Patterns, Volume *2*

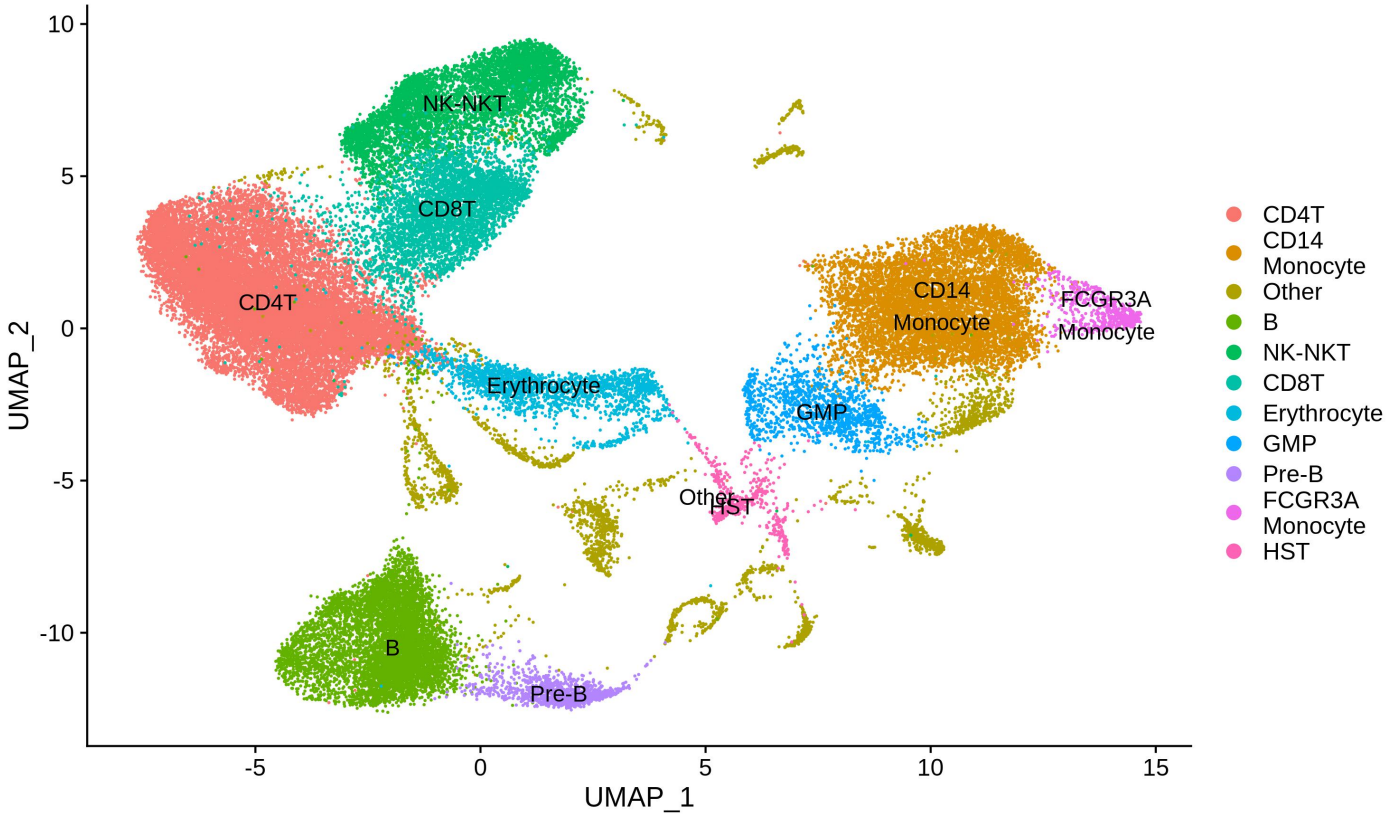
Supplemental information

Noise regularization removes correlation artifacts in single-cell RNA-seq data preprocessing

