On model based clustering of RNA-seq expression data

Kushal K Dey Matthew Stephens

Abstract

We propose a model based approach to cluster the reads level expression for bulk RNA and single cell RNA-seq data. It is similar to the topic model in Natural Language Processing and the admixture model in Population Genetics, that assigns grades of cluster membership to each sample. This approach provides us with easily interpretable cluster visualization and detects the underlying structure in the data better than distance based approaches. It also extracts the important genes that drive the clusters and provides measures of model fit to assess the strength of clustering. Further, we show that this method is robust under low coverage of reads. We apply this method on the GTEx tissue level bulk RNA expression data as well as two single cell RNA-seq data. Our methods are implemented in a R package **CountClust**, available at `https://github.com/kkdey/CountClust`.

1 Introduction

Clustering of samples based on gene expression data is a popular exploratory mechanism in bulk RNA-seq and single cell RNA-seq (scRNA-seq) experiments. It aids in quality control and helps in understanding the heterogeneity across tissue samples (bulk RNA-seq) or single cells (scRNA-seq). Usually the clustering techniques more commonly used in RNA-seq literature are distance based, primarily hierarchical and k-means clustering (see Jaitin *et al* 2014 [12], Buettner *et al* 2015 [21], GTEx Consortium paper [4]). However, the data obtained from the RNA sequencing experiments are counts data, representing the number of reads mapping to different genes. and there exist model based clustering methods based on counts which seem to be directly applicable to the RNA-seq reads data.

The clustering model we propose in this paper is similar to the topic model approach, widely used in Natural Language Processing (see Blei, Ng and Jordan 2003 [24], Blei and Lafferty 2009 [25]), which is derived from the Admixture model in population genetics (see Pritchard, Stephens and Donnelly 2000 [6]).This approach models each tissue sample as having some proportion of its reads coming from each cluster. This assumption is biologically sensible since in reality, each tissue sample indeed is a mixture of different cell types and presumably, the clusters under this model could be driven by the cell types. Also, such graded membership approach is capable of representing more continuous cluster patterns.

Deconvolution techniques using marker genes are popularly used to learn about the the concentration of different cell types in a cell mixture and cell type signature expression profiles. Our technique is analogous to blind deconvolution approach which estimates the cell type proportions and cell type signatures jointly (see Schwartz *et al* 2010 [29] and Repsilber *et al* 2010 [28]), except that we operate under Poisson model framework to model the counts data. This method can also be extended to semi-supervised deconvolution that assumes some information of cell type signature expression (see Shen-Orr *et al* 2010 [26], Qiao *et al* 2012 [27]).

In this paper, we demonstrate that for RNA-seq (bulk or single cell) data with known structural patterns, such count clustering approach identifies the structure better than hierarchical clustering. It also allows one to interpret each cluster by providing information about genes that are playing a significant role in driving the clusters and these genes may be important from both biological and medical standpoint. Also we show our method to be robust even for low coverage data as might be the case for single cell RNA-seq (scRNA-seq) data. We illustrate the performance of our method on GTEx tissue level bulk-RNA seq data as well as on two single cell data (due to Jaitin *et al* 2014 [12] and Deng *et al* 2014 [13]) .

2 Methods and Materials

2.1 Data preprocessing

We assume that the data from an RNA-seq experiment have been summarized by a table of counts $C_{N\times G} = ((c_{ng}))$, where c_{ng} is the number of RNA-seq reads from sample n that mapped to gene (or transcript) g. Such a table of counts is obtained by processing the BAM or FASTQ files obtained from sequencing machines and currently, there is an effort to make such reads table, ready for statistical analysis, publicly available (check the ReCounts website [2]). Before applying the graded membership model on the reads data, we remove the genes with 0 or same count of matched reads across all samples as they are non-informative for clustering. We also remove any sample or gene with NA values of reads and spike-in control genes, as the latter may create bias due to their typically having high number of reads mapped to them [1].

