



UPPSALA
UNIVERSITET

UPTEC X 22034

Examensarbete 30 hp

01–23

Single-cell RNA-seq mapping of chicken leukocytes

An investigation into single-cell transcriptomics as an alternative to traditional immunological methods within non-traditional model organisms

Matilda Maxwell



Civilingenjörsprogrammet i molekylär bioteknik



UPPSALA
UNIVERSITET

Single-cell RNA-seq mapping of chicken leukocytes

Matilda Maxwell

Abstract

The immune system is a complex infrastructure where many cells interact with each other and perform duties depending on their type and function. When using traditional immunological methods in studying non-traditional model organisms, such as birds, challenges arise. These are often associated with a lack of knowledge surrounding the organism in question—particularly, the expected types of leukocytes, their cell-specific marker genes, and associated reagents. Single-cell transcriptomics allows us to study the immune system at the level of each singular cell and create a profile of each cell present in a sample without as much prior knowledge of the organism. This project aimed to investigate the possibility of using single-cell transcriptomics as an alternative to traditional laboratory methods in avian immunology and using this as a basis for further research into avian medicine. The study was performed by sequencing the mRNA in approximately twenty thousand individual chicken blood cells from 4 healthy adult birds, performing unsupervised clustering of the cells, and attempting to annotate clusters based on expression profiles. Most of this study has been performed using the *R*-based package Seurat and 10 x genomics software *Cell Ranger*.

Putative cell types discovered include expected populations such as several different T-cells, B-cells, monocytes, thrombocytes, red blood cells, and cells in various stages of the cell life cycle. After computational analysis, the number of cells per cell type corresponds to laboratory analysis of the cell types performed prior to sequencing by fluorescence-activated cell sorting. This indicates that the in-silico annotation of putative cell types is consistent with the known cell types in the samples.

This study of chicken leukocytes highlights the possibility of the usage of single-cell transcriptomics within non-traditional model organism immunology. It shows that using modern single-cell sequencing and existing software, sequencing-based characterisation of immune cells is possible and could prove a robust option in immunology study cases where traditional methods are limited.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Rober Söderlund Ämnesgranskare: Staffan Svärd

Examinator: Pascal Milesi

Unveiling the hidden functions of the immune system

Popular Science Summary

Matilda Maxwell

The immune system is the collection of cells that protect our body from infection and disease in various ways. The immune system is a complex interplay between lots of different types of cells that perform a range of different functions in the body (Sompayrac 2015). In humans, much research has been performed, and more is underway to find useful information about the immune system and its mechanisms (Varadé *et al.* 2021). Unfortunately, the same is not true for all other animals.

Chickens are one of the world's most important and widely used livestock animals (Conway 2020), yet little medical research has been conducted on them. The welfare of chickens is of great interest from an economic point of view for farmers but also from a sustainability point of view. Healthy animals provide both better yields and live longer (Glisson *et al.* 2013). Therefore, the national veterinary institute (SVA) is working on developing new techniques in chicken medicine. Unfortunately, when working with less studied animals, such as chickens, methods traditionally used in the lab are limited because they rely on existing knowledge about the cells (Adan *et al.* 2017). In such cases, it may be useful to approach the problem from a different angle. Instead of studying cells based on what we know about them, we want to learn as much as possible about all cells without preconceptions. One can think of it as instead of deciding in advance what you want to look at in a set of cells; you look first at what cells are available.

We do this using a technique called “single-cell RNA sequencing” (scRNA-seq). This means that all the genes that a cell is using are found. This gives us information about what each individual cell in the immune system is doing (Eberwine *et al.* 2014). Using this information, we can map every cell in a blood sample and determine which cells are there. Cells are mapped by grouping the most similar cells. This gives us a map of all the cells. The map comprises many small islands (or clusters as they are called) of cells, with all cells doing the same thing forming one island. By then looking at what the different islands have in common and what distinguishes them from the other islands, you can determine what kind of cells they consist of. In this way, we have been able to identify several types of immune cells found in chickens and investigate what their functions might be.

Contents

Abstract.....	3
Popular Science Summary.....	5
Abbreviations.....	9
1 Background.....	11
1.1 Aims of this project.....	11
1.2 Importance of livestock medicine.....	11
1.3 Immunology in non-traditional model organisms.....	12
1.3.1 Avian immune cells.....	12
1.4 Single-cell RNA sequencing.....	14
1.4.1 Technology.....	15
1.5 Bioinformatic tools.....	15
1.5.1 Cell ranger.....	15
1.5.2 Integration of multiple datasets.....	16
1.5.3 Principle component analysis.....	17
1.5.4 Nearest neighbour algorithms.....	18
1.5.5 The Louvain algorithm and modularity.....	18
1.5.6 Uniform Manifold Approximation and Projection.....	19
1.5.7 Wilcoxon rank sum test.....	19
2 Materials and methodology.....	20
2.1 Sample preparation (performed at SVA).....	20
2.2 Data.....	21
2.2.1 Reference files.....	21
2.2.2 Sequencing output.....	21
2.3 Read counting.....	21
2.4 Data preparation.....	21
2.4.1 Doublet removal.....	22
2.4.2 Quality Control.....	22
2.4.3 Data normalisation.....	22
2.4.4 Data integration.....	22
2.5 Cluster determination.....	23
2.5.1 PCA.....	23
2.5.2 Cluster determination.....	23
2.5.3 UMAP.....	23
2.6 Cluster annotation.....	24
2.6.1 DE analysis.....	24
2.6.2 GO terms.....	24

Re-analysis.....	24
2.7 Marker evaluation.....	24
3 Results	24
3.1 T-cells	27
3.2 B-cells.....	29
3.3 Monocytes.....	31
3.4 Proliferating cells.....	33
4 Discussion	34
4.1 Technical limitations.....	36
4.2 Biological limitations.....	37
4.3 Future work	37
5 Conclusion.....	38
6 Ethical permits.....	38
7 Acknowledgements.....	38
8 Supplementary material	39
Appendix A Quality metrics.....	49
Appendix B Comparison of clustering algorithms.....	51
Appendix C Resolution comparison	52
Appendix D UMAP values comparison	53
Appendix E Marker gene quality control	54
E.1 TARP and TRBV65.....	54
E.2 FOXP3.....	54
Appendix F Sample compositions.....	55

Abbreviations

ANNOY– approximate nearest neighbour algorithm oh yeah

CD - Cluster of differentiation

cDNA - copy DNA

CTL - Cytotoxic T-cell

DC - Dendritic cell

DE - Differential expression

FACS - Fluorescence-activated cell sorting

GO - Gene ontology

Ig - Immunoglobulin

JCHAIN - joining chain of multimeric IgA and IgM

KNN - K-nearest neighbours

mRNA - Messenger RNA

MHC - Major histocompatibility complex

NK - Natural killer

PC - Principal component

PCA - Principal component analysis

PCR - Polymerase chain reaction

RBC - Red blood cell

SNN - Shared nearest neighbour

SLM - Smart local moving

SVA - Statens veterinärmedicinska anstalt
- national veterinary institute

scRNA-seq - Single-cell RNA sequencing

STAR - Spliced Transcripts Alignment to a Reference

TCR - T-cell receptor

TARP - TCR gamma alternate reading frame

UMAP - Uniform Manifold
Approximation and Projection

UMI - Unique molecular identifier

UPPMAX - Uppsala Multidisciplinary
Center for Advanced Computational
Science

1 Background

Chickens are both globally and locally in Sweden important animals in agriculture, yet the research surrounding their health and immune system is underdeveloped. To ensure a sustainable and efficient poultry industry it is crucial to safeguard the health of poultry. The national veterinary institute (SVA) has recently developed a method for analysing leukocytes in blood samples from chickens (Wattrang *et al.* 2020). Still, necessary reagents and antibodies for the types of white blood cells in chickens are currently lacking (Ratcliffe 2006), hindering effective, laboratory-based research into chicken immunology.

This project is part of a larger project at SVA intending to expand the knowledge of the immune system in chickens to enable better usability of blood cell analysis in poultry medicine.

1.1 Aims of this project

This project aims to investigate the possibility of using single-cell RNA-sequencing (scRNA-seq) to 1. help identify leukocytes that are already known today and infer their functions; 2. identify unknown populations of leukocytes and infer their functions. This would enable further research into infectious diseases in hens and work as a prerequisite for the development of treatments against these.

The project aim has been fulfilled by producing a map of chicken leukocytes, annotated with respect to their putative cell type and studied with respect to their marker gene expression. The mapping has been achieved by scRNA-seq of cells present in chicken blood samples. The cells are mapped to the extent possible with respect to their cell type, the fraction of cells belonging to each type, and their function.

1.2 Importance of livestock medicine

Poultry farming is an integral part of agriculture today, both on a global scale and locally in Sweden. The primary purpose of poultry farming is for food through the production of eggs and meat. Poultry is the most numerous farm animal in the world, and chickens made up almost 40% of meat production worldwide in 2020. As the global population continues to increase, so will poultry meat and egg consumption. The consumption of poultry is expected to increase, especially in lower-income countries, due to the cheap nature of poultry compared to other meats (Conway 2020). This means that it is crucial to ensure a sustainable and profitable poultry industry, and with that comes ensuring a good health status in poultry.

1.3 Immunology in non-traditional model organisms

Studying immunology is vital for understanding an individual's health, and the study of leukocytes has long been used in clinical diagnostics and research in human and animal medicine practices. The composition of leukocytes in the blood is affected by and can be indicative of disease (Wattrang *et al.* 2020) or physiological stress in an individual (Scanes 2016). Among mammals, the current leading method of leukocyte study is flow cytometry (fluorescence-activated cell sorting (FACS)) (Maecker *et al.* 2012). Where a cell is analysed using visible light scatter as it flows through a laser beam. Cells are also often labelled (usually with fluorophores) before flow cytometry, allowing for many parameters to be studied, such as a cell's genetic content, protein content, or antibody affinity (Adan *et al.* 2017).

Within non-mammals however, studying differences in the composition of the immune system can cause difficulties when using traditional methods. Such is the case in chickens, where leukocytes that are not present in mammals and the presence both nucleated red blood cells (RBCs) and thrombocytes (platelets) complicates the usage of these techniques (Kaiser & Balic 2015). These differences in leukocyte identity lead to a need for different reagents used for studying chicken-specific markers, such as immune cell surface receptors, transcription factors, and cytokines. Currently, access to such reagents for chickens is limited, making traditional immunology methods expensive and time-consuming.

1.3.1 Avian immune cells

The avian immune system has many similarities to that of mammals but also some key differences (Kaiser & Balic 2015). When trying to identify different cell populations, marker genes are used. These are genes that are highly expressed in one population and allow for distinguishing populations from each other (Kiselev *et al.* 2019). Below are listed some common leukocytes in chickens and their potential marker genes.

1.3.1.1 T-cells

The T-cells are a type of immune cell developed in the thymus and defined by their carrying of the T-cell receptor (TCR). The T cell receptor is responsible for antigen recognition in the T-cells. The receptor is a complex consisting of subunits, some constant, and some variable. The constant subunits form the CD3 (cluster of differentiation 3) complex, consisting of the subunits CD3 γ , CD3 δ , and CD3 ϵ . In chickens, these are encoded by two genes, CD3 γ/δ and CD3 ϵ (Smith & Göbel 2022). The variable subunits are either α/β or γ/δ subunits (Charles A Janeway *et al.* 2001). The two major lineages of T-cells are α/β T-cells and γ/δ T-cells. These are divided into sub-populations based on other surface molecules. These include among others the α/β T-cells CD4 positive T-cells (also T helper cells) and CD8 positive T-cells (cytotoxic T-cells (CTL)). CD8 consists of 2 subunits called CD8 α and CD8 β . These are either expressed as a CD8 $\alpha\alpha$ or a CD8 α/β heterodimer on the cell surface. In the chicken, there are populations of both α/β TCR+ as well as γ/δ TCR+ that express either of these CD8 types (Smith & Göbel 2022). The regulatory T-cells (Tregs) are defined in mammals as CD4,

CD25, and FOXP3 positive (Shanmugasundaram & Selvaraj 2011). The transcription factor FOXP3 has only recently been discovered within the chicken genome and is expected to make the identification of Tregs in chickens easier due to its high Treg specificity (Burkhardt *et al.* 2022). Natural killer (NK) cells in chickens are not well studied, and conclusive information regarding the NK cell receptors in chickens is unavailable (Straub *et al.* 2013). A marker gene for all NK cells for chickens is still missing, but it is known that NK cells are CD3 cell surface negative and have a higher expression of CD107 (LAMP1) (Meijerink *et al.* 2021).

1.3.1.2 B-cells

The B-cells in birds are developed in an organ not present in mammals, the bursa of Fabricius. Lymphoid precursors mature within the bursa before travelling to the blood. The purpose of the B-cells is to produce antibodies as a reaction to pathogens. B-cell specificity is mediated by their immunoglobulin genes. Immunoglobulin (Ig) is a protein complex consisting of light and heavy chains. Chickens have one immunoglobulin light chain gene and three heavy chain genes (the genes encode the μ , α , and ν Ig heavy chains of the IgM, IgA, and IgY immunoglobulins, respectively). Cytokines largely influence mature B-cells, and receptors for these can be used as B-cell identifiers. One such receptor is the BAFF- receptor (TNFSF13B). Other important B-cell markers include CD40, and cytokines IL7, IL10, and IL2 (Ratcliffe & Härtle 2022). The chicken B-cells also have a unique expression of Bu-1 (Gilmour *et al.* 1976) and expression of BLIMP1 and PAX5, which are involved in maturing of plasma and B-cells, respectively (Nutt *et al.* 2007). SOX5 is also expressed in B-cells and is associated with late-stage B-cell differentiation and plasmablast differentiation (Rakhmanov *et al.* 2014).

Plasma cells are ultimately differentiated B-cells capable of secreting vast amounts of immunoglobulin (Ratcliffe & Härtle 2022). Plasma cells can be identified by their strong expression of genes related to this activity, such as JCHAIN and BLIMP-1 (PRDM1). JCHAIN stands for “joining chain of multimeric IgA and IgM” and is a protein responsible for binding activity in dimers of Immunoglobulin M and Immunoglobulin A (Frutiger *et al.* 1992). BLIMP-1 is a transcription factor involved in regulating the development of a variety of immune cells and is expressed within all antibody-secreting cells. Within B- cells, it is involved in plasma cell differentiation and is only upregulated in plasma and plasmablasts cells (Nutt *et al.* 2007).

1.3.1.3 Monocytes and dendritic cells (DC)

Monocytes and dendritic cells (DCs), like B-cells, are antigen-presenting (Sutton *et al.* 2022). Antigen presentation is the expression of antigens on the outer surface of a cell (Cruse *et al.* 2004). The antigen is bound to the cell surface through interaction with either Major histocompatibility complex (MHC) class I or II molecules (Sutton *et al.* 2022). Several genes are used to separate DCs from monocytes. The cytokines FLT3 and XCR1 are involved in generating DCs but are not expressed in monocytes.

1.3.1.4 Granulocytes

The primary granulocytes in chickens are the heterophils. These are considered analogues to neutrophils in mammals. Heterophils are responsible for initiating the acute inflammatory response and producing substances with antimicrobial properties (Kogut 2022). They do this by recognising pathogens using surface receptors (pattern recognition receptors such as toll-like receptors and C-type lectin receptors) (Genovese *et al.* 2013). Their antimicrobial properties include phagocytosis, degranulation, antimicrobial peptides, and extracellular traps (Kogut 2022). Genes associated with these receptors and activities can be used as markers for heterophils.

1.3.1.5 Thrombocytes

Thrombocytes (platelets) are blood cells responsible for stopping bleeding. In mammals, these cells have no nuclei, but in birds, they are nucleated and carry genetic material. These are the most numerous white blood cell in chickens and have more immune functions than in mammals. These are defined by their expression of a homolog of mammalian integrin (CD41/CD61) (Astill *et al.* 2022 s. 8).

1.4 Single-cell RNA sequencing

Like all cells, immune cells achieve their differences in form and function by expressing genes at different levels to produce the proteins they need to do their jobs. Studying gene expression can tell us much about what is happening in a tissue at a given moment (NHGRI 2022). The problem with looking at all the gene expression in tissue (bulk RNA) is that we cannot tell which cells in a sample are doing what. This is when scRNA-seq is useful. It allows for all expressed genes in only one cell to be studied and can give information about what that cell is doing at that time (Eberwine *et al.* 2014).

