Visualizing the Structure of 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

* mstephens@uchicago.edu

Abstract

Grade of membership models, also known as "admixture models", "topic models" or "Latent Dirichlet Allocation", are a generalization of cluster models that allow each othersample 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. By applying these methods to several example RNA-seq applications we demonstrate their utility in identifying and summarizing structure and heterogeneity. Applied to data from the GTEx project on 51 human tissues, the approach highlights similarities among biologically-related tissues and identifies distinctively-expressed genes that 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 the Bioconductor package `CountClust`.

Author Summary

The gene expression profile of a biological sample (either tissue or single cells) results from a complex interplay between many biological processes. Consequently, for example, distal tissue samples may be similar in their gene expression profiles because they share biological processes. Our goal here is to illustrate that grade of membership (GoM) models – an approach widely used in population genetics to cluster admixed individuals who have ancestry from multiple populations – provide an attractive approach for clustering biological samples of RNA sequencing data. The GoM model allows each biological sample to have partial memberships in multiple biological-distinct clusters, in contrast to traditional clustering methods that partition samples into distinct subgroups. We also provide methods for identifying genes that are distinctively expressed in each cluster to help biologically interpret the results. Applied to a dataset of 51 human tissues, the GoM approach highlights similarities among biologically-related tissues and identifies distinctively-expressed genes that recapitulate known biology. Applied to gene

expression data of single cells from mouse preimplantation embryos, the approach highlights both discrete and continuous variation through early embryonic development stages, and genes involved in a variety of relevant processes. Our study highlights the potential of GoM models for elucidating biological structure in RNA-seq gene expression data.

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 [5–7]. For example, clustering of genes can identify genes that are working together or are 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 [8, 9] 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 [10], 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 [16], and in document clustering [34, 35] 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" [34]. GoM models have also recently been applied to detect mutation signatures in cancer samples [33].

Although we are not the first to apply GoM-like models to gene expression data, previous applications have been primarily motivated by a specific goal, "cell type deconvolution", which involves using cell-type-specific expression profiles of marker genes to estimate the proportions of different cell types in a mixture [40, 42, 43]. Specifically, the GoM model we use here is analogous to – although different in detail from – blind deconvolution approaches [38, 39, 41] which estimate cell type proportions and cell type signatures jointly (see also [36, 37] for semi-supervised approaches). Our goal here is to demonstrate that GoM models can be useful much more broadly for understanding structure in RNA-seq data – not only to deconvolve mixtures of cell types. For example, in our analysis of human tissue samples from the GTEX project below, the GoM model usefully captures biological heterogeneity among samples even though the inferred grades of membership are unlikely to correspond precisely to proportions of specific cell types. And in our analyses of single-cell expression data the GoM model highlights interesting structure, even though interpreting the grades of membership as "proportions of cell types" is clearly inappropriate because each sample is a single cell! Here we are exploiting the GoM as a flexible extension of traditional cluster models, which can capture "continuous" variation among cells as well as the more "discrete" variation captured by cluster models. Indeed, the extent to which variation among cells can be described in terms of discrete clusters versus 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.

Methods Overview

We assume that the RNA-seq data on N samples has 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 g (or other unit, such as transcript or exon) [14]. The GoM model is a generalization of a cluster model, which allows that each sample has some proportion ("grade") of membership, in each cluster. For RNA-seq data this corresponds to assuming that each sample n has some proportion of its reads, q_{nk} coming from cluster k. In addition, each cluster k is characterized by a probability vector, $\theta_{k\cdot}$, whose gth element represents the relative expression of gene g in cluster k. The GoM model is then

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim \text{Multinomial}\left(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG}\right), \tag{1}$$

where

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

The number of clusters K is set by the analyst, and it can be helpful to explore multiple values of K (see Discussion).

To fit this model to RNA-seq data, we exploit the fact that exactly the same GoM model is commonly used for document clustering [34]. This is because, just as RNA-seq samples can be summarized by counts of reads mapping to each possible gene in the genome, document data can be summarized by counts of each possible word in a dictionary. Recognizing this allows existing methods and software for document clustering to be applied directly to RNA-seq data. Here we use the R package `maptpx` [15] to fit the GoM model.

Fitting the GoM model 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" [17], which is named for its widespread use in visualizing the results of the *Structure* software [16] 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 Fig 1 for example.

To help biologically interpret the clusters inferred by the GoM model we also implemented methods to identify, for each cluster, which genes are most distinctively differentially expressed in that cluster (see Methods). Functions for fitting the GoM model, plotting the structure plots, and identifying the distinctive ("driving") genes in each cluster, are included in our R package `CountClust` [46] available through Bioconductor [31].

Results

Bulk RNA-seq data of human tissue samples

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/utilities/gene_names_all_gtex.txt).

We fit the GoM model to these data, with number of clusters $K = 5, 10, 15, 20$. For each K we ran the fitting algorithm three times and kept the result with the highest log-likelihood. Fig 1(a) shows the Structure plot for $K = 20$, with results for other K in S1 Fig. (See also http://stephenslab.github.io/count-clustering/project/src/tissues_tSNE_2.html for an alternative visualization using a 2-dimensional projection with t-SNE [22, 23].)

In all cases results reflect the known division of samples into tissues: that is, samples from the same tissue tend to have similar cluster membership proportions. As might be expected, increasing K highlights finer structure in the data, with tissues that cluster together with smaller K being subdivided into distinct subgroups for larger K. For brevity we focus on results for $K = 20$ (Fig 1(a)). Here some tissues are represented by essentially a single cluster (e.g. Pancreas, Liver), whereas other tissues are represented as a mixture of multiple clusters (e.g. Thyroid, Spleen). Furthermore, the results highlight biological similarity among some tissues by assigning similar membership proportions to samples from those tissues. For example, samples from several different parts of the brain often have similar memberships, as do the arteries (aorta, tibial and coronary) and skin samples (sun-exposed and un-exposed).

