**Introduction**

**Single Cell RNA Sequencing**

Eukaryotic tissues are complex environments made up of a host of cell types, each playing important roles in tissue function and development[1]. Although cellular biology has progressed until very recently through the study of bulk populations of cells, these techniques represent average cellular behaviour and therefore obscure the action of rare cell populations and interactions between subpopulations. In order to study rare cell types, populations of cells can be purified through the use of fluorescence activated cell sorting (FACS)[2]. However, it has been shown that populations which are homogenous in terms of cell surface markers can in fact be transcriptionally heterogeneous and FACS can therefore obscure subpopulations of cells whose biological function can vary[3, 4]. Therefore, it has become increasingly evident that a deep understanding of tissue function requires the ability to study biological systems at single cell resolution. Recent strides in sequencing technologies mean it is possible to obtain genome wide transcriptome data from single cells using high throughput protocols [1, 5-7](scRNA-seq). scRNA-seq measures the distribution of expression levels for each gene across a population of cells and therefore allows for the study of biological questions in which cell specific changes in transcriptome are of importance. The cellular resolution of scRNA-seq and its genome wide scope mean that scRNA-seq can address questions that are intractable using other methods. These questions can relate to cell type identification, transcriptional heterogeneity within a population[1], the pseudo-temporal development[8] of cells along a differentiation trajectory or the fine response of a cell population in response to external stimuli. ScRNA-seq is therefore a tool of central importance in the stem cell field.

In general, scRNA-seq protocols requires both the isolation and processing of large numbers of single cells, during which only a portion of the transcriptome of each cell may be captured[1, 5]. Until recently, these factors limited the scope of scRNA-seq to answer biological questions involving the action of rare populations, who’s function may depend on a large number of factors and who’s study required the passage of large numbers of cells. Drop-seq was a protocol developed to process large numbers of isolated cells whilst keeping track of the identity of each cell[6]. The Drop-seq protocol utilises microfluidics to capture cells within droplets alongside microbeads which contain distinct chemical barcodes (**Figure 1**). The unique chemical barcodes attached to microbeads allow specific transcripts to be associated with specific cells in a high throughput manner and modern platforms such as the 10x Genomics platform which utilise this approach allow five thousand single cells to be processed at once. Other platforms[9-12] such as Smart-seq2 have been developed which allow much deeper sequencing of single cells. However, these techniques are both more expensive and can only process a few hundred cells at one time. Platforms such as 10x Genomics are therefore popular when studying rare populations.

The 10x Platform suffers from limitations when studying populations of cells harvested from multiple donor organisms. When studying rare populations of cells or the response of biological systems to external stimuli, the heterogeneity of response between individual donors can be of biological relevance. When characterising the transcriptional and functional property of a cell type, for example, the biological response of that cell type to external stress is not likely to be identical between individuals. To isolate properties that are conserved between individuals and identify those properties which are confined only to a single donor, it is therefore important that within a given experiment, multiple individuals are considered. Concurrently, we also wish to avoid batch effects in which separate donors are processed individually meaning that technical variation between scRNA-seq experiments could be confused with *bona fide* biological variation. This necessitates processing biological material from multiple donors within a single scRNA-seq experiment and it is currently not possible to keep track of the donor identity (ID) of each individual cell during subsequent gene expression analysis. Keeping track of donor ID during single cell experiments is of central importance when studying response to external stimuli and therefore relevant when studying, for example, immune response to infection or inflammation.

**Figure 1 –** Drop-seq experimental procedure. Cells are first dissociated from each other and then capture in droplets with microbeads.

****

In this project I develop a method to study individual donor contributions to high throughput scRNA-seq experiments involving multiple donors. Within a population of cells harvested from the tissues of multiple donors, cells of the same type from different donors could be expected to express the same genes. However, unless each donor is genetically identical, the DNA encoding those genes will vary from donor to donor. Therefore, the transcribed RNA within cells from different donors will also show heterogeneity. A standard 10x Genomics scRNA-seq experiment produces a file containing all sequenced RNA transcripts with each transcript assigned to a specific cell. In order to assign a likely donor ID to each cell I utilise somatic mutations present within these transcripts to locate genomic sites which vary between cells. Using these sites, I assign a genotype to cell at each genomic site and subsequently cluster cells based on shared mutational profiles in order to assign each cell to a likely donor. In order to test the method on a real biological system I will use recently generated data from the Cvejic lab, utilising zebrafish as a model organism to study the function and evolution of vertebrate innate immune systems.

**Innate Lymphoid Cells**

Vertebrate immune systems consist of an innate arm, which responds immediately to challenge, and an adaptive arm which responds via acquired antigen receptors. In mammals, B and T cells contribute to adaptive immunity whilst myeloid cells form the innate immune system[13, 14]. Recently discovered innate lymphoid cells (ILCs) represent a rare population of lymphocytes. Unlike T and B cells, ILCs don’t express antigen receptors or undergo clonal expansion when stimulated[15]. ILCs instead sense and respond to environmental cues via cytokine receptors. It has been shown that during homeostasis, humans and mice contain four populations of ILCs: natural killer (NK) cells, and three types of helper ILCs (ILC1, ILC2 and ILC3). NK cells directly kill cells infected with intracellular pathogens and bear resemblance to cytotoxic T cells. Helper ILCs bear similarity to classes of T helper (Th) cells. Th1 and ILC1 cells both contribute to defence against intracellular pathogens. Meanwhile Th2 and ILC2 cells contribute to defence against helminths and venoms. Finally, ILC3 and Th17 promote immunity against bacteria and funghi[15-17].

In humans and mouse, ILC1, ILC2 and ILC3 are classified based on their transcription factor (TF) and cytokine secretion profiles as well as phenotypic cell-surface markers [15, 16, 18]. Due to lack of available antibodies, it is not possible to use phenotypic cell surface markers with zebrafish and therefore the Cvejic lab used high throughput scRNA-seq to transcriptionally characterise the innate zebrafish immune system. Since ILCs were initially identified in mice [15, 16] through the use of Rag1-/- mouse lines, which are not able to produce adaptive lymphocytes, Rag1-/- fish were similarly used under homeostatic and immune challenged conditions. In particular, to study the behaviour of ILCs in zebrafish, they established three short-term inflammation models to trigger cytokine expression of potential ILCs in the zebrafish gut. Injection of *Vibrio Anguillarum* was used to induce type 3 immunity, whilst *Anisakis Simplex* was used to induce type 2 immunity and finally PBS was used as a control. Six hours after immune challenge, scRNA-seq analysis was used to study the makeup of innate lymphoid populations within the fish. Using these inflammation models and ILC marker genes, the Cvejic lab identified an ILC1-like subpopulation as well as a rare naïve ILC population which could dynamically respond upon immune challenge and become ILC2- or ILC3-like. In each of the three inflammation models, cells were clustered based on differential gene expression to identify the active cell types within the immune challenged fish (**Figure 2**).