The use of scRNA-seq for mapping cells is a relatively new method that contrasts with classical laboratory methods in that much of the work can be performed bioinformatically. Similar mappings of leukocytes have been performed on horses (Patel *et al.* 2021) and zebrafish (Chan *et al.* 2022), among others. These studies are similar to this project since the immune systems of these animals are also not previously well studied and mapped. ScRNA-seq enables a new way to classify and define cell types in a sample, and scRNA-seq can circumvent problems that arise with laboratory methods for cell type determination. An example of such a problem is that flow cytometry, a common practice for determining cell populations (Maecker *et al.* 2012), relies on existing knowledge of cell-type specific markers on the cell and access to antibodies against these markers (Adan *et al.* 2017). Such information is often available in well-known organisms, but this knowledge can be lacking in less studied organisms. When this is the case, scRNA-seq is a good method to study cell type, as it is largely independent of previous knowledge of the organism (one still requires a good genome assembly and annotation for reference) (Patel *et al.* 2021).

1.4.1 Technology

The data in this project was produced using 10x genomics chromium Single Cell 3' v3.1 Dual Index Gene Expression solution (10X genomics 2022a). The 10 x chromium instruments use a gel bead-in-emulsion technology referred to as next GEM. Cell samples and barcoded gel beads are loaded onto a chromium microfluidic chip, with one gel bead, and one cell joined in each GEM. Inside each GEM, the cell is lysed, and the barcode is attached to the molecule of interest (in this case, mRNA). The barcode is used to capture the identity of the cell in each gem, and molecules are tagged with unique molecular identifiers (UMI). The gems are then collected, copy DNA (cDNA) is generated through reverse transcription, pooled, and sequence libraries are generated (10X genomics 2021a). The libraries are standard Illumina paired-end constructs and start and end with sequences for binding to an Illumina flow cell. (10X genomics 2020a.)

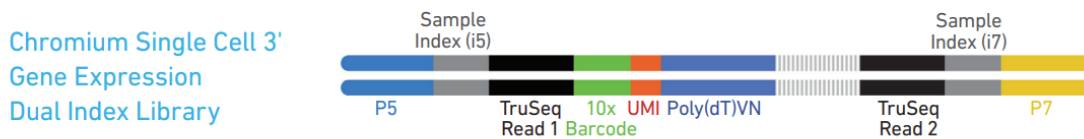


Figure 1 Composition of the Single Cell 3' Dual Index sequencing libraries. The region with grey lines corresponds to the cDNA insert. This is a 90 bp long region. It is flanked by sequences required for sequencing. (10X genomics 2020b).

The sequencing is then performed using standard sequencing procedures. In this project, it was done using the Illumina NovaSeq 6000 system (Illumina 2022a). The sequencing was performed with read lengths 28+10+10+90bp. As seen in Figure 1, these correspond to Cell barcode & unique molecular identifier (UMI) (Read 1, 28 bp), Sample Index (i7 Index 10 bp), Sample Index (i5 Index, 10 bp), and Insert (Read 2, 90 bp), respectively. The insert corresponds to the biological mRNA sequence (10X genomics 2021b). The UMI is a short sequence used to tag each molecule (sequence) in a library. This allows for indexing of the molecules and makes it possible to keep track of the original molecule from which the PCR amplicons stem (Illumina 2022b).

1.5 Bioinformatic tools

The main bioinformatic tools used in this study are *Cell Ranger* (10X genomics 2020c) and *Seurat* (Hao *et al.* 2021). Seurat is implemented in R as a library and includes many functions specifically developed to enable the analysis of single-cell data. The theoretical background of some of the implemented functions is described below.

1.5.1 Cell ranger

Cell Ranger works by aligning reads to the reference genome and counting reads that align to loci annotated as transcriptomic (exonic and intronic reads) in the reference. Before aligning all the reads to the genome, they are trimmed to remove poly-A and template switch oligo

sequences (10X genomics 2018a) from the sequence's 3' and 5' end, respectively, to avoid these causing confounding mappings.

The reads are aligned to the reference genome using *Spliced Transcripts Alignment to a Reference* (STAR) software (Dobin *et al.* 2013). STAR works in two phases, seed searching and clustering/stitching/scoring. The software performs the seed searching by searching for the longest matching substring of the read in the genome. The minimum mappable length decides how short the substring can be. The algorithm starts the seed searching from the first base of the read. Whatever part of the read remains unmapped gets used again for searching. This is done until the whole read is mapped. This way, STAR quickly aligns reads in a splice-aware manner, stopping the alignment when a splice junction is reached and continuing after. STAR also allows for alignments with indels and mismatches. It does so by extending the substring of the read. If that still does not give a sufficient match, the algorithm will drop the part of the read that has no match in the genome. The algorithm then moves into the second phase, where it stitches together the substrings that were successfully aligned. These are then clustered together by proximity to a set of anchors. All seeds that map within a window surrounding an anchor are stitched together. A scoring scheme is implemented, and the read combination with the highest score is assumed as the best alignment (Dobin *et al.* 2013).

The reads that STAR has confidentially mapped to transcriptomic regions are kept (the version of *Cell Ranger* used in this study does by default keep exonic and intronic reads in order to map unsliced reads as well, (10X genomics 2022b)) and aligned to the transcriptome (the transcriptome is produced as a part of the *Cell Ranger* reference). After mapping, *Cell Ranger* uses the annotation file to sort reads into exonic, intronic, or intergenic based on the position in the genome. These reads are then used downstream in the UMI counting (10X genomics 2020d). During the UMI counting, *Cell Ranger* groups the reads mapped as transcriptomic together based on their barcode, UMI, and gene annotation. In groups with the same gene and barcode but slightly different UMIs, the UMIs with lower support are corrected to match the high-support UMI. This is because small errors are likely to be introduced by the sequencing and cause UMIs from the same origin to differ slightly. If several groups of reads have the same barcode and UMI but different gene annotations, they are assigned the annotation with the highest support. This correction is done since all reads with the same (or highly similar) UMIs should stem from the same original sequence (post cDNA amplification). Each observed combination of gene, UMI, and barcode is used as a UMI count (10X genomics 2020d).

Cell Ranger count then gives filtered and unfiltered gene-barcode matrix files, among other outputs. The filtered matrices exclude barcodes representing non-cell-containing GEMs (10X genomics 2018b) and are used for this study.

1.5.2 Integration of multiple datasets

Seurat includes methods to integrate multiple datasets, this is done to match up shared cell populations between datasets to use as one reference. This is necessary since data from

different origins (data can be sourced from different individuals, technologies or modalities (Efremova & Teichmann 2020)) may include the same cell populations but represent them differently, causing them to not be classified as one population downstream (Stuart *et al.* 2019). As detailed by Stuart and co-workers, the data integration procedure consists of feature selection, anchor identification, and dataset merging.

Features in the datasets are used to identify anchors; since the anchors should be in a matched biological state, they would have a matched feature pattern. The anchor prediction does not require the full set of features; instead, a subset of heterogeneous features is used. The most variable features in each dataset are identified to use for downstream analyses since high variability across cells represents heterogeneous features. Heterogeneous features are used since they are likely to represent a strong biological signal unique to a cell type. These are found by identifying features that are outliers from the mean variability. The most variable features for each dataset are prioritised so that those that occur across multiple samples are prioritised (Stuart *et al.* 2019). These are then used for downstream analysis (anchor identification).

An anchor represents two cells from two datasets predicted to be of the same biological origin. The anchors are identified by dimensionality reduction on two datasets, followed by finding the K-nearest neighbours (KNN) for each cell. Mutual nearest neighbours are identified, where two cells are contained within each other's nearest neighbourhoods. These mutual nearest neighbours are used as anchors (Stuart *et al.* 2019).

The anchors are then scored and weighted. A weight matrix is constructed that defines the association between the cells in the dataset and the anchors. They are based on the distance between the query cell and anchor and the anchor score. Including the score ensures that good anchors with high scores are weighted higher than poorer-scored anchors. This weight matrix is then used to correct the expression matrix, which is used downstream as a normalised scRNA-seq matrix. When integrating more than two datasets, the datasets are integrated pairwise. Anchor identification is done pairwise, and the distances between datasets are computed as the total nr of cells in the smaller datasets divided by the number of anchors in the two datasets. Hierarchical clustering on the pairwise distances between all datasets is done to determine in which order to merge the datasets (Stuart *et al.* 2019).

1.5.3 Principle component analysis

Principal component analysis (PCA) is a technique for increasing the interpretability of a dataset without losing biological information. It is done by linearly transforming multidimensional data to a 2-dimensional coordinate system in a way that preserves as much variance as possible in the dataset. The largest variance is found by constructing a matrix from the multidimensional data and finding the largest eigenvalue of the said matrix (Jolliffe & Cadima 2016).

1.5.4 Nearest neighbour algorithms

K-nearest neighbours (KNN) is a supervised learning algorithm used to classify data points. It is built on the assumption that data points found near each other are similar. The goal is to identify the k nearest neighbours for a query point and use the class of these to define the class of the query point. The algorithm computes the distance between the query point and all other data points in the set using the straight line between the two points (the Euclidian distance). It then labels the query point based on the value of the majority of its k nearest neighbours. If k is set to 5, it will look at the values of the 5 points closest, using the calculated distance metrics, to the query and assign the point the value that occurs most in these 5. (IBM 2022).

In the case of large datasets, we want to avoid computationally heavy calculations. Therefore, instead of an algorithm that calculates and stores pairwise distances between all points, as KNN does, we opt to use a subset of the points and calculate the distances to those. This results in an approximation of the nearest neighbours rather than a perfect calculation (Apache Software Foundation 2022). The approximate nearest neighbour algorithm oh yeah (ANNOY) algorithm, does this by building a guide tree of random projections (Osika 2022). Each node in the tree represents a split of the dataset in half using a hyperplane, chosen by sampling two points and finding their equidistant (a plane that passes through their midpoint) using a Euclidian distance measure. Each tree includes $n * k$ (n = nr of trees in the forest here 50, k = nr of nearest points to calculate, in this case, 20) nodes. The tree construction is done n times to build a forest.

SNN algorithm clusters points based on their nearest neighbours; points are clustered if they have many overlapping nearest neighbours (Kumari *et al.* 2016). The similarity is calculated using the Jaccard index (Dodge 2008). The Jaccard index measures the similarity between finite datasets by calculating the intersection of the datasets divided by the union of the datasets. The cut-off for an acceptable Jaccard index is set to 1/15. Any edges in the SNN graph with a lower index than this are removed from the graph (Hoffman 2022b).

1.5.5 The Louvain algorithm and modularity

The Louvain algorithm and is used to define cluster boundaries in the dimensionality-reduced data (Hoffman 2022c). The Louvain algorithm is a modularity optimisation-based community detection algorithm (Blondel *et al.* 2008). It works by iteratively maximising a modularity score for each community and is a fast and memory-efficient algorithm (Lu *et al.* 2014).

Modularity is a measure used to determine how well nodes (here cells) in a network divide into modules (clusters). A network with high modularity has a high degree of connections between nodes within a module and sparse connections between modules. Modularity optimisation is a commonly used method for determining community structure in networks. Biological networks tend to have high degrees of modularity, making modularity optimisation appropriate in biological research (Newman 2006).

Modularity has been shown to suffer a resolution limit and may be unable to detect modules smaller than a certain scale which depends on the total network size since the mathematically found optimal partition may not capture the actual community structure. (Fortunato & Barthélemy 2007). Because of this, modularity optimisation methods with tuneable resolutions have been developed. On the other hand, these methods tend to split subgraphs at high resolutions or merge subgraphs at low resolutions. This leads to the need for resolution evaluation by the user (Lancichinetti & Fortunato 2011).

Clustree (Zappia & Oshlack 2018) can be used for evaluation of resolutions by gives plots consisting of clustering trees that show how the clusters change between iterations of a clustering parameter (here, resolutions). The plot consists of a tree with nodes where each node represents a cluster. The clusters are connected to show how the data points (here, cells) move between them. It can help the user tell which clusters are distinct and unstable and change depending on settings (Zappia 2020).

1.5.6 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that allows the visualisation of high-dimensionality data in a 2-dimensional plot. UMAP works by constructing a graph that works as a high-dimensional representation of the data. Points are connected based on a set radius, where other points whose radius overlaps with the radius of the first point connect to that point. Points are weighted based on how large the radius is, with bigger radii having a lower likelihood. The high-dimensional graph is then converted to a low-dimensional graph. This works by calculating similarity scores for points to preserve high-dimension clustering. The similarity scores are based on the distances between the points in each considered dimension. The similarity score is based on a negative exponential curve where the sum of the similarity scores for all points other than the considered point equal $\log_2(n.\text{neighbours})$. The score for a neighbouring point is equal to the y-value of this curve given the x-value of a neighbouring point. The neighbouring points are placed on the x-axis based on their distance from the considered point. The similarity scores are used to initialise a low-dimensional graph with all points on one axis. The location of the points on the axis is adjusted to maximise distances between points that are not in the same high-dimensional cluster (McInnes *et al.* 2020).

1.5.7 Wilcoxon rank sum test

The Wilcoxon rank sum test (also the Mann-Whitney U test) is a non-parametric alternative to the two-sample t-test for non-normal data and allows us to test a null hypothesis. The test aims to see if two distributions are shifted from each other by differences in median expression. The null hypothesis is that the two populations have the same distribution with the same median (Ford 2017); in this case no differences in expression distributions between the groups (Whitley & Ball 2002).

2 Materials and methodology

The project is part of a bigger project at the Swedish veterinary institute in which blood had been collected from healthy chickens, and scRNA-seq performed in collaboration with SciLifeLab. Comparative data from the same blood samples e.g. studies of chicken immune cells using immunofluorescence staining and flow cytometric analysis were also available.

The methodology of this project was performed in two parts. One was the read counting performed through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) (Lundberg 2022) using 10x genomics software *Cell Ranger* (10X genomics 2020c) (Version: 7.0.0). *Cell ranger count* was used to generate filtered feature-barcode matrices used in the project's second part.

The second part represented the bulk of the analyses. It was performed predominantly in the *Seurat* (Hao *et al.* 2021) (Version: 4.3.0) package in R (Version: 4.2.2). This part includes the data preparations, clustering, and visualisation steps.

2.1 Sample preparation (performed at SVA)

Blood samples were taken from four healthy adult female birds (24-week-old laying hen hybrids, Bovans Robust) kept in a clean but not sterile environment as blood donors at SVA. The blood was collected using regular blood sampling from the jugular vein into sample tubes with heparin as an anticoagulant.

Leukocytes in the samples, were separated from RBCs using gradient centrifugation on Ficoll (GE Healthcare) according to established protocols. Chicken platelets (thrombocytes) are nucleated and, therefore, contain mRNA. They may constitute a considerable amount of the cells obtained after gradient centrifugation on Ficoll (20-70%). Thus, the proportion of platelets in the samples was reduced to obtain better sequencing of rarer populations of white blood cells. For this immunomagnetic cell separation with EasySep PE Selection Kit (StemCell Technologies, protocol no. 28898), methodology was used according to protocols previously used with chicken white blood cells at SVA (Wattrang *et al.* 2019). This was done using an antibody to integrin CD41/61 (OriGene) that is expressed exclusively on platelets in chicken blood cells (Lacoste-Eleau *et al.* 1994). In addition to sequencing, the purified cell preparations were also analysed by immunofluorescence and flow cytometry to determine the proportion of various known types of leukocytes. (Wattrang *et al.* 2020).

Preparation of libraries from approximately nine thousand cells per bird was performed on the SNP/SEQ platform at SciLifeLab from fresh purified leukocytes. The libraries were used for 3' RNA-seq with Chromium NextGEM Single Cell 3' v3.1 kit (10x Chromium), and subsequent sequencing (Illumina) was performed at SciLifeLab (Frejd 2022). The library preparation resulted in approximately five thousand cells per bird. Sequencing was performed using a NovaSeq SP flow cell (Illumina) to an average depth of 35,000 reads/cell.

2.2 Data

The raw and resulting data, as well as the used scripts, can be found in the supplementary materials (<https://github.com/maxwelma/exjobb>).

2.2.1 Reference files

The read counting was performed using the existing reference genome and genome annotation for the chicken, scientific name *Gallus gallus* (NCBI taxa: 9031. Assembly: RefSeq GCF_016699485.2; GenBank GCA_016699485.1). The current genome assembly is from 2021 and was assembled by the Vertebrate Genomes Project. The bird used was of the heterogametic sex (female) and a cross between a maternal broiler (bGalGal2) and a paternal white leghorn layer (bGalGal3). The sequencing was performed using a combination of long and short-read sequencing of the offspring as well as the parents to create a high-quality reference genome (NCBI 2022a).

