Thèse de doctorat

Exploring human T cells functional diversity using single-cell RNAseq: methodological and biological strategies

Thèse de doctorat en Génétique, Omiques, Bioinformatique et Biologie des systèmes

ED 474 Frontières de l'Innovation en Recherche et Education
Thèse dirigée par Vassili Soumelis

Thèse présentée et soutenue publiquement à Paris, le 8 octobre 2021, par

Elise AMBLARD

Composition du jury:

Aurélien DE REYNIES PhD, PU-PH, Université de Paris	Président du jury
Pierre SAINTIGNY MD-PhD, Médecin spécialiste, Centre de Recherche en Cancérologie de Lyon	Rapporteur
Céline VALLOT PhD, CRCN, Institut Curie	Rapportrice
Pierre MILPIED PhD, CRCN, Centre d'Immunologie de Marseille-Lumigny	Examinateur
Magali RICHARD PhD, CRCN, Université Grenoble Alpes	Examinatrice
Vassili SOUMELIS MD-PhD, PU-PH, Université de Paris	Directeur de thèse

Synthèse

Titre : Exploration de la diversité fonctionnelle des cellules T dans les données de séquençage d'ARN en cellule unique à l'aide d'outils méthodologiques et biologiques

Résumé : Le séquençage d'ARN en cellule unique (scRNAseq) est une technique jeune. Elle consiste à faire une photographie instantanée des ARN d'une cellule. Après une phase timide d'adoption, son usage s'est généralisé. La richesse de ces données permet de disséquer finement la biologie du vivant, en accédant à des informations telles que l'hétérogénéité d'une population ou la caractérisation différentielle des cellules saines et malades.
Cependant, le scRNAseq est à double tranchant. Bien que cela se soit démocratisé, le travail à la paillasse est encore perfectible, car on ne capture que 5 à 20% des ARN. Quant au travail à l'ordinateur, il constitue un défi : les données sont bruitées car immergées dans un espace à grande dimension, et car la capture des ARN est incomplète.
La nouveauté du scRNAseq n'a pas encore permis l'émergence de standards d'analyse communs. Au contraire, il y a une explosion des algorithmes. Cependant, il y a un schéma consensuel : importation et nettoyage de la matrice d'expression cellules $\times$ gènes, normalisation et réduction de dimension. Les données en dimensions réduites permettent de faire de la visualisation, de l'agrégation ou de l'inférence de trajectoire. Enfin les groupes sont annotés. Cette dernière étape d'interprétation est particulièrement critique mais souvent biaisée.
Je me suis attachée dans ma thèse à deux aspects de l'analyse des données scRNAseq : l'aspect méthodologique, et l'interprétation.
J'ai d'abord étudié le bruit dimensionnel, autrement appelé la malédiction de la dimensionnalité. Cette malédiction complique l'analyse en brouillant les différences entre points proches et lointains. L'analyse scRNAseq, qui repose sur la production de graphes de voisinage, est nécessairement pénalisée par cette malédiction qui déforme les graphes. L'astuce habituelle consiste donc à réduire la dimension. Il existe un autre effet de la malédiction, le phénomène de hubness, qui est aussi nocif, car il déforme le graphe des plus proches voisins. Toutefois ce phénomène peut être corrigé. J'ai évalué l'ampleur du phénomène de hub dans les données omiques, ainsi que l'effet de la correction de hub sur la performance de l'analyse scRNAseq. Le phénomène de hub est bien présent, en particulier dans les matrices caractérisées par une grande dimension intrinsèque, et l'analyse de ces jeux de données en particulier bénéficie de la réduction de hub, avec une performance optimale dans l'espace de dimension effective maximale. Bien que cela ne semble être qu'un algorithme de plus dans la jungle déjà existante, c'est surtout le changement de paradigme qui est singulier, puisqu'on modifie conceptuellement une étape consensuelle, la réduction de dimension.
Ensuite, je me suis intéressée aux cellules T, d'abord via le prisme des lymphocytes T régulateurs. Ces cellules, définies initialement par leur fonction, sont difficile à isoler chez l'homme. En partant de l'hypothèse qu'il y a potentiellement décorrélation entre le phénotype et la fonction, j'ai ensuite élargi mon cadre d'étude à l'ensemble des cellules T en questionnant le paradigme actuel de lignée. J'ai adopté une approche supervisée afin de capturer la fonctionnalité des cellules T avec le scRNAseq. A l'aide de modules fonctionnels pré-définis, je peux relier chaque cellule à sa/ses fonction/s. J'ai d'abord prouvé l'apport de cette approche par rapport à un pipeline non supervisé. Ensuite, j'ai caractérisé les différences fonctionnelles entre cellules T issues d'un tissu sain ou cancéreux. Nous avons aussi implémenté cette méthode pour l'analyse de cellules dendritiques de

patients souffrant de la Covid-19, après sélection des modules fonctionnels idoines. Cette stratégie peut donc être appliquée pour d'autres types de cellules que les cellules T, d'autres pathologies que le cancer, et même dans un contexte physiologique, afin de cartographier les fonctions des cellules immunitaires.

Mots clés : séquençage d'ARN en cellule unique, analyse de données omiques, grande dimension, fonctionnalité, approche supervisée, cancer, immunologie, bioinformatique.

Synthesis

Title: Exploring human T cells functional diversity using single-cell RNAseq: methodological and biological strategies

Abstract: Single-cell ARN sequencing (scRNAseq) is a fairly young technique. It makes a snapshot of a single-cell transcriptome. After a slow start, its usage became more systemic. Indeed, the richness of the data enables a fine dissection of a living organism's biology. It gives access to an unprecedented amount of information to better understand cell heterogeneity, or quantify the differences between physiological and pathological states. However, the single-cell approach is a double-edged sword. In spite of its democratisation, the wet lab part is still perfectible, as we are currently only able to capture 5 to 20% of the reads. Regarding the computer process, it is challenging: scRNAseq data is noisy as its effective dimension is high, and because of the incomplete capture.
Because scRNAseq is still a young technique, not enough time elapsed for analysis standards to emerge. On the contrary, there is an exponential increase in the number of analytical tools. However, there is a common pipeline: load the genes $\times$ cells count matrix -or its transpose-, filter out outlier cells and genes, normalise and reduce the dimension. From the data projected onto a smaller subspace, the next steps can be clustering, trajectory inference or visualisation. Finally, the different clusters or trajectory nodes are annotated. This last step, where we interpret the data, is critical but unfortunately often biased.
In this thesis, I focused on two aspects of the analysis of scRNAseq data: a methodological aspect, and the interpretation step.
First, I studied dimensional noise, alternatively called the curse of dimensionality. The curse complicates the analysis. It blurs the differences between close and far away data points. Since analysing scRNAseq relies heavily on the production of neighbor graphs, the performance will be degraded by the curse, which distorts the graphs. The usual trick is to reduce the dimension. However, the blurring, or concentration, of distances is not the only effect of dimensional noise. An additional phenomenon called the hubness phenomenon is also detrimental to the analysis as it distorts nearest neighbors graphs. While measure concentration cannot be corrected in high dimensional spaces, hubness can. I quantified the magnitude of the hubness phenomenon in omics data, and the effect of correcting for hubness on the performance of scRNAseq analysis. scRNAseq data is indeed "hubby", especially the datasets with a high intrinsic dimension. The performance when analysing the latter would be improved upon hubness correction, with the best performance reached in the space with the highest effective dimension. I reckon that it might be perceived as just another tool in the already existing jungle, but I believe that the change of paradigm is really interesting, as we modified conceptually one of the most performed step of the analysis, the dimension reduction.
Second, I focused more specifically on T cells, through the prism of regulatory T cells. Those cells have a precise functional definition, while there is no strong consensus on the population's markers for humans. I hypothesized that there might be a decorrelation between function and phenotype and I decided to extend my study to all T cells, since the lineage paradigm is also questionable here. I did a supervised analysis of scRNAseq data in order to better unveil T cells' functionality. After defining functional modules, I can link each cell to its function/s. First, I assessed the novelty of the approach, by comparing it to the unsupervised pipeline. Then, I characterized the functional differences between T cells from a healthy or a cancer tissue. We also implemented this method to analyse dendritic cells from Covid-19 patients, scoring functions exerted by dendritic cells. This

strategy can be applied for other immune cells, other diseases, and even in a physiological setting, so as to functionally map immune cells.

Keywords: scRNAseq, omics data analysis, high dimension, hubness, functionality, supervised approach, cancer, immunology, bioinformatics

Remerciements

En premier lieu, je remercie mon jury, qui a accepté la lourde tâche de relire ce manuscrit, et d'évaluer le travail de ces dernières années : Céline Vallot et Pierre Saintigny, qui ont bien voulu être rapporteurs de cette thèse, ainsi que Magali Richard, Aurélien de Reyniès et Pierre Milpied. Je suis également reconnaissante envers Servier, à travers Olivier Nosjean et Antoine Bril qui m'ont permis de mener à bien ce travail.

Je remercie aussi mon directeur de thèse, Vassili Soumelis, de m'avoir accueillie dans son équipe, et même dans ses consultations d'hématologie ! Merci de m'avoir laissé mener mes projets et mes collaborations avec une grande liberté. Bien évidemment, merci à toute la team Soumelis passée et présente pour leur enthousiasme et leur bienveillance : dans le désordre, Charlotte (& Nicolas !), Alain, Sarantis, Caroline, Léa, Coline, Philémon, Clémence, Iris, Salima, Lilith, Faezeh, Arturo, Jasna, Maeva, Emna, Alba, Justine, Daniel. En particulier, merci à Pierre, Floriane, Camille, Maude, Melissa et Lucile pour avoir répondu patiemment à chacune de mes questions quand j'étais (un peu) perdue. Je tiens aussi à remercier les équipes avec qui j'ai pu collaborer aussi intensément: Andrei Zinovyev et Jonathan à Curie, Antonio Rausell, Akira, et Loredana à Imagine.

Je profite aussi de cette page pour exprimer ma gratitude envers Brice Gaudillière et Pascal Priollet, qui ont, par leurs personnalités, leur aménité et leur passion, profondément inspiré et soutenu mon parcours académique.

Quant à mes amis, merci Julie pour les soirées à thème, les distractions extra-doctorales, pour m'avoir suivie en séjour d'escalade, de ski de rando ou de rédaction. Merci Regimb pour ton aide, ton soutien, tes encouragements, tes cours de surf et de LaTeX, qui n'ont pas encore porté leurs fruits... J'ai hâte d'être à ta soutenance ! Merci à mes amis de l'escalade qui m'ont si souvent écoutée râler, et si souvent encouragée, que ce soit pour attraper la prise d'après ou pour la thèse: Luc, Nils, Paulin, Gabriel, Maud, Tibo, Jeanne. Je vous laisse un peu de répit d'ici la prochaine (thèse)... Merci Kraj pour ta patience quand il a fallu m'inculquer quelques notions sur la RMT. Merci à la formidable bande de fidèles copines, sobrement surnommée ❤: l'astucieuse Lili, la fantasque Riri, la sportive Cel, la calme Myl, et l'audacieuse Clem. En vrac, merci aussi à Max Piffoux pour tes conseils et ton côté aussi apaisant que de l'aloe vera, Vidalou et Hugues pour votre générosité et votre gentillesse, Yara pour sa présence affectueuse pendant mes débuts à Curie, et Julien qui est si inspirant.

Merci aussi à la meilleure équipe de grimpeurs qui ne grimpe plus beaucoup, et aux autres amis de l'X : PA, Xavier, JP-san, Ziem, Max, Yassine pour m'avoir laissé house-sitter à Villejuif, Clément pour m'avoir traînée à la danse, Lau pour ses visites chocolatées et gegrilltes-gemüse-ées, la fratrie Zakine pour me faire faire des abdos à force de rigoler, Augustin Braun qui me rappelle sporadiquement qu'une thèse peut durer 9 ans, Nicolas Cliche pour accepter avec beaucoup d'indulgence que je ne travaille jamais mon violon, et que je viens toujours sans mes partitions, Jonas et Anne que j'ai retrouvés avec plaisir à Curie, Max André pour les dîners et les petits-déjeuners annuels, Mich parce qu'il swing du tonnerre !

Je me sens privilégiée d'avoir pu faire ma thèse au sein du CRI, grâce auquel j'ai pu bien m'entourer : l'inimitable ekirpe et leurs légendaires eskirpades, Hugo (et Loulou), et Judith, mon (ex-)voisine préférée ! Évidemment, merci infiniment à Anthony pour ses encouragements et son soutien précieux, indispensable et indéfectible pendant ces années de thèse à rebondissements, et pour m'avoir reboostée à chaque fois que j'en ai eu besoin (c'est-à-dire souvent !).

Enfin, merci à ma famille. Une mention spéciale à mon cousin JB qui m'a laissé l'envahir cet hiver, et à mon tout nouveau beau-frère, Jeanpo, pour son aide toujours experte mais parfois cryptique et ergodique. Merci à mes parents qui m'ont donné les clés pour me construire et m'épanouir, et ont été de vaillants supporters depuis toujours; à ma grande soeur Irène, qui est repassée en tête pour me montrer la voie et qui m'écoute -le plus souvent- déblatérer avec patience; à ma petite soeur Esther, qui a elle aussi développé ses qualités d'écoute à mon côté, ce qui a peut-être contribué à hauteur de 1% à faire d'elle une future psychiatre exceptionnelle, et qui me nourrit, dans tous les sens, avec sa gentillesse.

Contents

List of Abbreviations **xi**

List of Figures **xiii**

Preamble **16**

Origins of the project . . . 16

Synopsis of the manuscript . . . 18

I Introduction **19**

1 Cancer and immunity **20**

1.1 T cell biology . . . 20

1.1.1 Brief historical introduction to immunology . . . 20

1.1.2 The T cell story . . . 21

1.1.3 The lineage paradigm . . . 23

1.1.3.1 Cytotoxic cells . . . 23

1.1.3.2 Helper cells . . . 23

1.1.3.3 Regulatory cells . . . 24

1.1.4 T cells are plastic . . . 26

1.1.4.1 Fluctuating phenotypes and functions . . . 26

1.1.4.2 Plasticity in disease . . . 27

1.1.4.3 Questioning the lineage paradigm . . . 28

1.2 Cellular cross-talks in the Tumor Microenvironment . . . 29

1.2.1 Historical perspective on cancer . . . 29

1.2.2 Cancer hallmarks . . . 30

1.2.3 The Tumor Microenvironment . . . 31

1.2.4 Plasticity of immune cells in cancer . . . 32

1.2.5 Focus on the regulatory cells' example . . . 32

1.3 The 'Why' and 'How' of single-cell RNAseq . . . 33

1.3.1 Why using single-cell RNAseq . . . 33

1.3.2 Methodological background . . . 33
1.3.3 How to perform single-cell RNAseq... . . . 35
1.3.3.1 ...from the wet experiment... . . . 35
1.3.3.2 ...to the dry experiment . . . 37

2 Single-cell RNAseq analytical challenges 41
2.1 Modelling noises . . . 41
2.1.1 Biological noise . . . 41
2.1.2 Technical noise . . . 43
2.1.3 Discriminating technical from biological noises . . . 44
2.2 Sparsity . . . 46
2.2.1 Reminder on the terminology . . . 46
2.2.2 Zero-inflated... . . . 46
2.2.2.1 Modelling zero-inflation . . . 46
2.2.2.2 Imputation . . . 48
2.2.3 ...or not ? . . . 48
2.3 Dealing with high dimensional data . . . 50
2.3.1 A new type of data . . . 50
2.3.2 New analytical methods . . . 50
2.3.3 Curse and blessings of dimensionality . . . 50
2.3.3.1 Definition . . . 50
2.3.3.2 Avoiding effects of the curse of dimensionality . . . 52
2.3.3.3 Targeting the measure concentration effect . . . 52
2.3.3.4 Targeting noise in the data . . . 53
2.3.3.5 Avoiding hubness . . . 53

3 Single-cell RNAseq biological interpretation 55
3.1 Balancing between technical and biological accuracy . . . 55
3.1.1 A humongous amount of analytical tools... . . . 55
3.1.2 ...to benchmark . . . 56
3.1.2.1 Performance scores . . . 56
3.1.2.2 Good practices in benchmarking analytical tools . . . 57
3.2 Looping on old knowledge . . . 58
3.2.1 Gene list enrichment . . . 58
3.2.2 Differential Gene Expression . . . 61
3.2.3 single-cell RNAseq atlases . . . 61
3.2.4 Automated annotation of single-cell RNAseq data . . . 62
3.3 Creating single-cell RNAseq-based knowledge . . . 63
3.3.1 Detection of new cell types . . . 63

3.3.2 Validation of new cell types . . . 64

4 Objectives of the thesis 65

4.1 First focus: tackling the curse of dimensionality in order to improve the performance of scRNAseq analysis pipelines . . . 65

4.2 Second focus: Dissect the functional diversity of single-cell RNAseq of T cell in cancer with a supervised functional approach . . . 66

4.3 Third focus: Context-dependant approach enable to unveil T cell functions: the example of regulatory T cells in cancer . . . 66

II Results 68

5 Hubness reduction improves clustering and trajectory inference in single-cell transcriptomic data 69

5.1 Synopsis of the hubness study . . . 69

5.2 Article . . . 69

6 Supervised analysis of single-cell RNAseq data to functionally classify T cells in cancer 115

6.1 Introduction and statement . . . 115

6.2 Functional modules . . . 116

6.2.1 Construction . . . 116

6.2.2 Description . . . 120

6.2.3 Comparison with previous knowledge . . . 122

6.3 Scoring functions in single-cell RNAseq data . . . 123

6.3.1 Encoding . . . 123

6.3.2 Control with bulk RNAseq data . . . 127

6.4 Assessing the added value of the functional modules . . . 128

6.4.1 Mutual information between ground truth labels and module scoring . . . 130

6.4.2 Mutual information between cluster labels and module scoring . . . 130

6.5 Mapping tumoral T cells functions . . . 133

6.5.1 Barcoding cells . . . 133

6.5.2 Interpretation . . . 135

6.6 Conclusion . . . 138

6.7 Perspectives . . . 139

7 Meta-analysis of regulatory T cells in cancer: highlighting their prognostic role in a context-dependent manner 144

7.1 Rationale behind the meta-analysis . . . 144

7.2 Article . . . 144

III General discussion and perspectives **162**

8 General considerations and prospects **163**

8.1 On technical aspects 163

8.1.1 The hubness project 163

8.1.2 Validity and reproducibility of dry lab experiments 164

8.1.3 Accumulate data or create new tools? 165

8.1.4 Creating scRNAseq-based knowledge 166

8.2 On biological aspects 166

8.2.1 The functional project 166

8.2.2 Regulatory T cells in cancer 167

8.2.3 Implementing scRNAseq in onco-immunology, from the bench to the bedside 168

IV Annexes **183**

9 Single-cell RNAseq of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity **184**

10 Meta-analysis of human cancer single-cell RNAseq datasets using the fully integrated IMMUcan database **208**

V Résumé en français **238**

List of Abbreviations

APC	Antigen Presenting Cell
BC	Before Christ
BCL	Base Class Library
bp, kb, Gb	base pair, kilo base, Giga base
CD	Cluster of Differentiation
cDNA	complementary DNA
DC	Dendritic Cell
ddNTP, dNTP	di-deoxynucleotide triphosphate, deoxynucleotide triphosphate
DGE	Differential Gene Expression
DL	Deep Learning
FACS	Fluorescence Activated Cell Sorting
GSEA	Gene Set Enrichment Analysis
HVG	Highly Variable Genes
IL	InterLeukin
IVT	In Vitro Transcription
k-NN	k-Nearest Neighbors
MCA	Multiple Correspondence Analysis
MHC	Major Histocompatibility Complex
ML	Machine Learning
mRNA	messenger RNA
NGS	Next-Generation Sequencing
PC	Principal Component
PCA	Principal Component Analysis

PCR	Polymerase Chain Reaction
QC	Quality Control
RMT	Random Matrix Theory
RT	Reverse Transcription
scRNAseq	single-cell RNA sequencing
STAT	signal transducer and activator of transcription
T_{FH}	Follicular Helper T cell
T_{H}	Helper T cell
T_{REG}	Regulatory T cell
TCR	T Cell Receptor
TF	Transcription Factor
TIME	Tumor Immune MicroEnvironment
TME	Tumor MicroEnvironment
UMAP	Uniform Manifold Approximation Projection
UMI	Unique Molecular Identifier

List of Figures

1 Manhattan and Euclidean metrics 17

1.1 Schematic view of human T cell differentiation 22
1.2 T_{REG} diversity mirrors the diversity of T_H 25
1.3 Graphical model of T_H reprogramming 27
1.4 Cellular signal integration in immune cells 28
1.5 Cancer-related discoveries timeline 29
1.6 Cancer hallmarks 30
1.7 The TME 31
1.8 Bulk versus single-cell RNAseq 33
1.9 Sanger sequencing 34
1.10 Sequencing costs over years 34
1.11 Wet scRNAseq protocol 35
1.12 Available scRNAseq protocols 36
1.13 Increase of scRNAseq datasets cardinality 37
1.14 Dry scRNAseq protocol 37
1.15 Choosing a cluster's minimal size 39
1.16 Batch correction 40

2.1 Gene expression variability 42
2.2 Waddington's landscape 42
2.3 scRNAseq count modelization 44
2.4 AE neural network for scRNAseq 45
2.5 Sparsity in scRNAseq datasets 47
2.6 Zero-inflation depends on the technique 49
2.7 Measure concentration in scRNAseq 51
2.8 Dimension reduction with PCA 53

3.1 Number of scRNAseq analytical tools 56
3.2 Silhouette score on scRNAseq data 57
3.3 Proteomic versus transcriptomic truth 59
3.4 The dynverse platform 60

3.5 GSEA score calculation . . . 60
3.6 The balance between sequencing depth and cell number . . . 63
3.7 RaceID picks out outlier genes . . . 64

6.1 Functional modules cardinality . . . 120
6.2 Gene-level specificity . . . 120
6.3 Function-level specificity . . . 121
6.4 Functional modules overlap . . . 122
6.5 Encodings correlation . . . 127
6.6 Control with pure bulk RNAseq data . . . 128
6.7 Functional genes are informative . . . 129
6.8 UMAP from 3 different sets of genes . . . 129
6.9 Effector UMAP mixes ground truth labels . . . 130
6.10 Classical UMAP mixes functional scoring . . . 131
6.11 Encoding cluster scores distribution . . . 132
6.12 Inter-cluster differences are smaller than expected . . . 132
6.13 Intra-cluster differences are higher than expected . . . 133
6.14 Fraction of cells positive for each functional module . . . 134
6.15 Similarity between the 4 partitions . . . 135
6.16 Encoding cell scores distribution . . . 136
6.17 Consensus UMAP . . . 137
6.18 Functions correlate differentially depending on the tissue of origin . . . 138

"L'expérience est une lanterne accrochée dans le dos, qui n'éclaire que le chemin parcouru."

Confucius

Preamble

Origins of the project

As a wannabe medical student, I have always been fascinated by the human body, its physiology and its response to diseases. As a wannabe researcher in "hard" sciences, I have been passionate about studying mathematics and physics. I wanted to combine these two interests in an interdisciplinary project when it was time for me to start a PhD work. After my masters I was still a complete novice, and interested by many subjects, but I wanted to have at least:

- ❒ an interdisciplinary project,
- ❒ involving "hard" sciences,
- ❒ and human health, with possibly a focus on cancer,
- ❒ and on immunology,
- ❒ in an hospital environment that would foster interactions with medical doctors.

When I started to interact with Vassili Soumelis in order to start my PhD within his team at Institut Curie, I had the freedom to write my own subject. I wanted to work on onco-immunology, and I was in particular intrigued by regulatory T cells. It has been a bulky field of research, and I had stumbled upon a few interesting articles that attempted at describing their role in cancer and their therapeutic value. I also remembered working during one of my master theses in the Nolan Lab in the US on single-cell RNA sequencing data (scRNAseq), and being fascinated, and overwhelmed, by the wealth of information that this technique would bring.

The initial project has been to do a fine characterization of regulatory T cells and the differences between the tumor compartment or the healthy counterpart, the juxtatumor compartment. I would use scRNAseq data to shed light on those differences. In particular, I wanted to study the interplay between effector and regulatory T cells. When I started to work on public data, I quickly realised the challenges inherent to the analysis of scRNAseq data. One of the main challenges was to cope with the curse of dimensionality, a set of strange, and usually unwanted phenomena that happen in high-dimensional spaces like the space of genes used to describe the cells in scRNAseq data.

So I decided I should first tackle this curse, starting with one of its most common effect: the measure concentration. If you describe objects, for example cats, you can quantify different features to do so. You can start collecting features such as weight, height or length of hair, and use them to accurately discriminate for example kittens from adult cats. But if you accumulate features, for example tail length, breadth and height of

head or ears, size of the pupils, and so on, an unsupervised approach will usually perform worse than with a couple of features, even though we have the intuition that the more information, the better. The reason behind this remarkable observation is the measure concentration. It relates to the fact that in high-dimensional spaces, the contrast between similar and dissimilar data points, for example adult cats and kittens, vanishes. It has also been observed that metrics do not suffer equally from the measure concentration, with for example the Manhattan metric being less sensitive than the Euclidean one (Figure 1), even if the difference in sensitivity shrinks with the dimension.

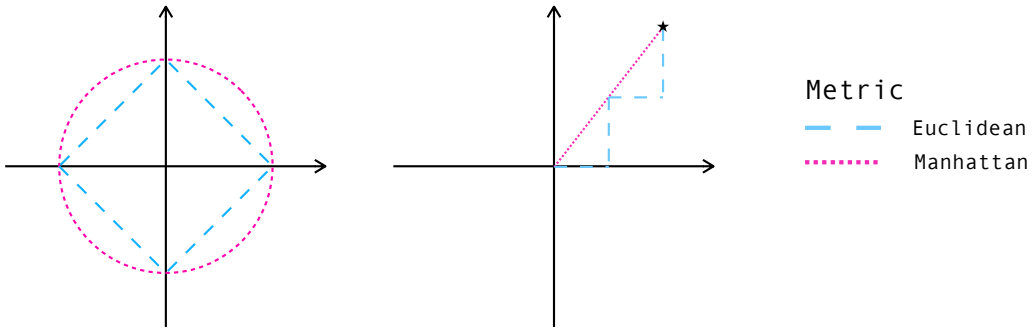


Figure 1: Unit circle (left) and point norm (right) for the Manhattan and Euclidean metrics.

The first idea that I tested was to do metrics learning in order to propose a metric better adapted to scRNAseq data than the Euclidean metric. For this task, we need a cost function with constraints: either we have ground truth labels and we can define constraints by stating that cells with the same label should be similar, while cells with different labels should not, or we can use other similarity metrics to write constraints. At this point, I estimated that it could not work as the constraints would be flawed: ground truth labels are rare, and their granularity is too coarse, while using existing metrics that suffer from measure concentration would be like a snake biting its own tail. Additionally, each trained metric would be optimal for a specific question and a specific dataset, so that the approach cannot be generalized. Going back to our cat example, and using the features weight, height, length of hair and number of hair colors, the metric should focus on height and weight if we want to distinguish babies from adults, but on hair length if we want to separate a Sphynx from an Angora cat, and on number of hair colors to separate an albino from a healthy cat.

Instead of tackling the measure concentration that is *in fine* hardly correctible in high-dimensional spaces, I studied instead the hubness phenomenon. It is an another dimension-related effect that distorts k-nearest neighbors (k-NN) graphs. To take another illustrative example, a neighbor graph is for example the Parisian subway map: each subway station is connected to other close stations, and we can count the number of connections, or connectivity degree, for one node. Some stations have a particularly high connectivity degree like République (degree = 10) or Châtelet (degree = 15), while others are barely connected like Olympiades (degree= 1) or new unconnected stations like Sevran-Livry (degree= 0).

In parallel of this methodological work, I would not forget the fact that improving the raw performance of an analysis is worthwile but does not necessarily improve the interpretability of the data and the biological conclusions. Furthermore, I was still interested by working specifically on T cells, and even more on immune cells from the tumor microenvironment, as I was puzzled by the difficulty of interpreting that kind of data.

I was also intrigued by the plasticity of the different T cell types. My first focus has been the regulatory T cell population. I wondered at why this population was so hard to grasp, and I took the prism of evaluating its impact on cancer prognosis. Indeed, there is a plethora of articles attempting at characterizing the link between Tregs and cancer prognosis, while there is no consensus, and even less in a pan-cancer approach. To tackle this issue, I worked on a meta-analysis in order to see whether I could extract some hints that would either explain the discrepancy or shed light on Tregs' role in cancer.

Since I wanted to take advantage of scRNAseq data, because it is both rich and available, I also worked on the analysis and functional interpretability of T cells from scRNAseq data and I postulated that we could describe each cell as a collection of functions, in order to better dissect functional diversity.

Synopsis of the manuscript

I will introduce in the first part notions about immunology, oncology and scRNAseq: why improving the technique is crucial but also challenging.

Then I will introduce the three research questions I tried to answer during my PhD thesis and the results I obtained for each of them. I worked hand in hand with Andrei Zinovyev and Jonathan Bac for the hubness project, and with Antonio Rausell and Akira Cortal for the functional project. Finally, I will conclude on those results and outline the prospects opened by the present work.

In the annex, I show additional projects that I did or collaborated on during my PhD: annotation of dendritic cells from Covid-19 patients, with the functional approach developed in the second part of the results, that was done in the team with Melissa Saichi (published), and a scRNAseq database for cancer datasets constructed in collaboration with European partners from the ImmuCAN project (in preparation).

Part I

Introduction

Chapter 1

Cancer and immunity

1.1 T cell biology

In vertebrates in general and in humans in particular, two systems would protect against pathogenic encounters: the innate immune system and the adaptive one. While innate immunity triggers an immediate and non-specific response, the adaptive immune system is able to recognize and target specific patterns, called antigens. Among the cells belonging to the adaptive immunity are T cells, which will be the focus of this section. A major source for the upcoming section will be the seminal textbook *Janeway's immunobiology* [1].

1.1.1 Brief historical introduction to immunology

The fact that an encounter with a pathogen confers protection for forthcoming ones has been known for a long time, although the underlying mechanisms were unveiled many centuries later. Thucydides (ca 460-400 BC) reported in his *History of the Peloponnesian War* that "nobody would be infected for the second time and die from the disease" (II, 51) during the Athenian plague, that has been described either as a typhus or smallpox epidemic, depending on the hypothesis [2]. Later on, there is the striking example of smallpox and variolation. Smallpox was a serious disease since it was highly contagious, deadly and debilitating for survivors, leading to blindness, joint or skin damage, encephalitis, etc. In Constantinople, Lady Mary Wortley Montagu (1689-1762) found out about an old Chinese tradition that consisted in inoculating, either in the nose or through the skin, the pus of smallpox pustules. Various sources estimate it has existed since somewhere between the VIth and the XVIth centuries [3]. Montagu brought back this procedure to England, from where it spread to other European countries, such as France, under the influence of Voltaire among others [4], as a successful way of protecting against smallpox.

Edward Jenner (1749-1823) is considered as the founder of modern immunology upon proving in 1796 that the inoculation of cowpox, a mild and bovine version of smallpox, could protect against smallpox. He named this procedure vaccination, from the Latin word for cow, *vacca*, which describes the injection an attenuated form of the pathogen in order to produce what he called artificial immunity. Thanks to vaccination, the World Health Organization declared the eradication of smallpox in 1980[1].

[1] https://www.who.int/features/2010/smallpox/en/

Following Jenner's work, Louis Pasteur (1822-1895) designed a vaccine against other diseases, namely cholera and rabies [5]. The basis to understand the processes underlying the success of vaccination was provided by the work of Robert Koch (1843-1910), Emil von Behring (1854-1917), Shibasaburo Kitasato (1853-1931) and Jules Bordet (1870-1961), who discovered respectively the existence of pathogens [6], the "anti-toxic activity" of the serum (due to antibodies) [7], and the complement. These findings mirror the former contention between two theories: the cellular theory -supported e.g. by Ilya Ilyich Mechnikov (1845-1916)- and the humoral theory -supported e.g. by Paul Ehrlich (1854-1915)-, that culminated in 1908, when the two front runners were awarded the Nobel Prize in Physiology or Medicine[2].

1.1.2 The T cell story

Ehrlich supported the humoral theory, and was the first to hypothesize the existence of antibodies and side-chains (or cell membrane receptors) in 1900 [8, 9]. He predicted that antibodies were excess secreted side-chains [10], and thus caught up with Bordet's opinion that every secreted active substance is of cellular origin. He is also the first to make a discrimination between immunological self and non-self, leading to the concept of auto-immunity (or *horror autotoxicus*) and immune regulation. In the wake of these suggestions, James B Murphy (1890-1930 ?) claimed that lymphocytes were the cellular basis of antibodies [11]. It is quite surprising that his discoveries went completely unnoticed[3].

Only half a century later, James Gowans (1924-2020) (re)discovered the cells responsible for adaptive immunity by studying the circulation of lymphocytes, after he was told that "if [he] can find out where they go, [he] can find out what they do" [12]. Only then, the discovery was widely acknowledged. Upon this result, James Miller (1931-) apprehended the existence of two populations, that he later called the T and B lymphocytes, because of where they are produced: T cells in the thymus and B cells in the bone marrow. B cells are the cells that effectively produce antibodies, while T cells interact with and help them [13]: in that respect, it is obvious that communication and cooperation between immune cells is of paramount importance in the physiological immune system. After this paradigm shift, Miller ushered in a new research field, viz. T cell biology.

Focusing on T cells, we can further classify them into different categories, be it activation status (naive vs. effector vs. memory cells), functional (helper vs. cytotoxic cells) or phenotypic categories (type 1 vs. type 2 vs. type 9).

Let us go through the different stages in the life of a T cell [1, 14]. The young T cell is naive: it is small, 5-7 μm in diameter, with low transcriptional activity and few organelles. The effector T cell is in its prime after the crucial encounter with an antigen and the recognition of a specific epitope presented by a self cell surface protein, the Major Histocompatibility Complex (MHC). The conversion to an effector phenotype is elicited by a set of stimulation cues:

1. Capture and display of an epitope by the MHC of an Antigen Presenting Cell (APC),

2a. Physical interaction between the epitope-MHC complex and the T Cell Receptor (TCR) specific for this epitope, among a repertoire of 10^8 TCRs,

[2] https://www.nobelprize.org/prizes/medicine/1908/summary/
[3] He does not even have his Wikipedia article!

2b. Expression of co-stimulatory molecules, both membrane-bound and secreted, by the APC. In the case of cytotoxic cells, an additional activation step by a so-called helper cell is usually required.

The effector cell goes through the following steps, after the antigenic encounter: it stops migrating, its size increases up to 10-20 μm, there is a burst in transcriptional activity and a high proliferation rate giving rise to thousands of clones. Effector helper cells stay in the lymphoid tissue to activate other immune cells, such as B cells and $CD8^+$ cytotoxic T cells, while cytotoxic cells migrate towards the site of infection. Finally, after antigen clearance, some elder T cells retain the memory of the pathogenic encounter, while most of the effector cells die. The memory T cell is the basis of immunological memory, as it will be activated faster than naive cells upon reinfection.

Zooming yet further on effector and memory cells, it has been observed that there are several phenotypes, related to two main classes: the helper class that is $CD4^+$, and the cytotoxic class that is $CD8^+$. Helper cells exhibit a plethora of subsets that orchestrate different parts of the immune response, and so do cytotoxic cells. Since cytotoxic subsets mirror helper subset, we will merely describe the latter. The first discovered subsets were the T_H1 and T_H2 populations, shortly followed by other populations though, such as T_H9, T_H17, T_H22, T_{FH} (for follicular helper) and T_{REG} (for regulatory) [1, 15]. Each of these subsets arises from a precise set of stimulation cues and is specific to a peculiar class of pathogens (Figure 1.1). The cytotoxic counterparts are the Tc1, Tc2, Tc9 subsets and so on [16]. Other T cell populations have been described that we will not discuss in the scope of this manuscript: $CD4^+CD8^+$, $CD4^-CD8^-$, $\gamma\delta$ T cells, or NK-T cells.

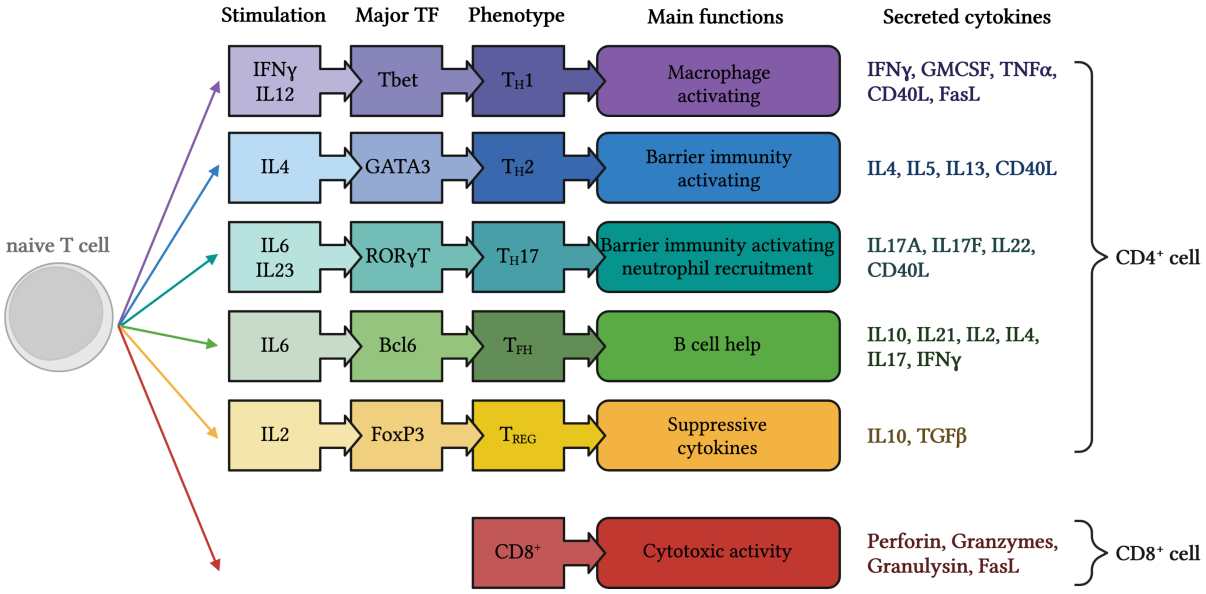


Figure 1.1: Schematic view of human T cell differentiation. A naive T cell differentiates into different subsets according to the context, in order to yield an appropriate response. Depending on major stimulatory cytokines (first column), different transcription factors (second column) are activated, that would lead to a phenotype (third column) specific of a particular context (forth column), producing appropriate effector cytokines (fifth column).

1.1.3 The lineage paradigm

The different populations are linked to specific pathogenic contexts and functions. They are now classically differentiated according to the expression of phenotypic markers and cytokine secretion patterns.

1.1.3.1 Cytotoxic cells

While an extracellular pathogen will be targeted by antibodies or the complement system, rules are different for intracellular pathogens such as viruses, that are not directly accessible. In that case, infection can only be cleared by the destruction of the infected cell itself. The cytotoxic activity is carried on mostly by $CD8^+$ T cells, although $CD4^+$ T cells sometimes acquire cytotoxic properties [17, 18][4]. The killing role is exerted via the immunological synapse through which the cytotoxic cell sends its death signals to the target. The immune synapse itself is formed upon recognition of the cognate antigen presented at the surface of an infected cell by the MHC to the TCR.

The first killing method is the extrinsic pathway of apoptosis, in which the killer cell activates the death receptors of the target by producing the corresponding ligands. The ligand-receptor binding initiates the deadly signalling cascade. The range of death ligands includes FasL, TNFα and LTα.

The second killing method is the intrinsic pathway. It initiates in the absence of survival signals, or as a response to toxic stimuli. The stimulus, either positive (noxious signal) or negative (no survival signal), triggers the same cascade as the extrinsic pathway. The cytotoxic cell can produce noxious stimuli such as cytotoxic granules.

The apoptotic cell is then broken down and ingested by phagocytic cells. One achievement of cytotoxic cells is that they perform quietly: they do not damage nearby cells as they accurately target the cell of interest, nor do they modify the inflammatory status of the milieu.

There is a third *modus operandi*: by secreting various bystander cytokines. IFNγ directly targets viral replication and inhibits it. It also increases the number of displayed MHC proteins in order to better flag infected cells. Finally, it recruits macrophages. TNFα and LTα participate as well in macrophage activation. Macrophages are indeed important as they can also act as APCs.

1.1.3.2 Helper cells

The helper cells have a supporting role, as indicated by their transparent denomination, in the sense that they help other cells, such as B cells and $CD8^+$ cells, to mount an efficient immune response. Let us describe the main helper populations: T_H1, T_H2, T_H17, and T_{FH}, although other populations were reported, such as T_H9 and T_H22. These populations are differentiated according to the class of pathogens that they respond to, and the cytokines they produce; they arise from different combinations of stimulatory cytokines (Figure 1.1).

T_H1 was discovered along with T_H2. It was first observed that helper cells encompass heterogeneous cells [19] and described in murine clones, as subsets with differentiated cytokine secretion patterns [20]. It was later associated to specific pathogenic contexts [21].

[4] Incidentally, this is a first hint of a flaw in the current population model that we will challenge later on.

Naive T cells differentiate into T_H1 cells in the presence of IFNγ and IL-12, generally in the context of intracellular pathogens [22]. These cytokines activate in turn transcription factors (TF) of the signal transducer and activator of transcription (STAT) family, viz. STAT1 and STAT4. The two STATs are able to stimulate the expression of T-bet and IFNγ: the former is the key TF of the T_H1 lineage and the latter one of its signature cytokines [23]. It will also enhance the expression of other genes, such as IL-10, IL-21 or ICOS. Signature proteins include, but are not restricted to, IFNγ, CXCR3, IL-2 or TNFα [15, 16].

Usually opposed to T_H1, the T_H2 population is specialized in eliminating helminths and is induced from naive T cells by IL-2 and IL-4 [15, 22]. IL-4 triggers the activation of STAT6, that is the intermediate to more than 80% of IL-4-regulated genes. Among its targets are GATA3, RUNX1 and BATF, with the first one being the major lineage TF for the T_H2 population [24]. The following cytokines are also produced upon T_H2 differentiation: IL-4, IL-5, IL-9, IL-10, IL-13, CCR4, or TNFα [16, 22].

The third helper subset is the T_H17 population. It is characterized by the production of IL-17 upon IL-23 stimulation [25]. These cells fight extracellular bacteria and fungi. The differentiation into this lineage is triggered by TGFβ together with IL-6, IL-21 and IL-23 that activate STAT3. The TF binds notably to the *Il17* locus, but also to the *Rorc*, *Irf4*, *Il23r* and *Il6ra* loci [23]. Its signature cytokines are IL-10, IL-17A, IL-17F, IL-22, TNFα, CCR4 and CCR6 and the major lineage TF is RORγt [15, 16].

Regarding T_{FH}, it provides help to B cells, locates in the germinal center and is identified by CXCR5 and PD1 [22]. The precise requirements for its differentiation have not been fully elucidated yet, but a good candidate is IL-6. The main TF is Bcl6, that initiates the expression of CXCR5, the receptor for CXCL13 expressed by stromal cells from the B-cell follicle. Other proteins expressed by T_{FH} are ICOS, whose ligand is produced by B cells, and IL-21, that stimulates the proliferation and differentiation of B cells into antibody-producing plasma cells. These three proteins are essential to enable the co-localization and communication with B cells.

Let us briefly mention the case of T_H9 and T_H22: T_H9 are IL-9-producing cells that differentiate from naive T cells after stimulation with IL-2, IL-4 and TGFβ. These 3 cytokines activate STAT5 and STAT6, which bind to the *Il9* promoter. The role of the pleiotropic cytokine IL-9 is not yet completely understood and might play a role in both protective immunity and immunopathological diseases [26]. On the other hand, T_H22 produce IL-22, their phenotype is acquired over stimulation of naive T cells with IL-1β, IL-6 and TNFα and the major TF is AHR. These cells work to enhance innate defences, mostly in skin where they reside [27].

All these subsets are tightly intertwined by positive and negative feedback loops: they share stimulation cytokines, TFs, and effector cytokines, and this is a first lead to explain the volatile commitment of a effector cell to one or the other lineage [23], as well as the fragility of a monolithic view of effector subsets. They also have mutually exclusive signalling cascades: for example IFNγ produced by T_H1 cells inhibits the proliferation of T_H2 cells, while IL-4 produced by T_H2 cells inhibits the proliferation of T_H1 cells.

1.1.3.3 Regulatory cells

The existence of T_{REG} was already hypothetized by Ehrlich at the beginning of the XXth century, when he understood the need for a regulatory mechanism of the immune system

in order to avoid what he called *horror autotoxicus*, although T_{REG} would not be the only safeguard of the immune system, which includes also the innate immunity e.g. [28]. This intuition was dusted off later in the seventies with a $CD8^+$ suppressor population [29], and the formal proof of the existence of a regulatory population was exhibited by Sakaguchi in 1995 [30]. These cells are functionally defined: cells with negative immune regulatory properties. This definition encompasses highly heterogeneous subpopulations, including $CD8^+$ T_{REG} , and B_{REG}, that we will not discuss here. Focusing on $CD4^+$ regulatory cells, we observed several sources of heterogeneity: ontogenic, functional, phenotypic, etc.

The ontogenic diversity originates from the two different modes of production for T_{REG} cells. The first ontogenic subset is produced in the thymus, and termed natural, or thymic T_{REG} (nT_{REG} or tT_{REG} respectively). The second subset stems from circulating naive T cells under specific stimulation cues, and is called peripheral *in vivo* or induced *in vitro* T_{REG} (pT_{REG} or iT_{REG} respectively) [31]. These two subsets have different functions [31], phenotypes and epigenomes [32], although it is not completely elucidated because of the scarce existence of differential markers, especially in humans, where Helios has been suggested as a good candidate to mark nT_{REG} but has not been widely adopted yet.

There is also a major functional diversity. There is for example a plethora of suppressing mechanisms, depending on whether the T_{REG} targets an APC or another T cell, or whether the suppression is contact-dependant or -independent (cf the introduction of the article in the results chapter 7). Another source of functional variety is the existence of T_{REG} populations that remarkably mirror the helper subsets, viz. T_H1-, T_H2-, T_H17-like T_{REG} and T_{FR} (for follicular regulatory) [33]. These T_H-like regulatory cells express the same TFs as their helper counterparts: T_H1-like T_{REG} are $Tbet^+$, T_H2-like T_{REG} $GATA3^+$, and T_H17-like T_{REG} $ROR\gamma t^+$. They also arise from the same contexts and are specifically regulating each of the matching helper populations (Figure 1.2) [34].

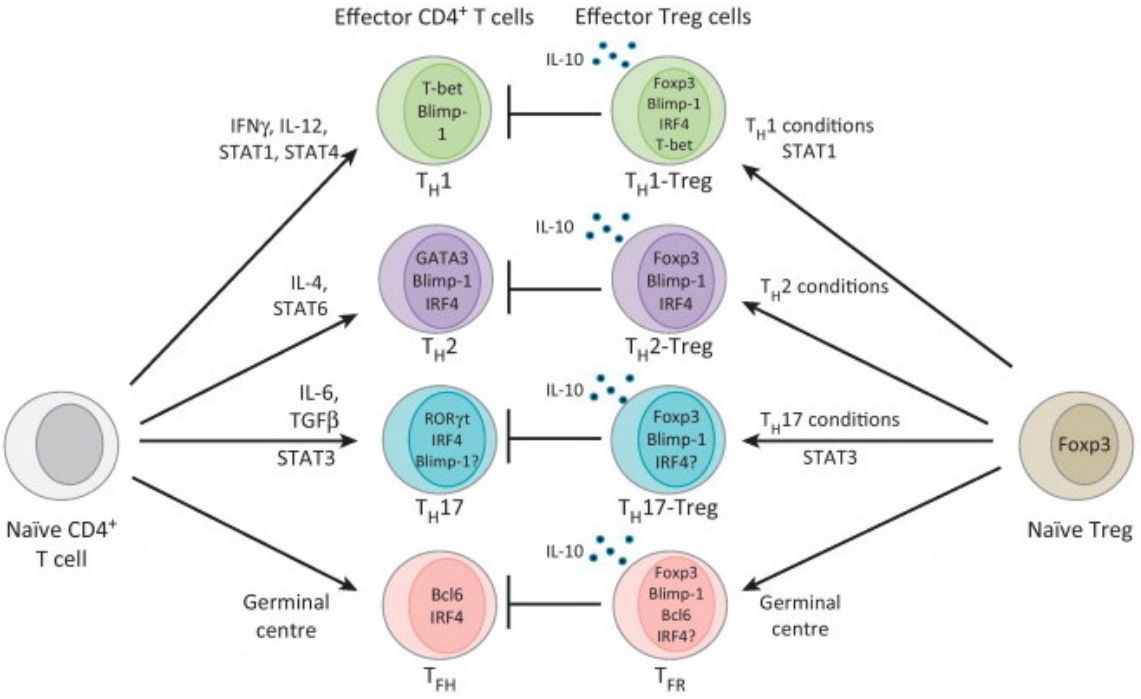


Figure 1.2: T_{REG} diversity mirrors the diversity of T_H. Adapted from [33].

Despite this heterogeneity, the T_{REG} population shares common traits. In the mouse, they are $CD4^+CD25^+$. In humans, although there are different flavors, the most consensual markers are CD4, CD25, FOXP3 and absence of CD127. These cells are elicited upon IL-2- and TGFβ-mediated activation of STAT5 that binds to the *Foxp3* locus and increases their survival chances by modulating the anti-apoptotic protein BCL2 and the receptor for IL-2, namely CD25 (or IL-2RA) [23]. Classical proteins expressed by T_{REG} include CD25, CTLA4, IL-10 and TGFβ [15, 16]. In the same vein as for helper cells, the lineage (in)stability depends on positive and negative feedback loops.

This diversity and plasticity led us to the following question: how can we better describe these phenotypic and functional shifts?

1.1.4 T cells are plastic

1.1.4.1 Fluctuating phenotypes and functions

We can make several statements about the regulatory T cell, while they stand for all T cells as well:

- T_{REG} is a heterogeneous population,
- T_{REG} commitment to its lineage is unstable [35],
- The commitment strength itself is not determined *a priori*, but depends on the dynamic context [32, 36],
- T_{REG} plasticity is undirected, meaning that the regulatory cell can loose its functionality or regain it,
- The role of T_{REG} plasticity and the fraction of stochasticity in the process are not fully elucidated yet.

Indeed, T_{REG} adapt to the cellular environment, in particular to their different targets, as well as to the soluble environment, as they are exposed to a range of cytokines. Let us examine few examples of this plasticity:

- T_{REG} homeostasis relies on the signaling through the IL-2-IL-2R couple. Upon deprivation of IL-2 from the milieu, the $FOXP3^+$ population contracts and the level of expression of FOXP3 decreases [36].
- Comparing T_{REG} cultured in the presence or absence of IL-6, the cells loose the expression of FOXP3 when fed with IL-6 [37].
- Mesenchymal stem cells (or MSCs) induce the conversion of naive T cells into iT_{REG} and stabilise the expression of FOXP3, transforming the methylation landscape of the iT_{REG} into an nT_{REG}-like landscape [38].
- T_{REG} are able to convert to T_H1- or T_H17-like cells, under specific stress-associated conditions [39].

Similarly, helper cells exhibit evolving phenotypes as well. A first proof of this plasticity is the change of status of cytokines that were considered as unequivocal. Once

thought to be specific for T_H2, IL-10 is in fact secreted by many subsets, while IFNγ can be produced by T_H2 or T_H17 cells [40]. A even stronger challenge to the monolithic helper subsets concept is the co-expression of master TFs, such as Tbet$^+$FOXP3$^+$ or RORγt$^+$FOXP3$^+$ cells [40], or the fact that one TF can give way to another, enabling T_H17 cells to evolve into T_H1 or T_H2 cells upon stimulation with IL-12 or IL-4, respectively, or T_H2 cells to convert into an IFNγ^+Tbet$^+$ T_H1-like population [41–43]. In fact, virtually all switches are possible (Figure 1.3).

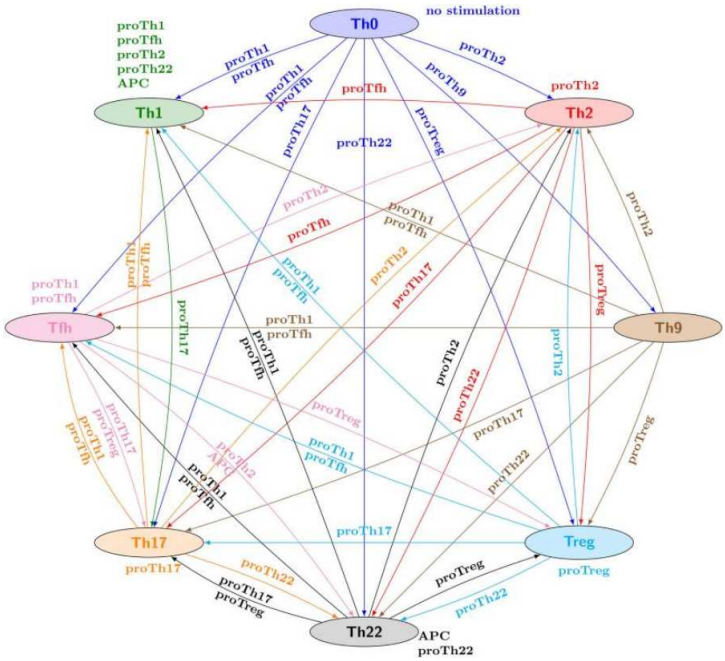


Figure 1.3: Graphical model of T_H reprogramming. Some subsets in the graph are strong attractors, such as T_H1 and T_H2, while others are highly plastic, such as T_H9. From [44].

1.1.4.2 Plasticity in disease

In a disease context, the complex local environment implies a multiplication of fate signals and thus an increased degree of plasticity. In particular in the case of cancer, there is an additional difficulty due to cancer cells modifying the milieu and rerouting immune cells towards their interest, i.e. immune escape, as explained in an upcoming section (section 1.2.2). We have mentioned among others the T_H17 switch, that require precise cytokines. The presence or absence of these cytokines depends from the dynamic context. The dynamic nature of the milieu is an increasing function from a physiological to a pathogenic situation. In order to better understand this idea, let us go through the stages during the course of an infection. It is elicited by a breach in the natural barriers, including the skin or mucosa, allowing the entrance of a pathogen. The pathogen is detected by the immune system via pathogen-associated molecular patterns, that are recognized as a danger signal. Pathogen-activated innate cells start to produce cytokines and chemokines that will in turn initiate inflammation, activate Innate Lymphoid Cells, and recruit more innate effector cells as well as APCs. Depending on its entry site, the molecular patterns detected by the innate immune system and the infection evolution, the pathogen will trigger a specific response, dominated by one of the helper subset: T_H1, T_H2, T_H9, T_H17 or T_H22 (Figure 1.4). After clearance of the pathogen, the immune system returns to

homeostasis, with the help of T_{REG}.

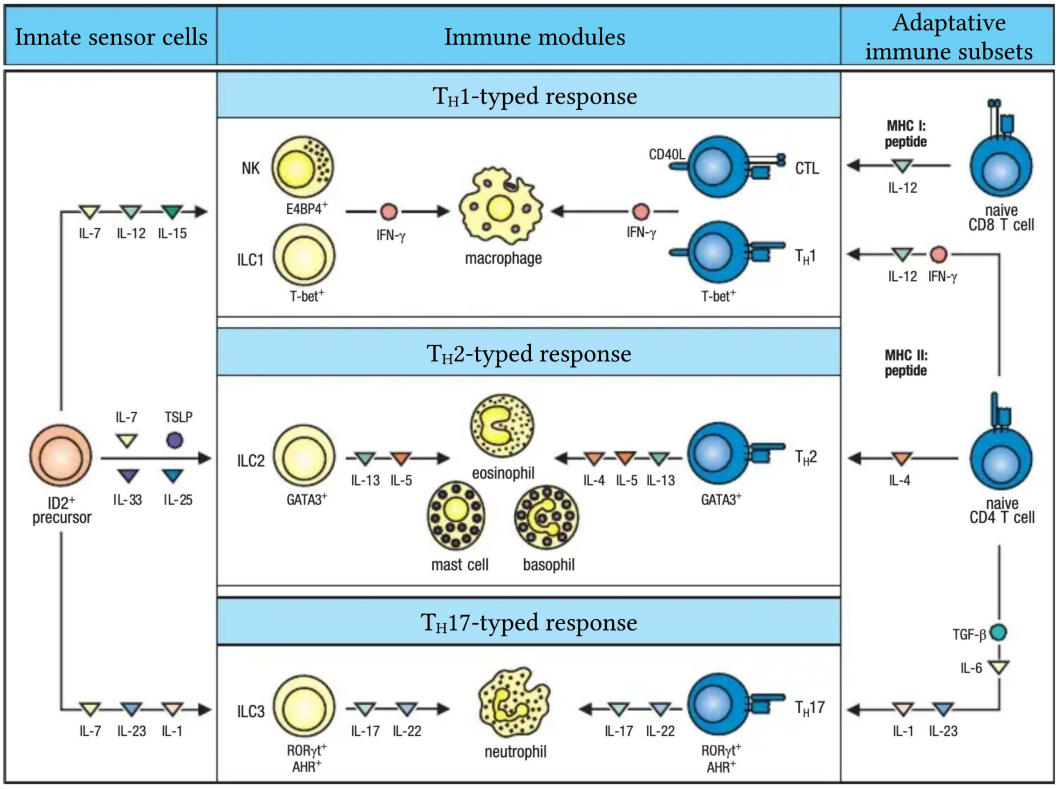


Figure 1.4: Signal integration by innate sensor cells leads to the emergence of different immune modules. Adapted from [1].

The plasticity is therefore the keystone for appropriate immune responses in disease and homeostasis in health [45, 46].

1.1.4.3 Questioning the lineage paradigm

We, and other [40], believe that the current lineage paradigm does not represent faithfully the dynamic relationships between the different poles of functions exerted by T cells. Attempts have been made in order to describe the complexity of T cells subsets, for example thanks to computer modelling of intricate molecular networks [44, 47], or to an extensive characterization and segmentation of the tangible phenotypes [16, 48]. Coming back to our question of how to better describe phenotypic and functional shifts, we hope that this doctoral work will help answer it, by attempting to propose a new point of view in classifying T cells.

1.2 Cellular cross-talks in the Tumor Microenvironment

1.2.1 Historical perspective on cancer

Cancer has been known for a long time, with a first mention of it in the Edwin Smith Papyrus, one of the oldest medical treaty, that should be at least 3,500 years old: "If thou examinest a man having bulging tumours on his breast and thou findest that swelfings have spread over his breast [...], they have no granulation, they form no fluid [...], and they are bulging to thy hand [...]. There is no treatment" [49]. Interestingly, (one of) the Hieratic graphics for "tumor"  or  is not specific of what we call nowadays a tumor, but occurs elsewhere in the surgical treatise with the meanings abscess, ulcer or rash, while the adjective "bulging" clarifies its use at this point of the text. There is also early physical evidence of tumors in mummies from ancient Egypt, and even traces of bone damage, depicted by onco-archaeologists, evocative of cancer on the skeleton of other species as old as dinosaurs or prehistoric men [50]. The name cancer derives from the ancient greek *karkinos*: Hippocrates (ca 460-370 BC), who coined the term, thought that the external ulcerating manifestations of tumors looked like a crab. There is yet again a notion of severity depicted in one of his aphorisms: "That which medicine does not heal, the knife frequently heals; and what the knife does not heal, cautery often heals; but when all these fail, the disease is incurable." The usage of this word was later clarified by Galen (ca 129-216) who decided to term only malignant tumors as *karkinomas* (non-ulcerating) or *karkinos* (ulcerating), while regrouping all tumors, including benign ones, under the greek word for swelling, *onkoi* [51]. Tumoral lumps were described as an irregular, abnormal growth of tissue strongly irrigated by blood vessels.

After this epiphany of discoveries in the Antiquity, knowledge about cancer did not significantly increase for 1,500 years. However, the description of the disease started to improve, backed up by a better understanding of human anatomy and physiology with the generalization of dissection. For example, Henri-François Le Dran (1685–1770) hypothetized the mechanism of metastatic spread via the lymphatic system. From the XVIth century onward, new findings started to accumulate exponentially in parallel with increased technological possibilities, such as enhanced imaging (Figure 1.5). The first Nobel prize in medicine related to cancer was awarded in 1926 and is now regularly dedicated to cancer-related researches, roughly every 20 years.

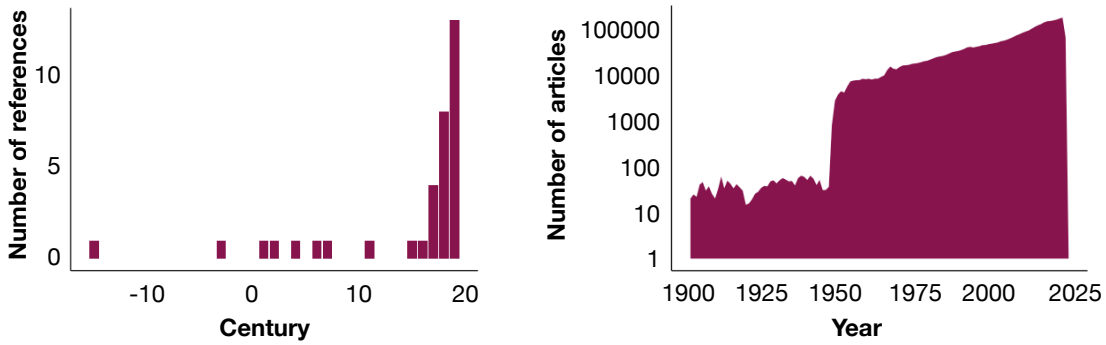


Figure 1.5: Amount of cancer-related discoveries, from 1600 BC to the XIXth century (left panel) and for the last century (right panel).

Regarding the curing panoply, it has been for a long time dominated by the surgical approach. Depending on the stage and location, it was known early on that a surgical

excision could be beneficial or detrimental. Other options would include cauterisation (associated or not to surgery), bloodletting, vegetal or animal ointments, among others. The current curing arsenal still includes parts of the ancient approaches, such as surgery or plants that were part of the pharmacopoeia and that have a proven anti-cancer activity, for example *Arisaema tortuosum* [52, 53].

1.2.2 Cancer hallmarks

In the seminal articles from Hanahan and Weinberg [54, 55], the authors delineated major factors as necessary and sufficient conditions to explain tumor behaviour. In particular, and as depicted by others [50, 56], cancer is understood as an age-related event, and as such implies an accumulation of rate-limiting probabilistic events. Evidence for this accretion of events comes from the fact that the genome of cancer cells is altered in multiple sites. These genomic alterations lead progressively to a phenotype that displays invasive features identified as major hallmarks. In the first article, the authors outlined six features: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evading apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion & metastasis. In the revised version of this list, Hanahan and Weinberg added metabolic reprogramming, immune escape, inflammation and genetic instability as key ingredients to explain the emergence of the aforementioned characteristics in candidate cancer cells (Figure 1.6).

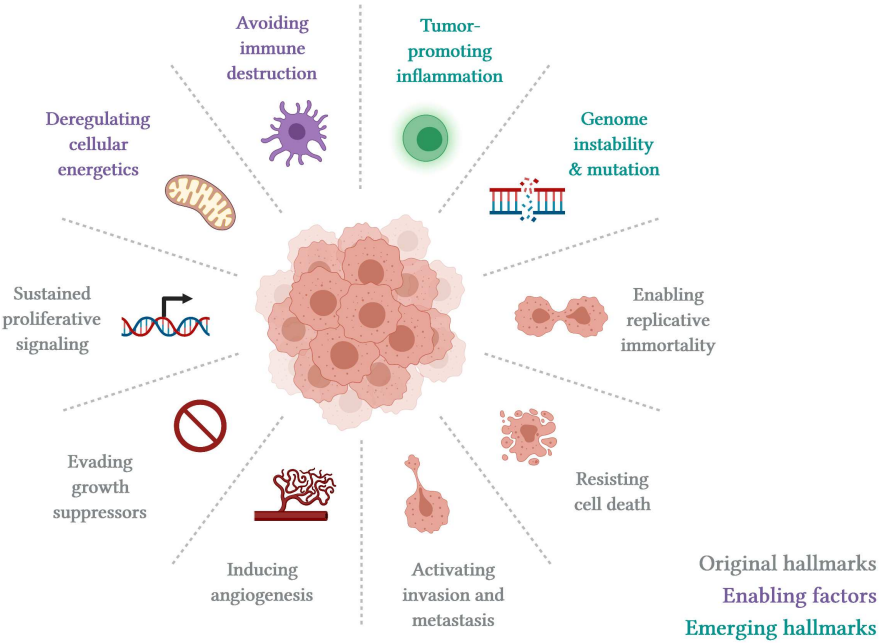


Figure 1.6: Cancer hallmarks as outlined in [55][5].

The rationale behind this surprising simplification of such a complex disease could come from the finding that the number of possible phenotypes is restricted by evolutionary trade-offs, in order for the organisms to be sub-optimal at multiple tasks [57]. For example, if an organism performs 2 tasks and displays 3 traits, then the possible phenotypes would stand on a straight line connecting the 2 archetypal phenotypes that are optimal for a single task, in the three-dimensional space of the 3 traits.

[5]this image and all images marked with an (*) were made with Biorender

Nevertheless, the conditions that lead to this supposedly frugal phenotype, combining these 10 major hallmarks, are complex because a tumor is a collection of heterogeneous cells. Those heterogeneous cells encompass cancer cells, but also stromal cells, fibroblasts, immune cells, as well as their cancer-associated flavors, since these cells are plastic and evolve as a function of the milieu they are exposed to. The mixture of all these cells is called the Tumor MicroEnvironment (TME). Although the conditions that lead to cancer are not completely clear yet, we can describe how the TME contributes to the maintenance of the 10 hallmarks described by Hanahan and Weinberg.

1.2.3 The Tumor Microenvironment

The TME plays a crucial role in the context of cancer, by providing support to cancer cells, e.g. by favoring tumorigenesis [55]. The first appearance of the term "TME" in 1988 [58] touched on the potential impact of the local environment on chemo- and radio-therapies efficiency, although it only evoked the blood component. In a subsequent mention of the TME, the author explains more carefully the impact of blood perfusion, as influencing the acidity and the nutrient input, including oxygen input [59]. Later on, more components were added, progressively drawing a clearer picture of the TME (Figure 1.7):

- blood and lymphatic vasculature,
- stroma: stromal cells forming the connective tissue (fibroblasts, epithelial and endothelial cells) [60], extracellular matrix [61],
- soluble environment [62], e.g. chemokines [63] or angiogenic factors [64],
- immune infiltrate, e.g. lymphocytes [65], natural killer cells, macrophages [63].

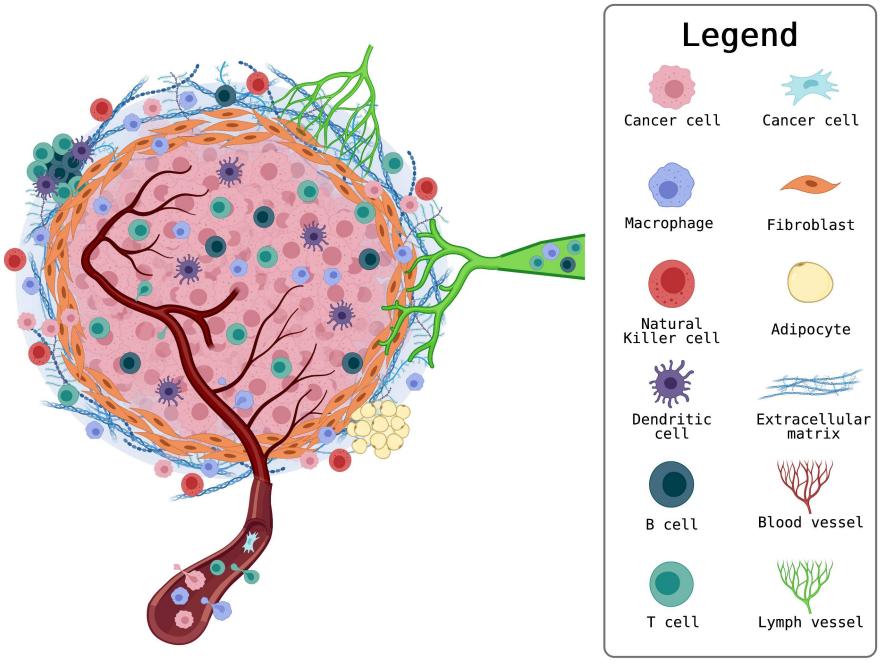


Figure 1.7: Raw scheme of the tumor microenvironment(*).

These components, in particular the immune TME (TIME), are not passive. On the contrary they have an effect on cancer editing, growth, invasion, immune escape, angiogenesis, treatment, in all kind of cancers (solid or liquid, primary or metastatic) and models (in vivo or in vitro) [1]. Contributing to further complexity, the different components interact not only with tumor cells but also with each other: for example, the degree of infiltration of murine T cells, defined as $CD90^{+}$, is smaller in more hypoxic regions [66]. Considering metastasis, there is also the idea of a favorable environment that should foster cancer cells, with the "Seed and Soil" hypothesis formulated by Paget (1855-1926), when he noticed that breast metastases would favor some sites like the liver over other like the spleen [67].

In the clinic, the TIME is playing an increasing role as it has been discovered to have a prominent part in classification, prognosis, or treatment choice. E.g., the Immunoscore has been translated in the clinic as a useful tool to predict relapse in colon cancer patients, thus helping to orient medical treatment and monitoring [68]. Another example are immune bio-markers to predict response to immunotherapy: PD-L1 is an (imperfect) marker to predict response to immunotherapy in melanoma [69] while it has been approved for head and neck squamous cell carcinomas [70]. Investigate the interplay between the TIME and cancer cells is a challenging and ongoing field of research. One of multiple reasons is that the already volatile commitment to a given immune phenotype [44] is potentiated nearby tumor cells, that are able to manipulate the latter in a highly entangled manner.

1.2.4 Plasticity of immune cells in cancer

The concept of cell plasticity has emerged from and challenged a more traditional linear view of differentiation, as discussed above (paragraph 1.1.4). It goes along with the idea that cells can and need to adapt to their environment [71, 72]. This is especially true when dealing with cancer, because of the tremendous complexity of the local milieu. More precisely, immune cells are tricked to support or ignore cancer cells [1]. Depending on the stimulation cues, they adapt to the milieu, leading to a large spectrum of phenotypes and functions, compared to the juxtatumor, considered as the healthy counterpart [73]. To embrace further the concept of constrained phenotypes, one could hypothesize that more tasks to be performed leads to more available phenotypes [57].

It is a reciprocal game though: the immune compartment has also an impact on cancer cells, on evolution and response to treatment depending on the subtle balance between inflammation and anti-inflammation, inhibition and activation [74, 75]. The treatment itself modifies the TIME [76]. A haunting question is therefore: how to accurately describe this versatile environment?

1.2.5 Focus on the regulatory cells' example

In the case of T_{REG}, the question of the versatility is of utmost importance. It is exemplified by the attempts at deciphering their role with regard to cancer prognosis. So far, there is no clear answer, since it would depend on the cancer type, the TME and the TIME, and the study itself [77]. A first manner to solve this ambivalence could be to better take into account the context (Chapter 7). Another strategy could be to focus more on the different functions exerted by T cells instead of the broad phenotypes, although phenotypes partly reflect functions. This could be achieved thanks to omics technologies, and in particular thanks to transcriptomic studies.

1.3 The 'Why' and 'How' of single-cell RNAseq

We mentioned in the previous section that the TME is a highly complex milieu. One tool that has been thoroughly used to describe this complexity is single-cell RNA sequencing (scRNAseq).

1.3.1 Why using single-cell RNAseq

There are two main advantages of using scRNAseq over bulk RNAseq or other omics methods.

Genomic studies allow to characterize DNA sequences and their variations. For example, it has been used to map the human genome [78], or in the field of evolutionary studies [79], drug target identification [80, 81] or virology [82]. But it is not possible to study cellular phenotype, nor to measure gene expression or RNA editing. Since the functional product of a gene is the protein, the ideal experiment would be to do proteomics. Unfortunately, current proteomics methods do not allow to detect an extensive number of proteins as they are not fully mature yet [83]. Instead, a satisfying proxy is to measure RNA, although its expression dynamic does not fully match the one of the protein.

Compared to bulk, scRNAseq presents several advantages: a first technical aspect is that it requires less RNA material: 1 ng vs. 1 μg. But the conceptual change is even more important, as it allows to study cell-level transcriptomic changes. Such changes are of paramount importance for example in rare cell type identification or developmental studies [84] (Figure 1.8). The fields that require dramatically a single-cell approach relate indeed to cellular heterogeneity and include, but are not restricted to, immunology, oncology, neurology or embryology [85, 86].

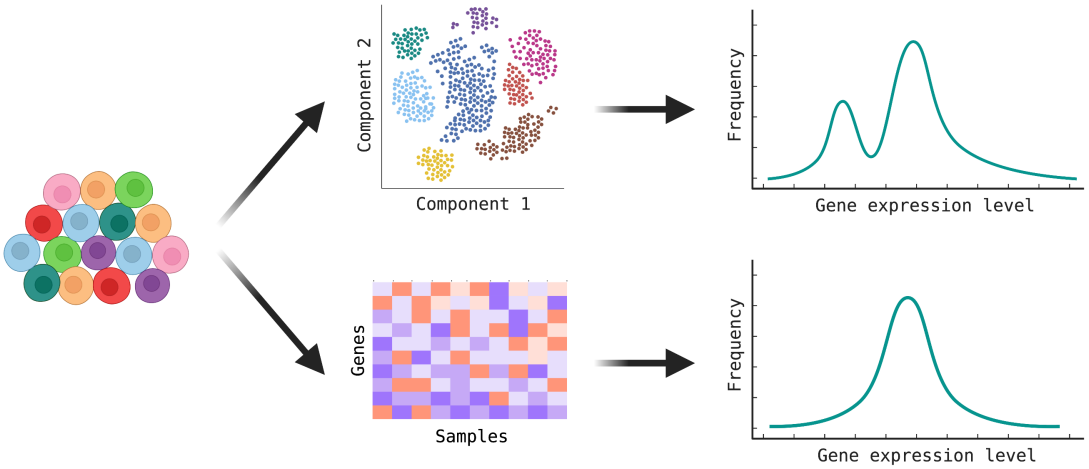


Figure 1.8: Comparison of bulk versus single-cell measures(*).

1.3.2 Methodological background

scRNAseq is a fairly new technology, as the latest innovation in a series of sequencing applications. The first sequencing method was invented by Sanger in the seventies [87, 88], based on the principle of chain termination: the nucleotide polymerisation mix

contains deoxynucleotide triphosphates (dNTPs) as well as di-deoxynucleotide triphosphates (ddNTPs) that will be randomly incorporated in the newly synthesised DNA fragment, and interrupt the synthesis. The ddNTPs are labelled, either with fluorescence or radioactivity. The fragments are then separated by size before reading the labeled nucleotide (Figure 1.9). Quality of the sequencing can be evaluated with a phred score, that indicates the confidence for each base [89].

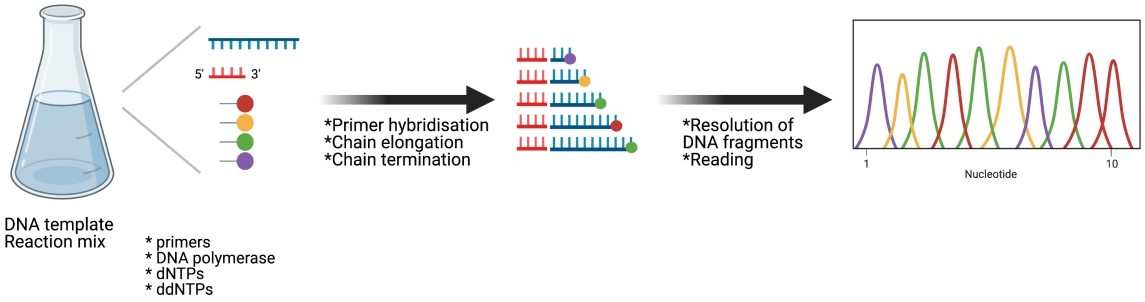


Figure 1.9: Sanger sequencing graphical protocol(*).

After roughly 40 years of Sanger sequencing, Next-Generation Sequencing (NGS), also called massive parallel sequencing, appeared at the dawn of the new millennium, unleashing a whole new area in sequence-driven research. Although NGS is constrained by the small size of read compared to Sanger sequencing (a couple hundreds vs. maximum 1,000 bp), the number of nucleotides it can process is increased by a 10^6-fold order of magnitude (couple kb vs. couple Gb) [90]. The second major change is conceptual: most NGS methods rely on sequencing-by-synthesis [91] instead of the chain termination method, meaning that the reading is done along with the polymerisation [92]. Finally, this revolution also enabled to shrink costs drastically (Figure 1.10).

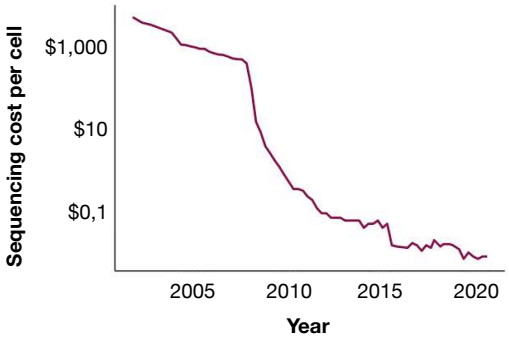


Figure 1.10: The sequencing cost has diminished exponentially over time (data from genome.gov).

The last step in this evolution, which is the scope of this doctoral work, is the combination of NGS with single-cell technologies, to extract the RNA count information on a per-cell basis: scRNAseq. Adoption has been really broad, happening in fields such as neurosciences [93], developmental biology [94], immunology [95] or oncology [96]. The reasons for this widespread use are the possibility to resolve heterogeneous cell mixtures, discover new cell types, or get mechanistic insights of physiological and pathological conditions.

1.3.3 How to perform single-cell RNAseq...

Since the first scRNAseq protocol in 2009 [97, 98], a plethora of tools has been developed for each step, from the cell isolation, the production of the libraries, to the sequencing, the alignment, and up to the analysis of the resulting count matrix, although there is no firm consensus on the best methods to use [99, 100].

1.3.3.1 ...from the wet experiment...

The outlines for the bench experiment are [101] (Figure 1.11):

1. Collection of the sample,
2. Digestion and preparation of a single-cell suspension,
3. Separation of individual cells via a microfluidic device or in wells,
4. RNA extraction,
5. Adjunction of molecular tags and reverse transcription (RT),
6. Amplification of the complementary DNA (cDNA) strand,
7. Fragmentation or tagmentation.

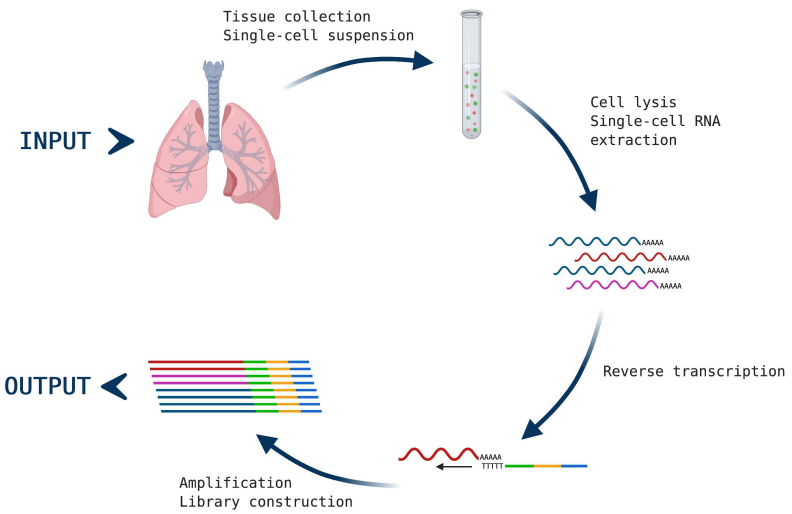


Figure 1.11: Wet part of the single-cell RNA sequencing protocol(*).

The global output of this experiment is a library of cDNA fragments ready for sequencing. The steps that are specific of single-cell handling include steps 2 to 5, while steps 6 and 7 stem from sequencing procedures. Let us discuss briefly these different steps, though omitting steps 1 and 2, which are out of the scope of this manuscript.

There are currently 2 methods for the high-throughput separation of individual cells (step 3.): droplet-based or plate-based procedures. In droplet-based methods, the single-cell suspension is mixed with barcoded beads: a microfluidic device co-encapsulates one bead with one cell. The nanoliter droplet is the factory where the cell is lysed and its

mRNAs captured and barcoded by the bead [102]. Then, the RT and the amplification are carried in parallel after breaking the droplets. In the plate-based method, the single-cell suspension is poured into a plate, at the rate of one cell per well. The pouring is performed either by FACS, for example in the Smart-seq2 protocol [103], or by mean of a microfluidic chip with the C1 platform [104, 105]. Again, the lysis and the binding of primers is performed in the well, as well as the generation of cDNA. Depending on the protocol, the next steps are carried on the plate (e.g. Smart-seq2 [103]) or the cDNAs can be pooled (e.g. CEL-seq2 [106]). For all protocols, the reaction mix lyses the cells and primes their RNA (step 4. and 5.). Primers contain a tail of poly-T that will bind to the polyadenylated tail of the mRNA. What is attached to this poly-T tail depends on the protocol: some include cell or well barcodes in order to do pooled reactions while retaining the information about the origin of each strand, or Unique Molecular Identifiers (UMIs) that enable the unique count of each reverse-transcribed mRNA, and a promoter for the amplification step [107].

The amplification (step 6.), by more than 1 million fold, comes in two flavors: it can be done with Polymerase Chain Reaction (PCR) or In Vitro Transcription (IVT). The latter requires a specific promoter, the T7 promoter, to recruit the T7 RNA polymerase that will synthesize a RNA fragment matching the cDNA template [108]. Thus, it implies an additional step of RT at the end of the amplification. The amplification is linear: for one strand of cDNA, we synthesize one strand of RNA at each cycle. On the other hand, the PCR would synthesize directly cDNA fragments from the template, thanks to a DNA polymerase and a PCR primer [109, 110]. Contrary to IVT, PCR amplification is exponential, since we double the number of cDNA strands at each cycle.

The final step for the preparation of cDNA libraries is the fragmentation or tagmentation of the strands, to make the shotgun sequencing itself possible (step 7.). In the tagmentation process, the cDNA strand is cut into fragments of 100-500 bp after PCR [111], while RNA fragmentation is done after IVT [106, 112], to make fragments of a couple of bp.

The different solutions are included in the several protocols available, such as **10X**, Drop-seq, Mars-seq, CEL-seq, Smart-seq, etc (Figure 1.12), and they are regularly updated and optimized. Constant progress in cell isolation automation, mRNA capture, RT and amplification has led to ever-increasing sequencing precision and datasets cardinality [86] (Figure 1.13).

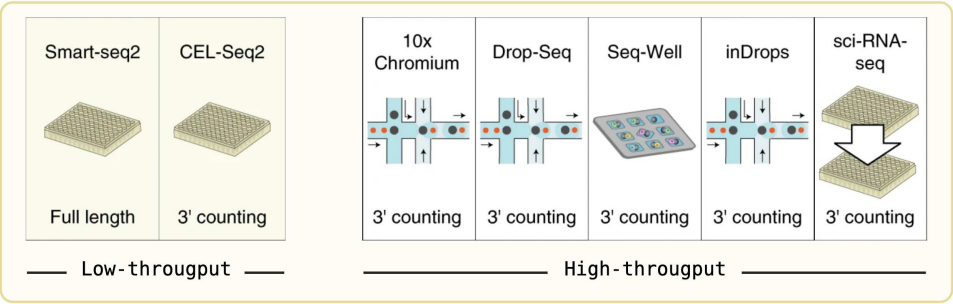


Figure 1.12: scRNAseq protocols. From [113].

The critical parts of the wet protocol lie in dealing with minimal amount of biological material, exacerbated by the fact that RNA is fragile. Because of the fragility, the conversion to a more stable DNA strand is required but lead to a first source of uncertainty, as

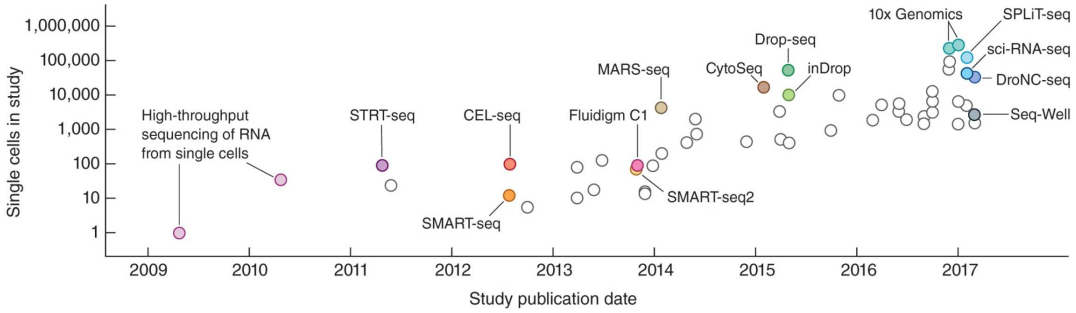


Figure 1.13: Datasets' cardinality increases exponentially. From [86].

the capture of mRNA is not fully efficient: according to different estimations, only 5-20% of the total mRNA is captured [101, 114]. It is unclear whether there is randomness or not in the capture efficiency of the different genes depending on their features: length, GC content, etc. A second source of uncertainty comes from the amplification step: depending on the amplification technique, there is a possibly infinite fold-change in the amplification rate of the different genes [115–118]. A key point in the downstream analysis of these data will be to consider that technical and biological noises coexist, especially when integrating datasets, in order to disentangle the two sources of noise [119–121] (paragraph 2.1.3).

1.3.3.2 . . . to the dry experiment

The dry protocol takes as input the cDNA libraries, while the outputs are diverse: depending on the research question, it can be trajectories, clusters, classes, new cell types, as well as new markers, etc [84, 85] (Figure 1.14).

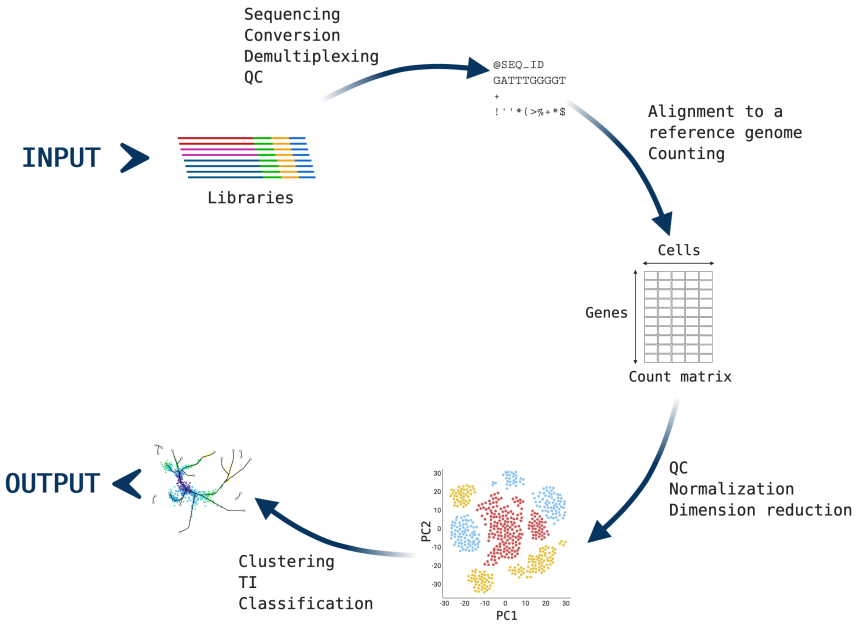


Figure 1.14: Dry part of the single-cell RNA sequencing protocol(*).

The sequencing step is mostly preempted by illumina machines, while other devices exist, such as SOLiD or 454 sequencers among others [122]. The raw sequencing data comes in the form of so-called BCL files. These files contain the nucleotides sequence,

and a phred score for each base call. BCL files are then converted, and eventually demultiplexed, into a readable format called FASTQ, that contains basically the same information as the BCL file. Demultiplexing is required whenever different libraries were mixed for the sequencing step, in order to separate them from each other. Then, FASTQ files undergo quality control (QC) in order to eliminate adapters and low quality bases before alignment. The alignment step compares the reads to a reference genome in order to map each read to a given gene, before counting the number of reads assigned to each gene. The count matrix, where one axis contains the cells or barcodes and the other the genes or features, is then used for downstream analyses [123, 124].

Although there is no analysis standards yet [99], the classical workflow includes preprocessing, dimension reduction, visualisation, cell assignment and gene identification. There is a myriad of tools for each of these steps (about 1,000 as of July 2021), which makes it difficult to navigate through the analysis and keep up-to-date, although the field is dominated by two platforms: Seurat [125] in R and Scanpy [126] in python. The choice of a platform, or other tools, as well as the programming language, conditions the analysis itself.

The preprocessing step includes QC, normalisation and batch correction whenever needed. The QC is meant to spot doublets or multiplets, which happen when two or more cells co-localize in the same droplet or well, and damaged cells. The basic strategy to identify multiplets is to remove barcodes with large numbers of counts and detected genes compared to the rest of the distribution, although more sophisticated strategies exist, such as Scrublet for example, which simulates doublets and then trains a classifier in order to detect real doublets [127]. A second QC filters out cells with a comparatively high percentage of mitochondrial counts, considered as damaged cells. A third QC eliminates cells with a low number of counts and detected genes, to further remove poor quality cells.

A last (or first) QC focuses on the genes, and is usually done when loading the count matrices, to remove genes that are present in no or $n << N$ cells, N being the typical size of a cluster. To help choose n, we can make a quick experiment. Let us assume that all cells that fall within the same cluster are rigorously identical, and that the capture efficiency is of c=10%, meaning we capture 10% of the total RNA in each cell. The estimated fraction of detected mRNA molecules for a cluster of size N will be:

$$\mathbb{P} = (1 - (1 - c)^N) \tag{1.1}$$

Therefore, in order to observe a reasonable fraction of the reads in each cluster, let us say 80%, we should collect a minimum of $N = 15$ cells per cluster (Figure 1.15). Obviously, this very simple model does not recapitulate the complexity of the data. In particular, the uniformity assumption is wrong, even if we can reasonably assume that it remains correct if we consider only the highly expressed genes, which are the genes of interest, while the remaining genes would contain most of the transcriptional stochasticity (or biological noise). Also, the genes of interest are expressed at high levels in the corresponding cluster and therefore suffer less from dropout (or technical noise). Thus, we come up with an updated description of the capture of marker genes, by updating c with a higher value $\tilde{c}$:

$$\mathbb{P} = (1 - (1 - \tilde{c})^N) \tag{1.2}$$

Hence, choosing for example $\tilde{c} = 0.2$, we need a minimum of $N = 7$ cells per cluster to collect 80% of the reads (Figure 1.15). Finally, considering the default value to define marker genes in the Seurat platform, i.e. a gene that is detected in at least 10% of the

cells in the cluster of interest, and making a last assumption that all k clusters have the same size, and all cells have the same amount of marker reads M, we refine further our model:

$$\mathbb{P}_{k,N,M,\tilde{c}} = \left[1 - \sum_{j=0}^{\lfloor \frac{N}{10k} \rfloor} \left[1 - \frac{\binom{M\frac{k-1}{k}}{M\tilde{c}}}{\binom{M}{M\tilde{c}}} \right]^{j} \left[\frac{\binom{M\frac{k-1}{k}}{M\tilde{c}}}{\binom{M}{M\tilde{c}}} \right]^{1-j} \right]^{k} \tag{1.3}$$

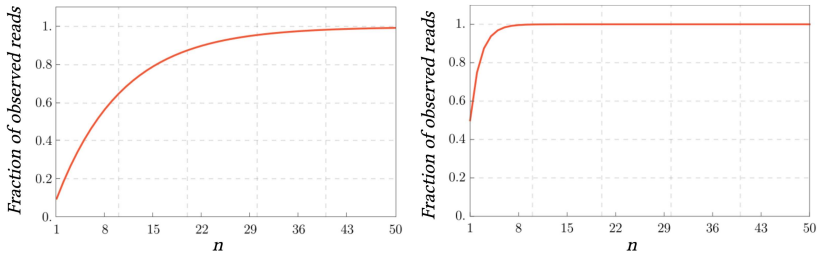


Figure 1.15: How many cells per cluster do we need in order to have a given amount of information? Considering the model from equation 1.1 (left panel) or equation 1.2 (right panel).

Unfortunately, equation 1.3 is computationally intractable, so we would stick to the value of a minimal cluster size of $N = 7$. Thus, a reasonable value for n could be around 2-4, while the default value in Seurat and Scanpy is $n = 3$.

This small experiment illustrates the difficulty of setting user-based parameters in the analysis: more generally, the QC is based on a manual choice of appropriate data-dependent thresholds regarding the number of genes and cells, the count depth (or library size) and the fraction of mitochondrial RNA, and as such should be permissive, in order to not eliminate cells or genes of interest, even if it means going back to the QC step later in the analysis [128]. It happens often since the required quality cannot be determined beforehand, but from the evaluation of downstream performances. Indeed, in some situations, cells with a large library size might be simply large cells, while quiescent cells would have a small library size, among other examples. If a significant proportion of cells is filtered out, let us say above 10%, or above the expected doublet rate for example, it might indicate an overall poor quality of the data, but also could be due to the biology of the cells. After the threshold-based QC, one might also compute a Principal Component Analysis (PCA), or alternatively an Uniform Manifold Approximation Projection (UMAP) to remove subsequent outliers [99, 129].

After QC, the next preprocessing step is normalization. The intentions behind normalization are:

- ↬ to correct for technical variation,
- ↬ to reduce zero-inflation, and eventually make the data look more like normally-distributed,
- ↬ required for dimension reduction.

There are 2 major categories of normalization tools: linear or non-linear. Linear tools attempt to correct the count depth between cells, assuming that they all have an *a priori* similar library size, while non-linear tools attempt at describing the data distribution over several assumptions. The linear tools derive mostly from bulk, and correct for technical variation. The second normalization step is a log-transformation of pseudo-counts

(counts+1), in order to reduce zero-inflation and the skewness. Finally, there is an optional normalization over the genes, sometimes required for dimension reduction, although its utility is still disputed.

Briefly, three other optional steps include batch correction, regression of biological effect such as sex or cell cycle, and imputation. The batch correction is applied in order to eliminate technical drift between two different batches that would not co-localize otherwise (Figure 1.16). The regression step is meant to eliminated confounding factors that would blur the downstream analysis by adding unwanted variability. Lastly, imputation is sometimes performed as a way to correct for the dropout. It usually takes advantage of neighbor cells to reconstruct missing information.

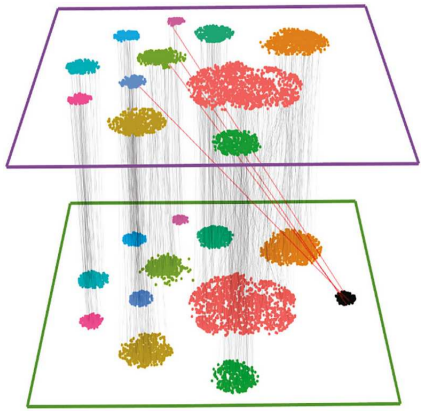


Figure 1.16: Comparison and matching of two batches for batch correction. From [121].

Once the preprocessing is done, there is the dimension reduction step, generally after feature selection such as Highly Variable Genes (HVG) extraction. HVG selection is meant to select interesting genes, remove noise and speed up computation. It is classically done by choosing the genes with high ratio variance over mean in expression bins. The most common dimension reduction tool is PCA, or a flavor of PCA adapted to single-cell data [130]. The reduced data is the basis for most of the further applications : visualisation, clustering, trajectory inference, while we usually go back to the raw data for Differential Gene Expression (DGE) analysis.

Chapter 2

Single-cell RNAseq analytical challenges

scRNAseq data belongs to the field of big data, as it is characterized by the measure of thousands of features over thousands of cells. This huge amount of data is a double-edged sword [131–133]. Bioinformaticians, when dealing with the data, face 3 main challenges: the noise, the dropout and the high dimension. To tackle these challenges, researchers came up with a myriad of tools, although there is no consensus on the best strategy [99]. So far, the guideline has been to design one pipeline per research question and dataset, in order to take into account the high variability of scRNAseq data.

2.1 Modelling noises

2.1.1 Biological noise

We already evoked in the preceding chapter the existence of biological noise, but we shall now explain what it is. There is an obvious "macroscopic noise" in gene expression, exemplified by the fact that cells from a same organism, with the same genetic code, have different fates: an epithelial cell, a muscle cell or a neuron from a unique individual have the same DNA, but completely different shapes, organelles, protein content and functions. It has been suggested that the "microscopic" noise [134], i.e. the variability in gene expression of supposedly identical cells, participate to the "macroscopic noise" [135]. In [136], the authors consider 4 sources of variability:

- ↬ the stochastic nature of biochemical reactions governing the process of transcription-translation-degradation (i),
- ↬ the internal state of the cell, e.g. cell cycle (e),
- ↬ external stimuli, e.g. homeoprotein gradients (e),
- ↬ point genetic mutations (e).

