Supplementary Information for "Axes of inter-sample variability among transcriptional neighborhoods reveal disease-associated cell states in single-cell data"

Contents

Supplementary Tables **2**

Supplementary Figures **13**

Supplementary Tables

Dataset	# cells	# steps	Avg. nbhd. size
RA	27216	3	336.43
Sepsis	102814	4	1503.33
TB	500089	7	12183.74

Supplementary Table 1: Neighborhood characteristics for each dataset analyzed. For each of the three datasets analyzed in this paper, we show the number of random walk steps selected by CNA as well as the average across neighborhoods of the number of cells required to capture 70% of the mass of each neighborhood, estimated by sampling 500 random neighborhoods.

Pathway	Adjusted P	Enrichment
REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND _PEPTIDE_LOADING_OF_CLASS_I_MHC	7.76e-06	-0.89
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	7.76e-06	-0.85
REACTOME_INTERFERON_GAMMA_SIGNALING	7.76e-06	-0.79
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	7.76e-06	-0.76
REACTOME_INTERFERON_SIGNALING	7.76e-06	-0.73
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION	7.76e-06	-0.70
REACTOME_DOWNSTREAM_TCR_SIGNALING	7.76e-06	-0.65
REACTOME_TCR_SIGNALING	7.76e-06	-0.59
REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION	7.76e-06	-0.59
REACTOME_CELLULAR_RESPONSES_TO_STRESS	7.76e-06	-0.50
REACTOME_NEUTROPHIL_DEGRANULATION	7.76e-06	-0.50
REACTOME_SIGNALING_BY_NOTCH	7.76e-06	-0.49
REACTOME_CELLULAR_RESPONSES_TO_EXTERNAL_STIMULI	7.76e-06	-0.47
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING _PRESENTATION	7.76e-06	-0.45
REACTOME_CELL_CYCLE_CHECKPOINTS	7.76e-06	-0.44

Supplementary Table 2: Gene sets most strongly enriched among gene expression correlations to NAM-PC2. Pathway names are listed, along with FDR-adjusted FGSEA P-values and enrichment scores indicating directionality of enrichment with respect to NAM-PC2. Among all gene sets with the minimum P-value, the top ten—based on ranking by degree of enrichment—are shown. The most enriched gene sets include gene sets related to antigen cross-presentation on MHC-II and interferon signaling. These suggest that the fibroblasts with low values along NAM-PC2 may be activated fibroblasts with high MHC class II gene expression and interferon-mediated signaling; such fibroblasts have been previously hypothesized to be activated in response to interferon gamma from lymphocytes and to play important roles in synovial tissue inflammation.[1]

Phenotype		Reyes et al.	CNA		
Cases	Controls	MS1 FDR	Global P	# NAM-PCs	# nbhds with FDR< 10%
Int-URO, URO, Bac-SEP, ICU-SEP	Control, Leuk-UTI, ICU-NoSEP	NA	7.0×10^{-5}	2	50,696
URO, Int-URO	Control, Leuk-UTI	$< 10^{-3}$	2.8×10^{-4}	2	25,875
Bac-SEP, ICU-SEP	Control	$< 10^{-3}$	5.0×10^{-4}	2	0
URO, Int-URO Bac-SEP, ICU-SEP	ICU-NoSEP	0.27	0.86	3	0
Leuk-UTI	Control	6.0×10^{-2}	0.21	4	0
Int-URO	Control	$< 10^{-3}$	5.9×10^{-4}	2	0
URO	Control	$< 10^{-3}$	1.5×10^{-2}	4	0
Bac-SEP	Control	3.0×10^{-3}	0.12	2	0
ICU-SEP	Control	$< 10^{-3}$	2.6×10^{-3}	2	0
ICU-NoSEP	Control	$< 10^{-3}$	1.6×10^{-2}	2	0

Supplementary Table 3: Sub-cohort assessments for sepsis dataset. The original authors did not perform an aggregated sepsis versus non-sepsis association test. Rather, they compared every patient group to healthy controls (6 tests), in addition to the following tests: Int-URO and URO vs. Control or Leuk-UTI; Bac-SEP and ICU-SEP vs. Control; Int-URO, URO, Bac-EP and ICU-SEP vs. ICU-NoSEP. We performed association tests for these same 9 patient groupings using CNA as well as an aggregated sepsis vs. no sepsis phenotype (Int-URO, URO, Bac-SEP, and ICU-SEP vs. Leuk-UTI and Controls). The aggregated phenotype analysis (sepsis vs. no sepsis) was used for downstream analysis.

Pathway	Adjusted P-value	Enrichment
PID_RAC1_PATHWAY	2.06e-04	0.71
PID_PDGFRB_PATHWAY	1.33e-03	0.52
PID_TOLL_ENDOGENOUS_PATHWAY	1.62e-03	0.74
PID_ERBB1_DOWNSTREAM_PATHWAY	1.79e-03	0.54
PID_CDC42_PATHWAY	2.16e-03	0.61
PID_TXA2PATHWAY	2.68e-03	0.64
PID_IL6_7_PATHWAY	1.54e-02	0.58
PID_IL8_CXCR2_PATHWAY	1.77e-02	0.62
PID_AMB2_NEUTROPHILS_PATHWAY	2.01e-02	0.65
PID_LYSOPHOSPHOLIPID_PATHWAY	2.71e-02	0.59
PID_P38_ALPHA_BETA_PATHWAY	3.03e-02	0.63
PID_THROMBIN_PAR1_PATHWAY	3.07e-02	0.59
PID_CASPASE_PATHWAY	3.19e-02	0.56
PID_EPO_PATHWAY	3.88e-02	0.60
PID_ECADHERIN_NASCENT_AJ_PATHWAY	4.66e-02	0.58

Supplementary Table 4: Gene sets most enriched genes that globally distinguish expanded from depleted cells populations in sepsis. Gene-set enrichment analysis was performed where the input rank list of genes was computed based on the correlation across cells between gene expression and neighborhood coefficients for the sepsis phenotype. FGSEA p-values are shown. Many of these gene sets have established links to sepsis: *PDGFRB* knockout attenuates brain inflammation in sepsis,[2] ERBB1 (EGFR) inhibition in vivo blocks septic shock,[3] TLR signaling contributes to cytokine storms in sepsis,[4] and *CDC42* is beneficially upregulated in sepsis to restore endothelial barrier function and decrease edema.[5]

Pathway	Adjusted P-value	Enrichment	Comparator	Population
PID_TELOMERASE_PATHWAY	1.25e-02	-0.57	Cluster	BS1
PID_RAC1_PATHWAY	3.45e-02	-0.55	Cluster	BS1
PID_HDAC_CLASSII_PATHWAY	4.15e-02	-0.61	Cluster	BS1
PID_IL12_2PATHWAY	1.74e-02	-0.54	Cluster	BS2
PID_HIF1_TFPATHWAY	1.72e-02	-0.55	Cluster	DS1
PID_ERBB1_DOWNSTREAM_PATHWAY	3.75e-02	-0.45	Cluster	DS1
PID_RAC1_PATHWAY	1.35e-02	-0.57	Cluster	DS2
PID_TOLL_ENDOGENOUS_PATHWAY	4.62e-04	-0.80	Cluster	MS4
PID_RAC1_PATHWAY	5.29e-03	-0.63	Cluster	MS4
PID_CDC42_PATHWAY	1.63e-02	-0.59	Cluster	MS4
PID_IL6_7_PATHWAY	3.66e-02	-0.58	Cluster	MS4
PID_AVB3_OPN_PATHWAY	3.91e-02	-0.64	Cluster	MS4
PID_IL12_2PATHWAY	2.38e-04	-0.69	Cluster	TS1
PID_CD8_TCR_PATHWAY	4.25e-04	-0.64	Cluster	TS1
PID_CD8_TCR_DOWNSTREAM_PATHWAY	1.36e-03	-0.65	Cluster	TS1
PID_AMB2_NEUTROPHILS_PATHWAY	4.98e-03	-0.71	Cluster	TS1
PID_CXCR4_PATHWAY	8.75e-03	-0.52	Cluster	TS1
PID_TCR_PATHWAY	1.54e-02	-0.54	Cluster	TS1
PID_IL12_STAT4_PATHWAY	3.12e-02	-0.63	Cluster	TS1
PID_HIF1_TFPATHWAY	3.00e-02	-0.54	Cluster	TS2
PID_IL2_PI3K_PATHWAY	4.76e-02	-0.61	Cluster	TS2

Supplementary Table 5: Gene sets distinguishing cluster sub-populations depleted among patients with sepsis from the non-associated cells in the same cluster (*e.g.*, the depleted population of TS1 is compared to the rest of TS1). This table lists the numerical within-cluster gene-set enrichments underlying Supplementary Figure 9. FGSEA p-values are shown.

