PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Scalable multi-sample single-cell data analysis by Partition-Assisted Clustering and Multiple Alignments of Networks

Ye Henry Li[1☯], Dangna Li[2☯], Nikolay Samusik[3], Xiaowei Wang[4], Leying Guan[4], Garry P. Nolan[3], Wing Hung Wong[4,5]*

1 Structural Biology Department and Public Policy Program, Stanford University, Stanford, United States of America, **2** Institute for Computational and Mathematical Engineering, Stanford University, Stanford, United States of America, **3** Department of Microbiology and Immunology, Baxter Laboratory, Stanford University, Stanford, United States of America, **4** Statistics Department, Stanford University, Stanford, United States of America, **5** Department of Biomedical Data Science, Stanford University, Stanford, United States of America

☯ These authors contributed equally to this work.
* whwong@stanford.edu

Citation: Li YH, Li D, Samusik N, Wang X, Guan L, Nolan GP, et al. (2017) Scalable multi-sample single-cell data analysis by Partition-Assisted Clustering and Multiple Alignments of Networks. PLoS Comput Biol 13(12): e1005875. https://doi.org/10.1371/journal.pcbi.1005875

Editor: Florian Markowetz, University of Cambridge, UNITED KINGDOM

Received: March 28, 2017

Accepted: November 6, 2017

Published: December 27, 2017

Data Availability Statement: The datasets generated in our study are publicly available on https://community.cytobank.org under PAC-MAN Dataset.

Funding: This research was supported by the Stanford SIGF-BioX fellowship (https://biox.stanford.edu/) and the NIH T32 GM007276 training grant to YHL and the NIH R01GM109836, NSF-DMS1330132, and NSF-DMS1407557 grants to WHW. The funders had no role in study design,

Abstract

Mass cytometry (CyTOF) has greatly expanded the capability of cytometry. It is now easy to generate multiple CyTOF samples in a single study, with each sample containing single-cell measurement on 50 markers for more than hundreds of thousands of cells. Current methods do not adequately address the issues concerning combining multiple samples for subpopulation discovery, and these issues can be quickly and dramatically amplified with increasing number of samples. To overcome this limitation, we developed Partition-Assisted Clustering and Multiple Alignments of Networks (PAC-MAN) for the fast automatic identification of cell populations in CyTOF data closely matching that of expert manual-discovery, and for alignments between subpopulations across samples to define dataset-level cellular states. PAC-MAN is computationally efficient, allowing the management of very large CyTOF datasets, which are increasingly common in clinical studies and cancer studies that monitor various tissue samples for each subject.

Author summary

Recently, the cytometry field has experienced rapid advancement in the development of mass cytometry (CyTOF). CyTOF enables a significant increase in the ability to monitor 50 or more cellular markers for millions of cells at the single-cell level. Initial studies with CyTOF focused on few samples, in which expert manual discovery of cell types were acceptable. As the technology matures, it is now feasible to collect more samples, which enables systematic studies of cell types across multiple samples. However, the statistical and computational issues surrounding multi-sample analysis have not been previously examined in detail. Furthermore, it was not clear how the data analysis could be scaled for hundreds of samples, such as those in clinical studies. In this work, we present a scalable analysis pipeline that is grounded in strong statistical foundation. Partition-Assisted

data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Clustering (PAC) offers fast and accurate clustering and Multiple Alignments of Networks (MAN) utilizes network structures learned from each homogeneous cluster to organize the data into data-set level clusters. PAC-MAN thus enables the analysis of a large CyTOF dataset that was previously too large to be analyzed systematically; this pipeline can be extended to the analysis of similarly large or larger datasets.

Introduction

Analyses of CyTOF data rely on many of the tools and ideas from flow cytometry (FC) data analysis, as CyTOF datasets are essentially higher dimensional versions of flow cytometry datasets. Currently, the most widely used method in FC is still human hand-gating, as other methods often fail to extract meaningful subpopulations of cells automatically. In hand-gating, we draw polygons or other enclosures around pockets of cell events on a two-dimensional scatterplot to define subpopulations and cellular states that are observed in the data. This process is painfully time-consuming and requires advance knowledge of the marker panel design, the quality of the staining reagents, and, most importantly, *a priori* what cell subpopulations to expect to occur in the data. When presented with a new set of marker panels and biological system, the researcher would find it difficult to delineate the cell events, especially in high-dimensional and multi-sample datasets.

The inefficient nature of hand-gating in flow cytometry motivated algorithmic development in automatic gating. Perhaps the most popular is flowMeans[1], which is optimized for FC and can learn subpopulations in FC data[2] in an automated manner; however, it has not been successfully applied to CyTOF data analysis. Currently, most data analysis tools created for flow cytometry data analyses are not easily applicable for high-dimensional datasets[3]. An exception is SPADE, which was developed and optimized specifically for the analysis of CyTOF datasets[3]. flowMeans and SPADE constitute the leading computational methods in cytometry, but as shown later in this work, their performance may become sub-optimal when challenged with large and high-dimensional datasets. There are also other recent clustering-based tools that utilize dimensionality reduction and projections of high-dimensional data, however, these tools do not directly learn the subpopulations for all the cell events, and may be too slow to complete data analysis for an increasing amount of samples.

In this study, we address the data analysis challenges in two major steps. First, we propose the partition-assisted clustering (PAC) approach, which produces a partition of the k-dimensional space (k = number of markers) that captures the essential characteristic of the data distribution. This partitioning methodology is grounded in a strong mathematical framework of partition-based high-dimensional density estimation[4–7]. The mathematical framework offers the guarantee that these partitions approximate the underlying empirical data distribution; this step is faster than the recent k-nearest neighbor-based method [8] and is essential to the scalability of our clustering approach to analyze datasets with many samples. The clustering of cells based on recursive partitioning is then refined by a small number of k-means style iterations before a merging step to produce the final clustering.

Secondly, the subpopulations learned separately in multiple different but related datasets can be aligned by marker network structures (multiple alignments of networks, or MAN), making it possible to characterize the relationships of subpopulations across different samples automatically. The ability to do so is critical for monitoring changes in a subpopulation across different conditions. Importantly, in every study, batch effect is present; batch effects shift subpopulation signals so that the means can be different from experiment to experiment.

PAC-MAN naturally addresses batch effects in finding the alignments of the same or closely related subpopulations from different samples.

PAC-MAN finds homogeneous clusters efficiently with all data points in a scalable fashion and enables the matching of these clusters across different samples to discover cluster relationships in the form of clades.

Results/Discussion

PAC

PAC has two parts: partitioning and post-processing. In the partitioning part of PAC, the data space is recursively divided into smaller hyper-rectangles based on the number of data points in the locality (Fig 1A). The partitioning is accomplished by either Bayesian Sequential Partition (BSP) with limited look-ahead (Fig 1A and 1B) or Discrepancy Sequential Partition (DSP) (Fig 1A); these are two fast variants of partition-based density estimation methods previously developed by our group [4–7], with DSP being the fastest. BSP and DSP divide the sample space into hyper-rectangles with uniform density value in each of them. The subsetting of cells according to the partitioning provides a principled way of clustering the cells that reflects the characteristics of the underlying distribution. In particular, each significant mode is captured by a number of closely located rectangles with high-density values (Fig 1C). Although this method allows a fast and unbiased localization of the high-density regions of the data space, we should not use the hyper-rectangles directly to define the final cluster boundaries for two reasons. First, real clusters are likely to be shaped elliptically, therefore, the data points in the corners of a hyper-rectangle are likely to be incorrectly clustered. Second, a real cluster is often split into more than one closely located high-density rectangles. We designed post-processing steps to overcome these limitations: 1) a small number of k-means iterations is used to round out the corners of the hyper-rectangles, 2) a merging process is implemented to ameliorate the splitting problem, which is inspired by the flowMeans algorithm. The details of post-processing are given in the Materials and Methods. The resulting method is named b-PAC or d-PAC depending on whether the partition is produced by BSP or DSP.

MAN

An approach to analyze multiple related samples of CyTOF data is to pool all samples into a combined sample before detection of subpopulations. This is a natural approach under the assumptions that there are no significant batch effects or systematic shifts in cell subpopulations across the different samples. However, such assumptions may not hold due to one or more of the following reasons:

1. **Dataset size and instruments used.** Large number of samples usually means the samples were collected on different days with different experimental preparations. Many steps can introduce significant shifts in measurement levels.
2. **Staining reagents.** Reagents such as antibodies, purchased from different vendors and batch preparations can affect the overall signal. While saturation of reagents in the protocol could help eliminate the batch effects in the staining procedure, this approach is costly and might not work for all antibodies, especially those with poor specificity.
3. **Normalization beads stock.** While normalization beads[9] help to control for the signal level, especially within one experiment, the age of the beads stock and their preparation could lead to significant batch effects. In addition, there are different types of normalization beads and normalization calculations.

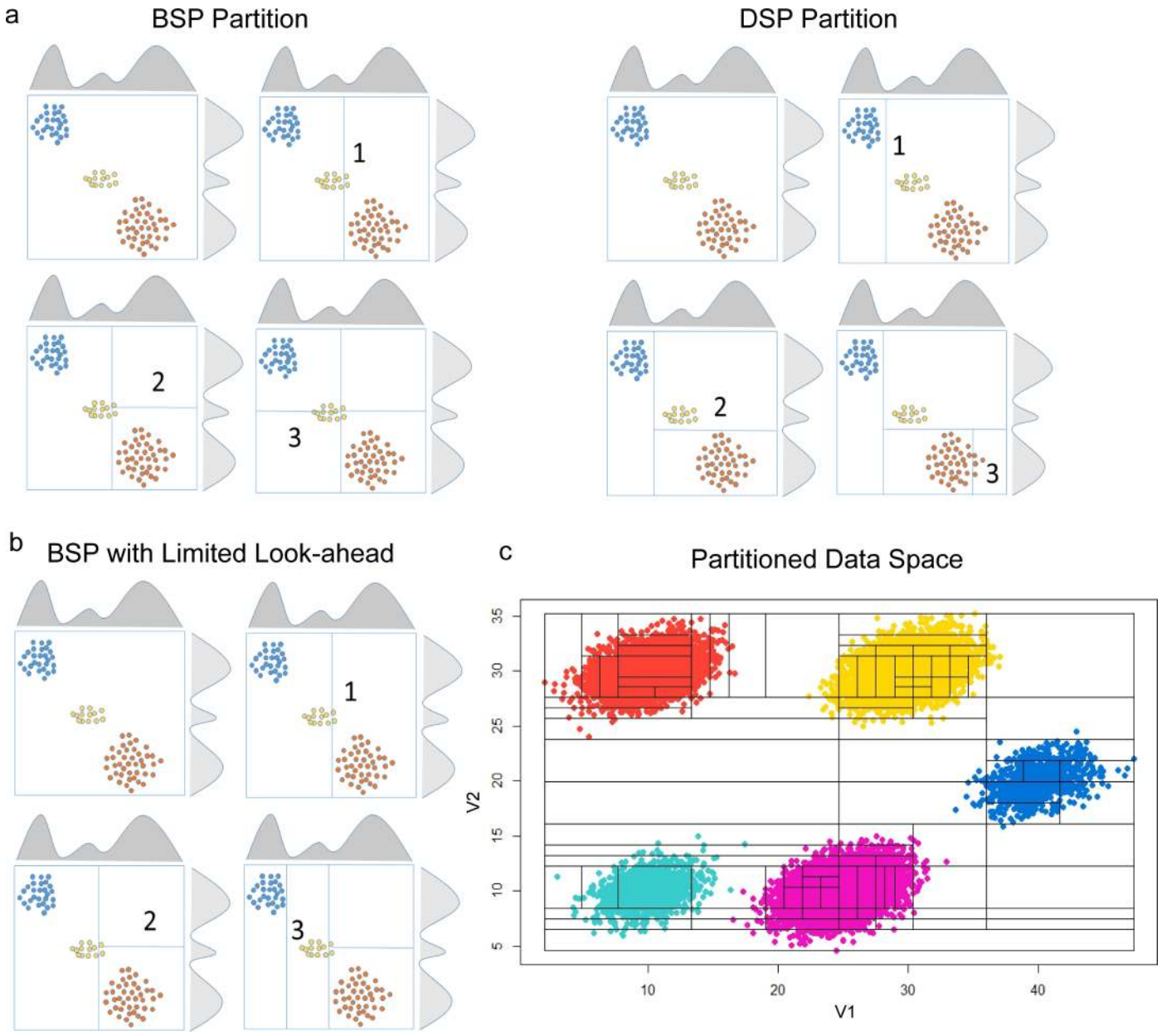

Fig 1. PAC recursively partitions the data space to obtain rational initialization structure. Partition-based methods estimate data density by cutting the data space into smaller rectangles recursively. Shown in parts a-b are three clusters of points, the data marginal densities, and several partition scenarios. The data space (box) is partitioned in sequential steps denoted by numbers on the cut lines. Only the first three partition cuts are shown in parts a and b. (a) Bayesian Sequential Partition (BSP) is a Bayesian procedure that maximizes the posterior of the density estimation by dividing the data space via binary partitions; these partitions occur in the middle of the bounded region. On the other hand, Discrepancy Sequential Partition (DSP) performs division other than the mid-point; here, the division is guided by the discrepancy score through a series of tests of uniformity in point distribution, and the procedure stops when discrepancies are smaller than a set threshold. (b) In the (one-step) look-ahead version of BSP partition, the algorithm cuts the data space for all potential cuts plus one step more (steps 2 and 3), and it finds the optimal future scenario (after step 3). In comparison to the (sub-optimal) BSP scenario (one of many scenarios) illustrated in part (a), the scenario in (b) segregates the gold cluster much better, and it is a preferred cut to make in the continuation of partitioning procedure. In theory, BSP can produce sequential partitions for a pre-set number of steps ahead; however, to maintain computational feasibility, we implemented the one-step look-ahead BSP for this work. (c) The partitioning of simulated data space containing five subpopulations; the hyper-rectangles surround high-density areas, approximating the underlying distribution.

https://doi.org/10.1371/journal.pcbi.1005875.g001

4. **Human work variation.** While many researchers are studying the same system (e.g., immune system), different protocols and implementation by different researchers, who sometimes perform experimental steps slightly differently, can lead to batch effects.
5. **Subpopulation dynamics.** The subpopulation centers can move from sample to sample due to treatments on the cells in treatment-control studies or perturbation studies. General practice is to cluster by phenotypic markers.
6. **Sample background.** If the data came from different cell lines or individuals in a clinical study, the measurement levels and proportions of cell subpopulations would be expected to change from sample to sample. Without expert scrutiny, it would be difficult to make sense of the data with current data analysis tools.

Could we extract shared information that allows us to interpret cross-sample similarities and differences? We note that efforts were made to analyze cross-sample relationships in a previous publication [10], in which the data was carefully collected with barcode reagents in uniform staining, which enable pooling of the data for downstream analysis. Experimentally, it would be difficult to up-scale the barcoding and uniform staining control to a larger number of samples. Furthermore, previous efforts were dependent on down-sampling of the data points, which would significantly affect the clustering results. While it is possible, through careful experimental design and cross-sample controls, to establish uniform staining for a small pooled sample data analysis, there is a need to resolve the above batch effect difficulties for studies that require scalability, such as in the clinical setting in which hundreds of patient blood samples are collected at different times.

To ameliorate the difficulties of potential high-dimensional cluster shifts and scalability, we have designed an alternative approach that is effective in the presence of substantial systematic between-sample variation. In this approach, each sample is analyzed separately (by PAC) to discover within-sample subpopulations. As an exploration step, we over-partition to capture both large and small subpopulations in high-dimension. The subpopulations from all samples are then compared to each other based on a pairwise dissimilarity measure designed to capture the differences in within-sample distributions (among the markers) across two subpopulations. Using this dissimilarity, we perform bottom-up hierarchical clustering of the subpopulations to represent the relationship among the subpopulations. The resulting tree of subpopulations is then used to guide the merging of subpopulations from the same sample, and to establish linkage of related subpopulations from different samples. We note that the design of a dissimilarity measure (Materials and Methods) that is not sensitive to systematic sample-to-sample variation is a novel aspect of our approach. The merging of subpopulations from the same sample is also important, as it offers a way to consolidate any over-partitioning that may have occurred during the initial PAC analysis of each sample. We emphasize that, as with the usage of all statistical methods, the user must utilize samples or datasets that are considered as good as possible and that the sample comparisons make biological sense; interpretation of the analysis results rely on the researchers to collect data with validated reagents for all samples. In general, sensible data would come from 1) samples that are carefully prepared to not include contamination of cells from other tissues, 2) cytometry panel with validated markers that enable the observation of known, coherent cell subpopulations in the tissue samples (important for determining the number of PAC clusters to explore in the partition step to control for aggressive over-partitioning), 3) successful execution of standard cytometry experiment protocol, and 4) collection of data to achieve enough cell events (important for building stable network structures). These steps would ensure the reproducibility of PAC-MAN data analysis. In addition, any novel subpopulation discovery or difference between samples

observed should be validated with downstream experiments (perhaps using low-dimensional flow cytometry and sorting methods).

