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 [14]. 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

An outline for results (under consideration)

- Form two separate subsections, one for the GTEx Version 4 data and the other for the single cell Zeisel data.
- For GTEx data, give a figure comprising of 4 Structure plots for different Ks, may be $2, 5, 10, 15$. Fix the thinning parameter p_{thin} to say 0.0001. Also record the log likelihoods (Bayes factors) for each of the 4 models, as reported by **maptpx**.
- Have one figure showing the robustness of the clustering method on the thinning parameter p_{thin}. Fix $k = 10$ and vary p_{thin} to be 0.0001, 0.001 and 0.01.

- One t-SNE plot for GTEx samples (with and without admixture in the same plot). Should this be in results or in discussions? Also the t-SNE probably would require an electronic supplemental file as I would need the **qtlcharts** highlighting for those plots.
- The GTEX brain samples Structure plot for $K = 4$ that shows the neuron cell types in brain cerebellum and cerebellar hemisphere. That is to show that the clusters are driven by cell types.
- Gene annotations for the GTEx significant genes (for brain) and also for the general set up (to decide on which K to fix). Use Bayes Factor?
- The Structure plot for Zeisel single cell data. again Multiple $K = 2, 5, 7, 10$.
- Gene annotations for the Zeisel single cell data. Need to choose the optimal K. Use Bayes Factor?
- t-SNE plot of the admixture proportions??..Is that required? Depends on how we present t-SNE. If this goes to discussion, we will avoid it here

5 Discussions

5.1 Normalization issue

A common practice in RNA-seq literature is to normalize the counts data by the library size before applying any differential analysis or clustering methods to it. Depending on sequencing machine/ lane use or change in sequencing depth, it may happen that some samples have very high counts across most genes, while some other samples may have very low counts of reads across all genes. This can lead to severe bias in statistical analysis if not accounted for. A way to counter this, given the raw counts, is to use the CPM (counts per million) normalized data [4] [15]. We define the CPM normalized data X_{ng} as

$$X_{ng} = \frac{c_{ng}}{\left[\frac{L_n}{10^6}\right]}$$

Here L_n is the library size or the total sum of the counts of all reads for the sample n. Though we apply our clustering algorithm on raw counts, we claim that CPM-normalization is intrinsic to the method. In mathematical terms, we are trying to model

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

A very naive estimate of p_{ng} would be

$$\hat{p}_{ng} = \frac{c_{ng}}{L_n} = \frac{X_{ng}}{10^6}$$

Therefore clustering with respect to p_{ng} ideally takes into account the variation at the X_{ng} level or normalized counts level instead of at c_{ng} level. From an information theoretic point of view, it seems that all information required for the clustering is contained in X_{ng}'s.

5.2 Admixture model vs Hierarchical model

Admixture model is a model based soft clustering method and hierarchical model is a non-model based tree clustering method and it is difficult to find a common measure that can effectively compare these two clustering schemes. We use real data to compare between the two methods. We drew 50 tissue samples from the pool of Muscle-Skeletal and Heart- Left Ventricle samples in GTEx Version 4 gene counts data with thinning parameter (p_{thin}) equal to 0.0001 and compared the heatmap of the admixture proportions q computed from the Admixture model with $k = 2$ clusters, with the hierarchical clustering heatmap on the counts data. It seems that admixture reduces the noise in the high dimensional data and does a better job at segregating the tissue samples corresponding to Muscle-Skeletal and Heart Left-Ventricle (Fig ??).

5.3 Batch effects

One of the important factors that may impact the clustering or differential analysis in tissue level RNA-seq or single cell RNA-seq analysis are technical or batch effects. These batch effects may stem from data coming from different laboratories or even for a single laboratory experiment, there may be effects due to the sequencing lane used , or the plate chosen for the experiment or the amplification process adopted. There has been a growing concern among biostatisticians today regarding how to deal with batch effects [17] [18]. For our clustering method as well, batch effects are an important concern. We present a case study where we used the Admixture model on a massively paralllel single cell RNA-seq data obtained from mouse spleen by Jaitin *et al* 2015, with the aim to replicate the clustering results reported in the paper [16]. However, following the experimental metadata provided by the authors, we observed that the data coming from the same sequencing or amplification batch seemed to have very similar patterns and it seemed that the biological effects may be confounded with the batch effects (Fig ??). To top that there was also a complete confounding between the amplification batch and then sequencing batch. So, one needs to be cautious about jumping to conclusions about clustering patterns just by looking at the Structure plot. We strongly suggest careful investigation of the experimental details to determine if there are any batch effects and also gene annotations to observe if the clustering method is indeed driven by genes that have biological functions relevant to the clustering patterns.

5.4 Performance analysis of the Admixture model

We carried out a few simulation studies to analyze the performance of the Admixture model under different scenarios. The main focus was to figure out how sensitive the method is to the admixture proportions or the relative gene expression differences. For instance, if we have data coming from 2 clusters, then under how well does our model do under different choices of the admixture proportions and the gene expression. We first consider a scenario where we have 1000 samples and 500 genes and the admixture proportion for sample n is of the form $(\frac{n}{1000}, \frac{1000-n}{1000})$ and the allele frequency vectors θ_1 and θ_2 for the two clusters are given by

$$\theta_1 = (0.01, 0.05, \frac{0.94}{498}, \frac{0.94}{498}, \cdots, \frac{0.94}{498})$$
$$\theta_2 = (0.05, 0.01, \frac{0.94}{498}, \frac{0.94}{498}, \cdots, \frac{0.94}{498})$$

The true admixture graph and the estimated admixture graph on the simulated counts table for 1000 samples and 500 genes is provided in Fig ?? (top panel). The top two significantly enriched genes for the clustering (as per Subsection 3.3) were found to be genes 1 and 2, which is clearly the case.

Now we consider a second scenario with the same set up as before but now the admixture proportion for n th sample is of the form $(0.4 + 0.2 \times \frac{n}{1000}, 0.6 - 0.2 \times \frac{n}{1000})$. This means that the variation in admixture proportions is less compared to the previous set up. The true admixture graph and the estimated admixture graph on the simulated counts table under this set up is presented in Fig ?? (bottom panel). The top significantly enriched genes for the clustering were found to be 483 and 224 which are way off. This shows that keeping the gene expression the same, admixture model is more successful in detecting clusters with large differences in admixture proportions between them.

