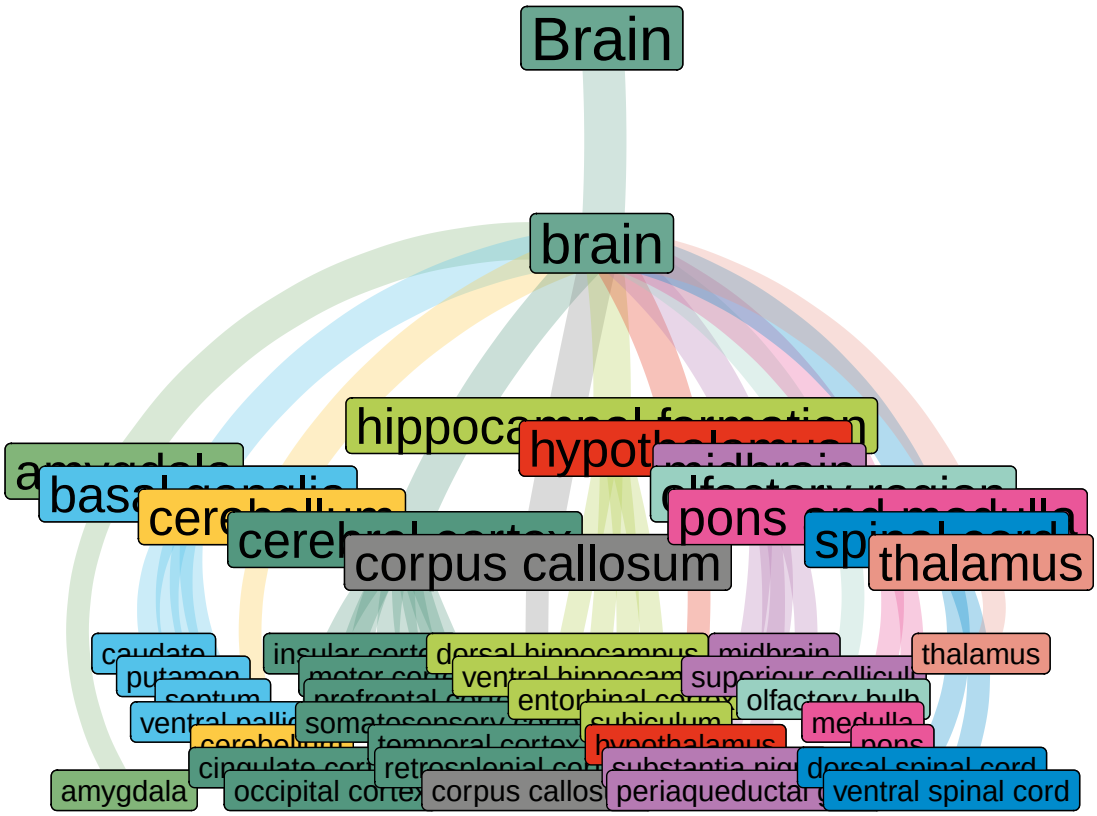
QC of pig data

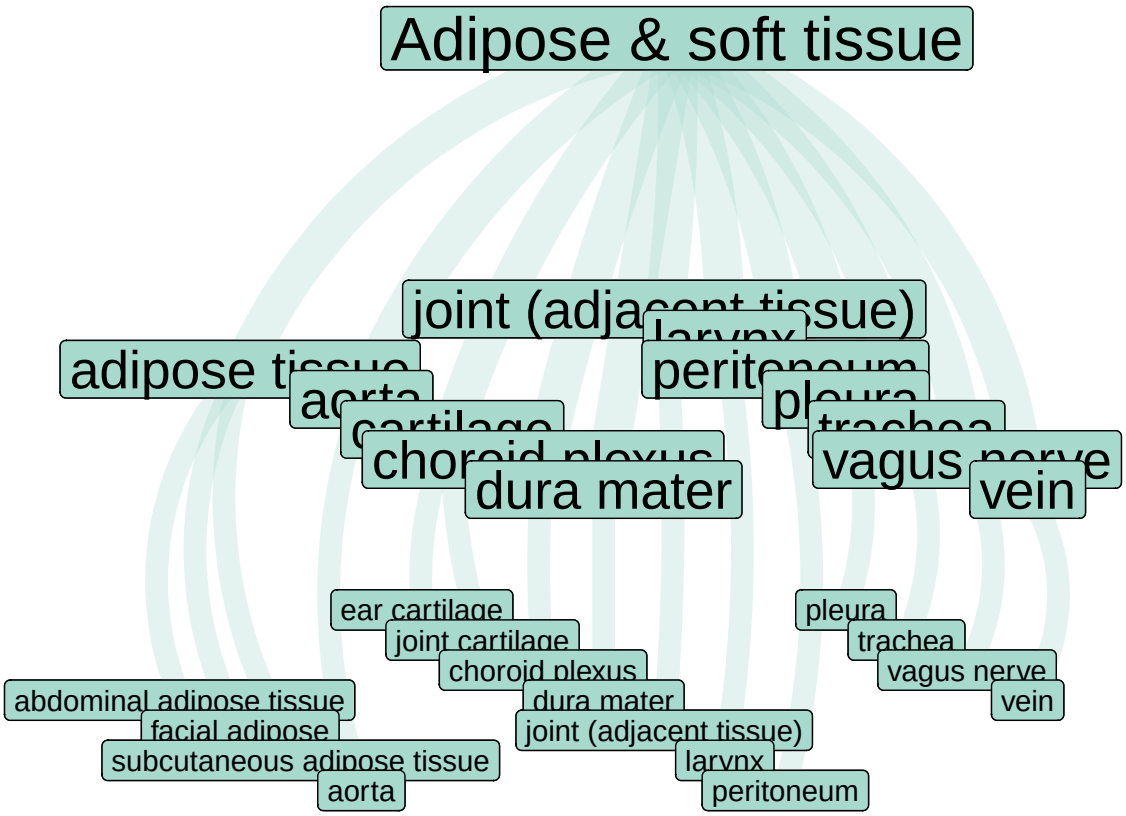
Max Jonatan Karlsson

21 October 2019

Tissue grouping hierarchy

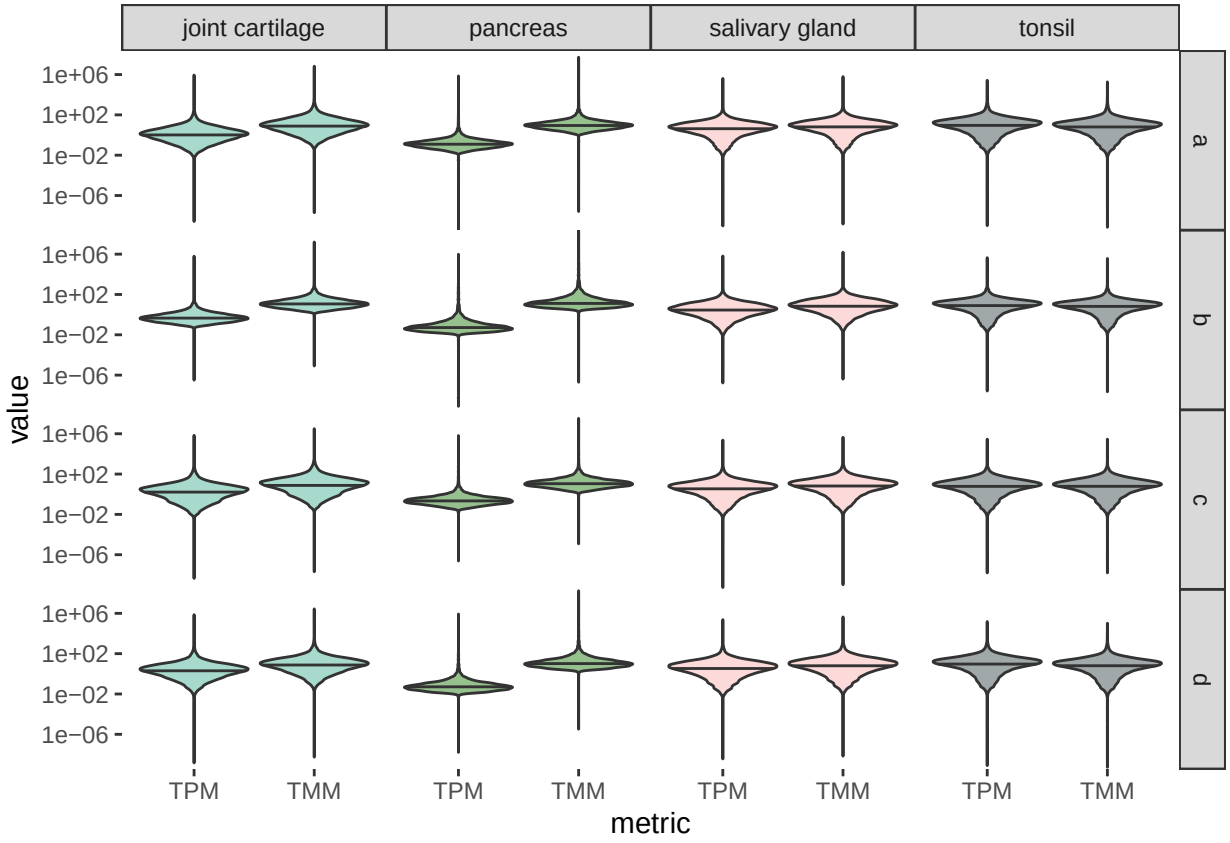
The following two plots describe the hierarchy of tissues for the Brain, and Adipose & soft tissue organs, and how they are grouped together. Similar plots for remaining organs are available in the file "Pig tissue groupings.pdf". The levels in the plot are from top to bottom as follows: Organ, Consensus tissue, Region (Only for brain), Tissue.



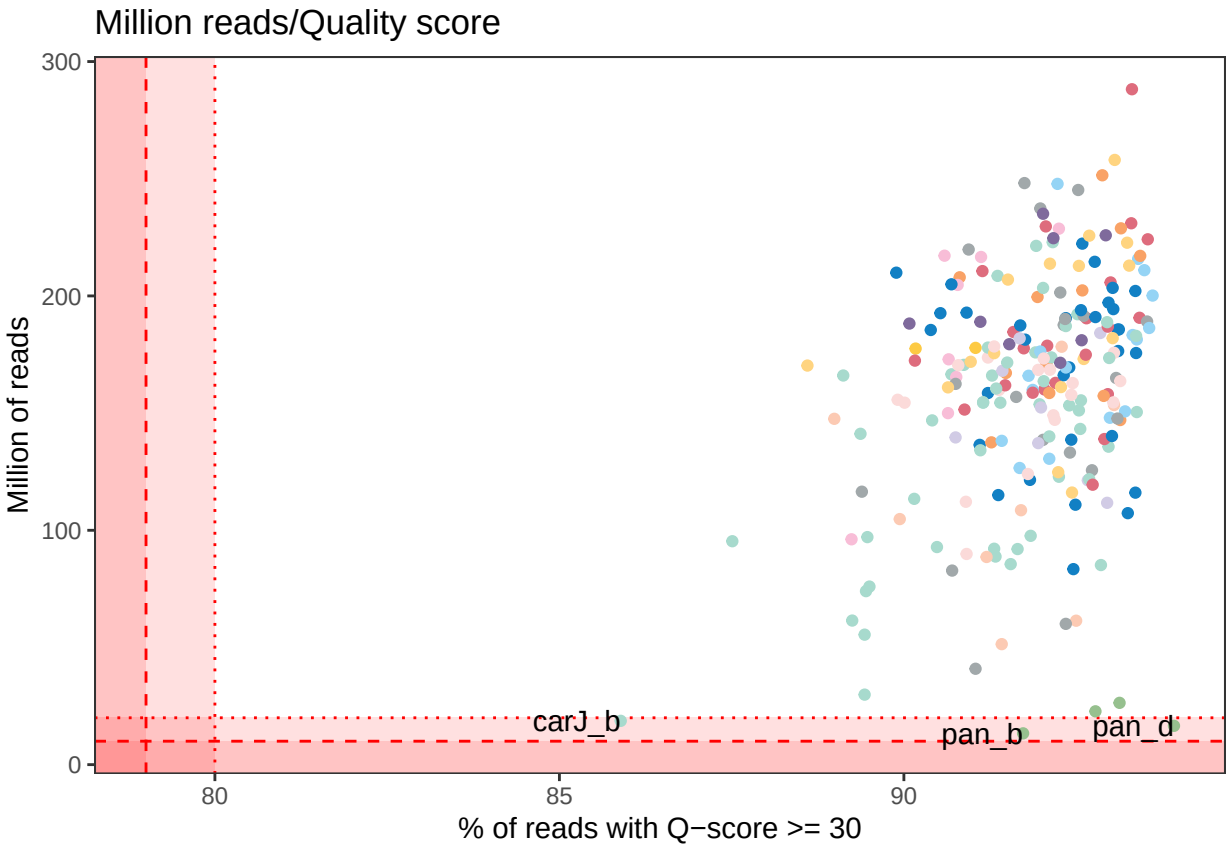

Adipose & soft tissue
joint (adjacent tissue)
larynx
adipose tissue
peritoneum
aorta
pleura
cartilage
trachea
choroid plexus
vagus nerve
dura mater
vein
ear cartilage
pleura
joint cartilage
trachea
choroid plexus
vagus nerve
abdominal adipose tissue
dura mater
vein
facial adipose
joint (adjacent tissue)
subcutaneous adipose tissue
larynx
aorta
peritoneum

Normalization

All samples were TMM normalized together with a median sample (median expression for all genes) as a reference distribution. Only protein coding genes are included. The plot below shows the distribution of expression values before and after TMM normalization for some selected tissues. Only transcripts with non-zero tpm have been included.

