Immunogenetics (2017) 69:439–450
DOI 10.1007/s00251-017-1002-x

ORIGINAL ARTICLE

Identification of innate lymphoid cells in single-cell RNA-Seq data

Madeleine Suffiotti[1,2] · **Santiago J. Carmona**[1,2] · **Camilla Jandus**[1] · **David Gfeller**[1,2]

Received: 28 February 2017 / Accepted: 11 May 2017 / Published online: 22 May 2017

Abstract Innate lymphoid cells (ILCs) consist of natural killer (NK) cells and non-cytotoxic ILCs that are broadly classified into ILC1, ILC2, and ILC3 subtypes. These cells recently emerged as important early effectors of innate immunity for their roles in tissue homeostasis and inflammation. Over the last few years, ILCs have been extensively studied in mouse and human at the functional and molecular level, including gene expression profiling. However, sorting ILCs with flow cytometry for gene expression analysis is a delicate and time-consuming process. Here we propose and validate a novel framework for studying ILCs at the transcriptomic level using single-cell RNA-Seq data. Our approach combines unsupervised clustering and a new cell type classifier trained on mouse ILC gene expression data. We show that this approach can accurately identify different ILCs, especially ILC2 cells, in human lymphocyte single-cell RNA-Seq data. Our new model relies only on genes conserved across vertebrates, thereby making it in principle applicable in any vertebrate species. Considering the rapid increase in throughput of single-cell RNA-Seq technology, our work provides a computational framework for studying ILC2 cells in single-cell transcriptomic data and may help exploring their conservation in distant vertebrate species.

Keywords Innate lymphoid cells · Single-cell RNA-Seq · Immune cell type evolution · Immunogenomics · Cell type predictions

Introduction

Innate lymphoid cells (ILCs) form a family of lymphocytes that are involved in innate immunity and tissue remodeling (Eberl et al. 2015; Sonnenberg and Artis 2015). NK cells represent the most frequent type of ILCs and have been intensively studied in immunology. Non-cytotoxic ILCs, often simply referred to as ILCs, have been discovered much more recently and are broadly divided into three subtypes—ILC1, ILC2, and ILC3—which mirror the Th1, Th2, and Th17 subtypes in the CD4 T cell lineage both at the transcriptional and functional levels, albeit lacking somatic T cell receptor (TCR) rearrangement (Diefenbach et al. 2014). At the transcriptomic level, ILCs have been analyzed in mouse (Robinette et al. 2015; Seehus et al. 2015; Chea et al. 2016; Shih et al. 2016) and human (Björklund et al. 2016; Kløverpris et al. 2016; Koues et al. 2016).

ILCs are distinguished from other lymphocytes by the absence of specific lineage markers (Artis and Spits 2015) and isolating them requires mainly negative sorting with antibodies specific for markers of other immune cell types. From a technical point of view, this results in complex sorting protocols that are not easy to use, are prone to contamination of sorted cells by unwanted cell types, and may alter the transcriptome of sorted ILCs. From an evolutionary point of view, absence of specific markers makes it difficult to trace the conservation of these cell types. Indeed, conservation analysis of immune cell types across vertebrate species has mainly relied on gene orthology. For instance,

Electronic supplementary material The online version of this article (doi:10.1007/s00251-017-1002-x) contains supplementary material, which is available to authorized users.

✉ David Gfeller
david.gfeller@unil.ch

[1] Ludwig Centre for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland

[2] Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland

genes like CD8, CD19, RAG, or TCR have been found in all jawed vertebrate species, supporting the fact that T and B cells are conserved and show similar TCR/B cell receptor (BCR) recombination mechanisms (Boehm 2012). Other immune cell types, like NK cells, are more difficult to study based on gene orthology since many of their markers in mammalian species are not conserved. In the case of ILCs, indirect evidences based on conservation of ID2—a DNA-binding protein involved in ILC development and blocking transcription factors involved in B and T cell development (Delconte et al. 2016), or PLFZ—a transcription factor expressed in early development of ILCs (Constantinides et al. 2014; Ishizuka et al. 2016), suggest that they could be present in all vertebrates (Vivier et al. 2016). However, many genes involved in ILC development or showing differential expression in ILCs are also expressed by other cell types and it is unclear whether their conservation can be used as a proxy for ILC subtype conservation.

Single-cell RNA-Seq is a powerful approach for studying immune cell populations without relying on specific markers and antibodies (Proserpio and Lönnberg 2016; Neu et al. 2017). Recently, we and others have shown that unbiased single-cell transcriptome profiling enables studying conservation of immune cells, and these studies have led to the discovery of a novel population of NK-like cells in zebrafish (Moore et al. 2016; Carmona et al. 2017).

Here, we develop and validate a novel approach to identify and study ILCs at the gene expression level using single-cell transcriptomics data. Our proposed framework combines single-cell RNA-Seq analysis of bulk lymphocytes together with clustering and a novel multi-class gene expression-based cell type classifier. More specifically, we first train a logistic regression-based classifier on mouse ILC and NK gene expression data (Robinette et al. 2015) and validate it in single-cell RNA-Seq data of human NK and ILCs (Björklund et al. 2016). Importantly, our classifier only relies on genes that are broadly conserved across vertebrate species. We then show that ILCs, and especially ILC2, can be identified from a pool of bulk lymphocytes single-cell transcriptomics data by first separating T, B, and NK/ILCs with unsupervised clustering and then applying our new classifier on cells in the NK/ILC cluster. As lymphocytes can be isolated from peripheral blood with density gradient centrifugation followed by flow cytometry based on light scatter measurements, our approach should be applicable in most vertebrate species with a sequenced genome and availability of peripheral blood samples. Overall, our work provides a computational framework for the use of single-cell RNA-Seq technology to identify and study ILC2 cells at the transcriptomic level without relying on species-specific antibodies and complex cell sorting procedures.

Materials and methods

Datasets' description

Mouse ILC expression data (Robinette et al. 2015) Microarray raw data (Affymetrix Mouse Gene 1.0 ST array) were downloaded from GEO (accession code GSE37448). We employed three ILC1 samples ($CD127^+$ ILC1 samples extracted from spleen), two ILC2 samples (from small intestine), three ILC3 samples ($NKp46^+Rorgt^{hi}$), and three NK samples ($CD127^-$ cell samples extracted from spleen). We normalized and log transformed the data using the robust multi-array average (RMA) method from the *Affy* package.

Human ILC single-cell gene expression data (Björklund et al. 2016) Single-cell RNA-sequencing data of human tonsils (Illumina HiSeq 2000) were downloaded from GEO (accession code GSE70580), including 127 ILC1, 139 ILC2, 308 ILC3, and 74 NK cells from three different donors. Raw data were normalized and log2 transformed using voom method in *limma* package (Ritchie et al. 2015). Batch effect from donors was corrected using ComBat method in *sva* package (Parker et al. 2014).

Human T and B cell single-cell RNA-Seq data (Tirosh et al. 2016) Processed single-cell RNA-sequencing data from melanoma tumors were downloaded from GEO (accession code GSE72056). Among lymphocytes, this dataset contained 51 NK cells, 512 B cells and 2040 T cells (based on the annotation provided by the authors for the different lymphocytes), as well as 50 not-annotated (NA) CD45+ cells. Gene expression levels were available as TPM (transcript per million).