Pathway	Adjusted P-value	Enrichment	Comparator	Population
PID_RAC1_PATHWAY	1.86e-02	-0.54	Major cell type	BS1
PID_ERBB1_DOWNSTREAM_PATHWAY	4.40e-02	-0.44	Major cell type	BS1
PID_RAC1_PATHWAY	1.86e-02	-0.54	Major cell type	BS1
PID_RAC1_PATHWAY	2.93e-02	-0.54	Major cell type	DS2
PID_CDC42_PATHWAY	2.94e-02	-0.54	Major cell type	DS2
PID_HIF1_TFPATHWAY	3.42e-02	-0.54	Major cell type	DS2
PID_ERBB1_DOWNSTREAM_PATHWAY	3.72e-02	-0.47	Major cell type	DS2
PID_RAC1_PATHWAY	7.40e-04	-0.61	Major cell type	MS4
PID_HIF1_TFPATHWAY	5.19e-03	-0.56	Major cell type	MS4
PID_AMB2_NEUTROPHILS_PATHWAY	9.29e-03	-0.63	Major cell type	MS4
PID_ERBB1_DOWNSTREAM_PATHWAY	9.51e-03	-0.46	Major cell type	MS4
PID_IL8_CXCR2_PATHWAY	1.62e-02	-0.58	Major cell type	MS4
PID_CDC42_PATHWAY	2.13e-02	-0.51	Major cell type	MS4
PID_TXA2PATHWAY	2.60e-02	-0.53	Major cell type	MS4
PID_TRKR_PATHWAY	3.01e-02	-0.54	Major cell type	MS4
PID_PDGFRB_PATHWAY	3.39e-02	-0.41	Major cell type	MS4
PID_HDAC_CLASSII_PATHWAY	3.95e-02	-0.56	Major cell type	MS4
PID_TOLL_ENDOGENOUS_PATHWAY	4.17e-02	-0.59	Major cell type	MS4
PID_IL6_7_PATHWAY	4.51e-02	-0.50	Major cell type	MS4
PID_MYC_ACTIV_PATHWAY	2.49e-02	-0.57	Major cell type	TS1
PID_CD8_TCR_PATHWAY	4.38e-02	-0.57	Major cell type	TS1
PID_MYC_ACTIV_PATHWAY	2.49e-02	-0.57	Major cell type	TS1
PID_CD8_TCR_PATHWAY	4.38e-02	-0.57	Major cell type	TS1
PID_MYC_ACTIV_PATHWAY	2.49e-02	-0.57	Major cell type	TS1
PID_CD8_TCR_PATHWAY	4.38e-02	-0.57	Major cell type	TS1
PID_MYC_ACTIV_PATHWAY	2.49e-02	-0.57	Major cell type	TS1
PID_IL12_2PATHWAY	1.57e-04	-0.65	Major cell type	TS2
PID_AMB2_NEUTROPHILS_PATHWAY	1.02e-03	-0.69	Major cell type	TS2
PID_CD8_TCR_PATHWAY	3.10e-03	-0.54	Major cell type	TS2
PID_HIF1_TFPATHWAY	3.33e-03	-0.57	Major cell type	TS2
PID_CD8_TCR_DOWNSTREAM_PATHWAY	3.80e-03	-0.56	Major cell type	TS2
PID_IL12_STAT4_PATHWAY	1.43e-02	-0.60	Major cell type	TS2
PID_RAC1_PATHWAY	2.43e-02	-0.52	Major cell type	TS2
PID_HDAC_CLASSII_PATHWAY	3.48e-02	-0.57	Major cell type	TS2
PID_TXA2PATHWAY	3.69e-02	-0.53	Major cell type	TS2
PID_CASPASE_PATHWAY	4.37e-02	-0.51	Major cell type	TS2

Supplementary Table 6: Pathways distinguishing within-cluster populations depleted among patients with sepsis from the remaining cells in the same major cell type (*e.g.*, the depleted population of TS1 is compared to all T cells). This table lists the numerical within-major-cell-type gene-set enrichments underlying Supplementary Figure 9. FGSEA p-values are shown.

	T	B	Mono	NK	DC
NAM-PC1	-0.08	0.15	-0.13	0.55	-0.10
NAM-PC2	0.84	0.53	-0.89	0.30	0.41
NAM-PC3	-0.23	-0.11	0.23	-0.05	-0.22
NAM-PC4	0.25	-0.15	-0.21	0.18	-0.06
NAM-PC5	-0.03	0.10	0.01	-0.07	0.27

Supplementary Table 7: Relationships of NAM-PCs to abundances of major cell types. Correlations between each sample's fraction of cells from a given major cell type and that sample's loading along each of the first five NAM-PCs of the sepsis dataset.

	T	B	Mono	NK	DC
NAM-PC1	0.51	0.24	0.31	1.7e-06	0.42
NAM-PC2	4.0e-18	5.1e-06	1.1e-22	0.01	6.5e-04
NAM-PC3	0.06	0.39	0.06	0.68	0.09
NAM-PC4	0.05	0.22	0.1	0.15	0.64
NAM-PC5	0.82	0.44	0.92	0.57	0.03

Supplementary Table 8: Statistical significance of relationships of NAM-PCs to abundances of major cell types. Analytical P-values corresponding to the correlations shown in Supplementary Table 7. (See Methods.)

Pathway	Adjusted P-Value	Enrichment	NAM-PC
PID_RAC1_PATHWAY	1.55e-02	0.52	NAMPC1
PID_HIF1_TFPATHWAY	1.60e-02	0.52	NAMPC1
PID_AMB2_NEUTROPHILS_PATHWAY	1.65e-02	0.61	NAMPC1
PID_RAC1_PATHWAY	8.08e-04	-0.72	NAMPC2
PID_IL12_2PATHWAY	1.17e-03	0.61	NAMPC2
PID_CD8_TCR_PATHWAY	1.18e-03	0.62	NAMPC2
PID_CD8_TCR_DOWNSTREAM_PATHWAY	1.63e-03	0.63	NAMPC2
PID_CDC42_PATHWAY	2.32e-03	-0.67	NAMPC2
PID_TOLL_ENDOGENOUS_PATHWAY	2.61e-03	-0.80	NAMPC2
PID_IL2_STAT5_PATHWAY	1.36e-02	0.69	NAMPC2
PID_PDGFRB_PATHWAY	1.48e-02	-0.51	NAMPC2
PID_IL8_CXCR2_PATHWAY	2.28e-02	-0.67	NAMPC2
PID_ENDOTHELIN_PATHWAY	3.15e-02	-0.63	NAMPC2
PID_RAC1_PATHWAY	1.34e-03	0.61	NAMPC3
PID_ERBB1_DOWNSTREAM_PATHWAY	6.40e-03	0.50	NAMPC3
PID_TOLL_ENDOGENOUS_PATHWAY	6.53e-03	0.70	NAMPC3
PID_IL6_7_PATHWAY	1.88e-02	0.56	NAMPC3
PID_PDGFRB_PATHWAY	2.46e-02	0.45	NAMPC3
PID_PTP1B_PATHWAY	2.61e-02	0.58	NAMPC3
PID_AMB2_NEUTROPHILS_PATHWAY	3.09e-02	0.62	NAMPC3
PID_CD8_TCR_DOWNSTREAM_PATHWAY	3.71e-02	-0.35	NAMPC3
PID_IL12_2PATHWAY	7.11e-04	0.73	NAMPC4
PID_CD8_TCR_PATHWAY	7.11e-04	0.74	NAMPC4
PID_CD8_TCR_DOWNSTREAM_PATHWAY	7.11e-04	0.84	NAMPC4
PID_IL2_STAT5_PATHWAY	1.59e-03	0.80	NAMPC4
PID_TOLL_ENDOGENOUS_PATHWAY	1.58e-02	-0.72	NAMPC4
PID_IL2_PI3K_PATHWAY	2.36e-02	0.71	NAMPC4
PID_RAC1_PATHWAY	2.75e-02	-0.54	NAMPC4
PID_IL2_1PATHWAY	3.54e-02	0.59	NAMPC4
PID_TCR_PATHWAY	4.14e-02	0.54	NAMPC4
PID_CD8_TCR_PATHWAY	1.13e-03	-0.67	NAMPC5
PID_HIF1_TFPATHWAY	4.14e-02	-0.57	NAMPC5
PID_CD8_TCR_PATHWAY	1.13e-03	-0.67	NAMPC5

Supplementary Table 9: Gene sets most strongly enriched among gene expression correlations to each of the first five NAM-PCs in the sepsis dataset. Pathway names are listed, along with FDR-adjusted FGSEA P-values and enrichment scores indicating directionality of enrichment with respect to the given NAM-PC.