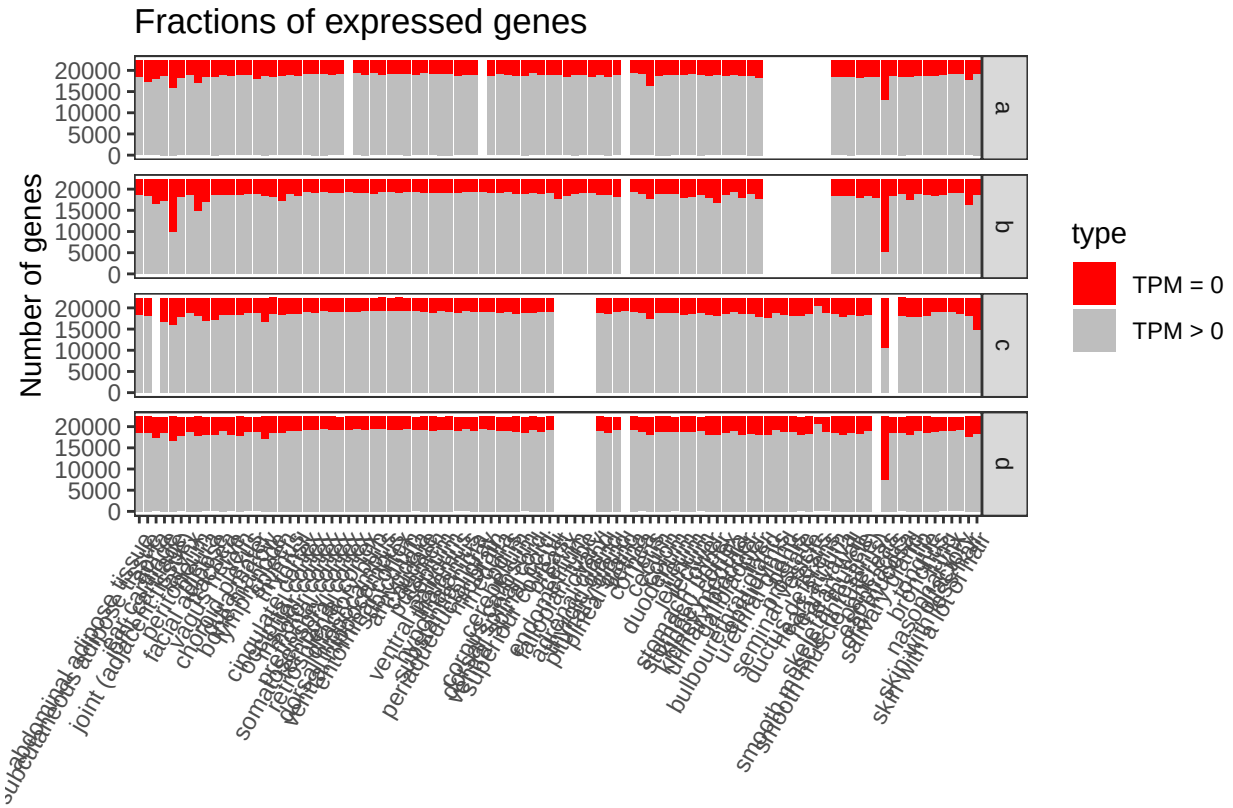


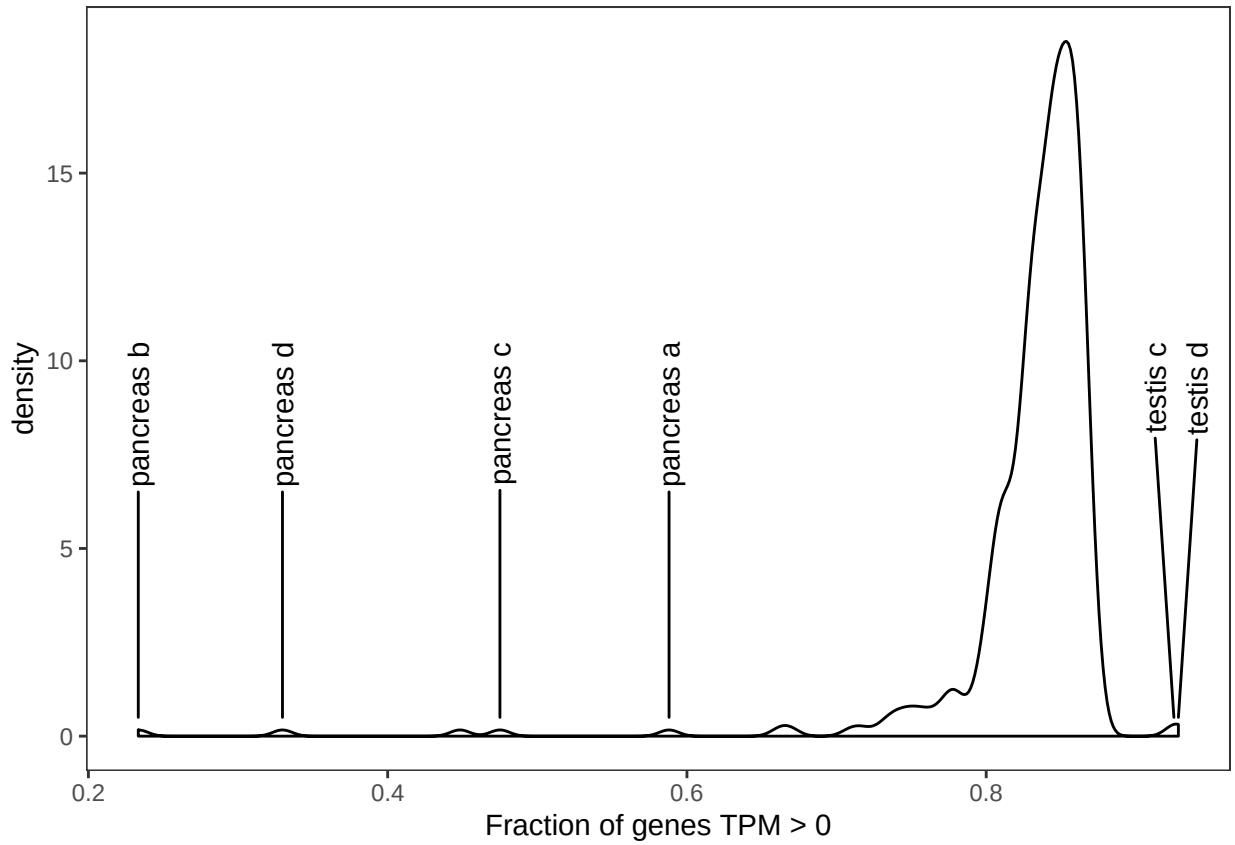
RNA and sequencing quality control

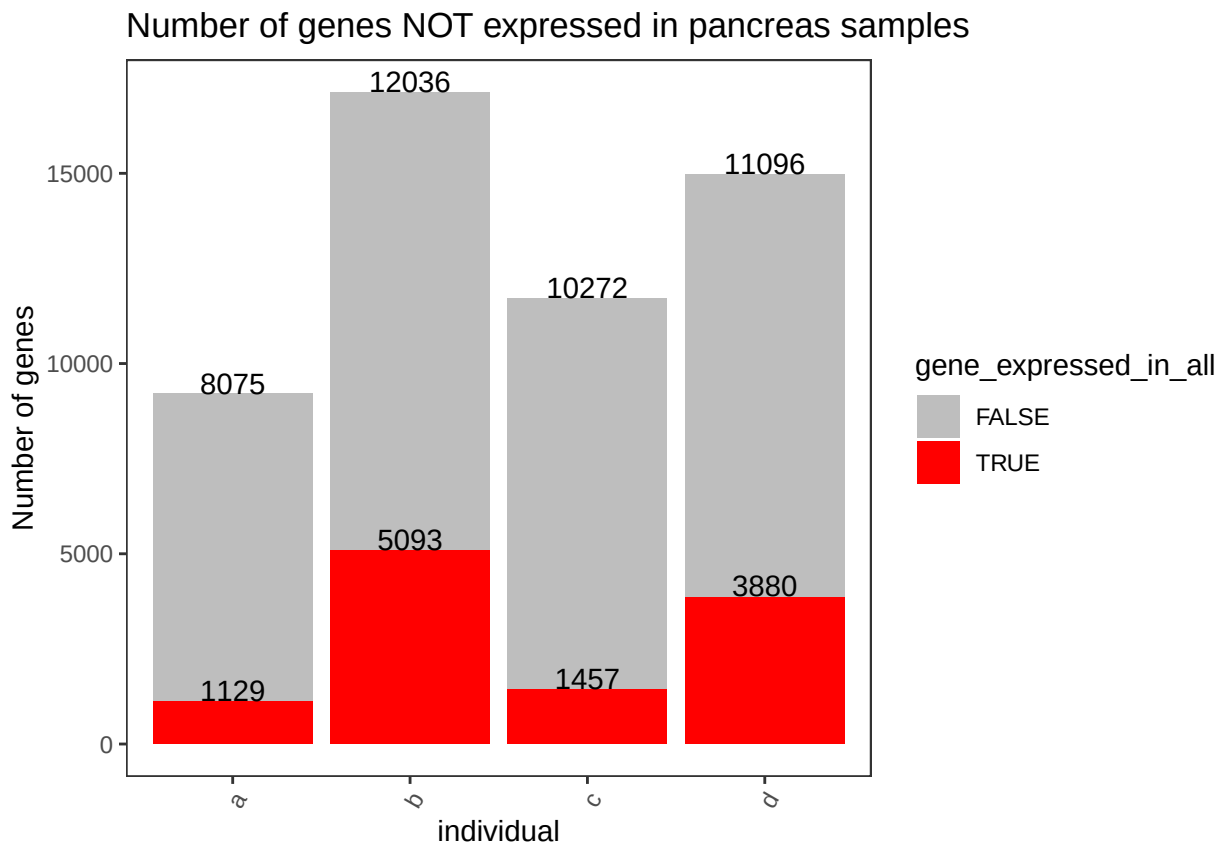


Fraction of genes with zero expression

The following plots describe the number of genes with TPM = 0 in each sample. We see that many genes are missing for pancreas samples - much more than for any other sample (except one joint adjacent tissue sample that will be removed further ahead). This could be due to autodigestion of the sample by the nucleases present in the pancreas. Pancreas samples b and d lack most genes, and will thus be removed.

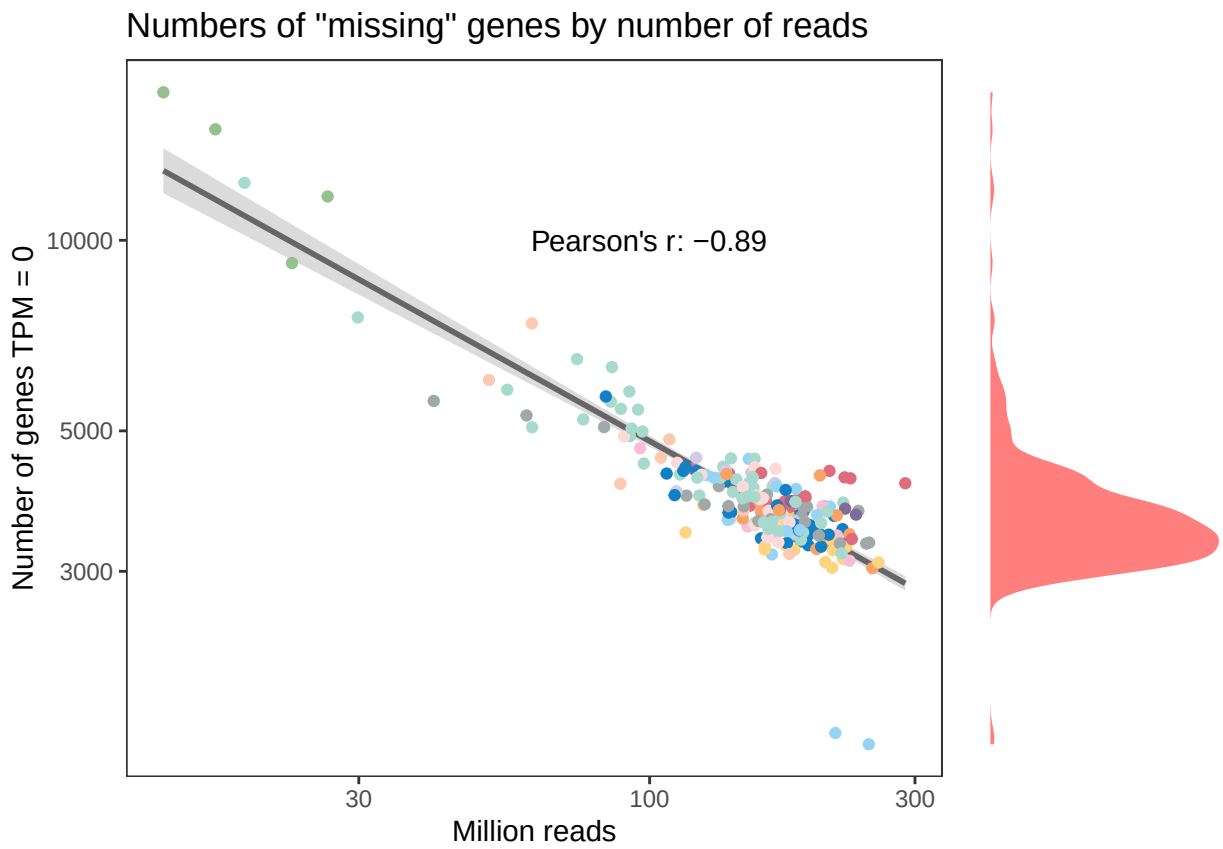

Fractions of expressed genes
Number of genes
20000
15000
10000
5000
0
a
b
c
d
type
TPM = 0
TPM > 0


density
15
10
5
0
pancreas b
pancreas d
pancreas c
pancreas a
testis c
testis d
0.2
0.4
0.6
0.8
Fraction of genes TPM > 0