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 Fig 1a (see also S1 Table). Reassuringly, many results align with known biology. For example, the purple cluster (cluster 18), which distinguishes Pancreas from other tissues, is enriched for genes responsible for digestion and proteolysis, (e.g. *PRSS1*, *CPA1*, *PNLIP*). Similarly the yellow cluster (cluster 12), which primarily distinguishes Cell EBV Lymphocytes from other tissues, is enriched with genes responsible for immune responses (e.g. *IGHM*, *IGHG1*) and the pink cluster (cluster 19) which mainly shows up in Whole Blood, is enriched with genes related hemoglobin complex and oxygen transport (e.g. *HBB*, *HBA1*, *HBA2*). Further, Keratin-related genes characterize the skin cluster (cluster 6, light denim), Myosin-related genes characterize the muscle skeletal cluster (cluster 7, orange), etc. The royal purple cluster (cluster 1) has memberships in most tissues and the genes distinguishing the cluster seem to be responsible for nucleus and nucleoplasm related functionality. 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 genes in the red cluster (cluster 3), which is common to Breast Mammary tissue, Adipose Subcutaneous and Adipose Visceral, are related to adipocytes and/or fatty acid synthesis; and the top genes in the salmon cluster (cluster 4), which is common to the Gastroesophageal Junction, Esophagus Muscularis and Colon Sigmoid, are related to smooth muscle.

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 only three clusters to all brain tissues (clusters 1,2 and 9 in Fig 1(a)), and we suspected that additional substructure might be uncovered by analyzing the brain samples separately with larger K. Fig 1(b) shows the Structure plot for $K = 6$ 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 (lime green), which is enriched with genes related to cell periphery and communication (e.g. *PKD1*, *CBLN3*) as well as 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 (yellow-orange), the top defining gene for the cluster

being *MBP* which is involved in myelination of nerves in the nervous system [44]. Another driving gene, *GFAP*, participates in system development by acting as a marker to distinguish astrocytes during development [4].

The remaining samples all show membership in multiple clusters. Samples from the putamen, caudate and nucleus accumbens show similar profiles, and are distinguished by strong membership in a cluster (cluster 4, bright red) whose top driving gene is *PPP1R1B*, a target for dopamine. And cortex samples are distinguished from others by stronger membership in a cluster (cluster 2, turquoise in Fig 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 [3].

Single-cell RNA-seq data

Recently RNA-sequencing has become viable for single cells [11], and this technology has the promise to revolutionize understanding of intra-cellular variation in expression, and regulation more generally [12]. 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 100 and $10,000$ to mimic the lower sequencing depth of a typical single cell experiment. For the thinned GTEx data the Structure plot for $K = 20$ preserves most of the major features of the original analysis on unthinned data (S1 Fig). For the accuracy comparisons with distance-based methods, both methods suffer reduced accuracy in thinned data, but the GoM model remains superior (S3 Fig). For example, when thinning by a factor of $10,000$, 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* [24] and Deng *et al* [25].

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 Fig S7. Our hope was that applying our method to these data would identify, and perhaps refine, the cluster structure evident in [24] (their Fig 2A and 2B). However, our method yielded rather different results (Fig 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 [28] and in single cell data specifically [29]. 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 [26, 27]. However, why these batch effects are not evident in Fig 2A and 2B of [24] is unclear.

Deng et al, 2014

Deng *et al* collected single-cell expression data of mouse preimplantation embryos from the zygote to blastocyst stage [25], with cells from four different embryos sequenced at each stage. The original analysis [25] focuses 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 [25] suggested perhaps 6-7 clusters, and we focus here on results with $K = 6$.

The results from the GoM model (Fig 4) clearly highlight changes in expression profiles that occur through early embryonic development stages, and enrichment analysis of the driving genes in each cluster (Table 3, S4 Table) indicate that many of these expression changes reflect important biological processes during embryonic preimplantation development.

In more detail: Initially, at the zygote and early 2-cell stages, the embryos are represented by a single cluster (blue in Fig 4) that is enriched with genes responsible for germ cell development (e.g., *Bcl2l10* [54], *Spin1* [55]). Moving through subsequent stages the grades of membership evolve to a mixture of blue and magenta clusters (mid 2-cell), a mixture of magenta and yellow clusters (late 2-cell) and a mixture of yellow and green (4-cell stage). The green cluster then becomes more prominent in the 8-cell and 16-cell stages, before dropping substantially in the early and mid-blastocyst stages. That is, we see a progression in the importance of different clusters through these stages, from the blue cluster, moving through magenta and yellow to green. By examining the genes distinguishing each cluster we see that this progression reflects the changing relative importance of several fundamental biological processes. The magenta cluster is driven by genes responsible for the beginning of transcription of zygotic genes (e.g., *Zscan4c-f* show up in the list of top 100 driving genes : see `https://stephenslab.github.io/count-clustering/project/src/deng_cluster_annotations.html`), which takes place in the late 2-cell stage of early mouse embryonic development [57]. The yellow cluster is enriched for genes responsible for heterochromation *Smarcc1* [58] and chromosome stability *Cenpe* [59] (see S4 Table) . And the 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 Fig 4) dominate. The orange cluster is enriched with genes involved in the formation of trophectoderm (TE) (e.g., *Tspan8*, *Krt8*, *Id2* [52]), while the purple cluster is enriched with genes responsible for the formation of inner cell mass (ICM) (e.g., *Pdgfra*, *Pyy* [53]). Thus these two clusters are consistent with the cell lineages at early blastocyst (32-cell stage). These findings, however, are incongruent with the literature of three cell types within the blastocyst (ref). The primitive endoderm (PE) and the epiblast (EPI) cells are formed from the ICM cells in mid to late blastocyst [2].

To explain the blastocyst results, we performed additional analysis investigating embryos within the blastocyst stages and compared to a previous single-cell qPCR

study for mouse preimplantation development [52]. Their data revealed three distinct clusters of cell samples within the blastocyst based on 48 marker genes that are of known function in early development or known differential expression within the blastocyst. They found these three clusters to be enriched for genes responsible for the formation of TE, PE, and EPI. To compare to their results, we fitted the GoM model on the same 48 marker genes with three clusters (**??**) and identified the driving genes of each cluster (**??**). Unlike Guo et al, there is not a clear relationship between the three GoM clusters and the three cell types within the blastocyst, neither in the driving genes nor in the transcriptional expression patterns (**??**). Other visualization tools, however, also reveal a two-cluster pattern (S15 Fig). [Say something about the technical artifacts in qPCR study versus in RNA-study?]