Model construction, validation, and gene signature definition

To identify ILCs among a mixed population of NK cells and ILCs single-cell RNA-Seq data, we trained a multi-class gene expression-based classification model with mouse gene expression data from (Robinette et al. 2015) (model 1 in Fig. 1). Specifically, we first performed differential gene expression analysis with *limma* (Ritchie et al. 2015) to select the individually most discriminative genes for each cell type (NK, ILC1, ILC2, and ILC3) in mouse. Differential expression analysis was done by pairwise comparisons of the samples from each cell type against the samples from all other cell types, and the 11 most upregulated genes based on log-fold change values were selected for each cell type. As ILC1 and NK cells show high transcriptional similarity, we further included the two most down regulated genes between NK and ILC1 samples (i.e., Inpp4 and As3mt) in the gene signature for NK cells. Genes such as the killer cell lectin like receptors and pseudogenes (e.g., Gm4761) were then removed from the list,

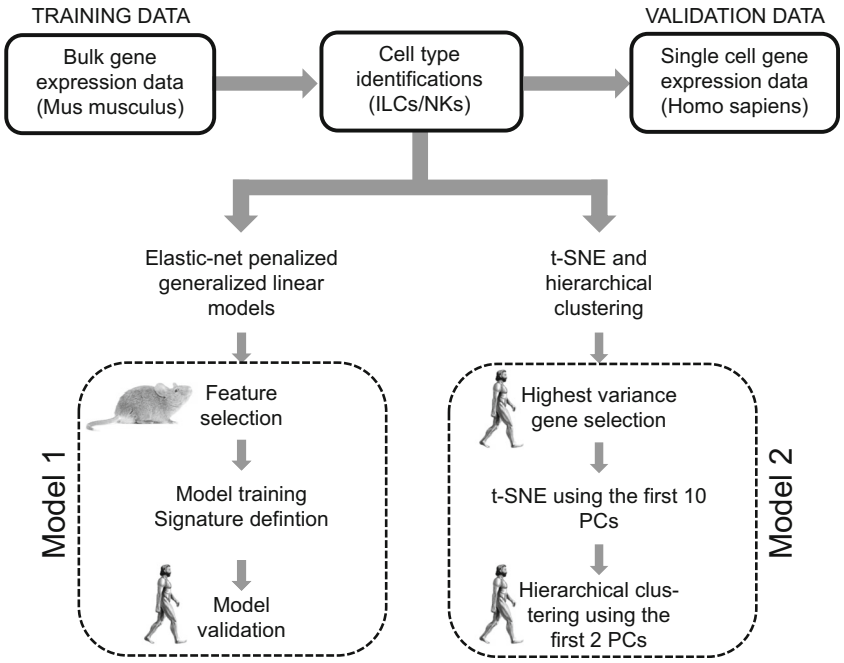

Fig. 1 NK/ILC cell type classifiers. Starting from mouse gene expression data of sorted NK, ILC1, ILC2, and ILC3 cells (Robinette et al. 2015), we trained a multi-class gene expression-based classification model (*model 1*) and validated it on human single-cell RNA-Seq data (Björklund et al. 2016). As an alternative, we also tested direct unsupervised clustering of human single-cell RNA-Seq data of NK and ILCs (*model 2*). *PC* stands for principal component

as they are not conserved across vertebrates. We also added some genes known to be specific for ILC subsets in mouse (i.e., Tnf for ILC1, Il9r, and Ccl3 for ILC2 as well as Gzma, Gzmb, and Gzmk for NK cells) to obtain a preliminary list of 28 genes. We then applied a multinomial logistic regression model (*glmnet*, default parameters for lasso optimization) (Friedman et al. 2010) on the mouse gene expression data. This model operates as a feature selection, and 16 genes were removed from the signature. Finally, we verified the presence of orthologs for each of the remaining genes. Genes without orthologs were replaced by conserved and highly correlated genes that may therefore show similar discriminative power for gene expression-based cell type predictions (i.e., Ccl3 with Il12rb2, Il13 with Alox5, Il9r with Pparg and Rxrg, Igfbp7 with Rorc, and Tnf with Tnfrsf21). The final list of genes is shown in Table 1, first column. The logistic regression model was retrained with this conserved gene signature (α = 0.8). The output of the model consists of posterior probabilities of belonging to each cell type. These probabilities are computed as

$$P(k|x) = \frac{e^{\left(\beta_{0k}+\beta_k^T x\right)}}{\sum_{k'=1}^{4} e^{\left(\beta_{0k'}+\beta_{k'}^T x\right)}},$$

where $k = 1\ldots4$ stands for the cell types ILC1/2/3 and NK, x stands for the expression of the signature genes (first column in Table 1) and β_k stands for the coefficients of the model for each cell type (second column in Table 1 shows non-zero coefficients for each cell type). β_{0k} corresponds to intercept values. Supplemental Fig. 2 displays an example of implementation of the code in R to carry out these predictions. Being aware that the single-cell samples may not be robustly classifiable due to their high intra-group heterogeneity, we used a conservative threshold for posterior probability of 0.7 in the contingency tables (Tables 2 and 3).

To benchmark our performances, we applied our model to human NK and ILCs single-cell RNA-Seq data (Björklund et al. 2016). We used receiver operator characteristic (ROC) analysis and calculated the area under the curve (AUC) as measure of performance of our model. As the output of our multi-class model is a score (posterior probability) for each class (cell type), we decomposed the problem into four pairwise comparisons (one class vs. the rest). We used the *ROCR* package to plot ROC curves and calculate AUC values for each class.

Next, an unsupervised clustering approach was implemented using t-SNE (Van der Maaten and Hinton 2008) for dimensionality reduction of gene expression space, followed by hierarchical clustering, as in Björklund et al. (2016) (model 2 in Fig. 1). Briefly, low dimensional cell representations produced by t-SNE were used to build a dendrogram (R pvclust package, using default parameters with average linkage). Then, the dendrogram was cut in order to obtain four clusters (one for each cell type). To assign each cluster to a cell type, we considered all possible associations (24 possibilities) and computed the AUC for each of them and kept the best performing one. This assignment resulting in the highest AUC was used in Fig. 2.

