Biclustering using sparse factor analysis to understand the human immune system

TRIAD II dataset

Kath Nicholls

Supervised by Chris Wallace and Ken Smith

Overview

- Why biclustering?
- Data processing
- Interpreting results of factor analysis

TRIAD II data

- RNA-Seq data
- 15 immune system cell types
- 6 immune-mediated diseases

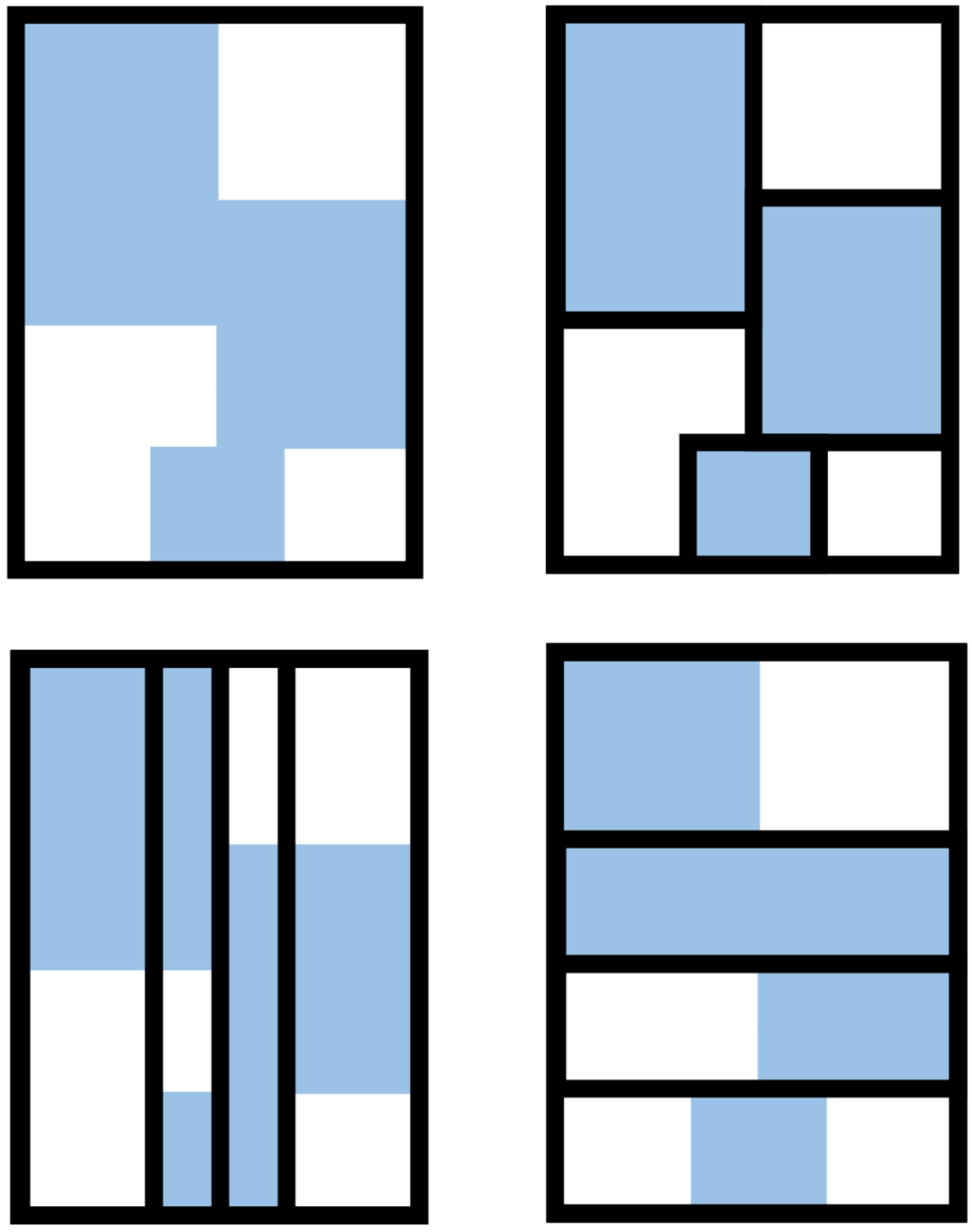
TRIAD II data

	CD16MACS	CD16FACS	CD14	CD19Naive	CD19mem	plasmablast	CD4Naive	CD4mem	CD4Treg	CD8	CD8Naive	CD8Mem	NK	cDC	pDC	
AAV-MPO	2	15	8	0	1	0	8	8	8	0	8	8	8	8	8	90
AAV-PR3	2	15	8	0	0	0	8	8	8	0	7	8	8	8	8	88
IBD-CD	0	0	12	1	0	1	12	12	13	19	13	13	13	12	13	134
IBD-UC	0	0	17	1	0	2	17	17	17	22	16	16	17	16	17	175
IPF	0	14	15	1	0	0	15	15	15	15	15	15	15	14	15	164
SLE	0	12	8	0	0	1	8	8	8	0	8	8	8	8	8	85
Unknown	0	0	1	1	3	0	1	1	0	0	0	0	0	1	0	8
Healthy	8	24	25	0	1	0	25	25	25	25	26	26	25	25	25	285
	12	80	94	4	5	4	94	94	94	81	93	94	94	92	94	1029

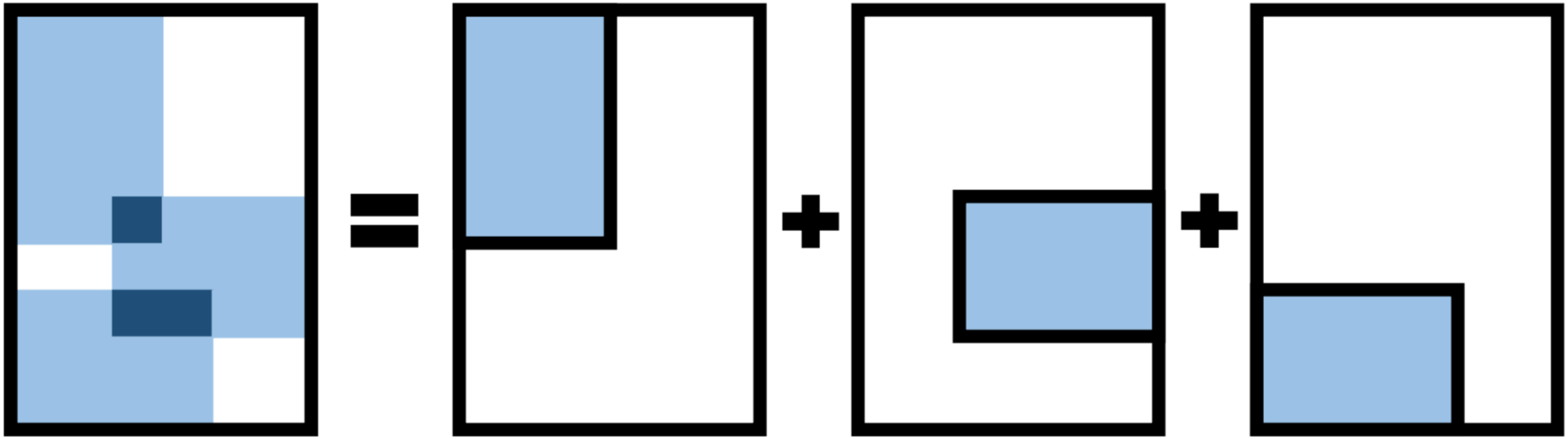
High-dimensional data

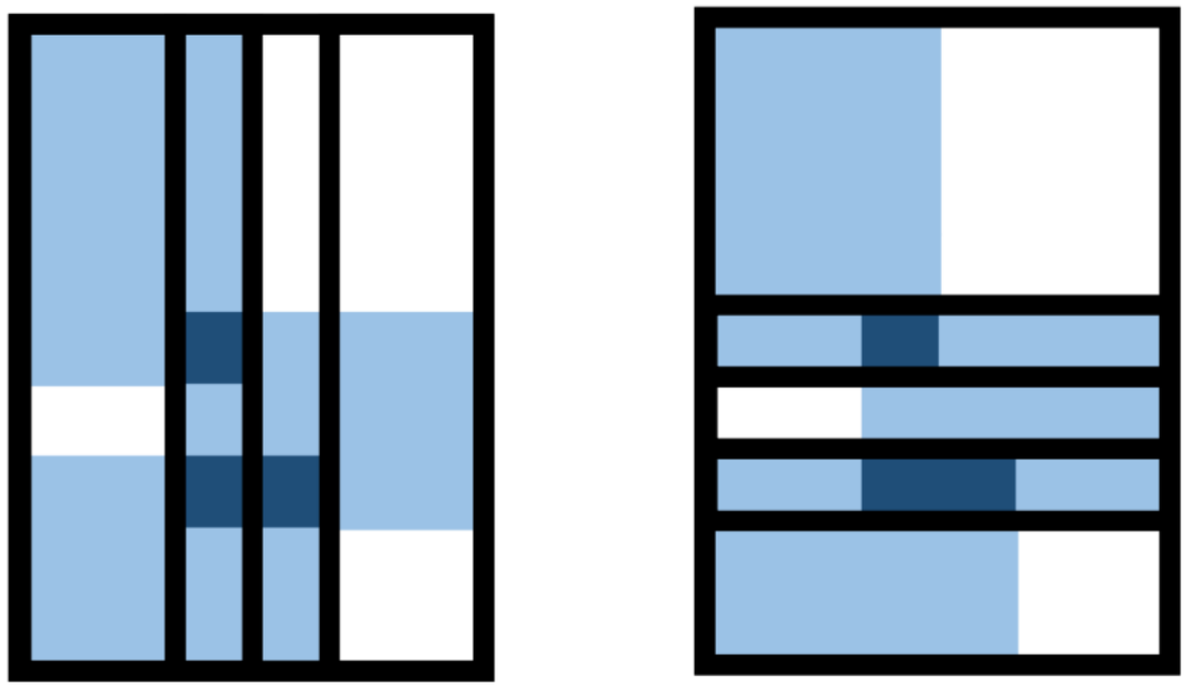
- 1029 samples
- ~ 40,000 genes and pseudogenes

