HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2022 April 21.

Published in final edited form as:

Nat Biotechnol. 2022 March ; 40(3): 355–363. doi:10.1038/s41587-021-01066-4.

Covarying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics

Yakir Reshef[1,2,3,4,5,*], **Laurie Rumker**[1,2,3,4,5,*], **Joyce B. Kang**[1,2,3,4,5], **Aparna Nathan**[1,2,3,4,5], **Ilya Korsunsky**[1,2,3,4,5], **Samira Asgari**[1,2,3,4,5], **Megan B. Murray**[6], **D. Branch Moody**[3], **Soumya Raychaudhuri**[1,2,3,4,5,7,**]

[1.] Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA

[2.] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[3.] Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[4.] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[5.] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[6.] Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, USA

[7.] Versus Arthritis Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK

Abstract

As single-cell datasets grow in sample size, there is a critical need to characterize cell states that vary across samples and associate with sample attributes like clinical phenotypes. Current statistical approaches typically map cells to clusters then assess differences in cluster abundance. We present covarying neighborhood analysis (CNA), an unbiased method to identify associated cell populations with greater flexibility than cluster-based approaches. CNA characterizes dominant axes of variation across samples by identifying groups of small regions in transcriptional space—termed neighborhoods—that covary in abundance across samples, suggesting shared

** Correspondence to: Soumya Raychaudhuri, 77 Avenue Louis Pasteur, Harvard New Research Building, Suite 250, Boston, MA 02446, USA. soumya@broadinstitute.org; 617-525-4484 (tel); 617-525-4488 (fax).

*These authors jointly led this work.

Author Contributions:

YR, LR, and SR designed and conceptualized the study. YR and LR designed and implemented the algorithm, and performed simulations. YR, LR, and JK performed analysis of real data.

AN, SA, and IK provided input on methodologic design and real data analysis. DBM, MM, AN, and IK offered dataset-specific expertise. YR, LR, and SR composed the manuscript with input from the remaining authors.

Competing interests:

Soumya Raychaudhuri serves as a consultant for Gilead, Pfizer, Janssen, and Rheos Medicines and is a founder for Mestag, Inc. Ilya Korsunsky serves as a consultant for Mestag Inc.

function or regulation. CNA performs statistical testing for associations between any sample-level attribute and the abundances of these covarying neighborhood groups. Simulations show that CNA enables more sensitive and accurate identification of disease-associated cell states than a cluster-based approach. When applied to published datasets, CNA captures a Notch activation signature in rheumatoid arthritis, identifies monocyte populations expanded in sepsis, and identifies a novel T-cell population associated with progression to active tuberculosis.

Introduction

High-dimensional profiling of single cells is a central tool for understanding complex biological systems[1]. Cells gathered from distinct samples are used to characterize cell states that associate with a sample attribute like a clinical phenotype or experimental perturbation. Current methods for analyzing multi-sample single-cell datasets typically impose a global transcriptional structure on the dataset by partitioning cells into groups through clustering[2, 3]. The data are then analyzed solely through this lens by asking whether a sample attribute is associated with expansion or depletion of any clusters. Such approaches assume the underlying biology is well captured by the imposed structure and often require substantial tuning of parameters such as clustering resolution[4].

Here we present covarying neighborhood analysis (CNA), a method for characterizing dominant axes of inter-sample variability and conducting association testing in single-cell datasets without requiring a pre-specified transcriptional structure. The core notion of CNA is the value of granular analysis of neighborhoods—very small regions in transcriptional space—with aggregation of neighborhoods according to their covariance across samples. We posit that groups of neighborhoods that change in abundance together across samples are likely to represent biologically meaningful units that share function, regulatory influences, or both. CNA can be used to define these covarying neighborhood groups and then identify statistical associations between them and any sample-level attribute. One published method, MELD, has already demonstrated the potential of neighborhood-scale abundance information in datasets with small sample size [5]; however, this method does not provide a framework for determining statistical significance in order to differentiate true from false discoveries. As we show, the large number of neighborhoods in many single-cell datasets makes well-powered association testing at this granularity a challenge. CNA addresses this challenge by leveraging the extensive covariance structure that we show exists across neighborhoods. As a result, CNA offers both a data-dependent, parsimonious representation of single-cell data and well-powered and accurate association testing.

By testing simulated sample attributes in real single-cell data, we demonstrate that CNA is well calibrated and, compared to cluster-based analysis, detects diverse signals with improved power and ability to correctly recover the cell populations driving those signals. We then apply CNA to three published datasets[6–8], demonstrating that it both refines and expands upon the associations previously found using standard approaches.

Results

Overview of Methods

Covarying neighborhood analysis (CNA) relies on a representation of each sample in a single-cell dataset by its abundance of cells across neighborhoods. To construct this representation we begin with a cell-cell similarity graph that captures all cells from all samples. This graph can be created from any representation chosen by the user, such as gene-expression principal components or canonical covariates for a multimodal dataset, typically processed with a batch correction tool [9–12]. We then define one neighborhood per cell m in the dataset: every other cell m' belongs to the neighborhood anchored at cell m according to the probability that a random walk in the graph from m' will arrive at m after s steps (Methods; Figure 1A). CNA chooses s in a data-dependent manner to minimize neighborhood size while ensuring that neighborhoods are not dominated by cells from only a few samples. Supplementary Figure 1 and Supplementary Table 1 show example neighborhoods and average neighborhood sizes for the real datasets analyzed in this paper.

We aggregate this information into a *neighborhood abundance matrix* (NAM) whose n,m-th entry is the relative abundance of cells from sample n in neighborhood m (Figure 1B–C). We then apply principal components analysis to the NAM to define neighborhood groups whose abundances change in concert across samples (Figure 1D). For each NAM principal component (NAM-PC), the neighborhoods with positive loadings tend to have high abundance together in the samples for which the neighborhoods with negative loadings have low abundance. Likewise, the sample loadings for each NAM-PC yield information about the extent to which that NAM-PC's pattern of covarying neighborhoods appears in each sample.

NAM-PCs can be used to characterize transcriptional changes that comprise the axes of greatest variation in neighborhood abundances across samples. They can also be used to test for associations between these transcriptional changes and a per-sample attribute of interest, *e.g.*, a clinical attribute, genotype, or experimental condition. To perform this test, we model the attribute value for each sample as a linear function of the sample's loadings on the first k NAM-PCs, where k is chosen in a data-dependent manner to optimize model performance without overfitting (Methods). We report a p-value for this association by permuting attribute values within experimental batches to obtain a null distribution.

Finally, we define the specific cell populations driving any detected associations. We do so by using the neighborhood loadings on the first k NAM-PCs and the estimated per-PC effect sizes from our linear model to estimate per-neighborhood correlations between neighborhood abundance—as captured by the first k NAM-PCs—and the sample attribute (Methods). We report false discovery rates (FDRs) for each per-neighborhood association by again permuting attribute values within experimental batches to obtain null distributions. We refer to the abundance correlation between the attribute and the neighborhood anchored at each cell as the *neighborhood coefficient* of that cell. We control for sample-level confounders, such as demographic variables, technical parameters and batch effects, by linearly projecting them out of the NAM and the attribute prior to association testing (Methods). We have released open-source software implementing the method (**URLs**).

CNA requires no parameter tuning, and it has favorable runtime properties: given a nearest-neighbor graph, computing the NAM and conducting permutation-based association testing takes <1 minute (and 579MB memory) for a dataset of >500,000 cells and >250 samples.

Performance assessment with simulations

We used real single-cell data with simulated per-sample attributes (Supplementary Figure 2) to assess CNA's calibration (type I error) and to compare CNA's statistical power (type II error) against cluster-based analysis. This published dataset of 259 patients previously infected with *Mycobacterium tuberculosis* contains 500,089 memory T cells in a canonical correlation analysis (CCA)-based per-cell joint representation of whole-transcriptome mRNA and abundances of 31 surface proteins[6]. In addition to assessing calibration and power, we also assessed CNA's ability to recover the precise cell populations underlying an association by computing the correlation between the per-cell ground-truth values used to create the simulated attribute and effect sizes estimated by the method; we refer to this quantity as "signal recovery" (Methods; Supplementary Figure 3).

To assess type I error, we simulated sample attributes without true associations to the data and found CNA was well-calibrated. We first permuted patient age across all samples and observed a $p<0.05$ global association in 41/1000 trials (type I error rate at $\alpha=0.05$ of 0.041 ± 0.013; Supplementary Figure 4). We next permuted patient ages within experimental batches and observed $p<0.05$ for 44/1000 trials (type I error 0.044 ± 0.013; Supplementary Figure 4). Finally, to simulate extreme batch effects, we selected batches at random and for each randomly selected batch assigned case status to the samples in the selected batch and control status to all other samples. We observed $p<0.05$ for 60/1000 trials (type I error 0.060 ± 0.015; Supplementary Figure 4).

To assess CNA's power and signal recovery, we simulated sample attributes with true associations to different types of cell populations and compared CNA's performance to that of a cluster-based association test using Mixed-effects modeling of Associations of Single Cells (MASC)[13]; MASC offers greater power than a t-test or linear model by accounting for per-cell information[14]. For CNA, power was defined as the proportion of simulations with global $p<0.05$. For MASC, power was defined as the proportion of simulations for which at least one cluster achieved $p<0.05/[\text{total clusters}]$. Cluster-based analysis is sensitive to the choice of parameters such as the resolution parameter, and users typically explore a range of resolutions before selecting one[4]. To reflect this, we ran MASC using four different clustering resolutions. We aggregated power results across these resolutions by taking the minimum p-value and correcting for the four resolutions tested. We aggregated signal recovery results by taking the average signal recovery across the tested resolutions (Methods).

We simulated three signal types, each at a variety of noise levels: 1) cluster abundance, where the attribute is a sample's abundance of cells from a given cluster (matching the cluster-based analysis model; Figure 2A); 2) global gene expression program (GEP), where the attribute is a sample's average use of a GEP across all cells (Figure 2B); and 3) cluster-specific GEP, where the attribute is a sample's average use of a GEP across cells in one cluster (Figure 2C). We used principal components computed from the matrix of

cells-by-canonical variables for the whole dataset or for cells within a cluster as our global and cluster-specific GEPs, respectively.

CNA had superior power over cluster-based analysis to detect global GEP and cluster-specific GEP signals, while retaining comparable power for cluster abundance signals (Figure 2A–C). These conclusions also hold with respect to the best-performing individual clustering resolution: for the global GEP and cluster-specific GEP signals CNA had better power than cluster-based analysis at the best-performing resolution, and for cluster abundance signals CNA had comparable power to cluster-based analysis run on the best-performing clustering resolutions, including the ground-truth resolution used to define the clusters (Supplementary Figure 5).