2.2 Model overview

The method we adopt allows each sample to have a grade of membership in the underlying clusters. This type of approach was first used in population genetics to model population samples as having parts of its genome derived from different clusters that represented ancestries (see Pritchard, Stephens and Donnelly 2000 [6]). Later, a similar approach was used to model documents to have grades of membership in different clusters, representing topics, depending on the frequencies of different words used (see Blei, Ng and Jordan 2003 [24]). In RNA-sequencing data too, each sample is viewed as having grades of membership in different underlying subpop-

ulations (for e.g. cell types for tissue samples, cell phases for single cells etc).

We assume that the row vector of counts for each sample n across the genes follows a multinomial distribution.

$$c_{n*} \sim Mult(c_{n..}, p_{n*})$$

where c_{n*} is the count vector for the n th sample, $c_{n..}$ is the sum of the counts in the vector c_{n*}, and p_{n*} is the probability that a read coming from sample n would get assigned to one of the G genes. The main idea is that this read could come from hidden subpopulations (may be cell types for tissue level expression study or cell cycle phases for single cell study) and its probability of getting assigned to some gene g may depend on which subpopulation it comes from. Denote the probability that the sample is coming from the k th subpopulation by q_{nk} ($q_{nk} \geq 0$ and $\sum_{k=1}^{K} q_{nk} = 1$ for each n). Given that the sample is coming from the kth subgroup, the probability of a read being matched to the gth gene is given by θ_{kg} ($\theta_{kg} \geq 0$ and $\sum_{g=1}^{G} \theta_{kg} = 1$ for kth subgroup). Then one can write

$$p_{ng} = \sum_{k=1}^{K} q_{nk}\theta_{kg} \qquad \sum_{k=1}^{K} q_{nk} = 1 \qquad \sum_{g=1}^{G} \theta_{kg} = 1$$

This model has $N \times (K-1) + K \times (G-1)$ many unconstrained parameters, which is much smaller than the $N \times G$ data values of counts. Usually for RNA-seq samples N varies in the region of 100s to 1000s and G ranges from $10,000$ to $50,000$ (depending on the underlying species and the types of genes tokenized) and $K << \{N, G\}$.

Due to the widespread use of topic models in Natural Language Processing literature, many softwares are available for fitting such models on document collections. Here, we use a Maximum a posteriori (MAP) estimation procedure implemented in R package **maptpx**(see Taddy 2012 [5]).

2.3 Visualization

A nice way to visualize the amount of relatedness among the samples is through the Structure plot, which is a popular tool to visualize the admixture patterns in population genetics based on SNP/ microsatellite data (see Rosenberg *et al* 2002 [7]). For each sample n, q_{nk} represents the relative abundance of hidden subgroup k in the sample. The Structure plot assigns a color to each of the subgroups and then presents a vertical stacked barchart for each individual, with stack heights representing subgroup proportions and each subgroup assigned a particular color. If the colored patterns of two bars are similar, then the two samples must be closely related.

Another visualizing tool we recommend is t-distributed Stochastic Neighbor Embedding (t-SNE), which is well-suited for visualizing the high dimensional datasets on 2D, preserving the relative distance between samples in high dimension to a fair extent in 2D (see L.J.P. van der Maaten [10] and L.J.P. van der Maaten and Hinton [9]). t-SNE shows us which samples are close to each other when the data is projected on 2D. t-SNE is not a clustering tool and unlike Structure plot, does not show the relative abundance patterns of different subgroups in the sample. However, both Structure plot and t-SNE give a lot more interpretable visualization of the clustering compared to the heatmap and hierarchical clustering (see Results for illustration).