Rational initialization for PAC increases clustering effectiveness

Appropriate initialization of clustering is very important for eventually finding the optimal clustering labels; PAC works well because the implicit density estimation procedure yields rational centers to learn the modes of sample subpopulations. When tested on the hand-gated CyTOF data on the bone marrow sample in [11], compared to kmeans alone, PAC gives lower total sums of squares and higher F-measures in the subpopulations (Fig 2A and 2B). In the comparison to kmeans, we utilized random kmeans initialization by Lloyd (and Forgy), which uses random initialization, and also kmeans++ initialization, which uses a more advanced initialization [12,13]. The process of rational initialization also helps PAC to converge in 50 iterations (Fig 3) in post-processing, whereas k-means performs very poorly even after 5000 iterations (Fig 4). Through the lens of t-SNE plots (Fig 4), the PAC results are more similar to the hand-gating results, while the k-means, flowMeans, and SPADE clustering results perform poorly. In flowMeans, several large subpopulations are merged. SPADE's separation of points is inconsistent and highly heterogeneous, probably due to its down-sampling nature. On the other hand, by inspection, PAC obtains similar separation for both the major and minor subpopulations as the hand-gating results.

PAC is consistently better than flowMeans and SPADE for simulated datasets and hand-gated cytometry datasets

In the systematic simulation study, we challenged the methods with different datasets with varying number of dimensions, number of subpopulations, and separation between the

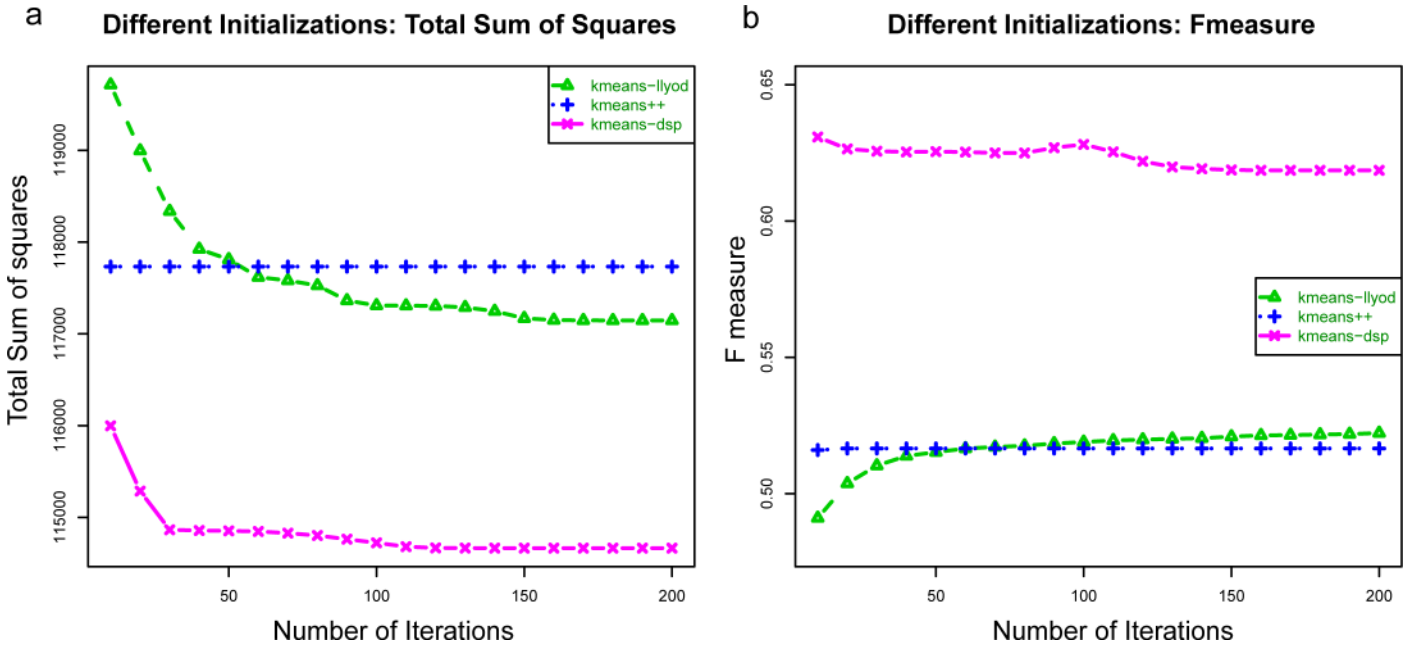

Fig 2. Rational initialization is better than random initialization. The hand-gated CyTOF data (see S1 Fig) is used for illustration. Commonly, kmeans algorithms utilize initialization via the Lloyd's algorithm or kmeans++ algorithm. In comparison, (a) the overall sum of squares error is lower and (b) the F-measure is higher for DSP with kmeans versus the classic kmeans initialization algorithms. The rational initialization helps anchor the cluster starting points, and become very important for the fast convergence of PAC (Fig 3).

https://doi.org/10.1371/journal.pcbi.1005875.g002

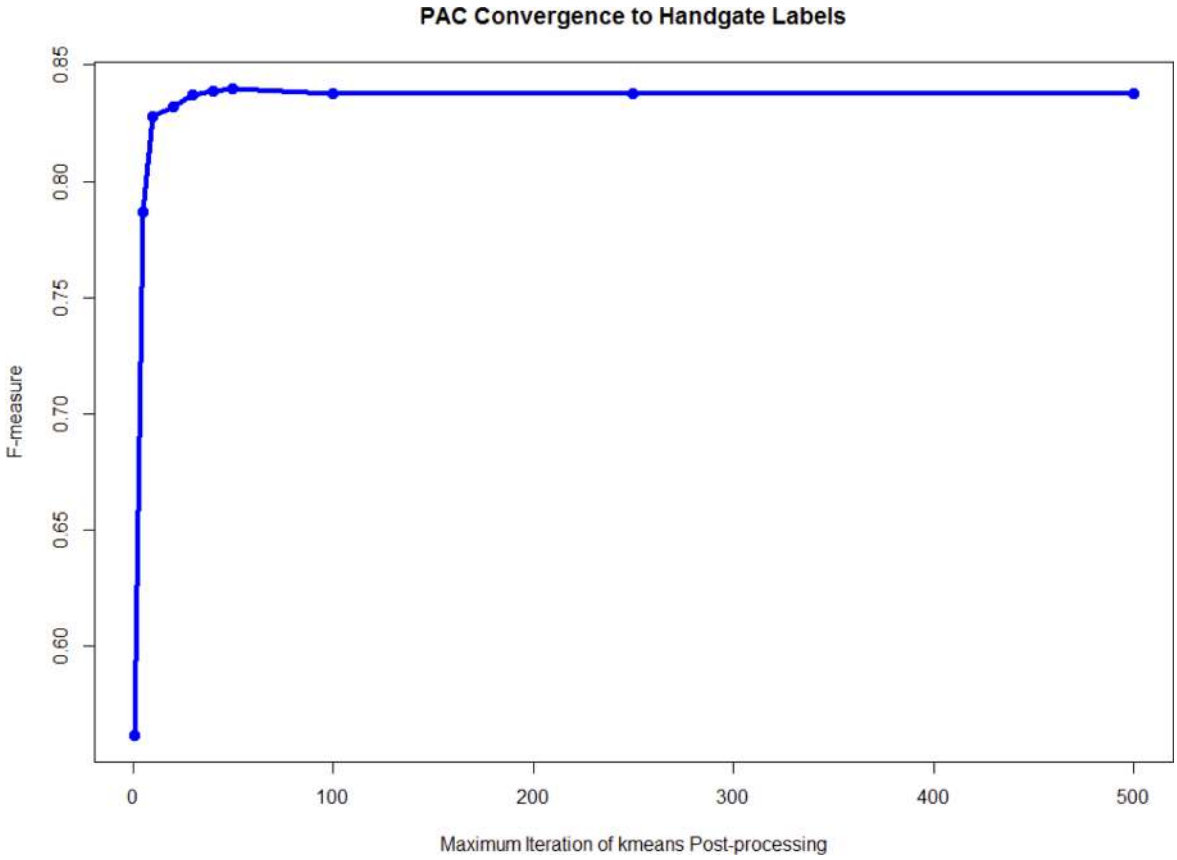

Fig 3. Rational initialization, minimal kmeans post-processing iterations, and merging give fast convergence. We use the hand-gated CyTOF data for illustration. The data space is first partitioned into 50 hyperrectangles, which is about twice (recommended setting) the expected number of subpopulations (24). Next, the number of kmeans iterations was varied followed by flowMeans style merging. The convergence of PAC toward the hand-gated results, or ground truth, is fast due to the informative anchoring of cluster centers around high-density regions by the rational initialization. It takes less than 50 post-processing kmeans iterations for the PAC to achieve convergence. This efficiency allows the PAC method to scale to handle the clustering of large samples.

https://doi.org/10.1371/journal.pcbi.1005875.g003

subpopulations. The F-measure and p-measures for the PAC methods are consistently equal or higher than that of flowMeans and SPADE (Table 1 and S2A Fig). For some higher dimension cases in which the subpopulation separation is relatively small, SPADE failed to cluster. In addition, we observe that flowMeans gives inconsistent F-measures for similar datasets (Table 1), which may be due to the convergence of k-means to a local minimum without a rational initialization.

Next, we tested the methods based on published hand-gated cytometry datasets to see how similar the estimated subpopulations are to those obtained by human experts. We applied the methods on the hematopoietic stem cell transplant and Normal Donors datasets from the FlowCAP challenges[2] and on the subset of gated mouse bone marrow CyTOF dataset (Dataset 9) recently published[11]. The gating strategy of the CyTOF dataset is provided in S1 Fig. The dataset and expert gating strategy are the same as described earlier[14]. Note that in the flow cytometry data, the computed F-measures are slightly lower than that reported in FlowCAP; this is due to the difference in the definition of F-measures. Overall, the PAC outperforms flowMeans and SPADE by consistently obtaining higher F-measures (Table 1). In particular, in the CyTOF data example, PAC generated significantly higher F-measures (greater than 0.82) than flowMeans and SPADE (0.59 and 0.53, respectively). In addition, PAC

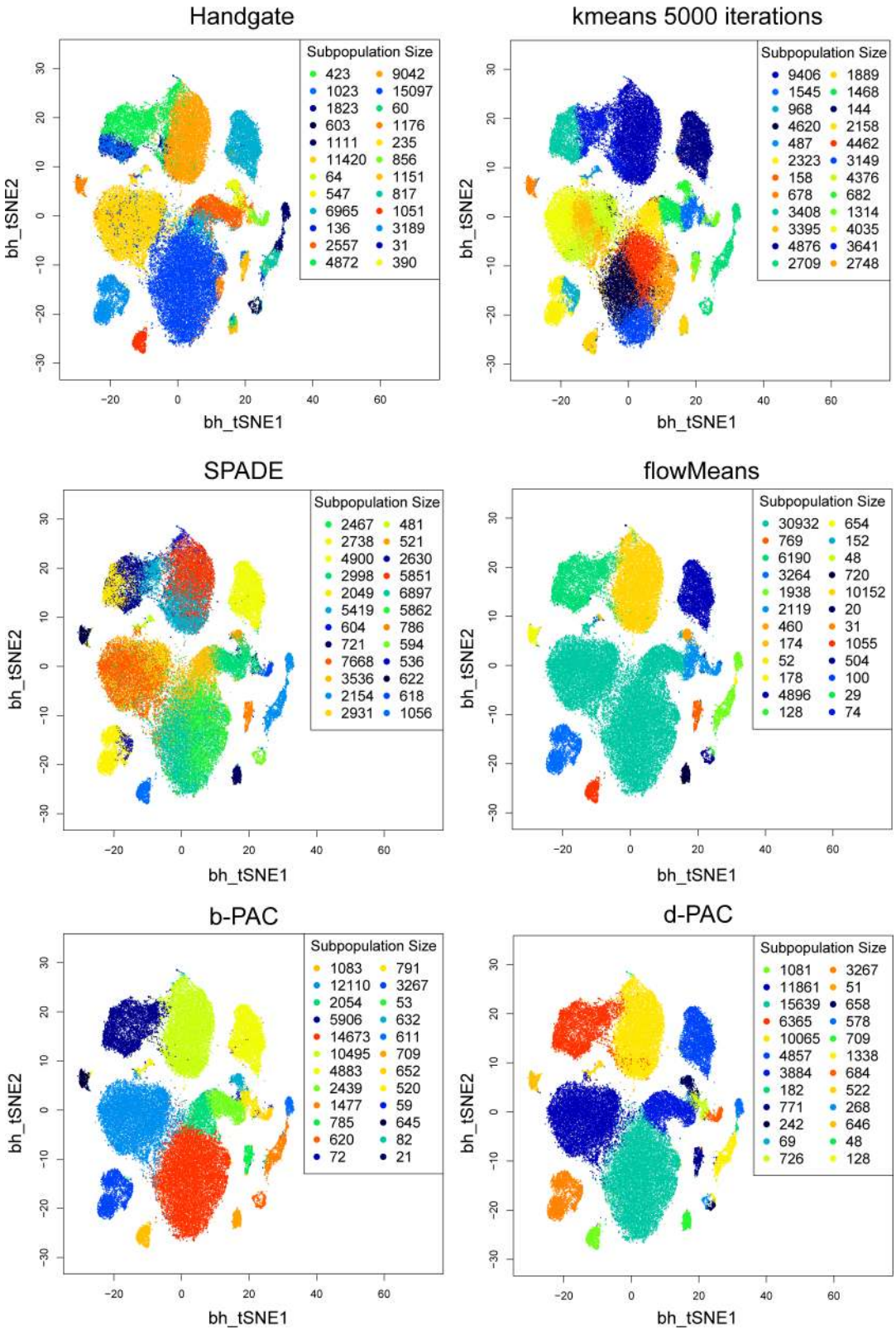

Fig 4. t-SNE visualization of clustering methods. We compare the clustering results between hand-gate, (Lloyd's) kmeans, SPADE, flowMeans, bPAC, and dPAC labels. Each t-SNE plot contains all gated cell events from the hand-gated CyTOF data with different set of colored labels. The colored labels denote different subpopulations within each

plot; however, the colors do not have cross-plot meaning. The subpopulation numbers for all methods were set to be the same as that of hand-gated results (24 subpopulations). PAC methods achieve a significantly better convergence to the hand-gate labels than alternative methods.

https://doi.org/10.1371/journal.pcbi.1005875.g004

gives higher overall subpopulation-specific purities (S2B Fig and S1 Table). These results indicate that PAC gives consistently good results for both low and high-dimensional datasets. Furthermore, PAC results match human hand-gating results very well. The t-SNE 'islands' in the plots are well-colored by the PAC methods, demonstrating that both major and minor/rare subpopulations are captured. The consistency between PAC-MAN results and hand-gating results in this large data set confirms the practical utility of the methodology.

We use t-SNE plots heavily for visualization because t-SNE is a great visualization tool. It is reasonable to ask whether one can obtain good subpopulations by performing cluster analysis on the low-dimensional data points output by t-SNE. Currently, this alternative approach is computationally expensive and not scalable as existing t-SNE implementations cannot be scaled to millions of high-dimensional points, restricting this analysis approach to only hundred of thousands of points in practice. In the downstream, hierarchical clustering or kmeans clustering could be applied; however, hierarchical clustering is very expensive due to the maintenance of a distance matrix during calculations (cannot be easily performed for data with more than thousands of points), while kmeans clustering does not give satisfactory results (S7 Fig) due to the 'flattened' geometry of the high-dimensional points in the t-SNE embedding. Thus, embedding is good for visualization but it is not supposed to capture all information of clusters efficiently. In CyTOF data analysis, we recommend performing PAC methods on the dataset, and utilize t-SNE plots to visualize the clustering results with a subset of points for confirmation.

Separate-then-combine outperforms pool approach when batch effect is present

It is natural to analyze samples separately then combine the subpopulation features for downstream analysis in the multiple samples setting. However, we need to resolve the batch effects. Two distinct subpopulations could overlap in the combined/pooled sample, such as in the case when the data came from two generations of CyTOF instruments (newer instrument elevates the signals). On the other hand, in cases with changing means, two subpopulations can evolve together such that their means change slightly, but enough to shadow each other when samples are merged prior to clustering.

We introduce Multiple Alignments of Networks to resolve the management issue surrounding the organization of homogeneous clusters found in the PAC step (Fig 5). First, we consider the overlapping scenario (Fig 6A). When viewed together in the merged sample, the right subpopulation from sample 1 overlaps with the left subpopulation in sample 2 (Fig 6B left panel). There is no way to use expression level alone to delineate the two overlapping subpopulations (Fig 6B right panel). By learning more subpopulations using PAC, there are some hints that multiple subpopulations are present (Fig 6C). Despite these hints, it would not be possible to say whether the shadowed subpopulations relate in any way to other distinct subpopulations.