CNA also had superior signal recovery relative to cluster-based analysis for all three signal types (Figure 2A–C). Moreover, for global GEP signals, CNA's signal recovery was superior to signal recovery for cluster-based analysis even at the best-performing clustering resolution (Supplementary Figure 5). For cluster abundance signals, the only resolution parameter choice that obtained superior signal recovery to CNA was the one used to create the simulated cluster signals. For cluster-specific GEPs, the only resolution outperforming CNA was the finest resolution tested, which included 72 clusters; all other resolutions were less accurate, and two had signal recovery near zero (Supplementary Figure 5). In downsampled versions of the TBRU data at lower sample sizes (18 batches/*N=107*, 12 batches/*N=71*, and 8 batches/*N=48*), CNA generally continued to outperform the cluster-based comparator although the latter gained an advantage for the causal cluster signal type at lower sample sizes (Supplementary Figure 6). In a second smaller dataset of patients with and without sepsis (*N=65*), as well as in downsampled versions of this smaller dataset (*N=40*, *N=20*), CNA outperformed the cluster-based comparator method across all three signal types (Supplementary Figure 7).

CNA captures Notch activation gradient implicated in rheumatoid arthritis

To assess whether CNA can detect important biological structure in real data, we applied CNA to 27,216 fibroblast scRNA-seq profiles from synovial joint tissue of six rheumatoid arthritis (RA) patients and six patients with osteoarthritis[8]. The original publication, also by our group, used trajectory analysis to uncover a fibroblast trajectory corresponding to endothelial Notch signaling and found expansion of Notch-activated fibroblasts in RA. This prior study also identified two fibroblast clusters—representing the lining versus sublining synovium regions—and demonstrated sublining fibroblast expansion in RA.

CNA identifies NAM-PC1 as the dominant signal in this dataset: NAM-PC 1 explains 39% of the variance in the NAM while no other NAM-PC explains more than 12%. NAM-PC1 reflects Notch activation: cells' expression of *PRG4*—an established Notch-response gene in the synovial joint tissue[15]—was most strongly correlated with their anchored neighborhoods' NAM-PC1 loadings (Pearson r=0.79, p<1e-10), followed by expression of *FN1* (Pearson r=0.71, p<1e-10), a signaling molecule shown to regulate Notch[16]. Further, two Notch gene sets were significantly enriched among all gene correlations to NAM-PC1 ("Vilimas NOTCH1 targets up" and "Reactome signalling by NOTCH", FDR=0.0073 and FDR=0.019, respectively). Moreover, NAM-PC1 has a stronger correlation than the

published trajectory to the experimentally defined Notch activation score from the original paper (Spearman r=0.56 vs r=0.43, p<0.01 by bootstrapped permutation test, Figure 3A–C). CNA's focus on inter-sample abundance covariance information was useful for uncovering this structure: PC1 from naive transcriptional PCA of the cells-by-genes expression matrix has a low correlation (Spearman r=0.22) with Notch activation (Figure 3D). Notably, NAM-PC1 detected the Notch activation signal without the parameter tuning required by trajectory analysis.

NAM-PC1 largely separates the sublining and lining clusters (t-test p<1e-10) because sublining cells generally have higher Notch activation[8], but CNA shows that Notch activation variation exists within these clusters. Neighborhood loadings on NAM-PC1 are correlated to the Notch activation scores of their anchor cells even within each cluster (Pearson r=0.36 lining cluster with p<0.001, Pearson r=0.33 sublining cluster with p<0.001; Figure 3E–F). NAM-PC2 appears to reflect an axis of fibroblast activation in response to interferon [17] (Supplementary Table 2). Low sample size precludes detailed interpretation of further NAM-PCs in this dataset.

CNA identified RA-associated cell populations (global p=0.02) that recapitulate the coarse cluster-based associations but more precisely reflect the driving Notch mechanism. Nearly all cells to which CNA assigned significantly positive neighborhood coefficients (99.9% of 5,181 total cells at FDR<0.05) belong to the sublining cluster, and nearly all cells assigned significantly negative neighborhood coefficients belong to the lining cluster (96.8% of 7,169 total cells at FDR<0.05). However, CNA assigned some sublining-cluster cells to the depleted population, and these cells have lower Notch activation gene expression than other sublining-cluster cells. Likewise, CNA assigned some lining-cluster cells with higher Notch activation to the expanded population (Figure 3E–F). Therefore, CNA adds informative granularity beyond the cluster-based associations.

CNA refines sepsis-associated blood cell populations

To assess CNA's ability to identify granular case-control associations in a dataset with many cell types, we next applied CNA to scRNA-seq profiles of 102,814 peripheral blood mononuclear cells (PBMCs) from 29 patient with sepsis and 36 patients without sepsis. The published analysis[7] compared patients with and without sepsis in several sub-cohorts, *e.g.*, among intensive care patients and among emergency department patients. Using a clustering of the data (Figure 4A), this analysis identified expansion of a monocyte state "MS1" in sepsis in multiple sub-cohorts (Supplementary Table 3). Our re-analysis compares patients with and without sepsis across the full cohort using CNA and a MASC cluster-based analysis.

CNA found significant changes in sepsis compared to control samples (global p=7e-5) and identified a population expanded in sepsis (19,991 monocytes at FDR<0.05; Figure 4B). This population overlapped with MS1 but contained cells from other clusters: 56% of cells in CNA's expanded population were in MS1 while 44% were in clusters MS2, MS3, and MS4. CNA's expanded population contained 75% of all MS1 cells. In contrast, our cluster-based analysis of the same sepsis phenotype found that no cluster was significantly associated, though MS1 did have the smallest p-value (p=0.26; Figure 4C). Therefore, our

results support the original finding but demonstrate that the published clusters partition transcriptional space in a manner that reduces power to detect the sepsis association in the full cohort.

CNA's cluster-free delineation of sepsis-associated cell states implicates known sepsis-relevant pathways. Gene expression correlations to per-cell neighborhood coefficients were most highly enriched for the RAC1 activation gene set (FDR=2.5e-4, r=0.63 between summed gene set expression and neighborhood coefficients), a known sepsis-associated pathway[18] whose suppression has therapeutic benefit in septic encephalopathy[19] (Figure 4D). The other most significantly enriched gene sets also have established sepsis associations (Supplementary Table 4).

Strikingly, CNA identified considerable within-cluster heterogeneity in this dataset: eight of the fifteen published clusters included clear subpopulations with distinct degrees—and even directions—of associations to sepsis (Figure 4E–H; Supplementary Figure 8). For example, MS4 contains both a significantly expanded and a significantly depleted subpopulation (FDR<0.05; Figure 4F). Both of these associations were obscured by aggregating these subpopulations together. In the published analysis, clustering resolution was tailored to each cell type (e.g. Leiden 0.6 for T cells, 0.4 for monocytes). In contrast, CNA does not require parameter tuning to detect associated populations.

To explore local contrasts in gene expression between cluster sub-populations implicated in sepsis by CNA and closely-related but non-sepsis-associated cells, we conducted differential expression contrasting these sub-populations both to their respective clusters and to their respective major cell types. We found that many of the gene expression programs detected through global analysis (*e.g.*, RAC1) also distinguish each cluster's depleted sub-population from similar but non-depleted cells. However, as shown in Supplementary Figure 9, this analysis also revealed gene expression programs that uniquely typify sepsis-associated populations from specific clusters (see Methods, Supplementary Table 5, and Supplementary Table 6). For example, the depleted sub-populations of BS1 and MS4 are negatively enriched in Class II Histone Deacetylase Complex (HDAC) activity, consistent with literature showing that HDAC activity increases in sepsis and also that inhibition of a class II HDAC therapeutically increases B cell and monocyte populations in patients with sepsis [20]. Further, IL-12 signaling gene sets were negatively enriched in the depleted sub-populations of DS1, DS2, MS4, and TS2, matching a known elevation of serum IL-12 among septic patients [21,22]. Finally, telomerase pathway genes are negatively enriched in the depleted sub-population of BS1; telomere lengths are known to be shortened in patients with sepsis [23].

For comparison, we also ran MELD on this dataset. MELD per-cell abundance relationship scores to sepsis were correlated with CNA's neighborhood coefficient values (r=0.6, Supplementary Figure 10). In contrast to CNA, however, MELD does not assess significance for these scores and produced patterns of scores on randomly permuted case-control labels that also appear to have nontrivial structure (Supplementary Figure 10). When we applied a permutation-based approach identical to the one used by CNA to assess significance at the neighborhood level, none of the individual per-cell MELD scores were significant at

FDR<0.05 (Supplementary Figure 10). This highlights the power advantage of CNA's use of inter-sample covariance information.

To understand the dominant axes of inter-sample variation in this dataset, we examined gene expression and cell-type abundance correlates of each of the first five NAM-PCs. We found that some NAM-PCs do indeed capture information related to broad cell-type populations (Supplementary Figure 11, Supplementary Table 7, and Supplementary Table 8). For example, NAM-PC 1 has statistically significant positive correlation with NK cell abundance (Pearson r=0.55, p=1.7e-6) while NAM-PC2 has statistically significant positive correlation with B-cell abundance (Pearson r=0.53, p=5.1e-6) and T-cell abundance (Pearson r=0.84, p=4.0e-18) and negative correlation with macrophage abundance (Pearson r=−0.89, p=1.1e-22). In contrast, other NAM-PCs, such as NAM-PC3, do not exhibit statistically significant correlation with broad cell type abundance, suggesting they capture finer-scale structure. At the level of pathways, we found several immune activation and inflammation-related gene sets enriched in these NAM-PCs that largely recapitulated the gene sets we identified as relating to sepsis (Supplementary Figure 11 and Supplementary Table 9). These included some gene sets, such as RAC1 signaling, that were enriched in multiple NAM-PCs and others, such as IL12 signaling mediated by STAT4, that were enriched only in one NAM-PC.

CNA captures diverse associations in tuberculosis dataset

We next applied CNA to a larger and more richly phenotyped dataset: 500,089 memory T cells from 259 patients in a tuberculosis progression cohort[6]. (This dataset, recently published by our group, was also used above for simulations.) The published analysis employed 31 clusters to compare patients previously infected with *Mycobacterium tuberculosis* who rapidly developed symptoms ('progressors', N=128) to those who sustained latent infections ('non-progressors', N=131).

The NAM-PCs in this dataset appear to carry biologic meaning. For example, neighborhood loadings on NAM-PC1 correlate strongly across cells with a previously-defined transcriptional signature of "innateness," the degree of effector function in each cell (r=0.81; Figure 5A)[24, 25], and individual gene correlations to NAM-PC1 also reflect this (Supplementary Table 10). This result shows that individuals vary to a substantial degree in their average T cell "innateness." Moreover, NAM-PC1 sample loadings were nearly identical when computed using protein profiling, mRNA profiling, or the joint CCA representation of this data (Figure 5B); by contrast, PC1 sample loadings from naive PCA of each data type were far less correlated (Figure 5B). Across the three modalities, 50% of total variance in each NAM was explained by the top 5–10 PCs (out of 271; Figure 5C), suggesting NAM-PCs offer a parsimonious representation of this dataset.