Next we present a phase diagram analysis to show that the sensitivity of the admixture model depends both on how close the admixture proportions of the two subgroups or clusters are as well as how close the relative gene expression patterns for the two clusters are. For this analysis, we again assume that we have $K = 2$ clusters, we chose the number of samples to be 200 and we vary over the number of genes to be $50, 100$ and 200. We assume two phase parameters, α and γ. α is the phase parameter for the admixture proportions and we assume that for sample n, the admixture proportion is of the form $(\alpha, 1 - \alpha)$. On the other hand, for G genes, we assume that the phase parameter $\gamma < \frac{2}{G}$ and the relative gene expression θ_1 and θ_2 are of the form

$$\theta_1 = \left(\gamma, \frac{2}{G} - \gamma, \frac{1}{G}, \frac{1}{G}, \cdots, \frac{1}{G}\right)$$
$$\theta_2 = \left(\frac{2}{G} - \gamma, \gamma, \frac{1}{G}, \frac{1}{G}, \cdots, \frac{1}{G}\right)$$

We varied over α and γ, for each choice generated a counts table with 200 samples and G genes and then carried out admixture. We then observed whether the estimated admixture proportions match with the true proportions and whether we are detecting the significantly enriched genes for clustering, in this case genes 1 and 2. If Ω is the $N \times K$ true topic proportion matrix and $\hat{\Omega}$ be the estimated topic proportion matrix. We say *omega match* if

$$min\left(||\hat{\Omega} - \Omega||_2, ||1 - \hat{\Omega} - \Omega||_2\right) < 0.05$$

We say *theta match* if the top two genes detected by the clustering method to be driving the clusters (as per Subsection 3.3) are genes 1 and 2. We present the phase diagram to see for which values of α and γ, there is *omega match* or *theta match* or both (Fig ??).

Finally, it must also be emphasized that if there are more than two clusters in the data and we fit just 2 clusters, then the clusters obtained may not be driven by any of the real clusters but by a mix of the clusters present. We performed a simulation scenario where we have samples coming from a mix of 5 clusters with varying admixture proportions (true admixture proportion design shown in Fig ?? (a)). The relative gene expression profiles of 500 genes considered for the 5 clusters were assumed to be

$$\theta_1 = (0.01, 0.02, \frac{0.97}{498}, \frac{0.97}{498}, \cdots, \frac{0.97}{498})$$
$$\theta_2 = (\frac{0.98}{498}, \frac{0.98}{498}, \cdots, \frac{0.98}{498}, 0.01, 0.01)$$
$$\theta_3 = (\frac{0.97}{498}, \frac{0.97}{498}, \cdots, \frac{0.97}{498}, 0.01, 0.02)$$
$$\theta_4 = (0.01, 0.01, 0.1, 0.2, \frac{0.68}{498}, \frac{0.68}{498}, \cdots, \frac{0.68}{498})$$
$$\theta_5 = (0.1, 0.1, 0.1, 0.1, 0.1, \frac{0.5}{495}, \frac{0.5}{495}, \cdots, \frac{0.5}{495})$$

We drew 1000 samples from the above set up of admixture proportions and relative gene expression of the clusters for 500 genes and then applied the admixture model to the counts data. The true and the estimated Structure plot are presented in Fig ??.

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

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.

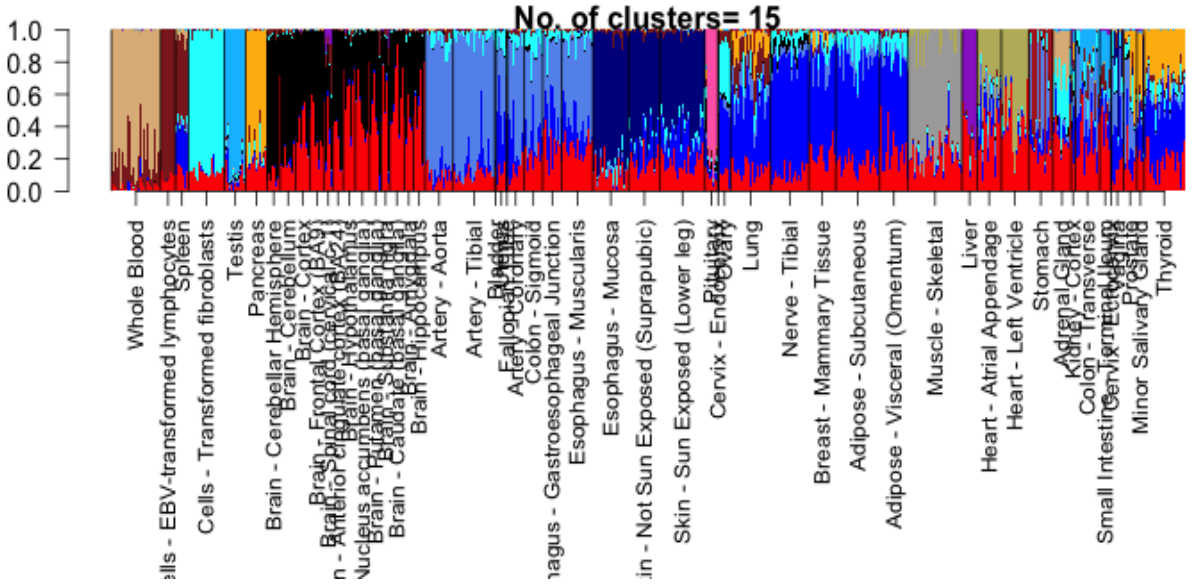


Figure 1. Structure plot of the admixture proportions (with 15 topics/clusters) for all the tissue samples in GTEX Version 4 data based on 5000 genes with the highest mean expression. Some of the tissues form very distinct clusters, for instance Whole Blood, Pancreas, Skin, Arteries etc while there is a lot of similarity in cluster patterns between Muscle Skeletal and Heart Left Ventricle, or among the different sub-tissues of the Brain.

