1 Outline

The outline for this paper (in the format of George M. Whitesides)

Title:

Authors: Kushal Dey and Matthew Stephens

2 Introduction

- *objectives of the work*: to devise a completely unsupervised method to cluster the samples (tissue or single cell samples) into biologically meaningful sub-types based on the RNA-seq gene counts data
- *justification of objectives* :
  1. People have mainly used hierarchical clustering from GTEx consortium paper to most single cell RNA seq papers I have come across. We have evidence Admixture model does better than hierarchical clustering from a biological viewpoint (see structure.beats.hierarchical.html).
  2. Hierarchical clustering does not give us directly the genes that drive the clusters, Admixture model does, and it also provides us with a log likelihood to fix how many clusters to choose, based on Bayes factor.
  3. We can predict the admixture proportions of cell types in any new sample coming in, so we can easily cluster new samples in cancer biopsy where the sub-types may involve cancer or non-cancer samples.
- *Background*
  1. The BackSpin algorithm used by Zeisel et al. Claim is it does better than hierarchical but not model based (also not convincingly proven to be better)
  2. Use of downsampling and then modified hierarchical clustering scheme as applied by Jaitin et al.
  3. Mainly, people have used hierarchical clustering scheme
  4. Population genetics uses Admixture model on a regular basis. We think we can generalize that to RNA-seq data. The only question is do we really see the tissue samples as cell type admixture, as we observe individuals as population admixture. The answer seems to be yes.
- *Guidance to the reader*
  1. The Structure plot and t-SNE plots for GTEx tissues and for Zeisel data. Much better visualization than the regular heatmaps that we tend to see in RNA-seq papers.
  2. The Structure plot analysis for Brain samples that shows 80% one cluster in cerebellum tissue samples and then from gene annotations, it is revealed this cluster is indeed associated with synaptic activities implying it must be neuronal cell types.

This is pretty cool because we have a priori knowledge from cell type specific markers that around 80% of cells in cerebellum are neurons.

3. Also the strategy is similar to the topic model strategy in natural language processing and it is a really nice technique to use for RNA-seq datasets clustering.

3 Methods and Materials

3.1 Data preprocessing

RNA-seq experiments usually provide us with a set of FASTQ files that contain the nucleotide sequence of each read and a quality score at each position, which can be mapped to reference genome or exome or transcriptome. The output of this mapping is usually saved in a SAM/BAM file using SAMtools [2] . This task is primarily accomplished by *htseq-counts* by Sanders et al 2014 [1] or *featureCounts* [R package **Rsubread**] by Liao et al 2013 [3]. RNA-seq raw counts are the basis of all statistical workflows, be it exploration or differential expression analysis [**edgeR** [4], **limma** [5]]. There is a growing trend to make the analysis ready raw counts tables openly accessible for statistical analysis. ReCount is a online site that hosts RNA-seq gene counts datasets from 18 different studies [6] along with relevant metadata. We start with such gene count datasets and assume that we have samples (say N) along the rows and the genes (say G) along the columns. Before we apply our methods, we remove the genes with 0 count of matched reads across all samples, implying that these genes are probably not expressed in any sample and hence non-informative for the clustering or differential analysis of the samples. We also remove the samples or genes with NA values of reads, if any. Additionally we also remove any ERCC spike-in controls as they may create bias to the biological clustering patterns. For illustration, we have applied our method on a single-cell RNA seq data due to Zeisel et al (2015) [7] and GTEx Version 4 gene counts data [8]. The GTEx data is a tissue sample data and the reads are recorded for multitude of cells present in the tissue sample. This can lead to really large values of read counts, in particular for highly expressed genes. To reduce the model over-dispersion and to make the analysis comparable to single cell datasets, we applied a thinning mechanism to the GTEx data. If C_{ng} is the gene count for g th gene in tissue sample n, then we define the thinned counts as

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

where p_{thin} is the thinning probability. W chose p_{thin} to be of the order of the ratio of the total number of reads mapped to a single cell experiment (in this case Zeisel et al (2015) data for instance) and the total number of reads in the GTEx dataset, which turned out to be approximately 0.0001. To check for robustness of our clustering algorithm, we varied p_{thin} to be $0.01, 0.001, 0.0001$ (see Fig).

3.2 Methods Overview

We use a topic model approach due to Matt Taddy (package **maptpx**) to perform the clustering of the samples based on RNA-seq reads data [9]. We denote this matrix of counts by $C_{N\times G}$ where

N is the total number of samples (tissue/single cell) and G is the number of genes. We assume that the row vector of counts for each sample n across the genes is multinomially distributed.