2.4 Cluster annotation

A question of considerable biological interest is which genes are significantly differentially expressed across the clusters, or in other words, which genes are driving the clustering. To answer this, we fix one cluster and for each gene, define a distance metric between that cluster and any of the other clusters, based on the cluster expression profile of the gene, namely the θ values. For cluster k, we define the distance from cluster l based on expression profile of gene g to be

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

This is similar to the Kullback Leibler divergence of the Poisson distribution with parameter θ_{lg} from another Poisson distribution with parameter θ_{kg}. For each cluster k, we define the divergence measure for gene g as

$$Div^g[k] = \min_{l \neq k} KL^g[k,l]$$

The higher the divergence measure, the more significant is the role of the gene in the clustering. We choose a small subset of around 50-100 genes with highest values of $Div^g[k]$ for each k and this set of genes can be viewed to be the most important genes driving the cluster k.

Once the most important driving genes for each cluster k have been extracted, we perform gene annotations on them using **mygene** R Bioconductor package (due to Mark A, Thompson R and Wu C 2014 [19]). We check if the driving genes for a particular cluster are associated with some specific biological functionality. This would validate whether the subgroups are biologically relevant.

3 Results

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

To assess structure in these data we applied the grade-of-membership model with $K = 10, 12, 15$. Although results differ with K, many of the primary patterns were consistent across K. Here, for brevity, we focus on results for $K = 15$, shown as a Structure plot in **Figure 1(a)** (see also an alternative visualization using a 2-dimensional projection with t-sne [9], [10], in **Supplementary Fig 1** `http://htmlpreview.github.io/?https://github.com/stephenslab/count-clustering/blob/master/project/src/tissues_tSNE.html`). Reassuringly, much of the structure highlighted by these results follows the known division of samples into tissues: that is, samples that come from the same tissue tend to have similar grades of membership across clusters. Some tissues are represented by essentially a single cluster (e.g. Pancreas, Whole Blood), whereas other tissues are represented as a mixture of multiple clusters (e.g. Thyroid). Furthermore, the results highlight biological similarity among some tissues by assigning samples from those tissues similar membership proportions. For example, samples from different parts of the brain have similar memberships, as do the arteries (Artery-aorta, Artery-tibial and Artery-coronary) and skin (Skin Not Sun Exposed- suprapubic and Skin Sun Exposed- lower leg). Samples from the tibial nerve cluster with the adipose tissues (Adipose Subcutaneous and Adipose Visceral (Omentum)), possibly reflecting contamination of those samples with adipose cells.

To biologically interpret the clusters in **Fig 1**, we performed cluster annotation (see Methods and Materials). **Tab ??** presents the gene IDs, names and a short summary of the functions of top 3 driving genes for each cluster. The cluster annotation in **Tab ??** highlights tissue specific functions and pathways. We observe that *PRSS1* (protease serine 1), *CPA1* (carboxypeptidase) and *PNLIP* (pancreatic lipase) are the top three genes that drive the cluster separating Pancreas from the other tissues. Similarly, *HBB* (hemoglobin, beta), *HBA2* (hemoglobin, alpha 2) and *HBA1* (hemoglobin, alpha 1) seem to be the top three genes that distinguish Whole Blood and drive a separate cluster from the rest.

Although global analysis of all tissues is useful for highlighting major structure in the data, it is less well suited to identifying structure within tissues or among similar tissues. Therefore, we ran the analysis on samples from a particular tissue or similar tissues. **Fig 1(b)** shows the Structure plot for $K = 4$ on just the Brain samples. Clearly, we see additional patterns across the Brain tissues in this plot compared to the global Structure plot in **Fig 1(a)**. Brain Cerebellum and Cerebellar hemisphere seem to have $80 - 85\%$ membership proportion in green cluster

which seems distinctive . Recent stereological approaches have shown that rat cerebellum contains $> 80\%$ neurons (Herculano-Houzel and Lent 2005) [18], much higher than other parts of the brain. We performed cluster annotation and observed that the pivotal genes that separated out the green cluster in brain cerebellum and cerebellar hemisphere are SNAP25 (synaptosomal-associated protein, 25kDa), ENO2 (enolase 2- gamma, neuronal) and CHGB (chromogranin B), all of which are associated with neuronal activities (see **Supplementary Table 1**). Therefore, the green cluster does seem to be driven by neuronal cell types, though not completely determined by them (since this cluster is almost non-existent in other parts of the Brain).