Name	Type	Correlation	P.value
CD62L	Protein	0.15	<1e-10
SELL	Gene	0.24	<1e-10
NKG7	Gene	-0.81	<1e-10
GZMH	Gene	-0.76	<1e-10
CCL5	Gene	-0.70	<1e-10
GZMA	Gene	-0.67	<1e-10
GNLY	Gene	-0.67	<1e-10

Supplementary Table 10: Gene expression correlations to neighborhood loading on NAM-PC1 of the TBRU dataset reflecting a spectrum of "innateness". Correlations were computed across cells between gene expression and each cell's anchored neighborhood's loading on NAM-PC1 from the joint CCA representation of the TBRU dataset. Cells with higher transcriptional 'innateness' signature tended to have lower loadings on NAM-PC1. Consistent with this pattern, CD62L/*SELL* expression is positively correlated with neighborhood loading on NAM-PC1 and expression levels of effector molecules are negatively correlated with neighborhood loading on NAM-PC1. Naive P-values were computed analytically by treating cells as observations.

Name	Type	Correlation	P.value
XIST	Gene	-0.41	<1e-10
RPS4Y1	Gene	0.58	<1e-10
DDX3Y	Gene	0.35	<1e-10
UTY	Gene	0.30	<1e-10
TTTY15	Gene	0.26	<1e-10
KDM5D	Gene	0.18	<1e-10
USP9Y	Gene	0.16	<1e-10

Supplementary Table 11: NAM-PC4 neighborhood loadings in the TBRU dataset correlate most strongly with expression of sex chromosome genes. Correlations across cells between gene expression and each cell's anchored neighborhood's loading on NAM-PC4 of the non-harmonized mRNA representation of the TBRU dataset. The genes whose expression was most correlated or most anti-correlated with cell loading, shown here, were located on sex chromosomes. Naive P-values were computed analytically by treating cells as observations.

Name	Type	Correlation	P.value
CD4	Protein	-0.27	<1e-10
CD8	Protein	0.33	<1e-10

Supplementary Table 12: Differential expression of selected surface proteins on NAM-PC2. Correlations across cells between CD4 and CD8 protein expression and each cell's anchored neighborhood's loading on NAM-PC2 of the non-harmonized mRNA representation of the TBRU dataset. Previous literature on sex differences in the immune system has documented decreased CD4+/CD8+ ratio among males compared to females.[6]

Ratio	Correlation to NAM-PC2	P-value
Regulatory T cells / Total CD4+ T cells	0.16	8.83e-3
CD4+ T cells / CD8+ T cells	-0.42	2.69e-13

Supplementary Table 13: Correlations across samples between NAM-PC2 and abundance of cell populations known to vary with sex. We computed for each sample the fraction of its CD4+ cells that were T-regulatory as well as the ratio between its CD4+ cell abundance and its CD8+ cell abundance. We correlated each of these ratios across samples with each sample's loading on NAM-PC2 of the non-harmonized mRNA representation of the TBRU dataset. Samples with higher loadings on NAM-PC2 – which are more likely to come from patients identified as male, as shown in Figure 5 – are more likely to have a lower ratio of CD4+ T cells to CD8+ T cells as well as a higher fraction of CD4+ T cells that are T-regulatory cells. Previous literature on sex differences in the immune system has documented decreased CD4+/CD8+ ratio and increased T-regulatory cells among males compared to females.[6]

Component	Gene Expression PC Cor.	NAM-PC Cor.
1	0.0544	0.0246
2	0.2544	0.4557
3	-0.0005	0.0389
4	0.0731	0.7351
5	-0.1420	-0.0183
6	-0.0462	-0.2847
7	-0.2168	-0.0351
8	-0.2799	0.0668
9	-0.3571	0.0382
10	0.1858	0.1147
11	-0.2302	-0.0092
12	0.3657	-0.0433
13	0.1208	-0.0239
14	-0.2933	-0.0042
15	-0.0893	-0.0571
16	-0.5972	-0.0260
17	-0.5689	0.0251
18	-0.8125	0.0608
19	0.8397	0.0789
20	0.7194	-0.0851

Supplementary Table 14: Correlation to donor sex for each of the top 20 NAM-PCs and each of the top 20 gene expression PCs. Correlations for NAM-PCs were computed between sample loadings on each NAM-PC and donor sex. For gene expression PCs, which do not have sample loadings, we aggregated the cell-loadings to the sample level by taking the mean across all cells from a given sample; these values were then correlated to donor sex. We note that while the NAM-PC sample loadings are orthogonal to each other, the per-sample values generated by our averaged gene expression PCs are not. As a result, squared correlations can be summed across NAM-PCs to obtain a joint prediction R^2 but the same cannot be done with the gene expression PCs since these are correlated with each other across samples.

Name	Type	Correlation	P.value
GZMH	Gene	0.58	<1e-10
NKG7	Gene	0.55	<1e-10
FGFBP2	Gene	0.53	<1e-10
GNLY	Gene	0.49	<1e-10
CD244/2B4	Protein	0.44	<1e-10
ZEB2	Gene	0.43	<1e-10
CCL4	Gene	0.39	<1e-10
CD279/PD-1	Protein	0.20	<1e-10
CD29	Protein	0.14	<1e-10
CD161	Protein	-0.17	<1e-10
CD27	Gene	-0.21	<1e-10
KLRB1	Gene	-0.23	<1e-10
IL7R	Gene	-0.25	<1e-10
AQP3	Gene	-0.26	<1e-10
CD127/IL-7R	Protein	-0.28	<1e-10
CD26	Protein	-0.39	<1e-10
CD196/CCR6	Protein	-0.41	<1e-10

Supplementary Table 15: Genes and proteins associated with CNA populations for TB progression. Correlations between expression of selected proteins/genes and the per-cell neighborhood coefficient computed by CNA for TB progressor status. Naive P-values were computed analytically by treating cells as observations.

Attribute tested	Covariates
Age	Weight, # scars, Winter blood draw TB progression
Height	Weight, % European ancestry, Male sex Edu. below high school, Winter blood draw, TB progression
Weight	Age, Height, % European ancestry Edu. below high school, TB progression
% European ancestry	Height, Weight, Edu. below high school Winter blood draw, TB progression
# scars	Age, BCG scar, Male sex TB progression
BCG scar	# scars, Male sex, Winter blood draw
BCG vaccine	-
Male sex	Height, # scars, BCG scar TB progression
Edu. below high school	Height, Weight, % European ancestry
High SES	-
Low SES	-
Medium SES	-
Smoking status	-
Alcohol use	-
Underweight	-
Spring blood draw	Winter blood draw, TB progression
Winter blood draw	Age, Height, % European ancestry BCG scar, Spring blood draw

Supplementary Table 16: Covariate control for TB phenotypes tested. For each sample attribute analyzed using CNA and MASC, we list the sample attributes included as covariates in the analysis.

Phenotype	CNA P	Var. Exp.	Clusters+	Clusters-	CNA+	CNA-	% Replic.	% Novel
Age	< 1.0e-06	47%	67882	41228	51032	118976	68	44
Winter	< 1.0e-06	26%	0	90946	30826	49112	44	50
Sex	2.0e-06	28%	17085	77483	28507	147172	65	35
Ancestry	3.9e-04	16%	NA	NA	6114	37230	NA	100

Supplementary Table 17: Survey of associations in TB dataset. Four per-sample outcome variables from the TB dataset were found to have global associations using CNA, of which one did not have a global association by cluster-based association testing using MASC. Global P-values from CNA were subjected to Bonferroni correction for the number of sample attributes in the survey. Global associations for each tested outcome by CNA are tabulated, along with the percent of outcome variance explained by CNA. The total number of cells in all expanded clusters identified by MASC is also shown, where each cluster is considered significant only after Bonferroni correction for the number of clusters in the analysis. Likewise, the total cells in all depleted clusters is shown. The number of cells whose neighborhoods were found to be positively correlated in abundance with the outcome by CNA at an FDR 5% threshold are shown, as well as the number of cells whose neighborhoods were found to be negatively correlated with the outcome by CNA at an FDR 5% threshold. Finally, to illustrate the scope of recapitulated and novel associations, for the phenotypes for which local association were found by both methods we show the fraction of all cells found to belong to expanded or depleted clusters that were also assigned to associated populations by CNA ("% Replicated") and the fraction of all cells assigned to associated populations by CNA that did not belong to expanded or depleted clusters ("% Novel").

