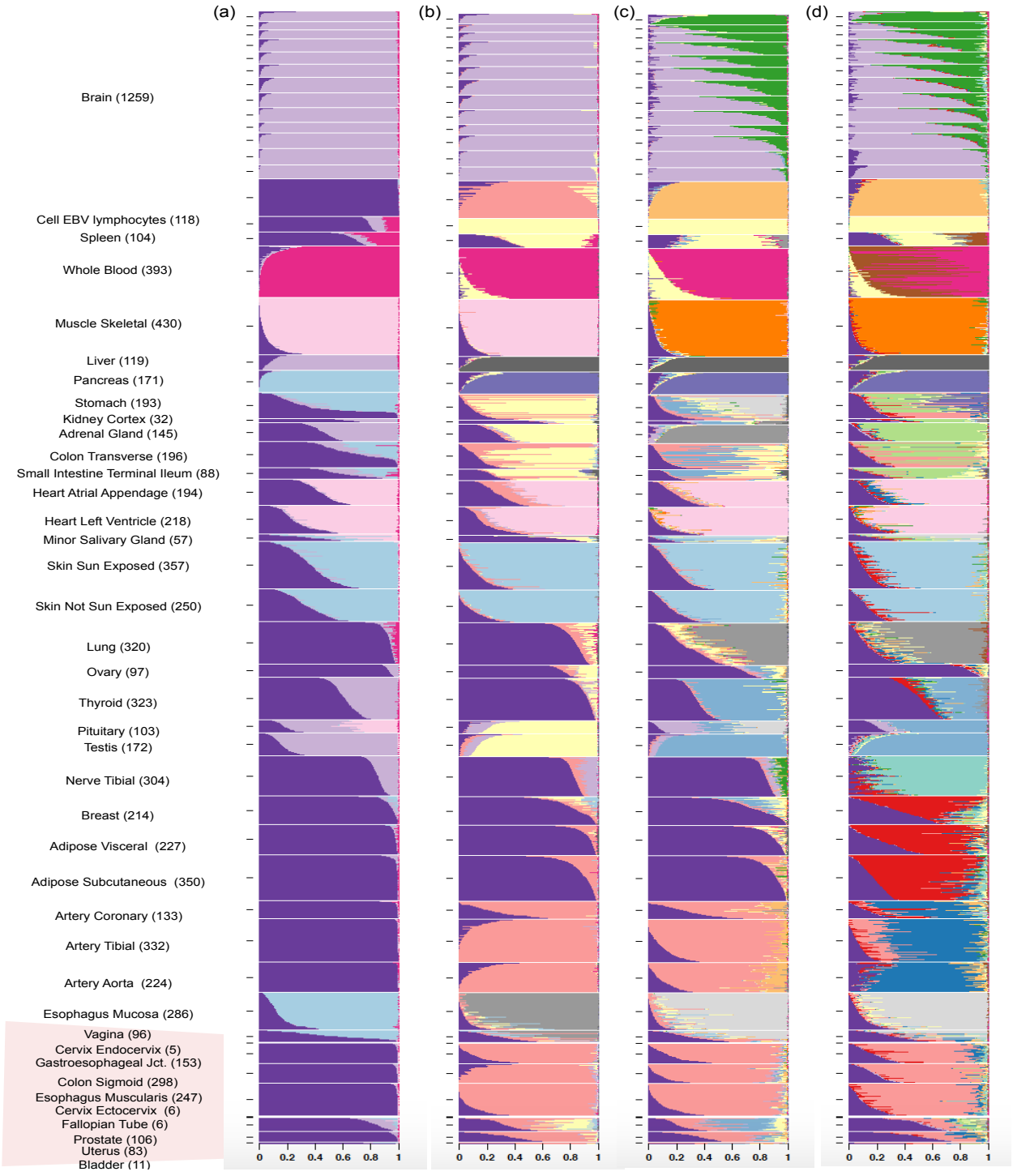
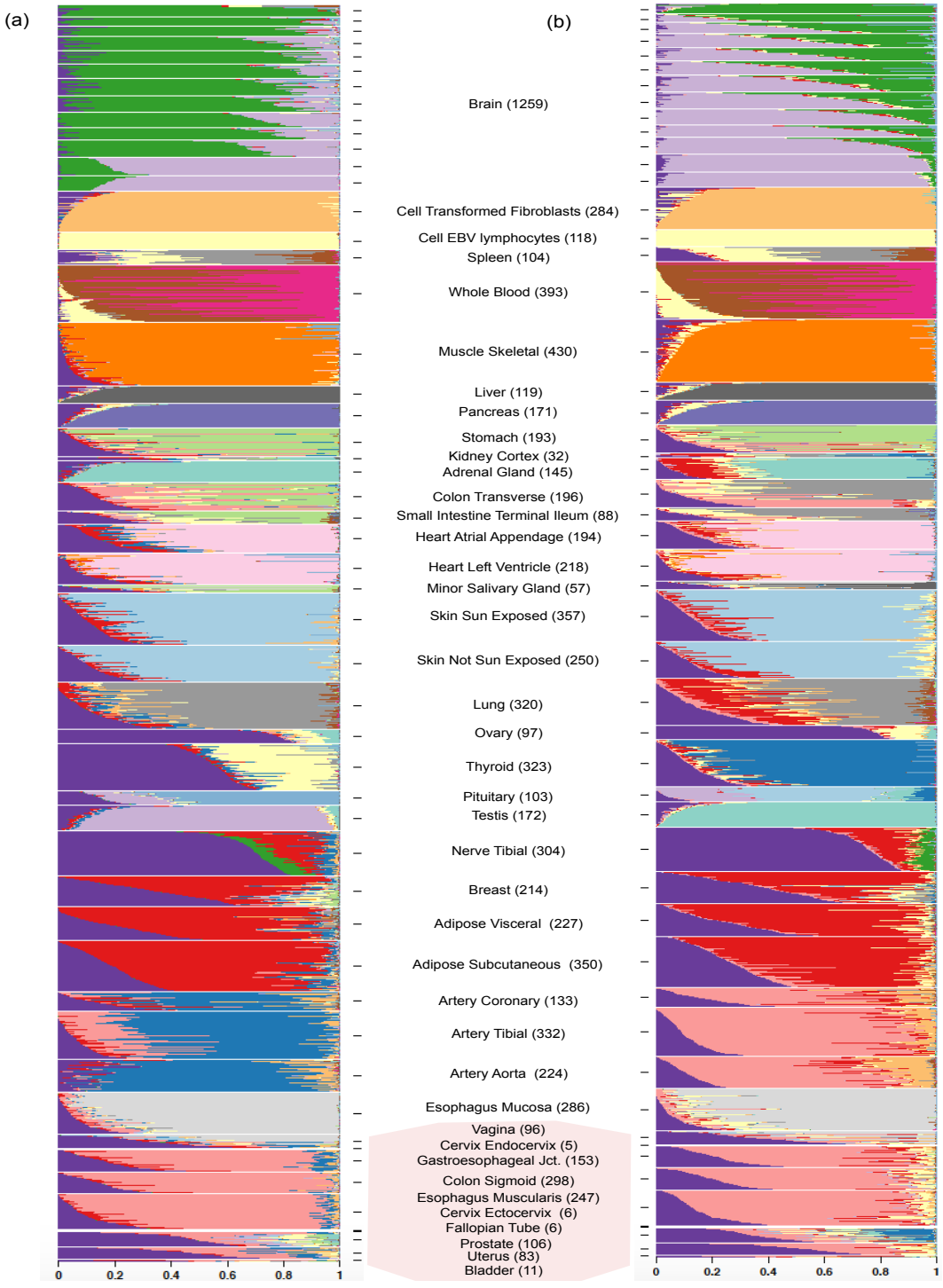
1 Supplementary figures

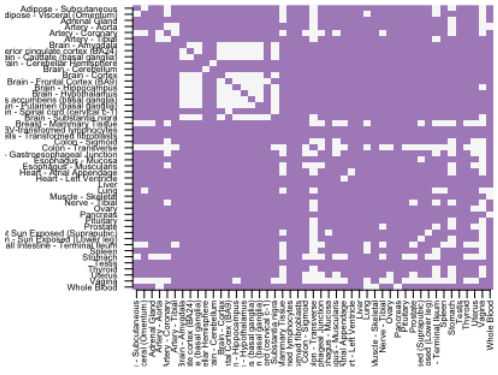
S1 Fig. Structure plot of GTEx V6 tissue samples for (A) $K = 5$**, (B)** $K = 10$**, (C)** $K = 15$**, (D)** $K = 20$**.** Some tissues form a separate cluster from the other tissues from $K = 5$ onwards (for example: Whole Blood, Skin), whereas some tissue only form a distinctive subgroup at $K = 20$ (for example: Arteries).


(a)
(b)
(c)
(d)
Brain (1259)
Cell EBV lymphocytes (118)
Spleen (104)
Whole Blood (393)
Muscle Skeletal (430)
Liver (119)
Pancreas (171)
Stomach (193)
Kidney Cortex (32)
Adrenal Gland (145)
Colon Transverse (196)
Small Intestine Terminal Ileum (88)
Heart Atrial Appendage (194)
Heart Left Ventricle (218)
Minor Salivary Gland (57)
Skin Sun Exposed (357)
Skin Not Sun Exposed (250)
Lung (320)
Ovary (97)
Thyroid (323)
Pituitary (103)
Testis (172)
Nerve Tibial (304)
Breast (214)
Adipose Visceral (227)
Adipose Subcutaneous (350)
Artery Coronary (133)
Artery Tibial (332)
Artery Aorta (224)
Esophagus Mucosa (286)
Vagina (96)
Cervix Endocervix (5)
Gastroesophageal Jct. (153)
Colon Sigmoid (298)
Esophagus Muscularis (247)
Cervix Ectocervix (6)
Fallopian Tube (6)
Prostate (106)
Uterus (83)
Bladder (11)
0 0.2 0.4 0.6 0.8 1

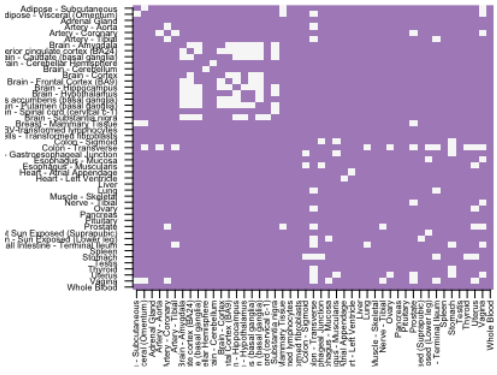
S2 Fig. Structure plot of GTEx V6 tissue samples for $K = 20$ in two runs under different thinning parameter settings. (A) $p_{thin} = 0.01$ and (B) $p_{thin} = 0.0001$. The structural patterns in these two plots closely resemble the structural patterns in Fig 1(a), though there are a few differences from the unthinned version.