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

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 idea here is that this read could be coming from some cell type for the tissue level expression study (or from some cell cycle phase for the single cell case study) and its probability of getting assigned to some gene g will depend on which cell type (cell cycle phase) it comes from. In general, we may assume that the read is coming from one of the several (say K) underlying classes/groups, which are not observed. Denote the probability that the sample is coming from the k th subgroup by q_{nk} ($q_{nk} \geq 0$ and $\sum_{k=1}^{K} q_{nk} = 1$ for each n) and the probability of a read coming from the kth subgroup, to be matched to the gth gene, 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 in all $N \times (K-1) + K \times (G-1)$ many unconstrained parameters, which is much smaller than the NG many counts data we have. Usually $K << min\{N, G\}$ and for RNA-seq datasets, N is usually in the region of 100s to 1000s and G range from $20,000$ to $50,000$. To estimate the model, a Maximum a posteriori (MAP) based approach is used. It assumes the priors

$$q_{n*} \sim Dir(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K})$$
$$\theta_{k*} \sim Dir(\frac{1}{KG}, \frac{1}{KG}, \cdots, \frac{1}{KG})$$

For better estimation stability, the usual parameters of the model are converted to natural exponential family parameters to which one can apply the EM algorithm (see Taddy 2012 [9]). The value of the Bayes factor for the model with K clusters compared to the model with 1 cluster, is recorded for each K, and the optimal K is chosen by running the clustering method for different choices of K and then choosing the one with maximum Bayes factor. The two main outputs from this method are the $Q_{N\times K}$ topic proportion matrix and $F_{K\times G}$ relative gene expression for each cluster.

3.3 Post processing analysis

For each n, q_{nk}'s which will give an idea about the relative abundance of individual subgroups (cell functional groups or cell types) represented in the sample (single cell or tissue respectively). If two samples n and $n^{'}$ are very close, say both coming from the same tissue for the tissue level data, then we expect q_{n*} and $q_{n^{'}*}$ to be very close too. A nice way to visualize the amount of relatedness among the samples is through the Structure plot due to Pritchard Lab, which

is a popular tool to visualize the admixture patterns in population genetics based on SNP/ microsatellite data [10] [11]. The Structure plot assigns a color to each of the subgroups and then presents a vertical barplot for each individual, which is fragmented by the subgroup proportions and colored accordingly. If the colored patterns of two bars are similar, then the two samples must be closely related. The other visualizing tool we use is t-distributed Stochastic Neighbor Embedding (t-SNE) due to Laurens van der Maaten, 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 [12] [13].

The other question of 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 each gene and then look at the KL divergence matrix of one cluster/subgroup k relative to other cluster/subgroup $k^{'}$, which we call $KL^{g}_{K\times K}$. This matrix is symmetric and has all diagonal elements 0 as the divergence of a cluster with respect to itself is 0. Next we define the divergence measure for gene g as

$$Div(g) = \underset{k}{max}\ \underset{l\neq k}{min}\ KL^{g}[k,l]$$

$$K_{div}(g) = arg\ \underset{k}{max}\ \underset{l\neq k}{min}\ KL^{g}[k,l]$$

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 $Div(g)$ and put the gene in the $K_{div}(g)$ th cluster/subgroup. Then we perform gene annotations for the top genes in each subgroup using **mygene** R Bioconductor package [20]. We then try to see if the significant genes in a particular subgroup/cluster are associated with some specific biological functionality. This would indicate if the subgroups are actually biologically relevant or not. For instance, for GTEx tissue sample data, if the clusters are indeed driven by cell types, then the top genes for these clusters will probably be associated with proteins related to functions for that particular cell type.

4 Results

The admixture model was applied on the V4 Genotype Tissue Expression (GTEx) project read counts data. We first used the **eQtlBma** software due to Flutre *et al* [23] to detect quantitative trait loci for gene expression (eQTLs) and then extracted all the cis gene -cis eQTL pairs from the data along with their posterior effect sizes for each tissue. We removed all gene-SNP pairs which had NA value reported in the posterior effect sizes for any of the tissues. Such NA values may result from ?? (small sample sizes, not all genotypes recorded?). At the end of the above filtering procedure, we were left with 16407 cis-genes. We extracted out these genes from the reads matrix and applied our admixture model on the read counts matrix for 8555 samples and these 16407 cis-genes. We present the Structure plot for admixture model fit with $K = 15$ in **Fig 1**. The Structure plot reflects the homogeneity among the samples coming from the same tissue and also gives us an idea about which tissues have similar patterns of gene expression. For example, the different Brain tissues seem to cluster together, the same being true for the

arteries (Artery-aorta, Artery-tibial and Artery-coronary). But interestingly, Muscle Skeletal and Heart tissues (Heart Left Ventricle and Heart Atrial Appendage) seem to be very close in their clustering patterns. The fact that in terms of gene expression patterns, Muscle Skeletal and Heart Left Ventricle are close is also evident from the fact that hierarchical clustering fails to separate out these two tissues **Fig 3**. Besides the Structure plot, we feel another nice approach of visualizing the clustering patterns is using the t-SNE (Supplemental Fig **??**) [12] [13]. Using the expression matrix of the genes for the clusters (the θ matrix), we obtained the set of top driving genes for each cluster. In **Table ??**, we present the gene names, the proteins they code and a short summary of their functions, obtained from the **mygene** package in R [20]. As can be seen from the table, *PRM2*(protamine2), *PRM1* (protamine1) and *PHF7* (PHD finger protein 7) are the top three genes that drive the cluster which separates out testis from the other tissues in the admixture cluster model in **Fig 1**. Similarly, *HBB* (hemoglobin, beta), *HBA2* (hemoglobin, alpha 2) and *HBA1* (hemoglobin, alpha 1) seem to be the top three genes that distinguish the whole blood and for a separate cluster from the rest.