The annotation (annotation release ID: 106) used is from 2021/2022. The annotation was performed using the NCBI genome annotation pipeline, which automatically annotates genes, transcripts, and proteins on genomes. The annotation was performed using the same genome as above (bGalGal1.mat.broiler.GRCg7b) (NCBI 2022b).

2.2.2 Sequencing output

The sequencing data consists of four files per sample, two for each of the two lanes used. Each lane has two reads, where read 1 corresponds to barcodes and read 2 to cDNA sequences. The MultiQC report of the data shows that all sequences scored well in the general metrics, meaning that the raw data was of good quality.

2.3 Read counting

Read counting from the raw data was performed using *Cell Ranger* standard workflow (10X genomics 2020e). First, a *Cell Ranger* reference was created from the reference (bGalGal1.mat.broiler.GRCg7b) genome (.fasta) and annotation (.gtf) file using *Cell Ranger mkref* (script: make_ref.sh). Then read counting was performed for each sample using *Cell Ranger count* (script: count.sh). The outputted matrix files (data/processed/count_output/sample_name/outs/filtered_feature_bc_matrix) were used for downstream analyses.

2.4 Data preparation

Before continued analyses, the data needed to be processed. Putative doublets and damaged cells were removed to avoid skewing the results. Cells with high mitochondrial expression were removed as they are likely to represent damaged cells. The high mitochondrial count indicates that cytoplasmic mRNA has leaked out of the cell, leaving only the mRNA in the

mitochondria. Damaged cells are also indicated by low count depth and few detected genes (Luecken & Theis 2019).

2.4.1 Doublet removal

Doublets are errors that occur in droplet-based single-cell sequencing, where two cells have been encapsulated in the same droplet and received the same barcode. Doublets are characterised by a very high UMI count per barcode since they include double the amount of genetic material as one cell. However, not every sample with a high UMI is a doublet; some may be cells with very high activity (Luecken & Theis 2019). Therefore, using doublet prediction algorithms rather than simple UMI count cut-offs is preferable.

Putative doublets were removed from the matrices using the *DoubletDetection* (Gayoso *et al.* 2019) package in python. *DoubletDetection* utilises the creation of computer-generated doublets and compares the expression profiles of the synthetic doublets to the cells in the matrix and estimates based on the comparison which cells are likely to be doublets. The script predicts doublets and edits the barcode list size to remove barcodes corresponding to putative doublets. The doublet removal was performed through UPPMAX using a *conda* environment.

2.4.2 Quality Control

After doublet removal, the matrices were loaded into R and converted to Seurat objects. The data was further filtered in R. The filtering was done to ensure that only viable cells were used for downstream analyses.

Cells with a higher mitochondrial percentage than 20%, cells with a lower feature count than 300, and features that appear in less than three cells were filtered. RBCs were also filtered out based on HB-gene expression; cells with more than 5% HB genes were filtered out. Figures are available in Appendix A.

2.4.3 Data normalisation

After filtering, the data were normalised. This was done to compensate for potential differences in gene expression between cells. The normalisation was performed in *Seurat* using the *SCT-transform* independently on each sample. The SCT transform is a normalisation method developed for single-cell data. It uses regularised negative binomial regression to normalise UMI count. It is better at correcting for technical factors than other normalisation methods when handling scRNA count data while conserving biological heterogeneity. The data is returned as a Seurat object with the normalised data stored in assay “SCT” (Hafemeister & Satija 2019).

2.4.4 Data integration

In this project, four samples have been used. Since the differences between them are not of interest and they are studied as one entity, the four samples were integrated into one dataset. The data was integrated on 3000 variable features found using *SelectIntegrationFeatures()*. Seurat then identifies cells between datasets that are in a matched biological state and

assigning these as anchors *FindIntegrationAnchors()* (Hoffman 2022a). These anchors are then used for integration using *IntegrateData()*. The datasets are integrated pairwise by creating a weight matrix based on the association between the cells and the anchors. The weight matrices are used to correct the original expression matrix (Stuart *et al.* 2019).

2.5 Cluster determination

After data preparations, the primary analyses of the data were performed.

2.5.1 PCA

Dimensionality reduction was performed using PCA on the integrated data. The PCA was performed using the variable features associated with the assay. In this case, they are the same as were calculated for the data integration. The cumulative proportion of variance was calculated and used to determine how many principal components (PCs) to keep. The cut-off was set so that 90% of variance would be conserved; this resulted in 27 PCs being used for downstream analyses for the complete dataset.

2.5.2 Cluster determination

Seurat implements a combination of PCA, graph-based clustering, and the Louvain algorithm to determine clusters in the dataset (Kiselev *et al.* 2019). After PCA, the nearest neighbours for each cell are computed, and Seurat constructs a nearest neighbour graph and a shared nearest neighbour (SNN) graph (Hoffman 2022b). These are stored in the Seurat object.

After PCA, the nearest neighbours for each cell were computed using *FindNeighbors()* in *Seurat*. In Seurat, the distances between cells are calculated using the PCs established earlier (Hoffman 2022d). The nearest neighbour representation is in the form of a graph, where each data point is represented by a node (a cell) connected by a set of k edges to its neighbours (Levine *et al.* 2015). The data points are not classified in this step with regard to grouping, but rather the groupings are formed when calling *FindClusters()*, using the nearest neighbour graph calculated by *FindNeighbors()* (Hoffman 2022d).

Then the *FindClusters()* function was called to establish cluster boundaries using the Louvain algorithm. The function allows the user to choose other algorithms (Louvain algorithm, Louvain algorithm with multilevel refinement, Smart local moving algorithm (SLM), Leiden algorithm). The algorithm was chosen based on the comparison in Appendix B. The number of principal components was used as the dimensions of reduction. The *FindClusters()* resolution was evaluated using *clustree* (Zappia & Oshlack 2018). The default resolution of 0.8 was used when analysing the entire dataset (see Appendix C for rationale).

2.5.3 UMAP

UMAP was run using *RunUMAP()* in *Seurat*, and the default parameters were used (see Appendix D for rationale). It is used for the visualisation of cell clustering. The number of dimensions used was the selected number of principal components.

2.6 Cluster annotation

Annotations of the clusters were performed using a combination of manual annotation by inspection of differentially expressed genes by cluster and GO term enrichment by cluster.

2.6.1 DE analysis

DE is used to compare cell groups based on their gene expression. Seurat calculates differentially expressed genes for a set of cells compared to another set of cells using the Wilcoxon Rank Sum test as the default (Hoffman 2022e). The marker gene-based annotation was performed by looking at DE profiles per cluster. This was done by sub-setting the cells by cluster identity and running the *FindMarkers()* function in *Seurat* on these cells.

2.6.2 GO terms

The gene ontology (GO) term enrichment analysis was performed using the R-library gprofiler2 (Kolberg *et al.* 2020). Gprofiler2 works by interfacing the web-based GO term tool g:profiler (Raudvere *et al.* 2019). The functionality used in g:profiler is g:GOSt, which performs an enrichment analysis based on gene lists from each cluster. It returns several types of over-representation information, including GO terms, Kegg terms, biological pathways, cellular components, etc. The source of biological information is the Ensembl database. The evaluation of enrichments is performed using the cumulative hypergeometric probability (also Fisher's one-tailed test). The GO terms were created per cluster using *gost()* with the top 100 differentially expressed genes as the query. The GO search was performed against the reference database for *gallus gallus*.

Re-analysis

The major clusters were all re-analysed using the above steps on the corresponding cell subsets. This was done to investigate if any biologically significant subgroups could be found within the original clusters.

2.7 Marker evaluation

Some marker genes were controlled and evaluated due to unexpected biological signal (see Appendix E).

3 Results

The results from the analyses consist of annotated UMAP visualisations of both the whole set of cells and subsets of cells based on their annotated cell type. The data was annotated using marker genes and GO terms. The putative cell types of the main clusters can also be seen in Figure 2 (A). The expression patterns of the genes in the data are shown in Figure 3. The

central marker genes used and their corresponding cell types are shown in Table 1. The putative cell types identified by these marker genes with the addition of some additional genes described in the sections below were:

- T-cells - CD3+
 - CD8+ T-cells
 - CD4+ T cells
 - Tregs - CTLA4+ and IL2RA+ (CD25)
 - γ/δ T-cells – TARP +
 - “Cytolytic cells”- GNLY+, FASLG+, GZMA+ and GZMM+
- B-cells – Bu-1+
 - Late-stage differentiating B cells - SOX5+
 - Possible Pro B-cells
- Monocytes – MMR1L4+
- Red blood cells – HBBA+
- Thrombocytes – ITGA2B/ITGB3+
- Proliferating cells
 - Plasma cells – JCHAIN+
- Possible basophils – CD63+ HDC+

When comparing the per cell type fractions (Figure 2 (B)), it was found that they concurred with the fractions found experimentally before sequencing (see Appendix F), as well as that the fraction of cell groups per sample is similar across samples.

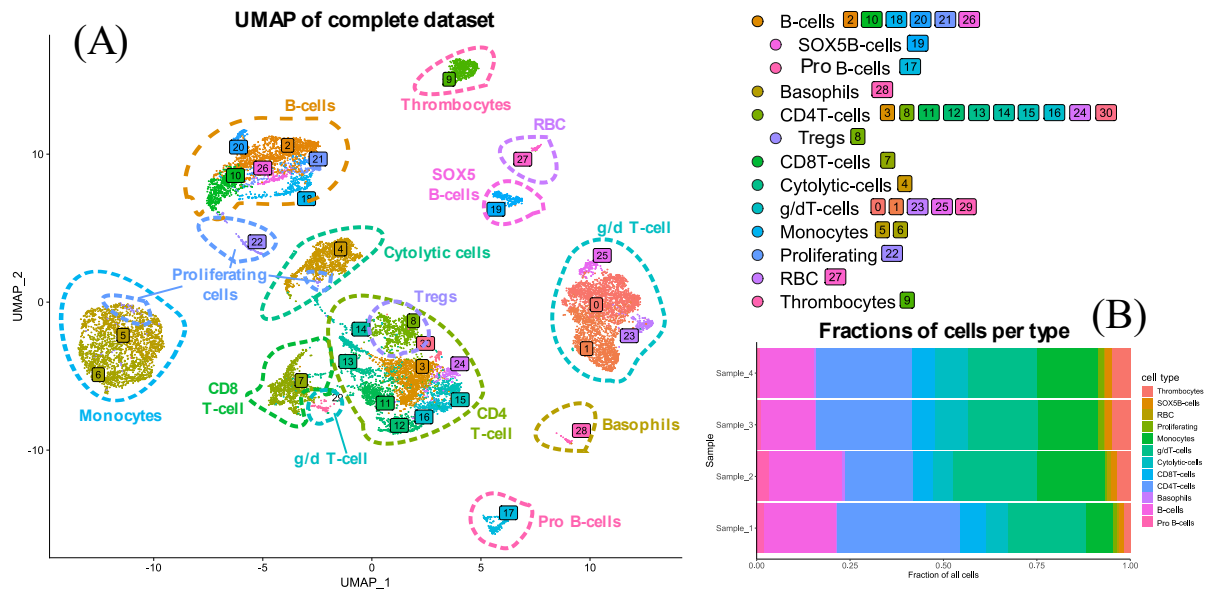


Figure 2 (A) UMAP Visualization of the 16 936 studied cells in a 2D space after clustering with putative cell type annotations for the main clusters. Cells (points) are coloured based on their cluster identity. Putative cell types have been annotated using manual annotation from marker expression and gene ontology of expressed genes. (B) Fractions of cell types per sample.

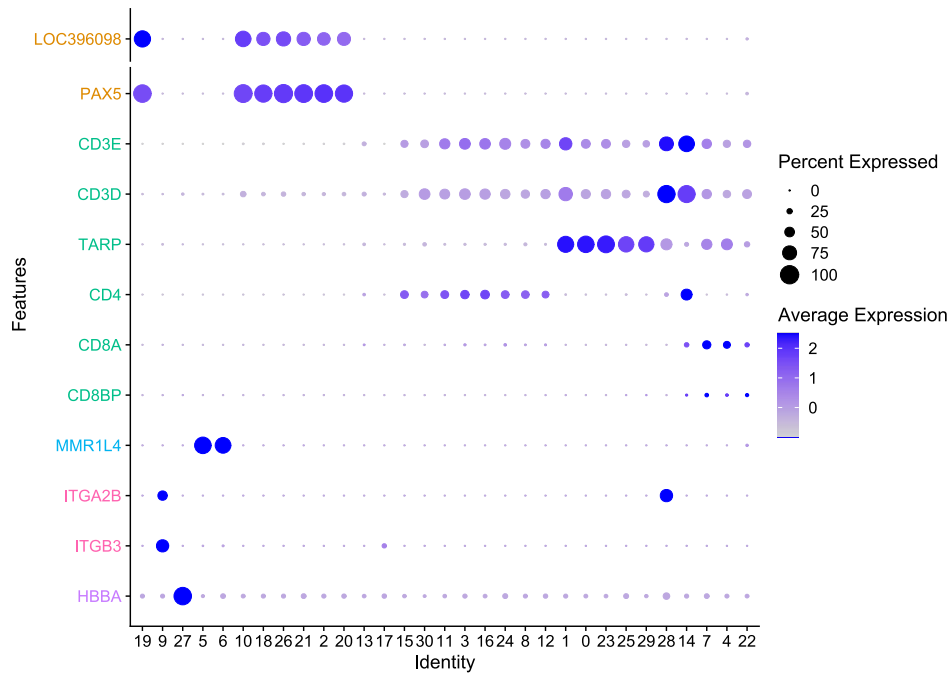


Figure 3 Dot plot of expression of marker genes based on cluster identity. The radius of the dot corresponds to percentage of cells in cluster expressing the gene, and colour intensity corresponds to scaled expression values. The colour of the gene corresponds to the cell type it annotates. Genes in orange are B-cell markers, genes in green are T-cell markers, genes in blue are monocyte markers, genes in pink are thrombocyte markers, genes in purple are red blood cell markers.

Table 1 Selected marker genes used for annotation of general cell types.

Marker gene	Name in dataset	Cell type
chB6/Bu-1	LOC396098	B-cell https://pubmed.ncbi.nlm.nih.gov/8662088/
SRY-box transcription factor 5	SOX5	B-cell (https://pubmed.ncbi.nlm.nih.gov/24945754/)
Paired box 5	PAX5	B-cell (https://pubmed.ncbi.nlm.nih.gov/34301800/)
CD3 epsilon subunit of T-cell receptor complex	CD3E	T-cell https://www.ncbi.nlm.nih.gov/gene/916
CD3 delta subunit of T-cell receptor complex	CD3D	T-cell https://www.ncbi.nlm.nih.gov/gene/916
TCR gamma alternate reading frame protein	TARP	Gamma delta T-cell
cluster of differentiation 4	CD4	T-cell (https://www.ncbi.nlm.nih.gov/gene/920)
Cluster of Differentiation 8a	CD8A	T-cell (https://www.ncbi.nlm.nih.gov/gene/925)
Cluster of Differentiation 8b pseudogene	CD8BP	T-cell https://www.ncbi.nlm.nih.gov/gene/926
macrophage mannose receptor 1-like 4	MMR1L4	Monocyte (https://www.ncbi.nlm.nih.gov/gene/4360)

integrin subunit alpha 2b / CD41	ITGA2B	Thrombocyte (https://pubmed.ncbi.nlm.nih.gov/8896225/)
integrin subunit beta 3 /CD61	ITGB3	Thrombocyte (https://pubmed.ncbi.nlm.nih.gov/8896225/)
hemoglobin subunit epsilon 1	HBBA	Red blood cell https://www.ncbi.nlm.nih.gov/gene/396485

3.1 T-cells

Clusters 0,1,23,25, 3,8,11,12,13,14,15,16,24,30, 7, 29, and 4 were annotated as T-cells based on their expression of T-cell marker genes (see Figure 3). Figure 4 (B) shows some general T-cell markers in black and pink. Among the T-cells, subgroups of cells have been identified based on the expression of different T-cell-type specific markers. The groups that have been possible to identify are:

- “Cytolytic cells” (cluster 4). A probable mixture of cytotoxic T-cells (CTLs) (TCR α/β +CD8 $\alpha\beta$ +) and other cytolytic cells, such as NK cells and cytolytic γ/δ T cells. These have been identified using genes associated with cytolytic functions, GNLY, FASLG, GZMA and GZMM, indicated in Figure 4 (B).
- γ/δ T-cells (clusters 0, 1, 23 and 25). As shown in Figure 4 (A), the putative γ/δ T-cells form their own distinct cluster separate from the other T-cells. These were identified using the TARP (TCR gamma alternate reading frame) gene and expression of the TCR δ -chain (LOC121110951) along with genes associated primarily with TCR γ/δ + cells indicated in green in Figure 4 (B).
- α/β T-cells, the α/β T-cells (TRBV6-5) consist of several types of T-cells. The subcategories of the α/β T-cells have been challenging to identify but likely include CD8 $\alpha\beta$ + cells (Cluster 7), cells expressing CD8A and CD8BP but do not express CD4, these might also include CTL that are not strongly expressing cytolytic genes.
 - CD4+ T-cells (clusters 3, 8, 11, 12, 13, 14, 15, 16, 24 and 30)
 - Treg (cluster 8), putatively assigned to this CD4+ cluster by high expression of CTLA4 and IL2RA in some of the cells in the cluster. In resting cells, these genes are primarily expressed in Tregs (Haddadi & Negahdari 2022).
 - Clusters of cells in different stages of cellular activity.

