

1
2
3
4 Scalable Multi-Sample Single-Cell Data Analysis by Partition-Assisted Clustering and Multiple
5 Alignments of Networks
6
7
8

9 Ye Henry Li^{1,¶}, Dangna Li^{2,¶}, Nikolay Samusik³, Xiaowei Wang⁴, Leying Guan⁴, Garry P.
10 Nolan³, Wing Hung Wong^{4,5,*}

11
12 ¹Structural Biology Department and Public Policy Program, Stanford University, Stanford, USA.
13

14 ²Institute for Computational and Mathematical Engineering, Stanford University, Stanford, USA.
15

16 ³Department of Microbiology and Immunology, Baxter Laboratory, Stanford University,
17 Stanford, USA

18 ⁴Statistics Department, Stanford University, Stanford, USA.

19 ⁵Department of Biomedical Data Science, Stanford University, Stanford, USA

20

21 *Corresponding author.
22 Email: whwong@stanford.edu

23

24 [¶]These authors contributed equally to this work.
25

26

27 **Abstract**

28

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

41

42 **Author Summary**

43

44 Recently, the cytometry field has experienced rapid advancement in the development of
45 mass cytometry (CyTOF). CyTOF enables a significant increase in the ability to monitor 50 or
46 more cellular markers for millions of cells at the single-cell level. Initial studies with CyTOF
47 focused on few samples, in which expert manual discovery of cell types were acceptable. As the
48 technology matures, it is now feasible to collect more samples, which enables systematic studies
49 of cell types across multiple samples. However, the statistical and computational issues
50 surrounding multi-sample analysis have not been previously examined in detail. Furthermore, it
51 was not clear how the data analysis could be scaled for hundreds of samples, such as those in
52 clinical studies. In this work, we present a scalable analysis pipeline that is grounded in strong
53 statistical foundation. Partition-Assisted Clustering (PAC) offers fast and accurate clustering and
54 Multiple Alignments of Networks (MAN) utilizes network structures learned from each
55 homogeneous cluster to organize the data into data-set level clusters. PAC-MAN thus enables the
56 analysis of a large CyTOF dataset that was previously too large to be analyzed systematically;
57 this pipeline can be extended to the analysis of similarly large or larger datasets.

58

59

60

61 Introduction

62

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

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

86 In this study, we address the data analysis challenges in two major steps. First, we
87 propose the partition-assisted clustering (PAC) approach, which produces a partition of the k-
88 dimensional space (k=number of markers) that captures the essential characteristic of the data
89 distribution. This partitioning methodology is grounded in a strong mathematical framework of
90 partition-based high-dimensional density estimation[4–8]. The mathematical framework offers
91 the guarantee that these partitions approximate the underlying empirical data distribution; this
92 step is faster than the recent k-nearest neighbor-based method [9] and is essential to the
93 scalability of our clustering approach to analyze datasets with many samples. The clustering of
94 cells based on recursive partitioning is then refined by a small number of k-means style iterations
95 before a merging step to produce the final clustering.

96 Secondly, the subpopulations learned separately in multiple different but related datasets
97 can be aligned by marker network structures (multiple alignments of networks, or MAN),

98 making it possible to characterize the relationships of subpopulations across different samples
99 automatically. The ability to do so is critical for monitoring changes in a subpopulation across
100 different conditions. Importantly, in every study, batch effect is present; batch effects shift
101 subpopulation signals so that the means can be different from experiment to experiment. PAC-
102 MAN naturally addresses batch effects in finding the alignments of the same or closely related
103 subpopulations from different samples.