Subsequently, to evaluate and compare the performances of the two approaches in more plausible biological conditions, we reduced the frequencies of ILCs compared to NK samples (over-represented NK cells). We randomly selected five ILC1, five ILC2, and five ILC3 cells in 10 replicates, combined them

Table 1 Orthologs of signature genes

Gene Symbol	Coefficients	*Xenopus tropicalis* (ENSXETG0000)	*Anolis carolinensis* (ENSACAG0000)	*Gallus gallus* (ENSGALG0000)	*Danio rerio* (ENSDART0000)	*Elephant shark* (NCBI Gene ID)	*Lamprey* (ENSPMAG0000[a])
ILC1—*intercept*	−3.703						
Cd3e	1.291	0010119	0007255	0007416	Detected in other bony fish	EOG090B0C9R[b]	Not conserved
Cd28	0.413	0033572	0004486	0008669	0147200	103176828	Not conserved
ILC2—*intercept*	3.374						
Il12rb2	−0.300	0012397	XP_016854485.1[a]	0011222	0052157	103189128	0006307
Alox5	0.557	0007490	0013726	0005857	0043085	103185827	0003663
Pparg	0.479	0017422	0013360	0004974	0031848	103185037	0002615
Rxrg	0.387	0004750	0007923	0003406	0004697	103184778	0007822
ILC3—*intercept*	2.257						
Il23r	0.016	XP_012816978.1[a]	0027663	0011212	0052158	103189608	Not conserved
Chad	0.327	0006832	0005236	0019761	0045071	103175646	0007797
Tnfrsf21	0.311	0011317	0013394	0016719	0001807	103182320	0007989
Rorc	0.357	0002131	0004098	0025988	0087195	Not conserved	Not conserved
NK—*intercept*	−1.928						
Tnfrsf21	−0.304	0011317	0013394	0016719	0001807	103182320	0007989
As3mt	0.943	0031996	0029427	0008133	0027572	103171996	0008466
Inpp4b	1.449	0021313	0002735	0009880	0075201	103181285	Not conserved
Gzmk	−0.164	0031829	0015449	0013546	0090380	Putative ortholog (Akula et al. 2015)	Not conserved

Column 1: signature genes used in our NK/ILC classifier. Column 2: coefficients of the regression model of Eq. 1. Column 3–8: conservation across vertebrate species. Cd3e is not found in zebrafish, but has been reported in many other bony fish species (Alabyev et al. 2000; Koppang et al. 2010)

[a] Cases where the protein was only annotated in NCBI

[b] Cases where the proteins was only annotated in OrthoDB

Table 2 Contingency table for the identification of NK and ILCs in human single-cell RNA-Seq data using the full dataset from (Björklund et al. 2016)

		Observed groups			
		ILC1	ILC2	ILC3	NK
Predicted groups	ILC1	97 (47%)[a]	19	76	14
	ILC2	4	51 (91%)[a]	1	0
	ILC3	3	4	161 (93%)[a]	5
	NK	11	0	0	44 (80%)[a]
	NP	12	65	70	11

Percentages stand for the precision values (i.e., fraction of true positives among the predictions). NP stands for cells that are not predicted.

[a] $P < 0.05$

with all NK cells (74 cells) and applied both the classification model and the unsupervised clustering approach to these data. Results in Fig. 2b, d and Table 3 correspond to the average over the 10 replicates.

Ortholog identification

In order to define a model that can be applicable across vertebrates, we specifically selected genes that are conserved across vertebrates. We verified the orthology in one species of amphibians (*Xenopus tropicalis*), reptiles (*Anolis carolinensis*), birds (*Gallus gallus*), and bony fish (*Danio rerio*) and cartilaginous fish (*Callorhinchus milii*, i.e., elephant shark). Orthologous gene identification was performed by querying of the BioMart database from Ensemble (http://www.biomart.org) and OMAbrowser (http://www.omabrowser.org) (Altenhoff et al. 2015). In cases where no orthologs were found, we performed gene name-based searches in NCBI, UniProtKB/TrEMBL, and UniProtKB/SwissProt databases, protein sequence-based searches using Psi-BLAST, and finally, literature searches. We reported Ensemble IDs, and if not available, NCBI Reference Sequence and/or literature references in Table 1.

Table 3 Contingency table for the identification of NK and ILCs in human single-cell RNA-Seq data using only five cells of each ILC subtype (average over ten random choices of ILCs)

		Observed groups			
		ILC1	ILC2	ILC3	NK
Predicted groups	ILC1	3.7 (19%)	1.1	1.1	14
	ILC2	0.1	1.6 (89%)[a]	0.1	0
	ILC3	0.2	0.1	1.9 (26%)	5
	NK	0.4	0	0	44 (99%)[a]
	NP	0.6	2.2	1.9	11

Percentages stand for the precision values (i.e., fraction of true positives among the predictions). NP stands for cells that are not predicted.

[a] $P < 0.05$

Analysis of lymphocyte single-cell RNA-Seq data

To show that ILCs can be identified in single-cell transcriptomics data of mixed lymphocyte populations, we retrieved human T and B cells single-cell RNA-Seq data (see above) and combined them with human ILCs and NK single-cell RNA-Seq data. We selected 250 T cells and 60 B cells to have realistic proportions of lymphocytes from peripheral blood. We further included 50 CD45+ cells without annotation to mimic more realistic situation with cells that cannot be easily annotated from their transcriptional profile, and may include some ILCs. As these single-cell RNA-Seq data consists of transcript per million (TPM), single-cell RNA-Seq NK/ILC data were transformed in TPM by dividing raw counts by the gene length and normalizing. Batch effects for the two datasets were corrected using ComBat method (Parker et al. 2014). We then applied t-SNE on the combined dataset. We selected the 3000 most variable genes (same parameters as before) and extracted the three clusters shown in Fig. 3a from the hierarchical clustering.

Results

Mouse gene signature for NK/ILC identification

NK cells and ILCs originate from the same precursors and show higher transcriptional similarity compared with other immune cell types (Robinette et al. 2015). To probe whether NK cells and ILCs can be distinguished based on transcriptional signatures, we set out to train a classifier (model 1 in Fig. 1). We first performed differential gene expression analysis of mouse bulk gene expression data from sorted ILCs and NK cells (Robinette et al. 2015) and selected the top 11 upregulated genes for each cell type (see the "Materials and methods" section). As our purpose was to define a gene signature that allows us to distinguish NK/ILCs and can be applied in different species, we filtered out several genes that were clearly not conserved across vertebrates (e.g., KIR genes, pseudogenes) and defined a preliminary subset of potentially discriminating genes (28 genes, see the "Materials and methods" section). We then used this preliminary signature from mouse NK/ILC data to train a multinomial logistic regression model to predict the different cell types (ILC1/2/3 and NK, see the "Materials and methods" section). Using standard feature selection, we were able to further restrict our final gene signature to 13 conserved genes (Table 1). Details about the regression model are discussed in the "Materials and methods" section, and the exact values of the different coefficients are shown in Table 1. The signature genes selected by our model to distinguish NK cells and ILCs at the transcriptomic level consist of two genes (Cd3e and Cd28) for ILC1, four genes (Il12rb2, Alox5, Pparg, and

