Preliminary analysis of 10X Genomics scRNA-seq Data from the Bresnick Lab

Jurijs Nazarovs

Keles Lab

May 16, 2018

Data information

We consider 4 replicates from Bresnick Lab: mutant(A and B) and wild type (F and I) and conduct an initial comparison with a published 10X data (Peripheral blood mononuclear cells (PBMCs)) to see the quality and similarity between data sets (our and published).

Along the presentation published 10x dataset referenced as Real and accessible here:
https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k

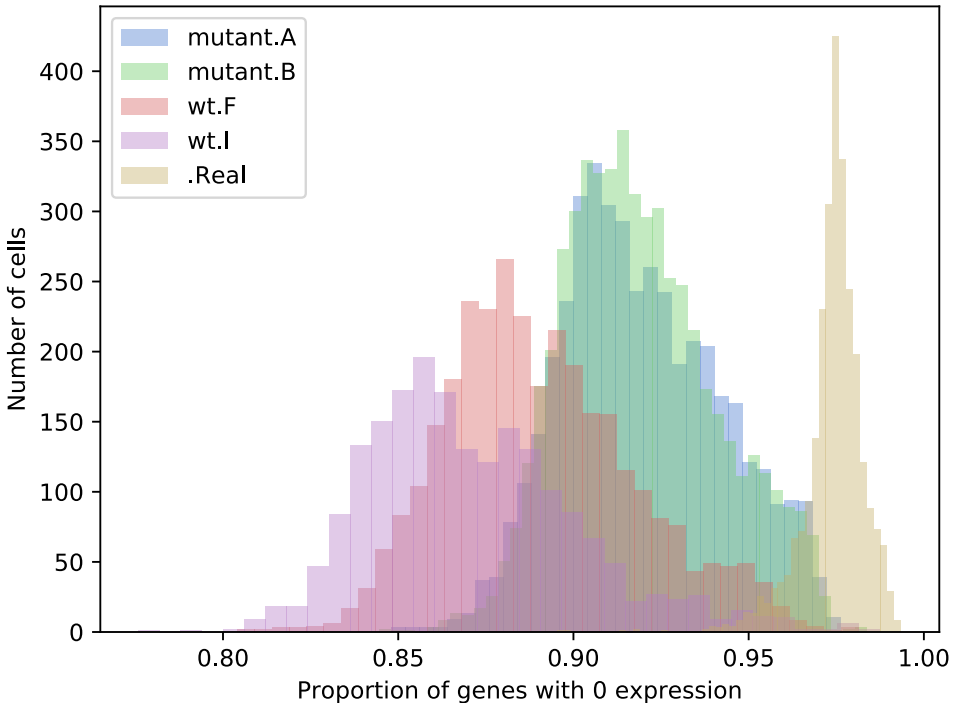
More published 10x datasets can be found here:
https://support.10xgenomics.com/single-cell-gene-expression/datasets

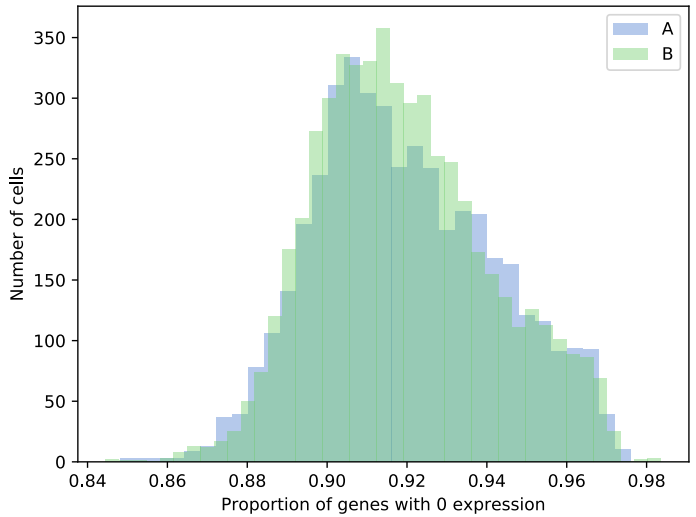
For initial comparison we use raw data - unnormalized data after CellRanger processing (also, we have a umi-version of analysis.)

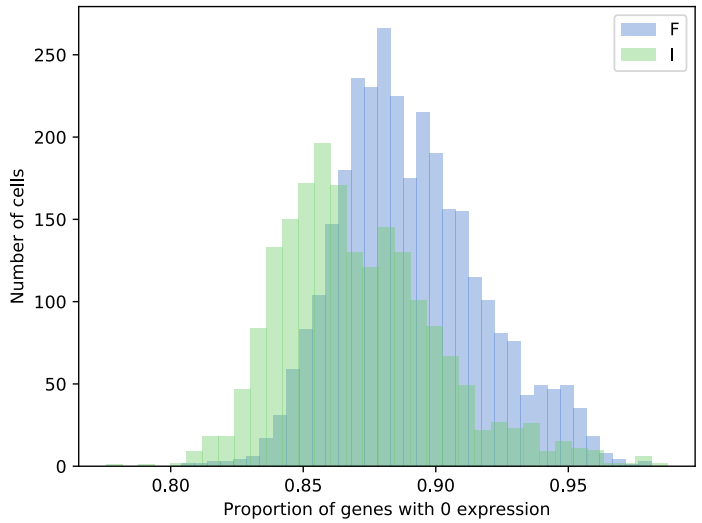
	A	B	F	I	Real
Condition	Mutant	Mutant	Wild type	Wild type	
Number of genes	27877	27877	27877	27877	32552
Number of cells	4351	5439	3119	1985	2700

Comparison with other 10X Data

Proportion of genes with 0 expression across cells:



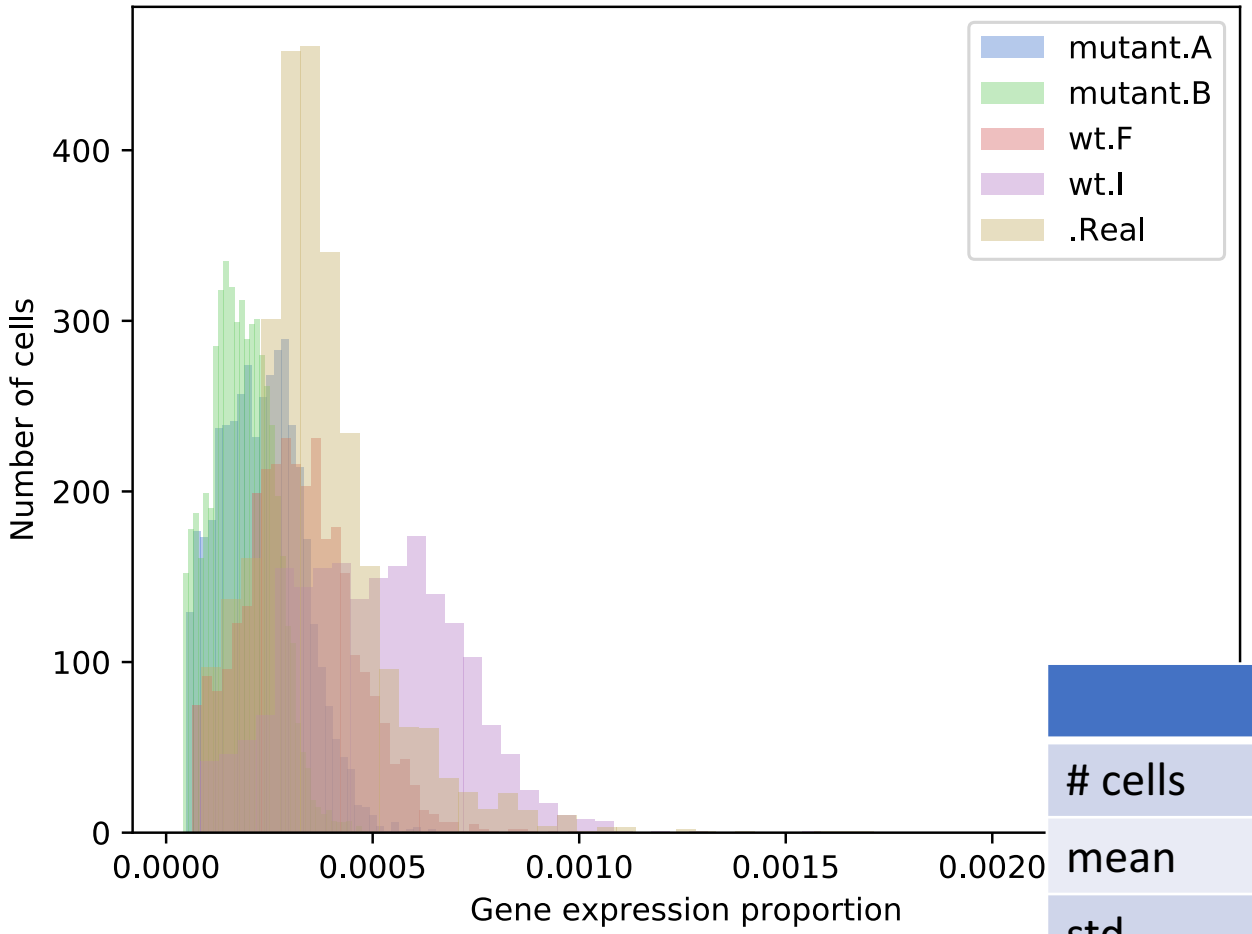




	A	B	F	I	Real
# cells	4351	5439	3119	1985	2700
min	0.8480	0.8445	0.8041	0.7758	0.8954
median	0.9169	0.9168	0.8871	0.8662	0.9750
max	0.9761	0.9836	0.9813	0.9873	0.9935

Raw data

Normalized gene expression proportion per cell with respect to total expression



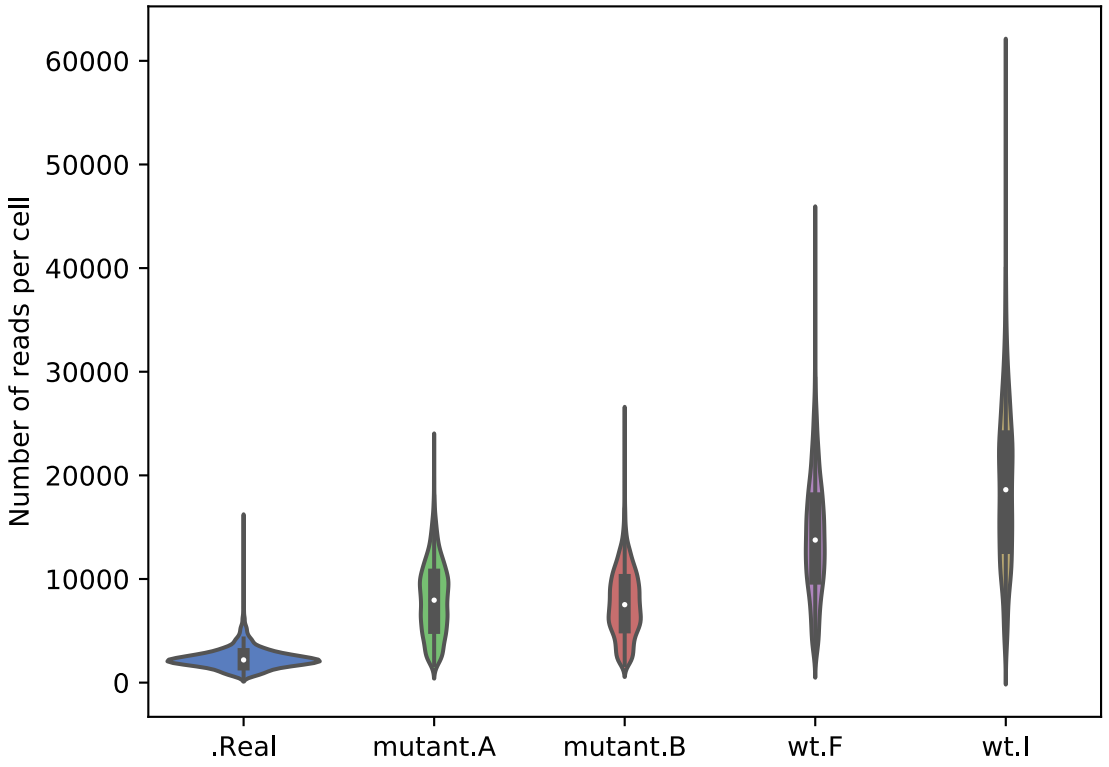
We consider the total gene expression per cell normalized with respect to total gene expression for all experiments (sum of all umi-counts)

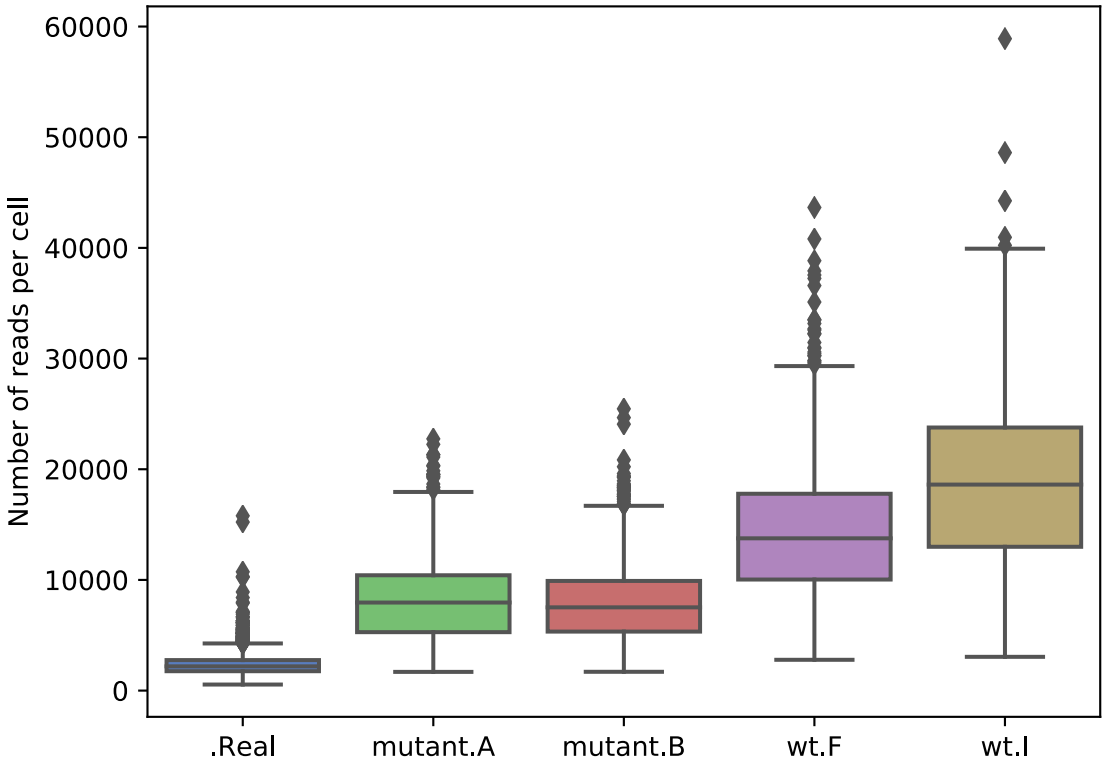
Question: Why is variance so different between mutant and wildtype?
Possible answer: if we have more duplicates per cell per specific experiment, then variance for this experiment is less (because we have less unique values)

	A	B	F	I	Real
# cells	4351	5439	3119	1985	2700
mean	0.00023	0.000184	0.000321	0.000504	0.000370
std	0.0001	0.000077	0.000132	0.000199	0.000171

Raw data

Sequencing depths of cells across samples





Q: Why are mutants lower than wild-types? Is there an experimental/biological reason or this is due to sequencing step?

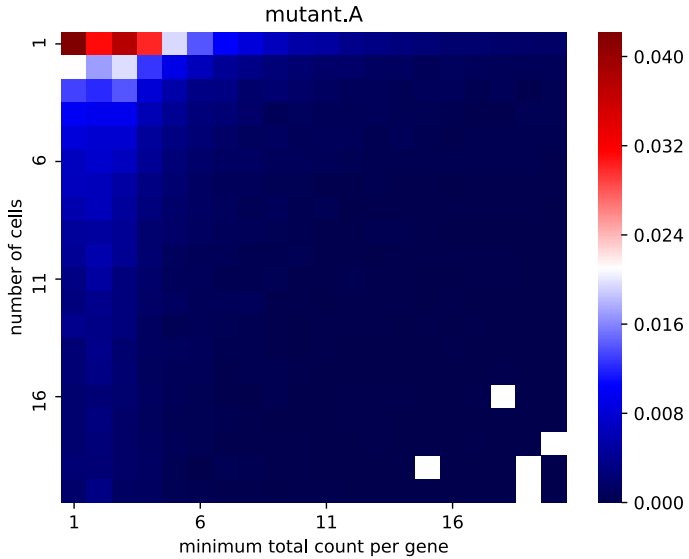
Note: Despite that mutant have twice more cells, it has higher percentage of 0 expressed genes
(Median: 91%, 91% vs 88%, 86%)

Raw data

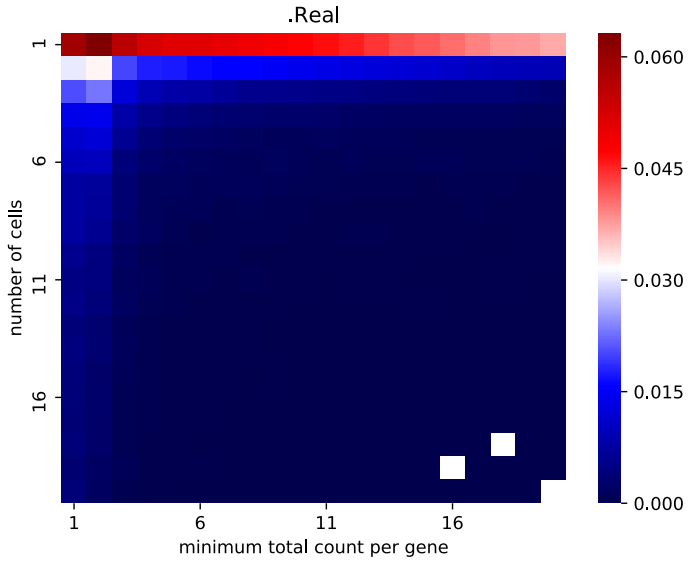
Pre-processing for analysis

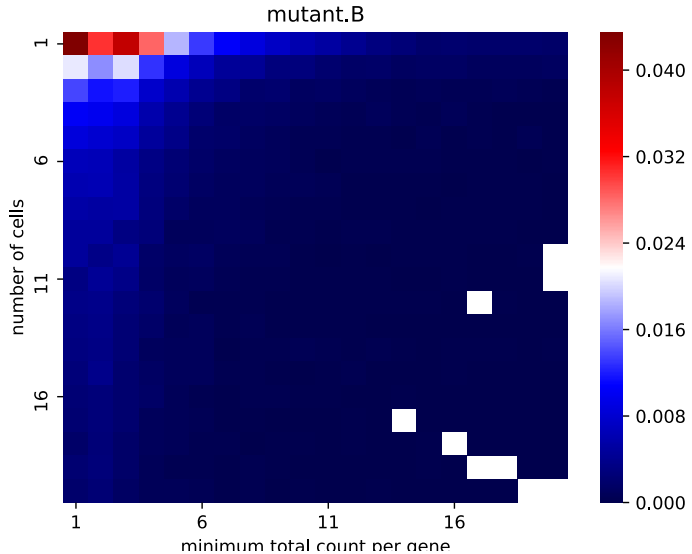
1) Filter genes which do not have at least (min.reads counts in min.cells).
2) Normalize by sequencing depth (data for each cell is divided by the total depth of the cell and then multiplied by “T”, where T is median cell depth across all samples.
3) Log base 10 transformation (+ 1 to avoid log(0)).
4) We will apply 1 & 2 for each sample separately and then take union of the remaining genes as the final set of genes will keep in the analysis.