A more detailed annotation of the clusters is available in the supplementary material.

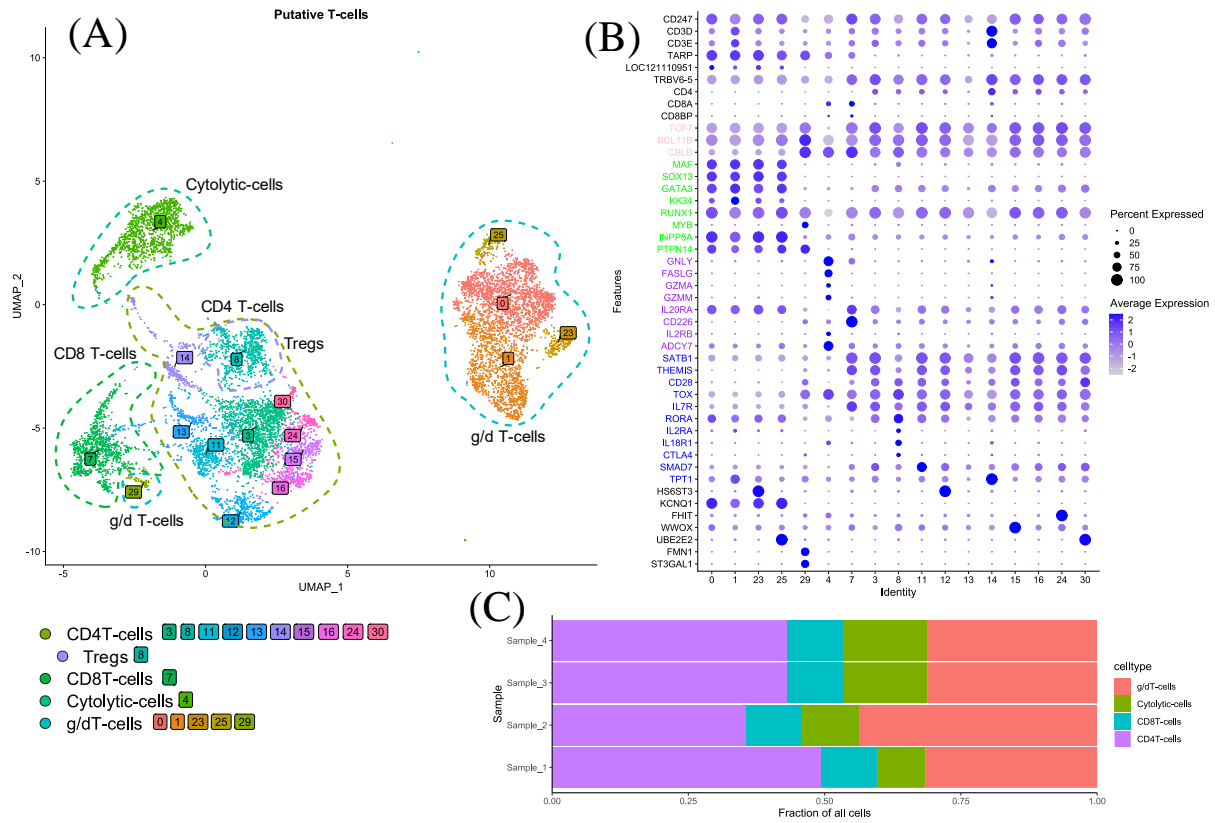


Figure 4 Subset of data corresponding to T-cells. (A) UMAP for these cells with putative cell types annotated based on the differential expression within the subcluster. (B) Expression levels of relevant genes for establishing of cell types are visualised in form of a dot plot. Expression values are scaled within the plot. The first twelve genes (black and pink) are genes used to identify T-cells. The following genes are used to identify subtypes of T-cells with genes in green primarily associated with gamma/delta T-cells, genes in purple are primarily associated with cytotoxic cells (CD8+), genes in blue are primarily associated with CD4+ T-cells, and the last seven genes in black are distinctly expressed in some clusters with functions so far not directly associated with T-cells. (C) Fractions of T-cells per sample.

Cluster 4, putative cytolytic cells, was re-clustered to investigate cell types within the cluster (Figure 5). Here, cells in sub-cluster 2 expressed high levels of TARP, indicating that it contains γ/δ T-cells. Of the genes indicative of cytolytic activity GNLY showed the highest expression in this sub-cluster. Cells in sub-cluster 1 showed a high expression of CD3D, CD3E and TRBV6-6 as well as CD8A and CD8BP, indicating that this sub-cluster contained CTL. In this sub-cluster all the studied genes indicative of cytolytic activity showed significant expression and GZMA and GZMM had the highest expression and FASLG the lowest. Cells in sub-cluster 0 showed low expression of CD3D, CD3E, TARP and TRBV6-5 indicating that this subcluster comprises low numbers of T-cells while high expression of CD247 (CD3 ζ) suggests that a majority of cells might be NK cells. NK-cells are cytolytic non-T-cells of lymphoid origin and it is known that they can have separate expression of this transmembrane part of the T-cell receptor complex (Lanier 2001). Of the studied genes indicative of cytolytic activity cells in this cluster showed a high expression of FASLG.

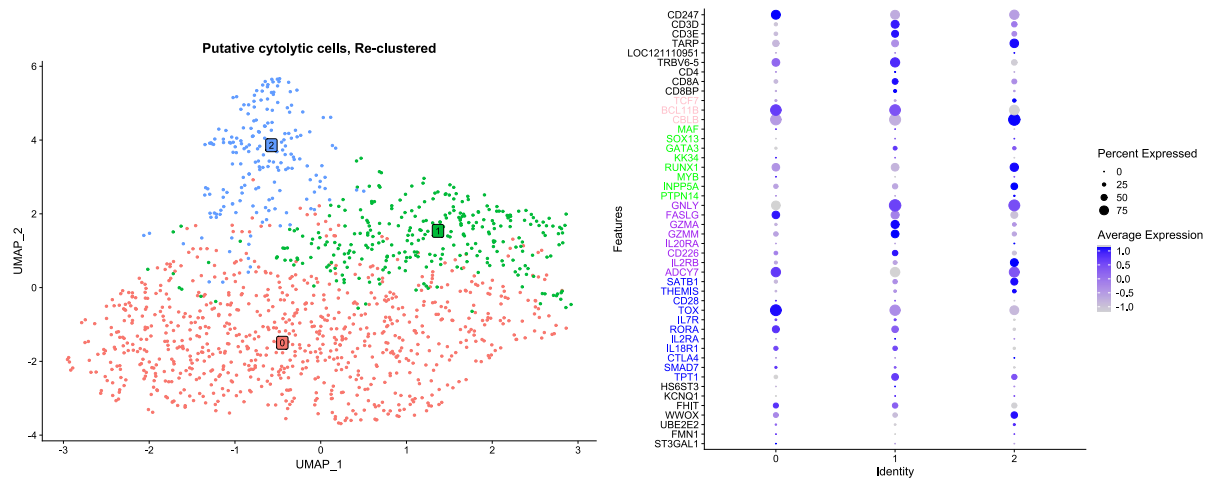


Figure 5 Subset of data corresponding to cytolytic cells. (A) UMAP for these cells after re-clustering with putative cell types annotated based on the differential expression within the subcluster. (B) Expression levels of relevant genes for establishing of cell types are visualised in form of a dot plot. Expression values are scaled within the plot. The first twelve genes (black and pink) are genes used to identify T-cells. The following genes are used to identify subtypes of T-cells with genes in green primarily associated with gamma/delta T-cells, genes in purple are primarily associated with cytotoxic cells (CD8+), genes in blue are primarily associated with CD4+ T-cells, and the last seven genes in black are distinctly expressed in some clusters with functions so far not directly associated with T-cells.

3.2 B-cells

The B-cells have been identified using the expression of Bu-1 (LOC396098). Clusters 2, 10, 18, 19, 20, 21 and 26 showed in addition to expression of Bu-1 a high expression of other typical B-cell associated genes such as the B-cell receptor genes CD79A (LOC121108878) and CD79B, immunoglobulin light chain, IGLL1, and IgA, VH26L1, as well as genes associated with B-cell receptor signaling, EVI2A (Li *et al.* 2014), VAV2 (Turner 2002) and LYN (Brian & Freedman 2021). In addition, cells in these clusters expressed B-cell transcription factors PAX5, EBF1 and TCF4, and MHCII genes BLB1, BLB2 and CD74. Thus, the B-cell identity of cells in these clusters was quite clear. Moreover, the expression pattern of known B-cell associated genes was similar between these clusters with the exception of cluster 19 and cluster 10. A striking feature of cluster 19 was a high expression of SOX5 that may indicate that cells in this cluster 19 could be B-cells in late-stage development (Rakhmanov *et al.* 2014). In addition, cells in cluster 19 showed the highest expression of Bu-1, EVI2A, IRAG2, BHLHE41, PTPRJ, MAML3 and BANK1 among the B-cell clusters. In mammals BHLHE41 (Kreslavsky *et al.* 2018) and PTPRJ (Skrzypczynska *et al.* 2016) are associated to so-called B1 B-cells and MAML3 has been associated with so-called marginal zone B-cells (Wu *et al.* 2007). Cells in cluster 10 showed the highest expression of B-cell receptor genes CD79A and CD79B, IGLL1, MHCII genes, CXCR4, BAFF (TNFSF13B), BAFF-receptor (TNFRSF13C) and HMGB1 among the B-cell clusters. This could indicate that cells in cluster 10 are activated by antigen/s.

Cluster 17 does not express typical B-cell genes, e.g. Bu-1 or CD79A/B, but was putatively identified as pro B-cells from the expression of ATP11A that in mice is expressed in

developing B-cells but not after the pro B-cell stage (Segawa *et al.* 2018) and FNIP1 that is expressed during B-cell development (Iwata *et al.* 2017). It is possible that cluster 17 contains a mixture of pro-B-cells and myeloid cells since cells share expression of some genes with clusters 5 and 6 (data not shown). The expression of these genes may also be due to the mutual activity of antigen presentation that both cell types perform. Hence, currently the annotation of cluster 17 is unsure.

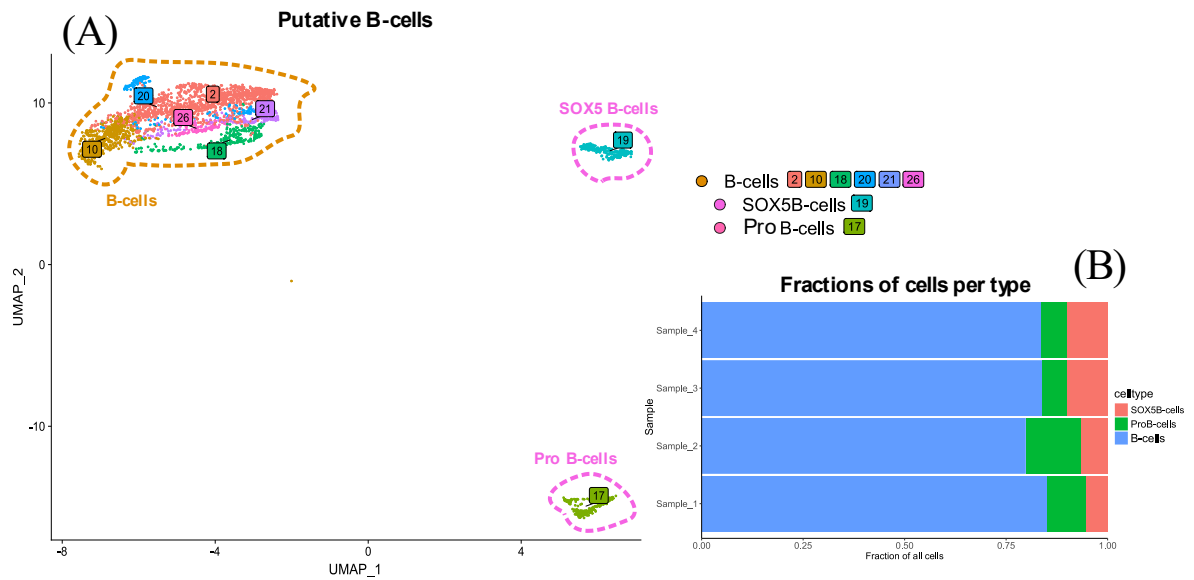


Figure 6 Putative B-cells (Bu-1+ cells). (A) UMAP for these cells with putative cell types were established based on the differential expression of Bu-1 within the data. (B) Fractions of B-cell types.

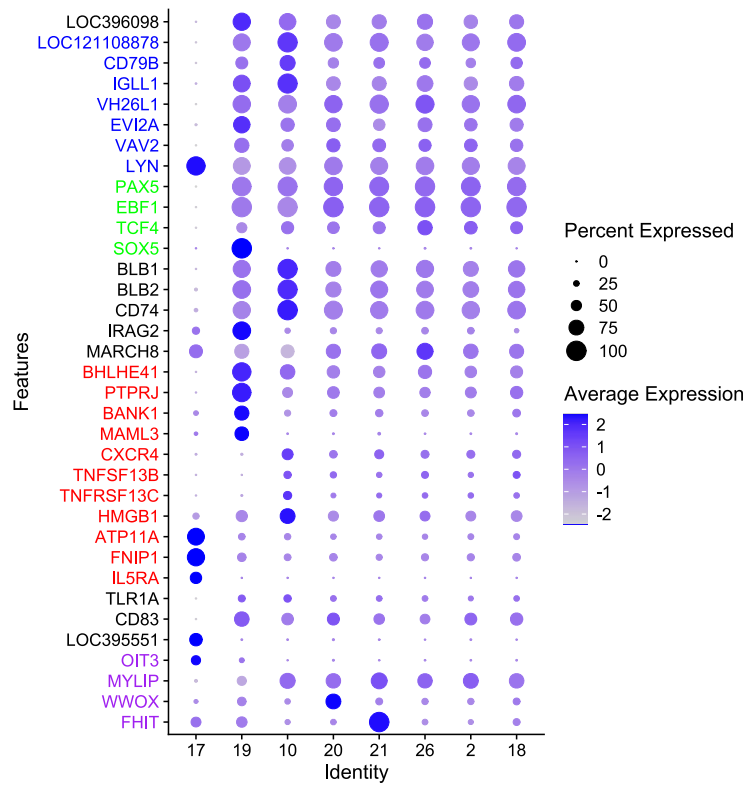


Figure 7 Dot plot of expression levels of relevant genes for establishing of cell types for clusters annotated as B-cells. Expression values are scaled within the plot. The first gene corresponds to BU-1. The blue genes are genes associated with B-cell receptor and B-cell receptor signalling. The following green genes are B-cell associated transcription factors. The black genes and MHCII genes. The red genes are general B-cell associated genes. The following black genes are general immune genes. The purple genes are differentiating for the cluster but not directly B-cell associated.

3.3 Monocytes

The distinct cluster formed by clusters 5 and 6 (Figure 2 (A)), which likely consists of monocytes, was annotated using marker gene MMR1L4 expression. From the dot plot in Figure 3, it can be seen that clusters 5 and 6 express MMR1L4. The primary clustering shows two populations, corresponding to a population with relatively high MMR1L4 expression and relatively low MMR1L4 expression when comparing clusters 5 and 6 (Figure 8). Cluster 5 expresses more elevated levels of MHCII-associated genes (CD74, BLB1, BLB2) and is likely to comprise cells with active MHCII antigen-presenting capacity. Cluster 6 also comprised some cells with high MMP9 expression. This gene is not typically associated with monocytes in the chicken and suggests this cluster also comprised other cells of putative myeloid origin, likely heterophils (Sekelova *et al.* 2017).

