A simple, scalable approach to building a cross-platform transcriptome atlas

Paul W Angel[1], Nadia Rajab[1], Yidi Deng[2], Chris M Pacheco[1], Tyrone Chen[1], Kim-Anh Lê Cao[2], Jarny Choi[1], and Christine A Wells[1]

[1]**Centre for Stem Cell Systems, The University of Melbourne, Melbourne, Victoria, Australia**

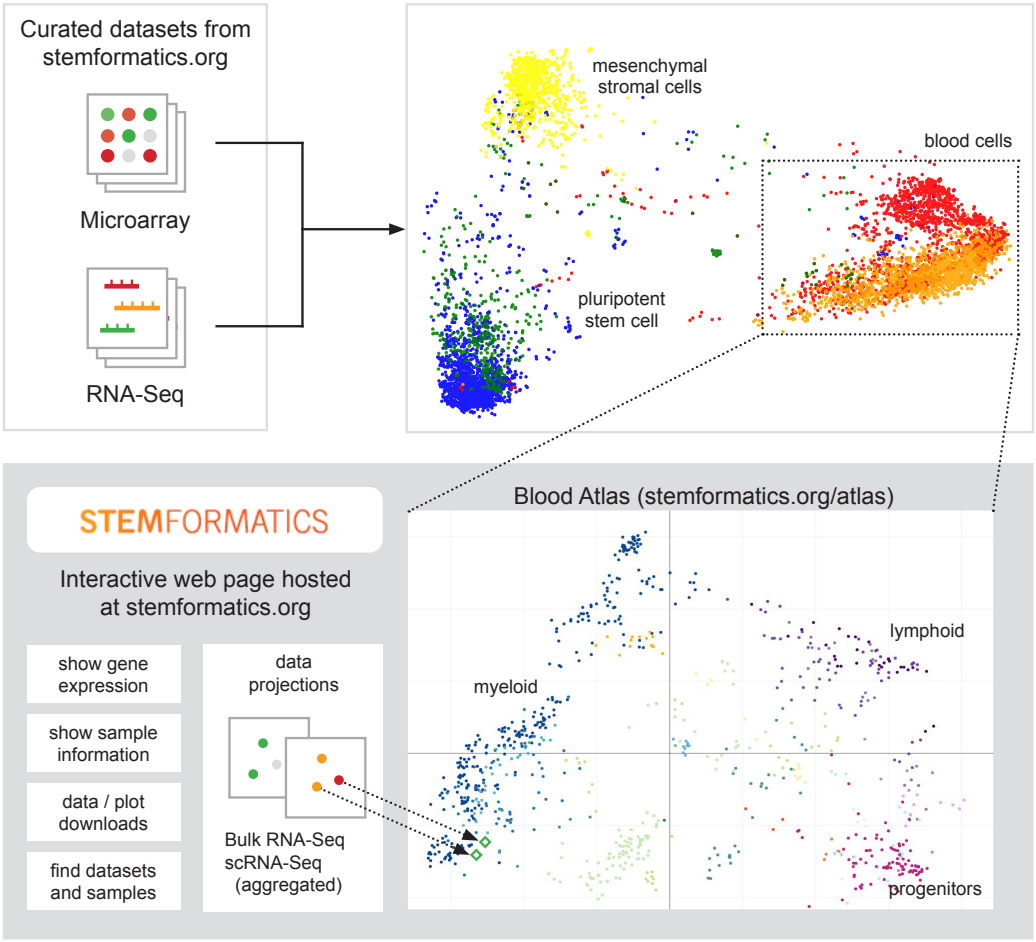
[2]**Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria, Australia**

ABSTRACT

Gene expression atlases have transformed our understanding of the development, composition and function of human tissues. New technologies promise improved cellular or molecular resolution, and have led to the identification of new cell types, or better defined cell states. But as new technologies emerge, information derived on old platforms becomes obsolete. We demonstrate that it is possible to combine a large number of different profiling experiments summarised from dozens of laboratories and representing hundreds of donors, to create an integrated molecular map of human tissue. As an example, we combine 850 samples from 38 platforms to build an integrated atlas of human blood cells. We achieve robust and unbiased cell type clustering using a variance partitioning method, selecting genes with low platform bias relative to biological variation. Other than an initial rescaling, no other transformation to the primary data is applied through batch correction or renormalisation. Additional data, including single-cell datasets, can be projected for comparison, classification and annotation. The resulting atlas provides a multi-scaled approach to visualise and analyse the relationships between sets of genes and blood cell lineages, including the maturation and activation of leukocytes in vivo and in vitro.

In allowing for data integration across hundreds of studies, we address a key reproduciblity challenge which is faced by any new technology. This allows us to draw on the deep phenotypes and functional annotations that accompany traditional profiling methods, and provide important context to the high cellular resolution of single cell profiling. Here, we have implemented the blood atlas in the open access Stemformatics.org platform, drawing on its extensive collection of curated transcriptome data. The method is simple, scalable and amenable for rapid deployment in other biological systems or computational workflows.

Keywords: gene expression, data integration, variance modelling, leukocyte



Graphical abstract: Recursive approach to generating a multi-scaled atlas. Top panel: The method integrates data from all cell types in the Stemformatics database, and shows clear division of samples into global categories of stromal, pluripotent or blood (inset) cell types. Bottom panel: Integration of only the blood cell subsets provides a blood atlas. Projection of external samples (green) onto the blood atlas. Samples are coloured by curated annotations derived from the original studies, and can be viewed at Stemformatics.org

INTRODUCTION

RNA profiling has been a mainstay descriptor of cellular systems for over two decades, but methods for measuring transcript abundance have changed dramatically over this period. The field was revolutionised by microarrays, which allowed simultaneous hybridisation and colourmetric read out for a catalogue of known genes (Schena et al., 1995). Microarrays were rapidly adopted because they were a fast, inexpensive and simple way to measure the transcriptional output of a biological system. However, the need to predefine sequences to be interrogated, and a linear range constrained by the stoichiometry

of probe and target meant that microarray platforms were rapidly superseded by RNA sequencing (RNAseq) technologies. Now the most prevalent experimental platform, the range of detected transcripts is determined by the number of tags counted in a sequencing run and cellular resolution is determined by the complexity of the profiled population (Cloonan et al., 2008; Forrest et al., 2014). Increased resolution has escalated rapidly with the advent of single-cell RNA sequencing (scRNAseq) technologies (Stubbington et al., 2017; Regev et al., 2017). Although some platforms are being refactored and repurposed, such as the reinvention of hybridisation-based platforms for spatial profiling (Eng et al., 2019), successive technologies become rapidly redundant, as does the data generated on them.

There is a need to move past information gathering and to move towards build new knowledge frameworks. Yet technology change drives much recursive data derivation. This represents a massive intellectual and financial investment by research groups and funders on data that is not adequately being reused, despite its availability in data repositories (Athar et al., 2018; Lizio et al., 2016). A wealth of information still resides in data generated on obsolete technologies: these collectively represent a large back catalogues of carefully phenotyped cells and meticulous experimental systems that can be viewed one study at a time in platforms such as ArrayExpress (Papatheodorou et al., 2019). A major barrier to data reuse is the computational capacity to directly integrate and compare successive technologies. While the drivers for platforms are increased sensitivity and resolution of systems-scale measurements, it remains difficult to benchmark the new against the old.

Drawing on curated knowledge commons is particularly important when new platforms, such as scRNAseq, rely on annotations from post-hoc analysis rather than starting with well phenotyped cells. The methods that are most commonly used to integrate scRNAseq with different platforms rely on projection or harmonisation of different data types onto a reference scRNAseq data set, and are designed to compare data in a pairwise manner, so are not easily scaled to include many experimental series (Stuart et al., 2019). In order to take advantage of the back-catalogue of phenotype-gene expression data, we need new approaches to combine experimental series from several different platforms and across multiple studies.

Combining RNAseq with the microarray is particularly challenging because data are acquired in a continuous (microarray) or discrete (RNAseq) manner, and the number of genes captured in a single cell may be orders of magnitude less than that measured in a population. While it is most common to combine data from the same microarray platform (e.g. Novershtern et al. (2011) and Hawrylycz et al. (2012)) or RNAseq (e.g. Frazee et al., 2011; Leek, Johnson, et al., 2012.) combining different types of platforms is less common (e.g. see also Rohart et al. (2016)). Combining microarrays with RNAseq has been previously attempted (Thompson et al., 2016; Taroni and Greene, 2017), however, these methods focus on global normalisation, which has a major impact on stability and scalability when new data is imported. Many normalisation approaches that account for platform variance require prior identification of sample groups that are expected to harmonise together. This can introduce class biases, whilst also enforcing

such strong transformations to data structure that meaningful biological signal is removed - these are acknowledged problems with batch correction methods such as COMBAT (Johnson et al. (2006)), and RUV-III (Gagnon-Bartsch and Speed (2012)). Class imbalance is typically encountered when attempting to merge a small number of data sets. For example, when benchmarking a new sample type against an existing exemplar the lack of common or appropriate reference samples in the comparison, as well as prior designation of sample class in the normalisation structure can lead to spurious claims of cell-type similarity. This could be addressed if new data could be compared to a reference atlas series, but no such benchmark exists.

Here, we use the Stemformatics catalog (Choi, Pacheco, et al., 2018), which has curated hundreds of studies, to assess the extent that platform impacts on expression variance for each gene. This challenged our initial assumption that accounting for batch necessitates an adjustment to every gene expression value. We selected a subset of genes with low attributable platform variance to compile samples from many studies, resulting in a reference atlas that reflects cell properties that are independent of mode of measurement. By including sufficient representation across different cell types we gain insights into the behaviour of related cell types, whilst also providing a platform for further analysis (e.g. comparisons between disease and normal states, or between in vitro and in vivo models); and to benchmark new platforms, including scRNAseq.

MATERIAL AND METHODS

In designing this method, the effect of platform is assumed to be systematic variation, and other batch effects will be averaged out by the multiple datasets covering the biology. We test these assumptions by leveraging the large collection of data in Stemformatics which samples different platforms and numerous cell types. The method introduced here assesses each gene independently to quantify the impact of experimental platform on that gene's expression across the whole data series. Genes with low platform effect are selected for subsequent analyses.

Data Curation

All data used to compile the blood atlas was curated for data quality, and for method of cell isolation and phenotyping. This metadata is captured in the Stemformatics annotation table (available to download at https://www.stemformatics.org/atlas/blood), and includes tissue source, antibody profiles where bead or FACs isolation is used, age of donor (fetal, neonatal, adult). Cells that are profiled directly from tissue are annotated as an in vivo source; mature cells isolated from blood or bone marrow and cultured for any period of time are labelled ex vivo; and cells differentiated in the laboratory from hematopoietic progenitor (typically mobilised peripheral blood, bone marrow or cord blood) or from a pluripotent cell source are labelled in vitro. This information is available to the viewer in the Stemformatics implementation of the

blood atlas; primary data sources and publications are linked from every data set page.

Note that in early iterations of the atlas, T-cell subsets isolated using negative selection alone were found to have a high monocyte contamination when compared to T-cells isolated using flow cytometry gates, as evaluated by high expression of myeloid marker gene profiles CD14, CD16 and HLADR. Therefore samples isolated using negative-selection methods were excluded from the atlas unless further purification and phenotyping was provided by the authors.