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 (Fig 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.

Existing visualization tools

Common methods for visualizing the structure of gene expression data include PCA, multidimensional scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), and distance-based methods such as hierarchical clustering. All of these methods, except for distance-based methods, are techniques for dimension reduction of multivariate data. The visualizations produced by these methods may vary and depend on the structure of sample variation in the dataset. To highlight their differences and similarities, we provide visualizations produced by these methods for all the example datasets in the supplementary materials.

PCA and MDS map multivariate gene expression data to linear dimensions which reveal global variation between samples. The sample coordinates in PCA/MDS dimensions are traditionally presented in two-dimensional scatter plots. For the GTEx data, the PCA plot of the first two dimensions reveals that the whole blood samples are the most distinct among all the GTEx tissue samples, followed by the brain tissue samples. The fifth PC dimension further suggests that transformed fibroblasts and transformed lymphocytes are distinct from the other tissue samples (**??**). Although PCA plots may be useful in spotting global sample variation, the visualization of all PCA dimensions is difficult to interpret. Currently, there does not exist a compact visualization for representing all the PCA/MDS dimensions for high dimensional data, except for making series of bi-scatter plots.

t-SNE was specifically developed for compact visualization using two-dimensional plots. For the GTEx data and the mouse embryos datasets, the solutions produced by t-SNE appear to produce accurate clustering of samples that correspond to their tissue labels. However, these visualizations provide little insight into the similarity between GTEx tissues. Applying to the GTEx brain data, t-SNE indicates no structure among

the brain samples and the visualization does not group brain samples by subregions. Interesting, the t-SNE visualization of the GTEx brain data is similar to the dendrogram produced by hierarchical clustering.

Hierarchical clustering is a distance-based clustering method that is frequently used to detect the substructure of gene expression data. The method computes pairwise distance between samples and applies agglomerative clustering technique to produce a dendrogram visualization. The number of distances to be computed increases exponentially with the number of samples. For instance, the GTEx visualization using hierarchical clustering requires xx calculations. We performed calculations on a xx computer and were unable to generate the distance matrix for the GTEx v6 data due to the limited computing power. However, previously GTEx pilot study analyzed xx tissue samples. Their visualization revealed that whole blood samples and brain tissue samples are the out-groups of all the GTEx tissue samples. These results are consistent with the PCA/MDS visualization. Furthermore, we used the GTEx data to test whether the GoM model is more accurate in detecting substructure than distance-based clustering methods. 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 81% of comparisons, compared with 29% for the distance-based method (Fig 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.

Another method that has close connection to the GoM models is sparse factor analysis (SFA). See [21] for discussion of relationships among these methods in the context of inferring population genetic structure. The GoM model requires all factors to be positive and their loadings summing up to 1. While, SFA is more flexible than the GoM model and allows the factors to be sparse as well as to hold either positive or negative loadings. However, the flexibility of SFA comes at the cost of making the results harder to interpret or to visualize. We have developed methods to visualize results from SFA for this particular application, but we feel that the visualization produced by SFA does not have the immediate appeal of the Structure plot. On the other hand, the flexibility of SFA in modeling both positive and negative contribution of underlying factors might allow the model to capture features that cannot be easily captured by the GoM models. Future work is needed to investigate the application of SFAT for detecting the structure of gene expression data.

[Say something about NMF here?] Although people have applied NMF to expression data, they never used it to interpret the results and look at visualization as we have.

In summary, our examples illustrate that the GoM models provide a more visually and biologically appealing summary of the data than the other dimension reduction techniques. In particular, for the GTEx data, the GoM models combined with Structure plot provides a convenient visualization of tissue samples in 20 dimensions corresponding to 20 clusters. While, PCA, MDS and t-SNE provide solutions that are difficult to visualize more than two dimensions at one time. For the GTEx brain tissue samples, PCA and MDS produce a solution in which most of the brain regions are not well distinguished except for Cerebellum and Cerebellar Hemisphere, and t-SNE fails to uncover a clustering pattern of the brain tissue samples (**??**). While, the GoM model revealed possibly four cell types underlying the brain subregions that are suggested as similar in PCA and MDS. Finally, applying to embryonic cell samples collected at each developmental stage, PCA and MDS recover a pattern of sample variation that is consistent with the developmental stages from which the embryos were collected. t-SNE groups cell samples by their embryonic stage. At the same time, the GoM models revealed a rich admixture of cell types that reflect important biological processes during embryonic preimplantation development.

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 the results for the data from Jaitin *et al* [24] and Deng *et al* [25]: 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 stomach samples share memberships in common with both the pancreas (purple) and the adrenal gland (light green). This pattern can be seen in the Structure plot (Fig 1) but would be hard to discern from a standard hierarchical clustering.

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 [25] ($n = 259; K = 6$), through 33 minutes for the data from [24] ($n = 1,041; K = 7$) to 3,370 minutes for the GTEx data ($n = 8,555; K = 20$). 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 [49] 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 [17]. 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 [20], 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 [20] and we would expect the same to happen here for RNA-seq data.

Methods and Materials

Model Fitting

We use the `maptpx` R package [15] 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 [15] for details.

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 [22, 23]. Compared with the Structure plot our t-SNE plots contain less information, but can better emphasize 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. `http://stephenslab.github.io/count-clustering/project/src/tissues_tSNE_2.html`). This may be particularly helpful in settings where no external information is available to aid in making an informative Structure plot.

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 [18].) 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 [30].

For each cluster k, we filter out a number of genes (top 100 for the Deng *et al* data [25] and GTEx V6 data [13]) 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/`) [50] [51]. See Table 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.

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 dendrogram 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.

Fig 2 shows, for each pair of tissues, whether each method successfully partitioned the samples into the two tissue groups.

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.

Code Availability

Our methods are implemented in an R package `CountClust`, available as part of the Bioconductor project at `https://www.bioconductor.org/packages/3.3/bioc/html/CountClust.html`. The development version of the package is also available at `https://github.com/kkdey/CountClust`.

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

Acknowledgments

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

the analyses described in this manuscript were obtained from: the GTEx Portal on 10/19/2015 and dbGaP accession number phs000424.v6.p1.

