Clustering RNA-seq expression data using grade of membership models

Kushal K Dey[1] Chiaowen Joyce Hsiao[2] Matthew Stephens[1,2]

[1] Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA; [2] Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA

Keywords: Admixture model, Grade of membership model, Latent Dirchlet Allocation, RNA-seq, single cell RNA-seq

Corresponding Author: Email mstephens@uchicago.edu; kkdey@uchicago.edu

Abstract

Grade of membership models (also known as "admixture models" or "Latent Dirichlet Allocation") are a generalization of cluster models that allow each sample to have membership in multiple clusters. These models are widely used in population genetics to model admixed individuals who have ancestry from multiple "populations", and in natural language processing to model documents having words from multiple "topics". Here we illustrate the potential for these models to cluster samples of RNA-seq gene expression data, measured on either bulk samples or single cells. The approach provides attractive visual summaries of the primary structure in several example data sets, and in our quantitative comparisons is more accurate than distance-based approaches in separating samples from different human tissues. We also provide methods to identify the genes that are most distinctively expressed in each cluster. The methods are implemented in an R package **CountClust**, available at `https://github.com/kkdey/CountClust`.

1 Introduction

Ever since large-scale gene expression measurements have been possible using micro-arrays, clustering – of both genes and samples – has played a major role in their analysis [1] [2] [3]. For example, clustering of genes can identify genes that are working together or co-regulated, and clustering of samples is useful for quality control as well as identifying biologically-distinct subgroups. A wide range of clustering methods have therefore been employed in this context, including distance-based hierarchical clustering, k-means clustering, and self-organizing maps (SOMs); see for example [4] [5] for reviews.

Here we focus specifically on cluster analysis of samples (as opposed to clustering of genes). Traditional clustering methods for this problem attempt to partition samples into distinct groups that show "similar" expression patterns. While partitioning samples in this way has intuitive appeal, it seems likely that the structure of a typical gene expression data set will be too complex to be fully captured by such a partitioning. Motivated by this, here we analyse expression data using grade of membership models [6], which generalize clustering models to allow each sample to have partial membership in multiple clusters. That is, they allow that each sample has a proportion, or "grade" of membership in each cluster. Such models are widely used in population genetics to model admixture, where individuals can have ancestry from multiple populations [12], and in document clustering ([27, 28]) where each document can have membership in multiple topics. In these fields the grade of membership models are often known as "admixture models", and "topic models" or "Latent Dirichlet Allocation" [27].

In the context of RNA-seq expression data, the grade of membership model allows that each sample has some proportion of its RNA-seq reads coming from each cluster. For typical bulk RNA-seq experiments this assumption could be motivated by a simple – or perhaps simplistic – argument: each sample is a mixture of different cell types, and so clusters could represent cell types, and the membership of a sample in each cluster could represent the proportions of each cell type present. This is similar to the idea of "deconvolution" methods that use cell-type-

specific expression profiles of marker genes to estimate the concentration of different cell types in a mixture [33]. And, indeed, the grade of membership model we use here is analogous to – although different in detail from – blind deconvolution approaches [31,32] which estimate cell type proportions and cell type signatures jointly (see also [29, 30] for semi-supervised approaches). However, we believe that the grade of membership model can be useful more generally to elucidate structure in expression data. For example, in single-cell expression data treating each sample as a "mixture of cell types" is clearly inappropriate, and yet we see value in the idea that there may be some "continuous" variation in cell types, rather than (or perhaps in addition to) the purely discrete variation captured by cluster models. Indeed, the extent to which variation among cells can be described in terms of discrete clusters vs more continuous populations seems a fundamental question that, when combined with appropriate single-cell RNA-seq data, the grade of membership models used here may ultimately help address. Further, even for bulk RNA-seq data, we argue that grade of membership models may yield interesting insights into heterogeneity among samples even if the inferred cluster membership do not correspond precisely to proportions of specific cell types, as may often happen in practice.

Interestingly, although we have not previously seen grade of membership models applied to RNA-seq data, several software packages for doing this already exist! [1] This is because the Latent Dirichlet Allocation model from [27], which is widely used to cluster documents based on their word counts, is based on a multinomial model that applies naturally and immediately to RNA-seq data. Whereas documents are characterized by counts of each possible word in a dictionary, RNA-seq samples are characterized by counts of reads mapping to each possible gene (or other unit, such as transcript, or exon) in the genome. Thus many software packages available for document clustering will also be applicable to RNA-seq data. Here we use the R package `maptpx` [11] to fit these models, and we add functionality for visualizing the results and annotating clusters by their most distinctive genes to help biological interpretation. These methods are implemented in the R package `count-clust` available from `https://github.com/kkdey/CountClust`.

2 Results

To briefly summarize the approach, assume RNA-seq data have been summarized by a table of counts $C_{N\times G} = (c_{ng})$, where c_{ng} is the number of reads from sample n mapped to gene (or transcript) g [10]. The grade of membership (GoM) model assumes that

$$c_{n\cdot} \sim \text{Mult}(c_{n+}, p_{n\cdot}) \tag{1}$$

where $c_{n\cdot}$ denotes the count vector for the nth sample, $c_{n+} := \sum_g c_{ng}$, and $p_{n\cdot}$ is a probability vector (non-negative entries summing to 1) whose gth element represents the relative expression

[1] While preparing this work for publication we became aware of recent independent work by [35] applying grade of membership models to RNA-seq data.

of gene g in sample n. The model further assumes that

$$p_{ng} = \sum_{k=1}^{K} q_{nk}\theta_{kg} \tag{2}$$

where $q_{n\cdot}$ is a probability vector whose kth element represents the grade of membership (or "membership proportion") of sample n in cluster k, and $\theta_{k\cdot}$ is a probability vector whose gth element represents the relative expression of gene g in cluster k. The number of clusters K is set by the analyst, and it can be helpful to explore multiple values of K (see Discussion).

Fitting this model (see Methods) results in estimated membership proportions q for each sample, and estimated expression values θ for each cluster. We visualize the membership proportions for each sample using a "Structure plot" [13], which is named for its widespread use in visualizing the results of the *Structure* software [12] in population genetics. The Structure plot represents the estimated membership proportions of each sample as a stacked barchart, with bars of different colors representing different clusters. Consequently samples that have similar membership proportions have similar amounts of each color. See Figure 1 for example.

2.1 Clustering human tissue samples using bulk RNA expression

We begin by illustrating the GoM model on bulk RNA expression measurements from the GTEx project (V6 dbGaP accession phs000424.v6.p1, release date: Oct 19, 2015, `http://www.gtexportal.org/home/`). These data consist of per-gene read counts from RNA-seq performed on 8555 samples collected from 450 human donors across 51 tissues, lymphoblastoid cell lines, and transformed fibroblast cell-lines. We analyzed $16,069$ genes that satisfied filters (e.g. exceeding certain minimum expression levels) that were used during eQTL analyses by the GTEx project (gene list available in `https://github.com/stephenslab/count-clustering/blob/master/project/utilities/gene_names_GTEX_V6.txt`).