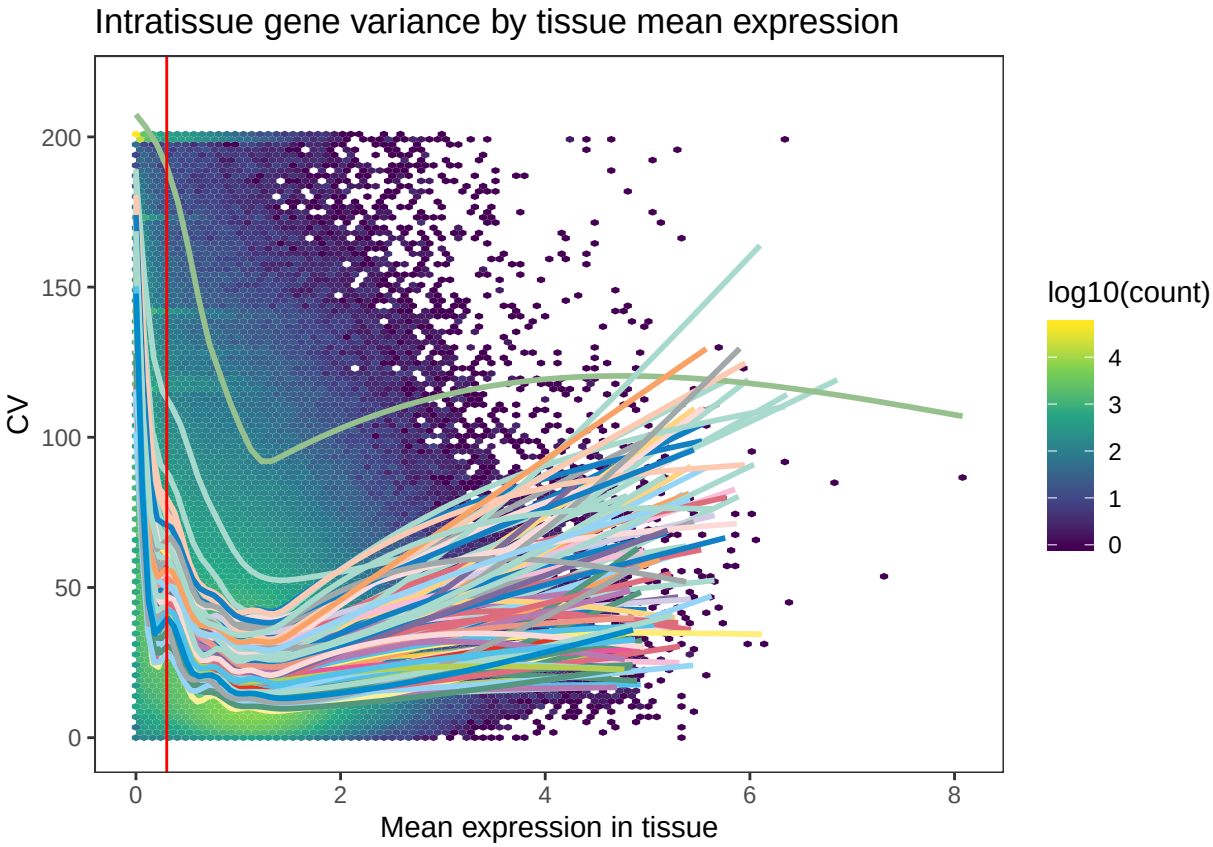


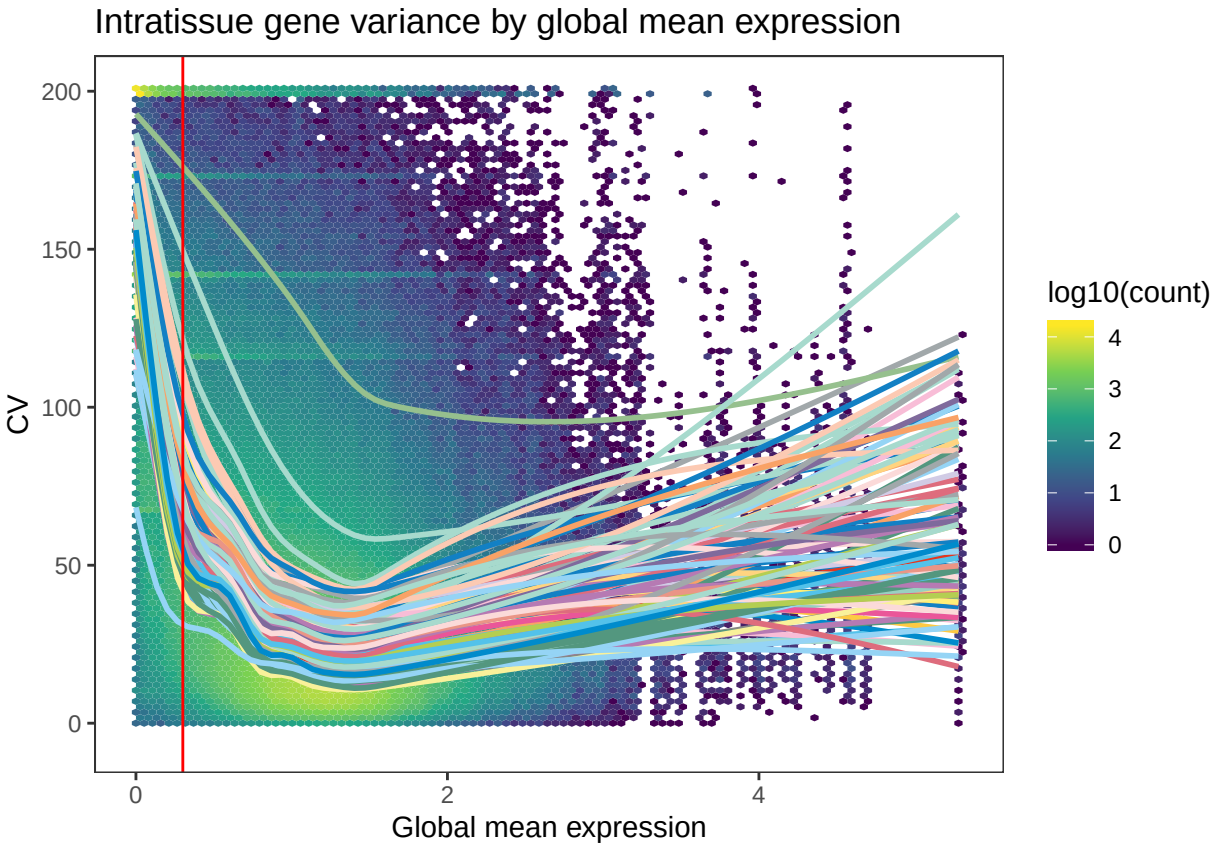
The following plots show how the number of "missing" genes in each sample is correlated with the sequencing quality metrics. We see that there is a strong correlation between number of reads and the number of genes with TPM = 0. This suggests that the issue with pancreas is simply sequencing depth.


Numbers of "missing" genes by % of reads >Q30
Pearson's r: −0.2
Number of genes TPM = 0
10000
5000
3000
90
% of reads >Q30



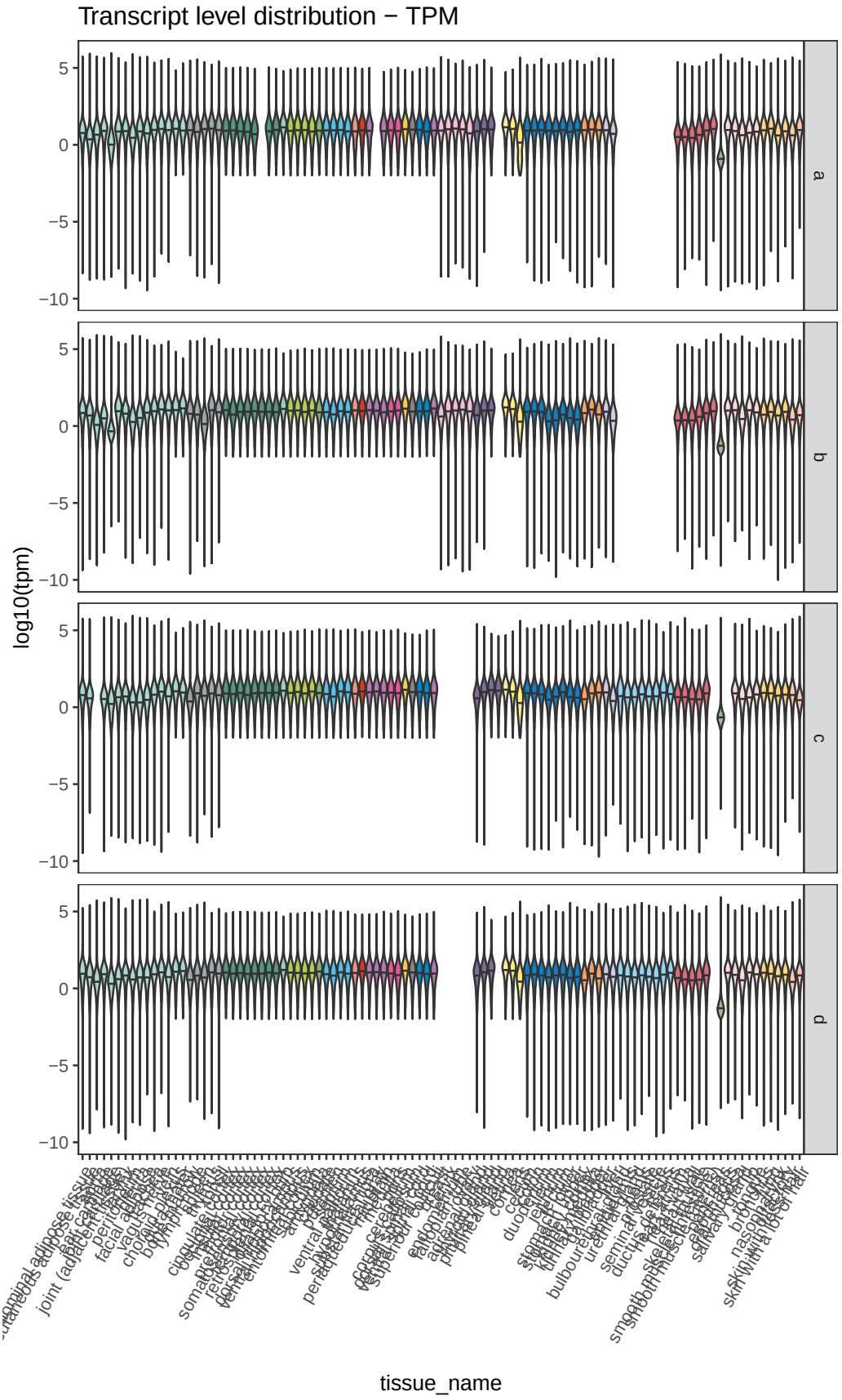
The following plots show the gene-wise variance between replicates by the genes' mean expression. We see pancreas and joint cartilage as outliers - again demonstrating that pancreas and joint cartilage b will need to be removed.

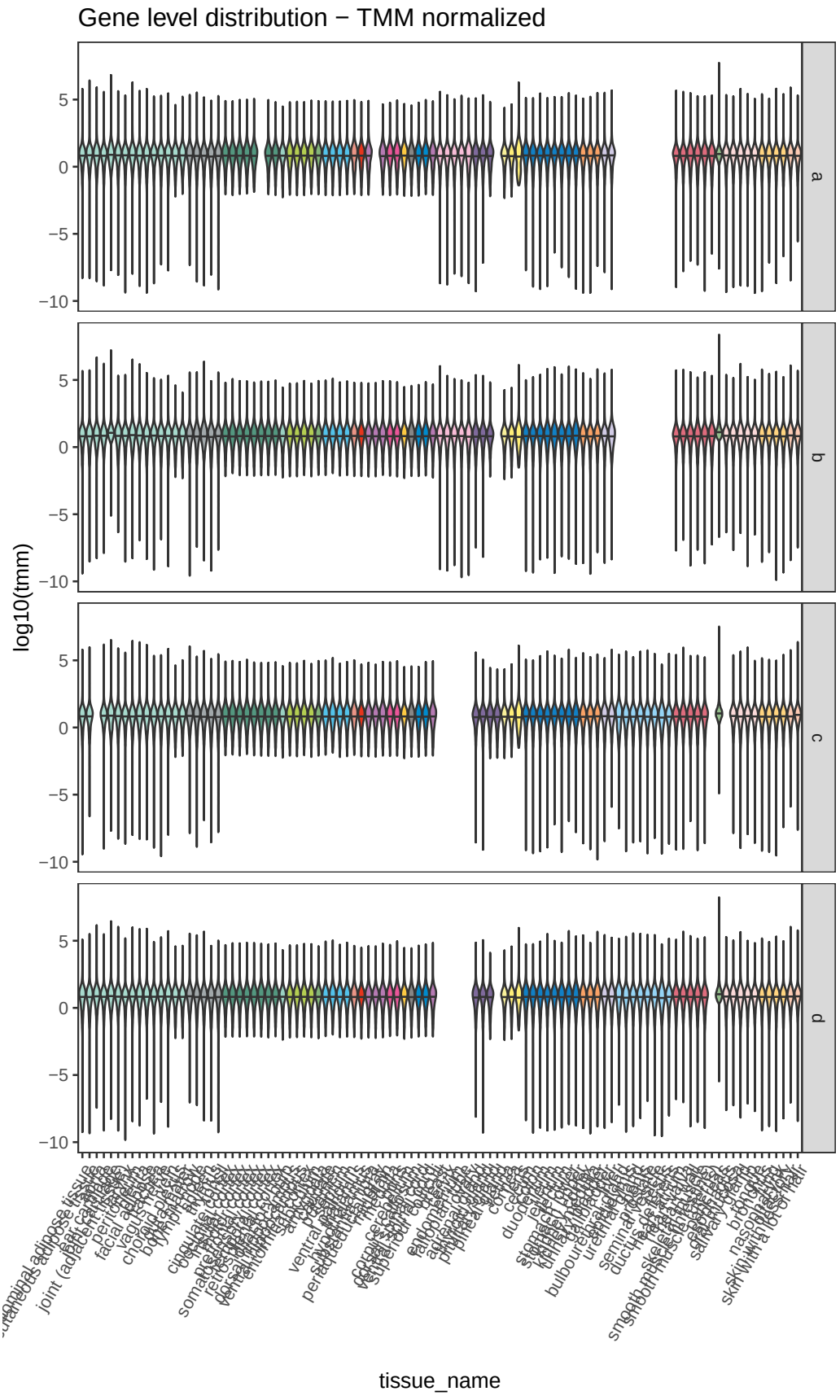

Intratissue gene variance by tissue mean expression
CV
200
150
100
50
0
0
2
4
6
8
Mean expression in tissue
log10(count)
4
3
2
1
0


Intratissue gene variance by global mean expression
CV
200
150
100
50
0
0
2
4
Global mean expression
log10(count)
4
3
2
1
0

