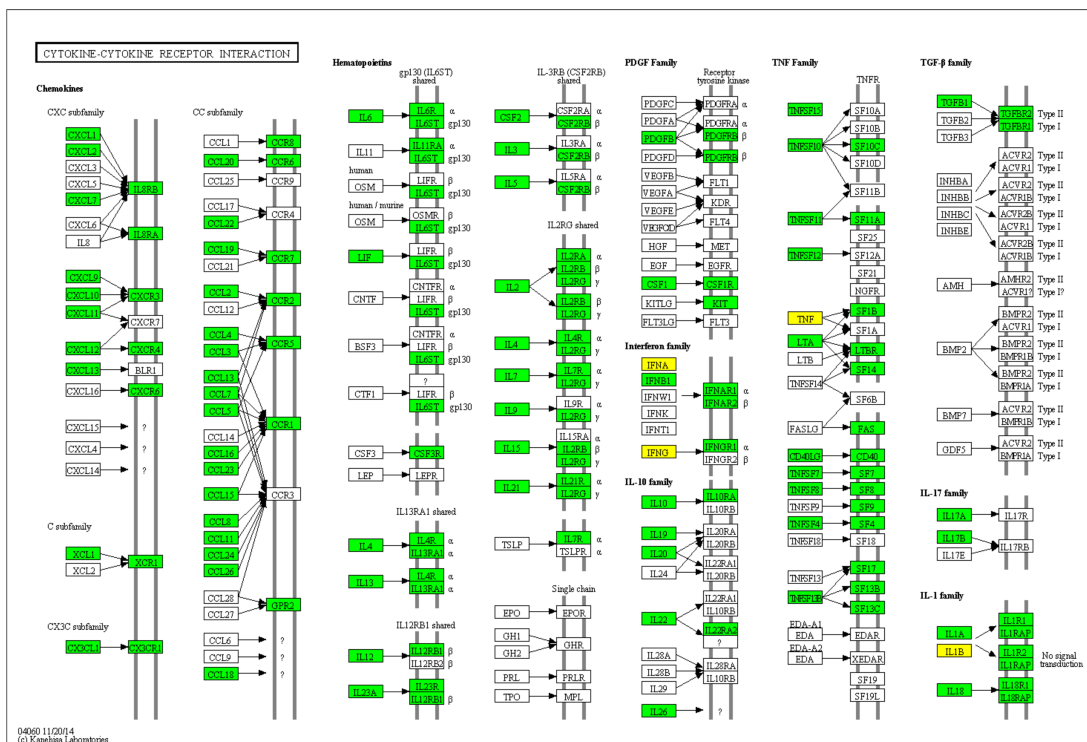
The Human Immunology v2 gene code set was selected as it covers 25 immunology-related gene networks, including the major cytokine and TLR induced gene expression pathways as illustrated by the use of KEGG charts (**Figure 55**). The codeset contains a total of 594 probes including 15 housekeeping gene candidates.

A



B





**Figure 55 : Gene expression pathways used to select NanoString Immunology panel.** KEGG database pathway analysis of (A) NF- $\kappa$ B, (B) TNFA, (C) Cytokine-Cytokine Receptor, and (D) TLR signaling pathways, with genes included in NanoString analysis colored green, and effector cytokines (IFN $\beta$ , IFN $\gamma$ , IL1 $\beta$ , TNFA) studied herein colored yellow. Genes in white were not represented on the NanoString codeset

### **3.2.1. nCounter codeset by Nanostring Technologies**

The first step consists of direct mRNA hybridization with color-barcoded probe pairs (around 50 bases length). The reporter probe carries the fluorescent signal, while the capture probe (coupled to a biotin) allows the complex to be immobilized in a streptavidin-coated glass slide for data collection. Each color-coded barcode is attached to a single target-specific probe corresponding to a gene of interest. The color codes carry six positions and each position can be one of four colors, thus allowing for a large diversity of tags that can be mixed together in a single well for direct hybridization to target.

Total mRNA prepared as described above, were diluted with RNase-free water at 20ng/ $\mu$ l directly in the 12-strip provided by NanoString. 100ng (5 $\mu$ l) of total RNA from each sample were hybridized with massive excess of pair probes set provided by the Human Immunology kit v2 according to the manufacturer's instructions.

### **3.2.2. Automated sample processing and analysis**

The specific formed complexes are then purified and immobilized in a complete automated manner by the nCounter® Prep Station instrument and certified reagents into a cartridge (for up to 12 samples per run) as described below.

After 16 h of hybridization, excess probes are washed away using a two steps magnetic bead-based purification. Magnetic beads coated with short nucleic acid sequences that are complementary to the capture probe and the reporter probes are used sequentially. Intermediate wash steps allow the removal of excess reporter probes and capture probes respectively. Finally, the purified Target/Probe complexes are eluted and immobilized in the cartridge for data collection.

Data Collection is then performed by the nCounter® Digital Analyzer that will collect the fields of view (FOV) per flow cell using a microscope objective and a CCD camera yielding data of hundreds of thousands of target molecule counts. Digital images are processed on the nCounter Digital Analyzer and the barcode counts are tabulated in a comma separated value (CSV) format.

### **3.3. Technical validation**

#### **3.3.1. Imaging quality control**

From the csv output file given by the digital analyzer quality control (QC) metrics can be calculated and used to validate the sample.

The first one refers to the percentage of fields of view (FOVs) successfully counted by scan of a lane and reflects the quality of the camera imaging. The second refers to the binding density per lane. When too many probes are present, the instrument is not able to distinguish each and every probe present in the lane indicating that the count results may not be accurate.

In the case of flags for those parameters, the sample was re-scanned or re-hybridized in the case of flag persistence.

#### **3.3.2. Internal positive probes**

Each nCounter assay contains a variety of positive control probes targeting molecules added at pre-determined concentration by NanoString during production of the CodeSet. A positive control linearity QC performs a correlation analysis in log<sub>2</sub> space between the concentrations of added targets and the resulting counts. Low correlation values may indicate an issue with the hybridization reaction (i.e. due to the presence of inhibitors, volume pipetting issue) and/or instrument performance.

#### **3.3.3. Internal negative probes**

Like the positive control probes, each nCounter assay contains negative control probes that should not match with any target gene from the sample processed. Those External RNA Controls Consortium (ERCC) developed by an ad-hoc group of academic, private, and public organizations hosted by the National Institute of Standards and Technology (NIST) are commonly used by different technology platforms. Then a limit of detection QC can be calculated by using the positive and negative control probes for each nCounter assay.

In our study, the eight negative probes used were:

NEG\_A, ERCC\_00096.1 NEG\_B, ERCC\_00041.1 NEG\_C, ERCC\_00019.1  
NEG\_D, ERCC\_00076.1 NEG\_E, ERCC\_00098.1 NEG\_F, ERCC\_00126.1

NEG\_G, ERCC\_00144.1 NEG\_H, ERCC\_00154.1

Negative control analysis was performed to determine the background for each sample. Of note, we observed variable expression of two negative control probes (NEG B, NEG F), which cross-reacted with bacterial nucleic acid present in two of the TruCulture stimulation systems (HKSA and BCG, respectively, **Figure 10C-D** « section Study 1 »), and thus these probes were excluded from use in data analysis.

### **3.4. mRNA content normalization**

#### **3.4.1. Criteria for housekeeping gene selection**

The ultimate goal of gene expression analysis is to reliably determine changes in transcript abundance across all our donors and under our different stimulation conditions. Whole blood is a complex tissue that does not allow the control of cell number and composition across donors. In addition, the mRNA content can change independently across stimulations. Even if the fixed amount of total RNA used in each experiment reduces errors, the normalization to internal reference (housekeeping) genes help correct for these differences in numbers of cells, absolute mRNA content and sample preparation (i.e. pipetting variability).

Usually one or more housekeeping genes are chosen as reference. In our Nanostring panel 15 candidate genes were provided. In our context, the challenge was to select the ones that display uniform expression across the immune cell types (i.e. lymphocytes, neutrophils) and under all stimulation conditions (i.e. bacteria, viruses).

#### **3.4.2. Housekeeping gene selection using gNorm R package**

To identify by an unbiased and reproducible way stable internal genes responding to the criteria cited above, we have apply the established geNorm method (Vandesompele et al., 2002) in our entire dataset. I want to emphasize the fact that this calculation has to be done for each final dataset to be normalized as each candidate housekeeping genes will not be impacted in the same manner by experimental conditions or cell sources.

Over the 15 proposed housekeeping genes, four have been selected (RPL19, TBP, POLR2A, and HPRT1) for final mRNA content normalization (**Figure 10A-B** « section Study 1 »).

### 3.5. Final data normalization pipeline using the nSolver® software

Data was imported into nSolver™ analysis software (version 2.5) for quality checking and normalization of data according to NanoString® analysis guidelines.

A first step of technical normalization using the internal positive controls permits to correct potential sources of variation associated to the platform. To do so, we calculated for each sample the geometric mean of the positive probe counts. A scaling factor for a sample was a ratio of the average across all geometric means and the geometric mean of the sample. For each sample, we multiplied all gene counts by the corresponding scaling factor.

Next, for each sample we calculated the background level as the median +2 SD across the 6 negative probe counts. For each gene in a sample we subtracted the background level. Then the content normalization was performed with nSolver™ Analysis Software that calculates the geometric means over the four housekeeping genes selected as described above in the entire dataset.

## 4. Selection criteria for gene expression analysis

The codeset contains a total of 594 probes (15 correspond to suggested housekeeping genes), of which 572 probes were included in downstream analysis after removing probes mapping to multiple genes and probes aligning to polymorphic regions with greater than two SNPs (**Table 18**). To this end, the probes were mapped against the human genomic sequence (GRCh37/hg19) with GSNAP (Wu and Nacu, 2010), a splice-aware aligner. To detect splice junctions GSNAP was applied with two criteria: (1) detection of novel splice sites, and (2) detection of known splice sites. 573 out of 594 probes mapped with 100% identity to the genome. 12 probes mapped with 1-2 mismatches in the middle of the sequence, 8 probes were misaligned in the first/last 1-9 bp, and 1 probe did not map at all (PECAM1 located on HG183\_PATCH). The misaligned probes were realigned manually using BLASTN against Ab-initio cDNAs database. Of the 594 probes 15 mapped to more than one genomic location (see **Table 18**). We removed from further analysis KIR\_Activating\_Subgroup\_1 probe, which mapped to 3 different genomic locations, as well as three other KIR probes that mapped to multiple locations: KIR\_Activating\_Subgroup\_2, KIR\_Inhibiting\_Subgroup\_1, and KIR\_Inhibiting\_Subgroup\_2. Next, we used Bioconductor biomaRt package (Durink et al., 2005) version 2.24.0 to query Ensembl (release 75) (Ficek et al., 2014) and to retrieve exonic



variants that mapped to the same regions as the NanoString probes. We considered only the SNPs with minor allelic frequency (MAF)  $> 0.05$  (in all individuals from phase 1 dataset of 1000 Genomes project). 48 probes showed the presence of 1-2 SNPs in their sequence. HLA-DRB1, HLA-DQA1 and HLA-DQB1 probes contained 4, 9, and 13 SNPs, respectively, and were therefore removed from further analysis (**Table 18**).













probe	chr	SNPs in probe (MAF)	ensembl gene IDs	hgnc_symbol	probek sequence
MR1	1	r2070770 (0.067)	ENSG00000196549	MMIE	GCATTTAGTGTGACAGTGTTCGAGAAAGGATGTTCTTCGAGCTTTCGCAAGTGTGCAGAGGTTGACCTTGGTATTTGCATATTTCTGG
MRL1	1	181030910-181031009	ENSG00000153029	MR1	AAACGTGTAATCTTCACTGAGGCTCATAGAGAGGAAAGTAACTCCATCCAGTCAAGTGTGGTCTGCAAGGATGTGTGGTGGAGGCTGT
MR1C	10	17865434-17865474; 17869541-17869599	ENSG00000183748	MR1	CTATGACACAGCAATCTGTCCAGAGGTTTGAAGACATGTATACCTTCTACAGTAACTGCCAATGGAACACTGTCGATCCGGTGTGAGTCT
MR2	11	18112405-18112445; 18116495-18116553	ENSG00000120586	MR1C	CTATGACACAGCAATCTGTCCAGAGGTTTGAAGACATGTATACCTTCTACAGTAACTGCCAATGGAACACTGTCGATCCGGTGTGAGTCT
MS4A1	11	60230502-60230594; 60231761-60231785	ENSG00000156738	MS4A1	GTTGATGAGTGTCCAGCAGGGATGTCCACCATCTGAGCTGTGTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT
MS1	8	16026369-16026739; 16032696-16032784	ENSG00000398945	MS1	AGAACAACTGAACTG
MU1	1	155158620-155158885; 155159770-155159734	ENSG00000273088; ENSG00000185469	MU1	CCCTAACGATACGATGCTAGAGGATTTCTGACGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
MX1	21	42817936-42818035	ENSG00000157601	MX1	GCCTTAACTAGACACACTCTCTCTCATGGAAGAAGAAGTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
MP1	3	36363845-38183946	ENSG00000170396	MP1	ACGTTTCTAGTTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
MP108	11	34195053-34195151	ENSG00000193845	MP108	GGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTT
MNR1	22	37257240-37257245; 37260082-37260171; 37260961-37260969	ENSG00000103465	MNR1	GGAATGTTTGGCTTAAACAACACTG
NCR1	19	55420813-55420882; 55421378-55421407	ENSG00000189430	NCR1	GATGTTTGGCTTAAACAACACTG
NEF1C1	18	77221365-77221366; 77222450-77222457	ENSG00000131196	NEF1C1	GATGTTTGGCTTAAACAACACTG
NEF1C2	20	50095030-5009629	ENSG00000102096	NEF1C2	GAAGCACTTGGAAAGAACAACACTG
NF1	16	6821752-6821752	ENSG00000107236	NF1	GTCTTCAAGGTTTCCATATCAACGAGGTGAGCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NF1L3	4	103516047-103516136	ENSG00000165030	NF1L3	GTCTTCAAGGTTTCCATATCAACGAGGTGAGCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NFKB1	9	94171517-94171518	ENSG00000109320	NFKB1	AGGTTATGCTTCCCAACTAGACTGATTTACTTGAATGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NFKB2	10	104157739-104157838	ENSG00000077150	NFKB2	ATCTCCGGGGCATCAACACTGAGACTGCTGAAGTCTGCAAGGATTTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NFKB3A	14	35871322-35871366; 35871600-35871654	ENSG00000109096	NFKB3A	GGAATGAGGAGATGATGACAGACTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NFKB3B	3	10157672-10157690; 101578162-101578229	ENSG00000144802	NFKB3B	ATTGGTCCGATGGCTTCCAGTGAAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGTCTGAGT
NLRP3	1	247583274-247583279; 247586532-247586625	ENSG00000106100	NLRP3	AGAGATGATTTCATGCTG
NO1	7	30465155-30465254	ENSG00000167207	NO1	GGCCAGAAATTTTAAAGAGGGCTCCTAGGGCTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
NO2	16	50766583-50766682	ENSG00000192480	NO2	GCCAGAAATTTTAAAGAGGGCTCCTAGGGCTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
NO3	17	26114730-26114829	ENSG00000148430	NO3	TUCCCTGGGTCATATGCTCCCAAGTTGATGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
NOTCH1	9	139413925-139414017; 139417302-139417308	ENSG00000148400	NOTCH1	CTGCCAGTTCACGCGGCGAATCTGAGGAAATATGCGAATTTCTCGAAGAAACAACGCAGAACGGGGGCTGTCTGTGAGGAGGGGCTGCAACCT
NOTCH2	1	12049114-120491189; 120491630-120491683	ENSG00000134250	NOTCH2	AAGGACTCCAAATTTTAGAGGTATGCTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTGCTGTG
NTS2	6	8618105-18618143; 86184953-86184959	ENSG00000135318	NTS2	ATTCGGGTTTGAATGGATTAACACTG
PAX5	9	36839793-36839892	ENSG00000159692	PAX5	CTCCAGAGGACAATCTGGGAGATGCTG
PDCD1	2	242795002-242795101	ENSG00000188389	PDCD1	CTTCCAGGCTGTGCTGTGCTG
PDCD2	6	5107800-510803; 5524533-5524601; 5534745-5534751	ENSG00000197466	PDCD2	TGTCGACTGTGGCAAGTCTATCAAAATACAGCAATCTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PDR1	22	39620485-39620584	ENSG00000109311	PDR1	GATTCCTCTTCTG
PDRFB	5	149514473-149514572	ENSG00000113721	PDRFB	GTGGGCTTCCTCTG
PCAM1	HG183	PATCH6:2437740-6243775; 62440509-62440572	ENSG00000261371	PCAM1	ATCTGCATCGAGATTTGACAAAGTTGCAAGGAAGAACAACACTGAGTATGTTGAGTGAATGCTGCTCCAGCGCGAGTATCTGATGATGCTGCC
PGR	1	20713271-207112888; 207113894-207113841	ENSG00000182896	PGR	CTCCAGACTCTGCAAGACTGAGTATGTTGCTG
PLA2G2	1	20302086-20302185	ENSG00000182857	PLA2G2	TGCTGTGAGGAAGTGGGCTGTGGTGGGAAGTTTCTCTGTGAGGAAGTTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
PLA2G3	1	20246921-20246942; 20246791-20246866	ENSG00000187894	PLA2G3	GAGTGTGTGAGGAAGTGGGCTGTGGTGGGAAGTTTCTCTGTGAGGAAGTTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGGATTTCTGAGG
PLAUR	10	44165948-44165957	ENSG00000104432	PLAUR	GTGAGGAGGAGTTCGACACTG
PLM	15	74290357-74290456	ENSG00000104054	PLM	TGAGGAGGAGTTCGACACTG
POU2F2	19	42599586-42599614; 42599667-42599761	ENSG00000208277	POU2F2	GAGCAGGAAGAGAGGACAGAGTGTGCTG
PRAP	3	12458286-12458385	ENSG00000123170	PRAP	GAAGCAAGAGTGGCTCCTCCGCACTTTCAGGCGCTG
PBP1	4	74853026-74853091; 74853234-74853267	ENSG00000169736	PBP1	GGAAAGGACTTCAGCAAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGTGTGAGT
PRDM1	6	106547328-106547427	ENSG0000057657	PRDM1	CTCCCTCCCAACCAAGTCTGCTGTGCTG
PRK1	10	72357375-72357474	ENSG00000180644	PRK1	ACTGTTTGGGAGAGTGGGTTTACAGCTAACTCCGCTG
PRKDC	3	53223958-53224017; 53226124-53226163	ENSG00000163932	PRKDC	ACATAAATGCACTTCGCAAGAGGGGTTTACAGCTAACTCCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PSMB10	16	67970140-67970135; 67970301-67970324	ENSG00000205220	PSMB10	CACTGCTGTTGCTTCTGCAAGAGGGGTTTACAGCTAACTCCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PSMB5	14	23495185-23495284	ENSG00000109804	PSMB5	ACATGCTGTTGCTTCTGCAAGAGGGGTTTACAGCTAACTCCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PSMB7	9	127167604-127167703	ENSG00000136930	PSMB7	GTTCAATGTTGCAAGGCTTACATGGAAGGATTTGAGGAGGAGTGTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PSMB8	6	32808762-32808824; 32809308-32809344	ENSG00000204264	PSMB8	ACTACAGAGCACTATCTGAGAGGCTTTGCAATGTTACAGCAAGAGAGGTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGAT
PSMB9	6	32809808-32809911; 32809101-32809236	ENSG00000204065	PSMB9	TGAGGATTTGAGGAGGATTTGAGGAGGCTTTGCAATGTTACAGCAAGAGGATTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGAT
PSMC2	6	338234967-338235010; 33823631-33823668	ENSG00000130366	PSMC2	CTCCAGCTTTCTGAGGAGGATTTGAGGAGGCTTTGCAATGTTACAGCAAGAGGATTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGATTTGGGAGAGGAT
PTAF1	16	74338211-74338292; 74339187-74339204	ENSG00000193806	PTAF1	CCTACGGAAGTCTGAGAGGAGTACCAGCAATCTGCGGGCAGCAGCAAGAGTGGTCTAGAGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGG
PTAFR	1	2476525-2476624	ENSG00000169403	PTAFR	GGGGGGGGGAGATCAGATGCTACTACTCTG
PTGER4	5	40681884-40681962; 40691881-40691901	ENSG00000174522	PTGER4	GCTCAAAAGTGGGATCTTAACTCTTCTAATGCAAGAGGCTTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PTGS2	1	196646959-196646962; 186647393-186647488	ENSG00000273756	PTGS2	GTTCAAAAGTGGGATCTTAACTCTTCTAATGCAAGAGGCTTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
PTK2	8	141813731-141813749; 141828376-141828453; 141828979-141828981	ENSG00000169398	PTK2	GGTTCAGTGGTATTTCAGTGGGATGGAAGTGGGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PTPN2	18	12859232-12859253; 12884072-12884149	ENSG00000134242	PTPN2	CTCCAGCACTCCACCACTGAGAGGATTCAGAGGATTCAGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PTPN22	1	114357476-114357575	ENSG00000111679	PTPN22	TGTCGAGAGAGGAGGATTCAGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PTPRC	12	7069314-7069402; 7069507-7069517	ENSG00000081237	PTPRC	AGAGGAAATTTGTCTGTTGATGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PTPRC_all	1	198608227-198608267; 198608356-198608414	ENSG00000081237	PTPRC_all	PTTCCTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
CD4SR1	1	198608444-198608471; 198641476-198641502; 198672427-198672471	ENSG00000081237	CD4SR1	PTTCCTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
CD4SR3	1	198608444-198608471; 198641476-198641502; 19866584-19866588	ENSG00000081237	CD4SR3	PTTCCTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
CD4SR4	1	198608444-198608471; 198641476-198641502; 19866584-19866588	ENSG00000081237	CD4SR4	PTTCCTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
PCARD	16	12129209-12129310	ENSG00000194890	PCARD	CCTATGACGATTTTAAAGAGGGCTTCTTCTCAGTACAGCAAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
RAP1	3	12626452-12626480; 12626621-12626691	ENSG00000192155	RAP1	CTATGACGATTTTAAAGAGGGCTTCTTCTCAGTACAGCAAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
RAP2	11	36515169-3651635	ENSG00000176997	RAP2	CTATGACGATTTTAAAGAGGGCTTCTTCTCAGTACAGCAAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
RAG2	11	36515169-3651635	ENSG00000176997	RAG2	CTATGACGATTTTAAAGAGGGCTTCTTCTCAGTACAGCAAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
BARRE5	11	63813806-63813905	ENSG00000133321	BARRE5	GATGCTG
RELA	11	65427634-65427686; 65429158-65429204	ENSG00000174039	RELA	GATGCTG
RELB	19	45506225-45506254; 45513072-45513080; 45515194-45515254	ENSG00000104856	RELB	CATCTCGCCCTTCCAGGAGGATGAGTATGGGATCCCACTCCACATCCCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
RORC	1	151748324-151748310; 151785423-151785434	ENSG00000143365	RORC	CTCATACCTCCATCCAGGCTCTCAAGAAGTAAGTACAGACTGATCACTGAGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
RUNX1	21	36231859-36231875; 36252854-36252916	ENSG00000159216	RUNX1	CACTCATGAAGCAACTGCGATTTGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
S100A8	1	153362701-153362719; 153362871-153362951	ENSG00000163240	S100A8	CCATGAGGAAATTTCCATGCTCTCAGGAGTACCTGCGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
S100A9	1	153330790-153330889	ENSG00000163220	S100A9	AACATGAGACCATCACTCAACTCCCGAATCTTGGAGCTGGGGCAAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGATTTGAGGAGGAT
S101P1	1	10170542-101705341	ENSG00000170989	S101P1	CTGTGCTG
SELL	1	169696900-169696999	ENSG00000079988	SELL	ATCTACCTAACAAGTCTGTGCTTCTG
SEI	1	169696900-169696999	ENSG00000079988	SEI	ATCTACCTAACAAGTCTGTGCTTCTG
SEL1	1	169660327-169660426	ENSG00000184404	SEL1	CTAACCTGAAGTATTGGGGCTG
SERP1G	11	109015817-109015916	ENSG00000110876	SERP1G	CTAACCTGAAGTATTGGGGCTG
SERP1G1	11	57367415-57367514	ENSG00000149131	SERP1G1	CTAACCTGAAGTATTGGGGCTG





VTN	17	26894753-36894862	rs704(0.42)	ENSG00000109072; ENSG00000109072; ENSG00000123171	VIN; SEBOX	AGTGGCTGCTCTTCCAGGTGAAGGAGGAACTTGGGAGCCAACTATGATGACTACAGATGGACTGGCTTGTGCTGCCACTGTGMACCCIJA
XBP1	22	29192142-29192180; 29193065-29193125		ENSG00000100219	XBP1	GGAGTTAAGACACGGCTTGGGATGGATGCCTGGTGTGAAGAGGAGGGAGAGCCAGGGGAAATGAAGTGAAGGCAAGTGGCGGGCTGTGTGATGCC
XCL1	1	168550714-168550813		ENSG00000143184	XCL1	GTACTCTTTGGATGGTAAACATATGGAAAGATTCGGCTAGTGTGCTATCAGAGTGAAGCTATATCAATCTCTCTTGAAGTCCAGCTGTGTAATGG
XCR1	3	46062864-46062963	rs78405233 (0.00037)	ENSG00000173578	XCR1	GTCTCACTTCCGACACACTTCCCAAGGTGTTCCTCAAGGTGTTCCTCCACGACTCTGTAAGTCTCACTGGCTGCTCTTCCGAGTGGCAAGACCT
ZAP70	2	98351062-98351161		ENSG00000115085	ZAP70	GGAGCTAGGACAGCAAGCTCTTCCGAGCGCGTAACTCTCCTAGCTGCTGCACTTGAAGCTGCTGGCTGGGCACTTGGCTGAGTGGCAGGCGGCTG
ZBTB16	11	114027112-114027156; 114057674-114057728		ENSG00000109906	ZBTB16	TCCTGGATAGTTGGGGCTGAGAAATGCACTACTGCTCACTGAGGGGCTCAAGCCCTTGTGTGATCAATGAGGGCTGACAGTTCGAAAGGAGA
ZEB1	10	3180971-3180970		ENSG00000148516	ZEB1	TACAAAATGGGGTTCCTACTGGTGGGCCATACAGCAACCAAGTCTCTCAGGGATGGGCAAGCTGTTGTCTGCCAACAGCTGTGCTGCAAGAGC
ABCF1	6	30550920-30550965; 30551385-30551438		ENSG00000204574	ABCF1	GATGTCTCCCGCCAGCCAGTGTAGAAAATGCACTTGAACATCTGGAACTTGGCAAGCTTTTGGTGTGTGGAGGATACATCCCAAGGAGTCTGCTAATGGAC
ALAS1	3	52242212-52242263; 52245298-52245346		ENSG00000233330	ALAS1	GGGGATGGGATGGAGTCAAGCAAAATGGCAATCTTGGAACTTGGCAAGCTTTTGGTGTGTGGAGGATACATCCCAAGGAGTCTGCTAATGGAC
EEF1G	7	13279577-13279676		ENSG00000106594	CHCHD3	CTTCAATGACTGCATCTCACTCACTGCAATGTTCCAGCGATGGGAAGCTTGGCCAGTGTGATCTTTTGGAAACCAACAATAAC
EEF1G	11	62327592-62327665; 62327754-62327779		ENSG00000254772; ENSG00000255508	EEF1G; MIR3654	CTTCAATGACTGCATCTCACTCACTGCAATGTTCCAGCGATGGGAAGCTTGGCCAGTGTGATCTTTTGGAAACCAACAATAAC
GFPD	X	153761873-153761884; 153762250-153762337		ENSG00000160211	GFPD	ACACATGGCTGGTATTCTCCTCACCTGAGAGGCTTGGCAAGTGGCTGGGGGGCTATTCGATGATTTGGGATTCGGGATTCCTCCGGGAGTGGTA
GAPDH	7	6647095-6647162; 6647267-6647298	rs1803621 (0.39)	ENSG00000111640	GAPDH	CACTCCCACTTTTGAAGCTGGGGCTGGCATTTGCTCCACAGCACTTGTCAAGCTATTCTGGTATGCAAAAGTAATGGCAACAAGG
GUSB	2	6543521-6543533; 65439282-65439348		ENSG00000169919	GUSB	CGTCTGATGTGTGCTGTGGCAACAGCCTGGTCCACTAGAACTGCTGGCTACTTGAAGATGTTGATGCTGCACACAAATCTTGGAGCC
HPRT1	X	13360932-13360931		ENSG00000165704	HPRT1	TGTGATAGAGGATGGGAGCCATCACAATGACCCCTGTGCTCAAGGGGCTAAATCTTCTGACCTGCTGGATACATGAAAGACATG
OAZ1	19	2271473-2271529; 2271781-2271823		ENSG00000104904	OAZ1	GTTGGGGAGGGAATAGTGAAGGATCAGAACTTTTCACTTACTTCTCTGCTGATGGGCTGAATGTAAGAGGAAGCACTAACTCCAAAGCA
POLR1B	2	113332760-113332859		ENSG00000125630	POLR1B	GGAAACTGGCCTAGAATACTTGTGGATGTTAAGGCTGTGCTGACTAATTTCTATGGACCAGAGAGTTATATAGTCATCAGTGGGCTAGAA
POLR2A	17	7411719-7411794; 7412165-7412286	rs228132 (0.13)	ENSG00000181222	POLR2A	TTCGAGAGCCAAAGACTCTTCTGCTTCTCTTGGGGCAATGCTGCTGCTGATGGGAGGAGGAGGAGGATTTCTGTGCTGCTGCTGGAGCAT
PP1A	7	48841377-48841455		ENSG00000196262	PP1A	GGAAATTGAANAATGAGCAACACTGGCCATGCTGCTCACTGCTGTAATGATTAATCTGAGGCAAGACCCT
SDHA	5	224531-224630		ENSG00000173578	SDHA	TGGAGGGCAGGCTGGGAGGCTGATTGGCTTCTTGGAGCGGGTTTAAACAGCATGTTCACAGCTGTTCTACAGGTCACACTGTTGA
TBP	6	17087546-170873720; 170876006-170876030		ENSG00000112592	TBP	ACATGAACTCGTGTGAACCTGACTAAAGACCTTGGCTGTCGCCAAGCCGAATATATCCGAGGGTTCGCGGTATCATCAGAGA
TUBB	6	3069072-30690805; 30691117-30691132		ENSG00000196230	TUBB	TGGTGGATCAGAACTGGGACCAAGACTGTTGGCTAGGCTCTTTTGGCCAGATTTTGGACCAACAACCTTGTATTTGGTCAAGTGGGCGAGG
TUBB	8	3020971-30209816		ENSG00000127589	TUBB81	TGGTGGATCAGAACTGGGACCAAGACTGACTGTTGGCTGCTTGGCCAGCTTTTGGCCAGATTTAGACCAACAACCTTGTATTTGGTCAAGTGGGCGAGG
RP19	17	37359260-37359359		ENSG00000108298	RP19	CCATGCCGAATGCCAAGAGTGCATGGATGAGGAAATGAGGATTTTGGCCGGCTGCTCAGAAATACGCTGAACTTAAGAAAGTGCATGCCCA

\* - Excluded from analysis

HK - Used as house keeping genes for data normalization

Table 18 : Gene expression probes

List of genes analyzed by Nanostring nCounter technology including chromosome number, probe map position, SNPs in probe, Ensemble gene IDs, and probe sequence.

Remark : The original table has been split into 9 parts to facilitate reading.

## 5. Immune cell phenotype analysis and data normalization

### 5.1. Human material and staining protocol

Whole blood samples were collected from the 1,000 healthy, fasting donors from Milieu Interieur on Li-heparin, every working day from 8 to 11AM, 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 days on average after the first visit (range: 7 to 44 days). After receipt, samples were kept at room temperature prior to sample staining. Details on staining protocols can be found elsewhere (Hasan et al., 2015).

The same procedure has been followed for the replicative study (for a single visit) on a period going from March 2016 to June 2016 in South San Francisco, USA.

### 5.2. Reproducibility testing and assay development

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

### 5.3. Cytometric analyses

Ten 8-color flow cytometry panels were developed. Details on staining antibodies can be found in **Table 8** and gating strategies are described in **Figures 56-65**. 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 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, 107 cell counts, 89 MFI and 8 ratios. We excluded from subsequent analyses all cell proportions and 35 immunophenotypes that were measured several times on different panels and were exported for quality controls (**Table 9**). A total of 168 flow cytometry measurements were thus analysed, including 76 cell counts, 89 MFI and 3 ratios (**Table 9**). Problems in flow cytometry processing, such as abnormal lysis or staining were systematically flagged by trained experimenters, which resulted in 8.67% missing data among the 168,000 measured immunophenotype values.

For the replicative study, a preliminary set of three out of the 10 panels have been performed. The panels selected were for the characterization of Natural Killer cells (panel 4), of general populations (panel 5) and of polymorphonuclear cells (panel 7). The same protocol has been applied with some modifications due to differences in the instrumentation setting or reagent availability. First, due to non-commercially available anti-CD32 antibody coupled with the PE-Cy7 (custom tagging for the *Milieu Intérieur* study from Becton Dickinson, USA), in the panel 7, the anti-CD32-PE-Cy7 has been replaced by the same clone, same provider in APC and the anti-CD203c-APC has been replaced by the same clone coupled with PE-Vio770 (Miltenyi Biotech, Germany). All the antibodies were titrated on the instrument dedicated to the study (FacsCantoII, Becton Dickinson) and the new panel validated to identify the population of interest (i.e. eosinophils). At the difference of the MacsQuant instrument used for the French study, the cell count could not be directly assessed on the FacsCantoII. In this order, the clinical WBC differential total count was used as total cell number reference to calculate the children population counts. The analysis has been then performed using the same gating strategies and then exported in csv tables previously described using FlowJo 10.2.

T cell panel

Phenotypes:

- T cells (CD3<sup>+</sup>)
- CD4<sup>+</sup>
- CD4<sup>+</sup>CD8 $\alpha$ <sup>+</sup>
- CD8 $\beta$ <sup>+</sup>
- DN

- In CD4<sup>+</sup> and CD8 $\beta$ <sup>+</sup>:*
- Naïve CD27<sup>+</sup>CD45RA<sup>+</sup>
  - CM CD27<sup>+</sup>CD45RA<sup>-</sup>
  - EM CD27<sup>-</sup>CD45RA<sup>-</sup>
  - EMRA CD27<sup>-</sup>CD45RA<sup>+</sup>

- In each subpopulation:*
- HLA-DR MFI
  - CCR7 MFI

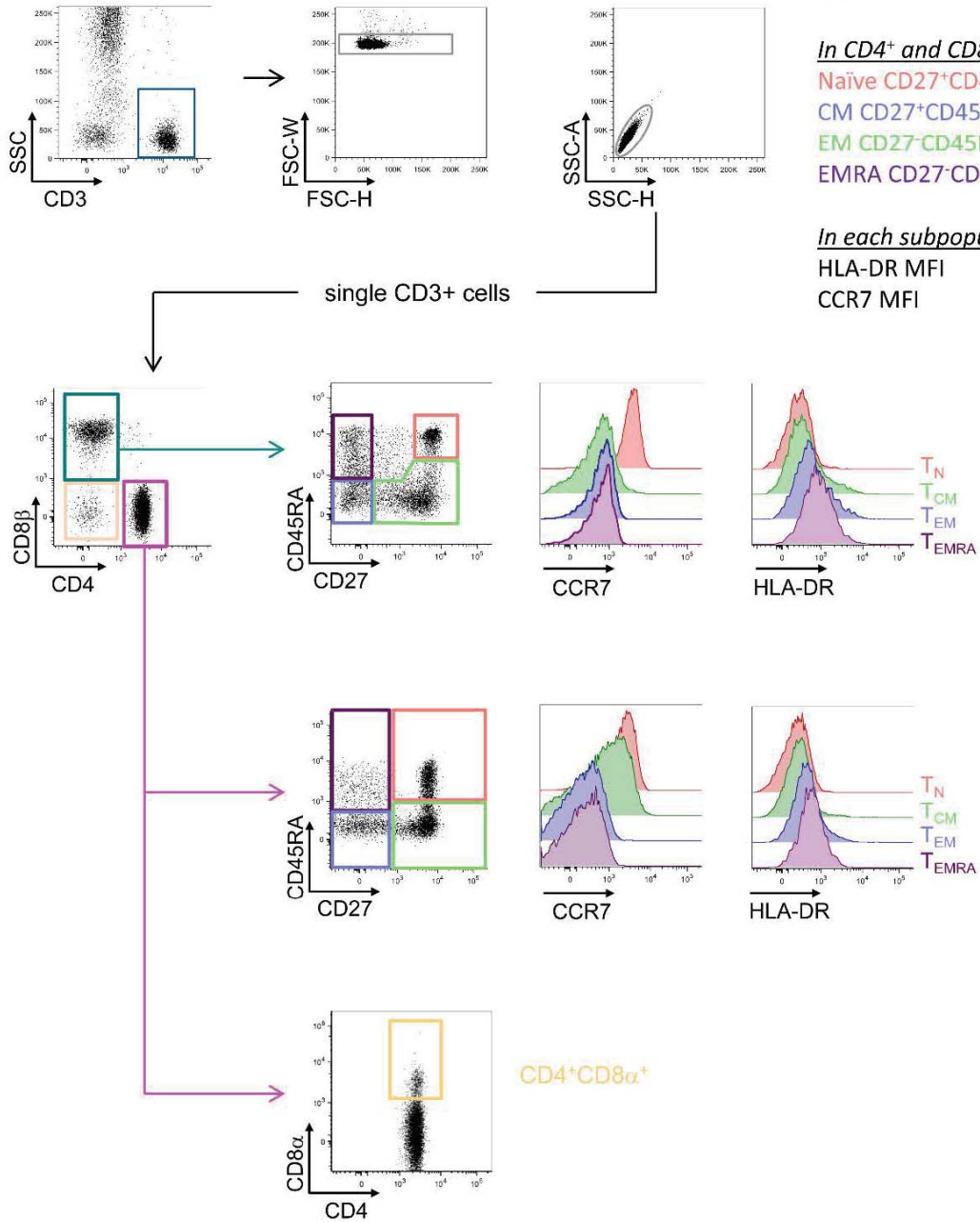


Figure 56 : Gating strategy for the T cell flow cytometry panel (Panel 1).

## Treg panel

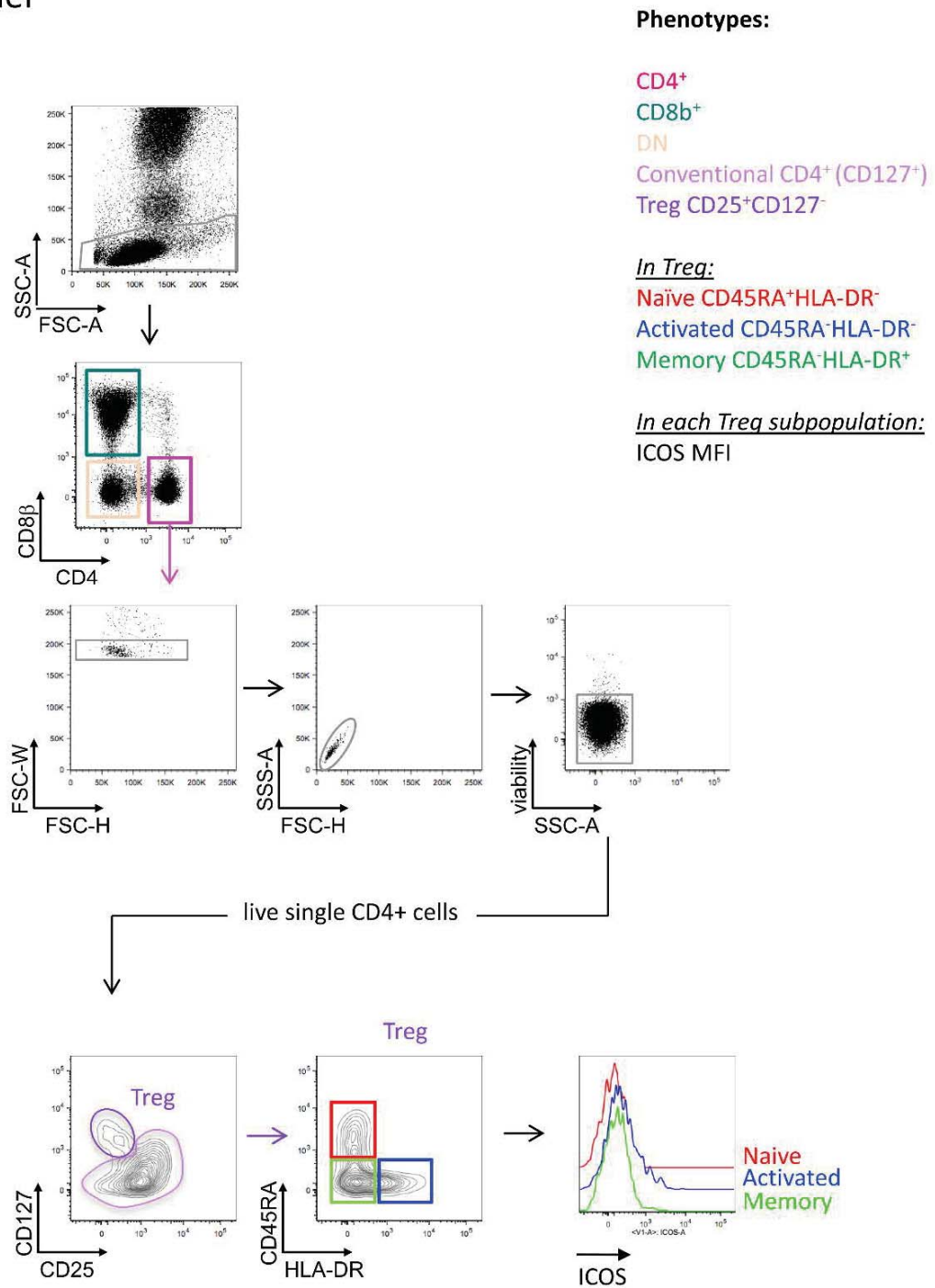


Figure 57 : Gating strategy for the *Treg* flow cytometry panel (Panel 2).

# MAIT/NKT panel

## Phenotypes:

- T cells (CD3<sup>+</sup>)
- TCR  $\gamma\delta$ <sup>+</sup>
- CD8 $\beta$ <sup>+</sup>
- CD4<sup>+</sup>
- DN

- In T cell subsets:*
- MAIT TCRV24<sup>+</sup>CD161<sup>hi</sup>
  - NKT TCRV7.2<sup>+</sup>CD161<sup>+</sup>

- In each MAIT & NKT subpopulation:*
- HLA-DR MFI

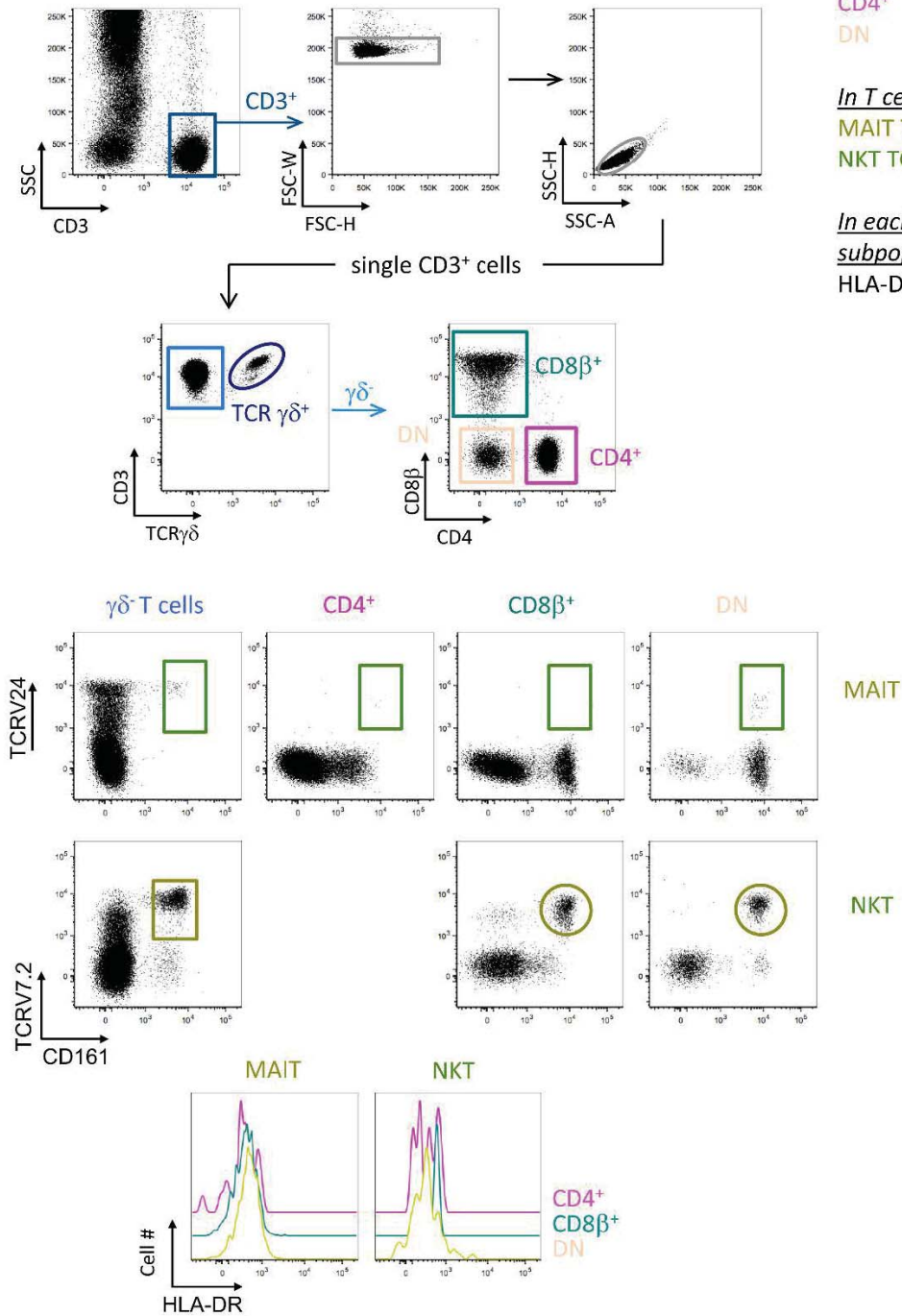


Figure 58 : Gating strategy for the MAIT/NKT flow cytometry panel (Panel 3).

# NK panel

## Phenotypes:

NK cells (NKp46<sup>+</sup>)  
 CD56<sup>hi</sup>CD16<sup>lo/-</sup>  
 CD16<sup>hi</sup>CD56<sup>lo</sup>

## In each NK subpopulation:

CD69 MFI  
 CD8 $\alpha$  MFI  
 HLA-DR MFI

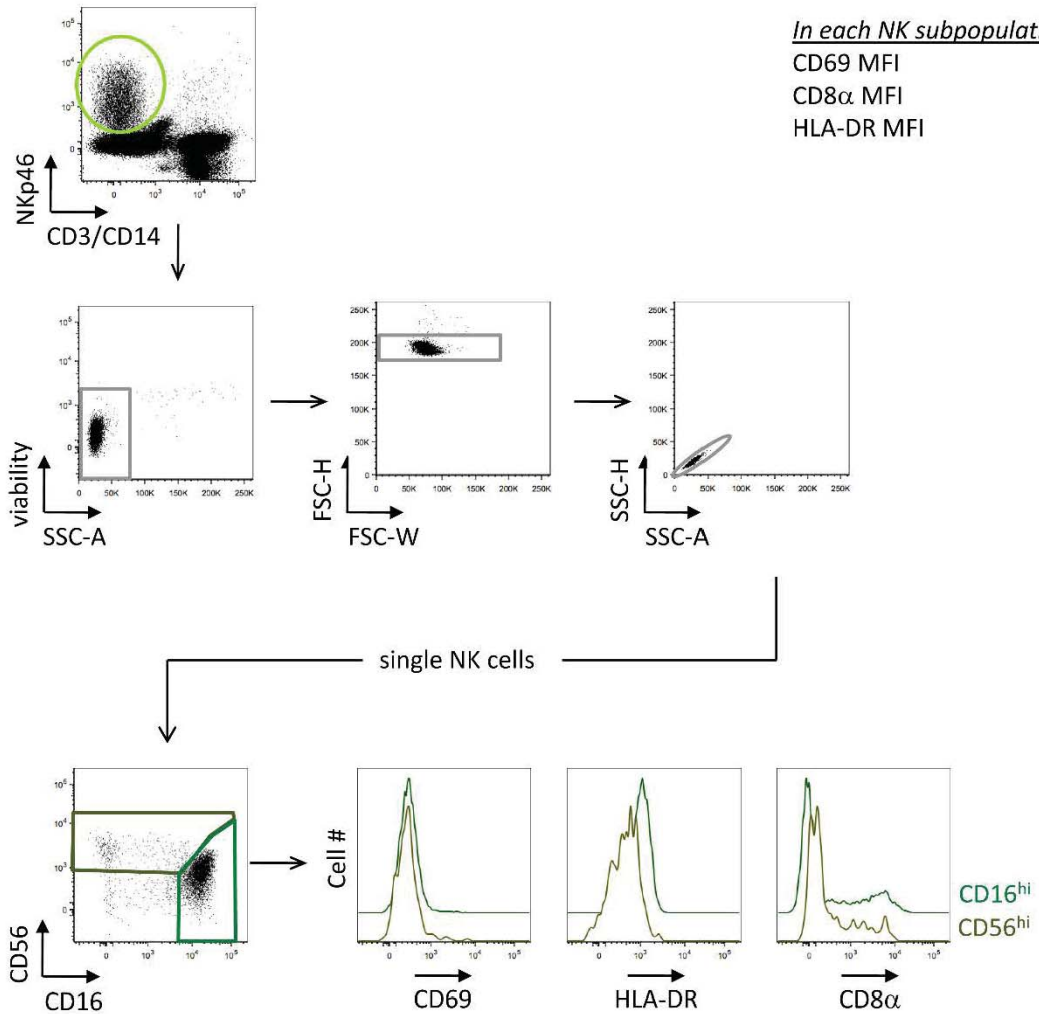


Figure 59 : Gating strategy for the NK flow cytometry panel (Panel 4).



Lineage panel

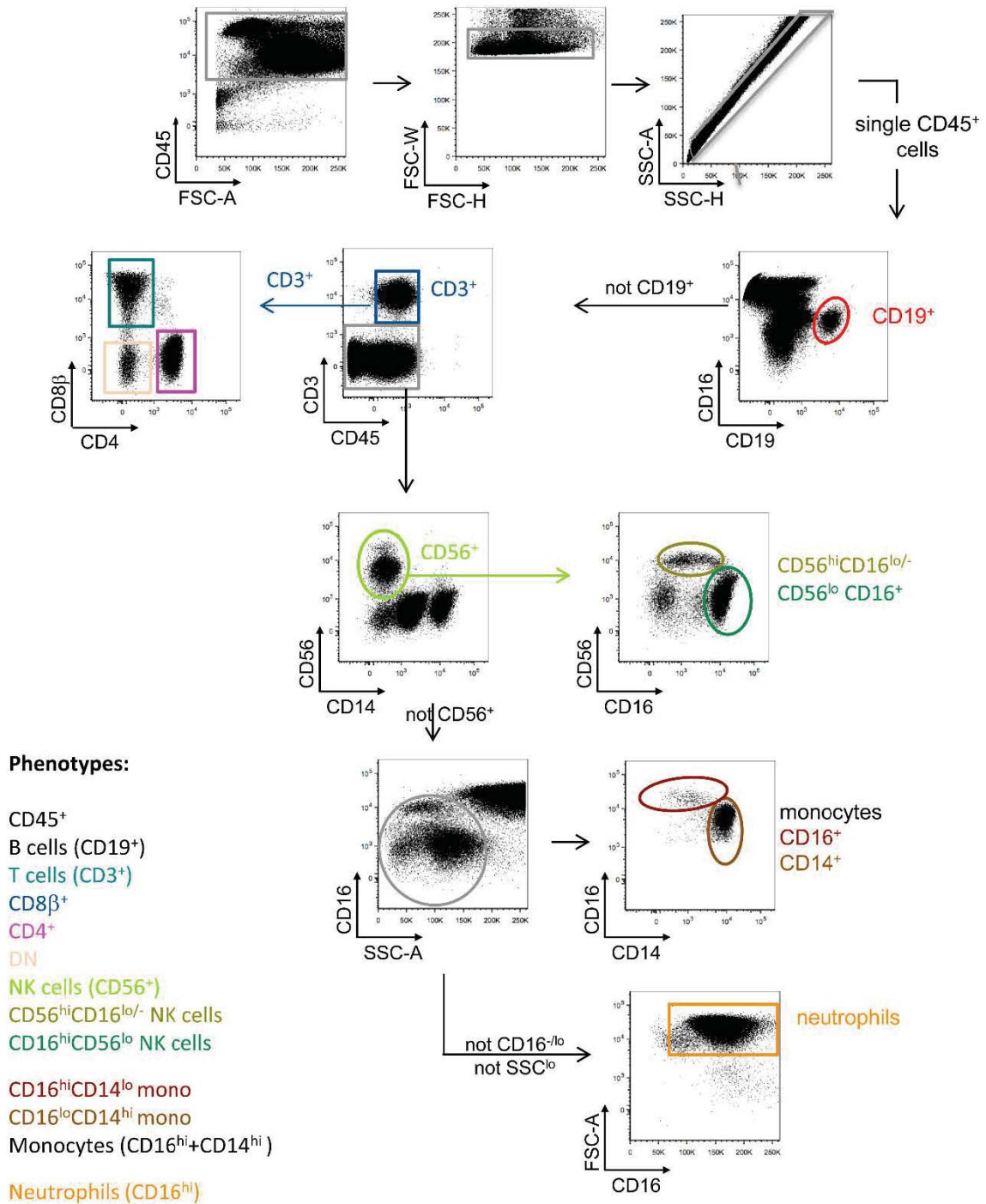


Figure 60 : Gating strategy for the lineage flow cytometry panel (Panel 5).

## B cell panel

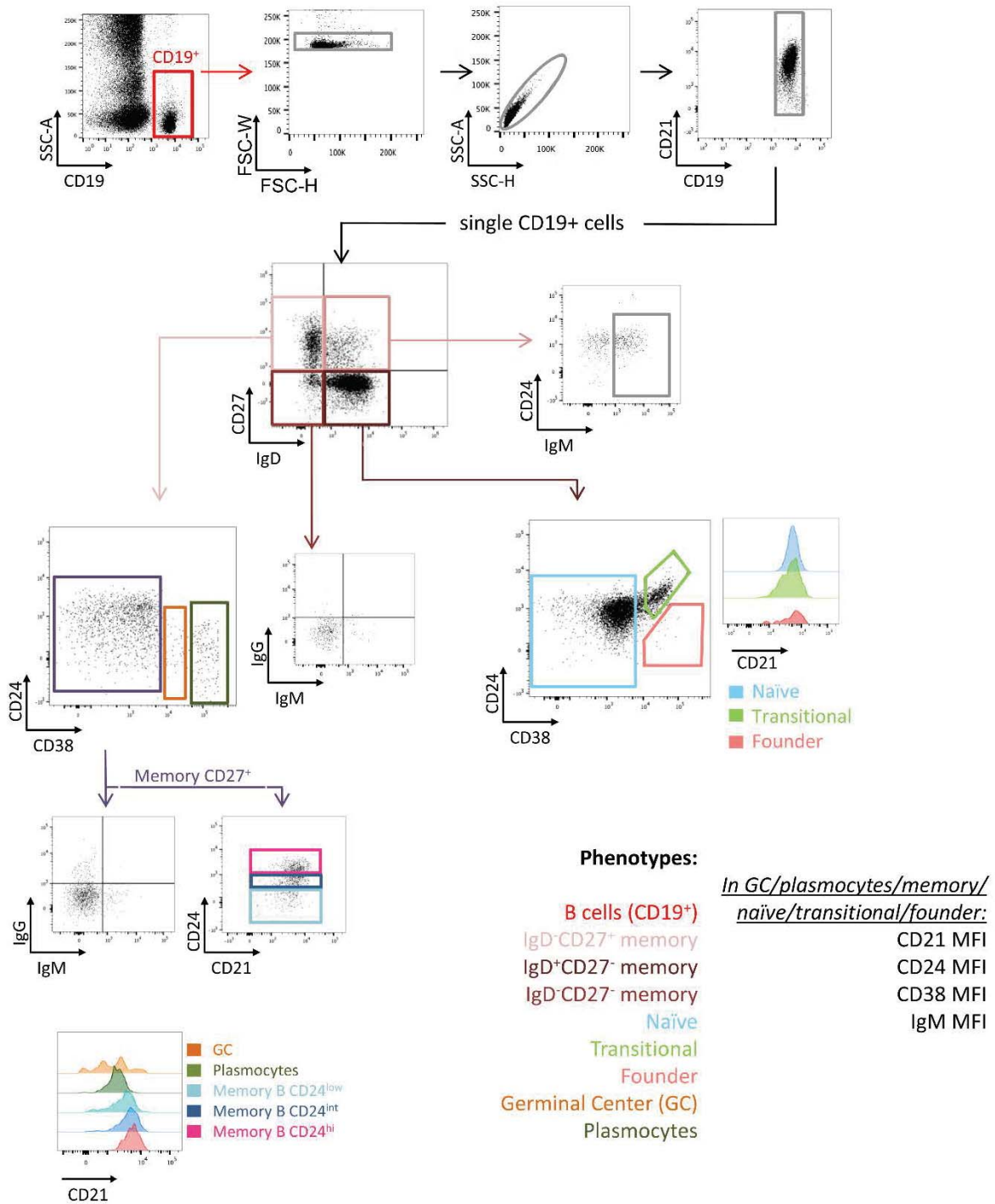


Figure 61: Gating strategy for the B cell flow cytometry panel (Panel 6).

## PMN panel

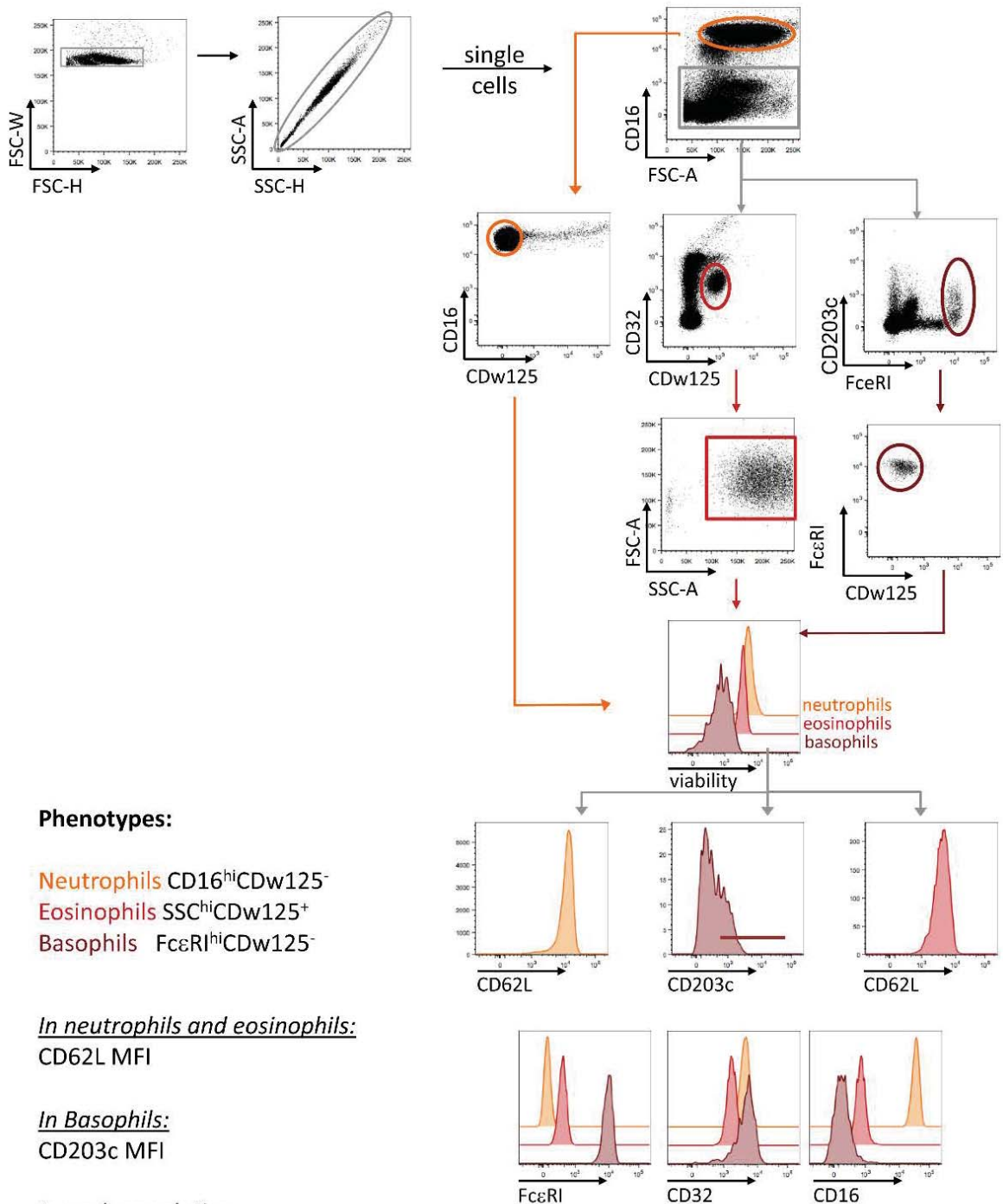


Figure 62 : Gating strategy for the PMN flow cytometry panel (Panel 7).

DC panel

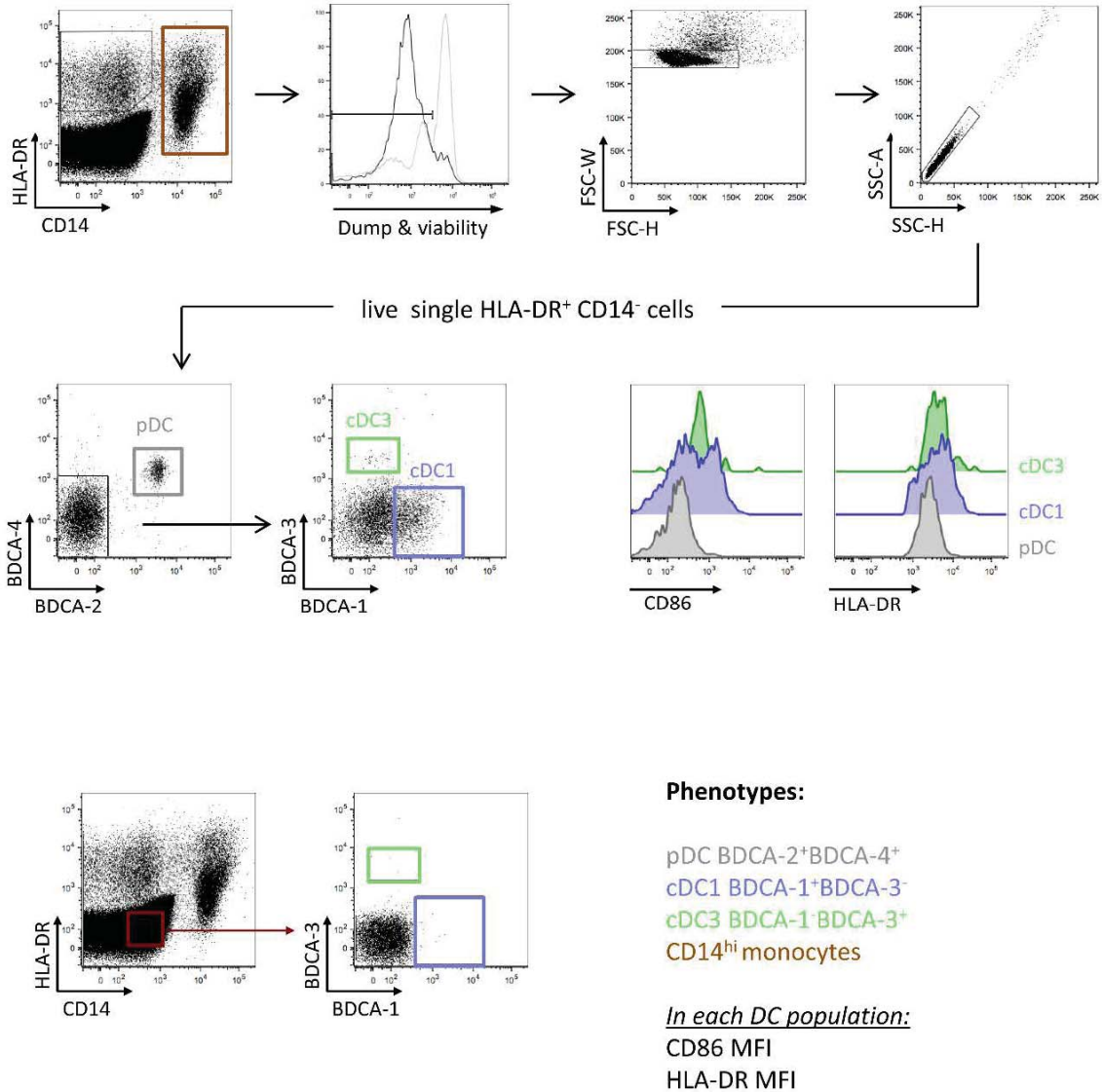


Figure 63 : Gating strategy for the DC flow cytometry panel (Panel 8).

Th panel

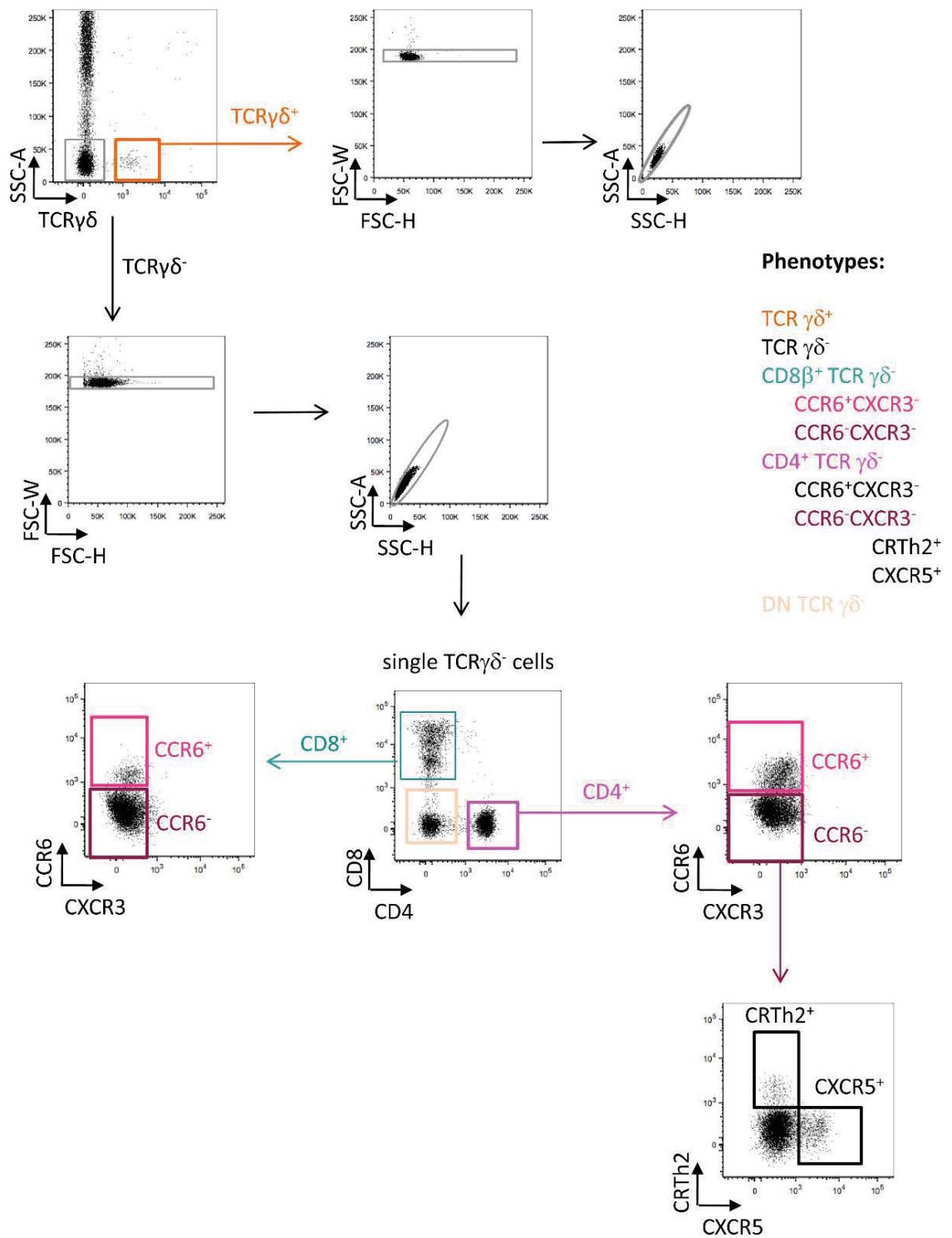


Figure 64 : Gating strategy for the *Th* flow cytometry panel (Panel 9).

## ILC panel

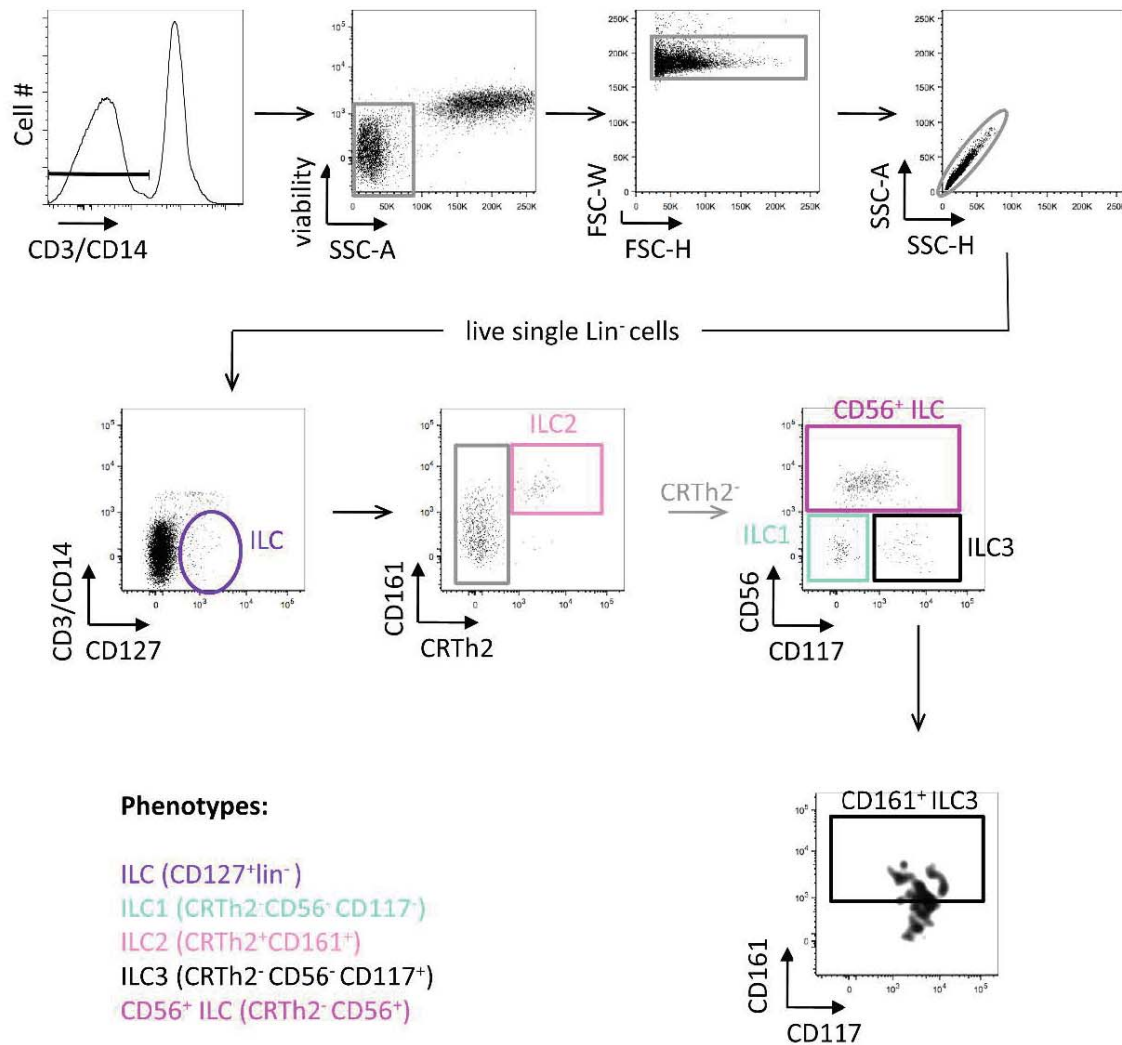


Figure 65 : Gating strategy for the ILC flow cytometry panel (Panel 10)

The following method sections have been applied on the *Milieu Interieur* dataset only.