To assess whether NAM-PCs can detect nuanced transcriptional shifts that span a broad range of cell types, we re-computed NAM-PCs for this dataset without the upstream batch correction from the published analysis, which has the potential to eliminate subtle biologic variation[26, 27]. Indeed, in the mRNA data we found that NAM-PC2 and -PC4 correlate strongly with sex (joint R^2=0.76; Figure 5D). Neighborhood loadings on NAM-PC4 indeed capture sex chromosome gene expression (Supplementary Table 11)—which differentiates

otherwise very similar cells from individuals with different sex chromosomes across all cell types—while NAM-PC2 captures cell states known to vary in abundance with sex[28] (Figure 5E, Supplementary Figure 12, Supplementary Table 12, and Supplementary Table 13). While sex information is also encoded in naive gene expression PCs, it is captured primarily in later PCs: the total predictive power of the first four naïve gene expression PCs for sex was R^2=0.05 as compared with R^2=0.76 for the first four NAM-PCs. When we instead expanded to the first 20 naïve gene expression PCs, we found that the PC most strongly correlated with sex was PC-18 (R=0.84; Supplementary Table 14). Thus, NAM-PCs better prioritize—*i.e.*, capture in earlier PCs—expression variation that is relevant to inter-sample differences (*e.g.*, sex) rather than intra-sample differences (*e.g.*, cell cycle) that may not differ strongly across samples.

We next analyzed the primary phenotype, TB progression, using identical data processing and covariate control to the published analysis (Methods), which defined 31 clusters (Figure 6A) and found two clusters with Th17- and innate-like character (“C-12” and “C-20”, respectively) to be depleted among progressors. CNA found a significant global association (CNA global p=0.0015) driven by a depleted population (FDR<0.05) as well as an expanded population (FDR<0.05). CNA’s depleted population overlapped with the previously published C-12 (86% of cluster) and C-20 (64% of cluster) but also contained many cells from additional, phenotypically-similar clusters (74% of depleted population; Figure 6B). Overall, this population had similar characteristic proteins and genes to the cluster-based depleted population (Figure 6C, Supplementary Table 15) but contained substantially more cells.

In contrast to the cluster-based analysis, CNA identified a population of cytotoxic cells expanded among progressors (Figure 6B–C; Supplementary Table 15), consistent with prior work describing the interplay between cytotoxic cells and mycobacteria[29]. These cells were predominantly captured by two clusters: 72% were from cluster C-23 (“CD4+ cytotoxic”) and 27% were from C-22 (“CD4+ CD161+ cytotoxic”). Tested individually, these clusters show weak evidence of association with progressor status (p=0.013 and 0.022, respectively) and do not pass multiple testing correction for the 31 clusters total. With a single test, CNA detected an associated population of functionally similar cells that had been split across multiple clusters.

In the original publication, the association to progressor status was only significant after unbiased mRNA profiling was combined with targeted surface protein quantification in a multimodal representation. Given our observed correlations among NAM-PCs across data modalities, we speculated that CNA might identify this association in unbiased mRNA data alone, and indeed it does (global p=4.5e-3).

Finally, we conducted a survey for associations between the single-cell data (multimodal representation) and 17 sample-level attributes besides progressor status (Methods). With control for confounders and multiple testing (Methods; Supplementary Table 16), we found global associations for age (p<1e-6; Figure 6D–E), season of blood draw (p<1e-6; Figure 6F), genetic ancestry (p=1.8e-4; Figure 6G), and sex (p=3e-6) (Supplementary Figure 13, Supplementary Table 17). These results align with the published cluster-based analysis,

which also found evidence of these associations, and demonstrate that CNA can detect associations to a variety of signals, including demographic, environmental, and genetic factors. On average, CNA chose 19 (out of 271; 7%) NAM-PCs to explain 26% of variance in these attributes, a 3.7x enrichment, suggesting that NAM-PCs are a parsimonious, phenotype-relevant representation of this complex dataset. The cell populations associated by CNA are juxtaposed with those found by a conventional cluster-based approach with identical covariate control in Supplementary Figure 13.

Some of the distinguishing cell states CNA finds associated with these attributes are established in the literature, while others are less well elucidated. We find older age is associated with higher CD8+/CD4+ ratio, decreased costimulatory molecule expression, and greater effector memory character relative to central memory character [30] (Figure 6E; Supplementary Table 18). CNA highlights a shift toward more Th2 character relative to Th1 character during the winter season in contrast to studies of seasonality in other locations [31], and identifies an expanded CD8+ central memory population and depleted CD4+ cytotoxic population with increasing European genetic ancestry (Supplementary Table 19 and Supplementary Table 20). Cluster-based analyses with identical covariate control produced generally similar results, but implicated fewer cells in each association than CNA (Supplementary Figure 13, Supplementary Table 17).

Discussion

In this work we introduced CNA, a method to characterize dominant axes of abundance variation across samples in a single-cell dataset and to identify with greater flexibility and granularity cell populations whose abundance correlates with sample attributes of interest. CNA offers improved power and signal recovery over traditional cluster-based analysis while remaining robust to experimental artifacts and providing control for sample-level confounders, and it does so without requiring parameter tuning or long computation times. CNA can be used to study diverse sample attributes, enabling improved understanding of disease pathology, risk, and treatment.

In addition to their utility for testing for associations to sample-level attributes, NAM-PCs themselves appear to carry biological meaning: for example, our analyses revealed NAM-PCs that correspond to Notch signaling, memory T cell innateness, and a sex chromosome gene signature. For NAM-PCs without clear biological interpretation, characterizing the cellular functions and/or regulatory influences that unify these covarying neighborhood groups could yield insight into basic biology. Covarying neighborhood groups may, for example, delineate cell states most relevant to context-dependent cellular processes such as gene regulation and cellular metabolism.

CNA offers a versatile framework that can be easily extended to other data modalities. We highlight datasets of scRNA-seq and multimodal mRNA-and-protein profiling, but CNA can be extended to any modality for which cell-cell graphs can be built such as single-cell ATAC-seq epigenome profiling or mass cytometry protein profiling. For some of these applications, NAM factorization with approaches besides PCA, such as non-negative matrix factorization or independent components analysis, may be useful. For example,

decomposition methods that do not enforce orthogonality among components could result in components with clearer correspondence to individual cellular programs. Such exploration, however, is beyond the scope of this work.

CNA has several limitations. First, because its emphasis is explicitly on inter-sample variation, CNA's power and signal recovery do degrade with sample size as demonstrated in our simulations. Second, due to its signal type-agnostic methodology, CNA may also be less powerful than more constrained models at lower sample sizes specifically when the underlying biology matches those models. Third, though existing approaches for biological annotation of clusters and trajectories can be applied to CNA populations and NAM-PCs, respectively, such approaches typically seek a single explanatory signal; an associated population or NAM-PC might capture multiple related processes, and a given biological process may be captured by multiple NAM-PCs. Fourth, while cell assignments into clusters are discrete, CNA's neighborhoods have probabilistic distributions in transcriptional space. As a result, it is not always obvious where the boundary of a CNA-associated population lies, or whether such a boundary exists.

Despite these limitations, CNA is a sensitive way to identify disease states and drivers of variation across samples in single-cell datasets that is unique in taking advantage of inter-sample variation. As single-cell datasets grow in sample volume, methods that use and characterize inter-sample information at fine-scale transcriptional resolution will become crucial to realize the promise of single-cell technologies.

Methods

Covarying Neighborhood Analysis

Intuition—Covarying neighborhood analysis is built on the idea of a *transcriptional neighborhood*, a very small subset of transcriptional space, typically much smaller than would arise from traditional clustering. Our method is based on two intuitions about transcriptional neighborhoods. First, because of the neighborhoods' granularity, any meaningful variation across samples will result in differential abundance of one or more neighborhoods across samples. Second, neighborhoods covary in abundance across samples because of shared function and/or regulatory influences. The first intuition leads us to represent multi-sample single-cell data using the *neighborhood abundance matrix* (NAM), a samples-by-neighborhoods matrix that describes the relative abundance of each neighborhood in each sample. The second intuition leads us to analyze the NAM using principal components analysis. The resulting principal components reveal sets of neighborhoods that covary in abundance across samples as well as samples with similar abundance profiles across neighborhoods. We use this information to find structure and conduct association testing with sample-level attributes like clinical information, genotypic information, or experimental conditions. The remainder of our technical material establishes notation and assumptions, then provides detailed descriptions of (i) definition of transcriptional neighborhoods, (ii) construction and quality control of the NAM, (iii) PCA of the NAM controlling for batch and covariates, and (iv) association testing.

Notation and assumptions—Let X be an $M \times G$ matrix representing a single-cell dataset with M cells and G cell-level features such as genes. Let N be the number of distinct samples to which the cells belong, and for every cell m and every sample n let $C(n)$ be the set of cells belonging to the n-th sample. We assume that X has already undergone quality control and (if desired) batch correction and that a nearest neighbor graph construction algorithm—such as the UMAP nearest neighbor algorithm—has been run on X to produce a sparse, weighted $M \times M$ adjacency matrix A whose m,m'-th entry indicates the similarity between cells m and m' in the graph.

Definition of transcriptional neighborhoods—For each cell in our dataset, we define a transcriptional neighborhood anchored at that cell using the sense of locality provided by the nearest neighbor graph. That is, two cells are considered close to each other if it is "easy" to reach one from the other in the graph. A natural way to define neighborhoods through this lens is to stipulate that two cells are in the same neighborhood to the extent that a random walk on the graph would be likely to reach one from the other.

More formally, we define a random walk whose transition probabilities are proportional to the entries of $I+A$, where I is the $M \times M$ identity matrix. (The addition of the identity adds self-loops to the UMAP graph.) That is, the probability that the walk moves from cell m' to cell m in one step is given by

$$\widetilde{A}_{m',m} := \frac{(I+A)_{m',m}}{1+\Sigma_{m''}A_{m',m''}}.$$

For some number of steps s, we then define the extent to which the m'-th cell belongs in the neighborhood of the m-th cell as the probability that a random walk starting at the m'-th cell will end up at the m-th cell after s steps. This is given by

$$P^s{}_{m' \to m} := \left(e^{m'}\right)^T \widetilde{A}^s e^m$$

where $\widetilde{A}$ is the matrix whose entries are given by $\widetilde{A}_{m',m}$ and e_m is a length-M vector whose m-th entry equals one and whose other entries are all zero, and $e^{m'}$ is similarly defined. As we discuss in detail below, the number of steps s is chosen to minimize neighborhood size while ensuring that neighborhoods are not dominated by a small number of samples.

