UNIL | Université de Lausanne
Faculté de biologie
et de médecine

MelanomX
Tumour Microenvironment
Crosstalk in Melanoma
Adaptive Resistance

Identification of immune cells in melanoma with single-cell RNA-sequencing

Master Thesis in Medicine N° 3377

Alexandre WICKY

Co-tutors: Prof. David Gfeller and Prof. Olivier Michielin, Department of Oncology, CHUV-UNIL

Expert: Prof. Giovanni Ciriello, Department of Computational Biology, UNIL

Lausanne, December 15th, 2016

ACKNOWLEDGMENT

While I am the only author mentioned in this report, there are many great people who remarkably contributed to this project.

First, I would like to take the opportunity to thank David Gfeller for his direct supervision and guidance during the whole process of this thesis. Especially for the patience and pedagogy he showed while he was explaining me some maths. I am also greatly thankful to Santiago Carmona and Julien Racle, two Post-docs at Gfeller's lab, who show me the way of coding and with whom I had a great time discussing about science. They became more than colleagues, they are now true friends. I am profoundly grateful for these invaluable moments I spent at the lab with the whole team. It has been an extraordinary period of learning, not only bioinformatics but also on a personal level.

I also want to thank Olivier Michielin for giving me the opportunity to work on this project and for his advices and support concerning my future medical career. In addition I would like to thank the MelanomX team for their brilliant collaboration and Giovanni Ciriello, my thesis expert, for his participation in this project.

Finally there are my family and Lara who support me and feed me during the writing process of this report. Thank you all ☺

TABLE OF CONTENTS

ABSTRACT

This project aims to determine the cellular identity of the immune cells composing the tumor microenvironment of a melanoma mouse model using single-cell RNA-sequencing (scRNA-seq). The recent development in the scRNA-seq technology has now made the full transcriptome of individual cells within reach and therefore enables us to characterize the identity of cells according to their full gene expression. However the substantial stochasticity and high dropout rates of scRNA-seq data challenge the process of cell type identification. To overcome these limitations I apply here a computational approach that takes advantages of immune transcriptomic repositories to define cell type specific gene markers in order to quantify the cellular identity of immune cells in melanoma. The performance of the identification algorithm has been first tested with cross-validation analysis of sorted cell bulk gene expression profiles. Then, the cell type identification algorithm has been applied on scRNA-seq data and revealed that the melanoma immune microenvironment from our mouse model is composed of a large proportion of T and NK cells.

Keywords: Single-cell RNA-seq, Bioinformatics, Immunology, Melanoma

INTRODUCTION

Tumor is a complex ecosystem, not only composed of cancer cells but also with various other cells such as endothelial cells, cancer-associated fibroblasts or immune cells[1,2]. The latter appear to play an important role in the development or control of the tumor and the promising results of immunotherapy against cancer have confirmed this major role[3-6]. Therefore analyzing the composition and the interplay of cells in the tumor microenvironment (TME) is crucial for understanding how tumor grow and resist to the destruction by the immune system.

In this attempt to characterize the immune landscape of tumors, single-cell RNA sequencing (scRNA-seq) represents a powerful method to profile simultaneously each individual cells by obtaining their full transcriptome[7,8]. This technical achievement has been made possible by the combination of cell isolation like with the fluidigm-C1 technology[9] or Fluorescence-activated cell sorting (FACS) and the improvement of high-throughput sequencing technology.

With the availability of the whole transcriptome for a given cell, it now opens the gate of defining the cellular identity according to the expression of all genes and not only based on cell surface proteins markers or morphology as it has been predominantly done so far in biology. There are currently two commonly used methods to differentiate cells among each other: supervised and unsupervised

clustering technique. While supervised clustering separates cells based on a priori information (categories are known and assignation depends on the expression on some known gene markers for example), unsupervised classification in the other hand requires no a priori information and will separate cells according to their intrinsic propriety (for instance the expression of all their genes) but can be highly influenced by batch effect[10,11]. The purpose of each method is also different: where unsupervised clustering would be more suited for de novo cell type identification, supervised clustering is more appropriate for validation of known cell types from transcriptomic samples.

Enumeration and classification of cell types in immunology has been of great concern since the beginning of the field, as illustrated by the number of cluster of differentiation (CD) proteins discovered and used as markers so far. Some of these classifications remain controversial, in particular for rare cell types but the major hematopoietic cell types are well validated and described. Gene expression profiles of well-defined immune cell types are also freely accessible in a website (www.immgen.org) from the Immunological Genome Project (ImmGen)[12]. The ImmGen project is composed of immunologists and bioinformaticians that have used highly standardized cell sorting strategy (FACS) and sequencing procedure (Micro-array) to provide well curated expression profiles of murine immune cells.

By taking advantage of the immune profiles published by ImmGen I have proposed to describe the immune cell population in a murine tumor model of melanoma from the MelanomX project. However to define the identity of a particular cell type in tumor with supervised classification, I needed a method to define cellular identity by selecting cell type specific genes. I found an answer to this problem with the paper published by Birnbaum & Kussell[13] and Efroni *et al*[14]. They have proposed a method where they quantified the cellular identity by using gene markers that were first selected and which received a score reflecting the amount of cell type specific information they carried. In other words, this method allows to classify genes according to their specificity for a given cell type and therefore allows to define subsets of genes that can be used for supervised classification of cells. The number of marker genes to use remains an open question, because with single-cell RNA-seq, the amount of gene expressed is rather low and stochastic and by using too few markers we could miss to detect one cell type. On the other hand, by using too many markers we risk to lower the specificity of the cell type identification. An optimal number of markers is therefore needed.

Through this project I have selected immune cells profiles from ImmGen and defined their associated marker genes in order to potentially identify these cells later within scRNA-seq data from the MelanomX project. In order to validate the method I carried out a first phase of cross-validation with

the leave one out method and many key parameters of the identification algorithm used were studied in order to optimize the identification process for further classification of single-cell transcriptomic data.

RESULTS

PART I: SELECTION OF IMMUNE CELL TRANSCRIPTOMIC PROFILES FROM IMMGEN

I selected 110 microarray expression data over the ~270 mice leukocytes subsets available in the ImmGen.org website. The criteria to select these profiles in particular was their global representability of emblematic hematopoietic cell types, the exclusion of genetically modified mice and the evaluation of the experiment, the body location and cell-sorting strategy. Monocytes or bone marrow precursors were not selected because I was mostly interested in more differentiated cells, typically those that we can find in tissue or tumors. Moreover monocytes are somewhat present in the data set in the form of Macrophages or DCs as these cells arise from differentiated monocytes. Sadly, no Myeloid-derived-suppressor cells (MDSC) sample was available in ImmGen and the tumor-infiltrating lymphocytes (TILs) samples were not available for downloading despite my emails.

In order to refine the immune transcriptomic profiles collected, a first step was to visualize the clustering of profiles in 2 dimensions using three main dimensionality reduction methods (t-Distributed Stochastic Neighbor Embedding (t-SNE), Multidimensional Scaling (MDS) and Principal Component Analysis (PCA)) (Figure 1). Then some profiles were excluded or reassigned based on the clustering groups. For example in Figure 1A, a particular B cell sample clusters with activated CD8 T cells on the right of the graph. This sample has been sorted differently from other B cells and has been defined as germinal center B cells, likely to cell-cycle just after activation. Therefore to avoid overlapping of clusters, this sample has been removed for further computation. This example depicts the selection method used and further analysis has been made to exclude or characterize other cell types. For example the definition of M1 or M2 Macrophages relies mainly on the expression of Stat1 and Stat6 respectively (even though the separation of M1 or M2 macrophages is controversial in immunology because it is not a definitive lineage commitment but it is more seen as a phenotypic plasticity)[15]. Thus these two groups were also defined by using specific gene expression such as Stat1 and Stat6 (showed in Figure 2).

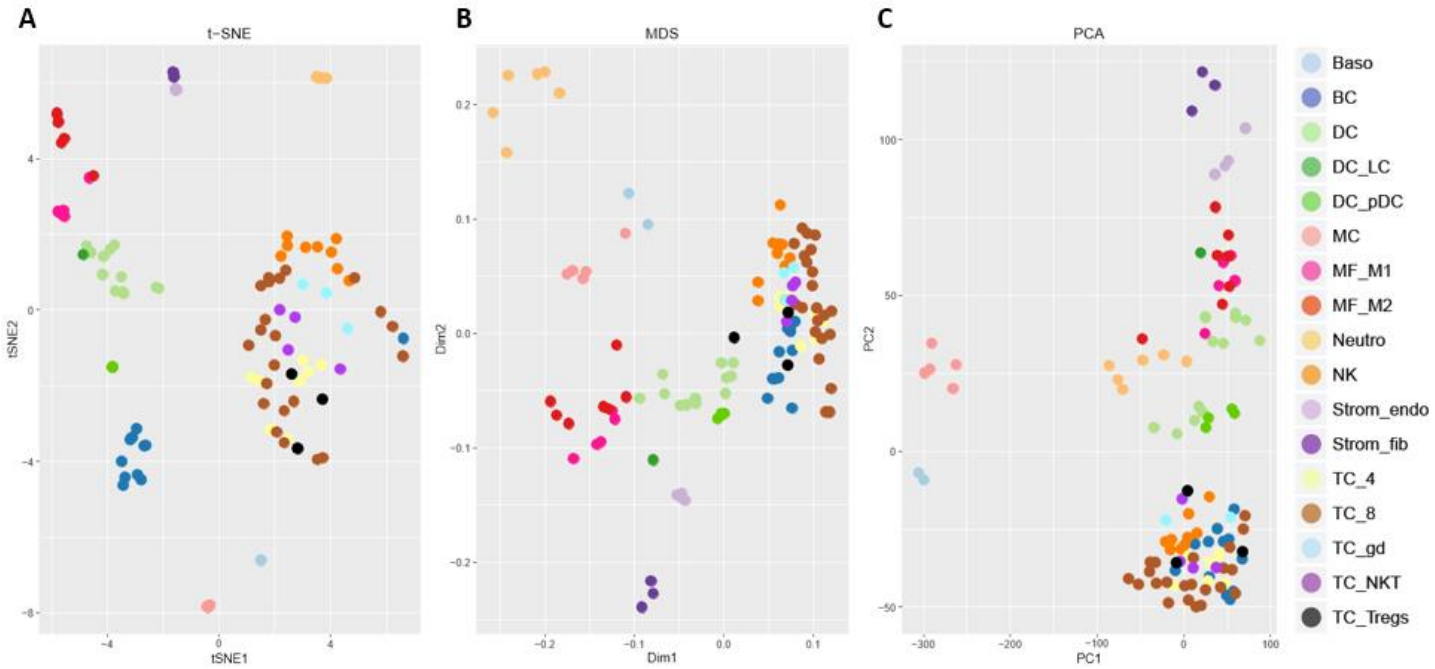

Figure 1. Dimensionality reduction for 110 ImmGen immune transcriptomic profiles. Three methods to reduce the dimensionality have been used: t-SNE, Multidimensional scaling (MDS) and Principal component analysis (PCA). 18 major cell types are show on this graph according to the color legend on the right. In addition to the evident abbreviation of immune cells, here are the less obvious ones: Baso = Basophil, BC = B cells, DC_LC = Langerhans cells, MC = Mast cells, MF_M1 = M1 "inflammatory" Macrophages, MF_M2 = M2 "tissue-repair" Macrophages, Strom_endo = Stromal endothelial cells, Strom_fim = Stromal fibroblast cells, TC_4 = T cells CD4+, TC_gd = T cells gamma-delta