(a)
(b)
Brain (1259)
Cell Transformed Fibroblasts (284)
Cell EBV lymphocytes (118)
Spleen (104)
Whole Blood (393)
Muscle Skeletal (430)
Liver (119)
Pancreas (171)
Stomach (193)
Kidney Cortex (32)
Adrenal Gland (145)
Colon Transverse (196)
Small Intestine Terminal Ileum (88)
Heart Atrial Appendage (194)
Heart Left Ventricle (218)
Minor Salivary Gland (57)
Skin Sun Exposed (357)
Skin Not Sun Exposed (250)
Lung (320)
Ovary (97)
Thyroid (323)
Pituitary (103)
Testis (172)
Nerve Tibial (304)
Breast (214)
Adipose Visceral (227)
Adipose Subcutaneous (350)
Artery Coronary (133)
Artery Tibial (332)
Artery Aorta (224)
Esophagus Mucosa (286)
Vagina (96)
Cervix Endocervix (5)
Gastroesophageal Jct. (153)
Colon Sigmoid (298)
Esophagus Muscularis (247)
Cervix Ectocervix (6)
Fallopian Tube (6)
Prostate (106)
Uterus (83)
Bladder (11)
0
0.2
0.4
0.6
0.8
1

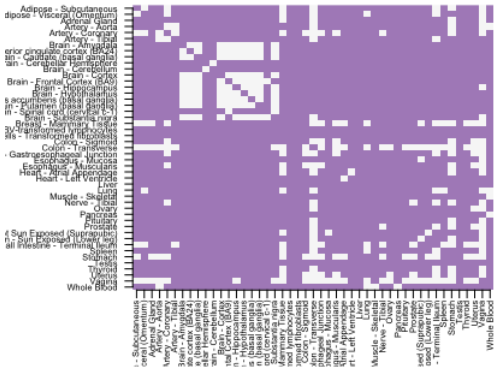
S3 Fig. A comparison of "accuracy" of hierarchical clustering vs. GoM on thinned GTEx data, with thinning parameters of $p_{thin} = 0.01$ and $p_{thin} = 0.001$. For each pair of tissue samples from the GTEx V6 data we assessed whether or not each clustering method (with $K = 2$ clusters) separated the samples according to their tissue of origin, with successful separation indicated by a filled square. Thinning deteriorates accuracy compared with the unthinned data (Fig 2), but even then the model-based method remains more successful than the hierarchical clustering in separating the samples by tissue or origin.



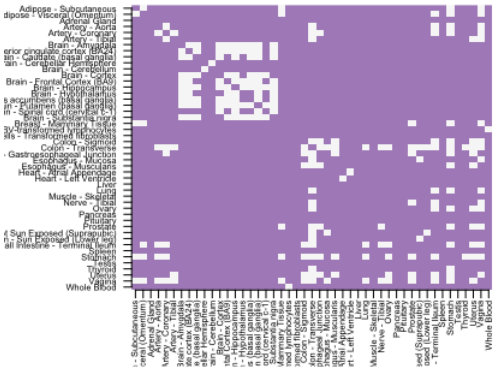
(a) hierarchy thin 0.01



(b) GoM thin 0.01

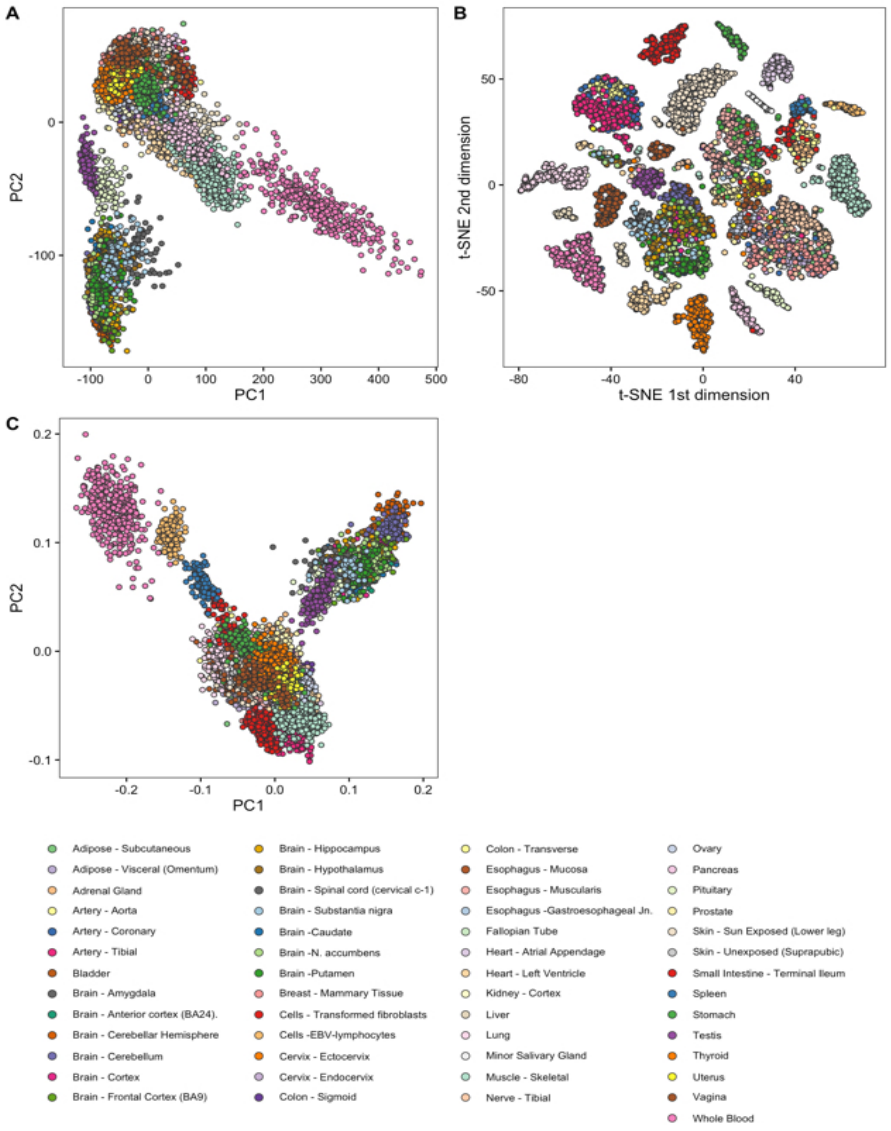


(c) hierarchy 0.001

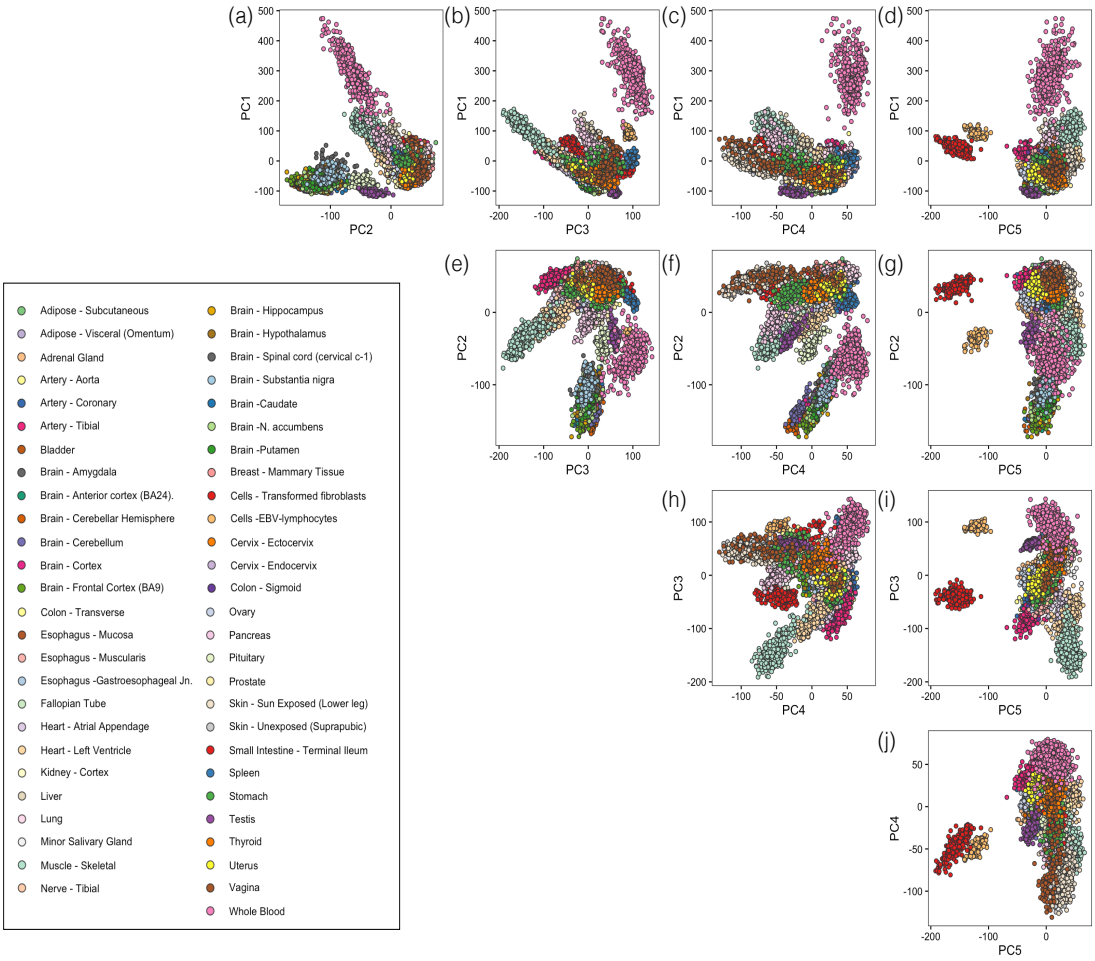


(d) GoM thin 0.001

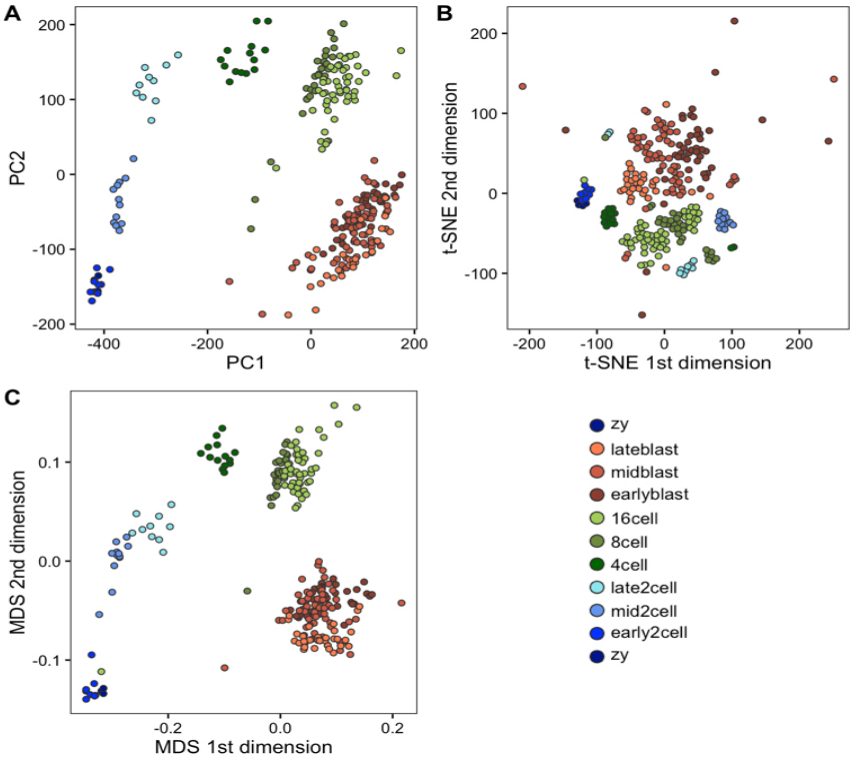
S4 Fig. GTEx V6 tissue samples visualization using (A) principle component analysis, (B) t-SNE, and (C) multidimensional scaling. The colors represent different tissue types.