## 5.4. Data transformation

### 5.4.1. 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 controlling immune cell levels, we removed these outlier values. Outliers were detected using a distance-based algorithm, chosen based on the high skewness of the distributions of flow cytometry measurements. A value was considered an outlier if the distance to the closest point in the direction of the mean of the distribution were more than 20% of the total range of the sample. All points more distant from the mean than these points were also considered as outliers. A total of 127 singleton values were removed using this method.

### 5.4.2. Adjustment for batch effects

Linear regression and visual inspection were used to evaluate the impact of batch effects on immunophenotypes. Two batch effects were considered: the day at which samples were processed (8 to 12 samples per day, from September 2012 to August 2013) and the hour at which blood samples were drawn (from 8h to 11h in the morning). We observed that sample processing day has a substantial impact on MFIs, while hour of blood draw impacts a limited number of cell counts, mainly CD16<sup>hi</sup> NK cells (**Figure 33**). We adjusted all cell counts for the effect of the hour of blood draw using linear regression, and adjusted MFIs for the effect of processing days using the ComBat non-parametric Bayesian framework (Johnson and Li, 2007). This algorithm adjusts for batch effects by leveraging multivariate correlations among response variables. We did not include variables of interest in the ComBat model, because none was significantly different across sample processing days, with the exception of smoking ( $P = 0.002$ ). To ensure the highest precision, we imputed MFI immunophenotypes prior to ComBat correction, using the random forest based missForest R package (Stekhoven and Bühlmann, 2012).

## 6. Statistical analysis, Data visualization and Softwares

### 6.1. Dimension reduction: Principal Component Analysis

Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) are commonly used method for identifying the smallest number of uncorrelated variables (principal components) that explain the maximum amount of variance where the eigenvalues are the variances of the principal components.

For a comprehensive overview of PCA and the exploratory analysis using dual PCA and the accompanying PCA biplots we refer to Fontes (Fontes, 2012). Before applying PCA, the variables (mRNA expression levels) were log transformed, mean centered per donor, to avoid inter-donor variability obscuring inter-stimuli responses, and finally the variables were scaled to unit variance. The mean centering per donor is in accordance with the paired structure in the data and paired t-tests or ANOVA were performed throughout. Scaling to unit variance prevents large variances in the data from obscuring the correlation structure in the data. Q values, which are defined as false discovery rate (FDR) adjusted p values (Benjamini et al., 1995), were used to define statistical significance. Plots were exported from the software Qlucore Omics Explorer 3.1.

### 6.2. Dimension reduction: factor analysis

The goal in factor analysis is to explain the covariance between the variables for example using a correlation matrix from a separate analysis. This approach allows to standardize the variables measured by different scales or if the variances differ widely between variables.

Correlation circles were generated by computing the median value across the 25 donors, for each of the considered 44 genes; we then transposed the data matrix so as to consider the four stimulation conditions as the four PCA dimensions; finally, the vectors representing the TLR stimuli (generated accordingly) were projected onto the four-dimensional PCA. The respective 2D PCA projection plots were made with the R package 'FactoMineR' (version 1.28) to compute PCA scores and projected coordinates.



### 6.3. Silhouette score analysis

Silhouette analysis was used to study the separation distance among the TLR and microbial stimuli. The k-means clustering was performed using the Open CV library (Bradski and Kaehler, 2008); with settings equal to 100 iterations and 500 attempts and the silhouette scores were computed (Steinhaus, 1956 ; Bradski and Kaehler, 2008). Cluster number was selected based on the number of stimuli represented in the PCA (k=7 for TLR, k=8 for microbes). Note, silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters; a value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

### 6.4. Bootstrapped hierarchical clustering

Bootstrapped hierarchical clustering analysis was performed using the 'pvclust' R package (version 1.3-2) using a Spearman-based dissimilarity metric. 1000 trees were sampled to evaluate the robustness of each cluster. Correlation matrices were plotted using the R graphics package ggplot2 (version 1.0.0) on the R platform (version 3.1.1).

### 6.5. Support vector machine

The goal of linear support vector machines is to try to differentiate two groups of points using a hyperplane. Once the algorithm is performed, we can identify the most important variables that help the most for the differentiation of both groups, using the normal vector that defines the obtained hyperplane: the variables with highest loadings (in absolute value) are the most important.

Stimulus signatures, consisting of gene lists specific for each of the 4 cytokines, were created by training a SVM for each individual stimulus vs. null. This approach was used to define stimulus signatures set by a discrete number of variables. In order to discover reasonably complex gene interaction networks among the 4 stimuli, the SVMs were optimized from 12-57 gene subsets (approximately 2%-10% of the total number of genes). For all 4 cytokine

stimuli the optimal classifier determined by the SVM cross validation scheme corresponded to the smallest gene set size. The identified gene lists had perfect accuracy upon 10 repeated complete SVM test runs. This shows that a small number of selected variables can predict the specific stimulus used. The small overlap between the established stimulus signatures indicates that the selection strikes a reasonable balance between capturing the complexity in the data and at the same time identifying those important individual genes. The open source C++ software library OpenCV was used to build and evaluate the SVMs (Burges, 1998; Chang and Lin, 2011). For comparison, a kNN classifier was also tested, using the implementation in the OpenCV library with default parameter settings, which gave exactly the same stimulus signatures.

## **6.6. Non-genetic factors analysis on immune cell phenotypes**

### **6.6.1. Correlation matrix of cell counts**

The correlation matrix among cell counts (**Figure 35**) was estimated using the sample correlation matrix on the residuals of the immunophenotypes from a regression model that included non-genetic covariates selected using the stability selection algorithm (described below) on the 40 demographic covariates, together with the batch variables. The cell counts were modeled using a log-normal distribution.

### **6.6.2. Stability selection of analyzed environmental variables**

A total of 147 variables were chosen from the demographic variables available in the *MI* cohort (Thomas et al., 2015), based on their relevance to the immune traits. These included, among others, variables related to smoking, CMV infection, vaccination history, childhood diseases, nutrition and lifestyle, clinically related variables, and sociological variables such as income and education. Of these, 40 variables were chosen for confirmatory cross-sectional analysis (**Table 7**). These variables were selected based on their distribution (i.e., categorical variables with only rare levels, such as infrequent vaccines, were excluded) and on their levels of correlation with other demographic variables (e.g., height and BMI).

The dependency matrix among the 40 demographic variables (**Figure 39**) 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.

### 6.6.3. Analysis of the impact of non-genetic factors on immunophenotypes

We ran an ANOVA for each of the 40 non-genetic treatment variables and each immunophenotype. Thus, a total of 6,720 models were fitted, and tests were performed. Models were fitted to complete cases. We considered all of the tests as one multiple testing family and we used the false discovery rate (FDR) as error rate. We did not attempt to make a causal model for all variables and we thus kept the amount of controls small to ensure interpretability of the study, and to make it easier to reproduce. We included age, sex and CMV seropositivity as controls for all models (**Figure 40**), except when they were the treatment variable to be tested (**Figure 37**). Age and gender were included as covariates because they affect many of the other non-genetic variables and because it is reasonable to assume that they are not affected by any of the other non-genetic variables, and their inclusion will thus not give rise to any spurious correlation. CMV seropositivity was included because it has been shown to strongly affect some immunophenotypes in the literature. We also included as covariates genome-wide significant SNPs for corresponding immunophenotypes (**Table 10**). We assume that the genetic variables (i.e., genome-wide significant SNPs) do not affect the non-genetic variables, and are not affected by the non-genetic variables or the immunophenotypes, and were thus included to reduce the residual variance of the models.

The response variables analyzed consisted of immune cell counts, ratios of immune cell counts, and cell marker mean fluorescence intensities (MFI). For the immune cell counts, we considered two types of error distributions: log-normal and negative binomial. We chose the log-normal distribution based on residual plots and AIC measures. Values less than or equal to zero were considered to be missing. This excluded 1,208 points, 990 of which came from only three immunophenotypes: HLA-DR<sup>+</sup> CD56<sup>hi</sup> NK cells, HLA-DR<sup>+</sup> CD4<sup>-</sup> CD8<sup>-</sup> MAIT cells and the number of HLA-DR<sup>+</sup> CD4<sup>-</sup> CD8<sup>-</sup> NKT cells. For the ratio of cell counts and MFIs we considered normally- and log-normally distributed residuals errors. We chose to model the residuals errors as being normally distributed, also based on AIC measures and residual plots. We considered a test as significant if the FDR was less than 1%. If a test was

significant, we estimated the standard errors of the tested parameter using robust sandwich estimation and we constructed false coverage rate adjusted confidence intervals. Sandwich estimation was performed using the sandwich package in R.

## **6.7. Genetic-related analysis on circulating cell population parameters**

### **6.7.1. 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).

HumanExome SNP call rate was lower than 97% in 11 donors, which were thus removed from this dataset. We filtered out from both datasets SNPs that: (i) were unmapped on dbSNP138, (ii) were duplicated, (iii) presented a low genotype clustering quality (GenTrain score  $< 0.35$ ), (iv) presented a call rate  $< 99\%$ , (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 and 87,960 SNPs for the HumanOmniExpress and HumanExome BeadChips, respectively. The two datasets 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  $< 99\%$ . 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%, validating that no problem occurred during DNA sample processing. The final dataset included 732,341 QC-filtered genotyped SNPs.

### **6.7.2. 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 are not identical-by-state between all possible pairs of subjects, obtained with KING (Manichaikul et al., 2010). Genetic structure was estimated with the Principal Component Analysis (PCA) implemented in EIGENSTRAT

(Patterson et al., 2006). For comparison purposes, the analysis was performed on 261,827 independent SNPs and 1,723 individuals, which include the 1,000 *Milieu Intérieur* subjects together with a selection of 723 individuals from 36 populations of North Africa, the Near East, western and northern Europe (Behar et al., 2010).

### 6.7.3. Genotype imputation

Prior to imputation, we phased the final SNP dataset with SHAPEIT2 (Delaneau et al., 2013) using 500 conditioning haplotypes, 50 MCMC iterations, 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 that have C/G or A/T alleles. Genotype imputation was performed by IMPUTE v.2 (Howie et al., 2009), 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  $\leq 0.8$ , 43,737 duplicated SNPs, 955 monomorphic SNPs, and 449,903 SNPs with missingness  $>5\%$  (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.

### 6.7.4. Genome-wide association analyses

Prior to genome-wide association analyses we imputed all missing values of cell count immunophenotypes using the random forest-based missForest R package. MFI immunophenotypes were already imputed prior to batch correction. We included as a covariate in the GWAS any non-genetic variable that was confidently identified as associated with a specific immunophenotype. These covariates were imputed prior to this analysis to reduce the large loss of sample size, which would have occurred if individuals with missing samples was removed for all 40 covariates.

To select non-genetic variables, we used stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) with elastic net regression as support estimator. For each immunophenotype, we considered a linear model including each of the 40 variables ([Table](#)

7). We estimated for each variable the probability of the parameter corresponding to that variable being zero when optimizing the elastic net cost function. The variables were then chosen to be controls in the GWAS by thresholding this probability. It has been shown that this procedure, with the right threshold, controls the false discovery rate of selected variables (Shah and Samworth, 2013). We assumed that no covariate is affected by individual genotypes and the purpose of this control is thus to increase the power of the GWAS by reducing the residual variance.

To reduce the risk of false positives, we transformed immunophenotypes to normality conditional on the selected non-genetic covariates. Immunophenotypes with non-positive values were transformed using a Yeo-Johnson transformation and immunophenotypes with only positive values were transformed using a Box-Cox transformation. The transformations were performed using modifications of the functions in the *car* package in R.

Univariate genome-wide association study was conducted for each imputed, transformed immunophenotype using the linear mixed model implemented in GEMMA (Zhou et al. 2014). Genetic relatedness matrices (GRM) were estimated for each chromosome separately, using the 21 other chromosomes, to exclude from the GRM potentially associated SNPs (i.e., "leave-one-chromosome" approach; see Yang et al., 2014). A conditional GWA analysis was also carried out for each of the 14 immunophenotypes that showed the strongest genome-wide significant signals ("main immunophenotypes" in **Table 10**), by including as a covariate the genotypes of the most strongly associated SNP. A multivariate GWAS was conducted on a set of 6 candidate immunophenotypes (i.e., number of HLA-DR<sup>+</sup> memory T cells), using GEMMA linear mixed model adjusted with covariates that were selected for all the six traits. For all genome-wide association analyses, a conservative genome-wide significant threshold of  $P < 10^{-10}$  was used, to account for testing multiple SNPs and immunophenotypes.

We confirmed the power to identify large-effect genotype-phenotype associations in the Milieu Intérieur cohort by replicating well-known genetic associations with non-immune traits, including OCA2/HERC2 genes with eye and hair color (rs12913832,  $P = 6.7 \times 10^{-138}$  and  $8.5 \times 10^{-18}$ , respectively), SLC45A2 with hair color (rs16891982,  $P = 3.2 \times 10^{-9}$ ), UGT1A gene cluster with bilirubin levels (rs6742078,  $P = 2.6 \times 10^{-75}$ ), SLC2A9 with uric acid levels (rs6832439,  $P = 4.3 \times 10^{-14}$ ), and CETP with HDL levels (rs711752,  $P = 4.5 \times 10^{-8}$ ).

We included as a covariate in the GWAS any non-genetic variable that was confidently identified as associated with a specific immunophenotype. To select these non-genetic variables, we used stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) with elastic net regression as support estimator. Immunophenotypes with

non-positive values were transformed using a Yeo-Johnson transformation and immunophenotypes with only positive values were transformed using a Box-Cox transformation, implemented in the *car* R package. Univariate genome-wide association study was conducted for each imputed, transformed immunophenotype using the linear mixed model implemented in GEMMA (Zhou and Stephens, 2014). A conditional GWAS analysis was then carried out for each of the 14 immunophenotypes that showed a genome-wide significant signal, by including as a covariate the genotypes of the most strongly associated SNP. A multivariate GWAS was conducted on a set of 6 candidate immunophenotypes (i.e., number of HLA-DR<sup>+</sup> memory T cells), using GEMMA linear mixed model adjusted with covariates that were selected for all the six traits. For all genome-wide association analyses, a conservative genome-wide significant threshold of  $P < 10^{-10}$  was used, to account for testing multiple SNPs and immunophenotypes.

## 6.8. HLA typing and association tests

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

## 6.9. Expression quantitative trait loci analysis

The nCounter® Human Immunology v2 gene code set (NanoString technologies) was used to measure gene expression in non-stimulated whole blood of the 1,000 *Milieu Intérieur* subjects. The Human Immunology v2 gene code set was selected, as it covers most of the immunity-related genes found in the genomic regions identified by our genome-wide association study. This data will be described in details in a separate work (B.P., A.U., L.Q.-M., M.L.A., unpublished data). As previously indicated, expression probes that bind to cDNAs in which at least 3 known common SNPs segregate in humans were removed from the analyses (including HLA-B, HLA-DRB1, HLA-DQA1 and HLA-DQB1).

After quality control filters described previously, mRNA levels were available for 984 individuals at 52 candidate genes, *i.e.*, immunity-related genes in a 1-Mb window around the 14 genome-wide significant associations identified in this study. For each sample, probe counts were  $\log_2$  transformed, normalized and adjusted for batch effects (data not shown). For the 52 candidate genes, expression quantitative trait loci (eQTL) mapping was performed in a 1-Mb window around corresponding association signals, using the linear mixed model implemented in GEMMA (Zhou and Stephens, 2014), in the same conditions as for immunophenotypes.

### 6.10. Decomposition of the proportion of variance explained

For each immunophenotype, we included all significant non-genetic factors (**Figure 40A**) and both genome-wide significant ( $P < 10^{-10}$ ) and suggestive ( $P < 5 \times 10^{-8}$ ) genetic factors in a unique linear model. The contribution of each significant variable 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. The difference in contribution to explained variance between innate and adaptive MFIs was tested using a generalized least squares model, where we used overall genetic contribution to explained variance per MFI as response variable, and had an indicator variable for if the MFI variable was innate or adaptive. The covariance matrix between the response variables was estimated using the sample covariance among the MFI immunophenotypes.

## 7. Data mining tool: R Shiny (Interactive Web Application) development

To complement this work interactive web applications are provided that allows an extensive exploration of the dataset presented in these studies.

Concerning the transcriptomic analysis, the application presents four different types of analytical visualizations: PCA, boxplots, hierarchical clustering and a searchable reference table. Each visualization contains default settings that match the figures presented in the manuscript. Visualization controls enable the user to navigate the entire dataset following



their own scientific interests. The interactive table provides reference values, based on the 25 healthy donors, which can be directly browsed using a selected method (median expression values, coefficient of variations or q-values from paired t-tests as compared to the Null condition). More details about each data visualization tool can be found within the web application. The application was implemented using the Open Source R platform. It makes use of its Shiny package (version 0.12.2), ggplot2 package (version 1.0.0), dplyr package (version 0.4.3) and tinyr package (version 0.3.1). All visualization and analysis methods are accessible through a web browser, without the need to install any additional software, or possess knowledge of a programming language. The application is available at: <https://www.synapse.org/MilieuInterieur> (<http://dx.doi.org/10.7303/syn7059574>).

For the second study, we provide the distribution, ranges and statistics of all batch-corrected immune cell counts (**Table 9**), thereby facilitating comparisons with cytometry data collected as part of routine clinical practice. Values can be accessed through a user-friendly web application ([http://104.236.137.56:3838/LabExMICytometryBrowser\\_ShinyApp/](http://104.236.137.56:3838/LabExMICytometryBrowser_ShinyApp/), draft Shiny application available for review), which can be queried based on personal characteristics such as age or gender.

**LIST OF ILLUSTRATIONS**

**LIST OF TABLES**

## List of illustrations

Figure 1 : Activator signals and function profile of NK cells, ILC1, ILC2 and ILC3 subtypes.....	23
Figure 2 : Platelets and leukocytes as partners in innate immunity. ....	25
Figure 3 : Subsetting CD4 <sup>+</sup> T cell responses based on T <sub>H</sub> cell polarization.....	30
Figure 4 : TNF Signals through TNFR1 and TNFR2 receptors.....	36
Figure 5 : Interferon receptors and activation of classical JAK–STAT pathways by type I and type II interferons. ....	39
Figure 6 : Schematic representation of receptor structures and main signaling pathways of the PRR families. ....	41
Figure 7 : Cellular localization of major PRRs families.....	45
Figure 8 : Schematic overview of workflow from blood draw to gene expression analysis .....	59
Figure 9 : Quality control measures for gene expression analysis.....	60
Figure 10 : Internal control evaluation for gene expression analysis.....	61
Figure 11 : Distinct gene expression signature induced by cytokine stimulation.....	63
Figure 12 : IFNA and IFNB show overlapping gene expression profiles.....	64
Figure 13 : Silhouette score comparison from the clusters identifying the cytokine stimulation induced response using 572 gene versus 44 genes .....	65
Figure 14: Distinct gene expression signature induced by cytokine stimulation.....	67
Figure 15 : Interactions and outlier responses among the cytokine-induced gene expression signatures (1) .....	69
Figure 16 : Interactions and outlier responses among the cytokine-induced gene expression signatures (2) .....	70
Figure 17 : Projection of TLR stimuli onto PCA analysis as defined by 4 effector cytokines.....	72
Figure 18 : TLR induced gene expression can be represented as a function of cytokine-induced gene signatures.....	73
Figure 19 Four cytokine gene induced expression upon TLR stimulation .....	74
Figure 20 : Distinct and variable response to TLR agonist stimulation can be captured using the cytokine induced 44 gene signature .....	75
Figure 21 : Comparison of the silhouette score from the clusters identifying the response to TLR agonists using 572 gene versus 44 gene signature .....	76
Figure 22 : Distinct and variable response to microbial stimulation can be captured using the cytokine induced 44 gene signature .....	77
Figure 23 : Comparison of the silhouette score from the clusters identifying the response to microbial stimulation using 572 gene versus 44 gene signature. ....	78
Figure 24 : Correlation among TNFSF10 stimulus-induced gene expression helps to trace cytokine loops .....	79
Figure 25 : Correlation among IRAK3 stimulus-induced gene expression helps to trace cytokine loops .....	80
Figure 26 : Microbial induced lymphokines are absent from TLR and cytokine gene expression signatures.....	82
Figure 27 : Gene expression patterns not captured by four effector cytokine induced changes.....	83
Figure 28 : Graphical abstract summarizing the study 1.....	87
Figure 29 : Immune cells and cell-surface markers measured in the <i>MI</i> cohort .....	101
Figure 30 : Distributions of 169 raw immunophenotypes measured in the <i>MI</i> cohort .....	102
Figure 31 : Repeatability and reproducibility of semi-automated flow cytometry measurements .....	103
Figure 32 : Reproducibility of semi-automated flow cytometry measurements.....	104
Figure 33 : Sample hour effect on measured immunophenotypes .....	105
Figure 34 : Effect of sampling over time on measured immunophenotypes .....	106
Figure 35 : Transformed distributions of the 168 immunophenotypes studied in the <i>MI</i> cohort. ....	107
Figure 36 : Correlations among circulating levels of immune cell populations in the <i>MI</i> cohort. ....	109
Figure 37 : Age, sex and CMV infection impact immune cell counts in healthy individuals. ....	111

Figure 38 : Age and sex impact protein levels of immune cell markers in healthy individuals. ....	112
Figure 39 : Correlation matrix among 40 demographic variables analyzed in this study.....	114
Figure 40 : Tobacco smoking strongly impacts immune cell counts in healthy individuals. ....	115
Figure 41 : Smoking strongly impacts protein levels of immune cell markers in healthy individuals. .....	116
Figure 42 : Genetic relatedness and structure in the <i>MI</i> cohort. ....	119
Figure 43 : Genome-wide significant associations with immunophenotypes measured in the <i>MI</i> cohort. .....	120
Figure 44 : Local association signals for the 14 genome-wide significant hits associated with immunophenotypes measured in the <i>MI</i> cohort. ....	129
Figure 45 : Local association signals for the 7 genome-wide significant hits identified by conditional GWAS of the 14 immunophenotypes showing strong genetic association in the <i>MI</i> cohort.....	134
Figure 46 : Local association signals detected by multi-trait GWAS of immunophenotypes measured in the <i>MI</i> cohort. ....	134
Figure 47 : Proportion of variance of innate and adaptive cell parameters explained by non-genetic and genetic factors. ....	137
Figure 48 : Antibody titration for flow cytometry analysis of the replicative study.....	140
Figure 49: Modified flow cytometry panels used for the replicative study. ....	141
Figure 50 : Method comparison for cell count assessment. ....	142
Figure 51: Comparison of monocyte and lymphoid population frequency across studies. ....	144
Figure 52 : SNPs controlling circulating immune cell parameters. ....	146
Figure 53 : Schematic representation of donor recruitment for the <i>Milieu Intérieur</i> study. ....	162
Figure 54 : Distinct protein inflammatory signatures for whole microbe, TLR and cytokine stimulation.....	165
Figure 55 : Gene expression pathways used to select NanoString Immunology panel. ....	171
Figure 56 : Gating strategy for the <i>T cell</i> flow cytometry panel (Panel 1). ....	188
Figure 57 : Gating strategy for the <i>Treg</i> flow cytometry panel (Panel 2). ....	189
Figure 58 : Gating strategy for the <i>MAIT/NKT</i> flow cytometry panel (Panel 3). ....	190
Figure 59 : Gating strategy for the <i>NK</i> flow cytometry panel (Panel 4). ....	191
Figure 60 : Gating strategy for the <i>lineage</i> flow cytometry panel (Panel 5). ....	192
Figure 61: Gating strategy for the <i>B cell</i> flow cytometry panel (Panel 6). ....	193
Figure 62 : Gating strategy for the <i>PMN</i> flow cytometry panel (Panel 7). ....	194
Figure 63 : Gating strategy for the <i>DC</i> flow cytometry panel (Panel 8). ....	195
Figure 64 : Gating strategy for the <i>Th</i> flow cytometry panel (Panel 9). ....	196
Figure 65 : Gating strategy for the <i>ILC</i> flow cytometry panel (Panel 10) ....	197

## List of tables

Table 1 : Characteristic and function of major circulating innate cells .....	20
Table 2 : Characteristic and function of major circulating adaptive cells.....	27
Table 3 : Expression and function of major inflammatory cytokines .....	34
Table 4: Complex and purified PRR Stimuli used in the Whole-Blood stimulation systems .....	46
Table 5 : Comparison of different gene expression profiling technologies for whole blood analysis...	61
Table 6 : Cytokine gene signature that defines transcriptional response to IFNB, IFNG, IL1B and TNFA .....	68
Table 7 : Demographic variables included in the <i>MI</i> study .....	94
Table 8 : Staining antibodies used for the ten 8-color flow cytometry panels.....	95
Table 9 : List of the immunophenotypes measured in this study.....	100
Table 10 : Genome-wide signals of association with immunophenotypes in the <i>MI</i> cohort. ....	118
Table 11 : Genome-wide signals of association with immunophenotypes in the <i>MI</i> cohort, with their consequences on gene expression, protein sequence and human diseases and traits.....	123
Table 12 : Other genome-wide significant ( $P < 10^{-10}$ ) or suggestive ( $P < 5 \times 10^{-8}$ ) signals of association with immunophenotypes in the <i>MI</i> cohort .....	125
Table 13 : .....	131
Table 14 : Associations of HLA classical alleles with immunophenotypes in the <i>MI</i> cohort .....	132
Table 15 : Demographic parameters of the replicative collection .....	138
Table 16: Examples of allele frequency distribution within the replicative collection.....	139
Table 17 : Innate and adaptive immune stimuli used for development of whole-blood stimulation system.....	166
Table 18 : Gene expression probes .....	185

## APPENDICES

---

### **1. PHD PUBLICATION (STUDY 1)**

***URRUTIA ET AL ., CELL REPORTS 2016 :***  
*DEFINING THE BOUNDARIES OF A HEALTHY  
IMMUNE RESPONSE USING STANDARDIZED  
IMMUNE MONITORING TOOLS*

### **2. ADDITIONAL INFORMATIONS (STUDY 1)**

*TECAN Script For Automated RNA Purification*

### **3. ADDITIONAL PUBLICATIONS**

***(MILIEU INTERIEUR) :***

***THOMAS ET AL ., CI 2015***

***DUFFY ET AL ., IMMUNITY 2014***

***HASAN ET AL ., CI 2015***

***CHEN ET AL ., CI 2015***

## **PhD PUBLICATION (Study 1)**

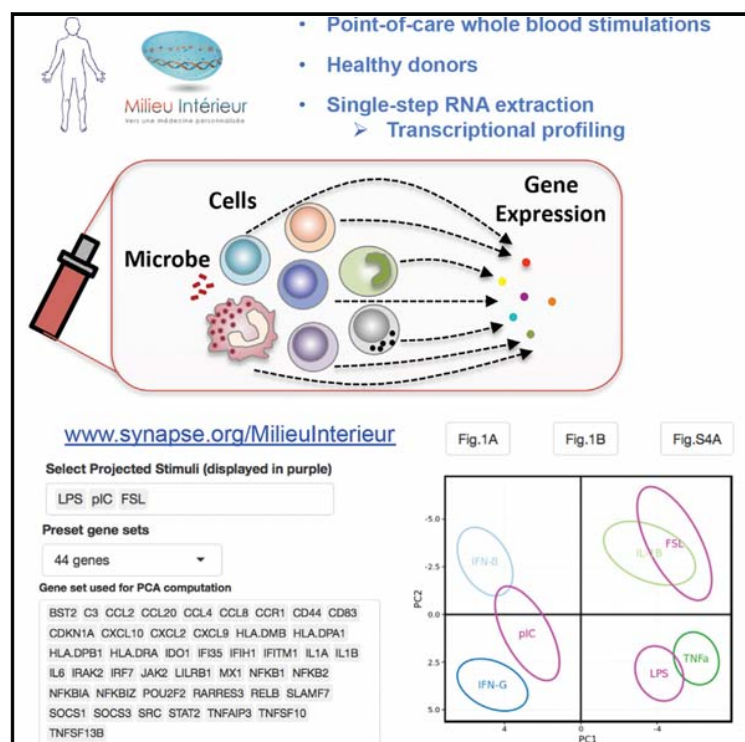
**Urrutia et al ., Cell Reports 2016 :**

*DEFINING THE BOUNDARIES OF A HEALTHY  
IMMUNE RESPONSE USING STANDARDIZED  
IMMUNE MONITORING TOOLS*

# Cell Reports

## Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses

### Graphical Abstract



### Authors

Alejandra Urrutia, Darragh Duffy, Vincent Rouilly, ..., Lluís Quintana-Murci, Matthew L. Albert, Milieu Intérieur Consortium

### Correspondence

quintana@pasteur.fr (L.Q.-M.), albertm@pasteur.fr (M.L.A.)

### In Brief

Urrutia et al. test the hypothesis that responses to TLR ligands or microbes can be captured by induced cytokine signatures. They identified 44 genes that improved segregation of complex stimuli. They also provide reference values that reflect natural variation of immune responses among humans and launched a companion online R-Shiny application for interactive data mining.

### Highlights

- Standardized whole-blood stimulation, single-step RNA extraction, and gene expression
- Response to TLR ligands and microbes can be captured by induced cytokine signatures
- A 44-gene signature allows deconvolution of complex responses
- Online R-Shiny application permits interactive data-mining

### Accession Numbers

GSE85176



Urrutia et al., 2016, Cell Reports 16, 2777–2791  
September 6, 2016 © 2016 The Authors.  
<http://dx.doi.org/10.1016/j.celrep.2016.08.011>

CellPress



# Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses

Alejandra Urrutia,<sup>1,2,10,11</sup> Darragh Duffy,<sup>1,2,3,11</sup> Vincent Rouilly,<sup>3</sup> Céline Posseme,<sup>3</sup> Raouf Djebali,<sup>3</sup> Gabriel Illanes,<sup>3,4,5</sup> Valentina Libri,<sup>3</sup> Benoit Albaut,<sup>6</sup> David Gentien,<sup>6</sup> Barbara Piasecka,<sup>3</sup> Milena Hasan,<sup>3</sup> Magnus Fontes,<sup>4,7</sup> Luis Quintana-Murci,<sup>8,9,\*</sup> Matthew L. Albert,<sup>1,2,3,10,12,\*</sup> and Milieu Intérieur Consortium

<sup>1</sup>Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris 75015, France

<sup>2</sup>INSERM U1223, Paris 75015, France

<sup>3</sup>Center for Translational Research, Institut Pasteur, Paris 75015, France

<sup>4</sup>IGDA, Institut Pasteur, Paris 75015, France

<sup>5</sup>Centro de Matemática, Facultad de Ciencias, Universidad de la República, 11200 Montevideo, Uruguay

<sup>6</sup>Institut Curie, Centre de Recherche, Département de recherche translationnelle, Plateforme de Génomique, Paris 75005, France

<sup>7</sup>Centre for Mathematical Sciences, Lund University, 221 00 Lund, Sweden

<sup>8</sup>Laboratory of Human Evolutionary Genetics, Department of Genomes and Genetics, Institut Pasteur, Paris 75015, France

<sup>9</sup>CNRS URA3012, Paris 75015, France

<sup>10</sup>Department of Cancer Immunology, Genentech Inc., San Francisco, CA 94080, USA

<sup>11</sup>Co-first author

<sup>12</sup>Lead Contact

\*Correspondence: [quintana@pasteur.fr](mailto:quintana@pasteur.fr) (L.Q.-M.), [albertm@pasteur.fr](mailto:albertm@pasteur.fr) (M.L.A.)

<http://dx.doi.org/10.1016/j.celrep.2016.08.011>

## SUMMARY

Systems approaches for the study of immune signaling pathways have been traditionally based on purified cells or cultured lines. However, *in vivo* responses involve the coordinated action of multiple cell types, which interact to establish an inflammatory microenvironment. We employed standardized whole-blood stimulation systems to test the hypothesis that responses to Toll-like receptor ligands or whole microbes can be defined by the transcriptional signatures of key cytokines. We found 44 genes, identified using Support Vector Machine learning, that captured the diversity of complex innate immune responses with improved segregation between distinct stimuli. Furthermore, we used donor variability to identify shared inter-cellular pathways and trace cytokine loops involved in gene expression. This provides strategies for dimension reduction of large datasets and deconvolution of innate immune responses applicable for characterizing immunomodulatory molecules. Moreover, we provide an interactive R-Shiny application with healthy donor reference values for induced inflammatory genes.

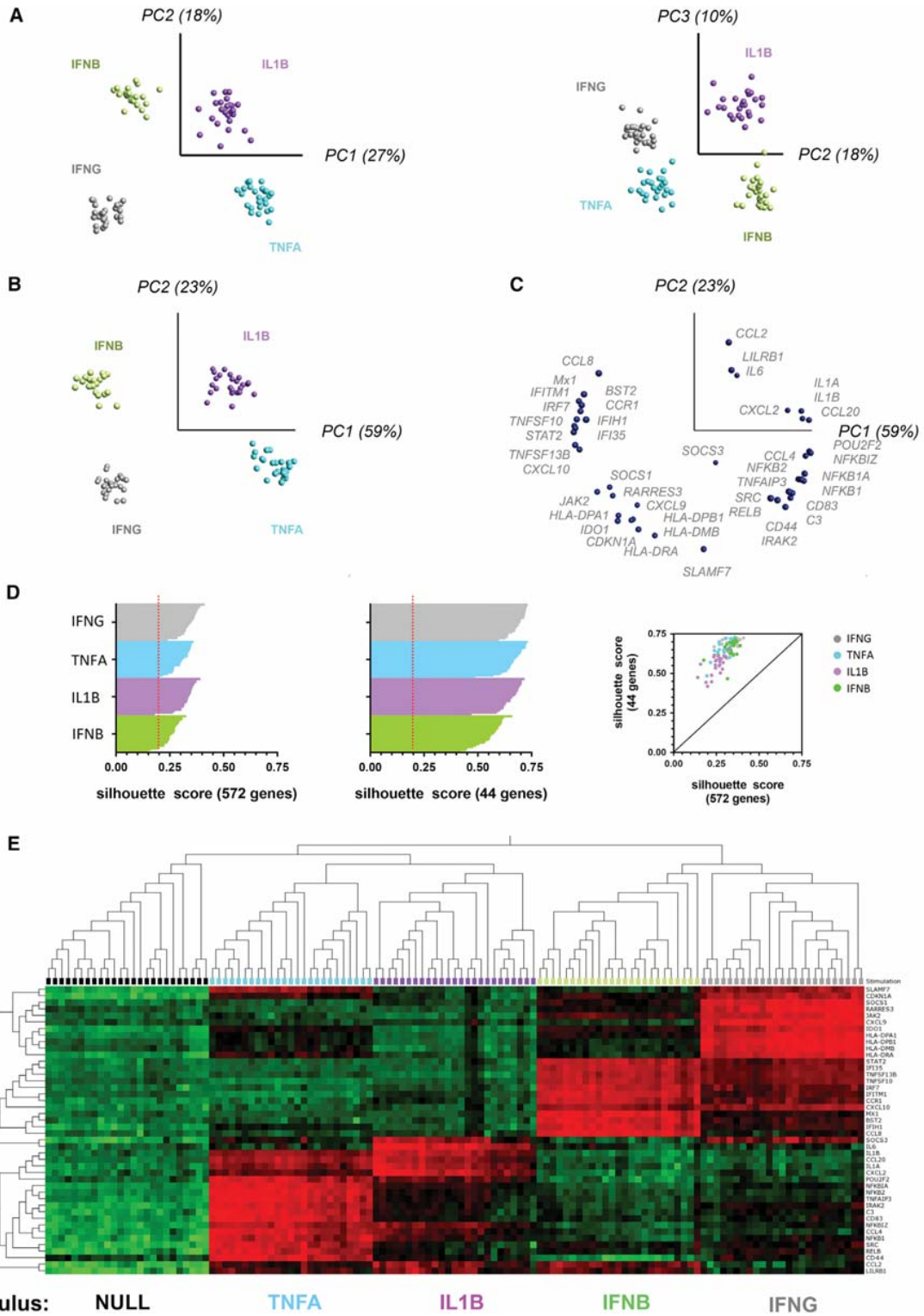
## INTRODUCTION

The initiation of inflammatory responses is typically triggered by a local event engaging sentinel cells, leading to the subsequent recruitment and accumulation of leukocytes. This process can

result in the elimination of the initial cause of tissue disruption, the clearance of dying cells, and establishes a path toward tissue resolution. Cytokines mediate cell-to-cell communication, acting to recruit immune cells to inflammatory microenvironment and drive the required effector mechanisms. Despite the inherent complexity of these processes *in natura*, analyses of inflammation have typically focused on the decision-making circuits within cells, and, in most cases, have been restricted to single cell types (Amit et al., 2009; Jovanovic et al., 2015; Lee et al., 2014). Several other studies have assessed *in vivo* responses to vaccination, typically performing sampling over time to assess induced protein, mRNA expression, and seroconversion (Banchereau et al., 2014; Li et al., 2014; Tsang et al., 2014). While informative, these latter approaches permit the testing of only one stimulation condition per individual and are restricted to qualified or experimental vaccines. To properly account for inter-individual variability in the deconvolution of complex immune responses, both simple (synthetic or purified ligand) and complex (live or heat-killed microbe), stimulations must be performed in the same donor and at the same time, and standardized approaches for all steps from sample collection to analysis must be applied.

To test the hypothesis that responses to Toll-like receptor ligands or whole microbes can be captured by the transcriptional signature of key effector cytokines, we employed a standardized whole-blood stimulation approach with an automated single-step RNA extraction and hybridization gene array readout. Stimulations were performed at the point-of-care, using syringe-based medical devices (TruCulture tubes), in a pilot study that consisted of 25 well-characterized healthy individuals of European ancestry (Thomas et al., 2015). Previously, we reported the testing of protein signatures present in the culture supernatant (Duffy et al., 2014). Herein, we used the cell pellets extracted





(legend on next page)

from the TruCulture stimulation systems to define the transcriptional response to clinically relevant cytokines; interferon-alpha 2A (IFN- $\alpha$ ), interferon-beta 1 (IFN- $\beta$ 1), interferon-gamma (IFN- $\gamma$ ), tumor necrosis factor alpha (TNF- $\alpha$ ), and interleukin 1-beta (IL-1 $\beta$ ). By defining unique and distinct gene expression signatures of cytokine-induced transcription, it was possible to test the clustering and classification of responses to Toll-like receptor (TLR) agonists or whole microbes (including heat killed [HK] gram-negative bacteria, HK gram-positive bacteria, HK fungi, live mycobacteria and viruses). Our results demonstrate the ability to define complex stimuli in terms of the underlying cytokine loops. Moreover, we provide reference values that reflect the degree of naturally occurring variation of immune responses among healthy individuals originating from a homogeneous European background. These data have been made available as a reference for the community, accessible through an online R-Shiny application that permits data-mining using the analytical methods presented.

## RESULTS

### Distinct Transcriptional Signatures Induced by the IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$ , and TNF- $\alpha$ Cytokines

To perform *ex vivo* stimulation while preserving physiological cellular interactions, we utilized syringe-based medical devices for activating immune cells present in whole blood. Based on initial dose-finding studies, quality assurance, solubility, and stability testing<sup>8</sup>, we prioritized stimuli for development in TruCulture whole-blood collection and culture devices (Myriad RBM). After 22 hr stimulation, insertion of a valve separator yielded a cell pellet that was stabilized in Trizol LS and stored at  $-80^{\circ}\text{C}$  for subsequent mRNA expression analysis utilizing the NanoString nCounter technology (Figure S1A). Due to the Trizol content in our samples and to minimize pre-analytical biases, we established an automated mRNA single-step chloroform-free extraction protocol (Tecan script provided on-line, see <http://www.pasteur.fr/labex/milieu-interieur>). Direct comparison with conventional RNA extraction protocols indicated excellent correlation in gene expression counts between the two extraction methods (Spearman's rank-order correlation,  $r_s > 0.99$ , Figure S1B). Expression data were normalized with nSolver Analysis Software (NanoString), using four housekeeping genes: *RPL19*, *TBP*, *POLR2A*, and *HPRT* (Figures S1C–S1F). These four housekeeping genes were selected following the application of the geNorm method (Vandesompele et al., 2002), an established

algorithm for identifying stable housekeeping genes. The selection of these genes is supported by their strong correlations pre- and post-stimulation ( $r_s > 0.9$ ) across the 25 donors, in contrast with those housekeeping genes that were discarded ( $r_s < 0.7$ ) (Figure S1D and data not shown). The overall rationale for the selection of the NanoString platform, as compared to other transcriptional profiling strategies, is presented in Table S1. This choice was validated by the high reproducibility of the data obtained when experiments were performed at different times or at separate institutional core facilities ( $r_s > 0.98$ , Figure S1B).

To assess the signatures induced by cytokine stimulation, we analyzed the expression data of a total of 572 genes in the 25 donors, using unsupervised principal component analysis (PCA) (Figure 1A). The PCA revealed strong clustering of stimuli-specific responses, with the first three principal components (PCs) explaining 55% of the total variance; PC1 separated IL-1 $\beta$  and TNF- $\alpha$  from IFN- $\beta$  and IFN- $\gamma$ , and PC2 distinguished TNF- $\alpha$  from IL-1 $\beta$  and IFN- $\beta$  from IFN- $\gamma$ . Of note, the response to IFN- $\alpha$  was also tested and found to be similar to that of the IFN- $\beta$  response (t test with  $q < 0.05$  reported no variables as significantly different between the two stimuli) (Figure S3), and therefore, IFN- $\alpha$  was excluded from further analyses.

To reduce the dimensionality of the data and exclude genes that did not contribute to unique cytokine-induced signatures, we next defined the differential gene expression for each stimulus with respect to the null control using linear support vector machine (SVM) approaches (Burges, 1998). This enabled us the selection of predictive cytokine gene signatures from gene lists ranked according to a paired t test (individual stimulus versus null condition). Bootstrapping of data in the SVM training phase ensured robust results (details provided in the Experimental Procedures). The union of the selected cytokine gene signatures yielded a set of 44 genes that separated the four cytokine stimuli (Table 1). The resulting PCA projection revealed that the four stimulation conditions could be separated into four clearly distinct clusters based on the expression levels of these 44 genes, with PC1 and PC2 capturing 82% of the total variance (Figure 1B). The 44 genes are represented on a biplot—a synchronized dual projection of the variables that drive the loading of the PC vectors (Figure 1C). To quantify the improved clustering provided by this approach, we calculated silhouette scores, i.e., a measure of the distance between the respective k-means clusters, reported for each sample based on the likelihood to localize into one cluster as compared to any of the three other defined clusters. Comparison between the scores that

### Figure 1. Distinct Gene Expression Signature Induced by Cytokine Stimulation

Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with IFN- $\beta$  (pale green), IFN- $\gamma$  (gray), IL-1 $\beta$  (purple), and TNF- $\alpha$  (turquoise). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes employing Qlucore Omics Explorer v3.1. Prior to applying PCA, values for each of the 572 mRNA were log transformed, centered to a mean value of zero across each donor, and scaled to unit variance. The four cytokine stimuli are indicated by the colored circles and the vector position of each of the 25 donors is represented.

(A) Left: PC1 versus PC2. Right: PC2 versus PC3. The percentage of variance captured by each PC is indicated.

(B) PCA on filtered gene expression data; first for differential gene expression (paired t test comparing each cytokine with null and a q value cut-off of  $10^{-3}$ ); followed by the classification of samples using linear support vector machine (SVM) approaches, and genes ranked according to a paired t test, yielding a union gene set of 44 genes.

(C) A bi-plot of the 44 gene set variable PCA is depicted.

(D) Silhouette scores for each cytokine IFN- $\beta$  (green), IFN- $\gamma$  (gray), IL-1 $\beta$  (purple), and TNF- $\alpha$  (turquoise) based on the complete 572-gene set and the selected 44-gene set.

(E) Hierarchical clustering of the donors based on the filtered gene list and four cytokine stimuli and Null condition showing the unique and overlapping expression.

**Table 1. Cytokine Gene Signature that Defines Transcriptional Response to IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$ , and TNF- $\alpha$**

Gene Name	Associated Cytokine	q Value (Stim versus Null)	q Value (ANOVA on Four Cytokine Stimuli)
BST2	IFN- $\beta$	$4.3 \times 10^{-43}$	$4.6 \times 10^{-43}$
C3	TNF- $\alpha$	$1.6 \times 10^{-64}$	$2.3 \times 10^{-64}$
CCL2	IL-1 $\beta$	$3.1 \times 10^{-21}$	$2.7 \times 10^{-21}$
CCL20	IL-1 $\beta$	$6.3 \times 10^{-62}$	$1.6 \times 10^{-61}$
CCL4	TNF- $\alpha$	$4.1 \times 10^{-57}$	$7.9 \times 10^{-57}$
CCL8	IFN- $\beta$	$8.8 \times 10^{-53}$	$9.6 \times 10^{-53}$
CCR1	IFN- $\beta$	$8.8 \times 10^{-33}$	$9.0 \times 10^{-33}$
CD44	TNF- $\alpha$	$3.2 \times 10^{-58}$	$5.8 \times 10^{-58}$
CD83	TNF- $\alpha$	$1.4 \times 10^{-59}$	$2.4 \times 10^{-59}$
CDKN1A	IFN- $\gamma$	$1.2 \times 10^{-41}$	$1.3 \times 10^{-41}$
CXCL10	IFN- $\beta$	$5.3 \times 10^{-51}$	$5.7 \times 10^{-51}$
CXCL2	IL-1 $\beta$	$8.4 \times 10^{-39}$	$7.5 \times 10^{-39}$
CXCL9	IFN- $\gamma$	$4.0 \times 10^{-43}$	$4.0 \times 10^{-43}$
HLA-DMB	IFN- $\gamma$	$3.8 \times 10^{-61}$	$2.5 \times 10^{-61}$
HLA-DPA1	IFN- $\gamma$	$4.2 \times 10^{-51}$	$3.5 \times 10^{-51}$
HLA-DPB1	IFN- $\gamma$	$3.5 \times 10^{-51}$	$2.7 \times 10^{-51}$
HLA-DRA	IFN- $\gamma$	$4.0 \times 10^{-45}$	$3.9 \times 10^{-45}$
IDO1	IFN- $\gamma$	$2.2 \times 10^{-61}$	$1.2 \times 10^{-61}$
IFI35	IFN- $\beta$	$3.4 \times 10^{-55}$	$2.7 \times 10^{-55}$
IFIH1	IFN- $\beta$	$2.3 \times 10^{-54}$	$2.2 \times 10^{-54}$
IFITM1	IFN- $\beta$	$5.7 \times 10^{-49}$	$6.1 \times 10^{-49}$
IL1A	IL-1 $\beta$	$1.1 \times 10^{-59}$	$2.0 \times 10^{-59}$
IL1B	IL-1 $\beta$	$7.8 \times 10^{-83}$	$2.9 \times 10^{-82}$
IL6	IL-1 $\beta$	$1.9 \times 10^{-67}$	$6.9 \times 10^{-67}$
IRAK2	TNF- $\alpha$	$4.8 \times 10^{-62}$	$9.8 \times 10^{-62}$
IRF7	IFN- $\beta$	$3.4 \times 10^{-56}$	$2.3 \times 10^{-56}$
JAK2	IFN- $\gamma$	$5.8 \times 10^{-51}$	$5.0 \times 10^{-51}$
LILRB1	IL-1 $\beta$	$1.6 \times 10^{-37}$	$1.5 \times 10^{-37}$
MX1	IFN- $\beta$	$1.4 \times 10^{-61}$	$6.3 \times 10^{-62}$
NFKB1	IL-1 $\beta$	$8.7 \times 10^{-52}$	$1.0 \times 10^{-51}$
NFKB2	TNF- $\alpha$	$2.0 \times 10^{-64}$	$3.6 \times 10^{-64}$
NFKBIA	TNF- $\alpha$	$2.6 \times 10^{-67}$	$3.2 \times 10^{-67}$
NFKBIZ	IL-1 $\beta$	$2.1 \times 10^{-61}$	$3.5 \times 10^{-61}$
POU2F2	IL-1 $\beta$	$1.8 \times 10^{-70}$	$6.6 \times 10^{-70}$
RARRES3	IFN- $\gamma$	$2.2 \times 10^{-49}$	$2.1 \times 10^{-49}$
RELB	TNF- $\alpha$	$1.8 \times 10^{-40}$	$1.9 \times 10^{-40}$
SLAMF7	IFN- $\gamma$	$9.0 \times 10^{-62}$	$2.5 \times 10^{-62}$
SOCS1	IFN- $\gamma$	$1.6 \times 10^{-42}$	$1.6 \times 10^{-42}$
SOCS3	TNF- $\alpha$	$9.0 \times 10^{-62}$	$3.6 \times 10^{-55}$
SRC	TNF- $\alpha$	$2.3 \times 10^{-57}$	$4.3 \times 10^{-57}$
STAT2	IFN- $\beta$	$4.3 \times 10^{-55}$	$3.8 \times 10^{-55}$
TNFAIP3	TNF- $\alpha$	$2.7 \times 10^{-59}$	$4.9 \times 10^{-59}$
TNFSF10	IFN- $\beta$	$2.3 \times 10^{-58}$	$9.9 \times 10^{-59}$
TNFSF13B	IFN- $\beta$	$1.7 \times 10^{-57}$	$1.0 \times 10^{-57}$

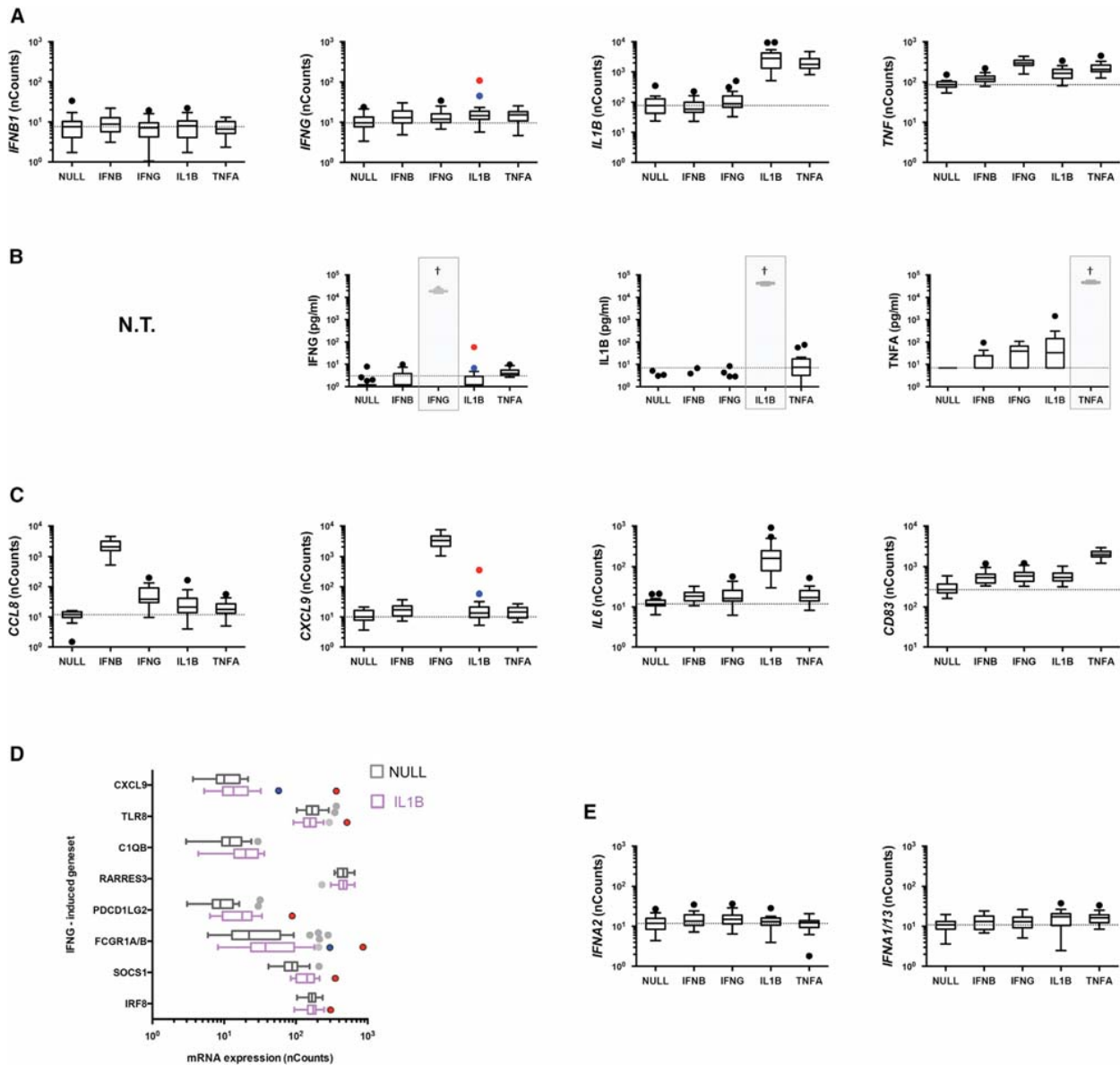
The union set of 44 genes as selected for each cytokine stimulus using linear support vector machine (SVM) approaches and paired t tests with respect to the null control. The q values for each cytokine as compared to the Null (paired t tests) and within the four cytokines (multi-group ANOVA) are shown.

were based on the complete 572 gene set versus the selected 44 gene set revealed a higher score with reduced dimensionality of the feature list and a focus on those most highly discriminating genes (Figure 1D). While our analyses revealed specific cytokine gene signatures, there was modest overlap in the induced gene lists when the stimulation conditions were compared to the null (Figure 1E). Hierarchical clustering of the filtered gene list displayed the unique and overlapping gene expression for the four cytokine groups (Figure 1E).

To examine the intersection among cytokine-induced genes, we first analyzed the induction of IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$  and TNF- $\alpha$  gene expression. While none of the four cytokines triggered high levels of type I or type II IFN expression (Figure 2A), IL-1 $\beta$  and TNF- $\alpha$  both induced high expression of IL-1 $\beta$  mRNA, and all four cytokine stimuli induced modest expression of TNF- $\alpha$  (Figure 2A). These data suggest potential cross-talk among the pathways and highlight a strong feed-forward inter-cellular spread of IL-1 $\beta$  signaling. While this has been previously shown (Dinarello et al., 1987), there is no mechanistic understanding of how IL-1 $\beta$  activates the inflammasome and triggers caspase-1 activation. Unexpectedly, this analysis revealed two outlier individuals who showed high expression levels of IL-1 $\beta$ -induced IFN- $\gamma$  (marked by red and blue dots, Figure 2A). To establish if the observed high levels of IFN- $\gamma$  expression resulted in higher protein secretion, we re-analyzed our previously published protein dataset (Duffy et al., 2014) generated using samples from the same donors and indeed, the two individuals showed the highest levels of IFN- $\gamma$  protein in the culture supernatants (Figure 2B). The presence of recombinant protein that was used as the stimulus restricted the interpretation of potential positive feedback loops for the given protein (these data points are masked by a gray box, Figure 2B). In addition to the induction of IFN- $\gamma$  by the two outlier individuals, we also observed higher expression of several IFN- $\gamma$ -induced genes, as compared to the other donors studied (Figures 2C–2E). Together, these data support the concept that the induced innate responses include the spreading of signals through cytokine feedback loops and potential cross-talk among the inter-cellular pathways.

### Variable Responses to TLR and Microbe Stimulation Are Captured by Induced Cytokine Response

During vaccination or acute infection, the immune system is exposed to agonists that stimulate Toll-like receptor (TLRs) signaling. In such conditions, small numbers of cells are engaged, triggering in turn the production of cytokines that spread the inflammatory response. To test this concept, we evaluated whether the induced transcriptional responses to the four effector cytokines are capable of capturing the diversity of seven well-defined TLR agonists (Duffy et al., 2014): FSL-1 (FSL, also known as Pam2C) that engages the TLR2-TLR6 heterodimer; poly IC (pIC) that engages TLR3; lipopolysaccharide (LPS) that engages TLR4; flagellin (FLA) that engages TLR5; gardiquimod (GARD) that engages TLR7; R848 that engages both TLR7 and TLR8; and CpG-2216 oligonucleotide (ODN) that engages TLR9. Limiting doses of the respective agonists were selected to more closely reflect in vivo responses and to ensure that we were working within the linear range of physiological responses

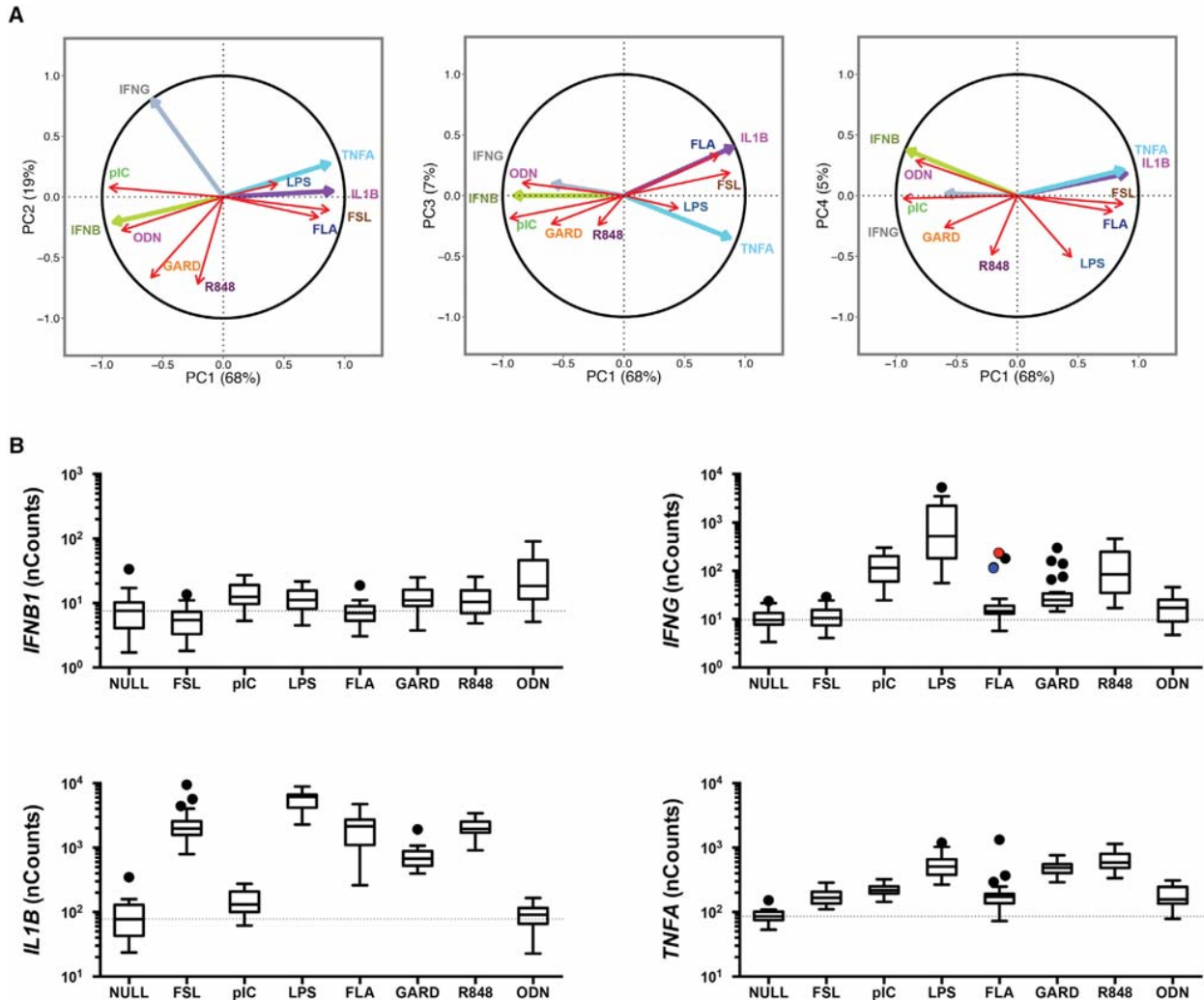


**Figure 2. Interactions and Outlier Responses among the Cytokine-Induced Gene Expression Signatures**

(A–D) Whole blood from 25 healthy donors was stimulated using the Null, IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$ , and TNF- $\alpha$  stimulation conditions. mRNA gene expression (A), absolute nCounts, or induced protein expression (B) are plotted for each of the four genes or gene products: *IFN- $\beta$ 1*, *IFN- $\gamma$* , *IL-1 $\beta$* , and *TNF- $\alpha$*  (N.T. signifies not tested for IFN- $\beta$  protein; gray shaded boxes mask those protein assays that are detecting the input stimulus in the TruCulture tube). mRNA expression for the most differentially expressed gene is shown (C), one per cytokine stimulus as reported in Table 1 gene list. Top IFN- $\gamma$ -induced gene expression is shown for the Null (gray) and IL-1 $\beta$  (purple) stimuli (D). Data are represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation. Two individual outliers (identified by their induction of *IFN- $\gamma$*  expression in response to IL-1 $\beta$  stimulation) are indicated using blue and red circles, respectively. (E) Cytokine stimulation does not induce expression of IFN- $\alpha$  genes. Box-whisker Tukey plots of IFN- $\alpha$ 2 and IFN- $\alpha$ 1/13 mRNA expression following stimulation with NULL, IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$ , and TNF- $\alpha$ . Dotted line indicates median null value.

(please refer to Duffy et al., 2014 or <http://www.milieuinterieur.fr/en> for details on the dose and source of these reagents). To assess potential similarity in gene expression, we projected the data from each of the seven TLR stimuli onto a fixed PCA coordinate, which was defined by the eigenvectors and eigenvalues

of the optimized PCA of the four cytokine-induced mRNA expression data (44 genes defined in Figure 1C). Strikingly, two of the TLR stimuli clustered with a defined cytokine—FLA and FSL vectors both projected onto the IL-1 $\beta$  cluster (Figures S4A and S4B). ODN eigenvectors projected into the IFN- $\beta$  quadrant,



**Figure 3. TLR-Induced Gene Expression Can Be Represented as a Function of Cytokine-Induced Gene Signatures**

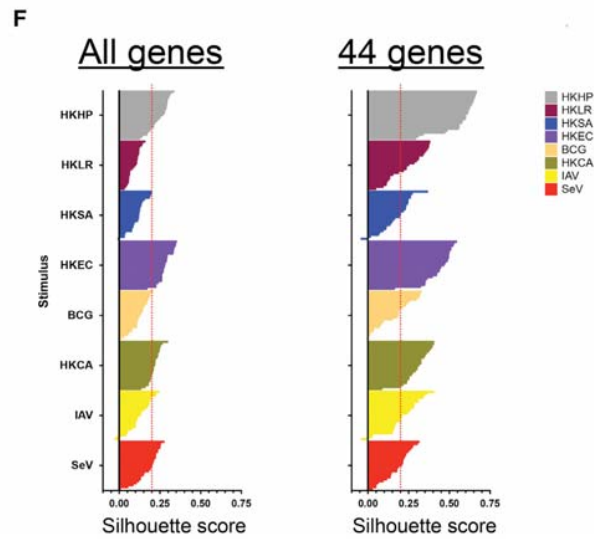
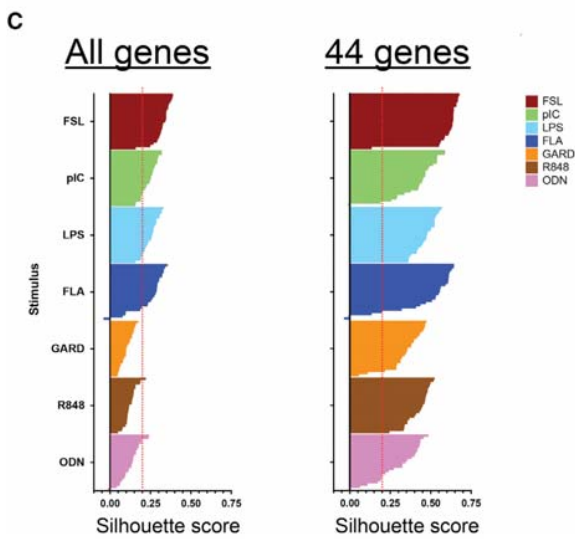
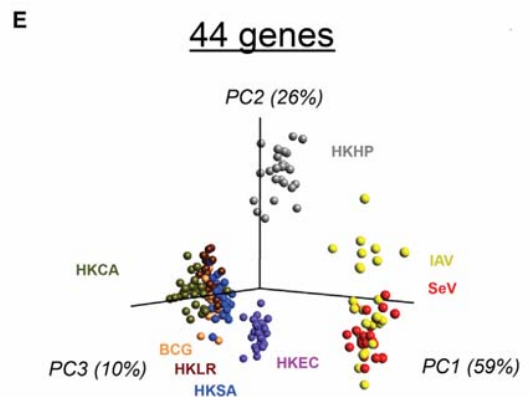
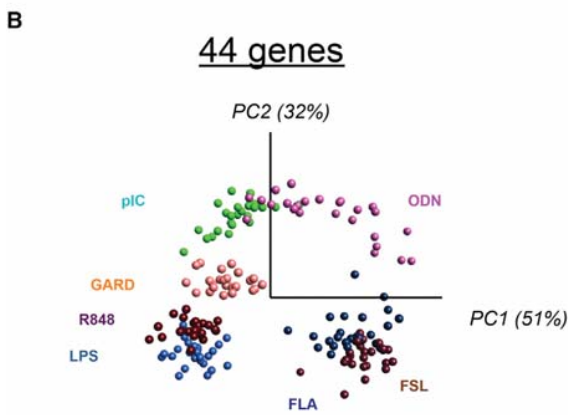
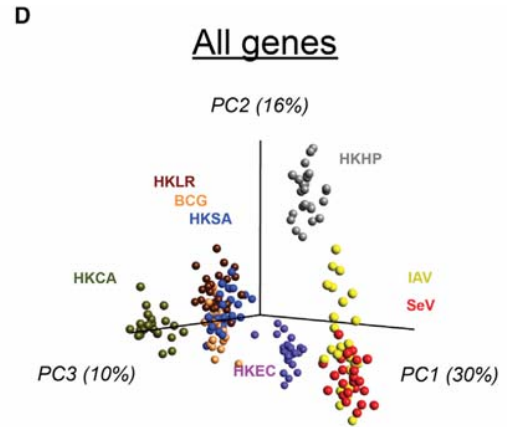
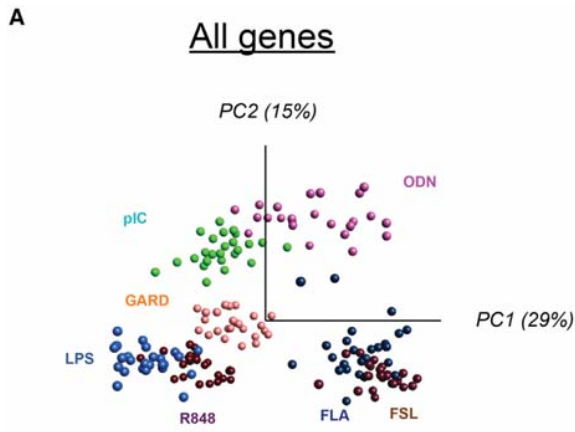
(A) Correlation circles of unit length were constructed using the 44 gene set and PCA loadings were obtained using the gene expression dataset from the four cytokine stimuli (as defined in Table 1). The vectors for TLR-induced gene expression signatures were generated from the median value for the 25 donors, projected onto the correlation circles across the four PC.

(B) *IFN- $\beta$ 1*, *IFN- $\gamma$* , *TNF- $\alpha$* , and *IL-1 $\beta$*  gene expression nCounts are shown for the Null and TLR stimulation conditions. Data are represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation. Two individual outliers (identified by their induction of *IFN- $\gamma$*  expression in response to IL-1 $\beta$  stimulation, Figure 2A) are indicated using blue and red circles, respectively.

with an inter-donor variance in the intensity of gene expression (Figure S4A), which was consistent with our previous study of induced proteins. This analytical approach can be further explored using the online user interface (<http://www.synapse.org/MilieuInterieur>, <http://dx.doi.org/10.7303/syn7059574>).

We next represented the data on a correlation circle, as an alternative for visualizing the relationships among stimuli (Figure 3A), allowing us the projection of all TLR stimulation conditions across the four PC axes. When two stimulation vectors are close to the unit circle and are proximal to each other, then they are positively correlated (e.g., FLA and FSL). By contrast, if they are orthogonal to each other, they are not correlated

(e.g., FLA and R848). Alternatively, when a stimulation vector is close to the center (e.g., LPS in PC1 versus PC2), it means that information is carried in the other axes (e.g., in the case of LPS almost all variance is carried by PC3 and PC4). Collectively, these data suggest that FLA- and FSL-induced transcriptional signatures are highly correlated to the IL-1 $\beta$  stimulation response; piC, GARD, R848, and ODN are correlated with type I or type II IFN stimulation; and LPS is intermediate between the two. These results were consistent with the TLR induced expression of *IFN- $\beta$ 1*, *IFN- $\gamma$* , *IL-1 $\beta$* , and *TNF- $\alpha$*  (Figure 3B). One unanticipated result was the similarity between FLA and FSL and the IL-1 $\beta$  gene expression signature. In the case of FLA,



(legend on next page)

we suggest this may be occurring due to the engagement of the intracellular sensor NLRC4, in turn activating caspase-1 (Gay et al., 2014); however, the mechanisms underlying FSL activation of the inflammasome also remains uncharacterized. Notably, these analyses also identified the two outlier individuals discussed above, who showed high expression levels of FLA-induced *IFN- $\gamma$*  (blue and red dots, Figure 3B).

We applied the same approach to characterize several less well-studied agonists. These included whole  $\beta$ -glucan particles (WGP) derived from *Saccharomyces cerevisiae*, known to engage Dectin-1 and lacking TLR-stimulating activity (Li et al., 2007); lipoarabamanin (LAM), a component of mycobacterial cell walls and an inducer of TLR2; and calcium pyrophosphate dihydrate crystals (CPPD), the etiological agent of pseudogout (Martinon et al., 2006), and a stimulator of NLRP3. Consistent with inflammasome activation, CPPD mapped to the IL-1 $\beta$  cluster, and similar to FSL1, we demonstrate that the LAM-induced gene expression overlaid the IL-1 $\beta$  gene set (Figure S4B). By contrast, WGP induced an mRNA expression signature that projected between IL-1 $\beta$  and TNF- $\alpha$ . Extension of this method may support the classification of unknown adjuvants or innate stimuli.

Next, we performed unsupervised PCA on the TLR-stimulated gene expression data using the entire 572-gene set (Figure 4A). The first two PCs, capturing 44% of the total variance, segregated all TLR stimuli with the exception of FLA and FSL (shown to have similar gene expression patterns), and to a lesser extent LPS and R848. The clustering achieved with the entire dataset was then compared to a PCA plot built using the 44-gene signature, selected for the four effector cytokines (Table 1). Strikingly, the vectors built from the cytokine-gene set fully captured the diversity of responses among the TLR stimuli (Figure 4B). Moreover, the cytokine-optimized gene set provided improved definition of the clusters, as indicated by a higher silhouette scores (Figure 4C). This is most evident for the improved discrimination of LPS from R848 (Figure 4B, see PC2; and an increase in the median silhouette score from 0.26 to 0.46 for LPS and from 0.11 to 0.35 for R848 samples, Figure 4C). These observations support the hypothesis that, in situations of limited agonist concentration and heterogeneous cell types, the characteristic TLR gene signatures can be identified by a limited set of cytokine-induced genes. From the perspective of population-based studies, this introduces the concept that a handful of highly discriminatory gene expression responses are sufficient

to distinguish the transcriptional landscape activated by TLR pathways.

To test the robustness of this prediction, we subsequently evaluated the gene expression patterns induced by whole microbes, first using the entire 572-gene set (Figure 4D). The microbes included heat-killed *Escherichia coli* O111:B4 (HKEC), *Staphylococcus aureus* (HKSA), *Lactobacillus rhamnosus* (HKLR), *Helicobacter pylori* (HKHP), *Candida albicans* (HKCA), a clinical preparation of live bacillus Calmette-Guerin (BCG), H1N1 attenuated influenza A/PR8 (IAV), and Sendai virus (SeV). The first three principal components, capturing 56% of the total variance, segregated samples from the viral stimuli and HKEC from the other microbes in PC1; HKHP was separated by PC2; and the remaining microbes falling along PC3 with HKCA being distinguishable from HKLR, HKSA, and BCG. Again, we demonstrated improved clustering when using the 44-gene set, as defined by the response to the four effector cytokines (Figure 1B,C). Strikingly, when using the 44-gene set, the variance captured by the first three principle components reached 95% (Figure 4E). Indeed, even with whole microbe stimulation—representing a higher level of biological complexity due to the activation of multiple signaling pathways—we obtained improved silhouette scores for k-means clustering across all stimuli when the PCA was based on the 44-gene set (Figure 4F). For example, the clustering of HKHP samples improved from a median silhouette score of 0.27 to 0.52, when applying the selected 44-gene set in place of the complete 572 genes. Notably, HKLR, HKSA, and BCG were less distinguishable, likely a result of common agonist activity and similar levels of induced cytokines. IAV and SeV also co-segregated for similar reasons. Nonetheless, a doubling of the median silhouette score indicated that, here too, a focused feature list improved clustering of the data. In light of these results, we conclude that a standardized sample collection combined with precise measurement of induced gene expression supports a massive reduction in the dimensionality of the data space, while preserving the ability to discriminate the inflammatory trigger as well as the variability among human donors.