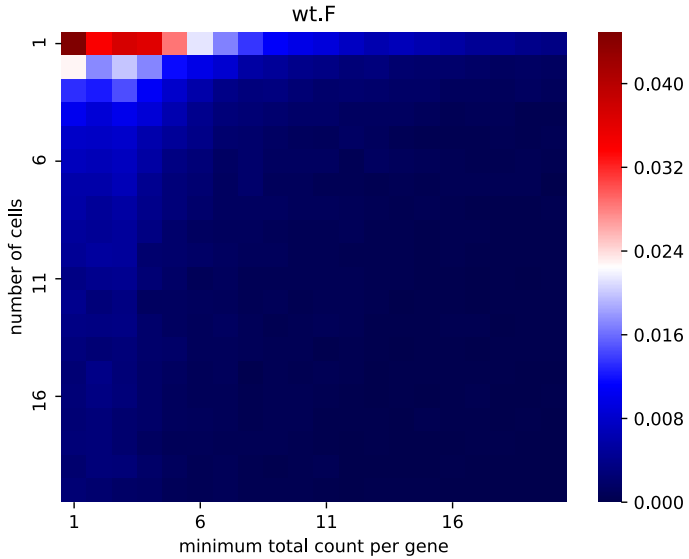
Distribution of proportion of genes with minimal counts – useless

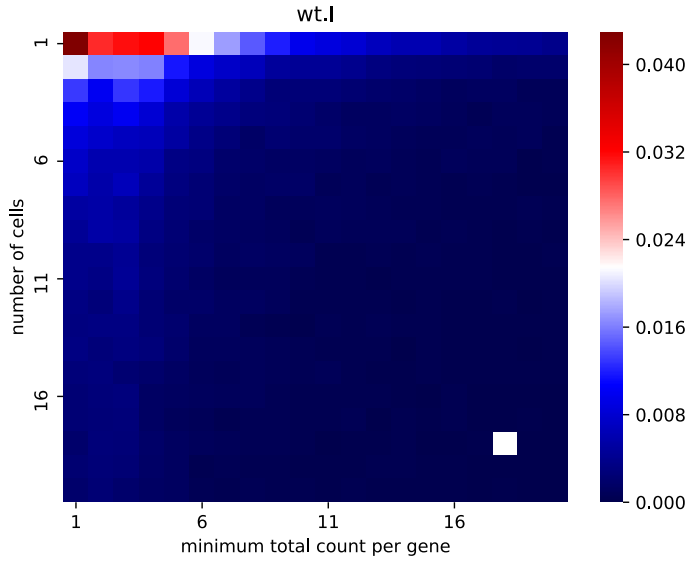


Number of selected genes/
number of genes per cell



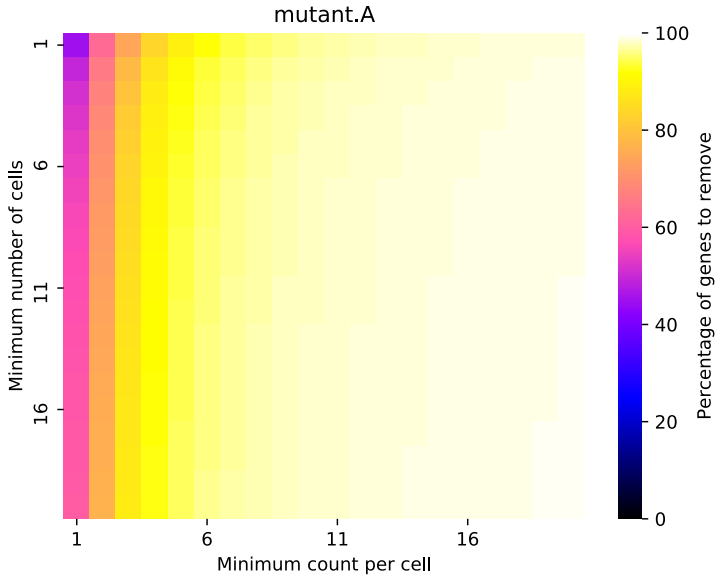


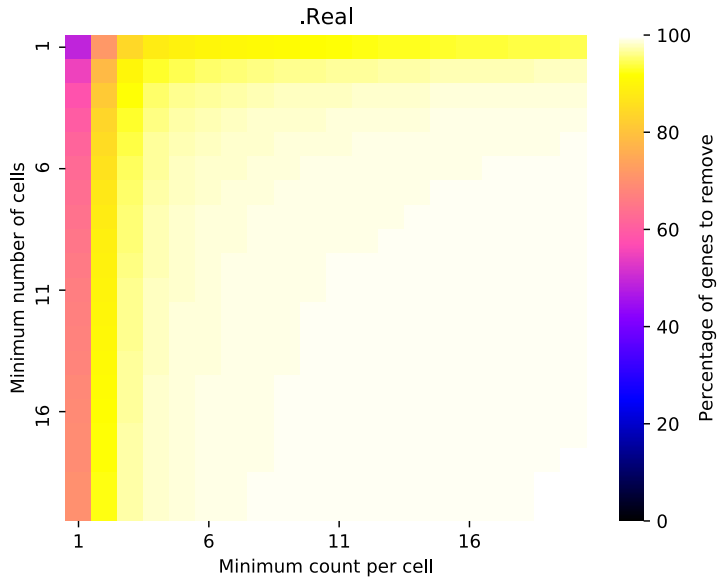


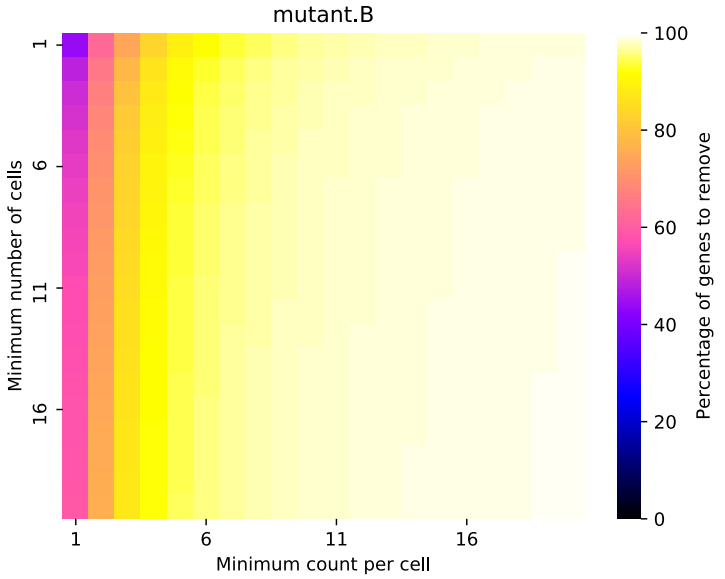


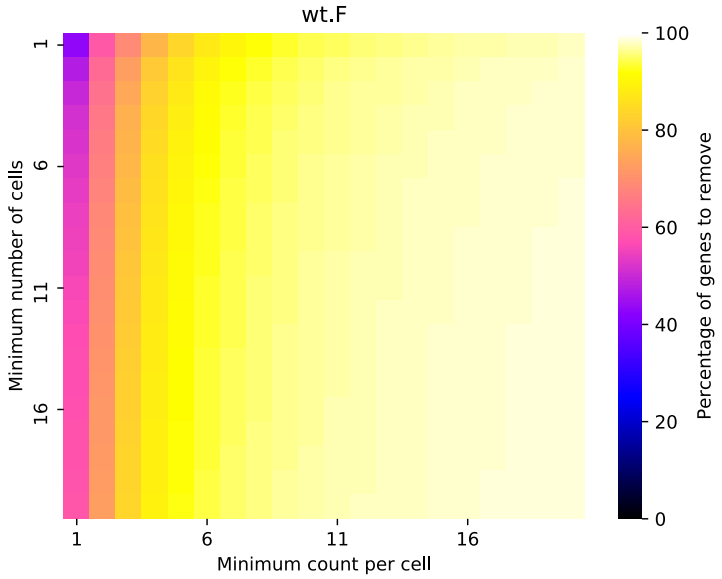
Raw data

Percentage of removing genes: at least n counts, at least n cells

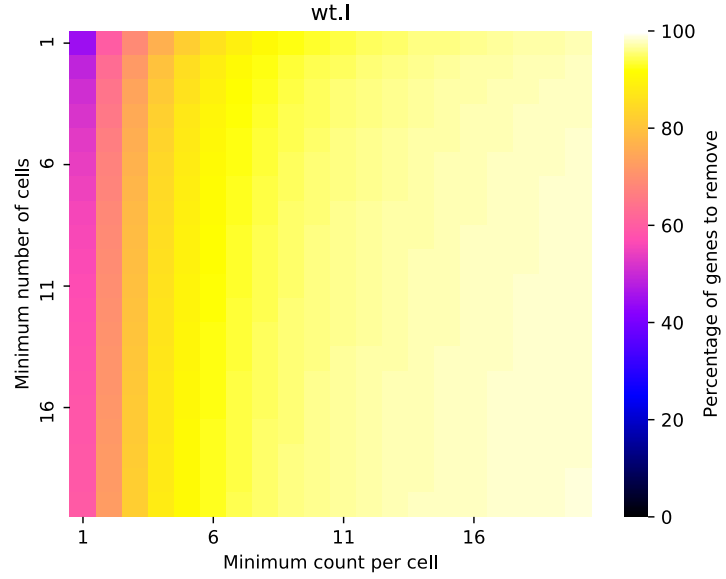








Select minimum read = 2
and at least 2 cells



Raw data

Short summary about 0 expressed genes

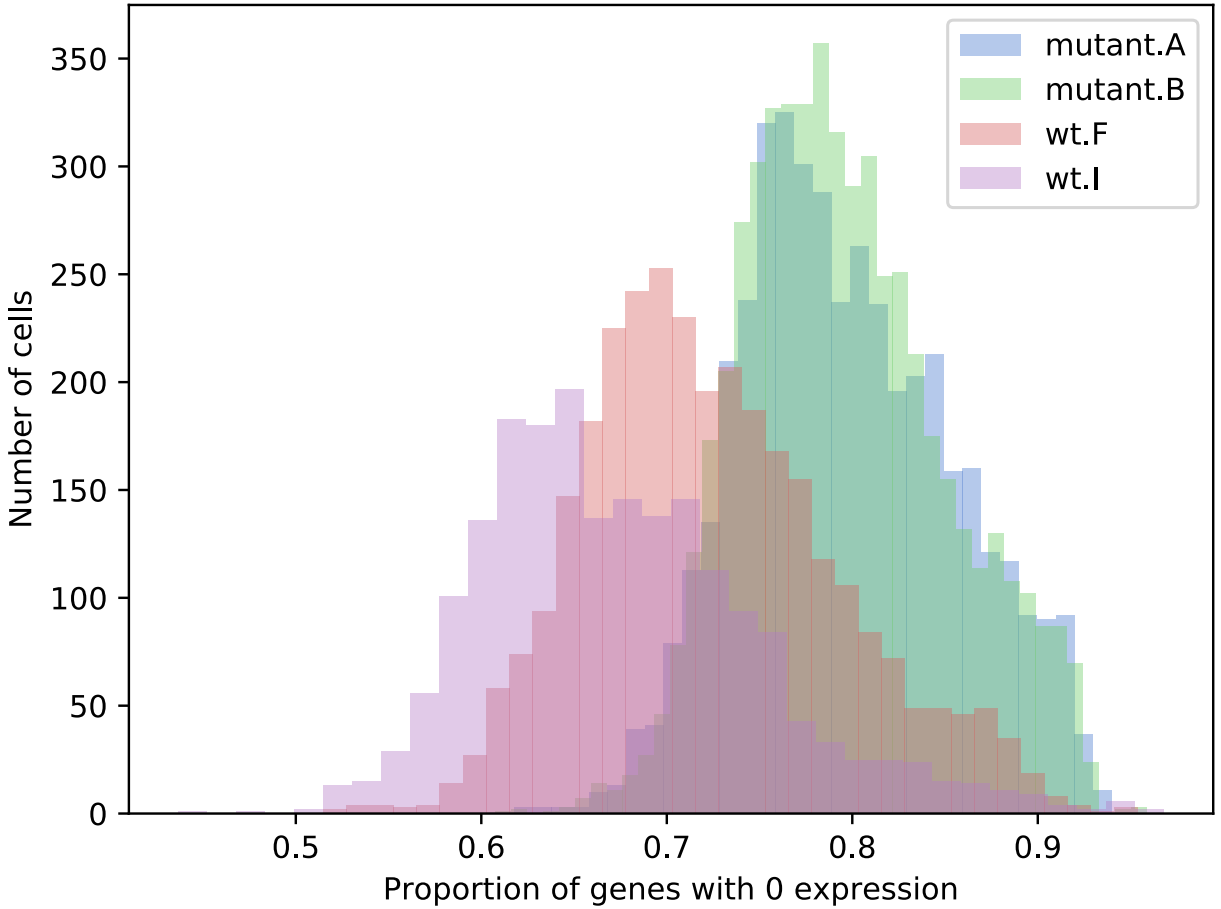
Sample	% of 0 expressed genes	Number of remaining genes
Mutant - A	45.05	15318
Mutant - B	44.11	15579
Wild type - F	43.16	15845
Wild type - I	44.65	15429
Real	49.15	16550

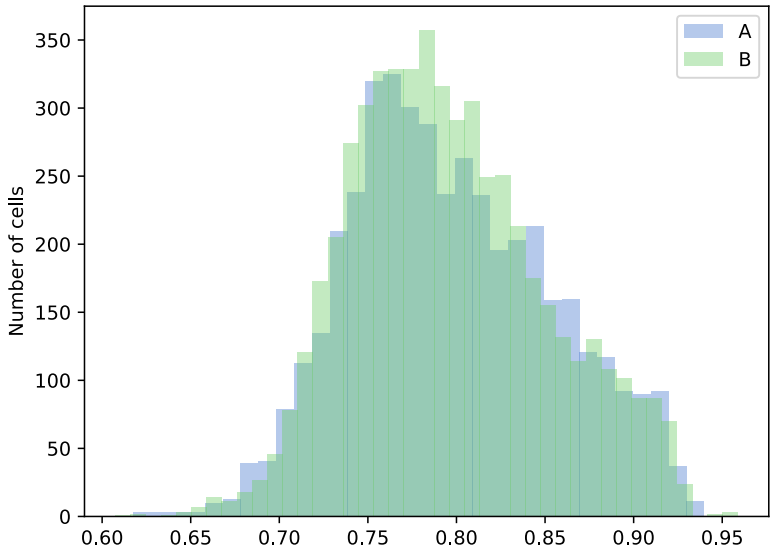
Results on min.read-min.cell filtering

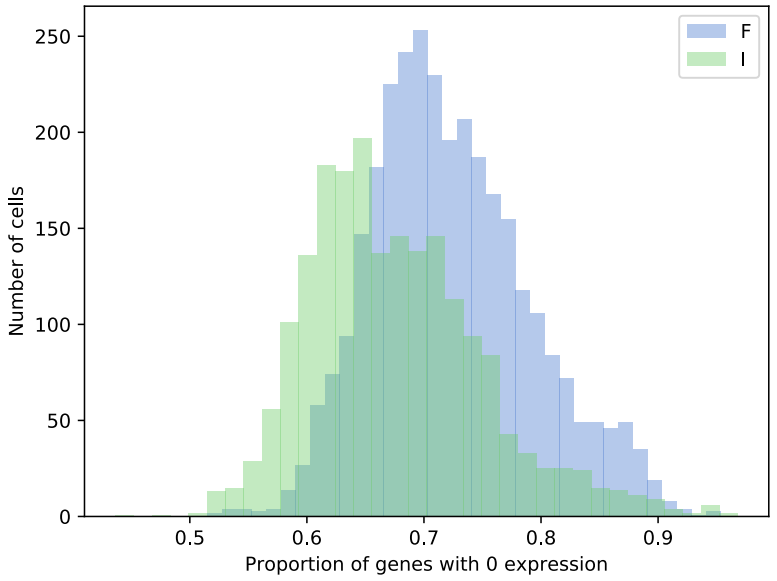
Remove genes which do not have a count of at least 2 in at least 2 cells.