There is a further distinction, viz. between intrinsic and extrinsic noise. Intrinsic noise will affect co-regulated genes within one cell, while extrinsic noise will affect differently two cells, either at a global level or at the level of a specific pathway. Each noise can be related to one of the four sources of variability (indicated by (i) for intrinsic or (e)

for extrinsic in the list above), and the extrinsic noise usually exceeds the intrinsic one [135, 137].

This variability in gene expression is readily observed with techniques such as Fluorescence Activated Cell Sorting (FACS): different cells have different levels of expression as illustrated by the spread of the dot plot (Figure 2.1). Noise can have no or dramatic effect depending on the magnitude and duration of the fluctuation. For example, it has been hypothesized that intrinsic noise govern the development of the sense of smell, by influencing the choice of a single odorant receptor in each olfactory neuron [138]. In this situation, noise is an advantage, while it can also blur the accuracy of cellular processes, by perturbing the temporal shape of proteins or mRNAs expression, at the single-cell or at the population level [139].

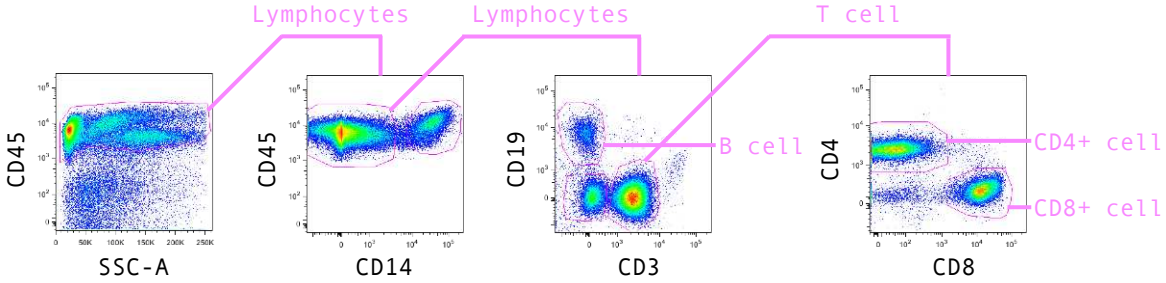


Figure 2.1: Variability in gene expression is observable with FACS.

We mentioned that "microscopic" noise may participate to what we called "macroscopic noise", which is merely plasticity and leads to cellular heterogeneity. Although it is not completely clear how noise is controlled [137], one of the possible mechanisms is that it depends on the stress [134, 136]: noise increases with stress, as a way to increase the chances of reaching a fit phenotype. This is reconcilable with the rigid map of Waddington's landscape [140], by extending it while imagining a flattening of the peaks under stress, that would create a sea of phenotypes instead of separated flows (Figure 2.2) [135]. Another putative control mechanism is to duplicate genes in order to smooth the associated noise [134, 137].

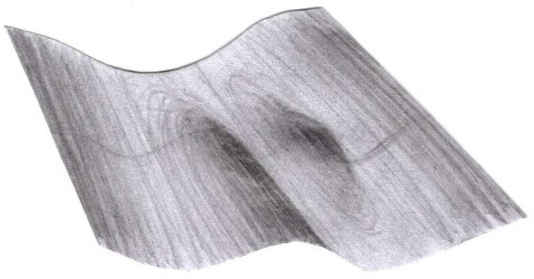


Figure 2.2: Modified Waddington's landscape to include expression noise by flattening peaks. Courtesy of J. Amir Tahmasseb.

Regarding scRNAseq, RNA expression is noisy because of its production and degradation rates: RNA is produced during transcriptional bursts, periods of intense transcription, followed by transcriptional inactivity, and degraded during both stochastic and

deterministic processes. The transcription efficiency depends on the burst frequency (ratio of the duration of transcriptional activity to inactivity) and burst size (number of transcripts made during the burst) [135].

2.1.2 Technical noise

However, even without considering the fact that the measurand, i.e. the object and the quantity being measured, is intrinsically noisy, every measure suffers from noise, be it the measure of the Earth-Sun distance (149,597,870.7 ± 8,000,000 km) or of the radius of a proton (0.83 ± 0.1 fm). Measure uncertainty has three causes.

The first cause relates to quantum mechanics and was first reported by Heisenberg (1901-1976) [141]. It states that it is not possible to determine with exactitude simultaneously the position and the speed of an wave-like object. More precisely, if one wants to increase the precision of the position measurement σ_x, it will be at the expense of the speed measurement accuracy σ_s, and vice-versa:

$$\sigma_x \sigma_s \geq \frac{\hbar}{2}, \tag{2.1}$$

$$\hbar \approx 1.054 \times 10^{-34} \text{ J·s},$$

where $\hbar$ is the reduced Planck constant. The so-called Heisenberg's uncertainty principle is observed in all wave-like systems, which is not the case of RNA.

The second cause of uncertainty is called the observer effect, and refers to the fact that performing a measure usually alters the measurand. Famous thought experiments testing this effect are Schrödinger's cat and Wigner's friend [142], while this effect is also observed in other fields such as social sciences, where the measure depends on the biases of the observer and the measurand: e.g. the Hawthorne effect (survey participants would modify their behavior because they know they are observed) or the observer-expectancy effect (observers would modify unconsciously the behavior of survey participants). In our case, the observer effect happens mostly at the very first step of the wet lab experiment, when sampling a tissue and its RNA.

The last cause of uncertainty relates to the finite precision of the measuring instrument and of the operator [143]. This effect can be systematic and random, and produces a probability distribution of successive measurements.

In the scRNAseq field, the technical noise is mostly visible via imprecise read counting. It comes from basically all steps of the wet protocol: an inappropriate sampling and digestion of the tissue of interest (observer effect), the poor capture of RNA fragments (finite precision of the technique), as well as errors in RT (observer effect) and amplification (finite precision of the technique). The critical and most noisy step is probably the RNA capture, since only a fraction (5-20% depending on the estimations) of all fragments of a cell are observed, while it is not clear yet whether the capture is random or not [135]. It is even worse with decreasing amounts of starting biological material [144]: less RNA material leads to an increased amount of noise. Hence, genes with low read count exhibit a stronger noise than genes with a high read count, leading to within-cell noise variability. There is also a between-cell variability in RNA capture, exemplified by the differences in library size. The second critical step is the amplification, which causes non-linear distortions, and this is again especially the case for poorly abundant genes [145]. This deformation of the count matrix, as compared to an expected amount of counts,

severely impacts the downstream analysis, for example the DGE analysis, or the pooling of different datasets suffering from batch effect.

2.1.3 Discriminating technical from biological noises

We mentioned already that the critical advantage of scRNAseq over bulk lies in the possibility of exploring differences between seemingly similar cells. Therefore, it is of paramount importance to be able to distinguish biological from technical noise in order to take advantage of the single-cell approach [146]. Different approaches have been implemented in order to disentangle meaningful biological variations from detrimental technical discrepancies, that we classify into three categories.

The first family of strategies considers each cell × gene data point as a random variable and tries to fit it with parametric statistical models, usually making an assumption over the technical variability, in order to spot biology-related overdispersion. More precisely, it is classically observed that there is a quadratic polynomial relation between the mean μ_i and the variance σ_i (or the coefficient of variation $CV = \frac{\sigma}{\mu}$) of all d genes $\mathcal{G} = \{g_i; i \in [\![1, \dots, d]\!]\}$. This polynomial relation is well recapitulated by a Poisson distribution [135, 144–146], a negative binomial [120, 147] or a log-normal distribution [73] (Figure 2.3).

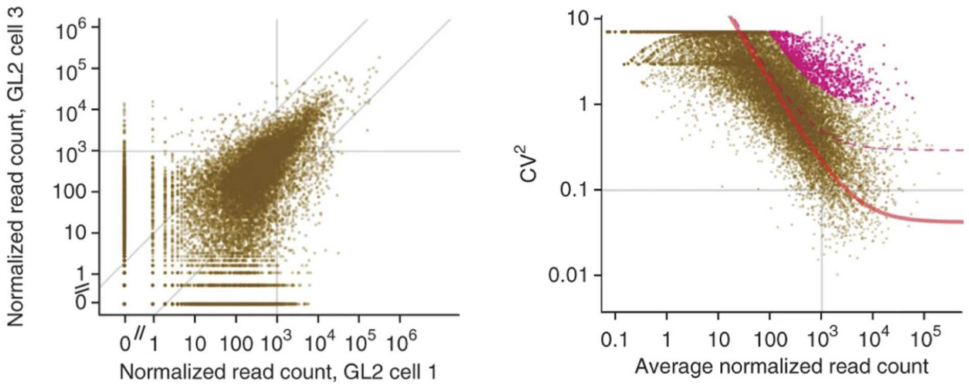


Figure 2.3: Observed scRNAseq counts can be statistically modelled. Scatter plot for normalized read counts for all genes $\mathcal{G}$ from two cells (left panel). CV-Mean scatter plot with HVG highlighted in magenta, as genes that significantly deviates from a Poisson distribution, i.e. well above the dashed line (right panel). From [144].

The second family jumps directly to the interpretation by extracting only biologically meaningful genes: for example, one can extract correlated gene sets, compute a PCA with the latter genes and select the gene sets associated to overdispersed PCA, i.e. PCA that would explain a higher percentage of variance with its first PC than would a PCA with a random gene set [148].

The last family relies on a fairly new possibility: denoising the data using generative autoencoder neural networks, which are praised as a more universal and flexible approach. The network learns simultaneously a low-dimensional representation of the data (encoding or convolution) and how to infer the data back in the original space (decoding or deconvolution) [149–152]. Briefly, it works by minimizing a cost function that measures the drift between the original count matrix and the low-dimensional representation. Concurrently, it learns the method used to make the low-dimensional representation and invert it, in

order to reconstruct a denoised count matrix in the original space, from the denoised projection. While it does not explicitly model the technical versus biological noise, this approach has proven its validity and effectiveness, especially since it is an all-in-one approach: the encoded matrix is used for clustering, trajectory inference or visualisation, and the decoded matrix is used for imputation or DGE analysis (Figure 2.4).

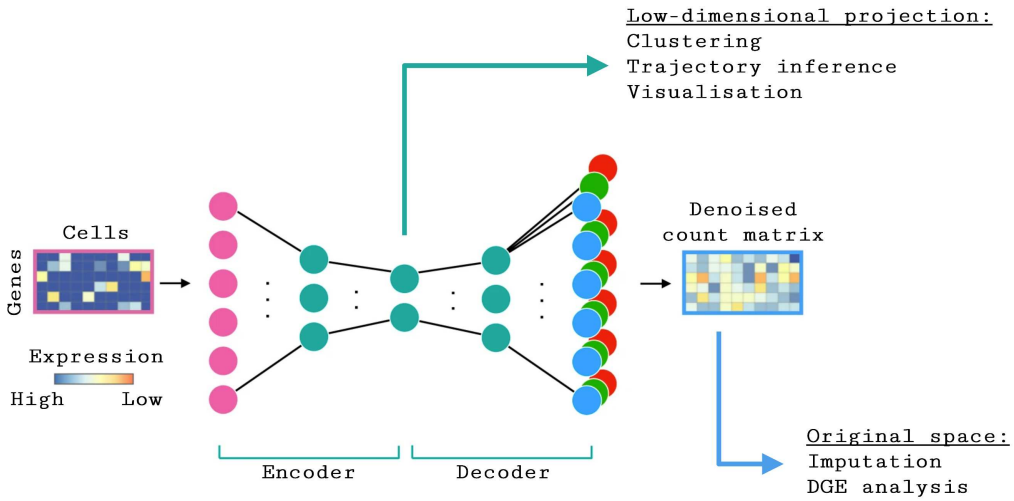


Figure 2.4: Autoencoder neural network adapted to scRNAseq data. Adapted from [150].

Except for the autoencoder strategy where all the steps of the analysis are performed simultaneously, the disentangling of the technical from the biological noise is done during the preprocessing, most specifically during the normalisation step that would include the parametric noise models (Figure 1.14). Some normalisation approaches can be straightforward, such as taking into account the mere library size of each cell. On a side note, we did not mention in this section imputation methods, such as MAGIC [153] or SAVER [154], that also rely on a modelling of the data. Imputation algorithms also aim at correcting technical noise by targeting one specific aspect, the dropout. We will discuss these methods in the next section, along with the sparsity challenge.

2.2 Sparsity

Sparsity is a peculiar aspect of technical noise and is a prominent challenge in the analysis of scRNAseq data [131].

2.2.1 Reminder on the terminology

Let us first echo the remark made in [146] about the fact that there is an imprecise terminology in the scRNAseq field.

Dropout refers to the zeros in scRNAseq data but means either observed zeros (i.e. all counts x_{ij} that are effectively null in the observed count matrix $\mathbb{X}$), or artificial additional zeros (i.e. genes that were expressed but not detected at all), or even all underestimated counts. In this manuscript, I will use the term *dropout* with the meaning of additional zeros in the observed count matrix $\mathbb{X}$. Sparsity, coined by the economist Harry Markowitz (1927-) is the percentage of observed zeros in $\mathbb{X}$. Hence, we cannot distinguish in sparse counts true zeros from dropouts. Finally, missing data would represent all reads that were not detected, leading to gene counts below their true levels. Another important distinction should be clarified regarding scRNAseq models, as it causes more confusion. Using the classification defined in [146], there are three types of models:

- measurement models $p(\mathbb{X}|M)$, connecting $\mathbb{X}$ to the true expression matrix M,
- expression models $p(M)$, modelling M,
- observation models $p(\mathbb{X}) = p(\mathbb{X}|M) \cdot p(M)$, combination of an expression and a measurement model, to model $\mathbb{X}$.

2.2.2 Zero-inflated...

scRNAseq count matrices suffer from a high sparsity: a typical count matrix would be at least 50% sparse (Figure 2.5). The magnitude of sparsity depends on the scRNAseq platform used and the sequencing depth: high-throughput droplet-based experiments are sparser than plate-based experiments, with a sparsity that can approach 100%. It also depends on the true gene expression levels, with a higher chance for poorly expressed genes to be dropped (Figure 2.5).

From this observation, one could easily jump to the conclusion that there is a zero-inflation, i.e. a higher proportion of zeros than expected, although an unknown fraction of the zeros are due to the genuine biology. If this is true, zero-inflation should be taken into account, either by adding a zero-inflated component in the models, or by imputing missing data and hence correcting for dropout.

2.2.2.1 Modelling zero-inflation

Many models that attempt at correcting technical noise include a zero-inflated component, e.g. by using a Bernoulli process that decides whether the real count is observed or dropped, on top of the technical noise component.

For example, in the Splatter package, their framework includes a zero-producing component to simulate scRNAseq data [155]. More precisely, dropout is optionally added

Name	GEO code	Cell number	Gene number	Method	Year	Species	Droplet/plate
Freytag	GSE115189	3372	58302	10X	2018	Human	Droplet
Tian	GSE111108	4000	33456	10X	2018	Human	Droplet
Baron1	GSE84133	8569	20125	inDrop	2016	Human	Droplet
Baron2	GSE84133	1886	14878	inDrop	2016	Mouse	Droplet
Zeisel	GSE60361	3005	19972	C1	2015	Mouse	Plate
Guo	GSE99254	12346	23370	Smart-seq2	2018	Human	Plate
Zhang	GSE108989	8530	23370	Smart-seq2	2018	Human	Plate
Zheng	GSE98638	3636	23389	Smart-seq2	2017	Human	Plate

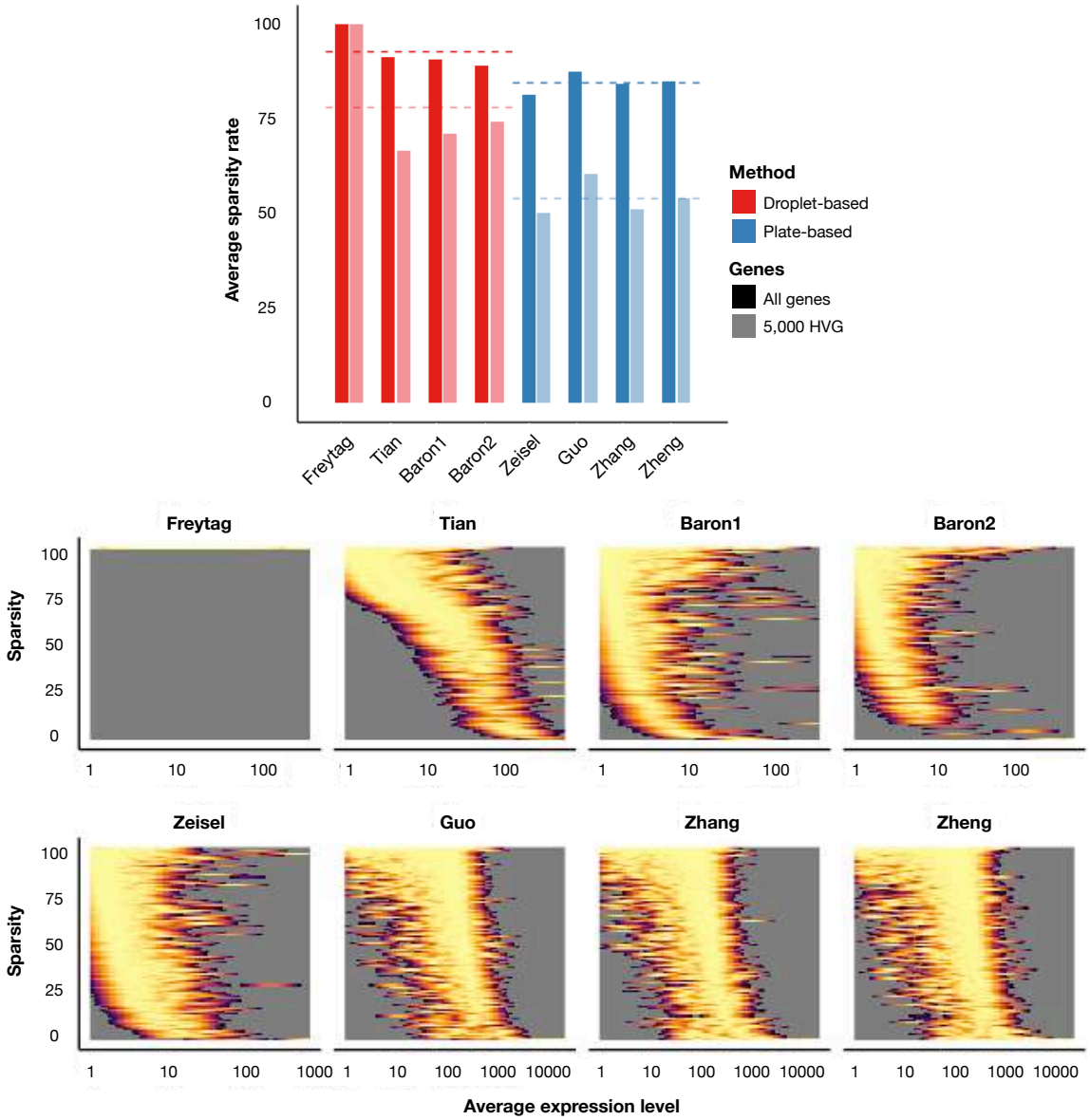


Figure 2.5: scRNAseq count matrices suffer from sparsity at high rates. Example with 8 datasets (top panel). They have a high sparsity rate, even higher for droplet-based assays, and even when keeping only 5,000 HVGs (middle panel). The sparsity strongly depends on the average expression level for each gene (bottom panel).

on top of signal simulation with a logistic regression in order to decide the fraction of

counts to drop per gene g_i, and the counts $x_{ij,j\in\{[\![1,\dots,n]\!]\}}$ of gene g_i across the n cells are dropped following a Bernoulli distribution.

On the other hand, ZIFA [156] models the technical noise with a Gaussian component while dropout is added, s.t. it follows a square exponential decay as a function of the mean gene expression level:

$$d_i = \exp(-\lambda \mu_i^2), \tag{2.2}$$

where λ is a parameter to tune and μ_i the average expression level of gene g_i calculated on log-transformed non-zero pseudocounts (counts+1).

Other examples includes ZINB-WaVE [120], SCDE [145] or MAST [157], among others.

2.2.2.2 Imputation

Instead of modelling the zero-inflated component, one could also attempt at recovering missing data by performing imputation. There are currently 3 classes of imputation methods.

The first class focuses actually only on dropout events and not on the whole spectrum of missing data. It identifies dropouts to correct them, by modelling the data generation mechanism and attributing to each count x_{ij} a probability of being a dropout. Such methods include for example SAVER [154], scImpute [158] or Biscuit [73].

The second class smooths the data by averaging cell expression profiles, taking advantage of the information brought by neighbouring, hence supposedly similar, cells. MAGIC is one example [153]. The main weakness of this approach is the circularity of information though: using noisy data from other cells to reconstruct an expression profile is indeed flawed as it can generate false positives or irreproducible differential expression. This can be avoided by adding external information [159].

Finally, the third class reconstructs a denoised and imputed count matrix from a latent representation, obtained either via matrix factorisation, or with neural networks: DCA [150], scVI [149] or ZIFA [156] fall into this category (Figure 2.4).

2.2.3 . . . or not ?

However, and following an intriguing blog post from Valentin Svensson entitled "Droplet scRNAseq is not zero inflated"[1], the fact that scRNAseq is zero-inflated is itself questionable, especially when considering UMI data [130, 160] (Figure 2.6).

Indeed, while the data exhibits a considerable amount of zeros due to both biological zeros and dropouts, one could argue that it is also 1-inflated (because of all genes that had more than 1 count in a cell but for which only 1 read has been detected), 2-inflated, 3-inflated, etc. Additionally, there is nothing in the wet lab protocol that would justify the existence of an independent zero-producing mechanism [146]. Finally, following the precise terminology about the different ways of modelling the data mentioned above (section 2.2.1), and referring to measurement models, a non-zero-inflated Poisson distribution or a negative binomial one would nicely account for the apparent zero-inflation of the data (but hence also for the 1-inflation, 2-inflation, etc) [146, 161]. Of note, it does not rule out the validity of zero-inflated models, for example for expression of observation models.

[1] https://www.nxn.se/valent/2017/11/16/droplet-scrna-seq-is-not-zero-inflated

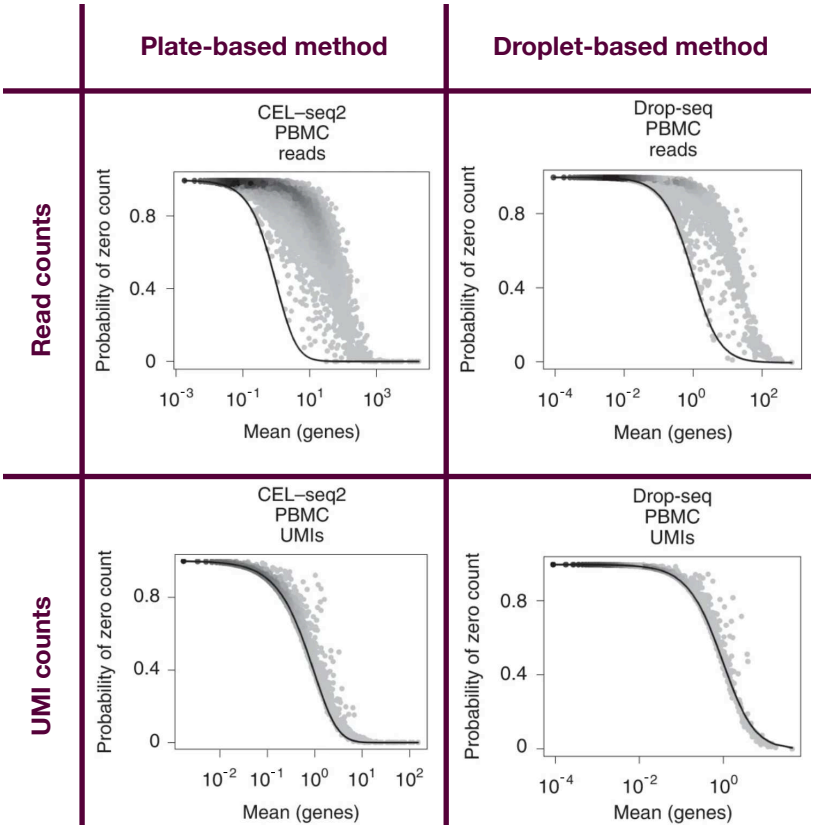


Figure 2.6: UMI count data is not zero-inflated, as opposed to read count data: UMI tags allow to trace back reads to the original mRNA molecule, thus correcting partly for the amplification bias that artificially widens the gap between zero and non-zero expressed genes. From [160].

While non-zero-inflated measurement models are verified for UMI counts data, non-UMI count matrices still exhibit zero-inflation: UMI counting deflates the amplification bias that causes the unexpectedly high proportion of zeros by collapsing all the reads referring to one RNA molecule, but it is not the case for read count data (Figure 2.6). With that respect, different measurement models are used for the mean-sparsity relationship of read counts data, such as the Michaelis-Menten function [162]. As we mentioned in section 2.1.3, these models are then used to distinguish technical noise from biological variability.

Besides from sparsity, there is another observation: the data is scarce. The scarcity comes from the data being high-dimensional: for each measurand, which is a single cell in our case, we measure thousands of features. While traditional techniques would quantify a couple of features, technological progress, be it in storage capacities, analysis power or ever-increasing quantification performances, led to the emergence of a new field called big data, that we shall discuss in the section below.

2.3 Dealing with high dimensional data

Big data, along with machine learning (ML) and deep learning (DL) methods, are buzzing words, and encompass quite broad and fuzzy meanings.

2.3.1 A new type of data

With technological advances, we produce and store more and more data. Because of the whopping amount of features being collected, such data cannot be analysed manually but require a computer or even a computing infrastructure. Let us take an illustrative example with the autonomous vehicle. Such a car needs to be able to analyse what is in front of it: crossing roads, pedestrians, trees, sidewalks, etc. It will do so by collecting images from its environment. The collection of consecutive images is a huge matrix, one row per image and one column per pixel, as each image can be described by its pixels' sequence. A typical matrix would contain an order of magnitude of 10^6 pixels times the number of images in the movie, around 10^3 images per minute. One quickly grasps that this huge amount of data is not analysable by a human operator, furthermore in a short amount of time as to make quick decisions when driving a vehicle. Such datasets belong to the field of big data, which includes areas such as weather forecast or robotics.

Coming back to biology, even considering longstanding techniques such as FACS, current FACS apparatuses measure several colors, hence representing tens to hundreds of features. In the case of omics data, we collect thousands of features, and also the number of cells per study increases, reaching nowadays a couple of millions of cells[2] (Figure 1.13). The massive size of count matrices, 10^4 genes $\times 10^3 - 10^6$ cells, requires a computational analysis.

2.3.2 New analytical methods

ML usually refers to all the methods designed to handle big data, in which the computer learns to recognize striking features that are useful to form groups in the data. DL is a subsection of ML, grouping algorithms that use neural networks. Supervised methods use labels, while unsupervised approach try to preserve the original structure.

In the case of scRNAseq, ML and DL algorithms are used for all the steps of the analysis: dimension reduction, clustering, visualisation, trajectory inference, etc.

When dealing with big data, not only the tremendous size of the datasets poses challenges, but also the fact that high-dimensional spaces are suffering from a set of strange and counter-intuitive phenomena that we discuss below. These phenomena are specific of the high-dimension, and do not happen in the regular 3-dimensional everyday space.

2.3.3 Curse and blessings of dimensionality

2.3.3.1 Definition

Datasets with thousands of observations and even more features are a double-edged sword. The obvious interest of big data is the massive resource it represents. Unfortunately, we

[2] https://www.nxn.se/single-cell-studies/gui

are currently not able to deal with this amount of information, especially since it is not only too complex for the human brain, but also noisy. There are in fact two sources of noise in high-dimensional settings:

- ↬ the "classical" noise, that we already discussed before in section 2.1, i.e. the noise due to the measure as well as the noise intrinsic to the measurand,
- ↬ a "dimensional" noise, encompassing all phenomena happening in the big data universe, but not in low-dimensional spaces.

These blurring dimension-related phenomena are termed as the curse of dimensionality, which was coined by Richard Bellman (1920-1984). The best known effect is the measure concentration. This refers to the fact that the range of values for pairwise distances shrinks with the dimension. In other words, all pairwise distances become similar, making it hard to distinguish similar from dissimilar data points as their contrast vanishes. Luckily, there is another side of the coin: as the geometry is simplified in high-dimensional spaces, there are also positive effects, called blessings of dimensionality [163]. For example, all data points are linearly separable. Hence, it actually depends on the downstream application whether the high-dimensional geometry will be a blessing or a curse. In the scRNAseq situation, since we are trying to group similar cells, the analysis will rather suffer from the curse of dimensionality (Figure 2.7) [128].

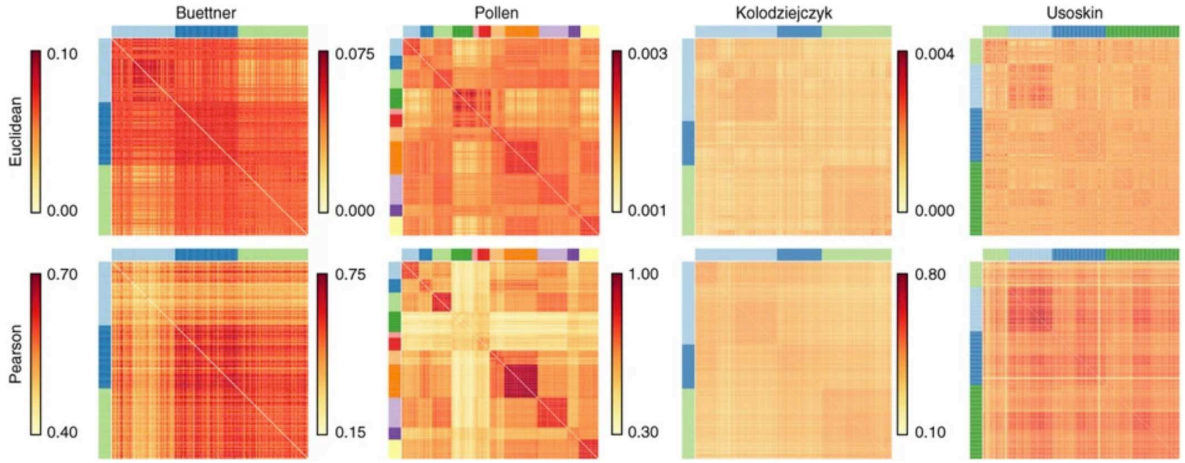


Figure 2.7: Measure concentration in scRNAseq datasets, using pairwise Euclidean distance and Pearson correlation. Colored bars on top and left of each heatmap indicate the ground truth labels, showing that the inter-group similarity is not unequivocally higher than the intra-group one. From [164].

We mentioned measure concentration, but there are other effects of the curse. One of them is that big data is usually sub-sampled or scarce. To reasonably sample a 1-dimensional segment, let us say we need ca. 10 points. For a 2-dimensional plane, we would need 10^2 data points. To sample a d-dimensional volume, we need 10^d data points. No dataset has such a high cardinality, thus they are all scarce.

Another effect is the hubness phenomenon. While training music recommendations algorithms, it has been observed that some tracks would be recommended particularly often, or rather at an abnormally high frequency [165]. It means that in the corresponding directed neighbors graph, some data points are neighbors of many other points. More precisely, one can compute in a k-Nearest Neighbors (k-NN) graph the in- and out-degree

of each node. The distribution of in-degrees gets skewed to the right when the dimension increases. All the points in this right fat tail have a surprisingly high in-degree, i.e. an in-degree $d_i >> k$, k being the value used to build the k-NN graph, and the expectation for the in-degree value. This is the manifestation of what is called the hubness phenomenon, and data points in the right fat tail are called hubs [166]. Since the k-NN graph is distorted, it is expected that "hubby" k-NN graphs would lead to a worse performance of k-NN graph-based algorithms.

2.3.3.2 Avoiding effects of the curse of dimensionality

A reasonable strategy, and the one that has been chosen so far in scRNAseq analysis, to minimize the detrimental effects of the high dimension has been to reduce the dimension. As depicted in a recent tutorial on scRNAseq analysis [99], feature selection and dimension reduction is an inevitable step of the analysis. This is especially true for visualisation where the number of dimensions is reduced to 2 or 3.

There are several recipes for the feature selection (that usually leaves the dimension of the same order of magnitude), and the dimension reduction (for which we usually keeps a couple of tens of dimensions). This strategy has the advantage of tackling all noxious effects of the high dimension, while it poses the risk of loosing information contained in the dimensions that are removed. This is worrying in our single-cell case, compared to bulk: while the first few principal components (PCs) of the data explain a consequent amount of the variability in bulk data, this is not true anymore for single-cell data.

The feature selection has 2 technical purposes: reducing the size of the count matrix, and speeding up downstream dimension reduction algorithms, and one biological purpose: selecting biologically relevant genes [128]. To this end, interesting features are selected according to their mean expression and variance. The user has to choose a threshold, either on the number of HVGs to retain (e.g. 2,000 in the Seurat pipeline), or on the authorized interval for mean expression and variance. Selecting more HVGs might result in a higher noise, but reduces the risk of removing biologically relevant information [128].

A common method for the dimension reduction has been PCA and its flavors such as GLM-PCA [130], but it is included in a wider field of research, viz. manifold learning. The manifold assumption hypothesizes that the original data lies onto a significantly lower-dimensional manifold that would recapitulate it perfectly. This hypothesis is supported by the intuition that single-cell expression profiles depend on a set of limited and redundant molecular reactions and coordinated gene modules, a weighted combination of which yields the different cell states and types [167].

2.3.3.3 Targeting the measure concentration effect

Apart from reducing the dimension, there is another ruse to mitigate the measure concentration. It has been observed that some metrics would be less sensitive to this phenomenon: for example the L_p (quasi)norm of the form (Figure 1):

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p} \tag{2.3}$$

$$x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$$

is all the more sensitive as $p \in \mathbb{R}^{+*}$ is high [168].

A budding field is the area of metrics learning, while it is not conceivable to implement it for scRNAseq since it is mostly a supervised approach and the majority of the datasets do not have ground truth labels. Fortunately, there are other options, such as constructing a corrected affinity matrix by using kernel functions, as done in SIMLR [164, 169], for which we need to choose the number of clusters though, or by using alternative metrics [168], although it is still debated whether it would indeed improve the analysis [170].

2.3.3.4 Targeting noise in the data

Under the manifold assumption, noise is considered as a dimension-related perturbation that is eliminated upon dimension reduction. In the case of PCA, there are a few rules of thumb to choose the number of PCs that should be retained in order to retain signal and remove noise: the elbow plot, the jackstraw procedure, or an *a priori* choice (Figure 2.8). The elbow method relies on spotting an elbow in the plot of the percentage of variance explained by each PC, assuming that further PCs bring negligible information. The jackstraw procedure attributes a p-value to each PC, in order to choose significant PCs. An *a priori* choice can be e.g. to keep 50 PCs, or to keep the n PCs that explain more than a tenth of the variance explained by the first PC, n being considered as the intrinsic dimension, i.e. the dimension of the lower-dimensional manifold. There are other heuristics for other dimension reduction tools such as the Independent Component Analysis [171].

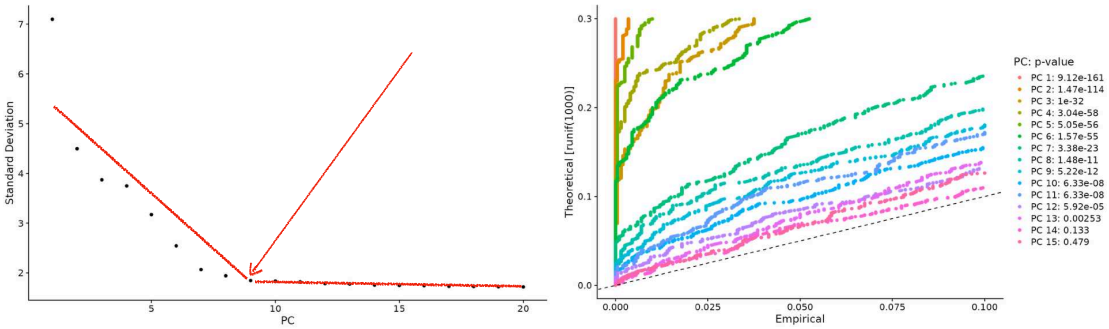


Figure 2.8: Choosing the number of principal components to retain. Using the Elbow plot (left panel) or the Jackstraw plot (right panel). From Seurat website.

There is a last method, that do not relies on heuristics, but on a theory that attempts at describing the behavior of random matrices, viz. the Random Matrix Theory (RMT). The rationale is to consider the count matrix as a mixture of a random matrix and a matrix containing the biological signal. Hence, all PCs, or eigenvectors, that deviate from the behavior predicted by RMT are considered as a part of the signal-containing matrix, while the eigenvectors following the Marcenko-Pastur distribution can be discarded: this approach specifically aims at distinguishing signal from noise in the count matrix [172].

2.3.3.5 Avoiding hubness

Regarding hubness, there are specific techniques aimed at reducing it. The idea behind these methods are to make each node in the k-NN graph wobble slightly according to their in-degree: a node that has a high in-degree wobbles towards a less dense region in order to loosen its links to neighbor nodes, while a node with a small in-degree wobble

towards a more dense region in order to strengthen its links with other nodes. There is currently a total of 4 hub reduction graph-correcting methods: Mutual Proximity (MP), Local Scaling (LS) and a variant LS_{nicdm} and DisSimilarity Local (DSL) [173].

MP models pairwise distances $d_{i,j\in\{1,\dots,n\}\setminus i}$ of a set of n points with random variables X_i that depict the distribution of distances between x_i and all other points:

$$\text{MP}(d_{i,j}) = 1 - P(X_i > d_{i,j} \cap X_j > d_{i,j}) \tag{2.4}$$

where P is the joint probability density function.

LS takes into account the local neighborhood:

$$\text{LS}^k(d_{i,j}) = 1 - \exp(-\frac{d_{i,j}}{r_i^k}\frac{d_{i,j}}{r_j^k}) \tag{2.5}$$

where k refers to the size of the local neighborhood, and r_i^k is the distance of x_i to its k-th neighbor.

The variant LS_{nicdm} uses the average distance to the k neighbors instead of the mere distance to the k-th neighbor:

$$\text{NICDM}^k(d_{i,j}) = \frac{d_{i,j}}{\sqrt{\mu_i^k \mu_j^k}} \tag{2.6}$$

where μ_i^k is the average distance of x_i to its k nearest neighbors.

DSL uses local centroids $c^k(\bullet)$ to reduce hubness:

$$\text{DSL}^k(x_i, x_j) = \|x_i - x_j\|_2^2 - \|x_i - c^k(x_i)\|_2^2 - \|x_j - c^k(x_j)\|_2^2 \tag{2.7}$$

where the local centroid is estimated as the barycenter of the k nearest neighbors of x_i:

$$c^k(x_i) = \frac{1}{k}\sum_{x_j \in \text{kNN}(x_i)} x_j$$

MP uses all data points to correct for hubness, while DSL uses local centroids and LS and its variant LS_{nicdm} local neighborhoods. All these methods output a less "hubby" k-NN graph that can then be inputed in the various algorithms used in scRNAseq analysis such as community-detection-based clustering or visualisation tools.

Chapter 3

Single-cell RNAseq biological interpretation

3.1 Balancing between technical and biological accuracy

At the dawn of scRNAseq technologies, analytical tools were inspired from bulk pipelines, but then specific methods have been developed for single-cell studies, especially because of the aforementioned challenges (chapter 2). With the democratisation of single-cell measures, there has been a tremendous increase in the number of available algorithms, making it hard to navigate this sea of tools. We want to point out two particular points that we deemed of interest:

- ↬ A substantial number of tools are never or barely re-used,
- ↬ Benchmarking tools is a complicated task, since there are no standards, although good practices are emerging [174].

3.1.1 A humongous amount of analytical tools...

Out of the 982 tools tracked by the scRNA-tools website [175] [1] as of June 24, 2021 (1,056 at the end of September 2021), more than 30% were never cited (337 to be precise), while there is only a small fraction that have been cited more than 100 times, this being related to the 3,652 occurrences for the search term "single-cell RNA seq" in PubMed (Figure 3.1).

A possible explanation is that there is a disconnection between bioinformatics and biology teams that stay in their respective ivory towers: some tools are highly technical and improve theoretical performance but not biological interpretability. At the other end of the spectrum, some analytical practices are questionable as they do not tune correctly hyperparameters, while it might have a massive impact on the downstream results [176]. It is also probably because the analytical landscape is dominated by two behemoths that are easy-to-use integral pipelines: Seurat [125] and Scanpy [126].

[1] https://www.scrna-tools.org/

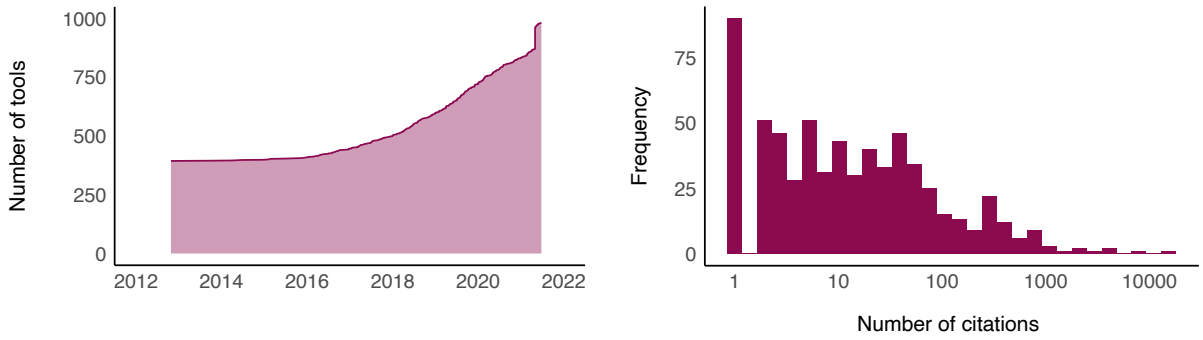


Figure 3.1: Tools for scRNAseq analysis are produced at a fast pace (left panel) but poorly diffused throughout the community (right panel).

3.1.2 . . . to benchmark

We see two reasons that could possibly explain the difficulty to benchmark new tools.

3.1.2.1 Performance scores

The benchmark of emerging tools has been inspired by the broader field of supervised and unsupervised ML research. Let us go with the example of evaluating the performance of the clustering task.

For the unsupervised approach, one can use performance metrics that quantify the inter-cluster separability versus the intra-cluster coherence. Such scores include the to-be-maximised silhouette (equation 3.1), the to-be-maximised Calinski-Harabasz (equation 3.2) or to-be-minimised Davies-Bouldin (equation 3.3) indexes. They rely exclusively on the coordinates of the data points and their cluster labels and characterize the physical overlap between the different labels.

It is not completely clear yet whether the unsupervised scores make sense, given that the coordinates in the original or projected spaces are noisy. A supplementary Achilles' heel is that they are well suited for convex clusters but not anymore if the shape is concave. A bean-shaped cluster for example would lower the score even if there is no overlap with other clusters (Figure 3.2). Obviously, since not all clusters are expected to be convex, therefore unsupervised scores should be used carefully.

For the supervised approach, the rationale is to compare the match between ground truth labels and clusters. There is a range of supervised scores: the ARI (Adjusted Rand Index) (equation 3.4), the homogeneity score (equation 3.5), or the Jaccard index (equation 3.6), among others. These scores evaluate whether two partitions overlap (ARI, Jaccard index), or whether a query partition looks like a reference one (homogeneity score), and should usually be maximised.

In this case, the caveat is related to the definition of ground truth. While the definition itself is shaky as it might be subjective, depending on the markers used to sort the sequenced populations, there is another mistrust: it is not obvious that transcriptome-defined populations would mirror cell populations defined with proteins. In other words, there is a doubt about the fact that the transcriptomic truth should reflect the proteomic one. In [177], the authors sorted T_{REG} vs T_{CONV} based on the expression of FOXP3 (ground truth labels), but these two populations merged back in the clusters found with

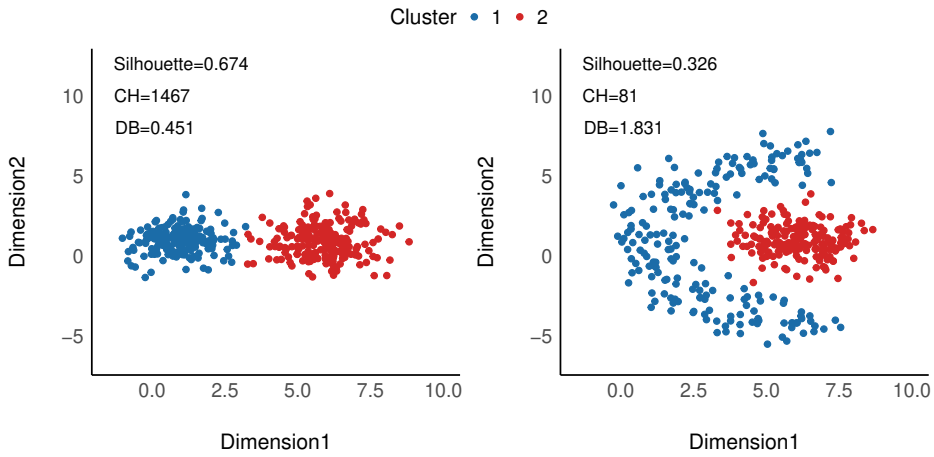


Figure 3.2: Unsupervised scores performance relates to cluster shape. Concave-shaped clusters' scores, such as bean-shaped (right panel), are worse even if the between-clusters separation is the same as for convex-shaped clusters (left panel).

the analysis of the transcriptome (cluster labels) (Figure 3.3).

3.1.2.2 Good practices in benchmarking analytical tools

Given the aforementioned weaknesses of evaluating an algorithm's performance, benchmarks should be carefully conducted. In particular, a proper benchmark study should have the following minimal characteristics [174]:

- ↬ unbiased (for example tuning parameters for some methods and not for others),
- ↬ reproducible,
- ↬ the choice of methods and parameters should be discussed,
- ↬ the benchmark should rely on a collection of well-characterized dataset, real and simulated,
- ↬ the choice of performance metrics should be explained,
- ↬ the pros and cons of each method tested have to be mentioned in order to give recommendations on how to choose one method over the others,
- ↬ the benchmark should enable future extensions.

These good practices are for example illustrated by the dynverse platform (Figure 3.4).

There is an additional common flaw in current benchmark studies: optimism [178]. Every single new method claims that it outperforms previous ones. Recurring hints, even unconscious, are to tune extensively every parameter of the new tool while using default ones for the others, or choosing particular datasets and performance metrics. In line with this observation, benchmarks performed by independent teams, or neutral benchmarks, usually fail to reproduce these claims. For this reason, one should favor neutral benchmarks to make a proper selection of an analytical tool.

Unsupervised scores

Let $\mathbb{X} = \{x_1, x_2, \ldots, x_N\}$ be a set of N data points, each data point belongs to one of the K clusters with a label $C(i)$.

$$I_k = \{i \in [\![1, N]\!] | C(i) = k\}$$

$$a_i = \frac{1}{|I_{C(i)}| - 1} \sum_{j \in I_{C(i)}, j \neq i} d(x_i, x_j),\ b_i = \min_{k' \in [\![1,K]\!], k' \neq k} \frac{1}{|I_{k'}|} \sum_{i' \in I_{k'}} d(x_i, x_{i'})$$

$$s_{sil}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$$\mu_k = \frac{1}{|I_k|} \sum_{i \in I_k} x_i,\ \mu = \frac{1}{N} \sum_{i=1}^{N} x_i,\ \bar{\delta}_k = \frac{1}{|I_k|} \sum_{i \in I_k} d(x_i, \mu_k)$$

$$B = \sum_{k=1}^{K} |I_k| \|\mu_k - \mu\|,\ W_k = \frac{1}{|I_k|} \sum_{i \in I_k} \|x_i - \mu_k\|$$

$$S_{sil} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} s_{sil}(i) \tag{3.1}$$

$$S_{CH} = \frac{(N-K)B}{(K-1) \sum_{k=1}^{K} W_k} \tag{3.2}$$

$$S_{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \left(\frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})} \right) \tag{3.3}$$

3.2 Looping on old knowledge

Interpreting scRNAseq is a highly challenging task, as the data is confounded by nuisance factors such as variation in capture efficiency and sequencing depth. Once the signal has been extracted, there is an extra fact to explain the difficulty of the interpretation: scRNAseq quantifies the transcriptome, while our knowledge mostly relies on the study of proteins. Assuming that the transcriptomic measure is an accurate proxy for the quantification of proteins, the interpretation step of the data takes advantage of previous knowledge to identify clusters or stages in a trajectory. There are three complementary strategies to annotate the data, be it at the cell- or the cluster-level.

3.2.1 Gene list enrichment

The first strategy is to use signatures for cell types that are assumed to be present in the data. These signatures come from third-party sources such as the literature, bulk or single-cell data. To compute an enrichment score of a signature in a given cell or group of cells, one can perform a hypergeometric test or alternatively a Gene Set Enrichment

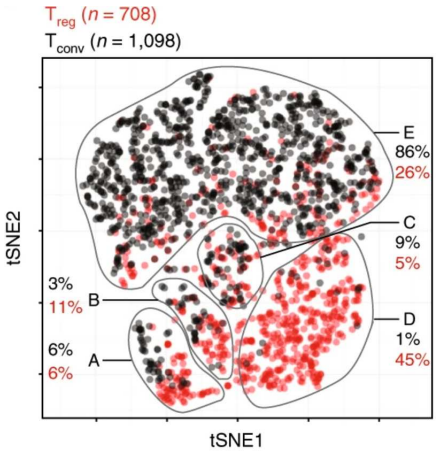


Figure 3.3: Ground truth labels are mixed in transcriptomic cluster labels. From [177].

Supervised scores

Let $T = \{T_1, T_2, \dots T_t\}$ be the ground truth partition, $C = \{C_1, C_2, \dots C_K\}$ the cluster partition, N the number of data points and $M = \{m_{ij}\}_{i \in [\![1,t]\!], j \in [\![1,K]\!]}$ the contingency table where $m_{ij} = |T_i \cap C_j|$.

$$a_i = \sum_{j=1}^{K} m_{ij}, \; b_j = \sum_{i=1}^{t} m_{ij}$$

$$H(T|C) = -\sum_{j=1}^{K}\sum_{i=1}^{t} \frac{m_{ij}}{N} \log\left(\frac{m_{ij}}{\sum_{\tilde{i}=1}^{t} m_{\tilde{i}j}}\right), \; H(T) = -\sum_{i=1}^{t} \frac{\sum_{j=1}^{K} m_{ij}}{t} \log\left(\frac{\sum_{j=1}^{K} m_{ij}}{t}\right)$$

$$ARI = \frac{\sum_{ij} \binom{m_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{N}{2}} \tag{3.4}$$

$$h = \begin{cases} 1 & \text{if } H(T) = 0 \\ 1 - \frac{H(T|C)}{H(T)} & \text{else} \end{cases} \tag{3.5}$$

$$J(T, C) = \frac{|T \cap C|}{|T \cup C|} \tag{3.6}$$

Analysis (GSEA). These two analyses rely on the existence of gene set databases, such as PanglaoDB[2] [179] or MSigDB[3] [180].

GSEA enrichment score indicates whether a set of genes $\mathcal{S}$ is overrepresented at the top or bottom of a ranked gene list [180]. The running-sum statistic decreases monotonously along the gene list except when it encounters a gene belonging to $\mathcal{S}$, upon which it is incremented by 1 unit. The enrichment score is finally derived by a Kolmogorov–Smirnov-like statistic, as the maximum deviation to 0 (Figure 3.5). The

[2] https://panglaodb.se
[3] https://www.gsea-msigdb.org/gsea/msigdb/

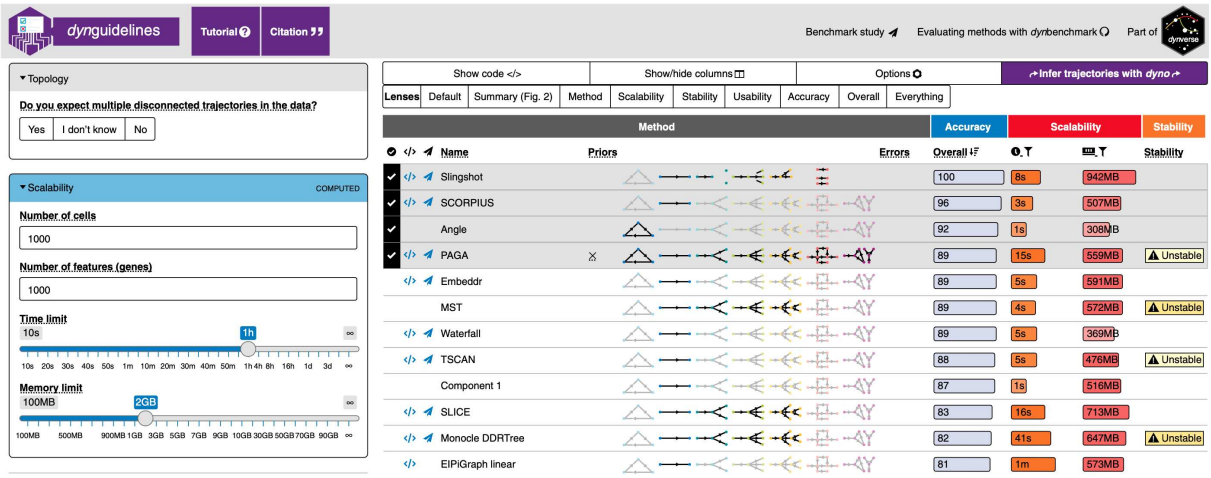


Figure 3.4: The dynverse platform provides recommendations for selecting a trajectory inference method.

score is then used to compare cells or groups for a given signature or to compare signatures for a given cell or group.

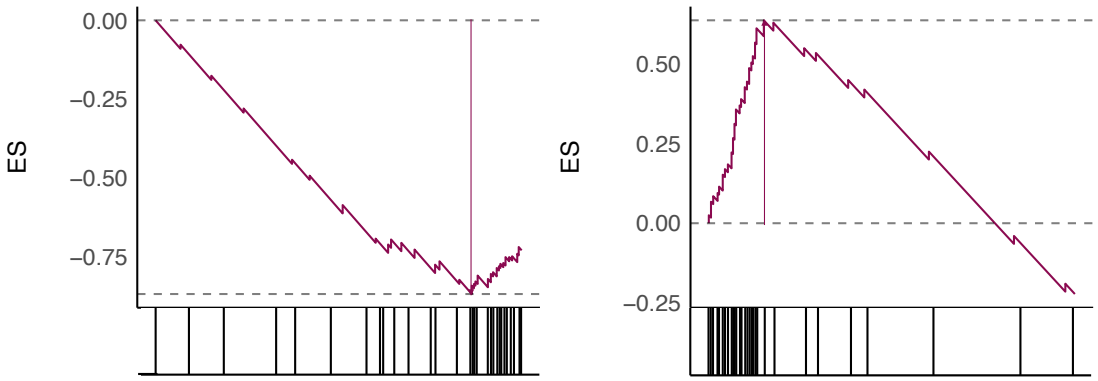


Figure 3.5: GSEA enrichment score. The score is depicted by the length of the arrow and quantifies whether a gene set is enriched at the top (left panel) or bottom (right panel) of a ranked gene list.

The hypergeometric test compares the ranked gene list to the hypergeometric distribution which is the probability of having s successes -drawing s genes of interest- in n draws -in n ranked genes- without replacement in a population of size N -in the total of N genes- containing exactly S objects with the interesting feature -containing exactly S genes of interest. The probability mass function of a random variable X following a hypergeometric distribution is:

$$P(X = s) = \frac{\binom{S}{s}\binom{N-S}{n-s}}{\binom{N}{n}} \tag{3.7}$$

In our case, $s = |\mathcal{S}|$, while we consider only a subset of features of size $n \ll N$. The output of the hypergeometric test is a p-value that indicates what is the most probable class or cell type for a given cell or group.

There is a third way to use gene sets to annotate cells: classifiers. Classifiers such as

Garnett [181], CaSTLe [182] or Moana [183] use ML tools such as linear Support Vector Machines to classify cells. Basically, this ML algorithm maximises the separation between the different classes. Once the boundaries for each class are learnt -training phase, new cells can easily be projected, and would fall in one or another of the classes.

There is a main drawback to the enrichment method: we need trustworthy signatures and the corresponding cell types should be present in the data. Hence, it is not possible to label new cell types with this approach. Even worse, if the choice of the signatures tested do not correspond to the cell types in the dataset, the annotation will be highly biased, as there is no rule of thumb to decide whether an enrichment score above a threshold should be trusted. Also, the quality of the annotation depends strongly on the quality of the signatures. Finally, we make again the assumption that there is a perfect match between the signatures that are usually based on bulk samples sorted on the basis of proteic markers and the transcriptomic truth.

3.2.2 Differential Gene Expression

The second strategy is to do a DGE analysis: this common approach consists in testing each gene and evaluate whether it is significantly up- or down-regulated in a cluster of interest versus all the other clusters, for example with a t-test. Genes are then ranked for each cluster according to the p-value, the log-fold change between the cluster of interest and the rest of the cells and the level of expression. The last step is a manual sort and investigation of the marker genes based on the existing corpus of knowledge, in order to link each gene to a known cell type for the annotation. Interestingly, it is possible to identify new cell types, as compared to the gene set enrichment strategy.

There are several downsides to performing a DGE analysis. The first one is that the annotation is hardly automatable, tedious and operator-dependent, although it is partially facilitated by exploiting gene ontologies [128, 133]. Secondly, the list of marker genes is highly dataset-dependent as they stem from a comparison between the cluster of interest and the other cells present in the data. Thus, depending on the cell types that would be sequenced, marker genes would be different. Lastly, it can only be performed at the group- and not the cell-level, while one of the advantages -although under-exploited- of scRNAseq is to work at the single-cell scale.

3.2.3 single-cell RNAseq atlases

The last strategy to annotate cells or groups is by capitalising on previous scRNAseq datasets that were carefully annotated to use them as a reference. There is a growing body of these atlases, such as the Human Cell Atlas[4] [184] that contains already 13.5 million cells or the murine Tabula Muris[5] [185] with 100,000 cells. Atlases are of paramount importance to better dissect and understand human -or other species- health and disease. New data can be projected onto an atlas (Figure 1.16). There are several technical possibilities to project, including Seurat's Azimuth[6], CelliD [186], or scmap [187]. Most of these techniques rely on the discovery of the nearest cell in the query dataset for each cell in the reference dataset. It is also possible to train a classifier on a reference dataset and use its results to annotate a query dataset.

[4] https://www.humancellatlas.org/
[5] https://tabula-muris.ds.czbiohub.org/
[6] https://azimuth.hubmapconsortium.org/

The major disadvantage of this approach is its sensitivity to the non-negligible batch effect, as well as the risk of loosing new cell types from the query. We face also once again the issue we have had with gene sets: if the reference does not contain the same cell types as the query, the projection will fail, and will not necessarily issue a warning flag.

3.2.4 Automated annotation of single-cell RNAseq data

For all the above-mentioned tools, we saw that the critical part of the annotation is to carefully choose the reference, be it a reference dataset, or reference signatures, or reference articles to uncover the meaning of marker genes. Because this step relies heavily on prior knowledge about the content of the data, it is usually poorly replicable, although there is at least one example of a meta-analysis that has tried to tediously harmonize annotations across different studies for $CD8^+$ cells in the TME [188]. Another major caveat relates to cluster-, or cell state-based approaches, as the annotation would also depend on the quality of the clustering, while it has been reported that clusters could contain more than one identified cell type [189]. Finally, it is a time-consuming process.

There are tools that claim to overcome some or all of those pitfalls, by annotating automatically datasets, using references asserted as exhaustive [190]. There are 2 main categories for such tools: either they do a DGE analysis and compare marker genes to reference gene sets, or they project new data onto an annotated reference. Let us go through few examples of reference databases and automated annotation pipelines. CellMatch, used by the scCATCH tool, contains 353 cell types and 686 subtypes, both murine and human, and scCATCH works at the cluster-level [189]. It combined previous databases: CellMarker[7] made from a manual curation of the literature, MCA[8] and CancerSEA[9] made from scRNAseq datasets, and CD Marker Handbook[10]. Cell types that scored the highest are used to annotate clusters. For a cell-level analysis, the CellMatch database can be injected in another tool, cellassign [191], which computes a probability score for each cell to belong to a given class. cellassign is capable of detecting new cell types by designating them as unassigned, although it is not able to label them. Similarly, CelliD [186] annotates at the cell-level as well while putting aside unannotated cells. Another extensive database is PanglaoDB [179] that is a repository of scRNAseq datasets with almost 4.5 million human cells, coupled with an analytical tool for automated annotation of new datasets, alona [192], and it works at the cluster-level. SingleR[11] is another reference database constructed from pure bulk RNAseq expression profiles, coupled with an annotation pipeline [193]. scANVI [194] is an extension of the scVI pipeline [149] that aims at projecting unannotated datasets onto reference data by taking advantage of neural networks in order to merge the two datasets in a latent space used for the construction of a k-NN classifier. It is also capable of annotating a whole dataset based on available labels, that would describe only a fraction of the cells.

Large databases are useful to quickly annotate a new dataset but should be used with caution, and annotations should be double checked, as the caveats mentioned above are less important but not fully eliminated. The hope is that atlases underway will serve as a reference.

[7] http://biocc.hrbmu.edu.cn/CellMarker
[8] https://figshare.com/articles/MCA_DGE_Data/5435866
[9] http://biocc.hrbmu.edu.cn/CancerSEA
[10] http://static.bdbiosciences.com/documents/cd_marker_handbook.pdf
[11] https://comphealth.ucsf.edu/app/singler

3.3 Creating single-cell RNAseq-based knowledge

One of the tremendous advantage of scRNAseq is to work at the cell-level scale, thus offering the possibility of discovering previously unnoticed new cell types or states.

3.3.1 Detection of new cell types

It is expected that new cell types represent rare populations. It is usually implicitly accepted that rare cell types' discovery is facilitated by larger datasets with a higher sequencing depth, in order to increase the raw numbers of rare cells although the precision plateaus [195] (Figure 3.6).

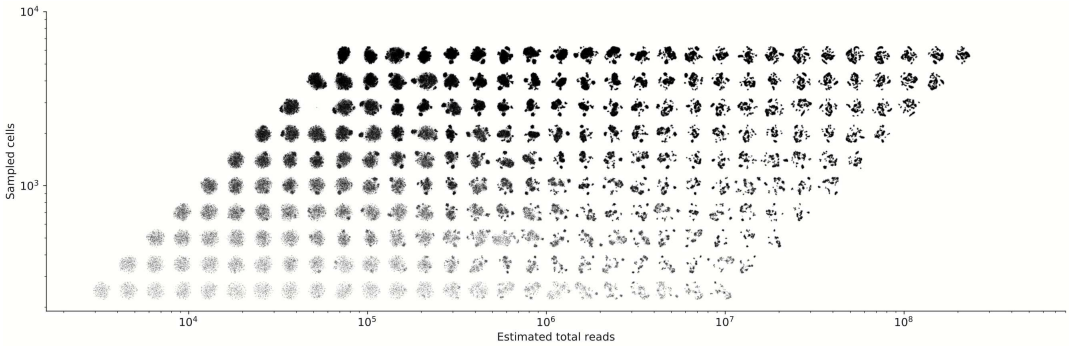


Figure 3.6: t-SNE granularity increases as a function of sequencing depth and number of cells up to a plateau. From [195].

Additionally, we need specific tools that are able to detect scarce populations, especially if bringing back to the mind the small experiment that we did in section 1.3.3.2: lowly-expressed genes, that are potential markers for rare populations, are usually filtered out right at the beginning of the analysis. While regular clustering algorithms impose to choose the number of clusters, either directly, for example with a k-mean approach, or indirectly, for example with the Louvain [196] or Leiden [197] algorithms which request the user to choose a resolution parameter that is a monotonous function of the resulting cluster number, this poses the following problem: it is not obvious that all regions of the data display the same heterogeneity. Thus, it appears detrimental to use the same resolution or number of clusters homogeneously across the whole data space. A single resolution opens the way for over-fitting some regions while under-fitting some others. To rephrase it, this could lead to some clusters being highly heterogeneous, containing different cell types and probably rare populations, and other clusters which would not give a satisfactory DGE analysis, as they would be too similar to other clusters. To overcome this issue, some methods have been specifically designed for the screening of rare populations.

There are 3 different methods to solve this issue. The first method is to have a local cluster-number-related parameter instead of a global one, to decide for the number of clusters: TooManyCells [198] or PanoView [199] use this strategy. The second method identifies outlier genes specific for rare populations: RaceID does the identification of those genes after a regular clustering using a mean-variance negative binomial model for each cluster [200] (Figure 3.7), GiniClust spots outlier genes at the beginning of the analysis using the Gini index [201]). A third method is to identify outlier cells: FiRE

[202] attributes a rareness score to every cell.

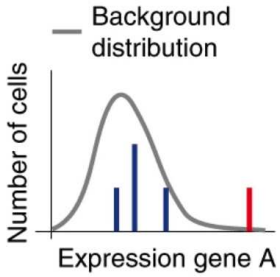


Figure 3.7: Outlier genes such as gene A are detected as the ones not following a negative binomial distribution in the RaceID algorithm. From [200].

On top of those solutions, we can decipher whether a cluster is homogeneous or not, in which case the clustering should be performed again with a higher resolution. Such algorithms evaluate a cell or a gene parameter related to the entropy. Entropy is a measure of disorder, i.e. the higher the entropy, the higher the chaos. A cluster with a high dispersion of entropies should potentially be considered as being a collection of meaningful to-be-divided sub-clusters. ROGUE [203] evaluates the gene entropy to assess whether it yields, in a single cluster, an expected entropy or if it falls outside of the expectation. Depending on the number of high-entropy genes, the cluster is considered as pure or not, although it depends on a choice of a threshold. scEntropy [204], on the other hand, evaluate the cellular entropy, and goes even further by infusing the entropy information directly to the clustering algorithm.

Upon the clustering of rare, and potentially novel, cell types, one has to conduct a DGE analysis, in order to identify each cluster. Clusters that cannot be annotated based on its marker genes to an existing cell type, using existing literature, databases and annotation tools (see previous section), are considered to be new cell types until proven otherwise.

3.3.2 Validation of new cell types

As most of the fields in scRNAseq analysis, there is no gold standard on how to proceed to validate a new cell type. It is strongly recommended though, that new populations discovered with scRNAseq should go through a compulsory validation step at the bench, in order to demonstrate their functional specificity [133]. In order to study a new cell type in the wet lab, one has to (i) identify specific surface markers that would enable to sort, for example with FACS, the new population. The validity of the sorting should be checked with a (ii) subsequent scRNAseq experiment, before performing (iii) adequate functional assays, morphological evaluations, screening of secreted proteins, and assessment of the anatomical compartment or spacialisation [205, 206].

Additionally, we believe that standardised cell ontologies should help to ascertain the novelty of a given cluster, by enabling faster comparisons between a scRNAseq result and existing knowledge.

Chapter 4

Objectives of the thesis

In the introduction, I have outlined several challenges in the analysis of scRNAseq data, technical as well as biological. The data is very noisy, because of the technique but also because of the genuine biology. In order to take advantage of scRNAseq data, we need to disentangle the two sources of noise.

- ✻ Regarding the technical noise, I have been especially intrigued by the dimensional noise, usually termed as the curse of dimensionality, and how to eliminate it or at least mitigate it. The curse of dimensionality is critical, since the analysis starts with the noisy high-dimensional count matrix. It is also interesting to tackle it if we make the hypothesis that noise is mostly dimension-related, as assumed by imputation methods.
- ✻✻ But I was also worried by the gap between highly technical and numerous solutions and their use within biology teams and the biological interpretation. A disturbing amount of analytical tools are never used, while some analysis are performed without proper use of the tools. Additionally, the interpretation remains manual, time-consuming and subjective.

For my thesis, I wanted to challenge current paradigms in the field of scRNAseq and T cell biology, using a point of view both methodological to confront ✻ and immunological/biological to confront ✻✻. A third part of the results focuses on a specific biological question: why the prognostic role of T_{REG} in cancer is fuzzy? This work stresses the importance of deciphering T cell functions, in order to better encompass T cell complexity. It has been the basis of questioning the lineage paradigm.

4.1 First focus: tackling the curse of dimensionality in order to improve the performance of scRNAseq analysis pipelines

The curse of dimensionality has a strong negative impact on the analysis of scRNAseq data, as it blurs the contrast between small and large pairwise distances. Since pairwise distances or similarities are the main ingredient to form clusters or draw trajectories, it is of utmost importance to take it into account. In addition, the curse of dimensionality happens in high-dimensional spaces, but in fact high dimension could mean already 10. As a consequence, retaining 20 PCs does not guarantee that the "low-dimensional" data is exempt from the curse anymore.

One effect of the curse is the measure concentration, that I discussed in the introduction. Unfortunately, the measure concentration can hardly be corrected in high-dimensional spaces, while the hubness phenomenon, another curse-related effect is correctible. Yet, it is interesting to work in the high-dimensional space since:

- it contains all the signal,
- the "low-dimensional" space might still suffers from the curse of dimensionality.

We deemed that it would be interesting to study the hubness phenomenon in scRNAseq data, with the golden thread that has been to verifying the interest of working in higher-dimensional spaces than is usually done. We investigated it to prove whether hubness affects scRNAseq data and whether correcting it is useful, especially in higher-dimensional spaces, for the performance of scRNAseq analysis.

While it will be just one out of the thousand tools already existing, I believe that a major interest lies in the fact that we directly tackle the curse of dimensionality, or at least one of its effects, instead of avoiding it. This is intellectually satisfying, but also judicious, as we avoid discarding signal as it would be the case with a drastic dimension reduction. This is especially valid in the current context, where there are almost no consensual and valid method to choose the dimensionality of the low-dimensional manifold.

4.2 Second focus: Dissect the functional diversity of single-cell RNAseq of T cell in cancer with a supervised functional approach

This second project was triggered by the observation that interpretability of the data is still puzzling. I started from the hypothesis that there is a potential decorrelation between functions and the current immune cell classification that is used nowadays to annotate cells in scRNAseq data. This hypothesis is supported by the fact that the current lineage paradigm is being questioned by new discoveries, such as the fact that cells can transition from one lineage to another, or express TFs from two different lineages.

Therefore, we suggest a supervised approach of scRNAseq data to analyse the functionality of T cells, avoiding to go through the annotation step performed with the existing classification. I carefully designed functional modules in order to score T cell functions in every cell of the count matrix.

First, I verified the added value of this approach, compared to the unsupervised pipeline. Then I assessed the functions of T cells from the tumor or the juxtatumor. The goal of this approach is to be able to determine the functions of T cells, but also of other immune cells, in order to produce a functional atlas of the different tissues, in a physiological or pathological condition.

4.3 Third focus: Context-dependant approach enable to unveil T cell functions: the example of regulatory T cells in cancer

This project started following the intriguing observation that the role of Tregs in cancer with respect to prognosis was ambiguous. Since there is barely any cohort which would

evaluate the prognosis with scRNAseq, but rather flow cytometry or immunohistochemistry, I thought to take advantage nevertheless of the existing data galore. While the data differs from single-cell count matrices, it is interesting as well: it emphasizes the importance of better characterizing T cell functions (Tregs in our case), here via the prism of the context.

For the meta-analysis, I chose 5 cancer types, I selected relevant experimental articles and collected all parameters relevant to describe the context: treatment, tissue, markers, quantification method, etc. I systematically evaluated the effect of 3 context-related parameters on the evaluation of the prognosis, in order to show that it would improve the consensus on Treg prognosis role, as well as to extract a clearer picture of Treg role for cancer prognosis.

Part II

Results

Chapter 5

Hubness reduction improves clustering and trajectory inference in single-cell transcriptomic data

5.1 Synopsis of the hubness study

In order to study the effect of the hubness phenomenon in scRNAseq data, we first evaluated its magnitude in omics data. We used bulk RNAseq data as well to probe systematically the effect of sparsity and intrinsic dimension on hubness, to better understand the mechanisms driving the emergence of hubness. Intrinsic dimension relates to the minimal number of dimensions needed to accurately describe the data. We observed that omics data is sensitive to hubness, all the more so given a high sparsity and a high intrinsic dimension. Since hubness is related to the dimension, it is plausible that hubness follows the same trend as effective, but also intrinsic dimension.

While we used classical methods to probe the hubness phenomenon, we realised that no method was reliable to capture correctly hub cells, so we designed one, based on the size of the hub cells' neighborhoods.

Using our new method to retrieve hubs, we studied the nature of the latter. My intuition has been that hub cells would be the archetypical profile of the corresponding clusters, but it proved wrong. On the contrary, there is no biological nor technical differences between hubs and regular cells, or between antihubs and regular cells. The only trait that we could observe is that hubs stand in dense regions, near cluster centers, while antihubs are in scarce regions, on the outer margin of clusters. This is useful, as it means that hubs could serve for centroid initialisation.

To evaluate the usefulness of hubness correction, we infused different k-NN graphs to clustering, trajectory inference and visualisation algorithms: hub-corrected or not. We observed that "hubby" datasets, corresponding to datasets with a high intrinsic dimension, would particularly benefit from hubness correction, as seen with higher supervised performance metrics for the clustering and trajectory inference tasks, and higher unsupervised performance metrics for the visualisation task.

5.2 Article

Bioinformatics, 2021, 1–7
https://doi.org/10.1093/bioinformatics/btab795
Advance Access Publication Date: 23 November 2021
Original Paper

Gene expression

Hubness reduction improves clustering and trajectory inference in single-cell transcriptomic data

Elise Amblard [1,†], Jonathan Bac[2,3,4,†], Alexander Chervov[2,3,4], Vassili Soumelis[1] and Andrei Zinovyev [2,3,4,5,*]

[1]Université de Paris, INSERM, HIPI, F-75010 Paris, France, [2]Institut Curie, PSL Research University, F-75005 Paris, France, [3]INSERM, U900, F-75005 Paris, France, [4]CBIO-Centre for Computational Biology, Mines ParisTech, PSL Research University, 75006 Paris, France and [5]Laboratory of Advanced Methods for High-Dimensional Data Analysis, Lobachevsky University, 603000 Nizhny Novgorod, Russia

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Anthony Mathelier

Received on June 10, 2021; revised on November 5, 2021; editorial decision on November 15, 2021; accepted on November 17, 2021

Abstract

Motivation: Single-cell RNA-seq (scRNAseq) datasets are characterized by large ambient dimensionality, and their analyses can be affected by various manifestations of the dimensionality curse. One of these manifestations is the hubness phenomenon, i.e. existence of data points with surprisingly large incoming connectivity degree in the data-point neighbourhood graph. Conventional approach to dampen the unwanted effects of high dimension consists in applying drastic dimensionality reduction. It remains unexplored if this step can be avoided thus retaining more information than contained in the low-dimensional projections, by correcting directly hubness.

Results: We investigated hubness in scRNAseq data. We show that hub cells do not represent any visible technical or biological bias. The effect of various hubness reduction methods is investigated with respect to the clustering, trajectory inference and visualization tasks in scRNAseq datasets. We show that hubness reduction generates neighbourhood graphs with properties more suitable for applying machine learning methods; and that it outperforms other state-of-the-art methods for improving neighbourhood graphs. As a consequence, clustering, trajectory inference and visualization perform better, especially for datasets characterized by large intrinsic dimensionality. Hubness is an important phenomenon characterizing data point neighbourhood graphs computed for various types of sequencing datasets. Reducing hubness can be beneficial for the analysis of scRNAseq data with large intrinsic dimensionality in which case it can be an alternative to drastic dimensionality reduction.

Availability and Implementation: The code used to analyze the datasets and produce the figures of this article is available from https://github.com/sysbio-curie/schubness.

Contact: andrei.zinovyev@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell omics profiling revolutionized many fields of modern molecular biology, providing more direct ways to study such biological phenomena as differentiation (Trapnell, 2015), development (Blakeley *et al.*, 2015), heterogeneity of cancer cell populations and related resistance to treatment (Aynaud *et al.*, 2020; Tirosh *et al.*, 2016). However, the analysis of single-cell RNA sequencing (scRNAseq) datasets remains challenging, amplifying the difficulties already encountered in the analysis of bulk omics measurements as well as introducing new ones, specific to single-cell technologies (Kiselev *et al.*, 2019; Lähnemann *et al.*, 2020).

RNAseq datasets have large ambient dimensionality close to 10^4 (number of unique genes) by order of magnitude. The expression profiles of individual genes are coupled through networks of linear and non-linear dependencies. This makes the intrinsic dimensionality (ID) of the data point cloud much lower. For example, if all genes were linearly correlated to one common latent factor, then all cells would be located on a line segment in the multi-dimensional space, and the ID would be equal to one. Real gene expression datasets are influenced by more than one latent factor, and, intuitively, the number of distinct latent factors corresponds to the global ID (GID) of the data (Kairov *et al.*, 2017). Previous studies suggest GID estimates for scRNAseq

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

data vary from 3–4 to few tens (Albergante *et al.*, 2019; Aynaud *et al.*, 2020).

The difficulties of dealing with many dimensions in data analysis are broadly referred to as the 'curse of dimensionality'. Talking about the curse becomes relevant when the logarithm of the number of data points is less than the ID of the data (Bac and Zinovyev, 2019; Gorban and Tyukin, 2018). In practice, it means that certain manifestations of the dimensionality curse might appear starting with an ID as low as 10. These manifestations are diverse: among the most known is the concentration of distances quantified as the vanishing contrast between 'close' and 'far' distances. Several approaches were proposed to compensate for the undesirable effect of distance concentration (Luecken and Theis, 2019; Wang *et al.*, 2018). However, in practice, it was demonstrated that it cannot be avoided by global modifications of the data space metric (Mirkes *et al.*, 2020).

In most current analysis workflows, scRNAseq datasets are subjected to drastic dimensionality reduction before applying unsupervised machine learning methods (Wolf *et al.*, 2018). This common practice aims at reducing possible manifestations of the curse of dimensionality, at the cost of neglecting signals that are potentially contained in higher dimensions. Moreover, for trajectory inference (TI), the dimensionality is frequently reduced to 2 or 3 (Saelens *et al.*, 2019). This is in striking contrast with the observation that in a typical scRNAseq dataset, the first few tens of principal components can explain only a small fraction of total variance (e.g. 5–10%). Thus, it remains unclear if the practice of reducing the dimension of scRNAseq data eliminates useful information, and whether it is the only way to fight the dimensionality curse.

Yet another manifestation of the curse of dimensionality is the hubness phenomenon. It has been described that in high-dimensional space some points might be surprisingly popular among the k-Nearest Neighbours (k-NN) of other points. This observation was formalized in Radovanovic *et al.* (2010), coining the term 'hubness'. The hubness of a point is the in-degree of the corresponding node in the k-NN graph. The distribution of hubness scores as a function of the dimension shifts to the right when the dimension increases, forming a fat Lévy-type power–law tail (Supplementary Fig. S1), which contains the hubs.

Dealing with hubness is crucial when exploiting k-NN graphs, which are an essential ingredient of most of current computational approaches for scRNAseq analysis (Wolf *et al.*, 2018). Presence of hubs in the k-NN graph impacts their expected properties: e.g. it can change the structure of geodesic distances along the graph. Hubs (and antihubs, i.e. points with a null in-degree) make the structure of k-NN graphs heterogeneous in terms of connectivity, which can violate required assumptions to apply graph-based algorithms. It is surprising that the hubness properties of scRNAseq neighbourhood graphs and their impact on downstream analyses has never been studied so far.

Hubness reduction methods aim at explicitly reducing the hubness of the k-NN graph, usually through specific transformations of the distance matrix (Feldbauer *et al.*, 2018). Interestingly, hubness reduction could be used as a replacement for dimensionality reduction. In this study, we hypothesized that hubness reduction methods can be beneficial for scRNAseq analyses relying on k-NN or other neighbourhood graphs. We systematically evaluate the effect of hubness reduction on clustering, TI and visualization tasks in scRNAseq data, using previously established benchmark datasets. Finally, we specify the conditions in which hubness reduction is expected to be beneficial.

2 Materials and methods

2.1 Datasets collection

Bulk datasets were downloaded from the ARCHS[4] and TCGA website. The ARCHS[4] expression matrix was limited to bulk gene expression profiles and downsampled to 2000 samples. For TCGA, we took the breast (BRCA) and renal (KIRC) datasets as the largest RNASeq datasets in TCGA, containing more than 600 samples. To evaluate clustering and TI, we gathered scRNAseq datasets with gold- and silver-standard labels used in previous benchmarks (Abdelaal *et al.*, 2019; Duò *et al.*, 2018; Gulati *et al.*, 2020; Krzak *et al.*, 2019; Saelens *et al.*, 2019; Sun *et al.*, 2019; Tian *et al.*, 2019) (SupplementaryFig. S26 and Supplementary Table S1).

2.2 Hubness quantification and reduction

2.2.1 Evaluating hubness

From the bulk or single-cell datasets, we performed Principal Component Analysis (PCA) using log-transformed data, while retaining only the 10 000 most variable genes. The k-NN graph was computed for a number of PCs ranging from 2 to the number of cells minus 1 and k ranging from 5 to 100, but most of the results are shown for $k = 10$. From the k-NN graph, the in-degree or hubness score X_i is calculated for each cell i.

Let k be the value used to build the k-NN graph, X the distribution of hubness scores, μ its mean and σ its standard deviation.

The $2k$ estimator counts the number of points with a hub score above $2k$. The *Mean* estimator counts the number of points with a hub score larger than three standard deviations above the mean. The *Antihub* estimator is the number of points having zero hub score. The *Asymmetry* estimator counts the percentage of unidirectional edges in the k-NN graph. The *Skewness* estimator is calculated as $S_k = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$. The *Max* estimator is the maximum hub score observed in the distribution, divided by the cardinality of the dataset.

Of note, in the case of scRNAseq datasets, large proportions of cells are antihubs, while the tail of the distribution of in-degree follows a power law that can lead to variance divergence. As a consequence, the distribution of in-degrees is strongly skewed and the usual threshold-based methods for defining hubs can be misleading. We suggest a novel definition of hubs based on the size of the incoming neighbourhood. For m data points with the largest in-degree we calculate the number of data points $n(m)$ that have at least one of these m putative hubs in their nearest neighbours. We call these $n(m)$ cells the reverse-covered cells. We compute the increment $\frac{n(m)-n(m-1)}{N}$ and choose such m when the increment drops below a threshold α, where the threshold value is chosen such that the number of hubs would equal zero in the projection onto the first two principal components. Because of redundancy of hubs in terms of the data points they cover, the α threshold can be crossed several times, so we select the largest m above the increment threshold, such that any further increase of m by one does not increase the coverage by more than α, see Supplementary Figure S6.

We used the Python package scikit-hubness to measure skewness and to reduce hubness (Feldbauer *et al.*, 2018). This package offers four methods for reducing hubness, that produce a hub-corrected k-NN graph: Mutual Proximity (MP), Local Scaling (LS) and its variant LS-NICDM (Non-Iterative Contextual Dissimilarity Measure) and DisSimLocal (DSL) (Feldbauer and Flexer, 2019; Schnitzer *et al.*, 2012) (Supplementary Methods).

2.2.2 Dropout simulation

We simulated dropout in two different ways. First, the dropout rate is a fixed constant for all samples, and we drop only non-zero gene counts. Second, we used the tool from R library Splatter (Zappia *et al.*, 2017) to add dropout in a more realistic way, reproducing the distribution of scRNAseq data values.

2.2.3 Intrinsic dimensionality

We evaluated ID using PCA with the scikit-dimension package (Bac *et al.*, 2021). Global ID is defined as the number of eigenvalues of the covariance matrix exceeding a tenth of the largest eigenvalue. We consider that datasets with a GID above 25 are high dimensional (high-ID datasets). Mean local ID is defined as the mean of ID values computed for the 100-nearest neighbourhood of each point.

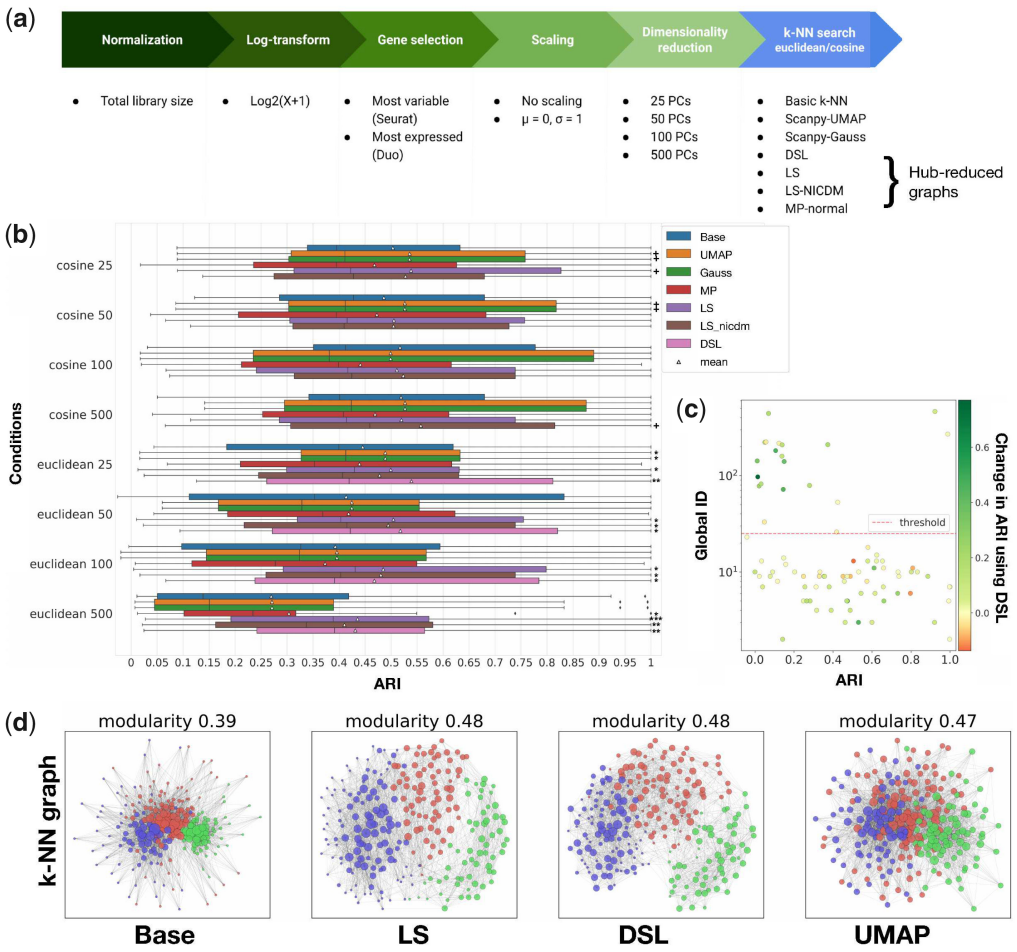


Fig. 1. Evaluation of hubness reduction effect on clustering performance. (**a**) Preprocessing workflow with the different conditions used to construct various k-NN graphs upstream of the clustering task. (**b**) ARI scores for high-ID datasets, as a function of the metric, dimension and k-NN graph production method used; example with the Seurat recipe, scaling and the Leiden algorithm. Relative differences for individual datasets are shown in Supplementary Figure S12. (**c**) Relationship between GID, ARI and improvement in the clustering score using the hubness reduction method DSL. (**d**) Selected example of Leiden clustering on a scRNAseq dataset (GSE60783), using Euclidean distance, 50 PCs and a 15-NN graph. UMAP k-NN graph does not reduce the skewness of the in-degree distribution as compared with hub-reduced graphs. The modularity is improved for the UMAP and hubness-reduced graphs compared to the base one. Each colour represents a ground truth class of data points and point size is proportional to the in-degree in the respective k-NN graph. *P*-values are indicated following the mapping: '+' indicates *P*-value between 0.05 and 0.1, '*' indicates *P*-value between 0.01 and 0.05, '**' indicates *P*-value between 0.001 and 0.01 and '***' indicates *P*-value below 0.001. Each condition is compared with the base k-NN graph

2.3 Clustering

We processed the datasets with Scanpy (Wolf *et al.*, 2018). We systematically tested two recipes for the preprocessing (Duo or Seurat) with two different metrics (Euclidean and cosine dissimilarity), using scaling or not, and four values for the number of PCs to retain (25, 50, 100 and 500). The following k-NN graphs were computed: simple (base) k-NN graph, four hubness-reduced graphs, using the hubness reduction methods from the scikit-hubness package, and two methods to compute neighbourhood graphs from the Scanpy package (Coifman *et al.*, 2005; McInnes *et al.*, 2018).

The Duo recipe consists in log-normalizing the data, keeping the 5000 most variable genes and normalizing again.

The Seurat recipe log-normalizes the data as well and selects the variable genes according to a set of thresholds: variable genes with a mean between 0.0125 and 3, and a dispersion above 0.5. The data are normalized again after the gene filtering step.

The clustering was done on the seven k-NN graphs with the Leiden algorithm (De Meo *et al.*, 2011). The number of nearest-neighbours was set to the square root of dataset cardinality. Since the graph-based clustering methods do not allow choosing the exact number of clusters, we tuned the resolution parameter to get the ground truth number of clusters. We started with a resolution of 1.5 and limited the search of the resolution to the interval [0, 3]. We then performed iterative clustering, with a maximum of 20 iterations and a resolution which would increase or decrease in a dichotomous manner (Supplementary Methods).

We used the Adjusted Rand Index (ARI) and the homogeneity scores to evaluate the quality of clustering (Rosenberg and Hirschberg, 2007). The best score value is 1 for both measures (Supplementary Methods).

2.4 Trajectory inference

We used the implementation of PAGA from Scanpy. Same combinations of preprocessing steps, metrics and clustering algorithm have been used as described in the clustering section.

We used the R toolbox dynverse to compute three quality metrics on each trajectory: correlation, F1_branches and featureimp_wcor (Supplementary Methods). We also computed an overall score of the three quality metrics which is the arithmetic mean of the latter (Saelens *et al.*, 2019).

2.5 Statistics

We carried out paired *t*-tests to compare the differences in performance between the different k-NN graph production methods using the base k-NN graph as a reference. We made randomization tests

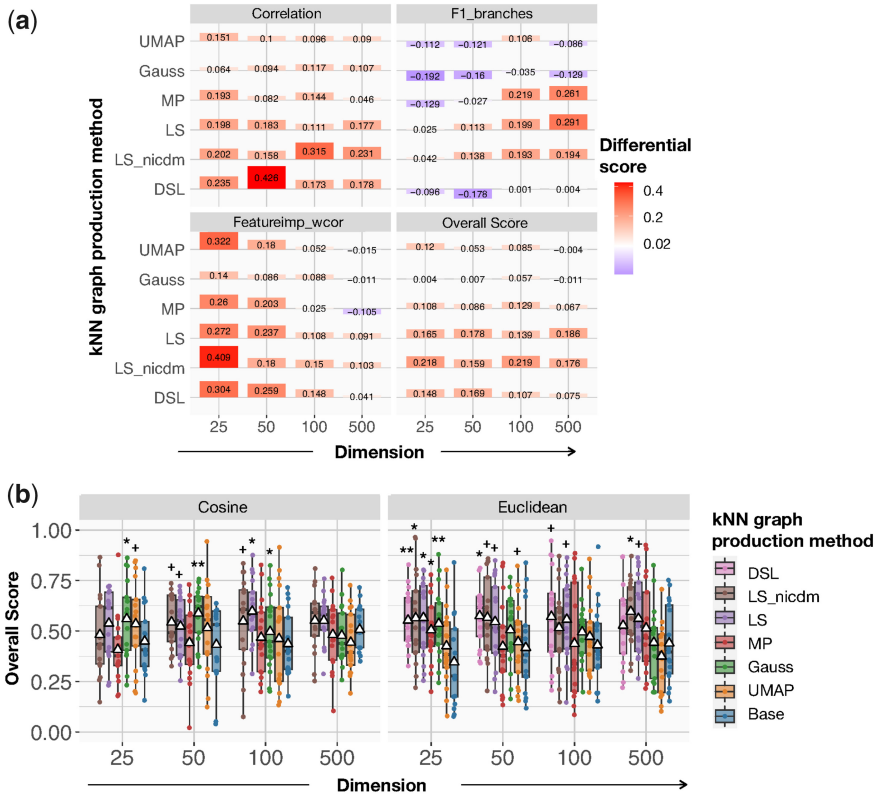


Fig. 2. Trajectory inference (TI) improvement from application of hubness reduction. (**a**) Differential TI quality scores (taking as reference the base score) for the three TI quality metrics and the average score as a function of the dimension and the k-NN graph production method, calculated for the high-ID datasets [8,19]; example with the Seurat recipe, Euclidean metric and Leiden algorithm. (**b**) Average TI quality score of all datasets, as a function of the dimension and the k-NN graph production method, calculated for the high-ID datasets [8,19]; example with the Seurat recipe and Leiden algorithm. *P*-values are indicated following the mapping: '+' indicates *P*-value between 0.05 and 0.1, '*' indicates *P*-value between 0.01 and 0.05, '**' indicates *P*-value between 0.001 and 0.01 and '***' indicates *P*-value below 0.001. Each condition is compared with the base k-NN graph

to compare the estimated hubness level between sequencing data and other real-life datasets from the openML repository.

3 Results

3.1 RNA-seq data are prone to hubness

As a first step of our analysis, we hypothesized that increased dropout rate in scRNASeq datasets as compared with the bulk datasets, could lead to increasing data hubness. In order to validate this hypothesis, we started with three bulk RNASeq datasets from The Cancer Genome Atlas (TCGA) and ARCHS[4] repositories and simulated increasing dropout (see Section 2). We used previously developed tools to quantify the magnitude of hubness (Feldbauer and Flexer, 2019; Low *et al.*, 2013) (see Section 2). The skewness (k-skewness) and the asymmetry estimators increase with the number of PCs retained before reaching a plateau, for all datasets (Supplementary Figs S2A and S3C). Two other hub estimators, the Maximum and the Mean estimators (see Section 2), behave similarly with respect to the dimension (Supplementary Fig. S3A and D). From those observations, we concluded that there are hubs in RNASeq data, which appear already at intermediate dimensions, namely 10 PCs.

To investigate the link between sparsity and hubness, we studied their respective correlation with the global intrinsic dimension (GID, see Section 2). Since sparsity correlates to GID ($R = 0.93$, $P < 0.0001$, Spearman correlation) and GID to hubness, defined with the asymmetry estimator, at $k = 10$ and considering 100 dimensions ($R = 0.81$, $P < 0.0001$, Spearman correlation), we can assume that the effect of sparsity on hubness is at least partially due to an increased GID (Supplementary Fig. S4A). This observation confirms the intuition that hubness is a high dimension-related effect.

We analysed a diverse collection of scRNAseq datasets (Duò *et al.*, 2018) to measure hubness, using the same hub estimators we applied to bulk datasets and reproduced their evolution for increasing dimensionality (Supplementary Figs S2B and S3B). We concluded that scRNAseq is prone to hubness as well, starting already at around 10 PCs. We also showed that the scRNASeq datasets are characterized by stronger hubness on average than a wide collection of 500 real-life and synthetic datasets obtained from the OpenML repository (Bac *et al.*, 2021; Vanschoren *et al.*, 2014) (Supplementary Fig. S5).

We also investigated the link between sparsity and GID in scRNAseq, and observed that sparsity is not sufficient to explain the variations of GID. We uncovered three parameters that influence GID: sparsity, cardinality of the dataset and the signal-to-noise ratio (SNR) (see Supplementary Methods). The SNR and the cardinality are two dependent parameters ($R = 0.87$, $P = 0.0026$, Spearman correlation) (Supplementary Fig. S4B), so we computed the correlation between GID and the composite parameter ratio of sparsity to SNR, which appeared to be significant ($R = 0.77$, $P = 0.015$, Spearman correlation) (Supplementary Fig. S4C).