Ruoyu Zhang, Gurinder S. Atwal, and Wei Keat Lim

Figure S1. HCA single cell data used for this study

A



B

Cluster	0	1	2	3	4	5	6	7	8	9
Cell type	CD4T	CD14 Monocyte	B	NK-NKT	CD8T	Erythrocyte	GMP	Pre-B	FCGR3A Monocyte	HST
Cell number	16936	7413	6534	5847	4467	1974	1347	1052	583	598
Top 10 markers	IL7R	S100A9	CD79A	GNLY	GZMK	HBB	MPO	CD79B	LST1	SPINK2
	LTB	S100A8	CD74	NKG7	RGS1	AHSP	ELANE	HIST1H1C	IFITM3	AVP
	TRAC	S100A12	IGHD	GZMB	CCL4	CA1	PRTN3	TCL1A	AIF1	SOX4
	NOSIP	LYZ	MS4A1	FGFBP2	DUSP2	HBD	AZU1	SOX4	FCGR3A	KIAA0125
	LEPROTL1	FCN1	IGHM	GZMH	CMC1	PRDX2	LYZ	VPREB3	COTL1	ANKRD28
	PIK3IP1	CXCL8	HLA-DQB1	PRF1	CCL5	HBA1	CTSG	CD24	FCER1G	IGLL1
	CD3D	TYROBP	HLA-DRA	CST7	GZMA	BLVRB	RETN	NEIL1	SERPINA1	PRSS57
	LDHB	VCAN	HLA-DRB1	KLRD1	CST7	HBA2	RNASE2	IGHM	S100A11	PRDX1
	MAL	CSTA	HLA-DPA1	CCL5	IL32	TUBA1B	LGALS1	PCDH9	SAT1	H2AFY
	CD3E	NAMPT	HLA-DQA1	KLRF1	KLRB1	TUBB	H2AFZ	VPREB1	PSAP	SERPINB1

Figure S2. PPI enrichment of randomly sampled gene pairs.

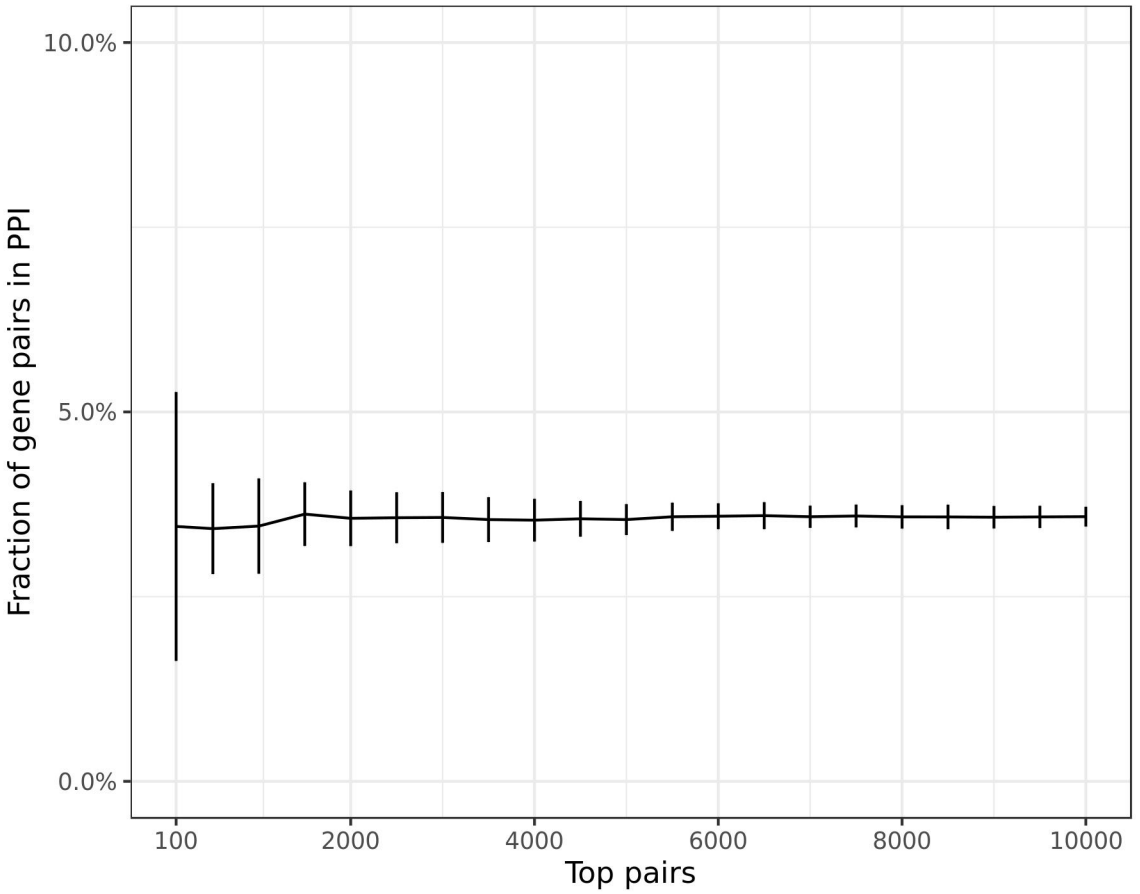


Figure S3. Gene-gene correlation coefficients before and after noise regularization.