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

107

108 **Results and Discussion**

109

110 **PAC**

111 PAC has two parts: partitioning and post-processing. In the partitioning part of PAC, the
112 data space is recursively divided into smaller hyper-rectangles based on the number of data
113 points in the locality (Fig 1a). The partitioning is accomplished by either Bayesian Sequential
114 Partition (BSP) with limited look-ahead (Fig 1a and 1b) or Discrepancy Sequential Partition
115 (DSP) (Fig 1a); these are two fast variants of partition-based density estimation methods
116 previously developed by our group [4–8], with DSP being the fastest. BSP and DSP divide the
117 sample space into hyper-rectangles with uniform density value in each of them. The subsetting of
118 cells according to the partitioning provides a principled way of clustering the cells that reflects
119 the characteristics of the underlying distribution. In particular, each significant mode is captured
120 by a number of closely located rectangles with high-density values (Fig 1c). Although this
121 method allows a fast and unbiased localization of the high-density regions of the data space, we
122 should not use the hyper-rectangles directly to define the final cluster boundaries for two
123 reasons. First, real clusters are likely to be shaped elliptically, therefore, the data points in the
124 corners of a hyper-rectangle are likely to be incorrectly clustered. Second, a real cluster is often
125 split into more than one closely located high-density rectangles. We designed post-processing
126 steps to overcome these limitations: 1) a small number of k-means iterations is used to round out
127 the corners of the hyper-rectangles, 2) a merging process is implemented to ameliorate the
128 splitting problem, which is inspired by the flowMeans algorithm. The details of post-processing
129 are given in the Materials and Methods. The resulting method is named b-PAC or d-PAC
130 depending on whether the partition is produced by BSP or DSP.

131

132 **Fig 1: PAC recursively partitions the data space to obtain rational initialization structure.**
133 (a) Partition-based methods estimate data density by cutting the data space into smaller
134 rectangles. Bayesian Sequential Partition (BSP) divides the data space via binary partition in the
135 middle of the bounded region, while that of Discrepancy Sequential Partition (DSP) occur at the
136 location that balances the data point uniformly on both sides of the cut. The numbers denote
137 sequential order of partitions. Since DSP adapts to the data points, it converges on the estimated
138 density faster than BSP. (b) In the (one-step) look-ahead of version of partition, the algorithm
139 cuts the data space for all potential cuts plus one step more (steps 2 and 3), and it finds the
140 optimal future version (after step 3), which determines the actual cut (step 2). (c) The
141 partitioning of simulated data space containing five subpopulations; the hyper-rectangles
142 surround high-density areas, approximating the underlying distribution.

143

144 **MAN**

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

- 150 1) *Dataset size and instruments used.* Large number of samples usually means the samples
151 were collected on different days with different experimental preparations. Many steps can
152 introduce significant shifts in measurement levels.
- 153 2) *Staining reagents.* Reagents such as antibodies, purchased from different vendors and
154 batch preparations can affect the overall signal. While saturation of reagents in the
155 protocol could help eliminate the batch effects in the staining procedure, this approach is
156 costly and might not work for all antibodies, especially those with poor specificity.
- 157 3) *Normalization beads stock.* While normalization beads[10] help to control for the signal
158 level, especially within one experiment, the age of the beads stock and their preparation
159 could lead to significant batch effects. In addition, there are different types of
160 normalization beads and normalization calculations.
- 161 4) *Human work variation.* While many researchers are studying the same system (e.g.,
162 immune system), different protocols and implementation by different researchers, who
163 sometimes perform experimental steps slightly differently, can lead to batch effects.
- 164 5) *Subpopulation dynamics.* The subpopulation centers can move from sample to sample
165 due to treatments on the cells in treatment-control studies or perturbation studies. General
166 practice is to cluster by phenotypic markers.
- 167 6) *Sample background.* If the data came from different cell lines or individuals in a clinical
168 study, the measurement levels and proportions of cell subpopulations would be expected

169 to change from sample to sample. Without expert scrutiny, it would be difficult to make
170 sense of the data with current data analysis tools.