We aimed to assess the cluster performance of the graded membership method on lower coverage data as is observed in single cell RNA seq (scRNA seq) data. To compare the cluster performance between bulk-RNA coverage and single cell-RNA coverage, we thinned the GTEx reads data. If c_{ng} is the counts of number of reads mapping to gene g for sample n for the original data, then the thinned counts are given by

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

where p_{thin} represents the proportional coverage of a single cell RNA-seq data compared to the bulk RNA-seq data. We compared the total library size between the GTEx tissue level data and the single cell data due to Jaitin *et al* [12] and observed that $p_{thin} = 0.0001$. We then used this thinning parameter and fitted the clustering model on t_{ng}s. **Fig 1(c)** presents the Structure plot for $K = 15$ for the thinned data. Many of the features from **Fig 1(a)** are preserved even after thinning, for instance the Brain tissues clustering together, Adipose tissues and Nerve Tibial also showing similar patterns. This implies that the clustering model is robust to the coverage of the data.

We next sought to demonstrate more quantitatively the utility of the model based clustering compared to other non model based clustering methods such as hierarchical clustering. In **Fig 2**, we consider every pair of tissues from the list of tissues in GTEX with number of samples > 50. Then we generated a set of 50 samples randomly drawn from the pooled set of samples coming from these two tissues and then observed whether the hierarchical and the admixture were separating out samples coming from the two different tissues. The same remains true for thinned GTEx data under different choices of thinning parameters. Check **Fig 5** for demonstration.

We applied the model on a couple of single cell datasets due to Jaitin *et al* [12] and Deng *et al* [13].

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

Figures 2A and 2B of [12]. However, our method yielded rather different results (Figure 3). Each cell was assigned to multiple clusters, and the cluster membership vectors strongly correlated with amplification batch (which in turn is strongly correlated with sequencing batch). The fact that batch effects are detectable in data like these is not particularly surprising. There is a growing recognition of the importance of batch effects in high-throughput data generally [16] and in single cell data specifically [17]. And indeed, dimension reduction methods such as the ones we use here can be helpful in controlling for such effects [14] [15]. However, why these batch effects are not visible in the original analyses [12], is unclear.

Deng *et al* collected expression data from individual cells from zygote to blastocyst stages of mouse preimplantation development [13]. Deng *et al*'s analysis focussed particularly on allele-specific expression from the two contributing mouse strains (CAST/EiJ and C57BL/6J). Here we analyze the counts of the two alleles combined. Visual inspection of the Principal Components Analysis in [13] suggested 6-7 clusters, so we fit the cluster model with $K = 6$. The results (Figure 4) clearly highlight the structure in the different development stages starting from zygote, through early/mid/late 2 cells, 4 cells, 8 cells, 16 cells, and early/mid blastocyst to finally late blastocyst. Specifically, cells that are from the same stage show similar cluster membership proportions. Further, many of the clusters show notable trends through the stages. For example, membership in the red cluster is non-existent in early stages, starts in the 4-cell stage, becomes more prominent in the 8-16 cell stages, drops substantially in the early and mid-blastocyte stages, and is essentially absent in the late blastocytes. More generally, cluster memberships for cells from adjacent stages tend to be more similar to one another than those for cells from distant stages.

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

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

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

4 Discussions

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

Since our method is model based, it provides an optimal K for the model fit. However, one has to run the model on the data for a range of K's and that is not always practical when running the model on large datasets as in RNA-seq reads data. For a single run with the algorithm running till the successive log posterior increase is less than 0.01, the computation time for the algorithm was approximately 33.4 mins for Jaitin *et al* [12] single cell experiment ($K = 7$), 16.2 mins for Deng *et al* single cell experiment ($K = 6$) and 3297.4 mins for GTEx V6 data ($K = 15$).