Construction and quality control of the neighborhood abundance matrix—With neighborhoods defined, we transform our dataset into a matrix of samples by neighborhoods whose n,m-th entry is the relative abundance of neighborhood m in sample n, i.e., the NAM. To formally define the NAM, we first let

$$R_{n,m} := \sum_{m' \in C(n)} P^s{}_{m' \to m}$$

be the length-M vector representing the total number of expected cells from the n-th sample that would arrive in the m-th neighborhood from our random walk. The NAM is given by normalizing the rows of R to sum to one, i.e.,

$$Q_{n,m} = \frac{R_{n,m}}{\Sigma_m R_{n,m}}$$

These entries can be computed very quickly using iterative sparse matrix multiplication, taking under one minute for a dataset with 500K cells and over 250 samples.

Choosing the length of the random walk: When selecting the number of steps s in the random walk that defines the NAM, our guiding principle is that s should be chosen in a data-dependent manner to minimize neighborhood size, thereby retaining informative granularity, while ensuring neighborhoods are not dominated by cells from a few samples. We quantify this by measuring, for each neighborhood, the kurtosis of its respective column of the NAM: a large kurtosis indicates a small number of samples dominates the relevant neighborhood. With increasing timesteps, as neighborhoods expand to incorporate more cells, kurtosis decreases. To achieve an appropriate balance, we allow our random walk to continue until either the median kurtosis across neighborhoods is less than 8 (the kurtosis of a uniform distribution over only 10% of samples) or the median kurtosis across neighborhoods decreases by less than 3 (the kurtosis of the normal distribution) over consecutive time steps.

Removing neighborhoods with strong batch effects: If batch information is available, we remove neighborhoods dominated by one or a few batches by averaging the rows of the NAM within each batch to produce a batches-by-neighborhoods matrix and then computing for each neighborhood the kurtosis of its respective column of this new matrix. We discard all neighborhoods with kurtosis greater than twice the median value across all neighborhoods. Because this removal of individual neighborhoods may not eliminate subtle batch effects spread across many neighborhoods, we also control for batch as a covariate in our linear model-based framework, as described below.

Conditioning on sample-level covariates (including batch information if available): If there are sample-level covariates whose influence on X we do not wish to be represented among the principal components of the NAM, we linearly project them out of each column of the NAM, i.e., we regress each column of the NAM on the sample-level covariates that are supplied and replace it with the residuals arising from that regression. If there is batch information available, this can also be done with a one-hot encoding of batch IDs to further remove subtle batch effects. In this case, we use ridge regression with an automatically chosen ridge parameter to account for the typically large number of batches relative to samples.

PCA of the NAM while conditioning on batch and covariates—Once the NAM is constructed, principal components analysis yields the decomposition

$$\overline{Q} = U D V^T$$

where $\overline{Q}$ is the NAM with columns standardized to have mean zero and variance one; U is a matrix whose i-th column contains the i-th left singular vector, which has one entry per sample; D is the diagonal matrix of singular values; and V is a matrix whose i-th column contains the i-th right singular vector, which has one entry per neighborhood. Each of the right singular vectors identifies neighborhoods that covary in abundance across samples. Each of the left singular vectors identifies samples that have similar abundance profiles across neighborhoods.

Association testing—CNA quantifies association of the NAM to a given sample-level attribute in two ways: i) a global quantification of the fraction of variance in the attribute explained by the single-cell data, with an associated p-value, and ii) a local estimate of the correlation between the attribute and each neighborhood's abundance across samples, with associated false discovery rates.

Global association test: Let y be a length-N vector containing a sample-level attribute of interest such as clinical information, genotypic information, or experimental condition, and suppose we want to associate y with the inter-sample variation in X. Because the left singular vectors of the NAM, i.e., the columns of U, each contain one number per sample, we can do this in a simple linear model in which each sample is an observation. That is, for some number k of principal components, we can fit the model

$$y = U^k \beta^k + \epsilon$$

where U^k denotes the first k columns of U, β^k is a length-k vector with one coefficient per principal component, and ϵ represents mean-zero noise.

To choose k in a flexible and automatic way, we fit the above model for four different values of k ranging from $[N/50]$ through $min([N/50], N/5)$ where $[...]$ denotes the ceiling function. For each value of k, we compute a multivariate F-test p-value for the null hypothesis $H_0: \beta^k = 0$, and we choose the value k^* that yields the minimal p-value. Thus larger values of k are only selected if they provide increased predictive power for y beyond what we would expect simply from their providing more degrees of freedom to the model.

If covariates were residualized out of the NAM, these are residualized out of y prior to fitting the model. Similarly, if batch information was residualized out of the NAM, it is likewise residualized out of y prior to fitting the model using the same ridge parameter.

To obtain a p-value for global association, we perform the above procedure, including selection of k, on a large number of empirical null instantiations (1,000 by default) obtained by permuting the values of y within each batch of the dataset. We then use the resulting set of p-values, of which there is one per null instantiation, as our null distribution.

Local association test: A natural notion of neighborhood-level effect size would be the correlation between the m-th column of the NAM and y. However, the NAM is very high-dimensional and so these correlations are noisy. Instead, we therefore compute a "smoothed correlation" between neighborhood m and y, i.e., the correlation between the m-th column of the rank-k* representation of the NAM and y. Mathematically, this equals

$$\gamma := V^{k^*} D^{k^*} \beta^{k^*}$$

where D_k denotes the upper-left $k \times m$ submatrix of D, V_k denotes the first k columns of V, and β^{k*} is the estimate from the model for global association. We refer to the entries of γ as the *neighborhood coefficients* for the sample attribute y. To assess statistical significance, we again compute null versions of γ and use these to estimate an empirical false discovery rate for a variety of magnitudes of correlation by comparing the number of entries of γ with magnitude above a given threshold to the average number of entries in the null versions of γ with magnitude above that threshold.

Assessing performance with simulations

To assess the calibration and power of our method, we conducted simulations using two real single-cell datasets. Our primary simulations were conducted with data from the tuberculosis research unit (TBRU) cohort[6]. This dataset consists of M=500,089 memory T cells from N=271 samples profiled with CITE-Seq[32], which simultaneously provides both single-cell RNA-seq and single-cell quantification of 31 surface proteins. We also conducted supplementary simulations on data from $M = 102,814$ CD45+ peripheral blood mononuclear cells (PBMCs) from $N = 65$ samples profiled with single-cell RNA-seq. (Further aspects of these datasets are described elsewhere in the Methods.) All simulations in the TBRU dataset used a 20-dimensional canonical correlation analysis-based representation of each cell generated by the original publication reflecting information shared by the RNA-seq and surface protein modalities. This representation was subsequently run through Harmony[9] to remove sample-specific and batch effects. We first describe our simulations for quantifying type 1 error, generation of simulated non-null attributes, and analysis of those non-null attributes in our primary simulations using the TBRU dataset at full sample size ($N = 271$). We then describe how we adapted these procedures for our supplementary simulations using the smaller sepsis dataset ($N = 65$) as well as downsampled versions of the two datasets (18 batches/N=107, 12 batches/N=71, and 8 batches/N=48 for TBRU; N=40 and N=20 for sepsis).

Quantifying type 1 error—For the simulation in Supplementary Figure 4A, we simulated 1,000 independent null attributes by permuting an existing sample attribute in the TBRU dataset (age at sample collection) across all samples in the dataset. For the simulation in Supplementary Figure 4B, we simulated 1,000 independent null attributes by permuting this same sample attribute in the TBRU dataset (age at sample collection) within each batch. This was done to preserve whatever batch effects may be present in the data in our null attributes. For the simulation in Supplementary Figure 4C, we created 1,000 independent null attributes with maximal batch effect by selecting 1,000 batches $\{b_{1,\ldots,} b_{1000}\}$ randomly

with replacement and setting the i-th attribute to equal one for all the samples in batch b_i and zero otherwise.

In all of the above simulations, we used CNA to obtain a p-value for association to the single-cell data for each attribute, accounting for possible batch effects. This yielded 1,000 p-values in each case.

Generation of simulated non-null attributes—For Figure 2, we simulated three signal types in the full sample-size TBRU dataset: cluster abundance (Figure 2A), global gene expression program (Figure 2B), and cluster-specific gene expression program (Figure 2C). In each case, we added gaussian noise to each simulated attribute to achieve signal-to-noise ratios of {0.01,0.1,0.2,...,0.9,1}.

For the cluster abundance signal type (Figure 2A), we clustered the data using the Leiden algorithm[4] with the same resolution (2.0) used by the authors of the original TBRU study[6]. We then removed any clusters lacking at least 10 samples with at least 50 cells each, and we removed any clusters whose abundance had correlation greater than 0.25 to membership in any batch. This reduced the number of clusters from 26 to 24. For each remaining cluster, we computed the abundance of that cluster per sample. For each of our 11 signal-to-noise ratios, we then simulated 10 independent attributes by summing this attribute with the appropriate amount of gaussian noise. This resulted in 24×10=240 attributes per noise level.

For the global gene expression program signal type (Figure 2B), we treated the 20 per-cell harmonized canonical variables as each representing the activity of a gene expression program across cells. For each canonical variable, we computed the average value of that variable across all cells in each sample. For each of our 11 signal-to-noise ratios, we then simulated 10 independent attributes by summing this attribute with the appropriate amount of gaussian noise. This resulted in 20×10=200 attributes per noise level.

For the cluster-specific gene expression program signal type (Figure 2C), we first clustered the cells using the Leiden algorithm with a resolution of 1.0 and filtered the clusters using the same criteria as in the cluster abundance simulation. (We used a coarser clustering resolution here because this signal type is driven by intra-cluster variability rather than inter-cluster variability, so we wanted to use larger clusters.) For each of the 10 largest clusters, we then computed the top 3 principal components of all the harmonized canonical variables among only the cells in that cluster, which we treated as each representing activity of a cluster-specific gene expression program. For each of these principal components, we computed the average value of that component across all the cells in each sample, assigning cells outside the cluster in question a score of zero. For each of our 11 signal-to-noise ratios, we then simulated 10 independent attributes by summing this attribute with the appropriate amount of gaussian noise. This resulted in 10×3×10=300 attributes per noise level.

Analysis of simulated non-null attributes—For each signal type, we analyzed the simulated attributes using i) CNA accounting for possible batch effects, and ii) MASC with the recommended inclusion of sample-level and batch-level random effects. MASC requires

a set of clusters whose abundance it assesses for correlation with the attribute, so we ran 4 different versions of MASC using 4 different sets of clusters. These were created by running Leiden clustering on our data with resolution parameters of 0.2, 1, 2, and 5, resulting in 3, 15, 26, and 72 clusters, respectively.