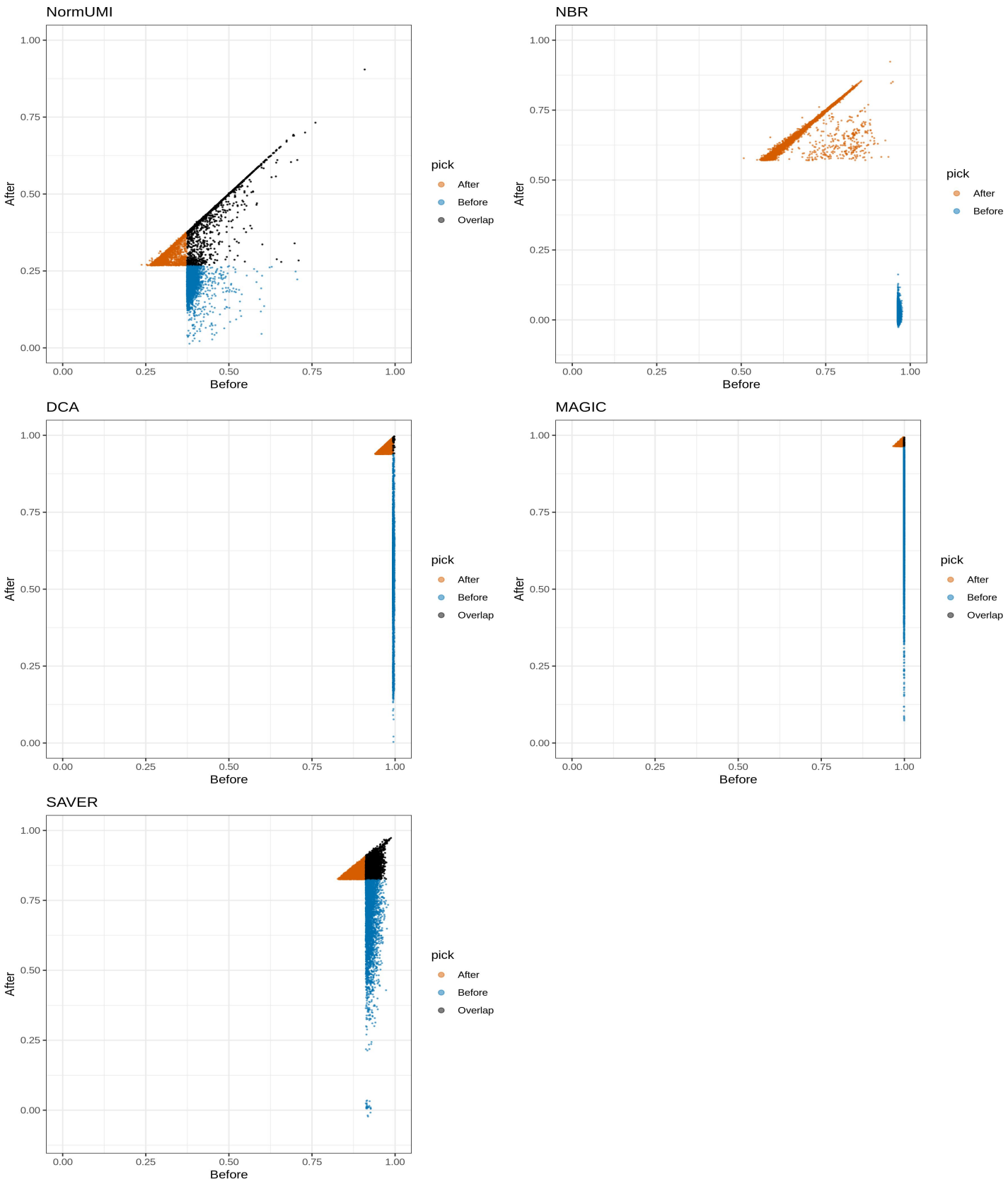


Figure S4. Overlap of the top 5000 gene pairs before and after noise regularization in same method