PAC-MAN resolves the overlapping issue by analyzing the samples separately (Fig 7). In the case in which we do not know *a priori* the number of true subpopulations, we learn three subpopulations per sample (Fig 7A). The network structures of the subpopulations discovered are presented in Fig 7B and 7C and we see that the third subpopulations from the two samples share the same network structures, while the first two subpopulations of the two samples differ

Table 1. F-measure comparisons of methods on simulated and hand-gated cytometry datasets.**

Data	Analysis Methods			
	flowMeans	SPADE	d-PAC	b-PAC
5_10_40_100k	0.79	0.64	0.94	0.94
5_20_40_100k	0.9	0.73	0.94	0.94
10_5_30_100k	0.74	0.93	0.93	0.97
10_10_30_100k	0.97	0.88	0.98	0.98
10_10_40_100k	0.92	0.95	0.98	0.98
10_20_30_100k	0.88	0.76	0.9	0.91
10_20_40_100k	0.94	0.93	0.95	0.95
10_40_30_100k	0.42	0.55	0.7	0.7
20_5_20_100k	0.75	0.71	0.91	0.9
20_5_30_100k	0.76	0.98	0.99	0.99
20_5_40_100k	0.72	0.85	1.00	1.00
20_10_40_100k	0.25	0.96	0.97	0.97
20_20_40_100k	0.93	0.91	0.92	0.93
35_10_10_100k	0.77	N/A	0.82	0.83
35_10_10_200k	0.56	N/A	0.88	0.88
35_20_10_100k	0.60	N/A	0.70	0.70
35_20_10_200k	0.60	N/A	0.67	0.72
35_10_30_100k	1.00	0.93	1.00	1.00
35_20_30_100k	0.94	N/A	1.00	1.00
35_5_40_200k	0.96	0.89	0.99	0.99
35_10_20_200k	0.96	0.93	1.00	1.00
35_10_40_200k	0.93	0.79	0.96	0.96
40_10_10_100k	0.81	N/A	0.85	0.87
40_10_10_200k	0.73	N/A	0.90	0.90
40_20_10_100k	0.61	N/A	0.71	0.69
40_20_10_200k	0.60	N/A	0.69	0.67
40_10_20_100k	1.00	0.90	1.00	1.00
40_10_20_200k	1.00	0.93	1.00	1.00
40_10_25_100k	0.94	0.94	1.00	1.00
40_20_30_100k	0.95	0.92	1.00	1.00
40_10_25_200k	0.96	0.94	0.99	0.99
50_10_10_100k	0.78	N/A	0.88	0.88
50_10_10_200k	0.80	N/A	0.89	0.88
50_20_10_100k	0.63	N/A	0.73	0.71
50_20_10_200k	0.63	N/A	0.71	0.72
50_10_30_100k	0.96	0.94	1.00	1.00
50_20_20_200k	0.97	N/A	0.98	0.98
50_20_30_200k	0.95	N/A	1.00	1.00
Stem Cell (6 dimensions, 5 subpopulations)	0.98	0.41	0.98	0.91
NDD (12 dimensions, 8 subpopulations)	0.8	0.77	0.79	0.8
CyTOF (39 dimensions, 24 subpopulations)	0.59	0.53	0.84	0.82

F-measure is calculated using the original hand-gate labels and the estimated labels generated by each analysis method. The true-positives are found if the methods assign the same labels to points belonging to the same subpopulation in the hand-gated data. The more true-positives found, the higher the F-measure, which ranges from 0 to 1, with 1 being the highest. Partition-based methods perform consistently well on data ranging from 5 to 50 dimensions. In the simulations, d-PAC and b-PAC perform just as well or better than flowMeans and SPADE. flowMeans gives drastically different F-measures for the cases 20_10_40_100k and 20_20_40_100k: 0.25386 vs. 0.92518; this large difference is likely due to the random initiation of cluster centers. In the hand-gated datasets, SPADE has the worst performance. Ultimately, the performance of flowMeans and SPADE deteriorate for the 39-dimensional real CyTOF data, while d-PAC and b-PAC perform consistently well.

In this table, simulated data have the following convention: a_b_c_d, where a denotes the number of dimensions/markers, b denotes the number of subpopulations, c denotes the edge size of the hypercube for data generation, and d denotes the number of cells. The clustering problem becomes harder as the number of subpopulations increases and the data space volume decreases. We report the results for simulated cases that worked for all methods, except for higher-dimension cases in which the clusters are nearby and SPADE failed to cluster; these SPADE results are denoted with N/A.

** In this table, 1.00 means a number which rounds up to 1.00.

https://doi.org/10.1371/journal.pcbi.1005875.t001

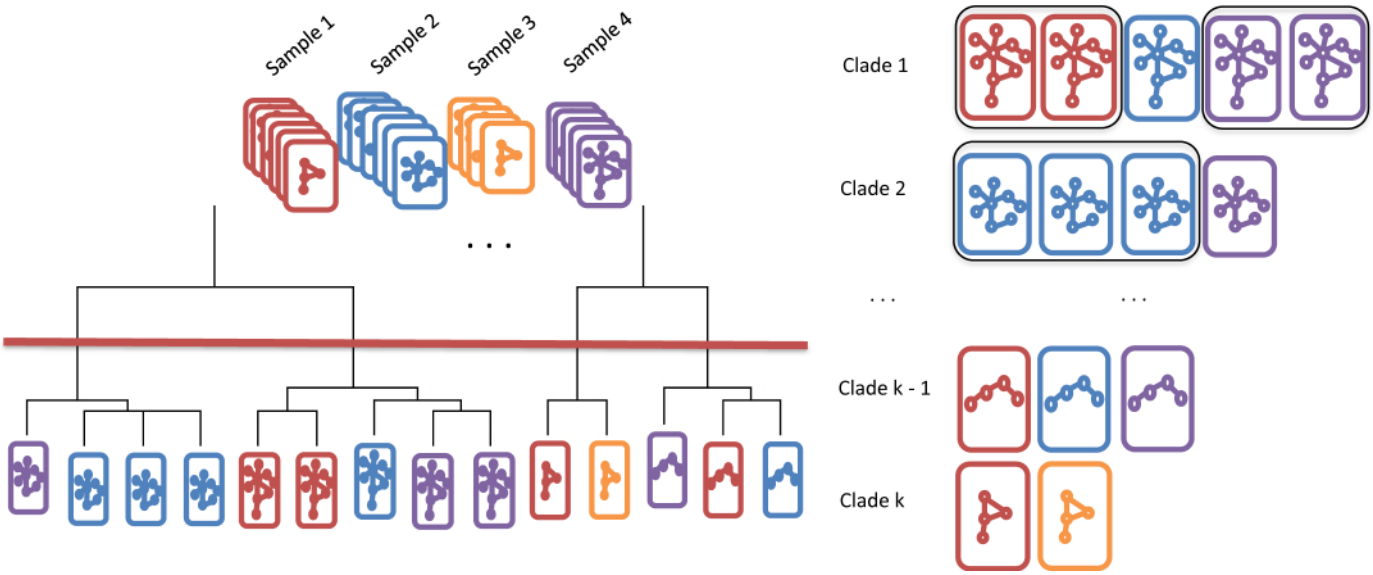

Fig 5. Schematic analogy of MAN. Consider a deck of networks (in analogy to cards), with each "suit" representing a sample and each "rank" representing a unique network structure. The networks are aligned by similarity and organized on a dendrogram. The tree is cut (red line) at the optimal level (by elbow point analysis, see S8 Fig) to output *k* clades. Within each clade, the network structures are similar or the same. If the same sample has multiple networks in the same clade, then these networks are merged (black box around same cards).

https://doi.org/10.1371/journal.pcbi.1005875.g005

by only one edge; these respective networks are clustered together in the dendrogram (Fig 8A right panel; subpopulation indexes are suffixes on the dendrogram). By utilizing the networks, the clades that represent the same and/or similar subpopulations of cells can be established. Clustering by network structures alone resolves the points in the data (Fig 8A, left panel). In contrast, alignment by marker (gene expression) levels cannot resolve the batch effect (Fig 8B).

Next we consider the case with dynamic evolution of subpopulations that models the treatment-control and perturbation studies. The interesting information is in tracking how subpopulations change over the course of the experiment. In the simulation, we have generated two subpopulations that nearly converge in mean expression profile over the time course (Fig 9). The researcher could lose the dynamic information if they were to combine the samples for clustering analysis. As in the previous case, we could use PAC to learn several subpopulations per sample (Fig 10). Then, with the assumption that there are two evolving clusters from data exploration, we align the subpopulations to construct clades of same and/or similar subpopulations (Fig 11 left panel) based on the network structural information (S3 Fig). With network and expression level information in the alignment process, the two subpopulations or clades can be resolved naturally (Fig 11 right panel).

Network and expression alignment is better than network or expression alignment alone

With networks in hand, we could further characterize the relationships between subpopulations across samples. However, the alignment process needs to work well for true linkage to be established. We could align by network alone, by expression (or marker) means, or both. Fig 11 presents these alternatives in comparison. By using all the subpopulation networks, the results still contain subsets of misplaced cells (11 left panel). This is because small clusters of cells have noisy underlying covariance structure; therefore, the networks cannot be accurately inferred. These structural inaccuracies negatively impact the network clustering. The (mean)

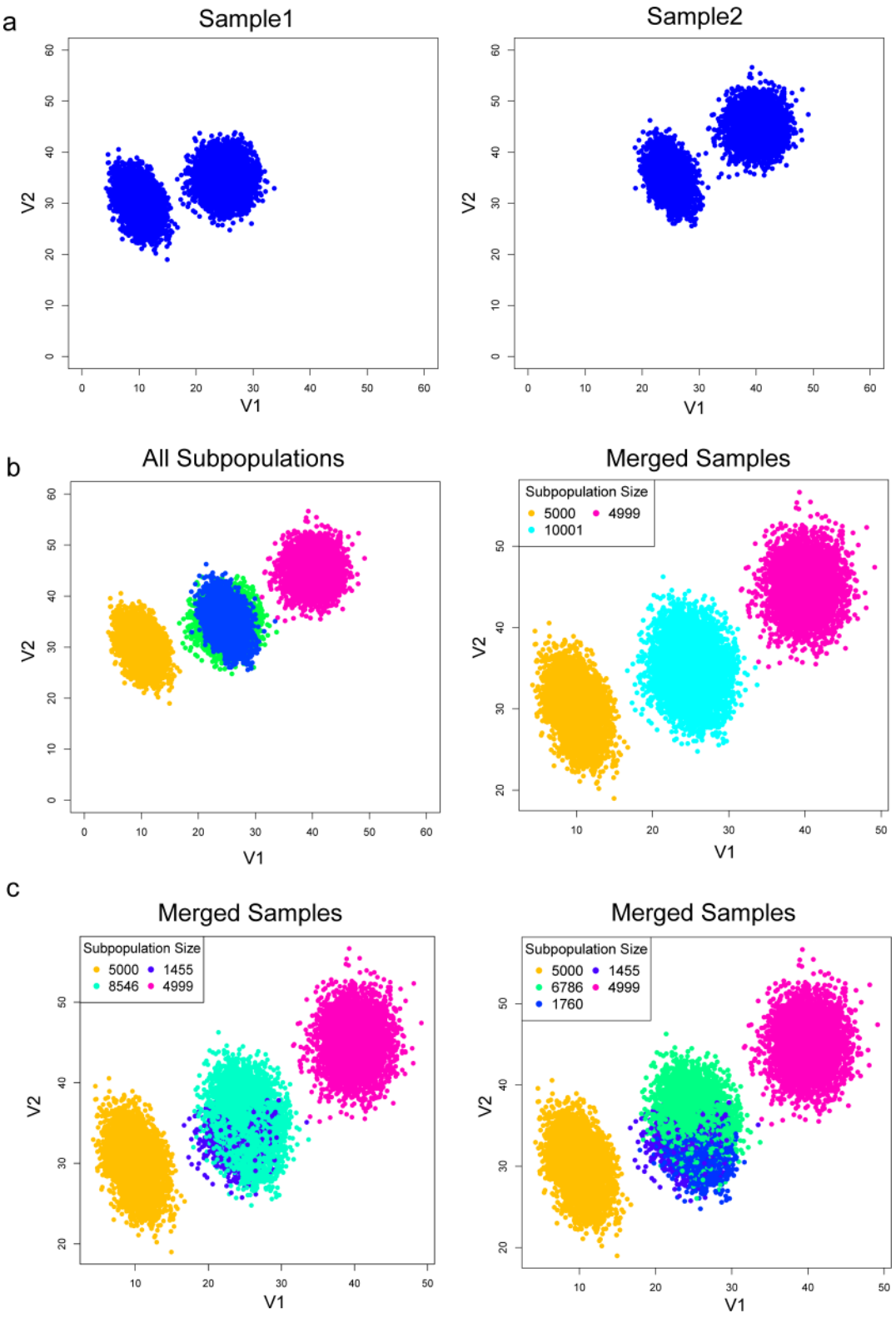

Fig 6. Simple batch effect scenario. A simple batch effect dataset was simulated and visualized. This data has 5 dimensions, with 2 informative dimensions for visualization. (a) Two simulated data samples with the same subpopulations. The means shifted (up in sample 2) due to measurement batch effect. (b) When the samples are combined, as in the case of analyzing/pooling all samples together, two different subpopulations overlap (left panel).

The overlapped subpopulations cannot be distinguished by clustering (right panel). (c) PAC could be used to discover more subpopulations, however, the hints of the present of another subpopulation do not help to resolve the batch effect. Thus, in this case, it is necessary to analyze the samples separately and then find relationships between the subpopulations across the samples.

https://doi.org/10.1371/journal.pcbi.1005875.g006

marker level approach also does not work well (Fig 11 center panel) due to the subpopulation mean shifts across samples. On the other hand, the sequential approach works well (Fig 11 right panel). In the sequential approach, larger (>1000) subpopulations' networks are utilized for the initial alignment process. Next, the smaller subpopulations, which have noisy covariance, are merged with the closest larger, aligned subpopulations. Thus, more subpopulations could be discovered upstream (in PAC), and the network alignment would work similarly as the smaller subpopulations, which could be fragments of a distribution, do not impact the alignment process (S4A and S4B Fig). Moreover, in the network inference step, unimportant edges can negatively impact the alignment process (S4C Fig) in the network-alone case. Biologically, this means that edges that do not constrain or define the cellular state should not be utilized in the alignment of cellular states. Effectively, the threshold placed on the number of edges in the network inference controls for the importance of the edges. Thus, the combined alignment approach works well and allows moderate over-saturation of cellular states to be discovered in the PAC step so that no advance knowledge of the exact number of subpopulations is necessary. It is important to note that we have not utilized high-dimensional mutual information for network structure inference, which is computationally intensive. It may be possible that there exist complex relationships between more than two markers that could yield different network structures for two subpopulations that otherwise would have the same network structure. However, in our analysis of cytometry data, pairwise mutual information with downstream processing yields robust characterization of the cellular state relationships between subpopulations.

PAC-MAN efficiently outputs meaningful data-level subpopulations for mouse tissue dataset

We use the recently published mouse tissue dataset [11] to illustrate the multi-sample data analysis pipeline. The processed dataset contains a total of more than 13 million cell events in 10 different tissue samples, and 39 markers per event (S2 Table). The original research results centered on subpopulations discovered from hand-gating the bone marrow tissue data to find 'landmark' subpopulations; the rest of the data points were clustered to the most similar landmark subpopulations. While this enables the exploration of the overall landscape from the perspective of bone marrow cell types within an acceptable time frame, a significant amount of useful information from the data remains hidden; a larger dataset would make it infeasible to analyze by manual gating and existing computational tools to learn the relationships of the cellular states among all samples. In addition, a natural question is how well do the bone marrow cell types represent the whole immune system?

In contrast to the one-sample perspective, using d-PAC-MAN, the fastest approach by our comparison results, we can perform subpopulation discovery for each sample automatically and then align the subpopulations across samples to establish dataset-level cellular states. On a standard Core i7-44880 3.40GHz PC computer, the single-thread data analysis process with all data points and optimization takes about two hours to complete, which is much faster than alternative methods. With multi-threading and parallel processing, the data analysis procedure can be completed very quickly. As mentioned earlier, PAC results for the bone marrow

