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 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], a task 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. Such gene counts datasets are the inputs for our clustering algorithm.

In the preprocessing step before applying our method, we remove the genes with 0 or same count of matched reads across all samples (non-informative genes), any sample or gene with NA values of reads and ERCC spike-in controls, as the latter may create bias due to their typical very high expression (number of reads mapped to them). For illustration, we applied our method GTEx Version 6 tissue level gene counts data [8] and on a couple of single cell data due to Zeisel *et al* [7] and Jaitin *et al* [16].

3.2 Model 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]. Let us denote the gene counts matrix as $C_{N \times G}$ 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 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 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 then 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 in all $N \times (K-1) + K \times (G-1)$ many unconstrained parameters, which is much smaller than the $N \times G$ data values of counts. Usually $K << min\{N, G\}$, N in the region of 100s to 1000s and G ranging from $10,000$ to $50,000$ (depending on the species the RNA-seq data is coming from and whether the set of genes recorded include non-protein genes or not). To estimate the model, a Maximum a posteriori (MAP) based approach is used (see Taddy 2012 [9]).

3.3 Visualization

For each n, q_{nk}'s which will give an idea about the relative abundance of individual subgroups (which may be driven by 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.

Another visualizing tool we recommend 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]. t-SNE provides some sense about which samples are closer to each other when the data is projected on 2D. But on the flipside, it 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).

3.4 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 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 observe 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

We begin by illustrating our method on the tissue level data from the GTEx project (V6 dbGaP accession phs000424.v6.p1, release date: Oct 19, 2015, `http://www.gtexportal.org/home/`) read counts data. RNA-seq data was obtained from 8555 samples collected from 450 donors across 51 tissues and 2 cell-lines. We collected a set of 16069 cis-genes that fulfilled some quality criteria (gene list available in `https://github.com/stephenslab/count-clustering/blob/master/utilities/gene_names_GTEX_V6.txt`).

Fig 1 presents the Structure plot for admixture model fit for $K = 15$. The Structure plot clearly highlights the similarity among the samples coming from the same tissue and also gives an idea about which tissues have similar patterns of gene expression. It is pretty nice that all the Brain tissues seem to cluster together and show very similar patterns, the same being true for the arteries (Artery-aorta, Artery-tibial and Artery-coronary). Interestingly, Nerve Tibial and Adipose tissues (Adipose Subcutaneous and Adipose Visceral (Omentum)) also seem to be very close in their clustering patterns.

As observed from the Structure plot, some tissues seem to be assigned to separate clusters for $K = 15$, (e.g: Pancreas, Whole Blood), but other tissues are represented as an "admixed" version of multiple clusters (e.g: Thyroid). But in these latter cases, the samples coming from the same tissue all seem to be cluster together as they have very similar patterns of "admixing" of different clusters. A different way of visualizing the results, which highlights the clustering and separation of the different tissues on a 2D projection space (see **Supplementary Fig 1 [url]**).

In order to get a sense of the biological interpretation of the clusters in **Fig 1**, we performed cluster annotation (see Methods and Materials). In **Tab ??**, we present the gene IDs, names and a short summary of their functions, obtained from the **mygene** package in R [20]. As seen from the table, *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 in the admixture 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. Overall it seems from **Tab ??** that cluster annotation is highlighting tissue specific functions and pathways.

Fig 1 Structure plot and the corresponding cluster annotation mainly highlighted inter tissue variation in the clusters, which is pretty logical because the cells in different tissues are pretty differentiated. We were curious if we can extract some cell type information if we focus on samples from a particular tissue or similar tissues. **Fig 2** shows the Structure plot for $K = 4$ on just the Brain samples data. Brain Cerebellum and Cerebellar hemisphere stand out in the plot as we see one cluster (blue) explaining around $80 - 85\%$ admixture proportion. Recent stereological approaches have shown that rat cerebellum contains $> 80\%$ neurons (Herculano-Houzel and Lent 2005) [19], much higher than other parts of the brain. We performed cluster annotation (Supplementary Table 1) and observed that the pivotal genes that separated out the blue cluster in brain cerebellum and cerebellar hemisphere were SNAP25 (synaptosomal-associated protein, 25kDa), ENO2 (enolase 2- gamma, neuronal) and CHGB (chromogranin B) all of which were associated with neuronal activities. It is definitely not true that the blue cluster represents the neuron cell type as it does not show up in other parts of the Brain, but it seems to be driven by neuronal cell type.