The standard Stemformatics processing pipelines were implemented, where data was assessed for linear range/library size, RNA species, and RNA degradation using 5'/3' signals where appropriate for the profiling method. Datasets were excluded if they showed evidence of over-amplification, incomplete data availability in the public databases GEO or ArrayExpress, incomplete sample metadata or identified sample-swaps, or where experimental design was confounded. Details of the Stemformatics data curation pipeline are available (Wells et al., 2013; Choi, Pacheco, et al., 2018).

Data transformation

Combining datasets measured on microarray platforms and RNAseq presents two main difficulties. Firstly, each platform produces data on a different scale, i.e. they measure abundance in different units. Secondly, microarrays are composed of gene probes, which are physically different and may be in principle measure transcripts not represented by alternate array models. These problems are addressed in two stages presented below, a data transformation stage and a gene filtering stage.

Only genes measurable in all of the available platforms are used to construct the atlas. In this instance we start with 13,661 genes. Expression values from RNAseq (RPKM) or microarray are transformed to the same scale. Microarrays have a component of lowly expressed genes at a non-zero value, whereas lowly expressed genes within RNAseq data can be exactly zero. Thus data structure (discrete vs. continuous) and sensitivity are quite different. Gene expression for each samples is transformed into rank percentile values - the highest expression gene is assigned a value 1 and the lowest receives a value 0. Values in between are uniformly spaced accorded to the rank of the genes expression. Tied values are given the same rank, which is average of their would-be ranks if they were not tied. Note that this scheme is scalable because the inclusion of new samples only requires that they are given the same ranked transformation, avoiding the need to continually renormalise the entire data series. The analysis of the influence of the rank transformation on the platform effect can be found in the Supplementary Section S1.3.

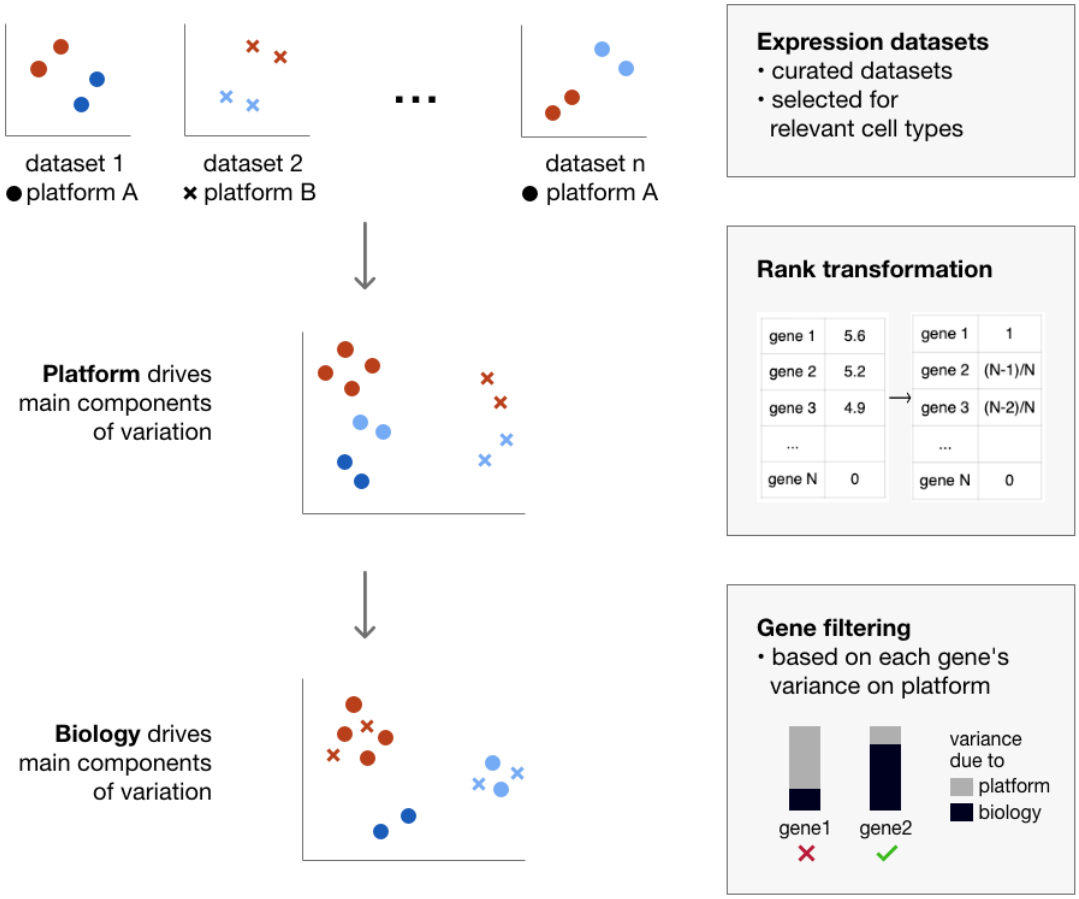


Figure 1. The blood atlas is constructed by integrating many independent curated datasets. Top row: the individuals PCAs of a set of quality-controlled independet datasets. These datasets are measured on a different platforms. Middle row: genes are rank transformed in order to move the expression distributions from the different platforms onto the same distribution. However, after running a PCA on the transformed data a platform clustering is still present. Bottom row: genes are univariately assessed for platform dependence, and filtered in order to keep only genes with a low fraction of the variance dependent upon platform. The resulting PCA then shows clustering based biological features.

Principal component analysis (PCA) is performed to collate samples after the percentile transformation, in order to find reliable global structure (Moon et al., 2017). As in Figure 1 middle row, there is a clear platform effect in the clustering of the samples, which must be suppressed. We estimate the platform effect on each gene by fitting a univariate linear model with platform as an independent variable,

$$y = X_p \beta_p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

where y is the expression of a single gene across all samples, X_p indicates membership of the platform

with coeffecient β_p. The variance attributable to platform is defined as

$$\sigma_p^2 = var(X_p\beta_p),$$

and the total variance

$$\sigma_{\text{Total}}^2 = \sigma_p^2 + \sigma_\varepsilon^2.$$

therefore, the proportion of variance attributable to platform is

$$\frac{\sigma_p^2}{\sigma_{\text{Total}}^2}.$$

In practice this is implemented this using the variance partitioning package Hoffman and Schadt, 2016a, with a single fixed effect (platform). This model is a fixed effect analysis of variation (ANOVA).

The distribution of variance attributable to platform is shown in Supplementary Figure S1. Approximately 25% of genes examined were seriously impacted by platform, that is more than half of their variance was attributable to platform. Most genes were not overwhelmed by their method of measurement. In order to select the genes with minimal dependence upon platform a threshold of 0.2 of the variance of a gene is required. The PCA was constructed from this gene subset. The resulting PCA is shown in Figure 2 and effectively removes platform dependence. The process reducing platform dependence when lowering the threshold is outlined in Supplementary S1.1.1 and Figure S2. All PCA generation was implemented via the python scikit-learn package (Pedregosa et al., 2011).

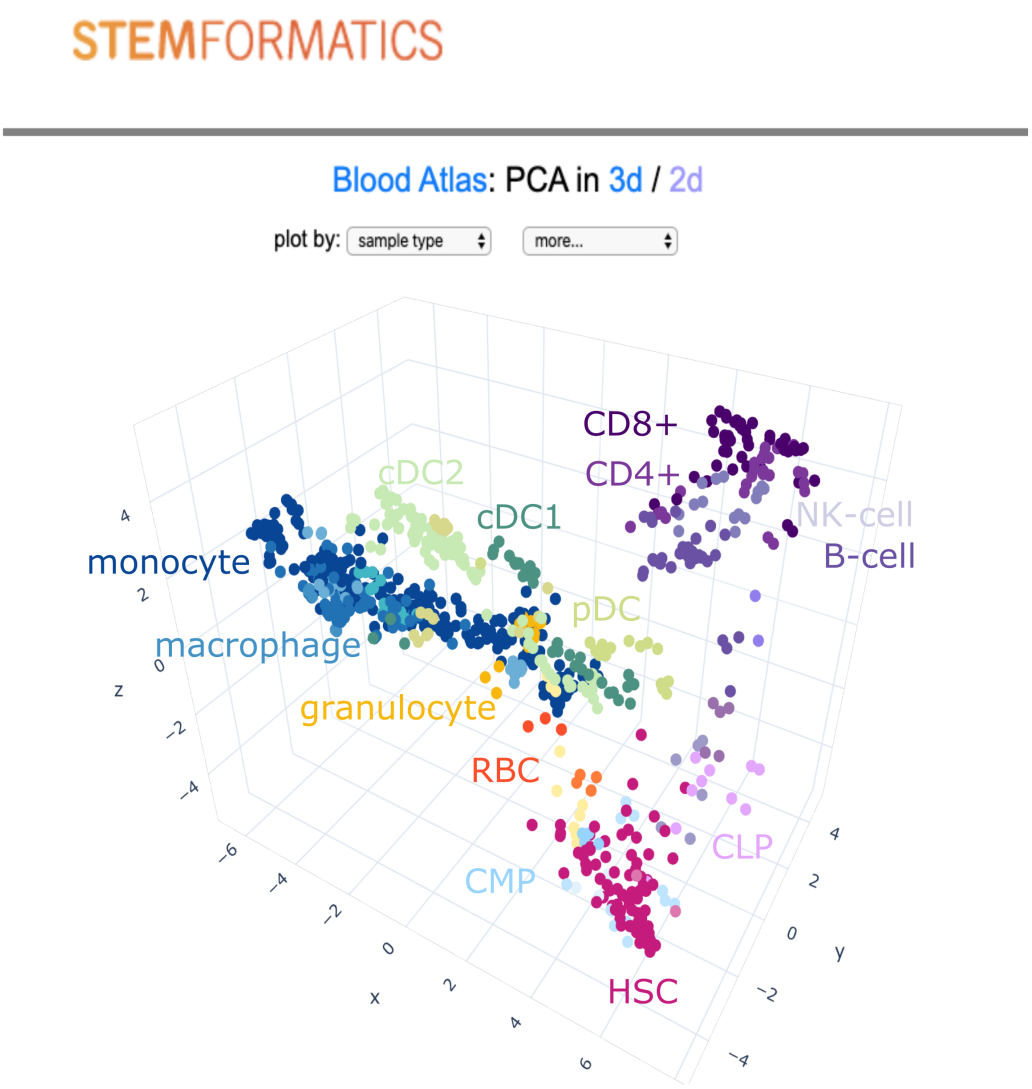


Figure 2. The S4M blood atlas, in which each point is a sample from one of the 38 independent datasets. There are 3700 genes used in construction of the PCA. The colour indicates the annotated cell type. Progenitors sit in a region in the corner, while the the myeloid and lymphocyte arms separate out. The lymphocyte region includes both T and B cells. Dendritic cells sit in the cloud in the center if derived from an in vivo source, or cord-blood derived DC sit in a group.

Comparison of filtered and non filtered genes based on variance partitioning

To assess the effect of gene filtering in our approach, we partitioned the variance of all genes from the original data set (13661 genes) and calculated the variance explained by the 'Class' (sample source, progenitor type or cell type) and Platform using a linear mixed model (LMM) as follows.

Ranks percentiles were transformed using the probit function to fit the normality assumption of a LMM. For each gene i, $i = 1, \ldots, 13661$, we fitted a linear mixed model of the form

$$y_i = \mu + Z^{(1|Class)}\alpha_{Class} + Z^{(1|Platform)}\alpha_{Platform} + \varepsilon$$

with variance components $\alpha_{Class} \sim N\left(0, I\sigma^2_{class}\right)$ and $\alpha_{Platform} \sim N\left(0, I\sigma^2_{Platform}\right)$.