Crucially, the identification of ILC-like populations within zebrafish required the analysis of thousands of cells from multiple fish under each of the three conditions studied by Cvejic et al. This requirement necessitated the use of a high throughput platform such as 10x Genomics. However, since response to immune challenge will vary between individuals, it is important to investigate whether individuals are particularly over or underrepresented within each of the transcriptional clusters identified by Cvejic et al. In particular, if an observed cell type response is comprised only of cells from a single individual then that cluster of cells is likely to represent an individual specific response rather than one conserved across individuals. Using the transcriptional data generated from each of the short-term inflammation models, I use the single cell genotyping protocol described above to firstly assign cells into clusters based on likely donor ID. Using these IDs, I study the breakdown of each of the clusters identified by Cvejic et al into different genotypes. I find that depending upon the immune challenge, different donors present a very different immune response in some cases. Most cell types identified by the Cvejic lab are composed from multiple different fish. However, even some more common cell types, as identified by scRNA-seq, show a definite skew towards some donors, highlighting the heterogeneous immune response. This method of genotyping represents a novel approach to preserving individual donor information within scRNA-seq experiments whilst still requiring high throughput.

**Figure 2 –** In each of the three inflammation models, cells were clustered based on their gene expression profiles. Cell types were subsequently identified using previously published marker genes allowing the transcriptional characterisation of entire zebrafish innate immune system.

******

**Methods**

**Experimental Protocol**

Each of the three short-term inflammation models was carried out using the Cell Ranger pipeline from the 10x Genomics platform. FAC sorting of cells harvested from *Tg(lck:EGFP)* Rag1-/- transgenic reporter lines was shown to label a population of lymphoid cells which showed no expression of T cell specific genes and therefore represented an innate lymphoid population. Each experiment consisted of three fish (nine total) which were injected intraperitoneally with PBS, *Vibrio Anguillarum* and *Anisakis Simplex* in experiments 1, 2 and 3 respectively. Cells were subsequently harvested from the zebrafish gut six hours post immune challenge. All cells from a given experiment were processed together for scRNA-seq within a single Cell Ranger experiment. In total, after quality control, 3211 single *lck:EGFP+* cells from the guts of PBS-injected, 3626 cells from *A. simplex*-injected and 3487 cells from *V. anguillarum*-injected *rag1-/-* zebrafish were collected and processed. Each cell from each experiment carries a specific cellular barcode from the Drop-seq experiment. All of my subsequent analysis was done using a list of barcodes from each experiment alongside the aligned transcript data generated from each Cell Ranger run.

**Computational Method**

All scripts from this project have been uploaded to an online repository at <https://github.com/dhall1995/genotyping_scRNAseq>. Briefly these consist of:

**Barcode Extraction** - From the bulk dataset containing aligned reads from all cells, for each cell barcode which has passed quality control, *Extract\_barcodes.sh* creates a separate file containing only transcripts originating from that cell and indexes them utilising the online SAMtools package (<https://github.com/samtools/samtools>). Thus, for each experiment a new directory is created containing separated datasets for each of the cells that have passed quality control from the initial gene expression analysis performed by Cvejic et al. For subsequent analysis, the resulting data sets are also merged creating a dataset containing all transcripts from only quality controlled cells.

**Initial Variant Calling** – The resulting merged dataset of quality-controlled cells from each experiment is now treated as transcriptional data from only one cell using the script *Variant\_Calling\_per\_Experiment.sh*. Variant genomic sites are identified de novo within this ‘cell’ through the use of the *mpileup* and *call* functions from the BCFtools package (<https://github.com/samtools/bcftools>). The resulting variant call format (VCF) file contains all genomic sites which show variation amongst the transcripts from quality-controlled cells. Poor quality variants are then filtered out based on criteria suggested on the SAMtools github (<https://samtools.github.io/bcftools/howtos/variant-calling.html>):

* Phred-scaled Quality score < 50
* **Or** (Read Position Bias < 0.5 & Quality Score < 70)
* **Or** (Allele Count < 5 & Quality Score < 70)
* **Or** (Maximum number of reads supporting an indel (IDV) <= 3)
* **Or** (IDV/(Read Depth) <= 0.3)

This results in a VCF file containing quality controlled variants called *de novo* from each experiment.

**Variant Calling per Cell –** The VCF file produced by the initial variant calling is now used as a reference to genotype each cell at each quality-controlled variant site using the *variant\_calling\_per\_barcode.sh* script. Due to the sparsity of scRNA-seq data, many sites won’t be present within transcripts from a given cell and such sites are lefts as *NA*. The resulting VCFs are merged through the use of the BCFtools *merge* function to create a single VCF file containing the genotype of each cell at each variant site identified in that experiment. The variants are finally filtered to exclude those variant sites present in <5% of cells. This resulted in 2592 variants identified within the *PBS* experiment, 2778 variants within the *Vibrio* experiment and 2879 variants within the *Simplex* experiment.