Name	Type	Correlation	P.value
CD28.1	Protein	0.50	<1e-10
CD27.1	Protein	0.38	<1e-10
CD196/CCR6	Protein	0.12	<1e-10
CD8a	Protein	0.04	<1e-10
KLRB1	Gene	-0.10	<1e-10
CD4.1	Protein	-0.11	<1e-10
CCL4	Gene	-0.46	<1e-10
CST7	Gene	-0.57	<1e-10
FGFBP2	Gene	-0.60	<1e-10
GNLY	Gene	-0.61	<1e-10
GZMA	Gene	-0.70	<1e-10
GZMH	Gene	-0.70	<1e-10
CCL5	Gene	-0.70	<1e-10
NKG7	Gene	-0.72	<1e-10

Supplementary Table 18: Genes and proteins associated with CNA populations for age. Correlations between intensity (resp. expression level) of selected proteins (resp. genes) and the per-cell neighborhood coefficient computed by CNA for age. Naive P-values were computed analytically by treating cells as observations.

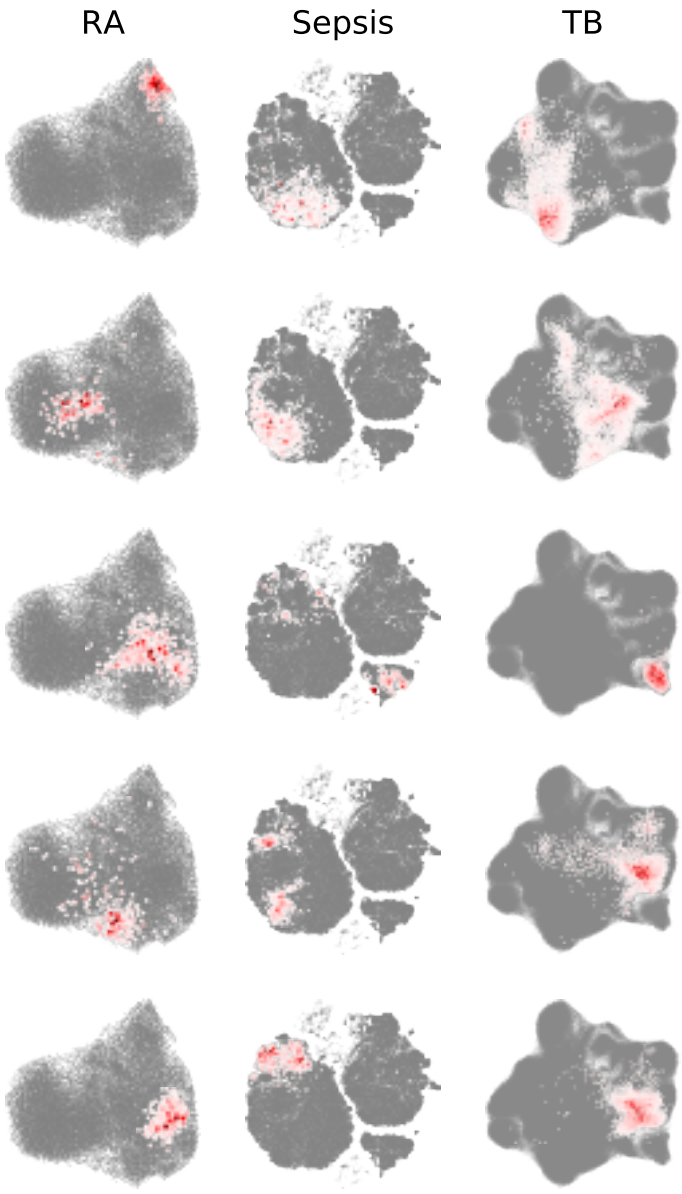
Name	Type	Correlation	P.value
CD194/CCR4	Protein	0.29	<1e-10
GATA3	Gene	0.28	<1e-10
KRT1	Gene	0.27	<1e-10
ANXA1	Gene	0.26	<1e-10
IL7R	Gene	0.24	<1e-10
STAT1	Gene	-0.04	<1e-10
IFNG	Gene	-0.08	<1e-10
CD183/CXCR3	Protein	-0.12	<1e-10
CTLA4	Gene	-0.17	<1e-10
TBC1D4	Gene	-0.19	<1e-10
CD27	Gene	-0.25	<1e-10
TIGIT	Gene	-0.26	<1e-10
LIMS1	Gene	-0.28	<1e-10

Supplementary Table 19: Genes and proteins associated with CNA populations for winter blood draw. Correlations between intensity (resp. expression level) of selected proteins (resp. genes) and the per-cell neighborhood coefficient computed by CNA for winter blood draw. Naive P-values were computed analytically by treating cells as observations.

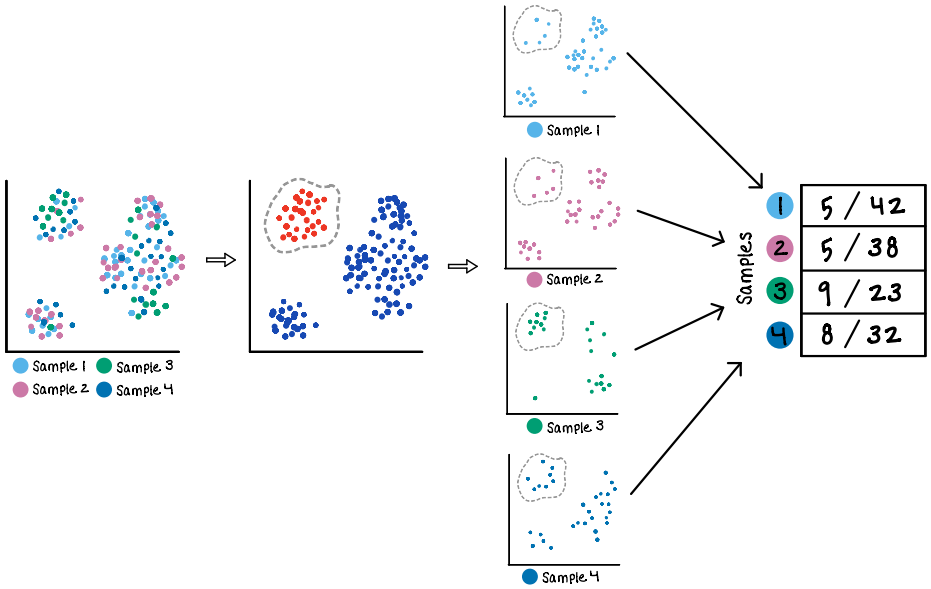
Name	Type	Correlation	P.value
CD194/CCR4	Protein	0.47	<1e-10
LTB	Gene	0.46	<1e-10
LDHB	Gene	0.27	<1e-10
SELL	Gene	0.25	<1e-10
CD62L	Protein	0.24	<1e-10
CD8a	Protein	0.04	<1e-10
CD4.1	Protein	-0.11	<1e-10
CD29	Protein	-0.21	<1e-10
CD244/2B4	Protein	-0.52	<1e-10
GNLY	Gene	-0.61	<1e-10
GZMA	Gene	-0.70	<1e-10
CCL5	Gene	-0.70	<1e-10
GZMH	Gene	-0.70	<1e-10
NKG7	Gene	-0.72	<1e-10

Supplementary Table 20: Genes and proteins associated with CNA populations for European ancestry. Correlations between intensity (resp. expression level) of selected proteins (resp. genes) and the per-cell neighborhood coefficient computed by CNA for European ancestry. Naive P-values were computed analytically by treating cells as observations.

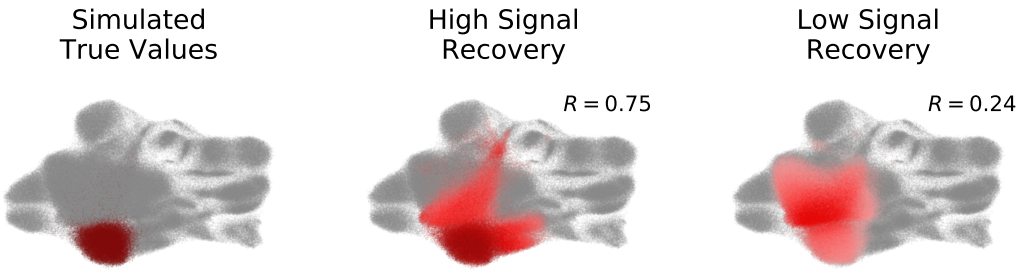
Supplementary Figures



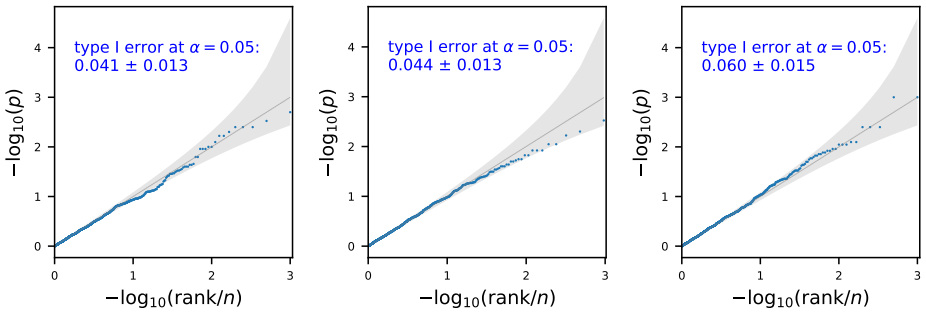
Supplementary Figure 1: Example neighborhoods in three real datasets. Each column shows five example neighborhoods from the indicated dataset among the three datasets analyzed in the paper. Cells are colored according to their degree of belonging to the neighborhood. In each case, only the "bulk" of each neighborhood is shown: that is, only cells with belonging above a certain threshold are shown, with the threshold set such that 70% of the mass of the neighborhood is included.