Transcript level distribution

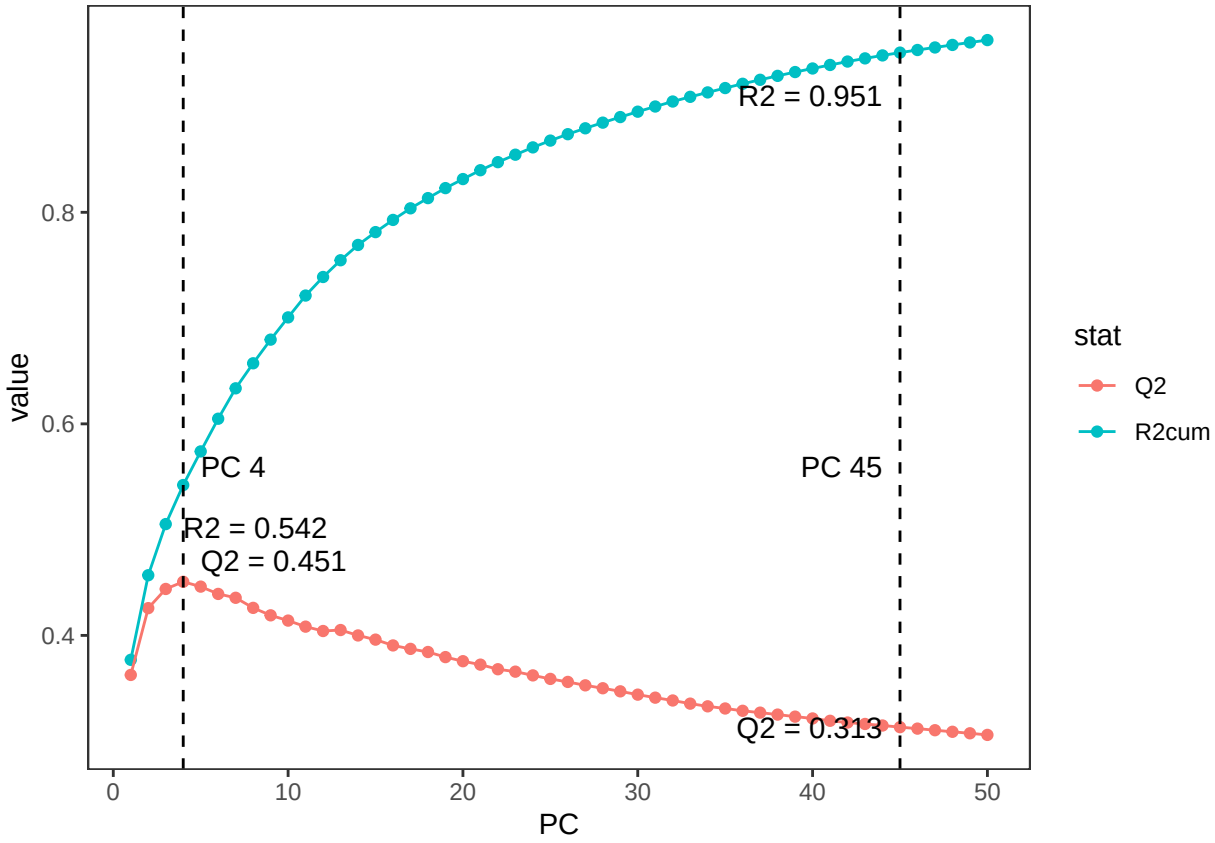
Here we see the transcript level distribution in the four individuals (a, b, c, d) first for TPM values and then TMM values. Only transcripts with non-zero tpm have been included.


Transcript level distribution − TPM
log10(tpm)
5
0
−5
−10
a
b
c
d
tissue_name

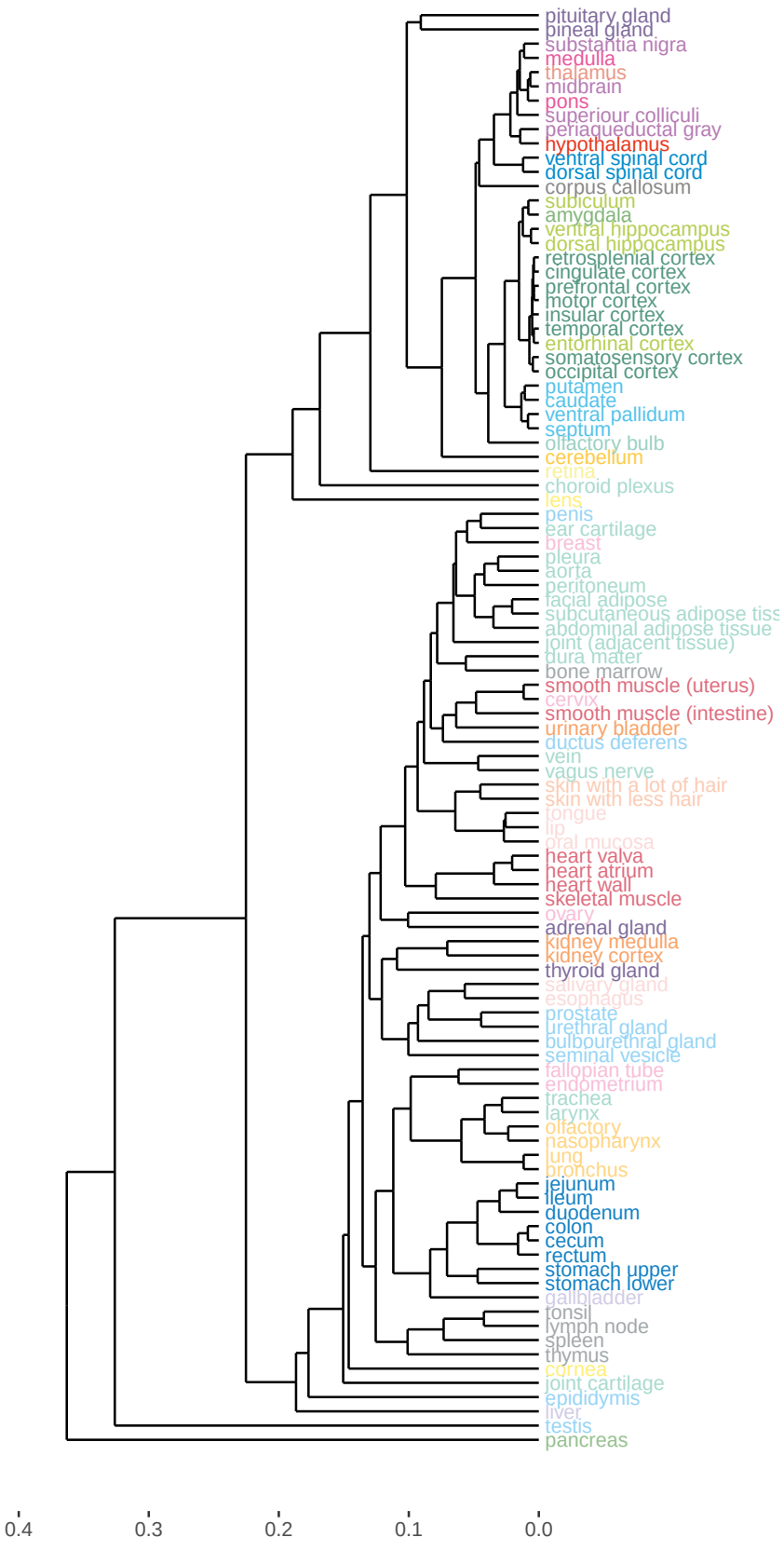

Gene level distribution − TMM normalized
log10(tmm)
5
0
−5
−10
a
b
c
d
tissue_name

Tissue wise clustering

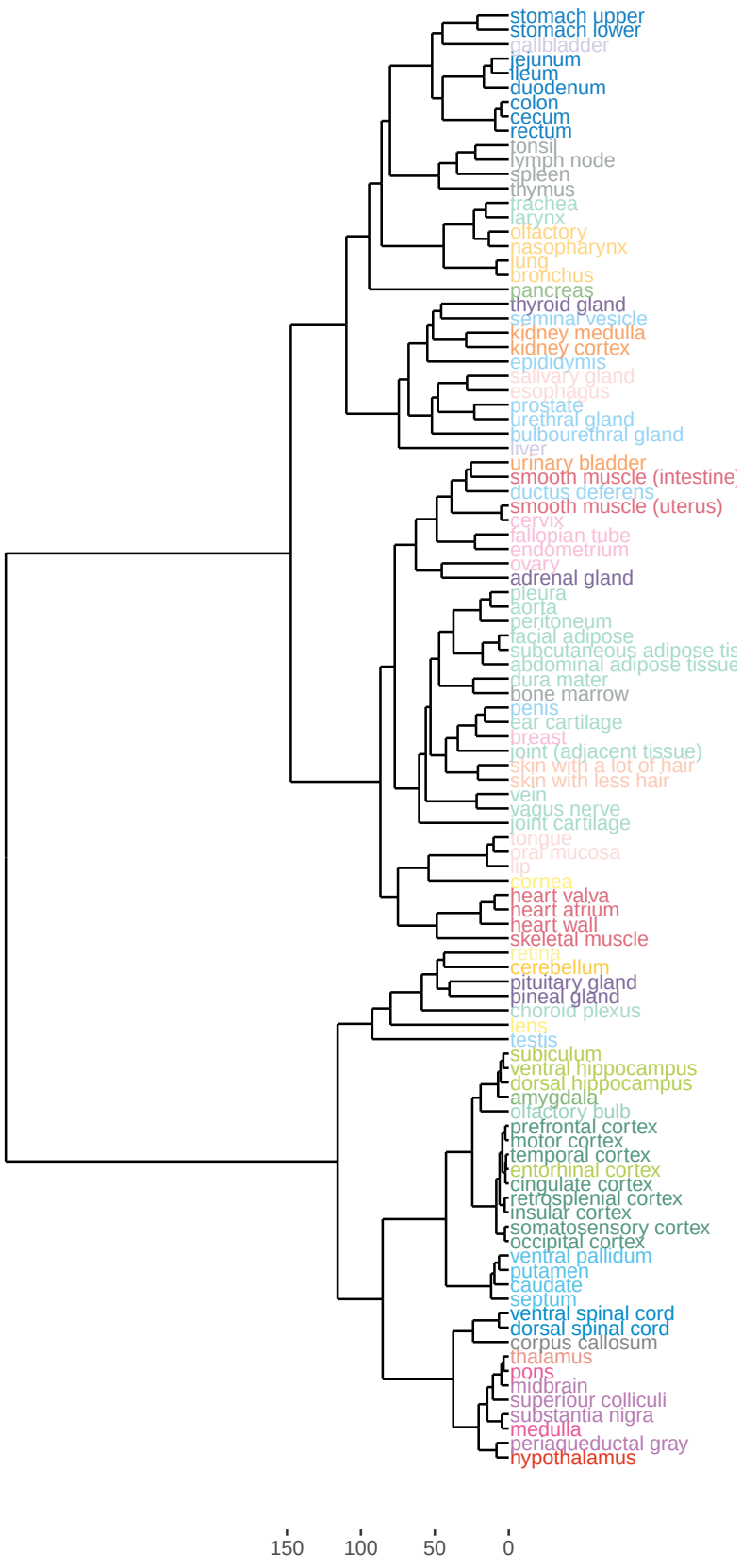
For each tissue, avarages were taken for each gene and a PCA was performed on the normalized data to reduce the number of dimensions in the data. This helps to remove noise and keep only the structures in the data that separates the features of our tissues from each other. Below, we see a plot showing R2 (The fraction of explained variance) and Q2 (crossvalidated R2, can approximately be explained as the amount of information the PCA model contains). We see that 95% R2 is reached at component 39, and we will thus choose 39 components as our dimensionally reduced data. We choose this arbitrary cutoff as we want to remove noise but still keep a majority of the structure in the data.



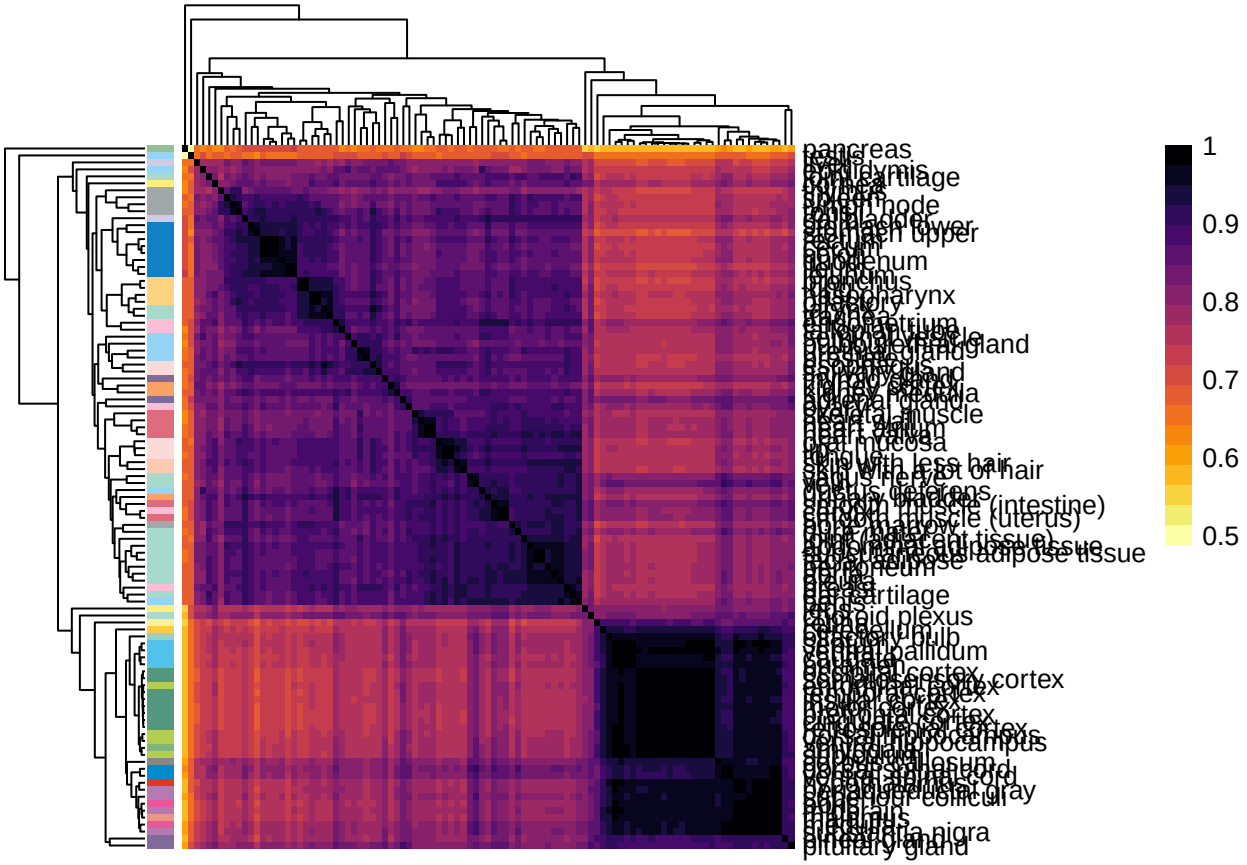
Below we see a dendrogram built from 1 - Spearman's rho between the tissues using all genes.


