Model-based clustering of RNA-seq expression data using grade of membership models

Kushal K Dey[1] 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: graded membership model, GTEx bulk-RNA, single cell RNA-seq

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

Abstract

We argue for the benefits of model-based approach to clustering expression data from bulk RNA and single cell RNA-seq data. It is similar to the topic model in Natural Language Processing and the admixture model in Population Genetics, that assigns grades of cluster membership to each sample. This approach provides us with easily interpretable cluster visualization and detects the underlying structure in the data better than distance based approaches. It also extracts the important genes that drive the clusters and provides measures of model fit to assess the strength of clustering. Further, we show that this method is robust under low coverage of reads. We apply this method on the GTEx tissue level bulk RNA expression data as well as two single cell RNA-seq data. Our methods are implemented in a 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 has played a major role in their analysis [1]. For example, clustering can aid in quality control, and helps elucidate heterogeneity among samples. However, although gene expression measurement technologies have moved on – with gene expression measurements now routinely made with sequencing technologies (RNA-seq), sometimes at the resolution of single cells (sc-RNA-seq) – clustering methods used have remained largely the same. Indeed, the most commonly-used clustering techniques in gene expression studies remain the distance-based hierarchical clustering methods used in early microarray work [1]. This contrasts with many other fields, where model-based clustering methods have become widely used, and in many cases the method of choice (e.g. [6]). Our goal here is to argue that such model-based approaches also provide an attractive approach to cluster analysis of RNA-seq data, both bulk and single-cell. In particular we illustrate the potential for "grade of membership models" [2] to elucidate structure in both bulk and single-cell RNA-seq expression data.

Grade of membership models are generalizations of clustering models that allow each sample to have partial membership in more that one cluster. 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 [6], and in document clustering ([18,19]) where each document can have membership in multiple "topics". In these fields the grade of membership models are often known as "admixture models", and "topics models" or "Latent Dirichlet Allocation" [18].

Here we apply grade of membership models to RNA-seq expression data. In this context these models allow that each sample has some proportion of its RNA-seq reads coming from each cluster. For 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 [24]. And, indeed, the grade of membership model we use here is analogous to blind deconvolution approaches [22, 23] which estimate cell type proportions and cell type signatures jointly (see also [20, 21] for semi-supervised approaches), although our models differ from these previous works in the use of a multinomial model for the RNA-seq read counts. 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, several software packages capable of fitting these grade of membership models to RNA-seq data are already available! This is because the models used in natural language processing for clustering documents based on their word counts are multinomial models that apply 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. Here we use the R package `maptpx` [5] 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

2.1 Clustering human tissues based on 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 and LCL 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/utilities/gene_names_GTEX_V6.txt`).

To assess structure in these data we applied the grade-of-membership 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 [9], [10], in **Supplementary Fig 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 (Artery-aorta, Artery-tibial and Artery-coronary) and skin (Skin Not Sun Exposed- suprapubic and Skin Sun Exposed- lower leg). 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 (Adipose Subcutaneous and Adipose Visceral (Omentum)). Indeed, many tissues show membership in the red "Adipose" cluster, possibly reflecting, at least in some cases, contamination with adipose cells.

Each cluster in our model 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 Cluster Annotation, in Methods and Materials). **Tab 1** summarizes the results of this cluster annotation (top three genes) for the GTEX analysis in Figure 1a. Again, reassuringly, the results consistently align with known biology. For example, the top three genes driving the light brown cluster, 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 grey cluster, 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 burlywood cluster which is common to Lung, Spleen and Small Intestine - Terminal Ileum, code for surfactant proteins B, A2 and A1.

Although global analysis of all tissues is useful for highlighting major structure in the data, it may be less effective at identifying finer-scale structure within tissues or among similar tissues. For example, here our 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. **Fig 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**). The spinal cord samples also show consistently strong membership in a single cluster, whose top 3 defining genes are MBP (involved in myelination [25]) and MYH11 and ACTA2, both of which are .. Genetic mutations in the latter two genes can cause TAAD Thoracic aortic aneurysms/dissection [26]. The remaining samples all show membership in multiple cluster, with cortex samples being distinguished from other samples by stronger membership in a cluster (blue in figure) whose top 3 defining genes include UCHL1, which is expressed in neurons and is required for normal synaptic and cognitive function [27].

2.2 Quantitative comparison with hierarchical clustering

Since most clustering of gene expression datasets is performed using distance-based hierarchical clustering methods, we compared the accuracy of these distance-based methods with the model-based approach. Specifically, for each pair of tissues in the GTEX data we assessed whether or not each clustering method correctly partitioned samples into the two groups (See methods for details.) Overall the model-based method succeeded in 59% of the comparisons, compared with 29% for the distance-based method; Figure 2.