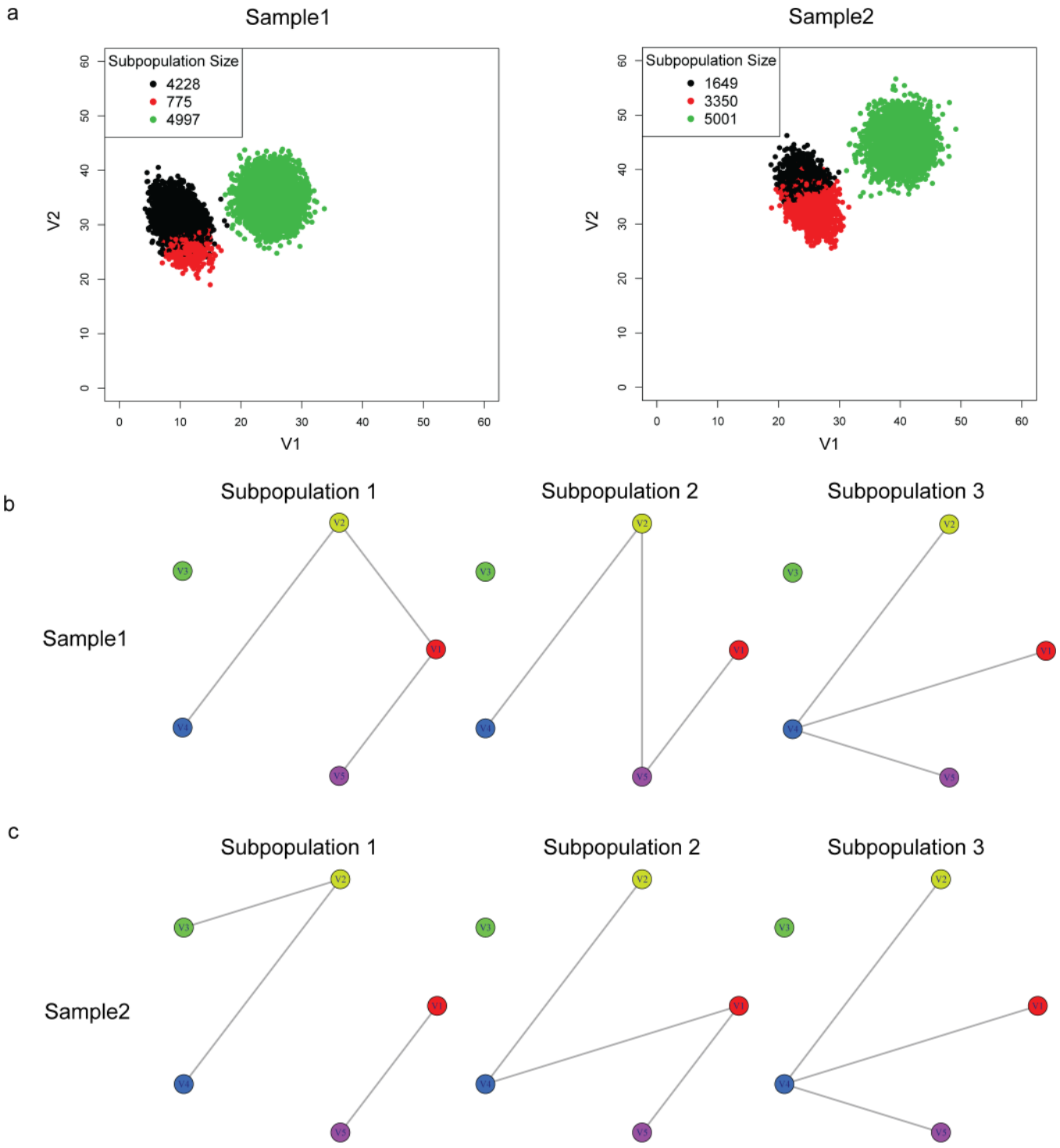

Fig 7. Calculation of sample clusters and their underlying network structures. (a) In the batch effect simulation data, PAC was used to discover several subpopulations per sample without advanced knowledge of the exact number of subpopulations. Here, the colors denote the different clusters within each sample. Panels (b)-(c) show the networks of the subpopulations in both samples 1 and 2, respectively, that are discovered in (a). In these networks, the nodes denote the markers (or genes) measured (in this simulation data, the dimensions are named V1, V2,. . ., V5). The edges denote correlative relationship in terms of mutual information. These networks can be grouped by similarities to organize the subpopulations across samples. In the PAC-MAN implementation, the alignment is based on Jaccard dissimilarity network structure, and we organize the networks with hierarchical clustering of the Jaccard scores.

https://doi.org/10.1371/journal.pcbi.1005875.g007

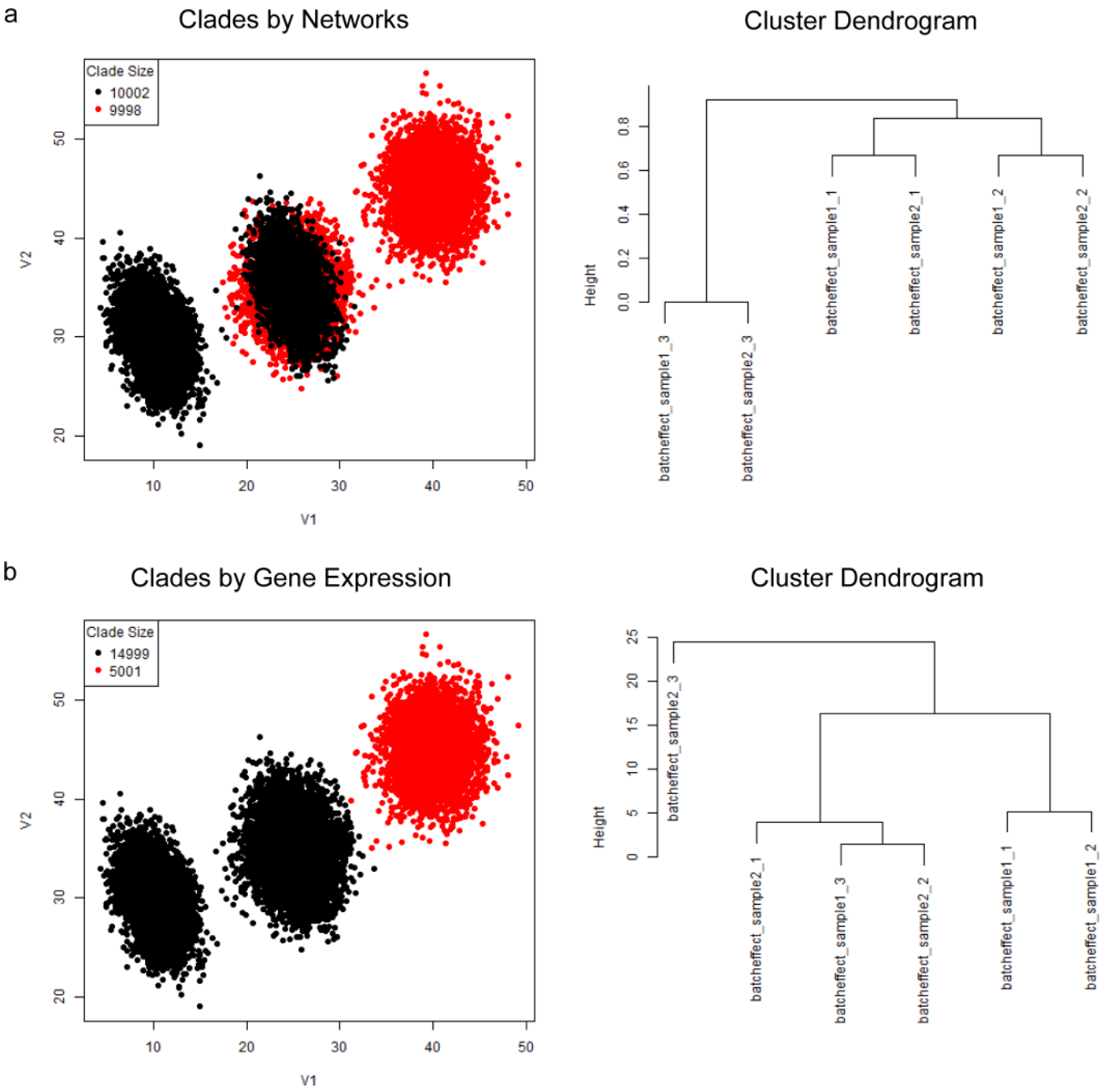

Fig 8. Resolution of batch effects for simple batch effect scenario. Network alignments allow the resolution of mean shift batch effect. (a) Resolution of batch effect by networks of all subpopulations discovered. In the left panel, the colors denote subpopulations that are aligned by network structures. The overlapped subpopulations are correctly labeled. The right panel shows the hierarchical clustering of the subpopulations' networks via Jaccard dissimilarities. These subpopulations are the same as those in Fig 7. (b) Resolution of batch effect by marker levels. Alternative to alignment by network, marker levels (subpopulation centroids) can be used. However, the overlap of the different subpopulations from the two samples makes it impossible to resolve the mean shift in this simulated data. The hierarchical clustering of the centroids organize the subpopulations differently than that in part (a).

https://doi.org/10.1371/journal.pcbi.1005875.g008

subsetted data from this dataset matches closely to that of the hand-gated results. This accuracy provides confidence for applying PAC to the rest of the dataset.

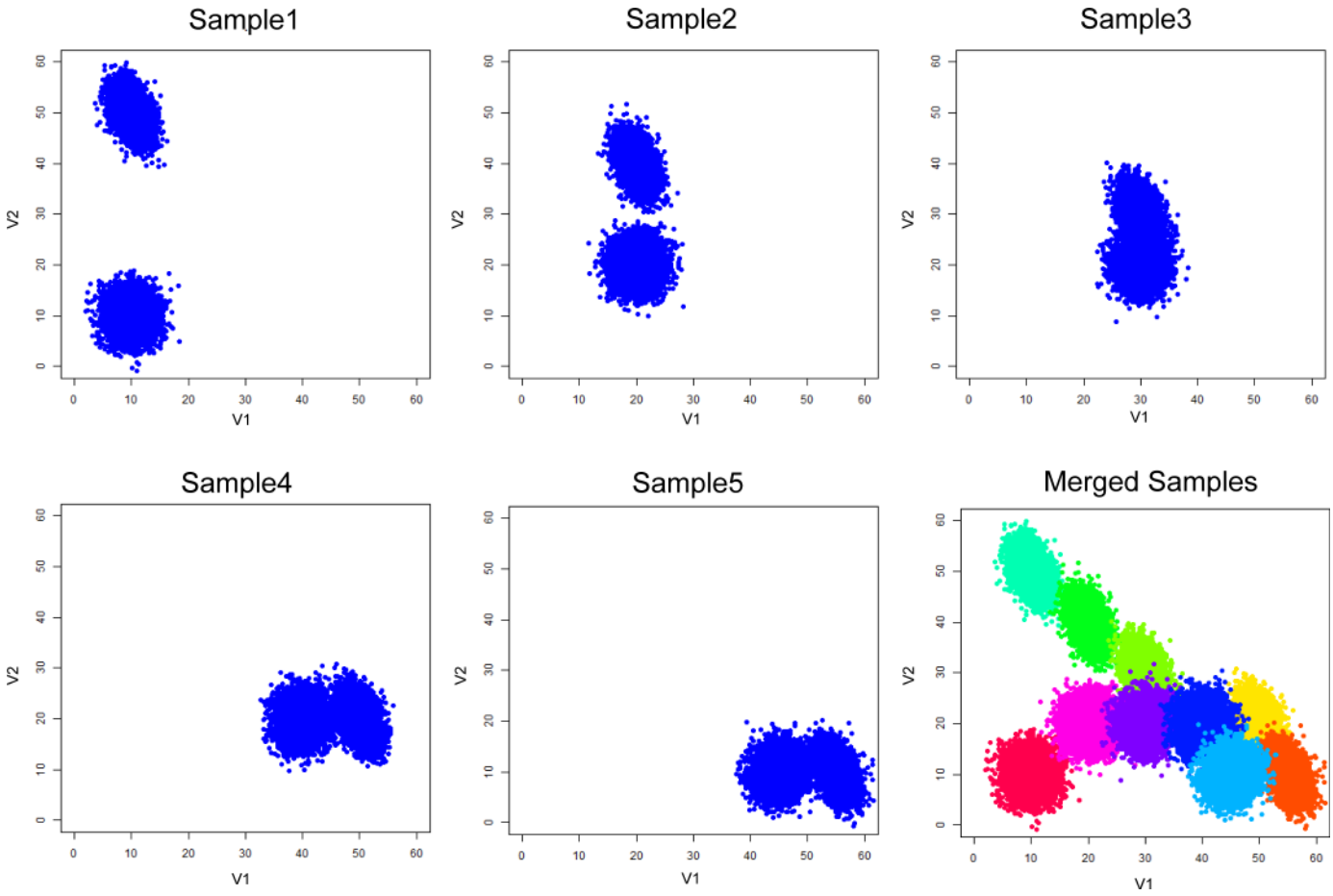

Fig 9. Dynamic batch effect scenario. Two subpopulations, in blue color, migrate in a time series fashion (begins in sample 1, and progresses through samples 2, 3, 4, and 5). In this simulation data, the dimensions are named V1, V2,. . ., V5, and V1 and V2 are the informative dimensions. The two sample subpopulations almost converge by mean shifts through the time series. The bottom right panel shows the subpopulations pooled into one figure; the colors denote subpopulations.

https://doi.org/10.1371/journal.pcbi.1005875.g009

Figs 12 and 13 show the t-SNE plots for subpopulation discovered (top panel of each sample) and the representative subpopulation established (bottom panel of each sample) for the entire dataset. In the PAC discovery step, we learn 50 subpopulations per sample without advance knowledge of how many subpopulations are present. This moderate over-partitioning of the data samples leads to a moderate heterogeneity in the t-SNE plots. From tests, we have found that learning 2–3 times the expected number of subpopulations in the sample works well; it is important to emphasize that aggressive over-partitioning is suboptimal because it creates very small subpopulations that have unstable covariance structures, which removes these small clusters data points from network alignment. Next, the networks are inferred for the larger subpopulations (with number of cell events greater than 1000), and the networks are aligned for all the tissue samples. To choose the optimal number of total subpopulations to output, we perform the elbow point test at this step, in which we calculated the within cluster standard deviations while varying the number of subpopulations outputted for the entire dataset. The elbow point rests at 130 clusters (S8 Fig), and we outputted 130 representative subpopulations, also called clades, for the entire dataset to account for the traditional immunological cellular states and sample-specific cellular states present. Within samples, the subpopulations that cluster together by network structure are aggregated. The smaller subpopulations (<1,000

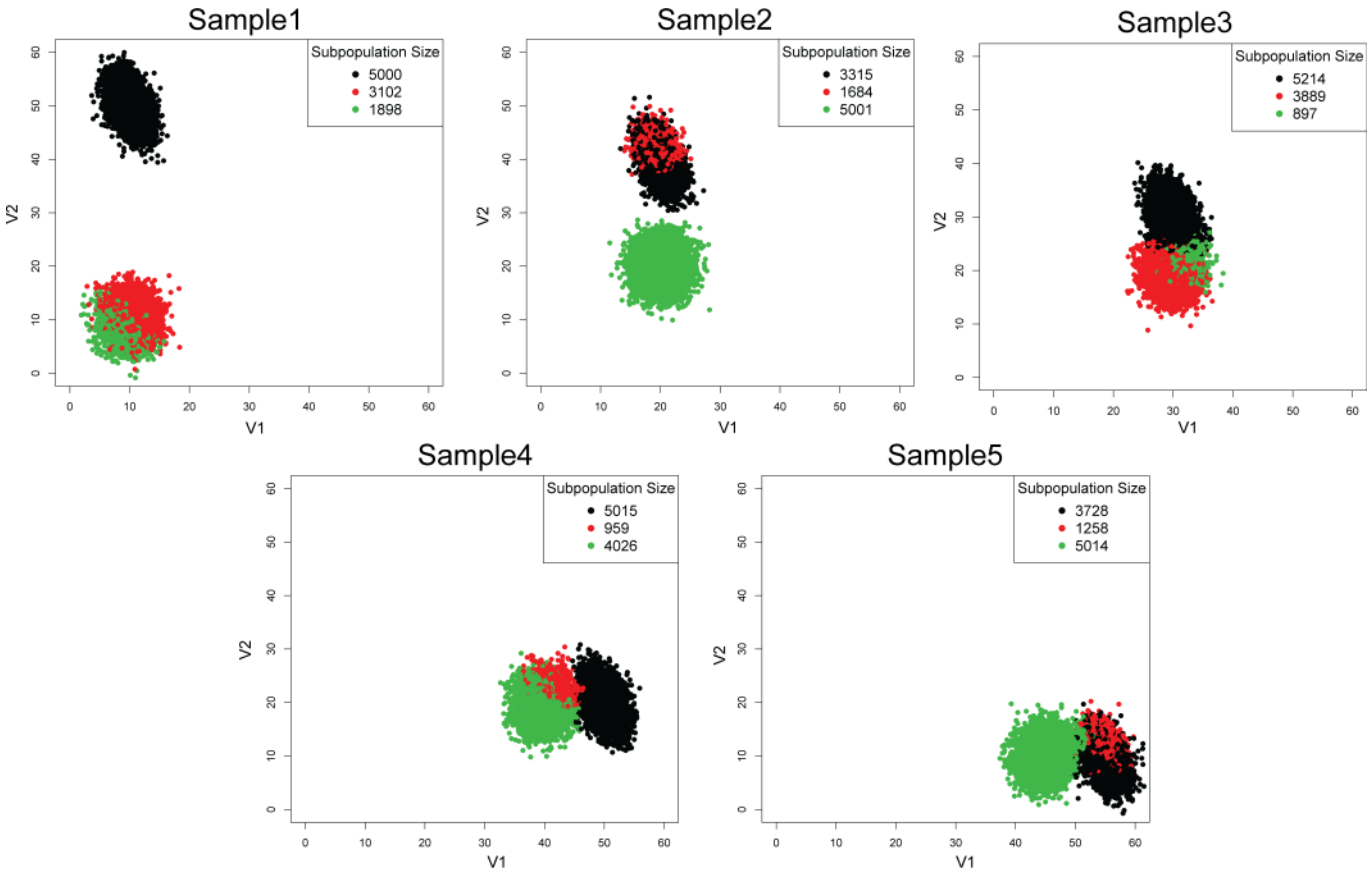

Fig 10. PAC clustering on dynamic batch effect scenario samples. We used PAC to discover several subpopulations per sample without advanced knowledge of the number of subpopulations present. The colors within each sample denote a distinct PAC subpopulation, but the colors have no meaning across samples.

https://doi.org/10.1371/journal.pcbi.1005875.g010

cells each, not involved in network alignment) are either merged to the closest larger subpopulation or establish their own sample-specific subpopulation by expression alignment. We attempt to assign these very small subpopulations back with larger clades by grouping all subpopulations within each sample into 5 expression-level clusters (using cluster centroids), and thus we kept the larger subpopulations and a maximum of 4 minor sample-specific subpopulations for each tissue sample. Subpopulations with less than 100 cell events were discarded. The representative subpopulations (143 total including sample-specific minor subpopulations) follow the approximate distribution of the cell events on the t-SNE plots and the aggregating effect cleans up the heterogeneities due to over-partitioning in the PAC step.