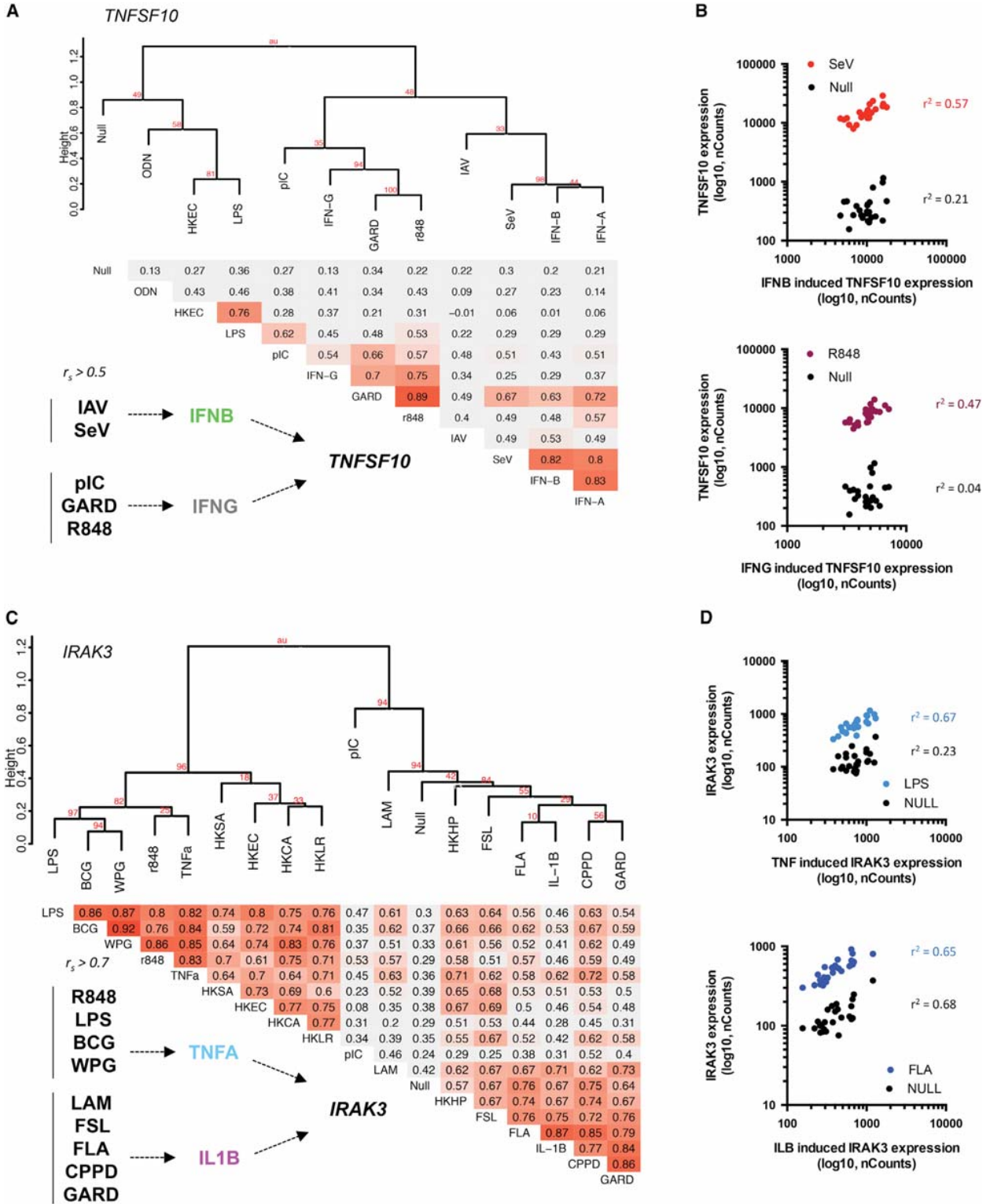
### Inter-individual Variable Gene Expression Supports Tracing of Cytokine Loops

We next extended the concept of correlation among the stimulation conditions to shed light onto possible cytokine loops

#### Figure 4. Distinct and Variable Response to TLR Agonist and Microbial Stimulation Can Be Captured Using the Cytokine-Induced 44-Gene Signature

(A and B) Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with FSL (maroon), pIC (green), LPS (light blue), FLA (dark blue), GARD (orange), R848 (brown), and ODN (pink). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes (A), PC1 versus PC2 (the percentage of variance captured by each PC is indicated). A parallel PCA was constructed using the mRNA expression data from the filtered set of 44 cytokine-induced genes (from gene lists reported in Table 1) (B), PC1 versus PC2 (the percentage of variance captured by each PC is indicated). (C) Silhouette scores were determined for each sample based on k means clustering ( $k = 7$ ). Samples are plotted according to TLR stimulus. The red-line indicates a silhouette score of 0.2 (considered a strong fit). The median silhouette score for 572-gene set was 0.19; and for the 44-gene set it was 0.45. (D and E) Whole-blood stimulation was also performed using HKHP (gray), HKLR (brown), HKSA (blue), HKEC (purple), HKCA (gray-green), BCG (orange), IAV (yellow), and SeV (red). PCA was used to project mRNA expression data from 572 genes (D); and the parallel PCA was constructed using the mRNA expression data from 44 genes. (F) Silhouette scores were determined for each sample based on k means clustering ( $k = 8$ ). Samples are plotted according to microbial stimulus. Note that IAV and SeV were mixed among two clusters (not depicted); and two samples were misclustered using the 572-gene set versus five samples misclustered using the 44-gene set (not depicted). The red-line indicates a silhouette score of 0.2 (considered a strong fit). The median silhouette score for 572-gene set was 0.18; and for the 44-gene set it was 0.26.





(legend on next page)

involved in individual gene expression. This approach provides an exploratory analysis of possible cell-to-cell interactions that can be tested in future experimental studies. Spearman correlation matrices and hierarchical clustering, based on a connected correlation dissimilarity metric, were performed for each gene, and results were bootstrapped to ensure the identified correlations were robust. Using these outputs, we identified cases where the variable responses to TLR or microbe stimulations could be explained by the inter-individual gene expression variance observed when using one of the four cytokine stimuli. To illustrate this observation, the dendrogram depicting the clusters of Spearman correlations and a table indicating the respective  $r_s$  coefficients are shown for *TNFSF10* (Figure 5A). A cut-off value of 2-fold expression change greater than the null condition was utilized for inclusion of stimuli in the cluster. Interestingly, the viral stimuli clearly clustered with type I IFN stimulation, with SeV showing a high correlation with IFN- $\beta$ -induced *TNFSF10* ( $r_s = 0.82$ ); whereas GARD and R848 clustered with IFN- $\gamma$  ( $r_s = 0.7$  and  $0.75$ , respectively) (Figures 5A and 5B). As a second example, *IRAK3* is shown, illustrating distinct clustering of bacterial/TLR stimuli with TNF or IL-1 $\beta$  (Figures 5C and 5D). Schematic depictions of the putative stimulus-induced cytokine-mediated expression of *TNFSF10* or *IRAK3* are shown with dotted line arrows provided for illustrative purposes. This analytical approach allows us to predict the distinct cytokine loops that drive common gene expression following stimulation by TLR agonists or microbes. While this modeling approach to population-based data must be experimentally validated, we highlight the possibility that inter-individual variance can be utilized as a means to identify causal pathways driving gene expression, which will support future experimental inquiry.

### Microbial Gene Expression Is Defined by Lymphocyte-Derived Cytokines

Although the four cytokines studied herein represent major effector pathways in host response and disease pathogenesis, we were cognizant of additional upstream factors that help to specify the inflammatory reaction. To identify other potential effector cytokines, we generated a list of genes upregulated by each stimulus as compared to the null condition (stimulus > null, paired t test  $q < 10^{-3}$ ) and then merged the resulting gene lists for the four cytokines, the seven TLR, and the eight microbial stimuli. A Venn diagram depicts the overlap and intersections in gene expression for these three groups, respectively (Figure 6A). Additionally, we calculated the median gene expression for each stimulus and generated heat maps, clustering by both genes and samples, using either the set of genes that were expressed after microbial but not cytokine stimulation (Figure 6B); TLR but not cytokine stimulation (Figure S6A); and microbial but not TLR stimulation (Figure S6B). Strikingly, the

complex stimuli induced a subset of genes indicative of lymphocyte activation. This subset of genes included: (1) transcription factors such as *FoxP3* (highly induced after bacterial stimulation), *EOMES* (induced by HKCA) and *GATA3* (induced by BCG); (2) cytolytic effectors such as *GZMA* (highly induced by HKEC); and (3) anti-microbial genes such as *NOS2* (induced after bacterial stimulation), *DEFB103A* (induced by BCG) and *HAMP* (highly induced by HKEC) (Figure 6B). Additionally, we detected the differential induction of 18 cytokines, which included *IL2* (induced by HKSA, BCG, HKCA, IAV, and SeV), *CSF2* (highly induced by HKCA), and *IL22* (induced after bacterial and HKCA stimulation) (Figure 6C). As indicated by the comparison with Staphylococcal enterotoxin B (SEB) stimulation and consistent with the presence of microbial antigen-specific T cells within the repertoire of healthy donors (Becattini et al., 2015; Geiger et al., 2009), these cytokine genes likely reflect the activation of lymphocyte subsets (Figure 6C). The characterization of these lymphocyte-derived cytokines may further establish the role of feed-forward cytokine loops in the deconvolution of microbial-induced gene signatures.

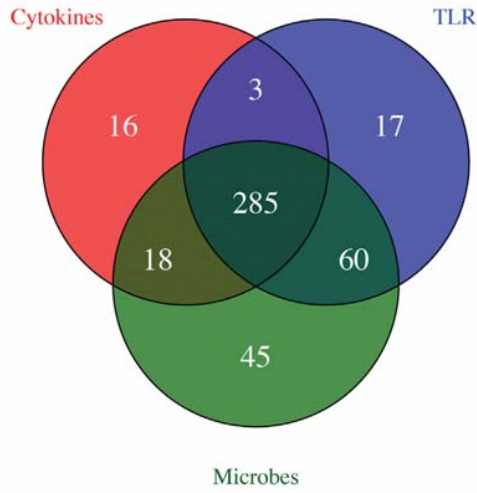
### DISCUSSION

In this study, we aimed at testing if standardized whole-blood stimulation systems can support the identification of a handful of genes that are capable of deconvoluting complex responses to immune stimulation. We utilized medically relevant stimuli to determine their inflammatory signatures and, in doing so, established the degree of naturally occurring variation present in a population of well-defined healthy donors of European descent. The definition of host immune responses to adjuvants and microbial agents, and subsequent characterization of inter-individual variability in the human population, is of major fundamental interest and provides the necessary foundation for understanding human health and disease pathogenesis. Although functional tests are routinely used in laboratory investigation (Folds and Schmitz, 2003), the standardization of such assays has been challenging. While whole-blood assays are more biologically relevant and introduce less experimental bias than, for example, PBMC stimulation, they are not without technical challenges in particular due to the high levels of globin RNA and enzyme-inhibiting compounds (e.g., heparin interference of reverse transcriptase) (Chaussabel et al., 2010). Previous efforts have focused on removing the globin RNA before downstream analysis, however, these processes can introduce, in turn, higher levels of technical variance as compared to what was achieved with our data generation pipeline (Shin et al., 2014). Specifically, the innovation brought forward in this study is an automated single-step RNA extraction method from whole blood, which minimized pre-analytical bias and generated highly reproducible results when

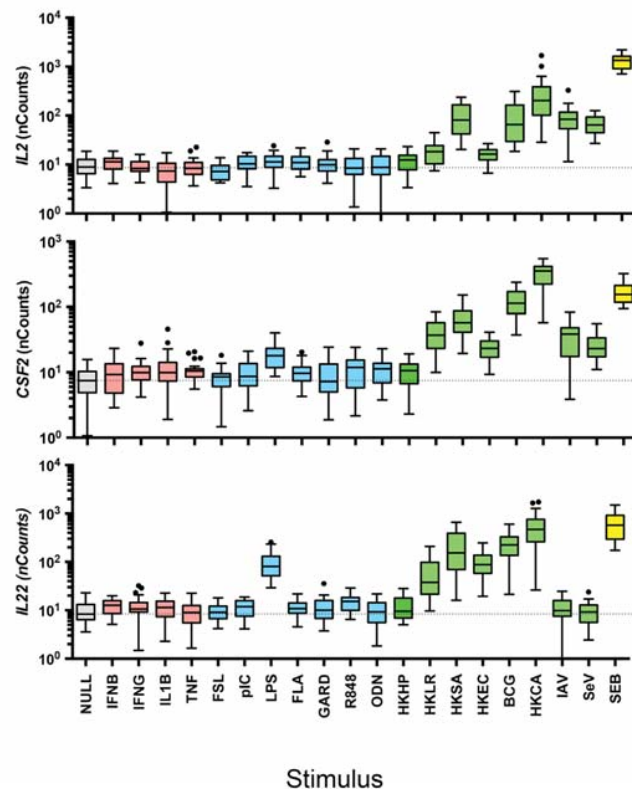
#### Figure 5. Correlation among Variable Stimulus-Induced Gene Expression Helps to Trace Cytokine Loops

Gene expression data from all 23 stimulation conditions were used to generate Spearman correlation matrices and hierarchical cluster analysis followed by bootstrapping. The dendrograms shown depict clustering of stimuli based on Spearman correlations for *TNFSF10* (A) or *IRAK3* (C) and the associated triangular matrix indicates the respective pairwise  $r_s$  coefficients. Scatter plots for indicated stimulation pairs are shown. Each dot represents the absolute nCount for a single individual of the 25 healthy donors tested for *TNFSF10* (B) or *IRAK3* (D). Red numbers at the intersection of the dendrogram branches indicate approximately unbiased (au) p values, reported as percentage for 1,000 sampled dendrograms. Color scale on tables indicates strength of correlation. Proposed schematics for stimulus-driven cytokine-induced gene expression is proposed using indicated cut-off for  $r_s$ .

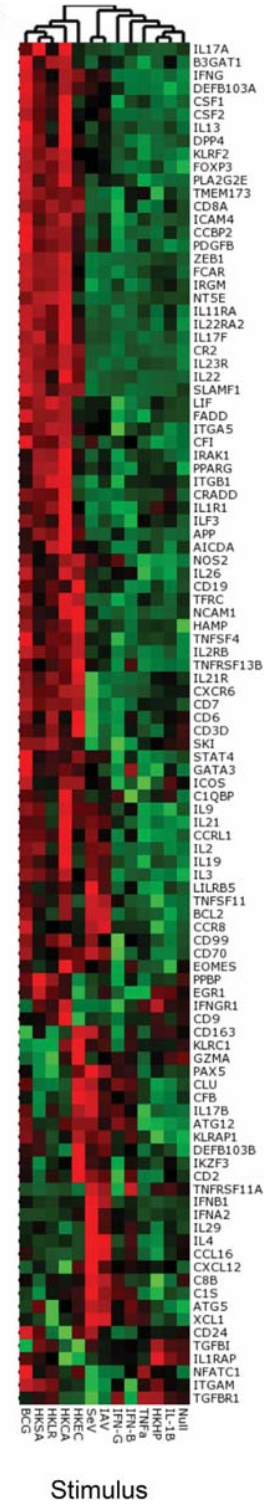
A



C



B



(legend on next page)

using a gene hybridization read-out. These solutions are essential for multicenter population-based studies, as well as for assays with ambitions for clinical deployment.

Using the reference data presented herein, we tested the hypothesis that responses to TLR ligands or whole microbes can be captured by the transcriptional signature of key effector cytokines. We tested a total of 23 stimulation systems, all built into whole-blood syringes for point of care sampling. Using linear SVM learning, it was possible to identify a 44-gene set, selected based on their ability to differentially cluster cytokine-induced genes. Strikingly, these same genes, when applied to the stratification of responses to TLR ligands or microbes, resulted in improved discrimination among the stimuli as indicated by a marked improvement in silhouette scores. In the era of an increased use of whole-genome transcriptional profiles, our results suggest that limiting the pre-analytical bias introduced by cell separation and non-standardized stimulation protocols may be more important than obtaining greater numbers of measured genes. In addition to sample collection and data analysis standardization, we minimized intrinsic variability by the recruitment of donors of Western European ancestry (third generation born in Metropolitan France). Furthermore, we minimized pre-analytic or environmental sources of variability by applying highly precise inclusion and exclusion criteria (Thomas et al., 2015). To restrict other sources of variability, in addition to the standardization of the assay systems, all donors were sampled at the same time of day (09:00–11:00), during the same week, and in the same location. Such a reliable monitoring of induced immune gene expression responses permitted the classification of inflammatory and host immune responses based on the variance observed in healthy donors.

In addition to defining detailed healthy reference ranges to be considered in future clinical studies, this work permitted the identification of a number of outlier responses. This included two individuals that responded to FLA or IL-1 $\beta$  by producing IFN- $\gamma$  and in turn the induction of IFN- $\gamma$ -stimulated genes. Following from this observation, we extended the approach of tracing cytokine loops and gene expression pathways, using inter-individual variance and correlation among the stimulation signatures as a means to deconvolute complex transcriptional responses. This approach may also support the future classification of unknown adjuvants, innate stimuli, new pathogenic agents or the stratification of disease and treatment response. If extended to the study of disease states, it may be possible to classify, for example, subsets of rheumatoid arthritis patients that are responsive to IL-1 $\beta$  versus TNF- $\alpha$  blockade (Gibbons and Hyrich, 2009; McInnes and Schett, 2007).

This reference dataset and the applied analytical approach offers a useful resource to the community, nevertheless, several specific limitations should be highlighted. First, some of the employed TLR stimuli may engage secondary pathways in addition to their commonly ascribed receptors. Notably, the observation that FLA is highly correlated with the IL-1 $\beta$ -induced gene signature suggests that it may also trigger NLRC4 within the whole-blood stimulation systems. This may occur within neutrophils, which express high levels of the NLRC4 inflammasome and release IL-1 $\beta$  (Chen et al., 2014). If correct, it would also help to explain why, despite the high prevalence of dominant-negative forms of TLR5 in Europeans (Barreiro et al., 2009; Hawn et al., 2003), all 25 donors showed an induced response after FLA stimulation (Barreiro et al., 2009). Alternatively, TLR sensor pathways on platelets and neutrophils may be unique in their ability to engage caspase-1 (Hayashi et al., 2003). We also observed that IAV and SeV were highly correlated with pIC, suggesting that the latter is engaging RIG-I like receptors (RLRs) in addition to TLR3. We also acknowledge that, in the natural setting, human immune responses typically occur in mucosal tissues and, as such, stromal cells and tissue resident immune populations such as macrophages and ILCs may need to be considered to fully apply our dataset to physiologic and pathologic responses. Lastly, our analyses consider a single analytical time point only, thus capturing a snapshot of the complexity inherent in dynamic immune responses.

Finally, it is our aim with this resource paper to highlight the growing need to make data more accessible and easier to explore. In line with recent efforts (Gorenshteyn et al., 2015; Speake et al., 2015), we have thus developed an online R-Shiny application software that will allow readers to fully query the dataset based on their specific questions. This application software was built as a direct companion to the presented analyses with publically available R-scripts and downloading options for gene expression data. In sum, the data resource presented here and the available online tools provide a foundation for association studies, kinetic analyses, and in vivo mechanistic experimentation. For example, it remains to be established how the inter-individual variation in gene expression that we identified here is accounted for by host genetic variants (i.e., expression quantitative trait loci [eQTLs]), specifically in cases where gene expression variation is altered upon activation with certain immune stimuli (i.e., response/interaction eQTLs). Conceptually, the strategy to trace inter-cellular cytokine driven gene expression may support such future eQTL association studies, especially in cases where inter-cellular *trans*-eQTL are identified. From a practical viewpoint, the tools will support a path toward more targeted immune monitoring from whole blood, enabling

#### Figure 6. Microbial-Induced Lymphokines Are Absent from TLR and Cytokine Gene Expression Signatures

(A) Gene expression data from all 23 stimulation conditions were used to generate stimulus-induced signatures (stimulus > null, paired t test with q value cut-off of  $10^{-3}$ ). The union sets of cytokine (IFN- $\beta$ , IFN- $\gamma$ , IL-1 $\beta$ , TNF- $\alpha$ ); TLR (FSL, pIC, LPS, FLA, GARD, R848, ODN); and microbes (HKHP, HKLR, HKSA, HKEC, HKCA, BCG, IAV, SeV) were generated. The Venn diagram indicates the number of shared and unique genes among the three groups of stimuli.

(B) Hierarchical clustering of the donors and genes based on the 105 genes present in the union set of microbes but not cytokines was performed. The median gene expression value was used for each stimulus, with variables log-transformed, mean-centered per donor, and scaled to unit variance. NB, the dendrogram for clustering of genes not shown.

(C) Representative gene expression data are shown for *IL2*, *CSF2*, and *IL22* for each stimuli, as well as Staphylococcal enterotoxin B (SEB) stimulation for reference. Data are represented as box-whisker Tukey plots. Dotted lines indicate the median value for the Null stimulation.

the use of standardized approaches that capture the common variation within the human population.

## EXPERIMENTAL PROCEDURES

### Donors

Samples were obtained as part of the Milieu Intérieur Healthy Donor Cohort (<https://www.clinicaltrials.gov/>; NCT01699893). The study protocol was designed and conducted in accordance with the ethical principles of the Declaration of Helsinki and Good Clinical Practices as outlined in the ICH Guideline for Good Clinical Practices. The data were collected under pseudo-anonymized conditions: the identity of the subject is coded in a way that does not allow third-party persons to detect the identity of the person. All subjects (12 male, 13 female) aged 30–39 years old, gave informed consent and were considered as healthy based on medical history, clinical examination, laboratory results, and electrocardiography (ECG). More specific details on criteria to define healthy can be found in previously published work (Thomas et al., 2015).

### TruCulture Stimulation

TruCulture tubes were prepared in batch with the indicated stimulus, resuspended in a volume of 2 ml buffered media and maintained at  $-20^{\circ}\text{C}$  until use. Blood was obtained from the antecubital vein using a 60 ml syringe containing sodium-heparin (50 IU/ml final concentration). Within 15 min of collection, 1 ml of whole blood was distributed into pre-warmed TruCulture tubes, inserted into a dry block incubator, and maintained at  $37^{\circ}\text{C}$  ( $\pm 1^{\circ}\text{C}$ ), room air for 22 hr ( $\pm 15$  min). After incubation, a valve was inserted to separate cells from the supernatant and to stop the stimulation reaction. Upon removal of the liquid supernatant, cell pellets were resuspended in 2 ml Trizol LS (Sigma), vortexed for 2 min, and rested for 10 min at room temperature (RT) before  $-80^{\circ}\text{C}$  storage.

### High-Throughput Standardized RNA Extraction

Samples were randomized and extracted in groups of 95. Cell pellets in Trizol LS were thawed on ice 60 min prior to processing. To complete thawing and RNA release, tubes were vortexed twice for 5 min at 2,000 rpm. Before processing, a centrifugation ( $3,000 \times g$  for 5 min at  $4^{\circ}\text{C}$ ) of the thawed samples was performed to pellet the cellular debris generated during the Trizol lysis. The barcoded tubes were loaded in the rack module of the Freedom EVO platform (TECAN) and scanned for sample traceability. For extraction, a modified protocol of the NucleoSpin 96 RNA tissue kit (Macherey-Nagel) was developed and adapted to the Freedom EVO integrated vacuum system. The detailed script for the operation of the TECAN system is provided online (<http://www.milieuinterieur.fr/en>). In brief, 600  $\mu\text{l}$  of clarified phase of the Trizol lysate was transferred to a deep well plate preloaded with 900  $\mu\text{l}$  of 100% ethanol. The binding mixture was transferred into the silica membrane plate. The columns were washed with buffers MW1 and MW2 ( $\times 2$ ) and RNA eluted into 0.5 ml 2D barcoded tubes (ThermoScientific) using 60  $\mu\text{l}$  RNase-free water. As an internal control of the extraction process, a tube containing a defined quantity of spiked RNA was included in each run. To avoid unnecessary freeze and thaw of the RNA, distinct aliquots for quality control and gene expression analysis were prepared, and all aliquots were frozen at  $-80^{\circ}\text{C}$  until use.

### RNA Quality Controls

RNA concentration was estimated using Qubit RNA HS Assay Kit (Life Technologies) according to the protocol provided by the manufacturer. An automated RNA integrity assessment was performed using the Standard RNA Reagent Kit on a LabChipGX (Perkin Elmer). The RNA quality score (RQS) was calculated using the LabChip System software, and all samples with a RQS greater than four were processed for gene expression analysis.

### Selection Criteria for Gene Expression Analysis

NanoString nCounter, a hybridization-based multiplex assay, was selected after comparison with multiple gene expression technologies (microarray, qPCR-based methods) (Table S1). All assays were performed at the genomic platform (Institut Curie), with the exception of the cross platform control comparison performed at Institut Pasteur, Paris. The Human Immunology v2 gene

code set was selected as it covers 25 immunology-related gene networks as illustrated by the use of KEGG charts (Figure S2). The code set contains a total of 594 probes (15 correspond to housekeeping genes), of which 572 probes were included in downstream analysis after removing probes mapping to multiple genes and probes aligning to polymorphic regions with greater than two SNPs (Table S2). To this end, the probes were mapped against the human genomic sequence (GRCh37/hg19) with GSNAP (Wu and Nacu, 2010), a splice-aware aligner. A total of 573 out of 594 probes were mapped with 100% identity to the genome. Twelve probes mapped with one to two mismatches in the middle of the sequence, eight probes were misaligned in the first/last 1–9 bp, and one probe did not map at all (PECAM1 located on HG183\_PATCH). The misaligned probes were realigned manually using BLASTN against Ab-initio cDNAs database. Of the 594 probes, 15 mapped to more than one genomic location (see Table S2). We removed from further analysis KIR\_Activating\_Subgroup\_1 probe, which mapped to three different genomic locations, as well as three other KIR probes that mapped to multiple locations: KIR\_Activating\_Subgroup\_2, KIR\_Inhibiting\_Subgroup\_1, and KIR\_Inhibiting\_Subgroup\_2. Bioconductor biomaRt package (Durinck et al., 2005) version 2.24.0 was used to query Ensembl (release 75) (Flicek et al., 2014) and retrieve exonic variants that mapped to the same regions as the NanoString probes. We considered only SNPs with minor allelic frequency  $>0.05$  (1000 Genomes Project). Forty-eight probes showed the presence of one to two SNPs in their sequence. HLA-DRB1, HLA-DQA1, and HLA-DQB1 probes contained 4, 9, and 13 SNPs, respectively, and were therefore removed from further analysis.

### Gene Expression Analysis

Total mRNA were diluted with RNase-free water at 20 ng/ $\mu\text{l}$  in the 12-strip provided by NanoString. We analyzed 100 ng (5  $\mu\text{l}$ ) of total RNA from each sample using the Human Immunology kit v2 according to manufacturer's instructions. Each sample was analyzed in a separate multiplexed reaction including in each, eight negative probes and six serial concentrations of positive control probes. Negative control analysis was performed to determine the background for each sample. Of note, we observed variable expression of two negative control probes (NEG B, NEG F), which cross-reacted with bacterial nucleic acid present in two of the TruCulture systems (HKSA and BCG, respectively, Figures S1D and S1E), and thus these probes were not used for data normalization. Data was imported into nSolver analysis software (version 2.5) for quality checking and normalization of data. A first step of normalization using the internal positive controls permitted correction of potential sources of variation associated with the technical platform. To do so, we calculated for each sample the geometric mean of the positive probe counts. A scaling factor for a sample was a ratio of the average across all geometric means and the geometric mean of the sample. For each sample, we multiplied all gene counts by the corresponding scaling factor. Next, for each sample we calculated the background level as the median  $+2$  SD across the six negative probe counts. For each gene in a sample, we subtracted the background level. Finally, to normalize for differences in RNA input we used the same method as in the positive control normalization, except that geometric means were calculated over four housekeeping genes (RPL19, TBP, POLR2A, and HPRT1). These genes were selected using geNorm method (Vandesompele et al., 2002), an established approach for identification of stable housekeeping genes, from the 15 candidate genes provided by NanoString.

### Statistical Analysis, Data Visualization, and Software

Principal component analysis (PCA) or singular value decomposition (SVD) was used to decompose the data matrix in a way that is amenable for dimension reduction (Alter et al., 2000). The decomposition was used to orthogonally project both the rows and the columns of the data matrix into lower dimensional space in an optimal way—optimal signifying the retention of as much of the original variance in the dataset as possible. For a comprehensive overview of PCA and the exploratory analysis using dual PCA and the accompanying PCA biplots, we refer to Fontes (2012). Before applying PCA, the variables (mRNA expression levels) were log-transformed, mean-centered per donor, to avoid inter-donor variability obscuring inter-stimuli responses, and finally the variables were scaled to unit variance. The mean-centering per donor is in accordance with the paired structure in the data and paired *t* tests or

ANOVA were performed throughout. Scaling to unit variance prevents large variances in the data from obscuring the correlation structure in the data. Q values, which are defined as false discovery rate (FDR)-adjusted p values (Benjamini and Hochberg, 1995), were used to define statistical significance.

Correlation circles were generated by computing the median value across the 25 donors, for each of the considered 44 genes; we then transposed the data matrix to consider the four stimulation conditions as the four PCA dimensions; finally, the vectors representing the TLR stimuli were projected onto the four-dimensional PCA. The respective 2D PCA projection plots were made with the R package “FactoMineR” (version 1.28) to compute PCA scores and projected coordinates. Silhouette analysis was used to study the separation distance among the TLR and microbial stimuli. K means clustering was performed using the Open CV library (Bradski and Kaehler, 2008); with settings equal to 100 iterations and 500 attempts and the silhouette scores were computed (Bradski and Kaehler, 2008; Steinhaus, 1956). Cluster number was selected based on the number of stimuli represented in the PCA ( $k = 7$  for TLR,  $k = 8$  for microbes). Note, silhouette coefficients near +1 indicate that the sample is far away from the neighboring clusters; a value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters, and negative values indicate that those samples might have been assigned to the wrong cluster. Bootstrapped hierarchical clustering analysis was performed using the “pvclust” R package (version 1.3-2) using a Spearman-based dissimilarity metric. One thousand trees were sampled to evaluate the robustness of each cluster. Correlation matrices were plotted using the R graphics package ggplot2 (version 1.0.0). Plots were exported from the QluCore Omics Explorer 3.1 or created using the ggplot2 package (version 1.0.0) on the R platform (version 3.1.1).

Stimulus signatures, consisting of gene lists specific for each of the four cytokines, were created by training a support vector machine (SVM) for each individual stimulus versus null. This approach was used to define stimulus signatures set by a discrete number of variables. In order to discover reasonably complex gene interaction networks among the four stimuli, SVMs were optimized from 12–57 gene subsets (~2%–10% of the total gene number). For all four cytokine stimuli, the optimal classifier determined by the SVM cross validation scheme corresponded to the smallest gene set size. The identified gene lists had perfect accuracy upon ten repeated complete SVM test runs. This shows that a small number of selected variables can predict the specific stimulus used. The small overlap between the established stimulus signatures indicates that the selection strikes a reasonable balance between capturing the complexity in the data and at the same time identifying those important individual genes. The open source C++ software library OpenCV was used to build and evaluate the SVMs (Burges, 1998; Chang and Lin, 2011). For comparison, a kNN classifier was also tested, using the implementation in the OpenCV library with default parameter settings, which gave exactly the same stimulus signatures.

### R Shiny (Interactive Web Application) Development

To complement this manuscript, we provide an interactive web application that allows exploration of the dataset presented in this study. The application presents four different types of analytical visualizations: PCA, boxplots, hierarchical clustering, and a searchable reference table. For each visualization, we provide default settings that match figures presented in the manuscript. Visualization controls enable the user to navigate the entire dataset following their own scientific interests. The interactive table provides reference values, based on the 25 healthy donors, which can be directly browsed using a selected method (median expression values, coefficient of variations or q values from paired t tests as compared to the Null condition). The application was implemented using the Open Source R platform, Shiny package (version 0.12.2), ggplot2 package (version 1.0.0), dplyr package (version 0.4.3), and tinyr package (version 0.3.1). All visualization and analysis methods are accessible through a web browser, without the need to install any additional software or possess knowledge of a programming language and is available at <https://www.synapse.org/MilieuInterieur> (<http://dx.doi.org/10.7303/syn7059574>).

### ACCESSION NUMBERS

The accession number for the expression profiling by array data reported in this paper is GEO: GSE85176.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and two tables and can be found with this article online at <http://dx.doi.org/10.1016/j.celrep.2016.08.011>.

### CONSORTIUM

The members of the Milieu Intérieur Consortium are Laurent Abel, Andres Alcover, Kalla Astrom, Philippe Bousso, Pierre Bruhns, Ana Cumano, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Jacques Fellay, Antonio Freitas, Odile Gelpi, Ivo Gomperts-Boneca, Serge Hercberg, Olivier Lantz, Claude Leclerc, Hugo Mouquet, Sandra Pellegrini, Stanislas Pol, Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis, Frédéric Tangy, Eric Tartour, Antoine Toubert, Marie-Noëlle Ungeheuer, Lluís Quintana-Murci, and Matthew L. Albert.

### AUTHOR CONTRIBUTIONS

A.U. performed experiments, analyzed data, and wrote the paper. V.R., G.L., and B.P. analyzed data. C.P., R.D., V.L., B.A., D.G., and M.H. performed experiments. D.D., M.F., L.Q.-M., and M.L.A. designed the study, analyzed data, and wrote the paper. M.F., L.Q.-M., and M.L.A. contributed equally.

### ACKNOWLEDGMENTS

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). We thank Stéphanie Thomas for management of the Milieu Intérieur Consortium, and Haiyin Chen, Sebastian Amigorena, and Shannon Turley for critical review of the manuscript.

Received: January 19, 2016

Revised: May 31, 2016

Accepted: August 2, 2016

Published: August 25, 2016

### REFERENCES

- Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97, 10101–10106.
- Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326, 257–263.
- Banchereau, R., Baldwin, N., Cepika, A.-M., Athale, S., Xue, Y., Yu, C.I., Metang, P., Cheruku, A., Berthier, I., Gayet, I., et al. (2014). Transcriptional specialization of human dendritic cell subsets in response to microbial vaccines. *Nat. Commun.* 5, 5283.
- Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5, e1000562.
- Becattini, S., Latorre, D., Mele, F., Foglierini, M., De Gregorio, C., Cassotta, A., Fernandez, B., Kelderman, S., Schumacher, T.N., Corti, D., et al. (2015). T cell immunity. Functional heterogeneity of human memory CD4<sup>+</sup> T cell clones primed by pathogens or vaccines. *Science* 347, 400–406.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300.
- Bradski, G., and Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library* (O’Reilly).

- Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167.
- Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. TIST* 2.
- Chaussabel, D., Pascual, V., and Banchereau, J. (2010). Assessing the human immune system through blood transcriptomics. *BMC Biol.* 8, 84.
- Chen, K.W., Groß, C.J., Sotomayor, F.V., Stacey, K.J., Tschopp, J., Sweet, M.J., and Schroder, K. (2014). The neutrophil NLR4 inflammasome selectively promotes IL-1 $\beta$  maturation without pyroptosis during acute Salmonella challenge. *Cell Rep.* 8, 570–582.
- Dinarello, C.A., Ikejima, T., Warner, S.J., Orencole, S.F., Lonnemann, G., Cannon, J.G., and Libby, P. (1987). Interleukin 1 induces interleukin 1. I. Induction of circulating interleukin 1 in rabbits in vivo and in human mononuclear cells in vitro. *J. Immunol.* 139, 1902–1910.
- Duffy, D., Rouilly, V., Libri, V., Hasan, M., Beitz, B., David, M., Urrutia, A., Bisiaux, A., Labrie, S.T., Dubois, A., et al.; Milieu Intérieur Consortium (2014). Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. *Immunity* 40, 436–450.
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440.
- Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* 42, D749–D755.
- Folds, J.D., and Schmitz, J.L. (2003). 24. Clinical and laboratory assessment of immunity. *J. Allergy Clin. Immunol.* 111 (2, Suppl), S702–S711.
- Fontes, M. (2012). Statistical and knowledge supported visualization of multivariate data. In *Analysis for Science, Engineering and Beyond*, K. Åström, L.-E. Persson, and S.D. Silvestrov, eds. (Springer Berlin Heidelberg), pp. 143–173.
- Gay, N.J., Symmons, M.F., Gangloff, M., and Bryant, C.E. (2014). Assembly and localization of Toll-like receptor signalling complexes. *Nat. Rev. Immunol.* 14, 546–558.
- Geiger, R., Duhen, T., Lanzavecchia, A., and Sallusto, F. (2009). Human naive and memory CD4<sup>+</sup> T cell repertoires specific for naturally processed antigens analyzed using libraries of amplified T cells. *J. Exp. Med.* 206, 1525–1534.
- Gibbons, L.J., and Hyrich, K.L. (2009). Biologic therapy for rheumatoid arthritis: clinical efficacy and predictors of response. *BioDrugs* 23, 111–124.
- Gorenshteyn, D., Zaslavsky, E., Fribourg, M., Park, C.Y., Wong, A.K., Tadych, A., Hartmann, B.M., Albrecht, R.A., García-Sastre, A., Kleinstein, S.H., et al. (2015). Interactive Big Data Resource to Elucidate Human Immune Pathways and Diseases. *Immunity* 43, 605–614.
- Hawn, T.R., Verbon, A., Lettinga, K.D., Zhao, L.P., Li, S.S., Laws, R.J., Skerrett, S.J., Beutler, B., Schroeder, L., Nachman, A., et al. (2003). A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. *J. Exp. Med.* 198, 1563–1572.
- Hayashi, F., Means, T.K., and Luster, A.D. (2003). Toll-like receptors stimulate human neutrophil function. *Blood* 102, 2660–2669.
- Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347, 1259038.
- Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343, 1246980.
- Li, B., Cramer, D., Wagner, S., Hansen, R., King, C., Kakar, S., Ding, C., and Yan, J. (2007). Yeast glucan particles activate murine resident macrophages to secrete proinflammatory cytokines via MyD88- and Syk kinase-dependent pathways. *Clin. Immunol.* 124, 170–181.
- Li, S., Roupheal, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., et al. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* 15, 195–204.
- Martinon, F., Pétrilli, V., Mayor, A., Tardivel, A., and Tschopp, J. (2006). Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* 440, 237–241.
- McInnes, I.B., and Schett, G. (2007). Cytokines in the pathogenesis of rheumatoid arthritis. *Nat. Rev. Immunol.* 7, 429–442.
- Shin, H., Shannon, C.P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., Wilson-McManus, J.E., Ng, R.T., McManus, B.M., and Tebbutt, S.J.; PROOF Centre of Excellence Team (2014). Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS ONE* 9, e91041.
- Speake, C., Presnell, S., Domico, K., Zeitner, B., Bjork, A., Anderson, D., Mason, M.J., Whalen, E., Vargas, O., Popov, D., et al. (2015). An interactive web application for the dissemination of human systems immunology data. *J. Transl. Med.* 13, 196.
- Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci. Fr.* 4, 801–804.
- Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.-G., Fauchoux, N., Sayegrih, S., Vray, M., et al. (2015). The Milieu Intérieur study - an integrative approach for study of human immunological variance. *Clin. Immunol.* 157, 277–293.
- Tsang, J.S., Schwartzberg, P.L., Kotliarov, Y., Biancotto, A., Xie, Z., Germain, R.N., Wang, E., Olnes, M.J., Narayanan, M., Golding, H., et al.; Baylor HIPC Center; CHI Consortium (2014). Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* 157, 499–513.
- Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3, RESEARCH0034.
- Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26, 873–881.

**Cell Reports, Volume 16**

**Supplemental Information**

**Standardized Whole-Blood Transcriptional  
Profiling Enables the Deconvolution  
of Complex Induced Immune Responses**

**Alejandra Urrutia, Darragh Duffy, Vincent Rouilly, Céline Posseme, Raouf Djebali, Gabriel Illanes, Valentina Libri, Benoit Albaud, David Gentien, Barbara Piasecka, Milena Hasan, Magnus Fontes, Lluís Quintana-Murci, Matthew L. Albert, and Milieu Intérieur Consortium**



## Supplementary Figure Legends

**Table S1. Comparison of different gene expression profiling technologies for whole blood analysis.**

**Table S2. Gene expression probes.** List of genes analyzed by Nanostring nCounter technology including chromosome number, probe map position, SNPs in probe, Ensemble gene IDs, and probe sequence

**Figure S1. Quality control measures for gene expression analysis.** (A) Schematic overview of workflow from blood draw to gene expression analysis. (B) Comparison between mRNA counts for single step extraction protocol and standard extraction protocol utilizing chloroform step, for nCounter analysis at 2 separate time point (75 days apart), and at 2 different locations (Institut Curie, Paris and Institut Pasteur, Paris) (Representative examples are shown and  $r_s^2$  is reported, based on a Spearman correlation). (C) Mean of mRNA counts (log scale) for the 4 selected house keeping genes (HPRT1, POLR2A, RPL19, TBP) across the different stimulation conditions for the 25 donors included in the study. (D) Comparison of mRNA counts (linear scale) for two geNorm selected genes (left plots) versus two candidate house keeping genes (right plots) upon TNFA and SeV stimulation (E) Box-whisker Tukey plots for the negative control probe counts. (F) Example of Neg B and Neg\_F probes from TruCulture stimuli LPS, HKSA, and BCG.

**Figure S2. Gene expression pathways used to select NanoString Immunology panel.** KEGG database pathway analysis of (A) NF- $\kappa$ B, (B) TNFA, (C) Cytokine-Cytokine Receptor, and (D) TLR signaling pathways, with genes included in NanoString analysis colored green, and effector cytokines (IFNB, IFNG, IL1B, TNFA) studied herein colored yellow. Genes in white were not represented on the NanoString codeset.

**Figure S3. IFNA and IFNB show overlapping gene expression profiles.** (A) Whole-blood stimulation was performed on 25 healthy donors using TruCulture systems pre-loaded with IFNA (red), IFNB (pale green), IFNG (grey), IL1B (purple), and TNFA (turquoise). Principle component analysis (PCA) was used to project mRNA expression data from 572 genes (the percentage of variance captured by each PC is indicated). (B-C) Hierarchical cluster analysis of the donors and gene expression following stimulation with IFNA, IFNB, and NULL control (black) identified 58 genes commonly down regulated (B) and 212 genes commonly

upregulated (C) (ANOVA test, q value <  $10^{-3}$ ). Each donor is color-coded revealing that in most instances, individual donors clustered for IFNA / IFNB responses.

**Figure S4. Projection of TLR stimuli onto PCA analysis as defined by 4 effector cytokines.**

(A) PCA defined by the eigenvectors and eigenvalues as based on the four-cytokine induced mRNA expression data of the 44 genes defined in Table 1. Ellipses representing 95% confidence interval (CI) were constructed and replaced the individual samples. Projected sample vectors of TLR stimuli (shown in red) for each of the 25 donors (FSL, pIC, LPS, FLA, GARD, R848, ODN), individually projected onto the first 3 PC vectors, using the 44 selected genes (B) Projection of different synthetic ligands (WGP, LAM, CPPD) onto the PCA as defined by four-cytokine induced mRNA expression.

**Figure S5. Projection of microbial stimuli onto fixed PCA analysis defined by 4 effector cytokines.**

PCA defined by the eigenvectors and eigenvalues and optimized based on the four-cytokine induced mRNA expression data (44 genes defined in Table 1). Ellipses representing 95% confidence interval (CI) were constructed and replaced the individual samples. Projected sample vectors (shown in red) for microbial stimuli for each of the 25 donors (HKHP, HKSA, HKLR, HKEC, BCG, HKCA), individually, projected onto the first 3 PC vectors, using the 44 selected genes.

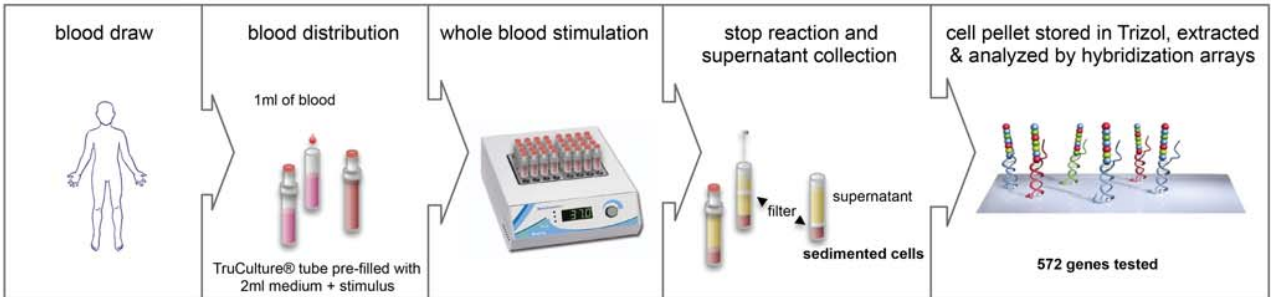
**Figure S6. Gene expression patterns not captured by four effector cytokine induced changes.** Hierarchical cluster analysis of donors and gene expression based on genes expressed after microbial stimulation but not cytokine stimulation (A), and TLR but not cytokine stimulation (B) as defined by firstly by a paired T test for all stimuli versus null (q <  $10^{-3}$ ) and merging the gene lists.

Table S1

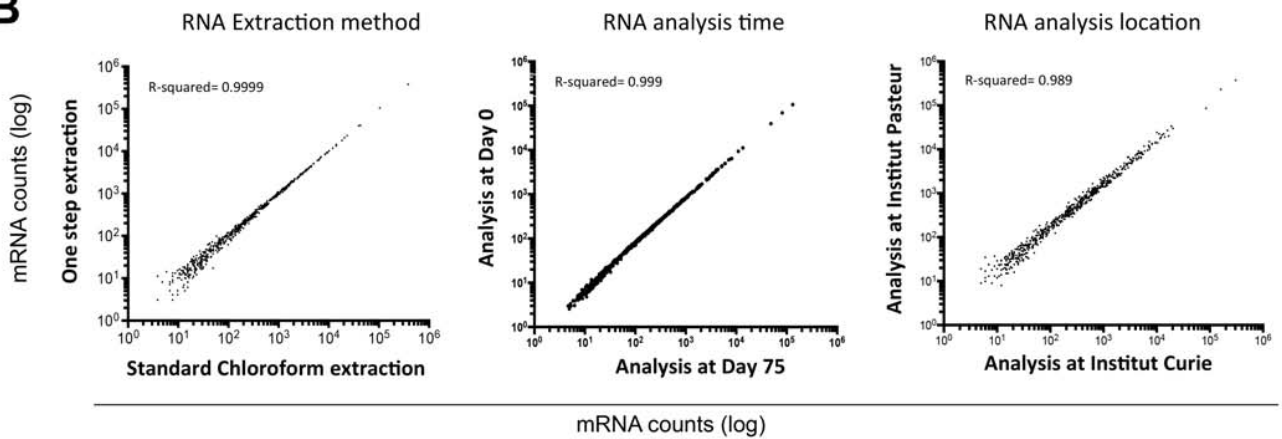
Technology	Method	Gene set	Adaptable for use with whole blood
Fluidigm Biomark	Microfluidic-based PCR	Up to 90 genes	<ul style="list-style-type: none"> <li>• cDNA preparation source of possible pre-analytical bias</li> <li>• PCR reaction inhibited by presence of heparin (and other serum factors) in TruCulture systems</li> </ul>
Nanostring nCounter	RNA Hybridization	Up to 780 genes	<ul style="list-style-type: none"> <li>• Direct quantification of mRNA expression (i.e., no cDNA preparation, no amplification)</li> <li>• suitable for single-step extracted RNA</li> </ul>
Affymetrix microarray	cDNA Hybridization	<b>Whole genome</b>	<ul style="list-style-type: none"> <li>• requirement for high quality RNA</li> <li>• cDNA preparation source of possible pre-analytical bias</li> </ul>
Ion Torrent / Illumina	NGS	<b>Whole genome</b>	<ul style="list-style-type: none"> <li>• requirement for high quality RNA</li> <li>• globin RNA present in whole blood dominate read count (~70% of total RNA)</li> </ul>

Figure S1

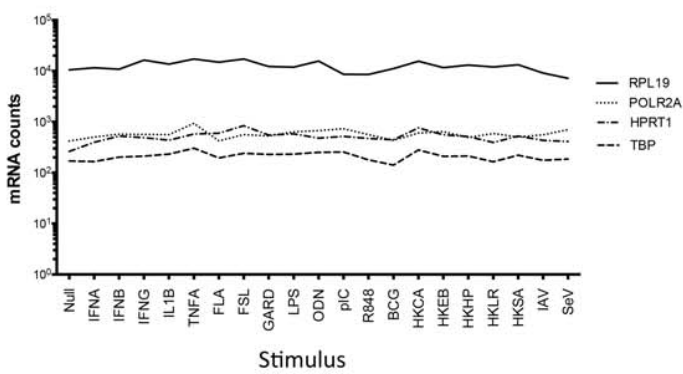
**A**



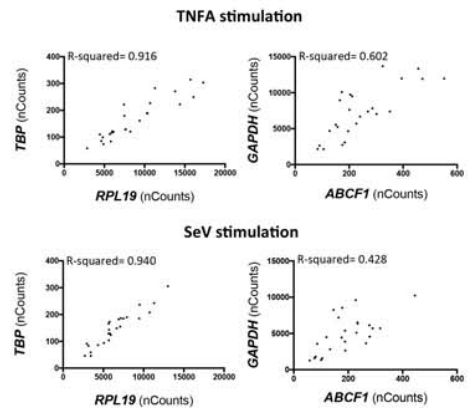
**B**



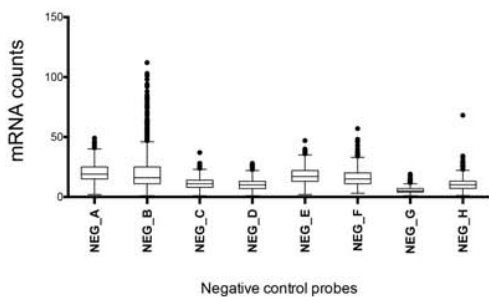
**C**



**D**



**E**



**F**

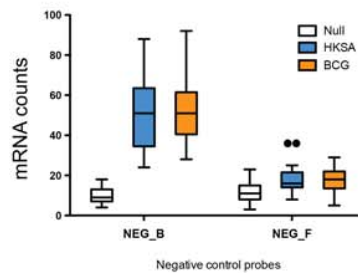
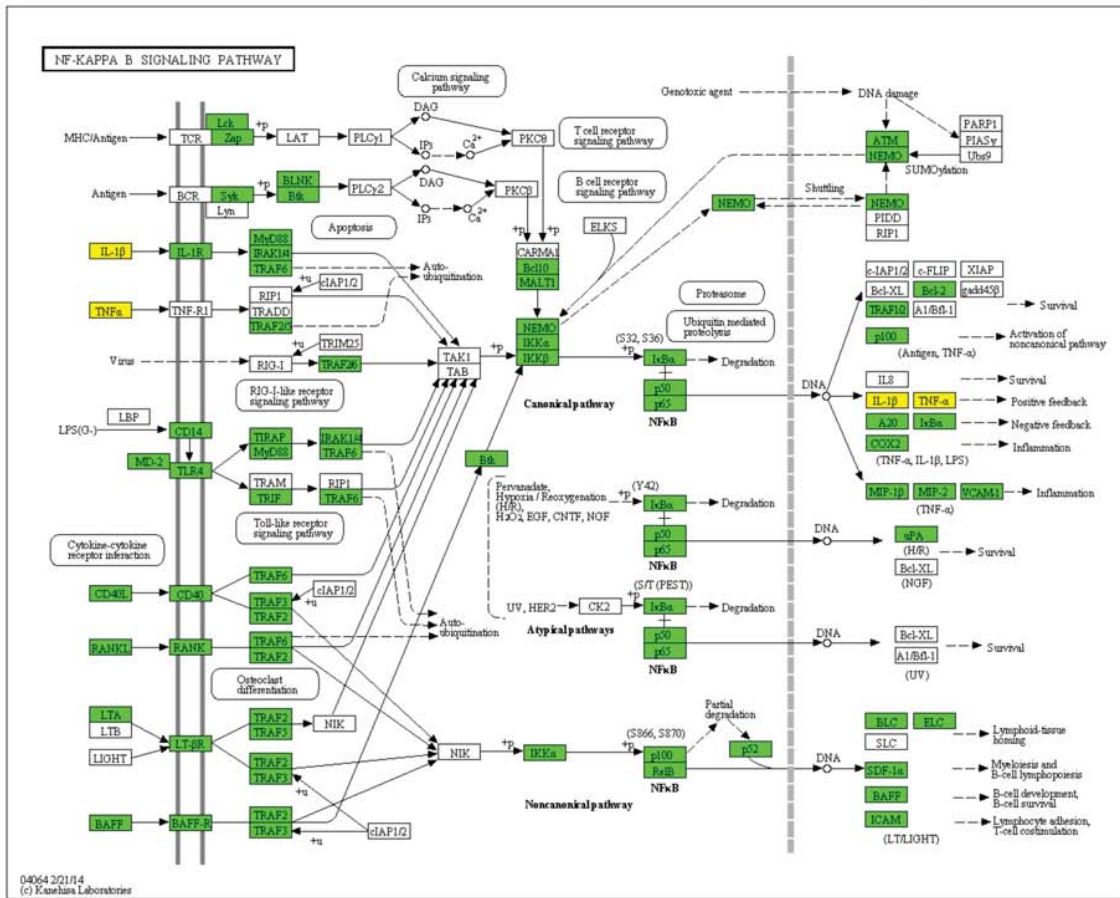


Figure S2

A



B

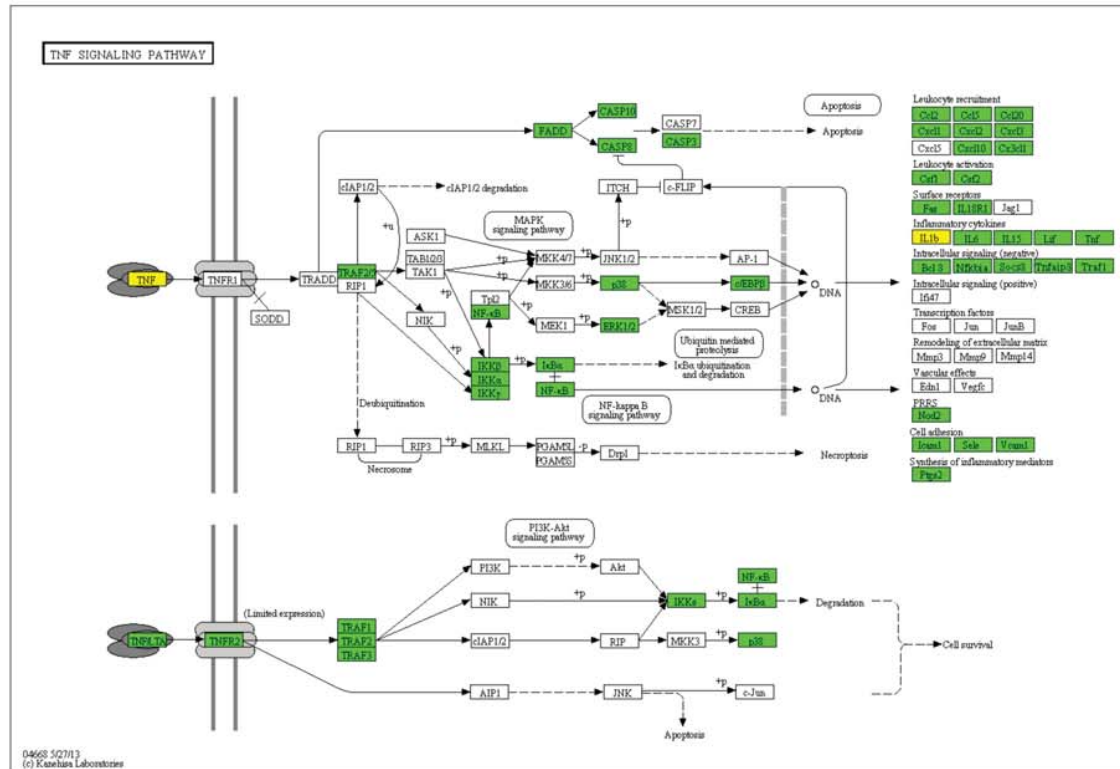






Figure S4





Figure S5

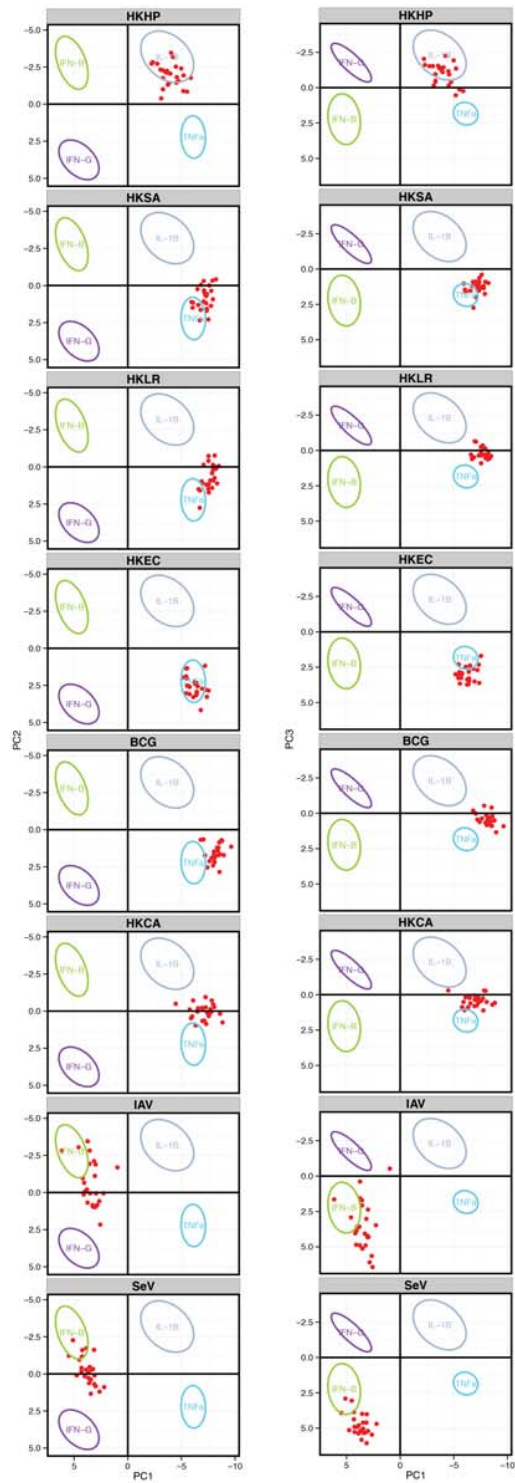
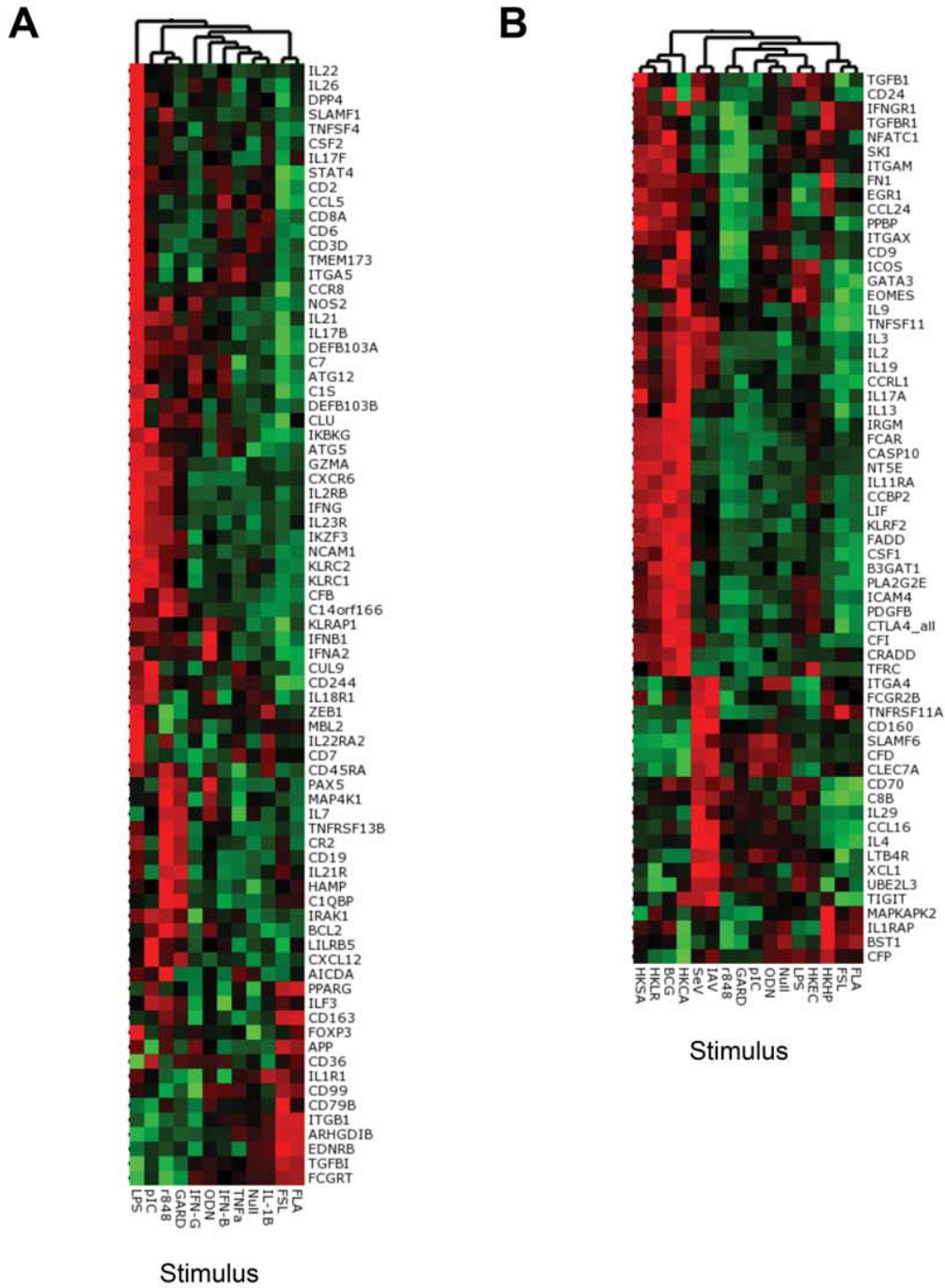





















Figure S6










# **ADDITIONAL INFORMATIONS (STUDY 1)**









**TECAN Script For Automated RNA Purification**

1	Comment	Extraction miRNA a partir de sang lyser dans Trizol
2	Set Variable	ech "Nbr d'echantillons ?" "Nbr d'echantillons ?", 1 - 96
3	Set Variable	Z "Position des tips dans la premiere boite ?" "Position des tips dans la premiere boite ?", 1 - 96
4	Set Variable	Pos_DW "Position de départ dans le deep well 96" "Position de départ dans le deep well 96", 1 - 96
5	Set Variable	Pos_Filter "Position de départ dans la plaque de silice" "Position de départ dans la plaque de silice", 1 - 96
6	Set Variable	Vol Lys "Volume de lysat ?" "Volume de lysat ?", 350 - 700
7	User Prompt	"Verifier les bidons sources" sound : no
8	User Prompt	"placer les 48 premiers tubes de 1 à 48 Le robot s'arrêtera pour les suivants de 49 à 96: presser Retry" sound : no
9	User Prompt	"Placer 3 boites de tips 1000 a filtre en grid 3 Placer les bac ETOH, MW1,MW2 en grid 9 Placer les bacs MW2 , H2O en grid 10 Placer les echantillons centrifugés du grid 11 a 16" sound : no
10	User Prompt	"Placer le deep well sur shaker en grid 20 Montage du vacuum : placer en position Rear la cale 6 par dessus poser la wash plate puis le bloc C et enfin poser la plaque de silice 96 RNA tissue" sound : no
11	Condition	Pos_Filter > 96 FIN
12	Condition	Pos_DW+ech-1 > 96 FIN
13	Wash Tips	 10.0 + 5.0 ml
14	Comment	ETOH 100%
15	Set Variable	sample_cnt = ech
16	Get DiTis	 DiTi 1000ul Filter LiHa
17	Begin Loop	12 times "volume ethanol"
18	Aspirate	  1.5*Vol Lys µl Ethanol "ETOH" (Col. 1, Rows 1-8)
19	Set Variable	WELL_OFFSET = Pos_DW-1

20	Dispense	 	1.5*Vol Lys µl Ethanol free dispense "Labware8" (Col. 1, Rows 1-8) , 1 option
21	Set Variable		WELL_OFFSET = 0
22	End Loop		"volume ethanol"
23	Set Variable		sample_cnt = -1
24	Drop DiTis		"Diti Waste"
25	Wash Tips		10.0 + 5.0 ml
26	Comment		Transfert echantillon sur silice
27	Set Variable		sample_cnt = ech
28	Begin Loop		12 times "volume transfert"
29	Get DiTis		DiTi 1000ul Filter LiHa
30	Aspirate	 	Vol Lys µl EchTrizol "Labware33" (Col. 1, Rows 1-8) , 1 option
31	Set Variable		WELL_OFFSET = Pos_DW-1
32	Dispense	 	Vol Lys µl Mix LyseTrizol "Labware8" (Col. 1, Rows 1-8) , 1 option
33	Set Variable		WELL_OFFSET = 0
34	Set Variable		WELL_OFFSET = Pos_DW-1
35	Aspirate	 	820 µl >> EchTrizol << "Labware8" (Col. 1, Rows 1-8) , 1 option
36	Set Variable		WELL_OFFSET = Pos_Filter-1
37	Dispense	 	820 µl >> EchTrizol << "lavage" (Col. 1, Rows 1-8) , 1 option
38	Set Variable		WELL_OFFSET = 0
39	Drop DiTis		"Diti Waste"
40	End Loop		"volume transfert"
41	Comment		Filtration
42	Set Variable		sample_cnt = -1
43	Wash Tips		10.0 + 5.0 ml








44	Vacuum Separation	SetPressureDiff(200)
45	Vacuum Separation	ApplyVacuumRear(200)
46	Start Timer	1
47	Wait for Timer	Timer 1 : 120 sec
48	Vacuum Separation	VentRear()
49	Vacuum Separation	DeactivateSystem()
50	Comment	Second Transfert echantillon sur silice
51	Set Variable	sample_cnt = ech
52	Begin Loop	12 times "volume transfert"
53	Get DiTis	 DiTi 1000ul Filter LiHa
54	Set Variable	WELL_OFFSET = Pos_DW-1
55	Aspirate	  820 µl >> EchTrizol << "Labware8" (Col. 1, Rows 1-8) , 1 option
56	Set Variable	WELL_OFFSET = Pos_Filter-1
57	Dispense	  820 µl >> EchTrizol << "lavage" (Col. 1, Rows 1-8) , 1 option
58	Set Variable	WELL_OFFSET = 0
59	Drop DiTis	 "Diti Waste"
60	End Loop	"volume transfert"
61	Set Variable	sample_cnt = -1
62	Wash Tips	 10.0 + 5.0 ml
63	Comment	Filtration
64	Vacuum Separation	SetPressureDiff(200)
65	Vacuum Separation	ApplyVacuumRear(200)
66	Start Timer	1
67	Wait for Timer	Timer 1 : 120 sec

68	Vacuum Separation	VentRear()
69	Vacuum Separation	DeactivateSystem()
70	User Prompt	"verification colmatage avant 1er lavage" sound : no
71	Comment	Lavage MW1
72	Set Variable	sample_cnt = ech
73	Get DiTis	 DiTi 1000ul Filter LiHa
74	Begin Loop	12 times "volume ethanol"
75	Aspirate	  600 µl Ethanol "MW1" (Col. 1, Rows 1-8)
76	Set Variable	WELL_OFFSET = Pos_Filter-1
77	Dispense	  600 µl Ethanol "lavage" (Col. 1, Rows 1-8) , 1 option
78	Set Variable	WELL_OFFSET = 0
79	End Loop	"volume ethanol"
80	Set Variable	sample_cnt = -1
81	Drop DiTis	 "Diti Waste"
82	Wash Tips	 10.0 + 5.0 ml
83	Comment	Filtration
84	Vacuum Separation	SetPressureDiff(200)
85	Vacuum Separation	ApplyVacuumRear(200)
86	Start Timer	1
87	Wait for Timer	Timer 1 : 120 sec
88	Vacuum Separation	VentRear()
89	Vacuum Separation	DeactivateSystem()
90	Comment	Lavage MW2
91	Set Variable	sample_cnt = ech

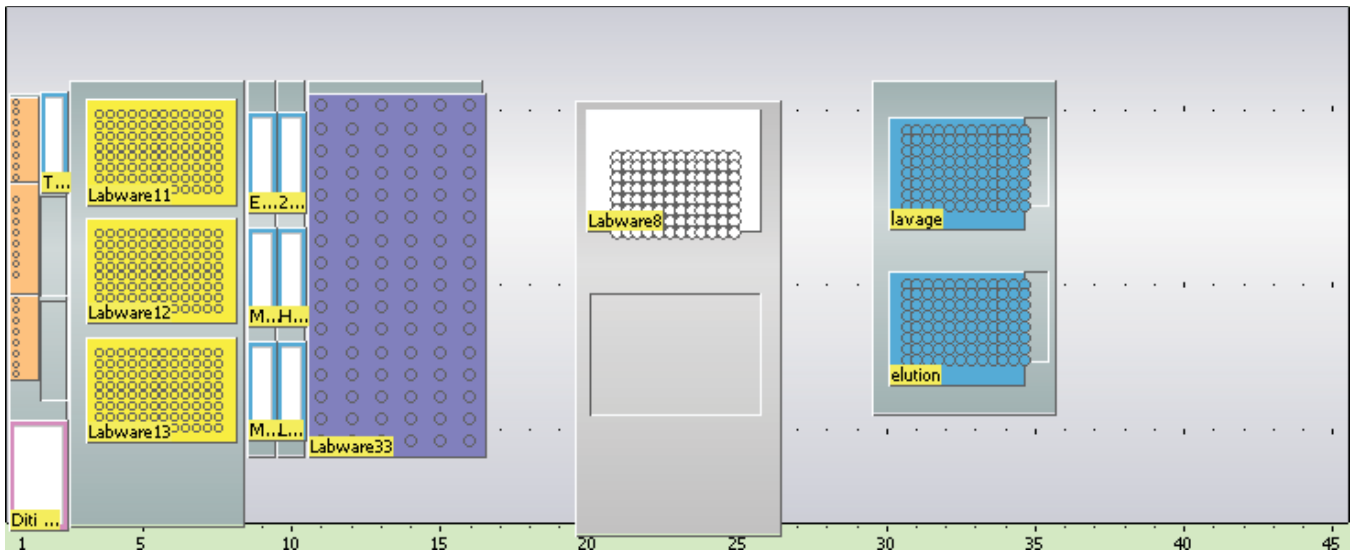
92	Get DiTis		DiTi 1000ul Filter LiHa
93	Begin Loop		12 times "volume ethanol"
94	Aspirate		900 µl Ethanol "MW2" (Col. 1, Rows 1-8)
95	Set Variable		WELL_OFFSET = Pos_Filter-1
96	Dispense		900 µl Ethanol "lavage" (Col. 1, Rows 1-8) , 1 option
97	Set Variable		WELL_OFFSET = 0
98	End Loop		"volume ethanol"
99	Set Variable		sample_cnt = -1
100	Drop DiTis		"Diti Waste"
101	Wash Tips		10.0 + 5.0 ml
102	Comment		Filtration
103	Vacuum Separation		SetPressureDiff(200)
104	Vacuum Separation		ApplyVacuumRear(200)
105	Start Timer		1
106	Wait for Timer		Timer 1 : 60 sec
107	Vacuum Separation		VentRear()
108	Vacuum Separation		DeactivateSystem()
109	Comment		2eme Lavage MW2
110	Set Variable		sample_cnt = ech
111	Get DiTis		DiTi 1000ul Filter LiHa
112	Begin Loop		12 times "volume ethanol"
113	Aspirate		900 µl Ethanol "2MW2" (Col. 1, Rows 1-8)
114	Set Variable		WELL_OFFSET = Pos_Filter-1
115	Dispense		900 µl Ethanol "lavage" (Col. 1, Rows 1-8) , 1 option



116	Set Variable	WELL_OFFSET = 0
117	End Loop	"volume ethanol"
118	Set Variable	sample_cnt = -1
119	Drop DiTis	 "Diti Waste"
120	Wash Tips	 10.0 + 5.0 ml
121	Comment	Filtration
122	Vacuum Separation	SetPressureDiff(200)
123	Vacuum Separation	ApplyVacuumRear(200)
124	Start Timer	1
125	Wait for Timer	Timer 1 : 60 sec
126	Vacuum Separation	VentRear()
127	Vacuum Separation	SetPressureDiff(600)
128	Vacuum Separation	ApplyVacuumRear(600)
129	Start Timer	1
130	Wait for Timer	Timer 1 : 30 sec
131	Vacuum Separation	VentRear()
132	Vacuum Separation	DeactivateSystem()
133	User Prompt	"Enlever la wash Plate Repositionner la plaque" sound : no
134	Comment	Filtration sechage
135	Vacuum Separation	SetPressureDiff(600)
136	Vacuum Separation	ApplyVacuumRear(600)
137	Start Timer	1
138	Wait for Timer	Timer 1 : 600 sec

139	Vacuum Separation	VentRear()
140	Vacuum Separation	DeactivateSystem()
141	User Prompt	"Positionner la plaque de silice sur les tube matrix d'elution sur le vac en front" sound : no
142	Comment	Elution
143	Set Variable	sample_cnt = ech
144	Get DiTis	 DiTi 1000ul Filter LiHa
145	Begin Loop	12 times "volume ethanol"
146	Aspirate	  75 µl Ethanol free dispense "H2O" (Col. 1, Rows 1-8)
147	Set Variable	WELL_OFFSET = Pos_Filter-1
148	Dispense	  75 µl >> Water free dispense << "elution" (Col. 1, Rows 1-8) , 1 option
149	Set Variable	WELL_OFFSET = 0
150	End Loop	"volume ethanol"
151	Set Variable	sample_cnt = -1
152	Drop DiTis	 "Diti Waste"
153	Wash Tips	 10.0 + 5.0 ml
154	Comment	incubation
155	Start Timer	1
156	Wait for Timer	Timer 1 : 120 sec
157	Comment	Filtration
158	Vacuum Separation	SetPressureDiff(600)
159	Vacuum Separation	ApplyVacuumFront(600)
160	Start Timer	1
161	Wait for Timer	Timer 1 : 120 sec
162	Vacuum Separation	VentFront()

Vacuum Separation	DeactivateSystem() 163
User Prompt	"Aller mettre les eluats a 4°C" sound : no 164
Comment	FIN 165



## **ADDITIONAL PUBLICATIONS**

***(MILIEU INTERIEUR) :***

- 1. Thomas et al ., CI 2015**
- 2. Duffy et al ., Immunity 2014**
- 3. Hasan et al ., CI 2015**
- 4. Chen et al ., CI 2015**



available at [www.sciencedirect.com](http://www.sciencedirect.com)

Clinical Immunology

[www.elsevier.com/locate/yclim](http://www.elsevier.com/locate/yclim)



## The *Milieu Intérieur* study — An integrative approach for study of human immunological variance



Stéphanie Thomas<sup>a,b,c</sup>, Vincent Rouilly<sup>a,d</sup>, Etienne Patin<sup>e,f</sup>,  
Cécile Alanio<sup>a,b,c</sup>, Annick Dubois<sup>g</sup>, Cécile Delval<sup>g</sup>,  
Louis-Guillaume Marquier<sup>h</sup>, Nicolas Fauchoux<sup>h</sup>, Seloua Sayegrih<sup>h</sup>,  
Muriel Vray<sup>i</sup>, Darragh Duffy<sup>a,b,c</sup>, Lluís Quintana-Murci<sup>e,f,\*</sup>,  
Matthew L. Albert<sup>a,b,c,j,\*\*</sup>, for The *Milieu Intérieur* Consortium<sup>¶</sup>

The *Milieu Intérieur* Consortium is composed of the following team leaders:  
Laurent Abel<sup>1</sup>, Andres Alcover, Philippe Bouso, Pierre Bruhns, Ana Cumano,  
Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano,  
James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga,  
Antonio Freitas, Odile Gelpi, Ivo Gomperts-Boneca, Serge Hercberg<sup>2</sup>,  
Olivier Lantz<sup>3</sup>, Claude Leclerc, Hugo Mouquet, Sandra Pellegrini,

**Abbreviations:** AbdoCM, abdominal circumference; ALP, alkaline phosphatase; ALT, alanine aminotransferase; ANSM, Agence Nationale de sécurité du médicament et des produits de santé; AST, aspartate aminotransferase; BASO, basophil;  $\beta$ -HCG, beta-human chorionic gonadotropin; BILI, bilirubin; BMI, body mass index; BUN, blood urea nitrogen; Ca, calcium; Cl, chloride; CLT, clinical laboratory test; CMV, cytomegalovirus; CREAT, creatinine; CRF, case report form; CRO, clinical research organization; CRP, C-reactive protein; CVD, cardiovascular disease; df, degree of freedom; DYSBP1, diastolic blood pressure; EBV, Epstein–Barr virus; ECG, electrocardiogram; eCRF, electronic case report form; EOS, eosinophil; GFR, glomerular filtration rate; GGT, gamma-glutamyl transpeptidase; GLUC, glucose; HAS, human serum albumin; HBV, hepatitis B virus; hCG, human chorionic gonadotropin; HCO<sub>3</sub>, bicarbonate; HCT, hematocrit; HCV, hepatitis V virus; HDL, high density lipoprotein; HGB, hemoglobin; HIV, human immunodeficiency virus; HSA, human serum albumin; HTLV, human T cell lymphotropic virus; IgA, Immunoglobulin A; IgE, Immunoglobulin E; IgG, Immunoglobulin G; IgM, Immunoglobulin M; INSEE, Institut National de la statistique et des études économiques; IQR, interquartile range; K, potassium; LabEx, Laboratoire d'Excellence; LDL, low density lipoprotein; LYMPH, lymphocyte; MAMP, microbial-associated molecular pattern; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; mM, glucose level; MONO, monocyte; Na, sodium; NEUTR, neutrophil; P, phosphorus; PBMC, peripheral blood mononuclear cells; PCA, principal component analysis; PHOS, phosphorus; PLT, platelet; RBC, red blood cell; RT, room temperature; SNP, single nucleotide polymorphisms; SOPs, standard operating procedures; SYSBP1, systolic blood pressure; TCHOL, total cholesterol; TG, triglyceride; TPROT, total protein; TRIGLY, triglyceride; UA, urinalysis; UAC, uric acid; V0, visit 0; V1, visit 1; V2, visit 2; WBC, white blood cell.