**Genotype Clustering –** I made use of several R packages in order to process the VCF file produced using the shell scripts for each experiment. In particular, I firstly used the VariantAnnotation package [19] for initial processing of the VCF file and filtered out variants which did not originate from single-nucleotide polymorphisms (SNPs). For each processed experiment I created an matrix containing the genotype of each cell at each variant site using the snpStats package [20]. In order to run further analysis on the matrix, genotypes are numerically encoded as:

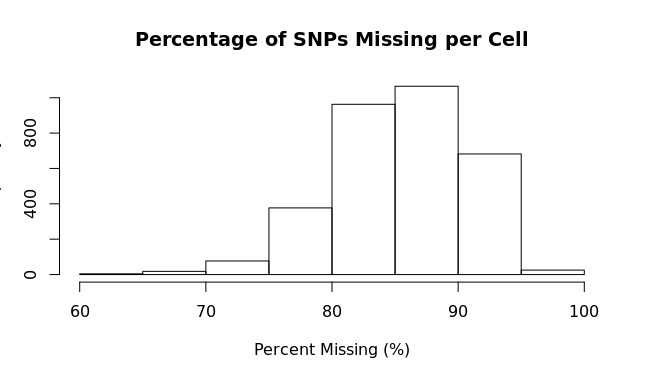
* 0 = Homozygous reference allele
* 1 = Heterozygous
* 2 = Homozygous alternative allele
* NA = Missing genotype

Due to the sparsity of scRNA-seq data, a large amount of data was missing (**Figure 3**) and I therefore made use of probabilistic PCA[21] using the pcaMethods package[22] for dimensionality reduction of the resulting matrix. Probabilistic PCA assigns missing values based on a maximum likelihood estimator and is thus tolerant to sparse data. Finally, I used the Mclust package (<https://CRAN.R-project.org/package=mclust>) for clustering of the cells by principle component. Optimal clustering of cells was expected to result in four clusters per experiment (one cluster per individual and one cluster of undetermined cells). Using Mclust, the optimal number of clusters is calculated for several clustering algorithms and the best fitting model is used to classify cells. The best model is chosen based on minimising the total uncertainty in cell cluster classification based off of the gaussian mixture models used by Mclust.

**Figure 3 a)** – Histogram of percentage of missing cell genotypes per SNP, PBS experiment

**b)** – Histogram of percentage of missing SNPs per cell, PBS experiment. There is a cut-off at ~95% due to the quality control processing of SNPs.

**a)**

**b) **

**Comparison with Previously Identified Clusters**

The above analysis was completed for each experiment and the subsequent genotype clustering was used to evaluate the donor composition of clusters identified by Cvejic et al. (**Figure 2**). Using the contributions of cells from each donor to each observed cluster, I created an contingency matrix, , for each experiment where:

In order to test whether donor cells were distributed at random throughout the observed cell types, I made use of Fishers Exact Test [23]. Fishers Exact Test calculates the probability that two categorical variables are independent of each other given a contingency matrix. A p-value is generated for the observed distribution of genotypes under the null hypothesis that no cell type shows significant donor bias.

In particular, in order to quantify the observed skew within each of the observed cell type clusters I used Fishers exact test on each cluster versus all others. Specifically, for each cell type , I created a matrix of the following form:

Where and are the number of cells in cell type originating from individual 1, individual 2 and individual 3 respectively. Similarly, and are the number of cells not within cell type originating from individual 1, individual 2 and individual 3 respectively. I thus tested whether cells from each donor could be considered to be distributed randomly within cell type given the number of cells of that type that are present and the total number of cells not of type . Monte Carlo simulations[24] were used where made necessary due to very large or very small numbers of cells. Further to this, I excluded cells whose genotype was not well defined. This analysis was repeated for all clusters from all experiments using the R Fisher package.

Finally, I generated bar plots using R to visualise the distribution of donor cells within different clusters. Bar plots were generated for both the frequency of each donor within each cluster as well as the proportional contribution of each donor to each cluster.

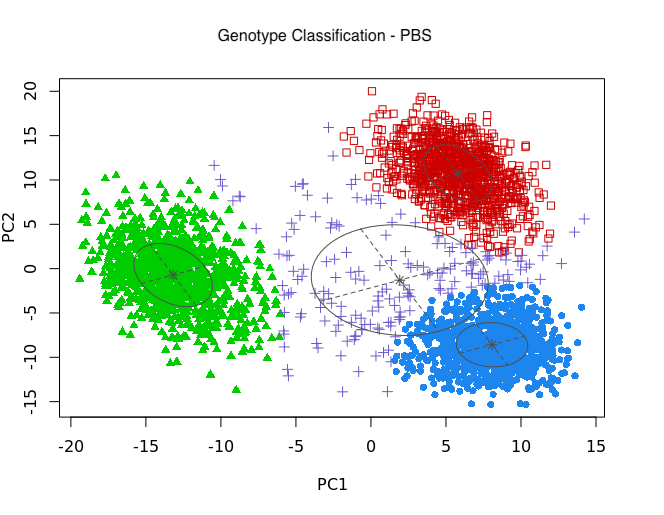
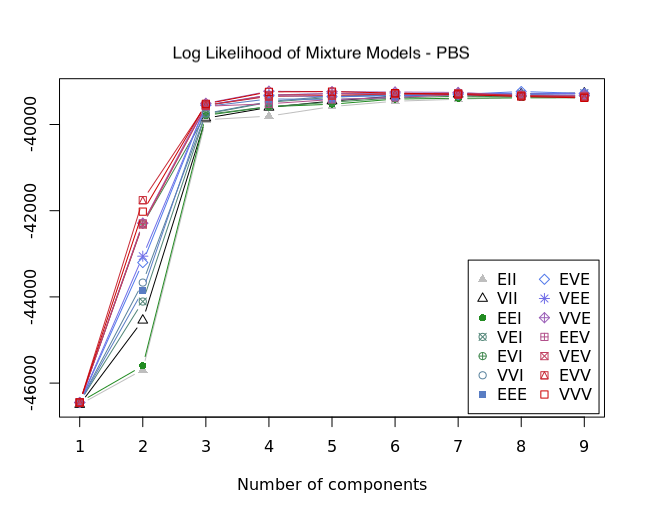
**Results**

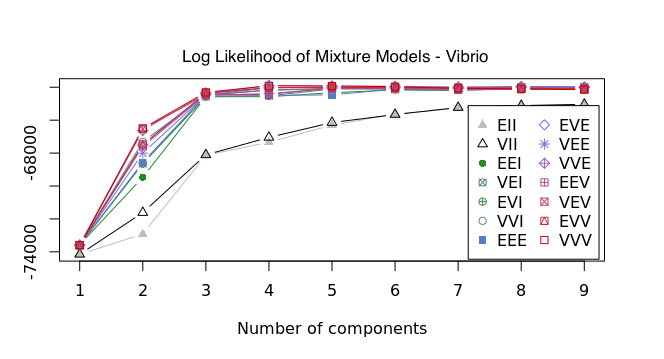
**Algorithm classifies cells well, if not optimally using Mclust**