To estimate power for a given signal type, we computed for each method and for each signal-to-noise ratio the fraction of tests in which the method reported a p-value <0.05. For CNA, we used the global p-value. For MASC, we used the lowest p-value for any individual cluster, multiplied by the number of clusters to achieve multiple testing correction. We quantified uncertainty in our power estimates by computing empirical standard errors for our estimate of this mean. In Figure 2, we aggregated the 4 versions of MASC into one p-value by computing the minimum p-value across all 4 MASC clustering resolutions and Bonferroni correcting for 4 tests. Supplementary Figure 5 shows results at the level of individual MASC clustering resolutions.

To define a notion of signal recovery for each method, we first utilized each method to obtain per-cell estimates of correlation to the attribute as follows. For CNA, we used the neighborhood coefficient for each cell, which is the per-neighborhood correlation to the attribute from the neighborhood for which that cell is the anchor. For MASC at a given clustering resolution, we assigned to each cell the signed effect size beta that MASC estimated for that cell's parent cluster. We then defined ground truth per-cell scores for each signal type such that the noiseless version of each attribute would be obtained by averaging the per-cell scores of all the cells in each sample: for cluster abundance signals, we assigned a score of one to cells in the causal cluster and a score of zero to other cells; for global gene expression program signals, we used the per-cell values of the canonical variable in question; and for cluster-specific gene expression program signals, we used the per-cell values of the principal component in question, with a score of zero assigned to all cells outside the cluster.

To estimate signal recovery for a given signal type, we then computed for each method and for each signal-to-noise ratio the average per-attribute correlation between the method's reported per-cell scores and the ground truth per-cell scores for that attribute. Thus, signal recovery takes values between −1 (worst) and 1 (best). We emphasize that this is distinct from correlating a method's estimated per-sample attribute value with the ground-truth simulated sample attributes. This latter approach would be an assessment of each method's accuracy as a predictor of the sample attributes, whereas our approach assesses something more challenging: the ability of each method to identify the underlying causal cells driving the sample attribute in question. We quantified uncertainty in our estimate of signal recovery by computing empirical standard errors for our estimate of this mean. In Figure 2, we aggregated the 4 versions of MASC into one p-value by averaging their accuracies. Supplementary Figure 5 shows results at the level of individual MASC clustering resolutions.

Simulations at lower sample size—We adapted our simulation framework to the sepsis dataset at full sample size (*N=65*) by Leiden clustering these cells at resolutions of 0.2, 1, 2, and 5 as in our TBRU simulations. This produced four sets of clusters as input to the cluster-

based comparator method MASC. As in our TBRU simulations, the Leiden 2 clustering was used to construct causal cluster and cluster-specific GEP simulated signals. We generated all three simulated signal types as described above, with the following two modifications: we used standard principal components rather than canonical variables to construct global GEP signals, and we included a broader range of noise levels, namely {0.01, 0.1, 0.2, …, 0.9, 0.99} proportion of variance in the final simulated attributes explained by additive noise. We ran CNA and MASC on all simulated signals without batch information or covariates, paralleling the original published analysis of this dataset. This yielded the top two rows of Supplementary Figure 7.

We then created downsampled versions of both the sepsis dataset and the TBRU dataset. For the TBRU dataset, we randomly selected samples for inclusion by batch to obtain three smaller datasets with 18 batches/*N=107*, 12 batches/*N=71*, and 8 batches/*N=48*. For the sepsis dataset, we randomly selected by sample for inclusion to obtain two smaller datasets with *N=40* and *N=20*. For sepsis, we re-computed PCA on the smaller datasets in order to construct new nearest-neighbor graphs. For TBRU, we did not recompute the CCA-based representation of the cells included after downsampling before we constructed new nearest neighbor graphs; therefore, there was some information leakage from the full TBRU dataset to the downsampled datasets. For each downsampled dataset, we re-clustered the cells anew at Leiden {0.2, 1, 2, 5} as input to MASC and again used Leiden 2 clustering to construct causal cluster and cluster-specific GEP simulated signals. We then ran our simulation framework on these smaller datasets to complete Supplementary Figure 6 and Supplementary Figure 7.

Analyses of real data

We analyzed three real datasets: synovial fibroblasts from patients with rheumatoid arthritis versus osteoarthritis (N=12 samples, M=27,216 cells)[8]; peripheral blood mononuclear cells from patients with and without sepsis (N=65 samples, M=102,814 cells)[7]; and memory T cells from patients in a large tuberculosis progression cohort (N=271 samples, M=500,089 cells)[6].

Analysis of rheumatoid arthritis dataset—We obtained the rheumatoid arthritis dataset from the authors of a study of synovium of patients with rheumatoid arthritis (RA) and osteoarthritis (OA). The dataset consisted of M=27,216 synovial cells from N=12 samples profiled with single-cell RNAseq and processed with the Harmony algorithm to mitigate sample-specific effects. The cells also had cluster labels—identifying lining versus sub-lining populations—assigned by the authors of the original study. Mirroring the original study, we filtered this dataset to fibroblasts only. We then applied CNA to this dataset without covariate or batch correction to produce an NAM with principal components as well as a p-value for global association to RA/OA status and neighborhood-level correlations to RA/OA status with corresponding FDRs.

To assess whether NAM-PC1 was related to Notch activation, we obtained the per-cell Notch activation scores defined using the experimentally derived Notch activation gene set from the original study and computed the correlation across cells between these scores and

the NAM-PC1 neighborhood loading assigned to each cell's neighborhood. We computed a p-value for whether this correlation was significantly greater than the corresponding correlation to Notch activation scores of per-cell pseudotime values by bootstrapping over samples to create a null distribution for the difference between the magnitudes of the two correlations.

To assess enrichment of Notch gene sets along NAM-PC1, we computed correlations for the top 5,000 most variable genes in the dataset between expression level in each cell and the cells' anchored neighborhood loadings on NAM-PC1. Using these per-gene correlations as our input ranked list, we computed the enrichment of gene sets containing the term "NOTCH" from MSigDB's "C7" catalogue of immune-related gene sets. We used R's FGSEA package[33] with a maximum gene set size of 500, a minimum size of 15, and 100,000 permutations. This approach to gene set enrichment analysis along an NAM-PC, which we refer to as "NAM-PC gene set enrichment analysis", was also employed to investigate biological processes reflected by NAM-PC2 using gene sets from the Reactome database in MSigDB's "C2" catalogue.

Analysis of sepsis dataset—We downloaded the sepsis dataset of Reyes *et al.* from the Broad Institute Single Cell Portal. The dataset consisted of M=102,814 CD45+ peripheral blood mononuclear cells (PBMCs) from N=65 samples profiled with single-cell RNA-seq. The study consisted of three clinical cohorts: i) patients presenting to the emergency department (ED) with urinary tract infection, divided into patients with leukocytosis but no organ dysfunction (Leuk-UTI, N=10), urosepsis (Int-URO, N=7) and persistent urosepsis (URO, N=10); ii) bacteremic patients with sepsis in hospital wards (Bac-SEP, N=4); and iii) patients admitted to the intensive care unit (ICU) with either sepsis (ICU-SEP, N=8) or no sepsis (ICU-NoSEP, N=7). There were also 19 healthy controls (Control, N=19). In total, the study included 29 patients with sepsis (Int-URO, URO, and Bac-SEP) and 36 patients without sepsis (Control, Leuk-UTI, and ICU-NoSEP).

To maximize comparability between the published analysis and ours, we analyzed the data following the same preprocessing steps as the original authors, namely: we removed cells with <100 unique molecular identifiers (UMIs) and genes with expression in <10 cells, and then we log-normalized the counts and filtered out genes with mean expression <0.0125 or dispersion <0.5. The dataset also includes some samples that are enriched for dendritic cells; following the original analysis, we included these in the dataset but did not assign them phenotype labels so that they could be included in the unsupervised portion of the analysis but would not directly affect any of the association analyses.

The original publication conducted case-control comparisons within 9 different subgroups of sepsis patients and controls (e.g., {URO, Int-URO} vs {Control, Leuk-UTI}). We conducted the same 9 association tests using CNA (without batch information or covariates, following the original study) and found good qualitative agreement; see Supplementary Table 5.

We then ran CNA on the aggregate phenotype of "any sepsis", for which sepsis was defined as {Int-URO, URO, Bac-SEP, ICU-SEP} and non-sepsis was defined as {Control, Leuk-UTI, ICU-NoSEP}. To assess for gene set enrichment in association with sepsis, we

computed correlations for the top 5,000 most variable genes between expression level in each cell and the cells' anchored neighborhood correlations to sepsis. Using these per-gene correlations as our input ranked list, we computed the enrichment of gene sets from the Pathways Interaction Database ("PID") stored in MSigDB's "C2" catalogue. We used R's FGSEA package[33] with parameters as above.

To assess for heterogeneity within the 15 author-defined cell states, we then examined the distribution of CNA-estimated neighborhood correlations to the "any sepsis" phenotype within each published cell-state (see Supplementary Figure 8). We identified MS4, TS2, and BS1 as the most visually striking examples of bimodality of these correlations within individual clusters.

To biologically characterize the cluster sub-populations identified by CNA, we performed differential expression and pathway enrichment analysis as follows. We compared the depleted sub-populations within each cluster found by CNA to have intra-cluster heterogeneity in effect size (TS1, TS2, MS4, DS1, DS2, BS1, BS2) to i) the remaining cells of the same published cluster (*e.g.*, TS1) and ii) the remaining cells of the same major cell type (*e.g.*, T cells). Then, using these per-gene correlations as our input ranked lists, we computed the enrichment of gene sets from the Pathways Interaction Database ("PID") stored in MSigDB's "C2" catalogue and plotted these values in Supplementary Figure 9. We used R's FGSEA package [33] with parameters as above.

We produced interpretations of each of the first five NAM-PCs in this dataset in two ways: first, we computed the correlation across samples between each NAM-PC's sample loadings and the abundances per sample of each of the five main cell populations in the dataset— T cells, B, cells, monocytes, NK cells, and dendritic cells— with corresponding analytic p-values based on a beta-distributed null. We plotted these values as a heatmap in Supplementary Figure 11, retaining only those correlations that achieved nominal significance. Second, we conducted NAM-PC gene set enrichment analysis (see above) for gene sets from the Pathways Interaction Database ("PID") stored in MSigDB's "C2" catalogue and plotted these values in Supplementary Figure 11.

Analysis of tuberculosis dataset—We obtained the pre-processed TBRU dataset directly from the authors of the index study[7]. This dataset consisted of M=500,089 memory T cells from N=271 samples that were profiled with CITE-seq[32], which simultaneously provides both single-cell RNA-seq and single-cell quantification of 31 surface proteins. The TBRU cohort was designed to identify correlates of progression to active tuberculosis infection compared to latent tuberculosis infection. Accordingly, approximately half the samples come from patients who had active TB at enrollment (4 – 7 years before the single-cell data were collected) and half the samples come from household contacts of these patients who developed latent infections after enrollment. The dataset also contains sample-level attributes such as age, sex, weight, and ancestry imputed from genotype information about each sample. See Supplementary Table 6 for the full list of sample-level information we analyzed with CNA.

