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 Dirichlet Allocation, bulk RNA-seq, single cell RNA-seq

Corresponding Author: Email mstephens@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. We also provide methods to help interpret the clusters, by identifying genes that are distinctively expressed in each cluster. Together these methods provide an attractive summary of the primary structure in several example RNA-seq applications. Applied to data from the GTEx project on 51 human tissues, the approach highlight similarities among biologically-related tissues and identifies distinctively-expressed genes that largely recapitulate known biology. Applied to single-cell expression data from mouse preimplantation embryos, the approach highlights both discrete and continuous variation through early embryonic development stages, and highlights genes involved in a variety of relevant processes – from germ cell development, through compaction and morula formation, to the formation of inner cell mass and trophoblast at the blastocyte stage. 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, clustering – of both genes and samples – has played a major role in their analysis [3] [4] [5]. 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 [6] [7] for reviews.

Here we focus on cluster analysis of samples, rather than clustering of genes (although our methods do highlight sets of genes that distinguish each cluster). 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 (GoM) models [8], 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 [14], and in document clustering ([30,31]) where each document can have membership in multiple topics. In these fields GoM models are often known as "admixture models", and "topic models" or "Latent Dirichlet Allocation" [30].

In the context of RNA-seq expression data, the GoM 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 – and 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 [36]. And, indeed, the GoM model we use here is analogous to – although different in detail from – blind deconvolution approaches [34,35] which estimate cell type proportions and cell type signatures jointly (see also [32, 33] for semi-supervised approaches). However, we believe that the GoM model can be useful more generally to elucidate structure in expression data - not only as a way to deconvolve mixtures of cell types. For example, in single-cell expression data, where each sample is a single cell, 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 is a fundamental question that, when combined with appropriate single-cell RNA-seq data, the GoM models used here may ultimately help address. Further, even for bulk RNA-seq data, we argue that GoM models may yield interesting insights into heterogeneity among samples even if the inferred cluster memberships do not correspond precisely to proportions of specific cell types, as may often happen in practice.

Interestingly, although we have not previously seen GoM models applied to RNA-seq data, several software packages for doing this already exist! [1] This is because of a parallel between clustering samples based on RNA-seq counts, and clustering documents based on word counts, which means that many existing software packages for document clustering can be applied directly to RNA-seq data. Specifically, the Latent Dirichlet Allocation model from [30], 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. Here we use the R package `maptpx` [13] to fit these models, and we add functionality for conveniently visualizing the results and annotating clusters by their most distinctive genes to help biological interpretation. These methods are implemented in the R package `CountClust` available from `https://github.com/kkdey/CountClust`.

2 Results

In brief, our approach starts by summarizing RNA-seq data 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 [12]. We fit a

[1] While preparing this work for publication we became aware of ongoing independent work by [38] applying GoM models to RNA-seq data.

GoM model to this table of counts, which 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 of gene g in sample n. The GoM 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" [15], which is named for its widespread use in visualizing the results of the *Structure* software [14] 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-seq

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 $8,555$ 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 `http://stephenslab.github.io/count-clustering/project/src/gene_annotation_2.html`).

We applied the GoM model to these data, with $K = 10, 12, 15$. Many of the primary patterns were consistent across these K, and so 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 [20], [21], in **Supplementary Figure 1** `http://stephenslab.github.io/count-clustering/project/src/tissues_tSNE_2.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 membership proportions 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 similar membership proportions to samples from those tissues. 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 this "Adipose" cluster (cluster 1, purple in Figure), possibly reflecting, at least in some cases, contamination with adipose cells.

To help biologically interpret results we implemented methods to identify the genes and genetic processes that characterize each cluster (see Methods). Table 1 summarizes results for the GTEx results in Figure 1a (see also Supplementary Table 1). Reassuringly, many results align with known biology. For example, the light red cluster (cluster 4), which distinguishes Testis from other tissues, is enriched for genes responsible for sexual reproduction and spermatogenesis etc, (e.g.: *PRM2*, *PRM1*, *PHF7*). Similarly the light green cluster (cluster 10), which primarily distinguishes Whole Blood from other tissues, is enriched with genes responsible for immune responses (e.g. *CSF3R*, *IFITM2*) and hemoglobin complex and oxygen transport (e.g. *HBB*, *HBA1*, *HBA2*). Further, digestion-related and proteolysis-related genes characterize the Pancreas cluster (cluster 13, light violet), Keratin-related genes characterize the skin cluster (cluster 6, sky blue), Myosin-related genes characterize the muscle skeletal cluster (cluster 9, green), etc. In cases where a cluster occurs in multiple tissues these biological annotations may be particularly helpful for understanding what is driving this co-membership. For example, the top four 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, A1 and C).

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, which is enriched with genes related to cell periphery and communication (e.g. *PKD1*, *CBLN3*)as well genes expressed largely in neuronal cells and playing a role in neuron differentiation (e.g. *CHGB*). The spinal cord samples also show consistently strong membership in a single cluster, the top defining gene for the cluster being *MBP* which is involved in myelination of nerves in the nervous system [37]. The other driving gene *GFAP* takes part in system development by acting as a marker to distinguish astrocytes during development [2].

The remaining samples all show membership in multiple clusters, with cortex samples being distinguished from other samples by stronger membership in a cluster (cluster 3, turquoise in Figure 1(b)) whose distinctive genes include *ENC1*, which interacts with actin and contributes to the organisation of the cytoskeleton during the specification of neural fate [1].

2.2 Quantitative comparison with hierarchical clustering

We hypothesized that the model-based GoM approach might be more accurate in detecting substructure than distance-based methods, and 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. This presumably reflects the general tendency for model-based approaches to be more efficient that distance-based approaches, provided that the model is sufficiently accurate.

2.3 Clustering of single-cell RNA-seq data

Recently RNA-sequencing has become viable for single cells [9], and this technology has the promise to revolutionize understanding of intra-cellular variation in expression, and regulation more generally [10]. 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 illustrate the potential for the GoM model to be applied to single cell data.

To be applicable to single-cell RNA-seq data, methods must be able to deal with lower sequencing depth than in bulk RNA experiments: single-cell RNA-seq data typically involve substantially lower effective sequencing depth compared with bulk experiments, due to the relatively small number of molecules available to sequence in a single cell. Therefore, as a first step towards demonstrating its potential for single cell analysis, we checked robustness of the GoM model to sequencing depth. Specifically, we repeated the 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 2). For the accuracy comparisons with distance-based methods, both methods suffer reduced accuracy in thinned data, but the GoM model remains superior. For example, when thinning by a factor of $1,000$, the success rate in separating pairs of tissues is 0.32 for the GoM model vs 0.10 for hierarchical clustering.

Having established its robustness to sequencing depth, we now illustrate the GoM model on two single cell RNA-seq datasets, from Jaitin *et al* [22] and Deng *et al* [23].

2.3.1 Jaitin et al, 2014

Jaitin *et al* sequenced over $4,000$ single cells from mouse spleen. Here we analyze $1,041$ 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 1,040 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 [22] (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. And, more generally, these later batches are collectively more similar to one another than they are to the earlier batches (0-4).

The fact that batch effects are detectable in these data is not particularly surprising: there is a growing recognition of the importance of batch effects in high-throughput data generally [26] and in single cell data specifically [27]. And indeed, both clustering methods and the GoM model can be viewed as dimension reduction methods, and such methods can be helpful in controlling for batch effects [24] [25]. However, why these batch effects are not evident in Figures 2A and 2B of [22] is unclear.

2.3.2 Deng et al, 2014

Deng *et al* collected single-cell expression data of mouse preimplantation embryos from the zygote to blastocyst stage [23], with cells from four different embryos sequenced at each stage. The original analysis [23] focusses on trends of allele-specific expression in early embryo development. Here we use the GoM model to assess the primary structure in these data without regard to allele-specific effects (i.e. combining counts of the two alleles). Visual inspection of the Principal Components Analysis in [23] suggested perhaps 6-7 clusters, and we focus here on results with $K = 6$.

