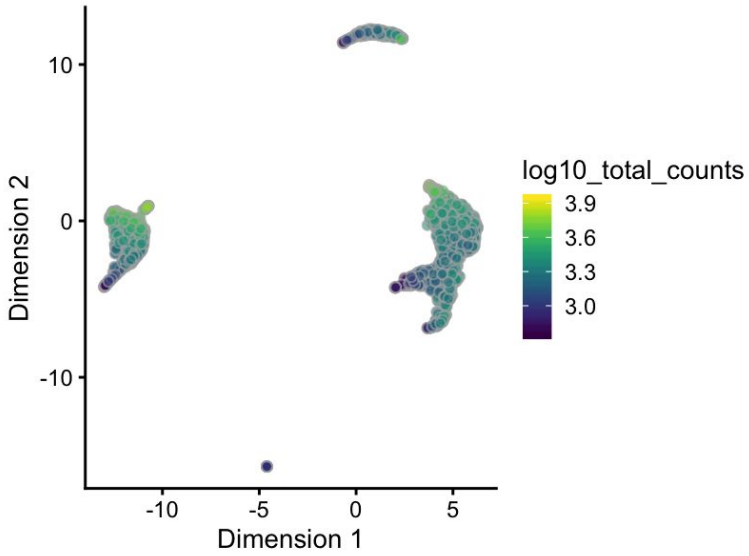
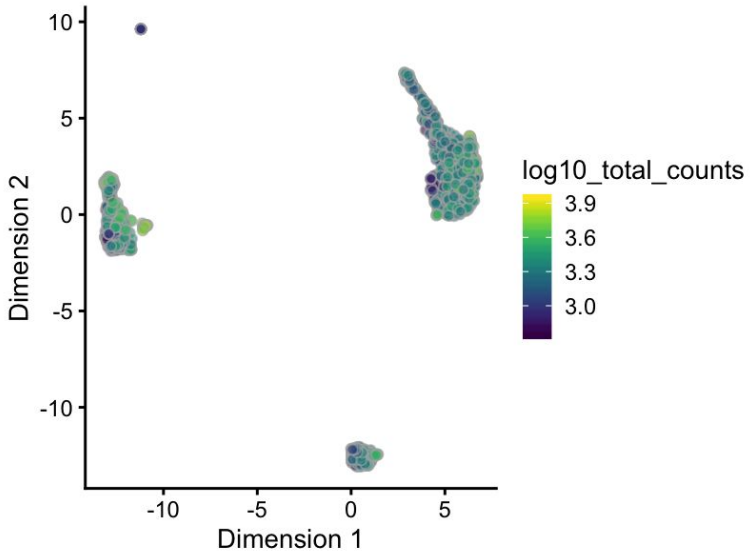
NormJam: Controls

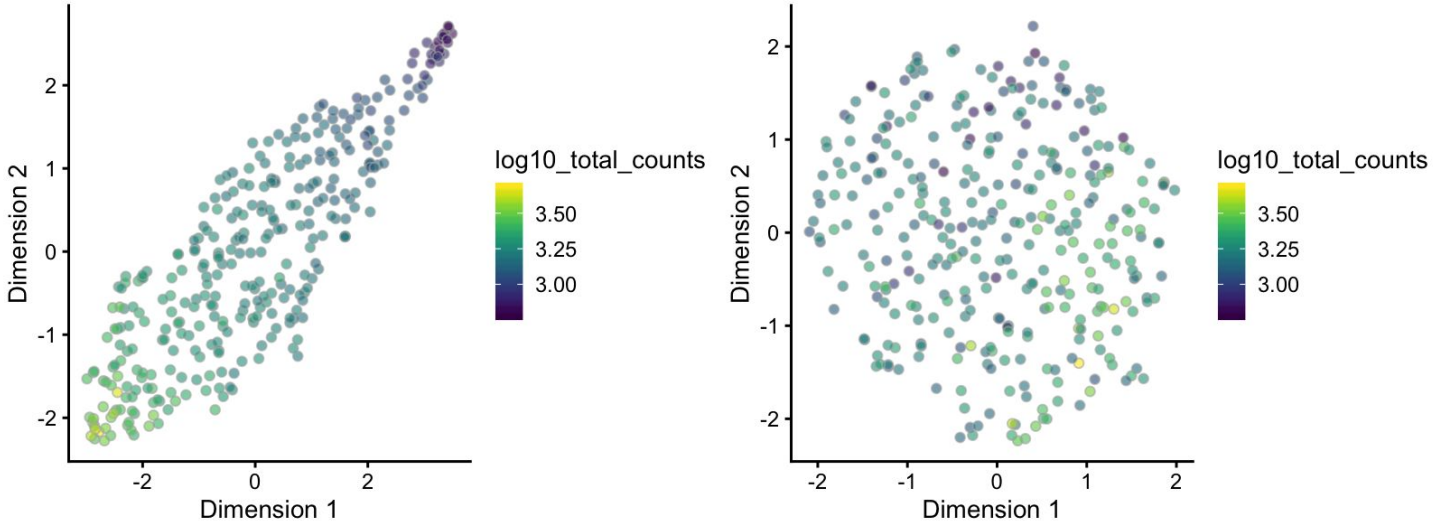
Why normalize?

With no norm, good clusters with false structures.