171

172 Could we extract shared information that allows us to interpret cross-sample similarities
173 and differences? To ameliorate these difficulties, we have designed an alternative approach that
174 is effective in the presence of substantial systematic between-sample variation. In this approach,
175 each sample is analyzed separately (by PAC) to discover within-sample subpopulations. Over-
176 partitioning in this step is allowed in order not to miss small subpopulations in high dimension
177 due to lack of prior knowledge. The subpopulations from all samples are then compared to each
178 other based on a pairwise dissimilarity measure designed to capture the differences in within-
179 sample distributions (among the markers) across two subpopulations. Using this dissimilarity, we
180 perform bottom-up hierarchical clustering of the subpopulations to represent the relationship
181 among the subpopulations. The resulting tree of subpopulations is then used to guide the merging
182 of subpopulations from the same sample, and to establish linkage of related subpopulations from
183 different samples. We note that the design of a dissimilarity measure (Materials and Methods)
184 that is not sensitive to systematic sample-to-sample variation is a novel aspect of our approach.
185 The merging of subpopulations from the same sample is also important, as it offers a way to
186 consolidate any over-partitioning that may have occurred during the initial PAC analysis of each
187 sample. We emphasize that, as with the usage of all statistical methods, the user must utilize
188 samples or datasets that are considered as good as possible; interpretation of the analysis results
189 rely on the researchers to collect data with validated reagents for all samples.

190

191 **Rational initialization for PAC increases clustering effectiveness**

192 Appropriate initialization of clustering is very important for eventually finding the
193 optimal clustering labels; PAC works well because the implicit density estimation procedure
194 yields rational centers to learn the modes of sample subpopulations. When tested on the hand-
195 gated CyTOF data on the bone marrow sample in (14), compared to k-means alone, PAC gives
196 lower total sums of squares and higher F-measures in the subpopulations (Fig 2a and 2b). This
197 process also helps PAC to converge in 50 iterations (Fig 3) in post-processing, whereas k-means
198 performs very poorly even after 5000 iterations (Fig 4). Through the lens of t-sne plots (Fig 4),
199 the PAC results are more similar to the hand-gating results, while the k-means, flowMeans, and
200 SPADE clustering results perform poorly. In flowMeans, several large subpopulations are
201 merged. SPADE's separation of points is inconsistent and highly heterogeneous, probably due to
202 its down-sampling nature. On the other hand, by inspection, PAC obtains similar separation for
203 both the major and minor subpopulations as the hand-gating results.

204

205 **Fig 2. Rational initialization is better than random initialization.** The hand-gated CyTOF
206 data (see S1 Fig) is used for illustration. In this case, (a) the overall sum of squares error is lower
207 and (b) the F-measure is higher for PAC.

208

209 **Fig 3. Rational initialization and minimal kmeans post-processing iterations give fast**
210 **convergence.** The convergence of PAC toward the hand-gated results, or ground truth, is fast. It
211 takes less than 50 downstream post-processing kmeans iterations for the PAC to achieve
212 convergence.

213

214 **Fig 4. Visualization and comparison of clustering results by t-sne plots.** Each t-sne plot
215 contains the same 10,000 cell events from the hand-gated CyTOF data with different set of
216 colored labels drawn. Note that the colors are informative only within each panel. These labels
217 are from kmeans, SPADE, flowMeans, b-PAC, and d-PAC. The subpopulation numbers for all
218 methods were set to be the same as that of hand-gated results. PAC methods achieve a
219 significantly better convergence to the hand-gate labels than alternative methods.

220

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

223 In the systematic simulation study, we challenged the methods with different datasets
224 with varying number of dimensions, number of subpopulations, and separation between the
225 subpopulations. The F-measure and p-measures for the PAC methods are consistently equal or
226 higher than that of flowMeans and SPADE (Table 1 and S2a Fig). In addition, we observe that
227 flowMeans gives inconsistent F-measures for similar datasets (Table 1), which may be due to the
228 convergence of k-means to a local minimum without a rational initialization.

229

230 **Table 1. F-measure Comparisons of Methods on Simulated and Hand-gated Cytometry**
231 **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	0.967	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_5_40_200k	0.96	0.89	0.99	0.99
35_10_40_200k	0.93	0.79	0.96	0.96
<hr/>				
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

232

233 F-measure is calculated using the original hand-gate labels and the estimated labels generated by
234 each analysis method. The true-positives are found if the methods assign the same labels to
235 points belonging to the same subpopulation in the hand-gated data. The more true-positives
236 found, the higher the F-measure, which ranges from 0 to 1, with 1 being the highest. Partition-
237 based methods perform consistently well on data ranging from 5 to 39 dimensions. In the
238 simulations, d-PAC and b-PAC perform just as well or better than flowMeans and SPADE.
239 flowMeans gives drastically different F-measures for the cases 20_10_40_100k and
240 20_20_40_100k : 0.25386 vs. 0.92518; this large difference is likely due to the random initiation
241 of cluster centers. In the hand-gated datasets, SPADE has the worst performance. Ultimately, the
242 performance of flowMeans and SPADE deteriorate for the 39-dimensional real CyTOF data,
243 while d-PAC and b-PAC perform consistently well.