3.2 Hubs are not artefact cells

We assessed whether hubs have different properties compared with other cells by looking at various quality control (QC) metrics: number of genes, total number of features, dropout rate, entropy (see Supplementary Methods), position in low-dimensional projections. We retrieved hubs using our reverse-coverage method (see Section 2). There was no clear difference in the distributions of QC metrics between hubs, antihubs and other (normal) cells (Supplementary

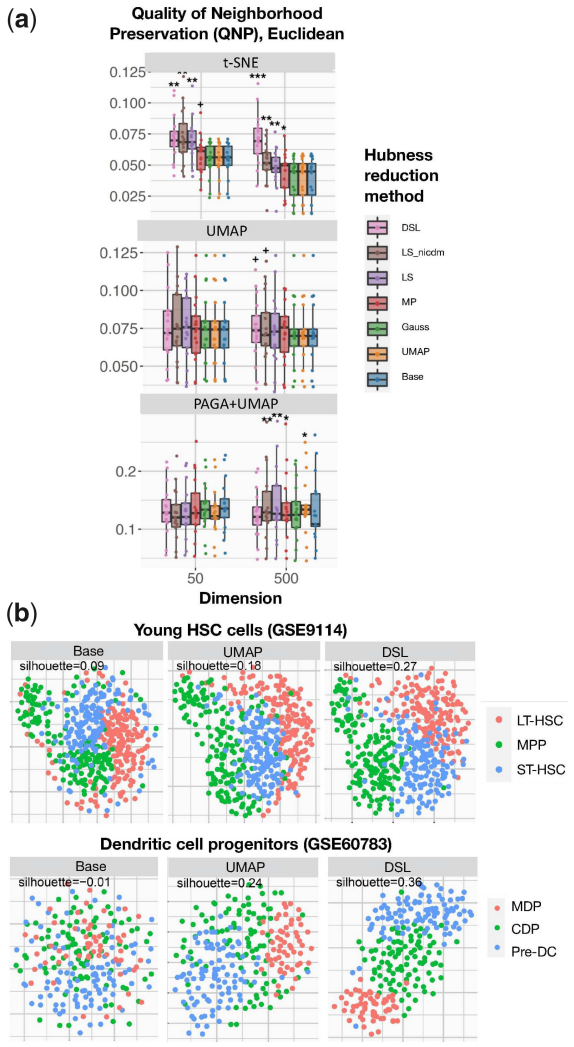


Fig. 3. Hubness reduction affecting non-linear embedding algorithm performance (visualization task) for high-ID datasets. (**a**) Quality of point neighbourhood preservation (QNP) after applying various hubness reduction methods to the point neighbourhood graph used in various embedding methods (t-SNE, UMAP, PAGA+UMAP). *P*-values are indicated following the mapping: '+' indicates *P*-value between 0.05 and 0.1, '*' indicates *P*-value between 0.01 and 0.05, '**' indicates *P*-value between 0.001 and 0.01, '***' indicates *P*-value below 0.001. Each method is compared with the base k-NN graph. Complete analysis of QDM and QNP metrics of various hubness reduction methods in visualization tasks is provided in Supplementary Figure S24. (**b**) Examples of visualizations obtained by replacing the neighbourhood graph in UMAP, corrected or not corrected for hubness, with estimation of the silhouette score for the ground truth labels for the low-dimensional projection. The neighbourhood graphs were computed in 500-dimensional space, and embedded using PAGA+UMAP approach with Leiden clustering

Figs S2C and D and S7). Removal of hubs made new hubs emerging (Supplementary Text, Supplementary Fig. S2F). These observations suggest that hub or antihub cells do not form a distinct group of cells: they are not biological or technical artefacts, but merely a consequence of high dimensionality.

3.3 Hubness reduction improves clustering accuracy

We studied the effect of hubness reduction on the clustering of scRNAseq data using labelled datasets collected from previous clustering benchmark studies. To compare them in an uniform manner, we processed them using Scanpy (Wolf *et al.*, 2018) according to standard steps with several combinations of parameters: normalization, log-transformation, gene selection, scaling, dimensionality reduction (Fig. 1a, see Section 2).

Hubness reduction was applied to generate corrected k-NN graphs. We compared these with the uncorrected k-NN graph as well as the neighbourhood graphs provided by two methods from Scanpy (Gauss and UMAP) (Coifman *et al.*, 2005; McInnes *et al.*, 2018). Our results show that GID and hubness are important parameters to consider when clustering scRNAseq data. Datasets with higher GID, i.e. above 25, had generally lower ARI scores, whereas low and high scores were possible for lower GID (Fig. 1c). The two exceptions of high-ID datasets with high scores correspond to the only two simulated datasets included in our benchmark (Duò *et al.*, 2018).

High-ID datasets are also the ones prone to hubness in the Euclidean space: indeed the mean local ID (LID) correlates with k-skewness. Although there is no direct correlation between k-skewness and ARI, it is clear that high GID and hubness need to be taken into account when clustering (Supplementary Fig. S9C).

Clustering after hubness reduction was most useful for high-ID datasets (Supplementary Figs S9B and S11), performing better than the uncorrected k-NN graphs (Fig. 1b and c, Supplementary Figs S12 and S13). Interestingly, the highest average ARI and homogeneity scores were achieved using hubness reduction and 500 PCs. The use of cosine dissimilarity to build the k-NN graph resulted both in lower hubness and higher clustering accuracy than the Euclidean distance, as expected from previous literature (Schnitzer *et al.*, 2014) (Supplementary Fig. S9A). This provides a rationale to consider cosine dissimilarity and related metrics (e.g. the angular distance) as more appropriate to cluster scRNAseq data. It also indicates that a less stringent dimension reduction can yield better clustering performance. For the case of low-ID datasets, the benefit of doing hubness reduction is not obvious anymore but should be evaluated individually (Supplementary Fig. S17).

Dimension- and hub-reduction both mitigate negative effects of high GID on downstream analysis. Our study suggests that these two procedures have complementary effects, with hubness reduction allowing to reduce the dimension less stringently. We also observe from Figure 1d and Supplementary Figure S10 that the improvement in clustering performance is accompanied by more homogeneous densities and a reduced skewness of the k-NN graph, while the k-NN graph constructed with UMAP corrects density inhomogeneity but not high skewness (Supplementary Text). We also tested this hypothesis on bulk datasets, where hubness reduction improved modularity in the absolute majority of the datasets (Supplementary Fig. S17).

3.4 Hubness reduction improves trajectory inference

To evaluate the effect of hubness reduction on the performance of TI in scRNAseq data, we generated various k-NN graphs as input for the TI task, with or without hubness reduction. We used the Partition-based Graph Abstraction (PAGA) (Wolf *et al.*, 2019) method to do TI since it was ranked as the top-performing tool in a large-scale benchmark (Saelens *et al.*, 2019). It is also appropriate in our study since it uses a k-NN graph as the input. The following quality scores have been utilized: correlation to evaluate the relative position of cells along the trajectory, F1_branches to compare branch assignment and featureimp_wcor to measure the respective importance of differentially expressed features while constructing the trajectory (see Section 2). We also calculated an average score. We observed that the inferred trajectories were closer to the ground truth in most high-ID cases when TI was performed on a hub-reduced k-NN graph rather than using the base or the Scanpy k-NN graphs, in terms of the overall summary score and regardless of the combination of preprocessing parameters (Fig. 2b, Supplementary Text). We display one example of preprocessing parameters combination (with the Seurat recipe, the Euclidean metric and the Leiden algorithm) in Figure 2a to show the improvement of the various TI quality scores compared with the base k-NN graph. There are no clear patterns revealing that the increase in the quality of TI would be due to a specific increase in one of the three quality metrics: in

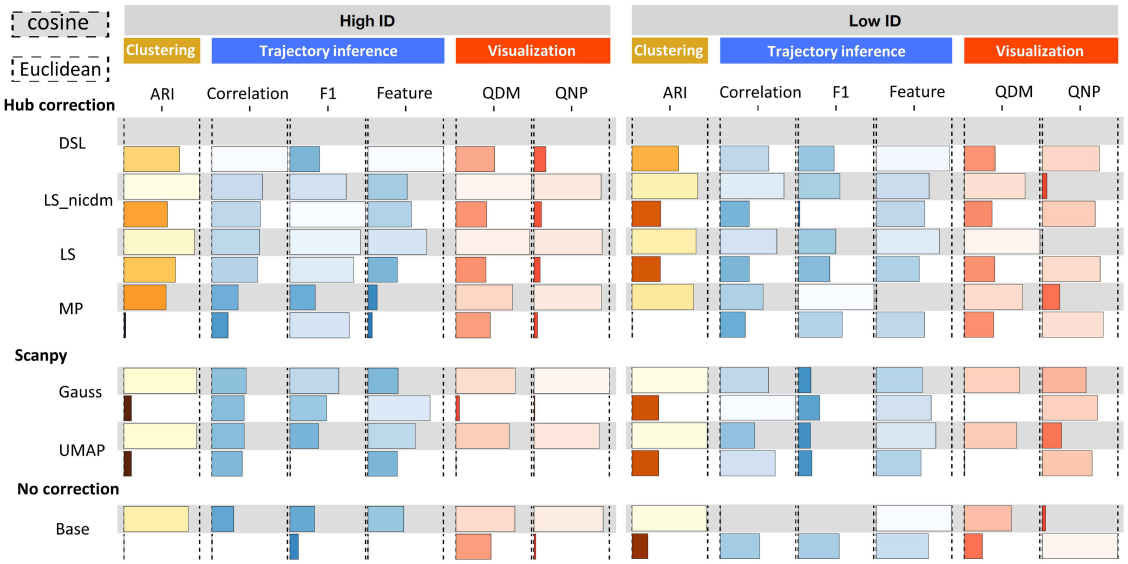


Fig. 4. Average across multiple conditions of quality metrics scores for three tasks of single-cell data analysis (clustering, trajectory inference and visualization), as a function of the type of k-NN graph used, and the data metric (cosine dissimilarity or Euclidean distance), for high-ID and low-ID datasets

fact, it depends strongly on the preprocessing (Supplementary Figs S20 and S21).

If we consider all the different preprocessing combinations and all datasets together, we can study the respective efficacy of each hubness reduction method. For the TI task and considering the largest GID, we observed that the quality of the TI done after applying the two LS-based hubness reduction methods is the highest, shortly followed by DSL then MP. Going back to the datasets characterized by a low GID, it is not clear anymore what is the best hubness reduction method to improve TI (Supplementary Figs S18 and S19). As a consequence, we suggest that one should test hubness reduction, with a preference for DSL and LS methods, to reach the best performance evaluated by scoring the clustering results with the silhouette score and the biological interpretation, for high-GID datasets (Fig. 2b, Supplementary Text).

3.5 Low-dimensional embeddings upon hubness reduction

Since the popular visualization methods t-SNE and UMAP are based on embedding the data point neighbourhood graph, we evaluated the impact of hubness reduction on the quality of the resulting visualisation. We verified that in a model distribution that suffer from hubness, there is a lower quality of the projection in terms of cost functions and the Quality of Distance Mapping (QDM) and the Quality of point Neighbourhood Preservation [QNP; see Supplementary Methods and (Gorban and Zinovyev, 2010)], as compared with a distribution that suffer less from hubness (Supplementary Figs S23 and S24, Supplementary Text).

Similarly, we saw a moderate improvement of the visualization task performed after hubness reduction for high-ID datasets, for at least one hubness reduction method and especially when using the Euclidean metric (Fig. 3). There was only one use case for which hubness reduction was not beneficial: when we projected the data with UMAP after PAGA initialization and with the cosine dissimilarity. For low-ID datasets, the benefit of applying hubness reduction was not clear with our data (Supplementary Fig. S24).

4 Discussion

We have shown that transcriptomic data, both bulk and single-cell can suffer from hubness. Using bulk RNASeq data, we observed that this sensitivity positively correlates with sparsity, probably because sparsity positively influences the GID. In scRNAseq data, we found a positive correlation between sparsity and hubness, even if this effect is mitigated by the cardinality and the signal-to-noise ratio. It would be interesting to explore other factors explaining differences in hubness.

To quantify hubness, we used methods previously introduced in the literature Feldbauer and Flexer (2019); Low *et al.* (2013), but found that defining points as hubs in scRNAseq data can be non-trivial, especially in high dimensions. We introduced a definition of hubs based on the proportion of points having them as their closest neighbours.

We studied the nature of hubs, showing that they are not artefact cells or cells with specific biological properties. However, they have a topological utility, in the sense that they tend to be located close to the cluster centres and can be used for initialization of the clustering Tomasev *et al.* (2013).

We evaluated existing techniques of hubness reduction that modify local metrics with respect to their effect on the quality of clustering, TI and visualization. The summary of this evaluation is provided in Figure 4. We show that hubness reduction can be beneficial, especially for the datasets characterized by high GID, probably because they suffer more from hubness. We noticed that cosine dissimilarity produces k-NN graphs that are less prone to the hubness phenomenon, compared with the more widely used Euclidean distance. It appears that hubness reduction is complementary to dimension reduction, allowing one to retain more principal components than is usually done.

However, the available hubness reduction methods differ in efficacy, with mutual proximity (MP) method showing generally poor improvement. Our hypothesis to explain its poor performance for the clustering and TI tasks compared with the three other methods is that MP uses all pairwise distances to correct for hubness. On the contrary, other methods take advantage of the local neighbourhoods which may explain their better efficiency (Supplementary Fig. S25).

Besides the known set of hubness reduction methods that we benchmarked in this article, there exists other approaches also aiming at improving the properties of neighbourhood graphs in high dimensions. To mention few, UMAP dimensionality reduction method is built on a modified point neighbourhood graph, and shared nearest neighbours (SNN) graphs are introduced to compensate asymmetry of neighbourhood relations in high dimensions. We compared both UMAP (see Figures of this manuscript) and SNN (data not shown) approaches to the standard set of hubness reduction methods, and found that their advantages in the standard tasks

of single-cell data analysis are limited compared with the explicit hubness reduction approaches benchmarked in this study.

Acknowledgement

The authors thank the ABiMS platform for the access to their computing resources.

Funding

This work was partially supported by the French government under management of Agence Nationale de la Recherche as part of the 'Investissements d'Avenir' program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute); the Ministry of Science and Higher Education of the Russian Federation (Project No. 075-15-2021-634); Association Sciences et Technologie—Groupe de Recherche Servier and the doctoral school Frontières de l'Innovation en Recherche et Education-Programme Bettencourt.

Conflict of Interest: none declared.

Data Availability

Data are available via Zenodo under DOI 10.5281/zenodo.4597151. Code has been uploaded on GitHub in the schubness repository (https://github.com/sysbio-curie/schubness).

References

Abdelaal,T. *et al.* (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.

Albergante,L. *et al.* (2019) Estimating the effective dimension of large biological datasets using Fisher separability analysis. In: *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary. IEEE, pp. 1–8.

Aynaud,M.-M. *et al.* (2020) Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single-cell resolution. *Cell Rep.*, **30**, 1767–1779.e6.

Bac,J., and Zinovyev,A. (2019) Lizard brain: tackling locally low-dimensional yet globally complex organization of multi-dimensional datasets. *Front. Neurorobotics*, **13**, 110.

Bac,J. *et al.* (2021) Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, **23**, 1368.

Blakeley,P. *et al.* (2015) Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development (Cambridge, England)*, **142**, 3613.

Coifman,R.R. *et al.* (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci.*, **102**, 7426–7431.

De Meo,P. *et al.* (2011) Generalized Louvain method for community detection in large networks. In: *2011 11th International Conference on Intelligent Systems Design and Applications*, Cordoba, Spain. pp. 88–93.

Duò,A. *et al.* (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, **7**, 1141.

Feldbauer,R., and Flexer,A. (2019) A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge Inf. Syst.*, **59**, 137–166.

Feldbauer,R. *et al.* (2018) Fast approximate hubness reduction for large high-dimensional data. In: *2018 IEEE International Conference on Big Knowledge (ICBK)*, Singapore. pp. 358–367.

Gorban,A., and Tyukin,I. (2018) Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Phil. Trans. R. Soc. A*, **376**, 20170237.

Gorban,A.N., and Zinovyev,A. (2010) Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural Syst.*, **20**, 219–232.

Gulati,G. *et al.* (2020) Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, **367**, 405–411.

Kairov,U. *et al.* (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, **18**, 712.

Kiselev,V. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.

Krzak,M. *et al.* (2019) Benchmark and parameter sensitivity analysis of scRNAseq clustering methods. *Front. Genet.*, **10**, 1253.

Lähnemann,D. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.

Low,T. *et al.* (2013) *The Hubness Phenomenon: Fact or Artifact? In towards Advanced Data Analysis by Combining Soft Computing and Statistics*, **Vol. 285**. Springer, Berlin, Heidelberg, pp. 267–278.

Luecken,M., and Theis,F. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.

McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861.

Mirkes,E. *et al.* (2020) Fractional norms and quasinorms do not help to overcome the curse of dimensionality. *Entropy (Basel, Switzerland)*, **22**, 1105.

Radovanovic,M. *et al.* (2010) Hubs in space: popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, **11**, 2487–2531.

Rosenberg,A., and Hirschberg,J. (2007) V-measure: a conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.

Schnitzer,D. *et al.* (2012) Local and global scaling reduce hubs in space. *J. Mach. Learn. Res.*, **13**, 2871–2902.

Schnitzer,D. *et al.* (2014) Choosing the metric in high-dimensional spaces based on hub analysis. In: *European Symposium on Artificial Neural Networks. Bruges, Belgium.*

Sun,S. *et al.* (2019) Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol.*, **20**, 269.

Tian,L. *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.

Tirosh,I. *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.

Tomasev,N. *et al.* (2013) The role of hubness in clustering high-dimensional data. *IEEE Trans. Knowledge Data Eng.*, **26, 739–751.**

Trapnell,C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.

Vanschoren,J. *et al.* (2014) Openml: networked science in machine learning. *SIGKDD Explorations*, **15**, 49–60.

Wang,B. *et al.* (2018) SIMLR: A tool for large-scale single-cell analysis by multi-kernel learning. *Proteomics*, **18**, 2.

Wolf,F. *et al.* (2018) Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.

Wolf,F. *et al.* (2019) PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, **20**, 59.

Zappia,L. *et al.* (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.

Supplementary Files for the manuscript Hubness reduction improves clustering and trajectory inference in single-cell transcriptomic data

Amblard E, Bac. J et al.

1 Supplementary Methods

1.1 Hubness quantification

1.1.1 Hubness of generic data

We compared the sensitivity to hubness for generic data, by collecting the 501 datasets from the openML repository[1] that contained more than 1,000 samples and using the scikit-hubness Python package to measure skewness and hub occurrence fraction.

1.1.2 Hubness of simple model data distribution

We generated in Python Gaussian and uniformly sampled from hypercube data distributions with 10,000 samples, in spaces of dimension 2, 10, 50 and 500. Then we compute the 10-NN graph to retrieve the in-degree of each point, and show the distribution of in-degrees with 200 bins, averaged over 100 i.i.d. iterations for each dimensionality value and the data distribution.

1.1.3 Signal-to-Noise-Ratio (SNR) evaluation

To quantify the SNR, we assumed that the distribution of the eigenvalues from the cell-cell covariance matrix follows a Marcenko-Pastur distribution, except for a few eigenvalues that contain the signal of the data. As a consequence, we derive that the fraction of eigenvalues following the Marcenko-Pastur distribution is a good estimation of the noise magnitude, while the fraction of eigenvalues outside this distribution is a proxy for the signal magnitude. Instead of fitting the Marcenko-Pastur distribution, we designed a simpler proxy for SNR, that proved to be satisfactory in our experiments(that means it was enough to catch the dependence of GID to the ratio between thus defined SNR and the sparsity):

$$\mathrm{SNR} = \frac{\max(\mathbb{X})}{\mathrm{median}(\mathbb{X})}$$

where $\mathbb{X}$ is the distribution of the eigenvalues of the cell-cell covariance matrix.

1.1.4 QC measurements

We used the Seurat library to compute the UMAP and PCA projections, as well as the total number of features and the number of unique genes. The sparsity rate is the percentage of zeros in the expression matrix. To estimate the transcriptomic single-cell entropy, we used the scEntropy tool.[LSL20] To compute the stability of hub identity, we sampled 10 times 90% of the cells and looked at the mean percentage of hubs from the original data that were recovered in the resampled data to evaluate the intrinsic nature of hubs.

[1] https://www.openml.org/search?type=data

1.1.5 Hub positions with respect to the center of model data distributions

To evaluate hubs' position, we generated in Python Gaussian and uniformly sampled in hypercube distributions with 10^5 points each, in different spaces of dimension 2, 5, 8, 10 and 20. We used the scikit-hubness library to construct the 10-NN graph and get the hubness score. From those scores we computed the average distance to the coordinate origin and its rank for the data points with a hubness score above a given threshold. To make the average computation more robust, we considered only the averages calculated over more than 100 points. For the average distance M_t, we normalize it by the dimension:

$$M_t = \frac{1}{n} \sum_{\{i; s_i \geq t\}} \sqrt{\frac{n_i}{D}}$$

where t is the threshold on the hubness score, D the dimension, s_i the hubness score of point i and n_i its norm.

1.2 Hubness reduction

1.2.1 Hubness reduction methods

Mutual Proximity models pairwise distances $d_{i,j \in \{1,\dots,n\} \setminus i}$ of a set of n points with random variables X_i that depict the distribution of distances between x_i and all other points, then:

$$\mathrm{MP}(d_{i,j}) = 1 - P(X_i > d_{i,j} \cap X_j > d_{i,j})$$

where P is the joint probability density function.
Local Scaling is calculated using the pairwise distance $d_{i,j}$ and takes into account the local neighborhood:

$$\mathrm{LS}^k(d_{i,j}) = 1 - \exp(-\frac{d_{i,j}}{r_i^k}\frac{d_{i,j}}{r_j^k})$$

where k refers to the size of the local neighborhood, and r_i^k is the distance of point x_i to its k-th neighbor.
The variant LS-NICDM uses the average distance to the k neighbors instead of the mere distance to the k-th neighbor:

$$\mathrm{NICDM}^k(d_{i,j}) = \frac{d_{i,j}}{\sqrt{\mu_i^k \mu_j^k}}$$

where μ_i^k is the average distance of point x_i to its k nearest neighbors.
DisSimLocal uses local centroids $c^k(\bullet)$ to reduce hubness:

$$\mathrm{DSL}^k(x_i, x_j) = \|x_i - x_j\|_2^2 - \|x_i - c^k(x_i)\|_2^2 - \|x_j - c^k(x_j)\|_2^2$$

where the local centroid is estimated as the barycenter of the k nearest neighbors of x_i:

$$c^k(x_i) = \frac{1}{k} \sum_{\mathrm{x_j} \in \mathrm{kNN(x_i)}} x_j$$

1.2.2 Hubness and measure concentration

To evaluate the impact of hubness reduction on k-skewness and measure concentration in a general case, we generated two types of distributions: one or two Gaussian blobs in 10, 50 and 100 dimensions, with 5,000 points per blob, over 10 iterations. We used the scikit-hubness Python package to reduce hubness and measure k-skewness. For the measure concentration, we evaluated it as:

$$\text{Conc} = \frac{1}{N}\sum_i \frac{D^i_{max} - D^i_{min}}{D^i_{max}}$$

where D^i_{max} and D^i_{min} are the maximum, resp. the minimum pairwise distances for point i. These distances were calculated either considering all points, or only the 50 nearest neighbors.

1.3 Clustering

1.3.1 Tuning of clustering resolution

Algorithm 1 Resolution tuning

```
step ← 0
min ← 0
max ← 3
while step < max_step do
    resol ← min + (max−min)/2
    Perform clustering with parameter resol
    if cluster > truth then
        max ← resol
    end if
    if cluster < truth then
        min ← resol
    end if
    if cluster == truth then
        return resol
    end if
    step ← step + 1
end while
return resol
```

1.3.2 Evaluation of clustering accuracy

ARI and homogeneity (h) scores are calculated as:

$$\text{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}}$$

$$h = \begin{cases} 1 & \text{if } H(P_2) = 0 \\ 1 - \frac{H(P_2|P_1)}{H(P_2)} & \text{else,} \end{cases}$$

where

$$H(P_2|P_1) = -\sum_{j=1}^{|P_1|}\sum_{i=1}^{|P_2|}\frac{n_{ij}}{N}\log\left(\frac{n_{ij}}{\sum_{i=1}^{|P_2|} n_{ij}}\right)$$

$$H(P_2) = -\sum_{i=1}^{|P_2|}\frac{\sum_{j=1}^{|P_1|} n_{ij}}{|P_2|}\log\left(\frac{\sum_{j=1}^{|P_1|} n_{ij}}{|P_2|}\right)$$

where P_1 and P_2 are the two partitions, n_{ij} is the value of the i-th row and j-th column in the contingency table, and a_i, resp. b_j, is the sum of the values sitting on the i-th row, resp. j-th column, of the contingency table.

1.3.3 Model distribution simulating strongly heterogeneous data point density

We generated a 10-dimensional Gaussian distribution containing 10,000 points. The data point cloud was separated in two parts by a hyperplane of coordinates x=0, x being the first axis. From the right half hyperball, we randomly pick 100 points and discard the others. We then constructed the k-NN graphs with the scikit-hubness Python package, with or without hubness reduction and used the k-NN graph to estimate the density.[LA13] Briefly, it is possible to evaluate the following quantity from the unweighted k-NN graph:

$$\mathbb{D} = \log(p(X_{target})) - \log(p(X_{source}))$$

where X_{source} and X_{target} are two data points, and p is the local density. First, we determine the shortest path γ between X_{source} and X_{target} in the k-NN graph using the Dijkstra algorithm. Then for each intermediate point X_i in the path, we get from the k-NN graph the quantities:

$$\text{Left}_\gamma(X_i) = |\text{Out}(X_i) \cap \text{In}(X_{i-1})|$$

$$\text{Right}_\gamma(X_i) = |\text{Out}(X_i) \cap \text{In}(X_{i+1})|$$

where $\text{In}(X_i)$ and $\text{Out}(X_i)$ are the in- and out-neighborhoods of X_i. From this point, the density estimate along the path γ is:

$$\mathbb{D} = C\sum_{X_i \in \gamma}[\text{Right}_\gamma(X_i) - \text{Left}_\gamma(X_i)]$$

where C is a constant depending on k and the number of dimensions.[Li11] In our case we fixed the source at the center of the Gaussian hyperball and randomly sampled 60 targets in each half hyperball, to be able to compare the estimates from the two half hyperballs by calculating the average of the density estimates for both half hyperballs.

1.3.4 Clustering modularity evaluation in bulk RNA-seq data

We took the mouse collection of datasets from the ARCHS[4] data repository, retaining only those containing more than 300 samples. Without any other filters, it represents a total of 148 datasets. To compute the modularity for each dataset, the data was log-transformed then projected or not in the PCA space with 50 components. From that we compute the k-NN graph with the cosine dissimilarity, with or without hub reduction done with the LS and MP methods only to reduce computation time. Finally we applied Louvain clustering algorithm using different k-NN graphs and computed the modularity Q using the Python library igraph:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i - c_j)$$

where m is the number of edges in the k-NN graph, A_{ij} is the element of the adjacency matrix on the i-th row and j-th column, k_i is the in-degree of point i, c_i its cluster identity and δ the Dirac function.

1.4 Trajectory inference

1.4.1 Evaluation of trajectories

Correlation is calculated from the geodesic distance and quantifies the correlation between the relative distances of a given cell in the reference and the predicted trajectories:

$$\text{Correlation} = \frac{1}{n}\sum_i \text{corr}(X_i, Y_i)$$

where X_i is the distribution of relative geodesic distances to cell i in the reference trajectory and Y_i the distribution of relative geodesic distances to i in the prediction.
F1_branches computes the similarity of branch membership between two trajectories, by mapping each cell to its closest branch:

$$\text{Jaccard}(b, b') = |\frac{b \cap b'}{b \cup b'}|$$

$$\text{Recovery} = \frac{1}{|B|}\sum_{b\in B} \text{max}_{b'\in B'}\text{Jaccard}(b, b')$$

$$\text{Relevance} = \frac{1}{|B'|}\sum_{b'\in B'} \text{max}_{b\in B}\text{Jaccard}(b, b')$$

$$\text{F1} = \frac{2}{\frac{1}{\text{Recovery}} + \frac{1}{\text{Relevance}}}$$

where B and B' are the two branch partitions for the reference and the predicted trajectories.
For the calculation of featureimp_wcor, the geodesic distances of all cells to all milestones in the trajectory are computed, then predicted with a Random Forest. From the Random Forest, we retrieve the importance of each gene for the prediction in the two trajectories in order to compute a weighted Pearson correlation, with the weights depending on the mean importance in the reference trajectory:

$$m_{ref} = \frac{\sum_g (R_g^{ref})^2}{\sum_g R_g^{ref}}$$

$$m_{pred} = \frac{\sum_g R_g^{ref} R_g^{pred}}{\sum_g R_g^{ref}}$$

$$s_{ref} = \frac{\sum_g R_g^{ref}(R_g^{ref} - m_{ref})^2}{\sum_g R_g^{ref}}$$

$$s_{pred} = \frac{\sum_g R_g^{ref}(R_g^{pred} - m_{pred})^2}{\sum_g R_g^{ref}}$$

$$s = \frac{\sum_g R_g^{ref}(R_g^{ref} - m_{ref})(R_g^{pred} - m_{pred})}{\sum_g R_g^{ref}}$$

$$wcor_{feat} = \frac{s}{\sqrt{s_{ref} s_{pred}}}$$

where R_g is the importance of gene g in the predicted or reference trajectory.

1.4.2 Trajectory stability

We used the same methodology described in a previous benchmark[SCTS19] to evaluate the stability of PAGA. Briefly, we sample 95% of the cells and genes iteratively and evaluate the differences between two successive trajectories, doing 10 iterations and using the correlation and F1_branches metrics, but excluding featureimp_wcor which is not stable on the identity. To compare two successive iterations we compute both metrics using the common cells and genes, on the n+1 iteration, using the n-th iteration as the reference. We get a stability score for each metric. To compare the stability across the different datasets and conditions, we normalize the scores, such that the correlation and F1branches have the same magnitude for each dataset. Briefly, for each dataset and each metric, we transform the scores to get $\sigma = 1$ and $\mu = 1$, then apply the unit probability density function of the normal distribution. We then compute the arithmetic mean of the two metrics. To speed the computation of the stability, we just ran it on the Sun et al. datasets[SZMZ19] and we did not compute the two scanpy k-NN graphs.

1.5 Visualisation task

1.5.1 Generating n-cubes and n-spheres

We generated n-cubes and n-spheres in Python, using the packages scikit-dimension and numpy. We generated 10 sets for each distribution, each containing 5,000 points, embedded in spaces of various dimensions in the range [10, 50, 100].

1.5.2 Low dimension projections: t-SNE, UMAP, PAGA+UMAP

We used the following Python libraries to compute the projections: sklearn for t-SNE, umap for UMAP, and scanpy for PAGA + UMAP. For t-SNE, we use `metric='precomputed'` and `perplexity=50.0` for the single-cell experiment (and the default values for the model experiment). For PAGA+UMAP, we set `init_pos='paga'` when running `scanpy.tl.umap`. For all projections, we set `n_components=2`.

1.5.3 Correlation metric QDM and QNP

Quality of Distance Mapping quantifies the correlation of pairwise distances, only retaining a subset of the latter. We compute first what is called "natural PCA"[GZ10] on the reference: the pair of most distant points (i_1, j_1) represents the first components. Then, for the n+1 component (i_{n+1}, j_{n+1}), it is such that i_{n+1} is the most distant to the set of previous components $S_n = \{i_1, \ldots, i_n, j_1, \ldots, j_n\}$ and j_{n+1} is the point of S_n closest to i_{n+1}. We used this set of pair to compute the QDM:

$$\text{QDM} = \text{corr}(d_{i,j}\hat{d}_{i,j})$$

where $d_{i,j}$ is the distance in the reference space, $\hat{d}_{i,j}$ in the projection and we compute the correlation using the set of components S_n from the natural PCA. We took n=1000 for the tests with the hypercube and the hypersphere, and n

equals to the number of cells for the tests with the single-cell datasets.
Quality of point Neighborhood Preservation computes the intersection of the neighborhoods in the reference and projection:

$$\mathrm{QNP_k} = \frac{1}{k}\sum_{i=1}^{N}\frac{|S_i^k \cap \hat{S}_i^k|}{N}$$

where S_i^k, resp. $\hat{S}_i^k$, is the neighborhood of point i in the reference, resp. the projection, k is the size of the neighborhood and N the number of points. For the hypercube and hypersphere test, we took k in the range [10, 50, 100] and for the single-cell data, k equals the square root of the cardinality.

1.5.4 Cost function: Kullback-Leibler divergence and cross-entropy

The KL divergence is the cost function used for the t-SNE algorithm and is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{i,j} P_{ij}\log\frac{P_{ij}}{Q_{ij}}$$

$$P_{j|i} = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\sigma_i})}{\sum_{k\neq i}\exp(-\frac{\|x_i - x_k\|^2}{2\sigma_i})}, P_{ij} = \frac{P_{j|i} + P_{i|j}}{2N}$$

$$Q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k - y_l\|^2)^{-1})}$$

where x_i, resp. y_i, is the vector of the i-th point in the reference, resp. the projected, space, σ_i is a parameter that is entirely determined by the choice of the perplexity in the t-SNE algorithm and N is the number of points.
The cross-entropy is the cost function in UMAP:

$$\mathrm{CE}(P,Q) = \sum_{i,j}[P_{ij}\log\frac{P_{ij}}{Q_{ij}} + (1-P_{ij})\log\frac{1-P_{ij}}{1-Q_{ij}}]$$

$$Q_{ij} = \frac{1}{1+a\|y_i - y_j\|^{2b}}$$

where y_i is the vector of the i-th point in the projected space, a and b are two parameters entirely determined by the choice of a min_dist in UMAP and P_{ij} is the membership strength of the 1-simplex between the i-th and the j-th points.

2 Supplementary Results

2.1 Hubs are not artefact cells

In some datasets, hubs and antihubs had more often a higher dropout rate, a higher number of unique genes detected or a lower number of total features compared to normal cells but this observation was not reproducible across all datasets; and there is no observable trend at all for the entropy. Regarding their position, the hubs and antihubs are scattered across the whole projection in the two-dimensional UMAP and PCA embeddings, in the sense that they do not form a distinct cluster or a set of outliers, although hubs seem to be located in denser regions (Supplementary Figures 2C and 7B). In order to empirically rationalize the position of hubs, we looked at the following model distributions in various dimensions: Gaussian and uniformly sampled hypercube. It was observed that for the Gaussian data, the distance to the data center decreased when the node degree increased except for the smallest dimensions which were less sensitive to hubness as expected. For the uniformly sampled hypercube data, in small dimensions the hubs were far from the origin (which has been observed before[LBSN13]), nevertheless for high enough dimensions, hubs concentrated again near the data space origin (Supplementary Figure 8). Since the hubness phenomenon is specific to high dimensional data, we can conclude that hubs tend to concentrate near the cluster centers. We also looked at the stability of hubs upon resampling of 90% of the cells and proved their poor stability: there is in most cases less than 25% of hubs in common between the original data and the resampled one, whatever the dimension and the metric used to compute the k-NN graph are (Supplementary Figure 2E). It proves that being a hub is not an intrinsic property of the cells. Lastly, we studied the hubness of the data after removing the hubs identified by the reverse-coverage approach, and observed that the k-NN graph remains asymmetrical, meaning that new hubs appeared in the data (Supplementary Figure 2F). It also serves as a proof that hubness can not be reduced by merely removing hubs: more elaborated techniques are needed to correct the skewed k-NN graph.

2.2 Hubness reduction improves clustering accuracy

We looked at hubness reduction considering only gold-standard datasets. Even if the improvement in ARI and homogeneity scores is less strong than if we consider only high-ID datasets, hubness reduction remains a useful step to perform, especially for the Euclidean distances, and considering a number of PCs above 25 (Supplementary Figure 15). We tested the Louvain clustering algorithm, which yielded similar results as compared to Leiden (corresponding figures are available on our Zenodo repository, DOI 10.5281/zenodo.4597151). Number of nearest-neighbors was set to the square root of the dataset cardinality (see Methods). We also explored the effect of hubness correction on local density inhomogeneities. We evaluated the strength of the density correction in a model high-dimensional Gaussian distribution which has been inhomogeneously sampled (see Supplementary Methods). Briefly, we take a Gaussian ball in 10 dimensions and remove 98% of the points in one of the half hyperball (Supplementary Figure 16A). As a mere consequence, the mean density of each half hyperball is different. We show that the density evaluated from the unweighted k-NN graph is more uniform after hub correction (Supplementary Figure 16C,D). The local neighborhood relations are also better represented after the hubness reduction, in the sense that close points fall back in the same neighborhood (Supplementary Figure 16B).

2.3 Hubness reduction improves trajectory inference

Most of the trajectories were improved. Exceptions were some combinations of preprocessing parameters used with 25 PCs (namely the Duo recipe with the cosine dissimilarity). Some combinations used with 100 PCs were not improved with hubness reduction either (namely the cosine dissimilarity with the Duo recipe and Leiden algorithm) (Supplementary Figures 20, 21). In total, only 9 out of 32 combinations of preprocessing parameters failed to yield better overall performance with hubness reduction. Out of these 9 combinations, 4 were computed with 25 PCs. When hub reduction is applied on the datasets embedded in lower dimensional spaces, e.g. 25 PCs, it is actually not surprising that hubness reduction has a weaker effect since the magnitude of hubness itself is smaller. Also, 8 combinations were computed with the cosine dissimilarity, which we know from previous experiments exhibit less initial hubness compared to the Euclidean metric (Supplementary Figure 9A). To conclude, we noticed that the benefit of hubness reduction was much higher when using a high number of PCs and the Euclidean metric, which is coherent with the observations of the clustering task. Briefly, we also noticed a slight improvement in the TI performance if we consider the low dimensional datasets from Sun et al.[SZMZ19] This is again especially true for the Euclidean metric, except for the preprocessing done with the Duo recipe and the Leiden algorithm (Supplementary Figures 18, 19). This is interesting compared to the clustering task, for which the dimensionality of the datasets was an important parameter to decide whether hubness reduction would be beneficial or not.

2.4 Low-dimensional embeddings upon hub reduction

To evaluate the impact of hubs and hubness reduction on the visualisation task, we designed two tests. Firstly, we used two distributions, the n-cube and the n-sphere, to evaluate the impact of the hubness phenomenon on the goodness-of-fit between the projection and the original data. The second test comprises scRNAseq data to quantify the quality of the projection before or after hubness reduction, in the same vein as what we did for the clustering and TI tasks. Here, we evaluated two visualisation algorithms, namely t-SNE[vdMH08] and UMAP,[MHM18] that are widely used within the single-cell community. We used randomly sampled n-cubes and n-spheres (see Supplementary Figure 22A), assuming that the n-cube exhibits hubs, while the n-sphere does not, or to a lesser extent (see Supplementary Figure 22B).[LBSN13] Thus, we can estimate whether the presence of hubs impedes projecting the data onto a smaller (e.g. 2-dimensional) space. We quantify the goodness-of-fit

by looking at the respective cost functions of the visualisation algorithms: Kullback-Leibler (KL) divergence and cross-entropy (CE) (see Supplementary Methods), as well as at two metrics measuring correlation: the Quality of Distance Mapping (QDM) and the Quality of point Neighborhood Preservation (QNP; see Supplementary Methods). The projection is the best possible whenever it minimizes the cost function and maximizes correlation. Our hypothesis is that the projection for a n-sphere will be of better quality than the one for a n-cube, because hubs distort the pairwise distance matrix used to compute t-SNE or UMAP. Regarding the cost functions, we note that they are designed to point towards the direction of the gradient descent for a given aim, but not as an absolute reference of the goodness-of-fit. We observed that QNP and QDM correlation metrics were always higher for the n-sphere than for the n-cube, both for t-SNE and UMAP, and irrespective of the number of dimensions or neighbors tested (see Supplementary Figure 23). For the cost functions, and keeping in mind the fact that they focus on specific structures in the data, we see that the KL divergence and the CE are smaller for the n-sphere than the n-cube, except for the CE computed after UMAP (see Supplementary Figure 23). We explain it by the fact that CE attributes a high importance to hubs and antihubs and thus the existence of these specific points accelerates the minimization of CE while performing the gradient descent. We reinforce this explanation with Figure 1d, where we observe that the UMAP k-NN graph keeps the hubs at the center and the antihubs at the border, as in the base projection. Consequently, the k-NN graph structure with hubs is easier to preserve in the sense of the CE, even if the projection is overall of worse quality.
Then we switched to single-cell datasets, using the same set of high-ID data as for the TI task, and tested the various k-NN graphs (the two Scanpy graphs and the four hub-reduced ones), but excluding the base one, that were projected in the UMAP, UMAP initialized with PAGA (PAGA + UMAP), or t-SNE spaces. Here, we evaluated the fit only with QNP and QDM metrics. To reduce the computation time, we evaluated less preprocessing combinations, using only the Seurat recipe, the Leiden clustering algorithm, scaling and 25, 50, 100 or 500 PCs.

Supplementary References

[GZ10] Alexander Gorban and Andrei Zinovyev. Principal manifolds and graphs in practice: From molecular biology to dynamical systems. *International journal of neural systems*, 20:219–32, 06 2010.

[LA13] U. Luxburg and M. Alamgir. Density estimation from unweighted k-nearest neighbor graphs: A roadmap. *Advances in Neural Information Processing Systems*, 01 2013.

[LBSN13] Thomas Low, Christian Borgelt, Sebastian Stober, and Andreas Nrnberger. The Hubness Phenomenon: Fact or Artifact? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, volume 285, pages 267–278. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-30277-0 978-3-642-30278-7.

[Li11] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4, 01 2011.

[LSL20] Jingxin Liu, You Song, and Jinzhi Lei. Single-cell entropy to quantify the cellular order parameter from single-cell RNA-seq data. *Biophysical Reviews and Letters*, 15:1–15, 02 2020.

[MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

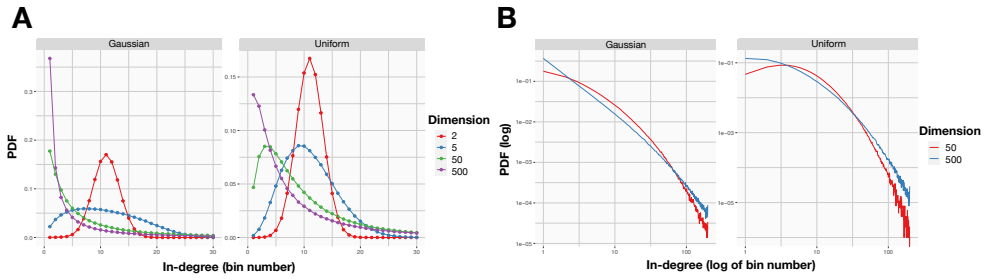
[SCTS19] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37:1, 05 2019.

[SZMZ19] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, 20(1): 269, 2019.

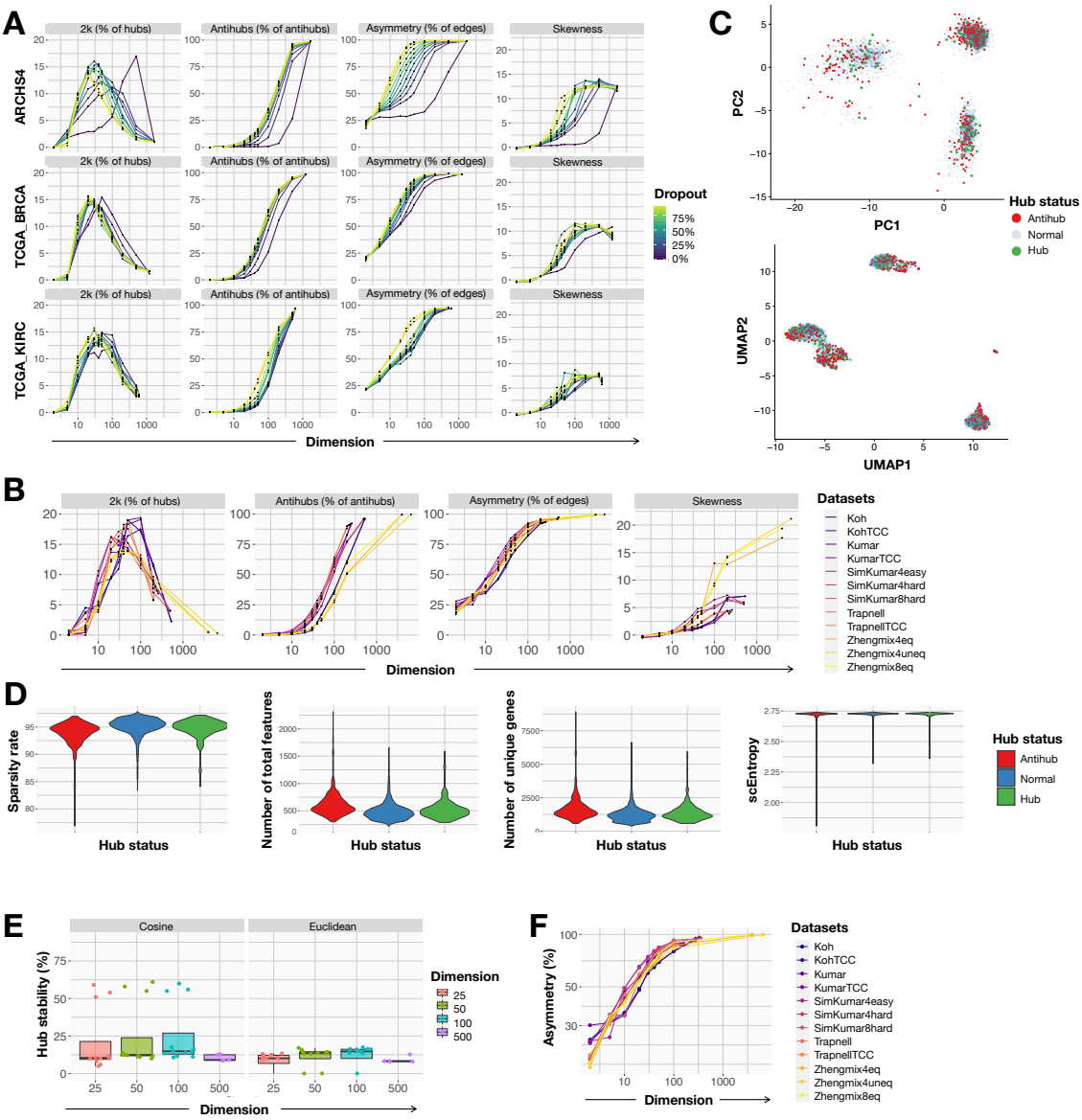
[vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

3 Supplementary Figures

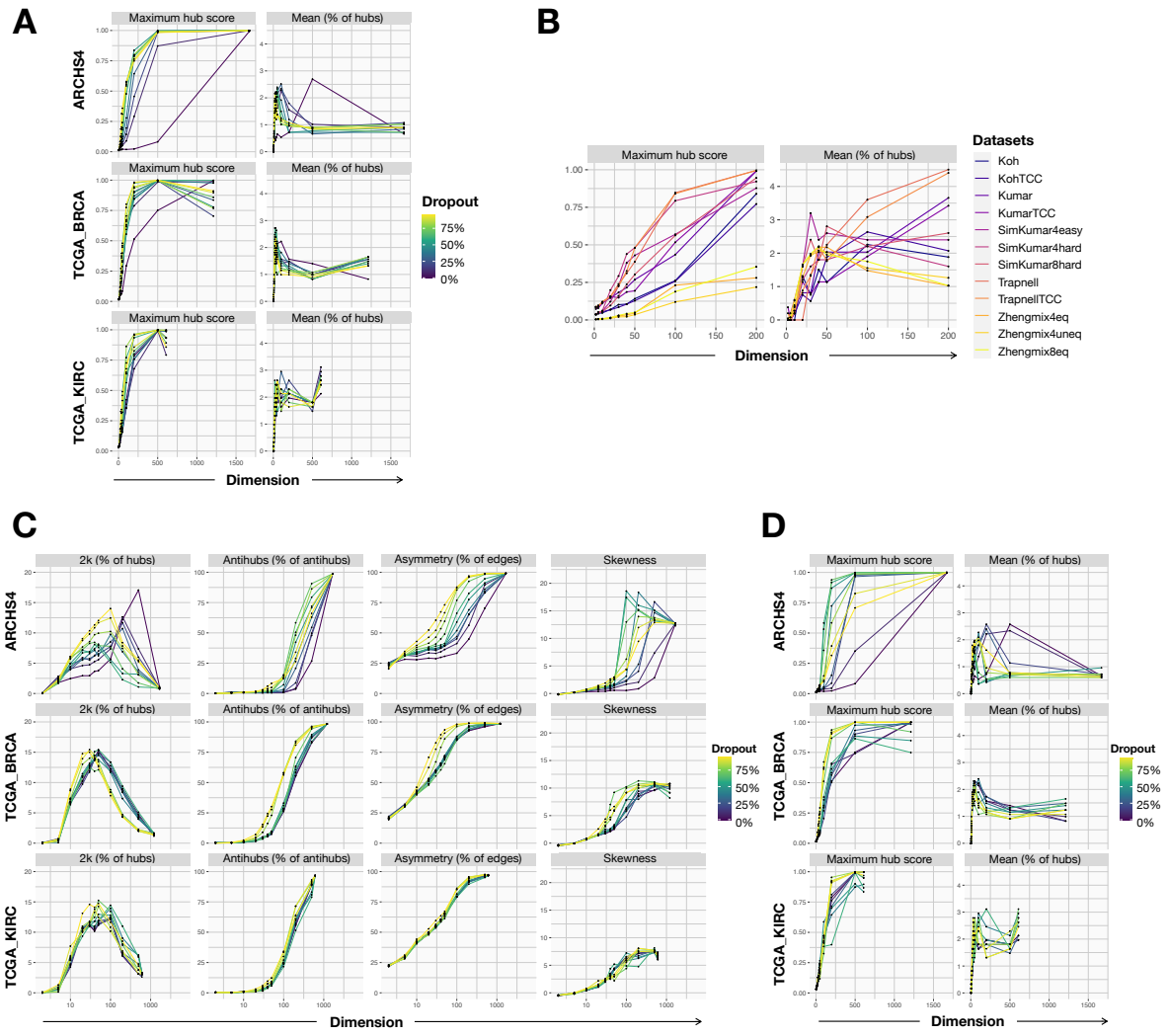
3.1 Hubness in sequencing data



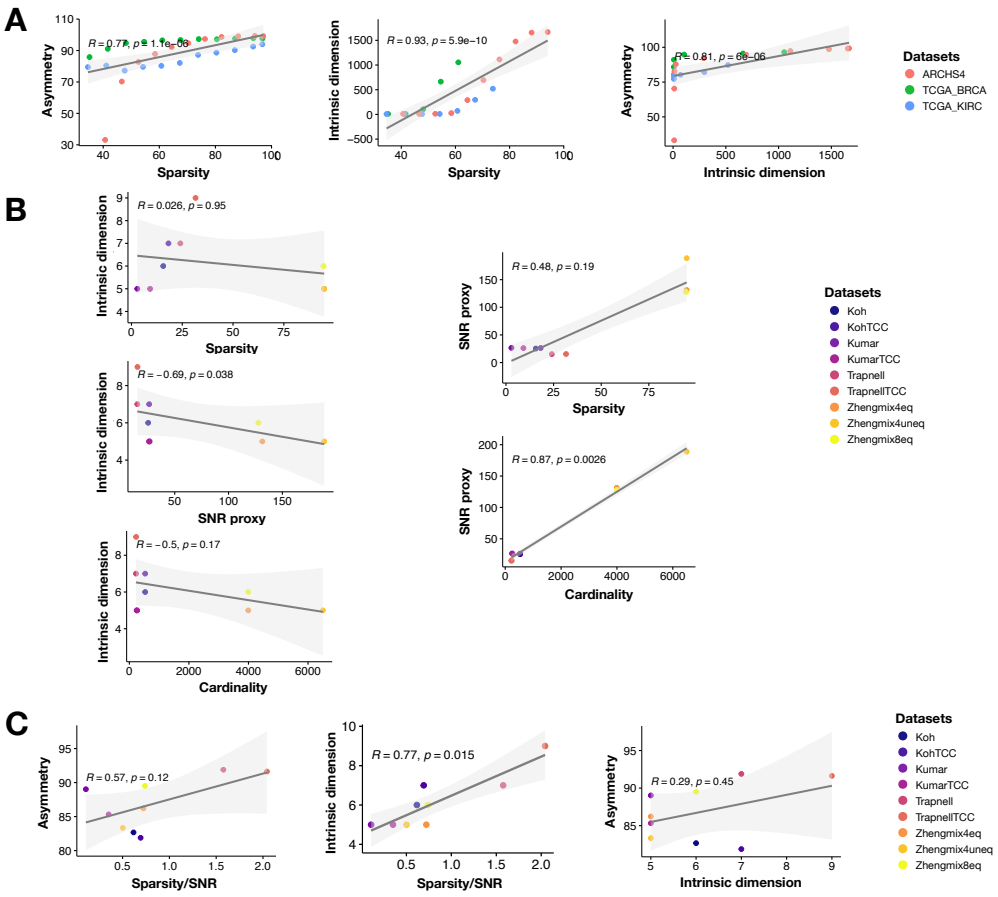
Supplementary Figure 1: Distribution of in-degree values. **(A)** Density of in-degrees for Gaussian (left panel) and uniformly sampled hypercube (right panel) distributions, with dimensions from 2 to 500; a fat tail appears with higher dimensions in both distributions. **(B)** Log-log plot of in-degree distributions for Gaussian (left panel) and uniformly sampled hypercube (right panel) distributions, with dimensions from 2 to 500; the fat tail can be linearly approximated, in both distributions.



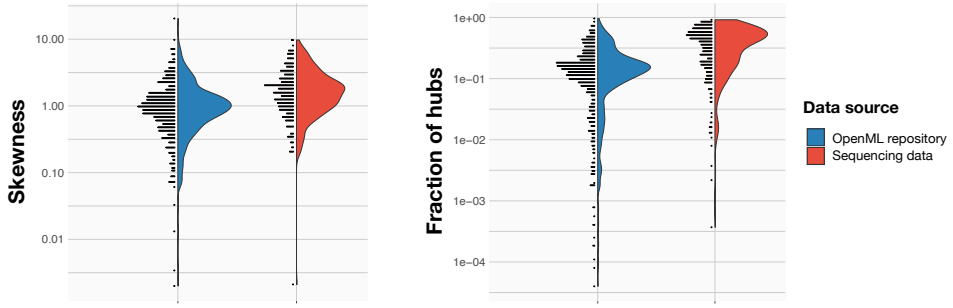
Supplementary Figure 2: Hubness in sequencing data. We quantify hubness with 4 different estimators: percentage of hubs in the data defined as cells with an in-degree above 2k, using the same value for k as the one chosen to build the k-NN graph (first column), percentage of antihubs (second column), asymmetry of the k-NN graph (third column), skewness of the in-degree distribution (fourth column). The quantification is shown as a function of the dimension after PCA reduction **(A,B)**. Hubness quantification methods are applied to 3 bulk datasets, with various rates of simulated dropout **(A)**, or to single-cell datasets **(B)**. **(C)** Position of antihubs and reverse-coverage hubs in the PCA and UMAP projections; example with the Zhengmix4eq single-cell dataset. **(D)** Quality control metrics measured on antihubs, normal cells and hubs defined by reverse-coverage: dropout rate distribution (first column), number of total features counted (second column), number of unique genes detected (third column), single-cell entropy (scEntropy) distribution (fourth column); example with the Zhengmix4eq single-cell dataset. **(E)** We count the proportion of reverse-coverage hubs that are in common between the original data and resampled data upon random removal of 10% of the cells, using scRNA-seq datasets (Duo et al., 2018). **(F)** Asymmetry of the k-NN graph over dimension upon removal of reverse-coverage hubs, with scRNA-seq datasets (Duo et al., 2018), in order to evaluate the resulting magnitude of the hubness phenomenon.



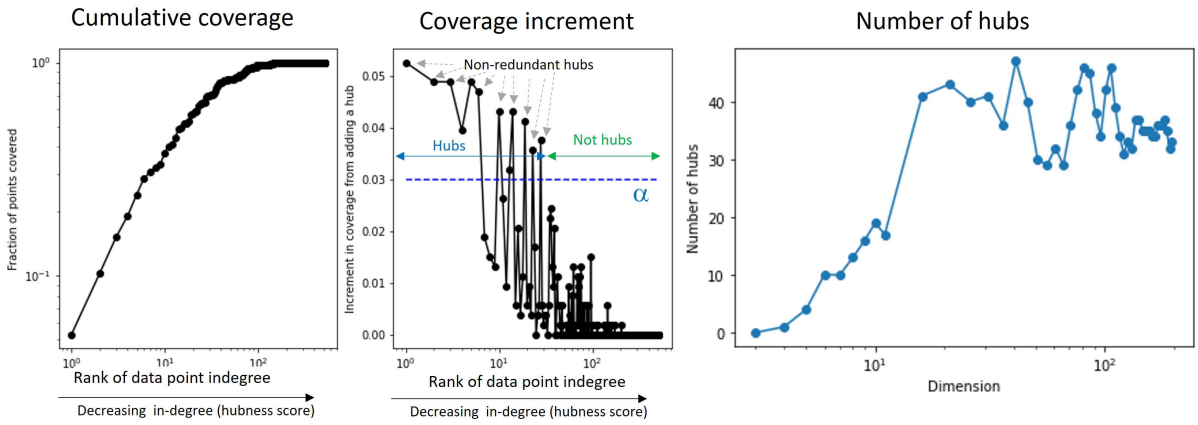
Supplementary Figure 3: Hubness in sequencing data. We quantify hubness with 2 alternative different estimators: maximum hubness score (first column), percentage of hubs as cells with an in-degree above $\mu + 3\sigma$ (second column). The quantification is shown as a function of the dimension after PCA reduction **(A,B)**. Hubness quantification methods are applied to 3 bulk datasets, with various rates of simulated dropout **(A)**, or to single-cell datasets from Sun et al., 2019 **(B)**. **(C)** Classical hubness quantification methods applied to bulk datasets, with various rates of Splatter-simulated dropout. **(D)** Alternative hubness quantification methods applied to bulk datasets, with various rates of Splatter-simulated dropout.



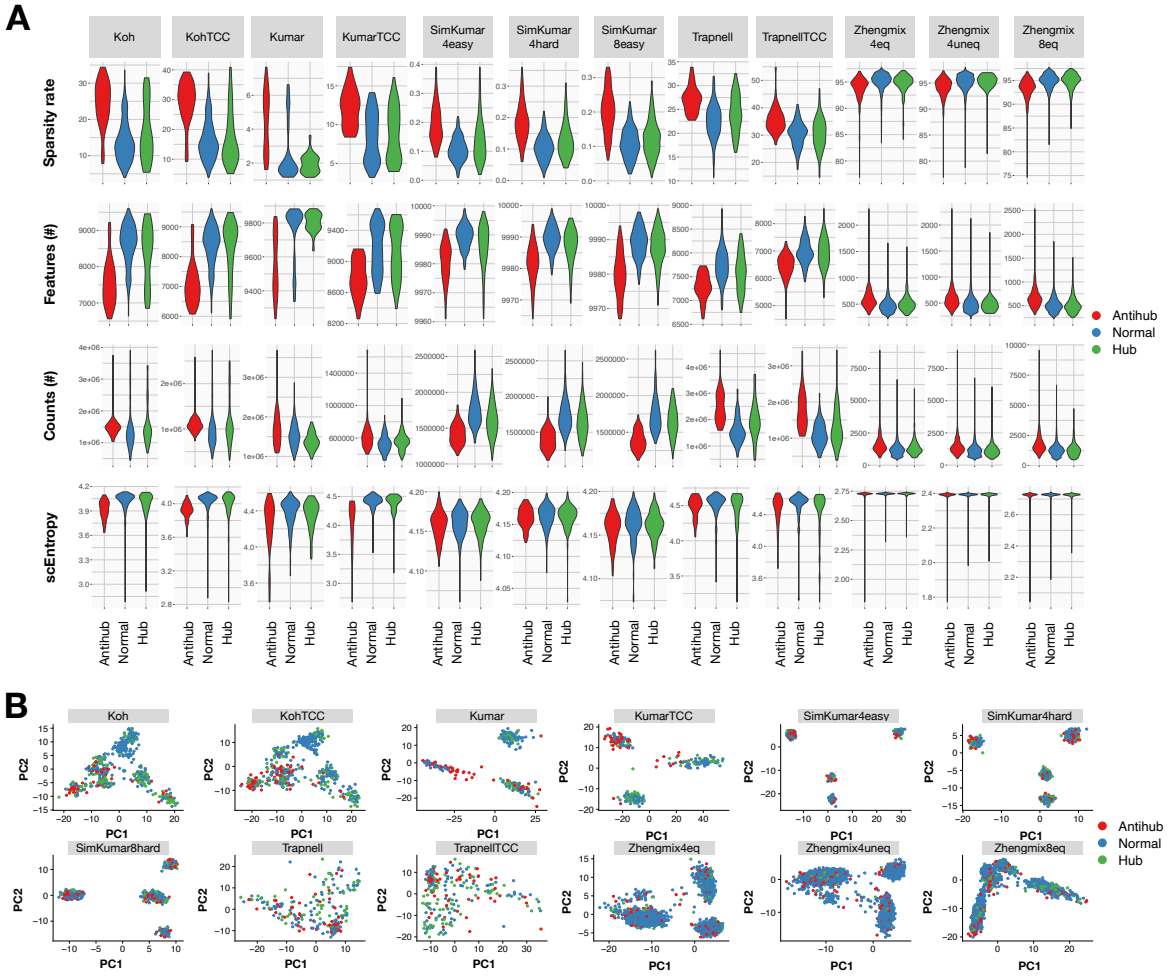
Supplementary Figure 4: Emergence of hubness relates to sparsity, SNR and GID. **(A)** We investigate the link between hubness and sparsity, by showing the Pearson correlation of sparsity to GID and of GID to hubness, using the bulk datasets with various rates of simulated dropout. **(B)** In the first column, we show the correlation between three parameters and GID: sparsity (first row), SNR (second row) and cardinality (third row). In the second column, we test the independence of these three parameters: SNR and sparsity are independent (first row) while SNR and cardinality are dependent (second row). **(C)** We investigate the link between hubness and the ratio of sparsity to SNR, by showing the Pearson correlation of the ratio sparsity/SNR to GID and of GID to hubness, using the 9 real single-cell datasets from Sun et al., 2019.



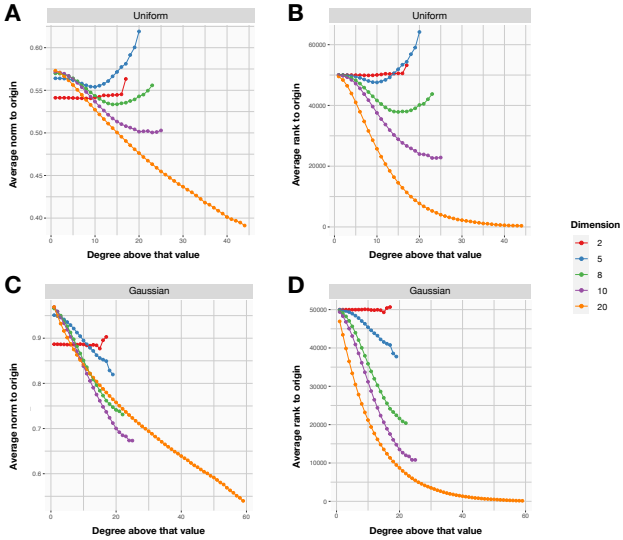
Supplementary Figure 5: Susceptibility to hubness across datasets from various sources. Left plot: k-Skewness for 501 datasets from the openML repository and sequencing data. p-value¡10^{-4}, randomization test. Right plot: Fraction of hubs evaluated as data points with an in-degree above $2k$. p-value¡10^{-4}, randomization test



Supplementary Figure 6: Reverse-coverage hub-retrieving method. **Left:** Size of the reverse coverage (or percentage of reverse-covered data) as a function of the number of putative hubs, i.e. the number of cells with the highest in-degrees, computed for a single-cell dataset from Sun et al., 2019. **Middle:** Increment of the size of the reverse coverage (proportion of new data points covered from adding a new hub). **Right:** Dependence of the number of identified hubs on the dimensionality of the dataset reduced by PCA.

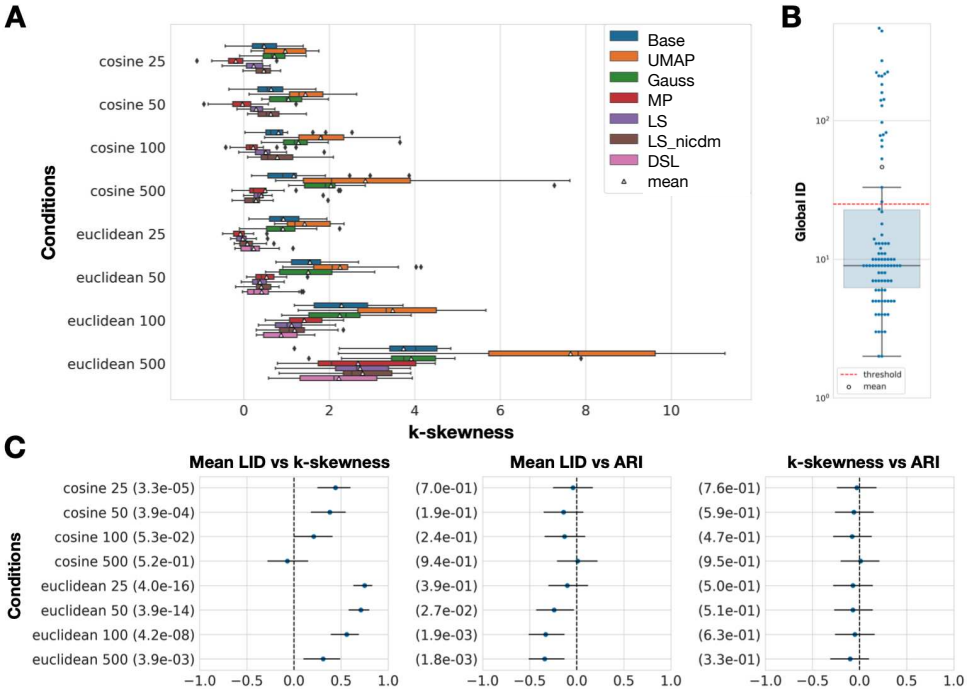


Supplementary Figure 7: Quality control metrics. **(A)** Quality control metrics distribution for hubs, antihubs and normal cells on the 12 datasets from Sun et al., 2019: dropout rate (first row), number of total features (second row), number of unique genes (third row), single-cell entropy (last row). **(B)** PCA projections showing the positions of hubs, antihubs and normal cells for the 12 Duo datasets.

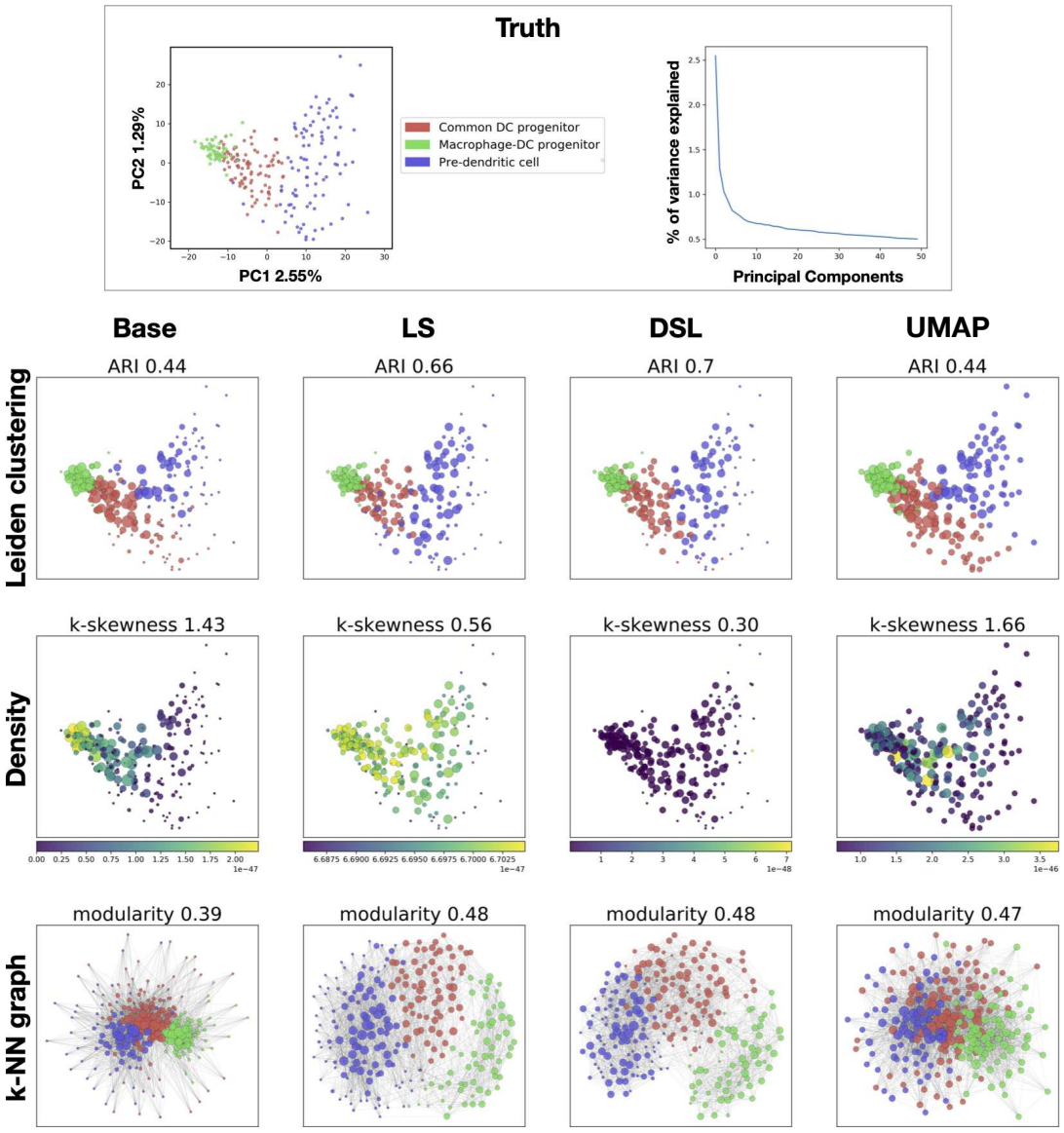


Supplementary Figure 8: Hubs positions. Position of hubs for uniform **(A,B)** and Gaussian **(C,D)** data distribution. **(A,C)** Average norm of points with an in-degree above the abscissa value. **(B,D)** Average ranking to the origin of points with an in-degree above the abscissa value.

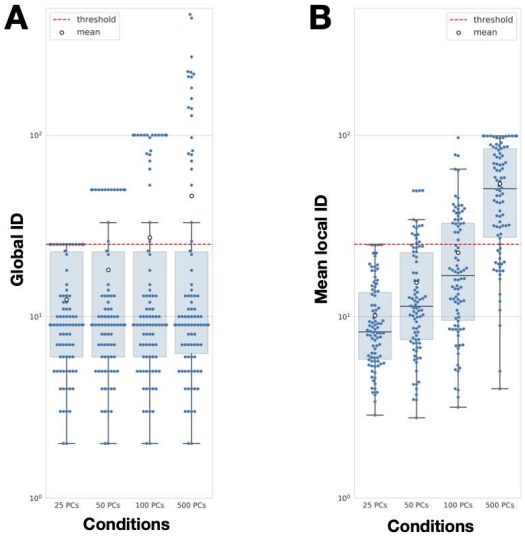
3.2 Hubness reduction improves clustering



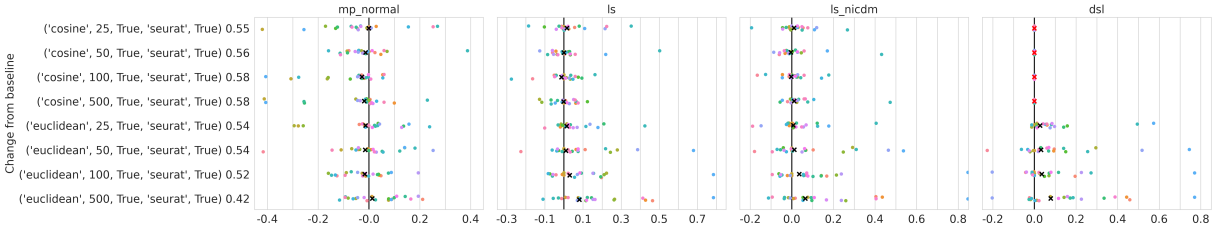
Supplementary Figure 9: Evaluation of hubness reduction effect on clustering performance. **(A)** k-skewness of high-ID datasets, as a function of the metric, dimension and k-NN graph production method used; example with the Seurat recipe and scaling. **(B)** Measure of the global ID (GID) of all datasets, used to define the high-ID datasets. **(C)** Pearson correlation coefficients distribution (p-value in parentheses) between k-skewness, ARI and mean local ID (LID) for all datasets, calculated with the base k-NN graph, using the Seurat recipe, scaling and the Leiden algorithm.



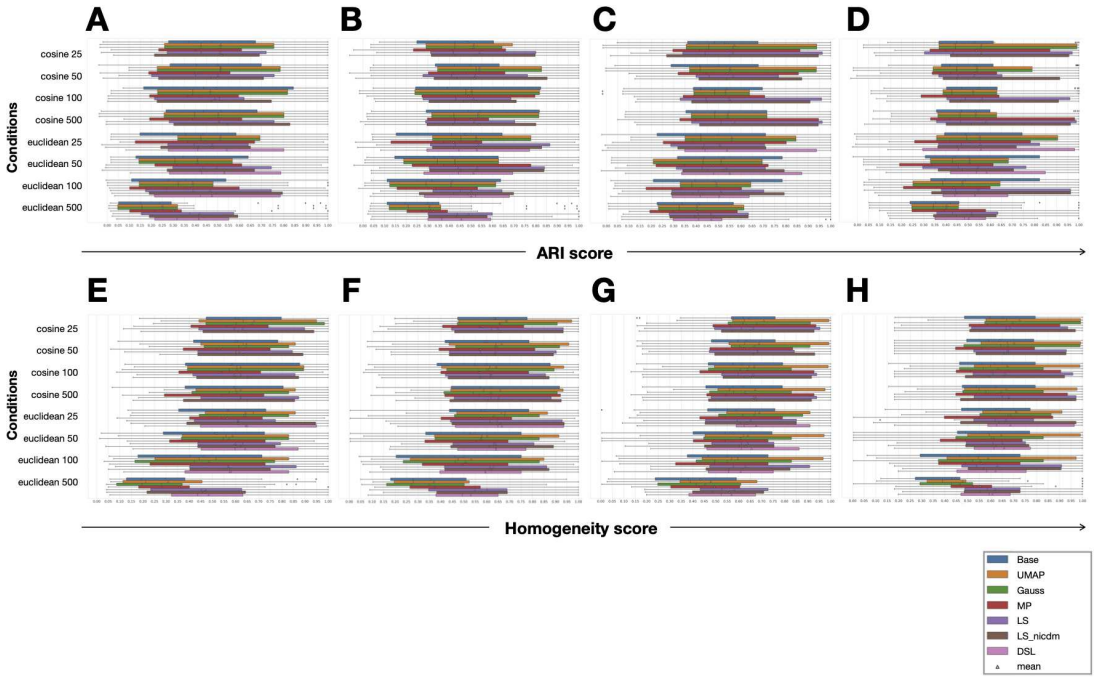
Supplementary Figure 10: A selected example of Leiden clustering on a scRNA-seq dataset with FACS-labelled mouse blood dendritic cells (GSE60783), using Euclidean distance, 50 PCs and 15-NN graph. Clustering with the usual k-NN graph (Base) or the UMAP k-NN graph (scanpy_umap) results in lower ARI while Local Scaling (LS) and DisSimLocal (DSL) k-NN graphs yield better accuracy. Both hubness-reduced and UMAP k-NN graphs produce more uniform Gaussian kernel density estimates. However, unlike hubness reduction, UMAP k-NN graph does not reduce the skewness of the in-degree distribution. The modularity is improved for the UMAP and hubness-reduced graphs compared to the base one, although the UMAP graph looks more intricate by eye. Point size is proportional to the in-degree in the respective k-NN graph.



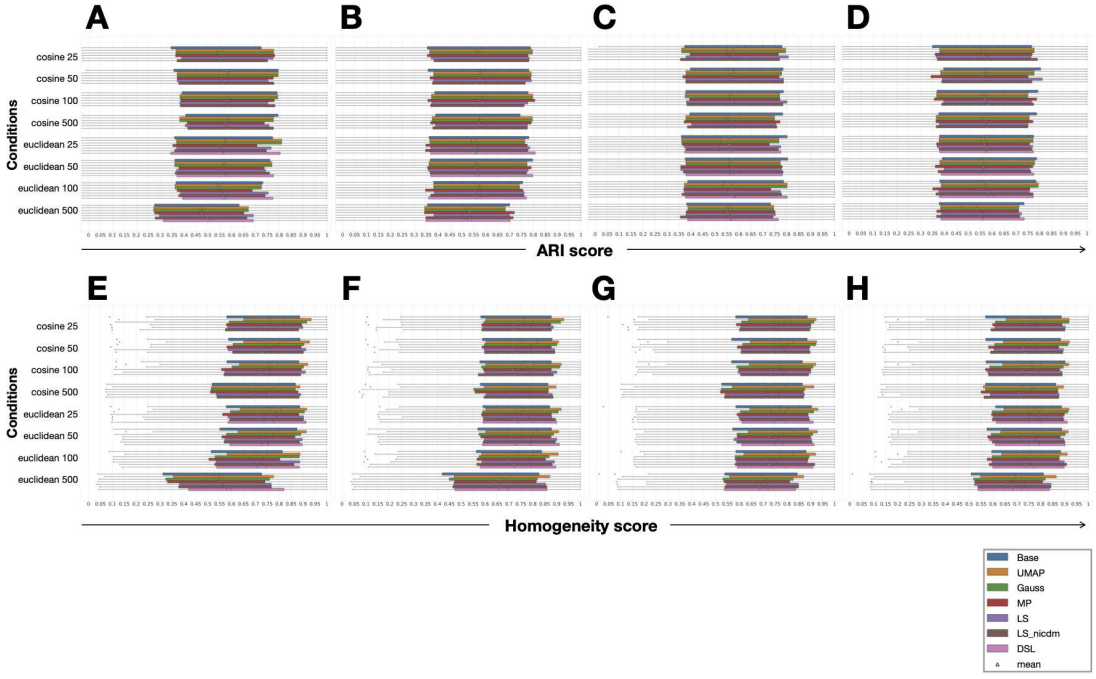
Supplementary Figure 11: GID **(A)** and mean LID **(B)** of datasets estimated with PCA, the Seurat preprocessing and scaling.



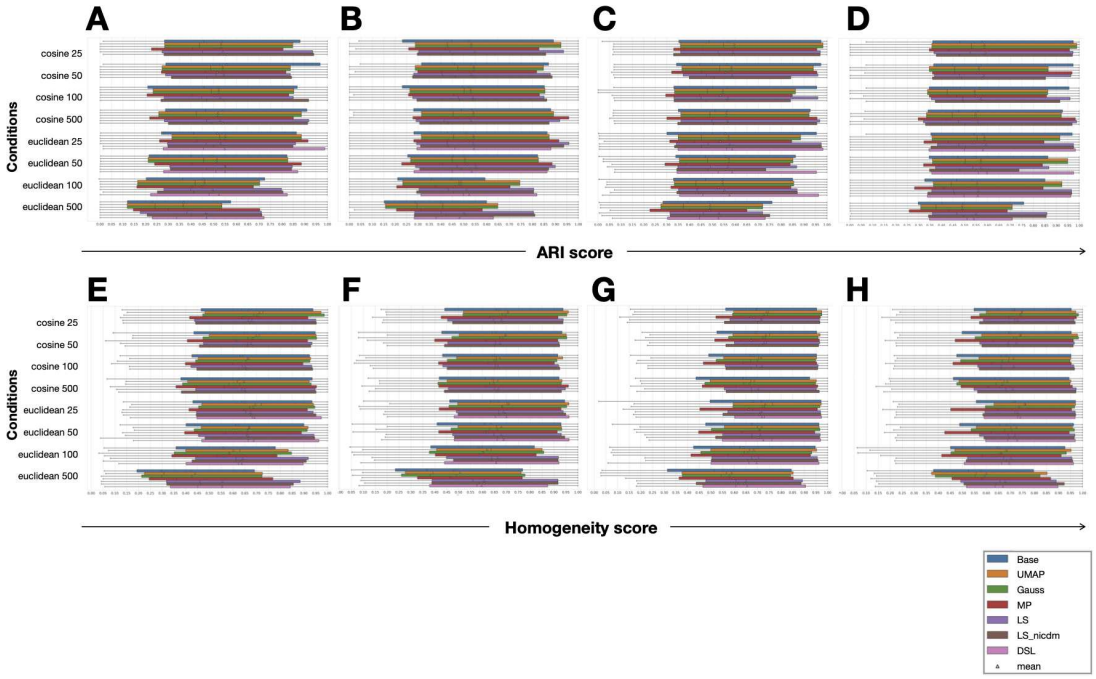
Supplementary Figure 12: Relative difference in ARI, with the base k-NN graph performance as reference, for the Seurat preprocessing and the Leiden algorithm for the high-ID datasets.



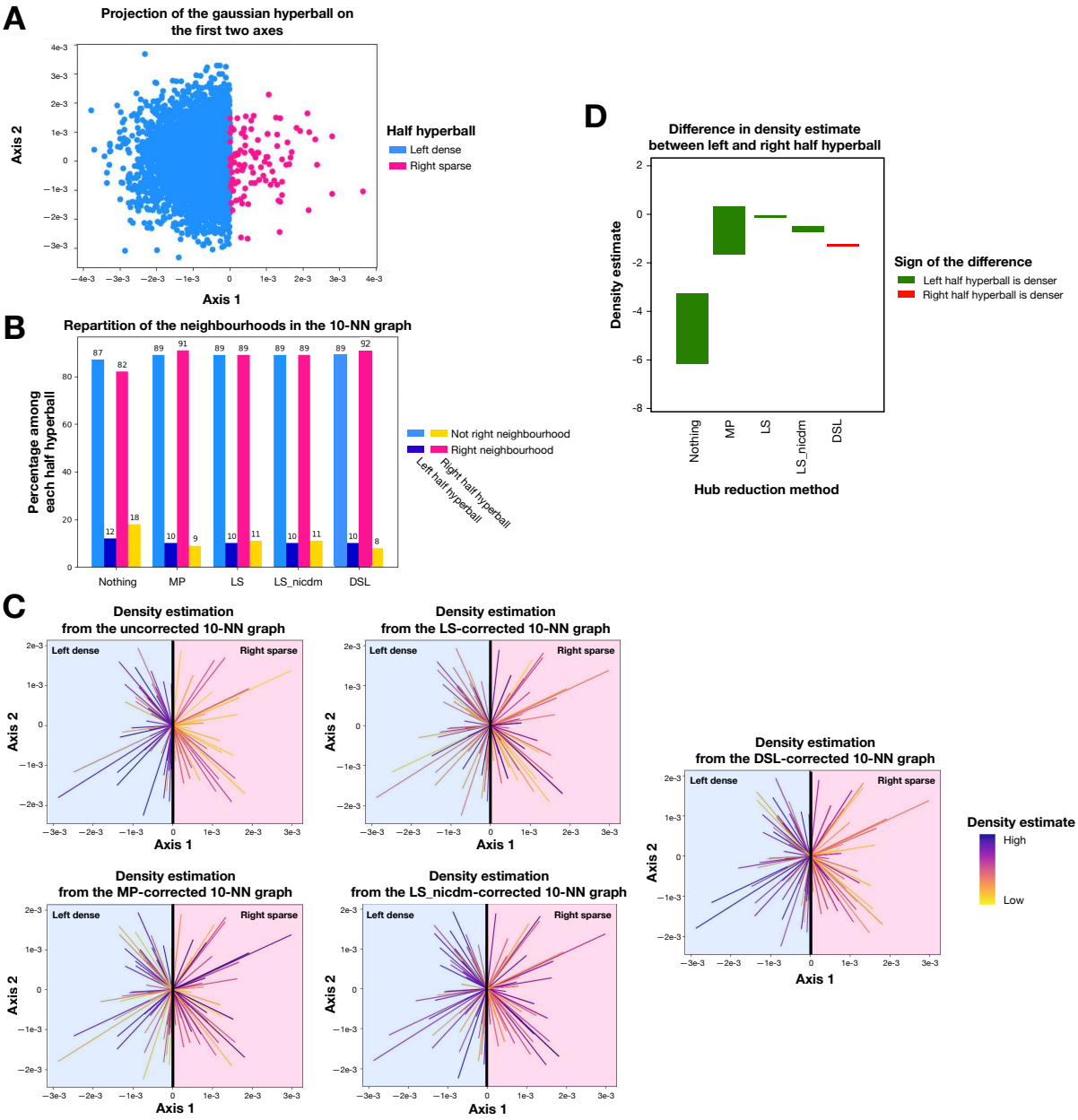
Supplementary Figure 13: Clustering scores done with the Seurat **(A,B,E,F)** or Duo **(C,D,G,H)** preprocessing with **(A,C,E,G)** or without **(B,D,F,H)** scaling, for high-ID datasets: ARI scores **(A,B,C,D)** and homogeneity scores **(E,F,G,H)**.



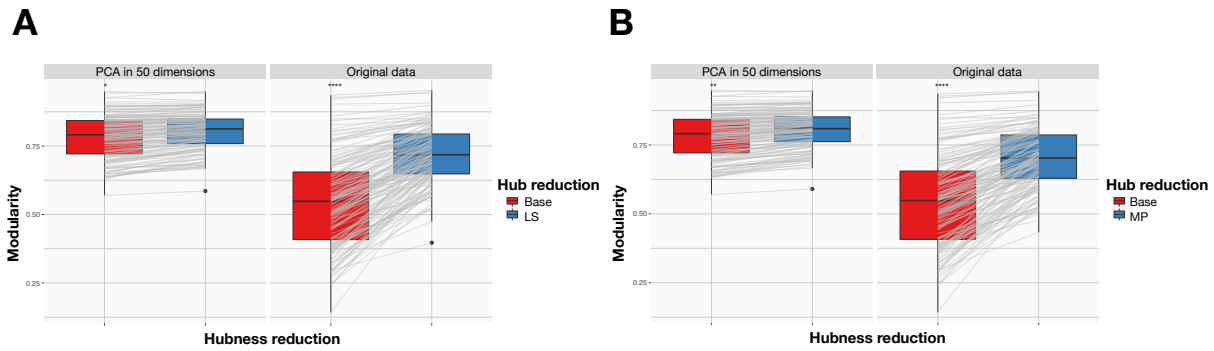
Supplementary Figure 14: Clustering scores done with the Seurat **(A,B,E,F)** or Duo **(C,D,G,H)** preprocessing with **(A,C,E,G)** or without **(B,D,F,H)** scaling, for low-ID datasets: ARI scores **(A,B,C,D)** and homogeneity scores **(E,F,G,H)**.



Supplementary Figure 15: Clustering scores done with the Seurat **(A,B,E,F)** or Duo **(C,D,G,H)** preprocessing with **(A,C,E,G)** or without **(B,D,F,H)** scaling, for gold-standards datasets: ARI scores **(A,B,C,D)** and homogeneity scores **(E,F,G,H)**.

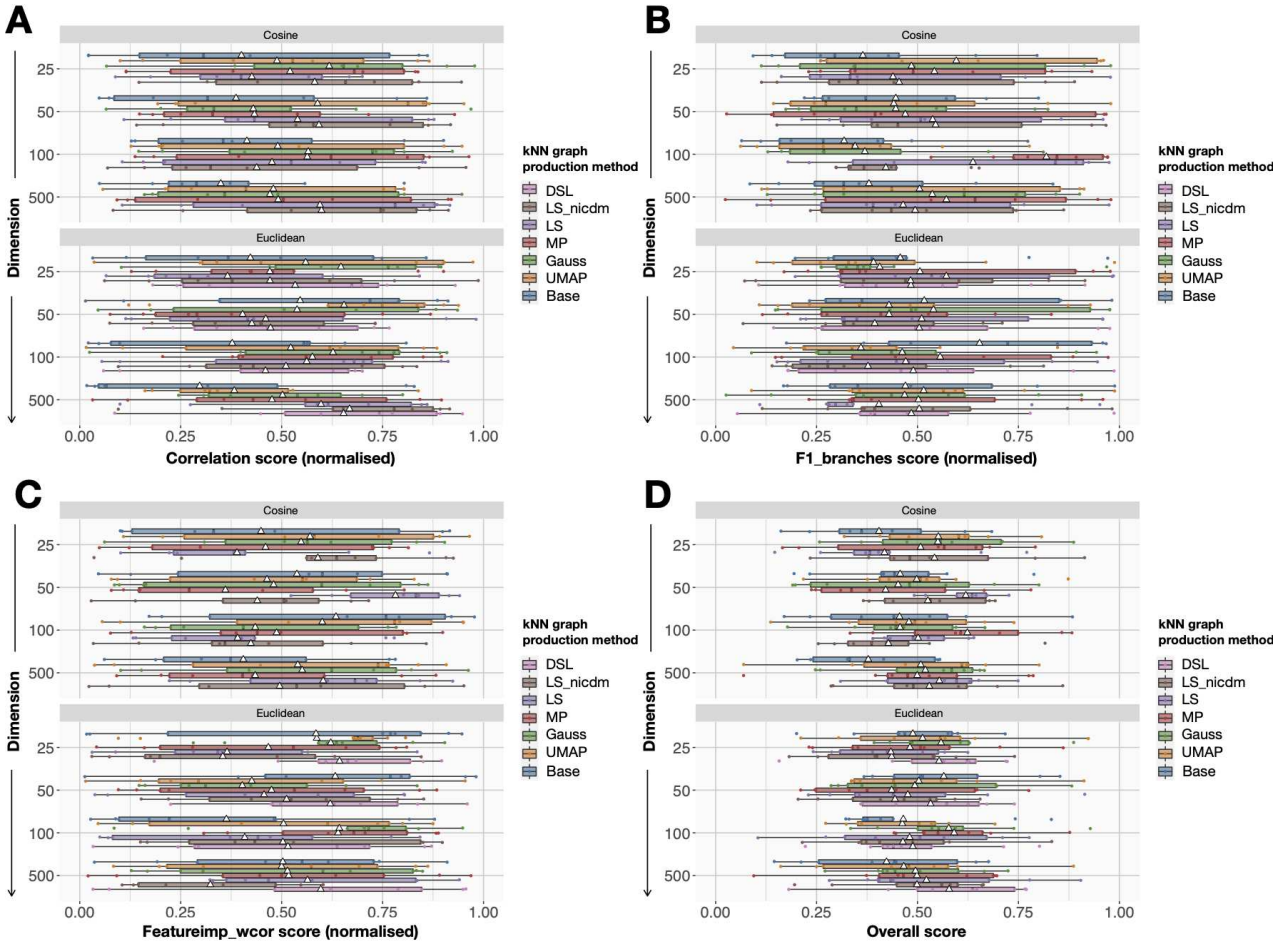


Supplementary Figure 16: Correcting heterogeneous densities with hubness reduction. **(A)** Gaussian ball in 10 dimensions, with 5000 points in the half hyperball left of the hyperplan x=0 and 100 points in the right half hyperball, projected on the first two dimensions. **(B)** Proportion of points of each half hyperball in the neighborhood of the right half hyperball. **(C)** Visualization of the density estimate calculated from the unweighted k-NN graph before and after hubness reduction. The source of the density estimate is at the center of the Gaussian ball and the targets are picked randomly in each half hyperball. Each line connect the source and a target and its color represents the density estimate. The pale background colors represent the two half spaces: blue for the left one, pink for the right one. **(D)** Difference in the density estimates between the left and right half hyperballs. Each edge of a bar is the mean density estimate in one of the half hyperballs. If the rectangle is green, the lower border is the estimate from the right half hyperball; if it is red, it is from the left one.

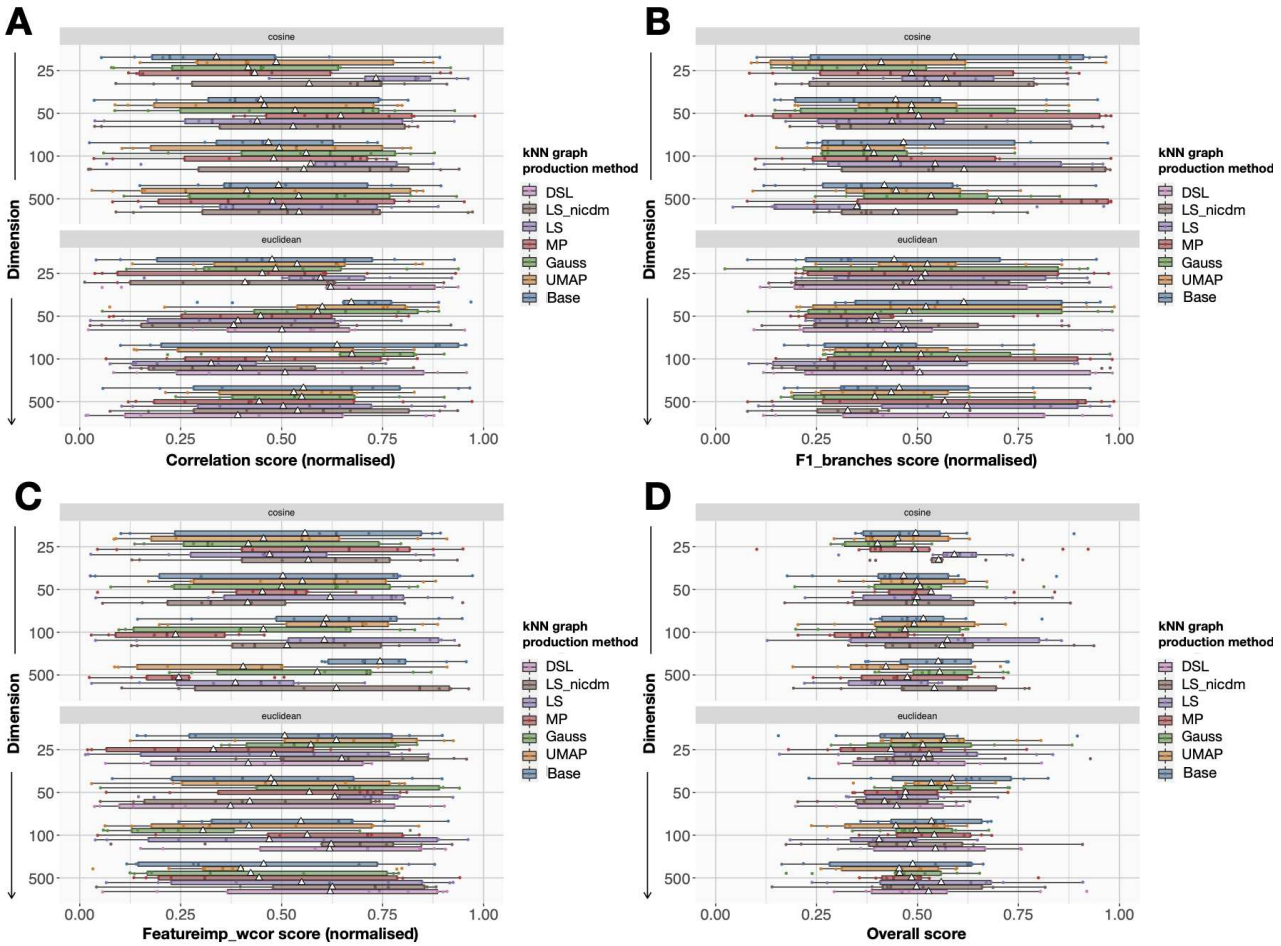


Supplementary Figure 17: Modularity improvement upon hubness reduction. Per-dataset modularity of the Louvain clustering with (left panel) or without PCA (right panel). Comparison between the modularity with or without hubness reduction, performed with the LS **(A)** or the MP algorithm **(B)**.