To assess structure in these data we applied the GoM model with $K = 10, 12, 15$. Although results differ with K, many of the primary patterns were consistent across K. Here, for brevity, we focus on results for $K = 15$, shown as a Structure plot in **Figure 1(a)** (see also an alternative visualization using a 2-dimensional projection with t-sne [17], [18], in **Supplementary Figure 1** `http://stephenslab.github.io/count-clustering/project/src/tissues_tSNE.html`). Reassuringly, much of the structure highlighted by these results follows the known division of samples into tissues: that is, samples from the same tissue tend to have similar grades of membership across clusters. Some tissues are represented by essentially a single cluster (e.g. Pancreas, Whole Blood), whereas other tissues are represented as a mixture of multiple clusters (e.g. Thyroid). Furthermore, the results highlight biological similarity among some tissues by assigning samples from those tissues similar membership proportions. For example, samples from different parts of the brain have similar memberships, as do the arteries (aorta, tibial and coronary) and skin (sun-exposed and un-exposed). Samples from the tibial nerve have small but consistent amounts of membership in common with brain tissues, as well as larger

amounts in common with the adipose tissues. Indeed, many tissues show membership in cluster 1 (purple) or "Adipose" cluster, possibly reflecting, at least in some cases, contamination with adipose cells.

Each cluster is characterized by a vector that contains the mean expression level for each gene. To help biologically interpret each cluster we annotate it by identifying the genes whose expression levels most strongly distinguish that cluster from the others (see Methods). **Table 1** summarizes the results of this cluster annotation (top three genes in each cluster) for the GTEx analysis in Figure 1a. Again, reassuringly, many results align with known biology. For example, the top three genes driving the light violet cluster (cluster 13), which distinguishes Pancreas from other tissues, are *PRSS1*(protease serine 1), *CPA1* (carboxypeptidase) and *PNLIP* (pancreatic lipase), all of which are intimately involved in pancreatic function. Similarly, the top three genes driving the light green cluster (cluster 10), which distinguishes Whole Blood, are all hemoglobin genes, *HBB* (hemoglobin, beta), *HBA2* (hemoglobin, alpha 2) and *HBA1* (hemoglobin, alpha 1). Similarly, spermatogenesis and sperm-related genes characterize the Testis cluster, Keratin-related genes characterize the skin cluster, Myosin-related genes characterize the muscle skeletal cluster, etc. In cases where a cluster occurs in multiple tissues these annotations may be particularly helpful for understanding what may be driving this co-membership. For example, the top three genes in the light orange cluster (cluster 8), which is common to Lung, Spleen and Small Intestine - Terminal Ileum, all code for surfactant proteins (specifically, B, A2 and A1).

Although global analysis of all tissues is useful for highlighting major structure in the data, it may miss finer-scale structure within tissues or among similar tissues. For example, here the global analysis allocated similar cluster memberships to all brain tissues, and we suspected that these tissues may exhibit substructure that could be uncovered by analyzing the brain samples separately. **Figure 1(b)** shows the Structure plot for $K = 4$ on only the Brain samples. The results highlight much finer-scale structure compared with the global analysis. Brain Cerebellum and Cerebellar hemisphere are essentially assigned to a separate cluster, whose top 3 defining genes are SNAP25 (synaptosomal-associated protein, 25kDa), ENO2 (enolase 2- gamma, neuronal) and CHGB (chromogranin B), all of which are associated with neuronal activities (**Supplementary Table 1**). This cluster seems likely to reflect the expected high concentration of Cerebellar granule cells in cerebellar samples. . The spinal cord samples also show consistently strong membership in a single cluster, whose top 3 defining genes are MBP (involved in myelination [34]) and MYH11 and ACTA2, both of which play a key role in contraction of smooth muscle cells. (Genetic mutations in both MYH11 and ACTA2 can cause Thoracic aortic aneurysms/dissection [36].) The remaining samples all show membership in multiple clusters, with cortex samples being distinguished from other samples by stronger membership in a cluster (blue in figure) whose distinctive genes include UCHL1, which is expressed in neurons and is required for normal synaptic and cognitive function [37].

2.2 Quantitative comparison with hierarchical clustering

The GoM model is, in many ways, complimentary to, rather than only competing with, distance-based hierarchical clustering methods. Nonetheless, we hypothesized that the model-based nature of the GoM approach might also provide greater accuracy in detecting substructure than distance-based methods. We used the GTEx data to test this hypothesis. Specifically, for each pair of tissues in the GTEx data we assessed whether or not each clustering method correctly partitioned samples into the two tissue groups (see Methods). The GoM model was substantially more accurate in this test, succeeding in 86% of comparisons, compared with 39% for the distance-based method; Figure 2.

2.3 Clustering of single-cell RNA-seq data

Recently RNA-sequencing has become viable for single cells [7], and this technology has the promise to revolutionize understanding of intra-cellular variation in expression, and regulation more generally [8]. Although it is traditional to describe and categorize cells in terms of distinct cell-types, the actual architecture of cell heterogeneity may be more complex, and in some cases perhaps better captured by the more "continuous" GoM model. In this section we aim to illustrate the potential for the GoM model to be applied to single cell data.

Single-cell RNA-seq data typically involve substantially lower effective sequencing depth compared with bulk experiments, due to the lower number of molecules available to sequence in a single cell. To check robustness of the GoM model to lower sequencing depth we repeated analyses above after thinning the GTEx data by a factor of $10,000$ to mimic the lower sequencing depth of a typical single cell experiment. For the thinned GTEx data the Structure plot for $K = 15$ preserves most of the major features of the original analysis on unthinned data (**Supplementary Figure 6**). For the accuracy comparisons with distance-based methods, both methods suffer reduced accuracy in thinned data, but the model-based method retains its superior performance. For example, when thinning by a factor of 1,000 the success proportion in separating tissues is 0.10 for hierarchical clustering and 0.32 for GoM model.

Now we apply the GoM models to two single cell RNA-seq datasets, from Jaitin *et al* [19] and Deng *et al* [20].

Jaitin *et al* sequenced over 4000 single cells from mouse spleen. Following the original authors protocol, we also filtered out 16 genes that they found to show significant batch-specific expression. Here we analyze 1041 of these cells that were categorized as $CD11c+$ in the *sorting markers* column of their data (`http://compgenomics.weizmann.ac.il/tanay/?page_id=519`), and which had total number of reads mapping to non-ERCC genes greater than 600. (We believe these cells correspond roughly to the 1040 cells in their Figure S7.) Our hope was that applying our method to these data would identify, and perhaps refine, the cluster structure evident in [19] (their Figures 2A and 2B). However, our method yielded rather different results (Figure 3), where most cells were assigned to have membership in several clusters. Further, the cluster membership

vectors showed systematic differences among amplification batches (which in these data is also strongly correlated with sequencing batch). For example, cells in batch 1 are characterized by strong membership in the orange cluster (cluster 5) while those in batch 4 are characterized by strong membership in both the blue and yellow clusters (2 and 6). Some adjacent batches show similar patterns - for example batches 28 and 29 have a similar visual "palette", as do batches 32-45. The fact that batch effects are detectable in data like these is not particularly surprising. There is a growing recognition of the importance of batch effects in high-throughput data generally [23] and in single cell data specifically [24]. And indeed, dimension reduction methods such as the ones we use here can be helpful in controlling for such effects [21] [22]. However, why these batch effects are not evident in Figures 2A and 2B of [19] is unclear to us.

