Decoding cellular communications and interactions between immune cells by using single-cell approaches

Pierre Bost

To cite this version:


HAL Id: tel-03371870
https://theses.hal.science/tel-03371870
Submitted on 9 Oct 2021

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.
Decoding cellular communications and interactions between immune cells by using single-cell approaches

By

Pierre Bost

PhD in Biology

Advisors:

Benno Schwikowski and Ido Amit

Defended on the 7th of October 2020

with the following jury:

Mr. Sten Linnarsson (Professor) Karolinska Institute Referee
Mr. Martin Guiliams (Professor) Ghent University Referee
Mrs. Sonia Garel (DR) École Normale Supérieure President of the jury
Mrs. Leïla Perié (CR) Curie Institute SU delegate
Mr. Thomas Walter (DR) Curie Institute Examiner
Mr. Benno Schwikowski (DR) Pasteur Institute Thesis supervisor

Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-nd/3.0/
# Contents

Aknowledgments vii

General introduction 1

1 Cellular immune communications ...................................... 3
1.1 Introduction to cellular communications .......................... 3
1.1.1 General principles of cellular communications ............... 3
1.1.2 Description of common cellular communication pathways ... 5
1.1.3 The importance of studying cellular communications at the single cell level ...................................................... 7
1.2 Cellular communications in immunology .......................... 9
1.2.1 The immune system is made up of various cellular components that have to communicate ................................. 9
1.2.2 Immune cells communicate through soluble factors and direct contact ................................................................. 13
1.3 Role of immune communications in health and diseases .......... 13
1.3.1 Dysregulations of immune communication trigger or fuel several diseases .......................................................... 13
1.3.2 Immunotherapy has revolutionized the treatment of various diseases ................................................................. 16
1.4 Current challenges in the study of immune communications .... 18

2 An introduction to single-cell genomic technologies ................. 21
2.1 A historical perspective of the single-cell genomic field .......... 21
2.1.1 The rise of the single-cell genomic field (2009-2013) .......... 22
2.1.2 The maturation and scaling up of the technologies (2014-2016) 24
2.1.3 The era of functional assays, multi-omics and single-cell atlases (2016-2019) ............................................................. 25
2.2 The single-cell RNA sequencing technology ........................ 28
2.2.1 MARS-seq 2.0: Example of a typical 3’ biased scRNA-seq protocol ................................................................. 28
2.2.2 Comparison of MARS-seq with other protocols .................. 31
2.3 Processing and analysis of the scRNA-seq data ..................... 33
Aknowledgments
General introduction

Contents

1 Cellular immune communications ................................ 3
  1.1 Introduction to cellular communications ....................... 3
  1.2 Cellular communications in immunology ........................ 9
  1.3 Role of immune communications in health and diseases ........ 13
  1.4 Current challenges in the study of immune communications .... 18

2 An introduction to single-cell genomic technologies ............. 21
  2.1 A historical perspective of the single-cell genomic field ...... 21
  2.2 The single-cell RNA sequencing technology ..................... 28
  2.3 Processing and analysis of the scRNA-seq data .................. 33
  2.4 Contribution of the single-cell technologies to Immunology ... 37
1 Cellular immune communications

1.1 Introduction to cellular communications

1.1.1 General principles of cellular communications

Every cells, ranging from the solitary bacteria to the highly social mammalian cells, have the ability to adapt to their surrounding and ever-changing environment. Such adaption relies on three distinct steps: the sensing of the environmental signals/changes, the signal processing and the changes of cellular processes. While unicellular organisms mostly responds to changes in local nutrient concentration, temperature or osmolarity, cells from multicellular organisms have to integrate multiple signals coming from other cells. Those cellular communications control many aspects of the cell behavior, including cell survival, division, metabolism and locomotion and are therefore an essential aspect of the multicellular organisms biology.

Communication between cells is mostly mediated by soluble factors that can either be proteins, lipids, glucids and even in some cases, gazes. Some of those molecules can operate over large distance, while others only affect direct cellular neighbors. Cells can detect those factors thanks to proteins called receptors, usually located (but not always) at the cell membrane, which can physically bind to the soluble factor. The binding of factors (also called ligands) will trigger physical or chemical changes on the receptor, a process often called ‘receptor activation’. The activated receptor will in turn activate one or more intracellular signaling pathways which then process and integrate the different signals detected by the cell to activate or inhibit specific proteins. Those proteins that lies at the end of the signaling pathways are called the ‘effector proteins’ and implement the changes needed in the cell behavior. Effector proteins can be metabolic enzymes, transcription factors or cytoskeletal proteins and will respectively alter the cell metabolism, gene expression and shape/locomotion. It is important to keep in mind that mammalian cells have the ability to respond to a huge diversity of signals: indeed thanks to genomic analysis, more than 1500 genes that encode receptors have been identified in the human genome, each gene having the ability to generate several receptors through alternative splicing and post-translational modifications.

Those receptors can detect incoming signals emitted by cells at varying distances, ranging from direct neighbor to cells from other organs. This variability is used to classify cellular signaling events into four main forms: contact-dependent, paracrine, synaptic and endocrine (Figure 2). While we previously mentioned that communication between cells is mostly mediated by soluble factors, some factors are physically bound to the membrane and only acts on the direct neighboring cells. Such contact-dependent signaling is especially important during immune responses and development. However in most of the cases, the signaling factor is soluble and diffuses around the producing cell. In such case only cells in the local environment are affected: such signaling events are called paracrine signaling and are used in various biological
context, including immune response, development, metabolism regulation and tissue repair.

While sufficient for short range communication, paracrine and contact dependant signaling are not enough for large-scale organisms such as mammalian who need fast and long-range signaling pathways. This gap is filled through the use of synaptic signaling: Specific and highly complex cells called nerve cells, or neurons extend long cellular sections (called axons) that enable them to contact target cells far away through the propagation of an electric impulse. At the very end of the axons, the electric impulse is converted into the release of soluble factors called neurotransmitters that will then activate a specific cell. A common alternative to the neuronal signaling is the endocrine signaling. In such case, specific cells, called endocrine cells, produce a signal molecule, called hormone, that is released in the bloodstream. The blood carries the hormone over long distances, allowing it to activate cells in every possible organ.

Figure 1: Example of a simple signaling pathway activated by an a soluble factor. Adapted from (Alberts, 2015)
As we have seen, evolution has granted to mammal cells highly efficient and diverse ways to communicate over both short and long distances, and therefore the ability of the cells and the whole organism to react to sudden variations in their environment.

1.1.2 Description of common cellular communication pathways

As mentioned earlier, cell signaling regulates every possible aspects of the life of a cell, including its metabolism, moves and division. An entire book would not be enough to describe all the signaling pathways that have been identified in mammals, so only essential pathways that will be mentioned and discussed later in the manuscript will hereby be described.

Among the several families of signaling receptors, enzyme coupled receptors are membranous proteins with an extracellular domain that can recognize and bind to a set of ligands, and
an intracellular domain that either possesses intrinsic enzyme activities or is physically associated with an enzyme. Within this family of receptors, several sub-families have been identified based on the type of enzyme associated with the receptor. One of the most common family is the Receptor Tyrosine Kinase (RTK) family, i.e receptors that phosphorylate substrates at a tyrosine amino acid. While receptors in this family react to extremely diverse ligands ranging from growth factors to hormones, they all share the same molecular mechanisms. In the absence of ligand, RTK are monomers, but the binding of ligands brings two monomers close by and allows the tyrosine kinase domain or the physically associated enzyme to phosphorylates the other RTK. Such phosphorylation has two effects: first it usually increase the kinase activity of the receptor, second it generate docking sites for intracellular signaling proteins, therefore resulting in the creation of a large protein complex that can transmit signals to multiple effector proteins (enzymes, cytoskeleton components and transcription factors).

![Image of the JAK-STAT pathway and its activation by cytokines.](image)

Figure 3: The JAK-STAT pathway and its activation by cytokines. The binding of the cytokine to two receptor monomers allows the associated JAK to be close enough to phosphorylate each other on tyrosines and make them fully active. Once activated, they phosphorylate the receptor itself allowing the recruitment and subsequent phosphorylation of STAT proteins. The phosphorylated STAT proteins form a dimer that can enter the nucleus and activates the transcription of specific genes. Adapted from (Alberts, 2015).

Cytokines are a broad category of small globular proteins that are produced by immune
cells and modulate the immune response. A large part of cytokine receptors are members of the RTK family, and more precisely of the JAK-STAT pathway (Figure 3). Those receptors are indeed physically bound to cytoplasmic tyrosine kinases called Janus Kinase (after the two-faces Roman god) which phosphorylate and activate the transcriptional activity of effectors called STAT (Signal Transducer and Activator of Transcription). Upon phosphorylation by JAK enzymes, STAT proteins form a dimer and migrate to the nucleus to activate the transcription of specific genes. In mammal, six different STAT proteins have been identified, each of them being phosphorylated by a specific set of receptor and JAK, and activating the expression of specific genes. For instance Interferon-gamma (IFNg) binds to the interferon gamma receptor which is associated to JAK1 and JAK2: those two JAK enzymes will phosphorylate STAT1 and allow the transcription of proteins required for immune cell activation and bacteria destruction.

While a significant parts of the cytokines act on cells through JAK-STAT signaling pathways, several cytokines relies on a completely different pathway called the NF-kB pathway (NF-kB stands for Nuclear Factor kappa-light-chain enhancer of activated B cells) (Figure 4). It can be activated by several cytokines such as Tumor Necrosis Factor alpha (TNFa) or Interleukin 1 beta (IL1B) through their respective receptors, but can also be triggered by the binding of pathogen components to specific proteins called Toll-like receptors (TLRs). Upon binding of one of those ligands to receptor monomers, the monomers form dimers or trimers and large conformationnal changes allow the phosphorylation of the IKK complex. Once phosphorylated the IKK complex gain the ability to phosphorylate another protein complex made of two NF-kB proteins and of the inhibitory IkB protein. Phosphorylation of this complex induces the degradation of IkB and the liberation of the NF-kB dimer that migrates to the nucleus and activates the transcription of inflammatory genes. Similarly to STAT proteins, five different NF-kB proteins have been identified in mammals and can form a variety of homo and heterodimers, each of them able to activate the transcription of specific genes.

1.1.3 The importance of studying cellular communications at the single cell level

Over the years, multiple signaling pathways have been identified and studied using cellular and molecular biology as well as genetic tools. While those approaches were highly efficient to identify the cellular and molecular components involved in those pathways, they were unable to quantify efficiently those signaling events. More importantly, measures were performed at the whole population and not at the single-cell level, therefore hiding the behavior of individual cells. While this seems anecdotal, a key study published in 1998 revealed that single-cell measurement of the MAPK pathway was necessary to reveal the all-or-none response of oocytes to incoming progesterone, a phenomena hidden when bulk measurements were performed (Ferrell and Machleder, 1998) (Figure 5).

While such results was known back to the late 90’s, the corresponding approach was
Figure 4: The NF-kB pathway and its activation by TNFa. The binding of TNFa causes a conformational change of the aggregated cytosolic tails of the receptors, which then recruit several signaling proteins, resulting in phosphorylation and activation of IKB kinase kinase (IKK). IKK is a heterotrimer composed of two kinase subunits and a regulatory subunit called NEMO. The two kinases phosphorylate IKB, inducing its degradation and the liberation of the NF-kB dimer. Upon liberation, the dimer enters the nucleus and activates the transcription of specific inflammatory genes. Adapted from (Alberts, 2015)

generated only 10 years later due to technical limitation. Indeed the original 1998 study used Xenopus Oocytes, large cells (1.3 mm) that can easily be individually handled, unlike other mammal cells. Critical progresses in the field of fluorescent imaging as well as digital image processing allowed a first surge in the study of cellular communication: starting for the end of the 2000’s, several elegant studies revealed that similarly to the activation of the MAPK pathway in oocytes, the NF-kB pathway is activated in an all-or-one manner (Tay et al., 2010) (Figure 6). Surprisingly, the use of live imaging revealed that NF-kB pathway exhibits an oscillatory behavior due to a negative feedback loop and that such oscillations are required for the expression of key inflammatory and immune genes. A similar phenomenon can be observed in other pathways, such as the p53 pathway, involved in the response to
General introduction

Figure 5: Value of single-cell analysis for cellular signaling studies. (A) Mean activation of the MAPK pathway of a set of Xenopus’ oocytes with increasing doses of progesterone. (B) Two different models compatible with the observation made in the panel (A). In the first one (upper panel) MAPK activation is homogenous across cells and the activation level of each cell progressively increase with progesterone dose. In the second (bottom panel), cells tend to display an all-or-none response and it is the proportion of activated cells that increases with the dose. Adapted from (Alberts, 2015; Ferrell and Machleder, 1998)

DNA damages and UVs, therefore highlighting the importance of studying cell signaling at the single-cell level (Lahav et al., 2004).

The recent surge in single cell genomic and proteomic technologies now allows the measurement of key cellular properties (RNA and protein expression level, cell size, signaling protein phosphorylation...) on thousands of cells and therefore represents a unique opportunity to understand and dissect the mechanisms of cellular signaling events.

1.2 Cellular communications in immunology

1.2.1 The immune system is made up of various cellular components that have to communicate

All living organisms have to defend themselves against infections by harmful invaders, called pathogens, which can be microbes (bacteria, viruses or fungi) or even multicellular organisms. Even bacteria fight against specific viruses called bacteriophages (or phages) using a set of proteins called restriction factors. Restriction factors are produced over bacteriophage infections and are able to block them in a non specific manner through the degradation of their genetic material. Such form of immunity is called ‘innate immunity’ and is characterized by it speed and limited pathogen specificity. In parallel, bacteria have developed a system called the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) system, which
confers resistances to specific foreign genetic elements, such as phages and plasmids. Such system identifies phage sequences during infections, integrate them to the bacterial genome and destroy the corresponding phage sequences using specific endonucleases guided by those integrated sequences. This systems provide a highly specific but slow immune response to pathogenic agent, a form of immunity known as the ’adaptive immune response’.

Vertebrates and more precisely mammals also relies on those two forms of immunity to protect themselves against various pathogens. While this subsection was not written with the objective of providing an exhaustive description of the mammal immune system (which would then be hundreds pages long), it will describe the most relevant aspects of immunology related to my PhD. For interested readers, I would suggest the excellent textbook Janeway’s Immunology (Murphy and Weaver, 2017).

Both the innate and the adaptive immune response relied on a large network of highly dedicated cells, called immune cells. Those cells can classified into 7 major groups of cells :

1. Neutrophils: the most common immune cell in blood and are typical phagocytes, i.e cells able to ingest harmful foreign agents or dying cells to protect the body. They are part of the innate immune system and are among the first cells to migrate to infection sites to destroy the pathogen.

2. Monocytes: those large circulating cells represents between 2 and 10% of the blood immune cells. Upon infection they migrate to the site of infection and differentiate into
effector cells, that is to say macrophages and dendritic cells. In parallel, they produce large amount of inflammatory molecules that activates and fuels the immune response.

3. Macrophages: unlike the two previous cell types, those cells do not circulate in the bloodstream but are instead resident in various organs. These large phagocytes can engulf and destroy pathogens like neutrophils and are therefore part of the innate immune system, but they also participate to multiple biological processes such as dead cell removal, iron and fatty acid metabolisms. Based on their location, they harbor different names and sometimes unique properties and functions with Kupffer’s cells in the liver, Langerhan’s cells in the skin and microglia in the brain.

4. Dendritic cells (DCs): those cells phagocyte and process pathogen materials (also called antigen) in order to present them to the T lymphocytes. As antigen presentation allows to activate antigen/pathogen specific T cells and therefore the adaptive immune system, DCs fill the gap between the innate and adaptive immune system.

5. T cells: unlike the other cell types mentioned earlier, T cells are not myeloid cell but lymphocytes, i.e small cells (7µm of diameter) with a relatively large nucleus and an absence of granules. T cells specifically express a protein complex called the T-Cell Receptor (TCR) that can bind and recognize specific antigens presented by DCs or macrophages. Each T cell expresses a specific TCR that is generated through a special process of controlled DNA recombination and mutation. Therefore each T cell can recognize a unique sets of antigens. Upon binding and activation of the TCR by a cognate antigen, T cells actively proliferate and starts to express several effector molecules. This activation allow them to either directly kill cell infected by intracellular pathogens (those are called cytotoxic T cells, or CD8+ T cells) or to activate other immune cells such as macrophages and B cells (T helper cells, also called CD4+ T cells).

6. B cells: similarly to T cells, each B cells express a specific germline encoded receptor called the B-Cell Receptor. When the BCR recognize a soluble antigen, the cell actively divides and produces a soluble form of the BCR, called antibody. Antibodies bind to viruses and prevent them to enter and infect host cells, but can also bind to bacteria or bigger pathogens to enhance their phagocytosis by macrophages. Together with T cells, B cells are the two arms of the adaptive immune response.

7. Innate lymphoid cells (ILCs): those cells were only identified recently due to their scarcity in blood and similarity with the T lymphocytes. They are defined as lymphocytes which do not express any germline encoded receptor (TCR or BCR). The most common ILCs are Natural Killer (NK) cells, cytotoxic cells that eliminate infected, stressed and cancer cells in a non specific manner (antigen independent).
It is worth noting that non specialized cells can also significantly contribute to the immune response: for instance epithelial cells and fibroblast can induce potent antiviral response, while neurons can directly control the innate immune response in the lung and in the skin. This point will be discussed later in the manuscript.

Each of the aforementioned cell types have specific properties and functions and are not sufficient on their own to protect the body against pathogens. Therefore immune cells have to tightly communicate in order to mount an efficient immune response. An excellent example is the activation of macrophages by T cells. Macrophages can easily internalized bacteria and other pathogens, but are not always able to physically destroy them. In such cases, T cells have to interact with the macrophages and activate them. Activated macrophages can then...
produce potent and highly toxic chemical compounds that destroy the phagocyted pathogen.

1.2.2 Immune cells communicate through soluble factors and direct contact

As described above, immune cells have an intrinsic need to communicate during the immune response to coordinate in a adapted manner. This can be done by soluble factors on short (paracrine) or long distances (endocrine), as well as by direct contact. In this manuscript, endocrine signaling will not be discussed as during my thesis I have focused on immune communication happening at the scale of a tissue.

Paracrine signaling between immune cell is done through small globular proteins specifically expressed by immune cells called cytokines. Each one is only produced and released upon specific conditions, in specific tissues and by specific cell types. Moreover, the effect of each cytokine is unique and will also vary depending on the targeted cell. Lastly, cells are usually not activated by a unique cytokine at a given time but rather by a cocktail of cytokines produced by different cell types, therefore increasing the complexity of the immune cell signaling network.

Let’s take as an example the interferon-gamma cytokine (IFNg), already mentioned earlier. IFNg is specifically produced by lymphocytes during infection: the activation of CD8+ T cells and of specific CD4+ T cells called Th1 cells, through the triggering of their TCR, induces the production and release of IFNg, while ILCs can also release IFNg upon stimulation by the two cytokines IL12 and IL18. Upon release, the cytokine binds to the IFNg receptor and triggers the transcription of many genes including specific immune effector genes. Those genes contribute to an increased antigen presentation, a more efficient destruction of phagocyted bacteria and an expanded recruitment of immune cells at the site of infection. Worth noting, IFNg can also alter the expression of various metabolic enzymes and chromatin regulator, therefore affecting the target cells on a long time scale.

While IFNg has a highly conserved effect across cell types, it can induce cell type specific effects in both immune and non-immune cells. IFNg is indeed able to activate Th1 cells, i.e T helper cells specialized against intra-cellular bacteria while it inhibits Th2 cells, T helper cells dedicated to the elimination of parasitic worms (helminths). It is also able to favor angiogenesis by acting on endothelial cells and even regulates the sensibility of adipocytes (lipid stocking cells) to insulin.

1.3 Role of immune communications in health and diseases

1.3.1 Dysregulations of immune communication trigger or fuel several diseases

While early immunology studies did not rely on modern sequencing, genetic and optical tools they were yet able to show how crucial cellular signaling is essential to mount an efficient and suited immune response, mainly through the prism of clinical immunology. Indeed, the study
of several genetic immunodeficiencies such as hyper IgM syndrome or X-linked hypohidrotic ectodermal dysplasia and immunodeficiency (XL-EDA-ID) have revealed that a mutation in a single gene is sufficient to deeply affect the whole immune system. For instance, mutation of the CD40 gene completely prevents the production of high-quality antibodies and therefore make the patient extremely sensitive to various bacteria and fungi (hyper IgM syndrome) while a single mutation of the NEMO gene, a component of the NK-kB pathway, results in a decreased innate and adaptive immune response (XL-EDA-ID) (Geha and Notarangelo, 2016).

While the study of such diseases revealed that the genetic alteration of a single signaling actor is enough to perturb the whole immune system, those diseases are extremely rare and not really clinically relevant. It is however possible to found common diseases in which subtle changes in the immune cell signaling can dramatically alter its clinical outcome. Let’s consider leprosy, the human disease caused by the mycobacteria *Mycobacterium leprae* with more than 200,000 cases every year over the globe. The classical clinical features in leprosy patients is the association of cutaneous lesions, neuropathologic changes and deformation (Figure 10).
However, clinical symptoms can dramatically vary depending on the type of immune response to the mycobacteria. Clinical spectrum is usually divided into two polar forms: tuberculoïd and lepromatous leprosy. Tuberculoïd leprosy is associated with mild clinical symptoms due to the production of IFNg by T-cells and the efficient elimination of the phagocytosed bacilli by macrophages. In the lepromatous form, IFNg production by T-cells is lacking and other cytokines, called Th2 cytokines such as IL4, IL5 and IL10 are secreted by T-cells. This leads to a prominent production of antibody, a form of immunity inefficient against *Mycobacterium leprae* as it is a mandatory intracellular pathogen, and therefore to extensive tissue damages across the whole body.

It is currently unclear why some patients develop rather a Th1 instead of a Th2 response to the exact same pathogen, although large scale genetic analysis have revealed a possible contribution of the host genetic (Geha and Notarangelo, 2016).
1.3.2 Immunotherapy has revolutionized the treatment of various diseases

Early in the last decade, Stanford researchers theorized the concept of the 'Eroom’s law': for a fixed budget, the number of FDA approved drugs is divided by two every 9 years. According
to them, this was due to an increasingly tighter regulation and the saturation of possible druggable targets (Scannell et al., 2012). Such model had fitted to the data over the last 50 years and was expected to hold in the next decades. However this prophecy turned out to be false, partly due to the emergence of new and highly profitable treatments based on the manipulation of the immune system signaling and communications: immunotherapies (Ringel et al., 2020).

Among them, the most successful are checkpoint inhibitors and are used to treat various cancers. Those drugs are synthetic antibodies that bind to and inhibit specific membrane proteins called immune checkpoints. Such proteins have the ability to inhibit T cell response and are often express by cancer cells to protect them from the immune system (Pardoll, 2012). Therefore, blocking those checkpoints will result in an increased immune response toward the tumor. Those treatments have significantly increased the survival rate and life expectancy of patients suffering from deadly and untreatable cancers such as melanoma and lung cancers (Pardoll, 2012; Brahmer et al., 2015) (Figure 11).

Simultaneously, antibody based treatments that targets inflammatory cytokines such as TNFα (Smolen and Aletaha, 2015; Smolen et al., 2013), IL6, IL17 and IL23 were successfully developed and approved for the treatment of various auto-immune diseases such as psoriasis, rheumatoid arthritis and Crohn’s disease (Figure 12). Those drugs have an opposite effect
when compared to the checkpoint inhibitors: instead of triggering or fueling the immune response of the patient, they aim to block the excessive immune response, the key feature of auto-immune diseases. Such treatments have display extensive efficacy and are now commonly used to treat the mentioned diseases, and similar drugs are now being actively developed and improved.

While such drugs have revolutionized key fields of medicine, the underlying mechanisms explaining their efficacy remain elusive and significant side effects have been observed for many of those drugs, including life-threatening ones for checkpoint inhibitors, due to an uncontrolled activation of the immune system. Moreover, efficacy of those drug can significantly vary across patients, especially for checkpoint inhibitors, strongly limiting their efficacy.

Drugs based on the manipulation of immune cross-talks therefore represent powerful new clinical tools, yet much need to be done to improve the safety as well as the efficacy of those drugs, including a better understanding of the basic events of immune signalling.

1.4 Current challenges in the study of immune communications

Over the last 50 years, studies based on molecular and cellular biology have proven to be extremely useful to identify the components involved in immune signaling and communica-
General introduction

tion. While this knowledge vastly improved the understanding of many pathologies such as cancers and auto-immune diseases, it is not sufficient to fully apprehend basic properties of the immune system or to explain the pathophysiology of key diseases like neurodegenerative diseases. In this section I will describe several current challenges in the field of basic and applied immunology that I tried to answer during the three years of my thesis.

The first and most basic question is the following one: how is the specificity of an immune response toward a given pathogen created and then transmitted to the whole immune system? Indeed, for each pathogen a unique response is produced, with specific cell type recruited at the site of infection and specific cytokines released. Some have speculated that each pathogen triggers the activation of a specific set of innate receptors (Tan et al., 2014), therefore explaining the specificity of the innate response. However such model seems to only explain the specific expression of well known cytokines such as TNFa, IL12 or IL10 and does not explain the involvement of specific cell types. This topic will mostly be addressed in chapter I.

Viruses are unique pathogens: unlike bacteria and fungi, they are unable to replicate on their own and must infect host cells to reproduce. During evolution, all living species have developed a biologic arsenal to protect themselves against viruses, including mammals. In this arm-race, viruses gained the ability to manipulate the host immune response at the cellular response and therefore increase their chance of survival. Such host-virus interactions has mostly been studied in-vitro and are poorly described in in-vivo settings. In the context of the devastating COVID-19 pandemic, we developed a new approach allowing to detect and track virally-infected cells and study how it affects immune cell signaling. This approach is described in chapter II.

Some cytokines have been extensively studied over the years due to their key role in the immune response: we know when and by which cells they are produced, together with their target cells and their effects. However such 'cataloging' approach does not provide quantitative information about those cytokines: at which time and space scale do they act? How far are the target cells? How does the distance to the cytokine producing cells affect cell signaling? Many researchers have hypothesized that understanding the quantitative length and time scales at which cytokines function could provide new insights into the mechanisms by which immune responses is fine-tuned (Altan-Bonnet and Mukherjee, 2019). While such approach is still in its infancy and requires a strong multidisciplinary background, it can reveal essential features of immune signaling. This theme will be mostly discussed in chapter III, done in collaboration with the team of Philippe Bousso.

The last part of my thesis is dedicated to the study of interactions between immune and non immune cells, and more precisely neuro-immune communications: more and more studies are indeed revealing how much entangled the nervous and immune systems are. The nervous system can indeed directly modulate the immune response by activating or inhibiting innate immunity upon infection. On the other hand, the immune system has been identified as a key player in numerous neurodegenerative diseases such as Alzheimer Disease (AD). The clear role
of each brain immune cell in such diseases is unclear, as most the studies focusing on microglia (resident macrophages of the brain). In collaboration with the laboratory of Michal Schwartz, we studied the interactions between immune cells and the oligodendrocytes during AD as well as in other diseases. This last work is described in chapter IV.

While most of those topics are currently being addressed using ’classical’ cellular and molecular biology approaches, the originality of my thesis consist in the constant use of technologies measuring biological features at the single-cell level, also called single-cell technologies, that will be extensively described in the next part of the introduction.
2 An introduction to single-cell genomic technologies

2.1 A historical perspective of the single-cell genomic field

Microscopy is the cornerstone of cellular biology: created and developed during the XVIIth century by scientific pioneers such as Robert Hook, it has revealed the existence of the cells as the most basic unit of life. Critically, it has also unveiled the striking diversity and heterogeneity of cells in a given organism and the imperious need to study this heterogeneity to understand biological processes. A good illustration is the work of Ramón y Cajal, a Spanish neuroscientist who extensively described the shape diversity of cells from the nervous system (Cajal, 1924) and whose work laid the foundation of neurosciences (Figure 13). While the development of new staining procedures allowed the study of various cellular features trough the staining of specific compounds or organelle, microscopy is a barely quantitative approach and can only assess at the same time a very limited set of features.

![Figure 13: Drawing done by Cajal after the observation of a chicken nervous system. A high diversity of size and shape among the cells can be observed. Adapted from (Cajal, 1924).](image)

In opposition to microscopy, molecular biology and genomic approaches can faithfully measure up to thousands of features: gene expression levels can be assessed through RNA-sequencing (RNA-seq), chromatin accessibility by Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), and DNA mutation by DNA-sequencing (DNA-seq). However, this is done at the expense of the cellular resolution: hundreds of cells are usually required
to perform such experiments, therefore preventing their use to study cellular heterogeneity. This situation drastically changed with the emergence of the single-cell genomic field which combine the advantage of both microscopy (single-cell resolution) and genomic (high number of measured features).

In the following section I will adopt a chronological approach to describe the development of the single-cell genomic field and introduce key technical notions that will be described more thoroughly later in the manuscript. It is important to keep in mind that the single-cell genomic field is a constantly and rapidly-evolving field and encompasses a large number of approaches and methods. The proposed historical perspective of the field will therefore not be fully exhaustive and will rather highlight essential technical and conceptual progresses. It will also focus on single-cell RNA-sequencing (scRNA-seq) as it is the most widespread single-cell genomic technology.

2.1.1 The rise of the single-cell genomic field (2009-2013)

At the end of the 2000’s, constant improvement of RNA-sequencing (RNA-seq) allowed to perform sequencing with extremely small amount of input RNA, between 0.1 and 1 µg of total RNA. However this was not sufficient to analyze the transcriptome of a unique cell, which typically ranges between 1 to 50 pg, based on cell type. Moreover RNA-seq of individual cells (also termed as single-cell RNA-seq or scRNA-seq) requires the physical isolation of the cells, a complex and labor intensive task at that time. Altogether this meant that efficient RNA-seq of individual cells was beyond scientist reach at this time and required significant technological breakthroughs to be possible.

The first real single-cell genomic experiment took place in 2009 and consisted in the sequencing of total messenger RNA of mouse blastomeres, large cells with a rich RNA content (Tang et al., 2009). Cells were manually isolated and an improved PCR/RNAseq protocol was used to amplify the initial amount of RNA and bypass the existing threshold of minimal input RNA required for sequencing. However this approach was not scalable and only few large cells were sequenced.

In 2011, a milestone paper (Islam et al., 2011) (Figure 14) introduced a key notion: the multiplexing of the cellular RNA molecules (Figure 15). The multiplexing term refers to the practice of processing and analyzing multiple samples at once. In single-cell genomic experiment, this is done by the addition of molecular barcodes to the RNA or cDNA before pooling material from different cells together. By tagging the RNA/cDNA of individual cells and pooling them before performing further steps of the sequencing protocol, the multiplexing strategy dramatically decreases the amount of reagent needed (i.e the cost per cell) while increasing the throughput of the experiment. Indeed, Islam et al were able to sequenced more than 90 cells in a single experiment, a major breakthrough at the time.

This multiplexing strategy was widely adopted and coupled with new cDNA amplification
protocol that could amplify extremely low amount of input material without the biases of previous methods. This gave birth to more robust and reproducible protocols such as the CEL-seq protocol that uses In-Vitro Transcription (IVT) to amplify cDNA in a linear manner and provides a highly quantitative measurement of the transcriptome of hundreds of individual cells (Hashimshony et al., 2012). Other protocols based on improved PCR amplification instead of IVT, such as the SMART-seq protocol (Ramsköld et al., 2012), were also developed and increased the number of available effective scRNA-seq protocols.

If amplification and multiplexing protocols started to be functional around 2012, the second major technical challenge in scRNA-seq, that is to say the cells physical isolation, remained unsolved. Until 2013 this was done in a manual and labor intensive manner through manual pipetting. An easy-to-implement strategy was to use the already existing fluorescence-activated cell sorting (FACS) method to isolate cells into microwell plates. With such strategy, hundreds of cells can theoretically be sorted in a single plate in a few minutes, dramatically
increasing the throughput of the experiment. However such flow cytometers are expensive and require highly trained staff. Commercial tool were therefore simultaneously developed and in 2013, Fluidigm released the microfluidic C1 system that can capture and isolate up to 96 cells in half an hour (Brennecke et al., 2013) (Figure 15).

In 2013, the combination of FACS or C1 system with the new sequencing protocols allowed the high-quality RNA sequencing of up to a hundred of cells in a single cell experiment. However, it is important to keep in mind that such tools and technology were only available to a reduced set of laboratories and required highly trained and dedicated staff, thus limiting their spread.

2.1.2 The maturation and scaling up of the technologies (2014-2016)

In 2014, the single-cell genomic field entered a new era with the publication of the first scRNA-seq protocol allowing to sequence thousands of cells in a single-cell experiment, the MARS-seq (Massively Parallel Single-Cell RNA-Seq) protocol (Jaitin et al., 2014) (Figure 15). While MARS-seq is based on the already existing CEL-seq protocol, it introduced several major technological advances such as the use of 384 well plates, liquid-handling robotics and of an additional round of plate-based barcoding, raising to 4000 the number of cells sequenced in an unique experiment. In addition to this increase of throughput, MARS-seq also introduced the use of Unique Molecular Identifiers (UMIs) in the single-cell genomic field, in order to remove RNA amplification biases (Kivioja et al., 2012) and improve the quality of the produced data.

![Figure 16: Principle of the droplet-based single-cell sequencing. Adapted from (Macosko et al., 2015).](image)

Only one year later, droplet based sequencing was introduced and further increased the
scale of scRNA-seq experiments while significantly simplifying the isolation of the cells (Figure 15). By using microfluidic devices, thousands of individual cells can easily be encapsulated in tiny water droplets with a nanoliter volume (Figure 16). Each droplet contains a unique barcoded bead that is able to capture the RNA of each cell. Droplets are then broken and RNA/beads complexes are used to build single-cell libraries through a PCR-based amplification protocol. The droplet based approach was simultaneously implemented by the Drop-seq (Macosko et al., 2015) and InDrop (Klein et al., 2015) protocols and progressively became the most popular approach in the field.

The availability of these new protocols allowed to analyze and dissect biological systems that were poorly understood, due to the lack of high-resolution single-cell technologies. The generation of myeloid cells in the bone marrow (Paul et al., 2015), the diversity of brain cortex and hippocampus cells (Zeisel et al., 2015) and the immune dynamic of individual cells (Shalek et al., 2014) were successfully dissected thanks to these technologies, hence validating their ability to analyze complex biological systems.

In parallel to the introduction of these new scRNA-seq protocols, it is worth noting that single-cell versions of other genomic assays started to be available. In 2015, a single-cell version of the ATAC-seq (Buenrostro et al., 2015), Chromatin ImmunoPrecipitation sequencing (ChIP-Seq) (Rotem et al., 2015) and Chromosome Conformation Capture (HiC) (Nagano et al., 2015) protocols were simultaneously released. All of these new protocols consisted in adaption of already established scRNA-seq protocols. Indeed the new scATAC-seq protocol was based on the C1 system while the scCHIP-seq one was relying on the Drop-seq technology, explaining their quick development and availability.

At the end of 2015, the single-cell genomic field was thriving: better and cheaper protocols were available while it became possible to measure other genomic features than RNA level at a single-cell resolution. Moreover, new concepts and cell types were already identified by scRNA-seq, thus validating the interest of the approach.

2.1.3 The era of functional assays, multi-omics and single-cell atlases (2016-2019)

Like any other genomic techniques, scRNA-seq is a purely descriptive method: it faithfully describes the whole RNA contents of individual cells but can not be used to directly infer cellular mechanisms such as transcriptional regulation. In the late 2016, two independent teams came with a similar idea to fill this gap: by combining the CRISPR-Cas9 genome modifying technology with scRNA-seq, they were able to study the effects of specific genes knock out on the transcriptome of individual cells in a high-throughput manner (Jaitin et al., 2016; Dixit et al., 2016; Adamson et al., 2016) (Figure 17). As a proof of concept, these two technologies were successfully used to dissect the Unfolded Protein Response and the monocyte differentiation pathway, both in-vitro and in-vivo. An other noteworthy and functional approach developed at this time is the single-cell joined profiling of lineage and transcriptome: during development
General introduction

Figure 17: CRISP-seq approach. Cas9 expressing mouse cells are infected with a pool of lentiviral vectors containing different simple guide RNA (sgRNA). Cells will therefore express a specific sgRNA, or a combination of several sgRNAs. ScRNA-seq allows to simultaneously measure the transcriptome and identify the sgRNA(s) of each cell: the effect of each sgRNA can therefore be assessed on a single-cell level and possible epistasis studied at a high throughput. Adapted from (Jaitin et al., 2016).

of an organism, an inducible CRISPR-Cas9 system edits a specific transgene, therefore recording cell lineage under the form of a mutated barcode (Raj et al., 2018). Meanwhile, scRNA-seq allows to identify the final state of the cell. While requiring highly modified model organisms, such approach can simultaneously characterize molecular identities and lineage histories of thousands of cells during development and disease.

An other common criticism toward scRNA-seq is that it only measures a proxy of protein gene expression (RNA expression level) and not the real protein expression level. While several studies have shown a strong correlation between mRNA and protein expression level at the population level, the validity of this hypothesis at the single-cell level is questionable. In 2017, by using antibody tagged with specific oligonucleotide sequences, two independent teams were able to solve this challenge by measuring the protein expression level of dozens of proteins in parallel to the conventional scRNA-seq analysis. Interestingly such approach can be easily transposed to most scRNA-seq protocol, including commercial solutions (Stoeckius et al., 2017; Peterson et al., 2017) (Figure 18). Indeed, antibody-bound oligonucleotide sequences act as synthetic transcripts that are captured during most large-scale oligoT-based scRNA-seq library preparation protocols. Such approach is now widely adopted, especially in the immunology field and up to 200 proteins can currently be measured simultaneously.

All technologies mentioned since the beginning of this chapter are extremely powerful
experimental tools, however by 2018 only a limited number of laboratories and genomic platforms used them. While this limited diffusion could be partially explained by the cost of single-cell sequencing, the real cause was that no efficient commercial solution was available. Each laboratory willing to perform scRNA-seq had therefore to implement from scratch a chosen protocol, a time- and money-consuming task. This has changed in 2018 with release of
the Chromium® technology by the 10X Genomics company (Zheng et al., 2017): this droplet-based scRNA-seq technology enables encapsulation of tens of thousands of single cells within minutes with no specific expertise needed, therefore allowing any laboratory to easily perform scRNA-seq studies. This technology is now used in more than half of the scRNA-seq studies and has significantly contributed to the democratization of scRNA-seq (Svensson et al., 2019b) (Figure 19).

This joint increase of scRNA-seq technology throughput and availability finally allowed the launching of large-scale projects with the aim to establish a complete atlas of all cells in an organism. Among them, the most developed project is the Human Cell Atlas with more than two and a half million cells sequenced from 10 different organs, sampled from more 150 healthy donors. While such projects are still in their infancy and will likely require several years to complete, they hold great promises and could have a scientific impact similar to the human genome project.

2.2 The single-cell RNA sequencing technology

Now that we have described the emergence of the single-cell genomic field and provided an overview of the different existing scRNA-seq methods we can focus on the detailed functioning of widely used scRNA-seq platforms. While highly technical, those details are essential to understand several approaches developed during my PhD, especially the Viral-Track tool (Bost et al., 2020a). We will use the example of the MARS-seq 2.0 protocol (Keren-Shaul et al., 2019) to introduce several key notions and techniques used in most of the scRNA-seq approaches and to detail the four distinct steps of any scRNA-seq protocols: the physical isolation of the cell and the barcoding of their cellular RNAs, the pooling of the tagged cDNA, the amplification of the cDNA and the final sequencing step (Figure 20).

![Figure 20: The four main steps of any scRNA-seq protocols.](image)

2.2.1 MARS-seq 2.0: Example of a typical 3’ biased scRNA-seq protocol

In order to explain how common scRNA-seq protocol work, we will use as an example the second version of the massively parallel single-cell RNA sequencing protocol, also known as
the MARS-seq 2.0 protocol (Keren-Shaul et al., 2019) (Figure 21). This protocol is a plate based protocol, which means that single cells are isolated into individual wells of a 384-well plate using flow cytometry. Those wells contain lysis buffer and RNAse inhibitor: the cells are therefore immediately destroyed and their RNA released while the RNAse inhibitors protect the RNA integrity.

A Retro-Transcription (RT) step is then performed to simultaneously convert the RNA into DNA and barcode the content of each well: to do so each well contains a unique RT primer successively composed of a polyT sequence, a well/cell-specific barcode (BC), a Unique

Figure 21: Schematic highlighting the different steps in the MARS-seq 2.0 protocol. Adapted from (Keren-Shaul et al., 2019)
Molecular Identifier (UMI), an Illumina sequencing primer and finally a T7 promoter. By using this RT barcode, all polyadenylated RNA (mostly mRNA) are retro-transcribed and the corresponding cDNAs now have a unique BC and UMI attached to their 3’ end. Moreover the cDNA can now be amplified using In-Vitro Transcription (IVT) as a T7 promoter is also located at the very end of the cDNA molecule.

The content of each well is then pooled in a unique eppendorf tube and the complementary strand of the cDNA molecule is synthesized. The cDNA is then amplified by IVT: the T7 RNA polymerase linearly amplify the cDNA and produces amplified RNA (aRNA) that is devoid of the T7 promoter. The aRNA is then fragmented and a plate specific barcode is added to allow the pooling of multiple plates before sequencing. A second RT step is finally performed to get sequenceable cDNA and add a second Illumina barcode. Lastly, the final cDNA product is further amplified by PCR using the Illumina primers to have enough sequencing input material.

The cDNA is sequenced using pair-end sequencing: 15 nucleotides are sequenced in the 3’ end of the cDNA (call read 1) while 75 are sequenced in the 5’ end (read 2). Read 1 will therefore contains the BC and UMI while the read 2 will simultaneously hold the plate barcode and a fragment of the original RNA sequence: each sequenced cDNA molecule can be assigned to a unique well and plate (and therefore cell) and to a unique genomic region. The use of UMIs further improve the quantification of individual genes in each cell: as each RNA molecule is amplified several times by IVT and PCR, significant amplification biases can be introduced and decrease the overall quality of the quantification. As we know from which RNA molecule each cDNA is generated from, it is possible to collapse all cDNA assigned to the same cell and with the same UMI.

It is crucial to keep in mind the limitations of such scRNA-seq protocol: first due to the use of polyT based RT primers, only polyadenylated RNA molecules are captured and sequenced. Therefore several classes of RNA, such as the long non-coding RNA or some viral RNA molecules can not be tracked by the MARS-seq protocol, but also nearly by all scRNA-seq protocols, as they also all rely on polyT primers. A second important limitation is known as the 3’ bias. During the protocol, the aRNA are fragmented and the obtained fragments have a mean length of 200/300 nucleotides (Keren-Shaul et al., 2019). The genomic part of the read 2 will therefore be located close to the polyA tail, that is to say in 3’ UTR region, and a clear bias of coverage can therefore be observed (Figure 22).

This phenomenon is called the 3’ bias and can observed in most of the scRNA-seq protocols, including the droplet based protocols such as the Chromium and Drop-Seq protocols. This usually does not affect the quantification of gene expression but limits their ability to distinguish different isoforms of the same gene. Other complex analysis, such as the imputation of mutations, can not be performed due to the limited coverage outside of the 3’ UTR region of transcripts, a neglectable part of the genome.

MARS-seq 2.0 is therefore a typical 3’ biased single-cell protocol that can comprehensively
General introduction

2.2.2 Comparison of MARS-seq with other protocols

As mentioned above, MARS-seq is a plate-based protocol and not a droplet based one, like the widely spread Chromium and Drop-seq platforms. This can be seen as a significant drawback of the protocol, as current droplet-based protocols can be encapsulate thousands of cells in a few minutes whereas several minutes are needed to sort a single MARS-seq plate (384 cells). Moreover, unlike droplet based isolation, FACS requires significant technical expertise. This is offset by the ability of plate-based methods to isolate rare cell population: by staining for specific membrane proteins, it is possible to isolate rare cell population and exclude the well characterized abundant cell populations, avoiding the unnecessary sequencing of those cells and limiting the cost of single-cell experiments (Keren-Shaul et al., 2019) (Figure 23). It is also worth noting that the data recorded by the flow cytometer (size of the cell, membrane protein expression) can be linked to the genomic data: similarly to CITE-seq and REAP-seq, expression of a dozen of proteins can be assessed, but in the case of MARS-seq without additional sequencing step. Lastly, MARS-seq can easily be combined with other technologies to answer to specific biological questions:

1. Combined with PhotoActivatable fluorescent reporter, it can specifically sequenced cells from a given spatial niches (NICHE-seq) and was successfully used to dissect the cellular composition of various splenic niches (Medaglia et al., 2017)).

Figure 22: Coverage of the human Metapneumovirus N gene. The data have been generated by scRNA-seq of COVID-19 patients bronchoalveolar lavages using the 10X Genomics Chromium technology. Adapted from (Bost et al., 2020a).
2. As mentioned earlier, it can be coupled with genome editing tools such as CRISPR-Cas9 to assess the effect of specific gene KO in a high throughput manner (CRISP-seq) (Jaitin et al., 2016).

Like most of the scRNA-seq protocols, MARS-seq is a 3' biased protocol. Protocols that do not suffer from this bias are known as full-length protocols and the most used one is the SMART-seq2 (Switching Mechanism At the end of the 5’-end of the RNA Transcript - Sequencing) (Picelli et al., 2014). This protocol offers a significantly better yield than any other 3' biased protocol (higher number of genes detected in similar cells) and can efficiently detect multiple isoforms in the same cell. However this come with a significant cost: such protocols are (up to recently) not compatible with the use of UMIs and are significantly more expensive than 3' biased one (20 times more expensive). Those technologies are therefore
only use in specific context, typically when studying human tumor biopsies (low number of available cells and high heterogeneity) (Tirosh et al., 2016; Neftel et al., 2019).

MARS-seq is therefore a powerful and efficient 3' biased scRNA-seq protocol that can isolate rare and precious cells and can easily be combined with other approaches when needed. These properties will be extensively used in several papers presented in this manuscript.

2.3 Processing and analysis of the scRNA-seq data

2.3.1 From the raw sequencing data to the expression table

Like any sequencing experiments, the result of a single-cell genomic experiment consist in a fastq file, i.e a list of nucleotide sequences and their respective quality score. Such data are not interpretable as is and need to be heavily processed before any real analysis. In the case of scRNA-seq, the expected result of data processing is an expression table containing the number of RNA molecules found in each cell for each gene. I will introduce here the basics of single-cell genomic data processing by using scRNA-seq as an example.

As mentioned earlier, all scRNA-seq experiments rely on pair-end sequencing: both the 5' and 3' end of the resulting cDNA molecules are sequenced. One of the end contains the cellular and UMI barcode, while the other end corresponds to the initial RNA molecule. Here we will consider that the 5' end (R1 read) contains the cell and UMI barcode while the 3' (R2 read) corresponds to the initial RNA molecule. The R1 and R2 reads are therefore processed in parallel to simultaneously identify for each pair of reads the cell from which it come from and the location of the initial captured RNA.

The first step therefore consists in the extraction of the cellular barcode from R1 reads: depending on the technology this barcode usually ranges from 8 to 16 nucleotides and each well or droplet has a unique corresponding barcode. By knowing the expected length and location of the barcode in the R1 read, those are easily extracted and can be then compared to a barcode whitelist. This whitelist either come from the known list of cellular barcodes put in the wells (for plate-based scRNA-seq) or is estimated by looking at the frequency of barcodes and detecting barcodes which occurrence is higher than background signal (for droplet based scRNA-seq). Such density based approach allows to remove barcodes corresponding to empty droplets or sequencing artifacts. It is worth noting that in both cases, the barcode whitelist can be used to correct wrong barcodes: indeed during the multiple amplification steps by IVT or PCR, the low fidelity enzymes used to amplify the genomic material can generate errors in the barcodes. Different approaches have been proposed to correct such errors, the most common one being to consider that every barcode with a Hamming distance to an identified barcode equal to one has been generated from this barcode. Other strategies, mostly graph-based approaches, can also be used to identify and correct more complex errors but at the cost of an increased computational burden. In a similar fashion, UMIs can also be extracted.
from the R1 read and error correction strategy can also be applied with the only difference being that this is done for each individual cells.

Figure 24: Comparison of the different graph-based approaches used for UMI and cellular barcode correction. Each node represents a barcode sequence and the inner number the amount of the given barcode. For a full description of each method, see the original UMI-tools paper (Smith et al., 2017). Adapted from (Smith et al., 2017).

Now that each read pair has been assigned to a unique cell or removed from the analysis if no corresponding cell could be found, the genomic sequences from R2 reads are aligned to the reference genome of the organism of interest. This is done by using splicing-aware aligners, such as STAR or HISAT, that can align millions of reads in a few hours and take into account splicing events. R2 reads that do not align to any genome location, or to too many of them are usually discarded from analysis. Then, by combining the mapping information with the corrected cellular barcode (and if available the UMIs), it is possible to compute the number of reads/UMIs assigned to each gene for each cell. Those information are stored in a matrix, usually called the UMI expression table, that can be then analyzed to generate biological hypothesizes.

2.3.2 Analyzing and interpreting scRNA-seq expression table

The raw UMI expression table is matrix where each column corresponds to a cell (sample) and each row to a gene (feature). Like for any RNA-seq experiments, entries of this matrix are integers that typically follow a negative binomial distribution parametrized by the total number of cellular UMIs and the gene of interest. For a more detailed description of the modeling of scRNA-seq expression, please see the mathematical appendix of (Bost et al., 2020a). Due to the broad number of known genes and profiled cells in a scRNA-seq experiment, UMI tables are large matrices with millions of entries and are therefore hard to analyze due
to their high-dimensionality and the prohibitive computational cost required for each step of the analysis. However, this is balanced by two interesting properties:

- Their high sparsity: between 60 and 80% of the matrix entries are zeros, due to the low proportion of mRNA captured by the scRNA-seq protocols. Lowly expressed genes are therefore often not detected, a situation sometimes called 'dropout' even though this term is controversial. This sparsity allows the use of specific algorithms that significantly reduces the computation time required for basic operation such as column-wise sum or eigenvalues computation.

- Their low intrinsic dimensionality: similarly to bulk RNA-seq, high correlation between the expression of genes with similar biological functions (or expressed by specific cell types) can be observed in scRNA-seq data. This can be used to reduce the dimensionality of the data before performing any further analysis in order to reduce computational cost and improve the interpretability of the analysis.

Now that I have described the basic properties of scRNA-seq data, I will introduce the computational approach used in my different papers to analyze scRNA-seq data. Several pipelines for scRNA-seq data analysis have been developed over time, each of them being based on a different mathematical model and having specific advantages and drawbacks. In the context of my PhD, is decided to use a simple and easy to modify pipeline, Pagoda2 (Lake et al., 2018). It is highly similar to several pipelines and recapitulate several features of the commonly used Seurat pipeline (Stuart et al., 2019). For an exhaustive description of the existing approaches, please see (Luecken and Theis, 2019). Like most scRNA-seq pipelines, Pagoda2 relies on five steps:

1. Filtering of the data: before any real analysis, low quality features and samples have to be removed. Typically dead cells, empty droplets or doublets are filtered out, while genes that are expressed at an extremely low level can be removed. Dead cells are identified by quantifying the proportion of mitochondrial RNA (mt-RNA) among the total sequenced RNA pool as mt-RNA is released in the cytoplasm upon apoptosis. Empty droplets/wells and doublets are identified by looking at the total number of UMIs sequenced as they exhibit an abnormally low or high (respectively) number of total UMIs compared to real cells.

2. Normalization of gene expression: Each count in a UMI expression matrix results from the successful capture, RT and sequencing of a unique mRNA molecule. Therefore gene count for identical cells can vary due to the inherent stochasticity of each of these steps. Thus, when gene expression is compared between cells based on count data, significant differences can arise solely due to sampling effects. Normalization procedures try to address this issue by scaling count data to obtain relative gene expression abundances...
between cells. Many normalization methods have been developed for bulk RNA-seq analysis and have been transposed to the scRNA-seq field: Pagoda2 and many other pipelines relies on count depth scaling, i.e dividing the expression by the total number of sequenced UMIs for a given cell.

3. Selection of Highly Variable Genes (HVGs): many, not to say most of the genes, are not involved in the biological process studied and will be expressed at a similar level by all the cells. Such genes are not informative and can therefore be removed before performing any real analysis. To do so the variance and the mean of each gene is computed, and a local polynomial regression model is fitted to account for the dependency of the variance toward the mean, a characteristic of the negative binomial distribution. Residuals of the regression are computed and are considered as the 'corrected variance'. The genes with the highest corrected variance are usually selected and kept for further analysis. A different approach developed in my thesis consist in computing the proportion of cells in which a gene is not detected as well as mean gene expression. By performing a local regression, it is possible to identify genes with an excess of zeros with regards to their mean expression: those genes have therefore a highly 'concentrated' expression pattern, where a few cells are containing most of the gene UMIs. Genes with the highest zeros excess are then selected for further analysis.

4. Dimensionality reduction: following HVGs selection, the dimensions of normalized expression table can be further reduced by dedicated dimensionality reduction algorithms. These algorithms embed the expression matrix into a low-dimensional space which is designed to capture the underlying structure of the data in as few dimensions as possible, with the reduced dimensions being generated through linear or non-linear combinations of the initial gene expression vectors. Such approach works well on scRNA-seq data as they are inherently low dimensional and can easily capture the major trends of a scRNA-seq dataset. Dimensionality reduction is a field on its own, but two approaches have gained popularity in the single-cell community: Principal Component Analysis (PCA) and Diffusion Map (DM). Pagoda2 uses PCA reduction as it can be easily computed from a sparse matrix (Lake et al., 2018)

5. Clustering: now that the data have been cleaned, the next step consists in organizing cells with similar expression patterns into clusters. This is usually done by using graph-based clustering approaches, also known as community detection methods. Such methods start by computing the K-Nearest Neighbor (KNN) graph. In this graph, each node corresponds to a cell, and each cell is connected to its K most similar cells in the reduced space, based on a given metric (typically euclidean, cosine or correlation distance). Those graphs capture the underlying topology of the expression data. Densely connected regions of graph, i.e communities, are then detected using community detec-
tion algorithms, such as the Louvain or Leiden algorithms (Blondel et al., 2008; Traag et al., 2019). This approach was pioneered by (Levine et al., 2015) and initially developed for flow and mass cytometry data analysis but the high consistency of the resulting clustering with manual annotations and its high speed made it a popular tool to analyze scRNA-seq tools.

6. Differential Expression (DE) analysis: once that several clusters of cells have been identified, one might wonder which are the genes specifically expressed by one of the cell cluster. To do so, several approaches have been developed, most of them being inspired by bulk RNA-seq DE tools. As the current trend consists in sequencing more cells at a lower depth (Svensson et al., 2019a), therefore generating 'binary' data. I therefore used a model based on such binary variables to study DE between conditions. This approach is described in (Blecher-Gonen et al., 2019; Bost et al., 2020a) and relies on a specific binomial regression model called the complementary log-log regression model (Ananth and Kleinbaum, 1997).