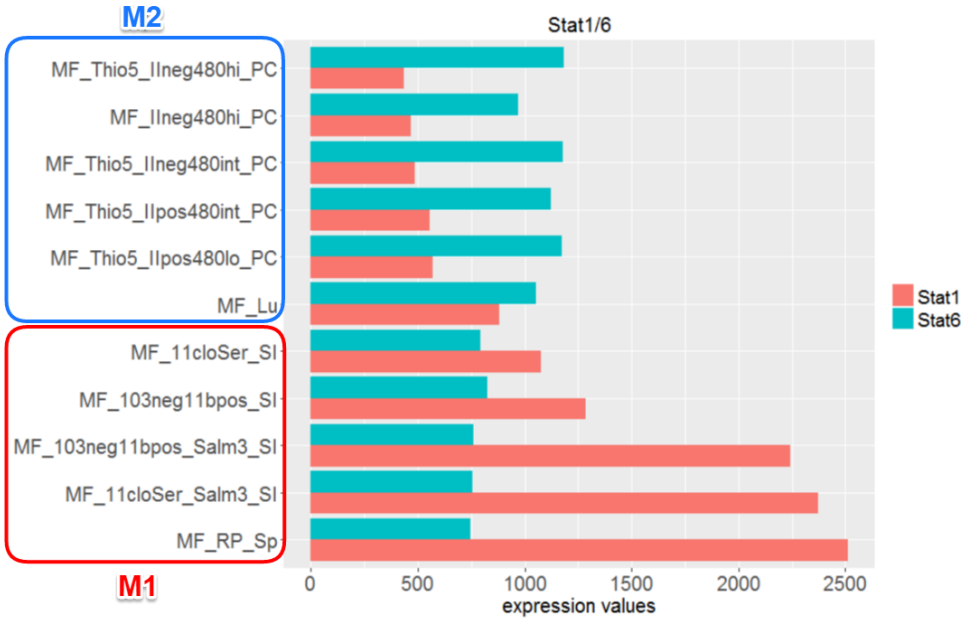

Figure 2. Stat1 and 6 expression in macrophages subsets. M1 Macrophages stands for an inflammatory profile while M2 is associated with the "tissue repair" profile of Macrophage. Expression values are relative expression values from the ImmGen microarray and y-axis labels corresponds to ImmGen samples names. Values were normalized across 3 independent microarrays and true gene expression was considered above 120.

As observed in Figure 1, the lymphoid samples tend to cluster together when all samples are analyzed in the same time. Thus I have separated lymphoid and myeloid samples and compare their pattern of organization based on expression profiles with the t-SNE method for each group (Figure 3). On Figure 3A, T cells subsets display good clusters of NK and B cells however naïve CD4 and CD8 T cells appear to cluster together. In consequence these clusters have been renamed as naïve T cells in celltypes3 (Table 1). It should be noted that Treg samples did not cluster closely together and were rather dispersed along the CD4 T cells memory cluster. This group was nevertheless maintained in the analysis despite its dispersion. In contrast to lymphoid cells, myeloid cells showed a better clustering according to cell types. The only exception is the relative separation of pDC from the DC cluster.

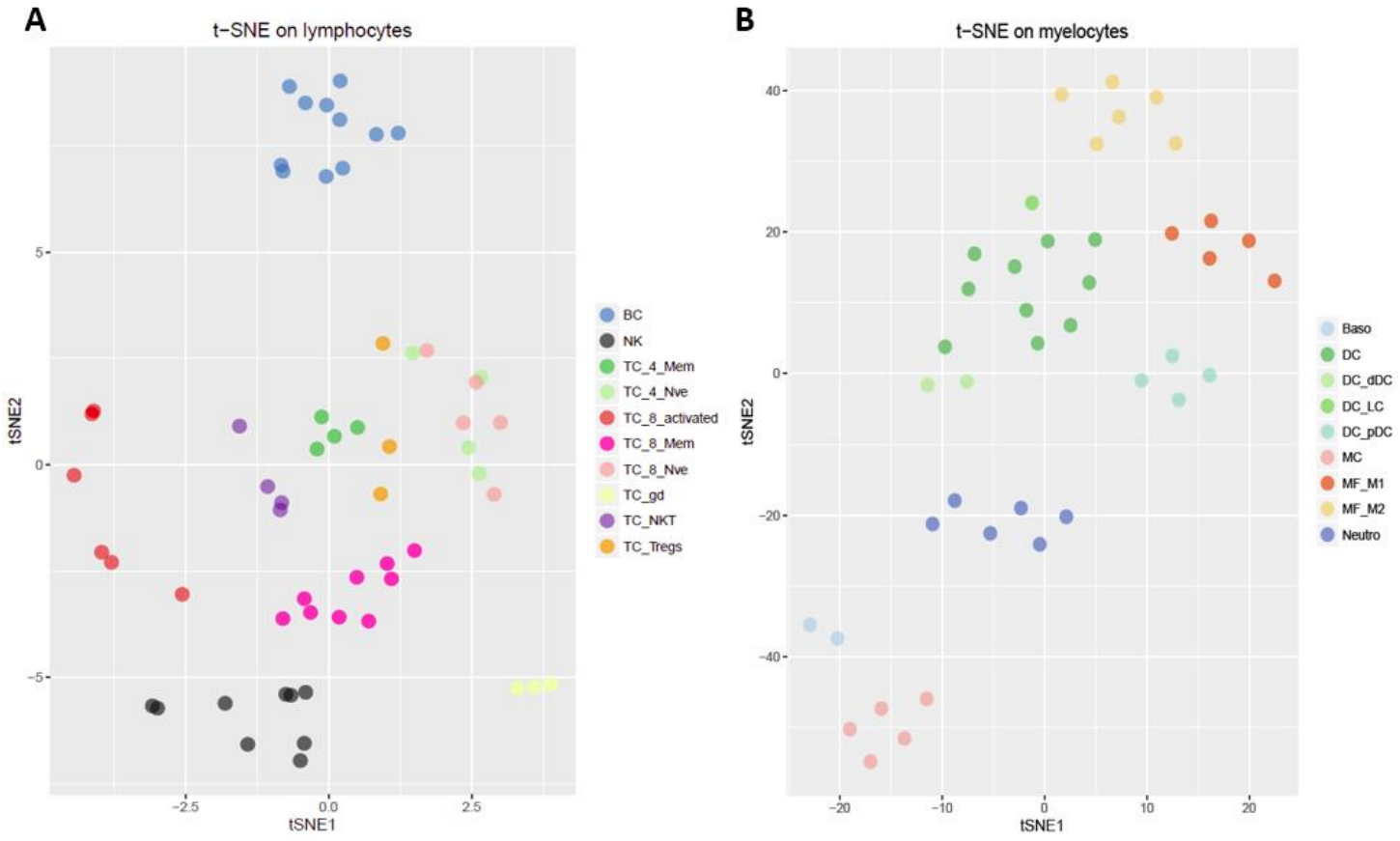

Figure 3. T-SNE dimensionality reduction for ImmGen lymphoid and myeloid cells transcriptomic profiles.
TC_4_Nve = CD4 T cells naïve. TC_8_Nve = CD8 T cells naïve.

PART II: DETERMINATION OF CELL TYPE GENE MARKERS AND CELLULAR IDENTITY

DEFINING CELL TYPE-SPECIFIC SCORE (SPEC SCORES)

In order to identify gene markers for defining a cellular identity, cell type specific scores called Spec scores were calculated as described by Birnbaum and Kussell[13] (see mathematical explanation with an idealized example in Figure 4). These Spec scores were calculated relatively to other cell types and therefore depended on the level of definition of cell types. For example, a given gene would not have the same value of Spec if it is limited to the classification of lymphoid/myeloid cells or if one wishes to obtain more precise information with subgroups such as Treg or pDCs for instance. The Spec scores were therefore established for different level of cell type definition (n=3, Table 1) using 106 expression data derived from ImmGen microarray profiles. The first five gene markers with the highest Spec score for ten different cell types are given in Table 2. We can observe the presence of emblematic gene markers for lymphoid cells such as Cd3 or Cd27 for T cells, Cd19 for B cells and killer cell lectin-like receptor for NK cells. In the other hand, the five better markers for myeloid cells are less common, but there is also less specific markers known in immunology for myeloid cells subgroups than for lymphoid cell subgroups. Similar findings were observed for stromal cells with the exception of Tek, an endothelial cell-speific gene, whose presence as specific marker can be explained by the fact that the identity "stromal cells" is composed of 4 profiles of endothelial cells and 3 profiles of fibroblasts.