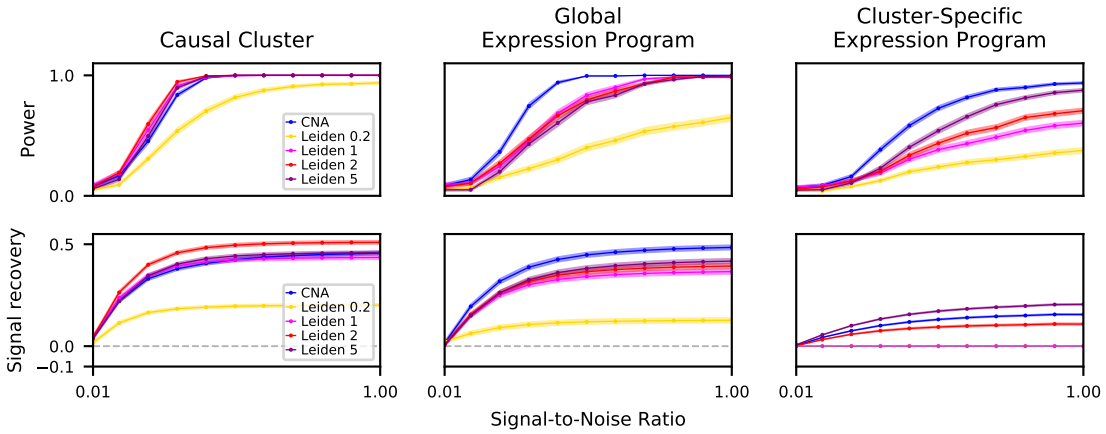
Supplementary Figure 2: Schematic overview of simulation framework. Our simulation framework begins by taking all of the cells in our dataset (**left**) and assigning a ground-truth effect size to each cell (**middle**). We then set each sample's simulated attribute value to be the average effect size across all the cells in the sample (**right**). Finally, noise is added to the ground-truth set of simulated sample attribute values before these values are provided to CNA and the cluster-based comparator method. In this example, there are four samples and we illustrate the construction of a causal cluster signal where the attribute being simulated is the abundance of the top-left cell population in each sample. This is accomplished by setting the ground-truth effect size of all cells in that cluster to 1 and the ground-truth effect size of all cells in the remaining clusters to 0.



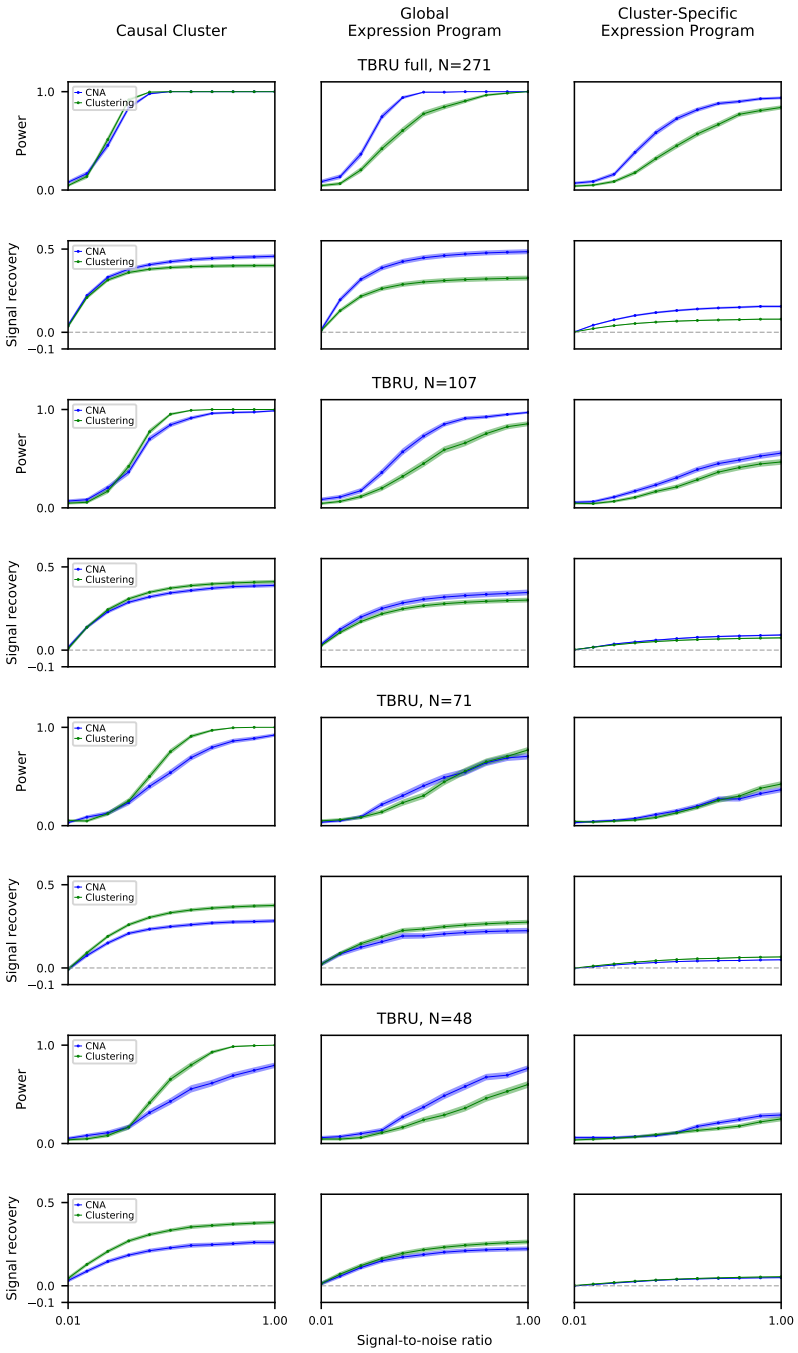
Supplementary Figure 3: Signal recovery, illustrated. Consider a simulated signal of the cluster abundance type, where true values per cell are assigned to be 1 for all cells in the selected cluster and 0 elsewhere (**Left**). Our resulting per-sample values are the fraction of that sample's cells from the selected cluster. A model estimate of this causal population with high signal recovery closely approximates the true direction and degree of abundance association to the simulated per-sample values across transcriptional space (**Center**). A model estimate with lower signal recovery does not accurately identify the true causal population (**Right**).