Deng *et al* collected expression data from individual cells from zygote to blastocyst stages of mouse preimplantation development [20]. Deng *et al*'s analysis focussed particularly on allele-specific expression from the two contributing mouse strains (CAST/EiJ and C57BL/6J). Here we analyze the counts of the two alleles combined. Visual inspection of the Principal Components Analysis in [20] suggested 6-7 clusters, so we fit the cluster model with $K = 6$. The results (Figure 4) clearly highlight the structure in the different development stages starting from zygote, through early/mid/late 2 cells, 4 cells, 8 cells, 16 cells, and early/mid blastocyst to finally late blastocyst. Specifically, cells that are from the same stage show similar cluster membership proportions. Further, many of the clusters show notable trends through the stages. For example, membership in the green cluster is non-existent in early stages, starts in the 4-cell stage, becomes more prominent in the 8-16 cell stages, drops substantially in the early and mid-blastocyte stages, and is essentially absent in the late blastocytes. More generally, cluster memberships for cells from adjacent stages tend to be more similar to one another than those for cells from distant stages.

Examining the clustering results by embryo highlights apparent embryo-level effects in the early stages (Figure 4): that is, cells from the same embryo sometimes showed distinctive differences from other embryos. For example, the two cells from one of the 2-cell embryos (check) shows much stronger membership in the magenta cluster than other 2-cell embryos, and four cells from one of the 4-cell embryos (embryo 4) shows consistently more yellow membership than the other 4-cell embryos.

Finally, the results indicate a few samples that appear to be outliers - for example, a cell from a 16-cell embryo that looks like a very early stage cell (zygote or early 2-cell), and a cell from an 8-stage embryo that looks rather different from any of the others.

Notably, for both these single-cell data sets, most cells are assigned to a combination of more than one cluster, rather than a single cluster (the exception being the very early-stage cells in data from Deng et al). This highlights the potential utility for GoM models to capture structure in single cell data that might be missed by simpler cluster-based approaches.

The codes and scripts for reproducing results in this paper are available from `http://stephenslab.github.io/count-clustering/`.

3 Discussion

Our goal here is to highlight the potential for model-based clustering methods, and particularly GoM models, to elucidate structure in RNA-seq data from both single cell sequencing and bulk sequencing of pooled cells. We also provide tools to identify which genes are most distinctively expressed in each cluster, to aid interpretation of results. As our applications to the GTEx data illustrate, these methods have the potential to highlight biological processes underlying the cluster structure identified.

The GoM model has several advantages over distance-based hierarchical methods of clustering. At the most basic level model-based methods are often more accurate than distance-based methods. Indeed, in our simple test on the GTEx data the model-based GoM approach more accurately separated samples into "known" clusters. However, there are also other subtler benefits of the GoM model. Because the GoM model does not assume a strict "discrete cluster" structure, but rather allows that each sample has a proportion of membership in each cluster, it can provide insights into how well a particular dataset really fits a "discrete cluster" model. For example, consider our results for the data from [19] and [20]: in both cases most samples are assigned to multiple clusters, although the results are closer to "discrete" for the latter than the former. The GoM model is also better able to represent situation where there is not really a single clustering of the samples, but where samples may cluster differently at different genes. For example, in the GTEx data, the lung samples share memberships in common with both the spleen and adipose-related tissues. This pattern is clearly visible in the Structure plot (Figure 1) but would be hard to discern from a standard hierarchical clustering.

GoM models also have close connections with dimension reduction techniques such as factor analysis, principal components analysis and non-negative matrix factorization. All of these methods can also be used for RNA-seq data, and may often be useful. See [16] for discussion of relationships among these methods in the context of inferring population genetic structure. While not arguing that the GoM model is uniformly superior to these other methods, we believe our examples illustrate the appeals of the approach. In particular, we would argue that for the GTEx data, the Structure plot (Figure 1) combined with the cluster annotations (Table 1) provide a more visually and biologically appealing summary of the data than would a principal components analysis.

Fitting GoM models can be computationally-intensive for large data sets. For the datasets we considered here the computation time ranged from 12 minutes for the data from [20] ($n = 259; K = 6$), through 33 minutes for the data from [19] ($n = 1041; K = 7$) to 3,297 minutes for the GTEx data ($n = 8,555; K = 15$). Computation time can be reduced by fitting the model to only the most highly expressed genes, and we often use this strategy to get quick initial results for a dataset. Because these methods are widely used for clustering very large document datasets there is considerable ongoing interest in computational speed-ups for very large datasets, with "on-line" approaches capable of dealing with millions of documents [38] that could be useful in the future for very large RNA-seq datasets.

A thorny issue that arises when fitting these types of model is how to select the number of clusters, K. Like many software packages for fitting these models, the `maptpx` package implements a measure of model fit that provides one useful guide. However, it is worth remembering that in practice there is unlikely to be a "true" value of K, and results from different values of K may complement one another rather than merely competing with one another. For example, seeing how the fitted model evolves as K increases is one way to capture some notion of hierarchy in the clusters identified [13]. More generally it is often fruitful to analyse data in multiple ways using the same tool: for example our GTEx analyses illustrate how analysis of subsets of the data (in this case the brain samples) can complement analyses of the entire data.

The version of the GoM model fitted here is relatively simple, and could certainly be embellished. For example, the model allows the expression of each gene in each cluster to be a free parameter, whereas we might expect expression of most genes to be "similar" across clusters. This is analogous to the idea in population genetics applications that allele frequencies in different populations may be similar to one another [15], or in document clustering applications that most words may not differ appreciably in frequency in different topics. In population genetics applications incorporating this idea into the model, by using a correlated prior distribution on these frequencies, can help improve identification of subtle structure [15] and we would expect the same to happen here for RNA-seq data.

4 Methods and Materials

4.1 Model overview

We assume the RNA-seq data have been summarized by a table of counts $C_{N\times G} = (c_{ng})$, where c_{ng} is the number of reads from sample n mapped to gene (or transcript) g [10]. We remove genes g with all zero counts ($c_{ng} = 0$ for all n), and use the `maptpx` R package [11] to fit the grade of membership (GoM) model, also known as "Latent Dirichlet Allocation" (LDA). This model assumes the RNA-seq counts for each sample follow a multinomial distribution

$$c_{n\cdot} \sim \mathrm{Mult}(c_{n+}, p_{n\cdot}) \tag{3}$$

where $c_{n\cdot}$ denotes the count vector for the nth sample, $c_{n+} := \sum_g c_{ng}$, and $p_{n\cdot}$ is a probability vector (non-negative entries summing to 1) whose gth element represents the relative expression of gene g in sample n. The model further assumes that

$$p_{ng} = \sum_{k=1}^{K} q_{nk}\theta_{kg} \tag{4}$$

where $q_{n\cdot}$ is a probability vector whose kth element represents the grade of membership of sample n in cluster k, and $\theta_{k\cdot}$ is a probability vector whose gth element represents the relative expression of gene g in cluster k. The `maptpx` package fits this model using an EM algorithm to perform Maximum a posteriori (MAP) estimation of the parameters q and θ. See [11] for details.

4.2 Visualizing Results

In addition to the Structure plot, we have also found it useful to visualize results using t-distributed Stochastic Neighbor Embedding (t-SNE), which is a method for visualizing high dimensional datasets by placing them in a two dimensional space, attempting to preserve the relative distance between nearby samples [17,18]. Compared with the Structure plot our t-SNE plots contain less information, but can better emphasise clustering of samples that have similar membership proportions in many clusters. Specifically, t-SNE tends to place samples with similar membership proportions together in the two-dimensional plot, forming visual "clusters" that can be identified by eye (e.g. Supplementary Figure 1). This may be particularly helpful in settings where no external information is available to aid in making an informative Structure plot.