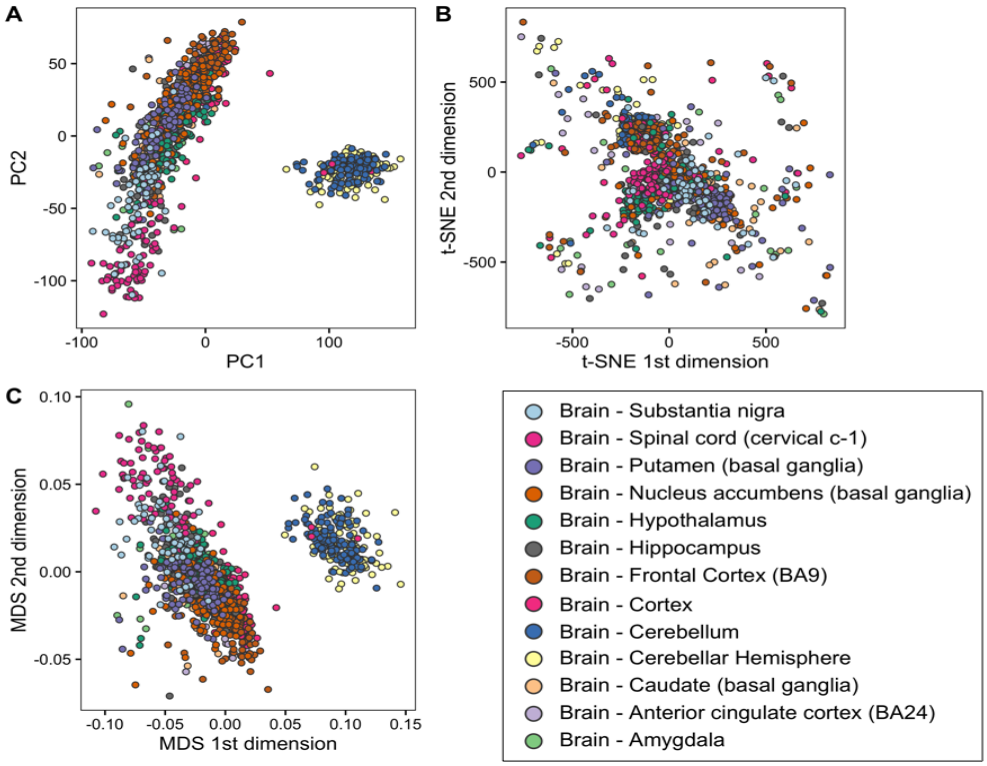
S5 Fig. Top five principal components (PC) for GTEx V6 tissue samples. Scatter plot representation of the top five PCs of the GTEx tissue samples. Data was transformed to log2 counts per million (CPM).



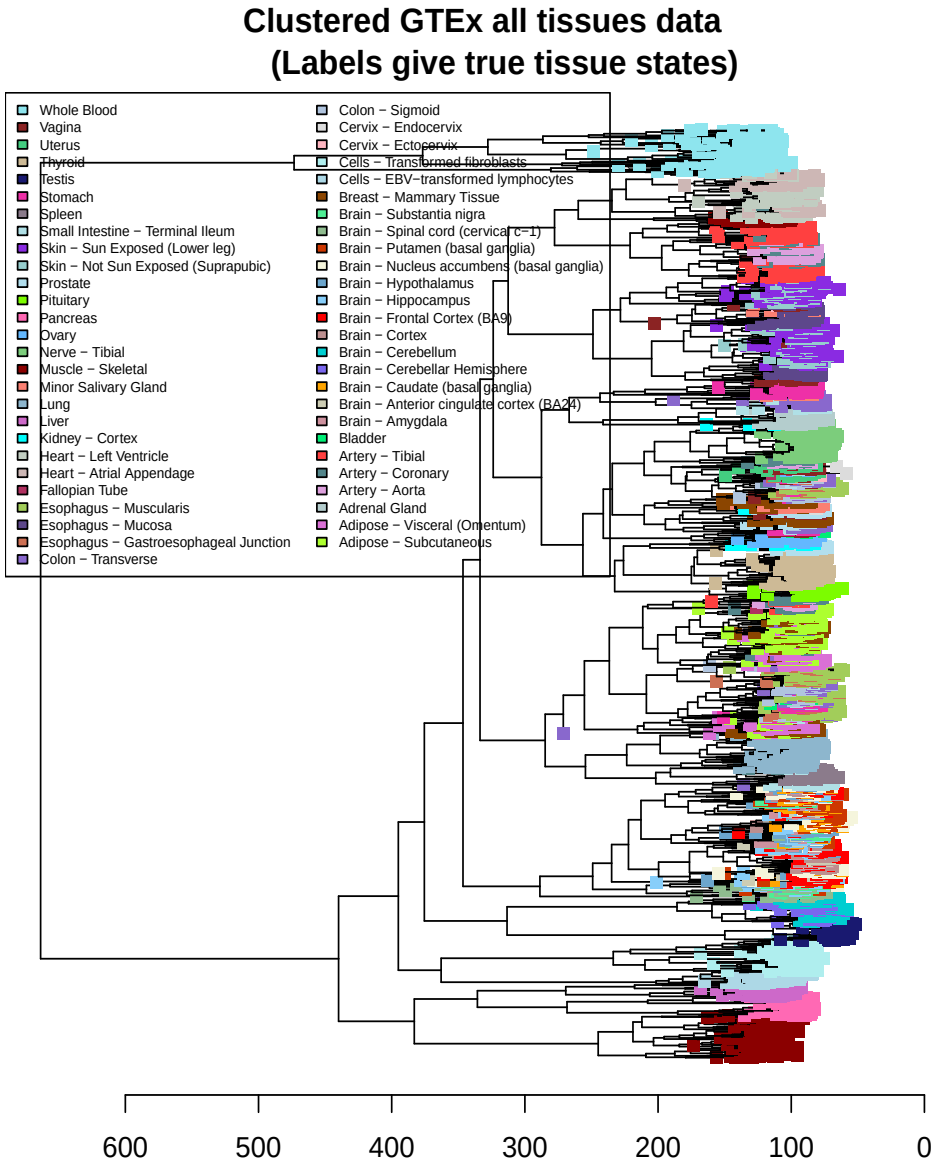
S6 Fig. Visualizing mouse pre-implantation embryos data from Deng et al (2014) using (A) Principle Component Analysis, (B) t-SNE and (C) Multidimensional Scaling. The colors represent different developmental stages.



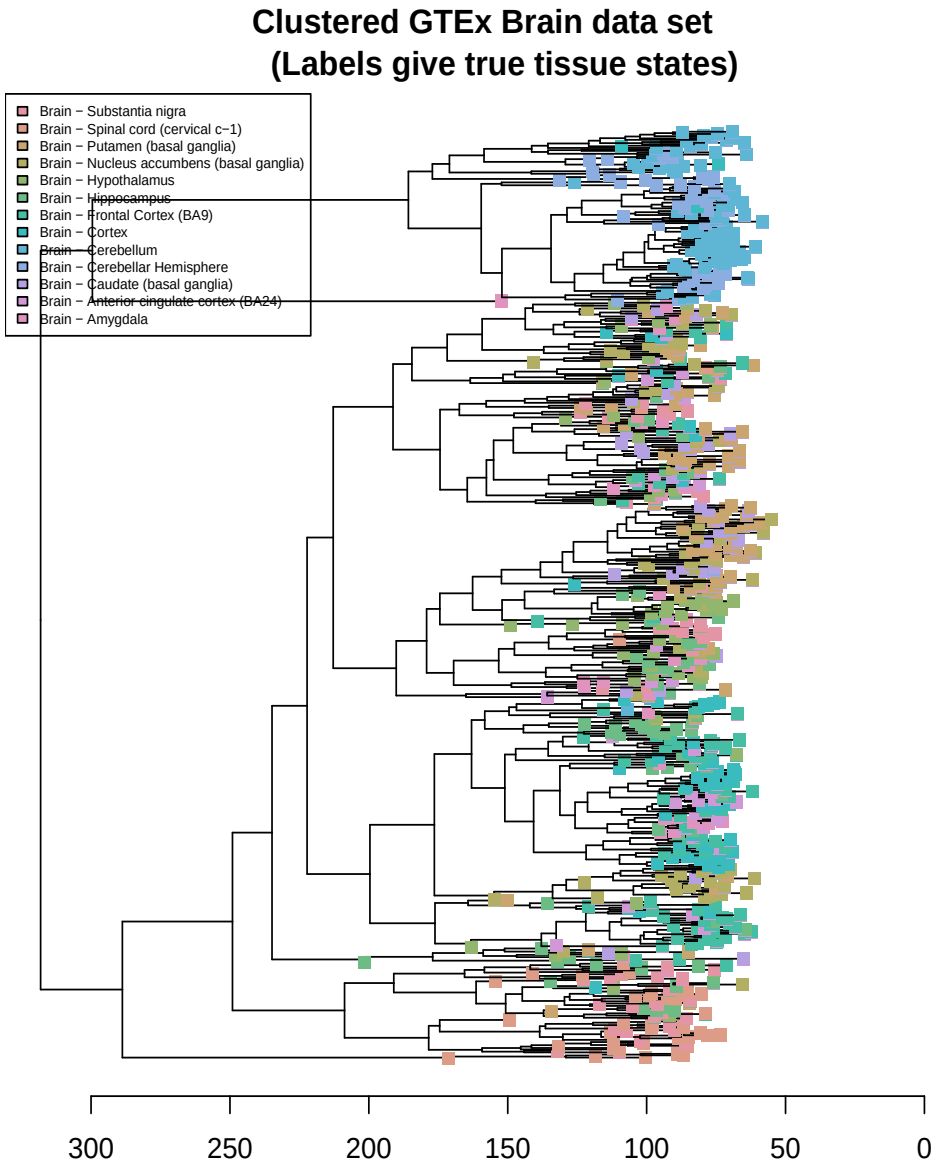
S7 Fig. GTEx brain tissue samples visualization using (A) principle component analysis, (B) t-SNE, and (C) Multidimensional scaling. The colors represent the 13 different brain tissue types. In (A) and (B), the majority of the tissue samples are distinct from Cerebellum tissue samples (the cluster of samples located on the right side of the plot). While, in (C), most tissue samples are located at the enter of the plot and are similar to each other in the t-SNE dimensions.



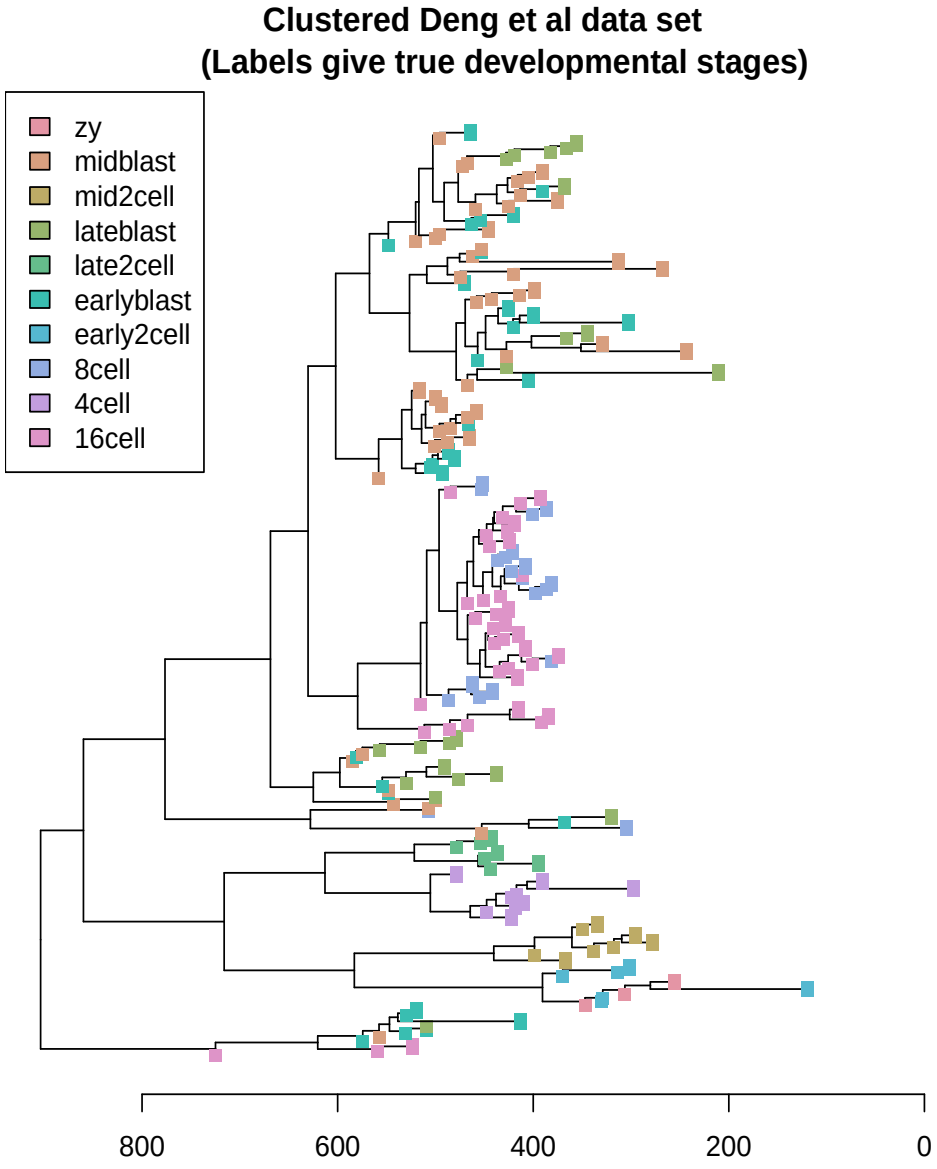
S8 Fig. Dendrogram visualization of hierarchical clustering results of GTEx V6 tissue samples. Hierarchical clustering of Euclidean distance based on complete linkage was applied to 8,555 tissue samples from the GTEx V6 data. Data was transformed to log counts per million (CPM) prior to clustering. Complete linkage method was used to plot the Dendrogram. The colors represent different tissue types. Samples from different tissues seem to cluster together, but any further patterns. However, because of the large number of samples, patterns of structural variation between tissue samples remain difficult to detect.