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

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

References

1. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR and Oliver B. *Synthetic spike-in standards for RNA-seq experiments.* Genome Res, 2011:21, 1543-1551

2. Frazee AC, Langmead B, Leek JT. *ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets.* BMC Bioinformatics, 2011, 12:449.

3. Amit Zeisel, Ana B. Muoz-Manchado, Simone Codeluppi, Peter Lnnerberg, Gioele La Manno, Anna Jurus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science 6 March 2015: 347 (6226), 1138-1142.

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

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

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

7. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW. The genetic structure of human populations. Science 2002, 298: 2381-2385.

8. Anil Raj, Matthew Stephens, and Jonathan K. Pritchard. *fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets.* Genetics.2014 197:573-589.

9. L.J.P. van der Maaten and G.E. Hinton. *Visualizing High-Dimensional Data Using t-SNE.* Journal of Machine Learning Research, 2008: 2579-2605.

10. L.J.P. van der Maaten. *Accelerating t-SNE using Tree-Based Algorithms.* Journal of Machine Learning Research, 2014:3221-3245.

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

12. Jaitin DA, Kenigsberg E et al. *Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types.* Science, 2014: 343 (6172) 776-779.

13. Deng Q, Ramskold D, Reinius B and Sandberg R. *Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells.* Science, 2014: 343 (6167) 193-196.

14. Jeffrey T. Leek and John D.Storey *Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis* PLoS Genet 3(9): e161. doi:10.1371/journal.pgen.0030161

15. O. Stegle, L. Parts , M. Piipari, J. Winn and R. Durbin *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nat Protoc. 2012 Feb 16;7(3):500-7. doi: 10.1038/nprot.2011.457.

16. Jeffrey T. Leek, Robert B. Scharpf, Hctor C Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature Reviews Genetics 11, 733-739.

17. Stephanie C Hicks, Mingxiang Teng and Rafael A Irizarry *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

18. Herculano-Houzel S and Lent R. *Isotropic fractionator: a simple, rapid method for the quantification of total cell and neuron numbers in the brain.* J Neurosci. 2005 Mar 9;25(10), 2518-21.

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

20. Grn D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. *Single-cell messenger RNA sequencing reveals rare intestinal cell types.* Nature. 2015 Sep 10;525(7568), 251-5.

21. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC and Stegle O. *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells* Nature Biotechnology 2015, 33, 155?160, doi:10.1038/nbt.3102

22. Palmer C, Diehn M, Alizadeh AA and Brown PO. *Cell-type specific gene expression profiles of leukocytes in human peripheral blood.* BMC Genomics 2006, 7:115.

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

24. Blei DM, Ng AY and Jordan MI. *Latent Dirichlet Allocation* Journal of Machine Learning Research 2003, 3, 993-1022

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

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

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

28. 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. BMC bioinformatics 2010, 11(1), 27+