☆ One sentence summary: This report presents the first demographic data from the *Milieu Intérieur* Consortium, which has established a 1000-person healthy population-based study, for assessing the genetic and environmental determinants of human immunologic variance.

\* Correspondence to: L. Quintana-Murci, Unit of Human Evolutionary Genetics, CNRS URA3012/Institut Pasteur, 28, rue du Dr. Roux, 75724 Paris Cedex 15, France.

\*\* Correspondence to: M.L. Albert, Unit of Dendritic Cell Immunobiology, Inserm U818/Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 85 48.

E-mail addresses: [quintana@pasteur.fr](mailto:quintana@pasteur.fr) (L. Quintana-Murci), [albertm@pasteur.fr](mailto:albertm@pasteur.fr) (M.L. Albert).

¶ unless otherwise indicated, partners are located at Institut Pasteur, Paris.

<http://dx.doi.org/10.1016/j.clim.2014.12.004>

1521-6616/© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Stanislas Pol<sup>4</sup>, Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis<sup>3</sup>, Frédéric Tangy, Eric Tartour<sup>5</sup>, Antoine Toubert<sup>6</sup>, Marie-Noëlle Ungeheuer, Lluís Quintana-Murci<sup>§</sup>, Matthew L. Albert<sup>§</sup>

<sup>1</sup> Hôpital Necker, France

<sup>2</sup> Université Paris 13, France

<sup>3</sup> Institut Curie, France

<sup>4</sup> Hôpital Cochin, France

<sup>5</sup> Hôpital Européen George Pompidou, France

<sup>6</sup> Hôpital Saint-Louis, France

<sup>a</sup> Center for Human Immunology, Institut Pasteur, Paris, France

<sup>b</sup> Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris, France

<sup>c</sup> INSERM U818, France

<sup>d</sup> Center for Bioinformatics, Institut Pasteur, Paris, France

<sup>e</sup> Laboratory of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France

<sup>f</sup> CNRS URA3012, France

<sup>§</sup> PIRC, Institut Pasteur, Paris, France

<sup>h</sup> Biotrial, Rennes, France

<sup>i</sup> Unit of Emerging Diseases Epidemiology, Institut Pasteur, Paris, France

<sup>j</sup> INSERM UMS20, France

Received 17 July 2014; accepted with revision 1 December 2014

Available online 3 January 2015

## KEYWORDS

Healthy donor;  
Cohort design;  
Immune phenotypes;  
Baseline serologic data;  
Metabolic syndrome;  
CMV

**Abstract** The *Milieu Intérieur* Consortium has established a 1000-person healthy population-based study (stratified according to sex and age), creating an unparalleled opportunity for assessing the determinants of human immunologic variance. Herein, we define the criteria utilized for participant enrollment, and highlight the key data that were collected for correlative studies. In this report, we analyzed biological correlates of sex, age, smoking-habits, metabolic score and CMV infection. We characterized and identified unique risk factors among healthy donors, as compared to studies that have focused on the general population or disease cohorts. Finally, we highlight sex-bias in the thresholds used for metabolic score determination and recommend a deeper examination of current guidelines. In sum, our clinical design, standardized sample collection strategies, and epidemiological data analyses have established the foundation for defining variability within human immune responses.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

## 1. Introduction

Susceptibility to infections, disease severity, and response to medical therapies or vaccines are highly variable from one individual to another. Medical practices and public health policies typically take a 'one size fits all' model for disease management and drug development. This approach ignores individual heterogeneity in immune responses that likely impacts the response to therapy or the efficiency and development of side effects secondary to vaccine or treatment administration. Due to the complexity of immune responses at

the individual and population level, it has been challenging thus far to define the borders of a healthy immune system as well as the parameters (genetic, epigenetic, and environmental) that drive its naturally-occurring variability. In particular, such assessments require large sample sizes, consensus for defining "healthy", and standardized protocols for sample recruitment. In this context, the *Milieu Intérieur* Consortium initiated in September 2012 a cross-sectional healthy population-based study called "*Genetic & Environmental Determinants of Immune Phenotype Variance: Establishing a Path Towards Personalized Medicine (ID-RCB Number: 2012-A00238-35)*".

The overall aim of the *Milieu Intérieur* study is to assess the factors underlying immunological variance within the general healthy population. The primary objective is to

<sup>§</sup> co-coordinators of the *Milieu Intérieur* Consortium. Additional information can be found at: <http://www.pasteur.fr/labex/milieu-interieur>.

define genetic and environmental factors that contribute to the observed heterogeneity in immune responses. This will be realized by characterizing and integrating (i) every-day life habits through an extensive questionnaire; (ii) genomic variability using genome-wide SNP genotyping and whole-exome sequencing; (iii) metagenomic diversity based on sequence analysis of bacterial, fungal and viral populations in fecal and nasal samples; (iv) induced transcriptional and protein signatures by whole microbes, microbial-associated molecular pattern (MAMP) agonists, medically relevant cytokines, or stimulators of the T cell response; and (v) variability in levels of circulating immune cell populations based on flow cytometry. The secondary objective is to establish a cell bank, including EBV-transformed B cell lines and fibroblasts from genetically annotated healthy individuals for use in mechanistic studies. To achieve the above-mentioned objectives, a total of 1000 healthy volunteers, descendants of mainland French persons for at least three generations, split equally by sex (1:1 sex ratio) and stratified across five-decades of life were recruited.

Herein, we present the socio-demographic and biological parameters that define our healthy donor cohort. Through unbiased statistical approaches, we identified known sex- and age-associated phenotypes, thus confirming the overall integrity of the data and validating our population sample as a reference for the healthy French population. Additional analyses provided new insight into the definition and risk factors of metabolic syndrome. Finally, we identified dependent and independent variables among the collected meta-data, results that will be applied to future association studies. This unique healthy donor population study may ultimately serve as a control reference sample for future disease based studies.

## 2. Materials and methods

### 2.1. Study objectives

In the context of a French scientific initiative, financed through the Investissement d'Avenir as part of a *Laboratoire d'Excellence* (LabEx) research program, the *Milieu Intérieur* Consortium was developed with the objective to define the determinants of human immune variance.

### 2.2. Clinical protocol and implementation

The clinical study was approved by the *Comité de Protection des Personnes – Ouest 6* (Committee for the protection of persons) on June 13th, 2012 and by the French *Agence nationale de sécurité du médicament* (ANSM) on June 22nd, 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](https://clinicaltrials.gov/ct2/show/study/NCT01699893) (study# NCT01699893).

Our strategy to define the parameters of a healthy population included the gathering of a working group composed of experts representing different clinical (medical biology, regenerative medicine, allergy, pediatrics, nutrition, psychiatry, lab medicine) and scientific (immunology, genetics, epidemiology, methodology, sociology, gut microbiota)

specialties to help establish the criteria for qualifying an individual as a “healthy” donor while preserving the feasibility of recruitment and permitting robust statistical analysis. Specifically, this working group discussed the general eligibility criteria to pre-screen subjects (age, sex, BMI, self-reported ancestry, relatedness with the other subjects) and identified specific exclusion criteria that may impact the immune system and/or study procedures (e.g., chronic diseases known to involve the immune system, subjects with skin disorders that would compromise skin biopsy, etc.). Known medical, physiological, and behavioral factors with potential to affect immune cell activities or the microbiota environment were thoroughly reviewed and retained on the basis of their impact on the objective of our project, while preserving the feasibility of enrollment. We further considered the prevalence of donor characteristics, excluding those phenotypes that are below 1% in the population (e.g., peanut allergy), to ensure sufficient power for association studies. Efforts were made to avoid the selection of individuals following too conservative criteria (i.e., “super healthy” population), as this would compromise the underlying purpose of the study. A Scientific Advisory Board helped to develop and refine the study protocol, donor information and consent forms. They also provided oversight for ensuring consistency in screening, enrollment, body site sampling, and compliance with regulatory and data management requirements.

Laboratory protocols were standardized and staff members were trained in sample preparation protocols. Two risk assessments audits were conducted during the training period to refine sample handling and technical protocols. The clinical study opened at the investigator site (Biotrial, Rennes, France) on September 7th, 2012 and the first sample was collected on September 17th, 2012. All subjects provided informed consent prior to enrollment in the study. Subjects received compensation for their participation.

### 2.3. Subject screening and recruitment

A pre-existing donor database composed of ~110,000 donors was used for pre-screening potential participants in accordance with the study criteria. Additional advertising and website recruitment campaigns were launched in order to complete strata not sufficiently represented in the donor database. Eligibility was assessed by telephone interview and confirmed during a preliminary information meeting about the objectives of the research. Interested participants that met pre-screening criteria returned for the enrollment visit (referred to as V0). During V0, eligibility criteria were assessed in two stages: first, based on demographical data and clinical examination; and second, by analysis of blood and urine samples that were sent for clinical laboratory testing (Table S1). During the course of their participation in the *Milieu Intérieur* project, subjects were informed and encouraged to participate in a non-interventional French nutritional survey, *Etude Nutrinet-Santé* ([www.etude-nutrinet-sante.fr](http://www.etude-nutrinet-sante.fr)) [25].

Upon receiving the clinical laboratory results, and confirming that all inclusion and exclusion criteria were respected, all subjects were invited to return for the inclusion visit (referred to as V1). Based on a defined randomization strategy, 500 subjects participated in a second visit (referred

to as V2) for repeat sampling (Fig. 1). V0 and V1 were scheduled with a 4–14 day interval; and V2 took place 14–42 days after V1. Of those that were randomized for repeat collections, 340 donors consented for a skin biopsy at V1 (n.b. the number of subjects with skin biopsy at V1 was restricted due to technical constraints). The financial compensation for participating in V0 was 50€, 150€ for V1, 100€ for V2 and 50€ for the skin biopsy.

## 2.4. Cross-sectional study

The *Milieu Intérieur* sample is composed of 1000 healthy volunteers, descendants of mainland French persons for at least three generations, stratified according to sex with a 1:1 ratio (500 subjects by sex); and age (5 decades of age: [20–29], [30–39], [40–49], [50–59] and [60–69] years, with 200 subjects per stratum). Subjects were randomized for a single or repeated collection (50% per stratum returned for V2). All donors were recruited by Biotrial Inc., a clinical research organization (CRO) based in Rennes, France. From September 17th, 2012 to August 8th, 2013, a total of 1238 donors were screened and 1012 healthy donors were enrolled. Twelve donors withdrew, so the final sample collection was composed of 1000 persons.

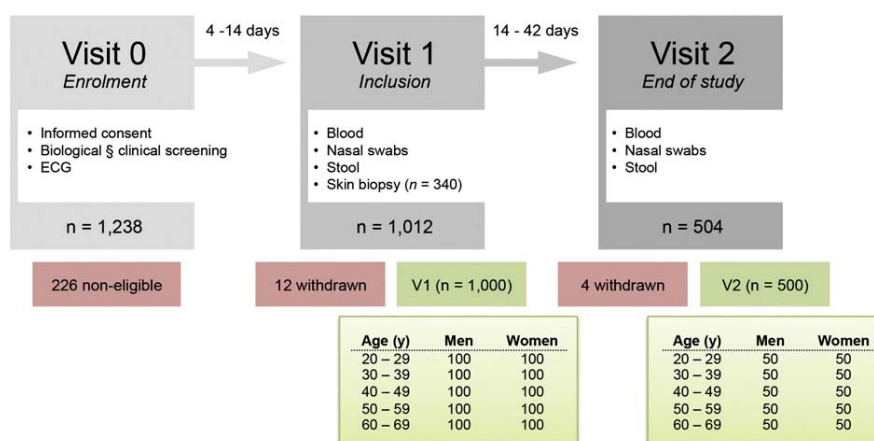
## 2.5. Inclusion/exclusion criteria

The study design concerned the definition of “healthy” in accordance with the goal to maximize our ability to associate genetic and epigenetic variation with defined phenotypes. This was achieved by establishing a detailed list of inclusion and exclusion criteria that ensured the recruitment of volunteers with a minimally perturbed immune

system. Briefly, donors could not have evidence of, or report a history of neurological or psychiatric disorders, or severe/chronic/recurrent pathological conditions. Other exclusion criteria included: history or evidence of alcohol abuse, recent use of illicit drugs (including cannabis), recent vaccine administration, and recent use of immune modulatory agents. To avoid the influence of hormonal fluctuations in women during the peri-menopausal phase, only pre- or post-menopausal women were included. To avoid the presence of population structure in our study population (i.e., highly variable genetic backgrounds due to different ancestry), which would impact upon the power to detect genotype-to-phenotype associations, we restricted our study to individuals of European-descent, i.e., French citizens whose ancestry for three generations was of Metropolitan French origin (i.e., the subject’s parents and grandparents were born in continental France).

## 2.6. Physical and clinical laboratory testing

After initial evaluation for recruitment criteria, additional physical examination and clinical laboratory testing were performed at visit V0 in order to fully include the donors. Donor BMI was restricted to  $\geq 18.5$  and  $\leq 32$  kg/m<sup>2</sup>. 20 mL of blood sample (collected at V0 and V2, for repeat sampling) was used for clinical chemistry, hematologic and serologic assessments. A urinary human chorionic gonadotropin (hCG) test was performed on female donors, and urine toxicology screens for cannabinoid use, proteinuria and glycosuria were conducted on all donors. All clinical laboratory assessments were performed at the certified *Laboratoire de biologie médicale, Centre Eugene Marquis* (Rennes, France).



**Figure 1** Schematic representation of donor recruitment for the *Milieu Intérieur* study. To include 1000 healthy persons stratified according to sex (500 men, 500 women) and age (200 donors per decade of life, 20–69 years of age), we enrolled a total of 1238 individuals at visit 0 (Enrollment). Of those screened, 226 donors were considered non-eligible for reasons of consent withdrawal ( $n = 54$ ), past medical history ( $n = 67$ ), identification of an exclusion criteria during the onsite physical examination ( $n = 54$ ), or during laboratory testing ( $n = 51$ ) (see Fig. S1). An additional 16 donors withdrew consent in the course of the study. During visit 1, whole blood, fecal samples and nasal swabs were collected. Punch biopsies of the skin were obtained from 340 of these donors. Half of the subjects were randomly selected (stratified by age and sex) to return for a visit 2, when repeat sampling of whole blood, fecal samples and nasal swabs was performed. Detailed medical histories and questionnaires were completed from all donors, recorded by medical personnel using an electronic case report form.



Biochemistry tests, immunological analysis, and viral serologies were performed on serum-separator tubes using AU 400 Olympus, DXC 660 I, Advanced 2020, DXI (Beckman Coulter), UF 50 Sysmex (Biomérieux) and Modular E170 (Roche) analyzers; Modular E170 (Roche), IRMA (Immunotech), RIA (Labodia), and Hydrasys (Sebia) systems; and DXI immunoassay system (Beckman Coulter), respectively. Hematology analysis were performed on EDTA tubes and coagulation tests were performed on citrate tubes using LH750 (Beckman Coulter) and STA-R (Stago) analyzers respectively.

## 2.7. Sample collections and storage

Blood, nasal swabs and stool samples were collected from all donors according to established protocols. For 500 individuals, samples were collected at V1 only; and for the remaining 500 donors, samples were collected at V1 and V2 – separated by 14–42 days – thus providing validation samples to be used in phenotypic studies. For donors randomized for two sample collections, biopsies of the skin were performed once at V1, in 340 donors.

From each volunteer, 20 mL of blood was collected into 2 Na Heparin tubes, and 5 mL of blood into 1 EDTA tubes for cytometric studies and banking of DNA, respectively. These tubes were maintained at 18–25 °C, during daily transport to Institut Pasteur (Paris), and processed within 6 h of collection. An additional 50 mL of blood was collected using a pre-heparinized large-bore syringe. This sample was aliquoted into 40 – 1 mL TruCulture® tubes within 15 min of collection. The TruCulture® systems were developed to provide reproducible induction of innate or adaptive immune responses and are described elsewhere [26]. After stimulation, the liquid supernatants from the TruCulture® tubes were aliquoted, and the cell pellet was stabilized in Trizol. Both samples were stored at –80 °C.

Stool samples were collected in a double-lined sealable bag with the outer bag containing a GENbag Anaer atmosphere generator (Aerocult, Biomerieux), used to maintain anaerobic conditions, and an anaerobic indicator strip (Anaerostest®, Merck Millipore) to record maintenance of the anaerobic atmosphere. Subjects were asked to produce the fecal specimen at their home within 24 h before their scheduled visits (V1, V2). Upon reception at the clinical site, the specimen was aliquoted into cryotubes and stored at –80 °C.

Nasal swabs were obtained with sterile, dry flocked swabs (FLOQSwab™). Right and left nostrils were sampled separately. All swabs were stored in stabilization media and frozen at –80 °C.

Skin punch biopsies were performed under local anesthesia. The biopsy was taken using a sterile single use biopsy punch (7 mm \* 3 mm round dermal punch). The material collected was shipped the same day of the collection at 4 °C to Genethon (Evry, Ile de France, France) where human fibroblast cell lines were generated and aliquots stored.

The processing of each donor involved the production and registration of more than 180 tubes. To ensure effective traceability of all samples, a customized software system was developed for managing 2D barcoded tubes. A central sample database has been established to aggregate all sample information for each donor, visit, and sample type.

## 2.8. Case report forms

Detailed medical histories and questionnaires collecting general information about socio-demographic, lifestyle and family health history were recorded in an electronic case report form. For example, the questionnaire collected information concerning family status, income, occupational status and educational level, smoking habits, alcohol intake, sleeping habits, depressive symptoms, family medical history and nutritional behavior and habits (for details, see supplementary material: case report forms).

## 2.9. Statistical analyses

Statistical analysis was performed using the Open Source R Software, version 3.0.1 [27]. All statistical graphics were generated using the 'ggplot2' package, version 0.97 [28]. The hierarchical clustering of our continuous explanatory variables was based on the Spearman's correlation score (Rs) using the 'hclust' function available from the base functions. Random Forest (RF) models [2] were built using the 'randomForest' package (version 4.6–7). For each RF model built (sex, age categories, smoking status) a forest of 1000 trees was computed, and the 'mtry' parameter was set to be the square root of the number of available explanatory variables. When investigating outliers in our sample, we used a z-score based criterion. For a given metric, we considered a donor as an outlier if its measurement was 3 standard deviations away from the mean of the whole sample. Principal component analysis (PCA) on the outlier cases of our dataset was performed with the 'FactoMineR' package version 1.25. Regression analyses were conducted using the *glm* function in R. Levels of immunoglobulins were log-transformed and standardized, prior to regression analyses. The representative nature of the cohort was assessed by stratified sampling: 500 individuals were sampled 10,000 times among all cohort participants, in order to match the proportions of males and females and of 10-decades age groups observed in the general population. Public data from the *Institut National de la Statistique et des Etudes Economiques* (INSEE; National Institute of Statistics and Economic Studies) were retrieved for the entire *Ille-et-Vilaine* French department and the city of Rennes (<http://www.insee.fr/en/default.asp>).

## 3. Results

### 3.1. Sample and data overview

From September 17th, 2012 to August 8th, 2013, a total of 1238 donors were screened and 1012 healthy donors were enrolled (Fig. 1). The reasons for excluding the 226 pre-screened donors included withdrawal of consent ( $n = 54$ ), as well as medical history ( $n = 67$ ), physical exam findings ( $n = 54$ ) or laboratory test results ( $n = 51$ ) that were not in accordance with the defined inclusion or non-exclusion criteria (Fig. S1, Table S1). Questionnaires were completed and clinical laboratory testing was performed at visit V0 (Tables 1–2, Tables S1–3). Among those enrolled, 12 donors withdrew consent during the collection phase of the protocol.

This resulted in a final set of 1000 subjects, with all donors having completed an evaluation at visit V1 and 50% of them (500 subjects) returning for evaluation at visit V2. During V1, 340 had a skin biopsy.

Donor recruitment was conducted in the vicinity of the city of Rennes, in the *Ille-et-Vilaine* French region. We first compared the socio-economic characteristics of the *Milieu Intérieur* cohort to those of the general population of this region (Table S4), after adjustment to match regional age and sex stratification (see Methods section). We observed ~10% higher unemployment levels in the *Milieu Intérieur* cohort (16.9% with 95% confidence interval (CI) [14.7%–19.1%]), when compared to the *Ille-et-Vilaine* region or the

city of Rennes (6.0% and 8.2%, respectively) (Table S5). The cohort also contained a higher proportion of retired persons (16.6% [15.1%–17.9%] versus 8.1% and 4.9%) (Table S5). Among employed people, socio-professional categories of the *Milieu Intérieur* donors were biased towards more employees and fewer laborers. We also observed that the educational level of the *Milieu Intérieur* donors was generally higher. Finally, 42.5% [39.5%–45.4%] of participants were renters, a value that is intermediate between those of the *Ille-et-Vilaine* region and the city of Rennes (33.1% and 59.5%, respectively) (Table S5), consistent with the fact that donors reside in Rennes as well as in surrounding areas.

**Table 1** Sample collections obtained from study subjects.

	Visit 0	Visit 1	Visit 2
	20 mL	87 mL	83 mL
<b>Whole blood collection</b>			
CLT <sup>a</sup> Complete blood count: RBC count, HCT, HGB, MCV, MCH, WBC count, NEUTRO, MONO, LYMPHO, EOS, BASO, PLT count	X		X
CLT Blood electrolytes: Na, K, Ca, P, Cl, HCO <sub>3</sub>	X		
CLT Liver function tests: HSA, ALP, AST, ALT, GGT, BILI, TPROT	X		
CLT Inflammation: CRP	X		X
CLT Renal function tests: BUN, CREAT, UA	X		
CLT Lipids/metabolism: GLUC, TCHOL, LDL, HDL, TRIGLY	X		
CLT Serology: HBV (HBs Ag), HCV (anti HCV IgG, viral load if Ab+), HIV (anti-HIV IgM, IgG), CMV (anti-CMV IgG), HTLV-1 (anti-HTLV-1 IgG), influenza (anti-Influenza IgG)	X		
CLT Immunoglobulin electrophoresis: serum immunoglobulin concentrations (IGM, IGG, IGA, IGE)	X		
R <sup>b</sup> Immunophenotyping (Na Heparin tube): cytometric analysis for major subsets of immune cells in circulation		X	X
R Functional immune stimulation (Na Heparin syringe): TruCulture tubes ×40		X	X
R Genetic tests (EDTA tube): TruCulture tubes ×40		X	
<b>Urine collection</b>			
	>5 mL		>5 mL
CLT Biochemistry: proteinuria, glycosuria (dipstick)	X		
CLT Pregnancy test: bHCG concentration (women only)	X		X
CLT Toxicology: cannabinoids	X		
<b>Fecal sample collection</b>			
		>100 g	>100 g
R Enterotyping: bacterial, viral, fungal strains		X	X
<b>Nasal swab</b>			
		2 swabs	2 swabs
R Enterotyping: bacterial, viral, fungal strains		X	X
<b>Biopsy</b>			
	7 mm punch		
R Punch biopsy of skin	X (n = 340)		

<sup>a</sup> CLT, clinical laboratory test; RBC, red blood cell; HCT, hematocrit; HGB, hemoglobin; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; WBC, white blood cell, NEUTRO, neutrophil; LYMPHO, lymphocyte; EOS, eosinophil; BASO, basophil; PLT, platelet; Na, sodium; K, potassium; Ca, calcium; P, phosphorus; Cl, chloride; HCO<sub>3</sub>, bicarbonate; HSA, human serum albumin; ALP, alkaline phosphate; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transpeptidase; BILI, bilirubin; TPROT, total protein; CRP, C-reactive protein; BUN, blood urea nitrogen; CREAT, creatinine; UA, urinalysis; GLUC, glucose; TCHOL, total cholesterol; LDL, low density lipoprotein; HDL, high density lipoprotein; TRIGLY, triglycerides; HBV, hepatitis B virus; HCV, hepatitis C virus; HIV, human immunodeficiency virus; CMV, cytomegalovirus; HTLV, human T cell lymphotropic virus; βHCG, beta-human chorionic gonadotropin.

<sup>b</sup> R, Research tests.

**Table 2** Socio-demographic information for study subjects.

Donor characteristics	Total (n = 1000)		Male (n = 500)		Female (n = 500)		20–29 years (n = 200)		30–39 years (n = 200)		40–49 years (n = 200)		50–59 years (n = 200)		60–69 years (n = 200)	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
<b>BMI *</b>																
18 < BMI ≤ 25	635	63.5	283	56.6	352	70.4	160	80	135	67.5	128	64	111	55.5	101	50.5
25 < BMI ≤ 30	300	30	181	36.2	119	23.8	34	17	55	27.5	54	27	72	36	85	42.5
30 < BMI ≤ 32	65	6.5	36	7.2	29	5.8	6	3	10	5	18	9	17	8.5	14	7
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
<b>Education §</b>																
No diploma	38	3.8	20	4	18	3.6	6	3	4	2	10	5	10	5	8	4
Primary school certificate only CAP, BEP, apprenticeship certificate, BEPC (High school diploma equivalent)	46	4.6	22	4.4	24	4.8	0	0	1	0.5	1	0.5	11	5.5	33	16.5
Baccalaureate or technician’s certificate	332	33.2	170	34	162	32.4	34	17	47	23.5	84	42	84	42	83	41.5
Higher education (no professional degree)	268	26.8	130	26	138	27.6	64	32	65	32.5	44	22	50	25	45	22.5
Higher education (Masters, PhD, engineer’s diploma, MD, etc.)	156	15.6	75	15	81	16.2	49	24.5	36	18	31	15.5	26	13	14	7
Total	160	16	83	16.6	77	15.4	47	23.5	47	23.5	30	15	19	9.5	17	8.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
<b>Employment</b>																
Steady job	510	51	247	49.4	263	52.6	75	37.5	148	74.0	155	77.5	118	59	14	7
Unemployed	158	15.8	91	18.2	67	13.4	38	19	43	21.5	37	18.5	34	17	6	3
Student	74	7.4	41	8.2	33	6.6	71	35.5	3	1.5	0	0	0	0	0	0
Looking for first job	16	1.6	5	1.0	11	2.2	14	7	1	0.5	1	0.5	0	0	0	0
Housewife/househusband	21	2.1	2	0.4	19	3.8	2	1	4	2.0	5	2.5	8	4	2	1
Retired	215	21.5	114	22.8	101	20.2	0	0	0	0.0	2	1	40	20	173	86.5
NA	6	0.6	0	0.0	6	1.2	0	0	1	0.5	0	0	0	0	5	2.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Full-time	406	40.6	219	43.8	187	37.4	65	32.5	115	57.5	122	61	96	48	8	4
Part-time	135	13.5	38	7.6	97	19.4	28	14	38	19	36	18	24	12	9	4.5
Not answered	459	45.9	243	48.6	216	43.2	107	53.5	47	23.5	42	21	80	40	183	91.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Exclusively during the day	369	36.9	155	31	214	42.8	59	29.5	102	51	110	55	86	43	12	6
Exclusively during the night	36	3.6	12	2.4	24	4.8	7	3.5	10	5	10	5	8	4	1	0.5
Without fixed hours	135	13.5	90	18	45	9	27	13.5	41	20.5	37	18.5	26	13	4	2
Not answered	460	46	243	48.6	217	43.4	107	53.5	47	23.5	43	21.5	80	40	183	91.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
<b>Socio-professional category</b>																
Farmer	10	1	6	1.2	4	0.8	1	0.5	2	1	1	0.5	2	1	4	2
Artisans, tradesman or company director	46	4.6	36	7.2	10	2	4	2	11	5.5	10	5	7	3.5	14	7
Senior executive or independent profession	42	4.2	29	5.8	13	2.6	1	0.5	15	7.5	4	2	7	3.5	15	7.5
Middle management	113	11.3	59	11.8	54	10.8	3	1.5	16	8	27	13.5	28	14	39	19.5
Employee	507	50.7	197	39.4	310	62	82	41	116	58	115	57.5	106	53	88	44
Labourer	100	10	76	15.2	24	4.8	11	5.5	19	9.5	19	9.5	27	13.5	24	12
Other categories (e.g. artist, clergy, soldier, police officer)	60	6	48	9.6	12	2.4	7	3.5	11	5.5	18	9	15	7.5	9	4.5
Not answered	122	12.2	49	9.8	73	14.6	91	45.5	10	5	6	3	8	4	7	3.5

(continued on next page)



Table 2 (continued)

Donor characteristics	Total (n = 1000)		Male (n = 500)		Female (n = 500)		20–29 years (n = 200)		30–39 years (n = 200)		40–49 years (n = 200)		50–59 years (n = 200)		60–69 years (n = 200)	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
<b>Contraception †</b>																
Intrauterine device (IUD)					72	14.4	7	7	29	29	29	29	7	7	0	0
Oral Contraception					117	23.4	55	54	37	37	20	20	5	5	0	0
Male or female condomn					94	18.8	29	28	26	26	32	32	7	7	0	0
Tubal ligation					7	1.4	0	0	2	2	5	5	0	0	0	0
Other method of contraception					14	2.8	7	7	3	3	2	2	2	2	0	0
None					13	2.6	4	4	3	3	5	5	1	1	0	0
Not answered (or not asked)					183	36.6	0	0	0	0	7	7	78	78	100	100
Total					500	100	102	100	100	100	100	100	100	100	100	100

\* Study inclusion criteria set limits for BMI.

<sup>§</sup> The certificat d'études primaires (CEP) was a diploma awarded at the end of elementary primary education in France (from 11 to 13 years inclusive until 1936) and certifying that the student had acquired basic skills in writing, reading, mathematics, history, geography and applied sciences. It was officially discontinued in 1989.

<sup>†</sup> Questions were posed to pre-menopausal women only. Multiple choice was allowed.

### 3.2. Analysis of sex-, age-, and smoking habit- associated biological parameters

A total of 328 variables were obtained from the study questionnaire (see Case Report Form). The physical examination and clinical laboratory analyses were assembled into a data warehouse using LabKey [1]. To validate the data collected in our study, we first tested our ability to identify known biological correlates of sex, age or smoking-habits. To achieve this, we utilized a discovery-based approach. With the initial aim of reducing the complexity of the biochemical, hematologic and serologic data – thereby increasing the power of our association studies – we correlated all quantitative values from clinical laboratory data for the 1000 donors to each of the other variables using a Spearman's correlation matrix. Results were clustered and plotted using a dendrogram to represent the relationships between variables, with *height* (ordinate axis) being inversely related to the correlation coefficient (Fig. 2A). For pairs or groups of variables that showed high correlation ( $\text{height} < 0.3$ , equivalent to  $r_s > 0.67$ ), we selected one representative variable (indicated by red star). Next, we utilized the standard machine learning Random Forest (RF) approach [2], applied to the dataset in order to identify the variables that are most important to correctly classify donors based on sex (Fig. 2B) or age (Fig. 2C). Of note, bootstrap aggregation (also referred to as bagging) of data was selected due to its stability and accuracy in statistical classification and regression. This approach, which reduces variance and avoids overfitting, can be applied to a variety of binary data (e.g., male vs. female) and continuous variables (e.g., age). Using this method, we found that serum creatinine (CREAT) concentration, hematocrit (HCT) and height are the features that are most predictive of sex; and lower glomerular filtration rate (GFR), higher plasma low density lipoprotein (LDL) concentration and higher systolic blood pressure (SYSBP1) to be most associated with

age. These results were validated using univariate tests (Table S3) and representative box-plots are shown for the most significant variables (Figs. 2D, E).

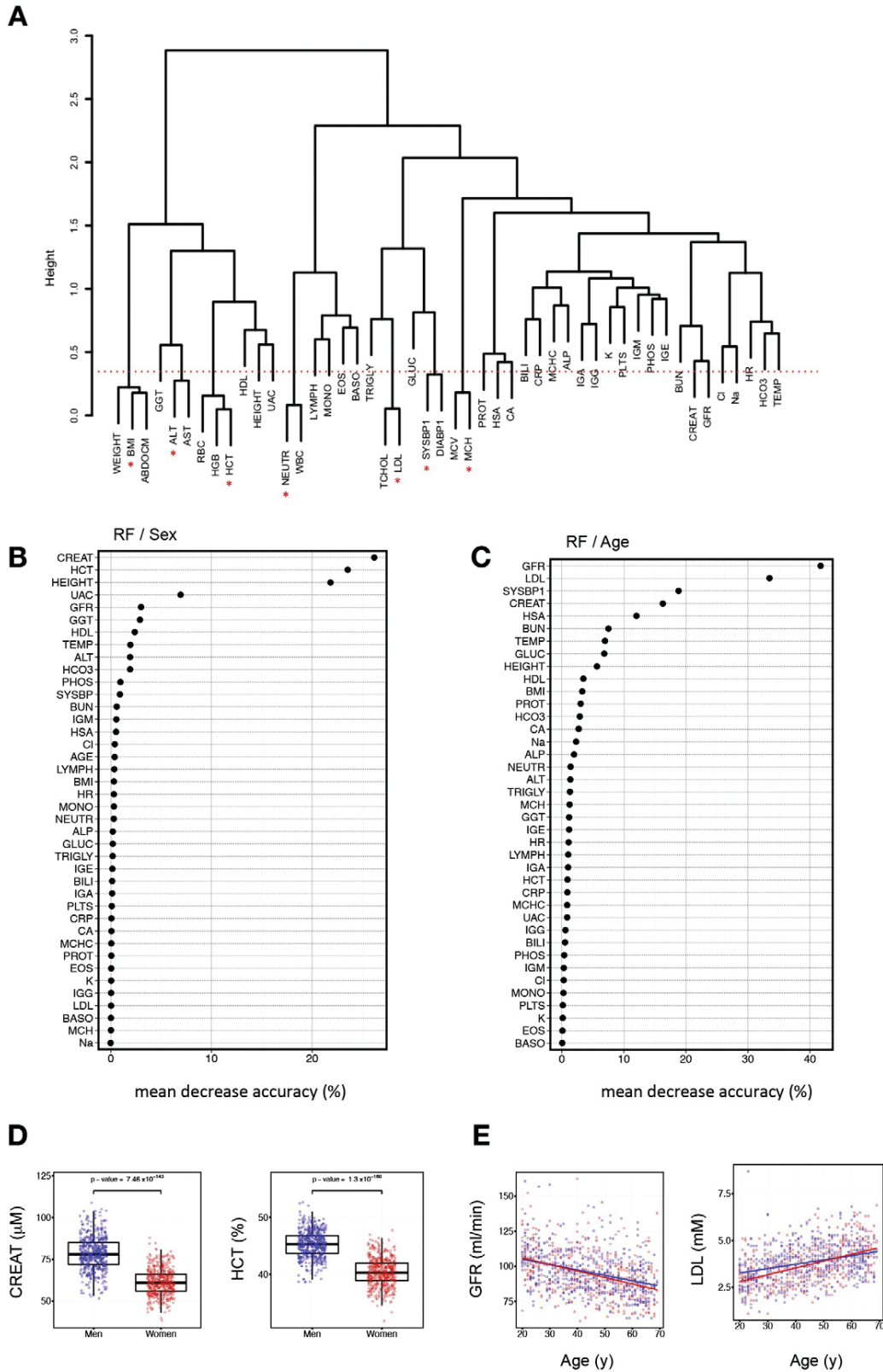
To assess other determinants, while controlling for sex- and age-associated effects on clinical laboratory data, we used a linear regression model, considering sex and age as independent covariates. This permitted us to examine the features predictive of smoking habits, again employing Random Forest analysis to segregate non-smokers, not exposed to second-hand smoke ( $n = 394$ ) from active smokers ( $n = 208$ ) present in our sample (Fig. 3A). Validating prior findings [3–6], we report that serum IgG and bilirubin concentrations were lower in smokers as compared to non-smokers (Figs. 3A, B); whereas monocyte, neutrophil and lymphocyte numbers were higher in smokers as compared to non-smokers (Figs. 3A, C). These observations may be related to lower antioxidant concentrations [7], and a diminished adherence of leucocytes to blood vessel walls [8]. Interestingly, a comparison of non-smokers and prior smokers present in our sample ( $n = 251$ ) indicates that smoking cessation restores the biochemical and immunological phenotypes associated with non-smokers (Figs. 3B, C). Together, these data highlight that our sample population can be used to study associations in the general French population and can be compared to prior study cohorts.

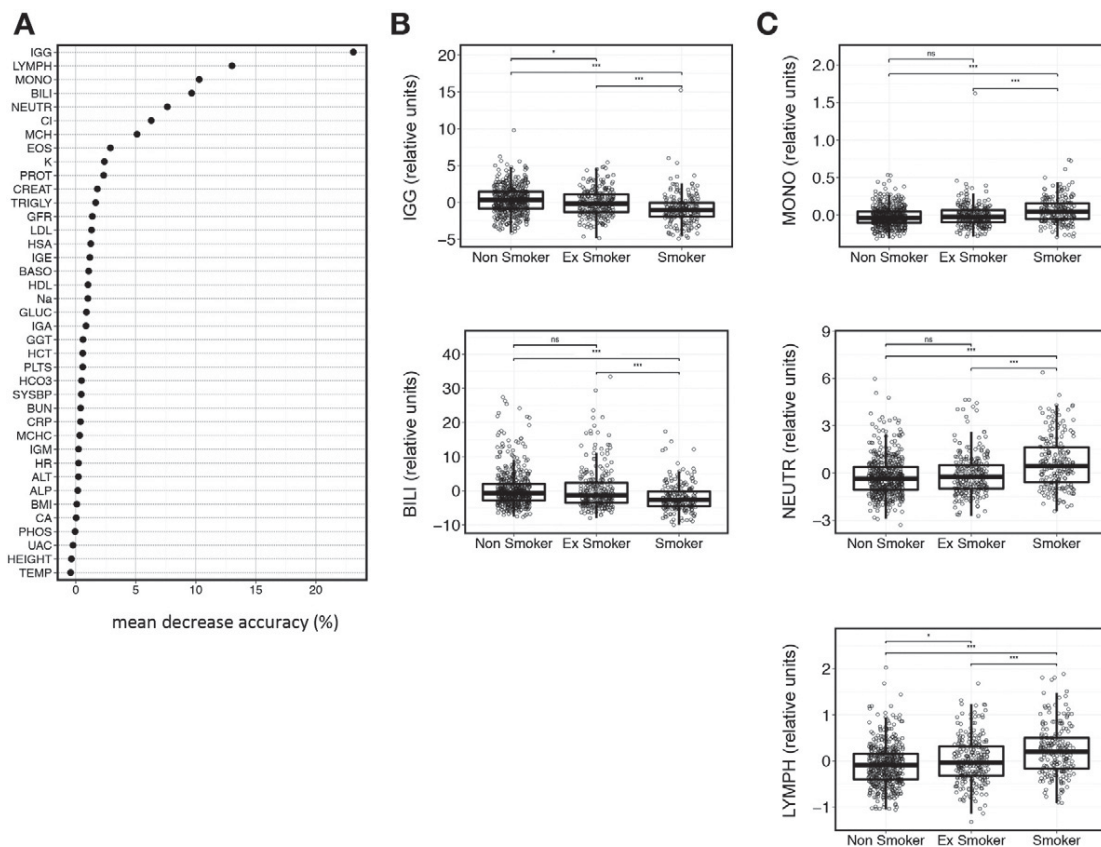
### 3.3. Smoking habits confer increased risk for metabolic syndrome among healthy donors

Over the past two decades, there has been increasing concern about the prevalence of obesity and its association with diabetes and cardiovascular disease (CVD), and their link to metabolic syndrome [9]. While several assessment scores have been established, the metabolic syndrome score is now a widely applied measure. Metabolic syndrome is most commonly defined by six variables: increased abdominal circumference

(AbdoCM > 94cm European men, > 80cm European women), elevated systolic blood pressure (SYSBP ≥ 130mmHg), elevated diastolic blood pressure (DYSBP ≥ 85mmHg), elevated triglyceride levels (TG ≥ 1.7mM), diminished levels of high

density lipoprotein (HDL < 1 mM men, < 1.3 mM women) and glucose concentration (≥ 6.1 mM) [10]. We thus analyzed donors for these six criteria, using accepted cut-values for European men and women, and for each criterion, data was



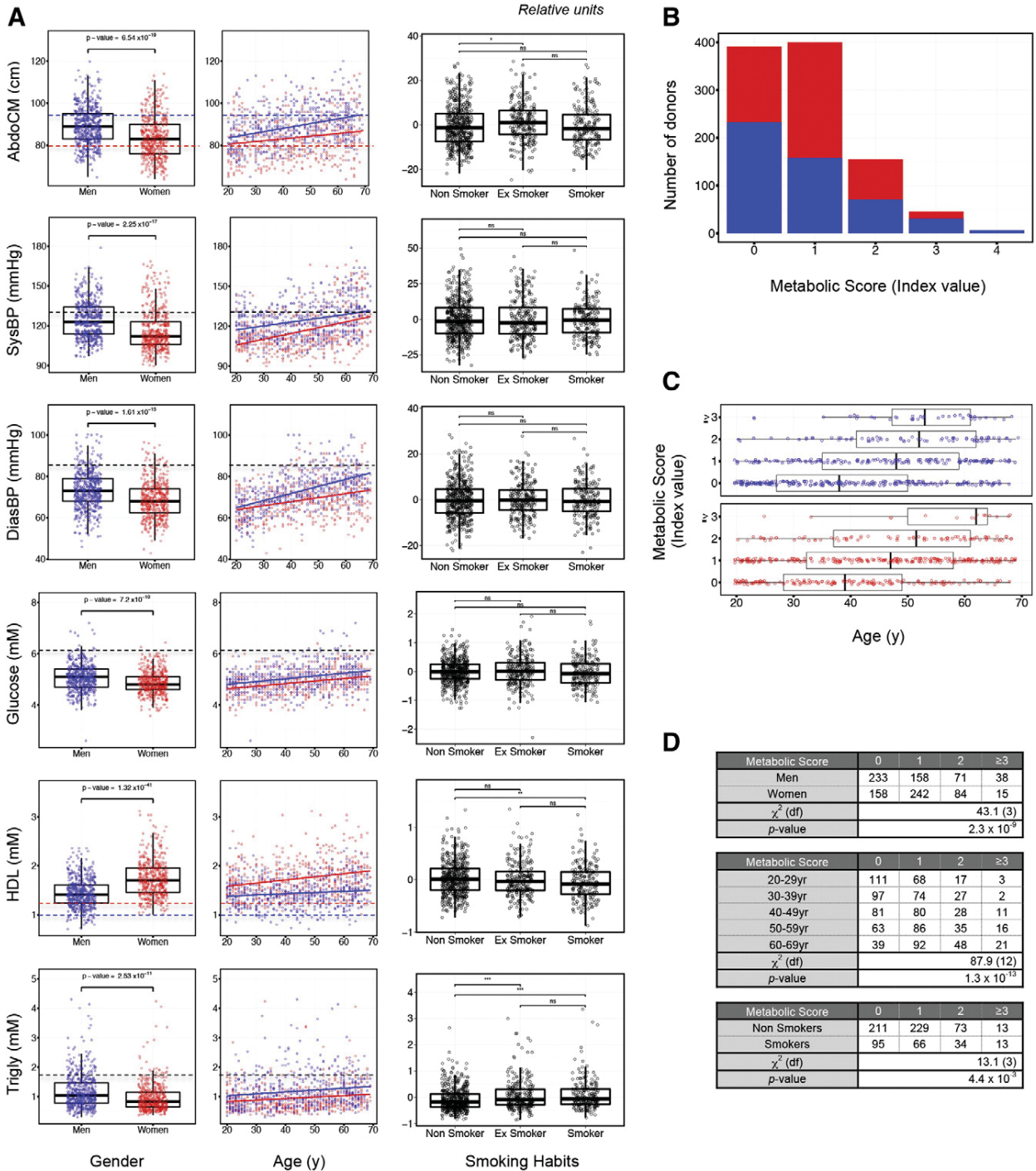


**Figure 3** Tobacco use associated with lower IgG, bilirubin concentrations and higher number of circulating monocytes, neutrophils and lymphocytes. (A) The same variables selected from Fig. 2A were regressed for sex and age, then subjected to Random Forest analysis to identify variables that discriminate active smokers ( $n = 208$ ) from non-smokers with no reported passive smoking ( $n = 394$ ). The random forest classification had an estimated error rate of 24.5% on the out-of-bag error. Variables are reported according to their impact on the out-of-bag error (percentage of mean decrease accuracy). (B, C) The top variables found to be lower (B) or higher (C) in smokers as compared to non-smokers are shown, with the inclusion of ex-smokers with no reported passive smoking ( $n = 201$ ) as an additional group. Variables that measured importance through permutation are depicted: serum IgG concentration (IGG, g/L); bilirubin concentration (BILI,  $\mu\text{M}$ ); the absolute number of monocytes (MONO,  $\times 10^3/\mu\text{L}$ ), neutrophils (NEUTR,  $\times 10^3/\mu\text{L}$ ) and lymphocytes (LYMPH,  $\times 10^3/\mu\text{L}$ ). Individual donors are represented by an open black circle. The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, the lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR). A student  $t$ -test was used to determine statistical differences between two groups for the given variables ( $p$ -value indicated and bracket defining the two groups being compared, \*,  $p < 0.05$ ; \*\*\*,  $p < 0.001$ ; ns, not significant).

**Figure 2** Unbiased assessment of the clinical laboratory data revealed expected sex-, and age-associations. (A) Biological measurements from the electronic case report forms (eCRF) and clinical laboratory data were evaluated using Spearman's correlation matrix and plotted using a dendrogram. For subsequent data mining, representative variables were selected (red star) from pairs or groups of factors showing high correlation (height  $< 0.3$ , indicated by dotted red line). (B, C) Random Forest method was employed to identify variables that discriminate men and women (B) or age, used as a continuous variable (C). The random forest classification on sex indicated an estimated error rate of 2.4% based on the out-of-bag error, while the random forest regression on age had a mean of squared residuals of 95.1, with 54% variance explained. Variables are reported according to their impact on the out-of-bag error (percentage of mean decrease accuracy). (D, E) The top two variables for sex (D) or age (E) that measured importance through permutation are depicted: creatinine concentration (Creat,  $\mu\text{M}$ ); hematocrit (HCT, %); glomerular filtration rate (GFR, mL/min); and low density lipoprotein concentrations (LDL, mM). Individual donors are represented by an open circle (blue, men; red, women). The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR); and a student  $t$ -test was used to determine statistical differences between men and women for the given variables ( $p$ -value indicated at the top of each graph) (D). Regression lines indicate the respective curve for men and women and results of univariate statistical analyses can be found in Table S2 (E).

reported in relation to sex and age (Fig. 4A). Notable differences in the diastolic, systolic blood pressure, triglyceride and glucose levels were observed, with men having significantly higher levels than women (n.b., comparisons were made for those criteria in which reference values were similar between men and women). In all instances, biologic measures showed a significant increase with advancing age. Quantitative laboratory data were again regressed out for sex and age effects, and the component variables were evaluated among smokers, ex-smokers and non-smokers (Fig. 4A).

To generate a composite metabolic score, 1 point was assigned for each of the assessed variables, taking blood pressure elevation as a single value (i.e., elevated SYSBP and/or DIASBP = 1 point) [11]. The index value for the metabolic score indicated that 400 individuals (40%) had at least one positive criterion, 155 donors (15.5%) had a score of 2, and 53 donors (5.3%) had a score of  $\geq 3$ , despite meeting all criteria for being a healthy donor (Fig. 4B). Notably, women had a higher probability of scoring  $\geq 1$  due to the low threshold for abdominal circumference for





European women (299 women vs. 135 men being above the respective abdominal circumference cut value) (Table S6). Indeed all variables that constitute the metabolic score, with the exception of HDL, showed a sex bias (Fig. S2). As a result, we observed a significant association between sex and the metabolic index ( $\chi^2 = 43.1$ , degrees of freedom (df) = 3,  $p = 2.3 \times 10^{-9}$ ; Fig. 4B); and there was a significant increase in median age when donors were stratified based on metabolic index ( $\chi^2 = 87.9$ , df = 12,  $p = 2.3 \times 10^{-13}$ , Fig. 4C). While smoking habits did not impact each of the individual variables, there was a significant relative risk increase associated with smoking as compared to non-smokers, after regressing out sex and age (Fig. 4D,  $\chi^2 = 13.1$ , df = 3,  $p < 0.005$ ; Table S6). Thus, smoking habits are an independent risk factor for the metabolic syndrome, distinct from its known association with CVD.

### 3.4. Sex, age and relationship status are risk factors for altered immunological status

Common infections and abnormal levels of immunoglobulins are conditions that may alter the immunological state of individuals. In our study population, circulating levels of immunoglobulins (i.e., IgM, IgG, IgE and IgA) were quantified, as well as influenza- and cytomegalovirus (CMV)- specific IgG antibodies (Table S3). We investigated association of demographic, socio-economic variables and/or lifestyle habits with these serological parameters (Table 2). Regression analyses were used to identify independent predictors among the 73 available variables.

Using univariate regression analysis, we found that positive detection of anti-influenza virus IgG was significantly associated with higher stature ( $p = 5.7 \times 10^{-4}$ ), sex (incidence of 86.1% and 77.8% in men and women;  $p = 1.0 \times 10^{-3}$ ), a higher weight ( $p = 2.1 \times 10^{-3}$ ) and a younger age ( $p = 2.3 \times 10^{-2}$ ). However, only sex and age remained significantly associated with influenza specific IgG when multiple regression analyses were performed, including sex, age, height and weight predictors ( $p = 1.1 \times 10^{-3}$ ,  $2.4 \times 10^{-2}$ , 0.85 and 0.17, respectively). Similarly, when stratifying by sex, parameters such as height and weight were no longer associated with

anti-influenza virus IgG ( $p > 0.05$ ). More interestingly, a younger age was found to be associated with infection only in women ( $p_{\text{women}} = 7.8 \times 10^{-3}$  vs.  $p_{\text{men}} = 0.91$ ). Indeed, the sex ratio in IgG-samples was 0.90 in individuals between 20 and 39 years, while it dropped to 0.53 in people between 40 and 69 years. Together, and in accordance with previous findings [12], our analyses support the notion that men are at higher risk of being positive for anti-influenza IgG, and suggests a female-specific influence of age on influenza infection.

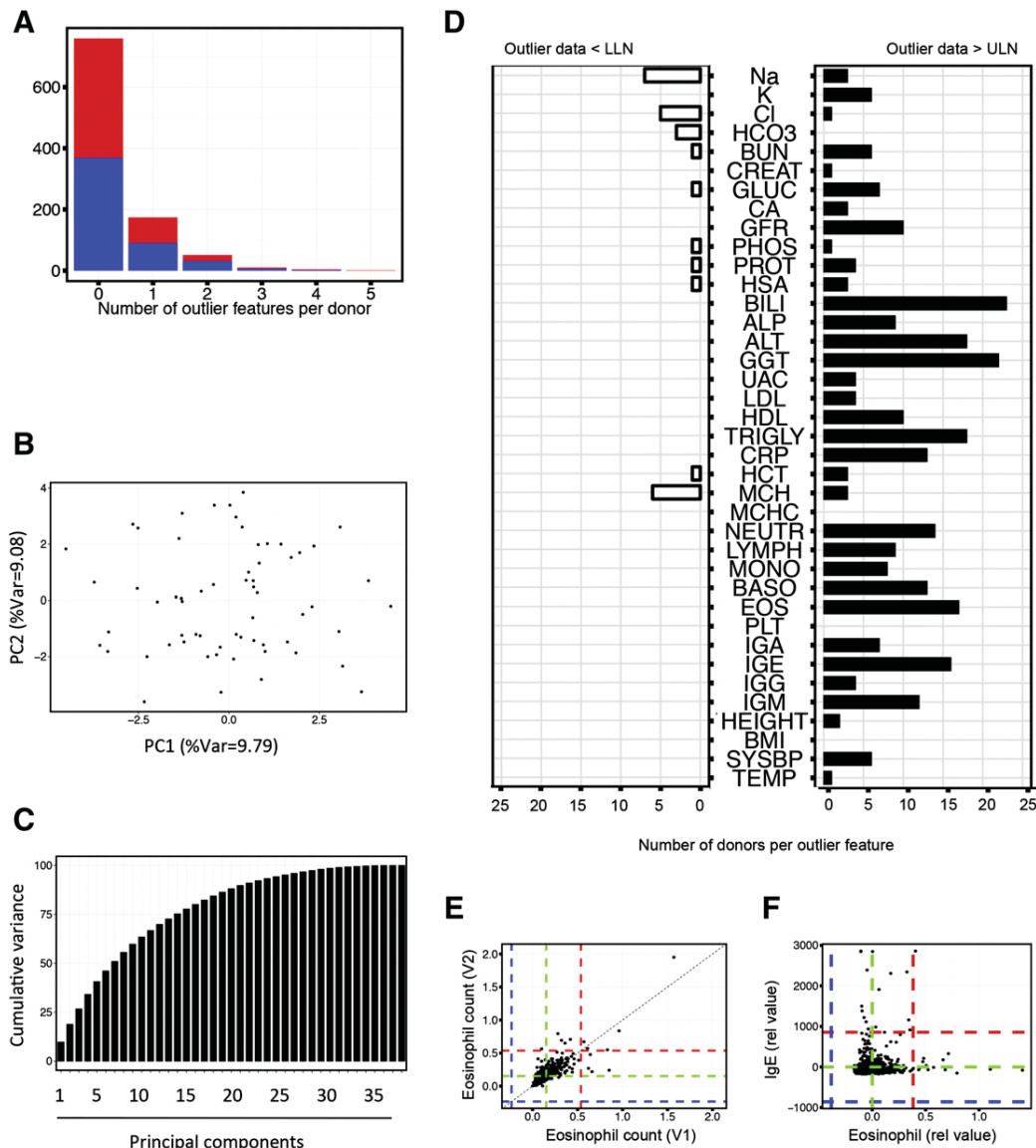
Conversely, factors associated with positive detection of anti-CMV IgG were an older age ( $p = 7.6 \times 10^{-7}$ ), the consumption of raw fruits and vegetables ( $p = 4.0 \times 10^{-4}$  and  $p = 1.6 \times 10^{-4}$ , respectively), being female ( $p = 1.7 \times 10^{-3}$ ), a shorter sleep duration ( $p = 1.5 \times 10^{-3}$ ), single status ( $p = 5.8 \times 10^{-3}$ ) and a lower stature ( $p = 1.0 \times 10^{-2}$ ), by univariate regression analysis. All these factors remained significantly associated in a multiple regression analysis, with the exception of height and the consumption of raw fruits (Table S7). While age was consistently associated with CMV infection in both men and women ( $p_{\text{men}} = 2.1 \times 10^{-2}$  and  $p_{\text{women}} = 3.8 \times 10^{-3}$ ), consumption of raw vegetables ( $p_{\text{men}} = 0.59$  and  $p_{\text{women}} = 8.2 \times 10^{-3}$ ), relationship status ( $p_{\text{men}} = 0.48$  and  $p_{\text{women}} = 4.6 \times 10^{-3}$ ) and hours of sleep ( $p_{\text{men}} = 0.13$  and  $p_{\text{women}} = 7.7 \times 10^{-2}$ ) were significant (or trended towards significance) in women only. By contrast, the association of CMV infection with being single in men was restored when restricting the analysis to men who have children ( $p_{\text{men}} = 1.9 \times 10^{-2}$ ; Fig. S2).

Next, the different classes of immunoglobulins were evaluated for their association with available demographic data, using multiple regression of the most significant univariate predictors. Elevated IgG levels were associated with smoking ( $p = 1.2 \times 10^{-13}$ ; Fig. 3), influenza virus infection ( $p = 5.2 \times 10^{-4}$ ), multivitamins consumption ( $p = 8.2 \times 10^{-3}$ ) and being a woman ( $p = 3.1 \times 10^{-2}$ ). The three former factors remained significant (or trended towards significance) in males and females, when considered separately. Elevated IgM levels were associated with being a woman ( $p = 5.0 \times 10^{-4}$ ) and with lower BMI ( $p = 1.8 \times 10^{-2}$ ). Elevated IgE levels were associated with being a man ( $p = 1.4 \times 10^{-6}$ ), younger age ( $p = 1.4 \times 10^{-3}$ ), exposure to silica

**Figure 4** Among healthy donors, being a male, increasing age and tobacco use are independent risk factors for higher metabolic score index value. (A) Variables that are used for determining an individual's metabolic score are plotted individually, representing differences between men and women; across age, as a continuous variable; or among non-smokers, ex-smokers and smokers. The variables included abdominal circumference (AbdCM, cm), systolic blood pressure (SysBP, mm Hg), diastolic blood pressure (DysBP, mm Hg), glucose levels (mM), high density lipoprotein concentration (HDL, mM), and triglyceride concentrations (Trigly, mM). For depiction of sex and age associations, individual donors are represented by an open circle (blue, men; red, women). For smoking habit associations, data was regressed for sex and age, data is plotted as relative units for respective plots, and individual donors are represented by an open black circle. Dotted black lines indicate reference values for European population; and where relevant sex-dependent reference indicators are used (blue dotted line, men; red dotted line, women). (B) The metabolic score was calculated for each donor and plotted to represent number of donors having indicated index values. Bar graphs indicate men (blue) and women (red). (C) Age association with metabolic score index values is shown, indicating men (blue circles) in the top plot, and women (red circles) in the bottom plot. (D) Contingency tables are shown for indicated comparisons and results from  $\chi^2$  testing are reported. The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, the lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR) (A, C). Where indicated a student t-test was used to determine statistical differences between two groups for the given variables ( $p$ -value indicated and bracket defining the two groups being compared, \*,  $p < 0.05$ ; \*\*,  $p < 0.01$ ; ns, not significant).

( $p = 1.0 \times 10^{-2}$ ), smoking ( $p = 4.7 \times 10^{-2}$ ), and a familial history of atopy ( $p = 5.0 \times 10^{-2}$ ). When stratifying these analyses by sex, exposure to silica was significant in men

only ( $p_{\text{men}} = 4.6 \times 10^{-3}$  and  $p_{\text{women}} = 0.48$ ). Finally, elevated IgA levels were associated with an older age ( $p = 5.3 \times 10^{-5}$ ), being a man ( $p = 5.5 \times 10^{-3}$ ) and non-smoking ( $p = 2.1 \times 10^{-2}$ ).



**Fig. 5** Outlier data maps primarily to liver function tests and complete blood count measurements, yet shows underlying structure based on measured variables. (A) For each donor, and for each measurement, an outlier status was assessed based on a z-score criteria. An aggregated score was computed for each donor to represent the number of times a given donor had been flagged as an outlier. The distribution of aggregated outlier cases among our cohort is shown. Colors indicate sex (men, blue; women, red). (B, C) A Principal Component Analysis (PCA) on 62 donors presenting >2 outlier cases has been performed. The scatterplot shows the projection of the donors onto the plane composed of the first 2 principal components, capturing 18.8% of the variance (B). The cumulative variance from the principal components is represented (C). (D) For each measurement considered in our analysis, we represented the number of donors that had been flagged as an above-the-range (on the right), or below-the-range outlier (on the left). (E) The absolute number of eosinophils is represented for V1 and V2, shown as representative data for the measured variables. Values have been regressed-out to take into consideration age and sex effects. Individual donors are represented by a black dot. A dotted line depicts the theoretical ideal correlation between the 2 visits. Dotted blue and red lines show the lower and upper threshold, respectively, as defined by our z-score based outlier detection. The green line shows the mean measurement. (F) The relationship between the eosinophil count and IgE level is shown. Both values have been regressed-out to take into consideration age and sex effects. Individual donors are represented by a black dot. Blue and red dotted lines show the lower and upper threshold, respectively, as defined by our z-score based outlier detection. The green line shows the mean measurement.

### 3.5. Outlier phenotypes showed independence among the measured variables

Despite the stringent criteria used for the recruitment of healthy donors (Table S1), we observed donors presenting extreme values within the observed range of biological measures. We identified 241 donors (24.1%) with clinical laboratory values that were outliers with respect to at least one variable, as defined by z-score based criteria (Fig. 5A). In only 66 individuals (6.6%), we observed two or more outlier events (Fig. 5A). To assess possible structure among the outlier events, we analyzed the data from those 66 individuals with outlier values for one or more laboratory tests, and projected the data using principal component analysis (PCA) (Fig. 5B). The dataset showed a lack of structure, which could also be observed by the broad distribution of variance across the top 35 component axes (Fig. 5C).

To interrogate the variables for which donors had outlier events, we plotted the number of donors per feature (Fig. 5C). Interestingly, liver function tests (e.g., ALT, GGT, BILI) and circulating immune cell counts (e.g., EOS, NEUTR, BASO) were highly represented among the feature space. We also observed higher numbers of donors (>15) with outlier TRIGLY and IGE levels. Selected variables were re-tested in the 500 donors sampled at V2, allowing the evaluation of repeatability. As shown, 4 of 8 (50%) of the donors with elevated numbers of EOS during V1 also showed higher levels at V2 (Fig. 5E). These data reinforce the added value of repeat testing for spurious outlier clinical laboratory data, but may also indicate the impact of environmental determinants on transient biochemical or cell number elevations. Finally, we investigated a possible association between EOS and IgE concentrations, as both are associated with allergic phenotypes. In support of the conclusions of the PCA, EOS count and IgE concentrations showed no correlation among healthy donors.

Our findings collectively help to define and validate the constitution of a healthy reference population, which will serve as a foundation for understanding and quantify the extent to which phenotypic variation in immune responses is under genetic or environmental control.

## 4. Discussion

The immune system is responsible for maintaining a healthy state, preventing infection and maintaining homeostasis. For some individuals, however, immune dysfunction can occur and results in increased susceptibility to infections, inflammation, autoimmunity, allergy or even cancer. Moreover, such individual heterogeneity in the immune response may have a major impact on the likelihood to respond to therapy or the development of side effects secondary to vaccine administration. Most prior studies aiming to understand the extent to which variation in immune responses is associated with immunopathology *sensu lato* have taken a disease-based approach, from which considerable insight into immune mechanisms have been obtained. Nonetheless, to utilize this information in diagnosis and disease management, the definition of the baseline parameters for immune function across the human “healthy” population is required.

To achieve this goal, the *Milieu Intérieur* Project aims to provide a foundation for defining perturbations in an individual's immune system responses.

The *Milieu Intérieur* clinical study was designed and performed in healthy volunteers to develop a diverse sample collection with wide ranging associated meta-data. Ultimately, the generation of genetic data (based on genome-wide genotyping and whole exome sequencing) and multiple phenotypes (molecular, cellular and organismal) in available samples of the study cohort will produce a rich data warehouse. This will allow data mining studies for associations and consequently increase our knowledge of the different factors involved in the regulation of immune responses. During the design of the clinical study, we encountered the challenge of defining the genuine meaning of being “healthy” according to rational and measurable parameters. As such, strict criteria for enrolling donors were established, taking into consideration both recruitment feasibility, and the statistical power provided by a 1000-persons study, covering 10 strata (segregated across sex and age, by decade). While some exclusion factors were easy to apply, such as chronic infections (e.g., HCV) or severe disease (e.g., cancer, autoimmunity, etc.), others were more challenging, such as the boundary for allergic individuals, those that are exposed to known toxins (e.g., cigarettes), and persons with presyndromic signs (e.g., hypertension). Although the use of reference values for hematological, biochemical and serology parameters, commonly accepted in the clinic to define the healthiness of an individual, was considered as inclusion/exclusion criteria, there was a concern about the potential loss of extreme phenotypes. We thus chose cut off values that might indicate the requirement for medical follow-up (e.g., liver enzyme concentrations > 3 × ULN). Factors affecting the immune system and/or the composition of microbiota were also considered, including pre-term birth, current and prior exposure to medical treatments (e.g., aspirin), or the use of homeopathic medicaments (e.g., essential oils). Ultimately, we settled on the allowance of parameters expected to be present in >5% of the sampled individuals, and excluded any condition that necessitated past or current medical treatment. Detailed personal and family medical histories were systematically recorded, and associated meta-data will be used to define genetic, immunologic, and enterotype associations; and/or to regress out potential confounding factors. We hope that this set of criteria will help the international community taking steps towards a consensus definition of a healthy status for immunologic studies.

In considering selection biases linked to cross-sectional population-based studies [13], we consider several potential sources. The primary sources of selection bias are selective survival with fixed exposure in time (e.g., older donor survival effect); and non-fixed exposure in time (e.g., smoking, diet, alcoholic intake, professional exposure are variable in time). With respect to the survival effect bias, we in fact see this as an opportunity, as evidence for an age-associated narrowing of immunologic, genetic and enterotype variation may point towards a core signature of healthy status. To address non-fixed exposure, we highlight that our complete questionnaires provide an overview of both current and past exposures/habits. An additional caveat is that the healthy volunteers were selected from a pre-existing donor database, curated by a Clinical Research Organization. These volunteers

may be more “health-conscious” than *one-off* volunteers. This recruitment strategy may also explain the higher percentage of out-of-work persons, and the higher level of education as compared to the local *Ille-et-Vilaine* population.

Following our initial validation of known associations of health and clinical laboratory/immunological parameters, we investigated correlates with the metabolic syndrome index score. Among the general adult population, it is estimated that 20–25% meet the criteria for having metabolic syndrome (index score  $\geq 3$ ), so the identification of risk factors is central to establishing public health initiatives. While metabolic syndrome has been carefully evaluated in the context of disease settings, few studies have investigated healthy donors for risk factors. Interestingly, in a “healthy” setting, our study revealed that sex (i.e., being a man), aging and active cigarette smoking are each independent risk factors for an elevated metabolic score index. While epidemiological data support our findings for men and age as associated risk factors, the evidence for smoking as an independently associated variable (i.e., measured after regressing out sex and age) has been so far controversial. Our data indicate that four of the six individual component biologic variables are not statistically different for smokers as compared to non-smokers, however the global score supported its association with metabolic syndrome. Previous published studies, focusing primarily on individuals presenting overweight and obesity, showed an additive effect for smoking as an associated risk factor [14]. Conversely, other studies have failed to detect such associations and at least one study conducted among Turkish women found a protective effect of smoking on metabolic syndrome [15]. This has been attributed in part to the use of different definitions of metabolic syndrome. A recent meta-analysis evaluated data from 13 prospective studies for which primary data was available ( $n = 56,691$  participants overall), and in a dose-response analysis, active smoking habits was positively associated with risk of metabolic syndrome (pooled relative risk [RR] = 1.26, 95% CI: 1.1–1.44) [16]. Our results, which differ from previous studies in that they are based on healthy donors, provide additional support for their findings and are consistent with experimental data indicating that cigarette smoking modifies hormone levels (e.g., cortisol), which in turn may result in the establishment of a more “insulin-resistant” state or the increase in waist circumference, a result of deposition visceral fat mass [17].

Our investigations of the metabolic syndrome score in healthy donors also revealed a troubling sex-bias. With the sole exception of HDL levels, the reference values for women and men have differing cut-values. For abdominal circumference, the effects are dramatic with nearly 60% women being considered above the threshold value as compared to 27% for men. Given the wide application of the metabolic syndrome score since 2001 [18], we suggest that the threshold values be adjusted for sex-associated differences. Notably, this has been done in USA populations, however it must be considered whether the adjusted thresholds have been set based on a shared definition of health, or instead due to the epidemic of obesity that is currently raging in first-world countries. Indeed these indicators impact public health initiatives and treatment endpoints, and therefore must be properly calibrated and correlated to real endpoints of health and disease. It is

our hope that the *Milieu Intérieur* project will contribute to the identification of genetic, enterotype and immune response associations to metabolic score and other health indicators, possibly leading to the innovation of personalized algorithms.