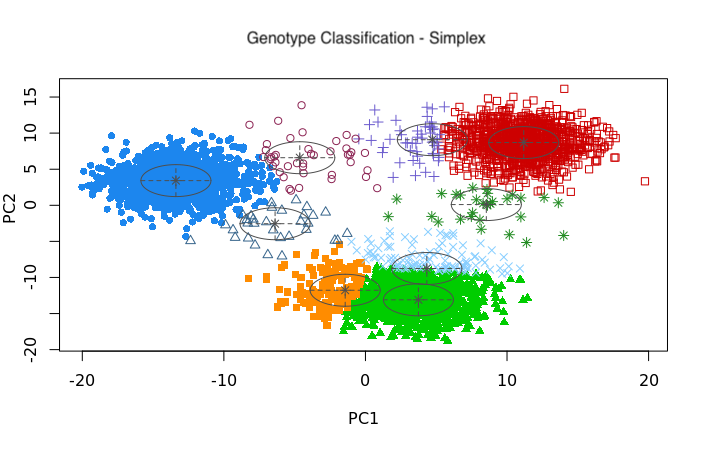
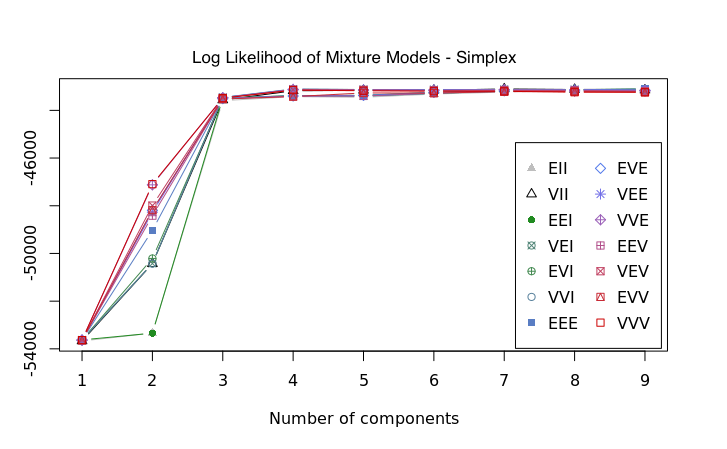
Using the given algorithm, cells within each experiment could be seen visually to separate into three clusters based off of the first two principle components (**Figure 4**). In the PBS experiment, four clusters were identified by the Mclust algorithm as the optimal clustering of cells using two principle components (**Figure 4a,b**) and resulted in 94.2% (3027/3211) of cells being assigned to a donor cluster. In particular, I identified three donor clusters with 973, 1029 and 1025 cells respectively as well as an undetermined cluster of 184 cells. Within the *Vibrio* and *Simplex* experiments, Mclust identified five and nine clusters respectively as the ‘optimal’ clustering based on minimising uncertainty in classification (**Figure 4b,d**). However, visual inspection of the clustering in these cases showed that three donor clusters were clearly visible in each case. Since both cases showed similar likelihoods for four clusters as for the ‘optimal’ number (**Figure 4c,e**) I analysed the clustering further in both the *Vibrio* and *Simplex* experiments. For the *Vibrio* case, clustering based off of the three principle components (instead of two) again recovers 4 clusters as the optimal clustering (**Figure 5a**) and I therefore used these clusters to define donors in the *Vibrio* experiment. This resulted in 96.1% (3353/3487) of cells being assigned to a donor cluster for the *Vibrio* experiment. The three donor clusters contained 986, 1075 and 1292 cells respectively whilst 134 cells were left undetermined. Finally, for the *Simplex* case, using multiple principle components did not result in four clusters as optimal. Therefore, I fixed four clusters using two principle component and selected the optimal classification using gaussian mixture models defined by Mclust (**Figure 5b**). This resulted in 96.0% (3481/3626) of cells being assigned to a donor cluster. The donor clusters in this case contained 1067, 1225 and 1189 cells respectively whilst 145 cells were left undetermined. These results demonstrate that overall each experiment contained an even distribution of cells from each fish.

**Figure 4 a), c), e) –**Initial genotype clustering of cells using the Mclust algorithm. PBS, *Vibrio* and *Simplex* experiments respectively.

**b), d), f) –** Log-likelihood of different cluster numbers using a variety of models. PBS, *Vibrio* and *Simplex* experiments respectively.

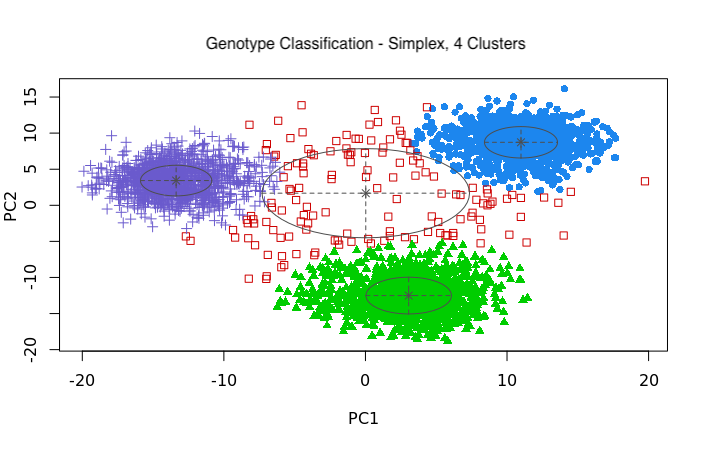
**a) b)**

**c) d) **

**e)  f)**

**Figure 5 a) –** Optimal genotype clustering of cells from the *Vibrio* experiment using the first three principle components generated using the probabilistic PCA algorithm.