2.3 Robustness to sequencing depth

RNA-sequencing of single-cell 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. Therefore, we checked robustness of our results above to sequencing depth. Specifically we repeated analyses using thinned data designed to mimic lower sequencing depth: if c_{ng} is the counts of number of reads mapping to gene g for sample n for the original data, then the thinned counts are given by

$$t_{ng} \sim Bin(c_{ng}, p_{thin}) \tag{1}$$

where we used $p_{thin} = 0.0001$. With this value of p_{thin} total sequencing depth is similar to the per-sample depth in the single-cell data from [11] considered below. (See **Fig ??** [check!] and **Fig 5** for results with other values of p_{thin}.)

For the thinned GTEX data the Structure plot for $K = 15$ (**Fig 1(c)**) preserves most of the major features of the original analysis on unthinned data (**Fig 1(a)**). For the comparisons with distance-based methods, both methods suffer reduced accuracy for the thinned data, but the model-based method retains its superior performance (xx vs yy).

2.4 Clustering of single-cell RNA-seq data

To examine the potential for model-based clustering of single cell RNA-seq data we examined two datasets, from Jaitin *et al* [11] and Deng *et al* [12].

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 Figures 2A and 2B of [11]. However, our method yielded rather different results (Figure 3). Each cell was assigned to multiple clusters, and the cluster membership vectors strongly correlated with amplification batch (which in turn is strongly correlated with sequencing batch). 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 [15] and in single cell data specifically [16]. And indeed, dimension reduction methods such as the ones we use here can be helpful in controlling for such effects [13] [14]. However, why these batch effects are not visible in the original analyses [11], is unclear.

Deng *et al* collected expression data from individual cells from zygote to blastocyst stages of mouse preimplantation development [12]. 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 [12] 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 red 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 grade-of-membership models to capture structure in single cell data that might be missed by simpler cluster-based approaches.

3 Discussions

We have presented a model based clustering approach for RNA-seq (bulk or single cell) read counts data which models each sample as having a mixed membership in different clusters and also helps identifying genes driving the clusters, which may be of significant bio-medical importance. Our approach is an alternative to the distance based methods of clustering, for instance hierarchical clustering, and it seems to outperform the latter in separating biologically meaningful groups in our tests.

Since our method is model based, it provides an optimal K for the model fit. However, one has to run the model on the data for a range of K's and that is not always practical when running the model on large datasets as in RNA-seq reads data. For a single run with the algorithm running till the successive log posterior increase is less than 0.01, the computation time for the algorithm was approximately 33.4 mins for Jaitin *et al* [**?**] single cell experiment ($K = 7$), 16.2 mins for Deng *et al* single cell experiment ($K = 6$) and 3297.4 mins for GTEx V6 data ($K = 15$).

So far, in our studies, we fitted the model on the entire RNA-seq reads data, comprising of all the genes. In reality, most of the genes will not be informative about the clusters and an efficient variable selection algorithm, if incorporated with the clustering algorithm, can lead to significant speed up without much loss of information.This is a future direction to this work we are interested in. Another point of biological interest would be to perform cluster annotation of genetic pathways which would be more meaningful as genes often act together with other genes in pathways related to different activities.

The methods discussed in this paper are implemented in the package **CountClust** available on Github (https://github.com/kkdey/CountClust) which is a wrapper package of **maptpx** due to Matt Taddy [5].

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 [4]. We remove genes g with all zero counts ($c_{ng} = 0$ for all n), and use the `maptpx` R package [5] 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

$$c_{n\cdot} \sim \text{Mult}(c_{n+}, p_{n\cdot}) \tag{2}$$

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

$$p_{ng} = \sum_{k=1}^{K} q_{nk}\theta_{kg} \tag{3}$$

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 [5] for details.

4.2 Visualizing Results

We visualize results using a "Structure plot" [7], which is named for its widespread use in visualizing the results of the "structure" software [6] in population genetics. The Structure plot represents each GoM vector $q_{n\cdot}$ as a vertical stacked barchart, with bars of different colors representing membership proportion in each cluster (e.g. Figure 1). If the colored patterns of two bars are similar, then the two samples have similar membership proportions. The Structure plot is particularly helpful when external information is available on each sample that can be used to order or group the samples in an informative way.