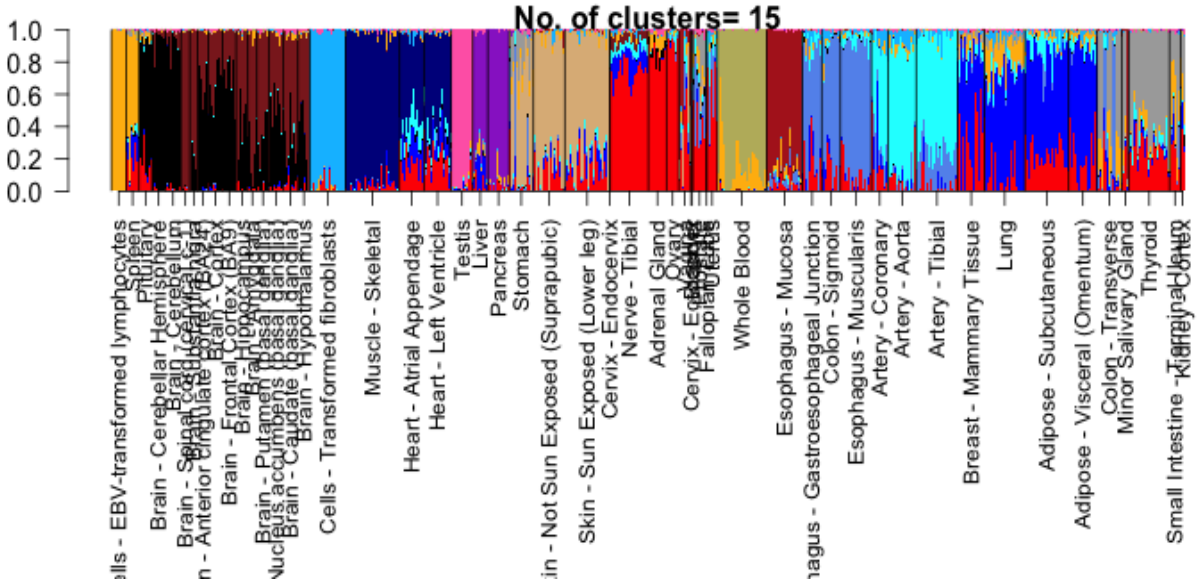


Figure 2. Structure plot of the admixture proportions (with 15 topics/clusters) for all the tissue samples in GTEX Version 4 data based on 16407 cis genes from the eQTL study conducted by the GTEx Consortium. Many of the patterns in this Structure plot are retained from the Structure plot based on the 5000 genes with highest mean expression.

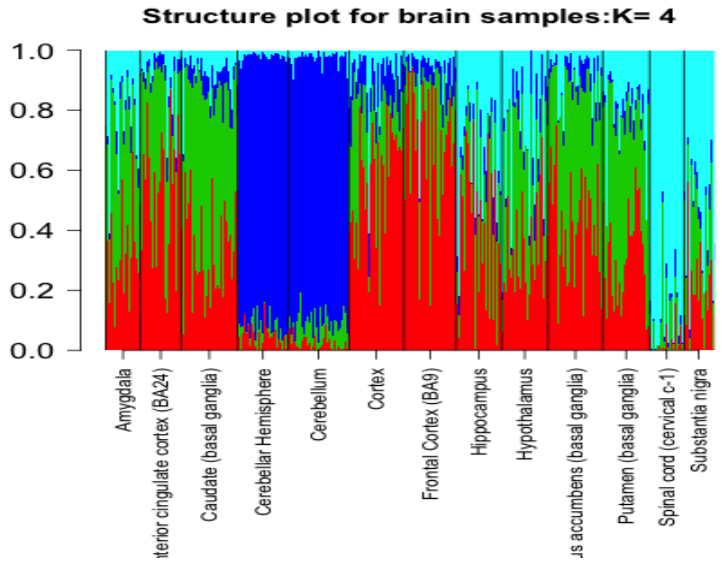
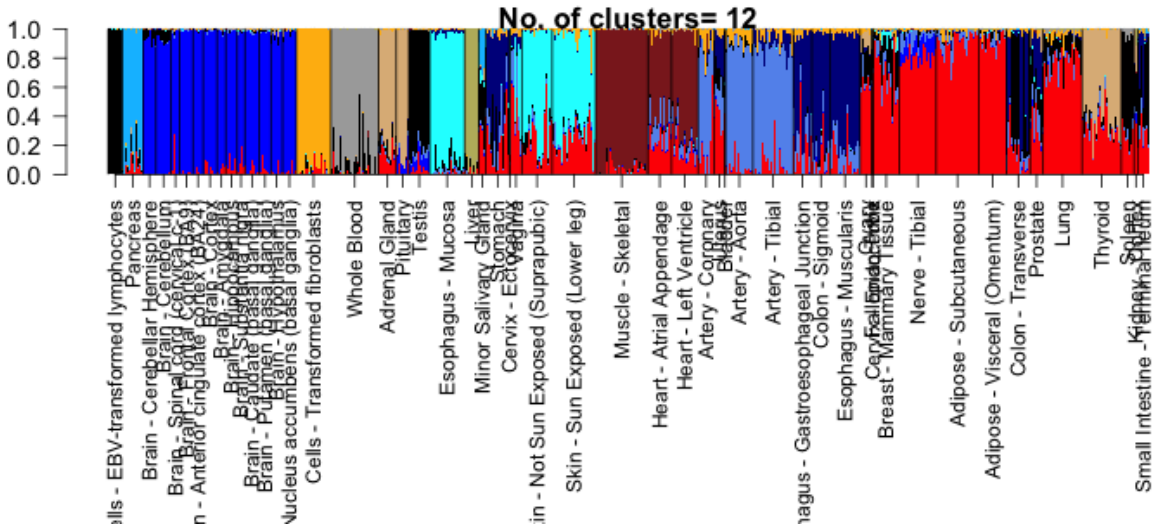
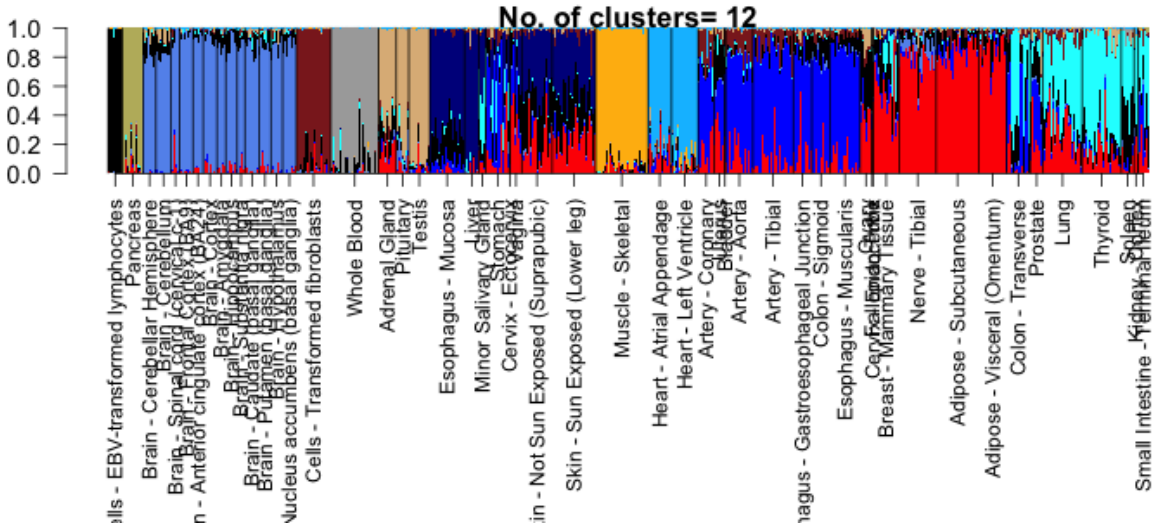