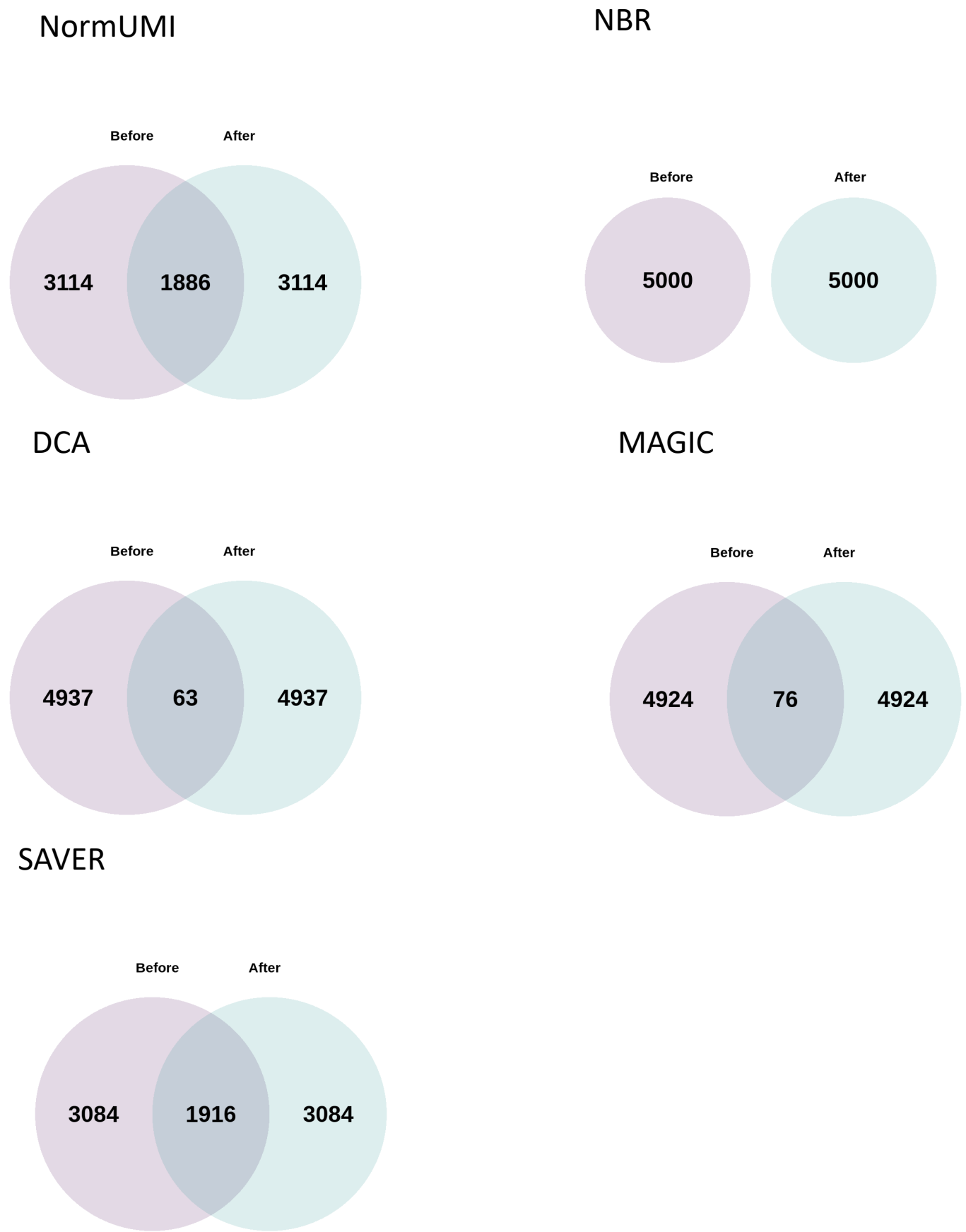
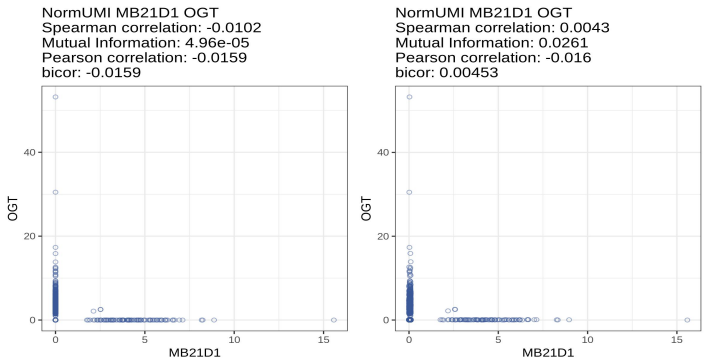
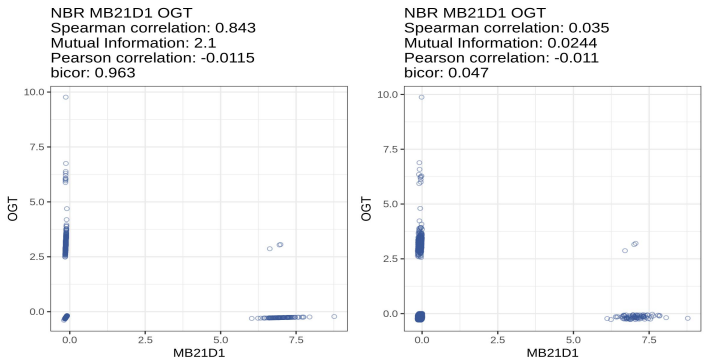