Cytomegalovirus (CMV) infection is one of the most common infections of the general population [19], with a seroprevalence of 43% in Europe [20] and 50% in the US [21]. CMV is known to be transmitted through direct contact with infected bodily fluids, including urine, saliva, tears, but also blood and semen. A large number of studies have evaluated risk factors for CMV infection among pregnant women, but few have studied CMV incidence in a well-defined healthy donor population. We confirm that being single is an important risk factor for men and women, due to an increased number of partners [22], while having children has no direct impact. This challenges the notion that CMV transmission often occurs from children to adults and we suggest instead that adult-to-adult transmission is more common. Longitudinal studies will be required to confirm these observations, and to confirm that exposure to children varies with marital status. We identified another factor of interest: the consumption of raw vegetables. To our knowledge, this habit has not been previously described as a risk factor for CMV infection, and challenges the view that CMV is mainly transmitted by direct contact. Interestingly, recommendations to prevent infection usually include the avoidance of food sharing with young children [23]. While this might suggest that indirect contact is a risk factor, there had been little evidence to support this public health measure. One study has evaluated the duration of CMV viability on environmental surfaces and found that the virus could remain viable for 6 h on wet surfaces, including crackers [24]. Together, our observation supports that CMV transmission from food sharing has been underestimated and should be thoroughly evaluated in order to adapt preventive behaviors.

Interestingly elderly donors in our cohort showed a higher percentage (44%) of CMV negative individuals than previously reported studies (20–30% CMV- for >60 years old) [20]. This likely reflects the “healthy” status of our donors, as defined by stringent inclusion criteria, in contrast with previous studies that were performed on the general population. This is also reflected across the entire cohort, which was 64.8% CMV-, higher than other reported studies in both Europe (43%) [20] and the US (50%) [21]. Identifying host and environmental factors that may lead to increased resistance to CMV infection throughout life could have major implications for cardiovascular disease, sepsis and healthy aging.

To summarize, our study outlined herein provides an initial overview of the *Milieu Intérieur* cohort, which we believe constitutes a rich source of information and materials that will ultimately help to characterize and define topics relating to immunity, genetics, environment and lifestyle behaviors.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.clim.2014.12.004>.

## Conflict of interest statement

The authors declare that there are no conflicts of interest.

## Acknowledgments

This work benefited from the 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). We acknowledge the advice of the "healthy donor" working group that helped us to define the inclusion/exclusion criteria: Laurent Abel, François Aubin, Raphaëlle Bourdet-Sicard, Ingrid Callies, Miguel Fenoy, Bridget Holmes, Laurence Mathivet, Cédric Ménager, Stéphane Reynier, Manfred Schmoltz, Muriel Vray). Scientific advisory board members that supported the trial design discussions included: Hyam Levitsky, Philip Greenberg, John Marioni, Jonathan Braun, Adrian Hayday and Thomas Joos. We also thank Dusko Ehrlich and Joël Doré for advice on stool sample collection procedures. We would like to acknowledge all the Biotrial team of doctors, nurses, technicians and project managers for their expertise and professionalism. We also thank the donors for their contribution to the study.

## References

- [1] E.K. Nelson, B. Piehler, J. Eckels, A. Rauch, M. Bellew, P. Hussey, S. Ramsay, C. Nathe, K. Lum, K. Krouse, D. Stearns, B. Connolly, T. Skillman, M. Igra, LabKey Server: an open source platform for scientific data integration, analysis and collaboration, *BMC Bioinformatics* 12 (2011) 71.
- [2] L. Breiman, Random forests, *Mach. Learn.* 45 (2001) 5–32.
- [3] A. Gulsvik, M.K. Fagerhoi, Smoking and immunoglobulin levels, *Lancet* 1 (1979) 449.
- [4] S.A. McMillan, J.P. Douglas, G.P. Archbold, E.E. McCrum, A.E. Evans, Effect of low to moderate levels of smoking and alcohol consumption on serum immunoglobulin concentrations, *J. Clin. Pathol.* 50 (1997) 819–822.
- [5] H.A. Schwertner, Association of smoking and low serum bilirubin antioxidant concentrations, *Atherosclerosis* 136 (1998) 383–387.
- [6] M.R. Smith, A.L. Kinmonth, R.N. Luben, S. Bingham, N.E. Day, N.J. Wareham, A. Welch, K.T. Khaw, Smoking status and differential white cell count in men and women in the EPIC-Norfolk population, *Atherosclerosis* 169 (2003) 331–337.
- [7] M. Tsuchiya, A. Asada, E. Kasahara, E.F. Sato, M. Shindo, M. Inoue, Smoking a single cigarette rapidly reduces combined concentrations of nitrate and nitrite and concentrations of antioxidants in plasma, *Circulation* 105 (2002) 1155–1157.
- [8] M.M. Rahman, I. Laher, Structural and functional alteration of blood vessels caused by cigarette smoking: an overview of molecular mechanisms, *Curr. Vasc. Pharmacol.* 5 (2007) 276–292.
- [9] R.S. Padwal, A.M. Sharma, Prevention of cardiovascular disease: obesity, diabetes and the metabolic syndrome, *Can. J. Cardiol.* 26 (Suppl. C) (2010) 18C–20C.
- [10] K.G. Alberti, R.H. Eckel, S.M. Grundy, P.Z. Zimmet, J.I. Cleeman, K.A. Donato, J.C. Fruchart, W.P. James, C.M. Loria, S.C. Smith Jr., Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity, *Circulation* 120 (2009) 1640–1645.
- [11] P.W. Wilson, R.B. D'Agostino, H. Parise, L. Sullivan, J.B. Meigs, Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus, *Circulation* 112 (2005) 3066–3072.
- [12] S.L. Klein, A. Hodgson, D.P. Robinson, Mechanisms of sex disparities in influenza pathogenesis, *J. Leukoc. Biol.* 92 (2012) 67–73.
- [13] M. Delgado-Rodriguez, J. Llorca, Bias, *J. Epidemiol. Community Health* 58 (2004) 635–641.
- [14] H. Cena, M.L. Fonte, G. Turconi, Relationship between smoking and metabolic syndrome, *Nutr. Rev.* 69 (2011) 745–753.
- [15] A. Onat, H. Ozhan, A.M. Esen, S. Albayrak, A. Karabulut, G. Can, G. Hergenc, Prospective epidemiologic evidence of a "protective" effect of smoking on metabolic syndrome and diabetes among Turkish women—without associated overall health benefit, *Atherosclerosis* 193 (2007) 380–388.
- [16] K. Sun, J. Liu, G. Ning, Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies, *PLoS One* 7 (2012) e47791.
- [17] K. Linder, F. Springer, J. Machann, F. Schick, A. Fritsche, H.U. Haring, G. Blumenstock, M.B. Ranke, N. Stefan, G. Binder, S. Ehehalt, Relationships of body composition and liver fat content with insulin resistance in obesity-matched adolescents and adults, *Obesity (Silver Spring)* 22 (2014) 1325–1331.
- [18] S.M. Grundy, H.B. Brewer Jr., J.I. Cleeman, S.C. Smith Jr., C. Lenfant, Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition, *Circulation* 109 (2004) 433–438.
- [19] J. Bodurtha, S.P. Adler, W.E. Nance, Seroepidemiology of cytomegalovirus and herpes simplex virus in twins and their families, *Am. J. Epidemiol.* 128 (1988) 268–276.
- [20] P.R. Lubeck, H.W. Doerr, H.F. Rabenau, Epidemiology of human cytomegalovirus (HCMV) in an urban region of Germany: what has changed? *Med. Microbiol. Immunol.* 199 (2010) 53–60.
- [21] S.L. Bate, S.C. Dollard, M.J. Cannon, Cytomegalovirus seroprevalence in the United States: the national health and nutrition examination surveys, 1988–2004, *Clin. Infect. Dis.* 50 (2010) 1439–1447.
- [22] K.B. Fowler, R.F. Pass, Sexually transmitted diseases in mothers of neonates with congenital cytomegalovirus infection, *J. Infect. Dis.* 164 (1991) 259–264.
- [23] G.J. Demmler-Harrison, Congenital cytomegalovirus: public health action towards awareness, prevention, and treatment, *J. Clin. Virol.* 46 (Suppl. 4) (2009) S1–S5.
- [24] J.D. Stowell, D. Fortin-Passoni, E. Din, K. Radford, D. Brown, A. White, S.L. Bate, S.C. Dollard, S.R. Bialek, M.J. Cannon, D.S. Schmid, Cytomegalovirus survival on common environmental surfaces: opportunities for viral transmission, *J. Infect. Dis.* 205 (2012) 211–214.
- [25] E. Kesse-Guyot, V. Andreeva, K. Castetbon, M. Vernay, M. Touvier, C. Mejean, C. Julia, P. Galan, S. Hercberg, Participant profiles according to recruitment source in a large Web-based prospective study: experience from the Nutrinet-Sante study, *J. Med. Internet Res.* 15 (2013) e205.
- [26] D. Duffy, V. Rouilly, V. Libri, M. Hasan, B. Beitz, M. David, A. Urrutia, A. Bisiaux, S.T. Labrie, A. Dubois, I.G. Boneca, C. Delval, S. Thomas, L. Rogge, M. Schmolz, L. Quintana-Murci, M.L. Albert, Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli, *Immunity* 40 (2014) 436–450.
- [27] RC Team, R: A Language and Environment for Statistical Computing, in: R Foundation for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [28] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York, 2009.

# Functional Analysis via Standardized Whole-Blood Stimulation Systems Defines the Boundaries of a Healthy Immune Response to Complex Stimuli

Darragh Duffy,<sup>1,2,3,14</sup> Vincent Rouilly,<sup>1,4,14</sup> Valentina Libri,<sup>1,14</sup> Milena Hasan,<sup>1</sup> Benoit Beitz,<sup>1</sup> Mikael David,<sup>1</sup> Alejandra Urrutia,<sup>1,2,3</sup> Aurélie Bisiaux,<sup>2,3</sup> Samuel T. LaBrie,<sup>5</sup> Annick Dubois,<sup>6</sup> Ivo G. Boneca,<sup>7,8</sup> Cécile Delval,<sup>6</sup> Stéphanie Thomas,<sup>1,2,3</sup> Lars Rogge,<sup>1,9</sup> Manfred Schmolz,<sup>10</sup> Lluís Quintana-Murci,<sup>11,12,15,\*</sup> and Matthew L. Albert<sup>1,2,3,13,15,\*</sup>  
for The *Milieu Intérieur* Consortium

<sup>1</sup>Center for Human Immunology, Institut Pasteur, 75015 Paris, France

<sup>2</sup>INSERM U818, 75015 Paris, France

<sup>3</sup>Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, 75015 Paris, France

<sup>4</sup>Center for Bioinformatics, Institut Pasteur, 75015 Paris, France

<sup>5</sup>Myriad Rules Based Medicine, Inc., Austin, TX 78759, USA

<sup>6</sup>Center for the Integration of Clinical Research, Institut Pasteur, 75015 Paris, France

<sup>7</sup>Laboratory of Biology & Genetics of the Bacterial Cell Wall, Department of Microbiology, Institut Pasteur, 75015 Paris, France

<sup>8</sup>INSERM, Equipe Avenir, 75015 Paris, France

<sup>9</sup>Laboratory of Immunoregulation, Department of Immunology, Institut Pasteur, 75015 Paris, France

<sup>10</sup>Myriad Rules Based Medicine, Inc., 72770 Reutlingen, Germany

<sup>11</sup>Laboratory of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, 75015 Paris, France

<sup>12</sup>CNRS URA3012, 75015 Paris, France

<sup>13</sup>INSERM UMS20, 75015 Paris, France

<sup>14</sup>These authors contributed equally to this work

<sup>15</sup>These authors contributed equally to this work

\*Correspondence: [quintana@pasteur.fr](mailto:quintana@pasteur.fr) (L.Q.-M.), [albertm@pasteur.fr](mailto:albertm@pasteur.fr) (M.L.A.)

<http://dx.doi.org/10.1016/j.immuni.2014.03.002>

## SUMMARY

Standardization of immunophenotyping procedures has become a high priority. We have developed a suite of whole-blood, syringe-based assay systems that can be used to reproducibly assess induced innate or adaptive immune responses. By eliminating preanalytical errors associated with immune monitoring, we have defined the protein signatures induced by (1) medically relevant bacteria, fungi, and viruses; (2) agonists specific for defined host sensors; (3) clinically employed cytokines; and (4) activators of T cell immunity. Our results provide an initial assessment of healthy donor reference values for induced cytokines and chemokines and we report the failure to release interleukin-1 $\alpha$  as a common immunological phenotype. The observed naturally occurring variation of the immune response may help to explain differential susceptibility to disease or response to therapeutic intervention. The implementation of a general solution for assessment of functional immune responses will help support harmonization of clinical studies and data sharing.

## INTRODUCTION

The immune system is responsible for maintaining a healthy state, ensuring beneficial cohabitation with microbiota, and pre-

venting infection. Immune system dysfunction is often associated with increased susceptibility to infection, inflammation, autoimmunity, or even development of cancer. Moreover, individual heterogeneity in the immune response can have important medical consequences, such as the likelihood to respond to anti-infectious therapy, the efficiency of vaccine administration, or the development of side effects secondary to treatment. Because of the complexity of immune responses at both the individual and population level, it has not been possible, thus far, to define the boundaries of a “healthy immune response” or its naturally occurring variability. Most studies have taken a disease-based approach, from which considerable insight into immune mechanisms has been obtained. Nonetheless, to utilize this information in diagnosis and disease management, the assessment of a healthy functional immune response within the human population is required. Specifically, there is an unmet need for reliable and reproducible assay systems for studying human immune responsiveness. In other words, we must overcome technical challenges and preanalytical error in order to assess the true variability in functional immune responses. Only then will immunologists be positioned to contribute to the promises of personalized medicine, applying simple-to-use technologies that provide in-depth understanding of the phenotypic variance of immune responses in the human population.

Human innate or adaptive immune responsiveness is typically studied *in vitro*, thereby permitting the evaluation of multiple stimulation conditions in parallel. Standard laboratory practice is to transport collected blood to a centralized facility, thereby allowing isolation of peripheral blood mononuclear cells (PBMCs) by Ficoll-Hypaque gradient centrifugation by trained

personnel (Folds and Schmitz, 2003). Stimulation can be performed immediately, but often cells are cryopreserved in order to batch test samples (Maecker et al., 2012). In addition to this process being labor intensive, there is a risk of sample contamination by microbial components (e.g., bacterial endotoxin). Moreover, sample handling results in variable loss of cells and cryopreservation diminishes functional responsiveness, also in a nonlinear and/or nonreproducible way (Chen et al., 2010). Although whole-blood human lymphocyte assays were first innovated in 1975 (Eskola et al., 1975), they have not been widely used in scientific research or clinical evaluation of functional immune responses. Notably, direct measurements made in whole blood have the advantages of minimizing contamination and sample handling. Moreover, maintaining total leukocytes (e.g., polymorphonuclear cells) and platelets in a plasma matrix may provide a more accurate reflection of natural responsiveness to immune stimuli (Chen et al., 2010; De Groot et al., 1992; Ida et al., 2006; Kirchner et al., 1982; Schmolz et al., 2004).

Herein, we report the development of 27 whole-blood stimulation systems, built into syringe-based medical devices that may be utilized in point-of-care approaches and tested in 25 ethnically well-defined individuals of European ancestry. With these stimulation conditions, we define the boundaries of a healthy immune response to complex stimuli (i.e., whole microbes), including gram-negative bacteria, gram-positive bacteria, mycobacteria, fungi, and live viruses (Krishna and Miller, 2012; Miettinen et al., 2008; Stuyt et al., 2003; Stuyt et al., 2001; Zhao et al., 2007). In addition, we developed assay systems to study the response to purified or synthetic ligands for the major innate host response pathways, including those triggered by the Toll-like receptors (TLRs) (Alexopoulou et al., 2001; Diebold et al., 2004; Gantner et al., 2003; Godaly and Young, 2005; Hayashi et al., 2001; Hemmi et al., 2000, 2002; Jurk et al., 2002; Liu-Bryan et al., 2005; Takeuchi et al., 1999; Zhao et al., 2007), nucleotide-binding domain and leucine-rich repeat containing molecules (NLRs) (Allen et al., 2009; Ichinohe et al., 2009), and C-type lectin-like receptors (CLRs) (Brown et al., 2003). To directly evaluate the variable responses after cytokine receptor signaling, we also tested several clinically employed cytokines such as interferon-alpha (IFN- $\alpha$ ), interferon-beta (IFN- $\beta$ ), interferon-gamma (IFN- $\gamma$ ), tumor necrosis factor-alpha (TNF- $\alpha$ ), interleukin 1-beta (IL-1 $\beta$ ), and interleukin 23 (IL-23) (Dinarelo, 2012; González-Navajas et al., 2012; Oppmann et al., 2000; Plataniias, 2005; Zheng et al., 2013). Finally, we utilized direct T cell receptor cross-linking (anti-CD3+anti-CD28) and super-antigen stimulation as two distinct mechanisms for eliciting T cell activation (Smith-Garvin et al., 2009). By quantifying the stimulus-induced production of cytokines, chemokines, and growth factors, it was possible to establish specific protein signatures for each stimulation system and we establish reference values as well as an estimate of variation among healthy individuals originating from a homogeneous ethnic background. These tools and the data set provided will be a valuable resource for the immunologic community. Moreover, we propose that through coordinated use of validated assay systems and open sharing of data sets, it will be possible to rapidly implement measures of functional immune responsiveness into clinical studies and medical practice.

## RESULTS

### Reproducible Whole-Blood Assays for Assessing Innate and Adaptive Immune Responses

To establish *in vitro* assay systems that preserve physiological cellular interactions, we developed syringe-based medical devices that can be used for activating immune cells present in whole blood. Fifty-four stimuli were considered for study and evaluated for sterility, solubility, dose response, short and long-term stability, and reproducibility (exclusion criteria are detailed in [Supplemental Experimental Procedures](#) available online). Based on initial testing, we prioritized 27 stimuli for development in TruCulture whole-blood collection and culture devices (Myraid RBM) (Table 1). In brief, during the manufacturing process (certified according to ISO 13845) of the TruCulture collection syringes, the indicated stimuli were dissolved in the proprietary TruCulture cell medium (2 ml per tube). These TruCulture systems were then frozen and stored at  $-20^{\circ}\text{C}$  until use. After thawing to room temperature, the collection syringes were filled with 1 ml whole blood and incubated for 22 hr ( $\pm 10$  min) in room air at  $37^{\circ}\text{C}$  ( $\pm 1^{\circ}\text{C}$ ), utilizing a bench-top heating block (VLMH GmbH). After immune stimulation, insertion of a valve separator (an integral part of the TruCulture system) yielded a culture supernatant that was aliquoted and stored at  $-80^{\circ}\text{C}$  for subsequent multiplex protein immunoassay testing (Figure S1A).

For all stimuli, we selected low and highly induced protein analytes that could be measured and used for dose-finding studies. We selected dose concentrations for the stimuli that maximized the ability to detect low-expressed proteins, while taking precautions not to exceed the upper limit of the biologic range for highly expressed proteins. Representative data for one microbe, MAMP, and T cell stimulus is shown (Figure S1B), and the selected dose for all assays can be found in Table 1. We further validated our assay systems by serially testing individual responsiveness to immune stimulation, repeating the measurements four times at the same time point (Figure S1C) and four times over a 25-day time period (Figure S1D). As represented by the data of lipopolysaccharide (LPS)-induced responses, 25 of the 27 stimuli induced protein signatures with intraindividual coefficients of variance (CVs) ranging from 5 to 14 (Figure 1D, Table S1, and data not shown). The two exceptions were calcium pyrophosphate dihydrate crystals (CPPD) and whole glutan particles (WGP), both of which are particulate agonists that were difficult to homogenize in liquid suspension and resulted in higher technical variation (Table S1). Regarding the response to LPS (Figure S1D), substantial variability could be observed among the three donors tested. Donor G showed high levels of IL-6 but intermediate induction of IFN- $\gamma$ . By contrast, donor H showed the highest production of IFN- $\gamma$  but the lowest induction of IL-6. Additional quality-control data, including selection of anticoagulant used, can be found in the [Supplemental Experimental Procedures](#) and Table S2.

### Quantitative and Qualitative Differences in Healthy Donor Responses to Immune Stimulation

To demonstrate the utility of our whole-blood stimulation systems, we recruited 25 healthy volunteers of European ancestry, aged 30–39 and stratified by gender (13 women, 12 men).

**Table 1. Innate and Adaptive Immune Stimuli Used for Development of Whole-Blood Stimulation Systems**

Stimulus	Abbreviation	Concentration	Supplier	Sensor or Receptor	Reference
Null	∅		NA		
<b>Microbe</b>					
HK <i>E. coli</i> 0111:B4	HKEC	10 <sup>7</sup> bacteria	Invivogen	complex	Takeuchi et al., 1999
HK <i>S. aureus</i>	HKSA	10 <sup>7</sup> bacteria	Invivogen	complex	Krishna and Miller, 2012
HK <i>L. rhamnosus</i>	HKLR	10 <sup>7</sup> bacteria	Invivogen	complex	Miettinen et al., 2008
BCG (Immucyst)	BCG	3 × 10 <sup>5</sup> bacteria	Sanofi Pasteur	complex	Means et al., 1999; Godaly and Young, 2005; Randhawa et al., 2011
HK <i>H. pylori</i>	HKHP	10 <sup>7</sup> bacteria	Invivogen	complex	Zhao et al., 2007
HK <i>C. albicans</i>	HKCA	10 <sup>7</sup> bacteria	Invivogen	complex	Brown et al., 2003; Gantner et al., 2003
Influenza A virus (live)	IAV	100 HAU	Charles Rivers	complex	Diebold et al., 2004; Kato et al., 2006; Ichinohe et al., 2009; Allen et al., 2009
Sendai virus (live)	SeV	10 HAU	Charles Rivers	Rig-I and Mda/5	Yoneyama et al., 2005; Kato et al., 2005
<b>MAMP</b>					
C12-IE-DAP	DAP	4 µg/ml	Invivogen	NOD1	Chamaillard et al., 2003
CPPD	CPPD	100 µg/ml	Invivogen	NLRP3 and TLR2	Liu-Bryan et al., 2005; Martinon et al., 2006
FSL-1	FSL	2 µg/ml	Invivogen	TLR2/6	Shibata et al., 1997; Okusawa et al., 2004
Poly I:C	pIC	20 µg/ml	Invivogen	TLR3	Alexopoulou et al., 2001
LPS-EB (ultrapure)	LPS	10 ng/ml	Invivogen	TLR4	Poltorak et al., 1998; Shimazu et al., 1999
Flagellin-ST	FLA	0.25 µg/ml	Invivogen	TLR5	Hayashi et al., 2001
Gardiquimod	GARD	3 µM	Invivogen	TLR7	Hemmi et al., 2002
R848	R848	1 µM	Invivogen	TLR7 and TLR8	Jurk et al., 2002
ODN 2216	ODN	25 µg/ml	Invivogen	TLR9	Hemmi et al., 2000; Krieg, 2002
lipoarabinomannan	LAM	10 µg/ml	Invivogen	Mannose R, CD36	Józefowski et al., 2011; Sieling et al., 1995
WGP	WPG	40 µg/ml	Invivogen	Dectin-1	Goodridge et al., 2011
<b>Cytokines</b>					
IFN-α <sub>2b</sub> (Intron A)	IFN-A	1,000 IU/ml	Merck	IFNAR	González-Navajas et al., 2012
IFN-β (Betaseron)	IFN-B	1,000 IU/ml	Bayer	IFNAR	González-Navajas et al., 2012
IFN-γ (Imukin)	IFN-G	1,000 IU/ml	Boehringer Ingelheim	IFNγR	Platanias, 2005
TNF-α	TNF-A	10 ng/ml	Miltenyi Biotech	TNFR	Kolb and Granger, 1968
IL-1β	IL-1B	25 ng/ml	Peptotec	IL1R	March et al., 1985
IL-23	IL-23	50 ng/ml	Miltenyi Biotech	IL23R	Oppmann et al., 2000
<b>T Cells</b>					
α-CD3 + α-CD28	CD3+CD28	0.4 µg/ml + 0.33 µg/ml	RND Systems + Beckman Coulter	TCR	Smith-Garvin et al., 2009
Enterotoxin SEB	SEB	0.4 µg/ml	Bernhard Nocht Institute	TCR and MHC II	Fleischer and Schrezenmeier, 1988

Abbreviations are as follows: HK, heat killed; HAU, hemagglutinin units; IU, international units.

The 28 stimulation conditions used for the preparation of TruCulture tubes are listed, with the indicated dose and commercial supplier. Stimuli are ordered based on four categories: whole microbe, MAMP, cytokine, and T cell agonist. See also Figure S1 and Tables S1, S2, S3, and S4.

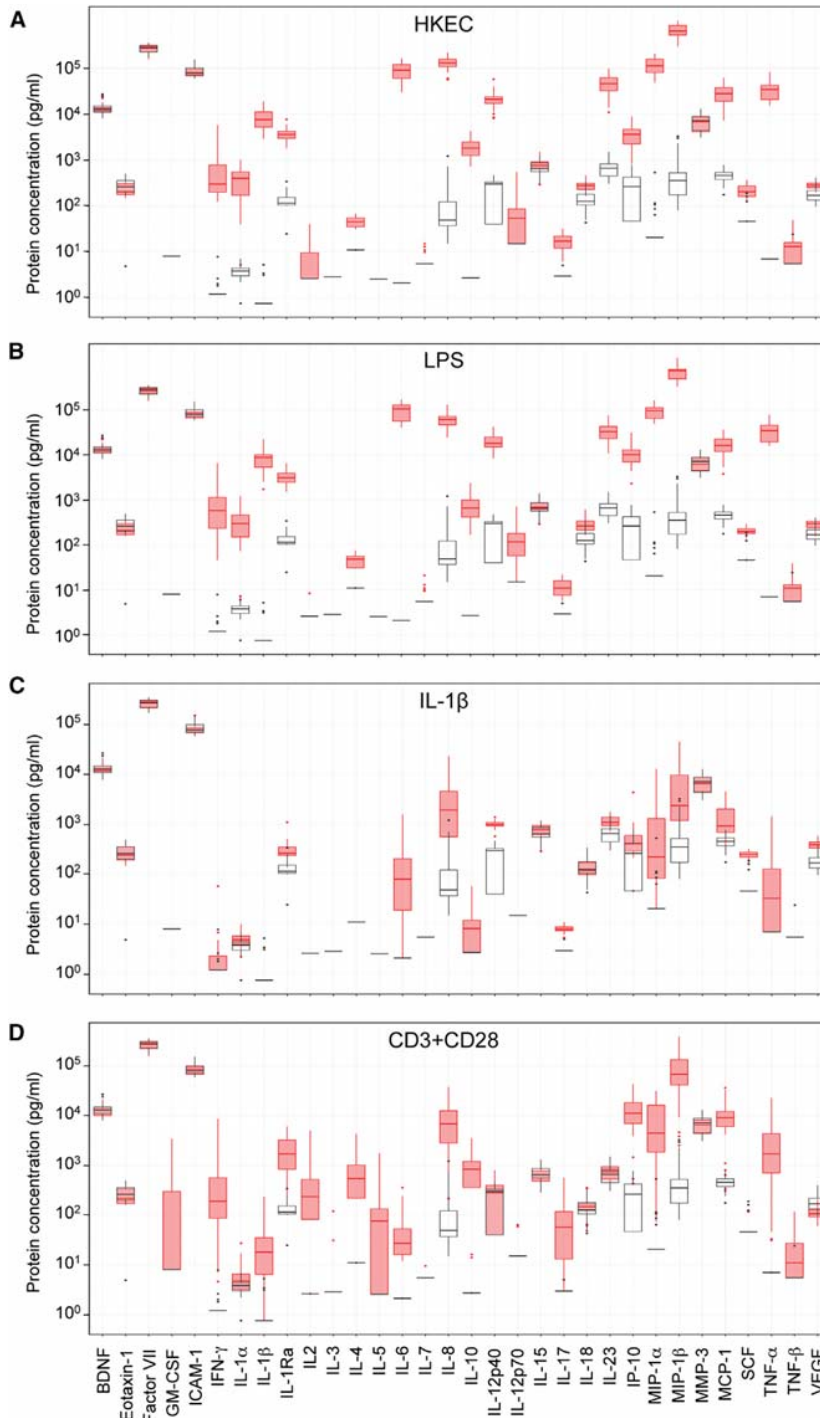
Samples were collected, processed, and analyzed as described (Figure S1A). Luminex assays employed in the study are listed and the lower limit of quantification (LLOQ) and least detectable doses (LLD) for each assay are indicated (Table S3). In order to assess the overall signatures induced in the 28 conditions, we plotted the concentration of the measured analytes across all donors (four representative stimulation systems are shown: HKEC, LPS, IL-1β, and CD3+CD28, with the null response overlaid in each graph; Figure 1). Notably, we achieved a range of

induced biologic responses, spanning, in some instances, greater than 1,000-fold as compared to the null condition (e.g., IL-6, MIP-1α). Importantly, the stimulations achieved by the assay systems did not exceed the measured biologic limit (as defined by a plateau in the response of selected analytes), and a broad range of induced protein responses were observed. Several protein analytes remained unchanged across all stimulation systems (i.e., IL-7, MMP-3, and sICAM-1) and were therefore removed from further analysis.



## Immunity

### Defining Boundaries of a Healthy Immune Response



**Figure 1. Dynamic Range of Stimulation Systems**

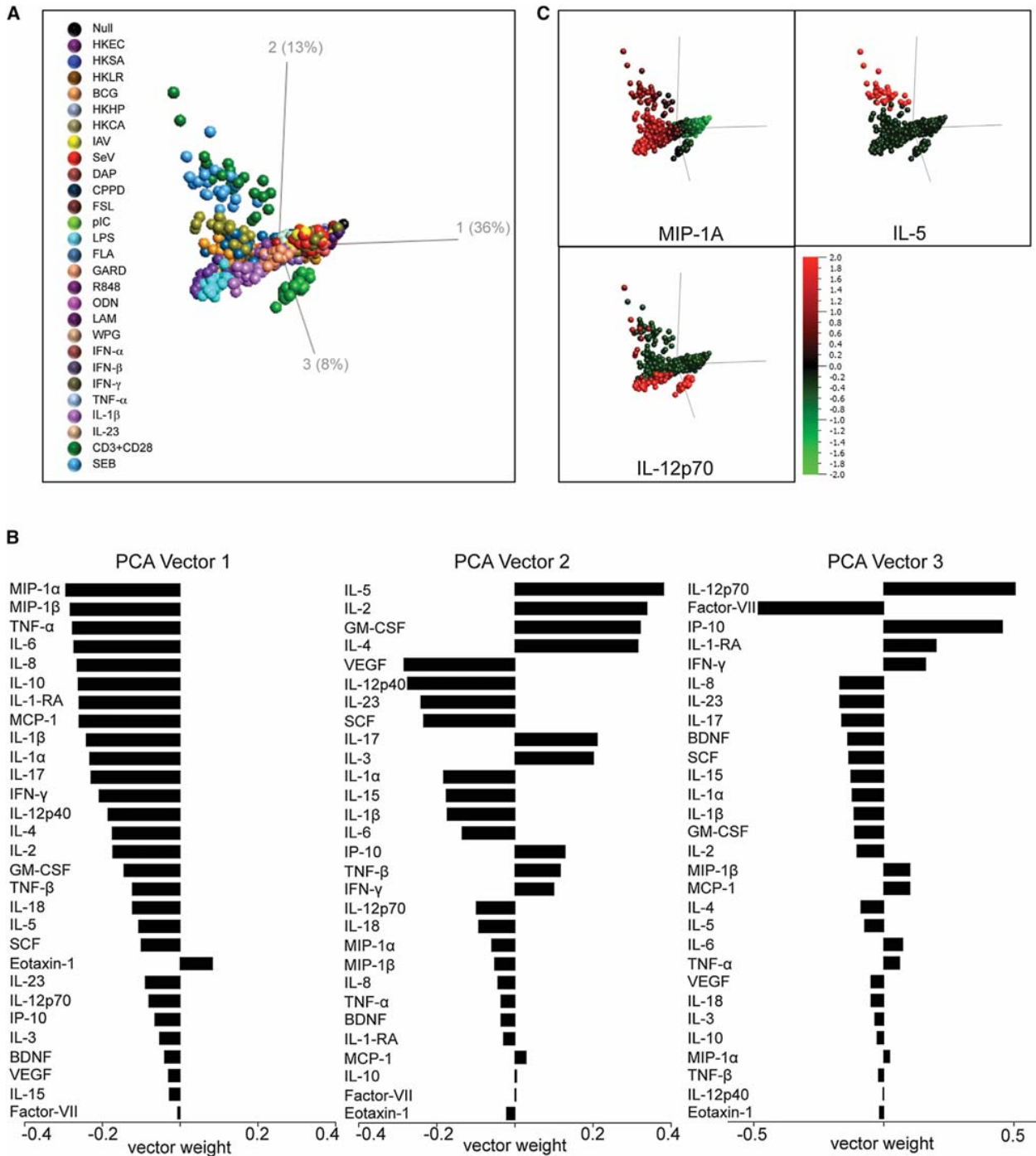
Box-whisker plots indicate the induced protein response for 25 healthy donors for 4 representative stimuli: HKEC (A), LPS (B), IL-1 $\beta$  (C), and CD3+CD28 (D). Induced responses are in red, and the null response is overlaid in gray. Protein analytes are reported in pg/ml and listed alphabetically. The median is represented by the horizontal line, the interquartile range (IQR) by the box, and the whiskers represent 1.5  $\times$  IQR. Data beyond the end of the whiskers are outliers and plotted as points.

each of the 25 donors is represented (Figure 2A). The PCA revealed strong stimulus-specific clusters, with the first three principal component (PC) vectors explaining 57% of the total variance (Figures 2A and 2B). Highlighting the presence of a common, core signature for the induced innate response, we found that PC1 is composed by the contribution of chemokines and cytokines: MIP-1 $\alpha$ , MIP-1 $\beta$ , TNF- $\alpha$ , IL-6, IL-8, IL-10, IL-1RA, and MCP-1. Interestingly, PC2 separated stimuli that induced an adaptive immune signature and was mainly driven by IL-5, IL-2, GM-CSF, and IL-4. Stimuli that were directed toward the second principal component axis included CD3+CD28, SEB, and to a lesser extent *Candida albicans* (HKCA), bacillus Calmette-Guerin (BCG), and *Staphylococcus aureus* (HKSA). PC3 was composed of IL-12p70 and IP-10 as induced analytes and Factor-VII as a suppressed factor. The signature achieved by pIC stimulation could be easily separated across this third principal component axis. Illustrating how the calculated vectors relate to the overall PCA, we superimposed expression data for the top analyte of each vector on the PCA plot (Figure 2C).

To further validate our approach and to explore the underlying architecture of the PCA, we focused on the two T cell stimulation systems: CD3+CD28 and SEB stimulation. When analyzed separately, the two stimulation conditions could be easily distinguished with a PCA plot that was based on 12 protein analytes that showed statistical differences

We next analyzed the data by unsupervised principal component analysis (PCA), employing Qlucore Omics Explorer 2.3 (Andersson et al., 2005). Prior to applying PCA, values for each of the 29 protein analytes were centered to a mean value of zero and scaled to unit variance. The 27 stimuli and null control are indicated by the filled circles and the vector position of

(CD3+CD28 versus SEB, MW q value < 0.05) (Figures 3A–3C). Interestingly, CD3+CD28 and SEB induced similar amounts of T cell cytokines (e.g., IL-4, IL-5); however, the distinct mechanisms of activation—unique action on the TCR signaling pathway as compared to cross-linking of MHCII and the TCR, respectively (Fleischer and Schrezenmeier, 1988)—accounted for higher

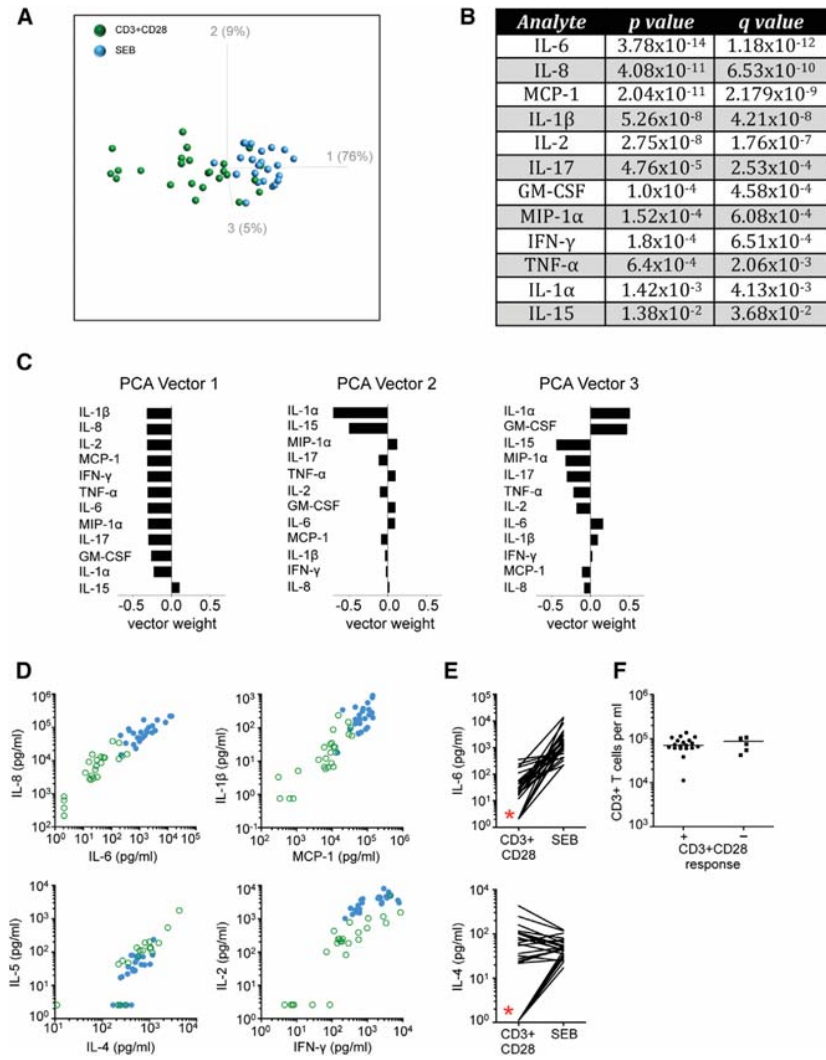


**Figure 2. Distinct Inflammatory Signatures for Stimulation Systems**

(A) Principal component analysis (PCA) was performed on the data set obtained from 25 healthy donors. Each colored circle represents one of the 28 different whole-blood stimulation conditions, and the PCA was run with data obtained from the analysis of 29 proteins. The PCA plot shown captures 57% of the total variance within the selected data set (PCA1, 36%; PCA2, 13%; PCA3, 8%).

(B) The contribution of each protein analyte to the three principal component axes of the PCA plot are shown. (The positioning of the bars is arbitrary and is not considered negative or positive except in relation to the other analytes.)

(C) The protein analyte contributing most strongly to each of the three principal component axes was overlaid on the PCA plot. A heat map indicates the relative expression of the indicated protein analyte (red indicating high levels of expression, green indicating low levels of expression).



**Figure 3. CD3+CD28 and SEB Induce Distinct Inflammatory Signatures**

(A) Principal component analysis (PCA) was performed on the data set obtained from 25 healthy donors for the response to CD3+CD28 (green circles) and SEB (blue circles) stimulation systems, and discriminating protein analytes (q value [ANOVA FDR adjusted p value] < 0.05) were incorporated in the analysis. The PCA plot shown captures 90% of the total variance within the selected data set.

(B) The induced responses to whole-blood stimulations with CD3+CD28 and SEB were compared and the 12 differentially expressed proteins were identified (ANOVA q value < 0.05).

(C) The contribution of each protein analyte to the three principal component axes of the PCA plot are shown. (The positioning of the bars is arbitrary and is not considered negative or positive except in relation to the other analytes.)

(D) Correlation plots highlight differentially and similarly expressed proteins after whole-blood stimulations with CD3+CD28 (open green circles) or SEB (closed blue circles).

(E) Pairwise comparison for IL-6 and IL-4 concentration is shown after whole-blood stimulations with CD3+CD28 and SEB. Black lines indicate individual donors. Red star highlights non-responders to CD3+CD28 stimulation.

(F) The number of CD3<sup>+</sup> T cells per ml of whole blood in CD3+CD28-positive and -negative responders. See also Figure S2.

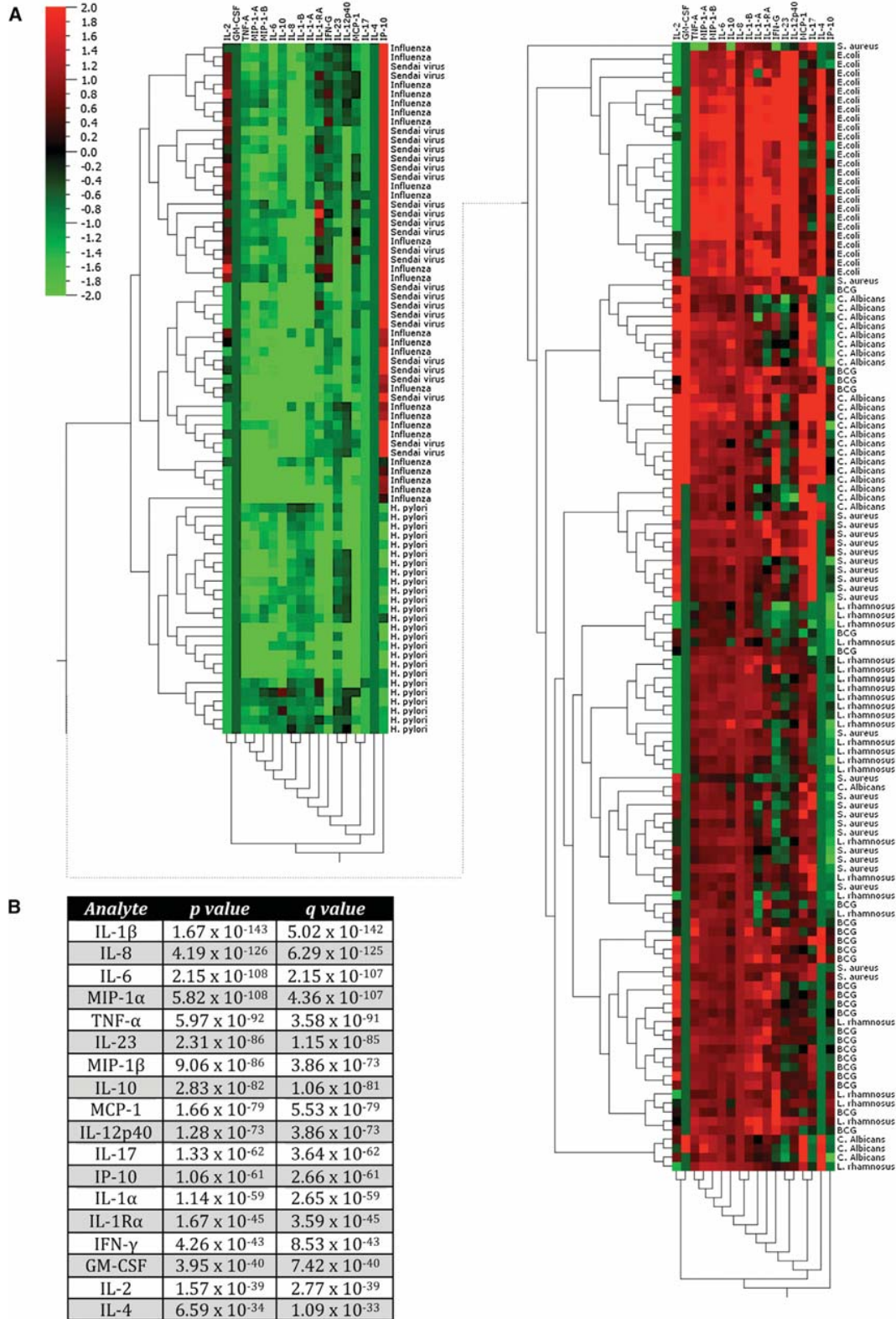
concentrations of SEB-induced IL-6, IL-8, MCP-1, and IL-1 $\beta$  (Figure 3D). This was confirmed experimentally by intracellular flow cytometry staining, which showed that after SEB, but not CD3+CD28 stimulation, the MHCII-expressing monocytes were induced to express IL-8, MCP-1, and IL-1 $\beta$  protein (Figure S2). Interestingly, we observed that 6 of the 25 donors failed to produce IFN- $\gamma$ , IL-2, IL-4, or IL-6 in response to CD3+CD28 stimulation, whereas all donors were capable of responding to SEB (Figures 3D and 3E). These data highlight the ability to reliably measure cell-cell interactions within the whole-blood stimulation conditions. Moreover, this approach permitted the identification of healthy donors that were unable to respond fully to anti-CD3 (clone UCHT1) stimulation, despite the binding of the antibody to donor T cells as confirmed by flow cytometry (Figure 3F).

**Distinct Inflammatory Signatures Induced by Whole Microbes, Microbe-Associated Agonists, or Cytokine Stimulation**

Complex stimuli used in our stimulation systems included heat-killed *Escherichia coli* O111:B4 (HKEC), *Staphylococcus*

(Table 1). The *E. coli* used was derived from a strain that causes acute diarrhea in babies (Viljanen et al., 1990). *H. pylori* is also a human pathogen and is the main cause of ulcer disease and stomach cancer in humans (Wroblewski et al., 2010). Healthy donors may be carriers for *S. aureus* or *C. albicans*, but in some instances these agents may be the cause of human disease (e.g., in immunologically compromised individuals) (Gow et al., 2012; Otto, 2009). BCG is used as a vaccine in order to protect humans from childhood tuberculosis and is the standard of care for treatment of bladder cancer (NB: all donors received BCG vaccination) (Kawai et al., 2013; Romano and Huygen, 2012). *L. rhamnosus* is considered to be a transient inhabitant of humans and is present in some yogurt preparations (Borriello et al., 2003). Most humans are exposed to H1N1 IAV as a result of seasonal epidemics or through vaccination; and to serve as a contrast to IAV, we selected SeV, which does not infect humans yet triggers an innate inflammatory response (Kato et al., 2005; Norrby et al., 1992).

To characterize the patterns of protein analytes induced by such complex stimuli, we performed hierarchical clustering and



(legend on next page)

focused on the 18 most highly discriminating read-outs, all with a  $q$  value  $< 10^{-30}$  (Figures 4A and 4B). This approach separated IAV and SeV from the other stimuli, based on the induction of high IP-10 levels. HKEC was the most potent stimulus, marked by the highest levels of the pyrogenic cytokines TNF- $\alpha$ , IL-6, and IL-1 $\beta$ , as well as high levels of IL-12p40 and IL-23 (Figures 4A and S3). There was some overlap among the donor responses to HKCA, HKSA, HKLR, and BCG, but distinct patterns could be discerned. HKCA induced high amounts of GM-CSF in 20 of 25 donors, significantly higher than any other microbial stimulus (KW  $p < 1 \times 10^{-7}$ ). As observed in the overall PCA plot (Figure 2), we could discriminate HKSA, BCG, and HKCA based on their induction of IL-2, possibly a reflection of our donor population having been previously exposed to these microbial agents (Figures 4A and S3). We also noted an interesting pattern of expression for IL-12p40, IL-12p70, and IL-23; most stimuli triggered IL-12p40, but only HKEC triggered IL-12p70 in a large number of donors (16 of 25) (Figure S3). Additionally, for some stimuli there appeared to be a bimodal pattern of induced responses, in particular the HKHP-, IAV-, and SeV-induced TNF- $\alpha$  and IL-12p40 and the HKEC-, HKLR-, HKCA-, IAV-, and SEV-induced IL-2 response (Figure S3).

We next selected the TLR agonists for analysis, because we achieved extensive coverage of this class of host sensors (Table 1). With the exception of the TLR1-TLR2 heterodimer, we were able to validate stimulation systems for the known TLR receptors expressed by humans (NB: Pam3CSK4 was evaluated, but failed short-term stability testing). FSL-1 (FSL, and also known as Pam2C) is a synthetic diacylated lipoprotein mimicking an agonist present in *Mycoplasma salivarium* (Okusawa et al., 2004). A high-molecular-weight, vaccine-grade poly IC (pIC) was used to activate TLR3. Ultrapure lipopolysaccharide (LPS) derived from *E. coli* O111:B4 was used to stimulate TLR4. For TLR5, we selected ultrapure flagellin (FLA), extracted from *Salmonella* Typhimurium. To uniquely stimulate TLR7, we utilized the vaccine-grade preparation of Gardiquimod (GARD), an imidazoquinoline compound, and we used a related molecule, R848 (also vaccine-grade) as a stimulator with mixed agonist activity for TLR7 and TLR8. The TLR9 agonist selected was the class A CpG-2216 oligonucleotide (ODN), a fully synthetic oligonucleotide that contains unmethylated CpG dinucleotides within a particular sequence.

As shown in the PCA plot, it was possible to segregate the TLR stimuli based on the induced protein signatures. We selected the 11 most significant proteins (identified by ANOVA,  $q$  value  $< 10^{-60}$ ), which allowed us to capture 93% of the measured variance in the response to TLR stimulation (Figure 5A). The notable exception was the overlap between FSL and FLA, which may be explained by the similar cellular expression of TLR2/6 and TLR5 on circulating monocytes (Mäkelä et al., 2009) and the use of a common MyD88-dependent signaling pathway (Takeda et al.,

2003). R848 and GARD also showed a similar signature, yet the two could be segregated based on the overall higher levels of induced cytokines/chemokines and the increased number of donors that produced IL-12p70 to GARD (Figure S4A), a likely reflection of TLR8 engagement on monocytes (Bekeredjian-Ding et al., 2006). LPS triggered the strongest inflammatory response, as shown by the significantly higher levels of pyrogenic cytokines induced (TNF- $\alpha$ , IL-1 $\beta$ , and IL-6 higher for LPS as compared to the other stimuli, KW  $p < 1 \times 10^{-7}$ , Figure S4A). One caveat was a sampling bias in the selection of analytes measured, because the Luminex panels were oriented toward LPS-induced responses. We also highlight the relatively weak response induced by ODN, which we believe results from the agonist being quenched by the whole-blood matrix. Alternatively, it could be a reflection of the low numbers of plasmacytoid dendritic cells that are present within 1 ml of whole blood. However, upon removal of ODN from the analysis, the remaining TLR ligands kept their unique position (apart from FSL and FLA) within the PCA and a similar level of variance (94%) was captured (Figure S4B). In addition, we were able to distinguish ODN from the null condition based on eight analytes (ODN versus null, MW  $q < 0.05$ ; Figures S4C and S4D), with the most induced protein being IP-10, a reflection of type I IFN being produced as a result of TLR9 stimulation (Krug et al., 2004). From the initial PCA plot (Figure 2), pIC stimulation could be distinguished by its unique inflammatory signature. This is recapitulated when pIC is compared to the other TLR agonists; the pattern of protein expression being remarkable for the high levels of IL-12p70 and the complete absence of induced IL-10. A bimodal distribution was again seen for some cytokines; in particular, a certain number of donors failed to produce IL-12p40 after FSL, FLA, or ODN and others did not produce IL-12p70 after LPS or R848 (Figure S4A).

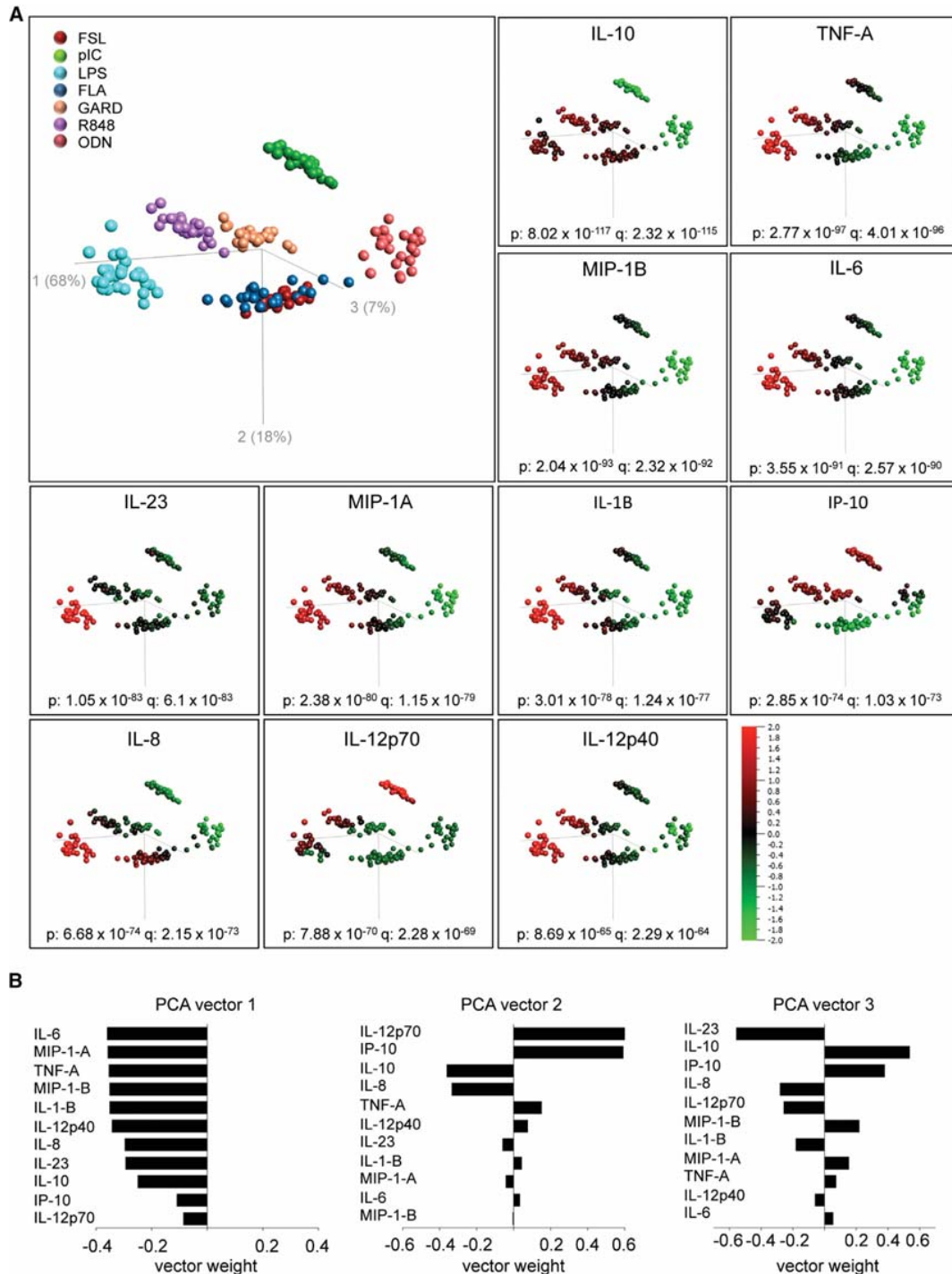
To provide insight into the variable response to stimulation via cytokine receptors, we exposed cells to IFN- $\alpha_{2a}$ , IFN- $\beta$ , IFN- $\gamma$ , TNF- $\alpha$ , IL-1 $\beta$ , or IL-23. The latter four cytokines were also measured as one of the proteins assessed in the multiplex luminex assays. As such, we had an internal control that donors were stimulated with similar cytokine concentrations; additionally, it was important to exclude the measured variable from the stimulation signature (Figures S5). As expected because of their use of the same IFN- $\alpha/\beta$  receptor, the signatures for IFN- $\alpha_{2a}$  and IFN- $\beta$  were identical and also similar to that seen for IFN- $\gamma$ , which induces a common set of interferon-stimulated genes (ISGs) and proteins (Der et al., 1998). These data also permit deconvolution of some of the more complex signatures. For example, we highlight that pIC results in the induction of both TNF- $\alpha$  and IL-1 $\beta$  (Figure S4), both of which can induce IL-10 when they are used as stimuli (Figures S5); yet there was a clear absence of IL-10 induction upon pIC stimulation (Figure S4). Although this may be a reflection of lower levels

#### Figure 4. Unique Inflammatory Signatures Induced by Complex Microbial Stimulation

(A) Hierarchical clustering was performed on the data set obtained from 25 healthy donors, restricting the analysis to whole-blood stimulation by heat-killed (HK) *E. coli* (HKEC), HK *S. aureus* (HKSA), HK *L. rhamnosus* (HKLR), *Bacillus Calmette-Guérin* (BCG), HK *H. pylori* (HKHP), HK *C. albicans* (HKCA), influenza A virus (IAV), and Sendai virus (SeV). A heat map is shown, based on the 18 most differentially induced proteins as defined by ANOVA  $q$  values.

(B) The 18 most differentially expressed proteins that were used to define microbe stimulation-specific signatures are listed in order of statistical significance (cutoff value was determined by ANOVA,  $q$  value  $< 10^{-30}$ ).

See also Figure S3.



**Figure 5. Segregation of TLR Agonists Based on Their Induced Protein Signatures**

(A) Principal component analysis (PCA) was performed on the data set obtained from 25 healthy donors. Analysis was restricted to the 7 whole-blood stimulations that contained TLR agonists (FSL, pIC, LPS, FLA, GARD, R848, ODN). Each colored circle represents a different whole-blood stimulation condition, and the PCA was run with the 11 most differentially induced proteins (cutoff value was determined by ANOVA, q value <  $10^{-60}$ ). The PCA plot shown captures 85% of the total variance within the selected data set. Expression levels for each of the 11 protein analytes was overlaid on the PCA plot. A heat map indicates

(legend continued on next page)

of pIC-induced TNF- $\alpha$  and IL-1 $\beta$ , we favor the interpretation that TRIF activation or perhaps the high levels of STAT1 signaling results in suppression of IL-10 expression (Saraiva and O'Garra, 2010).

### Characterization of Naturally Occurring Variance to Immune Stimulation

The development of reliable stimulation systems for monitoring immune responses permits the establishment of reference ranges for healthy individuals. Moreover, it permits the classification of inflammatory and host immune responses based on the variability among healthy donors as well as the identification of responses outside the defined reference range. To rapidly visualize variance among our 25 donors, we plotted the induced responses on a radar plot (Figure 6): absolute concentrations of the induced proteins are plotted along the spokes of the plot; lines trace the induced protein signature from individual donors; the shaded gray polygon indicates the median value of the null condition; and the black circles mark the induced fold change over the median null value. Data have been sorted so the least induced protein is at the top of the radar plot, with increasing fold change plotted in a clockwise manner. Analytes were excluded from the signature if the absolute value of the median fold change (stimulation/null) was <1.3. Data from the LPS stimulation system are shown and all plots can be found as an online Excel file (Document S2).

For the LPS-induced signature, we highlight that several induced cytokines and chemokines showed limited interindividual variance (CV < 50%). By contrast, other analytes showed high variance; for example, IL-12p70 and IFN- $\gamma$  showed a range that spanned more than two orders of magnitude with CVs of 106% and 132%, respectively (Table S4). Additionally, this representation permitted the identification of two individuals that were outliers in their failure to produce measurable amounts of IL-1 $\alpha$  in response to LPS (red star, Figure 6A). Notably, the rest of the signature was intact. To explore this finding further, we studied induced IL-1 $\alpha$  across the entire stimulation space. Data from four consecutive donors are shown (including one of the outlier individuals identified), and we compared the response for three proteins that showed a high correlation (IL-1 $\alpha$ , IL-1 $\beta$ , and IL-6). Expression of the IL-1 receptor antagonist (IL-1Ra) is also shown, because it is involved in the IL-1 pathway. Strikingly, none of the stimuli used triggered detectable levels of IL-1 $\alpha$  production by donor 203 (or donor 312, not shown); this was in contrast to the induction of IL-1 $\beta$  and IL-6, which was within the range of values reported for the other donors tested (Figure 6B and Table S4). Given the importance of IL-1 $\alpha$  in sterile inflammation and disease pathogenesis, we believe that our findings will be of general interest. Moreover, we highlight the value of utilizing standardized measures for host immune responses, thus enabling the identification of interindividual variance and extreme phenotypes among human populations.

### DISCUSSION

The definition of host immune responses to microbial agents is of major interest and facilitates an increased understanding of human health and disease pathogenesis. Although functional tests are routinely used in laboratory investigation (Folds and Schmitz, 2003), the standardization of assays has been challenging. Indeed, there exist few examples of standardized systems for measuring induced immune response in human population-based studies or clinical practice. This study aimed at testing whole-blood stimulation systems for medically relevant stimuli to determine the inflammatory signature and characterize the naturally occurring variation present in a population of healthy donors of European descent. The robust definition of the boundaries of a healthy immune response at the population level is an indispensable prerequisite to subsequently understand how perturbations in these responses correlate and in some instances account for a pathologic state. Our approach utilizes a practical solution to monitoring induced immune responses and requires only 1 ml of blood per stimulation system and a 37°C heating block, maintained in room air. Additionally, there is minimal sample handling and specialized technical experience is not required.

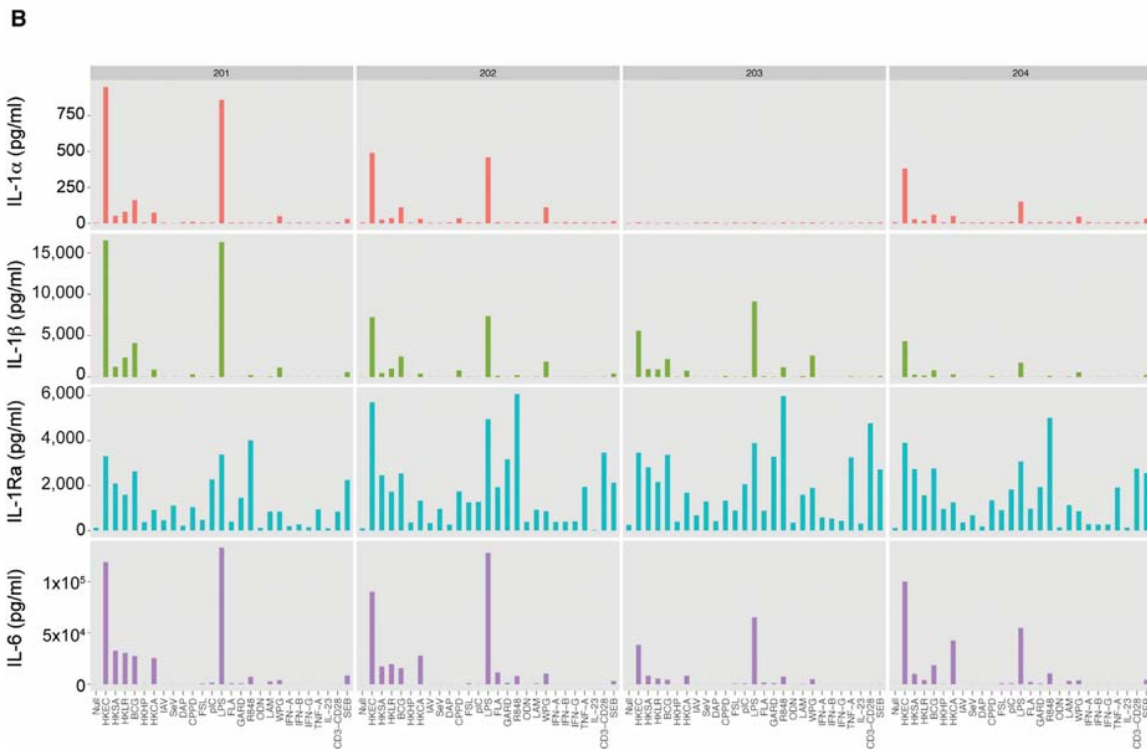
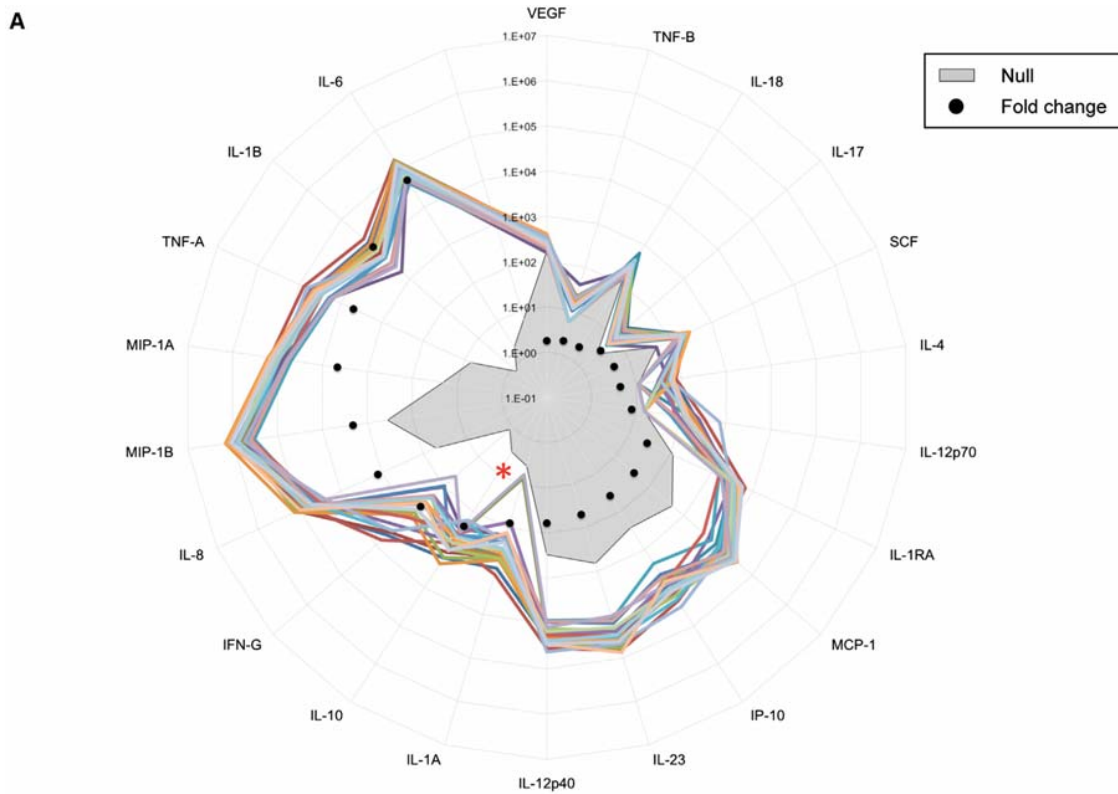
The concept of utilizing whole-blood assays for assessing leukocyte function was first introduced by Ruuskanene and colleagues in 1975 (Eskola et al., 1975), used at that time for monitoring PHA and ConA-induced lymphocyte proliferation. Digel and colleagues extended this approach to the study of cytokines in 1983 (Digel et al., 1983), reporting the use of whole-blood stimulation with SEA and anti-CD3 antibodies, followed by the measurement of type I and type II interferons. Over the last three decades, whole-blood cultures have been utilized for probing various aspects of the immune response (Chen et al., 2010; De Groote et al., 1992; Ida et al., 2006; Kirchner et al., 1982; Nerad et al., 1992; Pott et al., 2009). However, several problems have persisted, including the ill-defined period of time between blood draw and cell culture and the requirements for specialized lab equipment (e.g., tissue culture hoods, CO<sub>2</sub> incubators). One notable exception has been the clinical development of the QuantiFERON TB Gold In-Tube (QFT-G IT) assay (Santin et al., 2012), which has been approved for the diagnosis of latent tuberculosis infection. QFT-G IT measures the induction of IFN- $\gamma$  production in whole blood after *in vitro* stimulation with *Mycobacteria tuberculosis* antigens.

In this study, we report the development and testing of 27 stimulation systems, built into whole-blood syringes. We aimed to test our assay system via a broad array of immune stimuli, including bacteria, fungi, and viruses; agonists specific for defined innate immunity sensors; clinically employed cytokines; and activators of T cell immunity. With the exception of two assay systems (CPPD and WGP), the coefficient of variance was low, in the range of 5%–14%, with long-term stability of up to 12 months. The endpoints chosen for evaluating the

the relative expression of the indicated protein analyte (red indicating high levels of expression, green indicating low levels of expression). ANOVA p and q values are reported.

(B) The contribution of each protein analyte to the three principal component axes of the PCA plot are shown. (The positioning of the bars is arbitrary and is not considered negative or positive except in relation to the other analytes.)

See also Figure S4.



(legend on next page)



inflammatory signature consisted of a selection of inducible cytokines, chemokines, and growth factors. Importantly, dose-finding studies ensured that the induced responses did not exceed the biologic limit, as indicated by the broad range of analyte concentrations observed across the different stimulation conditions. Unique, specific signatures were identified for most stimulation systems. As an initial validation of lymphocyte activation, we observed an expected T cell signature (e.g., induction of IL-5, IL-2, GM-CSF, IL-4) when using anti-CD3 and anti-CD28 or SEB stimulation. These two signatures were clearly separated from the other stimuli in the global analysis and could be distinguished from each other, a result of SEB activation of MHCII-expressing cells. We also identified six healthy donors who failed to respond to CD3+CD28 stimulation despite binding of the CD3 antibody to T cells. This confirmed earlier reported findings and might be due to polymorphisms in the Fc $\gamma$ RII expressed by monocytes (Ceuppens et al., 1985; Tax et al., 1983) that bind mouse IgG1 (the subclass of the UCHT-1 anti-CD3 clone used), or it might be due to other as yet unidentified common genetic variants. Interestingly, we also observed a modest “lymphocyte” signature when we utilized BCG, HKSA, and HKCA stimulation. Specifically, these three stimuli induced low levels of IL-2, which was not observed when we used HKEC, HKLR, or HKHP.

When directly comparing whole microbes, we could identify clear signatures for HKEC (strong induction of pyrogenic cytokines and high expression of both IL-12p70 and IL-23), HKCA (based on the stimulation of GM-CSF expression), and the two viral stimuli, IAV and SeV (triggering the highest levels of IP-10). Notably, the interindividual variance was highest for microbial-induced IL-2 and IFN- $\gamma$  (CV in range of 78%–165%, 60%–183%, respectively), HKEC-induced IL-12p70 (CV = 155%), HKHP-induced IL-10 (CV = 216%), and BCG or HKCA-induced GM-CSF (CV = 175% and 216%, respectively). These differences are presumed to be due to a combination of host genetic factors and environmental, both internal and external, exposures (Newport et al., 2004). We may also consider that prior exposure and/or carrier state (e.g., colonization by HKCA) might account for differential memory responses (e.g., lymphocyte activation or antibody opsonization of the microbe) (Zielinski et al., 2012), in turn impacting the magnitude of the inflammatory response. We also highlight the relatively weak response to HKHP stimulation, which is probably due to the bacterium harboring an extensively modified lipid A moiety as part of its LPS, which reduces by >1,000-fold its TLR4 agonist activity (Cullen et al., 2011) and a flagellin that is poorly recognized by TLR5 (Gewirtz et al., 2004; Lee et al., 2003). Furthermore, stimulation of TLR2 by *H. pylori* mediates a tolerogenic response (Sun et al., 2013), potentially contributing to the weak response induced by HKHP.

To capture a more precise measure of the innate response, we also utilized purified or synthetic MAMPs known to engage the TLR, NLR, or CLR families of microbial sensors. These pathways have been heavily investigated over the past two decades and efforts are underway to establish some of the selected ligands as adjuvants for vaccine formulation. For NF- $\kappa$ B-induced cytokines (e.g., TNF- $\alpha$ , IL-1 $\beta$ , IL-6, MIP-1 $\alpha$ , MIP-1 $\beta$ , IL-8, and IL-12p40), we found a similar pattern of expression across the different stimuli. LPS was unique in its induction of IL-23, whereas pIC induced the highest levels of IL-12p70. Although our xMAP testing did not evaluate many interferon-induced proteins, the levels of IP-10 were consistent with the endosomal TLRs being more robust stimulators of IRF3 than the surface receptors (Blasius and Beutler, 2010). DAP was a relatively weak stimulus, possibly because of poor membrane permeability, though we were able to observe a consistent induction of NF- $\kappa$ B-dependent chemokines or cytokines. Although they were less reliable than other stimulation systems, we were able to detect strong signatures by using CPPD and WGP, both of which showed high induction of IL-1 $\beta$  and measureable levels of IL-18, a likely result of inflammasome activation. Of note, the interindividual variance for many of the induced proteins was greater than the intraindividual variance of the assay systems. The LAM signature was notable for the highest interindividual variance in IL-10 among the different stimuli used (range 2.7–1,100 pg/ml; CV = 158% for LAM-induced IL-10). Our data are consistent with mRNA and protein expression patterns that have been evaluated via transcriptional profiling, ELISA, or Luminex on specific stimulated cell types, such as human monocytes and dendritic cells (Huang et al., 2001; Kwissa et al., 2012; Torri et al., 2010). Our study, however, represents a systematic evaluation of pattern recognition receptor (PRR) activation that takes into consideration the complex cellular interactions occurring in whole blood and serum matrix components, which might be closer to the natural conditions in which immune responses are provided.