Sample	% of genes that satisfy criteria	% of genes that are intersect	Number of remaining genes
Mutant - A	65.79	60.5	10999
Mutant - B	65.46	60.5	10999
Wild type - F	62.51	60.5	10999
Wild type - I	62.94	60.5	10999
Real	78.12	-	-

Proportion of genes with 0 expression across cells after filtering



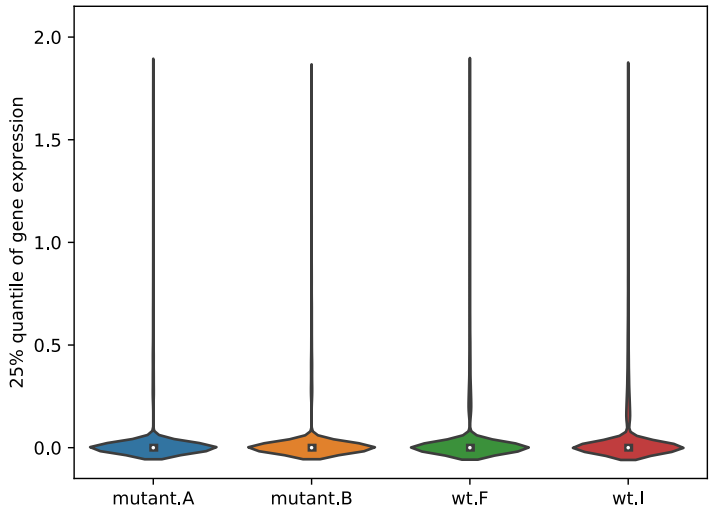


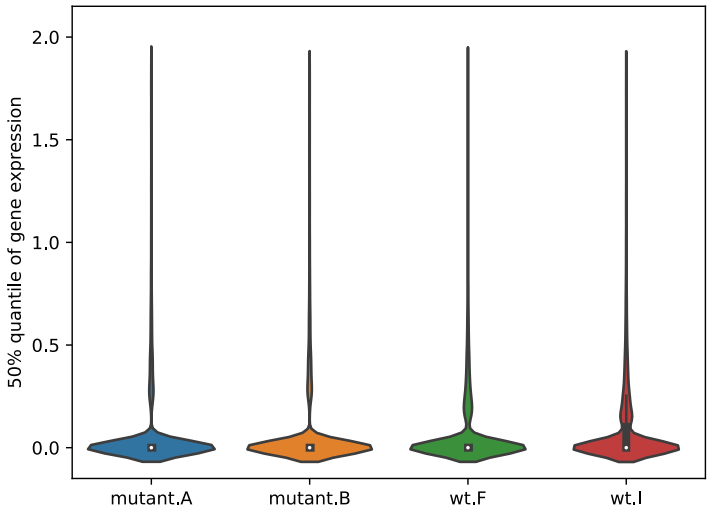


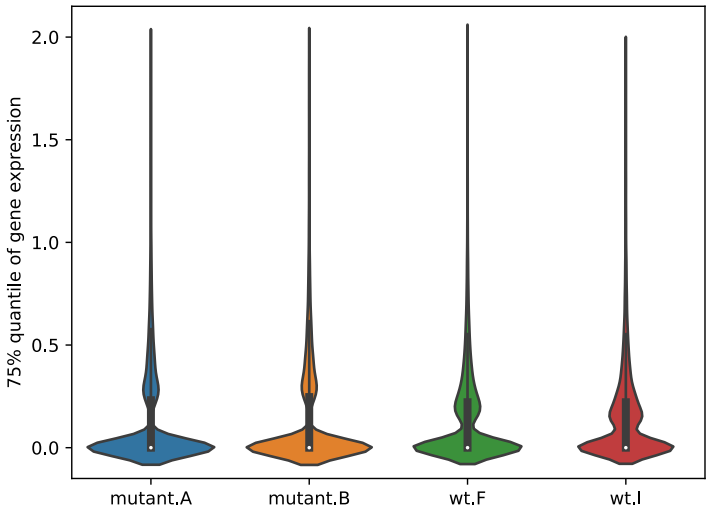
Filtered data

Distribution of gene expression quantiles - useless

Consider distribution of gene expression for every gene for all cells. Select one of quantiles: 25%, 50%, 75%.
Plot distribution of quantiles for every sample.

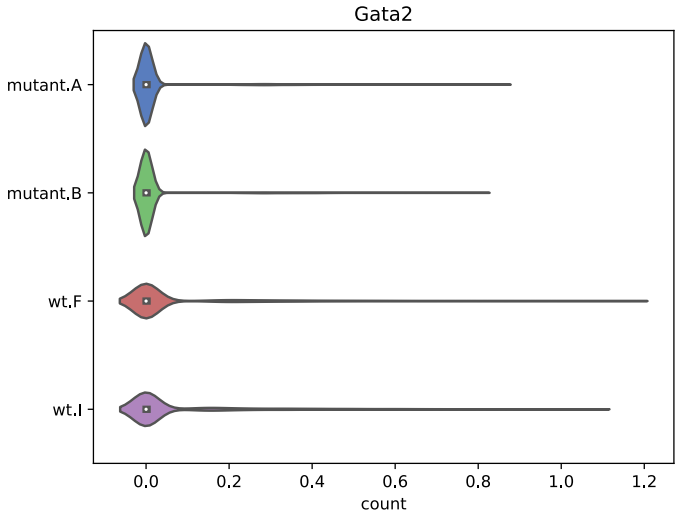


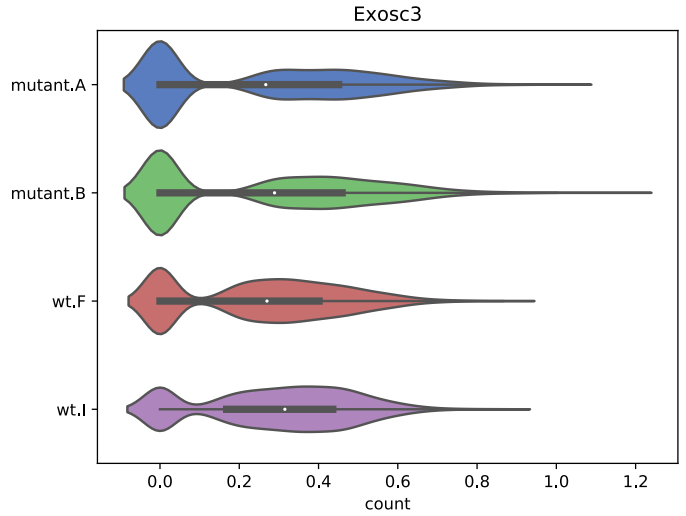


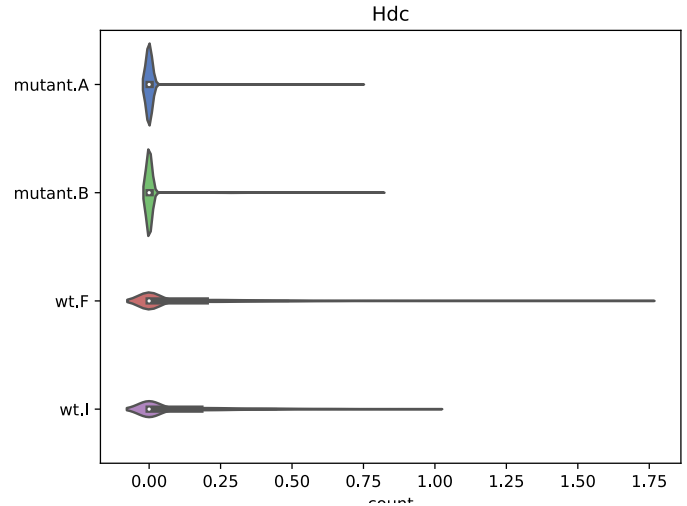


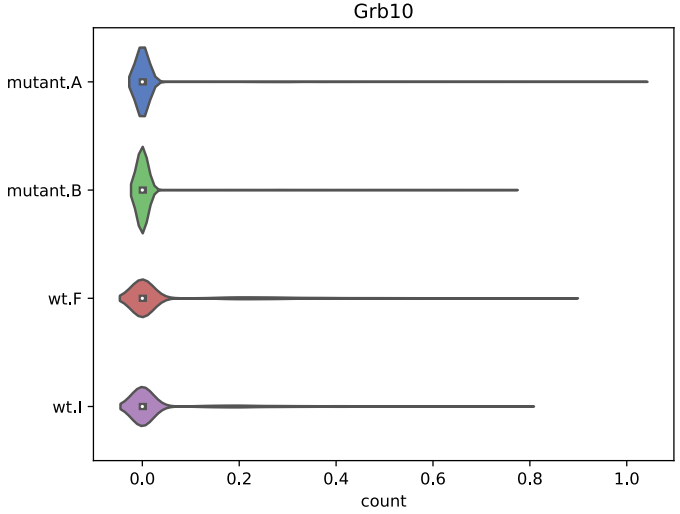
We can see that 75% quantile of gene expression is higher in mutant then in wildtype

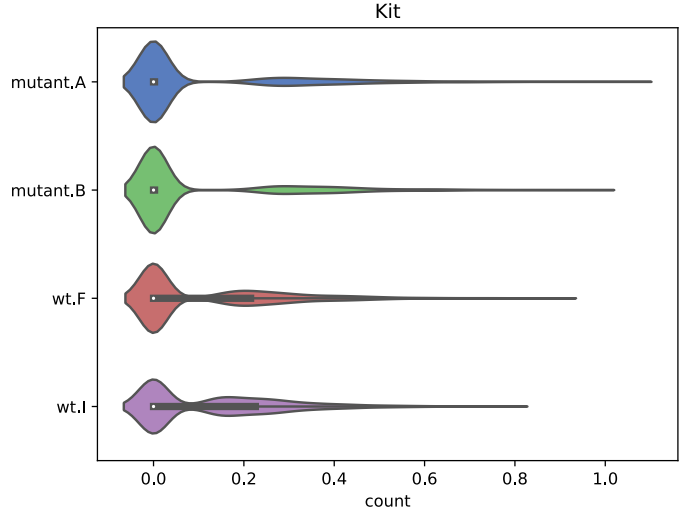
Comparison of expression distribution of different genes - useless

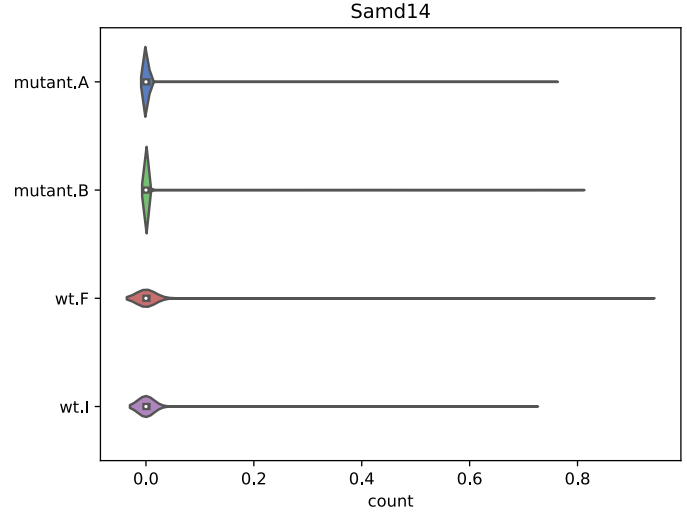












Gata2, Hdc, Grb10 and Samd14 have different distributions between wildtype and mutant.
They have a higher concentration of 0 in mutant.

Dimension reduction

t-sne

T-sne has an ability to preserve local structure of data. That is, if points are close to each other in high dimension, most likely they are closed to each other in low dimension – shown by k-means applied to a raw imputed data and plotted results using t-sne.

To plot t-sne results for different samples on same axis, we concatenate all experiments together and later plot results of t-sne separately with respect to original cell barcodes.

Data corresponding to combined experiments has following dimension: (10999, 14894)

PCA

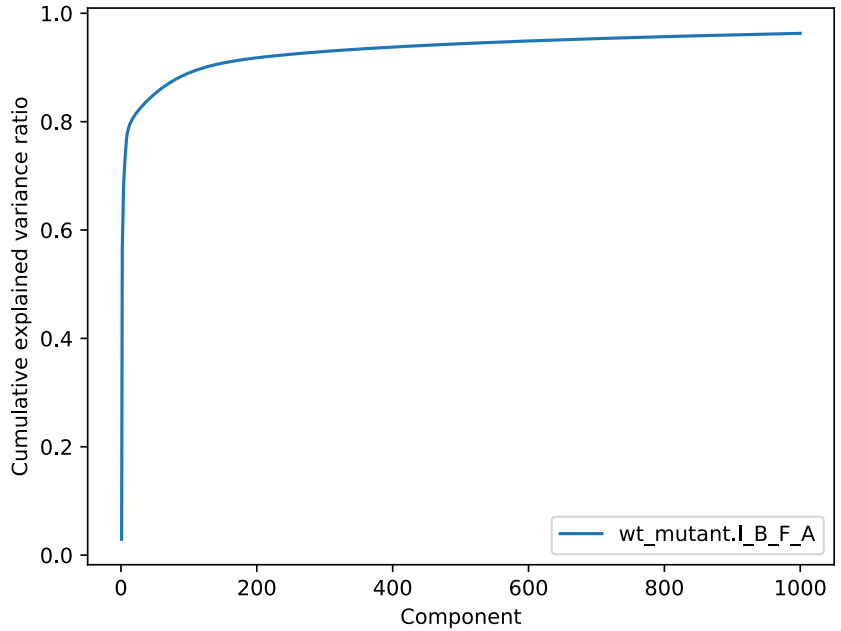
Running t-sne on such big data takes a lot of time and resources. Thus, prior to t-sne run we execute PCA first to decrease dimension.

PCA

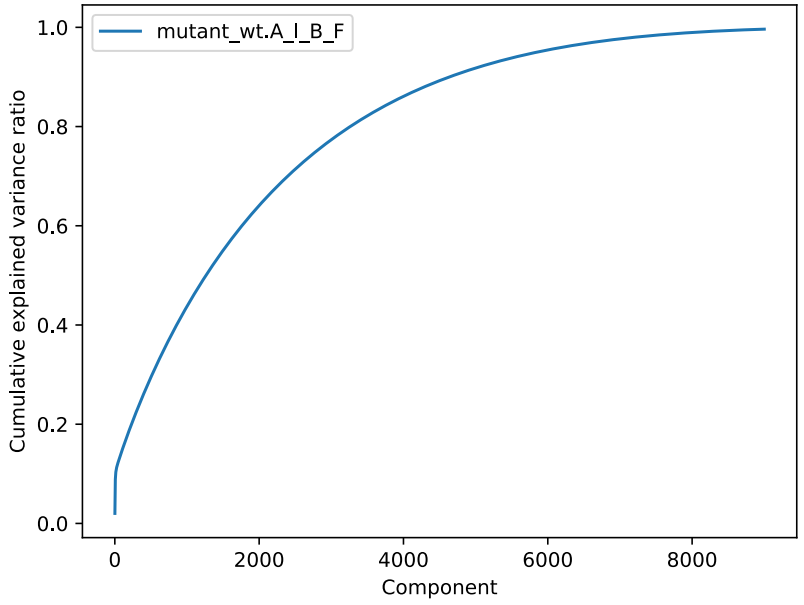
We apply PCA with number of components from 1 to 1000 to chose the best amount in respect of percentage of explained variance.

We apply PCA for both data sets: with log10(1+x) transformation and without

No log



log



For data set with no log transformation with 900 components we achieve 96% of explained variance

PCA

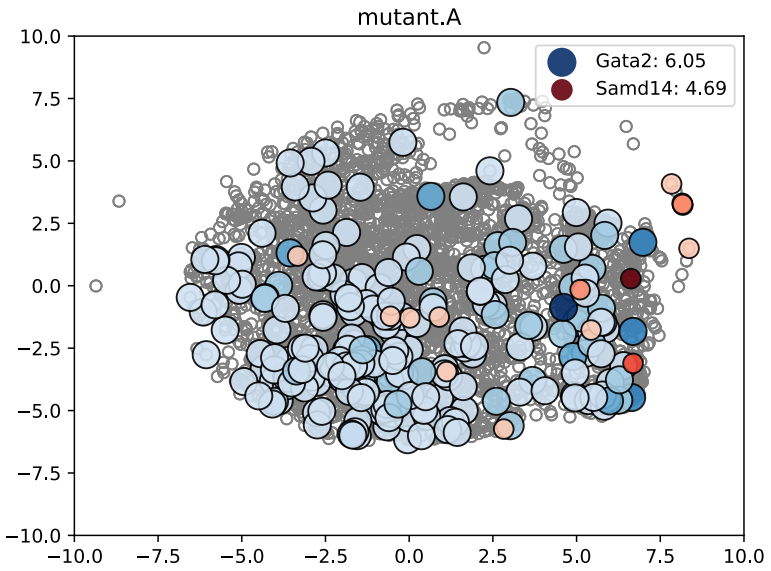
Explanation of why for data set with no log transformation with 900 components we achieve 96% of explained variance, but for log-transformed we have to select around 6k components, to achieve at least 80% of variance.

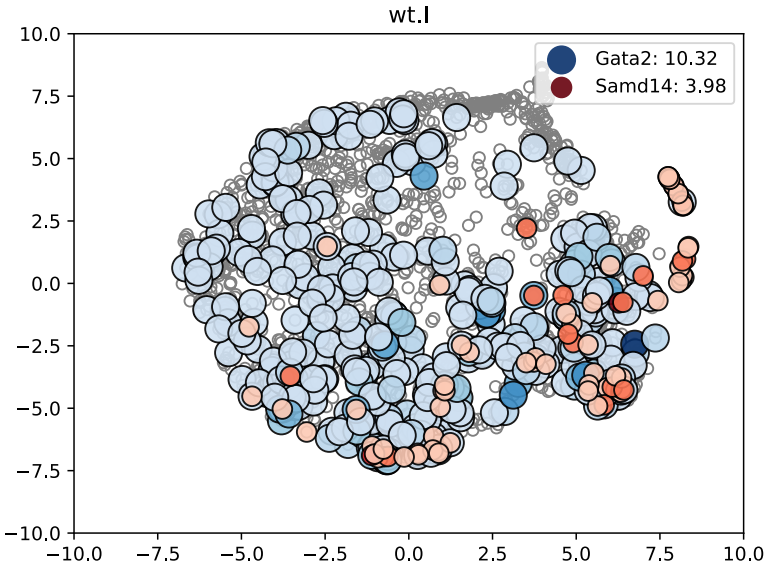
The reason for this is that different variables (genes) have VERY different variances. RNA-seq data are ultimately counts of RNA molecules, and the variance is monotonically growing with the mean (think Poisson distribution). So the genes that are highly expressed will have high variance whereas the genes that are barely expressed or detected at all, will have almost zero variance.