Figure 3. Structure plot of the admixture proportions (with 4 clusters) for the brain tissue samples drawn from GTEX Version 4 data. Quite clearly, brain cerebellum and cerebellar hemisphere seem to be dominated by the blue cluster while the Spinal cord and Substantia nigra by the cyan cluster. Prior marker based approaches have verified (?) that 80% of cells in brain cerebellum correspond to neurons. 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 (Subsection 3.3).

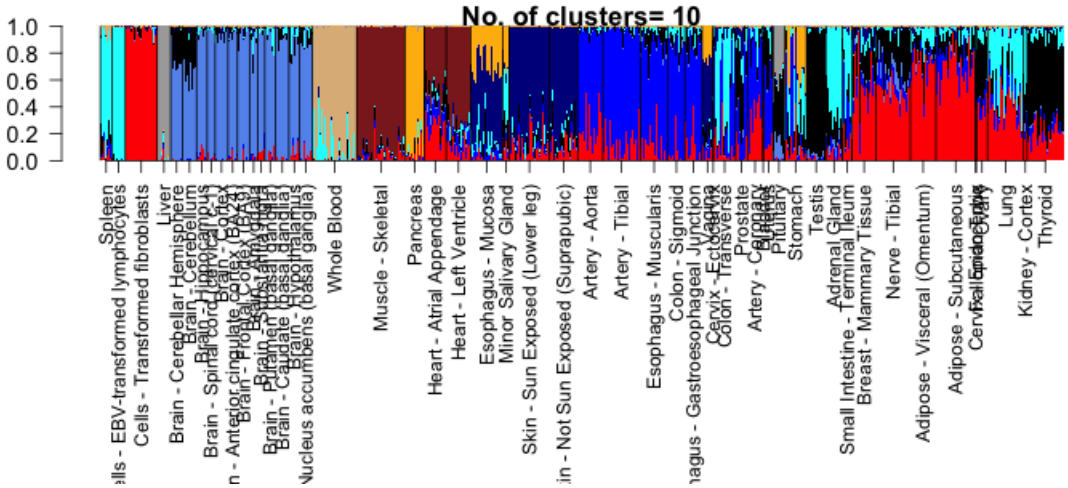


(a) $p_{thin} = 0.001$

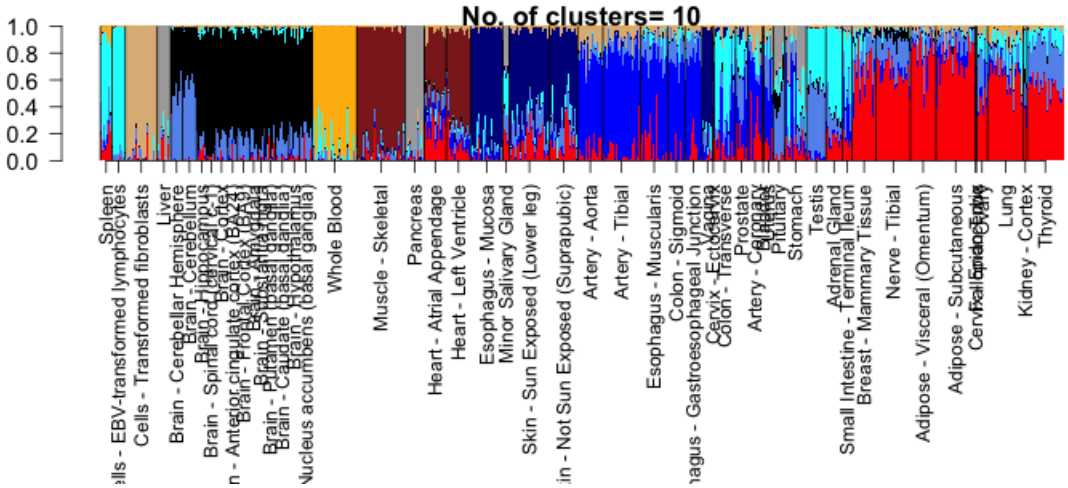


(b) $p_{thin} = 0.0001$

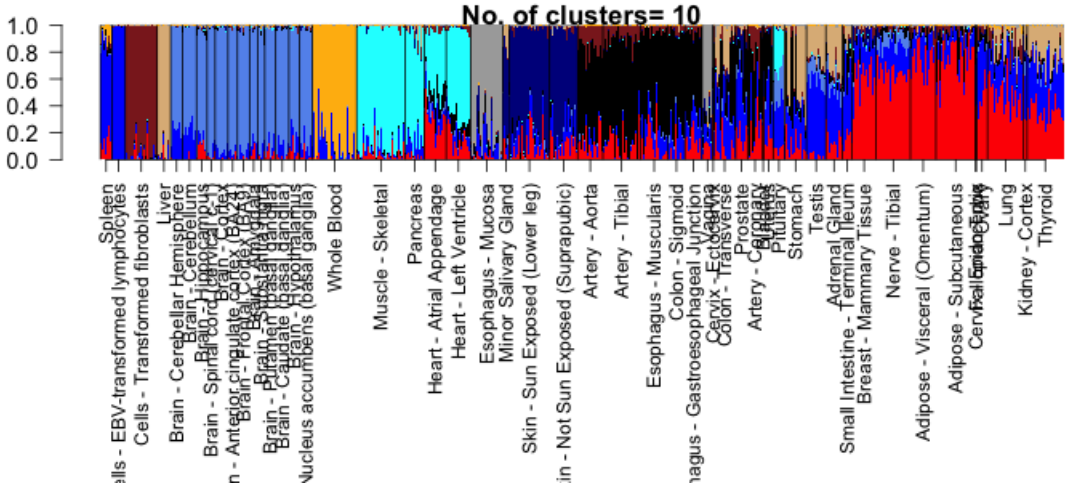
Figure 4. Structure plot of all tissue samples in GTEx version 4 data for K=12 under different values of the thinning parameter p_{thin}. The three values of the thinning parameter chosen are 0.01, 0.001 and 0.0001. It seems the results are pretty robust, though some of the cluster patterns seem to change. For instance, Muscle Skeletal and Heart tissue samples separate out at $p_{thin} = 0.0001$ but cluster together at $p_{thin} = 0.001$ for the same number of topics.