One of the most interesting results of our study was the identification of 2 of 25 donors who did not release IL-1 $\alpha$  after stimulation with any of the 27 different stimuli. Despite the failure to detect IL-1 $\alpha$ , all remaining chemokine and cytokine signatures were intact for these donors, including the production of IL-1 $\beta$ . Two distinct but related genes, *IL1A* and *IL1B*, encode for IL-1 $\alpha$  and IL-1 $\beta$ , respectively (Dinarello, 2009). Both bind the same surface receptor, and both are antagonized by the soluble protein IL-1Ra. Notably, IL-1 blockade, by means of IL-1RA or neutralizing antibodies, has become central to the clinical management of rheumatologic diseases and hereditary systemic autoinflammatory disorders (Dinarello et al., 2012). IL-1 $\alpha$  is expressed by most cells and because of the lack of a signal peptide it is not readily secreted. Intracellular IL-1 $\alpha$  is preformed

#### Figure 6. Interindividual Variance in the Response to LPS Stimulation

(A) Radar plot representation of the LPS-induced response obtained from 25 healthy donors. Analytes are represented as picograms per milliliter (pg/ml) and ordered clockwise in increasing fold change (as compared to null). Each donor is represented by a colored line, connecting the concentration of measured protein analytes. The gray polygon depicts the median value of the null response for the 25 donors. Black dots indicate the fold change as compared to the median value of the null response. Analytes with a median fold change (stimulation/null) >1.3 or <-1.3 were included. A red asterisk highlights the identification of two donors in which IL-1 $\alpha$  was not induced above background.

(B) Histogram plots representing the IL-1 $\alpha$ , IL-1 $\beta$ , IL-1R $\alpha$ , and IL-6 response for 4 consecutive donors are shown for the Null condition and 26 whole-blood stimuli (NB: the IL-1 $\beta$  stimulation tube was omitted from analysis because it confounds the measurement of IL-1 $\beta$ ).

See also Figure S5.

and bioactive; as such, its release from damaged cells is considered to be one of the first steps in the initiation of so-called sterile inflammation. In our study, we detected measurable amounts of IL-1 $\alpha$  in the culture supernatant after whole-blood stimulation with HKEC, HKSA, HKLR, BCG, HKCA, SeV, CPPD, LPS, R848, and WPG (defined by median fold change over null stimulation > 1.3). Future studies will be required in order to identify how the genetic makeup of the host, including common polymorphisms in the European population, may account for the failure of the two donors to release IL-1 $\alpha$  in the setting of multiple immune stimulations.

In summary, the whole-blood collection stimulation systems presented allow the definition of induced inflammatory signatures for a broad range of innate and adaptive stimuli, helping to address the urgent need for monitoring of functional immune responses in a reliable and reproducible manner. Moreover, we have identified preliminary boundaries for the natural variation in the induced immune protein phenotypes, setting the basis for a better understanding of the meaning of a healthy immune response. These tools will support integrative and systems-level human population-based studies (Braga-Neto and Marques, 2006) aimed at defining the genetic and/or environmental determinants of natural or disease-induced variation in immune responsiveness.

## EXPERIMENTAL PROCEDURES

### Donors

Samples were obtained as part of the *Milieu Intérieur* Healthy Donor Cohort. Details are provided in [Supplemental Experimental Procedures](#) and can be found at <http://www.clinicaltrials.gov> (identifier NCT01699893).

### Whole-Blood Stimulation

TruCulture tubes were prepared in batch with the indicated stimulus, resuspended in a volume of 2 ml buffered media, and maintained at  $-20^{\circ}\text{C}$  until time of use. Blood was obtained from the antecubital vein by a 60 ml syringe containing sodium-heparin (50 IU/ml final concentration). Within 15 min of collection, 1 ml of whole blood was distributed into each of the prewarmed TruCulture tubes, inserted into a dry block incubator, and maintained at  $37^{\circ}\text{C}$  ( $\pm 1^{\circ}\text{C}$ ) room air for 22 hr ( $\pm 15$  min). At the end of the incubation period, tubes were opened and a valve was inserted in order to separate the sedimented cells from the supernatant and to stop the stimulation reaction. Liquid supernatants were aliquoted and immediately frozen at  $-80^{\circ}\text{C}$  until the time of use.

### Multianalyte Profiling and Identification of Inflammatory Signatures

Supernatants from whole-blood stimulation systems were analyzed with Luminex xMAP technology. Samples were measured according to CLIA guidelines (validated by guidelines set forth by the USA Clinical and Laboratory Standards Institute). The 32 measured analytes were organized on three multiplex arrays, and a single batch of reagents was used for testing all samples. The least detectable dose (LDD) for each assay was derived by averaging the values obtained from 200 runs with the matrix diluent and adding 3 standard deviations to the mean. The lower limit of quantification (LLOQ) is determined based on the standard curve for each assay and is the lowest concentration of an analyte in a sample that can be reliably detected and at which the total error meets CLIA requirements for laboratory accuracy. For analytes tested, the LDD and LLOQ can be found in [Table S3](#). The lower assay limit (LAL) is the lowest value read out after application of the standard curve and use of curve-fitting algorithms. In most instances, the LAL is less than the LDD and the LLOQ. For data mining, individual values below the LAL were replaced with a value that is 50% of the lowest value measured in the data set.

### Statistical Analysis and Data Visualization

Principal component analysis (PCA), agglomerative hierarchical clustering, and ANOVA testing were performed with Qlucore Omics Explorer, v.2.3 (Qlucore). We report ANOVA-based p values, and to correct for multiple testing we report false discovery rate (FDR)-adjusted ANOVA p values, called q values. Dot plot graphs and two-way correlation plots were compiled with GraphPad Prism v.6.0. Correlation matrices and bar graphs were calculated with R v.2.15.1 and drawn with graphical package ggplot2 v.0.9.3.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, four tables, and one data set and can be found with this article online at <http://dx.doi.org/10.1016/j.immuni.2014.03.002>.

## CONSORTIA

*Milieu Intérieur* Consortium team leaders: Laurent Abel, Andres Alcover, Philippe Bouso, Ana Cumano, Marc Daéron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Antonio Freitas, Ivo G. Boneca, Serge Hercberg, Olivier Lantz, Claude Leclerc, Hugo Mouquet, Sandra Pellegrini, Stanislas Pol, Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis, Frédéric Tangy, Eric Tartour, Antoine Toubert, Marie-Noëlle Ungeheuer, Lluís Quintana-Murci, and Matthew L. Albert. Additional information can be found at <http://www.milieuinterieur.fr/en>.

## ACKNOWLEDGMENTS

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). We also acknowledge A. Pugsley for his support in obtaining funding and the logistical aspects of initiating the project.

Received: July 11, 2013

Accepted: January 15, 2014

Published: March 20, 2014

## REFERENCES

- Alexopoulou, L., Holt, A.C., Medzhitov, R., and Flavell, R.A. (2001). Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. *Nature* **413**, 732–738.
- Allen, I.C., Scull, M.A., Moore, C.B., Holl, E.K., McElvania-TeKippe, E., Taxman, D.J., Guthrie, E.H., Pickles, R.J., and Ting, J.P. (2009). The NLRP3 inflammasome mediates in vivo innate immunity to influenza A virus through recognition of viral RNA. *Immunity* **30**, 556–565.
- Andersson, A., Olofsson, T., Lindgren, D., Nilsson, B., Ritz, C., Edén, P., Lassen, C., Råde, J., Fontes, M., Mörsé, H., et al. (2005). Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations. *Proc. Natl. Acad. Sci. USA* **102**, 19069–19074.
- Bekeredjian-Ding, I., Roth, S.J., Gilles, S., Giese, T., Ablasser, A., Hornung, V., Endres, S., and Hartmann, G. (2006). T cell-independent, TLR-induced IL-12p70 production in primary human monocytes. *J. Immunol.* **176**, 7438–7446.
- Blasius, A.L., and Beutler, B. (2010). Intracellular toll-like receptors. *Immunity* **32**, 305–315.
- Borriello, S.P., Hammes, W.P., Holzapfel, W., Marteau, P., Schrezenmeir, J., Vaara, M., and Valtonen, V. (2003). Safety of probiotics that contain lactobacilli or bifidobacteria. *Clin. Infect. Dis.* **36**, 775–780.
- Braga-Neto, U.M., and Marques, E.T., Jr. (2006). From functional genomics to functional immunomics: new challenges, old problems, big rewards. *PLoS Comput. Biol.* **2**, e81.
- Brown, G.D., Herre, J., Williams, D.L., Willment, J.A., Marshall, A.S., and Gordon, S. (2003). Dectin-1 mediates the biological effects of beta-glucans. *J. Exp. Med.* **197**, 1119–1124.

- Ceuppens, J.L., Meurs, L., and Van Wauwe, J.P. (1985). Failure of OKT3 monoclonal antibody to induce lymphocyte mitogenesis: a familial defect in monocyte helper function. *J. Immunol.* *134*, 1498–1502.
- Chamaillard, M., Hashimoto, M., Horie, Y., Masumoto, J., Qiu, S., Saab, L., Ogura, Y., Kawasaki, A., Fukase, K., Kusumoto, S., et al. (2003). An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. *Nat. Immunol.* *4*, 702–707.
- Chen, J., Bruns, A.H., Donnelly, H.K., and Wunderink, R.G. (2010). Comparative in vitro stimulation with lipopolysaccharide to study TNF $\alpha$  gene expression in fresh whole blood, fresh and frozen peripheral blood mononuclear cells. *J. Immunol. Methods* *357*, 33–37.
- Cullen, T.W., Giles, D.K., Wolf, L.N., Ecobichon, C., Boneca, I.G., and Trent, M.S. (2011). *Helicobacter pylori* versus the host: remodeling of the bacterial outer membrane is required for survival in the gastric mucosa. *PLoS Pathog.* *7*, e1002454.
- De Groote, D., Zangerle, P.F., Gevaert, Y., Fassotte, M.F., Beguin, Y., Noizat-Pirenne, F., Pirenne, J., Gathy, R., Lopez, M., Dehart, I., et al. (1992). Direct stimulation of cytokines (IL-1  $\beta$ , TNF- $\alpha$ , IL-6, IL-2, IFN- $\gamma$  and GM-CSF) in whole blood. I. Comparison with isolated PBMC stimulation. *Cytokine* *4*, 239–248.
- Der, S.D., Zhou, A., Williams, B.R., and Silverman, R.H. (1998). Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* *95*, 15623–15628.
- Diebold, S.S., Kaisho, T., Hemmi, H., Akira, S., and Reis e Sousa, C. (2004). Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* *303*, 1529–1531.
- Digel, W., Marcucci, F., and Kirchner, H. (1983). Induction of interferon gamma in leucocyte cultures of the peripheral blood of mice. *J. Interferon Res.* *3*, 65–69.
- Dinareello, C.A. (2009). Immunological and inflammatory functions of the interleukin-1 family. *Annu. Rev. Immunol.* *27*, 519–550.
- Dinareello, C.A. (2012). Keep up the heat on IL-1. *Blood* *120*, 2538–2539.
- Dinareello, C.A., Simon, A., and van der Meer, J.W. (2012). Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. *Nat. Rev. Drug Discov.* *11*, 633–652.
- Eskola, J., Soppi, E., Viljanen, M., and Ruuskanen, O. (1975). A new micro-method for lymphocyte stimulation using whole blood. *Immunol. Commun.* *4*, 297–307.
- Fleischer, B., and Schrezenmeier, H. (1988). T cell stimulation by staphylococcal enterotoxins. Clonally variable response and requirement for major histocompatibility complex class II molecules on accessory or target cells. *J. Exp. Med.* *167*, 1697–1707.
- Folds, J.D., and Schmitz, J.L. (2003). 24. Clinical and laboratory assessment of immunity. *J. Allergy Clin. Immunol. Suppl.* *111*, S702–S711.
- Gantner, B.N., Simmons, R.M., Canavera, S.J., Akira, S., and Underhill, D.M. (2003). Collaborative induction of inflammatory responses by dectin-1 and Toll-like receptor 2. *J. Exp. Med.* *197*, 1107–1117.
- Gewirtz, A.T., Yu, Y., Krishna, U.S., Israel, D.A., Lyons, S.L., and Peek, R.M., Jr. (2004). *Helicobacter pylori* flagellin evades toll-like receptor 5-mediated innate immunity. *J. Infect. Dis.* *189*, 1914–1920.
- Godaly, G., and Young, D.B. (2005). *Mycobacterium bovis* bacille Calmette Guerin infection of human neutrophils induces CXCL8 secretion by MyD88-dependent TLR2 and TLR4 activation. *Cell. Microbiol.* *7*, 591–601.
- González-Navajas, J.M., Lee, J., David, M., and Raz, E. (2012). Immunomodulatory functions of type I interferons. *Nat. Rev. Immunol.* *12*, 125–135.
- Goodridge, H.S., Reyes, C.N., Becker, C.A., Katsumoto, T.R., Ma, J., Wolf, A.J., Bose, N., Chan, A.S., Magee, A.S., Danielson, M.E., et al. (2011). Activation of the innate immune receptor Dectin-1 upon formation of a 'phagocytic synapse'. *Nature* *472*, 471–475.
- Gow, N.A., van de Veerdonk, F.L., Brown, A.J., and Netea, M.G. (2012). *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. *Nat. Rev. Microbiol.* *10*, 112–122.
- Hayashi, F., Smith, K.D., Ozinsky, A., Hawn, T.R., Yi, E.C., Goodlett, D.R., Eng, J.K., Akira, S., Underhill, D.M., and Aderem, A. (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* *410*, 1099–1103.
- Hemmi, H., Takeuchi, O., Kawai, T., Kaisho, T., Sato, S., Sanjo, H., Matsumoto, M., Hoshino, K., Wagner, H., Takeda, K., and Akira, S. (2000). A Toll-like receptor recognizes bacterial DNA. *Nature* *408*, 740–745.
- Hemmi, H., Kaisho, T., Takeuchi, O., Sato, S., Sanjo, H., Hoshino, K., Horiuchi, T., Tomizawa, H., Takeda, K., and Akira, S. (2002). Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nat. Immunol.* *3*, 196–200.
- Huang, Q., Liu, D., Majewski, P., Schulte, L.C., Korn, J.M., Young, R.A., Lander, E.S., and Hacohen, N. (2001). The plasticity of dendritic cell responses to pathogens and their components. *Science* *294*, 870–875.
- Ichinohe, T., Lee, H.K., Ogura, Y., Flavell, R., and Iwasaki, A. (2009). Inflammasome recognition of influenza virus is essential for adaptive immune responses. *J. Exp. Med.* *206*, 79–87.
- Ida, J.A., Shrestha, N., Desai, S., Pahwa, S., Hanekom, W.A., and Haslett, P.A. (2006). A whole blood assay to assess peripheral blood dendritic cell function in response to Toll-like receptor stimulation. *J. Immunol. Methods* *310*, 86–99.
- Józefowski, S., Sobota, A., Pawiowski, A., and Kwiatkowska, K. (2011). *Mycobacterium tuberculosis* lipoarabinomannan enhances LPS-induced TNF- $\alpha$  production and inhibits NO secretion by engaging scavenger receptors. *Microb. Pathog.* *50*, 350–359.
- Jurk, M., Heil, F., Vollmer, J., Schetter, C., Krieg, A.M., Wagner, H., Lipford, G., and Bauer, S. (2002). Human TLR7 or TLR8 independently confer responsiveness to the antiviral compound R-848. *Nat. Immunol.* *3*, 499.
- Kato, H., Sato, S., Yoneyama, M., Yamamoto, M., Uematsu, S., Matsui, K., Tsujimura, T., Takeda, K., Fujita, T., Takeuchi, O., and Akira, S. (2005). Cell type-specific involvement of RIG-I in antiviral response. *Immunity* *23*, 19–28.
- Kato, H., Takeuchi, O., Sato, S., Yoneyama, M., Yamamoto, M., Matsui, K., Uematsu, S., Jung, A., Kawai, T., Ishii, K.J., et al. (2006). Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* *441*, 101–105.
- Kawai, K., Miyazaki, J., Joraku, A., Nishiyama, H., and Akaza, H. (2013). *Bacillus Calmette-Guérin* (BCG) immunotherapy for bladder cancer: current understanding and perspectives on engineered BCG vaccine. *Cancer Sci.* *104*, 22–27.
- Kirchner, H., Kleinicke, C., and Digel, W. (1982). A whole-blood technique for testing production of human interferon by leukocytes. *J. Immunol. Methods* *48*, 213–219.
- Kolb, W.P., and Granger, G.A. (1968). Lymphocyte in vitro cytotoxicity: characterization of human lymphotoxin. *Proc. Natl. Acad. Sci. USA* *61*, 1250–1255.
- Krieg, A.M. (2002). CpG motifs in bacterial DNA and their immune effects. *Annu. Rev. Immunol.* *20*, 709–760.
- Krishna, S., and Miller, L.S. (2012). Host-pathogen interactions between the skin and *Staphylococcus aureus*. *Curr. Opin. Microbiol.* *15*, 28–35.
- Krug, A., Luker, G.D., Barchet, W., Leib, D.A., Akira, S., and Colonna, M. (2004). Herpes simplex virus type 1 activates murine natural interferon-producing cells through toll-like receptor 9. *Blood* *103*, 1433–1437.
- Kwissa, M., Nakaya, H.I., Oluoch, H., and Pulendran, B. (2012). Distinct TLR adjuvants differentially stimulate systemic and local innate immune responses in nonhuman primates. *Blood* *119*, 2044–2055.
- Lee, S.K., Stack, A., Katzwitsch, E., Aizawa, S.I., Suerbaum, S., and Josephans, C. (2003). *Helicobacter pylori* flagellins have very low intrinsic activity to stimulate human gastric epithelial cells via TLR5. *Microbes Infect.* *5*, 1345–1356.
- Liu-Bryan, R., Pritzker, K., Firestein, G.S., and Terkeltaub, R. (2005). TLR2 signaling in chondrocytes drives calcium pyrophosphate dihydrate and monosodium urate crystal-induced nitric oxide generation. *J. Immunol.* *174*, 5016–5023.
- Maecker, H.T., McCoy, J.P., and Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. *Nat. Rev. Immunol.* *12*, 191–200.

- Mäkelä, S.M., Strengell, M., Pietilä, T.E., Osterlund, P., and Julkunen, I. (2009). Multiple signaling pathways contribute to synergistic TLR ligand-dependent cytokine gene expression in human monocyte-derived macrophages and dendritic cells. *J. Leukoc. Biol.* *85*, 664–672.
- March, C.J., Mosley, B., Larsen, A., Cerretti, D.P., Braedt, G., Price, V., Gillis, S., Henney, C.S., Kronheim, S.R., Grabstein, K., et al. (1985). Cloning, sequence and expression of two distinct human interleukin-1 complementary DNAs. *Nature* *315*, 641–647.
- Martino, F., Pétrilli, V., Mayor, A., Tardivel, A., and Tschopp, J. (2006). Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* *440*, 237–241.
- Means, T.K., Wang, S., Lien, E., Yoshimura, A., Golenbock, D.T., and Fenton, M.J. (1999). Human toll-like receptors mediate cellular activation by *Mycobacterium tuberculosis*. *J. Immunol.* *163*, 3920–3927.
- Miettinen, M., Veckman, V., Latvala, S., Sareneva, T., Matikainen, S., and Julkunen, I. (2008). Live *Lactobacillus rhamnosus* and *Streptococcus pyogenes* differentially regulate Toll-like receptor (TLR) gene expression in human primary macrophages. *J. Leukoc. Biol.* *84*, 1092–1100.
- Nerad, J.L., Griffiths, J.K., Van der Meer, J.W., Endres, S., Poutsika, D.D., Keusch, G.T., Bennish, M., Salam, M.A., Dinarello, C.A., and Cannon, J.G. (1992). Interleukin-1 beta (IL-1 beta), IL-1 receptor antagonist, and TNF alpha production in whole blood. *J. Leukoc. Biol.* *52*, 687–692.
- Newport, M.J., Goetghebuer, T., Weiss, H.A., Whittle, H., Siegrist, C.A., and Marchant, A.; MRC Gambia Twin Study Group (2004). Genetic regulation of immune responses to vaccines in early life. *Genes Immun.* *5*, 122–129.
- Norby, E., Kövamees, J., Blixenkrone-Möller, M., Sharma, B., and Orvell, C. (1992). Humanized animal viruses with special reference to the primate adaptation of morbillivirus. *Vet. Microbiol.* *33*, 275–286.
- Okusawa, T., Fujita, M., Nakamura, J., Into, T., Yasuda, M., Yoshimura, A., Hara, Y., Hasebe, A., Golenbock, D.T., Morita, M., et al. (2004). Relationship between structures and biological activities of mycoplasmal diacylated lipopeptides and their recognition by toll-like receptors 2 and 6. *Infect. Immun.* *72*, 1657–1665.
- Oppmann, B., Lesley, R., Blom, B., Timans, J.C., Xu, Y., Hunte, B., Vega, F., Yu, N., Wang, J., Singh, K., et al. (2000). Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* *13*, 715–725.
- Otto, M. (2009). *Staphylococcus epidermidis* – the ‘accidental’ pathogen. *Nat. Rev. Microbiol.* *7*, 555–567.
- Platanias, L.C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* *5*, 375–386.
- Poltorak, A., He, X., Smirnova, I., Liu, M.Y., Van Huffel, C., Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C., et al. (1998). Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science* *282*, 2085–2088.
- Pott, G.B., Chan, E.D., Dinarello, C.A., and Shapiro, L. (2009). Alpha-1-antitrypsin is an endogenous inhibitor of proinflammatory cytokine production in whole blood. *J. Leukoc. Biol.* *85*, 886–895.
- Randhawa, A.K., Shey, M.S., Keyser, A., Peixoto, B., Wells, R.D., de Kock, M., Lerumo, L., Hughes, J., Hussey, G., Hawkrige, A., et al.; South African Tuberculosis Vaccine Initiative Team (2011). Association of human TLR1 and TLR6 deficiency with altered immune responses to BCG vaccination in South African infants. *PLoS Pathog.* *7*, e1002174.
- Romano, M., and Huygen, K. (2012). An update on vaccines for tuberculosis - there is more to it than just waning of BCG efficacy with time. *Expert Opin. Biol. Ther.* *12*, 1601–1610.
- Santin, M., Muñoz, L., and Rigau, D. (2012). Interferon- $\gamma$  release assays for the diagnosis of tuberculosis and tuberculosis infection in HIV-infected adults: a systematic review and meta-analysis. *PLoS ONE* *7*, e32482.
- Saraiva, M., and O’Garra, A. (2010). The regulation of IL-10 production by immune cells. *Nat. Rev. Immunol.* *10*, 170–181.
- Schmolz, M., Hurst, T.L., Bailey, D.M., Powell, J.R., Forsey, R.J., Thompson, J.M., Williams, C., and Pawelec, G. (2004). Validation of a new highly standardized, lab-independent whole-blood leukocyte function assay for clinical trials (ILCS). *Exp. Gerontol.* *39*, 667–671.
- Shibata, K., Hasebe, A., Sasaki, T., and Watanabe, T. (1997). *Mycoplasma salivarium* induces interleukin-6 and interleukin-8 in human gingival fibroblasts. *FEMS Immunol. Med. Microbiol.* *19*, 275–283.
- Shimazu, R., Akashi, S., Ogata, H., Nagai, Y., Fukudome, K., Miyake, K., and Kimoto, M. (1999). MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J. Exp. Med.* *189*, 1777–1782.
- Sieling, P.A., Chatterjee, D., Porcelli, S.A., Prigozy, T.I., Mazzaccaro, R.J., Soriano, T., Bloom, B.R., Brenner, M.B., Kronenberg, M., Brennan, P.J., et al. (1995). CD1-restricted T cell recognition of microbial lipoglycan antigens. *Science* *269*, 227–230.
- Smith-Garvin, J.E., Koretzky, G.A., and Jordan, M.S. (2009). T cell activation. *Annu. Rev. Immunol.* *27*, 591–619.
- Stuyt, R.J., Netea, M.G., Kim, S.H., Novick, D., Rubinstein, M., Kullberg, B.J., Van der Meer, J.W., and Dinarello, C.A. (2001). Differential roles of interleukin-18 (IL-18) and IL12 for induction of gamma interferon by staphylococcal cell wall components and superantigens. *Infect. Immun.* *69*, 5025–5030.
- Stuyt, R.J., Kim, S.H., Reznikov, L.L., Fantuzzi, G., Novick, D., Rubinstein, M., Kullberg, B.J., van der Meer, J.W., Dinarello, C.A., and Netea, M.G. (2003). Regulation of *Staphylococcus epidermidis*-induced IFN-gamma in whole human blood: the role of endogenous IL-18, IL-12, IL-1, and TNF. *Cytokine* *21*, 65–73.
- Sun, X., Zhang, M., El-Zataari, M., Owyang, S.Y., Eaton, K.A., Liu, M., Chang, Y.M., Zou, W., and Kao, J.Y. (2013). TLR2 mediates *Helicobacter pylori*-induced tolerogenic immune response in mice. *PLoS ONE* *8*, e74595.
- Takeda, K., Kaisho, T., and Akira, S. (2003). Toll-like receptors. *Annu. Rev. Immunol.* *21*, 335–376.
- Takeuchi, O., Hoshino, K., Kawai, T., Sanjo, H., Takada, H., Ogawa, T., Takeda, K., and Akira, S. (1999). Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. *Immunity* *11*, 443–451.
- Tax, W.J., Willems, H.W., Reekers, P.P., Capel, P.J., and Koene, R.A. (1983). Polymorphism in mitogenic effect of IgG1 monoclonal antibodies against T3 antigen on human T cells. *Nature* *304*, 445–447.
- Torri, A., Beretta, O., Ranghetti, A., Granucci, F., Ricciardi-Castagnoli, P., and Foti, M. (2010). Gene expression profiles identify inflammatory signatures in dendritic cells. *PLoS ONE* *5*, e9404.
- Viljanen, M.K., Peltola, T., Junnila, S.Y., Olkkonen, L., Järvinen, H., Kuistila, M., and Huovinen, P. (1990). Outbreak of diarrhoea due to *Escherichia coli* O111:B4 in schoolchildren and adults: association of Vi antigen-like reactivity. *Lancet* *336*, 831–834.
- Wroblewski, L.E., Peek, R.M., Jr., and Wilson, K.T. (2010). *Helicobacter pylori* and gastric cancer: factors that modulate disease risk. *Clin. Microbiol. Rev.* *23*, 713–739.
- Yoneyama, M., Kikuchi, M., Matsumoto, K., Imaizumi, T., Miyagishi, M., Taira, K., Foy, E., Loo, Y.M., Gale, M., Jr., Akira, S., et al. (2005). Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J. Immunol.* *175*, 2851–2858.
- Zhao, Y., Yokota, K., Ayada, K., Yamamoto, Y., Okada, T., Shen, L., and Oguma, K. (2007). *Helicobacter pylori* heat-shock protein 60 induces interleukin-8 via a Toll-like receptor (TLR)2 and mitogen-activated protein (MAP) kinase pathway in human monocytes. *J. Med. Microbiol.* *56*, 154–164.
- Zheng, Y., Humphry, M., Maguire, J.J., Bennett, M.R., and Clarke, M.C. (2013). Intracellular interleukin-1 receptor 2 binding prevents cleavage and activity of interleukin-1 $\alpha$ , controlling necrosis-induced sterile inflammation. *Immunity* *38*, 285–295.
- Zielinski, C.E., Mele, F., Aschenbrenner, D., Jarrossay, D., Ronchi, F., Gattorno, M., Monticelli, S., Lanzavecchia, A., and Sallusto, F. (2012). Pathogen-induced human TH17 cells produce IFN- $\gamma$  or IL-10 and are regulated by IL-1 $\beta$ . *Nature* *484*, 514–518.



available at [www.sciencedirect.com](http://www.sciencedirect.com)

Clinical Immunology

[www.elsevier.com/locate/yclim](http://www.elsevier.com/locate/yclim)



# Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping



Milena Hasan<sup>a</sup>, Benoit Beitz<sup>a</sup>, Vincent Rouilly<sup>a,b</sup>, Valentina Libri<sup>a</sup>,  
Alejandra Urrutia<sup>a,c,d</sup>, Darragh Duffy<sup>a,c,d</sup>, Lydie Cassard<sup>a</sup>,  
James P. Di Santo<sup>e</sup>, Estelle Mottez<sup>a</sup>, Lluís Quintana-Murci<sup>f,g</sup>,  
Matthew L. Albert<sup>a,c,d,h,\*</sup>,  
Lars Rogge<sup>a,i,\*\*</sup> for The *Milieu Intérieur* Consortium

<sup>a</sup> Center for Human Immunology, Institut Pasteur, Paris, France

<sup>b</sup> Center for Bioinformatics, Institut Pasteur, Paris, France

<sup>c</sup> INSERM U818, France

<sup>d</sup> Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris, France

<sup>e</sup> Laboratory of Innate Immunity, Department of Immunology, Institut Pasteur, Paris, France

<sup>f</sup> Laboratory of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France

<sup>g</sup> CNRS URA3012, France

<sup>h</sup> INSERM UMS20, France

<sup>i</sup> Laboratory of Immunoregulation, Department of Immunology, Institut Pasteur, Paris, France

Received 4 July 2014; accepted with revision 18 December 2014

Available online 6 January 2015

## KEYWORDS

Immunophenotyping;  
Cytometry;  
Whole blood;

**Abstract** Immunophenotyping by multi-parametric flow cytometry is the cornerstone technology for enumeration and characterization of immune cell populations in health and disease. Standardized procedures are essential to allow for inter-individual comparisons in the context of population based or clinical studies. Herein we report the approach taken by the *Milieu Intérieur* Consortium,

**Abbreviations:** DC, dendritic cell; CM, central memory; CV, coefficient of variation; EM, effector memory; EMRA, effector memory expressing CD45RA; FcR, Fc receptor; FMO, fluorescence minus one; FSC, forward scatter; FVD, Fixable Viability Dye; MFI, mean fluorescence intensity; NK, natural killer cell; PBMC, peripheral blood mononuclear cell; PMN, polymorphonuclear cell; PMT, photomultiplier tube; RBC, red blood cell; RT, room temperature; SOPs, standard operating procedures; SSC, side scatter; T<sub>N</sub>, naïve T cell

\* Correspondence to: M.L. Albert, Unit of Dendritic Cell Immunobiology, Inserm U818, Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 85 48.

\*\* Correspondence to: L. Rogge, Immunoregulation Unit, Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 80 21. E-mail addresses: [albertm@pasteur.fr](mailto:albertm@pasteur.fr) (M.L. Albert), [lars.rogge@pasteur.fr](mailto:lars.rogge@pasteur.fr) (L. Rogge).

<http://dx.doi.org/10.1016/j.clim.2014.12.008>

1521-6616/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Automation;  
Standardization

highlighting the standardized and automated procedures used for immunophenotyping of human whole blood samples. We optimized eight-color antibody panels and procedures for staining and lysis of whole blood samples, and implemented pre-analytic steps with a semi-automated workflow using a robotic system. We report on four panels that were designed to enumerate and phenotype major immune cell populations (PMN, T, B, NK cells, monocytes and DC). This work establishes a foundation for defining reference values in healthy donors. Our approach provides robust protocols for affordable, semi-automated eight-color cytometric immunophenotyping that can be used in population-based studies and clinical trial settings.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Multiparametric flow cytometry is widely used for phenotyping immune cell populations in human blood samples. The abundance of reagents and growing technical innovations in the field of cytometry (e.g., mass cytometry, imaging cytometry and spectral analyzers) has further enhanced the enthusiasm for applying these approaches to the management of patients and the phenotyping of healthy individuals. Flow cytometric techniques have been used for more than two decades in clinical laboratories for the enumeration of CD4<sup>+</sup> and CD8<sup>+</sup> T cells, in the diagnosis of AIDS [1,2], and also in the characterization of lymphoma and leukemic immune cell expansions [3,4]. However, the implementation of standardized procedures within academic research laboratories has recently become a concern for the community, as the absence of such standardization has precluded comparison between studies and experimental settings. Indeed, greater attention is now placed on the requirement for optimized approaches and harmonization of methods [5,6]. Several international initiatives have supported increased standardization of flow cytometry protocols and applications across multiple laboratories that share common scientific or clinical interests. These include the EuroFlow Consortium, which focuses on laboratory procedures for the phenotyping of malignant leukocytes [7,8]; the Human Immunology Project Consortium (HIPC) and European Network for Translational Immunology Research and Education (ENTIRE), which are working together to develop panels for the phenotyping healthy donors [5,6,9]; the ONE study consortium, which is addressing cellular phenotyping in the setting of transplantation [10]; and the Association for Cancer Immunotherapy (CIMT), which have established proficiency panels for different cell populations [11].

The reproducibility of cytometric data depends on five principle criteria: sample type, sample handling, choice of reagents, instrument selection and qualification, and data analysis. In three coordinated reports, we detail the steps that have been taken by the *Milieu Intérieur* Consortium to control for the pre-analytic aspects of cellular phenotyping (reported here), to optimize the analysis of multi-dimensional data [Chen et al. co-submission], which applied together have allowed the characterization of immune phenotype variation in a population of healthy donors [Urrutia et al., in preparation]. Our approach to immune cell phenotyping supports our Consortium's long-term efforts in utilizing cytometric data as a quantitative intermediate phenotype for association studies.

Only with accurate and reproducible methodologies can we begin to establish, integrate and share large data-warehouses of phenotypic and genetic data.

Several prior and ongoing efforts have contributed to the challenge of harmonizing methods in academic research laboratories. Particular attention has been given to sample type, with comparative assessments of fresh or frozen purified peripheral blood mononuclear cells (PBMCs) and whole blood [12–14]. Additional parameters that have been considered include panel design [6,8,10], the use of liquid, lyophilized or freeze-dried reagents [15] and the calibration and settings for the optical bench of multi-laser cytometers that permit longitudinal, multi-user or inter-laboratory standardization [7]. In academic studies, however, there is less attention given to the variability introduced by sample handling. In many instances, sample collection is not proximal to core facilities and despite the use of standard operating procedures (SOPs), studies have not evaluated the impact of manual sample handling on the measured cellular phenotypes such as size, granularity and activation state.

In this report, we detail the steps that were taken to establish a robust protocol for immunophenotyping from 100  $\mu$ l of fresh whole blood, using four eight-color cytometry panels. We present the design of cytometry panels used for phenotyping and quantifying major cell populations present in human blood – T cells, B cells, NK cells, monocytes, dendritic cells, neutrophils, basophils and eosinophils. These data establish the foundation for the analysis of six hundred healthy donors, analyzed over a six-month time interval.

## 2. Materials and methods

### 2.1. Human subject materials, reagents and instrumentation used

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. Samples collected as part of the *Milieu Intérieur* population based study were procured by investigators working at BioTrial, Rennes [Thomas

et al., in preparation], and tracking procedures were established in order to ensure temperature controlled delivery to Institut Pasteur, Paris within 6 h of blood draw (Supplementary Protocol #1). In all cases, whole blood was collected using Li-heparin as an anti-coagulant and maintained at room temperature (18–25°) until processing. The cells were stained using commercially available monoclonal antibodies (Table S1), according to the operating procedure developed as described herein. The standardized protocol is provided (Supplementary Protocol #2). We evaluated three red blood cell lysis reagents to optimize the staining protocol: BD FACS lysing solution (BD Biosciences, ref. 349202), Red Blood Cell lysis (Miltenyi Biosciences, ref. 130-094-183) and RBC lysis buffer (eBioscience Inc., ref. 00-4333-57). Fc-receptor blocking antibodies were used (eBioscience ref. 14-9161). Dead cells were excluded using the Fixable Viability Dye (FVD) eFluor 506 (eBioscience, ref. 65-0866) in the PMN and DC panels. For establishing compensation matrices, Mouse (BD CompBead Set Anti-mouse Ig, k ref. 552843) and Rat (BD CompBead Set Anti-rat/hamster Ig, k ref. 552845) compensation beads were employed. The acquisition of cells was performed using two MACSQuant analyzers (Serial numbers 2420 & 2416), each fit with identical three lasers and ten detector (FSC, SSC and eight fluorochrome channels) optical racks. Calibration of instruments was performed using MacsQuant calibration beads (Miltenyi, ref. 130-093-607). The semi-automated staining was performed using the Evo-150 liquid handling system (Tecan). A detailed script for the semi-automated sample processing is provided in (Online Supplementary Data File #1, <http://www.milieuinterieur.fr/en>).

## 2.2. Staining protocol for cytometric analysis

Whole blood (2 mL) was washed by mixing fresh whole blood and PBS at a 1:1 ratio, followed by centrifugation at 500g for 5 min at 18–22 °C (room temperature). Washed blood and pre-mixed liquid reagents were loaded onto the Freedom Evo 150 liquid handling system. The supernatant was aspirated and discarded, followed by the addition of fresh PBS taking it to the same final volume as input whole blood. Antibody premixes were prepared, shortly spun (about 20 s) and 100 µL of the resuspended cells was aliquoted into tubes containing the pre-mixed antibody cocktail. The samples were shortly vortexed and incubated 20 min in the dark at room temperature (RT). In samples stained with the PMN and DC panels 1 mL of 1x viability dye solution was added, followed by incubation for 30 min in the dark at 4 °C. Thereafter, 1 ml of cold PBS (4 °C) was added to the tubes, which were centrifuged for 5 min at 500g and the supernatant was aspirated. All samples, irrespective of the panel used, were resuspended in 2000 µL of 1x RBC lysing solution, shortly vortexed and incubated 15 min at RT protected from light. After centrifugation for 5 min at 500g, the supernatant was aspirated, the samples were resuspended in 240 µL PBS and immediately acquired on the cytometer.

## 2.3. Data analysis and statistical methods

Flow cytometry data were generated using MACSQuantify™ software version 2.4.1229.1 and saved as .mqd files

(Miltenyi). The files were converted to FCS compatible format and analyzed by FlowJo software version 9.5.3. Statistical graphs were prepared with the R Software version 3.0.1 (Ref.: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>), using the ggplot2 graphical package version 0.9.3.1 [16].

## 3. Results

### 3.1. Panel design

To enable detection, enumeration and phenotyping of major leukocyte populations present in circulation – PMNs, T cell, B cells, NK cells, monocytes and DCs – we designed four 8-color cytometry panels. The “lineage” panel covered the major cell populations, providing a reference for comparison with other consortia and served as an internal control for other panels (Fig. 1A). The “PMN” panel enabled the classification of neutrophils (CD16<sup>+</sup>FcεRIα<sup>-</sup> cells), basophils (FcεRIα<sup>+</sup>CD16<sup>-</sup>) and eosinophils (CDw125<sup>+</sup>) (Fig. 1B). Activation status of neutrophils was assessed by CD62L expression, and used as a marker of healthy donor status. The “T cell” panel was designed to classify CD4<sup>+</sup> and CD8<sup>+</sup> naïve (T<sub>naïve</sub>), central memory (T<sub>CM</sub>), effector memory (T<sub>EM</sub>) and EMRA<sup>+</sup> T cell (T<sub>EMRA</sub>) subsets, utilizing the relative expression levels of CD27, CD45RA and CCR7 (Fig. 1C) [17]. By combining anti-CD8α and anti-CD8β antibodies within the same panel, we were able to distinguish CD8αα, CD8αβ and CD4 CD8αα T cells [18] (Fig. 1C). Information on the activation status of T cells was obtained by surface expression of HLA-DR. The “DC” panel delineates three principle subsets of dendritic cells in peripheral blood: plasmacytoid dendritic cells (pDCs), BDCA-1<sup>+</sup> and BDCA-3<sup>+</sup> conventional dendritic cells (herein referred to as cDC1 and cDC3, respectively) (Fig. 1D).

### 3.2. Selection of reagents and optimization of semi-automated staining procedure

Careful selection of antibody clones and optimal combinations of compatible fluorochromes is key to establish robust flow cytometry panels. We worked with three antibody suppliers (BD Biosciences, eBioscience and Miltenyi), who provided us with several clones and fluorochrome combinations for each antigen of interest. As previously reported by others [6], significant differences were observed between the different reagents despite their targeting the same cell surface protein. Our selection criteria were (i) specificity of the signal, as based on the staining index that is defined as the difference between the positive and the negative populations and the spread of the negative population [19]; (ii) signal resolution; (iii) availability of desired fluorochrome; (iv) fluorochrome stability (tandem dyes); (v) price and availability of single lot of reagents for cohort study; and, when possible, (vi) availability of CE-IVD format. A complete list of tested antibodies and notable observations concerning their staining performance are reported in Table S1. Two examples (anti CD14 and anti CD8β) are shown to illustrate our testing and selection procedures (Figure S1).

A Lineage				B PMN			
fluorochrome	specificity	clone (vendor)	mg/100mL of blood	fluorochrome	specificity	clone (vendor)	mg/100mL of blood
eF450	CD3	SK7 (eBio)	0.03	eF450	CD62L	Dreg 56 (BD)	0.125
V500	CD14	M5E2 (BD)	0.5	eF 506	FVD	— (eBio)	*
FITC	CD56	NCAM16.2 (BD)	0.06	FITC	FCeR1a	AER-37/ CRA1 (eBio)	0.06
PE	CD45	HI30 (BD)	0.13	PE	CDw125	A14 (BD)	0.12
PerCP-Cy5.5	CD16	3G8 (BD)	0.06	PerCP-Cy5.5	CD16	3G8 (BD)	0.002
PE-Cy7	CD8b	SID8BEE (eBio)	0.13	PE-Cy7	CD32	FLI8.26 (BD)	3.06
APC	CD19	SJ25C1 (BD)	0.13	APC	CD203c	FR3-16A11 (Milttenyi)	0.8
APC-H7	CD4	SK3 (BD)	0.06				

C T cell				D DC			
fluorochrome	specificity	clone (vendor)	mg/100mL of blood	fluorochrome	specificity	clone (vendor)	µg/100µL of blood
eF450	CD3	SK7 (eBio)	0.25	VioBlue	CD14	TUK4 (Milttenyi)	1.5
V500	HLA-DR	L243/G46-6 (BD)	0.13	eF 506	FVD	/ (eBio)	*
FITC	CD45RA	L48 (BD)	0.5	VioGreen	CD19	LT19 (Milttenyi)	0.8
PE	CD8a	BW135/80 (Milttenyi)	0.8	VioGreen	CD3	BW264/56 (Milttenyi)	0.2
PerCP-eF710	CD27	O323 (eBio)	0.03	FITC	CD1c	AD5-8E7 (Milttenyi)	0.7
PE-Cy7	CD8b	SID8BEE (eBio)	0.13	PE	CD301	AC144 (Milttenyi)	0.6
APC	CCR7	FR11-11E8 (Milttenyi)	1.5	PerCP	HLA-DR	AC122 (Milttenyi)	1
APC-H7	CD4	SK3 (BD)	0.06	PE-Vio770	CD86	FM95 (Milttenyi)	0.6
				APC	CD304	AD5-17F6 (Milttenyi)	0.5
				APC Vio770	CD141	AD5-14H12 (Milttenyi)	0.5

**Figure 1** Organization of panels for whole blood immunophenotyping. Four eight-color panels were established in order to quantify and characterize the major leukocyte populations in circulation. (A) The lineage panel consisted of markers for T cell, B cell, NK cell and monocyte populations. (B) The polymorphonuclear cells (PMN) permitted classification of neutrophils, basophils and eosinophils. (C) The T cell panel assessed CD4<sup>+</sup> and CD8<sup>+</sup> naive, central memory (T<sub>CM</sub>), effector memory (T<sub>EM</sub>) and effector memory RA<sup>+</sup> (T<sub>EMRA</sub>) subsets. (D) The dendritic cell (DC) panel classified the three major DC subsets – pDCs, cDC1 and cDC3. Selection of fluorochrome, clone, vendor and optimal dilution for 100 µL of whole blood used in the study is indicated. FVD, fixed viability dye, was used at a 1:1000 dilution.

### 3.3. Gating strategies

#### 3.3.1. The lineage panel

For the characterization of major leukocyte populations, we first identified CD45<sup>+</sup> hematopoietic cells, followed by exclusion of doublets (Fig. 2A). Subsequently, B cells were gated as CD19<sup>+</sup>CD16<sup>-</sup>, and T cells were identified as CD19<sup>-</sup> cells followed by CD3<sup>+</sup> staining, then analyzed for the expression of CD4 and CD8 (Fig. 2B). Within the CD3<sup>-</sup> cells, NK cells were identified as CD56<sup>+</sup> and analyzed for their expression of CD16 and CD56. In the population of CD56<sup>-</sup> cells, CD16<sup>hi</sup>SSC<sup>low</sup> cells were selected in order to segregate monocytes from neutrophils. Further gating identified CD14<sup>+</sup>CD16<sup>int</sup> monocytes and CD14<sup>low</sup>CD16<sup>hi</sup> monocytes. Neutrophils were defined as CD16<sup>hi</sup>SSC<sup>hi</sup> (Fig. 2B).

#### 3.3.2. The PMN panel

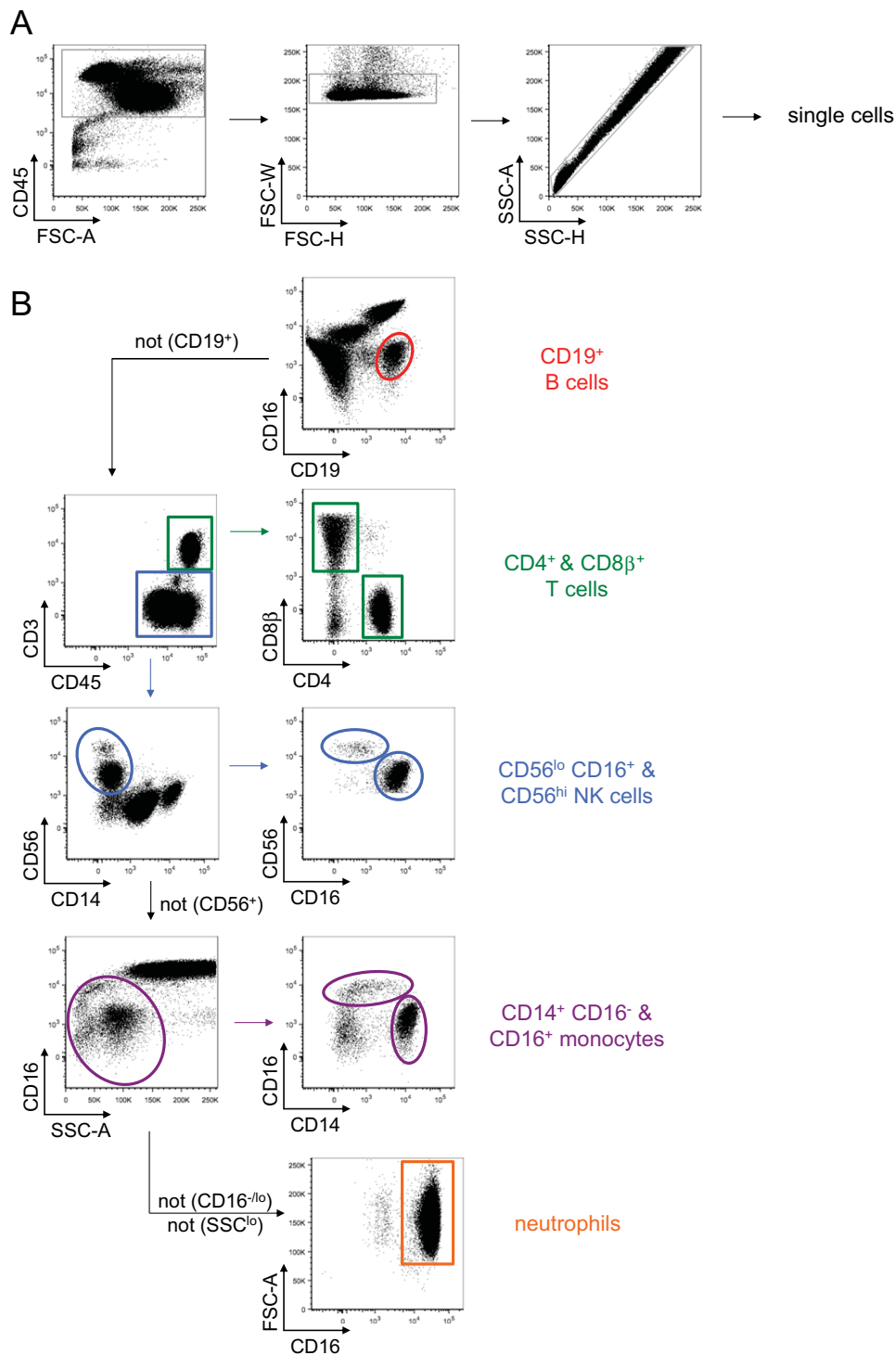
To characterize granulocytes populations, doublets were first excluded (Fig. 3A) and neutrophils were identified as

CD16<sup>hi</sup>CDw125<sup>-</sup> live cells. We also assessed the expression of CD62L within this cell population as a marker of activation (Fig. 3B). Basophils and eosinophils were gated within the CD16<sup>low/-</sup> cells as FcεR1α<sup>+</sup>CD203c<sup>+</sup> and CDw125<sup>+</sup>, respectively (Fig. 3C). Of note, we highlight a difference in the staining of different subpopulations of PMN for Fixable Viability Dye (FVD) (Fig. 3A–C), using saponin treated cells as a positive control for dead cells (Fig. 3D).

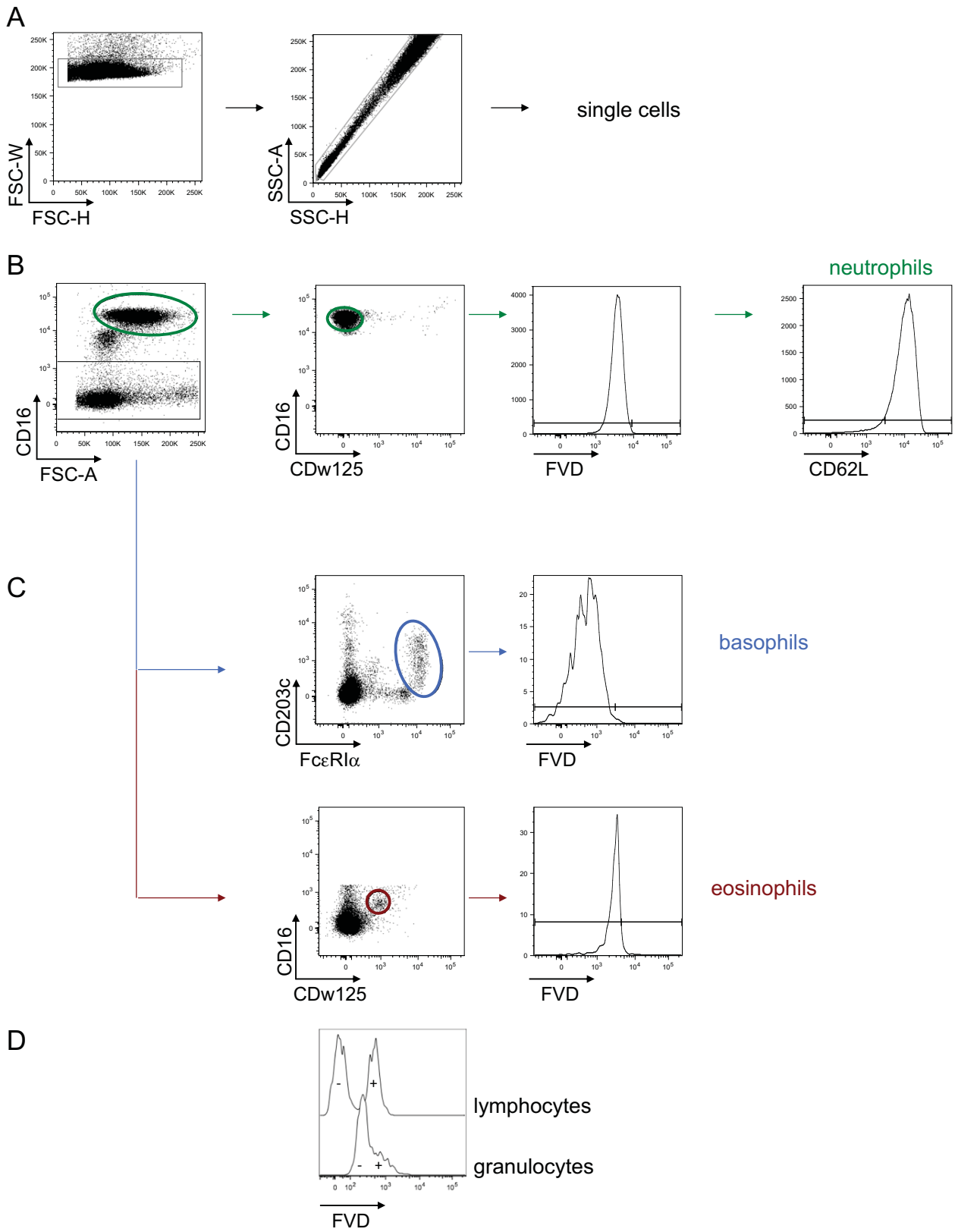
#### 3.3.3. The T cell panel

T cells were identified as CD3<sup>+</sup> cells (Fig. 4A). Upon exclusion of doublets (Fig. 4A), CD4<sup>+</sup> and CD8β<sup>+</sup> were gated and analyzed. We characterized naive (T<sub>N</sub>), central memory (T<sub>CM</sub>), effector memory (T<sub>EM</sub>) and effector memory expressing RA (T<sub>EMRA</sub>) subpopulations of both T cell subsets, based on their expression of CD45RA and CD27 [17,20] (Fig. 4B). T<sub>N</sub> and T<sub>CM</sub> cells have also been defined by the expression of CCR7 [21]. We therefore assessed the expression of CCR7 by these cell populations. The activation status was determined





**Figure 2** Gating strategy for the lineage panel. (A) CD45<sup>+</sup> cells were identified. Doublets were excluded using forward scatter height (FSC-H) and forward scatter width (FSC-W) followed by side scatter height (SSC-H) and side scatter area (SSC-A). (B) Expression of CD19 (red gate) was used to identify B cells. T cells were identified within the CD19<sup>-</sup> cells based on CD3 expression (green gate) and analyzed for the expression of CD4 and CD8β. Within the CD3<sup>-</sup> cells (blue gate), NK cells were identified as CD56<sup>+</sup> and analyzed for their expression of CD16 and CD56. In the population of CD56<sup>-</sup> cells, CD14<sup>+</sup> and CD16<sup>+</sup> monocytes were gated within the CD16<sup>+/low</sup> SSC<sup>low</sup> cells (violet gates). Finally, neutrophils were identified as CD16<sup>+</sup>SSC<sup>hi</sup> (orange gate).



by HLA-DR expression. In addition, the CD4<sup>+</sup> T cell population expressing CD8 $\alpha$  was identified (Fig. 4B).

### 3.3.4. The DC panel

To characterize DCs, we first gated on HLA-DR<sup>+</sup>CD14<sup>-</sup> and excluded dead cell doublets, and CD3<sup>+</sup>, CD19<sup>+</sup> or CD14<sup>+</sup> lineage positive cells using a cocktail of reagents (Fig. 5A). pDC, cDC1 and cDC3 populations were identified as BDCA4<sup>+</sup>BDCA2<sup>+</sup> (CD304<sup>+</sup>CD303<sup>+</sup>), BDCA1<sup>+</sup> (CD1c<sup>+</sup>) and BDCA3<sup>+</sup> (CD141<sup>+</sup>), respectively (Fig. 5B). The activation status of the three DC subsets was assessed by their expression of HLA-DR and the costimulatory molecule CD86 (Fig. 5C). The position of gates to define cDC subsets was determined using HLA-DR<sup>-</sup>CD14<sup>-</sup> cells as a negative control (Fig. 5D).

## 3.4. Standard operating procedures

To optimize the pre-analytical steps of immunophenotyping, we evaluated different conditions for each step of the protocol (Supplementary Protocol #2). We focused in particular on the temperature and duration of blood storage, as well as on red blood cell lysis and the staining protocol. Finally, to make the procedure amenable to a large cohort study, we implemented a semi-automated procedure using liquid handling robotics.

### 3.4.1. Sample

The treatment of blood samples has a large impact on cytometry data [15]. One of the biggest considerations was analysis of fresh blood as compared to freezing samples and batching analysis. To assess potential differences in results obtained by immunophenotyping of fresh whole blood versus thawed PBMC in our experimental conditions, we compared results from eighteen healthy donors. We did not observe a major difference in B cell frequencies (Figure S2A, upper panel) or of CD19 mean fluorescence intensity (MFI) values (Figure S2A, lower panel). Lower frequencies of CD3<sup>+</sup> lymphocytes and a significantly lower expression level of CD3 were observed in frozen/thawed PBMCs compared to whole blood (Figure S2A). While the frequency of CD4<sup>+</sup> and CD8<sup>+</sup> lymphocytes was similar in thawed PBMC and whole blood (Figure S2A, upper panels), we noted higher and more variable MFI values of CD4, and lower expression of CD8 $\beta$  in PBMCs as compared to fresh whole blood samples (Figure S2B, lower panels). Analysis of thawed PBMCs revealed lower CD56 expression, but no alteration in frequencies of NK cells (Figure S2B). Other differences were also noted, again with the observation that freeze/thaw introduced variance, especially in monocyte and DC populations (Figure S2C and S2D, lower panels). Based on these data, and the experience

of other consortia [4,6], we utilized fresh whole blood samples for our study.

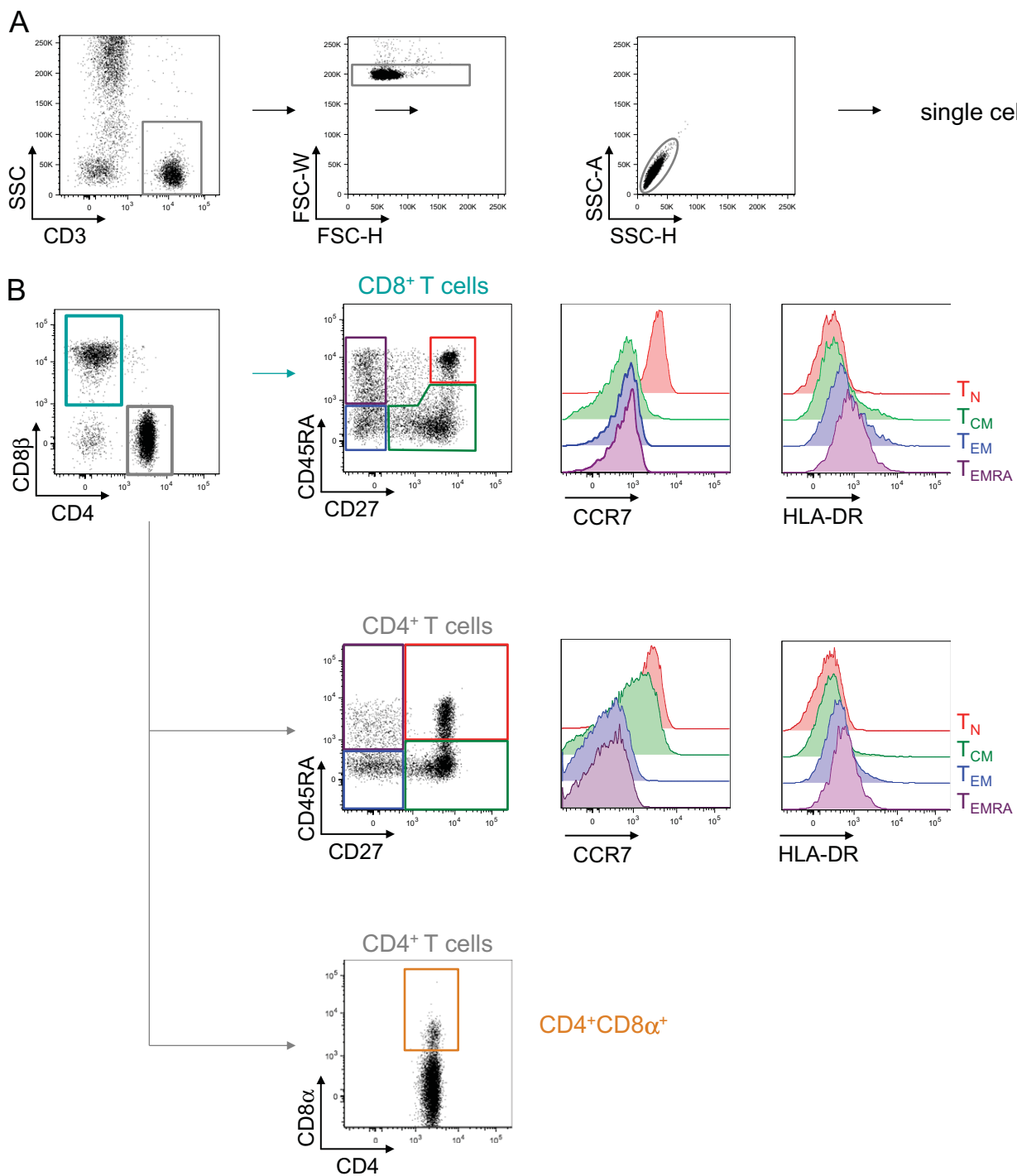
To assess the impact of time between blood draw and staining, we analyzed blood from three healthy donors at four different time-points: immediately after blood draw, and 2 h, 7 h and 24 h post-blood draw. The aliquots were kept at room temperature (18–25 °C) until the analysis. Staining patterns of the analyzed immune cell populations did not change within the first 7 h. However, further delay in time of sample staining and analysis (24 h) had a non-negligible impact on the size and granularity of cells, with an additional population of FCS<sup>high</sup>/SSC<sup>low</sup> cells appearing 24 h after collection (Fig. 6). Furthermore, there was a striking impact on the activation status of dendritic cells after 24 h, as observed by an increased expression of HLA-DR on the surface of cDC3 cells (Fig. 6).

We also evaluated the impact of the time between sample collection and processing on the cell numbers of selected immune populations. No differences were observed in the T cell, B cell, neutrophil or dendritic cell numbers (Fig. 7). The notable exception was the number of neutrophils in one of the three donors. Together, our data, based on cell phenotyping and enumeration studies, clearly showed that 6 h post-blood draw is the maximum permitted delay.

### 3.4.2. Staining protocol

Fresh whole blood samples were washed to eliminate soluble antibodies and other molecules that may interfere with staining. The duration of antibody incubation and staining temperature was evaluated (data not shown). Since blocking of FcR did not have a significant impact on the results (data not shown), it was not included in the staining protocol. The staining was followed by red blood cell lysis. We tested three red blood cell lysis solutions and identified the BD solution to be the most efficient, with reagents from other suppliers being either less efficient (Miltenyi) or slower in achieving RBC lysis (eBioscience) (data not shown). Additionally, we tested different criteria that impact lysis conditions and the reproducibility of the results: duration of incubation in lysis solution, use of mixing and the implementation of a wash step. We highlight that an additional wash step was introduced after red blood cell lysis to ensure complete elimination of the lysis solution, which also showed better preservation of size and granularity characteristics of leukocytes. The staining protocol, established for immunophenotyping of 100  $\mu$ l of fresh whole blood, is detailed in Supplementary Protocol #2. All tested antibodies were titrated to fit the experimental conditions described in the protocol. To minimize variation of fluorescent signal intensity, only one lot of each antibody was used for staining throughout the whole study.

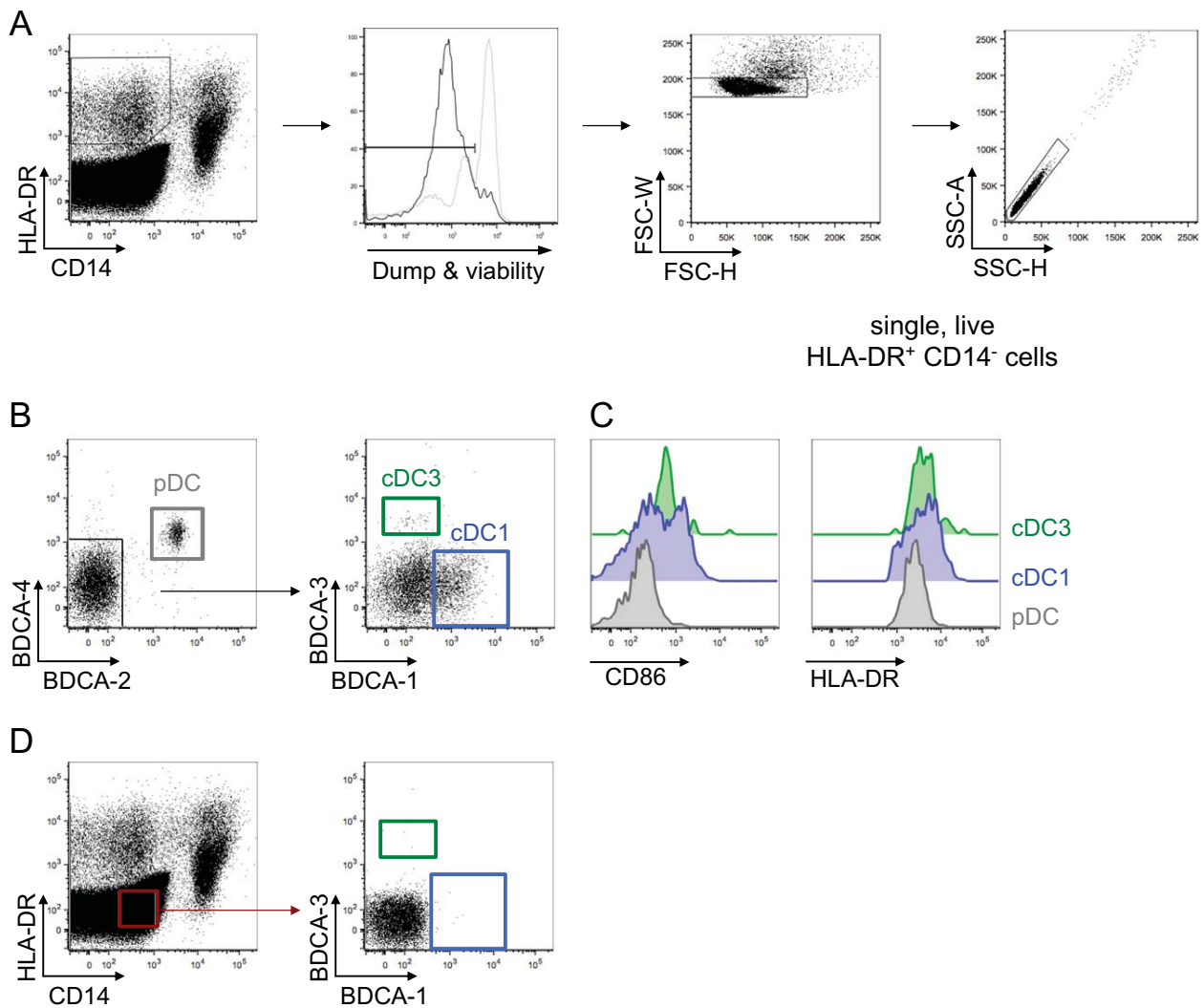
**Figure 3** Gating strategy for the PMN panel. (A) Doublets were excluded from the analysis using FSC-W/FSC-H and SSC-A/SSC-H parameters. (B) Neutrophils were identified based on their high expression of CD16 (green gate), with hierarchical gating to select cells with low levels of CDw125 expression, low levels of FVD and high expression of CD62L. (C) CD16<sup>low/-</sup> cells were independently valuated for high Fc $\epsilon$ RI $\alpha$  expression and intermediate CD203c expression (blue gate), a phenotype characteristic of basophils; or intermediate expression of CDw125 and CD16 (red gate), hallmarks of eosinophils. (D) Whole blood was incubated with 1% saponin for 1 min, washed with PBS and stained with FVD. Granulocytes and lymphocytes were gated based on their size and granularity. The different levels of auto-fluorescence of lymphocytes and granulocytes used to set the gates in (B) and (C) are shown.



**Figure 4** Gating strategy for the identification of T cell populations. (A) CD3<sup>+</sup> cells were identified. Doublets were excluded using FSC-W/FSC-H and SSC-A/SSC-H parameters. (B) Subsequent phenotypic analysis identified CD4<sup>+</sup> and CD8<sup>+</sup> T cell subsets based on CD4 and CD8β expression, respectively. T<sub>Naive</sub> (CD27<sup>+</sup>CD45RA<sup>+</sup> cells, and red shaded histograms), T<sub>CM</sub> (CD27<sup>+</sup>CD45RA<sup>-</sup> cells, and green shaded histograms), T<sub>EM</sub> (CD27<sup>-</sup>CD45RA<sup>-</sup> cells, and blue shaded histograms) and T<sub>EMRA</sub> (CD27<sup>-</sup>CD45RA<sup>+</sup> cells, and violet shaded histograms) were based on surface expression of CD27 and CD45RA. On each of the eight respective populations, CCR7 and HLA-DR expression was analyzed and plotted as a histogram. (C) The CD4<sup>+</sup>CD8β<sup>-</sup> cells expressing CD8α were identified (orange gate).

While we were finalizing our study a comprehensive report by Kalina et al. described the efforts of the Euroflow Consortium to standardize cytometry protocols. We noted that our

independently established procedures were very similar to the ones described by Kalina et al. and further emphasized the importance of reagent selection and staining conditions [7].



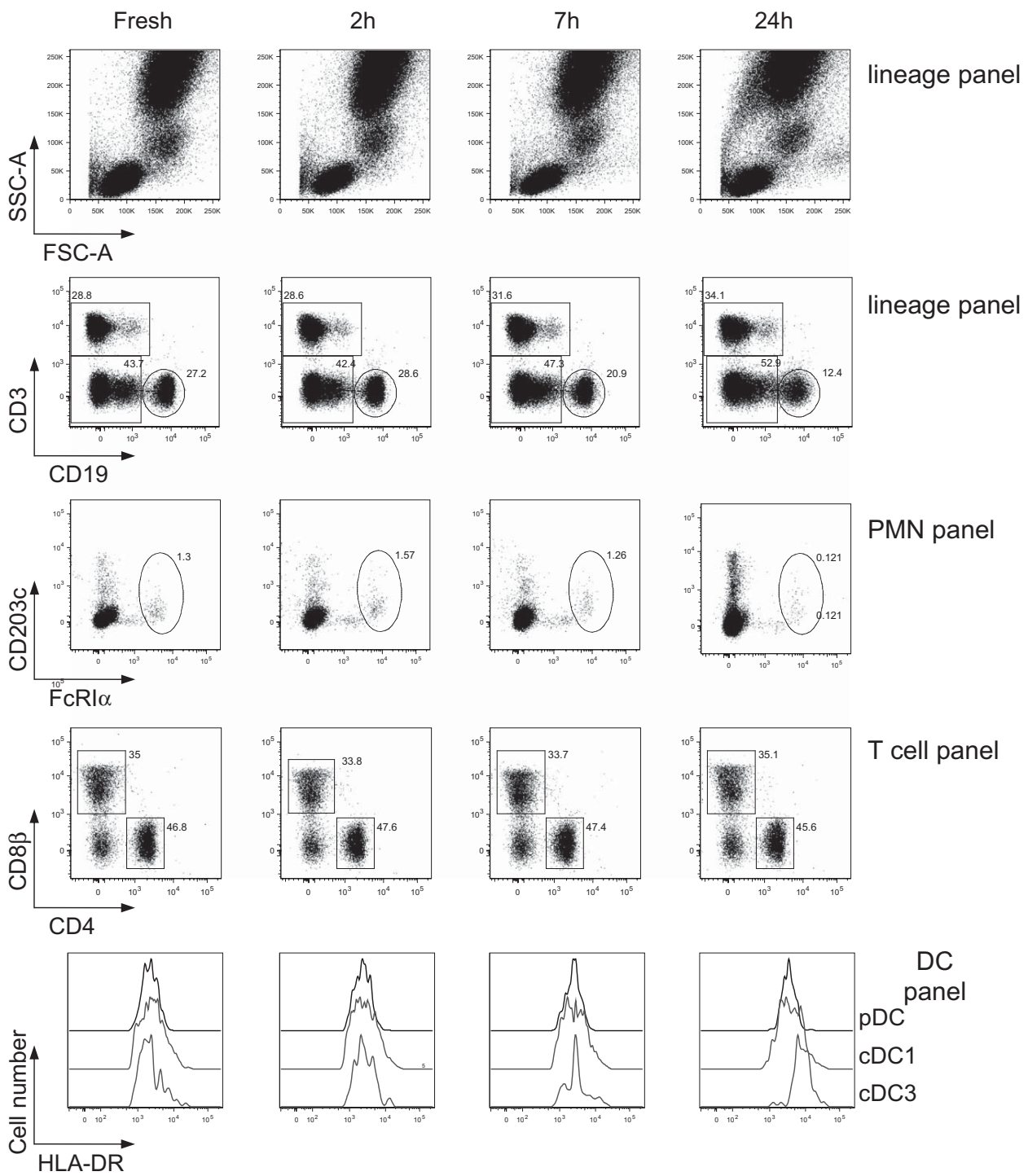
**Figure 5** Gating strategy to identify dendritic cell populations. (A) HLA-DR<sup>+</sup>CD14<sup>-</sup> cells were identified and lineage-marker expressing cells and dead cells were excluded, followed by selection of single cells based upon FSC and SSC properties. (B) pDCs were identified by co-staining with BDCA-2 and BDCA-4 (gray gate). (C) BDCA2<sup>-</sup>BDCA4<sup>-</sup> cells were further characterized as cDC1 and cDC3 based on their expression of BDCA1 (blue gate) or BDCA3 (green gate), respectively. (C) All three DC subsets were evaluated for their expression of CD86 and HLA-DR with comparative histograms shown for the respective subsets. (D) HLA-DR<sup>-</sup>CD14<sup>-</sup> cells were used as an internal control for establishing the MFI cut-off values used for identification of the cDC1 subset.

