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

* 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 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 the BioConductor package `CountClust`.

Author Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

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 [3–5]. 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 ([31, 32]) 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" [31].

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 [37]. And, indeed, the GoM model we use here is analogous to – although different in detail from – blind deconvolution approaches [35, 36] which estimate cell type proportions and cell type signatures jointly (see also [33, 34] 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 [31], 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

[1] While preparing this work for publication we became aware of ongoing independent work by [39] applying GoM models to RNA-seq data.

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 Bioconductor package `CountClust` [40].

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 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 Fig 1 for example.

Fig 1. GTEx tissue samples grades of membership. (A) Structure plot of estimated membership proportions for GoM model with $K = 15$ 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).

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 Fig 1(A) (see also an alternative visualization using a 2-dimensional projection with t-SNE [20], [21], in Supplemental Fig `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 **??** summarizes results for the GTEx results in Fig 1(A) (see also $S1_{Table}).Reassuringly, manyresultsalignwithknownbiology.Forexample, thelightredcluster($
PRM2, PRM1, PHF7).Similarlythelightgreencluster(cluster10), whichprimarilydistinguishes
relatedandproteolysis -
relatedgenescharacterizethePancreascluster(cluster13, lightviolet), Keratin -
relatedgenescharacterizetheskincluster(cluster6, skyblue), Myosin -
relatedgenescharacterizethemuscleskeletalcluster(cluster9, green), etc.Incaseswhereacluster
membership.Forexample, thetopfourgenesinthelightorangecluster(cluster8), whichiscommon
TerminalIleum, allcodeforsurfactantproteins(specifically, B, A2, A1andC).$

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. Fig 1(B) shows the Structure plot for $K = 4$ on only the Brain samples. The results highlight much finer-scale structure compared with the global analysis. Brain Cerebellum and Cerebellar hemisphere are essentially assigned to a separate cluster, 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 [38]. 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 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 [1].

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; 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.

Fig 2. Compare accuracy of GoM model versus 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.

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 (S2$_{Fig}$).$Fortheaccuracycomparisonswithdistance-basedmethods, bothmethodssufferreducedaccuracyinthinneddata, buttheGoMmodelremainss$

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].

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 Fig 2(A)-(B)). 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).

Fig 3. Jaitin *et al* single-cell sample estimated membership proportions. Structure plot of estimated membership proportions for GoM model with $K = 7$ clusters fit to $1,041$ single cells from [22]. 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.

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 Fig 2(A)-(B) of [22] is unclear.

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 (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 **??**) indicate that many of these expression changes reflect important biological processes during embryonic preimplantation development.

Fig 4. Deng *et al* single-cell sample estimated membership proportions. Structure plot of estimated membership proportions for GoM model with $K = 6$ 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 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 **??** for more extensive annotation results). See text for discussion of biological processes driving these results.

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* [48], *Spin1* [49]). 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* [51]), which takes place in the late 2-cell stage of early mouse embryonic development. The yellow cluster is enriched for genes responsible for heterochromation *Smarcc1* [52] and chromosome stability *Cenpe* [53]. 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 outer trophoblast cells (e.g., *Tspan8*, *Krt8*, *Id2* [46]), while the purple cluster is enriched with genes responsible for the formation of inner cell mass (e.g., *Pdgfra*, *Pyy* [47]). Thus these two clusters are consistent with the two cell lineages, the trophectoderm and the primitive endoderm, that make up the majority of the cells of the blastocyst [50]. 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 (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.

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 (Fig 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 (Fig 1) combined with the cluster annotations (Table **??**) 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 [43] 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.

Materials and Methods

Model Fitting

We use the `maptpx` R package [13] to fit the GoM model (Eq (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.

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 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. **??**). 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 [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/`) [44, 45]. See Tables **??** - **??** and Table **??** for the top significant gene ontologies driving each cluster in the GTEx V6 data and the Deng *et al* data respectively.

Supporting Information

S1 Fig. Structure plot of the thinned GTEx V6 data with K=15. (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.

S2 Fig. Compare "accuracy" of GoM model versus hierarchical clustering on thinned GTEx data. (A)-(B) thinning parameter $p_{thin} = 0.01$, and (C)-(D) $p_{thin} = 0.0001$. 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.

S1 Table. Cluster Annotations GTEx V6 data of all tissue samples with top gene summaries.

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

S3 Table. Deng et al (2014) cluster top GO annotations.) (A) Blue cluster (Cluster 1) , (B) Magneta cluster (Cluster 2) , (C) Yellow cluster (Cluster 3) , (D) Green cluster (Cluster 4) , (E) Purple cluster (Cluster 5) , (F) Orange cluster (Cluster 6) .

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.

We thank Matt Taddy, Amos Tanay and Effi Kenigsberg 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. Hernandez MC , Andres-Barquin PJ , Martinez S , Bulfone A , Rubenstein JL , Israel MA. ENC-1: a novel mammalian kelch-related gene specifically expressed in the nervous system encodes an actin-binding protein. 1997 *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. Cluster analysis and display of genome-wide expression patterns. 1998 *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. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. 1999 *Science*, 286(5439): 531-7

5. Alizadeh AA1, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. 2000 *Nature*, 403(6769): 503-11

6. D'haeseleer P. How does gene expression clustering work? 2005 *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. Latent class representation of the grade of membership model. 2006 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. 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/`

30. 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

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

32. 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.*

33. 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

34. 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)

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

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

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

38. 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

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

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

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

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

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

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

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

46. 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. 2010 *Developmental Cell.* 18(4), 675-685

47. 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). 2007 *BMC Developmental Biology*. 7(92), 1-13

48. 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. 2009 *Biology of Reproduction*. 81(3), 497-506.

49. Evsikov AV, De Evsikova C. Gene expression during the oocyte-to-embryo transition in mammals. 2009 *Molecular Reproduction and Development*. 76, 805-818.

50. Rossant J. Development of the extraembryonic lineages. 1995 *Seminars in Developmental Biology*. 6(4), 237-247.

51. 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. 2007 *Developmental biology*. 307(2), 539-550.

52. 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. 2009 *Stem cells*. 27(12), 2979-91.

53. 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. 2002 *Developmental cells*. 3(3), 351-365.