4.3 Cluster annotation

To help biologically interpret the clusters, we developed a method to identify which genes are most distinctively differentially expressed in each cluster. (This is analogous to identifying "ancestry informative markers" in population genetics applications [**?**].) Specifically, for each cluster k we measure the distinctiveness of gene g with respect to any other cluster l using

$$\mathrm{KL}^g[k,l] := \theta_{kg}\, log\frac{\theta_{kg}}{\theta_{lg}} + \theta_{lg} - \theta_{kg}, \tag{5}$$

which is the Kullback–Leibler divergence of the Poisson distribution with parameter θ_{kg} to the Poisson distribution with parameter θ_{lg}. For each cluster k, we then define the distinctiveness of gene g as

$$D^g[k] = \min_{l \neq k} \mathrm{KL}^g[k,l]. \tag{6}$$

The higher $D^g[k]$, the larger the role of gene g in distinguishing cluster k from all other clusters. Thus, for each cluster k we identify the genes with highest $D^g[k]$ as the genes driving the cluster k. We annotate the genes driving each cluster with biological functions using the `mygene` R Bioconductor package [25].

4.4 Comparison with hierarchical clustering

Distance based hierarchical clustering methods are the most commonly used clustering techniques for gene expression data. To compare between the grade of membership model and the distance based hierarchical clustering algorithm, we used both methods to samples from pairs of tissues from the GTEx project, and assessed which methods separated samples according to tissue. For each pair of tissues we randomly selected 50 samples from the pool of all samples coming from these tissues. For the hierarchical clustering approach we cut the dendogram at $K = 2$, and checked whether or not this cut partitions the samples into the two tissue groups.

(We applied hierarchical clustering using Euclidean distance, with both complete and average linkage; results were similar and so we showed results only for complete linkage.) For the model-based approach we analysed the data with $K = 2$, and sort the samples by their membership in cluster 1. We then partitioned the samples at the point of the steepest fall in this membership, and again we check whether this cut partitions the samples into the two tissue groups.

Figure 2 shows, for each pair of tissues, whether each method successfully partitioned the samples into the two tissue groups using these approaches.

4.5 Thinning

We used "thinning" to simulate lower-coverage data from the original higher-coverage data.. Specifically, if c_{ng} is the counts of number of reads mapping to gene g for sample n for the original data, we simulated thinned counts t_{ng} using

$$t_{ng} \sim Bin(c_{ng}, p_{thin}) \tag{7}$$

where we used p_{thin} is a specified thinning parameter.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 10/19/2015 and dbGaP accession number phs000424.v6.p1.

The authors also thank Matt Taddy, Amos Tanay and Effi Kenigsberg for helpful discussions.

Disclosure Declaration

The authors have no conflict of interest.

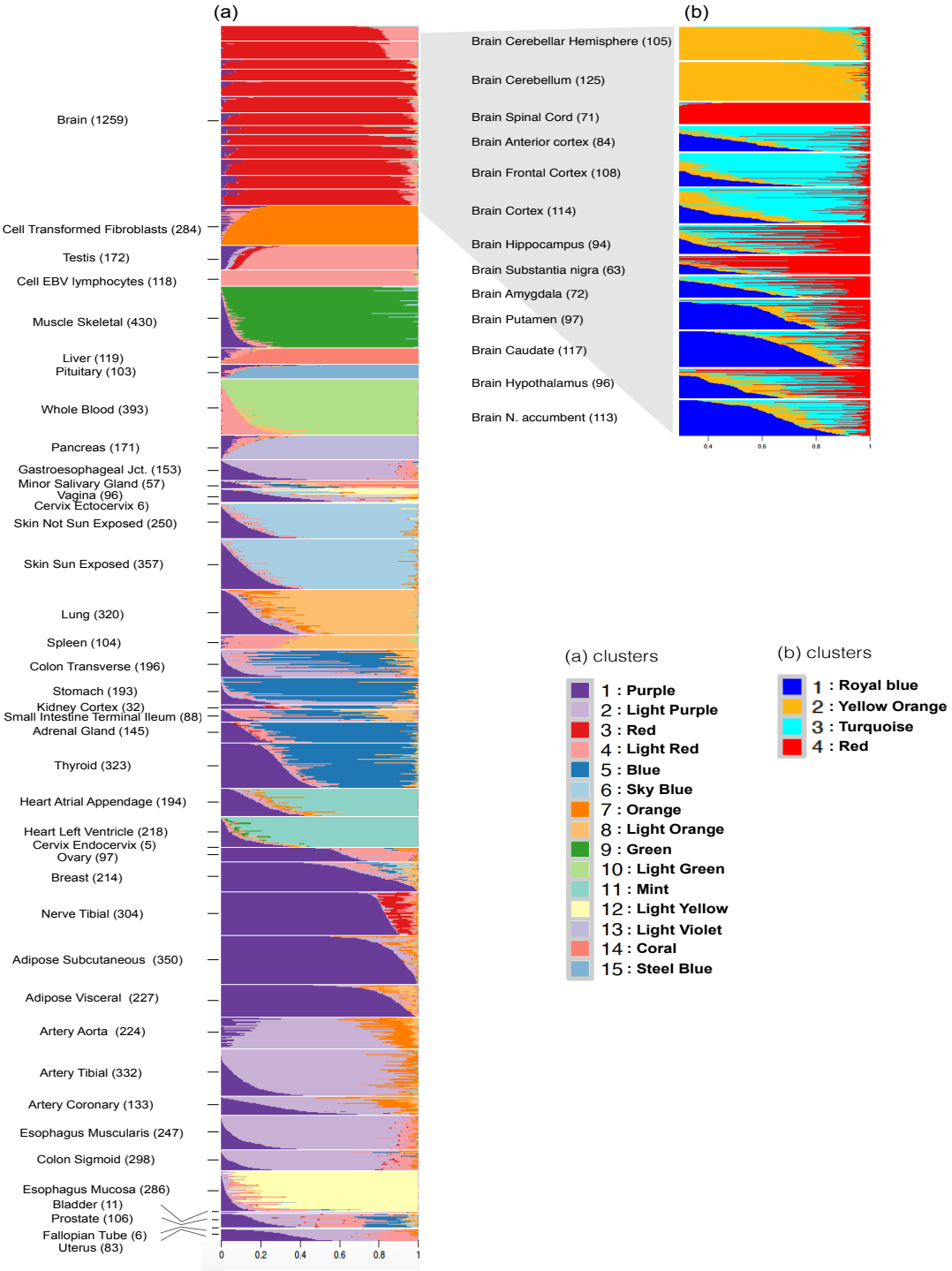
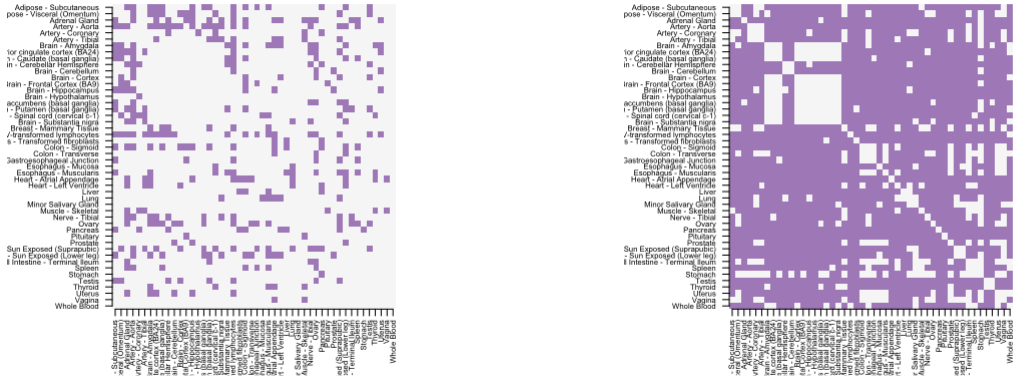