The cell type clades are the representative subpopulations for the entire dataset, and they could either be present across samples or in one sample alone. Their distribution is visualized by a heatmap (Fig 14). While the bone marrow sample contains many cell types, only a subset of them are directly aligned to cell types in other samples, which means using the bone marrow data as the reference point leaves much information unlocked in the dataset. Therefore, the data suggests that the bone marrow cell types are not adequate in representing all cell types in the immune system. The cell types in the blood and spleen samples have various alignments with cell types in other samples. The lymph node samples share many clades likely due to the connection through the lymphatic vessels; the small intestine and colon samples also share many clades, probably due to closeness in location and biological function. Nevertheless,

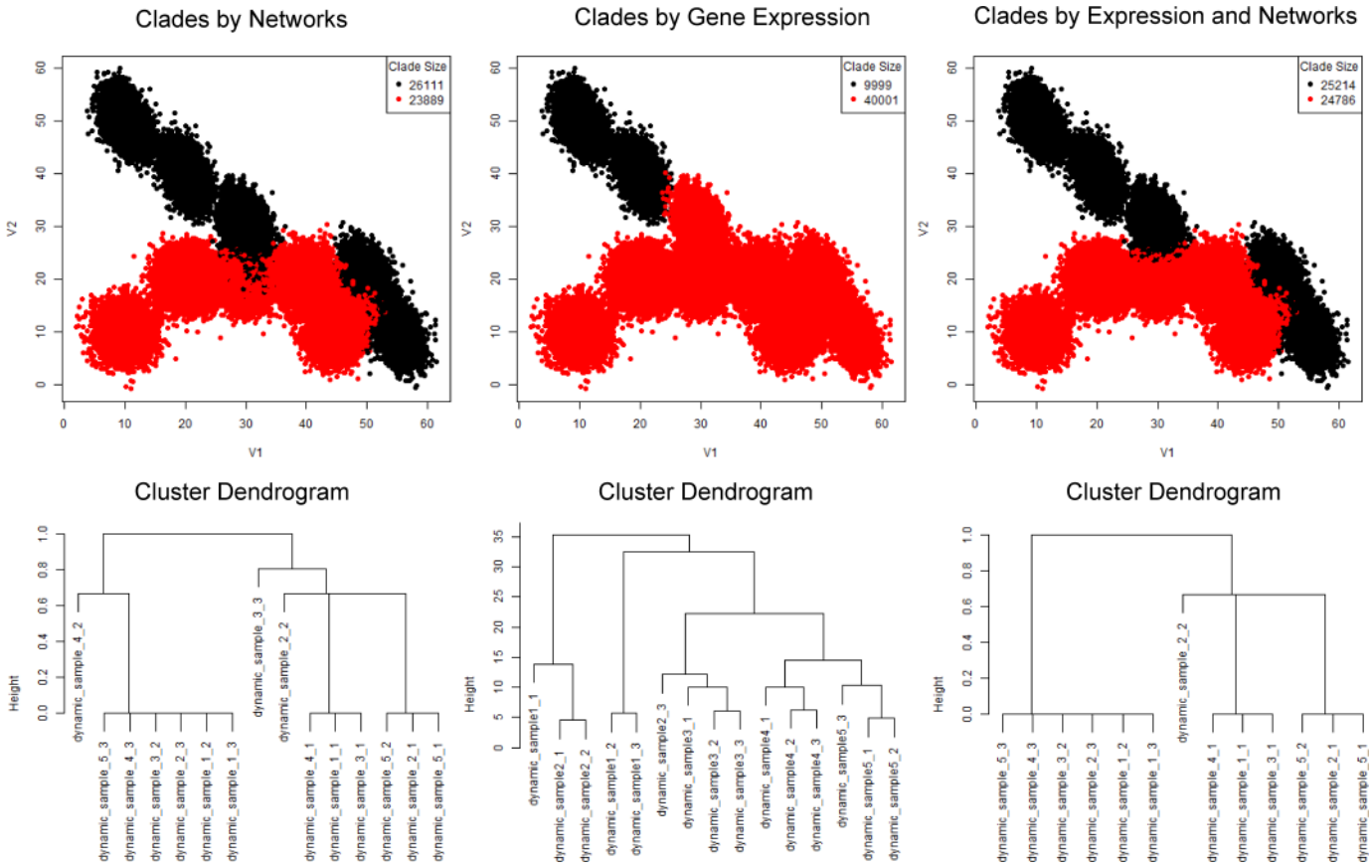

Fig 11. Resolution of dynamic batch effects scenario. Comparison of PAC-MAN results between representative clades (number of clades set to 2). Using network structures (left panel) or expression information (middle panel) alone does not resolve the dynamic information. On the other hand, the dynamic information is resolved first by alignments of networks of larger subpopulations and then by merging smaller subpopulations (with unstable network structures) by expression into the aligned clades (right panel).

https://doi.org/10.1371/journal.pcbi.1005875.g011

the results show that the tissue samples do not share exactly the same clades, suggesting that the immune system cells have different states in different organs. On the other hand, the thymus sample has few clades shared with other samples, which may be due to its functional specificity.

PAC-MAN style analysis can be applied to align the tissue subpopulations by their means instead of network similarities (S5 Fig). As done previously, 143 overall representative clades (130 network clades + 13 minor sample-specific subpopulations) were outputted. The same aggregating effect is observed (S5A Fig), and this is due to the organization from dataset-level variation in the means. Comparing to the network alignment, the means linkage approach has more subpopulations per sample; the subpopulation proportion heatmap (S5B Fig) shows more linking. Although the bone marrow sample subpopulations co-occur in the same clades slightly more with other sample subpopulations, this sample does not co-occur with many clades in the dataset. Thus, a PAC-MAN style analysis with means linkage also harvests additional information from the entire dataset.

In general, the means alignment approach gives many more clades per sample than that of the network alignment PAC-MAN approach. In fact, the network approach has 88 linkages while the means approach has 270 linkages. The linkage plot (S6A Fig) shows that the low linkages occur slightly more frequently for the network approach. One consequence is that the network approach aggregates PAC subpopulations within sample more frequently; for instance,

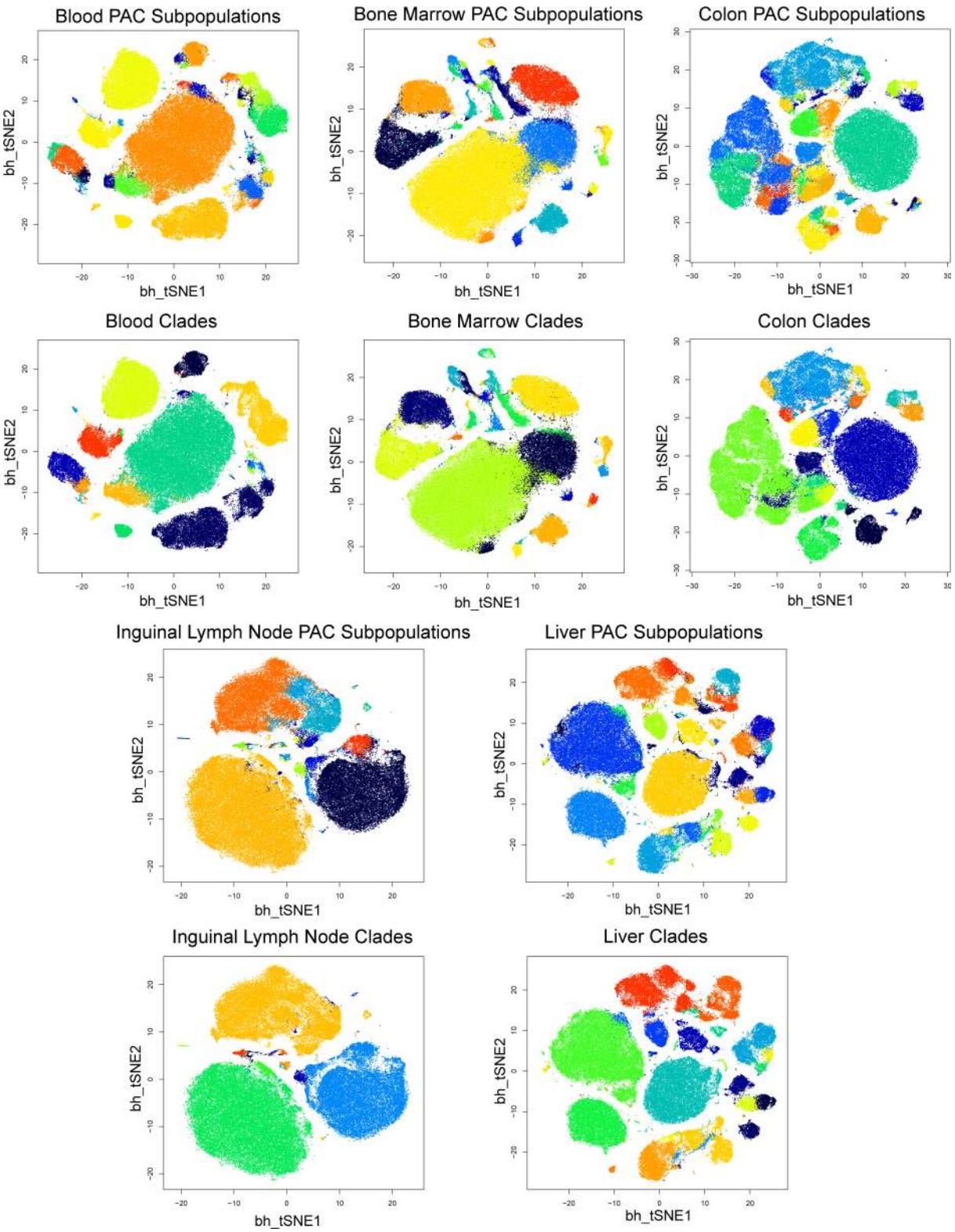

Fig 12. Visualization of PAC vs. PAC-MAN results for blood, bone marrow, colon, inguinal lymph node, and liver samples. The PAC (explorative clustering) and PAC-MAN (data-level cellular states) results are presented for each sample in column-wise fashion. Each tissue sample's t-SNE plots were generated using 100,000 randomly drawn cell events for that sample. The results from PAC (top panel) and PAC-MAN (bottom panel) steps are presented in pairs. Initial PAC discovery was set to 50 subpopulations

without advanced knowledge of the number of subpopulations in each sample. In MAN, 130 network clades (optimal number from elbow point analysis) were outputted, and the cellular states are defined by expression (marker signal), network structure, and dataset-level variation. This composite definition of cellular state naturally aggregates the PAC clusters to yield smaller number of subpopulations in less variable samples. S11 Fig is a higher resolution version of Fig 12 with subpopulation and clade labels.

https://doi.org/10.1371/journal.pcbi.1005875.g012

in the thymus sample, the network approach yields 13 clades (and 2 minor sample-specific subpopulations) while the means approach yields 39 clades.

After aggregating, the clade sizes (with unique participants per sample) are plotted (S6B Fig). The network approach tends to find fewer linkages, as more clades have sizes of less than 4, while the means approach has more clades than the network approach with clade sizes greater than 4. The network approach is more conservative due to the additional constraints from network structures. Conventionally, in the cytometry field, only the means are considered in the definition of cellular states. The network alignment is more stringent in the establishment of linkages; the network PAC-MAN approach defines cellular states with the additional information from network structures, and it has the effect of constraining the number of linkages between samples while finding linkages for subpopulations that are distant in their means.

Further studies are needed to combine the information from both the marker level and network structures to organize the cellular states discovered in cytometry datasets, for example, through a weighted score based on the means and network alignments. In this study, we demonstrated that the covariance and network structures built from subpopulations are valuable and can be utilized to organize data-level cellular state relationships.

Network hubs provide useful annotations

To further characterize the cell types, we annotate the clades within each sample using the top network hub markers, which constrain the cellular states. The full network structure annotation, along with average expression profiles, is presented in S3 Table. The clade information is presented in the ClusterID column. The annotations for cells across different samples but within the same clades share hub markers. For example, in clade 1 for the blood and bone marrow samples, the cells share the hub markers Ly6C and CD11b. In the bone marrow sample, one important set of subpopulations is the hematopoietic stem cell subpopulations. One such subpopulation is present as clade 33 with the annotation F4/80.CD16/32.Sca1.cKit and is about 1.18 percent in the bone marrow sample. Clade 33 is only present in the bone marrow sample, indicating that the PAC-MAN pipeline defines this as a sample-specific and coherent subpopulation using dataset-level variation. The thymus contains a large subpopulation clade 124 (84.07 percent) that is characterized as CD5.CD43.CD3.CD4, suggesting it to be the maturing T-cell subpopulation.

Constellation plot combines clade and signal information

PAC-MAN generates both the clade and subpopulation signal (or expression) information. Fig 14 visualizes the occurrence and proportions of representative subpopulations in the dataset. To understand the expression levels of the markers for the subpopulation, a heatmap is constructed (Fig 15 and S14 Fig). In high-dimension, the subpopulations can form regions in which similar cellular states are next to each other. Do subpopulations belonging to the same clade occupy the same region? In addition, what is the spatial spread of subpopulations belonging to the same clade? To visualize the clade relationships between subpopulations in the dataset, we construct the constellation plot (Fig 16). First, the centroids of the discovered subpopulations are inputted into a t-SNE visualization processing, which projects and separates the centroids onto a 2D

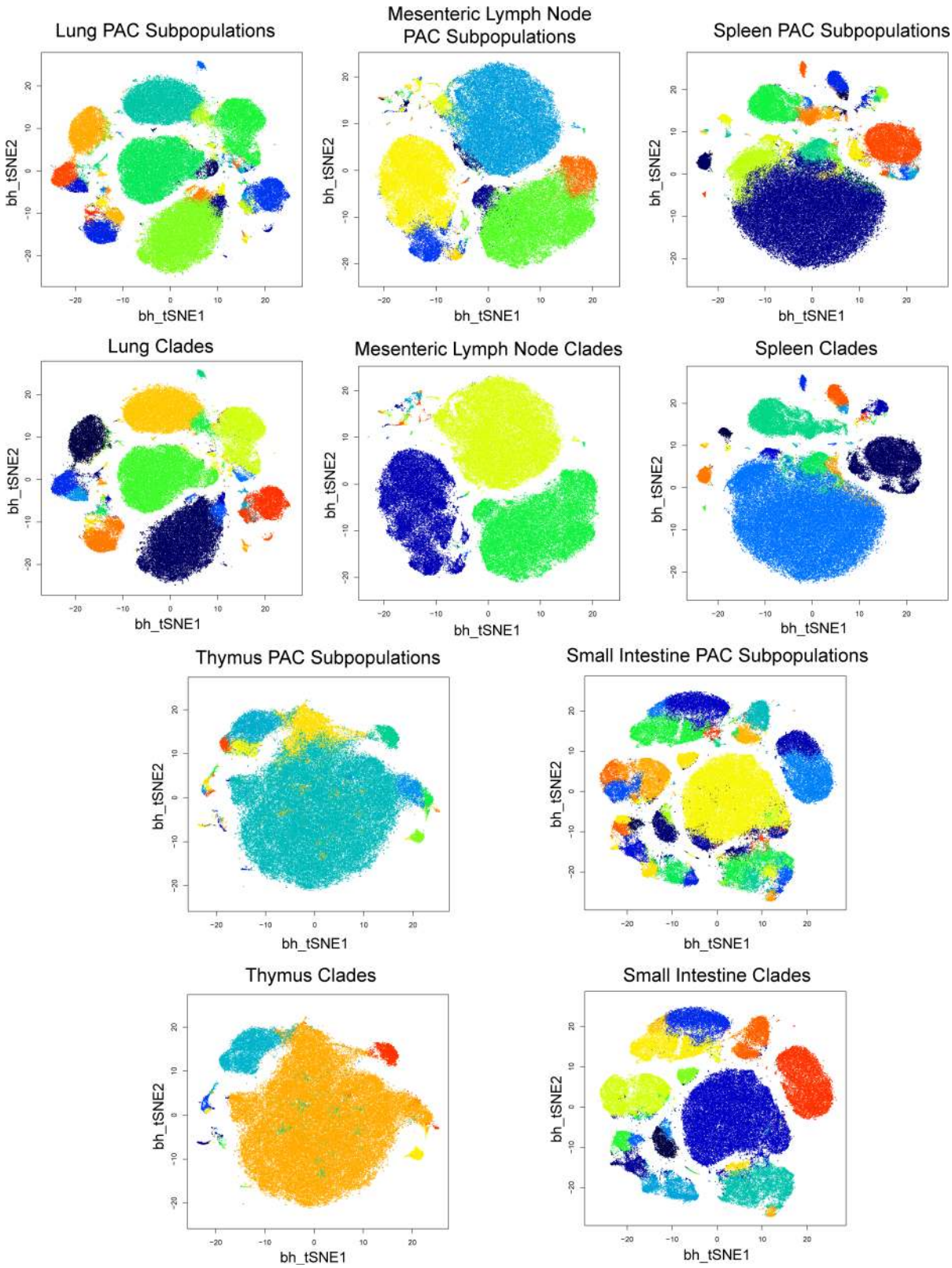

Fig 13. Visualization of PAC vs. PAC-MAN results for lung, mesenteric lymph node, spleen, thymus, and small intestine samples. The settings and descriptions are the same as those in Fig 12. Continuation of visualization of PAC-MAN results for the mouse tissue data. S12 Fig is a higher resolution version of Fig 13 with subpopulation and clade labels.