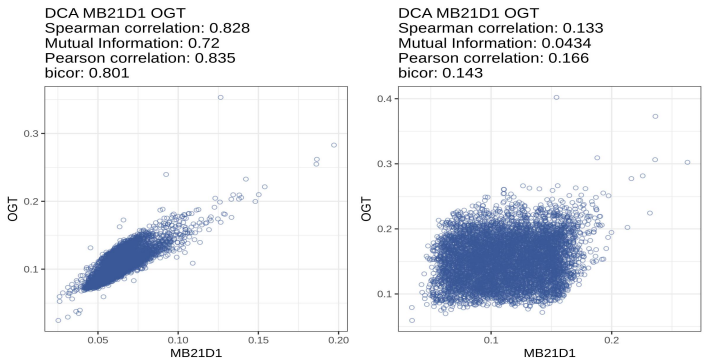
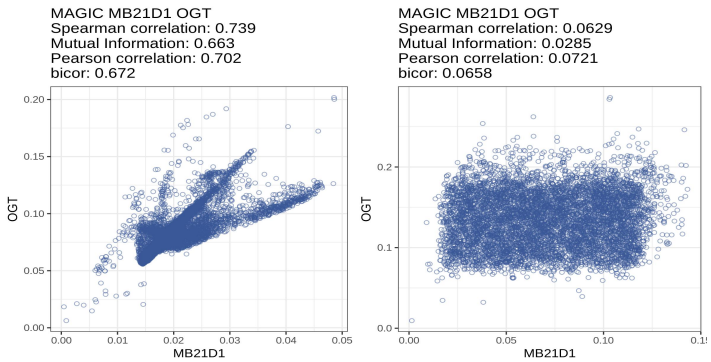