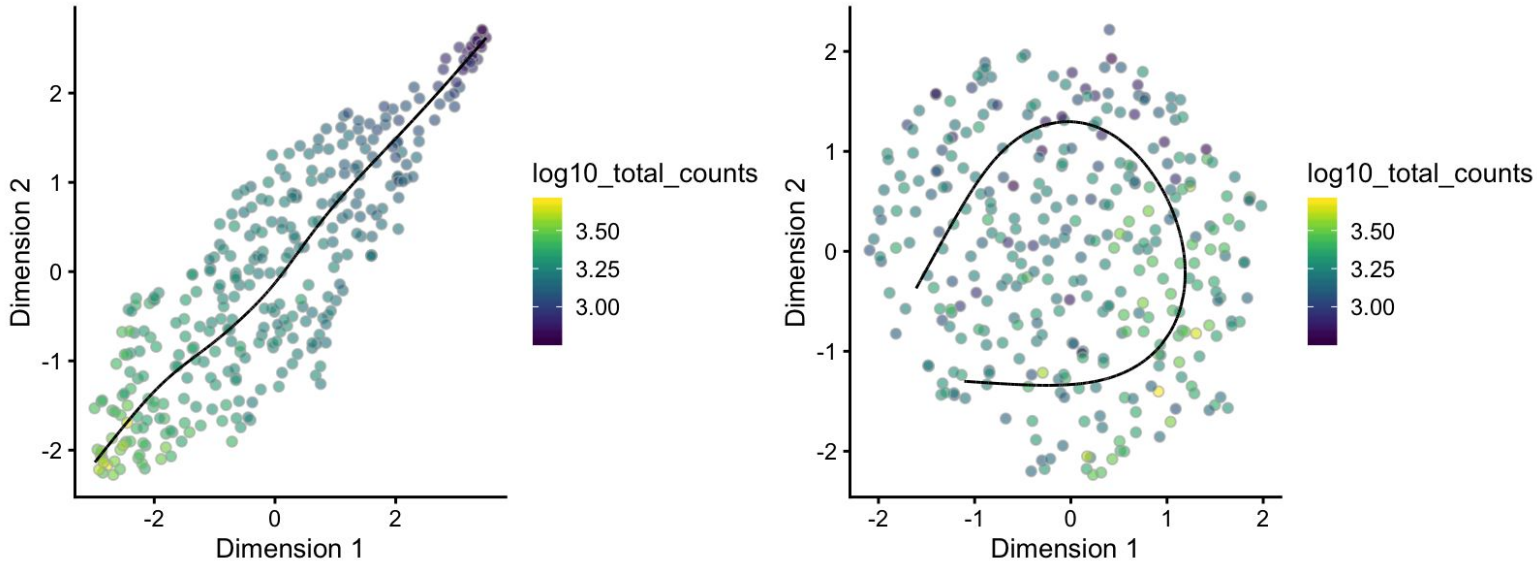




B-cells


Dimension 2
2
1
0
-1
-2
-2
0
2
Dimension 1
log10_total_counts
3.50
3.25
3.00
Dimension 2
2
1
0
-1
-2
-2
-1
0
1
2
Dimension 1
log10_total_counts
3.50
3.25
3.00