(a) Run 1

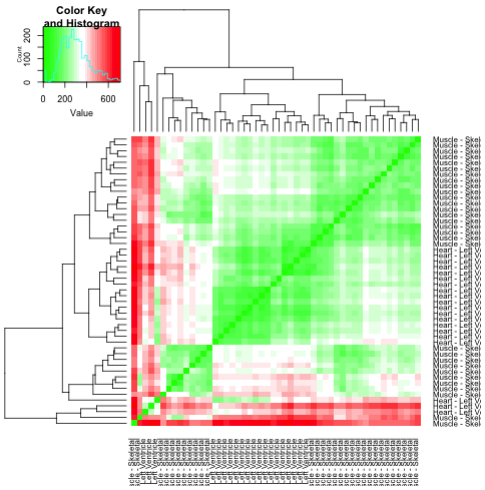


(b) Run 2

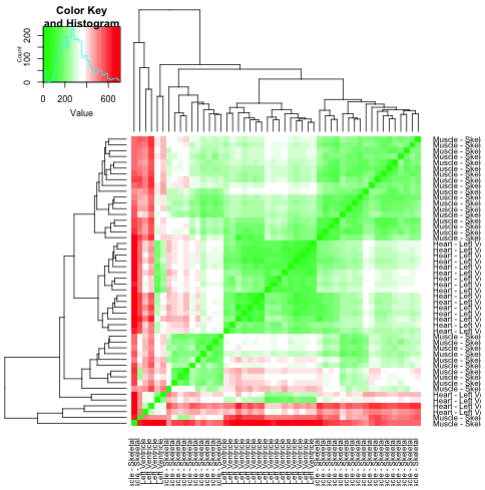


(c) Run 3

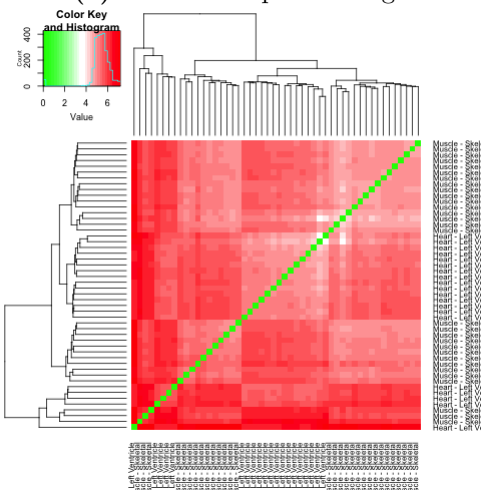
Figure 5. Structure plot of all tissue samples in 3 runs of the GTEx version 4 data for for K=10 for the thinning parameter $p_{thin} = 0.01$. For the 3 runs, the datasets are randomly generated from the actual counts data using $p_{thin} = 0.01$, so the datasets are different across the 3 runs. However, it seems the results are pretty robust.