2.4 Contribution of the single-cell technologies to Immunology

Single-cell technologies are expensive, time-consuming and often require extensive training to be used. However, such investments are usually worth it and the information provided by single-cell technologies have been crucial to solve many long-standing problems. In particular, immunology has extensively benefited from the development of scRNA-seq over the last years. Indeed, scRNA-seq has successfully been used to study the immune response from in-vitro, in-vivo and clinical samples. In this section, I will present major advances in the field of immunology that rely on single-cell technology, especially on scRNA-seq, but only in an illustrative and not exhaustive manner.

![Diagram](image)

Figure 25: Advances in immunology enabled by scRNA-seq. Adapted from (Avraham et al., 2015; Stubbington et al., 2017).
Due to the limited throughput and sensitivity of the first scRNA-seq technologies, scRNA-seq was first applied to in-vitro immunology experiments. In an early milestone paper (Shalek et al., 2014), mouse dendritic cells were stimulated in-vitro by LPS (a bacterial component triggering a strong innate immune response) before being sequenced at various times post-stimulation. Computational analysis revealed that while all cells display an activated phenotype at later time points, only a small fraction of cells are initially activated by LPS and produce cytokines to activate the neighbor cells in a paracrine fashion. One year later, a second paper extended this work by analyzing the immune response to an intracellular bacteria at the single cell level (Avraham et al., 2015). By using a plate-based scRNA-seq approach coupled with fluorescently labeled bacteria, it was possible to identify and sort the infected cells. Again, significant heterogeneity could be observed among the cells especially for the interferon pathway. By using mutant bacteria, the authors showed that the stochastic expression of a bacterial transcription factor controls the response of the host at the single-cell level. Both studies had a significant impact as they show that even in the simplest biological settings (uniform cell population and a simple in-vitro stimulation), an unexpected and significant heterogeneity can be observed and should be taken into account.

With the introduction of high-throughput scRNA-seq (Macosko et al., 2015; Jaitin et al., 2014), more complex, i.e in-vivo, immunological systems started to be profiled. Innate lymphoid cells (ILCs), were among the first immune cells to be studied in-vivo by scRNA-seq. Indeed those were only discovered recently and their substantial rarity, heterogeneity and plasticity made them extremely hard to study by conventional approaches, making them an ideal target for scRNA-seq studies. Several scRNA-seq studies of ILCs were successively published and allowed a better classification and understanding of those cells (Gury-BenAri et al., 2016; Bjorklund et al., 2016). (Un)fortunately ILCs were not the only problematic immune cell population: Dendritic Cells (DCs) are rare yet highly diverse and their classification based on a combination of morphology, physical properties and molecular markers suffers from a lack of robustness. Similarly to ILCs, deep scRNA-seq studies by multiple teams helped to build a more robust and coherent classification of the DCs, as well as their possible regulation and development mechanisms (Villani et al., 2017; Brown et al., 2019). In both cases, scRNA-seq was essential to establish a robust classification of those rare yet key immune populations.

As mentioned above, progresses in fundamental immunology have lead to the development of immunotherapy. While extremely efficient against a long list of diseases, the precise mode of actions of those drugs are poorly understood, and only a subset of patients usually respond to the treatment. ScRNA-seq has therefore been used to better understand the complex behavior of the immune system in such diseases, that is to say mostly cancer and autoimmune diseases. Breakthrough papers successfully map the immune infiltrate in different tumors such as melanoma (Tirosh et al., 2016) and glioblastoma (Neftel et al., 2019) and identified the mechanisms of melanoma immunotherapy resistance (Jerby-Arnon et al., 2018). Similarly, the immune ecosystem of ulcerative colitis and inflammatory bowel disease, two
similar auto-immune diseases were dissected, and the mechanism underlying the resistance to the common anti-TNF therapy identified (Smillie et al., 2019; Martin et al., 2019). Altogether, scRNA-seq can be used to directly dissect and understand extremely complex diseases and pave the way to improved therapies and treatment.
Chapter I

Contribution of the pathogen positive cells in the initiation and tailoring of the immune response

1 A suited adaptive immune response is generated in the lymph node

Immune responses to pathogens may take a number of forms based on the nature and properties of the pathogenic agent: while intracellular pathogens such viruses or mycobacteria require a cytotoxicity-based response to eliminate infected cells, the response to helminths (parasitic worms) is based on an increased mucus production and tissue repair. Therefore, the immune system must be able to generate a tailored and adapted response to any incoming threat. This challenge is solved thanks to CD4+ T cells, also known as T-helper cells (Th). As mentioned in the introduction, naive Th can be activated by Antigen Presenting Cells (APCs) such as dendritic cells (DCs), macrophages or to a lesser extent by monocytes and neutrophils, that are loaded with pathogen-derived antigens. Based on the cytokines produced by the APCs during the stimulation, the naive Th cells will differentiate into different effector subsets, each one of them expressing a unique master transcription factor and producing different cytokines (Yamane and Paul, 2013).

Specifically, IL12 and IFNg promote differentiation into Th1 cells, Th cells dedicated to fight intracellular infections through an intensive production of IFNg and other effector molecules in a Tbet dependant manner. Th2 differentiation is supported by the IL2 and IL4 cytokines, in the context of anti-helminth responses. Th2 cells express the transcription factor Gata3, allowing an extensive production of IL4, 5 and 13 that promote antibody production, mucus secretion and eosinophils activation. The inflammatory cytokines IL6 and IL23 favor Th17 differentiation, a subset devoted to fight extracellular bacterial infection through the
Contribution of the pathogen positive cells in the initiation and tailoring of the immune response

Figure I.1: T-helper cell subsets function and generation process. Each Th subset is promoted by a specific set of cytokines that trigger the expression of key transcription factor. The resulting Th subsets produce themselves unique sets of cytokines. Adapted from (Yamane and Paul, 2013).

secretion of IL17A, IL17F (hence their name) and IL22 in a Rorgt dependant manner. Lastly, induced T-regulatory cells (iTreg), Foxp3 expressing cells, limit the expansion and activation of the T-cell response mainly through the release of IL10 and TGF-B (Yamane and Paul, 2013) (Figure I.1).

This differentiation process takes place in lymph nodes, highly structured organs that promote communications between the innate and adaptive immune cells. During an immune response, myeloid cells (monocytes, dendritic cells, neutrophils) are activated and migrate from the site of infection (typically the skin) to the closest lymph node through afferent lymphatic vessels. In parallel, lymphoid cells (mostly naive T cells) enter the lymph node from the blood through the High-Endothelial-Venule (HEV) (Figure I.2). In this way, the pathogen derived antigens are brought in contact with the T cells, allowing to select T cells reactive to the pathogen and specifically activate and induce their proliferation.

While this differentiation process has been extensively studied, both in-vitro (Grandclaudon et al., 2019) and in-vivo (Connor et al., 2014; Linton et al., 2003; Hsieh et al., 1993), and is now used as an example of immune cell communication, how the APCs produce the correct polarizing cytokines at the right time remains unclear. In-vitro studies have revealed that APCs can recognize a variety of pathogen associated molecular patterns (PAMPS) thanks to
Contribution of the pathogen positive cells in the initiation and tailoring of the immune response

Figure I.2: Naive T cells encounter antigen presenting cells in the lymph nodes. DCs (and other APCs) enter the lymph node through afferent lymphatic vessels while blood lymphocytes enter through the HEV. Pathogen specific T-cells are quickly activated and start to proliferate while the others leave the lymph node by efferent lymphatic vessels. Adapted from (Murphy and Weaver, 2017).

different innate receptors such as Toll-like receptors (TLR). Each PAMP will bind to a specific receptor and triggers the expression of unique sets of genes, including the polarizing cytokines mentioned earlier (Walsh and Mills, 2013). External factors can also affect the activation profiles of the APCs as cytokines produced at the site of infection by other cells can affect APCs activation state (Walsh and Mills, 2013). To add an additional layer of complexity, multiple cell types can be simultaneously loaded with pathogen derived antigens and the exact contribution of each APC type to T cell activation is hard to deconvolve. Indeed, some models assume that only one cell type is providing all signals required for T-cell activation (antigen loaded on MHC proteins) and polarization (cytokines) while other claim that these two signals can be provided independently by different cell types (Chow et al., 2016).
Contribution of the pathogen positive cells in the initiation and tailoring of the immune response

If lymph node immune cell populations have been well described, their large plasticity and variability in their activation states, as well as the paucity of specific, yet key, cell types make their study a formidable challenge. Standard approaches, such as flow cytometry, can only handle a limited set of cell types, and therefore a different strategy is required. Using non-targeted scRNA-seq might also not be an optimal strategy as the pathogen derived antigen loaded cells (antigen positive cells) only represent a small fraction (<1%) of already scarce immune populations in the lymph node and only a few thousands cells can be sequenced in a single experiment. Therefore, a smart strategy to isolate those cells combined with the resolution of scRNA-seq methods has to be implemented to study how a tailored immune response is generated in the lymph node.

2 Tracking the antigen positive cells to understand their heterogeneity and decipher cellular communication networks

As explained above, tracking the rare antigen positive cells (i.e cells loaded with pathogen-derived antigens) is required to understand the processes underlying the generation of an adapted immune response in the lymph nodes. A common approach is to label the pathogen or antigen of interest with a fluorophore, inject it to the animal or the cells of interest and then purify the antigen positive cells by FACS. Such approach can be easily coupled with plate-based scRNA-seq and has already been used in a milestone paper published in 2015 (Avraham et al., 2015). In this paper, mouse macrophages were generated in-vitro and then infected by the facultative intracellular bacteria Salmonella typhimurium. Using a clever experimental design based on fluorescent bacteria (Figure I.3), the authors were able to distinguish, isolate and sequence uninfected, successfully infected and unsuccessfully infected cells. Infected macrophages displayed a unique transcriptional phenotype compared to bystander macrophages. Moreover, a significant heterogeneity in the inflammatory and interferon response among the infected cells was observed and was elegantly associated with the stochastic expression of a bacterial gene (Figure I.3). Therefore, this in-vitro paper highlights the need for a single-cell study of antigen-positive cells.

In this chapter and its associated paper, we use scRNA-seq coupled with fluorescent labeling to show the heterogeneity of antigen positive cells and infer their cellular communication networks. Indeed, antigen positive cells can be assumed to be the initiators of the immune response as they are early and major producers of cytokines during an infection. Moreover, a significant fraction of the antigen positive cells are also APCs and thus modulate the adaptive immune response in the lymph node. Therefore by performing scRNA-seq of lymph node cells, including antigen positive cells, it should be possible to detect cytokines (or other cell
Contribution of the pathogen positive cells in the initiation and tailoring of the immune response

Figure I.3: Approached used in (Avraham et al., 2015) to study the heterogeneity of cells phagocytizing the intracellular bacteria *Salmonella typhimurium*. Bacteria express the Green Fluorescent Protein (GFP) and are labeled by the red dye pHrodo which binds to the cell wall of bacteria and increases in fluorescence in the low pH environment of macrophage lysosomes (left panel). Due to rapid degradation of GFP in case of bacteria death, relative red and green fluorescence can be used to distinguish uninfected macrophages and macrophages containing live or dead bacteria. The stochastic expression of the bacterial transcription factor PhoP regulating LPS synthesis determines the expression of type I IFN genes by infected macrophages (right panel). Adapted from (Avraham et al., 2015).

...communication related genes) whose expression is induced by the phagocytosis of the antigen. Once the cytokines have been found, one may look for expression of genes known to be induced by the cytokines of interest and the cells expressing them, therefore inferring a cellular cross-talk. As a cytokine can itself induces the expression of other cytokines (e.g IL12 induces the production of IFNg by T cells) this process can be repeated several times. It is worth noting that in theory this strategy can only be used with data sampled at different times during the immune response in order to correctly estimate the time order of the inferred cell cross-talk. Lastly, we also point out that the symmetric strategy can be applied by first identifying an interesting transcriptional signature induced by a cytokine and then identify the likely cellular source.
**Contribution of the pathogen positive cells in the initiation and tailoring of the immune response**

Figure I.4: Strategy used to decipher cellular communication networks in (Blecher-Gonen et al., 2019).

---

### 3 Single-cell analysis of diverse pathogen responses defines a molecular roadmap for generating antigen-specific immunity

This chapter includes the first paper of my PhD and describes the single-cell analysis of the immune response in a mice model of immunization. In this model, various inactivated pathogens are injected to the ear as it is drained to a single lymph node. As the pathogens were fluorescently labeled, we were able to sort and sequence both antigen positive and negative immune cells at various time points. We successfully applied the analytical strategy to those data and identified the cellular cross-talks required for the generation of an efficient Th1 response against mycobacteria.
Single-Cell Analysis of Diverse Pathogen Responses Defines a Molecular Roadmap for Generating Antigen-Specific Immunity

Authors
Ronnie Blecher-Gonen, Pierre Bost, Kerry L. Hilligan, ...,
Benno Schwikowski,
Franca Ronchese, Ido Amit

Correspondence
fronchese@malaghan.org.nz (F.R.),
ido.amit@weizmann.ac.il (I.A.)

In Brief
Single-cell analysis of lymph node response to diverse pathogen challenges reveals that antigen-specific immunity is driven by transcriptional changes that are mostly restricted to antigen-carrying dendritic and monocyte populations, suggesting their key role in the initiation of immune responses.

Highlights
- scRNA-seq of pathogen-challenged lymph node identifies key cell types and pathways
- Antigen-specific immunity is driven by antigen-carrying dendritic cells and monocytes
- Antigen-specific cytokines drive population heterogeneity at the cellular level
- Crossstalk between monocytes and NK cells promotes the induction of Th1 response
Single-Cell Analysis of Diverse Pathogen Responses Defines a Molecular Roadmap for Generating Antigen-Specific Immunity

Ronnie Blecher-Gonen,1,8,9 Pierre Bost,1,2,3,8 Kerry L. Hilligan,4,5,8 Eyal David,1 Tomer Meir Salame,6 Elsa Roussel,4 Lisa M. Connor,4 Johannes U. Mayer,3 Keren Bahar Halpern,7 Beáta Toóth,7 Shalev Itzkovitz,7 Benno Schwikowski,2 Franca Ronchese,4,* and Ido Amit1,10,*

1Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel
2Systems Biology Group, Center for Bioinformatics, Biostatistics, and Integrative Biology (C3Bl) and USR 3756, Institut Pasteur CNRS, Paris 75015, France
3Sorbonne Université, Complexité du vivant, F-75005 Paris, France
4Malaghan Institute of Medical Research, Wellington 6201, New Zealand
5Department of Pathology and Molecular Medicine, University of Otago Wellington, Wellington 6201, New Zealand
6Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot 76100, Israel
7Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel
8These authors contributed equally
9Present address: Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
10Lead Contact
*Correspondence: fronchese@malaghan.org.nz (F.R.), ido.amit@weizmann.ac.il (I.A.)
https://doi.org/10.1016/j.cels.2019.01.001

SUMMARY

The immune system generates pathogen-tailored responses. The precise innate immune cell types and pathways that direct robust adaptive immune responses have not been fully characterized. By using fluorescent pathogens combined with massively parallel single-cell RNA-seq, we comprehensively characterized the initial 48 h of the innate immune response to diverse pathogens. We found that across all pathogens tested, most of the lymph node cell types and states showed little pathogen specificity. In contrast, the rare antigen-positive cells displayed pathogen-specific transcriptional programs as early as 24 h after immunization. In addition, mycobacteria activated a specific NK-driven IFNγ response. Depletion of NK cells and IFNγ showed that IFNγ initiated a monocyte-specific signaling cascade, leading to the production of major chemokines and cytokines that promote Th1 development. Our systems immunology approach sheds light on early events in innate immune responses and may help further development of safe and efficient vaccines.

INTRODUCTION

Immune responses to infectious agents are orchestrated, flexible, and robust processes that involve a large number of cell types and pathways tailored for combating the insult at hand (Yamané and Paul, 2013).

Lymph nodes are structured immune organs that facilitate communication between innate and adaptive immune cells. Resting lymph nodes contain distinct T and B cell areas interspersed with resident innate populations including conventional dendritic cells (cDCs), plasmacytoid DCs (pDCs) (Merad et al., 2013), and macrophages (Baratin et al., 2017; Gray and Cyster, 2012). Cells from peripheral tissues enter lymph nodes through the afferent lymphatics and are guided to the appropriate location by chemokine gradients (Lian and Luster, 2015). In the steady state, incoming cells are predominantly migratory DCs that promote tolerance to self-antigens (Hawiger et al., 2001), whereas monocytes are rare and rapidly transit the lymph node via lymphatics and blood (Jakubzick et al., 2013).

Tissue inflammation can substantially influence lymph node composition. The influx of tissue and blood-derived cell populations creates distinct niches that facilitate cell-cell contact and regulate the conditioning and function of neighboring cells (Oyer-Yaniv et al., 2017). DCs loaded with pathogen-derived antigens increase in number and activation status and provide the critical components required to activate and guide CD4+ T cells toward a specific phenotype: antigen presentation, co-stimulation, and cytokines (Zhu et al., 2010). Specifically, interleukin (IL)-12 promotes Th1 differentiation (Isaev et al., 1999), OX40L co-stimulation supports Th2 differentiation (Linton et al., 2003), and IL-6 and IL-23 favor Th17 differentiation (Zhu et al., 2010). These DC-derived signals can act directly on T cells, as well as on diverse innate immune populations, which in turn also contribute to shaping CD4+ T cell function (Zhu et al., 2010). Further, upon inflammation, monocytes and neutrophils are rapidly recruited to tissues and lymph nodes (Hampton et al., 2015; Tamoutounour et al., 2013) and can contribute to antigen transport and the cytokine environment (Hohl et al., 2009). Depending on the cytokine milieu and antigenic signals, monocytes can differentiate into macrophages or DC-like cells (Goldszmid et al., 2012; Menezes et al., 2016).

Lymph node cell populations are well described. However, the large variability in activation states within each population,
together with the low frequency of specific cell populations, make it difficult to apply traditional technologies, which are limited to a finite number of markers, to monitor early immune responses with the required degree of sensitivity and precision. Recently developed single-cell RNA sequencing (scRNA-seq) technologies have the capacity to unravel cellular heterogeneity with high resolution and accuracy across complex tissues and organs (Giladi and Amit, 2018), identify new transcriptional states, and assess the response of specific cell populations to environmental cues (Cohen et al., 2018).

Here, we used massively parallel scRNA-seq (MARS-seq) (Jaitin et al., 2014; Paul et al., 2015) combined with fluorescently labeled pathogens (Connor et al., 2014), flow cytometry, and single-molecule RNA in situ hybridization (smFISH) (Lyubimova et al., 2013) to generate an extensive innate cellular atlas and communication network following immunization with inactivated bacteria (Mycobacterium smegmatis; Ms), helminths (Nippostrongylus brasiliensis; Nb), or fungi (Candida albicans; Ca). Fluorescent labeling of pathogens together with index sorting and flow cytometry enabled us to track bystander cells as well as cells that directly interact with pathogens. We found that pathogen specificity was coded mostly among antigen-carrying cells, with few antigen-negative populations showing condition-specific changes. One such condition-specific antigen-negative population was natural killer (NK) cells, which were major producers of IFNγ and increased in number only following Ms stimulation. Depletion of NK cells and IFNγ revealed that monocytes were the main responders to these signaling cues. IFNγ signaling strongly activated secretion of the chemokines Cxcl9 and Cxcl10 by monocytes to further Th1 development.

Our work uncovers the innate immune heterogeneity underlying a wide range of pathogens and demonstrates that studying highly complex cell-cell communication networks by scRNA-seq is feasible and yields robust and valuable data (Cohen et al., 2018). Such approaches could be efficiently applied to other complex immunological systems where intercellular communications are essential, such as tumors, developing immune organs, and inflammatory reactions within tissues.

RESULTS

A Comprehensive Map of Lymph Node Innate Responses to Pathogens

To better understand the contribution of different immune subsets toward mounting specific pathogen-directed responses, we collected single-cell profiles from the auricular lymph nodes of mice exposed to inactivated pathogens (Camberis et al., 2013) (Figure 1A). Mice were injected with fluorescently labeled M. smegmatis (Ms), N. brasiliensis (Nb), or C. albicans (Ca), which are well known to induce CD4+ T cell responses characterized by IFNγ, IL-4, or IL-17 production, respectively (Camberis et al., 2013; Igýártó et al., 2011; Junqueira-Knpis et al., 2013).

Pathogen doses were selected by performing dose titration experiments, in which cytokine production by CD4+ T cells was assessed 5 days post immunization (e.g., IFNγ+ Th1 cells for Ms) (Figure S1A). Inactivated pathogens were used to maintain stable conditions across stimulations and time points and prevent the production of pathogen-secreted molecules that may modify the immune responses.

To avoid biases stemming from cell surface markers or selective tissue dissociation, we combined a broad gating strategy that focused on the innate immune compartment (CD45+/CD3−/TCRβ−/CD19−) and permissive tissue dissociation protocols, for massively parallel scRNA-seq (STAR Methods; Figure S1B). MARS-seq analysis of 15,296 cells from 2 biological replicates of control conditions and two time points after injection (day 0, 1, and 2) generated a comprehensive repertoire of defined lymph node innate immune cell populations. Data from 13,827 cells that passed quality control are shown in Figure 1B. Quality control measurements of our analyses are shown in Figures S1C–S1I and Table S5. To quantify cell type numbers in a robust and systematic manner, we performed unsupervised clustering on the entire single-cell data including all time points and pathogens (Cohen et al., 2018; Jaitin et al., 2014), which revealed that the lymph node harbors a complex innate immune cell compartment (Figures 1B and 1C). This included four populations of DCs: migratory DCs, two lymph node-resident cDC populations (Guilliams et al., 2014), and pDCs. The migratory DCs were the most abundant DC type (~20% of total innate immune cells) and were characterized by high expression of Fscn1 and the chemokine receptor Ccr7 (Figure 1C), both related to DC migration and maturation (Ohi et al., 2004; Yamakita et al., 2011), as well as the chemokines Ccl22 and Ccl17. Lymph node-resident cDC1 (5% of the cells) expressed high levels of MHC-II transcripts (H2-Dmb2), the adhesion molecule Cadm1, and the chemokine receptor Xcr1. Lymph node-resident cDC2 (11% of the cells) highly expressed MHC-II transcripts, Cd209a and Cd209d, which are part of the DC-SIGN family of molecules (Geijtenbeek et al., 2000) but low levels of the monocyte marker genes Fcgr2b and Ccr2 (Figure 1B). These genes were strongly expressed in cluster 1 (corresponding to monocytes) but not in cluster 4, supporting the characterization of cluster 4 as LN-resident cDC2 (Figure 1C). Finally, pDCs (7% of the cells) were identifiable by the expression of canonical pDC markers including Siglec, Ly6d, and the chemokine receptor Ccr9 (Swiecki and Colonna, 2015). Monocytes and macrophages were also abundantly detected. Monocytes constituted about 21% of the innate population and strongly expressed Fcgr1, Fcgr2b, and Ccr2. The
macrophage population was rare (2% of cells) and was defined by high expression of complement genes (C1qa, C1qb, and C1qc), ApoE, and MHC-II transcripts, a profile that is consistent with the recently described T cell zone (T2) macrophages (Baratin et al., 2017) (Figures 1B and 1C; Table S1). Two other relatively rare cell types were identified, namely neutrophils (5% of total cells) and plasma cells (2% of total cells), characterized by the expression of S100a8 and S100a9, and immunoglobulin genes (Igk and Igλ2), respectively (Figures 1B and 1C; Table S1). Cells expressing B and T cell markers (CD19, and CD3 and TCRβ, respectively) were excluded from our sorting; however, the dataset contains some contaminating lymphocytes (cluster 5), which highly express the T cell markers Tcf7, Trbc2, and Thy1. More detailed clustering of these cells revealed that they consist of two distinct cell types—type-2 innate lymphocytes (ILC2) and T cells (Figure S2A). Index sorting data of key protein markers further supports the gene-expression-based characterization of lymph node cells (Figures 1B, lower panel, and S2B).

Pathogen-induced cell-cell signaling can result in altered cell state or the active recruitment of specific cell type/s through chemokine gradients. Therefore, we looked for significant changes in the proportion of individual cell populations or clusters over time and across conditions (Figure 1D). The proportion of monocytes significantly increased following immunization with all pathogens when compared to controls, whereas NK cell abundance specifically increased 2 days after Ms injection. This analysis did not detect any other major cell populations that were associated with a specific pathogen. MARS-seq results were further validated by flow cytometry (Figures S2C–S2E). Finally, we projected the single-cell data shown in Figure 1B into a low dimensional embedding using the UMAP algorithm (McInnes and Healy, 2018), which revealed a global structure coherent with our clustering (Figure S2F) and an absence of clear condition specificity (Figures S2A–2C and S3G–S3H; Table S2). Using the single-cell expression profiling and index sorting data, we constructed a robust gating strategy for identifying migratory DCs, monocyte, and neutrophil populations by flow cytometry. Migratory DCs were gated as MHC-II<sup>hi</sup>/CD86<sup>pos</sup> cells, monocytes as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6C<sup>int</sup>, and neutrophils as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6G<sup>−</sup> cells. Ly6G-gating of neutrophils confirmed their phenotype as MHC-II<sup>int</sup>/Ly6<sup>−</sup>G<sup>−</sup> and CD11b<sup>+</sup> (Figure S3I).

The largest group of Ag+ cells was migratory DCs, which consisted of two sub-clusters; migratory DC (1) and (2). These clusters were not associated with a specific treatment, but migratory DC (2) appeared slightly more abundant on day 2 than day 1 (Figure 2A). It is important to note that the two DC subsets described in Figure 2 both correspond to migratory DC2 cells, with flow cytometry profile: MHC-II+/CD11c+/CD103−. Differential (>4-fold) and statistically significant (corrected p value < 0.01) genes were identified in Table S2. Using the single-cell expression profiling and index sorting data, we constructed a robust gating strategy for identifying migratory DCs, monocyte, and neutrophil populations by flow cytometry. Migratory DCs were gated as MHC-II<sup>hi</sup>/CD86<sup>pos</sup> cells, monocytes as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6C<sup>int</sup>, and neutrophils as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6G<sup>−</sup> cells. Ly6G-gating of neutrophils confirmed their phenotype as MHC-II<sup>int</sup>/Ly6<sup>−</sup>G<sup>−</sup> and CD11b<sup>+</sup> (Figure S3I).

Antigen-Positive Cells Are Highly Pathogen Specific

Apart from the aforementioned changes in the abundance of monocytes and NK cells, we could not detect pathogen-specific clusters based on gene expression profiles. We reasoned that pathogen-specific cell populations and gene expression patterns may be limited to the rare cells that had interacted with and phagocytosed antigen. To assess this, we injected fluorescently labeled pathogens and specifically sorted the Af488-antigen-positive (Ag+) cells for MARS-seq analysis 1 or 2 days postinfection. The Ag+ cells represented a small fraction of the lymph node innate compartment, accounting for only 2% to 5% of all CD19−/CD3−/TCRβ− cells, depending on the immunization type and timing (Figure S1B). Overall, 8,256 Ag+ cells were sequenced, and 6,399 cells passed quality control (Figures S3A–S3F). We identified six major populations of Ag+ cells, which could be grouped into four main lineages: migratory DCs, neutrophils, macrophages, and monocytes (Figures 2A, S2, S3G, S4A, and S4B). No resident cDCs were identified among the Ag+ populations. Substantial heterogeneity was observed within each cell population, and many of the clusters showed significant condition specificity (Figures 2A–2C and S3G–S3H; Table S2). Using the single-cell expression profiling and index sorting data, we constructed a robust gating strategy for identifying migratory DCs, monocyte, and neutrophil populations by flow cytometry. Migratory DCs were gated as MHC-II<sup>hi</sup>/CD86<sup>pos</sup> cells, monocytes as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6C<sup>int</sup>, and neutrophils as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6G<sup>−</sup> cells. Ly6G-gating of neutrophils confirmed their phenotype as MHC-II<sup>int</sup>/Ly6<sup>−</sup>G<sup>−</sup> and CD11b<sup>+</sup> (Figure S3I).

The largest group of Ag+ cells was migratory DCs, which consisted of two sub-clusters; migratory DC (1) and (2). These clusters were not associated with a specific treatment, but migratory DC (2) appeared slightly more abundant on day 2 than day 1 (Figure 2A). It is important to note that the two DC subsets described in Figure 2 both correspond to migratory DC2 cells, with flow cytometry profile: MHC-II+/CD11c+/CD103−. Differential (>4-fold) and statistically significant (corrected p value < 0.01) genes were identified in Table S2. Using the single-cell expression profiling and index sorting data, we constructed a robust gating strategy for identifying migratory DCs, monocyte, and neutrophil populations by flow cytometry. Migratory DCs were gated as MHC-II<sup>hi</sup>/CD86<sup>pos</sup> cells, monocytes as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6C<sup>int</sup>, and neutrophils as CD11b<sup>hi</sup>/MHC-II<sup>int</sup>/Ly6G<sup>−</sup> cells. Ly6G-gating of neutrophils confirmed their phenotype as MHC-II<sup>int</sup>/Ly6<sup>−</sup>G<sup>−</sup> and CD11b<sup>+</sup> (Figure S3I).

Monocytes were the second most abundant Ag+ population and included two sub-clusters: Monocytes (1) and (2), which were highly condition specific and differed in the expression of hundreds of genes. Monocytes (1) expressed Plac8, Ly6C2, and Ccr2 and were transiently enriched 1 day after Nb injection, whereas Monocytes (2) expressed Cxcl9 and Cxcl16 and were enriched 2 days after Ms injection (Figures 2A and S2C; Table S2). Index sorting data indicated that MHC-II and CD11c protein levels, as well as cell size (FSC.A), were low in Monocytes (1) and elevated in Monocytes (2) (Figures 2A and S4B), suggesting that Monocytes (2) may be a population of monocytes differentiating into DCs or macrophages. Monocytes (2) also expressed higher...
levels of molecules involved in antigen presentation including Cd40, Cd86, and especially Il12b (Figures 2A and 2B) and expressed high levels of S100a8 and S100a9. While neutrophils are usually not considered to be antigen-presenting cells (APCs), their ability to transport mycobacteria to the lymph node and facilitate T cell responses has been described (Abadie et al., 2005). Finally, Ag+ macrophages, characterized by high levels of Pla2g2d and the cytokine Il18, were found primarily 2 days after Ca immunization (Figures 2A, 2D, and S4A).

Flow cytometry data validated our MARS-seq results and showed that the ratio between the two major Ag+ cell types varied dramatically between the three antigen stimulations, with Nb+ and Ca+ cells comprising mainly migratory DCs, and Ms+ cells comprising equal proportions of migratory DCs and monocytes. Neutrophils were present in Ms but were less frequent after Nb immunization (Figures 2D and 2E).

To further dissect the specific effects of antigens on innate cells, we combined single-cell data from total (CD45+/CD3-/TCRβ-/) and Ag+ (Ag+/CD45+/CD3-/TCRβ-/) cells and analyzed the over-dispersed gene modules across monocytes or migratory DCs (Fan et al., 2016). When applied to the monocyte population, this approach revealed several gene modules corresponding to Th1 or Th2 conditions and monocyte-to-macrophage transitions (Figures S4D–S4F). These programs were strongly regulated by the type of pathogen and the antigen status of the cell. For example, the expression of Il12b and Tnf, two cytokines involved in Th1 polarization, were specific to Ms day 1 and Ms day 2 Ag+ monocytes (Figures 2F and S4D), while the expression of Mrc1, which is associated with M2 polarization (Gundra et al., 2014) was enriched in Nb day 2 Ag+ Monocytes (Figures 2F and S4E). Some modules were specifically enriched in Ag+ cells, regardless of the type of pathogen, such as the macrophage differentiation module (Figure S4F), which included the cathepsin genes Ctsd and Ctsb (Figures 2F and S4G).

The same approach applied to migratory DCs revealed a module specific to Ag+ DC (Figure S4H) that included the chemokines Ccl17 and Ccl22, as well as the Notch ligand Dll4 (Figures S4H and S4I). The expression of the canonical co-stimulatory molecule Cd86 in migratory DCs was also strongly associated with the antigen status of the cell but not with the type of pathogen (Figure S4I), an observation that was consistent with flow cytometry data (Figure S4C). These changes were largely shared by all Ag+ migratory DCs, suggesting a lower degree of condition-associated heterogeneity in migratory DCs compared to Monocytes. Together, our data demonstrate a high degree of condition specificity among Ag+ cells and their gene expression profiles, highlighting a molecular roadmap for generating pathogen-specific responses.

**M. smegmatis Immunization Activates a Specific IFNγ Program in NK Cells**

An in-depth analysis of migratory DCs identified Ms-specific expression of four chemokines, Cxcl1, Cxcl2, Cxcl13, and Ccl3 on day 1, suggesting that Ms+ DCs may trigger pathogen-specific recruitment of other immune cells to the lymph node (Figures S5A and S5B). Indeed, we observed a significant and reproducible enrichment of NK cells (corrected p value < 8.5 x 10^-12 for each replicate) 2 days after immunization with Ms but not with Nb or Ca (Figure 3A; Table S5). This observation was supported by flow cytometry experiments, which showed elevated NK cell numbers after Ms but not Nb immunization (corrected p value = 7.3 x 10^-7). Furthermore, the number of NK cells was found to peak on day 2 after Ms immunization and decrease thereafter (Figure 3B). These results suggest that NK recruitment to the lymph node is a pathogen-dependent and temporally coordinated process.

To gain insight into the activation status of NK cells in different conditions, we compared the transcriptional profile of NK cells after Ms or control treatment. This analysis revealed significant up-regulation of Ifng (corrected p value = 7.6 x 10^-6), a hallmark...
gene of Th1-type immune responses, as well as the chemokines Cxcl9 and Cxcl10 (Figure 3C). Other up-regulated transcripts included Gzmb, which is elevated upon NK cell activation, and Xcfl, a chemokine that is transcribed, but not translated, by resting NK cells (Crozat et al., 2010) and is known to recruit XCR1+DC in the early stages of inflammation (Dorner et al., 2009) (Figure 3C; Table S3). Flow cytometry confirmed the strong and specific up-regulation of the IFNγ-stimulated molecules CXCL9 and Ly6A/E in monocytes from Ms-treated mice, but not Nb or control treated mice, which was detectable as early as 1 day after Ms immunization (Figure 3H). Expression of CXCL9 and Ly6A/E was also detected on TZ macrophages using flow cytometry (Figures 3I and S5G). In contrast, the CD169+ macrophages were Ly6A/E positive but did not express CXCL9 (Figure 3I). Cxcl9 is a highly IFNγ-specific response gene (Groom and Luster, 2011), while Ly6A/E can be up-regulated in response to type-1 IFN signaling (Khan et al., 1993). These results suggest that TZ macrophages are the main macrophage target of IFNγ. Altogether, these data show that lymph node NK cells are specifically recruited and activated during the first 2 days after Ms immunization and may modify the lymph node niche by producing IFNγ that triggers the specific activation of monocytes and TZ macrophages.

**IFNγ-Producing NK Cells Are Required for Optimal Immune Responses to *M. smegmatis***

To validate the inferred cell communication between NK cells and monocytes following Ms stimulation, we depleted NK cells and neutralized IFNγ with monoclonal antibodies prior to Ms immunization. MARS-seq and flow cytometry were performed on both total and Ag+. CD45+/CD19−/CD3−/CD8− cells 2 days post-immunization (Figure 4A; QC measurements are shown in Figure S4B). The antibody-mediated depletion of NK cells efficiently reduced NK cell numbers by approximately 90%, as validated by flow cytometry and MARS-seq analysis (Figures S6A and S6C). MARS-seq data showed the same populations as in previous experiments, which were reproducible both for the cell ratios and cell states (Figures 4B and S6B). We then analyzed the cell types affected by NK/IFNγ signaling by computing the number of differentially expressed (DE) genes between NK/IFNγ-depleted mice and controls for each cell type (Figures 4C and 4D). Consistent with data in Figure 3, monocytes were the populations most significantly impacted by NK/IFNγ signaling. Both Ag+ and Ag− populations failed to up-regulate Ifng response genes, including Cxcl9, Cxcl10, Gbp2, Stat1, and Ly6a, after Ms immunization (Figures 4E and 4F; Table S4). Since the effects of NK/IFNγ signaling were similar between Ag+ and

---

**Figure 4. IFNγ Is Required for Optimal Immune Response to *M. smegmatis***

(A) Schematic of the experimental procedure used to analyze Ms-stimulated lymph node immune cells following depletion of NK cells and blocking of IFNγ.

(B) Clustering of 2,485 innate immune cells from lymph nodes of NK/IFNγ-depleted or isotype-control mice, 2 days after Ms or PBS injection. Heatmap includes 22 marker genes identified by the PAGODA2 pipeline (top panel). The lower panel presents clustering analysis of 1,077 antigen-positive cells from the same samples, featuring 20 marker genes identified by the PAGODA2 pipeline.

(C) and (D) Number of differentially expressed (DE) genes between NK/IFNγ-depleted or isotype control mice for each cluster of antigen positive (C) or total (D) cells. Uregulated (white bars) and downregulated (black bars) genes are shown separately.

(E and F) Volcano plots showing the relative gene expression between NK/IFNγ-depleted and isotype-control-treated antigen-positive (E) or total (F) monocytes. Statistically significant genes (corrected p value < 0.01) are shown in red. See also Table S4.

(G) Flow cytometric analysis of CXCL9, Ly6A/E, and CD11c expression in monocytes and TZ macrophages from NK1.1/IFNγ-depleted or isotype-control-treated mice. Percent values in individual mice are shown on the left, and representative flow plots of three independent experiments are shown on the right. Each dot corresponds to one mouse. A one-way ANOVA with Holm-Sidak’s post-test was used to calculate p values (**p < 0.0001). Error bars represent SEM.

(H) Flow cytometric analysis assessing cytokine (IFNγ and IL-17A), transcription factor (Tbet and Gata3) and CXCPR expression by CD4+ T cells from PBS-treated mice, and NK1.1/IFNγ-depleted or isotype-control-treated mice injected with Ms or Nb. Representative percent values and flow plots in individual mice are shown; each dot corresponds to one mouse. A one-way ANOVA with Holm-Sidak’s post-test was used to calculate p values (**p < 0.0001, *p < 0.05, NS p > 0.05). Error bars represent SEM.

See also Figure S6 and Table S4.
Ag+ monocytes, we conclude that the expression of IFNγ-regulated genes was not primarily controlled by antigen status.

One role of IFNγ is to promote the differentiation of monocytes into more efficient APCs (Goldszmid et al., 2012). Indeed, NK/IFNγ depletion led to a significant decrease in the expression of MHC-II by monocytes, as seen in the index sorting data (Figure S6F). Furthermore, flow cytometry showed that the expression of the IFNγ-regulated molecules CXCL9 and Ly6A/E was significantly lower in monocytes (corrected p values < 10^-10) and TZ macrophages from NK/IFNγ-depleted mice compared to isotype controls (Figure 4G).

Finally, we evaluated the impact of blocking this early cell communication event on CD4+ T cell priming. Th1 polarization after Ms immunization was greatly impaired by NK/IFNγ depletion, with the numbers of IFNγ+ and Tbet+ CD4+ T cells both significantly reduced compared to isotype controls (adjusted p values < 10^-5) (Figures 4H and S6G). The number of T follicular helper (Tfh) cells was not decreased in the absence of NK cells and IFNγ, but their expression of Tbet was significantly reduced (adjusted p value < 10^-4) (Figures 4H and S6G). The number of IFNγ+ cells in the lymph node after Ms immunization was not directly investigated in our study. However, our data indicate that several cell types and signals may participate in this process. Ag+ migratory DCs rapidly expressed Cc10, a chemokine that binds CCR5 on Th1 cells. This observation suggests that early IFNγ is essential for Th1 cell polarization in this model, which is consistent with previous observations utilizing other Th1 models (Martin-Fonseca et al., 2004).

The precise mechanism of NK cell recruitment and activation in the lymph node after Ms immunization was not directly investigated in our study. However, our data indicate that several cell types and signals may participate in this process. Ag+ migratory DCs rapidly expressed Cc10, a chemokine that binds CCR5 on NK cells and promotes their migration in vitro (Bernardini et al., 2008). Cc10 expression was followed by monocyte expression of Cxcl9 and Cxcl10, two IFNγ-induced chemokines that act on the CXCR3 receptor expressed on NK cells and activated CD4+ cells (Wendel et al., 2008). These observations suggest a possible feedforward loop, in which IFNγ originating from NK cells migratory DCs, monocytes, macrophages, and neutrophils. In contrast, monocytes dramatically modified their transcriptional profile in a context-specific manner: Ms+ monocytes specifically upregulated genes related to Th1 polarization, including Il12b, Tnf, and Ccl3/4/5 (O'Garra, 1998), whereas Nb+ monocytes specifically expressed Mrc1, a gene related to uptake of helmint antigens (Sundra et al., 2014). Monocytes are considered to be limited in their ability to prime naive T cells (Tamoutounour et al., 2013). Monocyte expression of high levels of Th1-polarizing cytokines such as Il12b suggests that they may have a role in reinforcing and directing Th1 cell responses originally primed by DCs. This “division of labor” model has been supported by several previous studies in the context of Th1 polarization (Chow et al., 2016; Schreiber et al., 2013) and is clearly documented in the single-cell data from Ms-treated mice presented in this study. We also report that TZ macrophages, a new cell type recently reported to contribute to apoptotic cell disposal in the lymph node T cell zone (Baratin et al., 2017), can also participate in the immune response to Ms by responding to IFNγ and producing chemokines such as CXCL9.

We also observed a significant increase in NK cell abundance specifically after Ms immunization, which overexpressed effector molecules including Ifng and Gzmβ, as well as the chemokines Cxcl9, Cxcl10, and Xcl1, and were located within T cell-rich regions of the lymph node. Depletion of NK cells and IFNγ prevented the later activation of Ms-specific monocyte gene programs and impaired Tbet and IFNγ expression by CD4+ T cells. This observation suggests that early IFNγ is essential for Th1 cell polarization in this model, which is consistent with previous observations utilizing other Th1 models (Martin-Fonseca et al., 2004).

The precise mechanism of NK cell recruitment and activation in the lymph node after Ms immunization was not directly investigated in our study. However, our data indicate that several cell types and signals may participate in this process. Ag+ migratory DCs rapidly expressed Cc10, a chemokine that binds CCR5 on NK cells and promotes their migration in vitro (Bernardini et al., 2008). Cc10 expression was followed by monocyte expression of Cxcl9 and Cxcl10, two IFNγ-induced chemokines that act on the CXCR3 receptor expressed on NK cells and activated CD4+ cells (Wendel et al., 2008). These observations suggest a possible feedforward loop, in which IFNγ originating from NK...
cells induces monocyte expression of Cxcl9 and Cxcl10 to further enhance NK cell recruitment to the lymph node (Martínez-Fontecha et al., 2004). In addition, monocyte expression of Cxcl9 and Cxcl10 may contribute to the appropriate localization and optimal differentiation of CD4+ T cells into Th1, as was reported in a previous study (Groom et al., 2012).

Together, our data suggest that the initiation of the immune response requires cis-activation by direct pathogen-APC contact. Trans-activating signals, such as context-specific inflammatory mediators secreted by activated innate immune cells, guide the immune response toward the required phenotype. This conclusion is in line with previous studies showing that inflammatory mediators alone are unable to promote a full immune response (Sporri and Reis e Sousa, 2005).

In conclusion, we show that scRNA-seq can be effectively used to resolve complex populations that are heterogeneous in lineage, antigen uptake, and activation status, providing a global and unbiased atlas of lymph node cell composition during early immune responses. We also show that in combination with cellular techniques, scRNA-seq can be successfully used to generate and validate hypotheses regarding the cellular communication between immune populations and the mechanisms underlying key immunological pathways. The next challenge will undoubtedly be to understand the complexity of immune responses within the context of the tissue and its 3D structure. New experimental methods including temporal and spatial transcriptomic and proteomic (Goltsev et al., 2018), as well as intercellular enzymatic labeling coupled with scRNA-seq (Giladi and Amit, 2018) may be powerful additional tools to solve complex tissue organization at the single-cell level.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Mice
  - Preparation of Antigens
- **METHOD DETAILS**
  - Immunizations and In Vivo Treatments
  - Preparation of Cells
  - Flow Cytometry
  - Single Cell Sorting
  - Massively Parallel Single-Cell RNA-seq Library preparation (MARS-seq)
  - Single-Molecule RNA Fluorescent In Situ Hybridization (smFISH)
  - Single Cell RNA-Seq Data Clustering and Analysis
  - Robust Detection of Differentially Expressed Genes between Cell Populations in Single Cell RNA-Seq Data
  - Detection of Enriched Cell Types
  - Analysis of Index Sorting Data
  - Analysis of Over-Dispersed Pathways
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
- **DATA AND SOFTWARE AVAILABILITY**

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and six tables and can be found with this article online at https://doi.org/10.1016/j.cels.2019.01.001.

ACKNOWLEDGMENTS

We thank members of the I.A. and F.R. labs as well as Alice Balfour for discussions. We thank Tal Bigdari for artwork. I.A. is supported by the Chan Zuckerberg Initiative, the HHMI International Scholar award, the European Research Council Consolidator Grant (ERC-COG) 724471-HemTree2.0, the Israel Science Foundation (703/15), the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine, the Helen and Martin Kimmel award for innovative investigation, a Minerva Stiftung research grant, the Israel Ministry of Science, Technology, and Space, the David and Fela Shapell Family Foundation, the NeuroMac DFG/Transregional Collaborative Research Center Grant, and the Abramson Family Center for Young Scientists. I.A. is the incumbent of the Alan and Laraine Fischer Career Development Chair. This work was also funded by an Independent Research Organization grant from the Health Research Council of New Zealand to the Malaghan Institute (grant number HRC14/1003). K.L.H. was supported by a Ph.D. scholarship from the University of Otago Wellington, Wellington. P.B. was supported by a Ph.D. scholarship from the Ecole Normale Supérieure, Paris.

AUTHOR CONTRIBUTIONS

R.B.-G. and K.L.H. conceived, designed, performed, and analyzed experiments and wrote the manuscript. P.B. conceived, designed, and analyzed experiments; developed computational methods; performed bioinformatic analysis. T.M.S., E.R., L.M.C., J.U.M., K.B.H., B.T., and S.I. contributed to experiments and wrote the manuscript. E.D. contributed to data processing and analysis; and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 27, 2018
Revised: November 20, 2018
Accepted: January 14, 2019
Published: February 13, 2019

REFERENCES


## STAR METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD16/32</td>
<td>BD Biosciences</td>
<td>Cat#:553141; RRID:AB_394656</td>
</tr>
<tr>
<td>(clone 2.4G2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD3</td>
<td>BD Horizon</td>
<td>Cat#:564379; RRID:AB_2783484</td>
</tr>
<tr>
<td>BV786 (clone 145-2C11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD4</td>
<td>BD Horizon</td>
<td>Cat#:565151; RRID:AB_2687549</td>
</tr>
<tr>
<td>BV605 (clone RM4-5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD11b</td>
<td>BD Horizon</td>
<td>Cat#:564443; RRID:AB_2783485</td>
</tr>
<tr>
<td>BV737 (clone M1/70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD11c</td>
<td>BD Horizon</td>
<td>Cat#:563735; RRID:AB_2783486</td>
</tr>
<tr>
<td>BV786 (clone HL3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD44</td>
<td>BD Horizon</td>
<td>Cat#:564392; RRID:AB_2783487</td>
</tr>
<tr>
<td>BUV737 (clone IM7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD86</td>
<td>BD Biosciences</td>
<td>Cat#:745716; RRID:AB_2743194</td>
</tr>
<tr>
<td>BUV395 (clone GL-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD326</td>
<td>BD Horizon</td>
<td>Cat#:563134; RRID:AB_2783489</td>
</tr>
<tr>
<td>BV711 (clone G8.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse Ly6A/E</td>
<td>BD Horizon</td>
<td>Cat#:562730; RRID:AB_2783490</td>
</tr>
<tr>
<td>PECF594 (clone D7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse MHCII</td>
<td>BD Horizon</td>
<td>Cat#:563413; RRID:AB_2783491</td>
</tr>
<tr>
<td>BV605 (clone M5/114)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse NK1.1</td>
<td>BD Horizon</td>
<td>Cat#:564143; RRID:AB_2783492</td>
</tr>
<tr>
<td>BV650 (clone PK136)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse TCRb</td>
<td>BD Horizon</td>
<td>Cat#:562840; RRID:AB_2783493</td>
</tr>
<tr>
<td>(clone H57-597)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse TCRb</td>
<td>Biolegend</td>
<td>Cat#:100243; RRID:AB_2563946</td>
</tr>
<tr>
<td>biotin (clone H57-597)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD3</td>
<td>Biolegend</td>
<td>Cat#:101239; RRID:AB_11125575</td>
</tr>
<tr>
<td>CD19 biotin (clone 6D5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD19</td>
<td>Biolegend</td>
<td>Cat#:115503; RRID:AB_313638</td>
</tr>
<tr>
<td>biotin (clone 17A2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse MHCII</td>
<td>Biolegend</td>
<td>Cat#:116415; RRID:AB_1953308</td>
</tr>
<tr>
<td>PerCP/Cy5.5 (clone AF6-120.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD45</td>
<td>Biolegend</td>
<td>Cat#:103115; RRID:AB_312980</td>
</tr>
<tr>
<td>APC/Cy7 (clone 30-F11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD11b</td>
<td>Biolegend</td>
<td>Cat#:109203; RRID:AB_314264</td>
</tr>
<tr>
<td>BV560 (clone M1/70)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD11c</td>
<td>Biolegend</td>
<td>Cat#:117307; RRID:AB_317677</td>
</tr>
<tr>
<td>PE (clone N418)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse NK1.1</td>
<td>eBioscience</td>
<td>Cat#:48-5941-82; RRID:AB_2783501</td>
</tr>
<tr>
<td>E450 (clone PK136)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD103</td>
<td>BD Biosciences</td>
<td>Cat#:557495; RRID:AB_386732</td>
</tr>
<tr>
<td>PE (clone M290)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD64</td>
<td>Biolegend</td>
<td>Cat#:139304; RRID:AB_10612740</td>
</tr>
<tr>
<td>PE (clone X54-5/7.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD169</td>
<td>Biolegend</td>
<td>Cat#:142416; RRID:AB_2565620</td>
</tr>
<tr>
<td>AF594 (clone 3D6.112)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD279</td>
<td>Biolegend</td>
<td>Cat#:109110; RRID:AB_572017</td>
</tr>
<tr>
<td>PECy7 (clone RMPI1-30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse Ly6C</td>
<td>Biolegend</td>
<td>Cat#:128018; RRID:AB_1732082</td>
</tr>
<tr>
<td>PECy7 (clone HK1.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse KLRG1</td>
<td>eBioscience</td>
<td>Cat#:17-5893-81; RRID:AB_2783507</td>
</tr>
<tr>
<td>APC (clone 2F1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CD8</td>
<td>in house preparation</td>
<td>n/a</td>
</tr>
<tr>
<td>FITC (clone 2.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti- mouse IL-4</td>
<td>BD Biosciences</td>
<td>Cat#:554435; RRID:AB_395391</td>
</tr>
<tr>
<td>PE (clone 11B11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse IFNγ</td>
<td>BD Biosciences</td>
<td>Cat#:554413; RRID:AB_398551</td>
</tr>
<tr>
<td>APC (clone XMG1.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse IFNγ</td>
<td>BD Biosciences</td>
<td>Cat#:554411; RRID:AB_395375</td>
</tr>
<tr>
<td>FITC (clone XMG1.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse IL-17A</td>
<td>eBioscience</td>
<td>Cat#:515606; RRID:AB_1877135</td>
</tr>
<tr>
<td>PECy7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CXCL9</td>
<td>Biolegend</td>
<td>Cat#:644810; RRID:AB_2200542</td>
</tr>
<tr>
<td>AF647 (clone MIG-2F5.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse Tb7b</td>
<td>Biolegend</td>
<td>Cat#:551960; RRID:AB_394301</td>
</tr>
<tr>
<td>PE (clone 4B10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse CXCR5</td>
<td>BD Biosciences</td>
<td>Cat#:551960; RRID:AB_394301</td>
</tr>
<tr>
<td>Biotin (clone 2G8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-NK1.1 (clone PK136, depleting antibody)</td>
<td>BioXcell</td>
<td>Cat#:BP0036; RRID:AB_2783514</td>
</tr>
<tr>
<td>anti-IFNγ (clone XMG1.2, neutralizing antibody)</td>
<td>BioXcell</td>
<td>Cat#:BP0055; RRID:AB_2783515</td>
</tr>
<tr>
<td>Mouse IgG2a Isotype control</td>
<td>BioXcell</td>
<td>Cat#:BP0085; RRID:AB_2783516</td>
</tr>
<tr>
<td>Rat IgG1 Isotype control</td>
<td>BioXcell</td>
<td>Cat#:BP0088; RRID:AB_2783517</td>
</tr>
</tbody>
</table>

### Critical Commercial Assays

- LIVE/DEAD Fixable Blue
  - Molecular Probes, Invitrogen | Cat#:L32105
- BD Cytofix/Cytoperm kit
  - BD Pharmingen | Cat#:554715
- Transcription Factor staining kit
  - eBioscience | Cat#:00-5523-00
- UltraComp eBeads
  - eBioscience | Cat#:01-2222-41
- CD11b Microbeads
  - Miltenyi Biotech | Cat#:130-049-601

(Continued on next page)
**CONTACT FOR REAGENT AND RESOURCE SHARING**

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ido Amit (ido.amit@weizmann.ac.il).

**EXPERIMENTAL MODEL AND SUBJECT DETAILS**

**Mice**

C57BL/6 mice were purchased from Jackson Laboratories and bred and housed at the Malaghan Institute of Medical Research animal facility or the Weizmann Institute of Science animal facility, under specific pathogen-free conditions. Female mice, 6-8 weeks
of age, were used for all experiments. Experimental protocols were approved by the Weizmann Institute of Science or the Victoria University of Wellington Animal Ethics Committee and were performed according to institutional guidelines.

**Preparation of Antigens**

*M. smegmatis (Ms)* was grown in Luria-Bertani broth overnight at 37°C. Bacteria were washed thrice in PBS containing 0.05% Tween 80 and heat-killed at 75°C for 1 hour. *N. brasiliensis* (Nb) infective L3 larvae were prepared, washed in sterile PBS and killed by three freeze-thaw cycles as described (Camberis et al., 2013). *C. albicans* (Ca) was kindly provided by Prof. Richard Cannon (University of Otago, New Zealand). Yeast were propagated by inoculating sterile yeast extract-peptone-dextrose broth and incubating the culture under agitation at 30°C for 72 hours. Yeast was washed thrice in PBS and heat-killed at 75°C for 1 hour. For fluorescent labelling of antigens, non-viable Ms, Nb or Ca were incubated for 10-15 minutes in 0.05M NaHCO3 buffer and 0.1mg AF488 NHS Ester (Molecular Probes, Invitrogen) and then washed with 0.1M Tris buffer.

**METHOD DETAILS**

**Immunizations and In Vivo Treatments**

Mice were anesthetized and antigens administered by intradermal injection into the ear pinna. PBS was injected into the ear pinna of control animals. Where indicated, mice were treated with depleting or neutralizing antibodies or their isotype controls to assess the requirements for specific cells or cytokines. Anti-NK1.1 (PK136; 200 μg/dose), anti-IFNγ (XMG1.2; 500 μg/dose), Mouse IgG2a Isotype control and Rat IgG1 Isotype control (all from BioXCell) were injected 24 hours prior to immunization and on the day of immunization.