B-cells show artificial trajectory

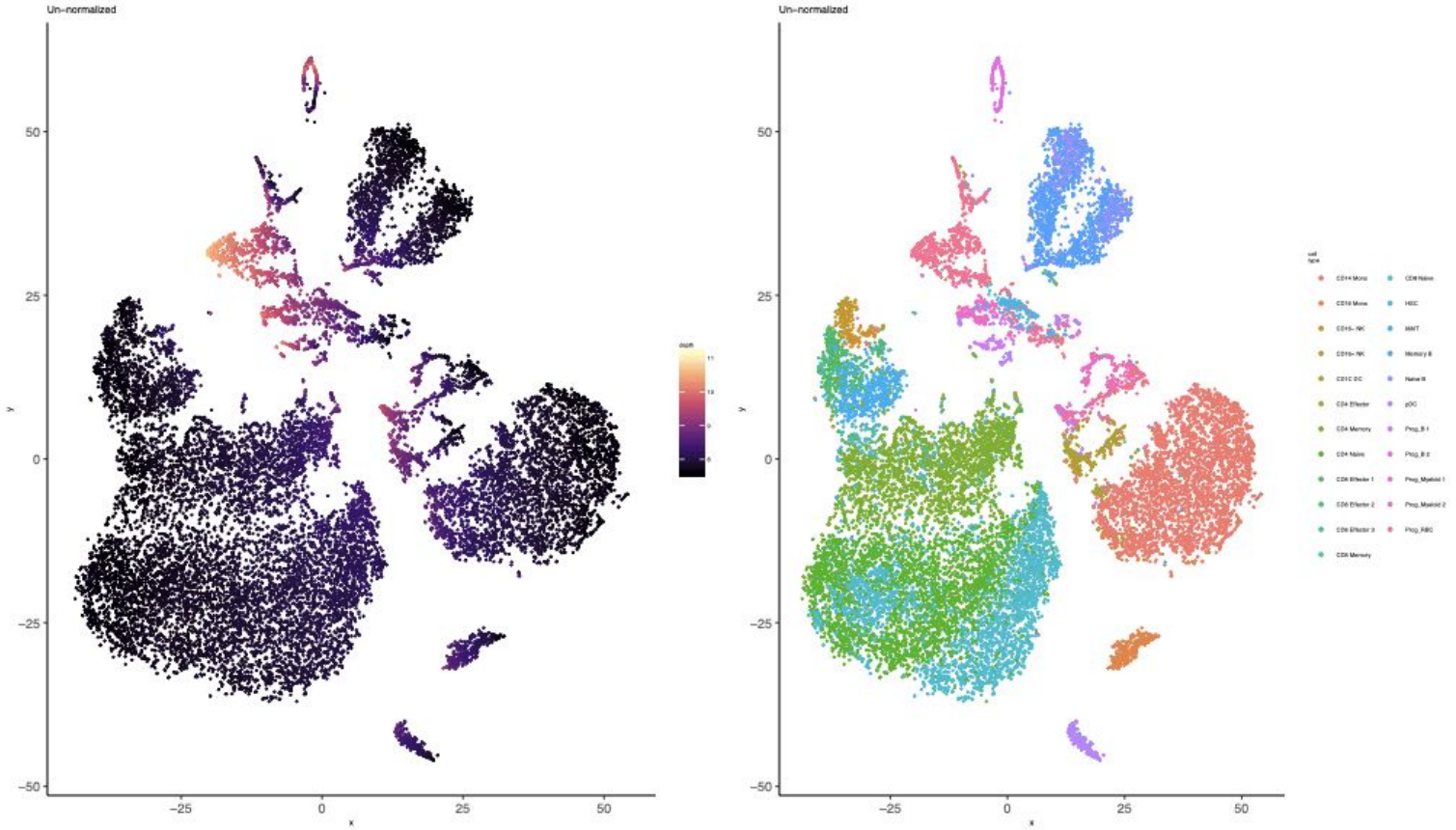


Enrichment using MSigDB

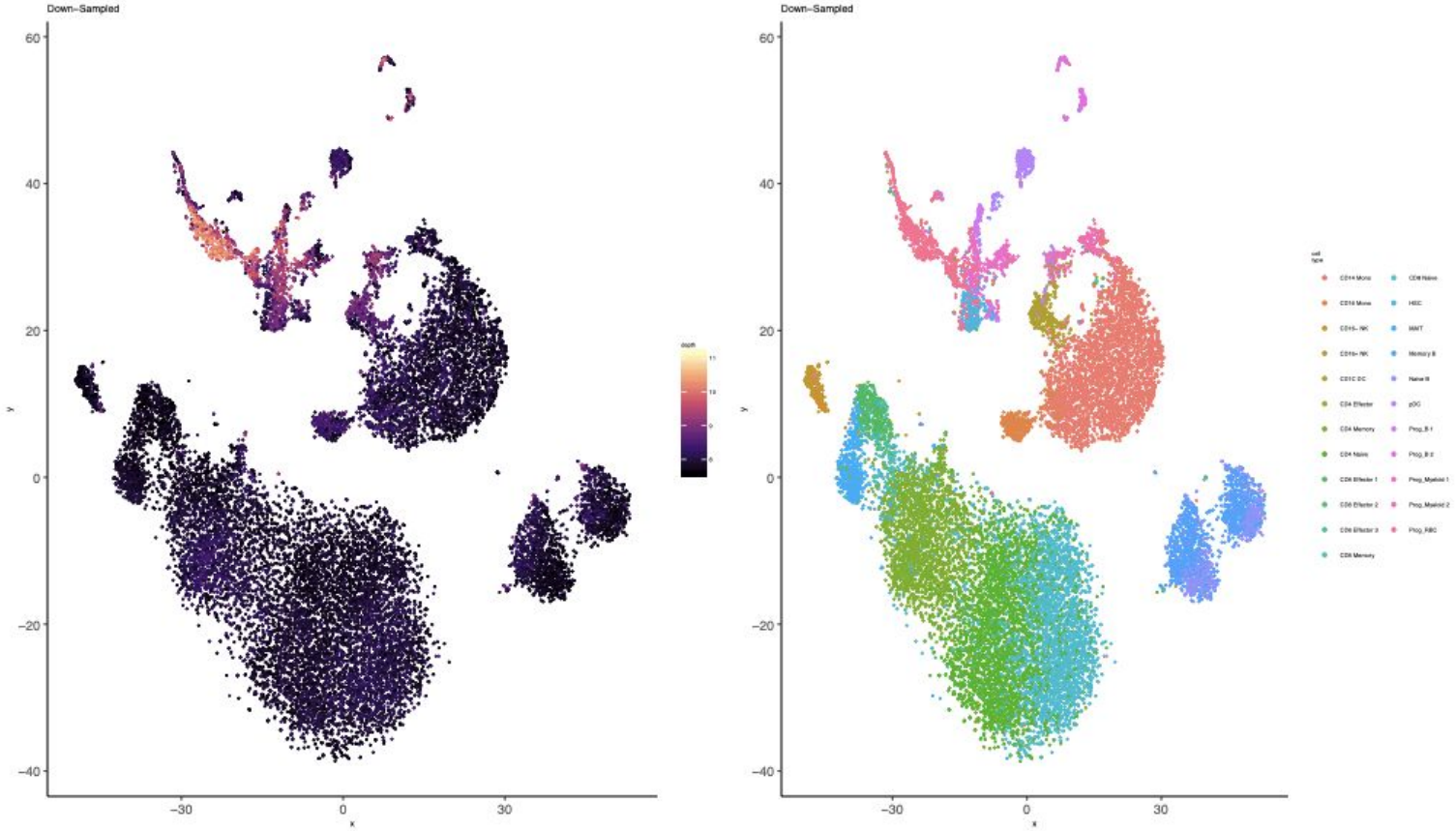
Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
GO_COTRANSLATIONAL_PROTEIN_TARGETING_TG_TO_MEMBRANE [104]	The targeting of proteins to a membrane that occurs during translation. The transport of most secretory proteins, particularly those with more than 100 amino acids, into the endoplasmic reticulum lumen occurs in this manner, as does the import of some proteins into mitochondria. [ISBN:0716731363, PMID:10512867, PMID:16896215]	9		$5.71\ e^{-17}$	$5.39\ e^{-13}$
GO_CYTOSOLIC_RIBOSOME [113]	A ribosome located in the cytosol. [GOC:mtg_sensu]	9		$1.24\ e^{-16}$	$5.39\ e^{-13}$
GO_ESTABLISHMENT_OF_PROTEIN_LOCALIZATIATION_TO_ENDOPLASMIC_RETICULUM [117]	The directed movement of a protein to a specific location in the endoplasmic reticulum. [GOC:mah]	9		$1.71\ e^{-16}$	$5.39\ e^{-13}$
GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_IC_PROCESS_NONSENSE_MEDIATED_DECAY [120]	The nonsense-mediated decay pathway for nuclear-transcribed mRNAs degrades mRNAs in which an amino-acid codon has changed to a nonsense codon; this prevents the translation of such mRNAs into truncated, and potentially harmful, proteins. [GOC:krc, GOC:ma, PMID:10025395]	9		$2.16\ e^{-16}$	$5.39\ e^{-13}$
GO_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_MIC_RETICULUM [142]	A process in which a protein is transported to, or maintained in, a location within the endoplasmic reticulum. [GOC:mah]	9		$1.02\ e^{-15}$	$2.04\ e^{-12}$
GO_STRUCTURAL_CONSTITUENT_OF_RIBOSOME [185]	The action of a molecule that contributes to the	9		$1.15\ e^{-14}$	$1.87\ e^{-11}$