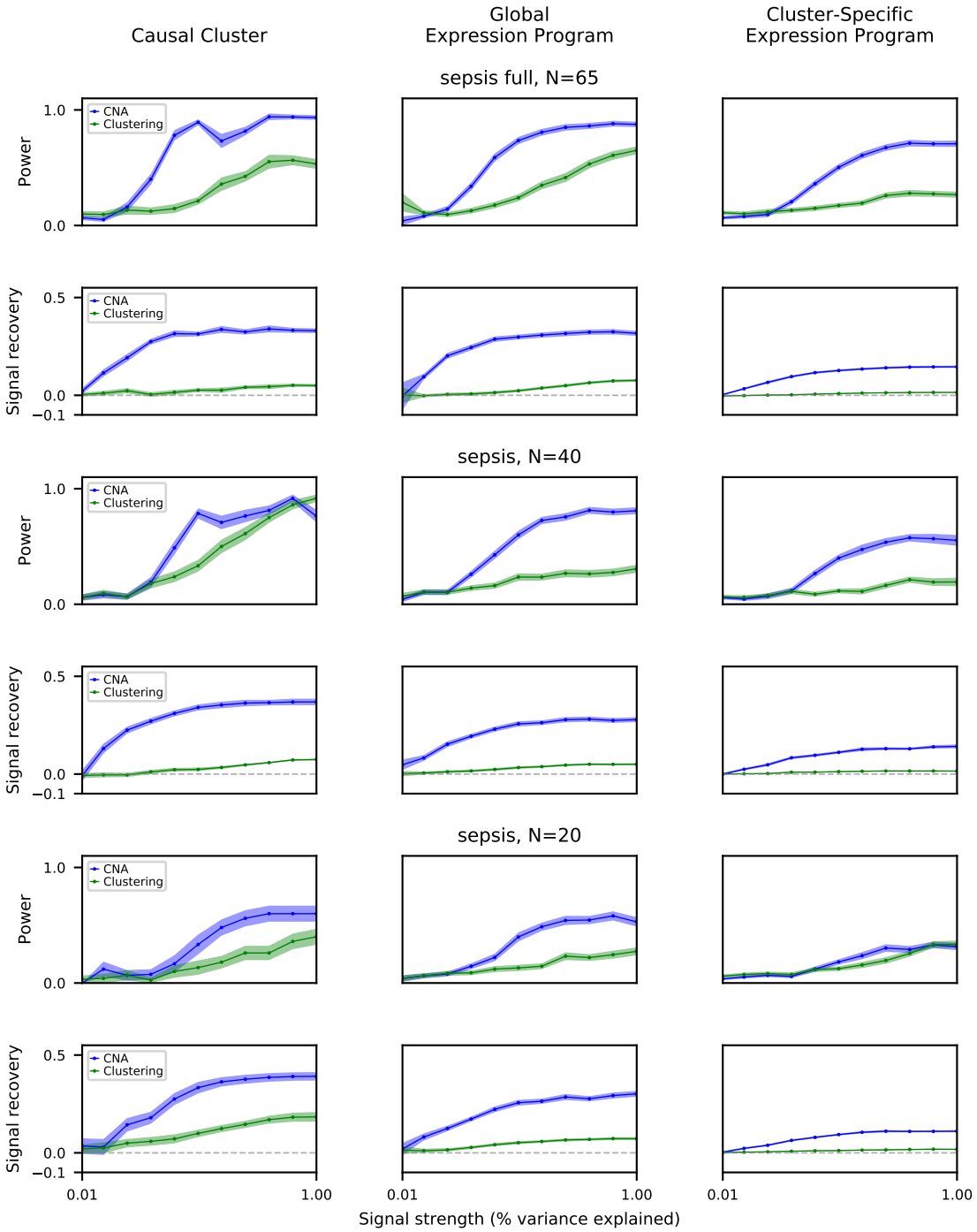
Supplementary Figure 4: Calibration of CNA. P-values from 1,000 trials when CNA is conducted with simulated per-sample outcomes of: patient age values permuted randomly across the dataset (**Left**), or patient age values permuted within batch, to test calibration under moderate batch effects (**Middle**). We then tested calibration under extreme batch effects (**Right**) simulating outcomes with a value of 1 for all samples in a given batch and 0 otherwise, across 1,000 randomly chosen batches. A 95% confidence interval under the null is shown in grey.



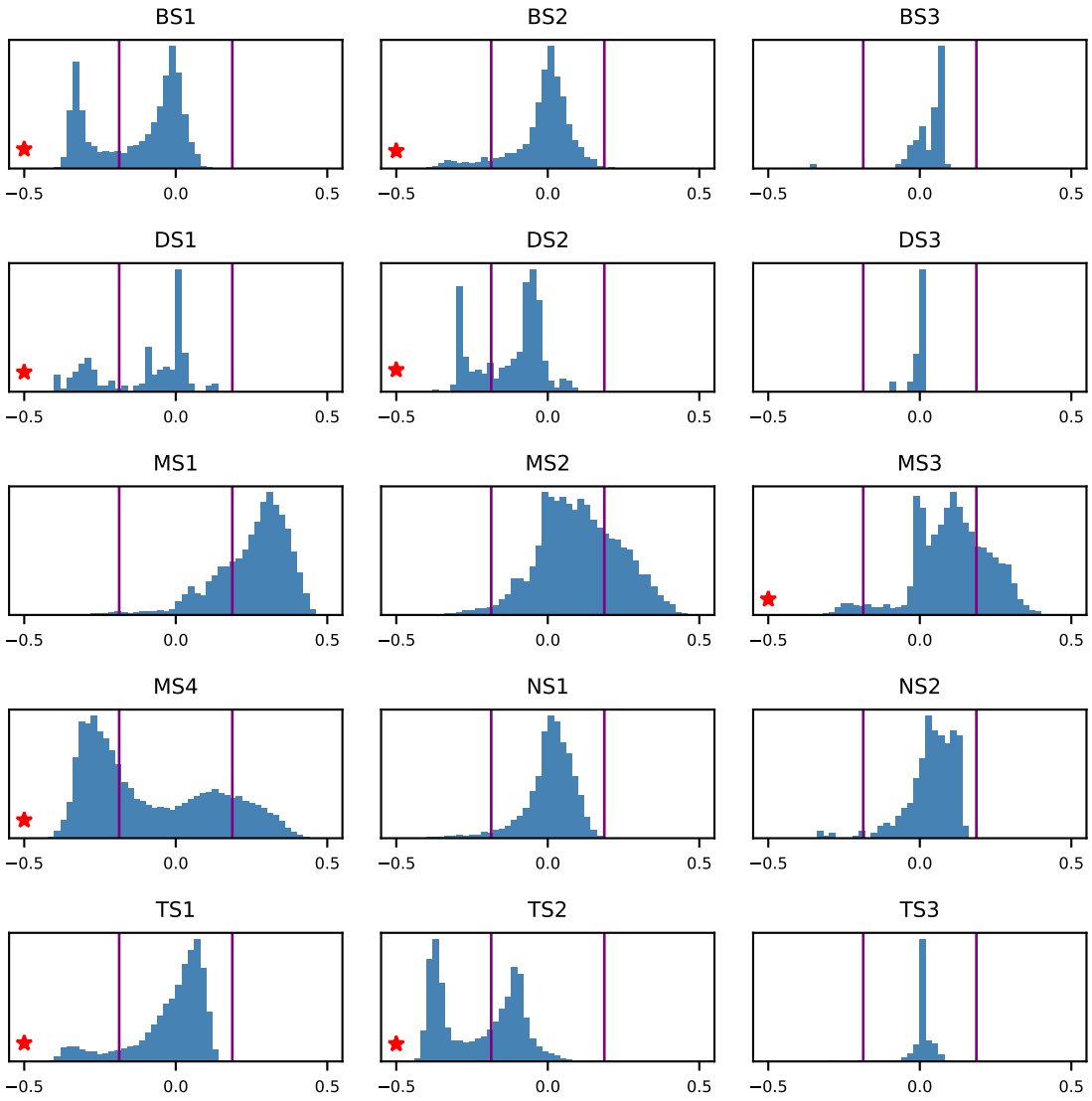
Supplementary Figure 5: Power and signal recovery shown separately by clustering resolution. Three signal types – causal clusters, global gene expression programs, and cluster-specific gene expression programs – were simulated, corresponding to the left, middle and right columns, respectively. For each signal type, we show (**top**) the relative power of CNA versus a cluster-based approach across a range of signal:noise ratios, and (**bottom**) the relative signal recovery of CNA versus a cluster-based approach across a range of noise levels. For power and signal recovery, we plot the mean across all simulations at the given signal-to-noise ratio, as well as the standard error around the mean.