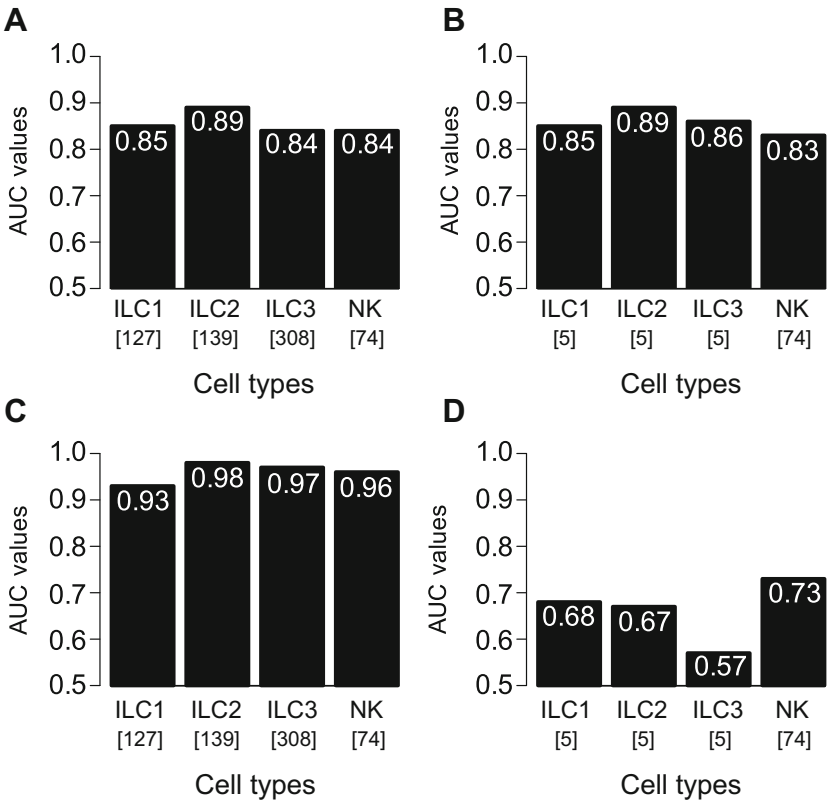

Fig. 2 AUC values for predictions of NK cells and ILCs in human single-cell RNA-Seq data. **a**, **b** Results of our multi-class classification model (model 1) on all human NK and ILCs from (Björklund et al. 2016) (**a**), as well as a more realistic mixture of human NK cells and ILCs (**b**). **c**, **d** Results of the unsupervised clustering (model 2) on the same datasets used in **a** and **b**. *Numbers in brackets* indicate the total number of cells from each cell type used in this validation

Rxrg) for ILC2, four genes (Il23r, Chad, Tnfrsf21, and Rorc) for ILC3, and four genes (Tnfrsf21, As3mt, Inpp4b, and Gzmk) for NK cells. In our model, Tnfrsf21 was selected both as a positive marker for ILC3 and negative marker for NK cells (coefficients with opposite sign in Table 1). Table 1 shows orthologs of signature genes in *X. tropicalis* (amphibians), *A. carolinensis* (reptiles), *G. gallus* (birds), *D. rerio* (bony fish), and *C. milii* (cartilaginous fish), which were all included in the selection of signature genes. We also show orthologs in *Petromyzon marinus* (jawless vertebrate). As expected from our model construction, these genes were conserved in all jawed vertebrate species, including cartilaginous fish, but many of these genes were not conserved in lamprey, indicating that the approach proposed in this work is not applicable in jawless vertebrate species.

Although the protein complex Cd3 is specific for T cells, the Cd3 epsilon chain (Cd3e) was shown to be specifically upregulated in ILC1 in mouse at the mRNA level (Robinette et al. 2015). This highlights the fact that genes optimally selected for transcriptional signatures of immune cells may not necessarily correspond to cell surface protein markers with a known functional role. Cd3e gene orthologs were found in all the species we selected, except for zebrafish. However, Cd3e was found in other ray-finned fish (Koppang et al. 2010) and cartilaginous fish genomes (Alabyev et al. 2000), suggesting that it is generally conserved in fish species. The majority of known ILC2-specific marker genes consist of a group of cytokines or cytokine receptors (Il13, Il5, Il4, Il9r, and Il12rb) (Robinette et al. 2015), most of which were excluded from our signature because of the rapid evolution these molecules undergo (Brocker et al. 2010) and consequently the lack of orthologs outside mammals (Kaiser et al. 2004). However, Il12rb2 was kept as we could find orthologous gene across vertebrates. Interleukin-12 receptor beta 2 subunit is a cell surface receptor, whose expression is regulated by different cytokines in ILC2 (Lim et al. 2016) and was recently shown to be clinically relevant in human diseases (Silver et al. 2016). Among ILC2-specific markers, Alox5, Pparg, and Rxrg genes were also selected by our approach and confirm the emerging role of bioactive lipid mediators in regulating ILC2 biology (Konya and Mjösberg 2016; Moltke von et al. 2017). ILC3-specific genes include well-known ILC3 markers as Il23r and Rorc, the last one being constitutively expressed in this subtype (Sonnenberg and Artis 2015). Tumor necrosis factor receptor superfamily such as Tnfrsf21, also known as death receptor 6, plays an important role in ILCs for the development and homeostasis of lymphoid organs, a peculiar function of the ILC3 subtype (Šedý et al. 2015), anti-tumor responses, and promote inflammation (Sonnenberg and Artis 2015). Chondroadherin coding gene (Chad) was already identified as a signature gene for ILC3 subtype in mouse (Vivier et al. 2016).

Differential gene expression analysis revealed that the most discriminant NK genes consists of several classical

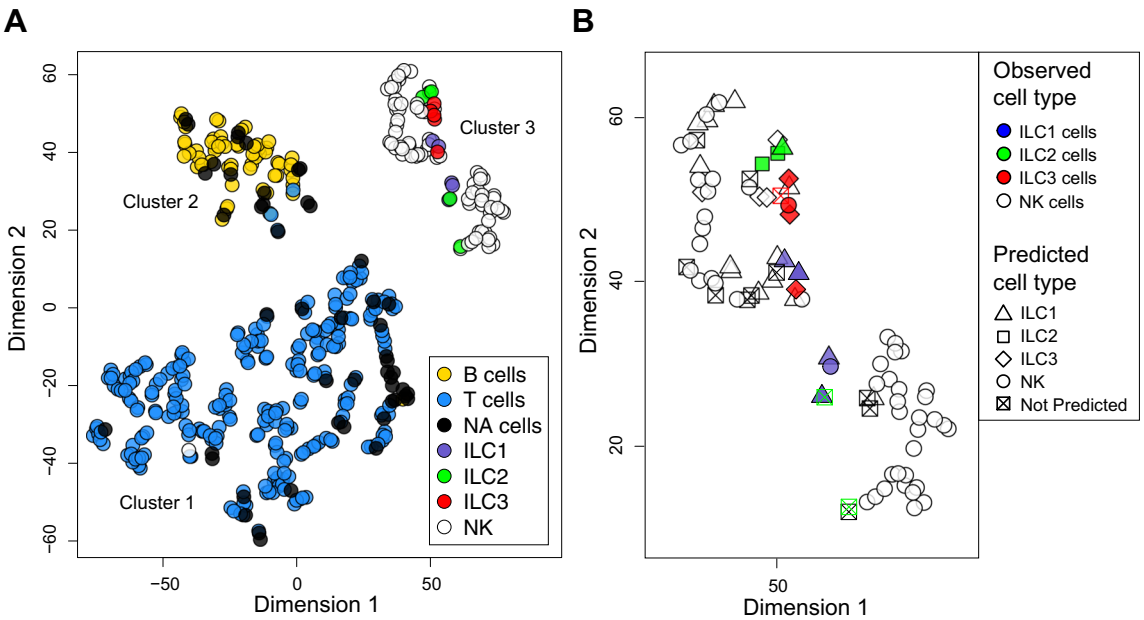

Fig. 3 Identifications of ILCs in single-cell RNA-Seq data from a pool of lymphocytes. **a** Two-dimensional representation (t-SNE) of a pool of T cells (n = 250), B cells (n = 60), and non-annotated cells (n = 50) from (Tirosh et al. 2016), as well as NK cells (n = 74) and ILCs (five of each type) from (Björklund et al. 2016). Clusters identified by hierarchical clustering show that T (*cluster 1*), B (*cluster 2*), and NK/ILCs (*cluster 3*) form distinct clusters. Reversely, ILC and NK are difficult to distinguish with clustering when NK cells largely outnumber ILCs (see also Fig. 2d). **b** Zoom on cluster 3, including examples of predictions of our classification model (model 1). *Squares with crosses* indicate cells that could not be assigned to any cell type by our predictor