Figure S5. Negative control gene pair before (left) and after (right) noise regularization









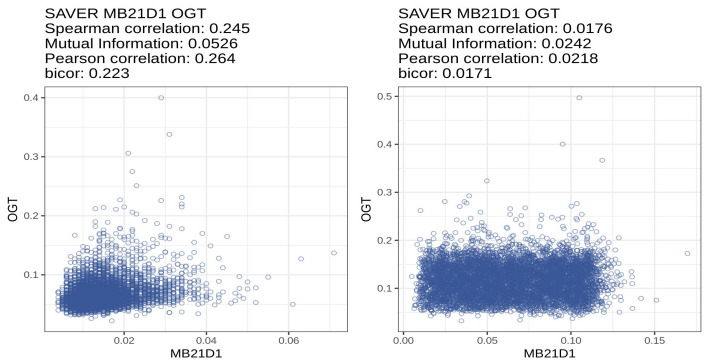
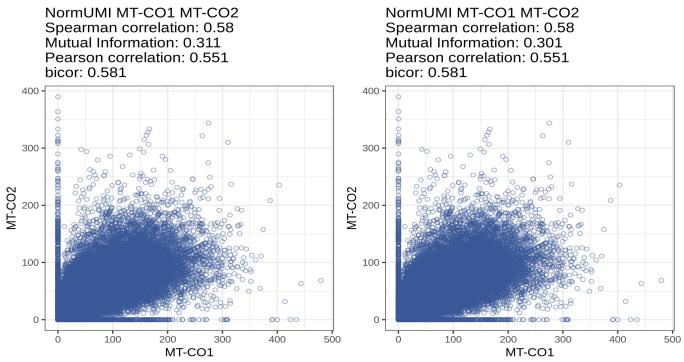
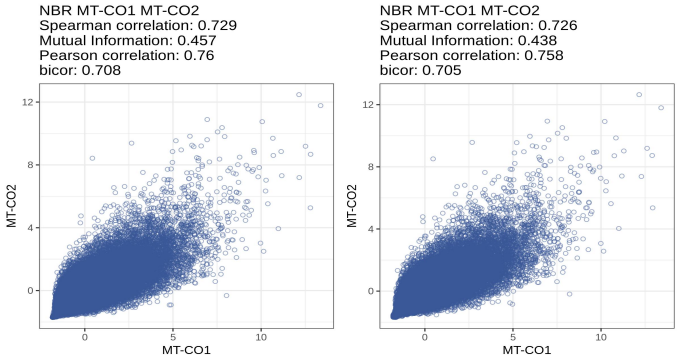
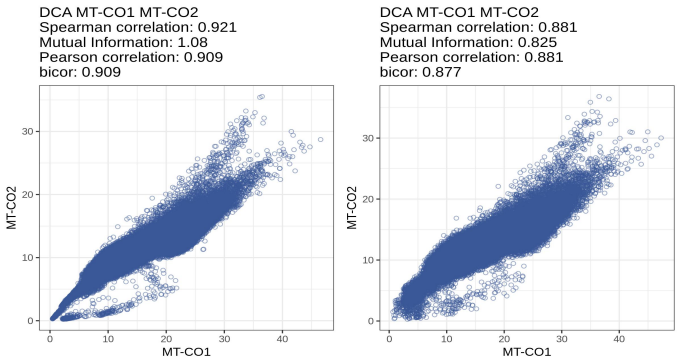
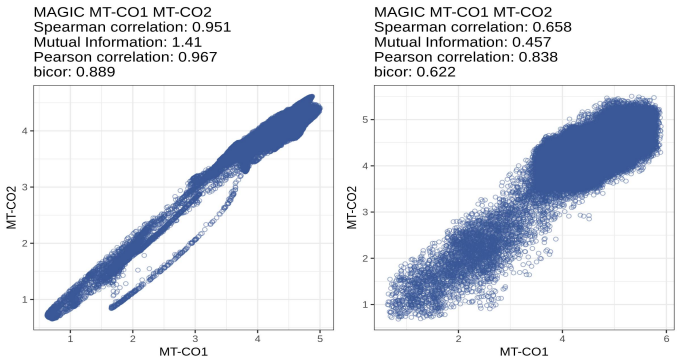


Figure S6 Positive control pair MT-CO1, MT-CO2 before (left) and after (right) noise regularization









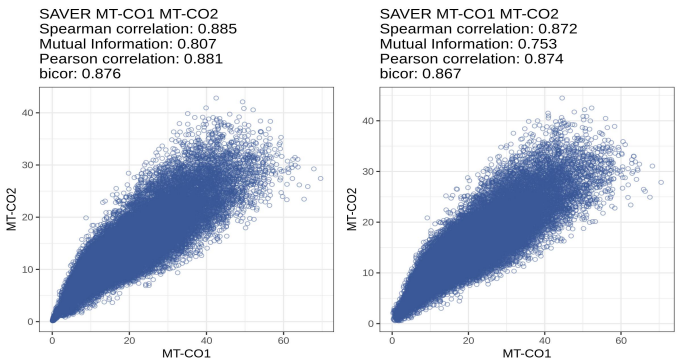


Figure S7. Positive control pair S100A8, S100A9 before (left) and after (right) noise regularization