A field of very active interest in recent times is to estimate the proportion of different cell types in different tissues. Marker based approaches are usually adopted to validate for different cell types and get a sense of the abundance of different cell types in the tissue samples [**?**] [22]. The admixture model is a marker free method to obtain clusters driven by cell types. When applied on the full GTEX samples data, it is not evident because of the inter tissue variation is pretty strong. But when we apply the admixture model on just the Brain samples data, we see one cluster explaining around $80-85\%$ admixture proportion in Brain Cerebellum and Cerebellar hemisphere (see **Fig 2**) . This seems encouraging because it has been found using stereological approaches that rat cerebellum contains $> 80\%$ neurons (Herculano-Houzel and Lent 2005) [19]. We subsequently performed gene annotations for a few top genes driving the different clusters (Supplementary Table 1) and observed that the pivotal genes that separated out this particular cluster in brain cerebellum and cerebellar hemisphere from the rest had synaptic activities.

Besides the biologically novel fact that admixture model seems generates clusters driven by cell types, evidence seems to suggest this method purely as a clustering technique seems to outperform the hierarchical clustering which is the most commonly used approach of clustering in RNA-seq and single cell seq literature. In **Fig 3**, 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. For both the approaches, we used a color-coding scheme in case of complete separation of the two tissues in the pooled set of samples. We find that admixture model is more successful in separating out different tissues in general, compared to the hierarchical clustering technique. The admixture model is essentially a count based modeling approach and seems to handle low counts and zero counts much better than the hierarchical method which is a more general approach to clustering. Since the RNA-seq data and in particular single cell Seq data have lots of low counts and zero counts, the admixture model seems to be more suited for such data compared to hierarchical clustering method.

Currently there is a lot of interest in single cell sequencing as it is more informative about individual cell expression profiles compared to the RNA-seq on tissue samples. We were curious to see how stable the Admixture results are if the GTEx RNA-seq data is viewed at the scale of a single cell data. We achieve the latter by thinning the GTEx data under thinning parameter $p_{thin} = 0.0001$ which is the order of scale obtained by dividing the total library size of the Jaitin *et al* [**?**] with respect to the library size of the GTEx V4 read counts data. We fitted the admixture model for $K = 12$ on the thinned data and the Structure plot for the fitted model is presented in **Fig 4**. It seems that most of the features observed in **Fig 1** seem to be retained, for instance- the Brain samples clustering together, Whole blood and Testis forming separate clusters, Muscle skeletal and Heart tissue samples showing very similar patterns etc. However, thinning indeed shrinks the small differences across tissues and makes it more difficult to distinguish between tissues, as evident from the comparative study of hierarchical and admixture models, analogous to **Fig 3**, for thinned data with thinning parameters $p_{thin} = 0.001$ and $p_{thin} = 0.0001$ in **Fig 6**. One can see that with thinning, the performance of admixture model in separating the tissues deteriorates but encouragingly, it seems that admixture does outperform the hierarchical clustering even under thinned data.

Finally, we applied the admixture model on a couple of single cell datasets due to Jaitin *et al* [**?**] and Zeisel *et al* [7]. Jaitin *et al* sequenced around 4000 single cells from mouse spleen, where the cells were a heterogeneous mix enriched for expression of CD11c marker. The goal of their study was to separate out the B cells, NK cells, pDCs and monocytes. The sequencing was carried out in different amplification and sequencing batches. However the biological effect in their study was completely confounded with the amplification and sequencing batch effects. We present the Structure plot corresponding to the admixture model fit for $K = 7$ for the Jaitin *et al* data with the samples arranged by their amplification batch **Fig 5** (top panel). Since the batch effects and biological effects are confounded, it is difficult to interpret whether the clusters are driven by biology or by technical effects. Zeisel *et al* analyzed the single cell data obtained from mouse cortex and hippocampus and obtained 47 molecularly distinct subclasses, comprising all known major cell types in the region. They also identified many marker genes informative about cell types, morphology and location. We fitted admixture model for $K = 10$ on their data and we arranged the Structure samples as per their subclass assignment **Fig 5** (bottom panel).The subclasses as depicted by them did seem to show pretty homogeneous patterns overall under Structure, but it was interesting that some of the samples in Oligo4 subclass seemed to show more heterogeneity - the proportion of red cluster seemed high for a few samples compared to others. Also the first few samples under Oligo6 seemed to show patterns similar to some of Oligo4 samples with lower red cluster proportion. These samples in Oligo6 were pretty different in pattern from the rest of Oligo6 samples which had no trace of red cluster. Since within each group, the samples are ordered in the same order as reported in the dataset, there is high likelihood, adjacent samples may be coming from same plate or may be sequenced in same lane etc, all of which can lead to similar patterns due to technical effects. The main highlight of **Fig 5** is that one must be careful about interpreting Admixture results or any clustering results, as there is a possibility of batch effects driving the clusters instead of true biological effects.

There has been a growing concern among biostatisticians today about how to deal with batch effects [17] [18].