The results from the GoM model (Figure 4) clearly highlight changes in expression profiles that occur through early embryonic development stages, and enrichment analysis of the top driving genes in each cluster (Table 3) indicate that many of these expression changes reflect important biologal processes during embryonic preimplantation.

In more detail: Initially, at the zygote stage, the embryos are represented by a single cluster (blue in Figure 4) that is enriched with genes responsible for germ cell development (e.g., *Bcl2l10* [46], *Spin1* [47]), which is essential to the zygote and 2-cell stages. Continuing on to the cleavage division stages, the grades of membership change to include a mixture of blue and magenta clusters at the mid 2-cell stage, a mixture of magenta and yellow clusters at the late 2-cell stage, and then a mixture of yellow and green at the 4-cell stage. The green cluster then becomes more prominent in the 8-6 cell stages, before dropping substantially in the early and mid-blastocyst

stages. This green cluster is enriched for cytoskeletal genes (e.g., *Fbxo15*) and cytoplasm genes (e.g., *Tceb1*, *Hsp90ab1*), all of which are essential for compaction at the 8-cell stage and morula formation at the 16-cell stage. Finally, during the blastocyst stages two new clusters (purple and orange in Figure 4) dominate. The orange cluster is enriched with genes involved in the formation of outer trophoblast cells (e.g., *Tspan8*, *Krt8*, *Id2* [44]), while the purple cluster is enriched with genes responsible for the formation of inner cell mass (e.g., *Pdgfra*, *Pyy* [45]). These two clusters are therefore consistent with the two cell lineages, the trophectoderm and the primitive endoderm, that make up the majority of the cells of the blastocyst [48]. Interestingly, however, the cells do not appear to fall into two distinct and clearly-separated populations – at least, not in terms of their expression patterns – but rather show a continuous range of memberships in these two clusters, even in the late blastocyst stage.

In addition to these trends across development stages, the GoM results also highlight some embryo-level effects in the early stages (Figure 4). Specifically, cells from the same embryo sometimes show greater similarity than cells from different embryos. For example, while all cells from the 16-cell stage have high memberships in the green cluster, cells from two of the embryos at this stage have memberships in both the purple and yellow clusters, while the other two embryos have memberships only in the yellow cluster.

Finally, we note that, like clustering methods, the GoM model can be helpful in exploratory data analysis and quality control. Indeed, the GoM results highlight a few single cells as outliers. For example, a cell from a 16-cell embryo is represented by the blue cluster - a cluster that represents cells at the zygote and early 2-cell stage. Also, a cell from an 8-stage embryo has strong membership in the purple cluster - a cluster that represents cells from the blastocyst stage. It would seem prudent to consider excluding these cells from subsequent analyses of these data.

3 Discussion

Our goal here is to highlight the potential for 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 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 Jaitin *et al* [22] and Deng *et al* [23]: 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 the 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 [19] 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 [23] ($n = 259; K = 6$), through 33 minutes for the data from [22] ($n = 1,041; K = 7$) 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" (sequential) approaches capable of dealing with millions of documents [41] 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 [15]. 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 [18], 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 [18] and we would expect the same to happen here for RNA-seq data.

4 Methods and Materials

4.1 Model Fitting

We use the `maptpx` R package [13] to fit the GoM model (1,2), which is also known as "Latent Dirichlet Allocation" (LDA). The `maptpx` package fits this model using an EM algorithm to perform Maximum a posteriori (MAP) estimation of the parameters q and θ. See [13] 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 [20, 21]. 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 [16].) Specifically, for each cluster k we measure the distinctiveness of gene g with respect to any other cluster l using

$$\mathrm{KL}^g[k,l] := \theta_{kg}\, log\frac{\theta_{kg}}{\theta_{lg}} + \theta_{lg} - \theta_{kg}, \tag{3}$$

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

$$D^g[k] = \min_{l \neq k} \mathrm{KL}^g[k,l]. \tag{4}$$

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 biological functions of these individual genes using the `mygene` R Bioconductor package [28].

For each cluster k, we filter out a number of genes (top 100 for the Deng *et al* data [23] and GTEx V6 data [11]) with highest $D^g[k]$ value and perform a gene set over-representation analysis of these genes against all the other genes in the data representing the background. To do this, we used ConsensusPathDB database (`http://cpdb.molgen.mpg.de/`) [42] [43]. See Tables 1 - 2 and Table 3 for the top significant gene ontologies driving each cluster in the GTEx V6 data and the Deng *et al* data respectively.

4.4 Comparison with hierarchical clustering

We compared the GoM model with a distance-based hierarchical clustering algorithm by applying both methods to samples from pairs of tissues from the GTEx project, and assessed their accuracy in separating 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 GoM model we analysed the data with $K = 2$, and sorted 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 checked 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.

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

$$t_{ng} \sim Bin(c_{ng}, p_{thin}) \tag{5}$$

where p_{thin} is a specified thinning parameter.

4.6 Code Availability

Our methods are implemented in an R package **CountClust**, available at `https://github.com/kkdey/CountClust`.

Code for reproducing results reported here is available at `http://stephenslab.github.io/count-clustering/`.

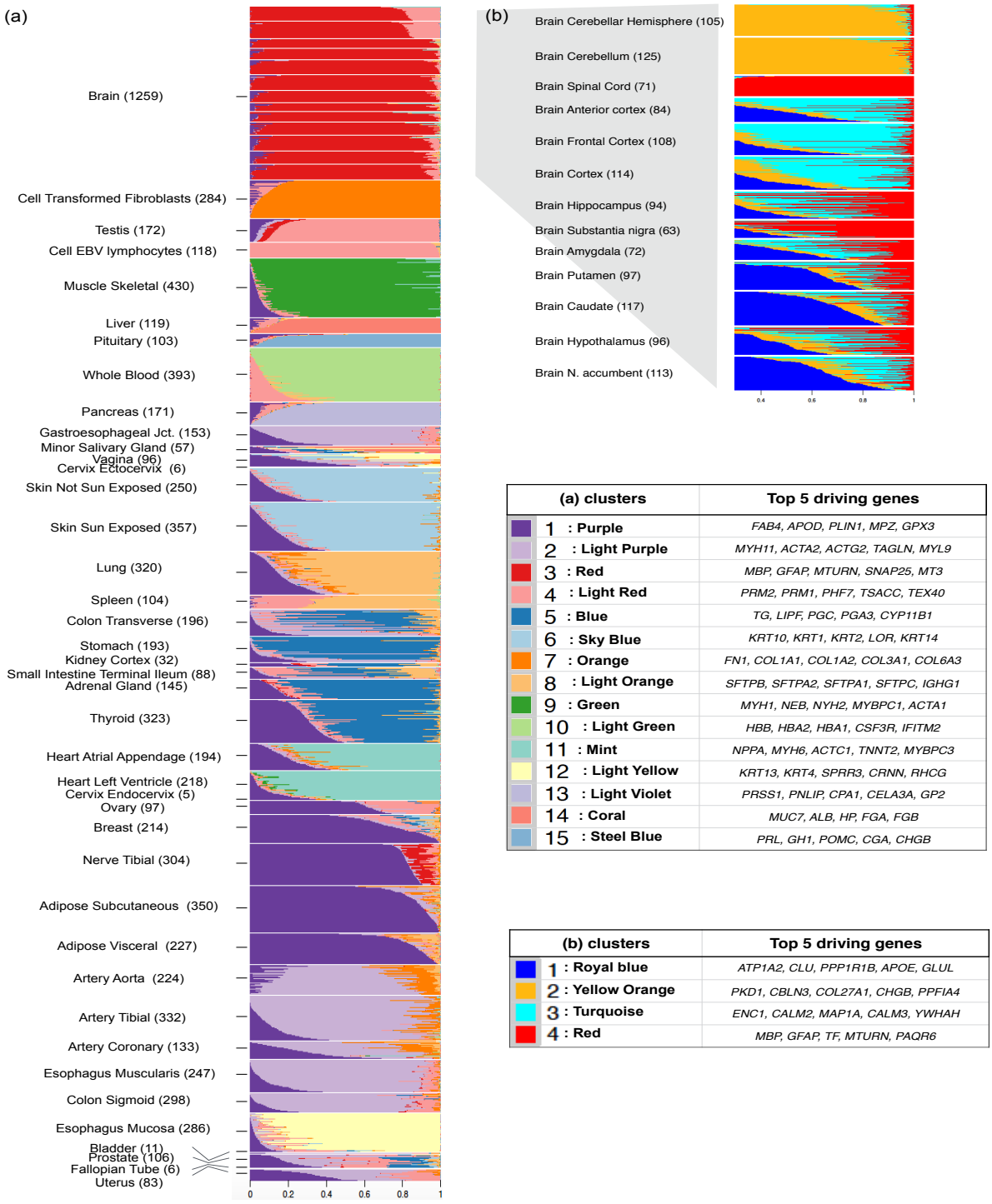
Acknowledgements

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.