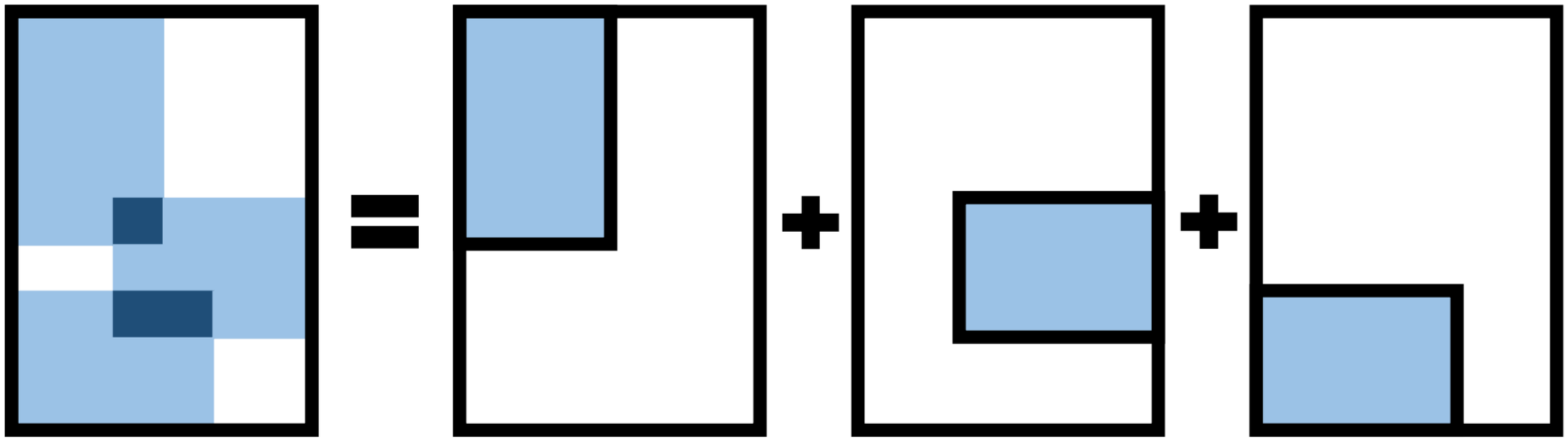
Biclustering



Factor analysis







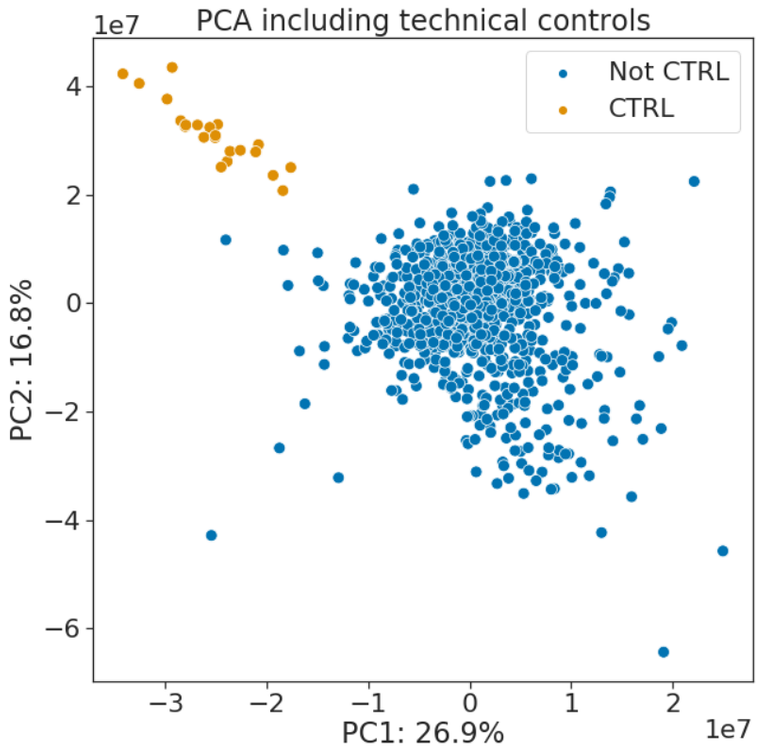
$$\underset{(p\times n)}{\mathrm{M}} = \underset{(p\times K)}{\begin{bmatrix} | & | & & | \\ l_1 & l_2 & \dots & l_k \\ | & | & & | \end{bmatrix}} \underset{(K\times n)}{\begin{bmatrix} — & f_1 & — \\ — & f_2 & — \\ & \vdots & \\ — & f_k & — \end{bmatrix}} + \underset{(p\times n)}{\varepsilon}$$

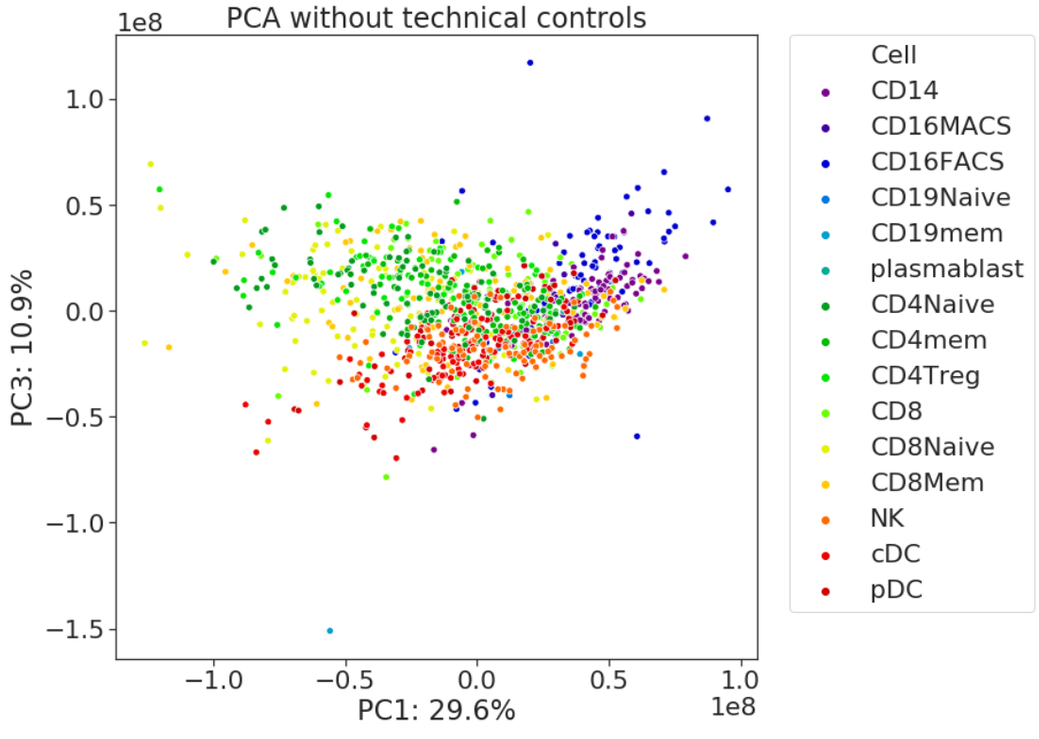
Overview

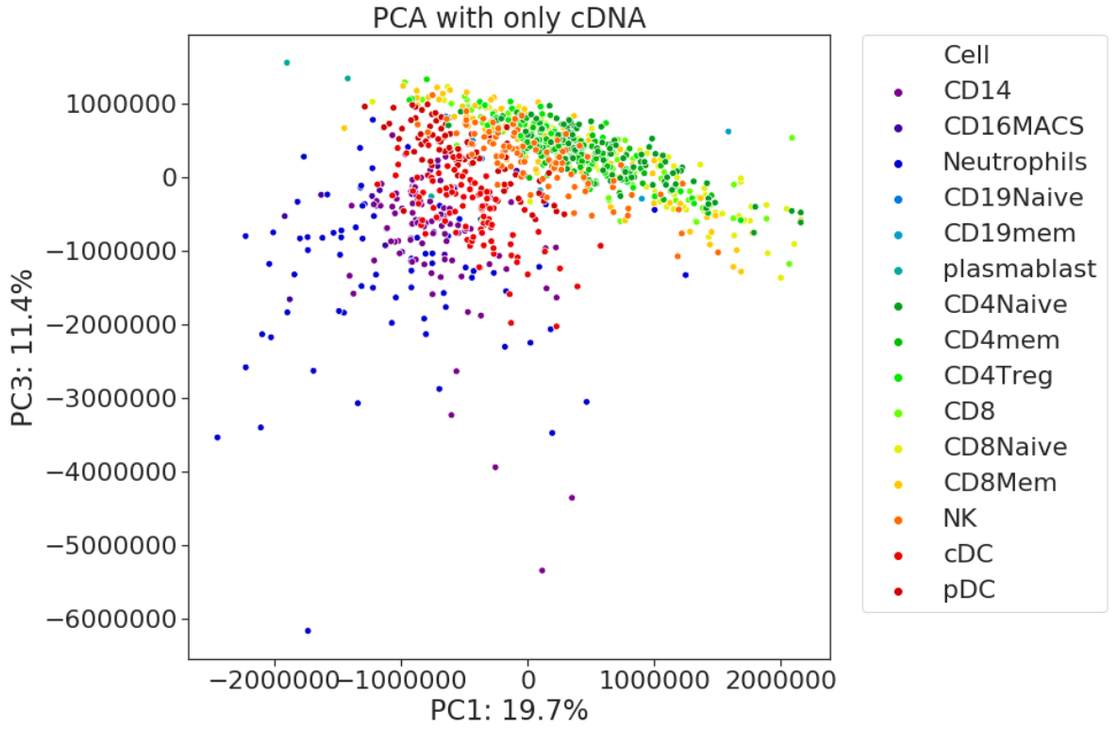
- Why biclustering?
- Data processing
- Interpreting results of factor analysis

Data processing

- Trimmed reads with prinseq
- Quantified using kallisto (tpm)
- Discarded technical controls
- Discarded ncRNA transcripts