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 [9, 10]. 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. Specifically, for each cluster k we measure the distinctiveness of gene g with respect to any other cluster l using

$$\mathrm{KL}^g[k,l] := \theta_{kg}\ log\frac{\theta_{kg}}{\theta_{lg}} + \theta_{lg} - \theta_{kg}, \tag{4}$$

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

$$D^g[k] = \min_{l \neq k} \mathrm{KL}^g[k,l]. \tag{5}$$

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 [**?**].

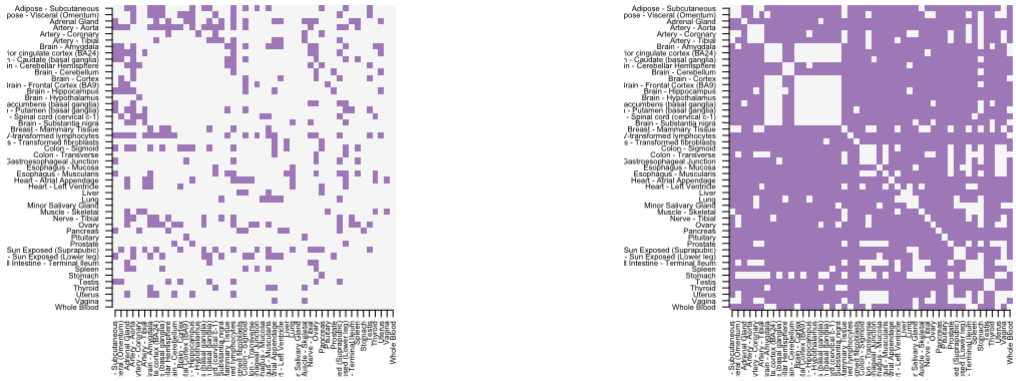
4.4 Comparison with hierarchical clustering

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.

../plots/fig1_paper/fig1.pdf

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). **(c)**: As in a), but with data thinned so that the coverage is closer to a typical single cell RNA-seq data ($p_{thin} = 0.0001$). The clustering patterns are largely similar to a), but slightly noisier.



(a) hierarchy thin 0.1 **(b)** admixture thin 0.1

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 xx comparisons and the hierarchical clustering in yy comparisons.

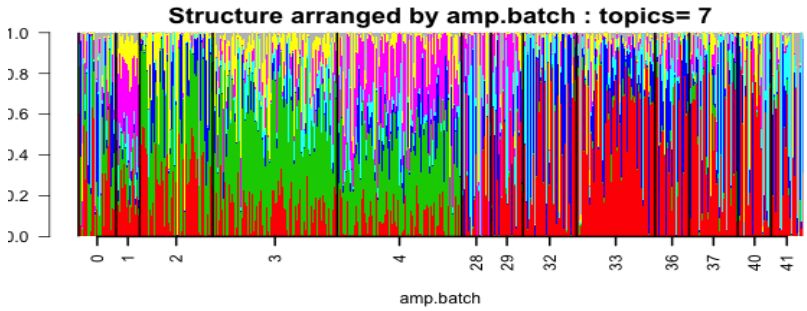


Figure 3. Structure plot of estimated cluster membership proportions for $K = 7$ clusters fit to 1041 single cells from [11]. The samples are ordered so that samples from the same amplification batch are adjacent. In this analysis the samples do not appear to form clearly-defined clusters, with each 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.