pituitary gland
pineal gland
substantia nigra
medulla
thalamus
midbrain
pons
superiour colliculi
periaqueductal gray
hypothalamus
ventral spinal cord
dorsal spinal cord
corpus callosum
subiculum
amygdala
ventral hippocampus
dorsal hippocampus
retrosplenial cortex
cingulate cortex
prefrontal cortex
motor cortex
insular cortex
temporal cortex
entorhinal cortex
somatosensory cortex
occipital cortex
putamen
caudate
ventral pallidum
septum
olfactory bulb
cerebellum
retina
choroid plexus
lens
penis
ear cartilage
breast
pleura
aorta
peritoneum
facial adipose
subcutaneous adipose tiss
abdominal adipose tissue
joint (adjacent tissue)
dura mater
bone marrow
smooth muscle (uterus)
cervix
smooth muscle (intestine)
urinary bladder
ductus deferens
vein
vagus nerve
skin with a lot of hair
skin with less hair
tongue
lip
oral mucosa
heart valva
heart atrium
heart wall
skeletal muscle
ovary
adrenal gland
kidney medulla
kidney cortex
thyroid gland
salivary gland
esophagus
prostate
urethral gland
bulbourethral gland
seminal vesicle
fallopian tube
endometrium
trachea
larynx
olfactory
nasopharynx
lung
bronchus
jejunum
ileum
duodenum
colon
cecum
rectum
stomach upper
stomach lower
gallbladder
tonsil
lymph node
spleen
thymus
cornea
joint cartilage
epididymis
liver
testis
pancreas
0.4
0.3
0.2
0.1
0.0

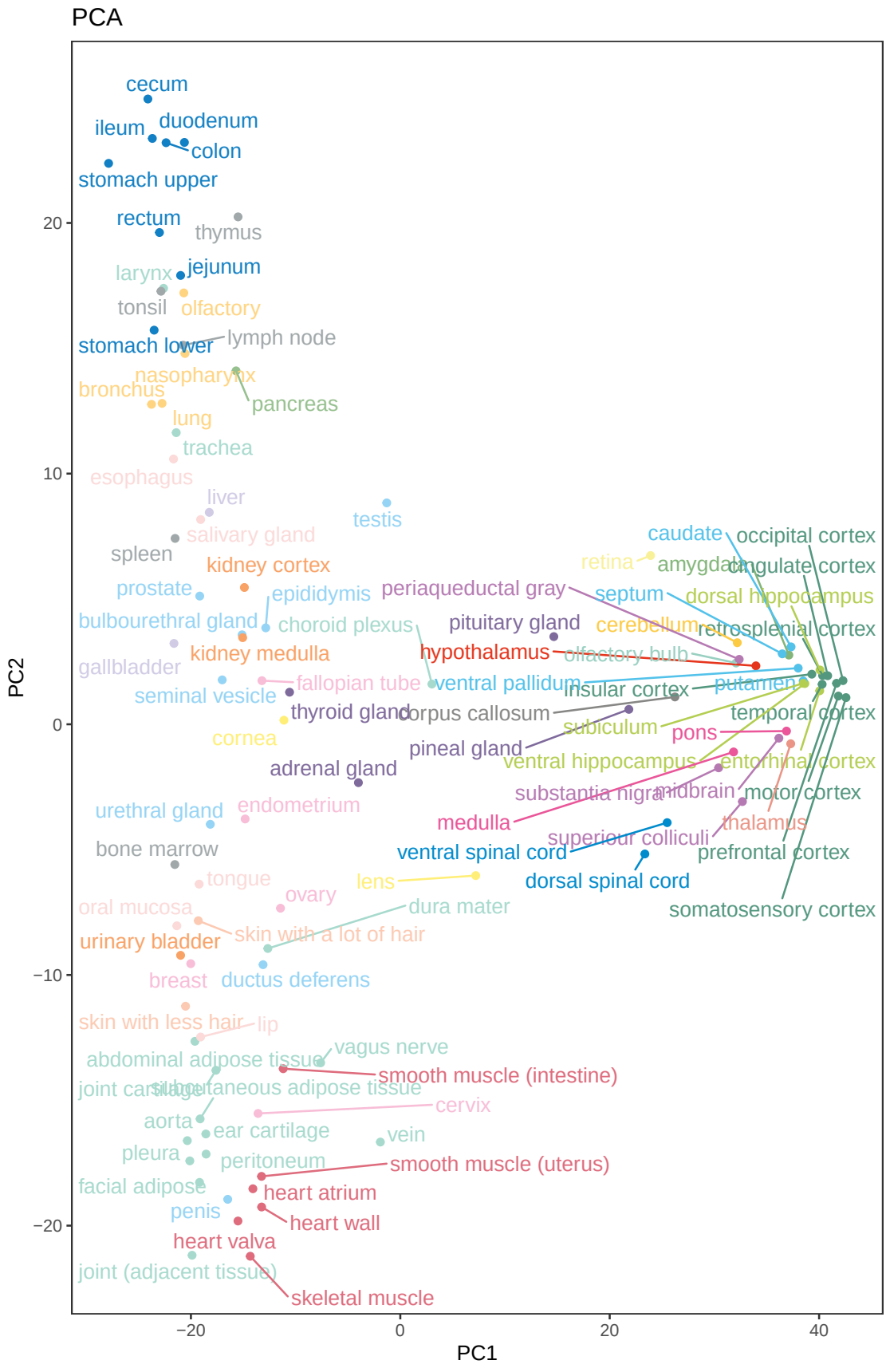
Below is an alternative (and complementary) clustering of tissues using Ward clustering of PCA scores in the 39 selected components. This type of clustering is based on minimizing the variance within each cluster and thus takes distances into account, which correlation does not.


stomach upper
stomach lower
gallbladder
jejunum
ileum
duodenum
colon
cecum
rectum
tonsil
lymph node
spleen
thymus
trachea
larynx
olfactory
nasopharynx
lung
bronchus
pancreas
thyroid gland
seminal vesicle
kidney medulla
kidney cortex
epididymis
salivary gland
esophagus
prostate
urethral gland
bulbourethral gland
liver
urinary bladder
smooth muscle (intestine)
ductus deferens
smooth muscle (uterus)
cervix
fallopian tube
endometrium
ovary
adrenal gland
pleura
aorta
peritoneum
facial adipose
subcutaneous adipose tis
abdominal adipose tissue
dura mater
bone marrow
penis
ear cartilage
breast
joint (adjacent tissue)
skin with a lot of hair
skin with less hair
vein
vagus nerve
joint cartilage
tongue
oral mucosa
lip
cornea
heart valva
heart atrium
heart wall
skeletal muscle
retina
cerebellum
pituitary gland
pineal gland
choroid plexus
lens
testis
subiculum
ventral hippocampus
dorsal hippocampus
amygdala
olfactory bulb
prefrontal cortex
motor cortex
temporal cortex
entorhinal cortex
cingulate cortex
retrosplenial cortex
insular cortex
somatosensory cortex
occipital cortex
ventral pallidum
putamen
caudate
septum
ventral spinal cord
dorsal spinal cord
corpus callosum
thalamus
pons
midbrain
superiour colliculi
substantia nigra
medulla
periaqueductal gray
hypothalamus
150
100
50
0

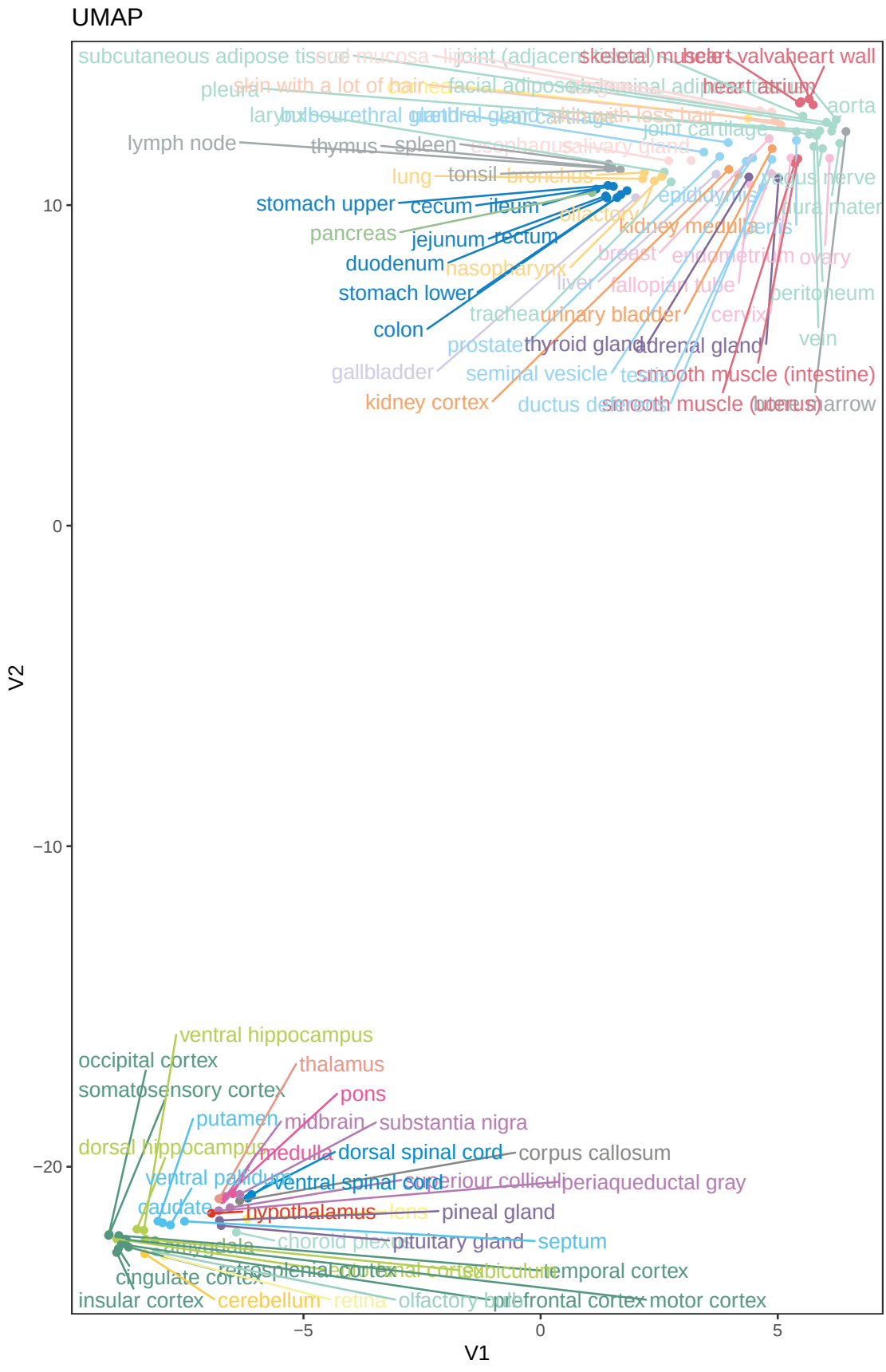
The following heatmap shows spearman correlation between tissues.



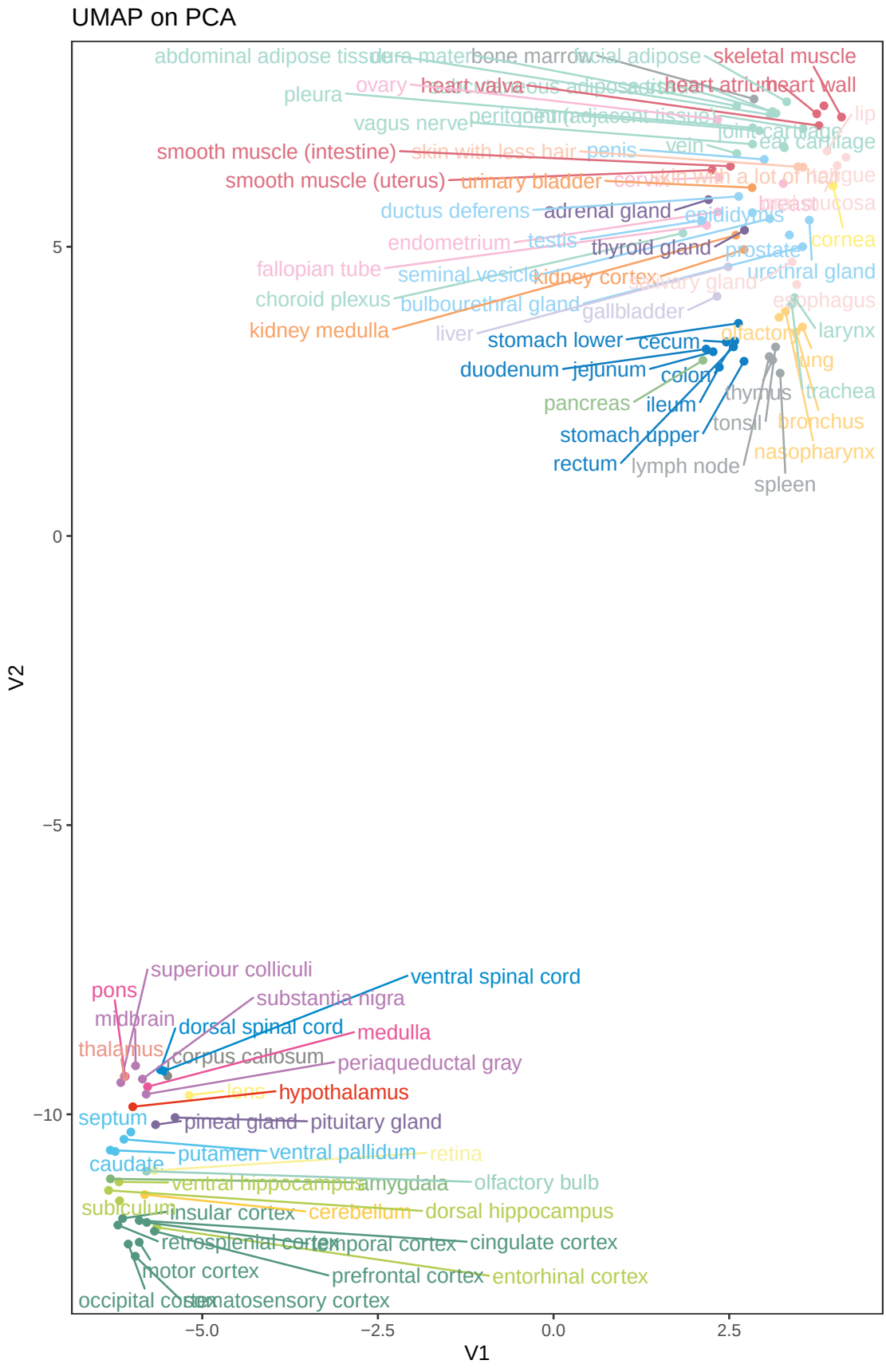
Below we see three plots showing the twodimensional projections of the data using three different strategies:
1. PCA