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

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

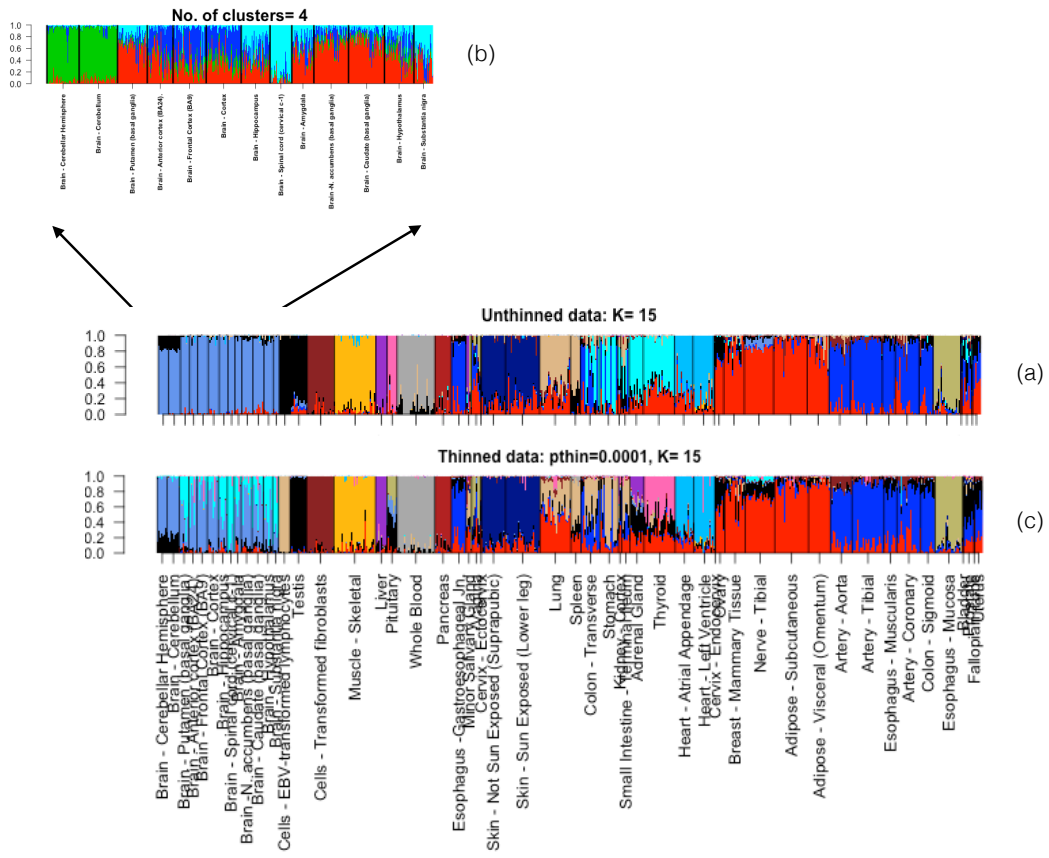
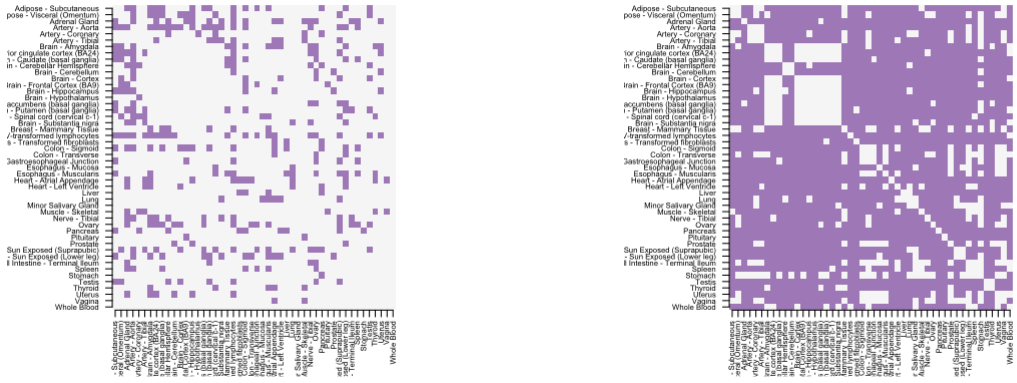


Figure 1. **(a)**: Structure plot of the admixture proportions from model fit with 15 topics/clusters for the 8555 tissue samples coming from 53 tissues in GTEX V6 data based on 16069 cis genes derived using a quality assessment procedure (check Results) . Note that the samples coming from the same tissue have similar admixing patterns. Tissues of same origin, for instance all the Brain tissues, all the arteries seem to cluster together. Also, some other tissues, presumably not of same origin, show markedly similar clustering patterns - for instance Breast mammary tissue, Nerve Tibial and Adipose tissues are very similar in clustering patterns. **(b)**: Structure plot of the cluster memberships from model fit with 4 clusters on just the brain tissue samples. We see more clustering patterns across the Brain tissues in this plot, that was masked by inter tissue variation in (a). Brain cerebellum and cerebellar hemisphere seem to be represented by the green cluster while the Spinal cord and Substantia nigra have high grades of membership in cyan cluster. **(c)**: Structure plot of all tissue samples in GTEx V6 data thinned data with $p_{thin} = 0.0001$ for $K = 15$. The thinning parameter has been chosen so that the GTEx RNA-seq data can be interpreted at the same scale as a scRNA-seq data. The clustering patterns are slightly more noisy compared to the non-thinned version in (a), but overall, the similarity patterns across the tissues are retained.