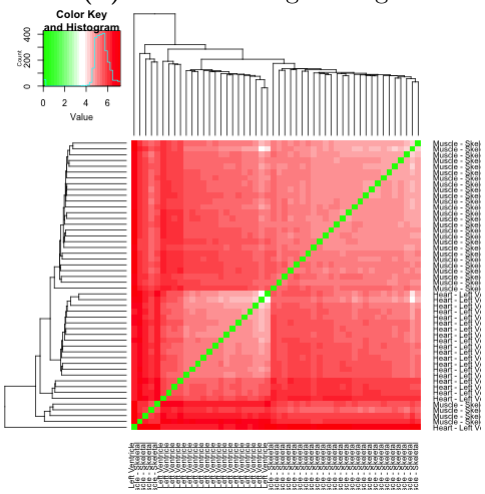
(a) counts- complete linkage



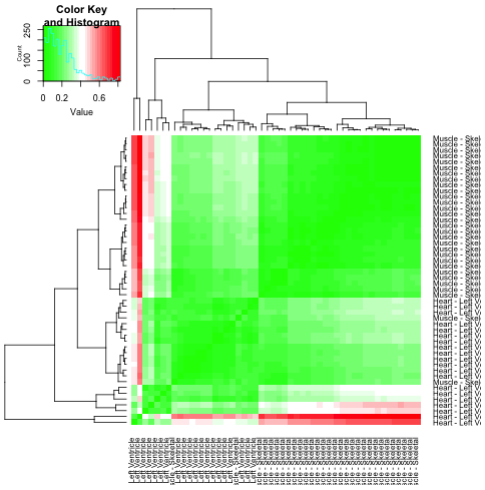
(b) counts- average linkage



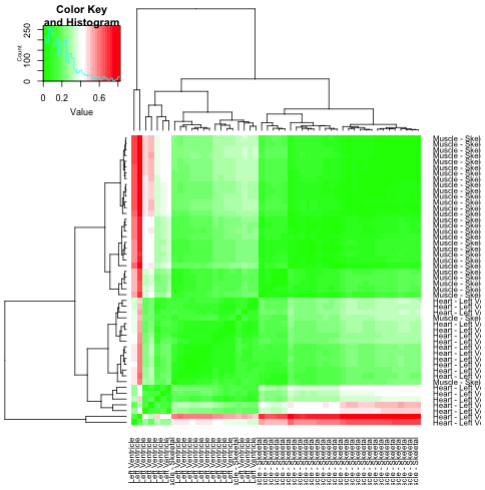
(c) cpm- complete linkage



(d) cpm- average linkage

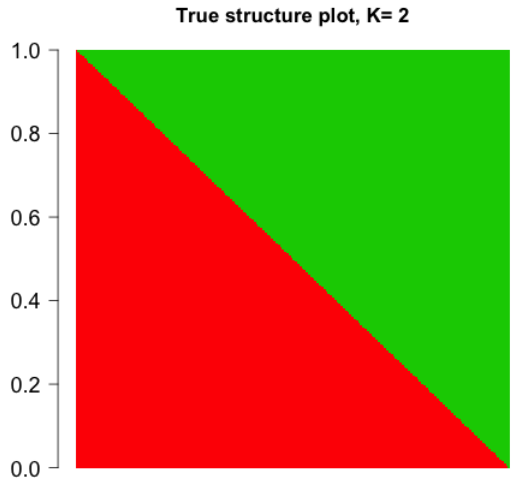


(e) admix prop- complete linkage

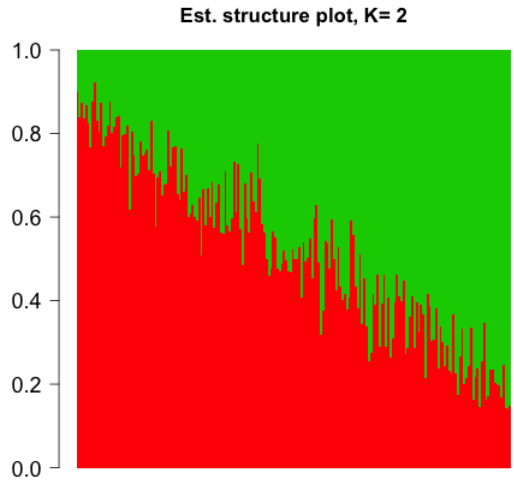


(f) admix prop- average linkage

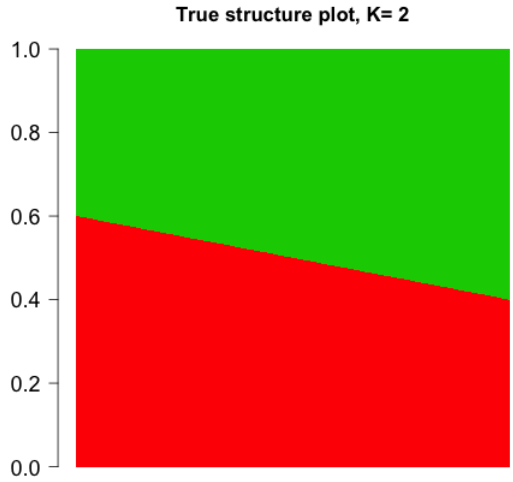
Figure 6. Comparison of heatmap of the counts data (*top panel*), the cpm normalized data (*middle panel*) and the admixture proportions data (*bottom panel*) on a randomly drawn 50 samples from the pool of Muscle-Skeletal and Heart-Left Ventricle samples in GTEx Version 4 RNA-seq counts data. The distance method used euclidean and the linkage used was average linkage. Color scale provided in the figure. It seems that for admixture model heatmap, all the Muscle-skeletal and Heart Left-Ventricle samples cluster separately, while for the hierarchical



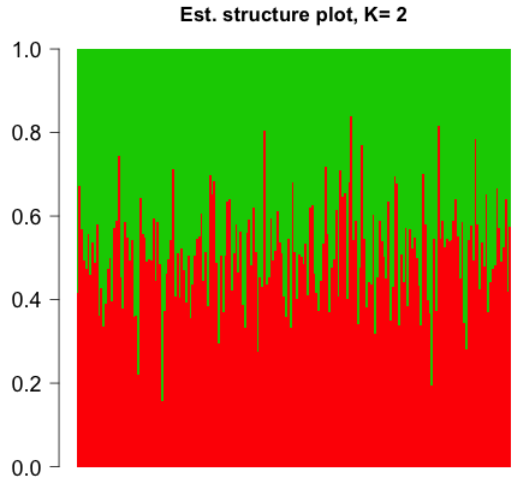
(a) Set up 1 (true structure)



(b) Set up 1 (est. structure)

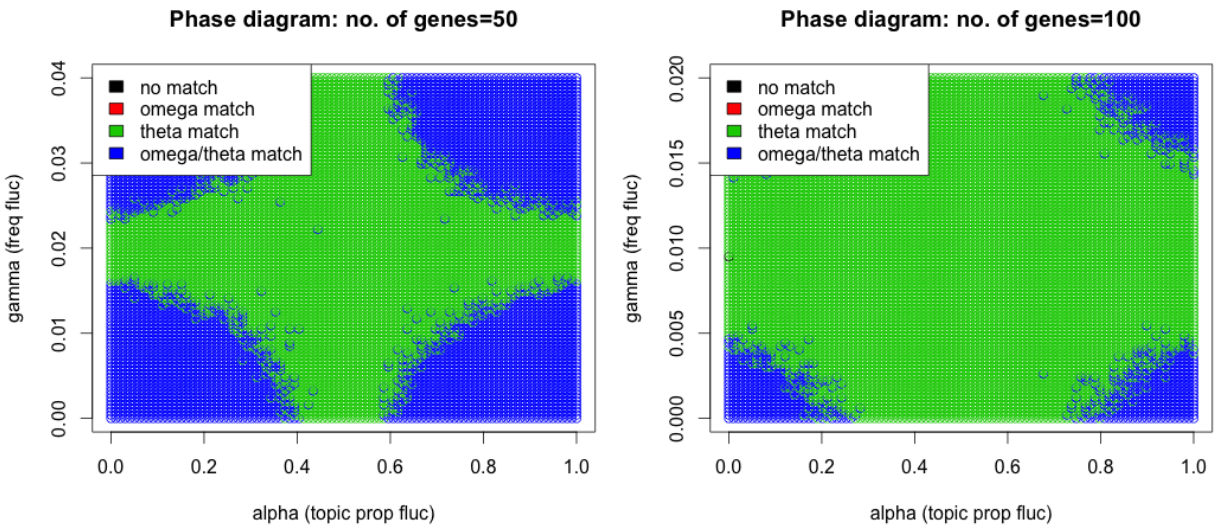


(c) Set up 2 (true structure)