The authors of the original study used canonical correlation analysis to create a 20-dimensional representation of each cell that incorporated information shared by the RNA-seq and surface protein modalities. This representation was subsequently run through Harmony[9] to remove potential sample-specific and batch effects. To maximize comparability between the published analysis and ours, we used this representation in all analyses unless stated otherwise.

Unsupervised analysis: We computed the initial NAM by running CNA on the full dataset with correction only for batch and per-sample averages of i) the percent mitochondrial reads (pMT) of each cell, and ii) the number of unique molecular identifiers (nUMI) for each cell. To identify biological processes corresponding to NAM-PCs, we then computed the correlation per-gene (the top 5,000 most variable in the dataset) and per-protein between expression level in each cell and the cells' anchored neighborhood loadings on each NAM-PC. To evaluate the extent to which NAM-PC2 reflects known sex differences in T cell populations, we computed per-sample the ratio of total cell fraction from CD4+-labeled clusters to total cell fraction from CD8+-labeled clusters using cluster assignments with cell-type annotations from the original publication of this dataset [6]. We then computed the correlation (and corresponding analytical p-value using a beta-distributed null) between CD4+/CD8+ ratio and sample loading on NAM-PC2. We also computed per-sample the ratio of total cell fraction in the T-regulatory CD4+ cluster to total cell fraction from all CD4+-labeled clusters, and computed the correlation (and p-value) between this ratio and sample loading on NAM-PC2. Finally, we computed the correlation across cells between NAM-PC2 neighborhood loadings and binary cell membership in CD4+-labeled clusters, CD8+-labeled clusters, and the T-regulatory cluster. (See Supplementary Figure 12, Supplementary Table 12, and Supplementary Table 13.)

Association analysis for TB progression phenotype: We analyzed the TB progression attribute with CNA, controlling for the same covariates that the authors of the original study used in their analysis: pMT, nUMI, age, age squared, sex, season of blood draw, and percent European ancestry. We retained cells whose neighborhood coefficients showed correlation to TB progression at FDR < 0.05. These cells clearly segregated into two contiguous groups in UMAP space: a depleted population and an enriched population. We examined the genes, among the top 5,000 most variable, and the surface proteins whose expression per-cell was most highly correlated with the cells' anchored neighborhoods' estimated abundance correlations to the TB phenotype.

Association survey across many sample-level attributes: We one-hot encoded all categorical attributes and standardized all continuous attributes. We then removed any attributes with missing values for >10% of samples, any one-hot categorical attributes with fewer than 20 individuals represented, and one from every attribute pair with a correlation >0.75. Seventeen of the attributes were retained after this step. We then determined for each of these 17 attributes y which others (including TB progression status) had a nominally significant ($p<0.05$) correlation to y and included those as covariates when analyzing y. Using the resulting selected covariates, shown in Supplementary Table 6, we ran CNA. Using the 31 clusters previously identified in this data, we ran a per-cluster association

test with identical covariate control for each attribute. We added multiple hypothesis testing correction across clusters, and, for both cluster-based analysis and CNA, across the 17 attributes tested. For each attribute with a globally-significant association by CNA, we examined the genes, among the top 5,000 variable genes, and the surface proteins whose expression per anchor cell was most highly correlated with corresponding neighborhood coefficients to the given attribute.

Data Availability

All data analyzed during this study were available through three previously-published articles [6–8].

Code Availability

An open-source repository containing code for running CNA is available at https://github.com/immunogenomics/cna, an open-source repository containing code underlying all figures and tables is available at https://github.com/immunogenomics/cna-display, and an open-source repository containing code underlying all simulations is available at https://github.com/immunogenomics/cna-sim.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Anika Gupta, Dylan Kotliar, Yang Luo, Nghia Millard, Miguel Reyes, Saori Sakaue, Fan Zhang, the members of the CGTA discussion group, and the Raychaudhuri lab for helpful discussions and feedback. This work is supported in part by funding from the National Institutes of Health (UH2AR067677, U19 AI111224, U01 HG009379, and 1R01AR063759). SA was supported by the Swiss National Science Foundation (SNSF) postdoctoral mobility fellowships P2ELP3_172101 and P400PB_183823 and NIH T32 grant T32HG010464. LR was supported in part by the NIH-NHGRI T32 (5T32HG2295–17). JK was supported by award Number T32GM007753 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

1. Kashima Y, et al. , Single-cell sequencing techniques from individual to multiomics analyses. Experimental & Molecular Medicine, 2020. 52(9): p. 1419–1427. [PubMed: 32929221]
2. Andrews TS, et al. , Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc, 2021. 16(1): p. 1–9. [PubMed: 33288955]
3. Luecken MD and Theis FJ, Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol, 2019. 15(6): p. e8746. [PubMed: 31217225]
4. Traag VA, Waltman L, and van Eck NJ, From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep, 2019. 9(1): p. 5233. [PubMed: 30914743]
5. Burkhardt DB, et al. , Quantifying the effect of experimental perturbations at single-cell resolution. Nat Biotechnol, 2021.

6. Nathan A, et al. , Multimodal memory T cell profiling identifies a reduction in a polyfunctional Th17 state associated with tuberculosis progression. bioRxiv, 2020: p. 2020.04.23.057828.
7. Reyes M, et al. , An immune-cell signature of bacterial sepsis. Nat Med, 2020. 26(3): p. 333–340. [PubMed: 32066974]
8. Wei K, et al. , Notch signalling drives synovial fibroblast identity and arthritis pathology. Nature, 2020. 582(7811): p. 259–264. [PubMed: 32499639]
9. Korsunsky I, et al. , Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods, 2019. 16(12): p. 1289–1296. [PubMed: 31740819]
10. Butler A, et al. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology 36(5): 411–420.
11. Haghverdi L, et al. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36(5): 421–427. [PubMed: 29608177]
12. Liu J, et al. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. Nature Protocols 15(11): 3632–3662. [PubMed: 33046898]
13. Fonseka CY, et al. , Mixed-effects association of single cells identifies an expanded effector CD4(+) T cell subset in rheumatoid arthritis. Sci Transl Med, 2018. 10(463).
14. Millard N, et al. , Maximizing statistical power to detect clinically associated cell states with scPOST. bioRxiv, 2020: p. 2020.11.23.390682.
15. Liu Z, et al. , Notch signaling in postnatal joint chondrocytes, but not subchondral osteoblasts, is required for articular cartilage and joint maintenance. Osteoarthritis Cartilage, 2016. 24(4): p. 740–51. [PubMed: 26522700]
16. Wang X and Astrof S, Neural crest cell-autonomous roles of fibronectin in cardiovascular development. Development, 2016. 143(1): p. 88–100. [PubMed: 26552887]
17. Zhang F, et al. (2019). “Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry.” Nat Immunol 20(7): 928–942. [PubMed: 31061532]
18. Sanlioglu S, et al. , Lipopolysaccharide induces Rac1-dependent reactive oxygen species formation and coordinates tumor necrosis factor-alpha secretion through IKK regulation of NF-kappa B. J Biol Chem, 2001. 276(32): p. 30188–98. [PubMed: 11402028]
19. Pan C, et al. , Suppression of the RAC1/MLK3/p38 Signaling Pathway by β-Elemene Alleviates Sepsis-Associated Encephalopathy in Mice. Frontiers in Neuroscience, 2019. 13(358).
20. von Knethen A and Brüne B (2019). Histone Deacetylation Inhibitors as Therapy Concept in Sepsis. Int J Mol Sci 20(2).
21. Wu H-P, et al. (2011). Serial increase of IL-12 response and human leukocyte antigen-DR expression in severe sepsis survivors. Critical Care 15(5): R224. [PubMed: 21939530]
22. Steinhauser ML, et al. (1999). Multiple Roles for IL-12 in a Model of Acute Septic Peritonitis. The Journal of Immunology 162(9): 5437–5443. [PubMed: 10228022]
23. Oliveira NM, et al. (2016). Sepsis induces Telomere Shortening: a Potential Mechanism Responsible for Delayed Pathophysiological Events in Sepsis Survivors? Molecular Medicine 22(1): 886–891. [PubMed: 27925632]
24. Gutierrez-Arcelus M, et al. , A genome-wide innateness gradient defines the functional state of human innate T cells. bioRxiv, 2018: p. 280370.
25. Cano-Gamez E, et al. , Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4(+) T cells to cytokines. Nat Commun, 2020. 11(1): p. 1801. [PubMed: 32286271]
26. Luecken M, et al. , Benchmarking atlas-level data integration in single-cell genomics. bioRxiv, 2020: p. 2020.05.22.111161.
27. Stuart T and Satija R, Integrative single-cell analysis. Nature Reviews Genetics, 2019. 20(5): p. 257–272.
28. Klein SL and Flanagan KL, Sex differences in immune responses. Nature Reviews Immunology, 2016. 16(10): p. 626–638.
29. Silva CL, et al. (2001). “Cytotoxic T cells and mycobacteria.” FEMS Microbiology Letters 197(1): 11–18. [PubMed: 11287139]

30. Li M, et al. (2019). “Age related human T cell subset evolution and senescence.” Immunity & Ageing 16(1): 24. [PubMed: 31528179]

31. Shirai T, et al. (2003). “TH1-biased immunity induced by exposure to Antarctic winter.” J Allergy Clin Immunol 111(6): 1353–1360. [PubMed: 12789239]

32. Stoeckius M, et al. , Simultaneous epitope and transcriptome measurement in single cells. Nature Methods, 2017. 14(9): p. 865–868. [PubMed: 28759029]

33. Korotkevich G, et al. , Fast gene set enrichment analysis. bioRxiv, 2021: p. 060012.

Methods References

1. Nathan A, et al. , Multimodal memory T cell profiling identifies a reduction in a polyfunctional Th17 state associated with tuberculosis progression. bioRxiv, 2020: p. 2020.04.23.057828.