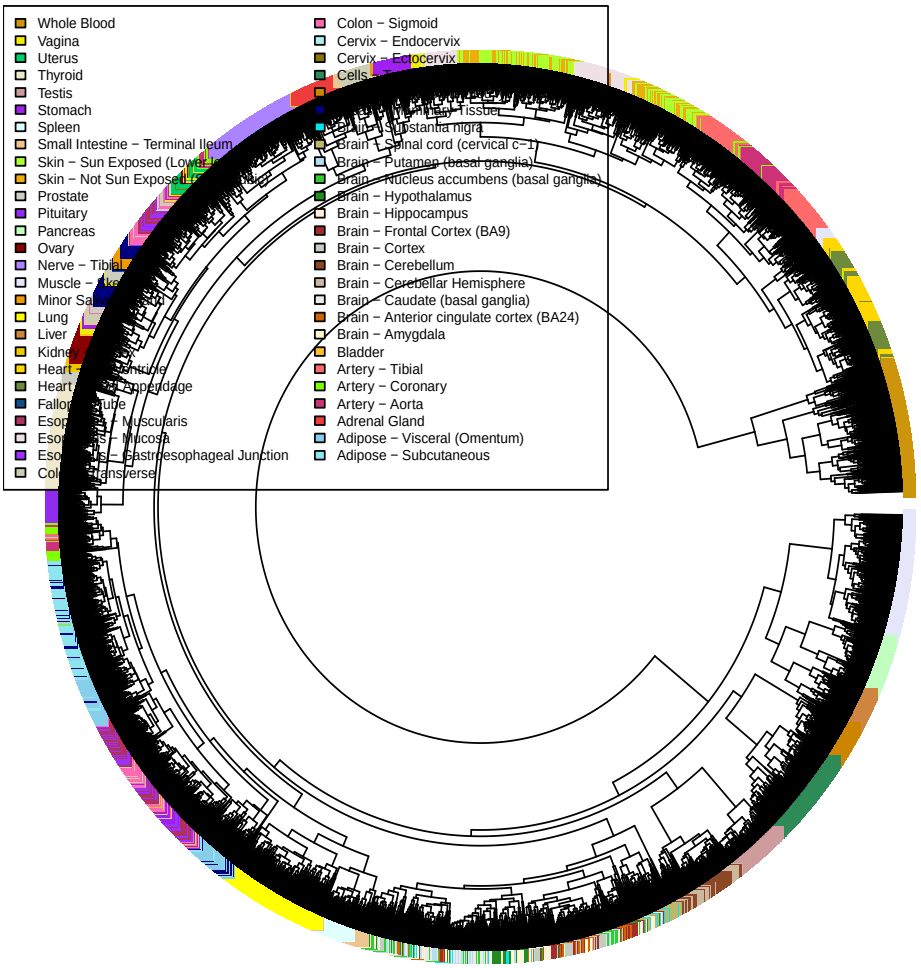
S9 Fig. Hierarchical clustering results visualization of GTEx brain tissue samples. Hierarchical clustering of Euclidean distance based on complete linkage was applied to 1,259 GTEx brain tissue samples. Data was transformed to log counts per million (CPM) prior to clustering. Complete linkage method was used to plot the Dendrogram. The colors represent different brain tissue types. We note that the sample labels in the Dendrogram are not easy to read because of the large sample size. However, there seems to be a separation between cerebellum, cerebellar hemisphere, spinal cord, and substantia nigra from other parts of the brain.



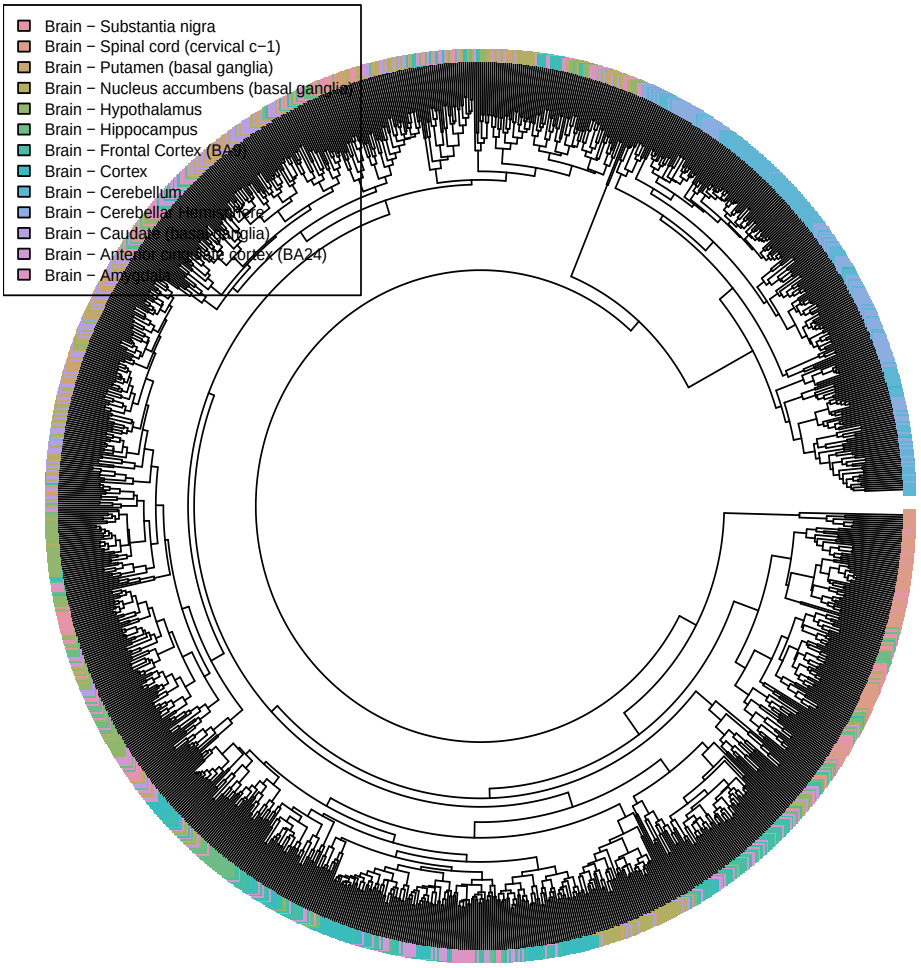
S10 Fig. Dendrogram visualization of hierarchical clustering results of mouse pre-implantation embryos data from Deng et al (2014). Hierarchical clustering of Euclidean distance based on complete linkage was applied to 259 single cell samples from Deng et al., (2014). Data was transformed to log counts per million (CPM) prior to clustering. The colors represent different developmental phases. Although samples from different developmental stages seem to cluster together, the samples are not arranged in a continuum according to their developmental stages.



S11 Fig. Circular dendrogram of hierarchical clustering results for GTEx V6 tissues samples. Alternative visual representation of S8 Fig. Circular representation of the dendrogram makes the tissue labels easier to read. Samples from different tissues seem to cluster together. But, because of the large number of samples, patterns of structural variation between tissue samples remain difficult to detect.

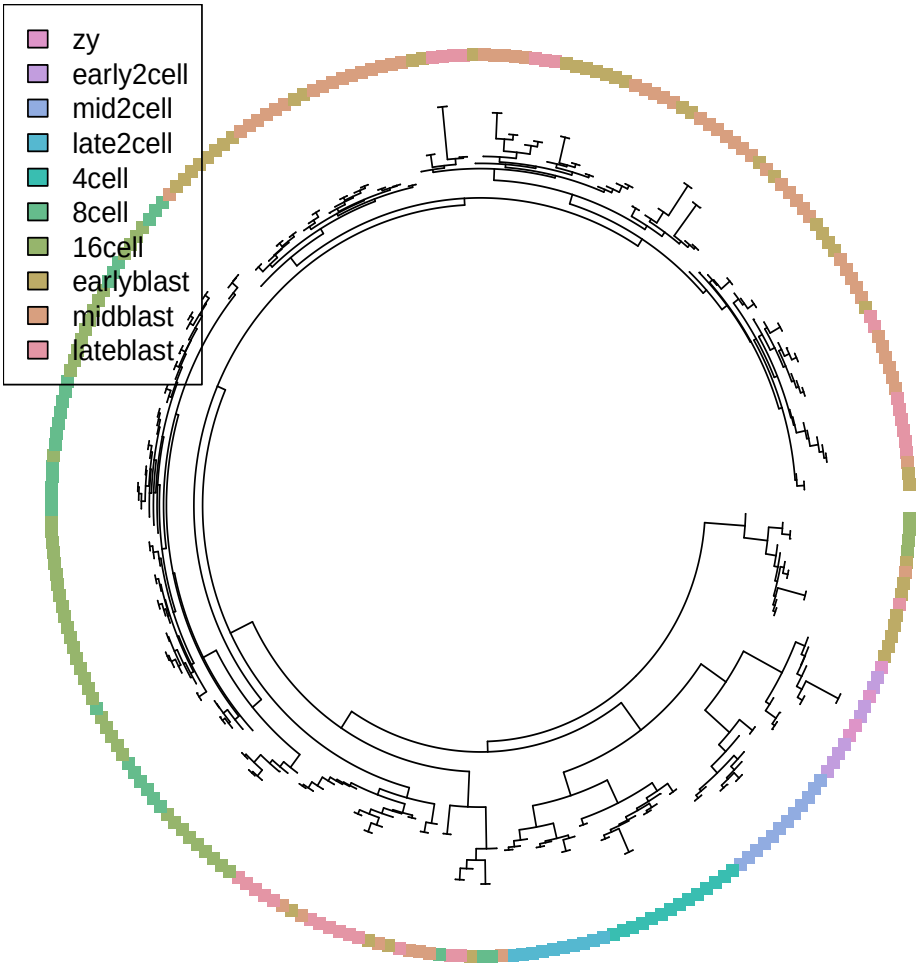


S12 Fig. Circular dendrogram of hierarchical clustering results for GTEx brain tissues samples. Alternative visual representation of S9 Fig. Circular representation of the dendrogram makes the tissue labels easier to read. Samples from different brain regions seem to cluster together. In particular, Brain Cerebellar, Cerebellar Hemisphere, Brain Spinal Cord, and Substantia Nigra seem to cluster together and separate from samples from other brain regions. But, because of the large number of samples, patterns of structural variation between tissue samples remain difficult to detect.