5 Discussions

We suggest a model based clustering approach for RNA seq or scRNA seq data that takes as input the read counts matrix over the samples and genes and number of clusters to fit (K), and gives as output the admixture proportions matrix for all the samples and the relative expression profiles of the genes in each of the K clusters. It also provides us with a model log-likelihood that can be used to choose the optimal K to fit. However for genetic data, it is more recommended to observe the clustering patterns over a range of values of K to observe how the patterns change as we increase K. The clustering method is pretty fast as it uses EM algorithm along with quasi-Newton updates to speed up the iterations. The clustering proportions obtained as output can be viewed using a Structure plot or the t-SNE plot that give a much better visual representation of the clustering patterns than heatmap or PCA. As per model specifications, ideally the clusters should be driven by the cell types and we already have seen some evidence in support of that in **Fig 2** when the model was applied on brain samples. Besides the cluster proportions, the model also provides the user with the relative expression profile of all genes in each of these clusters, from which it is easy to figure out which genes have significantly high expression in one or more clusters compared to the other clusters or in other words, are informative in driving the clusters. The user can select these cluster driving genes and annotate them to get a better understanding of the biological significance of the clusters. Even purely as a clustering technique, our method outperforms hierarchical method in separating out the samples belonging to distinct classes (in case of the GTEx data, the different tissues). So, overall we feel our model has a number of advantages over the standard methods of clustering used in RNA-seq or scRNA-seq literature, like hierarchical clustering, in terms of cluster quality, biological validation, visualization and interpretation. The clustering along with the Structure plot representation based on the sample metadata is implemented in package **CountClust** available on Github (https://github.com/kkdey/CountClust) which is a wrapper package of **maptpx** due to Matt Taddy [9].

Future works

- It will be worthwhile to see if instead of finding out cluster driving genes, we can find out cluster driving gene pathways. which would have significance from a biomedical standpoint.
- Since many of the genes are not informative for the clustering, we may try to impose a variable selection preprocessing or incorporate that in our model suitably so that it will extract out only the genes that are informative about the clusters and will also speed up the model fitting.
- The admixture proportions may be useful for determining the mixing weights for the prior covariance matrices in the eQtlbma.

- We may have important metadata on the samples (for instance the individual from whom the sample came from) or on the genes (for instance the gene length, GO or KEGG annotations, GC content etc) which we have not incorporated in our clustering model so far. We may want to do so in future (Alex and Trevor are pretty interested in this because they want to incorporate the bird bodymass)

References

1. S Anders, T P Pyl, W Huber. *HTSeq : A Python framework to work with high-throughput sequencing data.* Bioinformatics, 2014, in print; online at doi:10.1093/bioinformatics/btu638

2. Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. *The Sequence alignment/map (SAM) format and SAMtools.* Bioinformatics, 2009, 25, 2078-9. [PMID: 19505943]

3. Liao Y, Smyth GK and Shi W. *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.* Nucleic Acids Research, 2013, 41, pp. e108.

4. Robinson MD, McCarthy DJ and Smyth GK. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010, 26, pp. -1.

5. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK. *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015, 43(7), pp. e47.

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

7. 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.

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

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

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

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

12. 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.

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

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

15. Law CW, Chen Y, Shi W, Smyth GK. *voom: precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biology. 2014;15(2):R29.

16. 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.

17. 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.

18. 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

19. 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.

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

21. 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.

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

Cluster	Gene names	Proteins	Summary
cluster 1,red (nerve, adrenal)	ENSG00000160882	cytochrome P450, family 11, subfamily B, polypeptide 1	catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids, mutations cause congenital adrenal hyperplasia due to to 11-beta-hydroxylase deficiency.
	ENSG00000148795	cytochrome P450, family 17, subfamily A, polypeptide 1	catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids, muttaions associated with associated with isolated steroid-17 alpha-hydroxylase deficiency, pseudo-hermaphroditism, and adrenal hyper-plasia
	ENSG00000158887	myelin protein zero	encodes a major structural protein of peripheral myelin, mutations re-lated to autosomal dominant form of Charcot-Marie-Tooth disease type 1 and other polyneuropathies.
cluster 2, blue (adipose and lung)	ENSG00000168878	surfactant protein B	an amphipathic surfactant protein es-sential for lung function and home-ostasis after birth, muttaions cause pulmonary alveolar proteinosis, fatal respiratory distress in the neonatal pe-riod.
	ENSG00000168484	surfactant protein C	hydrophobic surfactant protein essen-tial for lung function and homeostasis after birth, associated with pulmonary alveolar proteinosis, interstitial lung disease in older infants, children, and adults.
	ENSG00000185303	surfactant protein A2	encode pulmonary-surfactant associ-ated proteins, mutations associated with idiopathic pulmonary fibrosis.
cluster 3, shallow blue (colon and esophagus)	ENSG00000163017	actin, gamma 2, smooth muscle, enteric	involved in various types of cell motil-ity and maintenance of the cytoskele-ton, constituent of the contractile ap-paratus and muscle tissues.
	ENSG00000133392	myosin, heavy chain 11, smooth muscle	functions as a major contractile protein, chromosomal rearrangement is associated with acute myeloid leukemia of the M4Eo subtype.
	ENSG00000107796	actin, alpha 2, smooth mus-cle, aorta	play a role in cell motility, structure and integrity, associated with aortic aneurysm familial thoracic type 6.