https://doi.org/10.1371/journal.pcbi.1005875.g013

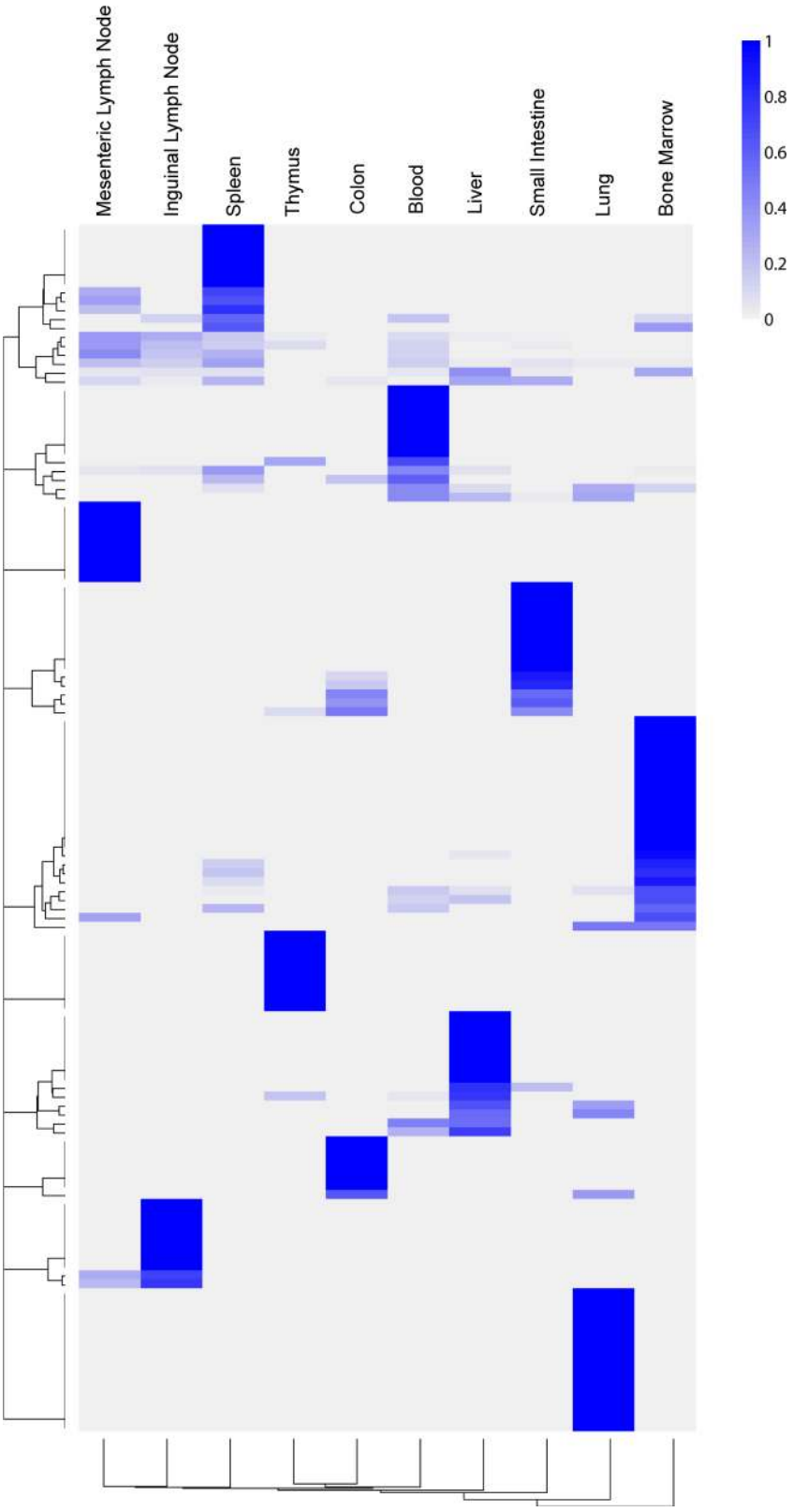

Fig 14. Heatmap of clade proportions across the tissue samples. Sample-specific clades have a value of 1, while shared clades have proportions spread across different samples. Physiologically similar samples share more clades. S13 Fig is a higher resolution version of Fig 14 with clade labels.

https://doi.org/10.1371/journal.pcbi.1005875.g014

plane. Next, the clades are color-coded such that 1) grey color indicates sample-specific clade and 2) non-grey colors indicate clades with multiple sample representation. Finally, we group the subpopulations in each clade by drawing lines to connect the closest clade subpopulation on the 2D plane, analogous to the visualization of stars by constellation nomenclature.

The constellation plot is useful in looking at the spread of the clades in relation to other subpopulations. For example, clade 10, which contains subpopulations that are CD45+CD3+CD5+CD8+, and clade 8, which contains subpopulations that are CD45+CD3+CD5+CD4+, are T cells (S14 Fig); these two clade groups exist next to each other in the constellation plot, but they do not overlap. Clade 2 is in a region that contains CD45+CD19+B220+ subpopulations, which signify B cells. Furthermore, within each clade, the subpopulation networks are similar and contain similar hub genes. For instances, clades 2 and 8 represent data-level subsets of T cells and B cells, respectively; clades 2 and 8's networks are presented in Figs 17 and 18. Each clade has its unique network structures and a set of hub markers. Overall, in this analysis, we observe that clades defined by signal levels and network structures tend to occupy defined regions in high-dimensional space. Certainly, not all cell types are present in all tissue samples, and those immune cell subsets that are similar enough to be in the same clade may differ due to their tissue-specific, local environmental factors.

Conclusion

We have presented the PAC-MAN data analysis pipeline. This pipeline was designed to remove major roadblocks in the utilization of existing and future CyTOF datasets. First, we established a quick and accurate clustering method that closely matches expert gating results; second, we demonstrated the management of multiple samples by handling mean shifts and batch effects across samples. We demonstrated that the inter-marker relationship in the form of mutual information networks is extremely useful in defining cellular states. The alignment of network structures allows researchers to find relationships between cells across samples without resorting to pooling of all data points. PAC-MAN allows the cytometry field to harvest information from the increasing amount of CyTOF data available. It is important to standardize multi-sample data analysis with automation so that discoveries based on multi-sample CyTOF datasets from different laboratories do not depend on the experts' manual gating strategies and the grouping of subpopulations that is constrained by non-systematic computations. Furthermore, due to PAC-MAN's generality, this pipeline can be utilized to analyze large datasets of high-dimension beyond the cytometry field.

Materials and methods

Partition-assisted clustering has two parts

1. Partitioning: a partition method (BSP[5] or DSP[7]) is used to learn N initial cluster centers from the original data.
2. Post-processing: A small number (m) of k-mean iterations is applied to the rectangle-based clusters from the partitioning, where m is a user-specified number. We used m = 50 in our examples. After this k-means refinement, we merge the N clusters hierarchically until the desired number of clusters (this number is user-specified) is reached. The merging is based on a given distance metric for clusters. In the current implementation, we use the same distant metric as in flowMeans[1]. That is, for two clusters X and Y, their distance D(X,Y) is defined as:

$$D(X, Y) = \min\{(\bar{x} - \bar{y})^T S_x^{-1}(\bar{x} - \bar{y}), (\bar{x} - \bar{y})^T S_y^{-1}(\bar{x} - \bar{y})\} \quad (1)$$

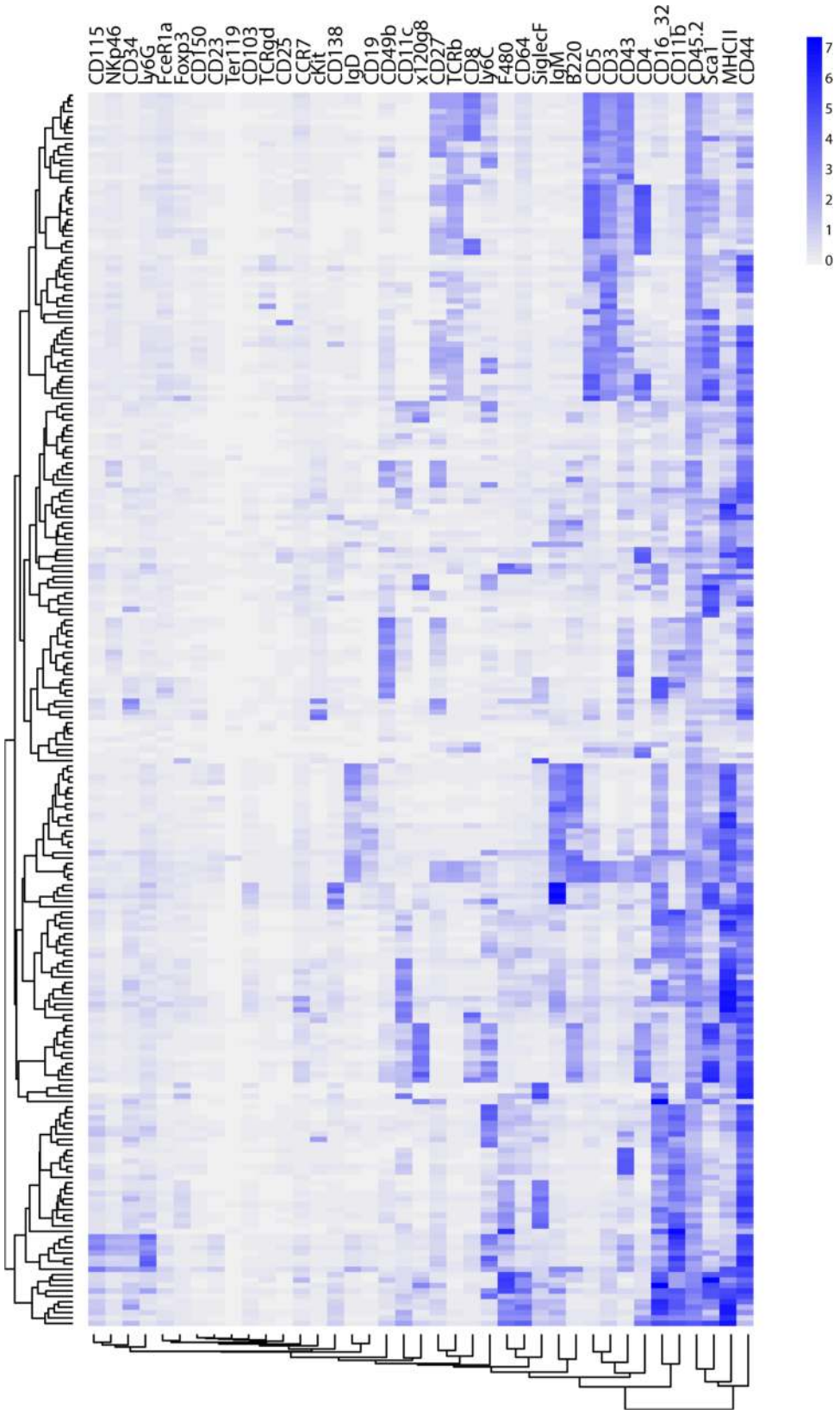
CD115
NKp46
CD34
Ly6G
FceR1a
Foxp3
CD150
CD23
Ter119
CD103
TCRgd
CD25
CCR7
cKit
CD138
IgD
CD19
CD49b
CD11c
x120g8
CD27
TCRb
CD8
Ly6C
F480
CD64
SiglecF
IgM
B220
CD5
CD3
CD43
CD4
CD16_32
CD11b
CD45.2
Sca1
MHCII
CD44
7
6
5
4
3
2
1
0

Fig 15. Heatmap of average subpopulation expression levels in all tissue samples. The expression heatmap illustrates the average expression of PAC-MAN-discovered subpopulations. The subpopulations are grouped by hierarchical clustering, and subpopulations close in expression space are organized into blocks. S14 Fig is a higher resolution version of Fig 15 with clade labels.

https://doi.org/10.1371/journal.pcbi.1005875.g015

where $\bar{x}, \bar{y}$ are the sample mean of cluster X and Y, respectively. S_x^{-1} is the inverse of the sample covariance matrix of cluster X. S_y^{-1} is defined similarly. In each step of the merging process, the two clusters having the smallest pairwise distance will be merged together into one cluster.

Partition methods

There are two partition methods implemented in the comparison study: d-PAC and b-PAC. The results are similar, with d-PAC being the faster algorithm. Fig 1A illustrates this recursive process.

d-PAC is based on the discrepancy density estimation (DSP)[7]. Discrepancy, which is widely used in the analysis of Quasi-Monte Carlo methods, is a metric for the uniformity of points within a rectangle. DSP partitions the density space recursively until the uniformity of points within each rectangle is higher than some pre-specified threshold. The dimension and the cut point of each partition are chosen to approximately maximize the gap in uniformity of two adjacent rectangles.

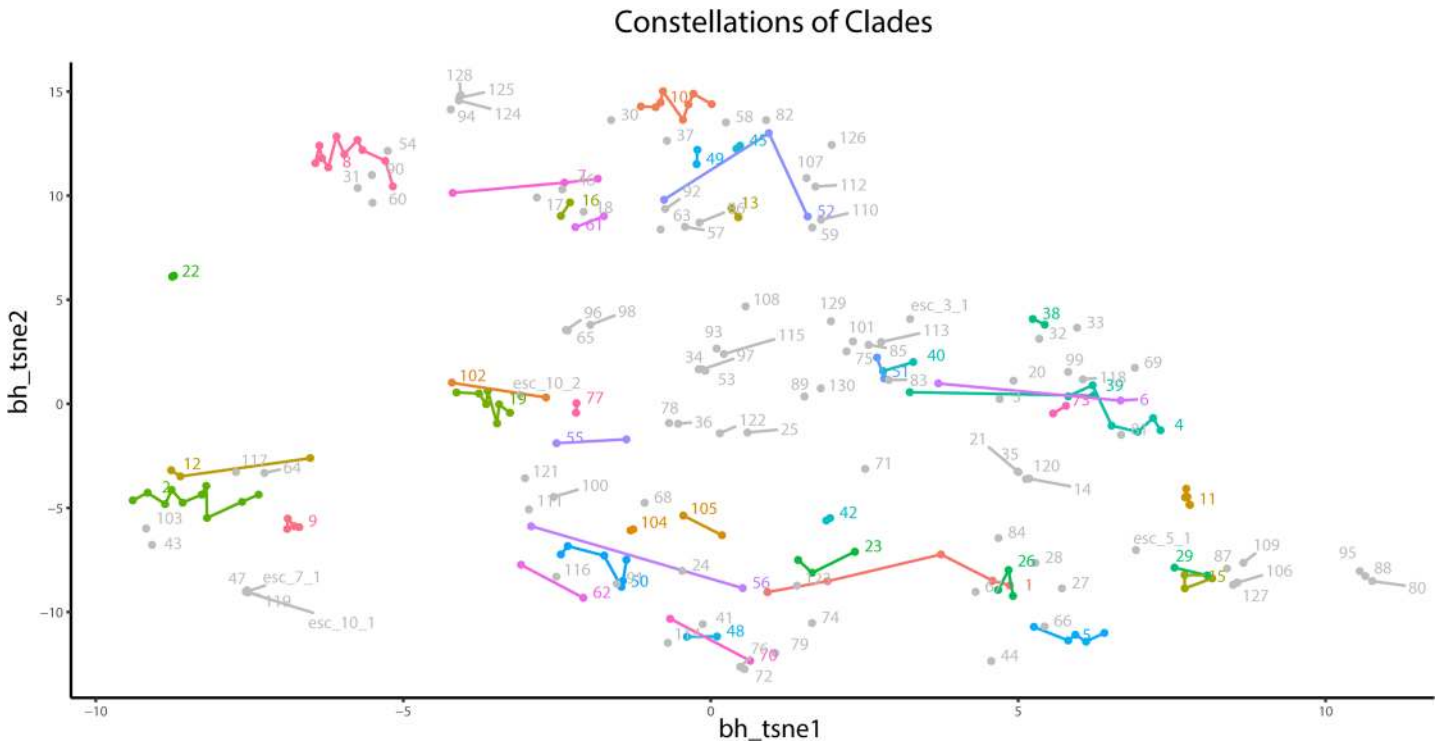

Fig 16. Constellation plot of clades. The constellation plot is designed to visualize both the expression and clade information of discovered subpopulations. Here, the centroids (average expression) of PAC-MAN-discovered subpopulations in the example tissue dataset are projected onto a t-SNE 2D space. Clades that only occur in one sample are colored grey. The non-grey clades occur in at least two samples, with unique colors and clade identification denoting each clade. On the constellation plot, the closest multi-sample clade subpopulations are connected by a straight line.