mammalian NK receptors such as killer-cell inhibitory receptor (KIR), which were excluded from the signature because of their poor conservation beyond mammals. Among the marker genes for NK cells, our model selected the genes encoding inositol polyphosphate 4-phosphatase (Inpp4b) and granzyme k (Gzmk) (Akula et al. 2015). Inn4b belongs to a class of inositol phospholipid signaling enzymes shown to hydrolyze the SHIP products PI(3,4)P2 and PI(3)P. Inn4b gene is evolutionary closely linked to the IL-15 gene (Wang et al. 2007), a key regulator of NK cell activity (Ranson et al. 2003). Although very little is known on the serine protease Granzyme K, it has been reported to mediate caspase-independent killing by distinct subsets of NK cells (Jiang et al. 2011). Importantly, the signature genes in Table 1 were selected for optimal predictions of NK and ILCs based on gene expression data and may not all have a functional role in ILCs at the protein level.

Identification of NK/ILCs in human single-cell RNA-Seq data

To validate our NK/ILC classifier, we took advantage of a recent single-cell RNA-Seq dataset of fluorescence activated cell sorted human NK cells and ILCs (Björklund et al. 2016). This dataset consists of 74 NK, 127 ILC1, 139 ILC2, and 308 ILC3 cells. Predictions based on our model trained in mouse were carried out for each single cell based on expression of the signature genes, and predictive performance of the classifier was assessed with ROC curves (see the "Materials and methods" section). Overall, we obtained good AUC values for all classes, and especially for the ILC2 group ($AUC_{ILC2} = 0.89$) (Fig. 2a). A contingency table summarizing cell type classification is shown in Table 2. We observe an overall concordance of 72% between the predicted classes and the observed classes when using the whole single-cell dataset. ILC2 and ILC3 show the lowest number of false positives and the highest precision (percentages in Table 2), suggesting that one could study these cell types at the transcriptomic level using cell type predictions performed on mixtures of NK and ILCs single-cell RNA-Seq data. On the other hand, NK and ILC1 show lower precision suggesting that these subsets are more difficult to distinguish. Interestingly, ILC1 and NK cells show the strongest similarity both at the transcriptional (Robinette et al. 2015) and functional levels (Spits et al. 2016).

Our aim is to identify ILCs among realistic mixtures of NK/ILCs from peripheral blood. In such a case, the number of ILCs compared to NK cells is considerably lower although high variability can be observed between species, tissues, and conditions (e.g., infections). Therefore, we simulated a more plausible scenario by randomly selecting five cells from each ILC cell type and mixing them with the whole set of NK cells. This was repeated 10 times and performance was assessed (Fig. 2b). Importantly, as our classifier is applied separately to each cell, cell type prediction for

each individual cell is not influenced by the relative proportion of other cell types. Table 3 represents the contingency table summarizing predictions for these unbalanced mixtures of NK and ILCs (average over 10 replicates). The best predictions are again obtained with ILC2. Interestingly, for ILC2, the same sensitivity (35%, fraction of ILC2 cells correctly predicted, second column in Tables 2 and 3) and the same precision (89 and 91%, fractions of true positives among the predicted ILC2 cells) values are obtained as before. These results suggest that even in the presence of a much lower frequency of ILCs compared to NK cells, ILC2 cells can be accurately predicted and studied at the gene expression level (see below for a discussion about the minimal number of cells needed). Reversely, the precision was much lower for ILC1 and ILC3 cell types, suggesting that our predictions should be taken with caution for these two cell types and may not be reliable enough to study ICL1 and ILC3 cells at the gene expression level.

The global higher performance in detecting ILC2 subtype could be explained by the fact that this is the most homogeneous group (Robinette et al. 2015). It has also been shown that ILC1 and ILC3 present at least three distinct subtypes (Robinette et al. 2015; Björklund et al. 2016) and that these groups show bidirectional plasticity in inflamed intestinal tissues (Robinette et al. 2015; Bernink et al. 2015). Moreover, ILC1 cells show highest transcriptional similarity with NK cells (Robinette et al. 2015), and may therefore be difficult to identify in a mixed population of NK/ILCs based only on their transcriptome.

Another widely used approach for cell type identification in single-cell RNA-Seq data of mixed populations is unsupervised clustering of cells based on transcriptional profile similarity (Björklund et al. 2016) (model 2 in Fig. 1). We assessed the performance of the unsupervised clustering approach (t-SNE dimensionality reduction followed by hierarchical clustering) for grouping cells of the same type using both the full dataset of NK and ILCs (as in Fig. 2a) and the more realistic mixtures used in Fig. 2b (see the "Materials and methods" section for details). For the full dataset (i.e., with similar numbers of cells from each cell type), AUC values obtained for each cell type were very good (Fig. 2c). However, when performing clustering on unbalanced, and therefore more realistic cell type proportions, the performance of this approach dropped significantly (Fig. 2d). Altogether, these results show that our supervised approach trained in mouse outperforms the unsupervised clustering approach and allows more precise identification of ILC subtypes among unbalanced NK/ILC populations when NK cells greatly outnumber ILCs. We note that this is a known limitation of unsupervised clustering approaches when some cell types are found in much lower proportion than others and are not clearly distinct at the transcriptional level.

Identification of ILCs in mixed lymphocyte populations using single-cell RNA Seq data

We next asked whether ILCs could be identified from a mixed population of lymphocyte using single-cell RNA-Seq data. Previous studies have shown that the major types of lymphocytes (T, B, NK cells) can be separated by unsupervised clustering (Heng and Painter 2008; Tirosh et al. 2016; Carmona et al. 2017). To quantitatively explore this approach in the context of ILC identification, we combined single-cell RNA-seq data from human T and B cells obtained from melanoma tumors (Tirosh et al. 2016) (including cells without annotation to mimic more realistic situations) together with human NK/ILCs single-cell RNA-seq data (see the "Materials and methods" section). The proportions of T, B, and NK cells roughly mimic those expected in peripheral blood. Five cells from each ILC subtype were further included. We then applied t-SNE on the combined datasets using the 3000 most variable genes (Fig. 3a). Interestingly, the three clusters recapitulate the known cell types: cluster 1 contained mainly T cells, cluster 2 contained mainly B cells, and cluster 3 consists of a mixture of NK cells and ILCs. Our results therefore confirm that major lymphocyte cell types can be accurately distinguished in single-cell RNA-Seq with clustering techniques.