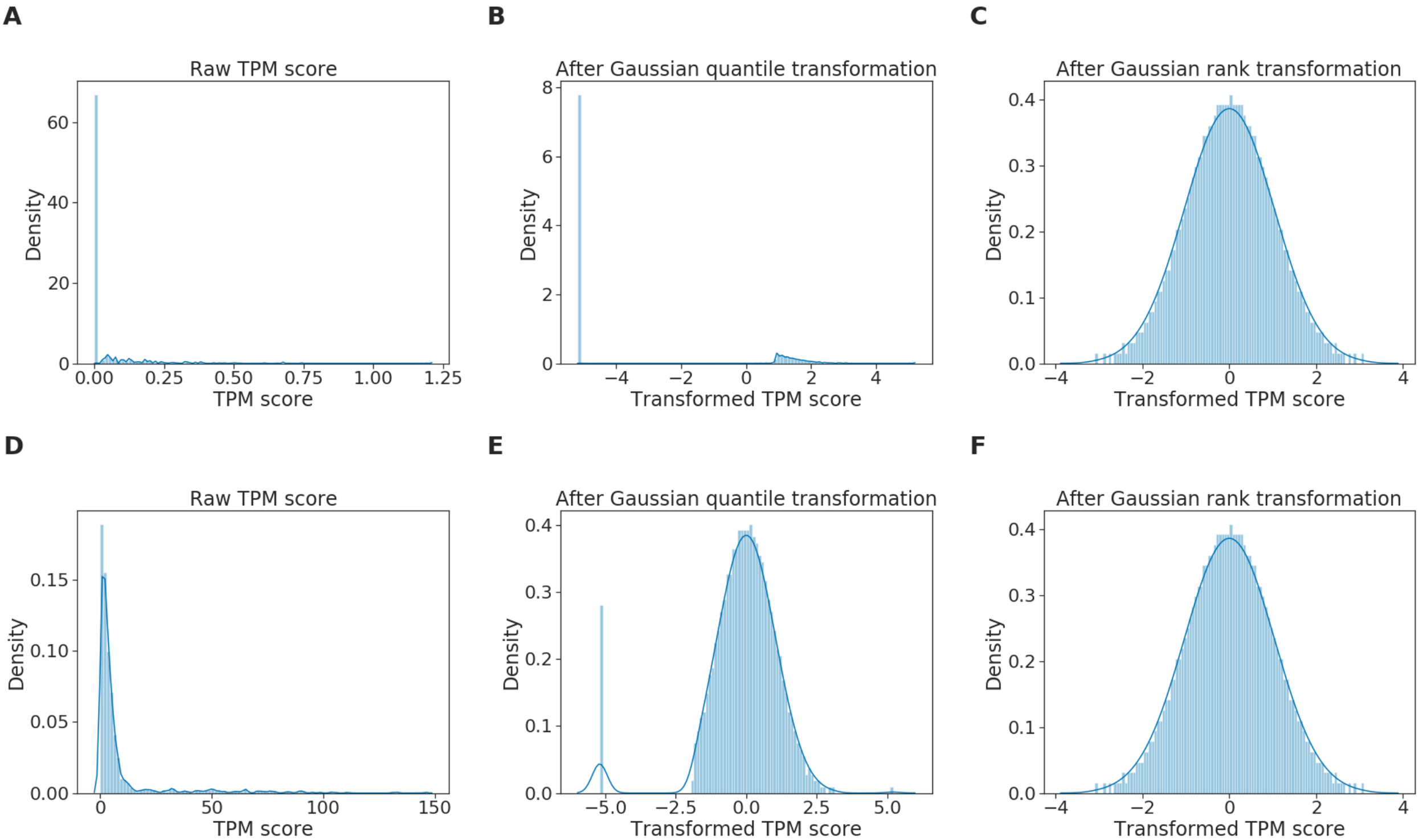






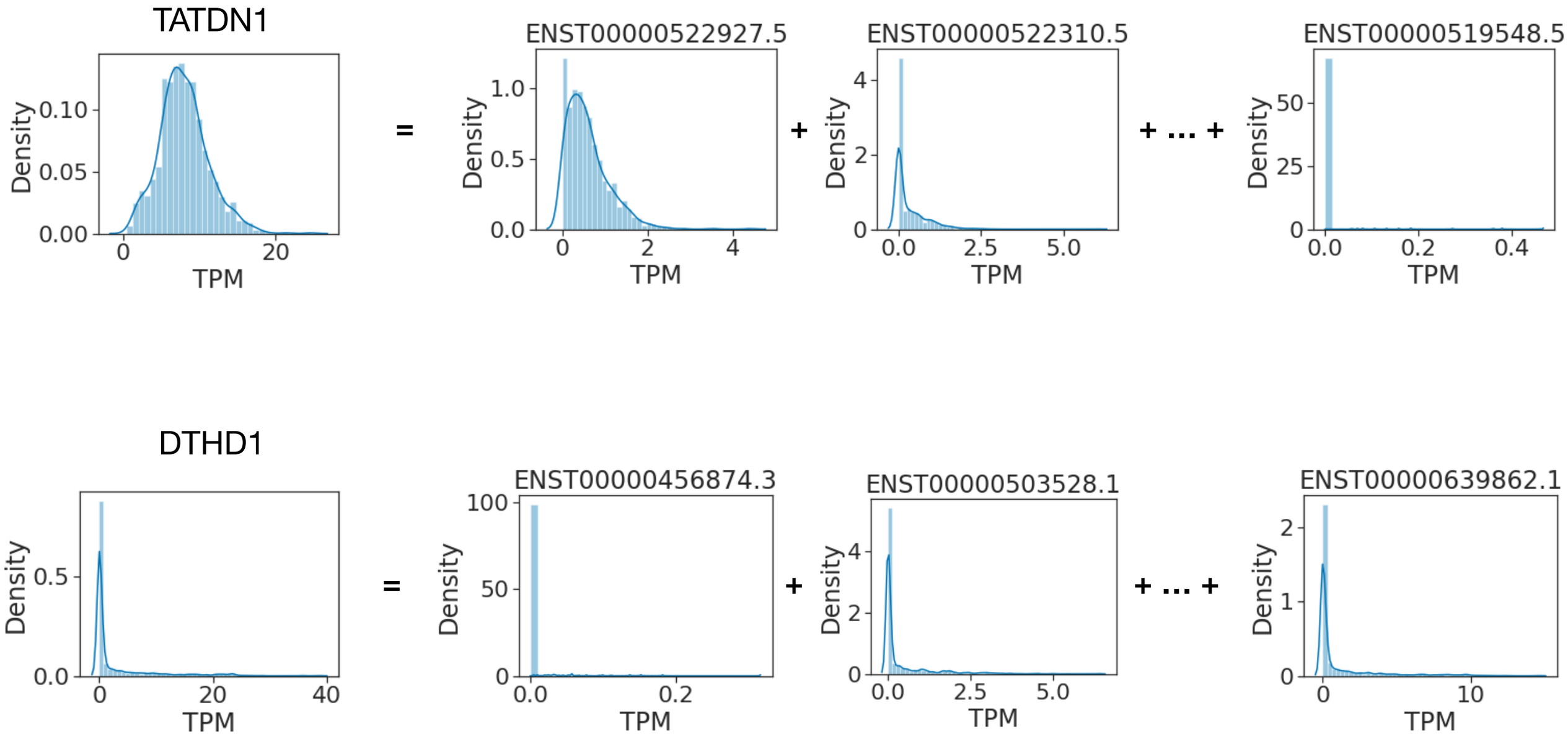
kallisto: Bray et al. 2016, prinseq: Schmeider & Edwards 2011