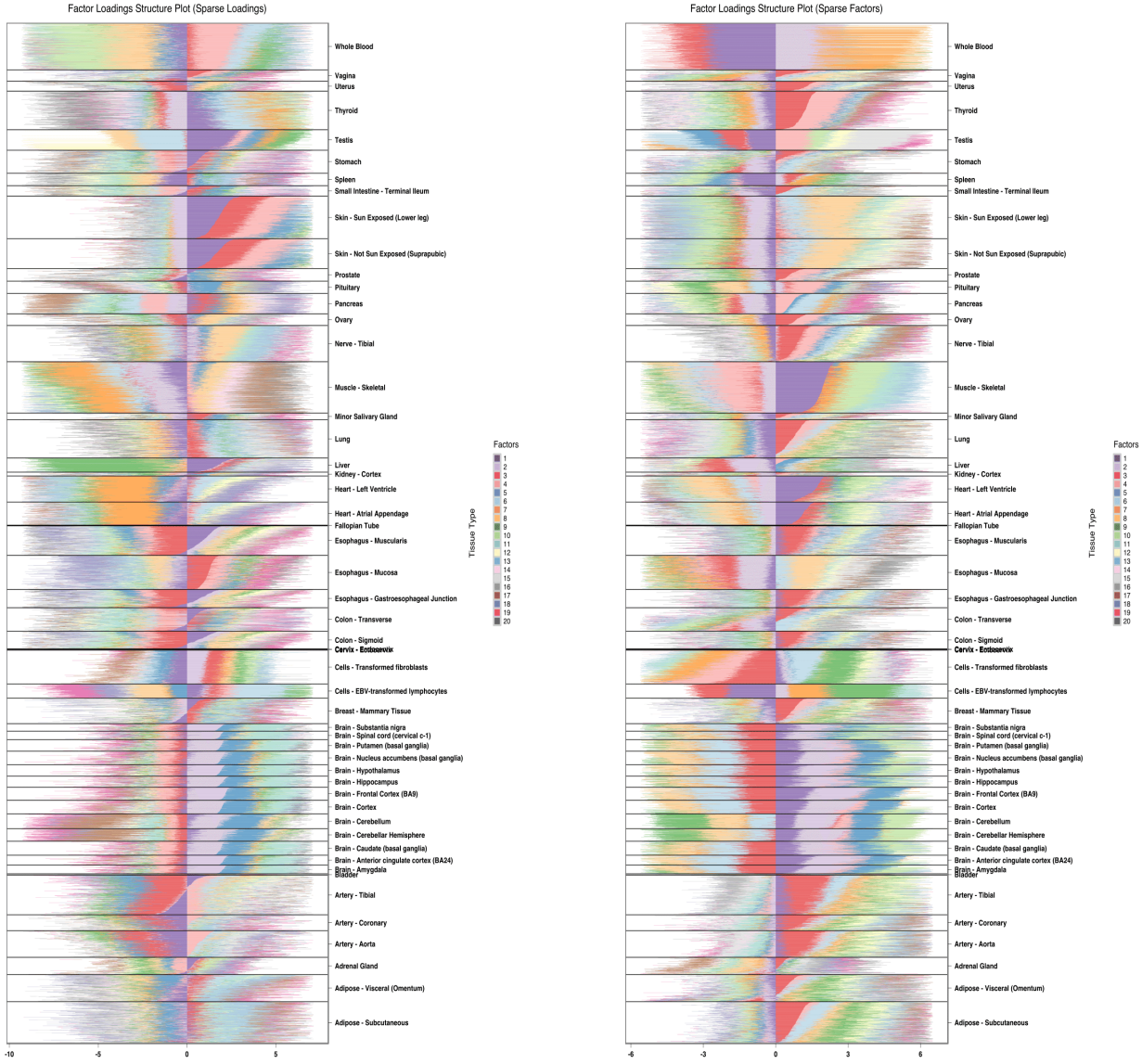


S13 Fig. Note that the labels in this case are difficult to interpret and also the hierarchical clustering fails to represent the continuum among the different developmental phases as in PCA or the GoM model.

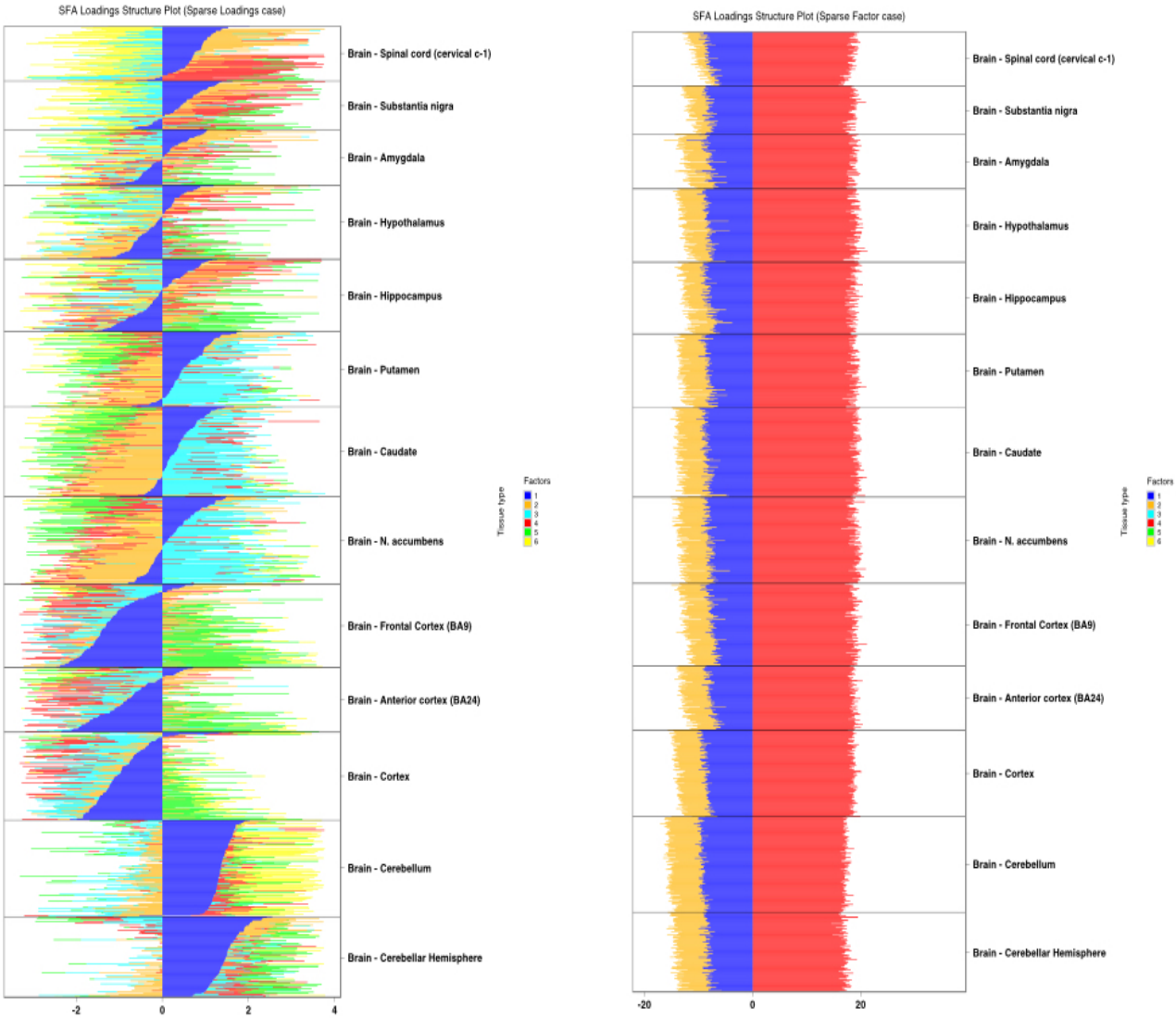
Circular dendrogram of hierarchical clustering results for mouse pre-implantation embryos data from Dent et al (2014). Alternative visual representation of S10 Fig. Circular representation of the dendrogram makes the sample labels easier to read. Although samples from different developmental stages seem to cluster together, the samples are not arranged in a continuum according to their developmental stages.



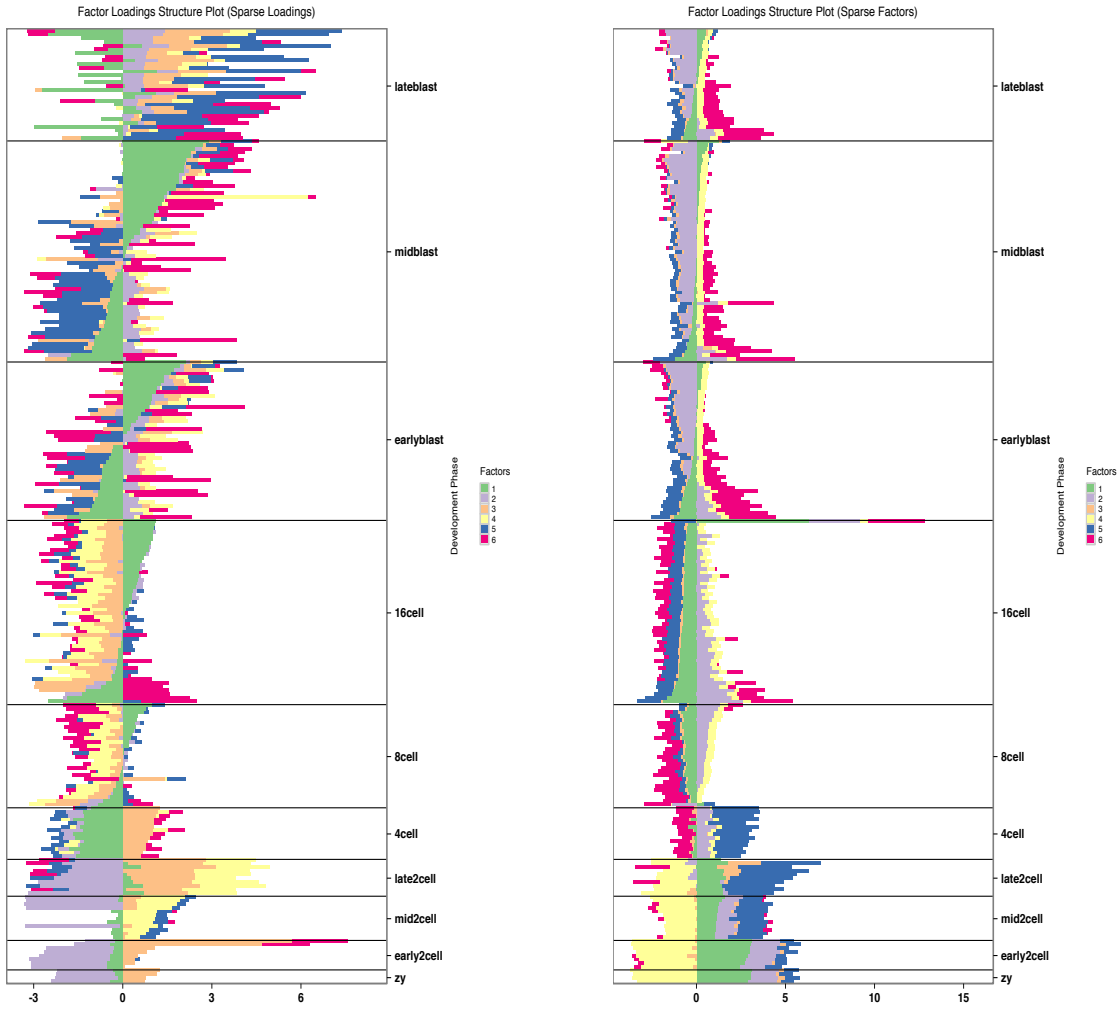
S14 Fig. Sparse Factor Analysis loadings visualization of GTEx V6 tissue samples. The colors represent the 20 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of (A) when the loadings are sparse and (B) when the factors are sparse.