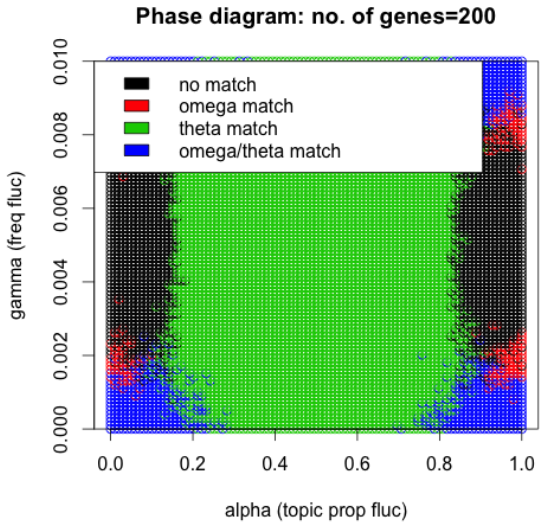


(d) Set up 2 (est. structure)

Figure 7. A case study to check for the performance of the admixture model under two set ups, one where the true admixture proportions vary a lot from 0 to 1 across samples for the two clusters and the other where the true admixture proportions vary mildly around 0.5 (from 0.4 to 0.6) for the two clusters. It is found that the admixture model is able to distinguish the clusters better in the first set up compared to the second. Even from gene annotations point of view, it is found that the admixture model is able to extract the truly significantly enriched genes for the clusters in the first set up but fails to do so in the second set up.

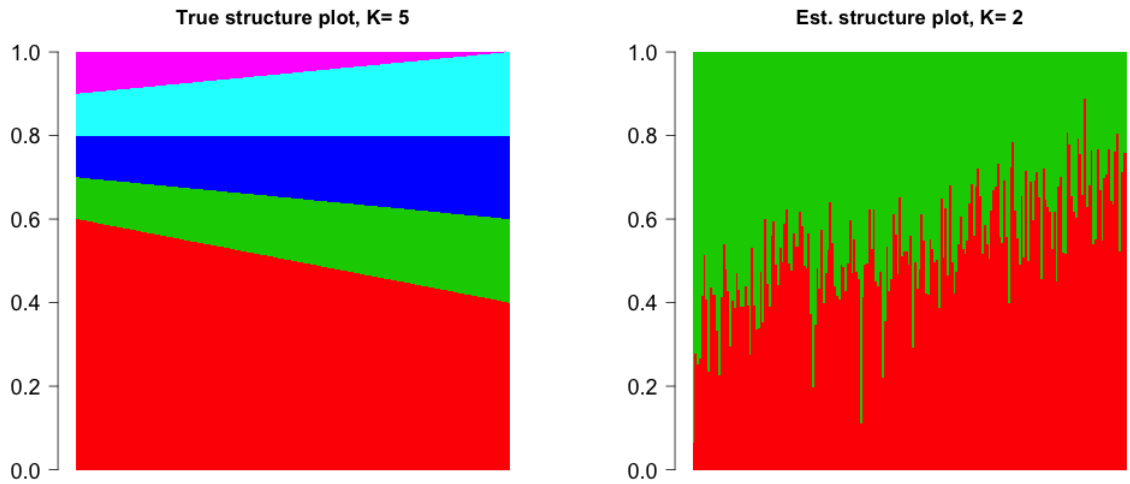


(a) G=50 **(b)** G=100



(c) G=200

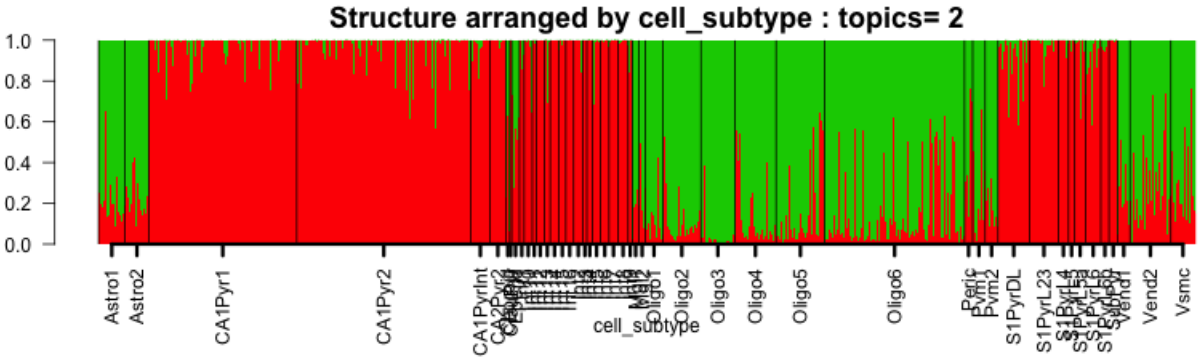
Figure 8. Phase diagram analysis for number of samples $N = 200$ and three choices of G, 50, 100 and 200. Note that as G increases, the gene expression signal and the distance between the relative gene expression between the two clusters decreases. As a result, the performance of the admixture model deteriorates. Also, for a broad part of the phase space, though the admixture model does not manage to determine the admixture proportions well enough (*omega match* does not hold) but it is able to extract the actually important genes driving the clusters (*theta match* holds).



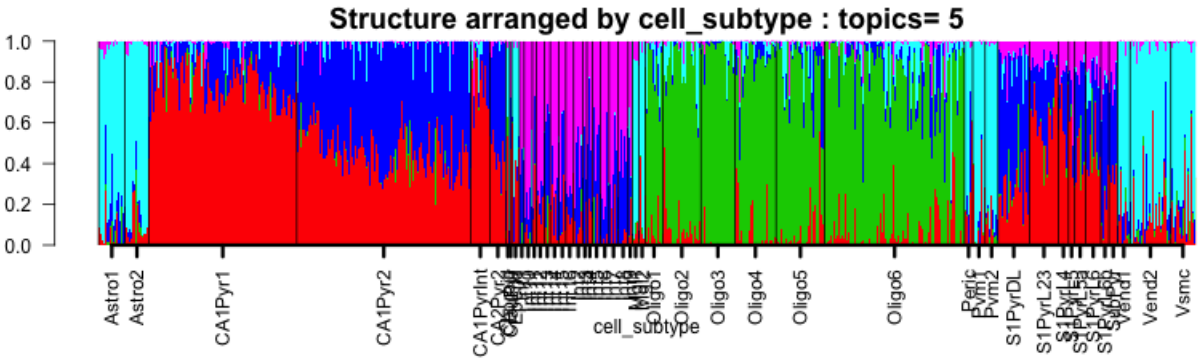
(a) Set up 3 (true structure)