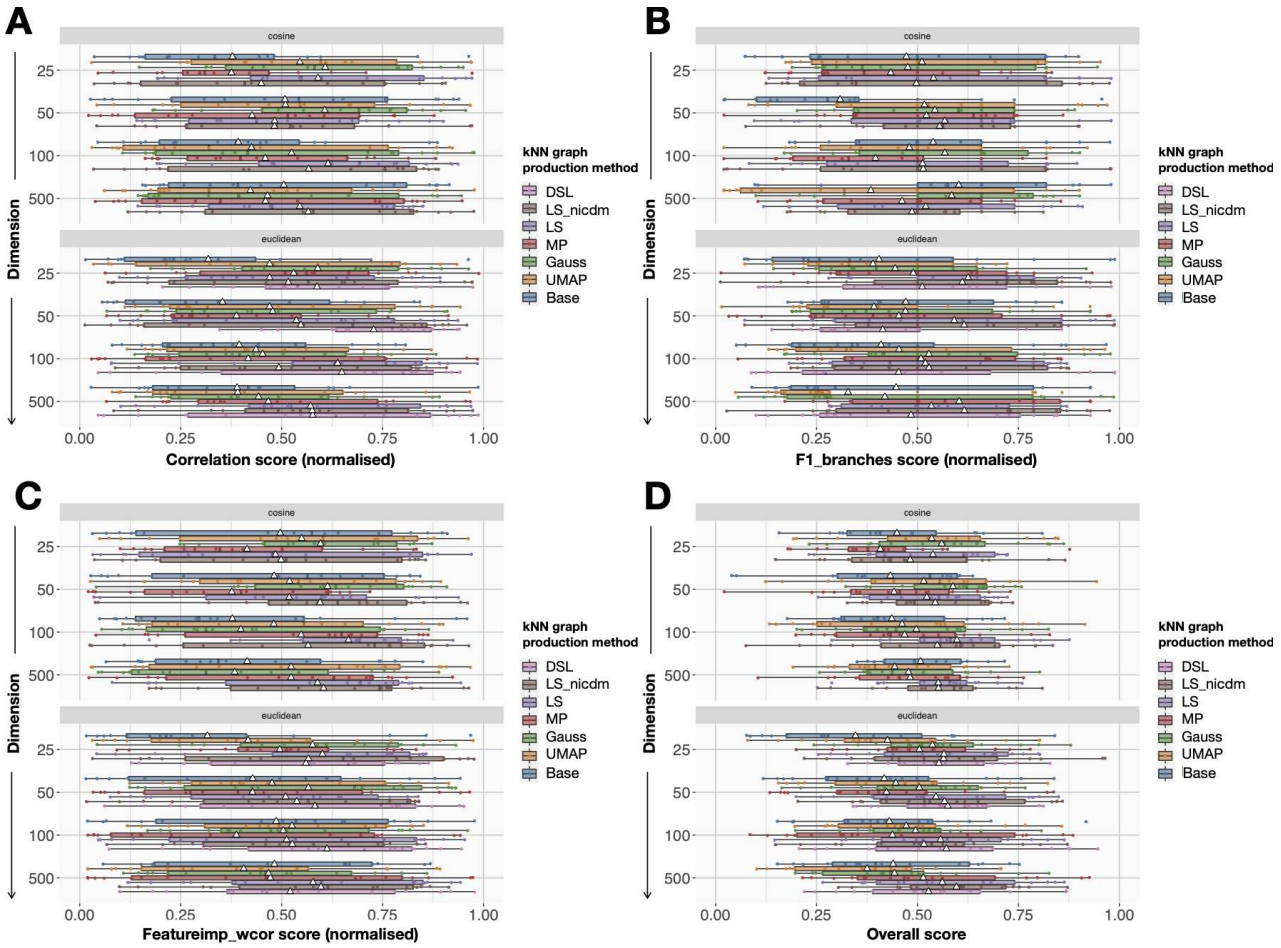
3.3 Hubness reduction improves TI



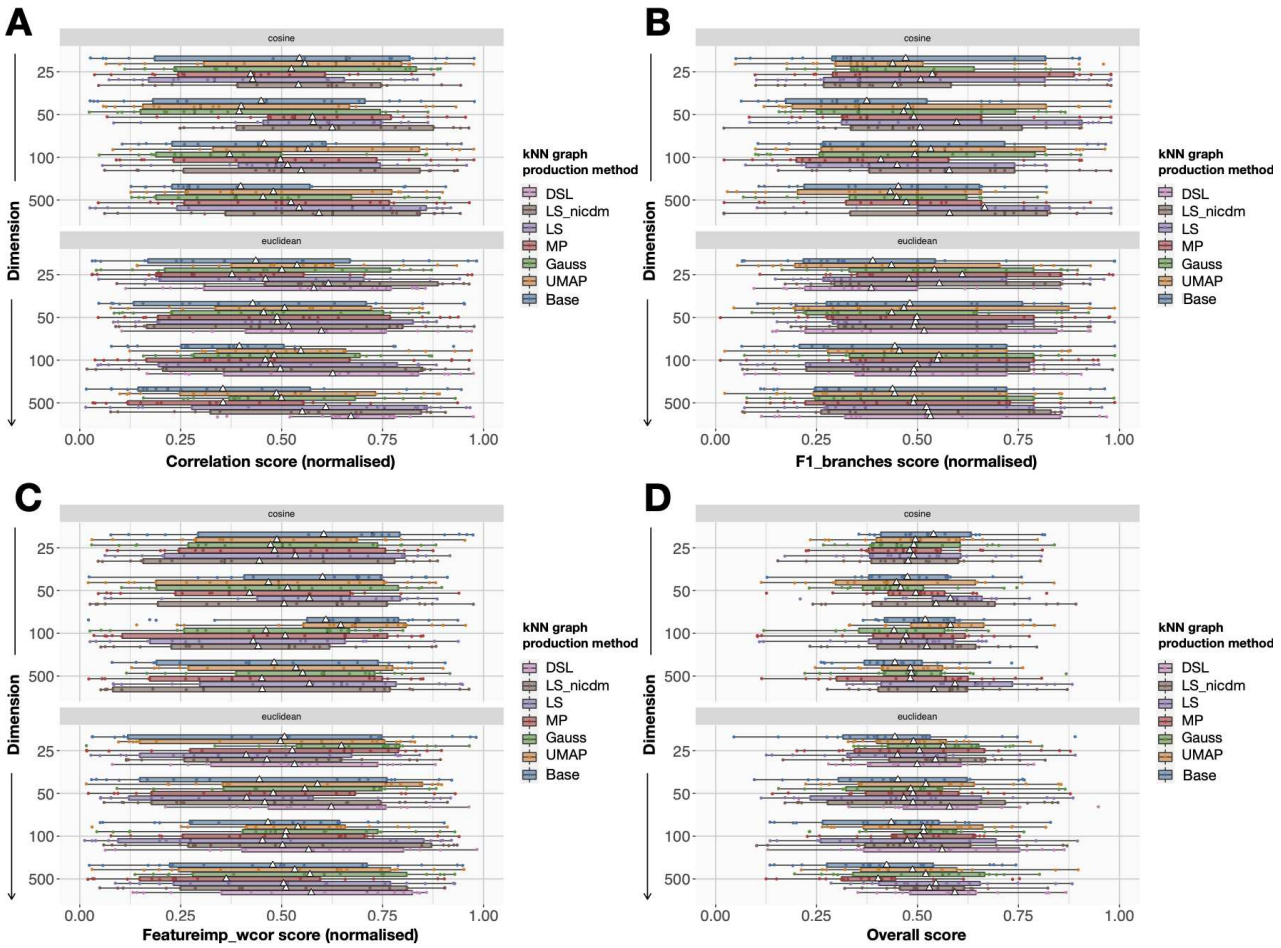
Supplementary Figure 18: Per-dataset trajectory inference scores on low-ID datasets, using the Seurat recipe and Leiden clustering. **(A)** Detailed correlation scores. **(B)** Detailed F1_branches scores. **(C)** Detailed featureimp_wcor scores. **(D)** Detailed overall score.



Supplementary Figure 19: Per-dataset trajectory inference scores on low-ID datasets, using the Duo recipe and Leiden clustering. **(A)** Detailed correlation scores. **(B)** Detailed F1_branches scores. **(C)** Detailed featureimp_wcor scores. **(D)** Detailed overall score.

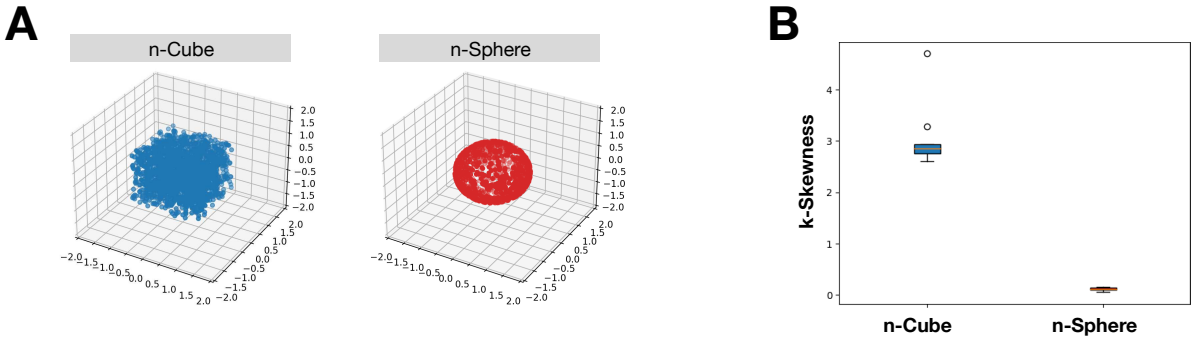


Supplementary Figure 20: Per-dataset trajectory inference scores on high-ID datasets, using the Seurat recipe and Leiden clustering. **(A)** Detailed correlation scores. **(B)** Detailed F1_branches scores. **(C)** Detailed featureimp_wcor scores. **(D)** Detailed overall score.

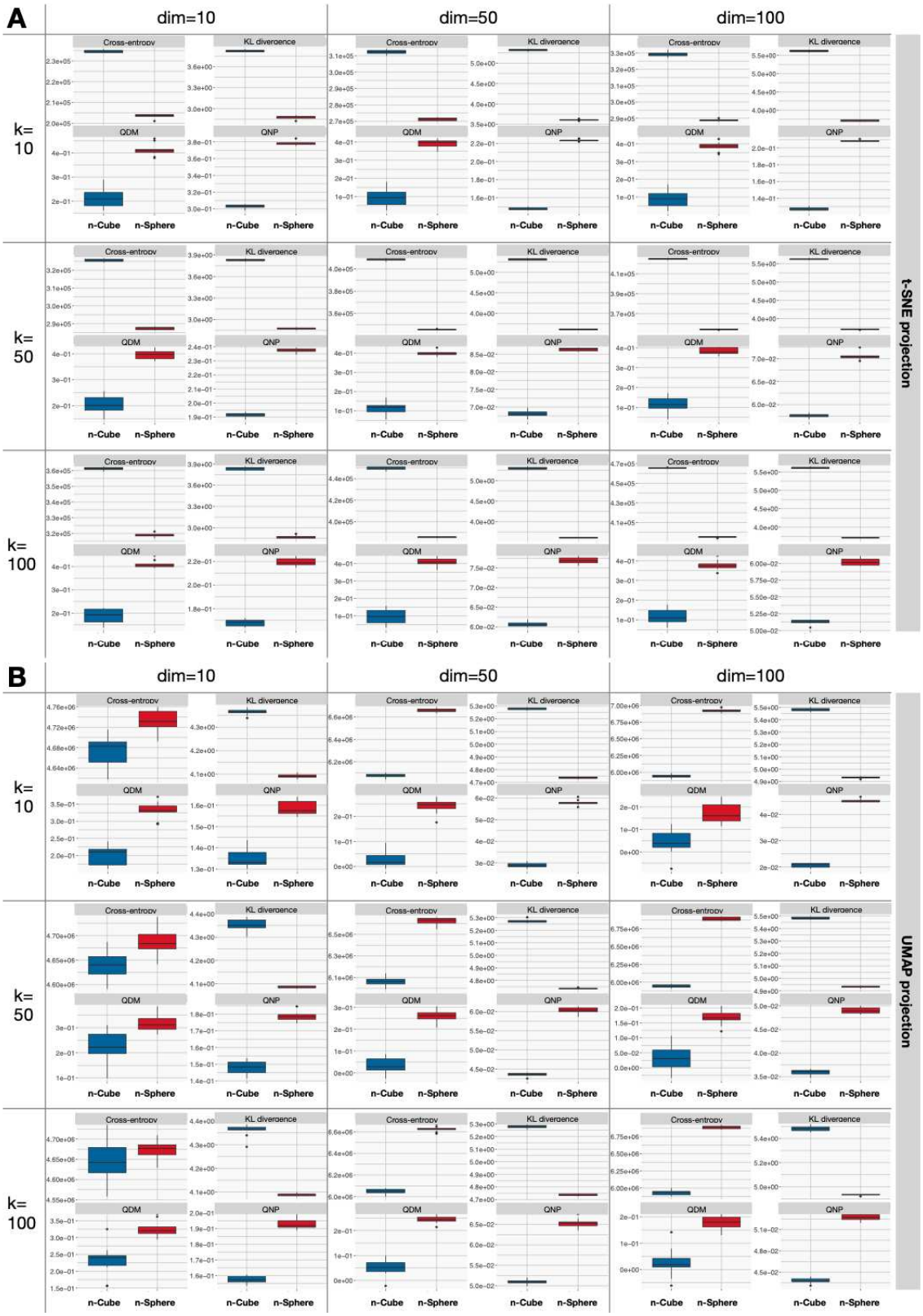


Supplementary Figure 21: Per-dataset trajectory inference scores on high-ID datasets, using the Duo recipe and Leiden clustering. **(A)** Detailed correlation scores. **(B)** Detailed F1_branches scores. **(C)** Detailed featureimp_wcor scores. **(D)** Detailed overall score.

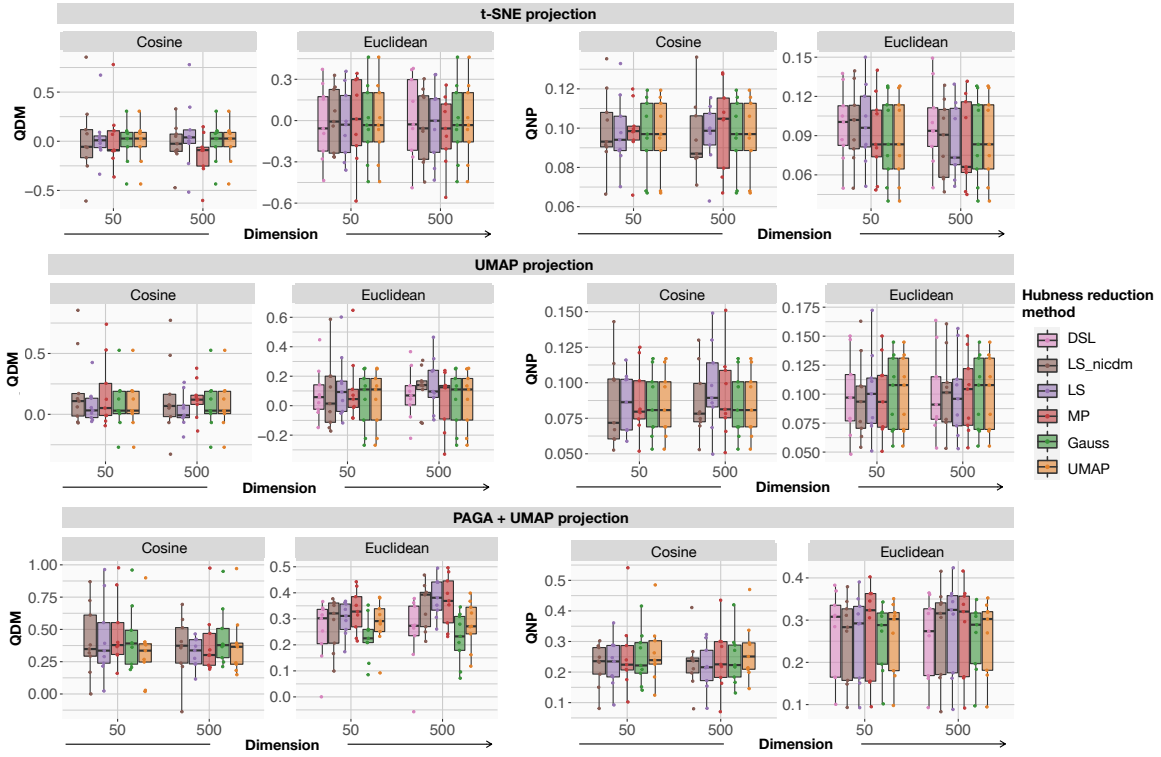
3.4 Hubness reduction improves visualisation



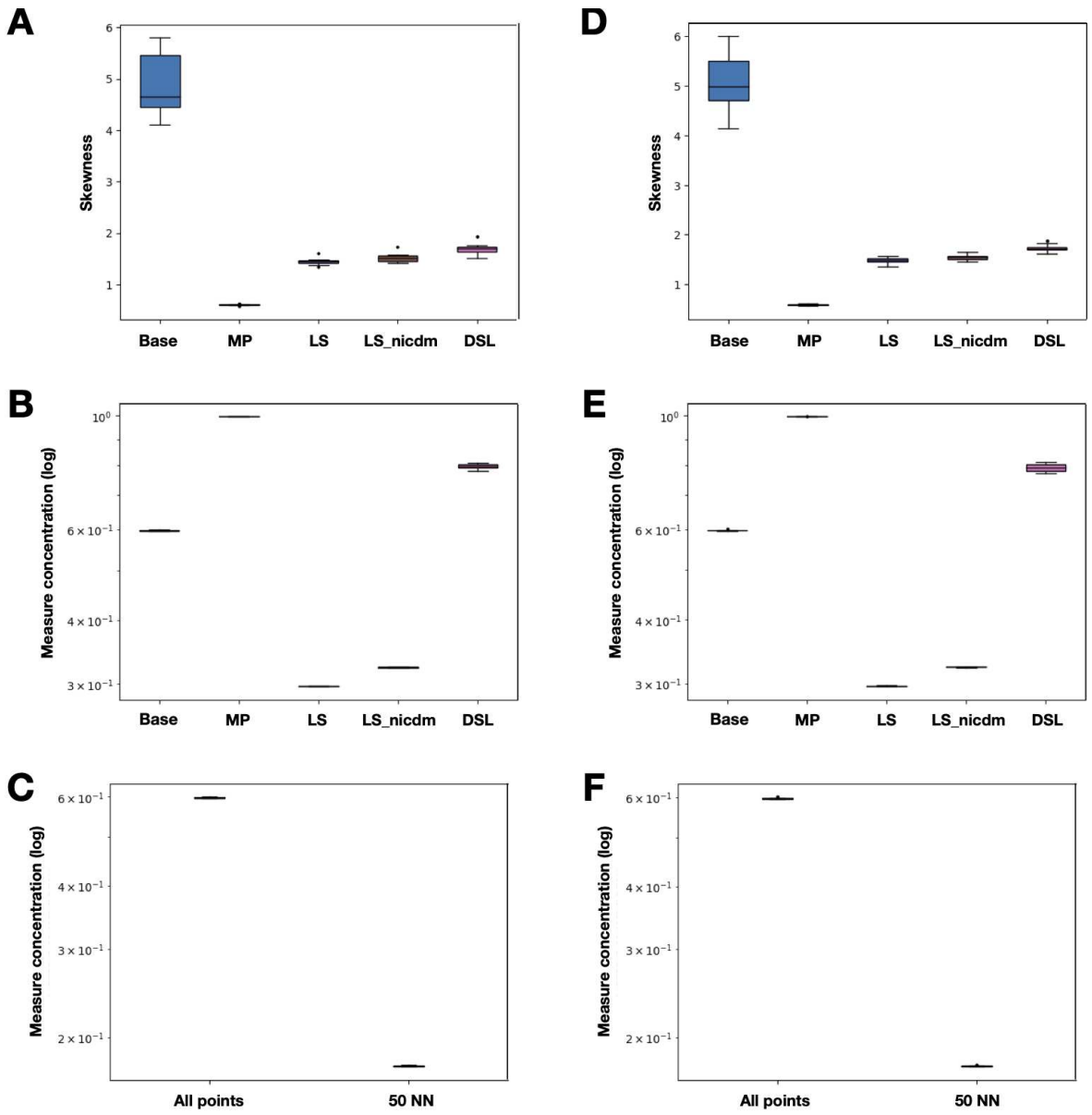
Supplementary Figure 22: n-Cube and n-Sphere. **(A)** 3D plot of a 3-Cube (left) and a 3-Sphere (right). **(B)** k-Skewness of a 50-Cube and a 50-Sphere, each containing 5,000 points, and k=50, computed 10 times.



Supplementary Figure 23: Visualisation quality metrics of the n-Cube and the n-Sphere. We show the n-cube and the n-sphere after t-SNE **(A)** or UMAP **(B)** projections, with different values for the size of the neighborhood k and the number of dimensions n

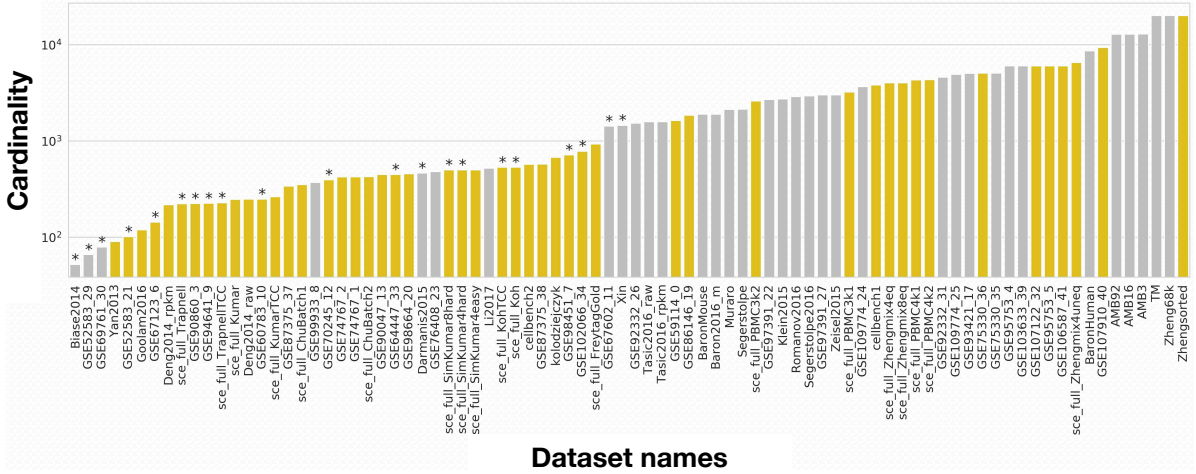


Supplementary Figure 24: QDM and QNP before or after hubness reduction, evaluated after various visualisation algorithms, compared to the PCA with 500 PCs, for low-ID datasets. We project low-ID datasets either with t-SNE (top row), UMAP (middle row) or PAGA+UMAP (bottom row) and evaluate QDM (left column) and QNP (right column). The different projections are computed either with the cosine dissimilarity or the Euclidean metric, and using the two Scanpy k-NN graphs or the four hub-reduced graphs.



Supplementary Figure 25: Evaluation of the impact of hubness correction on hubness and measure concentration. We computed Gaussian distributions, either a single blob **(A,B,C)** or two distinct blobs **(D,E,F)**. Skewness of the data with or without hubness reduction **(A,D)**. Global measure concentration with or without hubness reduction **(B,E)**. Measure concentration without hubness reduction, either considering all points, or the 50 nearest neighbors **(C,F)**

3.5 Datasets



Supplementary Figure 26: Benchmark datasets collected from previous studies. Colors represent gold and silver standards, stars mark high ID. Note that the figure contains less datasets than listed in the supplementary table because there are overlaps between Gulati et al., 2020 and other benchmark studies

Dataset name in benchmark study	Sequencing protocol	Number of cells	Number of features	Number of cluster labels	Description	Ref.	Benchmark study	Label type	Used to evaluate
Koh	SMARTer	531	48981	9	FACS purified H7 human embryonic stem cells in different differention stages	GSE85066	Duo et al. PMC6134335	FACS	Clustering
KohTCC	SMARTer	531	811938	9	FACS purified H7 human embryonic stem cells in different differention stages	GSE85066	Duo et al. PMC6134335	FACS	Clustering
Kumar	SMARTer	246	45159	3	Mouse embryonic stem cells, cultured with different inhibition factors	GSE60749	Duo et al. PMC6134335	Culture conditions	Clustering
KumarTCC	SMARTer	263	803405	3	Mouse embryonic stem cells, cultured with different inhibition factors	GSE60749	Duo et al. PMC6134335	Culture conditions	Clustering
SimKumar4easy	Synthetic dataset	500	43606	4	Simulation using different proportions of differentially expressed genes	PMC6134335	Duo et al. PMC6134335	Simulated data	Clustering
SimKumar4hard	Synthetic dataset	499	43638	4	Simulation using different proportions of differentially expressed genes	PMC6134335	Duo et al. PMC6134335	Simulated data	Clustering
SimKumar8hard	Synthetic dataset	499	43601	8	Simulation using different proportions of differentially expressed genes	PMC6134335	Duo et al. PMC6134335	Simulated data	Clustering
Trapnell	SMARTer	222	41111	3	Human skeletal muscle myoblast cells, differention induced by low-serum medium	GSE52529	Duo et al. PMC6134335	Culture conditions	Clustering
TrapnellTCC	SMARTer	227	684953	3	Human skeletal muscle myoblast cells, differention induced by low-serum medium	GSE52529	Duo et al. PMC6134335	Culture conditions	Clustering
Zhengmix4eq	10x	3994	15568	4	Mixtures of FACS purified peripheral blood mononuclear cells	SRP073767	Duo et al. PMC6134335	FACS	Clustering
Zhengmix4uneq	10x	6498	16443	4	Mixtures of FACS purified peripheral blood mononuclear cells	SRP073767	Duo et al. PMC6134335	FACS	Clustering
Zhengmix8eq	10x	3994	15716	8	Mixtures of FACS purified peripheral blood mononuclear cells	SRP073767	Duo et al. PMC6134335	FACS	Clustering
GSE59114	Smart-seq2	1622	7539	3	Mouse Blood Phenotypes Aging HSCs (Smart-seq2)	GSE59114	Gulati et al. PMID: 31974247	FACS	Clustering
GSE74767	SC3-seq	212	28796	7	Human (SC3-seq)	GSE74767	Gulati et al. PMID: 31974247	Cell lines	Clustering
GSE74767	SC3-seq	421	28796	13	Macaque Embryo Timepoints Blastocyst timepoints (SC3-seq)	GSE74767	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE90860	C1	223	42832	3	Mouse Brain Timepoints Cortical interneurons (C1)	GSE90860	Gulati et al. PMID: 31974247	Timepoints	Clustering/TI/Visualisation
GSE95753	10x	6000	27933	14	Mouse Brain Phenotypes Dentate gyrus phenotypes (10x)	GSE95753	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE95753	10x	6000	27933	8	Mouse Brain Timepoints Dentate gyrus timepoints (10x)	GSE95753	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE67123	Tang et al.	143	24028	5	Mouse Embryo Timepoints Embryonic HSCs (Tang et al.)	GSE67123	Gulati et al. PMID: 31974247	Timepoints	Clustering/TI/Visualisation
GSE98451	CEL-seq	714	12479	5	Mouse Uterus Timepoints Endometrium (CEL-seq)	GSE98451	Gulati et al. PMID: 31974247	Timepoints	Clustering/TI/Visualisation
GSE99333	Smart-seq2	369	23420	4	Mouse Adrenal medulla Phenotypes Peripheral glia (Smart-seq2)	GSE99333	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE94641	Plate-seq	225	33327	4	Mouse Brain Timepoints Medial ganglionic eminence (C1)	GSE94641	Gulati et al. PMID: 31974247	Timepoints	Clustering/TI/Visualisation
GSE60783	C1	248	15752	3	Mouse Blood Phenotypes Dendritic cells (C1)	GSE60783	Gulati et al. PMID: 31974247	FACS	Clustering/TI/Visualisation
GSE67602	C1	1422	25932	5	Mouse Skin Phenotypes Hair epidermis (C1)	GSE67602	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering/TI/Visualisation
GSE70245	C1	394	23955	8	Mouse Blood Phenotypes HSPCs (C1)	GSE70245	Gulati et al. PMID: 31974247	FACS	Clustering/TI/Visualisation
GSE90047	Smart-seq2	447	40829	7	Mouse Liver Timepoints Hepatoblast (Smart-seq2)	GSE90047	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE75748	C1	1018	19095	6	Human Embryo Phenotypes hESC in vitro (C1)	GSE75748	Gulati et al. PMID: 31974247	FACS	Clustering
GSE52529	C1	170	46077	3	Human Muscle Phenotypes HSMM (C1)	GSE52529	Gulati et al. PMID: 31974247	Culture conditions	Clustering/TI/Visualisation
GSE85066	C1	498	30670	9	Human Embryo Phenotypes Mesoderm (C1)	GSE85066	Gulati et al. PMID: 31974247	FACS	Clustering
GSE93421	10x	5000	16957	10	Human Blood Phenotypes Peripheral blood (10x)	GSE93421	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE36552	Tang et al.	85	20012	6	Human Embryo Phenotypes Pre-implant human embryo (Tang et al.)	GSE36552	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE86146	Smart-seq2	1844	24153	17	Human Embryo Timepoints Germ cells (Smart-seq2)	GSE86146	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE98664	RamDA-seq	456	47515	5	Mouse Embryo Timepoints mESC in vitro (RamDA-seq)	GSE98664	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE52583	C1	101	23093	4	Mouse Lung Timepoints Lung development (C1)	GSE52583	Gulati et al. PMID: 31974247	Timepoints	Clustering/TI/Visualisation
GSE97391	inDrop	2684	28205	4	Mouse Brain Phenotypes Direct in vitro neuron (inDrop)	GSE97391	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE76408	CEL-seq	480	23460	6	Mouse Intestine Phenotypes Lgr5-CreER intestine (CEL-seq)	GSE76408	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE109774	10x	3652	13526	11	Mouse Blood Phenotypes Bone marrow (10x)	GSE109774	Gulati et al. PMID: 31974247	FACS (+clustering)	Clustering
GSE109774	Smart-seq2	4897	17479	8	Mouse Blood Phenotypes Bone marrow (Smart-seq2)	GSE109774	Gulati et al. PMID: 31974247	FACS (+clustering)	Clustering
GSE92332	Smart-seq2	1522	20108	9	Mouse Intestine Phenotypes Intestine (Smart-seq2)	GSE92332	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE97391	inDrop	2996	28205	7	Mouse Brain Phenotypes Standard in vitro neuron (inDrop)	GSE97391	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE45719	Smart-seq2	286	22431	13	Mouse Embryo Phenotypes Pre-implant mouse embryo (Deng et al.)	GSE45719	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE52583	C1	66	23093	3	Mouse Lung Phenotypes AT2/AT1 lineage (C1)	GSE52583	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering/TI/Visualisation
GSE69761	C1	79	35016	5	Mouse Lung Phenotypes Lung fibroblast (C1)	GSE69761	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering/TI/Visualisation
GSE92332	Drop-seq	4581	15971	15	Mouse Intestine Phenotypes Intestine (Drop-seq)	GSE92332	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE107122	Drop-seq	5998	21201	3	Mouse Brain Timepoints Neural stem cells (Drop-seq)	GSE107122	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE64447	C1	447	24480	4	Mouse Bone Phenotypes Skeletal stem cells (C1)	GSE64447	Gulati et al. PMID: 31974247	FACS	Clustering/TI/Visualisation
GSE102066	C1	781	13762	8	Human Brain Timepoints In vitro NPCs (C1)	GSE102066	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE75330	C1	5050	23226	12	Mouse Brain Phenotypes Oligodendrocyte phenotypes (C1)	GSE75330	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE75330	C1	5050	23226	23	Mouse Brain Timepoints Oligodendrocyte timepoints (C1)	GSE75330	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE87375	Smart-seq2	338	40829	6	Mouse Pancreas Timepoints Pancreatic alpha cell (Smart-seq2)	GSE87375	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE87375	Smart-seq2	575	40829	7	Mouse Pancreas Timepoints Pancreatic beta cell (Smart-seq2)	GSE87375	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE103633	Drop-seq	21612	28065	2	Planaria Organism Phenotypes Whole planaria (Drop-seq)	GSE103633	Gulati et al. PMID: 31974247	Markers (+clustering)	Clustering
GSE107910	Drop-seq	9307	19530	8	Mouse Thymus Timepoints Thymus (Drop-seq)	GSE107910	Gulati et al. PMID: 31974247	Timepoints	Clustering
GSE106587	Drop-seq	39505	23974	12	Zebrafish Organism Timepoints Early zebrafish (Drop-seq)	GSE106587	Gulati et al. PMID: 31974247	Timepoints	Clustering
FreytagGold	10x	925	58302	3	Human lung adenocarcinoma cell lines	GSE111108	Freytag et al. PMC6124389, Sun et al. PMC6902413	FACS	Clustering
PBMC3k	10x	3205	58302	11	Human	SRP073767	Freytag et al. PMC6124389, Sun et al. PMC6902413	FACS	Clustering
PBMC4k	10x	4292	58302	11	Human	SRP073767	Freytag et al. PMC6124389, Sun et al. PMC6902413	FACS	Clustering
Baron (Mouse)	inDrop	1886	14861	13	Mouse pancreas	GSE84133	Abdelaal et al. PMC6734286	Markers (+clustering)	Clustering
Baron (Human)	inDrop	8569	17499	14	Human pancreas	GSE84133	Abdelaal et al. PMC6734286	Markers (+clustering)	Clustering
Muraro	CEL-Seq2	2122	18915	9	Human pancreas	GSE85241	Abdelaal et al. PMC6734286	FACS (+clustering)	Clustering
Segerstolpe	SMART-Seq2	2133	22757	13	Human pancreas	E-MTAB-5061	Abdelaal et al. PMC6734286	Markers (+clustering)	Clustering
Xin	SMARTer	1449	33889	4	Human pancreas	GSE81608	Abdelaal et al. PMC6734286	Markers (+clustering)	Clustering
CellBench1	10X chromium	3803	11778	5	Mixture of five human lung cancer cell lines	GSE118767	Abdelaal et al. PMC6734286	Cell lines	Clustering
CellBench2	CEL-Seq2	570	12627	5	Mixture of five human lung cancer cell lines	GSE118767	Abdelaal et al. PMC6734286	Cell lines	Clustering
TM	SMART-Seq2	54865	19791	55	Whole Mus musculus	GSE109774	Abdelaal et al. PMC6734286	FACS (+clustering)	Clustering
AMB	SMART-Seq v4	12832	42625	4/22/110	Primary mouse visual cortex	GSE115746	Abdelaal et al. PMC6734286	FACS (+clustering)	Clustering
Zheng sorted	10X CHROMIUM	20000	21952	10	FACS-sorted PBMC	SRP073767	Abdelaal et al. PMC6734286	FACS	Clustering
Zheng 68K	10X CHROMIUM	65943	20387	11	PBMC	SRP073767	Abdelaal et al. PMC6734286	Markers (+clustering)	Clustering
Baron_m2016	inDrop	1886	14861	13	Mouse pancreas	GSE84133	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Klein2015	inDrop	2712	24027	4	Embryonic stem cells	GSE65525	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Zeisel2015	STRT-Seq UMI	3005	19972	9	Mouse cortex and hippocampus	GSE60361	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Darmanis2015	C1	466	21630	9	Human brain	GSE67835	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Deng2014_raw	Smart-Seq, Smart-Seq2	268	21297	6	Mouse embryo	GSE45719	Krzak et al. PMC6918801	Timepoints	Clustering
Goolam2016	Smart-Seq2	124	28147	4	Mouse embryo	E-MTAB-3321	Krzak et al. PMC6918801	Timepoints	Clustering
Kolodziejczyk2015	SMARTer	704	32225	3	Mouse embryonic stem cells	E-MTAB-2600	Krzak et al. PMC6918801	Culture conditions	Clustering
Li2017	SMARTer	561	43055	9	Human colorectal tumors	GSE81861	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Romanov2016	C1	2881	21143	7	Mouse hypotalamus	GSE74672	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Tasic2016_raw	SMARTer	1679	21617	17	Mouse cortical cells	GSE71585	Krzak et al. PMC6918801	FACS (+clustering)	Clustering
Deng2014_rpkm	Smart-Seq, Smart-Seq2	268	22958	5	Mouse embryo	GSE45719	Krzak et al. PMC6918801	Timepoints	Clustering
Segerstolpe2016	Smart-Seq2	3514	25525	15	Human pancreatic islet cells	E-MTAB-5061	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Tasic2016_rpkm	SMARTer	1679	24057	17	Mouse cortical cells	GSE71585	Krzak et al. PMC6918801	FACS (+clustering)	Clustering
Yan2013	Tang et al.	90	20214	6	Human embryo	GSE36552	Krzak et al. PMC6918801	Timepoints	Clustering
Biase2014	SMARTer	56	25737	4	Mouse embryo	GSE57249	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
Treutlein2014	SMARTer	80	23271	5	Mouse lung epithelium	GSE52583	Krzak et al. PMC6918801	Markers (+clustering)	Clustering
ChuBatch1	SMARTer	350	19097	5	Human	GSE75748	Sun et al. PMC6902413	FACS	Clustering/TI/Visualisation
ChuBatch2	SMARTer	425	19097	6	Human	GSE75748	Sun et al. PMC6902413	FACS	Clustering/TI/Visualisation
Schlitzer	Fluidigm	238	4480	3	Mouse DCs from the BM	GSE60783	Sun et al. PMC6902413	FACS	TI/Visualisation
Petropoulos	Smart-Seq2	1289	8772	5	Human embryo	E-MTAB-3929	Sun et al. PMC6902413	Timepoints	TI/Visualisation
LiM	Smart-Seq2	649	4777	8	Male fetal gonads	GSE86146	Sun et al. PMC6902413	Timepoints	TI/Visualisation
LiF	Smart-Seq2	666	5319	12	Female fetal gonads	GSE86146	Sun et al. PMC6902413	Timepoints	TI/Visualisation
ZhangBeta	Smart-Seq2	562	6138	7	Mouse pancreatic beta cells	GSE87375	Sun et al. PMC6902413	Timepoints	TI/Visualisation
ZhangAlpha	Smart-Seq2	322	6138	6	Mouse pancreatic alpha cells	GSE87375	Sun et al. PMC6902413	Timepoints	TI/Visualisation
GuoF	Tang et. al.	100	8772	5	Human female primordial germ cells	GSE63818	Sun et al. PMC6902413	Timepoints	TI/Visualisation
GuoM	Tang et. al.	166	8772	5	Human male primordial germ cells	GSE63818	Sun et al. PMC6902413	Timepoints	TI/Visualisation
KowalczykYoung	Smart-Seq	493	2227	3	Mouse stem cells	GSE59114	Sun et al. PMC6902413	FACS	TI/Visualisation
KowalczykOld	Smart-Seq	873	2815	3	Mouse stem cells	GSE59114	Sun et al. PMC6902413	FACS	TI/Visualisation
Hayashi	RamDA-seq	414	23658	5	Mouse ES cells	GSE98664	Sun et al. PMC6902413	Timepoints	TI/Visualisation
ShalekLPS	Smart-seq	504	4158	5	Mouse DCs	GSE48968	Sun et al. PMC6902413	Timepoints	TI/Visualisation
Trapnell	SMARTer	290	8772	4	Human skeletal muscle myoblasts cells	GSE52529	Sun et al. PMC6902413	Timepoints	TI/Visualisation
Olsson	SMARTer	316	3594	3	Mouse stem cells	GSE70245	Sun et al. PMC6902413	FACS	TI/Visualisation

Supplementary Table 1: Table of all benchmark datasets' technical characteristics used in our study. Rows in gold are gold-standard, the rest is silver-standard.

	Dataset name	Cardinality	Number of features	Sparsity
1	1m	3892	33	41
2	20_newsgroups_drift	19460	989	89
3	2dplanes	25536	10	35
4	a3a	24947	121	89
5	a9a	35277	121	89
6	ada_agnostic	4512	6	31
7	ada	4146	47	79
8	AI4I2020	10000	12	50
9	Ailerons	13750	40	0
10	allbp	3656	6	0
11	allrep	3656	6	0
12	Allstate_Claims_Severity	100000	14	2
13	aloi	100000	128	76
14	amazon-commerce-reviews	1480	9878	85
15	auml_eml_1_c	3845	678	91
16	auml_eml_1_d	3734	10	1
17	auml_url_1	100000	500	0
18	auml_url_2	57379	12	40
19	auml_url_3	13529	77	29
20	auml_web_1	5785	30	34
21	autoUniv-au4-2500	2500	58	14
22	autoUniv-au7-1100	1100	12	29
23	avila	20867	12	9
24	BachChoralHarmony	5665	16	55
25	bank-marketing	4521	7	29
26	bank32nh	8192	32	3
27	bankmarketing	40864	10	14
28	Bike_Sharing_Demand	17358	8	11
29	Bike	4435	11	8
30	Bioresponse	3751	1776	84
31	blocks	5393	10	1
32	BMC_TrainingData	100000	31	67
33	BNG_2dplanes_	38905	10	35
34	BNG_adult_	100000	14	41
35	BNG_Ailerons_	100000	40	0
36	BNG_anneal_	100000	38	73
37	BNG_Australian_	100000	14	16
38	BNG_auto_price_	100000	14	0
39	BNG_autoHorse_	100000	17	2
40	BNG_autos_	100000	15	0
41	BNG_baseball_	100000	15	0
42	BNG_bridges_version1_	100000	12	19
43	BNG_cholesterol_	100000	6	18
44	BNG_cleveland_	100000	6	10
45	BNG_colic_	100000	7	0
46	BNG_cpu_act_	100000	21	0
47	BNG_cpu_small_	100000	12	0
48	BNG_credit-a_	100000	6	0
49	BNG_credit-g_	100000	7	25
50	BNG_cylinder-bands_	100000	18	6

51	BNG_dermatology_	100000	34	58
52	BNG_elevators_	100000	18	11
53	BNG_eucalyptus_	100000	14	7
54	BNG_heart-c_	100000	6	10
55	BNG_hepatitis_	100000	6	0
56	BNG_ionosphere_	100000	34	24
57	BNG_JapaneseVowels_	100000	14	0
58	BNG_kr-vs-kp_	100000	36	29
59	BNG_labor_	100000	8	2
60	BNG_letter_	100000	16	0
61	BNG_libras_move_	100000	90	0
62	BNG_lymph_	100000	18	38
63	BNG_mfeat-fourier_	100000	76	0
64	BNG_mfeat-zernike_	100000	47	0
65	BNG_mushroom_	100000	22	26
66	BNG_mv_	78732	7	0
67	BNG_page-blocks_	100000	10	0
68	BNG_pbc_	100000	10	0
69	BNG_pendigits_	100000	16	0
70	BNG_pharynx_	100000	10	36
71	BNG_primary-tumor_	73897	17	36
72	BNG_puma32H_	100000	32	0
73	BNG_pwLinear_	36430	10	35
74	BNG_satellite_image_	100000	36	0
75	BNG_segment_	41714	19	56
76	BNG_solar-flare_	27034	12	38
77	BNG_sonar_	100000	60	0
78	BNG_soybean_	100000	35	56
79	BNG_spambase_	12480	57	93
80	BNG_SPECT_	100000	22	53
81	BNG_trains_	100000	27	53
82	BNG_vehicle_	100000	18	0
83	BNG_vote_	12268	16	47
84	BNG_vowel_	100000	10	0
85	BNG_waveform-5000_	100000	40	0
86	BNG_wine_	100000	13	0
87	BNG_wine_quality_	100000	11	0
88	BNG_wisconsin_	100000	32	0
89	BNG_zoo_	100000	16	59
90	bot-iot-all-features	100000	35	40
91	Brazilian_houses	10118	8	15
92	Buzzinsocialmedia_Twitter	100000	77	10
93	BuzzinsocialmediaTomsHardware	23998	97	54
94	car-evaluation	1728	21	71
95	cardiotocography	2115	34	43
96	CD4	8946	61	72
97	Census-Income	100000	13	53
98	Chaos_detection_in_Duffing_system	100000	25	7
99	christine	5418	1595	6
100	churn	5000	16	5
101	CIFAR_10_small	20000	3072	0
102	CIFAR_10	60000	3072	0
103	clean2	6598	168	1
104	Click_prediction_small	31113	5	35
105	cnae-9	1052	686	99
106	coil2000	8261	85	57
107	cold	2435	24	0
108	compas-two-years	1605	7	51
109	connect-4	67557	42	81
110	covertype	92107	14	22

111	CovPokElec	100000	16	13
112	cpu_act	8192	21	19
113	cpu_small	8192	12	5
114	cpu	8189	6	7
115	creditcard	100000	29	0
116	CreditCardFraudDetection	100000	30	0
117	CreditCardSubset	13901	30	0
118	crimecommunitynums2	1994	126	20
119	ct–slice–localization	53437	376	49
120	dataset_sales	10730	14	15
121	dataset_time_7	19242	6	1
122	default_credit_card_p	30000	23	13
123	default–of–credit–card–clients	29944	23	11
124	Devnagari–Script	91975	785	48
125	Diabetes130US	100000	13	34
126	diau	2435	24	0
127	dilbert	10000	2000	0
128	dionis	100000	55	2
129	dis	3656	6	0
130	dna	3001	180	75
131	dtt	2435	24	0
132	Edge_Embedding	100000	30	3
133	eeg–eye–state	14980	14	0
134	electrical–grid–stability	10000	13	0
135	elevators	16599	9	0
136	EMNIST_Balanced	100000	727	64
137	enron	1561	1001	93
138	eye_movements	10936	24	18
139	fabert	8237	796	99
140	fars	74176	14	19
141	Fashion–MNIST	70000	784	50
142	fbis_wc	2455	2000	92
143	first–order–theorem–proving	5342	50	17
144	freMTPL2freq	100000	8	22
145	fried	40768	10	0
146	GAMETES_Epistasis_2–Way_1000atts_0_4H_EDM–1_EDM–1_1	1600	1000	58
147	GAMETES_Epistasis_2–Way_20atts_0_1H_EDM–1_1	1599	20	56
148	GAMETES_Epistasis_2–Way_20atts_0_4H_EDM–1_1	1597	20	59
149	GAMETES_Epistasis_3–Way_20atts_0_2H_EDM–1_1	1599	20	63
150	GAMETES_Heterogeneity_20atts_1600_Het_0_4_0_2_50_EDM–2_001	1592	20	64
151	GAMETES_Heterogeneity_20atts_1600_Het_0_4_0_2_75_EDM–2_001	1599	20	55
152	gas–drift	13910	128	0
153	gas–turbine–2011	7411	11	0
154	gas–turbine–2012	7628	11	0
155	gas–turbine–2013	7152	11	0
156	gas–turbine–2014	7151	11	0
157	gas–turbine–2015	7384	11	0
158	GeographicalOriginalofMusic	1059	73	0
159	GesturePhaseSegmentationProcessed	9873	32	0
160	gina_agnostic	3468	970	69
161	gina_prior	3468	636	76
162	gina_prior2	3468	636	76
163	gina	3153	970	69
164	gisette	7000	5000	87
165	GTSRB–HOG01	51831	1568	4
166	GTSRB–HOG03	51831	2916	9
167	GTSRB–HueHist	51831	256	49
168	guillermo	19884	4282	46
169	har	10299	540	7
170	heat	2435	24	0

171	helena	65192	27	0
172	Higgs	100000	28	8
173	hill-valley	974	100	3
174	hiva_agnostic	4206	1429	91
175	hls4ml_lhc_jets_hlf	100000	15	0
176	house_16H	22783	16	14
177	house_sales_reduced	21518	19	21
178	house_sales	21604	19	21
179	Hyperplane_10_1E-3	100000	10	0
180	ijcnn	100000	22	41
181	image	2000	135	0
182	IMDB_drama	100000	1001	98
183	Indian_pines	9144	220	0
184	internet-firewall	45742	11	57
185	Ishwar	2205	17	4
186	isolet	7797	617	8
187	jannis	83733	54	0
188	JapaneseVowels	9961	14	1
189	jasmine	2984	8	3
190	jungle_chess_2pcs_endgame_elephant_elephant	2351	13	23
191	jungle_chess_2pcs_endgame_lion_elephant	4704	20	17
192	jungle_chess_2pcs_endgame_lion_lion	2352	18	19
193	jungle_chess_2pcs_endgame_panther_elephant	4704	14	21
194	jungle_chess_2pcs_endgame_panther_lion	4704	20	17
195	jungle_chess_2pcs_endgame_rat_rat	3660	13	22
196	Kaggle_bike_sharing_demand_challange	10886	9	16
197	kc1	1192	21	20
198	KDDCup99	100000	34	69
199	KEGGMetabolicReactionNetwork	19197	26	22
200	Klaverjas2018	100000	32	25
201	kr-vs-kp	3196	36	24
202	Kuzushiji-49	100000	784	69
203	la1s_wc	3204	13172	99
204	la2s_wc	3071	12411	99
205	langLog	1256	919	84
206	LED_50000_	100000	24	46
207	led24	3199	24	45
208	letter	18668	16	2
209	Long	4477	19	1
210	lotto	1153	11	20
211	LottoMaster-144	11739	33	6
212	madeline	3140	259	0
213	madelon	2600	500	0
214	MagicTelescope	18905	10	0
215	malicious_urls	13529	77	29
216	mc1	2006	37	29
217	medical_charges	100000	5	1
218	Mercedes_Benz_Greener_Manufacturing	2652	312	82
219	meta_stream_intervals_arff	45164	71	16
220	Methane	100000	30	11
221	mfeat-factors	1994	213	1
222	mfeat-fourier	1994	76	0
223	mfeat-karhunen	1994	64	0
224	mfeat-pixel	1994	240	39
225	mfeat-zernike	1896	47	0
226	microaggregation2	20000	20	10
227	mlr_knn_rng	100000	7	9
228	mlr_rpart_rng	91287	7	4
229	mnist_784	70000	716	79
230	mnist_rotation	62000	784	0

231	mofn-3-7-10	1024	10	50
232	mtp	4449	198	14
233	musk	6581	166	1
234	mv	40768	7	0
235	mydata	3892	33	41
236	NASA_PHM2008_1	45918	20	9
237	NASA_PHM2008	45918	18	17
238	NELL	7767	757	99
239	nfl_games	16274	8	26
240	Node_Embedding	100000	30	3
241	nomao	32050	83	2
242	numerai28_6	96320	21	0
243	obesity-level-indicators	2086	8	8
244	oh0_wc	1003	3177	98
245	oh10_wc	1050	3236	98
246	one-hundred-plants-margin	1600	64	24
247	one-hundred-plants-shape	1600	64	0
248	one-hundred-plants-texture	1599	64	37
249	online-shoppers-intention	11994	14	40
250	OnlineNewsPopularity	39644	59	28
251	optdigits	5620	63	48
252	OVA_Uterus	1545	10935	0
253	ozone_level	2528	72	5
254	ozone-level-8hr	2526	72	1
255	page-blocks	5393	10	2
256	parity5_plus_5	1024	10	50
257	parkinsons-telemonitoring	5875	22	3
258	ParkinsonSpeechDatasetwithMultipleTypesofSoundRecordings	1039	29	7
259	pc1	948	20	16
260	pc2	1404	35	22
261	pc3	1436	37	18
262	pc4	1343	36	23
263	PCam	969	27648	0
264	pendigits	10992	16	13
265	philippine	5832	257	1
266	phish_url	13529	77	29
267	PhishingWebsites	5785	30	54
268	PieChart3	1077	37	15
269	PizzaCutter3	1043	37	15
270	poker-hand	100000	10	16
271	poker	100000	10	16
272	pokerhand	6141	5	8
273	pol	14958	27	73
274	post-operative	45742	11	57
275	Premier_League_matches	2961	16	4
276	premier_league_with_tda	2565	20	1
277	puma32H	8192	32	0
278	qsar-biodeg	1052	41	44
279	QSAR-TID-10003	1080	1018	94
280	QSAR-TID-100044	1269	1018	93
281	QSAR-TID-100080	1124	1024	94
282	QSAR-TID-100417	1191	1024	94
283	QSAR-TID-10051	977	1024	94
284	QSAR-TID-10131	1887	1024	95
285	QSAR-TID-10140	2517	1024	94
286	QSAR-TID-10142	2535	1024	93
287	QSAR-TID-101464	1012	1024	94
288	QSAR-TID-10188	3592	1024	94
289	QSAR-TID-10197	2883	1024	94
290	QSAR-TID-10209	1067	1017	94

291	QSAR-TID-10266	1864	1024	94
292	QSAR-TID-10280	2771	1024	95
293	QSAR-TID-103	1012	1022	94
294	QSAR-TID-10315	882	1004	93
295	QSAR-TID-10378	1209	1022	94
296	QSAR-TID-10434	3405	1024	94
297	QSAR-TID-10475	964	1024	93
298	QSAR-TID-10495	1682	1024	94
299	QSAR-TID-10498	1611	1024	94
300	QSAR-TID-105	1107	1022	94
301	QSAR-TID-10502	1302	1017	94
302	QSAR-TID-10526	1469	1023	94
303	QSAR-TID-10529	2003	1023	93
304	QSAR-TID-10548	1018	1011	94
305	QSAR-TID-10576	3598	1024	94
306	QSAR-TID-10580	1707	1023	93
307	QSAR-TID-106	1118	1019	94
308	QSAR-TID-10627	2161	1024	94
309	QSAR-TID-10695	1234	1024	94
310	QSAR-TID-10696	1021	1024	94
311	QSAR-TID-107	2459	1024	95
312	QSAR-TID-10781	1937	1024	94
313	QSAR-TID-108	2527	1024	95
314	QSAR-TID-10811	1415	1024	94
315	QSAR-TID-10839	1804	1024	94
316	QSAR-TID-10849	1501	1024	94
317	QSAR-TID-10906	942	1024	94
318	QSAR-TID-10907	1388	1024	94
319	QSAR-TID-10918	1198	1024	94
320	QSAR-TID-10938	1768	1024	94
321	QSAR-TID-10979	1608	1024	94
322	QSAR-TID-10980	5454	1024	94
323	QSAR-TID-11	5259	1024	94
324	QSAR-TID-11017	1171	1024	94
325	QSAR-TID-11024	2137	1024	94
326	QSAR-TID-11082	1618	1024	94
327	QSAR-TID-11109	1814	1023	94
328	QSAR-TID-11140	2998	1024	95
329	QSAR-TID-11149	1422	1024	94
330	QSAR-TID-11225	2303	1024	93
331	QSAR-TID-11242	1067	1024	94
332	QSAR-TID-11280	1380	1023	95
333	QSAR-TID-11290	1029	1011	95
334	QSAR-TID-11300	1400	1018	95
335	QSAR-TID-11336	1106	1020	94
336	QSAR-TID-114	3146	1024	94
337	QSAR-TID-11407	1431	1024	94
338	QSAR-TID-11408	1151	1024	94
339	QSAR-TID-11451	2328	1024	94
340	QSAR-TID-11473	1451	1020	94
341	QSAR-TID-11522	1322	1015	93
342	QSAR-TID-11534	1582	1023	94
343	QSAR-TID-11536	1197	1019	93
344	QSAR-TID-11575	1314	1015	93
345	QSAR-TID-11631	1110	1021	94
346	QSAR-TID-11635	797	1023	94
347	QSAR-TID-11638	1128	1024	94
348	QSAR-TID-11678	2409	1024	94
349	QSAR-TID-11720	994	1024	94
350	QSAR-TID-11727	1099	1001	95

351	QSAR-TID-11736	1463	1024	94
352	QSAR-TID-11755	1061	1024	94
353	QSAR-TID-118	1204	1017	92
354	QSAR-TID-11902	1150	1024	94
355	QSAR-TID-11942	953	1017	94
356	QSAR-TID-11969	1196	1024	94
357	QSAR-TID-12071	1730	1024	94
358	QSAR-TID-12090	1265	1024	94
359	QSAR-TID-12227	1324	1016	94
360	QSAR-TID-12247	1298	1024	94
361	QSAR-TID-12252	2707	1024	93
362	QSAR-TID-12261	1761	1024	94
363	QSAR-TID-12268	1165	1018	95
364	QSAR-TID-124	1615	1022	94
365	QSAR-TID-12476	926	999	95
366	QSAR-TID-125	1220	1023	94
367	QSAR-TID-12512	2520	1024	94
368	QSAR-TID-12592	2363	1024	94
369	QSAR-TID-12666	2080	1024	94
370	QSAR-TID-12687	1459	1022	95
371	QSAR-TID-127	1360	1023	94
372	QSAR-TID-12724	1216	1024	94
373	QSAR-TID-128	1186	1022	94
374	QSAR-TID-12824	963	1003	93
375	QSAR-TID-12840	1544	1024	94
376	QSAR-TID-129	3538	1024	94
377	QSAR-TID-12947	1249	1024	94
378	QSAR-TID-12967	2623	1024	94
379	QSAR-TID-130	2764	1024	94
380	QSAR-TID-13000	3156	1024	94
381	QSAR-TID-133	2738	1024	94
382	QSAR-TID-136	3486	1024	94
383	QSAR-TID-137	3105	1024	94
384	QSAR-TID-138	1253	1020	94
385	QSAR-TID-14037	3738	1024	95
386	QSAR-TID-163	2242	1023	94
387	QSAR-TID-17080	1077	1013	93
388	QSAR-TID-17084	1459	1018	94
389	QSAR-TID-17085	1150	1024	95
390	QSAR-TID-174	1814	1024	95
391	QSAR-TID-188	2012	1024	93
392	QSAR-TID-191	3750	1024	94
393	QSAR-TID-194	4741	1024	94
394	QSAR-TID-197	1169	1024	94
395	QSAR-TID-19905	2762	1024	93
396	QSAR-TID-20014	2483	1024	94
397	QSAR-TID-20151	1274	1023	95
398	QSAR-TID-20154	912	1015	95
399	QSAR-TID-20174	1090	1016	95
400	QSAR-TID-219	1367	1023	95
401	QSAR-TID-226	1310	1020	94
402	QSAR-TID-227	1096	1023	95
403	QSAR-TID-234	2020	1024	93
404	QSAR-TID-235	1428	1023	95
405	QSAR-TID-238	1289	1024	94
406	QSAR-TID-246	978	1018	94
407	QSAR-TID-247	1093	1013	94
408	QSAR-TID-249	1347	1020	94
409	QSAR-TID-25	1695	1024	94
410	QSAR-TID-250	2124	1022	94

411	QSAR-TID-252	3702	1024	94
412	QSAR-TID-259	3841	1024	94
413	QSAR-TID-278	2033	1023	94
414	QSAR-TID-280	3057	1024	94
415	QSAR-TID-30017	1025	1024	94
416	QSAR-TID-30036	1086	1024	94
417	QSAR-TID-30043	1038	1024	94
418	QSAR-TID-35	1524	1024	94
419	QSAR-TID-36	1504	1024	94
420	QSAR-TID-42	1203	1023	94
421	QSAR-TID-43	1426	1024	94
422	QSAR-TID-47	1673	1024	95
423	QSAR-TID-50	1278	1023	94
424	QSAR-TID-51	2917	1024	94
425	QSAR-TID-56	1411	1024	95
426	QSAR-TID-61	1771	1024	95
427	QSAR-TID-65	1267	1024	96
428	QSAR-TID-72	4635	1024	94
429	QSAR-TID-8	1632	1024	94
430	QSAR-TID-87	3681	1024	94
431	QSAR-TID-9	4544	1024	94
432	QSAR-TID-90	1782	1024	95
433	RandomRBF_0_0	100000	10	0
434	re0_wc	1400	2868	98
435	re1_wc	1618	3726	99
436	RelevantImagesDatasetTEST	100000	27	27
437	reuters	1923	243	88
438	riccardo	18201	4284	48
439	ringnorm	7400	20	0
440	robert	10000	7200	1
441	Santander_transaction_value	4458	4692	97
442	SantanderCustomerSatisfaction	100000	200	0
443	satellite_image	6435	36	0
444	Satellite	5100	36	0
445	satimage	6430	36	0
446	sbox_sar	10000	32	3
447	scene	2398	294	1
448	scm1d	9803	280	0
449	scm20d	8966	61	0
450	segment	2086	19	17
451	semeion	1593	256	67
452	SensIT-Vehicle-Combined	98500	100	0
453	SensorDataResource	127591	25	11
454	sgemm-gpu-kernel	100000	18	31
455	shill-bidding	6321	11	25
456	Sick_numeric	3711	29	55
457	slashdot	3740	1076	99
458	spambase	4194	57	77
459	Speech	3686	400	0
460	splice	2990	60	23
461	spo	2435	24	0
462	spoken-arabic-digit	100000	14	4
463	steel-plates-fault	1941	33	22
464	SVHN_small	9927	3072	0
465	svmguide3	1243	22	21
466	sylva_agnostic	14395	40	1
467	sylva_prior	14395	108	78
468	sylvine	5124	20	0
469	test_hate_vects3	1557	3454	100
470	texture	5473	37	0

471	thyroid–allbp	2716	6	1
472	thyroid–allhyper	2716	6	1
473	thyroid–allhypo	2716	6	1
474	thyroid–allrep	2716	6	1
475	thyroid–ann	3709	21	68
476	thyroid–dis	2716	6	1
477	topo_2_1	8860	261	19
478	treasury	1049	15	0
479	TurkiyeStudentEvaluation	3977	33	16
480	twonorm	7400	20	0
481	USPS	1424	256	13
482	vehicle_sensIT	98500	100	0
483	volkert	58298	142	21
484	w1a	34704	267	95
485	wall–robot–navigation	5456	24	0
486	wap_wc	1560	8338	98
487	Waterstress	1187	20	0
488	waveform–5000	5000	40	0
489	WaveformDatabaseGenerator	5000	22	2
490	webdata_wXa	35277	121	89
491	wind	6574	14	1
492	wine_quality	5318	11	0
493	wine–quality–red	1359	11	1
494	wine–quality–white	3961	11	0
495	Wine	1359	12	1
496	WorkersCompensation	74645	7	29
497	wq	1060	16	5
498	yeast_ml8	2417	103	0
499	yeast	1453	8	25
500	Yolanda	100000	100	0
501	yprop_4_1	8774	212	86

Supplementary Table 2: Table of the datasets collected from the openML repository.

Chapter 6

Supervised analysis of single-cell RNAseq data to functionally classify T cells in cancer

Selected content of this chapter is a part of a publication in preparation. This project has been done in close collaboration with the team of Antonio Rausell in Imagine.

6.1 Introduction and statement

The current unsupervised strategy to annotate T cells in scRNAseq is mainly based on the recognition of markers, either via an enrichment analysis or differentially expressed genes. Incidentally, it is usually cumbersome to identify classical helper populations in scRNAseq datasets, such as T_H1, T_H2, etc. The gene sets or knowledge used to annotate cells rely on phenotypic markers that were used to isolate cells with classical techniques such as FACS. Supervised approach, where one takes advantage of previously annotated scRNAseq datasets, depends also implicitly on unsupervised approaches that have been used in the first place to annotate reference datasets and/or atlases. We identified 3 technical caveats with this annotation methodology: it is hardly reproducible, subjective, and time-consuming. In particular, the poor reproducibility stems from the heterogeneity of signatures or gene sets used for the annotation. This heterogeneity is easily observed: the intersection of 3 bulk transcriptomic signatures for Treg of respective gene lengths 294 [HDM+20], 136 [BKG+16] and 31 [PWS+16] contains only 10 genes. The second caveat refers to the fact that different teams would rely on different knowledge to annotate the data, in the sense that the differential expression relies on the choice of user-dependant parameters, and that the manual selection of marker genes out of the list of differentially expressed genes is biased. Lastly, the annotation is time-consuming, because of the manual review of the literature step.

We believe there is an additional stone in the shoe of the way of biologically analysing the data: the current annotation of the data is a chimera of phenotype and functions. For example, some articles decorrelate functions and phenotypes: they would use DGE analysis and GSEA to manually functionally annotate clusters, while annotating phenotypes separately [MKCK+21, LXW+21, JASC+18]. It embodies the present confusion about the lineage paradigm and how to interpret it with the new body of data brought by scRNAseq experiments. The blurring of lines between the different lineages and the functions they

would exert is exemplified in [ZZK+18], where regulatory cells, sorted upon the basis of their regulatory phenotypes, are clustered in different groups. If we make the assumption that the transcriptome profile mirrors the protein profile and thus is an accurate proxy for the cell function, it means that different clusters represent different functions: hence the regulatory population, sorted as an homogeneous group, comprises different functions. There is a dichotomy between the classical lineages, such as the effector vs. the regulatory lineages, and their functions.

We want to attempt to resolve this dichotomy by using a supervised approach. Using carefully-designed functional modules, we believe we could better understand the functionality of T cells in light of scRNAseq data.

Firstly, we need to design those modules, and conceive a scoring method. Secondly, we should quantify the overlap between our supervised approach and an unsupervised pipeline, to verify whether it is a redundant or novel approach. Lastly, we should be able to map T cell functions, and we used cancer datasets, as we believe that it should be an interesting archetype to test the method. Indeed, our intuition is that the highly dynamic TME should increase the diversity of functions [ACP+18]. We hope to better recapitulate the complexity of T cells with this new classification by easing the description of functional shifts, as well as to increase interpretability of scRNAseq T cell data.

6.2 Functional modules

6.2.1 Construction

We outlined 15 functions, that formed 15 functional modules.

- → 4 functions dedicated to communication and support: B cell, Monocyte-Macrophage (MM), dendritic cell (DC) and T cell help, termed BC-, MM-, DC- and TC-help,
- → 4 functions towards attracting other immune cells: B cell, Monocyte-Macrophage (MM), DC and T cell attraction, termed BC-, MM-, DC- and TC-attraction,
- → 1 function for the production of anti-microbial peptides: AMP,
- → 1 function for apoptosis,
- → 1 function for antiviral capacities,
- → 1 function for T cell trafficking: homing,
- → 1 function for cytotoxicity,
- → 1 function for immune suppression,
- → 1 function for proliferation.

I attributed to each module the corresponding relevant genes: each gene that belongs to a given module is an effector for the function. I parsed the literature to fill up the modules. I selected mostly experimental articles, with a rigorous proof of the contribution of a given gene to the function of interest, or few reviews from experts in the field.

I collected 232 effector genes, with 172 unique genes, from 191 articles. I verified that they were all tabulated with their EntrezID symbol, using an online symbol checker tool. Each gene has been confirmed on the basis of a mean of 2 articles (Table 6.1).

Gene	Function	References
CXCL13	BC-attraction	12093871, 29880013, 16516453, 31002794
CXCL9	BC-attraction	18561120
CXCL10	BC-attraction	18561120
CXCL11	BC-attraction	18561120
CXCL12	BC-attraction	12093871, 19804625, 18561120, 15749730
CCL19	BC-attraction	12093871, 19804625, 18561120, 16516453
CCL21	BC-attraction	12093871, 19804625, 18561120, 15749730, 16516453
CCL20	BC-attraction	29880013, 31166050, 29375554, 19804625, 18561120, 15749730
CCL19	DC-attraction	24725321, 29563613, 25753266, 18379575, 15001175, 11489962
CCL21	DC-attraction	24725321, 29563613, 25753266, 18379575, 15001175, 11489962
CCL3	DC-attraction	29563613, 11489962
CCL5	DC-attraction	29563613, 11489962
CCL2	DC-attraction	29563613, 11489962
CCL20	DC-attraction	29563613, 25130722
CCL27	DC-attraction	29563613
CXCL12	DC-attraction	29563613
RARRES2	DC-attraction	29563613
C5	DC-attraction	29563613
C1QA	DC-attraction	29563613
PLG	DC-attraction	29563613
IL18	DC-attraction	29563613
CCL4	DC-attraction	29563613
CCL8	DC-attraction	29563613
CCL7	DC-attraction	29563613
CCL11	DC-attraction	12218106, 9558100, 9269754, 9561368
CCL23	DC-attraction	9269754, 15978562, 10536111, 20956349, 10360972
CCL24	DC-attraction	12218106, 9558100, 9269754, 9561368
CCL28	DC-attraction	12218106, 9558100, 9269754, 9561368, 21937703
CCL2	MM-attraction	19215821, 26635790
PLEKHO1	MM-attraction	23747421
CCL3	MM-attraction	28499492
CCL4	MM-attraction	28499492
CCL1	MM-attraction	17947648
CXCL12	TC-attraction	20364260, 29894310, 15634883
TNF	TC-attraction	23494522, 22138716, 24636534
CCL5	TC-attraction	17177831, 12960247, 11261794, 22138716
CXCL9	TC-attraction	12960247
CXCL10	TC-attraction	12960247
CXCL11	TC-attraction	12960247

Gene	Function	References
CD40LG	BC-help	26276638, 31002794
CD86	BC-help	31002794
CD84	BC-help	26276638, 31002794
IL21	BC-help	26276638, 31002794
AICDA	BC-help	26276638
IL2	BC-help	15450978
IL4	BC-help	26276638, 31002794, 15450978
IL10	BC-help	26276638
SEMA4C	BC-help	30150988
EPHB4	BC-help	30150988
EPHB6	BC-help	30150988
IFNG	BC-help	30150988, 31002794
ICOS	BC-help	30150988, 31002794
TNFSF13	BC-help	16187941
CCL20	BC-help	29375554
SH2D1A	BC-help	26276638, 31002794
PRDM1	BC-help	26276638
TNFSF13B	BC-help	16187941, 31002794
ITGB2	BC-help	29669250, 28939548
IL1	DC-help	18981105
IL2	DC-help	15450978
IL4	DC-help	15450978, 18981105
IL5	DC-help	15450978
IL10	DC-help	15450978
IFNG	DC-help	15450978, 18981105
CD40LG	DC-help	18981105, 26781939, 22539281
LAT	DC-help	15450978
TNF	DC-help	18981105
CSF2	DC-help	18981105
FYB1	DC-help	23918975
ITGB2	DC-help	27501450, 15450978, 26781939
XCL1	DC-help	29563613, 22100876
CTLA4	DC-help	26781939
TNFRSF4	DC-help	10637280, 9378971
TNF	MM-help	NBK27101, 26635790
IFNG	MM-help	NBK27101, 26635790
IL4	MM-help	26635790
IL13	MM-help	26635790
IL17A	MM-help	26635790
IL10	MM-help	26635790, 15784460
IL15	MM-help	24942581
CSF2	MM-help	24942581
IFNG	TC-help	26781939
IL2	TC-help	26781939, 20856822, 22343569, 22539281
LTA	TC-help	11907234, 24698108, 11145686
IL10	TC-help	30423297, 26781939
TGFB1	TC-help	26781939
CD40LG	TC-help	26781939
IL21	TC-help	26781939, 22539281
TNFSF14	TC-help	11994431
IL7	TC-help	22539281
IL15	TC-help	22539281

Gene	Function	References
IL26	AMP	27337042
GNLY	AMP	10462738, 11023496
GZMB	AMP	20536558
IFNG	AMP	28618267
CCL1	AMP	22837760
CCL2	AMP	22837760
CCL11	AMP	22837760
CCL13	AMP	22837760
CCL14	AMP	22837760
CCL15	AMP	22837760
CCL17	AMP	22837760
CCL18	AMP	22837760
CCL19	AMP	22837760
CCL20	AMP	12149255, 27631019, 22837760
CCL21	AMP	22837760
CCL22	AMP	22837760
CCL25	AMP	22837760
CCL28	AMP	26062132, 22837760
XCL1	AMP	22837760
CXCL1	AMP	22837760
CXCL2	AMP	22837760
CXCL3	AMP	22837760
CXCL6	AMP	18443119, 26062132, 22837760
CXCL7	AMP	26062132, 22837760, 21101183
CXCL9	AMP	26062132, 22837760
CXCL10	AMP	26062132, 22837760
CXCL11	AMP	26062132, 22837760
CXCL12	AMP	22837760
CXCL13	AMP	22837760
CXCL14	AMP	22837760
CAMP	AMP	11358975, 24287494, 22429567, 22837760
IFITM3	Antiviral	30914370, 30662816, 22046135, 24699674, 23358889, 22328912
IFITM1	Antiviral	23358889, 22328912
IFITM2	Antiviral	23358889, 22328912
EIF2AK2	Antiviral	1382142, 22328912
OAS1	Antiviral	20038200, 22328912
MX1	Antiviral	22328912
APOBEC3G	Antiviral	22328912
TRIM5	Antiviral	22328912
ISG15	Antiviral	22328912
ADAR	Antiviral	22328912
BST2	Antiviral	22328912
RSAD2	Antiviral	22328912, 30684519
DDX58	Antiviral	22328912, 21478870
DDX60	Antiviral	22328912
GBP1	Antiviral	22328912
GBP2	Antiviral	22328912
HPSE	Antiviral	22328912, 21478870
IFIH1	Antiviral	22328912, 21478870
IFIT1	Antiviral	22328912
IFIT2	Antiviral	22328912
IFIT3	Antiviral	22328912
IFIT5	Antiviral	22328912
ISG20	Antiviral	22328912, 31600344
MX2	Antiviral	22328912
PML	Antiviral	22328912

Gene	Function	References
BAX	Apoptosis	8918887, 11163212, 10395708, 10814794, 11281652, 31324752
BAK1	Apoptosis	11163212, 31324752, 30536008, 30334018
BCL2	Apoptosis	15110520
BID	Apoptosis	9727492, 10982793, 8918887, 9873064, 23834359
CASP3	Apoptosis	9422506, 9422513, 10814794, 11281652, 11279545, 23834359
CASP6	Apoptosis	24727569,10438520, 11279545, 15321985
CASP7	Apoptosis	31687791, 28367243, 23834359, 14583630
CASP8	Apoptosis	9464839, 9727492, 10982793, 11281652
CASP9	Apoptosis	11281652, 23834359
TNFRSF1A	Apoptosis	11314015, 11752172, 11861282
TNFRSF10A	Apoptosis	9082980, 9430228, 9430227
TNFRSF10B	Apoptosis	9311998, 9430228, 9430227, 8994832, 9285725
TNFRSF21	Apoptosis	9714541
FAS	Apoptosis	2787530, 2469768, 1713127, 10814794, 11752172
FADD	Apoptosis	7538907, 7536190
CAD	Apoptosis	9422506, 9422513
DFFA-DFFB	Apoptosis	9108473
PTK2B	Apoptosis	7519637
RIPK1	Apoptosis	7538908
TRADD	Apoptosis	7758105, 9714541
PRF1	Cytotoxicity	16405653, 12093006, 28499492, 20536553, NBK27101, 10462738, 3077301, 19337143, 24394640, 28166682
GZMB	Cytotoxicity	16405653, 28499492, 12093006, 20536553, NBK27101, 10462738, 19337143, 20536558, 24394640, 28166682
GZMA	Cytotoxicity	16405653, 12093006, 28499492, 20536553, NBK27101, 10462738, 19337143, 28166682
GZMH	Cytotoxicity	12093006, 28499492
GZMK	Cytotoxicity	12093006, 28499492
TNF	Cytotoxicity	12093006, 28499492, NBK27101, 3077301
IFNG	Cytotoxicity	12093006, 28499492, NBK27101, 3077301
FASLG	Cytotoxicity	12093006, 28499492, 20536553, NBK27101, 10462738, 19337143, 24394640, 10415024, 26635790
TNFSF10	Cytotoxicity	27265595, 10415024
ITGB2	Cytotoxicity	28499492
GNLY	Cytotoxicity	10462738, 19337143, 24394640
LTA	Cytotoxicity	3077301

Gene	Function	References
ITGB1	Homing	1380034, 12234367, 11261794, 22275188, 14708592
ITGB2	Homing	17624950, 11261794, 22275188, 22138716, 12234367
CXCR3	Homing	14632748, 22138716, 17291292
CCR1	Homing	14632748, 22138716
CCR2	Homing	22138716
CCR5	Homing	22138716, 17291292
CXCR4	Homing	22275188, 17291292
ITGA4	Homing	12234367, 17291292, 11261794, 22275188, 19808049
SELPLG	Homing	19808049, 22138716, 14708592, 17181631
CCR8	Homing	17181631
CCR4	Homing	19808049, 16516453
CCR10	Homing	14708592
SELL	Homing	20146713, 1705015, 22138716
CCR7	Homing	20146713, 18379575, 15001175, 25753266
S1PR1	Homing	20146713
KLF2	Homing	20146713
CXCR5	Homing	28499492, 12851649
ITGB7	Homing	17291292, 19808049
CCR9	Homing	17291292, 19808049, 14708592
CCR6	Homing	21376174
ITGAE	Homing	17291292
TNFRSF1A	Homing	24636534
TNFRSF1B	Homing	24636534
ENTPD1	Immune suppression	27313580, 27590281, 27851913, 19737784
NT5E	Immune suppression	18566595, 27313580, 19737784
CTLA4	Immune suppression	21945485, 17632406, 29141660, 19737784
LGALS9	Immune suppression	25065622, 18566595, 19737784
TIGIT	Immune suppression	25994968, 27851914, 24745333
ITGB2	Immune suppression	19737784
MT-CO2	Immune suppression	19900843
EBI3	Immune suppression	27590281, 27851913, 19737784
IL27	Immune suppression	19737784
IL12A	Immune suppression	27590281, 27851913, 19737784
LGALS1	Immune suppression	23219401, 19737784
TNFRSF18	Immune suppression	16557261, 30484986
TGFB1	Immune suppression	23219401, 18566595, 11535631, 19737784, 16557261
VEGFA	Immune suppression	16557261
IL10	Immune suppression	23219401, 18566595, 19737784, 16557261
LRRC32	Immune suppression	25904740, 30674536, 31357555
LAG3	Immune suppression	26013006
PDCD1	Immune suppression	17304234, 17606980, 19426216
ICOS	Immune suppression	24312642
AREG	Immune suppression	23333074, 27432879
ITGB8	Immune suppression	25127859, 10.1101/2020.05.14.084913
IL2RA	Proliferation	17383196, 10952731
IL2RB	Proliferation	17383196, 10952731
IL2	Proliferation	21889323
IL7R	Proliferation	29616038, 19637200
KLF2	Proliferation	24874925
LAT	Proliferation	29616038
MYC	Proliferation	24731854
TNFSF14	Proliferation	19702559
MAPK1	Proliferation	12801802

Table 6.1: Table of the functional modules including references.

6.2.2 Description

The mean length of a functional module is 15 genes, with the shortest module being MM-attraction with 5 genes and the longest being AMP with 31 genes (Figure 6.1). We looked at each function to evaluate its specificity with regard to the other modules, to answer the question whether the effector genes of a given module are only members of this module, or if they belong to other functions.

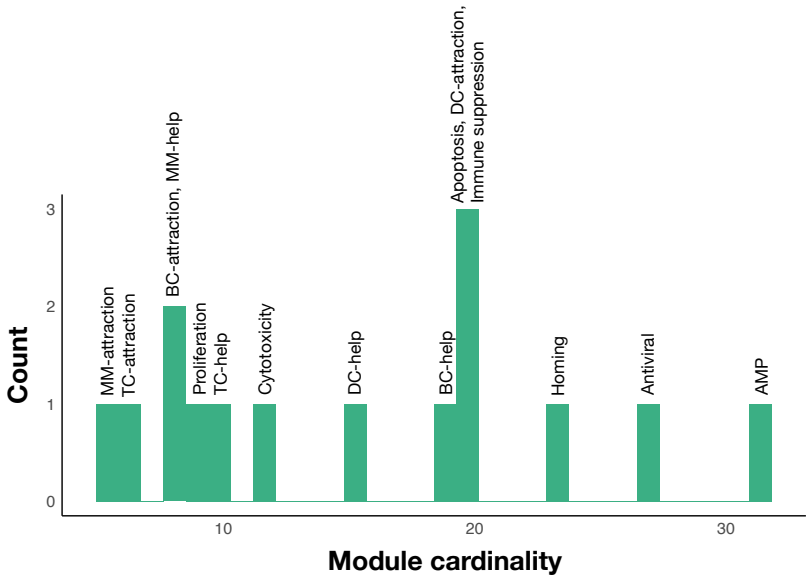


Figure 6.1: Histogram for the cardinality of functional modules.

Starting with the genes, the vast majority of them is unmatched: 136 out of 172, i.e. almost 80%, belong to a unique module (Figure 6.2a). The remaining 36 genes appear in a mean of almost 3 modules, and the most ubiquitous gene takes part in 6 functions (Figure 6.2b).

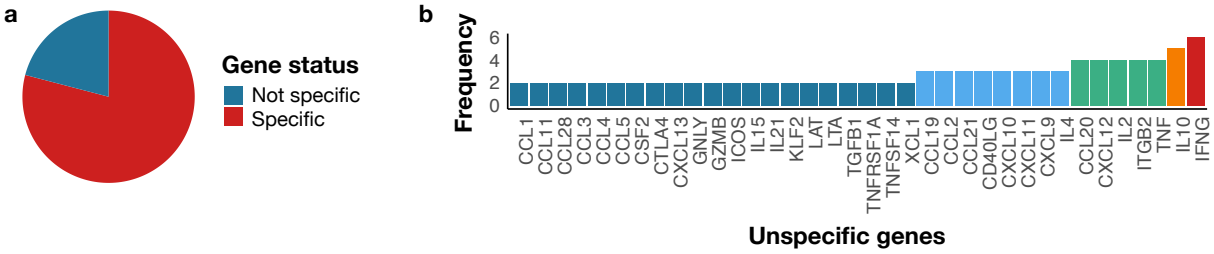


Figure 6.2: Gene-level specificity. (**a**) Specificity status of the 172 unique effector genes. (**b**) Number of occurrences for each unspecific gene.

At the function level, I designed a specificity score to quantify the singularity of each module. For a function $\mathcal{F}$, it is calculated as:

$$Spec_{\mathcal{F}} = \frac{1}{|\mathcal{F}|} \sum_{g \in \mathcal{F}} \frac{1}{\Omega_g},$$

where Ω_g is the number of occurrences of gene g in all functions. A score of 100% means that all genes are found only in $\mathcal{F}$, while a score of 50% means that all the genes are found

in another function, or two thirds of them belong to three additional functions, etc. Most of the functions are reasonably specific, with a score above 60% (Figure 6.3a). In fact, the less specific functions are the help and the attraction functions that share many cytokines and chemokines with each other. Moreover, the specificity correlates with the cardinality: a populated function is more specific than a scarce one, although the correlation is weak ($R^2 = 0.46$), p-val= 0.003) (Figure 6.3b).

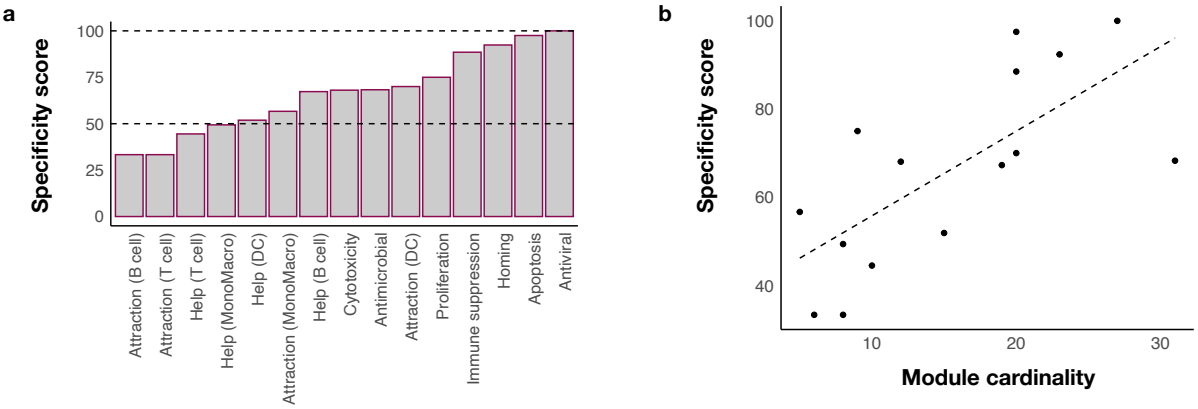


Figure 6.3: Function-level specificity. (**a**) Specificity score of the functional modules. (**b**) Module cardinality and specificity are weakly correlated ($R^2 = 0.46$, p-val= 0.003).

If we look more carefully at the overlap between functions, we can better understand the intertwining between functions, especially within the 'help' group and the 'attraction' group. I computed the intersection size between two functions $\mathcal{F}1$ and $\mathcal{F}2$ and normalized it by $|\mathcal{F}1|$. On the heatmap, $\mathcal{F}1$ is the horizontal function and $\mathcal{F}2$ the vertical function (Figure 6.4). Thus, we observe that 100% of the BC-attraction genes belong to the AMP module as well, while 25% of the AMP genes are in the BC-attraction function. Furthermore, we observe that there is a proximity between all 'attraction' modules and with the AMP module and between all 'help' modules.

We chose nonetheless to analyse the functions within the 'help' meta-module and the 'attraction' + AMP meta-module individually, instead of grouping them, in order to retain as much as granularity as possible, even if we expect similar scoring for those two meta-modules.

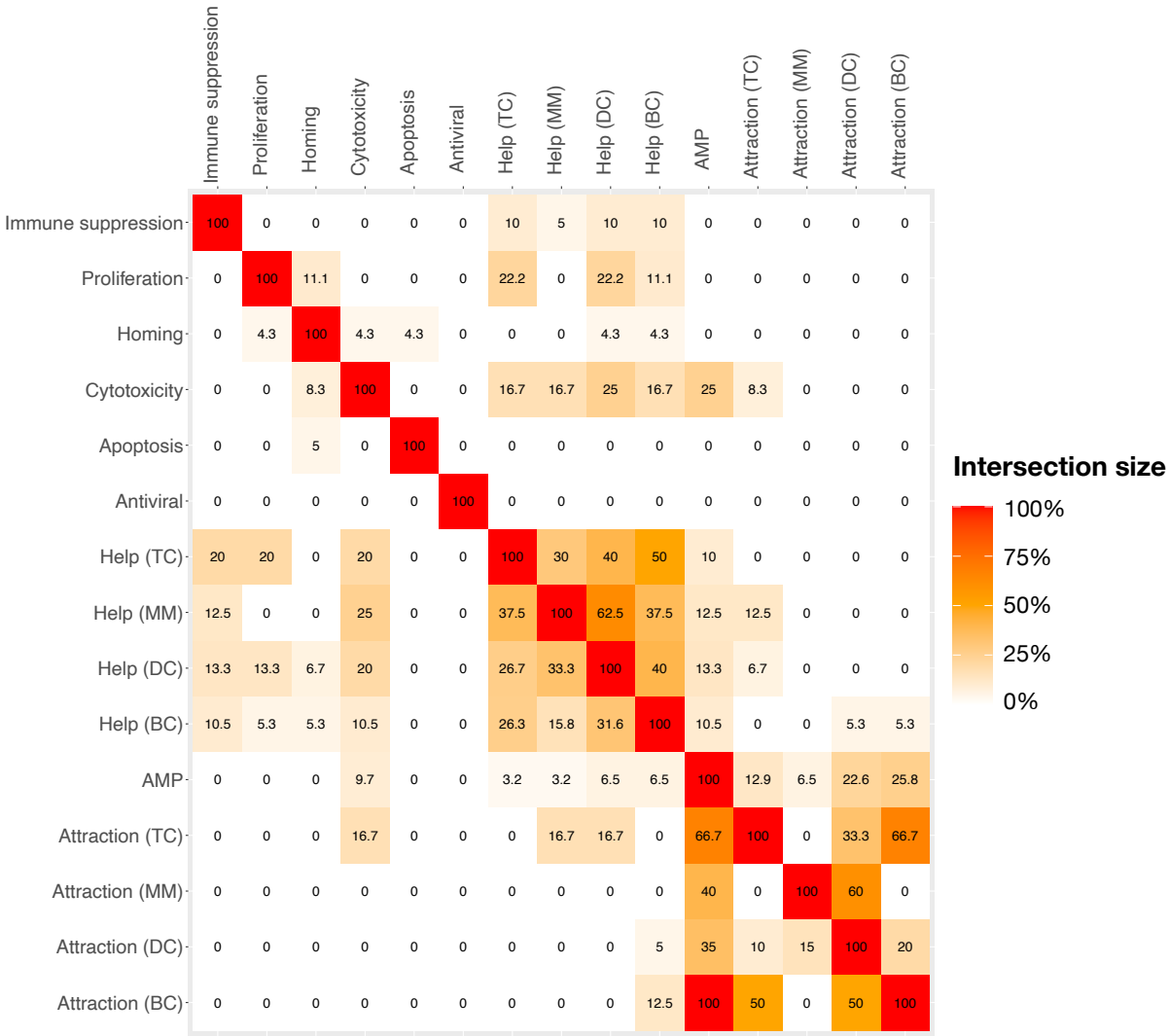


Figure 6.4: Heatmap of the pairwise overlap between functional modules.

6.2.3 Comparison with previous knowledge

I used the Metascape tool [ZZP+19] to confirm the goodness of fit between effector genes and the said function. Listing only the two first GO terms enriched in each module, we observe a match between modules and GO terms, except for the 'help' meta-module, and the AMP (but the antimicrobial term is the third most enriched GO term) (Table 6.2).

Module	GO term	Description	% of genes	-log10 p
BC-attraction	GO:0030593	neutrophil chemotaxis	87.5	16.25
	GO:0072676	lymphocyte migration	87.5	-12.67
DC-attraction	GO:0060326	cell chemotaxis	85	30.51
	GO:0002686	negative regulation of leukocyte migration	25	7.43
MM-attraction	GO:0048245	eosinophil chemotaxis	80	12.13
	GO:0008360	regulation of cell shape	60	5.82
TC-attraction	GO:0071674	mononuclear cell migration	83.3	7.05
BC-help	GO:0031294	lymphocyte costimulation	26.32	6.9
	GO:0031341	regulation of cell killing	26.3	5.88
DC-help	GO:0002694	regulation of leukocyte activation	80	14.12
	GO:0022409	positive regulation of cell-cell adhesion	66.7	13.52
MM-help	GO:1902107	positive regulation of leukocyte differentiation	75	9.18
	GO:1903706	regulation of hemopoiesis	87.5	8.99
TC-help	GO:0051251	positive regulation of lymphocyte activation	90	12.38
	GO:0001819	positive regulation of cytokine production	90	11.9
AMP	GO:0070098	chemokine-mediated signaling pathway	77.4	50.91
	GO:0048247	lymphocyte chemotaxis	58	37.4
Antiviral	GO:0051607	defense response to virus	92.6	44.7
	GO:0048525	negative regulation of viral process	77.8	36.43
Apoptosis	GO:0071214	cellular response to abiotic stimulus	52.6	11.91
	GO:0071550	death-inducing signaling complex assembly	26.3	11.53
Cytotoxicity	GO:0010942	positive regulation of cell death	66.7	10.03
	GO:0019835	cytolysis	33.3	7.12
Homing	GO:0050900	leukocyte migration	65.2	17.17
	GO:0045123	cellular extravasation	30.4	10.21
Proliferation	GO:0048872	homeostasis of number of cells	55.6	5.23
Immune suppression	GO:0050863	regulation of T cell activation	65	16.04
	GO:0007162	negative regulation of cell adhesion	45	9.41

Table 6.2: Modules' GO terms enrichment.

6.3 Scoring functions in single-cell RNAseq data

6.3.1 Encoding

Together with Antonio Rausell, we tested 7 methods to encode each function in each cell, ranging from a basic binary encoding to more sophisticated continuous encodings.

→ Binary encoding: the function is either present (1) or absent (0).

→ Seurat-inspired encodings: mean count of the module's genes, either corrected [SFG+15] (Seurat1) or left untouched (Seurat2).

→ CelliD-inspired encodings: mean inverse rank of the module's genes given by the *GetCellGeneRanking* function of the CelliD package [CMSR21] (CelliD1), or GSEA score computed with the *RunCellGSEA* CelliD function (CelliD2), or distance to the module barycenter in the space of MCA (Multiple Correspondence Analysis)(MCA).

→ Geometric encoding: geometric mean of the percentages $P_{ij,i\in\mathcal{F}}$ of the module's genes $\mathcal{F}$. P_{ij} is the fraction of cells that have a smaller count than x_{ij} for gene i (x_{ij} is the number of reads of gene i in cell j).

Binary: For each cell, we removed the null counts, then we sorted the remaining genes by their level of expression in a decreasing fashion and we discarded the bottom half. Each function is coded by 1 if one of its genes is detected in the genes retained, and 0 otherwise.

```
binary_coding <- function(X, module_list, percentage) {
  top_gene_expressed <- apply(X, 2, function(x) {
                              sorted_genes <- order(x, decreasing=T);
                              keep_nb <-length(x[x!=0])*percentage/100;
                              retain <- names(x)[sorted_genes][1:keep_nb];
                              return(retain)})
  coded <- matrix(0, nrow=length(module_list), ncol=ncol(X))
  rownames(coded) <- names(module_list)
  colnames(coded) <- colnames(X)
  for (i in names(module_list)) {
    features <- module_list[[i]]
    coded[i,] <- sapply(top_gene_expressed, function(x)
                        ifelse(any(features %in% x), 1, 0))
  }
  return(data.frame(coded))
}
```

Seurat1: The Seurat *AddModuleScore* function associates to each gene of a signature a control set of genes from the same expression bin. The score of a signature is the mean expression of the signature's genes minus the mean expression of all the control genes. The score is normalized so that is varies between 0 and 1.

```
AddModuleScore1 <- function(X, module_list, ctrl=100) {
  features <- module_list
  features <- lapply(features, function(x) {
    missing.features <- setdiff(x, rownames(X))
    return(intersect(x, rownames(X)))
  })
  cluster.length <- length(features)
  pool <- rownames(X)
  data.avg <- Matrix::rowMeans(X[pool, ])
  data.avg <- data.avg[order(data.avg)]
  data.cut <- cut_number(data.avg + rnorm(n = length(data.avg))/1e+30,
                         n = 24, labels = FALSE, right = FALSE)
  names(data.cut) <- names(data.avg)
  ctrl.use <- vector(mode = "list", length = cluster.length)
  for (i in 1:cluster.length) {
    features.use <- module_list[[i]]
    for (j in 1:length(features.use)) {
      ctrl.use[[i]] <- c(ctrl.use[[i]],
                         names(sample(data.cut[which(data.cut == data.cut[features.use[j]])],
                         size = ctrl, replace = FALSE)))
    }
  }
  features.scores <- matrix(data = numeric(length = 1L), nrow = cluster.length, ncol = ncol(X))
  for (i in 1:cluster.length) {
    features.use1 <- module_list[[i]]
    features.use2 <- ctrl.use[[i]]
    tmp1 <- X[features.use1,]
    tmp2 <- t(sapply(1:length(features.use1),
                  function(x) Matrix::colMeans(X[features.use2[(ctrl*(x-1)+1):(ctrl*x)],])))
    features.scores[i, ] <- Matrix::colMeans(tmp1-tmp2)
  }
  rownames(features.scores) <- names(module_list)
  colnames(features.scores) <- colnames(X)
  return(features.scores)
}
seurat1_coding <- function(X, module_list) {
  coded <- AddModuleScore1(X, module_list)
  coded <- pbapply(coded, 1, function(x) {
                  tmp<-(x-min(x));
                  return(tmp/max(tmp))})
  return(data.frame(t(coded)))
}
```

Seurat2: We simply computed the mean expression level of all genes from each module. The score is normalized so that it varies between 0 and 1.

```
AddModuleScore2 <- function(X, module_list) {
  features <- module_list
  features <- lapply(features,function(x) {
    missing.features <- setdiff(x, rownames(X))
    if (length(missing.features) > 0) {
      warning("The following features are not present in the object: ",
              paste(missing.features, collapse = ", "))
    }
    return(intersect(x, rownames(X)))
  })
  cluster.length <- length(features)
  pool <- rownames(X)
  features.scores <- matrix(data = numeric(length = 1L), nrow = cluster.length,
                            ncol = ncol(X))
  for (i in 1:cluster.length) {
    features.use <- module_list[[i]]
    data.use <- X[features.use, , drop = FALSE]
    features.scores[i, ] <- Matrix::colMeans(x = data.use)
  }
  rownames(features.scores) <- names(module_list)
  colnames(features.scores) <- colnames(X)
  return(features.scores)
}
seurat2_coding <- function(X, module_list) {
  coded <- AddModuleScore2(X, module_list)
  coded <- pbapply(coded, 1, function(x) {
                tmp<-(x-min(x));
                return(tmp/max(tmp))})
  return(data.frame(t(coded)))
}
```

CelliD1: CelliD ranks all genes in each cell according to their distance to the cell in the MCA space. CelliD1 computes the mean inverse rank, so that a higher score means a higher expression of the function. The scores are normalized afterwards, to restrict the range to the interval $[0, 1]$.

```
cellid1_coding <- function(X, module_list) {
  seurat <- Seurat::CreateSeuratObject(X)
  ranking <- CellID::GetCellGeneRanking(seurat, dim=seq(30))
  max_rank <- length(ranking[[1]])
  coded <- pbsapply(module_list, function(y) {sapply(ranking, function(x) {
                    sum(max_rank+1-which(names(x) %in% y))/length(y)})})
  coded <- data.frame(coded)
  colnames(coded) <- names(module_list)
  return(data.frame(t(coded)))
}
```

CelliD2: CelliD2 uses the rankings computed in the MCA space to perform GSEA. The scores are also normalized afterwards to the interval $[0, 1]$.

```
cellid2_coding <- function(X, module_list) {
  seurat <- Seurat::CreateSeuratObject(X)
  GSEA <- CellID::RunCellGSEA(seurat, pathways = module_list, minSize=1, dim=seq(30))
  ES_matrix <- GetGSEAMatrix(GSEA, metric = "ES")
  coded <- data.frame(ES_matrix[,rownames(seurat@meta.data)])
  coded <- coded[names(module_list),]
  return(coded)
}
```

MCA: The barycenter for each function is defined in the MCA space as the barycenter of its genes. The MCA scores represent the distance between each cell and the barycenter of a given function in the MCA space. Afterwards, the score is normalized and inverted, in order to be maximal for a high expression of the function.

```
distance_to_barycenter <- function(matrix, barycenters) {
  dist_mat <- sapply(1:ncol(barycenters), function(y)
                     sapply(matrix, function(x)
                            dist(rbind(t(x),t(barycenters[y])))))
  return(dist_mat)
}
mca_coding <- function(X, module_list) {
  seurat <- Seurat::CreateSeuratObject(X)
  seurat <- NormalizeData(seurat)
  seurat <- ScaleData(seurat, features = rownames(seurat))
  seurat <- RunMCA(Baron)
  coordinates <- seurat@reductions$mca
  archetypes <- data.frame(pblapply(module_list, function(x) {
                               genes_common <- x[x %in% rownames(coordinates@feature.loadings)];
                               return(ifelse(rep(is.null(dim(coordinates@feature.loadings[genes_common,])),
                                                 50),
                                             coordinates@feature.loadings[genes_common,],
                                             colMeans(coordinates@feature.loadings[genes_common,]))})))
  coordinates <- cbind(t(coordinates@cell.embeddings), archetypes)
  df <- data.frame("MCA1"=unname(t(coordinates[1,])),
                   "MCA2"=unname(t(coordinates[2,])),
                   "Cell"=sapply(colnames(coordinates), function(x)
                                 ifelse(x %in% gsub("-",".",names(module_list)),x,"normal")),
                   "Cell2"=sapply(colnames(coordinates),function(x)
                                 ifelse(x %in% gsub("-",".",names(module_list)),
                                        "archetypical",
                                        "normal")))
  #ggplot(df, aes(x=MCA1, y=MCA2, color=Cell)) + geom_point()
  #ggplot(df, aes(x=MCA1, y=MCA2, color=Cell2)) + geom_point()
  # Get distances
  coded <- distance_to_barycenter(data[,1:ncol(seurat)],data[,(ncol(seurat)+1):ncol(data)])
  colnames(coded) <- names(module_list)
  return(t(coded))
}
```

Geometric: I computed P_{ij}, the fraction of cells that have a smaller count than x_{ij} for gene i, and I took the geometric mean of one module's percentages $P_{ij,i\in\mathcal{F}}$.

```
geom_mean <- function(matrix, gene_set) {
  filter_matrix <- matrix[gene_set,]
  product <- apply(filter_matrix, 2, function(x) prod(x, na.rm = T))
  return(nthroot(product, length(gene_set)))
}
percentage_coding <- function(X, module_list, ctrl=20) {
  features <- module_list
  features <- lapply(features, function(x) {
                     return(intersect(x, rownames(X)))
  })
  cluster.length <- length(features)
  pool <- rownames(X)
  data.avg <- Matrix::rowMeans(X[pool, ])
  data.avg <- data.avg[order(data.avg)]
  data.cut <- cut_number(data.avg + rnorm(n = length(data.avg))/1e+30,
                         n = 24, labels = FALSE, right = FALSE)
  names(data.cut) <- names(data.avg)
  ctrl.use <- vector(mode = "list", length = cluster.length)
  for (i in 1:cluster.length) {
    features.use <- module_list[[i]]
    for (j in 1:length(features.use)) {
      ctrl.use[[i]] <- c(ctrl.use[[i]],
                         names(sample(data.cut[which(data.cut == data.cut[features.use[j]])],
                         size = ctrl, replace = FALSE)))
    }
  }
  ctrl.use <- unique(unlist(ctrl.use))
  pool2 <- unique(c(ctrl.use, unique(unlist(features)))) # Reduce the size to speed the computation
  coded1 <- pbapply(X[pool2,], 1, function(x)
                    sapply(x, function(y) mean(y<x, na.rm = T)))
  rownames(coded1) <- colnames(X)
  coded2 <- data.frame(t(pbsapply(module_list,function(x) geom_mean(coded1,x))))
  return(data.frame(t(coded2)))
}
```

I evaluated the proximity between the 7 encodings described above, looking at the Spearman correlation on the data provided by the SingleR *MonacoImmuneData* function. All correlations were reassuringly positive, but below 0.5 except for the 4 encodings Seurat1, Seurat2, CelliD1 and CelliD2. MCA and binary encodings were the most dissimilar, with mean Spearman correlation to the other encodings of 0.16 and 0.23 respectively, shortly followed by the geometric encoding (mean correlation of 0.27), Seurat1 (mean correlation of 0.48), Seurat2 (mean correlation of 0.49), CelliD1 (mean correlation of 0.50) and CelliD2 (mean correlation of 0.51) (Figure 6.5).