Gaussian transformations

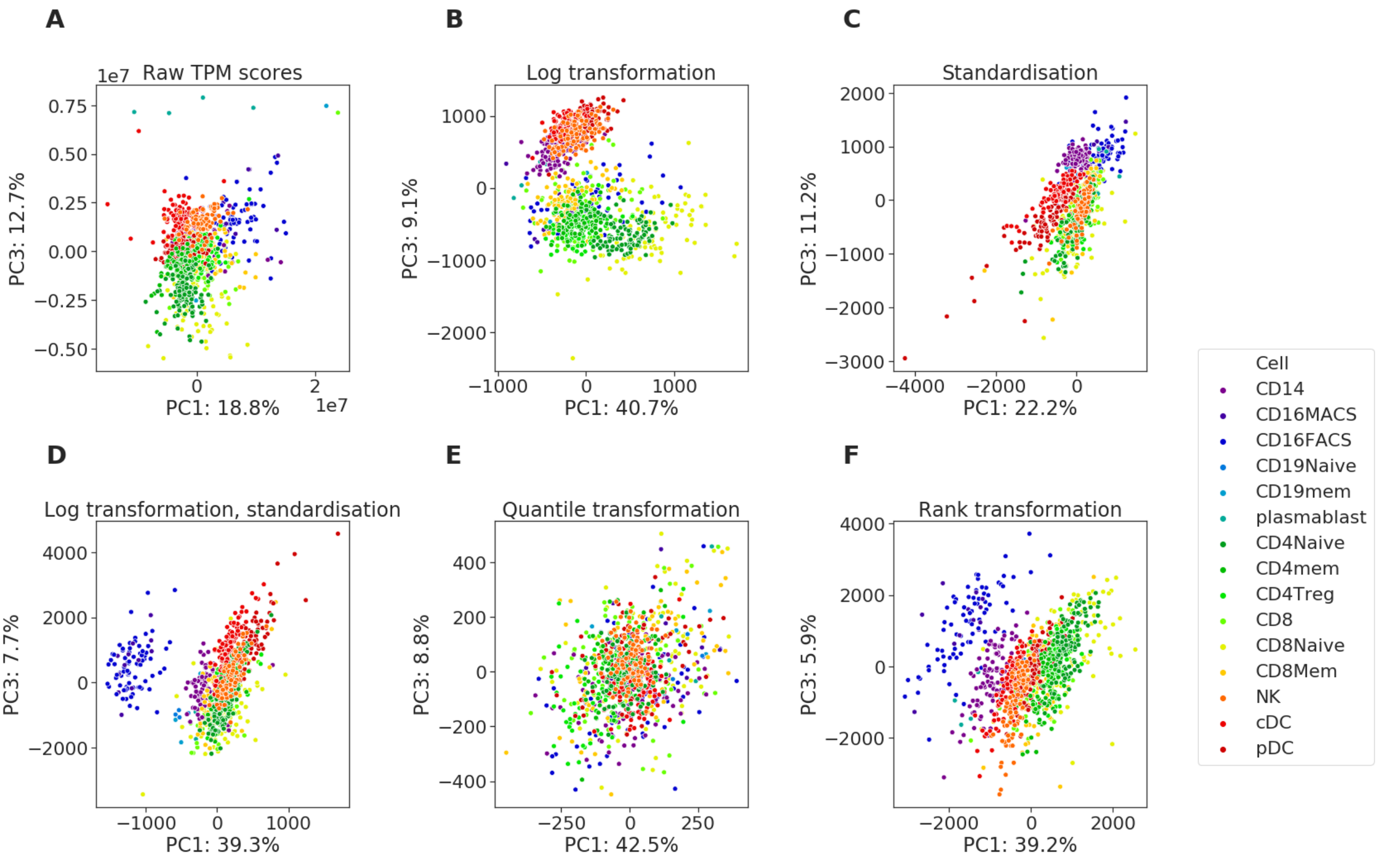


Merging transcripts to genes

- BicMix advises Gaussian rank normalisation



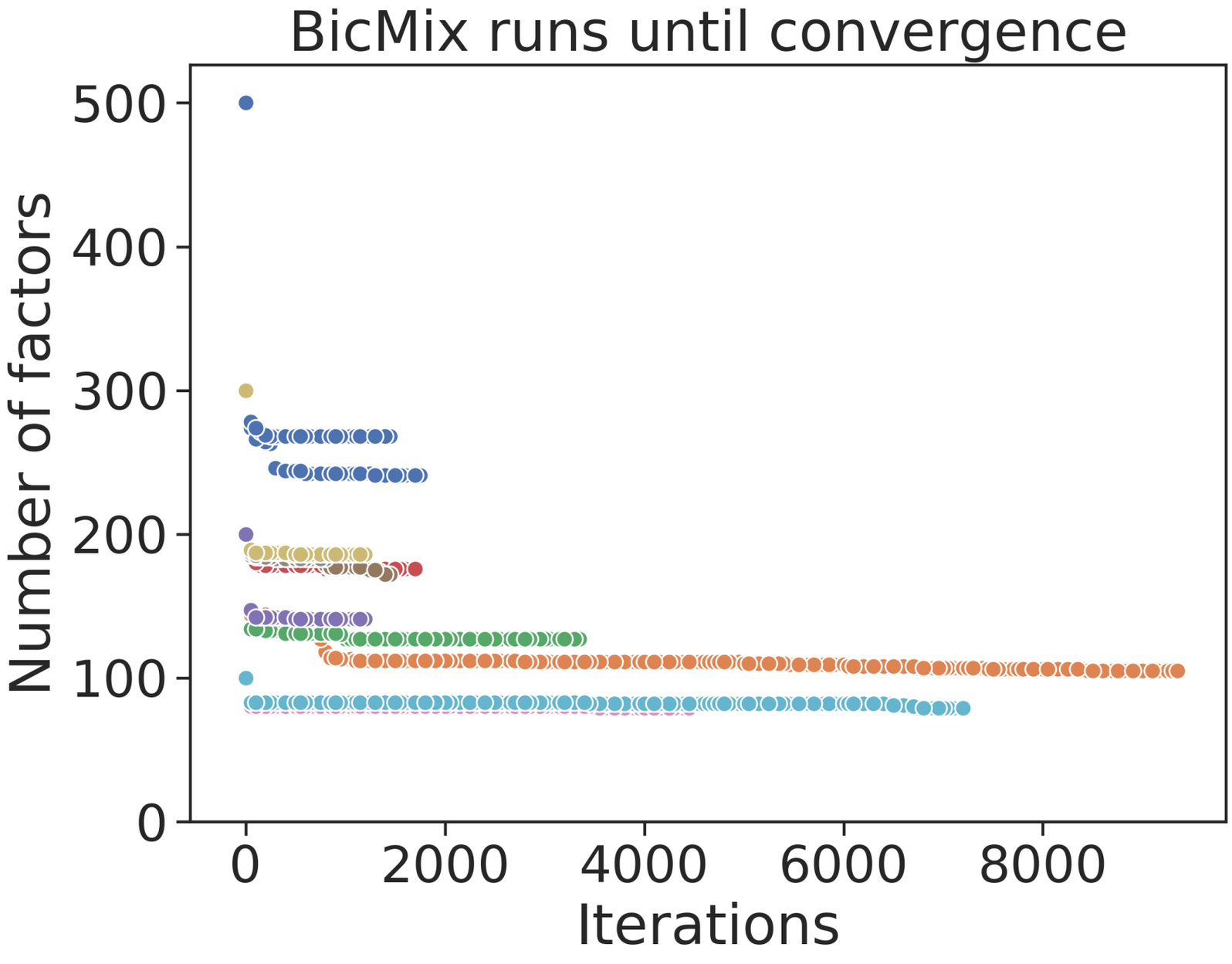
Transformations (without sparse genes)



Overview

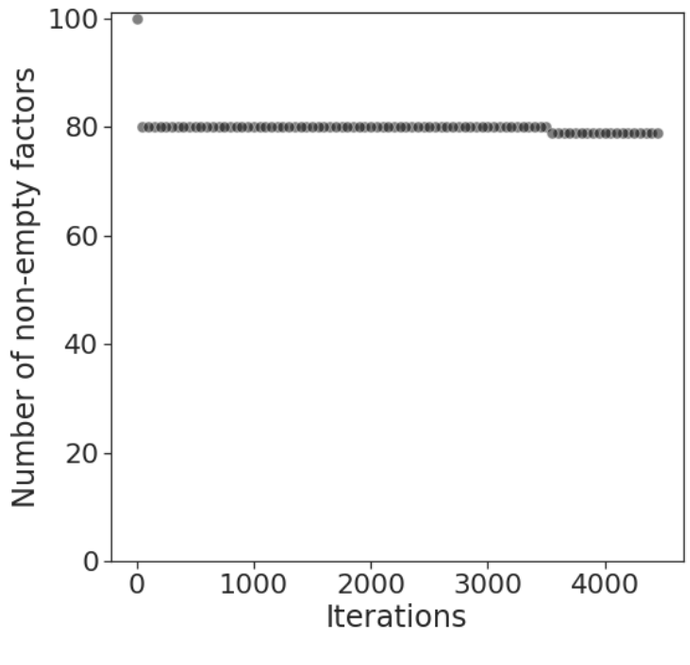
- Why biclustering?
- Data processing
- Interpreting results of factor analysis