The proportion of variance explained by class effect and platform effect was evaluated with the variancePartition R package (Hoffman and Schadt, 2016b) and genes were ordered according to their estimated class/platform variance ratio $\frac{\hat{\sigma}^2_{class}}{\hat{\sigma}^2_{Platform}}$.

Clustering

In order to define regions of cohesive biology and test the stability of the atlas, we applied K-Means clustering to the principal components, which represent the coordinates of each sample in the 3D space. It is implemented via the sci-kit learn packages (Pedregosa et al., 2011). To provide a comparison, agglomerative (bottom up hierarchical) clustering is also implemented via the sci-kit learn package. Euclidean distance and Ward linkage was used in the Agglomerative algorithm. The two re-sampling schemes used were a jackknife re-sampling (leave-one-dataset-out), and bootstrap re-sampling performed 500 times.

Both cluster algorithms require the number of clusters, k, as an input parameter. Multiple values of k are assessed via a stability analysis based on re-sampling (described in Supplementary Section S1.2), and the optimal k value was chosen as soon as the stability measure started to decrease. The stability measured used is the H-index, outline in section S1.2.

Projection of External Data

To project new data sets onto the atlas, we transform the data as previously described into percentile values. Only genes selected in the construction of the original atlas are retained. The original PCA defines the graph coordinates system defined by principal components. Each component is defined by a linear combination of genes, with each gene receiving a weight, also known as it's loading. Applying these coefficients to new data produces a coordinate in the PCA space for projection. The PCA and transformation is done with the scikit-learn (Pedregosa et al., 2011).

If genes are missing from the projection data, they are given the lowest rank. These missing genes often result from slightly different genome annotations: microarrays particularly suffer from outdated probe annotations resulting in absent or misrepresentation of genes used to construct the atlas. If a large proportion of genes are missing, this will distort the projection, thus it is advisable to use caution when applying old or uncommon microarray platforms built on outdated genome versions. Note that

Stemformatics workflows include alignment of microarray probes to the current genome version for gene annotation purposes.

Single cell RNAseq expression data

Single cell RNAseq expression data was sourced from Galen et al. (2019). A pseudo-bulk aggregation method was used to aggregate cells belonging to the same cluster, and where the cluster identity was taken from the original publication Galen et al. (2019). Each cluster was randomly divided into subgroups such that each projected 'sample' had the same number of cells within it, and transcript reads from these cells were pooled to create a single pseudo-bulk sample for that subgroup. The subgroup has the same identity as the original group, so might be expected to project into the same region of the atlas. These pseudo-bulk samples were projected onto the atlas in the same manner as described above.

Implementation and Code Availability

The Blood atlas and accompanying myeloid subset are available as interactive plots at www.stemformatics.org/atlas/blood and www.stemformatics.org/atlas/imac. These pages contain a number of features to help users navigate the atlas and perform useful functions:

- Interactive PCA with 3d/2d toggle.
- Colour by sample group, such as progenitor type or cell type.
- Show gene expression profile as a colour gradient.
- View gene expression and colour by sample group side by side.
- Project RNA-Seq dataset hosted at Stemformatics after a search.
- Project one's own dataset by providing expression and sample files as text files.
- Show and find which samples from which datasets make up the atlas.
- Download relevant data files (rank transformed expression and annotation tables) used by the atlas.
- Download plot in custom size.

Python code which can be used to manipulate the atlas data, to recreate the PCA for example, is available at https://bitbucket.org/stemformatics/s4m_pyramid/src/master/scripts/atlas.py.

RESULTS

Recursive Application and Unsupervised K-Means clustering upon the blood data

The method relies on several assumptions: (1) the biology of interest can be represented by the expression of many genes, (2) across platforms, some genes are measured less consistently than others, but there is a subset of genes where platform contributes substantially less to gene expression variability than the biology of interest, and therefore (3) the biology of a cell can be meaningfully described at several scales by identifying subsets of molecular attributes that are selected on cross-platform performance.

The method in its simplest implementation is agnostic to the presence of a biological signal or other confounding technical variables, but these can be subsequently applied to assess the major sample groupings. Here, 13661 genes common to 5 platforms were filtered for expression variance across 850 samples taken from 38 blood data sets. 3700 genes with low platform variance were subsequently used in a PCA to visualise the behaviour of samples relative to the platform or study that they were sampled from. The outcome of the these steps are shown in Figure 2, where each point represents one sample and the plots show cohesive grouping of similar cell types drawn from different platforms and independent studies. Supplementary Figure S3 shows that most genes retained in the atlas explain a high proportion of variance related to either sample source, progenitor type or cell type compared to platform.

At a global level, the PCA shows clear separation of progenitor types, lymphocytes, and myeloid lineages. The uniformity and stability of these sample groups was confirmed by K-means cluster analysis (see Supplementary Tables S2 and Table S1). The tables S2 shows results of the stability analysis performed over a range of k for the K-Means and Agglomerative algorithms. The most stable k, as measured by the median of the H-index of the clusters, is highlighted in yellow. In the top right hand column of 3 shows the most stable clustering on all of the blood (including myeloid, lymphocyte and progenitors) is run with $k = 6$. The annotated cell identities in Table S1 show that cluster 1, in the bottom corner, contains the progenitors, Cluster 2 captures lymphocytes and contains the majority of B, T and NK cells. The myeloid lineage is split over a three distinct clusters: Cluster 4 containing circulating monocytes and granulocytes, Cluster 5 predominantly cultured monocytes and tissue-resident macrophages, and Cluster 6 containing dendritic cells.

The large number of different myeloid cell types drives a resolution favouring these subsets. It follows that the resolution of biologically interesting subtypes requires representation from several data sets, and may not resolve if the major biological signal is driven by cell classes that are disproportionately represented. We address this using recursive application of the method, on subsets of samples captured in specific regions of the original graph. This allows for ever finer detail and identification of nuanced cell phenotypes, with the limiting factor being the availability of enough data for the biological subset of interest. By using a recursive approach, we view the atlas as a series of blood hierarchies, starting with

the most broad categorisation, and moving through smaller sample groupings to find more detailed cell types. For example, in order to resolve the lymphocytes better, the lymphocytes and myeloid arms were isolated from each other and the technique repeated on each. These separate graphs are shown in Figure 3. The PCA on the 728 myeloid samples is performed from the clusters containing progenitors, circulating monocytes and granulocytes, and dendritic cells (and approximately 3600 genes). We further observed differences between circulating and cultured monocytes, naive or activated states, and distinctions between primary or in vitro derived cells (Figure 3 Myeloid). In contrast the PCA on the lymphocytes only include 255 samples from clusters containing lymphocytes and progenitors. The lower number of samples makes it more difficult to resolve structure and results in only approximately 2400 genes being included. Despite the lower number of samples, Figure 3 Lymphocyte shows evident separation of the T and B Cells along the z-axis, as their difference is now strong enough to exceed the platform effect in our gene filter step, however the atlas lacks sufficient samples describing B-cell maturation or identifying phenotypically distinct T-cell classes. At each iteration, a robust global clustering is found for that scale, and only that scale. By stitching these together, the true multiscale nature of the myeloid arm of the blood hierarchy emerges, and more molecular detail is revealed. The two examples provided here can be further explored in the blood atlas (https://www.stemformatics.org/atlas/blood) and the myeloid subsets in the iMAC atlas (https://www.stemformatics.org/atlas/imac) (Rajab et al., 2019).

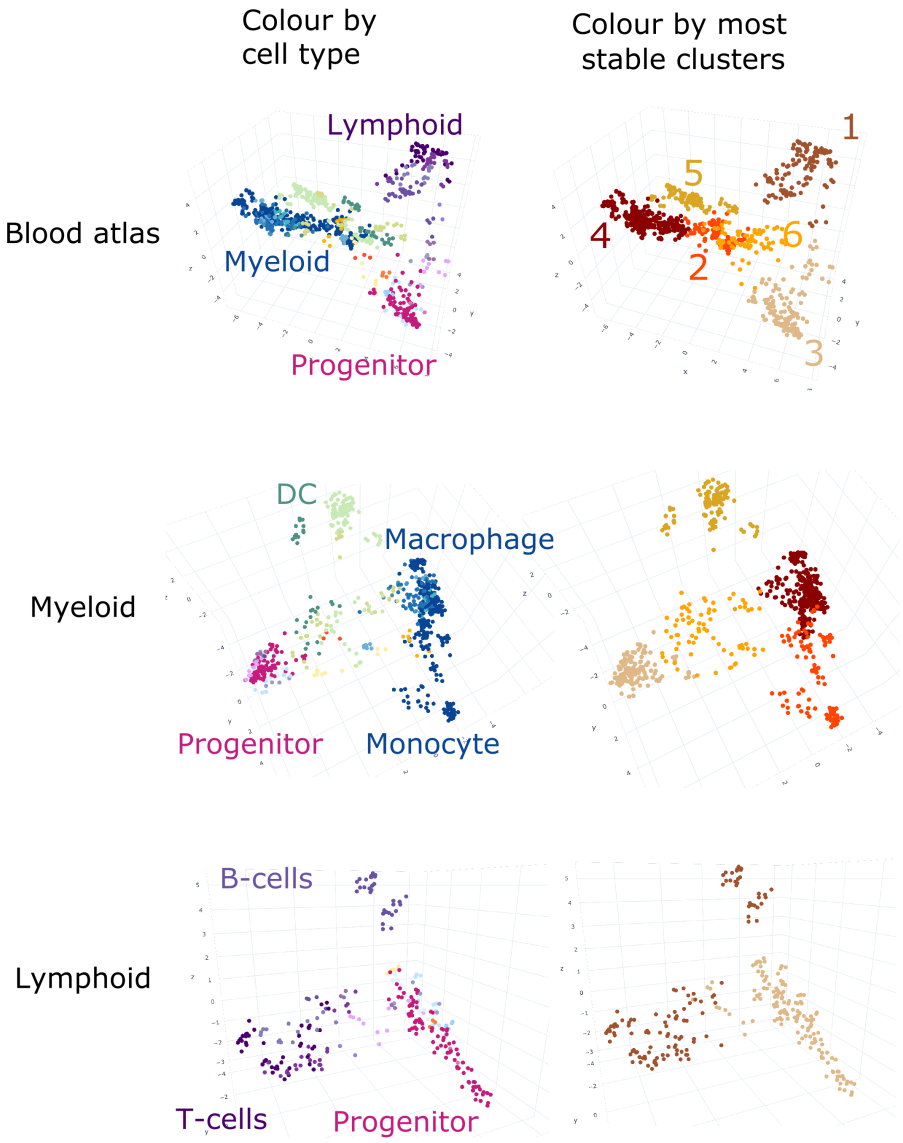


Figure 3. These panels show the results of repeated application of gene filtering and PCA upon the annotated blood samples in S4M. Each point is a sample, with colour indicative of annotated cell type (left column), or cluster identity (right column). The top left shows application to all blood samples as in Figure 2, while the top right shows the robust clusters defined upon these coordinates. The identities of these clusters are in Supplementary Table S1. Their highlighted colours are propagated to the middle and bottom panels to display the behaviour of these clusters subsequent to recursive application. The middle row shows the PCA when variance modelling and filtering is applied only to the myeloid lineage clusters (2,3,4,5 and 6). The myeloid PCA shows the clusters defining monocyte, macrophages and dendritic cells separating into distinct regions. The bottom row shows the variance modelling, filtering and PCA upon the lymphoid lineage clusters (1 and 3). Now the increased resolution splits the lymphocyte cluster, 1, into more detailed subsets containing either T or B cells.