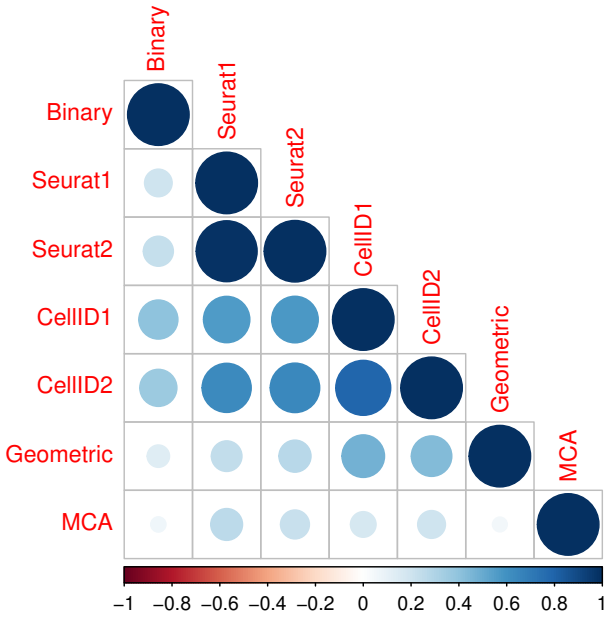


Figure 6.5: Spearman correlation between the 7 encoding methods.

6.3.2 Control with bulk RNAseq data

We performed a positive control of our encoding methods using bulk RNAseq data. I collected bulk RNAseq data from pure populations, using the data provided with the *MonacoImmuneData* function from the SingleR package. I could test BC-help by extracting the scores for Tfh, cytotoxicity by extracting the scores for $CD8^+$ T cells, immune suppression by extracting the scores for Treg, and homing by extracting the scores for memory T cells. The positive control proved useful, as it allowed to eliminate meaningless encoding methods: we could not validate the binary, MCA and geometric methods. Binary encoding showed no difference between the two cell populations in any of the 4 tests. MCA encoding could not retrieve the proper trend when testing BC-help and homing. Geometric encoding showed no difference for the BC-help and the immune suppression test (Figure 6.6).

For the next steps, I will only continue to test the 2 Seurat and the 2 CelliD encodings.

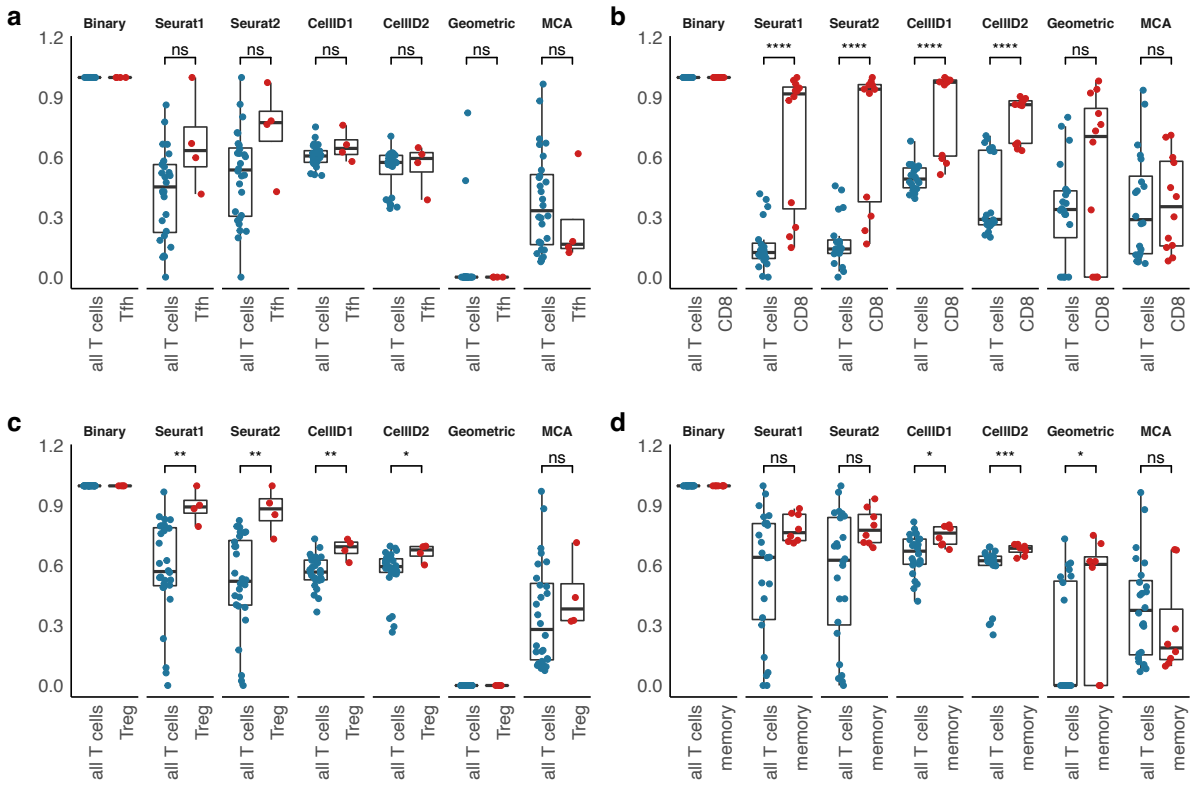


Figure 6.6: Testing the functional encoding methods with pure bulk populations. (**a**) The BC-help module is expected to score higher in Tfh. (**b**) The cytotoxic module is expected to score higher in $CD8^+$ T cells. (**c**) The immune suppression module is expected to score higher in Treg. (**d**) The homing module is expected to score higher in memory T cells.

6.4 Assessing the added value of the functional modules

I used the following datasets for my experiments: GSE99254 [GZZ+18], GSE108989 [ZYZ+18], GSE98638 [ZZY+17]. They are respectively from NSCLC, colorectal cancer and liver cancer patients, contain the tumor, juxtatumor and blood tissues, and were sorted before the sequencing into 4 populations: $CD8^+$, $CD4^+CD25^{low}$, $CD4^+CD5^{medium}$ and $CD4^+CD25^{high}$ cells. With these datasets, I have been verifying that the functional modules approach would bring new information, as compared to ground truth or cluster labels.

First I quantified the fraction of the information in the data brought by the 172 effector genes, by evaluating the variance explained by the functional genes, as opposed to the variance explained by random genes from the same expression bin in the PCA, the MCA or the UMAP spaces. The effector genes explain a higher fraction of the information contained in the first 50 PCs, as compared to the same number of random genes (Figure 6.7). Visually, the projection seems to remain meaningful if calculated from the PCA or the MCA coordinates obtained using only effector genes as compared to random genes from the same expression bin (Figure 6.8).

From Figures 6.7 and 6.8, we concluded that the effector genes represent a reasonable amount of the information contained in the scRNAseq count matrix. Furthermore, intu-

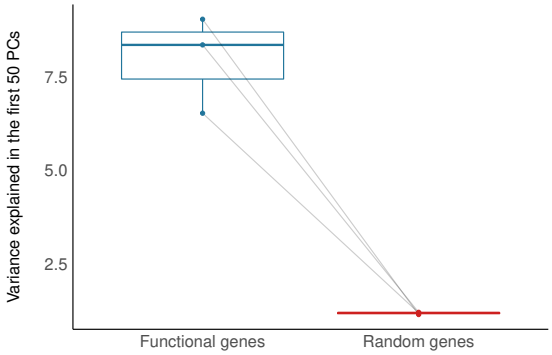


Figure 6.7: Variance explained by functional genes exceeds variance explained by the same number of random genes from the same expression bin.

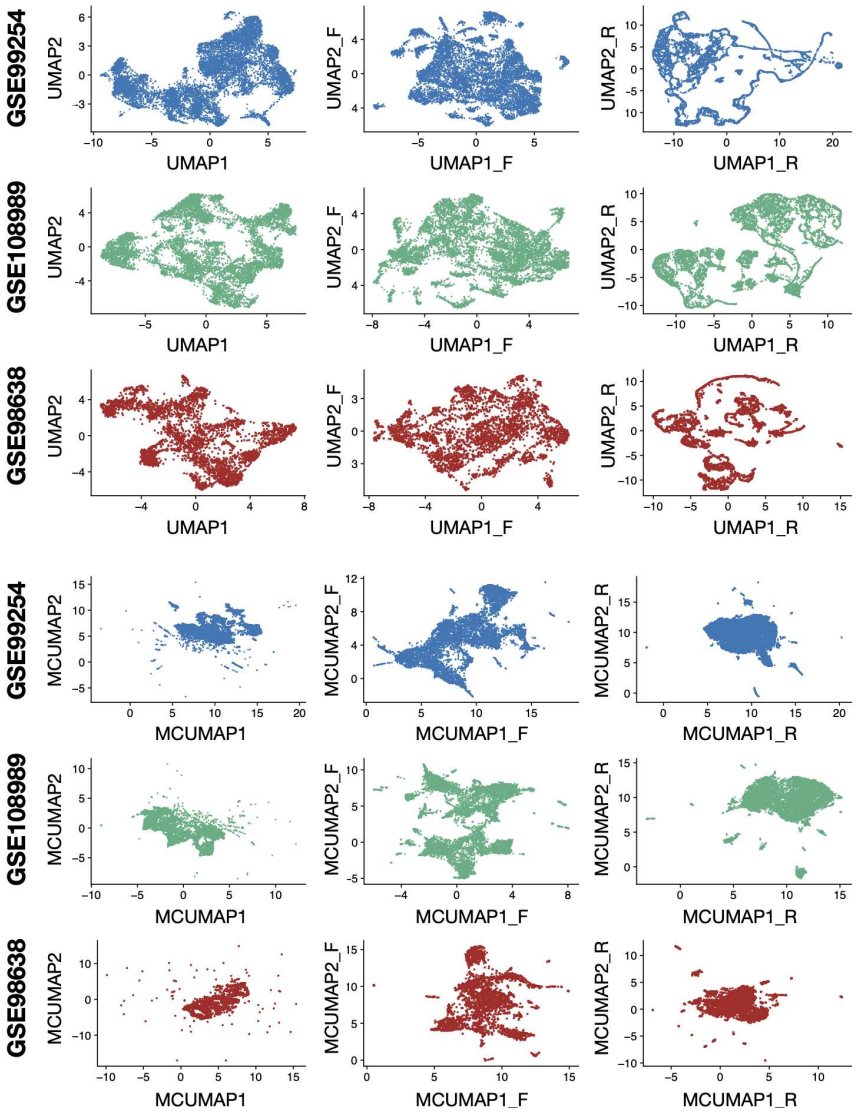


Figure 6.8: UMAP computed from the PCA (first 3 rows) or MCA (second 3 rows) spaces, using all genes (first column), effector genes (second column) or random genes (third column). Visually, the UMAP computed from the PCA and with the effector genes is similar to the one computed with all genes, while the one with random genes is completely different.

ition based on the UMAP projection, either regular UMAP or MCA-based MCUMAP, seems promising.

6.4.1 Mutual information between ground truth labels and module scoring

We selected the 3 datasets described above as they contain T cells from the TME and they display ground truth labels that are suited for our experiments, since they sorted cytotoxic $CD8^+$ cells and regulatory $CD4^+CD25^{high}$ cells, that will help us guide our exploration.

We checked the overlap between ground truth labels and functional encodings by trying to fit a generalized linear model for binomial data between the latter. We could not validate any of these models, based on their p-value that were all above 0.05. We also computed the silhouette score, using ground truth labels and UMAP coordinates computed using the effector genes, with the caveat that silhouette should work best for convex clusters. Here again, it is another proof that there is no information shared between ground truth labels and functional modules, with low silhouette scores (0.043, -0.029 and 0.081 resp. for GSE99254, GSE108989 and GSE98638, Figure 6.9).

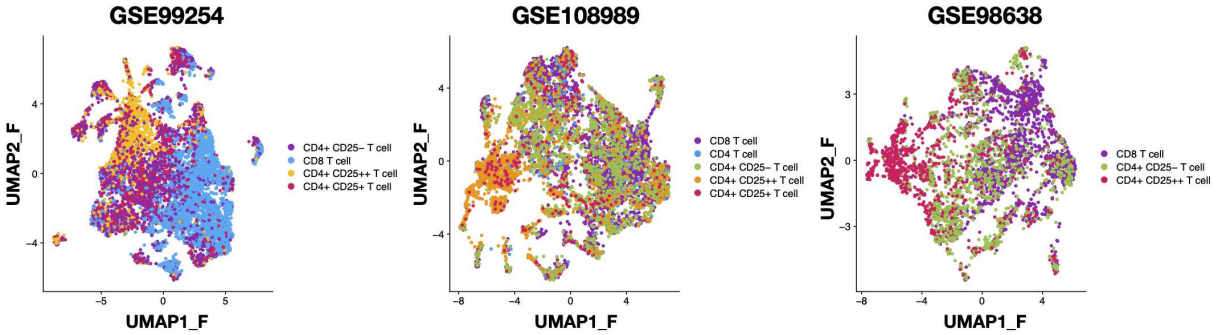


Figure 6.9: Ground truth labels are mixed in the effector genes UMAP.

Lastly, we can visualise the encoding directly on the classical UMAP computed with all genes. We observed that while some functions, such as immune suppression, exhibit a gradient along a given direction of the 2D UMAP projection, it is not the case for all functions, such as proliferation (Figure 6.10).

6.4.2 Mutual information between cluster labels and module scoring

We did a second sanity check in order to assess the novelty of the functional encoding approach: we evaluated the overlap between unsupervised cluster labels and modules encoding.

The first control that we conducted has been to quantify the mutual information between PCA coordinates and encodings (and not simply the raw effector genes as done in section 6.4). I compared encodings to PCA coordinates since the latter are a common ingredient for clustering, trajectory inference or visualisation algorithms. I used a straightforward linear model, and assessed its validity:

$$Score_{\mathcal{F}} \sim \sum_{i=1}^{20} \alpha_i PC_i$$

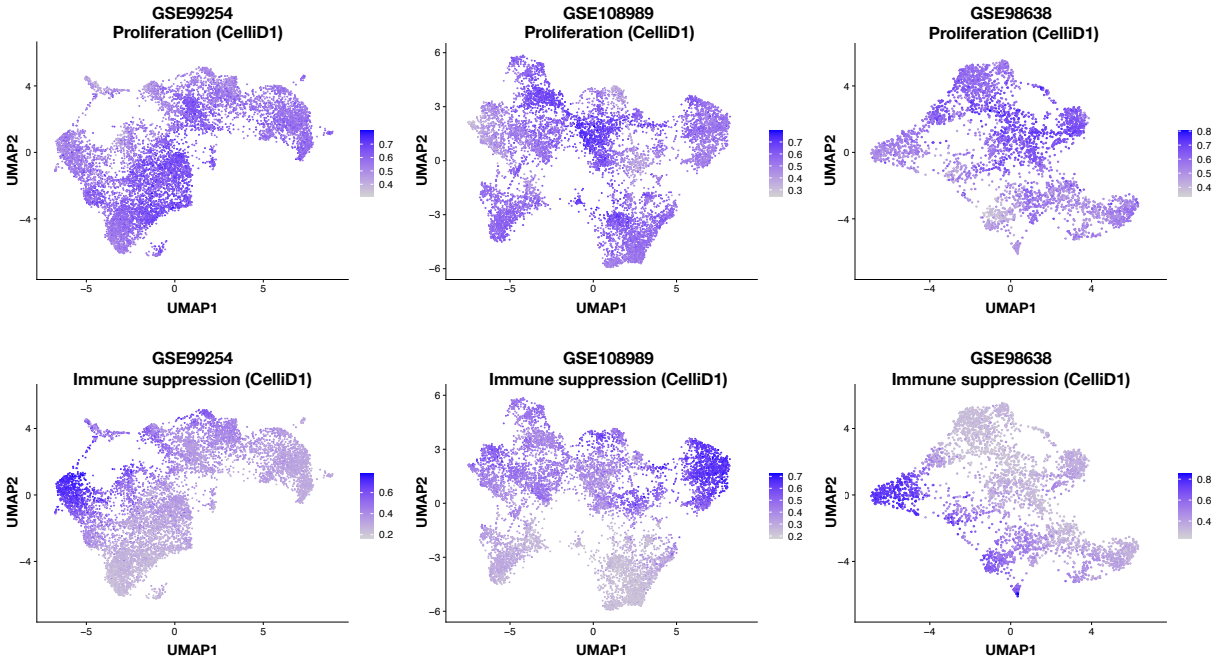


Figure 6.10: Some functions do not mix on the classical UMAP projection, while some others do.

where $Score_{\mathcal{F}}$ is a cell function encoding, PC_i is the cell coordinate vector of the i-th PC and α_i the associated coefficient in the linear model. Again, no model has been validated on the basis of R^2 values above 0.9.

The second control that we did has been to evaluate (i) whether the difference in functional module scores between clusters was not significant (or less significant than by chance), which could mean that cells with similar functional patterns would be clustered separately, and (ii) whether the intra-cluster variance is higher than by chance for each function, which could mean that cells with differential functional patterns would be regrouped in the same cluster. I used the cluster labels provided by the authors of the different datasets, that they obtained from an unsupervised analysis of their data. For test (i), we verified for each encoding method and the 3 datasets (but we show only GSE99254 and a subset of 2 functions for the sake of space) that the scores were on average more similar across clusters than expected (Figures 6.11, 6.12).

For test (ii), we also verified that the intra-cluster variance is higher than expected on average (Figure 6.13).

From the results of these tests, it confirms the hypothesis that the functional encoding approach is innovative as compared to the unsupervised analysis pipeline.

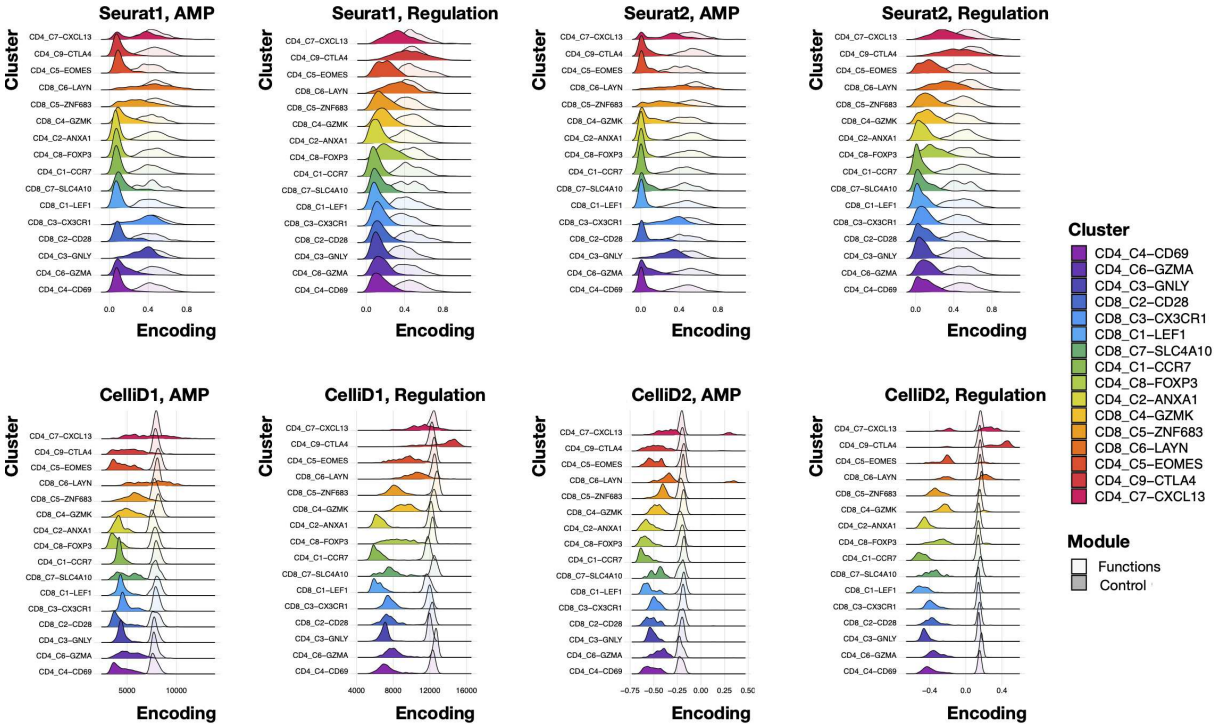


Figure 6.11: Encoding scores distribution for 2 functions in dataset GSE99254.

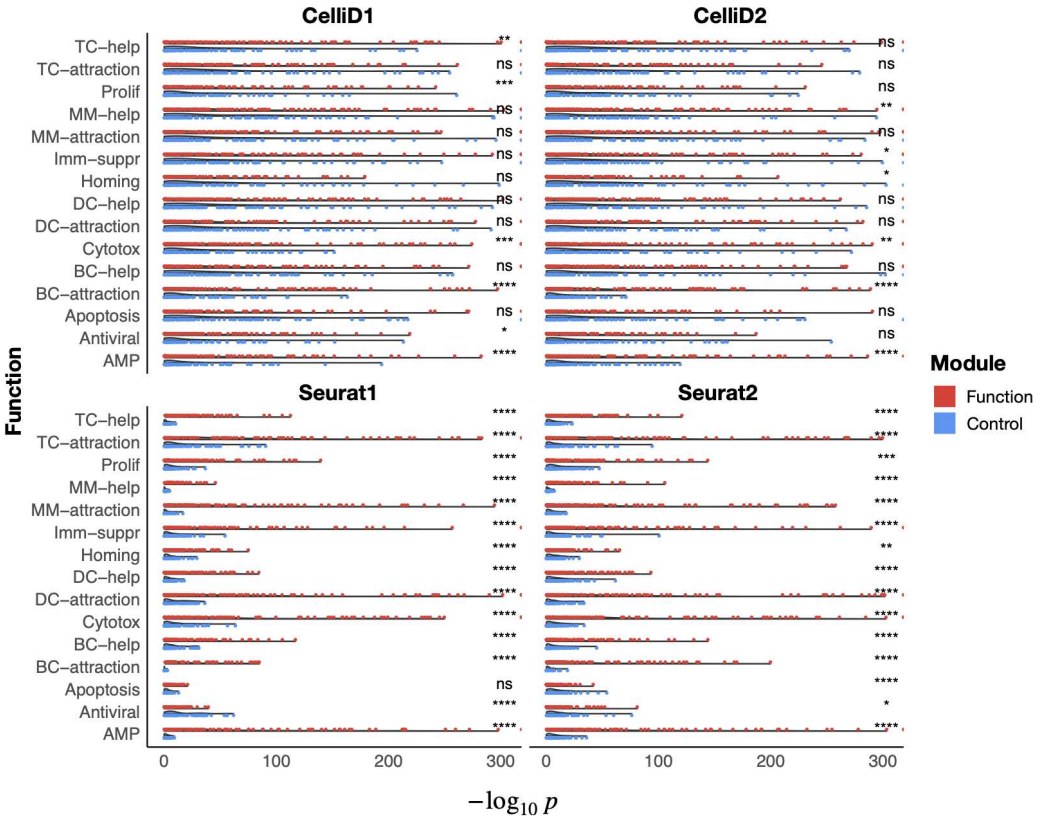


Figure 6.12: Inter-cluster differences are smaller for functional encodings than expected on average.

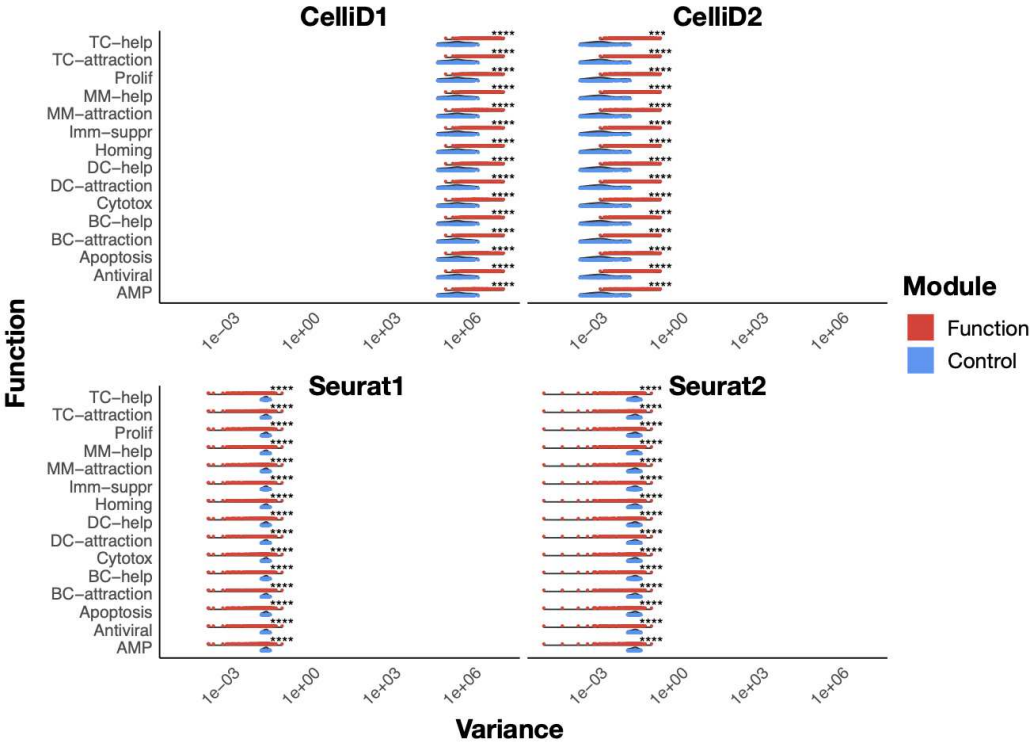


Figure 6.13: Intra-cluster differences are higher for functional encodings than expected on average.

6.5 Mapping tumoral T cells functions

For a first glance of the data and the encodings, we looked at the fraction of cells positive for each single function, as a function of the tissue of origin. We deemed a cell as positive if its score for a given function was above 0.5. The first observation that we made is that cells in the tumor are more functional than in the juxtatumor, in the sense that there is a higher proportion of cells positive for single functions in the tumor, with an average surplus of 0.49% of cells in the tumor as compared to the juxtatumor, across encoding methods and datasets (Figure 6.14).

Regarding reproducible trends across all datasets and encodings, there is a higher fraction of cells exerting immune suppression, or DC or MM attraction in the tumor than in the juxtatumor (Figure 6.14).

6.5.1 Barcoding cells

Although it is possible to barcode cells individually, we chose to implement clustering in the first instance, in order to be able to give an overall picture of the data. For the clustering, I considered 4 methods using the Seurat [SFG+15] clustering function based on the Louvain algorithm, except the last one for which I used the CiteFuse package [KLG+20] and the Louvain algorithm:

→ classical unsupervised clustering (method 1),

→ unsupervised clustering using only functional effector genes (method 2),

→ clustering on the scores for the 15 modules (method 3),

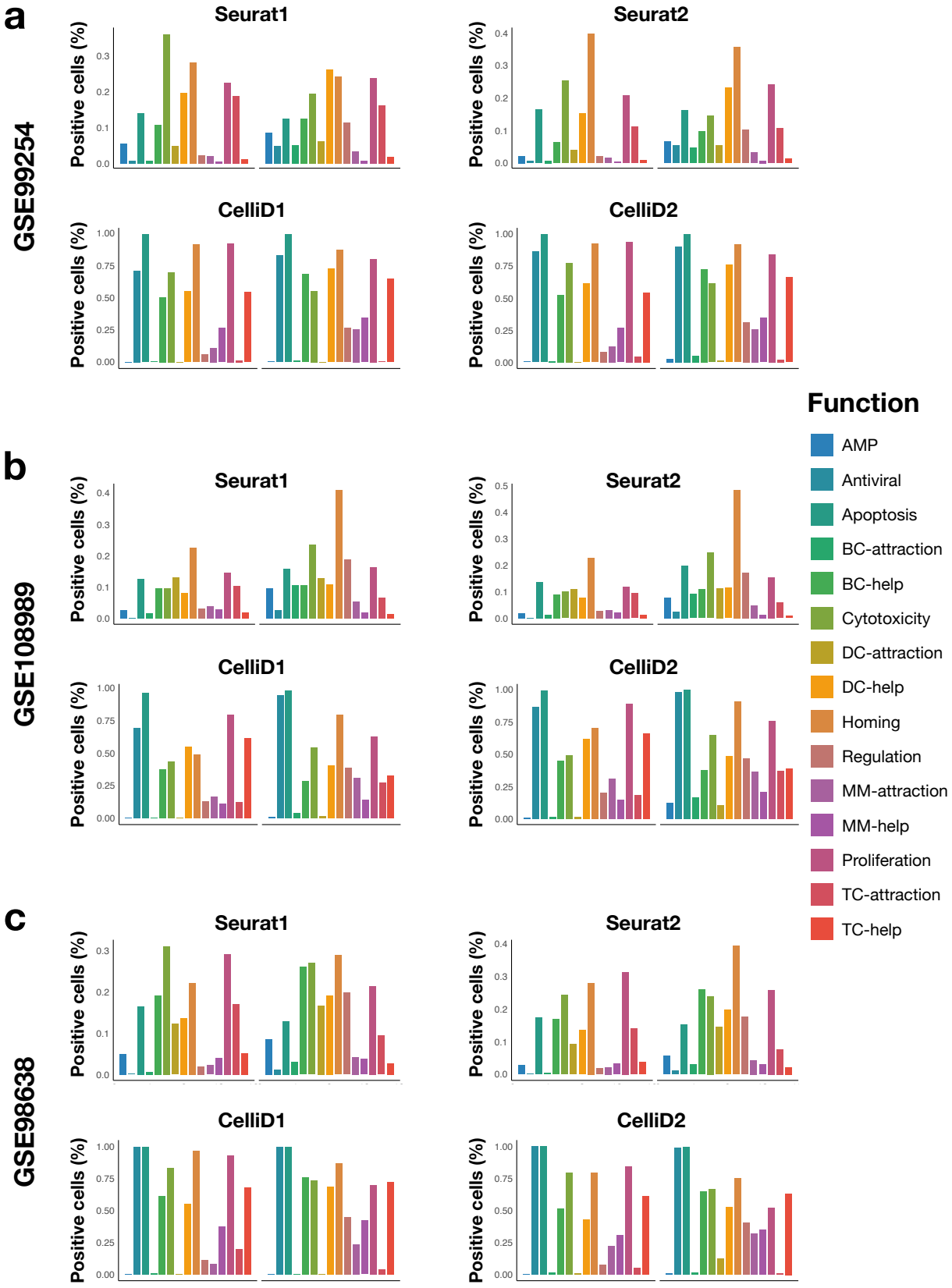


Figure 6.14: Fraction of cells positive for each function for the 4 encoding methods and 3 datasets. GSE99254 (**a**), GSE108989 (**b**) and GSE98638 (**c**). "Regulation" stands for immune suppression.

→ consensus clustering, using the RNA and the encoding information (method 4).

The first method quickly turned out to not be valid, since we already proved that classically obtained clusters could not overlap with the functional information brought by the encodings. Hence functionally barcoding these clusters is meaningless. Regarding the second method, we realised that its labels were quite in agreement with the labels obtained via the first method, so we assumed it was not a good strategy either (Figure 6.15). The third and forth methods were also in agreement with each other, and differed from the first 2 methods (Figure 6.15), with the advantage of the third being more scalable, but the forth probably more informative. We decided to use the forth method, but we did not rule out the possibility to use later the encoding-based clustering because of the problematic scalability of the consensus clustering method that we used, especially in a context where dataset cardinality is increasing fast: the limiting factor is the construction of the consensus affinity matrix based on the similarity network fusion method [WMD+14].

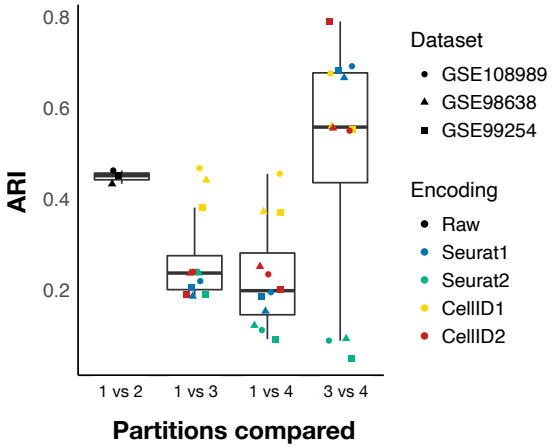


Figure 6.15: Similarity, using the ARI, between the 4 clustering methods listed.

6.5.2 Interpretation

To barcode the different clusters obtained with the forth method, we opted for a straightforward manner, considering clusters as positive for a function if the median cluster score for this function is above 0.5 for the Seurat encodings, or 0.6 for the CelliD encodings. We chose these thresholds based on the scores distribution (Figure 6.16).

Based on these choices, we investigated functional barcodes in order to compare the juxtatumoral versus the tumoral tissues. We decided to discard the Seurat encodings at this step because they were not informative: most of the clusters could not be functionally barcoded, as they were mostly negative for all functions, and the few positive clusters would be positive only for a single function (cytotoxicity for Seurat1, AMP, homing, or TC-attraction for Seurat2). We also excluded for now CelliD2 encodings on the basis of the UMAP projection (Figure 6.17).

For the CelliD1 encoding, we noticed the following trends: a slightly above-average number of clusters are more enriched in the tumor compared to the juxtatumor (58% in average, being 64%, 45% and 56% of the clusters in GSE99254, GSE108989 and GSE98638 respectively). Together with the fact that the Gini index is repeatedly higher for the juxtatumor than the tumor, it is indicative of a higher functional diversity in the tumor

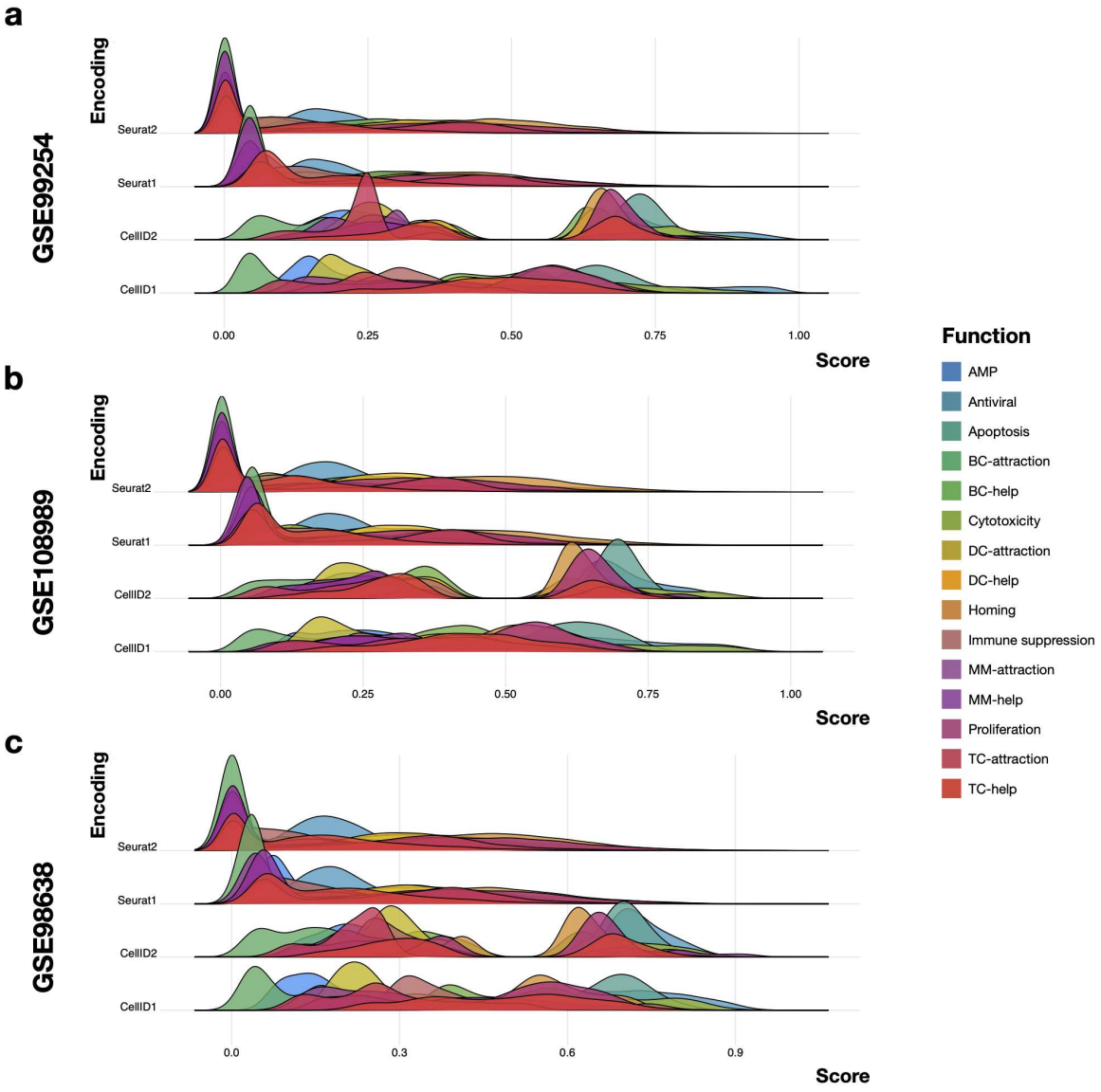


Figure 6.16: Distribution of the scores for the 4 encoding methods and 3 datasets. GSE99254 (**a**), GSE108989 (**b**) and GSE98638 (**c**).

than in the juxtatumor. This is coherent with previous observations of a higher spread (termed "phenotypic volume") of the phenotypes in the tumor compared to healthy tissue [ACP+18]. The mean number of functions exerted in a cluster is of 5 for clusters enriched in the juxtatumor and 6.2 for clusters enriched in the tumor. Additionally, we observed that functional patterns are more conserved across tumoral than juxtatumoral tissues: out of the 14 clusters that were enriched in the juxtatumor compared to the tumor, only 1 was common to GSE99254 and GSE98638, while out of the 14 tumor-related clusters, 3 were common to 2 datasets and 1 to the 3 datasets. This is aligned with the idea that cancer types share hallmarks [HW00, HW11]. Lastly, we observed 5 cluster types that would be enriched in the tumor of one dataset and in the juxtatumor of another dataset. Notably, the conflicts were always with the NSCLC article (GSE99254).

Regarding individual clusters, the tumor was enriched in clusters exerting immune suppression and homing at the same time, and MM attraction. Furthermore, we can study the correlations between the functions in the tumor or the juxtatumor. We observed that

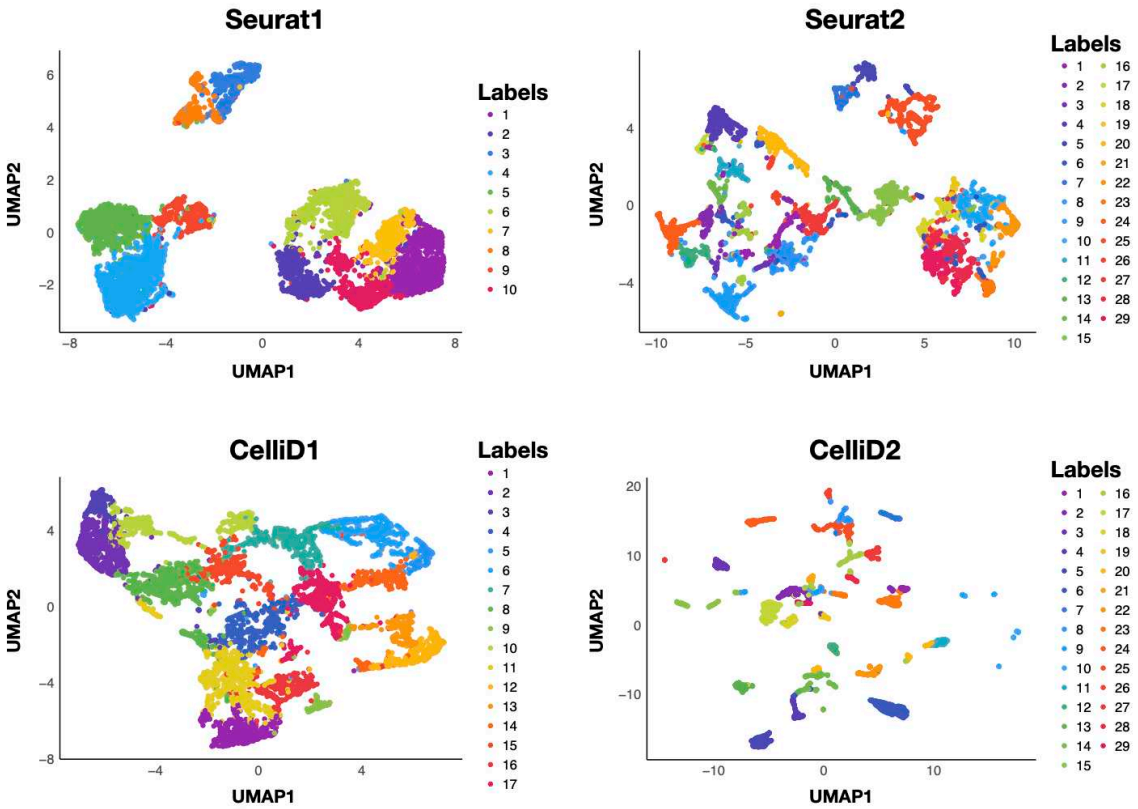


Figure 6.17: UMAP computed on the consensus matrix, with consensus clustering labels (showing GSE99254).

there is a stronger intertwining of functions in the tumor as compared to the juxtatumor, especially for the immune suppression, the MM-help and -attraction functions (Figure 6.18).

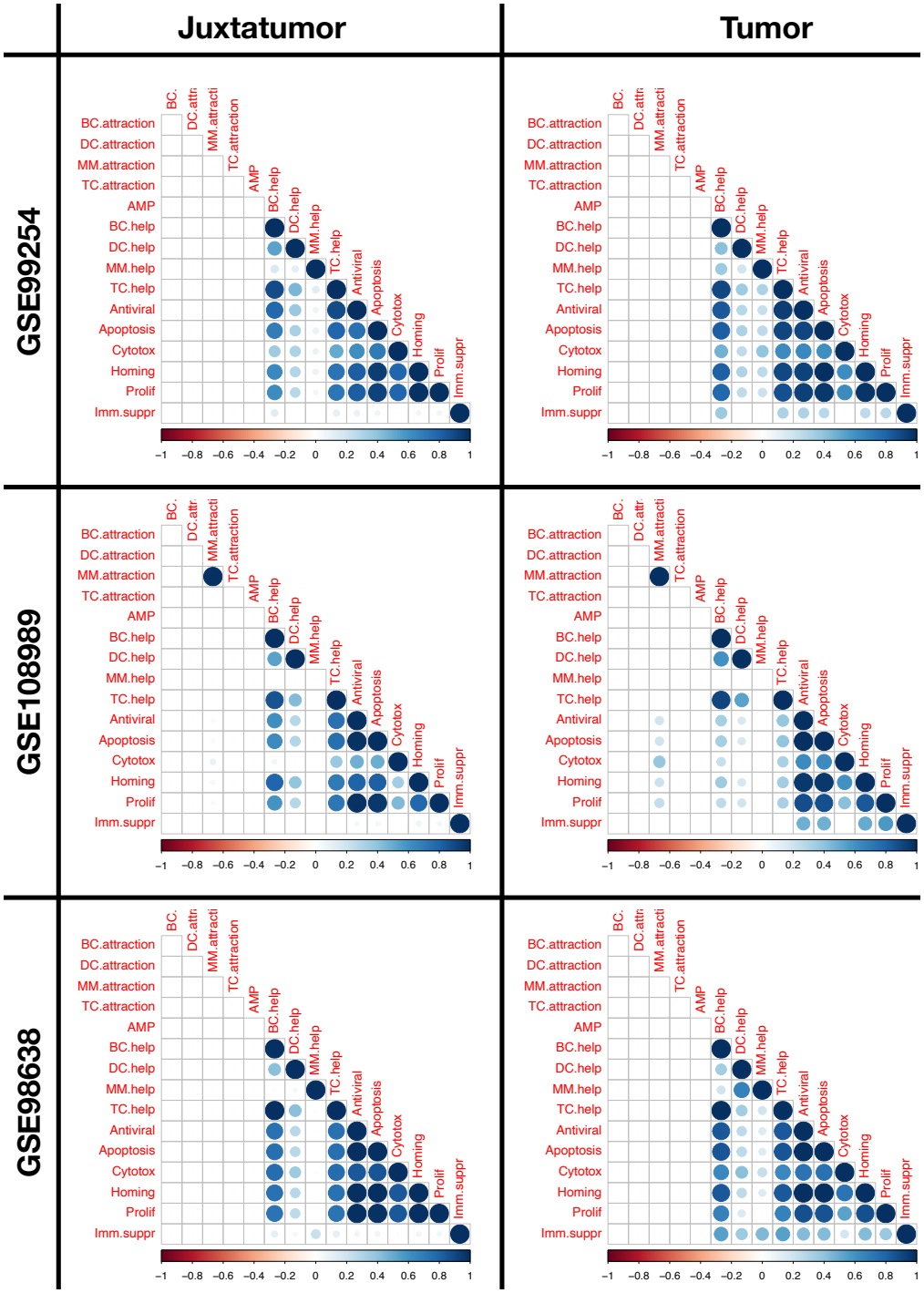


Figure 6.18: Functions are differentially correlated according to the tissue of origin.

6.6 Conclusion

We propose a novel supervised approach to analyse scRNAseq datasets of T cells. We suggest this approach in the conceptual frame of proposing a new classification of T cells based on their functions, as a way to challenge the current lineage paradigm.

First, we defined T-cell-related functions, outlining 15 of them: 4 help functions toward B cells, dendritic cells, monocytes-macrophages and T cells, 4 attraction functions

towards the same 4 subsets, anti-microbial peptide production, antiviral, apoptosis, cytotoxicity, homing, immune suppression and proliferation. I selected from the literature all genes that were effector for those functional modules.

Then, I verified that the supervised approach would not overlap with the unsupervised one, by quantifying the fit between the two approaches. I tested 7 methods to score the functional modules in each cell. The agreement was negligible between the supervised and unsupervised strategies. It was evaluated either by linear models between classical labels and encoding information, or by looking at the inter- and intra-cluster encoding variance in order to assess the possibility of having cells with similar functional patterns in different clusters, or cells with different functional patterns in the same cluster.

With our onco-immune datasets, we could explore the differences between the tumor and juxtatumor. We observed that the tumor exhibited a higher number of functions, in the sense that more cells exert functions and cells exert also more functions in the tumor than in the juxtatumor [ACP+18]. Tumoral cells were also more scattered across a higher diversity of clusters, while tumors are more conservative with regard to their functional patterns across cancer types [HW00]. However, some functional patterns are found alternatively enriched in the juxtatumor or in the tumor. Lastly, the tumor is globally more immune suppressing and attracting other immune cells than the juxtatumor which looks more like a dormant tissue.

The functional approach offers the following benefits: first of all it is supervised. In order to establish trust in our signatures, I verified whether it was in agreement with GO terms, and I did a positive control with bulk RNAseq data. Thus, making the assumption that the signatures were sufficiently carefully curated, the supervised approach avoids the tedious exploration step. Secondly, it simplifies the interpretation of the data with a straightforward reading of the functions and the functional patchworks present in the data. Thirdly, it simplifies the comparison across datasets, as we did with the 3 datasets that we tested.

6.7 Perspectives

However, we believe our method could be further improved, especially the last steps. In particular, we should test alternative ways to barcode clusters. We chose to use a threshold on the median function score, but we could explore other ways: for example, it would be interesting to use at least a fuzzy threshold or to find a solution to retain quantitative information regarding the magnitude of the score. Investigate more extensively barcoding methods could enable to explore more in depth Seurat encodings, that we discarded since they were not informative with our barcoding method. We also chose one clustering method, the consensus method, as we deemed it as more informative. First, we could explore why the consensus clustering agrees strongly with the encoding-based clustering, while it does not with the classical clustering, although we suspect it is due to similarity networks having a higher weight for encodings than for the RNA information. Second, we could also try to label the clusters we obtained with the RNA information, notwithstanding the supervised approach we implement here, in order to further characterize the added value of our novel approach. Third, we did not exploit here the advantage of scRNAseq which is to work at the single-cell scale. It should be one of the key focus to improve the method.

We also discussed in the introduction the fact that the RNA information would often

be used as a proxy for the missing protein information. Our approach relies fully on the assumption that the presence of the RNA of an effector gene indicates the corresponding functionality. Hence, all computational findings should be carefully confirmed at the bench.

References

[ACP+18] Elham Azizi, Ambrose J. Carr, George Plitas, Andrew E. Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, Kristy Choi, Rachel M. Fromme, Phuong Dao, Peter T. McKenney, Ruby C. Wasti, Krishna Kadaveru, Linas Mazutis, Alexander Y. Rudensky, and Dana Pe'er. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174(5): 1293–1308.e36, August 2018. ISSN 00928674.

[BKG+16] Ravikiran Bhairavabhotla, Yong C. Kim, Deborah D. Glass, Thelma M. Escobar, Mira C. Patel, Rami Zahr, Cuong K. Nguyen, Gokhul K. Kilaru, Stefan A. Muljo, and Ethan M. Shevach. Transcriptome profiling of human FoxP3+ regulatory T cells. *Human Immunology*, 77(2):201–213, February 2016. ISSN 1879-1166.

[CMSR21] Akira Cortal, Loredana Martignetti, Emmanuelle Six, and Antonio Rausell. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nature Biotechnology*, pages 1–8, April 2021.

[GZZ+18] Xinyi Guo, Yuanyuan Zhang, Liangtao Zheng, Chunhong Zheng, Jintao Song, Qiming Zhang, Boxi Kang, Zhouzerui Liu, Liang Jin, Rui Xing, Ranran Gao, Lei Zhang, Minghui Dong, Xueda Hu, Xianwen Ren, Dennis Kirchhoff, Helge Gottfried Roider, Tiansheng Yan, and Zemin Zhang. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine*, 24(7):978–985, 2018. ISSN 1546-170X.

[HDM+20] Barbara Höllbacher, Thomas Duhen, Samantha Motley, Maria M. Klicznik, Iris K. Gratz, and Daniel J. Campbell. Transcriptomic Profiling of Human Effector and Regulatory T Cell Subsets Identifies Predictive Population Signatures. *ImmunoHorizons*, 4(10):585–596, October 2020. ISSN 2573-7732.

[HW00] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, January 2000. ISSN 0092-8674, 1097-4172.

[HW11] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, March 2011. ISSN 0092-8674, 1097-4172.

[JASC+18] Livnat Jerby-Arnon, Parin Shah, Michael S. Cuoco, Christopher Rodman, Mei-Ju Su, Johannes C. Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, Jia-Ren Lin, Shu Wang, Bokang Rabasha, David Liu, Gao Zhang, Claire Margolais, Orr Ashenberg, Patrick A. Ott, Elizabeth I. Buchbinder, Rizwan Haq, F. Stephen Hodi, Genevieve M. Boland, Ryan J. Sullivan, Dennie T. Frederick, Benchun Miao, Tabea Moll, Keith T. Flaherty, Meenhard Herlyn, Russell W. Jenkins, Rohit Thummalapalli, Monika S. Kowalczyk,

Israel Cañadas, Bastian Schilling, Adam N. R. Cartwright, Adrienne M. Luoma, Shruti Malu, Patrick Hwu, Chantale Bernatchez, Marie-Andrée Forget, David A. Barbie, Alex K. Shalek, Itay Tirosh, Peter K. Sorger, Kai Wucherpfennig, Eliezer M. Van Allen, Dirk Schadendorf, Bruce E. Johnson, Asaf Rotem, Orit Rozenblatt-Rosen, Levi A. Garraway, Charles H. Yoon, Benjamin Izar, and Aviv Regev. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell*, 175(4):984–997.e24, November 2018.

[KLG+20] Hani Jieun Kim, Yingxin Lin, Thomas A Geddes, Jean Yee Hwa Yang, and Pengyi Yang. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*, 36(14):4137–4143, July 2020.

[LXW+21] Yuechen Luo, Changlu Xu, Bing Wang, Qing Niu, Xiuhua Su, Yingnan Bai, Shuxian Zhu, Chunxiao Zhao, Yunyan Sun, Jiali Wang, Maolan Liu, Xiaolei Sun, Ge Song, Haidong Cui, Xiaoli Chen, Huifang Huang, Haikun Wang, Mingzhe Han, Erlie Jiang, Lihong Shi, and Xiaoming Feng. Single-cell transcriptomic analysis reveals disparate effector differentiation pathways in human Treg compartment. *Nature Communications*, 12(1):3913, June 2021.

[MKCK+21] Ksenia Magidey-Klein, Tim J. Cooper, Ksenya Kveler, Rachelly Normand, Tongwu Zhang, Michael Timaner, Ziv Raviv, Brian P. James, Roi Gazit, Ze'ev A. Ronai, Shai Shen-Orr, and Yuval Shaked. IL-6 contributes to metastatic switch via the differentiation of monocytic-dendritic progenitors into prometastatic immune cells. *Journal for ImmunoTherapy of Cancer*, 9 (6):e002856, June 2021.

[PWS+16] Anne M. Pesenacker, Adele Y. Wang, Amrit Singh, Jana Gillies, Youngwoong Kim, Ciriaco A. Piccirillo, Duc Nguyen, W. Nicholas Haining, Scott J. Tebbutt, Constadina Panagiotopoulos, and Megan K. Levings. A Regulatory T-Cell Gene Signature Is a Specific and Sensitive Biomarker to Identify Children With New-Onset Type 1 Diabetes. *Diabetes*, 65(4):1031–1039, April 2016. ISSN 1939-327X.

[SFG+15] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, May 2015. ISSN 1546-1696.

[WMD+14] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, March 2014.

[ZYZ+18] Lei Zhang, Xin Yu, Liangtao Zheng, Yuanyuan Zhang, Yansen Li, Qiao Fang, Ranran Gao, Boxi Kang, Qiming Zhang, Julie Y. Huang, Hiroyasu Konno, Xinyi Guo, Yingjiang Ye, Songyuan Gao, Shan Wang, Xueda Hu, Xianwen Ren, Zhanlong Shen, Wenjun Ouyang, and Zemin Zhang. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*, 564(7735):268–272, December 2018.

[ZZK+18] David Zemmour, Rapolas Zilionis, Evgeny Kiner, Allon M. Klein, Diane Mathis, and Christophe Benoist. Single-cell gene expression reveals a land-

scape of regulatory T cell phenotypes shaped by the TCR. *Nature Immunology*, 19(3):291–301, March 2018.

[ZZP+19] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10:1523, April 2019.

[ZZY+17] Chunhong Zheng, Liangtao Zheng, Jae-Kwang Yoo, Huahu Guo, Yuanyuan Zhang, Xinyi Guo, Boxi Kang, Ruozhen Hu, Julie Y. Huang, Qiming Zhang, Zhouzerui Liu, Minghui Dong, Xueda Hu, Wenjun Ouyang, Jirun Peng, and Zemin Zhang. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell*, 169(7):1342–1356.e16, June 2017. ISSN 0092-8674.

Chapter 7

Meta-analysis of regulatory T cells in cancer: highlighting their prognostic role in a context-dependent manner

7.1 Rationale behind the meta-analysis

Since the different experimental articles or the reviews published on the role of regulatory T cells with respect to cancer prognosis could not reach a consensus, we deemed that integrating additional parameters could help resolve the inconsistency of the results. In particular, and echoing with the context-dependant plasticity and heterogeneity of these cells, we decided to describe the context simultaneously with the effect of Tregs.

We collected several parameters that could be useful to describe the context: the treatment, the technique used to quantify Tregs, the markers used to delineate the population as well as additional proteins detected in Tregs, and the cells associated negatively or positively to Tregs. We first attempted at drawing a clearer picture of the markers used to delineate Tregs in cancer studies. We then used 3 parameters to evaluate the importance of the context interplay with Tregs' role in cancer prognosis: the markers used to define Tregs or Tregs subsets, the tissue in which Tregs were quantified and whether the quantification was raw or a ratio of Tregs to another immune population. We also selected 5 cancer types, according to their consensus on Tregs' prognostic role: rather positively linked to prognosis (gastric and colorectal cancers), rather negatively linked (breast and Non Small Cell Lung cancers) or unclear (ovarian cancer). For each parameter, we evaluated whether taking it into account would improve the consensus, and whether we could outline precise conclusions regarding specific subsets or tissues for example.

It proved useful as it helped improve the consensus of the different studies included in the meta-analysis, and provided guidelines on how to better understand Treg prognostic role. Namely, the activated subset ($CD45RO^+$ Tregs) was consistently negatively linked with cancer prognosis, for all cancer types. This analytical strategy can be replicated for other immune cells for which their role is ambiguous, such as Th17 cells.

7.2 Article

Treg in cancer: a meta-analysis to define their prognostic role in a context-dependent manner

Elise Amblard[1,2] **and Vassili Soumelis**[1,2,3]

***For correspondence:**
elise.amblard@inserm.fr (EA)

[1]Université de Paris, Inserm U976, F-75006 Paris, France; [2]IRSL, F-75010 Paris, France; [3]Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Saint-Louis, Laboratoire d'Immunologie, F-75010, Paris, France

Abstract Assessing cancer prognosis is a challenging task, given the heterogeneity of the disease. Multiple features (clinical, environmental, genetic) have been used to serve this purpose. The Tumor Immune MicroEnvironment (TIME) is one of those key features, and describing the impact of TIME numerous components on cancer prognosis is an active field of research. Using the human TIME to assess prognosis is difficult, given the complexity of the context within the tumor micro-environment, with the example of regulatory T cells (Tregs). Tregs have a seemingly ambiguous prognostic role, characterized as negative, positive or neutral across studies. Focusing on five different cancer types (breast, colorectal, gastric, lung and ovarian cancers), we clarified how to define Tregs and use them to assess cancer prognosis by taking into account the context through the following parameters: the Treg subset, their anatomical location, and their neighboring cells. With the meta-analysis of these three parameters, we were able to clarify Tregs clinical role by recontextualizing them: we could delineate the fact that CD45RA$^-$ Tregs had a reproducible negative effect on prognosis across cancer types, and also better understand the meaning of the anatomical location of Tregs as well as neighboring cells on deciphering their prognostic value. Thus, we made a contribution to the question whether Tregs' role depends on the cancer type or not by favoring the pan-cancer answer. Additionally, we set up guidelines to improve the design of future studies addressing the physiopathological role of Tregs in cancer.

Introduction

In the last decades, cancer research has been strongly focusing on immunity in order to disentangle the links spanning the Tumor Immune MicroEnvironment (TIME), to resolve the mutual interactions and understand why immune cells fail to eradicate malignant tumors. In particular, the field of immunotherapy has been growing fast, with the goal of boosting the immune system in fighting cancer cells. There is evidence showing that the TIME plays a key role in predicting clinical evolution in humans, from tumorigenesis [Hanahan and Weinberg, 2011, Bissell and Hines, 2011] to global prognosis [Hsu et al., 2010, Tosolini et al., 2011], risk of metastasis [Olkhanud et al., 2011, Toh et al., 2011], and response to treatment [Binnewies et al., 2018]. Some TIME features, such as the Immunoscore, are used in the clinic and at the bench to classify tumors [Galon et al., 2012, Thorsson et al., 2018] for many different cancer types: prostate, breast, lung, colorectal, melanoma, among others.

Many reviews are summarizing the prognostic role of immune cells from the TIME or the peripheral circulation in different cancer types [Fridman et al., 2012, 2017, Ahrends and Borst, 2018]. In

Fridman et al. [2012], they are summarizing the role of the following immune cell subsets: cytotoxic $CD8^+$ cells, T helper $CD4^+$ cells as well as regulatory T cells (Tregs). For each of the different populations, the authors list the articles establishing a link between immune cell type and cancer prognosis (positive, negative or neutral) in patients. It appears that Tregs have a very versatile and the most ambiguous prognostic role: depending on the cancer type and on the study, Treg is either a good or bad prognostic factor, or has no impact.

The role of Tregs is indeed highly complex: it contributes to the maintenance of peripheral tolerance and suppress auto-immunity and inflammation, but is also preventing anti-tumor immunity. It is a population characterized by its strong heterogeneity.

One source of heterogeneity is their origin. A portion of Tregs originates from the thymus and is released into the peripheral circulation: the thymic or natural Tregs (nTregs) [Curotto de Lafaille and Lafaille, 2009]. The remaining population of Tregs develops after stimulation of naive peripheral $CD4^+$ T cells under specific conditions of antigen exposure and co-stimulation and is described as peripheral or induced Tregs (pTregs or iTregs). In humans, nTregs exhibit a demethylation of the Treg-specific demethylated region (TSDR) of the Foxp3 promoter which leads to a very stable expression of FOXP3; freshly differentiated iTregs and pTregs, on the other hand, are methylated at the TSDR [Mohr et al., 2018], explaining a more plastic phenotype compared to nTreg and a more volatile commitment to the regulatory lineage, although chronically stimulated iTregs are also demethylated at the TSDR. In the context of cancer, the claim is that the majority of tumor-infiltrating Tregs are mostly pTregs, diverted to a regulatory phenotype by the local microenvironment [Xydia et al., 2021].

A second source of heterogeneity lies in the variety of suppressing mechanisms: Tregs might exert their functions on Antigen Presenting Cells (APCs) or other T cells, in a contact-dependent or -independent manner. When targeting APCs, Tregs are able to induce them to be poor antigen presenters. This modulation of APC phenotype leads further to $CD4^+$ T cells developing a regulatory phenotype and impaired response of $CD8^+$ T cells. Tregs are also able to interact directly with effector T cells, via various mechanisms: they can either kill T cells by releasing perforins or granzymes [Grossman et al., 2004], impair their functions by releasing inhibitory cytokines such as IL-10, IL-35 [Liu et al., 2011], and TGF-β [Nakamura et al., 2001], or perturb their metabolism. Each of these mechanisms is elicited by context-specific cues, thus triggering a myriad of modes of action for regulation, and expanding further Tregs' diversity.

Tregs can also be dissected into subsets with a higher or smaller potential for suppression, whose power is eventually targeted towards specific cell populations such as Th1, Th2, Th17, Th22, etc.

In the context of cancer, Tregs are key players since they might modulate host response to the tumor, as well as host response to therapies: they can either be a crucial target for therapy or jeopardize or improve the response to treatments directed toward other cell types.

In the present meta-analysis, we aimed at clarifying the role for Tregs with regard to human cancer prognosis, by adopting a context-dependent approach to the problem. To the best of our knowledge, the context was never put on center stage in any of the past studies, while the tumor itself is usually extensively considered: tumor site, type, or stage for example [Shang et al., 2015]. Also, the fact that Tregs are often poorly defined [Frydrychowicz et al., 2017, Whiteside, 2014, 2019] is usually overlooked. Most of the articles investigating the prognostic role of Tregs would study either the size of the Treg compartment, comparing its variations at different stages of the disease or against healthy controls, or its ratio to another given cell population (immune or not), and neglect the diversity of Tregs. Here, we wanted to exploit different parameters, which relate to Tregs context, to better understand their role in prognosis. Namely, we fetched, whenever possible, the following parameters: i. the markers used for Treg definition, ii. the anatomical location, iii. the technique used to identify Tregs, iv. the study of cells in the same local environment as Tregs, that we called neighboring cells.

First, we checked if there is a consensus in the cancer literature on which Treg markers to use. Then we investigated the link between Tregs, their context and cancer prognosis, to see whether

we could improve the consensus on the prognostic role of Tregs and thanks to which parameter.

Results

Treg definition is fuzzy in human cancer literature

While parsing the human Treg literature, we found that there was no strong consensus on the markers used to define Tregs. In mouse, FOXP3 is an unequivocal marker for Tregs, whereas human FOXP3 is also expressed transiently by effector cells or by ex-Tregs [Allan et al., 2007, Sharma et al., 2013, Wang et al., 2007], and some Treg subsets do not express it at all [Otsubo et al., 2011, Boldt et al., 2014] or at low levels [Miyara et al., 2009]. Historically, the definition of Tregs is functional: they were first described as T cells that were regulating immunity by exerting suppression, thus even including the possibility of $CD8^+$ Tregs [Endharti et al., 2005, Chaput et al., 2009]. But this functional description comprises heterogeneous sub-populations: there is both an ontogenic and a phenotypic heterogeneity, linked to a plethoric diversity in functions [Miyara et al., 2009].

In our analysis, we focused on human studies and in order to span the whole spectrum of Treg prognostic values, we decided to study five cancer types: Tregs from breast and lung cancers (Non Small Cell Lung Cancer, NSCLC) are negatively associated with the clinical outcome, while it is the opposite for gastric and colorectal cancers, and there is seemingly no gain in using Tregs for prognostic application in ovarian cancer Fridman et al. [2017]. We decided to focus only on $CD4^+$ Tregs, since there was not enough material to investigate $CD8^+$ Tregs' role in cancer: we found only three articles about $CD8^+$ Tregs [Arruvito et al., 2014, Waniczek et al., 2017, Chakraborty et al., 2017] and one about $CD4^+CD8^+$ Tregs [Sarrabayrouse et al., 2014] among the 341 articles that we analyzed (second red stage in Figure 1).

To evaluate the strength of the consensus for Treg definition, we pooled all articles for the five cancer types mentioned above, for a total of 129 publications (n=23, 27, 24, 35 and 20 articles for breast, colorectal, gastric, lung, and ovarian cancers, respectively, see Methods for article selection, Supplementary Table 1 and Figure 1).

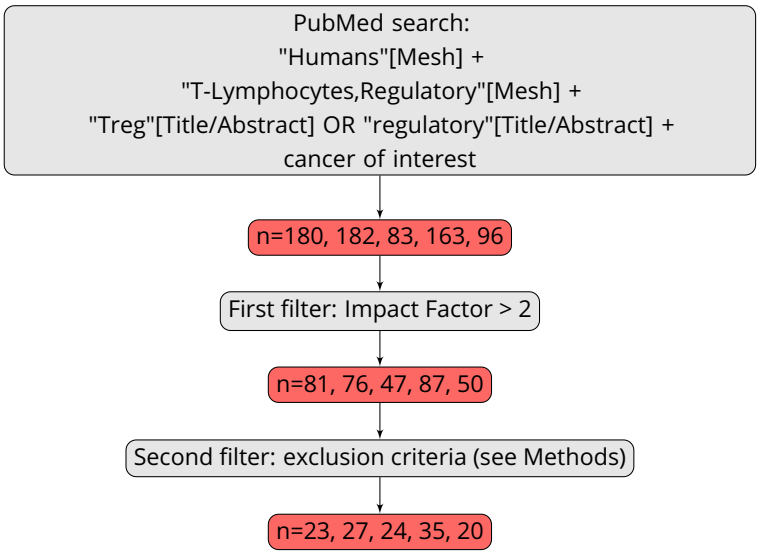


Figure 1. Article selection strategy. n represented the numbers of articles retained at each step, listed as follow: breast, colorectal, gastric, lung and ovarian cancers.

Aiming at outlining markers for the regulatory population, we removed the articles that were focusing on subsets only, to end up with a total of N=112 articles studying the whole Treg population. Regarding Treg detection methods, we distinguished two methods: either Fluorescence-Activated Cell Sorting (FACS) for 60% of the publications or ImmunoHistoChemistry (IHC) and whether it was

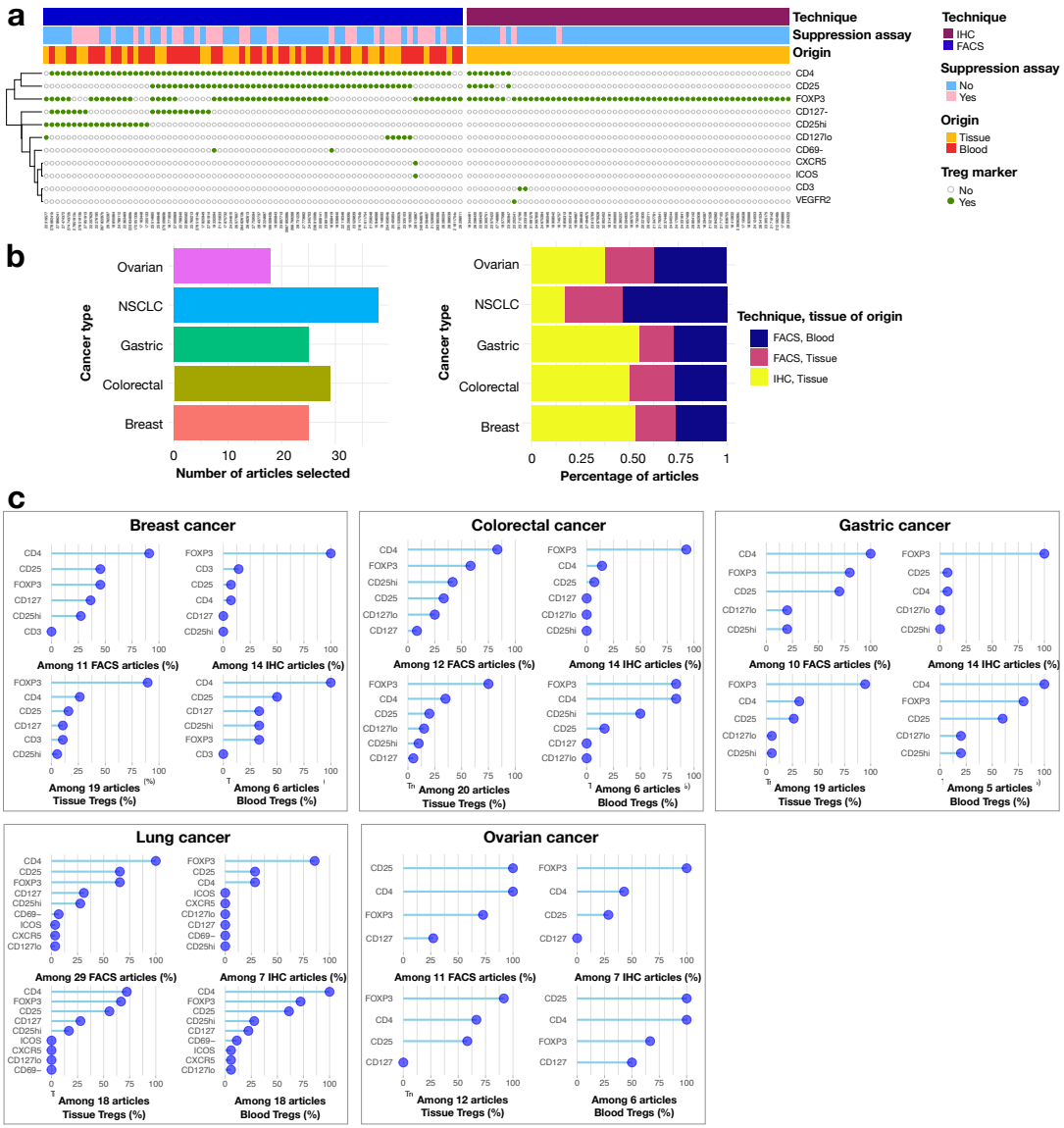


Figure 2. Treg markers used to identify Tregs in the cancer literature. **(a)** The heatmap draws a better comprehension of the commonly used markers and combinations of them for Treg definition, based on 112 articles from the cancer literature. Each row stands for a marker and each column for an article. The heatmap on the left recapitulates markers used in FACS along with a clustering on the markers in order to recapitulate common co-occurrences. The right heatmap summarizes markers used in IHC. It also displays whether the suppressing capacities of Tregs were tested in a functional assay, and from which tissue the Tregs come from. **(b)** Histogram of number of articles included per cancer type (left), normalized histogram per cancer type, with the color referring to the technique used for Treg detection and the tissue of origin (right). **(c)** Lollipop graphs depicting the magnitude of use of each Treg marker, depending on cancer type, technique for Treg detection and tissue of origin.

coupled with a working suppression assay (Figure 2a). Respectively 41% and 5% of the studies based on FACS and IHC performed a conclusive suppression assay. CD4 and CD25 were routinely used markers for FACS studies, as well as CD127 and the transcription factor FOXP3.Furthermore, the combination of CD4 and CD25 was very common. Regarding other markers, such as CXCR5 or CD69, their use was more anecdotal, and they were always used in combination to one or more of the classical markers CD4, CD25 and FOXP3. Some studies applied thresholds on the expression

of certain markers: e.g. 29% of the FACS articles in which Tregs were detected with CD25, used $CD25^{high}$ instead of $CD25^{+}$, and 25% used $CD127^{low}$ instead of $CD127^{-}$. For IHC, the ubiquitous marker was FOXP3, with a sparse use of CD4 and CD25. The only study that did not use FOXP3 to delineate Tregs, but CD4 and CD25, did so as the authors were investigating the expression of FOXP3 not only in Tregs but also in cancer cells [Kim et al., 2013]. The use of multiple markers is also notably rare for IHC, due to technical challenges.

We consecutively studied individual cancer types. Keeping in mind that we could not collect the same amount of articles for each cancer (Figure 2b, left panel), we first noted that within each cancer type, there was an equal proportion of articles using FACS or IHC for Treg detection. The only exception was lung cancer, since a larger fraction of related articles relied on blood samples, i.e. using FACS only (Figure 2b, right panel). Across cancer types and for IHC, the consensus was on a generalized use of FOXP3 staining. Regarding FACS, the diversity of markers was broader, with some studies using a high number of markers: e.g. the lung and colorectal studies, using respectively 9 and 6 different markers across studies. In fact, there was a strong correlation between the number of FACS articles and the number of markers used per cancer type (R^2=0.94, p-value=0.005, Pearson correlation, Supplementary Figure 1). Interestingly, we observed that FOXP3 was not commonly used to define Tregs with FACS in breast cancer, as compared to the other four cancer types: it was used in less than half of the articles. Lastly, there was no strong difference in Treg definition between blood and tissue Tregs (Figure 2c).

To conclude on the phenotypic definition of Treg in cancer studies, the consensus was clear for IHC-stained Tregs, while it was more blurred for FACS studies. FOXP3 is a popular marker, as well as the combinations of CD4/CD25 or CD4/CD25/FOXP3, with a conclusive use of a threshold, either on the expression of CD25 or CD127.

Context-dependent prognostic role of Tregs

This meta-analysis focused on the importance of the context, in terms of neighboring cells and anatomical location, and aimed at deciphering the contextual prognostic effect of Tregs in cancer. We studied independently 3 context-related parameters: the investigated Treg population, the anatomical location, and the Treg quantification method. We first evaluated whether we could improve the consensus on the prognostic role of Treg, considering successively the information brought by each of these three parameters, and each cancer type separately. Then, we tried to establish whether a higher granularity, considering a peculiar parameter, could lead to more reproducible conclusions on the link between Tregs and cancer across cancer types. Of note, we included in our meta-analysis n=3996, 6040, 2015, 2359, and 1754 patients from the combined studies for breast, colorectal, gastric, lung and ovarian cancers respectively.

Treg prognostic role depends on the Treg population

The lack of a harmonized combination of markers to delineate Tregs increased the difficulty in drawing conclusions about their role in cancer and their prognostic impact. A second pitfall comes from the fact that some studies only looked at specific regulatory sub-populations. We decided to evaluate how much the study of Treg subsets (Figure 3a) could help in increasing consensus on Treg prognostic role, stratifying the data on cancer types. Depending on the cancer type, we did not have the same extent of subsets' studies: for the ovarian cancer, we could not find any article focusing on subsets (Figure 3b), so we did not investigate this further. To measure the consensus on prognosis between different studies, we used the normalized Shannon entropy, or rather one minus the entropy, and the Fleiss' kappa, in order to quantify the agreement. To do so, we used a three-step approach: i. we considered all studies for one cancer type simultaneously, ii. we considered only the studies claiming to investigate the whole Treg population and iii. we computed the Shannon entropy and the Fleiss' kappa separately for each Treg subset and calculated the weighted mean, the weights being the number of patients. In the calculation of this mean, the whole Treg population was counted as a subset. The agreement coefficient ranges between 0 and

1, and is close to 1 if there is a consensus, or close to 0 if there is not. Our results show that there is no clear trend whether we considered all studies together, or whether we only took the studies using the global Treg population. But there is a clear increase of consensus if we consider the mean of the entropy for each subset (Figure 3c, Supplementary Figure 2). This implies that looking only at Treg subsets might explicitly improve the link between Tregs and cancer prognosis.

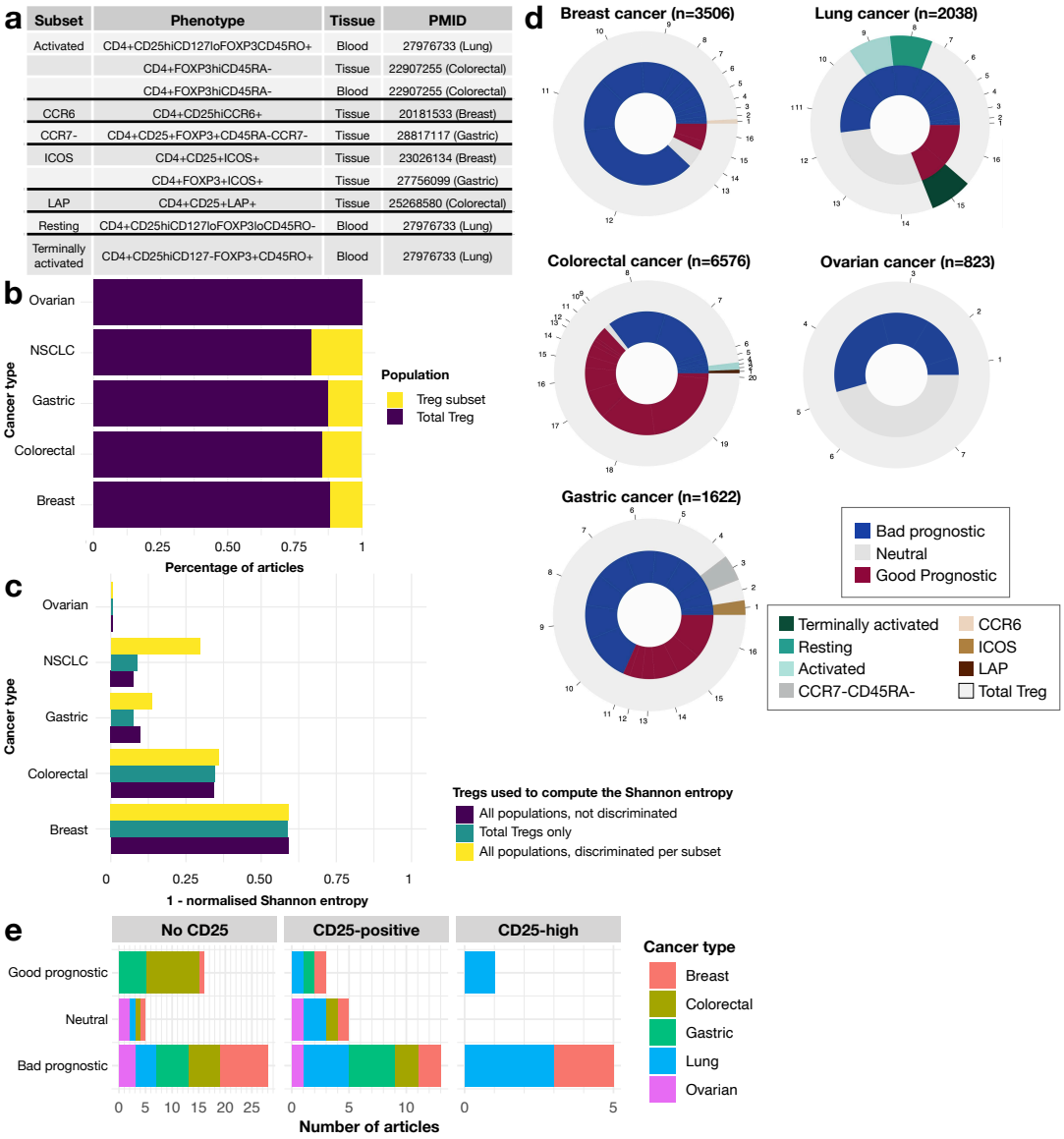


Figure 3. Subset diversity in cancer Tregs. **(a)** Summary of the various Treg subsets studied in the cancer literature used in this meta-analysis. **(b)** Frequency of articles studying either Treg as a whole or a specific subset. **(c)** One minus the normalized Shannon entropy for each of the five cancer types, and for each case: all articles together (dark blue), only the articles looking at the whole Treg population (dark green), mean of the entropies calculated for each population type (yellow). **d)** Pie chart of the prognostic impact of Tregs, as a function of the type of population used in the analysis, for each cancer type. Each numbered portion is an article and its size reflects the number of patients included in the study. **e)** Barplot showing the prognostic evaluation of Tregs from articles using either $CD25^{high}$ (right), $CD25^{+}$ (middle), or no CD25 (left) to delineate Tregs.

Interestingly, none of the articles looking at Treg subsets found a neutral role for Tregs in cancer prognosis. Also, all subsets with an activated or similar phenotype, as well as the resting population

(Figure 3a) were all negatively linked to the prognosis, independently of the cancer type (Figure 3d). However, subsets-focused studies represented a small fraction of the global cohort size in each cancer type (<1%, 2%, 7%, 23% for breast, colorectal, gastric and lung cancer respectively). This negative link was observed even in colorectal cancer, for which the consensual claim is that Tregs have a positive impact on the clinical outcome (Figure 3d). On the other end, the terminally activated regulatory fraction (Figure 3a) was of good prognosis, but it was studied in only one lung cancer publication (Figure 3d).We also explored in this meta-analysis the effect of the widely used $CD25^{high}$ marker, considering i. articles using $CD25^{high}$ (n=6), versus ii. articles using the mere positivity of CD25 (n=21), or iii. no CD25 (n=49). The rationale behind this exploration comes from the hypothesis that $CD25^{high}$ is a reliable marker of regulatory cells, as it eliminates contaminating activated CD4 helper cells [Saito et al., 2016]. Out of the six articles using the $CD25^{high}$ marker, five of them negatively linked Tregs to cancer prognosis (Figure 3e). The single article depicting a positive link considered terminally activated Tregs from the blood [Kotsakis et al., 2016], that we already described above as a sub-population of good prognosis (Figure 3d). An interesting, although less striking observation is about the $CD25^{+}$ population, that displayed a slightly stronger consensus (1−Shannon entropy=0.166) than the Tregs not delineated with CD25 (1−Shannon entropy=0.164), albeit less than the $CD25^{high}$ fraction (1−Shannon entropy=0.35) (Supplementary Figure 3). This strongly suggests that the $CD25^{high}$ fraction is the one of interest, while also being reproducibly linked to a negative cancer prognosis.

Treg prognostic value is context-dependent as observed when analysing ratios to other cells of the local environment

Some articles looked into the correlation between Tregs and other cell populations from the same environment. More precisely, this approach represented 36%, 23%, 34%, 19% and 29% of breast, colorectal, gastric, lung and ovarian cancer papers respectively. $CD8^{+}$ T cells were always positively associated to Tregs, whatever the cancer type was, and most articles looking at neighboring cells looked into this association. The same positive correlation was true for tumor cells, $CD3^{+}$ cells, cancer-associated fibroblasts (CAFs), follicular helper T cells (Tfh) and pre-dendritic cells (DCs) across cancer types. Positive correlations with other cell types such as myeloid cells in general, myeloid derived suppressor cells (MDSCs) in particular, macrophages and tumor-associated macrophages (TAMs) were described in just one breast cancer and one ovarian cancer article. Natural killer cells (NKs) and Th17 cells were negatively correlated to Tregs in one and three lung cancer studies respectively, as well as $FOXP3^{+}$ tumor cells in one colorectal cancer article (Figure 4a).

To study the role of neighboring cells on the Tregs' impact on prognosis, we applied the same methodology as above, adding the information of whether the authors of each article used the absolute Treg quantification information or a ratio of Treg over another cell population. The ratio-focused studies encompassed 25% of the total cohort, but with different weights for each cancer type: 26%, 6%, 21%, 11% and 41% for breast, colorectal, gastric, lung and ovarian cancer respectively. Again, the consensus was better if we use ratios whenever evaluating the influence of Tregs on the prognosis (Figure 4b). This iss coherent since ratios would partially take into consideration other components of the local environment and thus better recapitulate the complexity of the local environment. Strikingly, some ratios always correlated with bad prognosis (Treg/$CD8^{+}$ T cells), while some others always correlated with good prognosis (Treg/Th17 cells), whereas there was no trend for the remaining ratios: Treg/$CD4^{+}$ T cells or Treg/T cells (Supplementary Figure 4).

Tumor tissue Tregs have a clearer prognostic role than blood Tregs

Lastly, we also studied the role of Tregs that were detected in patients' peripheral blood or directly in the organ at stake, and even from specific parts of the tumor. Since there was no common base to name the different parts of the tumor, we merged the different denominations used by the different authors: intra-epithelial (or nest) vs stroma, intra-tumoral vs peri-tumoral. Nest designated cells surrounded by cancer cells, while stroma designated cells from the tumor stroma, i.e. cellular

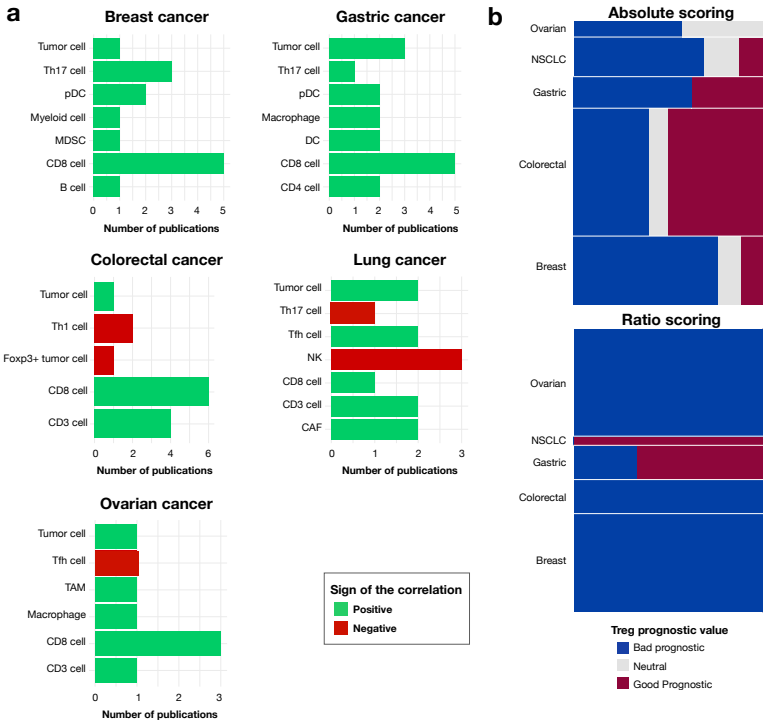


Figure 4. Interplay between Tregs and neighboring cells. **(a)** Correlation between Tregs and other cell populations, depending on each cancer type; number of articles depicting the different correlations. **(b)** Treemap of the prognostic value of Tregs depending on the quantification: either the absolute quantification (top panel) or quantification via a ratio scoring (considering all ratios Treg/neighboring cell) (bottom panel). The length of each bar represents the proportion of patients for each prognostic type per cancer type and the height of each rectangle relates to the proportion of patients from each cancer type compared to the patients from all cancer types (n=14,565 for the absolute quantification, n=3,653 for the ratio quantification).

patches almost free of cancer cells within the tumor. Peri-tumoral transparently meant cells at the margin, as opposed to intra-tumoral, referring to cells within the tumor center. We measured the consensus as described in the Methods section and we observed that the prognosis agreement per anatomical location was better than considering all anatomical locations together or even considering the undifferentiated tumor piece (Figure 5a). Overall, we observed that the anatomical location is a crucial parameter to better understand the role of the Treg population in each cancer type, since there is a strong decrease in entropy when we factored the location in the evaluation of the prognostic value (Figure 5a, Supplementary Figure 5). In particular, for lung cancer, the highest consensus is reached when taking into account only the tumor piece as a whole. This could be partly due to the fact that results based on blood Tregs showed the highest discrepancy in terms of prognostic conclusions and the majority of lung cancer studies are based on blood Treg detection. We suspect that blood results are particularly ambiguous since Tregs from blood samples are all delineated with FACS, for which the consensual markers are more blurred than with IHC. Additionally, we suspect that blood samples do not reflect the TIME as well as tissue samples, thus it does not recapitulate well the context.

For breast cancer, Triple Negative Breast Cancer was the only situation for which tissue Tregs were of good prognosis, while the neutral case is only met for an article with a very small cohort of patients (n=40) [Bailur et al., 2015], or when the authors considered peri-tumoral Tregs [Liu et al., 2012]. For the other cancer types, the interpretation was more cumbersome, as Tregs from different parts of the tissue also exhibited dual conclusions (Figure 5b). Regarding colorectal cancer, almost all anatomical locations exhibited simultaneously Tregs with a positive or negative link to cancer prognosis, depending on the article, except for the juxtatumoral site and blood Tregs. The

same is observed for tumoral or intra-tumoral Tregs from gastric cancer, and for blood and stromal Tregs from lung cancer (Figure 5b). However, all articles except one that concluded to a positive link between Tregs and cancer prognosis used IHC, and thus could not distinguish between the different Treg fractions that we described above. Furthermore, only one out of the fifteen articles used a ratio quantification.

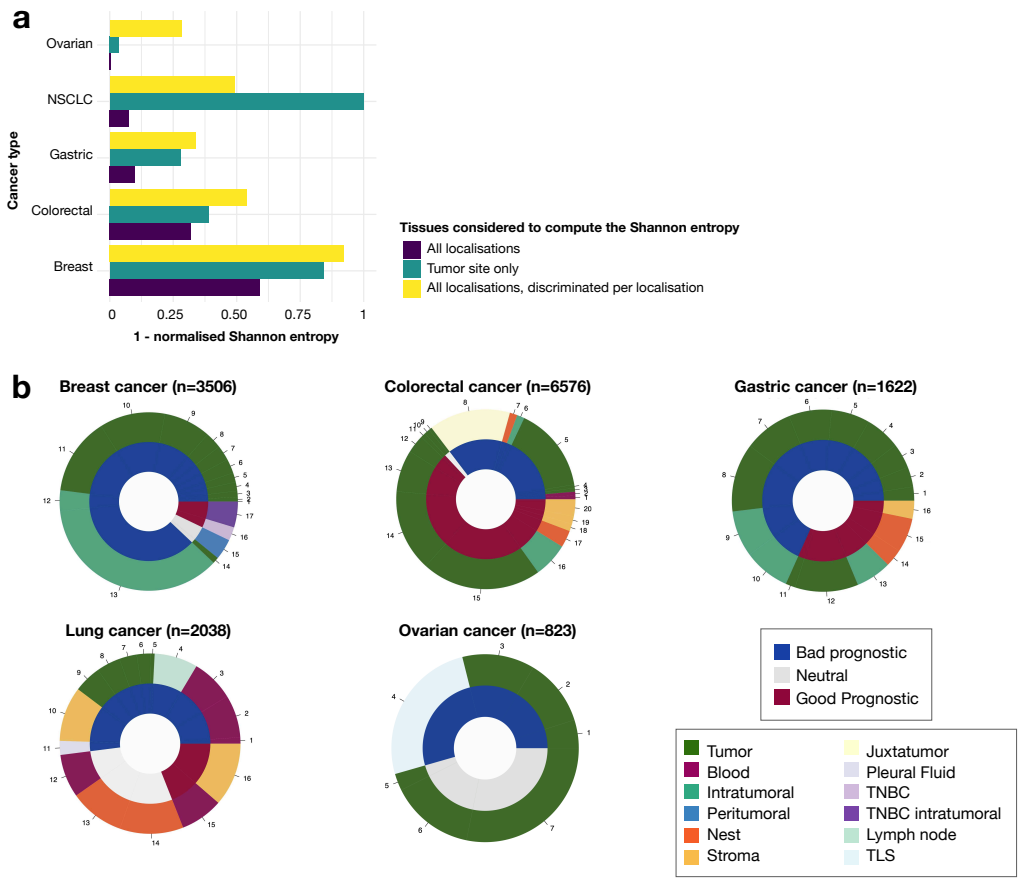


Figure 5. Interplay between Tregs, their anatomical locations and their prognostic use. **(a)** Histogram showing one minus the normalized Shannon entropy for each cancer type and for each group according to anatomical location: all locations together (dark blue), tumor site (dark green), mean of the entropies calculated for each anatomical location (yellow). **(b)** Pie charts of the prognostic value depending on the anatomical location for each cancer type. Each numbered portion is an article and its size reflects the number of patients included in the study. (TNBC: Triple Negative Breast Cancer; TLS: Tertiary Lymphoid Structure)

Discussion

The heterogeneity of regulatory phenotypes is a key parameter to explain, at least partially, the apparent discrepancy of the impact of Tregs in cancer prognosis. In this meta-analysis, we showed that explicit description of Tregs and their afferent context could help in better understanding their clinical role, as compared to considering only tumor characteristics such as site, stage, etc [Blatner et al., 2013, Fridman et al., 2012, 2017, Mao et al., 2016, Whiteside, 2014]. We took into account three different factors that could interfere with Treg physiopathological role: the different regulatory subsets studied to evaluate the prognosis, their quantification method and the different anatomical microenvironment, and we focused on five different cancer types: breast, colorectal, gastric, lung and ovarian cancers, collecting a total of 129 articles. The amount of information for each of those parameters was substantially different, hence we could not directly compare the

respective importance of these parameters, but we saw that considering them separately did increase the resolution on the link between Tregs and cancer prognosis, in most of the situations considered.

We could also draw reproducible links with respect to the prognosis in some particular cases: regarding the Treg subsets, the activated and resting sub-populations were always found to be of bad prognosis [Saito et al., 2016]. Regarding neighboring cells, we observed that the ratio Treg/$CD8^+$ T cells was also of bad prognosis, while Tregs were anti-correlated to $CD8^+$ T cells independently on the cancer type. It was interesting to observe that these conclusions proposed a reunited view of Tregs role across cancer types. Our methodology might serve for other cell types such as Th2 or Th17 cells, for which their role is not completely clear as well [Fridman et al., 2017], although to a lesser extent than Tregs.

We could not examine in this meta-analysis the cohorts' characteristics, foremostly because the details were very scarce and not standardized. Among the 64 articles that were informative for the prognosis, only 16 had the following minimal information: timeframe for the cohort's creation, median follow-up, mean (or median) age of the patients, sex balance, and tumor stage. Those clinical parameters are of paramount importance to decipher prognostic factors and we therefore suggest to improve the generation and management of meta-data, as a way to increase the quality of meta-analysis, but also the reproducibility.

We integrated three parameters in our analysis, but missed others parameters forming alternative contexts, such as treatment, because the information was too sparse, but an exciting perspective could be to consider this metadata as well [Jiménez-Sánchez et al., 2020, Hamy et al., 2019].