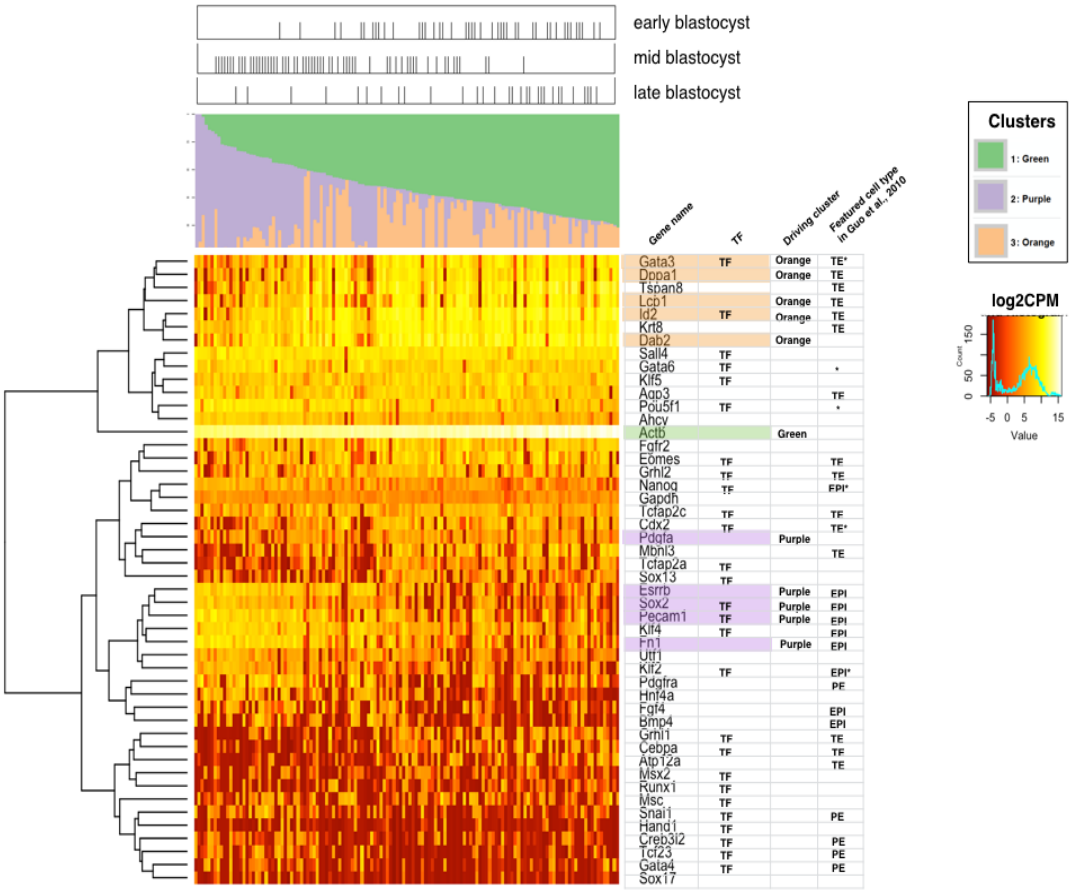
S15 Fig. Sparse Factor Analysis loadings visualization of GTEx brain tissue samples. The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of (A) when the loadings are sparse and (B) when the factors are sparse.



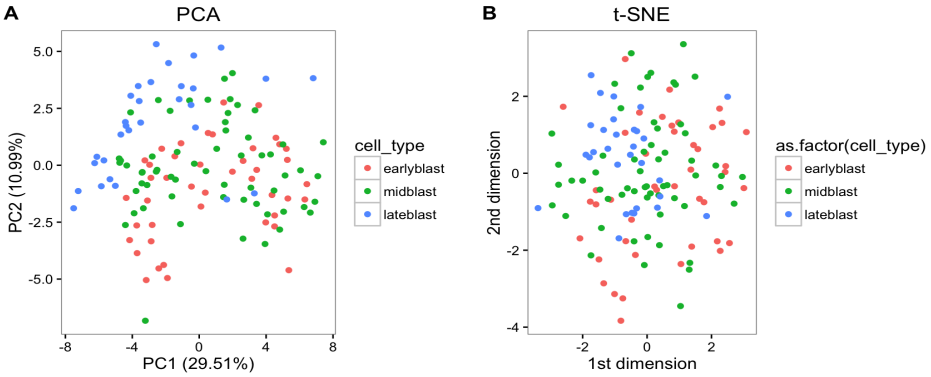
S16 Fig. Sparse Factor Analysis loadings visualization of mouse pre-implantation embryos from Deng et al., (2014). The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of (A) when the loadings are sparse and (B) when the factors are sparse.



S17 Fig. Additional GoM analysis of Deng et al (2014) data including blastocyst samples and 48 blastocyst marker genes. We considered 48 blastocyst marker genes (see Guo et al., 2010 for inclusion criteria) and fitted GoM model with $K = 3$ to 133 blastocyst samples. In the Structure plot, blastocyst samples are arranged in ascending order from right to left by their membership proportions in the Green cluster. The panel located above the Structure plot shows the corresponding pre-implantation stage from which blastocyst samples were collected. The heatmap located below the Structure plot represents expression levels of the 48 blastocyst marker genes (log2 CPM), and the corresponding dendrogram represents the results of hierarchical clustering based on complete lineage method. The table located to the right of the expression heatmap display gene information, showing, from left to right, 1) genes that are known as transcription factor, 2) the driving GoM cluster if the gene was among the top five driving genes, and 3) the featured cell type (TE: trophecoderm, EPI: epiblast, PE: primitive endoderm) that was found in Guo et al., 2010.

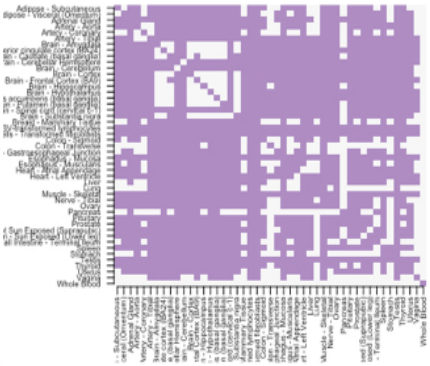


S18 Fig. Visualization of PCA and t-SNE results of mouse pre-implantation embryos data from Deng et al (2014) using 48 blastocyst marker genes

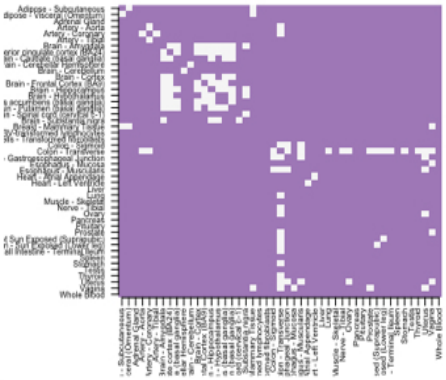


S19 Fig. Comparison between GoM model and hierarchical clustering under different scenarios of data transformation. We used GTEx V6 data for model performance comparisons. Specifically, for every pair of the 53 tissues, we assessed the ability of the methods to separate samples according to their tissue of origin. The subplots of heatmaps show the results of evaluation under different scenarios. Filled squares in the heatmap indicate successful separation of the samples in corresponding tissue pair comparison. (A) Hierarchical clustering on log2 counts per million (CPM) transformed data using Euclidean distance. (B) Hierarchical clustering on the standardized log2-CPM transformed data (transformed values for each gene was mean and scale transformed) using the Euclidean distance. (D) Hierarchical clustering on counts data with the assumption that, for each gene the sample read count c_{ng} has a variance $\bar{c}_g + 1$ that is constant across samples. And, the the gene-specific variance $\bar{c}_g + 1$ was used to scale the distance matrix for clustering. (C) GoM model of $K = 2$ applied to counts. (E) Hierarchical clustering applied to adjusted count data. Each gene has a mean expression value of 0 and variance of 1. Taken together, these results suggest that regardless of the different data transformation scenarios, the GoM model with $K = 2$ is able to separate samples of different tissue of origin, better than hierarchical cluster methods.

(a)



(b)



(c)

(d)

(e)

2 Supplementary tables

S1 Table. Cluster Annotations of GTEx V6 data with top driving gene summaries.

Cluster	Top Driving Genes	Gene names	Gene Summary
1, Royal purple	*NEAT1*	nuclear paraspeckle assembly transcript 1	produces a long non-coding RNA (lncRNA) transcribed from the multiple endocrine neoplasia locus, regulates genes involved in cancer progression.
	CCNL2	cyclin L2	regulator of the pre-mRNA splicing process, as well as in inducing apoptosis by modulating the expression of apoptotic and antiapoptotic proteins.
	SRSF5	serine/arginine-rich splicing factor 5	encodes proteins of serine/arginine (SR)-rich family, involved in mRNA export from the nucleus and in translation.
2, Light purple	*SNAP25*	synaptosomal-associated protein, 25kDa	this gene product is a presynaptic plasma membrane protein involved in the regulation of neurotransmitter release.
	FBXL16	F-box and leucine-rich repeat protein 16	members of F-box protein family, which interact with SKP1 through the F box, and they interact with ubiquitination targets through other protein interaction domains.
	SLC17A7	neurochondrin	encodes proteins expressed in neuron-rich regions; associated with the membranes of synaptic vesicles and functions in glutamate transport.
3, Red	*FABP4*	fatty acid binding protein 4	encodes the fatty acid binding protein found in adipocytes, takes part in fatty acid uptake, transport, and metabolism.
	PLIN1	perilipin 1	protein encoded by this gene coats lipid storage droplets in adipocytes, thereby protecting them until they can be broken down by hormone-sensitive lipase.
	FASN	fatty acid synthase	catalyze the synthesis of palmitate from acetyl-CoA and malonyl-CoA, in the presence of NADPH, into long-chain saturated fatty acids.
4, Salmon	*ACTG2*	actin, gamma 2, smooth muscle, enteric	involved in various types of cell motility and in the maintenance of the cytoskeleton.
	MYH11	myosin, heavy chain 11, smooth muscle	protein encoded by this gene is a smooth muscle myosin belonging to the myosin heavy chain family, functions as a major contractile protein, converting chemical energy into mechanical energy through the hydrolysis of ATP.
	SYNM	synemin	protein has been found to form a linkage between desmin, which is a subunit of the IF network, and the extracellular matrix, and provides an important structural support in muscle.
5, Denim	*RGS5*	regulator of G-protein signaling 5	encodes a member of the regulators of G protein signaling (RGS) family, associated with retinal arterial macroaneurysm.
	MFGE8	milk fat globule-EGF factor 8 protein	encodes a preproprotein that is proteolytically processed to form multiple protein products, been implicated in wound healing, autoimmune disease, and cancer
	ITGA8	synemin	Proteins generated mediate numerous cellular processes including cell adhesion, cytoskeletal rearrangement, and activation of cell signaling pathways.
6, Light denim	*KRT10*	keratin 10	encodes a member of the type I (acidic) cytokeratin family, mutations associated with epidermolytic hyperkeratosis.
	KRT1	keratin 1, type II	specifically expressed in the spinous and granular layers of the epidermis with family member KRT10 and mutations in these genes have been associated with bullous congenital ichthyosiform erythroderma.
	KRT2	keratin 2, type II	expressed largely in the upper spinous layer of epidermal keratinocytes and mutations in this gene have been associated with bullous congenital ichthyosiform erythroderma.
7, Orange	*NEB*	nebulin	encodes nebulin, a giant protein component of the cytoskeletal matrix that coexists with the thick and thin filaments within the sarcomeres of skeletal muscle, associated with recessive nemaline myopathy.
	MYH1	myosin, heavy chain 1, skeletal muscle, adult	a major contractile protein which converts chemical energy into mechanical energy through the hydrolysis of ATP.
	MYH2	myosin, heavy chain 2, skeletal muscle, adult	encodes a member of the class II or conventional myosin heavy chains, and functions in skeletal muscle contraction.