PCA
cecum
ileum
duodenum
colon
stomach upper
rectum
thymus
larynx
jejunum
tonsil
olfactory
stomach lower
lymph node
nasopharynx
bronchus
pancreas
lung
trachea
esophagus
liver
testis
salivary gland
spleen
kidney cortex
prostate
epididymis
periaqueductal gray
bulbourethral gland
choroid plexus
pituitary gland
kidney medulla
hypothalamus
gallbladder
fallopian tube
ventral pallidum
seminal vesicle
thyroid gland
corpus callosum
cornea
pineal gland
adrenal gland
urethral gland
endometrium
medulla
bone marrow
ventral spinal cord
tongue
lens
dorsal spinal cord
ovary
oral mucosa
dura mater
urinary bladder
skin with a lot of hair
breast
ductus deferens
skin with less hair
lip
vagus nerve
abdominal adipose tissue
smooth muscle (intestine)
joint cartilage
subcutaneous adipose tissue
cervix
aorta
ear cartilage
vein
pleura
peritoneum
smooth muscle (uterus)
facial adipose
heart atrium
penis
heart wall
heart valva
joint (adjacent tissue)
skeletal muscle
caudate
occipital cortex
retina
amygdala
cingulate cortex
septum
dorsal hippocampus
cerebellum
retrosplenial cortex
olfactory bulb
insular cortex
putamen
temporal cortex
subiculum
pons
ventral hippocampus
entorhinal cortex
substantia nigra
midbrain
motor cortex
thalamus
superiour colliculi
prefrontal cortex
somatosensory cortex
20
10
0
−10
−20
PC2
−20
0
20
40
PC1

2. UMAP

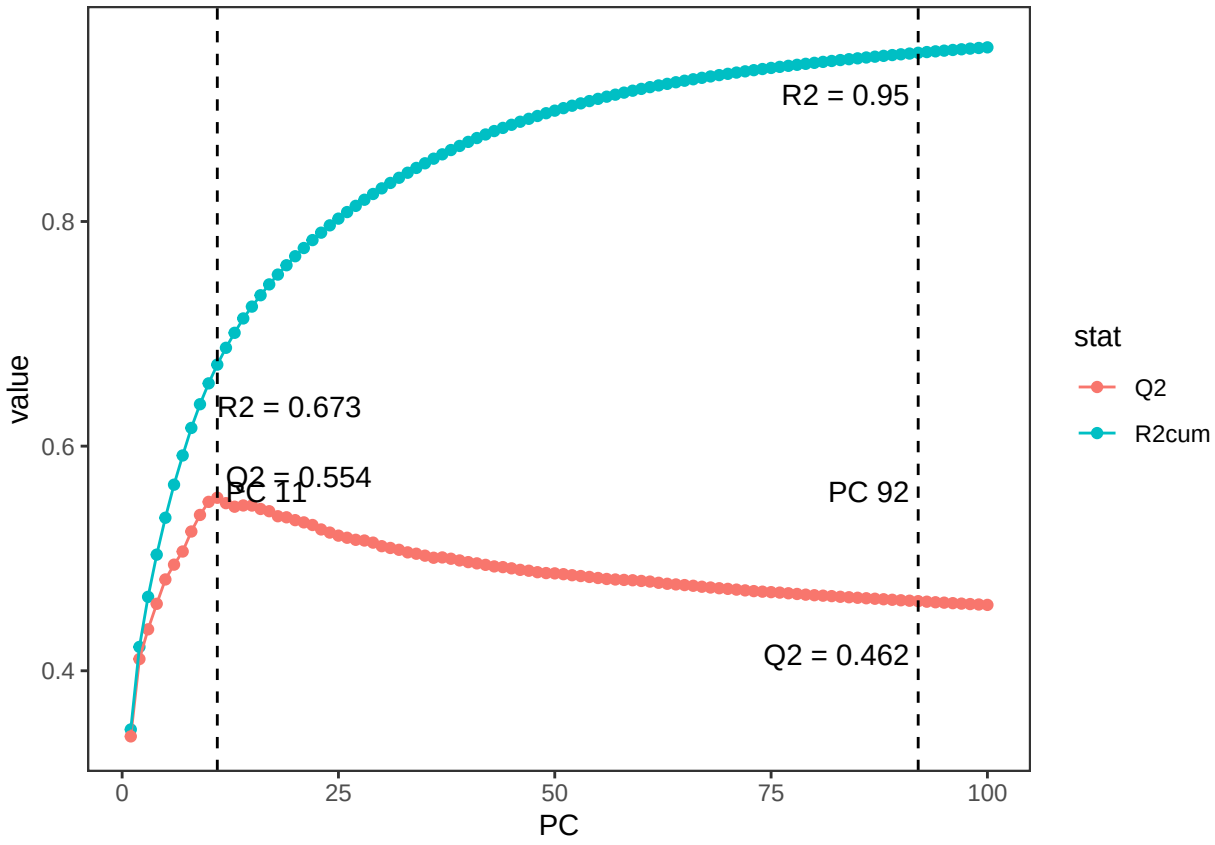

UMAP
V2
V1
10
0
−10
−20
−5
0
5
subcutaneous adipose tissue
skeletal muscle
heart valvaheart wall
pleura
facial adipose tissue
heart atrium
aorta
joint cartilage
lymph node
thymus
spleen
salivary gland
lung
tonsil
bronchus
vagus nerve
stomach upper
cecum
ileum
epididymis
dura mater
pancreas
jejunum
rectum
kidney medulla
duodenum
nasopharynx
breast
endometrium
ovary
stomach lower
liver
fallopian tube
peritoneum
trachea
urinary bladder
cervix
colon
prostate
thyroid gland
adrenal gland
vein
gallbladder
seminal vesicle
testis
smooth muscle (intestine)
kidney cortex
ductus deferens
smooth muscle (uterus)
bone marrow
ventral hippocampus
occipital cortex
thalamus
somatosensory cortex
pons
putamen
midbrain
substantia nigra
dorsal hippocampus
medulla
dorsal spinal cord
corpus callosum
ventral pallidum
ventral spinal cord
periaqueductal gray
caudate
hypothalamus
lens
pineal gland
choroid plexus
pituitary gland
septum
amygdala
cingulate cortex
temporal cortex
insular cortex
cerebellum
retina
olfactory bulb
frontal cortex
motor cortex

3. UMAP on 39 PCs from PCA

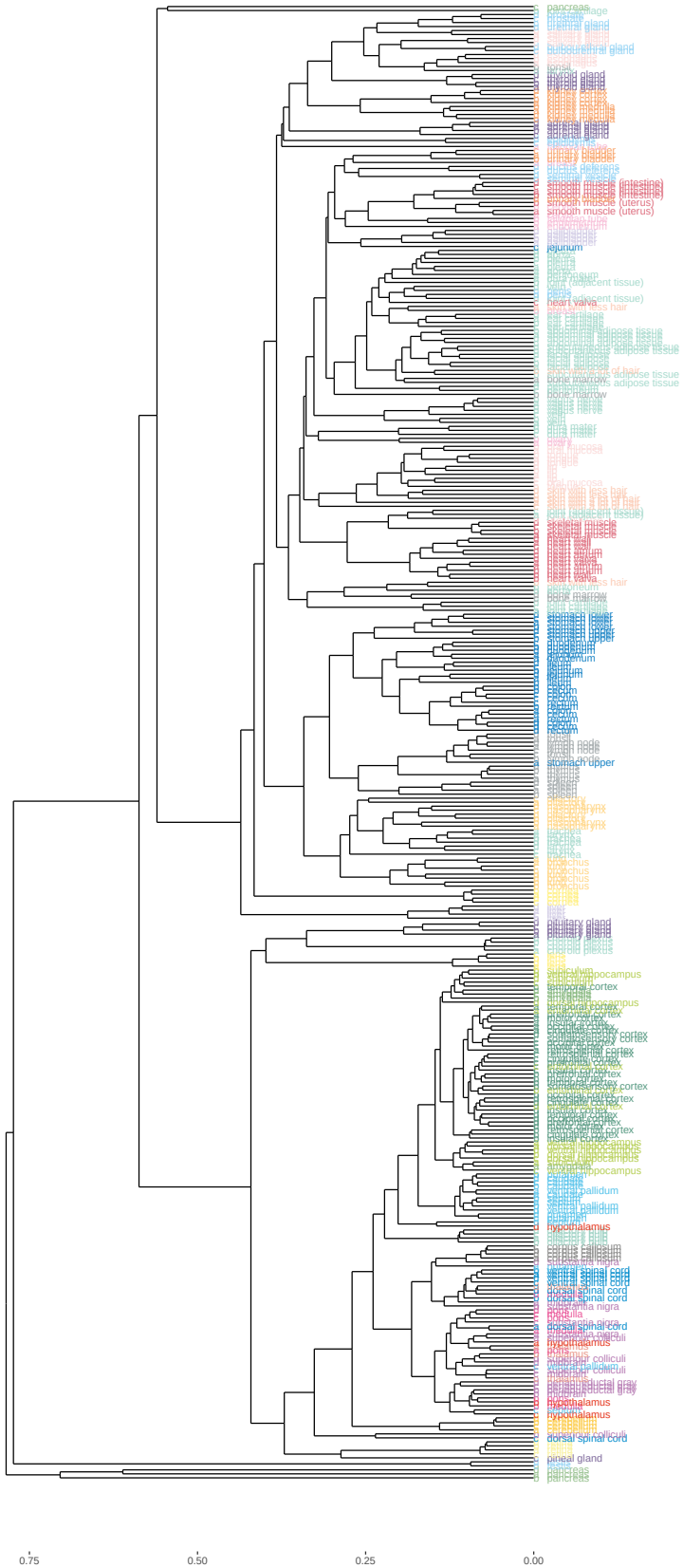

UMAP on PCA
abdominal adipose tissue
dura mater
bone marrow
facial adipose
skeletal muscle
pleura
ovary
heart valva
heart atrium
heart wall
vagus nerve
peritoneum
lip
joint cartilage
ear cartilage
smooth muscle (intestine)
skin with less hair
penis
vein
smooth muscle (uterus)
urinary bladder
cervix
skin with a lot of hair
tongue
ductus deferens
adrenal gland
epididymis
breast
endometrium
testis
thyroid gland
prostate
cornea
fallopian tube
seminal vesicle
kidney cortex
salivary gland
urethral gland
choroid plexus
bulbourethral gland
gallbladder
esophagus
kidney medulla
liver
stomach lower
cecum
olfactory
larynx
duodenum
jejunum
colon
lung
pancreas
ileum
thymus
trachea
tonsil
bronchus
stomach upper
rectum
lymph node
nasopharynx
spleen
superiour colliculi
ventral spinal cord
pons
substantia nigra
midbrain
dorsal spinal cord
medulla
thalamus
corpus callosum
periaqueductal gray
lens
hypothalamus
septum
pineal gland
pituitary gland
putamen
ventral pallidum
retina
caudate
ventral hippocampus
amygdala
olfactory bulb
subiculum
insular cortex
cerebellum
dorsal hippocampus
retrosplenial cortex
temporal cortex
cingulate cortex
motor cortex
prefrontal cortex
entorhinal cortex
occipital cortex
somatosensory cortex
5
0
−5
−10
V2
−5.0
−2.5
0.0
2.5
V1

Sample wise clustering

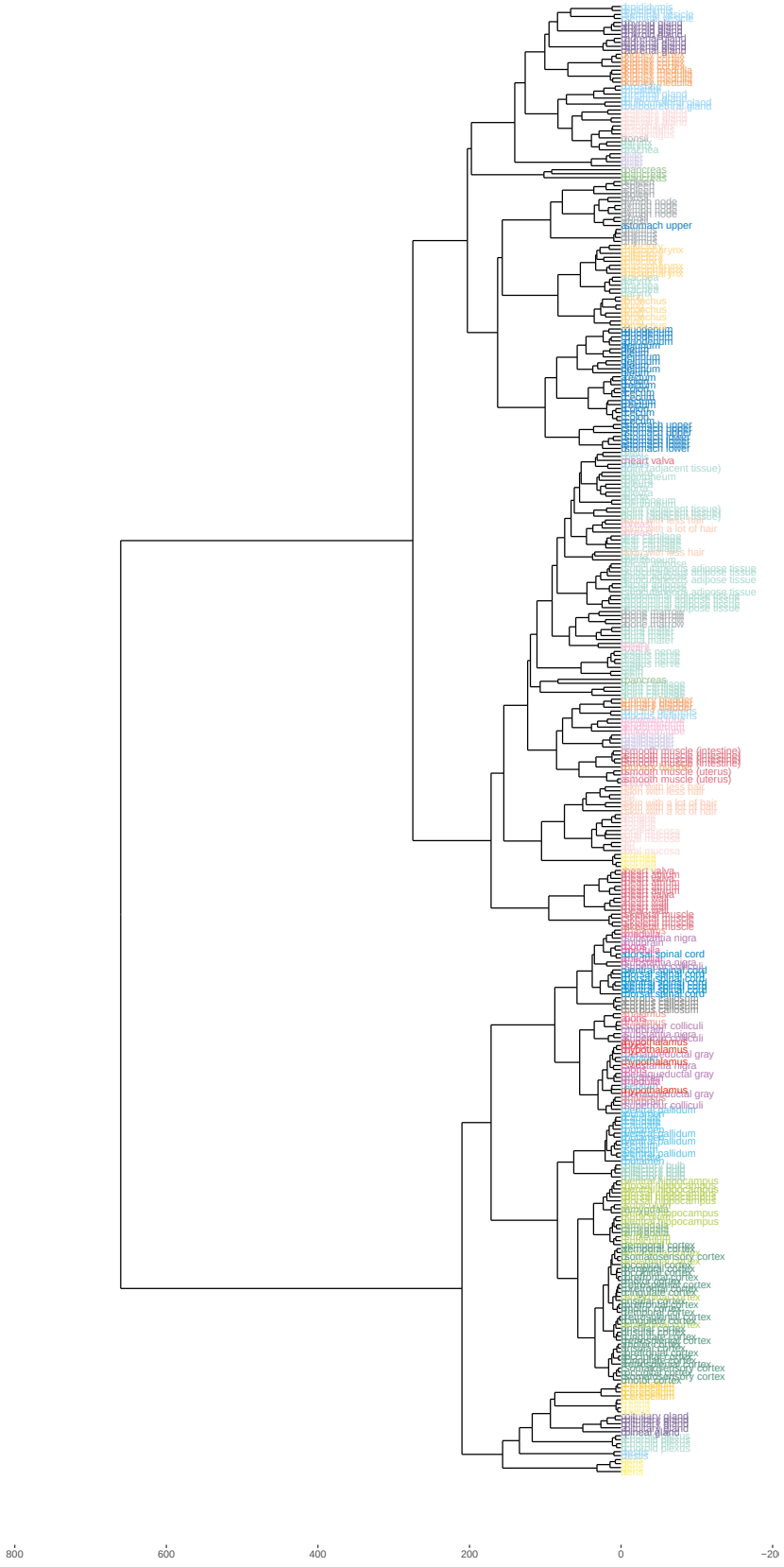
To investigate if we have any outliers or sample mixups we again perform a PCA, but this time on the full data. That is, we do not perform any avaraging between samples of the same tissue. We see that 92 PCs represent 95% of the variance.