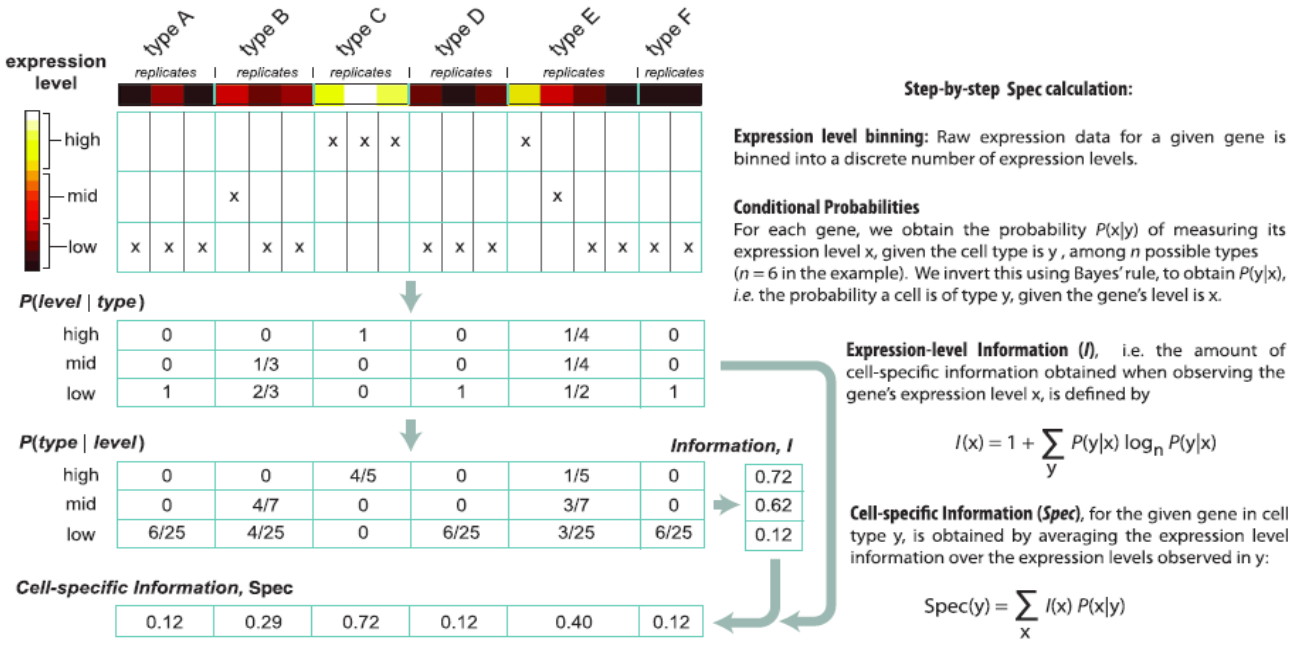

Figure 4. Example of the computation of Spec scores with an idealized profile of expression with one gene in 6 cell types with replicates. The mathematics is explained on the right and a schematic example of the binned expression of a gene in 6 different cell types on the left. Reprinted from Birnbaum KD and Kussell E: Measuring cell identity in noisy biological systems. Nucleic Acids Res. 2011.

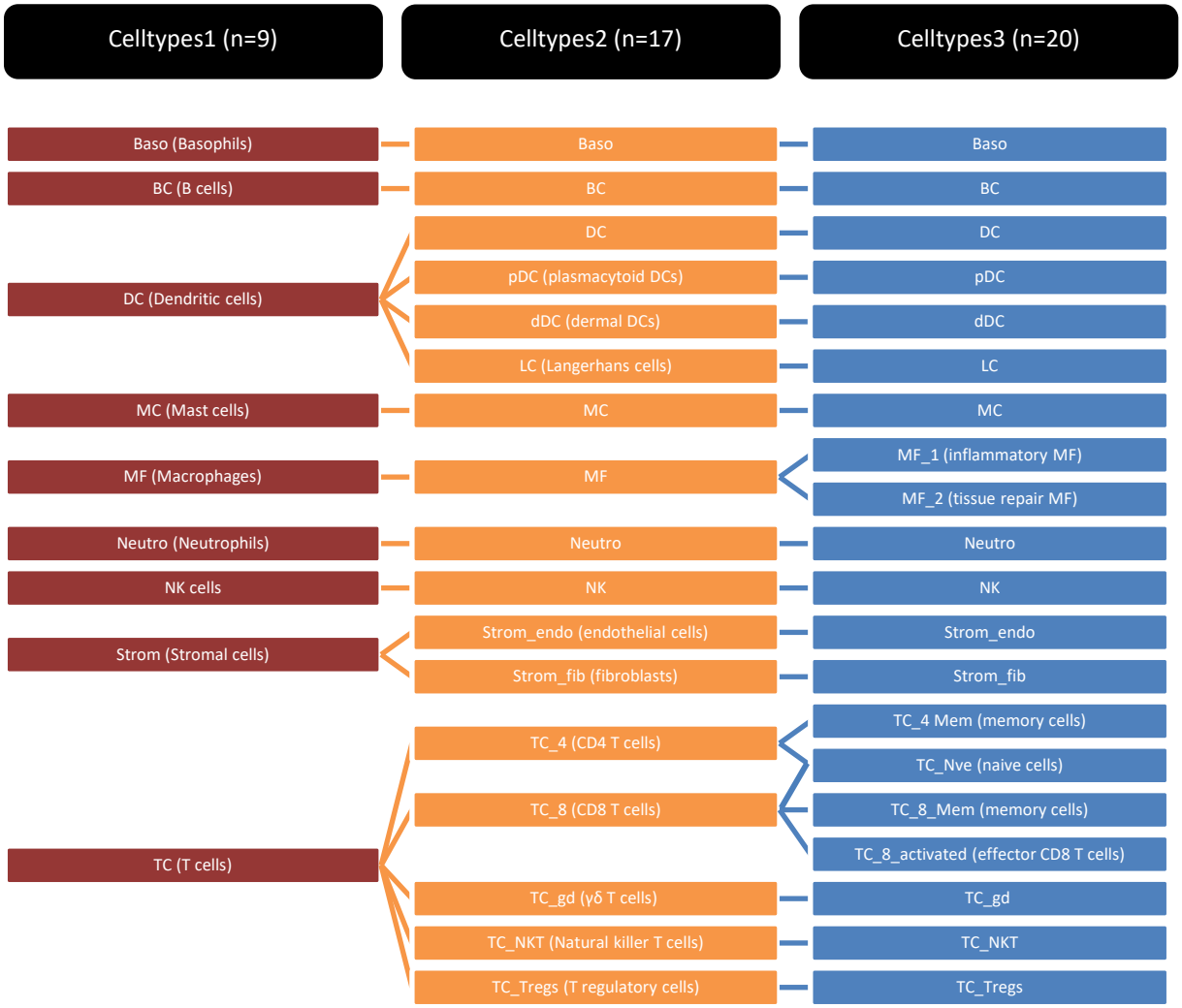

Table 1. Definition of immune cell type profiles used for identification.

CELL TYPE	GENE SYMBOL	SPEC	DESCRIPTION
T cells	Cd3g	1	CD3 antigen, gamma polypeptide [MGI:88333]
T cells	Cd3d	1	CD3 antigen, delta polypeptide [MGI:88331]
T cells	Cd3e	1	CD3 antigen, epsilon polypeptide [MGI:88332]
T cells	Bcl11b	1	B cell leukemia/lymphoma 11B [MGI:1929913] (involved in T cell development)[16]
T cells	Cd27	0.97	CD27 antigen [MGI:88326]
B cells	Pou2af1	1	POU domain, class 2, associating factor 1 [MGI:105086]
B cells	Cd79b	1	CD79B antigen [MGI:96431]
B cells	Pax5	1	paired box 5 [MGI:97489]
B cells	Cd19	1	CD19 antigen [MGI:88319]
B cells	Ms4a1	1	membrane-spanning 4-domains, subfamily A, member 1 [MGI:88321]
NK	Ncr1	1	natural cytotoxicity triggering receptor 1 [MGI:1336212]

NK	Klra9	1	killer cell lectin-like receptor subfamily A, member 9 [MGI:1321153]
NK	Klri2	1	killer cell lectin-like receptor family I member 2 [MGI:2443965]
NK	Klra10	1	killer cell lectin-like receptor subfamily A, member 10 [MGI:1321093]
NK	Klre1	1	killer cell lectin-like receptor family E member 1 [MGI:2662547]
MF	Tmem106a	1	transmembrane protein 106A [MGI:1922056]
MF	Cndp2	1	CNDP dipeptidase 2 (metallopeptidase M20 family) [MGI:1913304]
MF	Abhd12	1	abhydrolase domain containing 12 [MGI:1923442]
MF	Nhlrc3	1	NHL repeat containing 3 [MGI:2444520]
MF	C1qc	0.91	complement component 1, q subcomponent, C chain [MGI:88225]
DC	Flt3	0.94	FMS-like tyrosine kinase 3 [MGI:95559]
DC	Syngr2	0.86	synaptogyrin 2 [MGI:1328324]
DC	Paxbp1	0.82	PAX3 and PAX7 binding protein 1
DC	Uvrag	0.82	UV radiation resistance associated gene [MGI:1925860]
DC	Mycl	0.76	v-myc avian myelocytomatosis viral oncogene lung carcinoma derived
Neutro	S100a9	1	S100 calcium binding protein A9 (calgranulin B) [MGI:1338947]
Neutro	Lcn2	1	lipocalin 2 [MGI:96757]
Neutro	Chi3l1	1	chitinase-like 1
Neutro	Csf3r	1	colony stimulating factor 3 receptor (granulocyte) [MGI:1339755]
Neutro	Spatc1	1	spermatogenesis and centriole associated 1 [MGI:1921531]
MC	Mcpt4	1	mast cell protease 4 [MGI:96940]
MC	Olfr920	1	olfactory receptor 920 [MGI:3030754]
MC	Tpsb2	1	tryptase beta 2 [MGI:96942]
MC	Ctsg	1	cathepsin G [MGI:88563]
MC	Tpsab1	1	tryptase alpha/beta 1 [MGI:96943]
Baso	Hgf	1	hepatocyte growth factor [MGI:96079]
Baso	Mcpt8	1	mast cell protease 8 [MGI:1261780]
Baso	Htr1b	1	5-hydroxytryptamine (serotonin) receptor 1B [MGI:96274]
Baso	Dnase2b	1	deoxyribonuclease II beta [MGI:1913283]
Baso	Hist1h2ac	1	histone cluster 1, H2ac [MGI:2448287]
Stromal	Enpp2	1	ectonucleotide pyrophosphatase/phosphodiesterase 2 [MGI:1321390]
Stromal	Nfib	1	nuclear factor I/B [MGI:103188]
Stromal	Tek	1	endothelial-specific receptor tyrosine kinase [MGI:98664]
Stromal	Parva	1	parvin, alpha [MGI:1931144]
Stromal	Timp3	1	tissue inhibitor of metalloproteinase 3 [MGI:98754]

Table 2. First 5 gene markers with the highest Spec scores for 10 cell types from Celltypes1.