Figure 1. **(a)**: Structure plot of estimated cluster membership proportions for $K = 15$ clusters fit to 8555 tissue samples from 53 tissues in GTEx data. Each vertical bar shows the cluster membership proportions for a single sample, ordered so that samples from the same tissue are adjacent to one another. **(b)**: Structure plot of estimated cluster membership proportions for $K = 4$ clusters fit to only the brain tissue samples from GTEx. This analysis highlights finer-scale structure among the brain samples that is absent from (a).



(a) hierarchy method

(b) GoM method

Figure 2. A comparison of "accuracy" of hierarchical vs model-based clustering. For each pair of tissues from the GTEX data we assessed whether or not each clustering method (with $K = 2$ clusters) separated the samples according to their actual tissue of origin, with successful separation indicated by a filled square. Some pairs of tissues (e.g. pairs of brain tissues) are more difficult to distinguish than others. Overall the model-based clustering is successful in 86% comparisons and the hierarchical clustering in 39% comparisons.

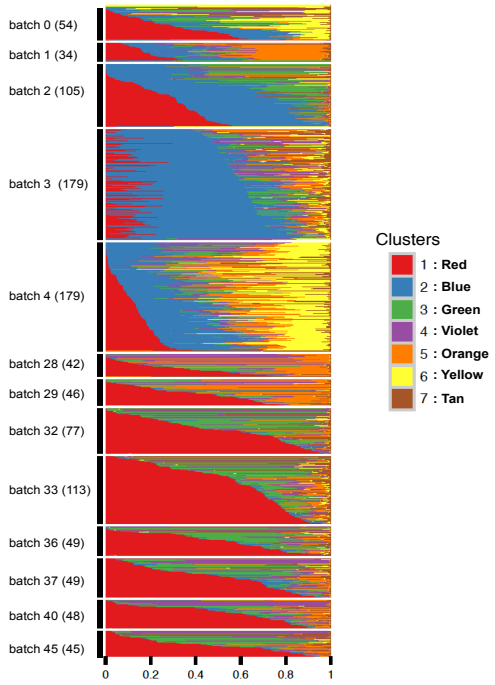


Figure 3. Structure plot of estimated cluster membership proportions for $K = 7$ clusters fit to 1041 single cells from [19]. The samples are ordered so that samples from the same amplification batch are adjacent. In this analysis the samples do not appear to form clearly-defined clusters, with each sample being allocated membership in several "clusters". Visually, samples from the same amplification batch tend to be assigned similar membership proportions, suggesting that batch effects are likely contributing to the inferred clustering.

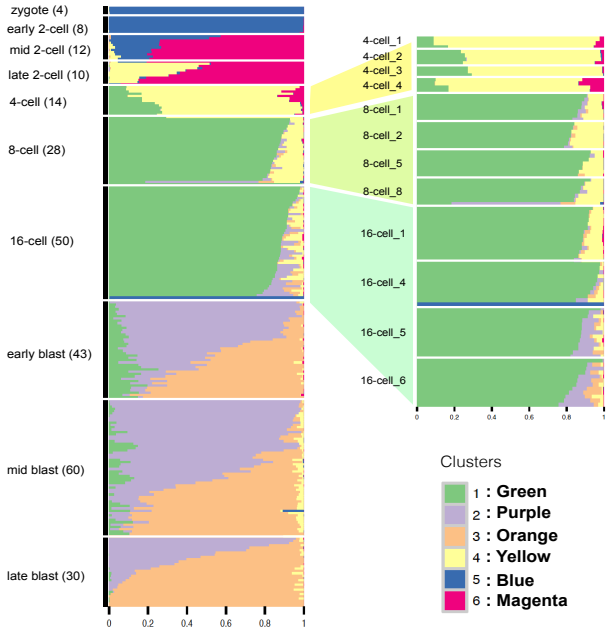


Figure 4. Structure plot of estimated cluster membership proportions for $K = 6$ clusters fit to 259 single cells from [20]. The cells are ordered by their preimplantation development phase (and within each phase, arranged in the same order as in the data file). While the very earliest developmental phases (*zygote/early2cell*) are essentially assigned to a single cluster, others are represented as a mix of two or more clusters. This illustrates the idea that structure of single-cell data may in some cases be better captured by a mixed membership model than by simple discrete clusters.