We thank Matt Taddy, Amos Tanay and Effi Kenigsberg for helpful discussions. We thank Po Yuan and John Blischak for helpful comments on a draft manuscript.

Disclosure Declaration

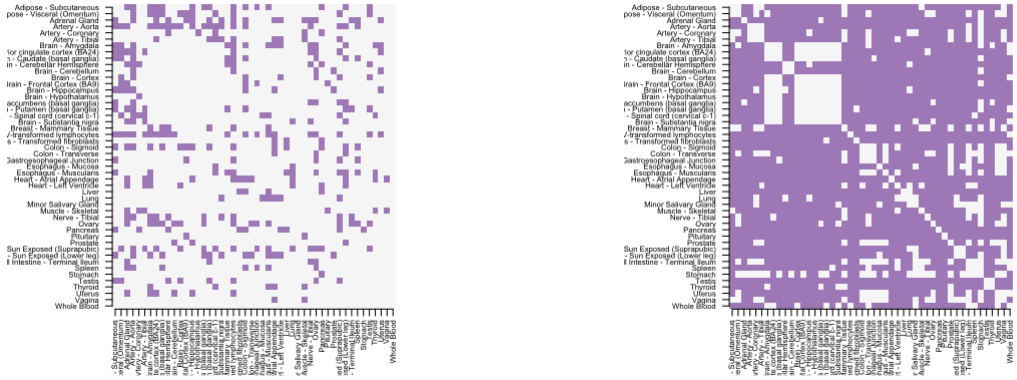
The authors have no conflict of interest.



(a) clusters	Top 5 driving genes
1 : Purple	*FAB4, APOD, PLIN1, MPZ, GPX3*
2 : Light Purple	*MYH11, ACTA2, ACTG2, TAGLN, MYL9*
3 : Red	*MBP, GFAP, MTURN, SNAP25, MT3*
4 : Light Red	*PRM2, PRM1, PHF7, TSACC, TEX40*
5 : Blue	*TG, LIPF, PGC, PGA3, CYP11B1*
6 : Sky Blue	*KRT10, KRT1, KRT2, LOR, KRT14*
7 : Orange	*FN1, COL1A1, COL1A2, COL3A1, COL6A3*
8 : Light Orange	*SFTPB, SFTPA2, SFTPA1, SFTPC, IGHG1*
9 : Green	*MYH1, NEB, NYH2, MYBPC1, ACTA1*
10 : Light Green	*HBB, HBA2, HBA1, CSF3R, IFITM2*
11 : Mint	*NPPA, MYH6, ACTC1, TNNT2, MYBPC3*
12 : Light Yellow	*KRT13, KRT4, SPRR3, CRNN, RHCG*
13 : Light Violet	*PRSS1, PNLIP, CPA1, CELA3A, GP2*
14 : Coral	*MUC7, ALB, HP, FGA, FGB*
15 : Steel Blue	*PRL, GH1, POMC, CGA, CHGB*

(b) clusters	Top 5 driving genes
1 : Royal blue	*ATP1A2, CLU, PPP1R1B, APOE, GLUL*
2 : Yellow Orange	*PKD1, CBLN3, COL27A1, CHGB, PPFIA4*
3 : Turquoise	*ENC1, CALM2, MAP1A, CALM3, YWHAH*
4 : Red	*MBP, GFAP, TF, MTURN, PAQR6*

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. Within each tissue, the samples are sorted by the proportional representation of the underlying clusters.**(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.

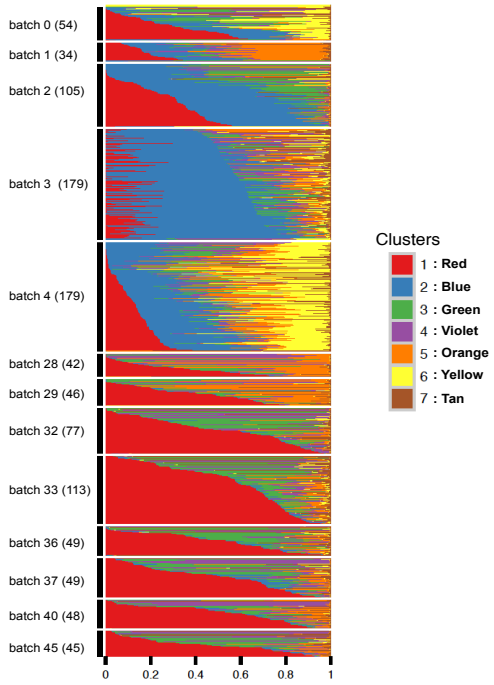


Figure 3. Structure plot of estimated cluster membership proportions for $K = 7$ clusters fit to $1,041$ single cells from [22]. The samples are ordered so that samples from the same amplification batch are adjacent and within each batch, the samples are sorted by the proportional representation of the underlying clusters. 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.