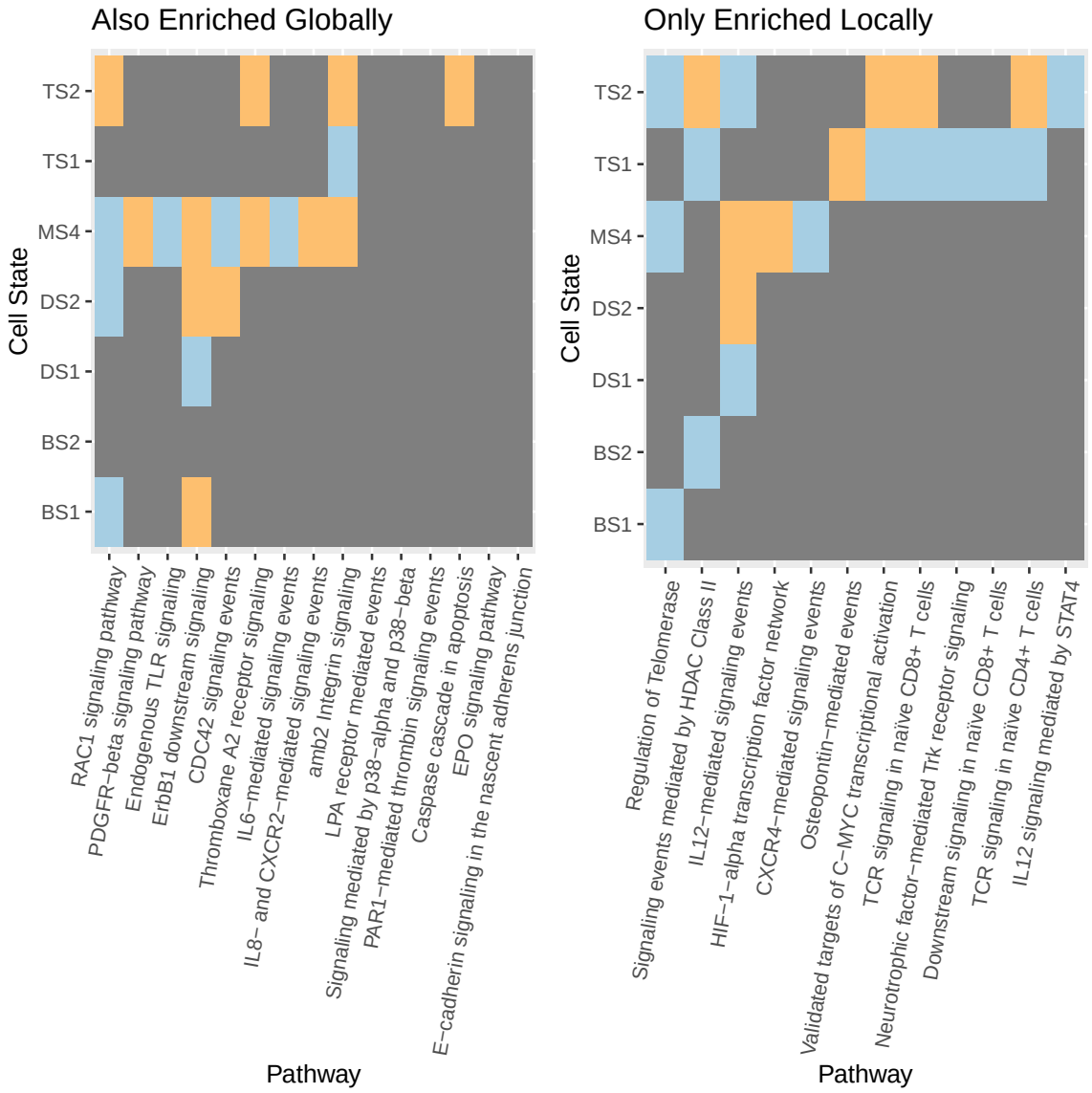
Supplementary Figure 6: Power and signal recovery in full and downsampled versions of the TBRU dataset. Three signal types were simulated analogously to Figure 2 for the full TBRU cohort ($N = 271$, **top two rows**) and three downsampled cohorts (18 batches/$N = 107$, **second two rows**; 12 batches/$N = 71$, **third two rows**; 8 batches/$N = 48$, **bottom two rows**). For each signal type and cohort size, we show (**top** within each pair of rows) the power of CNA versus a cluster-based approach across a range of noise levels, and (**bottom** within each pair of rows) the signal recovery of CNA versus a cluster-based approach across a range of noise levels. For power and signal recovery, we plot the mean across all simulations at the given signal-to-noise ratio, as well as the standard error around the mean. We note that the downsampled TBRU datasets include cell coordinates generated using CCA integration of mRNA and protein information at the full sample size, so there is some information leakage from the full dataset to the downsampled datasets.



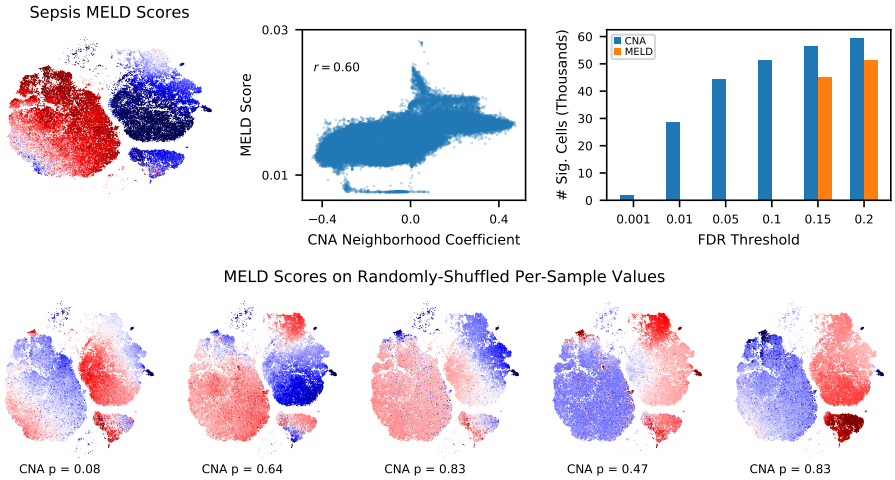
Supplementary Figure 7: Power and signal recovery in full and downsampled versions of the sepsis dataset. Three signal types were simulated analogously to Figure 2 for the full sepsis cohort ($N = 65$, **top two rows**) as well as two downsampled cohorts ($N = 40$, **middle two rows**; $N = 20$, **bottom two rows**). For each signal type and cohort size, we show (**top** within each pair of rows) the relative power of CNA versus a cluster-based approach across a range of noise levels, and (**bottom** within each pair of rows) the relative signal recovery of CNA versus a cluster-based approach across a range of noise levels. For power and signal recovery, we plot the mean across all simulations at the given signal strength, as well as the standard error around the mean.



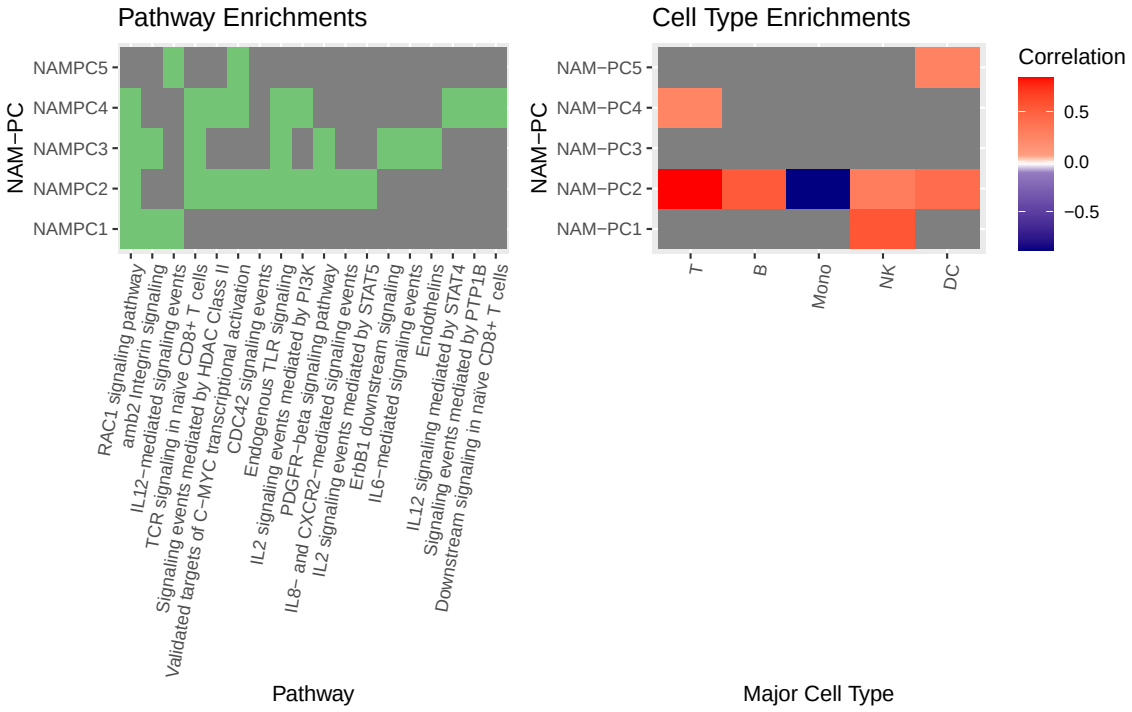
Supplementary Figure 8: Within-cluster heterogeneity in sepsis dataset. Histograms of neighborhood coefficients for the sepsis versus no-sepsis phenotype, within each published cluster for this dataset. The abundance correlation thresholds beyond which the anchor cell for each neighborhood was assigned to an expanded-in-sepsis population (right) or a depleted-in-sepsis population (left) are marked with purple vertical lines. Clusters containing sub-populations with distinct abundance associations to the sepsis phenotype are starred.