The paper is supported by the grant U01CA198933 from the NIH BD2K program.

We thank Matt Taddy, Amos Tanay and Effi Kenigsberg for helpful discussions. We thank Po-Yuan Tung, John Blischak and Jonathan Pritchard for helpful comments on the draft manuscript.

Disclosure Declaration

The authors have no conflict of interest.

Supporting Information

S1 Fig. Structure plot of GTEx V6 all tissue samples for (a) $K = 5$, (b) $K = 10$, (c) $K = 15$, (d) $K = 20$. Some tissues form a separate cluster from the rest of the tissues from $K = 5$ onwards (for example: Whole Blood, Skin), whereas some tissues only form a distinctive subgroup only at $K = 20$ (for example: Arteries).

S2 Fig. Structure plot of GTEx V6 all tissue samples K=20 in 2 runs under the thinning parameters settings (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 there are a few differences from the unthinned version.

S3 Fig. 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 (Fig 2), but even then the model-based method remains more successful than the hierarchical clustering in separating the samples by tissue or origin.

S4 Fig. GTEx data visualization of all tissue samples using (a) principle component analysis and (b) t-SNE (c) Multidimensional scaling. Samples of matching tissue types are indicated by points of matching color.

S5 Fig. Mouse embryo single cell sample visualization using (a) principle component analysis and (b) t-SNE (c) Multidimensional scaling. Single cell samples collected at the same developmental stage are indicated by points of matching color.

S6 Fig. Visualization of loadings from Sparse Factor Analysis on GTEx data. Stacked bar chart visualization of loadings Sparse Factor Analysis (SFA) on 8555 tissue samples across the 53 tissue types in GTEx V6 data. **(left)**: when the loadings are sparse. **(right)** when the factors are sparse.

S7 Fig. Visualization of loadings from Sparse Factor Analysis on Deng et al single cell developmental phase data. Stacked bar chart visualization of loadings Sparse Factor Analysis (SFA) on 259 single cell samples across developmental phases in mouse embryo for Deng *et al* data. **(left)**: when the loadings are sparse. **(right)** when the factors are sparse.

S8 Fig. Visualization of loadings from Sparse Factor Analysis on GTEx Brain data. Stacked bar chart visualization of loadings Sparse Factor Analysis (SFA) on 1259 samples from different Brain tissues in the GTEx V6 data. **(left)**: when the loadings are sparse. **(right)** when the factors are sparse.

S9 Fig. Comparison between GoM model and hierarchical in terms of power to separate samples from pairs of tissues. A comparison of accuracy of GoM model vs hierarchical clustering. Image plots to compare the GoM model with 4 different hierarchical clustering models on various transformations of the data. For each pair of tissues from the GTEx data we assessed whether or not each method (with $K = 2$ clusters) separated the samples precisely according to their actual tissue of origin, with successful separation indicated by a filled square. Very clearly, the GoM model seems to be more successful in separating pairs of tissues compared to any of the hierarchical clustering approaches. In SubFig (a), hierarchical clustering was performed on log counts per million (cpm) data using Euclidean distance. In SubFig(b), the log cpm data data was mean and scale transformed for each gene and then the hierarchical clustering was performed on the transformed data using the Euclidean distance. In SubFig (d), the hierarchical clustering was performed on counts data with the assumption the counts c_{ng} for each gene have a variance $\bar{c}_g + 1$, which we used to scale while computing distance matrix. In SubFig (e), we took the same scaled data as in SubFig(c), but we additionally performed mean and scale adjustments further so that all genes have expression of mean 0 and variance 1. In SubFig(c), GoM model is used to separate the tissues. Very clearly, GoM model seems to be performing better than any of the hierarchical methods.

S10 Fig. Dendrogram representation for Deng et al (2014) developmental phase single cell data. Dendrogram representation of the 259 cells collected across different developmental phases by Deng et al (2014). The hierarchical clustering in this case has been performed on the log counts per million (cpm) data. Note that the labels in this case are difficult to interpret and also the hierarchical clustering fails to represent the continuum among the different developmental phases as in PCA or the GoM model.

S11 Fig. Dendrogram representation for GTEx Brain tissue level data. Dendrogram representation of the 1259 GTEx brain tissue samples. The hierarchical clustering in this case has been performed on the log counts per million (cpm) data. Note that the labels in this case are not visible because of large sample size but largely it separates out Brain Cerebellar, Cerebellar hemisphere and Brain Spinal Cord, Substantia Nigra from other parts of the brain. But any other structure in the data is not easy to see or interpret.

S12 Fig. Dendrogram representation for GTEx all tissues level data. Dendrogram representation of the 8555 GTEx tissue samples across all the tissues. The hierarchical clustering in this case has been performed on the log counts per million (cpm) data. Very clearly, samples from different tissues seem to cluster together, but any further patterns, for instance, how similar different tissue types are, is hard to see.

S13 Fig. Top oder PC scatter plots for GTEx V6 all tissues data. Scatter plot representation of the different top order PCs of the GTEx samples based on the RNA-seq expression data (log cpm normalized).

S14 Fig. GoM results of Deng data including 48 marker genes. The top panel plots the Structure plot of $K = 3$. The bottom panel plots the expression heatmap of the 48 genes (log2 CPM values). The dendrogram is generated using hierarchical clustering with complete lineage and Euclidean distance.

S15 Fig. Visualization of PCA and t-SNE results of Deng data using 48 marker genes

S1 Table. Cluster Annotations of GTEx V6 data with top driving gene summaries.

S2 Table. Cluster Annotations of GTEx V6 Brain data with top driving gene summaries.

S3 Table. Cluster Annotations of Deng data with top driving genes.

S4 Table. Cluster Annotation of Deng data analysis using 48 genes with top driving gene summaries.

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 10/19/2015 and dbGaP accession number phs000424.v6.p1. The paper is supported by the grant U01CA198933 from the NIH BD2K program. We thank Matt Taddy, Amos Tanay, Effi Kenigsberg, Yoav Gilad and Jonathan Pritchard for helpful discussions. We thank Po-Yuan Tung and John Blischak for helpful comments on a draft manuscript.

Disclosure Declaration

The authors have no conflict of interest.