Cluster	Gene names	Proteins	Summary
cluster 4, black (brain)	ENSG00000259384	growth hormone 1	is expressed in the pituitary,member of the somatotropin/prolactin family of hormones, controls growth, mutations lead to short stature
	ENSG00000132639	synaptosomal-associated protein	involved in the regulation of neurotransmitter release
	ENSG00000115138	proopiomelanocortin	encodes a polypeptide hormone precursor, synthesized mainly in corticotroph cells of the anterior pituitar, hypothalamus, placenta, and epithelium, important for energy homeostasis, melanocyte stimulation, and immune modulation,associated with early onset obesity, adrenal insufficiency, and red hair pigmentation.
cluster 5,light blue (artery)	ENSG00000133392	myosin, heavy chain 11, smooth muscle	major contractile protein, converting chemical energy into mechanical energy through the hydrolysis of ATP
	ENSG00000143248	regulator of G-protein signaling 5	RGS proteins are signal transduction molecules involved in regulation of heterotrimeric G proteins by acting as GTPase activators.
	ENSG00000111341	matrix Gla protein	likely acts as an inhibitor of bone formation, defects causes Keutel syndrome.
cluster 6,deep blue (muscle heart)	ENSG00000143632	actin, alpha 1, skeletal muscle	produces highly conserved proteins that play a role in cell motility, structure and integrity, mutations cause nemaline myopathy type 3, congenital myopathy, diseases leading to muscle fibre defects
	ENSG00000104879	creatine kinase, muscle	protein encoded is cytoplasmic enzyme involved in energy homeostasis and serum marker for myocardial infarction.
	ENSG00000198125	myoglobin	encodes a member of the globin superfamily and is expressed in skeletal and cardiac muscles.
cluster 7,dark brown (brain)	ENSG00000197971	myelin basic protein	major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system
	ENSG00000131095	glial fibrillary acidic protein	encodes one of the major intermediate filament proteins of mature astrocytes, mutations casuses Alexander disease.
	ENSG00000180354	maturin, neural progenitor differentiation regulator homolog (Xenopus)	NA

Cluster	Gene names	Proteins	Summary
cluster 8, shallow yellow (skin stomach)	ENSG00000186395	keratin 10, type I	encodes a member of the type I (acidic) cytokeratin family, mutations associated with epidermolytic hyperkeratosis.
	ENSG00000096088	progastricsin	The protein is a digestive enzyme produced in the stomach, major component of gastric mucosa, associated with gastric cancer, Helicobacter pylori related gastritis.
	ENSG00000182333	lipase, gastric	encodes gastric lipase, responsible for fat digestion and digestion of triglycerides.
cluster 9, yellow (cell EBV)	ENSG00000211896	immunoglobulin heavy constant gamma 1 (G1m marker)	NA
	ENSG00000211893	immunoglobulin heavy constant gamma 2 (G2m marker)	NA
	ENSG00000019582	CD74 molecule, major histocompatibility complex, class II invariant chain	serves as cell surface receptor for the cytokine macrophage migration inhibitory factor (MIF)
cluster 10,grey (thyroid, small intestine)	ENSG00000042832	thyroglobulin	thyroglobulin produced predominantly in thyroid gland, synthesizes thyroxine and triiodothyronine, associated with Graves disease and Hashimotot thyroiditis.
	ENSG00000171195	mucin 7, secreted	encodes a small salivary mucin, aiding in speech, mastication, associated with ashthma
	ENSG00000115705	thyroid peroxidase	plays a central role in thyroid gland function, associated with congenital hypothyroidism, congenital goiter, IIA.
cluster 11,cyan cluster (cells fibroblasts)	ENSG00000115414	fibronectin 1	Fibronectin is involved in cell adhesion, embryogenesis, blood coagulation, host defense, and metastasis
	ENSG00000108821	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
	ENSG00000164692	collagen, type I, alpha 2	Same as above

Cluster	Gene names	Proteins	Summary
cluster 12,shallow green (Whole blood)	ENSG00000244734	hemoglobin, beta	mutant beta globin causes sickle cell anemia, absence of beta chain/ reduction in beta globin leads to thalassemia
	ENSG00000188536	hemoglobin, alpha 2	deletion of alpha genes may lead to alpha thalassemia
	ENSG00000206172	hemoglobin, alpha 1	deletion of alpha genes may lead to alpha thalassemia
cluster 13,light brown (esophagus mucosa)	ENSG00000171401	keratin 13, type I	keratins are intermediate filament proteins responsible for the structural integrity of epithelial cells
	ENSG00000163209	small proline-rich protein	NA
	ENSG00000143536	cornulin	play a role in the mucosal/epithelial immune response and epidermal differentiation
cluster 14,violet (liver pancreas)	ENSG00000204983	protease, serine 1	secreted by pancreas, associated with pancreatitis
	ENSG00000091704	carboxypeptidase A1	secreted by pancreas, linked to pancreatitis and pancreatic cancer
	ENSG00000169347	glycoprotein 2 (zymogen granule membrane)	secreted from intracellular zymogen granules and associates with the plasma membrane via GPI linkage
cluster 15, salmon (testis)	ENSG00000122304	protamine 2	Protamines are the major DNA-binding proteins in the nucleus of sperm
	ENSG00000175646	protamine 1	NA
	ENSG00000010318	PHD finger protein 7	This gene is expressed in the testis in Sertoli cells but not germ cells, regulates spermatogenesis.