https://doi.org/10.1371/journal.pcbi.1005875.g016

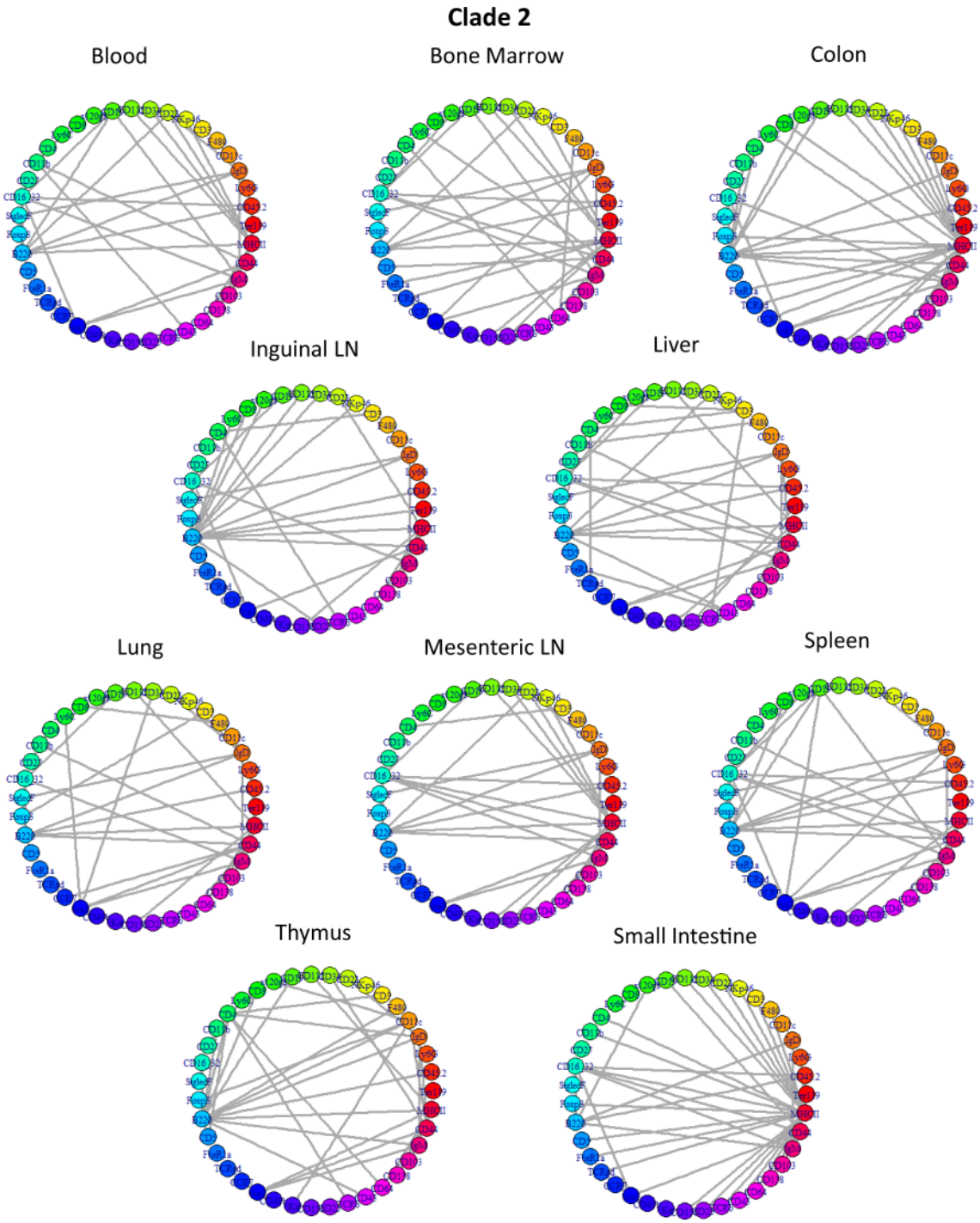

Fig 17. Network structures of Clade 2: B cells. In each network figure, the markers are denoted by nodes of different colors. These networks show the top edges that define the network structures for subpopulations in clade 2. The subpopulation network structures for each subpopulation in clade 2 show that certain markers, such as B220 and MHCII, are hubs across most, if not all, networks in this clade. This hub combination is consistent and unique to clade 2 (see Fig 18).

https://doi.org/10.1371/journal.pcbi.1005875.g017

BSP + LL is an approximation inference algorithm for Bayesian sequential partitioning density estimation (BSP)[5]. It borrows ideas from Limited-Look-ahead Optional Pólya Tree

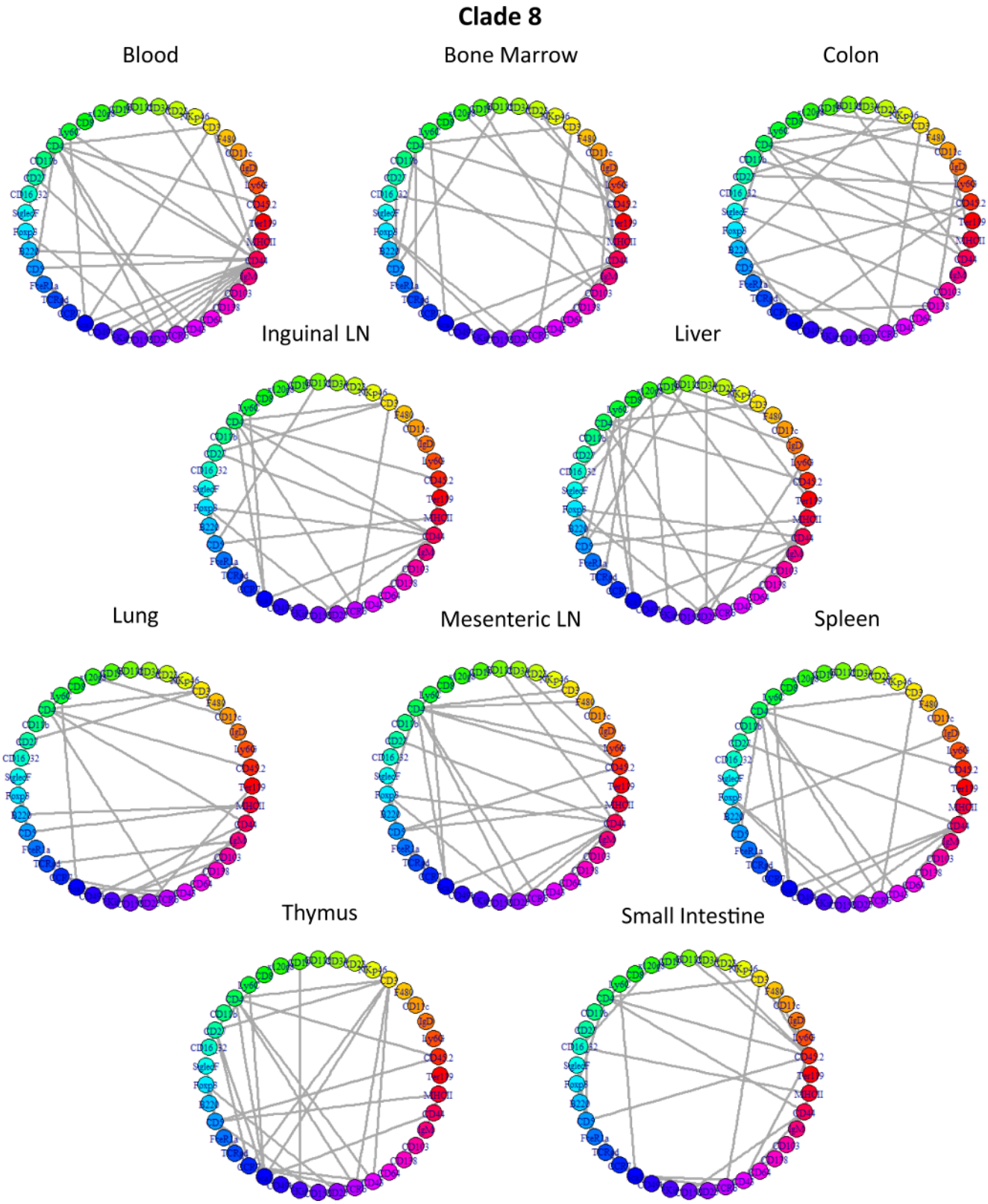

Fig 18. Network structures of Clade 8: CD4+ T cells. The set up is the same as in Fig 17. The subpopulation network structures for each subpopulation in clade 8 show that certain markers, such as CD3 and CD4, are hubs across most, if not all, networks in this clade. This hub combination is consistent and unique to clade 8.

https://doi.org/10.1371/journal.pcbi.1005875.g018

(LL-OPT), an approximate inference algorithm for Optional Pólya Tree[6]. The original inference algorithm for BSP looks at one level ahead (i.e. looking at the possible cut points one level deeper) when computing the sampling probability for the next partition. It then uses resampling to prune away bad samples. Instead of looking at one level ahead, BSP + LL looks at h levels ahead ($h > 1$) when computing the sampling probabilities for the next partition and does not do resampling (Fig 1B). In other words, it compensates the loss from not performing resampling with more accurate sampling probabilities. For simplicity, 'BSP + LL' is shortened to 'BSP' in the rest of the article.

F-measure

We use the F-measure for comparison of clustering results to ground truth (known in simulated data, or provided by hand-gating in real data). This measure is computed by regarding a clustering result as a series of decisions, one for each pair of data points. A true positive decision assigns two points that are in the same class (i.e. same class according to ground truth) to the same cluster, while a true negative decision assigns two points in different classes to different clusters. The F-measure is defined as the harmonic mean of the precision and recall. Precision P and recall R are defined as:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{2}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{3}$$

where TP is the total number of true positives, FP is the total number of false positives and FN is the total number of false negatives.

F-measure ranges from 0 to 1. The higher the measure, the more similar the estimated cluster result is to the ground truth. This definition of F-measure is different than that of FlowCAP challenge[2]. The use of co-assignment of labels in this definition is a more accurate way to compute the true positives and negatives.

Purity-measure (p-measure)

Most of the existing measurements for clustering accuracy aim at measuring the overall accuracy of the entire datasets, i.e. comparing with the ground truth over all clusters. However, we are also interested in analyzing how well a clustering result matches the ground truth within a certain class. Specifically, consider a population with K classes in the ground truth: $\{C_1, C_2, \ldots, C_K\}$. We construct a class-specific index called the purity measure, or p-measure for short, to measure how well our clustering result matches the ground truth. This index is computed as follows:

1. For each class C_k, look for the cluster that has the maximum number of overlapping points with this class, denoted by L_{i_k}.
2. Define

$$\mathrm{S}_1 = \frac{|C_k \cap \mathrm{L}_{i_k}|}{|\mathrm{L}_{i_k}|}, \ \mathrm{S}_2 = \frac{|C_k \cap \mathrm{L}_{i_k}|}{|C_k|} \tag{4}$$

where $|\cdot|$ denotes the number of points in a set.

3. The final P-index for class C_k is given by

$$P = \frac{2S_1S_2}{S_1 + S_2} \tag{5}$$

If we were to match a big cluster with a small class, even though the overlapping may be large, S_1 would still be low since we have divided the score by the size of the cluster in S_1. In addition, we are interested in knowing how many points in C_k are clustered together by L_{i_k}, which is measured by S_2.

Network construction and comparison

After PAC, the discovered subpopulations typically have enough cells for the estimation of mutual information. This enables the construction of networks as the basis for cell type characterization. In these networks, the nodes represent the markers monitored in the experiment, while the edges represent a correlation/mutual information dependence relationship between the marker levels. Computationally, it is not good to directly use the mutual information networks constructed this way to organize the subpopulations downstream. The distance measure used to characterize the networks could potentially give the same score for different network structures. Thus, it is necessary to threshold the network edges based on the strength of mutual information to filter out the noisy and miscellaneous edges. In this work, these subpopulation-specific networks are constructed using the MRNET network inference algorithm in the Parmigene [15] R package. The algorithm is based on mutual information ranking, and outputs significant edges connecting the markers. The top d edges (d is set to be 1x the number of markers in all examples) are used to define a network for the subpopulation. This process enables a careful calculation of the distance measure.

For each pair of subpopulation networks, we calculate a network distance, which is defined as follows. If G_1 and G_2 are two networks, let S be the set of shared edges and A be union of the of the edges in the two networks, then we define

$$\text{Similarity}(G_1, G_2) = \frac{|S|}{|A|} \tag{6}$$

where $|\cdot|$ denotes the size of a set.

This is known as the Jaccard coefficient of the two graphs. The Jaccard distance, or 1- Jaccard coefficient, is then obtained. This is a representation of the dissimilarity between each pair of networks; the Jaccard dissimilarity is the measure used for the downstream hierarchical clustering.

Cross-sample linkage of subpopulations

We perform agglomerative clustering of the pool of subpopulations from all samples. This clustering procedure greedily links networks that are the closest in Jaccard dissimilarity, and yields a dendrogram describing the distance relationship between all the subpopulations. We cut the dendrogram to obtain the k clades of subpopulations. Subpopulations from the same sample and falling into the same clade are then merged into a single subpopulation (Fig 5). This merging step has the effect of consolidating the moderate over-partitioning in the PAC step. No merging is performed for subpopulations from different samples sharing the same clade. In this way, we obtain k clades of subpopulations, with each clade containing no more

than one subpopulation from each sample. We regard the subpopulations within each clade as being linked across samples.

In the above computation, only subpopulations with enough cells to define a stable covariance are used for network alignment via the Jaccard distance; the rest of the cell events from very small subpopulations are then merged with the closet clade by marker profile via distance of mean marker signals. If the small subpopulations are distant from the defined clades, then a new sample-specific clade is created for these small subpopulations.

Elbow point analysis of optimal number of clades

To efficiently find the practical number of clades to output for PAC-MAN, we utilize the elbow point analysis approach. Initially in the PAC step, the sample points are clustered into 2–3 times the expected number of sample subpopulations expected by the researcher. Next, we calculate the within-cluster errors, or distance from the subpopulation centroid, for each cluster in all samples, and we obtain the within-cluster errors for all sample. This calculation is performed for a range of numbers of clades in MAN. Loess smoothing is applied to the average within-cluster errors over the numbers of clades, and the researcher determines the location of the elbow point, which is then inputted into the final network alignment.

Constellation plot analysis

To visualize the cellular state distribution in high-dimension, we construct the constellation plot. On the constellation plot, we observe two layers of information: the distribution of the clusters by expression level projection and the network similarities. By building the network structures and performing structural alignments, we remove extraneous connectivity between subpopulations that may appear close together in 'expression space' by grouping only subpopulations with strong network structural similarity. Those subpopulations that are in different clades but are close together on the constellation plot can be sample-specific subpopulations worth validating by future sorting and characterization experiments; these subpopulations are coherent clusters by expression and their network structures are different from those of other subpopulations.

In the constellation plot construction, Barnes-Hut t-SNE with default setting (perplexity of 30 and 1000 iterations) was run on the centroids (of expression/measurement signal) of the discovered clade subpopulations for the entire dataset after PAC-MAN; t-SNE plot projects and separates the centroids in two dimensions. Next, the clades are color-coded such that 1) grey color indicates sample-specific clade and 2) non-grey colors indicate clades with multiple sample representation. The subpopulations in each clade are grouped by lines connecting the closest clade subpopulation, analogous to the visualization of stars by constellation nomenclature.

Relative Euclidean distances (in the t-SNE embedding) between subpopulations and clade centers are utilized to prune away subpopulations that are far away within clades. For clades containing three or more subpopulations, the distances to clade centroids for each clade on the t-SNE plane are used as thresholds, and subpopulations that are more than twice (threshold constant multiplier) the average distance to their clade centroid are pruned. For clades with only two subpopulations, the distances between the subpopulations for each two-subpopulation clade are calculated; the mean of these distances for the two-subpopulation clades is used as a global threshold. Any two-subpopulation clade with separate larger than twice (threshold constant multiplier) this global threshold is pruned. The researcher also controls a maximum global separation threshold, and the pruning procedure uses the minimum of the

thresholds to determine the pruning of clade subpopulations. All pruned away subpopulations are given new clade designation (S9 Fig).

Network annotation of subpopulations

To annotate the cellular states, we first apply PAC-MAN to learn the dataset-level subpopulation/clade labels. Next, these labels are used to learn the representative/clade networks. The top hubs (i.e. the most connected nodes) in these networks are used for annotation. This approach has biological significance in that important markers in a cellular state are often central to the underlying marker network, which is analogous to important genes in gene regulatory networks; these important markers have many connections with other markers. If the connections were broken, the cell would be perturbed and potentially driven to other states.

Running published methods

To run t-SNE [16] a dimensionality reduction visualization tool, we utilized the scripts published here (https://lvdmaaten.github.io/tsne/). Default settings were used.

To run SPADE, we first converted the simulated data to fcs format using Broad Institute's free CSVtoFCS online tool in GenePattern[17] (http://www.broadinstitute.org/cancer/software/genepattern#).

Next, we carried out the tests using the SPADE package in Bioconductor R[18] (https://github.com/nolanlab/spade).

To run flowMeans, we carried out the tests using the flowMeans package in Bioconductor R[1] (https://bioconductor.org/packages/release/bioc/html/flowMeans.html).

In the comparisons, we selected only cases that work for all methods to make the tests as fair as possible.

To calculate the mutual information of the subpopulations, we use the infotheo R package (https://cran.r-project.org/web/packages/infotheo/index.html).

To run network inference, we use the mrnet algorithm in the parmigne R package [15]. (https://cran.r-project.org/web/packages/parmigene/index.html).

Code availability

The PAC R package can be accessed at: https://cran.r-project.org/web/packages/PAC/index.html

Simulated data for clustering analysis

To compare the clustering methods, we generated simulated data from Gaussian Mixture Model varying dimension, the number of mixture components, mean, and covariance. The dimensions range from 5 to 50. The number of mixture components is varied along each dimension. The mean of each component was generated uniformly from a d-dimensional hypercube; we generated datasets using hypercube of different sizes, but kept all the other attributes the same. The covariance matrices were generated as AA^T, where A is a random matrix whose elements were independently drawn from the standard normal distribution. The sizes of the simulated dataset range from 100k to 200k.