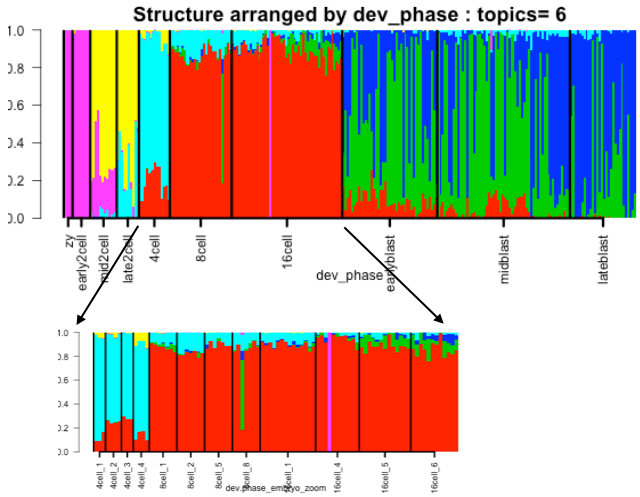
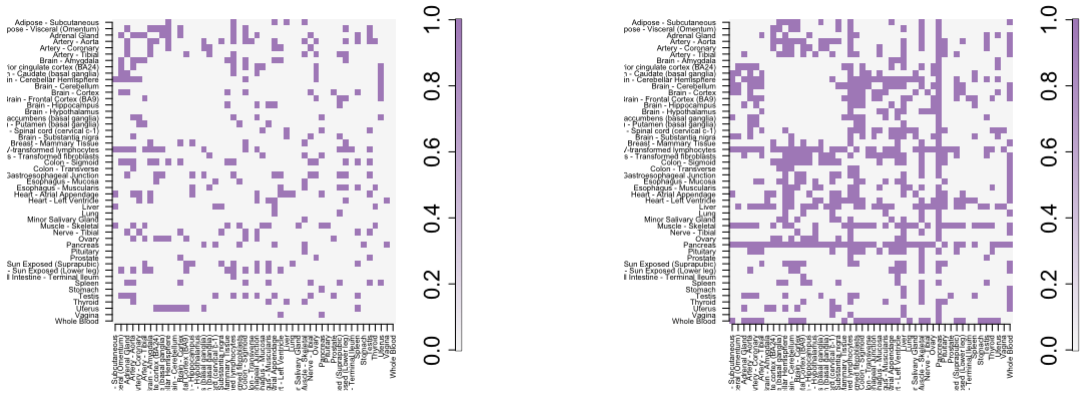


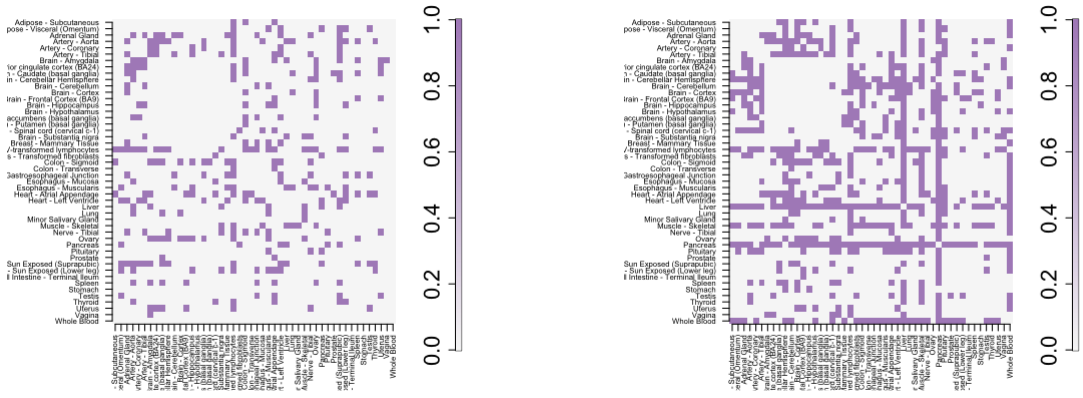
Figure 4. Structure plot of estimated cluster membership proportions for $K = 6$ clusters fit to xxx single cells from [12]. The cells are ordered by their preimplantation development phase (and within each phase, arranged in the same order as in the data file). While the very earliest developmental 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.

4.5 Supplemental figures



(a) hierarchy thin 0.01

(b) admixture thin 0.01



(c) hierarchy 0.001

(d) admixture thin 0.001

Figure 5. 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. As expected, thinning deteriorate accuracy compared with the unthinned data (Figure 3), but 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 red (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 blue (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 cornflowerblue (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 black (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 cyan (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 dark 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 brown4 (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 names	Proteins	Summary
cluster burlywood (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 darkgoldenrod (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 darkgray (Whole 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 deepskyblue (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 dark khaki (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 firebrick (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 dark orchid (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 hotpink (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.

4.6 Supplementary Table 1

Cluster	Gene names	Proteins	Summary
cluster 1,red	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, green	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,blue	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, cyan	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.

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. Erosheva EA. 2006. Latent class representation of the grade of membership model. Seattle: University of Washington.

3. The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*. 45(6): 580-585. doi:10.1038/ng.2653.

4. 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

5. Matt Taddy. 2012. On Estimation and Selection for Topic Models. *AISTATS 2012, JMLR W&CP 22*. (maptpx R package).

6. Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155.2, 945-959.

7. 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.

8. Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*. 197, 573-589.

9. van der Maaten LJP and Hinton GE. 2008. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.*. 2579-2605.

10. L.J.P. van der Maaten. 2014. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*. 3221-3245.

11. 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.

12. 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.

13. 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

14. 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.

15. 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.

16. 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

17. 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

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

19. 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.*

20. 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

21. 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)

22. 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+

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

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

25. 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

26. 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.

27. 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