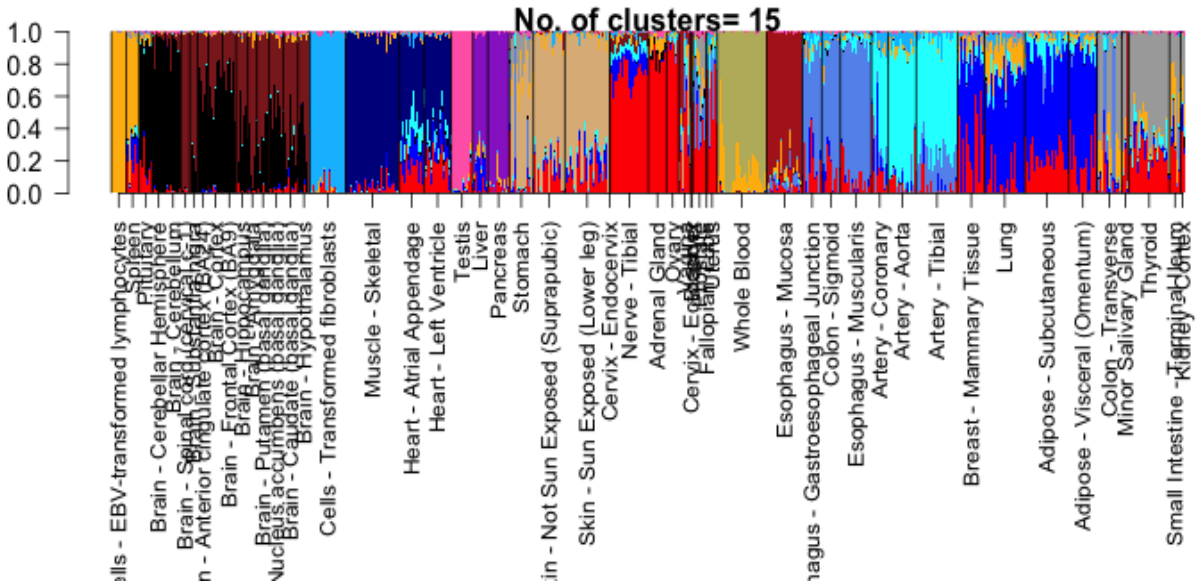


Figure 1. Structure plot of the admixture proportions (with 15 topics/clusters) for the 8555 tissue samples coming from 53 tissues in GTEX Version 4 data based on 16407 cis genes derived using eQtlbma (due to Flutre et al [23]). Note that there the samples coming from the same tissue are pretty homogeneous. Also tissues of same origin, for instance all the brain tissues, all the arteries seem to cluster together. There are other interesting patterns as well- for instance, Muscle Skeletal has similar clustering patterns as Heart tissues and Breast mammary tissue has similar patterns as Adipose tissues.

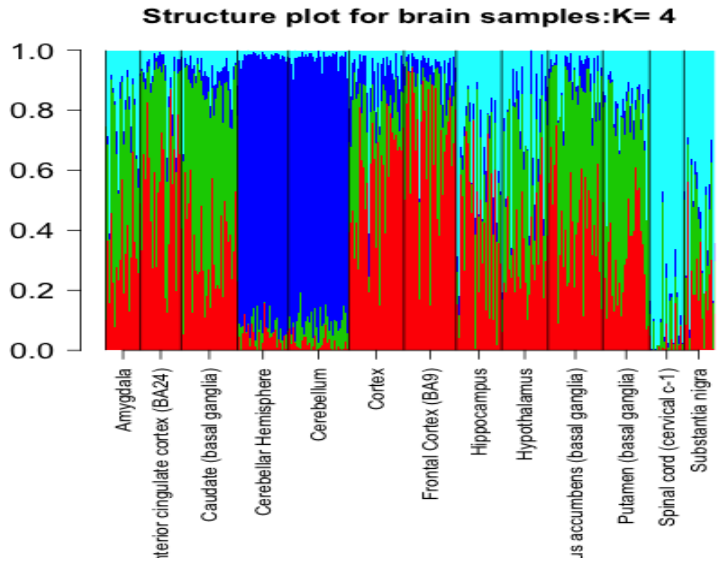
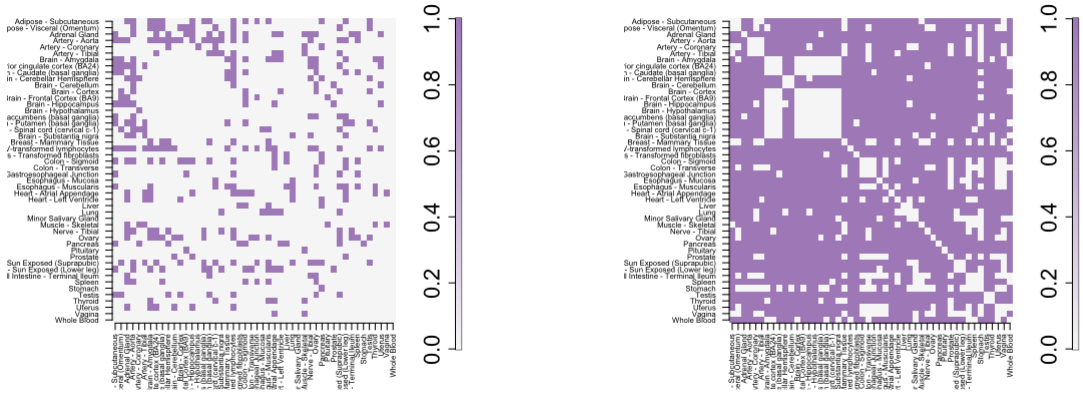


Figure 2. Structure plot of the admixture proportions (with 4 clusters) for the brain tissue samples drawn from GTEX Version 4 data. Quite clearly, brain cerebellum and cerebellar hemisphere seem to be dominated by the blue cluster while the Spinal cord and Substantia nigra by the cyan cluster. Prior marker based approaches have verified that $> 80\%$ of cells in brain cerebellum correspond to neurons [19]. So, the blue cluster seems to be driven by the neuron cell type. This fact is further attested by the gene annotations of the top genes driving the blue cluster (Supplementary Table 1).