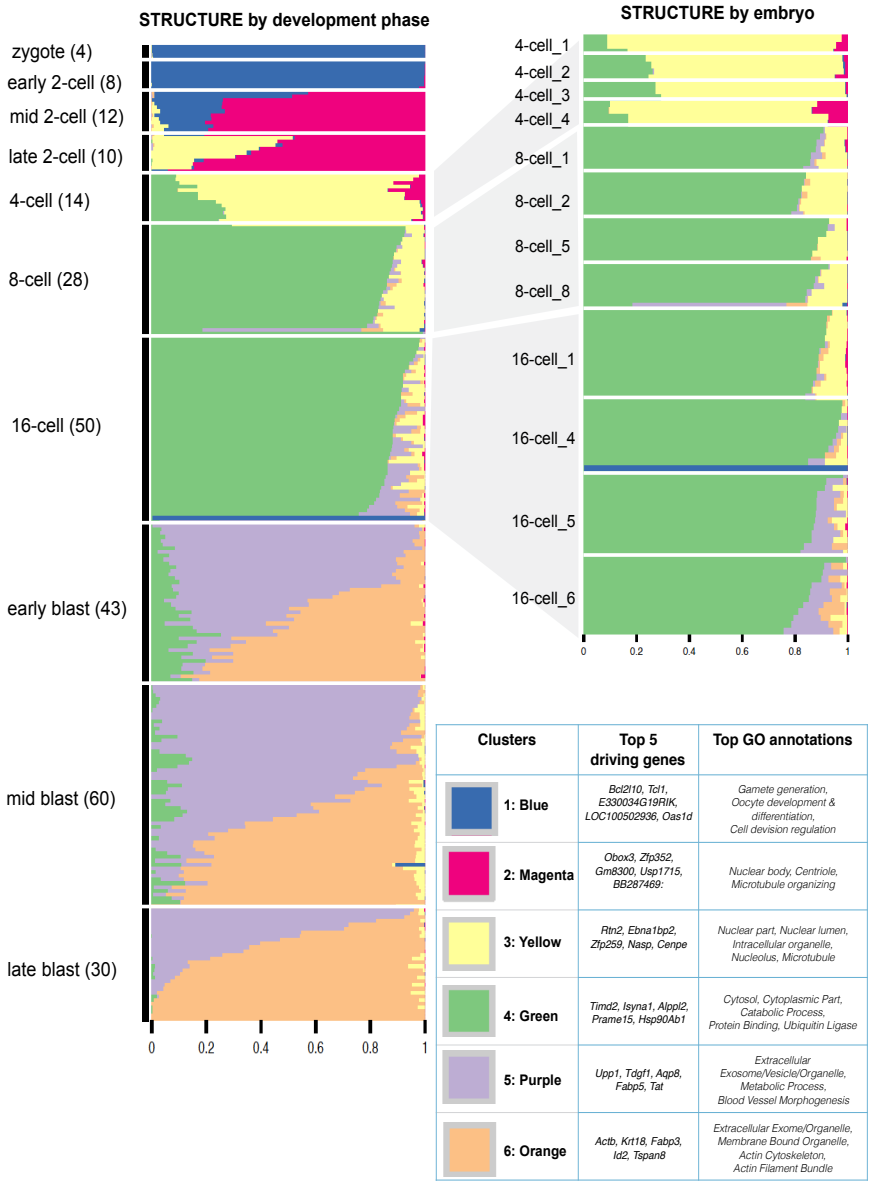


Figure 4. Structure plot of estimated cluster membership proportions for $K = 6$ clusters fit to 259 single cells from [23]. The cells are ordered by their preimplantation development phase (and within each phase, sorted by the proportional representation of the clusters). While the very earliest developmental 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 GoM model than by simple discrete clusters.

Table 1. Cluster Annotations GTEx V6 data (with GO annotations).

Cluster	Top 5 Driving Genes	Gene names	Top significant GO terms
cluster 1, purple	*FABP4*	fatty acid binding protein 4, adipocyte	GO:0005578 (proteinaceous extracellular matrix), GO:0005615 (extracellular space), GO:1901700 (response to oxygen-containing compound), GO:0005811 (lipid particle), GO:0009719 (response to endogenous stimulus), GO:0009611 (response to wounding)
	APOD	apoliprotein D	
	PLIN1	perilipin 1	
	MPZ	myelin protein zero	
	GPX3	glutathione peroxidase 3	
cluster 2, light purple	*MYH11*	myosin, heavy chain 11, smooth muscle	GO:0005925 (focal adhesion), GO:0005924 (cell-substrate adherens junction), GO:0015629 (actin cytoskeleton), GO:0001725 (stress fiber), GO:0006936 (muscle contraction), GO:0032432 (actin filament bundle)
	ACTA2	actin, alpha 2, smooth muscle, aorta	
	ACTG2	actin, gamma 2, smooth muscle, enteric	
	TAGLN	transgelin	
	MYL9	myosin light chain	
cluster 3, red	*MBP*	myelin basic protein	GO:0097458 (neuron part), GO:0007268 (synaptic transmission), GO:0007267 (cell-cell signalling), GO:0007399 (nervous system development), GO:0043005 (neuron projection), GO:0036477 (somatodendritic component), GO:0022008 (neurogenesis), GO:0030424 (axon)
	GFAP	glial fibrillary acidic protein	
	MTURN	maturin, neural progenitor	
	SNAP25	synaptosome associated protein 25kDa	
	MT3	metallothionein 3	
cluster 4, light red	*PRM2*	protamine 2	GO:0019953 (sexual reproduction), GO:0007276 (gamete generation), GO:0007283 (spermatogenesis), GO:0097228 (sperm principal piece), GO:0035686 (sperm fibrous sheath)
	PRM1	protamine 1	
	PHF7	PHD finger protein 7	
	TSACC	TSSK6 activating co-chaperone	
	TEX40	testis expressed 40	
cluster 5, blue	*TG*	thyroglobulin	GO:0042446 (hormone biosynthetic process), GO:0042445 (hormone metabolic process), GO:0008202 (steroid metabolic process), GO:0006694 (steroid biosynthetic process), GO:0006629 (lipid metabolic process), GO:0065008 (regulation of biological quality)
	LIPF	lipase F, gastric type	
	PGC	progastricsin (pepsinogen C)	
	PGA3	pepsinogen 3, group I (pepsinogen A)	
	CYP11B1	cytochrome P450 family 11 subfamily B member 1	

Cluster	Top 5 Driving Genes	Gene names	Top significant GO terms
cluster 6, sky blue	*KRT10* *KRT1* *KRT2* *LOR* *KRT14*	keratin 10, type I keratin 1, type II keratin 2, type II loricrin keratin 14, type I	GO:0043588 (skin development), GO:0008544 (epidermis development), GO:0060429 (epithelium development), GO:0042633 (hair cycle), GO:0042303 (molting cycle)
cluster 7, orange	*FN1* *COL1A1* *COL1A2* *COL3A1* *COL6A3*	fibronectin 1 collagen type I alpha 1 collagen type I alpha 2 collagen type III alpha 1 collagen type VI alpha 3	GO:0030198 (extracellular matrix organization), GO:0005578 (proteinaceous extracellular matrix), GO:0032963 (collagen metabolic process), GO:0005615 (extracellular space), GO:0030574 (collagen catabolic process)
cluster 8, light orange	*SFTPB* *SFTPA2* *SFTPA1* *SFTPC* *IGHG1*	surfactant protein B surfactant protein A2 surfactant protein A1 surfactant protein C immunoglobulin heavy constant gamma 1 (G1m marker)	GO:0002684 (positive regulation of immune system process), GO:0006955 (immune response), GO:0006952 (defense response), GO:0006959 (humoral immune response), GO:0002443 (leukocyte mediated immunity)
cluster 9, green	*MYH1* *NEB* *MYH2* *MYBPC1* *ACTA1*	myosin, heavy chain 1, skeletal muscle, adult nebulin myosin, heavy chain 2, skeletal muscle, adult myosin binding protein C, slow type actin, alpha 1, skeletal muscle	GO:0043292 (contractile fiber), GO:0030016 (myofibril), GO:0030017 (sarcomere), GO:0006936 (muscle contraction), GO:0003012 (muscle system process), GO:0015629 (actin cytoskeleton), GO:0008092 (cytoskeletal protein binding)
cluster 10, light green	*HBB* *HBA2* *HBA1* *CSF3R* *IFITM2*	hemoglobin, beta hemoglobin subunit alpha 2 hemoglobin subunit alpha 1 colony stimulating factor 3 receptor interferon induced transmembrane protein 2	GO:0006955 (immune response), GO:0071944 (cell periphery), GO:0005886 (plasma membrane), GO:0005833 (hemoglobin complex), GO:0005344 (oxygen transporter activity), GO:0015669 (gas transport), GO:0009611 (response to wounding)