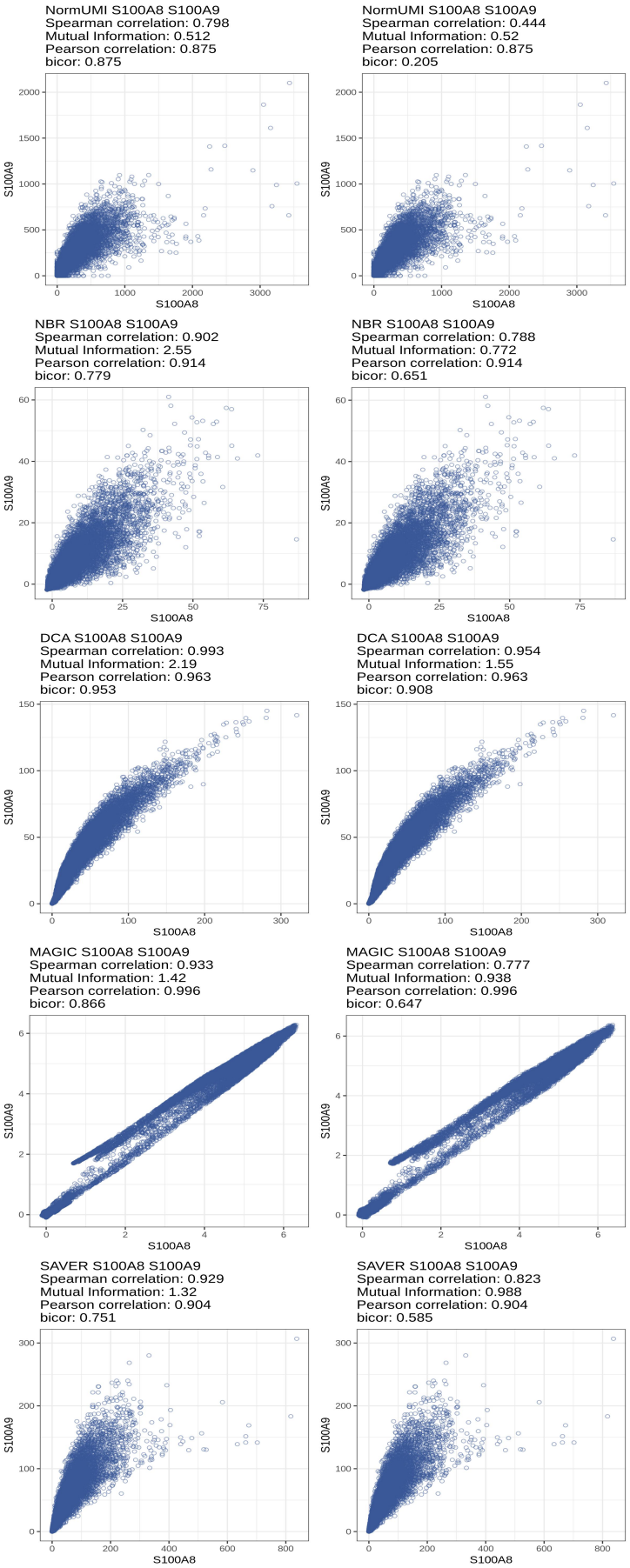
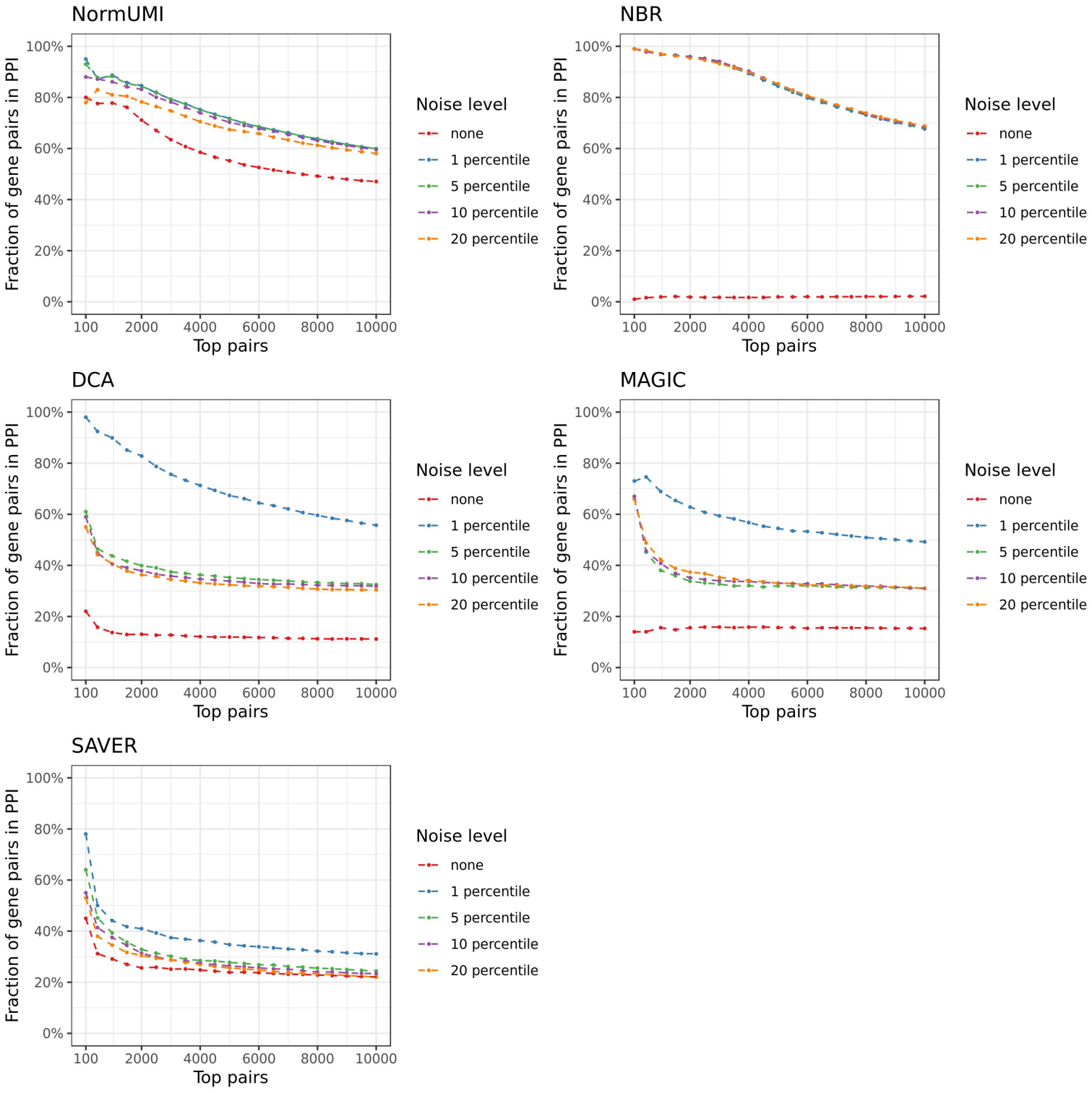