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
GO_MHC_PROTEIN_COMPLEX [25]	A transmembrane protein complex composed of an MHC alpha chain and, in most cases, either an MHC class II beta chain or an invariant beta2-microglobin chain, and with or without a bound peptide, lipid, or polysaccharide antigen. [GOC:add, GOC:jl, ISBN:0781735149, PMID:15928678, PMID:16153240]	6		$3.11\ e^{-15}$	$3.11\ e^{-11}$
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY [787]	A series of molecular signals initiated by the binding of a cytokine to a receptor on the surface of a cell, and ending with regulation of a downstream cellular process, e.g. transcription. [GOC:mah, GOC:signaling, PMID:19295629]	11		$1.58\ e^{-13}$	$7.92\ e^{-10}$
GO_RESPONSE_TO_CYTOKINE [1192]	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a cytokine stimulus. [GOC:sl]	12		$4.09\ e^{-13}$	$1.36\ e^{-9}$
GO_ER_TO_GOLGI_TRANSPORT_VESICLE_MEMBRMBRANE [62]	The lipid bilayer surrounding a vesicle transporting substances from the endoplasmic reticulum to the Golgi. [GOC:ai, GOC:ascb_2009,	6		$1.07\ e^{-12}$	$2.66\ e^{-9}$

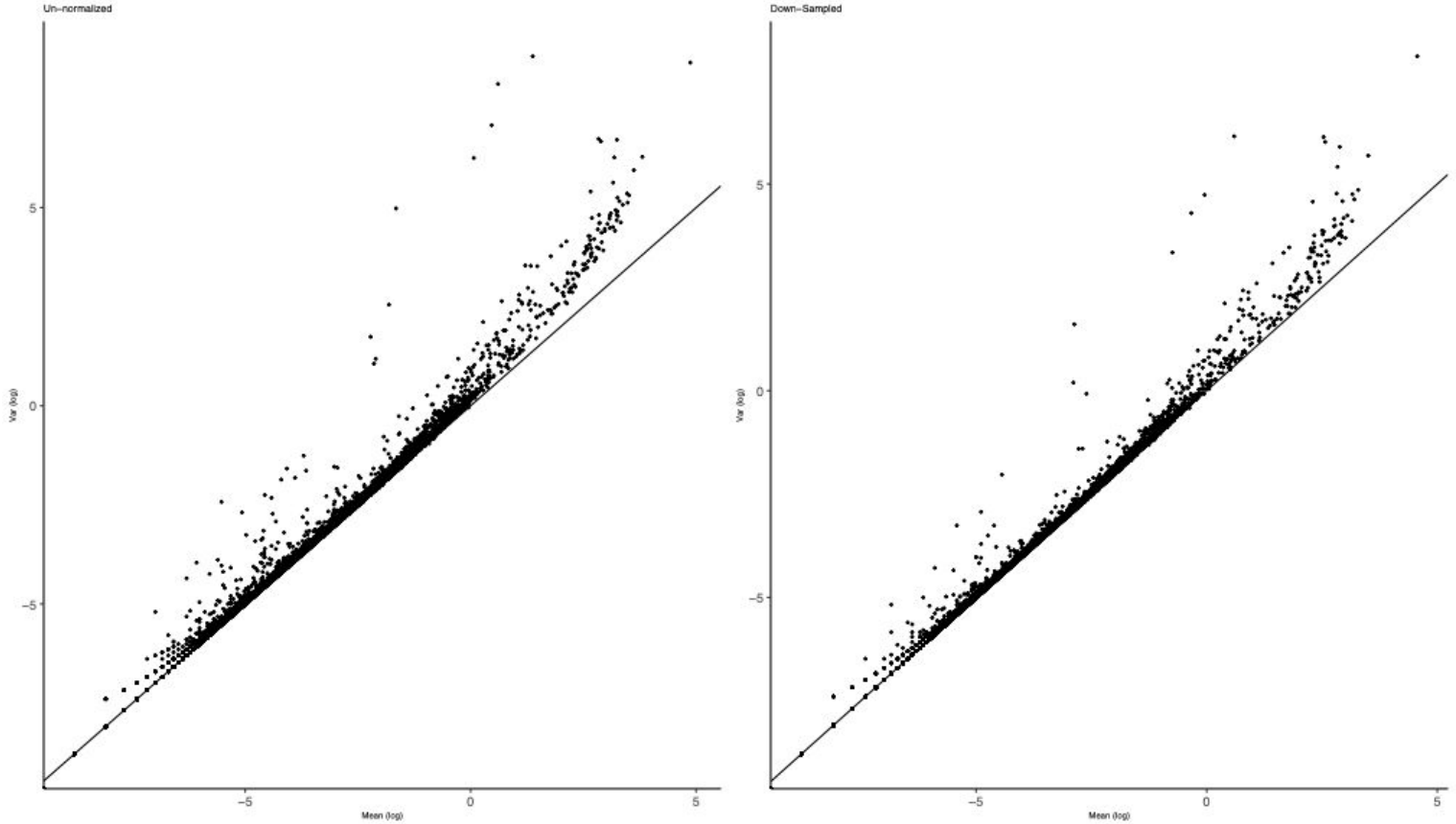
Un-normalized tSNE embedding



Normalized tSNE embedding



Mean-Variance relationship

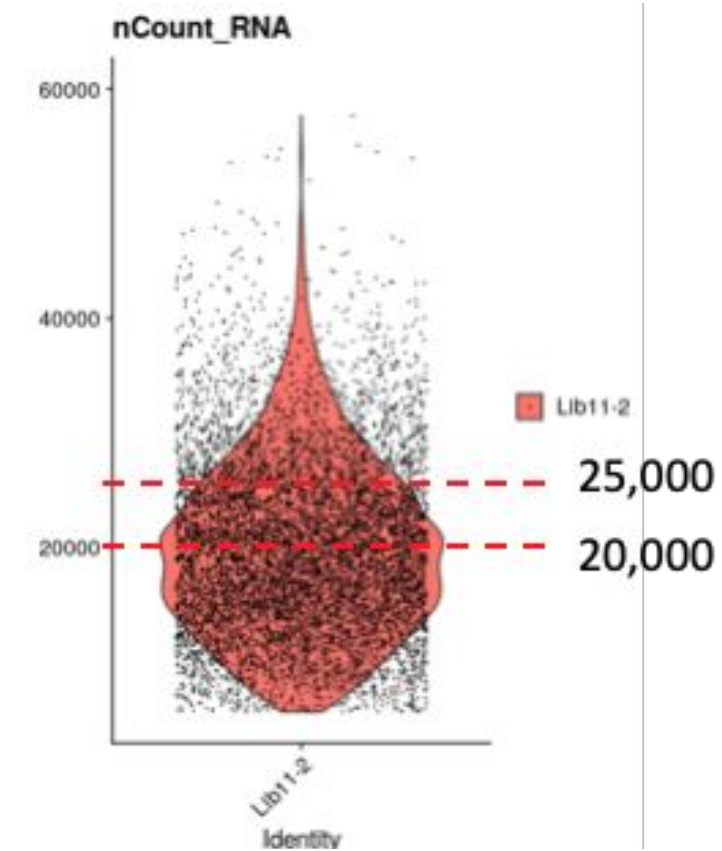


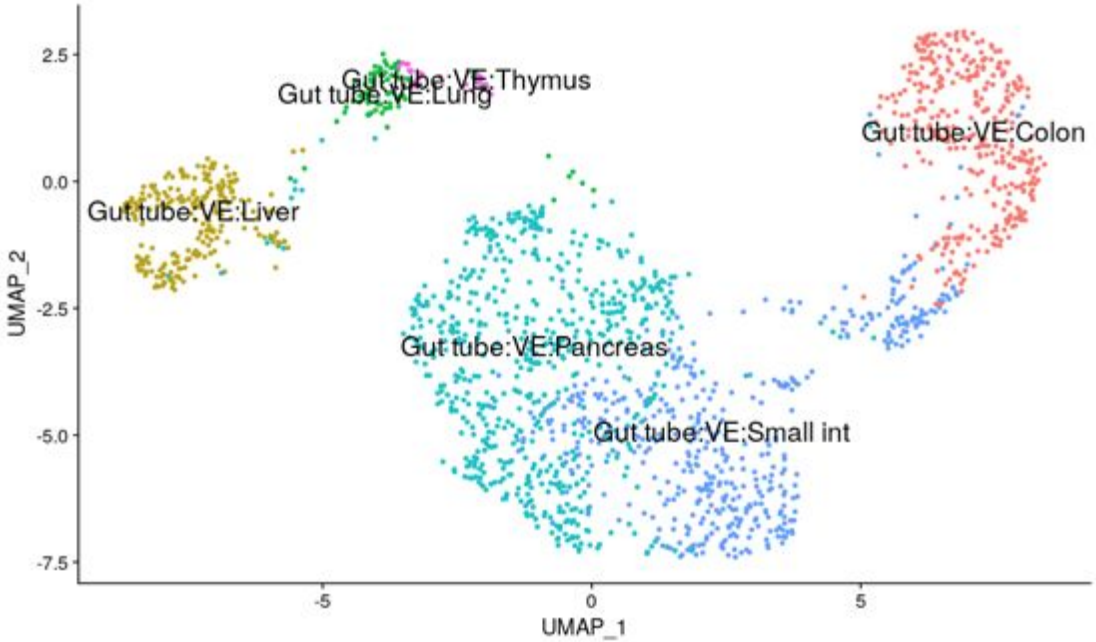
Mean-Variance Relationship

Negative controls show a linear mean-var relationship when conditioned on library size.

Biological data often shows an excess beyond Poisson, even in relatively pure populations.