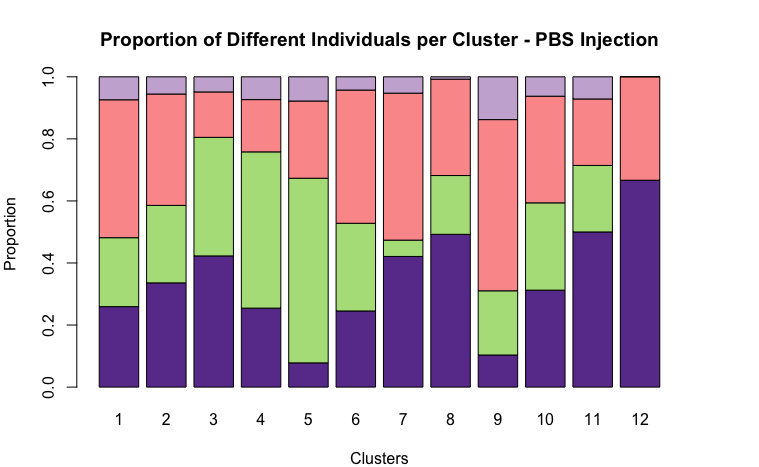
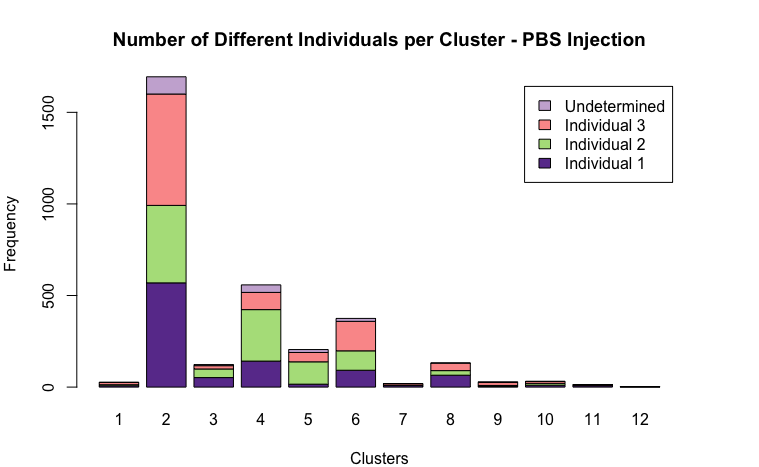
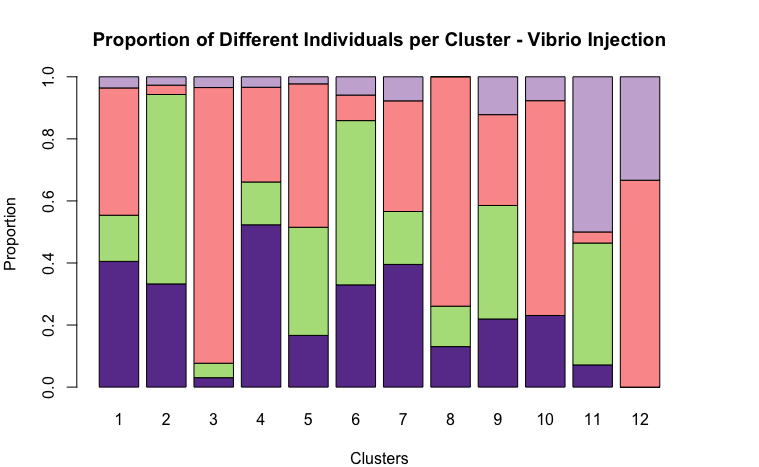
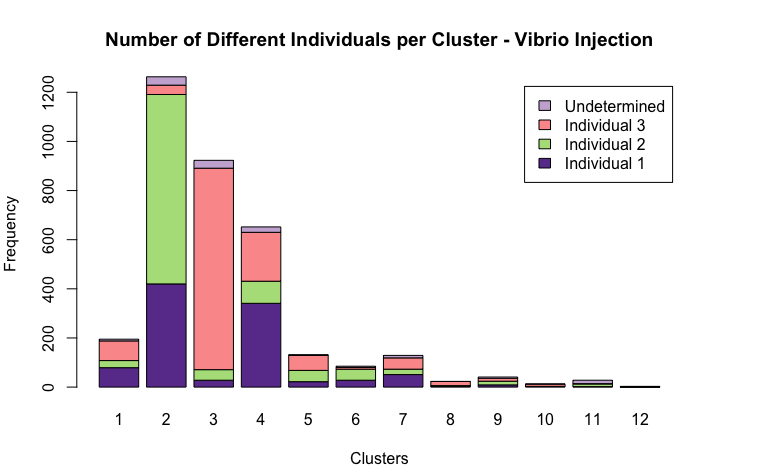
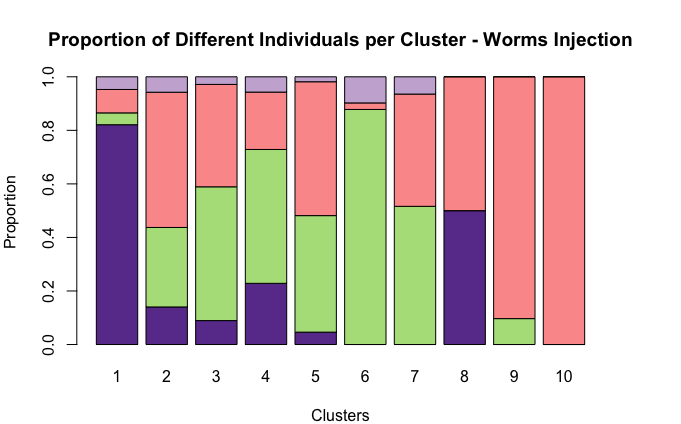
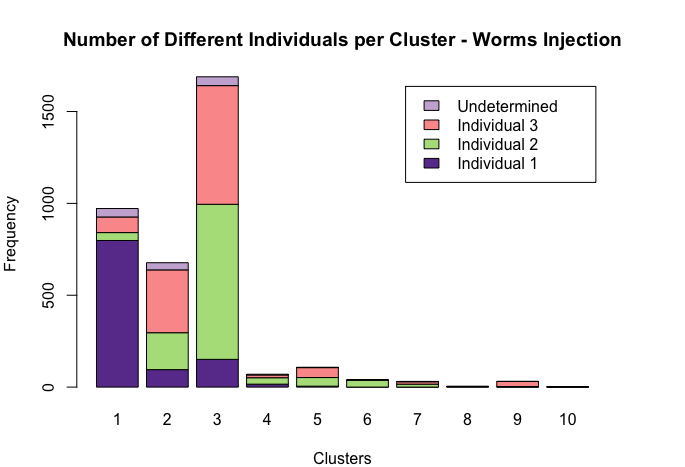
**b) –** Genotype clustering of cells from the *Simplex* experiment, forcing four clusters