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 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. The same remains true for data sets thinned to simulate the level of lower coverage data that might be observed in single cell experiments (the threshold parameter $p_{thin} = 0.0001$ taken on comparing the GTEx data with Jaitin *et al* data [16]). Check **Fig 6** for demonstration.

For the thinned data, though the cluster quality is poor compared to the original data, it still seems that the clustering patterns remain preserved to a great extent. **Fig 4** presents the Structure plot for $K = 12$ for the GTEx thinned data with $p_{thin} = 0.0001$. Many of the features from **Fig 1** are restored even after thinning, for instance the Brain tissues clustering together, Heart samples and Muscle Skeletal samples showing similar patterns. This implies that Admixture as a clustering technique is pretty robust to the coverage of the data.

We applied the admixture model on a couple of single cell datasets due to Jaitin *et al* [16] 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 aim of their study was to separate out the B cells, NK cells, pDCs and monocytes. However the biological effect in their study was completely confounded with the amplification and sequencing batches. **Fig 5** (top panel) presents the Structure plot for $K = 7$ for the Jaitin *et al* data with the samples arranged by their amplification batch (which was a refinement of the sequencing batch).

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. **Fig 5** (bottom panel) presents the Structure plot for $K = 10$ on their data, where the samples in Structure plot are grouped by their subclass assignment and are in the same order as the data presented within each group. It was interesting that the first few samples under Oligo6 seemed to show some "admixing" due to red cluster,which was not observed in other Oligo6 samples. 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 a possibility that adjacent samples may be coming from same plate or may be sequenced in same lane etc, all of which can lead to similar patterns.

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 technical 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.

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

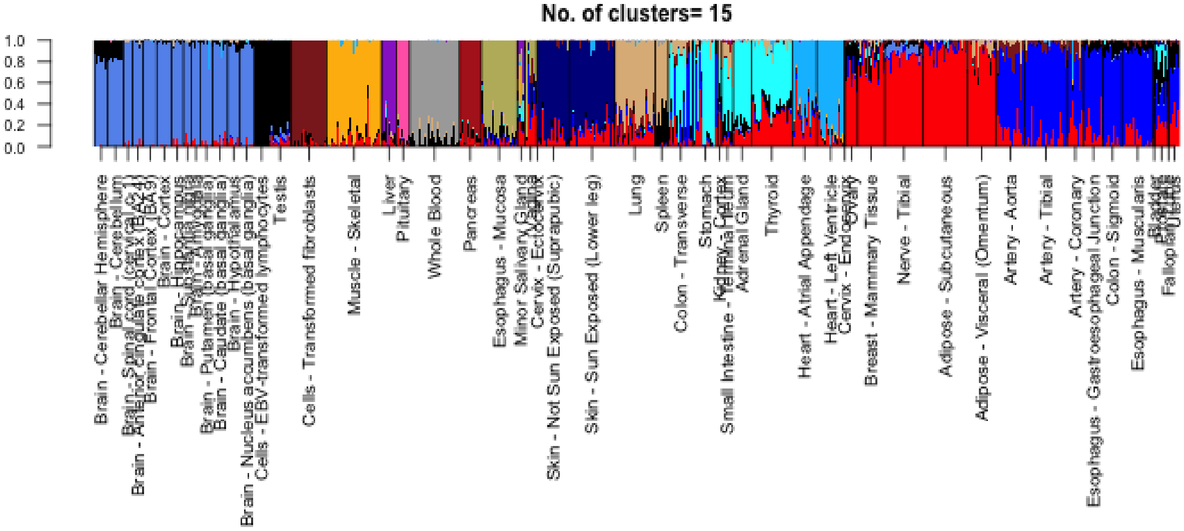


Figure 1. Structure plot of the admixture proportions (with 15 topics/clusters) for the 8555 tissue samples coming from 53 tissues in GTEX V6 data based on 16069 cis genes derived using PEER analysis MatrixEQTL . Note that the samples coming from the same tissue have similar admixing patterns. Tissues 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.