DETERMINATION OF AN INDEX OF CELLULAR IDENTITY (ICI)

After the computation of Spec scores with the ImmGen profiles, the genes with the highest Spec scores were chosen as marker in order to identify a cell type in other data sets such as single-cell expression data. In order to be as consistent as possible between cell types, the number of markers chosen depended on the values of Spec scores and was based on the cumulative sum of each Spec score that reached a constant (the first markers were added until their cumulative sum of Specs scores reached a predetermined value). This constant was arbitrarily fixed at 10 after evaluation of the results obtained during the cross validation process (see next part). This constant of 10 resulted in about 10 genes selected if the Spec scores were high and up to about 12 if they were low (meaning less specific). This number represented a good compromise, because single-cell expression data is likely to exhibit scarce genes expression and with too few markers we would endorse the risk of missing cell types. Additionally with too many markers we take the risk of adding less specific markers and therefore we might risk returning wrong profiles.

The identification of cell types from single-cell or even cell-sorted bulk expression data was obtained by using an index of identification (**ICI**) for every cell types predefined with the Spec computation. This score of identification for a given cell type (t) was calculated by multiplying the expression values (e_g) (TPM or micro-array expression values) of the predefined gene markers (g) by their corresponding Spec score ($s_{g,t}$) and then averaged by the number of markers used (n_t). The value was then weighted by the proportion of markers expressed and normalized from 0 to 1.

$$\mathrm{ICI}_t = \frac{\sum_g^{n_t} e_g * s_{g,t}}{n_t} * \frac{\sum_g^{n_t} expressed(g)}{n_t}$$

PART III: CROSS-VALIDATION

In order to evaluate the method of identification I used the method of the Leave-One Out Cross Validation (LOOCV) on the ImmGen samples. Multiple round of identification was made with a procedure that removed one ImmGen sample on each round, then the Specs score were calculated with the remaining dataset and the cell type identification (ICI) was performed on the removed sample. The results of this LOOCV are shown on Figures 5-6 with the immune profiles given in x-axis and the ImmGen samples on y-axis. It has to be mentioned that the Basophils and Langerhans cells samples were excluded of this validation because of their limited number of samples. We could clearly determine 8 general immune identities (celltypes1) for almost all ImmGen samples with a cumulative

Specs score of 10. With the exception of 2 DCs and 2 Macrophages samples that showed mixed identities between their immune profiles. However by taking the max value obtained in the cell type identity, their immune cell type was nevertheless correctly determined.

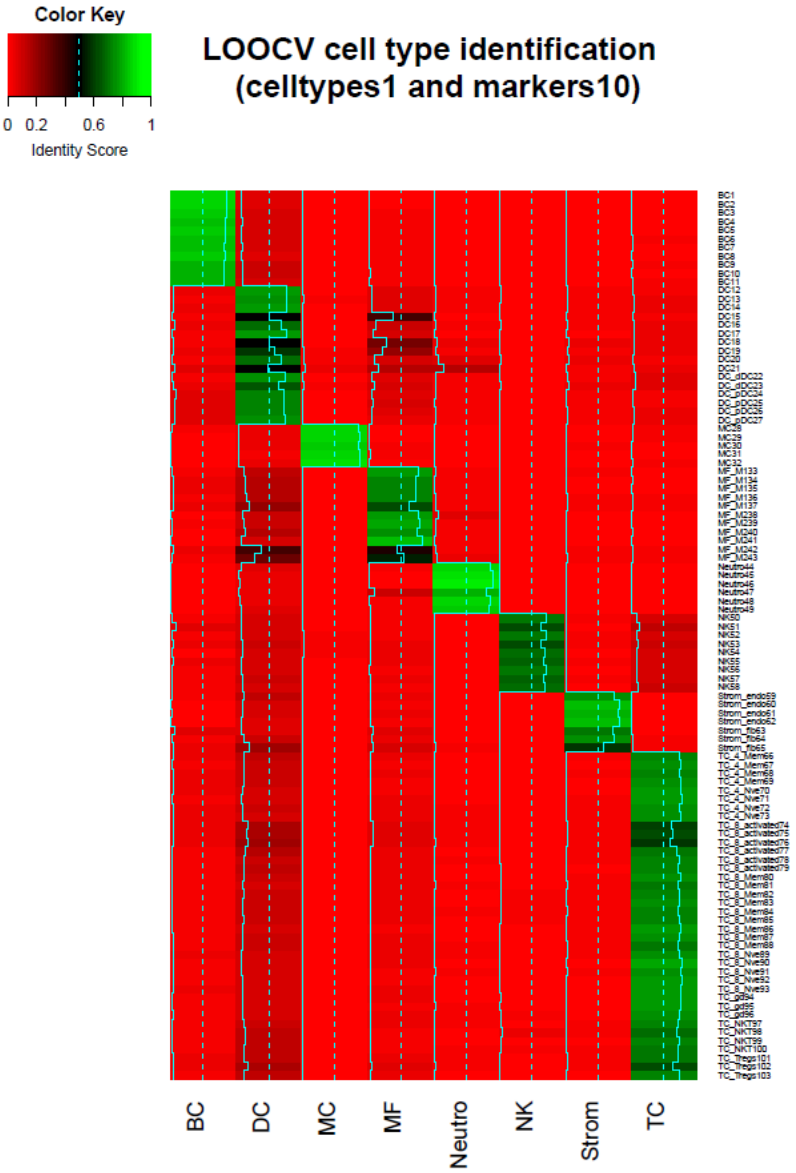

Figure 5 Index of cell identity for the Leave One Out Cross Validation (LOOCV) with cumulative markers information of 10 (markers10) and for a cell type definition of level 1 (celltypes1).

At a higher level of definition, with 14 different immune profiles, the algorithm could relatively well classify most of the profiles at the exception of CD4 and CD8 T cells that remained not clearly different, even though max values returned mostly correct identifications (Figure 6). Increasing the number of cell type to 17 with the differentiation of M1 and M2 macrophages and a different T cells classification

showed the same tendency for T cells and also decrease the identification value for DCs (Figure 7). However M1 and M2 macrophages could be differentiated to a certain extent.

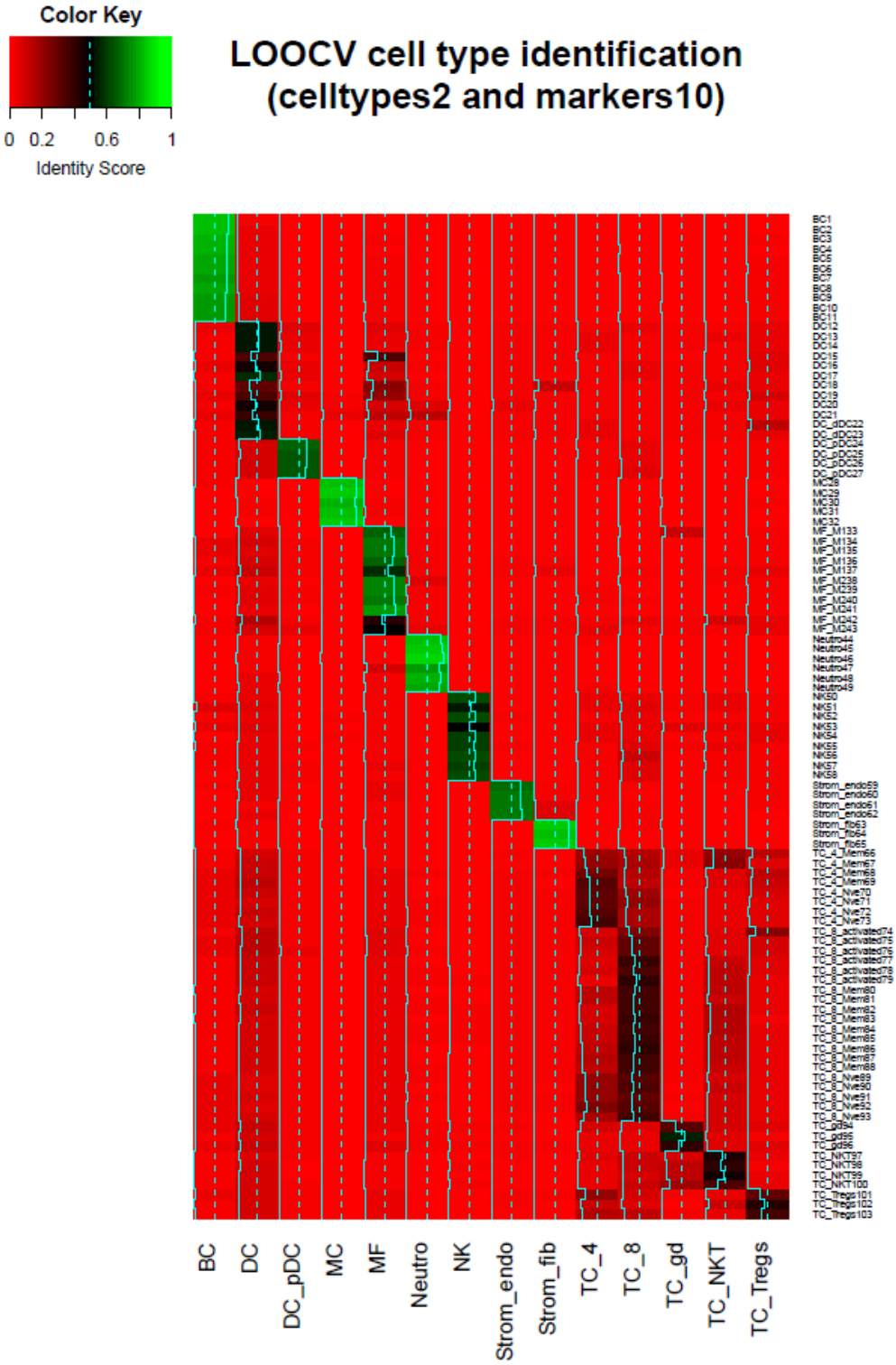

Figure 6. Index of cell identity for the Leave One Out Cross Validation (LOOCV) with cumulative markers information of 10 (markers10) and for a cell type definition of level 2 (celltypes2).