**Preparation of Cells**

To prepare single cell suspensions for MARS-seq and flow cytometry, auricular LNs were digested in IMDM containing 100μg/mL Liberase TL and 100μg/mL DNase I (both from Roche, Germany) for 20 minutes at 37°C. In the last 5 minutes of incubation, EDTA was added at a final concentration of 10mM. Cells were collected, filtered through a 70μm cell strainer, washed with IMDM and maintained strictly at 4°C.

For preparation of T cells, single cell suspensions were prepared from auricular LN by pressing cells through a 70μm cell strainer and washing with IMDM.

For assessment of T cell intracellular cytokines, single cell suspensions were cultured in foetal calf serum-supplemented IMDM (Invitrogen) in the presence of 50ng/mL PMA (Sigma-Aldrich), 1μg/mL ionomycin (Merck Millipore) and 1μL/mL GolgiStop™ (BD Pharmingen) for 5 hours.

For assessment of intracellular cytokines in monocyte, macrophage and NK cell populations, single cell suspensions were cultured in foetal calf serum-supplemented IMDM (Invitrogen) in the presence of 1μL/mL GolgiStop™ (BD Pharmingen) and 2μg/mL brefeldin A (Sigma-Aldrich) for 4 hours.

For assessment on LN macrophage populations, CD11b+ cells were enriched from the single cell suspension by positive magnetic selection. Briefly, cells were incubated with CD11b MicroBeads (Miltenyi Biotech) for 15 minutes at 4°C, washed and run through a MACS column (Miltenyi Biotech). CD11b+ cells were collected for further analysis.

**Flow Cytometry**

For staining of cell surface molecules, cells were suspended in anti-mouse CD16/32 (clone 2.4G2, affinity purified from hybridoma culture supernatant) to block Fc receptors prior to labelling with cocktails of fluorescent antibodies made up in PBS containing 2mM EDTA, 0.01% sodium azide and 2% FCS. Anti-CD3 (145-2C11), anti-CD4 (RM4-5), anti-CD11b (M1/70), anti-CD11c (HL3), anti-CD44 (IM7), anti-CD86 (GL-1), anti-CD326 (G8.8), anti-Ly6A/E (D7), anti-MHCII (M5/114), anti-NK1.1 (PK136) and anti-TCRβ (H57-597) were from BD Horizon. Anti-CD103 (M290) was from BD Biosciences. Anti-CD64 (X54-5/7.1), anti-CD169 (SD6.112), anti-CD279 (RMP1-30) and anti-Ly6C (HK1.4) were from BioLegend. Anti-KLRG1 (2F1) was from eBioscience. Anti-CD8 (2.43) was prepared in-house. Prior to running samples, cells were stained with DAPI (Molecular Probes, Invitrogen) or 7-AAD (Thermo Fisher Scientific) to identify dead cells.

For staining of intracellular cytokines, cells were stained with LIVE/DEAD® Fixable Blue (Molecular Probes, Invitrogen) prior to staining of cell surface molecules. Cells were then fixed and permeabilized using a BD Cytofix/Cytoperm kit (BD Pharmingen) and stained for intracellular cytokines. Anti-IL-4 (11B11) and anti-IFNγ (XMG1.2) were from BD Biosciences, anti-IL-17A from eBioscience (eBio17B7) and anti-CXCL9 (MIG-2F5.5) from BioLegend.

For staining of Tbet, cells were stained with LIVE/DEAD® Fixable Blue prior to staining of cell surface molecules. Cells were then fixed and permeabilized using a Transcription Factor staining kit (eBioscience) and stained for Tbet (4B10; BioLegend).

For staining of Tfh cells, cells were stained with anti-CXCR5 (clone 2G8; BD Biosciences) for 30 minutes at 37°C prior to staining for additional cell surface antigens.

Compensation was set in each experiment using UltraComp eBeads™ (eBioscience) and dead cells and doublets were excluded from analysis. All samples were collected on a LSRFortessa SORP™ or LSRII SORP™ flow cytometer (both from Becton Dickinson, San Jose, CA) and analyzed using FlowJo software (version 10, Treestar Inc).
Single Cell Sorting

Cells were sorted with FACSARIA-FUSION (BD Biosciences, San Jose, CA). Prior to sorting, all samples were filtered through a 70-mm nylon mesh. For the sorting of whole non-lymphocyte immune cell populations, samples were gated on CD45+ (APC/Cy7, Biolegend Inc. San Diego, CA), after exclusion of non-viable cells and doublets, then further gated on CD3e−/TCRγδ−/CD19− (all biotinylated, Biolegend Inc. San Diego, CA). For the isolation of antigen-positive cells, samples were further gated on AF488+ (cells from LN of a PBS- treated mouse were used to set the threshold for AF488 intensity). Other key cell-surface markers were recorded during the sorting, by using the FACS Diva 7 “index sorting” function, which was activated during single cell sorting. Following the sequencing and analysis of the single cells, each surface marker was linked to the genome-wide expression profile. Isolated cells were single cell sorted into 384-well cell capture plates containing 2 µl of lysis solution and barcoded poly(T) reverse-transcription (RT) primers for single-cell RNA-seq (Jaitin et al., 2014). Four empty wells were kept in each 384-well plate as a no-cell control for data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion into the lysis solution, snap frozen on dry ice, and stored at −80°C until processing.

Massively Parallel Single-Cell RNA-seq Library preparation (MARS-seq)

Single-cell RNA-seq libraries were prepared as previously described (Jaitin et al., 2014). In brief, mRNA from single cells sorted into capture plates were barcoded and converted into cDNA and then pooled using an automated pipeline. The pooled sample was linearly amplified by T7 in vitro transcription, and the resulting RNA was fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, RT, and PCR. Each pool of cells was tested for library quality and concentration as described previously (Jaitin et al., 2014).

Single-Molecule RNA Fluorescent In Situ Hybridization (smFISH)

Auricular LNs from Ms or PBS treated mice were harvested 2 days after immunization and fixed in 4% paraformaldehyde for 1 hr. LNs were then incubated overnight with 30% sucrose in 4% paraformaldehyde and embedded in OCT. 6 µm cryo-sections were used for hybridization. Probe libraries were designed and constructed as previously described (Lyubimova et al., 2013). Single molecule FISH probe libraries consisted of 44-83 probes of length 20 bps (see Table S9). Hybridizations with probes for Ifny, Gzma and Cd8e coupled to Cy5, Cy3 and TxRed respectively were performed overnight at 30°C, in a hybridization buffer containing 15% formamide. DAPI dye for nuclear staining was added during the washes. Images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a x100 oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). LN was scanned from one end to the other by imaging consecutive fields, 15 Z stacks of 0.3mm.

Single Cell RNA-Seq Data Clustering and Analysis

All scripts used for the paper are publicly available at https://github.com/PierreBSC/Lymph_node_project/.

In order to avoid removing populations of small-sized cells, a relatively low UMI threshold (350 UMIs) was used in the first instance to filter out low quality cells (Figure S1B). The 350 total UMIs threshold resulted in 742 genes detected per cell on average (95% interval: 148-1949). The distribution of total UMI number across sequencing batches, cell clusters and replicates, is shown in Figures S1C-S1F. Following cell clustering, additional low-quality cells that were rich in ERCC were removed from our analysis. Moreover, cells that belong to clusters representing less than 1% of the total cell population were also excluded from further analysis. Based on the dataset studied, genes with less than a given threshold of UMIs were removed. Single cell RNA-seq data analysis was performed using PAGODA2 pipeline (https://github.com/hms-dbmil/pagoda2) (Lake et al., 2018). Expression values and variance were scaled as described in (Lake et al., 2018). It is important to note that while the first version of PAGODA was relying on gene set analysis to perform cell clustering, PAGODA2 performs most of the analytical steps without any gene set analysis. The general approach used is similar to the one described at https://github.com/hms-dbmil/pagoda2/blob/master/vignettes/pagoda2.walkthrough.oct2018.md/.

First gene variance was corrected and highly variable genes identified using the adjustVariance function (gam.k=10). The 100 first components of Principal Component Analysis based on the 3000 most variable genes were then computed using the calculatePcaReduction function. Then a K-Nearest-Neighbors (KNN) graph was computed based on the PCA reduction through the makeKnnGraph function. Cosine distance was used at this stage to increase the robustness of the analysis. Lastly Multilevel and Infomap community detection clusterings were performed on the KNN graph using the getKnnClusters function and the multilevel.community and infomap.community functions from the igraph R package.

For large datasets (more than 10,000 cells), the K parameter for KNN graph computation was set to 40, Multilevel clustering method chosen and clusters with less than 1% of the cells removed. For small datasets (less than 10,000 cells), the k parameter was set to 20, Infomap clustering method used and clusters with less than 5% of the cells filtered out.

Once the cells were clustered, we identified all genes that were significantly expressed by cells from a given cluster compared to all other cells identified using the PAGODA2 function getDifferentialGenes that performs Kruskal-Wallis test, with default parameters (corrected Z-score > 3, upregulated.only=TRUE). An additional filter was then applied to select genes specifically expressed in a reduced number of cells. Briefly, the genome wide trend between the mean expression of a gene and the proportion of zero’s values is estimated using local regression (LOESS regression). Genes exhibiting a proportion of zeros bigger than expected for their mean
expression level (i.e. residuals of the regression) are considered as potential markers. For a given cluster, the final marker list hence consists of cluster specific differentially expressed genes and the genes with an excess of zeros. The resulting genes are then ranked based on their respective log2 Fold Change compared to other cells and the five top genes (or less due to the filtering) are used as marker genes for the heatmap.

Low dimensional embedding of the data was performed using UMAP algorithm (McInnes and Healy, 2018). The uwot package implementation was used (https://github.com/jlmelville/uwot). PCA dimensionality reductions computed using PAGODA2 was used as input data. N_neighbours parameter was set to 40, spread parameter to 6 and cosine metric was used.

To test for sequencing batch effects in our data, we used low dimensional embedding using UMAP (Figure S1G); which did not reveal any obvious batch effects.

**Robust Detection of Differentially Expressed Genes between Cell Populations in Single Cell RNA-Seq Data**

To robustly detect differentially expressed genes between two cell populations, we assumed that differences in expression were mostly due to changes in the proportion of cells expressing a set of genes, rather than a few cells highly expressing the same set of genes.

Briefly, gene expression was dichotomized (set to one if more than one UMI expressed or to zero otherwise). For each gene a logistic regression was then performed between the dichotomized gene expression and covariates of interest using the glm R core function. These covariates included library size (log10 scale), experimental conditions and sequencing batches if necessary. Statistical significance of the coefficient for each covariate was estimated using Wald’s test and multiple testing correction performed using Benjamini-Hochberg method (p.adjust R function). Genes with dichotomized expression not associated to library size (FDR < 0.01) were removed from further analysis.

To robustly estimate gene expression log fold change (LogFC) between two conditions, we computed the mean expression of each gene in both conditions. A LOESS regression between the mean expression of the two conditions was then performed using the loess R core function and the residuals were considered as the corrected LogFC. Visualization was achieved using a classical volcano-plot.

**Detection of Enriched Cell Types**

Conventional p-values cannot be calculated on two replicates per condition. However, it is possible to estimate how likely a deviation from an expected number of cells is. First, for every major cell population (>5% of the cells), the mean proportion p of this cell type is computed in PBS samples. Then, for each biological replicate, a one-sided binomial test is computed (binom.test R function, ‘greater’ = TRUE) using the total number of cells from this replicate and the mean proportion of this cell type in the control samples as parameters. The obtained p-values were then corrected using Bonferroni multiple testing correction (p.adjust R function). Using this approach, we identified the enrichment of a specific cell population that cannot be explained simply due to sampling noise. In cases when a third replicate was collected, this was included in an initial analysis, and gave similar results. For consistency, the manuscript presents two replicates from each condition. Lastly, we decided to focus on enrichment and not on depletion (hence the use of one-sided test) due to the relative nature of the quantification: an increase of the absolute number of one cell type will result in a general decrease of the other cell proportions.

**Analysis of Index Sorting Data**

Flow data were analyzed using the FlowCore R package (Hahne et al., 2009). Channels intensity were transformed and scaled using the Logicle transform. Transformation was performed using the logicleTransform and estimateLogicle functions with default parameters. GetIndexSort function was used to extract the location of the cells on each 384-well plaque.

**Analysis of Over-Dispersed Pathways**

In order to detect biologically meaningful gene modules/pathways in a robust manner, we performed pathway over-dispersion analysis using de-novo gene-sets. We first selected a given cell population and performed gene expression and variance normalization on the raw UMI counts as described above. The variable genes were selected using the getOdGenes function from PAGODA2 package with default parameters. These genes were then grouped into a given number of pathways (from 20 to 30 based on the number of variable genes) using hierarchical clustering (Ward linkage and correlation-based distance). Over-dispersion of these pathways was computed using the testPathwayOverdispersion function from the same package. The 10 most dispersed pathways were kept for visualization and further analysis.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

P-values for differential expression analysis between different clusters were calculated using the approach described in the previous section. In case of the comparison between two different populations, a gene was considered differentially expressed if it has an absolute logFC larger than 4 and a corrected p-value lower than 0.01. When cells from the same cell population where compared, a gene was considered differentially expressed if it had a corrected p-value lower than 0.01.
For analysis of flow cytometry data, statistical analyses were performed using Prism 7.0 GraphPad software. Data were analysed using Two-Way ANOVA with Sidak’s multiple comparisons test or One-Way ANOVA with Holm-Sidak’s post-test. p values <0.05 were considered statistically significant.

Quantification of the smFISH data was done using the TransQuant tool (Bahar Halpern and Itzkovitz, 2016). The numbers of Ifnγ, Gzma and Cd3e dots were counted over 7-12 0.3 mm Z stacks and normalized to the number of Z stacks * the area of the field. Dots larger than 40um were excluded from the analysis, since they represent background and not real smFISH signal. A cell was considered positive for Ifng or Gzma if more than 4 Ifng or Gzma dots were detected within the DAPI boundaries of the cell.

DATA AND SOFTWARE AVAILABILITY

The accession number of the datasets generated during the current study is: GSE: 125044 and is available via GEO at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125044.

All scripts used for the paper are publicly available at https://github.com/PierreBSC/Lymph_node_project/.
Supplemental Information

Single-Cell Analysis of Diverse Pathogen Responses Defines a Molecular Roadmap for Generating Antigen-Specific Immunity

Figure S1, related to Figure 1 | Sorting strategy and data quality and reproducibility. A. Percentage of IFNγ+ CD4 T cells, 5 days after immunization with increasing doses of inactivated Ms, Nb and Ca, respectively (top). Each point represents one mouse. Number of IL4+, IL-17A+ and IFNγ+ CD4+ T cells, 4 days after immunization with 4M CFUs of Ms, 300 larvae of Nb and 10M CFUs of Ca respectively (bottom). Bar graphs show mean +/- S.E.M for groups of ~30 mice. Statistical significance was determined using a One-Way or a Two-Way ANOVA with Sidak’s multiple comparisons test. NS = not significant, * p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001. Hash signs represent statistical significance relative to the PBS control. B. Gating strategy for single-cell index sorting of antigen-positive cells (PBS negative control in upper row, Nb day 2 in lower row). Data are representative of five experiments (PBS) or three experiments (Nb day2). C. Distribution of the total number of UMIs per cell among all sequenced cells. The dashed black line corresponds to the threshold of 350 UMIs. D-F. Distribution of the total UMI number across sequencing batches (D), across the clusters in Figure 1B (E), and across immunizations, time-points and replicates (F). The dashed black line in (D) corresponds to the threshold of 350 UMIs. G. Bar plot indicating the number of cells which pass QC cutoff across different conditions, time-points and replicates. H. UMAP embedding of total cells presented in Figure 1B. Each color corresponds to one sequencing batch. I. Linear regression showing the reproducibility of MARS-seq data between biological replicates. Examples of three major cell types from Ms day 2 samples are shown.
Figure S2, related to Figure 1 | Flow cytometry and alternate clustering methods validate findings in Figure 1. A. Detailed clustering of cluster 5 (Lymphocytes) from Figure 1B using tSNE clustering. B. Fluorescence intensity of selected index sorting markers across clusters presented in Figure 1B. C. Gating strategy for flow cytometry validation of the different populations identified in Figure 1B (left), and total cell number in the LN following the different immunizations. D–E. Flow cytometry analysis of the number of cells (D) and the percentage among non B/T cells of the different cell populations (E) identified in Figure 1B, across immunizations and time-points. Line graphs show mean ± S.E.M for groups of 5 mice. F–G. UMAP embedding of the cells shown in the Figure 1B heatmap. Cells are colored based on clustering in Figure 1B (F) or based on the different conditions and time-points (G).
Figure S3, related to Figure 2 | Antigen-positive cells express differential gene modules in a pathogen-specific manner. A. Distribution of the total number of UMIs per cell among antigen positive cells. The dashed black line corresponds to the threshold of 350 UMIs. B. Bar plot indicating the number of antigen positive cells which pass QC cutoff across different conditions, time-points and replicates. C.-E. Distribution of the total UMI number across sequencing batches (C), clusters in Figure 2A (D) and immunizations, time-points and replicates (E). F.-H. UMAP embedding of the antigen positive cells shown in the heatmap in Figure 2A. Cells are colored based on the sequencing batch (F) or on clustering in Figure 2A (G), or on the different conditions and time-points (H). I. Gating strategy used to identify cell subsets among antigen-positive cells by flow cytometry (related to the quantification appearing in Figure 2E). Data are representative of three independent experiments.
Figure S4, related to Figure 2 | Pathogen-specific gene modules are activated in monocytes and migratory DCs. A. Log2 gene expression of key molecules across antigen-positive cell types, as clustered in Figure 2A. B. Flow cytometry index data of the CD11b and forward scatter (FSC.A) parameters across clusters in Figure 2A and S4A. C. Flow cytometric analysis of CD86 expression on antigen-positive or antigen-negative cells, 1 day (left panel) or 2 days (right panel) after immunization with Ms, Nb or Ca, or after PBS treatment. Each point represents one mouse. A One-Way ANOVA with Holm-Sidak’s post-test was used to calculate p-values (* * * * * P<0.0001, *** P<0.001). Error bars indicate S.E.M. D-F. PCA score of a Th1 polarization pathway (D) or of an antigen positive pathway (E) or of a macrophage associated pathway (F) gene modules detected in monocytes 1 or 2 days following stimulation with Ms, Nb or Ca (right panel), and the weighting of genes contributing to the PCA score (left panel). G. Log2 expression of Cd40 and Csb genes in monocytes following different stimulations and at different time-points. Antigen-positive (plain border) and total cells (dashed border) cells are also compared. H. PCA score of a chemokine pathway gene module detected in migratory DCs 1 or 2 days following stimulation with Ms, Nb or Ca (right panel), and the weighting of genes contributing to the PCA score (left panel). I. Log2 expression of key molecules from the chemokine pathway presented in G as well as two other co-stimulatory molecules (Cd40 and Cd86), as expressed in migratory DCs following different stimulations and at different time-points. Antigen-positive (plain border) and total cells (dashed border) are also compared.
Figure S5, related to Figure 3 | M. smegmatis immunization activates a DC chemokine module and NK IFNγ production. A. PCA score of a chemokine pathway gene module detected in DCs 1 day following stimulation with Ms, but not with Nb or Ca (right panel), and the weighting of genes contributing to the PCA score (left panel). B. Log2 expression of key molecules from the pathway presented in A, as expressed in monocytes 1 day after Ms stimulation. Antigen-positive (filled bars) and total cells (hatched bars) are also compared. C. Gating strategy used to define IFNγ+ positive cells among NK cells as shown in Figure 3D. D. Flow cytometry analysis of the frequency of KLRG1+ cells among NK cells 3 days after immunization with Ms or Nb (left panels), or over 5 days following Ms immunization (right panels). Each point represents one mouse. One-Way ANOVA with Holm-Sidak’s post-test was used to calculate p values. Hash symbols denote significance compared to the PBS control. Asterisks denote significance between the indicated groups. (**P<0.0001, *** P<0.001, ** P<0.01, * P<0.05, NS P > 0.05). Error bars indicate S.E.M. E. Flow cytometry analysis of the number of IFNγ positive cells among various cell types in the LN after immunization with Ms over 5 days. Stacked bar graphs show mean +/- S.E.M. for groups of 10 mice at each time-point. F-G. Gating strategy used to define the various IFNγ positive cell populations as shown in Figure S5E (F) or to define the various monocyte and macrophage populations in Figure S3I (G). Data are representative of at least four samples.
Figure S6, related to Figure 4 | Depletion of NK cells and IFNγ does not significantly alter abundance of LN innate immune cells. A. Flow cytometric analysis of antibody-mediated NK cell depletion efficiency in peripheral blood. Percent values in individual mice are shown on the left; representative flow plots are shown on the right. Each dot corresponds to one mouse. Error bars indicate S.E.M. B-C. Abundance of antigen-positive (B) or total (C) cell subtypes in the lymph node of NK1.1/IFNγ depleted or isotype-control treated mice, 2 days after Ms immunization, as detected in MARS-seq data. D. Library size distribution of antigen positive (left) and total cells (right) in antibody-mediated NK depletion experiments. The dashed black line corresponds to the threshold of 350 UMIs. E. Reproducibility of MARS-seq data in antigen-positive monocytes between biological replicates of NK1.1/IFNγ depleted (right panels) or isotype control (left panels) treated mice. F. Comparison of MHC-II expression levels from MARS-seq index sorting data in total monocytes from NK1.1/IFNγ depleted or isotype-control treated mice, 2 days after Ms immunization. Rectangle correspond to points between the first and the third quartile. Whiskers corresponds to 1.5*IQR. G-H. Gating strategy used to define T cells (G) or T cell subtypes (H) in Figure 4H.
4 Contributions and limitations of our study

By combining fluorescent labeling of pathogens with scRNA-seq of the lymph node’s innate immune cells, we were able to study the early mice immune response to three different pathogens and to propose a complete model of the early immune response to mycobacterial infections (Figure I.5). In this model, DCs, monocytes and also to a minor extent neutrophils capture the antigen in the skin and migrate to the lymph node. Upon their early (24h) arrival in the lymph node, antigen positive DCs produce large amount of chemokines that are likely to recruit NK cells. Those NK cells will then release large amount of IFNγ and activate the monocyte/macrophage compartment by inducing CXCL9 and CXCL10, two Th1 promoting chemokines. IFNγ early production by NK cells is crucial for a correct Th1 polarization as shown by our depletion experiment which completely abrogate Th1 polarization. Interestingly, it seems that stimulation and polarization of the naive Th cells is performed by monocytes derived cells expressing high levels of IL12, TNF and costimulatory molecules and not by DCs.

![Figure I.5: Model of the early immune response to mycobacteria in the lymph node.](image)

While our joined experimental and analytical approach was effective at inferring immune cell cross-talks, it suffered from severe limitations (Figure I.6). The first one consist in the impossibility to directly study cell-cell interactions, and more precisely contact dependent interactions. Those cytokine-independant cross talks can not be efficiently studied with our approach as it is relying on the physical interaction between two cell types and not the ex-
pression and local release of cytokine. The second limitation is that our approach of antigen positive cells tracking can not be applied to human clinical samples as real pathogens will lack any reporter ability, such as fluorescence. An other critical flaw of our study is that it did not integrate the spatial distribution of the cells of interest even though this information is crucial to understand how cells communicate with each others. In addition to these conceptual limitations, we identified a more technical one: scRNA-seq low capture efficiency makes the detection of lowly expressed genes, such as transcription factors or cytokines challenging, therefore limiting the quality of our analysis. Lastly, this work only focused on cellular interactions between immune cells: multiple studies have reported key cellular cross talks between immune and non immune cells, especially interactions with neurons. Such neuro-immune cross-talks are driving the pathogenesis of both infectious (bacterial infections (Baral et al., 2018)) and non infectious diseases (Alzheimer’s disease (Heppner et al., 2015)) and have to be taken into account.

During my thesis, the first limitation was addressed by other researchers in two elegant papers (Pasqual et al., 2018; Giladi et al., 2020): in the first paper the authors developed a method called Labeling Immune Partnerships by SorTagging Intercellular Contacts (i.e LIPSTIC) (Pasqual et al., 2018), where a protein ligand/receptor pair of interest are genetically fused to either the SrtA enzyme or to a tag consisting of five N-terminal glycine residues (G5). Upon physical ligand-receptor interaction, the SrtA enzyme catalyze the transfer of a given substrate onto the G5-tagged receptor, typically a fluorescent ligand. The fluorescent cells can then be isolated and studied. LIPSTIC was successfully applied to the CD40/CD40L based DC/Th cells interaction, both in-vitro and in-vivo. While this method requires highly genetically engineered mice (modified for both the ligand and its cognate receptor), it is a powerful tool to study specific cellular interactions that happened at a given time. In opposition, the Physically Interacting Cells Sequencing (PIC-seq) aims to study cellular interactions that are taking place at a given time: through the use of carefully calibrated tissue disassociation pro-
tocols and staining of membrane proteins representing distinct cell type markers, the authors were able to sequence not single cells, but doublets which correspond to physically interacting cells. When combined with the sequencing of individual cells, it is possible to computationally deconvolve the expression profile of the doublets and identify doublet specific gene expression patterns. Such genes can be assumed to be induced by the physical interaction and therefore the result of cellular cross-talk. Due to the high quality of both works, such approach was not explored during my thesis and will not be discussed further in the manuscript.

Concerning the tracking of antigen positive cells in human samples, it was extensively studied and investigated in the case of viral infection in my thesis. A new computational approach to track virally infected cells, Viral-Track, was developed and successfully applied to study COVID-19 and its associated virus, SARS-CoV-2, resulting in two different papers. This work is fully described and discussed in chapter II of my thesis.

While scRNA-seq is a powerful tool, location of the cell is lost during the tissue dissociation step and can not be recovered. One may combine single-cell analysis with conventional imaging of the tissue through IF or smFISH staining in similar fashion to what was done in our study to identify the location of specific cell types. However, conventional microscopy is not adapted and in chapter III, we combined intravital imaging with bulk and scRNA-seq analysis to study how IFNγ affects the global immune environment of mouse and human tumors.

Finally, the last chapter of my thesis (chapter IV) focuses on neuro-immune interactions, and more precisely on the interactions between immune cells and oligodendrocytes in the context of various pathogenic contexts such as Alzheimer’s disease, multiple sclerosis or aging.
Chapter II

The tracking of virally infected cells

1 Why and how to study viral infections at the single-cell level?

Figure II.1: Examples of pathogen emergence or re-emergence. Red circles correspond to newly emerging diseases while blue ones correspond to re-emerging ones. It is worth noting that this figure was published in 2004, since then, several new pathogens have emerged, including SARS-CoV-2. Adapted from (Morens et al., 2004).

Despite continuous progress in health and medicine, infectious diseases still cause about one quarter of all human deaths worldwide, more than all cancer cases and only second to...
cardiovascular diseases. Recent pandemics (worldwide epidemic), such as AIDS in 1981, Ebola virus disease in 2014 and the latest, COVID-19, have shown that they remain a constant threat (Morens et al., 2004) (Figure II.1). Moreover, as climate and ecosystem changes dramatically fuel this trend, the situation is likely to worsen in the coming decades (Morens et al., 2004). While this threat has been overlooked so far, the dramatic human and economic damages done by the COVID-19 pandemic have been game-changing and it is likely that more efforts will be put on the study of these emerging pathogens, especially on emerging viruses.

In order to respond quickly and efficiently to these new diseases, it is crucial to gain knowledge of the causative virus and biological mechanisms involved in the disease. This knowledge span different fields and consists in knowing which cells are infected by the virus and how cells respond to the virus, how the virus affects infected cell behavior, how it replicates or how it manipulates the host immune response. To answer to these questions, several approaches have been developed and can be roughly grouped into three groups (Figure II.2):

- **In-vitro approaches**: host cells are cultivated *in-vitro* and then infected by the virus of interest. The response of the cells to the viral infection can easily be assessed, but also perturbed using various strategies including drugs or genetic perturbation of both the host and the virus. While easy to implement and providing valuable information about the early phases of the disease, such approaches are completely unable to model the systemic response of an organism, especially the adaptive immune response.

Figure II.2: Main approaches used to study viruses-induced diseases.

- **Animal models**: can be used to study the systemic response of the host. Possibility to perturb both the host and the virus. Animal models do not exist for all viral diseases. Physiological relevance might be limited. Low amount of available material. High biological variability. Impossible to perform biological perturbation.

- **Clinical research**: provides direct information on the pathogenesis of the disease.

**Pro**
- Easiest way to study the virus life cycle and dynamic.
- Can be used to study innate immunity against a given virus.

**Cons**
- Poor physiological relevance.
- Only recapitulate the first phase of the disease.

- In-*vitro* approaches: host cells are cultivated *in-vitro* and then infected by the virus of interest. The response of the cells to the viral infection can easily be assessed, but also perturbed using various strategies including drugs or genetic perturbation of both the host and the virus. While easy to implement and providing valuable information about the early phases of the disease, such approaches are completely unable to model the systemic response of an organism, especially the adaptive immune response. Worth
noting, recent pioneer works have shown that it is possible to use organoids, 3D cellular structures that can be established from induced pluripotent stem cells, to cultivate and study viruses in a controlled environment similar to real organs and tissues (Clevers, 2020).

- **Animal models:** to understand how a virus affects a whole organism, the simplest approach is to infect a model organism with this virus. By considering that the model organism’s response to the virus is similar to the human one, it is possible to infer the exact pathogenesis of the disease. Similarly to the *in-vitro* approach, both the host and the virus can be perturbed to validate specific hypotheses. However, it is important to keep in mind that good animal models do not always exist, and when they exist are not always easy to deal with: for instance after more than 30 years of research, no simple small animal model has been found for HIV, and only humanized mice and non-human primates such as macaques can be used, raising important ethical concerns (Hatziioannou and Evans, 2012).

- **Clinical research:** various biological samples can be directly obtained from living (biopsy) or deceased patients (autopsy) to understand the exact pathophysiology of the disease. While this approach bypasses the aforementioned limitations of both *in-vitro* and animal models studies, this comes at a great cost. For the sake of the patient, only limited amount of material can be collected, and sometimes the tissue of interest can only be reached through long, painful and risky biopsies (i.e, lung biopsies for viral pulmonary infections). Therefore, only a limited set of assays can be performed for each patient’s sample. In addition, confounding factors such as treatments, age and sex, can dramatically alter the biology of the sample and lead to wrong hypothesis if not carefully taken into account.

In this chapter, we hypothesize that scRNA-seq can be used to get the most out of precious human samples obtained from patients suffering from viral diseases. Indeed, unlike other methods, scRNA-seq requires extremely low amount of input material and can assess the state of each cellular population in a single experiment. Moreover, as most viruses express polyadenylated mRNA, scRNA-seq protocols can in theory simultaneously capture the host and the viral mRNA, an approach termed ’dual RNA-seq’ (Westermann et al., 2012) (Figure II.3). With a suited computational pipeline, this would allow to detect the cells infected by the virus and differentiate them from their neighboring but un-infected cells (sometimes termed bystander cells). It is worth noting, that such strategies can in theory capture the possible alterations of the host transcriptome induced by the viral infection.

The identification of the infected cell is indeed crucial: in the previous chapter we have seen that antigen positive cells are the initiator and drivers of the immune response. However, identifying and isolating those cells remains a challenge in human samples. While we
could track the pathogen fragments through fluorescent labeling in our mouse immunization model, this is not possible in human samples. Therefore, this approach could be of great interest to study the contribution of the infected cells in the immune response to viruses. This dual scRNA-seq approach has already been applied to mouse model of Influenza A infection (Steuerman et al., 2018), or to in-vitro infection of Herpes-Simplex-1 virus (HSV-1) (Drayman et al., 2019) but was never applied to human clinical samples. Moreover, no computational tool had been developed to easily analyze such data. We therefore developed a new computational tool called Viral-Track, that systemically scans for the presence of viral sequences in scRNA-seq sequencing data. Such approach allows not only to detect the expected virus, but also secondary and unexpected viral infections. As viral infections are suspected to trigger or fuel various non infectious diseases such as autoimmune diseases (Münz et al., 2009), our approach could be applied not only to ’classical’ viral diseases, but also to any disease which has been previously associated with a virus.

Figure II.3: Principle of dual RNA-seq. Adapted from (Westermann et al., 2012).
2 A new player in the game: the SARS-CoV-2 virus

Over the last two decades, emerging viruses were more or less successfully contained to their place of origin, sometimes at the cost of hundreds of medical staff lives (Sylvester Squire et al., 2017). In December 2019, a new respiratory virus called SARS-CoV-2 was detected in Wuhan (China) and quickly spread despite an aggressive lock-down policy. By march 2020, this epidemic became a pandemic as the virus was detected in nearly all countries. In mid-July 2020, more than half million people have died due to this virus.

This new virus is a member of the coronavirus family, a family of enveloped positive single-stranded RNA viruses that primarily infects the respiratory tract of various mammals, ranging from bats to humans and civets. It is highly similar to two recent emerging viruses, the SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) (with 80% and 50% homology, respectively) but is less deadly while having a longer incubation time, explaining the difficulty to limit its spread (Figure II.4).

Coronaviruses carry the largest genomes (between 26 to 32 thousands of bases) among all RNA virus families. Interestingly, each viral transcript owns a polyA tail and a 5’ cap, making them highly similar to eukaryotic and specifically human mRNA molecules and therefore easily detected and sequenced by conventional RNA-seq protocols (Kim et al., 2020) (Figure II.5). Upon cell entry, the genomic RNA is translated to produce nonstructural proteins from two open reading frames (ORFs), ORF1a and ORF1b. Those two ORFs encode two polypeptides that are cleaved into several nonstructural proteins, including a RNA-dependent RNA polymerase that is used for replication and transcription of the viral genome. Negative-sense RNA intermediates are generated to serve as the templates for the synthesis of positive-sense genomic RNA (gRNA) and subgenomic RNAs (sgRNAs). The gRNA are incorporated into virions while the sgRNAs are used to produced structural proteins, such as the spike
The tracking of virally infected cells

Figure II.4: Comparison of the three emerging coronaviruses and of Influenza A.

(S), envelope (E), membrane (M) and nucleocapsid (N) proteins. Accessory proteins are also translated from those sgRNA, but their role is less well-defined (Kim et al., 2020).

Figure II.5: Genome organization of SARS-CoV-2. Adapted from (Kim et al., 2020)

Similarly to SARS-CoV, SARS-CoV-2 entry into the cell relies on the binding to the Angiotensin-converting enzyme 2 (ACE2) membrane protein and the proteolytic activity of the serine protease TMPRSS2 (Hoffmann et al., 2020). These two genes are simultaneously expressed in the lung by type 2 pneumocytes, but also by ileal absorptive enterocytes, and nasal goblet secretory cells (Ziegler et al., 2020), allowing the virus to infect several tissues. However, unlike SARS-CoV and other respiratory viruses, SARS-CoV-2 does not trigger a
The tracking of virally infected cells robust interferon response but instead a strong inflammatory response with the induction of inflammatory cytokines (IL6, IL1B) and chemokines (CXCL9, CXCL16) (Blanco-Melo et al., 2020). This high degree of inflammation seems to be a driver of pathology, and several clinical trials have tried to limit this inflammation through the use of monoclonal antibodies such as Tocilizumab (anti-IL6) (Guaraldi et al., 2020) or the corticosteroid dexamethasone (Horby et al., 2020). How the virus affects the immune system remains unclear and most of the studies published focused on the blood and not on the lung immune response. Therefore, much need to be done to fully elucidate the pathogenesis of COVID-19 and identify new therapeutic targets.

3 Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients

The first paper of this chapter contains an extensive description of the tool developed to detect and analyze viral infection at the single-cell level, Viral-Track, as well as its benchmark on various mouse and human viral infections, such as Influenza A virus, Vesicular Stomatitis Virus (VSV), Lymphocytic Choriomeningitis Virus (LCMV) and Hepatitis B Virus (HBV). By collaborating with the Third’s Shenzhen people’s hospital in China, we were able to collect and sequence cells from COVID-19 patient bronchoalveolar lavages in the early phase of the COVID-19 pandemic, and were the first to provide a single-cell atlas of the immune response to SARS-CoV-2 virus in the human lung. This paper was published in the journal Cell in June 2020.
Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients

Graphical Abstract

Highlights

- Viral-Track: a computational framework to analyze host-viral infection maps
- Viral-Track sorts infected from bystander cells and reveals virus-induced expression
- SARS-CoV-2 infects epithelial cells and alters immune landscape in severe patients
- Co-infection of SARS-Cov-2 and hMPV affects monocytes and dampens interferon response

Authors

Pierre Bost, Amir Giladi, Yang Liu, ..., Benno Schwikowski, Zheng Zhang, Ido Amit

Correspondence

benno@pasteur.fr (B.S.), ido.amit@weizmann.ac.il (I.A.), zhangzheng1975@aliyun.com (Z.Z.)

In Brief

A computational framework that allows for the identification and characterization of virus-infected cells as well as bystander cell responses reveals how SARS-CoV-2 alters the immune responses of patients.
Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients

Pierre Bost,1,2,3,6 Amir Giladi,1,6 Yang Liu,4,6 Yanis Bendjelal,3 Gang Xu,4 Eyal David,1 Ronnie Blecher-Gonen,1 Merav Cohen,1 Chiara Medaglia,1 Hanjie Li,1 Aleksandra Deczkowska,1 Shuye Zhang,5 Benno Schwikowski,2,* Zheng Zhang,4,* and Ido Amit1,7,*

1Department of Immunology, Weizmann Institute of Science, Rehovot, Israel
2Systems Biology Group, Department of Computational Biology and USR 3756, Institut Pasteur and CNRS, Paris 75015, France
3Sorbonne Universite, Complexite du vivant, Paris 75005, France
4Institute for Hepatology, National Clinical Research Center for Infectious Disease, Shenzhen Third People’s Hospital, School of Medicine, Southern University of Science and Technology, Shenzhen 518112, Guangdong Province, China
5Shanghai Public Health Clinical Center and Institute of Biomedical Sciences, Fudan University, Shanghai 201508, China
6These authors contributed equally
7Lead Contact

*Correspondence: benno@pasteur.fr (B.S.), zhangzheng1975@aliyun.com (Z.Z.), ido.amit@weizmann.ac.il (I.A.)
https://doi.org/10.1016/j.cell.2020.05.006

SUMMARY

Viruses are a constant threat to global health as highlighted by the current COVID-19 pandemic. Currently, lack of data underlyng how the human host interacts with viruses, including the SARS-CoV-2 virus, limits effective therapeutic intervention. We introduce Viral-Track, a computational method that globally scans unmapped single-cell RNA sequencing (scRNA-seq) data for the presence of viral RNA, enabling transcriptional cell sorting of infected versus bystander cells. We demonstrate the sensitivity and specificity of Viral-Track to systematically detect viruses from multiple models of infection, including hepatitis B virus, in an unsupervised manner. Applying Viral-Track to bronchoalvelo-lavage samples from severe and mild COVID-19 patients reveals a dramatic impact of the virus on the immune system of severe patients compared to mild cases. Viral-Track detects an unexpected co-infection of the human metapneumovirus, present mainly in monocytes perturbed in type-I interferon (IFN)-signaling. Viral-Track provides a robust technology for dissecting the mechanisms of viral-infection and pathology.

INTRODUCTION

The development of efficient vaccines against viral pathogens is considered one of the biggest achievements of modern medicine and has significantly contributed to the increase in life expectancy worldwide. However, no vaccines exist for many life-threatening viruses such as HIV (Burton, 2019), Zika virus (Pierson and Diamond, 2018), or hepatitis C virus (HCV) (Bailey et al., 2019). Additionally, efficient broad-spectrum antiviral drugs are still missing, making infectious diseases a significant challenge for modern health systems. Viruses can also trigger or fuel non-infectious diseases such as cancer (Young and Rickinson, 2004) and are suspected to contribute to various other chronic diseases such as Alzheimer disease (Itzhaki, 2018) and various autoimmune disorders (Münz et al., 2009). The recent emergence of highly pathogenic viruses such as the Ebola virus and the emerging SARS-CoV-2 pandemic recalls the constant threat that viruses represent to global health. So far, the SARS-CoV-2 pandemic has caused a global financial and social catastrophe and is expected to make a significant long-lasting impact on human health (Zhu et al., 2020). Despite intensive research efforts, little is known thus far regarding the interaction of the SARS-CoV-2 virus with the human host and, as a consequence, no efficient treatment has been designed so far (Chen et al., 2020). Moreover, only few therapeutic targets have been identified, highlighting the urgency to develop additional strategies to dissect the virus-host interactions.

Single-cell RNA sequencing (scRNA-seq) is an emerging technology that has been extensively used to study several complex diseases, including cancer (Li et al., 2019), neurodegeneration (Keren-Shaul et al., 2017), and auto-immune (Zhang et al., 2019) and metabolic diseases (Jaitin et al., 2019), providing new insights and revealing new therapeutic targets and strategies (Yofe et al., 2020). In the context of infectious diseases, scRNA-seq studies identified the underlying cells and pathways interacting with various pathogens (Drayman et al., 2019; Shnayder et al., 2018; Steuerman et al., 2018; Zanini et al., 2018). During the immune response to a pathogen, a limited number of antigen-positive or infected cells initiate and modulate the host immune response (Blecher-Gonen et al., 2019), while most of the tissue response is propagated through cytokines, such as type I interferon (IFN) signaling, to bystander, uninfected cells. It is therefore essential to develop new analytical tools to identify the rare infected cells.

Please cite this article in press as: Bost et al., Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients, Cell (2020), https://doi.org/10.1016/j.cell.2020.05.006
in order to better understand complex host-virus interactions underlying these pathologies. Multiple experimental tools have been developed over the years to track virus-infected cells in vivo, characterize the cellular state of the infected cells, and differentiate them from their bystander neighbors. These include fluorescently labeled pathogens or pathogens expressing fluorescent proteins (De Baets et al., 2015; Blecher-Gonen et al., 2019), as well as reporter mice (Lienenklaus et al., 2009). However, in the case of human clinical samples, these tools are limited, making the pathogen-infected cells and viral reservoir cell types hard to detect.

Viruses exploit their host cells to first express viral genes, optimize the cellular environment, and then fully activate the viral replication program. Because scRNA-seq technologies rely on polyadenylated RNA isolation and amplification, current scRNA-seq methods can, in theory, detect viral RNA programs and therefore enable accurate identification of the bona fide infected cells and their unique properties at single-cell resolution. While such an approach has already been used to study both in vitro (Drayman et al., 2019; Shnayder et al., 2018) and in vivo infection models (Steuerman et al., 2018), no general computational framework has been developed to detect viruses and analyze host-viral maps in clinical samples. Here, we present a new computational tool, called Viral-Track, that is designed to systematically scan for viral RNA in scRNA-seq data of physiological viral infections using a direct mapping strategy. Viral-Track performs comprehensive mapping of scRNA-seq data onto a large database of known viral genomes, providing precise annotation of the cell types associated with viral infections. Integrating these data with the host transcriptome enables transcriptional sorting and differential profiling of the viral-infected cells compared to bystander cells. Using a new statistical approach for differential gene expression between infected and bystander cells, we are able to recover virus-induced programs and reveal key host factors required for viral replication. Viral-Track is able to annotate the viral program with high accuracy and sensitivity, as we demonstrate in several in vivo mouse models of infection, as well as human samples of hepatitis B virus (HBV) infection. Applying Viral-Track on bronchoalveolar lavage (BAL) samples from moderate and severe COVID-19 patients, we reveal the infection landscape of SARS-CoV-2 and its interaction with the host tissue. Our analysis shows a dramatic impact of the SARS-CoV-2 virus on the immune system of severe patients, compared to mild cases, including replacement of the tissue-resident alveolar macrophages with recruited inflammatory monocytes, neutrophils, and macrophages and an altered CD8+ T cell cytotoxic response. We find that SARS-CoV-2 mainly infects the epithelial and macrophage subsets. In addition, Viral-Track detects an unexpected co-infection of the human metapneumovirus in one of the severe patients. This study establishes Viral-Track as a broadly applicable tool for dissecting mechanisms of viral infections, including identification of the cellular and molecular signatures involved in virus-induced pathologies.

RESULTS

Viral-Track: An Unsupervised Pipeline for Characterization of Viral Infections in scRNA-Seq Data

All scRNA-seq computational packages implement a pipeline that initially aligns the sequenced reads to the expressed part of a reference host genome of the relevant profiled organism. Irrelevant reads, representing other organisms, primers, adaptors, template switching oligonucleotides, and other contaminants are then commonly discarded. We reasoned that during infection, and likely many other pathological processes, these reads can potentially carry valuable information about viral RNA that is discarded in this filtering step. In order to efficiently detect viral reads from raw scRNA-seq data in an unsupervised manner, we developed Viral-Track, an R-based computational pipeline (Figure 1A; STAR Methods). Briefly, Viral-Track relies on the STAR aligner (Dobin et al., 2013) to map the reads of scRNA-seq data to both the host reference genome and an extensive list of high-quality viral genomes (Stano et al., 2016). Because viral reads are highly repetitive and generate substantial sequencing artifacts, the viral genomes identified in Viral-Track with a sufficient number of mapped reads are then filtered, based on read mapping quality, nucleotide composition, sequence complexity, and genome coverage, to limit the occurrence of false-positives (STAR Methods). Due to the lack of high-quality viral genome annotations, Viral-Track includes de novo transcriptome assembly of the identified viruses using StringTie (Pertea et al., 2015). Finally, viral reads are demultiplexed, quantified using unique molecular identifiers (UMI), and assigned to unique viral transcripts and cells (Figures 1A and S1A). The Viral-Track algorithm has been designed to robustly handle various types of scRNA-seq datasets, as illustrated below, and is publicly accessible at https://github.com/PierreBSC/Viral-Track.

In order to evaluate the specificity and sensitivity of Viral-Track, we benchmarked Viral-Track on several scRNA-seq datasets (Table S1). These datasets include a large number of experiments we conducted, as well as published studies, that span several tissues (lung, spleen, liver, and lymph node) and a wide range of viruses: influenza A, lymphocytic choriomeningitis virus (LCMV), vesicular stomatitis virus (VSV), herpes simplex virus 1 (HSV-1), human immunodeficiency virus (HIV), and HBV. We first evaluated mouse lungs infected in vivo by influenza A virus and sequenced using MARS-seq2.0 (Keren-Shaul et al., 2019; Steuerman et al., 2018). Viral-Track analysis specifically detected the 8 distinct influenza A viral segments (NC_002016 to NC_002023 Refseq nucleotide sequences) from the specific infecting strain (H1N1 Puerto Rico 8 strain) (Figure 1B). We performed transcriptome assembly to test the feasibility of reconstructing the viral transcriptome from 3’-enriched scRNA-seq data. The results were highly coherent with the current knowledge of influenza A transcriptome, exemplified by Viral-Track’s ability to identify documented spliced transcript structures with single-nucleotide precision. For instance, we identified the exact location of the key splicing site on segment 7 that gives rise to M2 transcript and links nucleotides 51 and 740 (Dubois et al., 2014) (Figure 1C). Quantification of the number of viral reads across different experimental conditions was consistent with current knowledge of the disease, with lung stomal cells of non-immune lineages (CD45−) exhibiting a significantly higher viral load compared to immune cells (CD45+) (p = 0.039, two-tailed Welch’s t test) (Figure 1D).

As inbred mice lack the influenza-specific restriction factor Mx1, influenza A infection is extremely virulent in inbred mice (Haller et al., 1980). Moreover, all influenza A mRNA are capped
Please cite this article in press as: Bost et al., Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients, Cell (2020), https://doi.org/10.1016/j.cell.2020.05.006
and polyadenylated, making them an optimal substrate for scRNA-seq isolation and amplification protocols. We therefore evaluated the sensitivity and specificity of Viral-Track in a more challenging dataset. In this model, photoactivatable-GFP (PA-GFP) mice were infected with LCMV (Armstrong acute strain), a virus lacking strong poly(A) mRNA signals (Burrell et al., 2017), via injection to the footpad. 72 h post-infection, CD45+ splenic immune cells from different spatial niches (T zone, B zone, marginal zone, and total spleen) were profiled using the NICHE-seq technology (Medaglia et al., 2017). Even though the LCMV viral mRNAs are not polyadenylated, we detected mRNA molecules that converted to cDNA through priming of the MARS-seq oligo(dt) RT primer, and Viral-Track successfully identified the two viral segments (LCMV segment L [NC_004291] and S [NC_004294]) (Figure S1B), albeit the number of detected reads was an order of magnitude lower than the number observed in influenza A infection (Figure 1E). We detected viral reads in samples from the marginal zone, B zone, and the total spleen, but not in T zone samples, and marginal zone samples exhibited significantly higher viral load compared to B zone and total spleen samples (Figure 1E; p = 0.0067 and 0.0083 respectively, two-tailed Welch’s t test). This observation is in line with the biology of LCMV, which primarily infects macrophages and lymphocytes from the marginal zone of the spleen (Müller et al., 2002).

We next evaluated whether Viral-Track is sensitive to barcode swapping during Illumina-based scRNA-seq (Griffiths et al., 2018), which, in the case of viral RNA detection, can lead to the false assignment of viral reads to uninfected cells. To this end, we infected mice with one of two different viruses, LCMV and VSV, and performed MARS-seq2.0 on CD45+CD19−CD3− non-B/T cells from the auricular draining lymph node 1 day after infection (STAR Methods). All samples were sequenced concurrently to test for cross-sample viral read contamination. For both viruses, Viral-Track was able to identify the correct viral segments (Figures S1C and S1D), with no cross-contamination, evident by the absence of VSV reads detected in the LCMV-infected cells and vice versa (Figure S1E). We further generalized Viral-Track for commonly used scRNA-seq technologies and non-RNA viruses. We applied Viral-Track to scRNA-seq data from a recently publication of human primary cells infected ex vivo with HSV-1, a linear double-stranded DNA virus, generated using the Drop-seq platform (Drayman et al., 2019; Maciosko et al., 2019). We found that Viral-Track detected and identified correctly HSV-1 RNA specifically in the infected samples but not in the controls (NC_001806 Refseq nucleotide sequences) (Figures S1F and S1G). Finally, we analyzed scRNA-seq data of CD4+ T cells infected ex vivo with HIV-1 (Bradley et al., 2018), generated using the droplet-based chromium platform (Zheng et al., 2017). Viral-Track successfully identified HIV as the unique virus present in the infected samples (Figures S1H and S1I), but detected significant amounts of HV-1 viral reads in one control samples probably due to ambient contamination (Yang et al., 2020).

**Defining the Host Viral Interactions of HBV Using Viral-Track**

We further tested Viral-Track’s applicability for detecting viral reads in human clinical samples. For this purpose, we generated scRNA-seq data from a liver biopsy of an untreated hepatitis B patient and analyzed the data using Viral-Track. Viral-Track successfully identified HBV as the only virus present in the sample (Figure 1F) with 18,420 reads assigned to the HBV genome (NC_000377 Refseq sequence). Coverage analysis revealed a strong peak located at the 5’ end of the C gene, encoding for the main core protein, suggesting that the HBV virus is actively producing virions (Figure 1G). We then overlaid the viral data on the host transcriptome to identify infected and bystander populations. A total of 13,803 cells passed a lenient quality control, permitting apoptotic signals that may arise from viral infection. We identified several non-immune cell types (Figure S1J), including hepatocytes (expressing ALB and APOA2), as well as hepatocytes showing apoptotic signatures (ALB with high expression of mitochondrdial genes), sinusoidal endothelial cells (FCN2), and epithelial cells (KRT7). We also observed several subsets of immune cells such as B cells (MS4A1), plasma cells (M2B1), conventional dendritic cells 1 (cDC1; XCR1), and monocytes (cDC2; CCR2).

Figure 1. Viral-Track Retrieves Viral Reads in a Variety of Tissues, Viral Strains, and Sequencing Platforms

(A) Schematics of the Viral-Track approach. Single-cell sequencing data of cells from an infected tissue, containing infected and bystander cells are analyzed by Viral-Track. Viral-Track maps the sequenced reads to both the host reference genome and a database of viral genomes, overlaying infection status on top of the host transcriptional landscape.

(B) Results of Viral-Track analysis on scRNA-seq data from influenza A PR8-infected mouse lungs. For each viral segment, represented by a dot, the complexity of the sequences (measured by entropy, i.e., how repetitive are the mapped sequences) and the percentage of the segment that is mapped are plotted. Dark red dots correspond to viral segments of the influenza A PR8 strain and yellow dots to segments belonging to other H1N1 influenza strains. Viral segments with more than 50 mapped reads are shown.

(C) Coverage plot of the influenza A segment NC_002016 (influenza A PR8 segment 7). M2 transcript location estimated using StringTie is shown below with the splicing site position.

(D) Quantification of the number of reads assigned to influenza viral segments across experimental settings. Each dot corresponds to a technical replicate (384-well plate). Two-tailed Welch’s t test was used to compare viral load between CD45+ and CD45− cells (p = 0.039).

(E) Quantification of the number of reads assigned to LCMV viral segments in the different zones of the spleen. Each dot corresponds to a technical replicate (384-well plate). Two-tailed Welch’s t test was used to compare viral load between cells from the infected marginal zone to cells from the B zone or the whole spleen (p = 0.0067 and 0.0083 respectively).

(F) Result of Viral-Track analysis on scRNA-seq data from a HBV patient. For each viral segment, represented by a dot, the entropy of the sequence and the percentage of reads that is mapped is plotted. Green dots correspond to viral segments that passed quality control. Viral segments with more than 50 mapped reads are shown.

(G) Coverage plot of the HBV genome. Locations of the different viral genes from NCBI database are depicted at the bottom.

(H) Enrichment of infected cells across hepatic cell subsets (left panel); red line corresponds to an enrichment of one. Distribution of the number of HBV UMIs per cell in each cell subset (right panel).

See also Figure S1.
plasmacytoid dendritic cells (pDCs) (TCF4), and three different macrophage subsets (expressing TREM2, CD163, and FCN1, respectively). We observed a large diversity among the lymphocyte compartment with CD8+ T cells (CD8A), Th17 cells (CCR6, IL23A), γδ T cells (TRGC1), activated CD4 T cells (LEF1, OX40), natural killer (NK) cells (NKG7), and a distinct cluster of activated CD8+ T cells (CSF2 and TOX2). We analyzed infected cells using automated thresholding over the viral signal (Figure S1 J; STAR Methods). As expected, hepatocytes and apoptotic hepatocytes were strongly enriched among the infected cells (Figures 1H and S1K). We analyzed infected cells using automated thresholding over the viral signal (Figure S1 J; STAR Methods). As expected, hepatocytes and apoptotic hepatocytes were strongly enriched among the infected cells (Figures 1H and S1K).

Together, this extensive list of validations demonstrate that Viral-Track is a sensitive and accurate method to detect and identify, in an unsupervised manner, virus strains in diverse scRNA-seq samples, in different tissues, and at varying viral types and loads. Importantly, Viral-Track can be applied to human clinical samples to extract valuable insight into the biology of the host-virus interactions.