244 *Simulated data have the following convention: a_b_c_d, where a denotes the number of
245 dimensions/markers, b denotes the number of subpopulations, c denotes the size of the
246 hypercube for data generation, and d denotes the number of cells.

247 **from rounding up, not originally 1.00

248

249 Next, we tested the methods based on published hand-gated cytometry datasets to see
250 how similar the estimated subpopulations are to those obtained by human experts. We applied

251 the methods on the hematopoietic stem cell transplant and Normal Donors datasets from the
252 FlowCAP challenges[2] and on the subset of gated mouse bone marrow CyTOF dataset (Dataset
253 5) recently published[11]. The gating strategy of the CyTOF dataset is provided in Fig S1. The
254 dataset and expert gating strategy are the same as described earlier[12]. Note that in the flow
255 cytometry data, the computed F-measures are slightly lower than that reported in FlowCAP; this
256 is due to the difference in the definition of F-measures. Overall, the PAC outperforms
257 flowMeans and SPADE by consistently obtaining higher F-measures (Table 1). In particular, in
258 the CyTOF data example, PAC generated significantly higher F-measures (greater than 0.82)
259 than flowMeans and SPADE (0.59 and 0.53, respectively). In addition, PAC gives higher overall
260 subpopulation-specific purities (S2b Fig and S1 Table). These results indicate that PAC gives
261 consistently good results for both low and high-dimensional datasets. Furthermore, PAC results
262 match human hand-gating results very well. The consistency between PAC-MAN results and
263 hand-gating results in this large data set confirms the practical utility of the methodology.

264

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

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

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

282

283 **Fig 5. Schematic analogy of MAN.** Consider a deck of networks (in analogy to cards), with
284 each “suit” representing a sample and each “rank” representing a unique network structure. The
285 networks are aligned by similarity and organized on a dendrogram. The tree is cut (red line) at
286 the user-specified level to output the desired k clades. Within each clade, the network structures

287 are similar or the same. If the same sample has multiple networks in the same clade, then these
288 networks are merged (black box around same cards).

289

290 **Fig 6. Simple Batch Effect Scenario.** (a) Simulated data samples with two of the same
291 subpopulations. The means shifted due to measurement batch effect. (b) When the samples are
292 combined, as in the case of analyzing all samples together, two different subpopulations overlap
293 (left panel). The overlapped subpopulations cannot be distinguished by clustering (right panel).
294 (c) PAC could be used to discover more subpopulations, however, the hints of the present of
295 another subpopulation do not help to resolve the batch effect.

296

297 PAC-MAN resolves the overlapping issue by analyzing the samples separately (Fig 7). In
298 the case in which we do not know *a priori* the number of true subpopulations, we learn three
299 subpopulations per sample (Fig 7a). The network structures of the subpopulations discovered are
300 presented in Fig 7b-c and we see that the third subpopulations from the two samples share the
301 same network structures, while the first subpopulations of the two samples differ by only one
302 edge; these respective networks are clustered together in the dendrogram (Fig 8a right panel). By
303 utilizing the networks, the clades that represent the same and/or similar subpopulations of cells
304 can be established. Clustering by network structures alone resolves the majority of points in the
305 data (Fig 8a, left panel). Furthermore, as discussed next, by incorporating marker levels into the
306 alignment process, all the subpopulations can be resolved (Fig 8b).