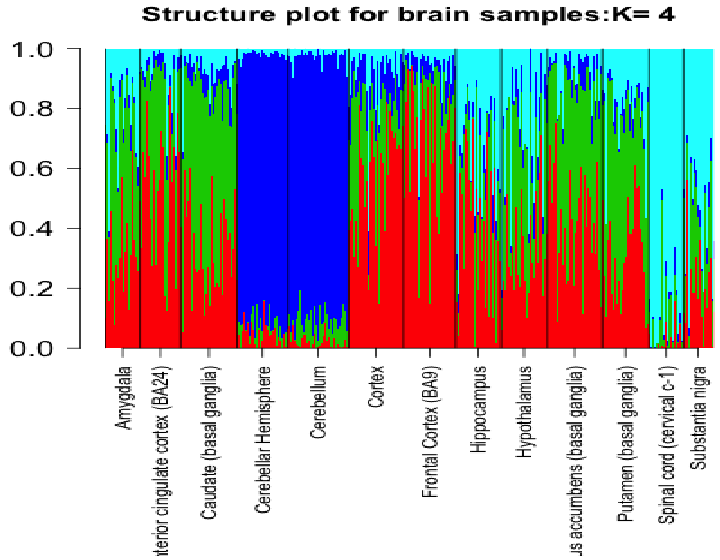
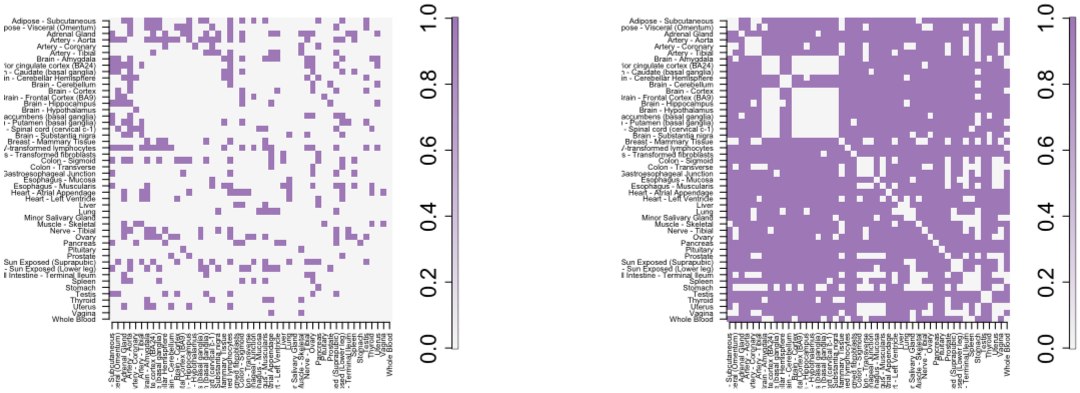


Figure 2. Structure plot of the admixture proportions (with 4 clusters) for the brain tissue samples drawn from GTEX Version 4 data. Quite clearly, brain cerebellum and cerebellar hemisphere seem to be dominated by the blue cluster while the Spinal cord and Substantia nigra by the cyan cluster. Prior 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.