Figure 3b shows details of cluster 3, where colors represent the actual cell types (as Fig. 3a) and NK/ILC cell type predictions are shown with different shapes. In this example, four out of five ILC1 cells, two out of five ILC2 cells, and three out of five ILC3 cells were correctly predicted (see Table 3 for the full statistics on several random choices of the ILCs). More importantly, among the predicted ILC2 cells, all of them are actual ILC2, and among the predicted ILC3 cells, half of them are actual ILC3. Although the ratio between NK cells and ILCs in the example of Fig. 3b is not exactly the one expected in bulk lymphocyte populations from peripheral blood, Fig. 2a, b and Tables 2 and 3 indicate that the performance of our model is not affected by the fraction of ILCs versus NK cells, suggesting that it may be applied to bulk lymphocytes single-cell RNA-Seq data from different tissues containing different proportions of NK and ILCs, provided enough cells are sequenced to capture some ILCs (see the next section).

We also note that single cells without annotation in (Tirosh et al. 2016) did cluster together with both B and T cells (black dots, Fig. 3a), suggesting that most of them correspond to T and B cells and very few of them to NK cells or ILCs, although these results could also be influenced by batch effects between the different studies of origin. To further investigate whether ILCs could be found in these melanoma samples, we selected the patient that displayed the highest number of NK cells (Mel60, 10 cells annotated as NK) (Tirosh et al. 2016). Low-dimensional visualization of CD45+ cells showed a small cluster enriched in NK cells, in addition to the expected T and B cell clusters (Supplemental Fig. 1). We then applied

our ILC predictor on cells found in the NK cluster, focusing on ILC2 since this cell type showed the highest precision. Unfortunately, none of the cells found in this cluster were predicted as ILC2, suggesting that the limited number of lymphocytes (a few hundreds) may prevent identification of ILCs in this dataset. This is in line with estimates of ILCs proportions in melanoma tumors where they are expected to form less than 0.5% of the total number of tumor infiltrating lymphocytes.

General strategy proposed for ILC identification in different vertebrate species

These results prompted us to propose a combined experimental and computational strategy for identification of ILCs that is theoretically applicable in any vertebrate species with a sequenced genome and a minimum body size suitable for peripheral blood extraction (Fig. 4). Starting from blood samples, the first step consists of separating lymphocytes and monocytes from other blood cells (erythrocytes, thrombocytes, or granulocytes) by density gradient centrifugation (step 1). Although non-mammalian erythrocytes and thrombocytes are nucleated cells, lymphocyte and monocyte densities differ from both erythrocytes and thrombocytes, which make their isolation by density gradient centrifugation possible (Haugland et al. 2012; Nagasawa et al. 2014). Lymphocytes are then separated from myeloid cells with flow cytometry based on light scatter measurements (step 2). The last experimental part of the proposed procedure consists of using single-cell RNA-Seq on the mixed population of lymphocytes (step 3). After standard RNA-Seq reads mapping and transcript quantification (Trapnell et al. 2009; Li and Dewey 2011), clustering is applied and major immune cell types (T, B, NK/ILC) are identified based on expression of known markers (step 4; see also Fig. 3a). Finally, ILC predictions are carried out with our classifier on the NK/ILC cell cluster (step 5). As this classifier was built with signature genes that are conserved in all major clades of vertebrate species, it is applicable in most vertebrates.

In Fig. 4, we further provide estimates of the number of cells needed for this approach. For instance, if we aim at correct identification of at least 10 ILC2 cells, it is advisable to start with roughly 50,000 lymphocytes and monocytes. This corresponds more or less to 10 μl of peripheral blood, suggesting that our approach could work well for many relatively large vertebrates for which such amount of blood can be easily extracted. For much smaller species, additional steps for increasing the number of lymphocytes should be used such as pooling together samples from multiple individuals. Typically, 25% of them will be lymphocytes (12,500) out of which 15% are expected to be NK/ILCs. Assuming a ratio of 1/20 for ILC/NK cells and more or less the same number of ILC1/2/3 cells (in practice, the fraction of ILCs can vary significantly between conditions, tissues, and species), we therefore expect to have roughly 30 ILC2 cells in these data. Finally, approximately 35% of ILC2 cells are correctly captured by our predictions (1.6/5, second column in Tables 2 and 3), which would lead on average to 10 detected ILC2 cells, with a purity (i.e., precision) of 90% (second line in Tables 2 and 3). Current datasets of immune cells in non-mammalian species clearly do not reach these numbers (Carmona et al. 2017), which prevents direct application of the proposed strategy. However, single-cell RNA-Seq technology is evolving extremely fast and sequencing tens of thousands of cells is increasingly becoming feasible (Macosko et al. 2015; Zheng et al. 2017), highlighting the importance and promise of computational tools to identify rare cell types such as ILCs in single-cell RNA-Seq data (Grün et al. 2015; Proserpio and Lönnberg 2016).

Discussion

The molecular characterization of different immune cell types in mouse and human is a cornerstone of modern immunology and has enabled detailed understanding of cellular mechanisms underlying immune recognition processes in infectious diseases and cancer. Unfortunately, much less is known about immune cell types in other vertebrate species and most of our

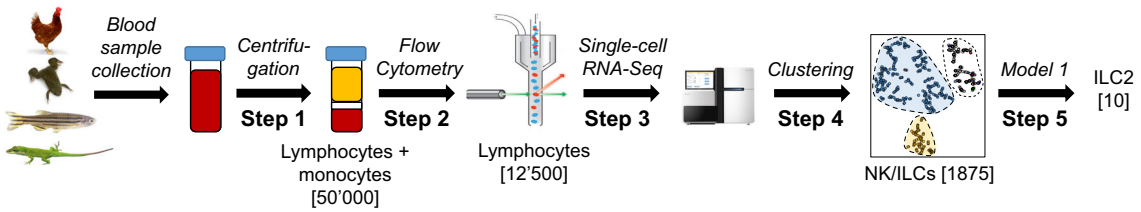

Fig. 4 Proposed strategy for identification of ILCs in peripheral blood samples from vertebrate species. Starting from peripheral blood, lymphocytes and monocytes are isolated by density gradient centrifugation (step 1). Lymphocytes are separated from myeloid cells with flow cytometry based on light scatter values (step 2). Mixed lymphocyte populations are sequenced at the single-cell level (single-cell RNA-Seq) (step 3). Cells are then clustered based on whole transcriptome similarity to identify T, B, and NK/ILC cell populations (step 4). Finally, our new classifier is applied on NK/ILC cells to identify ILC subtypes (step 5). Estimates for the number of cells at each step are shown in *brackets*

understanding of immune cell type conservation comes from orthology of mammalian immune marker genes. Orthology-based approaches are reasonable for cell types like T or B cells but are less suitable for other cell types like ILCs, which lack expression of specific lineage markers that entirely specific for ILCs (i.e., not expressed in other cell types).