### 3.4.3. Automation

All clinical laboratory tests use automation in sample processing and attempts have been made to implement automation in genomic assays (DNA/RNA extractions, genotyping, microarray assays, etc.). We decided to take advantage of automation in sample preparation for cellular immunophenotyping. To achieve this, we implemented our protocol using the EVO150 liquid handling platform (Tecan). The premix of antibodies was prepared manually on a daily basis, and all other steps for the staining protocol were performed using the liquid handling platform, with the exception of centrifugation. The pipetting scripts for the platform were created to enable staining of 4 to 12 samples, in parallel, in 96-deep well plates (Online Supplementary Data File #1, <http://www.milieuinterieur.fr/en>).

### 3.4.4. Setting of pre-acquisition parameters

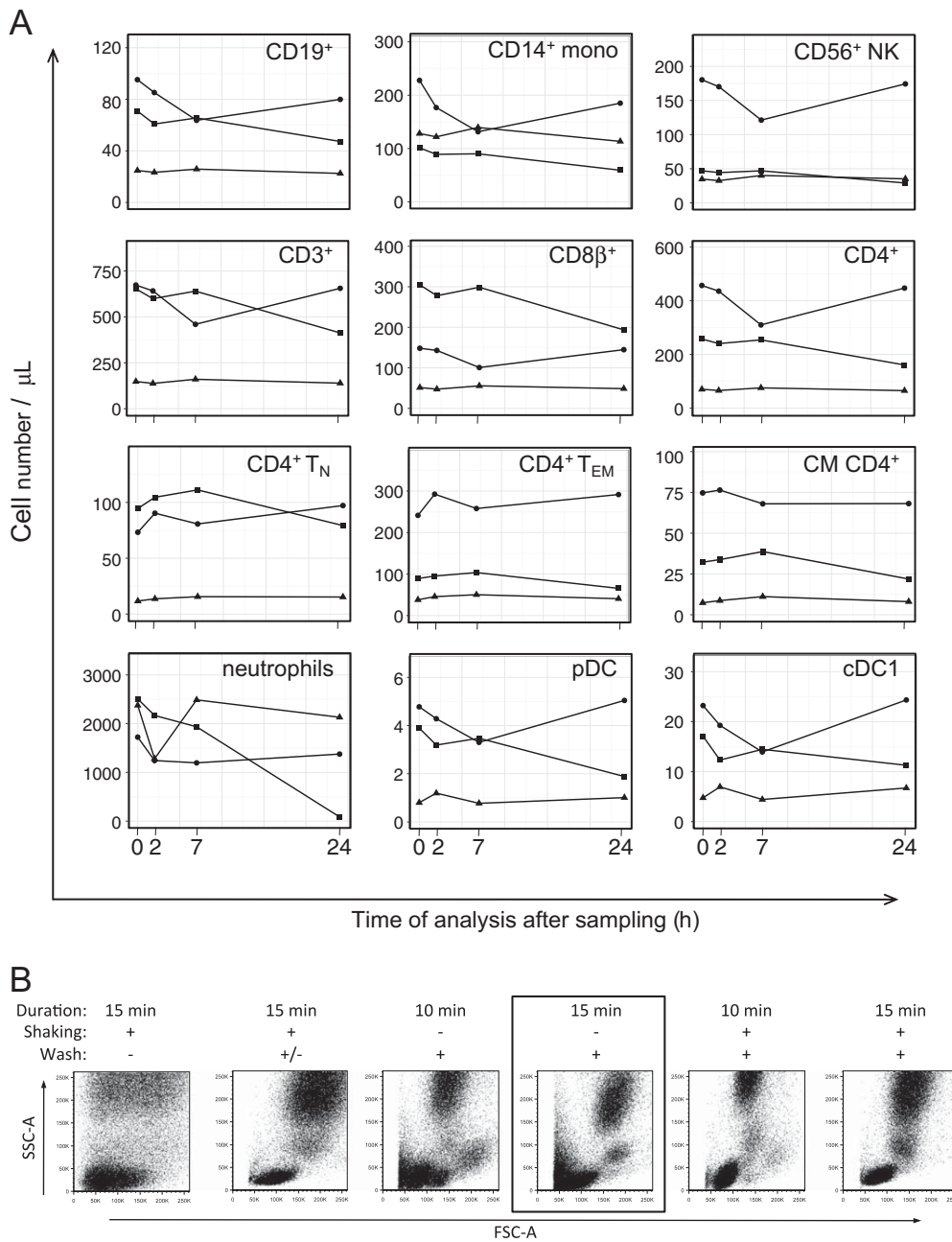
Our study complies with the MIFlowCyt requirements [22]. Dead cells were excluded using FVD in the PMN and DC panels, in which either rare populations needed to be identified, or a high autofluorescence of target cells was expected. A dump channel was included in the DC panel to exclude CD19<sup>+</sup>, CD3<sup>+</sup>, NKp46<sup>+</sup>, or CD14<sup>+</sup> cells to further improve specificity, and in accordance to prior studies [23]. For antibodies that showed weak signals and did not enable clear separation of positive from negative cell populations (e.g., CCR7), we used FMO (fluorescence minus one) staining to set the positive/negative cell gates. During the first two months of the study, compensation controls were run every day, using automatic hardware compensation on MACSQuant. The compensation beads were used to calculate the



**Figure 6** Optimization of sample processing and staining protocol. Whole blood was collected and stored at room temperature (18–25 °C) for 0 h, 2 h, 7 h or 24 h prior to the staining. Representative results of three independent experiments are shown for selected cell populations. The dot plots illustrate the impact of staining at different time-points after sampling of blood based on FSC/SSC characteristics and expression of selected cell surface markers, and on the MFI of the HLA-DR expressed on pDCs, cDC1 and cDC3.

compensation matrix for all antibodies, except for those labeled with Horizon V500 (BD) and dead cell marker (FVD eF506, eBio). For these reagents, cells were used as recommended by the suppliers. Consistent with EuroFlow consortium results [7], our compensation matrices did not

change (PMT values varied  $\leq \pm 5$  V, data not shown). We thus decided to run compensation controls bi-weekly, unless the PMT voltage values reported by the cytometer after the daily set-up varied for  $> \pm 5$  V from the values obtained during the prior compensation run. In order to control for



**Figure 7** Optimization of sample processing and staining protocol. (A) Blood was stained as described in Fig. 6 and the impact of delayed staining is represented for samples obtained from three donors. Individual donors are indicated by symbols (filled triangle, square, circle). (B) Lysis procedures were evaluated using 100  $\mu\text{l}$  of whole blood, mixed and incubated as indicated. The cells were acquired and their size and granularity characteristics were analyzed based on FCS and SSC parameters. The conditions selected for the study are highlighted by a black square.

cross-contamination, samples were plated with PBS filling every other well in the 96-well plates. None of the negative wells showed positive cells for any of the four panels (not shown).

### 3.4.5. Cytometer

Initial testing of panels was performed using an LSRII cytometer (BD) equipped with 4 lasers (488 nm, 405 nm and 630 nm and 658 nm). The design of our study (four 8-color

cytometry panels) required an instrument that enables automatic acquisition from 96-well plates, the acquisition of absolute cell counts and at least 3-lasers. We compared two cytometers that corresponded to these criteria and that were commercially available at the beginning of the study (initiated in 2012): the BD FACSVerser and Miltenyi MACSQuant. The cytometers were evaluated based on their hardware, sample acquisition performance, software robustness, quality control and post-installation support provided by the suppliers

**Table 1** Test of repeatability – technical replicates\*.

Cell type	Parameter	Stat	Panel	Donor					
				1	2	3	4	5	6
T cells	CD3+	Median (CV)	T	1019 (9.2)	527 (14.4)	271 (9.2)	537 (5.8)	736 (13.0)	623 (4.1)
			Lin	1020 (9.8)	530 (9.0)	260 (4.0)	536 (4.2)	694 (9.7)	603 (3.9)
	CD8+	Median (CV)	T	165 (9.0)	127 (14.7)	66 (8.1)	194 (6.2)	149 (14.9)	151 (5.6)
			Lin	175 (9.8)	135 (10.1)	63 (3.8)	209 (3.9)	157 (9.8)	160 (4.0)
Monocytes	CD4+	Median (CV)	T	768 (14.6)	359 (14.6)	170 (9.5)	768 (5.7)	516 (14.3)	347 (4.1)
			Lin	757 (9.6)	357 (8.3)	161 (4.1)	284 (4.5)	473 (9.6)	413 (3.9)
	CD14+	Median (CV)	Lin	334 (10.0)	96 (16.1)	216 (02.7)	220 (3.5)	284 (11.3)	149 (7.8)
			DC	283 (06.4)	85 (16.3)	228 (10.5)	227 (7.6)	154 (22.9)	144 (7.4)
DC	pDC	Median (CV)	DC	3 (09.9)	2 (15.2)	2 (8.7)	5 (11.2)	4 (21.5)	2 (11.8)
NK cells	CD56+	Median (CV)	Lin	152 (11.1)	95 (10.9)	96 (5.1)	122 (04.2)	142 (10.4)	100 (03.4)
Granulocytes	Neutrophils	Median (CV)	Lin	2040 (4.4)	650 (11.1)	797 (6.2)	1060 (7.1)	1770 (06.0)	1590 (5.8)
			PMN	2652 (6.3)	1247 (08.4)	1135 (3.0)	1643 (5.5)	2052 (10.3)	1802 (7.4)

\* Fresh blood samples from six healthy donors were divided in five aliquots each and immediately stained. Shown are absolute cell numbers (median value) of selected cell populations per 1  $\mu$ l of blood. For each immune cell population identified, the intra-panel coefficient of variation (CV) was calculated. In instances where cell populations could be identified by two different panels, both data are reported. The cell numbers of each indicated cell population obtained by PMN, T and DC panels were calculated upon normalization of the total cell number obtained by each panel to the total cell number as quantified by the lineage (Lin) panel.

(Supplementary Table 2). Although the hardware characteristics, quality control and the concept of the software were comparable between the two systems, MACSQuant showed significant advantages concerning the software robustness and performance. In addition, its cell counting feature was reproducible and easy to assess. The MACSQuant's SOP requires a daily quality control check using control beads to monitor the performance of instrument. The cytometer enables identification of "bank settings" for up to 5 different panels. These settings store the target values of MFI for five combinations of up to eight antibodies/fluorochromes (5 eight-color panels). On each occasion that the control beads are run the system automatically sets the PMT values so that the intensity of signal matches predefined MFI values, and thus overcomes fluctuations or decrease of laser power ensuring data reproducibility. Two MACSQuant cytometers were installed with a distinct bank settings programmed on the instruments for the panels used in the study.

### 3.5. Data analysis

We selected the FlowJo software to analyze data. In order to improve standardization of analysis, we created analysis

templates for each panel. A template consists of the gating strategy specific for the given panel, including a pre-defined table with parameters selected for statistical analysis. Magnetic gates were applied for the brightest and most clearly defined antigens to minimize bias introduced by manual repositioning of gates. Identical gate coordinates were selected to gate on the same cell populations across the four panels. The results obtained for each of the samples were verified by an operator prior to final validation. In order to minimize bias introduced by subjective analysis by different individuals, a given panel was analyzed by the same individual for all samples. The statistical parameters selected for the analysis included absolute cell number of each cell population of interest, its percentage in respect to relevant parent populations, gate coordinates to monitor fluorescence intensity and spread of fluorescent signal, and MFI values for cell populations in which activation markers were included (e.g., HLA-DR, CD86).

The bank settings programmed on the selected cytometers enable standardization of fluorescent signal, but do not enable standardization of cell size or granularity, features that vary based on sample handling (Supplementary Fig. 3). In order to take advantage of the standardization procedures of the newest cytometers, we omitted the

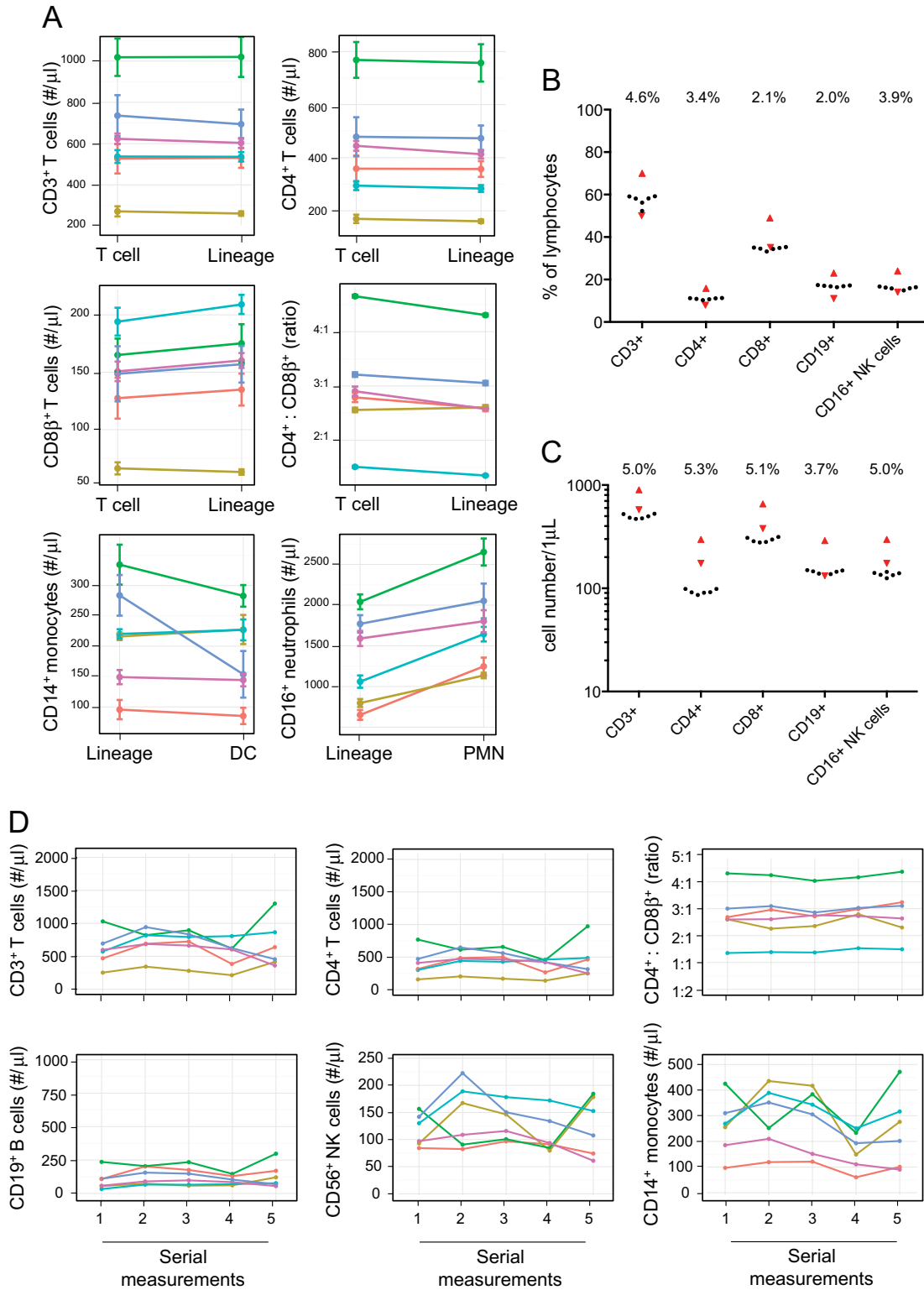
**Figure 8** Repeatability studies, longitudinal studies and panel-to-panel correlations illustrate assay stability. (A) Fresh blood samples from six individual donors were divided in five aliquots for repeatability study each and immediately stained using defined semi-automated procedures. The immune cell populations were identified and absolute cell numbers obtained using above-indicated gating strategies. The graphs show inter-panel comparison of the indicated cell populations as obtained by two different panels (lineage vs. DC panel, lineage vs. T cell and lineage vs. PMN panel). Each dot represents the median value of the five replicate tests for a given donor. Data from individual donors are represented by distinct colors. Standard deviation (SD) between the median values obtained by two panels is indicated. (B, C) Stabilized human blood samples (Eurocell) were analyzed in six independent experiments. Cells were stained using the lineage panel cocktail and labeled by the semi-automated procedure. The percentages of lymphocytes (B) and absolute cell counts per  $\mu$ l of blood (C) are shown for the indicated cell subsets. The target value range determined by the manufacturer on BD and Beckman Coulter flow cytometers is indicated by the red triangles. The CVs for serial measurements are indicated for each analyzed immune cell population. (D) Whole blood samples from six individual donors were collected and stained by the lineage panel cocktail using the semi-automated procedure. Serial measurements were performed at five different time-points over five months. Each dot represents the absolute cell number of indicated cell types during serial measurements. Individual donors are represented by distinct colors.



typical starting gates of FSC/SSC, and directly queried fluorescent markers, using FSC/SSC when needed to exclude doublets or dead cells (Figs. 2–5). The FlowJo templates created for and used in the study are provided (Online Supplementary Data File #2, <http://www.milieuinterieur.fr/en>).

### 3.6. Assay validation: technical replicates and robustness of the staining procedures

To define the variance in our immunophenotyping procedures, we performed experiments to assess repeatability



and reproducibility. To assess repeatability, we analyzed the same sample, in independent runs, by a single operator and run on a single cytometer. In our experimental setting, fresh blood samples from six healthy donors were separated into five aliquots and stained using each of the four panels and run on the liquid handling platform, followed by acquisition on the corresponding MACSQuant cytometer. The results were highly reproducible, with intra-panel CVs below 15% for most of the analyzed cell subsets, irrespective of their absolute counts (Table 1). Higher CVs were observed for pDCs in one of the six donors (21.5%) and for CD14<sup>+</sup> monocytes in two panels in two donors (16% and 23%). The higher variance observed for monocytes may result from the fact that CD14 can be expressed on other cell types, such as neutrophils [24].

Most of the immune cell populations (CD3<sup>+</sup>, CD4<sup>+</sup>, CD8<sup>+</sup> T cells, CD14<sup>+</sup> monocytes) were assessed independently using two different panels, and as such cross-panel comparisons were possible (Table 1 and Fig. 8A). We normalized the total cell counts of PMN, DC and T cell panels to those obtained by the lineage panel because it identifies all “core” cell populations and does not include an additional washing step that was used during the staining of dead cells in PMN and DC panels.

To assess reproducibility, we evaluated the stability of staining over time, an important consideration for large cohort studies. To provide a stable reference, we utilized commercially available stabilized blood, analyzed over a period of one month. These data showed reproducible results with CVs in the range of 2.0–5.3% (Fig. 8B and C). The percentage of each analyzed cell population was within the range of values indicated by the manufacturer (Fig. 8B). The total cell numbers obtained, however, were under the expected value, and consequently so were the total numbers of each analyzed cell populations (Fig. 8C). This difference may be explained by additional washing steps included in our protocol, which are not used to set the reference values indicated by the supplier of the stabilized blood. Another factor contributing to the difference may be the utilization of different cytometric platforms or the use of beads for cell enumeration, as compared to a volume based calibrator built into the cytometer itself. Finally, we analyzed blood samples from six donors across five different time points. As shown in Fig. 8D, the counts of most cell populations were stable over time. The differences observed for NK and monocyte cell number may be due to biological variation and intra-individual variance can be factored into the interpretation of our future population-based results.

#### 4. Discussion

Delivering on the promise of personalized medicine requires tools and techniques that allow both robust and reliable assessment of the immune status of individuals and comparisons between studies. Specifically, the adoption of universal, robust cytometric protocols will allow cross-population comparisons and the evaluation of the extent to which the proportion of different cell populations in patients presenting immunopathology deviate from “healthy” expectations. Flow cytometry is likely to play a key role due to recent technological advances in instrument design, and the availability of a large arsenal of reagents targeting specific molecules. Indeed, these two factors now permit low-cost,

real-time and deep phenotyping of immune cell populations. A notable concern for comparative studies is the pre-analytic variation (sample processing, reagent selection, and instrument parameters). Although several international consortia have begun to tackle this issue, additional efforts need to be taken in order to establish flow cytometry as a tool applicable in routine clinical laboratories.

We report our advances in standardizing pre-analytic procedures for flow cytometry for the *Milieu Intérieur* Consortium, a single-center study aiming to defining reference values of immune parameters in healthy individuals. Our challenge was to establish a standardized procedure for flow cytometry allowing the analysis of 15,000 samples by one operator in a single center. In this context, we considered automation as a solution to facilitate the workflow and to standardize the pre-analytic procedures for flow cytometry. Since clinical studies with large sample numbers involve repetitive work, implementation of automated procedures also eliminates possible error or variation caused by fatigued technical personnel. Perhaps most importantly, an automated procedure allows full traceability at each step (e.g., distribution of antibodies, wash solutions). This is of particular importance if flow cytometry is to be used in a quality-controlled environment, such as a clinical laboratory.

Our work revealed that implementation of automated procedures for flow cytometry is time-consuming and requires extensive testing. Automation also inevitably drives up costs because of expensive consumables and increased reagent use (e.g., dead-volumes in the robotic system). Although the robotic system used in our study operated without technical problems, we acknowledge that a certain amount of daily and weekly maintenance was required. The investment, however, was considered worthwhile based on the quality of the data obtained and it is our hope that others can benefit from the standardization of pre-analytic approaches for sample handling. We have provided complete access to our scripts, and encourage other Consortia to make such procedures available to the community in order to facilitate future improvements in the standardization of flow cytometry procedures.

Our study was inspired by the work of H. Maecker and the FOCIS Human Immunophenotyping Consortium, who identified technical variables in flow cytometric procedures requiring standardization [5,6]. Here and in the accompanying article [Chen et al., co-submission], we present our efforts to optimize staining procedures, selection of reagents, instrument set-up and data analysis. Based on the results of extensive antibody testing (Table S1) we selected reagents from different suppliers and thus have opted not to use the preconfigured lyophilized reagents, as suggested by Maecker et al. [6,15]. To rank the performance of the antibodies of the same specificity, we assessed the staining index and the fluorochrome stability. We observed a significant difference in the performance of different clones from different suppliers, results that reinforce the need to follow MIATA guidelines [25].

Studies involving large-scale or longitudinal immunophenotyping projects analyze either fresh whole blood samples directly at the recruitment site, or PBMCs that are separated from whole blood, frozen and shipped to the analysis laboratory for centralized analysis. Both approaches have advantages and inconveniences. Several studies have

demonstrated that the Ficoll purification can alter the composition and frequency of leucocyte population and the expression of certain surface markers [6,15,26]. The advantage of analyzing frozen PBMCs is the possibility to perform the phenotyping of the entire cohort in the centralized laboratory, eliminating possible errors resulting from all preanalytical steps (reagent preparation, fluorochrome stability, staining protocol, instrument set-up and performance, etc.). We compared these two approaches by analyzing fresh whole blood and comparing it to a portion of the sample that was used for PBMC isolation, frozen and later thawed for comparative analysis. Although several markers showed no observable differences (cell number and MFI), a considerable number of cell surface molecules were affected by the isolation and freezing/thawing procedures. In line with other studies, we support the use of whole blood for immune phenotyping studies when possible.

Our gating strategy builds on the characteristics of the new generation of cytometers that allow precise standardization of the fluorescent signals (voltage-dependent settings). The FSC/SSC gates that are commonly used as a first step in gating strategy were omitted due to variance that could not be controlled. This approach facilitated the rational setting of gates and permitted batch-analysis using FlowJo. As a result of these efforts, only minimal gate positions adjustment was required for a small number of cell populations. Additional standardization of post-analytic gating procedures is addressed in an accompanying manuscript [Chen et al., co-submission].

In conclusion, our efforts are in line with several international consortia, with high coherence in the staining protocols reported by the EuroCell Consortium [7]. These independent approaches converge on the use of whole blood and not frozen PBMCs, procedures for sample handling, and criteria for reagent selection. The new generation of flow cytometers, if properly set up and calibrated, allow precise standardization of fluorescent signals, thus enabling reliable results in longitudinal studies. We believe that the approach and protocols described here provide a rational basis to establish internationally standard operating procedures for immunophenotyping. This attention to standardized cytometric analysis is of paramount importance and will enable inter-institutional comparative studies in healthy and diseased populations.

## Conflict of interest statement

The author(s) declare that there are no conflicts of interest.

## Consortia

The *Milieu Intérieur* Consortium<sup>1</sup> is composed of the following team leaders: Laurent Abel (Hôpital Necker), Andres Alcover, Philippe Bousso, Pierre Bruhns, Ana Cumano, Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Antonio Freitas, Odile Gelpi, Ivo Gomperts-Boneca, Serge Hercberg (Université Paris 13), Olivier Lantz (Institut Curie), Claude

Leclerc, Hugo Mouquet, Sandra Pellegrini, Stanislas Pol (Hôpital Côchin), Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis (Institut Curie), Frédéric Tangy, Eric Tartour (Hôpital Européen George Pompidou), Antoine Toubert (Hôpital Saint-Louis), Marie-Noëlle Ungeheuer, Lluís Quintana-Murci<sup>2</sup>, and Matthew L. Albert<sup>2</sup>.

Additional information can be found at: <http://www.pasteur.fr/labex/milieu-interieur>.

## Acknowledgments

We thank Olivier Mary for organizing sample transport and Christophe Rousseau for purchasing. We thank BD, eBio, Miltenyi for providing the antibodies for testing, Traci Zell for her advice in the choice of fluorochromes in panel design and Cedric Ait Mansour for his help in the cytometer set-up. 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).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.clim.2014.12.008>.

## References

- [1] J.L. Fahey, H. Prince, M. Weaver, J. Groopman, B. Visscher, K. Schwartz, R. Detels, Quantitative changes in T helper or T suppressor/cytotoxic lymphocyte subsets that distinguish acquired immune deficiency syndrome from other immune subset disorders, *Am. J. Med.* 76 (1984) 95–100.
- [2] H. Kornfeld, R.A. Vande Stouwe, M. Lange, M.M. Reddy, M.H. Grieco, T-lymphocyte subpopulations in homosexual men, *N. Engl. J. Med.* 307 (1982) 729–731.
- [3] R.E. Duque, Flow cytometric analysis of lymphomas and acute leukemias, *Ann. N. Y. Acad. Sci.* 677 (1993) 309–325.
- [4] J.J. van Dongen, A. Orfao, C. EuroFlow, EuroFlow: resetting leukemia and lymphoma immunophenotyping. Basis for companion diagnostics and personalized medicine, *Leukemia* 26 (2012) 1899–1907.
- [5] H.T. Maecker, J.P. McCoy Jr., F.H.I. Consortium, M. Amos, J. Elliott, A. Gaigalas, L. Wang, R. Aranda, J. Banchereau, C. Boshoff, J. Braun, Y. Korin, E. Reed, J. Cho, D. Hafler, M. Davis, C.G. Fathman, W. Robinson, T. Denny, K. Weinhold, B. Desai, B. Diamond, P. Gregersen, P. Di Meglio, F.O. Nestle, M. Peakman, F. Villanova, J. Ferbas, E. Field, A. Kantor, T. Kawabata, W. Komocsar, M. Lotze, J. Nepom, H. Ochs, R. O'Lone, D. Phippard, S. Plevy, S. Rich, M. Roederer, D. Rotrosen, J.H. Yeh, A model for harmonizing flow cytometry in clinical trials, *Nat. Immunol.* 11 (2010) 975–978.
- [6] H.T. Maecker, J.P. McCoy, R. Nussenblatt, Standardizing immunophenotyping for the Human Immunology Project, *Nat. Rev. Immunol.* 12 (2012) 191–200.
- [7] T. Kalina, J. Flores-Montero, V.H. van der Velden, M. Martin-Ayuso, S. Bottcher, M. Ritgen, J. Almeida, L. Lhermitte, V. Asnafi, A. Mendonca, R. de Tute, M. Cullen, L. Sedek, M.B. Vidrales, J.J. Perez, J.G. te Marvelde, E. Mejstrikova, O. Hrusak, T. Szczepanski, J.J. van Dongen, A. Orfao, C. EuroFlow, EuroFlow standardization of flow cytometer

<sup>1</sup> Unless otherwise indicated, partners are located at Institut Pasteur, Paris.

<sup>2</sup> Co-coordinators of the *Milieu Intérieur* Consortium.

- instrument settings and immunophenotyping protocols, *Leukemia* 26 (2012) 1986–2010.
- [8] J.J. van Dongen, L. Lhermitte, S. Bottcher, J. Almeida, V.H. van der Velden, J. Flores-Montero, A. Rawstron, V. Asnafi, Q. Lecrevisse, P. Lucio, E. Mejstrikova, T. Szczepanski, T. Kalina, R. de Tute, M. Bruggemann, L. Sedek, M. Cullen, A.W. Langerak, A. Mendonca, E. Macintyre, M. Martin-Ayuso, O. Hrusak, M.B. Vidriales, A. Orfao, C. EuroFlow, EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes, *Leukemia* 26 (2012) 1908–1975.
- [9] D. Popadic, I. Anegon, D. Baeten, H. Eibel, T. Giese, P. Marits, E. Martinez-Caceres, F. Mascart, F. Nestle, R. Pujol-Borrell, E. Savic, C. Scheibenbogen, B. Seliger, S. Thunberg, M. Turina, F. Villanova, O. Winqvist, A.C. Wikstrom, Predictive immunomonitoring – the COST ENTIRE initiative, *Clin. Immunol.* 147 (2013) 23–26.
- [10] M. Streitz, T. Miloud, M. Kapinsky, M.R. Reed, R. Magari, E.K. Geissler, J.A. Hutchinson, K. Vogt, S. Schlickeiser, A.H. Kverneland, C. Meisel, H.D. Volk, B. Sawitzki, Standardization of whole blood immune phenotype monitoring for clinical trials: panels and methods from the ONE study, *Transplant. Res.* 2 (2013) 17.
- [11] L.K. McNeil, L. Price, C.M. Britten, M. Jaimes, H. Maecker, K. Odunsi, J. Matsuzaki, J.S. Staats, J. Thorpe, J. Yuan, S. Janetzki, A harmonized approach to intracellular cytokine staining gating: results from an international multicenter consortium proficiency panel conducted by the Cancer Immunotherapy Consortium (CIC/CRI), *Cytometry A* 83 (2013) 728–738.
- [12] V. Appay, S. Reynard, V. Voelker, P. Romero, D.E. Speiser, S. Leyvraz, Immuno-monitoring of CD8+ T cells in whole blood versus PBMC samples, *J. Immunol. Methods* 309 (2006) 192–199.
- [13] B. Hoffmeister, T. Bunde, I.M. Rudawsky, H.D. Volk, F. Kern, Detection of antigen-specific T cells by cytokine flow cytometry: the use of whole blood may underestimate frequencies, *Eur. J. Immunol.* 33 (2003) 3484–3492.
- [14] P. Renzi, L.C. Ginns, Analysis of T cell subsets in normal adults. Comparison of whole blood lysis technique to Ficoll–Hypaque separation by flow cytometry, *J. Immunol. Methods* 98 (1987) 53–56.
- [15] H.T. Maecker, A. Rinfret, P. D'Souza, J. Darden, E. Roig, C. Landry, P. Hayes, J. Birungi, O. Anzala, M. Garcia, A. Harari, I. Frank, R. Baydo, M. Baker, J. Holbrook, J. Ottinger, L. Lamoreaux, C.L. Epling, E. Sinclair, M.A. Suni, K. Punt, S. Calarota, S. El-Bahi, G. Alter, H. Maila, E. Kuta, J. Cox, C. Gray, M. Altfeld, N. Nougarede, J. Boyer, L. Tussey, T. Tobery, B. Brecht, M. Roederer, R. Koup, V.C. Maino, K. Weinhold, G. Pantaleo, J. Gilmour, H. Horton, R.P. Sekaly, Standardization of cytokine flow cytometry assays, *BMC Immunol.* 6 (2005) 13.
- [16] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York, 2009.
- [17] D. Hamann, P.A. Baars, M.H. Rep, B. Hooibrink, S.R. Kerkhof-Garde, M.R. Klein, R.A. van Lier, Phenotypic and functional separation of memory and effector human CD8+ T cells, *J. Exp. Med.* 186 (1997) 1407–1418.
- [18] M. Abuzakouk, J. Carton, C. Feighery, D.P. O'Donoghue, D.G. Weir, C. O'Farrelly, CD4+ CD8+ and CD8alpha+ beta- T lymphocytes in human small intestinal lamina propria, *Eur. J. Gastroenterol. Hepatol.* 10 (1998) 325–329.
- [19] H.T. Maecker, T. Frey, L.E. Nomura, J. Trotter, Selecting fluorochrome conjugates for maximum sensitivity, *Cytometry A* 62 (2004) 169–173.
- [20] I. Bretschneider, M.J. Clemente, C. Meisel, M. Guerreiro, M. Streitz, W. Hopfenmuller, J.P. Maciejewski, M.W. Wlodarski, H.D. Volk, Discrimination of T-cell subsets and T-cell receptor repertoire distribution, *Immunol. Res.* 58 (2014) 20–27.
- [21] F. Sallusto, D. Lenig, R. Forster, M. Lipp, A. Lanzavecchia, Two subsets of memory T lymphocytes with distinct homing potentials and effector functions, *Nature* 401 (1999) 708–712.
- [22] J.A. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, M. Goldberg, E.M. Goralczyk, B. Hyun, K. Jansen, T. Kollmann, M. Kong, R. Leif, S. McWeeney, T.D. Moloshok, W. Moore, G. Nolan, J. Nolan, J. Nikolich-Zugich, D. Parrish, B. Purcell, Y. Qian, B. Selvaraj, C. Smith, O. Tchuvatkina, A. Wertheimer, P. Wilkinson, C. Wilson, J. Wood, R. Zigon, International Society for Advancement of Cytometry Data Standards Task, R.H. Scheuermann, R.R. Brinkman, MIFlowCyt: the minimum information about a Flow Cytometry Experiment, *Cytometry A* 73 (2008) 926–930.
- [23] S. Attig, L. Price, S. Janetzki, M. Kalos, M. Pride, L. McNeil, T. Clay, J. Yuan, K. Odunsi, A. Hoos, P. Romero, C.M. Britten, C.-C.A.W. Group, A critical assessment for the value of markers to gate-out undesired events in HLA-peptide multimer staining protocols, *J. Transl. Med.* 9 (2011) 108.
- [24] S.D. Wright, R.A. Ramos, A. Hermanowski-Vosatka, P. Rockwell, P.A. Detmers, Activation of the adhesive capacity of CR3 on neutrophils by endotoxin: dependence on lipopolysaccharide binding protein and CD14, *J. Exp. Med.* 173 (1991) 1281–1286.
- [25] S. Janetzki, C.M. Britten, M. Kalos, H.I. Levitsky, H.T. Maecker, C.J. Melief, L.J. Old, P. Romero, A. Hoos, M.M. Davis, "MIATA" – minimal information about T cell assays, *Immunity* 31 (2009) 527–528.
- [26] A. Valle, N. Maugeri, A.A. Manfredi, M. Battaglia, Standardization in flow cytometry: correct sample handling as a priority, *Nat. Rev. Immunol.* 12 (2012) 864.



available at [www.sciencedirect.com](http://www.sciencedirect.com)

Clinical Immunology

[www.elsevier.com/locate/yclim](http://www.elsevier.com/locate/yclim)



# Automated flow cytometric analysis across large numbers of samples and cell types



Xiaoyi Chen <sup>a,b</sup>, Milena Hasan <sup>c</sup>, Valentina Libri <sup>c</sup>, Alejandra Urrutia <sup>c,d,e</sup>,  
Benoît Beitz <sup>c</sup>, Vincent Rouilly <sup>f</sup>, Darragh Duffy <sup>c,d,e</sup>, Étienne Patin <sup>i</sup>,  
Bernard Chalmond <sup>g,h</sup>, Lars Rogge <sup>c,i</sup>, Lluís Quintana-Murci <sup>j,k</sup>,  
Matthew L. Albert <sup>c,d,e,l,\*</sup>, Benno Schwikowski <sup>a,\*\*</sup>  
for the Milieu Intérieur Consortium

<sup>a</sup> Systems Biology Lab, Institut Pasteur, Paris, France

<sup>b</sup> Laboratory of Analysis Geometry and Modeling, Department of Mathematics, University of Cergy-Pontoise, Ile de France, France

<sup>c</sup> Center for Human Immunology, Institut Pasteur, Paris France

<sup>d</sup> INSERM U818, France

<sup>e</sup> Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris France

<sup>f</sup> Center for Bioinformatics, Institut Pasteur, Paris France

<sup>g</sup> University of Cergy-Pontoise, France

<sup>h</sup> CLMA, ENS-Cachan, France

<sup>i</sup> Laboratory of Immunoregulation, Department of Immunology, Institut Pasteur, Paris France

<sup>j</sup> Unit of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France

<sup>k</sup> CNRS URA3012, France

<sup>l</sup> INSERM UMS20, France

Received 4 July 2014; accepted with revision 20 December 2014

Available online 7 January 2015

## KEYWORDS

Flow cytometry;  
Multidimensional analysis;  
Population-based cohort;  
Automation;  
Standardization;  
Algorithms;

**Abstract** Multi-parametric flow cytometry is a key technology for characterization of immune cell phenotypes. However, robust high-dimensional post-analytic strategies for automated data analysis in large numbers of donors are still lacking. Here, we report a computational pipeline, called FlowGM, which minimizes operator input, is insensitive to compensation settings, and can be adapted to different analytic panels. A Gaussian Mixture Model (GMM)-based approach was utilized for initial clustering, with the number of clusters determined using Bayesian Information Criterion. Meta-clustering in a reference donor permitted automated identification of 24 cell types across four panels. Cluster labels were integrated into FCS files, thus permitting

*Abbreviations:* BIC, Bayesian Information Criterion; CV, coefficient of variation; DC, dendritic cell; EM, Expectation Maximization; FSC, forward scatter; GMM, Gaussian Mixture Model; MFI, mean fluorescent intensity; SSC, side scatter

\* Correspondence to: M. L. Albert, Unit of Dendritic Cell Immunobiology, Inserm U818, Institut Pasteur, 25, Rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 85 48.

\*\* Correspondence to: B. Schwikowski, Systems Biology Lab, Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 40 61 37 01.

*E-mail addresses:* [albertm@pasteur.fr](mailto:albertm@pasteur.fr) (M.L. Albert), [benno@pasteur.fr](mailto:benno@pasteur.fr) (B. Schwikowski).

<http://dx.doi.org/10.1016/j.clim.2014.12.009>

1521-6616/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

comparisons to manual gating. Cell numbers and coefficient of variation (CV) were similar between FlowGM and conventional gating for lymphocyte populations, but notably FlowGM provided improved discrimination of “hard-to-gate” monocyte and dendritic cell (DC) subsets. FlowGM thus provides rapid high-dimensional analysis of cell phenotypes and is amenable to cohort studies.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Flow cytometry is a key technology for the characterization of the cellular component of the immune system. Flow cytometers are able to simultaneously quantify different surface markers of single cells, allowing the identification and quantification of different immune cell subpopulations. In recent years, improvements in measurement speed and experimental automation have enabled comprehensive immunoprofiling of larger cohorts [1].

The gold standard for the analysis of raw flow cytometry data has until now remained “hand gating” (i.e., analysis through computer-assisted procedures for the classification of cells into single cell types using software tools such as FlowJo [2]). Each sample is analyzed by successively separating cell types by successive “gating” in a series of one- or two-dimensional projections. However, the manual operation is laborious and subject to biased visual inspection and gate adjustment. These concerns grow with increased numbers of measured phenotypic markers. Moreover, there is a major limitation in that information critical for accurate gating may not be present in the selected two-dimensional projections.

Here, we report a new method for analyzing multi-parametric flow cytometry, the need for which was motivated by the Milieu Intérieur study. This project aims at defining the genetic and environmental determinants of variable immunologic phenotypes in a healthy population [Thomas et al., co-submission]. Cell phenotyping constitutes one of the major data sets to be integrated into the data warehouse, and as such efforts were made to standardize each step of the sample collection, technical procedures and data analysis. A Companion paper highlights the pre-analytical semi-automated measures put in place for labeling and data generation [Hasan et al., co-submission]. This manuscript details the automated analytic workflow developed for the identification and analysis of 24 cell types across four 8-color cytometry panels.

Our work follows from a large number of computational approaches that have been developed for automated flow cytometry analysis. Recently, the FlowCAP study evaluated a range of approaches [13]. In all cases, however, the datasets used by these investigators were of a smaller scale than the ones in our study, in terms of samples studied (FlowCAP: up to 30 samples; here: 115 samples  $\times$  4 panels), and the number of events per experiment (FlowCAP: up to approximately 100,000 events; here: on average 300,000 events per FCS file). Due to these differences, we found that top-ranked FlowCAP approaches were inadequate to address the needs of our data sets. For example, the ADICyt approach [4] required more than 6 h for the analysis of a single sample. The FlowMeans software [5] was faster, but required manual assignment of cell types to each cluster in

every single sample. The recent X-Cyt approach [3] was designed explicitly to efficiently address the problem of larger numbers of samples. However, X-Cyt still requires the definition of a “partitioning scheme”, a series of mixture models whose sequence and parameters have to be manually configured and calibrated for each cell type of interest in any given analytic panel.

To support the analysis of the Milieu Intérieur cohort dataset, we developed a novel high-dimensional data analysis approach, which we refer to as FlowGM, utilizing fast algorithms that enable the standardized analysis of large numbers of samples. We describe its application to two representative 8-color panels with up to 11 cell populations classified per panel. Its principal feature is that, after the definition of global parameters in a reference sample (i.e., a one-time manual assignment of cell type labels to clusters), it is possible to automatically position and identify cell populations across the entire dataset. This approach will enable analysis of our large healthy donor cohort.

## 2. Materials and methods

### 2.1. Dataset

Four 8-color cytometry panels targeting major leukocyte populations across 115 individuals from different age groups and genders were designed to characterize the major immune cell populations (T cells, B cells, NK cells and monocytes), as well as subpopulations of T cells, dendritic cells (DC) and polymorphonuclear leukocytes (PMN). The standardized procedure of collection and treatment of the whole blood sample is described in [Hasan et al., co-submission]. For each of the four panels, technical replicates performed by five parallel blood samples obtained from three donors (“repeatability” studies from [Hasan et al., co-submission]) were generated to examine robustness of the experimental and computational protocols.

### 2.2. FlowGM cluster model

The input to FlowGM is a set of  $m$  sets of  $n$  quantitative measurements (“events”), formally,  $m$   $n$ -dimensional vectors. Clustering is based on a multivariate Gaussian Mixture Model (GMM) [6], which has the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^k \alpha_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

A GMM thus corresponds to a set of  $k$  clusters, each described by a cluster weight  $\alpha_j$  and an  $n$ -dimensional

Gaussian (normal) probability distribution, whose parameters  $\theta$  are its centroid  $\mu_j$ , and its extent and orientation,  $\Sigma_j$  in  $n$  dimensions. The weight of each cluster corresponds to the proportion of all cells assigned to it. Gaussian mixture models have been used for flow cytometry, but a particularity of FlowGM is that several such clusters can be used to model cells of one type that may not adequately be modeled by a single normal distribution.

### 2.3. Clustering cells using Expectation Maximization (EM)

Starting from an initial configuration, the degree of fit between the clusters and the data is quantified by a likelihood function. Each stage of an iterative optimization process (Expectation Maximization, EM) improves the likelihood in two steps [7]. In an E (Expectation) step, each event is assigned to (potentially, multiple) clusters whose location is close to the event. In an M (Maximization) step, the cluster parameters are optimized to fit the events assigned to it.

### 2.4. FlowGM workflow

**Step 1** Define pre-processing parameters (manual)  
To initialize automatic processing of Phase I, FlowGM requires the input of a few parameters, such as the choice of a reference sample, and the selection of potential pre-filtering and post-filtering parameters.

**Step 2** Perform pre-filtering (automatic)  
Automated pre-filtering helps eliminate noise (such as doublets) and/or “uninteresting” cells (i.e., Dump populations), which is of importance when the cell types of interest are rare. Two filters have been pre-configured: A doublet filter and a filter that eliminates cells that are negative relative to user-definable markers (based on two-component one- or two-dimensional GMMs). The filter eliminates the 95th percentile of the cluster corresponding to the “uninteresting” cells.

**Step 3** Determine the number of clusters (automatic)  
The number of clusters ( $k$ ) used to model the reference sample is determined by minimizing the Bayesian Information Criterion (BIC) [8]. The BIC represents a tradeoff maximizing the degree of fit between the cluster model and the data on the one hand (expressed by the likelihood  $p(\mathbf{x}|\theta)$ ), and, minimizing, on the other hand, model complexity (based on the number of clusters  $k$ ):

$$\text{BIC}_k = -2 \ln(p(\mathbf{x}|\theta)) + k \ln(m).$$

Specifically, we choose  $k$  that minimizes the average of  $\text{BIC}_k$  under 20 EM runs starting with random initial configurations.

**Step 4** Establish the reference clustering (automatic)  
Once the number  $k$  of clusters has been determined, FlowGM determines 100 random initial configurations of  $k$  clusters as starting points, and performs clustering using Expectation Maximization, as described in Section 2.4. The resulting clustering with the highest

likelihood is selected as the reference clustering in the second FlowGM phase.

**Step 5** Label reference clusters with cell types (manual)  
An operator defines the cell types of interest, and assigns one or more corresponding clusters to each such cell type (*labeling*). Thus, each cell type of interest corresponds to a set of clusters (*meta-cluster*).

**Step 6** Perform post-filtering (automatic, optional)  
This optional step offers the possibility of eliminating additional “uninteresting” events that remain in the clusters determined in Step 5 (analogous to a “dump” gate for conventional approaches and useful in focusing the clustering analysis). Two filters have been pre-configured: A dead cell filter (based on the Viability channel), and a “dump” filter that eliminates selected cells in specified meta-clusters. In both instances, the cells above or below a defined threshold are removed. This threshold is automatically determined as the 95th/99th percentiles of a fitted one-dimensional Gaussian distribution of a reference population along a pre-defined channel. The reference population may either be the meta-cluster itself, or a negative control that has been removed in the pre-filtering (Step 2).

**Step 7** Cohort samples: pre-filter and cluster by adjusting labeled reference clustering (automated)

After the reference sample has been processed in Steps 1–5, FlowGM processes all other samples in a fully automated manner. Pre-filtering proceeds as described for the reference donor (Step 2). FlowGM then determines the clustering using EM, as described in Section 2.4, starting with the labeled reference clustering (from Step 5) as the initial configuration. Finally, post-filtering is applied (if selected), as in Step 6.

### 2.5. Visualization of the resulting clusters in FlowJo

One innovation incorporated into FlowGM included the embedding of labels for each cluster and meta-cluster as additional attributes (numerical identifiers) for each cell in the FCS data file. This allows inspection of the different clusters in FlowJo [2] or other software that can analyze FCS data files.

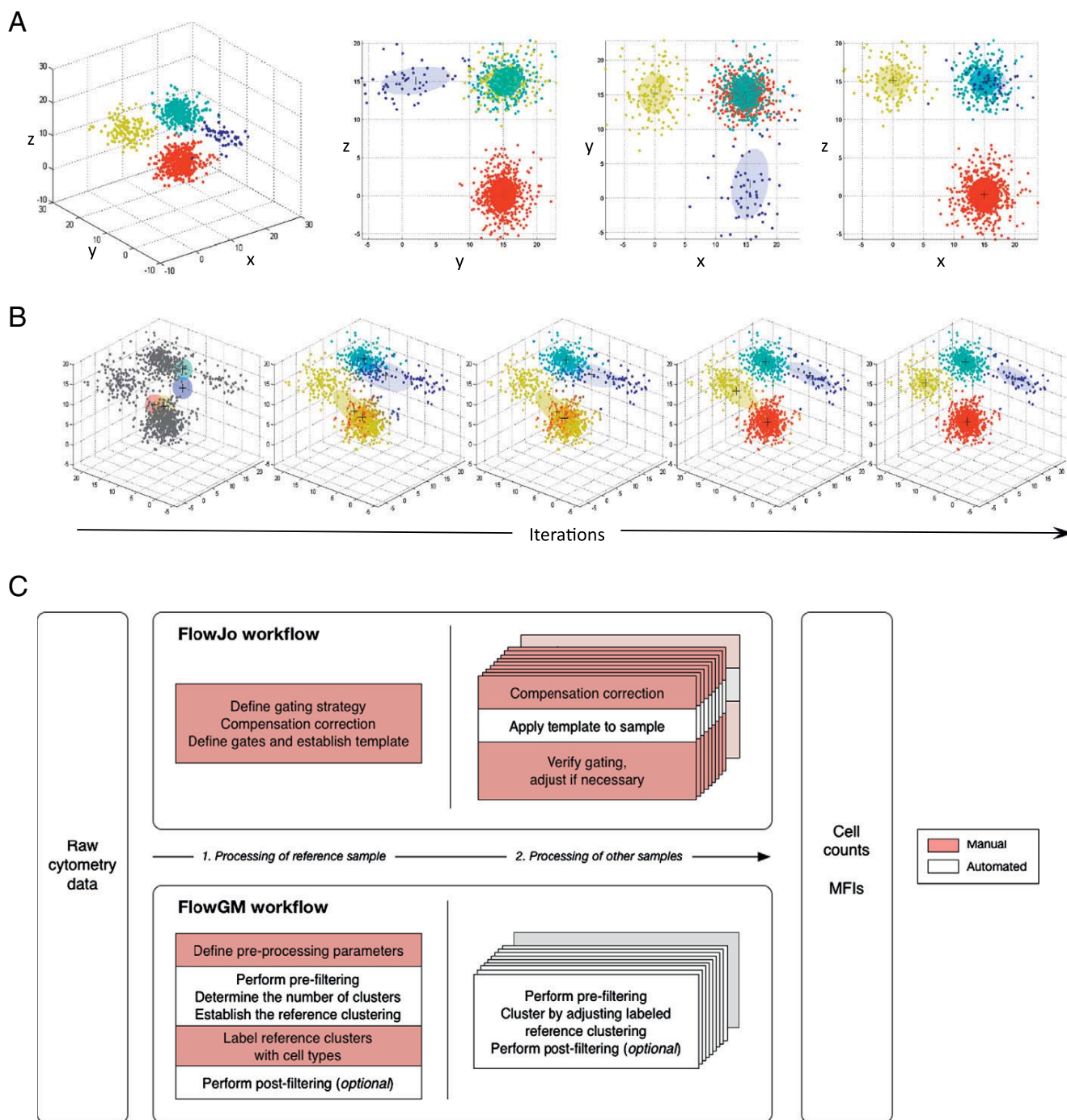
### 2.6. Software implementation

FlowGM was implemented using Matlab and Statistics Toolbox Release 2012b [9] and R (version 3.0.1) [10] flowCore package [11]. The visualization graphs were prepared with FlowJo software version 9.7.5.

## 3. Results

### 3.1. FlowGM workflow

Motivated by the need for high-quality analysis of a large flow cytometry data set, we developed the novel, and largely automated FlowGM data analysis approach. Its computational high-dimensional clustering approach avoids the limitations inherent to analysis based on two-dimensional projections (Fig. 1A). Experimental data is modeled as a mixture of



**Figure 1** Analytic approach for multidimensional clustering of multi-parameter cytometric data. (A) Four simulated clusters in 3D space that cannot be separated in any 2D projection. (B) Illustration of the expectation–maximization (EM) clustering algorithm using Gaussian mixture model (GMM) clusters, when applied to this data. Points are colored according to their posterior likelihood, the ellipsoid reflects cluster shape, ‘+’ indicates the cluster centroid, transparency of each ellipsoid reflects cluster weight. Five phases are shown: initial random parameter values, updated parameters after the first M-step, after two iterations, after ten iterations, and final solution. (C) FlowJo and FlowGM workflows.

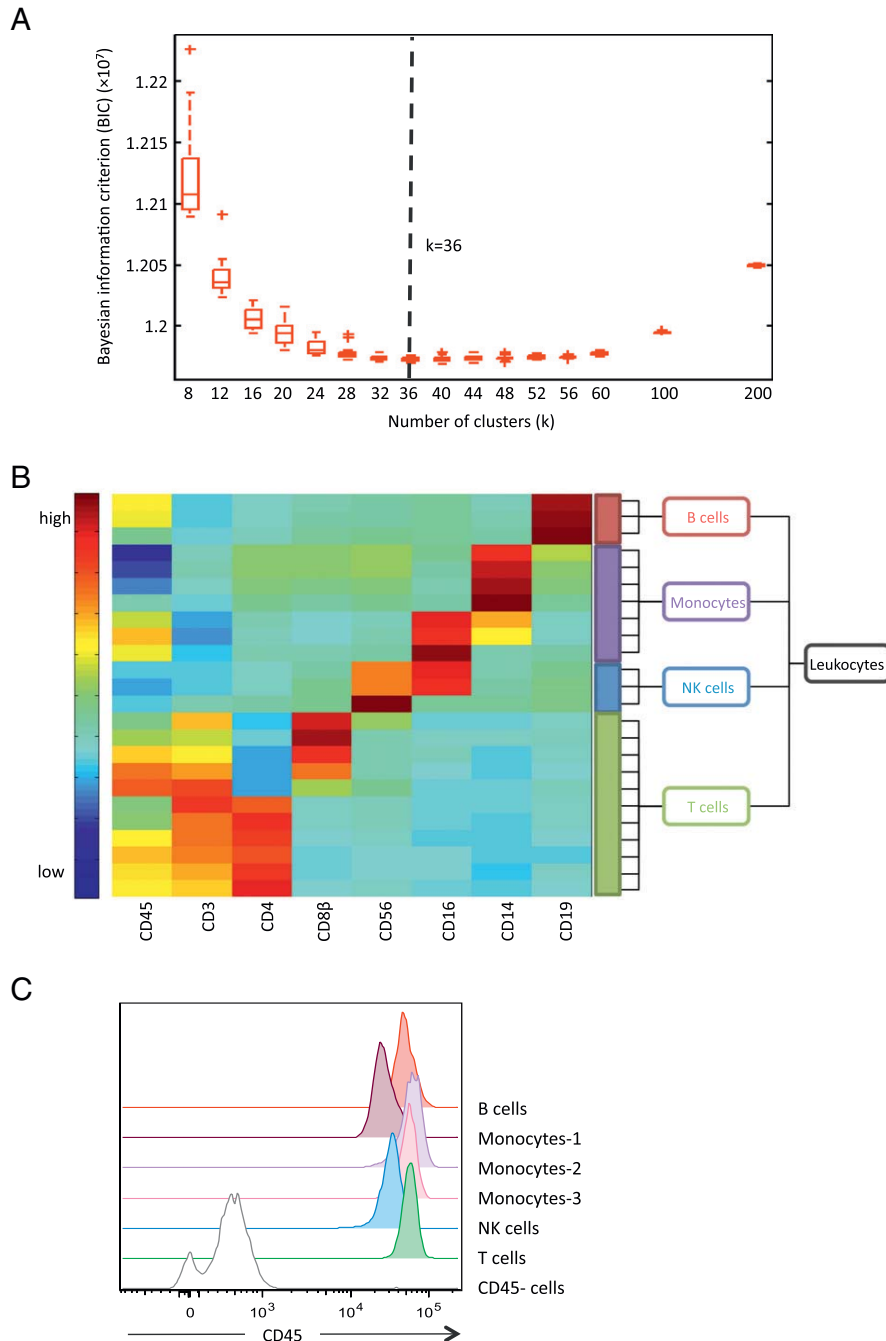
normal distributions (See [Materials and methods, Section 2.3](#)) and employs Expectation Maximization (EM) to iteratively adapt model parameters (Fig. 1B and [Materials and methods, Section 2.4](#)).

The overall operation of the FlowGM workflow can be understood on the basis of its similarities and differences relative to the current ‘gold standard’ manual FlowJo workflow (Fig. 1C). For both approaches, two phases can

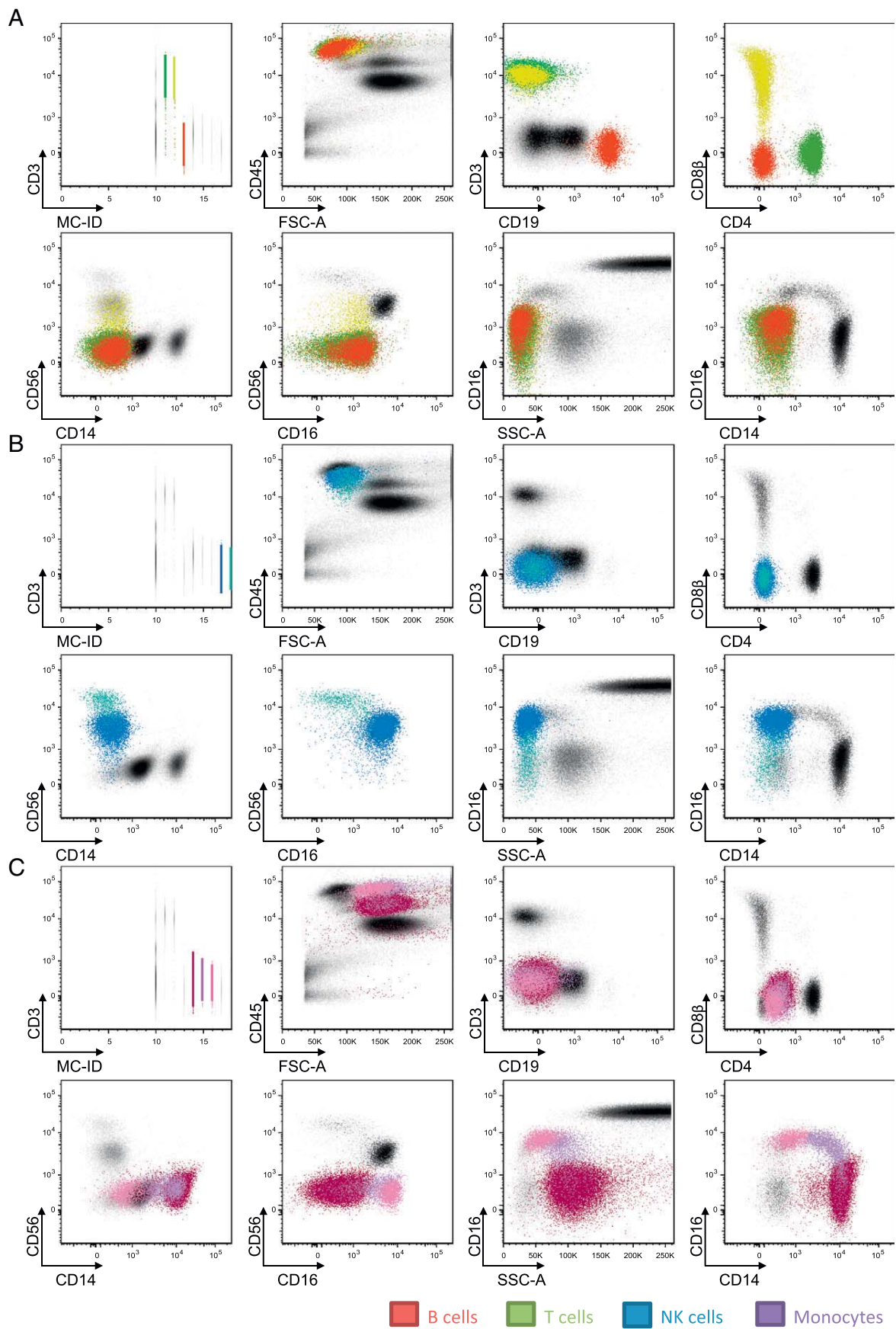


be distinguished. In the first phase, method parameters are calibrated on selected reference samples. In a second phase, all other samples are processed on the basis of the calibrated

parameters. To be suitable for large cohort studies, FlowGM was designed to minimize the manual per-sample effort in the second phase.



**Figure 2** Number of clusters and mapping to cell types. (A) The number of clusters  $k$  is determined with the minimum average Bayesian Information Criterion (BIC) when evaluated on 20 random initial solutions for each choice of  $k$ . For the lineage panel,  $k = 36$  is optimal. (B) User-based aggregation of FlowGM clusters into meta-clusters for immune cell type characterization with cluster centroid heat map (normalized coordinates). B cells are identified as  $CD19^+$ , T cells are identified as  $CD3^+$  with two subsets:  $CD4^+$  (T-1) and  $CD8\beta^+$  (T-2), NK cells are identified as  $CD56^+$  with two subsets:  $CD16^{hi}$  (NK-1) and  $CD56^{hi}$  (NK-2), monocytes are identified as three subsets:  $CD14^{hi}$  (Mono-1),  $CD14^{hi}CD16^{hi}$  (Mono-2) and  $CD14^{lo}CD16^{hi}$  (Mono-3). The manually assigned cell types are indicated on the right. (C) Distribution of CD45 intensity for different cell types of interest in the reference donor.



### 3.2. Identification of the major cell lineages by FlowGM

We first applied FlowGM to the lineage panel dataset [Hasan et al., co-submission]. Cells were stained with the markers CD45, CD3, CD4, CD8 $\beta$ , CD14, CD16, CD19, and CD56. Following the approach of the manual analysis by Hasan et al., we used forward and side scatter (FSC/SSC) solely to exclude doublets; the remainder of our data analysis is performed on the dimensions of the indicated eight markers. The number of events in the data files ranged from 106,000 to 787,000. After filtering out doublets, FlowGM estimated the optimal number of clusters  $k$  to be 36, using the BIC (see [Materials and methods, Section 2.4](#)) on the reference donor (Fig. 2A).

Once  $k$  was determined, FlowGM performed EM clustering 100 times, starting with different random initial configurations of  $k$  clusters. The clustering solution with the highest likelihood  $p(\mathbf{x}|\theta)$  constitutes the *reference clustering*, whose clusters were then manually labeled with the different cell types of interest (i.e., leukocyte subpopulations). The corresponding cluster centroids are represented as a heat map, with the assigned cell types indicated (Fig. 2B).

Note that only 24 of the 36 clusters corresponded to cell types of interest, and the color coding is chosen independently for each marker to resolve the entire spectrum of expression across these cell types (using the Matlab HeatMap function). For example, as CD45<sup>-</sup> cell populations were not of interest in this study, all selected cells were CD45<sup>+</sup> and as indicated by the normalization, the lowest and highest levels of CD45 expression were observed in monocytes and T cells, respectively (Figs. 2B, C).

To facilitate the understanding of our findings and permit user cross-validation, FlowGM allows the embedding of cluster IDs and meta-cluster IDs as additional channels (designated “C-ID” and “MC-ID”, respectively) into the FCS input file, permitting importation of all data into FlowJo (or other FCS-compatible software). FlowJo visualizations of the labeled FlowGM lineage clusters confirmed our GMM-based assignments (Fig. 3). By gating on MC-ID to select one FlowGM meta-cluster, it is possible to view the clustered cells in 2D projections that correspond to manual gating strategies. FlowJo visualizations of all 36 FlowGM clusters are shown in Fig. S1. Backgating is also possible: starting with manual gated data and examining where the captured events cluster in C-ID or MC-ID space (data not depicted).

### 3.3. Pre-filtering supports clustering of rare dendritic cell subsets

We next evaluated the performance of the method on rare subsets of cells (<1% of the total cell events). In addition to the elimination of doublets early in the analysis, we identified the need for pre-filtering of cells considered by the user as uninteresting – similar to the use of a “Dump” gate – only in the case of FlowGM the procedure is automated and thus

removes operator bias. Pre-filtering of the DC panel was based on a two-component, two-dimensional GMM that utilized data from CD14 and HLA-DR markers. Thresholds were automatically set at the 95th percentiles of the CD14/HLA-DR double-negative population (represented by the red line, Fig. 4A). The resultant cells were investigated using the FCS embedding feature of FlowGM, and inspection of representative files revealed accurate retention of desired HLA-DR<sup>+</sup> and/or CD14<sup>+</sup> cells (Fig. 4B).

Next, we estimated  $k$  using the BIC and defined a clustering solution using data from a reference donor (Fig. S2). Of the 40 clusters defined as the optimal fit, 22 were of interest and manual labeling of the meta-clustered data captured five myeloid cell subsets: cDC1, identified by their high BDCA2 MFI and low expression of CD14; pDCs, identified by the highest BDCA2 and BDCA4 MFIs; cDC3, identified by their expression of BDCA3; CD14<sup>lo</sup> monocytes, identified by the intermediate expression of CD14; and CD14<sup>hi</sup> monocytes, by the high CD14 MFI (Fig. S2B). Again, we highlight that the data represented in the heat map has been normalized, and in instances where all cell populations are positive for a given marker (i.e., HLA-DR), the normalization will scale values to span the range of marker expression. To illustrate the distributions of HLA-DR intensity, histogram plots for DCs and monocytes are shown (Fig. S2C).

Next, an initial post-filter removed dead cells from each meta-cluster, based on the Dump channel. A second post-filter removed cells from cDC1 and cDC3 populations based on expression of BDCA1 and BDCA3 respectively, of the CD14/HLA-DR double-negative population that was previously filtered out.

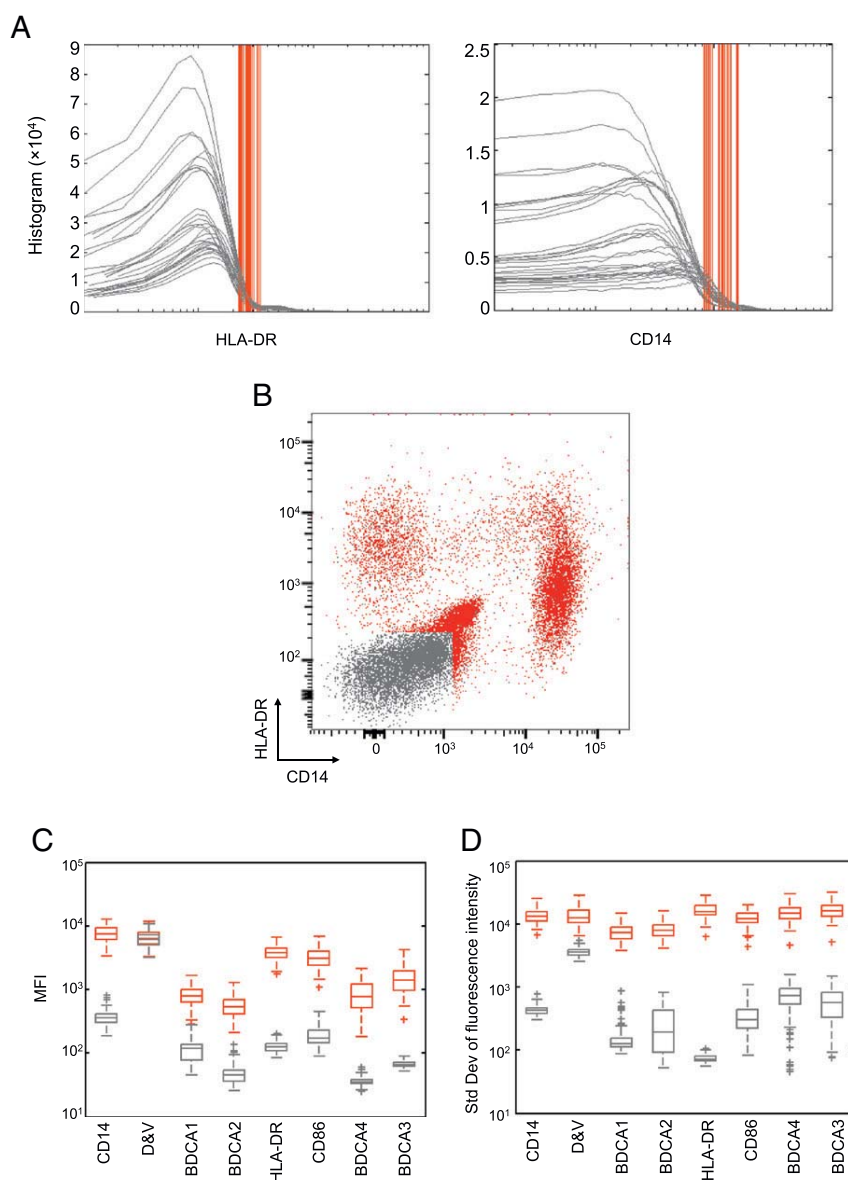
As a final validation step, we compared the level of marker expression between retained cells and events that were removed by the filtering process. Across all dimensions of the data set, we confirmed the efficacy of the pre-filtering approach (Figs. 4C, D). Additional visual confirmation can be found in the FlowJo-projected data, where meta-clustered data is overlaid on the total cell events in a representative file (Fig. S3).

### 3.4. FlowGM is robust to selection of reference donor and may be applied to uncompensated data

One potential concern with the FlowGM approach is the sensitivity of the clustering result to the choice of the reference sample in Step 1 (cf. [Section 2.4](#)). This is an important issue, as the resulting reference clustering will be used as the basis to cluster the data from all other samples. While practitioners may have a good intuition about which one of the input samples is “representative”, the degree of sensitivity to this choice could, in principle, be large.

We therefore investigated whether a more representative reference clustering based on a larger group of samples would be needed. To this end, we constructed 11 different clusterings: the originally chosen reference clustering (which we denote here by 1\*), and ten alternative reference clusterings

**Figure 3** Visualization of labeled meta-clusters in FlowJo Cluster IDs is incorporated into the FlowJo input file. Shown are meta-clusters with all principal manual gating steps, starting with SSC-A/Meta-Cluster ID (MC-ID). (A) The identified CD19<sup>+</sup> B cells (red) and CD4<sup>+</sup> (green) and CD8 $\beta$ <sup>+</sup> (yellow) subsets of CD3<sup>+</sup> T cells. (B) CD56<sup>hi</sup> (light blue) and CD16<sup>hi</sup> (dark blue) NK cell sub-populations. (C) CD14<sup>hi</sup>, monocytes (Mono-1, mauve) CD14<sup>hi</sup>CD16<sup>hi</sup> monocytes (Mono-2, lavender) and CD14<sup>lo</sup>CD16<sup>hi</sup> monocytes (Mono-3, light purple).



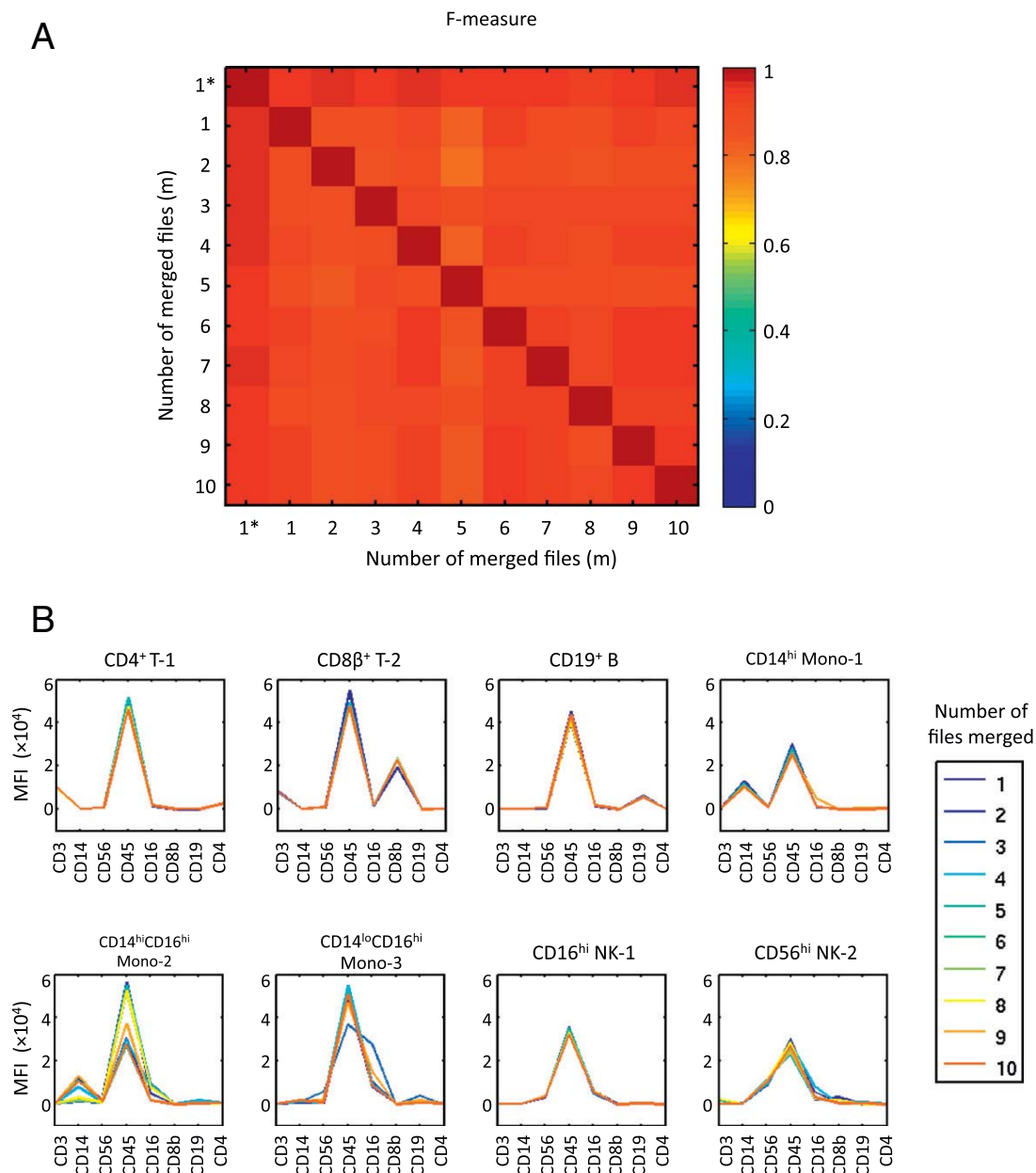
**Figure 4** Pre-filtering for analysis of rare cell populations. (A) Pre-filtering in dendritic cells (DC) by low expression of CD14 and HLA-DR. Red lines indicate the thresholds that were automatically determined using GMM. (B) Validation of pre-filtering using FlowJo visualization. (C) MFI of filtered (gray) and remaining (red) cells. Pre-filtered cells display a lower MFI in all channels except Dump. (D) Standard deviation of fluorescence intensity for the same cell population. Filtered cells display less variation.