4.6 Supplemental figures

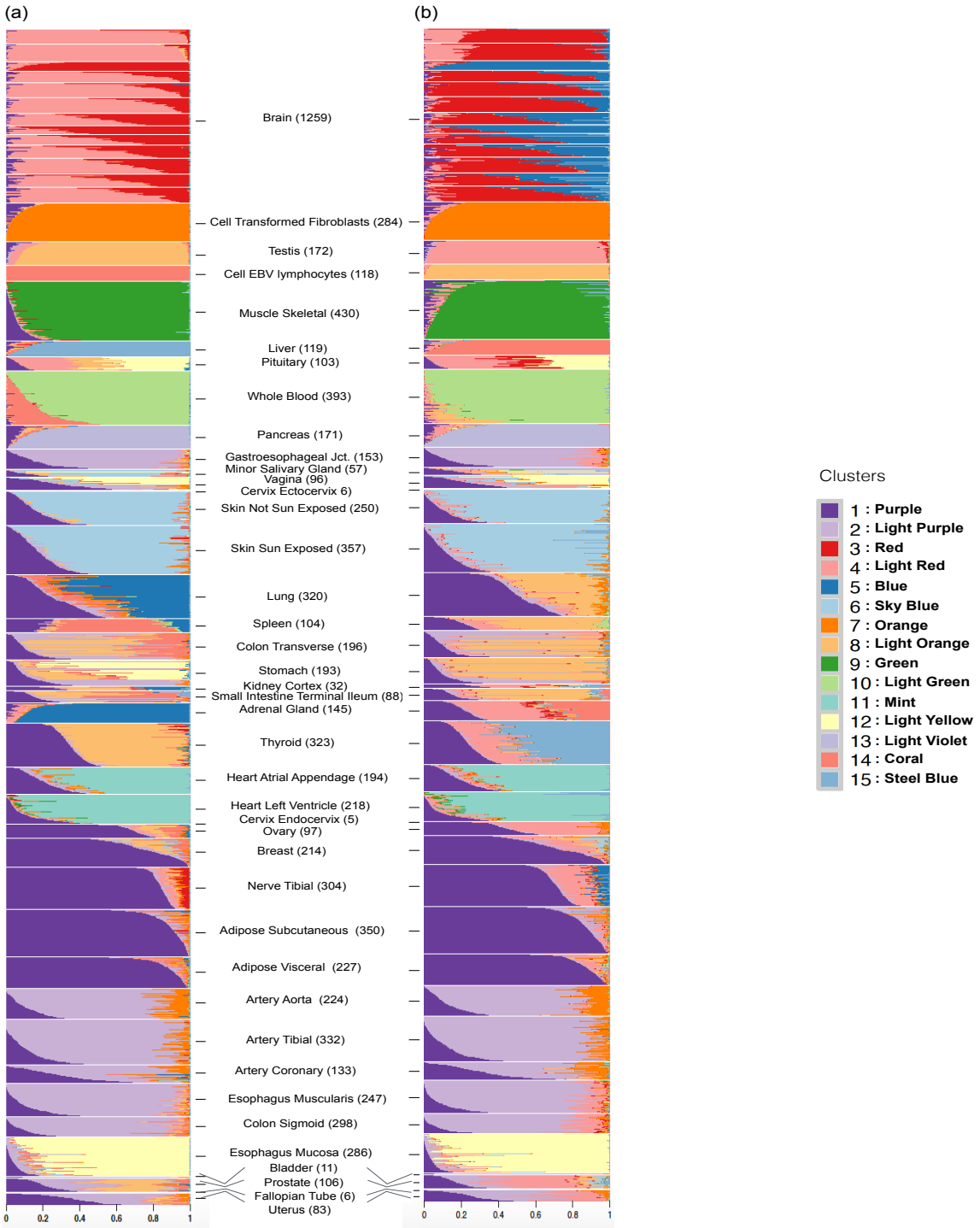
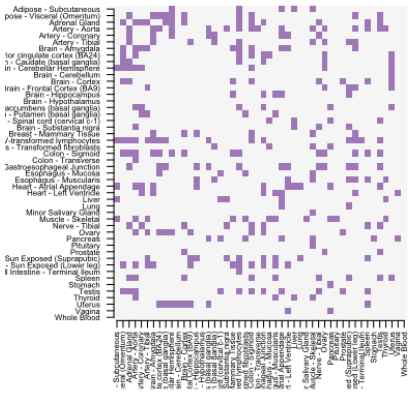
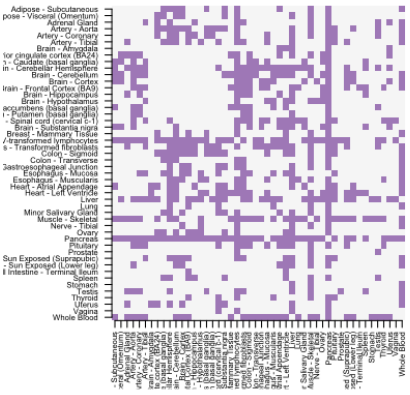


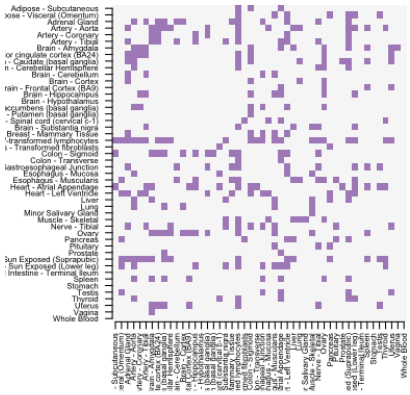
Figure 5. Structure plot of all tissue samples in 2 runs of the GTEx V6 data for K=15 for the thinning parameters (a) $p_{thin} = 0.01$ and (b) $p_{thin} = 0.0001$. The patterns in two plots closely correspond to the plot in **Fig** 1 (a), though are a bit more noisy than compared to the unthinned version.



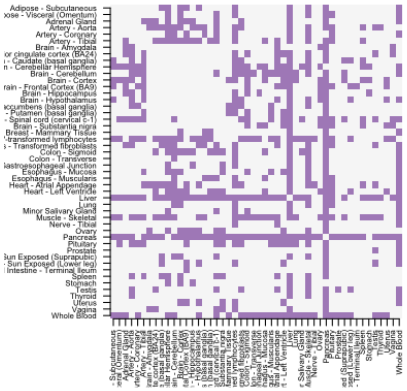
(a) hierarchy thin 0.01



(b) GoM thin 0.01



(c) hierarchy 0.001



(d) GoM thin 0.001

Figure 6. A comparison of "accuracy" of hierarchical vs model-based clustering on thinned GTEx data, with thinning parameter $p_{thin} = 0.001$ and $p_{thin} = 0.0001$. For each pair of tissues from the GTEx data we assessed whether or not each clustering method (with $K = 2$ clusters) separated the samples according to their actual tissue of origin, with successful separation indicated by a filled square. Thinning deteriorates accuracy compared with the unthinned data (Figure 2), but even then the model-based method remains more successful than the hierarchical clustering in separating the samples by tissue or origin.

Table 1. Cluster Annotations GTEx V6 data

Cluster	Gene names	Proteins	Summary
cluster 1, purple (Nerve, Adipose)	FABP4	fatty acid binding protein 4, adipocyte	FABP4 encodes the fatty acid binding protein found in adipocytes, roles include fatty acid uptake, transport, and metabolism
	APOD	apolipoprotein D	encodes a component of high density lipoprotein that has no marked similarity to other apolipoprotein sequences, closely associated with lipoprotein metabolism.
	PLIN1	perilipin 1	coats lipid storage droplets in adipocytes, thereby protecting them until they can be broken down by hormone-sensitive lipase.
cluster 2, light purple(Arteries, Esophagus)	MYH11	myosin, heavy chain 11, smooth muscle	functions as a major contractile protein, converting chemical energy into mechanical energy through the hydrolysis of ATP.
	ACTA2	actin, alpha 2, smooth muscle, aorta	protein encoded by this gene belongs to the actin family of proteins, which are highly conserved proteins that play a role in cell motility, structure and integrity, defects in this gene cause aortic aneurysm familial thoracic type 6.
	ACTG2	actin, gamma 2, smooth muscle, enteric	encodes actin gamma 2; a smooth muscle actin found in enteric tissues, involved in various types of cell motility and in the maintenance of the cytoskeleton.
cluster 3, red (Brain)	MBP	myelin basic protein	major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system
	GFAP	glial fibrillary acidic protein	encodes one of the major intermediate filament proteins of mature astrocytes, mutations casuses Alexander disease.
	SNAP25	synaptosomal-associated protein, 25kDa	this gene product is a presynaptic plasma membrane protein involved in the regulation of neurotransmitter release.
cluster 4, light red (Testis)	PRM2	protamine 2	Protamines are the major DNA-binding proteins in the nucleus of sperm
	PRM1	protamine 1	Protamines are the major DNA-binding proteins in the nucleus of sperm
	PHF7	PHD finger protein 7	This gene is expressed in the testis in Sertoli cells but not germ cells, regulates spermatogenesis.
cluster 5, blue (Thyroid, Stomach	TG	thyroglobulin	thyroglobulin produced predominantly in thyroid gland, synthesizes thyroxine and triiodothyronine, associated with Graves disease and Hashimotot thyroiditis.
	LIPF	lipase, gastric	encodes gastric lipase, an enzyme involved in the digestion of dietary triglycerides in the gastrointestinal tract, and responsible for 30 % of fat digestion processes occurring in human.
	PGC	progastricsin (pepsinogen C)	encodes an aspartic proteinase that belongs to the peptidase family A1. The encoded protein is a digestive enzyme that is produced in the stomach and constitutes a major component of the gastric mucosa, associated with susceptibility to gastric cancers.
cluster 6, sky blue (Skin)	KRT10	keratin 10, type I	encodes a member of the type I (acidic) cytokeratin family, mutations associated with epidermolytic hyperkeratosis.
	KRT1	keratin 1, type II	specifically expressed in the spinous and granular layers of the epidermis with family member KRT10 and mutations in these genes have been associated with bullous congenital ichthyosiform erythroderma.
	KRT2	keratin 2, type II	expressed largely in the upper spinous layer of epidermal keratinocytes and mutations in this gene have been associated with bullous congenital ichthyosiform erythroderma.
cluster 7, orange (Cells fibroblasts)	FN1	fibronectin 1	Fibronectin is involved in cell adhesion, embryogenesis, blood coagulation, host defense, and metastasis.
	COL1A1	collagen, type I, alpha 1	Mutations in this gene associated with osteogenesis imperfecta types I-IV, Ehlers-Danlos syndrome type and Classical type, Caffey Disease.
	COL1A2	collagen, type I, alpha 2	Mutations in this gene associated with osteogenesis imperfecta types I-IV, Ehlers-Danlos syndrome type and Classical type, Caffey Disease.