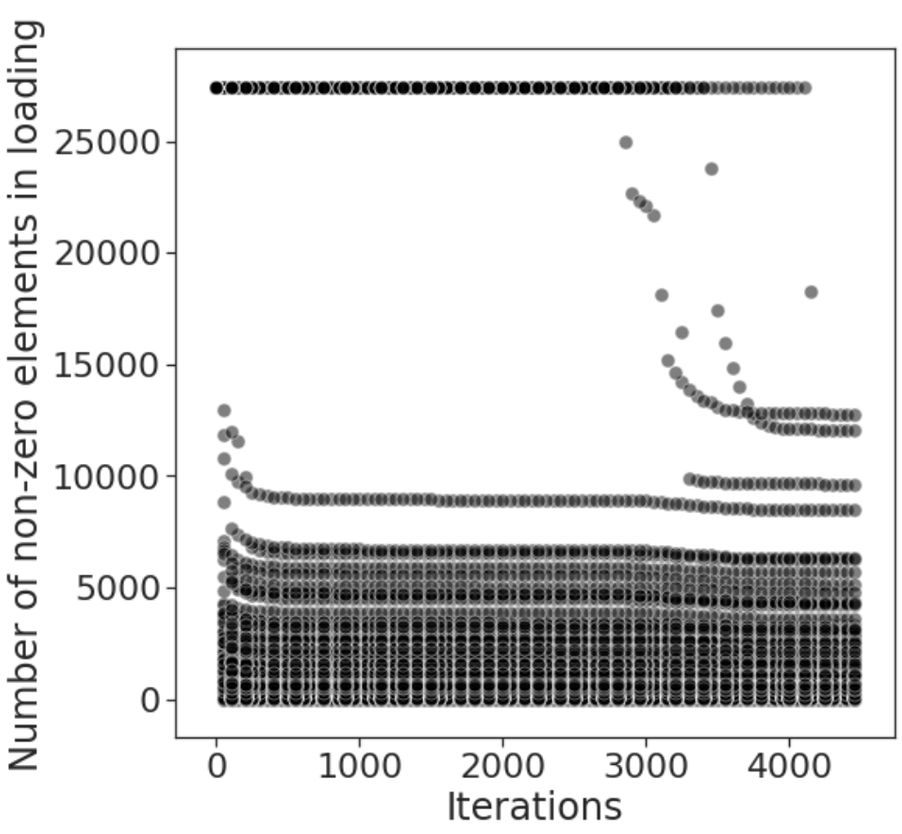
Running BicMix

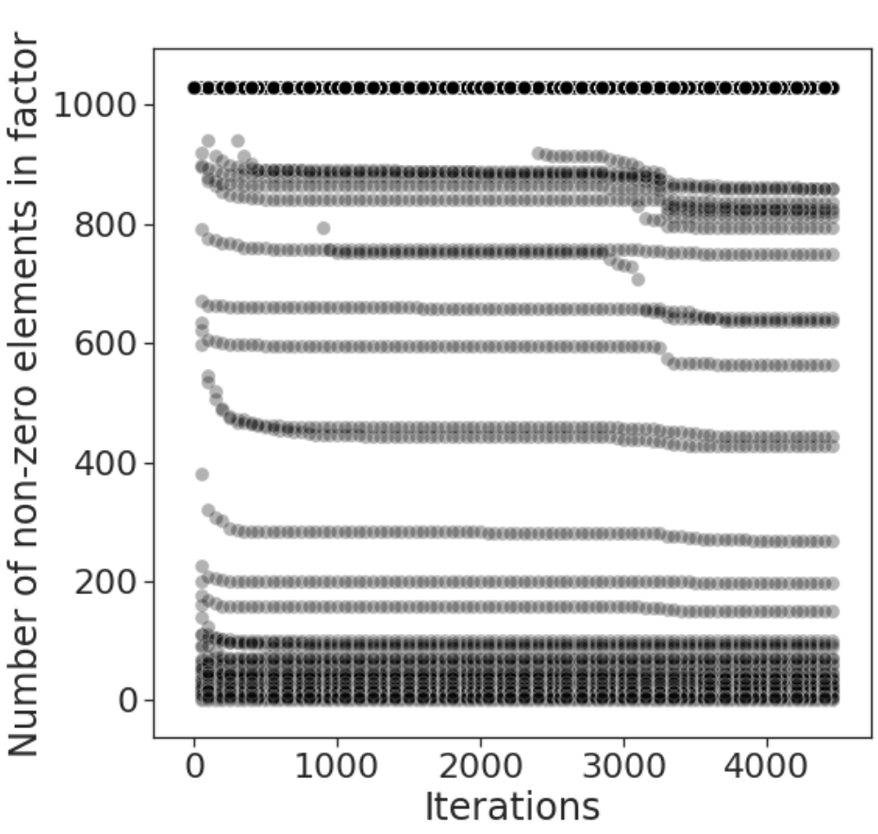


BicMix: Gao et al. 2014

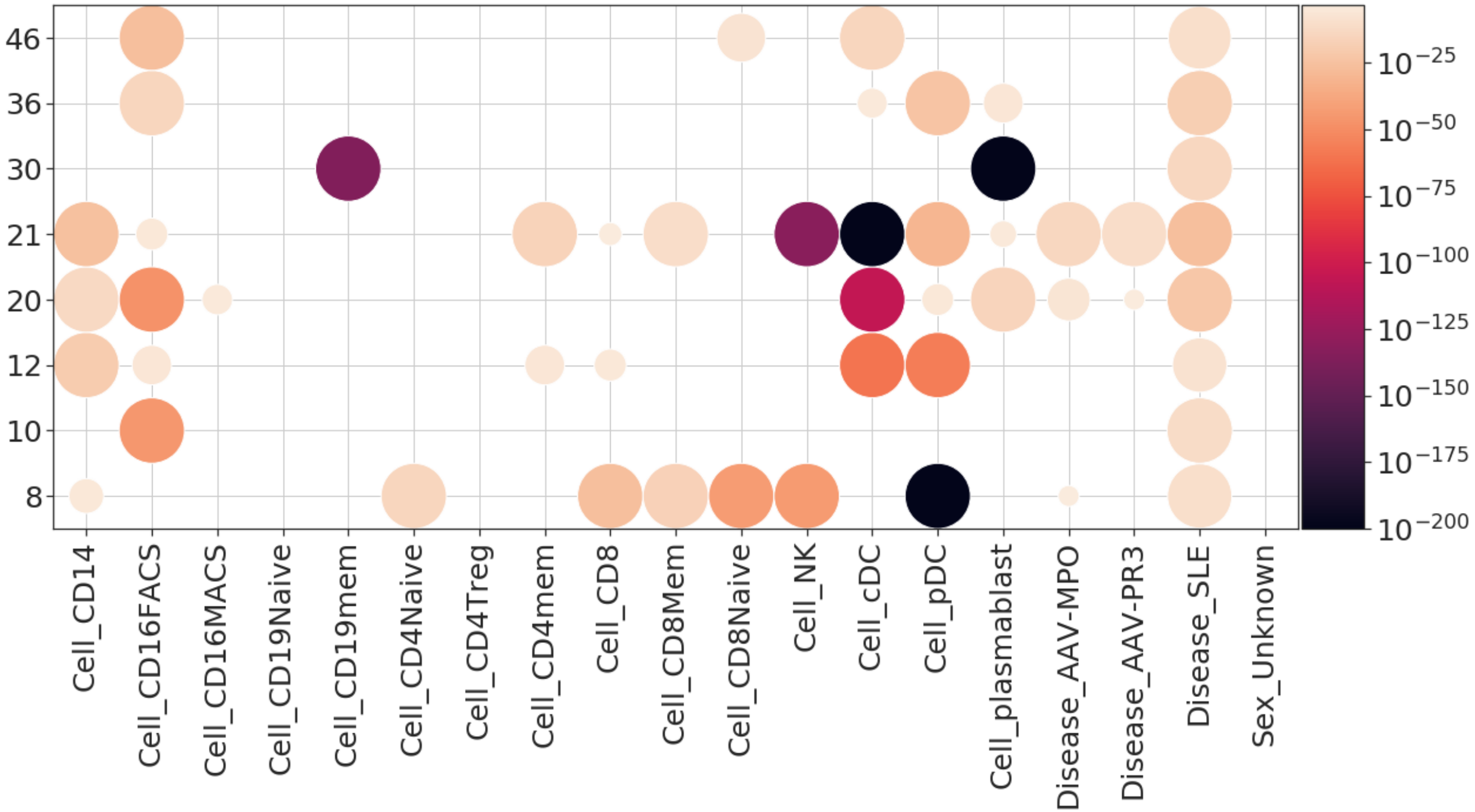
Focused on one run

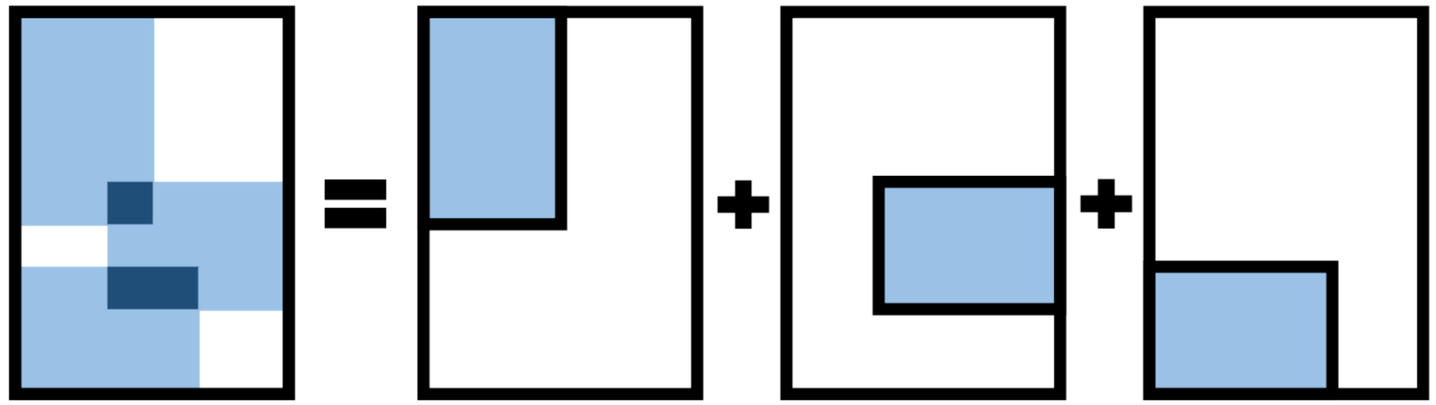




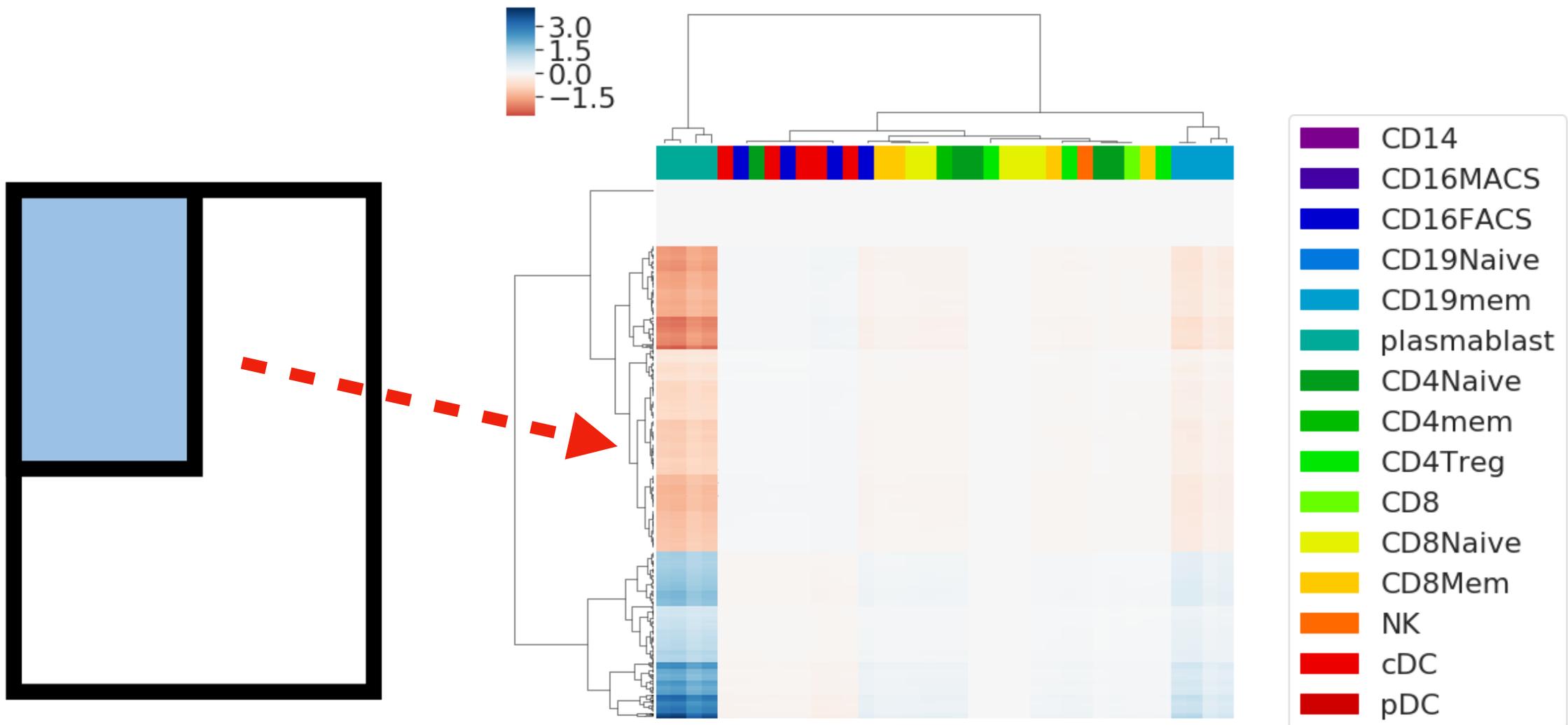


Factors associated with systemic lupus erythematosus

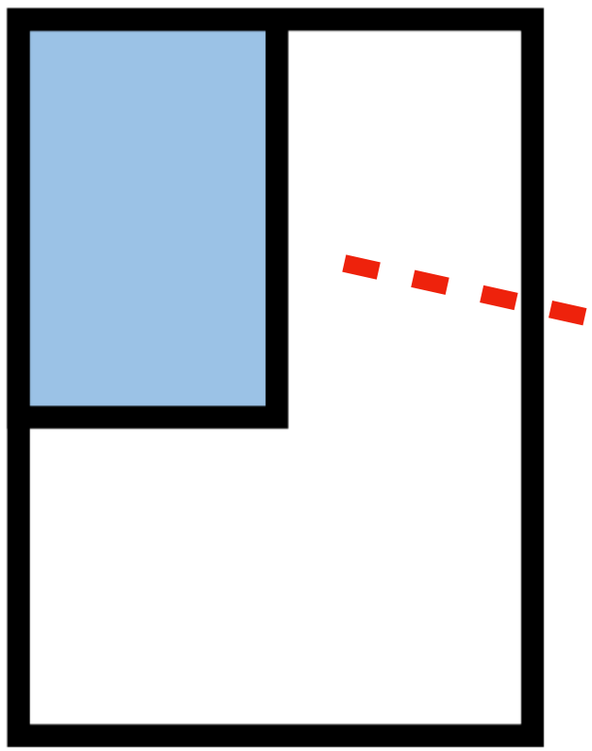