Cluster	Top 5 Driving Genes	Gene names	Top significant GO terms
cluster 11, mint	*NPPA* *MYH6* *ACTC1* *TNNT2* *MYBPC3*	natriuretic peptide A myosin, heavy chain 6, cardiac muscle, alpha actin, alpha, cardiac muscle 1 troponin T2, cardiac type myosin binding protein C, cardiac	GO:0022904 (espiratory electron transport chain), GO:0030017 (sarcomere), GO:0044449 (contractile fiber part), GO:0030016 (myofibril), GO:0045333 (cellular respiration), GO:0003015 (heart process), GO:0006091 (generation of precursor metabolites and energy), GO:0006936 (muscle contraction)
cluster 12, light yellow	*KRT13* *KRT4* *SPRR3* *CRNN* *RHCG*	keratin 13, type I keratin 4, type II small proline rich protein 3 cornulin Rh family C glycoprotein	GO:0008544 (epidermis development), GO:0031424 (keratinization), GO:0030855 (epithelial cell differentiation), GO:0065010 (extracellular membrane-bounded organelle), GO:0070062 (extracellular exosome), GO:1903561 (extracellular vesicle)
cluster 13, light violet	*PRSS1* *PNLIP* *CPA1* *CELA3A* *GP2*	protease, serine 1 pancreatic lipase carboxypeptidase A1 chymotrypsin like elastase family member 3A glycoprotein 2	GO:0007586 (digestion), GO:0004252 (serine endopeptidase activity), GO:0008233 (peptidase activity), GO:0044241 (lipid digestion), GO:0006508 (proteolysis),GO:0016160 (amylase activity)
cluster 14, coral	*MUC7* *ALB* *HP* *FGA* *FGB*	mucin 7, secreted albumin haptoglobin fibrinogen alpha chain fibrinogen beta chain	GO:0005615 (extracellular space), GO:0072562 (blood microparticle), GO:0065010 (extracellular membrane bound organelle), GO:0070062 (extracellular exosome), GO:0002526 (acute inflammatory response), GO:0031982 (vesicle)
cluster 15, steel blue	*PRL* *GH1* *POMC* *CGA* *CHGB*	prolactin 2 growth hormone 1 proopiomelanocortin glycoprotein hormones, alpha polypeptide chromogranin B	GO:0005179 (hormone activity), GO:0005148 (prolactin receptor binding), GO:0012505 (endomembrane system), GO:0046879 (hormone secretion), GO:0009914 (hormone transport), GO:0050432 (catecholamine secretion)

Table 2. Cluster Annotations GTEx V6 Brain data (with GO annotations).

Cluster	Top 5 Driving Genes	Gene names	Top significant GO terms
cluster 1, royal blue	*ATP1A2*	ATPase Na+/K+ transporting subunit alpha 2	GO:0044707 (single-multicellular organism process), GO:0005615 (extracellular space), GO:0048731 (system development), GO:0007154 (cell communication), GO:0007275 (multicellular organismal development), GO:0007267 (cell-cell signaling)
	CLU	clusterin	
	PPP1R1B	protein phosphatase 1 regulatory inhibitor subunit 1B	
	APOE	apoliprotein	
	GLUL	glutamate-ammonia ligase	
cluster 2, yellow orange	*PKD1*	polycystin 1, transient receptor potential channel interacting	GO:0005886 (plasma membrane), GO:0071944 (cell periphery), GO:0097458 (neuron part), GO:0030182 (neuron differentiation), GO:0007154 (cell communication), GO:0098794 (postsynapse), GO:0050803 (regulation of synapse structure/activity)
	CBLN3	cerebellin 3 precursor	
	COL27A1	collagen type XXVII alpha 1	
	CHGB	chromogranin B	
	PPFIA4	PTPRF interacting protein alpha 4	
cluster 3, turquoise	*ENC1*	ectodermal-neural cortex 1	GO:0007268 (synaptic transmission), O:0097458 (neuron part), GO:0007267 (cell-cell signalling), GO:0031175 (neuron projection development), GO:0030182 (neuron differentiation), GO:0042995 (cell projection)
	CALM2	calmodulin 2 (phosphorylase kinase, delta)	
	MAP1A	microtubule associated protein 1A	
	CALM3	calmodulin 3 (phosphorylase kinase, delta)	
	YWHAH	tyrosine 3-monooxygenase 5-monooxygenase activation protein	
cluster 4, red	*MBP*	mylein basic protein	GO:0043209 (myelin sheath), GO:0007275 (multicellular organism development), GO:0031982 (vesicle), GO:0048731 (system development), GO:0007272 (ensheathment of neurons), GO:0008366 (axon ensheathment)
	GFAP	glial fibrillary acidic protein	
	TF	transferin	
	MTURN	maturin, neural progenitor differentiation regulator homolog	
	PAQR6	progestin and adipoQ receptor family member VI	

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

Cluster	Top 10 Driving Genes	Gene names	Top significant GO terms
cluster 1, blue	*Bcl2l10*	Bcl2 like 10	GO:0007276 (gamete generation), GO:0032504 (multicellular organism reproduction), GO:0044702 (single organism reproduction), GO:0048477 (oogenesis), GO:0048599 (oocyte development), GO:0009994 (oocyte differentiation), GO:0051321 (meiotic cell cycle), GO:0006306 (DNA methylation), GO:0051302 (regulation of cell division)
	Tcl1	T cell lymphoma breakpoint 1	
	E330034G19Rik	RIKEN cDNA E330034G19 gene	
	LOC100502936	NA	
	Oas1d	$2^{'}$-$5^{'}$ oligoadenylate synthetase 1D	
	AU022751	expressed sequence AU022751	
	Spin1	spindlin 1	
	Khdc1b	KH domain containing 1B	
	D6Ertd527e	DNA segment, Chr 6, ERATO Doi 527, expressed	
	Btg4	B cell translocation gene 4	
cluster 2, magenta	*Obox3*	oocyte specific homeobox 3	GO:0016604 (nuclear body), GO:0005814 (centriole), GO:0044450 (microtubule organizing center part)
	Zfp352	zinc finger protein 352	
	Gm8300	predicted gene 8300	
	Usp17l5	NA	
	BB287469	expressed sequence BB287469	
	Rfpl4b	ret finger protein-like 4B	
	Gm2022	predicted pseudogene 2022	
	Gm5662	predicted gene 5662	
	Gm11544	predicted gene 11544	
	Gm4850	THO complex 4 pseudogene	
cluster 3, yellow	*Rtn2*	reticulon 2 (Z-band associated protein)	GO:0044428 (nuclear part), GO:0031981 (nuclear lumen), GO:0070013 (intracellular organelle lumen), GO:0005730 (nucleolus), GO:0005654 (nucleoplasm), GO:0003723 (RNA binding), GO:0005874 (microtubule), GO:0043229 (intracellular organelle)
	Ebna1bp2	EBNA1 binding protein 2	
	Zfp259	NA	
	Nasp	nuclear autoantigenic sperm protein (histone-binding)	
	Cenpe	centromere protein E	
	Rnf216	ring finger protein 216	
	Ctsl	cathepsin L	
	Tor1b	torsin family 1, member B	
	Ankrd10	ankyrin repeat domain 10	
	Lamp2	lysosomal-associated membrane protein 2	

Cluster	Top 10 Driving Genes	Gene names	Top significant GO terms
cluster 4, green	*Timd2*	T cell immunoglobulin and mucin domain containing 2	GO:0005829 (cytosol), GO:0044444 (cytoplasmic part), GO:1901575 (organic substance catabolic process), GO:0000151 (ubiquitin ligase com- plex), GO:0009056 (catabolic process), GO:0072655 (protein localization mitochondrion), GO:0044265 (cellular macromolecule catabolic process), GO:0051082 (unfolded protein binding), GO:0023026 (MHC class II protein complex binding), GO:0055131 (C3HC4-type RING finger domain binding)
	Isyna1	myo-inositol 1-phosphate synthase A1	
	Alppl2	alkaline phosphatase, placental-like 2	
	Prame15	preferentially expressed antigen in melanoma like 5	
	Hsp90ab1	heat shock protein 90 alpha (cytosolic), class B member 1	
	Fbxo15	F-box protein 15	
	Tceb1	transcription elongation factor B (SIII), polypeptide 1	
	Gpd1l	glycerol-3-phosphate dehydrogenase 1-like	
	Pemt	phosphatidylethanolamine N-methyltransferase	
	Hsp90aa1	heat shock protein 90, alpha (cytosolic), class A member 1	
cluster 5, purple	*Upp1*	uridine phosphorylase 1	GO:0070062 (extracellular exosome), GO:0043230 (extracellular organelle), GO:1903561 (extracellular vesicle), GO:0006950 (response to stress), GO:0006979 (response to oxidative stress), GO:0044710 (metabolic process), GO:0048514 (blood vessel morphogenesis), GO:0001944 (vasculature development), GO:0030198 (extracellular matrix organization)
	Tdgf1	teratocarcinoma-derived growth factor 1	
	Aqp8	aquaporin 8	
	Fabp5	fatty acid binding protein 5, epidermal, protects against atherosclerosis, diet-induced obesity, insulin resistance and experimental autoimmune encephalomyelitis	
	Tat	tyrosine aminotransferase, regulated by glucocorticoid and polypeptide hormones	
	Pdgfra	platelet derived growth factor receptor, alpha polypeptide	
	Pyy	peptide YY	
	Prdx1	peroxiredoxin 1	
	Col4a1	collagen, type IV, alpha 1.	
	Spp1	secreted phosphoprotein 1	
cluster 6, orange	*Actb*	actin, beta, involved in cell motility, structure, and integrity	GO:0065010 (extracellular membrane-bounded organelle), GO:0070062 (extracellular exosome), GO:0043230 (extracellular organelle), GO:1903561 (extracellular vesicle), GO:0031982 (vesicle), GO:0048468 (cell development), GO:0030036 (actin cytoskeleton and organization), GO:0032432 (actin filament bundle), GO:0005912 (adherens junction)
	Krt18	keratin 18	
	Fabp3	fatty acid binding protein 3, muscle and heart	
	Id2	inhibitor of DNA binding 2	
	T*span8*	tetraspanin 8	
	Gm2a	GM2 ganglioside activator protein	
	Lgals1	lectin, galactose binding, soluble 1	
	Adh1	alcohol dehydrogenase 1 (class I)	
	Lrp2	low density lipoprotein receptor-related protein 2	
	BC051665	cDNA sequence BC051665	