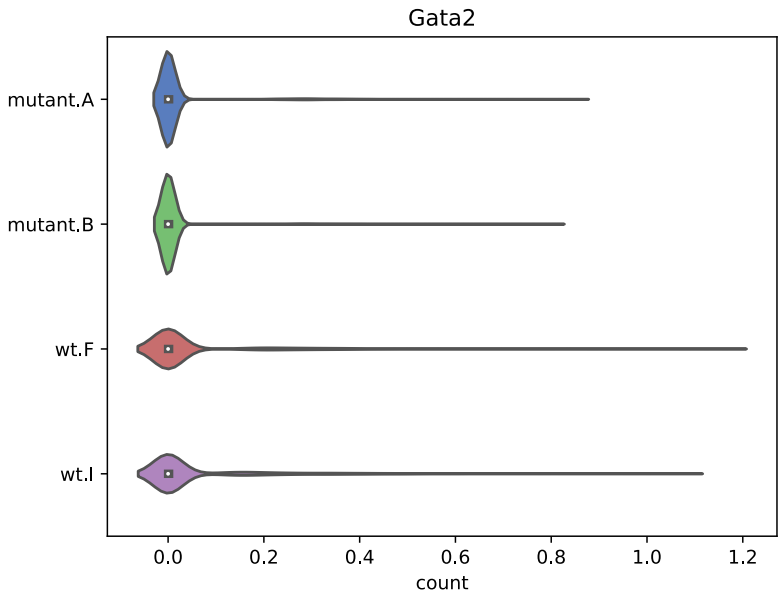
Without any transformations, there is one gene that *alone* explains above 40% of the variance (i.e. its variance is above 40% of the total variance). In this dataset, it happens to be this gene: https://en.wikipedia.org/wiki/Neuropeptide_Y which is very highly expressed (RPKM values over 100000) in some cells and has zero expression in some other cells. When you do PCA on the raw data, PC1 will basically coincide with this single gene.

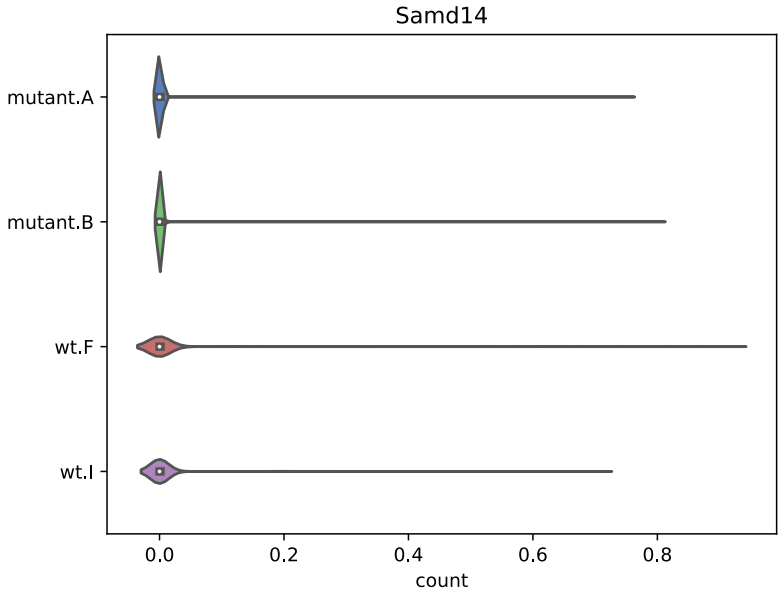
https://stats.stackexchange.com/questions/319794/why-does-log-transformation-of-the-rna-seq-data-reduce-the-amount-of-explained-v

T-sne: gene patterns – no log - useless



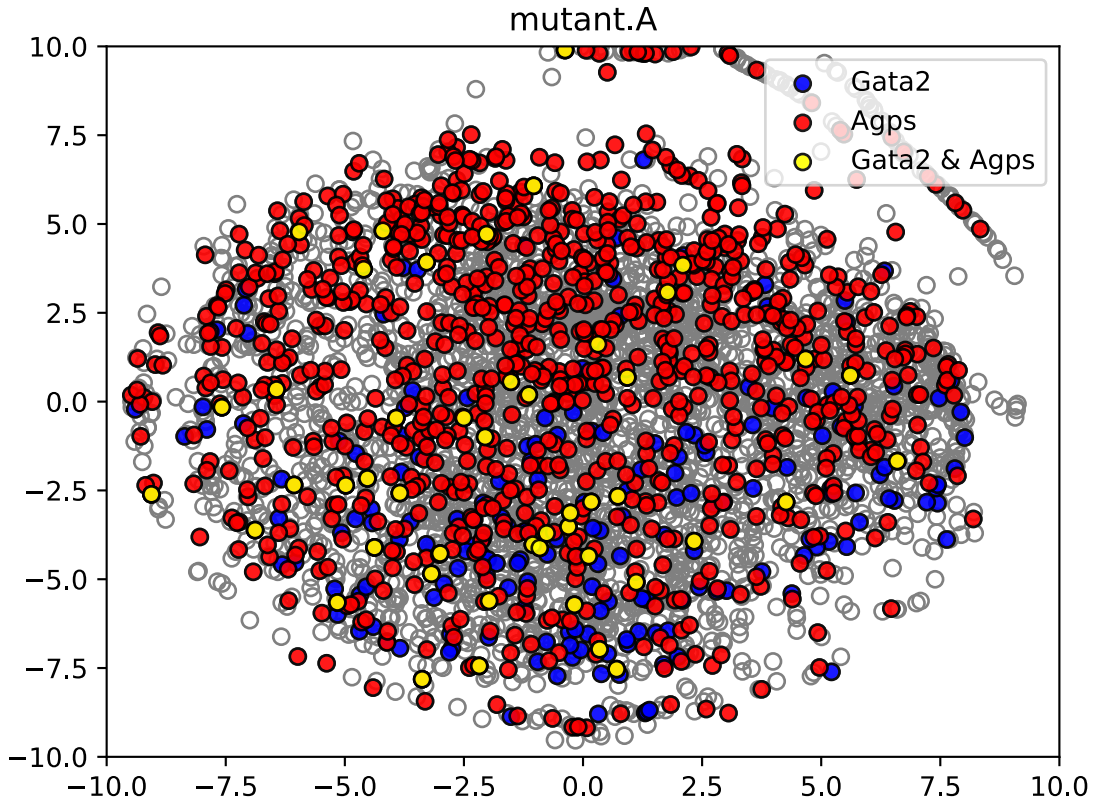


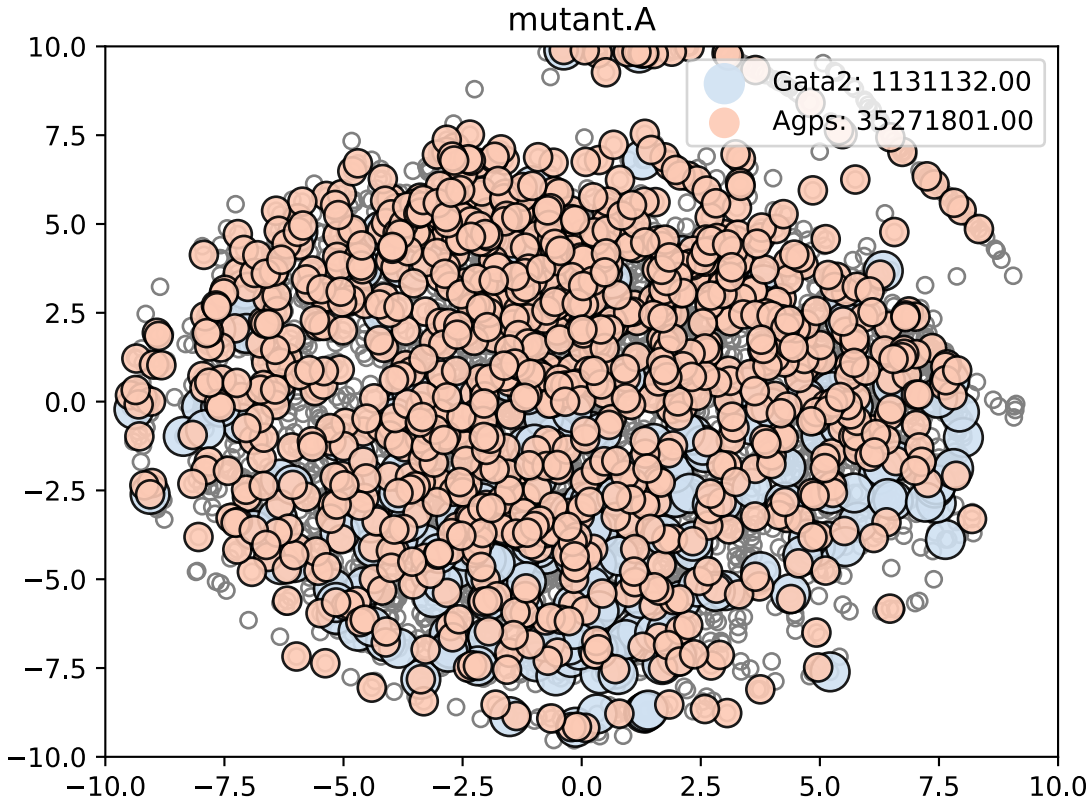




Pca = 4000, tsne = 2 – slide is useless,

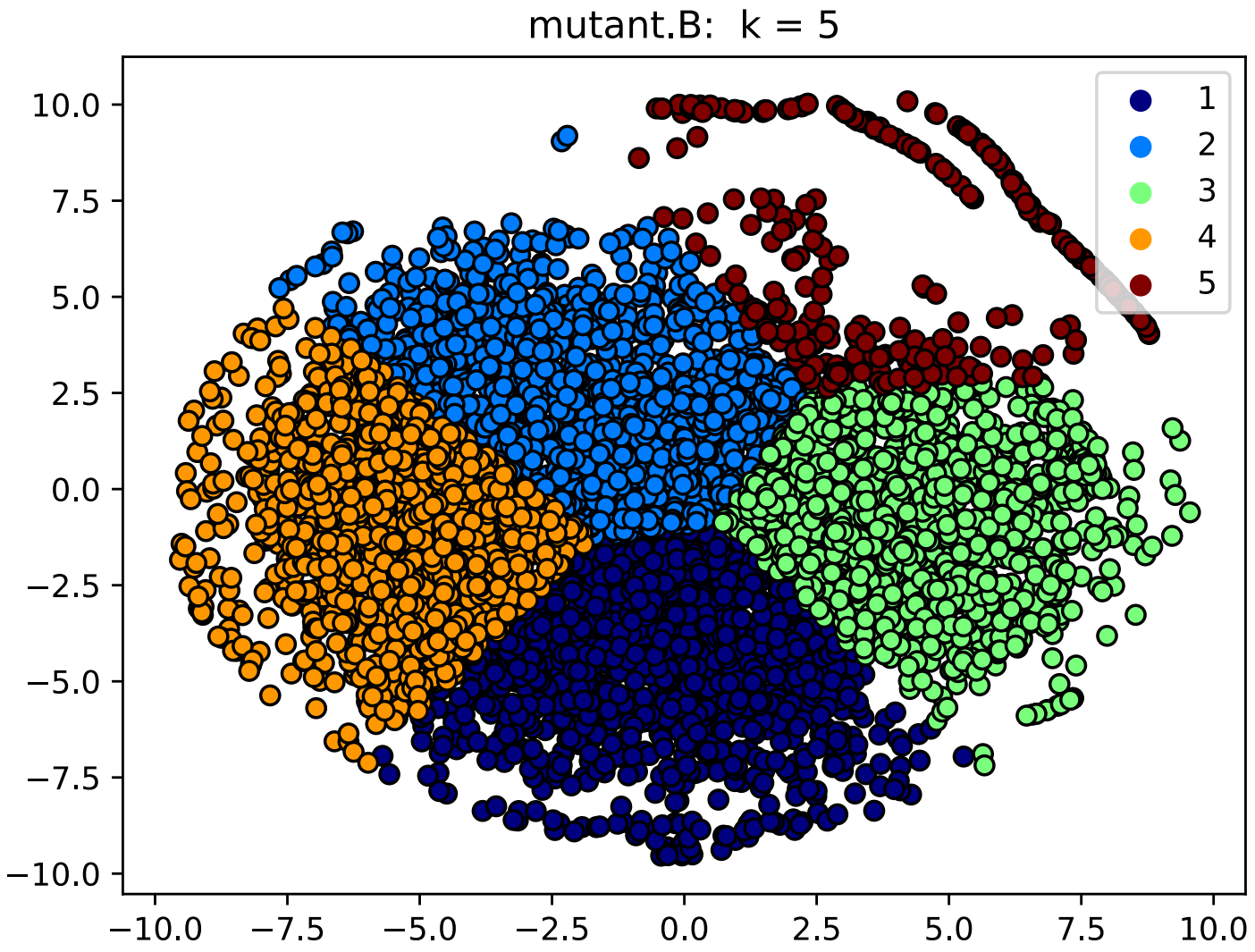
new way of pictures might be ok



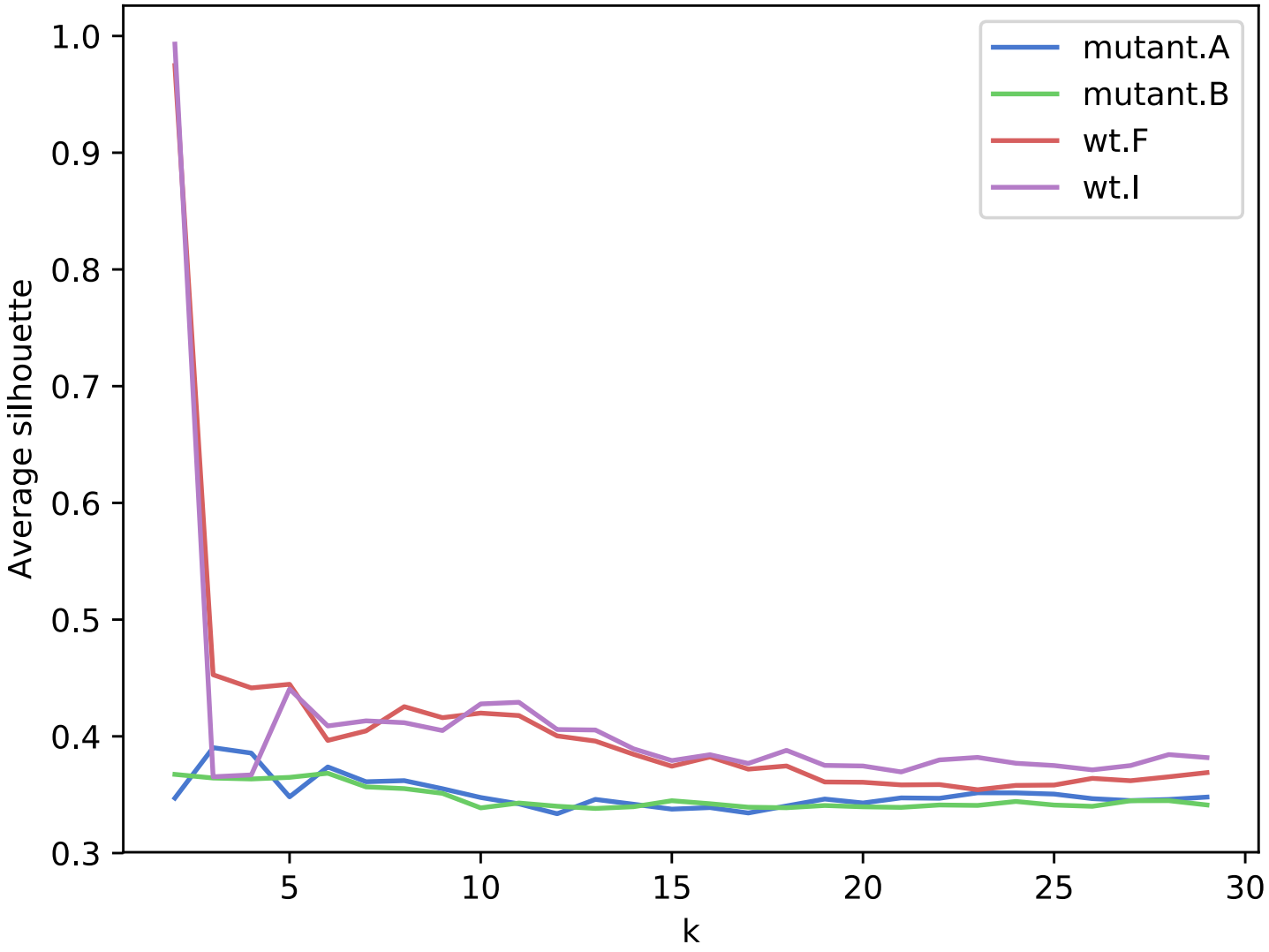


Useless

Obviously, k-means selects these clusters, because we got some sort of a circle



Pca = 4000, tsne = 2



Genes expressed as Gata2 – very conservative approach

Another way to reduce dimension is to do feature selection, by choosing genes, which have similar expression to Gata2. Namely, we would like to select genes which have no significant expression between samples of same condition and are as significant different between conditions as Gata2.