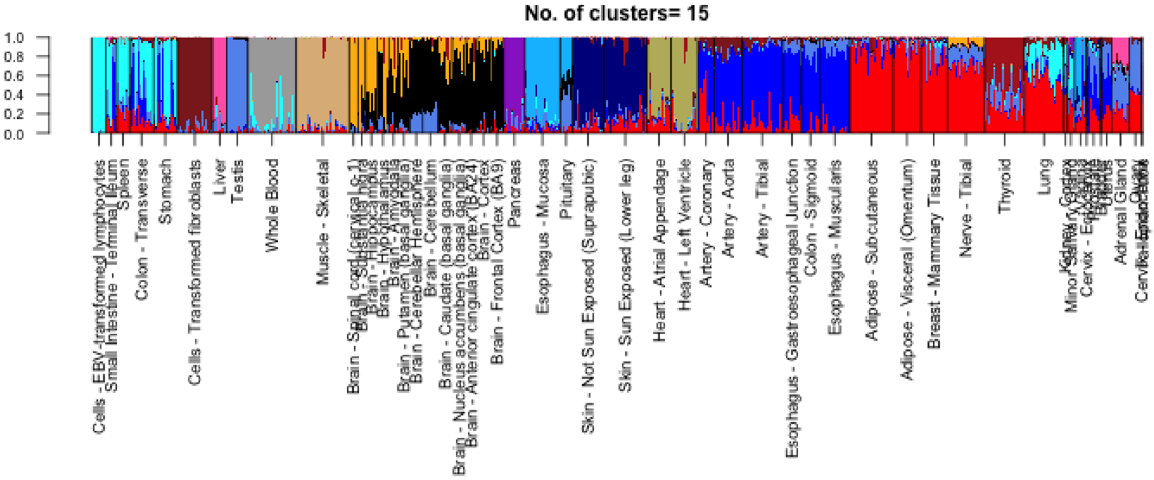
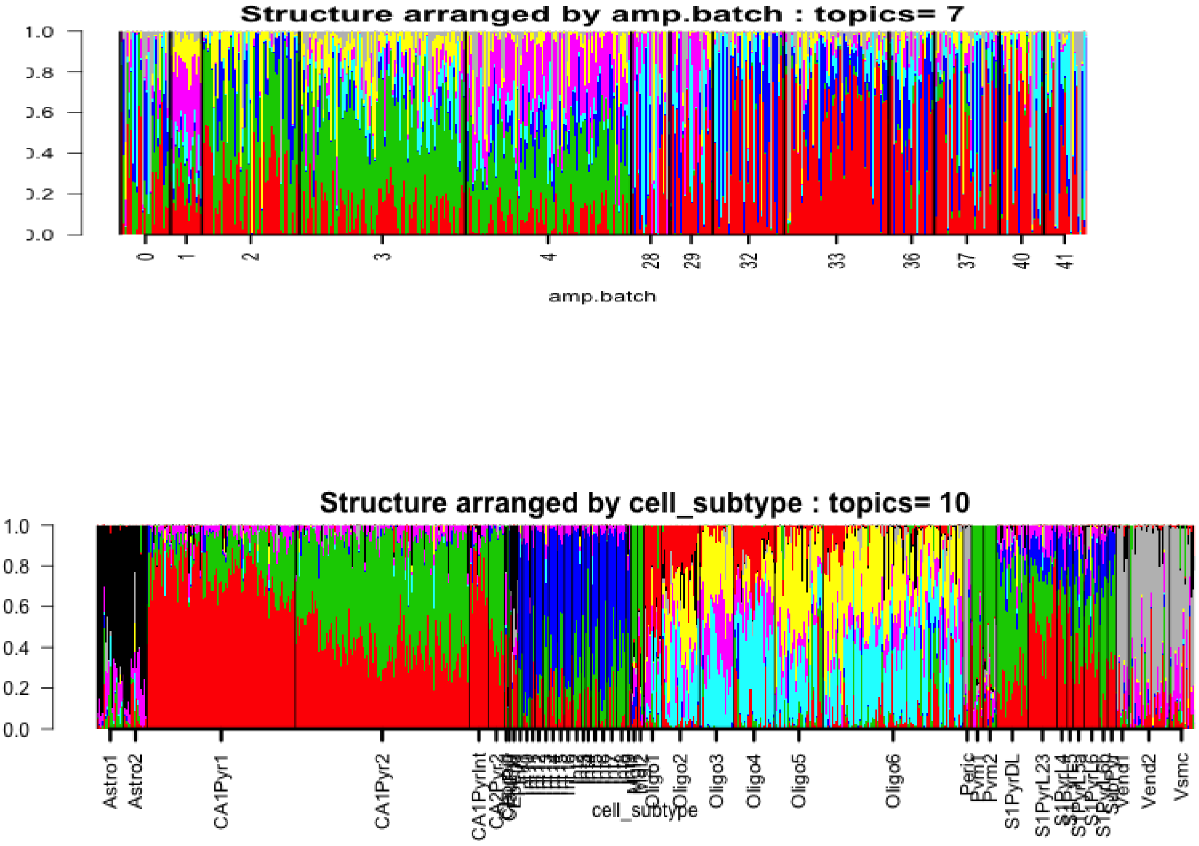