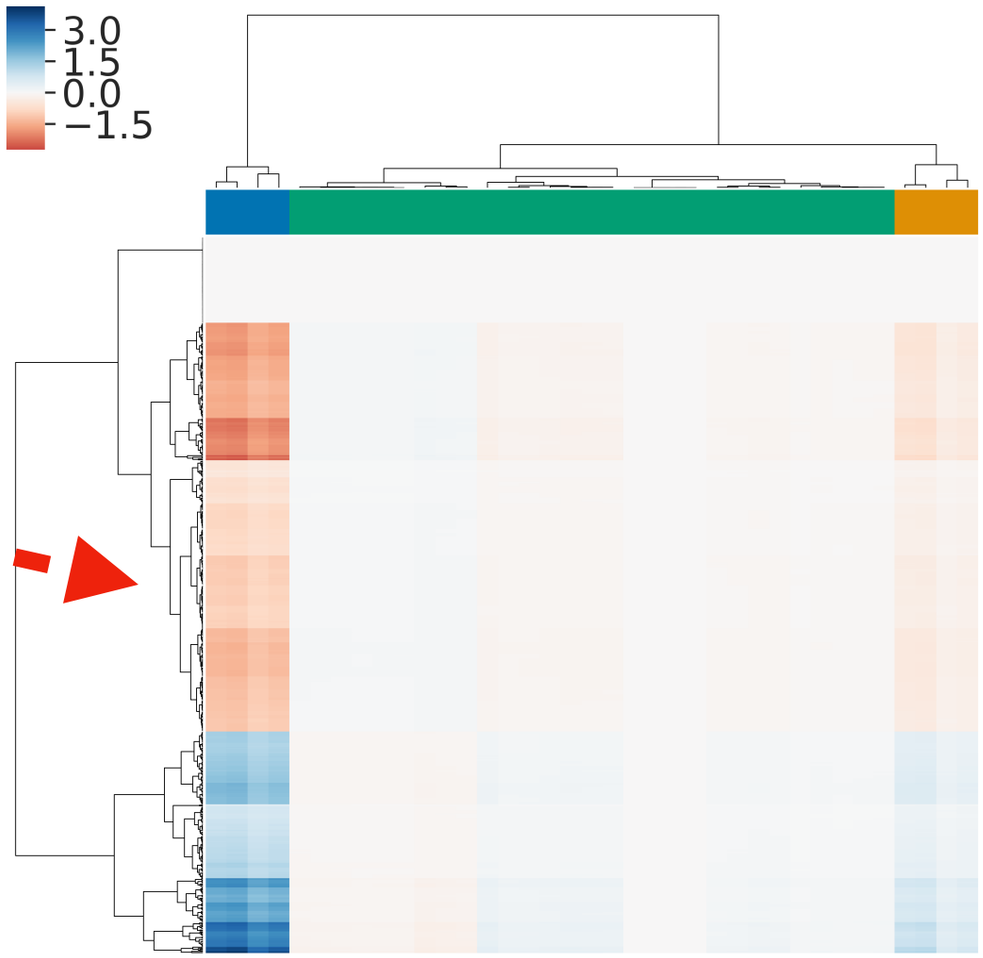
Expectation **Reality**

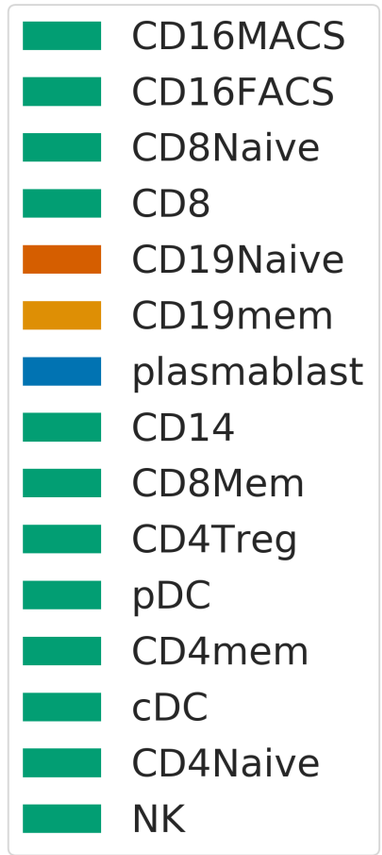


Expectation

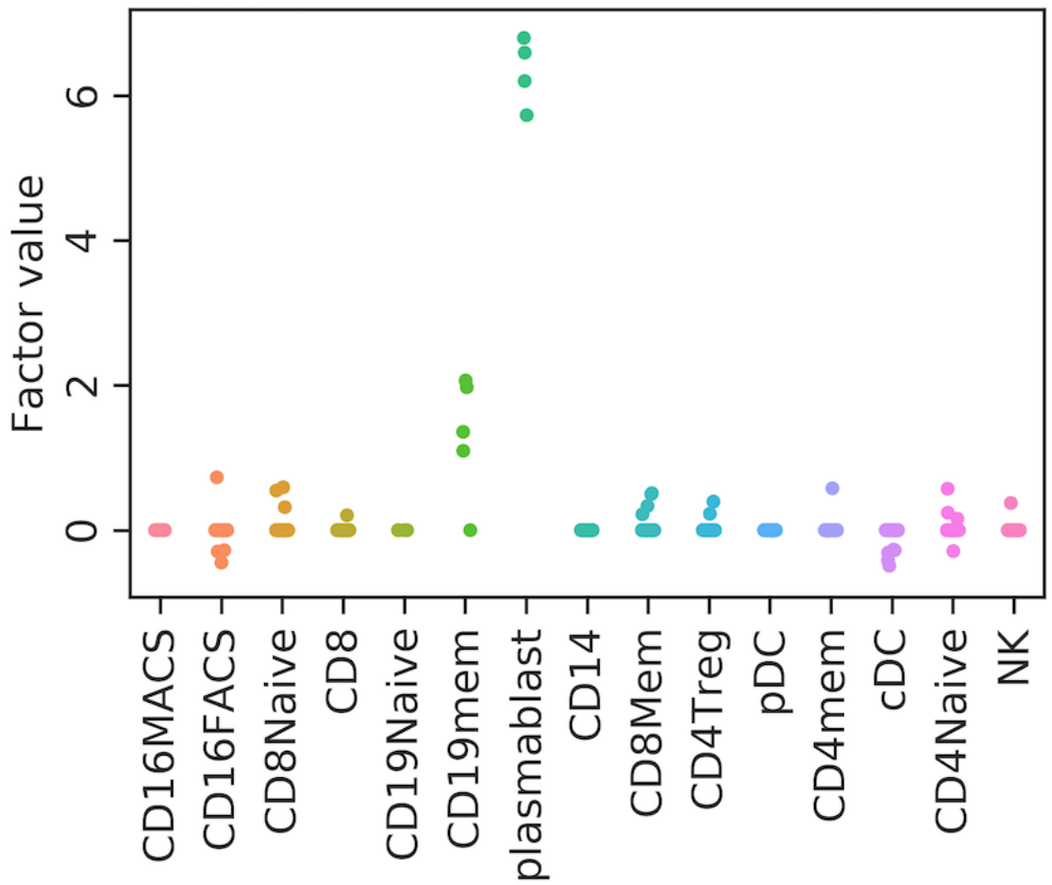


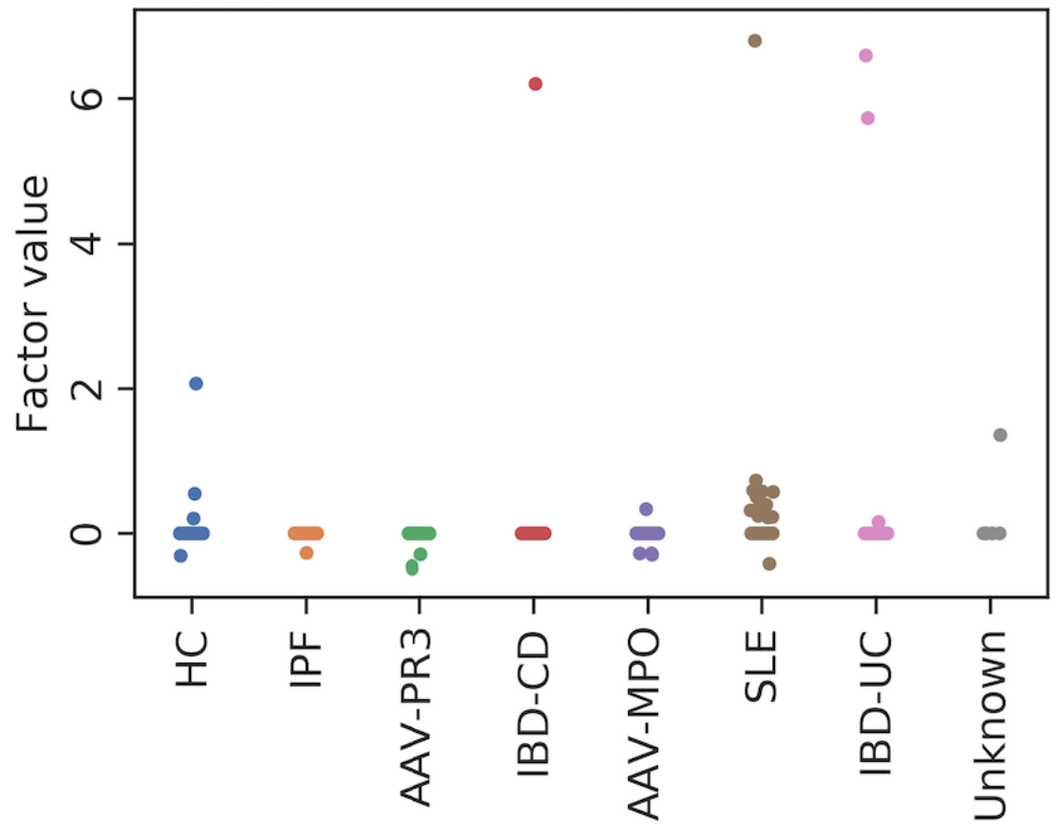
Reality





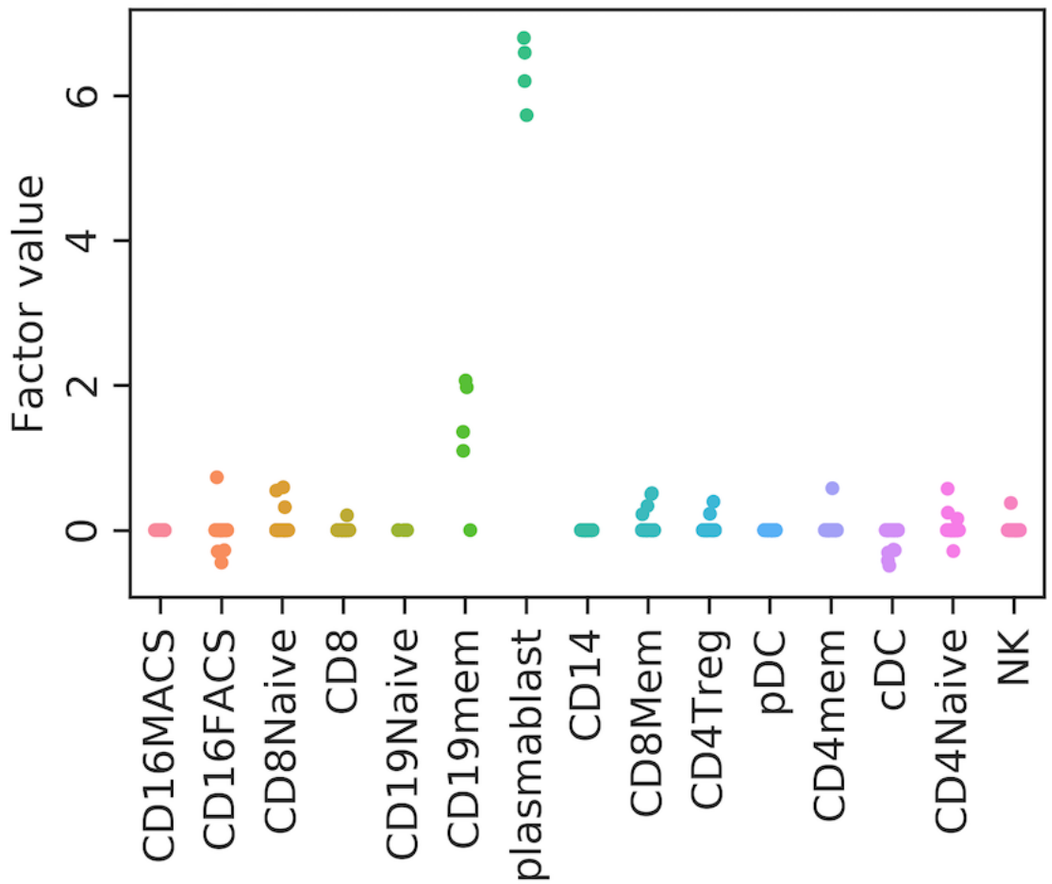
Factor 30

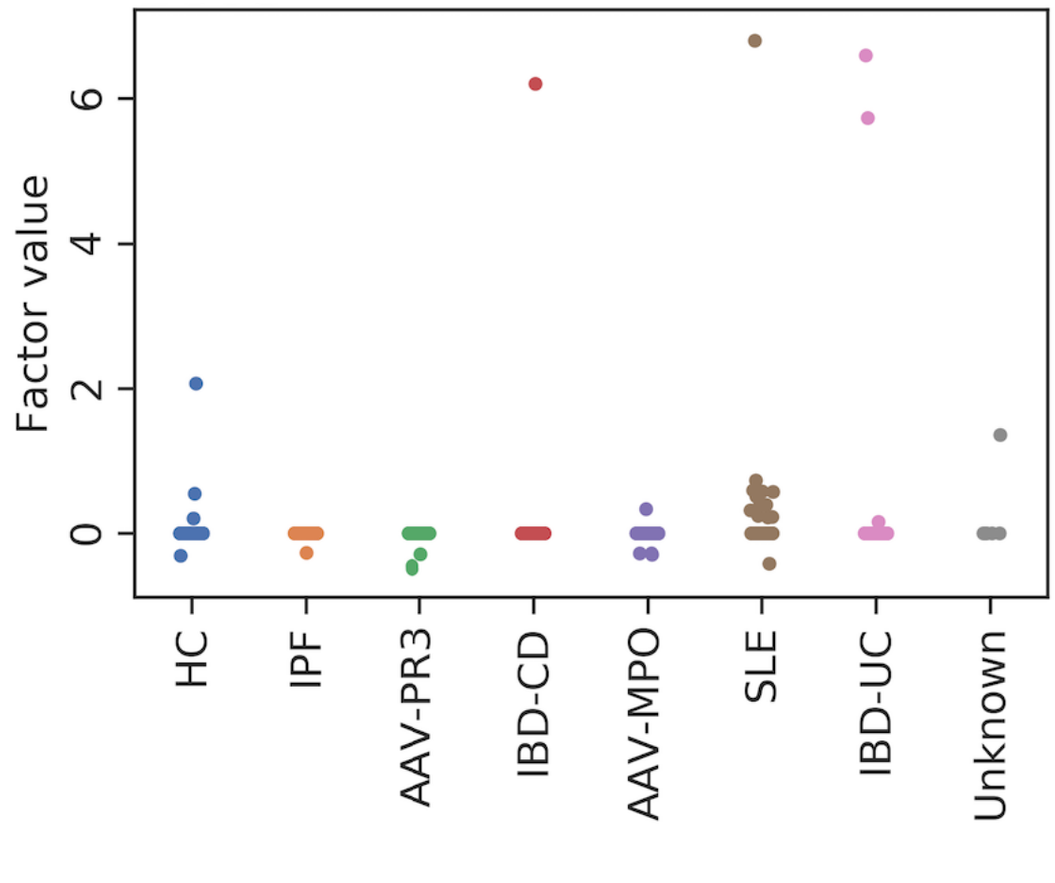




KEGG pathway linked to factor 30	Adjusted p-value
Fc gamma R-mediated phagocytosis	3.53×10^{-8}