Procedure:

Apply 6 Wilcoxon tests for every gene and select corresponding p-values:
(1) Within condition: wt.A vs wt.B, mutant.F vs mutant.I
(2) Between condition: wt.A vs mutant.F, wt.A vs mutant.I, wt.B vs mutant.F, wt.F vs mutant.I

We select rep_cut = min(1) for Gata2 and trt_cut = max(2) for Gata2

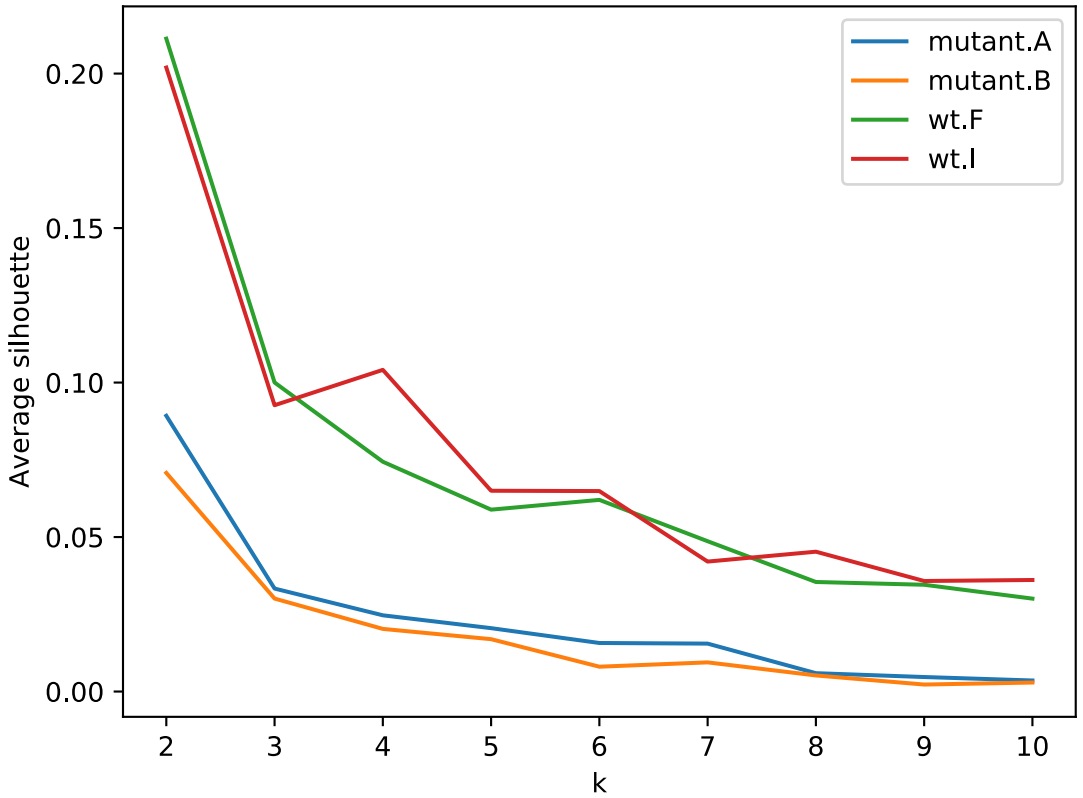
Final list of genes is based on genes with (min(1) > rep_cut and max(2) <= trt_cut)

Result:

Eventually we selected small number of genes

K-means clustering with selected genes – full data, no t-sne or other

We consider k-means clustering with running component from 1 to 10 and constructing an average silhouette measure for every cluster.

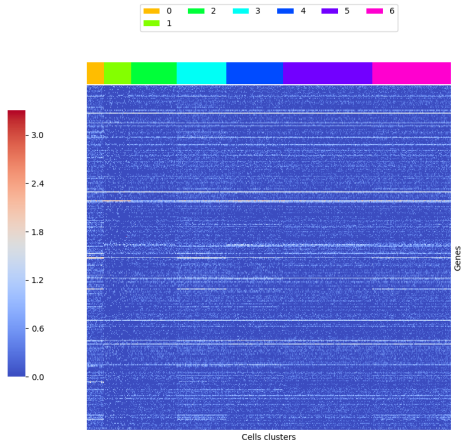


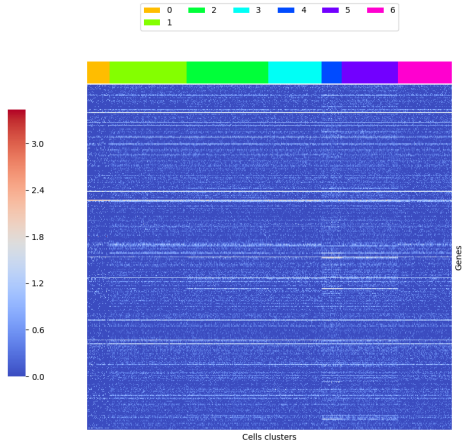
Unfortunately, the highest average silhouette measure corresponds to k = 2, which does not sound right

K-means clustering with selected genes – bad results

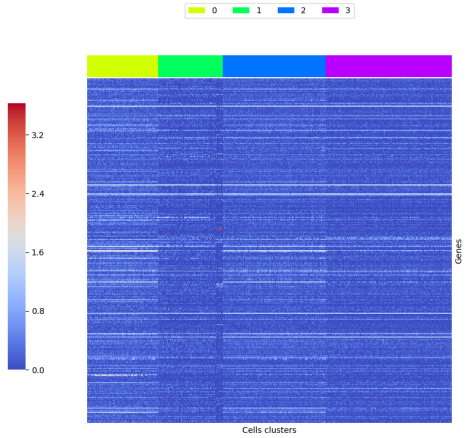
full data, no t-sne or other

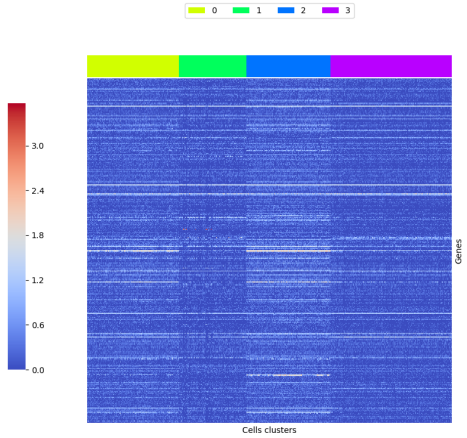
Mutant





Wt



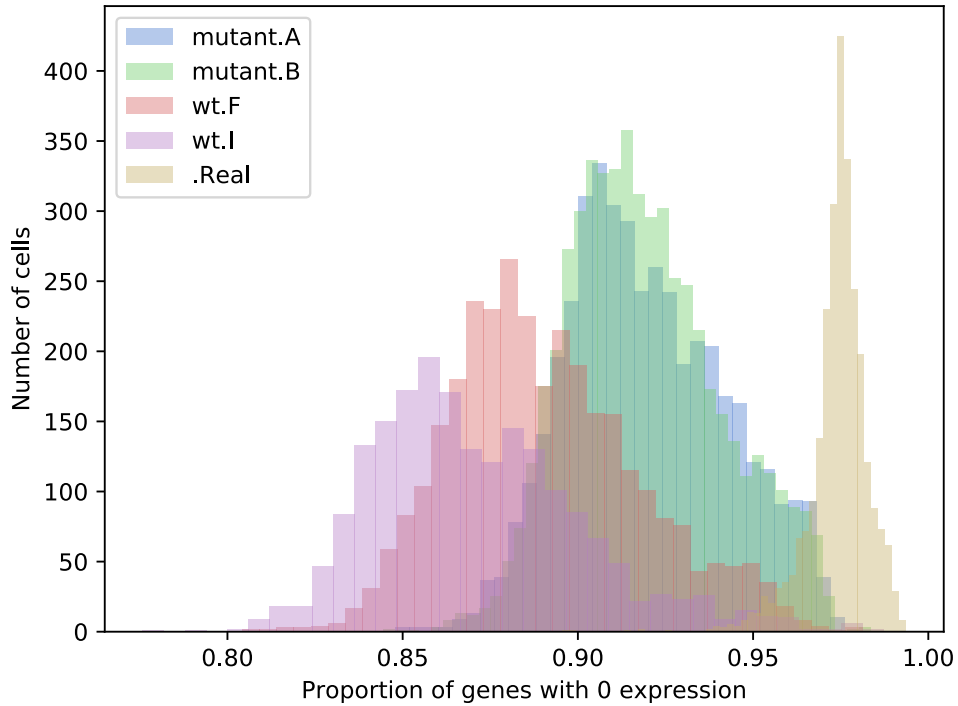


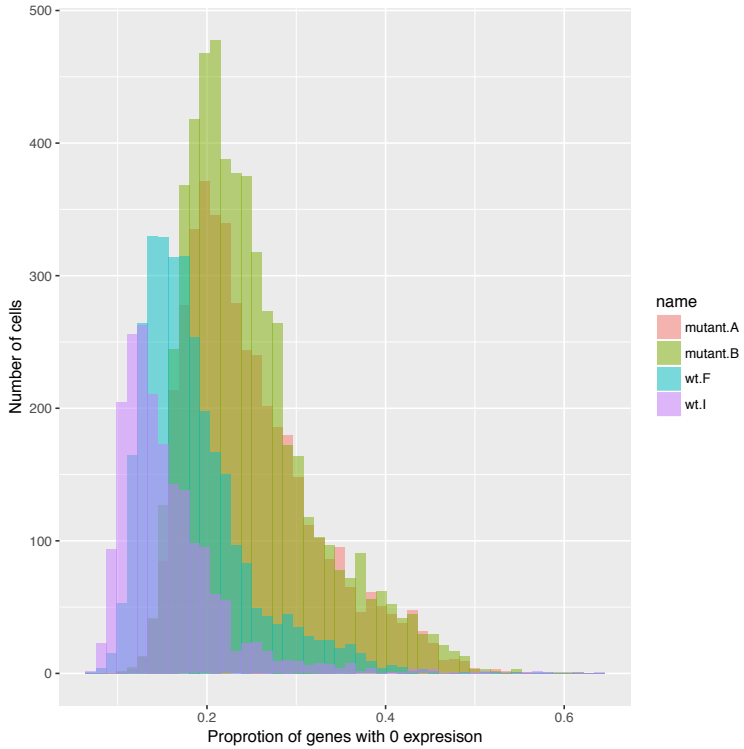
Pathways

Gene corresponds to several pathways. Pathway corresponds to several genes.
For every gene we take all pathways, and copy expression values to corresponding pathways. Then we aggregate pathways with sum and a mean (Sum provides slightly better results in clustering).

We get pathways based using R-package “gskb” and “mm_pathway“ pathways :
https://bioconductor.org/packages/release/data/experiment/html/gskb.html