References

1. Guo G, Huss M, Tong GQ, Wang C, SUn LL, Clarke ND, and Robson P. 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*,18(4): 675-685.

2. Rossant J, and Tam PP. 2009. Blastocyst lineage formation, early embryonic assymmetries and axis patterning in the mouse. *Development*,136(5): 701-13.

3. 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.

4. 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.

5. Eisen MB, Spellman PT, Brown PO and Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25): 14863-14868

6. 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

7. 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

8. D'haeseleer P. 2005. How does gene expression clustering work? *Nat Biotechnol*, 23(12):1499-501

9. 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.

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

11. 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.

12. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res.*, 25, 1491-1498.

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

14. 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

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

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

17. 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.

18. Rosenberg NA. 2005. Algorithms for selecting informative marker panels for population assignment. *J Comput Biol*. 12(9), 1183-201.

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

20. 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.

21. 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.

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

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

24. 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.

25. 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.

26. 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

27. 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.

28. 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.

29. 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

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

31. Gentleman, R., Bates, D., Bolstad, B *et al*. Bioconductor: a software development project. 2003. *Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston*. `https://bioconductor.org/`

32. 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

33. Shiraishi Y, Tremmel G, Miyano S and Stephens M. 2015. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.* 11(12): e1005657

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

35. 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.*

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

37. 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)

38. 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+

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

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

41. Wang N, Gong T, Clarke R, Chen L, Shih IeM, Zhang Z, Levine DA, Xuan J, Wang Y. 2015. UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics.* 31(1): 137-9

42. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II, Wang W. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. 2013. *Bioinformatics.* 29(15): 1865-71

43. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.*. 5(3): 29

44. 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

45. 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`.

46. Dey K, Hsiao J, Stephens M. 2016. `CountClust` : Clustering and Visualizing RNA-Seq Expression Data using Grade of Membership Models. *R package version 0.99.3*, `https://www.bioconductor.org/packages/3.3/bioc/html/CountClust.html`

47. 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.

48. 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

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

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

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

52. 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

53. 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

54. 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.

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

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

57. Falco G, Lee S, Stanghellini I, Bassey UC, Hamatani T, Ko MSH. Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental biology*. 307(2), 539-550.

58. Schaniel C, Ang YS, Ratnakumar K, Cormier C, James T, Bernstein E, Lemischka IR, Paddison PJ. Smarcc1/Baf155 couples self-renewal gene repression with changes in chromatic structure in mouse embroynic stem cells. *Stem cells*. 27(12), 2979-91.

59. Putkey FR, Cramer T, Morphew MK, Silk AD, Johnson RS, Mclntosh JR, Cleveland. Unstable Kinetochore-Microtubule capture and chromosomal instability following deletion of CENP-E. *Developmental cells*. 3(3), 351-365.