Control Library size



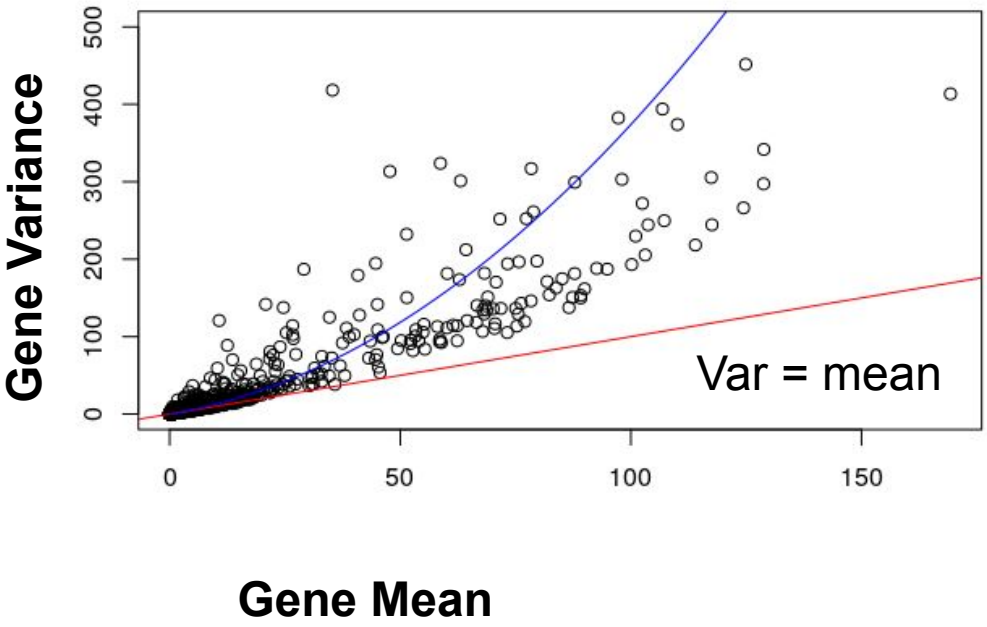


Nowotschin, S., Setty, M., Kuo, Y. Y., Liu, V., Garg, V., Sharma, R., ... & Church, D. M. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature, 569(7756), 361.