The simulated data are provided as (Datasets 1–6). Datasets 1–6 are for the PAC part. Dataset 1 contains data with 5 dimensions; Dataset 2 contains data with 10 dimensions; Dataset 3 contains data with 20 dimensions; Dataset 4 contains data with 35 dimensions; Dataset 5

contains data with 40 dimensions; and Dataset 6 contains data with 50 dimensions. The ground truth labels are included as separate sheets in each dataset.

When applying flowMeans, SPADE, and the PAC to the data, we preset the desired number of subpopulations to that in the data to allow for direct comparisons.

Gated flow cytometry data

Two data files were downloaded from the FlowCAP challenges[2]. One data file is from the Hematopoietic stem cell transplant (HSCT) data set; it has 9,936 cell events with 6 markers, and human gating found 5 subpopulations. Another data file is from the Normal Donors (ND) data set; it has 60,418 cell events with 12 markers, and human gating found 8 subpopulations. The files are the first ('001') of each dataset. These data files were all 1) compensated, meaning that the spectral overlap is accounted for, 2) transformed into linear space, and 3) pre-gated to remove irrelevant events. We used the data files without any further transformation and filtering. When applying flowMeans, SPADE, and the PAC to the data, we preset the desired number of subpopulations to that in the data to allow for direct comparisons.

Gated mass cytometry data

Human gated mass cytometry data was obtained by gating for the conventional immunology cell types using the mouse bone marrow data recently published[11]. The expert gating strategy is provided as S1 Fig. The gated sample subset contains 64,639 cell events with 39 markers and 24 subpopulations and it is provided as Dataset 9.

To test the performance of different analysis methods, the data was first transformed using the asinh(x/5) function, which is the transformation used prior to hand-gating analysis; For SPADE analysis, we utilize the asinh(x/5) option in the SPADE commands. The post-clustering results from flowMeans, SPADE, b-PAC, and d-PAC were then subsetted using the indexes of gated cell events. These subsetted results are compared to the hand-gated results.

Simulated data for MAN analysis

To test the linking of subpopulations, we generated simulated data from multivariate Gaussian with preset signal levels and randomly generated positive definite covariance matrices. There are two cases, batch effect and dynamic. Each simulated sample file has five dimensions, with two of these varying in levels; these are the dimensions that are visualized. Dataset 7 contains the data for general batch effects case and Dataset 8 contains the data for dynamic effects case. The ground truth labels are included as separate sheets in each dataset.

General batch scenario. Sample 1 represents data from an old instrument (instrument 1) while sample 2 represents data from a new instrument (instrument 2). There are two subpopulations per sample. These two subpopulations are the same, but their mean marker levels shifted higher up in sample 2 due to higher sensitivity of instrument 2 (Fig 6A). The subpopulations have different underlying relationships between the markers. In this simulated experiment, five markers were measured. Out of the five markers, two markers show significant shift, and we focus on these two dimensions by 2-dimensional scatterplots. In Fig 6A, the left subpopulation in sample 1 is the same as the left subpopulation in sample 2; the same with the right subpopulation. The same subpopulations were generated from multivariate Gaussian distributions with changing means with fixed covariance structure.

Dynamic scenario. Dynamic scenario models the treatment-control and perturbation studies. In the simulation, we have generated two subpopulations that nearly converge over the time course (Fig 9). The researcher could lose the dynamic information if they were to

combine the samples for clustering analysis. The related subpopulations were generated from multivariate Gaussian distributions with changing means with fixed covariance structure.

Raw CyTOF data processing

The researcher preprocesses the data to 1) normalize the values to normalization bead signals, 2) de-barcode the samples if multiple barcoded samples were stained and ran together, and 3) pre-gate to remove irrelevant cells and debris to clean up the data[9,19]. Gene expressions look like log-normal distributions[20]; given the lognormal nature of the values, the hyperbolic arcsine transform is applied to the data matrix to bring the measured marker levels (estimation of expression values) close to normality, while preserving all data points. Often, researchers use the asinh(x/5) transformation, and we use the same transformation for the CyTOF datasets analyzed in this study.

Mouse tissue data

In the Spitzer et al., 2015 dataset[11], three mouse strains were grown, and total leukocytes were collected from different tissues: thymus, spleen, small intestine, mesenteric lymph node, lung, liver, inguinal lymph node, colon, bone marrow, and blood. In each experiment, 39 expression markers were monitored. The authors used the C57BL6 mouse strain as the reference[11]; the data was downloaded from Cytobank, and we performed our analysis on the reference strain.

First, all individual samples were filtered by taking the top 95% of cells based on DNA content and then the top 95% of cells based on cisplatin: DNA content allows the extraction of good-quality cells and cisplatin level (low) allows the extraction of live cells. Overall, the top 90% of cell events were extracted. The filtered samples were then transformed by the hyperbolic arcsine (x/5) function, and merged as a single file, which contains 13,236,927 cell events and 39 markers per event (S2 Table).

Using PAC-MAN, we obtained 50 subpopulations in each sample, then, using elbow point analysis, we output 130 clades for the entire dataset. The 130 clades account for the traditional immune subpopulations and sample-specific subpopulations, which may include resident immune cells that are unique to certain tissues. In the network alignment step, smaller PAC subpopulations (<1,000 cells) are left out because they may not have stable covariance and network structures. We attempt to assign the left-out small subpopulations back to the dataset: hierarchical clustering of the cluster centroids (marker signals or expression level) was performed, and we limit the total number of unique small sample-specific subpopulation by generating 5 "expression" clades per sample in the clustering (the larger subpopulations with a maximum of four sample-specific minor subpopulations that have less than 1,000 cells). Subsequently, any clade with less than 100 cells was discarded. Subpopulation proportion heatmap was plotted to visualize the subpopulation-specificities and relationships across the samples. Network annotation was performed using the hub markers of each representative subpopulation in each sample. Finally, we plotted the expression heatmap for all the clades and the constellation plot to visualize the cross-sample clade relationships.

Supporting information

S1 Fig. Gating strategy of CyTOF data for methods comparison. Biaxial gating hierarchy for the mouse bone marrow CyTOF dataset. Gating strategy that was used to find 24 reference populations in the mouse bone marrow CyTOF data. Pre-gating step involved removal of doublets, dead cells, erythrocytes and neutrophils. Non-neutrophils population was either subject to cluster analysis by computational tools or subsequent gating. Dotted boxes represent 24

terminal gates that were selected as reference populations for the comparison analysis.
(TIF)

S2 Fig. Subpopulation purity of simulated and real CyTOF data. (a) Subpopulation-specific purity plot of 35-dimensional simulated data with 10 subpopulations. The blue points denote the differences between the p-measures of the partition-based method (either d-PAC or b-PAC) and flowMeans, while the red points denote the p-measure differences between the partition methods and SPADE. The horizontal line at 0 means no difference between the methods. Most of the blue and red points are above 0, indicating that the PAC generates purer subpopulations compared to the ground truth. The two subplots are very similar, which means that d-PAC and b-PAC give very similar p-measures. More precisely, the sum of differences between d-PAC and flowMeans and d-PAC and SPADE are 0.85 and 1.09, respectively; and the overall difference between b-PAC and flowMeans and b-PAC and SPADE are 0.84 and 1.08, respectively.
(b) Subpopulation-specific purity plot of the hand-gated CyTOF data. The same convention is used as in (S2A Fig). Again, more blue and red points are above 0, indicating that the partition-based methods generate purer subpopulations compared to the ground truth. There is a cluster of points below 0 occurring in the middle of the plot, suggesting that flowMeans and SPADE capture the mid-size subpopulations more similar to hand-gating than the partition-based methods. More specifically, flowMeans does better (p-measure difference of 0.1 or better; difference of less 0.1 is considered practically no difference) with finding subpopulations of GMP, CD8 T cells, MEP, CD4 T cells (compared to d-PAC), and Plasma cells, while SPADE does better with CD19+IgM- B cells, NK cells (compared to d-PAC), CD8 T cells, NKT cells, Basophils, Short-Term HSC, and Plasma cells. However, overall, PAC has a much better performance, as the absolute sum of points above 0 is higher than that of points below 0. More precisely, the sum of differences between d-PAC and flowMeans and d-PAC and SPADE are 1.21 and 1.45, respectively; and the overall difference between b-PAC and flowMeans and b-PAC and SPADE are 2.06 and 2.31, respectively. The difference table is provided in S1 Table.
(TIF)

S3 Fig. Networks inferred from subpopulations in the dynamic example simulated dataset. Fig 9 introduced the dynamic example in which five samples each having 2 true subpopulations captures the almost-convergence of means. Here the underlying network structures for the PAC discovered subpopulations (three per sample) in Fig 10 are presented.
(TIF)

S4 Fig. Comparison between aligning cross-sample subpopulations by network, expression profile, or both. (a) PAC can be used to discover more subpopulations, with the effect of more partitions from the true clusters. (b) When over-partitioning is present, network or expression profile alone cannot resolve the dynamic (or batch) effects due to noisy covariance for small fragments of distributions. However, first aligning the larger subpopulations with more stable covariance, and thus network structures, and then merge in the smaller subpopulations by expression profile resolves the effects. (c) If more irrelevant edges were introduced, network alignment would fail due to the negative impact of the miscellaneous edges; however, eliminating small subpopulations from the alignment step alleviates the increased edge count problem.
(TIF)

S5 Fig. PAC-MAN style linkage by means. (a) t-SNE plots of mouse tissue samples colored by representative subpopulations labels from linkage by means. (b) Subpopulation proportion heatmap of clades of samples from linkage by means.
(TIF)

S6 Fig. Comparison between network and means PAC-MAN. (a) PAC-discovered subpopulations are aggregated by MAN into clades; the number of PAC subpopulations/clades for the network and means PAC-MAN approaches are plotted. (b) After aggregating shared clades within samples, the number of shared clades for the entire dataset is plotted for the two PAC-MAN approaches.
(TIF)

S7 Fig. Clustering with t-SNE projected points. We use t-SNE plots heavily for visualization in our study. We tested the approach of clustering on t-SNE projected points using kmeans. We observe that, despite being a very valuable visualization tool, t-SNE points do not contain much information for defining well-separated clusters for the usual clustering algorithms that depend on Gaussian geometry. It is best to perform the clustering using all data points in the original high-dimensional space, and then use t-SNE to visualize a subset of the points (amount chosen with computational capacity to run t-SNE).
(TIF)

S8 Fig. Elbow point analysis to find practical optimal number of clades. Elbow point analysis is the most computationally feasible approach to find the optimal number of clades to output. We calculated the within-cluster errors (from the centroid) for each of the example tissue sample. Next, we averaged the within-cluster errors for all 10 tissue samples. This calculation was performed for a range of numbers of clades. Next, loess smoothing was applied to the average within-cluster errors over the numbers of clades. The elbow point occurs at 130 clades, highlighted by the vertical blue line.
(TIF)

S9 Fig. Pruned constellation plot. As descripted in Materials and Methods, relative distances between subpopulations and clade centers are utilized to prune away subpopulations that are far away within clades. Clades 6 and 39 were pruned by setting threshold constant multiplier at 2.
(TIF)

S10 Fig. t-SNE visualization of clustering methods. Higher resolution version of Fig 4.
(TIFF)

S11 Fig. Visualization of PAC vs. PAC-MAN results for blood, bone marrow, colon, inguinal lymph node, and liver samples. Higher resolution version of Fig 12 (in another color scheme) with subpopulation and clade labels.
(TIF)

S12 Fig. Visualization of PAC vs. PAC-MAN results for lung, mesenteric lymph node, spleen, thymus, and small intestine samples. Higher resolution version of Fig 13 (in another color scheme) with subpopulation and clade labels.
(TIF)

S13 Fig. Heatmap of clade proportions across the tissue samples. Higher resolution version of Fig 14 with clade labels.
(TIF)

S14 Fig. Heatmap of average subpopulation expression levels in all tissue samples. Higher resolution version of Fig 15 with clade labels.
(TIF)

S1 Table. Purity (p) measure differences in CyTOF comparison. p-measure differences in gated CyTOF data analysis comparison. The differences are shown for all the annotated cell subpopulations, which are ordered by their sizes. Overall, the PAC methods give more positive p-measures.
(XLSX)

S2 Table. Sample sizes in mouse tissue CyTOF dataset. The numbers of cells in the samples of Spitzer et al., 2015 CyTOF dataset. The data is from the C57BL6 mouse strain and a total of ten tissue samples are present. The raw column shows the number of cells prior to filtering by DNA and cisplatin values. The final cell counts are shown in the filtered file (3^{rd}) column.
(XLSX)

S3 Table. PAC-MAN subpopulation characterization output for mouse tissue CyTOF dataset. The full set of annotated results, along with mean expressions, subpopulation proportion and counts, are reported.
(XLS)

Acknowledgments

We thank the members of Wong Lab, in particular Tung-yu Wu, Chen-yu Tseng and Kun Yang, for critical feedback. We thank Karen Sachs for critical feedback and valuable discussions.

Author Contributions

Conceptualization: Ye Henry Li, Dangna Li, Xiaowei Wang, Leying Guan, Wing Hung Wong.

Data curation: Ye Henry Li, Dangna Li, Nikolay Samusik, Xiaowei Wang.

Formal analysis: Ye Henry Li, Dangna Li, Nikolay Samusik, Xiaowei Wang.

Funding acquisition: Wing Hung Wong.

Investigation: Ye Henry Li, Dangna Li, Nikolay Samusik, Xiaowei Wang, Leying Guan, Wing Hung Wong.

Methodology: Ye Henry Li, Dangna Li, Wing Hung Wong.

Project administration: Ye Henry Li, Garry P. Nolan, Wing Hung Wong.

Resources: Ye Henry Li, Dangna Li, Nikolay Samusik.

Software: Ye Henry Li, Dangna Li.

Supervision: Ye Henry Li, Garry P. Nolan, Wing Hung Wong.

Validation: Ye Henry Li, Dangna Li, Nikolay Samusik.

Visualization: Ye Henry Li, Dangna Li.

Writing – original draft: Ye Henry Li, Dangna Li, Xiaowei Wang, Wing Hung Wong.

Writing – review & editing: Ye Henry Li, Dangna Li, Wing Hung Wong.

References

1. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytometry A. 2011 Jan 1; 79A(1):6–13.

2. Aghaeepour N, Finak G, Consortium TF, Consortium TD, Hoos H, Mosmann TR, et al. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods. 2013 Mar; 10(3):228–38. https://doi.org/10.1038/nmeth.2365 PMID: 23396282
3. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011 Oct; 29(10):886–91. https://doi.org/10.1038/nbt.1991 PMID: 21964415
4. Wong WH, Ma L. Optional Pólya tree and Bayesian inference. Ann Stat. 2010 Jun; 38(3):1433–59.
5. Lu L, Jiang H, Wong WH. Multivariate Density Estimation by Bayesian Sequential Partitioning. J Am Stat Assoc. 2013 Dec 1; 108(504):1402–10.
6. Jiang H, Mu JC, Yang K, Du C, Lu L, Wong WH. Computational Aspects of Optional Pólya Tree. J Comput Graph Stat. 2015 Feb 13;0(ja):00–00.
7. Li D, Yang K, Wong WH. Density Estimation via Discrepancy Based Adaptive Sequential Partition. In: Advances in Neural Information Processing Systems. 2016. p. 1091–1099.
8. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods. 2016 Jun; 13(6):493–6. https://doi.org/10.1038/nmeth.3863 PMID: 27183440
9. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, et al. Normalization of mass cytometry data with bead standards. Cytometry A. 2013 May 1; 83A(5):483–94.
10. Zunder ER, Lujan E, Goltsev Y, Wernig M, Nolan GP. A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. Cell Stem Cell. 2015 Mar 5; 16(3):323–37. https://doi.org/10.1016/j.stem.2015.01.015 PMID: 25748935
11. Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, et al. An interactive reference framework for modeling a dynamic immune system. Science. 2015 Jul 10; 349 (6244):1259425. https://doi.org/10.1126/science.1259425 PMID: 26160952
12. Ostrovsky R, Rabani Y, Schulman LJ, Swamy C. The Effectiveness of Lloyd-type Methods for the K-means Problem. J ACM. 2013 Jan; 59(6):28:1–28:22.
13. Arthur D, Vassilvitskii S. K-means++: The Advantages of Careful Seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics; 2007. p. 1027–1035. (SODA '07).
14. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods. 2016 Jun; 13(6):493–6. https://doi.org/10.1038/nmeth.3863 PMID: 27183440
15. Sales G, Romualdi C. parmigene—a parallel R package for mutual information estimation and gene network reconstruction. Bioinformatics. 2011 Jul 1; 27(13):1876–7. https://doi.org/10.1093/bioinformatics/btr274 PMID: 21531770
16. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9(Nov):2579–605.
17. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006 May; 38(5):500–1. https://doi.org/10.1038/ng0506-500 PMID: 16642009
18. Linderman MD, Bjornson Z, Simonds EF, Qiu P, Bruggner RV, Sheode K, et al. CytoSPADE: high-performance analysis and visualization of high-dimensional cytometry data. Bioinformatics. 2012 Sep 15; 28(18):2400–1. https://doi.org/10.1093/bioinformatics/bts425 PMID: 22782546
19. Zunder ER, Finck R, Behbehani GK, Amir ED, Krishnaswamy S, Gonzalez VD, et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. Nat Protoc. 2015 Feb; 10(2):316–33. https://doi.org/10.1038/nprot.2015.020 PMID: 25612231
20. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res. 2005 Oct 1; 15(10):1388–92. https://doi.org/10.1101/gr.3820805 PMID: 16204192