Cluster	Top Driving Genes	Gene namese	Gene Summary
8, Light orange	*FN1*	fibronectin 1	Fibronectin is involved in cell adhesion, embryogenesis, blood coagulation, host defense, and metastasis.
	COL1A1	collagen, type I, alpha 1	Mutations in this gene associated with osteogenesis imperfecta types I-IV, Ehlers-Danlos syndrome type and Classical type, Caffey Disease.
	COL1A2	collagen, type I, alpha 2	Mutations in this gene associated with osteogenesis imperfecta types I-IV, Ehlers-Danlos syndrome type and Classical type, Caffey Disease.
9, Green	*MBP*	myelin basic protein	major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system
	GFAP	glial fibrillary acidic protein	encodes one of the major intermediate filament proteins of mature astrocytes, mutations casuses Alexander disease.
	CARNS1	carnosine synthase 1	catalyzes the formation of carnosine and homocarnosine, which are found mainly in skeletal muscle and the central nervous system, respectively.
10, Light green	*CYP17A1*	cytochrome P450 family 17 subfamily A member 1	encodes a member of the cytochrome P450 superfamily of enzymes, mutations in this gene are associated with isolated steroid-17 alpha-hydroxylase deficiency,20-lyase deficiency, pseudohermaphroditism, and adrenal hyperplasia.
	CYP11B1	cytochrome P450 family 11 subfamily B member 1	The protein encoded by this gene plays a key role in the acute regulation of steroid hormone synthesis by enhancing the conversion of cholesterol into pregnenolone, associated with congenital lipoid adrenal hyperplasia.
	GKN1	gastrokine 1	protein encoded by this gene is found to be down-regulated in human gastric cancer tissue as compared to normal gastric mucosa..
11, Turquoise	*MPZ*	myelin protein zero	specifically expressed in Schwann cells of the peripheral nervous system and encodes a type I transmembrane glycoprotein that is a major structural protein of the peripheral myelin sheath, mutations associated with autosomal dominant form of Charcot-Marie-Tooth disease type 1 and other polyneuropathies.
	APOD	apolipoprotein D	encodes a component of high density lipoprotein that has no marked similarity to other apolipoprotein sequences, closely associated with lipoprotein metabolism.
	PMP22	peripheral myelin protein 22	encodes an integral membrane protein that is a major component of myelin in the peripheral nervous system..
12, Yellow	*IGHM*	immunoglobulin heavy constant mu	IgM antibodies play an important role in primary defense mechanisms, Diseases associated with IGHM include agammaglobulinemia 1 and immunodeficiency 23.
	IGHG1	immunoglobulin heavy constant gamma 1 (G1m marker)	antigen binding functionality, diseases associated with IGHG1 include heavy chain deposition disease and chronic lymphocytic leukemia.
	IGHG2	immunoglobulin heavy constant gamma 2 (G2m marker)	antigen binding gene, diseases associated with IGHG2 include c2 deficiency.
13, Sky blue	*TG*	thyroglobulin	thyroglobulin produced predominantly in thyroid gland, synthesizes thyroxine and triiodothyronine, associated with Graves disease and Hashimotot thyroiditis.
	PRL	prolactin 2	encodes the anterior pituitary hormone prolactin. This secreted hormone is a growth regulator for many tissues, including cells of the immune system.
	PRM2	protamine 2	Protamines are the major DNA-binding proteins in the nucleus of sperm.
14, Light pink	*NPPA*	natriuretic peptide A	protein encoded by this gene belongs to the natriuretic peptide family, controls extracellular fluid volume and electrolyte homeostasis, mutations Mutations associated with atrial fibrillation familial type 6.
	MYH6	myosin, heavy chain 6, cardiac muscle, alpha	encodes the alpha heavy chain subunit of cardiac myosin, mutations cause familial hypertrophic cardiomyopathy and atrial septal defect 3
	TNNT2	protamine 2	protein encoded by this gene is the tropomyosin-binding subunit of the troponin complex, mutations in this gene have been associated with familial hypertrophic cardiomyopathy as well as with dilated cardiomyopathy.

Cluster	Top Driving Genes	Gene namese	Gene Summary
15, Light gray	*KRT13*	keratin 13, type I	protein encoded by this gene is a member of the keratin gene family, associated with the autosomal dominant disorder White Sponge Nevus.
	KRT4	keratin 4, type II	protein encoded by this gene is a member of the keratin gene family, associated with White Sponge Nevus, characterized by oral, esophageal, and anal leukoplakia.
	CRNN	cornulin	may play a role in the mucosal/epithelial immune response and epidermal differentiation.
16, Gray	*SFTPB*	surfactant protein B	an amphipathic surfactant protein essential for lung function and homeostasis after birth, muttaions cause pulmonary alveolar proteinosis, fatal respiratory distress in the neonatal period.
	SFTPA2	surfactant protein A2	Mutations in this gene and a highly similar gene located nearby, which affect the highly conserved carbohydrate recognition domain, are associated with idiopathic pulmonary fibrosis.
	SFTPA1	surfactant protein A1	encodes a lung surfactant protein that is a member of C-type lectins called collectins, associated with idiopathic pulmonary fibrosis.
17, Brown	*CSF3R*	colony stimulating factor 3 receptor	protein encoded by this gene is the receptor for colony stimulating factor 3, a cytokine that controls the production, differentiation, and function of granulocytes, mutations a cause of Kostmann syndrome
	MMP25	matrix metallopeptidase 25	proteins are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis.
	IL1R2	interleukin 1 receptor type 2	protein encoded by this gene is a cytokine receptor that belongs to the interleukin 1 receptor family.
18, Purple	*PRSS1*	protease, serine 1	secreted by pancreas, associated with pancreatitis
	CPA1	carboxypeptidase A1	secreted by pancreas, linked to pancreatitis and pancreatic cancer
	PNLIP	pancreatic lipase	encodes a carboxyl esterase that hydrolyzes insoluble, emulsified triglycerides, and is essential for the efficient digestion of dietary fats. This gene is expressed specifically in the pancreas.
19, Pink	*HBB*	hemoglobin, beta	mutant beta globin causes sickle cell anemia, absence of beta chain/ reduction in beta globin leads to thalassemia.
	HBA2	hemoglobin, alpha 2	deletion of alpha genes may lead to alpha thalassemia.
	HBA1	hemoglobin, alpha 1	deletion of alpha genes may lead to alpha thalassemia.
20, Dark gray	*ALB*	albumin	functions primarily as a carrier protein for steroids, fatty acids, and thyroid hormones and plays a role in stabilizing extracellular fluid volume.
	HP	haptoglobin	encodes a preproprotein, which subsequently produces haptoglobin, linked to diabetic nephropathy, Crohn's disease, inflammatory disease behavior and reduced incidence of Plasmodium falciparum malaria.
	FGB	fibrinogen beta chain	protein encoded by this gene is the beta component of fibrinogen, mutations may lead to several disorders, including afibrinogenemia, dysfibrinogenemia, hypodysfibrinogenemia etc.

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

Cluster	Top Driving Genes	Gene names	Gene Summary
1, Royal blue	*CLU*	clusterin	protein encoded by this gene is a secreted chaperone that can under some stress conditions also be found in the cell cytosol, also involved in cell death, tumor progression, and neurodegenerative disorders.
	OXT	oxytocin/neurophysin I prepropeptide	encodes a precursor protein that is processed to produce oxytocin and neurophysin I, involved in contraction of smooth muscle during parturition and lactation, cognition, tolerance, adaptation and complex sexual and maternal behaviour.
	GLUL	glutamate-ammonia ligase	catalyzes the synthesis of glutamine from glutamate and ammonia in an ATP-dependent reaction, associated with congenital glutamine deficiency, and overexpression of this gene was observed in some primary liver cancer samples.
2, Turquoise	*ENC1*	ectodermal-neural cortex 1	plays a role in the oxidative stress response as a regulator of the transcription factor Nrf2, may play role in malignant transformation.
	NCALD	neurocalcin delta	encodes a member of the neuronal calcium sensor (NCS), a regulator of G protein-coupled receptor signal transduction.
	YWHAH	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta	mediate signal transduction by binding to phosphoserine-containing proteins, associated with early-onset schizophrenia and psychotic bipolar disorder.
3, Lime green	*PKD1*	polycystin 1, transient receptor potential channel interacting	functions as a regulator of calcium permeable cation channels and intracellular calcium homoeostasis. It is also involved in cell-cell/matrix interactions and may modulate G-protein-coupled signal-transduction pathways.
	CBLN3	cerebellin 3 precursor	contain a cerebellin motif and C-terminal C1q signature domain that mediates trimeric assembly of atypical collagen complexes
	CHGB	chromogranin B	encodes a tyrosine-sulfated secretory protein abundant in peptidergic endocrine cells and neurons. This protein may serve as a precursor for regulatory peptides.
4, Red	*PPP1R1B*	protein phosphatase 1 regulatory inhibitor sub- unit 1B	encodes a bifunctional signal transduction molecule, may serve as a therapeutic target for neurologic and psychiatric disorders.
	RGS14	regulator of G-protein signaling 14	attenuates the signaling activity of G-proteins, increases the rate of conversion of the GTP to GDP.
	NCDN	neurochondrin	encodes a leucine-rich cytoplasmic protein, essential for spatial learning processes.
5, Yellow orange	*MBP*	myelin basic protein	protein encoded is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system.
	GFAP	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.
	TF	transferrin	transport iron from the intestine, reticuloendothelial system, and liver parenchymal cells to all proliferating cells in the body, involved in the removal of certain organic matter and allergens from serum.
6, Yellow	*IQGAP1*	IQ motif containing GTPase activating protein 1	interacts with components of the cytoskeleton, with cell adhesion molecules, and with several signaling molecules to regulate cell morphology and motility.
	A2M	alpha-2-macroglobulin	inhibits many proteases, including trypsin, thrombin and collagenase. A2M is implicated in Alzheimer disease (AD) due to its ability to mediate the clearance and degradation of A-beta, the major component of beta-amyloid deposits.
	C3	complement component 3	plays a central role in the activation of complement system, associated with atypical hemolytic uremic syndrome and age-related macular degeneration in human patients.