Cluster	Gene symbol	Gene name	Summary
cluster 8, light orange (Lung)	SFTPB	surfactant protein B	an amphipathic surfactant protein essential for lung function and homeostasis after birth, muttaions cause pulmonary alveolar proteinosis, fatal respiratory distress in the neonatal period.
	SFTPA2	surfactant protein A2	Mutations in this gene and a highly similar gene located nearby, which affect the highly conserved carbohydrate recognition domain, are associated with idiopathic pulmonary fibrosis.
	SFTPA1	surfactant protein A1	encodes a lung surfactant protein that is a member of C-type lectins called collectins, associated with idiopathic pulmonary fibrosis.
cluster 9, green (Muscle skeletal)	MYH1	myosin, heavy chain 1, skeletal muscle, adult	a major contractile protein which converts chemical energy into mechanical energy through the hydrolysis of ATP.
	NEB	nebulin	encodes nebulin, a giant protein component of the cytoskeletal matrix that coexists with the thick and thin filaments within the sarcomeres of skeletal muscle, associated with recessive nemaline myopathy.
	MYH2	myosin, heavy chain 2, skeletal muscle, adult	encodes a member of the class II or conventional myosin heavy chains, and functions in skeletal muscle contraction.
cluster 10, light green (Blood)	HBB	hemoglobin, beta	mutant beta globin causes sickle cell anemia, absence of beta chain/ reduction in beta globin leads to thalassemia.
	HBA2	hemoglobin, alpha 2	deletion of alpha genes may lead to alpha thalassemia.
	HBA1	hemoglobin, alpha 1	deletion of alpha genes may lead to alpha thalassemia.
cluster 11, mint (Heart)	NPPA	natriuretic peptide A	protein encoded by this gene belongs to the natriuretic peptide family, associated with atrial fibrillation familial type 6.
	MYH6	myosin, heavy chain 6, cardiac muscle, alpha	encodes the alpha heavy chain subunit of cardiac myosin, mutations in this gene cause familial hypertrophic cardiomyopathy and atrial septal defect 3.
	ACTC1	actin, alpha, cardiac muscle 1	protein encoded by this gene belongs to the actin family, associated with idiopathic dilated cardiomyopathy (IDC) and familial hypertrophic cardiomyopathy (FHC).
cluster 12, light yellow (Esophagus mucosa)	KRT13	keratin 13, type I	protein encoded by this gene is a member of the keratin gene family, associated with the autosomal dominant disorder White Sponge Nevus.
	KRT4	keratin 4, type II	protein encoded by this gene is a member of the keratin gene family, associated with White Sponge Nevus, characterized by oral, esophageal, and anal leukoplakia.
	CRNN	cornulin	may play a role in the mucosal/epithelial immune response and epidermal differentiation.
cluster 13, light violet (Pancreas)	PRSS1	protease, serine 1	secreted by pancreas, associated with pancreatitis
	CPA1	carboxypeptidase A1	secreted by pancreas, linked to pancreatitis and pancreatic cancer
	PNLIP	pancreatic lipase	encodes a carboxyl esterase that hydrolyzes insoluble, emulsified triglycerides, and is essential for the efficient digestion of dietary fats. This gene is expressed specifically in the pancreas.
cluster 14, coral (Liver)	MUC7	mucin 7, secreted	encodes a small salivary mucin, thought to play a role in facilitating the clearance of bacteria in the oral cavity and to aid in mastication, speech, and swallowing, associated with susceptibility to asthma.
	ALB	albumin	functions primarily as a carrier protein for steroids, fatty acids, and thyroid hormones and plays a role in stabilizing extracellular fluid volume.
	HP	haptoglobin	encodes a preproprotein, which subsequently produces haptoglobin, linked to diabetic nephropathy, Crohn's disease, inflammatory disease behavior and reduced incidence of Plasmodium falciparum malaria.
cluster 15, steel blue (Pituitary)	PRL	prolactin 2	encodes the anterior pituitary hormone prolactin. This secreted hormone is a growth regulator for many tissues, including cells of the immune system.
	GH1	growth hormone 1	expressed in the pituitary, play an important role in growth control, mutations in or deletions of the gene lead to growth hormone deficiency and short stature.
	POMC	proopiomelanocortin	synthesized mainly in corticotroph cells of the anterior pituitary, mutations in this gene have been associated with early onset obesity, adrenal insufficiency, and red hair pigmentation.

4.7 Supplementary Table 1

Table 2. Cluster Annotations GTEx V6 Brain data

Cluster	Gene symbol	Gene name	Summary
cluster 1, royal blue	ATP1A2	ATPase, Na+/K+ transporting, alpha 2 polypeptide	responsible for establishing and maintaining the electrochemical gradients of Na and K ions across the plasma membrane, mutations in this gene result in familial basilar or hemiplegic migraines, and in a rare syndrome known as alternating hemiplegia of childhood.
	CLU	clusterin	protein encoded by this gene is a secreted chaperone that can under some stress conditions also be found in the cell cytosol, also involved in cell death, tumor progression, and neurodegenerative disorders.
	DNAJB1	DnaJ (Hsp40) homolog, subfamily B, member 1	encodes a member of the DnaJ or Hsp40 (heat shock protein 40 kD) family of proteins, that stimulates the ATPase activity of Hsp70 heat-shock proteins to promote protein folding and prevent misfolded protein aggregation.
cluster 2, yellow orange	SNAP25	synaptosomal-associated protein, 25kDa	Synaptic vesicle membrane docking and fusion is mediated by SNAREs located on the vesicle membrane (v-SNAREs) and the target membrane (t-SNAREs), involved in the regulation of neurotransmitter release.
	ENO2	enolase 2 (gamma, neuronal)	encodes one of the three enolase isoenzymes found in mammals, is found in mature neurons and cells of neuronal origin.
	CHGB	chromogranin B	encodes a tyrosine-sulfated secretory protein abundant in peptidergic endocrine cells and neurons. This protein may serve as a precursor for regulatory peptides.
cluster 3, turquoise	CALM3	calmodulin 3 (phosphorylase kinase, delta)	is a calcium binding protein that plays a role in signaling pathways, cell cycle progression and proliferation.
	FBXL16	F-box and leucine-rich repeat protein 16	Members of the F-box protein family, such as FBXL16, are characterized by an approximately 40-amino acid F-box motif.
	UCHL1	ubiquitin carboxyl-terminal esterase L1	specifically expressed in the neurons and in cells of the diffuse neuroendocrine system. Mutations in this gene may be associated with Parkinson disease.
cluster 4, red	MBP	myelin basic protein	protein encoded is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system.
	MYH11	glial fibrillary acidic protein	encodes major intermediate filament proteins of mature astrocytes, a marker to distinguish astrocytes during development, mutations in this gene cause Alexander disease, a rare disorder of astrocytes in central nervous system.
	ACTA2	secreted protein, acidic, cysteine-rich (osteonectin)	encodes a cysteine-rich acidic matrix-associated protein, required for the collagen in bone to become calcified, in extracellular matrix synthesis and cell shape promotion, associated with tumor suppression.