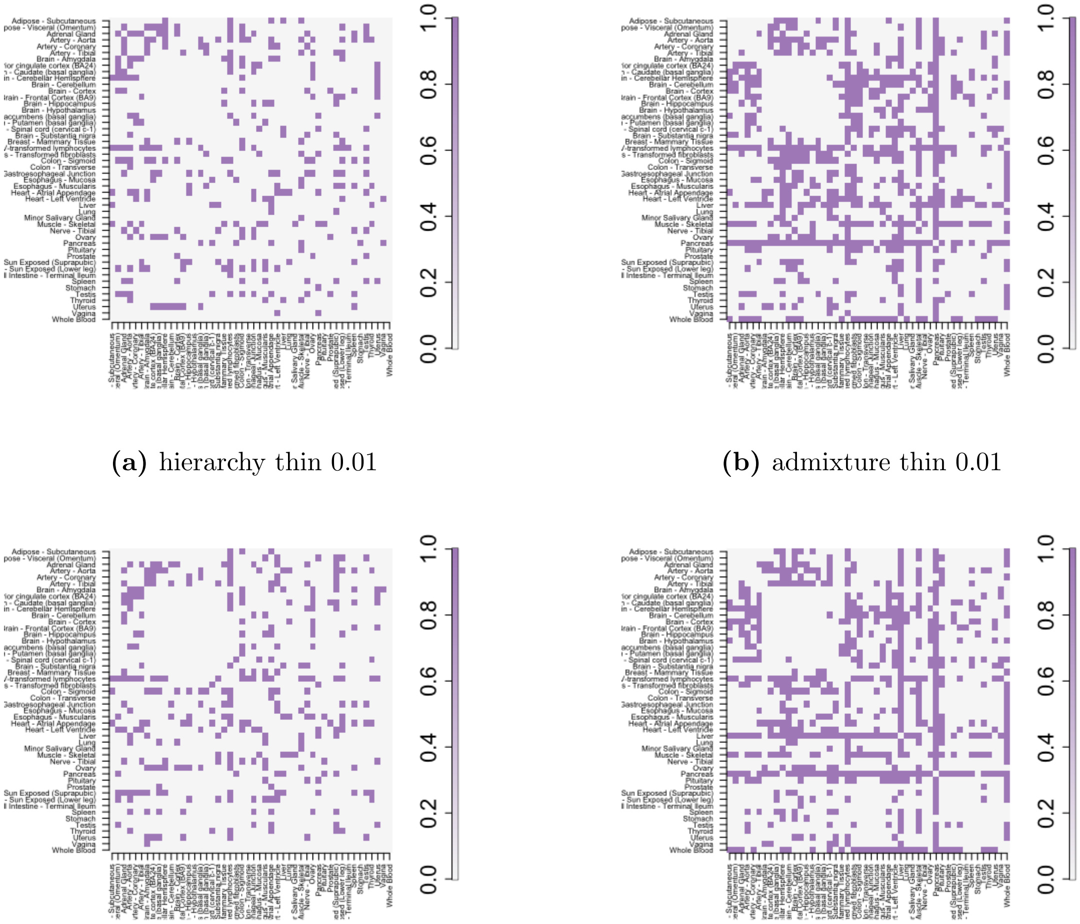
Figure 4. Structure plot of all tissue samples in GTEx V6 data thinned data with $p_{thin} = 0.0001$ for $K = 15$. The thinning parameter has been chosen so that the GTEx RNA-seq data can be interpreted at the same scale as a scRNA-seq data. The clustering patterns are more noisy compared to the non-thinned data in Fig 1, but overall, the similarity patterns across the tissues are retained. For instance, the brain tissues and the arteries still seem to be clustering together.



(a) $k = 10$

Figure 5. (*top panel*) Structure plot of the1041 single cells for K=7 of the Jaitin *et al* data [**?**] 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.

Cluster	Gene names	Proteins	Summary
cluster red (Nerve, Adipose)	ENSG00000170323	fatty acid binding protein 4, adipocyte	FABP4 encodes the fatty acid binding protein found in adipocytes, roles include fatty acid uptake, transport, and metabolism
	ENSG00000189058	apolipoprotein D	encodes a component of high density lipoprotein that has no marked similarity to other apolipoprotein sequences, closely associated with lipoprotein metabolism.
	ENSG00000166819	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)	ENSG00000133392	myosin, heavy chain 11, smooth muscle	functions as a major contractile protein, converting chemical energy into mechanical energy through the hydrolysis of ATP.
	ENSG00000107796	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.
	ENSG00000163017	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)	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.
	ENSG00000132639	synaptosomal-associated protein, 25kDa	this gene product is a presynaptic plasma membrane protein involved in the regulation of neurotransmitter release.