Here we propose and validate a novel strategy to identify and study ILCs (in particular ILC2 cells) at the transcriptomic level using single-cell RNA-Seq technology. In mouse and human, a major asset of our work is to overcome the long and delicate existing sorting procedures to isolate ILCs, which may alter their transcriptome and viability. In these two species, we also note that the high number of cells estimated in Fig. 4 may be significantly decreased by negative sorting of T and B cells, which is much less challenging than complete sorting of the different ILC subtypes. It is tempting to speculate that our approach may be also applied to single-cell suspensions obtained by mechanical or enzymatic dissociation of other immune infiltrated tissues (e.g., the gut, the lung), where ILCs are known to play an important role and can be difficult to physically isolate. In particular, the recently characterized role of ILCs in cancer indicates that these cells may be of therapeutic interest (Trabanelli et al. 2015; Vallentin et al. 2015; Tian et al. 2016). However, to apply the tools developed in this study, the number of single cell data needs to be large enough to clearly detect a NK/ILC cluster. Moreover, we cannot exclude that ILCs may change their expression profile in inflamed tissues compared to blood. As our model is trained using ILCs and NK cells derived from blood, further studies will be required to validate it for predictions of ILCs infiltrating solid tissues.

A second advantage of our proposed strategy is that it does not rely on species-specific antibodies or transgenic animals, and therefore may be applied in virtually any vertebrate species with a sequenced genome and availability of blood or immune infiltrated tissue samples (Fig. 4). Starting from peripheral blood samples, our approach takes advantage of the fact that lymphocytes can be isolated by successive application of density gradient centrifugation and flow cytometry (light scatter properties). The rest of the analysis builds upon the power of single-cell RNA sequencing and our new NK/ILC classifier. Nevertheless, for this evolutionary application, an important assumption of our approach is that ILCs, if present in non-mammalian species, display at least some similarity with mammalian ILCs, hence the use of the conserved signature genes. Based on our current understanding of immune cell types in vertebrate species, this assumption is realistic. For instance, even NK cells show conserved gene signatures at the gene expression level for cytoplasmic proteins, despite the low conservation of NK receptors (Carmona et al. 2017). Should ILCs show completely different patterns in non-mammalian species or display large transcriptional changes in different conditions and activation states, our model will fail to identify them. Therefore, we cannot guarantee successful application of the proposed strategy to any non-mammalian species. But in the absence of other approaches, we anticipate that our work provides an interesting possibility to investigate the conservation of ILCs in non-mammalian species.

We cannot exclude that the clustering shown in Fig. 3a between T, B, and NK/ILCs is partly influenced by the different study of origins of these cells. Unfortunately, there exists currently no single-cell RNA-Seq dataset with experimentally verified ILCs taken directly from a mixed population of lymphocytes. For this reason, we had to resort to immune cells obtained from different studies in order to validate our predictor. However, compelling evidence indicates that T, B, and NK cells display distinct gene expression signatures in many vertebrate species that can be captured by unsupervised clustering techniques (Heng and Painter 2008; Tirosh et al. 2016; Carmona et al. 2017) and that ILCs show stronger transcriptomic similarity with NK cells. Our results therefore indicate that starting from a pool of bulk lymphocytes from peripheral blood or other immune infiltrated tissues and using single-cell RNA-Seq technology, ILCs, and especially ILC2 may be accurately identified with our model.

Annotation of the NK/ILC cluster (step 4 in Fig. 4) may not be straightforward in some species, since typical NK receptors are poorly conserved across vertebrates. Other NK markers that may be used include FcRγ, Fasl, and Gzmb that are specifically expressed in NK cells among resting lymphocytes and are well conserved across vertebrates (Carmona et al. 2017). Moreover, standard markers of T and B cells are well conserved across vertebrates. Therefore, identification of the T and B cell clusters should be relatively easy and, by complementarity, the NK/ILC cluster could be identified. Another issue may arise with the quality of genome annotations. For instance, signature genes shown in Table 1 may not be annotated in some genomes with a low coverage. This may limit the application of our work to non-mammalian model species with a carefully annotated genome. Yet, the steady increase in sequencing throughput and the drop in sequencing costs lead to constant improvements in the quality of newly assembled genomes, suggesting that our approach will be increasingly applicable in different organisms.

Collectively, we propose and validate a novel approach to study ILCs, and especially ILC2, at the transcriptomic level without having to sort them with flow cytometry and without relying on species-specific antibodies. As our model only uses genes conserved in vertebrate species, this work may contribute to exploring the conservation of immune cell types in non-mammalian species. Here, we focused on ILCs as they are the latest discovered immune cell types and their conservation across vertebrates is still not fully established (Vivier et al. 2016). However, our framework may be extended to other immune cell subtypes (e.g., T helper cells, NKT cells,

regulatory B cells) by training new classifiers for these cell types based on gene expression data of sorted cells in mouse or human.

Acknowledgements D.G. is supported by the Ludwig Institute for Cancer Research (LICR) and the Center for Advanced Modeling Science (CADMOS); S.J.C. is supported by SystemsX (MelamonX grant). C.J is supported by the Swiss National Science Foundation (Ambizione Fellowship PZOOP3_161459).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akula S, Thorpe M, Boinapally V, Hellman L (2015) Granule associated serine proteases of hematopoietic cells—an analysis of their appearance and diversification during vertebrate evolution. PLoS One 10: e0143091. doi:10.1371/journal.pone.0143091

Alabyev BY, Guselnikov SV, Najakshin AM et al (2000) CD3epsilon homologues in the chondrostean fish *Acipenser ruthenus*. Immunogenetics 51:1012–1020

Altenhoff AM, Škunca N, Glover N et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–D249. doi:10.1093/nar/gku1158

Artis D, Spits H (2015) The biology of innate lymphoid cells. Nature 517: 293–301. doi:10.1038/nature14189

Bernink JH, Krabbendam L, Germar K et al (2015) Interleukin-12 and -23 control plasticity of CD127(+) group 1 and group 3 innate lymphoid cells in the intestinal lamina propria. Immunity 43:146–160. doi:10.1016/j.immuni.2015.06.019

Björklund AK, Forkel M, Picelli S et al (2016) The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol. doi:10.1038/ni.3368

Boehm T (2012) Evolution of vertebrate immunity. Curr Biol 22:R722–R732. doi:10.1016/j.cub.2012.07.003

Brocker C, Thompson D, Matsumoto A et al (2010) Evolutionary divergence and functions of the human interleukin (IL) gene family. Hum Genomics 5:30–55. doi:10.1186/1479-7364-5-1-30

Carmona SJ, Teichmann SA, Ferreira L et al (2017) Single-cell transcriptome analysis of fish immune cells provides insight into the evolution of vertebrate immune cell types. Genome Res 27(3): 451–461. doi:10.1101/gr.207704.116

Chea S, Schmutz S, Berthault C et al (2016) Single-cell gene expression analyses reveal heterogeneous responsiveness of fetal innate lymphoid progenitors to notch signaling. Cell Rep 14:1500–1516. doi: 10.1016/j.celrep.2016.01.015

Constantinides MG, McDonald BD, Verhoef PA, Bendelac A (2014) A committed precursor to innate lymphoid cells. Nature 508:397–401. doi:10.1038/nature13047

Delconte RB, Shi W, Sathe P et al (2016) The helix-loop-helix protein ID2 governs NK cell fate by tuning their sensitivity to interleukin-15. Immunity 44:103–115. doi:10.1016/j.immuni.2015.12.007

Diefenbach A, Colonna M, Koyasu S (2014) Development, differentiation, and diversity of innate lymphoid cells. Immunity 41:354–365. doi:10.1016/j.immuni.2014.09.005