4.8 Supplementary Table 2

Table 3. Cluster Annotations Deng et al (2014) data

Cluster	Gene symbol	Gene name/function
cluster 1, green	Timd2	T cell immunoglobulin and mucin domain containing 2
	Isyna1	myo-inositol 1-phosphate synthase A1
	Alppl2	alkaline phosphatase, placental-like 2
cluster 2, purple	Upp1	uridine phosphorylase 1
	Tdgf1	teratocarcinoma-derived growth factor 1
	Fabp5	fatty acid binding protein 5, epidermal, protects against atherosclerosis, diet-induced obesity, insulin resistance and experimental autoimmune encephalomyelitis
cluster 3, orange	Actb	actin, beta, involved in cell motility, structure, and integrity
	Krt18	keratin 18
	Fabp3	fatty acid binding protein 3, muscle and heart
cluster 4, yellow	Rtn2	reticulon 2 (Z-band associated protein)
	Hsp90ab1	heat shock protein 90 alpha (cytosolic), class B member 1
	Calm1	calmodulin 1, associated with catecholaminergic polymorphic ventricular tachycardia and long QT syndrome 14
cluster 5, blue	Bcl2l10	Bcl2 like 10
	Tcl1	T cell lymphoma breakpoint 1
	E330034G19Rik	RIKEN cDNA E330034G19 gene
cluster 6, magenta	Obox3	oocyte specific homeobox 3
	Btg4	B cell translocation gene 4
	Zfp352	zinc finger protein 352

References

1. Eisen MB, Spellman PT, Brown PO and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25): 14863-14868
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531-7
3. Alizadeh AA1, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC *et al*. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769): 503-11
4. D'haeseleer P. 2005. How does gene expression clustering work? *Nat Biotechnol*, 23(12):1499-501
5. Jiang D, Tang C, Zhang A. Cluster Analysis for Gene Expression Data: A Survey. *Microsoft Research*, http://research.microsoft.com/en-us/people/djiang/tkde04.pdf.
6. Erosheva EA. 2006. Latent class representation of the grade of membership model. Seattle: University of Washington.
7. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6, 377 - 382.
8. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res.*, 25, 1491-1498.
9. The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*. 45(6): 580-585. doi:10.1038/ng.2653.
10. Oshlack A, Robinsom MD, Young MD. 2010. From RNA-seq reads to differential expression results. *Genome Biology*. 11:220, DOI: 10.1186/gb-2010-11-12-220
11. Matt Taddy. 2012. On Estimation and Selection for Topic Models. *AISTATS 2012, JMLR W&CP 22*. (maptpx R package).
12. Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155.2, 945-959.
13. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. The genetic structure of human populations. *Science*. 298, 2381-2385.
14. Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*. 197, 573-589.
15. Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 164(4), 1567-87.
16. Engelhardt BE, Stephens M. 2010. Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. *PLOS Genetics*. DOI: 10.1371/journal.pgen.1001117.
17. van der Maaten LJP and Hinton GE. 2008. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.*. 2579-2605.
18. L.J.P. van der Maaten. 2014. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*. 3221-3245.
19. Jaitin DA, Kenigsberg E et al. 2014. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*. 343 (6172) 776-779.
20. Deng Q, Ramskold D, Reinius B, Sandberg R. 2014. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*. 343 (6167) 193-196.
21. Leek JT, Storey JD. 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis *PLoS Genet*. 3(9): e161. doi:10.1371/journal.pgen.0030161
22. Stegle O, Parts L , Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc.*. 7(3):500-7. doi: 10.1038/nprot.2011.457.

23. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. 11, 733-739.

24. Hicks SC, Teng M, Irizarry RA. 2015. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BiorXiv*. http://biorxiv.org/content/early/2015/09/04/025528

25. Mark A, Thompson R and Wu C. 2014. mygene: Access MyGene.Info services. *R package version 1.2.3.*.

26. Flutre T, Wen X, Pritchard J and Stephens M. 2013. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genet*. 9(5): e1003486. doi:10.1371/journal.pgen.1003486

27. Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*. 3, 993-1022

28. Blei DM, Lafferty J. 2009. Topic Models. *In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications . Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*.

29. Shen-Orr SS, Tibshirani R, Khatri, P, Bodian DL, Staedtler F, Perry NM, Hastie, T, Sarwal MM, Davis MM, Butte AJ. 2010. Cell typespecific gene expression differences in complex tissues. *Nature Methods*. 7(4), 287-289

30. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. 2012. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Comput Biol*. 8(12)

31. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M. 2010. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC bioinformatics*.11(1), 27+

32. Schwartz R, Shackney SE. 2010. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC bioinformatics*. 11(1), 42+

33. Lindsay J, Mandoiu I, Nelson C. 2013. Gene Expression Deconvolution using Single-cells `http://dna.engr.uconn.edu/bibtexmngr/upload/Lal.13.pdf`.

34. Hu JG, Shi LL, Chen YJ, Xie XM, Zhang N, Zhu AY, Zheng JS, Feng YF, Zhang C, Xi J, Lu HZ. 2016. Differential effects of myelin basic protein-activated Th1 and Th2 cells on the local immune microenvironment of injured spinal cord. *Experimental Neurology*. 277, 190-201

35. duVerle D, Tsuda K. 2016. cellTree: Inference and visualisation of Single-Cell RNA-seq data as a hierarchical tree structure. *R package version 1.1.0*, `http://tsudalab.org`.

36. Renard M, Callewaert B, Baetens M, Campens L, MacDermot K et al. 2013. Novel MYH11 and ACTA2 mutations reveal a role for enhanced TGFβ signaling in FTAAD *Int J Cardiol*. 165(2), 314-321.

37. Gong B, Cao Z, Zheng P, Vitolo OV, Liu S, Staniszewski A, Moolman D, Zhang H, Shelanski M, Arancio O. 2006. Ubiquitin Hydrolase Uch-L1 Rescues β-Amyloid-Induced Decreases in Synaptic Function and Contextual Memory *Cell*. 126(4), 775-788

38. Hoffman MD, Blei DM, Bach F. 2010. Online learning for latent Dirichlet allocation. *Neural Information Processing Systems*.