Dispersion persists after controlling for library size

Mean variance relation within same sequencing depth Liver cells

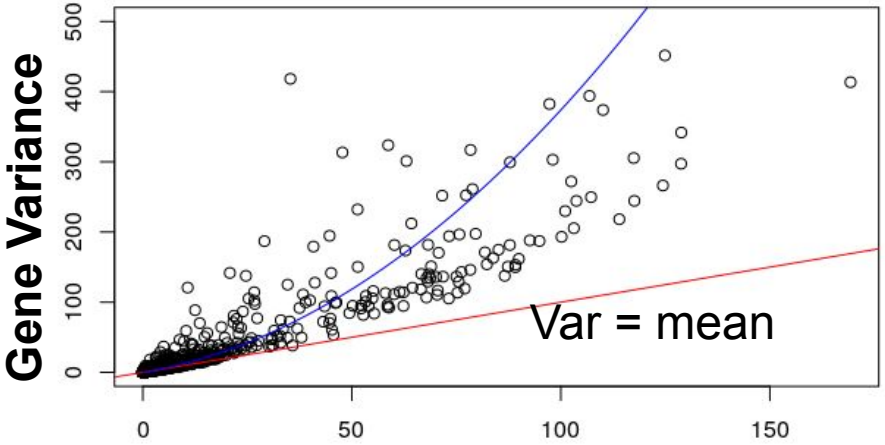
Var = mean + 37 * mean^2



Dispersion persists after controlling for library size

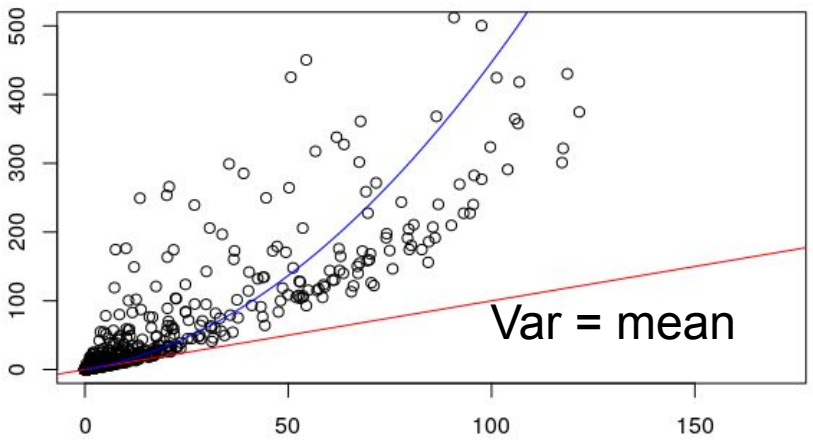
Liver cells

Var = mean + 37 * mean^2



Heterogeneous cells

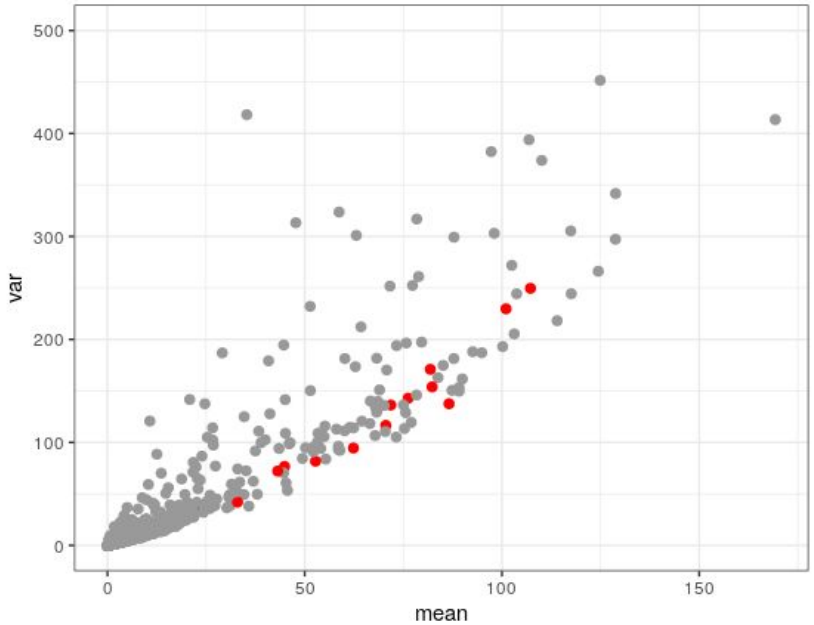
Var = mean + 29 * mean^2

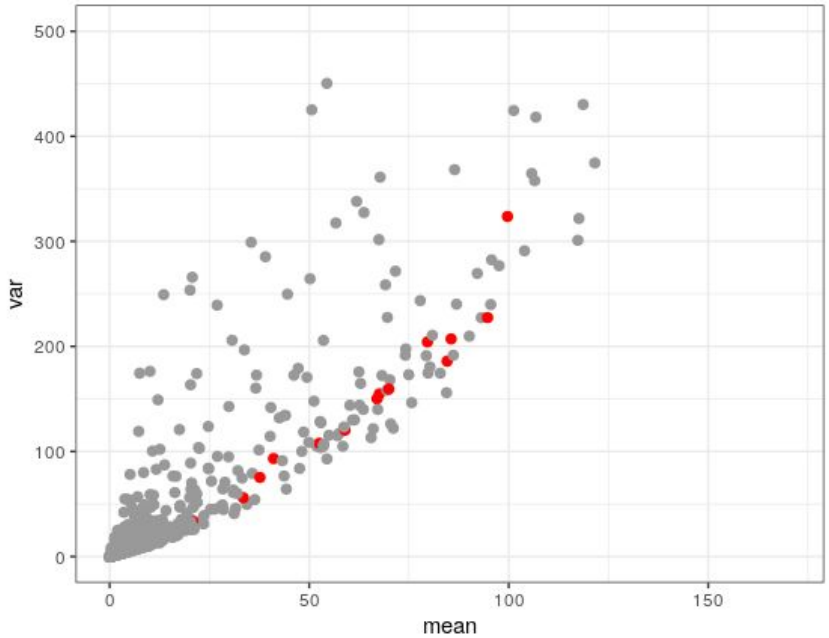


Gene Mean

Genes on the low-variance envelope

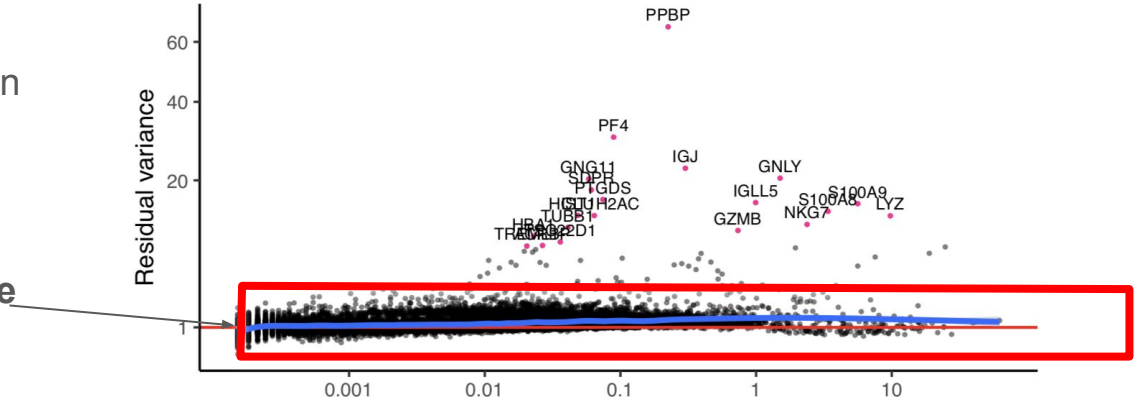
Genes with low variance in mouse across Tabula Muris Senis





SCTransform: Identifying low-variance genes

- Human datasets: Pancreas, brain (sNuc), bone marrow, PBMC, cord blood
- Identify 1000 genes that consistently exhibit **low-variance** in the data




```
 [1] "RPL36AL"  "RPL28"   "RPS9"    "SLC25A6"  "LRRC16B"  "RPL27A"  "UBE2D3"  "EDC4"    "FKBP8"   "EIF3K"   "RPS15"
[12] "ATP5J2"   "POLR2J"  "UBE2I"   "PPDPF"    "DDT"      "CHAF1B"  "AURKAIP1" "SRSF5"   "ARF1"    "RPL6"    "MELK"
[23] "PFDN5"    "RPS28"   "GNB2"    "TRAPPC1"  "KLHL23"   "RPLP2"   "RPL26"   "B4GALT2" "HNRNPK"  "ESPN"    "NACA"
```

MGI Mammalian Phenotype Level 4 2019

Bar G

Click the bars to sort. Now sorted by p-value ranking.

MP:0011100_preweaning_lethality,_complete_penetrance

MP:0013292_embryonic_lethality_prior_to_organogenesis

MP:0011096_embryonic_lethality_between_implantation_and_som

MP:0011094_embryonic_lethality_before_implantation,_complete

GO Cellular Component 2018

Bar Graph

Hover each row to see the overlapping genes.