(a) hierarchy thin 0.1

(b) admixture thin 0.1

Figure 2. A comparison of the hierarchical method with the admixture method. For each pair of tissues, we selected randomly 50 samples and then on the reads data for these 50 samples, we applied the hierarchical clustering method with complete linkage and Euclidean distance and then cut the tree at $K = 2$. We then observed if it separates out the samples coming from the two tissues, in case it does, we color the cell corresponding to that pair of tissues. We apply admixture model on the same data for $K = 2$. Then we fixed one cluster, observed the proportions for that cluster, sorted the samples based on the proportions for that cluster and separated out the samples at the point of maximum jump/fall in the proportions for that cluster. If that separates out the two tissues, we color the cell, else keep it blank. From the graph it seems that the admixture model has been far more successful in separating out different tissues compared to the hierarchical method.

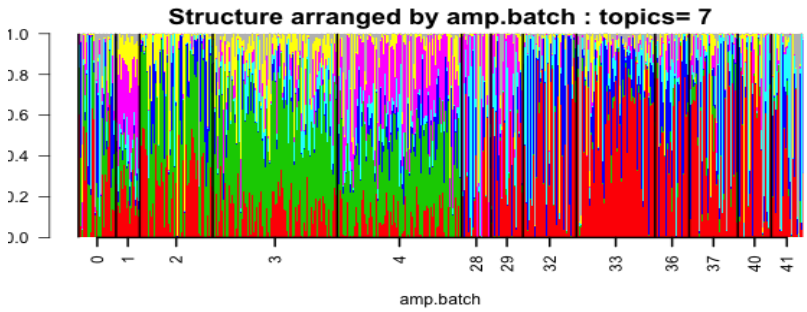


Figure 3. Structure plot of the1041 single cells for K=7 of the Jaitin *et al* data [12] arranged by the amplification batch. It is observed that the clustering patterns within each batch are similar and so, either the amplification batch is driving the clustering or it is confounded with the actual biological effects, making it difficult to interpret these clusters.