Viral-Track Identifies Infected versus Bystander Cells and Uncovers Virus-Induced Pathways

To further evaluate the accuracy of Viral-Track against a well-established model for tracking infection in single cells, we infected mice with a GFP-expressing LCMV virus (LCMV-GFP virus) (Medaglia et al., 2017). We performed MARS-seq on GFP+ splenocytes and infected MZB cells. See also Figure S2.
and total spleen cells 72 h post-infection and analyzed the sequenced cells (Figures S2A and S2B; STAR Methods). GFP+ cells were enriched for vUMI+ cells compared to total spleen (Figure S2A). We then calculated whether the cells positive for the LCMV-GFP signal (GFP+ cells) were similar to the ones designated by Viral-Track as containing viral UMIs (vUMI+). Following clustering and annotation, we observed similar proportions of GFP+ and vUMI+ cells across cell clusters (Figures 2A and S2C; R = 0.95, p = 9.0 * 10^-12), with monocytes, marginal zone B cells (MZBs), and macrophages being the major infected cell types. We then evaluated the transcriptional signatures within these two sets of cells by computing the Pearson correlation between each pair of cells. We observed similar distribution of Pearson correlation within the GFP+ and vUMI+ monocyte cells (Figure 2B) that was significantly higher (median correlation of 0.65, 0.64, and 0.51, respectively) than the correlation observed between GFP- vUMI+ bystander monocytes. We conclude that Viral-Track correctly identifies a homogeneous set of infected cells from in vivo scRNA-seq samples similar to the one identified by conventional reporter viruses, even in the more difficult scenario in which viral transcripts are poorly polyadenylated.

We next evaluated the ability of Viral-Track to detect host factors associated with virus replication. For this purpose, we developed a statistical method that detects differentially expressed genes based on data binarization and complementary log-log regression (STAR Methods; Methods S1). We used this approach to test for transcriptional differences between bystander and infected cells during spleen LCMV infection across the three main infected cell types: macrophages, MZB cells, and monocytes. We observed that MZB cells were the most influenced by the viral infection, compared to monocytes and macrophages (107, 42, and 3 genes upregulated, respectively; Z score >3) (Figure 2C).

We performed Gene Ontology enrichment analysis on the upregulated genes in MZB cells and observed a significant enrichment in several pathways, including "chromosome organization," "DNA replication," and "cell cycle," suggesting that LCMV triggers cell division in MZB cells (Figure 2D). Indeed, LCMV-infected MZB cells exhibited higher levels of cell cycle-related genes such as Smc2 (required for chromatin condensation), Cdc6 (regulator of DNA replication), and Stmn1 (regulator of mitotic spindle) (Figures 2E and S2D), but also fibrillin1 (Fbl), a host factor whose expression is required by several viruses (Deffrasnes et al., 2016) (Figure 2E). This is in line with a previous report highlighting the ability of LCMV to trigger an abortive form of cell division blocked in the G1 phase (Beier et al., 2019). Altogether, our results show that Viral-Track is sufficient to detect infected cells in vivo scRNA-seq data and infer the differential gene expression in infected versus bystander cells.

A Single-Cell Map of SARS-CoV-2 Infection in Mild and Severe Patients

COVID-19 is a viral disease caused by SARS-CoV-2 infection, which has recently been recognized as the cause for a pandemic (Wang et al., 2020a). Little is currently known about the course of the disease and how the virus interacts with the host immune system in its mild and severe manifestations. To gain insights on the infection course in humans, we performed scRNA-seq and Viral-Track analysis on BALF samples from three mild and six severe COVID-19 patients (Liao et al., 2020). In total, 50,615 cells passed quality control and were analyzed using the MetaCell algorithm (Baran et al., 2019) (Figure 3A; STAR Methods). Meta-cell analysis coarsely grouped the metacells into the myeloid, lymphoid, and epithelial lineages, and each lineage was further subdivided into smaller subsets (Figures 3A, 3B and S3A). Among epithelial cells, we identified epithelial progenitors (expressing SOX4), type II alveolar cells (AT2, expressing SFTPB), ciliated cells (FOXJ1), ionocytes (CFTR), goblet cells (MUC5B), and club cells (SCGB1A1; Figure S3B). Lymphoid cells consisted of several subtypes of CD4+ T cells, including naive CD4+ T cells (expressing CCR7), regulatory T cells (Treg, expressing FOXP3), and T follicular helper cells (Tfh, expressing CXCL13 and PDCD1), but also diverse CD8+ subsets, such as NK cells (NCAM1), resident memory CD8+ T cells (Tcm, CD8A, and ZNF683), effector CD8+ T cells (GZMA and GZMK), and cytotoxic CD8+ T cells (NLRC3, PRF1), as well as B cells (CD79A; Figure S3C). The myeloid compartment exhibited a high diversity of cell states, including neutrophils (FCGR3B), mast cells (CPA3), alveolar macrophages (FABP4), dendritic cells (DCs; FSCN1), and plasmacytoid DCs (pDC; TCF4) as well as a large diversity of monocytes (FCN1) and monocyte-derived macrophages (SPP1) sub-populations (Figure S3D). These results were robust across different analysis platforms (Liao et al., 2020).

Comparison of the cellular landscape of mild and severe patients revealed key differences in the composition of BAL...
samples (Figures 3B and 3C). We found changes to each of the three compartments (Figures 3D–3F and S3E–S3G). While alveolar macrophages and pDCs where enriched in the myeloid compartment in the mild patients, the severe patients’ myeloid cells were characterized by a patient-specific diversity associated with accumulation of neutrophils, FCN1 + monocytes, and monocyte-derived SPP1 + macrophages (Figures 3D and S3E). Additionally, NK cells and naive CCR7 + CD4 + T cells were consistently enriched across severe patients BAL, while ZNF683hi CD8 + Tm cells were specific to mild patients (Figures 3E and S3F). We also observed changes in the epithelial compartment, as severe patients exhibited higher numbers of club cells and AT2 cells (Figures 3F and S3G). By investigating expression patterns of shared gene expression programs, we observed that cytotoxic CD8 + cells and the CD4 + Th cells are the most proliferative compartments (Figure 3G), while a broad interferon type I response, a hallmark of viral response, is mainly expressed by neutrophils and, to a lesser extent, FCN1 + monocytes (Figure 3H). We next performed in-depth differential gene expression analysis between subsets characteristic of mild or severe patients. We found that CD4 + T cells in the severe patients exhibit a more naive phenotype, expressing higher levels of IL7R, CCR7, S1PR1, and LTb. The CD8 + Tm cells signatures are restricted to the mild patients and have higher levels of the effector molecules XCL1, ITGAE, CXCR6, and ZNF683 (Figure 3I). Comparing gene expression differences in myeloid types between severe and mild patients revealed disease severity-associated upregulation of inflammatory chemokine genes in SPP1 + monocyte-derived macrophages populations (CCL2, CCL3, CCL4, CCL7, and CCL8; Figure 3J), as well as genes associated with hypoxia or oxidative stress (HMOX1 and HIF1A), and downregulation of MHC class II (HLA-A and HLA-DQA1) and type I IFN genes (IFIT1 and OAS1). Alveolar macrophages displayed a severity-associated signature, including upregulation of the chemokines CCL18 and CCL4L2 and the cathepsins CTSL and CTSB (Figure 3J). Together, we identified dramatic differences between the mild and severe COVID-19 patients, including an inflammatory signature and a perturbed immune response associated with the severe manifestation of the COVID-19 disease. These also highlight potential immunotherapy treatment of the severe patients by targeting the hyper inflammatory response that is activated by inflammatory cytokines such as interleukin (IL)-6 and IL-8 (Liu et al., 2019) (Figure S3H).

Viral-Track Identifies Co-infection of SARS-CoV-2 with the Human Metapneumovirus

To characterize the in vivo crosstalk of SARS-CoV-2 with its human host, we applied Viral-Track on the data generated from the nine SARS-CoV-2 patients and the rich cellular landscape we identified. SARS-CoV-2 transcripts were detected in all six severe samples in variable amounts, ranging from less than 400 transcripts to more than 15,000 (Figures 4A and S4A). In contrast, no viral reads were detected in the three mild patients (Figure 4A). Coverage analysis revealed that the majority of the viral reads mapped to the 3’ end of the viral segment and corresponded to positive-stranded RNA (Figure 4B). This is in agreement with the coronaviruses transcription: due to a nested transcription process all genomic and subgenomic RNA molecules share the same 3’ end (Masters, 2006). We then analyzed the enrichment of vUMIs in the cell populations represented in the BAL samples. We observed a strong enrichment of viral reads in the ciliated and epithelial progenitor population, two known cellular targets of the virus, which express the main receptor of the SARS-CoV-2 virus ACE2, as well as TMPRSS2, a protease essential for SARS-CoV-2 entry (Figures 4C and S4B; Table S2) (Hoffmann et al., 2020). We also observed enrichment of SARS-CoV-2 reads in the SPP1 + macrophage population, suggesting either that SARS-CoV-2 can infect immune cells from the myeloid compartment or that SPP1 + macrophages phagocytose infected cells or viral particles. Differential gene expression analysis between vUMI + infected and vUMI + bystander SPP1 + macrophages in the patients with the highest viral load, revealed that infected macrophages have a higher expression of chemokines (CCL7, CCL8, and CCL18) and APOE, and a lower expression of TAOK1, a serine/threonine-protein kinase in the p38 MAPK cascade (Figure S4C). Interestingly, CD147 (also known as BSG), a potential new SARS-CoV-2 receptor (Wang et al., 2020b), is expressed by all cell types, including immune cells, suggesting alternative routes for the virus to infect these cells.

Often in cases of infectious diseases, the specific infecting virus is not known, or may be accompanied by co-infection with additional unknown viruses. Viral-Track applies an unsupervised mapping strategy and is optimally designed to systematically profile the source of infection or co-infections in human clinical samples. To our surprise, Viral-Track analysis of data from one of the severe patients (S1) revealed the presence of a second virus, the human metapneumovirus (hMPV) (NC_039199 Refseq sequence, Figure 4D) with more than one million reads mapped.
to hMPV in this specific patient. hMPV is a non-segmented, single-stranded, and negative-sense RNA virus that is responsible for upper and lower respiratory tract infections in mostly young (<5 years) children but can also target elderly as well as immunocompromised patients (Panda et al., 2014). hMPV has been implicated as a possible source of co-infection with the original SARS-CoV virus (Chan et al., 2003).

Coverage analysis revealed that most reads fall into the N, P, M, F, M2, SH, G, but not L, genes of hMPV (Figure 4E). We observed a typical pattern of biased scRNA-seq coverage, indicating that the N, P, M, F, M2, SH, and G genes are actively transcribed, and suggesting that the hMPV was active and replicating at the time of sample collection. Analysis of the viral UMI distribution across cells revealed a substantial viral load in a large subset of the cells, spanning hundreds to thousands vUMIs per infected cell (Figure 4F), independently of the total host UMIs in that cell (Figure S4D). We mapped the infected cells and characterized their distribution across cell types. The infected patient is characterized by high levels of monocytes and CD4+ T cells (Figure S4E). Unlike the SARS-CoV-2 virus infection map, hMPV-infected cells were highly enriched in the monocyte compartment but not in the epithelial and SPP1+ macrophage compartments (Figure 4G).

We tested whether the hMPV could alter the function of the infected monocytes, and therefore influence the course of the disease. Using Viral-Track, we detected a large number of up- and downregulated genes in infected monocytes compared to bystander monocytes (Figure 4H). Interestingly, several key receptor genes required for monocyte activation such as CD16 (FCGR3B), G-CSF receptor (CSF3R), and the formyl peptide receptor (FRP1) were downregulated in the infected compared to the bystander cells. Moreover, we observed a dramatic downregulation of type I interferon signaling and interferon stimulated genes (ISGs), including viral restriction factors, (e.g., IFIT3). A gene set enrichment analysis (Figure S4F) revealed a strong enrichment of interferon response genes in the downregulated gene set, suggesting that the hMPV is strongly downregulating the IFN response pathway. Several anti-inflammatory genes were upregulated, including LILRB4 (a potent inhibitor of monocyte activation) (Lu et al., 2009) and MITF, a transcription factor known to be a critical suppressor of innate immunity (Harris et al., 2018). Last, we observed a positive and significant association between total number of hMPV UMIs and production of type I IFN, highlighting that while hMPV dampens the response to type I IFN, production of this signal is highly restricted to a rare (~1%) population of cells with a high viral load (Figure S4G). Altogether, our analysis described the distribution of SARS-CoV-2-infected cells in patient’s BAL and revealed the presence of a viral co-infection by the hMPV that dampens the immune activation of the monocyte compartment in the infected patient.

Further large-scale analyses of mild versus severe patients need to be conducted to better understand if the co-infection is correlated or even causative in SARS-CoV-2 pathology.

**DISCUSSION**

The virosphere contains hundreds of thousands of species that constantly interact with their host cells. Over the years, several genomic techniques have been developed to detect virus-derived sequences in human samples. For instance, deep sequencing analyses are unbiased and sensitive in their ability to detect extremely rare viral sequences (Moustafa et al., 2017), but do not provide information about the infected cells and the cellular changes induced by the infection. Alternatively, it is possible to combine DNA probes with scRNA-seq to enrich for viral sequences and increase the sensitivity of the assays, but this requires prior knowledge of the viruses present in each sample (Zanini et al., 2018). Here, we present Viral-Track, a robust and unsupervised computational pipeline that can detect viral RNA in any scRNA-seq dataset without the need for experimental modifications or prior knowledge of the infecting agent. Viral-Track was benchmarked on data originating from various tissues, infected by viruses with marked differences in their RNA properties, and generated with different scRNA-seq platforms. We demonstrate that Viral-Track can readily provide essential information on infection status in clinical samples, identify infected cells, probe viral-induced transcriptional alterations, and reveal cases of co-infection. In practice, only 70%–85% of scRNA-seq reads map to the host genome and represent polyadenylated exonic host transcripts, whereas the remainder of the data is usually overlooked in analysis. We show that these unmapped scRNA-seq reads, in pathological human samples, potentially contain valuable information on viral infection and can be effectively used for viral genome assembly. Viral-Track can resolve complex cellular ecosystems perturbed by viral infection and provide an unbiased map of the infected cells, as well as the transcriptional perturbations induced by the virus at the single cell level. We combine Viral-Track with a novel statistical approach to detect differentially expressed genes from scRNA-seq data, therefore allowing the detection of gene expression changes triggered by viral infection and differentiating them from the more abundant bystander effects, such as type I IFN signaling, at the single cell level. Further advances will focus on applying Viral-Track on largescale datasets containing scRNA-seq data from dozens of samples, leading to robust single-cell viral metagenomic studies that characterize the viral evolution and interactions of virus-induced disease mechanisms with host genetics.

Here, we applied scRNA-seq and Viral-Track analysis to COVID-19 patient-derived samples to provide a cellular and viral atlas of the BAL lung cells from COVID-19 patients. This analysis revealed the diversity of the immune responses across COVID-19 patients and between mild and severe patients. We expect that as the pandemic keeps spreading and global research efforts grow, additional scRNA-seq samples from COVID-19 patients will be generated, including patients treated with emerging immunotherapies (Liu et al., 2019). Such an approach might help to solve key questions including the contribution of the humoral response (Iwasaki and Yang, 2020), the role of the IL6 pathway (Herold et al., 2020), and the immune memory induced by the virus (Promptetchara et al., 2020). Viral-Track can contribute to the global effort to identify the different cellular compartments that are targeted and affected by COVID-19 and other viruses and to detect possible co-infection by unexpected viruses. Co-infections are gaining recognition in the scientific and medical community as critical factors in disease prognosis (Zhang et al., 2020). So far, research focused mainly on co-infections of bacterial sources or of well-known viruses such as influenza A (Wu et al., 2020). Understanding the diversity of viral co-infections and their mechanisms of immune suppression at the cellular
and molecular level could therefore provide highly valuable information and lead toward possible therapeutic targets, especially for severe patients, whose treatment options are limited.

Limitations
Viral-Track is a new and powerful tool to decipher host-viral interactions. However, its impact is dependent on several factors, the most critical being the biochemical and pathophysiological properties of the virus. The absence of a poly(A) tail at the end of viral RNA molecules can significantly decrease their capture rate efficiency in current scRNA-seq techniques, as shown by the LCMV example. This may hinder Viral-Track’s ability to robustly identify infected cells or discern differential expression between infected and bystander cells in such viruses. Other properties of the viral RNA molecules, absence/presence of 5’ capping, nucleotide composition, or dependence on RNA binding proteins, may also affect capture efficiency, and as the technology develops, further research will focus on the classification of molecular features that facilitate or prevent virus identification by scRNA-seq. Notably, non poly(A)-based scRNA-seq techniques, such as RamDA-seq (Hayashi et al., 2018), can be potentially used when profiling these datasets.

Another limiting factor for Viral-Track’s applicability is the potential scarcity of viral reads and infected cells in the sample. As shown in our analysis of SARS-CoV-2-infected samples, only a limited number of viral reads are detected in some of the samples. This may be due to the specific stage of the disease (He et al., 2020), or sampling biases favoring mainly the lung immune populations, with lower representation of non-immune cells that are the primary targets of the virus. Therefore, future COVID-19 scRNA-seq studies should consider this limitation in their experimental design and aim for a better representation of the upper respiratory tissue and the lung parenchyma. Alternative approaches may rely on index sorting and single-cell transcriptome-trained sorting to design optimal gating strategies for capturing and enriching the stromal populations.

STAR Methods

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Mice
  - LCMV/VS induction
  - Subjects
- **METHOD DETAILS**
  - Lymph Node MARS-seq data generation
  - Influenza MARS-seq data generation
  - LCMV spleen MARS-seq data generation
  - 10X HBV liver data generation
  - 10X COVID-19 data generation

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Read mapping/alignment
  - Viral database and STAR Index building
  - Processing and filtering of the BAM files
  - Transcript reconstruction
  - MARS-seq data demultiplexing and UMI count
  - Drop-seq and 10X data download, pre-processing and demultiplexing
  - Analysis of the MARS-seq spleen LCMV dataset
  - Analysis of the 10X HBV liver dataset
  - Analysis of the COVID-19 BAL dataset
  - Testing for infection specificity in COVID-19 BAL dataset
  - Dichotomized differential gene expression analysis
  - Automate thresholding to detect HBV and hMPV infected cells
  - Gene set enrichment analysis

SUPPLEMENTAL INFORMATION
Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.05.006.

ACKNOWLEDGMENTS
We thank Dr. Noam Stern Ginosar and Dr. Yoav Golani for careful evaluation of the manuscript; Dr. Etienne Simon-Loriere, Thomas Jacquemont, and Alice Baltoulier for valuable advice; Táí Wiesel from the Scientific Illustration unit of the Weizmann Institute for artwork; and members of the Amit laboratory for discussions. I.A. is an Eden and Steven Romick Professorial Chair and supported by Merck KGaA (Darmstadt, Germany), the Chan Zuckerberg Initiative (CZI), an HHMI International Scholar award, the European Research Council Consolidator Grant (ERC-COG) 724471-HemTree2.0, an SCA award from the Wolfson Foundation and Family Charitable Trust, the Thompson Family Foundation, an MRA Established Investigator Award (509044), the Israel Science Foundation (703/15), the Ernest and Bonnie Beutler Research Program for Excellence in Genomic Medicine, the Helen and Martin Kimmel award for innovative investigation, a NeuroMac DFG/Transregional Collaborative Research Center grant, an International Progressive MS Alliance (PA-1604 08459), and an Adels Foundation grant. P.B. is supported by a PhD scholarship from the Ecole Normale Superieure, Paris. B.S. has received funding from the French Government’s Investissement d’Avenir program, Laboratoire d’Excellence “Integrative Biology of Emerging Infectious Diseases” (ANR-10-LABX-62-IBIEID). Z.Z. and Y.L. were supported by fundings from the National Natural Science Foundation of China (91442127 to Z.Z.; 81700540 to Y.L.).

AUTHOR CONTRIBUTIONS
P.B. designed and developed Viral-Track, performed various computational analyses, and wrote the manuscript. A.G. performed computational analyses and wrote the manuscript. Y.L., G.X., and S.Z. designed and performed experiments. Y.B. developed Viral-Track. E.D. contributed to data processing and analysis. R.B.-G., M.C., and C.M. designed and performed experiments. H.L. contributed to data analysis. A.D. contributed to data analysis, manuscript writing, and scientific communication. B.S. contributed to development of computational methods and bioinformatic analysis and wrote the manuscript. Z.Z. conceived, designed, and analyzed experiments and wrote the manuscript.

DECLARATION OF INTERESTS
The authors declare no competing interests.
REFERENCES


# STAR★METHODS

## KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antibodies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>santi-mouse TCRb biotin (clone H57-597)</td>
<td>Biolegend</td>
<td>Cat#:109203; RRID:AB_313426</td>
</tr>
<tr>
<td>anti-mouse CD3 biotin (clone 17A2)</td>
<td>Biolegend</td>
<td>Cat#:100243; RRID:AB_2563946</td>
</tr>
<tr>
<td>anti-mouse CD19 biotin (clone 6D5)</td>
<td>Biolegend</td>
<td>Cat#:115503; RRID:AB_313638</td>
</tr>
<tr>
<td><strong>Bacterial and Virus Strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vesicular Stomatitis Virus (VSV) Indiana Strain</td>
<td>In house</td>
<td>N/A</td>
</tr>
<tr>
<td>Lymphocytic choriomeningitis virus (LCMV)- Armstrong (Arm) strain</td>
<td>In house</td>
<td>N/A</td>
</tr>
<tr>
<td>LCMV-Arm-eGFP</td>
<td>In house</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Biological Samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COVID-19 BAL samples</td>
<td>Shenzhen Third People’s Hospital</td>
<td>N/A</td>
</tr>
<tr>
<td>HBV liver sample</td>
<td>Shenzhen Third People’s Hospital</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liberase TL</td>
<td>Roche</td>
<td>Cat#:5401020001</td>
</tr>
<tr>
<td>Dnase I, grade II</td>
<td>Roche</td>
<td>Cat#:10104159001</td>
</tr>
<tr>
<td><strong>Critical Commercial Assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromium Single Cell 3' Reagent Kit (v3 chemistry)</td>
<td>10X Genomics</td>
<td>1000075</td>
</tr>
<tr>
<td>Chromium Single Cell V(D)J Reagent Kits (v1 Chemistry)</td>
<td>10X Genomics</td>
<td>1000006</td>
</tr>
<tr>
<td><strong>Deposited Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw data files for the 10X COVID-19 and HBV patients</td>
<td>This paper</td>
<td>GEO: GSE145926</td>
</tr>
<tr>
<td>Raw data files for the LCMV/VSV single-cell RNA-seq</td>
<td>This paper</td>
<td>GEO: GSE149443</td>
</tr>
<tr>
<td><strong>Experimental Models: Organisms/Strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse: C57BL/6 WT</td>
<td>Jackson laboratories</td>
<td>RRID:IMSR_JAX:000664</td>
</tr>
<tr>
<td><strong>Software and Algorithms</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R (3.5.0)</td>
<td>The R project</td>
<td><a href="https://www.r-project.org">https://www.r-project.org</a></td>
</tr>
<tr>
<td>Python (3.6.5)</td>
<td>Python software foundation</td>
<td><a href="https://www.python.org">https://www.python.org</a></td>
</tr>
<tr>
<td>STAR (2.7.0)</td>
<td>Dobin et al., 2013</td>
<td><a href="https://github.com/alexdobin/STAR">https://github.com/alexdobin/STAR</a></td>
</tr>
<tr>
<td>Samtools (1.4.0)</td>
<td>Li et al., 2009</td>
<td><a href="http://www.htslib.org/download/">http://www.htslib.org/download/</a></td>
</tr>
<tr>
<td>StringTie (1.3.5)</td>
<td>Pertea et al., 2015</td>
<td><a href="https://ccb.jhu.edu/software/stringtie/">https://ccb.jhu.edu/software/stringtie/</a></td>
</tr>
<tr>
<td>UMI-tools (1.0.0)</td>
<td>Smith et al., 2017</td>
<td><a href="https://umi-tools.readthedocs.io/en/latest/">https://umi-tools.readthedocs.io/en/latest/</a></td>
</tr>
<tr>
<td>Pagoda2 (0.1.0)</td>
<td>Lake et al., 2018</td>
<td><a href="https://github.com/hms-dbmi/pagoda2/">https://github.com/hms-dbmi/pagoda2/</a></td>
</tr>
<tr>
<td>MetaCell (0.3.41)</td>
<td>Baran et al., 2019</td>
<td><a href="https://github.com/tanaylab/metacell">https://github.com/tanaylab/metacell</a></td>
</tr>
<tr>
<td>Cell Ranger (3.1.0)</td>
<td>N/A</td>
<td><a href="https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger">https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger</a></td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARS-seq reagents</td>
<td>Jaitin et al., 2014</td>
<td>N/A</td>
</tr>
</tbody>
</table>
RESOURCE AVAILABILITY

Lead Contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Ido Amit (ido.amit@weizmann.ac.il).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
The whole Viral-Track pipeline is freely available at https://github.com/PierreBSC/Viral-Track. The datasets generated during this study were deposited to the Gene Expression Omnibus (GEO) repository with accession codes GEO: GSE145926 and GSE149443.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice
C57BL/6 mice were purchased from Jackson Laboratories and bred and housed at the Weizmann Institute of Science animal facility, under specific pathogen-free conditions. Female mice, 6-8 weeks of age, were used for all experiments. Experimental protocols were approved by the Weizmann Institute of Science Ethics Committee and were performed according to institutional guidelines.

LCMV/VSV infections
For LCMV infection, 1x10⁵ Focus-Forming Units (FFUs) of the LCMV-Arm strain were injected. For VSV, 1x10⁶ Plaque-Forming Units (PFUs) of the VSV Indiana strain were used. Mice were anesthetized and viruses administered by intradermal injection into the ear pinna. 24h later, mice were sacrificed and auricular LN were harvested.

Subjects
This study was conducted according to the principles expressed in the Declaration of Helsinki. Ethical approval was obtained from the Research Ethics Committee of Shenzhen Third People’s Hospital. All participants provided written informed consent for sample collection and subsequent analyses.

METHOD DETAILS

Lymph Node MARS-seq data generation
To prepare single cell suspensions for MARS-seq and flow cytometry, auricular LNs were digested in IMDM containing 100mg/mL Liberase TL and 100mg/mL DNase I (both from Roche, Germany) for 20 minutes at 37°C. In the last 5 minutes of incubation, EDTA was added at a final concentration of 10mM. Cells were collected, filtered through a 70mm cell strainer, washed with IMDM and maintained strictly at 4°C. Cells were sorted with FACSARIA-FUSION (BD Biosciences, San Jose, CA). Prior to sorting, all samples were filtered through a 70-mm nylon mesh. Isolated cells were single cell sorted into 384-well cell capture plates containing 2 mL of lysis solution and barcoded poly(T) reversetranscription (RT) primers for single-cell RNA-seq (Jaitin et al., 2014). Four empty wells were kept in each 384-well plate as a no-cell control for data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion into the lysis solution, snap frozen on dry ice, and stored at –80°C until processing. Single-cell RNA-seq libraries were prepared as previously described (Jaitin et al., 2014). In brief, mRNA from single cells sorted into capture plates were barcoded and converted into cDNA and then pooled using an automated pipeline. The pooled sample was linearly amplified by T7 in vitro transcription, and the resulting RNA was fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, RT, and PCR. Each pool of cells was tested for library quality and concentration as described previously (Jaitin et al., 2014).

Influenza MARS-seq data generation
Full description of the protocol used to generate the Influenza A lung data can be found in Steuerman et al. (2018). Influenza PR8 H1N1 influenza virus (A/Puerto Rico/8/34) was cultivated in hen egg anion. 40μL of diluted virus (6x10⁷ PFU per mouse) were inoculated intranasaly to the mice, or 40μL of PBS for the control mice. Mice were killed 48 or 72h post infection and the lung perfused. Immune and none-immune cells were then extracted using two different extraction protocols before being single-cell sorted in 384-well plates and sequenced using the original MARS-seq protocol (Jaitin et al., 2014).

LCMV spleen MARS-seq data generation
Description for the full protocol used to generate the NICHE-Seq spleen data can be found in Medaglia et al. (2017). Briefly female mice received 1x10⁶ FFU of LCMV-Arm or LCMV-Arm-eGFP in the footpad. 72 hours after injection, spleens were harvested and forced through a 70μm mesh to form a single-cell suspension. Cells were then single-cell sorted using a SORP-aria into 384-well
plates containing lysis buffer before processing the plate according to the MARS-seq protocol (Jaitin et al., 2014). All infectious work was performed in designated Biosafety Level 2 (BSL-2) and BSL-3 workspaces in accordance with institutional guidelines.

**10X HBV liver data generation**

The approximately 1 cm long Liver biopsy was homogenized by mincing with scissors into smaller pieces (~0.5 mm² per piece). Then the tissue was transferred into 10 mL of enzyme mix consisting of 0.3 mg/ml collagenase type IV (Sigma, C9891) and DNase I (Sigma, D5025) for mild enzymatic digestion for 1 h at 37°C while shaking. 5 mL of Dulbecco’s phosphate-buffered saline (DPBS, Thermo, 14190250) supplemented with 5% FBS was added to interrupt digestion and dissociated cells in suspension were passed through a 40 μm strainer and centrifuged at 300 g for 5 min at 4°C. Erythrocytes were lysed using Ammonium-Chloride-Potassium (ACK, Thermo, A1049201), and finally cells were re-suspended in DPBS supplemented with 1% FBS at the concentration of 2, 000 cells/μl for scRNA-Seq. The single-cell capturing and downstream library constructions were performed using the Chromium Single Cell 3’/V3 library preparation kit according to the manufacturer’s protocol (10x Genomics). Full-length cDNA along with cell-barcode identifiers were PCR-amplified and sequencing libraries were prepared and normalized to 3 nM. The constructed library was sequenced on BGI MGISEQ-2000 platform. The Cell Ranger Software Suite (Version 3.1.0) was then used to perform sample de-multiplexing, barcode processing and single-cell 3’ UMI counting with human GRCh38 as the reference genome.

**10X COVID-19 data generation**

20 mL of BALF was obtained and placed on ice. Following BALF through a 100 μm nylon cell strainer to filter out lumps, the supernatant was centrifuged and the cells were re-suspended in the cooled RPMI 1640 complete medium. Then the cells were counted in 0.4% trypan blue, centrifuged and re-suspended at the concentration of 2 × 10⁶/ml for further use. Total 11 μl of single cell suspension and 40 μl barcoded Gel Beads were loaded to Chromium Chip A to generate single-cell gel bead-in-emulsion (GEM). The poly-adenylated transcripts were reverse-transcribed later. The single-cell capturing and downstream library constructions were performed using the Chromium Single Cell 5’ library preparation kit according to the manufacturer’s protocol (10x Genomics). Full-length cDNA along with cell-barcode identifiers were PCR-amplified and sequencing libraries were prepared and normalized to 3 nM. The constructed library was sequenced on BGI MGISEQ-2000 platform. Each sample was sequenced on a different sequencing run to avoid contamination between samples. The Cell Ranger Software Suite (Version 3.1.0) was then used to perform sample de-multiplexing, barcode processing and single-cell 5’ UMI counting with human GRCh38 as the reference genome. A more extensive description of the data generation process can be found in Liao et al. (2020).

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**Read mapping/alignment**

Reads were aligned using STAR 2.7.0 (Dobin et al., 2013) in the two-pass mode using the following parameters:--runThreadN was set to 14, --outSAMattributes to ‘NH HI AS NM XS’,--outSAMtype to ‘BAM SortedByCoordinate’,--outFilterScoreMinOverLread to 0.6, --outFilterMatchNminPverLread to 0.6, and--twopassMode to ‘Basic’.

**Viral database and STAR Index building**

As STAR performance drastically dropped when the reference index contains more than 10,000 scaffold/chromosomes, we decided to base our analysis on the limited, but high-quality, viruSITE database (Stano et al., 2016), derived from the NCBI Refseq database. The corresponding FASTA file was downloaded from the viruSITE website (http://www.virusite.org/archive/2019.1/genomes.fasta.zip). STAR indexes were built for both human and mouse samples using respectively the GRCh38 (hg38) and GRCm38 (mm10) reference genomes in addition with the whole viruSITE database. Both reference genomes were downloaded at http://www.ensembl.org/useast.ensembl.org/info/data/ftp/index.html?redirectsrc=http://www.ensembl.org/%2Finfo%2Fdata%2Fftp%2Findex.html. For the analysis of COVID-19 patients we added the official SARS-CoV-2 reference genome from the Refseq database (NC_045512.2) as it has not been added to the viruSITE database yet. In total this database contains 11988 viral segments from 9431 different viruses.

**Processing and filtering of the BAM files**

We empirically observed that viral genome sequences can contain highly repetitive subsequences and can therefore create false positive signal. Moreover, some viral genes can share a significant similarity with host genes and also generate mapping artifacts. To remove those, we implemented a strict filtering approach where for each viral segment, a list of mapping features are measured and used to estimate the quality of the mapping.

Following the alignment, the resulting BAM files were processed using the samtools toolbox (Li et al., 2009): first the BAM files were indexed using the samtools index command. Virus segment with more than 50 mapped reads were detected using the samtools idxstats command and a unique bam file was then created for each of the viral segment using the samtools view command. Each viral bam files were then loaded into an R environment using the readGAlignments() function from the GenomicAlignments package. Various features were then extracted to assess the quality of the mapping.
● The length of the longest mapped contig computed using the coverage() function.
● The percentage of the viral segment that is mapped, also computed using the coverage() function.
● The mean sequencing quality of the mapped reads.
● The number and percentage of uniquely mapped reads.
● The mean sequenced entropy of the mapped reads defined as follows: for each mapped read each nucleotide frequency was extracted using the alphabetFrequency() function of the Biostring package and averaged over the reads. Then the corresponding Shannon entropy was computed using napierian logarithm.

Empirically we determined that a mean sequence entropy bigger than 1.2, a coverage bigger than 5% and the longest contig bigger than three times the mean read length is sufficient to consider a viral segment to be present. This filter configuration eliminated all manually identified artifacts in the various benchmarked datasets and was used unchanged in the HBV and COVID-19 patient data analysis.

When using this strategy, we observed two different kinds of ‘contamination’:

● the first one consists of the detection of retroviruses specific to the sequenced host species: this is likely due to the expression of host endogenous retro-viral elements that highly similar to ‘real’ retroviruses.
● the second is the presence of a plant virus, the Tomato brown rugose fruit virus: this is an emerging virus that infects tomatoes and peppers and is endemic in Israel and Jordan. It is highly contagious and spreads easily. We detected this virus only in samples sequenced in Rehovot (Israel) suggesting that it was due to an airborne contamination.

To improve computation speed, this step was parallelised using the doParallel R package.

Transcript reconstruction
As viral genomes are poorly annotated, we decided to systemically reconstruct the transcriptome of each viral segment detected using the transcript assembler StringTie (Pertea et al., 2015). StringTie was used with default parameter except the minimum isomorph abundance parameter -i which was set to 0.01 to detect lowly abundant transcripts and the minimal distance between two transcript -g set to 5.

MARS-seq data demultiplexing and UMI count
In order to have a UMI-counting procedure adapted to viral genomes, i.e that distinguish spliced and un-spliced RNA molecule, we developed an in-house R script based on the GenomicRanges, GenomicAlignments and GenomicFeatures packages that used the same strategy as the commercial CellRanger toolkit. Briefly cell barcodes were extracted and compared with a cell barcode whitelist provided by the MARS-seq2 demultiplexing pipeline (Keren-Shaul et al., 2019): cell barcode that belong to the whitelist were kept while cell barcodes that did not belong to the whitelist but that has a highly similar barcode (Hamming distance equal to one, computed using the stringdist() function from the stringdist package) were corrected and kept. UMIs were also extracted and mono-nucleotide UMIs filtered out. Hamming distances between UMIs assigned to the same cell and the same gene were then computed similarly to cell barcodes and UMIs with a Hamming distance equal to one were aggregated and considered as redundant UMIs. Lastly the mapping file was loaded using the readAGAlignments() function from the GenomicAlignments package and reads were assigned to a specific viral gene using the findOverlaps() function from the same package. In case the read mapped to a given viral transcript but was not assigned to any viral gene, it was considered as coming from an un-spliced viral RNA molecule.

Drop-seq and 10X data download, pre-processing and demultiplexing
Fastq files were downloaded through the SRA Explorer tool (https://sra-explorer.info/#). Identification and correction of cellular barcode, as well as UMI demultiplexing was performed using UMI-tools 1.0.0 (Smith et al., 2017). First, cell barcodes were extracted and a putative whitelist computed using the umi_tools whitelist command with the parameters ‘--bc-pattern = CCCCCCCCCCCCCNNNNNNNNNNNNNNN-stdin ’ for the 10X data. For Drop-Seq data the same command is used except the--bc-pattern option set to CCCCCCCCCCCCCNNNNNNNNN. Collapsing of the UMIs is performed using the command umi_tools extract with parameters ‘--bc-pattern = CCCCCCCCCCCCCCCCCNNNNNNNNN --stdin --filter-cell-barcode’ on the 10X data and with the same command for Drop-seq data except for the--bc-pattern option set to ‘CCCCCCCCCCCCNNNNNNNNN’. Following the mapping of the reads to viral genomes and transcript assembly, the mapped reads were assigned to transcripts using the R package Rsubread through the function featureCounts() with default parameters. The command umi_tools tools count is then used to compute the final expression table with the following parameters:--per-gene--gene-tag = XT--assigned-status-tag = XS--per-cell.

Analysis of the MARS-seq spleen LCMV dataset
High-level analysis were performed using the R-based Pagoda2 pipeline (https://github.com/hms-dbmi/pagoda2/) (Lake et al., 2018) in addition to an in-house R script. Briefly UMI table were loaded and cells with less than 350 UMIs were removed. Lowly abundant genes (less than 100 UMIs) were also removed from analysis. Analysis of the filtered dataset was then performed similarly to our previous paper (Blecher-Gonen et al., 2019) by using the 1500 most variant genes and 100 PCs for dimensionality reduction. kNN graph was build with a parameter K equal to 30 and Louvain’s method used for clustering. Cluster marker genes were computed by using...
the getdiffGenes function with default parameters. Data were visualized using UMAP (McInnes et al., 2018) implemented by the uwot package.

**Analysis of the 10X HBV liver dataset**
High-level analysis were performed using the R-based Pagoda2 pipeline (https://github.com/hms-dbmi/pagoda2) (Lake et al., 2018) in addition to an in-house R script. Briefly UMI table were loaded and cells with less than 1,000 UMIs were removed. Lowly abundant genes (less than 50 UMIs) were also removed from analysis. Analysis of the filtered dataset was then performed similarly to our previous paper (Blecher-Gonen et al., 2019) by using the 1000 most variant genes and 100 PCs for dimensionality reduction. kNN graph was build with a parameter K equal to 30 and Louvain’s method used for clustering. Cluster marker genes were computed by using the getdiffGenes function with default parameters. Data were visualized using UMAP (McInnes et al., 2018) implemented by the uwot package.

**Analysis of the COVID-19 BAL dataset**
Upstream processing of reads was done with the CellRanger toolkit, resulting in a UMI table of 75,790 cells with a median UMI count of 2,442, and a median of 888 genes per cell. Cells with less than 500 UMI, or more than 50% mitochondrial genes were excluded.

We used the MetaCell package (Baran et al., 2019) to group single cells from all patients into groups of transcriptionally homogeneous groups, termed metacells. We first removed mitochondrial genes, ERCC, and the diverse immunoglobulin genes (IGH, IGK, and IGL).

Gene features for metacell covers were selected using the parameter Tvm = 0.4, total umi > 30, and more than 4 UMI in at least 3 cells (using the functions mcell_gset_filter_varmean, and mcell_gset_filter_cov). We excluded gene features associated with the cell cycle, stress response, type I interferon, and batch-specific genes via a clustering approach (using the functions mcell_gset_rpt_cor_.-anchors and mcell_gset_split_by_dsmat). To this end we first identified all genes with a correlation coefficient of at least 0.1 for one of the anchor genes TOP2A, MKI67, PCNA, MCM4, UBE2C, STMN1 (cell cycle), HSPA1B, HSPA1A, DNAJB1, HSPB1, HSPA6, FOS, JUN, CCL4, CCL4L2, MT1E, MT1X, MT1F, TYMS, GADPH, DUT, HMGB2 (stress and batch effect), IFIT1, IFI19, OAS1L, IRF7, IFI1, STAT1, and STAT3 (type I IFN). We then hierarchically clustered the correlation matrix between these genes (filtering genes with low coverage and computing correlation using a down-sampled UMI matrix) and selected the gene clusters that contained the above anchor genes. We thus retained 402 genes as features (Table S3). We used metacell to build a kNN graph, perform boot-strapped co-clustering (500 iterations; resampling 70% of the cells in each iteration), and derive a cover of the co-clustering kNN graph (K = 100). Outlier cells featuring gene expression higher than 4-fold than the geometric mean in the metacells in at least one gene were discarded.

Annotation of the metacell model was done using the metacell confusion matrix and analysis of marker genes. Detailed annotation within the myeloid, lymphoid and epithelial compartments was performed using hierarchical clustering of the metacell confusion matrix (Figure S3A) and supervised analysis of enriched genes. Metacells enriched for markers from more than one lineage (either T (TRBC2), myeloid (S100A8, C1QB), epithel (KRT18), and plasma cells (XBP1)) were marked as doublets and discarded from further analysis. We additionally discarded metacells of erythrocytes or plasma cells from further analysis.

To derive cell cycle and type I interferon response co-expressed gene modules, we used a clustering-approach as described in the previous paragraphs (using the functions mcell_gset_rpt_cor_anchors and mcell_gset_split_by_dsmat) on a set of cell cycle and interferon genes. We clustered, and manually inspected the resulting clusters, retrieving 72 cell-cycle related and 65 interferon related genes (Table S3).

To extract proportion of proliferating cells (Figure 3G), we calculated for each cells the number of cell-cycle related transcripts per 1,000 UMI. Cells with more than 8 transcripts were determined proliferating.

**Testing for infection specificity in COVID-19 BAL dataset**
To test for SARS-CoV-2 infection specificity in different cell populations, we computed for each metacell the total number of host UMIs (hUMI) and viral UMIs (vUMI) in the three severe patients (S1-3). We then computed for each metacell its expected vUMI count, based on its total UMI count (hUMI + vUMI) and the total vUMI proportion across all cells. Figure 4C shows log2 fold change between the observed and expected UMI in each metacell, after adding a regularization factor ( = 5) for each factor. Log2 fold change for the 27 subsets in Figure 3A, and calculated for each severe patient separately is shown in Table S2.

Testing for hMPV infection specificity was done in a similar manner. However, since UMI distribution across cells was abundant and heavy-tailed, we computed for each metacell the expected number of vUMI+ cells instead of its total vUMI count. A cell was determined vUMI+ if it had more than 10 viral UMI, as determined by automatic thresholding (Figure 4F). Figure 4G shows log2 fold change between the observed and expected vUMI+ cells in each metacell, after adding a regularization factor ( = 5) for each factor.

**Dichotomized differential gene expression analysis**
ScRNA-seq data are intrinsically noisy data with a large proportion of zeros values (previously called dropouts) due to limited sampling of the initial mRNA molecule pool. In addition, cell library size is a major cofounder variable, even after common normalization.
procedures such as TPM, especially for lowly expressed genes (Hafemeister and Satija, 2019). We therefore improved the method used in our former paper (Blecher-Gonen et al., 2019) that was based on logistic regression.

Briefly our method is based on the global trend of the field that consists in sequencing large amounts of cells but with a limited sequencing depth. Such approach will produce mostly ‘binary’ data and seem to be represent the best compromise on a cost/efficiency point of view (Svensson et al., 2019). So far, several statistical models have been used to model and analyze scRNA-seq count data, most of them being based on the zero-inflated negative-binomial (ZINB) distribution (Finak et al., 2015; Kharchenko et al., 2014). However, recent studies suggested that those models are too complex and introduce artificial complexity (Silverman et al., 2018; Svensson, 2020; Townes et al., 2019). We hypothesize that with such binary data, current models will not fit properly and more suited ones need to be developed.

We therefore developed a new approach based on the binomial complementary Log-log regression (cloglog model): once a given group of cells has been isolated, through Louvain’s clustering for instance (Blondel et al., 2008), we first dichotomized gene expression (if the normalized expression is bigger than 0 the gene is considered as expressed) and then computed a binomial Generalized Linear Model (GLM) with a complementary log log link function (cloglog) using the glm() R function. To mitigate the variation of the library size as well as the global effect of the infection (bystander effect), we include both variables in the regression model. The corresponding p value are then computed using a Likelihood Ratio Test (LRT) and then corrected using Benjamini Hochberg correction (Benjamini and Hochberg, 1995).

For a more comprehensive description of the approach please see Methods S1.

**Automate thresholding to detect HBV and hMPV infected cells**

In the case of the HBV and hMPV infections, we observed that cells could contain from one to several thousands UMIs. In order to know which cells were really infected and which one contain viral UMIs due to ambient contamination, we decided to apply Otsu’s thresholding after logarithmic transformation. Otsu’s method was implemented using an in-house R script (Otsu, 1979).

**Gene set enrichment analysis**

Gene set enrichment analysis was performed using the online GSEA tool https://www.gSEA-msigdb.org/gsea/index.jsp (Liberzon et al., 2015; Subramanian et al., 2005). The enrichment analysis was performed using the Hallmark and Gene Ontology biological process databases. False detection rate was set to 0.05. Only the top 10 most enriched terms were reported.
Figure S1. Benchmarking of Viral-Track on Diverse Infection Models, Related to Figure 1
A. Graph chart representing the different steps of the Viral-Track pipeline. B-D. Results of Viral-Track analysis performed on LCMV spleen, LCMV lymph node and VSV lymph node datasets, respectively. Viral segments with more than 50 mapped reads are plotted. (E). Number of detected LCMV (left panel) and VSV (right panel) reads in the different samples from the lymph node experiment. F. Results of Viral-Track analysis performed on the in-vitro HSV-1 data. G. Quantification of the number of HSV-1 reads in HSV-1 infected and control samples. (H). Results of Viral-Track analysis performed on the in-vitro HIV data. I. Quantification of the number of HIV reads in HIV infected and control samples. J. UMAP plot of the liver HBV data, dots are colored by cell subset assignment based on Louvain clustering. K. UMAP plot of the liver HBV data. infected cells are colored in orange and bystander cells in gray.
Figure S2. Comparison of Viral-Track Performance to Fluorescence Tagging Techniques, Related to Figure 2
A. Proportion of vUMI+ cells from total spleen and the LCMV-GFP+ population. B. UMAP plot of the spleen LCMV data, spots are colored based on Louvain clustering. C. UMAP plot of the spleen LCMV data, bystander cells are colored in gray, vUMI+ cells are colored in red and GFP+ cells in green. D. Mean gene expression in bystander and infected MZB cells. Genes with a log2FC bigger than 1 or lower than −1 and a corrected p value lower than 0.01 are colored in orange.
Figure S3. Detailed Molecular and Cellular Profiling of COVID-19 BAL Samples, Related to Figure 3
A. The confusion matrix of the MetaCell model shown in Figure 3A. Entries denote for each pair of metacells the propensity of cells from both metacells to be clustered together in a bootstrap analysis. B-D. Gene expression profiles of cells belonging to the epithelial (B), lymphoid (C), and myeloid (D). In A-D, color bars indicate association to 27 cell subsets depicted in Figure 3A. E-G. Quantification of the frequency of specific cell subsets in the myeloid (E), lymphoid (F), and epithelial (G) compartments, across the nine patients. Diamond marks patient S1, co-infected with the human Metapneumovirus (Figures 4D-4H). Horizontal lines indicate mean frequency. (H). Projection of IL6 and IL8 (CXCL8) expression on the 2D map shown in Figure 4A. Colors represent expression quantiles.
Figure S4. Viral-Track Performance on COVID-19 BAL Samples, Related to Figure 4

A. Results of Viral-Track analysis performed on samples with highest viral load (patients S2 and S3). B. Mean normalized expression of ACE2, TMPRSS2 and BSG across the 27 cell subsets. C. Log2 fold change between vUMI+ and vUMI- SPP1+ monocyte-derived macrophages in patient S2 (x axis) and patient S3 (y axis). D. Relation between total human and viral UMIs in cells from patient S1. E. Projection of cells from patient S1, co-infected with hMPV, on the metacell map from Figure 3A. F. Enrichment analysis of the downregulated genes in hMPV infected monocytes. G. Number of hMPV UMIs in cells producing type I IFN or not. P value was computed by fitting a logistic regression predicting if a cell would produce type I IFN using total host and viral UMIs.
1 scRNAseq data set : Mathematical primer

1.1 A quick review on scRNAseq data set modeling

Since the first generation of scRNAseq technologies, several attempts were made to model and analyze scRNAseq data sets using rigorous probabilistic models. scRNAseq data are hard to deal with for several reasons: they are high-dimensional (thousands of genes expressed) and highly sparse (more than 90 percent of the values are zeros). Moreover, intrinsic dimensionality and complexity of the data can be high due to the presence of dozens of cell types and states.

The first developed models were based on highly parametrized Zero-Inflated Negative Binomial (ZINB) distribution and were fitted using Bayesian derived methods (Finak et al., 2015; Kharchenko et al., 2014). However, the exponential increase in the number of sequenced cells made such techniques too computational heavy. Moreover, recent papers suggested that the sparsity observed in scRNAseq data is mostly due to limited sampling of the original RNA molecule pool and can be modeled using simple Negative Binomial (NB) distribution without using any zero inflation (Svensson, 2019; Silverman et al., 2018). Attempts were also made to fit multinomial distribution with success but at the cost of an increased mathematical and computational complexity (Townes et al., 2019).

As the global trend in the scRNAseq field favors in sequencing more cells very shallowly (15,000 reads per cells, corresponding to a thousand of Unique Molecular Identifiers (UMIs) per cell), development of robust and simple methods that can deal with such sparse data without using complex and over-fitting models is required (Svensson et al., 2019).
1.2 Mathematical notations

We will consider a scRNAseq experiment with m cells and p genes. Let U a m*p matrix where $U_{ij}$ corresponds to the number of UMIs identified for gene j in cell i. Total number of UMIs for cell i is called $L_i$. Directly $L_i = \sum_{j=1}^{p} U_{ij}$. We will refer to this value as the cell library size in the manuscript.

ScRNAseq data can be normalized using different strategies to mitigate the variations of cell library sizes. The most common one consist in computing the ratio of UMIs of a cell coming from a given gene. Then the normalized expression of gene j in cell i is therefore $\pi_{ij} = \frac{U_{ij}}{L_i}$. $\pi_{ij}$ is usually multiplied by $10^6$ to provide the count per million (CPM) of the gene. Log-transformation is then applied with a pseudo count of 1. The transformed variable is then expected to follow a normal distribution and is therefore equal to $LCPM_{ij} = \log(\pi_{ij} \times 10^6 + 1)$.

In the next part of the manuscript we will use discretized gene expression. If not stated otherwise, the discretized gene expression of gene j in cell i is $D_{ij}$ and is equal to:

$$D_{ij} = \begin{cases} 
1 & \text{if } U_{ij} > 0 \\
0 & \text{if } U_{ij} = 0 
\end{cases} \tag{1}$$

1.3 Shallow scRNAseq data are mostly binary

While first scRNAseq data sets consist in few cells sequenced with millions of reads, the global trend is to sequence more cells very shallowly (Svensson et al., 2019). We therefore wondered if such approach did not result in data sets that are mostly binary, i.e quantification of gene expression can be summarized as distinguishing cells expressing or not the gene of interest. To test this hypothesis we looked at the distribution of non-null values from a massively parallel single-cell RNA sequencing (MARS-seq, (Keren-Shaul et al., 2019)) experiment that consist in single-cell sequencing of different zones of the mouse spleen using a technique called NICHE-seq (Medaglia et al., 2017). MARS-seq relies on polyT primers and therefore generates 3’ biased but also shallow (mean library size around 15,000 reads) and can therefore be considered as representative of other methods such as the commercial 10X Chromium technology. The data set was processed using PAGODA2 pipeline (Lake et al., 2018) and cells were clustered using Louvain’s graph clustering method (Blondel et al., 2008).

For each of the 22 identified cell cluster types, we computed the mean proportion of non-null UMI counts that are bigger than one. While this value varies between clusters, we observed values between 46% and 65%, suggesting that the majority of the non-null values are equal to one. Moreover, these values were computed by using all genes, including highly expressed genes such as
ribosomal genes, and exogenous spike-in (ERCC). In practice, the genes of biological interest are not the most highly expressed but the most variable, thus reinforcing our hypothesis of binary data.

![Figure 1: Proportion of non-null values equal or bigger than one across cell clusters](image)

We then looked at the relationship between the cell library size and two features describing the amount of information available in each cell: the number of genes detected in the cell and the proportion of non-null UMI counts that are bigger than one. We observed for both measurements a linear relationship with library size when using logarithmic scales. However while a 40 fold increase in cell library size only doubles the amount of count values bigger than 1 (30% to 60%), it multiplies by 10 the amount of genes detected. Moreover, Pearson correlation was significantly higher between library size and the number of genes detected than between library size and the proportion of non-null count values bigger than one. We therefore conclude that an increase in cell library size results in a dramatic increase in the number of genes detected but only marginally affects our ability to distinguish different level of expression. Therefore the use of a discretized version of the data fully makes sense.

## 2 Description of the statistical models

### 2.1 First model: homogeneous gene expression

Let’s consider a gene expressed in a homogeneous cell population of $m$ cells. The gene of interest represents a proportion $\alpha$ of the original RNA molecule pool and each cell $i$ has $L_i$ UMIs sampled. We will consider that when one UMI is sampled from the RNA pool, it has a probability $\alpha$ of coming...
from the gene of interest and $1 - \alpha$ of not coming from that gene. As we assume here independent sampling, the probability of not detecting the gene in cell $i$ across all $L_i$ sampled molecules is equal to $(1 - \alpha)^{L_i}$. Therefore the probability of detecting it is equal to:

$$P(\text{Gene detected in cell } i) = 1 - (1 - \alpha)^{L_i} \quad (2)$$

This generative model can be used to identify the most suited statistical model. In our case we used a binomial regression with a Complementary Log-Log (cloglog) link function instead of the conventional logistic link function. The cloglog function is equal to:

$$F(x) = 1 - \exp(-\exp(x)) \quad (3)$$

Instead of using directly the library cell size as an explanatory variable, we used the log transformed library cell size with a coefficient set to 1 in addition to a simple intercept termed $\mu$. When injected to \ref{eq:log_to_cloglog} we obtained the following results:

$$F(\mu + \log(L_i)) = 1 - \exp(-\exp(\mu + \log(L_i))) \quad (4)$$

$$= 1 - \exp(-\exp(\mu) \cdot L_i)$$

Equation \ref{eq:gene_detection} can be rewritten:

$$P(\text{Gene detected in cell } i) = 1 - \exp(\log(1 - \alpha) \cdot L_i) \quad (5)$$
And therefore the $\alpha$ parameter can be derived from the intercept $\mu$ of the fitted model as:

$$\alpha = 1 - \exp(-\exp(\mu))$$ \hspace{1cm} (6)

### 2.2 Second model: heterogeneous gene expression

Now that we have established a link between a simple generative model and the cloglog regression, we can develop more complex models where gene expression can change across cells due to different variables (cell type, cellular stimulation...). To do so, we can simply use the formalism of Generalized Linear Models (GLM) and add a linear additive term before applying the cloglog link function. In the case of a single discrete explanatory variable with $p$ different values, we will use a coefficient variable $\beta$, a $p$ row vector, and a design matrix $M$ of size $n \times p$. Therefore:

$$P(Gene\ detected\ in\ cell\ i) = 1 - \exp(-\exp(\mu + \log(L_i) + \beta \ast M))$$ \hspace{1cm} (7)

In the case of simple variables which takes only two different values (stimulated and control cells), the ratio of gene expression level between the two groups (i.e of the $\alpha$ parameter) can be derived from equation 6:

$$Expression\ ratio = \frac{1 - \exp(-\exp(\mu + \beta))}{1 - \exp(-\exp(\mu))}$$ \hspace{1cm} (8)

### 3 Implementation of the model

#### 3.1 Model fitting with R

R is a powerful programming language and environment that is able to efficiently fit a large variety of statistical models, including GLMs, and to test statistical significance of variables used in the model. The models described in the previous sections can be easily fitted with the following command:

```r
model = glm(D~1+offset(log(L))+Variable, family = binomial(link="cloglog"))
```

Here we consider that $D$ corresponds to the discretized gene expression level, $L$ the total UMI count and $Variable$ a vector describing to which group belong each cell. This model needs to be fitted for each gene recursively in order to detect the most differentially expressed genes.