S3 Table. Cluster Annotations of Deng data with top driving genes.

	go.id	name	significant
1	GO:0007276	gamete generation	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
2	GO:0007292	female gamete generation	GDF9; BCL2L10; PABPC1L; BMP15; WEE2; DAZL; NOBOX
3	GO:0048609	multicellular organismal reproductive process	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
4	GO:0032504	multicellular organism reproduction	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
5	GO:0019953	sexual reproduction	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
6	GO:0044702	single organism reproductive process	GDF9; NOBOX; PABPC1L; BCL2L10; BMP15; CREB3L4; TGFB2; CASP8; RNF114; RGS2; PTTG1; TDRD12; WEE2; SPIN1; DAZL
7	GO:0048477	oogenesis	WEE2; GDF9; NOBOX; PABPC1L; DAZL
8	GO:0044703	multi-organism reproductive process	BCL2L10; GDF9; NOBOX; PABPC1L; RGS2; CREB3L4; RNF114; BMP15; PTTG1; TDRD12; WEE2; SPIN1; DAZL
9	GO:0048599	oocyte development	WEE2; GDF9; PABPC1L; DAZL
10	GO:0009994	oocyte differentiation	WEE2; GDF9; PABPC1L; DAZL
11	GO:0051321	meiotic cell cycle	H1FOO; WEE2; TDRD12; SPIN1; PTTG1; DAZL
12	GO:0001556	oocyte maturation	WEE2; PABPC1L; DAZL
13	GO:0006306	DNA methylation	TDRD12; H1FOO; TET3; ZFP57
14	GO:0051302	regulation of cell division	TGFB2; PTTG1; TXNIP; WEE2; CHEK1; DAZL
15	GO:0060255	regulation of macromolecule metabolic process	TGFB2; NOBOX; BPGM; UBE2D3; NFYA; CASP8; BMP15; TXNIP; TDRD12; GDF9; BCL2L10

S3 Table continued. Deng et al (2014) Cluster 2 (magenta) top GO annotations.

	go.id	name	significant
1	GO:0016604	nuclear body	YTHDC1; RBM8A; CDK12; PSME4; PPP1R8; HIPK1; TOPORS
2	GO:0005814	centriole	SFI1; PLK2; ROCK1; TOPORS
3	GO:0044450	microtubule organizing center part	SFI1; PLK2; ROCK1; TOPORS

S3 Table continued. Deng et al (2014) Cluster 3 (yellow) top GO annotations.

	go.id	name	significant
1	GO:0044428	nuclear part	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; TOR1B; MIOS; NR1H3; POLR3K
2	GO:0031981	nuclear lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1
3	GO:0070013	intracellular organelle lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1; DNTTIP2; ZBTB10; ZBTB17
4	GO:0043233	organelle lumen	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1
5	GO:0005730	nucleolus	XPO1; DNTTIP2; ESF1; WDR43; ZDHHC7; HEATR1; POLR1E; DDX24; POLR3K
6	GO:0005634	nucleus	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; TOR1B; MIOS; NR1H3; EIF5B; POLR3K
7	GO:0044446	intracellular organelle part	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; SLU7; NFYB; SLC25A36; ECE2
8	GO:0005654	nucleoplasm	MAD2L2; SMARCC1; PPRC1; SLU7; NFYB; POLR1E; MIOS; POLR3K; XPO1; ZBTB10; ZBTB17
9	GO:0003723	RNA binding	PPRC1; EIF5B; XPO1; DNTTIP2; WDR43; DDX10; EIF3C; BCLAF1; EBNA1BP2; RARS
10	GO:0003676	nucleic acid binding	SMARCC1; PPRC1; SLU7; NFYB; POLR1E; EIF5B; POLR3K; XPO1; DNTTIP2
11	GO:0043231	intracellular membrane-bounded organelle	MAD2L2; PTDSS2; SMARCC1; TOR1B; PPRC1; SLU7; NFYB; ESF1; ECE2; LMAN1L
12	GO:0043229	intracellular organelle	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; ARRDC1; SLU7; NFYB; ESF1; ECE2
13	GO:0005874	microtubule	WDR43; KLHL21; HAUS6; CENPE; TEKT2; RACGAP1; WDR81; BCL2L11; KIF20B
14	GO:0044822	poly(A) RNA binding	WDR43; DNTTIP2; ESF1; NXF1; DDX10; HEATR1; EIF3C
15	GO:0044424	intracellular part	MAD2L2; PTDSS2; SMARCC1; KLHL21; TOR1B; PPRC1; SNAPC4; POLR3K; ARRDC1; SLU7; NFYB; ESF1; WDR43; ECE2; LMAN1L

S3 Table continued. Deng et al (2014) Cluster 4 (green) top GO annotations.

	go.id	name	significant
1	GO:0005829	cytosol	PARG; UAP1; PSMB10; TCEB1; RPLP0; EIF5; CNBP; RPS3; PSAT1; AACS; PMM1; EXOSC7; EIF3I; SET; BHMT; BHMT2
2	GO:0044444	cytoplasmic part	PARG; UAP1; PSMB10; TCEB1; HSPA8; SERINC1; EIF5; CNBP; RPS3; PSAT1; GPD2; AACS; GPR137B; STIP1; PMM1; EXOSC7; VPREB3; PEX16
3	GO:0055131	C3HC4-type RING finger domain binding	HSPA8; PINK1; DNAJA1
4	GO:1901575	organic substance catabolic process	PSMB10; TCEB1; RPLP0; RPS3; GPD2; PINK1; EXOSC7; ALLC; BHMT; HSP90AB1; RPL13A; ATG7; CUL5; UBXN1; ZMPSTE24
5	GO:0000151	ubiquitin ligase complex	DNAJA1; RNF7; UBE2C; HSPA8; FBXO15; SUGT1; DCAF4; CUL5; FBXL20
6	GO:0072655	protein localization to mitochondrion	TIMM17A; BNIP3L; ARIH2; PEMT; SFN; PINK1; HSP90AA1; TIMM23
7	GO:1901564	organonitrogen compound metabolic process	PSMB10; RPLP0; SERINC1; EIF5; BHMT2; PINK1; EIF3I; ALLC; BHMT; MRPL22; RPL13A; ATG7; NUDT9; VNN1; CTSA; HK1
8	GO:0005737	cytoplasm	PARG; UAP1; PSMB10; TCEB1; HSPA8; SERINC1; EIF5; CNBP; RPS3; PSAT1; GPD2; AACS; GPR137B; STIP1; PMM1; EXOSC7
9	GO:0044265	cellular macromolecule catabolic process	EXOSC7; SUMO2; BNIP3L; ARIH2; PSMB10; TCEB1; RPLP0; UBXN1; HSP90AB1; RPL13A; RPS3; RNF7; PINK1
10	GO:0023026	MHC class II protein complex binding	HSP90AB1; HSP90AA1; HSPA8
11	GO:0051082	unfolded protein binding	DNAJA1; PTGES3; HSPA8; HSP90AB1; HSP90AA1; NPM1
12	GO:0009056	catabolic process	PSMB10; TCEB1; RPLP0; RPS3; GPD2; PINK1; EXOSC7; ALLC; WDR45; HSP90AB1; RPL13A
13	GO:0009057	macromolecule catabolic process	EXOSC7; SUMO2; BNIP3L; ARIH2; PSMB10; TCEB1; RPLP0; AZIN1; UBXN1; HSP90AB1; RPL13A
14	GO:0044248	cellular catabolic process	PSMB10; TCEB1; SUMO2; RPS3; GPD2; PINK1; EXOSC7; ALLC; WDR45; HSP90AB1
15	GO:0006626	protein targeting to mitochondrion	TIMM17A; BNIP3L; ARIH2; PEMT; PINK1; HSP90AA1; TIMM23

S3 Table continued. Deng et al (2014) Cluster 5 (purple) top GO annotations.

	go.id	name	significant
1	GO:0044710	single-organism metabolic process	PCK2; SAT1; EPHX2; NFATC4; CKB; PRDX6; MSH2; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; POGLUT1; FABP5; AKAP12; TDGF1; FBP2; SOX2
2	GO:0006950	response to stress	EPHX2; NFATC4; PRDX6; MSH2; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; FABP5; TDGF1; SOX2
3	GO:0065010	extracellular membrane-bounded organelle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4
4	GO:0070062	extracellular exosome	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4; PRKCI; RAC2; IDH1
5	GO:0043230	extracellular organelle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4
6	GO:1903561	extracellular vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; MARCKS; DPP4; PRKCI
7	GO:0042221	response to chemical	EPHX2; NFATC4; MFGE8; PRDX6; EPHA4; PROS1; PDGFRA; PRDX1; UBE2L6; TDGF1; SOX2
8	GO:0031988	membrane-bounded vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; SPARC
9	GO:0031982	vesicle	PCK2; EPHX2; MFGE8; CKB; PRDX6; PROS1; PRDX1; POGLUT1; FABP5; FBP2; TRAP1; PLOD2; DHRS4; SPARC
10	GO:0001525	angiogenesis	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; MEIS1; SPARC; COL4A2; COL4A1; FGF10; TDGF1
11	GO:0048514	blood vessel morphogenesis	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; ZFP36L1; MEIS1; SPARC; COL4A2; COL4A1; FGF10; TDGF1
12	GO:0001944	vasculature development	SAT1; PDGFRA; BMP4; NFATC4; MFGE8; FN1; ZFP36L1; MEIS1; PDPN; SPARC; COL4A2; COL4A1; FGF10; TDGF1
13	GO:0006979	response to oxidative stress	TAT; PDGFRA; BMP4; ETV5; TRAP1; PRDX6; IDH1; PARP1; AQP8; PRDX1; CRYGD
14	GO:0009725	response to hormone	PRKCI; GJA1; PDGFRA; BMP4; MFGE8; TAT; PLOD2; SPP1; IDH1
15	GO:0030198	extracellular matrix organization	PDGFRA; BMP4; JAM2; FN1; PLOD2; SPARC; SPP1; COL4A2; COL4A1; SERPINH1; DPP4

S3 Table continued. Deng et al (2014) Cluster 6 (orange) top GO annotations.

	go.id	name	genes
1	GO:0065010	extracellular membrane-bounded organelle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8; LCP1; UGP2
2	GO:0070062	extracellular exosome	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8; LCP1; UGP2
3	GO:0043230	extracellular organelle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11
4	GO:1903561	extracellular vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8
5	GO:0031988	membrane-bounded vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; TMSB4X; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2
6	GO:0031982	vesicle	MYH10; SLC2A3; GM2A; TSPAN8; ACTG1; TMSB4X; SDC4; TINAGL1; CRYAB; MSN; FABP3; PDZK1IP1; PRSS8; S100A11; DAB2; KRT8
7	GO:0008092	cytoskeletal protein binding	MYH10; TPM4; TMSB4X; CRYAB; MSN; TMSB10; FABP3; NDRG1; CALM1; FMNL2; MYH9; CAP1; TPM1; CDH1
8	GO:0015629	actin cytoskeleton	MYH10; CLIC4; MYH9; MYL12B; WDR1; CNN2; ARPC2; AHNAK; ACTN4; CRYAB; CAP1; TPM1; DSTN; ARPC5; TPM4
9	GO:0003779	actin binding	MYH10; TPM4; WDR1; CNN2; FMNL2; ARPC2; MYH9; CAP1; TPM1
10	GO:0048468	cell development	MYH10; CAPG; ACTG1; WDR1; CNN2; FMNL2; MYH9; ACTN4; SDC4; CAP1; TPM1; DSTN
11	GO:0030036	actin cytoskeleton organization	MYH10; CAPG; ACTG1; WDR1; CNN2; FMNL2; MYH9; ACTN4; SDC4; CAP1; TPM1
12	GO:0032432	actin filament bundle	MYH10; TPM4; MYL12B; CNN2; MYH9; CRYAB; TPM1; ACTN4; LCP1
13	GO:0005912	adherens junction	TJP2; MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4
14	GO:0070161	anchoring junction	TJP2; MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4
15	GO:0005925	focal adhesion	MYH9; ACTG1; CNN2; ARPC2; AHNAK; ACTN4; SDC4; CAP1; ARPC5

S4 Table. Cluster Annotation of Deng data analysis using 48 genes with top driving gene summaries.

Cluster	Top 5 Driving Genes	Top significant GO terms (function)[q-value]
Green	*Nanog, Pdgfa, Eomes, Fgfr2, Grhl2*	GO:0048568 (embryonic organ development)[9e-08], GO:0048468 (cell development)[4e-07], GO:0001890 (placenta development)[1e-06], GO:0051094 (positive regulation of developmental process)[1e-06], GO:0030097 (hemopoiesis)[1e-05]
Purple	*Creb3l2, Klf4, Fn1, Bmp4, Lcp1*	GO:0048864 (stem cell development)[4e-12], GO:0048863 (stem cell differentiation)[2e-11], GO:0009893 (positive regulation of metabolic process)[7e-10], GO:0009653 (anatomical structure morphogenesis)[4e-08]
Orange	*Tcfap2a, Klf5, Dppa1, Dab2, Gata4*	GO:0061061 (muscle structure development)[2e-13], GO:0060537 (muscle tissue development)[2e-12], GO:0048514 (blood vessel morphogenesis)[8e-12], GO:0007275 (multocellular orgasnimal development)