Figure S8. PPI enrichment after adding noise at different levels



Supplementary Figure Legend

Figure S1. HCA single cell data used for this study. (A) UMAP of 50,000 bone marrow cells, covering major immune cell types. (B) Cell type annotations, cell counts and the top 10 markers for the 10 biggest clusters in this dataset.

Figure S2. PPI enrichment of randomly sampled gene pairs. Gene pairs were randomly sampled and overlapped with PPI database to estimate the background enrichment level. The mean of the background enrichment is ~3.6%, error bar represents one standard deviation based on 20 random samplings.

Figure S3. Gene-gene correlation coefficients before and after noise regularization. From each preprocessing methods, top 5000 gene pairs (ranked by correlation coefficients) were selected before and after noise regularization, respectively. The top 5000 pairs before noise regularization were colored as blue, the top pairs selected after regularization were colored as brown, and the overlapped gene pairs were colored as black. The top 5000 gene pairs before regularization (blue and black dots together) had a wide range of correlations after regularization. On the contrary, the top 5000 gene pairs selected after regularization (red and black dots) were also highly correlated before the regularization.

Figure S4. Overlap of the top 5000 gene pairs before and after noise regularization in same method. Venn diagrams of the top 5000 gene pairs selected before and after noise regularization.

Figure S5. Negative control gene pair before (left) and after (right) noise regularization. Scatter plot of expression values of a negative control gene pair, OGT and MB21D1, before and after noise regularization. The oversmoothed data points were randomized and the correlations were effectively diluted after regularization

Figure S6. Positive control pair MT-CO1, MT-CO2 before (left) and after (right) noise regularization. Scatter plot of expression values of an experimentally validated interacting gene pairs: MT-CO1 & MT-CO2, before (left panel) and after (right panel) noise regularization. This gene pairs had high correlation before noise regularization and preserved high correlations with the added noise.

Figure S7. Positive control pair S100A8, S100A9 before (left) and after (right) noise regularization. Scatter plot of expression values of an experimentally validated interacting gene pairs: S100A8 & S100A9, before (left panel) and after (right panel) noise regularization. This gene pairs had high correlation before noise regularization and preserved high correlations with the added noise.

Figure S8. PPI enrichment after adding noise at different levels. Different level of noise is applied to regularize the data (1, 5, 10, 20 percentile of the expression level). Noise at 1 percentile of the expression level produces the optimal PPI enrichment.