#### 3.2 Testing for statistical significance

Now that the model is fitted, the statistical significance of the variable contribution can be tested using a Likelihood Ratio Test (LRT). This is done simply by writing:
Term significance = summary(model)

The statistical significance of the intercept and variable can then be extracted from the summary object. It is possible to remove genes whose model fitting did not succeed by throwing out models where the intercept term is not significant, i.e. the detection probability does not increase with library size. Lastly, when the p-values associated with the variable are computed for each gene of interest, they are corrected using Benjamini Hochberg correction (Benjamini and Hochberg, 1995) to remove false positive.
References


Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2018). Naught all zeros in sequence count data are the same. *bioRxiv.*


4 Deciphering the state of immune silence in fatal COVID-19 patients

The second paper of this chapter was done in collaboration with the university of Verona (Italy). We established a cohort of more than 30 COVID-19 patients and performed extensive scRNA-seq analysis of blood and bronchoalveolar lavage samples. By combining those data with clinical information, we identified cell types and transcriptional programs associated with the severity of the disease. In addition we studied viral co-infections and how they can alter the systemic immune response of the patients revealing that HSV-1 can infect the lung of severe COVID-19 patients. While writing this thesis, this paper was not published yet and was put on medrxiv repository.
Deciphering the state of immune silence in fatal COVID-19 patients

Pier Bost,1, 2, 3, 11 Francesco De Sanctis,4, 11 Stefania Canè,4, 11 Stefano Ugel,4 Katia Donadello,5 Monica Castellucci,6 David Eyal,1 Alessandra Fiore,4 Cristina Anselmi,4 Roza Maria Barouni,4 Rosalinda Trovato,4 Simone Caligola,4 Alessia Lamolinara,7 Manuela Iezzi,7 Federica Facciotti,8 Anna Rita Mazzariol,9 Davide Gibellini,9 Pasquale De Nardo,10 Evelina Tacconelli,10 Leonardo Gottin,5 Enrico Polati,5 Benno Schwikowski,2 Ido Amit,1, 12, * and Vincenzo Bronte4, 12, *

1Department of Immunology, Weizmann Institute of Science, Rehovot, Israel.
2Systems Biology Group, Department of Computational Biology and USR 3756, Institut Pasteur and CNRS, Paris 75015, France.
3Sorbonne Universite, Complexite du vivant, Paris 75005, France.
4Immunology Section, Department of Medicine, University and Hospital Trust of Verona, Verona 37134, Italy.
5Intensive Care Unit, Department of Surgery, Dentistry, Maternity and Infant, University and Hospital Trust of Verona, Verona 37134, Italy.
6The Center for Technological Platforms, University of Verona, Verona 37134, Italy.
7CAST- Center for Advanced Studies and Technology, University of G. D’Annunzio of Chieti-Pescara, Chieti 66100, Italy.
8Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milan 20139, Italy.
9Microbiology Unit, Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona 37134, Italy.
10Division of Infectious Diseases, Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona 37134, Italy.
11These authors contributed equally.
12These authors jointly supervised this work.

*Correspondence: ido.amit@weizmann.ac.il (IA); vincenzo.bronte@univr.it (VB)
Abstract

Since the beginning of the SARS-CoV-2 pandemic, COVID-19 has appeared as a unique disease with unconventional tissue and systemic immune features. While COVID-19 severe forms share clinical and laboratory aspects with various pathologies such as hemophagocytic lymphohistiocytosis, sepsis or cytokine release syndrome, their exact nature remains unknown. This is severely impeding the ability to treat patients facing severe stages of the disease. To this aim, we performed an in-depth, single-cell RNA-seq analysis of more than 150,000 immune cells isolated from matched blood samples and broncho-alveolar lavage fluids of COVID-19 patients and healthy controls, and integrated it with clinical, immunological and functional *ex vivo* data. We unveiled an immune signature of disease severity that correlated with the accumulation of naïve lymphoid cells in the lung and an expansion and activation of myeloid cells in the periphery. Moreover, we demonstrated that myeloid-driven immune suppression is a hallmark of COVID-19 evolution and arginase 1 expression is significantly associated with monocyte immune regulatory features. Noteworthy, we found monocyte and neutrophil immune suppression loss associated with fatal clinical outcome in severe patients. Additionally, our analysis discovered that the strongest association of the patients clinical outcome and immune phenotype is the lung T cell response. We found that patients with a robust CXCR6+ effector memory T cell response have better outcomes. This result is line with the rs11385942 COVID-19 risk allele, which is in proximity to the CXCR6 gene and suggest effector memory T cell are a primary feature in COVID-19 patients. By systemically quantifying the viral landscape in the lung of severe patients, we indeed identified Herpes-Simplex-Virus 1 (HSV-1) as a potential opportunistic virus in COVID-19 patients. Lastly, we observed an unexpectedly high SARS-CoV-2 viral load in an immuno-compromised patient, allowing us to study the SARS-CoV-2 *in-vivo* life cycle. The development of myeloid dysfunctions and the impairment of lymphoid arm establish a condition of immune paralysis that supports secondary bacteria and virus infection and can progress to “immune silence” in patients facing death.
Main

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the etiological agent of the novel coronavirus disease (COVID-19) outbreak in Wuhan city (China) [1] that is currently threatening worldwide health. Italy was the first European nation to be severely affected by COVID-19: the first death was reported on 21 of February, 2020 [2]; and as of 21st July 2020 more than 35,000 and 607,781 COVID-19-related deaths were registered in Italy and worldwide, respectively [3].

Many studies highlight different, stepwise patterns of diseases progression, characterized by mild to moderate features in most of the patients, with some of them who unfortunately progress to a more severe disease stage, which can lead to acute respiratory distress syndrome (ARDS), respiratory failure and eventually death [4, 5]. The contribution of host immune system in establishing the worse prognosis has been already confirmed by several clinical observations on SARS-CoV-2 and other SARSs-dependent diseases. Indeed, lymphopenia and release of pro-inflammatory cytokines such as CXCL10 (IP10), interleukin (IL)6, IL8, IL10, tumor necrosis factor (TNF)α and C-C motif chemokine ligand (CCL)2 are enlisted as hallmark of severe SARS-CoV2 infection and correlate with adverse clinical outcome [5-7]. Accordingly, among clinical parameters associated to critical outcome, multicenter analysis on hospitalized COVID-19 patients, established among clinical parameters associated to critical outcome not only age, co-morbidities and pre-existing diseases but also immune alterations such as an increased neutrophil to lymphocyte ratio [8], hinting that pathogenic disease characteristics of the disease are worsened in sub-optimally efficient and immune dysfunctional patients.

Whether a dysregulated host immune system characterized by the coexistence between pro-inflammatory and anti-inflammatory mediators represents a key feature of COVID-19 severe progression, a clear frame of the molecular mechanisms driving this imbalance is not elucidated yet. Indeed, ARDS experienced by COVID-19 patients has a unique signature that differs from ARDS caused by any other infective or traumatic insults [9]. More specifically, the increase in cytokine release in peripheral blood, often associated with disease severity [10] and commonly defined as “cytokine storm”, is only partially involved in COVID-19 patients. Indeed, IL6 plasma levels in COVID-19 severe patients are 10 to 40 fold lower than previously reported ARDS patients, and 1,000 fold lower compared to patients facing cytokine release syndrome following treatment with chimeric antigen receptor T cells [11, 12]. Thus, it is conceivable that SARS-CoV-2 infection may hijack host immune system in order to impair anti-viral immunity and trigger a chronic inflammation characterized, but not limited, by the accumulation of peculiar inflammatory cytokines that participate in acute lung injury in severe COVID-19 patients. Recent literature explored the ability of CoVs to skew cytokine release by af-
fecting IFN-dependent, anti-viral response towards other inflammatory pathways sustaining the activation of inflammasome [13]. Noteworthy, delayed IFN-I signaling impairs antigen-specific T cell responses and promotes high cytokine secretion in lung by incoming monocytes, resulting in vascular leakage and fatal disease in SARS-CoV-infected mice [14]; furthermore, type I IFN, T cells, and signal transducer and activator of transcription 1 (STAT1) are required for virus clearance and disease resolution in a mouse model of SARS-Cov 2 infection [15] and impaired IFN I activity results in worse outcome in human COVID-19 infected patients [16].

Severe COVID-19 patients display some shared features of sepsis, including secretion of inflammatory cytokines, neutrophil hyper-activation, reduced function of natural killer (NK) and dendritic cells (DC), altered monocyte activation and lymphopenia [7, 17]. Several high dimensional phenotypic and molecular approaches were deployed in order to dissect the biology of virus-immune system interaction during COVID-19 pathogenesis [7, 16, 18-22]. These analyses were performed on peripheral blood and peripheral blood mononuclear cells (PBMCs) isolated from patients with different disease severity. This strategy streamlines sample accessibility, identification of new peripheral predictive biomarkers, and allows comparison within different studies with the caveat of not considering the local microenvironment in which the virus is acting. Taking together, all these studies highlight the presence of an IFN signature in mild to moderate patients whereas evidence a sustained emergency myelopoiesis associated to an increase in immature neutrophils and monocytes with immune suppressive features in critically ill patients. Unfortunately, none of the past studies analyzed the immune regulatory properties of myeloid cells at functional level. A specific genetic locus including immune related genes (such as C-X-C motif chemokine receptor 6 - CXCR6), was found to be associated with worse prognosis in COVID-19 patients [23]. Other studies exclusively performed on broncho-alveolar lavage fluids (BAL), elucidated the sustained interplay between macrophages releasing inflammatory cytokines and lung epithelial cells in more severe COVID-19 disease stages [21], whereas pointed out highly clonally expanded CD8+ T cells in moderate patients [22]. Here we define a complete atlas of COVID-19 patients’ immune landscape integrating both local (lung) and systemic (blood) tissues, harmonizing molecular (single cell RNA sequencing, scRNA-seq), functional and clinical data, in order to dissect the complex interplay established by SARS-CoV-2 with host immune system. We find in severe patients the establishment of innate and adaptive dysfunctions, including loss of immune-suppression by various blood myeloid cells and the replacement of lung memory CD8+ T cells by naive T cells, suggesting a state of “immune silence” that correlates with a severe clinical manifestation and fatal outcome.

Establishment of BAL and blood-derived immune cell atlas obtained from COVID-19 patients
To gain insights into the immune deviation induced by SARS-CoV-2 virus in COVID-19 patients, we performed scRNA-seq analysis on BAL and matched peripheral blood samples obtained from 21 severe, COVID-19 patients admitted to Intensive Care Units (ICU) and on peripheral blood of 6 mild SARS-CoV-2 positive patients and 5 healthy donors (Figure 1a). Immunological features were assessed on the same cohorts, integrating 4 more mild SARS-CoV-2 patients, by multiplex ELISA, multiparametric flow cytometry and functional assay (see Methods). All the patients were hospitalized at the University Hospital Integrated Trust of Verona. The clinical characteristics of enrolled patients and healthy controls are summarized in table 1.

Following strict quality controls (Figure S1a-c), cells were analyzed using the Pagoda2 pipeline [24] and clustered using Leiden community detection method [25]. The number of analyzed high quality cells was comparable between groups (Figure S1d). Fourteen significant cell clusters (more than 1% of the cells) were identified, and gathered into four major cellular subsets based on their mean expression profiles (Figure 1b and C, Figure S1e): epithelial cells, lymphocytes, monocytes/macrophages and neutrophils. The epithelial cell compartment contained only one cell cluster, which was characterized by the expression of WFDC2, SPLI and keratin genes, like KRT8 and KRT19. 5. Different lymphocyte clusters were identified, namely B cells (CD79A, CD74), NK cells (CD247, GZMB, GNLY), CD8+ T cells (GZMA, CD8A), CD4+ T cells (LTB) and gamma-delta (γδ)-T cells (GZMH). Only three clusters of the monocyte/macrophage compartment were depicted in our dataset, monocytes (LYZ, VCAN, FCN1), M1-like (CTSB, CTSL) and M2-like macrophages (MRC1, ACP5, FBPI). Lastly, significant diversity of the neutrophil compartment was observed with 5 uncovered clusters. Those clusters could be differentiated based on their expression of key markers such as CD16B (FCGR3B) mostly expressed in neutrophil clusters (1), (2) and (3), interferon response genes (IFITM3 and IFIT3) in cluster (1), S100 calcium binding proteins (S100A8, S100A9 and S100A12) in clusters (3) and (5), CXCL8 in cluster (4) and inflammatory response genes (NFKBIA, IL1RN and SOD2) in cluster (5).

We observed a strong tissue specificity in the cell clusters distribution (Figure 1c and d). For instance, neutrophil clusters (4) and (5) were BAL specific, while clusters (2) and (3) were blood specific and cluster (1) could be found in both tissues. As expected, M1-like and M2-like macrophages could only be identified in BAL samples while monocytes were blood specific. Epithelial cells were limited to the BAL samples with low representation of the total cell population (1.1% of total cells).
We next looked for possible associations between the proportion of cell clusters and the clinical status of the patient. We observed that severe patients exhibited a significantly higher proportion of neutrophils in their blood samples compared to mild patients and healthy controls (Figure 1e, left panel). In contrast, the lymphocyte proportion was decreased in severe patients compared to mild patients and healthy controls (Figure 1e, middle panel) while the monocyte/macrophage compartment was not affected by disease severity (Figure 1e, right panel). As scRNA-seq is prone to biases for population proportion estimation, we validated our findings by performing blood cell counting and systemically looking for differences between the 3 groups of patients. Consistently with our scRNA-seq analysis, we observed significant differences in the neutrophil and lymphocyte population, but also a decreased erythrocyte number in severe patients (Figure S1f and g). No other cell population was significantly affected by the disease (Figure S1f).

COVID-19 is characterized by an excessive inflammatory response that is sometimes referred to “cytokine storm” and is deemed to drive the disease pathogenesis. As inflammatory macrophages are suspected to be the main producer of inflammatory cytokines such as IL6 and IL1β, we looked for possible enrichment of M1 macrophages in the BAL from patients who died due to COVID-19. While this cluster was indeed a significant producer of inflammatory cytokines such as osteopontin (SSP1, Figure 1f left panel), it was surprisingly associated with the survival of the patients (Figure 1f right panel), suggesting it as a predictive biomarker for better prognosis. Lastly, we investigated the concentration of cytokines in the plasma and systemically looked for differential concentration between the classes of patients. Following multiple testing correction, we observed that only three cytokines were significantly affected by the patient status: Vascular Endothelium Growth Factor - alpha (VEGF-A), IL6 and Interleukin -1 receptor Antagonist (IL1RA, encoded by the IL1RN gene). All three exhibited higher concentrations in severe and mild patients’ plasma compared to healthy control, whose levels were close to detection limit (Figure S1h to i).

Altogether, we comprehensively profiled more than 150.000 immune cells from blood and BAL sampled from COVID-19 patients. Coarse-grained clustering allowed us to detect a severity associated neutrophilia and lymphopenia, but also a SPP1+ M1-like macrophage population associated with severe patients’ survival. Interestingly, only the concentration of these three cytokines was associated with disease severity, suggesting that massive cytokine release in blood is not present in COVID-19 patients. However, the restricted size of our cohort significantly limits the statistical power of our analysis and might hinder the detection of disease-associated variables.

**Resting neutrophils are replaced by multiple SARS-Cov2 associated neutrophil types**
Neutrophils are the most common white blood cells and are the first cells to migrate to the site of infection. Our dataset contains 42,238 high quality blood neutrophils, therefore allowing an in-depth analysis. We performed a refined clustering of neutrophils, which identified 10 different clusters (Figure 2a and b), including a strongly distinct and rare subtype of CD66b (CEACAM8) and antibacterial peptide (LTF, DEFA3) expressing neutrophils, probably corresponding to low density neutrophils (LDNs) and were therefore defined as LDN-like cells. Among the other clusters, we observed both a resting neutrophil cluster (ICAM1, CXCL8) and an array of activated neutrophil clusters. Among them we identified an Interferon Stimulated Genes (ISGs; RSAD2, OAS2, IFIT1), a serine protease inhibitor (PI3 and SLPI), and a chemokine (CCL4, CCL3L3) expressing clusters, suggesting a strong heterogeneity of the neutrophil polarization across patients.

To identify any trend in the neutrophil compartment composition in an un-supervised fashion, we used the correspondence analysis (CA), a method similar to Principal Component Analysis but adapted to categorical data (see Methods). CA second component was able to stratify healthy controls, mild and severe patients (Figure 2c, S2a and S2b). By computing the correlation between CA dimension 2 and the proportion of each neutrophil cluster we observed that severe, and to a lesser extent mild patients, were associated to a replacement of resting neutrophils by multiple clusters including the ISGs, CD177 and PI3 expressing neutrophils (Figure 2d, S2c). Interestingly, LDN-like cells were only detected in both mild and severe patients, albeit at a low level (less than 5% of neutrophils) except in four severe patients (Figure 2D right panel). We systemically computed Pearson’s correlation between CA dimension 2 and each measured biological and clinical variable and identified IL6 and IL1RA concentration as the most positively correlated variables, with erythrocyte and partial CO2 concentration (pCO2) negatively correlating with CA dimension 2 (Figure S2d). Altogether, our refined analysis of blood neutrophils revealed that resting neutrophils are replaced by various neutrophil clusters in both mild and severe patients.

**Functional analysis of blood myeloid compartment reveals that immuno-suppression is a major predictor of COVID-19 patients’ survival**

Blood neutrophil compartment is widely affected in COVID-19, as highlighted by our scRNA-seq analysis. However, these cells are naturally over-represented in blood and most severe COVID-19 patients suffer from neutrophilia, therefore limiting the number of cells other than neutrophils sequenced in our blood samples and our ability to study them by scRNA-seq. Recent evidence suggested the presence of monocyte alteration in SARS-CoV-2 infected patients, mostly associated with
the expansion and accumulation of immunosuppressive monocytes [26]. Due to the low number of blood monocytes sequenced in our dataset (9103 cells, corresponding to less than 300 cells per patient) and the limited ability of scRNA-seq to provide functional information, we purified circulating CD14+ monocytes from fresh blood of some study subjects and used them to perform T-cell immunosuppression assays. Both cellular and supernatant-associated immune-suppressions were assessed. In addition, immuno-suppressive activities of both normal density neutrophils (NDN) and LDN supernatants were also measured.

All samples caused some degree of T cell suppression, with monocytes and monocyte supernatants exhibiting similar activity, while LDN and NDN exhibited the highest and lowest suppression activity, respectively (Figure S2e). Suppression rate by monocytes was significantly lower in healthy controls compared to both mild and severe patients but severe patients displayed a higher variance than mild patients, with the suppression rate ranging from 10% to nearly 80% (Figure 2e left panel). Surprisingly, this heterogeneity could be partly explained by the clinical outcome of the severe patients: while severe patients who survived displayed a high T-cell immuno-suppression by monocytes, monocytes of deceased patients were unable to dampen T-cell proliferation (Figure 2e, right panel). This inverse association between immuno-suppression and patient survival was not limited to monocytes as it could be observed, albeit less significantly, with both monocyte and LDN supernatants (Figure S2f).

Myeloid cells suppress T-cell activation through multiple strategies including anti-inflammatory cytokine secretion, nutrient depletion or immune checkpoint engagement [27]. To gain further insight in COVID-19 immune landscape, we profiled monocyte expression of PD-L1, ARG1 and HLA-DR by flow cytometry (Figure S2g). We observed a clear relation between mean ARG1 expression by monocytes and monocyte immunosuppressive function (Spearman ρ = 0.95; Figure 2f), which could be fitted using a modified Hill function (see Methods), revealing an extremely strong Hill coefficient (n=40.2). HLA-DR expression was also associated to immune suppression, but in a different manner compared to ARG1 (Figure 2g). HLA-DR mean expression and immune suppressive activity clustered three groups of patients: healthy controls with a high HLA-DR expression and low immuno-suppression; mild patients and severe patients who survived with both high suppression and high HLA-DR expression; a third group of severe patients with low suppression and HLA-DR expression. As more than half (7/12) of the patients from the last groups died, we hypothesize that this cluster corresponds to patients suffering from terminal immune dysfunctions and therefore at higher risk of fatal outcome. Lastly, we observed a limited association between PD-L1 and immuno-suppression (Spearman ρ= 0.57) (Figure S2h). Furthermore, the concentration of 20 different cytokines, including
both pro-inflammatory (IL6, TNFα) and anti-inflammatory ones (IL10) was assessed in monocyte supernatant; however, no cytokine highly correlated with immune suppression (absolute Spearman ρ lower than 0.4; Figure S2i). In summary, the immuno-suppressive activity of monocytes and other blood myeloid cells is a strong predictor of severe patient survival and is primarily associated with ARG1 expression, and to a lesser extent with PD-L1 but not with any specific cytokine secretion.

COVID-19 affects blood and lung lymphocyte compartments in a severity-dependent manner

The lymphocyte compartment is extremely heterogeneous and dynamic, since it contains various cell types with properties and functions that can evolve upon inflammation and infection. By re-clustering cells identified as lymphocytes, we were able to obtain a finer picture of their heterogeneity. We identified 14 clusters, including several effector and memory T cells, naïve T cells and activated γδ-T cells (Figure 3a and S3a). Interestingly we were able to identify a cluster of B cells expressing high level of TCF4 that was specific to the blood of patient 8, a patient suffering from Chronic Lymphocytic Leukemia (CLL). We assumed that those cells were neoplastic and were thus removed from the analysis. As expected, significant differences could be observed between blood and BAL lymphocytes, with memory, effector and dividing T cells mostly found in the BAL and NK cells in the blood (Figure 3a, right panel).

To identify trends in the lymphoid compartment composition in an un-supervised manner, we used the CA, as described above. The first CA dimension of the lymphoid population perfectly separated the blood and BAL samples (Figure 3b and S3b) and seemed to capture a major trend in the BAL lymphocyte population. We therefore computed the correlation between this dimension and the various clinical features measured and noticed a striking negative correlation with the Sequential Organ Failure Assessment (SOFA) score (Figure 3c) but not with other variables (Figure S3c). In opposition, a parallel analysis performed on the blood of severe patients identified no clinical parameters associated with CA dimension 2 (Figure S3d), strengthening the importance to analyze the main district affected by the disease. Among the six major BAL lymphocytes clusters (>5% of total BAL lymphoid cells clusters) CD8+ T resident memory (CD8+ Trm, expressing ZNF683 and ITGA1) had the strongest positive correlation with CA dimension 1 (R=0.79), followed by CD4+ T resident memory (CD4+ Trm) (R=0.17). Interestingly, CD4+ Trm expressed several immune checkpoints such as CTLA4 and PD-1 (encoded by PDCD1), suggesting that they can participate to the immune suppressive landscape (Figure 3d). The naïve T-cell cluster (IL7R, LEF1) had the strongest negative correlation with the CA dimension 1, and therefore the strongest positive association with the SOFA score (R=-0.66). Three other clusters were negatively associated with CA dimension 1: effector CD8+ T cells (R=-
0.30, LAG3 and CD27), dividing T cells (R=-0.16, MCM10, E2F8) and activated γδ-T cells (R=-0.13, XCL1/2 and TRDC).

To test whether CA was also able to capture meaningful variations in blood lymphocyte population, we analyzed potential association between patient status and CA dimension 2 as it seemed to capture an important trend in blood samples (Figure 3b). Interestingly, CA dimension 2 was significantly associated with patient clinical status with a higher value being specific for severe patients (Figure 3e, left panel). This dimension was associated with two populations of NK cells and two populations of γδ-T cells (Figure 3e and 3g), i.e. resting/activated NK cells, and resting/activated γδ-T cells. Both activated NK and γδ-T cells were associated with severe COVID-19 patients while resting cells were found in healthy controls and mild patients only (Figure 3e and S3e). Activated NK and γδ-T cells were associated with an increased expression of cytotoxicity genes and activation markers such as PRF1, NKG7, KLRG1 and CD247 (Figure 3g) while resting NK cells were featured by a higher expression of the inhibitory KIR receptors KIR2DL1 and KIR3DL2 and resting γδ-T cells by a higher expression of TNFα and DUSP8 (Figure 3g). Consistently with our initial analysis of the cytokine and blood cell count data, we found that CA dimension 2 was positively associated with IL1RA plasma concentration and neutrophil counts and negatively associated with erythrocyte count and hemoglobin concentration (Figure 3f). Taken together, our analysis of the lymphoid compartment revealed that the presence of a naïve T-cell population in the BAL is associated with high clinical severity, while the blood of severe COVID-19 is characterized by the activation of NK and γδ-T cells.

**rs11385942 may act trough the regulation of the memory lymphoid cell migration**

Expression profiles generated by scRNA-seq can be overlaid with genome-wide association studies (GWASs) to pinpoint specific cell types and identify potential cellular and molecular mechanisms explaining the described genetic associations [28]. In addition to the ABO group locus, a recent study found that another genomic locus is associated with the development of severe forms of COVID-19 [23]. However, how this locus contributes to the pathology is unclear. As six different genes were covered by the peak association (SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6, and XCR1), we looked at their expression in the four different cellular compartments in both blood and BAL samples. Strikingly, only CXCR6 was expressed at a detectable level (Figure 3h left panel, S3f) and specifically in BAL lymphocytes. When analyzing the expression pattern among BAL lymphocytes, it appeared to be expressed mostly by effector and memory T cells and not by the naïve T cell cluster (Figure 3h, right panel). As the risk allele is associated with a decreased expression of CXCR6 and that CXCR6 is expressed by key protective populations, we advance the hypothesis that patients with
the risk allele have a lower amount of the protective T cell populations in the lung, therefore increasing the risk of developing severe COVID-19 forms.

**Viral landscape of COVID-19 patients affects their immune profile**

At the time of ICU admission, most COVID-19 patients exhibited a low viral load in the lung, suggesting that the virus had been mostly eliminated and that at this time the pathology was mainly driven by an inappropriate immune response rather than by viral replication. To validate this hypothesis, we applied our recently published tool, Viral-Track [18] on the BAL sequencing data in order to quantify the SARS-CoV-2 viral load and detect possible secondary viral infections. Quantification of SARS-CoV-2 viral reads revealed that in most of severe patients (17/21), no SARS-CoV-2 reads could be found (Figure 4a). Out of the 4 SARS-CoV-2 positive patients, three had low levels of viral reads (few hundreds to few thousands reads) while patient 8 displayed more than 300,000 SARS-CoV-2 reads. Consistently with our previous scRNA-seq study of SARS-CoV-2, we observed that most reads were located in the very 3’ end of the viral genome (Figure 4b), probably due to the 3’ bias of our scRNA-seq method and to nested replication process of the virus [29]. Surprisingly, we also found that two patients had a significant amount of Herpex Simplex Virus 1 (HSV-1) reads (NCBI reference number: NC_001806) with respectively 265,392 and 46,229 viral reads in patients 4 and 25, respectively (Figure 4c). Coverage analysis revealed dozens of peaks that corresponds to viral genes, suggesting an active transcription of the viral genes, and likely an active viral replication (Figure 4d). We validated this finding by performing a multiplex PCR test for four different Herpesviridae, HSV-1, HSV-2, Human Cytomegalovirus (HCMV) and Varicella-Zoster Virus (VZV) on samples from patients 8 (CLL patient, negative control), 13 (negative control) and patient 25 (Figure S4a). None of the viruses was detected in samples from patient 8 and 13 but HSV-1 was specifically detected in BAL samples of patient 25 that were collected at two different time points during ICU permanence.

We then looked for a possible explanation about the un-expectedly high SARS-CoV-2 viral load of patient 8. As mentioned before, it appeared that patient 8 suffered from CLL, a B cell malignancy characterized by the accumulation of small, mature-appearing lymphocytes in the blood, the bone marrow and in the lymphatic systems [30]. CLL patients often suffer from immune deficiency, and more precisely, from hypogammaglobulinaemia, i.e a reduction in maturated and high-affinity antibodies. We therefore quantified the amount of IgG produced toward the RBD region of the viral spike protein (Figure 4E): while healthy controls and most of the mild patients lacked any RBD-targeting IgG, all severe patients but patient 8 had high levels of these immunoglobulins. Therefore, the lack
of an efficient antibody response might have prevented the complete clearance of the virus from the lung in this patient. We tried to look which BAL cells from patient 8 were infected by SARS-CoV-2. We observed that a reduced number of cells exhibit an extremely high number of viral UMIs (several hundreds to several thousands), but interestingly those cells were mostly apoptotic cells (high levels of mitochondrial reads), neutrophils or in some rare case, cells that expressed extremely low amount of human reads (Figure 4f, S4b-c). Lastly, we looked for the possible effects of high-viral loads on the immune system state: for this purpose, we performed a correspondence analysis on all blood cell populations. We observed that the first dimension was associated with the high viral load, as patients 4 and 8 had the highest score among all patients. Furthermore, this CA dimension correlated with a subset of blood neutrophils that specifically expressed Interferon Stimulated Genes (ISGs) such as IFIT1, RSAD2 and OAS3 (Figure 4h) but also PD-L1 (CD274), suggesting that a high viral load in the lung can significantly influence the blood immune landscape in COVID-19 patients.

Discussion

The real boundaries between sepsis induced by either SARS-CoV-2 or bacteria are still ill defined and need to be further demarcated. Indeed, hyper-inflammation associated with immune suppressive state are clinical characteristics shared among these pathologic conditions, which underscore a susceptibility state defined as sepsis-induced immunoparalysis [31]. In this disease phase, innate and adaptive dysfunctions cooperate for the ineffective clearance of the pathogen, vulnerability to secondary infections and reactivation of latent viruses [32].

Immunoparalysis is likely the ground for the HSV-1 superinfection observed in some ICU patients in our study. Recently, scRNA-seq unveiled a cluster of immature CD14+ monocytes, with low HLA-DR and expressing MPO, PLAC8 and IL1R2, in the blood of COVID-19 patients and similar cells were reported in sepsis [20, 33]. These findings shed light on the practice to measure the levels of HLA-DR in monocytes as marker for sepsis, preconized but never definitively proven as mortality biomarker for severe sepsis, especially in ICU patients [34, 35]. Considering the emerging cell heterogeneity, HLA-DR levels are insufficient to reach a predictive value for patient mortality, if not associated with functional analyses. Indeed, our study shows that HLA-DR reduction in monocytes, hallmark of many COVID-19 patients, can predict patient outcome only when is associated with an impairment of their immune regulatory properties (Figure 2g). More importantly, we found that loss of function in myeloid cells mirrors an immune pathological status progressing from immune paralysis to “immune silence” associated with higher susceptibility to death event (Figure 2e, Figure S2f). CD14+HLA-DRlow monocytes and ARG1*MPO*BPI− low density pre-neutrophils are expanded in
severe COVID-19 patients, a likely consequence of the pervasive emergency myelopoiesis triggered by SARS-Cov2 infection [33]. Based on gene expression profiles, these cell subsets were deemed to have immune suppressive features, which were not functionally addressed in the study. To the best of our knowledge, we show for the first time that immune suppression is a hallmark of COVID-19 evolution (Figure 2e) and could be the basis for gradual loss of effector/memory in favor of naïve T-cells (Figure 3d). Moreover, we defined a correlation between ARG1 presence and the immunosuppressive activity of monocytes (Figure 2f), with a minor contribution of PD-L1 (Figure S2g). Of note, CD14+ARG1+ immune suppressive monocytes were originally defined in patients with pancreatic cancer who also share increased levels in some inflammatory cytokines [36, 37].

Further studies are necessary to define whether these cells are immature precursors of the granulocytic/monocytic lineage or a new subset of circulating monocytes, as well as the extent of their overlap with the immature CD14+ cells described in COVID-19 patients and sepsis [20, 33]. Noteworthy, our study identifies, new markers associated with disease severity in the peripheral blood of patients, such as the proportion of activated/resting γδ-T cells (Figure S3E), which integrate with others already described and confirmed in our study (i.e NK cells and neutrophils activation [38, 39]), which might be responsible for healthy tissue damage [40]. This information has crucial consequences from clinical and biological point of views and it will support physicians in patient stratification. More importantly, the extensive molecular analysis performed in BAL samples reveals that lymphoid compartment landscape perfectly mirrors compromised healthy conditions of the patients and identifies in the increase of naïve T cells in spite of CD4+Trm and CD8+Trm the worst clinical scenario (Figure 3C-D). Naïve T cells unbalance is not clearly observed in peripheral blood of the same patients suggesting that systemic analysis, despite many advantages, will not provide a complete landscape of the disease. According to this, we highlight the importance of CXCR6, which has has been recently associated with less severe COVID-19 stage [23]. CXCR6, highly expressed in Trm cells (Figure 3h), orchestrates Trm cells partitioning within the lung directing them to the airways [41].

Collectively, the deep molecular, phenotypic, and functional myeloid and lymphoid characterization performed on peripheral and local districts points out that critically ill patients face a profound immune dysregulated status, which can support secondary bacteria and virus infection (table 1, Figure 4c, Figure S4a). Indeed, as sign of immune paralysis, monocytes in septic patients do not respond to LPS stimulation with the upregulation of NF-κB-dependent genes, including TNFα [33]. Although this was not investigated in our study, we nonetheless show evidence that monocytes from patients who had a fatal outcome in ICU were dysfunctional and had lost the immune regulatory properties. It is thus likely that monocytes in terminally ill patients are flawed in different biological responses,
possibly including the ability to differentiate into M1 macrophages in the lung, as observed in the BAL of these subjects. Together with the reset of lymphoid arm indicated by the relative abundance of naïve T cells, this configures a state of “immune silence” and supports the deploying of drugs that can “reawaken” host immune system. “Immune silence” could be related to the extensive immaturity of this cell population, as consequence of an abnormal and skewed myelopoiesis. Alternatively, it might be a pre-existing disorder making this group of patients unable to cope with the hyper-inflammatory state. Longitudinal studies are mandatory to dissect between these different disease developments.

Our data set the ground for the administration of drugs that aims at switching off and re-start host immune system in both adaptive and innate arms, as suggested by the efficacy of recently adopted dexamethasone in severe COVID-19 patients [42]. By combining scRNA-seq with functional assays, we were able to identify features associated with disease severity and clinical outcome of the patients. However, our study is not flawless and suffers from limitations, including experimental design ones. Firstly, this is not a longitudinal study. Secondly, the absence of BAL samples from mild patients prevents us to generalize and validate the observations made on severe ones. Moreover, the absence of BAL samples from healthy control deprives us of estimating the basal state and composition of the lung immune system. While it is technically possible to collect such samples, it raises important ethical question that need to be considered while designing COVID-19 cohorts. The third limiting factor is the nature of the BAL samples: briefly, a physiologic solution is introduced into the lower respiratory tract and then collected. Such method is able to capture infiltrating immune cells in spite of non-immune cell types (i.e. epithelial cells), as shown by recent autopsy reports [43]. Lung biopsy represents an interesting alternative to BAL but is far riskier, especially for patients suffering from ARDS, such as severe COVID-19 patients, and the amount of available material is limited. One may therefore try to collect samples from deceased patients, however as most patient died after several weeks in ICU, the biological features observed during the autopsy might either be linked to the disease itself or to the extended stay in ICU. In the case of lung biopsy, highly multiplexed imaging techniques such SeqFISH [44] or CODEX [45], but also spatial transcriptomic technologies [46] could be explored. Such approaches would provide gene expression spatial pattern at a near single-cell resolution and gain essential information concerning potential cellular interactions. Lastly, we observed that several patients had a significant interferon-induced polarization of blood neutrophils that could not be solely explained by the virus presence in the lung. Such polarizations might also be due to secondary infections, but mostly bacterial rather than viral. Therefore, an extensive metagenomic analysis of the BAL could reveal disease-associated bacterial landscape and be associated to specific immune phenotype.
Overall, our study expands the biological insight on the multifaceted virus-immune system interplay and provides a solid background to design and test new candidate drugs for severe COVID-19 patients.
Methods

Study subjects and clinical considerations

All 31 patients with COVID-19 and 5 healthy donors in this study were admitted, within the period from March 12th to April 20th 2020 to the University Hospital of Verona. At sampling, the stage of disease was categorized as mild (patients not requiring non-invasive/mechanical ventilation and/or admission to ICU) or severe (patients requiring admission to ICU and/or non-invasive/mechanical ventilation). This study includes a group of 21 severe COVID-19 patients admitted to ICU, 10 mild SARS-CoV-2 patients and 5 HDs. The clinical features of the 3 groups of subjects are recapitulated in table 1. The study has been designed with the purpose of defining a complete framework of COVID-19 patients’ immune landscape. To this aim we collected clinical (i.e. co-morbidities, pulmonary performances, outcome at dismissal from Verona Hospital) and laboratory (i.e. leukocyte subsets, P-ferritin, P-D-Dimer, C-reactive protein, P-fibrinogen quantification) information and integrated them with molecular (i.e. single cell transcriptomic analysis) proteomic (cytokines quantification and serology), phenotypic (myeloid characterization in terms of expression of immune suppression hallmarks) functional (myeloid immune suppressive assay) data.

Ethics approval statement.

All the patients (and/or initially their families) provided written informed consent before sampling and for the use of their clinical and biological data. This study was approved by the local ethical committee (Prot. n° 17963, and n° 51095, P.I. Vincenzo Bronte); informed consent was obtained from all the participants to the study.

Preparation of biological samples

For each patient approximately 20 ml of BAL fluid was obtained, stored at room temperature and processed within 2 hours in a BSL-3 laboratory. An unprocessed aliquot was used for bacterial culture. The BAL fluid was filtered 2 times through a nylon gauze and a 100-µm nylon cell strainer to remove clumps and debris. The supernatant was then washed with PBS 1x and centrifuged. RBCs were lysed with 4 mL of 0.2% NaCl solution (3 minutes, RT) and the reaction was blocked by adding 9 mL of 1.2% NaCl solution. The cells were washed with PBS 1x, re-suspended in RPMI 1640 medium supplemented with 5% bovine serum albumin and counted. Cell viability was determined by Trypan blue exclusion. BAL fluids of patients with COVID-19 infection contained a heterogeneous number of cells ranging from 0.83 x10⁶ to 22 x10⁶ of cells. Cells were re-suspended at a concentration of 1 x 10⁶/ml for single cell analysis. Peripheral blood (PB) from COVID-19 patients and HDs was collected in EDTA-coated tubes. 2 ml of PB was washed once with PBS 1x and the RBCs lysis was
performed twice adding 15 mL of 0.2% NaCl solution (3 minutes, RT) and the reaction was blocked by adding 35 mL of 1.2% NaCl solution. The cells were washed with PBS 1x, re-suspended in RPMI 1640 medium supplemented with 5% bovine serum albumin, filtered through a 100-μm nylon cell strainer and counted. Cell viability was determined by Trypan blue exclusion. Cells were re-suspended at a concentration of 1 × 10⁶/ml for single cell analysis.

SARS-CoV-2 detection, bacteria identification and validation of HSV-1 infection

Nucleic acids were extracted from BAL samples by Seegene Nimbus instrument (Seegene; Seoul, South Korea) and SARS-CoV-2 related genes (E, N and RdRP) were amplified by polymerase chain reaction (PCR) with the Allplex 2019-nCoV assay kit (Seegene) according to the manufacturer’s instructions. BAL samples from COVID-19 positive patients were processed for bacterial isolation. Briefly, samples were treated with 1% v:v Dithiothreitol (DTT) for 30 minutes at 25 °C Upon treatment, (20 μl of) the samples were streaked out on petri dishes containing different types of agar (Blood Agar, Chocolate Agar, Columbia Nalidix Acid agar, Mannitol Salt Agar, McConkey and Sabouraud Agar). Cultures were incubated for 24 hours at 37 °C and bacterial growth was evaluated as colony forming unit per ml (CFU/ml). Bacterial species were identified by MALDI-tof (VITEK-MS, BioMérieux; France) and antimicrobial susceptibility was tested by VITEK-2, BioMérieux; France) and Kirby-Bauer disk diffusion assay. Nucleic acids isolated from BAL samples were processed to multiplex PCR procedure for the simultaneous detection of HSV-1, HSV-2, HCMV and VZV, using an Allplex assay (Seegene) according to the manufacturer’s instructions.

Detection of cytokines and serology

Cytokines released by patients’ monocytes and neutrophils were quantified by Human Procarta-Plex™ Panel 1 multiplex (ThermoFisher Scientific, Waltham, MA, USA). The ELISA assay to detect Immunoglobulins (Ig) used fragment of the SARS-CoV2 spike glycoprotein (S-protein) as antigens based on a recently published protocol [47]. The Spike SARS-CoV2 glycoprotein receptor binding domain (RBD) was expressed in mammalian HEK293 cells at IEO, Milan by Drs. Marina Mapelli and Sebastiano Pasqualato as glycosylated proteins by transient transfection with pGACCS vectors generated in Dr. Krammer's laboratory. The constructs were synthesized using the genomic sequence of the isolated virus, Wuhan-Hi-1 released in January 2020, and contained codons optimized for expression in mammalian cells. Secreted proteins were purified from the culture medium by affinity chromatography, quantified and stored in liquid nitrogen in aliquots. The ELISA tests to detect IgG in patients’ sera used as antigens the recombinant fragments of the RBD of the Spike SARS-CoV2 glycoprotein. After binding of the proteins to a Nunc Maxisorp ELISA plate, patients’ sera to be
analyzed were applied to the plate to allow antibody binding, and then revealed with secondary anti-
human- IgG (BD) antibody conjugated to HRP. Reaction was revealed upon addition of TMB
(Merck). Optical density at 450 nanometers was measured on a Glomax (Promega) plate reader. All samples were tested and validated with an ELISA assay, as indicated in [48].

**Flow cytometry analysis**

Immunophenotype analysis on whole peripheral blood was performed according to standard proce-
dures in order to characterize monocyte subsets (defined as classical, CD14\textsuperscript{high} CD16\textsuperscript{low/dim}; intermediate CD14\textsuperscript{int} CD16\textsuperscript{+}; non classical CD14\textsuperscript{low/dim} CD16\textsuperscript{high}). Briefly, peripheral blood was incubated with FcR Blocking reagent (Miltenyi Biotec, Paris, FR) followed by the addition of: anti-human PE-conjugated CD56 (BD Bioscience, San Jose, CA, USA; clone NCAM16.2), FITC-conjugated-CD16 (Biolegend, San Diego, CA, USA; clone 3G8), PerCP-Cy5.5-conjugated CD3 (BD Bioscience, San Jose, CA, USA; clone UCHT1), PE.Cy7-conjugated HLA-DR (eBiosciences, ThermoFisher Scientific, Waltham, MA, USA; clone L243), APC.H7-conjugated CD14 (BD Bioscience, San Jose, CA, USA; clone M\textgreek{m}P9), Brilliant Violet 421™-conjugated PD-L1 (BD Bioscience, San Jose, CA, USA; clone MIH1) and aleza fluor 647-conjugated ARG1 (developed in our laboratory [36]) antibodies, Aqua LIVE/DEAD dye (ThermoFisher Scientific, Waltham, MA, USA), RBCs were lysed using Cal-
Lyse™ Lysing Solution (ThermoFisher Scientific, Waltham, MA, USA) in accordance with the manu-
ufacturer’s instructions. Samples were acquired with FACS Canto II (BD, Franklin Lakes, NJ, USA) and analyzed with FlowJo software (Tree Star, Inc., Ashland, OR, USA).

**Myeloid cell isolation, phenotypic characterization and functional assay**

Cells were isolated from EDTA-treated tubes (BD Biosciences, NJ, USA) and freshly separated by Ficoll-Hypaque (GE Healthcare, Uppsala, Sweden) gradient centrifugation. PBMCs were counted and the monocyte fraction (CD14\textsuperscript{+}) was further isolated by CD14- microbeads (Miltenyi), following manufacturer’s instructions. From the CD14\textsuperscript{+} fraction the CD66\textsuperscript{+} low density neutrophils (LDNs) were isolated by the sequential addition of CD66b-FITC antibody (BD Biosciences, NJ, USA) and microbeads anti-FITC (Miltenyi), following manufacturer’s instructions. The normal density neutrophils (NDNs) CD66b\textsuperscript{+} were isolated from the RBC layer by dextran density gradient followed by CD66b-antibody as described for LDNs. The purity of each fraction was evaluated by flow cytometry analysis. Samples with a purity greater than 95% were assessed for their suppressive ca-
pacity. 0.5x10\textsuperscript{6} cells of each cell type were plated in 24-well plates for 12 hours in complete RPMI supplemented with 10% FBS. At the end of the incubation, viability was evaluated by flow cytometry and Trypan blue assay, and both the supernatants and the cells were collected and cultured with CellTrace (Thermo Fisher Scientific) labeled PBMCs, stimulated with coated anti-CD3 (clone OKT-
3, eBioscience, Thermo Fisher Scientific) and soluble anti-CD28 (clone28.2, eBioscience, Thermo Fisher Scientific) for 4 days in 37°C and 8% CO2 incubator (for details refer to [49]. For the cells a ratio of 3:1 (target:effector) was used. At the end of the culture, cells were stained with anti-CD3-PE/Cy7 (UCHT1, eBioscience, Thermo Fisher Scientific) and CellTrace signal of lymphocytes was analyzed with FlowJo software (Tree Star, Inc. Ashland).

**Single-cell RNA-sequencing (scRNAseq)**

BAL and peripheral blood cells were isolated and prepared as described above. For each sample, cells were resuspended in RPMI supplemented with 5% FBS to a final concentration of 1000 cells per ml and processed using the 10x Genomics Chromium Controller and the Chromium NextGEM Single Cell 3’ GEM, Library & Gel Bead kit v3.1 (Pleasanton, California, United States) following the standard manufacturer’s instructions. In brief, 10,000 live cells were loaded onto the Chromium controller to recover 4,000 single cell GEMs per inlet uniquely barcoded. After synthesis of cDNA, sequencing libraries were generated. Final 10X library quality was assessed using the Fragment Analyzer High Sensitivity NGS kit (Agilent Technologies, Santa Clara, CA, USA) and then sequenced on the Illumina NextSeq500 (Illumina, San Diego CA, USA) generating 75 base pair paired-end reads (28bp read1 and 91bp read2) at a depth of 50,000 reads/cell.

**Data analysis and statistics**

*Generation of UMI tables.* Upstream processing of reads was done using the CellRanger toolkit with default parameters. SARS-CoV-2 (NCBI reference number: NC_045512.2) and human hg38 genomes were downloaded from NCBI website. SARS-CoV-2 GTF annotation file was downloaded from the UCSC and merged with the human GTF as an additional chromosome. ORF_10 Gene 3’ boundary extended by 100 bases to catch all reads that belong to this transcript.

*High-level analysis of scRNA-seq expression data.* ScRNA-seq expression data analysis were performed using the R-based Pagoda2 pipeline (https://github.com/hms-dbmi/pagoda2/) (29227469) in addition to an in-house R script. Briefly UMI table were loaded using the read.10x.matrices() function. Low quality cells were removed using the following strategy: cell with less than 500 UMIs and more than 20% of mitochondrial genes were removed. Two rounds of analysis were performed: in the first one, all filtered cells were used to identify the major cell types, then cells from each cellular compartments are analyzed individually to provide more detailed informations. For each analysis, the number of Highly Variables Genes (HVGs) was determined using the adjustVariance() function with the gam parameter set to 10. HVGs were selected using the following strategy: for each gene, its number of zeros and its mean expression are computed. A local polynomial model is then used to

---

All rights reserved. No reuse allowed without permission.
predict the number of zeros according to the log mean expression (loess function with degree parameter set to 2). The residuals of this model (excess of zeros) are then used to ranked the genes and the genes with the highest excess of the zeros are considered as the most HVGs. PCA reduction is then computed using the calculatePcaReduction() function. The number of computed PC was changed in each analysis due to variable number of cells and cellular heterogeneity. A K-nearest neighbor graph was then build with the function makeKnnGraph() with the K value set to 30 and the distance parameter set to ‘cosine’. In order to get high-quality cell clusters, we used the Leiden community detection implemented in the R package leiden, a wrapper of the python package leidenalg. The leiden() function was applied to the KNN graphs with default parameters for each analysis. Marker genes were identified using the getDifferentialGenes() function. UMAP low dimensional embedding was computed using the uwot R package, and more precisely the umap() function with the n_neighbors parameter set to 30, and the metric parameter set to ‘cosine’. In order to group clusters of cells in the first round of analysis, mean gene expression of the most variable genes was computed using the aggregate() function. Spearman’s correlation matrix was then computed using the cor() function with the method parameter set to ‘Spearman’. Hierarchical clustering was then performed on this matrix using Ward’s method and the resulting tree used to aggregate the cell clusters.

**Correspondence Analysis of the scRNA-seq data.** In order to identify trends in cellular composition across samples we used a multivariate technique called Correspondance Analysis (CA). CA is conceptually similar to principal component analysis but applies to categorical rather than continuous data. It is traditionally applied to contingency tables: CA decomposes the chi-squared statistic associated with this table into orthogonal factors. Because CA is a descriptive technique, it has the advantage of being applicable to tables whether or not the chi-squared statistic is appropriate. We used the R implementation of CA from the package FactoMineR (CA function) with default parameters. To determine the significant components we looked at the scree plot and selected the eigenvalues/component located before the elbow. To improve the quality of our analysis, we removed cell clusters corresponding to red blood cells, platelets and cancer cells from patient 8.

To detect clinical and biological variables associated with the computed correspondance components we used the following strategy: for cytokine concentrations, we first took the square root of the initial values to get normally distributed variables and then computed Pearson’s correlation with each component independently. For the other continuous variables (clinical scores, age, BMI…), Pearson’s correlation was directly computed. To test the association between CA component or a specific cell type proportion and a categorical variable (i.e. patient clinical status and survival) we either applied a Tukey’s range test (TukeyHSD() function) if the variable is not heavy tailed. If the cell proportions are clearly heavy-tailed, we applied a Kruskall-Wallis rank test.
Viral-Track analysis. To detect and study viruses in our scRNA-seq samples we used Viral-Track, a computational tool that screen the raw sequencing files to find viral reads (32479746). As previously described, processing of the file was performed using UMI-tool (28100584). First, cell barcodes were extracted and a putative whitelist computed using the umi_tools whitelist command with the parameters ‘--stdin —bc-pattern = CCCCCCCCCCCCCCNNN -log2stderr’. Following the mapping of the reads to viral genomes and transcript assembly, the mapped reads were assigned to transcripts using the R package Rsubread through the function featureCounts() with default parameters. The command ‘umi_tools count’ is then used to compute the final expression table with the following parameters:–per-gene–gene-tag = XT–assigned-status-tag = XS–per-cell.

In the case of patient 8, cells were not filtered on total host UMIs and proportion of MT UMIs but only on total combined host and viral UMIs to avoid removing apoptotic cells containing a high viral load but expressing few host genes.

Analysis of the serum cytokine, blood cell count and clinical data. Using a Cullen and Frey graph (descdist() function from the fitdistrplus package) we observed that both serum cytokine and blood cell count variables could be transformed into gaussian-like variables by applying a simple square root function and then used for further analysis. Association between blood cell counts or serum cytokine concentration and patient clinical status was assessed by fitting an ANOVA model to the transformed variables (aov() and anova() functions). Correction for multiple testing was done using the p.adjust function with parameter method set to ‘BH’. When correlations with a CA dimension were computed, the cor() function with default parameters was used. To validate the association between the SOFA score and the lymphoid CA dimension 1 we fitted a basic linear model with the lm() function and assessed the significance of the association by performing a Fisher test with the anova() function.

Analysis of the immuno-suppression, flow cytometry and cytokine secretion data. As both both flow cytometry and cytokine secretion data were extremely heavy-tailed we applied a logarithmic transformation with a pseudo count of 1 (log10(1+x)). Spearman correlations between protein MFI or cytokine concentration and immune-suppression was computed using the cor() function.

In order to model the relationship between ARG1 MFI and immune-suppression we used a function similar to the Hill function used in biochemistry:

\[ S = E_{\min} + \frac{E_{\max}}{1 + \left(\frac{K}{x}\right)^n} \]
Here $S$ corresponds to the immune-suppression, $x$ to the transformed ARG1 MFI, $E_{\text{min}}$ to the basal immune-suppression, $E_{\text{max}}$ to the maximal suppression that can be induced by ARG1, $K$ to the transformed ARG1 MFI required to get half of the maximal suppression $(E_{\text{min}}+E_{\text{max}})$ and $n$ the cooperation coefficient. This function was fitted using the nls() function with default parameters.

Quantitative variables indicated in table 1 were expressed as the median and interquartile range (IQR), qualitative ones as percentages. All statistical analysis were performed using R 3.6.1 on an Ubuntu 18.04 workstation.

**Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

**Data availability**

The authors declare that all the other data supporting the findings of this study are available within the article and its supplementary information files and from the corresponding authors upon reasonable request. All scripts used for the data analysis are available at [https://github.com/PierreBSC/Ve-rona_COVID19](https://github.com/PierreBSC/Ve-rona_COVID19).
References

Acknowledgement

This work was supported by Fondazione Cariverona (ENACT Project) and Fondazione TIM. We thank all patients who participated in this study and their families. We thank all the members of Immunology Section of Verona University Hospital who actively worked during the pandemic: Francesca Hofer, Varvara Petrova, Chiara Musiu, Cristina Frusteri, Veronica Batani, Alessandro Ghirardelli, Morena Martini, Fiorenza Paiola, Elena Lucchini, Claudia Pizzoli, Elena Chiesa, Oretta Gabrielli, Nadia Brutti, Monica Brentegani, Elisabetta Gallo, Giulio Fracasso, Tiziana Cestari, Ornella Poffe, Antonio Vella, Giovanna Zanoni, Silvia Sartoris, Riccardo Ortolani, Selena Gomirato, Daniel Lovato, Antonella Valentini and Claudia Italia. We thank the members of Division of Infectious Diseases who actively assist patients: Monica Brentegani, Alessandro Visentin, Amina Zaffagnini and Leonardo Motta. We thank the members of Intensive Care Unit who were actively involved in patients care and assistance: Sara Boschetti, Francesca Del Favero, Riccardo Boetti. We thanks Marina Mapelli and Sebastiano Pasqualato of European Institute of Oncology for technical support. We thank 10X Genomics Inc. and Carlo Erba to support the research. We deeply acknowledge the contribution of ‘Centro Piattaforme Tecnologiche’ of the University of Verona for samples sequencing. We dedicate this work to the memory of health care workers who have given their lives in the care of patients with COVID-19.