Stability

While the biological grouping in the atlas are visually compelling, in order to formally test the stability of the results we ran two types of clustering algorithm on the merged data - the K-Means algorithm and the Agglomerative (bottom-up hierarchical) algorithm. For each approach, we perturbed the underlying data with two re-sampling schemes and measured the stability of cluster membership using the H-index of the Jaccard similarity coefficient (e.g. Shannon et al., 2016). Both algorithms, K-Means and Agglomerative, require the number of clusters as an input parameter, and as this number is not known a priori, we test multiple values. The Supplementary Tables S2 and S3 show the results of the jackknife and bootstrap resampling carried out upon clusters found in the respective atlas of each row in 3. These tables list the median $\pm$ the maximum and minimum of the H-Index calculated on all clusters after re-sampling. The results of jackknife and bootstrap resampling are qualitatively very similar.

Our clustering defines biological groups by assigning class membership to samples based on their proximity in PCA coordinates. This is our preferred measure of structure because the principal components can (and may be expected to) change under random re-sampling. Groupings of samples ought to be preserved, regardless of coordinate system, if indeed the biological signal is stable and the relative proximities are conserved. If the clustering structure is genuine, stability can be expected up until the point where too many clusters are demanded, after which clusters will be artificially grouped and unstable. We can also expect that the different algorithms should produce similar results if our atlas is stable.

In Table S2 top row, algorithms with cluster numbers up to 6 performed the best. For the both the K-Means and Agglomerative algorithms, the median values in this range are about $\sim$0.9. This indicates that when re-sampling, the overall structure of the atlas is well preserved. Results for re-sampling the myeloid and lymphoid arms are shown in the middle and bottom rows of Table S2. The myeloid atlas is stable up until having approximately 5 clusters, at which point they have a high median H-indices of $\sim$0.9. The lymphocyte atlas is most stable with 4 clusters, but only has median H-index of 0.79 (K-Means) and 0.75 (Agglomerative), and is less stable than other graphs for all of the cluster numbers. This reflects the relatively smaller representation of lymphocyte samples within our data.

We also evaluated the variation of the set of selected genes under re-sampling for the atlas containing all of the blood. Over the 500 bootstrapped iterations the median percentage of genes in common with the true data is 93% , with a minimum of 86% and maximum of %96. For the leave-one-out re-samplings the median similarity is 97%, with a maximum of 99% and minimum of 88%. These indicate the set of genes used to generate the Atlas is also stable to perturbations.

External data can be projected onto the atlas

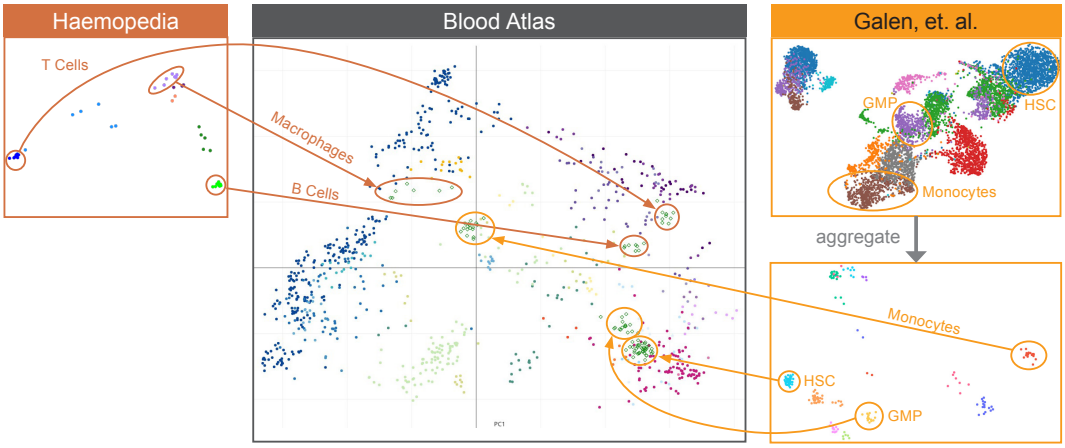


Figure 4. The projection of blood data from Haemosphere (https://www.haemosphere.org/) and Galen et al., 2019. The location of the projected data is consistent with the Blood Atlas

Allowing researchers to compare cell types is an important use-case for any robust transcriptional atlas which serves as a reference. This could be to validate or benchmark samples against the reference, to hypothesise about new cell types, or to find key regulators of differentiation. We have taken a simple approach of linearly projecting new data points onto the PCA space of the Blood Atlas for this type of comparison - as we allow users of the website to project their own data this way, a simple approach that users can easily understand is advantageous. We tested projections with a range of different datasets and data types that contain blood samples, to see if they produced expected results. The projections with bulk RNAseq datasets were highly reproducible for both myeloid and lymphocyte arms (Figure 4, Choi, Baldwin, et al. (2018)). For scRNA-Seq, simply projecting individual cells did not work very well, due to the fundamental difference in the distribution of values in the scRNAseq data compared to platforms. However, we have found that aggregating the samples to simulate pseudo-bulk samples did work well, as shown by Figure 4, (Galen et al., 2019).

For samples which are transcriptionally very different to blood cells, such as mesenchymal stromal cells, fibroblasts or neurons, projected coordinates are in the central region of the PCA (Figure S5). This region corresponds to coordinates where samples sit far away from all regions of the PCA. The web page contains some information about projections for the users, including a caution about interpreting a projection in this region, as well as information about the formats of files to use.

DISCUSSION

Gene filtering is an alternative to supervised normalisation

Transcriptional profiling was once a discovery platform used to find new molecules in well-established experimental systems (Graaf et al., 2016). It is also commonly used to assess cell composition of tissues, or benchmark new cell models by virtue of shared molecular patterns (Douvaras et al., 2017). Typically, researchers will "borrow" samples from other data sets, for example to benchmark or compare to their own experimental system to a previously published standard. Potential biases are introduced by the choice of reference, and this is further compounded by batch correction methods that require the analyst to make prior assumptions about the appropriate categorisation of samples. Methods that require prior determination of biological class force analysts to make a call about which variables are most important to promote or subtract, or how many biological classes are expected in the merged group. This may be desirable under some circumstances, but arguably less desirable when large data series are compiled, particularly if the normalisation approach inadvertently suppresses important variation across that data series. The addition of new data may require renormalisation of the entire series, limiting the number of comparisons. Without a standardised resource each study's comparator is different to the next, yet such approaches are expected to test the reproduciblity of individual studies. Here we show that the projection of new data onto a reference transcriptome atlas offers a straight forward solution.

In the example described here, the blood cell hierarchy, we demonstrate that when combining a large number of microarray and RNAseq data sets, a basic transformation and gene filter step is all that is required to extract prominent biological features. Supervised batch normalisation methods are very useful when applied to samples with well described properties, and when the split between sample class and technical batch is well balanced. Too often, however, batch and biology is confounded (reviewed in Leek, Scharpf, et al., 2010). Supervised normalisation seeks to rescue as many expression points (genes or probes) as possible, so applies a weighted adjustment across the entire gene set. Here we demonstrate that across dozens of data sets, representing hundreds of samples, the variability in gene expression attributable to platform profoundly impacts some, but not all genes. Therefore a weighted adjustment of expression where little prior batch effect is present has the potential to obscure genuine biology. Our approach does not seek to retain all expression measurements but rather constructs the atlas graph only with those expression values that escape a strong batch influence. This is achieved by taking many independent data sets, with minimal processing, to allow the dominant technical, experimental or biological trends to emerge from the combined data series. The resulting blood atlas demonstrably groups cells with common phenotypic attributes in an unbiased manner, and at several scales of resolution of cell type. Minimal processing also easily lends itself to an unsupervised method, which helps prevent over-fitting of sample classes or the biases associated with a restricted reference set.

An alternative method that is gaining attention in the integration of single cell datasets is canonical correlation analysis (CCA) (Butler et al., 2018). This method similarly uses a reduced feature selection approach, using only those variables (e.g. genes) that share a linear correlation structure across several data sets, to combine pairs of different experiments into an integrated series. CCA works best when there are a large number of data points in common between the samples to be combined. In contrast, we are combining many datasets of small sample size, such that any pair of individual datasets may lack overlapping cell classes, and in practise often are focused on one particular cell type, such as the humanised mouse models assessing tissue residency of dendritic cells (Haniffa et al., 2012) or an in-depth exploration of natural killer cell progenitors in fetal and adult tissues (Renoux et al., 2015). The correlation between genes and cell types is subsequently explored in using PCA.

The question of what is an appropriate normalisation must be assessed in light of the analysis question to be conducted. While there is clearly no 'one-size-fits-all' approach, we acknowledge that there are some limitations to our approach. Simplifying data on a ranked scale removes information about the scale of difference between two points. Consequently some information on gene-gene correlations is lost, although we do allow genes with the same value to keep the same rank. It is apparent from Supplementary section S1.3 that when combining data from platforms with different expression distributions, the benefits of performing the rank percentile transform outweighs the cons. By applying a gene filtering method, some biologically relevant genes will be removed from the analysis, and this may make it harder for a user to assess sample classes using a marker-based approach. By selecting genes with a low fraction of variance due to platform, we may we may lose resolution between some biological classes (insofar as variance indicates biological informativeness). Nevertheless we see in Figure 2 that enough information remains in order to extract a good deal of biological structure, and to find meaningful genes that are driving sample clustering. We also acknowledge that using PCA to review sample behaviour does not allow for examination of non linear relationships between genes or samples. Nevertheless, the advantages of rescaling and recursive filtering are clearly demonstrated here, and the resulting expression matrix would be suitable for other graphing or clustering approaches.

Recursive application to reveal fine-grained or coarse-grained atlas resolution

Blood is arguably the most accessible, and therefore the most comprehensively studied human tissue. The earliest attempts at finding unbiased molecular markers for different cell types came from the "Cluster of Differentiation" (CD) leukocyte markers (Bernard et al., 1984). In a community effort analogous to the atlas activities today, discovery of CD markers required over 50 laboratories undertaking an antibody screen against panels of blood cells without knowing what the antigen expressed by the cell is, nor what it does - the markers were adopted if they were able to reliably partition different cell types. CD markers are still used today - for example CD14 is a classical marker of monocytes and macrophages (Wright et al., 1990); CD4 and CD8 (Madakamutil et al., 2004; Sawada et al., 1994) have been adopted into

the naming convention of T-cell subsets. Nevertheless, very few of these markers are restricted to one cell type (Maddon et al., 1987), and more typically combinations of markers are required to categorise leukocytes.

Several atlas approaches were proposed to identify molecular markers of blood cell subtypes - these include the microarray profiles in Haematlas from the Bloodomics consortium (Watkins et al., 2009) and Haemopedia which compares RNAseq profiles between mouse and man (Choi, Baldwin, et al., 2018). While useful, most focus on profiling a small number of cell types in a large number of donors (e.g. QTL studies of monocyte gene expression) or a large number of well characterised cell types in a small number of donors (Novershtern et al., 2011; Prasad et al., 2014). However direct comparison between these projects is very difficult because of discrepancies in the way data is captured, and this is the problem addressed by the integrated atlas approach proposed here.

When considering what are the prominent biological features of any collection of data it is important to remember that 'prominent' is relative. The difference between lymphocytes and myeloid cells may be prominent when looking at blood cells only, but when compared to stem or stromal cells, it could be reasonably said that lymphoid and myeloid cells look very similar. The recursive approach is crucial to our analysis - at each level, the dominant global structure is retrieved and used to inform the next iteration, thus avoiding to impose global axis on all cells which may not reflect small scale structure. Therefore, the recursive approach is an intuitive way to map the cell landscape.