Another lead of investigation could be to cross-analyze all the information about treatment, cancer subtypes, anatomical location, definition of the global regulatory population and its subsets, and Treg quantification, to unravel an even clearer picture of Tregs' role in the TIME. An exciting way to answer these questions could be by taking advantage of -omics methods, to gather and cross-analyze even more information, for example relate to the inflammatory context. With the advent of -omics technologies, there is also hope that we could find core signature genes delineating the regulatory population. So far, it is not obvious though, since this signature remains dependent on the strategy used to capture the population of interest in the first place. Among three articles yielding gene lists of length 294, 136 and 31 respectively, the intersection contains only 10 genes, among which FOXP3, and CTLA-4 but not CD25 (IL2RA), that was not in the signature from Peranacker et al. [2016] although it was defined on $CD25^{high}$ regulatory cells (Supplementary Figure 6).

Our meta-analysis highlights how cancer Treg studies are designed differently, making a harmonic conclusion on Tregs' role in cancer prognosis very difficult. In light of our results, we suggest the following guidelines when studying Tregs in a cancer context: (i) focus on the quantification of $CD45RO^+FOXP3^{high}$ activated and $CD45RO^-FOXP3^{low}$ resting subsets, as well as on the $CD25^{high}$ fraction, (ii) quantify the Treg/CD8 ratio, (iii) carefully choose which sample to use (nest vs stroma or intra- vs peri-tumoral), (iv) comprehensively annotate clinical data. By following those steps, scientists and clinicians should be able to sketch a more plausible landscape of Tregs' clinical roles for any cancer type. A careful consideration of the same parameters should also be applied when using existing literature.

Methods

Articles selection

We searched consistently in PubMed for any article related to our topic, using the search words "Humans" [Mesh] AND "T-Lymphocytes, Regulatory" [Mesh] AND ("Treg" [Title/Abstract] OR "Tregs" [Title/Abstract] OR "regulatory T" [Title/Abstract]) and adding the neoplasms we were interested in, namely the breast, colorectal, gastric, lung and ovarian neoplasms, as a Mesh term. We added a second filter, only considering articles published in a journal with an impact factor above 2. We found a total of 81, 76, 47, 87 and 50 articles respectively for the breast, colorectal, gastric, lung

and ovarian cancers. Finally, we also added the following exclusion criteria: focus on other cells than human Tregs (mouse Tregs, regulatory B cells, $CD8^+$ Tregs), focus outside the primary tumor (metastasis, relapsed cancer, tumor lines, in-vitro systems), patients treated with immunotherapy or with a context of disimmunity, missing information (markers used, anatomical location, number of patients), review articles. We ended up with a total of 23, 27, 24, 35 and 20 articles per cancer type.

Agreement evaluation

We used the normalized Shannon entropy, and the Fleiss' kappa to evaluate the consensus. The Fleiss' kappa is calculated as:

$$\kappa = \frac{\sum_i N_i(N_i - 1)}{\sum_i N_i(\sum_i N_i - 1)},$$

where N_i is the number of raters opting for choice i. In our case, the choice is either $-1, 0$, or 1, for bad, neutral or good prognosis, and the number of raters stands for the number of patients. The Shannon entropy is calculated as:

$$1 - \mathrm{SE} = 1 + \frac{1}{\log_{10} \sum_\mathrm{j} 1} \sum_\mathrm{i} \frac{\mathrm{N_i}}{\sum_\mathrm{j} \mathrm{N_j}} \log_{10} \frac{\mathrm{N_i}}{\sum_\mathrm{j} \mathrm{N_j}}$$

Acknowledgments

We thank Cristina Ghirelli, Camille Chauvin, Lucile Massenet-Regad and Benoit Salomon for their careful proofreading of the manuscript. This work has been supported by Association Sciences et Technologie - Groupe de Recherche Servier.

References

T. Ahrends and J. Borst. The opposing roles of CD4 $^+$ T cells in anti-tumour immunity. *Immunology*, 154(4): 582–592, Aug. 2018. doi: 10.1111/imm.12941.

S. Allan, S. Crome, N. Crellin, L. Passerini, T. Steiner, R. Bacchetta, M. Roncarolo, and M. Levings. Activation-induced FOXP3 in human T effector cells does not suppress proliferation or cytokine production. *International Immunology*, 19:345–354, Feb. 2007. doi: 10.1093/intimm/dxm014.

L. Arruvito, F. Payaslián, P. Baz, A. Podhorzer, A. Billordo, J. Pandolfi, G. Semeniuk, E. Arribalzaga, and L. Fainboim. Identification and clinical relevance of naturally occurring human CD8+HLA-DR+ regulatory T cells. *Journal of Immunology*, 193(9):4469–4476, Nov. 2014. doi: 10.4049/jimmunol.1401490.

J. K. Bailur, B. Gueckel, E. Derhovanessian, and G. Pawelec. Presence of circulating Her2-reactive CD8+ T-cells is associated with lower frequencies of myeloid-derived suppressor cells and regulatory T cells, and better survival in older breast cancer patients. *Breast cancer research: BCR*, 17:34, Mar. 2015. doi: 10.1186/s13058-015-0541-z.

R. Bhairavabhotla, Y. C. Kim, D. D. Glass, T. M. Escobar, M. C. Patel, R. Zahr, C. K. Nguyen, G. K. Kilaru, S. A. Muljo, and E. M. Shevach. Transcriptome profiling of human FoxP3+ regulatory T cells. *Human Immunology*, 77(2): 201–213, Feb. 2016. doi: 10.1016/j.humimm.2015.12.004.

M. Binnewies, E. Roberts, K. Kersten, V. Chan, D. Fearon, M. Merad, L. Coussens, D. Gabrilovich, S. Ostrand-Rosenberg, C. Hedrick, R. Vonderheide, M. Pittet, R. Jain, W. Zou, T. K. Howcroft, E. Woodhouse, R. Weinberg, and M. Krummel. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature Medicine*, 24(5):541–550, May 2018. doi: 10.1038/s41591-018-0014-x.

M. Bissell and W. Hines. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nature Medicine*, 17(3):320–329, Mar. 2011. doi: 10.1038/nm.2328.

N. R. Blatner, F. Gounari, and K. Khazaie. The two faces of regulatory T cells in cancer. *Oncoimmunology*, 2(5): e23852, May 2013. ISSN 2162-4011. doi: 10.4161/onci.23852.

A. Boldt, K. Kentouche, S. Fricke, S. Borte, F. Kahlenberg, and U. Sack. Differences in FOXP3 and CD127 expression in Treg-like cells in patients with IPEX syndrome. *Clinical Immunology*, 153(1):109–111, July 2014. doi: 10.1016/j.clim.2014.04.001.

S. Chakraborty, A. K. Panda, S. Bose, D. Roy, K. Kajal, D. Guha, and G. Sa. Transcriptional regulation of FOXP3 requires integrated activation of both promoter and CNS regions in tumor-induced CD8+ Treg cells. *Scientific Reports*, 7(1):1628, May 2017. doi: 10.1038/s41598-017-01788-z.

N. Chaput, S. Louafi, A. Bardier, F. Charlotte, J.-C. Vaillant, F. Ménégaux, M. Rosenzwajg, F. Lemoine, D. Klatzmann, and J. Taieb. Identification of CD8+CD25+Foxp3+ suppressive T cells in colorectal cancer tissue. *Gut*, 58(4):520–529, Apr. 2009. doi: 10.1136/gut.2008.158824.

M. A. Curotto de Lafaille and J. J. Lafaille. Natural and adaptive foxp3+ regulatory T cells: more of the same or a division of labor? *Immunity*, 30(5):626–635, May 2009. ISSN 1097-4180. doi: 10.1016/j.immuni.2009.05.002.

A. T. Endharti, M. Rifa, Z. Shi, Y. Fukuoka, Y. Nakahara, Y. Kawamoto, K. Takeda, K.-i. Isobe, and H. Suzuki. Cutting Edge: CD8+CD122+ Regulatory T Cells Produce IL-10 to Suppress IFN-γ Production and Proliferation of CD8+ T Cells. *The Journal of Immunology*, 175(11):7093–7097, Dec. 2005. doi: 10.4049/jimmunol.175.11.7093.

W. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews. Cancer*, 12(4):298–306, 2012. doi: 10.1038/nrc3245.

W. Fridman, L. Zitvogel, C. Sautès-Fridman, and G. Kroemer. The immune contexture in cancer prognosis and treatment. *Nature Reviews. Clinical Oncology*, 14(12):717–734, Dec. 2017. doi: 10.1038/nrclinonc.2017.101.

M. Frydrychowicz, M. Boruczkowski, A. Kolecka-Bednarczyk, and G. Dworacki. The Dual Role of Treg in Cancer. *Scandinavian Journal of Immunology*, 86(6):436–443, Dec. 2017. doi: 10.1111/sji.12615.

J. Galon, F. Pagès, F. Marincola, H. Angell, M. Thurin, A. Lugli, I. Zlobec, A. Berger, C. Bifulco, G. Botti, F. Tatangelo, C. Britten, S. Kreiter, L. Chouchane, P. Delrio, H. Arndt, M. Asslaber, M. Maio, G. Masucci, M. Mihm, F. Vidal-Vanaclocha, J. Allison, S. Gnjatic, L. Hakansson, C. Huber, H. Singh-Jasuja, C. Ottensmeier, H. Zwierzina, L. Laghi, F. Grizzi, P. Ohashi, P. Shaw, B. Clarke, B. Wouters, Y. Kawakami, S. Hazama, K. Okuno, E. Wang, J. O'Donnell-Tormey, C. Lagorce, G. Pawelec, M. Nishimura, R. Hawkins, R. Lapointe, A. Lundqvist, S. Khleif, S. Ogino, P. Gibbs, P. Waring, N. Sato, T. Torigoe, K. Itoh, P. Patel, S. Shukla, R. Palmqvist, I. Nagtegaal, Y. Wang, C. D'Arrigo, S. Kopetz, F. Sinicrope, G. Trinchieri, T. Gajewski, P. Ascierto, and B. Fox. Cancer classification using the Immunoscore: a worldwide task force. *Journal of Translational Medicine*, 10(1):205, Dec. 2012. doi: 10.1186/1479-5876-10-205.

W. J. Grossman, J. W. Verbsky, W. Barchet, M. Colonna, J. P. Atkinson, and T. J. Ley. Human T regulatory cells can use the perforin pathway to cause autologous target cell death. *Immunity*, 21(4):589–601, Oct. 2004. doi: 10.1016/j.immuni.2004.09.002.

A.-S. Hamy, H. Bonsang-Kitzis, D. De Croze, E. Laas, L. Darrigues, L. Topciu, E. Menet, A. Vincent-Salomon, F. Lerebours, J.-Y. Pierga, E. Brain, J.-G. Feron, G. Benchimol, G.-T. Lam, M. Laé, and F. Reyal. Interaction between Molecular Subtypes and Stromal Immune Infiltration before and after Treatment in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 25(22):6731–6741, Nov. 2019. ISSN 1557-3265. doi: 10.1158/1078-0432.CCR-18-3017.

D. Hanahan and R. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, Mar. 2011. doi: 10.1016/j.cell.2011.02.013.

D. Hsu, M. Kim, B. Balakumaran, C. Acharya, C. Anders, T. Clay, H. K. Lyerly, C. Drake, M. Morse, and P. Febbo. Immune Signatures Predict Prognosis in Localized Cancer. *Cancer Investigation*, 28(7):765–773, July 2010. doi: 10.3109/07357900903095755.

B. Höllbacher, T. Duhen, S. Motley, M. M. Klicznik, I. K. Gratz, and D. J. Campbell. Transcriptomic Profiling of Human Effector and Regulatory T Cell Subsets Identifies Predictive Population Signatures. *ImmunoHorizons*, 4(10):585–596, Oct. 2020. doi: 10.4049/immunohorizons.2000037.

A. Jiménez-Sánchez, P. Cybulska, K. L. Mager, S. Koplev, O. Cast, D.-L. Couturier, D. Memon, P. Selenica, I. Nikolovski, Y. Mazaheri, Y. Bykov, F. C. Geyer, G. Macintyre, L. M. Gavarró, R. M. Drews, M. B. Gill, A. D. Papanastasiou, R. E. Sosa, R. A. Soslow, T. Walther, R. Shen, D. S. Chi, K. J. Park, T. Hollmann, J. S. Reis-Filho, F. Markowetz, P. Beltrao, H. A. Vargas, D. Zamarin, J. D. Brenton, A. Snyder, B. Weigelt, E. Sala, and M. L. Miller. Unraveling tumor–immune heterogeneity in advanced ovarian cancer uncovers immunogenic effect of chemotherapy. *Nature Genetics*, 52(6):582–593, June 2020. doi: 10.1038/s41588-020-0630-5.

M. Kim, T. Grimmig, M. Grimm, M. Lazariotou, E. Meier, A. Rosenwald, I. Tsaur, R. Blaheta, U. Heemann, C.-T. Germer, A. M. Waaga-Gasser, and M. Gasser. Expression of Foxp3 in colorectal cancer but not in Treg cells correlates with disease progression in patients with colorectal cancer. *PloS One*, 8(1):e53630, 2013. doi: 10.1371/journal.pone.0053630.

A. Kotsakis, F. Koinis, A. Katsarou, M. Gioulbasani, D. Aggouraki, N. Kentepozidis, V. Georgoulias, and E.-K. Vetsika. Prognostic value of circulating regulatory T cell subsets in untreated non-small cell lung cancer patients. *Scientific Reports*, 6:39247, Dec. 2016. doi: 10.1038/srep39247.

F. Liu, F. Tong, Y. He, and H. Liu. Detectable expression of IL-35 in CD4+ T cells from peripheral blood of chronic Hepatitis B patients. *Clinical Immunology*, 139(1):1–5, Apr. 2011. doi: 10.1016/j.clim.2010.12.012.

F. Liu, Y. Li, M. Ren, X. Zhang, X. Guo, R. Lang, F. Gu, and L. Fu. Peritumoral FOXP3$^+$ regulatory T cell is sensitive to chemotherapy while intratumoral FOXP3$^+$ regulatory T cell is prognostic predictor of breast cancer patients. *Breast Cancer Research and Treatment*, 135(2):459–467, Sept. 2012. doi: 10.1007/s10549-012-2132-3.

Y. Mao, Q. Qu, X. Chen, O. Huang, J. Wu, and K. Shen. The Prognostic Value of Tumor-Infiltrating Lymphocytes in Breast Cancer: A Systematic Review and Meta-Analysis. *PLOS ONE*, 11(4):e0152500, Apr. 2016. doi: 10.1371/journal.pone.0152500.

M. Miyara, Y. Yoshioka, A. Kitoh, T. Shima, K. Wing, A. Niwa, C. Parizot, C. Taflin, T. Heike, D. Valeyre, A. Mathian, T. Nakahata, T. Yamaguchi, T. Nomura, M. Ono, Z. Amoura, G. Gorochov, and S. Sakaguchi. Functional delineation and differentiation dynamics of human CD4+ T cells expressing the FoxP3 transcription factor. *Immunity*, 30(6):899–911, June 2009. doi: 10.1016/j.immuni.2009.03.019.

A. Mohr, R. Malhotra, G. Mayer, G. Gorochov, and M. Miyara. Human FOXP3+ T regulatory cell heterogeneity. *Clinical & Translational Immunology*, 7, Jan. 2018. doi: 10.1002/cti2.1005.

K. Nakamura, A. Kitani, and W. Strober. Cell contact-dependent immunosuppression by CD4(+)CD25(+) regulatory T cells is mediated by cell surface-bound transforming growth factor beta. *The Journal of Experimental Medicine*, 194(5):629–644, Sept. 2001. doi: 10.1084/jem.194.5.629.

P. Olkhanud, B. Damdinsuren, M. Bodogai, R. Gress, R. Sen, K. Wejksza, E. Malchinkhuu, R. Wersto, and A. Biragyn. Tumor-Evoked Regulatory B Cells Promote Breast Cancer Metastasis by Converting Resting CD4+ T Cells to T-Regulatory Cells. *Cancer Research*, 71(10):3505–3515, May 2011. doi: 10.1158/0008-5472.CAN-10-4316.

K. Otsubo, H. Kanegane, Y. Kamachi, I. Kobayashi, I. Tsuge, M. Imaizumi, Y. Sasahara, A. Hayakawa, K. Nozu, K. Iijima, S. Ito, R. Horikawa, Y. Nagai, K. Takatsu, H. Mori, H. D. Ochs, and T. Miyawaki. Identification of FOXP3-negative regulatory T-like (CD4+CD25+CD127low) cells in patients with immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome. *Clinical Immunology*, 141(1):111–120, Oct. 2011. doi: 10.1016/j.clim.2011.06.006.

A. M. Pesenacker, A. Wang, A. Singh, J. Gillies, Y. Kim, C. Piccirillo, D. Nguyen, N. Haining, S. Tebbutt, C. Panagiotopoulos, and M. K. Levings. A Regulatory T-Cell Gene Signature Is a Specific and Sensitive Biomarker to Identify Children with New-Onset Type 1 Diabetes. *Diabetes*, 65(4), Apr. 2016. doi: 10.2337/db15-0572.

T. Saito, H. Nishikawa, H. Wada, Y. Nagano, D. Sugiyama, K. Atarashi, Y. Maeda, M. Hamaguchi, N. Ohkura, E. Sato, H. Nagase, J. Nishimura, H. Yamamoto, S. Takiguchi, T. Tanoue, W. Suda, H. Morita, M. Hattori, K. Honda, M. Mori, Y. Doki, and S. Sakaguchi. Two FOXP3(+)CD4(+) T cell subpopulations distinctly control the prognosis of colorectal cancers. *Nature Medicine*, 22(6):679–684, June 2016. doi: 10.1038/nm.4086.

G. Sarrabayrouse, C. Bossard, J.-M. Chauvin, A. Jarry, G. Meurette, E. Quévrain, C. Bridonneau, L. Preisser, K. Asehnoune, N. Labarrière, F. Altare, H. Sokol, and F. Jotereau. CD4CD8$\alpha\alpha$ lymphocytes, a novel human regulatory T cell subset induced by colonic bacteria and deficient in patients with inflammatory bowel disease. *PLoS biology*, 12(4):e1001833, Apr. 2014. doi: 10.1371/journal.pbio.1001833.

B. Shang, Y. Liu, S.-j. Jiang, and Y. Liu. Prognostic value of tumor-infiltrating FoxP3 + regulatory T cells in cancers: a systematic review and meta-analysis. *Scientific Reports*, 5(1):15179, Oct. 2015. doi: 10.1038/srep15179.

M. D. Sharma, L. Huang, J.-H. Choi, E.-J. Lee, J. M. Wilson, H. Lemos, F. Pan, B. R. Blazar, D. M. Pardoll, A. L. Mellor, H. Shi, and D. H. Munn. An inherently bi-functional subset of Foxp3+ T helper cells is controlled by the transcription factor Eos. *Immunity*, 38(5):998–1012, May 2013. doi: 10.1016/j.immuni.2013.01.013.

V. Thorsson, D. Gibbs, S. Brown, D. Wolf, D. Bortone, T.-H. Ou Yang, E. Porta-Pardo, G. Gao, C. Plaisier, J. Eddy, E. Ziv, A. Culhane, E. Paull, I. A. Sivakumar, A. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J. Parker, L. Mose, N. S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S. Reynolds, R. Bowlby, A. Califano, A. Cherniack, D. Anastassiou, D. Bedognetti, Y. Mokrab, A. Newman, A. Rao, K. Chen, A. Krasnitz, H. Hu, and T. e. a. Malta. The Immune Landscape of Cancer. *Immunity*, 48(4):812–830.e14, Apr. 2018. doi: 10.1016/j.immuni.2018.03.023.

B. Toh, X. Wang, J. Keeble, W. J. Sim, K. Khoo, W.-C. Wong, M. Kato, A. Prevost-Blondel, J.-P. Thiery, and J.-P. Abastado. Mesenchymal Transition and Dissemination of Cancer Cells Is Driven by Myeloid-Derived Suppressor Cells Infiltrating the Primary Tumor. *PLoS Biology*, 9(9):e1001162, Sept. 2011. doi: 10.1371/journal.pbio.1001162.

M. Tosolini, A. Kirilovsky, B. Mlecnik, T. Fredriksen, S. Mauger, G. Bindea, A. Berger, P. Bruneval, W.-H. Fridman, F. Pagès, and J. Galon. Clinical Impact of Different Classes of Infiltrating T Cytotoxic and Helper Cells (Th1, Th2, Treg, Th17) in Patients with Colorectal Cancer. *Cancer Research*, 71(4):1263–1271, Feb. 2011. doi: 10.1158/0008-5472.CAN-10-2907.

J. Wang, A. Ioan-Facsinay, E. I. H. van der Voort, T. W. J. Huizinga, and R. E. M. Toes. Transient expression of FOXP3 in human activated nonregulatory CD4+ T cells. *European Journal of Immunology*, 37(1):129–138, Jan. 2007. doi: 10.1002/eji.200636435.

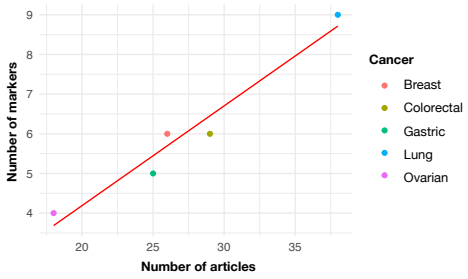
D. Waniczek, Z. Lorenc, M. Śnietura, M. Wesecki, A. Kopec, and M. Muc-Wierzgoń. Tumor-Associated Macrophages and Regulatory T Cells Infiltration and the Clinical Outcome in Colorectal Cancer. *Archivum Immunologiae Et Therapiae Experimentalis*, 65(5):445–454, Oct. 2017. doi: 10.1007/s00005-017-0463-9.

T. Whiteside. Regulatory T cell subsets in human cancer: are they regulating for or against tumor progression? *Cancer immunology, immunotherapy: CII*, 63(1):67–72, Jan. 2014. doi: 10.1007/s00262-013-1490-y.

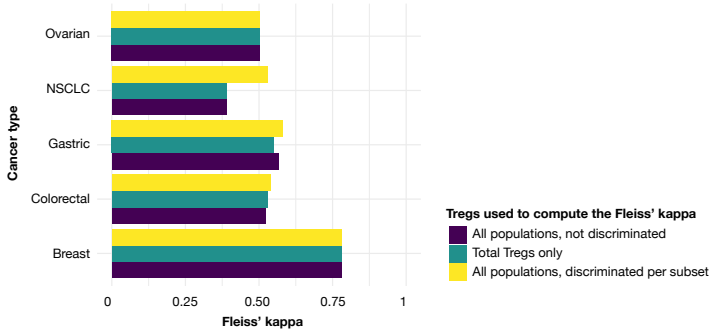
T. Whiteside. Human regulatory t cells (treg) and their response to cancer. *Expert Review of Precision Medicine and Drug Development*, 4:1–14, 07 2019. doi: 10.1080/23808993.2019.1634471.

M. Xydia, R. Rahbari, E. Ruggiero, I. Macaulay, M. Tarabichi, R. Lohmayer, S. Wilkening, T. Michels, D. Brown, S. Vanuytven, S. Mastitskaya, S. Laidlaw, N. Grabe, M. Pritsch, R. Fronza, K. Hexel, S. Schmitt, M. Müller-Steinhardt, N. Halama, C. Domschke, M. Schmidt, C. von Kalle, F. Schütz, T. Voet, and P. Beckhove. Common clonal origin of conventional T cells and induced regulatory T cells in breast cancer patients. *Nature Communications*, 12(1):1119, Feb. 2021. doi: 10.1038/s41467-021-21297-y.

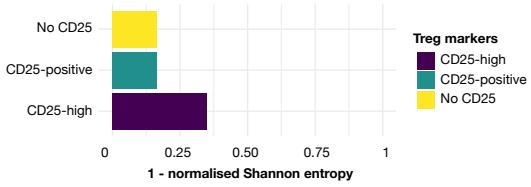
Supplementary Figures



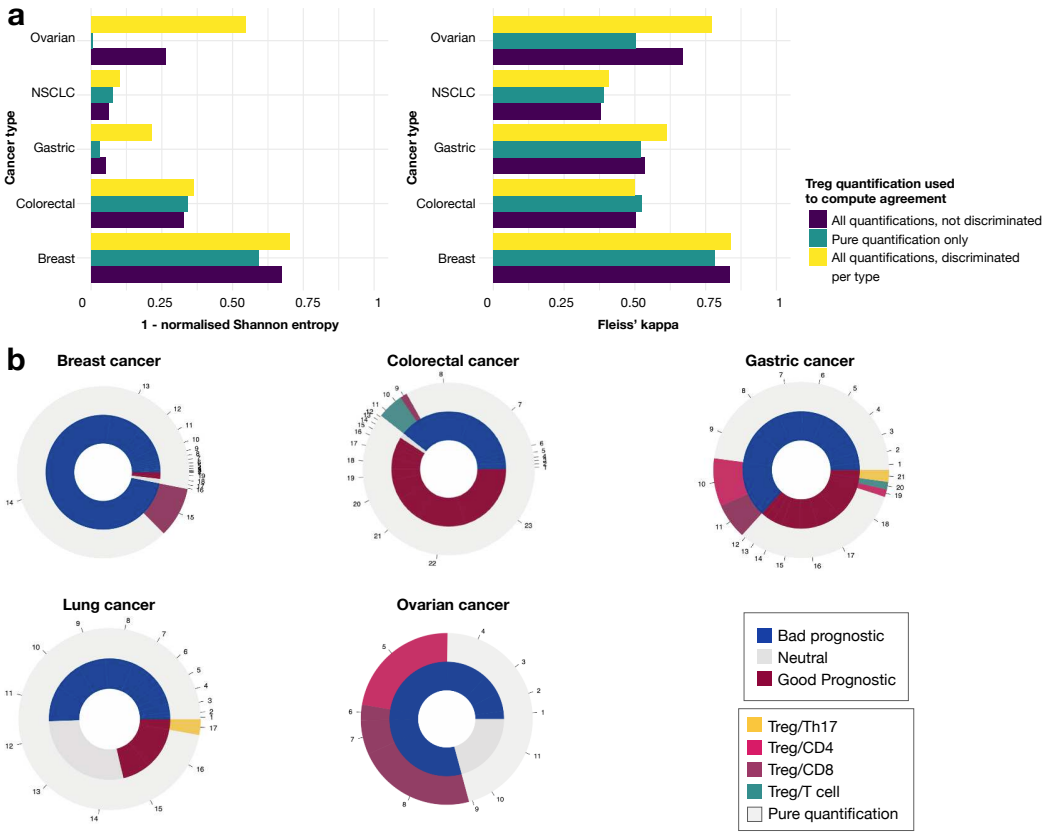
Supplementary Figure 1. Correlation between the number of markers used in each cancer type, and the number of articles used.



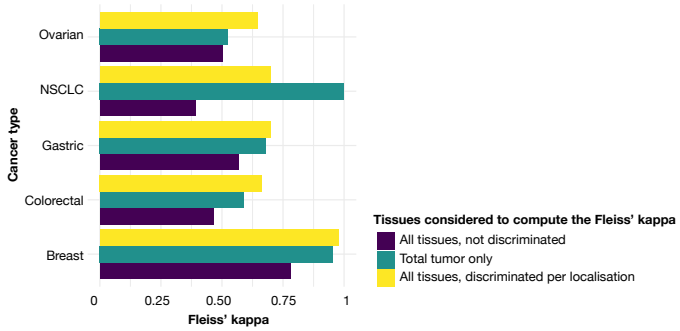
Supplementary Figure 2. Fleiss' kappa to determine the agreement between articles, stratified per regulatory subset.



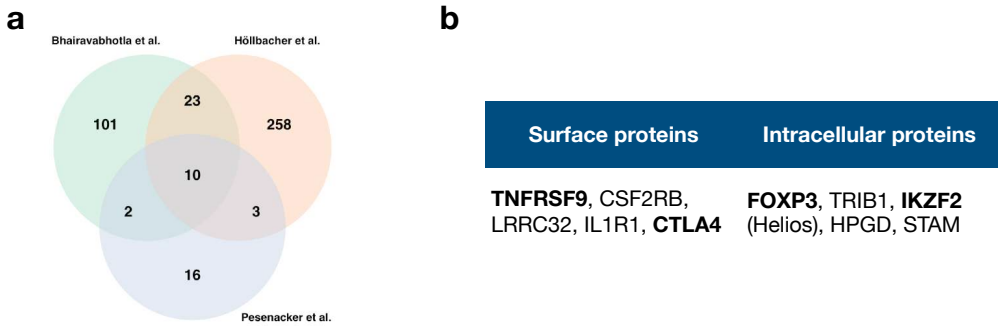
Supplementary Figure 3. Shannon entropy for the articles depending on the use of the marker CD25.



Supplementary Figure 4. Treg quantification's impact on its prognostic value. **(a)** Agreement between articles, stratified per quantification type: Shannon entropy (left), Fleiss' kappa (right). **(b)** Pie chart of the prognosis with the quantification information per cancer type. Each portion is an article and its size reflects the number of patients included.



Supplementary Figure 5. Fleiss' kappa to determine the agreement between articles, stratified per anatomical location.



Supplementary Figure 6. Overlap between transcriptomic regulatory signatures. **(a)** Venn diagramm of the gene signature for Tregs from Bhairavabhotla et al. [2016], Höllbacher et al. [2020], Pesenacker et al. [2016]. **(b)** Genes at the intersection of the three signatures. Genes in bold are recursively linked to the regulatory phenotype.

Cancer	Good prognostic	Neutral prognostic	Bad prognostic	Undetermined
Breast	28233108, 23075422	22842982, 25849846	24124553, 18820666, 17135638, 27851913, 23529839, 20181533, 23836289, 18413832, 23026134, 21717105, 22842982, 21521526, 27566250, 24562936	28388539, 17135638, 22760213, 27851913, 23529839, 20181533, 22836755, 19855964, 22116346, 23712790, 18294387
Colorectal	20386463, 19064967, 24005418, 19856313, 24675384, 24997850	23382847, 16740757	25268580, 26298011, 19064967, 21915633, 24005418, 22907255, 19577568, 19908042, 31681276	17205133, 22276195, 22319577, 23382847, 23613769, 25268580, 24064667, 20952660, 24005418, 22907255, 27851914, 22207629, 18985040, 25405854
Gastric	23807713, 24331841, 24170095, 21347781, 19732435, 32204925, 266799288		29804142, 28817117, 27756099, 24657498, 24040244, 22374482, 22083420, 21792941, 20221835, 19900843, 19153062	29804142, 28817117, 26782287, 24261990, 24170095, 22083420, 21528082, 21347781, 20422211, 19900843, 19153062, 18224687, 18087278
Lung	22300751, 23335103, 279767333	22300751, 23305175, 23891508	17099880, 20234320, 21719142, 22363469, 22608141, 23305175, 23891508, 27773662, 27851914, 279767331, 279767332	15846066, 16698419 17163448, 17825949, 18771959, 19148592, 19332094, 19597336, 21258248, 21611754, 21663645, 22363469, 24345703, 24780112, 26042578, 26280204, 26541534, 27000869, 27474372, 27866241, 28513867, 28731226
Ovarian		26077607, 20006900, 18314181	27748885, 26298430, 25365237, 24244610, 17875732, 16344461, 32902402	28437737, 27759594, 26482613, 25514665, 25416072, 23948613, 22865582, 22798340, 18166500, 18036640

Supplementary Table 1. List of PMID references used in this analysis

Part III

General discussion and perspectives

Chapter 8

General considerations and prospects

Rien dans la vie n'est à craindre, tout doit être compris. C'est maintenant le moment de comprendre davantage, afin de craindre moins.

Marie Skłodowska-Curie

When I started my thesis, I was fairly new to basically all the fields I worked on, and as such ready to be challenged and to challenge. I have been curious about many aspects of the areas I touched on, ranging from single-cell RNAseq data analysis *per se* to T cell biology or oncology. Now on the verge of concluding this work, I chose to focus specifically on some itching questions that I encountered along the past years.

8.1 On technical aspects

8.1.1 The hubness project

Conclusion For the last two years, I have been working on the hubness phenomenon. We first started to delineate the questions we wanted to explore: evaluate the magnitude of the phenomenon in sequencing data and the parameters influencing it, the nature of hub cells, and the effect of hubness on scRNAseq analysis. We used existing metrics to probe hubness in our data. While the current body of evidence indicates that hubness correlates to the dimension [166] and the local densities distribution [207], we additionally linked it to sparsity and intrinsic dimension of the data in sequencing count matrices. We also proposed a new method to uncover hub cells, since existing methods would not perform adequately in our data because of a whopping number of anti-hubs. The nature of those hub cells was seemingly similar to other "normal" cells or anti-hubs, which led us to think that hub cells were pure dimensional artefacts.

Since "hubby" datasets are those with a high intrinsic dimension, we verified that they would be the ones more impacted by hubness correction: for the clustering, trajectory inference and visualisation tasks, it was indeed the case, with higher performances upon hub correction. For low-intrinsic-dimension datasets, hubness correction could be beneficial or detrimental, with an overall performance improvement.

Discussion We explored the effects of hubness on scRNAseq data and its analysis. Regarding the study of the raw magnitude of hubness, we investigated leads to explain why there is a difference in magnitude with the sparsity and the intrinsic dimension, but it is only a partial explanation. We did not explore the local density lead which is challenging, but it could be worth it. I believe that it would be interesting to characterize further the nature of hubs. Regarding the study of the effect of hubness on scRNAseq analysis, we concluded that hubness correction was beneficial for datasets characterized by high intrinsic dimensions. We should explore further the biological impact of hubness correction on high-dimensional datasets, as well as the effect of hubness correction on low-intrinsic-dimension datasets, and why there is no direct correlation between the magnitude of hubness and the clustering performance, while it is the case for the visualisation task. Finally, we observed that the 4 hub correction methods that we used had dissimilar effects on the improvement of the clustering, trajectory inference and its stability, and visualisation. We suggested hypotheses to clarify it that should be tested, regarding the fact that one method uses all data points to correct for hubness, while the others are local.

Perspectives I believe that the main interest of studying hubness (apart from the fact that it proved useful in our benchmark) resides in the fact that we propose to tackle directly hubness instead of avoiding the curse-of-dimensionality related effect by reducing the dimension. More generally, the dimension reduction step in the analysis is critical and has been challenged extensively [126, 130, 149, 186, 208, 209]. The most common approach is to do a PCA, but there is already a question about the choice of the number of PCs to retain. The concern is to eliminate noise and redundant information while retaining signal. I feel that the current options are not sound, nor intellectually satisfactory, although an interesting option to separate the noise from the signal is to implement RMT [172]. Our hubness study is easy to implement, and it increases the possibility of retaining more signal, by retaining a higher number of dimensions and simultaneously alleviating dimensional noise. It opens also the way to a set of new strategies, that would directly target the effects of the curse of dimensionality instead of merely reducing the dimension [169]. However, it complicates the analysis by adding more steps to the current pipelines and we could further clarify the biological utility of hubness correction on datasets without ground truth.

8.1.2 Validity and reproducibility of dry lab experiments

Discussion There is a high emphasis put on the publication of precise methods by wet lab researchers (for example Cell Press implemented a new section, STAR★Methods, in order to ease the replication process[1]). The same effort is now done from the side of dry lab researchers, with many journals asking for the code. In the same vein as what is nowadays the norm for wet experiments, dry researchers should give access not only to the raw code but also to the session information with packages versions and to the hardware used. While it is necessary in order to reproduce results for all articles, it is of utmost importance when publishing a new method. Indeed, the aim of publishing the code is two-fold: it should increase the reproducibility of the experiments, but also enable to verify and test the code. For the hubness project, although we did not develop a tool *per se*, we ought to be as precise as possible in order enforce the validity of the benchmark. I worked hand in hand with Jonathan Bac, exchanging and proofreading

[1] https://www.cell.com/star-methods

bits of code. Since we used published packages (like Scanpy [126] or scikit-hubness [210]), we did not go through the tedious process of proofreading the functions that we used, although we got at least one example of a function that would not perform exactly like what was indicated in its documentation.

Perspectives In the present context of an increasing mistrust in science, and because there has been an ongoing replication crisis, researchers have a duty to make science more accessible. One way to achieve it is to be as transparent as possible, and this applies to the bioinformatic community as well. I think that we should implement a strict process in order to verify our scripts, since code writing is highly prone to errors. Ideally, systematic proofreading and tests should be mandatory. The minimal requirement in my opinion would be at least proofreading by a third party. There is an increasing concern about false data, or poor quality data in wet labs, but we should also feel concerned by the phenomenon in dry analyses and results. There are already few examples of research softwares that contained errors [211], while it could be alleviated with a rigorous process of code production. This is crucial, as false results might have devastating consequences such as the suspicion about vaccination due to a fraudulent 1988 Lancet article. Since a famous science motto is to doubt everything, we should also reasonably doubt our methods. After all, "the most important thing is to never stop questioning" (A. Einstein).

8.1.3 Accumulate data or create new tools?

Discussion During my thesis, I used mostly public scRNAseq datasets, as well as published tools dedicated to the analysis of scRNAseq data such as Seurat [125], scEntropy [204] or Scanpy [126]. I was stunned by the amount of objects, be it count matrices or tools, and I thought it was quite paradoxical. Without even discussing the technical progress leading to ever larger and more qualitative datasets, data is accumulating at a fast pace, while we are still developing tools in order to analyse the data already produced and extract as much information as possible from it. The fast production of objects is embodied by burgeoning databases that try to ease the navigation of users in the sea of objects: PanglaoDB[2], CancerSEA[3], scRNASeqDB[4], our own ImmuCANscDB[5] (see Annex 10), the scRNA-tools website[6], etc. Even with the help of databases, I believe that it makes the scRNAseq field hard to apprehend.

Perspectives As a bioinformatician, I would favor the option of tool's instead of data's accumulation for two main reasons. First, I think that there is still a lack, maybe not of tools, but at least of a strong consensus on the best pipeline to analyse the data, although one clear thing is that it is almost impossible to develop a pipeline that would work best for all datasets. But we do not even have, in most cases, heuristics to choose the tools suited to a given analysis, with the exception of platforms such as dynverse[7] [212]. In fact, neutral extensive benchmarks could be a solution to create these heuristics [174]. The second problem is more ethical: in a world with limited resources, accumulating data has

[2] https://panglaodb.se/
[3] http://biocc.hrbmu.edu.cn/CancerSEA/goBrowse
[4] https://bioinfo.uth.edu/scrnaseqdb/
[5] https://immucanscdb.vital-it.ch/
[6] https://www.scrna-tools.org/
[7] https://dynverse.org/

a cost, in terms of energy and rare materials, and the community should be aware of this issue and consider rationalizing the production of data, by checking beforehand whether the dataset they need is not already available. There is a massive amount of data that is underexploited, while I praise parsimony in science.

8.1.4 Creating scRNAseq-based knowledge

Discussion I raised above the problem of underexploited data. I faced many times during my PhD the problem of how to handle the data and extract the most out of it, with, running through it, the burning question of the utility of scRNAseq. In other words, I was bewildered, wondering how we could create knowledge while relying on previous one. An illustrative example of this problem is the benchmark. I discussed in the introduction the two ways of benchmarking tools, either with supervised or unsupervised scores. Regarding the unsupervised scoring metrics, I already mentioned their flaws. I am focusing here on the supervised scoring metrics, using ground truth labels. Ground truth labels make sense if we make the assumption that their truth matches the transcriptomic truth. If this is the case, then (i) the benchmark is valid, but (ii) the utility of scRNAseq diminishes because we already know the truth. If this is not the case, then there is a wealth of information but we cannot hardly access it, even less translate it into real life, because our benchmarks would not be validated for this situation. I often had the image of a snake biting its own tail when I thought about this inconsistency in the method.

This is one of the reasons why we attempted to develop a function-based classification of T cells that did not creates knowledge *per se*, but used what is already known to take advantage of scRNAseq data.

Perspectives Answering the question of the practical utility of scRNAseq is beyond my reach but I have few hints. The main one is that omics data speed up discoveries, because of the humongous amount of data that is creates: for example we can uncover new cell types or states [205], and even question the distinction between cellular states and types with the continuous approach of trajectory inference [213], discover new markers [72, 73], access to expression profiles (see Annex 9), etc. However, as I evoked it several times earlier in this manuscript, every discovery made with scRNAseq should be backed up by wet experiments [205, 206].

8.2 On biological aspects

8.2.1 The functional project

Discussion As I mentioned above, an answer to puzzling scRNAseq interpretation could be supervised analyses. While we showed extensively the added value of the functional modules analysis, we still need to improve technical aspects of the methodology. However, we could already interpret onco-immune datasets, by showing that the tumor tissue exhibits a higher number of functions than the dormant juxtatumor, including a higher expression of immune suppression and attraction functions. The increased functionality of tumor-resident T cells ties in with the idea that we described in the introduction of an increasing plasticity of T cells in the TME. Additionally, we observed that functional patterns were more conserved across different cancer types in the tumor, compared to

the juxtatumor. This functional approach has been extended to the study of other types of immune cells and other pathological contexts, such as dendritic cells in the blood of Covid-19 patients (see Annex 9), validating further the proof of concept.

Perspectives I believe this methodology could help to functionally map human tissues, in physiological or pathological conditions. Since it is supervised, the interpretation step can be automatised, speeding up and easing the analysis of scRNAseq data. However, it implies the following assumption: we consider that the detection of the RNA of an effector gene stands for the expression of the corresponding protein. This assumption has been challenged by the fact that there might be a decorrelation between protein and RNA expression profiles [214]. Again, this stresses the crucial need for a validation step at the bench [205], or at least for a confirmation of the protein expression. This will probably be enabled by the emergence of multi-omics techniques, such as CITE-seq [215].

Furthermore, since our tool is also usable at the scale of the cell (although we did not exploit this possibility in the present manuscript), it will help unlocking the full potential of scRNAseq. This makes even more sense with regard to the current trend of personalized medicine: clinicians and researchers are interested, not only in characterizing the differences between patients, but also within a patient, and this will be doable at the cell level.

8.2.2 Regulatory T cells in cancer

Discussion I have had a particular focus on regulatory T cells during my doctoral work, trying to understand their lability in cancer. While I first hypothetized that we could relate their prognostic value in cancer to whether the cancer was in a localization prone to inflammation, such as colon, rectum or head & neck, I soon realised that it was more intricate and decided that I should consider the global picture and take into account the complexity of the TME, in order to better describe their plastic role. So I tried to decipher whether taking the context into account would help to understand the prognostic value of Tregs in cancer. From the meta-analysis that I conducted, I would be tempted to say that it does. I studied 3 parameters (the localization of Tregs, the markers used to delineate the population and the quantification method) and showed that the consensus over the different articles that I included was increased upon considering the diversity in these parameters. To go even further, I believe that this kind of analysis could be done for other cells such as Th17 cells.

Perspectives Yet this meta-analysis eludes several questions. First I selected only 3 parameters, circumventing other sources of plasticity and divergence, such as the treatment. Second, I focused on $CD4^+$ regulatory cells, although I showed earlier that there is a blurring between the different lineages, and other cells might exert suppressing functions. Third, it would be interesting to try to model all cells at the same time in order to better recapitulate the complexity of the TME, that I described in the introduction. At least, it could be worth cross-analysing the 3 parameters that I used, although it is a challenging task given the heterogeneity of the studies that I included in the meta-analysis. In fact, it raises the issue of data production, storage and transmission. This is particularly critical in cancer studies, since it is a fast-moving field. We should not add heterogeneity of the data to the heterogeneity of the disease, of the diagnostic, of the patient care. Treg study

in cancer is an archetypical example, potentiated by the difficulty of capturing those cells. I think that firmer conclusions on the role of Tregs, and immune cells in cancer initiation, progression, metastasis, regression and response to treatment will require more technical progress, initiated by the advent of multi-omics, proteomics, and spatial sequencing, that I hope should help to resolve TME heterogeneity [216].

8.2.3 Implementing scRNAseq in onco-immunology, from the bench to the bedside

Immune therapies have recently been added to the cancer therapeutic arsenal, for example to treat melanoma [217]. Yet some patients would respond successfully, some would relapse on the contrary, and the factors governing success or failure have not been elucidated to this day. It exemplifies the blatant need to personalize treatment, but also to a better understanding of the tumor biology in order to orient personalized medicine. Obviously, single-cell omics technologies should help to meet the need of describing the TIME and cancer cells with a higher granularity [218].

So far, scRNAseq is used for research purposes only, and even bulk RNAseq is anecdotal, being used to spot specific mutations in target therapies in a few cancer centers. New results are gathered regarding the use of scRNAseq data from cancer patients (trials NCT04352777, NCT0406159), and several hospitals started to accumulate scRNAseq expression profiles, but it will take more years before it can be used at the bedside, because it remains an expensive and delicate technique, because biomarkers needs to be further tailored [69, 218], and because we still lack gold-standard for the analysis of the data.

Bibliography

[1] Kenneth Murphy & Casey Weaver. *Janeway's Immunobiology (9th edition)* (New York: Garland Science/Taylor & Francis Group, 2017).

[2] Habs, H. Epidemiologisch verwertbare Einzelangaben bei Thukydides. In *Die sogenannte Pest des Thukydides*, vol. 6 of *Sitzungsberichte der Heidelberger Akademie der Wissenschaften*, 21–23 (Springer, Berlin, Heidelberg, 1982).

[3] Boylston, A. The origins of inoculation. *Journal of the Royal Society of Medicine* **105**, 309–313 (2012).

[4] Voltaire. Sur l'insertion de la petite vérole (Lettre XI). In *Lettres philosophiques* (Paris, 1734).

[5] Pasteur, L. Méthode pour prévenir la rage après morsure. In *Comptes rendus hebdomadaires des séances de l'Académie des sciences* (Paris, 1885).

[6] Robert Koch. Die Ätiologie der Tuberkulose. *Berliner Klinische Wochenschrift* **15** (1882).

[7] Behring, E. v. & Kitasato, S. Über das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. *Deutsche Medizinische Wochenschrift* **49** (1890).

[8] Ehrlich, P. On immunity with special reference to cell life, the Croonian lecture. In *Proceedings of the Royal Society*, vol. 66 (London : Harrison and Sons, 1900).

[9] Ehrlich, P. Die Schutzstoffe des Blutes. *Deutsche Medizinische Wochenschrift* **27**, 913–916 (1901).

[10] Greenberg, S. A Concise History of Immunology (2003).

[11] Murphy, J. B. The lymphocyte in resistance to tissue grafting, malignant disease, and tuberculous infection. *Monographs of the Rockefeller Institute for Medical Research* **21** (1926).

[12] Gowans, J. The recirculation of lymphocytes from blood to lymph in the rat. *Journal of Physiology* **146**, 54–69 (1959).

[13] Miller, J. F. & Mitchell, G. F. The thymus and the precursors of antigen reactive cells. *Nature* **216**, 659–663 (1967).

[14] Alberts, B. *et al.* Helper T Cells and Lymphocyte Activation. *Molecular Biology of the Cell. 4th edition* (2002).

[15] Gagliani, N. & Huber, S. Basic Aspects of T Helper Cell Differentiation. *T-Cell Differentiation* 19–30 (2017).

[16] Mousset, C. M. *et al.* Comprehensive Phenotyping of T Cells Using Flow Cytometry. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **95**, 647–654 (2019).

[17] Rasoulouniriana, D. *et al.* A distinct subset of FcγRI-expressing Th1 cells exert antibody-mediated cytotoxic activity. *The Journal of Clinical Investigation* **129**, 4151–4164 (2019).

[18] Marshall, N. B. & Swain, S. L. Cytotoxic CD4 T Cells in Antiviral Immunity. *Journal of Biomedicine and Biotechnology* **2011** (2011).

[19] Tada, T., Takemori, T., Okumura, K., Nonaka, M. & Tokuhisa, T. Two distinct types of helper T cells involved in the secondary antibody response: independent and synergistic effects of Ia- and Ia+ helper T cells. *The Journal of Experimental Medicine* **147**, 446–458 (1978).

[20] Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A. & Coffman, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *Journal of Immunology* **136**, 2348–2357 (1986).

[21] Yamamura, M. *et al.* Defining protective responses to pathogens: cytokine profiles in leprosy lesions. *Science (New York, N.Y.)* **254**, 277–279 (1991).

[22] Hirahara, K. & Nakayama, T. CD4+ T-cell subsets in inflammatory diseases: beyond the Th1/Th2 paradigm. *International Immunology* **28**, 163–171 (2016).

[23] O'Shea, J. J., Lahesmaa, R., Vahedi, G., Laurence, A. & Kanno, Y. Genomic views of STAT function in CD4+ T helper cell differentiation. *Nature Reviews Immunology* **11**, 239–250 (2011).

[24] Elo, L. L. *et al.* Genome-wide Profiling of Interleukin-4 and STAT6 Transcription Factor Regulation of Human Th2 Cell Programming. *Immunity* **32**, 852–862 (2010).

[25] Aggarwal, S., Ghilardi, N., Xie, M.-H., de Sauvage, F. J. & Gurney, A. L. Interleukin-23 promotes a distinct CD4 T cell activation state characterized by the production of interleukin-17. *The Journal of Biological Chemistry* **278**, 1910–1914 (2003).

[26] Kaplan, M. H., Hufford, M. M. & Olson, M. R. The Development and in vivo function of TH9 cells. *Nature reviews. Immunology* **15**, 295–307 (2015).

[27] Jia, L. & Wu, C. The biology and functions of Th22 cells. *Advances in Experimental Medicine and Biology* **841**, 209–230 (2014).

[28] Signorini, V. *et al.* One year in review 2020: systemic lupus erythematosus. *Clinical and Experimental Rheumatology* **38**, 592–601 (2020).

[29] Thomas, Y. *et al.* Functional analysis of human T cell subsets defined by monoclonal antibodies. VI. Distinct and opposing immunoregulatory functions within the OKT8+ population. *The Journal of molecular and cellular immunology: JMCI* **1**, 103–113 (1984).

[30] Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *Journal of Immunology* **155**, 1151–1164 (1995).

[31] Shevach, E. M. & Thornton, A. M. tTregs, pTregs, and iTregs: similarities and differences. *Immunological Reviews* **259**, 88–102 (2014).

[32] Mohr, A., Malhotra, R., Mayer, G., Gorochov, G. & Miyara, M. Human FOXP3+ T regulatory cell heterogeneity. *Clinical & Translational Immunology* **7** (2018).

[33] Cretney, E., Kallies, A. & Nutt, S. L. Differentiation and function of Foxp3+ effector regulatory T cells. *Trends in Immunology* **34**, 74–80 (2013).

[34] Duhen, T., Duhen, R., Lanzavecchia, A., Sallusto, F. & Campbell, D. J. Functionally distinct subsets of human FOXP3+ Treg cells that phenotypically mirror effector Th cells. *Blood* **119**, 4430–4440 (2012).

[35] Bailey-Bucktrout, S. L. & Bluestone, J. A. Regulatory T cells: stability revisited. *Trends in immunology* **32**, 301–306 (2011).

[36] Rubtsov, Y. P. *et al.* Stability of the regulatory T cell lineage in vivo. *Science (New York, N.Y.)* **329**, 1667–1671 (2010).

[37] Guo, H. *et al.* Stability and inhibitory function of Treg cells under inflammatory conditions in vitro. *Experimental and Therapeutic Medicine* **18**, 2443–2450 (2019).

[38] Khosravi, M. *et al.* Induction of CD4+CD25+FOXP3+ regulatory T cells by mesenchymal stem cells is associated with modulation of ubiquitination factors and TSDR demethylation. *Stem Cell Research & Therapy* **9**, 273 (2018).

[39] Hua, J. *et al.* Pathological conversion of regulatory T cells is associated with loss of allotolerance. *Scientific Reports* **8**, 7059 (2018).

[40] O'Shea, J. J. & Paul, W. E. Mechanisms underlying lineage commitment and plasticity of helper CD4+ T cells. *Science* **327**, 1098–1102 (2010).

[41] Lee, Y. K. *et al.* Late Developmental Plasticity in the T Helper 17 Lineage. *Immunity* **30**, 92–107 (2009).

[42] Nakayamada, S., Takahashi, H., Kanno, Y. & O'Shea, J. J. Helper T cell diversity and plasticity. *Current Opinion in Immunology* **24**, 297–302 (2012).

[43] Lexberg, M. H. *et al.* Th memory for interleukin-17 expression is stable in vivo. *European Journal of Immunology* **38**, 2654–2664 (2008).

[44] Abou-Jaoudé, W. *et al.* Model checking to assess T-helper cell plasticity. *Frontiers in Bioengineering and Biotechnology* **2**, 86 (2014).

[45] Hirahara, K. *et al.* Mechanisms underlying helper T-cell plasticity: implications for immune-mediated disease. *The Journal of Allergy and Clinical Immunology* **131**, 1276–1287 (2013).

[46] Huber, S., Gagliani, N., O'Connor, W., Geginat, J. & Caprioli, F. CD4+ T Helper Cell Plasticity in Infection, Inflammation, and Autoimmunity. *Mediators of Inflammation* **2017**, 7083153 (2017).

[47] Carbo, A. *et al.* Computational modeling of heterogeneity and function of CD4+ T cells. *Frontiers in Cell and Developmental Biology* **2** (2014).

[48] Kunicki, M. A., Hernandez, L. C. A., Davis, K. L., Bacchetta, R. & Roncarolo, M.-G. Identity and Diversity of Human Peripheral Th and T Regulatory Cells Defined by Single-Cell Mass Cytometry. *The Journal of Immunology* **200**, 336–346 (2018).

[49] Breasted, J. *The Edwin Smith Surgical Papyrus, Volume 1: Hieroglyphic Transliteration, Translation, and Commentary.* Oriental Institute Publications (Chicago: The University of Chicago Press, 1931).

[50] Deeley, T. A brief history of cancer. *Clinical Radiology* **34**, 597–608 (1983).

[51] Faguet, G. B. A brief history of cancer: Age-old milestones underlying our current knowledge database. *International Journal of Cancer* **136**, 2022–2036 (2015).

[52] Rossier, L. Le cancer dans l'Antiquité. *Universitas* **2**, 16–19 (2011).

[53] André-Julien Fabre. Le cancer dans l'Antiquité: Les enseignements de Celse. *Histoire des sciences médicales* **42**, 63–70 (2008).

[54] Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).

[55] Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

[56] David, A. R. & Zimmerman, M. R. Cancer: an old disease, a new disease or something in between? *Nature Reviews Cancer* **10**, 728–733 (2010).

[57] Shoval, O. *et al.* Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science* **336**, 1157–1160 (2012).

[58] Schwartz, J. L. *et al.* X-Ray and cis-Diamminedichloroplatinum(II) Cross-Resistance in Human Tumor Cell Lines. *Cancer Research* **48**, 5133–5135 (1988).

[59] Rofstad, E. K. Influence of Cellular, Microenvironmental, and Growth Parameters on Thermotolerance Kinetics in Vivo in Human Melanoma Xenografts. *Cancer Research* **49**, 5027–5032 (1989).

[60] Fukumura, D. *et al.* Tumor Induction of VEGF Promoter Activity in Stromal Cells. *Cell* **94**, 715–725 (1998).

[61] Applegate, K. G., Balch, C. M. & Pellis, N. R. In Vitro Migration of Lymphocytes through Collagen Matrix: Arrested Locomotion in Tumor-infiltrating Lymphocytes. *Cancer Research* **50**, 7153–7158 (1990).

[62] Senger, D. & Perruzzi, C. Cell migration promoted by a potent GRGDS-containing thrombin-cleavage fragment of osteopontin. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1314**, 13–24 (1996).

[63] Negus, R. P. *et al.* The detection and localization of monocyte chemoattractant protein-1 (MCP-1) in human ovarian cancer. *Journal of Clinical Investigation* **95**, 2391–2396 (1995).

[64] O'Brien, T., Cranston, D., Fuggle, S., Bicknell, R. & Harris, A. L. Two Mechanisms of Basic Fibroblast Growth Factor-induced Angiogenesis in Bladder Cancer. *Cancer Research* **57**, 136–140 (1997).

[65] Waller, E. K. *et al.* Growth of primary T-cell non-Hodgkin's lymphomata in SCID-hu mice: requirement for a human lymphoid microenvironment. *Blood* **78**, 2650–2665 (1991).

[66] Lee, J., Fenton, B. M., Koch, C. J., Frelinger, J. G. & Lord, E. M. Interleukin 2 Expression by Tumor Cells Alters Both the Immune Response and the Tumor Microenvironment. *Cancer Research* **58**, 1478–1485 (1998).

[67] Paget, S. The distribution of secondary growths in cancer of the breast. *The Lancet* **133**, 571–573 (1889).

[68] Argilés, G. *et al.* Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology* **31**, 1291–1305 (2020).

[69] Michielin, O., van Akkooi, A. C. J., Ascierto, P. A., Dummer, R. & Keilholz, U. Cutaneous melanoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Approved by the ESMO Guidelines Committee: February 2002, last update September 2019. *Annals of Oncology* **30**, 1884–1901 (2019).

[70] Machiels, J.-P. *et al.* Squamous cell carcinoma of the oral cavity, larynx, oropharynx and hypopharynx: EHNS–ESMO–ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Annals of Oncology* **31**, 1462–1475 (2020).

[71] Campbell, D. J. & Koch, M. A. Phenotypic and functional specialization of FOXP3+ regulatory T cells. *Nature Reviews. Immunology* **11**, 119–130 (2011).

[72] De Simone, M. *et al.* Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. *Immunity* **45**, 1135–1147 (2016).

[73] Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293–1308.e36 (2018).

[74] Chew, V., Toh, H. C. & Abastado, J.-P. Immune Microenvironment in Tumor Progression: Characteristics and Challenges for Therapy. *Journal of Oncology* **2012**, 1–10 (2012).

[75] Roelands, J. *et al.* Immunogenomic Classification of Colorectal Cancer and Therapeutic Implications. *International Journal of Molecular Sciences* **18**, 2229 (2017).

[76] Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews. Cancer* **12**, 298–306 (2012).

[77] Fridman, W. H., Zitvogel, L., Sautès-Fridman, C. & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nature Reviews. Clinical Oncology* **14**, 717–734 (2017).

[78] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

[79] Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–1247 (2009).

[80] Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).

[81] Mannocci, L. *et al.* High-throughput sequencing allows the identification of binding molecules isolated from DNA-encoded chemical libraries. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17670–17675 (2008).

[82] Hawrami, K., Harper, D. & Breuer, J. Typing of varicella zoster virus by amplification of DNA polymorphisms. *Journal of Virological Methods* **57**, 169–174 (1996).

[83] Slavov, N. Increasing proteomics throughput. *Nature Biotechnology* 1–2 (2021).

[84] Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491–1498 (2015).

[85] Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**, 1–12 (2017).

[86] Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**, 599–604 (2018).

[87] Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–448 (1975).

[88] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467 (1977).

[89] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**, 175–185 (1998).

[90] Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264 (2008).

[91] Picelli, S. Single-cell RNA-sequencing: The future of genome biology is now. *RNA biology* **14**, 637–650 (2017).

[92] Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31–46 (2010).

[93] Ofengeim, D., Giagtzoglou, N., Huh, D., Zou, C. & Yuan, J. Single-Cell RNA Sequencing: Unraveling the Brain One Cell at a Time. *Trends in Molecular Medicine* **23**, 563–576 (2017).

[94] Semrau, S. *et al.* Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature Communications* **8**, 1096 (2017).

[95] Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews. Immunology* **18**, 35–45 (2018).

[96] Zhang, Y. *et al.* Single-cell RNA sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research* **40**, 1–17 (2021).

[97] Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

[98] Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature Protocols* **5**, 516–535 (2010).

[99] Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**, e8746 (2019).

[100] Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology* **38**, 747–755 (2020).

[101] Kolodziejczyk, A., Kim, J. K., Svensson, V., Marioni, J. & Teichmann, S. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell* **58**, 610–620 (2015).

[102] Macosko, E. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

[103] Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**, 171–181 (2014).

[104] Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods* **11**, 41–46 (2014).

[105] Tan, S. J. *et al.* A microfluidic device for preparing next generation DNA sequencing libraries and for automating other laboratory protocols that require one or more column chromatography steps. *PloS One* **8**, e64084 (2013).

[106] Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* **17**, 77 (2016).

[107] Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* **65**, 631–643.e4 (2017).

[108] Cazenave, C. & Uhlenbeck, O. C. RNA template-directed RNA synthesis by T7 RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 6972–6976 (1994).

[109] Saiki, R. K. *et al.* Enzymatic Amplification of β -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science* **230**, 1350–1354 (1985).

[110] Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I. & Khorana, H. G. Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of Molecular Biology* **56**, 341–361 (1971).

[111] Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research* **24**, 2033–2040 (2014).

[112] Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).

[113] Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology* **38**, 737–746 (2020).

[114] Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**, 163–166 (2014).

[115] Huang, H. *et al.* Non-biased and efficient global amplification of a single-cell cDNA library. *Nucleic Acids Research* **42**, e12–e12 (2014).

[116] Baugh, L. R., Hill, A. A., Brown, E. L. & Hunter, C. P. Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Research* **29**, e29–e29 (2001).

[117] Okino, S. T., Kong, M., Sarras, H. & Wang, Y. Evaluation of bias associated with high-multiplex, target-specific pre-amplification. *Biomolecular Detection and Quantification* **6**, 13–21 (2016).

[118] Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, 1–14 (2011).

[119] Gayoso, A. *et al.* scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv* 2021.04.28.441833 (2021).

[120] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9**, 284 (2018).

[121] Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

[122] Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135–1145 (2008).

[123] Kulkarni, A., Anderson, A. G., Merullo, D. P. & Konopka, G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology* **58**, 129–136 (2019).

[124] Jourdren, L., Bernard, M., Dillies, M.-A. & Le Crom, S. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics* **28**, 1542–1543 (2012).

[125] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

[126] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY : large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 1–5 (2018).

[127] Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281–291.e9 (2019).

[128] Nayak, R. & Hasija, Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics* **113**, 606–619 (2021).

[129] Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).

[130] Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**, 1–16 (2019).

[131] Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).

[132] Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews. Genetics* **16**, 133–145 (2015).

[133] Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273–282 (2019).

[134] Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**, 451–464 (2005).

[135] Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nature Reviews. Genetics* **20**, 536–548 (2019).

[136] Raser, J. M. & O'Shea, E. K. Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**, 2010–2013 (2005).

[137] Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science (New York, N.Y.)* **304**, 1811–1814 (2004).

[138] Serizawa, S., Miyamichi, K. & Sakano, H. One neuron-one receptor rule in the mouse olfactory system. *Trends in genetics: TIG* **20**, 648–653 (2004).

[139] Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).

[140] Waddington, C. H. Canalization of development and genetic assimilation of acquired characters. *Nature* **183**, 1654–1655 (1959).

[141] Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik* **43**, 172–198 (1927).

[142] Bong, K.-W. *et al.* A strong no-go theorem on the Wigner's friend paradox. *Nature Physics* **16**, 1199–1205 (2020).

[143] Bell, S. A. A beginner's guide to uncertainty of measurement. *Measurement Good Practice Guide* **11**, 1–30 (2001).

[144] Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**, 1093–1095 (2013).

[145] Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740–742 (2014).

[146] Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics* **53**, 770–777 (2021).

[147] Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**, 637–640 (2014).

[148] Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods* **13**, 241–244 (2016).

[149] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018).

[150] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10**, 390 (2019).

[151] Deng, Y., Bao, F., Dai, Q., Wu, L. F. & Altschuler, S. J. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nature Methods* **16**, 311–314 (2019).

[152] Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv* 791699 (2019).

[153] van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716–729.e27 (2018).

[154] Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods* **15**, 539–542 (2018).

[155] Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18**, 1–15 (2017).

[156] Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16**, 241 (2015).

[157] Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278 (2015).

[158] Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications* **9**, 997 (2018).

[159] Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Research* **7**, 1740 (2019).

[160] Cao, Y., Kitanovski, S., Küppers, R. & Hoffmann, D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nature Biotechnology* **39**, 158–159 (2021).

[161] Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147–150 (2020).

[162] Andrews, T. S. & Hemberg, M. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* **35**, 2865–2867 (2019).

[163] Kainen, P. C. Utilizing Geometric Anomalies of High Dimension: When Complexity Makes Computation Easier. In Kárný, M. & Warwick, K. (eds.) *Computer Intensive Methods in Control and Signal Processing: The Curse of Dimensionality*, 283–294 (Birkhäuser, 1997).

[164] Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods* **14**, 414–416 (2017).

[165] Aucouturier, J.-J. & Pachet, F. Improving timbre similarity: How high is the sky. In *Results in Speech and Audio Sciences* (2004).

[166] Radovanovic, M., Nanopoulos, A. & Ivanovic, M. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research* **11**, 2487–2531 (2010).

[167] Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology* **7**, 36–46 (2018).

[168] Aggarwal, C. C., Hinneburg, A. & Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Van den Bussche, J. & Vianu, V. (eds.) *Database Theory — ICDT 2001*, Lecture Notes in Computer Science, 420–434 (Springer, Berlin, Heidelberg, 2001).

[169] Wang, B. *et al.* SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. *Proteomics* **18** (2018).

[170] Mirkes, E. M., Allohibi, J. & Gorban, A. Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality. *Entropy (Basel, Switzerland)* **22** (2020).

[171] Kairov, U. *et al.* Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics* **18** (2017).

[172] Aparicio, L., Bordyuh, M., Blumberg, A. J. & Rabadan, R. A Random Matrix Theory Approach to Denoise Single-Cell Data. *Patterns* **1**, 100035 (2020).

[173] Feldbauer, R. & Flexer, A. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems* **59**, 137–166 (2019).

[174] Weber, L. M. *et al.* Essential guidelines for computational method benchmarking. *Genome Biology* **20**, 125 (2019).

[175] Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology* **14**, e1006245 (2018).

[176] Schneider, I. *et al.* Use of "default" parameter settings when analyzing single cell RNA sequencing data using Seurat: a biologist's perspective. *Journal of Translational Genetics and Genomics* **5**, 37–49 (2021).

[177] Zemmour, D. *et al.* Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nature Immunology* **19**, 291–301 (2018).

[178] Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R. & Boulesteix, A.-L. On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology* **22**, 1–8 (2021).

[179] Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019** (2019).

[180] Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

[181] Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nature Methods* (2019).

[182] Lieberman, Y., Rokach, L. & Shay, T. CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PloS One* **13**, e0205499 (2018).

[183] Wagner, F. & Yanai, I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv* 456129 (2018).

[184] Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

[185] Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

[186] Cortal, A., Martignetti, L., Six, E. & Rausell, A. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nature Biotechnology* 1–8 (2021).

[187] Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods* **15**, 359–362 (2018).

[188] van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human cancer: insights from single-cell analysis. *Nature Reviews. Cancer* **20**, 218–232 (2020).

[189] Shao, X. *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23** (2020).

[190] Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology* **20**, 194 (2019).

[191] Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods* **16**, 1007–1015 (2019).

[192] Franzén, O. & Björkegren, J. L. M. alona: a web server for single-cell RNA-seq analysis. *Bioinformatics* **36**, 3910–3912 (2020).

[193] Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**, 163–172 (2019).

[194] Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* **17**, e9620 (2021).

[195] Svensson, V., Beltrame, E. d. V. & Pachter, L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv* 762773 (2019).

[196] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

[197] Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).

[198] Schwartz, G. W. *et al.* TooManyCells identifies and visualizes relationships of single-cell clades. *Nature Methods* **17**, 405–413 (2020).

[199] Hu, M.-W. *et al.* PanoView: An iterative clustering method for single-cell RNA sequencing data. *PLoS computational biology* **15**, e1007040 (2019).

[200] Grün, D. Revealing dynamics of gene expression variability in cell state space. *Nature Methods* **17**, 45–49 (2020).

[201] Dong, R. & Yuan, G.-C. GiniClust3: a fast and memory-efficient tool for rare cell type identification. *BMC bioinformatics* **21**, 158 (2020).

[202] Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. *Nature Communications* **9**, 4719 (2018).

[203] Liu, B. *et al.* An entropy-based metric for assessing the purity of single cell populations. *Nature Communications* **11**, 3155 (2020).

[204] Liu, J., Song, Y. & Lei, J. Single-cell entropy to quantify the cellular transcriptome from single-cell RNA-seq data. *bioRxiv* (2019).

[205] Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science (New York, N.Y.)* **356**, eaah4573 (2017).

[206] Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).

[207] Low, T., Borgelt, C., Stober, S. & Nürnberger, A. The Hubness Phenomenon: Fact or Artifact? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, vol. 285, 267–278 (Springer, Berlin, Heidelberg, 2013).

[208] McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* (2018).

[209] Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t -SNE and UMAP. *Nature Biotechnology* **39**, 156–157 (2021).

[210] Feldbauer, R., Rattei, T. & Flexer, A. scikit-hubness: Hubness Reduction and Approximate Neighbor Search. *Journal of Open Source Software* **5**, 1957 (2020).

[211] Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences* **113**, 7900–7905 (2016).

[212] Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**, 547–554 (2019).

[213] Klimovskaia, A., Lopez-Paz, D., Bottou, L. & Nickel, M. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature Communications* **11**, 2966 (2020).

[214] Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009).

[215] Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865–868 (2017).

[216] Larsson, L., Frisén, J. & Lundeberg, J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods* **18**, 15–18 (2021).

[217] Fridman, W. H. Historique de l'immunothérapie. Changement de paradigme ? *Bulletin du Cancer* **103**, S122–S126 (2016).

[218] Kuksin, M. *et al.* Applications of single-cell and bulk RNA sequencing in onco-immunology. *European Journal of Cancer (Oxford, England: 1990)* **149**, 193–210 (2021).

Part IV

Annexes

Chapter 9

Single-cell RNAseq of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity

We implemented in this article the use of functional modules, though not the specific scoring and clustering methods that were developed later. It facilitated the interpretation of dendritic cells, emphasizing functional differences between different groups of Covid-19 patients and healthy subjects.

RESOURCE
https://doi.org/10.1038/s41556-021-00681-2

nature cell biology

Check for updates

Single-cell RNA sequencing of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity

Melissa Saichi[1,9], Maha Zohra Ladjemi[2,3,9], Sarantis Korniotis[1,9], Christophe Rousseau[2], Zakaria Ait Hamou[2,3], Lucile Massenet-Regad[1,4], Elise Amblard[1,5], Floriane Noel[1], Yannick Marie[6,7], Delphine Bouteiller[6], Jasna Medvedovic[1], Frédéric Pène[2,3] and Vassili Soumelis[1,8]✉

COVID-19 can lead to life-threatening respiratory failure, with increased inflammatory mediators and viral load. Here, we perform single-cell RNA-sequencing to establish a high-resolution map of blood antigen-presenting cells (APCs) in 15 patients with moderate or severe COVID-19 pneumonia, at day 1 and day 4 post admission to intensive care unit or pulmonology department, as well as in 4 healthy donors. We generated a unique dataset of 81,643 APCs, including monocytes and rare dendritic cell (DC) subsets. We uncovered multi-process defects in antiviral immune defence in specific APCs from patients with severe disease: (1) increased pro-apoptotic pathways in plasmacytoid DCs (pDCs, key effectors of antiviral immunity), (2) a decrease of the innate sensors *TLR9* and *DHX36* in pDCs and *CLEC9a*$^+$ DCs, respectively, (3) downregulation of antiviral interferon-stimulated genes in monocyte subsets and (4) a decrease of major histocompatibility complex (MHC) class II-related genes and MHC class II transactivator activity in cDC1c$^+$ DCs, suggesting viral inhibition of antigen presentation. These novel mechanisms may explain patient aggravation and suggest strategies to restore the defective immune defence.

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) infection is at the origin of coronavirus disease 2019 (COVID-19), characterized by a first phase of benign flu-like symptoms with an efficient control of the infection in most cases. In a second phase, disease aggravation may lead to acute respiratory failure, sepsis and death[1–6]. This is due to a multiplicity of factors: (1) an exacerbated inflammatory reaction, with systemic and organ-specific manifestations, (2) persistent viral load and (3) defective antiviral defence pathways[1–7]. Identifying the underlying cellular and molecular mechanisms is of paramount importance to understand COVID-19 physiopathology and guide the development of appropriate therapies.

Studies have characterized the systemic inflammatory response, revealing an excess production of inflammatory cytokines such as interleukin-6 (IL-6) and IL-1, tumour necrosis factor-α (TNF-α) and interferon-γ (IFN-γ)[2,8–22], suggesting new therapeutic targets. The endothelium may also contribute to the overt inflammatory reaction through the production of soluble mediators[23,24]. Anti-IL-6 compounds have given promising results in severe COVID-19[25–27]. However, the cellular mechanisms underlying the excessive inflammatory response remain mostly unknown.

Another unresolved question relates to the inefficiency of the innate and adaptive immune system to control the infection in patients with severe COVID-19. It has been suggested that production of IFN-α, a major antiviral cytokine, is decreased in these patients compared to those with moderate disease[6,9,21,28–30]. However, a recent study argued that increased IFN-α production might contribute to the pathogenic inflammatory response[17]. Other antiviral mechanisms and their cellular source remain to be studied.

Dendritic cells (DCs) form a family of innate antigen (Ag)-presenting cells (APCs) that contribute to the control of pathogens and subsequent presentation of pathogen-specific Ag to T cells[31]. Their study is challenging for three main reasons: (1) they are found in very low numbers in the circulation and in tissue, (2) they lack specific lineage-defining markers and (3) they include an ever-increasing number of subsets[31,32]. All DC subsets may potentially and variably contribute to modulating the inflammatory response following viral sensing, producing antiviral effector molecules and priming an Ag-specific adaptive immune response[33]. Plasmacytoid pre-DCs (pDCs) are a particular subset specialized in antiviral immunity through the production of large amounts of type I IFN[34]. Despite their central role in antiviral defence, the contribution of DCs to severe COVID-19 pathogenesis is not yet known.

In this paper we perform a high-resolution single-cell RNA-sequencing (scRNAseq) analysis of all APC subsets from fresh peripheral blood of patients with COVID-19. A pre-enrichment step enables the characterization of even rare DC subsets that were not captured in previous peripheral blood mononuclear cell (PBMC) scRNAseq studies[12,17,35]. We reveal previously unrecognized multi-process defects in patients with severe COVID-19[22,36,37].

Results

APC subset distribution in patients with COVID-19. To characterize the molecular profile of circulating APCs, we performed

[1]Université de Paris, INSERM U976, Paris, France. [2]Institut Cochin, INSERM U1016, CNRS UMR8104, Université de Paris, Paris, France. [3]Service de Médecine Intensive & Réanimation, Hôpital Cochin, Assistance Publique-Hôpitaux de Paris. Centre & Université de Paris, Paris, France. [4]Université Paris-Saclay, Saint-Aubin, France. [5]Université de Paris, Centre de Recherches Interdisciplinaires, Paris, France. [6]Institut du Cerveau (ICM), Plateforme de Génotypage Séquençage, Paris, France. [7]Sorbonne Universités, Université Pierre et Marie Curie, Paris, France. [8]AP-HP, Hôpital Saint-Louis, Laboratoire d'Immunologie-Histocompatibilité, Paris, France. [9]These authors contributed equally: Melissa Saichi, Maha Zohra Ladjemi, Sarantis Korniotis.
✉e-mail: vassili.soumelis@aphp.fr

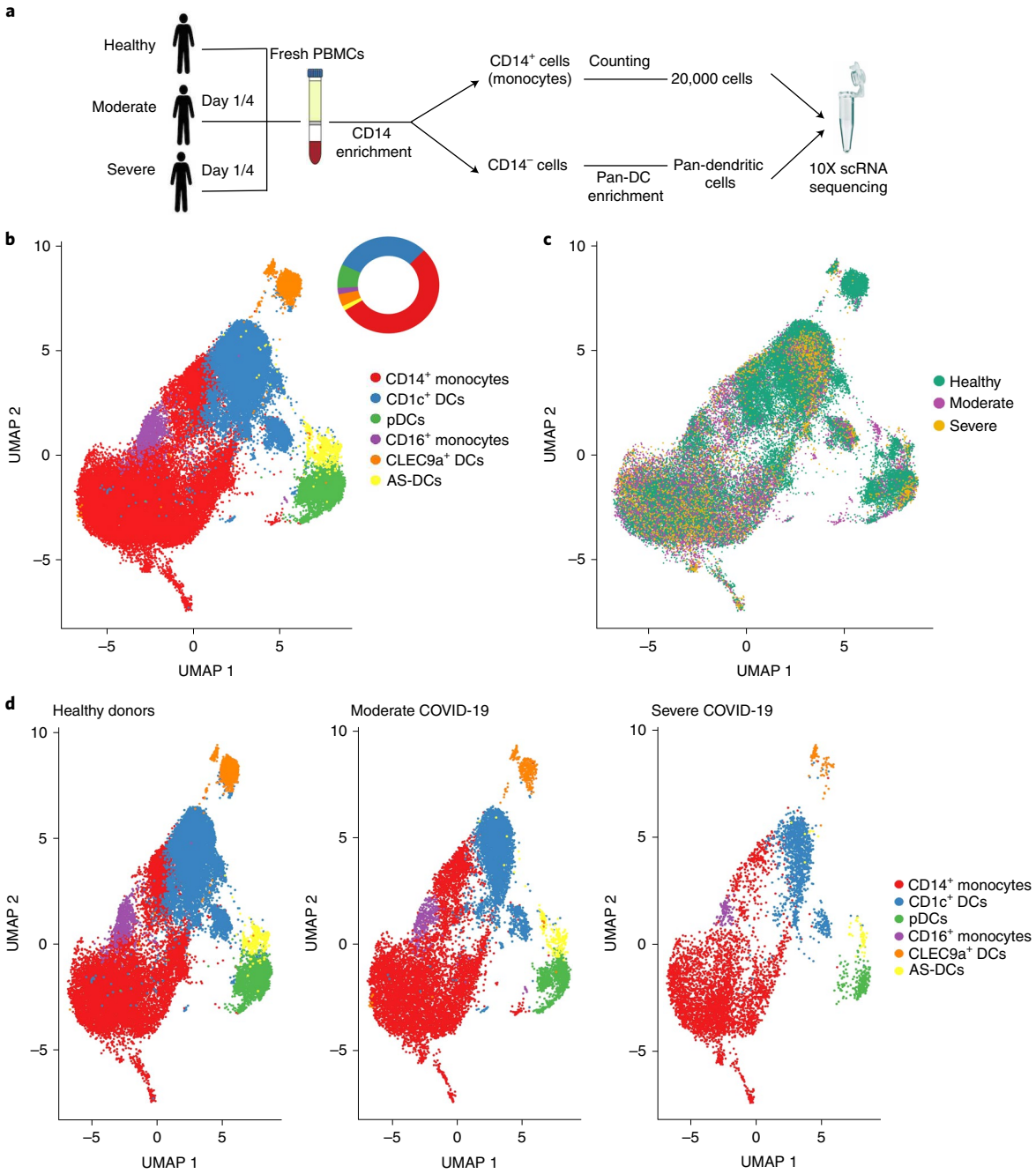


Fig. 1 | Circulating APC subset diversity in COVID-19 from the discovery set. a, Schematic of the experimental workflow. APCs were enriched from fresh PBMCs of healthy donors and patients with COVID-19 with either moderate or severe clinical symptoms at both day 1 and day 4 post hospital admission. The total APCs were sequenced using the 10X Genomics facility. **b**,**c**, Cellular maps of APC subsets ($n = 42{,}784$ cells) from the discovery set at single-cell resolution level displayed on UMAP dimension reduction based either on identified cell types (**b**) and severity (**c**). Proportions of the APC subtypes are displayed on the doughnut plot. **d**, UMAP plot of detected APC populations split by severity group (healthy controls and patients with moderate and severe COVID-19). The discovery set comprises a total of 12 samples ($n = 2$ controls, $n = 4$ moderate and $n = 6$ severe samples) collected from a total of seven patients and two healthy donors.

scRNAseq on freshly sampled APC-enriched PBMCs from five patients with moderate COVID-19 (non-mechanically ventilated, oxygen supply $<10\,l\,min^{-1}$) and ten patients with severe COVID-19 (mechanically ventilated or oxygen supply $\geq 10\,l\,min^{-1}$), at day 1 and day 4 following hospital and/or intensive care unit (ICU) admission, as well as four elderly healthy controls (HC) (Fig. 1a and Supplementary Tables 1 and 2). To obtain single-cell suspensions and minimize DC–DC and DC–T cell clusters and clumps, EDTA-containing medium was used for the enrichment steps in the first set of samples, which we further define as the 'discovery set'. This set is composed of a total of 12 samples from two HCs, three patients with moderate COVID-19 and four patients with severe COVID-19 from both day 1 and day 4 time points (results are presented in the figures and Extended Data Figs. 1 and 2). However, EDTA is known to decrease reverse transcription (RT) efficiency through RT deactivation and ion chelation,

resulting in reduced amounts of complementary DNA (cDNA) during amplification. We therefore validated the main results derived from the discovery set by using RPMI (EDTA-free) medium for the enrichment steps in a second set of samples, including a total of 15 samples (defined as the 'validation set') from two HCs, two patients with moderate COVID-19 and six with severe COVID-19 (Extended Data Figs. 3–7).

For each fresh sample, 25,000 cells (~20,000 monocytes and 5,000 total DCs) were loaded onto the 10X lane (10X Genomics technology) (Fig. 1a). As expected, more cells per sample were effectively sequenced in the EDTA (discovery set) than the RPMI (validation set) dataset (mean: 3,360 versus 2,528 cells, respectively) (Extended Data Fig. 3a), confirming that EDTA optimizes single-cell suspension efficiency for rare DC types. This retrospectively justified the importance of using two complementary experimental protocols, split into two independent datasets, to avoid biasing the results. All main findings were validated in both datasets, indicating the reproducibility and robustness to experimental procedures. Altogether, we analysed a total of 81,643 APCs, split into 42,784 cells in the discovery set and 38,859 cells in the validation set. The two sets were analysed separately after sample integration using Harmony[38]. Graph-based clustering (SNN-based), community detection and nonlinear dimension reduction, using uniform manifold approximation and projection (UMAP), were independently applied to both sets for cell cluster visualization. Manual annotation of the cell clusters using canonical gene signature markers for each APC subset established a comprehensive map of APCs in HCs and patients with COVID-19 in both sets (Fig. 1b,c and Extended Data Fig. 3b,c). In our discovery set, among the 42,784 APCs, we recovered six subsets: 22,690 CD14$^+$ monocytes; 866 CD16$^+$ monocytes; 13,252 CD1c$^+$ DCs; 1,754 CLEC9a$^+$ DCs; 3,538 pDCs; 684 Axl$^+$Siglec6$^+$ AS-DCs (Extended Data Fig. 3a). The validation set included 29,409 CD14$^+$ and 1,021 CD16$^+$ monocytes, 5,754 CD1c$^+$, 197 CLEC9a$^+$ DCs, 1,602 pDCs and 876 AS-DCs (Extended Data Fig. 3a). In both sets, APC populations were captured across all the collected samples (Supplementary Table 3).

The accurate identification of all six APC populations was confirmed by the expression of canonical markers defining each subset (Extended Data Fig. 3d). All DC populations expressed higher levels of human leukocyte antigen *HLA-DR* and *CD86* compared to monocytes (Extended Data Fig. 3d). None of the cells expressed *CD19* (B-cell marker), *GNLY* (natural killed (NK) marker) or *CD3E* (T-cell marker), validating the pure APC populations. CD14$^+$ monocytes expressed lineage-defining *CD14*, whereas CD16$^+$ monocytes expressed *FCGR3A*. *AXL* expression distinguished AS-DCs from pDCs, whereas *CD1c* and *CLEC9a* characterized the respective cDC subsets[39,40]. In both sets (discovery and validation), UMAP embeddings coloured by severity revealed the heterogeneity of APC distribution between the three groups (Fig. 1c). This was confirmed by splitting the UMAP embeddings per severity (Fig. 1d). Overall, our enrichment strategy allowed the efficient identification of all APC populations including the rare pDCs, AS-DCs and CLEC9A$^+$ DCs, enabling further molecular and phenotypic characterization.

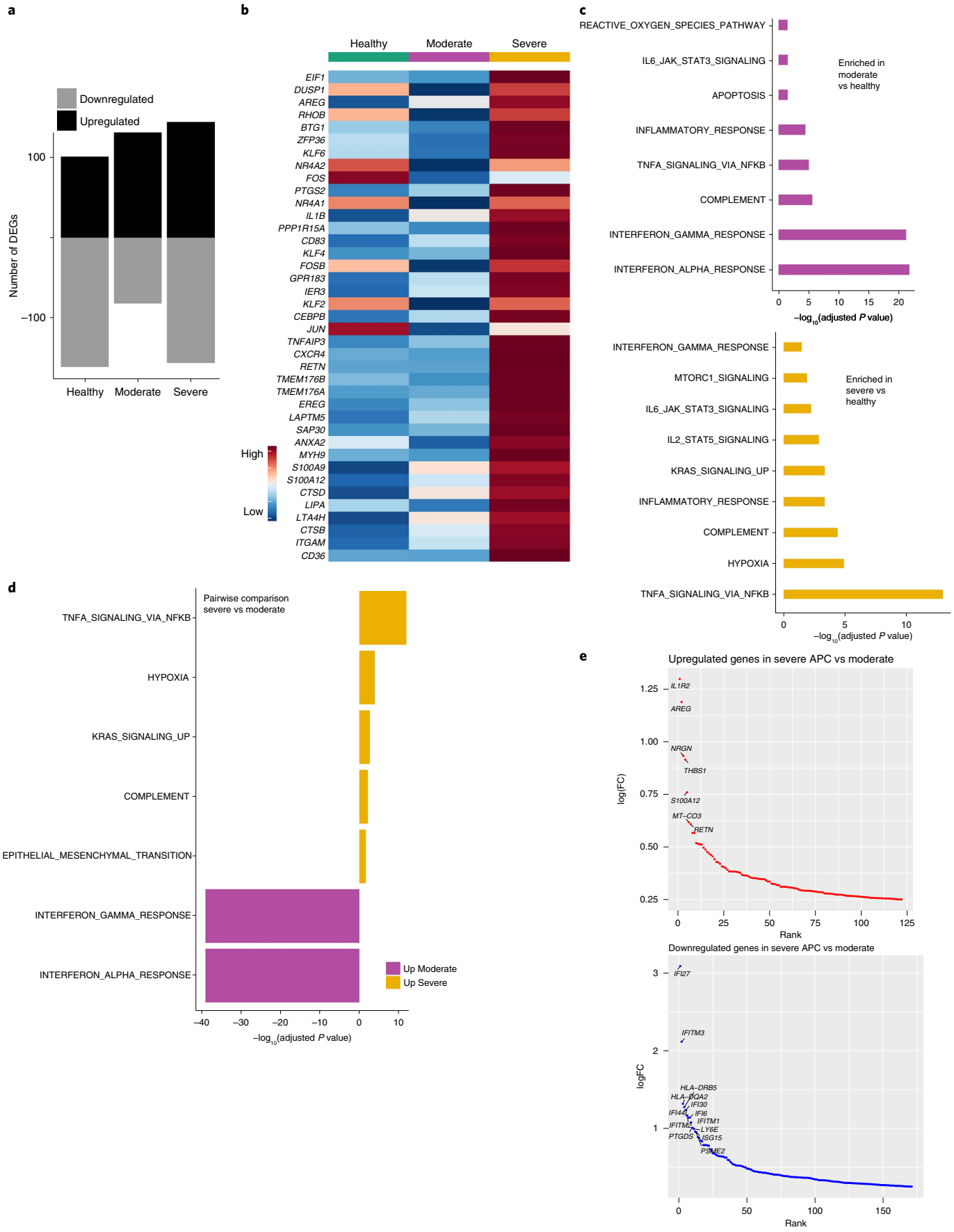
Inflammation-related pathways are hallmarks of COVID-19 APCs. We performed differential expression and pathway enrichment analyses among APC severity groups, revealing 368 differentially expressed genes (DEGs) among the three groups (absolute fold change > 1.4). Among them, 101 genes were upregulated in HCs (as compared to patients with moderate and severe COVID-19), 109 in patients with moderate COVID-19 and 134 in patients with severe COVID-19 as compared to the two other groups, respectively (Fig. 2a). The top 50 DEGs upregulated in severe APCs as compared to HCs and patients with moderate COVID-19 included pro-inflammatory molecules (*IL1B*, *CXCR4*), surface markers (*CD36*, *CD83*, *AREG*, *ITGAM*), enzymes (*CTSD*, *CTSB*) and secreted molecules (*RETN*, *EREG*, *ANXA2*) (Fig. 2b). Next, we sought to identify enriched pathways discriminating each severity group from HCs. We found enriched IFN-γ and IFN-α response pathways in APCs from patients with moderate COVID-19, whereas hypoxia and TNF-α signalling were enriched in patients with severe COVID-19 (Fig. 2c).

We next compared the enriched pathways upregulated in severe versus moderate COVID-19 and in moderate versus severe, respectively. We found that IFN-γ and IFN-α pathways could be used to discriminate moderate from severe APCs at the global level (Fig. 2d).

To allow for an accurate comparison between the two transcriptional signatures, we ranked the DEGs of the pairwise comparison according to decreasing fold change. Severe APCs significantly upregulated *AREG* (amphiregulin), *IL1R2* (IL-1 receptor), *NRGN* (calmodulin binding protein) and pro-inflammatory molecules (*S100A12*) (Fig. 2e). However, moderate APCs overexpressed interferon-stimulated genes (ISGs; *IFITM2*, *ISG15* and *IFI27*) and HLAII molecules (*HLA-DRB5* and *HLA-DQA2*), suggesting decreased Ag presentation and antiviral programs in severe as compared to moderate APCs (Fig. 2e). Similar observations were recovered from our validation set (Extended Data Fig. 4a–d). Additional upregulated genes in severe as compared to moderate APCs were found in the validation set, including *CXCL8*, *NAMPT* and *G0S2* (Extended Data Fig. 4e).

Defective IFN responses in COVID-19 APCs. Increases in inflammatory cytokines have been reported in COVID-19. We addressed the global contribution of APCs to the expression of inflammatory cytokines and their receptors. As compared to APCs derived from HCs, *IL1B*, *CXCL2*, *CXCL8* and *CCL3* were significantly increased, whereas *IL18* was decreased in both severity groups (Fig. 3a and Extended Data Fig. 1a). *TGFB1* and *IL10RA* expression decreased in severe, but not in moderate subsets, as compared to HCs (Fig. 3a and Extended Data Fig. 1a), whereas *IL6* was not detected in our discovery set (Extended Data Fig. 1a). Despite the low expression levels of most cytokines, we explored downstream biological pathways associated with inflammatory cytokine signalling (mainly *IL1B*, *IL6* and *TNF-α*). In comparison to APCs from HCs, both moderate and severe APCs showed higher score levels for hallmark inflammatory pathways, including 'IL6_JAK_STAT3', 'TGF-β', 'P53', 'TNFa_SIGNALLING_VIA_NFKB' and 'KRAS_SIGNALLING' (Fig. 3b).

Fig. 2 | Global increase in inflammation-associated pathways in COVID-19 APCs (discovery set). a, Barplot of the number of differentially expressed genes (DEGs) for each severity group (healthy versus patients with moderate and severe COVID-19; moderate versus healthy and severe; severe versus healthy and moderate). Upregulated (log fold change (FC) > 0.25) genes are shown in black, downregulated (log FC < −0.25) genes are shown in grey. **b**, Heatmap representation of the top upregulated genes in severe APCs, as compared to moderate and healthy groups. The *z*-score values of average expression levels of cells per severity group are colour-coded. **c**,**d**, Comparative analysis of enriched pathways from the upregulated genes in moderate or severe APCs as compared to healthy cells (**c**), as well as pairwise comparison of upregulated genes in moderate compared to severe (shown in pink) and upregulated genes in severe compared to moderate (shown in yellow) (**d**). Horizontal axes display the adjusted *P* values ($-\log_{10}$). **e**, Representation of ranked genes in descending order according to their absolute log FC, upregulated in moderate as compared to severe (red plot) and upregulated in severe as compared to moderate (blue plot). Top genes, with an absolute value of log FC above 0.5, are shown. In **a**–**e**, comparative analyses were performed on the discovery set ($n = 42{,}784$ cells), composed of $n = 2$ HC, $n = 4$ moderate and $n = 6$ severe samples. The two-sided Wilcoxon rank-sum test was used for comparison, *P* values were adjusted to multiple testing using 'Bonferroni' correction, and only genes with adjusted $P < 0.05$ were considered.


a
Downregulated
Upregulated
100
−100
Number of DEGs
Healthy
Moderate
Severe
b
Healthy
Moderate
Severe
EIF1
DUSP1
AREG
RHOB
BTG1
ZFP36
KLF6
NR4A2
FOS
PTGS2
NR4A1
IL1B
PPP1R15A
CD83
KLF4
FOSB
GPR183
IER3
KLF2
CEBPB
JUN
TNFAIP3
CXCR4
RETN
TMEM176B
TMEM176A
EREG
LAPTM5
SAP30
ANXA2
MYH9
S100A9
S100A12
CTSD
LIPA
LTA4H
CTSB
ITGAM
CD36
High
Low
c
REACTIVE_OXYGEN_SPECIES_PATHWAY
IL6_JAK_STAT3_SIGNALING
APOPTOSIS
INFLAMMATORY_RESPONSE
TNFA_SIGNALING_VIA_NFKB
COMPLEMENT
INTERFERON_GAMMA_RESPONSE
INTERFERON_ALPHA_RESPONSE
Enriched in moderate vs healthy
0
5
10
15
20
−log10(adjusted P value)
INTERFERON_GAMMA_RESPONSE
MTORC1_SIGNALING
IL6_JAK_STAT3_SIGNALING
IL2_STAT5_SIGNALING
KRAS_SIGNALING_UP
INFLAMMATORY_RESPONSE
COMPLEMENT
HYPOXIA
TNFA_SIGNALING_VIA_NFKB
Enriched in severe vs healthy
0
5
10
−log10(adjusted P value)
d
Pairwise comparison severe vs moderate
TNFA_SIGNALING_VIA_NFKB
HYPOXIA
KRAS_SIGNALING_UP
COMPLEMENT
EPITHELIAL_MESENCHYMAL_TRANSITION
INTERFERON_GAMMA_RESPONSE
INTERFERON_ALPHA_RESPONSE
Up Moderate
Up Severe
−40
−30
−20
−10
0
10
−log10(adjusted P value)
e
Upregulated genes in severe APC vs moderate
log(FC)
1.25
1.00
0.75
0.50
0.25
IL1R2
AREG
NRGN
THBS1
S100A12
MT−CO3
RETN
0
25
50
75
100
125
Rank
Downregulated genes in severe APC vs moderate
logFC
3
2
1
IFI27
IFITM3
HLA−DRB5
HLA−DQA2
IFI30
IFI44L
IFI6
IFITM1
IFITM2
LY6E
PTGDS
ISG15
PSME2
0
50
100
150
Rank

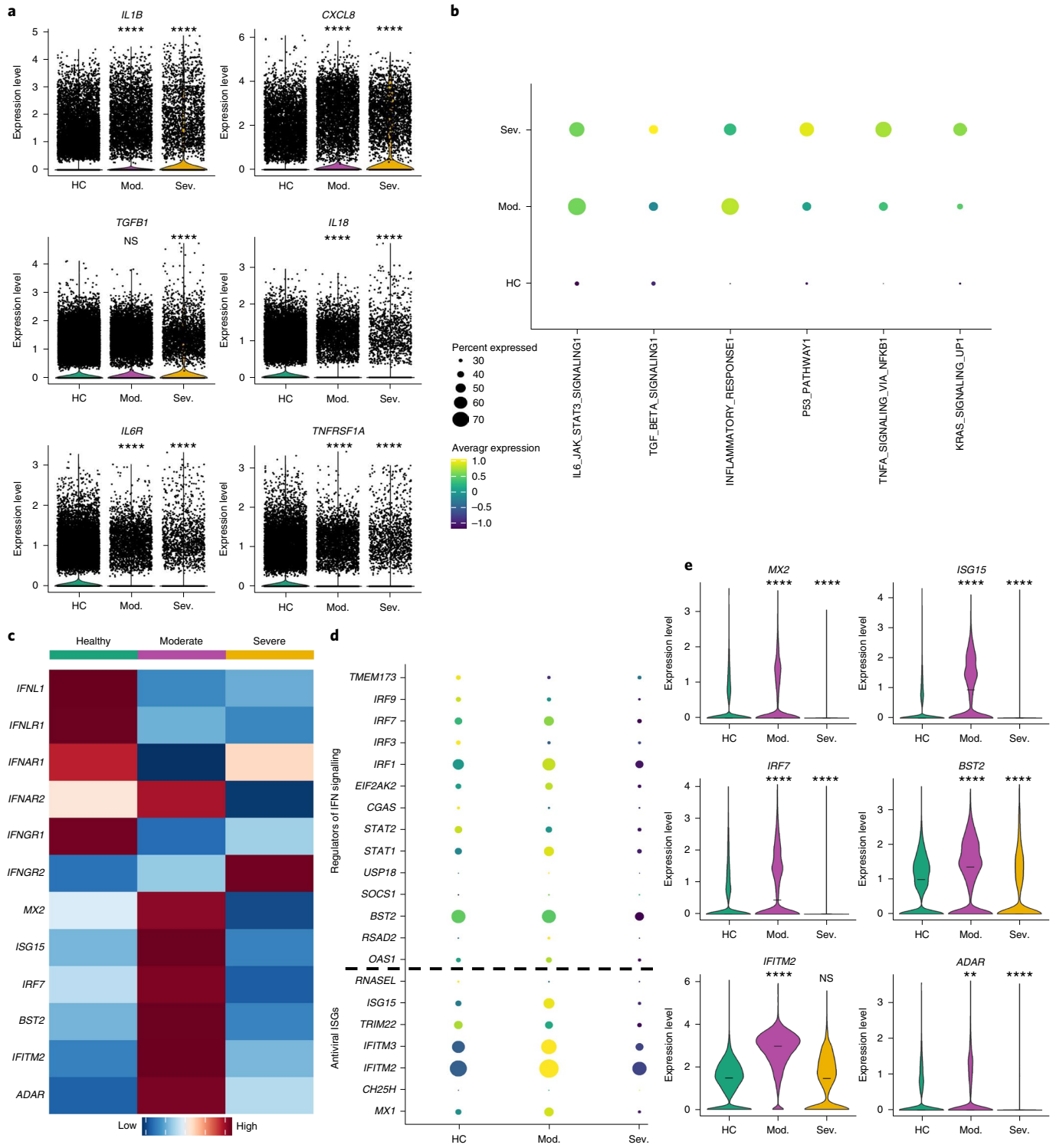


Fig. 3 | Activation of downstream pathways associated with pro-inflammatory cytokines is correlated with defective IFN responses in severe COVID-19 APCs from the discovery set. a, Violin plot representation of cytokine (*IL1B*, *TGFB1* and *IL-18*) receptor (*IL6R*, *TNFRSF1A*) and chemokine (*CXCL8*) gene expression levels detected by scRNAseq and comparison between severity groups. Each dot represents a cell and horizontal lines display the mean expression value. **b**, Dot plot of enrichment scores of pathways downstream of *IL-1B*, *IL-6* and *TGFB1* inflammatory cytokines; score levels are colour-coded, and the percentage of cells expressing the pathway score is size-coded. **c**, Heatmap representation of expression levels of IFN genes (ligands and receptors) and ISG expression levels in healthy, moderate and severe APCs. Expression levels are colour-coded. **d**, Dot plots of regulators of IFN signalling and antiviral ISG genes in HCs and patients with moderate and severe COVID-19. Expression levels are colour-coded, and the percentage of cells expressing the respective gene is size-coded. **e**, Violin plot representation of antiviral ISGs among the severity groups; the small horizontal line indicates the mean expression value for each plotted gene expression. In **a** and **e**, the violin plots were designed using the total APC subsets from the discovery set ($n = 42,784$ cells), composed of $n = 2$ HC, $n = 4$ moderate and $n = 6$ severe samples. Comparative analysis was performed using the two-sided Wilcoxon rank-sum test; *P* values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; NS, not significant.