2. Stoeckius M, et al. , Simultaneous epitope and transcriptome measurement in single cells. Nature Methods, 2017. 14(9): p. 865–868. [PubMed: 28759029]

3. Korsunsky I, et al. , Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods, 2019. 16(12): p. 1289–1296. [PubMed: 31740819]

4. Traag VA, Waltman L, and van Eck NJ, From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep, 2019. 9(1): p. 5233. [PubMed: 30914743]

5. Wei K, et al. , Notch signalling drives synovial fibroblast identity and arthritis pathology. Nature, 2020. 582(7811): p. 259–264. [PubMed: 32499639]

6. Reyes M, et al. , An immune-cell signature of bacterial sepsis. Nat Med, 2020. 26(3): p. 333–340. [PubMed: 32066974]

7. Korotkevich G, et al. , Fast gene set enrichment analysis. bioRxiv, 2021: p. 060012.

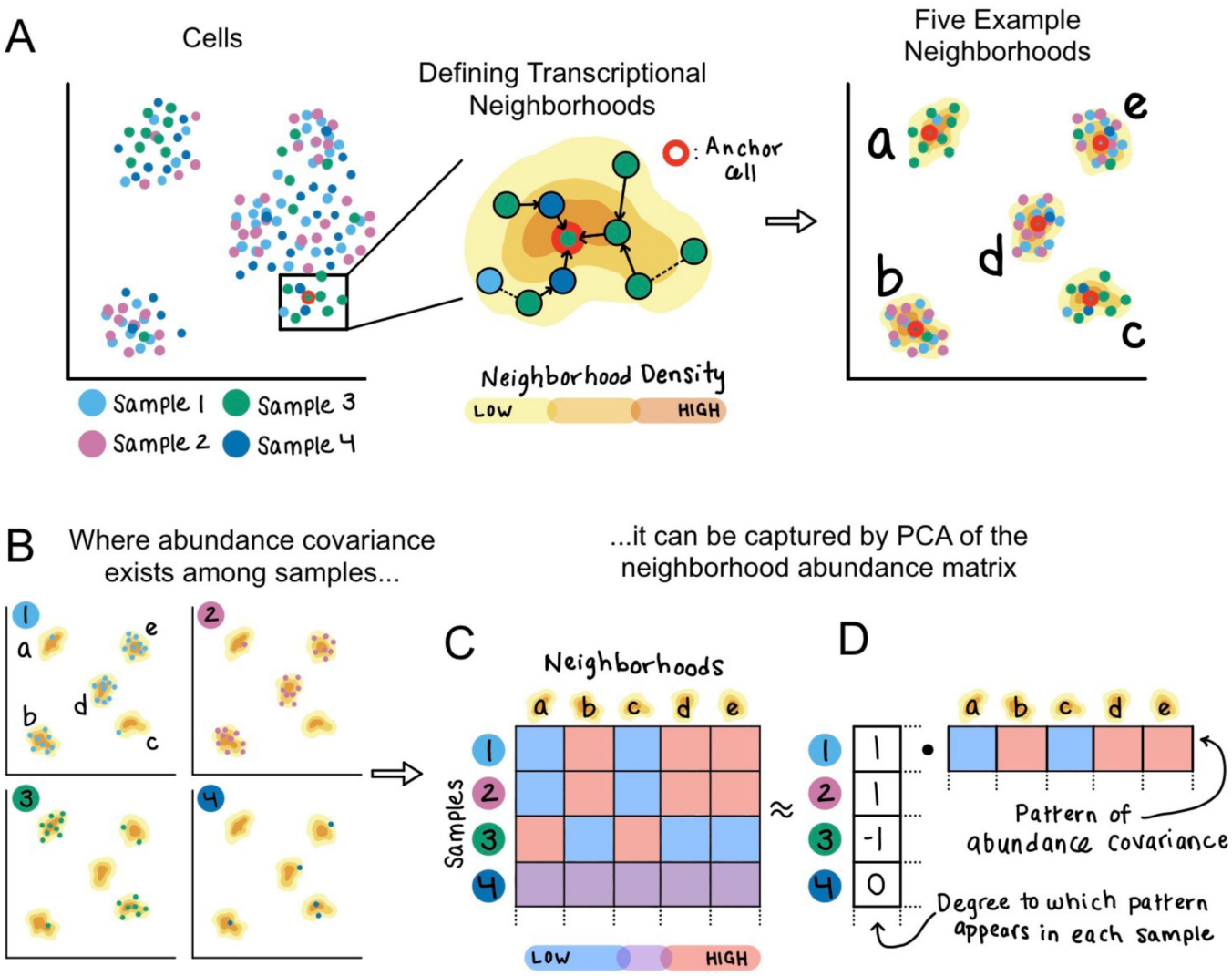

Figure 1: Method schematic.

(A) Given an example dataset of single cells sampled from four individuals, CNA defines one transcriptional neighborhood per cell in the dataset. Each other cell in the dataset belongs to this neighborhood according to the probability that a random walk in the cell-cell similarity graph from that cell will arrive at the neighborhood's anchor cell after a certain number of steps. Five example neighborhoods a-e are depicted. (**B**) Examining the representation of cells from each sample in these example neighborhoods identifies a pattern of abundance covariation. Neighborhoods b, d, and e tend to have a high abundance when neighborhoods a and c have low abundance, and vice versa. This covariation pattern appears in samples 1–3 but not in sample 4. (**C**) The neighborhood abundance matrix (NAM) quantifies the fractional abundance of cells in each neighborhood for each sample; we indicate higher abundance with red and lower abundance with blue. (**D**) Dominant patterns of abundance covariation across neighborhoods can be illuminated by factorizing the NAM, for example with PCA. The principal component corresponding to this example has per-neighborhood loadings that capture the neighborhood covariance pattern, as well as

per-sample loadings that reflect the degree to which the covariance pattern appears in each sample.

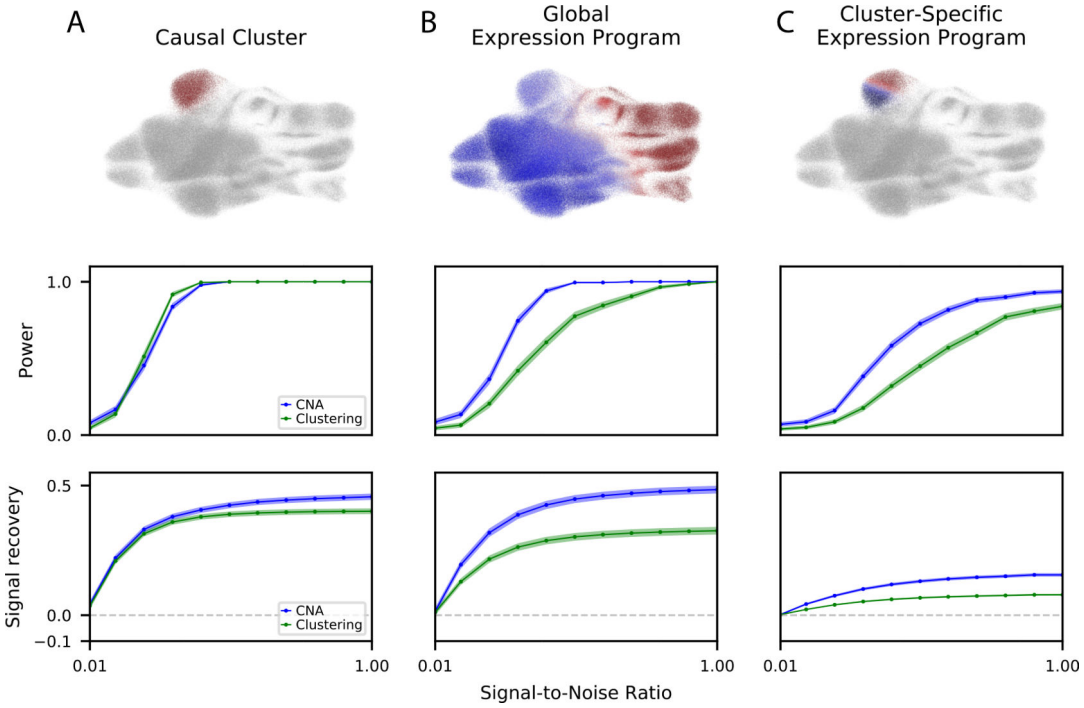

Figure 2: Power and signal recovery assessed in simulation.

We simulated three ground truth signal types: (**A**) causal clusters, (**B**) global gene expression programs, and (**C**) cluster-specific gene expression programs. For each signal type, we show (**top**) an example of the signal in UMAP space indicating the contribution of each cell to its respective sample's attribute value. with warmer colors indicating a positive contribution, cooler colors indicating a negative association, and grey indicating neutral contribution, (**middle**) the power of CNA versus a cluster-based approach (MASC) across a range of signal-to-noise ratios at $\alpha = 0.05$, and (**bottom**) the signal recovery of CNA versus a cluster-based approach across a range of signal-to-noise ratios. 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.

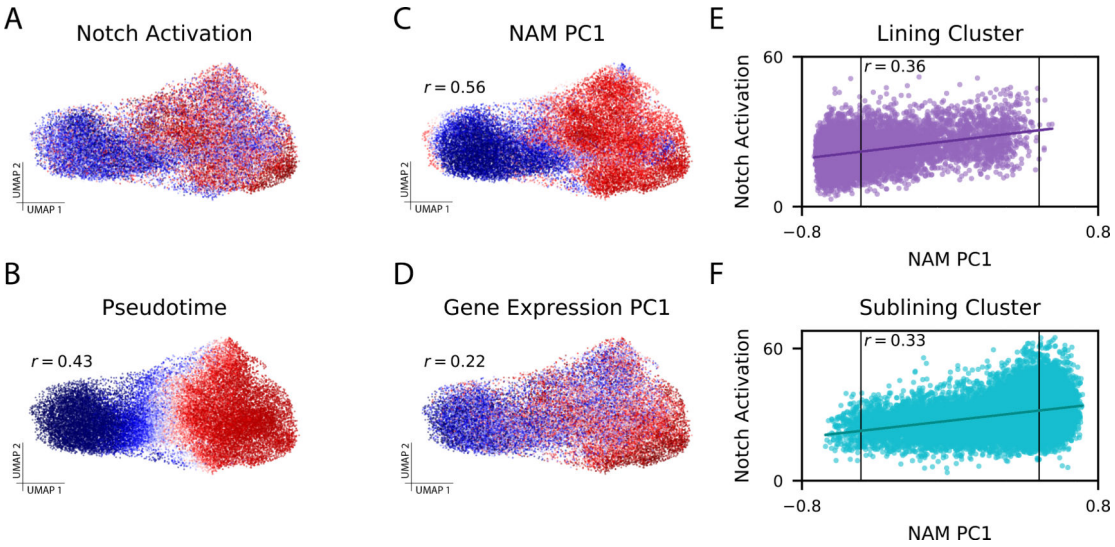

Figure 3: CNA captures Notch activation gradient in rheumatoid arthritis dataset.

(**A**) Experimentally-determined Notch activation score per fibroblast cell. (**B**) Pseudotime assignments per cell (Spearman r=0.43 to Notch activation score). (**C**) The loading on NAM-PC1 for each cell's anchored neighborhood (Spearman r=0.56 to Notch activation score). (**D**) Cell loadings on PC1 of the gene expression matrix (Spearman r=0.22 to Notch activation score). (**E**) Notch activation score per cell assigned to the lining cluster and (**F**) Notch activation score per cell assigned to the sublining cluster, each plotted against the anchored neighborhood's loading on NAM-PC1. The FDR<0.05 thresholds beyond which each neighborhood was considered expanded in RA (right) or depleted in RA (left) are marked with vertical lines on (E) and (F), highlighting that some cells from each cluster are included in the expanded population and the depleted population.