The multi-scaled nature of the biology highlights a second important limitation to our approach: the necessity for large amounts of diverse data, covering different cell types and experimental platforms. Subsampling regions of the atlas and applying a new round of gene filtering is a recursive approach that allows users to scale between global (all samples) or local cell comparisons. This extracts the most dominant structure at each resolution level, however with fewer samples we also approach the limits of our technique, and the results may become less robust. This can be observed in the lymphocyte arm of the atlas, which in the current iteration are represented by only a few data sets 2. The resolution of these cells is adequate at lineage level (B-cells vs T-cells) but with only 255 samples, it does not resolve subtypes of T Cells, such as CD4 or CD8. In contrast, resolution between different myeloid subsets is very high, and the emergent properties of the iMAC atlas highlight the impact of experimental handling or derivation method on the type of macrophage or DC studied.

Data projections and integration of single cell platforms

Given the advent and popularity now of single cell sequencing, future iterations will see the inclusion of single cell data. Deeper molecular characterisation of individual cells could be expected to better resolve functionally discrete populations, as well as provide new candidate markers for prospective cell isolation and characterisation. With the blood atlas method, we aim to provide a reference benchmark

that evaluates past transcriptomic data through a novel and relatively simple integration approach, and use this for comparisons to new data types, including scRNAseq of blood cells from different tissues. In the current iteration, we show the usefulness of projection of scRNA data onto the atlas, particularly for identification of blood cell types and annotation of scRNAseq clusters.

While we use the graph space obtained by the combined atlas series to project new data into the predefined state space, it's important to note that we are not using this to 'tune' new data sets into this space. Other graph smoothing methods have been described (Stuart et al., 2019; Weinreb et al., 2020), and particularly applied to the integration of single cell batches, where 'harmonisation' of the combined data is achieved by iterative weighting of gene expression in the introduced samples. Here, data set projections are used first to the reproduciblity of cell groups and group annotations using external, independent data. Secondly, projections of single cell expression data into prior annotated groups is used to lift atlas annotations over to the single cell experiment.

Projection strategies may provide additional benefits. Since we rank transform each sample before projection, each sample is treated independently to assess its similarity to the atlas cell types. Hence insights can be gained from data sets where batch effects are already confounding interpretation of data in the original experimental series - for example, where each sample class is obtained in a separate technical batch. In this instance, projection of each set of samples onto a reference atlas allows for examination of the experimental groups against an unbiased set of relevant cell classes. Projections may also inform trajectory analyses for scRNAseq datasets, without having to derive these trajectories de novo. For example, plotting single cell clusters from a differentiation series onto the blood atlas will allow better identification of haematopoietic cell lineages, or even suggest new pathways of differentiation, especially in cases where scRNAseq data come from cell types with low coverage within the Blood Atlas.

CONCLUSION

A shift from data collection on successive technologies, to integrated analyses across series of data offers an opportunity to view biological collections across a hierarchy of perspectives and information. In the example given here, we recapitulate the haemopoietic systems by combining 38 datasets, each describing detailed aspect of one part of that system in a small number of donors. The result is a multi-scaled tool to visualise and analyse the transcriptional relationships in the blood cell lineage. Recursive application of the method was demonstrated by the general categorisation seen in whole blood to the identification of specific myeloid cell types and activation states in the iMAC atlas. The projection of additional data onto the atlas, provides a tool for researchers to compare their own data to a robust reference collection. Projection of single cell data provides definitive annotations of blood cell clusters without prior assignment of marker genes in the scRNA-seq data. Implementation of the blood and iMAC atlases provides a simple web-based tool in the Stemformatics platform.

ACKNOWLEDGMENTS

The authors thank Matthew Rutar for early discussions on blood annotations, Jack Bransfield and Isha Nagpal for front end development of the Stemformatics server. Stemformatics was established through Australian Research Council Funding to Stem Cells Australia (SRI110001002)and to CAW (Future Fellowship FT150100330). KALC was supported by the National Health and Medical Research Council (NHMRC) Career Development fellowship (GNT1159458). PWA and JR are funded by NHMRC (GNT1181327) and (APP1186371) to CAW. NR is funded by the Centre for Stem Cell Systems and the CSIRO Synthetic Biology Future Science Platform. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

REFERENCES

Athar, Awais et al. (2018). "ArrayExpress update – from bulk to single-cell expression data". In: *Nucleic Acids Research* 47.D1, pp. D711–D715. ISSN: 0305-1048. DOI: 10.1093/nar/gky964. URL: https://doi.org/10.1093/nar/gky964.

Bernard, Alain et al. (1984). "Human Leucocyte Differentiation Antigens Detected by Monoclonal Antibodies". In:

Butler, Andrew et al. (2018). "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature Biotechnology* 36.5, pp. 411–420. ISSN: 1546-1696. DOI: 10.1038/nbt.4096. URL: https://doi.org/10.1038/nbt.4096.

Choi, Jarny, Tracey M Baldwin, et al. (2018). "Haemopedia RNA-seq: a database of gene expression during haematopoiesis in mice and humans". In: *Nucleic Acids Research* 47.D1, pp. D780–D785. ISSN: 0305-1048. DOI: 10.1093/nar/gky1020. URL: https://doi.org/10.1093/nar/gky1020.

Choi, Jarny, Chris M Pacheco, et al. (2018). "Stemformatics: visualize and download curated stem cell data". In: *Nucleic Acids Research* 47.D1, pp. D841–D846. ISSN: 0305-1048. DOI: 10.1093/nar/gky1064. URL: https://dx.doi.org/10.1093/nar/gky1064.

Cloonan, Nicole et al. (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing". In: *Nature Methods* 5.7, pp. 613–619. ISSN: 1548-7105. DOI: 10.1038/nmeth.1223. URL: https://doi.org/10.1038/nmeth.1223.

Douvaras, Panagiotis et al. (2017). "Directed Differentiation of Human Pluripotent Stem Cells to Microglia". eng. In: *Stem cell reports* 8.6, pp. 1516–1524. ISSN: 2213-6711. DOI: 10.1016/j.stemcr.2017.04.023. URL: https://pubmed.ncbi.nlm.nih.gov/28528700%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5470097/.

Eng, Chee-Huat Linus et al. (2019). "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: *Nature* 568.7751, pp. 235–239. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1049-y. URL: https://doi.org/10.1038/s41586-019-1049-y.

Forrest, Alistair R R et al. (2014). "A promoter-level mammalian expression atlas". In: *Nature* 507.7493, pp. 462–470. ISSN: 1476-4687. DOI: 10.1038/nature13182. URL: https://doi.org/10.1038/nature13182.

Frazee, Alyssa C et al. (2011). "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets". In: *BMC Bioinformatics* 12.1, p. 449. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-449. URL: https://doi.org/10.1186/1471-2105-12-449.

Gagnon-Bartsch, Johann A and Terence P Speed (2012). "Using control genes to correct for unwanted variation in microarray data". eng. In: *Biostatistics (Oxford, England)* 13.3, pp. 539–552. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxr034. URL: https://www.ncbi.nlm.nih.gov/pubmed/22101192%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3577104/.

Galen, Peter van et al. (2019). "Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity". In: *Cell* 176.6, 1265–1281.e24. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.01.031. URL: https://doi.org/10.1016/j.cell.2019.01.031.

Graaf, Carolyn A. de et al. (2016). "Haemopedia: An Expression Atlas of Murine Hematopoietic Cells". In: *Stem Cell Reports* 7.3, pp. 571–582. ISSN: 2213-6711. DOI: 10.1016/j.stemcr.2016.07.007. URL: https://doi.org/10.1016/j.stemcr.2016.07.007.

Haniffa, Muzlifah et al. (2012). "Human tissues contain CD141hi cross-presenting dendritic cells with functional homology to mouse CD103+ nonlymphoid dendritic cells". eng. In: *Immunity* 37.1, pp. 60–73. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2012.04.012. URL: https://pubmed.ncbi.nlm.nih.gov/22795876%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3476529/.

Hawrylycz, Michael J et al. (2012). "An anatomically comprehensive atlas of the adult human brain transcriptome". eng. In: *Nature* 489.7416, pp. 391–399. ISSN: 1476-4687. DOI: 10.1038/nature11405. URL: https://www.ncbi.nlm.nih.gov/pubmed/22996553%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243026/.

Hoffman, Gabriel E and Eric E Schadt (2016a). "variancePartition : interpreting drivers of variation in complex gene expression studies". In: *BMC Bioinformatics*, pp. 17–22. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1323-z. URL: http://dx.doi.org/10.1186/s12859-016-1323-z.

— (2016b). "variancePartition: interpreting drivers of variation in complex gene expression studies". In: *BMC bioinformatics* 17.1, p. 483.

Johnson, W Evan et al. (2006). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127. ISSN: 1465-4644. DOI: 10.1093/

biostatistics/kxj037. URL: https://doi.org/10.1093/biostatistics/kxj037.

Jones, Eric et al. (2001). *SciPy: Open source scientific tools for Python*. [Online; accessed ¡today¿]. URL: http://www.scipy.org/.

Law, Charity W et al. (2014). "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". eng. In: *Genome biology* 15.2, R29–R29. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-2-r29. URL: https://www.ncbi.nlm.nih.gov/pubmed/24485249%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053721/.

Leek, Jeffrey T, W Evan Johnson, et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics (Oxford, England)* 28.6, pp. 882–883. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts034. URL: http://europepmc.org/articles/PMC3307112.

Leek, Jeffrey T, Robert B Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data". In: *Nature Reviews Genetics* 11.10, pp. 733–739. ISSN: 1471-0064. DOI: 10.1038/nrg2825. URL: https://doi.org/10.1038/nrg2825.

Lizio, Marina et al. (2016). "Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals". In: *Nucleic Acids Research* 45.D1, pp. D737–D743. ISSN: 0305-1048. DOI: 10.1093/nar/gkw995. URL: https://doi.org/10.1093/nar/gkw995.

Madakamutil, Loui T et al. (2004). "CD8$\alpha\alpha$-Mediated Survival and Differentiation of CD8 Memory T Cell Precursors". In: *Science* 304.5670, 590 LP –593. DOI: 10.1126/science.1092316. URL: http://science.sciencemag.org/content/304/5670/590.abstract.

Maddon, P J et al. (1987). "Structure and expression of the human and mouse T4 genes". In: *Proceedings of the National Academy of Sciences* 84.24, pp. 9155–9159. ISSN: 0027-8424. DOI: 10.1073/pnas.84.24.9155. URL: https://www.pnas.org/content/84/24/9155.

Moon, Kevin R et al. (2017). "PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data". In: *bioRxiv*. DOI: 10.1101/120378. URL: https://www.biorxiv.org/content/early/2017/03/24/120378.

Novershtern, Noa et al. (2011). "Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis". In: *Cell* 144.2, pp. 296–309. ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.01.004. URL: https://doi.org/10.1016/j.cell.2011.01.004.

Papatheodorou, Irene et al. (2019). "Expression Atlas update: from tissues to single cells". In: *Nucleic Acids Research* 48.D1, pp. D77–D83. ISSN: 0305-1048. DOI: 10.1093/nar/gkz947. URL: https://doi.org/10.1093/nar/gkz947.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Prasad, Punit et al. (2014). "High-throughput transcription profiling identifies putative epigenetic regulators of hematopoiesis". In: *Blood*. ISSN: 0006-4971. DOI: 10.1182/blood-2013-02-483537.

eprint: http://www.bloodjournal.org/content/early/2014/02/27/blood-2013-02-483537.full.pdf. URL: http://www.bloodjournal.org/content/early/2014/02/27/blood-2013-02-483537.