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

Figure 3. 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.

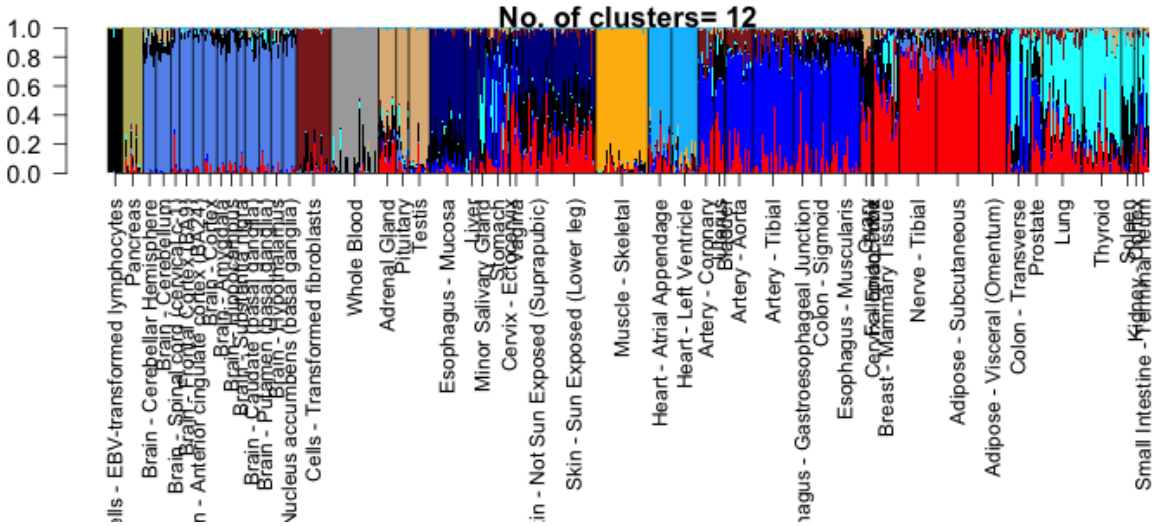
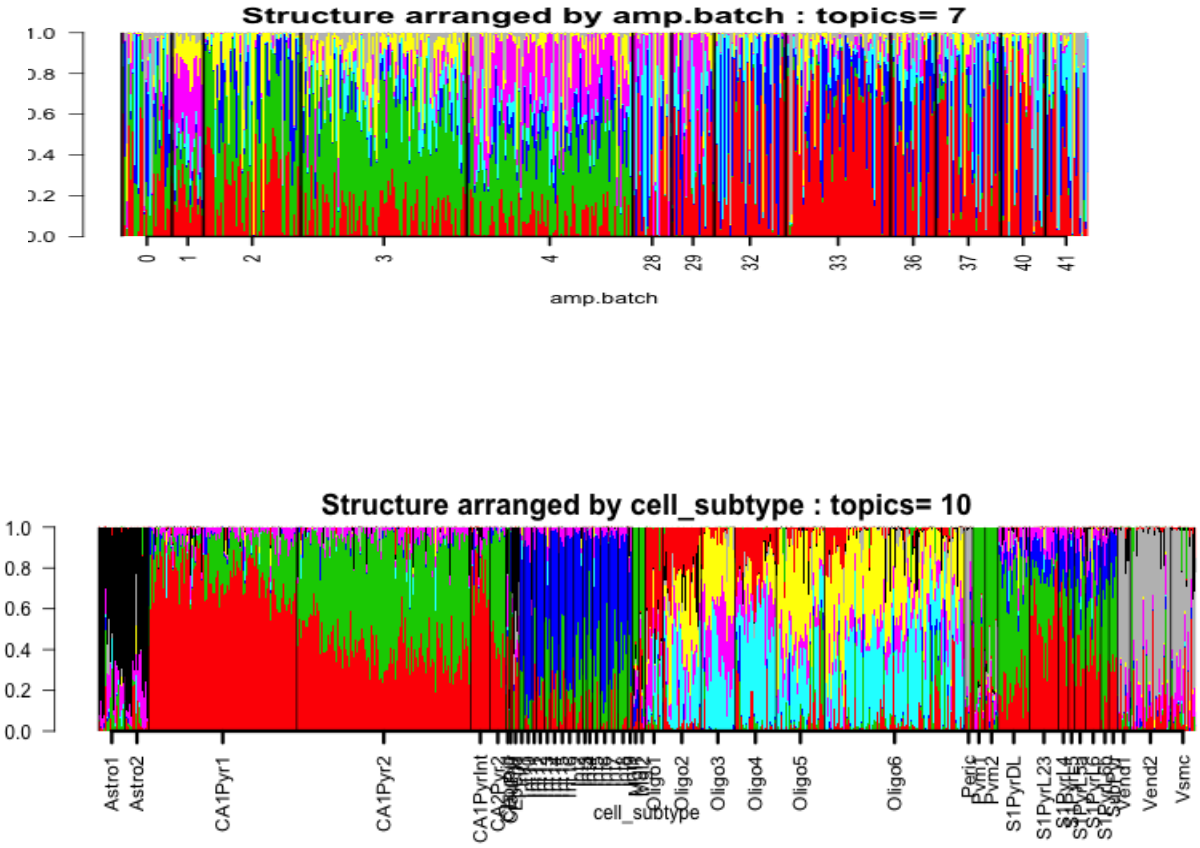