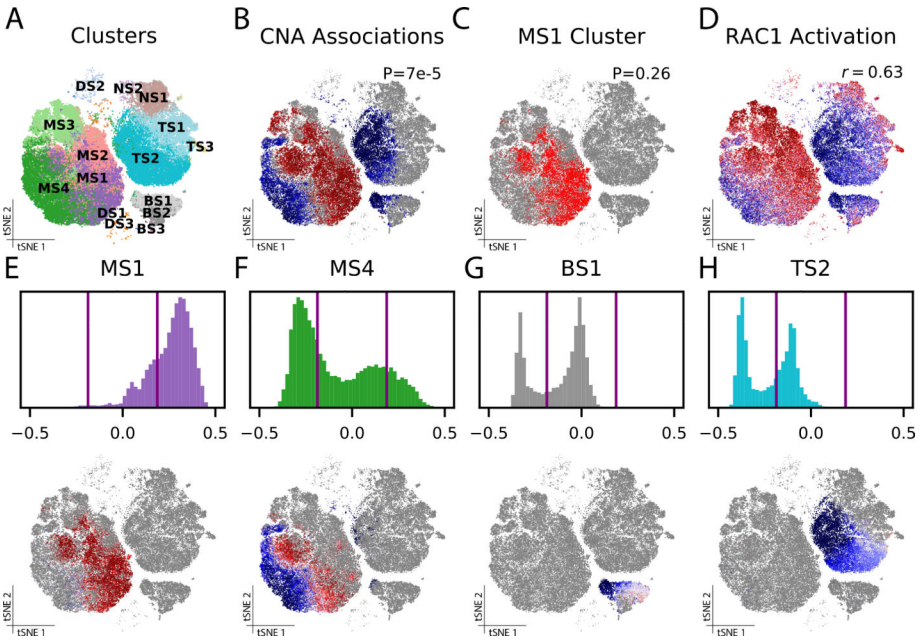

Figure 4: CNA refines sepsis-associated blood cell populations.

(**A**) Clusters from the published analysis. (**B**) Results of association test for sepsis across the whole cohort using CNA (CNA global p=7e-5): each cell is colored according to its neighborhood coefficient, with red indicating high correlation and blue indicating low correlation. (**C**) MS1, the cluster closest to approaching nominal significance (MASC p=0.26) in a cluster-based association test for sepsis across the whole cohort. (**D**) Cells colored according to their summed expression of genes in the RAC1 activation gene set. (**E-H**) The distribution of neighborhood coefficients to sepsis phenotype within several of the original clusters—MS1 (**E**), MS4 (**F**), BS1 (**G**) and TS2 (**H**)—are shown as histograms (top) and in tSNE space (bottom).

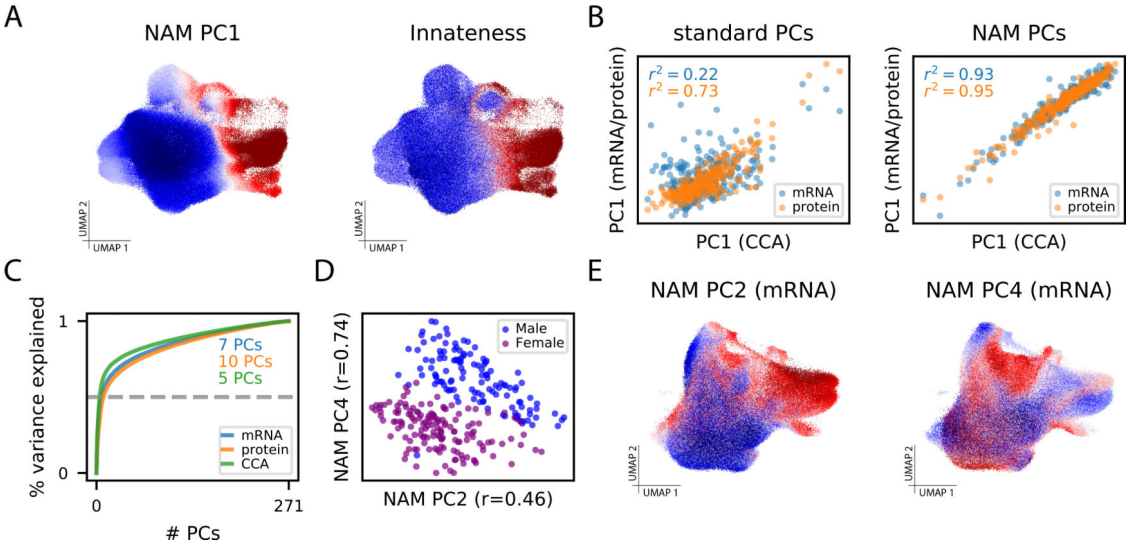

Figure 5: CNA characterizes biologically meaningful structure in TB dataset.

(**A**) UMAP with cells colored by their anchored neighborhood's loading on NAM-PC1 (left) or the transcriptional score of innateness from Gutierrez-Arcelus, *et al.* (right)[24]. (**B**) (Left) Sample loadings along the first PCs resulting from naive PCA of mRNA expression and protein expression plotted against the same loadings for the CCA-based joint mRNA/protein representation. (Right) Sample loadings along the first PCs resulting from PCA of the NAM generated from mRNA expression and protein expression plotted against the same loadings for the CCA-based joint mRNA/protein representation. (**C**) The cumulative percent of variance in the NAM explained by the NAM-PCs in each data modality. The minimum number of NAM-PCs needed to capture 50% of variance in each modality are highlighted. (**D**) Plot of sample loadings on NAM-PC2 and -PC4 colored by biological sex. (**E**) UMAP with cells colored by their anchored neighborhood's loading on NAM-PC2 (left) or NAM-PC4 (right).

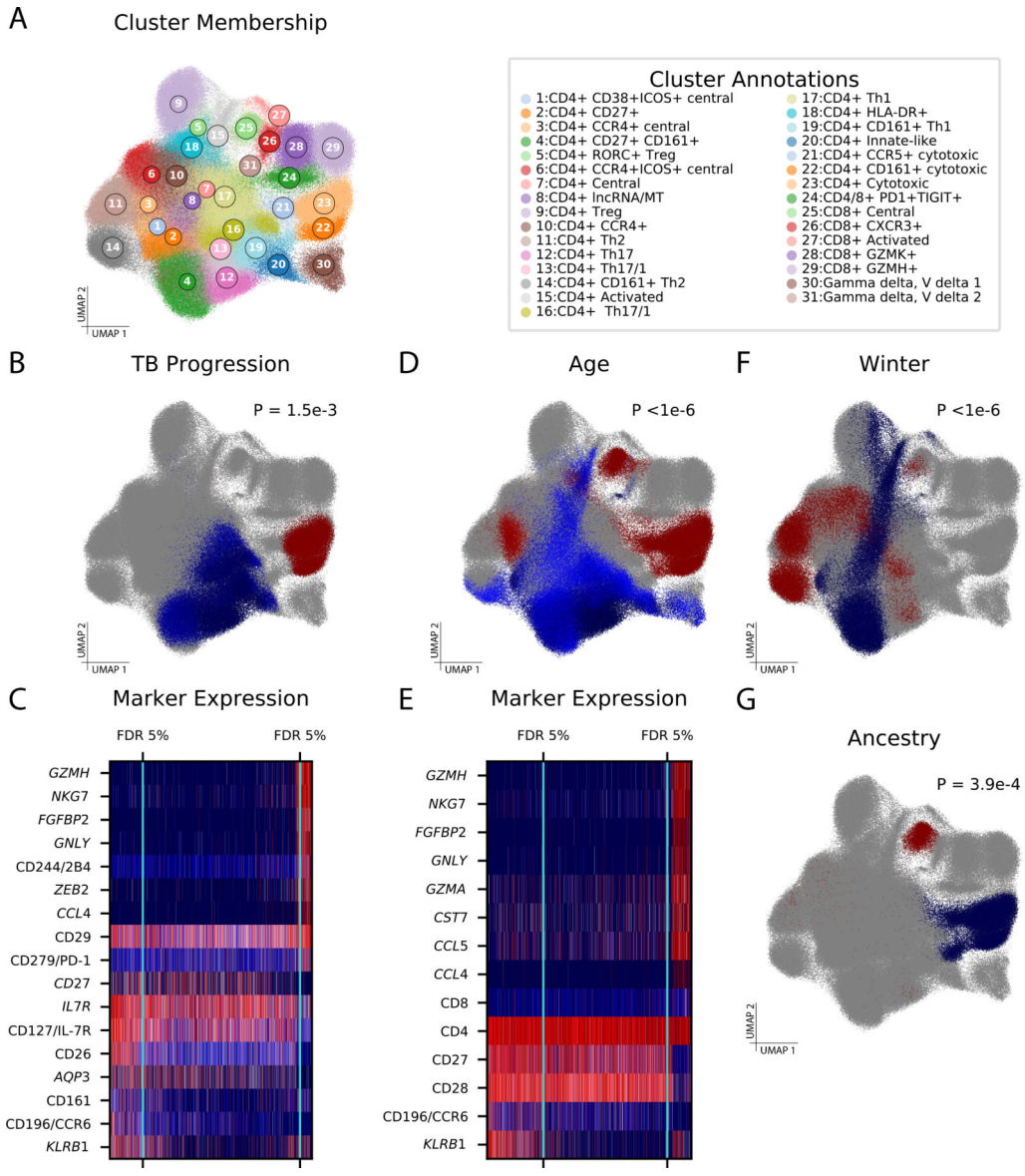

Figure 6: CNA improves characterization of diverse sample attributes in a tuberculosis cohort.

(**A**) UMAP of memory T cells colored by cluster assignment in the original study. (**B**) Cells in UMAP space, colored according to their neighborhood's abundance correlation to TB progressor versus non-progressor status, with red indicating high correlation and blue indicating low correlation. Cells whose neighborhood coefficients did not pass a FDR<0.05 threshold for association are shown in grey. CNA global p-value is shown. **(C)** A heatmap of expression for genes and proteins of biological interest across cells, with red indicating high expression and blue indicating low expression. The cells are ordered from left to right according to their neighborhood coefficients to TB progressor status, and the FDR <0.05 thresholds beyond which cells are included in the significantly depleted (left) or expanded (right) population are shown in aqua. (**D**) Populations expanded and depleted with increasing age. CNA global p-value is shown. **(E)** Heatmap of expression for genes and proteins of biological interest for the age association. (**F**) Populations expanded and

depleted among samples drawn during the winter season relative to samples drawn during other seasons. CNA global p-value is shown. (**G**) Populations expanded and depleted with increasing global fraction of European genetic ancestry. CNA global p-value is shown.