B cell receptor signaling pathway	4.15×10^{-6}
Regulation of actin cytoskeleton	1.36×10^{-5}
Chemokine signaling pathway	2.49×10^{-5}
Pathways in cancer	6.08×10^{-5}

Factor 30





Fc gamma R-mediated phagocytosis

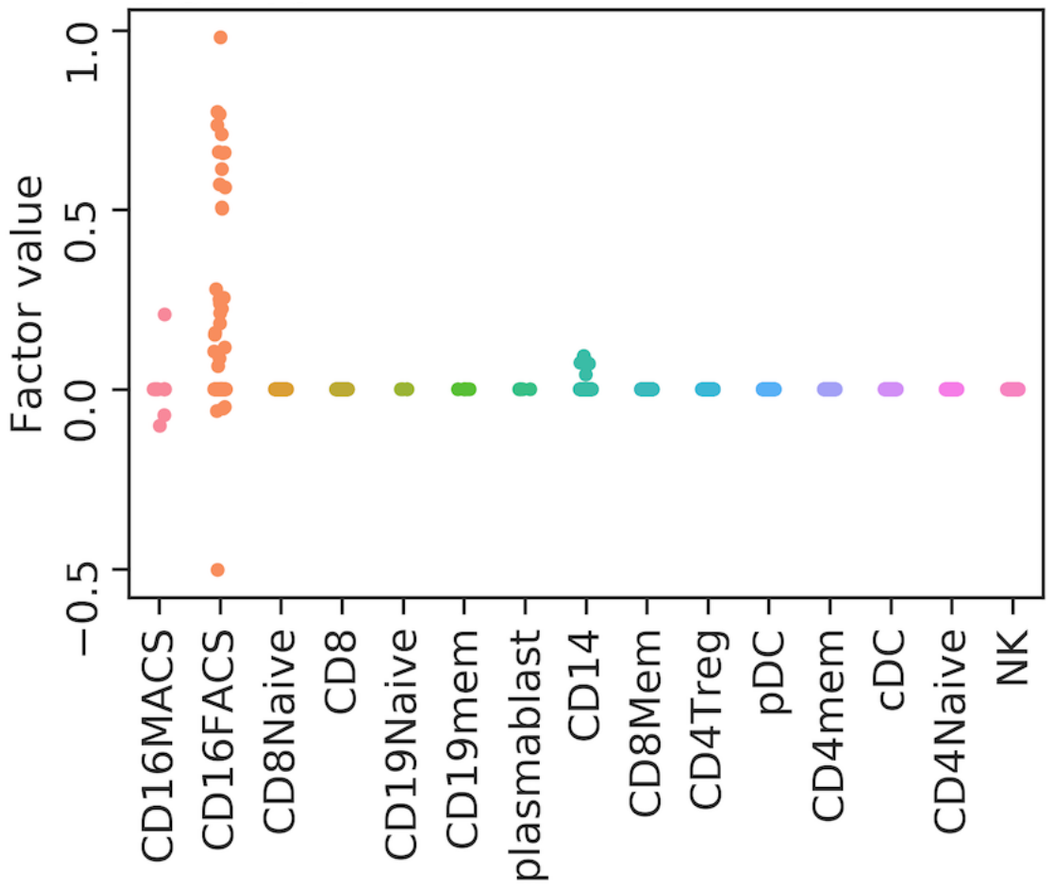
AKT1, AKT3, ARPC1B, ARPC3, ARPC4, ARPC5, ASAP1, CFL1, FCGR2A, FCGR2B, INPP5D, LIMK2, MAPK1, PTPRC, RAC2, VASP, VAV3, WAS