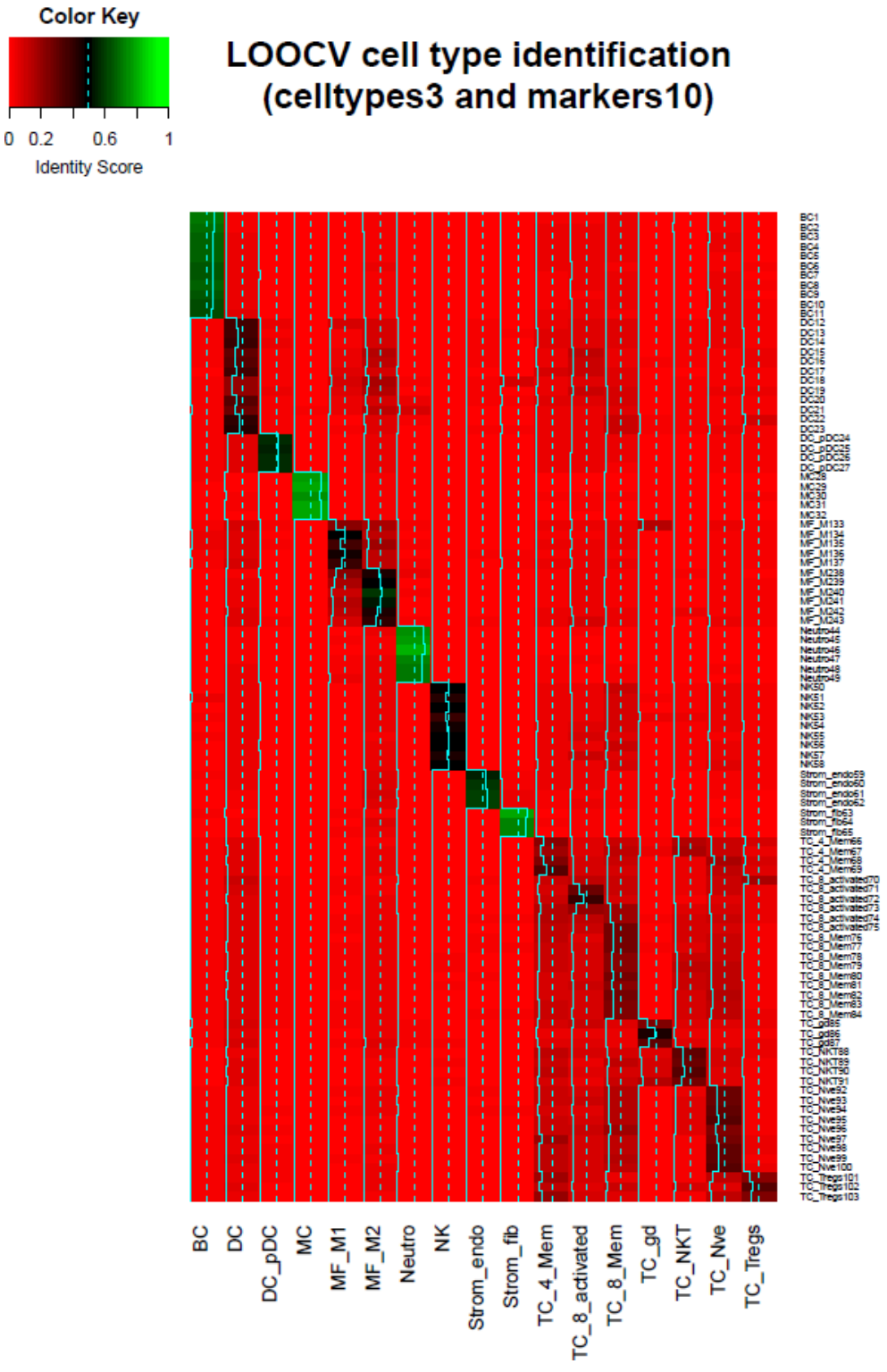

Figure 7. Index of cell identity for the Leave One Out Cross Validation (LOOCV) with cumulative markers information of 10 (markers10) and for a cell type definition of level 3 (celltypes3).

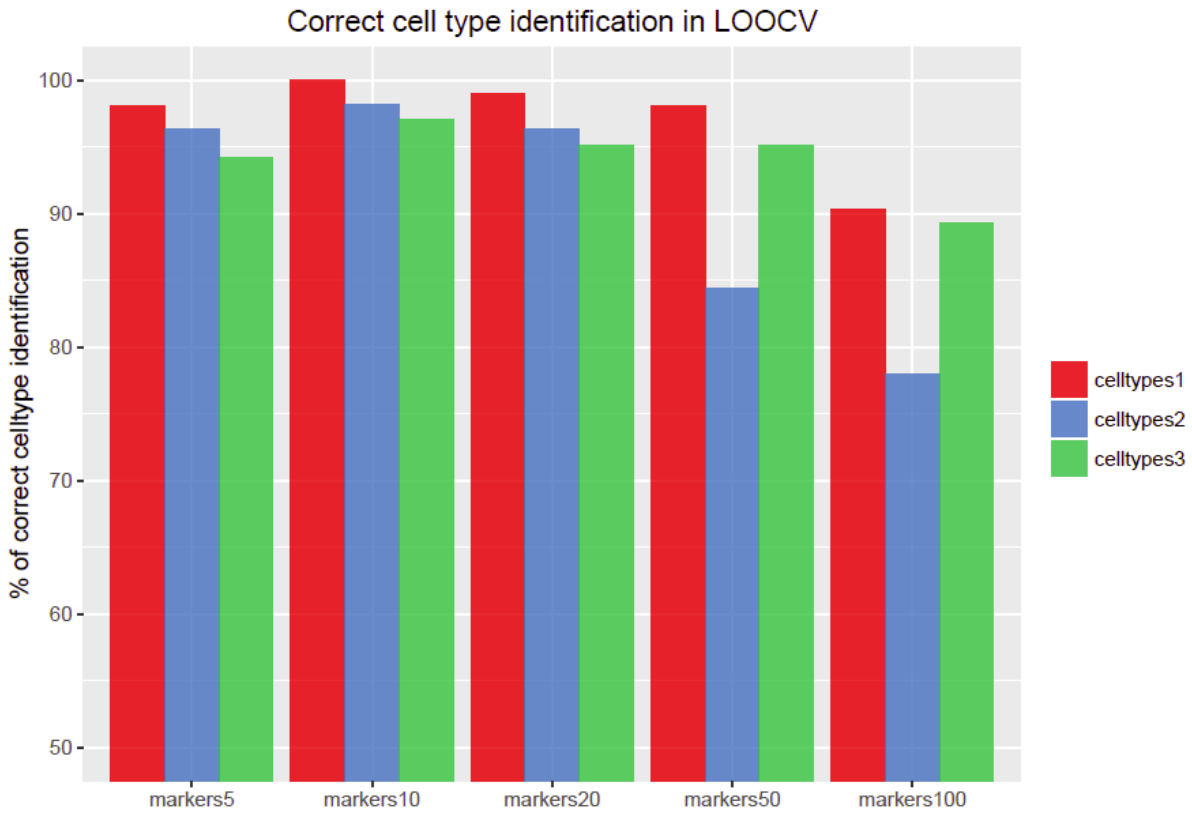

Figure 8. Correct cell type identification in LOOCV for different cell type definition and with different number of markers.

Ultimately the identification algorithm could correctly identify most of the samples if the max value was taken and the cell type definition remained general. The percent of correct identification (when the max value was corresponding to the actual identity of the sample) is given in Figure 8 for 5 different sets of markers used (with an increased number of markers from markers5 to markers100). We can observe a decreased trend of correct identification when augmenting the number of markers. This trend has also been observed by Efroni *et al*, when identifying plant cells.[14]

In order to further improve the identity predictions, I have proposed to use the identification algorithm sequentially by varying the predefined immune profiles. For example the samples identified as T cells by the algorithm identifying broad immune cell types are then analyzed more specifically with predefined immune profiles for determining only T cell subsets. The result of this identification with the LOOCV is shown on Figure 9. T cells subsets were globally identified correctly except for 5 samples with mixed identities. However, the identification remained mainly correct on the basis of the max value for determining the identity. Indeed we can deduce here that diminishing the number of cell types and focusing on one particular group such as T cells for the computation of Specs scores increases the power of discrimination of closely related samples.

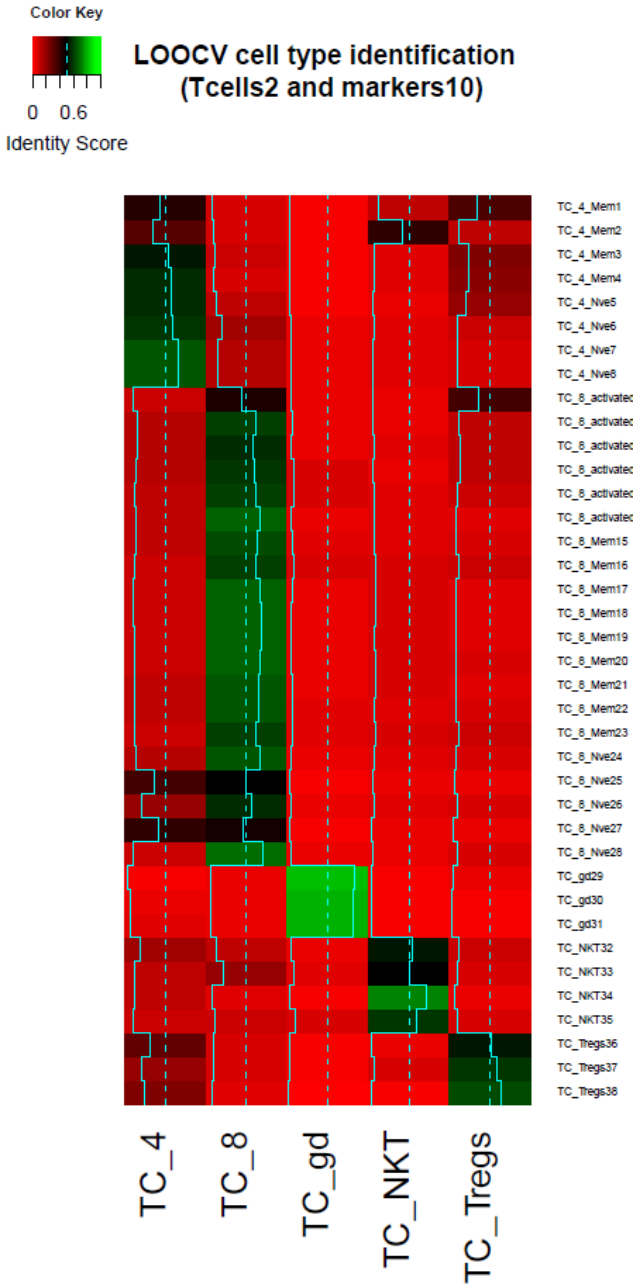

Figure 9. Index of cell identity for the Leave One Out Cross Validation (LOOCV) with cumulative markers information of 10 (markers10) and for T cells with a cell type definition of level 2 (Tcells2).