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

Cluster	Top 5 Driving Genes	Top significant GO terms (function)[q-value]
1. Royal purple	*NEAT1, IGFBP5, CCLN2, SRSF5, PNISR*	GO:0005654 (nucleoplasm)[2e-10], GO:0044822 (poly-A RNA binding)[3e-09], GO:0044428 (nuclear part)[1e-09], GO:0043233 (organelle lumen)[2e-08]
2. Light purple	*SNAP25, FBXL16, NCDN, SNCB, SLC17A7*	GO:0097458 (neuron part)[2e-25], GO:0007268 (synaptic transmission)[9e-18], GO:0030182 (neuron differentiation)[2e-14], GO:0022008 (neurogenesis)[1e-13], GO:0007267 (cell-cell signaling)[3e-13]
3. Red	*FABP4, PLIN1, FASN, GPX3, LIPE*	GO:0044255 (cellular lipid metabolism)[1e-09], GO:0006629 (lipid metabolism)[1e-09], GO:0006639 (acylglycerol metabolism)[3e-08], GO:0045765 (angiogenesis regulation)[4e-08], GO:0019915 (lipid storage)[4e-08]
4. Salmon	*ACTG2, MYH11, SYNM, MYLK, CSRP1*	GO:0043292 (contractile fiber)[3e-13], GO:0006936 (muscle contraction)[5e-12], GO:0030016 (myofibril)[5e-12], GO:0015629 (actin cytoskeleton)[2e-12], GO:0005925 (focal adhesion)[6e-11]
5. Denim	*RGS5, MGP, AEBP1, IGFBP7, MFGE8*	GO:0005578 (proteinaceous extracellular matrix)[4e-20], GO:0030198 (extracellular matrix)[2e-18], GO:0007155 (cell adhesion)[4e-14], GO:0001568 (blood vessel development)[4e-13]
6. Light denim	*KRT10, KRT1, KRT2, LOR, KRT14*	GO:0008544 (epidermis development)[3e-12], GO:0043588 (skin development)[5e-10], GO:0042303 (molting cycle)[8e-06], GO:0042633 (hair cycle)[7e-06], GO:0048513 (organ development)[6e-05]
7. Orange	*NEB, MYH1, MYH2, MYBPC1, ACTA1*	GO:0043292 (contractile fiber)[6e-52], GO:0030016 (myofibril)[1e-51], GO:0030017 (sarcomere)[5e-40], GO:0003012 (muscle system process)[2e-25], GO:0015629 (actin cytoskeleton)[1e-25]
8. Light orange	*FN1, COL1A1, COL1A2, COL3A1, COL6A3*	GO:0030198 (extracellular matrix)[6e-29], GO:0043062 (extracellular structure)[4e-29], GO:0032963 (collagen metabolism)[3e-16], GO:0030199 (collagen fibril organization)[1e-14], GO:0030574 (collagen catabolism)[1e-14]
9. Green	*MBP, GFAP, MTURN, HIPK2, CARNS1*	GO:0043209 (myelin sheath)[4e-07], GO:0007399 (nervous system development)[4e-05], GO:0008366 (axon ensheathment)[9e-05], GO:0044430 (cytoskeletal part)[1e-04], GO:0005874 (microtubule)[3e-04]
10. Light green	*CYP17A1, CYP11B1, PIGR, GKN1, STAR*	GO:0006694 (steroid biosynthesis)[2e-13], GO:0008202 (steroid metabolism)[1e-12], GO:0016125 (sterol metabolism)[1e-11], GO:0042446 (hormone biosynthesis)[1e-10], GO:0008207 (C21-steroid hormone metabolism)[3e-10]
11. Turquoise	*MPZ, APOD, PMP22, PRX, NGFR*	GO:0007272 (ensheathment of neurons)[4e-07], GO:0008366 (axon ensheathment)[7e-07], GO:0042552 (myelination)[7e-06], GO:0048856 (anatomical structure development)[1e-06], GO:0005578 (proteinaceous extracellular matrix)[1e-06]
12. Yellow	*IGHM, IGHG1, IGHG2, IGHG4, CD74*	GO:0006955 (immune response)[1e-18], GO:0002252 (immune effector process)[7e-18], GO:0003823 (antigen binding)[1e-15], GO:0019724 (B-cell mediated immunity)[5e-13], GO:0002684 (positive regulation immune system)[6e-13]
13. Sky blue	*TG, PRL, GH1, PRM2, PRM1*	GO:0019953 (sexual reproduction)[8e-10], GO:0048232 (male gamete generation)[2e-08], GO:0035686 (sperm fibrous sheath)[4e-06], GO:0005179 (hormone activity)[6e-05], GO:0042403 (thyroid hormone metabolism)[2e-04]
14. Light pink	*NPPA, MYH6, TNNT2, ACTC1, MYBPC3*	GO:0045333 (cellular respiration)[2e-34], GO:0022904 (respiratory electron transport)[8e-33], GO:0015980 (energy derivation by oxidation of organic compounds)[4e-30], GO:0031966 (mitochondrial membrane)[5e-26]
15. Light gray	*KRT13, KRT4, MUC7, CRNN, KRT6A*	GO:0070062 (extracellular exosome)[2e-23], GO:0043230 (extracellular organelle)[3e-23], GO:0031982 (vesicle)[3e-20], GO:0008544 (epidermis development)[2e-18], GO:0043588 (skin development)[1e-13]
16. Gray	*SFTPBβ, SFTPA1, SFTPA2, SFTPC, A2M*	GO:0001525 (angiogenesis)[5e-08], GO:0001944 (vasculature development)[2e-07], GO:0048514 (blood vessel morphogenesis)[2e-07], GO:0040012 (locomotion regulation)[4e-06], GO:2000145 (cell motility)[1e-05]
17. Brown	*CSF3R, MMP25, IL1R2, SELL, VNN2*	GO:0006955 (immune response)[8e-22], GO:0006952 (defense response)[9e-16], GO:0071944 (cell periphery)[7e-15], GO:0005886 (plasma membrane)[7e-15], GO:0050776 (regulation of immune response)[2e-12]
18. Purple	*PRSS1, CPA1, PNLIP, CELA3A, GP2*	GO:0007586 (digestion)[3e-14], GO:0004252 (serine-type endopeptidase activity)[4e-08], GO:0006508 (proteolysis)[6e-06], GO:0016787 (hydrolase activity)[6e-05], GO:0044241 (lipid digestion)[1e-04]
19. Pink	*HBB, HBA2, HBA1, FKBP8, HBD*	GO:0005833 (hemoglobin complex)[1e-13], GO:0015669 (gas transport)[4e-11], GO:0020037 (heme binding)[3e-07], GO:0031720 (haptoglobin binding)[3e-06], GO:0006950 (response to stress)[6e-04]
20. Dark gray	*ALB, HP, FGB, FGA, ORM1*	GO:0072562 (blood microparticle)[1e-27], GO:0043230 (extracellular organelle)[1e-24], GO:0044710 (single organism metabolism)[7e-20], GO:0019752 (carboxylic acid

Table 2. Cluster Annotations for GTEx V6 Brain data.

Cluster	Top 5 Driving Genes	Top significant GO terms
1. Royal blue	*CLU, OXT, GLUL, NDRG2, CST3*	GO:0043230 (extracellular organelle), GO:1903561 (extracellular vesicle), GO:0070062 (extracellular exosome), GO:0006950 (response to stress), GO:0031988 (membrane bound vesicle)
2. Turquoise	*ENC1, NCALD, YWHAH, KIF5A, NPTXR*	GO:0097458 (neuron part), GO:0008092 (cytoskeletal protein binding), GO:0031175 (neuron projection development), GO:0030182 (neuron differentiation), GO:0007268 (synaptic transmission)
3. Lime green	*PKD1, CBLN3, CHGB, COL27A1, ABLIM1*	GO:0005089 (Rho guanyl-nucleotide exchange factor activity), GO:0022008 (neurogenesis), GO:0035239 (tube morphogenesis), GO:0016604 (neuron body), GO:0006836 (neurotransmitter transport)
4. Red	*PPP1R1B, RGS14, NCDN, PDE1B, RAP1GAP*	GO:0065009 (regulation of molecular function), GO:0036477 (somatodendritic compartment), GO:0007268 (synaptic transmission), GO:0023051 (signaling regulation), GO:0010646 (cell communication regulation)
5. Yellow orange	*MBP, GFAP, TF, MTURN, SCD*	GO:0043209 (myelin sheath), GO:0007399 (nervous system development), GO:0007272 (ensheathment of neurons), GO:0048471 (perinuclear region of cytoplasm), GO:0010646 (cell communication regulation)
6. Yellow	*IQGAP1, A2M, C3, MYH7, TG*	GO:0072562 (blood microparticle), GO:0044449 (contractile fiber part), GO:0043230 (extracellular organelle), GO:0030017 (sarcomere), GO:0072376 (protein activation cascade)