(b) Set up 3 (est. structure)

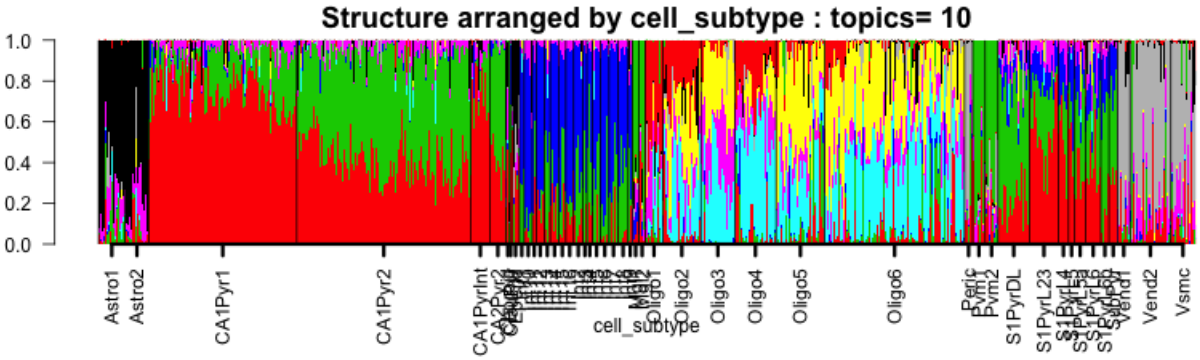
Figure 9. The true structure plot for a simulation set up with $N = 1000$ samples nd $G = 500$ genes with $K = 5$ clusters where the admixture proportions are given in (a). Counts data were generated under this true simulation set up. Then admixture model was fitted for $K = 2$ and the estimated Structure plot from the model fit is shown in (b).



(a) $k = 2$



(b) $k = 5$



(c) $k = 10$

Figure 10. Structure plot of all samples in 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. Here we present the Structure plots for number of topics $k = 2, 5, 10$.

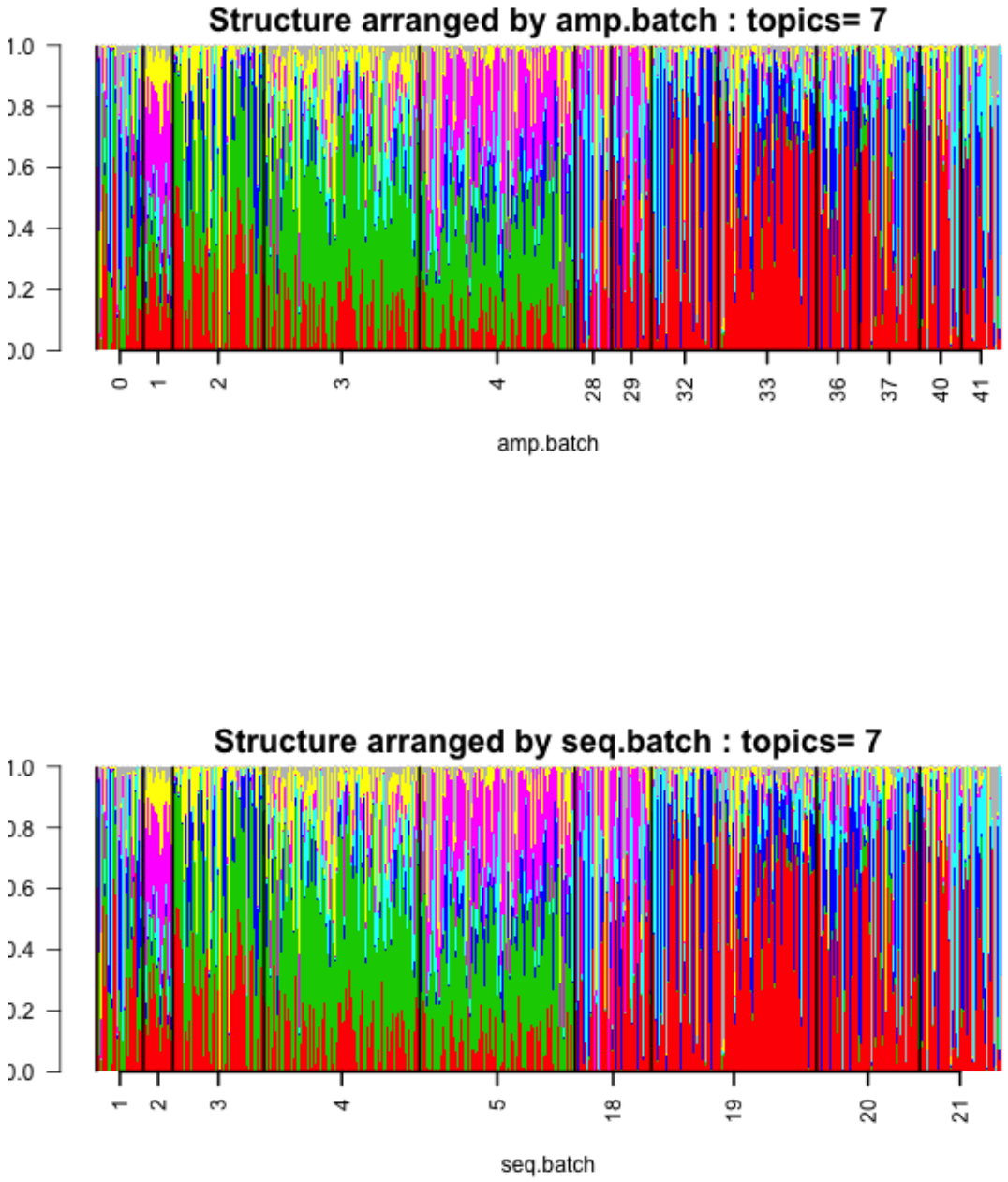


Figure 11. Structure plot of the 1041 single cells for K=7 used for clustering in Jaitin et al data [16] arranged by the amplification (*top panel*)and the sequencing batches (*bottom panel*). It was claimed that the 7 topics corresponded to 7 cell types. However there is a potential batch effect that may be confounded with the biological factors or may be driving the clusters altogether. Also there seems to be confounding between the sequencing batch and the amplification batch.