Proportion of 0-expressed genes is much smaller

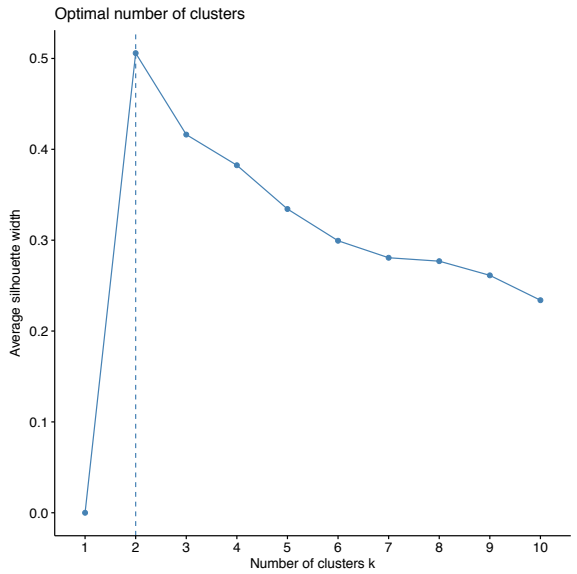


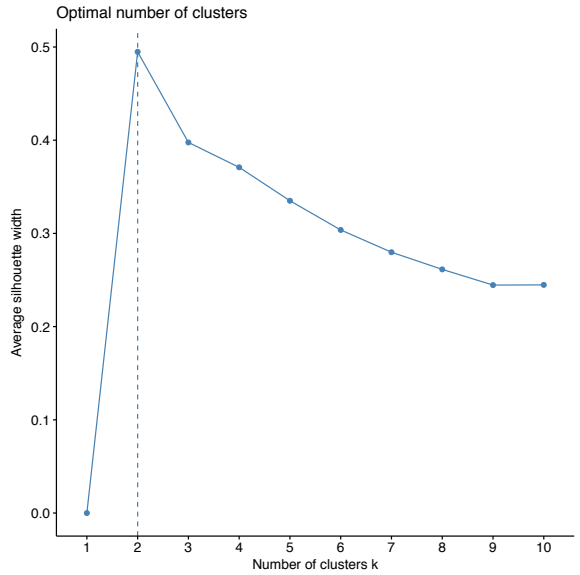


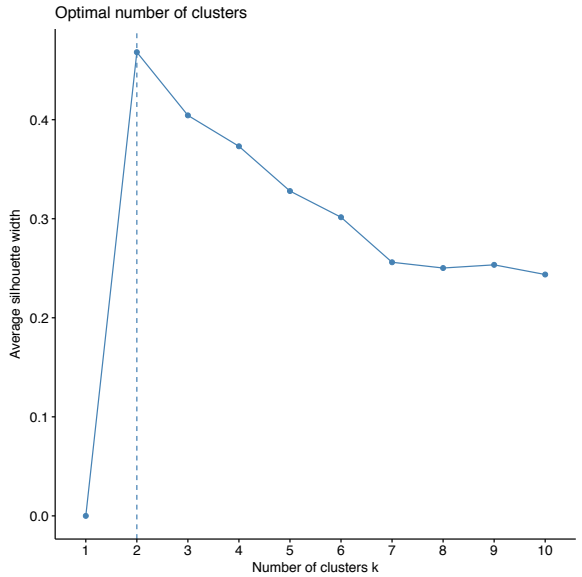
Mean and sum here are the same

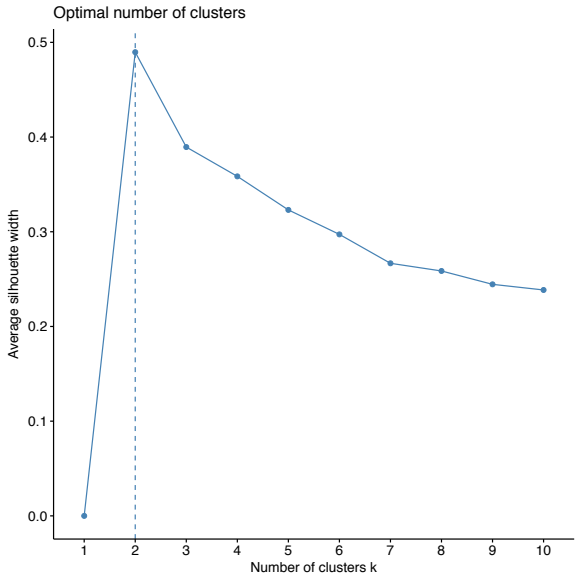
Clustering comparison

Pathways sum

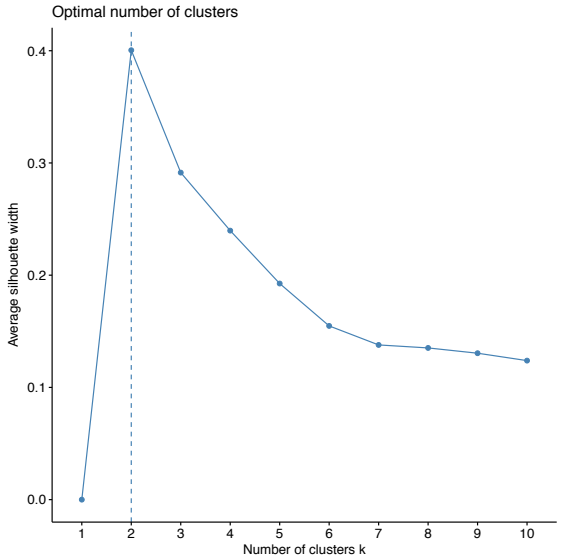


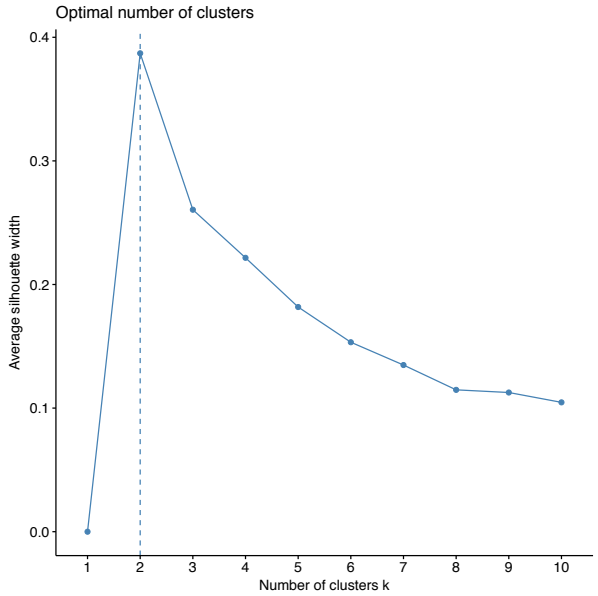


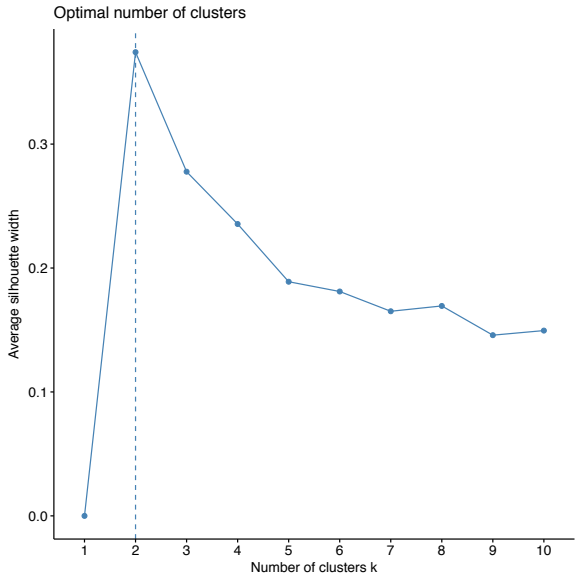


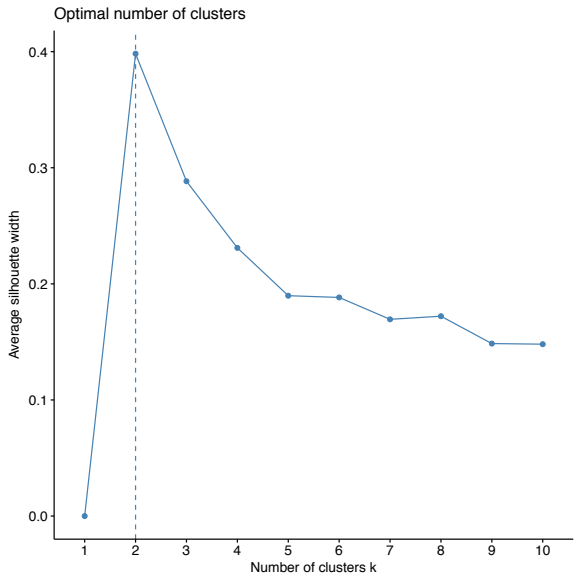


Pathways mean



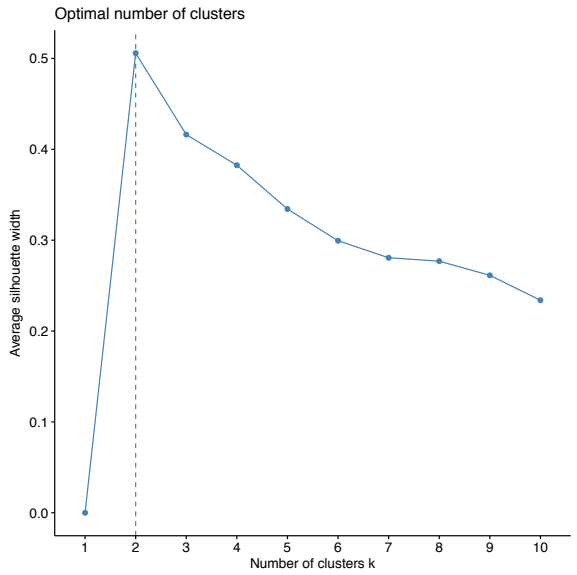


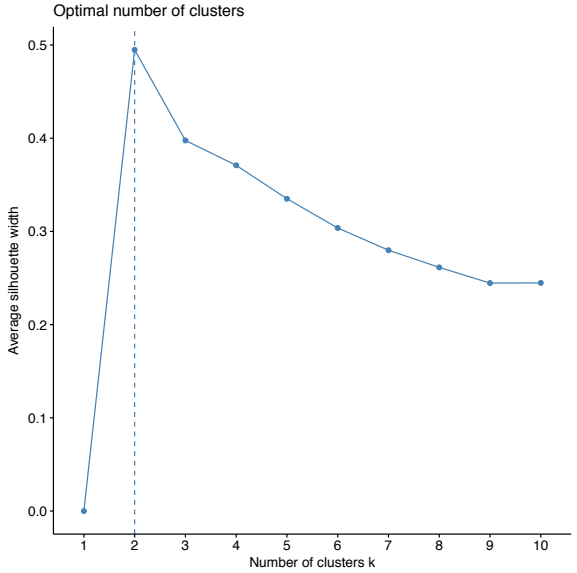


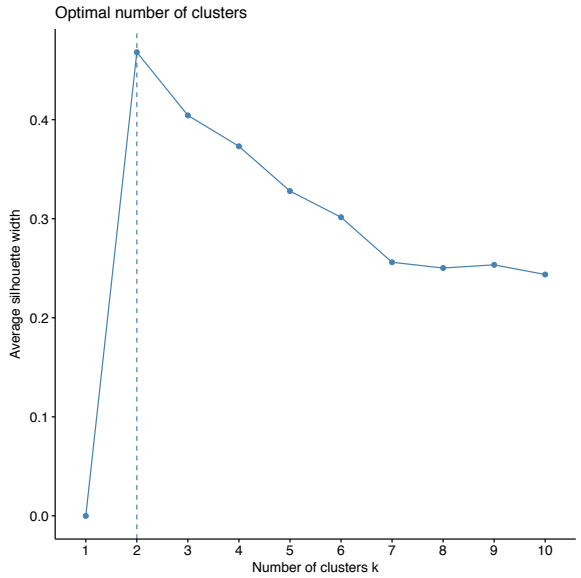


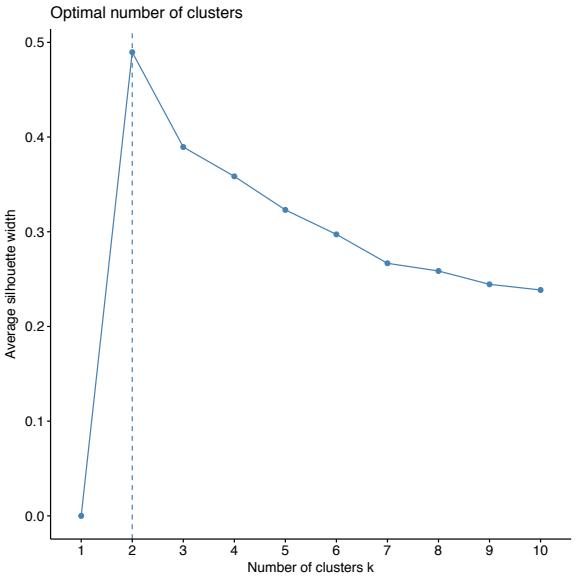
Clustering comparison

Pathways sum

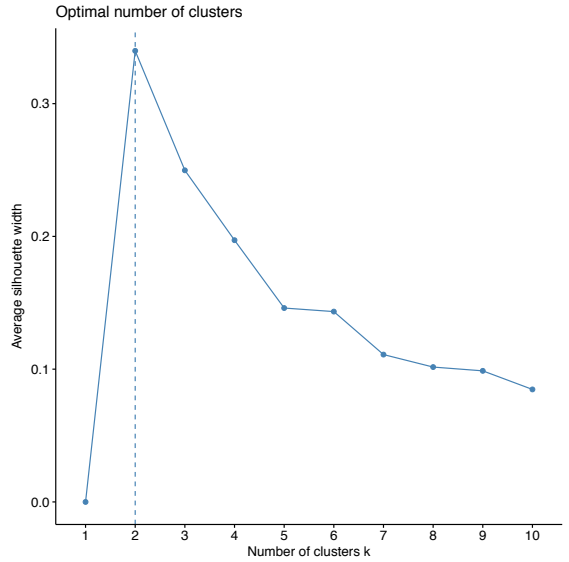


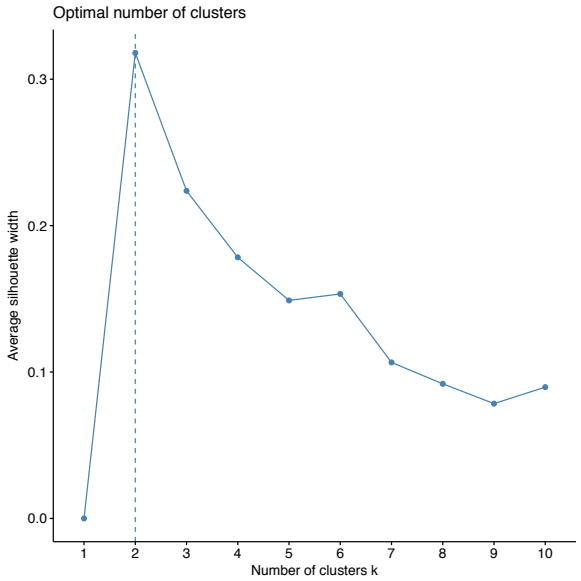


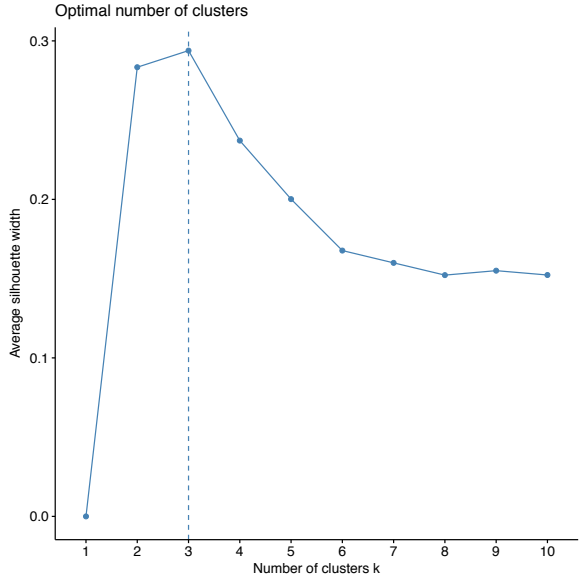


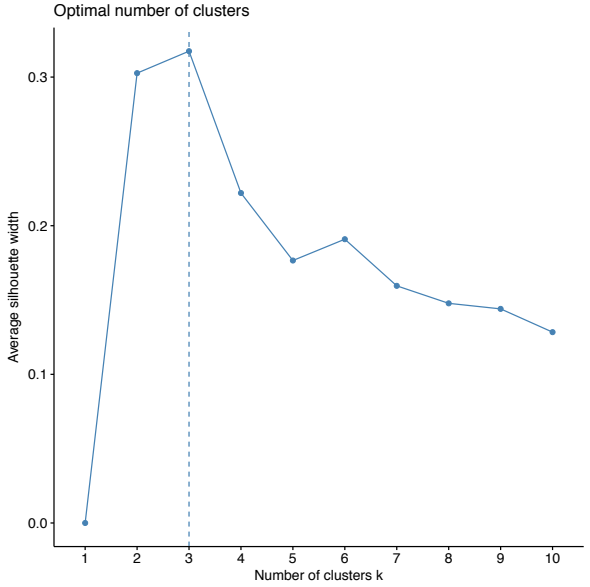


Raw data

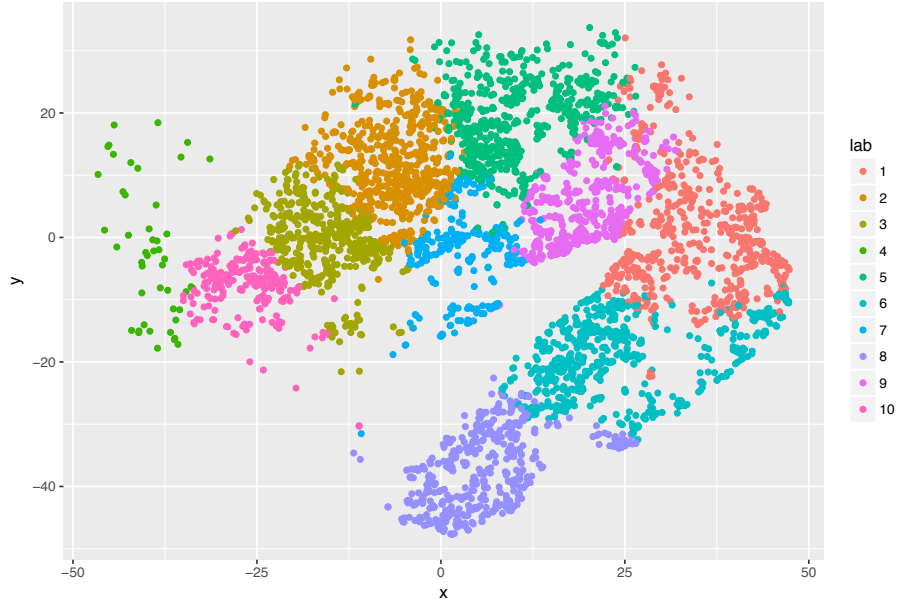








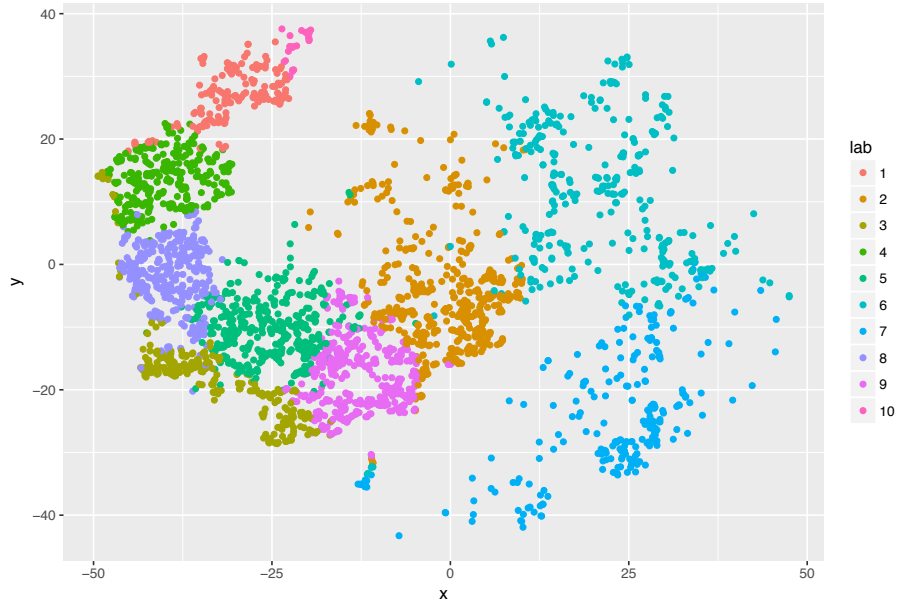
Clustering of pathways sum data and plotting labels on t-sne data