Author information

These authors contributed equally: Pierre Bost, Francesco De Sanctis and Stefania Canè

These author share senior contribution: Ido Amit and Vincenzo Bronte

Affiliations

Department of Immunology, Weizmann Institute of Science, Rehovot, Israel
Pierre Bost, David Eyal and Ido Amit

Systems Biology Group, Department of Computational Biology and USR 3756, Institut Pasteur and CNRS, Paris 75015, France
Pierre Bost and Benno Schwikowski

Sorbonne Universite, Complexite du vivant, Paris 75005, France
Pierre Bost
Immunology Section, Department of Medicine, University and Hospital Trust of Verona, Verona 37134, Italy
Francesco De Sanctis, Stefania Canè, Stefano Ugel, Alessandra Fiore, Cristina Anselmi, Roza Maria Barouni, Rosalinda Trovato, Simone Caligola and Vincenzo Bronte

Intensive Care Unit, Department of Surgery, Dentistry, Maternity and Infant, University and Hospital Trust of Verona, Verona 37134, Italy
Katia Donadello, Leonardo Gottin and Enrico Polati

The Center for Technological Platforms, University of Verona, Verona 37134, Italy.
Monica Castellucci

CAST- Center for Advanced Studies and Technology, University of G. D’Annunzio of Chieti-Pescara, Chieti 66100, Italy.
Alessia Lamolinara and Manuela Iezzi

Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, Milan 20139, Italy.
Federica Facciotti

Microbiology Unit, Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona 37134, Italy.
Anna Rita Mazzariol and Davide Gibellini

Division of Infectious Diseases, Department of Diagnostics and Public Health, University and Hospital Trust of Verona, Verona 37134, Italy.
Pasquale De Nardo and Evelina Tacconelli

Author Contribution
P.B. designed and developed Viral-Track, performed computational and statistical analysis and wrote the manuscript. F.D.S. performed sample processing, bio banking and single cell experiments, coordinated the study, data analysis and wrote the manuscript. S.Can. performed sample processing, bio banking, functional studies and data analysis. S.U. coordinated the study and wrote the manuscript; K.D. was responsible for biological specimens and patients’ data collection and revised the manuscript; M.C. performed single cell experiments; D.E. contributed to data processing and analysis; A.F., C.A., R.M.B. and R.T. performed sample processing, bio banking, functional studies and data analysis; S.Cal contributed to computational analysis; A.L., M.I. and F.F. performed cytokine and serological analysis and revised the manuscript; D.G., A.R.M., were responsible for microbiological analysis and revised the manuscript; P.D.N., E.T., L.G. and E.P. were responsible for biological specimens and patients’ data collection and revised the manuscript; B.S. contributed to development of
computational methods and bioinformatic analysis and revised the manuscript; I.A. and V.B. directed the project, conceived, designed the experiments, and wrote the manuscript.

Corresponding Authors
Ido Amit: ido.amit@weizmann.ac.il
Vincenzo Bronte: vincenzo.bronte@univr.it

Ethics declarations
Competing interests
The authors declare no competing interests.

Figure Legends

Figure 1. Establishment of BAL and blood-derived immune cell atlas obtained from COVID-19 patients. (a) Description of the cohort. (b) Two-dimensional UMAP embedding of the scRNA-seq data. Dots (cells) are colored according to their respective metacluster (Epithelial cells, lymphocytes, neutrophils and monocytes/macrophages). (c) Expression heatmap of the 14 significant clusters detected in our scRNA-seq dataset. The top 5 best markers were selected for each cluster. (d) Two-dimensional density plot of the UMAP embedding of the BAL (upper panel) or the blood (lower panel) cells. (e) Proportion of neutrophils, lymphocytes and monocytes/macrophages in blood samples across patient status. For each of the cell types, an ANOVA test was performed and corrected for multiple-testing by Bonferroni correction. Median and 5%-95% theoretical quantiles are shown. (f) Expression of the SPP1 (osteopontin) gene across cell types (left panel) and distribution of M1 macrophages among total BAL cells based on severe patient clinical outcome (right panel). A two-sided Welch’s t-test was performed to compare proportion between the two groups of patients (t = 2.2 with a degree of freedom equal to 16.5). Median and 5%-95% theoretical quantiles are shown.

Figure 2. ScRNA-seq and functional analysis of blood myeloid cells reveal unique features associated to patient status and outcome. (a) Expression heatmap of the 10 clusters identified among the blood neutrophils. (b) Two-dimensional UMAP embedding of the blood neutrophil. Cells are colored according to their cluster. (c) Scatter plot of the Correspondence Analysis of the blood neutrophil populations. (d) Proportion of resting neutrophils (left panel), ISGs neutrophils (middle panel) and LDN-like (right panel) among blood neutrophils according to patient clinical status. A Tukey’s
range test was used in the left panel while Kruskall-Wallis rank test was used in both middle and right panel. Median and 5%-95% theoretical quantiles are shown. (e) T-cell suppression ability of CD14+ monocytes according to clinical status (right) or to ICU outcome (severe patients only, right panel). A Tukey’s range test was used in the left panel while a two-sided Welch’s t-test was performed in the right panel (t = 4.9 with a degree of freedom equal to 11.8). Median and 5%-95% theoretical quantiles are shown. (f) Association between CD14+ monocytes ARG1 Mean Fluorescence Intensity (MFI) and immune suppression. The dashed line corresponds to the fitted Hill-like function. (g) Association between CD14+ monocytes HLA-DR Mean Fluorescence Intensity (MFI) and immune suppression.

Figure 3. ScRNA-seq analysis of blood and BAL lymphocytes identify cellular (a) Two-dimensional UMAP embedding of lymphocytes colored according to their cluster (left panel) or their tissue (right panel). (b) Correspondence Analysis of blood and BAL lymphocytes. (c) Association between SOFA score and CA dimension 1 score of BAL samples. (d) Expression heatmap of the BAL lymphocytes belonging to the cell clusters that are associated with CA dimension 1. The 10 best markers are shown for each cluster. (e) Distribution of CA dimension 2 score of blood samples according to clinical status (left panel), and proportion of resting (middle panel) and activated NK (right panel) according to clinical status. Tukey's range test was used to compute the shown p-values. Median and 5%-95% theoretical quantiles are shown. (f) Ranked Pearson correlation between biological features and CA dimension 2. (g) Expression heatmap of the blood lymphocytes belonging to the cell clusters that are associated with CA dimension 2. The 20 best markers are shown for each cluster. (h) Mean expression of CXCR6 across different tissues and cell types (left panel) and among the different BAL lymphocytes clusters associated with CA dimension 1.

Figure 4. Analysis of severe COVID-19 patients viral landscape. (a) Number of SARS-CoV-2 UMIs in each BAL sample. (b) Coverage plot of SARS-CoV-2 genome. (c) Viral-Track analysis of patients 4 and 25. (d) Coverage plot of HSV-1 genome. (e) Quantification of IgG targeting the RBD domain of the SARS-CoV-2’s spike protein. OD: optical density. (f) Mean number of SARS-CoV-2 UMIs across patient 8 cell clusters. (g) CA analysis of total blood cell population. (h) Expression heatmap of cells belonging to resting or ISGs neutrophils.
Table 1: Clinical characteristic of enrolled patients and healthy controls.

<table>
<thead>
<tr>
<th>CHARACTERISTICS</th>
<th>Healthy Controls N=5</th>
<th>Mild Patients N=10</th>
<th>Severe Patients (ICU) N=21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anagraphics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, yr: Median (IQR)</td>
<td>66 (64-73)</td>
<td>69 (56-80)</td>
<td>67 (58-70)</td>
</tr>
<tr>
<td>Male, no. (%)</td>
<td>4 (80)</td>
<td>6 (60)</td>
<td>17 (81)</td>
</tr>
<tr>
<td><strong>Coexisting disorder, no. (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any</td>
<td>2 (40)</td>
<td>10 (100)</td>
<td>17 (81)</td>
</tr>
<tr>
<td>Obesity</td>
<td>0 (0)</td>
<td>2 (22)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>2 (40)</td>
<td>10 (100)</td>
<td>11 (52)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0 (0)</td>
<td>3 (30)</td>
<td>7 (33)</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>0 (0)</td>
<td>2 (20)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>0 (0)</td>
<td>5 (50)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>0 (0)</td>
<td>2 (20)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Active malignancies</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>2 (10)</td>
</tr>
<tr>
<td><strong>Median (IQR) interval from symponts onset (S.O.) and Outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days from S.O. to Hospitalization</td>
<td>-</td>
<td>5 (2-6)</td>
<td>6 (4-7)</td>
</tr>
<tr>
<td>Days from S.O. to ICU admission</td>
<td>-</td>
<td>-</td>
<td>7 (6-10)</td>
</tr>
<tr>
<td>Days from S.O. to dismissal from ICU</td>
<td>-</td>
<td>-</td>
<td>38 (23-45)</td>
</tr>
<tr>
<td>Days from S.O. to dismissal from Verona Hospital (alive or dead)</td>
<td>-</td>
<td>21 (13-24)</td>
<td>43 (32-66)</td>
</tr>
<tr>
<td>Outcome, no. deaths (%)</td>
<td>-</td>
<td>1 (10)</td>
<td>8 (38)</td>
</tr>
<tr>
<td><strong>Clinical features at sampling</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APACHE score, Median (IQR)</td>
<td>-</td>
<td>-</td>
<td>23.5 (15-28.5)</td>
</tr>
<tr>
<td>SOFA score, Median (IQR)</td>
<td>-</td>
<td>2 (0.8-3.3)</td>
<td>6 (4-7)</td>
</tr>
<tr>
<td>pCO2 [35-40 mmHg], Median (IQR)</td>
<td>-</td>
<td>34 (31-41)</td>
<td>48 (41-52)</td>
</tr>
<tr>
<td>pO2 [80-100 mmHg], Median (IQR)</td>
<td>-</td>
<td>59 (55-64)</td>
<td>77 (69-97)</td>
</tr>
<tr>
<td>FiO2 %, Median (IQR)</td>
<td>-</td>
<td>26 (21-31)</td>
<td>60 (50-90)</td>
</tr>
<tr>
<td>P/F ratio mmHg, Median (IQR)</td>
<td>-</td>
<td>213 (194-267)</td>
<td>146 (71-177)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laboratory findings at sampling, Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin (135-160 g/L)</td>
</tr>
<tr>
<td>Leukocytes (4.3-10 10⁹/L)</td>
</tr>
<tr>
<td>Neutrophils (1.8-8 10⁹/L)</td>
</tr>
<tr>
<td>Lymphocytes (1.2-4 10⁹/L)</td>
</tr>
<tr>
<td>Monocytes (0.2-1 10⁹/L)</td>
</tr>
<tr>
<td>Eosinophils (&lt;0.45 10⁹/L)</td>
</tr>
<tr>
<td>Basophils (&lt;0.2 10⁹/L)</td>
</tr>
<tr>
<td>Platelets (150-400 10⁹/L)</td>
</tr>
<tr>
<td>C-reactive protein [inf. a 5 mg/L]</td>
</tr>
<tr>
<td>P-Ferritin [30 - 300 μg/L]</td>
</tr>
<tr>
<td>P-D-Dimer [inf. a 500μg/L]</td>
</tr>
<tr>
<td>Trombin clotting time (pt) 0.82-1.17 INR</td>
</tr>
<tr>
<td>P-fibrinogen [2.00 - 4.00 g/L]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microbiology analysis on BAL at sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung infection (CFU&gt;10⁶), no. (%)</td>
</tr>
<tr>
<td>Pseudomonas lung infection (CFU&gt;10⁶), no. (%)</td>
</tr>
</tbody>
</table>

IQR denotes interquartile range, PCO2 carbon dioxide partial pressure, PO2 oxygen partial pressure, FiO2 fraction of inspired oxygen, P/F PO2/FiO2, CFU denotes colony forming units.
Figure 2, Bost et al.
Figure 3, Bost et al.

(a) UMAP-2 and UMAP-1 visualizations of different cell types.

(b) Scatter plot showing differentiation between blood and BAL cells.

(c) Correlation graph with Pearson R = -0.66, p = 0.001.

(d) Heatmap with color intensity indicating correlation with CA.

(e) Statistical analysis comparing healthy control, mid, and severe conditions.

(f) Correlation with CA dimension 2.

(g) Heatmap showing correlation with CA-2 for NK cells and gd T-cells.

(h) Bar graph comparing CXCR6 expression in different conditions.
Supplementary Figure 1. Quality control of the scRNA-seq dataset and blood cell count analysis. (a) Distribution of total cellular UMIs (log10 scale). The vertical red dashed line corresponds to the 500 UMI threshold used to filter out low quality cells. (b) Distribution of the proportion of mitochondrial UMIs among total cellular UMIs. The vertical red dashed line corresponds to the 20% threshold used to filter out low quality cells. (c) Distribution of total gene UMIs (log10 scale). The vertical red dashed line corresponds to the 50 UMI threshold used to filter out lowly expressed genes. (d) Number of high quality cells that passed QC in the different samples types. Median and 5%-95% theoretical quantiles are shown. (e) Spearman’s correlation heatmap of the mean expression of the most variable genes in each single-cell cluster. (f) Effects of patient clinical status on blood cell counts. The p-values are computed by fitting and testing an ANOVA model (Methods). The vertical red dashed line corresponds to the 0.01 significance threshold used to identify cell types which blood count is affected by the clinical status. (g) Number of erythrocytes (left panel), neutrophils (middle panel) and lymphocytes (right panel) in blood based on patients clinical status. Displayed p-values were computed as described above. Median and 5%-95% theoretical quantiles are shown. (h) Effects of patient clinical status on serum cytokine concentration. The p-values are computed by fitting and testing an ANOVA model (Methods). The vertical red dashed line corresponds to the 0.01 significance threshold used to identify cytokines which serum concentration is affected by the clinical status. (i) Serum concentration of VEGF-A (left panel), IL6 (middle panel) and IL1RA (right panel) based on patients clinical status. Displayed p-values were computed as described above. Median and 5%-95% theoretical quantiles are shown.

Supplementary Figure 2. Dissection of the blood neutrophil compartment. (a) Percentage of variance explained by the 10 first components of the blood neutrophil CA. (b) Correspondence Analysis of blood neutrophils (first and second components). (c) Proportion of PI3+ (left panel) and CD177+ neutrophils (right panel) among total blood neutrophils based on patient clinical status. A Kruskall-Wallis rank test was used in both panels. Median and 5%-95% theoretical quantiles are shown. (d) Pearson correlation of biological variables with blood neutrophil CA dimension 2. (e) T-cell suppression by various blood myeloid cell or their supernatant. The p-value is computed by fitting and testing an ANOVA model (F = 16.0, degree of freedom equal to 3). Median and 5%-95% theoretical quantiles are shown. (f) T-cell suppression by monocyte supernatant (left panel) and LDN supernatant (right panel) of severe patients based on clinical outcome. P-values are computed
by applying a two-sided Welch’s t-test \((t=3.0\) and \(t=2.2\) with a degree of freedom of 12.9 and 11.6 respectively). Median and 5%-95% theoretical quantiles are shown. (g) Flow cytometry gating used to measure HLA-DR, ARG1 and PD-L1 (CD274) expression in monocytes. (h) Association between monocytes PD-L1 MFI and suppression. (i) Pearson correlation of monocyte secreted cytokine concentrations with monocyte immune-suppression.

**Supplementary Figure 3. Association between blood and BAL lymphocyte populations and patient clinical status.** (a) Expression heatmap of lymphocytes. The 16 displayed genes correspond to genes known to play an important role in lymphocyte biology. (b) Percentage of variance explained by the 10 first components of the lymphocyte CA. (c) Correlation between CA dimension 1 score of BAL samples and 5 clinical variables. (d) Correlation between CA dimension 2 score of blood samples and 5 clinical variables. (e) Proportion of resting (left) and activated (right) \(\gamma\delta\)-T cells among blood lymphocytes. Tukey’s range test was used to compute p-values. Median and 5%-95% theoretical quantiles are shown. (f) Expression of five severe COVID-19 genetically associated genes among tissues and cell types.

**Supplementary Figure 4. Study of the pulmonary viral landscape in severe COVID-19 patients.** (a) Results of the qPCR test for herpes infections. (b) Relationship between total host and SARS-CoV-2 UMIs in patient 8 BAL cells. (c) Percentage of variance explained by the 10 first components of the total blood CA.
The tracking of virally infected cells

5 Contributions and limitations of our dual scRNA-seq approach.

By combining conventional scRNA-seq experiments with a new computational tool, Viral-Track, we were able to study at a single-cell resolution the in-vivo behavior of several viruses, including Influenza A, LCMV, Hepatitis B, HSV-1, hMPV and last but not least SARS-CoV-2, the pathogen agent responsible of COVID-19. Indeed, we have extensively used this approach in the context of the COVID-19 pandemic to understand how SARS-CoV-2 infection alters the immune system and how those alterations were associated with the clinical phenotype. In addition, our studies have also pointed out the possible contribution of secondary viral infections, such as hMPV or HSV-1 infection and how they can affect both the local and systemic immune response.

Despite the critical information provided by both studies, our accumulated experience has led us to conclude that our approach suffers from both technical and critical flaws. As mentioned in the first manuscript (Bost et al., 2020a), most scRNA-seq protocols rely on polyA-based RT primers and therefore only capture polyAdenylated transcripts: while this is sufficient to capture most of the host mRNA, several families of viruses are lacking polyA tails at the end of their mRNA such as the Arenaviridae family that includes LCMV, Lassa hemorrhagic fever virus and other pathogenic viruses (Burrell et al., 2017). Moreover, limited transcript coverage and sensitivity of 3’ scRNA-seq methods also lessen the quality of the viral transcriptome analysis. Those problems can be addressed by the use of new scRNA-seq protocols, such as the RamDA-seq (Random Displacement Amplification sequencing) protocol, that can capture non-polyA transcripts and display a homogenous coverage along the transcripts (Hayashi et al., 2018). An alternative solution would be to use multiplexed transcriptomic imaging such as seqFISH (Shah et al., 2016), where sets of fluorescently labeled complementary probes are used to detect individual mRNA molecules. Those methods display extremely high sensitivity (90% of mRNA molecules captured compared to 10% by scRNA-seq) and can be designed to distinguish spliced and unspliced mRNA, but simultaneously measure only a limited set of transcripts. Moreover, seqFISH can be combined with standard immunofluorescence staining to image viral proteins. Therefore, spatial transcriptomic analysis of virally-infected cells could be an interesting alternative to dual scRNA-seq and could provide key information about the viral life cycles.

As pointed out by the two manuscripts, infected cells usually represent a small fraction of the total pool of sequenced cells, thus limiting the number of sequenced virally-infected cells. Moreover, many of them display an apoptotic phenotype, making the analysis of the host transcriptome challenging. A brute-force solution would be to simply increase the number of sequenced cells but at the price of increasing the cost of already expensive experiments. Instead, one could enrich for cell populations that are more likely to be infected: for instance in
the case of COVID-19, one might enrich for ACE2 expressing cells before the sequencing. As ACE2 is the main cellular receptor of SARS-CoV-2, SARS-CoV-2 infected cells are thus more likely to be sequenced. More appropriate types of samples can also be used: by directly sequencing lung biopsies instead of BAL, the proportion of ACE2+ cells will drastically increase and thus the probability of sequencing SARS-CoV-2 infected cells. However this is not always possible due to the clinical risks induced by such biopsies on severe patients (i.e pneumothorax for lung biopsy). While the underlying ethic of such approaches has to be questioned, the low amount of knowledge gained by blood and BAL studies should remind us how critical this point is.

Lastly we received significant feedback concerning our computational pipeline through the associated GitHub page and by email. In addition to the unavoidable bugs afflicting early versions of all softwares, we observed that several points should be improved, namely the user interface and the analysis results visualization, but more importantly, the transcript assembly and quantification. In Viral-Track, transcript assembly is performed using StringTie, a tool developed for eukaryote organisms. Due to the extreme complexity of viral genomes with frequent overlapping transcripts, such tool might be not adapted and will likely aggregate different transcripts. We hypothesized that instead of classical transcript assembly, peak detection algorithms should be used to detect viral transcriptomic features, similarly to what is done in ATAC and ChiP-seq analysis. Such approach has recently been described for scRNA-seq data and have revealed differential transcript usage among immune cells (Patrick et al., 2020). Therefore significant improvements of Viral-Track core methods and implementations could increase the value of Viral-Track analysis and provide critical new biological knowledge on emerging viral infections.
Chapter III

Quantitative study of IFNg effects in tumors

1 The necessity of a quantitative characterization of cellular cross-talks

As it has been shown in the previous chapters, scRNA-seq is an adequate tool to identify new cross-talks between two cell types. While offering an exhaustive list of the genes induced and repressed in the context of a given cross talk, scRNA-seq fails at providing a quantitative description of it. Different features of an immune cross-talk can be quantitatively assessed, especially for cytokine mediated communications:

1. The relationship between the cytokine concentration in the cellular environment and the response of the targeted cells. Usually, the corresponding curve is a sigmoid curve fully described by a maximal response and an effective concentration for half-maximum response (EC50) (Altan-Bonnet and Mukherjee, 2019). One might also study the nature of the activation: the dose-response curve is usually performed at the bulk level and we have seen in the subsection 1.1.3 that the same curve can be explained by completely different behaviors at the single-cell level. Therefore, knowing the nature of the activation (digital or analog) is also required to fully characterize the cellular cross-talk of interest.

2. A critical point is the typical distance at which each cytokine acts as it ranges from a few micrometers (autocrine communication) to a several centimeters (endocrine communications) (Alberts, 2015; Altan-Bonnet and Mukherjee, 2019). Secreted cytokines reach their target cell by the coupling of diffusion (random and spontaneous motion of molecules) and advection (active transport driven by the flow of a fluid). Several factors can affect and limit this process, including consumption and degradation by cells or reduced diffusion due to the transient binding of cytokines to the Extra Cellular Matrix (ECM) (Altan-Bonnet and Mukherjee, 2019). Measuring and modeling each of these
aspects allows to compute at which distance cytokines act and how it might be affected by local changes of cellular or ECM composition.

3. Lastly, the duration of a cellular communication can significantly affect the intensity and nature of the targeted cell response, both for contact dependent and paracrine communications (Riquelme et al., 2009; Braun et al., 2013). Measuring such time-scale is challenging yet essential as different genes might be induced by the same cytokine with different stimulation time (Thibaut et al., 2020).

While quantitative studies of immune communications are rare, they proved to be essential to understand specific phenomenon such as how immune spatial niches, i.e localized domains of high cytokine concentrations, are generated and their size tuned according to the immune context (Oyler-Yaniv et al., 2017).

Several tools are available to perform such quantitative studies: if the nature of the activation and cytokine dose-response curve can be assessed by basic in-vitro experiments, imaging is the tool of choice to study the spatial and temporal aspects of immune cross-talks. Both in vitro and in vivo imaging coupled with fluorescent reporters can reveal the activation state of the cells as well as the location of cytokine producing cells. Combining these two information allows to estimate the typical distance at which the cytokine is acting. The time-scale is more challenging to assess and will require live imaging of the cells: this can directly be done in-vitro but will require a specific imaging technique called intravital imaging in the context of in-vivo experiments. Intravital imaging allows to image living cells while they are in their original biological environment. It usually relies on multiphoton microscopy which confers a greater depth of penetration compared to single-photon microscopy (up to 500\(\mu\)m compared to 50-100 \(\mu\)m for single-photon microscopy). For an extensive introduction to multiphoton microscopy please see (Benninger and Piston, 2013).

2 Contribution of IFNg in the tumor microenvironment

IFNg is a primordial cytokine, as suggested by the extreme scarcity of individuals with a homozygous loss of function mutation in the encoding gene (Kerner et al., 2020). Mutation in this gene, or in other IFNg related genes such as the IFNg receptor 1 and 2 genes, STAT1 or IL12p40 (a sub-unit of the IL12 cytokine) results in a strong susceptibility of the patient toward mycobacterial infections, including the tuberculosis causative agent *Mycobacterium tuberculosis* but also to other mycobacteria species known as 'atypical mycobacteria' that usually do not cause infections in humans (Geha and Notarangelo, 2016).

In addition to its key antibacterial role, IFNg plays a critical role during the anti-tumor immune response (Ivashkiv, 2018) as shown by the spontaneous development of tumors in IFNg deficient mice (Kaplan et al., 1998). Cancer cells express highly mutated proteins that are considered as 'non-self' by the immune system and are termed 'neo-antigen'. CD8+ T cells
Quantitative study of IFNg effects in tumors

Figure III.1: Effects of IFNg in tumors. IFNg is a critical immune regulator of the tumor and has both anti-tumor effects (red box) by promoting antigen presentation and cancer cell apoptosis, and pro-tumor (blue box) by inducing the expression of immune checkpoints such as PDL1. Adapted from (Ivashkiv, 2018).

IFNg can recognize such neo-antigens through their TCR, triggering the killing of the target cells but also the release of IFNg. IFNg can also be produced by NK and ILC1s upon recognition of specific stress-induced ligands, such as MICA and MICB that are often found in tumors (Ivashkiv, 2018).

Several mechanisms have been proposed to explain the antitumor effects of IFNg: in addition to its strong cytotoxic and cytostatic effects, IFNg induces the expression of MHC genes and of several antigen-presentation related genes, therefore increasing the presentation of neo-antigens by the tumor cells and the efficacy of cytotoxic T-cells. Moreover, IFNg further strengthens the local Th1 polarization and stimulates the expression of inflammatory chemokines, triggering a massive recruitment of additional cytotoxic lymphocytes (Garcia-Diaz et al., 2017).

However, several studies have pointed out possible detrimental effects of IFNg on the anti-tumor immune response. Notably IFNg is a potent inducer of the immune checkpoints PD-L1 and PD-L2 by cancer cells (Garcia-Diaz et al., 2017), significantly dampening the T-cell response. Other negative regulators of the immune response are also induced by IFNg, including the suppressor of cytokine signalling 2 (SOCS2) signaling gene and the indoleamine 2,3-dioxygenase (IDO) enzyme (Ivashkiv, 2018), therefore creating a locally immuno-suppressive environment.
IFNg has therefore a dual effect on the anti-tumor immune response and its exact contribution \textit{in-vivo} is unclear. Moreover, a quantitative characterization of IFNg production and effects on the tumor microenvironment is currently lacking. As IFNg seems to significantly affects the efficacy of immune checkpoint blockades (Ivashkiv, 2018), such studies are therefore crucial and required to improve current immunotherapies.

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
\textbf{Personal contribution} \\
\hline
- Unlike in the other papers shown in this thesis, I am not a first co-author of this paper. I have only joined this project lately and contributed by analyzing both the imaging and genomic data displayed in the paper. Especially I have developed a set of scripts to quantify the STAT1 cellular activation from intravital imaging data and performed the secondary analysis of the human scRNA-seq datasets to generalize the observations made in mice to clinical human samples. \\
\hline
\end{tabular}
\end{table}
3 Bystander IFNγ activity promotes widespread and sustained cytokine signaling altering the tumor microenvironment

In this paper we performed an in-depth characterization of the IFNγ effects on the tumor microenvironment by combining intravital imaging, flow cytometry, bulk RNA-sequencing and re-analysis of previously published scRNA-seq datasets. Our study reveals that in opposition to the established model where lymphocytes are supposed to release IFNγ through an ‘immunological synapse’ targeting a unique target cell, IFNγ acts on the whole tumor and not on specific target cells. In addition, we observed that a sustained IFNγ stimulation is required to induce the expression of key genes expressed by cancer cells. Altogether our analysis suggests that a sustained and global IFNγ stimulation can be observed both in multiple mice tumor models and clinical samples from human tumors.
Bystander IFN-γ activity promotes widespread and sustained cytokine signaling altering the tumor microenvironment

Ronan Thibaut1,2, Pierre Bost3,4,5,6, Idan Milo1,6, Marine Cazaux1,2, Fabrice Lemaître1, Zacarias Garcia1, Ido Amit3, Béatrice Breart1, Clémence Cornuot1, Benno Schwikowski4 and Philippe Bousso1,2✉

The cytokine interferon (IFN)-γ produced by tumor-reactive T cells is a key effector molecule with pleiotropic effects during anti-tumor immune responses. Although IFN-γ production is targeted at the immunologic synapse, its spatiotemporal activity within the tumor remains elusive. In the present study, we report that, although IFN-γ secretion requires local antigen recognition, IFN-γ diffuses extensively to alter the tumor microenvironment in distant areas. Using intravital imaging and a reporter for STAT1 translocation, we provide evidence that T cells mediate sustained IFN-γ signaling in remote tumor cells. Furthermore, tumor phenotypic alterations required several hours of exposure to IFN-γ, a feature that disfavored local IFN-γ activity over diffusion and bystander activity. Finally, single-cell RNA-sequencing data from melanoma patients also suggested bystander IFN-γ activity in human tumors. Thus, tumor-reactive T cells act collectively to create large cytokine fields that profoundly modify the tumor microenvironment.

IFN-γ is a key soluble effector molecule during anti-tumor immune responses. Mice deficient for IFN-γ production are prone to develop spontaneous or carcinogen-induced tumors, highlighting the crucial importance of IFN-γ during immunosurveillance2. IFN-γ also plays an important role during tumor immunotherapy because the benefit of anti-programmed death 1 (anti-PD-1) therapy in mice has been shown to be dependent on IFN-γ production by CD8+ T cells and the associated interleukin (IL)-12 production by dendritic cells. Moreover, in melanoma patients, acquired resistance to anti-CTLA-4 therapy has been associated with mutations in the IFN-γ signaling pathway. Several mechanisms have been described to explain the beneficial impact of IFN-γ in the tumor microenvironment. For example, IFN-γ can exert direct cytotoxic or cytostatic effects on tumor cells, and contribute to tumor senescence and tumor ferroptosis. In addition, IFN-γ signaling on stromal cells can lead to reduced angiogenesis in the tumor bed, limiting nutrients and oxygen access. IFN-γ signaling can also lead to major histocompatibility complex (MHC) class I upregulation, thereby increasing tumor sensitivity to CD8+ T cell-mediated lysis. Finally, IFN-γ orchestrates the recruitment of natural killer (NK) cells, T cells and invariant NK T cells to the tumor by triggering production of the chemokines CXCL9, CXCL10 and CXCL11.

Conversely, intratumoral IFN-γ may have detrimental effects on the outcome of anti-tumor immune responses. In this regard, it has been reported that IFN-γ increases tumor cell genomic instability, a process that may favor tumor escape from the immune pressure. IFN-γ can also induce the expression of ligands for inhibitory receptors such as PD ligand 1 (PD-L1) and PD-L2 on stromal and tumor cells, contributing to the inhibition of tumor-infiltrating T cells. IFN-γ could also induce the deletion of tumor-reactive T cells by favoring activation-induced cell death. Finally, long-term IFN-γ exposure has been associated with deep transcriptomic changes in tumor cells that can contribute to the failure of immune checkpoint blockade therapy.

Despite our understanding of molecular and cellular mechanisms triggered by IFN-γ, we critically lack information on the spatiotemporal activity of this cytokine in the tumor microenvironment. Quantitative rules regulating cell–cell communication are only starting to be considered but are essential to understand complex immune responses. Fundamental questions remain to be addressed. Does IFN-γ act primarily in discrete areas or more widely in the tumor microenvironment? How is IFN-γ spatial activity shaped by the number of cytokine-producing T cells? What is the duration of IFN-γ exposure required to alter tumor cell function and phenotype? It is well established that IFN-γ is secreted by T cells in a directional manner toward the immunologic synapse formed with target cells. On T cell receptor (TCR) stimulation, IFN-γ production is also thought to be a brief (few hours’ long) event. Although these two features are expected to favor a local activity of IFN-γ around the sites of cytokine production, the extent of IFN-γ diffusion in tissue has yet to be fully understood. Indeed, using in vitro settings, it has been shown that the immunologic synapse does not perfectly restrict IFN-γ secretion, which accounts for bystander activity of IFN-γ in cells that are not in contact with cytokine-producing T cells. Evidence for bystander IFN-γ activity has also been reported in vivo in the mesenteric lymph node after *Toxoplasma gondii* infection and in the dermis infected with *Leishmania major*. Yet, the extent and functional consequences of IFN-γ diffusion in the microenvironment of distinct tumor types remain unclear.

In the present study, we investigated the spatiotemporal activity of IFN-γ in the tumor microenvironment using B cell lymphoma

1. Dynamics of Immune Responses Unit, Equipe Labellisée Ligue Contre le Cancer, Institut Pasteur, INSERM U1223, Paris, France. 2. University Paris Diderot, Sorbonne Paris Cité, Cellule Pasteur, Paris, France. 3. Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. 4. Systems Biology Group, Center for Bioinformatics, Biostatistics, and Integrative Biology and USR 3756, Institut Pasteur CNRS, Paris, France. 5. Sorbonne Université, Paris, France.
6. These authors contributed equally: Pierre Bost, Idan Milo. e-mail: philippe.bousso@pasteur.fr
and melanoma models. Using distinct approaches including mosaic tumors and intravital imaging with a real-time reporter of signal transducer and activator of transcription 1 (STAT1) nuclear translocation, we show that IFN-γ signaling occurs with extensive bystander activity throughout the tumor microenvironment, affecting both tumor and tumor-infiltrating immune cells. Although STAT1 activity in tumor cells was sustained, it did not require continuous interaction with T cells, confirming the extensive bystander activity of intratumoral IFN-γ. We provide evidence that prolonged IFN-γ exposure is in fact needed to alter tumor cell phenotype. Our results support a model in which IFN-γ regulates the tumor microenvironment by acting as a widespread and sustained cytokine field, which concentration is dependent on the collective T cell activity.

**Results**

**T cell-derived IFN-γ profoundly modifies the tumor microenvironment.** To characterize how intratumoral IFN-γ acts on tumor cells, we examined the expression of several markers, including MHC class I and PD-L1, in distinct tumor cell lines, on exposure to various concentrations of IFN-γ. Using models for Myc-driven B cell lymphoma (Eμ-myc)21 melanoma (B16.F10)22 and mammary tumor (E0771)22, we observed a dose-dependent upregulation of these markers for all cell lines (Fig. 1a,b, and see Extended data Fig. 1a). Phenotype changes on IFN-γ treatment followed an analog, rather than a digital, pattern with a concentration-dependent shift in expression of the whole population rather than a bimodal distribution (see Extended data Fig. 1b). At high concentrations, IFN-γ also induced substantial cell death of B16.F10 tumor cells and an even more pronounced effect when combined with tumor necrosis factor (TNF)-α (see Extended data Fig. 1c,c). By contrast, IFN-γ did not induce substantial cell death of Eμ-myc cells either alone or in combination with TNF-α, highlighting the diversity of the tumor response to this cytokine (see Extended data Fig. 1d,f).

To test whether these phenotypic changes also occur in response to intratumoral T cell-derived IFN-γ in vivo, we adoptively transferred OT-I CD8+ T cells into Rag2-/- mice bearing ovalumin+ (OVA+) Eμ-myc lymphomas and analyzed tumor phenotype in the bone marrow, a primary site of tumor growth (Fig. 1c, and see Extended data Fig. 2). We used recipients devoid of endogenous T cells for better control of the source of IFN-γ production. A fraction of OT-I T cells at the tumor site was found to produce IFN-γ (Fig. 1d). As shown in Fig. 1e,f, the presence of T cells resulted in an increased expression of MHC class I and PD-L1 on tumor cells. Phenotypic changes were also detected in tumor-infiltrating immune cells that expressed increased levels of MHC class I on their surface (Fig. 1g). All these effects appeared to be dependent on T cell-derived IFN-γ because they were abolished when IFN-γ-deficient OT-I T cells were transferred. Importantly, these results were recapitulated in Rag2-/-;γ-/- recipients that lack NK cells, suggesting that T cell-derived IFN-γ is sufficient to directly mediate phenotypic changes in tumor cells (Extended data Fig. 3). To extend these findings in a setting of chimeric antigen receptor (CAR) T cell therapy, we analyzed the phenotypic changes in lymphoma cells on transfer of IFN-γ either sufficient or deficient in anti-CD19 CAR T cells. The presence of wild-type (WT) but not IFN-γ-deficient CAR T cells resulted in a substantial increase in MHC class I and PD-L1 expression (Fig. 1h). Altogether, these results indicate that IFN-γ produced by T or CAR T cells has a broad impact on tumor and tumor-infiltrating immune cells.

**Tumor-infiltrating T cells selectively accumulate and arrest in tumor areas with cognate antigen expression.** Production of IFN-γ by T cells has been associated with stable interactions with antigen-presenting targets25. However, it is unclear whether IFN-γ produced by T cells acts on engaged targets or their close neighbors, or more extensively in the tumor microenvironment. To address this question, we wanted to create mosaic tumors in which only selective regions express a tumor antigen to spatially restrict IFN-γ production (Fig. 2a). We took advantage of our previous observation that tumor subclones tend to grow as anatomically segregated patches to create these disparate environments26. When we injected a mixture of antigen-positive (OVA+) and antigen-negative Eμ-myc tumors expressing distinct fluorescent proteins, we observed tumor patches with a typical radius of 100–500 μm in the bone marrow (Fig. 2b). Using intravital imaging, we found that most OT-I T cells selectively accumulated and arrested in antigen-positive areas where they displayed reduced velocity and increased confinement, and formed long-lasting interactions with tumor cells (Fig. 2b–g, and see Extended data Fig. 4 and Supplementary Video 1). The remaining T cells seen in antigen-negative tumor areas were rapidly migrating with no evidence for prolonged interactions (Fig. 2d–g, and see Extended data Fig. 4 and Supplementary Video 1). Some of the stable contacts with antigen-positive tumors most probably represent sites of cytokine secretion because previous studies have linked T cell arrest to IFN-γ secretion27. Even if it remains formally possible that migrating T cells also release small amounts of cytokine, the increased T cell numbers and residence time in antigen-positive areas would imply higher cytokine release in these regions. Of note, T cell arrest in antigen-positive areas was largely IFN-γ independent because prolonged interactions were also observed with IFN-γ-deficient T cells (see Extended data Fig. 4). In sum, this strategy provides a means to spatially control cognate interactions.
leading to IFN-γ production by T cells and to assess the occurrence of long-distance effects in antigen-negative areas.

**Strong bystander activity of T cell-derived IFN-γ in the tumor microenvironment.** Using the aforementioned setup, we asked how T cells impact on the phenotype of tumor cells located in antigen-positive versus antigen-negative areas. As IFN-γ induces MHC class I and PD-L1 upregulation, we relied on these readouts as a proxy for IFN-γ activity. We first used the Eµ-myc lymphoma model. When OT-I T cells were transferred in mice bearing antigen-negative tumors only, no changes in MHC class I and PD-L1 levels were observed (Fig. 3b–d) despite the presence of T cells in the tumor microenvironment (see Extended data Fig. 4e). This confirmed that local antigen restimulation is required for IFN-γ production. In mosaic tumors, the presence of T cells resulted in an increased expression of MHC class I and PD-L1. Importantly, both antigen-positive and antigen-negative tumor cells upregulated these markers to the same extent (Fig. 3e–d). The widespread effects mediated by IFN-γ could be due to the activity of numerous T cells, each of them acting on a limited number of bystander cells (Fig. 3e). Alternatively, efficient diffusion of IFN-γ may result in a relatively uniform intratumoral cytokine concentration, lead-

![Diagram](image-url)
Fig. 2 | Tumor antigen expression drives the selective accumulation and arrest of intratumoral T cells. a, Experimental setup. Rag2⁻/⁻ mice were injected with a 1:1 mixture of OVA-expressing and OVA-non-expressing Eµ-myc B lymphoma cells, labeled with CFP and YFP, respectively. On days 12-13, mice were injected with activated GFP⁺ OT-I T cells. After 2 d, intravital imaging of the bone marrow was performed. Scale bar, 50 µm. b–g, CD8⁺ OT-I T cells specifically accumulate and arrest in antigen-expressing cellular patches of mosaic tumors. b, Left: representative image of OVA⁺ (blue) or OVA⁻ (orange) tumor patches infiltrated with OT-I T cells (green). Scale bar, 50 µm. Right: time-lapse images (corresponding to the dashed squares) showing OT-I T cells (indicated by arrows) forming stable contacts with OVA⁺ Eµ-myc cells but not with OVA⁻ Eµ-myc cells. Scale bars, 15 µm (representing two independent experiments with n = 3 mice in each experiment). c, Individual tracks of OT-I T cells located either in OVA⁺ tumor patches (blue) or OVA⁻ tumor patches (orange). Numbers indicate distance in µm. d–f, Graphs show mean speed (d), straightness (e) and arrest coefficient (f) of individual OT-I T cells located in OVA⁺ or OVA⁻ tumor patches. Only tracks that were at least 5-min long were considered. g, Maximal duration of interaction of individual OT-I T cells with OVA⁺ or OVA⁻ Eµ-myc cells measured during 45-min-long videos. Results in c–g are from five videos obtained in two independent experiments with three mice per experiment. In d–g, each dot represents one individual track (n = 75 OT-I tracks in OVA⁺ tumor patches; n = 103 OT-I tracks in OVA⁻ tumor patches). Red lines indicate mean values (****P < 0.0001, two-tailed Mann-Whitney U-test).
ing to a homogeneous response, the intensity of which depends on the number of producing T cells (Fig. 3e). To distinguish between these two possibilities, we repeated our experiments with different numbers of T cells. We reasoned that, if T cells acted on a limited number of tumor cells, lowering the number of T cells should reduce the frequency of responding tumor cells without affecting the intensity of their response (best modeled by a bimodal distribution in tumor cell response) (Fig. 3e). By contrast, if IFN-γ primarily acted as a shared resource, reducing the number of T cells should decrease the intensity of the response in all cells (unimodal distribution of tumor cell response). Supporting the latter hypothesis, we found that a tenfold decrease in the number of transferred T cells resulted in a homogeneous reduction in MHC class I upregulation (Fig. 3f). Altogether, these experiments support the idea that IFN-γ diffuses extensively in the tumor microenvironment. As a result, the collective activity of IFN-γ-producing T cells generates a widespread field of IFN-γ, the concentration of which dictates phenotypic changes across the whole tumor. We measured that, when transferring $2 \times 10^5$ T cells, the ratio of OVA+ Eµ-myc cells to OT-I T cells was approximately 100:1. By comparing the fold increase in MHC class I levels observed in vivo with that seen in vitro (Fig. 1a), we estimated that this T cell density resulted in an intratumoral concentration of IFN-γ of around 1 nmol L$^{-1}$ (see Extended data Fig. 5a,b). This is obviously a rough estimate because several parameters could potentially differ between the in vitro and the in vivo settings (such as cytokine consumption or basal IFN-γ receptor (IFN-γR) expression). To test whether the activity of IFN-γ may influence tumor fate, we first assessed the effect of systemic IFN-γ delivery. We observed phenotypic changes in tumor cells, suggesting effective diffusion in vivo, and most importantly found reduced tumor burden in the bone marrow of treated mice (see Extended data Fig. 5c,d). This result suggests, that, in contrast to what was seen in vitro with Eµ-myc cells, IFN-γ exerts a negative effect on tumor development in vivo. This discrepancy may originate from an indirect effect of IFN-γ on host cells or from changes in IFN-γR expression/signalization on tumor cells in vivo. To test how T cell-derived IFN-γ may affect tumor growth in vivo, we relied on our mosaic tumor system and compared the impact of T cell-derived IFN-γ on the growth of both antigen-positive and antigen-negative tumors. When recipient mice were transferred with IFN-γ–competent T cells, we noted that antigen-positive (OVA+) cells were eliminated, but also that the antigen-negative tumor was partly controlled (see Extended data Fig. 5e,f). This partial control was lost on transfer of IFN-γ–deficient T cells (see Extended data Fig. 5e,f), supporting the idea that bystander IFN-γ activity may help limit tumor growth in vivo.

To extend these findings to solid tumors, we generated mosaic tumors using the B16.F10 melanoma model. In this setup, we again observed the hallmarks of IFN-γ signaling in both antigen-positives (OVA+ B16) and antigen-negative (B16) cells in the presence of T cells (Fig. 3g–i). Thus, extensive bystander activity of T cell–derived IFN-γ is a property that also pertains to solid tumors.

Widespread and sustained STAT1 signaling in the tumor microenvironment. The observed bystander changes in tumor phenotype prompted us to better characterize the spatiotemporal dynamics of IFN-γ signaling in the tumor microenvironment. We focused on STAT1 translocation in tumor cells, because this represents an early event after cytokine exposure. To this end, we generated a fluorescent reporter combining a STAT1–GFP (green fluorescent protein) fusion protein and a nuclear mCherry protein, and introduced this probe into Eµ-myc lymphoma cells. As expected, in the absence of stimulation, STAT1–GFP was cytoplasmic and excluded from the nucleus (Fig. 4a). On in vitro addition of IFN-γ, a more uniform (cytoplasmic and nuclear localization) distribution of STAT1–GFP was rapidly observed, confirming the translocation of a fraction of the STAT1–GFP pool (Fig. 4a). To quantify STAT1 translocation in multiple cells, we designed an automated image analysis procedure based on the detection of cell and nuclear shape and computing of Pearson’s correlation coefficients between STAT1–GFP and nuclear mCherry signals (see Extended data Fig. 6). With this strategy, statistically distinct translocation scores were observed for cells in the presence or absence of IFN-γ (Fig. 4b).

We therefore used these distinct patterns of STAT1 localization to identify cells with ongoing cytokine signaling in vivo by two-photon imaging. In a first model, we analyzed the response of H-2b+ B cell lymphoma in response to MataHari effector CD8+ T cells, which express a H-Y–specific transgenic TCR. In female Rag2–/– mice with established tumor expressing the fluorescent reporter, we predominantly observed a cytoplasmic-only distribution of STAT1–GFP, suggesting an absence of IFN-γ signaling (Fig. 4c–e). However, in mice that were also transferred with tumor-reactive MataHari effector CD8+ T cells, we observed evidence of cytokine signaling in a large number of tumor cells as detected by STAT1 translocation. This was evident from two-photon images (Fig. 4c,d) and confirmed by automated computing of STAT1 translocation scores (Fig. 4e). Importantly, STAT1 translocation was widespread and detected in tumor cells irrespective of whether or not they were contacting T cells. Consistently, there was no correlation between STAT1 translocation score in individual tumor cells and distance to the nearest T cells (Fig. 4f). Nuclear STAT1–GFP typically persisted throughout the imaging experiments (typically 1–2 h), suggesting sustained signaling (Fig. 4f, and see Supplementary Video 2). To extend these results, we repeated these experiments in a second model relying on OVA+ Eµ-myc B cell lymphoma and OT-I CD8+ T cells. In this setup also, nuclear STAT1–GFP was detected in the vast majority of tumor cells when OT-I effector CD8+ T cells were transferred but not in the absence of T cells (see Extended data...
Fig. 7a–d and Supplementary Video 3). To further demonstrate the ability of T cell-derived IFN-γ to signal in distant tumor cells, we created mosaic tumors containing antigen-negative tumor that expressed the STAT1–GFP reporter and fluorescently labeled antigen-positive tumor cells. As shown in Extended data Fig. 7e–h, we detected increased STAT1 activity in antigen-negative tumor cells when T cells and antigen-positive tumor cells were present compared with controls lacking antigen-negative tumor cells.

These results establish that STAT1 activity occurs at the tissue level rather than in discrete spots and indicate that STAT1 translocation does not require ongoing interactions with effector T cells.

Sustained IFN-γ signaling is required for alteration of tumor cell phenotype. Early events of IFN-γ signaling, including STAT1 phosphorylation and STAT1 translocation, are typically detected within minutes of cytokine exposure\(^3\). However, it is unclear whether...
transient or prolonged exposure to IFN-γ is required to alter tumor cell phenotype. We therefore designed an experiment in which Eμ-myc B lymphoma cells are cultured for a fixed period of time (24h), but exposed to IFN-γ during this period for various amounts of time (by washing out the cytokine and adding a blocking anti-IFN-γ monoclonal antibody to prevent any residual effects). RNA-sequencing (RNA-seq) was then performed for these different stimulation patterns (Fig. 5a). We confirmed that sustained IFN-γ exposure resulted in prolonged STAT1 translocation (see Extended Data Fig. 8). As shown in Fig. 5bc, the core of the response that included well-known IFN-γ-responsive genes (Cd74, Cd274, H2-K1, H2-D1, and so on) required at least 6–24h of cytokine exposure. Notably, cells exposed to IFN-γ for 1h were barely distinguishable from unstimulated cells. We confirmed these findings at the protein level for H2-Kb, H2-Dd and PD-L1 using flow cytometry (Fig. 5d). In these settings, 6–24h of cytokine exposure was needed for substantial upregulation of these molecules. The need for long-lasting cytokine exposure was not restricted to B lymphoma cells because we observed a similar requirement for B16.F10 melanoma cells (Fig. 5e). Altogether, these results suggested that sustained exposure to IFN-γ is needed to drive phenotypic changes in tumor cells, including MHC class I and PD-L1 upregulation. Our transcriptomic analysis also revealed a strong enrichment of interferon regulatory factor 8 (IRF8) and STAT1 motifs and that the expression of both transcription factors, IRF8 and STAT1, increased in response to IFN-γ (Fig. 5fg). This raises the possibility that a two-step signaling activation, in which STAT1 and IRF8 upregulation by IFN-γ subsequently potentiates the cellular response to this cytokine, accounts for the requirement for sustained cytokine exposure.

Probing IFN-γ activity in human melanoma. To test whether bystander IFN-γ activity also occurs in human tumors, we reanalyzed single-cell RNA-seq data from the immune infiltrate of eight non-treated melanoma patients. Overall, we found that CD8+ T cells were the main producers of IFN-γ because more than 80% of all IFN-γ mRNA molecules were specific to those cells (Fig. 6a).

We decided to investigate the impact of IFN-γ production on the tumor microenvironment by focusing on the monocyte/macrophage population as a sensor of intratumoral IFN-γ. Of note, the low number of tumor cells present in the samples precluded us from performing the same analysis on those cells. By performing overdispersion analysis, we identified one module (noted as pathway 7 in Extended data Fig. 9) specifically enriched in IFN-γ-responsive genes (Fig. 6b). Remarkably, the mean IFN-γ signature in monocytes/macrophages significantly correlated (R=0.91, P=0.0018; Fig. 6c) with the frequency of IFN-γ-producing CD8+ T cells. Similarly, when repeating the analysis with intratumoral neutrophils, we observed an IFN-γ signature that correlated with the fraction of IFN-γ-producing T cells (Fig. 6e,f). These data suggest that, as observed in our mouse models, IFN-γ signals in multiple cell types of the tumor microenvironment. The observation that neutrophils, which are not typical tumor antigen-presenting cells, also harbor the IFN-γ signature suggests that they have received IFN-γ signals in a bystander fashion.

We also noted that, in patients with a high density of IFN-γ-producing T cells, the distribution of the IFN-γ signature on monocytes and neutrophils appeared unimodally shifted toward higher values (Fig. 6g). This observation is compatible with relatively uniform IFN-γ activity in the tumor, which was also observed in head and neck squamous cell carcinoma (HNSCC) patients, as analyzed by single-cell RNA-seq in tumor cells (see Extended Data Fig. 10). Future studies assessing the spatial distribution of IFN-γ-responsive genes in relation to T cell infiltration will help refine our understanding of cytokine propagation and signaling in human tumors.

Discussion

In the present study, we established the spatiotemporal dynamics of intratumoral T cell–derived IFN-γ in models of B cell lymphoma and melanoma. We provided evidence that the collective activity of effector T cells, together with cytokine diffusion, drives a tumor-wide response. Moreover, the formation of a sustained field of IFN-γ was required to alter tumor cell phenotype.

The range of cytokine activity is dependent on the production rate and levels, diffusion and consumption by other cells. Our results support the idea that T cell production of IFN-γ generates a widespread cytokine field shared by most tumor cells and tumor-infiltrating immune cells. MHC class I expression was uniformly upregulated in the tumor microenvironment even in cells located at a distance (>100 µm) from the sites of cognate T cell interactions, a feature that may favor tumor lysis by cytotoxic T cells. Conversely, the expression of the inhibitory ligand PD-L1 was also upregulated by IFN-γ in a bystander fashion. Thus, both potentially beneficial and deleterious effects of IFN-γ appeared to be regulated by a cytokine field, rather than by very discrete cytokine hotspots.

At first, these results may appear at odds with the well-established notion that IFN-γ is produced by T cells at the immunologic synapse with target cells6,15,18, a phenomenon that should generate high local concentrations of IFN-γ at sites of antigen recognition and create extensive heterogeneity in cell responses. Our finding that prolonged exposure to IFN-γ is required to alter tumor cell phenotype probably provides an explanation for this apparent contradiction because these cytokine hotspots will probably dissipate.
due to diffusion well before they have the time to generate a local biological effect.

A direct consequence of this mode of action is that the activity of IFN-γ will be determined at the tissue level by the overall number of cytokine-producing cells. In this model, the collective activity of T cells is essential because it determines the overall cytokine concentration and the extent of tumor cell response, in a manner resembling quorum sensing. Such a collective mode of action was recently described for nitric oxide (NO)-producing macrophages, where a single macrophage has a negligible intrinsic

![Graphs and images showing the distribution and concentration of STAT1-GFP and Nucleus in different conditions:](image-url)
It has recently been shown that IFN-γ can bind to phosphatidylserine on the surface of tumor cells and be slowly released to prolong inflammation38. Such a mechanism, termed 'catch and release', activity, but a high number of NO-producing macrophages regulate activity at the tissue level in both NO-producing and NO-non-producing cells35,36.

**Fig. 5** | Sustained signaling is required to alter tumor cell phenotype. **a**, Experimental setup. Eμ-myc B lymphoma cells were stimulated with 5 ng ml\(^{-1}\) of IFN-γ. At 1h or 6h, the stimulation was blocked by adding 50 µg ml\(^{-1}\) of anti-IFN-γ monoclonal antibody. At 24h, cells were recovered for mRNA-seq. **b**, Heatmap of differentially expressed genes. Gene expression is normalized by row. **c**, Venn diagrams of differentially expressed genes between the various stimulated and control samples. **d,e**, Tumor cells were stimulated with 5 ng ml\(^{-1}\) of IFN-γ. At 1h or 6h, the stimulation was blocked by adding 50 µg ml\(^{-1}\) of anti-IFN-γ monoclonal antibody. At 24h, cells were recovered and analyzed by flow cytometry. Graphs represent H2-K\(^{b}\) (left), H2-D\(^{b}\) (middle) and PD-L1 (right) surface expression after indicated durations of stimulation on Eμ-myc (d) and B16.F10 (e) models. Each dot represents the mean of three technical replicates, representing three independent experiments. Unstim., unstimulated. **f,g**, Normalized enrichment score (NES) of the 10 most enriched motifs for the differentially expressed genes between 6-h-stimulated (f) and 24-h-stimulated (g) samples and control samples.
Fig. 6 | Assessing IFN-γ bystander activity in human melanoma samples. Measurement of the IFN-γ signature in the tumor microenvironment of melanoma patients using single-cell RNA-seq. **a** Contribution of distinct cell clusters to IFNG RNA unique molecular identifier (UMI) production. Data are compiled from all patients analyzed (n = 8). **b**, Gene contribution to the IFN-γ signature in the monocyte/macrophage cluster. **c**, Association between the percentage of IFN-γ+ cells (>1 UMI) among CD8+ cells and the IFN-γ score in monocytes/macrophages (R represents Pearson’s correlation coefficient; statistical significance was assessed using Fisher’s test). **d**, Distribution of the IFN-γ signature in monocytes/macrophages from three different patients harboring distinct indicated frequencies of IFN-γ+ CD8+ T cells (density in a.u., arbitrary units). **e**, Gene contribution to the IFN-γ signature in the neutrophil cluster. **f**, Association between the percentage of IFN-γ+ cells among CD8+ cells and the IFN-γ score in neutrophils (R represents Pearson’s correlation coefficient; statistical significance was assessed using Fisher’s test). **g**, Distribution of IFN-γ signature in neutrophils from three different patients harboring distinct indicated frequencies of IFN-γ+ CD8+ T cells.
may favor the formation of relatively stable cytokine fields by temporarily averaging cytokine availability in the tumor microenvironment. Binding of IFN-γ on the extracellular matrix should also achieve similar effects. It will be important to determine whether other cytokines rely on a similar global mode of signaling and assess how the superimposition of distinct cytokine fields may translate into complex biological responses at the tissue level, in particular in the context of the tumor microenvironment. This is particularly relevant for TNF-α which can act synergistically with IFN-γ to induce tumor cell death. Cytokine consumption plays an important role in shaping the spatial range of cytokine activity, and may generate fields of different sizes for distinct cytokines, possibly creating functional heterogeneity in the tissue.