1. ****
2. ****

**Assigned donor ID clusters show significant skew towards specific cell types**

Within each of the short-term inflammation protocols, Cvejic et al. find several transcriptional clusters of cells corresponding to different cell types in the Zebrafish gut (**Figure 2**). Looking at the genotype breakdown of cells within these clusters demonstrates an unequal donor contribution across virtually all clusters in every experiment (**Figure 6**). However, most transcriptional clusters are not the product of a single individual suggesting that whilst immune response is not homogenous, most cell types are at least active to some degree within all donors. Interestingly, the genotype makeup of clusters shows a more significant skew towards single donors within the immune challenge experiments (**Figure 6b,c**) than for the PBS control (**Figure 6a**). In particular, cluster 3 (923 cells) within the *Vibrio* experiment seemed largely dominated by individual 3 (820 cells, 88.8%) whilst individuals 1 (28 cells, 3.03%) and individual 2 (43 cells, 4.66%) are mostly absent from that cluster. Conversely, in cluster 2 (1263 cells), individual 3 contributes only 38 cells (3.00%) whilst individuals 1 and 2 contribute 420 (33.2%) and 771 (61%) cells respectively. Additionally, In the *Simplex* experiment clusters 9 and 10 are almost entirely made up of cells originating from a single donor.

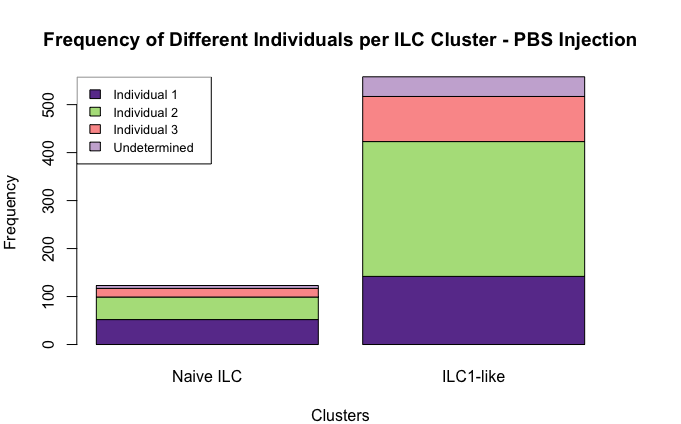
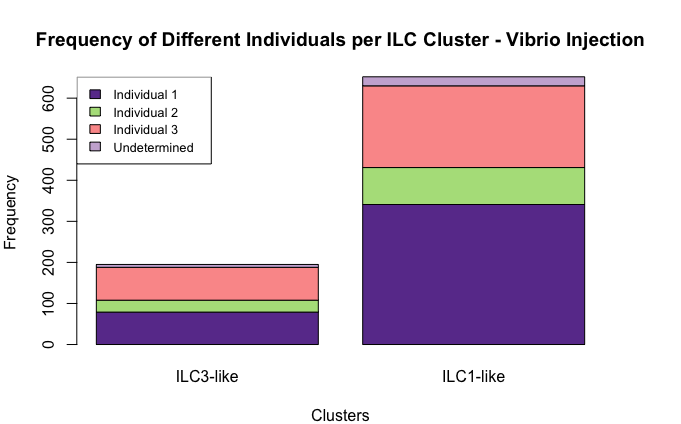
**Figure 6 –** Stacked bar plots showing the relative contribution (**left**) and relative proportion (**right**) of different donors within the cell type clusters. **a)** PBS **b)** Vibrio **c)** Simplex. Note that some smaller clusters are completely dominated by contributions from single donors or two donors.

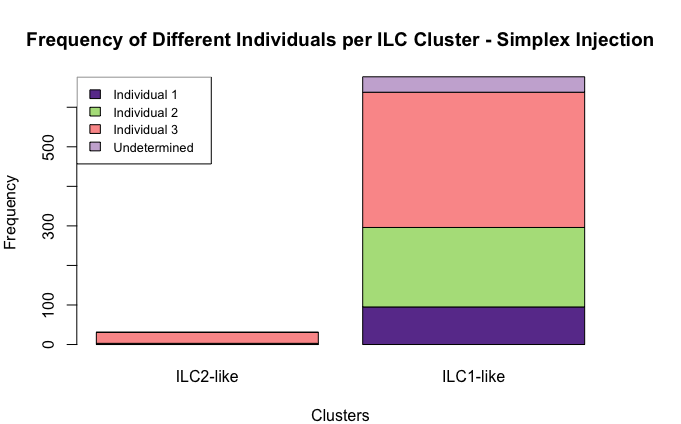
1. ****
2. ****
3. ****

**ILC-like Cells Mostly show contributions from all donors**

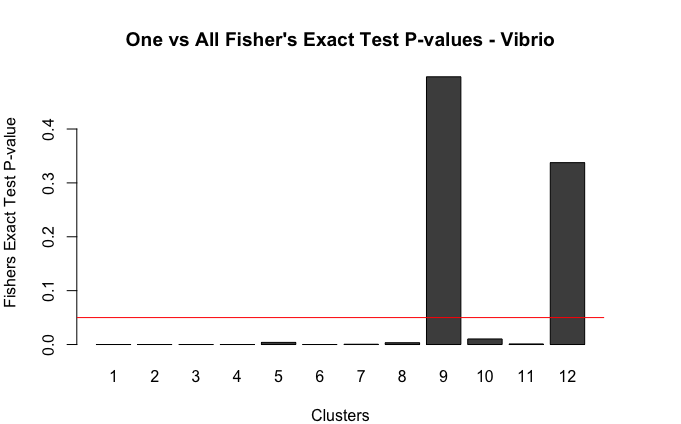
Cvejic et al established the existence of innate lymphoid populations within their single cell datasets. In particular, ILC1-like cells were seen within all three short-term inflammation protocols whilst naïve ILC cells within the PBS experiment were argued to convert upon specific immune challenge into the ILC2-like or ILC3-like cells seen within the *Simplex* and *Vibrio* cases respectively. Particular transcriptional clusters (**figure 2**) were associated with these cell types.Specifically, within the PBS experiment, clusters three and four represent the naïve ILC and ILC1-like cells respectively. Meanwhile, for the *Vibrio* experiment, cluster 4 was comprised of ILC-1 like cells whilst cluster 1 displayed ILC3-like cells. Finally, for the *Simplex* experiment, cluster 2 displayed ILC1-like cells and cluster 9, ILC2-like cells. Inspecting these pairs of clusters together shows that all ILC clusters except the ILC2-like cells in the *Simplex* experiment show a comparable contribution from all possible donors (**Figure 7**). The ILC2-like cluster is comparatively very small (31 cells, 28 Individual 3, 3 Individual 1) and therefore likely to show a larger skew in donor. Despite most clusters showing a non-zero contribution from all donors, almost all clusters show a significant p-value using Fishers Exact Test, demonstrating a definite skew towards a single donor (**Figure 8**). Surprisingly, cluster size displayed no obvious link with Fishers Exact Test results suggesting that even for more common cell types, the innate immune systems of individuals can demonstrate significant donor to donor variability.