Hierarchical clustering of individual samples. Individuals are displayed to the left and the tissue name to the right for each sample. Below we see a dendrogram built from 1 - Spearman's rho between the tissues using all genes.


0.75
0.50
0.25
0.00

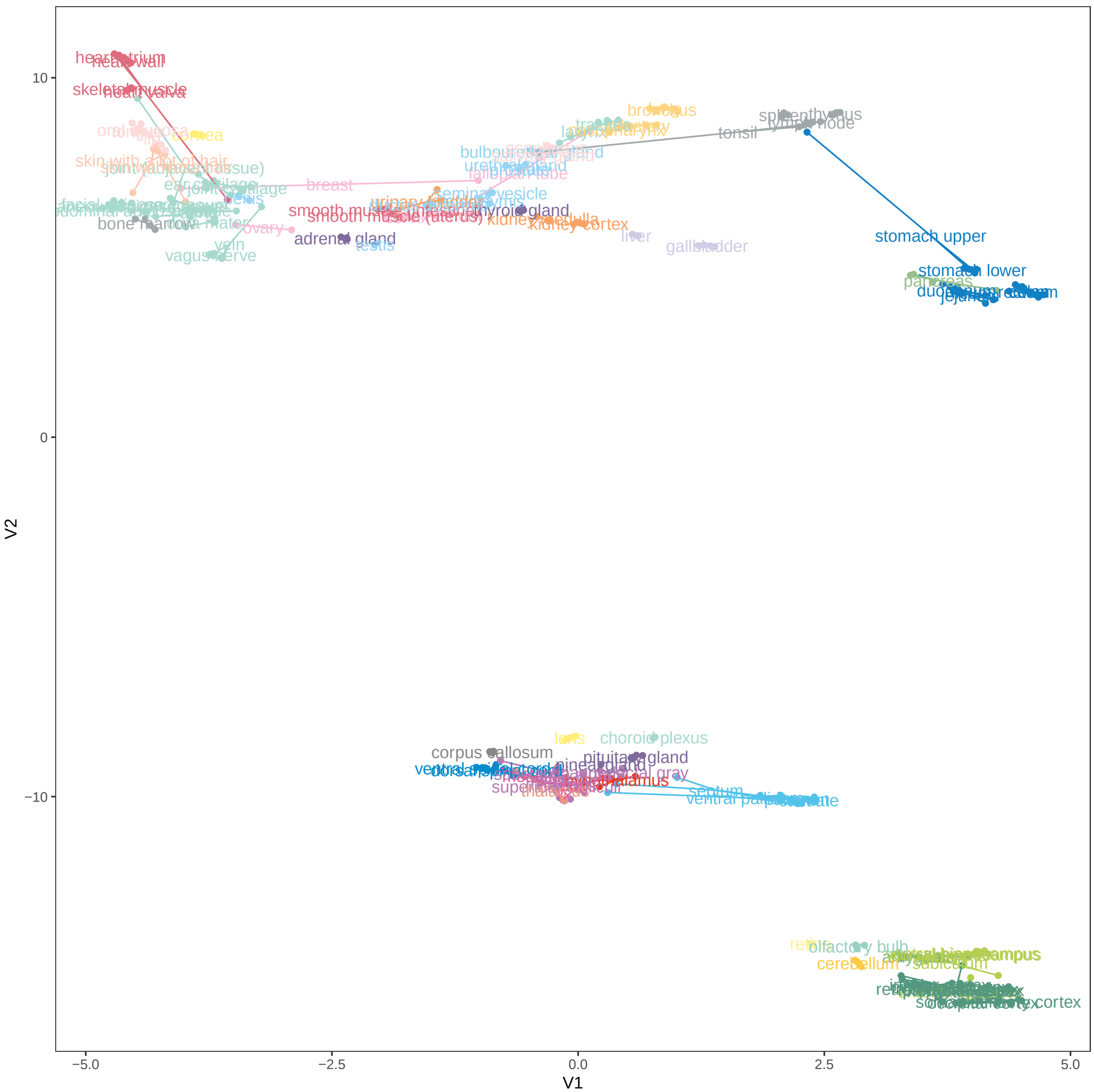
Below is an alternative (and complementary) clustering of tissues using Ward clustering of PCA scores in the 92 selected components. This type of clustering is based on minimizing the variance within each cluster and thus takes distances into account, which correlation does not.


800
600
400
200
0
−20

Heatmap displaying the clustering as well as the spearman correlation of individual is available in the file "Spearman heatmap dist clust.pdf".

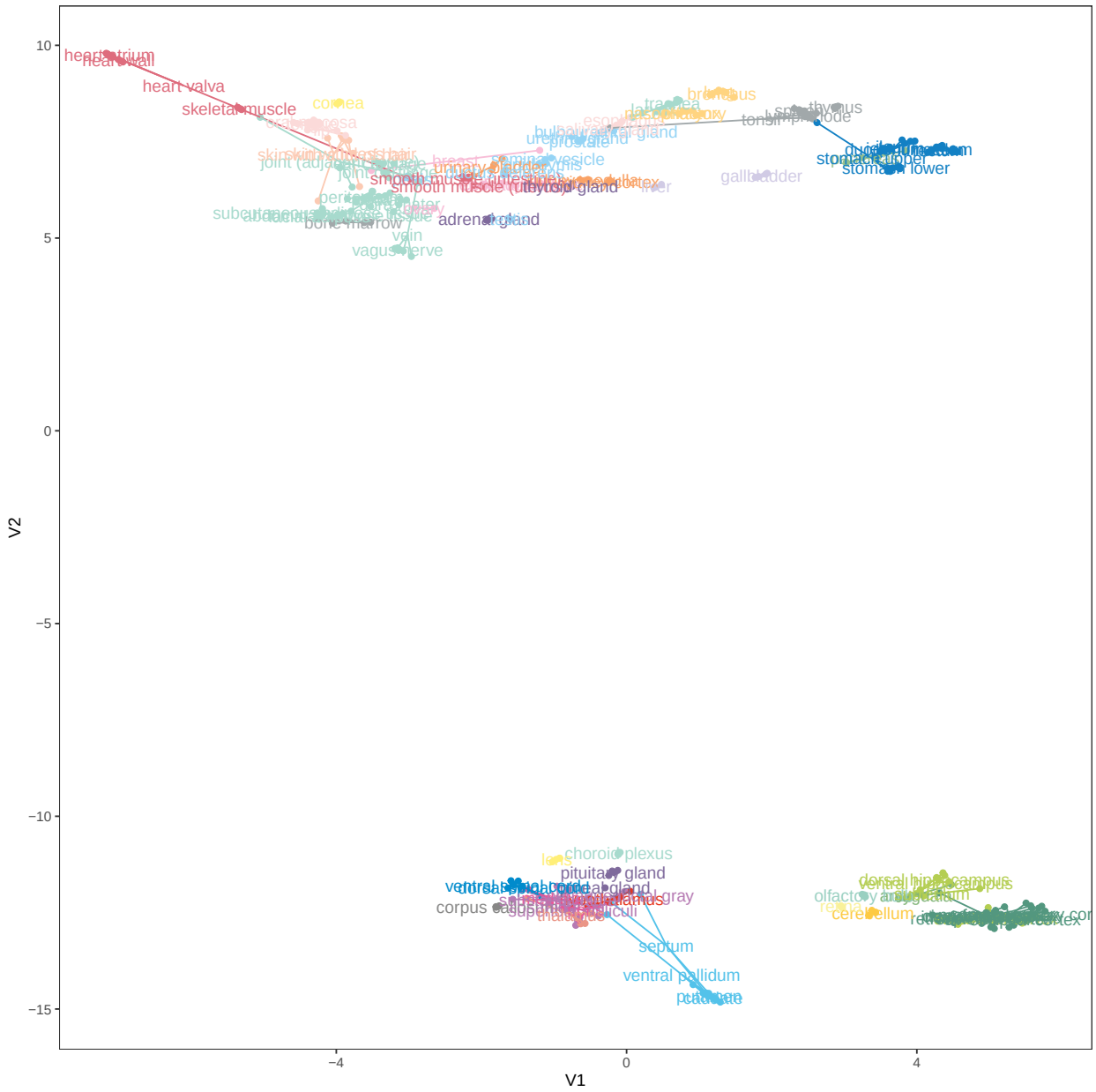
Below we see three plots showing the twodimensional projections of the data using three different strategies:
1. PCA

2. UMAP



10
0
−10
−5.0
−2.5
0.0
2.5
5.0
V1
V2
heart atrium
breast
tonsil
bronchus
stomach upper
stomach lower
liver
gallbladder
adrenal gland
bone marrow
vagus nerve
ovary
corpus callosum
choroid plexus
pituitary gland
septum
olfactory bulb
cerebellum
subiculum

3. UMAP on 92 PCs



Investigate sample mixup

Filtered samples

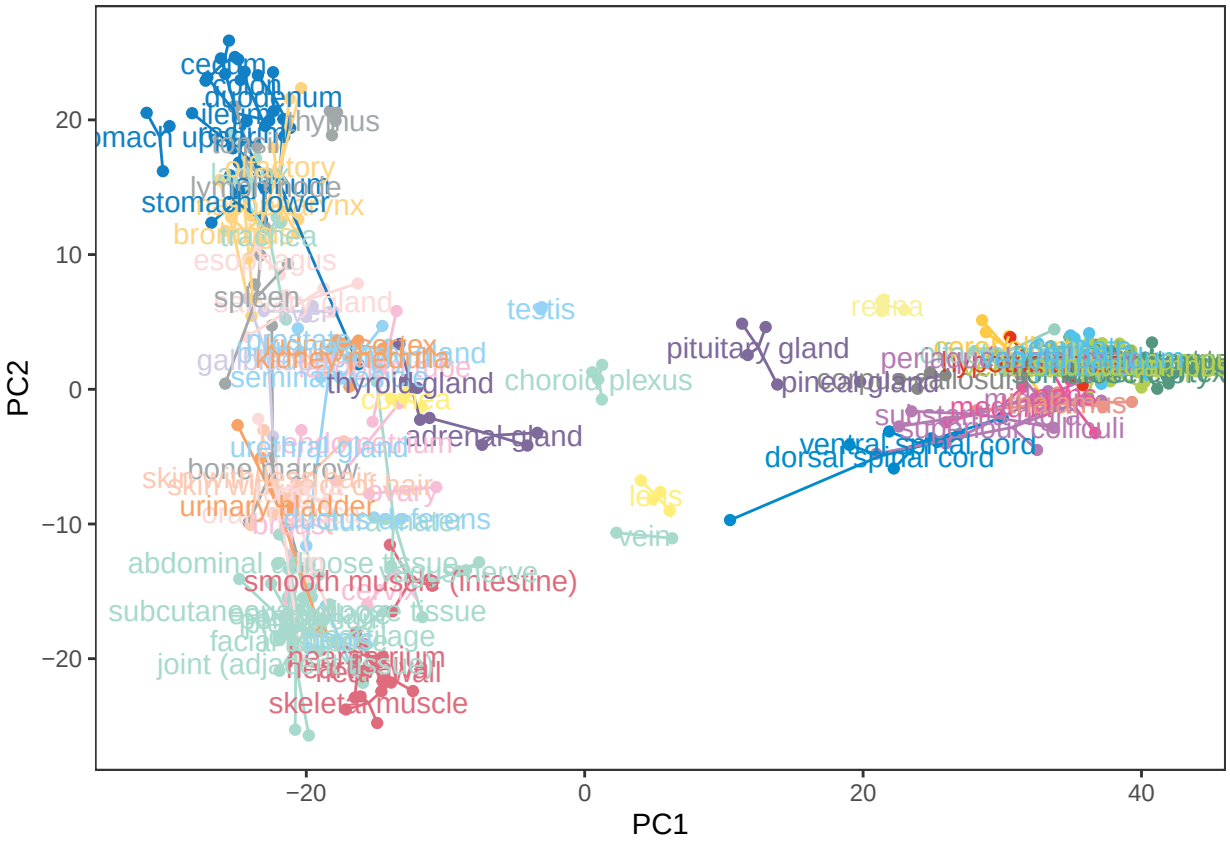
The following section describes which samples were removed from the data set. Summary plots for each individual tissue can be found in the file "Sample QC summary.pdf". Please see this file for plots showing how samples cluster in relation to their tissue and to the dataset as a whole.