In summary, our work provides new insights into the spatiotemporal activity of intratumoral IFN-γ with important implications for anti-tumor immune responses. Single-cell RNA-seq, as illustrated here, but also spatial transcriptomics, will help estimate the range of cytokine activity in human tumors. Finally, the use of systems immunology, together with the development of new probes to image cytokine propagation and activity in vivo, will be essential to establish the rules of cytokine-mediated communication at steady-state or during disease pathogenesis.

Methods

Mice and cell lines. The 6- to 8-week-old C57BL/6 mice were purchased from Charles River. Rag2<sup>−/−</sup>, Rag2<sup>−/−</sup>, γ<sup>−/−</sup>/γ<sup>−/−</sup> OT-I TCR, UBC-GFP Rag2<sup>−/−</sup>/OT-1 TCR, cyan fluorescent protein (CFP expressed under the actin promoter), Rag2<sup>−/−</sup> MHC class II restricted Uty<sub>246–254</sub> peptide (WMHHNMLDI) or H-2K<sub>b</sub>-restricted OVA<sub>257–264</sub> peptide (SIINFEKL) for 2 h in RPMI medium 1640–Glutamax at 37 °C. The rest of the cells were incubated at 37 °C in complete medium supplemented with 50 μg/ml of gentamicin. The two populations were mixed and cultured for 2–3 d.

Splenocytes from MataHari TCR transgenic mice or OT-I TCR transgenic mice were isolated and cultured in RPMI medium 1640 supplemented with 10% Fetal Bovine Serum (FBS), 2 mM L-glutamine, 100 units/ml penicillin, 100 μg/ml streptomycin, and 1 μg/ml of gentamicin. The two populations were mixed and cultured for 2–3 d.

Intratumoral two-photon imaging. Bone marrow imaging was performed as described<sup>3</sup>. Mice were anesthetized with a mixture of xylazine (Rompun, 10 mg/kg<sup>−1</sup>) and ketamine (Imalgène, 100 mg/kg<sup>−1</sup>), which was replenished hourly. During imaging, the mice temperature was maintained at 37 °C with a heating pad. Two-photon imaging was performed with an upright microscope (FV1000-510L, Olympus) and a ×25/0.5 numerical aperture, water-dipping objective (Olympus). Excitation was provided by an Insight DeepSee dual-laser (Spectra-Physics) system, the laser was tuned at 920 or 960 nm. The following filters were used: for fluorescence detection, CFP (483/32), GFP (520/35), YFP (542/27), background (593/35) and mCherry (624/40). To create time-lapse sequences, we typically scanned a 30- to 40-μm-thick volume of tissue at 5-μm Z-steps and 60-s intervals.

Image analysis. Videos were processed and analyzed using ImageJ or Fiji software (ImageJ 1.50). Videos and figures based on two-photon microscopy are shown as two-dimensional maximum intensity projections of three-dimensional data. For optimal contrast rendering, GFP was pseudo-colored in yellow in some images. Contact durations were calculated manually, based on close apposition between a given T cell and a tumor cell. To perform an unbiased analysis of STAT1–GFP translocation, we developed an in-house MATLAB script that first automatically detects cell envelope and cell nuclei, and then computes a translocation score. Briefly, both STAT1–GFP and nuclear mCherry channels are first binarized using Otsu’s method. Image closing with a circular structuring element is then applied to the STAT1–GFP binarized image, followed by a filling of the holes. Euclidean distance transform is then applied to both binarized images before using a watershed segmentation tool to segment the contiguous cell. For each identified cell, Pearson’s correlation between the STAT1–GFP and the mCherry signal is computed and considered as a good proxy for STAT1–GFP translocation (translocation score). This strategy was applied to both in vitro and in vivo data. After background subtraction, we added a step of contrast enhancement before image binarization: first, a Laplacian of Gaussian filter with a sigma parameter of size equal to the estimated size of the cells is used, followed by the extraction of the local peak with an h-dome transform<sup>9</sup>. Whenever the automated segmentation was not sufficiently accurate, we performed manual cell segmentation and computed Pearson’s correlation coefficient using the region of interest integrator and coco2 Fiji plugins.

Messenger RNA-seq and analysis. The mRNA seq was performed on cultured Ep-myc cells. The mRNAs were extracted using the RNeasy Mini-Kit (Qiagen) according to the manufacturer’s guidelines. RNA-seq libraries were prepared using the TrueSeq Stranded mRNA Library Preparation Kit, in accordance with the manufacturer’s instructions. Briefly, 500 ng of total RNA was used for purification and fragmentation of mRNA. Purified mRNA underwent first- and second-strand complementary DNA synthesis. Complementary DNA was then adenylated, ligated to Illumina sequencing adapters and amplified by PCR (using 10 cycles). Final libraries were evaluated using fluorescence-based assays including PicoGreen (Invitrogen), Qubit (Thermo Fisher) and Bioanalyzer (Agilent 2100). We sequenced on an Illumina HiSeq2500 sequencer (v4 chemistry) using 2 × 125 bpcap entry (bp) cycles. Ribosomal RNA abundance was measured by mapping with Bowtie<sup>2</sup>. Quality control of the reads was performed using Picard and RStudio<sup>4</sup>. The reads were aligned with STAR (v.2.5.2.a), and genes annotated in Gencode v.M16 were quantified using featureCounts (v.1.4.3-p1). Normalization and differential expression were done with the Bioconductor package DESeq2 (v.1.18.1). Differential expression analysis, as well as normalization, were performed using the DESeq2 package (v1.18.1) on R (v3.4.3). Analysis was performed according to the recommendation of the package with the same parameter, except for the log10-fold change shrinkage method parameter that was set to 'apeglm'. A gene was considered to be significantly differentially expressed if its corresponding adjusted P value was <0.01 and its absolute shrunk log(1/2) change > 1. Expression heatmaps were performed using the heatmap package, with Ward’s criterion used for the hierarchical clustering.

Enrichment of specific transcription factor-binding motifs in the promoter of upregulated genes was performed using the iloseq Cytoscape plugin<sup>8</sup>. miRnas were chosen as the reference species and all other parameters were set to default, except for the putative regulatory region and motif-ranking databases that were set to ‘500 bp upstream’.

Single-cell RNA-seq data clustering and analysis. The two single-cell RNA-seq datasets were downloaded on Gene Expression Omnibus (GEO) repository (accession numbers GSE123139 and GSE103322 for the melanoma and HNSCC data, respectively). Annotators were loaded using the GEOquery package (v2.68.0). As the HNSCC data were already preprocessed (as provided as log(tPM+1) data, TPM, transcripts per million), we used two slightly different settings for data preprocessing. In the case of the melanoma data, we processed as follows: to avoid removing populations of small-sized cells, a threshold of relatively low number of unique molecules (350 unique molecular identifiers or UMIs) was used in the first

312 NATURE CANCER | VOL 1 | MARCH 2020 | 302–314 | www.nature.com/natcancer
instance to filter out low-quality cells. Genes with fewer than 300 UMIs expressed were also removed to reduce computation cost and time. The filtered data were then analyzed using the PAGODA2 pipeline (https://github.com/lms-bmi/pagoda2/blob/master/vignettes/pagoda2 walkthrough oct2018.md). First, gene variance was corrected and highly variable genes identified using the adjust variance function (gam.k=10). The 10 first components of principal component analysis (PCA), based on the 3,000 most variable genes, were then computed using the calculateReduction function. Then a K-nearest-neighbors (KNN) graph was computed based on the PCA reduction through the makeKnnGraph function (k=30). Cosine distance was used at this stage to increase the robustness of the analysis. Last, Louvain community detection was performed on the KNN graph using the getKnnClusters function and the multilevel community functions from the igraph R package. Once the cells were clustered, we identified all genes that were significantly expressed by cells from a given cluster compared with all other cells identified using the PAGODA2 function getDifferentialGenes, which performs the Kruskal–Wallis test, with default parameters (corrected Z-score > 3, uncorrected only=TRUE). Using this gene list, we annotated the various cell clusters based on the specific expression of well-known immune markers (CD3, CD8A, TRBC2, and so on). Low-dimensional embedding of the data was performed using the UMAP algorithm. The uwot package implementation was used (https://github.com/lmldev/uwot). PCA dimensionality reductions computed using PAGODA2 were used as input data. The N_neighbors parameter was set to 40 and spread parameter to 4, and a cosine metric was used. In the case of the HNSCC, the data were already normalized and loaded into a PAGODA object. By default, the parameter set was used for the next steps of the analysis. The 50 first PCs were used, the 1,000 most variable genes were selected, the K parameter was set to 15 and a cosine distance was used.

Analysis of over-dispersed pathways. To detect biologically meaningful gene modules/pathways in a robust manner, we performed pathway over-dispersion analysis using de novo gene sets. We first selected a given cell population and performed gene expression and variance normalization on the raw UMI counts as described in the previous section. The variable genes were selected using the getOdGenes function from the PAGODA2 package with the 200 top genes selected. These genes were then grouped into 15 pathways using hierarchical clustering (Ward linkage and correlation-based distance). Over-dispersion of these pathways was computed using the testPathwayOverdispersion function from the same package with default parameters. Gene set enrichment was performed through a binomial test (binom.test function, alternative = ‘greater’) using the 200 most variable genes as the reference ‘universe’. The HALLMARK_INTERFERON_ALPHA_RESPONSE and HALLMARK_INTERFERON_GAMMA_RESPONSE gene sets were downloaded from the Gene Set Enrichment Analysis website (https://www.gSEA-msigdb.org/gsea/ msigdb/gene sets.jsp?collection=11). Multiple testing correction was performed using the Benjamini–Hochberg method (p.adjust R function).

Association between IFN-γ score and IFN-γ production. In the case of the melanoma dataset, the macrophage population was identified as the C/IQ/APOE-expressing cluster. The IFN-γ score was computed using the testPathwayOverdispersion function as described for the macrophage population as a mean value for each patient. Neutrophils were identified as the S100A8/S100A9/Lyr9-expressing cluster. Patients with fewer than 25 macrophages or neutrophils were removed from the analysis. The fraction of IFN-γ-expressing CD8+ T cells among T cells was computed by computing the number of cells annotated as coming from a ‘CD3+ gating belonging to a CD8+ T cell RNA-seq cluster, and has more than 0 UMI coming from the IFN-γ gene. The association between the two was assessed using a simple linear regression model (lm R function). The significance of the association was checked using Fisher’s test (binom.test function, alternative = ‘greater’) using the 200 most variable genes as the reference ‘universe’.

Statistics and reproducibility. All statistical tests were performed using GraphPad Prism v6. Points in graphs indicate individual mice, and lines indicate means. In bar graphs, bars indicate means and error bars indicate sample s.d. The statistical tests employed are detailed in the figure legends. No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during the experiments and outcome assessment.

Data availability
RNA-seq data reported in this paper are deposited in the National Center for Biotechnology Information GEO database (GSE140191). Previously published human single-cell RNA-seq data that were reanalyzed in the present study are available under accession codes GSE123199 and GSE103322. All other data supporting the findings of the present study are available from the corresponding author on reasonable request.

Code availability
The R and Matlab scripts used for the transcriptomic and image analysis are available on GitHub (https://github.com/PierreBSC/IFNG_Cancer_project).

Online content
Any methods, additional references, Nature Research reporting summaries, source data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s43018-020-0038-2.

Received: 28 May 2019; Accepted: 29 January 2020; Published online: 9 March 2020

References


Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Tumor cell responses to IFN-γ in vitro. a–b, Tumor cells respond to IFN-γ in an analogic manner in vitro. a E0771 mammary breast tumor cells were stimulated with indicated IFN-γ concentrations in vitro for 24h. H2-Kb (top), H2-Db (middle) and PD-L1 (bottom) surface expression was then analyzed by flow cytometry. Representative of two independent experiments. b, Representative histograms of H2-Db (left) and PD-L1 (right) surface expression in B16.F10 (top) or Eµ-myc (bottom) cells, after treatment with indicated IFN-γ concentrations for 24h. Representative of three independent experiments. c–f, Impact of IFN-γ and TNF-α on tumor cell death. B16.F10 (c,e) or Eµ-myc (d,f) cells were incubated with the indicated IFN-γ and/or TNF-α concentrations in vitro for 24h. Cells were then counted by flow cytometry. Representative of two independent experiments.
Extended Data Fig. 2 | Gating strategy for flow cytometry analysis of the tumor microenvironment. The figure depicts the gating strategies for identifying tumor cells, OT-I T cells, NK cells, monocytes and neutrophils present in the bone marrow of tumor bearing mice.
Extended Data Fig. 3 | NK cells are dispensable for tumor phenotypic changes upon T cell transfer. a, Experimental set-up. Rag2^−/−γc^−/− recipient mice were injected i.v with OVA-expressing Eμ-myc B lymphoma cells. On day 12-13, in vitro activated OT-I CD8^+ T cells were injected i.v. Two days later, the recipient bone marrow was harvested and analyzed by flow cytometry. b, Representative examples of histograms showing H2-K^b (left) and PD-L1 (right) surface expression on tumor cells isolated from the bone marrow of mice that were left untreated (filled grey) or injected with OT-I T cells (line, blue). c, H2-K^b (left), H2-D^q (middle) and PD-L1 (right) surface expression on tumor cells isolated from mice treated or not with OT-I T cells, as assessed by flow cytometry. Each dot represents one mouse with n=4 mice per group. Red lines indicate mean values. (* P<0.05, two-tailed Mann-Whitney U-test).
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | OT-I T cells arrest on antigen-positive tumor cells independently of IFN-γ. a, Rag2−/− mice were injected with a 1:1 mixture of OVA-expressing and non-expressing Eµ-myc B lymphoma cells, labeled with CFP and YFP, respectively. On day 12-13, mice were injected with activated GFP+ OT-I T cells. Two days later, intravital imaging of the bone marrow was performed. The graph shows the relationship between mean T cell velocity and straightness in the OVA+ (left) and OVA− (right) tumor areas. Each dot represents one T cell track (OVA− areas n = 103 tracks, OVA+ areas n=75 tracks).
b-d, Rag2−/− mice were injected with a 1:1 mixture of OVA-expressing and non-expressing Eµ-myc B lymphoma cells, labeled with CFP and GFP, respectively. On day 12-13, mice were injected with activated WT or IFN-γ-deficient OT-I T cells transduced to express the mCherry fluorescent protein. Two days later, intravital imaging of the bone marrow was performed. b, Representative image of OVA+ (blue) tumor patches infiltrated with IFN-γ-deficient OT-I T cells (red). Scale bar: 50 µm. Right. Time lapse images (corresponding to the dashed squares) showing IFN-γ-deficient OT-I T cells (pointed by arrows) forming stable contacts with OVA+ Eµ-myc cells. Scale bar: 15 µm. c-d, Both WT and IFN-γ-deficient OT-I T cells decelerate in OVA+ tumor areas. Graphs show mean velocities (c) and arrest coefficient (d), for individual WT or IFN-γ-deficient T cells in OVA+ (blue) and OVA− (orange) tumor areas. Each dot represents one track. Red lines indicate mean values. Data shown in b-d, are representative of two independent experiments with n=3 mice per group. (* P<0.05; ** P<0.01; *** P<0.001; two-tailed Mann-Whitney U-test). e, OT-I T cells infiltrate antigen-negative tumors in the bone marrow. Rag2−/− mice were injected with either Eµ-myc alone or a 1:1 mixture of Eµ-myc and OVA-expressing Eµ-myc B lymphoma cells (labeled with different fluorescent proteins). On day 12-13, recipients were injected with OT-I T cells or left untreated. Two days later, the bone marrow of the mice was recovered and analyzed by flow cytometry. The graph shows that OT-I T cells can efficiently infiltrate tumors that contain only antigen-negative cells. Each dot represents one mouse with n=4 mice per group. Red lines indicate mean values.
Extended Data Fig. 5 | T cell-derived IFN-γ can control bystander tumor cells. a–b, Estimating the intratumoral IFN-γ concentration upon T cell transfer. a, The ratio of OT-I T cells to OVA+Eµ-myc tumors in the bone marrow (measured at day 2) is graphed as a function of the total number of transferred T cells. Results are shown as mean±SD with n=5 mice per group. b, The fold change in H2-D^d levels measured in Eµ-myc tumors in vitro is graphed as a function of IFN-γ concentration. The graph was used to infer the putative IFN-γ concentration in vivo (dashed red line) upon transfer of 2x10^6 T cells.

c–f, Control of tumor burden by T cell-derived IFN-γ. c–d, Recipient female Rag2^-/- mice were injected with h-Y+Eµ-myc B lymphoma cells. After 3 weeks, mice were either injected i.v. with recombinant IFN-γ (1 µg) twice (one day apart) or left untreated. On day 2, tumor burden c, and phenotype d, in the bone marrow were assessed by flow cytometry. Each dot represents one mouse (n=3 and n=4 mice for the untreated and the treated group, respectively). Red lines indicate mean values. (* P<0.05, two-tailed unpaired t-test). e–f, Functional activity of intratumoral IFN-γ. e, Experimental set-up. Rag2^-/- mice were injected with a 1:1 mixture of Eµ-myc and OVA-expressing Eµ-myc B lymphoma cells (labeled with different fluorescent proteins). On day 12-13, recipients were injected with either WT or IFN-γ-deficient OT-I T cells or left untreated. Five days later, the bone marrow was recovered and analyzed by flow cytometry. f, Graphs show the residual numbers of antigen-positive (top) and antigen-negative (bottom) tumors following transfer of the indicated T cell populations. Each dot represents one mouse (n=6 mice per group). Red lines indicate mean values. (* P<0.05, ** P<0.01, two-tailed Mann-Whitney U-test). Representative of two independent experiments. Of note, small deviation from the initial ratio seen in the absence of T cells is most likely due to minor differences in tumor cell division time.
Extended Data Fig. 6 | Automated procedure for quantifying STAT1 translocation in tumor cells imaged in vitro or in vivo. a, The diagram recapitulates the various steps used for image processing and quantification (see methods for details). b, Representative example of cell segmentation from the STAT1-GFP signal in tumor cells. Scale bar: 50 µm.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Tumor-reactive CD8+ T cells mediate widespread and sustained STAT1 activity in the tumor microenvironment. a, Experimental set-up. Recipient Rag2−/−γ−/− mice were injected with OVA+ Eµ-myc B lymphoma cells expressing the STAT1-GFP reporter and a nuclear mCherry protein. After 3 weeks, activated CD8+ T cells bearing the OT-I TCR were injected i.v. One day later, recipients were subjected to intravital imaging of the bone marrow. b, STAT1-GFP is largely excluded from the nucleus in tumor cells developing in the absence of T cells. Representative two-photon images (scale bar: 20µm), highlighting two specific regions (insets, scale bar: 10µm). c, Detection of nuclear STAT1-GFP in T cell-infiltrated tumors. Representative two-photon images (scale bar: 20µm), highlighting two specific regions (insets, scale bar: 10µm). d, Translocation score was computed from two-photon images obtained in mice left untreated (no T cells) or transferred with OT-I T cells (n=10 cells per group, box plot showing the median, first and third quartile and min and max values, *** P<0.001, two-tailed Mann-Whitney U-test). Data in b–d are representative of two independent experiments.

e–h STAT1 signaling is detected at distance from antigen-positive tumor cells. e, Experimental set-up. Rag2−/− mice were injected with Eµ-myc cells expressing the STAT1 reporter either alone or mixed at a 1:1 ratio with OVA+ Eµ-myc (expressing mCFP). Two weeks later, all mice were adoptively transferred with activated OT-I T cells. After 2 days, mice were subjected to intravital imaging of the bone marrow. f, STAT1-GFP is largely excluded from the nucleus in tumor cells developing in the absence of antigen-positive tumors. Representative two-photon images (scale bar: 20µm), highlighting two specific regions. Scale bar, 20 µm. g, Detection of STAT1 translocation in antigen-negative tumor cells developing in the presence of antigen-positive tumor cells. Representative two-photon images (scale bar: 20µm), highlighting two specific regions. Scale bar, 20 µm. h, Translocation score was computed from two-photon images obtained in mice bearing STAT1-GFP-expressing tumor cells alone, or bearing mosaic tumors containing both STAT1-GFP and OVA+ tumor cells. (n=10 cells per group, box plot showing the median, first and third quartile and min and max values, ** P<0.01, two-tailed Mann-Whitney U-test). Data in f–h are representative of n=3 mice per group.
Extended Data Fig. 8 | Prolonged STAT1 translocation in tumor cells with sustained exposure to IFN-γ. a, Experimental set-up. Eμ-myc cells expressing STAT1-GFP and nuclear mCherry were cultured for one hour in presence of recombinant IFN-γ. Cells were washed and either incubated with anti-IFN-γ mAb (to limit cytokine exposure), or re-incubated with IFN-γ (to prolong cytokine exposure). After an additional hour, cells were imaged. b, STAT1 translocation score was computed automatically for all individual cells (unstimulated n=593 cells, IFN-γ n=583, anti-IFN-γ n=570 cells). Each dot represents one cell. Red lines indicate mean values. (**** P<0.0001, *** P<0.001, two-tailed Mann-Whitney U-test).
Extended Data Fig. 9 | Determination of an IFNG signature in monocyte/macrophage cluster. 15 over-dispersed gene sets were determined for the monocyte/macrophage cluster as explained in materials and methods. Graph shows the contribution of IFNG or type I IFN-related genes for each gene sets. Arrow highlights a gene signature that is specific of IFN-γ signaling (hereafter called IFNG signature).
Extended Data Fig. 10 | Distribution of IFN-γ signature in tumor cells from head and neck squamous cell carcinoma patients. a. Gene contribution to the IFNG signature in the tumor cells. b. Distribution of IFNG signature in tumor cells from 13 different patients with head and neck squamous cell carcinoma patients. Although the mean of the distribution varies from patient to patient, a relatively uniform distribution is observed in most patients.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- □ The statistical test(s) used AND whether they are one- or two-sided
- □ Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- □ A description of all covariates tested
- □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- □ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- □ Give P values as exact values whenever possible.
- □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- □ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

- For FACs experiments, samples were acquired using either BD FACSDiva software (BD Biosciences) or CytExpert 2.3 software (Beckman Coulter).
- For microscopy experiments, movies were acquired using FV315-SW software (Olympus)

Data analysis

- Flow cytometry data were analyzed using FlowJo v10.4.1 (Tree Star). Microscopy movies were created from raw data and analyzed using Fiji software (ImageJ v1.50i). Alternatively, analysis was performed on Imaris 7.4.2 (Bitplane) or Matlab2019b(R). Data were represented using GraphPad Prism v6.
- For RNAseq data, Bowtie2 was used to determine RNA abundance. Quality control was performed using Picard and RSeQC. Reads were aligned using STAR (version 2.5.2a) and quantified with featureCounts (version 1.4.3-p1). Further analysis was performed using R (version 3.4.3) with the DESeq2 package (version 1.18.1). Enrichment of specific transcription factors motifs was analyzed using the iRegulon Cytoscape plugin.
- For single-cell RNAseq data, data were analyzed using R once again, with the GEOquery (version 2.48.0), Pagoda2, ighraph packages.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Already published Single-cell RNAseq data were downloaded from GEO, accession number GSE123139
Field-specific reporting
Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/hr-reporting-summary-flat.pdf

Life sciences study design
All studies must disclose on these points even when the disclosure is negative.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>No sample size calculation was performed. Sample sizes were determined based on previous experiments in the laboratory, in order to get statistically significant results and limit mice number at the same time.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exclusions</td>
<td>For microscopy experiments, mice with in which the fluorescent signal to noise ratio was too low were not analyzed. The criteria for exclusion was the inability to discriminate unambiguously individual cell contours.</td>
</tr>
<tr>
<td>Replication</td>
<td>Experiments were repeated several times, as indicated in figure legends. All repetitions of the experiments yielded similar results. Mice were housed in the same facility, in the same room and were age-matched as best as possible in every experiments and between different experiments. At the time of injection, tumor cells were less than one month in culture and all fluorescent cell lines were prepared from the same original uncultored line.</td>
</tr>
<tr>
<td>Randomization</td>
<td>Age- and sex-matched mice were used. For experiments using R16.10 melanoma cells, different groups of mice were randomly attributed to different experimental conditions. For experiments using B cell lymphoma, different groups of mice were attributed to different experimental conditions according to the percentage of tumor cells in the blood, in order to have comparable groups in terms of tumor growth before T cells injection.</td>
</tr>
<tr>
<td>Blinding</td>
<td>The experimentators were not blinded to experimental conditions during experiments of experiments analysis. Experiments and analyses were often performed by the same person.</td>
</tr>
</tbody>
</table>

Reporting for specific materials, systems and methods
We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

<table>
<thead>
<tr>
<th>Materials &amp; experimental systems</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>☑ Antibodies</td>
<td>☑ CHIP-seq</td>
</tr>
<tr>
<td>☑ Eukaryotic cell lines</td>
<td>☑ Flow cytometry</td>
</tr>
<tr>
<td>☑ Palaeontology</td>
<td>☑ MRI-based neuroimaging</td>
</tr>
<tr>
<td>☑ Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>☑ Human research participants</td>
<td></td>
</tr>
<tr>
<td>☑ Clinical data</td>
<td></td>
</tr>
</tbody>
</table>

Antibodies
The following antibodies were used:
- anti-mouse PD-L1, biotin: clone 10F.9G2, BioLegend # 124305, lot #B220794, dilution 1/100
- anti-mouse PD-L1 BV785: clone 10F.9G2, BioLegend # 124331, lot #B253998, dilution 1/100
- anti-mouse H2-Kb-FITC, clone A6G8-85.5, Biolegend # 110519, lot #4248976, dilution 1/100
- anti-mouse H2-Kb-UV595, clone A6G8-85.5, BD Biosciences # 742864, lot #B227699, dilution 1/100
- anti-mouse H2-Db-APC, clone H9S5, BioLegend # 111514, lot #B242441, dilution 1/100
- anti-mouse CD8-UV595, clone 53-6-7, BD Biosciences # 563786, lot #B072932, dilution 1/200
- anti-mouse CD8-BV786, clone 53-6-7, BD Biosciences #563332, dilution 1/200
- anti-mouse INF-γ, purified: clone XM1G1.2, BioXcell #BED055, lot #63001711. Used at 50 μg/mL for in vitro blocking purposes
- anti-mouse IFN-γ APC, clone XM1G1.2, eBioscience #51-7311-82, lot #E08729-1131, dilution 1/100
- anti-mouse/human CD11b, clone M1/79, BioLegend #101228, lot #B260495, dilution 1/100
- anti-mouse NK1.1-PE, clone PK136, eBioscience #12-5941-83, lot #EG30343, dilution 1/100
- anti-mouse Ly6C-APC, clone 1A8, BioLegend #127610, lot #A55839, dilution 1/100
- anti-mouse Ly6C-BV605, clone HK1.4, BioLegend #128036, lot #B2194B4, dilution 1/100
Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
The Eu-myc cell line was derived in our laboratory from a male Eu-myc transgenic mouse. Eu-myc mice harbor a transgene with the c-myc gene under the IgH promoter and develop spontaneous B lymphoma. HEK cells were purchased at American Type Culture Collection. B16.F10 melanoma cells were a gift from Guy Shakhar (Weizmann Institute, Rehovot, Israel).

Authentication
HEK and B16.F10 cells were previously authenticated by ATCC (see vendor website). The phenotype of the cell lines used in the study was checked after thawing of the cells.

Mycoplasma contamination
HEK and B16.F10 cells were tested negative for mycoplasma contamination by PCR. Eu-myc cells were also tested negative at the exception of YFP+ cells which where not tested.

Commonly misidentified lines
No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals, ARRIVE guidelines recommended for reporting animal research

Laboratory animals
CS7Bl/6J Rag2tm1(Rag2-deficient mice), CS7Bl/6J Tg(UBC-GFP)30Scha Tg[TeraTcrib]1LOOM/Jb Rag1tm1Bal (Rag1-deficient; OT-I TCR transgenic, GFP-expressing mice), CS7Bl/6J Tg(Act-CFP)1 expressing CFP under the actin promoter), CS7Bl/6J Tg[TCR Matahari] Rag2tm1Fawa (Rag2-deficient, Matahari TCR transgenic) and CS7Bl/6J Rag2tm1Fawa Il2ghtm1Cgn (Rag2 and g common chain-deficient) were housed and bred in our facility.

For experiments involving OT-I cells transfer, age- and sex-matched male or female mice were used. For experiments involving Matahari T cells transfer, only age-matched female mice were used. Mice were 6 to 25 weeks old at the time of the experiments and all experiments were performed in agreement with relevant Institut Pasteur internal guidelines and regulations (CETEA committee, protocol number 170038).

Wild animals
The study did not involve wild animals

Field-collected samples
The study did not involved field-collected samples

Ethics oversight
All procedures were approved by the Institut Pasteur Committee on Animal Welfare (CETEA) under protocol code 170038, as well as the French Superior Teaching and Research Ministry

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
☒ All plots are contour plots with outliers or pseudocolor plots.
☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

For experiments using B16.F10 melanoma in vivo, whole tumors were recovered and cut into small parts, then digested with Collagenase and DNase I for 30 min at 37°C. Digested tumors were then further crushed on a cell strainer and tumor cells were recovered before antibody stainings and flow cytometry analysis.

For experiments using Eu-myc lymphoma in vivo, tumor growth was monitored by measuring tumor cell percentage in the blood. To do so, blood was recovered in EDTA and red blood cells were lysed using ACK solution before analysis at the cytometer. Once tumors were established, bone marrow cells were recovered by flushing one tibia and one femur with FACS buffer (PBS supplemented with 0.5% FCS) through a cell strainer. Single-cell suspensions were then recovered and stained before analysis by flow cytometry.
A majority of the data was acquired using a Cytoflex LX [serial number BA42021]. Some data were acquired using a BD Canto II.

**Software**
BD FACSDiva was used for BD Canto II-acquired data. For cytoflex-acquired data, CytExpert version 2.3 was used.

**Cell population abundance**
No cell sorting was performed during experiments.

**Gating strategy**
For experiments using B16.F10 melanoma cells, cell doublets were gated out based on SSC-A/SSC-H followed by an additional gating based on FSC-A/FSC-W. Live cells were isolated by FSC-A/SSC-A gating. Injected OT-I T cells were distinguished based on their GFP expression. B16.F10 cells expressing OVA or not were then gated on based on their expression of TdTomato and membrane CFP respectively. A back gating on FSC-A/SSC-A was used to confirm the discrimination of B16.F10 cells.

For experiments using Eμ-myc B cell lymphoma in vivo, cell doublets were excluded as described above. CFP and GFP were used to detect Eμ-myc expressing OVA or not respectively. CD8 staining enabled further distinction of GFP+ OT-I T cells from GFP+ tumor cells.

In a distinct panel, CD45 enabled distinction of hematopoietic cells. OVA+ Eμ-myc cells and OT-I T cells were gated out using GFP and CD8 expression as above. Likewise, OVA+ Eμ-myc cells were gated out by their CFP expression. Among remaining cells, myeloid cells were distinguished by their CD11b expression and further divided into monocytes (Ly6C High Ly6G-) and neutrophils (Ly6C+ Ly6G+). Finally, NK cells were identified as CD45+ CD11b– NK1.1+. A figure describing the gating strategy that was used to identify tumor, NK, OT-I T cells, as well as monocytes and neutrophils, is provided in the manuscript.

☑️ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.
4 Going deeper than conventional imaging to study immune cross-talks

Imaging technologies are essential tools to study cell communication as they provide the spatial location of each individual cell, including ligand-producing and responder cells, as shown by our study. Using intravital imaging, we show that IFNg has a broad activity on the tumor microenvironment in multiple mouse models of tumor, and does not act on a limited set of target cells. Moreover, we observed a persistent activation of the IFNg pathway in-vivo that results in the expression of key genes such as PD-L1, both in mouse and human tumors. In this paper we succeeded at combining image analysis of mice tumor with secondary analysis of previously published scRNA-seq datasets generated from human tumor samples to validate our findings. While efficient, a more straightforward approach would have been to directly perform in-depth imaging analysis of human cancer samples using highly multiplexed imaging methods.

Such methods have emerged in parallel to the development of single-cell genomic technologies: over the last years, several teams have developed highly-multiplexed imaging methods able to quantify dozens of transcripts or proteins at a cellular, or even sub-cellular resolution (Figure III.2). Indeed, conventional imaging methods can only measure a limited set of features as they rely on fluorescent dyes to detect the targeted transcripts or proteins. Thus, only up to 4-5 different features can be imaged simultaneously due to spectral overlap.

The first method to be developed was multiplexed (or cyclic) immunofluorescence (IF) (Gerdes et al., 2013; Lin et al., 2015). Samples are successively stained with a set of fluorophore-tagged antibodies, imaged before undergoing a step of fluorophore inactivation usually based on a powerful oxydant such as hydrogen peroxyde (Figure III.2, top panel). As up to a dozen of cycles can be performed, no less than 40 different proteins can be imaged in a single experiment. More efficient, the commercial solution CODEX (CO-Detection by indEXing) technology was released recently and further increase the number of simultaneously imaged proteins while simplifying the experimental protocol (Goltsev et al., 2018).

Multiplexed Ion Beam Imaging (MIBI) and Imaging Mass Cytometry (IMC) use a different strategy: antibodies used to stain the samples are labeled by heavy metals that are not found in biological organisms (transition metals or lanthanides) instead of fluorophores (Figure III.2, middle panel) (Baharlou et al., 2019). Using oxygen-based primary ion beam (for MIBI) or a laser (for IMC), small areas of the samples are ablated and then analyzed using a time-of-flight mass spectrometer. Heavy metals are detected and used to quantify the antibody they were bound to, and therefore the targeted proteins. Due to the limited signal overlap compared to fluorescence imaging, up to a hundred proteins can be imaged at the same time, but at the price of expensive reagents and devices (Baharlou et al., 2019).

Lastly, multiplexed version of single molecular fluorescence in situ hybridization (smFISH)
Figure III.2: Methods for highly multiplexed imaging. Three types of methods have been developed: multiplexed IF which relies on successive cycles of staining and bleaching, MIBI/IMC an imaging version of mass cytometry and SeqFISH/MERISH, two multiplexed versions of single molecular FISH.

(Raj et al., 2008) were developed to measure gene expression of hundreds to thousands of genes. By combining new enzyme-independent amplification signal methods and DNA probe removal by DNAs, two laboratory simultaneously developed similar protocols termed Sequential FISH (SeqFISH) (Shah et al., 2016) and Multiplexed Error Robust FISH (MERFISH) (Chen et al., 2015). Interestingly data generated by such methods can be compared to the ones generated by scRNA-seq using advanced analytical strategies, including autoencoder neural networks, thus allowing to efficiently annotate SeqFISH datasets or to infer the location of sequenced cells in a tissue (Zhu et al., 2018; Lopez et al., 2019).

All those methods represent a unique opportunity to study immune communication as they can track the production of multiple cytokines and their potential effects on thousands of cells simultaneously. However, the lack of adequate computational tools for spatial statistical analysis is a major hurdle to perform such analysis and so far only basic cellular neighborhood analysis have been used (Goltsev et al., 2018). Therefore, the development of those computational methods is essential to fully analyze high dimensional spatial data and better understand immune cell communications.
Quantitative study of IFNγ effects in tumors
Chapter IV

Contribution of oligodendrocytes to Alzheimer’s disease

1 An immune flavor in neurodegenerative diseases

Dementia is defined as a chronic or persistent disorder of the mental processes likely caused by brain disease or injury and featured by memory disorders, personality changes, and cognitive decline. It represents a massive cost to the society (more than US$600 billion in 2010) and its incidence keeps increasing due to the global aging of the population. Alzheimer Disease (AD) is the most common cause of dementia as it affects between 10 to 30% of the >65 years of age population and more than a third of >90 year old people. Other common dementia associated diseases include mainly vascular dementia and Parkinson’s disease. Unfortunately, no efficient therapies have been identified so far and treatment is mostly symptomatic and aims at improving patients quality of life while being poorly efficient (Masters et al., 2015; Heppner et al., 2015).

AD can be defined as a fatal, progressive and unremitting neurodegenerative disease that affects the cerebral cortex and the hippocampus. It is featured by the progressive accumulation of two different protein aggregates (Masters et al., 2015; Querfurth and LaFerla, 2010; Scheltens et al., 2016): amyloid-β (Aβ) plaques and tau (τ) neurofibrillary tangles. Aβ plaques are extracellular protein aggregates derived from the abnormal cleavage of the amyloid precursor protein (APP) by enzymes called secretases whereas τ neurofibrillary tangles are mostly located in neurons and produced through hyperphosphorylation of the microtubule associated protein τ. Such pathological structures are initially detected in the frontal and temporal lobes, but progressively spread over other brain areas (Figure IV.1) and cause significant neuronal toxicity.

Approximately 5% of AD cases are diagnosed before the age of 65, an AD form called early-onset Alzheimer’s disease. As more than 60% of those patients have a family history of AD, this pathology is sometimes called Familial Alzheimer’s Disease (FAD). Genetic studies of FAD
Figure IV.1: The pathological features of AD and their evolution over time. (A) Microscopy images from amyloid plaques and neurofibrillary tangles published in Spielmeyer’s classic textbook ‘Histopathologie des Nervensystems’. (B) Spatial and temporal description of Aβ plaques and neurofibrillary tangles evolution in the brain of AD patients. Adapted from (Masters et al., 2015)

have highlighted the dominant role of Aβ in FAD as the three mutations associated with FAD are linked to Aβ, namely the precursor gene APP and Presenilin 1 and 2 (PSEN1 and PSEN2), two components of the γ-secretase complex, allowing to develop the ‘amyloid hypothesis’. It is thought that abnormal production or cleavage of the APP protein triggers the release of a soluble form of Aβ that can then aggregate and form Aβ fibrils and plaques in the extracellular space. Such plaques can be cleared by the brain phagocytic compartment (microglia) or degraded by extracellular proteases released by astrocytes, but also by infiltrating myeloid cells. However in AD patients brain, toxic aggregated soluble forms of Aβ (Aβ oligomers) are thought to diffuse and significantly alter neuronal functions and survival. Through an unknown mechanism, Aβ oligomers induce the intracellular aggregation of τ proteins that damage and kill the neurons. The released neurofibrillary tangles can be taken up by neighbouring neurons further triggering neuronal death (Figure IV.2).

Unlike genetic studies of FAD, genome wide association studies (GWAS) of non familial AD have pointed out the possible contribution of immune cells in AD: indeed, several AD genetically associated genes, such as TREM2 or TYROBP, are specifically expressed by microglia, brain resident macrophages and are associated to immune and not neuronal processes (Deczkowska et al., 2018). Elegant in-vivo lineage tracing studies have established that adult microglia derive from primitive myeloid progenitors that arise early during embryological life and see through the whole brain and are distinct from monocyte-derived phagocyte cells (Ginhoux et al., 2010). Microglia are known to contribute to the brain homeostasis by phagocytizing
Contribution of oligodendrocytes to Alzheimer’s disease

Figure IV.2: Mechanisms of protein aggregation in AD. (1) APP is produced by neurons and cleaved to produce soluble Aβ. (2) Soluble Aβ aggregate to form Aβ fibrils and plaques (3) which are destroyed in a phagocytic or protease dependent manner. (4) In AD, toxic Aβ oligomers dissociate from plaques and are toxic to surrounding neurons and (5) induce tau aggregation within neurons. (6) Neurofibrillary tangles induce neuronal death, are released and can be internalized by neighboring neurons. Adapted from (Masters et al., 2015)

apoptotic cells and protein aggregates, promoting synapse pruning and regulating stem cells survival (Li and Barres, 2018). However, there is exact role in neurodegenerative conditions remain controversial: while being able to engulf and destroy protein aggregates, uncontrolled activation of microglia is thought to trigger neuronal phagocytosis, further increasing neuronal degeneration and cognitive decline (Li and Barres, 2018). ScRNA-seq studies have identified a new subset of microglia called Disease-Associated Microglia (DAM) in several AD mouse models (Keren-Shaul et al., 2017), but also in human AD, multiple sclerosis (MS), Amyotrophic lateral sclerosis and aging patients (Deczkowska et al., 2018). DAM are featured by the expression of several genes associated to lysosomal, lipid metabolism and phagocytic pathways, suggesting that their ability to clear plaques and neurofibrillary tangles is increased (Figure IV.3). In addition to microglia, recent studies have pointed out the possible contribution
of T-cells to neurodegeneration as cytotoxic T cells were enriched in AD patient brains and cerebrospinal fluid (Gate et al., 2020), as well as in neurogenic niches of aging human brains. (Dulken et al., 2019).

2 A quick introduction to oligodendrocytes biology, functions and involvement in AD

Discovered more than a century ago by multiple major histologists such as Virchow and Golgi, oligodendrocytes are the major myelinating cells of the central nervous system (CNS) and represent up to 10% of the brain cells (Kuhn et al., 2019). Like other glial (non neuron) cells, their main purpose is to support nervous system cells, and more precisely to support action potential transmission. This task is achieved through the massive production of myelin sheath, extensions of plasma membrane that wrap around neurons axons in a concentric manner (Figure IV.4) and are mostly (80%) composed of specific lipids such as sphingomyelin, but also of key proteins such as PLP1 (ProteoLipid Protein 1) and MBP (Myelin Basic Protein). Each oligodendrocyte can wrap around multiple axons, with a mean of 50 axons covered by each oligodendrocyte (Kuhn et al., 2019).

Myelin sheaths are essential for rapid action potential transmission through the axon, but are also required for the metabolic support of the ensheathed neurons. Due to their high
lipid content, myelin sheaths are electrical insulators and therefore locally isolate the wrapped axon regions. Action potential thus 'jumps' from regions devoid of myelin, i.e nodes of Ranvier (Figure IV.4) (Kuhn et al., 2019) to the next, in a process termed saltatory conduction. Action potentials are much faster when traveling on myelinated axons, and therefore an efficient myelination of the neurons is required for a fully functional nervous system. As mentioned above, oligodendrocytes provide a direct metabolic support to the wrapped neurons. Thanks to high expression levels of glycolytic and Krebs cycle’s enzymes, oligodendrocytes generate large amounts of lactate and transfer it to axons through various lactate transporters, namely MCT1, MCT2 and MCT4 (Philips and Rothstein, 2017; Kuhn et al., 2019). Due to its prominent position in the cellular metabolic network, lactate is directly metabolized to generate ATP in axons. This metabolic support is essential as shown by the dramatic effects of MCT1 depletion on neurons survival in spinal cord culture (Lee et al., 2012). Oligodendrocytes are
Contribution of oligodendrocytes to Alzheimer’s disease

therefore essential for the survival and proper functioning of neurons.

Due to their unique metabolic properties (namely high iron content), oligodendrocytes are highly sensitive to oxidative stress, but also to excess of neurotransmitters such as glutamate (excitotoxicity). Due to this vulnerability, oligodendrocytes are affected in various diseases. The common causes of oligodendrocyte alteration in the CNS include traumatic brain injury, ischaemia (restricted blood supply) and auto-immune diseases such as Multiple Sclerosis (MS). In MS, auto-reactive cytotoxic T cells target myelin sheaths, inducing oligodendrocytes death and severe cognitive and motor impairments. Some rare genetic diseases have also been identified as causing severe impairments of myelin sheath formation, such as the Pelizaeus–Merzbacher disease, caused by a mutation of the PLP1 gene, a major myelin component. Interestingly, involvement of oligodendrocytes as downstream targets can be observed in multiple psychiatric disorders such as schizophrenia, mood disorders and depression (Fields, 2008).

Over the years, alteration of myelin and white matter in the context of AD has been studied (Nasrabady et al., 2018): significant loss of myelin can be observed in post-mortem brain sections of AD patients while the number of oligodendrocytes is decreased. Accordingly, mouse models have reported similar observations with a decreased number of oligodendrocytes in the triple transgenic mice model of AD (3xTg-AD). However, the exact cellular role played by oligodendrocytes in the pathogenic cascade leading to confirmed AD is not clear: are oligodendrocytes cellular state altered in AD, and if yes how does it affect neurons functions? Interestingly, soluble Aβ has been shown to induce significant toxicity to in-vitro cultured oligodendrocytes by inducing oxidative stress. Therefore, pathogenic soluble Aβ could directly induce oligodendrocyte death and demyelination and thereby contribute to cognitive decline. However, many questions remain unanswered: are oligodendrocytes alterations a cause or a consequence of AD? how do Aβ plaques and NFT act on oligodendrocytes? Therefore, an extensive study of oligodendrocyte contribution to AD pathogenesis might provide critical informations and leads to new therapeutic targets.
3 A shared oligodendrocyte activation state associated with neuroinflammation and neurodegeneration.

In the manuscript contained in this section, we studied how oligodendrocytes transcriptional state was altered in the context AD. Our analysis revealed that similarly to microglia which differentiate into Disease Associated Microglia (DAM), oligodendrocytes can turn into Disease associated Oligodendrocytes (DOLs) and express a unique set of genes. Secondary analysis of previously published datasets revealed that DOLs can be found in multiple mouse models of AD but also in other pathological contexts, such as aging or multiple sclerosis. In-depth analysis suggested that DOLs induction in AD mice is based on a cross-talk between OLs and microglia, and that this cross talk relies on TREM2 in order to activate the microglia in an adequate fashion. As this project was only started at the end of my PhD (third/last year of PhD) and due to the extensive amount of time required to study neurodegenerative diseases, it was not finished at the time my thesis’ writing. The proposed manuscript is therefore not a finalized version and will likely change in the following months, to take into account the results of ongoing experiments.
A shared oligodendrocyte activation state associated with neuroinflammation and neurodegeneration.

Mor Kenigsbuch\textsuperscript{1, 2*}, Pierre Bost\textsuperscript{2, 3,4*}, Renana Hajbi\textsuperscript{5}, Benno Schwikowski\textsuperscript{3}, Elior Peles\textsuperscript{5}, Michal Schwartz\textsuperscript{1,\dag}, Ido Amit\textsuperscript{2,\dag}

\textsuperscript{1}Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel.
\textsuperscript{2}Department of Immunology, Weizmann Institute of Science, Rehovot, Israel.
\textsuperscript{3}Systems Biology Group, Center for Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur CNRS, Paris, France;
\textsuperscript{4}Sorbonne Université, Complexité du vivant, Paris, France.
\textsuperscript{5}Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.

* These authors contributed equally to this work.
\dag These authors jointly supervised this work.

Abstract

More than 45 million people worldwide have Alzheimer’s Disease (AD), a neurodegenerative disease featured by progressive loss of memory and other cognitive domains that leads to death within 3 to 9 years after diagnosis. While recent genetic and single-cell works have shown that microglia, the phagocytic cells of the brain, are involved in the pathogenesis of AD, the contribution of other glial cells remain unclear. To this aim, we performed an in-depth, single-cell RNA-seq analysis of more than 10,000 glial cells isolated from WT and AD mice brain, revealing that oligodendrocytes were the most affected cells. We unveiled an new oligodendrocyte cell state that we called ‘Disease associated OLigodendrocytes’, or DOLs. DOLs could only be found in AD mice brain, and their abundance increase with age. Interestingly, DOLs could be found in multiple mouse models of AD, but also in other neurodegenerative and neuroinflammatory pathological contexts, such as multiple sclerosis and aging. Lastly, DOLs are likely induced through a cross-talk with a specific subset of microglia in a TREM2 dependent manner.
Introduction

Alzheimer’s disease (AD) is the most common age-related neurodegenerative disease and form of dementia\(^1,2\). It is considered as one of the main public health challenges in Western countries due to population aging and estimates are that in 2050, 1 in 85 persons will be living with the disease\(^3\), placing a heavy personal and financial burden on care-givers and health and support systems. AD causes progressive memory decline and cognitive dysfunctions, behavioral and personality changes, resulting in a reduced quality of life and loss of independence of the patients\(^1\). The disease primarily affects the hippocampus and the cerebral cortex, resulting in significant neuronal death and brain volume loss\(^1\). Pathological hallmarks include accumulation of misfolded protein aggregates of Amyloid-Beta (A\(\beta\)), known as plaques, neurofibrillary tangles (NFT) caused by hyper-phosphorylated protein tau\(^3-5\) and extensive neuroinflammation\(^1,6\). Despite intensive research, no cure has been identified and disease management is complex and inefficient\(^1,2\).

While most of the AD research traditionally focused on neuronal damage, recent studies have highlighted the role played by non-neuronal cells in disease. Indeed, key genes in AD pathology identified through large scale Genome Wide Association Studies (GWAS) appeared to be expressed by non-neuronal cells and mainly by the dominant brain innate immune cells, the microglia\(^7\). Further studies based on single-cell RNA sequencing (scRNA-seq) confirmed this hypothesis\(^8\) and identified specific microglia state associated with disease progression (Disease Associated Microglia, DAM).

Compared to the microglia, the contribution to AD pathology of non-immune glial cells, such as astrocytes and oligodendrocytes, remains relatively under explored. Oligodendrocytes (OLs) are the Central Nervous System (CNS) myelin-forming cells and are required for efficient electrochemical signal transmission\(^9\). While OLs have been traditionally viewed to serve this sole purpose, recent research revealed other functions such as lipid metabolism\(^10\), axonal energy metabolism and trophic support\(^11,12\), CNS repair\(^11\) and immunomodulation\(^13\).
In the context of AD, study of OLs has been quite limited. Most of it focused on myelin alteration and breakdown observed in the disease. However, it is still unknown how OLs react to the pathology and its associated environment, and conversely how their phenotype, function, signaling pathways and interactions with other cell types are contributing to the pathology. This is in large part because a thorough mapping of the OLs landscape in AD is still missing.

In this study, we used single cell RNA sequencing (scRNA-seq) to study cellular heterogeneity among non-immune glia in the brain of an AD murine model. We found that the most striking alterations occur in oligodendrocytes and identified a novel oligodendrocyte cell-state that is present in AD but absent from healthy brains, termed Disease-associated Oligodendrocytes (DOLs) and increased with disease progression. In addition, we have identified similar cell states across other neuroinflammatory and neurodegenerative conditions, including experimental autoimmune encephalitis (EAE) and aging. Lastly, we show that differentiation of OLs into DOLs requires TREM2 and might be mediated through a stimulation by Interferon-gamma (IFNγ) and Tumor Necrosis Factor alpha (TNFa), as shown by in-vitro stimulation experiment. Overall, our study uncovers a conserved OLs response to neuroinflammation and neurodegeneration, and sheds light on possible OLs involvement in AD pathology.

Results:

1. Oligodendrocytes display striking transcriptional alterations in AD

To gain new insights about cellular heterogeneity within the non-immune compartment of the AD brain, we performed Massively parallel single cell RNA sequencing method (MARS-seq 2.0) on CD45- cells from whole brains of Wild-Type (WT) and 5xFAD AD mice, an AD mouse model carrying 5 familial AD mutations (see Methods). Following strict quality controls (Figure S1a-c) 10,690 cells passed quality control and no significant differences were observed between control and AD mouse libraries (Figure S1d). Cells were clustered based on their expression profile using the Pagoda2 pipeline and 17 clusters were detected and annotated.
based on established cell-type specific marker genes (Figure 1a-c). The major populations we identified were OLs (Plp1, Mbp), astrocytes (Slc1a2, Slc1a3), pericytes (Myl9, Vtm, Rgs5), endothelial cells (Ly6a, Ly6e), ependymal cells (Tmem212, Ccdc153) and Choroid plexus (CP) cells (Ttr, Enpp2), along other small populations (olfactory ensheathing cells (OEC), fibroblasts, GABA+ neurons). A small contamination by microglia (Hexb, Cx3cr1) and red blood cells (RBC, Hbb-bs, Hbb-bt) were also detected. Cell type composition was largely identical between AD and WT brains (Figure S1e), and we did not notice any significant change in abundance of any of the cell types, suggesting that AD does not drastically affect non-immune brain cell composition.

Next, we investigated which cellular population displays the deepest transcriptional alterations in AD by performing differential gene expression analysis between cells originating from AD or WT brains. In order to precisely and robustly quantify the number of Differentially Expressed genes (DEG), we used a binomial regression-based approach that we previously described\textsuperscript{21}. Strikingly, OLs were the most altered population with 98 DEG while other populations exhibited limited transcriptional changes (Figure 1d). Out of these DEGs, more than 75% were specifically altered in OLs (Figure 1e), highlighting an OLs-specific response that is not shared with other cell types. To check that this was not solely due to the number of OLs, we performed a power analysis and found that even by sampling a limited number of OLs, the number of DEG was still higher compared to other cell types (Figure 1f).

Altogether, these results suggest that OLs are the most impacted cells by AD within the non-immune non neuronal cell compartment.

2. Identification of a unique Oligodendrocyte state associated with Alzheimer’s Disease, DOLs.

Intrigued by the transcriptional changes in OLs in AD context, we next sought to elucidate whether these DEG reflect a change across all the OLs or whether they originate from a specific sub-population with an altered cellular state. To this end, we enriched for OLs among sequenced cells in order to achieve deeper profiling of the transcriptional changes. This was
carried out using the OL-specific marker Galactosylceramidase (GalC)\textsuperscript{22,23}, and sorting of GalC+ cells from WT and AD brains ranging from 6 to 24 months of age (Figure 2a and S2a). GalC+ sorted plates contained between 48-86\% OLs (Figure S2b) and the obtained cells were filtered and computationally pooled with the previously sequenced OLs, resulting in 6194 high quality sequenced OLs sampled from WT and AD mice at various time points (Figure S2c). Refined clustering analysis of these OLs revealed 9 different sub-clusters, each one characterized by a unique expression profile (Figure 2b-c). A small contamination of dying cells (2.1\% of total OLs) was identified by elevated expression of the established stress markers \textit{Jun} and \textit{Egr1} (Figure S2d) and was removed of downstream analysis.