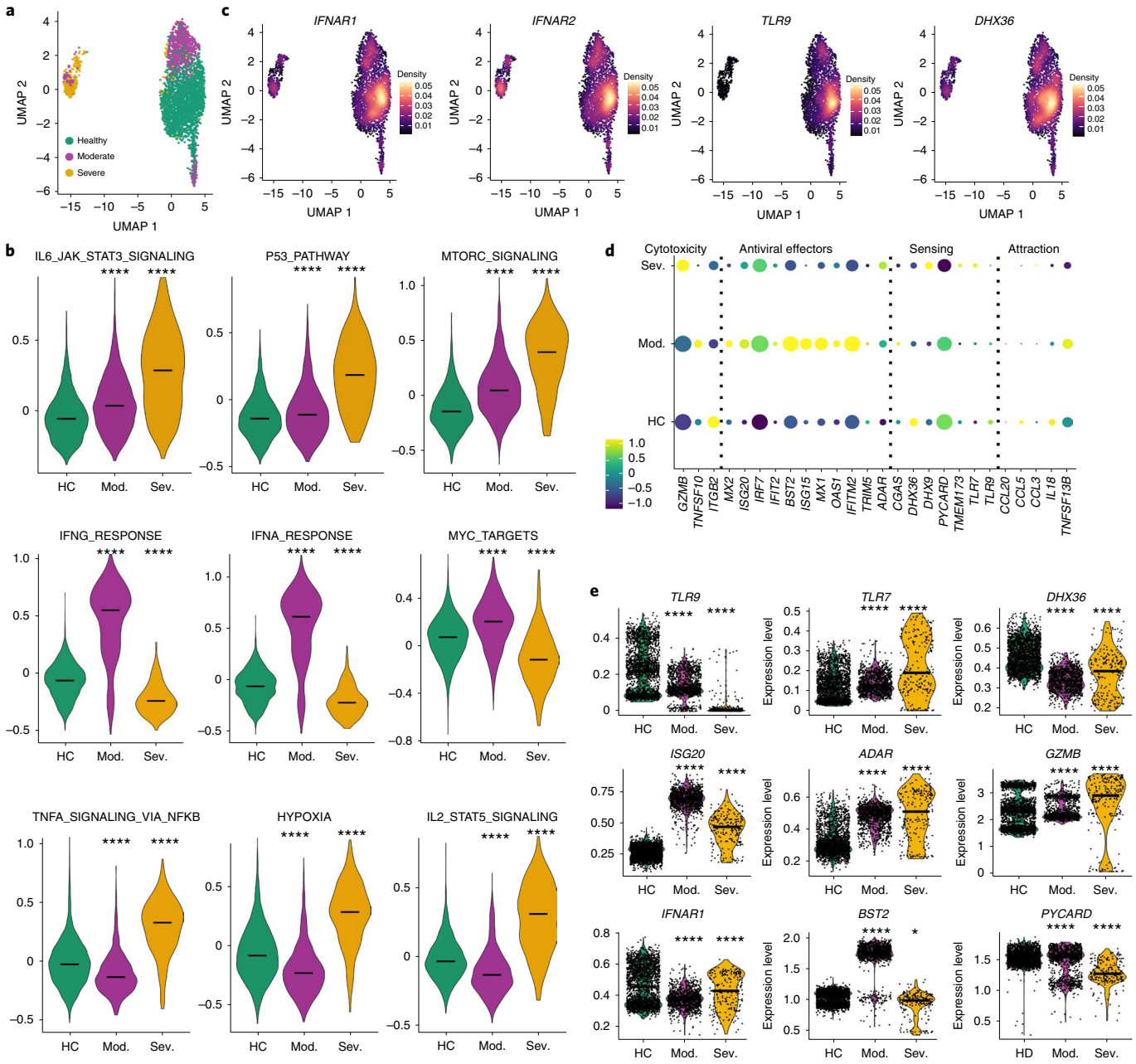


Fig. 4 | Multi-process defects in severe COVID-19 pDC effector pathways in the discovery dataset. a, pDCs displayed on UMAP ($n = 3{,}538$ cells) coloured by severity group of origin. **b**, Violin plot distribution of enrichment scores for enriched hallmark pathways from upregulated genes from pairwise comparison between the three severity groups; severity groups are colour-coded, each dot represents a cell and the horizontal line displays the mean value of the enrichment score of each given pathway **c**, UMAP representation of density scores corresponding to IFN-a receptors (*IFNAR1* and *IFNAR2*) and pDC sensors (*TLR9* and *DHX36*). Density levels were computed using Nebulosa and are colour-coded. **d**, Dot plot of in-house constructed pDC-related functional modules (cytotoxicity, antiviral effector molecules, innate sensing and attraction) and comparison among severity groups. Expression levels are colour-coded, and the percentage of cells expressing the respective gene is size-coded. **e**, Violin plot representation of genes involved in pDC defined biological functions between severity groups; each dot represents a cell, and horizontal lines display the mean expression value. In **b** and **e**, the violin plots were designed using the total pDC subsets from the discovery set, including $n = 2$ HC, $n = 4$ moderate and $n = 6$ severe samples. Comparative analysis was performed using the two-sided Wilcoxon rank-sum test, *P* values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

The IFN family of cytokines is one of the most important for innate and adaptive antiviral responses. We showed the expression levels of *IFNL1*, *IFNL1R*, *IFNAR1*, *IFNAR2*, *IFNA1*, *IFNGR1* and *IFNGR2* and further explored their distribution in the three severity groups via a scaled heatmap (Fig. 3c). Both IFN receptor types (*IFNAR1*, *IFNAR2*, *IFNGR1* and *IFNGR2*) were broadly expressed in the APC subsets, whereas detection of IFNL1 and IFNLR1 was patchy in our discovery dataset (Extended Data Fig. 3b). The heatmap representation indicated that severe APCs expressed lower levels of IFN molecules, suggesting a potential defect in IFN signalling (Fig. 3c). To further validate this hypothesis, we investigated the expression levels of ISGs. We observed higher expression levels of

ISGs (*MX2*, *ISG15*, *IRF7*, *BST2*, *IFITM2* and *ADAR*) in moderate APCs, but lower levels in severe APCs, supporting the hypothesis of defective antiviral programs contributing to the severity of COVID-19 (Fig. 3c). We further stratified a more exhaustive ISG signature according to their respective functions related to 'antiviral' and 'regulators of IFN signalling'. Moderate APCs displayed higher levels of these two ISG families compared to both severe and HC groups (Fig. 3d,e). These results suggest a global perturbation of IFN downstream functions in severe COVID-19 APCs.

Multi-process effector defects in severe COVID-19 pDCs. After having analysed COVID-19 APCs at the global level, we sought to decipher alterations occurring in specific APC subsets. To depict the alterations occurring in pDC subsets, we isolated and sub-clustered pDCs (Fig. 4a) and performed pairwise differential expression among the three severity groups. Pathway enrichment analysis using MsigDB hallmark signatures was conducted on the upregulated genes in each subset. Compared to both moderate and HC pDCs, severe pDCs were enriched for the 'TNFa_SIGNALING', 'IL2_STAT5' and 'HYPOXIA' signalling pathways. In parallel, compared to pDCs from patients with moderate COVID-19, pDCs from patients with severe COVID-19 were enriched in the 'IL6_JAK_STAT3', 'P53' and 'MTORC' signalling pathways (Fig. 4b). When comparing pDCs between patients with moderate and severe COVID-19, the most notably enriched pathways were related to IFN signalling (IFNG and IFNA response), along with MYC targets signalling pathways (Fig. 4b). We asked whether apoptosis and pro-inflammatory signalling signatures would be associated with changes in pDC innate sensing receptors, including *TLR9*, *DHX36*, *IFNAR1* and *IFNAR2*. We imputed the expression values to recover the signal from dropped-out features using Nebulosa (https://github.com/powellgenomicslab/Nebulosa), and plotted the density estimation values on UMAP embeddings (Fig. 4c). We observed zero-value density levels for *TLR9*, along with decreased density levels for *DHX36*, *IFNAR1* and *IFNAR2*, in pDCs from patients with severe COVID-19 (Fig. 4c). To explore whether these modulations may impact pDC functions, we defined four original functional modules using a literature-driven manual curation: 'immune cell attraction' (hereafter 'attraction') (18 genes), 'innate sensing' (12 genes), 'antiviral effector molecules' (23 genes) and 'cytotoxicity' (12 genes) (Fig. 4d). Each of these modules was crossed with the pDC expression matrix, and detected genes were depicted for each patient group (Fig. 4d). No major differences between groups were detected within the 'attraction' module. On the contrary, many genes in the 'innate sensing', 'antiviral effector molecules' and 'cytotoxicity' modules were detected in the three groups, and followed the same pattern: baseline in HCs, increased in patients with moderate COVID-19 and decreased in patients with severe COVID-19 (Fig. 4d,e and Extended Data Fig. 5). This was particularly striking for the viral sensors *TLR7*, *DHX9* and *DHX36*, the cytotoxic molecule *TNFSF10* and the antiviral effector IRF7. These results were supported by the downregulation of antiviral ISGs and innate sensors in pDCs from patients with severe COVID-19, including *BST2* and *PYCARD* (Fig. 4e), in both experimental datasets (Extended Data Fig. 5).

Coordinated transcriptional adaptation in monocyte subsets. Monocytes have been implicated in the physiopathology of severe sepsis and COVID-19. We performed dimensionality reduction through independent component analysis (ICA) and highlighted cells according to their severity group. We observed that IC1 clearly separated moderate from severe and HC CD14$^+$ monocytes, whereas IC2 distinguished HC from COVID-19 CD14$^+$ monocytes (Fig. 5a). The top 50 genes contributing to either IC1 or IC2 revealed distinct transcriptional signatures for the CD14$^+$ monocyte subsets identified in each severity group: the severe subset expressed higher levels of complement (*C1GC* and *C1GB*), B7 family (*VSIG4*) and *CD163*, which may function as an innate immune sensor and inducer of local inflammation. The moderate monocyte subset expressed increased levels of antiviral ISGs (*IFITM1*, *IFITM3*, *IFI27*, *MZB1* and *IFI6*) and the HLA-II gene (*HLA-DRB5*), suggesting an efficient antiviral program (Fig. 5b). Compared to HCs, several transcription factors (TFs) were downregulated in both moderate and severe groups, including the AP-1 superfamily (*FOS*, *JUNB* and *ZFP36*) and *DUSP1*, involved in MAPK dephosphorylation (Fig. 5b). Pathway enrichment analysis on the top 50 genes contributing to IC1 and IC2 identified key pathways that segregated COVID-19 CD14$^+$ monocytes from HCs (Fig. 5c). The 'complement', 'TNF-α', 'KRAS' and 'hypoxia' signalling pathways were upregulated in COVID-19 monocytes, whereas 'IFN-α' and 'IFN-γ' response signalling were decreased in the severe subset, as compared to the HC and moderate subsets (Fig. 5c). To estimate antiviral effector functions, we used our manually curated gene functional module across patient groups (Fig. 5d). We observed a decrease of almost all antiviral effector molecules in patients with severe COVID-19, as compared to either HCs or patients with moderate COVID-19, in both experimental datasets (Fig. 5d and Extended Data Fig. 6). In parallel, we subclustered CD16$^+$ monocytes and reduced the data dimension using UMAP projection to depict the corresponding clusters for each severity group (Fig. 5e). Differential expression between the three severity groups of this subset indicated similar trends as described in CD14$^+$ monocytes (Fig. 5b,f). This included overexpression of 'complement'-related genes (*C1QA*, *C1QB* and *C1GC*) by the severe subset, upregulation of antiviral ISGs (*ISG15*, *IFI6* and *IFI44L*) in the moderate subset, as compared to the HC subset (Fig. 5f). Overall, these disease-associated changes in CD16$^+$ paralleled those observed in CD14$^+$ monocytes, suggesting common adaptation mechanisms.

CLEC9A$^+$ DC- and AS-DC-specific transcriptional alterations. Thanks to our APC enrichment protocol, we could recover rare CLEC9a$^+$ DC and AS-DC subsets. Differential expression of AS-DC severity groups revealed significant upregulated genes in severe AS-DCs (*SEPT7* and *AREG*), compared to the moderate and HC subsets. We could also observe a significant downregulation of the *HLA-DQA2* gene and antiviral *IFI27* gene in severe, compared to moderate AS-DCs (Fig. 6a). In the search for upstream regulatory mechanisms, we inferred TF activity using the Dorothea algorithm[41] and scored the activity of each regulon using the Viper inference tool[42]. This identified a large number of highly variant TF activity scores (Fig. 6b). In moderate AS-DCs, we observed a higher activity scored for *IRF1*, *IRF9* and *STAT2*, reported to be involved in the ISG transcription cycle (Fig. 6b). In AS-DCs from patients with severe COVID-19, we found increased TF activities for *RELA*, *NFKB1*, *STAT5* and *STAT3*, indicative of a higher activation of NFKB/STAT signalling, potentially induced by the pro-inflammatory cytokines described in the 'APC subset distribution in patients with COVID-19' section, along with hypoxia activation, indicated by a higher activity of HIF1A (Fig. 6b).

DEGs among the CLEC9a$^+$ DC subclusters included specific transcriptional signatures segregating patients with moderate and severe COVID-19 from HCs (Fig. 6c). We remarkably observed a downregulation of HLA-II genes, including *HLA-DQB1* and *HLA-DPB1*, in severe as compared to HCs, along with a significant upregulation of a larger subset of ISGs, including *IRF1*, *IFI44L*, *IFI6*, *IFI27*, *IFITM2*, *IFITM3*, *IFI44L*, *ISG15* and *ISG20*, in moderate as compared to both HC and severe subsets (Fig. 6c). Expression values representation indicated a significant increase of *AREG* and *SEPT7* genes, which were also upregulated by severe AS-DCs (Fig. 6a,d). Most importantly, we noted a significant decrease of the *IFNGR1* CLEC9a$^+$ DC subset in patients with moderate and severe COVID-19 as compared to HCs (Fig. 6d), supporting a defective antiviral program.

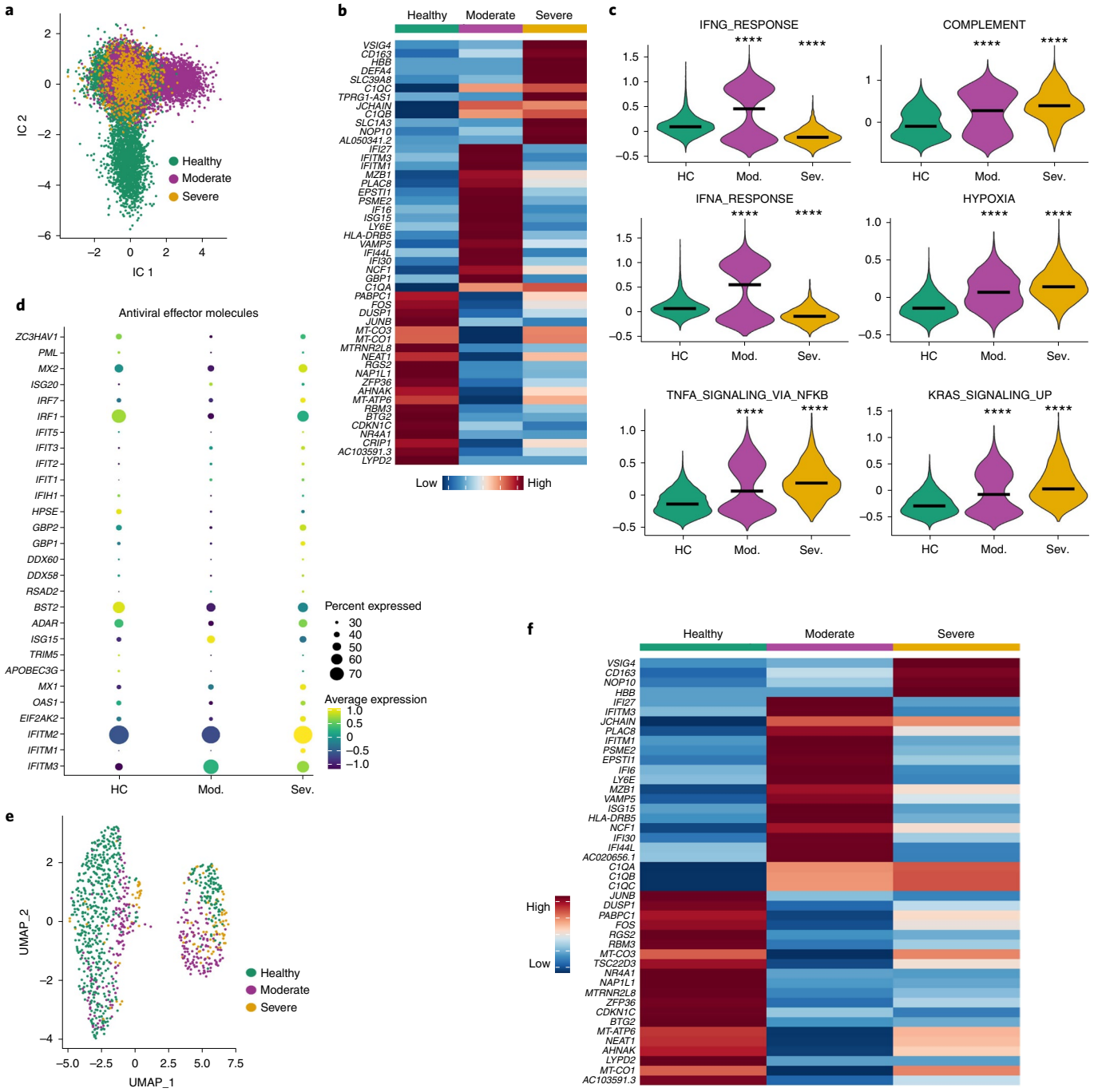


Fig. 5 | Dissection of inflammatory and antiviral response pathways in all monocyte (CD14⁺ and CD16⁺) subsets from the discovery set. a, ICA representation of CD14⁺ monocytes derived from all severity groups ($n = 22{,}690$ cells). IC1 and IC2 components allowed the separation of CD14⁺ monocytes according to severity. **b**, Heatmap representation of the top 50 unique genes contributing to either IC1 or IC2 in the CD14⁺ monocytes subset; *z*-scores of average expression levels are colour-coded. **c**, Violin plot distribution of enrichment score values of enriched pathways in the top 50 genes contributing to either IC1 or IC2; severity groups are colour-coded, each dot represents a cell and the horizontal line displays the mean value of the enrichment score of each given pathway. The violin plots were designed using the total CD14⁺ monocyte subsets from the discovery set, obtained from $n = 2$ HC, $n = 4$ moderate and $n = 6$ severe samples. Comparative analysis was performed using the two-sided Wilcoxon rank-sum test. *P* values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **d**, Dot plot of in-house constructed antiviral effector molecule modules across CD14⁺ monocytes and severity groups. Expression levels are colour-coded, and the percentage of cells expressing the respective gene is size-coded. **e**, UMAP representation of CD16⁺ monocytes ($n = 866$ cells) labelled according to severity group. **f**, Heatmap representation of top 50 DEGs among the three groups from the CD16⁺ monocytes subset; *z*-scores of average expression levels are colour-coded.

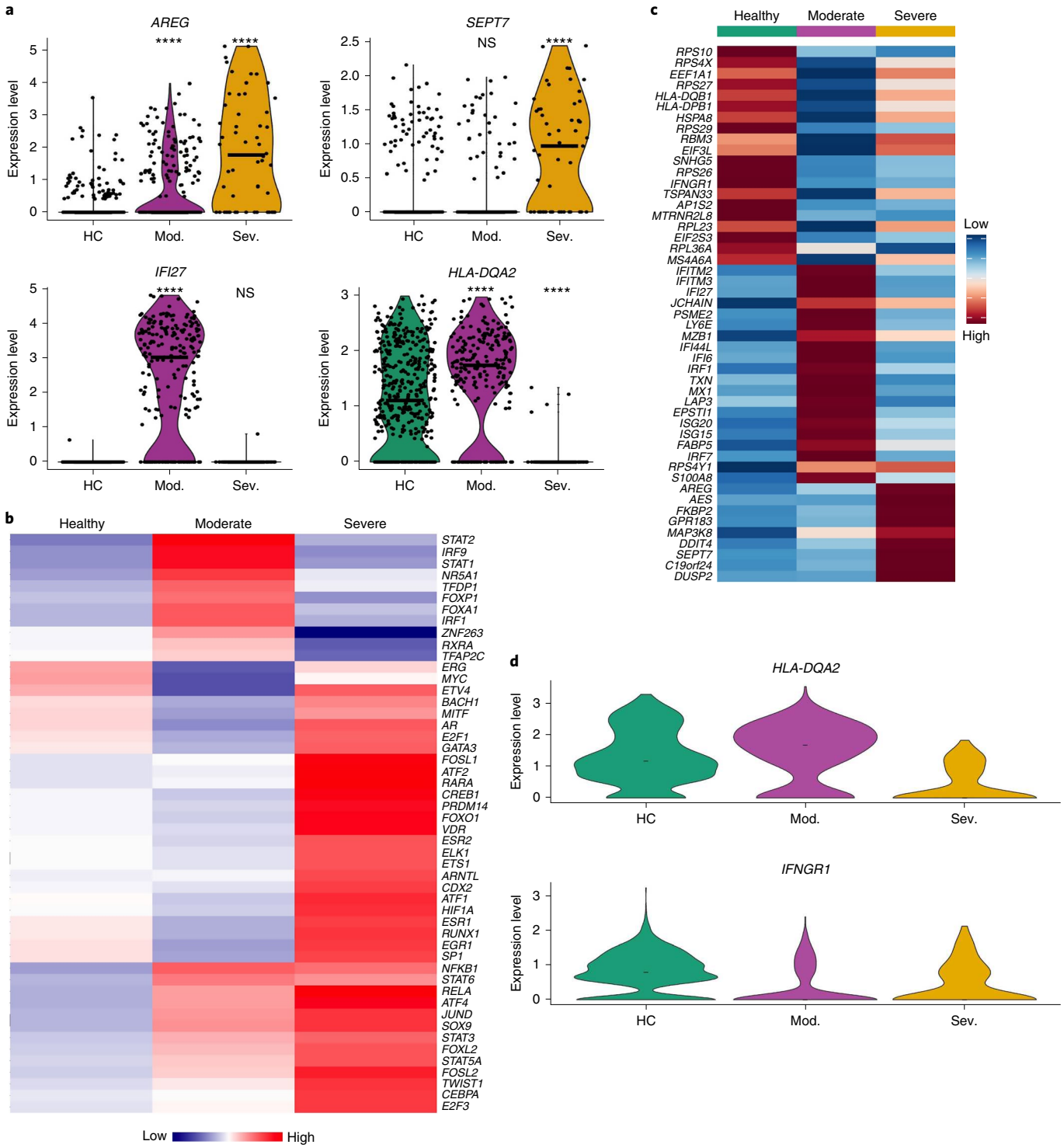


Fig. 6 | Molecular and functional modules in rare CLEC9A$^+$ and AS-DC subsets from the discovery set. a, Violin plot representation of the top upregulated (*AREG*, *SEPT7*)/downregulated (*IFI27*, *HLA-DQA2*) genes in severe as compared to moderate AS-DCs. The violin plots were designed using the total AS-DC subsets from the discovery set ($n = 684$ cells), obtained from $n = 2$ HC, $n = 4$ moderate and $n = 6$ severe samples. Comparative analysis was performed using the two-sided Wilcoxon rank-sum test; *P* values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; NS, not significant. **b**, Heatmap of top 50 highly variable TF activities among the three severity groups; the *z*-scores of TF activities are colour-coded. **c**, Heatmap representation of top 50 DEGs among the three severity groups isolated from the CLEC9a$^+$ DC subset (a total of 1,754 cells); *z*-scores of expression level values are colour-coded; **d**, Violin plot distribution of *IFNGR1* and *HLA-DQA2* genes among the three CLEC9a$^+$ DC severity groups.

Downregulation of MHC-II and CIITA activity in CD1C$^+$ DCs. We next focused on disease-induced alterations in CD1c$^+$ DCs. We first explored the gene expression levels of MHC-II-related genes (Fig. 7a). We noted a global decrease of HLA-II genes (mainly *HLA-DQA2* and *HLA-DRB5*) in patients with severe COVID-19 (Fig. 7a). Similar findings were reported for the validation set

(Extended Data Fig. 7). We then grouped the expression values of these MHC-II genes (*HLA-DRB1*, *HLA-DMA*, *HLA-DQA2*, *HLA-DRB5*, *HLA-DPB1*, *HLA-DQB1* and *HLA-DMB*), constructed a signature that we named the 'HLAII' module, and scored CD1c$^+$ DCs using the 'AddModuleScore' Seurat function. The 'HLAII' signature was significantly reduced in severe as compared to HC CD1c$^+$ DCs (Fig. 7b). This was associated with decreased expression of upstream MHC-II regulators (including *RFX5*, *RFXANK* and *CIITA*) in the severe group (Fig. 7b). Comparison of the scaled values revealed a reduction of *IRF1* and *RFX5* TF activities, mainly described to be involved in MHC-II gene synthesis, whereas C/EBP family member (*CEBPB* and *CEBPD*) TF activities, known to be involved in myeloid fate differentiation, were increased in severe subsets (Fig. 7c). We also noted higher TF activities for *RELA* (NFKB superfamily) and the AP-1 family in patients with severe COVID-19, including *FOSL1*, *FOSL2* and *JUN*, which regulate a large range of cellular processes, including cell survival, death and proliferation (Fig. 7c).

To further decipher the transcriptional changes occurring in DCs when transitioning from healthy to moderate and severe conditions, we conducted pseudo-temporal inference using Monocle3, using the UMAP embedding two-dimensional space of the DC subsets (Fig. 7d). The pseudotime tree revealed a continuous trajectory from healthy to moderate, and a marked transition to the severe subsets. This trajectory was correlated to pseudotime values (Fig. 7d, right). To recover the genes contributing to this transition tree, we conducted a graph-based test to assess the most significant genes. The top genes were associated with Ag presentation, including *B2M* and *HLA-DPA1*, along with genes related to ISG expression (*GABARAP* and *IFITM3*; Fig. 7e).

Given that MHC-II genes are involved in the DC–T cell interaction, we hypothesized that a more global dysfunction of DC–T cell communication may occur in COVID-19 APCs. To test this hypothesis, we applied our cell communication inference computational framework ICELLNET[43]. Using our 'reference partner cell' methodology, we inferred potential communication between each of the APC subsets and CD4$^+$ T cells in each of the disease groups (Fig. 8a). Cell connectivity networks revealed a global decrease in APC–T cell communication in patients with severe, as compared to moderate COVID-19 and HCs, predominantly in CD1c$^+$ DCs, CLEC9a$^+$ DCs and CD14$^+$ monocytes (Fig. 8a,b). We then explored the various molecular families that may explain this decrease. This revealed a dominant contribution of immune checkpoint molecules and cytokines for CD1c$^+$ DC–T cell communication (Fig. 8b), in particular decreased *JAG-NOTCH*, *CD80-CD28* and *CD48-CD2* interactions (Fig. 8c). Cytokines were mostly underlying the decrease in CD14$^+$ monocytes–T cell communication (Fig. 8b). As expected, signalling through HLA-II-related genes (*HLA-II/LAG3* pairs) was significantly decreased in moderate and severe subsets as compared to HCs. Among the cytokines, *IL10*-, *CCL5*- and *TGFb*-mediated interactions were predominantly damped in patients with severe COVID-19, which may contribute to immunopathology through excessive Th1 responses (Fig. 8c).

Persistent defects in severe COVID-19 APCs across samples. We also asked whether functional pathway alterations observed across APC subtypes were sustained over time. In parallel, we wanted to ensure that our main findings were not driven by a single patient and/or time point. We compared scRNAseq datasets generated at day 1 versus day 4 post hospital admission for each patient, in all severity groups. We used a focused approach, by selecting genes involved in previously identified altered functions, in APCs from patients with severe COVID-19, and systematically compared day 1 and day 4 expression levels. Most of the day 1 defects were sustained at day 4, in particular the low score of the HLA-II module in CD1c$^+$ DCs (Extended Data Fig. 2a) together with the decreased expression of the antiviral effector molecules in CD14$^+$ monocytes (Extended Data Fig. 2b) and increased 'apoptosis p53 pathway' in pDCs (Extended Data Fig. 2c). At a patient level, we observed that HC samples displayed similar score (or expression) levels for these three biological processes, whereas both moderate and severe samples displayed slight differences due to inter-individual heterogeneity. Overall, we confirmed that our findings were not associated with either a specific time point or a dominant single patient effect. However, this does not exclude changes in APC molecular profiles at later times in the course of moderate and severe COVID-19.

Discussion

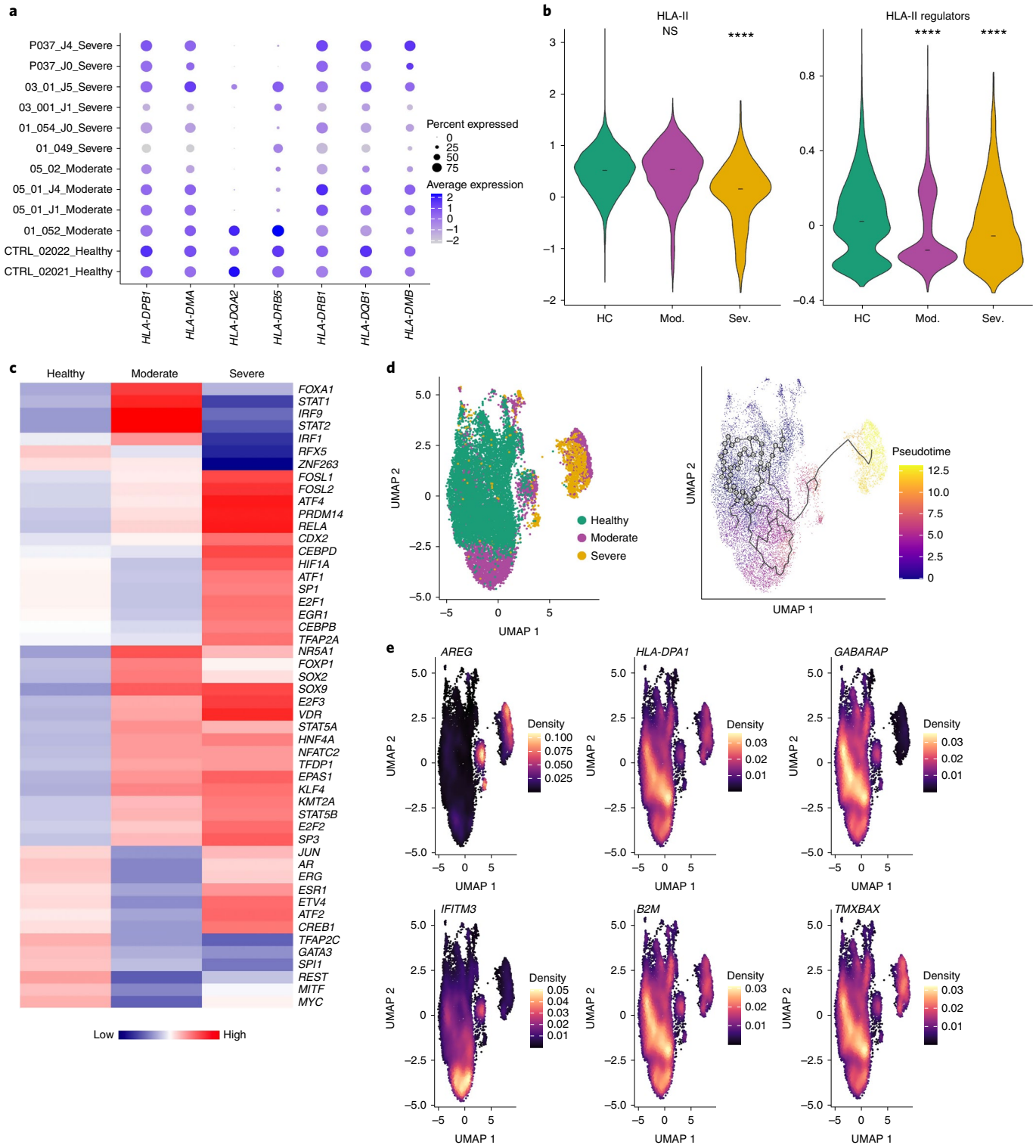
Severe COVID-19 harbours a complex physiopathology stemming from host–pathogen interactions evolving over time, and involves a large number of underlying cellular and molecular mechanisms. Hence, detailed studies on various immune cell compartments are required to obtain a global view of the process. DCs are central to immune responses by linking innate and adaptive immunity, in particular during infection[31]. DCs are rare cell types composed of multiple subsets[32], justifying dedicated studies to uncover putative dysfunctions. So far, very little is known about the role of DC subsets in COVID-19[44–46]. scRNAseq atlas studies of total PBMCs in patients with severe and moderate COVID-19 identified inflammatory monocytes defective for MHC-II molecules[12], as was previously shown in severe sepsis patients[47], and increased apoptosis pathways in both NK cells and monocytes[27,35,48–50]. So far, none of these studies were tailored to provide sufficient resolution into the DC compartments. The challenge is even greater knowing that some DC subsets, such as pDCs and CD141 (CLEC9A)$^+$ DCs, are depleted from the blood in severe COVID-19[45,51]. A recent study analysed PBMCs by scRNAseq, after DC enrichment in EDTA-containing medium, but focused only on the IFN pathway and ISGs[30]. Most of these studies utilized frozen/thawed PBMCs as a starting biological material, potentially inducing loss in some rare DC subsets. Through dedicated enrichment steps performed immediately after blood sampling (fresh samples), we were able to capture sufficient cell

Fig. 7 | Downregulation of MHC-II and upstream transcriptional regulators in severe COVID-19 CD1c$^+$ DCs of the discovery set. a, Dot plot distribution of HLA-II-related genes at the patient level within the CD1c$^+$ DC subset; expression levels are colour-coded, and the percentage of cells expressing the respective gene is size-coded. **b**, Violin plot distribution of HLA-II and the upstream regulators' (HLAII_Regulators) module scores among the three severity groups within the CD1c$^+$ DC subset; severity groups are colour-coded, each dot represents a cell and the horizontal line displays the mean value of the enrichment score of each given pathway. The violin plots were designed using the total CD1c$^+$ DC subsets from the discovery set obtained from $n=2$ HC, $n=4$ moderate and $n=6$ severe samples. Comparative analysis was performed using the two-sided Wilcoxon rank-sum test. *P* values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$; NS, not significant. **c**, Heatmap representation of top 50 highly variable TF activities between CD1c$^+$ DC severity groups; the *z*-score of activity scores is colour-coded. **d**, Pseudotime inference tree on UMAP embeddings (left) of the CD1c$^+$ DC subset using Monocle3; pseudotime values are colour-coded (right). **e**, UMAP representation of density scores for the top genes contributing to the pseudotime tree initially inferred in **d**. Density scores were computed using Nebulosa and are colour-coded. All statistical tests displayed in this figure were performed using the discovery set, comprising a total of 12 samples ($n=2$ controls, $n=4$ moderate and $n=6$ samples) collected from seven patients and two healthy donors.

numbers to define molecular profiles and identify specific defects in all known DC subsets.

As with most immune cells, DCs are not limited to a single function[31]. They play a key role in the first line of immune defence by sensing microbial pathogens, and also contribute to direct pathogen control through the production of antimicrobial peptides and antiviral effector molecules[52]. Other effector functions include the secretion of pro- and anti-inflammatory cytokines, and cytotoxic molecules[31]. Finally, they function as APCs to T cells, with which they communicate through secreted and surface molecules expressed within the immune synapse[53]. By using scRNAseq, and a combination of supervised and unsupervised bioinformatics methods, we were able to uncover defects in almost all of these processes, in specific APC subsets, associated with COVID-19 severity. This provides the first detailed molecular map of DC subsets and underlying molecular pathways in COVID-19.

Several studies have shown an increase of inflammatory cytokines in severe COVID-19, which may contribute to the severity of the disease[44]. Increased circulating levels of IL-1β and IL-6 were detected in patients with severe COVID-19[9–22,44]. However,



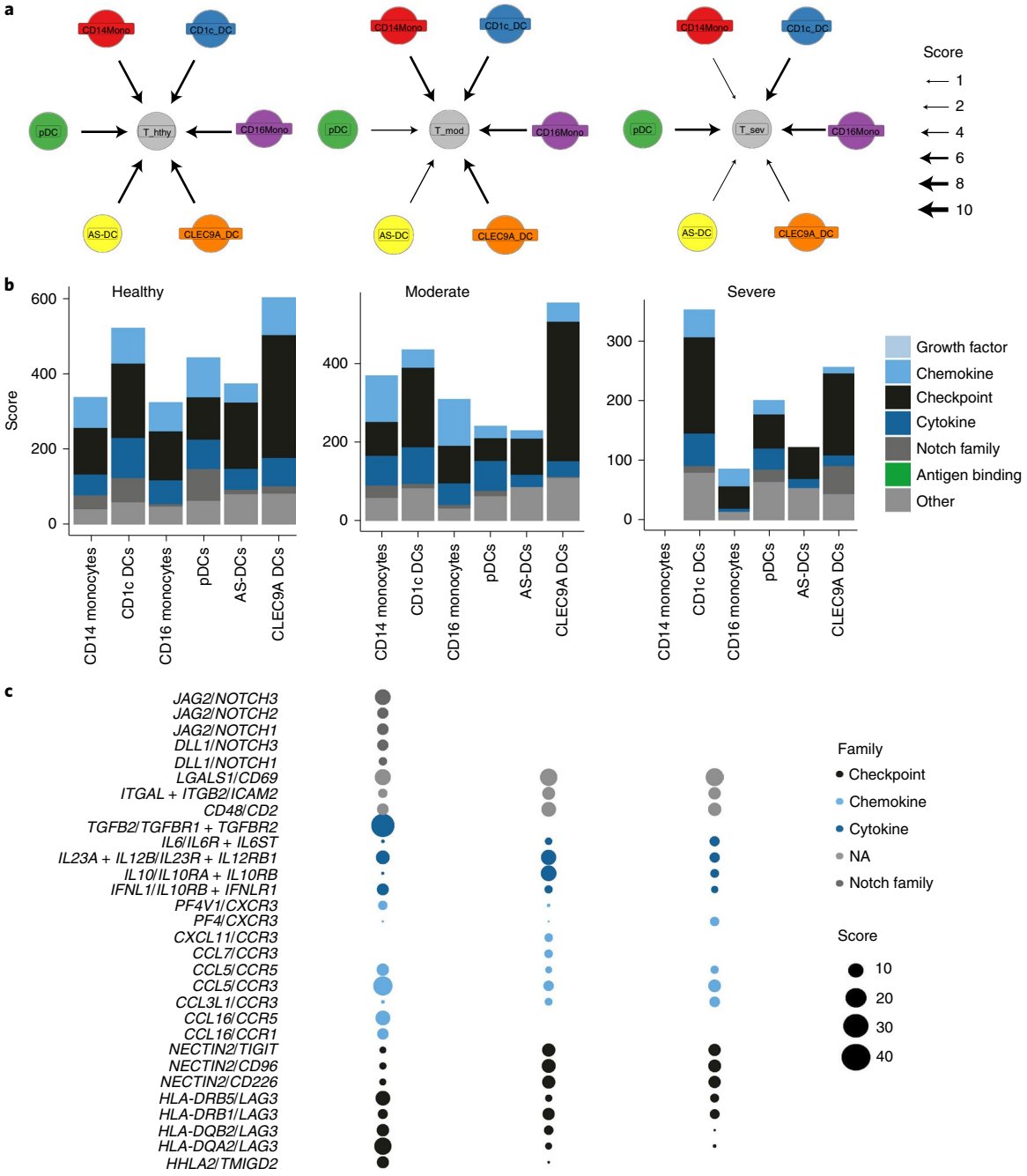


Fig. 8 | Perturbation of DC–T cell communication in patients with severe COVID-19 from the discovery set. a, Connectivity maps describing outward communication from APCs from the single-cell dataset at day 1 to T lymphocytes ($n=39$), according to patient severity (healthy, moderate and severe). The T lymphocyte transcriptomic profiles are from the Human Primary Cell Atlas, included in the ICELLNET R package. For APCs, average cluster gene expression profiles were considered. Only DC–T cell interactions are taken into account to compute the communication score (manually curated, $n=144$ interactions). **b**, Barplot of each communication score with contribution by families of communication molecules for outward communication from APCs to T lymphocytes. **c**, Focus on CD1c$^+$ DC outward communication to T lymphocytes, representing specific individual interaction scores that differ by at least 10 between patients with moderate and severe COVID-19 (cutoff chosen for the purpose of clarity).

the cellular source does not seem to be from circulating cells, but rather from inflammatory monocytes attracted to the lung[54], as well as endothelial cells[23,24]. Our study corroborates these findings for IL-6, with no significant expression detected across APC subsets. However, we did find increased expression of IL-1β, CXCL8 and CXCL2 in APCs at the global level, and this may contribute to systemic inflammation. In parallel, we observed increased TNF signalling in pDCs, but decreased in monocytes, suggesting that distinct APCs may respond differently to circulating inflammatory mediators.

Type I and III IFNs are critical antiviral cytokines[55]. APCs are a central source of IFN following viral sensing. Studies have shown that type I IFN responses are impaired in severe COVID-19[6,9,17,21,28–30,46,48,56], which may contribute to persistent viral load. Our data support these findings, as we did not detect any expression of *IFN-α* and *IFN-λ1* across all APC subsets. However,

we were also able to detect critical defects in the response to type I IFN. First, expression of *IFNAR1* and *2* was globally decreased in APC subsets from patients with severe COVID-19. Second, most downstream ISGs (both antiviral and regulators of IFN signalling) were expressed at lower levels in patients with severe COVID-19 compared to HCs, which themselves are expected to express low levels of ISGs given the absence of innate stimulation. Overall, the IFN pathway was defective in severe COVID-19 APCs at several levels: IFN production, receptor expression and downstream ISG responses.

pDCs are a cell type that is highly specialized in antiviral immunity, producing large amounts of all type I IFN[57]. Circulating pDCs have been shown to be diminished in COVID-19[51], but the underlying mechanisms remain unknown. We identified increased expression of pro-apoptotic molecules in pDCs from patients with severe COVID-19. This suggests that pDCs could be globally altered through increased cell death. In a separate study, we have shown that in vitro SARS-CoV-2 stimulation of pDCs from healthy donors leads to improved pDC survival[58], suggesting that the increased apoptosis we observed in pDCs from patients with severe COVID-19 was not due to direct virus-induced killing. In parallel, we detected several defects in various pDC functions: decreased innate sensing, through loss of *TLR7* and *DHX36*, which are key viral sensors[59], decreased antiviral effector functions and cytotoxicity. Hence, we report multi-process defects affecting key aspects of pDC antiviral functions. Interestingly, a recent study performed ex vivo stimulation of PBMCs of a patient with COVID-19 with TLR7/9 ligands, and showed decreased type I IFN production[30]. This provides an independent functional validation, while our study provides molecular mechanisms, in particular the increased pDC apoptosis and the decrease in *TLR7* expression.

Transcriptomic data, including scRNAseq, allow for the application of methods to infer TF activity, as a way to provide potential upstream mechanisms. We found that several important TF activities were decreased in CD1$^+$ DCs, suggesting defective immune effector functions in patients with severe COVID-19. STAT2 activity downregulation may indicate a deficiency to cross-present to CD8$^+$ T cells and license their cytotoxic function[60]. Subversion of DC immunogenicity by targeting STAT2 was observed in other viral infections. ZIKV evades type I IFN responses by antagonizing STAT2 phosphorylation[61]. The low estimated activity of *ESR1*, *CIITA*, *USF1* and *RFX5* in CD1$^+$ DCs may explain the decrease in MHC-II molecules we observed in patients with severe COVID-19, through decreased trans-activation of the MHC-II promoter[62,63]. Finally, the low activity of *EGR1* and *RUNX1* TF in CD1$^+$ DCs of patients with severe COVID-19 may contribute to an impaired function in CD8 T-cell activation and induction of IFN-γ[64,65]. Collectively, our results suggest that several aspects of CD1$^+$ DC effector functions may be altered through decreased activity of key TFs controlling MHC-II expression and T-cell stimulation.

Our study provides a unique insight into the physiopathology of APCs in severe COVID-19, uncovering previously unknown defects in multiple functional pathways, related to both innate and adaptive immunity. We were able to map molecular pathways in rare DC subsets, many of them previously unexplored in the context of COVID-19. Combined with studies in other anatomical sites[44], in particular the lung[54], and other disease severity stages, our results should contribute to a better understanding of COVID-19 immunopathology. They also open interesting perspectives for clinical applications. Simple molecular markers of defective APC subsets may be explored as prognostic and stratification biomarkers. This hypothesis echoes the immune pathology of bacterial sepsis, for which multiple defects in APCs have already been described[47,66]. A persistent decrease in circulating DCs, as well as monocyte deactivation as assessed by decreased *HLA-DR* expression or decreased *CD74* messenger RNA (mRNA) expression, are already known to be predictive of ICU-acquired superinfections in patients with bacterial sepsis[67,68]. It would be interesting to explore whether such markers, for example pDC apoptosis or CD1c$^+$ DC MHC-II downregulation, appear earlier in the course of COVID-19 and may predict aggravation. From a therapeutic standpoint, many innate adjuvants have been developed to target DC subsets[69,70], and could be considered as personalized immunotherapies depending on patient-specific DC dysfunction[69]. Finally, DCs are being considered in preventive vaccine development (ClinicalTrials.gov: NCT04386252). Ultimately, our study may form the ground for novel therapies to restore defective APC functions in patients with COVID-19.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended 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/s41556-021-00681-2.

Received: 3 September 2020; Accepted: 6 April 2021;
Published online: 10 May 2021

References

1. Wang, F. et al. The laboratory tests and host immunity of COVID-19 patients with different severity of illness. *JCI Insight* **5**, e137799 (2020).
2. Pedersen, S. F. & Ho, Y.-C. SARS-CoV-2: a storm is raging. *J. Clin. Invest.* **130**, 2202–2205 (2020).
3. Yan, R. et al. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
4. Hadjadj, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
5. Olagnier, D. et al. SARS-CoV2-mediated suppression of NRF2-signaling reveals potent antiviral and anti-inflammatory activity of 4-octyl-itaconate and dimethyl fumarate. *Nat. Commun.* **11**, 4938 (2020).
6. Galani, I.-E. et al. Untuned antiviral immunity in COVID-19 revealed by temporal type I/III interferon patterns and flu comparison. *Nat. Immunol.* **22**, 32–40 (2021).
7. Channappanavar, R. & Perlman, S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. *Semin. Immunopathol.* **39**, 529–539 (2017).
8. Ong, E. Z. et al. A dynamic immune response shapes COVID-19 progression. *Cell Host Microbe* **27**, 879–882 (2020).
9. Hadjadj, J. et al. Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* **369**, 718–724 (2020).
10. Sadanandam, A. et al. A blood transcriptome-based analysis of disease progression, immune regulation and symptoms in coronavirus-infected patients. *Cell Death Discov.* **6**, 141 (2020).
11. Laing, A. G. et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat. Med.* **26**, 1623–1635 (2020).
12. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
13. Del Valle, D. M. et al. An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat. Med.* **26**, 1636–1643 (2020).
14. Giamarellos-Bourboulis, E. J. et al. Complex immune dysregulation in COVID-19 patients with severe respiratory failure. *Cell Host Microbe* **27**, 992–1000 (2020).
15. Chua, R. L. et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 (2020).
16. Mangalmurti, N. & Hunter, C. A. Cytokine storms: understanding COVID-19. *Immunity* **53**, 19–25 (2020).
17. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5**, eabd1554 (2020).
18. Jøntvedt Jørgensen, M. et al. Increased interleukin-6 and macrophage chemoattractant protein-1 are associated with respiratory failure in COVID-19. *Sci. Rep.* **10**, 21697 (2020).
19. Lucas, C. et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020).
20. Mann, E. R. et al. Longitudinal immune profiling reveals key myeloid signatures associated with COVID-19. *Sci. Immunol.* **5**, eabd6197 (2020).
21. Blanco-Melo, D. et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* **181**, 1036–1045 (2020).

22. Zhu, L. et al. Single-cell sequencing of peripheral mononuclear cells reveals distinct immune response landscapes of COVID-19 and influenza patients. *Immunity* **53**, 685–696 (2020).
23. O'Sullivan, J. M., Gonagle, D. M., Ward, S. E., Preston, R. J. S. & O'Donnell, J. S. Endothelial cells orchestrate COVID-19 coagulopathy. *Lancet Haematol.* https://doi.org/10.1016/S2352-3026(20)30215-5 (2020).
24. Pons, S., Fodil, S., Azoulay, E. & Zafrani, L. The vascular endothelium: the cornerstone of organ dysfunction in severe SARS-CoV-2 infection. *Crit. Care* **24**, 353 (2020).
25. Schied, A., Trovillion, E. & Moodley, A. Sars-CoV-2 infection in a neutropenic pediatric patient with leukemia: addressing the need for universal guidelines for treatment of Sars-CoV-2 positive, immunocompromised patients. *Pedriat. Blood Cancer* https://doi.org/10.1002/pbc.28546(2020).
26. Somers, E. C. et al. Tocilizumab for treatment of mechanically ventilated patients with COVID-19. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa954 (2020).
27. Guo, C. et al. Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. *Nat. Commun.* **11**, 3924 (2020).
28. Lei, C. et al. Neutralization of SARS-CoV-2 spike pseudotyped virus by recombinant ACE2-Ig. *Nat. Commun.* **11**, 2070 (2020).
29. Zhang, Q. et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **370**, eabd4570 (2020).
30. Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* https://doi.org/10.1126/science.abc6261 (2020).
31. Banchereau, J. et al. Immunobiology of dendritic cells. *Annu. Rev. Immunol.* **18**, 767–811 (2000).
32. Collin, M. & Bigley, V. Human dendritic cell subsets: an update. *Immunology* **154**, 3–20 (2018).
33. Coquerelle, C. & Moser, M. DC subsets in positive and negative regulation of immunity. *Immunol. Rev.* **234**, 317–334 (2010).
34. Liu, Y.-J. IPC: professional type 1 interferon-producing cells and plasmacytoid dendritic cell precursors. *Annu. Rev. Immunol.* **23**, 275–306 (2005).
35. Zhang, J.-Y. et al. Single-cell landscape of immunological responses in patients with COVID-19. *Nat. Immunol.* **21**, 1107–1118 (2020).
36. Yao, C. et al. Cell-type-specific immune dysregulation in severely ill COVID-19 patients. *Cell Rep.* **34**, 108590 (2021).
37. Xu, G. et al. The differential immune responses to COVID-19 in peripheral and lung revealed by single-cell RNA sequencing. *Cell Discov.* **6**, 73 (2020).
38. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
39. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes and progenitors. *Science* **356**, eaah4573 (2017).
40. See, P. et al. Mapping the human DC lineage through the integration of high-dimensional techniques. *Science* **356**, eaag3009 (2017).
41. Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).
42. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
43. Noël, F. et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nat. Commun.* **12**, 1089 (2021).
44. Vabret, N. et al. Immunology of COVID-19: current state of the science. *Immunity* **52**, 910–941 (2020).
45. Zhou, R. et al. Acute SARS-CoV-2 infection impairs dendritic cell and T cell responses. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.3614132 (2020).
46. Zheng, C. et al. Risk-adapted treatment strategy for COVID-19 patients. *Int. J. Infect. Dis.* **94**, 74–77 (2020).
47. Delano, M. J. & Ward, P. A. Sepsis-induced immune dysfunction: can immune therapies reduce mortality? *J. Clin. Invest.* **126**, 23–31 (2016).
48. Yao, H. et al. Molecular architecture of the SARS-CoV-2 virus. *Cell* **183**, 730–738 (2020).
49. Schulte-Schrepping, J. et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* **182**, 1419–1440 (2020).
50. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
51. Sanchez-Cerrillo, I. et al. Differential redistribution of activated monocyte and dendritic cell subsets to the lung associates with severity of COVID-19. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.13.20100925 (2020).
52. Bedoui, S. & Greyer, M. The role of dendritic cells in immunity against primary herpes simplex virus infections. *Front. Microbiol.* **5**, 533 (2014).
53. Guermonprez, P., Valladeau, J., Zitvogel, L., Théry, C. & Amigorena, S. Antigen presentation and T cell stimulation by dendritic cells. *Annu. Rev. Immunol.* **20**, 621–667 (2002).
54. Zhou, Y. et al. Pathogenic T-cells and inflammatory monocytes incite inflammatory storms in severe COVID-19 patients. *Natl Sci. Rev.* **7**, 998–1002 (2020).
55. Broggi, A., Granucci, F. & Zanoni, I. Type III interferons: balancing tissue tolerance and resistance to pathogen invasion. *J. Exp. Med.* **217**, e20190295 (2020).
56. Stanifer, M. L. et al. Critical role of type III interferon in controlling SARS-CoV-2 infection in human intestinal epithelial cells. *Cell Rep.* **32**, 107863 (2020).
57. Vremec, D. et al. Production of interferons by dendritic cells, plasmacytoid cells, natural killer cells and interferon-producing killer dendritic cells. *Blood* **109**, 1165–1173 (2007).
58. Onodi, F. et al. SARS-CoV-2 induces human plasmacytoid predendritic cell diversification via UNC93B and IRAK4. *J. Exp. Med.* **218**, e20201387 (2021).
59. Zhang, Z. et al. DDX1, DDX21 and DHX36 helicases form a complex with the adaptor molecule TRIF to sense dsRNA in dendritic cells. *Immunity* **34**, 866–878 (2011).
60. Xu, J. et al. STAT2 is required for TLR-induced murine dendritic cell activation and cross-presentation. *J. Immunol.* **197**, 326–336 (2016).
61. Bowen, J. R. et al. Zika Virus antagonizes type I interferon responses during infection of human dendritic cells. *PLoS Pathog.* **13**, e1006164 (2017).
62. Mach, B., Steimle, V., Martinez-Soria, E. & Reith, W. Regulation of MHC class II genes: lessons from a disease. *Annu. Rev. Immunol.* **14**, 301–331 (1996).
63. Muhlethaler-Mottet, A., Di Berardino, W., Otten, L. A. & Mach, B. Activation of the MHC class II transactivator CIITA by interferon-γ requires cooperative interaction between Stat1 and USF-1. *Immunity* **8**, 157–166 (1998).
64. Rodríguez-Ubreva, J. et al. Prostaglandin E2 leads to the acquisition of DNMT3A-dependent tolerogenic functions in human myeloid-derived suppressor cells. *Cell Rep.* **21**, 154–167 (2017).
65. Singh, A., Svaren, J., Grayson, J. & Suresh, M. CD8 T cell responses to lymphocytic choriomeningitis virus in early growth response gene 1-deficient mice. *J. Immunol.* **173**, 3855–3862 (2004).
66. Boomer, J. S. et al. Immunosuppression in patients who die of sepsis and multiple organ failure. *JAMA* **306**, 2594–2605 (2011).
67. Grimaldi, D. et al. Profound and persistent decrease of circulating dendritic cells is associated with ICU-acquired infection in patients with septic shock. *Intensive Care Med.* **37**, 1438–1446 (2011).
68. Peronnet, E. et al. Association between mRNA expression of CD74 and IL10 and risk of ICU-acquired infections: a multicenter cohort study. *Intensive Care Med.* **43**, 1013–1020 (2017).
69. Bryant, C. E. et al. Dendritic cells as cancer therapeutics. *Semin. Cell Dev. Biol.* **86**, 77–88 (2019).
70. Saxena, M. & Bhardwaj, N. Turbocharging vaccines: emerging adjuvants for dendritic cell based therapeutic cancer vaccines. *Curr. Opin. Immunol.* **47**, 35–43 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Methods

Patient characteristics and recruitment into the study. Our study is compliant with all relevant ethical regulations regarding research involving human participants.This study was part of the DENDRISEPSIS project, aimed at investigating the functional profiles of APCs in patients with sepsis. The full study protocol can be accessed at https://clinicaltrials.gov/ct2/show/NCT03788772?term=dendrisepsis&cond=sepsis&draw=2&rank=1. The study was approved by the appropriate institutional review board and independent ethics committee (Comité de Protection des Personnes I (CPP), Rouen, France, ref: #2018-A01934-51). We included adult HCs, and patients with PCR-proven COVID-19 pneumonia, within 48 h of admission to ICU or to the pulmonology department from an urban tertiary care centre. Exclusion criteria were the following: haematological malignancy or significant history of bone marrow disease, HIV infection, any immunosuppressive drugs, bone marrow or solid organ transplant recipients, leucopenia ($<1{,}000\,mm^{-3}$) except if due to COVID-19 or pregnancy. With respect to HCs, exclusion criteria were the following: history of inflammatory disease, corticosteroid treatment at any dose and infection symptoms within the previous month. Informed consent was obtained from patients or next of kin. Patients were classified into moderate pneumonia if requiring oxygen supply of $<10\,l.min^{-1}$ and severe pneumonia if requiring invasive mechanical ventilation or oxygen supply of $\geq 10\,l.min^{-1}$. Patients were sampled at admission (day 1) and at day 4. HCs were sampled once. Detailed patient characteristics are provided in Supplementary Tables 1 and 2.

Cell purification. Blood samples (20 ml) were collected from each patient at days 1 and 4 post hospital admission, and from HCs. PBMCs were isolated by centrifugation on a density gradient (Lymphoprep, Proteogenix). After FICOLL (GE Healthcare and Lymphoprep StemCell) gradient centrifugation, total PBMCs were enriched in $CD14^+$ monocytes using human CD14 microbeads (Miltenyi Biotec) for positive magnetic selection according to the manufacturer's instructions. The negative fraction remaining after the positive selection of $CD14^+$ cells was used for pan-DC enrichment employing the EasySep human pan-DC enrichment kit (StemCell Technologies). Total pan-DCs were resuspended with 20,000 $CD14^+$ cells and sent for sequencing. Monocyte and pan-DC enrichments were performed immediately after sampling. To avoid DC–T cell clusters, which often form in DC-enriched preparations, EDTA-containing medium (DPBS 1×, 0.5% EDTA, 1% human serum) was used for sample enrichment in the first set, the 'discovery set'. The latter was composed of a total of 12 samples from two healthy donors, three patients with moderate COVID-19 and four with severe COVID-19 from both day 1 and day 4 time points. The reported results in the main figures (Figs. 1–8) along with Supplementary Figs. 1 and 2 were generated based on the discovery set. Because EDTA can decrease reverse transcription efficiency, we validated the findings derived from the discovery set by using RPMI for all enrichment steps (1640 + Glutamax, 2% BSA, 1% penicillin/streptomycin, 1% sodium puryvate, 1% minimum essential medium–non-essential amino acids) in a second set of samples, including a total of 15 samples, defined as the 'validation set'. The latter included two healthy donors, two patients with moderate COVID-19 and six with severe COVID-19. The results of the validation set are presented in Supplementary Figs. 3–7. All main findings were validated in both datasets.

Preparation and isolation of single-cell suspensions. Cell suspensions were subjected to gel bead emulsion using the Chromium 10X Genomics controller according to the manufacturer's guidelines. To perform scRNAseq after cDNA amplification, the concentration of each sample was measured using a Tapestation 2200 system (Agilent). To prepare the cDNA libraries for the 10X Genomics Chromium controller, we used the single-cell 3′ v3.1 kit. Quality control libraries were performed using the Tapestation 2200 (Agilent). An Illumina Novaseq6000 system (100-cycle cartridge) with a sequencing depth of at least 50,000 reads per cell was used for sequencing. The input number of cells was estimated at 20,000 cells per sample.

Quality control and pre-processing of expression matrices. The raw scRNAseq fastq files were processed using Cell Ranger 3.1.0 from 10X Genomics Technology and aligned to the Grch38 reference genome. Bam files and filtered expression matrices were generated using 'cellranger_count'. All expression matrices were loaded into R 4.0.0 using the 'Read10X' function from the Seurat library (https://github.com/satijalab/seurat) version 3.1.5. The latter library was used to perform the analysis.

Pre-processing steps were applied to remove genes expressed in fewer than 20 cells, and to remove cells with fewer than 50 genes or displaying more than 50% mitochondrial transcripts. To minimize technical confounding factors related to the sequencing steps, we evaluated the violin plot distribution of the number of unique molecular identifiers (nUMI), along with the total number of detected genes (nFeatures) per cell for all samples. Two upper cutoffs of 6,000 and 50,000 were manually set for the nUMi and nFeatures, respectively, for each sample. These quality control metrics filtered out low-quality cells. Normalization to 10,000 reads, centering and scaling were sequentially applied on the expression matrices to correct for the sequencing depth variability. To reduce the computational time for sample integration, we filtered out cells from cell types other than APCs. Cell type annotation is detailed in the section 'Manual annotation of cell types'). To decipher specific alterations occurring in each specific APC subset, we separately subclustered each cell subtype, scaled the data and applied graph-based clustering to obtain cell clusters. Genes encoding for immunoglobulins were removed before performing the subclustering step for each cell type to get rid of ambient RNA.

Integration of individual cell matrices into a merged expression matrix from all the samples. To allow comparison across severity states, we integrated the whole expression matrices from all the samples using the Harmony algorithm. Integration anchors, retrieved from the first 50 principal components using the 'FindIntegrationAnchors' Seurat function, were then used to integrate the datasets using the 'IntegrateData' function. This crucial step added an 'integrated' assay to the Seurat object. Scaling and principal component analysis dimension reduction were performed on the integrated assay with 50 principal components. High-resolution (resolution = 0.8) graph-based clustering and UMAP dimension reduction were conducted to retrieve and visualize cell clusters. ICA dimension reduction was specifically performed for $CD14^+$ monocytes, using 30 dimensions.

Manual annotation of cell types. Cells were manually annotated based on their expressing levels of their respective set of cell-type markers, defined as 'cell-type signatures'. For each cell-type signature, enrichment scores were computed using the 'AddModuleScore()' function per cell with 100 randomly selected control genes, split on 25 bins. Each cell cluster was annotated with a particular cell type if its signature score median value was >0. Cell-type signatures included the following: pDCs, expression of ('TCF4', 'CLEC4C', 'IRF7', 'IRF8', 'LILRA4', 'IL3RA', 'TLR9', 'SPIB'), cDCs ('ANPEP', 'CD1C', 'ITGAX', 'CST3', 'FCER1A'), monocytes ('CD14', 'FCGR1A', 'S100A12', 'FCGR3A', 'MS4A7', 'LYZ', 'CXCR3'), AS-DCs ('AXL', 'SIGLEC6', 'CD22'), NK cells ('NCAM1', 'FCGR3A', 'GNLY', 'XCL1', 'XCL2', 'NCR1', 'NKG7'), T cells ('CD3D', 'CD3E', 'CD3G'), B cells (CD19', 'MS4A1', 'CD79A', 'CD79B'), plasma cells ('IGHG2', 'IGHG1', 'IGLC2', 'IGHA1', 'IGHA2', 'IGHA3', 'JCHAIN', 'IGHM', 'XBP1', 'MZB1', 'CD38', 'IGLL5'), erythrocytes ('HBB', 'HBA1') and platelets (PPBP). Cells that were annotated as non-APC were discarded for each sample, before integration, to avoid high computational load during the integration step. For monocytes and cDCs, a subsequent classification of cells was performed according to their expression levels of monocytes and cDC subset markers (CD14 and FCGR3A for monocytes, CD1C and CLEC9A for cDCs).

Statistical analysis. Differential expression analysis between severity groups was performed using the 'FindAllMarkers' Seurat function, using the MAST test and a cutoff set to log FC > 0.3 to filter out false-positive DEGs. We regressed out the 'gender' confounding factor by using the 'MAST' test for comparative analysis and precising 'gender' as a latent variable. This 'gender' variable was added in the metadata slot for each cell from the discovery and validation sets: a cell is annotated as from a 'female' sample if the expression level of the *XIST* gene is higher than 0.1, otherwise the gender is annotated as 'male'. *P* values were corrected using the Bonferroni correction method. We only tested genes that were detected in a minimum fraction of 10% of each severity group. Median values of violin plot distributions of either gene expression levels or pathway-enrichment scores were compared using a Mann–Whitney–Wilcoxon ranked test, taking as a reference the HC. Note that the statistical calculations for the violin plot distributions are derived from the cell count in expression values/enrichment scores comparisons.

Pathway enrichment analysis. Pathway enrichment analysis was performed to seek for the perturbed or enriched pathways in severity groups, as compared to the HCs. Human MsigDB hallmark signatures (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) were loaded into the R session using the 'msigdbr' library version 7.0.1, and the category was set to 'H' for 'human'. The enrichment test was performed using the 'enricher' function from 'ClusterProfiler' version 3.16.0. Msig Database hallmark signatures were given as input to the 'enricher' function. The *P* values were corrected using the Bonferroni correction method. Encoding genes for each enriched pathway were extracted and used as the module to construct a 'pathway-score' signature using 'AddModuleScore' from the Seurat library.

TF activity inference. We sought to decipher the variation of TF activity between severity groups within particular cell types to avoid capturing differentially active TFs related to lineage markers. The Dorothea (https://saezlab.github.io/dorothea/) resource was used to infer TF activity. In this context of single-cell-level resolution, we constructed regulons based on the mRNA expression levels of each TF from a manually curated database, along with the expression level of its direct targets. In this context, TF activity is considered as a proxy of the transcriptional state of its direct targets. We created TF regulons using the 'dorothea_regulon_human' wrapper function from 'dorothea' library version 0.99.10, and chose 'A' and 'B' high-confidence TF selection. Viper scores were computed on the dorothea regulons, scaled and added as the 'Dorothea' slot on the integrated Seurat object. To allow comparison of TF score activities, mean and standard deviation values of the scaled viper scores were computed per severity group. TFs were ranked according to the variance of their corresponding viper scores. The top 50 highly variable scores per severity group ($n = 150$ TFs in total) were kept for visualization of their corresponding scores.

Manual construction of functional signatures. To evaluate the dysregulations occurring at the functional level for each APC subset from patients with COVID-19, we established a manually curated list of effector genes involved in specific APC functions: 'attraction', 'antiviral effector molecules' and 'cytotoxicity'. The signature construction relied on a thorough mining of existing literature, using a combination of MeSH terms and keywords on the PubMed search tool. Each selected molecule was considered an 'effector' of the related function if there was at least one experimental proof in a human setting. Overall, we outlined 12 'cytotoxicity' effector molecules, 29 'antiviral' effector molecules and 18 'attraction' effector molecules. 'Innate sensing' effectors included 13 genes (*DDX58*, *DHX58*, *CGAS*, *IFI16*, *AIM2*, *IRF3*, *TMEM173*, *NLRP3*, *PYCARD*, *TLR7*, *TLR9*, *DHX9* and *DHX36*), and were from refs. [71,72]. Both 'regulators of interferon signalling' and 'antiviral ISG' were implemented by literature mining from ref. [73].

Drop-out correction. To allow drop-out correction and imputation of missing values, we used Nebulosa (https://github.com/powellgenomicslab/Nebulosa) to represent density-based values on UMAP embeddings. This R package is designed to visualize features from single cells, using a kernel density estimation. It recovers the signal by incorporating the similarity between cells, allowing a convolution of the cell features. For pDCs from the discovery set, we specifically added a 'MAGIC_RNA' slot to the Seurat object using MAGIC[74] and specifically plotted the violin distribution of imputed values in Fig. 4e.

Pseudotime inference. For the CD1c$^+$ DC subset, we specifically computed pseudotime inference using Monocle3 (https://cole-trapnell-lab.github.io/), directly available using the Seurat Wrappers R package[75].

Analysis of intercellular communication networks. Communication scores were generated using the ICELLNET R package (https://github.com/soumelis-lab/ICELLNET/master). This library allows computation of cell–cell communication scores between cell subsets, given their corresponding transcriptomic profiles from the same or different datasets. Considering severity groups separately, only clusters including more than 15 cells were considered for the analysis. The average gene expression profiles of APC subset clusters were provided as input to the ICELLNET package, to compute communication scores between APC subsets and T lymphocytes for each severity group. As our datasets did not include T cells for the analysis, we used as reference the T-lymphocyte transcriptomic profile from the Human Primary Cell Atlas included in the ICELLNET package ($n = 39$). From the ICELLNET ligand–receptor interaction database, we only selected the 144 interactions known to be involved in DC–T communication[76]. Barplot and dot plot representations were generated to compare the proportions of communication type scores (checkpoint, cytokines, chemokines) among severity groups.

Statistics and reproducibility. Statistical analysis was performed using R (version 4.0.0). A two-sided Wilcoxon ranked-sum test was used to perform pairwise comparisons. To ensure the reproducibility of our main findings, we split our data analysis cohort into a discovery and a validation set. We reported our main findings from the discovery set and conducted similar analyses on the validation set.

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

Data availability

scRNAseq data that support the findings of this study have been deposited in the Gene Expression Omnibus under accession code GSE169346. Further information and requests for resources and reagents should be directed to and will be fulfilled by the V.S. This study did not generate new unique reagents.

Code availability

The R codes are publicly available on GitHub at https://github.com/MelissaSaichi/Covid_scRNAseq. All of the R packages that were used are available online.

References

71. Zhang, Z., Yuan, B., Lu, N., Facchinetti, V. & Liu, Y.-J. DHX9 pairs with IPS-1 to sense double-stranded RNA in myeloid dendritic cells. *J. Immunol.* **187**, 4501–4508 (2011).
72. Gaidt, M. M. et al. The DNA inflammasome in human myeloid cells is initiated by a STING-cell death program upstream of NLRP3. *Cell* **171**, 1110–1124.e18 (2017).
73. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu. Rev. Immunol.* **32**, 513–545 (2014).
74. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
75. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
76. Grandclaudon, M. et al. A quantitative multivariate model of human dendritic cell-T helper cell communication. *Cell* **179**, 432–447.e21 (2019).

Acknowledgements

We thank our patients for participating in this study. We also acknowledge N. Marin, M. Cojocaru, N. Carlier, J. Marey, D. Monnet and T.-A. Szwebel (all from Cochin Hospital) for their help with the inclusion of patients and controls. We thank A. Mohammed for performing the sequencing runs on Cell Ranger at the sequencing platform of the Institut du Cerveau et de la Moelle épinière (ICM). We thank M. Blot, F. Niedergang and E. Lauret from the Institut Cochin for their help during the setting up of the study. The COVID-19 emergency plan from Université de Paris supported this study regarding technical aspects for the efficient processing of sample sequencing. This study was supported by the ANR DENDRISEPSIS (ANR-17-CE15-0003), ANR APCOD (ANR-17-CE15-0003-01), Fast Grant for COVID-19 from the Mercatus Center, Université de Paris PLAN D'URGENCE COVID19 and Fund 101 grants. L.M. was supported by a PhD fellowship from La Ligue Contre le Cancer and E.A. by a PhD fellowship from Servier. We thank Fast Grant for COVID-19 from the Mercus Center for supporting F.N. and la Fondation pour la Recherche Médicale for supporting J.M.

Author contributions

V.S., J.M. and F.P. designed the study and defined the patient selection criteria. F.P. and Z.A.H. recruited the patients and healthy donors. M.Z.L., S.K. and C.R. performed the wet lab experiments. M.S. designed a workflow for the scRNAseq, analysed the data and generated the results and figures. L.M. performed the cell–cell communication analysis using ICELLNET. E.A. constructed the manually curated effector molecule signatures. F.N. helped with the bioinformatics analysis. D.B. and Y.M. performed and supervised the library construction and sequencing. V.S., M.S., M.Z.L., S.K. and F.P. wrote the manuscript. S.K. updated the bibliography. All authors provided feedback for the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

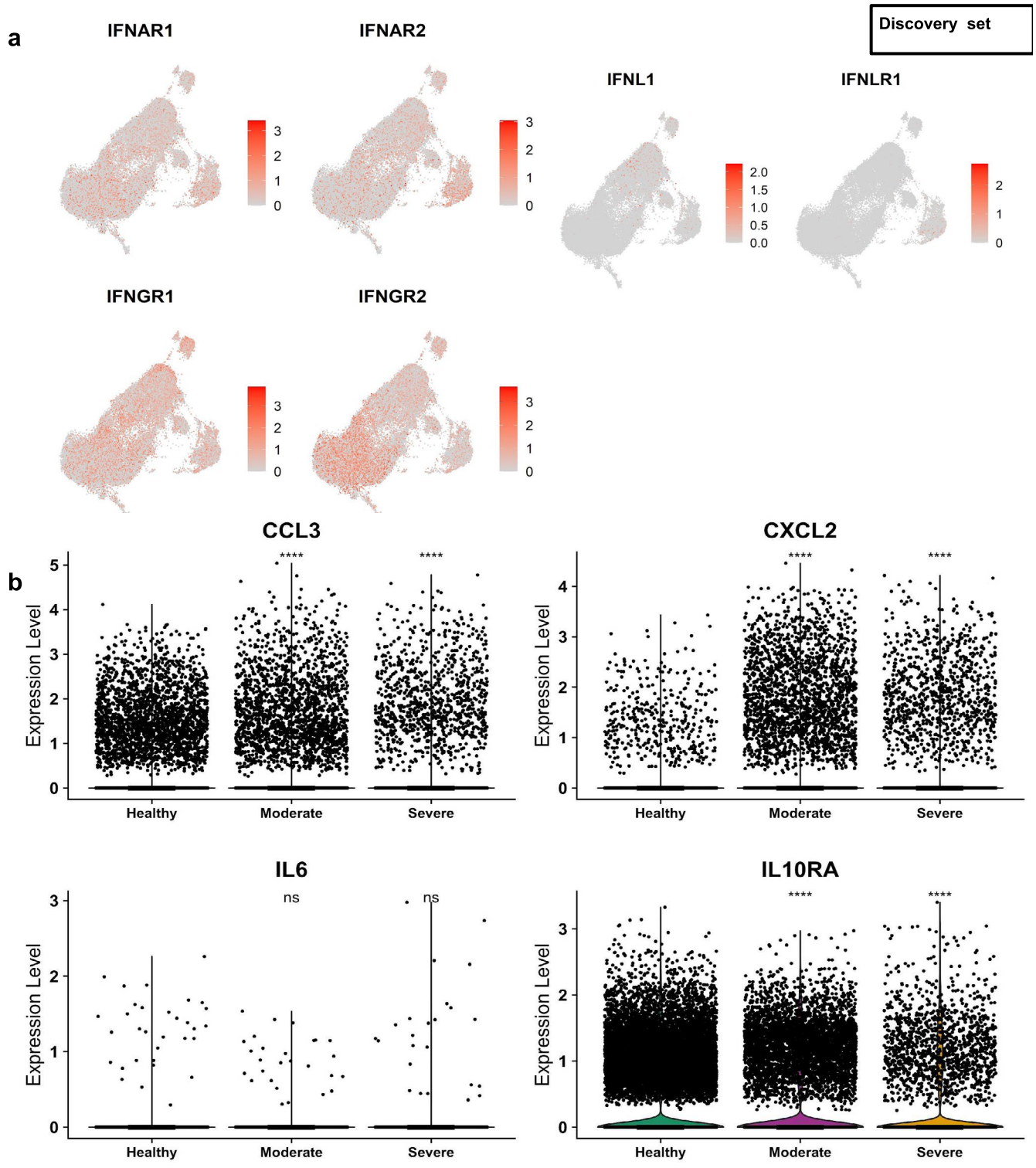
Extended data is available for this paper at https://doi.org/10.1038/s41556-021-00681-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41556-021-00681-2.

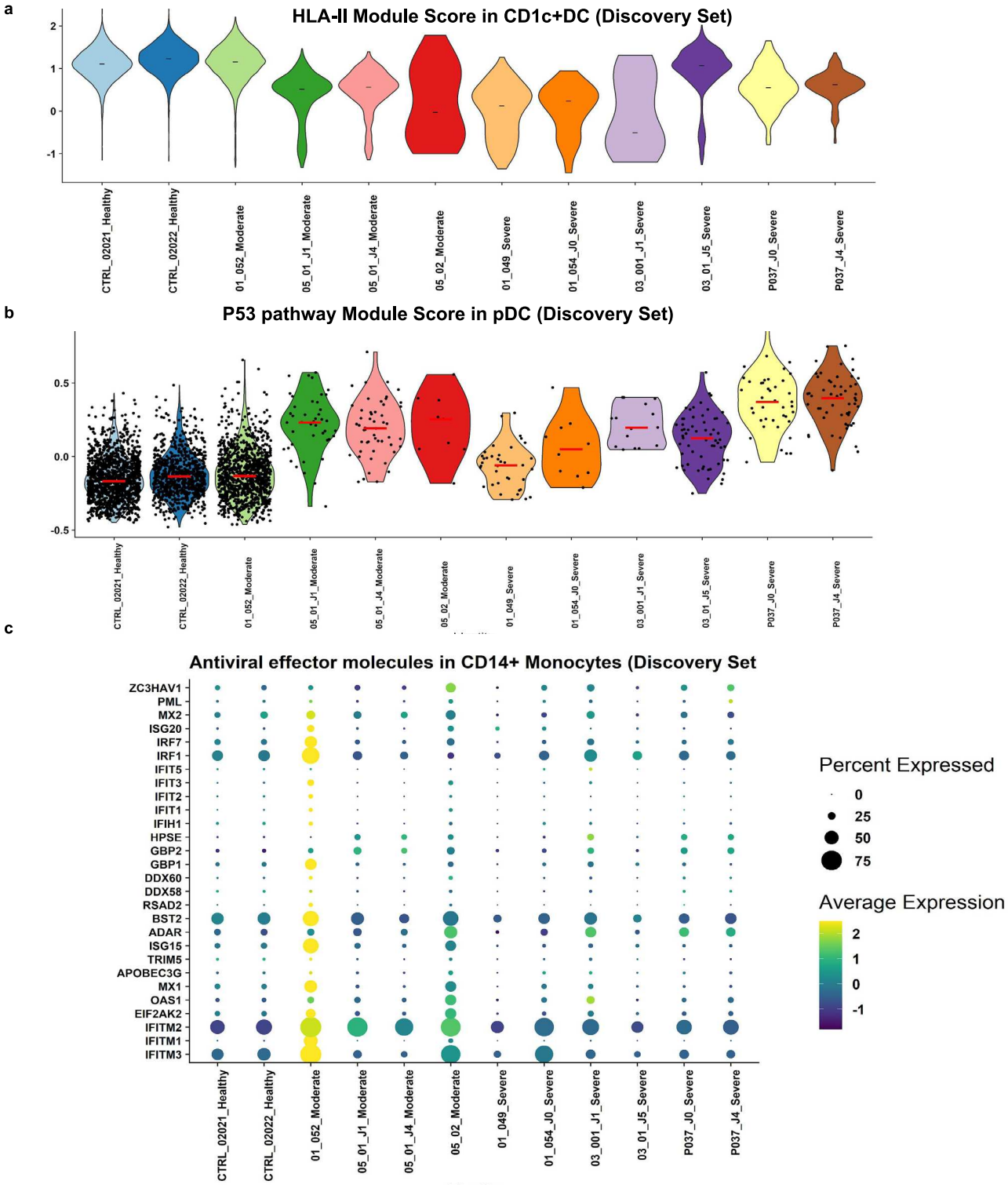
Correspondence and requests for materials should be addressed to V.S.

Peer review information *Nature Cell Biology* thanks Bryan Williams and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Pro-inflammatory defects in the discovery set. a, Umap representation of IFN subtypes expression values in the discovery set, expression levels are color coded; **b**, Violin representation of other pro-inflammatory cytokines in the discovery set which included: n=2 HC, n=4 moderate and n=6 severe samples; each dot represents a cell, horizontal lines display the mean expression value; Comparative analysis was performed using the two-sided Wilcoxon Rank-Sum test, P-values were adjusted to multiple testings using 'Bonferroni' correction. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.



Extended Data Fig. 2 | Maintenance of dysregulated patterns at time and patient levels. Violin plot representation of **a**. HLA-II Module Score in CD1c+DC, and **b**. P53 pathway Module Score in pDC, both from the discovery set (composed of: n=2 HC, n=4 moderate and n=6 severe samples); asterisks above moderate indicate significance of moderate versus control and asterisks above severe indicate significance of severe versus control. Comparative analysis was performed using the two-sided Wilcoxon Rank-Sum test, P-values were adjusted to multiple testings using 'Bonferroni' correction. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **c**, Dot Plot representation of antiviral effector molecules in CD14+ Monocytes across patients, Percentage of cells expressing the respective gene is size-coded.

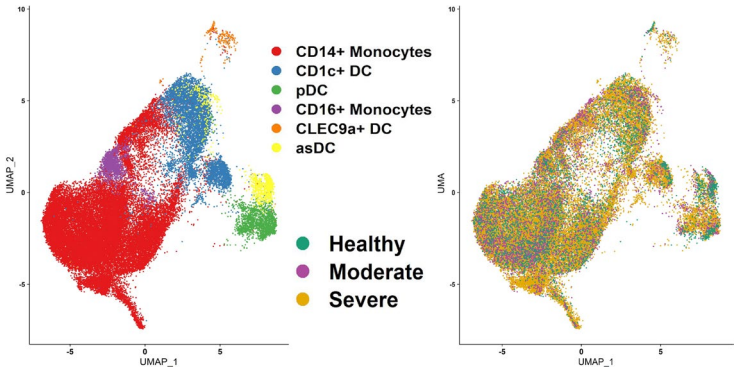
a

All sets

	EDTA (Discovery Set)	RPMI (Validation Set)
CD14+ Monocytes	22690	29409
CD1c+ DC	13252	5754
pDC	3538	1602
CD16+ Monocytes	866	1021
CLEC9a+ DC	1754	197
asDC	684	876

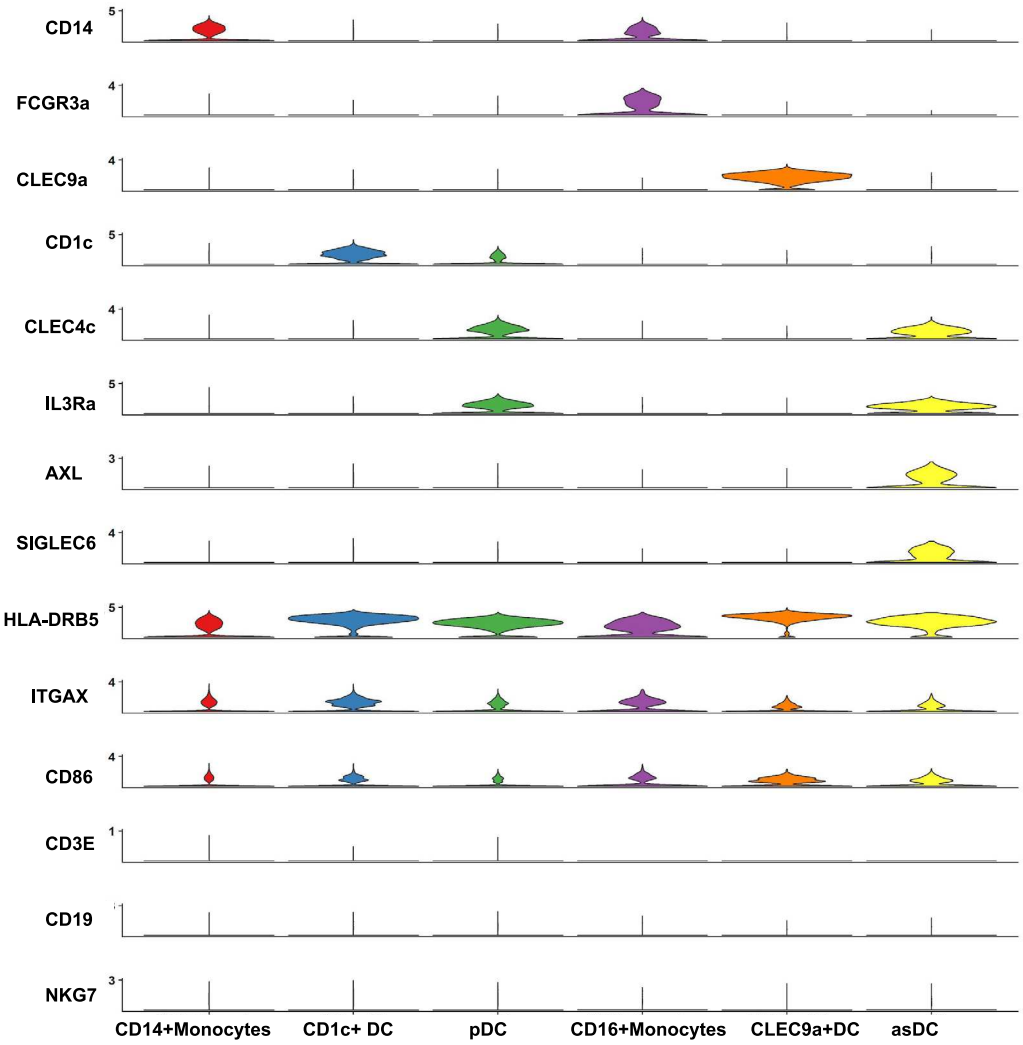
b

Validation set

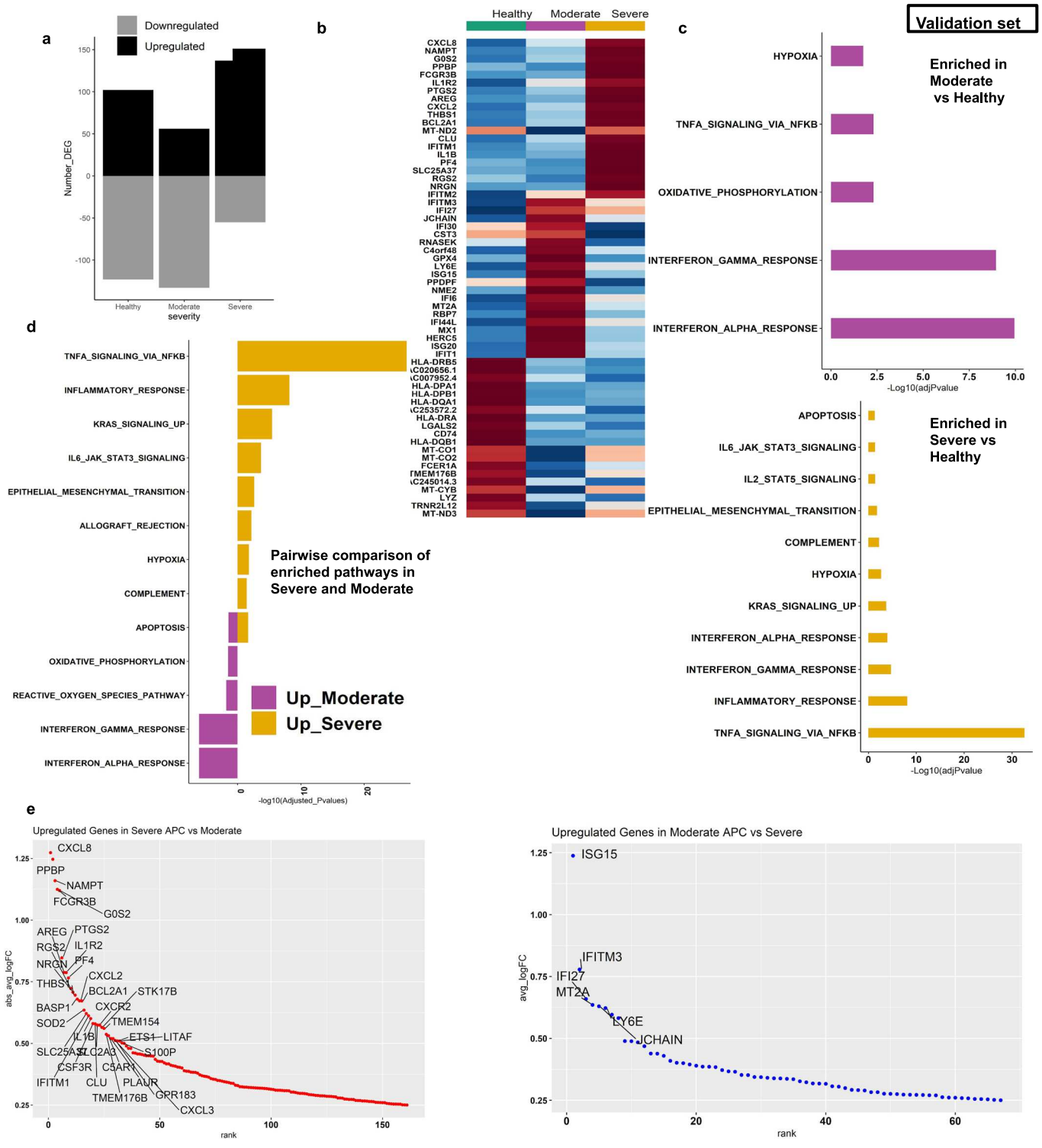


c

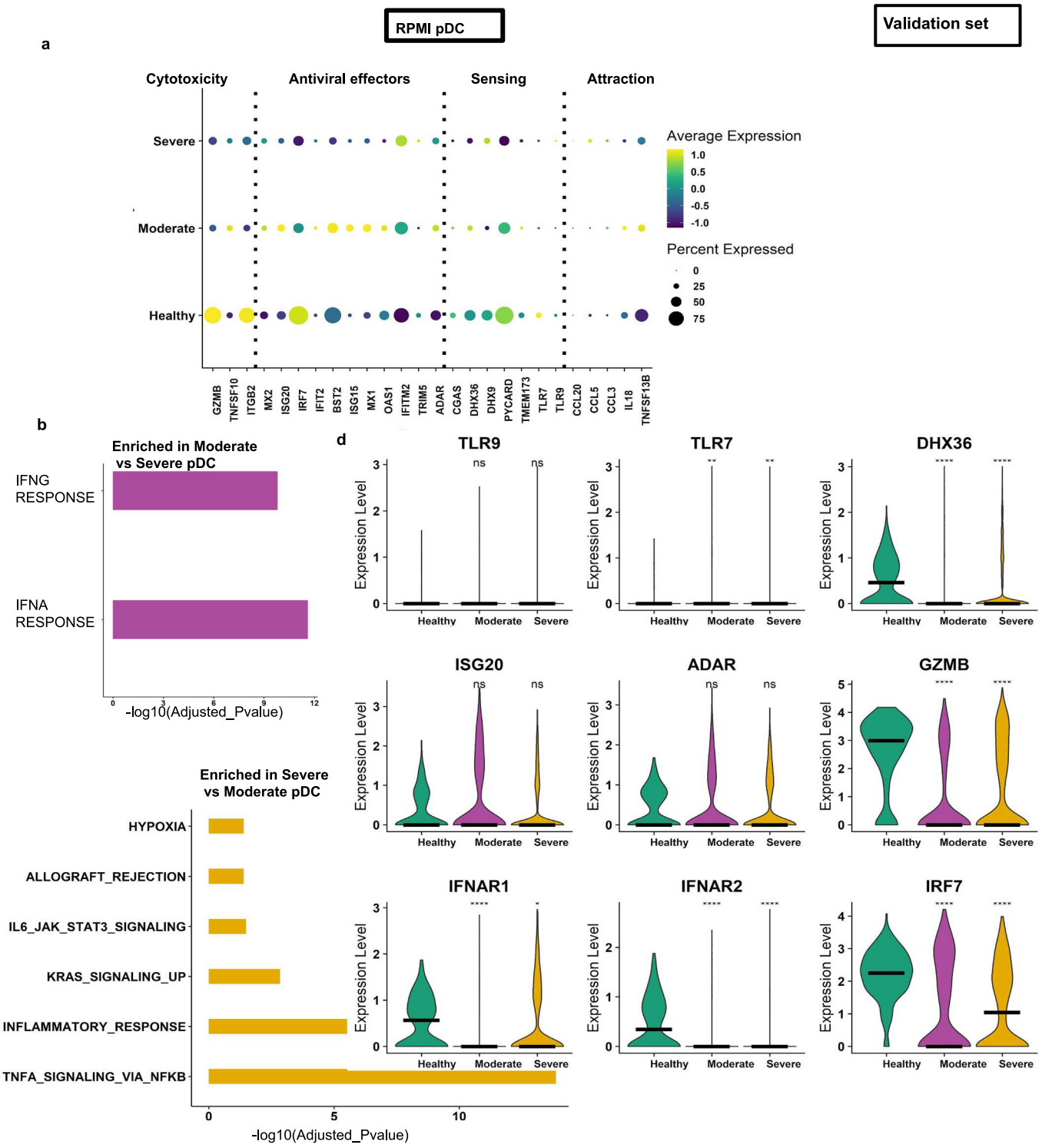
d



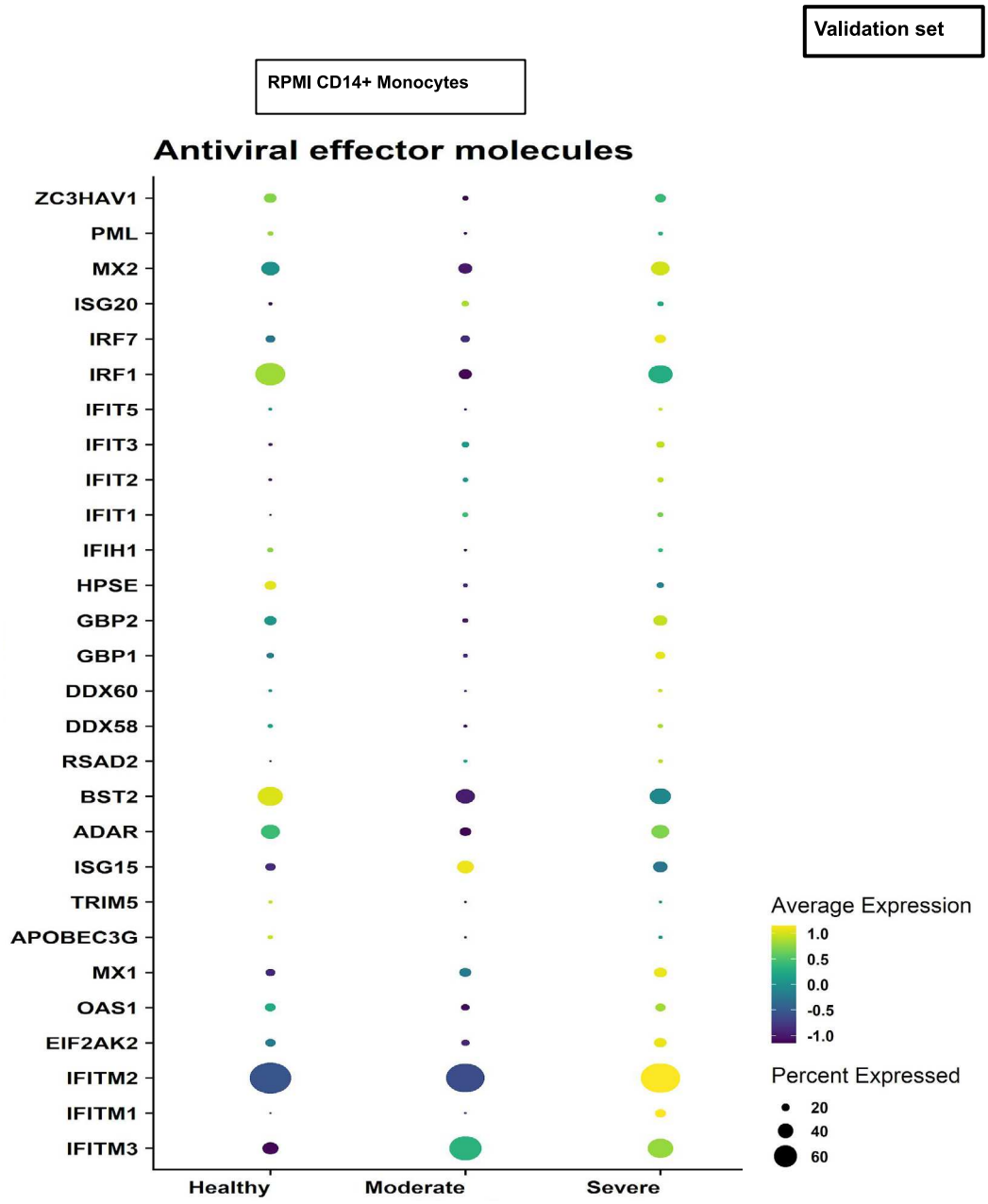
Extended Data Fig. 3 | Cellular map of APC subsets from the validation set. Cellular map of APC subsets at the single-cell resolution level from the validation set based on either APC subsets (**a**) or severity (**b**), **c** Proportions of APC subsets within the discovery and validation sets; **d**. Stuck Violin plot representation of canonical APC and non-APC markers for both discovery and validation sets. Validation set included: n=2 HC, n= 4 moderate and n=9 severe samples from a total of 2 healthy donors, 2 moderate and 6 severe patients.