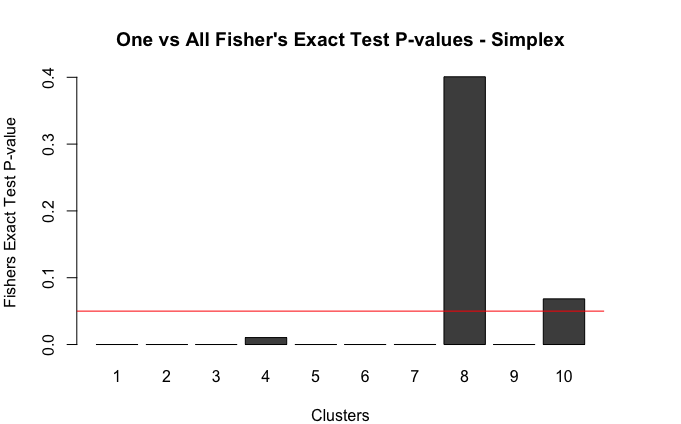
**Figure 7 –** Contributions of specific individuals to ILC-like populations discovered by Cvejic et al. **a)** PBS **b)** Vibrio **c)** Simplex. Most ILC clusters show contributions from all donors but ILC2 like cells have no representatives from individual 2 and are predominantly cells from individual 3.

1.  **b)** 

**c)**

**Figure 8 –** One vs All Fishers Exact Test p-value results for each cluster in each experiment. Most clusters show significant skew towards one donor. P-value 0.05 marked as red horizontal line on each plot. **a)** PBS **b)** Vibrio **c)** Simplex.

1. **b) **

**c)**

**Discussion**

Single Cell RNA-seq experiments are processing larger numbers of cells each year[1] and initiatives such as the human cell atlas[25] as well as already available repositories such as the Blood Atlas of Single Cells in zebrafish [26] mean that the transcriptional landscape of multiple biological systems is now accessible. When studying the biological behaviour of such systems, however, it is both scientifically desirable and technically necessary to consider data from multiple donors. Firstly, in order to understand the biology of tissues, multiple individuals are required to investigate those biological features which are necessary for tissue function and therefore conserved across healthy individuals. It is also of scientific interest to understand the variation between individuals since those variations may confer advantages or disadvantages to a specific donor and therefore be relevant in studying tissue malfunction. Secondly, multiple individuals are sometimes required simply to provide the necessary amount of biological material needed to study a subtle effect originating from a rare population of cells within a tissue. At the same time, technical difficulty as well as batch effects mean that tracking cell donor ID or processing and genotyping donors individually may be undesirable. The method generated in this project is therefore a useful tool to discern donor ID within existing as well as future scRNA-seq datasets.

**Computational Method is Broadly Successful but Requires Development**

From a technical perspective, the method I used worked extremely well to identify likely donor IDs. As shown in figure 4a, c and e, the probabilistic PCA method could be seen to visually split the cells very well into the three expected donor clusters in each experiment. However, the process is not fully autonomous since the method assumes knowledge about the number of donors. Indeed, without visual inspection of the clusters as well as the likelihood graphs shown in figure 4b, d and f, it would be unjustified to simply force the number of clusters to be four as I did for the *Simplex* experiment (figure 5b). Had the likelihood of four clusters in that experiment been significantly lower than the nine identified by Mclust then forcing the algorithm to select four clusters could result in significantly more misclassification of cells which could have large effects on further analysis.

Due to the sparsity of scRNA-seq data, similar care needs to be taken when assigning observed clusters to donors. In the case of tissues which may have undergone clonal expansion or contain tumour tissue, large clones of a single cell may express similar transcripts with the same mutational profile. Such a clone could be mistakenly identified as its own donor resulting in an extra cluster within the PCA plots. In that case, the only available information is the expected number of clusters (donors) and suspect clusters may have to be removed or assigned to another cluster simply via inspection. However, in my analysis I could exclude the possibility of clonal expansion since the observed clusters contained almost all varieties of cell types in each experiment. The low rate of failed assignment (i.e. undetermined donor) in each experiment meant I could confidently associate each cluster with an individual. In the future, this method could be extended to datasets involving more individuals to see if there is a theoretical upper limit to how many donors can be accurately identified.

**Innate Immune Response Shows High Degree of Heterogeneity between Individuals**

I observed a high degree of heterogeneity in donor contributions to each cluster in all cell types (figure 6, figure 8). Whilst individuals are expected to have comparable innate immune systems, the general state of an individuals’ immune system at any given time would be expected to fluctuate even without an applied stress such as injection of a pathogen. Therefore, the observed heterogeneity (figure 6) in all experiments is expected except in cases such as the *Vibrio* experiment clusters 2 and 3 where one or two donors dominate a very large cluster of cells (figure 6c, d). Examining the transcriptional properties of these clusters (Figure 2b, right panel) shows that the two clusters are actually extremely similar. The most significant difference between the two clusters is the high expression of the novel immune type receptor gene *nitr7b* within cluster 3. It’s therefore possible that, due to some inflammatory response, *nitr7b* was upregulated in a large group of cells from individual 3 that would otherwise have been in cluster 2. This would result in a relative abundance of cells from individual 3 in cluster 3, compensated by a relative sparsity of cells from individual 3 in cluster 2. Interestingly, across both clusters individual 2 and individual 3 have comparable numbers (814 and 858 respectively) of cells contributing. An analysis of *Simplex* experiment clusters 1 and 3 yields very similar results whereby the apparent disparity of individual 1 in cluster 3 is accounted for by the relative abundance of individual 1 cells within the transcriptionally similar cluster 1. It is therefore possible in that the two cell type clusters *Simplex* 1 and *Simplex* 3 actually represent the same biological cell type and the observed differences correspond to individual immune response. Similar arguments hold for the pair of clusters *Vibrio* 2 and *Vibrio* 3.