References

1. Hernandez MC , Andres-Barquin PJ , Martinez S , Bulfone A , Rubenstein JL , Israel MA. 1997. ENC-1: a novel mammalian kelch-related gene specifically expressed in the nervous system encodes an actin-binding protein. *J Neurosci.*,17(9): 3038-51.
2. Baba H, Nakahira K, Morita N, Tanaka F, Akita H, Ikenaka K. GFAP gene expression during development of astrocyte. *Dev Neurosci.*, 19(1):49-57.
3. Eisen MB, Spellman PT, Brown PO and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25): 14863-14868
4. 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
5. 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
6. D'haeseleer P. 2005. How does gene expression clustering work? *Nat Biotechnol*, 23(12):1499-501
7. 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.
8. Erosheva EA. 2006. Latent class representation of the grade of membership model. Seattle: University of Washington.
9. 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.
10. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res.*, 25, 1491-1498.
11. The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature genetics.* 45(6): 580-585. doi:10.1038/ng.2653.
12. 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
13. Matt Taddy. 2012. On Estimation and Selection for Topic Models. *AISTATS 2012, JMLR W&CP 22.* (maptpx R package).
14. Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics.* 155.2, 945-959.
15. 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.
16. Rosenberg NA. 2005. Algorithms for selecting informative marker panels for population assignment. *J Comput Biol.* 12(9), 1183-201.
17. Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics.* 197, 573-589.
18. 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.
19. 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.
20. van der Maaten LJP and Hinton GE. 2008. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.*. 2579-2605.
21. L.J.P. van der Maaten. 2014. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*. 3221-3245.
22. 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.

23. 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.

24. 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

25. 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.

26. 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.

27. 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

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

29. 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

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

31. 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*.

32. 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

33. 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)

34. 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+

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

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

37. 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

38. 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`.

39. 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.

40. 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

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

42. Kamburov A, et al. 2013. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*.

43. Pentchev K, et al. 2010. Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. *Bioinformatics*.

44. Guo G, Huss M, Tong GQ, Wang C, Sun LL, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*. 18(4), 675-685

45. Hou J, Charters AM, Lee SC, Zhao Y, Wu, MK, Jones SJM, Marra, MA, Hoodless PA. A systematic screen for genes expressed in definitive endoderm by Serial Analysis of Gene Expression (SAGE). *BMC Developmental Biology*. 7(92), 1-13

46. Yoon S, Kim E, Kim YS, Lee H, Kim K, Bae J, Lee K. Role of Bcl2-like 10 (*Bcl2l10*) in regulating mouse oocyte maturation. *Biology of Reproduction*. 81(3), 497-506.

47. Evsikov AV, De Evsikova C. Gene expression during the oocyte-to-embryo transition in mammals. *Molecular Reproduction and Development*. 76, 805-818.

48. Rossant J. Development of the extraembryonic lineages. *Seminars in Developmental Biology*. 6(4), 237-247.

5 Supplementary figures

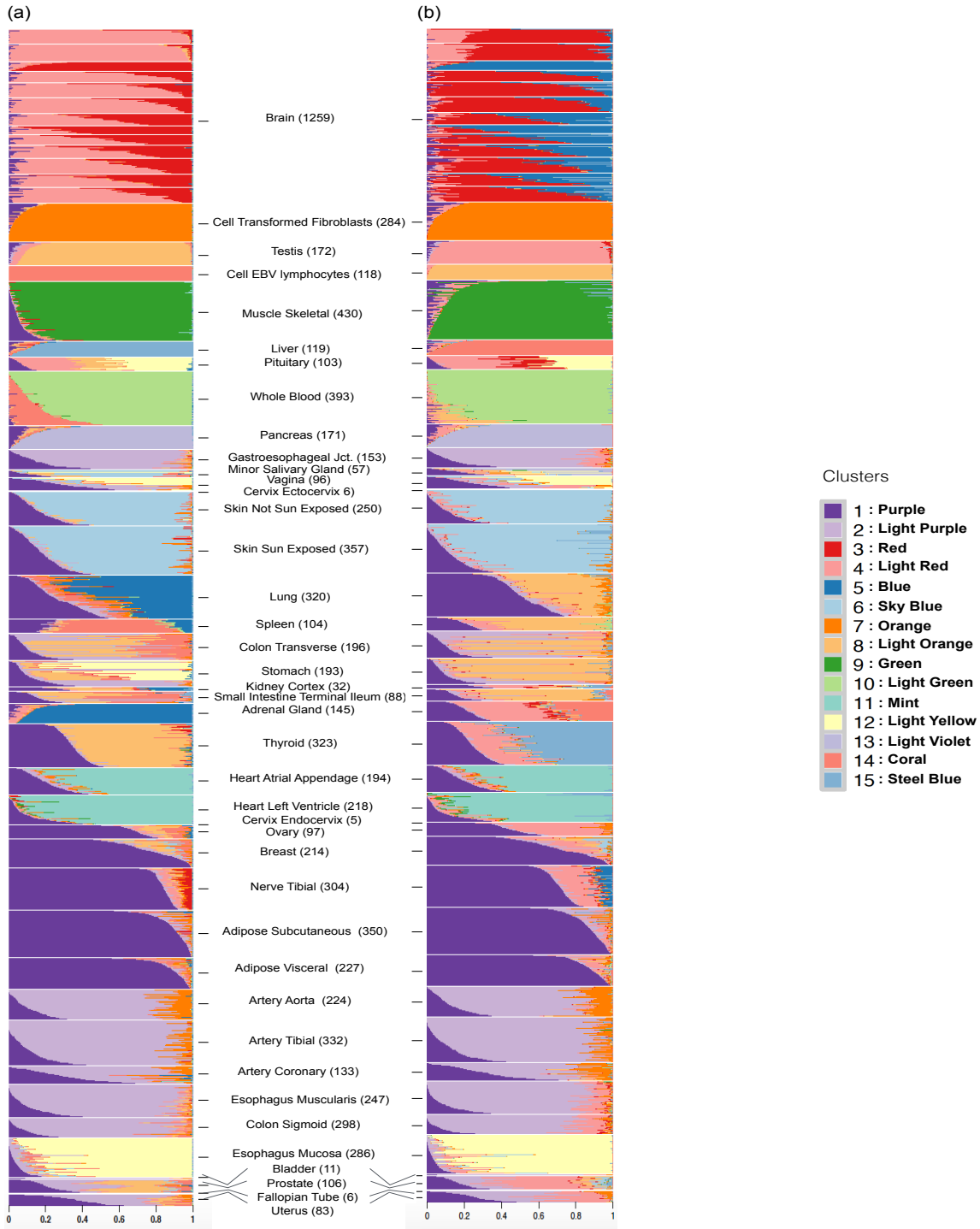
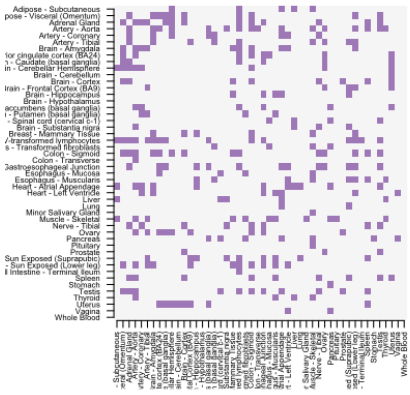
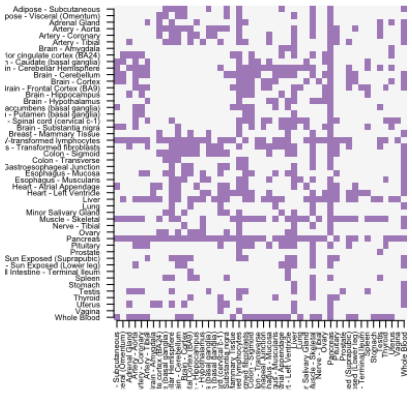