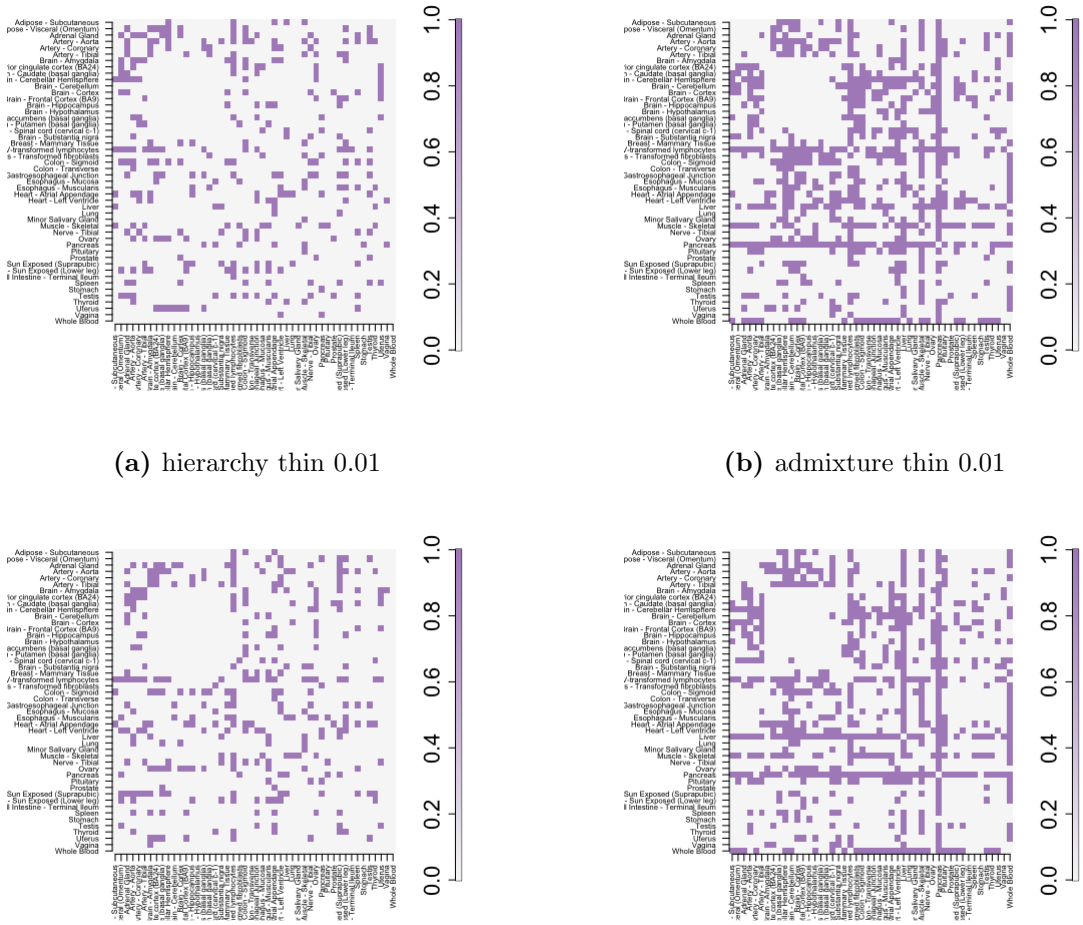
Figure 4. Structure plot of all tissue samples in GTEx version 4 data thinned data with $p_{thin} = 0.0001$ for K=12. The thinning parameter has been chosen so that the GTEx RNA-seq data can be interpreted at the same scale as a scRNA-seq data. It seems the results are pretty robust. This suggests that most of the structure in the clustering patterns we observed in textbfFig 1 is retained even after thinning.



(a) $k = 10$

Figure 5. (*top panel*) Structure plot of the1041 single cells for K=7 of the Jaitin *et al* data [16] arranged by the amplification batch. It is observed that the clustering patterns in each batch are pretty homogeneous 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. (*bottom panel*) Structure plot of all samples for $K = 10$ of Zeisel et al data [7], arranged by the cell subtype labels that were determined by the authors using their BackSpin algorithm and subsequent marker gene annotations. While the admixture patterns in cell subtypes are pretty homogeneous, the first few samples in Oligo6 show mild presence of red cluster and are pretty different from the rest of the samples in Oligo6 which do not show any trace of red cluster. These first few samples of Oligo6 look similar in pattern to some Oligo4 samples with mild red cluster presence. Oligo4 samples also shows some heterogeneity in terms of the proportion of red cluster present. This could either be due to misclassification of the Backspin algorithm, or some technical effects.

5.1 Supplemental figures



(a) hierarchy thin 0.01

(b) admixture thin 0.01

(c) hierarchy 0.001

(d) admixture thin 0.001

Figure 6. 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.