10 entries per page

Index	Name	P-value
1	cytosolic ribosome (GO:0022626)	5.693e-26
2	cytosolic large ribosomal subunit (GO:0022625)	4.944e-21
3	mitochondrial inner membrane (GO:0005743)	2.044e-34
4	large ribosomal subunit (GO:0015934)	2.157e-20

GO Molecular Function 2018

Bar Graph

Hover each row to see the overlapping genes.

10 entries per page

Index	Name	P-value
1	RNA binding (GO:0003723)	1.495e-53
2	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	6.918e-14
3	NADH dehydrogenase (quinone) activity (GO:0050136)	6.918e-14
4	oxidoreductase activity, acting on diphenols and related substances as donors, cytochrome as acceptor (GO:0016681)	0.00001527
5	ubiquinol-cytochrome-c reductase activity (GO:0008121)	0.00001527

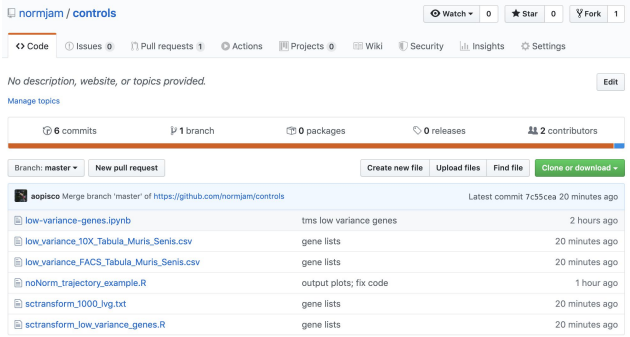
Using Tabula Muris Senis FACS(smartseq2)

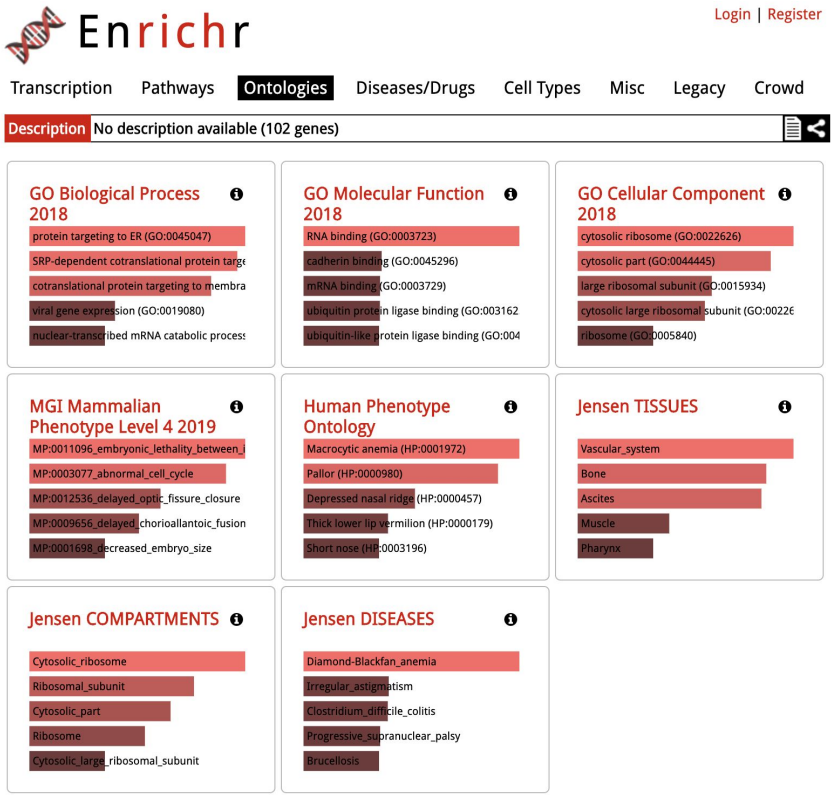
reading data in: adata = Tabula Muris Senis FACS

remove non-annotated cells: adata = adata[adata.obs['cell_ontology_class']!='nan']

n_obs × n_vars = 110824 × 22966

Raw data & Normalization: sc.pp.normalize_per_cell(adata, counts_per_cell_after=1e4)

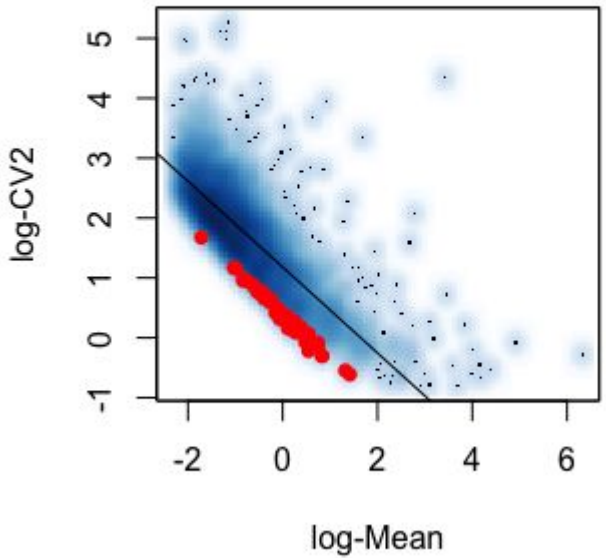




Zeisel et al dataset (astrocytes, oligos, microglia)

Separate analysis per group:

- `scran` normalisation
- Mean vs CV2 trend (log-scale)
- Residual CV2 wrt linear trend
- Genes ranked by residual CV2



Overlap in top 250, 500 and 1000 genes:

- Molecular function: RNA binding
- Mammalian phenotye: embryonic lethality

Tabula Muris / Zeisel overlap / Human datasets

- Gene enrichment are replicated across datasets / methods

- Overlap in genes: Rpls, Atps