Rajab, Nadia et al. (2019). "iMAC: An interactive atlas to explore phenotypic differences between ¡em¿in vivo¡/em¿, ex vivo and ¡em¿in vitro¡/em¿-derived myeloid cells in the Stemformatics platform". In: *bioRxiv*, p. 719237. DOI: 10.1101/719237. URL: http://biorxiv.org/content/early/2019/07/31/719237.abstract.

Regev, Aviv et al. (2017). "The Human Cell Atlas". In: *eLife* 6. ISSN: 2050-084X. DOI: 10.7554/eLife.27041. URL: http://europepmc.org/articles/PMC5762154.

Renoux, Virginie M et al. (2015). "Identification of a Human Natural Killer Cell Lineage-Restricted Progenitor in Fetal and Adult Tissues". eng. In: *Immunity* 43.2, pp. 394–407. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2015.07.011. URL: https://pubmed.ncbi.nlm.nih.gov/26287684.

Ritchie, Matthew E et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7, e47. DOI: 10.1093/nar/gkv007.

Rohart, Florian et al. (2016). "A molecular classification of human mesenchymal stromal cells". In: *PeerJ* 4, e1845. ISSN: 2167-8359. DOI: 10.7717/peerj.1845. URL: https://peerj.com/articles/1845.

Sawada, Shinichiro et al. (1994). "A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development". In: *Cell* 77.6, pp. 917–929. ISSN: 0092-8674. DOI: 10.1016/0092-8674(94)90140-6. URL: https://doi.org/10.1016/0092-8674(94)90140-6.

Schena, Mark et al. (1995). "Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray". In: *Science (New York, N.Y.)* 270, pp. 467–470. DOI: 10.1126/science.270.5235.467.

Shannon, Casey P et al. (2016). "SABRE: a method for assessing the stability of gene modules in complex tissues and subject populations". In: *BMC bioinformatics* 17.1, p. 460. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1319-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/27842512%20https://www.ncbi.nlm.nih.gov/pmc/PMC5109843/.

Stuart, Tim et al. (2019). "Comprehensive Integration of Single-Cell Data". In: *Cell* 177.7, 1888–1902.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.031. URL: https://doi.org/10.1016/j.cell.2019.05.031.

Stubbington, Michael J. T. et al. (2017). "Single-cell transcriptomics to explore the immune system in health and disease". In: *Science* 358.6359, pp. 58–63. ISSN: 0036-8075. DOI: 10.1126/science.aan6828. eprint: https://science.sciencemag.org/content/358/6359/58.full.pdf. URL: https://science.sciencemag.org/content/358/6359/58.

Taroni, Jaclyn N and Casey S Greene (2017). "Cross-Platform Normalization Enables Machine Learning Model Training On Microarray And RNA-Seq Data Simultaneously". In: *bioRxiv*, p. 118349. DOI: 10.1101/118349. URL: https://www.biorxiv.org/content/early/2017/03/21/118349.full.pdf+html.

Thompson, Jeffrey A et al. (2016). "Cross-platform normalization of microarray and RNA-seq data for machine learning applications". eng. In: *PeerJ* 4, e1621–e1621. ISSN: 2167-8359. DOI: 10.7717/peerj.1621. URL: https://www.ncbi.nlm.nih.gov/pubmed/26844019%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4736986/.

Watkins, Nicholas A. et al. (2009). "A HaemAtlas: characterizing gene expression in differentiated human blood cells". In: *Blood* 113.19, e1–e9. ISSN: 0006-4971. DOI: 10.1182/blood-2008-06-162958. eprint: http://www.bloodjournal.org/content/113/19/e1.full.pdf. URL: http://www.bloodjournal.org/content/113/19/e1.

Weinreb, Caleb et al. (2020). "Lineage tracing on transcriptional landscapes links state to fate during differentiation". In: *Science* 367.6479, eaaw3381. DOI: 10.1126/science.aaw3381. URL: http://science.sciencemag.org/content/367/6479/eaaw3381.abstract.

Wells, Christine A et al. (2013). "Stemformatics: Visualisation and sharing of stem cell gene expression". In: *Stem Cell Research* 10.3, pp. 387–395. ISSN: 1873-5061. DOI: https://doi.org/10.1016/j.scr.2012.12.003. URL: http://www.sciencedirect.com/science/article/pii/S1873506112001262.

Wright, S D et al. (1990). "CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein". In: *Science* 249.4975, 1431 LP –1433. DOI: 10.1126/science.1698311. URL: http://science.sciencemag.org/content/249/4975/1431.abstract.

S1 SUPPLEMENTARY METHOD

S1.1 Platform Effect Analysis and Gene Selection

S1.1.1 Choosing the platform-variance threshold

The appropriate value of the threshold is set by assessing the platform-batch effect after progressively lowering the threshold. A PCA using a threshold of 0.2 is shown in Figure S4 **D**. Compared to Figure S4 A the effect of platform is effectively suppressed and samples now cluster according to the annotated biological type. As the threshold is lowered, fewer genes pass the cut and are used to generate the PCA graphs. However, once platform is removed there is a range of thresholds for which the biological structure remains stable.

The effect of lowering the threshold is shown in Figures S4 and S2. Figure S4 depicts the changing PCA as the threshold is lowered from 0.8 down to 0.2. As this happens, it can be easily seen that the platforms, initially very separate in **A**, come to intermingle by **D**. This process is shown again in Figure S2. As we lower the threshold (moving down the y-axis), we measure the platform dependence of each of the first 10 PCA components independently using the Kruskal-Wallis H (KWH) Test (as implemented in Jones et al., 2001). This test quantifies the difference in median between the different populations. Darker boxes indicate less platform dependence, and vice-versa. Lowering the threshold has the effect of a) firstly moving the main platform dependence from component 1 to lower components, and b) eventually suppressing the effect over all components. This agrees with the visual inspection of the PCA after filtering genes with a range of thresholds. KWH Test values of $\sim$0.2 are present on the first 3 principal components when the threshold reaches ≤ 0.2. Empically, it is a good indicator that the platform effect is absent.

The platforms we use are Affymetrix HuGene and U133 Plus 2, RNA sequencing of any version, Illumina V4, Illumina V2 microarrays. These are the platforms with blood related datasets spanning several subtypes. Note that different versions of a platform, e.g. HuGene version 1 and 2, tend to cluster together as if they were one. The difference between them is small compared to the effect between the other platform types, and is also small compared to the effect of the biology. These differing versions are placed in the same category and treated as one platform. That leads to five platforms categories listed above.

S1.2 Stability and H-Index

Assessment of the stability of the gene selection using two random sampling techniques - bootstrap resampling directly upon the samples and leave-one-out resampling applied to datasets. Bootstrap resampling was performed 500 times. Multiple clustering methods were applied, in order to avoid results being dependent upon an idiosyncrasy of one algorithm. Our goal is not to assess clustering algorithms,

but rather provide evidence that the clustering structure present in the PCA is stable. Clustering is performing by KMeans and Hierarchical clustering implemented in cite Sklearn.

The H-Index (as described in (Shannon et al., 2016)) was calculated for each cluster, with the base case being the data with no resampling. At each iteration the process is repeated starting from the calculation of platform dependence: the univariate linear model calculated, thresholded with a value of 0.2 to select a subset of genes, PCA generated, and clustering algorithm applied. Tables S2 and S3 show the results for a selection of different clustering algorithms, and number of clusters. Results in the tables below show, for each combination of clustering algorithm and number of classes, the median H-index across all clusters, and the maximum and minimum of the H-index across all clusters.

We also can consider the stability of genes included under resampling. For each resample iteration, we calculated the percentage of genes that are still pass the cut.

Datasets are generated under different conditions and platforms, and in this way can be considered a single data point. Thus, resampling can be performed on either individual samples or datasets. This is further complicated as some datasets contribute a large number of samples. We list the results of leave-one-out resampling on datasets, followed by bootstrapping the samples, as bootstrapping datasets can have an extremely large effect on the underlying composition of the data.

Non-Blood Samples

Non-blood related samples were taken from publicly available data in Stemformatics. It was obtained by searching for sample types containing any one of the following terms: iPSC, embryonic, mesenchymal, mesoderm, fibroblast, pluripotent, neuron, astrocyte, adipocyte, melanocyte, epithelium, neural, endoderm, cardiomyocyte. The are a couple of iPSC derived blood datasets, which were excluded. The results of this search was 2093 samples from 140 datasets. The Stemformatics dataset id, the title of the relevant publication and the number of samples drawn from each dataset is listed in the table non_blood_datasets.tsv.

S1.3 Rank transformation corrects for platform variation between studies

The percentile rank transformation adopted in this study is advantageous as a non-parametric technique that does not require any specification of the potential batch sources or the experimental study design. This transformation is thus appropriate for any unknown confounding factor that was not reported in a given study, and can be considered as a normalisation technique in our atlas to enable between-study comparisons.

We simulated studies with 1000 samples, 10000 genes for 10 cell types and 4 batches. Amongst the 10000 genes, 100 to 200 genes were simulated as differentially expressed with a cell type effect. Count data were generated using negative binominal distributions with genewise specified dispersion trend described

by Law et al. (2014). Batch effects were then included in all genes by applying 4 different non-linear monotonic functions to the simulated data. Figure S6 shows that the percentile rank transformation outperforms two popular batch effect correction methods, limma and Combat (Ritchie et al., 2015; Johnson et al., 2006).

SUPPLEMENTARY FIGURES

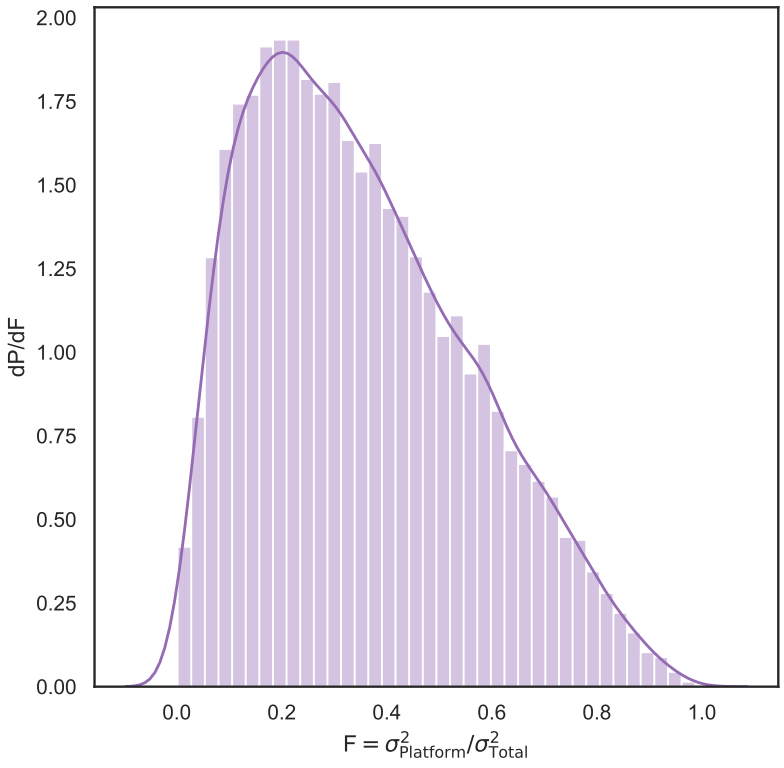


Figure S1. The distribution of the fraction of variance attributable to platform for the blood data. It is weighted towards low ratios, indicating that biological variation forms a major part of the signal.