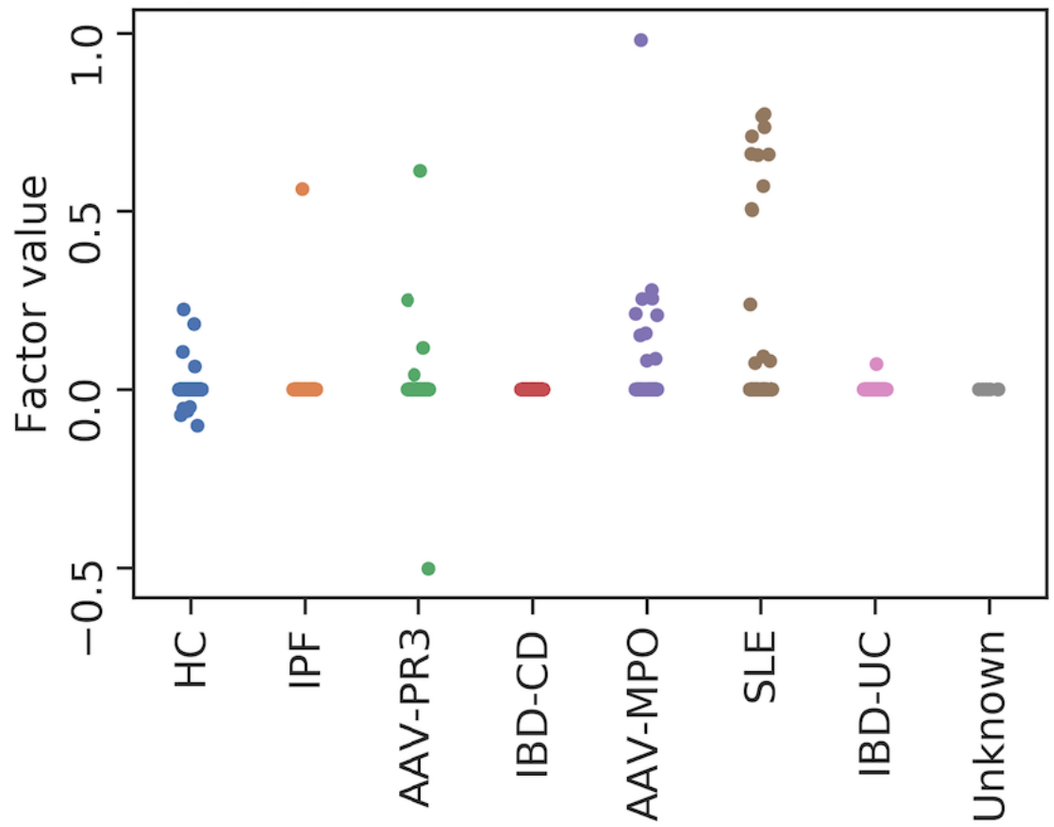
B-cell receptor signalling pathway

AKT1, AKT3, BLNK, CD19, DAPP1, FCGR2B, INPP5D, KRAS, MAPK1, NFATC3, PTPN6, RAC2, RASGRP3, VAV3

Genes in factor *and* pathway. Red indicates shared by both pathways

Factor 10



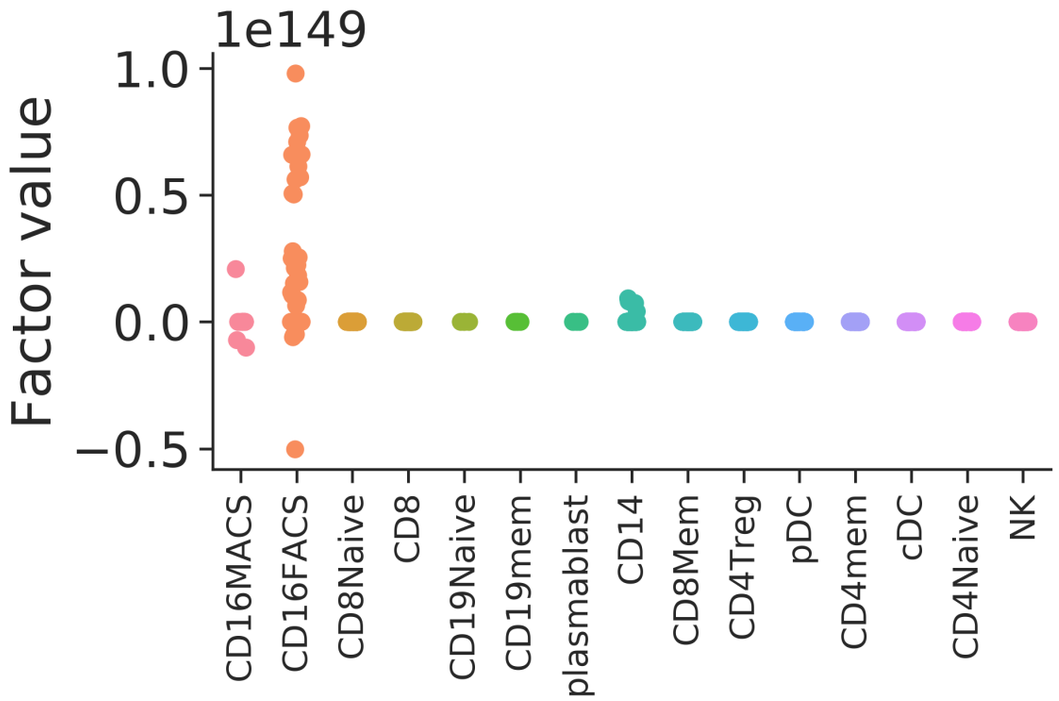


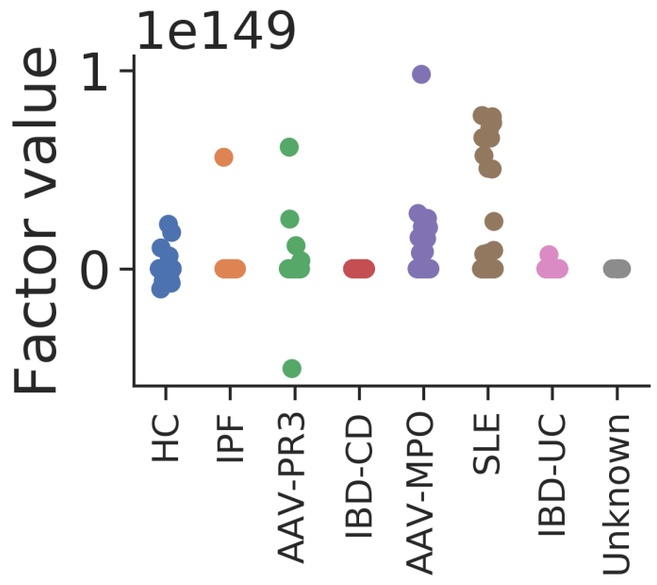
KEGG pathway linked to factor 10	Adjusted p-value
Viral carcinogenesis	4.28×10^{-8}
Herpes simplex infection	1.40×10^{-7}
Epstein-Barr virus infection	2.04×10^{-7}
Protein processing in endoplasmic reticulum	2.23×10^{-7}
NOD-like receptor signaling pathway	1.29×10^{-6}

KEGG pathway linked to factor 10	Adjusted p-value
Viral carcinogenesis	4.28×10^{-8}
Herpes simplex infection	1.40×10^{-7}
Epstein-Barr virus infection	2.04×10^{-7}
Protein processing in endoplasmic reticulum	2.23×10^{-7}
NOD-like receptor signaling pathway	1.29×10^{-6}
Measles	2.14×10^{-6}
Hepatitis B	7.74×10^{-6}
TNF signaling pathway	7.90×10^{-5}
Influenza A	1.47×10^{-4}
Alcoholism	2.58×10^{-4}

MolSigDB pathway	Total	In factor	Unadjusted p-value
Interferon gamma response	200	47	3.79×10^{-24}
Interferon alpha response	97	30	1.43×10^{-19}

Factor 10





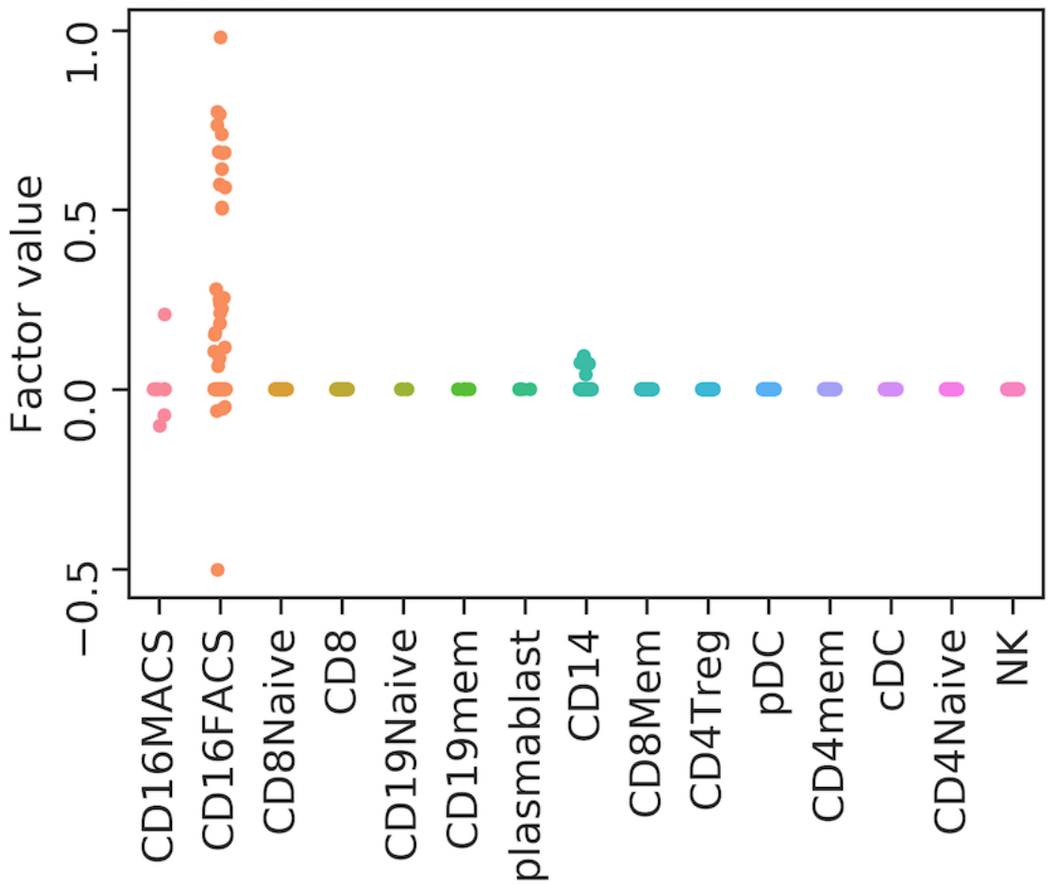
Viral carcinogenesis

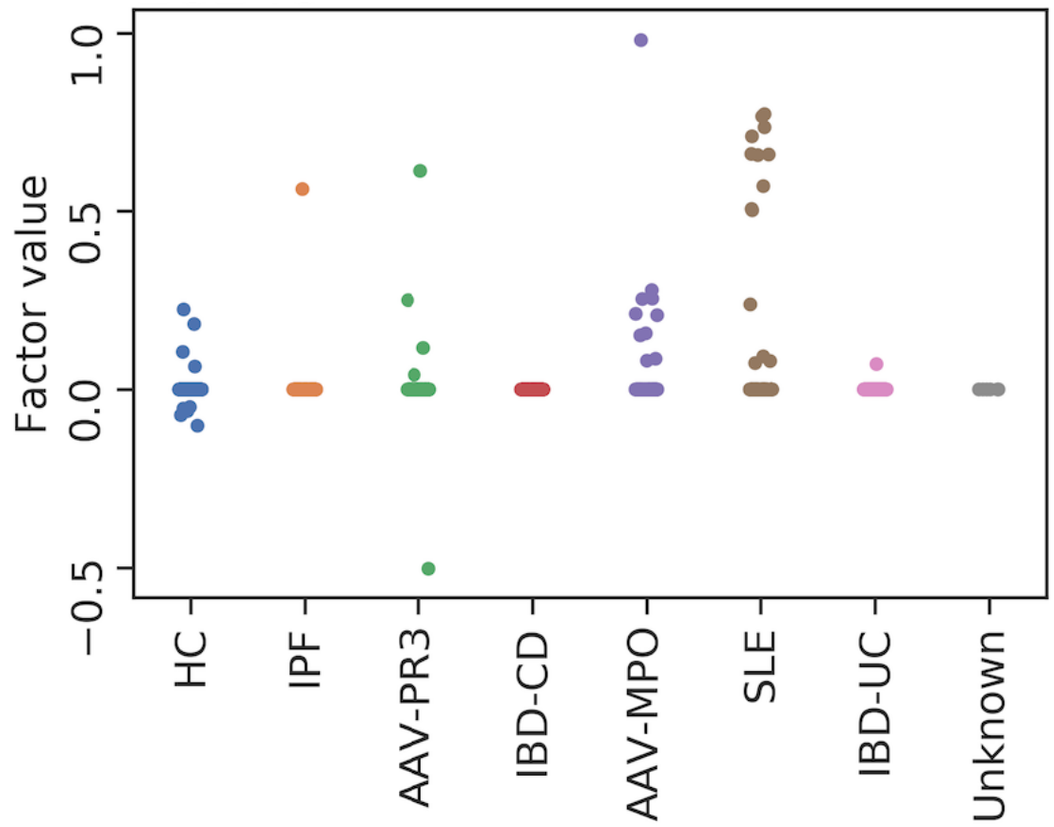
ACTN4, ATF4, CASP8, CCNA1, CREB3L2, EIF2AK2, HIST1H2BB, HIST1H2BJ, HIST1H4A, HIST1H4B, HIST1H4J, HIST2H4A, HLA-B, IRF7, JAK1, KAT2A, KRAS, MAPK1, NFKB1, NFKBIA, PRKACB, REL, RHOA, SP100, SRF, TRADD, UBR4, USP7, YWHAB, YWHAE, YWHAG

Herpes simplex virus infection

CASP8, CSNK2A1, CSNK2B, CUL1, DDX58, EIF2AK2, HLA-B, IFIH1, IFIT1, IRF7, JAK1, MAP3K7, NFKB1, NFKBIA, OAS1, OAS2, PML, PPP1CA, PPP1CB, PPP1CC, SP100, SRSF2, SRSF3, SRSF6, STAT1, TAP1, TLR2, USP7

Factor 10





Epstein-Barr virus infection

AKT2, ***BCL2***, *CCNA1*, *CSNK2A1*, *CSNK2B*, *DDX58*, *EIF2AK2*, *GSK3B*, ***HLA-B***, ***IRAK1***, ***JAK1***, *MAP2K7*, *MAP3K14*, *MAP3K7*, ***NFKB1***, *NFKBIA*, *POLR2B*, *POLR2C*, *POLR3GL*, *PRKACB*, *PSMD13*, *PSMD8*, ***RIPK1***, ***TNFAIP3***, ***TRADD***, *USP7*, *VIM*, ***YWHAB***, *YWHAE*, *YWHAG*

Bold means association with SLE found in MyGene.Info literature search

Conclusion

- Good potential for sparse factor analysis
- Possible improvements:
  - Full dataset
  - Convergence of BicMix
  - More starting factors