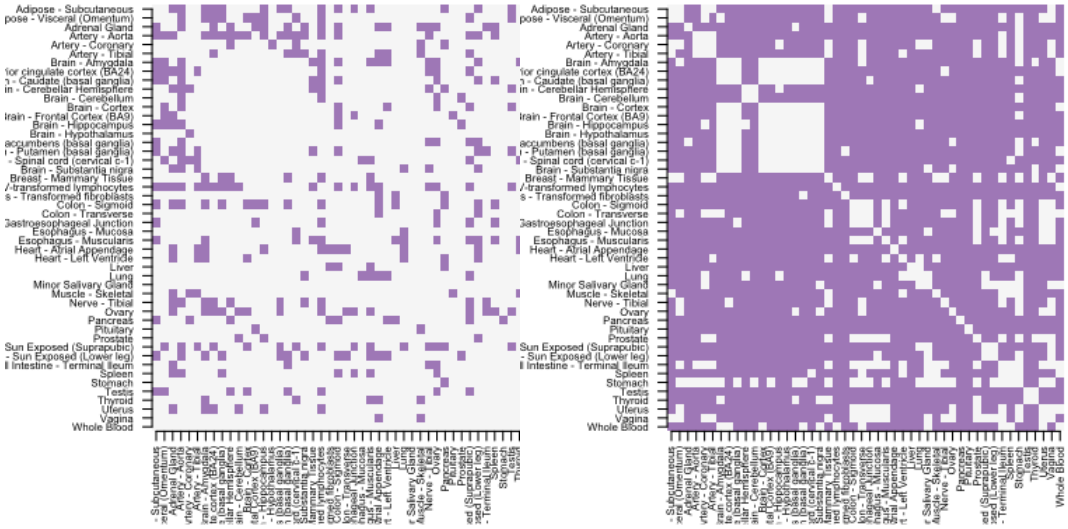
Table 3. Cluster Annotations for Deng et al (2014) data.

Cluster	Top 10 Driving Genes	Top significant GO terms
1. Blue	*Bcl2l10*, *E330034G19Rik*, *Tcl1*,*LOC100502936*, *Oas1d*, *AU022751*, *Spin1*, *Khdc1b*, *D6Ertd527e*, *Btg4*	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)
2. Magenta	*Obox3*, *Zfp352*, *Gm8300*, *Usp17l5*, *BB287469*, *Rfpl4b*, *Gm2022*, *Gm5662*, *Gm11544* , *Gm4850*	GO:0016604 (nuclear body), GO:0005814 (centriole), GO:0044450 (microtubule organizing center part)
3. Yellow	*Rtn2*, *Ebna1bp2*, *Zfp259*, *Nasp*, *Cenpe*, *Rnf216*, *Ctsl*, *Tor1b*, *Ankrd10*, *Lamp2*	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)
4. Green	*Timd2*, *Isyna1*, *Alppl2*, *Prame15*, *Fbxo15*, *Tceb1*, *Gpd1l*, *Pemt*, *Hsp90aa1*, *Hsp90ab1*	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)
5. Purple	*Upp1*, *Tdgf1*, *Aqp8*, *Fabp5*, *Tat*, *Pdgfra*, *Pyy*, *Prdx1*, *Col4a1*, *Spp1*	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)
6. Orange	*Actb*, *Krt18*, *Fabp3*, *Id2*, *Tspan8*, *Gm2a*, *Lgals1*, *Adh1* , *Lrp2*, *BC051665*	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)

(a) clusters	Top 5 driving genes
1 : Royal Purple	*NEAT1, IGFBP5, CCLN2, SRSF5, PNISR*
2 : Light purple	*SNAP25, FBXL16, NCDN, SNCB, SLC17A7*
3 : Red	*FABP4, PLIN1, FASN, GPX3, LIPE*
4 : Salmon	*ACTG2, MYH11, SYNM, MYLK, CSRP1*
5 : Denim	*RGS5, MGP, AEBP1, IGFBP7, MFGE5*
6 : Light denim	*KRT10, KRT1, KRT2, LOR, KRT14*
7 : Orange	*NEB, MYH1, MYH2, MYBPC1, ACTA1*
8 : Light orange	*FN1, COL1A1, COL1A2, COL3A1, COL6A3*
9 : Green	*MBP, GFAP, MTURN, H1PK2, CARNS1*
10 : Light green	*CYP17A1, CYP11B1, PIGR, GKN1, STAR*
11 : Turquoise	*MPZ, APOD, PMP22, PRX, NGFR*
12 : Yellow	*IGHM, IGHG1, IGHG2, IGHG4, CD74*
13 : Sky blue	*TG, PRL, GH1, PRM2, PRM1*
14 : Light pink	*NPPA, MYH6, TNNT2, ACTC1, MYBPC3*
15 : Light gray	*KRT13, KRT4, MUC7, CRNN, KRT6A*
16 : Gray	*SFTPB, SFTPA1, SFTPA2, SFTPC, A2M*
17 : Brown	*CSF3R, MMP25, IL1R2, SELL, VNN2*
18 : Purple	*PRSS1, CPA1, PNLIP, CELA3A, GP2*
19 : Pink	*HBB, HBA2, HBA1, FKBP8, HBD*
20 : Dark gray	*ALB, HP, FGB, FGA, ORM1*

(b) clusters	Top 5 driving genes
1 : Royal blue	*CLU, OXT, GLUL, NDRG2, CST3*
2 : Turquoise	*ENC1, NCALD, YWHAH, KIF5A, NPTXR*
3 : Lime green	*PKD1, CBLN3, CHGB, COL27A1, ABLIM1*
4 : Bright red	*PPP1R1B, RGS14, NCDN, PDE1B, RAP1GAP*
5 : Yellow orange	*MBP, GFAP, TF, MTURN, SCD*
6 : Yellow	*IQGAP1, A2M, C3, MYH7, TG*

Fig 1. (a): Structure plot of estimated membership proportions for GoM model with $K = 20$ clusters fit to 8555 tissue samples from 53 tissues in GTEx data. Each horizontal 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 membership proportions for $K = 4$ clusters fit to only the brain tissue samples. This analysis highlights finer-scale structure among the brain samples that is missed by the global analysis in (a).



(a) hierarchy method **(b)** GoM method

Fig 2. A comparison of accuracy of GoM model vs hierarchical clustering. For each pair of tissues from the GTEX data we assessed whether or not each method (with $K = 2$ clusters) separated the samples precisely 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 GoM model is successful in 86% comparisons and the hierarchical clustering in 39% comparisons.