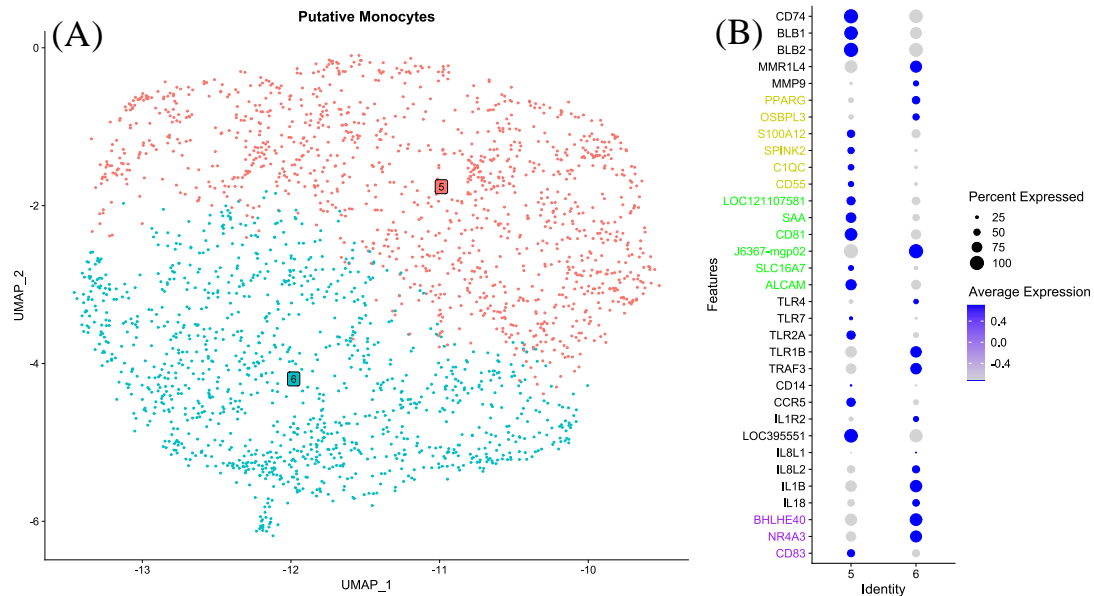


Figure 8 Subset of cells corresponding to putative monocytes (MMR1L4+ cells). UMAP for these cells with putative cell types were established based on the differential expression of xx within the subcluster (A). Expression levels of relevant genes for establishing cell types are visualised in a dot plot of monocyte and heterophil marker genes (B). The first black genes correspond to general marker genes associated with monocytes or heterophils. The following yellow genes are used to identify primarily subpopulations of monocytes with high expression of MMR1L4 and low expression of MHCII, and the green genes are used to identify monocytes with low expression of MMR1L4 and high expression of MHCII. The following black genes are markers, e.g pattern recognition receptors and cytokines, generally associated with immune sentinel cells. The purple genes are used to identify primarily heterophils.

After re-clustering the initial monocyte clusters 5 and 6, six subclusters (0-5) were indicated (Figure 9). Based on the differential expression of MMR1L4, exclusively expressed in chicken monocytes (Staines *et al.* 2014), MHCII genes and MMP9 that is expressed in chicken heterophils but not in chicken monocytes (Sekelova *et al.* 2017), it was suggested that subclusters 0, 1, 3, 4, and 5 comprised mainly monocytes and that subcluster 2 likely contained mainly heterophils. Cells in subcluster 2 also showed a high expression of BHLHE40 and NR4A3 that in mammals have been associated with neutrophils (Wang *et al.* 2022; Prince *et al.* 2017) Figure 9 (B). Moreover, the Re-clustering also revealed a polarisation of cells with high MMR1L4 expression and low/lower MHCII expression (cluster 0,1) and low MMR1L4 expression and high/higher MHCII expression (cluster 3,4,5), as seen in the original clustering. This analysis also showed that these putative monocyte subtypes differed in the expression of several genes, such as genes associated with direct antibacterial functions, S100A12 (Cunden *et al.* 2016) and SPINK2 (Dietrich *et al.* 2017), or proinflammatory responses, e.g., SAA (Rychlik *et al.* 2014) and cell activation, e.g., ALCAM (Bowen & Aruffo 1999).

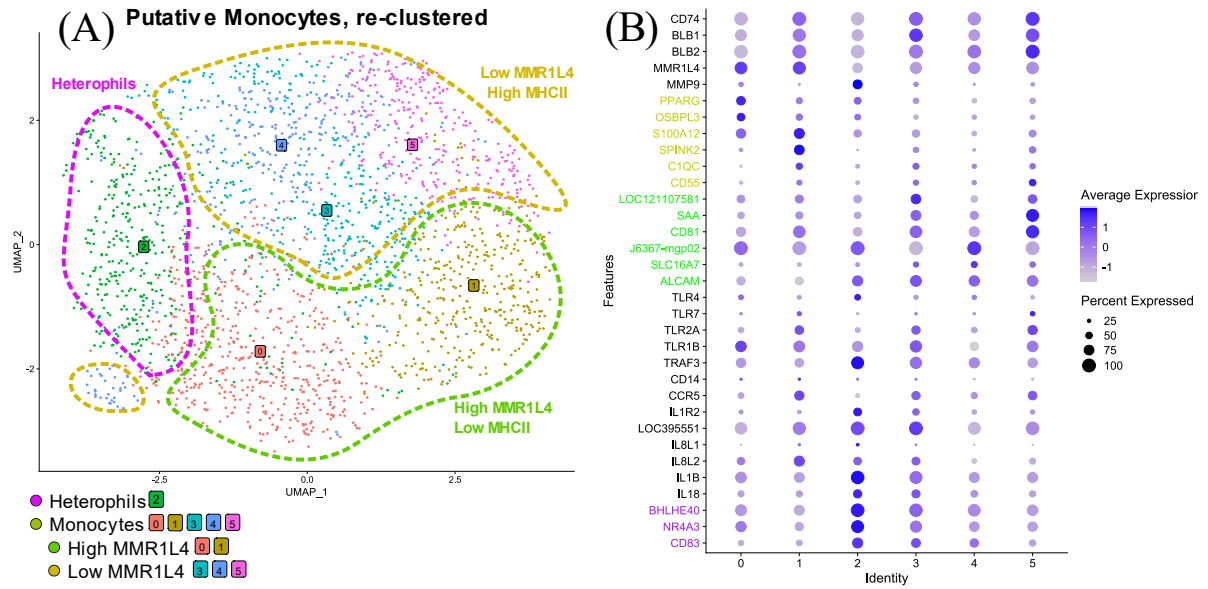


Figure 9 (A) Re-clustering of monocytes (MMR1L4+ cells) at resolution 0.8. (B) Dot plot of monocyte and heterophil marker genes. The first black genes correspond to general marker genes associated with monocytes and heterophils. The following yellow genes are used to identify primarily subpopulations of monocytes with high expression of MMR1L4 and low expression of MHCII, and the green genes are used to identify monocytes with low expression of MMR1L4 and high expression of MHCII. The following black genes are markers, e.g. pattern recognition receptors and cytokines, generally associated with immune sentinel cells. The purple genes are used to identify heterophils primarily.

3.4 Proliferating cells

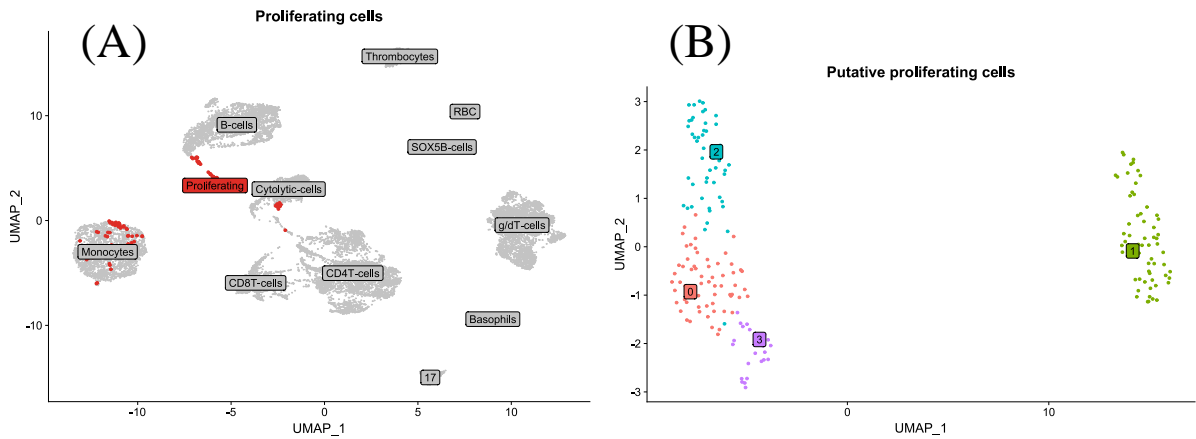


Figure 10 Subset of cells annotated as proliferating cells. (A) UMAP of all cells with proliferating cells highlighted in red. (B) Re-clustering of proliferating cells show 2 distinct groups of cells. Likely corresponding to lymphocytes (0,2,3) and myeloid cells (1).

The cells in Figure 10 have been assumed to be proliferating due to their high expression of proliferation-associated genes (cluster-specific gene expressions available in supplementary materials). The cells in this group form one cluster of cells belonging to three other primary cell types: Cytolytic cells, monocytes, and B-cells. Some of these cells also form a distinct

cluster not embedded in another cell population. When re-clustered, these cells again form 4 clusters (Figure 10 (B)).

When comparing the expression patterns of these clusters (Comparison within primary cluster 4), cluster 0 seemingly contains proliferating cells. Due to its closeness to clusters 2 and 3 and some expression of CD4 and CD28, these are likely lymphocytes. Cluster 2 expresses more T-cell-specific genes and seemingly consists of cytolytic γ/δ T-cells. Cluster 3 expresses typical B-cell genes, specifically a high expression of JCHAIN. Therefore, this cluster is believed to be plasma cells. Cluster 1 expresses monocyte markers and is probable MHCII monocytes.

When comparing the cluster's expressions with the whole dataset, more genes associated with cell proliferation are expressed. The GO terms also include terms that are ribosome, cytoskeleton, and translation associated. Therefore, it has been concluded that this cluster is a mixture of cells from different major cell groups involved in the same, probably proliferating, activity.

4 Discussion

This project aimed to investigate the possibility of using single-cell transcriptomics to infer knowledge about leukocytes in a non-traditional model organism, the chicken. Specifically, the aim was to investigate if it would be possible to 1. help identify leukocytes that are already known today and infer their functions. 2. identify unknown populations of leukocytes and infer their functions.

This study has shown that it is possible to use single-cell RNA sequencing to identify a large number of known leukocyte populations in chicken blood. We have identified populations of B-cells, T-cells, monocytes, thrombocytes, RBCs, and basophils. Among these populations, we have also identified sub-populations of cells. Some populations were more evident from their gene expressions, such as the monocytes and γ/δ T-cells, while others were difficult to distinguish, such as the different types of α/β T-cells. Clusters often did not express the expected gene patterns or genes for their putative cell types. We thought that re-clustering these subsets would help resolve sub-populations, but this was only the case for the monocytes, where the heterophils were embedded within the monocytes. This is likely due to their common origin as myeloid cells. The re-clustering did, however, not help resolve the α/β T-cells. Instead, re-clustering led to other properties of the cells taking precedence in the differential gene expression, which means that the cells were clustering on properties such as their cellular activity. This problem is inherent within this technique and is discussed further in 4.2.

Among the T-cell clusters we were able to infer some function and functional differences among the “cytolytic cells” with respect to the expression of some genes involved in the cytotoxic process (GNYL, FASLG, GZMA and GZMM). However, for most of the T-cell populations we were not able to either confirm or gain new information on cell function. Similarly as for the T-cell clusters, most identified B-cell clusters did not reveal any distinct phenotypical or immune functional differences with the exception of B-cell cluster 19. Within this cluster we identified unique or relatively higher expression of several genes associated with B-cell type/function. For instance the expression of SOX5 suggests that B-cells in cluster 19 are terminally differentiated (Rakhmanov *et al.* 2014). Moreover, cells in cluster 19 had a high expression of BHLHE41, PTPRJ and MAML3 that in mammals have been associated with B-1 B-cells (Kreslavsky *et al.* 2018; Skrzypczynska *et al.* 2016) and marginal zone B-cells (Wu *et al.* 2007). These B-cell types belong to the so-called “innate B-cells” and are involved in the primary immune response (Grasseau *et al.* 2020). To our knowledge, our results is the first indication of this type of B-cell in the chicken.

In contrast to the T and B-cell clusters, our extended analysis of monocytes revealed different phenotypic subsets of monocytes, i.e. those with high expression MMR1L4 of and low expression of MHCII compared to those with low expression of MMR1L4 and high expression of MHCII. In a study detecting cell surface expression of these two receptors by immunofluorescence labelling, two different populations of chicken spleen macrophages have previously been identified, MMR1L4^{high}MHCII^{low} and MMR1L4^{low}MHCII^{high}, respectively (Yu *et al.* 2020). Functional differences between these populations were then also identified where the MMR1L4^{high}MHCII^{low} population showed higher phagocytic capacity, higher migratory capacity, lower antigen presenting properties and lower expression of some pro-inflammatory cytokines compared to the MMR1L4^{low}MHCII^{high} population. Hence, our analysis showed that such phenotypic subpopulations also were present among chicken blood monocytes. Moreover, analysis of gene expression in the current monocyte clusters also indicated that functional differences correlating to those observed for the chicken spleen macrophage subsets were present. In addition, our analysis with more sub-clusters within the two general populations indicated that chicken monocytes might have more functional subsets. This would be in analogy with mammalian systems where e.g., for humans three major and potentially several more minor different functional subsets of circulating monocytes have been identified (Merah-Mourah *et al.* 2020). Thus, for the monocytes we have been able both to verify phenotypic and functional subtypes previously described with non-molecular methods as well as indicating further functional differences novel for the chicken.

Many populations that have been annotated have been of unclear type. For example, population 17 is believed to be pro-B cells but does not express typical B-cell genes, so this annotation is uncertain. Cluster 28 is annotated as basophils, but this cluster also expresses an unclear pattern that includes genes not expected in basophils. It is, therefore, possible that

some of the populations that were not confidentially annotated are different cell types than the ones postulated here, and some could be novel cell types.

This study shows that it is possible to study and classify immune cells without the use of laboratory methods such as flow cytometry. This is useful since flow cytometry, the leading laboratory method (Maecker *et al.* 2012), can introduce stress on the cells when they are sorted, which in turn can affect the expression profiles from these cells (van den Brink *et al.* 2017). It has also been seen that methods that rely on antibody binding can lead to a change in gene expression in the cell (Kornbluth & Hoover 1989). This means that the use of flow cytometry is not only complicated by reagent accessibility (see section 1.3) but also might be constraining the biological signal. During scRNA sequencing of cells some stress is also imposed on the cells, but generally less than with flow cytometry. This means that scRNA-seq could achieve better resolution than flow cytometry. Flow cytometry on the other hand allows for better cell population specificity, and enrichment of rare cell types in a way that is not possible with scRNA-seq (Nguyen *et al.* 2018).

4.1 Technical limitations

ScRNA-seq has opened up a world of possibilities for researchers interested in studying cellular activity. The technology does however come with limitations (Lähnemann *et al.* 2020). Unsupervised clustering often lies at the centre of these analyses. Unsupervised clustering both benefits and suffers from being unbiased towards what signal is used. This is what allows for novel biological states to be found, but it also means that discerning between real signal and noise becomes challenging. Due to the fact that each cell only expresses a small subset of all possible genes in the genome, the majority of gene counts will be zero in a single cell mRNA sequencing. This means that there is a high possibility for false signals, and technical noise. Since each cell is only sequenced once there is no technical replicate and, therefore, no good way to discern technical noise from the biological signal (Kiselev *et al.* 2019). It also becomes difficult to evaluate if the lack of gene expression in a cell is due to a real biological lack of signal of this gene or if these molecules have not been sequenced due to for example too low sequencing depth.