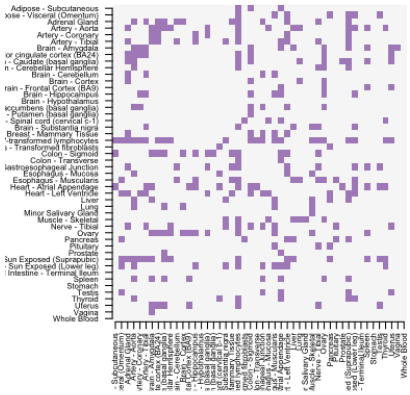
Figure 5. Structure plot of all tissue samples in 2 runs of the GTEx V6 data for K=15 for the thinning parameters (a) $p_{thin} = 0.01$ and (b) $p_{thin} = 0.0001$. The patterns in two plots closely correspond to the plot in **Fig** 1 (a), though are a bit more noisy than compared to the unthinned version.



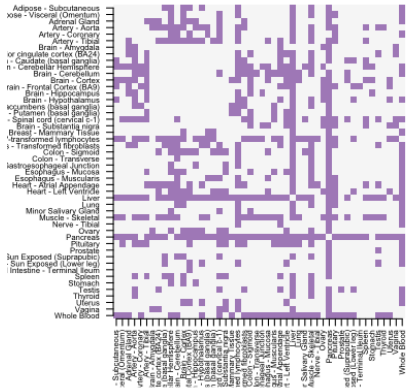
(a) hierarchy thin 0.01



(b) GoM thin 0.01



(c) hierarchy 0.0001



(d) GoM thin 0.0001

Figure 6. A comparison of "accuracy" of hierarchical vs model-based clustering on thinned GTEx data, with thinning parameter $p_{thin} = 0.01$ 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.

6 Supplementary Table 1

Table 4. Cluster Annotations GTEx V6 data (with top gene summaries).

Cluster	Top Driving Genes	Gene names	Gene Summary
cluster 1, purple	*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	*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	*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	*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	*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	*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	*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	Top Driving Genes	Gene namese	Gene Summary
cluster 8, light orange	*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	*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, l. green	*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	*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	*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	*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	*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	*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.

7 Supplementary Table 2

Table 5. Cluster Annotations GTEx V6 Brain data (with top gene summaries).

Cluster	Top Driving Genes	Gene names	Gene 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.
	PPP1R1B	protein phosphatase 1 regulatory inhibitor subunit 1B	encodes a bifunctional signal transduction molecule, may serve as a therapeutic target for neurologic and psychiatric disorders.
cluster 2, yellow orange	*PKD1*	polycystin 1, transient receptor potential channel interacting	functions as a regulator of calcium permeable cation channels and intracellular calcium homoeostasis. It is also involved in cell-cell/matrix interactions and may modulate G-protein-coupled signal-transduction pathways.
	CBLN3	cerebellin 3 precursor	contain a cerebellin motif and C-terminal C1q signature domain that mediates trimeric assembly of atypical collagen complexes
	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	*ENC1*	ectodermal-neural cortex 1	plays a role in the oxidative stress response as a regulator of the transcription factor Nrf2, may play role in malignant transformation.
	CALM2	calmodulin 2 (phosphorylase kinase, delta)	is a calcium binding protein that plays a role in signaling pathways, cell cycle progression and proliferation.
	MAP1A	microtubule associated protein 1A	involved in microtubule assembly, which is an essential step in neurogenesis, almost exclusively expressed in the brain.
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.
	GFAP	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.
	TF	transferrin	transport iron from the intestine, reticuloendothelial system, and liver parenchymal cells to all proliferating cells in the body, involved in the removal of certain organic matter and allergens from serum.

8 Supplementary Table 3

Table 6. Deng et al (2014) Cluster 1 (blue) top GO annotations.

	go.id	name	significant
1	GO:0007276	gamete generation	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
2	GO:0007292	female gamete generation	GDF9; BCL2L10; PABPC1L; BMP15; WEE2; DAZL; NOBOX
3	GO:0048609	multicellular organismal reproductive process	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
4	GO:0032504	multicellular organism reproduction	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
5	GO:0019953	sexual reproduction	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
6	GO:0044702	single organism reproductive process	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; CASP8; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
7	GO:0048477	oogenesis	WEE2; GDF9; NOBOX; PABPC1L; DAZL
8	GO:0044703	multi-organism reproductive process	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
9	GO:0048599	oocyte development	WEE2; GDF9; PABPC1L; DAZL
10	GO:0009994	oocyte differentiation	WEE2; GDF9; PABPC1L; DAZL
11	GO:0051321	meiotic cell cycle	H1FOO; WEE2; TDRD12; SPIN1; PTTG1; DAZL
12	GO:0001556	oocyte maturation	WEE2; PABPC1L; DAZL
13	GO:0006306	DNA methylation	TDRD12; H1FOO; TET3; ZFP57
14	GO:0051302	regulation of cell division	TGFB2; PTTG1; TXNIP; WEE2; CHEK1; DAZL
15	GO:0060255	regulation of macromolecule metabolic process	TGFB2; NOBOX; BPGM; UBE2D3; NFYA; CASP8; BMP15; TXNIP; TDRD12; GDF9; BCL2L10

Table 7. Deng et al (2014) Cluster 2 (magenta) top GO annotations.

	go.id	name	significant
1	GO:0016604	nuclear body	YTHDC1; RBM8A; CDK12; PSME4; PPP1R8; HIPK1; TOPORS
2	GO:0005814	centriole	SFI1; PLK2; ROCK1; TOPORS
3	GO:0044450	microtubule organizing center part	SFI1; PLK2; ROCK1; TOPORS

Table 8. Deng et al (2014) Cluster 3 (yellow) top GO annotations.

	go.id	name	significant
1	GO:0044428	nuclear part	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; TOR1B; MIOS; NR1H3; POLR3K
2	GO:0031981	nuclear lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1
3	GO:0070013	intracellular organelle lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1; DNTTIP2; ZBTB10; ZBTB17
4	GO:0043233	organelle lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1
5	GO:0005730	nucleolus	XPO1; DNTTIP2; ESF1; WDR43; ZDHHC7; HEATR1; POLR1E; DDX24; POLR3K
6	GO:0005634	nucleus	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; TOR1B; MIOS; NR1H3; EIF5B; POLR3K
7	GO:0044446	intracellular organelle part	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; SLU7; NFYB; SLC25A36; ECE2
8	GO:0005654	nucleoplasm	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1; ZBTB10; ZBTB17
9	GO:0003723	RNA binding	PPRC1; EIF5B; XPO1; DNTTIP2; WDR43; DDX10; EIF3C; BCLAF1; EBNA1BP2; RARS
10	GO:0003676	nucleic acid binding	SMARCC1; PPRC1; SLU7; NFYB; POLR1E; EIF5B; POLR3K; XPO1; DNTTIP2
11	GO:0043231	intracellular membrane-bounded organelle	MAD2L2; PTDSS2; SMARCC1; TOR1B; PPRC1; SLU7; NFYB; ESF1; ECE2; LMAN1L
12	GO:0043229	intracellular organelle	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; ARRDC1; SLU7; NFYB; ESF1; ECE2
13	GO:0005874	microtubule	WDR43; KLHL21; HAUS6; CENPE; TEKT2; RACGAP1; WDR81; BCL2L11; KIF20B
14	GO:0044822	poly(A) RNA binding	WDR43; DNTTIP2; ESF1; NXF1; DDX10; HEATR1; EIF3C
15	GO:0044424	intracellular part	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; SNAPC4; POLR3K; ARRDC1; SLU7; NFYB; ESF1; WDR43; ECE2; LMAN1L