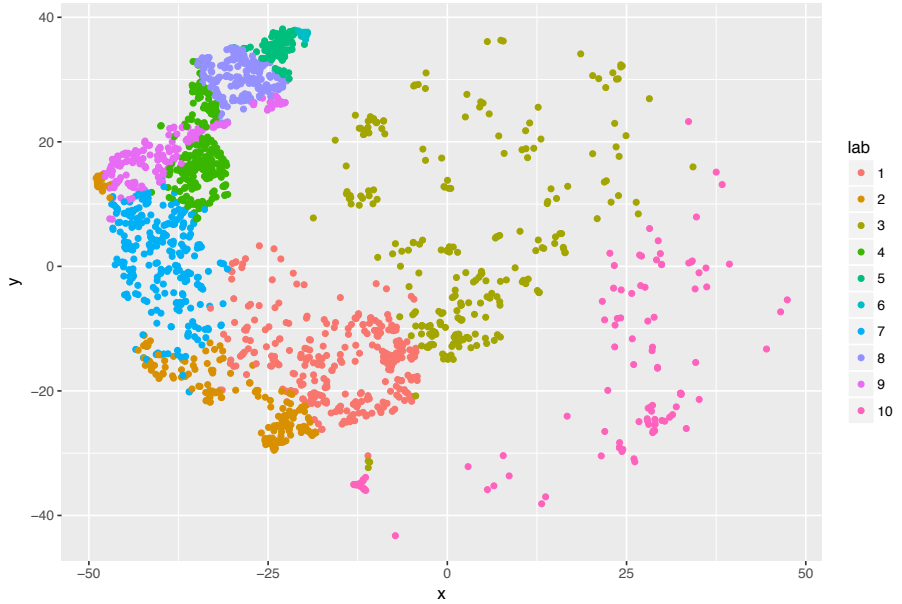
Mutant.A

Mutant.B

K –means and t-sne give same representation of clusters

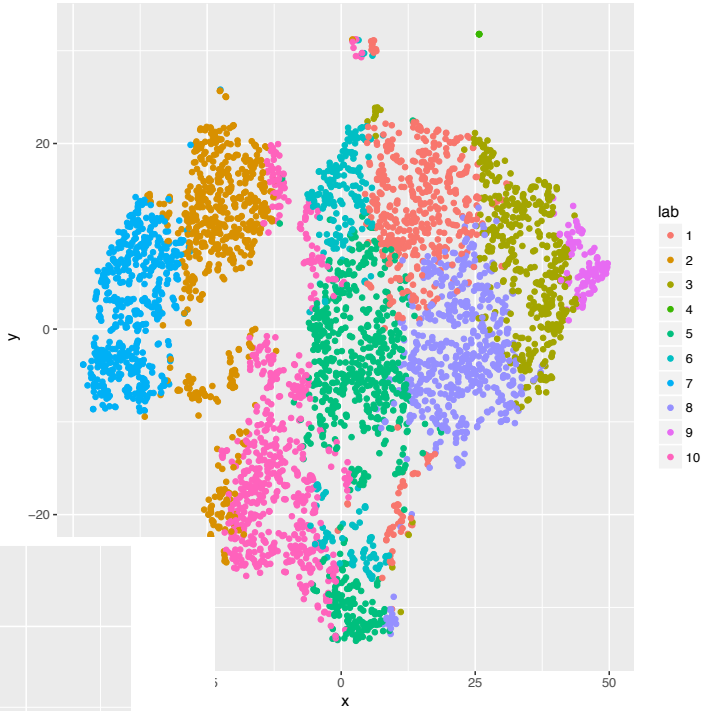


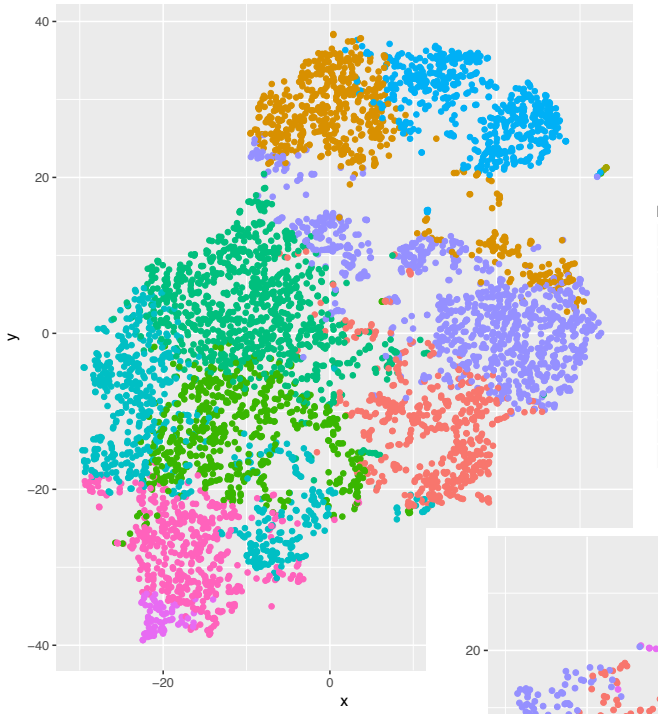
Wt.F

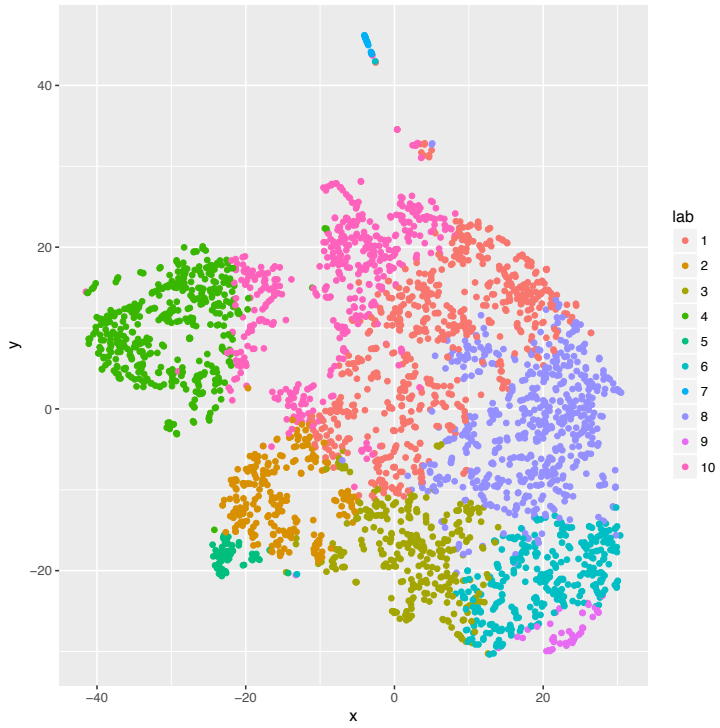


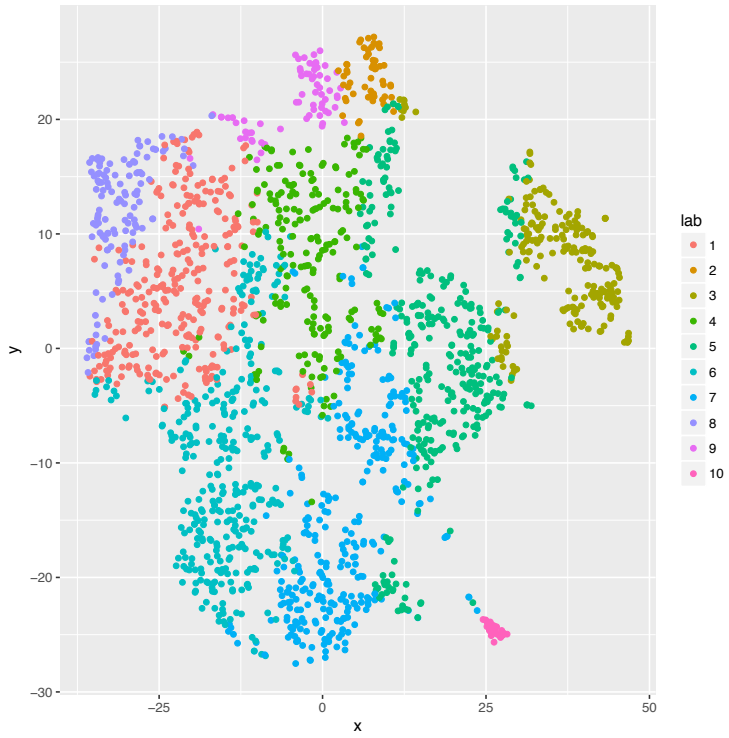
Wt.I

Clustering of raw data and plotting labels on t-sne data - old







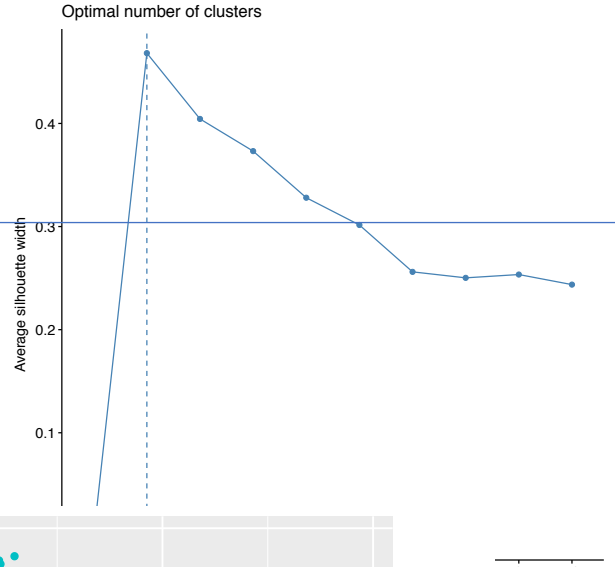


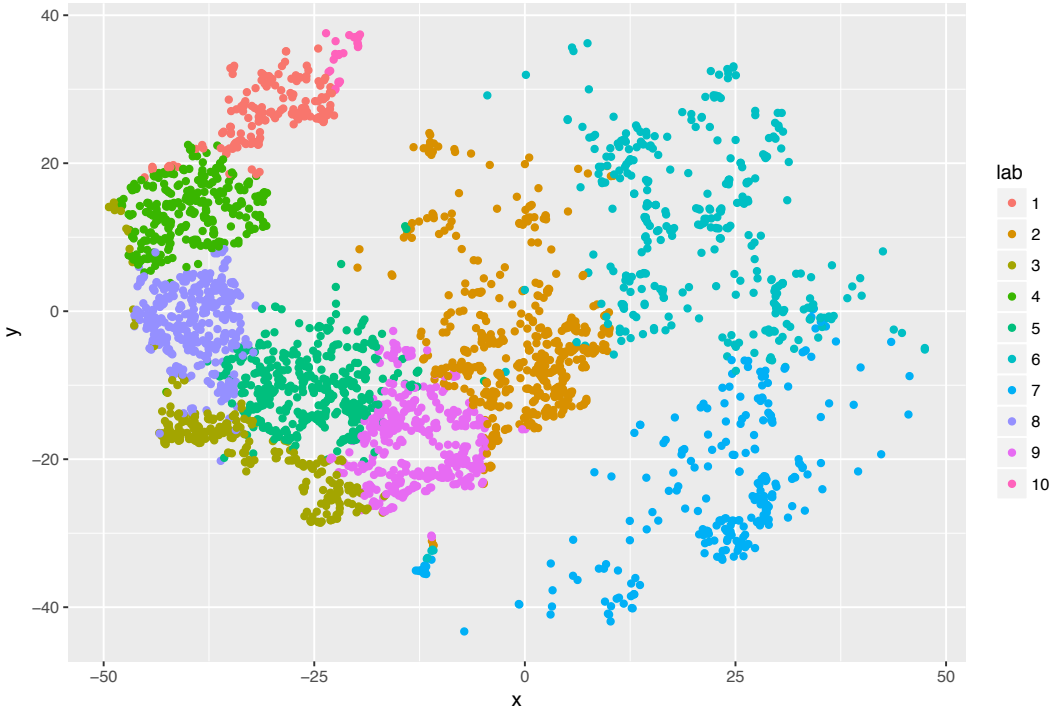
Imputation. MAGIC

Pathways is better than MAGIC

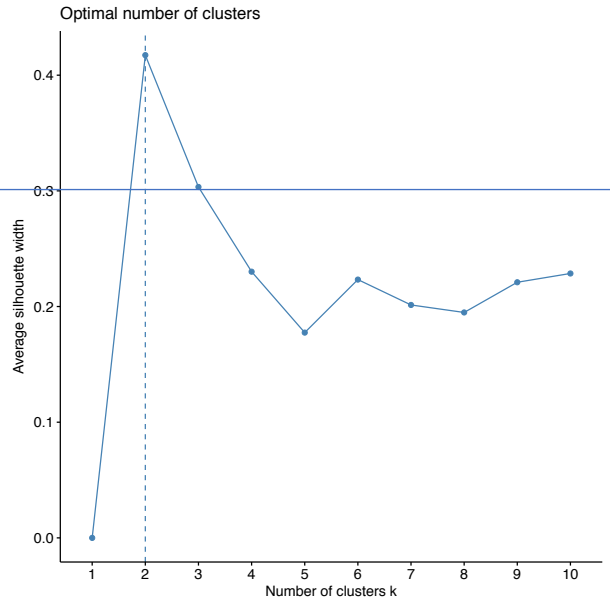
Unfortunately, MAGIC worked just with wt.F. So, compare result for imputation and pathways