The computational methods used also carry limitations. The main one is a lack of good validation methods. The current best method for validating the results is biologically, by working with cells where the expected types are known (Kiselev *et al.* 2019). This validation method becomes unhelpful when working with poorly studied organisms. Another problem with the computational methods is the user-set parameters (Kiselev *et al.* 2019). In this project, the user had to set the cluster resolution and umap's n.neighbours. The decision of these parameters strongly affects the outcome of the visualisations and cluster boundaries and introduces the possibility for a large variety of outcomes. Setting these parameters becomes a trade-off question between preserving the global structure of the data for finding primary cell

types and trying to resolve subpopulations from local structure. It is difficult to evaluate if a discovered community has clustered on signal indicating true cell-type difference or not.

4.2 Biological limitations

The cluster annotations have been carried out using a heavy focus on marker genes and existing knowledge surrounding potential genetic markers for leukocytes. A problem with this method is that the knowledge of genetic markers might not be represented in the cell's gene expression. Transient cell states and cellular activity, such as proliferation, can mask the relevant immunological signal (Kiselev *et al.* 2019). This is why we found that cells sometimes clustered based on other activity rather than their cell types, such as in cluster 4, where seemingly several types of cytolytic cells formed one cluster or in cluster 22, where several types of proliferating cells formed a cluster. Many marker genes are associated with surface proteins unique to the cell but most likely have little to do with their current activity. Thus cells in transient states or involved in other activities become challenging to annotate. This means that to fully map cells based on their expression profiles, we would require more specific knowledge surrounding cellular activity in immune cells. We would also need to be able to discern between confounding cell markers and those that signify a genuine cell-type difference.

Relying on manual annotation of cell populations also means that the project outcome depends on prior knowledge regarding existing populations. The annotation also becomes time-consuming, and the reproducibility of the annotations becomes poor (Lähnemann *et al.* 2020). This is currently a considerable limitation when using single-cell transcriptomics in this way, and the development of good resources for automatic annotation is needed. The use of GO term enrichment somewhat simplifies this, as GO terms for a cluster can give information about the primary cellular processes in a cluster but does not solve the issue of prior knowledge being needed about the cell types (Kiselev *et al.* 2019).

4.3 Future work

We found that for some leukocyte populations, e.g. T-cells, the expression of lineage marker genes such as CD4, CD8 α and CD8 β was low, and in some cases we could not find expression of expected marker genes e.g. FOXP3 for identification of Treg, despite special efforts to detect these genes. Consequently, some of the analysis of data was difficult to interpret and confident identification of cells was not possible. Deeper sequencing might improve detection of lineage marker genes and warrants evaluation improve analysis.

Moreover, single cell sequencing of purified cell populations of known identity, function and activation stage could provide cell-specific gene expression profiles that could improve the identification of samples with mixed cell populations such as in the current study. With such

information the data from the current experiment may easily be re-analysed and further information might be revealed.

5 Conclusion

This study was aimed to serve as a first investigation into the possible usage of single-cell genomics in avian immunology and as a basis for continued research into the chicken immune system and possible furthering of poultry medicine. It has been shown here that sequencing-based characterisation of immune cells is possible using currently available methods. It could prove a robust option in immunology study cases where traditional methods are limited by mapping the immune system using single-cell data and gaining novel insight into the type and function of immune cells.

6 Ethical permits

This study has involved animal blood collection, which must be performed in a way that minimises animal stress and pain. The blood sample collection was performed using existing ethical permits at SVA. Permit: Dnr 5.818-14692/2020.

7 Acknowledgements

I would like to thank my supervisor Robert Söderlund for his continued help and support throughout this project. I also want to thank researcher Eva Wattrang for helping with the interpretations of the results and providing the immunological background and expertise needed to complete the project, as well as Sonja Härtle (Ludwig-Maximilians-Universität München) for helping with the B-cell interpretation. I would also like to thank my subject reader Staffan Svärd, my examiner Pascal Milesi, and the course coordinator Lena Henriksson for dedicating their time to this coursework.

Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Computation and data handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The funding for this project was provided by The Royal Swedish Agricultural Academy and the Magnus Bergvall foundation.

8 Supplementary material

Link to data and scripts: <https://github.com/maxwelma/exjobb>

References

10X genomics. 2022a. Chromium Next GEM Single Cell 3' Reagent Kits v3.1(Dual Index) User Guide.

10X genomics. 2021a. Inside Chromium Next GEM Technology - 10xgenomics - PDF Catalogs | Technical Documentation. WWW-dokument 2021-: https://pdf.medicaexpo.com/pdf/10xgenomics/inside-chromium-next-gem-technology/128227-245904-_2.html. Hämtad 2022-12-16.

10X genomics. 2020a. Training Module - Chromium Single Cell 3' v3.1 - Dual Index. WWW-dokument 2020-: <https://pages.10xgenomics.com/sup-training-single-cell-gene-expression-dual-index.html>. Hämtad 2022-12-16.

10X genomics. 2020b. Technical Note – Sequencing Metrics & Base Composition of Single Cell 3' Dual Index Libraries • Rev A.

10X genomics. 2021b. Sequencing - Official 10x Genomics Support. WWW-dokument 2021-08-10: <https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/sequencing-requirements-for-single-cell-3>. Hämtad 2022-12-16.

10X genomics. 2020c. What is Cell Ranger? -Software -Single Cell Gene Expression - Official 10x Genomics Support. WWW-dokument 2020-: <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger>. Hämtad 2022-09-08.

10X genomics. 2018a. What is a template switch oligo (TSO)? WWW-dokument 2018-: <https://kb.10xgenomics.com/hc/en-us/articles/360001493051-What-is-a-template-switch-oligo-TSO->. Hämtad 2023-01-10.

10X genomics. 2022b. Recommendation on Including Introns for Gene Expression Analysis. WWW-dokument 2022-: <https://support.10xgenomics.com/docs/intron-mode-rec>. Hämtad 2022-12-16.

10X genomics. 2020d. Gene Expression Algorithms Overview. WWW-dokument 2020-: <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview#alignment>. Hämtad 2022-12-16.

10X genomics. 2018b. What is the difference between the filtered and raw gene-barcode matrix? WWW-dokument 2018-: <https://kb.10xgenomics.com/hc/en-us/articles/360001892491-What-is-the-difference-between-the-filtered-and-raw-gene-barcode-matrix->. Hämtad 2022-12-16.

10X genomics. 2020e. Cell Ranger count.

Adan A, Alizada G, Kiraz Y, Baran Y, Nalbant A. 2017. Flow cytometry: basic principles and applications. *Critical Reviews in Biotechnology* 37: 163–176.

Apache Software Foundation. 2022. ANN (Approximate Nearest Neighbor) | Ignite Documentation. WWW-dokument 2022-: <https://ignite.apache.org/docs/latest/machine-learning/binary-classification/ann>. Hämtad 2022-12-16.

Astill J, Wood RD, Sharif S. 2022. Chapter 8.3 - Thrombocyte functions in the avian immune system. I: Kaspers B, Schat KA, Göbel TW, Vervelde L (red.). *Avian Immunology* (Third Edition), s. 205–212. Academic Press, Boston.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008.

Bowen MA, Aruffo A. 1999. Adhesion molecules, their receptors, and their regulation: analysis of CD6-activated leukocyte cell adhesion molecule (ALCAM/CD166) interactions. *Transplantation Proceedings* 31: 795–796.

Brian BF, Freedman TS. 2021. The Src-family Kinase Lyn in Immunoreceptor Signaling. *Endocrinology* 162: bqab152.

Burkhardt NB, Elleder D, Schusser B, Krchlíková V, Göbel TW, Härtle S, Kaspers B. 2022. The Discovery of Chicken Foxp3 Demands Redefinition of Avian Regulatory T Cells. *The Journal of Immunology* 208: 1128–1138.

Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. 2001. T Cell-Mediated Immunity. *Immunobiology: The Immune System in Health and Disease*. 5th edition

Conway A. 2020. Latest poultry, egg market forecasts available in 2020 WATT executive guide to world poultry trends. WATT Global Media, Rockford

Cruse JM, Lewis RE, Wang H (red). 2004. 7 - ANTIGEN PRESENTATION. *Immunology Guidebook*, s. 267–276. Academic Press, San Diego.

Cunden LS, Gaillard A, Nolan EM. 2016. Calcium ions tune the zinc-sequestering properties and antimicrobial activity of human S100A12. *Chemical Science* 7: 1338–1348.

Dietrich MA, Słowińska M, Karol H, Adamek M, Steinhagen D, Hejmej A, Bilińska B, Ciereszko A. 2017. Serine protease inhibitor Kazal-type 2 is expressed in the male reproductive tract of carp with a possible role in antimicrobial protection. *Fish & Shellfish Immunology* 60: 150–163.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.

Dodge Y. 2008. *The Concise Encyclopedia of Statistics*.

Eberwine J, Sul J-Y, Bartfai T, Kim J. 2014. The promise of single-cell sequencing. *Nature Methods* 11: 25–27.

Efremova M, Teichmann SA. 2020. Computational methods for single-cell omics across modalities. *Nature Methods* 17: 14–17.

Ford C. 2017. The Wilcoxon Rank Sum Test | University of Virginia Library Research Data Services + Sciences. WWW-dokument 2017-: <https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/>. Hämtad 2023-01-12.

Fortunato S, Barthélemy M. 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 104: 36–41.

Frejd. 2022. About us. WWW-dokument 2022-07-15: <https://www.scilifelab.se/about-us/>. Hämtad 2022-12-16.

Frutiger S, Hughes GJ, Paquet N, Luethy R, Jaton JC. 1992. Disulfide bond assignment in human J chain and its covalent pairing with immunoglobulin M. *Biochemistry* 31: 12643–12647.

Gayoso A, Shor J, Carr AJ, Sharma R, Pe’er D. 2019. GitHub: DoubletDetection. doi 10.5281/zenodo.2678042.

Genovese KJ, He H, Swaggerty CL, Kogut MH. 2013. The avian heterophil. *Developmental & Comparative Immunology* 41: 334–340.

Gilmour DG, Brand A, Donnelly N, Stone HA. 1976. Bu-1 and Th-1, two loci determining surface antigens of B or T lymphocytes in the chicken. *Immunogenetics* 3: 549–563.

Glisson JR, McDougald LR, Nolan LK, Suarez DL, Nair VL. 2013. *Diseases of Poultry*. John Wiley & Sons

Grasseau A, Boudigou M, Le Pottier L, Chriti N, Cornec D, Pers J-O, Renaudineau Y, Hillion S. 2020. Innate B Cells: the Archetype of Protective Immune Cells. *Clinical Reviews in Allergy & Immunology* 58: 92–106.

Haddadi MH, Negahdari B. 2022. Clinical and diagnostic potential of regulatory T cell markers: From bench to bedside. *Transplant Immunology* 70: 101518.

Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* 20: 296.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. 2021. Integrated analysis of multimodal single-cell data. *Cell* 184: 3573-3587.e29.

Hoffman P. 2022a. Introduction to scRNA-seq integration. WWW-dokument 2022-: https://satijalab.org/seurat/articles/integration_introduction.html#performing-integration-on-datasets-normalized-with-sctransform-1. Hämtad 2023-01-13.

Hoffman P. 2022b. (Shared) Nearest-neighbor graph construction — FindNeighbors. WWW-dokument 2022-: <https://satijalab.org/seurat/reference/findneighbors>. Hämtad 2022-12-16.

Hoffman P. 2022c. Cluster Determination — FindClusters. WWW-dokument 2022-: <https://satijalab.org/seurat/reference/findclusters>. Hämtad 2022-12-16.

Hoffman P. 2022d. Seurat - Guided Clustering Tutorial. WWW-dokument 2022-: https://satijalab.org/seurat/articles/pbm3k_tutorial.html#cluster-the-cells-1. Hämtad 2023-01-13.

Hoffman P. 2022e. Gene expression markers of identity classes — FindMarkers. WWW-dokument 2022-: <https://satijalab.org/seurat/reference/findmarkers>. Hämtad 2022-12-16.

IBM. 2022. What is the k-nearest neighbors algorithm? | IBM. WWW-dokument 2022-: <https://www.ibm.com/se-en/topics/knn>. Hämtad 2022-12-16.

Illumina. 2022a. Welcome to immense discovery power. WWW-dokument 2022-: <https://www.illumina.com/systems/sequencing-platforms/novaseq.html>. Hämtad 2022-12-16.

Illumina. 2022b. Unique Molecular Identifiers (UMIs) | For sequencing accuracy. WWW-dokument 2022-: <https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-molecular-identifiers.html>. Hämtad 2022-12-16.

Iwata TN, Ramírez-Komo JA, Park H, Iritani BM. 2017. Control of B lymphocyte development and functions by the mTOR signaling pathways. *Cytokine & Growth Factor Reviews* 35: 47–62.

Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments | *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. WWW-dokument 2016-04-13: <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>. Hämtad 2022-12-16.

- Kaiser P, Balic A. 2015. Chapter 17 - The Avian Immune System. I: Scanes CG (red.). Sturkie's Avian Physiology (Sixth Edition), s. 403–418. Academic Press, San Diego.
- Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* 20: 273–282.
- Kogut MH. 2022. Subchapter 8.2 - Avian granulocytes. I: Kaspers B, Schat KA, Göbel TW, Vervelde L (red.). *Avian Immunology* (Third Edition), s. 197–203. Academic Press, Boston.
- Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. 2020. gprofiler2—an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* 9 (ELIXIR):
- Kornbluth J, Hoover RG. 1989. Anti-HLA Class I Antibodies Alter Gene Expression in Human Natural Killer Cells. I: Dupont B (red.). *Immunobiology of HLA: Volume II: Immunogenetics and Histocompatibility*, s. 150–152. Springer, Berlin, Heidelberg.
- Kreslavsky T, Wong JB, Fischer M, Skok JA, Busslinger M. 2018. Control of B-1a cell development by instructive BCR signaling. *Current Opinion in Immunology* 51: 24–31.
- Kumari S, Maurya S, Goyal P, Balasubramaniam SS, Goyal N. 2016. Scalable Parallel Algorithms for Shared Nearest Neighbor Clustering. 2016 IEEE 23rd International Conference on High Performance Computing (HiPC), s. 72–81.
- Lacoste-Eleau A-S, Bleux C, Quéré P, Coudert F, Corbel C, Kanellopoulos-Langevin C. 1994. Biochemical and Functional Characterization of an Avian Homolog of the Integrin GPIIb-IIIa Present on Chicken Thrombocytes. *Experimental Cell Research* 213: 198–209.
- Lancichinetti A, Fortunato S. 2011. Limits of modularity maximization in community detection. *Physical Review E* 84: 066122.
- Lanier LL. 2001. On guard—activating NK cell receptors. *Nature Immunology* 2: 23–27.
- Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. 2015. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162: 184–197.
- Li X-W, Rees JS, Xue P, Zhang H, Hamaia SW, Sanderson B, Funk PE, Farndale RW, Lilley KS, Perrett S, Jackson AP. 2014. New Insights into the DT40 B Cell Receptor Cluster Using a Proteomic Proximity Labeling Assay. *Journal of Biological Chemistry* 289: 14434–14447.
- Lu H, Halappanavar M, Kalyanaraman A. 2014. Parallel Heuristics for Scalable Community Detection.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, doi 10.15252/msb.20188746.

Lundberg M. 2022. About Uppsala Multidisciplinary Center for Advanced Computational Science - Uppsala Multidisciplinary Center for Advanced Computational Science - Uppsala University, Sweden. WWW-dokument 2022-05-25: <https://www.uppmax.uu.se/about-us/>. Hämtad 2022-12-16.

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, Pinello L, Skums P, Stamatakis A, Attolini CS-O, Aparicio S, Baaijens J, Balvert M, Barbanson B de, Cappuccio A, Corleone G, Dutilh BE, Florescu M, Guryev V, Holmer R, Jahn K, Lobo TJ, Keizer EM, Khatri I, Kielbasa SM, Korbel JO, Kozlov AM, Kuo T-H, Lelieveldt BPF, Mandoiu II, Marioni JC, Marschall T, Mölder F, Niknejad A, Raczkowski L, Reinders M, Ridder J de, Saliba A-E, Somarakis A, Stegle O, Theis FJ, Yang H, Zelikovsky A, McHardy AC, Raphael BJ, Shah SP, Schönhuth A. 2020. Eleven grand challenges in single-cell data science. *Genome Biology* 21: 31.