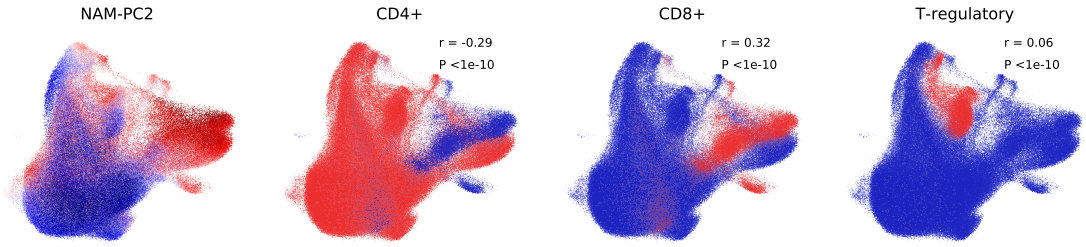
Supplementary Figure 9: Pathways distinguishing within-cluster populations depleted among patients with sepsis. For each cluster containing a distinct sub-population found by CNA to be depleted in sepsis, gene set enrichment analysis was used to characterize the pathways that distinguish that depleted population from closely-related cells. The clusters containing depleted populations are arrayed along the y-axis. Pathways are arrayed along the x-axis. All enrichments found were negative (*i.e.*, decreased use of the given program in the depleted population). Grey indicates lack of enrichment. Orange indicates enrichment of that pathway in the depleted population relative to all other cells in the same published cluster (*e.g.*, TS1). Blue indicates lack of enrichment with respect to cells in the same published cluster but enrichment of that pathway in the depleted population relative to all other cells of the same major cell type (*e.g.*, T cells). The enrichments shown are significant with FDR $< 5\%$. Several enriched pathways reflect patterns also observed globally across all depleted versus all expanded populations. These pathways are shown on the left and labeled "Also Enriched Globally." Other enriched pathways relevant to sepsis only emerge through examination of these more local contrasts between within-cluster depleted populations and closely related non-associated cells. Such pathways are shown on the right and labeled "Only Enriched Locally." For quantitative results of enrichment analyses, see Supplementary Table 5 and Supplementary Table 6.



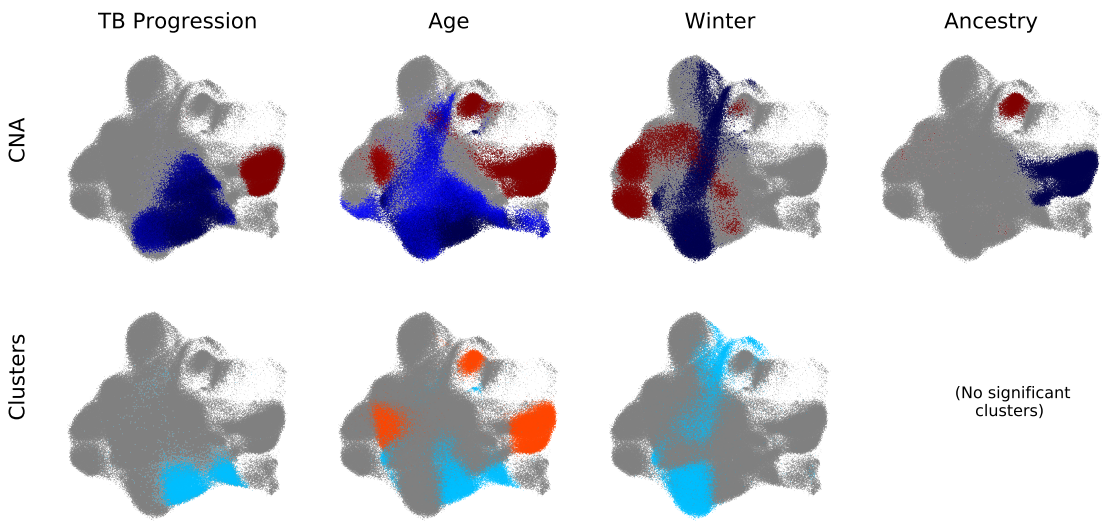
Supplementary Figure 10: Sepsis dataset analysis with MELD. We first applied the MELD algorithm to the sepsis dataset using the true sepsis vs non-sepsis sample attribute values. Per-cell scores from MELD are shown in tSNE space (**Top Left**). MELD scores were correlated with the neighborhood coefficients from CNA (Pearson R=0.60, **Top Middle**). MELD does not report a global significance metric, so we sought to determine whether the MELD score pattern from the sepsis attribute was more striking than MELD score patterns from null comparator attributes. We randomly permuted the sample sepsis vs non-sepsis attribute values five times and examined the resulting MELD scores, which appear to have non-trivial structure across transcriptional space (**Bottom Row**). For reference, we also ran CNA on these null attributes and all CNA global p-values were sub-significant. To evaluate the local significance of MELD scores per-cell from the true sepsis attribute, we applied a permutation-based approach identical to the one used by CNA to assess significance per-neighborhood. None of the individual per-cell MELD scores were significant at FDR $< 5\%$ (**Top Right**). More specifically, we generated 500 null distributions of MELD scores each resulting from a random permutation of the sample sepsis case-control labels. We then used these null instantiations to estimate the false discovery rate for MELD scores. We plot the number of cells (for MELD) and neighborhoods (for CNA) that pass an increasing FDR threshold.



Supplementary Figure 11: Pathway and cell-type enrichments for the first five NAM-PCs of the sepsis dataset. (**Left**) Plot of selected pathway enrichments among the genes most correlated with the neighborhood loadings for each NAM-PC, with green indicating significant enrichment at FDR $< 5\%$ and gray indicating no significant enrichment at FDR $< 5\%$. **Right:** Correlation across samples between the sample loadings for each NAM-PC and the abundances of each of the five main cell types in the dataset, with correlations not achieving nominal significance shown in gray.



Supplementary Figure 12: NAM-PC2 illustrated and compared to relevant cell populations. From left to right: cells in UMAP space colored by their anchored neighborhood's loading on NAM-PC2, with positive loadings in warmer color. Cells in UMAP space colored by membership in any CD4+ cluster (CD4+ cells in red). Cells in UMAP space colored by membership any CD8+ cluster (CD8+ cells in red). Cells in UMAP space colored by membership in the T-regulatory cell cluster (T-regulatory cells in red). We show the correlations between these three cell populations and NAM-PC2. Analytical P-values corresponding to these correlations were computed using a beta-distributed null.



Supplementary Figure 13: Contrasting cell populations implicated by CNA compared to cluster-based analysis. (**Top**) Cell populations implicated by CNA as depleted (blue) or expanded (red) in association with each sample attribute, arrayed left to right. (**Bottom**) Cell populations implicated by cluster-based analysis as depleted (blue) or expanded (orange) in association with each sample attribute. We note that although our cluster-based analysis did not recover significant local associations for ancestry, the cluster-based analysis in the original study did; this difference is primarily due to the two analyses using different ways of choosing which covariates to include as well as the more aggressive multiple-testing correction across phenotypes employed in our analysis.

References

[1] Fan Zhang et al. "Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry". In: *Nature immunology* 20.7 (2019), pp. 928–942.

[2] Lihui Duan et al. "PDGFRβ Cells Rapidly Relay Inflammatory Signal from the Circulatory System to Neurons via Chemokine CCL2". eng. In: *Neuron* 100.1 (Oct. 2018), 183–200.e8. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2018.08.030.

[3] Saurabh Chattopadhyay et al. "EGFR Kinase Activity Is Required for TLR4 Signaling and the Septic Shock Response". In: *EMBO reports* 16.11 (Nov. 2015), pp. 1535–1547. ISSN: 1469-221X. DOI: 10.15252/embr.201540337.

[4] V. Kumar. "Toll-like Receptors in Sepsis-Associated Cytokine Storm and Their Endogenous Negative Regulators as Future Immunomodulatory Targets". en. In: *International Immunopharmacology* 89 (Dec. 2020), p. 107087. ISSN: 1567-5769. DOI: 10.1016/j.intimp.2020.107087.

[5] J. Amado-Azevedo et al. "A CDC42-Centered Signaling Unit Is a Dominant Positive Regulator of Endothelial Integrity". en. In: *Scientific Reports* 7.1 (Aug. 2017), p. 10132. ISSN: 2045-2322. DOI: 10.1038/s41598-017-10392-0.

[6] Sabra L Klein and Katie L Flanagan. "Sex differences in immune responses". In: *Nature Reviews Immunology* 16.10 (2016), pp. 626–638.