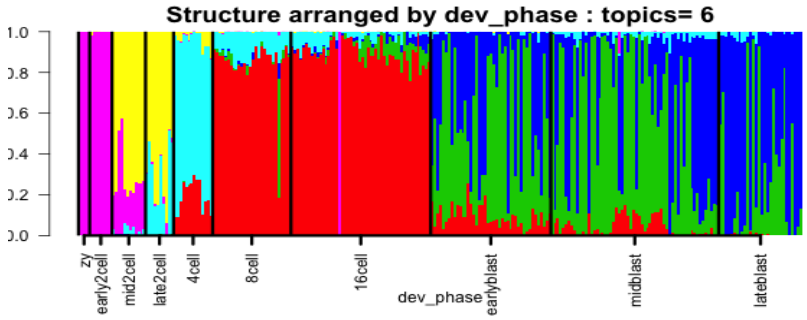
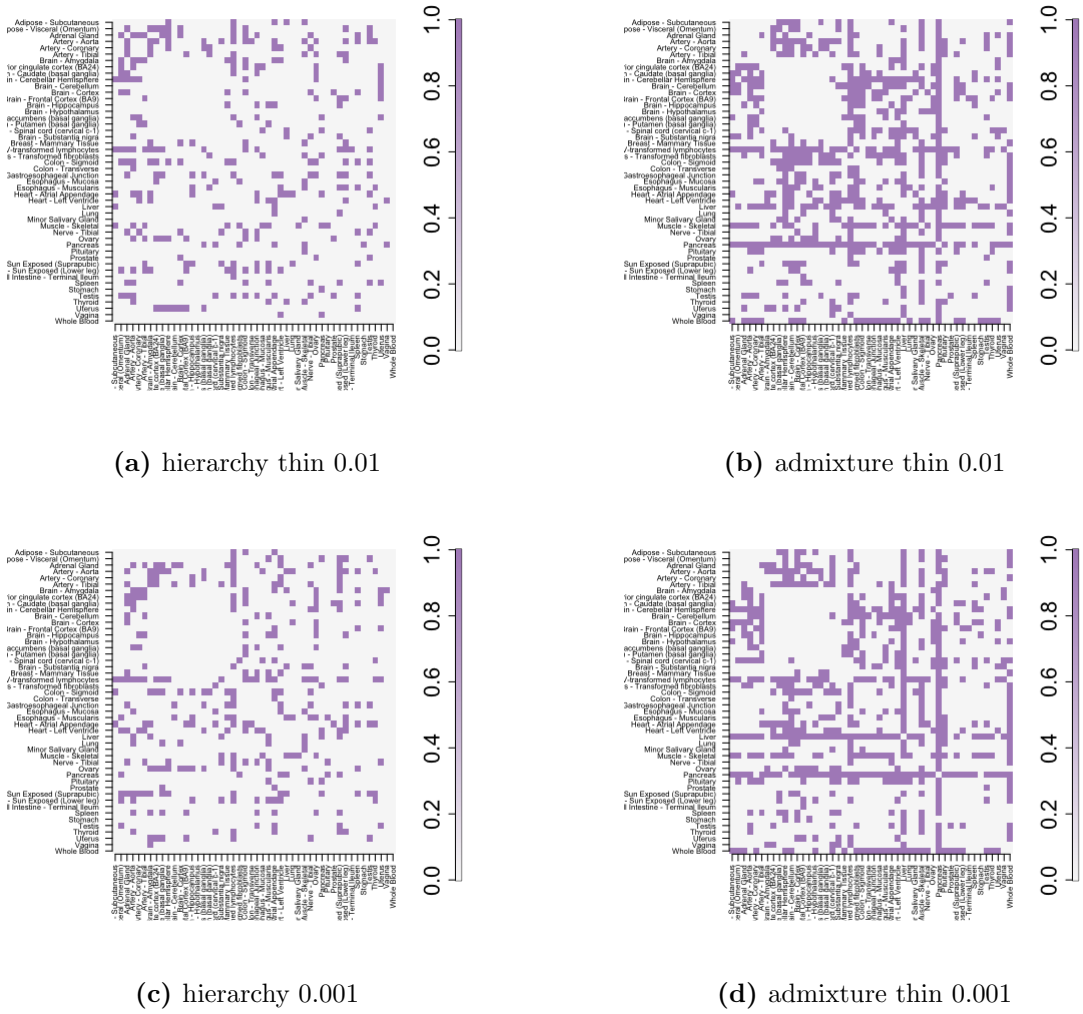


Figure 4. Structure plot of all samples for $K = 6$ of Deng et al data [13], arranged by the preimplantation development phases of the cells and within each phase, arranged in the same order as in the data. Some developmental phases are represented by a single cluster, for example- *zygote/early2cell* while some developmental phases can be written as a mix of two clusters or more clusters.

4.1 Supplemental figures



(a) hierarchy thin 0.01

(b) admixture thin 0.01

(c) hierarchy 0.001

(d) admixture thin 0.001

Figure 5. In this graph, we compare the hierarchical clustering method with the admixture method for thinned data with thinning parameters being $p_{thin} = 0.001$ and $p_{thin} = 0.0001$. The color coding scheme is similar to **Fig 3**. Note that the performance of the admixture indeed deteriorates from **Fig 3** in separating out the clusters as is expected. But it still outperforms the hierarchical clustering.

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

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

4.2 Supplementary Table 1

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