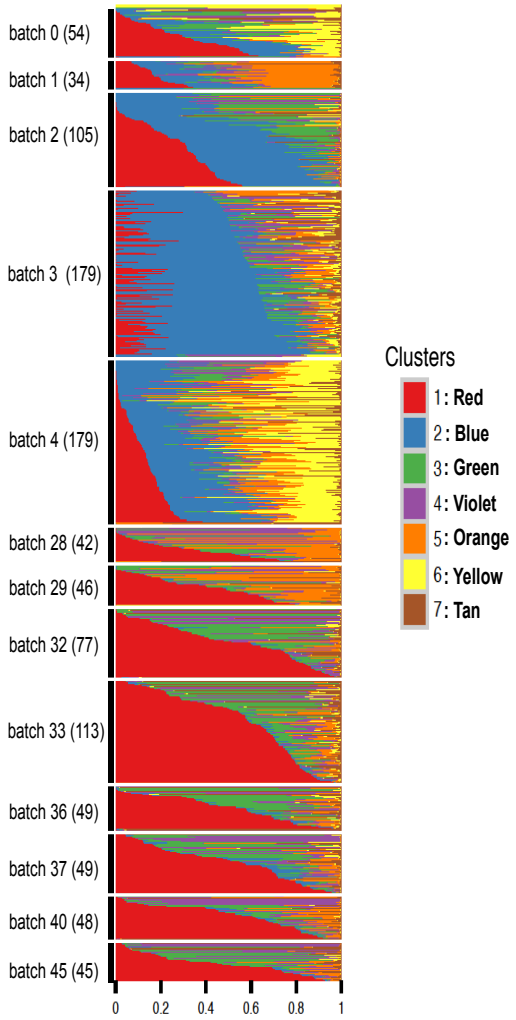


Fig 3. Structure plot of estimated membership proportions for GoM model with $K = 7$ clusters fit to $1,041$ single cells from [24]. The samples (cells) are ordered so that samples from the same amplification batch are adjacent and within each batch, the samples are sorted by the proportional representation 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". Membership proportions are correlated with batch, and some groups of batches (e.g. 28-29; 32-45) show similar palettes. These results suggest that batch effects are likely influencing the inferred structure in these data.

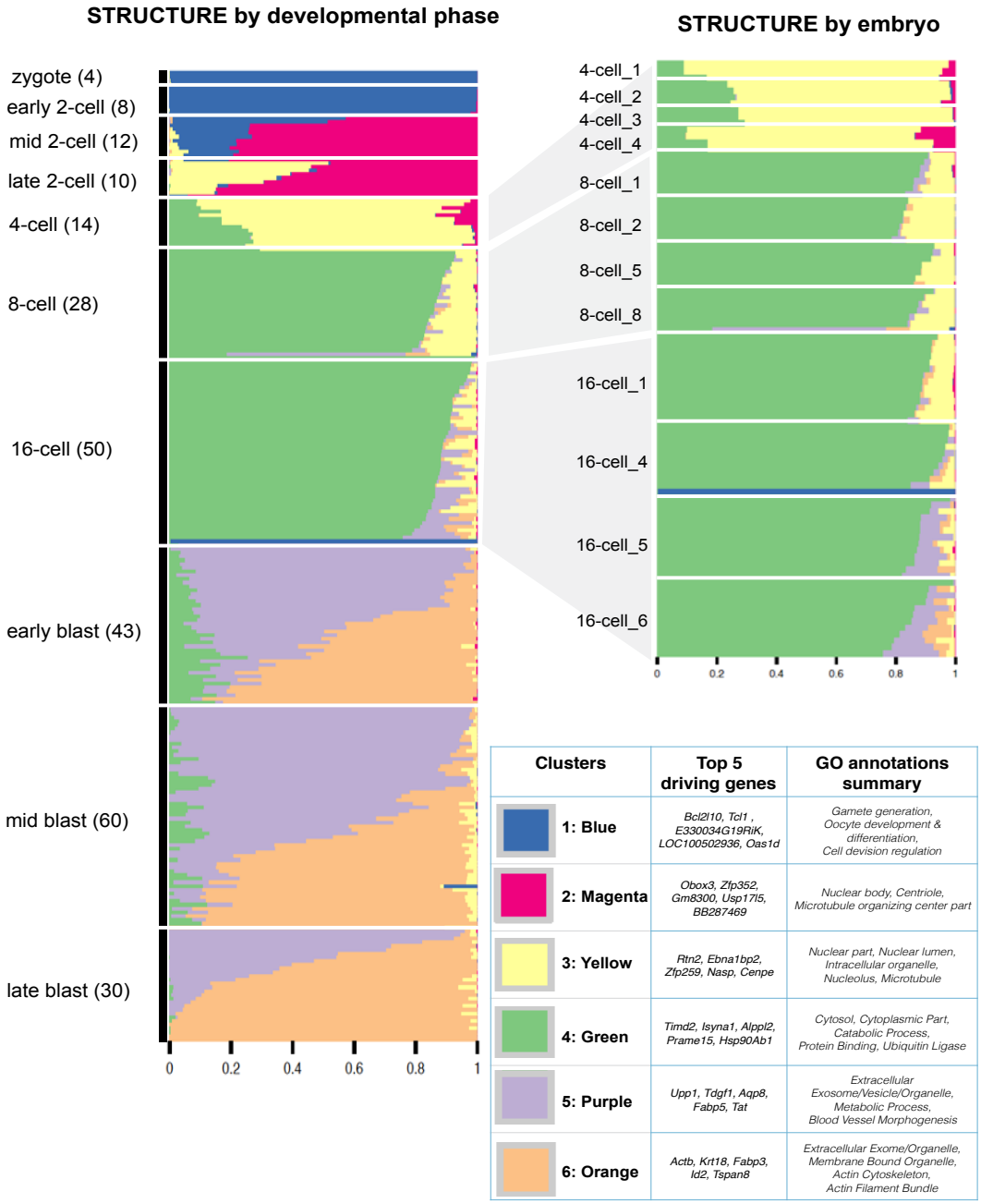


Fig 4. Structure plot of estimated membership proportions for GoM model with $K = 6$ clusters fit to 259 single cells from [25]. The cells are ordered by their preimplantation development phase (and within each phase, sorted by the proportional representation of the clusters). While the very earliest developmental phases (zygote and early 2-cell) are essentially assigned to a single cluster, others have membership in multiple clusters. Each cluster is annotated by the genes that are most distinctively expressed in that cluster, and by the gene ontology categories for which these distinctive genes are most enriched (see Table 3 for more extensive annotation results). See text for discussion of biological processes driving these results.