(1, ... 10) of increasing complexity, which were obtained by selecting a series of 10 samples from randomly chosen donors, and then merging the samples 1, ...,  $i$  for each  $i = 1, \dots, 10$ . Merging different samples without alignment can be expected to create reference clusterings that contain technical shifts, and thus could translate into significant variation in the clustering result.

For each possible pair of these 11 reference clusterings, we then determined the similarity of the two outcomes after clustering, using the F-measure [11,12] (Fig. 5A). Notably, the F-measure values were close to 1, independently, for all pairs of reference clusterings, indicating that the different reference clusterings did not translate into significantly different clustering outcomes. The locations of the resulting cell types for the

different reference clusterings were further represented in parallel coordinate plots (Fig. 5B). Except for the Mono-2 and Mono-3 populations, all coordinates match extremely well among the different reference clusterings across all dimensions. Together, these observations suggest that the choice of the initial reference clustering may not have a large impact on the resulting outcome.

We also investigated the impact of compensation. Routinely, automatic hardware compensation [Hasan et al., co-submission] is employed. Here, we compare the results of our approach on the same input data in an uncompensated state; machine-compensated; or machine-compensated and FlowJo-corrected. The computational analyses on these three datasets were initialized with the re-estimated parameters from the



**Figure 5** Differences in reference clustering do not impact cell type identification. Different reference clusterings are generated by merging data from one to ten randomly selected donors; solutions are then applied to 115 cohort donors. (A) Pairwise average similarity (F-measure) of solutions over 115 cohort donors after using different reference clusterings. (B) Mean fluorescence intensity (MFI) of each identified cell population from different reference clusterings.

reference clustering on machine-compensated data. The counts for three repeatability samples obtained from different dataset are shown (Fig. S4), and indicate that FlowGM is insensitive to instrument compensation, and therefore resistant to potential compensation error in the context of large datasets.

### 3.5. Benchmarking of FlowGM demonstrates its reliability and utility

To directly compare FlowGM clusters to manually gated data sets, we first calculated, for each hand-gated cluster

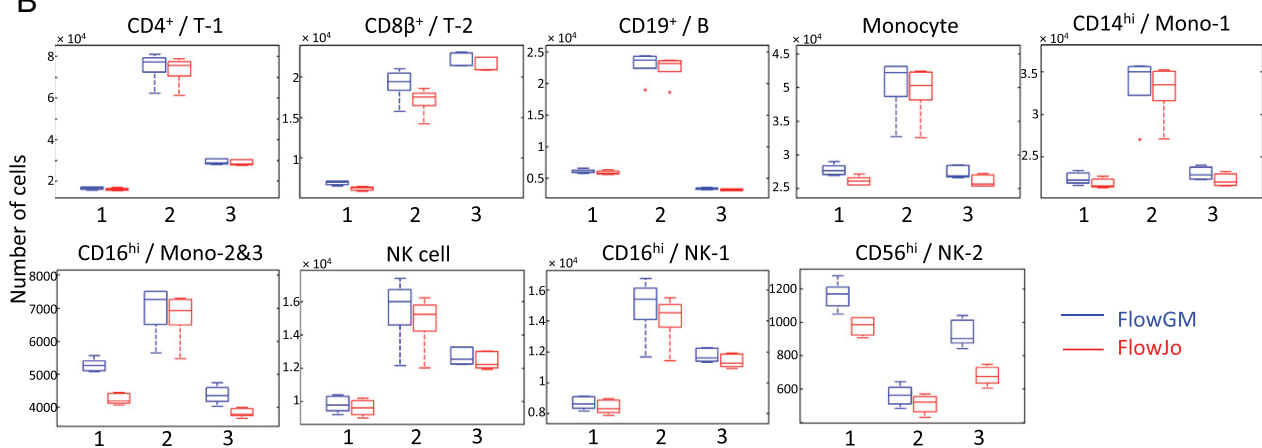
in the reference donor, the percentage of its events present in every other FlowGM cluster (Fig. 6A). The values indicated that, overall, the two approaches group events similarly. The one exception were monocytes, where FlowGM supported easy segregation of the CD14<sup>hi</sup>CD16<sup>hi</sup> sub-population of monocytes (Mono-2) from CD14<sup>lo</sup>CD16<sup>hi</sup> sub-monocytes (Mono-3), despite the lack of additional monocyte-specific markers (e.g., MCSF-1, CX<sub>3</sub>CR1, CCR2 PMID: 20832340).

We also studied the variability of manual and FlowGM-derived cell counts across the repeatability samples studied in Hasan et al. (Fig. 6B). We find that FlowGM results showed

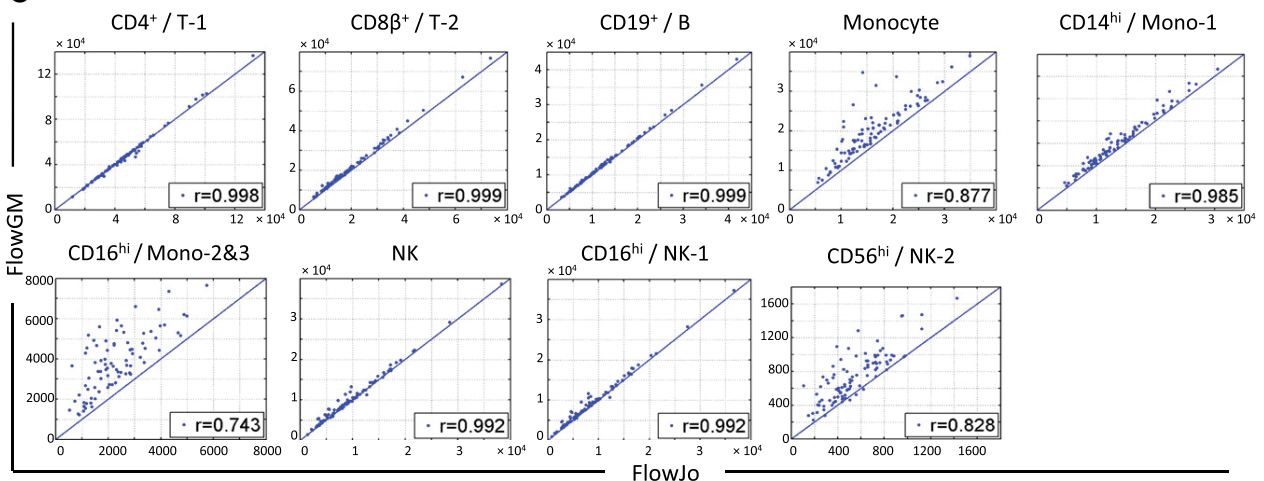
A

		FlowJo gates						
		CD4+	CD8+	CD19+	CD14 <sup>hi</sup> mono	CD16 <sup>hi</sup> mono	CD16 <sup>hi</sup> NK	CD56 <sup>hi</sup> NK
FlowGM meta-clusters	T - 1	99.8	0	0	0	0	0	0
	T - 2	0	98.7	0	0	0	0	0
	B	0	0	98.8	0	0	0	0
	Mono - 1	0	0	0	97.3	0	0	0
	Mono - 2	0	0	0	2.64	42.7	0	0
	Mono - 3	0	0	0	0	50.2	0	0
	NK - 1	0	0	0	0	3.8	96.2	1.5
	NK - 2	0	0	0	0	0	0.7	98.2
Unmapped clusters		0.2	1.3	1.2	0.06	3.3	3.1	0.3

B



C



**Figure 6** Comparison of manually gated data and FlowGM analysis. (A) Performance on reference donor: percentage of events in FlowJo cluster present in FlowGM clusters. (B) Performance on repeatability data: counts of each cell type for three donors with five replicates. The FlowGM results show a comparable CV with manually gated data. (C) Performance on 115 cohort donors: manually gated data and FlowGM analysis highly agree ( $r = 0.944$ ) on 115 cohort donors.

good agreement with the results from manual analyses. The slight bias for higher numbers from FlowGM may stem from the need for high-dimensional information to confidently assign certain events to cell types (as in the schematic

example shown, Fig. 1A). Coefficients of variation (CVs), which represent variation of data analysis and experimental variation, were at similar levels, further indicating the high accuracy of FlowGM analysis.

**Table 1** Repeatability.

	Donor <sup>a</sup> :	#1	#2	#3
Lineage	CD4+ T cells	16870 (4.4) <sup>b</sup>	77306 (9.9)	28838 (4.4)
	CD8 $\beta$ + T cells	6986 (3.8)	19408 (10.6)	21416 (4.0)
	CD19+ B cells	5983 (5.3)	23679 (9.8)	3325 (4.1)
	Monocytes	27615 (3.0)	42233 (11.0)	26894 (3.5)
	CD14 <sup>hi</sup> CD16 <sup>lo</sup> mono	22269 (3.2)	34969 (10.9)	22872 (3.2)
	CD14 <sup>hi</sup> CD16 <sup>hi</sup> mono	3196 (4.2)	3759 (11.7)	1436 (8.9)
	CD14 <sup>lo</sup> CD16 <sup>hi</sup> mono	2058 (3.3)	3505 (10.9)	2907 (5.3)
	NK cells	9803 (5.1)	15989 (12.9)	12534 (4.0)
	CD16 <sup>hi</sup> NK	8633 (4.9)	15424 (13.0)	11632 (3.7)
	CD56 <sup>hi</sup> NK	1171 (7.4)	565 (11.2)	902 (8.9)
T cell	CD4+ T cells	13172 (4.5)	64809 (16.4)	23450 (0.7)
	CD4+ T <sub>N</sub>	3043 (4.8)	23398 (13.8)	8961 (8.1)
	CD4+ T <sub>CM</sub>	8973 (4.4)	39350 (18.2)	13044 (3.6)
	CD4+ T <sub>EM</sub>	1044 (6.7)	3329 (18.3)	1250 (11.4)
	CD8 $\beta$ + T cells	5245 (5.7)	14847 (16.8)	15283 (3)
	CD8 $\beta$ + T <sub>N</sub>	553 (8.2)	5692 (16.8)	5903 (2.3)
	CD8 $\beta$ + T <sub>CM</sub>	2297 (6.2)	5737 (13.6)	5996 (7.7)
	CD8 $\beta$ + T <sub>EM</sub>	548 (10.2)	1181 (15)	1092 (21.2)
	CD8 $\beta$ + T <sub>EMRA</sub>	717 (5.1)	1206 (46.8)	954 (16)
	CD8 $\beta$ + 27 <sup>int</sup>	1036 (8.7)	1096 (23.3)	1516 (11.7)
DC	CD4+ CD8 $\alpha$ + T cells	153 (11.4)	770 (19.3)	539 (28)
	CD14+ monocytes	25232 (12.2)	29764 (4.4)	21287 (8.4)
	pDC	304 (18.5)	409 (4.1)	438 (5.0)
	cDC1	2159 (12.1)	5188 (3.9)	1677 (10.4)
PMN	cDC3	42 (30)	87 (16)	44 (8.1)
	Neutrophils	96062 (14.3)	188428 (13.0)	119529 (12.0)
	Basophils	1751 (11.4)	5878 (7.2)	2323 (11.6)
	Eosinophils	10483 (13.2)	18539 (10.6)	22329 (6.2)

<sup>a</sup> Fresh blood samples from three healthy donors were divided into five aliquots each and immediately stained using four antibody panels.

<sup>b</sup> Median absolute cell counts per 1 mL of blood in five independent analyses is represented for each cell population, as well as the corresponding coefficient of variation (CV).

Absolute counts and CVs for the repeatability data from all four panels are provided (Table 1). The estimation of the number of clusters and the resulting cluster positions, and assignments to cell types for the T cell and PMN panels are shown in Figs. S5 and S6 respectively. For the observed cell types, absolute counts were highly reproducible, with most CVs <15%. Compared to results of Hasan et al. [co-submission], the level of reproducibility of FlowGM was similar to the manual gating results across all four panels.

Finally, we used FlowGM-generated absolute cell counts of the lineage panel across 115 donors from the Milieu Intérieur cohort [Thomas et al., co-submission], comparing results to those obtained by manual gating. Again, results were highly concordant (Fig. 6C). The running time of the computational analysis for a single panel depends on the number  $n$  of measured events in each sample and the number  $k$  of clusters. For the panels analyzed here, the computation required 0.5 h (DC panel) and ~4 h (lineage panels) on a standard laptop PC.

#### 4. Discussion

The FlowGM flow cytometry approach was developed to address the need for fast, robust and high-quality analysis for

the Milieu Intérieur Consortium study. Our comprehensive validation study has shown that FlowGM has produced user-validated results whose quality is on par with, and in some cases, exceeds, the hand-gating approach. This is an exciting finding, as its simple computational approach does not require the expert knowledge and experience that is available to human operators. One important difference lies in the systematically higher number of events assigned to cell types by FlowGM, which suggests that the full dimensionality of the data, instead of two-dimensional views, allows for assigning cells that are unassigned in manual two-dimensional analysis due to the lacking dimensionality and user-bias. Another facet of this fundamental difference may be the observed ability of FlowGM to segregate subpopulations of monocytes without the need for an additional specific marker. Notably, separation of CD14<sup>lo</sup>CD16<sup>hi</sup> monocytes from NK cells and other cell populations was achieved by integrating information from all eight dimensions.

When comparing the design of FlowGM workflow to other computational clustering approaches, a characteristic difference lies in the choice to computationally model single cell types as mixtures of Gaussians, as opposed to single Gaussians, or other distributions, coupled with the incorporation of knowledge and experience of a human operator to define which clusters belong to the same cell

type (referred to herein as meta-clusters). This design may constitute a 'sweet spot' in cytometry workflow design: A fast and efficient overall workflow, combined with a mathematical model that is flexible enough to model experimental data well, the solution of a hard core problem (the assignment of cell types to clusters) using operator intervention, and the limitation of this intervention to a single reference sample, as the transposition of this knowledge to all other samples can be automated with high accuracy.

The minimization of operator intervention means not only significant savings in terms of manual effort, but also the elimination of variability between different samples introduced by subjective decisions, and a considerable improvement in transparency and reproducibility of the path from the samples to the absolute and relative cell counts. Furthermore, the facility with which results are accessible for human inspection using conventional tools, and the relative simplicity of the FlowGM approach itself imply a high level of accessibility to non-specialists that – we believe – will continue to play an important role in the evolution of the approach.

We believe that the FlowGM workflow is applicable to most other flow cytometry datasets, and anticipate that the need for fast, robust, and high-quality analysis of large cytometry datasets will only increase. Adaptations of the method may be required for heterogeneous samples, in which no single reference sample may be representative for all others, or in cases where certain subpopulations may be activated (e.g., disease populations). We believe that there are relatively straightforward approaches to extend FlowGM to automatically detect cases of inadequate fit, for example, through the introduction of additional reference donors (with recursive iteration of the manual Step 5). The increased availability of experimental datasets that have been acquired under standardized conditions may facilitate comparison and integration, which may lead to the necessary insights and technical developments to fully automate flow cytometry data analysis.

## 5. Conflict of interest statement

The authors declare that there are no conflicts of interest.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.clim.2014.12.009>.

## Acknowledgments

### Consortia

The Milieu Intérieur Consortium<sup>1</sup> is composed of the following team leaders: Laurent Abel (Hôpital Necker), Andres Alcover, Philippe Bouso, Pierre Bruhns, Ana Cumano, Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Odile Gelpi, Antonio Freitas, Ivo Gomperts-Boneca, Serge Hercberg

(Université Paris 13), Olivier Lantz (Institut Curie), Claude Leclerc, Hugo Mouquet, Sandra Pellegrini, Stanislas Pol (Hôpital Côtchin), Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis (Institut Curie), Frédéric Tangy, Eric Tartour (Hôpital Européen George Pompidou), Antoine Toubert (Hôpital Saint-Louis), Marie-Noëlle Ungeheuer, Lluís Quintana-Murci<sup>2</sup>, Matthew L. Albert<sup>3</sup>.

Additional information can be found at: <http://www.pasteur.fr/labex/milieu-interieur>.

## References

- [1] V. Orru, M. Steri, G. Sole, C. Sidore, F. Virdis, M. Dei, S. Lai, M. Zoledziewska, F. Busonero, A. Mulas, M. Floris, W.I. Mentzen, S.A. Urru, S. Olla, M. Marongiu, M.G. Piras, M. Lobina, A. Maschio, M. Pitzalis, M.F. Urru, M. Marcelli, R. Cusano, F. Deidda, V. Serra, M. Oppo, R. Pilu, F. Reinier, R. Berutti, L. Pireddu, I. Zara, E. Porcu, A. Kwong, C. Brennan, B. Tarrier, R. Lyons, H.M. Kang, S. Uzzau, R. Atzeni, M. Valentini, D. Firinu, L. Leoni, G. Rotta, S. Naitza, A. Angius, M. Congia, M.B. Whalen, C.M. Jones, D. Schlessinger, G.R. Abecasis, E. Fiorillo, S. Sanna, F. Cucca, Genetic variants regulating immune cell levels in health and disease, *Cell* 155 (2013) 242–256.
- [2] FlowJo, TreeStar Software, Ashland, OR.
- [3] X. Hu, H. Kim, P.J. Brennan, B. Han, C.M. Baecher-Allan, P.L. De Jager, M.B. Brenner, S. Raychaudhuri, Application of user-guided automated cytometric data analysis to large-scale immunoprofiling of invariant natural killer T cells, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 19030–19035.
- [4] Adicyt, Adinis S.R.O., Bratislava, Slovenia, 2013.
- [5] N. Aghaeepour, R. Nikolic, H.H. Hoos, R.R. Brinkman, Rapid cell population identification in flow cytometry data, *Cytometry A* 79 (2011) 6–13.
- [6] G.J. McLachlan, D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.
- [7] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *J. R. Stat. Soc. Ser. B* 39 (1977) 1–38.
- [8] G.E. Schwarz, Estimating the dimension of a model, *Ann. Stat.* 6 (1978) 461–464.
- [9] Matlab and Statistics Toolbox Release 2012b, The MathWorks Inc., Natick, Massachusetts, United States, 2012.
- [10] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [11] B. Ellis, P. Haaland, F. Hahne, N.L. Meur, N. Gopalakrishnan, J. Spidlen, flowCore: Basic structures for flow cytometry data, R package version 1.30.7 (2014).
- [12] C.J. Van Rijsbergen, *Information Retrieval*, 2nd ed., 1979. (Butterworth).
- [13] N. Aghaeepour, G. Finak, C.A.P.C. Flow, D. Consortium, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, R.H. Scheuermann, Critical assessment of automated flow cytometry data analysis techniques, *Nat. Methods* 10 (2013) 228–238.

<sup>1</sup> Unless otherwise indicated, partners are located at Institut Pasteur, Paris.

<sup>2</sup> Co-coordinator of the Milieu Intérieur Consortium.

<sup>3</sup> Co-coordinator of the Milieu Intérieur Consortium.



## REFERENCES

---

## REFERENCES

- Aggarwal, B. B., Gupta, S. C. and Kim, J. H. (2012). Historical perspectives on tumor necrosis factor and its superfamily : 25 years later, a golden journey. *Blood* 119, 651–665.
- Ahmad-Nejad, P., Hacker, H., Rutz, M., Bauer, S., Vabulas, R.M., and Wagner, H. (2002). Bacterial CpG-DNA and lipopolysaccharides activate Toll-like receptors at distinct cellular compartments. *Eur. J. Immunol.*32, 1958–1968.
- Akira, S., T. Hirano, T. Taga, and T. Kishimoto. (1990). Biology of multifunctional cytokines - IL-6 and related molecules (IL-1 and TNF). *Faseb Journal* 4, 2860-2867.
- Akira, S. and Takeda, K. (2004) Toll-like receptor signaling. *Nat. Rev. Immunol.* 4, 499–511.
- Akira, S., S. Uematsu, and O. Takeuchi. (2006). Pathogen recognition and innate immunity. *Cell* 124:783-801.
- Albert M.L., Jegathesan, M. Darnell, R. B. (2001). Dendritic cell maturation is required for the cross tolerization of CD8+ T cells. *Nat. Immunol.* 2, 1010-1017.
- Alexopoulou, L., Holt, A.C., Medzhitov, R., and Flavell, R.A. (2001). Recognition of double-stranded RNA and activation of NF- $\kappa$ B by Toll-like receptor 3. *Nature* 413, 732–738.
- Alexopoulou, L., Thomas, V., Schnare, M., Lobet, Y., Anguita, J., Schoen, R. T., Medzhitov, R., Fikrig, E. and Flavell, R. A. (2002). Hyporesponsiveness to vaccination with *Borellia burgdorferi* OspA in humans and in TLR1- and TLR2- deficient mice. *Nat. Med.* 8, 878-884.
- Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106.
- Altfeld, M., and Gale, M. (2015). Innate immunity against HIV-1 infection. *Nat. Immunol.* 16, 554–562.
- Amado, I.F., Berges, J., Luther, R.J., Mailhé, M., Garcia, S., Bandeira, A., Weaver, C., Liston, A., and Freitas, A.A. (2013). IL-2 coordinates IL-2-producing and regulatory T cell interplay. *J. Exp. Med.* 210, 2707–2720.
- Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326, 257–263.

- Andersen-Nissen, E., Smith, K.D., Strobe, K.L., Barrett, S.L., Cookson, B.T., Logan, S.M., and Aderem, A. (2005). Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc. Natl. Acad. Sci. USA* 102, 9247–9252
- Andrejeva, J., Childs, K. S., Young, D. F., Carlos, T. S., Stock, N., Goodbourn, S. and Randall, R. E. (2004) The V proteins of paramyxoviruses bind the IFN-inducible RNA helicase, MDA-5, and inhibit its activation of the IFN-beta promoter. *Proc Natl Acad Sci U S A*, 101, 17264-9.
- Andres, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G. and Nielsen, R.. (2009). Targets of Balancing Selection in the Human Genome. *Mol. Biol. Evol.* 26, 2755–2764.
- Antin, J.H., H.J. Weinstein, E.C. Guinan, P. McCarthy, B.E. Bierer, D.G. Gilliland, S.K. Parsons, K.K. Ballen, I.J. Rimm, G. Falzarano, D.C. Bloedow, L. Abate, M. Lebsack, S.J. Burakoff, and J.L.M. Ferrara. (1994). Recombinant human interleukin-1 receptor antagonist in the treatment of steroid-resistant graft-versus-host disease. *Blood* 84, 1342-1348.
- Appay, V., Van Lier, R. A. W., Sallusto, F. and Roederer, M. (2008). Phenotype and function of human T lymphocyte subsets: Consensus and issues. *Cytom. Part A.* 73, 975–983.
- Auffray, C. Sieweke, M. H. and Geissman, F. (2009). Blood monocytes development, heterogeneity, and relationship with dendritic cells. *Annu. Rev. Immunol.* 27, 669-692.
- Auron, P.E., A.C. Webb, L.J. Rosenwasser, S.F. Mucci, A. Rich, S.M. Wolff, and C.A. Dinarello. (1984). Nucleotide sequence of human monocyte interleukin-1 precursor cDNA. *PNAS-Biological Sciences* 81, 7907-7911.
- Bagaitkar, J., Demuth, D.R., and Scott, D.A. (2008). Tobacco use increases susceptibility to bacterial infection. *Tob. Induc. Dis.* 4, 1–10.
- Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M. et al.. (2016). Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* 165, 551–565
- Banchereau, R., Baldwin, N., Cepika, A.-M., Athale, S., Xue, Y., Yu, C.I., Metang, P., Cheruku, A., Berthier, I., Gayet, I., et al. (2014). Transcriptional specialization of human dendritic cell subsets in response to microbial vaccines. *Nat. Commun.* 5, 5283.
- Baranovski, B. M., Freixo-Lima, G. S., Lewis, E. C. and Rider, P. (2015). T Helper Subsets, Peripheral Plasticity, and the Acute Phase Protein,  $\alpha$  1-Antitrypsin. *Biomed Res. Int.*
- Barreiro, L. B. and Quintana-Murci, L. (2008) From evolutionary genetics to human

- immunology: how selection shapes host defence genes. *Nat. Rev. Gen.* 11, 17-30.
- Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5, e1000562.
- Bashirova, A.A., Martin, M. P., McVicar, D. W. and Carrington, M. (2006). The killer immunoglobulin-like receptor gene cluster : tuning the genome for defense. *7*, 277-300.
- Becattini, S., Latorre, D., Mele, F., Foglierini, M., De Gregorio, C., Cassotta, A., Fernandez, B., Kelderman, S., Schumacher, T.N., Corti, D., et al. (2015). T cell immunity. Functional heterogeneity of human memory CD4+ T cell clones primed by pathogens or vaccines. *Science* 347, 400–406.
- Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K. (2010). The genome-wide structure of the Jewish people. *Nature* 466, 238-243.
- Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M. F., Karl Skorecki, K., and Richard Villems, R., Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.
- Bevan, M. J. (1976). Minor H antigens introduced on H-2 different stimulating cells cross-react at the cytotoxic T cell level during *in vivo* priming. *J. Immunol.* 117, 2233-2238
- Boeckh, M., and Geballe, A.P. (2011). Science in medicine Cytomegalovirus: pathogen, paradigm, and puzzle. *J. Clin. Invest.* 121, 1673–1680.
- Boehm, T. and Bleul, C. C. (2007). The evolutionary history of lymphoid organs. *Nat. Immunol.* 8, 131–135.
- Boehm, T. and Swann, J. B. (2013). Thymus involution and regeneration: two sides of the same coin? *Nat. Rev. Immunol.* 13, 831–8.
- Boregaard N. (2010). Neutrophils, from marrow to microbes. *Immunity.* 33, 657-670.
- Boruchov, A.M., Heller, G., Veri, M., Bonvini, E., Ravetch, J. V, and Young, J.W. (2005). Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *J. Clin. Invest.* 115, 2914–2923.
- Bot, A. (2014). Innate immunity in normal and pathologic circumstances. *Int. Rev. Immunol.* 33, 441–442.
- Bradski, G., and Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library.*

- Brinkmann, V., Reichard, U., Goosmann, C., Fauler, B., Uhlemann, Y., Weiss, D. S., Weinrauch, Y. and Zychlinsky A. (2004). *Science* 303, 1532-1535.
- Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C.J.L., Furman, D., Shen-Orr, S., Dekker, C.L., Swan, G.E., Butte, A.J., et al. (2015). Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* 160, 37–47.
- Brown, G. D. and Gordon S. (2001) Immune recognition. A new receptor for beta-glucans. *Nature* 413:36–37
- von Bubnoff, D., Andrès, E., Hentges, F., Bieber, T., Michel, T., and Zimmer, J. (2010). Natural killer cells in atopic and autoimmune diseases of the skin. *J Allergy Clin Immunol* 125, 60–68.
- Buchholz, V.R., Neuenhahn, M., and Busch, D.H. (2011). CD8+ T cell differentiation in the aging immune system: until the last clone standing. *Curr. Opin. Immunol.* 23, 549–554.
- Buckwalter, M. R. and Albert, M.L. (2009). Orchestration of the immune response by dendritic cells. *Curr. Biol.* 19. 355-361.
- Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* 2, 121–167.
- Cao, X. (2016). Self-regulation and cross-regulation of pattern-recognition receptor signaling in health and disease. *Nat. Publ. Gr.* 16, 35–50
- Casanova, J., and Abel, L. (2005). Inborn errors of immunity to infection: the rule rather than the exception. *J. Exp. Med.* 202, 197-201.
- Casanova, J., and Abel, L. (2015). Disentangling Inborn and Acquired Immunity in Human Twins. *Cell* 160, 13–15.
- Cassard, L., Jönsson, F., Arnaud, S., and Daëron, M. (2012). Fc $\gamma$  Receptors Inhibit Mouse and Human Basophil Activation. *J. Immunol.* 189, 1–13.
- Cassatella, M.A., L. Meda, S. Bonora, M. Ceska, and G. Constantin. (1993). Interleukin-10 (IL-10) inhibits the release of proinflammatory cytokines from human polymorphonuclear leukocytes - evidence for an autocrine role of tumor-necrosis-factor and IL-1 beta in mediating the production of IL-8 triggered by lipopolysaccharide. *J Exp Med* 178, 2207-2211.
- Chamaillard, M., Hashimoto, M., Horie, Y., Masumoto, J., Qiu, S., et al. (2003). An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. *Nat. Immunol.* 4, 702–707.
- Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* TIST 2.

- Chaussabel, D., Pascual, V., and Banchereau, J. (2010). Assessing the human immune system through blood transcriptomics. *BMC Biol.* 8, 84.
- Chaussabel, D. and Baldwin, N. (2014). Democratizing systems immunology with modular transcriptional repertoire analyses. *Nat. Rev. Immunol.* 14, 271–80.
- Chehimi, J. et al. Dendritic cells and IFN- $\alpha$ -producing cells are two functionally distinct non-B, non-monocytic HLA-DR<sup>+</sup> cell subsets in human peripheral blood. *Immunology* 68, 486–90 (1989).
- Chen, W., Jin, N., Hardegen, N., Lei, K., Li, L., Marinos, N., McGrady, G. and Wahl, S.M. (2003). Conversion of peripheral CD4<sup>+</sup>CD25<sup>-</sup> naive T cells to CD4<sup>+</sup>CD25<sup>+</sup> regulatory T cells by TGF- $\beta$  induction of transcription factor foxp3. *J Exp Med.* 12, 1875–1886.
- Chen, K.W., Groß, C.J., Sotomayor, F.V., Stacey, K.J., Tschopp, J., Sweet, M.J., and Schroder, K. (2014). The neutrophil NLRC4 inflammasome selectively promotes IL-1 $\beta$  maturation without pyroptosis during acute Salmonella challenge. *Cell Rep.* 8, 570–582.
- Chensue, S.W., P.D. Terebuh, D.G. Remick, W.E. Scales, and S.L. Kunkel. (1991). In vivo biologic and immunohistochemical analysis of interleukin-1 alpha, beta, and tumor-necrosis-factor during experimental endotoxemia - kinetics, Kupffer cell expression, and glucocorticoid effects. *American Journal of Pathology* 138:395-402.
- Chin, J., and M.J. Kostura. (1993). Dissociation of IL-1 beta synthesis and secretion in human blood monocytes stimulated with bacterial cell wall products. *Journal of Immunology* 151:5574-5585.
- Cognasse, F., Nguyen, K. A., Damien, P., McNicol, A., Pozzetto, B. Hamzeh-Cognasse, H., and Garraud, O. (2015). The inflammatory role of platelets via their TLRs and Siglec receptors. *Front. Immun.* 6, 1-10.
- Cooper, S. and Shedden, K. (2003). Microarray analysis of gene expression during the cell cycle. *Cell & Chromosome*, 2:1.
- Crucian, B. E. and Sams, C. F. (1999). Whole blood activation results in enhanced detection of T cell and monocyte cytokine production by flow cytometry. *NASA Johnson Sp. Cent.*
- De Rosa SC, Herzenberg LA, Roederer M. (2001). 11-color, 13-parameter flow cytometry: Identification of human naive T cells by phenotype, function, and T-cell receptor diversity. *Nat Med.* 7, 245–248.
- Debre, P. and Forster, E. (2000). *Louis Pasteur*. Johns Hopkins University Press.
- Deenadayalan, A., Maddineni, P. and Raja, A. (2013). Comparison of whole blood and

PBMC assays for T-cell functional analysis. BMC Res. Notes 2–6 .

- Deforge, L.E., and D.G. Remick. (1991). Kinetics of TNF, IL-6, and IL-8 gene expression in LPS-stimulated human whole blood. *Biochemical and Biophysical Research Communications* 174, 18-24.
- DeForge, L.E., J.S. Kenney, M.L. Jones, J.S. Warren, and D.G. Remick. (1992). Biphasic production of IL-8 in lipopolysaccharide (LPS)-stimulated human whole blood. Separation of LPS- and cytokine-stimulated components using anti-tumor necrosis factor and anti-IL-1 antibodies. *J Immunol* 148, 2133-2141.
- Delaneau, O., Zagury, J.-F. and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*. 10, 5–6.
- Devaux, B., Wilson, K.J., Aguilar, B., Jorgensen, B., and Rothbard, J.B. (1995). Differential stability of HLA-DR alleles independent of endogenous peptides. *J. Immunol.* 155, 1921–1929.
- Didonato, J. A., Mercurio, F. and Karin, M. (2012). NF- $\kappa$ B and the link between inflammation and cancer. *Immunol. Rev.* 246, 379–400.
- Diebold, S.S., Kaisho, T., Hemmi, H., Akira, S., and Reis e Sousa, C. (2004). Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* 303, 1529–1531.
- Dinarello, C.A., Ikejima, T., Warner, S.J., Orencole, S.F., Lonnemann, G., Cannon, J.G., and Libby, P. (1987). Interleukin 1 induces interleukin 1. I. Induction of circulating interleukin 1 in rabbits in vivo and in human mononuclear cells in vitro. *J. Immunol. Baltim. Md* 1950 139, 1902–1910.
- Dubravec, D.B., D.R. Spriggs, J.A. Mannick, and M.L. Rodrick. (1990). Circulating human peripheral blood granulocytes synthesize and secrete tumor necrosis factor-alpha. *Proceedings of the National Academy of Sciences of the United States of America* 87:6758-6761.
- Duffy, D., Rouilly, V., Libri, V., Hasan, M., Beitz, B., David, M., Urrutia, A., Bisiaux, A., Labrie, S.T., Dubois, A., et al. (2014). Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. *Immunity* 40, 436–450.
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinforma. Oxf. Engl.* 21, 3439–3440.
- Dusseaux, M., Martin, E., Serriari, N., Pe, I., Premel, V., Louis, D., Milder, M., Bourhis, L.

- Le, Soudais, C., Treiner, E., et al. (2011). Human MAIT cells are xenobiotic-resistant, tissue-targeted, CD161hi IL-17-secreting T cells. *Blood* 117, 1250–1260.
- Eberl, G., Colonna, M., Di Santo, J. P., McKenzie, A. N. J. (2015). Innate lymphoid cells: A new paradigm in immunology. *Science* 348, 879-888.
- Eberl, G. (2016). Immunity by equilibrium. *Nat. Rev. Immunol.* 16, 524-532.
- Editorial, *Nature Medicine*. (2013). Of men, not mice. *Nat. Med.* 19, 379.
- Farrar, M. & Schreiber, R. (1993). The molecular cell biology of interferon-gamma and its receptor. *Annu. Rev. Immunol.* 11, 571–611.
- Faustman, D. L. and Davis, M. (2013). TNF receptor 2 and disease: autoimmunity and regenerative medicine. *Front. Immunol.* 4, 1-8.
- Fernandezbotran, R. (1991). Soluble cytokine receptors: Their role in immunoregulation. *Faseb Journal* 5:2567-2574.
- Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* 42, D749–D755.
- Folds, J.D., and Schmitz, J.L. (2003). 24. Clinical and laboratory assessment of immunity. *J. Allergy Clin. Immunol.* 111, S702–S711.
- Fontes, M. (2012). Statistical and Knowledge Supported Visualization of Multivariate Data. In *Analysis for Science, Engineering and Beyond*, K. Åström, L.-E. Persson, and S.D. Silvestrov, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 143–173.
- Forrester, J. S. & Bick-Forrester, J. (2005). Persistence of inflammatory cytokines cause a spectrum of chronic progressive diseases: Implications for therapy. *Med. Hypotheses* 65, 227–231.
- Forrester, J. S. & Libby, P. (2007). The Inflammation Hypothesis and Its Potential Relevance to Statin Therapy. *Am. J. Cardiol.* 99, 732–738.
- Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G. P., Riva, S., Clerici, M., Bresolin, N. and Sironi, M. (2009). Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *J. Exp. Med.* 206, 1395-08
- Furman, D., Jojic, V., Sharma, S., Shen-Orr, S.S., Angel, C.J.L., Onengut-Gumuscu, S., Kidd, B.A., Maecker, H.T., Concannon, P., Dekker, C.L., et al. (2015). Cytomegalovirus infection enhances the immune response to influenza. *Sci. Transl. Med.* 7, 281ra43.
- Gabay, C. and Kushner, I. (1999). Acute-phase proteins and other systemic responses to inflammation. *N. Engl. J. Med.* 6, 448-454.
- Gantner, B.N., Simmons, R.M., Canavera, S.J., Akira, S., and Underhill, D.M. (2003).



- Collaborative induction of inflammatory responses by dectin-1 and Toll-like receptor 2. *J. Exp. Med.* 197, 1107–1117.
- Garraud, O., Borhis, G., Badr, G., Degrelle, S., Pozetto, B., Cognasse, F. and Richard, Y. (2012). Revisiting the B-cell compartment in mouse and humans: more than one B-cell subset exists in the marginal zone and beyond. *BMC Immunol.* 13, 63.
- Garris, C.S., Blaho, V.A., Hla, T., and Han, M.H. (2014). Sphingosine-1-phosphate receptor 1 signaling in T cells : trafficking and beyond. *Immunology* 142, 347–353.
- Gawad, C., Koh, W. and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nat. Rev. Gen.* 17, 175-188.
- Gay, N.J., Symmons, M.F., Gangloff, M., and Bryant, C.E. (2014). Assembly and localization of Toll-like receptor signaling complexes. *Nat. Rev. Immunol.* 14, 546–558.
- Geiger, R., Duhon, T., Lanzavecchia, A., and Sallusto, F. (2009). Human naive and memory CD4+ T cell repertoires specific for naturally processed antigens analyzed using libraries of amplified T cells. *J. Exp. Med.* 206, 1525–1534.
- Gibbons, L.J., and Hyrich, K.L. (2009). Biologic therapy for rheumatoid arthritis: clinical efficacy and predictors of response. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.* 23, 111–124.
- Gilfillan, A. M. and Beaven, M. A. (2011). Regulation of mast cell responses in health and disease. *Crit. Rev. Immunol.* 31, 475-529
- Girardin, S. E., Boneca, I. G., Carneiro, L. A. M., Antignac, A., J'éhanno, M., et al. (2003). Nod1 detects a unique muropeptide from gram-negative bacterial peptidoglycan. *Science* 300, 1584–1587.
- Giri, J.G., R.C. Newton, and R. Horuk. (1990). Identification of soluble interleukin-1 binding protein in cell-free supernatants: Evidence for soluble interleukin-1 receptor. *Journal of Biological Chemistry* 265, 17416-17419.
- Gorenshteyn, D., Zaslavsky, E., Fribourg, M., Park, C.Y., Wong, A.K., Tadych, A., Hartmann, B.M., Albrecht, R.A., García-Sastre, A., Kleinstein, S.H., et al. (2015). Interactive Big Data Resource to Elucidate Human Immune Pathways and Diseases. *Immunity* 43, 605–614.
- Granowitz, E.V., B.D. Clark, J. Mancilla, and C.A. Dinarello. (1991). Interleukin-1 receptor antagonist competitively inhibits the binding of interleukin-1 to the type-II interleukin-1 receptor. *Journal of Biological Chemistry* 266, 14147-14150.
- Griffiths, S.J., Riddell, N.E., Masters, J., Libri, V., Henson, S.M., Wertheimer, A., Sims, S., Rivino, L., Larbi, A., David, M., et al. (2013). Age-Associated Increase of Low-

- Avidity Cytomegalovirus-Specific CD8<sup>+</sup> T Cells That Re-Express CD45RA. *J. Immunol.* 190, 5363–5372.
- Hardison S. E. and Brown G. D. (2012). C-type lectin receptors orchestrate antifungal immunity. *Nat Immunol.* 13, 817–822.
- van der Harst, P., Zhang, W., Leach, I.M., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., and Allayee, H. (2012). Seventy-five genetic loci influencing the human red blood cell. *Nature* 492, 369–375.
- Hasan, M., Beitz, B., Rouilly, V., Libri, V., Urrutia, A., Duffy, D., Cassard, L., Di Santo, J.P., Mottez, E., Quintana-Murci, L., et al. (2015). Semi-automated and standardized cytometric procedures for multi-panel and multi-parametric whole blood immunophenotyping. *Clin. Immunol.* 261–276.
- Hawn, T.R., Verbon, A., Lettinga, K.D., Zhao, L.P., Li, S.S., Laws, R.J., Skerrett, S.J., Beutler, B., Schroeder, L., Nachman, A., et al. (2003). A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. *J. Exp. Med.* 198, 1563–1572.
- Hayashi, F., Smith, K.D., Ozinsky, A., Hawn, T.R., Yi, E.C., Goodlett, D.R., Eng, J.K., Akira, S., Underhill, D.M., and Aderem, A. (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* 410, 1099–1103.
- Hayashi, F., Means, T.K., and Luster, A.D. (2003). Toll-like receptors stimulate human neutrophil function. *Blood* 102, 2660–2669.
- Heil, F., Hemmi, H., Hochrein, H., Ampenberger, F., Kirschning, C., Akira, S., Lipford, G., Wagner, H. and Bauer, S. (2004). Species-specific recognition of single-stranded RNA via Toll-like receptor 7 and 8. *Science* 303, 1526–1529
- Hemmi, H., Takeuchi, O., Kawai, T., Kaisho, T., Sato, S., Sanjo, H., Matsumoto, M., Hoshino, K., Wagner, H., Takeda, K., and Akira, S. (2000). A Toll-like receptor recognizes bacterial DNA. *Nature* 408, 740–745.
- Hemmi, H., Kaisho, T., Takeuchi, O., Sato, S., Sanjo, H., Hoshino, K., Horiuchi, T., Tomizawa, H., Takeda, K., and Akira, S. (2002). Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nat. Immunol.* 3, 196–200.
- Howie, B.N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* 5, e1000529.
- Hu, X., Deutsch, A.J., Lenz, T.L., Onengut-gumuscu, S., Han, B., Chen, W., Howson,

- J.M.M., Todd, J.A., Bakker, P.I.W. De, Rich, S.S., et al. (2015). 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.
- Ingersoll, M. A., Spanbroek, R., Lottaz, C., Gautier, E. L., Frankenberger, M., Hoffmann, R., Lang, R., Haniffa, M., Collin, M., Tacke, F., Habenicht, A. J. R., Ziegler-Heitbrock, L., and Randolph, G. J. (2010). Comparison of gene expression profiles between human and mouse monocyte subsets. *Blood* 115, 10–20.
- Ivashkiv, L. B. & Donlin, L. T. (2013). Regulation of type I interferon responses. *Nat. Rev. Immunol.* 14, 36–49.
- Janeway, C. A. Jr. (1989). Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* 54, 1-13.
- Jia, X., Han, B., Onengut-gumuscu, S., Chen, W., Concannon, P.J., Rich, S.S., Raychaudhuri, S., and de Bakker, P.I.W. (2013). Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* 8, e64683.
- Johnson, W.E., and Li, C. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127.
- Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347, 1259038.
- Kang, D. C., Gopalkrishnan, R. V., Wu, Q., Jankowsky, E., Pyle, A. M. and Fisher, P. B. (2002). mda-5 : An interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. *Proc. Natl. Acad. Sci. U. S. A.*, 99, 637-642.
- Kato, H., Takeuchi, O., Sato, S., Yoneyama, M., Masahiro Yamamoto, M., Matsui, K. and al.. (2006). Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* 441, 101–105.
- Kearley, J., Silver, J.S., Sanden, C., Liu, Z., Berlin, A.A., White, N., Mori, M., Kolbeck, R., and Humbles, A.A. (2015). Cigarette Smoke Silences Innate Lymphoid Cell Function and Facilitates an Exacerbated Type I Interleukin-33-Dependent Response to Infection. *Immunity* 42, 566–579.
- Kelso, A. (1998). Cytokines : Principles and prospects. *Immunol. and Cell Biol.* 76, 300-317.
- Kitchens, R. L., Thompson, P. a, Munford, R. S. & O’Keefe, G. E. (2003). Acute inflammation and infection maintain circulating phospholipid levels and enhance

- lipopolysaccharide binding to plasma lipoproteins. *J. Lipid Res.* 44, 2339–2348
- Kolaczowska, E., and Kubes, P. (2013). Neutrophil recruitment and function in health and inflammation. *Nat. Rev. Immunol.* 13, 159–175.
- Krieg, A.M. (2002). CpG motifs in bacterial DNA and their immune effects. *Annu. Rev. Immunol.* 20, 709–760.
- Kruppa, G., B. Thoma, T. Machleidt, K. Wiegmann, and M. Kronke. (1992). Inhibition of tumor-necrosis-factor (TNF)-mediated NF-kappa-B activation by selective blockade of the human 55-KDA TNF receptor. *Journal of Immunology* 148:3152-3157.
- Lang, C.H., C. Silvis, N. Deshpande, G. Nystrom, and R.A. Frost. (2003). Endotoxin stimulates in vivo expression of inflammatory cytokines tumor necrosis factor alpha, interleukin-1 beta,-6, and high-mobility-group protein-1 in skeletal muscle. *Shock* 19:538-546.
- Latz, E., Schoenemeyer, A., Visintin, A., Fitzgerald, K.A., Monks, B.G., Knetter, C.F., Lien, E., Nilsen, N.J., Espevik, T., and Golenbock, D.T. (2004). TLR9 signals after translocating from the ER to CpG DNA in the lysosome. *Nat. Immunol.* 5, 190–198.
- Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343, 1246980.
- Li, B., Cramer, D., Wagner, S., Hansen, R., King, C., Kakar, S., Ding, C., and Yan, J. (2007). Yeast glucan particles activate murine resident macrophages to secrete proinflammatory cytokines via MyD88- and Syk kinase-dependent pathways. *Clin. Immunol. Orlando Fla* 124, 170–181.
- Li, S., Roupheal, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., et al. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* 15, 195–204.
- Libri, V., Azevedo, R.I., Jackson, S.E., Di Mitri, D., Lachmann, R., Vukmanovic-, M., Yong, K., Kern, F., Soares, M.V.D., and Akbar, A.N. (2011). Cytomegalovirus infection induces the accumulation of short-lived , multifunctional CD4+ CD45RA+ CD27- T cells: the potential involvement of interleukin-7 in this process. *Immunology* 132, 326–339.
- Locksley, R.M., Killeen, N. and Lenardo M.J. (2001). The TNF and TNF receptor superfamilies: Integrating mammalian biology. *Cell* 104:487-501.
- Macallan, D.C., Wallace, D.L., Zhang, Y., Ghattas, H., Asquith, B., Lara, C. De, Worth, A.,

- Panayiotakopoulos, G., Griffin, G.E., Tough, D.F., et al. (2005). B-cell kinetics in humans: rapid turnover of peripheral blood memory cells. *Blood* 105, 3633–3641.
- Martin, N.G., Eaves, L.J., Kearsley, M.J., and Davies, P. (1978). The power of the classical twin study. *Heredity (Edinb)*. 40, 97–116.
- Mackay, F., H. Loetscher, D. Stueber, G. Gehr, and Lesslauer W. (1993). Tumor-necrosis-factor alpha (TNF-ALPHA)-induced cell adhesion to human endothelial cells is under dominant control of one TNF receptor type TNF-r55. *Journal of Experimental Medicine* 177, 1277-1286.
- Manichaikul, A., Mychaleckyj, J.C., Rich, S. S., Daly, K., Sale, M. and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873.
- Mantovani, A., Cassatella, M. A. & Costantini, C. (2011). Neutrophils in the activation and regulation of innate and adaptive immunity. *Nat. Publ. Gr.* 11, 519–531
- Mantovani, A. and Garlanda, C. (2013). Platelet-macrophage partnership in innate immunity and inflammation. *Nat. Publ. Gr.* 14, 768–770
- Martinon F, Burns K, Tschopp J. (2002). The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. *Mol Cell.* 2, 417–426.
- Martinon, F., Pétrilli, V., Mayor, A., Tardivel, A., and Tschopp, J. (2006). Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* 440, 237–241.
- McInnes, I.B., and Schett, G. (2007). Cytokines in the pathogenesis of rheumatoid arthritis. *Nat. Rev. Immunol.* 7, 429–442.
- Medzhitov, R., and C.A. Janeway. (1997). Innate immunity: The virtues of a nonclonal system of recognition. *Cell* 91, 295-298.
- Medzhitov, R. (2008). Origin and physiological roles of inflammation. *Nature* 454, 428-435.
- Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B* 72, 417–473.
- Miller, S. I., Ernst, R. K. and Bader, M. W. (2005). LPS, TLR4 and infectious disease diversity. *Nat Rev Microbiol.* 3, 36–46.
- Mohler, K.M., D.S. Torrance, C.A. Smith, R.G. Goodwin, K.E. Stremmer, V.P. Fung, H. Madani, and M.B. Widmer. (1993). Soluble tumor-necrosis-factor (TNF) receptors are effective therapeutic agents in lethal endotoxemia and function simultaneously as both TNF carriers and TNF antagonists. *J. Immun.* 151, 1548-1561.
- Monticelli, L.A., Sonnenberg, G.F., Abt, M.C., Alenghat, T., Ziegler, C.G.K., Doering, T.A.,

- Angelosanto, J.M., Laidlaw, B.J., Yang, C.Y., Sathaliyawala, T., et al. (2011). Articles  
Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza  
virus. *Nat. Immunol.* 12, 1045–1055.
- Morris, J., Williams, N., Rush, C., Govan, B., Sangla, K., Norton, R., and Ketheesan, N.  
(2012). *Burkholderia pseudomallei* triggers altered inflammatory profiles in a whole-  
blood model of type 2 diabetes-melioidosis comorbidity. *Infection and Immunity* 80,  
2089–2099.
- Morvan, M. G. and Lanier, L. L. (2015). NK cells and cancer: you can teach innate cells new  
tricks. *Nat Rev Cancer* 16, 7–19.
- Mosmann, T. R. and Coffman, R. L. (1989). Th1 AND Th2 CELLS: Different Patterns of  
Lymphokine Secretion Lead to Different Functional Properties. *Ann. Rev. Immunol* 7,  
145–173
- Motohashi, T., Miyoshi, S., Osawa, M., Eyre, H.J., Sutherland, G.R., Matsuda, Y., Nakamura,  
Y., Shibuya, A., Iwama, A., and Nakauchi, H. (2000). Molecular Cloning and  
Chromosomal Mapping of a Novel Protein Gene, M83. *Biochem Biophys Res  
Commun* 250, 244–250.
- Netea, M. G., van de Veerdonk, F. L., van der Meer, J. W. M., Dinarello, C. A. and Joosten,  
L. A. B. (2015). Inflammasome-Independent Regulation of IL-1-Family Cytokines.  
*Annu. Rev. Immunol.* 33, 49–77.
- Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. and Corcoran, L. M. (2015). The generation of  
antibody-secreting plasma cells. *Nat. Rev. Immunol.* 15, 160–171.
- Orme, I.M., Robinson, R.T., and Cooper, A.M. (2015). The balance between protective and  
pathogenic immune responses in the TB-infected lung. *Nat. Immunol.* 16, 57–63.
- Orrù, V., Steri, M., Sole, G., Sidore, C., Viridis, F., Dei, M., Lai, S., Zoledziewska, M.,  
Busonero, F., Mulas, A., et al. (2013). Genetic variants regulating immune cell levels  
in health and disease. *Cell* 155, 242–256.
- Ozinsky, A., Underhill, D. M., Fontenot, J. D., Hajjar, A. M., Smith, K. D., Wilson, C. B.,  
Schroeder, L. and Aderem, A. (2000). The repertoire for pattern recognition of  
pathogens by the innate immune system is defined by cooperation between Toll-like  
receptors. *PNAS* 97, 13766–71.
- Parham, P. (2005). MHC Class I molecules and KIRs in human history, health and survival.  
*Nat. Rev. Immun.* 5, 201-214.
- Park, S., and Rehermann, B. (2014). Review Immune Responses to HCV and Other Hepatitis  
Viruses. *Immunity* 40, 13–24.

- Paternoster, L., Standl, M., Waage, J., Baurecht, H., Hotze, M., Strachan, D.P., and Curtin, J.A. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* 47.
- Patterson, N., Price, A. L. and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet.* 2, e190.
- Patterson, S., Moran, P., Epel, E., Sinclair, E., Kemeny, M.E., Deeks, S.G., Bacchetti, P., Acree, M., Epling, L., Kirschbaum, C., et al. (2013). Cortisol Patterns Are Associated with T Cell Activation in HIV. *PLoS One* 8, e63429.
- Pellegrini, M., Bouillet, P., Robati, M., Belz, G.T., Davey, G.M., and Strasser, A. (2004). Loss of Bim Increases T Cell Production and Function in Interleukin 7 Receptor-deficient Mice. *J. Exp. Med.* 200, 1189–1195.
- Pennell, L.M., Galligan, C.L., and Fish, E.N. (2012). Sex affects immunity. *J. Autoimmun.* 38, J282–J291.
- Perussia, B. Fanning, V. and Trinchieri, G.(1985). A leukocyte subset bearing HLA-DR antigens is responsible for in vitro alpha interferon production in response to viruses. *Nat. Immun. Cell Growth Regul.* 4, 120-137.
- Pichlmair, A., Schulz, O., Tan, C. P., Naslund, T. I., Lijestrom., P., Weber, F. and Reis e Sousa, C. (2006). RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. *Science* 314, 997-1001.
- Platanias, L. C. (2005). Mechanisms of type-I and type-II interferon-mediated signaling. *Nat. Rev. Immunol.* 5, 375-386.
- Plato, A., Willment, J. and Brown, G. D. (2013). C-type lectin-like receptors of the dectin-1 cluster: ligands and signaling pathways. *Int. Rev. Immunol.* 32, 134–56
- Philpott, D. J., Sorbara, M. T., Robertson, S. J., Croitoru, K. and Girardin, S. E. (2014). NOD proteins: regulators of inflammation in health and disease. *Nat. Rev. Immunol.* 14, 9–23
- Poltorak A, He X, Smirnova I, LiuMY, Van Huffel C, Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C., Freudenberg, M., Ricciardi-Castagnoli, P., Layton, B. and Beutler, B. (1998). Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science* 282, 2085–88
- Prugnotte, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V. and Balloux F. (2005). Pathogen-Driven Selection and Worldwide HLA Class I Diversity. *Current Biology* 15, 1022–1027.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J.,

- Sklar, P., de Bakker, P. I. W., Daly, M. J. and Sham P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 59–575.
- Quintana-Murci, L. and Clark, A. G. (2013). Population genetic tools for dissecting innate immunity in humans. *Nat Rev. Immunol.* 13, 280-293.
- Rawlings, D. J., Sommer, K., Moreno-Garcia, M. E. (2006). The CARMA1 signalosome links the signaling machinery of adaptive and innate immunity in lymphocytes. *Nat Rev Immunol* 6, 799–812.
- Reichert, T., DeBruyère, M., Deneys, V., Tötterman, T., Lydyard, P., Yuksel, F., Chapel, H., Jewell, D., Van Hove, L., Linden, J., et al. (1991). Lymphocyte Subset Reference Ranges in Adult Caucasians. *Clin. Immunol. Immunopathol.* 60, 190–208.
- Reis e Sousa, C. (2006). Dendritic cells in a mature age. *Nat. Rev. Immunol.* 6, 476-483.
- Rocha, B. & Tanchot, C. CD8 T cell memory. *Semin. Immunol.* 16, 305–314 (2004).
- Roederer, M., Quaye, L., Mangino, M., Beddall, M.H., Mahnke, Y., Chattopadhyay, P., Tosi, I., Napolitano, L., Terranova Barberio, M., Menni, C., et al. (2015). The Genetic Architecture of the Human Immune System: A Bioresource for Autoimmunity and Disease Pathogenesis. *Cell* 161, 387–403.
- Rogers, N.C., Slack, E.C., Edwards, A.D., Nolte, M.A., Schulz, O., Schweighoffer, E., Williams, D.L., Gordon, S., Tybulewicz, V.L., Brown, G.D., and Reis, E.S.C. (2005). Syk-dependent cytokine induction by Dectin-1 reveals a novel pattern recognition pathway for C type lectins. *Immunity* 22, 507–517.
- Rosenberg, H. F., Dyer, K. D. & Foster, P. S. Eosinophils: Changing perspectives in health and disease. *Nat. Rev. Immunol.* 13, 9–22 (2013).
- Sallusto, F., Geginat, J., Lanzavecchia, A. (2004). Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu. Rev. Immunol.* 22, 745-763.
- Sancho, D. and Reis e Sousa, C. (2012). Signaling by myeloid C-type lectin receptors in immunity and homeostasis. *Ann Rev Immunol.* 30, 491–529.
- Sauce, D., and Appay, V. (2011). Altered thymic activity in early life : how does it affect the immune system in young adults ? *Curr. Opin. Immunol.* 23, 543–548.
- Schulz, O., Diebold, S. S., Chen, M., Naslund, T. I., Nolte, M. A., Alexopoulou, L., Azuma, Y. T., Flavell, R. A., Liljestrom, P. & Reis e Sousa, C. (2005) Toll-like receptor 3 promotes cross-priming to virus-infected cells. *Nature* 433, 887-92.
- Seok, J. Warren, H. S., Cuenca, A. G., Mindrinosa, M. N., Bakerc, H. V., Xu W. (et al.) and the Inflammation and Host Response to Injury, Large Scale Collaborative Research



- Program. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U. S. A.* 110, 3507–12.
- Serafini, N., Vosshenrich, C.A.J., and Di Santo, J.P. (2015). Transcriptional regulation of innate lymphoid cell fate. *Nat. Rev. Immunol.* 15, 415–428.
- Serhan, C. N. (2007). Resolution phase of inflammation: novel endogenous anti-inflammatory and proresolving lipid mediators and pathways. *Annu. Rev. Immunol.* 25, 101–137.
- Shah, R.D., and Samworth, R.J. (2013). Variable selection with error control : another look at stability selection. *J. R. Stat. Soc. Ser. B* 75, 55–80.
- Shi, C. and Pamer, E. G. (2011). Monocyte recruitment during infection and inflammation. *Nat. Rev. Immunol.* 11, 762–74
- Shimazu, R., Akashi, S., Ogata, H., Nagai, Y., Fukudome, K., Miyake, K. and Kimoto, M. (1999). MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J. Exp. Med.* 189, 1777–1782
- Shin, H., Shannon, C.P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., Wilson-McManus, J.E., Ng, R.T., McManus, B.M., Tebbutt, S.J., et al. (2014). Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PloS One* 9, e91041.
- Sims, J.E., M.A. Gayle, J.L. Slack, M.R. Alderson, T.A. Bird, J.G. Giri, F. Colotta, F. Re, A. Mantovani, K. Shanebeck, K.H. Grabstein, and S.K. Dower. (1993). Interleukin-1 signaling occurs exclusively via the type-I receptor. *PNAS* 90, 6155-6159.
- Sims, J.E., J.G. Giri, and S.K. Dower. (1994). 2 Interleukin-1 receptors play different roles in IL-1 actions. *Clinical Immunology and Immunopathology* 72, 9-14.
- Sornasse, T., Larenas, P. V., Davis, K. A., deVries, J. E. and Yssel, H. (1996). Differentiation and stability of T Helper 1 and 2 cells derived from naive human neonatal CD4 + T cells. *J. Exp. Med.* 184, 473-483.
- Speake, C., Presnell, S., Domico, K., Zeitner, B., Bjork, A., Anderson, D., Mason, M.J., Whalen, E., Vargas, O., Popov, D., et al. (2015). An interactive web application for the dissemination of human systems immunology data. *J. Transl. Med.* 13, 196.
- Spriggs, D., K. Imamura, C. Rodriguez, J. Horiguchi, and D.W. Kufe. (1987). Induction of tumor-necrosis-factor expression and resistance in a human breast tumor cell line. *Proc. Natl. Acad. Sci. U. S. A.*, 84:6563-6566.
- Sridharan, A., Esposito, M., Kaushal, K., Tay, J., Osann, K., Agrawal, S., Gupta, S., and Agrawal, A. (2011). Age-associated impaired plasmacytoid dendritic cell functions lead to decreased CD4 and CD8 T cell immunity. *Age* 33, 363–376.

- Stekhoven, D.J., and Bühlmann, P. (2012). MissForest — non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118.
- Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull Acad Pol. Sci Fr.* 4, 801–804.
- Sun L., Wu, J., Du, F., Chen, X. and Chen, Z. J. (2013). Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* 339, 786–91.
- Suzuki-Inoue, K., Kato, Y, Inoue, O., Kaneko, M. K., Mishima, K., Yatomi, Y., Yamazaki, Y., Narimatsu, H. and Ozaki, Y. (2007). Involvement of the snake toxin receptor CLEC-2, in podoplanin-mediated platelet activation, by cancer cells. *J Biol Chem.* 282, 25993–26001.
- Svajger, U., Anderluh, M., Jeras, M. and Obermajer, N. (2010). C-type lectin DC-SIGN: an adhesion, signaling and antigen uptake molecule that guides dendritic cells in immunity. *Cell Signal*;22:1397–1405.
- Sylwester, A.W., Mitchell, B.L., Edgar, J.B., Taormina, C., Pelte, C., Ruchti, F., Sleath, P.R., Grabstein, K.H., Hosken, N.A., Kern, F., et al. (2005). Broadly targeted human cytomegalovirus- specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. *J. Exp. Med.* 202, 673–685.
- Symons, J.A., P.R. Young, and G.W. Duff. (1995). Soluble type-II interleukin-1 receptor binds and blocks processing of IL-1 beta precursor and loses affinity for IL-1 receptor antagonist. *PNAS* 92, 1714-1718.
- Swain, S. L., McKinstry, K. K. & Strutt, T. M. (2012). Expanding roles for CD4+ T cells in immunity to viruses. *Nat. Rev. Immunol.* 12, 136–148.
- Szabo, S. J., Jacobson, N. G., Dighe, A. S., Gubler, U. and Murphy, K. M. (1995). Developmental commitment to the Th2 lineage by extinction of IL-12 signaling. *Immunity* 2, 665-675.
- Takahara, K., Yashima, Y., Omatsu, Y., Yoshida, H., Kimura, Y., Kang, Y.S., Steinman, R.M., Park, C.G., and Inaba, K. (2004). Functional comparison of the mouse DC-SIGN, SIGNR1, SIGNR3 and Langerin, C-type lectins. *Int. Immunol.* 16, 819–82
- Takeuchi, O., Kawai, T., Muhlradt, P. F., Morr, M., Radolf, J. D., Zychlinsky, A., Takeda, K. and Akira, S. (2001). Discrimination of bacterial lipoproteins by Toll-like receptor 6. *Int. Immunol.* 13, 933-940.
- Tartaglia, L.A., M. Rothe, Y.F. Hu, and D.V. Goeddel. (1993). Tumor necrosis factors cytotoxic activity is signaled by the p55 TNF receptor. *Cell* 73:213-216.

- Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.-G., Fauchoux, N., Sayegrih, S., Vray, M., et al. (2015). The Milieu Intérieur study - an integrative approach for study of human immunological variance. *Clin. Immunol.* 157, 277–293
- Tollerud, D.J., Clark, J.W., Morris Brown, L., Neuland, C.Y., Pankiw-Trost, L.K., Battner, W.A., and Hoover, R.N. (1989). The Influence of Age, Race, and Gender on Peripheral Blood Mononuclear-Cell Subsets in Healthy Nonsmokers. *J. Clin. Immunol.* 9, 214–222.
- Thoma-Uszynski, S., Stenger, S., Takeuchi, O., Ochoa, M.T., Engele, M., Sieling, P.A., Barnes, P.F., Rollinghoff, M., Bolcskei, P.L., Wagner, M., et al. (2001). Induction of direct antimicrobial activity through mammalian toll-like receptors. *Science* 291, 1544–1547.
- Trinchieri, G. and Sher, A. (2007). Cooperation of Toll-like receptor signals in innate immune defence. *Nat. Rev. Immun.* 7, 179–190.
- Trinchieri, G. (2010). Type I interferon: friend or foe? *J. Exp. Med.* 207, 2053–63
- Tsang E, Giannetti, A. M., Shaw D. Dinh, M., Tse, J. K., Gandhu, S., Ho, H., Wang, S., Papp, E., and Bradsaw, J. M. (2008). Molecular mechanism of the Syk activation switch. *J. Biol. Chem.* 283, 32650–32659.
- Tsang, J.S., Schwartzberg, P.L., Kotliarov, Y., Biancotto, A., Xie, Z., Germain, R.N., Wang, E., Olnes, M.J., Narayanan, M., Golding, H., et al. (2014). Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* 157, 499–513.
- Tsuchiya, M., Asada, A., Kasahara, E., Sato, E.F., Shindo, M., and Inoue, M. (2002). Smoking a Single Cigarette Rapidly Reduces Combined Concentrations of Nitrate and Nitrite and Concentrations of Antioxidants in Plasma. *Circulation* 105, 1155–1157.
- Vassallo, R., Tamada, K., Lau, J.S., Paula, R., and Chen, L. (2005). Cigarette Smoking Extract Suppresses Human Dendritic Cell Function Leading to Preferential Induction of Th-2 Priming. *J. Immunol.* 175, 2684–2691.
- Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3, RESEARCH0034.
- Vantourout, P. and Hayday, A. (2013). Six-of-the-best: unique contributions of  $\gamma\delta$  T cells to immunology. *Nat. Rev. Immunol.* 13, 88–100.

- Venet, F., Lukaszewicz, A.-C., Payen, D., Hotchkiss, R., and Monneret, G. (2013). Monitoring the immune response in sepsis: a rational approach to administration of immunoadjuvant therapies. *Curr. Opin. Immunol.* 25, 477–483.
- Vitti, J. J., Sharon R. Grossman, S. R., and Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annu. Rev. Genet.* 47, 97–120
- Vivier, E., Raulet, D.H., Moretta, A., Caligiuri, M.A., Zitvogel, L., Lanier, L.L., Yokoyama, W.M., and Ugolini, S. (2011). Innate or Adaptive Immunity? The Example of Natural Killer Cells. *Science* (80-. ). 331, 44–49.
- Voehringer, D. (2013). Protective and pathological roles of mast cells and basophils. *Nat. Rev. Immunol.* 13, 362–75
- Vrisekoop, N., Braber, I. Den, Boer, A.B. De, Ruiter, A.F.C., Crabben, S.N. Van Der, Schrijver, E.H.R., Spierenburg, G., Sauerwein, H.P., Hazenberg, M.D., Borghans, A.M., et al. (2008). Sparse production but preferential incorporation of recently produced naïve T cells in the human peripheral pool. *Proc. Natl. Acad. Sci. U.S.A.*, 105, 6115–6120.
- Watson, R. O., Bell, S. L., MacDuff, D. A., Kimmey, J. M., Diner, E. J., Olivas, J., Vance, R. E., Stallings, C. L., Virgin, H. W. and Cox J. S. (2015). Short Article The Cytosolic Sensor cGAS Detects Mycobacterium tuberculosis DNA to Induce Type I Interferons and Activate Autophagy Short Article The Cytosolic Sensor cGAS Detects Mycobacterium tuberculosis DNA to Induce Type I Interferons and Activate Autophagy. *Cell Host Microbe* 17, 811–819.
- Wen, T., Besse, J.A., Mingler, M.K., Fulkerson, P.C., and Rothenberg, M.E. (2013). Eosinophil adoptive transfer system to directly evaluate pulmonary eosinophil trafficking in vivo. *Proc. Natl. Acad. Sci.* 110, 6067–6072.
- Wertheimer, A.M., Bennett, M.S., Park, B., Uhrlaub, J.L., Martinez, C., Pulko, V., Currier, N.L., Nikolich-Zugich, D., Kaye, J., and Nikolich-Zugich, J. (2014). Aging and Cytomegalovirus Infection Differentially and Jointly Affect Distinct Circulating T Cell Subsets in Humans. *J. Immunol.* 192, 2143–2155.
- Whitacre, C.C. (2001). Sex differences in autoimmune disease. *Nat. Immunol.* 2, 777–780.
- Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinforma. Oxf. Engl.* 26, 873–881.
- Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genetics* 46, 100-106.

- Yatim, N., Jusforgues-Saklani, H., Orozco, S., Schulz, O., Barreira de Silva, R., Reis e Sousa, C., Green, D. R., Oberst, A. and Albert, M. L. (2015). RipK1 and NF- $\kappa$ B signaling in dying cells determines cross-priming of CD8<sup>+</sup> T cells. *Science* 350, 328-334.
- Yoneyama, M., Jogi, M. and Onomoto, K. (2015). Regulation of antiviral innate immune signaling by stress-induced RNA granules. *J. Biochem.* 159, 279–286.
- Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11, 407–409.
- Zúñiga-Pflücker, J. C. (2004). T-cell development made simple. *Nat. Rev. Immunol.* 4, 67-72.

## **Résumé :**

Le projet *Milieu Intérieur* a pour but d'identifier quels sont les facteurs génétiques et environnementaux qui ont un impact sur la variabilité immunitaire naturelle à l'échelle d'une population. Cette analyse multiparamétrique requière néanmoins d'utiliser des outils standardisés

Afin d'étudier la réponse immune induite, nous avons utilisé un système optimisé de seringue prête à l'emploi pour une stimulation *ex vivo* du sang et développé un protocole unique de quantification de l'ARN afin d'étudier la signature transcriptionnelle en réponse à des immuno-modulateurs. Par cette analyse nous avons testé l'hypothèse que la réponse à des composants complexes peut être définie par la signature ARN de cytokines clés. En utilisant une méthode statistique robuste, nous avons identifié 44 gènes capables d'optimiser la capture de la réponse à des stimulations plus complexes. Cette approche pourrait aider à la réduction dimensionnelle de larges données et la décomposition de réponses immunes innées et ainsi caractériser de nouvelles molécules immuno-modulatrices.

Dans une seconde étude, nous avons cherché à identifier les facteurs génétiques et environnementaux influençant le phénotype des cellules immunitaires circulantes. Pour cela, nous avons associé une étude semi-automatisée par cytométrie en flux des cellules du sang à une analyse du génotype pour les 1,000 donateurs inclus dans la cohorte. Nous avons observé que le tabac, l'âge, le genre et l'infection latente par le cytomégalovirus sont les facteurs impactant le plus la variabilité immunitaire. Cette étude a montré que les paramètres des cellules innées sont contrôlés par des facteurs génétiques alors que ceux des cellules adaptatives le sont plutôt par des expositions environnementales tout au long de la vie.

Des outils interactifs incluant ces nouvelles données de références accompagnent ces études.

L'ensemble de ces analyses montre que nous avons développé des outils performants pour une étude intégrative du système humain constituant une approche innovante vers une médecine personnalisée.

**Mots clés:** Immunophénotypage, Réponse ARN aux cytokines, Dissection de réponses complexes, Cytométrie en flux, Immunité innée, Association génétique.

## **Abstract : Defining the boundaries of a healthy immune response using standardized immune monitoring tools**

The project *Milieu Intérieur* aims to study the genetic and environmental factors that can have a major impact on occurring immunological variance in healthy human population. This characterization requires the use of standardized immunophenotyping technologies for integrating diverse, complex datasets. With this goal in mind, we used an optimized suite of standardized whole-blood stimulation systems to study the human induced immune response in physiological condition and developed a unique standardized protocol to analyze the ARN signatures upon whole-blood stimulation to test the hypothesis that responses to complex stimuli can be defined by the transcriptional signatures of key cytokines. We found 44 genes, identified using Support Vector Machine learning, which captured the diversity of complex innate immune responses with improved segregation between distinct stimuli. This provides new strategies for dimension reduction of large datasets and for deconvolution of innate immune responses, applicable for characterizing novel immunomodulatory molecules.

In a second related study, we aimed to identify the environmental and genetic factors driving innate and adaptive immune cell parameters in homeostatic conditions. To do so, we combined semi-automated flow cytometric analysis of blood leukocytes and genome-wide DNA genotyping in the 1,000 healthy donors included in the collection. We show that smoking, age, gender and latent cytomegalovirus infection, are main drivers of human variation (i.e. numbers of Treg and MAIT cells). These results demonstrated that innate cell parameters are strongly controlled by genetic factors, whereas adaptive cells are driven by life-long environmental exposures.

In addition, to help on the public data mining, we developed interactive R-Shiny application including healthy donor reference values for both studies.

All together, these results indicate that we developed powerful tools for human system biology approaches to support personalized medicine.

**Keywords:** Immune phenotyping, Cytokine gene signatures, Deconvoluting complex responses, Flow cytometry, Innate immunity, Genome-wide association study