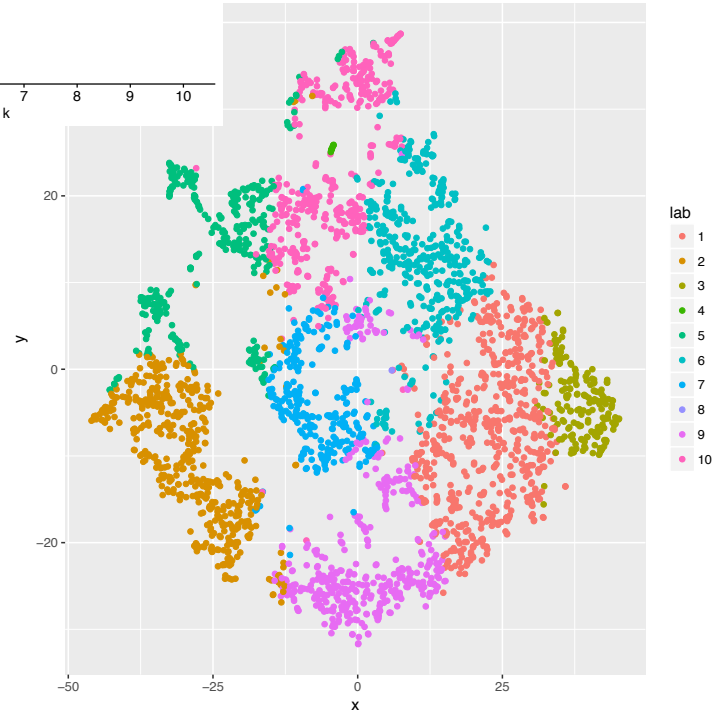
Pathways





MAGIC on raw data

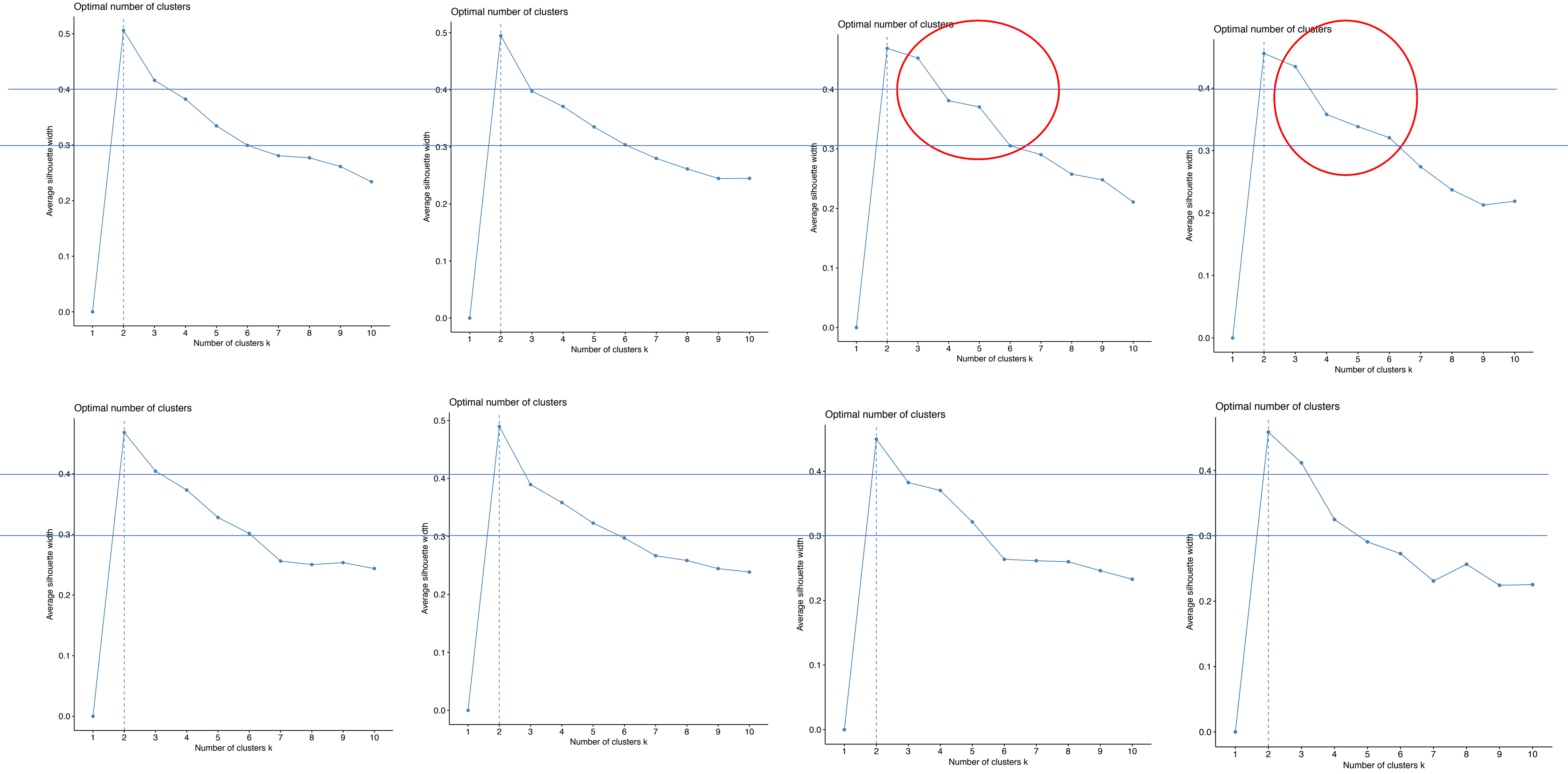




Imputation of sum pathways

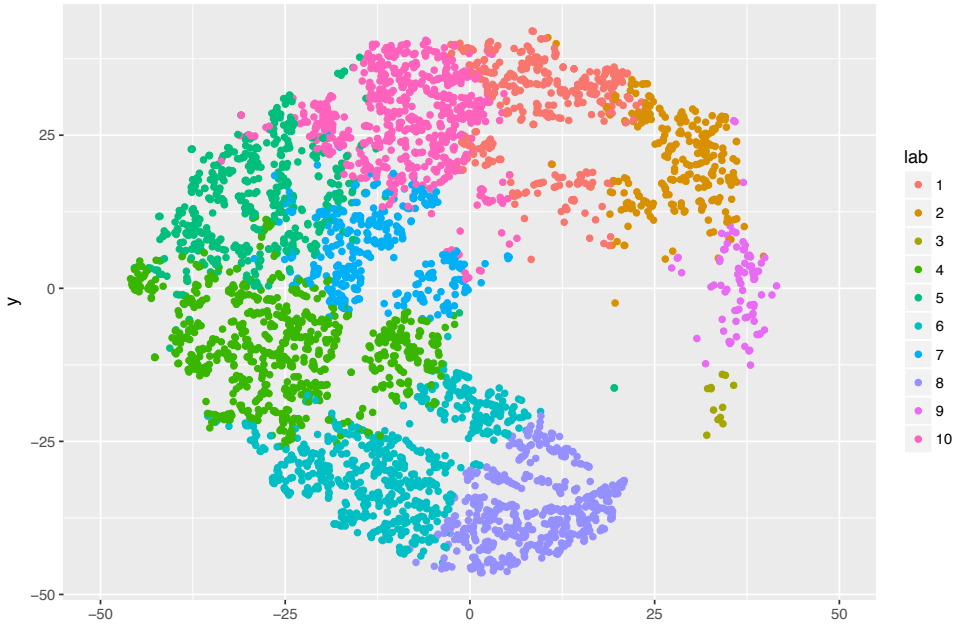
Pathways sum

Imputed data on pathways

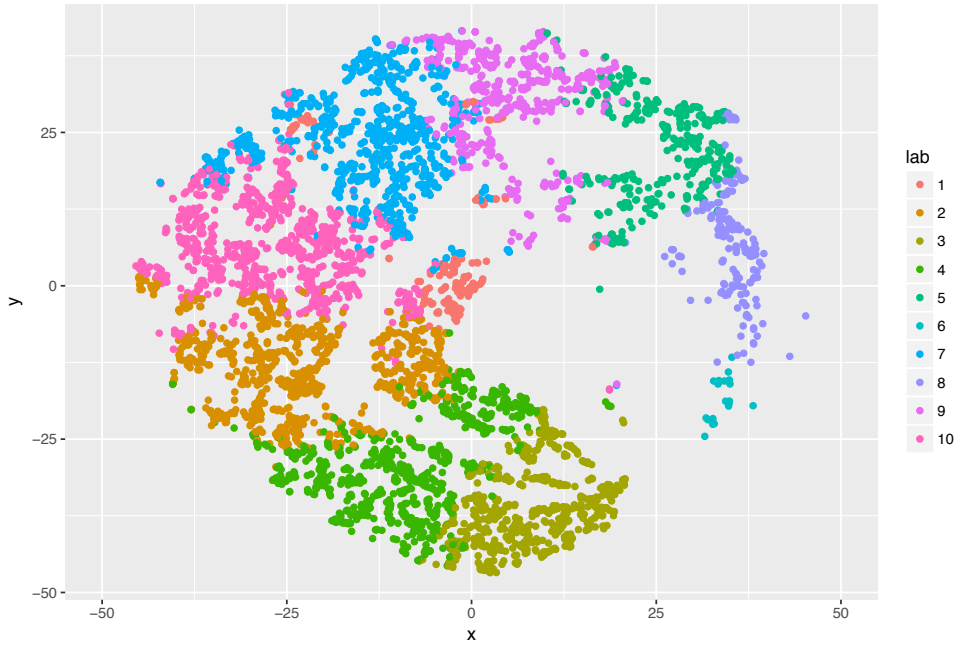


Clustering of MAGIC pathways and plotting labels on t-sne data

Mutant.A



Mutant.B



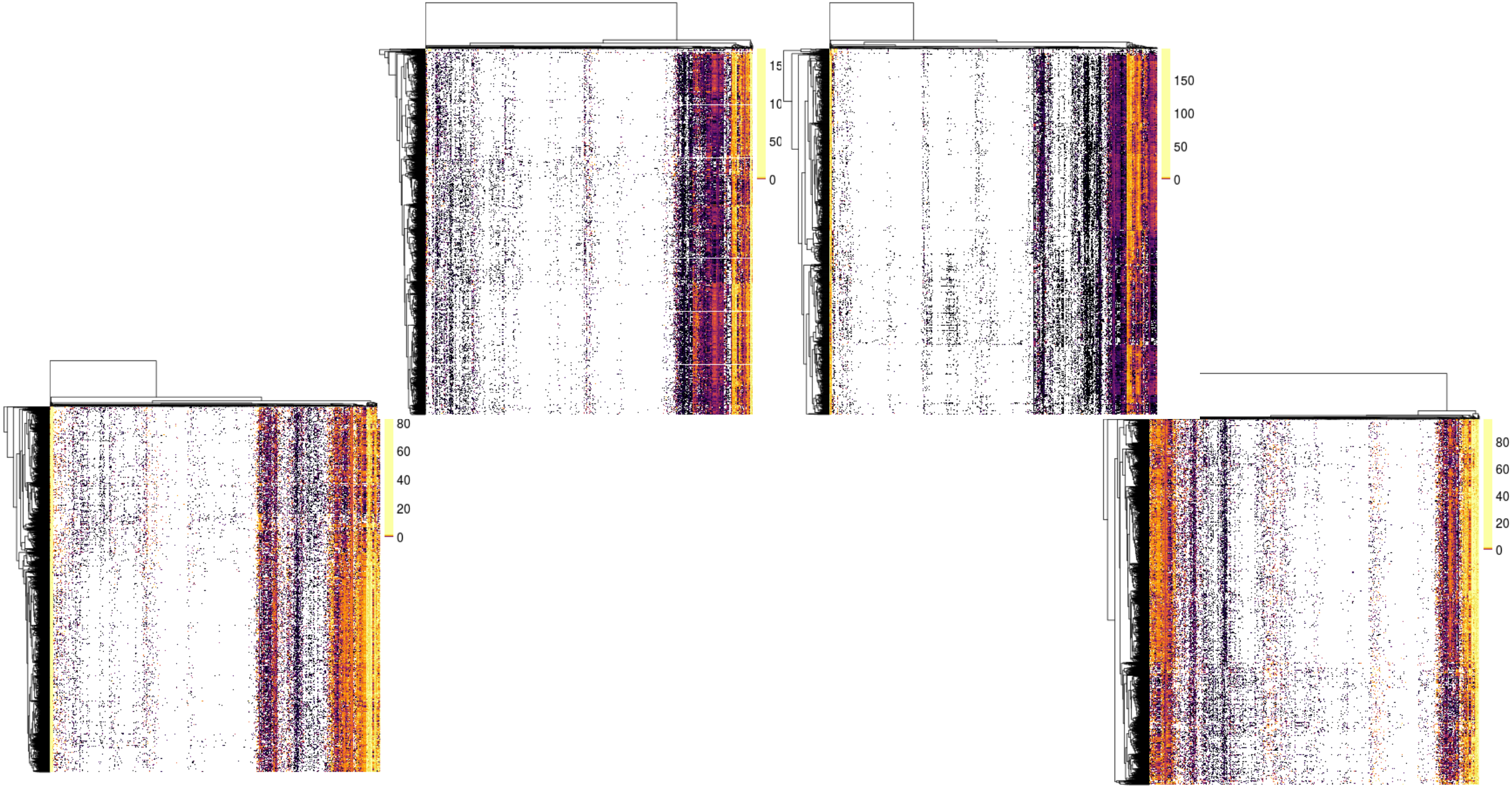
Wt.F

Wt.I

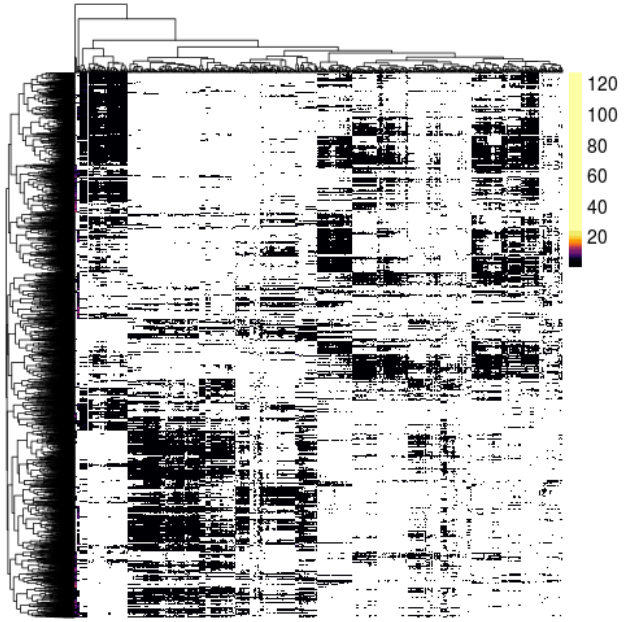
Results: pathways vs MAGIC vs pathways + MAGIC

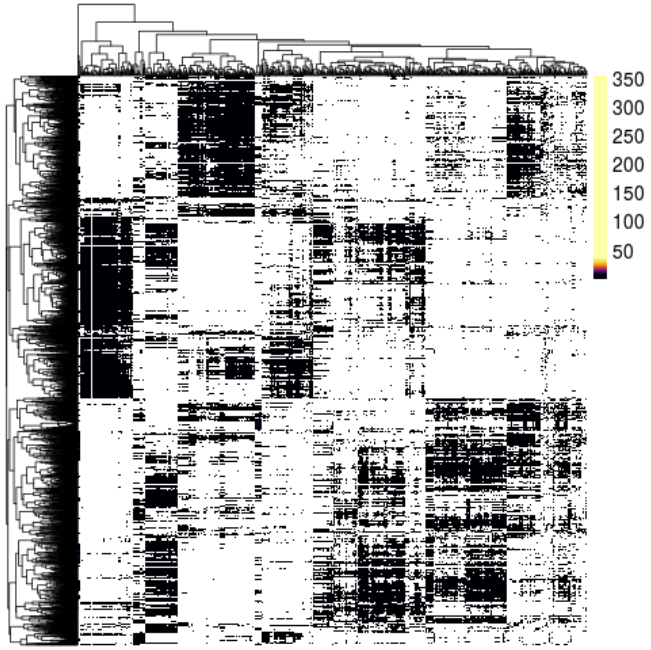
Pathways > MAGIC
Pathways Sum > Pathways Mean
Pathway + MAGIC ? Pathways

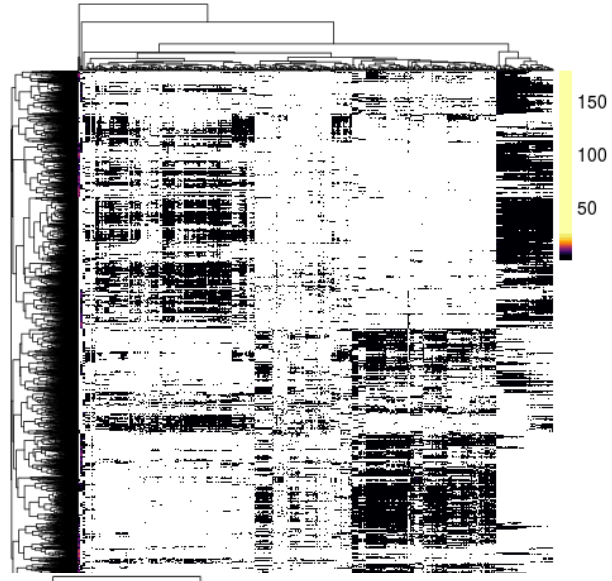
Pathway clustering rows and columns

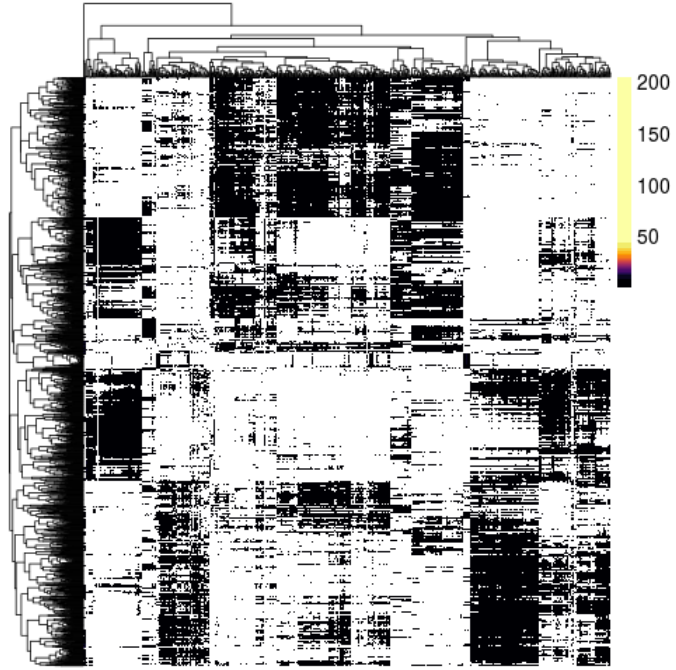


Pathway MAGIC clustering rows and columns









Following slides are appendix

Distance in clustering

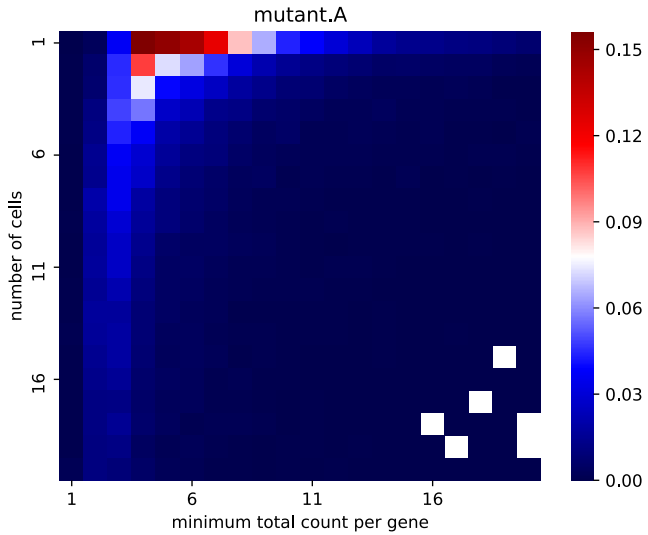
PCA
K means with pca
Zinbw method
Tsne do we need normalization?

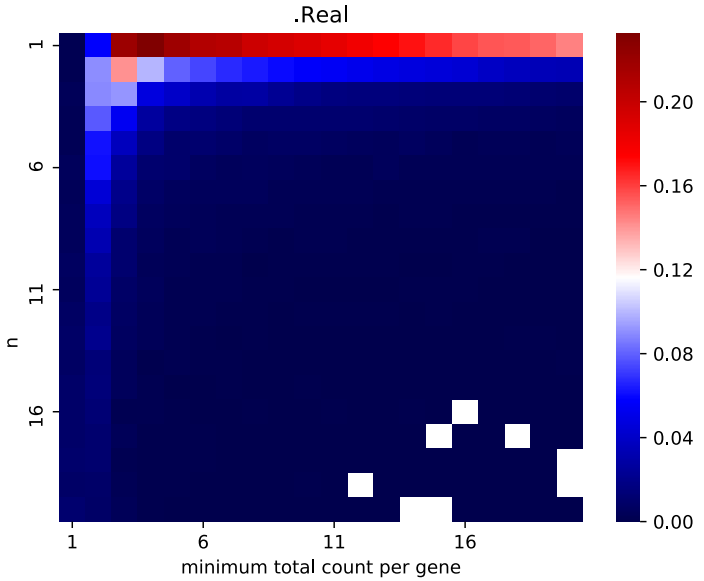
Identify de genes and then cluster based on these genes

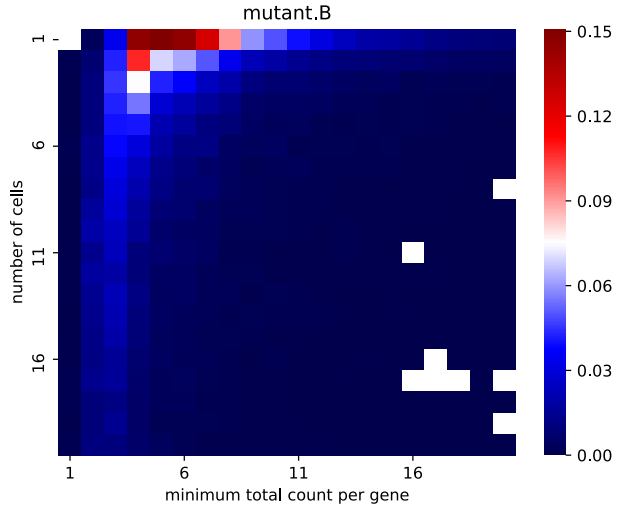
Sparse ct paper

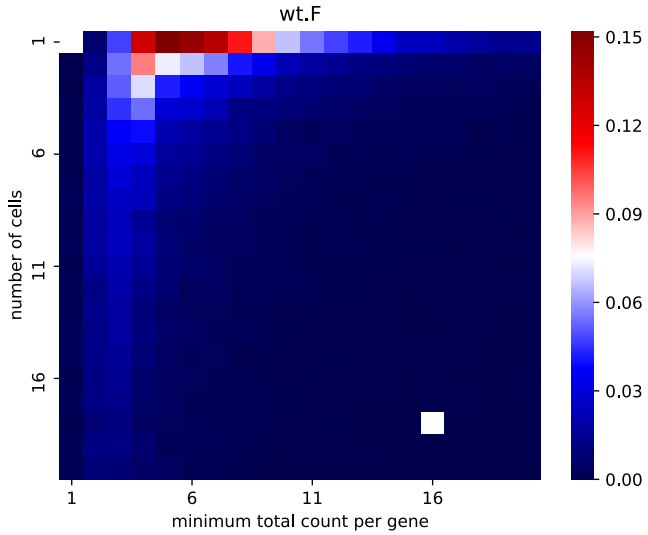
Given 2 condition of data how to identify clusters
And genes which drive to these clusters

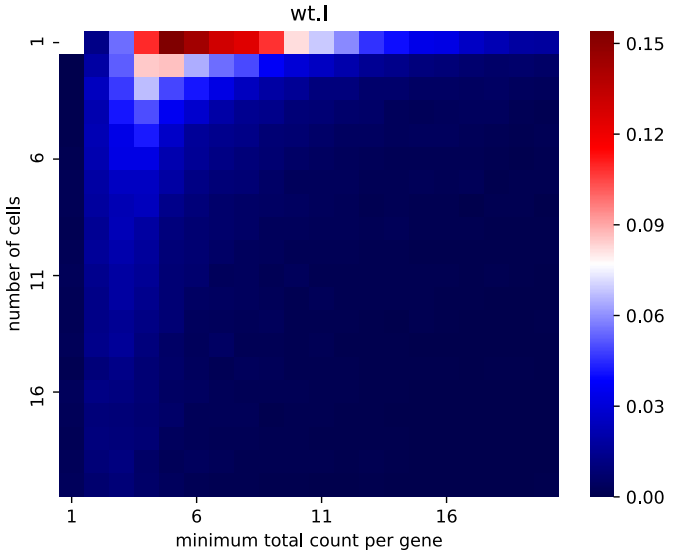
Distribution of percentage of genes with minimal counts







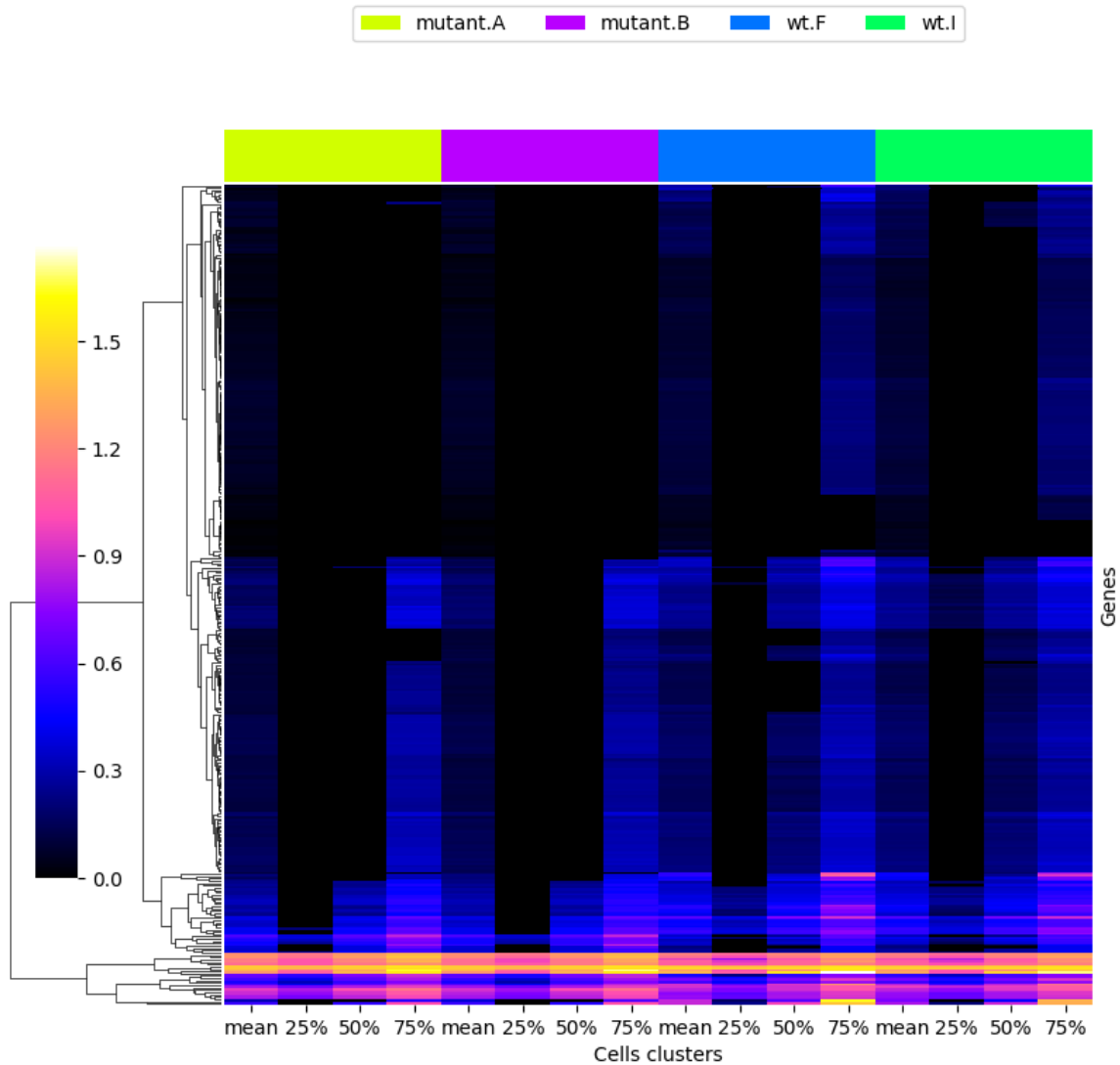




Filtered data

Distribution of gene expression across different samples

We try to see how different are features of distribution between samples



Fixed rows – genes (for every sample same order of rows) and now cluster cells