PART IV: IDENTIFICATION OF IMMUNE CELLS FROM SCRNA-SEQ DATA

To explore the immune cell composition of the TME, I used the scRNA-seq data from the MelanomX project (plate X61, CD45+ cells, cell size of 5-10 µm). In brief, the sequenced cells originated from a genetically-engineered melanoma mouse model called iBIP2 mouse that develops tumor mimicking human melanoma and which was treated with anti-CTLA-4 and anti-PD-1 antibodies. The cells were

rapidly dissociated and sorted with flow cytometry using the CD45+ markers to capture specifically hematopoietic live cells. Then the sorted cells were isolated with the Fluidigm C1 technology, libraries were prepared and analyzed with the Illumina HiSeq sequencing machines. Hereafter, the cells that passed quality control (n=42) were further analyzed with the identification algorithm presented in this project. As seen on Figure 10 most the captured cells had a T cell profile (n=24), 7 cells were identified as NK cells, 1 as a B cell, 2 as neutrophils and 2 others as DCs. The other cells that did not reach a threshold 0.6 for their identity score were considered as undetermined (n=6).

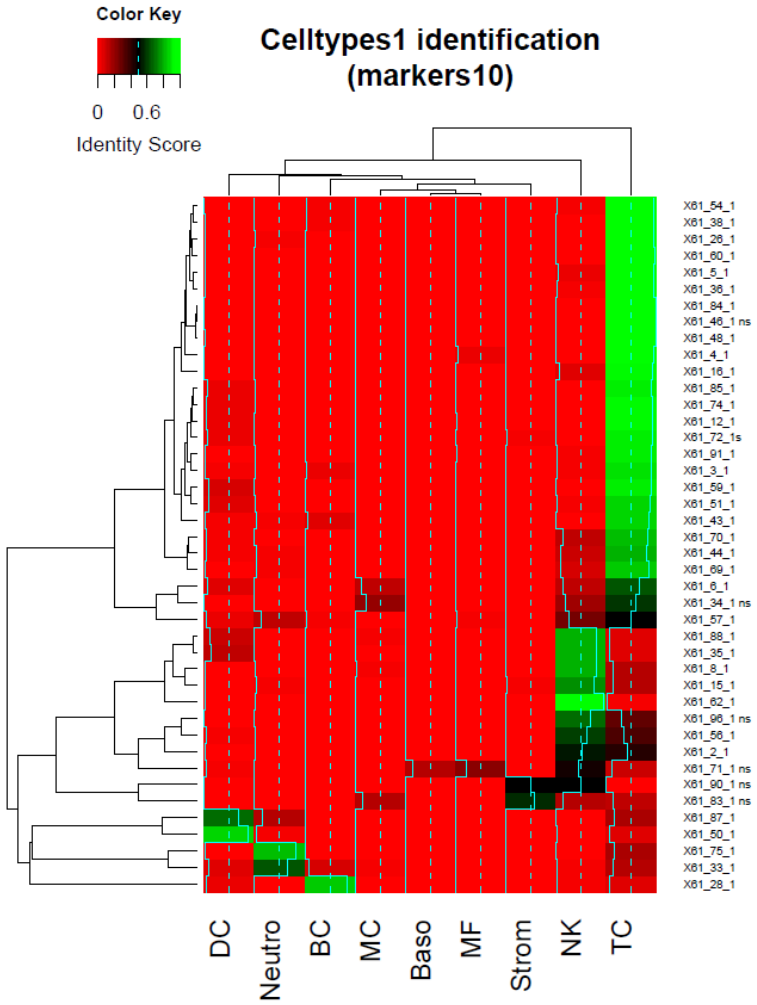

Figure 10. Immune cell type identification of CD45+ cells from the mouse melanoma tumor model. Identification was done with celltypes1 immune profiles and cumulative information threshold of 10 (markers10). All identification value were considered significant $P < 0.05$ (otherwise indicated "ns").

Further identification with celltypes2 profiles (Figure 11) could determine that most T cells were classified as Treg but this has to be confirmed with a more precise identification using only T cells

identification algorithm. Additionally the 7 NK cells were found again, as the 2 neutrophils, the B cell and one DC showed a pDC profile while the other DC did not get a sufficient identification score (even if the max value return a conventional DC profile).

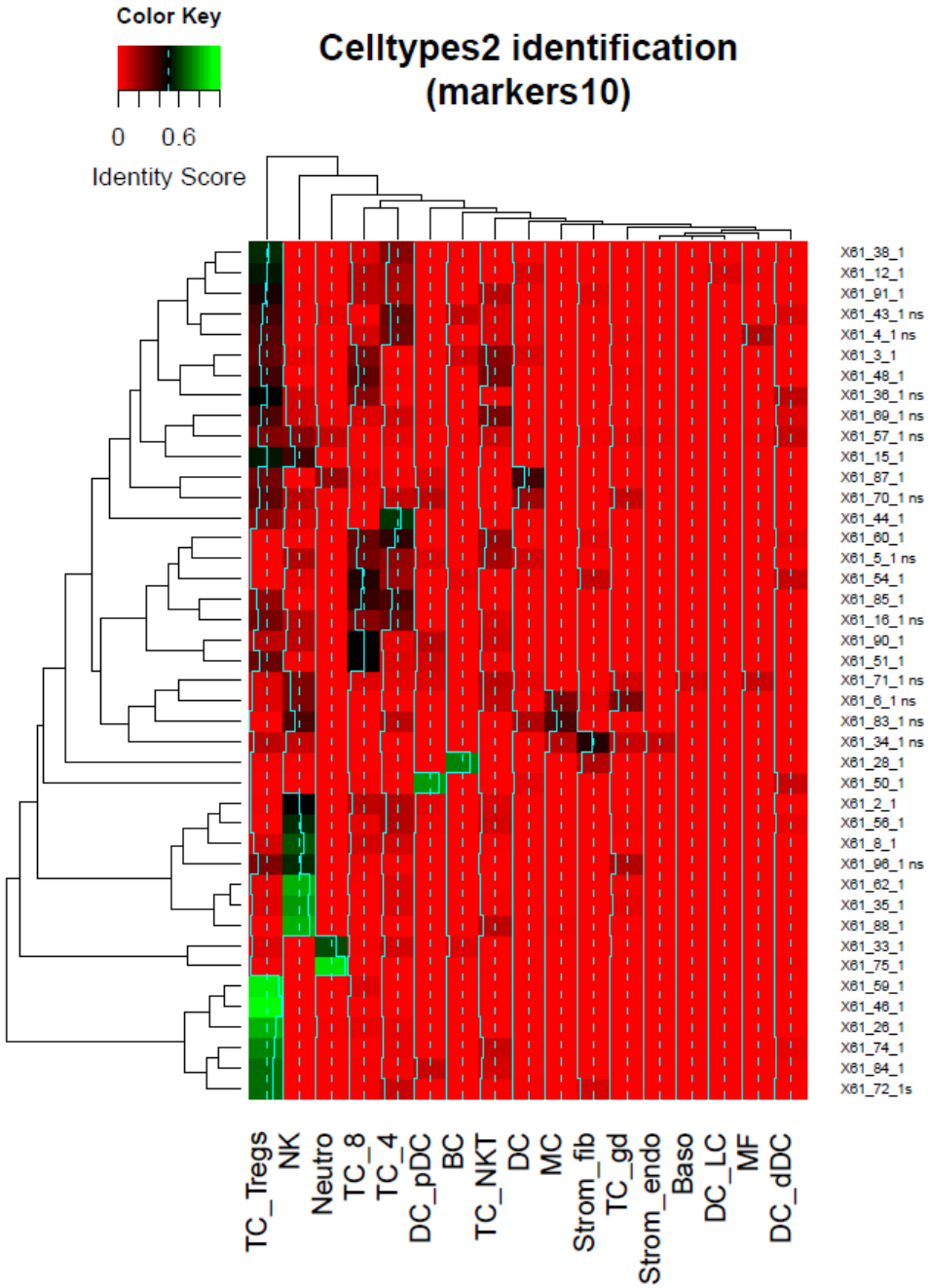

Figure 11. Immune cell type identification of CD45+ cells from the mouse melanoma tumor model using the celltypes2 profiles. Identification was done with a cumulative information threshold of 10 (markers10). All identification value were considered significant $P < 0.05$ (otherwise indicated "ns").

To confirm some identities, specific gene expression of emblematic immune markers has been evaluated one by one and is described hereafter (data not shown in this report except for T cells in the

next part). The cell identified as a B cell was the only one to express the Cd19 gene which is specific for B cells. In addition this cell express MHC class II genes as the two other cells identified as DCs. Moreover the pDC marker Spib was highly expressed in the X61_50 sample (identified as pDC).[17] Neutrophil marker (Csf3r, Fpr1, S100a8, S100a9, Cxcr2) were only expressed in the X61_33 and 75 cells. Interestingly the Ly6g (Gr-1) commonly used in FACS for defining neutrophils was not expressed in any of the cells. Additionally no clear macrophage markers were found, only sporadic expression of Adgre1 (F4/80), which is not really specific for macrophages. For cells identified as NK, numerous killer cell lectin-like receptor subfamily A (Klra) genes were found to be expressed in most of these cells and the Ncr1 markers was specifically expressed in these cells.

Finally using Tcells2 profiles on cells that was identified previously as T cells, returned again Treg as the main type of T cells and also showed 5 cells with a CD8 T cell profile, one CD4 T cells and one gamma-delta T cell (all with an identification score above 0.6) (Figure 12).