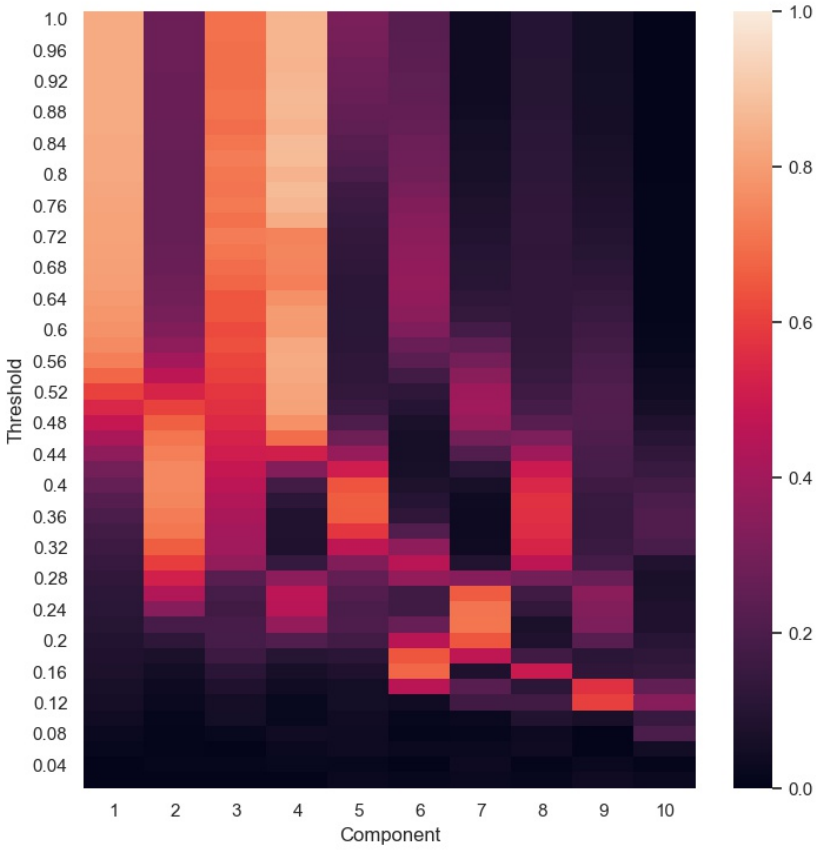
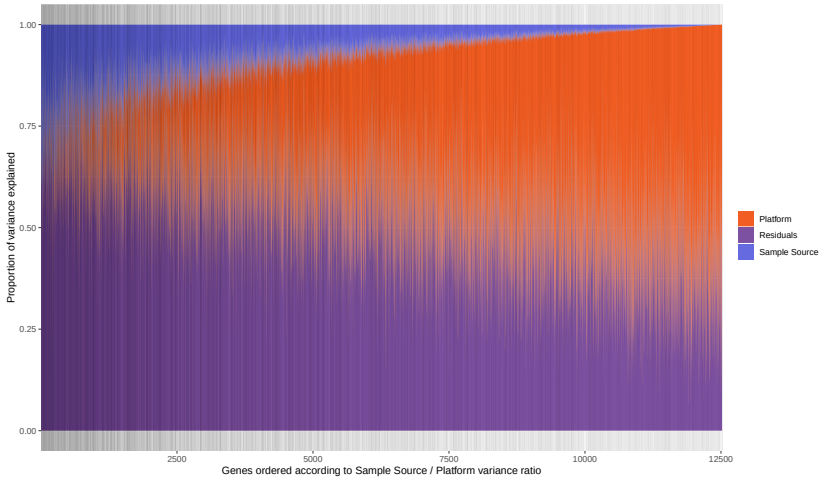
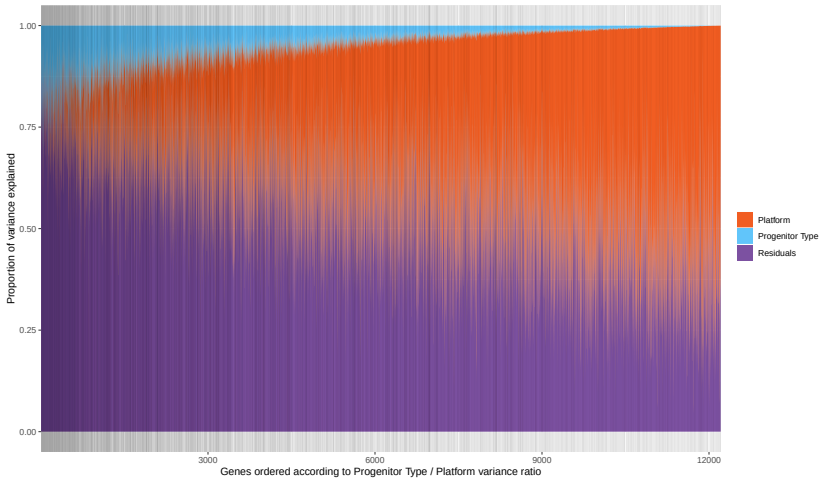


Figure S2. The change in value of the Kruskal-Wallis H Test for the first 10 components of the Blood PCA, as the platform-variance threshold is decreased. Lower values for the KWH test indicate less dependence of the component upon platform. Lowering the threshold has the effect of both moving platform-related components to lower components, and decreasing the overall dependence upon platform.

A



B



C

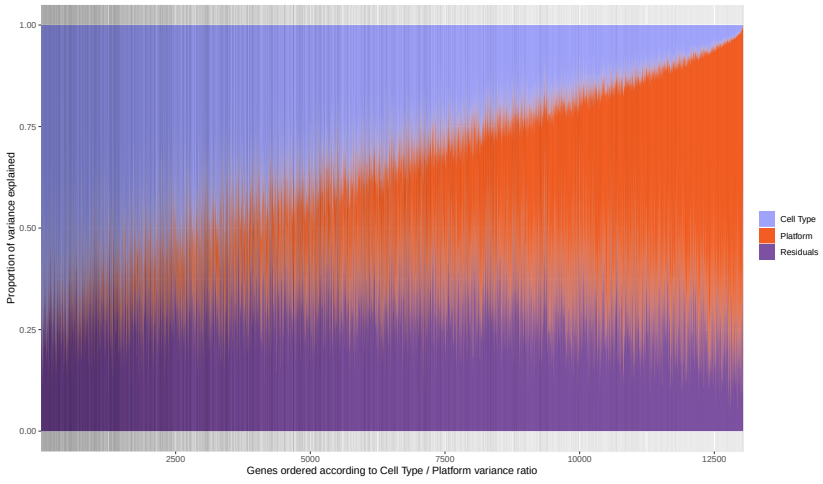


Figure S3. Proportion of variance explained by Platform, Residuals and either **A:** Sample Source, **B:** Progenitor type or **C:** Cell type assessed with a linear mixed model. Each gene is depicted as a vertical line on the x-axis, and genes are ranked according to the ratio Sample Source / Platform explained variance. Dark gray vertical lines indicate genes that were retained in the filtered data set.

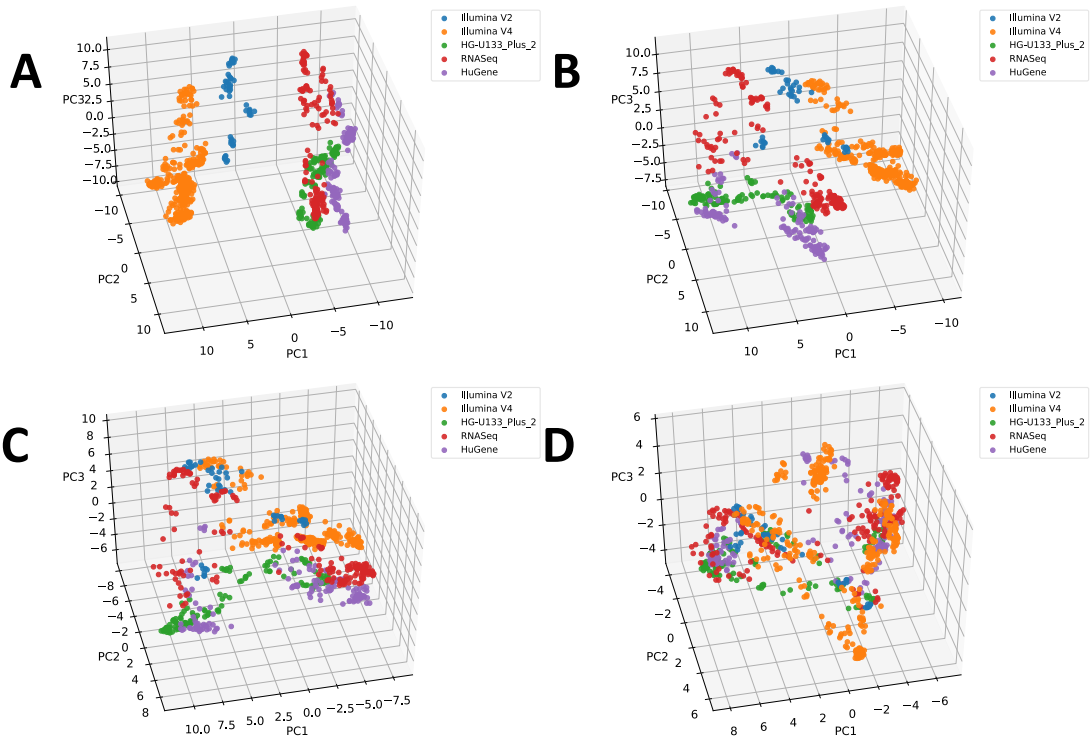


Figure S4. The PCA coordinates after filtering genes with a decreasing platform variance fraction threshold. A, the threshold is 0.8, B=0.6, C=0.4, D=0.2. As the threshold is lowered platform, initially separate, begin to merge.

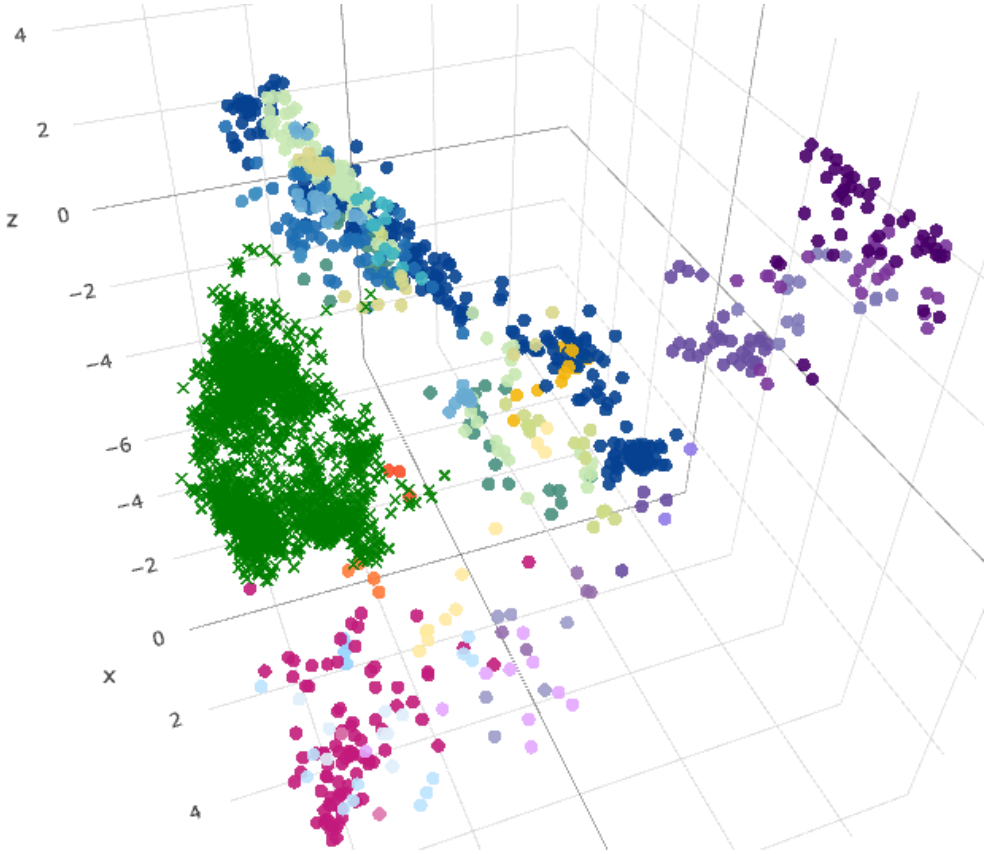


Figure S5. The projection of non-blood data (induced pluripotent stem cells, mesenchymal stem cells, fibroblasts, neurons) onto the Atlas. They are displayed as green crosses. They sit in a region low on component 2, a region not populated by either by the blood samples used to generate the atlas.