Table 9. Deng et al (2014) Cluster 4 (green) top GO annotations.

	go.id	name	significant
1	GO:0005829	cytosol	PARG; UAP1; PSMB10; TCEB1; RPLP0; EIF5; CNBP; RPS3; PSAT1; AACS; PMM1; EXOSC7; EIF3I; SET; BHMT; BHMT2
2	GO:0044444	cytoplasmic part	PARG; UAP1; PSMB10; TCEB1; HSPA8; SERINC1; EIF5; CNBP; RPS3; PSAT1; GPD2; AACS; GPR137B; STIP1; PMM1; EXOSC7; VPREB3; PEX16
3	GO:0055131	C3HC4-type RING finger domain binding	HSPA8; PINK1; DNAJA1
4	GO:1901575	organic substance catabolic process	PSMB10; TCEB1; RPLP0; RPS3; GPD2; PINK1; EXOSC7; ALLC; BHMT; HSP90AB1; RPL13A; ATG7; CUL5; UBXN1; ZMPSTE24
5	GO:0000151	ubiquitin ligase complex	DNAJA1; RNF7; UBE2C; HSPA8; FBXO15; SUGT1; DCAF4; CUL5; FBXL20
6	GO:0072655	protein localization to mitochondrion	TIMM17A; BNIP3L; ARIH2; PEMT; SFN; PINK1; HSP90AA1; TIMM23
7	GO:1901564	organonitrogen compound metabolic process	PSMB10; RPLP0; SERINC1; EIF5; BHMT2; PINK1; EIF3I; ALLC; BHMT; MRPL22; RPL13A; ATG7; NUDT9; VNN1; CTSA; HK1
8	GO:0005737	cytoplasm	PARG; UAP1; PSMB10; TCEB1; HSPA8; SERINC1; EIF5; CNBP; RPS3; PSAT1; GPD2; AACS; GPR137B; STIP1; PMM1; EXOSC7
9	GO:0044265	cellular macromolecule catabolic process	EXOSC7; SUMO2; BNIP3L; ARIH2; PSMB10; TCEB1; RPLP0; UBXN1; HSP90AB1; RPL13A; RPS3; RNF7; PINK1
10	GO:0023026	MHC class II protein complex binding	HSP90AB1; HSP90AA1; HSPA8
11	GO:0051082	unfolded protein binding	DNAJA1; PTGES3; HSPA8; HSP90AB1; HSP90AA1; NPM1
12	GO:0009056	catabolic process	PSMB10; TCEB1; RPLP0; RPS3; GPD2; PINK1; EXOSC7; ALLC; WDR45; HSP90AB1; RPL13A
13	GO:0009057	macromolecule catabolic process	EXOSC7; SUMO2; BNIP3L; ARIH2; PSMB10; TCEB1; RPLP0; AZIN1; UBXN1; HSP90AB1; RPL13A
14	GO:0044248	cellular catabolic process	PSMB10; TCEB1; SUMO2; RPS3; GPD2; PINK1; EXOSC7; ALLC; WDR45; HSP90AB1
15	GO:0006626	protein targeting to mitochondrion	TIMM17A; BNIP3L; ARIH2; PEMT; PINK1; HSP90AA1; TIMM23

Table 10. Deng et al (2014) Cluster 5 (purple) top GO annotations.

	go.id	name	significant
1	GO:0044710	single-organism metabolic process	PCK2; SAT1; EPHX2; NFATC4; CKB; PRDX6; MSH2; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; POGLUT1; FABP5; AKAP12; TDGF1; FBP2; SOX2
2	GO:0006950	response to stress	EPHX2; NFATC4; PRDX6; MSH2; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; FABP5; TDGF1; SOX2
3	GO:0065010	extracellular membrane-bounded organelle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4
4	GO:0070062	extracellular exosome	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4; PRKCI; RAC2; IDH1
5	GO:0043230	extracellular organelle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4
6	GO:1903561	extracellular vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4; PRKCI
7	GO:0042221	response to chemical	EPHX2; NFATC4; MFGE8; PRDX6; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; TDGF1; SOX2
8	GO:0031988	membrane-bounded vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; SPARC
9	GO:0031982	vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; SPARC
10	GO:0001525	angiogenesis	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; MEIS1; SPARC; COL4A2; COL4A1; FGF10; TDGF1
11	GO:0048514	blood vessel morphogenesis	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; ZFP36L1; MEIS1; SPARC; COL4A2; COL4A1; FGF10; TDGF1
12	GO:0001944	vasculature development	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; ZFP36L1; MEIS1; PDPN; SPARC; COL4A2; COL4A1; FGF10; TDGF1
13	GO:0006979	response to oxidative stress	TAT; PDGFRA; BMP4; ETV5; TRAP1; PRDX6; IDH1; PARP1; AQP8; PRDX1; CRYGD
14	GO:0009725	response to hormone	PRKCI; GJA1; PDGFRA; BMP4; MFGE8; TAT; PLOD2; SPP1; IDH1
15	GO:0030198	extracellular matrix organization	PDGFRA; BMP4; JAM2; FN1; PLOD2; SPARC; SPP1; COL4A2; COL4A1; SERPINH1; DPP4

Table 11. Deng et al (2014) Cluster 6 (orange) top GO annotations.

	go.id	name	genes
1	GO:0065010	extracellular membrane-bounded organelle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8; LCP1; UGP2
2	GO:0070062	extracellular exosome	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8; LCP1; UGP2
3	GO:0043230	extracellular organelle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11
4	GO:1903561	extracellular vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8
5	GO:0031988	membrane-bounded vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; TMSB4X; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2
6	GO:0031982	vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; TMSB4X; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8
7	GO:0008092	cytoskeletal protein binding	MYH10; TPM4; TMSB4X; CRYAB; MSN; TMSB10; FABP3; NDRG1; CALM1; FMNL2; MYH9; CAP1; TPM1; CDH1
8	GO:0015629	actin cytoskeleton	MYH10; CLIC4; MYH9; MYL12B; WDR1; CNN2; ARPC2; AHNAK; ACTN4; CRYAB; CAP1; TPM1; DSTN; ARPC5; TPM4
9	GO:0003779	actin binding	MYH10; TPM4; WDR1; CNN2; FMNL2; ARPC2; MYH9; CAP1; TPM1
10	GO:0048468	cell development	MYH10; CAPG; ACTG1; WDR1; CNN2; FMNL2; MYH9; ACTN4; SDC4; CAP1; TPM1; DSTN
11	GO:0030036	actin cytoskeleton organization	MYH10; CAPG; ACTG1; WDR1; CNN2; FMNL2; MYH9; ACTN4; SDC4; CAP1; TPM1
12	GO:0032432	actin filament bundle	MYH10; TPM4; MYL12B; CNN2; MYH9; CRYAB; TPM1; ACTN4; LCP1
13	GO:0005912	adherens junction	TJP2; MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4
14	GO:0070161	anchoring junction	TJP2; MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4
15	GO:0005925	focal adhesion	MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4; CAP1; ARPC5