Maecker HT, McCoy JP, Nussenblatt R. 2012. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology* 12: 191–200.

McInnes L, Healy J, Melville J. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Meijerink N, van Haarlem DA, Velkers FC, Stegeman AJ, Rutten VPMG, Jansen CA. 2021. Analysis of chicken intestinal natural killer cells, a major IEL subset during embryonic and early life. *Developmental & Comparative Immunology* 114: 103857.

Merah-Mourah F, Cohen SO, Charron D, Mooney N, Haziot A. 2020. Identification of Novel Human Monocyte Subsets and Evidence for Phenotypic Groups Defined by Interindividual Variations of Expression of Adhesion Molecules. *Scientific Reports* 10: 4397.

NCBI. 2022a. *Gallus gallus* isolate:bGalGal1 (ID 660757) - BioProject - NCBI. WWW-dokument 2022-: <https://www.ncbi.nlm.nih.gov/bioproject/660757>. Hämtad 2023-01-09.

NCBI. 2022b. *Gallus gallus* Annotation Report. WWW-dokument 2022-: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Gallus_gallus/106/#BuildInfo. Hämtad 2023-01-09.

Newman MEJ. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103: 8577–8582.

Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K. 2018. Experimental Considerations for Single-Cell RNA Sequencing Approaches. *Frontiers in Cell and Developmental Biology* 6: 108.

NHGRI. 2022. Gene Expression. WWW-dokument 2022-12-12:

<https://www.genome.gov/genetics-glossary/Gene-Expression>. Hämtad 2022-12-16.

Nutt SL, Fairfax KA, Kallies A. 2007. BLIMP1 guides the fate of effector B and T cells. *Nature Reviews Immunology* 7: 923–927.

Osika A. 2022. *spotify/annoy*.

Patel RS, Tomlinson JE, Divers TJ, Van de Walle GR, Rosenberg BR. 2021. Single-cell resolution landscape of equine peripheral blood mononuclear cells reveals diverse cell types including T-bet⁺ B cells. *BMC Biology* 19: 13.

Prince LR, Prosseda SD, Higgins K, Carlring J, Prestwich EC, Ogryzko NV, Rahman A, Basran A, Falciani F, Taylor P, Renshaw SA, Whyte MKB, Sabroe I. 2017. NR4A orphan nuclear receptor family members, NR4A2 and NR4A3, regulate neutrophil number and survival. *Blood* 130: 1014–1025.

Rakhmanov M, Sic H, Kienzler A-K, Fischer B, Rizzi M, Seidl M, Melkaoui K, Unger S, Moehle L, Schmit NE, Deshmukh SD, Ayata CK, Schuh W, Zhang Z, Cosset F-L, Verhoeyen E, Peter H-H, Voll RE, Salzer U, Eibel H, Warnatz K. 2014. High levels of SOX5 decrease proliferative capacity of human B cells, but permit plasmablast differentiation. *PloS One* 9: e100328.

Ratcliffe MJH. 2006. Antibodies, immunoglobulin genes and the bursa of Fabricius in chicken B cell development. *Developmental & Comparative Immunology* 30: 101–118.

Ratcliffe MJH, Härtle S. 2022. Chapter 4 - B cells, the bursa of Fabricius, and the generation of antibody repertoires. I: Kaspers B, Schat KA, Göbel TW, Vervelde L (red.). *Avian Immunology (Third Edition)*, s. 71–99. Academic Press, Boston.

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* 47: W191–W198.

Rychlik I, Elsheimer-Matulova M, Kyrova K. 2014. Gene expression in the chicken caecum in response to infections with non-typhoid *Salmonella*. *Veterinary Research* 45: 119.

Scanes CG. 2016. Biology of stress in poultry with emphasis on glucocorticoids and the heterophil to lymphocyte ratio. *Poultry Science* 95: 2208–2215.

Segawa K, Yanagihashi Y, Yamada K, Suzuki C, Uchiyama Y, Nagata S. 2018. Phospholipid flippases enable precursor B cells to flee engulfment by macrophages. *Proceedings of the National Academy of Sciences* 115: 12212–12217.

Sekelova Z, Stepanova H, Polansky O, Varmuzova K, Faldynova M, Fedr R, Rychlik I, Vlasatikova L. 2017. Differential protein expression in chicken macrophages and heterophils in vivo following infection with *Salmonella Enteritidis*. *Veterinary Research* 48: 35.

Shanmugasundaram R, Selvaraj RK. 2011. Regulatory T Cell Properties of Chicken CD4⁺CD25⁺ Cells. *The Journal of Immunology* 186: 1997–2002.

Skrzypczynska KM, Zhu JW, Weiss A. 2016. Positive Regulation of Lyn Kinase by CD148 Is Required for B Cell Receptor Signaling in B1 but Not B2 B Cells. *Immunity* 45: 1232–1244.

Smith AL, Göbel TW. 2022. Chapter 6 - Avian T cells: Antigen Recognition and Lineages. I: Kaspers B, Schat KA, Göbel TW, Vervelde L (red.). *Avian Immunology (Third Edition)*, s. 121–134. Academic Press, Boston.

Sompayrac LM. 2015. *How the Immune System Works*. John Wiley & Sons, Incorporated, Newark, UNITED STATES.

Staines K, Hunt LG, Young JR, Butter C. 2014. Evolution of an Expanded Mannose Receptor Gene Family. *PLoS ONE* 9: e110330.

Straub C, Neulen M-L, Sperling B, Windau K, Zechmann M, Jansen CA, Viertlboeck BC, Göbel TW. 2013. Chicken NK cell receptors. *Developmental and Comparative Immunology* 41: 324–333.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* 177: 1888–1902.e21.

Sutton K, Balic A, Kaspers B, Vervelde L. 2022. Chapter 8.1 - Macrophages and dendritic cells. I: Kaspers B, Schat KA, Göbel TW, Vervelde L (red.). *Avian Immunology (Third Edition)*, s. 167–195. Academic Press, Boston.

Turner M. 2002. B-cell development and antigen receptor signalling. *Biochemical Society Transactions* 30: 812–815.

van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, Robin C, van Oudenaarden A. 2017. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods* 14: 935–936.

Varadé J, Magadán S, González-Fernández Á. 2021. Human immunology and immunotherapy: main achievements and challenges. *Cellular & Molecular Immunology* 18: 805–828.

Wang L, Liu Y, Dai Y, Tang X, Yin T, Wang C, Wang T, Dong L, Shi M, Qin J, Xue M, Cao Y, Liu J, Liu P, Huang J, Wen C, Zhang J, Xu Z, Bai F, Deng X, Peng C, Chen H, Jiang L,

Chen S, Shen B. 2022. Single-cell RNA-seq analysis reveals BHLHE40-driven pro-tumour neutrophils with hyperactivated glycolysis in pancreatic tumour microenvironment. *Gut* gutjnl-2021-326070.

Wattrang E, Eriksson H, Jinnerot T, Persson M, Bagge E, Söderlund R, Naghizadeh M, Dalgaard TS. 2020. Immune responses upon experimental *Erysipelothrix rhusiopathiae* infection of naïve and vaccinated chickens. *Veterinary Research* 51: 114.

Wattrang E, Thebo P, Ibrahim O, Dalgaard TS, Lundén A. 2019. Parasite-specific proliferative responses of chicken spleen cells upon *in vitro* stimulation with *Eimeria tenella* antigen. *Parasitology* 146: 625–633.

Whitley E, Ball J. 2002. Statistics review 6: Nonparametric methods. *Critical Care* 6: 509–513.

Wu L, Maillard I, Nakamura M, Pear WS, Griffin JD. 2007. The transcriptional coactivator Maml1 is required for Notch2-mediated marginal zone B-cell development. *Blood* 110: 3618–3623.

Yu K, Gu MJ, Pyung YJ, Song K-D, Park TS, Han SH, Yun C-H. 2020. Characterization of splenic MRC1^{hi}MHCII^{lo} and MRC1^{lo}MHCII^{hi} cells from the monocyte/macrophage lineage of White Leghorn chickens. *Veterinary Research* 51: 73.

Zappia L. 2020. 2.1 The data.

Zappia L, Oshlack A. 2018. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* 7: giy083.

Appendix A Quality metrics

Before analysis the entire dataset consisting of 18 483 cells was filtered according to the steps in heading 0. The remaining 16 936 cells after filtering were analysed according to the steps in heading 2.5.

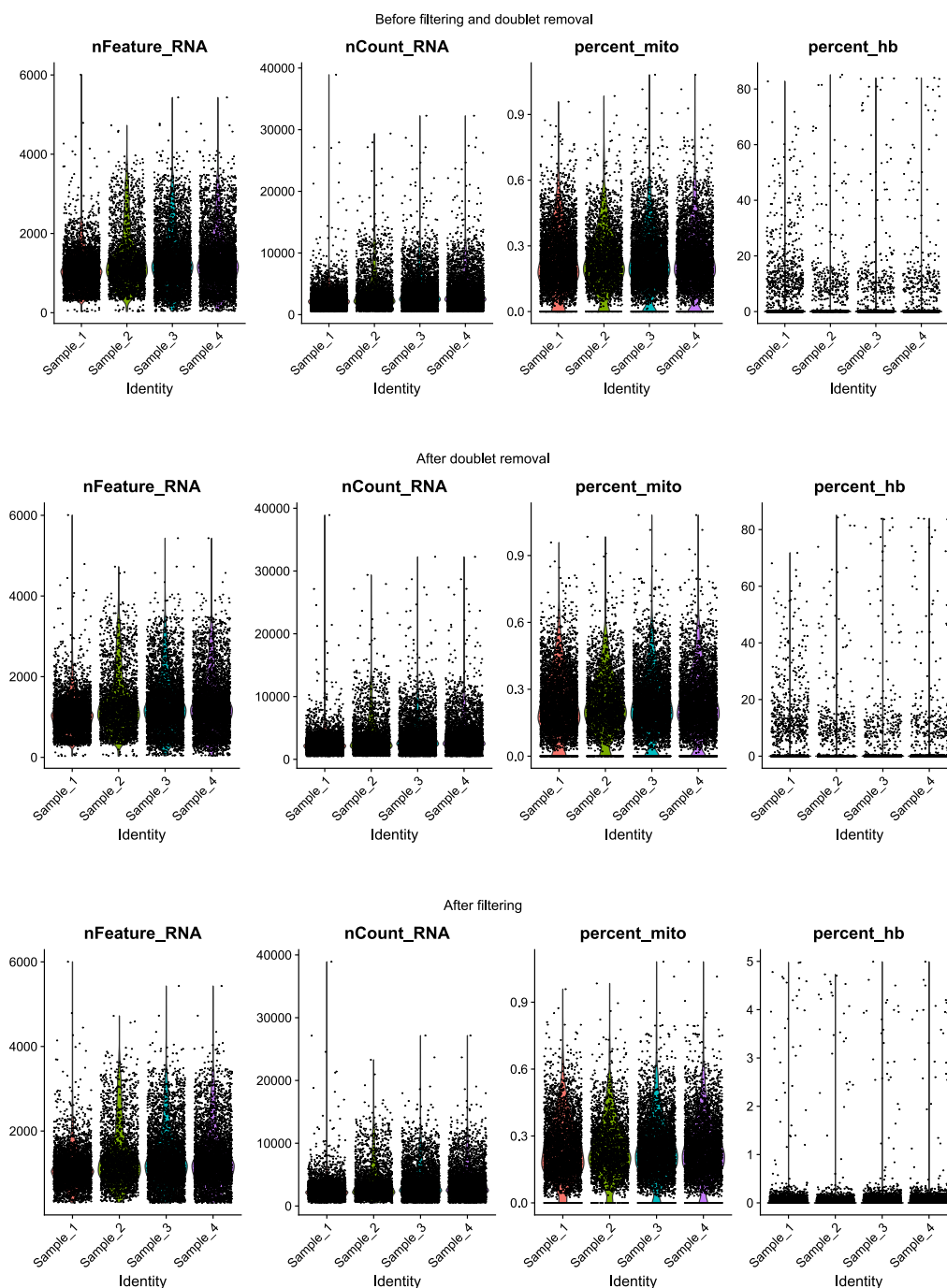


Figure 1 Violoin plots of considered parameters, from left to right, feature count/cell, read count/cell, percentage of mitochondrial genes/cell, percentage of red blood cell genes/cell, respectively. From top to bottom, raw data, after doublet removal and after filtering.

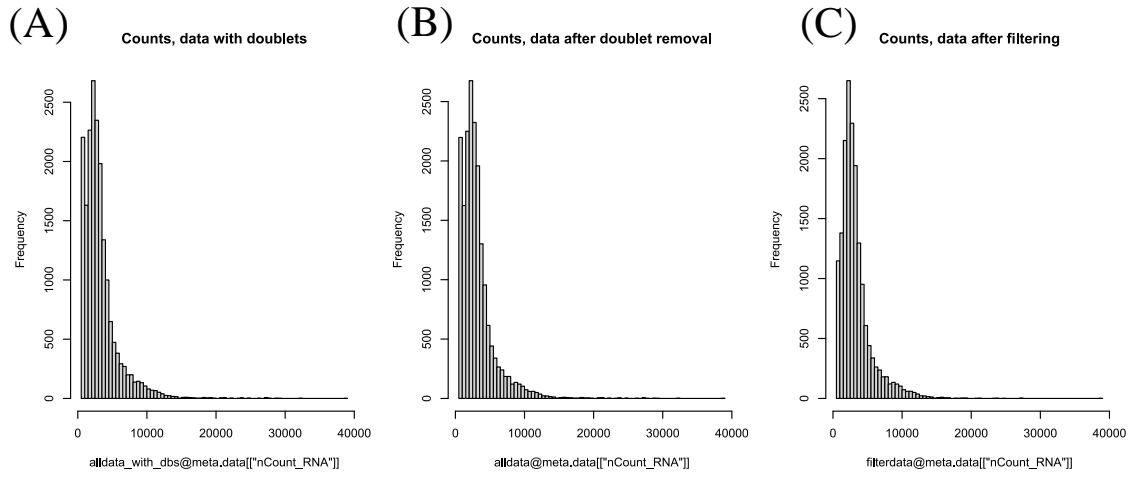


Figure 2 Histogram of read count per cell for whole dataset (A) before doublet removal, (B) before filtering and (C) after filtering.

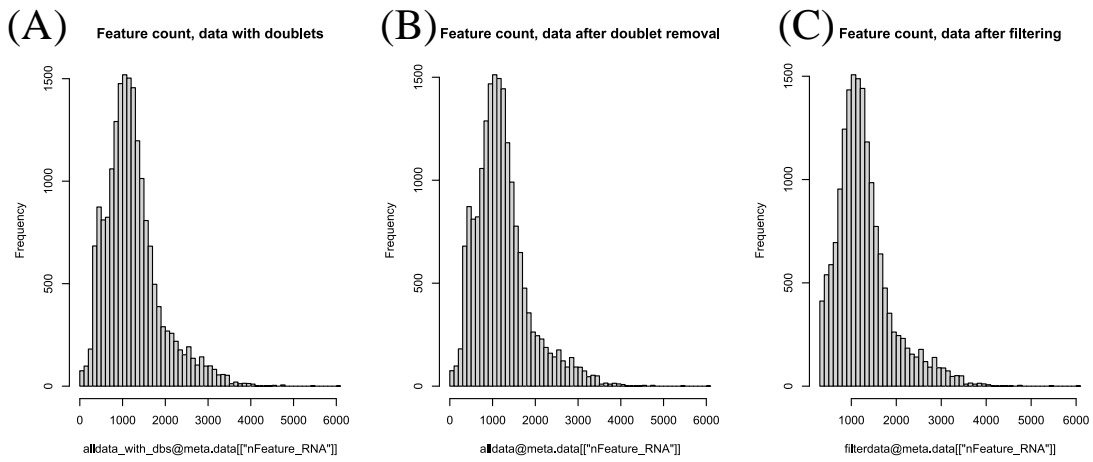


Figure 3 Histogram of feature count per cell for whole dataset (A) before doublet removal, (B) before filtering and (C) after filtering.