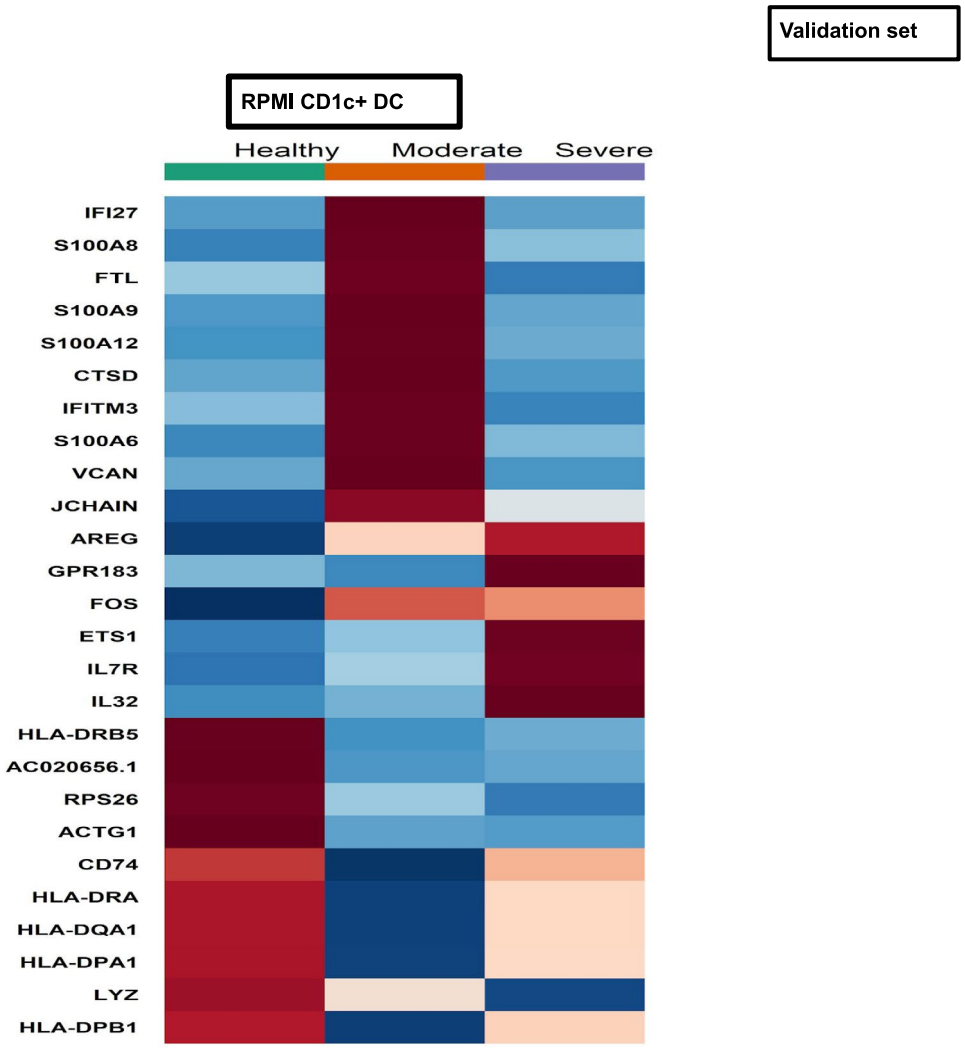
Extended Data Fig. 4 | Increased inflammatory pathways in APC validation set. a. Barplot of the number of Differentially Expressed Genes (DEG) for each severity group (Healthy versus moderate and severe patients; moderate versus healthy and severe; severe versus healthy and moderate). Up-regulated (logFoldChange > 0.25) genes are shown in black, down-regulated (logFoldChange < −0.25) genes are shown in grey; **b**. Heatmap representation of top up-regulated genes in severe APC from the validation set, as compared to moderate and healthy groups, z-score values of average expression levels of cells per severity group is color coded; Comparative analysis of enriched pathways from the upregulated genes in moderate or severe (**c**) APC as compared to healthy cells, as well as pairwise comparison of upregulated genes in moderate compared to severe (shown in pink), up-regulated genes in severe compared to moderate (shown in yellow) (**d**); horizontal axis displays the adjusted p-values (-log10), **e**. Representation of ranked genes by descendant order according to their absolute log Fold Change (log FC), upregulated in moderate as compared to severe (plot in red), upregulated in severe APC as compared to moderate (plot in blue).Top genes, with an absolute value of logFC above 0.5 are shown. Validation set included: n=2 HC, n= 4 moderate and n=9 severe samples from a total of 2 healthy donors, 2 moderate and 6 severe patients. The two-sided Wilcoxon Rank-Sum test was used for comparison, P-values were adjusted to multiple testing using 'Bonferroni' correction; and only genes with adjusted-P Values < 0.05 were considered.



Extended Data Fig. 5 | Global defects in pDC-related functions in the validation set. a. Dot plots of pDC-related functions 'Attraction','Innate sensing', 'Anti-viral effector molecules', 'Cytotoxicity' in pDC from HC, moderate and severe patients in the validation set. Expression levels are color-coded; Percentage of cells expressing the respective gene is size-coded, **b**. Comparative analysis of enriched pathways from the upregulated genes in moderate versus severe pDC (in pink),up-regulated genes in severe compared to moderate (shown in yellow); **c**. Violin plot representation of gene expression for IFN receptors (IFNAR1 and 2), IRF7, and anti-viral effector molecules. Asterisks above severe indicate *P* values for severe versus control; asterisks above moderate indicate significance of moderate versus control. Statistical tests were performed using the validation set, including: n=2 HC, n= 4 moderate and n=9 severe samples; Comparative analysis was performed using the two-sided Wilcoxon Rank-Sum test, P-values were adjusted to multiple testings using 'Bonferroni' correction. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.



Extended Data Fig. 6 | Defective anti-viral properties in CD14+ monocytes and CD1c+DC. Dot plots of 'antiviral effector molecules' in CD14+ monocytes from HC, moderate and severe patients in the validation set. Expression levels are color-coded; Percentage of cells expressing the respective gene is size coded.



Extended Data Fig. 7 | MHC-II antigen presentation defects in CD1c+DC. Heatmap representation of top 10 DEG (upregulated) for each severity group in CD1C+DC from the validation set.

Chapter 10

Meta-analysis of human cancer single-cell RNAseq datasets using the fully integrated IMMUcan database

We built a comprehensive single-cell RNAseq database for human cancer datasets. Apart from the utility of such a database, that groups in a single environment all the datasets that we could collect as well as the metadata, we show how it can be exploited with 2 use cases.

Meta-analysis of human cancer single cell RNAseq datasets using the fully integrated IMMUcan database

Authors

Jordi Camps[1], Floriane Noël[2], Robin Liechti[3], Lou Götz[3], Caroline Hoffmann[4], Lucile Massenet-Regad[2], Elise Amblard[2], Melissa Saichi[2], Mahmoud M Ibrahim[1], Jack Pollard[5], Jasna Medvedovic[2], Helge Gottfried Roider[1], Vassili Soumelis[2,6]

Affiliations

1 Bayer AG, Berlin, Germany

2 INSERM U976

3 Vital-IT group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

4 Institut Curie

5 Sanofi

6 APHP

Abstract

The development of single cell RNA-sequencing (scRNAseq) technologies has greatly contributed to deciphering the immune tumor microenvironment (TME) landscape, and a wealth of biological data is now publicly accessible. This represents a very valuable resource to researchers in the field, offering a reference for comparison of novel results, as well as opportunities for original meta-analysis studies. However, the massive amount of biological information renders its exploitation difficult in the absence of a well-structured and annotated resource. Marked heterogeneity and variability between studies in terms of cancer type, clinical context, technological platform, data quality, number and type of cells, create additional bottlenecks. We have developed a fully integrated scRNAseq database exclusively dedicated to human cancer. It gathers 119 studies on 45 different cancer types, annotated in 43 fields containing precise clinical, technological and biological information. We developed an original data processing pipeline organized in 4 steps: 1) data collection, 2) data processing (quality control, sample integration, cell clustering), 3) cell ontology tree of the TME) built and used to annotate the clusters in a supervised and manual manner, and 4) interface to analyze TME in a cancer type-specific or global manner. This integrated, accessible and user-friendly resource should be of great value to the biomedical community. It represents an unprecedented level of detailed annotation, offering vast possibilities for downstream exploitation of human cancer scRNAseq data for discovery and validation studies. The database is freely accessible at: https://immucanscdb.vital-it.ch.

Introduction

Tumor immunology has taken central stage due to the success of immunotherapy in a large number of clinical indications. However, deciphering the complexity of the tumor microenvironment (TME) remains an important challenge. It could help improve our knowledge of the cellular and molecular events taking place across tumor types, stages and anatomical location. It is also critical in order to further improve the efficiency and applications of current immunotherapies, alone or in combination, as well as developing novel strategies such as personalized medicine.

Single cell RNA sequencing (scRNAseq) technologies appeared as a unique way to explore the diversity of cellular phenotypes and underlying molecular pathways in a broad and unbiased manner. Their application to cancer studies has progressively increased over the years, from 1 study published in 2014 to 36 studies in 2020. This was facilitated by the commercial availability and standardization of the technology[1], as well as the development and validation of a large number of data analysis tools dedicated to scRNAseq[2,3]. Important biological discoveries were made through these approaches, including the characterization of the partial epithelial-to-mesenchymal transition (p-EMT) program in HNSCC[4], the identification of the cancer program involved in the resistance to checkpoint blockade immunotherapy in melanoma[5], and the identification of functional T cells states in lung cancer[6].

A large number of human cancers scRNAseq datasets have been published as resource and original research articles. This represents a major opportunity for biomedical discoveries, given that in most published articles the authors have addressed a limited number of hypotheses using a selected array of analysis methods. However, the wealth of datasets that have been generated is characterized by very important diversity and heterogeneity at several levels: 1) tumor types and clinical context, 2) technology and experimental protocols, 3) data analysis methods, 4) biological and clinical interpretation of the results. This creates major difficulties and bottlenecks in the exploitation of those datasets by independent teams in order to explore their own hypotheses and biomedical questions. As a result, a large amount of information remains unexploited and hardly accessible.

In order to facilitate the accessibility and reanalysis of scRNAseq datasets, institutional and collaborative initiatives led to the development of data resource solutions.

A number of scRNAseq data portals are available, including scRNASeqDB[7], SCPortalen[8], PanglaoDB[9], and JingleBells[10]. Such portals are useful in order to retrieve single cell studies according to specific search terms, across species and biological questions, in different types of diseases. Limited numbers of cancer-related studies are included, from cancer cell lines to animal models and human disease. There is no detailed annotation in relation to the technologies used, and most importantly to the large number of clinical features characterizing each dataset. This limits the applications and possibilities for meta-analysis in cancer biomedical and translational research.

CancerSEA and TISCH are the only two databases that we know of being cancer-specific. CancerSEA has focused on the identification of functional states, associated to specific gene signatures, based on 41 900 cancer single cells from 25 cancer types[11]. It combines datasets from human tumors, but also from cell lines and patient-derived xenografts (PDX). Clinical annotation is minimal, restricted to tumor type description. TISCH enables to browse through cancer scRNAseq datasets, both human (74 datasets) and mouse (5 datasets), in order to characterize the various cell types composing the TME and analyzing their gene expression and signatures[12].

Clinical annotation is limited to tumor type, primary versus metastatic, and treatment. The database functionalities allow comparison between the cellularity and gene expression of various datasets.

In this study we propose the first almost exhaustive and fully integrated scRNAseq database exclusively dedicated to human cancer, with a detailed clinical annotation, allowing connecting cell types and gene expression patterns to specific clinical patterns. The IMMUcan database offers a large number of functionalities for the analysis of multiple datasets. We hope it will become the gold standard reference tool to support cancer biomedical research, in the early discovery, hypothesis-generating, as well as validation settings.

Results

Literature-based creation of the IMMUcan scDB

The IMMUcan scRNAseq database (scDB) was created through 4 main steps: an exhaustive literature search for human cancer scRNAseq studies, a manual review and curation of each relevant article, the collection of the corresponding datasets through web repositories or by contacting the authors, the processing and integration of the datasets and all associated metadata to the IMMUcan scDB (Figure 1A).

Literature search was performed in PubMed and bioRxiv, using general terms such as “patient”, “cancer” and “single cell RNA sequencing” (methods), in order to minimize missing of relevant articles. This required the manual screening of 468 original articles to focus on human cancer datasets of malignant, immune and/or stromal cells, using single-cell RNA-seq technology coupled or not with other technologies. This led to a final selection of 131 studies to be further curated. Each relevant article was manually curated in order to extract a number of features covering bibliographic information, clinical characterization of the patient cohort, experimental protocol, scRNAseq methodology, and the description of available data and metadata (Figure 1A and table S1). These features were defined by a group of medical and surgical oncologists, biologists and bioinformaticians. Clinical features included the cancer type, disease ontology, cancer location, tissue type, number of patients, treatment type, representing a total of 9 annotation fields. Experimental features included tissue dissociation, enrichment markers (when relevant), reached cell types (when relevant), and total cell numbers. A number of methodological characteristics were also extracted, such as single cell isolation method, library construction, end bias, genome reference, alignment method, and expression value. Whenever possible, we used standard international vocabulary and technical terms. For example, cancer types and cell types were based on the human disease ontology[13] and cell ontology[14,15] respectively.

Dataset collection was performed by accessing public repositories such as GEO, SRA, EGA, and ArrayExpress. Raw and processed data were available for the majority of datasets. All available annotations, data and metadata were integrated into the IMMUcan scDB, and can be searched through a user-friendly interface (https://immucanscdb.vital-it.ch). A total of 95 publications were successfully integrated into our database, corresponding to 121 datasets, with information for all available annotation fields (Figure 1B). Fifty-one cancer indications were included, with a majority of melanoma (MEL) datasets (13 datasets, 192 patients), followed by Glioblastoma (GBM) (9 datasets, 81 patients), and colorectal cancer (CRC) (8 datasets, 85 patients) (Figure 1C). Rare tumor types included acute T cell leukemia, renal cell carcinoma and certain childhood tumors like medulloblastoma. The majority of the datasets were generated from single cell

suspensions with no prior enrichment (Unbiased) (53 datasets), followed by immune cell selection through CD45 enrichment (21 datasets), and T cell enrichment (13 datasets) (Figure 1C). Importantly, 18 different types of enrichment protocols were applied across different studies, making it an important parameter for hypothesis-driven search of appropriate datasets. Patient treatment was known and described for only 52,7% of the patients, corresponding to 60,2% of the datasets. This information also provides possibilities to test specific hypotheses in connection with the clinical setting. Last, the technology used to generate the data was very heterogeneous and included eleven technologies, with a dominance in Smart-seq2 and 10X Genomics technologies (Figure 1C).

Overall, our exhaustive literature search coupled to manual curation allowed a detailed annotation of each scRNAseq dataset in order to rapidly retrieve datasets of interest using specific search criteria, and to account for various levels of heterogeneity that may impact the analysis and interpretation of the different datasets.

Dataset processing and integration

In order to process the public single-cell datasets efficiently we developed an R-based pipeline that performed all necessary processing steps in a semi-automatic manner based on best practices in the field[16]. We called it scProcessor.

Each dataset was first cleaned and formatted as a count matrix with features as rows and cells as columns together with a metadata file that contains as much experimental information as was publicly available. These annotations could go from cell annotation and T cell clonotypes to patient information and biopsy site. We checked if the data were normalized, if not we applied a log-normalization. Cleaned datasets were further processed through quality control steps (Figure 2A and B), such as number of principal components used. A threshold of 250 was applied for the minimum number of detected genes, and a range of 5-20 for the maximum percentage of mitochondrial genes depending on the tissue that was analyzed[17] (Figure 2B). This led to the exclusion of 1% and up to 50% of the cells depending on how well the data was generated and cleaned by the data curator.

Multiple sample integration was then performed using Harmony[18]. This consisted in specifying a batch variable for every dataset, in most cases patient or sample, which gets corrected through Harmony. The result was a removal of the main technical effect bias in the datasets (Figure 2C).

Cell annotation was performed based on a supervised and unsupervised approach. Supervised cell labelling was done based on CHETAH[19], which is a rapid method that uses hierarchical clustering to assign cell scores based on a classification tree from the TME. CHETAH has shown to perform well in a benchmarking study on supervised cell annotation methods[20]. We used the integrated human TME scRNA-seq study defined by CHETAH with small adaptations (see methods) to assign cell identity labels in our database. In total we specified thirteen different cell types including regulatory T cells and plasmacytoid pre-dendritic cells (pDC) (see methods and Figure S1A). The added value of CHETAH is that it can also assign an identity to unknown cell types that do not fit with any of the pre-specified labels as well as intermediate cell labels, such as for T cells and stromal cells, which brings our total of specified cell labels to 21 (see methods). To annotate malignant cells in the relevant datasets, we used CopyKat[21] which predicts if each cluster is aneuploid or diploid based on copy number aberrations (see methods and Table S2). We then introduced a subsequent manual cluster annotation in order to correct mislabeled cell types (Table S2). Unsupervised clustering of aggregated datasets was performed using Louvain graph-

based clustering implemented in the Seurat package[22]. Cluster annotation was performed manually after having defined 3 cell ontology levels: major cell types, minor cell types, and immune cell types (Figure S1B). Clusters defined by CHETAH and manually, respectively, were associated with each other with the added benefit that we specified a larger number of cell types/ states in our manual annotation which provides a more in-depth view of the TME (Figure 2D).

Finally, differential expression analysis allowed for the identification of gene expression programs specific to a given cell ontology term. In addition, to get a confidence score on how important a gene was in a specified cluster, we pre-calculated a Shannon-index to allow for a significance ranking of datasets (see Methods). All Seurat objects were converted to h5ad files by sceasy[23] for easy loading into single-cell visualization platforms such as cellxgene[24]. Collectively, these data processing and analytical steps led to successful integration of 64 high quality datasets: 43 datasets were processed and 21 datasets were not processed due to the low number of cells or the unavailability of the data (not public or licensed).

The processed datasets can be downloaded as h5ad files and CSV files (average gene expression and differential gene expression).

Cell type-based exploration of the IMMUcan scDB

The large number of annotation features that we have integrated in the IMMUcan scDB offers extensive possibilities for initial filtering and selection of the most relevant studies and datasets to test a given hypothesis. Users can filter for relevant studies based on study-specific information such as cancer indication, treatment type, or technology. In addition, there is the possibility of filtering based on gene or cell type of interest. Gene search will rank datasets based on importance of the specified gene using the Shannon index, while for cell search, datasets are ranked based on the absolute number of cells from the selected cell type. This provides the user with a selection and prioritization of datasets in order to perform meaningful downstream meta-analysis.

To demonstrate the usefulness of the IMMUcan scDB we first focused on a cell-type specific use case. It is known that the TME can impact response to immune checkpoint blockade[25]. Therefore, we searched for datasets from immunotherapy-treated patient samples. We found thirteen datasets from 3 cancer indications: melanoma, basal cell carcinoma, and squamous cell carcinoma. We focused on the basal cell carcinoma dataset BCC_BIA_10X_GSE123813 since it contained more than 50,000 immune and non-immune cells from patients before and after anti-PD1 therapy. Selecting the dataset opens a panel called “UMAP plot”, where one can explore the cell types, marker genes and check the expression of genes of interest. In the panel, all cells are projected in the Uniform Manifold Approximation and Projection (UMAP) space, and colored following the supervised CHETAH annotation as standard (Figure 3A). To improve the responsiveness, a subset of 10,000 cells is visualized per dataset. However, this can easily be switched back to the original number of cells. The legend shows all groups colored in the UMAP plot by name and between brackets the number of cells per group. Cell annotation can be switched to other cell hierarchical levels such as immune, major or minor, and also coupled to additional metadata such as biopsy, treatment or patient.

Besides the UMAP plot, a summary stacked bar plot shows the proportion of the cell types in the dataset. By selecting a clinical annotation field, multiple stacked bar plots are constructed and enable the comparison of the differences in the cell type proportions (Figure 3B). It was previously observed that naive and memory B cells are increased in responders in melanoma[26]

and renal cell carcinoma[27]. Here, we provide evidence that B cells are also increased in anti-PD1 therapy responders in BCC (Figure 3B). On the contrary, Plasma cells are more abundant in non-responders. The UMAP and bar plots can be downloaded as png files. Below the plots, two gene tables are automatically loaded, one with a matrix of average gene expression per cell type and one with the differentially expressed genes based on the selected cell (Figure S2). By default, the genes are sorted alphabetically but can also be sorted by cell type to find the highest expressed genes for a given cell type. The table columns are gene abbreviation, gene full name, IDs from Ensembl, Uniport, NCBI and HGNC, as well as information regarding the differential expression such as average log fold change (FC), percentage of positive cells in the selected population (pct.1) and the other populations (pct.2), and adjusted p-value. By default, the table is ordered for ascending adjusted p-value and descending average log FC, but this can be adapted. The second table is available below the UMAP plot in the second tab, and corresponds to the differentially expressed genes (DEG) for each cluster, according to the annotation the user selected.

Each gene from both tables can be visualized as a violin plot by pressing the violin plot icon next to the gene name (Figure S2). A violin plot together with a boxplot display the gene expression per cell type. To improve the interpretation of these plots, the absolute cell number is represented as pie charts at the below the violin plots and the percentage of non-zero expressing cells appears in a mouseover (Figure 3C).

In addition, the expression of a selected gene in each individual cell can be visualized on the UMAP plot. This enables to explore the DEG between cell types according to the different available annotations. In our BCC_BIA_10X_GSE123813 example dataset, we highlighted the expression of the typical naive and memory B cells marker genes MS4A1 and CCR7, and the expression of IGHG1 and IGHG4 as plasma cells markers (Figure 3C).

Another panel references all study metadata information such as publication, cohort, technology and study metadata. Here, important links are embedded to other databases like the disease ontology, original publication and accession of the original raw and/or processed data. The IMMUcan scDB also provides several options to download the analysis performed, all plots that are created in the database can be downloaded as high quality png files.

Gene-based exploration of the IMMUcan scDB

Recently, a study of multiple bulk transcriptomic cancer datasets has shown that CXCL13 and CXCL9 could be used as a predictive biomarker of checkpoint immunotherapy response[28]. Using IMMUcan scDB features, we were interested in finding which cell types could express those 2 genes in the different cancer types. IMMUcan scDB allows the selection of datasets according to the user's gene of interest. It displays a heatmap of the gene mean expression in each cell type in every dataset. We looked for CXCL13 and observed that it is highly expressed in exhausted T cells and T follicular helper cells (T_{FH} cells, Tfh) in basal cell carcinoma (BCC), melanoma (MEL), non-small cell lung cancer (NSCLC), breast cancer (BC) and colorectal cancer (CRC) (Figure 4A). We then focused on the BCC_BIA_10X_GSE123813 dataset. We represented the level of expression of CXCL13 as a color gradient on the UMAP plot and we observed that cells with the highest expression, in red, corresponded to the Tfh and exhausted T cells subsets (Figure 4B).

The IMMUcan scDB also makes it possible to perform gene co-expression comparisons in the second panel called "gene X vs gene Y expression". Here the expression of two genes of interest in a given dataset can be queried and a scatter plot is created with one point per individual cell.

Cells are colored based on the selected cell type annotation category. The legend shows the annotation together with the number of cells that express both genes and a Pearson correlation p value in brackets for the given population. Since CXCL13 was expressed by exhausted T cells, we looked at PDCD1 (PD1), another marker associated to T cell exhaustion, to see if the 2 genes were co-expressed (Figure 4C top panel). Indeed, we observed a strong co-expression between CXCL13 and PDCD1 in exhausted CD8+ T cells and Tfh with a Pearson correlation coefficient of -0.15 and -0.32 for T_{FH} cells and exhausted CD8+ T cells, respectively (Figure 4C bottom panel). Below the scatter plot, Venn diagrams per cell type reflects the overlap of cells expressing gene X and gene Y (Figure 4D). For every cell type, the proportion of cells expressing at least one of the two genes is displayed as a pie chart, in the bottom-right corner (Figure 4D). The result of a hypergeometric test is available to assess the significance of the co-expression results and its p-value is visible on top of the pie chart (Figure 4D).

Litchfield et al. also discussed the role of CCR5 and CXCL9 as biomarkers[28]. In the BCC_BIA_10X_GSE123813 dataset, we observed that CCR5 and CXCL13 were co-expressed in T_{FH} cells and exhausted T cells (Figure S3A-B). However, as expected, CXCL9 and CXCL13 were not co-expressed (Figure S3C-D). CXCL9 was expected to be expressed in dendritic cells (DC) and macrophages[29]. In the BCC_BIA_10X_GSE123813 dataset, macrophages and 2 DC subtypes, CLEC9A+ DC and LAMP3+ DC express CXCL9 but not plasmacytoid pre-DCs (Figure 4E).

In other cancer types, such as melanoma, head and neck and lung cancer[30], CXCL9 is known to be expressed by macrophages. We went back to the gene-filtering feature of IMMUcan scDB to find whether there is macrophage or DC specific CXCL9 expression in different cancer types. We observed that CXCL9 was expressed by LAMP3+ DC in melanoma, hepatocellular carcinoma (HCC), BCC and lung adenocarcinoma (LUAD) (Figure 4F).

Using a gene-centric approach, IMMUcan scDB allowed us to quickly identify potentially novel cellular sources of CXCL9 and CXCL13 across tumor types.

In conclusion, here we present the IMMUcan scDB, a curated database of scRNA-seq studies of the human TME that is easily searchable and explorable. By means of 2 use cases we showed that the IMMUcan scDB is an efficient tool to validate observations from literature, to generate new hypotheses and to provide new insights.

Discussion

The number of scRNAseq studies in human cancer has increased exponentially in recent years. The first studies were performed to provide a large-scale description of tumor cells and TME, also referred to as an “atlas” view. Such studies extended from the most common tumor types (melanoma, breast cancer, non-small cell lung carcinoma) to rare cancers, such as atypical teratoid rhabdoid tumor[31] or rare molecular subtypes, such as triple negative breast cancer[32]. We can anticipate that scRNAseq “atlas” studies will continue to be published, focusing on an even broader diversity of tumor types, and probably including a larger number of patients and samples than initially done. Parallel to these descriptive studies, scRNAseq has been applied more recently to identify mechanisms of immune resistance[5], or a T cell-related signature associated to the response to immune checkpoint inhibitors[26]. Such hypothesis-driven studies should also grow in numbers and magnitude, with the diffusion and the increased accessibility to scRNAseq technologies. Another type of study design includes the comparison of different anatomical sites,

such as primary versus metastatic tumor location[4]. The number and diversity of past and most probably future scRNAseq studies justifies a resource that would be fully dedicated to human cancer datasets, in order to provide a detailed annotation, easy and efficient search functions, as well as multiple implemented methods for meta-analysis. We believe this to be the only way to cope with an anticipated number of several hundred datasets in the coming years. In this respect, we will pay particular attention to the prospective integration of newly published data sets according to the standardized strategy that we have established. Within the IMMUcan consortium we will maintain the database as much as possible with monthly updates. Additionally, we will soon add a feature enabling users to suggest new public datasets to add in the database.

Public data repositories offer access to an increasing number of large-scale ("omics") datasets, in particular genomics and transcriptomics. However, clinical annotation is often missing or reduced to a minimal amount of information, such as the tumor type. This greatly limits the possibilities for integration of clinical and biological data in the analysis process and interpretation of the results. Single cell portals, such as UCSC cell browser[33], Broad institute single cell portal (https://singlecell.broadinstitute.org/single_cell) or single cell expression atlas[34], do not include this level of annotation. Cancer scRNAseq databases such as CancerSea[11] or TISCH[12], include minimal clinical information, restricted to tumor type, primary or metastatic stage, and treatment type. In our study, we have gone through the manual process of extracting and mapping to reference ontologies detailed clinical features (9 items) associated to each patient cohort and datasets. This should allow biologists and clinicians to focus on datasets corresponding to a particular clinical setting, and to compare datasets across different clinical settings. Integrating this information in the analysis and interpretation process should also provide important insight into cell types, cell states and associated signatures.

Different from bulk transcriptomics analysis, scRNAseq generates data from a large number of cells even in individual samples. Assuming that cell numbers are sufficient, this offers the possibility for robust characterization of cellular clusters and associated gene expression programs in individual patients. In parallel, the aggregated analysis of several datasets fulfilling specific common conditions is also important in order to identify unifying patterns associated to a tumor type, a specific anatomical location, or a treatment effect. A recent study has constructed a "pan-cancer blueprint" of stromal cell heterogeneity using original scRNAseq data sets from 4 cancer types[35]. It revealed shared gene expression programs in infiltrating immune cells. In our IMMUcan scDB, we have implemented robust methodologies to integrate several samples in order to identify common patterns and increase statistical relevance to a given clinical setting. As a result, users may apply focused strategies on individual patient samples.

The IMMUcan scDB offers large possibilities for applications depending on the biomedical level of interest. By using exploratory analysis, users can utilize the database in an early discovery process in order to generate hypotheses for further validation. For example, comparison of cell type specific signatures from different clinical settings may reveal interesting mechanisms of immune activation or immune escape, or novel therapeutic targets. Data exploration can also be performed in a hypothesis-driven manner, in order to establish the expression pattern of specific genes or signatures according to different annotation terms. Last, our database can also be used to validate findings established in an independent study. The large and increasing number of scRNAseq datasets offers unique possibilities for cross-validation of results coming from different technologies, such as proteomics, genomics, or spatial transcriptomics.

Integrating such a large number of scRNAseq datasets into a single database has potential risks and limitations. As all literature-based resources, the quality of sample and dataset annotation

relies on the quality of the information provided in the original publication. In this respect, we have found tremendous heterogeneity in the way patient cohorts are described, both in the amount and in the quality of the clinical information. An important step forward would be the improvement and generalization of standardized terminologies, such as the human disease ontology[13] and cell ontology[14,15], as well as a more systematic and thorough clinical annotation within existing genomics data repositories, along with a unified storage procedure. The processing of scRNAseq datasets generated in different studies, using various tissue dissociation and enrichment protocols, as well as potentially different technological platforms, is certainly challenging and subject to technical biases. In our processing pipeline, we have implemented robust and validated methodologies at each step. We have selected Harmony as a method to reduce experimental bias in the process of multiple datasets integration[18]. Harmony uses reiterative clustering in order to remove batch effects between experiments and patients. From recent benchmarks studies on integration of single-cell RNA-sequencing data[16,36,37], Harmony was among the top performers and is recommended as integration method over methods such as CCA[38], scanorama[39] and MNNcorrect[40] for its good integration and short runtime. Users should be aware of all these limitations and possible biases and may use their own cross-validation methodologies in order to increase the robustness of their findings. Improving the performance of our data processing will remain a top priority in the coming years. We will survey the literature for any method that could work in synergy with the pipeline that we have established in order to control biases and increase data analysis quality. Overall, we believe that the power and possibilities offered by integrating such a large number of datasets largely outweighs the limitations and weaknesses inherent to any meta-analysis. We hope that our resource will facilitate the exploitation of publicly available scRNAseq datasets to address existing and novel challenges in human cancer research.

Material and methods

Literature search and dataset selection

We included search of peer-reviewed published dataset using Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/) as well as non-peer-reviewed studies using bioRxiv (https://www.biorxiv.org) databases. To include all studies falling into our criteria, we used ((cancer[Title/Abstract]) AND (patient)) AND (single cell RNA sequencing) key words in Pubmed, and “human cancer single-cell rna-sequencing” free-text keywords in bioRxiv. We applied a filter to select articles published from 2016-2021. We manually reviewed all the resulting article titles and abstracts to check for the relevance of each study to our database. According to the objectives of this database we focused on human cancer datasets, using single-cell RNA-seq technology coupled or not with other technologies such as Whole Exome/Genome Sequencing (WES/WGS), TCR-sequencing, Chip-sequencing or proteomic/CyTOF data. After selection of manuscripts with an applicable scRNA-seq dataset of human cancer patients, we only selected studies with more than 1000 cells for further curation of the data.

Definition of fields

We reviewed all available information from every study to select fields of information that would be relevant for the database. We then organized them into categories. The first category was related to the bibliographic information regrouping several fields like title, abstract and DOI of

the article and data accession information. The second category was related to the disease-specific attributes such as cancer localization, cancer type, number of patients, or treatment type. Then, the following fields precised the single-cell technology specific attributes (tissue dissociation method, enrichment markers used, enrichment cell types obtained, single cell isolation method, single cell entity, "Omic" type, clonal information, genotyping, library construction, end bias, library layout, reference genome used, alignment method, counting method, expression value format and cell amount). The two last categories are related to conclusions and free remarks, as well as metadata information availability.

We homogenized the terms, especially for disease ontology and treatment. Depending on the field, the information can be either free-text, a list from a controlled vocabulary, boolean values, or quantitative information.

Data access

After selection of 131 publications with an applicable scRNA-seq dataset of human cancer patients, we manually curated the processed data as from GEO, ArrayExpress, EGA and BioProject. Sixteen datasets were not available or under license. Every study that contained multiple experiments or cancer indications was split in separate datasets and only datasets with more than 1000 cells were selected for further curation leading to a total of 65 datasets across 54 different cancer indications. As means of completion, some datasets from heathy human tissues such as bone marrow peripheral blood mononuclear cells (PBMC) or similar were also included in order to compare tumor with healthy tissues. Datasets were downloaded in different forms, from counts matrices to h5ad files and different raw and normalized expression values. If available, raw count values were prioritized over normalized values for two reasons: 1) this would increase the comparability between studies and 2) transcripts per million (TPM) and reads per kilobase per million mapped reads (RPKM) normalization strategies are not optimal for single-cell studies[16]. A small fraction of datasets only provided TPM values; we choose to not convert these into raw counts because essential information like isoform lengths used to normalize were missing and would reverse normalizing them incorrect. Instead, we opted to document the processed values per study and provide full transparency. Certain metadata fields were standardized across studies like patient ID, biopsy, timepoint of treatment, response and if available, the original tissue annotation. The source code for processing all the collected datasets is available as a repository on github (https://github.com/soumelis-lab/IMMUcan).

Quality control and batch correction

To make sure cleaned datasets contained high-quality cells, first a quality control was conducted with cut offs for minimum detected genes and percentage of mitochondrial reads. For detected genes a standard cut off of 250 was used while for mitochondrial content this would range from 5 to 20% depending on the tissue[17]. Further processing was performed with Seurat v3[19] following current best practices[14]. To make sure batch effects would be removed for the appropriate datasets, we included an entropy-based method that quantifies the successfulness of batch separation. Batch is in most datasets defined as patient or sample depending on the experimental design. The entropy-based method is computed as follows: from a shared nearest neighbor graph on the 30 nearest neighbors of every cell j, the distribution q from batch m on the total batches M is calculated. Thereafter, the Shannon entropy Hj is calculated for every cell defined as:

$$Hj = -\sum_{m=1}^{M} q_j^m log q_j^m$$

A set of high entropies resembles a good mixture of batches whereas a low entropy resembles that the cells stay in the vicinity of their batch which indicates a batch effect. We decided to correct batch effect from an entropy below one. Batch correction was performed by Harmony[18] following standard procedures.

Supervised annotation and malignant cell prediction

Cells were annotated in a supervised fashion by CHETAH[19] based on a reference dataset provided by the authors with small changes to the provided annotation levels. We added an annotation group for pDCs changing the total number of cell types from twelve to thirteen. Cells were classified based on normalized counts with 500 genes used at every step at a threshold of 0.05.

Malignant cells are very heterogeneous and cannot be classified by a reference dataset. For datasets containing malignant cells, we called copy number aberrations with CopyKat[21]. We followed the standard procedures described by the authors. Normal cells were provided using CHETAH identifier to increase the prediction accuracy. The malignant clusters were then assigned as malignant based on the ratio of aneuploid over diploid together with a manual check if these clusters also expressed well known cancer genes (Table S2).

Annotation based on graph-based clustering

To make sure no erroneous annotation results are processed further, we checked manually for the validity of the results. We did this based on results of graph-based clustering that we performed through Seurat with a resolution of 1 and following the standard procedure. Plots based on marker genes of cell types from the TME were curated through bibliographic search (Table S2) were generated. Based on these output files, every cluster was evaluated and if necessary reassigned by us. In addition, we specified 3 annotation levels that were linked to the final cell annotation. We specified a major annotation based on ten cell types: endothelial, pericyte, fibroblast, B, plasma, myeloid, NK, T epithelial, malignant and other cells such as hepatocytes or melanocytes. Myeloid and T cells were further specified in the immune annotation level as macrophage, monocyte, mast cell, neutrophil, dendritic cell, granulocyte and cycling for myeloid cells and CD4+, CD8+ and cycling for T cells. Based on the Seurat clusters some cells were then even further annotated in the minor annotation. Macrophages were split in SPP1+ and C1QA+; dendritic cells in conventional, plasmacytoid, LAMP3+ and CLEC9C+; CD8+ T cells in naive, central memory, effector memory, effector, exhausted and mucosal associated; CD4+ T cells in naive, helper, follicular helper, helper 17, regulatory and activated regulatory T cells.

Gene entropy ranking and differential expression analysis

We use two techniques to prioritize genes, one is a gene ranking based on entropy, the other is a differential expression analysis between annotated cell types. Gene entropy ranking of genes is performed using the gene Shannon index as described in Ibrahim & Kramann 2019[41], performed on the most granular annotation which is the Seurat clustering. A specificity score is calculated for each gene in each cell cluster which combines the uniqueness of the gene to the cell cluster

and its expression level in that cluster using the function sortGenes in the genesorteR package setting binarizeMethod to “naive”. The scores range between 0 and 1, with a value of 1 indicating that a gene is expressed in all cells of a cluster and in none of the cells of any other cluster[41]. An entropy-like index is calculated on these scores; ranking genes by this index provides a ranking of “importance” of a gene in a dataset (ie. whether a gene is a unique marker in a given dataset). These gene ranks can then be compared across datasets to rank studies by gene query. Only genes with a specificity score adjusted p-value <0.1 as calculated by the function getPValues from the genesorteR package are ranked.

For the differential expression analysis, we subsampled every large dataset to 20,000 cells randomly, with a seed so that every operation would be repeatable. Differential testing was done on every annotation level, ranging from CHETAH annotation to minor annotation and was based on a non-parametric Wilcoxon rank sum test. The test was performed with Seurat for every cell type annotation versus the rest of the dataset. Other requirements were an output of only upregulated genes, genes had to be expressed in at least 10% of the cells with a log fold change of at least 0.25.

Web Portal

The web portal enables to query, browse, mine, visualize and download scRNAseq datasets normalized (batch corrected) and standardized by the processing pipeline. The front-end has been developed with the VueJS framework (https://vuejs.org/), the Bootstrap CSS Library (https://getbootstrap.com/), the echarts visualization library (http://echarts.apache.org/) and the d3js library (https://d3js.org/). The back-end has been developed with PHP and the SLIM framework (https://www.slimframework.com/). h5ad files are parsed with a custom Python script using the scanpy library (https://scanpy.readthedocs.io/). The web portal is freely accessible at: https://immucanscdb.vital-it.ch/

Acknowledgements

SIB would like to thank Jorge Molina (Core-IT, SIB) for his expert support in setting up the IT infrastructure of the web portal.

The IMMUcan project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821558. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and EFPIA (https://IMI.europa.eu).

We would like to thank Cristina Ghirelli for her help in reviewing the manuscript. We also would like to acknowledge our colleagues from IMMUcan consortium who gave us insightful feedback on the database.

Author contributions

J.P., H.G.R., J.M. and V.S. designed and supervised the study. V.S., J.C. and F.N. wrote the manuscript with input from all co-authors. J.C developed the analysis pipeline. J.C. and F.N. gathered, curated and analyzed the data with the help of C.H., L.M.-R., E.A., M.S., and M.M.I.. R.L. developed the database interface with the help of L.G..

Competing interests

Jordi Camps, Mahmoud M Ibrahim and Helge Gottfried Roider are currently employed at Bayer. Jack Pollard is a full-time employee at Sanofi. The remaining authors declare no competing interests.

References

1. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet. TIG* **34**, 666–681 (2018).

2. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

3. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

4. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

5. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984-997.e24 (2018).

6. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).

7. Cao, Y., Zhu, J., Jia, P. & Zhao, Z. scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells. *Genes* **8**, 368 (2017).

8. Abugessaisa, I. *et al.* SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.* **46**, D781–D787 (2018).

9. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).

10. Ner-Gaon, H., Melchior, A., Golan, N., Ben-Haim, Y. & Shay, T. JingleBells: A Repository of Immune-Related Single-Cell RNA–Sequencing Datasets. *J. Immunol.* **198**, 3375–3379 (2017).

11. Yuan, H. *et al.* CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* **47**, D900–D908 (2019).

12. Sun, D. *et al.* TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.* **49**, D1420–D1430 (2021).

13. Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, W541-545 (2011).

14. Diehl, A. D. *et al.* Hematopoietic cell types: prototype for a revised cell ontology. *J. Biomed. Inform.* **44**, 75–79 (2011).

15. Meehan, T. F. *et al.* Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).

16. Luecken, M. *et al. Benchmarking atlas-level data integration in single-cell genomics*.

http://biorxiv.org/lookup/doi/10.1101/2020.05.22.111161 (2020) doi:10.1101/2020.05.22.111161.

17. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA sequencing data quality control. *Bioinforma. Oxf. Engl.* (2020) doi:10.1093/bioinformatics/btaa751.

18. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

19. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **47**, e95–e95 (2019).

20. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).

21. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00795-2.

22. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

23. Cakir, B. *et al.* Comparison of visualization tools for single-cell RNAseq data. *NAR Genomics Bioinforma.* **2**, lqaa052 (2020).

24. *Chan Zuckerberg Initiative chanzuckerberg/cellxgene: An interactive explorer for single-cell transcriptomics data.*

25. Pitt, J. M. *et al.* Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *Ann. Oncol.* **27**, 1482–1492 (2016).

26. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).

27. Helmink, B. A. *et al.* B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).

28. Litchfield, K. *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596-614.e14 (2021).

29. Park, M. K. *et al.* The CXC Chemokine Murine Monokine Induced by IFN-γ (CXC Chemokine Ligand 9) Is Made by APCs, Targets Lymphocytes Including Activated B Cells, and Supports Antibody Responses to a Bacterial Pathogen In Vivo. *J. Immunol.* **169**, 1433–1443 (2002).

30. House, I. G. *et al.* Macrophage-Derived CXCL9 and CXCL10 Are Required for Antitumor Immune Responses Following Immune Checkpoint Blockade. *Clin. Cancer Res.* **26**, 487–504 (2020).

31. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* **51**, 1702–1713 (2019).

32. Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).

33. Speir, M. L. *et al. UCSC Cell Browser: Visualize Your Single-Cell Data*. http://biorxiv.org/lookup/doi/10.1101/2020.10.30.361162 (2020) doi:10.1101/2020.10.30.361162.

34. Papatheodorou, I. *et al.* Expression Atlas update: from tissues to single cells. *Nucleic*

Acids Res. gkz947 (2019) doi:10.1093/nar/gkz947.

35. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res.* **30**, 745–762 (2020).

36. Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y. & Hemberg, M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.* gkab004 (2021) doi:10.1093/nar/gkab004.

37. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).

38. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

39. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

40. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

41. Ibrahim, M. M. & Kramann, R. *genesorteR: Feature Ranking in Clustered Single Cell Data*. http://biorxiv.org/lookup/doi/10.1101/676379 (2019) doi:10.1101/676379.

Figure 1: Single-cell RNA-sequencing database workflow

A. Strategy used to create the IMMUcan SCdb. **B**. Overview of the home page of the database web interface. **C**. Statistics of the database content represented as lollipop plot, the information (cancer type, cell type enrichment, treatment and technology) is shown on the y-axis while the related number of datasets is shown on the x-axis. Point size correspond to the number of patients and the color-gradient represents the number of cells per 100000.

Figure 2: Dataset processing before integration into IMMUcan scDB

A. Dataset processing pipeline. **B**. Quality control plots for the CLL_IMM_10X_GSE111014 dataset: PCA standard deviation Elbow plot, the vertical red line indicates the number of principal components used for downstream analysis, Violin plot of the number of genes for each cell by patient, Violin plot of the number of counts for each cell by patient, Violin plot of the percentage of mitochondrial genes for each cell by patient, the horizontal red line indicates the threshold used to filter out the cells with a high percentage of mitochondrial genes. **C**. UMAP, PCA, and Harmony plot before and after batch correction by Harmony (colored by patient). **D**. UMAP plots with cells colored according to CHETAH (top-left) or immune annotation (bottom-left). Dotplot of matching cell annotations for CHETAH and immune (right panel), the color-gradient and point size represent the number of cells.

Figure 3: Cell-based exploration of IMMUcan scDB looking at B cells involvement in basal cell carcinoma response to anti-PD1 treatment

A. UMAP plot of BCC_BIA_10X_GSE123813 dataset (by cell type, by treatment response), the cells are colored according to their CHETAH annotation. **B**. Bar plots of the percentage of cells per cell types in the whole dataset, and per response to treatment status (Yes or No), the cell types are colored according to the CHETAH annotation. **C**. Violin plot combined with boxplot of the expression of two representative B cell (MS4A1, CCR7) and plasma cells markers (IGHG1, IGHG4), Pie chart representing the proportion of expressing cells ("non-zero") and below the absolute number. The colors correspond to CHETAH cell type annotation.

Figure 4: Gene-based exploration of the IMMUcan scDB using CXCL13, a predictive biomarker for immunotherapy response

A. Heatmap of CXCL13 expression across datasets (y-axis) and cell types (x-axis), cell types are defined according to the minor annotation. **B**. UMAP plot of BCC_BIA_10X_GSE123813 dataset colored by cell type (minor annotation, top) and CXCL13 expression (bottom). **C**. Co-expression plot of CXCL13 and PDCD1(PD1), cells are colored according to the minor annotation, top-panel plot displays all cell types, bottom-panel displays only exhausted CD8+ T cells (T CD8 ex) and T_{FH} cells, the legend indicates the cell type with the number of expressing cells and the Pearson correlation coefficient in brackets. **D**. Venn diagram showing the co-expression of CXCL13 and PDCD1by T_{FH} cells (left) and exhausted CD8+ T cells (T CD8 ex, right), the p-value of the hypergeometric test is in the top-right corner of each plot, a pie chart representing the proportion of expressing cells for one of the two genes is in the bottom-right corner of each plot. **E**. Violin plot of CXCL9 expression across cell types annotated and colored according to the annotation minor in BCC_BIA_10X_GSE123813 dataset. **F.** Heatmap of CXCL9 expression across datasets (y-axis) and cell types (x-axis), cell types are defined according to the minor annotation.

Supplementary Figure 1: Supervised and non-supervised cell type annotation strategies

A. t-distributed stochastic neighbor embedding representation of an example dataset colored by CHETAH annotation (left panel) and the corresponding CHETAH classification tree (right panel). **B.** Cell ontology classification structure used for the manual annotation. B: B cells, DC conv: conventional dendritic cells, DC plas: plasmacytoid dendritic cells, Mast: mast cells, NK: natural killer cells, Plasma: plasma cells, T CD4: CD4+ T cells, T CD8: CD8+ T cells, T reg ; regulatory T cells, Macro: macrophages, Mono: monocytes, DC: dendritic cells, DC CD1C cDC2: conventional type 2 dendritic cells, DC LAMP3: LAMP3+ dendritic cells, DC CLEC9A cDC1: conventional type 1 dendritic cells, T8n: naïve CD8+ T cells, T8cm: central memory CD8+ T cells, T8em: effector memory CD8+ T cells, T8eff: effector CD8+ T cells, T8ex: exhausted CD8+ T cells, T8mait: CD8+ Mucosal-Associated Invariant T Cells, T4n: naïve CD4+ T cells, T4h: CD4+ T helper cells, T4fh: CD4+ T follicular helper cells, T4h17: CD4+ T helper 17 cells, T4reg: CD4+ T regulatory cells.

Supplementary Figure 2: Preview of IMMUcan scDB interface when selecting a dataset. The dataset BCC_BIA_10X_GSE123813 was used as an example. The webpage is set on the UMAP tab and displays a UMAP representation of the dataset (top-left), a barplot of the distribution of cell types according to the treatment timepoint (top-right), and the differential gene expression tab (bottom).

Supplementary Figure 3: Use of the BCC_BIA_10X_GSE123813 of IMMUcan scDB to study the potential role of CCR5 and CXCL9 as biomarkers

A. Co-expression plot of CXCL13 and CCR5, cells are colored according to the minor annotation, the legend indicates the cell type with the number of expressing cells and the Pearson correlation coefficient in brackets. **B**. Venn diagram showing the co-expression of CXCL13 and CCR5 in exhausted CD8+ T cells (Tex) and T_{FH} cells (Tfh), the p-value of the hypergeometric test is in the top-right corner of each plot, a pie chart representing the proportion of expressing cells for one of the two genes is in the bottom-right corner of each plot. **C**. Co-expression plot of CXCL13 and CXCL9, cells are colored according to the minor annotation, the legend indicates the cell type with the number of expressing cells and the Pearson correlation coefficient in brackets. **D**. Venn diagram showing the co-expression of CXCL13 and CXCL9 in exhausted CD8+ T cells (Tex), T_{FH} cells (Tfh), cDC1, LAMP3+ DC (DC LAMP3) and Malignant cells, the p-value of the hypergeometric test is in the top-right corner of each plot, a pie chart representing the proportion of expressing cells for one of the two genes is in the bottom-right corner of each plot.

Supplementary Table 1: Information extracted as metadata for all datasets included in IMMUcan scDB

Categories of fields	Extracted metadata
Article information	First Author (name, surname)
	Date
	DOI
	PMID
	Journal
	Title
	Abstract
Sample information	Tissue
	Tissue ontology
	Cancer type
	Cancer type abbreviation
	disease ontology id
	Number of Patients
	Biopsy
	Healthy control group
	Matched biopsies
	Matched treatment
Treatment information	Treatment type
	Tissue/cell state
	Tissue dissociation
	Enrichment markers
	Enrichment cell types
	Enrichment abbreviation
	Single cell isolation
	Single cell entity
	Omic
	Clonal information

Data information	Library construction
	Library abbreviation
	End bias
	Library layout
	Genome reference
	Alignment method
	Counting method
	Expression value
	gene symbol/ensembl ID
	Accession processed data
	Accession raw data

Supplementary Table 2: Example of cluster annotation after CopyKat prediction (dataset GBM_UNB_SS2_GSE84465)

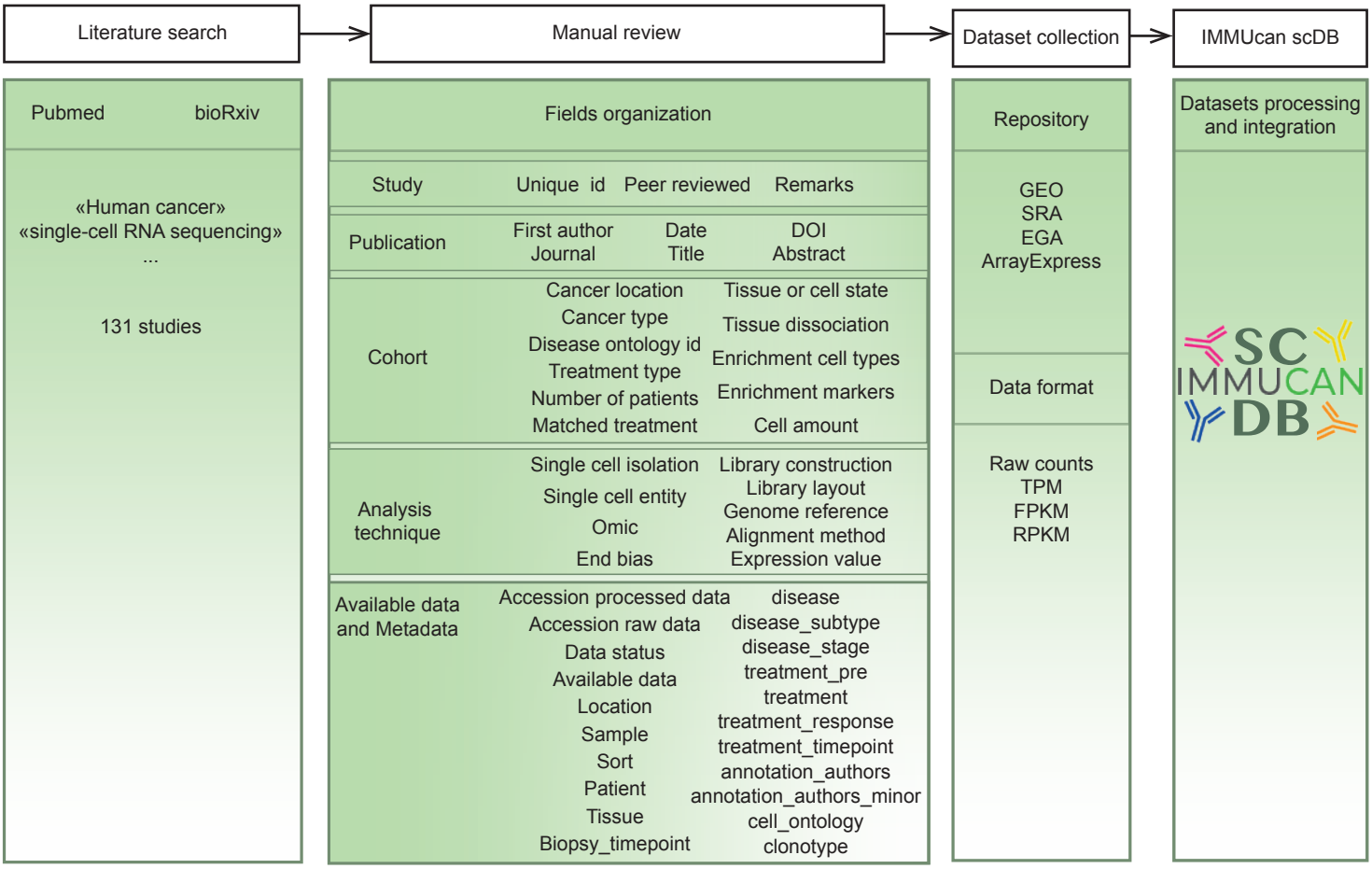
Abbreviations correspond to the manual cell type annotation. Ma: macrophages, mal: malignant cells

Seurat_clusters	annotation_CHETAH	fraction_CHETAH	copykat.pred	abbreviation
0	Macrophage	1	diploid	Ma
1	Macrophage	1	diploid	Ma
2	Node10	0,97	aneuploid	mal
3	Macrophage	1	diploid	Ma
4	Node10	0,87	aneuploid	mal
5	Macrophage	1	diploid	Ma
6	Node10	0,99	aneuploid	mal
7	Macrophage	0,94	diploid	Ma
8	Node10	0,9	aneuploid	mal
9	Node10	0,95	aneuploid	mal
10	Macrophage	1	diploid	Ma
11	Node10	0,98	aneuploid	mal
12	Macrophage	0,99	diploid	Ma
13	Macrophage	1	diploid	Ma
14	Macrophage	1	diploid	Ma
15	Node10	0,99	aneuploid	mal
16	Macrophage	1	diploid	Ma
17	Node10	0,92	aneuploid	mal
18	Node10	0,93	aneuploid	mal
19	Macrophage	0,88	diploid	Ma
20	Node10	0,95	aneuploid	mal
21	Node10	0,82	aneuploid	mal
22	Macrophage	0,93	diploid	Ma
23	Node10	0,85	aneuploid	mal

24	Macrophage	1	diploid	Ma
25	Node10	0,66	aneuploid	mal
26	Macrophage	0,44	diploid	Ma
27	Macrophage	0,96	diploid	Ma
28	Macrophage	1	diploid	Ma
29	Node10	0,91	aneuploid	mal
30	Node10	0,54	aneuploid	mal
31	myofibroblast	0,57	aneuploid	mal
32	Node10	1	aneuploid	mal
33	Fibroblast	1	aneuploid	mal

Figure 1

A


Literature search
Manual review
Dataset collection
IMMUcan scDB
Pubmed
bioRxiv
«Human cancer»
«single-cell RNA sequencing»
...
131 studies
Fields organization
Study
Unique id
Peer reviewed
Remarks
Publication
First author
Journal
Date
Title
DOI
Abstract
Cohort
Cancer location
Cancer type
Disease ontology id
Treatment type
Number of patients
Matched treatment
Tissue or cell state
Tissue dissociation
Enrichment cell types
Enrichment markers
Cell amount
Analysis technique
Single cell isolation
Single cell entity
Omic
End bias
Library construction
Library layout
Genome reference
Alignment method
Expression value
Available data and Metadata
Accession processed data
Accession raw data
Data status
Available data
Location
Sample
Sort
Patient
Tissue
Biopsy_timepoint
disease
disease_subtype
disease_stage
treatment_pre
treatment
treatment_response
treatment_timepoint
annotation_authors
annotation_authors_minor
cell_ontology
clonotype
Repository
GEO
SRA
EGA
ArrayExpress
Data format
Raw counts
TPM
FPKM
RPKM
Datasets processing and integration
SC IMMUCAN DB

B



C

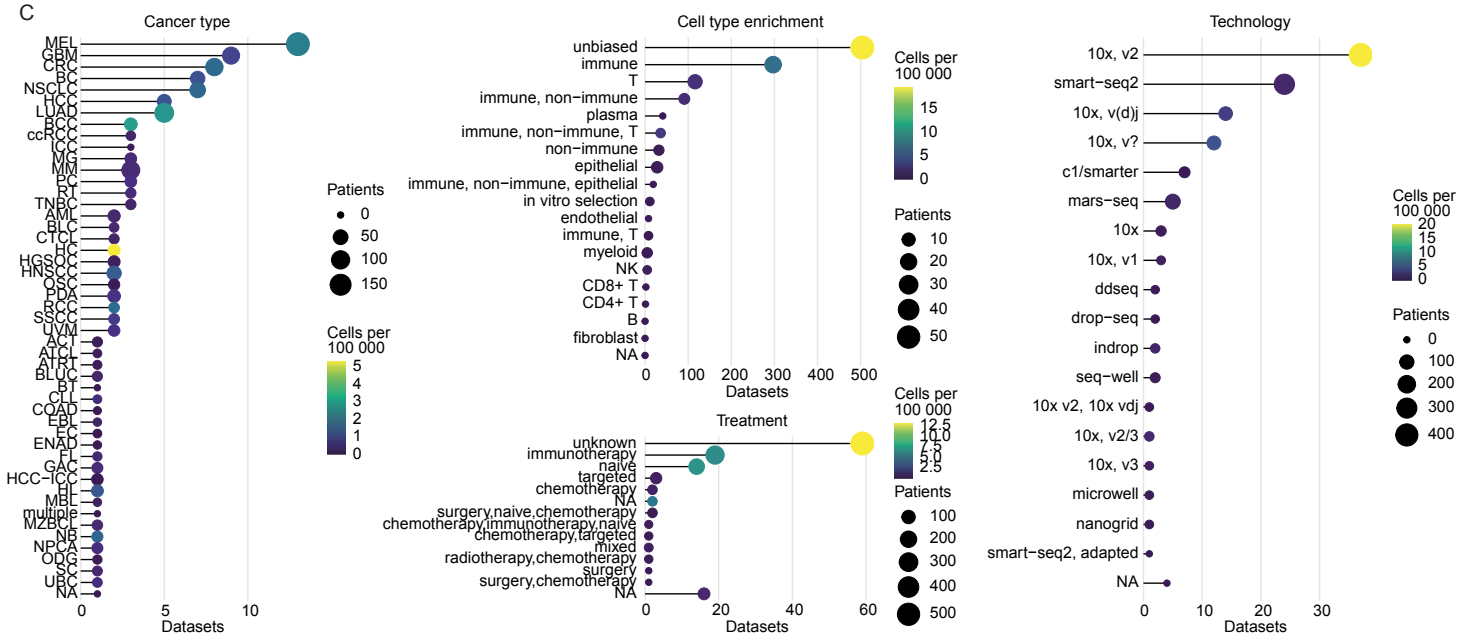

Cancer type
MEL
GBM
CRC
BC
NSCLC
HCC
LUAD
RCC
ccRCC
ICC
MG
MM
PC
PT
PC
TNBC
AML
BLC
CTCL
HC
HGSOC
HNSCC
OSC
PDA
RCC
SSCC
UVM
ACC
ATCL
ATRT
BLUC
BT
CLL
COAD
EBL
EC
ENAD
FL
GAC
HCC-ICC
HL
MEL
multiple
MZBCL
NB
NPCA
ODG
SC
UBC
NA
0
5
10
Datasets
Patients
0
50
100
150
Cells per 100 000
5
4
3
2
1
0
Cell type enrichment
unbiased
immune
T
immune, non−immune
plasma
immune, non−immune, T
non−immune
epithelial
immune, non−immune, epithelial
in vitro selection
endothelial
immune, T
myeloid
NK
CD8+ T
CD4+ T
B
fibroblast
NA
0
100
200
300
400
500
Datasets
Cells per 100 000
15
10
5
0
Patients
10
20
30
40
50
Treatment
unknown
immunotherapy
naive
targeted
chemotherapy
NA
surgery,naive,chemotherapy
chemotherapy,immunotherapy,naive
chemotherapy,targeted
mixed
radiotherapy,chemotherapy
surgery
surgery,chemotherapy
NA
0
20
40
60
Datasets
Cells per 100 000
12.5
10.0
7.5
5.0
2.5
Patients
100
200
300
400
500
Technology
10x, v2
smart−seq2
10x, v(d)j
10x, v?
c1/smarter
mars−seq
10x
10x, v1
ddseq
drop−seq
indrop
seq−well
10x v2, 10x vdj
10x, v2/3
10x, v3
microwell
nanogrid
smart−seq2, adapted
NA
0
10
20
30
Datasets
Cells per 100 000
20
15
10
5
0
Patients
0
100
200
300
400

Figure 2

A

QC	Principal Components selection [30-50] Minimum detected genes threshold [200-300] Maximum percentage of mitochondrial genes threshold [5-20]
Multiple sample integration	Harmony package
Clustering	Louvain graph-based clustering
Supervised cluster annotation	CHETAH cell annotation
Malignant cell prediction	CopyKAT
Cell ontology cluster annotation	3 levels of manual cell annotation (major, minor, immune)
Diffferential expression analysis	Wilcoxon test

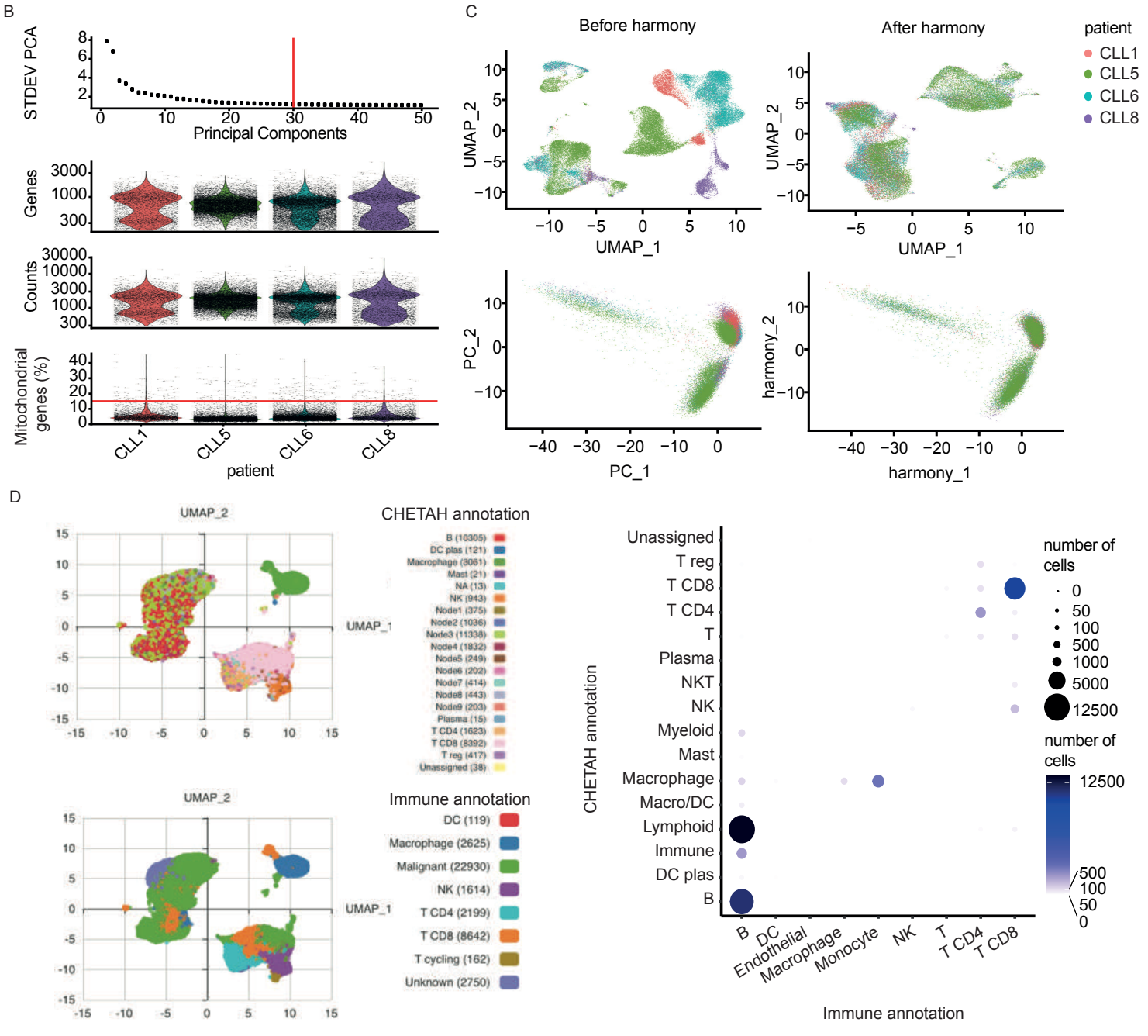


Figure 3

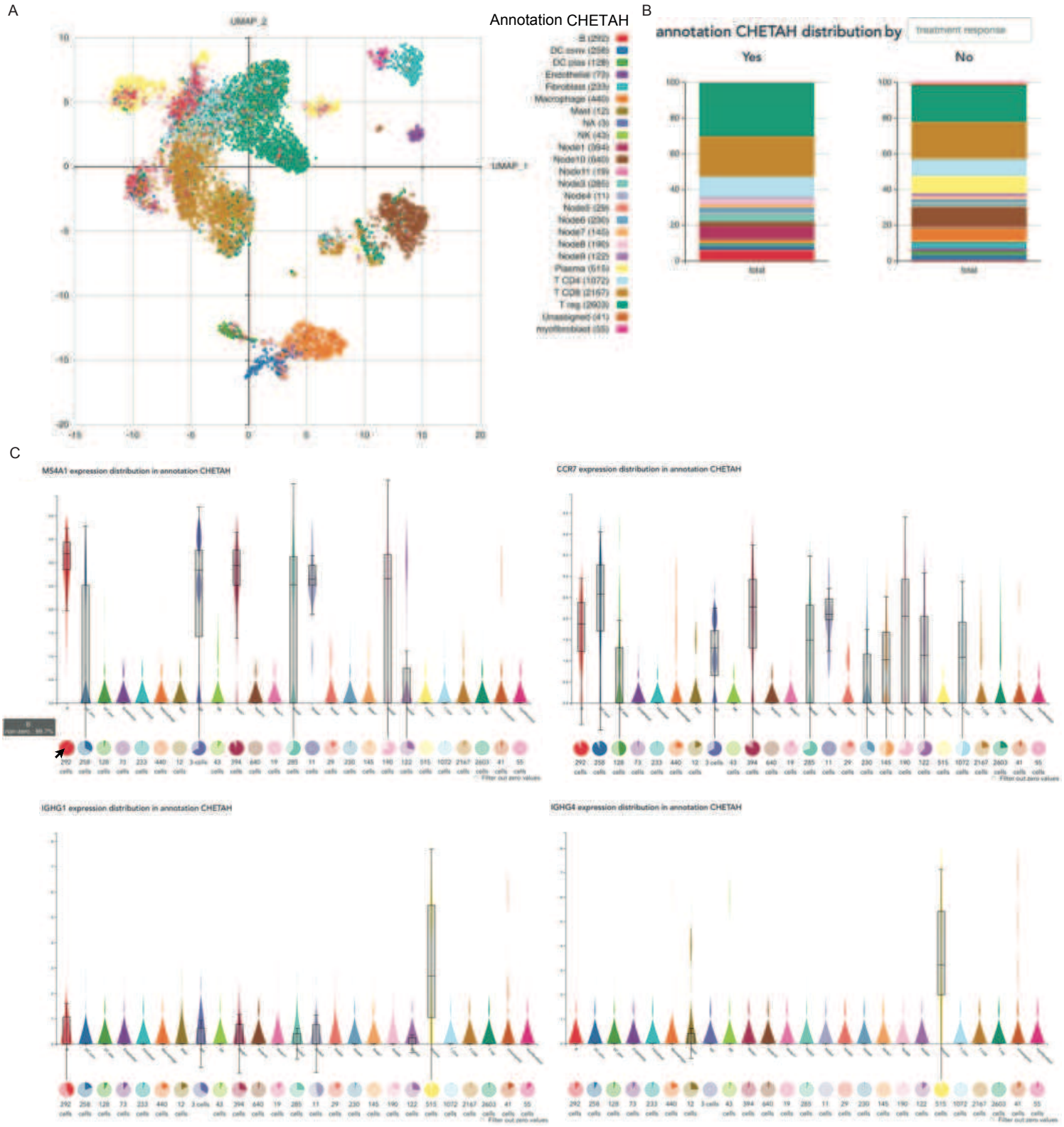
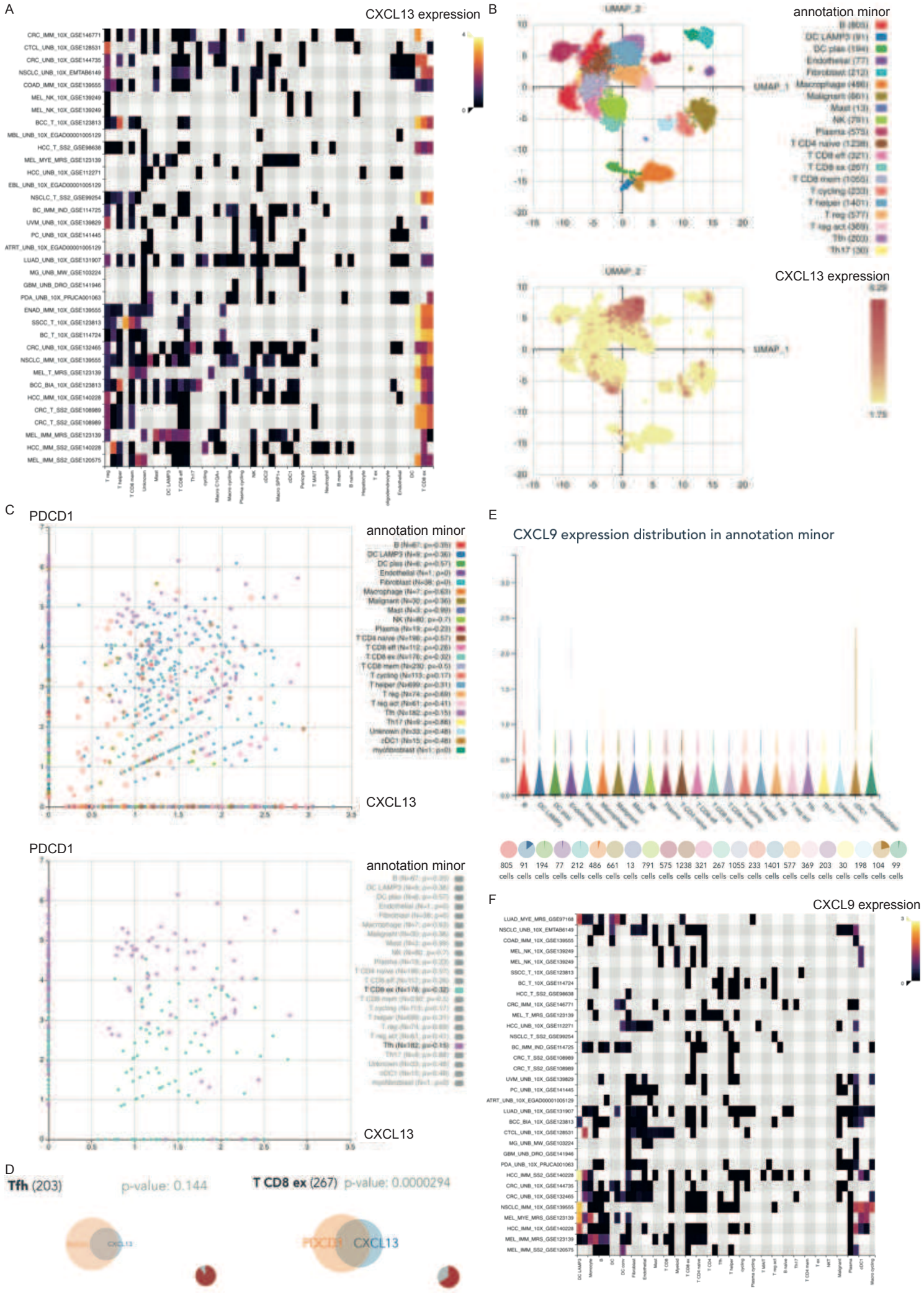
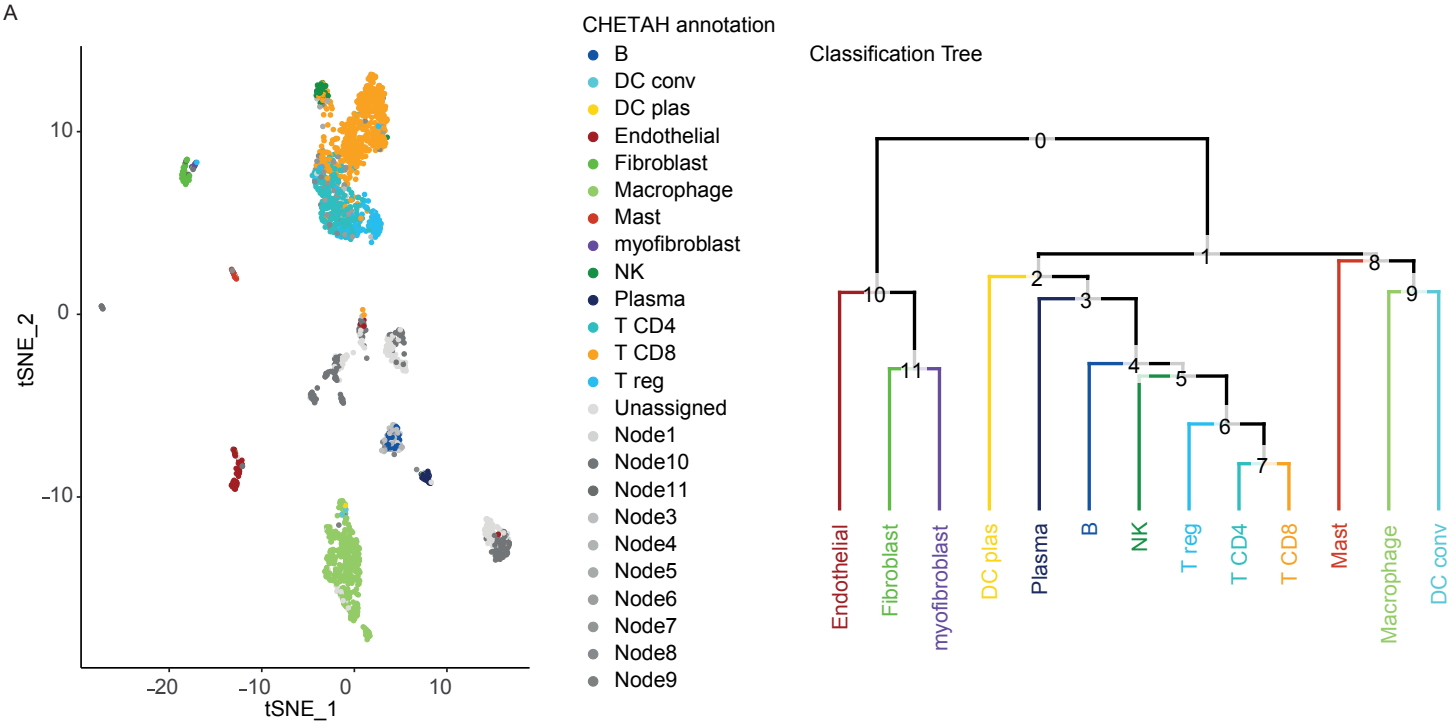

A
Annotation CHETAH
B
annotation CHETAH distribution by
treatment response
Yes
No
C
MS4A1 expression distribution in annotation CHETAH
CCR7 expression distribution in annotation CHETAH
IGHG1 expression distribution in annotation CHETAH
IGHG4 expression distribution in annotation CHETAH

Figure 4

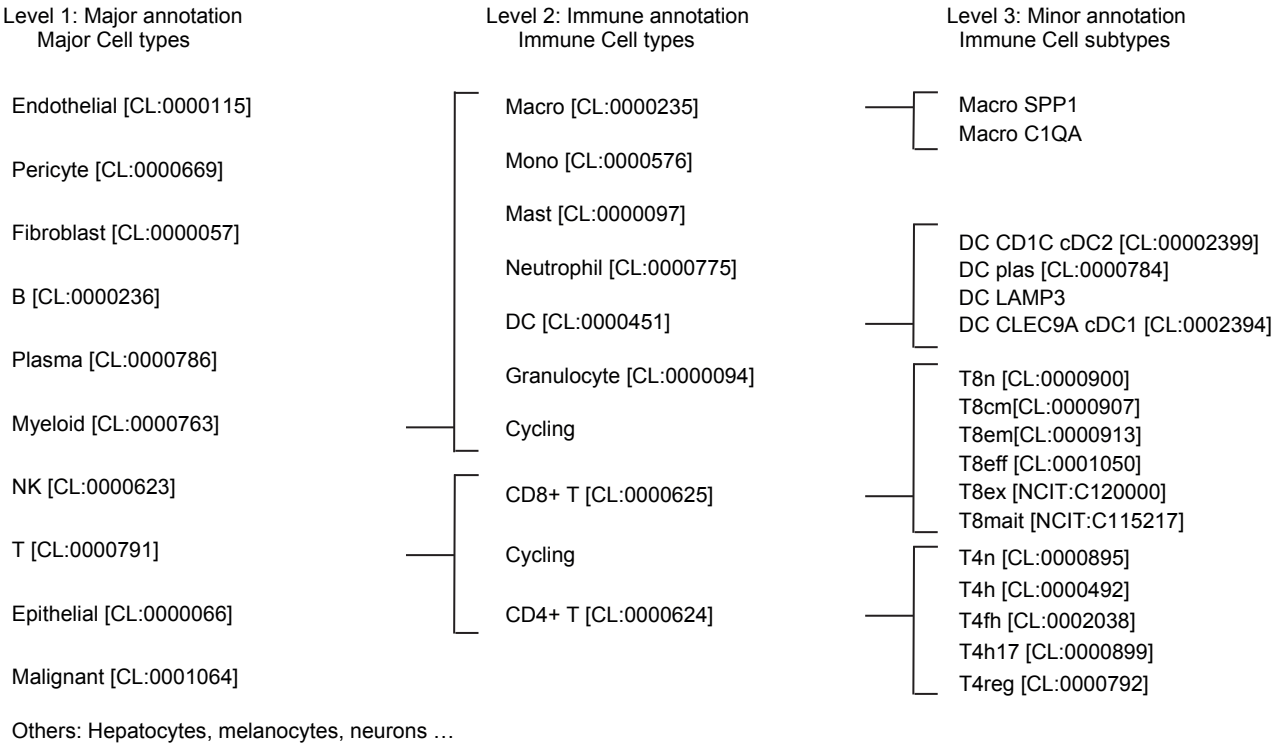

A
CXCL13 expression
B
annotation minor
UMAP_2
UMAP_1
CXCL13 expression
C
PDCD1
annotation minor
CXCL13
PDCD1
annotation minor
CXCL13
D
Tfh (203)
p-value: 0.144
T CD8 ex (267)
p-value: 0.0000294
PDCD1
CXCL13
E
CXCL9 expression distribution in annotation minor
805 cells 91 cells 194 cells 77 cells 212 cells 486 cells 661 cells 13 cells 791 cells 575 cells 1238 cells 321 cells 267 cells 1055 cells 233 cells 1401 cells 577 cells 369 cells 203 cells 30 cells 198 cells 104 cells 99 cells
F
CXCL9 expression

Supplementary Figure 1

A



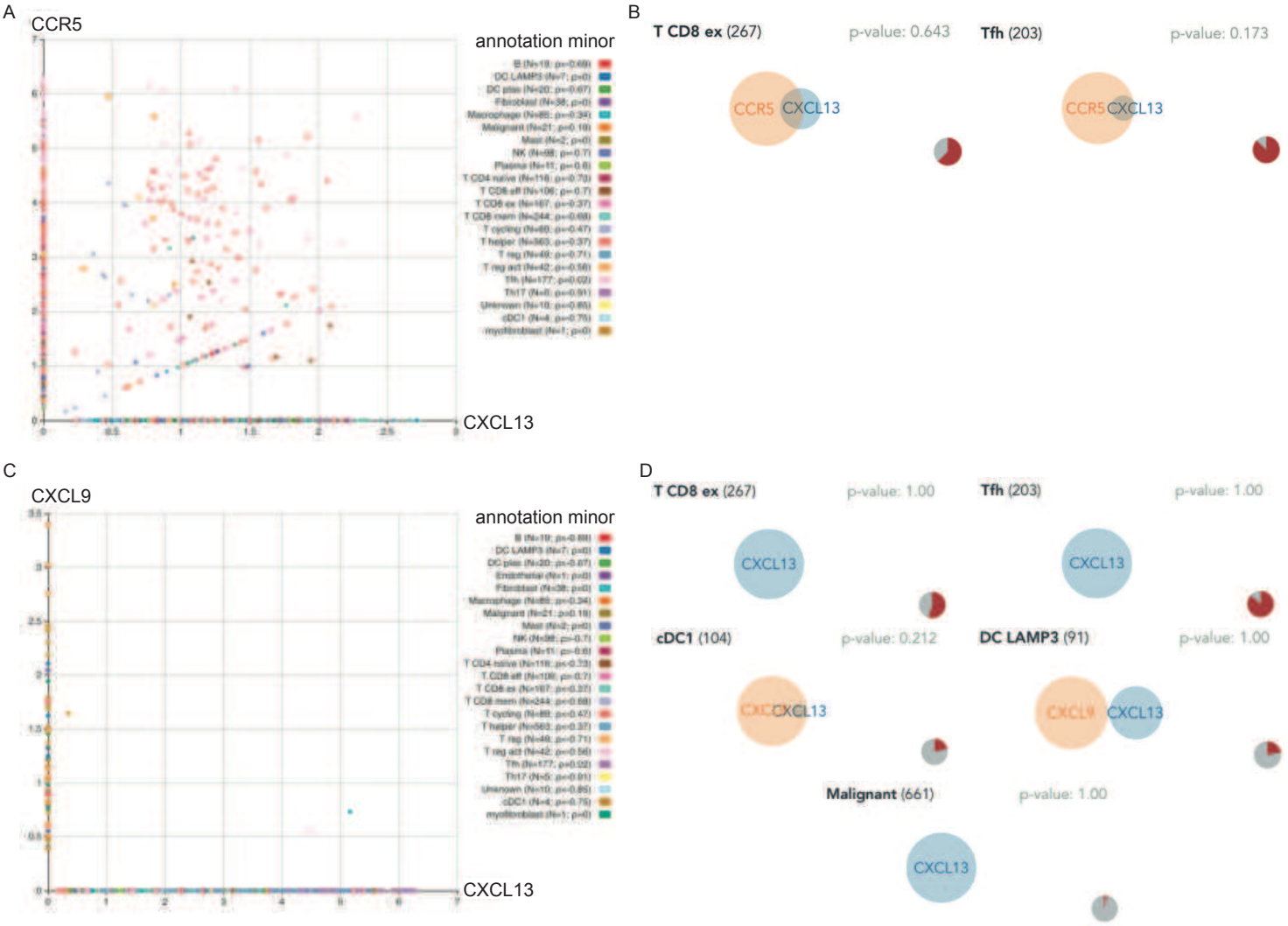
B



Supplementary Figure 2



Supplementary Figure 3



Part V

Résumé en français

Titre : Exploration de la diversité fonctionnelle des cellules T dans les données de séquençage d'ARN en cellule unique à l'aide d'outils méthodologiques et biologiques

Résumé : Le séquençage d'ARN en cellule unique (scRNAseq) est une technique jeune. Elle consiste à faire une photographie instantanée des molécules d'ARN messager contenues dans une cellule unique. Apparue en 2009, et après une phase timide d'adoption, son usage s'est généralisé, grâce à une simplification de la technique expérimentale et une baisse des coûts substantielle. Ces données sont plébiscitées pour leur richesse, qui permet de disséquer finement la biologie du vivant, en accédant à des informations telles que l'hétérogénéité d'une population ou la caractérisation différentielle des cellules saines et malades. En effet, le scRNAseq combine l'approche en cellule unique avec les techniques de séquençage de nouvelle génération qui permettent d'accéder, en théorie, à l'intégralité du matériel ARN de la cellule.

Cependant, le scRNAseq est à double tranchant. Bien que le travail à la paillasse se soit démocratisé, notamment grâce à l'apparition de kits commerciaux, il reste encore perfectible. En effet, on n'est capable de capturer, à l'heure actuelle, que 5 à 20% des ARN par cellule. Quant au travail à l'ordinateur, il constitue lui aussi un défi : les données sont fortement bruitées: ce bruit est dû non seulement au fait que les données sont immergées dans un espace à grande dimension, et donc souffrent de la malédiction de la dimensionalité et des phénomènes afférents, mais aussi au fait que la capture des ARN est incomplète. Ainsi, une analyse bioinformatique doit être capable de distinguer et séparer au mieux le bruit biologique intéressant du bruit technique qui parasite l'information.

Concernant la partie analytique, la nouveauté du scRNAseq n'a pas encore laissé suffisamment de temps aux équipes qui travaillent sur ces données pour élaborer des standards communs. Au contraire, on assiste à une explosion des algorithmes disponibles, majoritairement disponibles via R et python. Cependant, on retrouve schéma consensuel minimal : importation et nettoyage de la matrice d'expression cellules $\times$ gènes, normalisation et réduction de dimension. Les données en dimensions réduites permettent de faire de la visualisation, de l'agrégation ou de l'inférence de trajectoire. Enfin les groupes sont annotés. Cette dernière étape d'interprétation est particulièrement critique mais souvent biaisée car le plus souvent manuelle et donc biaisée.

Je me suis attachée dans ma thèse à deux aspects en particulier de l'analyse des données scRNAseq : l'aspect méthodologique, et l'interprétation.

J'ai d'abord étudié le bruit dimensionnel, autrement appelé la malédiction de la dimensionalité. Cette malédiction complique l'analyse en brouillant les différences entre points proches et lointains. Une manifestation classique de la malédiction est la concentration des distances, ce qui signifie que le ratio des distances extrémales tend vers 1. Autrement dit, la différence entre la distance maximale et la distance minimale observées dans un nuage de point tend vers 0. Comme l'analyse des données scRNAseq repose sur la production de graphes de voisinage, elle est nécessairement pénalisée par cette malédiction qui déforme des distances en atténuant les contrastes entre groupes ou noeuds de cellules. L'astuce habituelle consiste donc à réduire la dimension. Cependant, cette stratégie, si elle ne lasse pas d'être simple et efficace, soulève plusieurs questions, dont la principale est de savoir où tracer la ligne entre "éliminer du bruit" et "perdre de l'information". En plus de devoir faire un compromis non satisfaisant intellectuellement entre ces deux considérations, le bioinformaticien ne dispose souvent que de son "doigt mouillé" pour fixer un nombre de dimensions à garder. En outre, il existe aussi un autre effet, moins connu,

de la malédiction, qui s'appelle le phénomène de hubness. Il est aussi nocif, car il déforme le graphe des k plus proches voisins. Toutefois ce phénomène peut être corrigé avec des méthodes déjà existantes, qui s'attaquent soit à la correction du graphe des plus proches voisins, soit à la correction des inhomogénéités de densité locale à l'origine de l'émergence des hubs, soit à la réduction de la centralité spatiale. J'ai d'abord évalué l'ampleur du phénomène de hubness dans les données de séquençage, ainsi que l'effet de la correction de hub sur la performance de l'analyse scRNAseq, en appliquant les méthodes de correction du graphe des plus proches voisins. Le phénomène de hubness est bien présent, en particulier dans les matrices caractérisées par une grande dimension intrinsèque, et l'analyse de ces jeux de données en particulier bénéficie de la réduction de hubness, avec une performance optimale dans l'espace de dimension effective maximale. En particulier, nous nous sommes intéressés aux tâches de clustering, d'inférence de trajectoire et de visualisation, à l'aide de jeux de données dont la vérité était connue, c'est-à-dire que les cellules étaient déjà étiquetées. Bien que cela ne semble être qu'un algorithme de plus dans la jungle déjà existante, c'est surtout le changement de paradigme qui est singulier, puisqu'on modifie conceptuellement une étape consensuelle, la réduction de dimension. Il serait intéressant de regarder en particulier à quel point une réduction de hubness permettrait ou non d'améliorer l'interprétation biologique de données non étiquetées, qui se fait manuellement à la fin de l'analyse.

Ensuite, je me suis intéressée aux cellules T, d'abord via le prisme des lymphocytes T régulateurs. Ces cellules, définies initialement par leur fonction, sont difficile à isoler chez l'homme. J'ai d'abord formé l'hypothèse qu'il y a potentiellement décorrélation entre le phénotype et la fonction, et me suis donc intéressée au contexte. En prenant l'exemple des lymphocytes T régulateurs dans les cancers humains, j'ai montré le poids du contexte dans la détermination du rôle pronostic des lymphocytes T régulateurs dans cinq cancers: sein, poumon, ovarien, colorectal et gastrique. En effet, en partant d'une situation intriquée, dans laquelle il est difficile de démêler le rôle pronostic des lymphocytes T régulateurs dans le cancer humain, avec des articles expérimentaux contradictoires et des revues qui ne peuvent pas trancher clairement dans un sens ou l'autre, non seulement on améliore le consensus vis-à-vis de ce rôle pronostic pour chacun de ces cinq cancers pris individuellement, mais on peut mieux comprendre le rôle pronostic global des lymphocytes T régulateurs dans les cancers humains. A la lumière de la méta-analyse que j'ai conduite, je favorise donc l'hypothèse d'un rôle unique des lymphocytes T régulateurs pour le cancer, plutôt qu'un rôle spécifique en fonction de la localisation de la tumeur. En particulier, j'ai pu relever que les lymphocytes T régulateurs $CD45^-$ étaient systématiquement de mauvais pronostic, quel que soit le cancer étudié. En outre, j'ai observé que le tissu utilisé pour extraire et dénombrer les lymphocytes T régulateurs avait aussi une importance : il semble ainsi que les lymphocytes T régulateurs issus du sang ne soient pas utilisables pour la définition d'un rôle pronostic, tandis que c'est le cas pour les lymphocytes T régulateurs issus du tissu malade. De la même manière, quantifier les lymphocytes T régulateurs via un ratio par rapport à une autre population de cellules immunitaires permet de clarifier leur rôle pronostic, sans doute parce que la prise en compte d'une autre population cellulaire permet de rendre en partie la complexité du micro-environnement tumoral, et notamment du micro-environnement immunitaire. Une perspective de ce travail serait de prendre en compte d'autres paramètres ayant une influence sur le contexte, tel que le traitement. Il sera aussi possible d'appliquer cette méthodologie à d'autres types cellulaires pour lesquels le message pronostic demeure flou, tels que les Th2 et les Th17.

En utilisant toujours cette hypothèse de décorrélation entre le phénotype et la fonc-

tion, j'ai ensuite élargi mon cadre d'étude à l'ensemble des cellules T en questionnant le paradigme actuel de lignée. Le fil directeur a été de supposer qu'une classification, non plus phénotypique, mais fonctionnelle serait plus pertinente, en particulier dans des contextes pathologiques, tels que le contexte tumoral. Cette classification fonctionnelle permettrait en outre de résoudre l'incapacité du paradigme de lignée à prendre en compte de façon précise de la plasticité cellulaire. J'ai donc adopté une approche supervisée dans l'analyse des données de séquençage scRNAseq afin de capturer la fonctionnalité des cellules T. Nous avons d'abord défini un ensemble de 15 fonctions réalisées par les cellules T, telles que la cytotoxicité, la prolifération ou la migration. Chaque fonction, ou module fonctionnel, a ensuite été peuplé de gènes effecteurs. J'ai considéré un gène comme étant effecteur si et seulement si il permet de réaliser la fonction considéré. Cela signifie qu'il y a une preuve expérimentale que la fonction disparaît ou est atténuée (car suppléé potentiellement par d'autres gènes effecteurs) si le gène n'est plus fonctionnel. A l'aide de ces modules fonctionnels, j'ai pu relier chaque cellule à sa/ses fonction/s, en attribuant un score pour chacune des quinze fonctions par cellule. Ces scores en question ont été calculés de sept façons différentes, et j'ai pu valider quatre méthodes d'encodage. Une méthode a été éliminée car elle n'exhibait aucune différence entre les cellules, tandis que deux méthodes supplémentaires ont été exclues car elles n'ont pas passés les tests de contrôle positifs, effectués à l'aide de population pures séquencées en masse. J'ai d'abord prouvé la valeur ajoutée de cette approche par rapport à une analyse classique non supervisée. Plus précisément, j'ai évalué la quantité d'information mutuelle entre l'approche classique et l'approche fonctionnelle, par exemple en comparant les coordonnées des cellules en composantes principales ou en scores fonctionnels. J'ai aussi évalué l'entropie fonctionnelle des clusters obtenus par une analyse classique. Ensuite, j'ai effectué une étape de clustering consensuel, en mélangeant l'information de compte ARN et l'information fonctionnelle. J'ai finalement caractérisé les différences fonctionnelles entre cellules T issues d'un tissu sain ou cancéreux. On retrouve, par cette approche, que le tissu sain juxtatumoral est plus dormant et moins hétérogène que le tissu tumoral. L'intérêt de cette approche réside dans la facilité de l'interprétation des résultats obtenus après analyse. Par rapport à la méthode classique, on s'épargne l'étape fastidieuse et biaisée d'annotation, et notamment d'annotation fonctionnelle. En effet, dans un contexte clinique, il m'apparaît comme essentiel de relier les cellules récoltées dans les tissus malades à leur fonction plus qu'à leur identité cellulaire, car c'est leur fonction qui va guider la décision de savoir si ces cellules sont bénéfiques ou maléfiques, et donc si elles doivent être éliminées ou enrichies. Par exemple, les lymphocytes T régulateurs dans le micro-environnement tumoral vont être bénéfiques ou maléfiques, et devront donc être ciblés pour être éliminés ou enrichis, selon qu'ils seront ou non immuno-régulateurs. Nous avons aussi implémenté cette méthode dans un contexte pathologique additionnel, qui nous a servit comme preuve de concept : l'analyse de cellules dendritiques de patients souffrant de la Covid-19, après sélection des modules fonctionnels idoines, c'est-à-dire des fonctions effectuées par les cellules dendritiques. En résumé cette stratégie d'analyse fonctionnelle peut donc être appliquée pour d'autres types de cellules que les cellules T, d'autres pathologies que le cancer, et même dans un contexte physiologique, afin de cartographier les fonctions des cellules immunitaires.

Mots clés : séquençage d'ARN en cellule unique, analyse de données omiques, grande dimension, fonctionnalité, approche supervisée, cancer, immunologie, bioinformatique, méta-analyse.