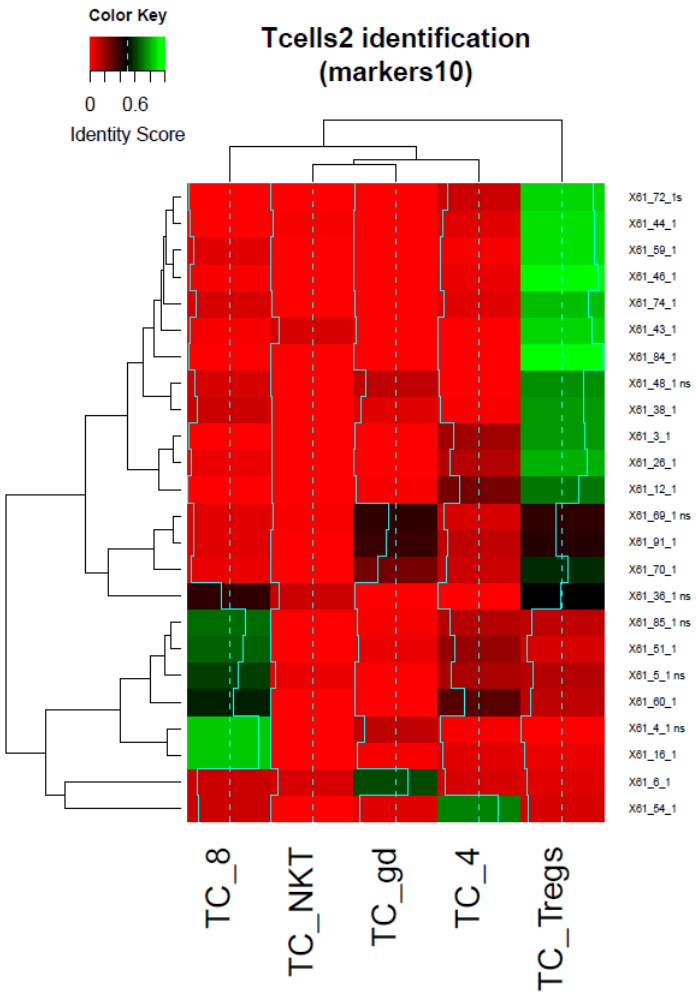

Figure 12. Further cell type classification using Tcells2 profiles on cells previously identified as T cells (in Figure 10). All identification value were significant $P < 0.05$ (otherwise indicated "ns").

To refine further the identification, specific genes expression for various T cell functions and identities are plotted for all identified T cells with Tcells2 (Figure 13). The y-axis represents the single cells labelled with the identification of Tcells2 when the identity score was above a threshold of 0.6 (otherwise they were labelled as “mixed”). The horizontal axis shows particular genes grouped according to their importance in cell type definition or function such as T cell activation for example. At first glance we can observe here the heterogeneity of single-cell gene expression that does not provide crystal clear pattern but rather rugged gene expression. We can nevertheless observe that cells identified as Treg tends to cluster together with a similar expression of Il2ra (CD25), a marker for Treg but also of activated T cell. Indeed these cells are also enriched in markers of T cells activation such as Icos, Ox40 or 4-1bb in addition to the Cd4 expression. The latter cells also display the expression of immune checkpoint genes such as Ctla4 or Pdcd1 and a transcription factor associated with the dysfunction of tumor-infiltrating T cells: Gata3.[18] Surprisingly, Foxp3, a critical driver of Treg development is only expressed in 4 cells but Gitr and Tnfrsf1b (Tnfr2) also specific for Treg were mostly found expressed in cells identified as Treg. Additionally, granzymes which are associated with effector functions were also expressed in some of the cells identified as Treg. However granzymes have been shown to be essential for Treg suppressive immune function in tumors.[19] Almost no cytokine (Il4, Il13, Il12b, Il6, Il21, Il17a, data not shown on the next figure) was found at the exception of Ifng in three cells of TC gamma-delta and mixed identities. Il10 was only found in the X61_72_1s sample, confirming its potent Treg immunosuppressive profile.

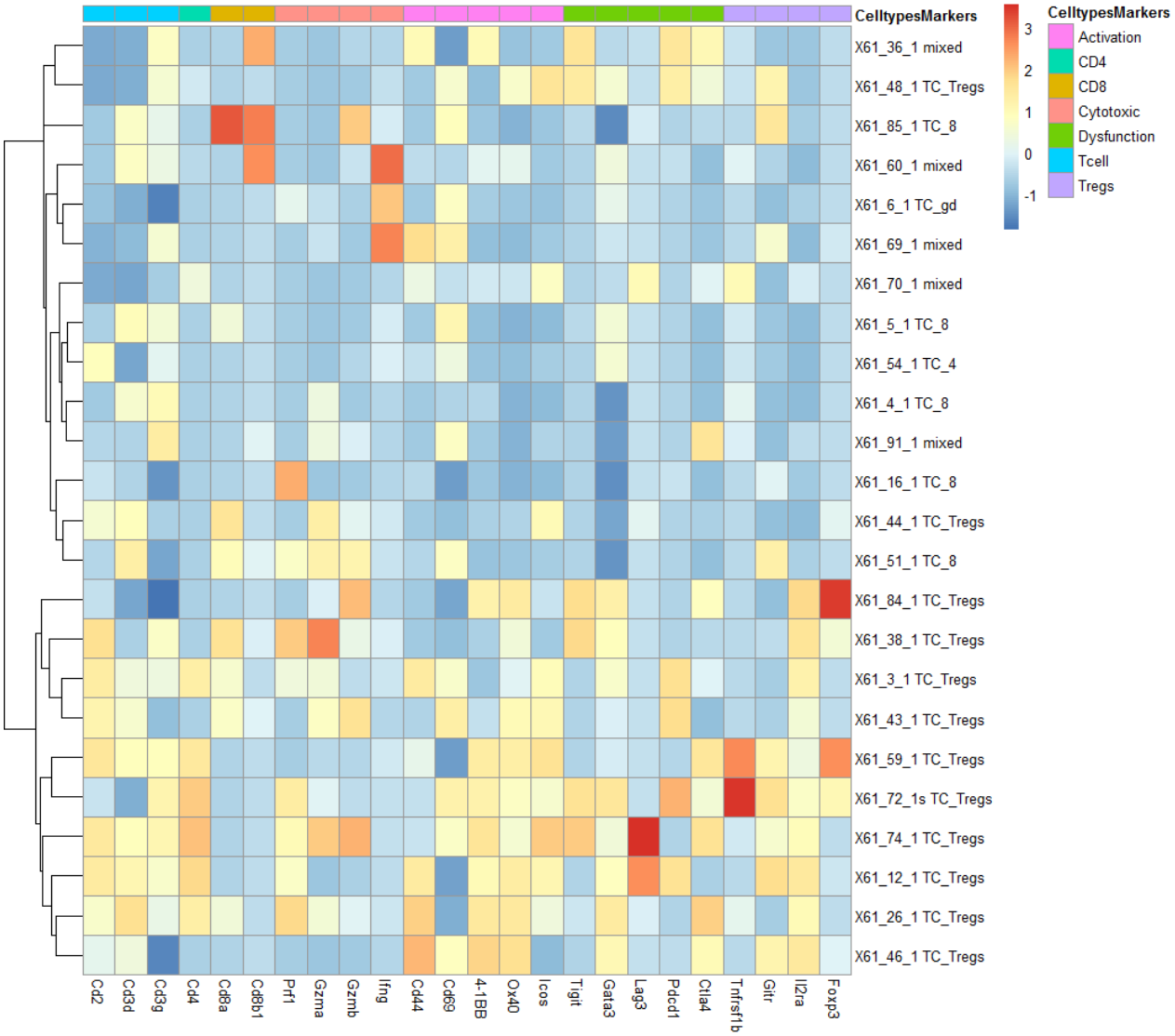

Figure 13. Heatmap of specific gene expression for various T cell functions and identities in cells identified as T cells by the identification algorithm. The y-axis represents the single cells labelled with the identification of Tcells2 when the identity score was above a threshold of 0.6 (otherwise they were labelled as "mixed"). The x-axis represents T cells emblematic genes. Gene expression is given in Z-score and cells are clustered on the basis of the Euclidian distance of the gene expressed.

DISCUSSION

As cellular identity can be defined by the molecular components of a cell, mainly mRNA and proteins, an approach measuring the whole content of mRNA in one cell as with scRNA-seq is indeed a major tool to gain insight of this identity. However the technical noise inherent with this technology and the apparent burst of transcription of single cell[20], forces us to reconsider the way we define a cell type with gene markers in scRNA-seq data. Here I used a method that incorporate all the cell types in one analysis and provide multiple sets of specific markers for various immune cell types to further quantify the cellular identity of scRNA-seq data. The results of the internal validation using the leave one out method has demonstrated the potential of this method in quantifying correctly most of the ImmGen samples of known identity (Figure 8). However the cross validation method has its own limit of inference because it uses only its own samples and moreover it uses microarray data which provide much more consistent expression data than scRNA-seq. Additionally the selection of samples based on clustering proximity has partially standardized the immune profiles and therefore it might have slightly over fitted the cross validation. I expect that the development of scRNA-seq will provide more profiles of various immune cells that has been previously cell-sorted with cell-surface markers for a better assessment of the algorithm performances (external validity).

The number of markers to use is also a crucial point of the analysis. I have evaluated empirically that a threshold of about 10 genes improved the number of correct identification with the cross validation (Figure 8). These results remain to be validated for single-cell, which probably contains a larger proportion of gene markers not expressed at all. Efroni *et al* also showed that a threshold of 20 was optimal for scRNA-seq data from plant cells and this threshold should be increased if the resolution of the data (ie the number of genes detected) decreased[14]. With our own scRNA-seq data I observed a slight decreased in the identification score when I increased the cumulative information threshold to 20 or 50 (data not shown). It appears that there is a tradeoff between choosing too few and too many markers. With too few markers there is an improvement of the identity score (particularly if a large number of gene is detected) but the risk of missing an identity by the absence of a gene marker or being bias by a sporadic gene expression of one gene is also more likely. Conversely, by including too many markers, some markers would be less specific and jeopardize the identification with false-positives.

Another compelling and perhaps not sufficiently investigated parameter is the equation quantifying the identity (ICI). As the equation rely on the level of expression of a particular gene, the results might be influenced by an extreme expression value. Even if the algorithm takes into account the absence of

other markers and the Spec score by weighting the final identification score, it is yet possible to have false positives because of sporadic gene expression with extreme values. To test this assumption I have empirically modified one gene marker expression and found that it would require an extreme value of approximately 15 times the max expression value observed in the whole data set to decrease the identity index from 1 to 0.5 approximately. Even if the impact is finally quite small we could imagine to use a binning strategy in order to prevent this phenomenon but here I have simply determining a stringent threshold of 0.6 from which identities are validated. Additionally, a mixed identity of two profiles with an identity score around 0.5 may still be informative and would require further analysis like comparing the results with unsupervised clustering techniques or evaluating specifically the expression of other gene markers.