**ILC-like Clusters Predominantly Contributed to by All Donors**

Within smaller cell type clusters, I expected donor contributions to show a higher degree of variability so it is therefore unsurprising that clusters such as *Vibrio* 12, PBS 12 and *Simplex* 10, each containing only three cells, are completely dominated by either one or two donors (figure 6, right panels). Encouragingly, within the ILC1-like clusters from each experiment, there are significant, if not equal, contributions from all donors (figure 7). Given the size of the ILC1-like clusters (558 cells, 652 cells and 677 cells in PBS, *Vibrio* and *Simplex* respectively), the distribution of donor cells suggests that ILC1-like cells are a persistent, if rare overall population of cells within all zebrafish blood. Similar arguments could be made about the naïve ILC population in PBS (123 cells) and the ILC3-like population of cells in *Vibrio* (195 cells). However, the ILC2-like cells within the *Simplex* experiment demonstrates that and potential ILC2-like response was weak enough to not be detectable within individual 1 and to produce only 3 cells from individual 2. It is possible that the immune response generated by the *Simplex* injection was less severe than that generated by *Vibrio* injection. However, in that case I would expect to see a naïve ILC population still present within individual 1 and this is not directly observed (although these naïve cells may in truth be part of the *Simplex* ILC1-like cluster). Alternatively, it could be that the comparatively large population of *Simplex* cluster 1 cells constituted the immune response of individual 1 to the *Simplex* injection. Cells in *Simplex* cluster 1 express ifng1-2 and granzymes suggesting they are NK type cells and this cluster is dominated by individual 1. Therefore, the ILC2-like response may require further investigation to reliably conclude whether it is simply an individual specific response or whether the *Simplex* injection did not excite enough of an immune response in individuals 1 and 2 stimulate production of ILC2-like cells.

Using donor ID information is a strong tool to establish individual specific contributions to scRNA-seq datasets. Although there are technical limitations to the method and some degree of care must be used when assigning donor ID, the tool gives a different perspective on cell type clusters and can therefore be used to either strengthen or call into question conclusions which could have otherwise been drawn. In the case of zebrafish ILCs, it is likely that the observed populations exist and function according to the framework laid out by Cvejic et al. However, using donor ID, the exact transition from naïve to ILC2-like cells upon immune challenge is shown to be highly specific to the individual and therefore will require closer investigation in future.

1. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Mol Cell, 2015. **58**(4): p. 610-20.

2. Fulwyler, M.J., *Electronic separation of biological cells by volume.* Science, 1965. **150**(3698): p. 910-1.

3. Wilson, N.K., et al., *Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations.* Cell Stem Cell, 2015. **16**(6): p. 712-24.

4. Guo, G., et al., *Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire.* Cell Stem Cell, 2013. **13**(4): p. 492-505.

5. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nat Methods, 2014. **11**(2): p. 163-6.

6. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-1201.

7. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nat Rev Genet, 2015. **16**(3): p. 133-45.

8. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.

9. Nichterwitz, S., et al., *Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling.* Nat Commun, 2016. **7**: p. 12139.

10. Picelli, S., et al., *Smart-seq2 for sensitive full-length transcriptome profiling in single cells.* Nat Methods, 2013. **10**(11): p. 1096-8.

11. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.

12. Serra, L., et al., *Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq.* Bio Protoc, 2018. **8**(4).

13. Boehm, T., *Evolution of vertebrate immunity.* Curr Biol, 2012. **22**(17): p. R722-32.

14. Riera Romo, M., D. Perez-Martinez, and C. Castillo Ferrer, *Innate immunity in vertebrates: an overview.* Immunology, 2016. **148**(2): p. 125-39.

15. Eberl, G., et al., *Innate lymphoid cells. Innate lymphoid cells: a new paradigm in immunology.* Science, 2015. **348**(6237): p. aaa6566.

16. Artis, D. and H. Spits, *The biology of innate lymphoid cells.* Nature, 2015. **517**(7534): p. 293-301.

17. Annunziato, F., C. Romagnani, and S. Romagnani, *The 3 major types of innate and adaptive cell-mediated effector immunity.* J Allergy Clin Immunol, 2015. **135**(3): p. 626-35.

18. Walker, J.A., J.L. Barlow, and A.N.J. McKenzie, *Innate lymphoid cells — how did we miss them?* Nature Reviews Immunology, 2013. **13**: p. 75.

19. Obenchain, V., et al., *VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.* Bioinformatics, 2014. **30**(14): p. 2076-8.

20. Sole, X., et al., *SNPStats: a web tool for the analysis of association studies.* Bioinformatics, 2006. **22**(15): p. 1928-9.

21. Zhao, J., P.L. Yu, and J.T. Kwok, *Bilinear probabilistic principal component analysis.* IEEE Trans Neural Netw Learn Syst, 2012. **23**(3): p. 492-503.

22. Stacklies, W., et al., *pcaMethods--a bioconductor package providing PCA methods for incomplete data.* Bioinformatics, 2007. **23**(9): p. 1164-7.

23. Fisher, R.A., *The Logic of Inductive Inference.* Journal of the Royal Statistical Society, 1935. **98**(1): p. 39-82.

24. Patefield, W.M., *Algorithm AS 159: An Efficient Method of Generating Random R &#xd7; C Tables with Given Row and Column Totals.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1981. **30**(1): p. 91-97.

25. Rozenblatt-Rosen, O., et al., *The Human Cell Atlas: from vision to reality.* Nature, 2017. **550**(7677): p. 451-453.

26. Athanasiadis, E.I., et al., *Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis.* Nat Commun, 2017. **8**(1): p. 2045.