Eberl G, Colonna M, Di Santo JP, McKenzie ANJ (2015) Innate lymphoid cells. Innate lymphoid cells: a new paradigm in immunology. Science 348:aaa6566–aaa6566. doi:10.1126/science.aaa6566

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22. doi:10.1109/TPAMI.2005.127

Grün D, Lyubimova A, Kester L et al (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525:251–255. doi:10.1038/nature14966

Haugland GT, Jordal A-EO, Wergeland HI (2012) Characterization of small, mononuclear blood cells from salmon having high phagocytic capacity and ability to differentiate into dendritic like cells. PLoS One 7:e49260. doi:10.1371/journal.pone.0049260

Heng TSP, Painter MW (2008) The immunological genome project: networks of gene expression in immune cells. Nat Immunol 9:1091–1094. doi:10.1038/ni1008-1091

Ishizuka IE, Chea S, Gudjonson H et al (2016) Single-cell analysis defines the divergence between the innate lymphoid cell lineage and lymphoid tissue-inducer cell lineage. Nat Immunol 17:269–276. doi:10.1038/ni.3344

Jiang W, Chai NR, Maric D, Bielekova B (2011) Unexpected role for granzyme K in CD56bright NK cell-mediated immunoregulation of multiple sclerosis. J Immunol 187:781–790. doi:10.4049/jimmunol.1100789

Kaiser P, Rothwell L, Avery S, Balu S (2004) Evolution of the interleukins. Dev Comp Immunol 28:375–394. doi:10.1016/j.dci.2003.09.004

Kløverpris HN, Kazer SW, Mjösberg J et al (2016) Innate lymphoid cells are depleted irreversibly during acute HIV-1 infection in the absence of viral suppression. Immunity 44:391–405. doi:10.1016/j.immuni.2016.01.006

Konya V, Mjösberg J (2016) Lipid mediators as regulators of human ILC2 function in allergic diseases. Immunol Lett. doi:10.1016/j.imlet.2016.07.006

Koppang EO, Fischer U, Moore L et al (2010) Salmonid T cells assemble in the thymus, spleen and in novel interbranchial lymphoid tissue. J Anat 217:728–739. doi:10.1111/j.1469-7580.2010.01305.x

Koues OI, Collins PL, Cella M et al (2016) Distinct gene regulatory pathways for human innate versus adaptive lymphoid cells. Cell 165:1134–1146. doi:10.1016/j.cell.2016.04.014

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 12: 323. doi:10.1186/1471-2105-12-323

Lim AI, Menegatti S, Bustamante J et al (2016) IL-12 drives functional plasticity of human group 2 innate lymphoid cells. J Exp Med 213: 569–583. doi:10.1084/jem.20151750

Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214. doi:10.1016/j.cell.2015.05.002

Moltke von J, O'Leary CE, Barrett NA et al (2017) Leukotrienes provide an NFAT-dependent signal that synergizes with IL-33 to activate ILC2s. J Exp Med 214:27–37. doi:10.1084/jem.20161274

Moore FE, Garcia EG, Lobbardi R et al (2016) Single-cell transcriptional analysis of normal, aberrant, and malignant hematopoiesis in zebrafish. J Exp Med 213:979–992. doi:10.1084/jem.20152013

Nagasawa T, Nakayasu C, Rieger AM et al (2014) Phagocytosis by thrombocytes is a conserved innate immune mechanism in lower vertebrates. Front Immunol 5:445. doi:10.3389/fimmu.2014.00445

Neu KE, Tang Q, Wilson PC, Khan AA (2017) Single-cell genomics: approaches and utility in immunology. Trends Immunol 38:140–149. doi:10.1016/j.it.2016.12.001

Parker HS, Leek JT, Favorov AV et al (2014) Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics 30:2757–2763. doi:10.1093/bioinformatics/btu375

Proserpio V, Lönnberg T (2016) Single-cell technologies are revolutionizing the approach to rare cells. Immunol Cell Biol 94:225–229. doi:10.1038/icb.2015.106

Ranson T, Vosshenrich CAJ, Corcuff E et al (2003) IL-15 is an essential mediator of peripheral NK-cell homeostasis. Blood 101:4887–4893. doi:10.1182/blood-2002-11-3392

Ritchie ME, Phipson B, Wu D et al (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47–e47. doi:10.1093/nar/gkv007

Robinette ML, Fuchs A, Cortez VS et al (2015) Transcriptional programs define molecular characteristics of innate lymphoid cell classes and subsets. Nat Immunol 16:306–317. doi:10.1038/ni.3094

Šedý J, Bekiaris V, Ware CF (2015) Tumor necrosis factor superfamily in innate immunity and inflammation. Cold Spring Harb Perspect Biol 7:a016279. doi:10.1101/cshperspect.a016279

Seehus CR, Aliahmad P, la Torre de B et al (2015) The development of innate lymphoid cells requires TOX-dependent generation of a common innate lymphoid cell progenitor. Nat Immunol 16:599–608. doi:10.1038/ni.3168

Shih H-Y, Sciumè G, Mikami Y et al (2016) Developmental acquisition of regulomes underlies innate lymphoid cell functionality. Cell 165:1120–1133. doi:10.1016/j.cell.2016.04.029

Silver JS, Kearley J, Copenhaver AM et al (2016) Inflammatory triggers associated with exacerbations of COPD orchestrate plasticity of group 2 innate lymphoid cells in the lungs. Nat Immunol 17:626–635. doi:10.1038/ni.3443

Sonnenberg GF, Artis D (2015) Innate lymphoid cells in the initiation, regulation and resolution of inflammation. Nat Med 21:698–708. doi:10.1038/nm.3892

Spits H, Bernink JH, Lanier L (2016) NK cells and type 1 innate lymphoid cells: partners in host defense. Nat Immunol 17:758–764. doi:10.1038/ni.3482

Tian Z, van Velkinburgh JC, Wu Y, Ni B (2016) Innate lymphoid cells involve in tumorigenesis. Int J Cancer 138:22–29. doi:10.1002/ijc.29443

Tirosh I, Izar B, Prakadan SM et al (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352:189–196. doi:10.1126/science.aad0501

Trabanelli S, Curti A, Lecciso M et al (2015) CD127+ innate lymphoid cells are dysregulated in treatment naïve acute myeloid leukemia patients at diagnosis. Haematologica 100:e257–e260. doi:10.3324/haematol.2014.119602

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. doi:10.1093/bioinformatics/btp120

Vallentin B, Barlogis V, Piperoglou C et al (2015) Innate lymphoid cells in cancer. Cancer Immunol Res 3:1109–1114. doi:10.1158/2326-6066.CIR-15-0222

Van der Maaten L, Hinton GE (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579

Vivier E, van de Pavert SA, Cooper MD, Belz GT (2016) The evolution of innate lymphoid cells. Nat Immunol 17:790–794. doi:10.1038/ni.3459

Wang T, Holland JW, Carrington A et al (2007) Molecular and functional characterization of IL-15 in rainbow trout *Oncorhynchus mykiss*: a potent inducer of IFN-gamma expression in spleen leukocytes. J Immunol 179:1475–1488

Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049. doi:10.1038/ncomms14049