This leads us toward the idea of a hierarchical identification process analysis, with multiple steps of identification to deepen and confirm the cell identity. As observed with the cross-validation in figure 6 and 9, the identification scores for T cells have increased when Specs scores were calculated with a more narrow set of cell types (i.e. restricted to T cells). This process has also been used for identifying particular subsets of T cells in the scRNA-seq data for example (Figure 10 and 13). Moreover we can observe that the identification scores revealed more distinct identities in Figure 12.

In an attempt to confirm the identities I have displayed the expression of particular genes among the identified cells (Figure 13). Interestingly the cells identified as Treg clustered together and displayed a relatively similar expression of Cd25 but most of them lack the expression of Foxp3. Hence, the accuracy of the Treg identification is challenged by this observation but other markers such as Ctla4, Gitr, TNfrsf1b, Ox40 well-known to be particularly expressed in Treg were also found in these cells[21]. Some of these markers are also part of the T cell activation module as Cd25 and therefore it is yet unclear whether these cells are truly Treg or activated T cells (theses two states are nevertheless not mutually exclusive). It has to be noticed that the Treg profile might also be related to the CD4 memory cells profile as their transcriptomes clustered closely in the unsupervised clustering analysis (Figure 3A). This example illustrates here the difficulty of determining precisely the cellular identity based on some genes selected from the literature because on one hand, the scientific information relative to the emerging field of single-cell analysis is somewhat incomplete and on the other hand, a great number of immune genes were simply not expressed at all. Over the 42 cells that passed the quality control, a median of 2618 unique genes were detected, which remains relatively low. In addition, the presence of unpredicted simultaneous expression of genes that are usually mutually exclusive has been observed such as Cd4 and Cd8a (in sample X61_3) or MHC class II genes (H2-DMb2 or H2-DMa) and Cd3/2/8a (X61_43). MHC class II expression in some T cells subsets such as Treg has been

demonstrated in human or rat but never in mice.[22,23] This observation would require more investigation in order to rule out a potential technical contamination of mRNA between wells during the Fluidigm C1 cell capture or unnoticed doublet cells for example.

In addition the mouse used in this study has been treated with the combination of checkpoint inhibitors (anti-CTLA-4 and anti-PD-1), which further complicates the interpretation. I was nevertheless able to demonstrate that the main immune cell types in our sample were T cells followed by NK cells and some neutrophils, dendritic and B cells. The small amount of myeloid cells found is explained by the cell size restriction of 5-10 µm used with the Fluidigm C1 integrated fluidic circuits that will likely exclude myeloid cells which are usually bigger than 10 µm. Therefore I would need more data to clarify the composition of the immune compartment in our mouse model of melanoma and especially scRNA-seq data from FACS sorted subpopulation of immune cells to validate definitely the algorithm. However the scripts are now available and ready to be embedded in an identification pipeline for large-scale analysis of mouse melanoma data. For human melanoma analysis I still require human immune cells repositories to compute Spec scores even though the immune system between human and mice is remarkably conserved. This represent the future goal of this project and human immune datasets are already under evaluation.[24,25]

Finally, we can now also relate of the striking heterogeneity observed for cells of presumed same identity and this observation appears to be the norm in single-cell studies. [21,2526] Hence, we can argue whether this heterogeneity comes from technical limitations or originate from biological variation. Obviously, the small amount of mRNA in a cell represents a major technical challenge and must be a substantial source of the noise but it also appears that biological variation between cells of apparent same identities play also its part. This evidence has been confirmed by other studies using bar coding which have shown that various cell fates emerge from apparently identical cells.[26,27] Hence, we can question whether this heterogeneity is in fact used by the organism to promote the diversity and the fine tuning of responses in a population of cells and if these variations are reflecting a more complex system. I believe that the development of single-cell technology will greatly help us in answering these questions.

MATERIALS AND METHODS

IMMGEN

Immune profiles were collected through the dataset retrieval section of the ImmGen website (www.immgen.org). Expression data of the selected immune profiles are available in a supplementary file. ImmGen immune profiles were obtained by averaging the expression of 3 biological replicates of 10,000 to 30,000 FACS-sorted immune cells from C57BL/6 mice. Data were generated using the Affymetrix 1.0 ST mouse microarrays. As the arrays contained duplicated probes, the max expression values was selected for defining only one expression for each gene.

MELANOMA TRANSGENIC MOUSE MODEL

The melanoma single-cell data originate from an engineered melanoma mouse model called iBIP2 (inducible human BRAFV600E, Ink4Arf floxed and PTEN floxed) that has been developed over a mouse with a genetic background of FVB at 70% and 30% of mixed background. These mice also have a CRE-ERT2 recombinase under the control of a melanocyte specific promoter (tyrosinase promoter). Thus they develop melanoma with similar genetic mutations that are observed in human and were also treated similarly to the clinical situation with the checkpoint inhibitors anti-CTLA-4 and anti-PD-1.

SINGLE-CELL RNA-SEQ DATA

Single-cell data were produced by the MelanomX team from the iBIP2 mice. Briefly, the tumors were subject to disaggregation, enzymatic digestion and red blood cells lysis followed by FACS-sorting of the CD45+ live cells. Then the cells were captured with the Fluidigm C1 technology using small plates (wells of 5-10 µM), followed by cDNA preparation and barcoding using the SMARTer protocol (Clontech). The following cDNA was sequenced by using the Illumina HiSeq instrument at the Lausanne Genomic Technologies Facility (LGTF). For this project I used the expression data of the X61 small plate with CD45+ cells. In total 2’732’700 reads could be mapped over the mouse genome (91% of total reads) with 1'177'365 reads mapping protein coding genes. Quality control filtered single cells with a minimal number of 10’000 reads, at least 500 genes detected and with a minimal proportion of 50% of reads that mapped protein coding region. In total 42 cells passed the quality control and I used TPM (tags per million) as expression values for cell type identification.

DATA AND SOURCE CODE

All scripts were performed using R version 3.3.1. All the R and bash code are available in a separate file. The procedure started with selecting ImmGen samples (expData.txt and cellTypesV3.txt) with various unsupervised clustering techniques (t-SNE, PCA, MDS). Then the selected transcriptomes (106 samples with each 21'755 genes) were splitted in matrix of 1000 genes (in specV3-7Analysis.R) for parallel calculation of Spec scores (according to Birnbaum and Efroni codes)[13,14] over the Vital-it cluster (http://www.vital-it.ch). The 1000 rows submatrix were loaded on the cluster and a bash script (clusterScriptV3.sh) was launched with the bsub command. The output files were merged to provide a matrix of Spec scores for every genes (with specV3-7Analysis.R or scX61-4.R for sc-RNA-seq analysis). The parallelization method with submatrix returned the exact same Spec results whether the matrix was taken fully or partially. Then, the names of Specs genes were then converted to ensembl identifier using the Bioconductor package Biomart and matched with the ones obtained in the scRNA-seq experiment (unmatched genes were simply removed for the scRNA-seq experiment). Significance was determined by randomly selecting an identical number of markers and performing an identification 1000 times to established a background distribution. The threshold for significance was defined as an identification value superior than the top 5% of the highest background identification and the p-value was adjusted with the false discovery rate (FDR). Cross-validation with the leave-one out method and Spec+ICI identification of the left out sample was also calculated on the cluster but with parallel calculation over the 103 samples removed one at a time (with the V3specClusterLOO.R and clusterScriptLOO.sh scripts, analysis of results: LOOCV_analysis.R).

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144:646-74.
2. Hanahan D, Coussens LM. Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. Cancer Cell 2012;21:309-22.
3. Topalian SL, Sznol M, McDermott DF, et al. Survival, Durable Tumor Remission, and Long-Term Safety in Patients With Advanced Melanoma Receiving Nivolumab. Journal of Clinical Oncology 2014;32:1020-+.
4. Hamid O, Robert C, Daud A, et al. Safety and Tumor Responses with Lambrolizumab (Anti-PD-1) in Melanoma. New England Journal of Medicine 2013;369:134-44.
5. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. Nature 2011;480:480-9.
6. Schreiber RD, Old LJ, Smyth MJ. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. Science 2011;331:1565-70.
7. Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods 2014;11:163-+.
8. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics 2013;14:618-30.
9. Jang JS, Simon VA, Feddersen RM, et al. Quantitative miRNA Expression Analysis Using Fluidigm Microfluidics Dynamic Arrays. Bmc Genomics 2011;12.
10. Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science 2014;343:776-9.
11. Islam S, Kjällquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research 2011.
12. Shay T, Kang J. Immunological Genome Project and systems immunology. Trends in Immunology 2013;34:602-9.
13. Birnbaum KD, Kussell E. Measuring cell identity in noisy biological systems. Nucleic Acids Res 2011;39:9093-107.
14. Efroni I, Ip PL, Nawy T, Mello A, Birnbaum KD. Quantification of cell identity from single-cell gene expression profiles. Genome Biology 2015;16.
15. Murray Peter J, Allen Judith E, Biswas Subhra K, et al. Macrophage Activation and Polarization: Nomenclature and Experimental Guidelines. Immunity 2014;41:14-20.
16. Wakabayashi Y, Watanabe H, Inoue J, et al. Bcl11b is required for differentiation and survival of alphabeta T lymphocytes. Nat Immunol 2003;4:533-9.
17. Sasaki I, Hoshino K, Sugiyama T, et al. Spi-B is critical for plasmacytoid dendritic cell function and development. Blood 2012;120:4733.
18. Singer M, Wang C, Cong L, et al. A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. Cell;166:1500-11.e9.
19. Cao X, Cai SF, Fehniger TA, et al. Granzyme B and perforin are important for regulatory T cell-mediated suppression of tumor clearance. Immunity 2007;27:635-46.
20. Satija R, Shalek AK. Heterogeneity in immune responses: from populations to single cells. Trends in Immunology 2014;35:219-29.
21. Pfoertner S, Jeron A, Probst-Kepper M, et al. Signatures of human regulatory T cells: an encounter with old friends and new players. Genome Biol 2006;7:R54.
22. Holling TM, Schooten E, van Den Elsen PJ. Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. Human Immunology 2004;65:282-90.
23. Baecher-Allan C, Wolf E, Hafler DA. MHC class II expression identifies functionally distinct human regulatory T cells. Journal of immunology (Baltimore, Md : 1950) 2006;176:4622-31.
24. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy Number Loss of the Interferon Gene Cluster in Melanomas Is Linked to Reduced T Cell Infiltrate and Poor Patient Prognosis. PLOS ONE 2014;9:e109760.
25. Hoek KL, Samir P, Howard LM, et al. A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. PLOS ONE 2015;10:e0118528.
26. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014;344:1396-401.