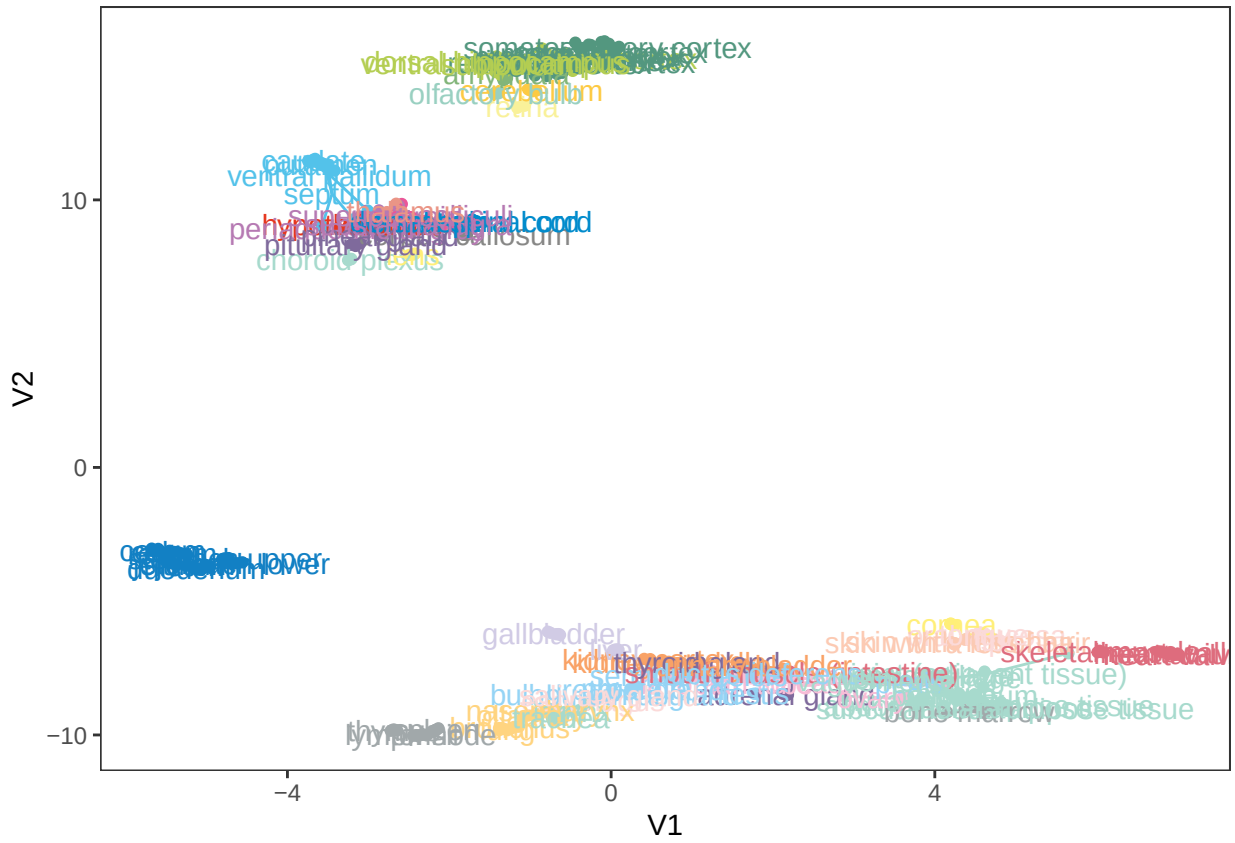
```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1 rows [17].
```

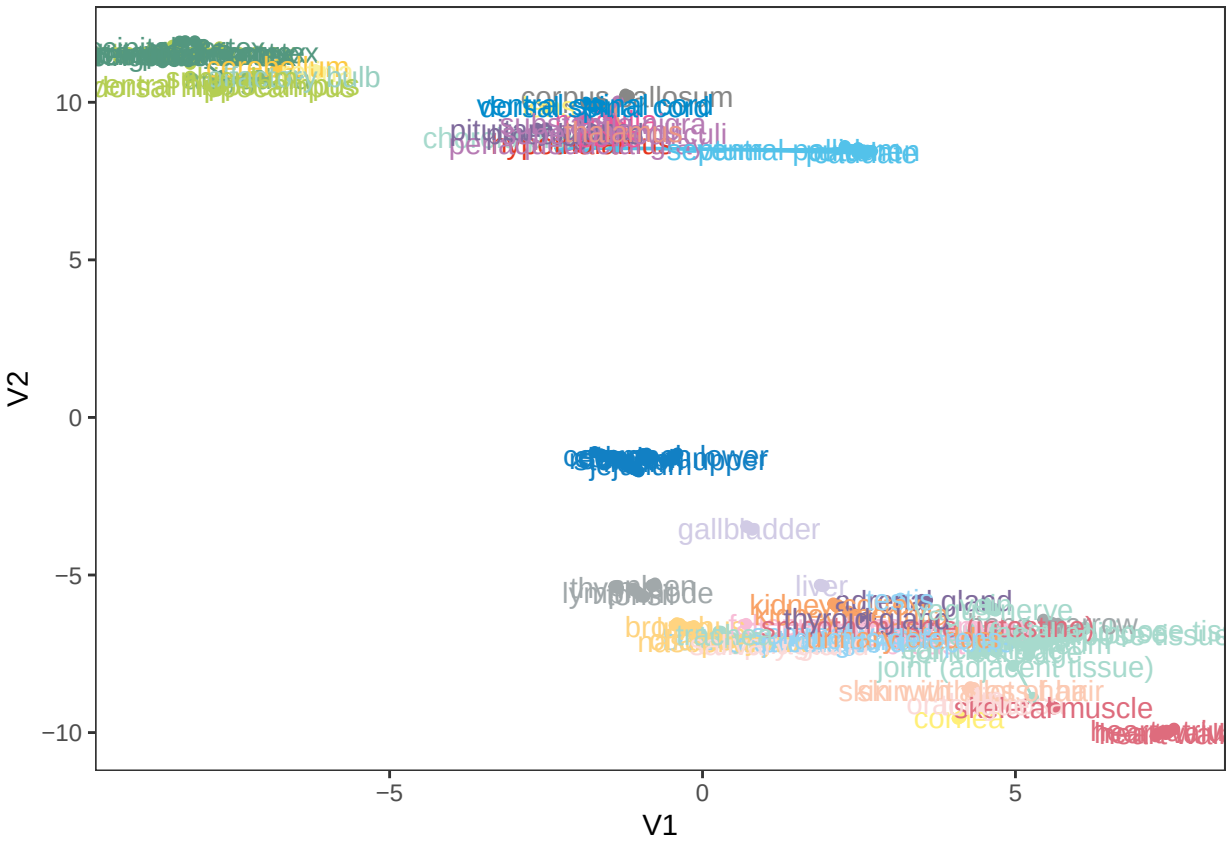
Table 1: Samples that were removed from the data set

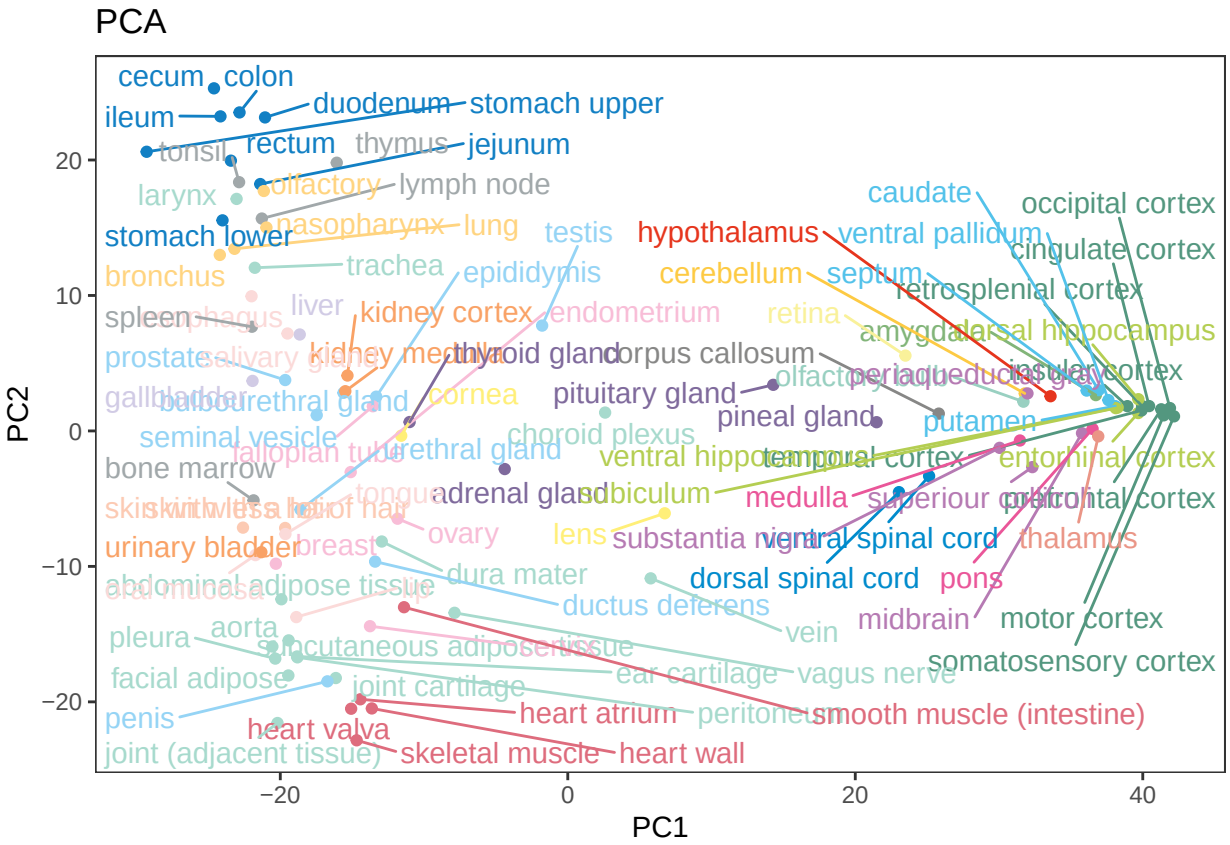
sample_ID	tissue	individual	comment
vein_c	vein	c	Wrong tissue collected, nerve was collected instead of vein as confirmed by histology
vein_d	vein	d	Clusters far from other vein samples
skiB_a	skiB	a	Clusters far from other skin samples
skiB_b	skiB	b	Clusters far from other skin samples
skiH_b	skiH	b	Clusters far from other skin samples
pan_a	pan	a	A large proportion of genes has TPM = 0
pan_b	pan	b	A large proportion of genes has TPM = 0
pan_c	pan	c	A large proportion of genes has TPM = 0
pan_d	pan	d	A large proportion of genes has TPM = 0
carJ_b	carJ	b	A large proportion of genes has TPM = 0 (and it therefore clusters with some pancreas
skiL_c	skiL	c	Histology showed the sample included a lot of hair folicles
smoU_a	smoU	a	Histology showed that the sample was mostly cervix, and clustering confirms this
smoU_b	smoU	b	Histology showed that the sample was mostly cervix, and clustering confirms this
heaV_c	heaV	c	Clusters far from other heart muscle samples
stoU_a	stoU	a	Clusters far from other GI samples and the sample included some esophagus in contrast
ton_b	ton	b	Clustering and histology shows that sample is not tonsil, but esophagus
OB-3X	OB-3X	NA	This sample was an extra “rouge sample” of olfactory bulb from one individual

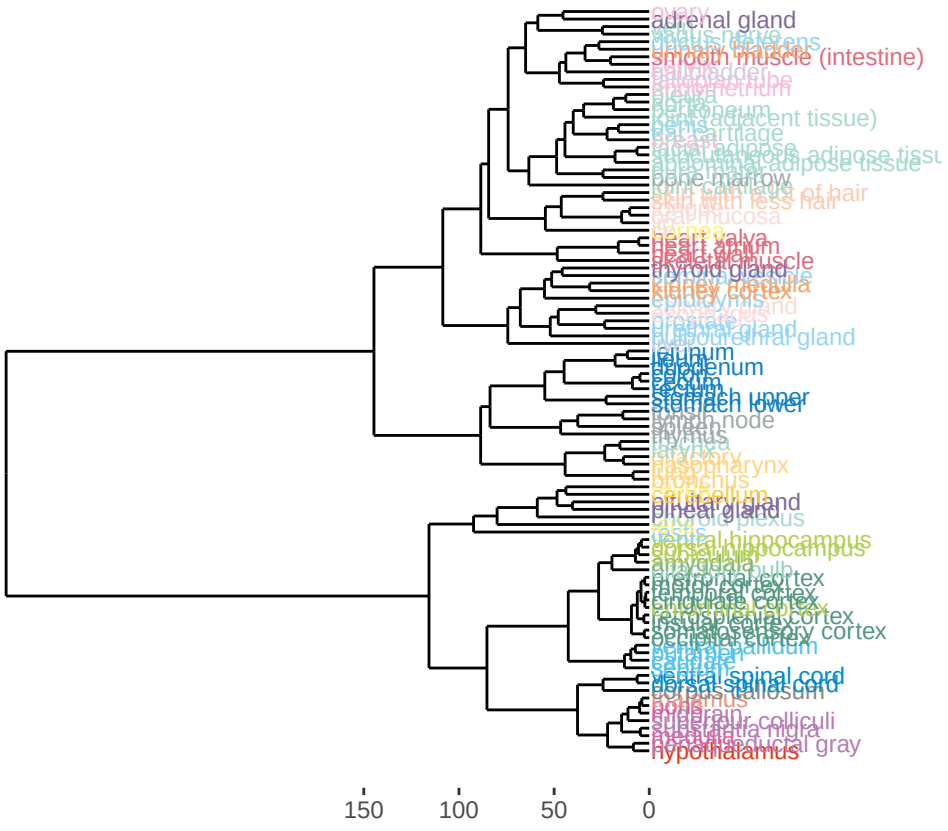
The following plots show the clustering after problematic samples have been excluded.




V2
10
0
−10
−4
0
4
V1
gallbladder
bone marrow


corpus callosum
gallbladder
liver
adrenal gland
thyroid gland
joint (adjacent tissue)
skeletal muscle
cornea
V1
V2
10
5
0
−5
−10
−5
0
5


PCA
cecum
colon
ileum
duodenum
stomach upper
tonsil
rectum
thymus
jejunum
larynx
olfactory
lymph node
caudate
occipital cortex
nasopharynx
lung
testis
hypothalamus
ventral pallidum
cingulate cortex
stomach lower
bronchus
trachea
epididymis
cerebellum
septum
retrosplenial cortex
spleen
liver
kidney cortex
endometrium
retina
amygdala
dorsal hippocampus
prostate
kidney medulla
thyroid gland
corpus callosum
gallbladder
bulbourethral gland
cornea
pituitary gland
pineal gland
putamen
seminal vesicle
choroid plexus
fallopian tube
urethral gland
bone marrow
adrenal gland
medulla
urinary bladder
breast
ovary
lens
substantia nigra
ventral spinal cord
thalamus
abdominal adipose tissue
dura mater
dorsal spinal cord
pons
ductus deferens
midbrain
motor cortex
pleura
aorta
vein
somatosensory cortex
facial adipose
joint cartilage
ear cartilage
vagus nerve
penis
heart atrium
smooth muscle (intestine)
heart valva
joint (adjacent tissue)
skeletal muscle
heart wall
20
10
0
−10
−20
−20
0
20
40
PC2
PC1


adrenal gland
smooth muscle (intestine)
hypothalamus
150
100
50
0