Cluster	Gene names	Proteins	Summary
cluster black (Testis)	ENSG00000122304	protamine 2	Protamines are the major DNA-binding proteins in the nucleus of sperm
	ENSG00000175646	protamine 1	Protamines are the major DNA-binding proteins in the nucleus of sperm
	ENSG00000010318	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	ENSG00000042832	thyroglobulin	thyroglobulin produced predominantly in thyroid gland, synthesizes thyroxine and triiodothyronine, associated with Graves disease and Hashimotot thyroiditis.
	ENSG00000182333	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.
	ENSG00000096088	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)	ENSG00000186395	keratin 10, type I	encodes a member of the type I (acidic) cytokeratin family, mutations associated with epidermolytic hyperkeratosis.
	ENSG00000167768	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.
	ENSG00000172867	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)	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	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)	ENSG00000168878	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.
	ENSG00000185303	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.
	ENSG00000122852	surfactant protein A1	encodes a lung surfactant protein that is a member of a subfamily of C-type lectins called collectins, associated with idiopathic pulmonary fibrosis.
cluster yellow (Muscle skeletal)	ENSG00000109061	myosin, heavy chain 1, skeletal muscle, adult	a major contractile protein which converts chemical energy into mechanical energy through the hydrolysis of ATP.
	ENSG00000183091	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.
	ENSG00000125414	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)	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 cyan (Heart)	ENSG00000175206	natriuretic peptide A	protein encoded by this gene belongs to the natriuretic peptide family, associated with atrial fibrillation familial type 6.
	ENSG00000197616	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.
	ENSG00000159251	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	Gene names	Proteins	Summary
cluster shallow green (Esophagus mucosa)	ENSG00000171401	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.
	ENSG00000170477	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.
	ENSG00000143536	cornulin	may play a role in the mucosal/epithelial immune response and epidermal differentiation.
cluster light brown (Pancreas)	ENSG00000204983	protease, serine 1	secreted by pancreas, associated with pancreatitis
	ENSG00000091704	carboxypeptidase A1	secreted by pancreas, linked to pancreatitis and pancreatic cancer
	ENSG00000175535	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)	ENSG00000171195	mucin 7, secreted	encodes a small salivary mucin, which is 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.
	ENSG00000163631	albumin	functions primarily as a carrier protein for steroids, fatty acids, and thyroid hormones and plays a role in stabilizing extracellular fluid volume.
	ENSG00000257017	haptoglobin	encodes a preproprotein, which is processed to yield both alpha and beta chains, which subsequently combine as a tetramer to produce haptoglobin, linked to diabetic nephropathy, Crohn's disease, inflammatory disease behavior, primary sclerosing cholangitis and reduced incidence of Plasmodium falciparum malaria.
cluster salmon (Pituitary)	ENSG00000172179	prolactin 2	encodes the anterior pituitary hormone prolactin. This secreted hormone is a growth regulator for many tissues, including cells of the immune system.
	ENSG00000259384	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.
	ENSG00000115138	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.

5.2 Supplementary Table 1

Cluster	Gene names	Proteins	Summary
cluster 1,red	ENSG00000160014	calmodulin 3 (phosphorylase kinase, delta)	is a calcium binding protein that plays a role in signaling pathways, cell cycle progression and proliferation.
	ENSG00000127585	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.
	ENSG00000154277	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 2,green	ENSG00000018625	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
	ENSG00000120885	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.
	ENSG00000132002	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 3, blue	ENSG00000132639	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
	ENSG00000111674	enolase 2 (gamma, neuronal)	encodes one of the three enolase isoenzymes found in mammals, is found in mature neurons and cells of neuronal origin
	ENSG00000089199	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 4, cyan	ENSG00000197971	myelin basic protein	protein encoded is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system.
	ENSG00000133392	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
	ENSG00000107796	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.