To check whether specific clusters were associated with the disease, we compared for each cluster its proportion among all OLs between AD and WT. While six clusters displayed similar proportions in AD and WT, this analysis revealed a strong enrichment of clusters 2 and 5 in AD (Figure 2d). Cluster 5 contained around 13\% of all OLs in AD, but only around 2\% in WT, from all ages. This cluster was also distinctly separated from the rest of the OLs in the Uniform Manifold Approximation and Projection (UMAP) (Figure 2c). Because of its strong association with the disease, we termed this transcriptional state “Disease associated OLigodendrocytes” (DOLs), and the unchanging populations collectively as “homeostatic OLs”. Cluster 2 was less abundant as it represented only 6\% of all AD OLs and less than 1\% of WT OLs. Interestingly, cluster 2 was mostly found in the two 24 months old AD mice, in which they represent between 9\% - 30\% of the OLs (Figure S2e). As this population was only found in the terminal phase of the disease and might result from interaction between AD and aging, we decided to focus on the DOLs which were present in earlier stages of the disease.

To examine whether DOLs emergence correlates with disease progression, we sorted and sequenced OLs (GalC+) from AD and age-matched WT mice in the ages of 6 (n=3(AD), 3(WT)), 8 (n=1, 1), 10 (n=6, 4), 15 (n=2, 2) and 24 months (n=2,1). Age-matched WT were used as control for age-dependent effects unrelated to the disease (Figure 2d). The proportion of DOLs among OLs increases with disease progression over time, ranging from 5\% for 6-month old mice.
to 20% for two-year old mice. In parallel, the proportion of DOLs in healthy mice never exceed 5% even in aged mice (Fig. 2d).

To identify the underlying biological process associated with the DOL population, we performed a Differential Expression (DE) analysis between DOLs and homeostatic OLs. This analysis identified 22 genes over-expressed by DOLs (p-value <0.01 and log2FC>1) that we called the DOL signature. This transcriptional signature is composed of upregulation of both immune and non-immune related genes. Upregulated immune related genes include complement component C4b, major histocompatibility complex I (MHC-I) genes (H2-D1, H2-K1, B2m) and the cytokine II-33. Several non-immune related genes were also identified, including genes with a well-established association with neuronal inflammation such as Serpina3n\textsuperscript{24} and Klk6\textsuperscript{25,26} (Figure 2e).

Lastly, to investigate the signaling pathways possibly controlling the induction of these genes, we performed motif enrichment analysis using iRegulon\textsuperscript{27}. Analysis results pointed at three major transcription factors families: the Irf/Stat factors, YY1/NF-kB pathway, and Sox9 transcription factor (Figure 2f). MHC-I genes and C4b were inferred to be regulated through Irf/Stat and YY1/NF-kB binding motifs while non-immune genes (Klk6, Sgk1) were associated to the Sox9 transcription factor (Figure 2g), a key transcription factor in oligodendrocytes.

These results indicate that the transcriptional changes observed in the OLs population in AD are a result of one AD-specific cellular state, DOLs, and not as a result of widespread state shift across the entire OLs population. DOLs express both immune and non-immune genes, likely induced by a limited sets of transcription factors including members of the NF-kB and IRF/STAT family. On average, 10% of the OLs population in AD are DOLs but they become more prominent with disease progression.

3. **DOLs are induced in multiple AD model and require a Trem2-based oligodendrocyte/microglia cross-talk.**

To check that DOLs were specific to our AD model, we decided to re-analyze a recently released dataset (GEO number: GSE153895) where hippocampal glial cells from various strains of 24th
month old mice were sequenced using 10X single-cell sequencing technology. WT, P301L, PS2/APP/P301L and PS2/APP/P301L/TREM2−/− strains were used, allowing to compare the effects of a pure tau-based AD model and of a mixed tau/Aβ AD model on OLs, but also to identify the contribution of the Trem2 gene (Figure 3a). Analysis of the data allowed to establish a comprehensive single-cell atlas of the glial cells (Figure 3b) that include more than 30,000 OLs and 13,000 microglial cells.

A large variety of OL states was observed as 9 clusters corresponding to OLs were identified (Figure S3a). Among them, we found a cluster characterized by the over-expression of DOL specific genes such as Serpina3n, C4b, Klk6, B2m and H2-D1 (Figure 3c). This ‘activated’ cluster was enriched in P301L and PS2/APP/P301L mice compared to WT mice (Figure 3d), supporting our hypothesis that DOLs can be found in various AD mouse models. Interestingly, DOLs were less frequent in PS2/APP/P301L/TREM2−/− mice (Figure 3d), suggesting that TREM2 is required for the induction of DOLs.

As TREM2 is mostly expressed by microglia28 (Figure S3b), we performed a second round of analysis to identify potential microglia population that were associated with the DOLs. We first looked at the microglia cluster corresponding to the so-called Disease Associated Microglia8 characterized by the high expression of Trem2, Cst7 and Tyrobp: this cluster could only be found in the brain of PS2/APP/P301L mice but not in the P301L and PS2/APP/P301L/TREM2−/− mice (Figure 3e). This is in agreement with previous observations showing that DAM induction is TREM2 dependent and cannot observed in tau-based AD mouse models. As the induction patterns of DAM and DOLs dramatically differ, we conclude that DOLs induction is Trem2 dependent but DAM independent.

We therefore screened for microglia population which abundance was correlated with DOL abundance. We identified a cluster of microglia strongly associated to DOL (R=0.94, Figure S3c and d) that was logically induced in both P301L and PS2/APP/P301L mice but not in the WT and PS2/APP/P301L/TREM2−/− mice (Figure 3f). This cluster was featured by the expression of several immune receptors such as Galectin-3 (Lgals3), Dectin-1 (Clec7a), Axl and Lilrb4a (Figure 3d), all of them being negative regulators of immunity29–31. Gene Set Enrichment
Analysis revealed that genes associated with oxidative phosphorylation were significantly upregulated (Electron Transport Chain and Oxidative Phosphorylation Wikipathway gene sets, Normalized Enrichment Score of 2.5 and and 1.8 respectively, p-value < 1.e-4) suggesting that those cells undergo a metabolic shift. Recent studies have shown that an oxidative metabolism is required for an efficient Aβ aggregate and apoptotic body clearance, suggesting that this ‘alternative’ microglia cluster corresponds to a highly-phagocytic and anti-inflammatory microglia subset.

Altogether, our analysis reveals that DOLs can be found in the brain of multiple AD mouse models and suggests that DOL induction likely relies on a microglia/OLs cross-talk in a TREM2 dependent manner.

4. Meta-analysis reveals OLs transcriptional states similar to DOLs in other neuroinflammatory and neurodegenerative conditions

As we observed DOLs in multiple mouse models of AD, we speculated they might be present in other neurodegenerative and neuroinflammatory conditions. To test this hypothesis, we conducted meta-analysis on three previously published mouse scRNA-seq datasets (Figure 4a). The first dataset was generated from the peak stage of Experimental Autoimmune Encephalomyelitis (EAE) mice spinal cord\textsuperscript{32}, modeling the acute phase of Multiple Sclerosis (MS). MS and EAE feature massive infiltration of immune cells into the CNS and cytotoxic attack of the OLs, resulting in their death and leading to de-myelination. The second dataset, also from EAE spinal cord, included data from the different stages of relapsing-remitting EAE: naive, priming, peak and remission phases, along with control injected with Complete Freund Adjuvant (CFA)\textsuperscript{33}. Lastly, the third dataset was generated from the subventricular zone (SVZ) of young (3 months old) and old mice (28-29 months old)\textsuperscript{34}. Aging by itself features neuroinflammation and neurodegeneration, albeit to a much smaller extent than AD\textsuperscript{35,36}.

We re-analyzed those datasets using the Pagoda2 pipeline\textsuperscript{20}. As described in their respective papers, all datasets were rich in OLs (1207, 2552 and 4683 OLs identified respectively), making
them suited for our analysis. We first computationally gated OLs, and then compared mean expression of the genes between pathogenic and non-pathogenic samples (Figure 4a). As expected, we found pathology associated genes in OLs in all three datasets (Figure 4b-d): DOL genes (marked in red) were among the top upregulated genes in all three datasets and gene set enrichment analysis (GSEA) revealed a significant enrichment in each one of them (Figure S4a-c). Interestingly, not all DOL genes were induced in each model, and those induced vary across datasets: for instance, $C4b$ is induced in acute EAE and aging model but not in relapsing-remitting EAE, while MHC-I genes are over-expressed in both EAE datasets but not in aging OLs.

To better characterize the contribution of DOL genes in those various pathological situations, we used a Principal Component Analysis (PCA)-based approach where a PCA of the OLs expression matrix is computed using only the DOLs genes\(^{37}\), thus allowing to check if a limited set of genes is able to distinguish cells in an un-supervised way. PCA of the acute EAE OLs was dominated by a first Principal Component (PC) (Figure S4d), that strongly distinguished OLs from healthy and sick mice (Figure 4e). Similarly, the first PC score of the relapsing-remitting EAE increases over time and peaked in the acute phase before decreasing during the remission phase (Figure 4f, S4e), suggesting that the DOL transcriptional program activity correlates with the intensity of auto-immune response directed against OLs. Interestingly, unlike in the acute EAE dataset, a second PC could also be considered as significant in this dataset but was not associated with the course of the disease (Figure S4g), suggesting that the DOL program could be split into two gene modules. Lastly, in the aging SVZ data (Figure S4f) we observed two significant PCs. While the first PC was not able to discriminate between young and old OLs, the second PC could (Figure 4g, S4h).

To understand the biological significance of these two different PCs, we looked at the correlation of the DOL gene expression and the two different PCs. In the relapsing-remitting EAE model we observed that the first PC was driven by the expression of MHC-I genes and other immune genes such as $Serpina3n$, and $C4b$, while PC2 correlated with $Il33$, $Cd9$, $Ptma$, $Gstp1$ and $Opalin$ expression (Figure 4h). On the other hand, those genes correlated with the PC1 of the aging SVZ
dataset while MHC-I and other immune genes correlated with PC2. We therefore conclude that the DOL transcriptional program can be split into two distinct modules: an immune-related and a non-immune program, and each pathology is dominated by one of the two.

Our meta-analysis therefore suggests that OLs cellular states similar to DOLs can be found in multiple pathologies where neuroinflammation and neurodegeneration are involved. In two different pathologies, the DOL signature could be split into an immune-related and a non-immune signature, with the immune program activity correlating with disease severity.

5. In-vitro screening for ligands able to induce DOLs

In order to identify ligands that can induce the differentiation of homeostatic OLs into DOLs, we performed in-vitro culture of primary mouse OLs and stimulated them by various ligands before sequencing the cell RNA using a shallow bulk sequencing approach (Figure 5a, Methods). As our promoter analysis revealed a significant enrichment of NF-kB and IRF/STAT promoters, we therefore decided to stimulate the OLs with cytokines known to activate those pathways, i.e Interferon-gamma (IFNg), Interferon-beta (IFNb), Tumor Necrosis Factor-alpha (TNFa) and Interleukin-1 beta (IL1b), and looked at their effects at various times post stimulation. In addition, we also used ligands found in AD brain, namely pathogenic Aβ under oligomers, fibrils and plaques forms, as well as blebs from apoptotic neurons. As IFNg was the most likely ligand due to its ability to induce MHC expression, three different biological replicates were performed.

Differential analysis was performed and the mean log2 fold change (log2FC) induced by each ligand was computed (Methods). Clustering of those ligand-induced signatures shows that IFNb and IFNg induce similar transcriptional changes in OLs while TNFa and IL1b signatures clustered together (Figure 5b). Lastly, signatures induced by apoptotic neurons and Aβ compounds clustered separately from cytokines (Figure 5b).

Unfortunately, no single ligand was able to induce the DOL genes in-vitro: while IFNg was able to induce the expression of the MHC genes and of some immune genes (C4b and Ifi27), it was not able to prompt the expression key DOL genes such as C63, Cd9, Serpina3n and Klk6. TNFa-
stimulated samples exhibit a higher expression of enzymes members of the DOL signature (Gstp1, Ctsb, Gpd1) and of Serpina3n but not of MHC-I genes.

Altogether our analysis suggests that OLs can not be turned into DOLs by a single ligand, but instead by a cocktail of cytokines, likely TNF and IFNγ. Other factors, likely secreted by the alternative microglia cluster mentioned above could also play a role and contribute to the emergence of DOLs in AD and other neurodegenerative diseases.

**Conclusion/Discussion**

The exact contribution to AD pathogenesis of the various glial cells (microglia, astrocytes, OLs) is still controversial: if alterations of each glial cell type could be observed in old and AD mice (ref paper Aleks, ref Disease-associated astrocytes in Alzheimer’s disease and aging, ref paper Colona), it is unknown if those changes are secondary or causative events of AD.

In this paper we described a novel transcriptional state of OLs that we called DOLs. This subset could be observed in all AD mouse models that we studied, but also in EAE and aging mice suggesting that DOLs represent a generic response of OLs to various stress, including neuronal death and neuro-inflammation. Our analysis of the TREM2 deficient AD mice revealed an important feature of the DOL: DOLs induction requires TREM2, a key membrane receptor of microglia involved in the phagocytosis of protein aggregates and apoptotic neurons. To our surprise, DOLs induction was however not dependent on DAM, a subset of activated microglia found in multiple neurodegenerative diseases, but instead seems to rely on an other subset of microglia that we named ‘alternative microglia’.

This subset is rarer and seems to have been less well described than DAM, but our analysis suggests that it is featured by the high expression of negative immune receptors (Galectin-3, Dectin-1, Axl) and metabolism based on oxidative phosphorylation instead of glycolysis.
Additional work is needed to better characterize this subset, but also to investigate the possible cross-talk between it and OLs.

The exact role of DOLs itself has also to be elucidated: additional experiments will have to be performed in order to check the effects of over- or loss of expression of DOL genes, such as Serpina3n, C4b or Il33. Lastly, a more functional characterization of the DOLs should be performed: OLs are featured by the production of myelin sheath, which structure can significantly be altered in neurodegenerative pathologies and can not be assessed by transcriptomic assays. Physical isolation of DOLs coupled with lipidomic or transmission electron microscopy could provide such informations and detail the role of DOLs in AD.

Overall, our study expands the biological insight on the contribution of glial cells to AD, and shed light upon the transcriptional changes of OLs in multiple neurodegenerative diseases and the possible cross-talk between microglia and OLs.

**Methods:**

**Mice**

Heterozygous 5XFAD transgenic mice (on a C57/BL6-SJL background) that overexpress familial AD mutant forms of human APP (the Swedish mutation, K670N/M671L; the Florida mutation, I716V; and the London mutation, V717I) and PS1 (M146L/L286V) transgenes under the transcriptional control of the neuron-specific mouse Thy-1 promoter\(^\text{19}\) (5XFAD line Tg6799; The Jackson Laboratory), were taken throughout adulthood in different time points as indicated in the text. Genotyping was performed by PCR analysis of tail DNA. Throughout the study, WT controls in each experiment were non-transgene littermates from the relevant tested mouse colonies. Mice were bred and maintained by the animal breeding center of the Weizmann Institute of Science. All experiments detailed herein complied with the regulations formulated by the Institutional Animal Care and Use Committee (IACUC) of the Weizmann Institute of Science.

**Brain dissociation for single cell suspension**
Mice were euthanized using an overdose of ketamine-xylazine, followed by transcardial perfusion with cold PBS and whole brains were harvested. Tissue was chopped to small pieces and subjected to 30 minutes of enzymatic digestion using papain in 37°C, followed by manual trituration using 5ml pipette and filtering through a 70µm cell strainer. Cells were pelleted at 800g for 5 min in 4°C and then suspended in ovomucoid protease inhibitor solution to stop papain activity. Cells were pelleted again in 600g, 5 min, 4°C, washed and myelin debris was removed by 30% percoll gradient. Then, cells were pelleted and washed again and subjected to 20 min Fc block (1:200), following 30 min cell surface staining.

Immediately before FACS reading, cells were washed, filtered through 70µm cell strainer, suspended in sorting buffer (PBS supplemented with 0.2mM EDTA pH8 and 0.5% BSA), and DAPI was added. note: DAPI was not used on fixed cells.

Samples were kept on ice at all times except for enzymatic digestion.

**Single cell sorting**

Cell populations were sorted using FACS-AriaIII (BD Biosciences, San Jose, CA). For the sorting of all the non-immune populations, samples were gated for CD45- (17-0451-82, eBioscience or 103115, biolegend, 1:200), while excluding debris (FSC-A vs SSC-A), dead cells (DAPI+) and doublets (FSC-A vs FSC-H). In order to enrich for different cell populations, the following markers were used: GalC (oligodendrocytes, FCMAB312F, Milli-Mark, 1:10), Ly6A/E (108133, Biolegend, 1:150). Isolated cells were single cell sorted into 384-well cell capture plates containing 2 mL of lysis solution and barcoded poly(T) reverse-transcription (RT) primers for single-cell RNA-seq\(^{18}\). Four empty wells were kept in each 384-well plate as a no-cell control during data analysis. Immediately after sorting, each plate was spun down to ensure cell immersion into the lysis solution, snap frozen on dry ice, and stored at –80°C until processed. To record marker level of each single cell, the FACS Diva “index sorting” function was activated during single cell sorting. Following the sequencing and analysis of the single cells, each surface marker was linked to the genome-wide expression profile. This methodology was used to optimize the gating strategy.
Single cell RNA library preparation by Massively parallel Single Cell RNA sequencing (MARS-seq)

Single-cell libraries were prepared using MARS-seq 2.0 protocol\textsuperscript{18}. In brief, mRNA from cell sorted into cell capture plates are barcoded, converted into cDNA, and pooled using an automated pipeline. The pooled sample is then linearly amplified by T7 in vitro transcription, and the resulting RNA is fragmented and converted into a sequencing-ready library by tagging the samples with pool barcodes and Illumina sequences during ligation, RT, and PCR. Each pool of cells was tested for library quality as described in MARS-seq 2.0 protocol.

Low level processing of the scRNA-seq data

MARS-seq libraries, pooled at equimolar concentrations, were sequenced using an Illumina NextSeq 500 or NovSeq 6000 sequencer, at a sequencing depth of 20K-50K reads per cell. Reads are condensed into original molecules by counting same unique molecular tags (UMI). We used statistics on empty-well spurious UMI detection to ensure that the batches we used for analysis showed a low level of cross-single-cell contamination (less than 3\%). Mapping of reads was done using HISAT (version 0.1.38\textsuperscript{(15)}; reads with multiple mapping positions were excluded. Reads were associated with genes if they were mapped to an exon, using the UCSC genome browser for reference. Exons of different genes that shared genomic position on the same strand were considered a single gene with a concatenated gene symbol. Cells with less than 500 UMIs were discarded from the analysis. All downstream analysis was performed in R.

ScRNA-seq data processing and clustering

ScRNA-seq expression data analysis were per-formed using the R-based Pagoda2 pipeline (https://github.com/hms-dbmi/pagoda2/) in addition to an in-house R script. Low quality cells were removed using the following strategy: cell with less than 350 UMIs and more than 30\% of mitochondrial genes were removed. The number of Highly Variables Genes (HVGs) was determined using the adjustVariance() function with the gam parameter set to 5. HVGs were selected using the following strategy: for each gene, its number of zeros and its mean expression are computed. A local polynomial model is then used to predict the number of zeros according to
the log mean expression (loess function with degree parameter set to 2). The residuals of this model (excess of zeros) are then used to rank the genes and the genes with the highest excess of the zeros are considered as the most HVGs. PCA reduction is then computed using the calculatePeaReduction() function. The number of computed PC was changed in each analysis due to variable number of cells and cellular heterogeneity. A K-nearest neighbor graph was then build with the function makeKnnGraph() with the K value set to 30 and the distance parameter set to ‘cosine’. Clusters were computed using the Louvain’s clustering approach implemented in the getKnnClusters() and multilevel.community() functions. Marker genes were identified using the getDifferentialGenes() function. UMAP low dimensional embedding was com-puted using the uwot R package, and more precisely the umap() function with the n_neighbors parameter set to 30, and the metric parameter set to ‘cosine’. In order to group clusters of cells in the first round of analysis, mean gene expression of the most variable genes was computed using the aggregate() function. Spearman’s correlation matrix was then computed using the cor() function with the method parameter set to ‘Spearman’. Hierarchical clustering was then performed on this matrix using Ward’s method and the resulting tree used to aggregate the cell clusters.

**Single cell differential gene expression analysis and power analysis**

To perform differential expression analysis, we used the dichotomized-based approach previously described in ‘Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients’\(^1\). Briefly, we first dichotomized gene expression (if the normalized expression is bigger than 0 the gene is considered as expressed) and then computed a binomial Generalized Linear Model (GLM) with a complementary log log link function (cloglog) using the glm() R function. To mitigate the variation of the library size as well as the global effect of the infection (bystander effect), we include both variables in the regression model. The corresponding p-values are then computed using a Likelihood Ratio Test (LRT) and then corrected using Benjamini Hochberg correction.

Power analysis was performed by randomly sampling 15 times a given proportion of the oligodendrocytes (80%, 70%, 60%, 50%, 20%, 10% and 5%) and then performing the dichotomized gene expression analysis described above.
**Promoter motif analysis**

Promoter analysis was performed using iRegulon\textsuperscript{27} (Cytoscape plugin, version1.3). The ‘Species and gene nomenclature’ parameter was set to ‘Mus musculus, MGI symbols’ and only motifs 500 bp upstream of the Transcription Start Sites (TSS) were used. All other parameters were set to default values.

**Gene Set Enrichment Analysis (GSEA)**

GSEA were performed using the liger (version 1.12) and the gskb (version 1.16)\textsuperscript{39} packages. Briefly, a gene set describing a list of biological pathways was loaded (mm_pathway object) from the gskb library and then used to analyze the log2FC value list using the bulk.gsea() function from liger.

**Meta-analysis of previously published scRNA-seq**

The three two EAE datasets used in the meta-analysis were downloaded from Gene Expression Omnibus (GEO) server while annotations of the samples was done using the getGEO() function from the GEOquery package (version 2.48). The first acute and relapsing-remitting EAE had the respective GEO accession number: GSE113973 and GSE118257. The aging SVZ 10X scRNA-seq data were kindly sent by the authors. Expression data were processed using the same approach as described above. A cell cluster was considered as corresponding to OLs if it specifically expressed a known OL marker such as Plp1 or Mbp. PCA based analysis were performed using the PCA() function from the FactoMineR package (version 1.41). All parameters of the PCA() function were set to default, except the parameter ‘scale’ which was set to ‘TRUE’. The DOL score is then simply the coordinate of each cell in the PC of interest.

**Analysis of bulk RNA-seq data**

Bulk sequencing data were analyzed using DESeq2 (version 1.24)\textsuperscript{40} and apeglm (version 1.60)\textsuperscript{41} packages. Samples with less than 2*10\textsuperscript{5} UMIs sequenced were removed from analysis, as well as
genes with less than 50 total UMIs. A DESeq2 object was created using the dds() function. The underlying statistical model was fit using the DESeq() function with parameter ‘fitType’ set to ‘parametric’. Due to the low number of technical replicates used in each experiment (3 samples per condition), the shrinked log2 fold changes were computed using the lfcShrink function with the parameter ‘type’ set to ‘apeglm’.

Correlation and expression heatmaps were drawn using the pheatmap() function from the pheatmap package (version 1.0.12). The clustering based on the shrink log2FC correlation matrix (Figure 5b) was used in the Log2FC heatmap of the DOL genes (Figure 5c).

**Immunofluorescence**

For immunofluorescence, mice were euthanized, intracardially perfused with PBS and brains were extracted and fixed in paraformaldehyde 2.5% over-night, then washed with PBS and immersed in 30% sucrose until sinking. 30µm free floating brain sections were cut using sliding microtome (Leica) and quenched using 100mM NH₄Cl for 1 hour at RT, blocked in PBS with 10% donkey serum and 0.5% Triton and then stained using the following primary antibodies: Olig2 (AB9610, Merck, 1:200), Serpina3n (AF4709, R&D Systems, 1:200), CNPase (MAB1580, Merck, 1:200), hAβ (BLG-803001, BioLegend, 1:200) in PBS with 5% donkey serum and 0.1% Triton. Secondary antibodies used are donkey anti-rabbit cy2 (Jackson ImmunoResearch #711-225-152), donkey anti-goat cy3 (Jackson ImmunoResearch #705-165-147) and donkey anti-mouse cy5 (Jackson ImmunoResearch #715-175-151). A confocal microscope (LSM 880) and Zeiss ZEN software were used for image capturing.

For DOLs-plaques association, 3 coronal sections were taken from each mouse in 3 different depths along the rostral-caudal axis. A 4x5 tile image with 20 Z-planes was imaged using 20x lens.

**Primary oligodendrocyte cultures**

Primary oligodendrocyte-precursor cells (OPC) were cultured as follows: P0-3 mouse pups were decapitated, cortices was extracted and put in DMEM+penicillin/streptomycin media on ice. Then, cortices were homogenized by two triturations with 19G needle, followed by 2 triturations
with 21G needle. Then, the suspension was centrifuged (200g, 5 min, room temp), suspended in Glia medium (DMEM, 10% heat-inactivated fetal bovine serum, 5% heat-inactivated horse serum, penicillin/streptomycin) and seeded in flasks. Glia medium was replaced every 3 days, and from the 6th day insulin (5ul/ml, Sigma l6634) was supplied. After 10 days, flasks were gently knocked and media was removed to remove microglia, followed by shaking over-night in 250RPM at 37°C. the following day, OPCs were purified from supernatant using MACS separation O4 beads and seeded over poly-D-lysine, poly-L-ornithine pre-coated plates. Media was replaced every 2 days.

**Culture stimulation, bulk RNA purification and library preparation**

Differentiated oligodendrocytes were stimulated after 7-8 days in culture. For cytokine stimulation, cells were added the cytokine at 100U/ml (IFNγ: R&D systems 485-M1-100, IL-1β: R&D systems 401-ML-10, TNFα: R&D systems 410-MT-25, IFNβ: pbl 12405-1) and collected for RNA-seq after 6, 12 and 24h. For Aβ stimulation, different aggregation states were prepared from HFIP-treated Aβ 1-42 (Bachem 4090148) as previously described42. Scrambled Aβ were used as control (Bachem 4064853). Cells were treated with aggregated Aβ at 20nM concentration for 24h. For apoptotic neurons stimulation, primary dorsal root ganglia (DRG) neurons were subjected to 4h of doxorubicin treatment, followed by media removal, wash, scraping of the cells, centrifugation, supernatant removal and suspension of apoptotic cell bodies in media. Cells were treated with apoptotic neurons for 24h.

Cells were collected and mRNA was purified using Dynabeads mRNA purification kit (Invitrogen). Libraries were prepared using a modified MARS-seq protocol for bulk.

Legends

Figure 1 : Identification of oligodendrocytes as the most strongly impacted cell population in 5xFAD mouse model. (a) Expression heatmap of glial cells from 5xFAD and WT mice. (b) Violin plot of known marker genes across the different cell clusters. (c) Two-dimensional UMAP embedding of the scRNA-seq data. Dots are colored based on the scRNA-seq clustering. (d) Number of DE genes between 5xFAD and WT mice across the major cell types. (e) Specificity of the DE genes in OLs. (f) Power analysis : number of DE genes based on the number of OLs used for the DE analysis.

Figure 2 : DOLs are a unique type of OL specific to AD mouse brain that accumulate during the disease. (a) Experimental strategy to study OLs. (b) Expression heatmap of OLs from 5xFAD and WT mice at various age. (c) Two-dimensional UMAP embedding of the scRNA-seq OLs data. Dots are colored based on the scRNA-seq clustering. (d) Comparison of the mean proportion of the different OL clusters between AD and WT mice. (e) Variation in the proportion of DOLs among OLs across time in WT (upper panel) and AD (lower panel) mice. (f) Volcano plot corresponding to the DE analysis between DOLs and other OLs. (g) Results of the promoter analysis by iRegulon.

Figure 3 : DOLs are induced in multiple AD model and require a Trem2-based oligodendrocyte/microglia cross-talk. (a) Underlying experimental design of the GSE153895 dataset. (b) Two-dimensional UMAP embedding of the scRNA-seq data. Dots are colored based on the scRNA-seq clustering. (c) Comparison of homeostatic and activated OLs transcriptomes. (d) Proportion of DOLs among OLs across mouse strains. (e) Proportion of DAM among microglia across mouse strains. (f) Proportion of alternative microglia among total microglia across mouse strains. (g) Comparison of homeostatic and activated microglia transcriptomes.

Figure 4 : OLs transcriptional states similar to DOLs are found in other neuroinflammatory and neurodegenerative conditions. (a) Analytical approach used to identify DOLs in other diseases. (b) Comparison of OLs transcriptome from control and EAE mice spinal cord. (c) Comparison of OLs transcriptome from control (CFA) and acute EAE mice spinal cord. (d) Comparison of OLs transcriptome from young and old mice SVZ. (e) Distribution of the DOL PC 1 score in control and EAE mice. (f) Distribution of the DOL PC 1 score in control (CFA) and at various stages of EAE. (g) Distribution of the DOL PC 2 score in young and old mice. (h) Pearson correlation of DOL gene expression level and relapse-remitting first and second PCs. (i) Pearson correlation of DOL gene expression level and aging SVZ first and second PCs.

Figure 5 : In-vitro screening for ligands able to induce DOLs. (a) Experimental set up used to study the effects of possible DOLs inducer. (b) Pearson correlation between the transcriptional changes induced by the different ligands used. (c) Log2FC of the DOL genes induced by the tested ligands.

Figure S1 associated to Figure 1. (a) Distribution of total cellular UMIs. (b) Distribution of total gene UMIs. (c) Distribution of the proportion of MT genes. (d) Comparison of the total cellular UMI distribution in cells from AD and WT mice (Kruskall-Wallis rank test). (e) Comparison of the proportion of MT genes in cells from AD and WT mice (Kruskall-Wallis rank test). (f) Comparison of the cluster proportion between AD and WT mice.
Figure S2 associated to Figure 2. (a) Gating strategy used to enrich for OLs. (b) Proportion of OLs isolated in each plate sequenced. (c) Number of OLs sequenced for each AD and WT mice. (d) Number of OLs sequenced for each mice based on their age. (e) Expression of Fos and Egr1 across the OL clusters. (f) Proportion of cluster 2 OLs in AD mice across time.

Figure S3 associated to Figure 3. (a) Heatmap showing Spearman correlation between the mean-expression of the 500 most variable genes in each cluster. (b) Expression of Trem2 gene across clusters. (c) Ranked Pearson correlation between DOLs and microglia cluster proportion. (d) Association between the proportion of alternative microglia and DOLs.

Figure S4 associated to Figure 4. (a) GSEA analysis plot corresponding to the EAE acute dataset. (b) GSEA analysis plot corresponding to the EAE relapse-remitting dataset. (c) GSEA analysis plot corresponding to the aging SVZ dataset. (d) Scree plot of the DOL PCA analysis in the EAE acute dataset. (e) Scree plot of the DOL PCA analysis in the EAE relapsing-remitting dataset. (f) Scree plot of the DOL PCA analysis in the aging SVZ dataset. (g) Distribution of the DOL PC 2 score in control (CFA) and at various stages of EAE. (h) Distribution of the DOL PC 1 score in young and old mice.
Figure 5, Kenigsbuch et al.

(a) Collection of OPCs → In-vitro differentiation and culture → Interferons (IFN-β, IFN-γ) Cytokines (TNFα) Apoptotic neurons

(b) Shallow RNA-sequencing

(c) Heatmap showing gene expression levels with color-coded intensity from green (low) to red (high)
Figure S1, Kenigsbuch et al.

(a) Histogram of total cellular UMIs (Log10).
(b) Histogram of total gene UMIs (Log10).
(c) Histogram of Proportion of MT genes (%).

(d) Boxplot of total cellular UMIs (Log10) for AD and WT.
(p = 0.267)

(e) Scatter plot of cell type proportions in AD mice vs. WT mice.

(p = 0.019)
4 Ongoing work and future experiments to unlock DOLs’ secrets

In this manuscript we discovered a new transcriptional state of oligodendrocytes characterized by the over-expression of both immune (MHC-I genes, IL33, SERPINA3N, C4B) and non immune genes (KLK6, CD9, CD63), that we named DOLs for ‘Disease-associated Oligodendrocytes’. We initially observed DOLs in an Aβ-based mouse model of AD, but we could also find them in other AD mouse models and in other pathological contexts such as Experimental Autoimmune Encephalomyelitis (a MS mouse model) and aging. Induction of DOLs seems to rely on an oligodendrocytes/microglia cross-talk, as the abundance of a specific subset of microglia, distinct from DAM, strongly correlates with DOLs abundance. While the features of this so-called ‘alternative’ microglia cluster are not clear, their induction seems to rely on TREM2 gene, similarly to DAM, suggesting an activation by AD ligands such as protein aggregates or apoptotic blebs.

We are currently exploring several paths in order to better understand how DOLs arise and what might be their contribution to AD pathogenesis. In addition, how they interact with microglia has yet to be determined. First, we performed extensive imaging of AD mice brain and simultaneously stained for Aβ plaques, homeostatic OLs and DOLs. Using these data, we are trying to know whether DOLs are spatially associated with plaques, as plaques are considered as local inflammatory niches (Figure IV.5, upper left panel). In parallel, we are trying to check that the ‘alternative’ microglia cluster we found in one dataset could also also be found in previously published datasets, like in the original DAM paper (Keren-Shaul et al., 2017). Additional experiments to better characterize those cells in-vivo are also required (including metabolic characterizations), and immuno-fluorescence imaging could be used to validate the existence of this sub-population (Figure IV.5, upper right panel). Lastly, we are trying to perform Physically Interacting Cells Sequencing (PIC-seq) of oligodendrocytes/immune cell doublets (Giladi et al., 2020): indeed, during our experiments we observed a significant amount of oligodendrocytes/immune cells (GALC+/CD45+) doublets. In addition, some of the DOL genes can be induced in-vitro by inflammatory cytokines. Therefore, PIC-seq represents an opportunity to study if interactions with immune cells can result into DOLs induction (IV.5, bottom panel).

All the aforementioned approaches will hopefully provide information about the mechanisms of DOL induction, but they will not explain what is the exact role of DOLs in AD pathogenesis. Indeed, we do not know if DOLs have a positive or a negative effect on the cognitive abilities of the AD mice: if DOLs are featured by the expression of known neuro-protective genes such as IL33 (?) and SERPINA3N (Vicuña et al., 2015), they also express genes with an unclear contribution to AD such as C4B and the MHC-I genes. We therefore considered two different strategies to study the contribution of DOLs to AD pathogenesis in-vivo: the
Contribution of oligodendrocytes to Alzheimer’s disease

Figure IV.5: Current working directions of the DOL project. (A) We are investigating the possible spatial association between DOLs and Aβ plaques through IF imaging of mouse and human AD brains. (B) Existence of the so-called ‘alternative’ microglia is currently investigated and their differences with homeostatic microglia and DAM is scrutinized. (C) PIC-seq is currently being calibrated in order to dissect interactions between immune cells and oligodendrocytes (Adapted from (Giladi et al., 2020)).

Cre-lop recombinase system (Bouabe and Okkenhaug, 2013) and adenoviral vectors with an oligodendrocyte specific promoter (Jonquieres et al., 2013). As both technologies can be used to block (or induce) the expression of a given gene specifically by oligodendrocytes, we should be able to block or stimulate the expression of a DOL gene by OLs and observe the induced physiological effects.

The first systems allows to selectively remove a gene from the mouse genome in a specific cell type but requires multiple rounds of breeding before getting the expected mouse strain. Once the breeding is over, the cognitive abilities of the mouse of the mice can be tested and compared with WT mice (Figure IV.6 left panel). Adenoviral vectors can be used to induce or block the expression of a given gene in a specific cell type through the use of particular promoter (for instance the promoter of the MBP gene in oligodendrocytes (Jonquieres et al., 2013)), and do not require multiple rounds of breedings but will only achieve a partial infection, and
Figure IV.6: Possible future axis of research to better understand the DOLs’ contribution to AD pathogenesis \textit{in-vivo}. (A) Breeding strategy to create mouse lacking the expression of a target gene specifically in oligodendrocytes. Such mice might have an AD-prone genetic background and used to study the effects of this KO on the disease development through behavioral tests. (B) Adenovirus vectors can be used to induce (or block) the expression of a given gene specifically in oligodendrocytes. AD mice treated by a specific adenovirus vector can provide information about the role of the target gene at a cellular level (through scRNA-seq) or at a more physiological level (through behavioral tests).

Therefore alteration, of the mouse brain cells. Following the infection, the cognitive abilities of the mice could be be tested and the transcriptome of the brain cell measured by MARS-seq. Both approaches require significant amounts of time to be implemented, but are the only way to study the exact role of DOLs \textit{in-vivo}.

Therefore, much remains to be done to elucidate the origin of DOLs and their contribution, but the various single-cell approaches aforementioned, as well as new emerging multiplexed imaging techniques could answer to these questions and pave the way to a better understanding of AD and of other neurodegenerative diseases.
General Conclusion

During my PhD, I have studied immune cell cross-talks in various contexts, ranging from mouse models of immunization (Blecher-Gonen et al., 2019), human viral infections (Bost et al., 2020a,b), mouse and human tumors (Thibaut et al., 2020) and neurodegeneration. Each time, single-cell technologies were used and were key in the understanding of the studied biological process. By combining standard scRNA-seq with other experimental approaches, such as pathogen tagging (Blecher-Gonen et al., 2019), in-vitro cell co-culture (Bost et al., 2020b) and intravital imaging (Thibaut et al., 2020), my collaborators and I have shown that:

- A tailored immune response to a given pathogen can be observed in the first 24/48 hours of the immunization process and is encoded by antigen-positive cells. In addition, early production of IFNg by NK cells is essential to induce a potent Th1 response in the lymph node.

- T-cell response in COVID-19 patients correlates with the severity of the disease while immune-suppression of blood myeloid cells predicts the clinical outcome of severe patients. Moreover, viral secondary infections could be observed in two cohorts and seemed to alter the patient immune response.

- IFNg does not act locally in tumors, but instead has a global and sustained effect on tumor cells and tumor-infiltrating immune cells.

- Homeostatic oligodendrocytes can turn into an activated subset in AD brain, that we named Disease associated OLigodendrocytes, and are likely induced by a specific subset of microglia. DOLs can be observed in multiple pathological contexts, including AD, MS and aging.

In parallel to these results, I would like to highlight that I have developed several computational tools, even though I have not published any ‘pure’ methodological paper. Indeed, as mentioned earlier in this thesis, data generated by scRNA-seq technologies are complex, large and hard to interpret. Therefore, dedicated computational tools and approaches were developed and tested to analyze the data generated across my projects. In order to make them accessible to the scientific community, the corresponding scripts and packages are freely available on my GitHub. This list of tools consists in:
Conclusion

- A stringent differential gene expression analysis method based on data binarization and binomial regression, initially developed in (Blecher-Gonen et al., 2019) and later improved in (Bost et al., 2020a).

- The Viral-Track pipeline, initially described in (Bost et al., 2020a) and applied on a larger scale in (Bost et al., 2020b) to 21 severe COVID-19 patients.

- The use of correspondence analysis to study large and heterogeneous scRNA-seq datasets generated from clinical cohorts. This approach was used to stratify COVID-19 patients using scRNA-seq data, as described in (Bost et al., 2020b).

- Large-scale secondary analysis of previously published scRNA-seq datasets to generalized observations to other models, diseases or species, as shown in (Thibaut et al., 2020) and in the DOL project.

- Image analysis scripts to automatically quantify STAT1 nuclear translocation at a single-cell resolution, in both in-vitro (conventional microscopy) and in-vivo (intravital imaging) settings as shown in (Thibaut et al., 2020).

Study of immune cell cross-talks is a thriving field, and will likely remain so in view of the constantly new success of immunotherapy to treat various diseases (Ständer et al., 2020; Traboulsee et al., 2020). The emergence of Chimeric Antigen Receptor T-cells (CAR T-cells), i.e genetically engineered T-cells expressing a synthetic receptor that binds a tumor antigen, as powerful treatments of liquid cancers (leukemia, lymphoma) further highlights the clinical interest of understanding those cellular interactions. However much remain to be done: interactions between immune and cancer cells inside the tumor are still poorly understood, alterations of the immune system leading to an aberrant immune response in infectious and auto-immune diseases remain mysterious, and interactions between the immune and the nervous system in neurodegenerative diseases is still puzzling. Therefore, much effort is still needed to address those questions and the development of new experimental and computational approaches will probably be required.

So what is the next step? New single-cell technologies are emerging, including single-cell proteomic by mass-spectrometry (Budnik et al., 2018; Marx, 2019), joined single-cell measurement of transcriptome and intracellular protein activity (INS-seq (Katzenelenbogen et al., 2020)) or extremely high throughput scRNA-seq (Datlinger et al., 2019). While improving our capacity to analyze complex biological systems, those techniques will not be game-changer as they will provide data similar to the one already available. The real change will likely come from the new spatial genomic and multiplexed imaging tools previously mentioned in the manuscript. While not being ‘real’ single-cell technologies, they can measure hundreds to thousands of variables with a cellular to sub-cellular spatial resolution, opening the door to multivariate spatial analysis. For instance, spatial transcriptomic has recently been used to
study transcriptional changes induced by the proximity to Aβ plaques in AD mice brain (Chen et al., 2020), while multiplexed IF imaging was used to elucidate the link between the spatial organization of the immune system in colorectal tumors and survival of the patients (Schürch et al., 2020).

Similarly to scRNA-seq in its early phase, the analysis of data generated by high-dimensional spatial techniques is challenging. In addition to the classical challenges when analyzing a new type of data such as data normalization, feature selection and batch correction, spatial data analysis suffers from a unique problem: unlike classical genomic data to which standard statistical test can be directly applied, the intrinsic autocorrelation between samples can perturb significance tests when the data are analyzed using standard correlation and regression techniques. Therefore all the statistical genomic analysis toolkit developed for bulk and single-cell genomic can not be applied and a new set of tools has yet to be developed. Interestingly, several recent papers do not seem to take this limitation into account, suggesting that secondary analysis could reveal new precious biological knowledge. So far, only a limited set of tools has been developed, (Svensson et al., 2018a; Arnol et al., 2019; Tanevski et al., 2020) and a significant amount of work will be required to develop appropriate and dedicated statistical tools, making it the main challenge for computational biologists in the coming years.
Bibliography


Medaglia, C., Giladi, A., Stoler-Barak, L., Giovanni, M. D., Salame, T. M., Biram, A.,
David, E., Li, H., Iannacone, M., Shulman, Z., and Amit, I. (2017). Spatial reconstruc-
tion of immune niches by combining photoactivatable reporters and scRNA-seq. *Science*,

Publishing Group.


responses: triggers of or triggered by autoimmunity? *Nature Reviews Immunology*, 9(4):246–
258. Number: 4 Publisher: Nature Publishing Group.

Nagano, T., Lubling, Y., Yaffe, E., Wingett, S. W., Dean, W., Tanay, A., and Fraser, P.
(2015). Single-cell Hi-C for genome-wide detection of chromatin interactions that occur
Nature Publishing Group.

changes in Alzheimer’s disease: a focus on myelin and oligodendrocytes. *Acta Neuropatho-
lógica Communications*, 6.

Neftel, C., Laffy, J., Filbin, M. G., Hara, T., Shore, M. E., Rahme, G. J., Richman, A. R.,
Silverbush, D., Shaw, M. L., Hebert, C. M., Dewitt, J., Gritsch, S., Perez, E. M., Gonz-
alez Castro, L. N., Lan, X., Druck, N., Rodman, C., Dione, D., Kaplan, A., Bertalan,
M. S., Small, J., Pelton, K., Becker, S., Bonal, D., Nguyen, Q.-D., Servis, R. L., Fung,
J. M., Mylvaganam, R., Mayr, L., Gojo, J., Haberler, C., Geyeregger, R., Czech, T., Slave,
I., Nahed, B. V., Curry, W. T., Carter, B. S., Wakimoto, H., Brastianos, P. K., Batchelor,
T. T., Stemmer-Rachamimov, A., Martinez-Lage, M., Frosch, M. P., Stamenkovic, I., Riggi,
N., Rheinbay, E., Monje, M., Rozenblatt-Rosen, O., Cahill, D. P., Patel, A. P., Hunter,
M. L. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for

Oyler-Yaniv, A., Oyler-Yaniv, J., Whitlock, B. M., Liu, Z., Germain, R. N., Huse, M., Altan-
cytokine propagation enables plasticity in cell-to-cell communication in the immune system.


List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example of a simple signaling pathway activated by an a soluble factor. Adapted from (Alberts, 2015)</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Main forms of intercellular signaling found in multicellular organisms. Adapted from (Alberts, 2015)</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>The JAK-STAT pathway and its activation by cytokines. The binding of the cytokine to two receptor monomers allows the associated JAK to be close enough to phosphorylate each other on tyrosines and make them fully active. Once activated, they phosphorylate the receptor itself allowing the recruitment and subsequent phosphorylation of STAT proteins. The phosphorylated STAT proteins form a dimer that can enter the nucleus and activates the transcription of specific genes. Adapted from (Alberts, 2015).</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>The NF-kB pathway and its activation by TNFa. The binding of TNFa causes a conformational change of the aggregated cytosolic tails of the receptors, which then recruit several signaling proteins, resulting in phosphorylation and activation of IKB kinase kinase (IKK). IKK is a heterotrimer composed of two kinase subunits and a regulatory subunit called NEMO. The two kinases phosphorylate IκB, inducing its degradation and the liberation of the NF-kB dimer. Upon liberation, the dimer enters the nucleus and activates the transcription of specific inflammatory genes. Adapted from (Alberts, 2015)</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>Value of single-cell analysis for cellular signaling studies. (A) Mean activation of the MAPK pathway of a set of Xenopus’ oocytes with increasing doses of progesterone. (B) Two different models compatible with the observation made in the panel (A). In the first one (upper panel) MAPK activation is homogenous across cells and the activation level of each cell progressively increase with progesterone dose. In the second (bottom panel), cells tend to display an all-or-none response and it is the proportion of activated cells that increases with the dose. Adapted from (Alberts, 2015; Ferrell and Machleder, 1998)</td>
<td>9</td>
</tr>
</tbody>
</table>
6 Single-cell imaging analysis reveals the NF-kB oscillation induced by (TNF-a). (a) Mouse fibroblasts expressing a p65 protein merged with a fluorescent protein (DsRed) are cultivated in vitro and stimulated in vitro. Activation of the NF-kB pathway is surveyed using fluorescence microscopy by looking at the cellular localization of the fluorescent protein. Red arrows highlight cells exhibiting an activation of the pathway. (b) Oscillation of the NK-kB activity in individual cells. Each curve corresponds to a unique cell. Adapted from (Tay et al., 2010)

7 The two arms of the adaptive immune system. Adapted from (Alberts, 2015).

8 IFNg production and signalling. Adapted from (Ivashkiv, 2018).

9 Effects of IFNg on immune and non-immune cells. (a) IFNg general effects shared by various cells. (b) Effects on various immune cells. (c) Effects on non-immune cells. Adapted from (Ivashkiv, 2018).

10 Clinical and immunological features of the two polar forms of leprosy. The two upper panels correspond to sections of leprosy lesions stained by hematoxylin and eosin, as well as the key features of the two forms. Cytokines expression patterns from patients displaying one the polar forms studied by Northen blots are displayed in the bottom panel. Adapted from (Geha and Notarangelo, 2016; Murphy and Weaver, 2017).

11 Survival curves of Advanced Squamous-Cell Non-Small-Cell Lung Cancer patients treated with chemotherapy only (Docetaxel curve) or chemotherapy plus an anti-PD1 antibody (Nivolumab curve). Nivolumab treatment significantly increases survival rate and was the first FDA approved anti-PD1 antibody. Adapted from (Brahmer et al., 2015).

12 Time before rheumatoid arthritis symptom increase in patients treated by methotrexate (MTX) only or MTX plus etanercept (anti-TNFa antibody) at two different doses. Anti-TNFa antibodies have significantly improved quality of life of patients suffering from rheumatoid arthritis, but are also efficient at treating other pathologies such as psoriasis, inflammatory bowel disease and ankylosing spondylitis. Adapted from (Smolen and Aletaha, 2015; Smolen et al., 2013).

13 Drawing done by Cajal after the observation of a chicken nervous system. A high diversity of size and shape among the cells can be observed. Adapted from (Cajal, 1924).

14 Exponential scaling of scRNA-seq experiments in the last decade. Adapted from (Svensson et al., 2018b).

15 List of technological breakthrough that made large-scale scRNA-seq possible. Adapted from (Svensson et al., 2018b).

16 Principle of the droplet-based single-cell sequencing. Adapted from (Macosko et al., 2015).
17 CRISP-seq approach. Cas9 expressing mouse cells are infected with a pool of lentiviral vectors containing different simple guide RNA (sgRNA). Cells will therefore express a specific sgRNA, or a combination of several sgRNAs. ScRNA-seq allows to simultaneously measure the transcriptome and identify the sgRNA(s) of each cell: the effect of each sgRNA can therefore be assessed on a single-cell level and possible epistasis studied at a high throughput. Adapted from (Jaitin et al., 2016).

26

18 Approach for simultaneous epitope and transcriptome measurement in single cells. Adapted from (Stoeckius et al., 2017; Peterson et al., 2017).

27

19 Number of scRNA-seq studies published per month stratified by method measurement in single cells. Adapted from (Svensson et al., 2019b).

27

20 The four main steps of any scRNA-seq protocols.

28

21 Schematic highlighting the different steps in the MARS-seq 2.0 protocol. Adapted from (Keren-Shaul et al., 2019).

29

22 Coverage of the human Metapneumovirus N gene. The data have been generated by scRNA-seq of COVID-19 patients bronchoalveolar lavages using the 10X Genomics Chromium technology. Adapted from (Bost et al., 2020a).

31

23 Comparison between droplet and plate based single-cell sequencing protocols. While droplet based approaches can sequence at a higher throughput, plate based approach can efficiently isolate rare cell population without over-sequencing abundant cell types. Adapted from (Keren-Shaul et al., 2019).

32

24 Comparison of the different graph-based approaches used for UMI and cellular barcode correction. Each node represents a barcode sequence and the inner number the amount of the given barcode. For a full description of each method, see the original UMI-tools paper (Smith et al., 2017). Adapted from (Smith et al., 2017).

34

25 Advances in immunology enabled by scRNA-seq. Adapted from (Avraham et al., 2015; Stubbington et al., 2017).

37

I.1 T-helper cell subsets function and generation process. Each Th subset is promoted by a specific set of cytokines that trigger the expression of key transcription factor. The resulting Th subsets produce themselves unique sets of cytokines. Adapted from (Yamane and Paul, 2013).

42

I.2 Naive T cells encounter antigen presenting cells in the lymph nodes. DCs (and other APCs) enter the lymph node through afferent lymphatic vessels while blood lymphocytes enter through the HEV. Pathogen specific T-cells are quickly activated and start to proliferate while the others leave the lymph node by efferent lymphatic vessels. Adapted from (Murphy and Weaver, 2017).
I.3 Approached used in (Avraham et al., 2015) to study the heterogeneity of cells phagocyting the intracellular bacteria *Salmonella typhimurium*. Bacteria express the Green Fluorescent Protein (GFP) and are labeled by the red dye pHrodo which binds to the cell wall of bacteria and increases in fluorescence in the low pH environment of macrophage lysosomes (left panel). Due to rapid degradation of GFP in case of bacteria death, relative red and green fluorescence can be used to distinguish uninfected macrophages and macrophages containing live or dead bacteria. The stochastic expression of the bacterial transcription factor PhoP regulating LPS synthesis determines the expression of type I IFN genes by infected macrophages (right panel). Adapted from (Avraham et al., 2015). ................................................................. 45

I.4 Strategy used to decipher cellular communication networks in (Blecher-Gonen et al., 2019) ................................................................. 46

I.5 Model of the early immune response to mycobacteria in the lymph node. ......................................................... 74

I.6 Limitations of our single-cell study of immune communications. Some of these limitations will be addressed in the next chapters of this thesis. ......................................................... 75

II.1 Examples of pathogen emergence or re-emergence. Red circles correspond to newly emerging diseases while blue ones correspond to re-emerging ones. It is worth noting that this figure was published in 2004, since then, several new pathogens have emerged, including SARS-CoV-2. Adapted from (Morens et al., 2004). ................................................................. 77

II.2 Main approaches used to study viruses-induced diseases. ......................................................... 78

II.3 Principle of dual RNA-seq. Adapted from (Westermann et al., 2012). ......................................................... 80

II.4 Comparison of the three emerging coronaviruses and of Influenza A. ......................................................... 82

II.5 Genome organization of SARS-CoV-2. Adapted from (Kim et al., 2020) ......................................................... 82

III.1 Effects of IFNg in tumors. IFNg is a critical immune regulator of the tumor and has both anti-tumor effects (red box) by promoting antigen presentation and cancer cell apoptosis, and pro-tumor (blue box) by inducing the expression of immune checkpoints such as PDL1. Adapted from (Ivashkiv, 2018). ......................................................... 163

III.2 Methods for highly multiplexed imaging. Three types of methods have been developed: multiplexed IF which relies on successive cycles of staining and bleaching, MIBI/IMC an imaging version of mass cytometry and SeqFISH/MERISH, two multiplexed versions of single molecular FISH. ......................................................... 197
IV.1 The pathological features of AD and their evolution over time. (A) Microscopy images from amyloid plaques and neurofibrillary tangles published in Spielmeyer’s classic textbook ‘Histopathologie des Nervensystems’. (B) Spatial and temporal description of Aβ plaques and neurofibrillary tangles evolution in the brain of AD patients. Adapted from (Masters et al., 2015) ........................................ 200

IV.2 Mechanisms of protein aggregation in AD. (1) APP is produced by neurons and cleaved to produce soluble Aβ. (2) Soluble Aβ aggregate to form Aβ fibrils and plaques (3) which are destroyed in a phagocytic or protease dependent manner. (4) In AD, toxic Aβ oligomers dissociate from plaques and are toxic to surrounding neurons and (5) induce tau aggregation within neurons. (6) Neurofibrillary tangles induce neuronal death, are released and can be internalized by neighboring neurons. Adapted from (Masters et al., 2015) .................. 201

IV.3 Model of microglia polarization toward the DAM state. It is a two-stage process where different genes are induced and repressed at each stage. Stage 2 DAM require the expression and activation of the TREM2 gene, likely by protein aggregates or dead cells. Adapted from (Deczkowska et al., 2018) .............. 202

IV.4 Organization of oligodendrocytes in the CNS. Oligodendrocytes generate plasma membrane protrusions that wrap around axons that are called myelin sheaths. ‘Naked’ axon region are usually referred to as nodes of Ranvier (Kuhn et al., 2019). ................................................................. 203

IV.5 Current working directions of the DOL project. (A) We are investigating the possible spatial association between DOLs and Aβ plaques through IF imaging of mouse and human AD brains. (B) Existence of the so-called ‘alternative’ microglia is currently investigated and their differences with homeostatic microglia and DAM is scrutinized. (C) PIC-seq is currently being calibrated in order to dissect interactions between immune cells and oligodendrocytes (Adapted from (Giladi et al., 2020)). ................................................................. 238

IV.6 Possible future axis of research to better understand the DOLs’ contribution to AD pathogenesis in-vivo. (A) Breeding strategy to create mouse lacking the expression of a target gene specifically in oligodendrocytes. Such mice might have an AD-prone genetic background and used to study the effects of this KO on the disease development through behavioral tests. (B) Adenovirus vectors can be used to induce (or block) the expression of a given gene specifically in oligodendrocytes. AD mice treated by a specific adenovirus vector can provide information about the role of the target gene at a cellular level (through scRNA-seq) or at a more physiological level (through behavioral tests). .................. 239