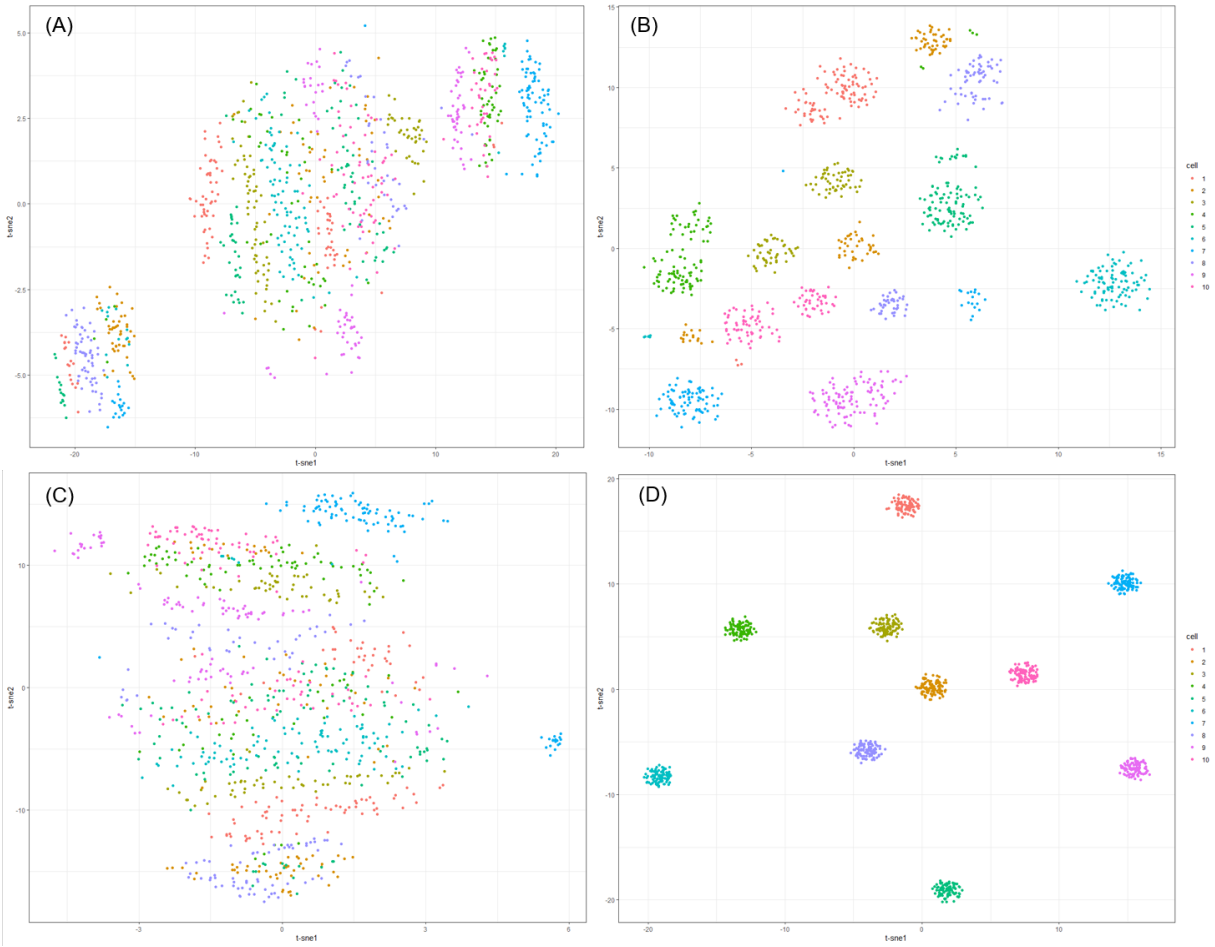


Figure S6. Comparisons of batch effect correction approaches on the simulated study described in Supplementary section S1.3. using t-SNE. Ten cell types are indicated by colors. **(A):** original count data include a batch effect across 4 platforms. **(B):** correction for platform effect with limma followed by voom transformation, **(C):** Combat and **(D):** percentile rank transformation.

SUPPLEMENTARY TABLES

Cluster Number	Cluster Name	Total Samples	Cell Type Identity
1	Lymphocyte cluster	120	B Cell (30/30), natural killer cell (15/15), T Cell (72/72), natural killer progenitor (3/4)
2	Circulating Monocyte and Granulocyte	126	monocyte (111/284), granulocyte (10/10), neutrophil (4/4), macrophage (1/104)
3	Progenitor	146	MK (4/4), erythrocyte (4/7), HPC (91/92), CMP (12/12) GMP (13/14), LP (22/25)
4	Macrophage	275	monocyte (173/284), macrophage (92/104), dendritic cell (10/172)
5	Dendritic Cell	106	dendritic cell (105/172), macrophage (1/104)
6	Mixed	77	dendritic cell (57/172), microglia (10/25), progenitor (10/96)

Table S1. List of clusters and annotations of the samples that belong to each clusters.

All Blood

N Clusters	K Means	Agglomerative
3	$0.73_{0.53}^{0.74}$	$0.74_{0.63}^{0.78}$
4	$0.83_{0.75}^{0.88}$	$0.75_{0.63}^{0.90}$
5	$0.88_{0.79}^{0.93}$	$0.88_{0.66}^{0.93}$
6	$0.91_{0.88}^{0.95}$	$0.90_{0.81}^{0.95}$
7	$0.90_{0.85}^{0.95}$	$0.85_{0.77}^{0.93}$
8	$0.83_{0.47}^{0.93}$	$0.82_{0.68}^{0.93}$
9	$0.83_{0.70}^{0.93}$	$0.80_{0.68}^{0.93}$
10	$0.81_{0.75}^{0.93}$	$0.79_{0.73}^{0.93}$

Myeloid

N Clusters	K Means	Agglomerative
3	$0.82_{0.82}^{0.85}$	$0.67_{0.54}^{0.69}$
4	$0.91_{0.90}^{0.92}$	$0.78_{0.67}^{0.92}$
5	$0.92_{0.89}^{0.93}$	$0.92_{0.87}^{0.94}$
6	$0.91_{0.86}^{0.93}$	$0.88_{0.68}^{0.94}$
7	$0.87_{0.87}^{0.93}$	$0.84_{0.69}^{0.94}$
8	$0.86_{0.72}^{0.94}$	$0.79_{0.64}^{0.94}$
9	$0.83_{0.56}^{0.94}$	$0.78_{0.34}^{0.92}$
10	$0.75_{0.60}^{0.90}$	$0.68_{0.3}^{0.88}$

Lymphoid

N Clusters	K Means	Agglomerative
3	$0.63_{0.61}^{0.76}$	$0.68_{0.68}^{0.75}$
4	$0.79_{0.75}^{0.84}$	$0.75_{0.69}^{0.84}$
5	$0.79_{0.63}^{0.79}$	$0.73_{0.47}^{0.79}$
6	$0.74_{0.63}^{0.84}$	$0.71_{0.54}^{0.79}$
7	$0.68_{0.63}^{0.84}$	$0.63_{0.41}^{0.82}$
8	$0.66_{0.63}^{0.84}$	$0.63_{0.47}^{0.82}$
9	$0.66_{0.35}^{0.84}$	$0.61_{0.47}^{0.79}$
10	$0.61_{0.37}^{0.79}$	$0.58_{0.46}^{0.79}$

Table S2. Results of the jackknife resampling stability analysis. Most stable number of clusters, the median H index, and their maximum/minimum H index as the superscript/subscript.

All Blood

N Clusters	K Means	Agglomerative
3	$0.59^{0.82}_{0.59}$	$0.71^{0.86}_{0.71}$
4	$0.89^{0.92}_{0.82}$	$0.85^{0.89}_{0.66}$
5	$0.92^{0.98}_{0.88}$	$0.90^{0.98}_{0.77}$
6	$0.97^{0.99}_{0.92}$	$0.96^{0.99}_{0.88}$
7	$0.96^{0.99}_{0.88}$	$0.90^{0.99}_{0.76}$
8	$0.82^{0.99}_{0.50}$	$0.80^{0.99}_{0.46}$
9	$0.82^{0.99}_{0.51}$	$0.74^{0.99}_{0.55}$
10	$0.85^{0.98}_{0.76}$	$0.79^{0.98}_{0.53}$

Myeloid

N Clusters	K Means	Agglomerative
3	$0.86^{0.88}_{0.77}$	$0.56^{0.86}_{0.48}$
4	$0.94^{0.96}_{0.91}$	$0.90^{0.98}_{0.80}$
5	$0.97^{0.99}_{0.92}$	$0.96^{0.99}_{0.88}$
6	$0.94^{0.99}_{0.88}$	$0.89^{0.99}_{0.74}$
7	$0.87^{0.99}_{0.84}$	$0.81^{0.98}_{0.71}$
8	$0.83^{0.95}_{0.59}$	$0.82^{0.94}_{0.54}$
9	$0.78^{0.96}_{0.57}$	$0.72^{0.96}_{0.47}$
10	$0.78^{0.97}_{0.6}$	$0.67^{0.96}_{0.49}$

Lymphoid

N Clusters	K Means	Agglomerative
3	$0.50^{0.91}_{0.47}$	$0.56^{0.90}_{0.55}$
4	$0.86^{0.86}_{0.84}$	$0.84^{0.90}_{0.75}$
5	$0.76^{0.89}_{0.58}$	$0.74^{0.91}_{0.60}$
6	$0.75^{0.92}_{0.65}$	$0.70^{0.93}_{0.65}$
7	$0.76^{0.95}_{0.50}$	$0.72^{0.95}_{0.50}$
8	$0.70^{0.96}_{0.54}$	$0.65^{0.95}_{0.58}$
9	$0.65^{0.94}_{0.50}$	$0.63^{0.95}_{0.38}$
10	$0.63^{0.91}_{0.51}$	$0.64^{0.93}_{0.45}$

Table S3. Results of the bootstrap resampling stability analysis. Most stable number of clusters, the median H index, and their maximum/minimum H index as the superscript/subscript.