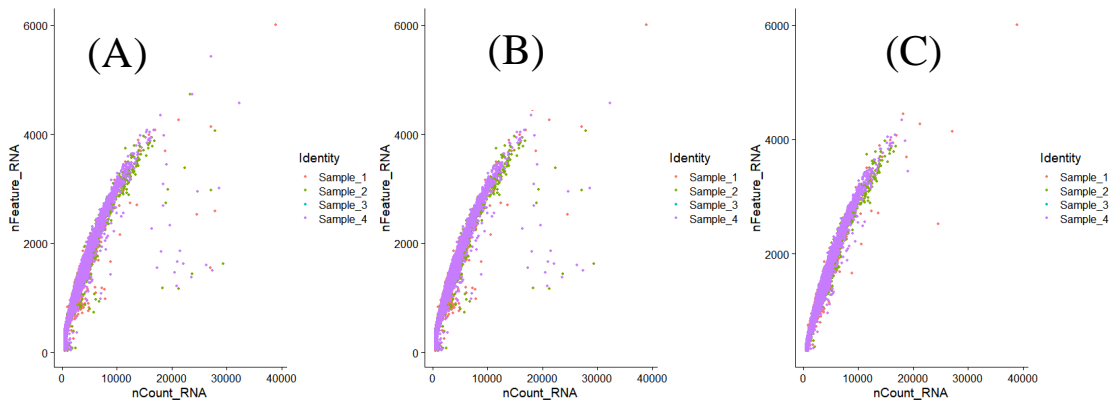


Figure 4 Scatterplot of number of features per cell vs number of reads per cell for whole dataset (A) before doublet removal, (B) before filtering and (C) after filtering.

Appendix B Comparison of clustering algorithms

The results of the first 3 of the clustering algorithms were compared, all other settings were set to the same values. There is no significant difference between the results of the standard Louvain algorithm and the Louvain algorithm multilevel refinement and SLM algorithm, except for the definition of more small clusters in the second. Since these internal cluster borders are not used at this stage but rather the major clusters are re-clustered, the definition of the sub-clusters is not essential to the analysis. Therefore, we chose to work with the standard Louvain algorithm when defining clusters since it is fast and shown to perform well for large datasets.

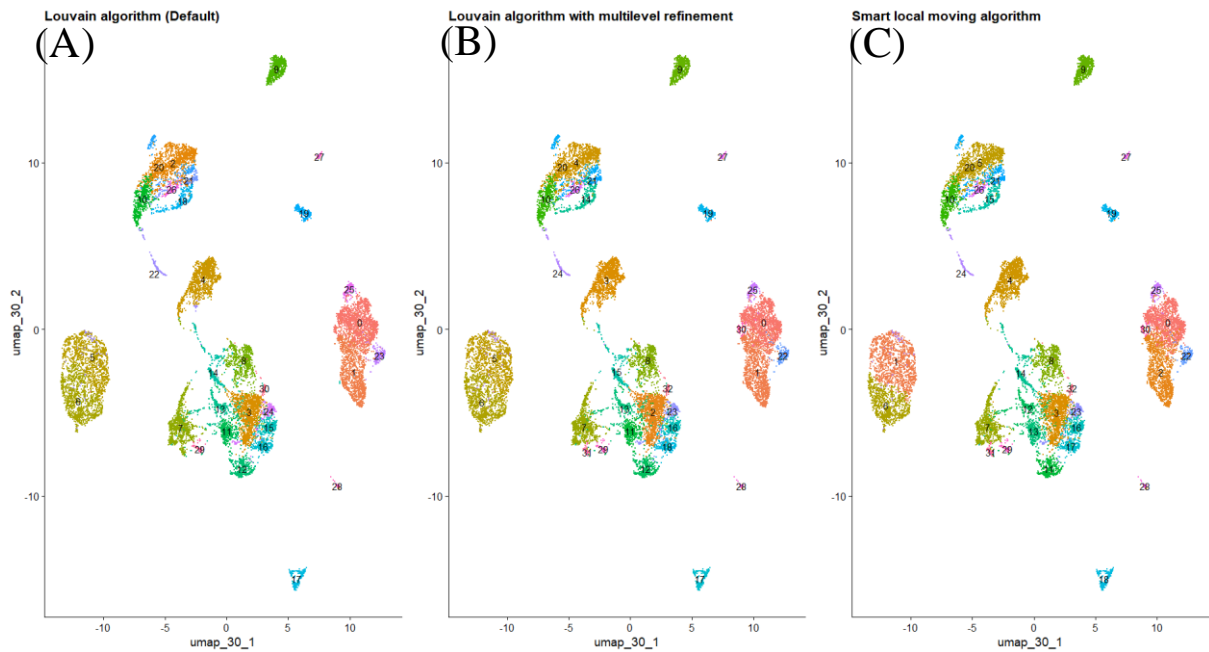


Figure 5 Cluster determination of the data using the standard Louvain algorithm (A) and the Louvain algorithm multilevel refinement (B) and the SLM algorithm.

Appendix C Resolution comparison

The cluster assignment is based on the resolution of the FindClusters() function. To avoid over clustering and compare differences in outcome based on resolution, *Clustree* is run (Zappia & Oshlack 2018). *Clustree* generates graphs of how samples move between clusters depending on resolution. The investigated resolutions lie between 0-1 and are evaluated using the stability index. This corresponds to the number of times a solution appears when running the method 100 times.

As seen in Figure 6 there is an increase in stability between resolutions 0-0.1 and incremental increases between the following resolution steps. Further increases in resolution also lead to more cells moving from stable to unstable states. Indicating that higher resolution is splitting up sub-graphs into smaller modules.

In order to not loose biological information and the merging of subgraphs that do not belong together we chose a resolution which generates smaller subclusters. Keeping in mind when annotating clusters that it is probable that these subclusters belong to the same higher order cluster. This way smaller clusters that carry biological information are not lost and the clustree graph is used to guide the choice of higher order clusters

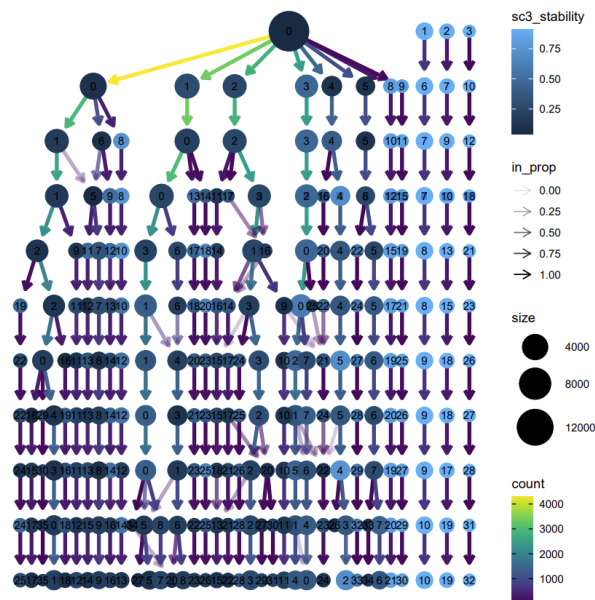


Figure 6 Clustree output for whole dataset, this shows the sequential increase in modules based on resolution. The colour of the node indicates stability, with a lighter colour indicating a more stable module.

Appendix D UMAP values comparison

The parameter `n.neighbors` can be used to choose where to lay the trade-off between local structures and global structures in the visualisation. Lower values of `n.neighbors` forces the algorithm to prioritise local structure while higher values put more focus on global structure of the data. It is important to choose a value for `n.neighbors` that preserves global structure without packing the points so densely that local structure is lost. The figure below (Figure 7) shows the dataset of this study at different values for `n.neighbors`.

By comparing the plots, we can see that the default value for `n.neighbors` of 30 yields a plot with defined global structure without too densely packing the points. Comparing the default plot to that of other values allowed only slight further insight into the structure of the data, but neither values below or above 30 provided an significantly increased insight into the data structure. Therefore, the default value of 30 was used in this study.

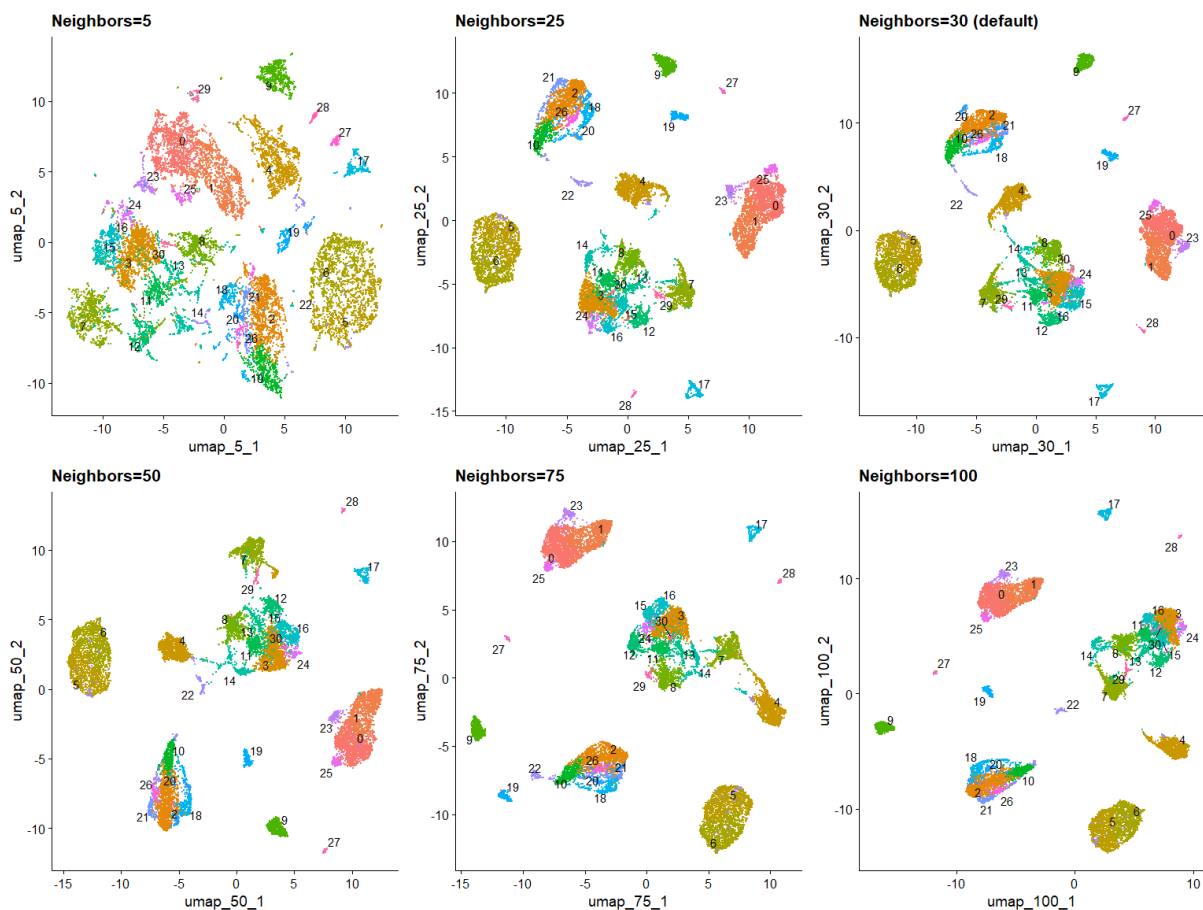


Figure 7 Comparison of UMAP results using different values for number of neighbouring points (`n.neighbors`). Larger values for `n.neighbors` give priority to global structure preservation while smaller numbers preserves more local structure by constraining connectedness between neighbouring points. The default value of 30 was used in this analysis.

Appendix E Marker gene quality control

Some marker genes that did not behave as expected were traced back for quality control in the original data. These were the genes TARP, TRBV65, and FOXP3.

E.1 TARP and TRBV65

TARP and TRBV65 were controlled by looking comparing were in the gene the annotated sequences occur. Since the sequencing is performed from the 3' end of the mRNA the reads are expected to map to the 3' end of the gene. This was true for TRBV65 but not for TARP. The histograms in Figure 8 show that the TARP sequences in a larger frequency than the TRBV65 sequences map to downstream parts of the gene. When performing short nucleotide blast on the TARP sequences it was also found that a large amount of reads did not hit against the expected gene in the database. This indicates that some of these mappings are probably faulty and might lead to erroneous interpretation of the expression profiles for the g/d T-cells.

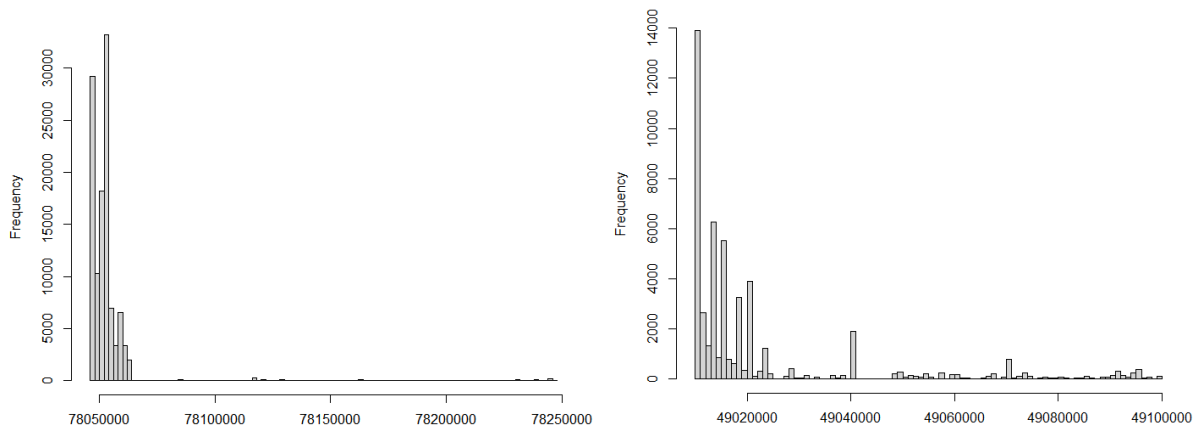


Figure 8 Histograms of sequence mapping for TRBV65 (A) and TARP (B) based on position on the chromosome. First position corresponds to starting position of the gene on the chromosome.

E.2 FOXP3

FOXP3 is a gene recently hypothesised to be present in chicken Tregs. It has been known to exist as a Treg transcription factor in mammals but was earlier unknown within chickens (Burkhardt *et al.* 2022). It is still currently missing from the reference annotation of the chicken genome. Therefore, it was controlled if any instances of reads belonging to this gene occurred. This was controlled both by manually adding the putative FOXP3 gene to the annotation file before read counting and looking for expression (in one sample). This led to no indication of significant expression. It was also controlled using short nt blast. This was done by constructing a blast database of all reads per sample and querying the mRNA sequence against this database. Neither this showed any indication of presence of FOXP3 genes in the samples. Due to this this gene was not used in this study.

Appendix F Sample compositions

Table 1 Comparison of fractions of cell types in total data per sample in (A) the computationally analysed data and (B) FACS-analysis of cells prior to sequencing. Large percentual differences are marked in yellow. These are probably due to issues in marker gene expressions (see Appendix E)

(A) % of total counts											
Cell type	Heterop hils	Tre g	MMR 1L4	LOC3 96098	TRBV 6-5	TARP	CD4	CD8A	CD8B P	ITGA 2B	ITGB 3
Sample 1	1,86	5,6 2	6,15	14,13	68,28	27,85	13,27	7,21	1,95	1,82	2,13
Sample 2	3,56	2,7 4	16,61	17,24	51,76	29,51	9,34	4,52	1,74	3,23	4,75
Sample 3	2,17	4,3 7	14,69	12,63	58,70	29,04	11,77	8,20	2,21	3,05	4,07
Sample 4	2,17	4,3 9	14,73	12,59	58,68	29,08	11,77	8,17	2,21	3,04	4,00
Total sum	2,31	4,4 5	12,77	13,75	60,11	28,81	11,78	7,35	2,07	2,75	3,65
(B) % of singlets											
Cell type	Heterop hils % singlet s	CD4C D25 %singl ets	MRC1 L-B %	Bu1 %	TCRab all %	TCRgd	CD4 all % singlet	CD8a %Singl ets	CD8b % singlet s	CD41/ 61 mean	
Sample 1	3,37	0,5 6	4,72	9,63	11,48	11,24	6,41	4,89	4,21	3,03	
Sample 2	3,92	0,3 2	9,87	6,37	9,21	10,06	5,88	3,56	3,58	2,38	
Sample 3	2,86	0,4 9	14,33	7,06	16,27	12,74	8,40	8,55	6,55	5,03	
Sample 4	3,40	0,5 4	14,50	6,94	16,27	16,61	8,49	9,22	7,17	5,46	
Total sum	3,39	0,4 8	10,86	7,50	13,31	12,66	7,30	6,55	5,37	3,98	