307

308 **Fig. 7. Calculation of sample clusters and their underlying network structures.** (a) PAC was
309 used to discover several subpopulations per sample without advanced knowledge of the exact
310 number of subpopulations. (b-c) The networks of the subpopulations in both samples discovered
311 in (a). Networks can be grouped by similarities to organize the subpopulations across samples;
312 the alignment is based on Jaccard dissimilarity network structure characterization matrix;
313 dendrogram of the hierarchical clustering results.

314

315 **Fig 8. Resolution of batch effects for simple batch effect scenario.** (a) Resolution of batch
316 effect by networks of all subpopulations discovered. (b) Resolution of batch effect first by
317 network structures of larger subpopulations and then by merging smaller subpopulations into the
318 aligned clades.

319

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

331

332 **Fig 9. Ground truth of simulated dynamic batch effect samples.** Two subpopulations, in blue
333 color, almost converge in time by mean shifts.

334

335 **Fig 10. PAC on dynamic batch effect scenario.** PAC discovers several subpopulations per
336 sample without advanced knowledge of the number of subpopulations present.

337

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

343

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

345 With networks in hand, we could further characterize the relationships between
346 subpopulations across samples. However, the alignment process needs to work well for true
347 linkage to be established. We could align by network alone, by expression (or marker) means, or
348 both. Figs 8 and 11 present these alternatives in comparison. By using all the subpopulation
349 networks, the results still contain subsets of misplaced cells (Figs 8a and 11 left panel). This is
350 because small clusters of cells have noisy underlying covariance structure; therefore, the
351 networks cannot be accurately inferred. These structural inaccuracies negatively impact the
352 network clustering. The (mean) marker level approach also does not work well (Fig 11 center
353 panel) due to the subpopulation mean shifts across samples. On the other hand, the sequential
354 approach works well (Figs 8b and 11 right panel). In the sequential approach, larger (>1500 in

355 batch effect case; >1000 in dynamic case) subpopulations' networks are utilized for the initial
356 alignment process. Next, the smaller subpopulations, which have noisy covariance, are merged
357 with the closest larger, aligned subpopulations. Thus, more subpopulations could be discovered
358 upstream (in PAC), and the network alignment would work similarly as the smaller
359 subpopulations, which could be fragments of a distribution, do not impact the alignment process
360 (S4a Fig and S4b Fig). Moreover, in the network inference step, unimportant edges can
361 negatively impact the alignment process (S4c Fig) in the network-alone case. Biologically, this
362 means that edges that do not constrain or define the cellular state should not be utilized in the
363 alignment of cellular states. Effectively, the threshold placed on the number of edges in the
364 network inference controls for the importance of the edges. Thus, the combined alignment
365 approach works well and allows moderate over-saturation of cellular states to be discovered in
366 the PAC step so that no advance knowledge of the exact number of subpopulations is necessary.

367

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

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

381 In contrast to the one-sample perspective, using d-PAC-MAN, the fastest approach by
382 our comparison results, we can perform subpopulation discovery for each sample automatically
383 and then align the subpopulations across samples to establish dataset-level cellular states. On a
384 standard Core i7-44880 3.40GHz PC computer, the single-thread data analysis process with all
385 data points takes about one hour to complete, which is much faster than alternative methods.
386 With multi-threading and parallel processing, the data analysis procedure can be completed very
387 quickly. As mentioned earlier, PAC results for the bone marrow subsetted data from this dataset
388 matches closely to that of the hand-gated results. This accuracy provides confidence for applying
389 PAC to the rest of the dataset.

390 Figs 11-12 show the t-sne plots for subpopulation discovered (top panel of each sample)
391 and the representative subpopulation established (bottom panel of each sample) for the entire

392 dataset. In the PAC discovery step, we learn 35 subpopulations per sample without advance
393 knowledge of how many subpopulations are present. This moderate over-partitioning of the data
394 samples leads to a moderate heterogeneity in the t-sne plots. Next, the networks are inferred for
395 the larger subpopulations (with number of cell events greater than 1000), and the networks are
396 aligned for all the tissue samples. We output 80 representative subpopulations or clades for the
397 entire dataset to account for the traditional immunological cellular states and sample-specific
398 cellular states present. Within samples, the subpopulations that cluster together by network
399 structure are aggregated. The smaller subpopulations (not involved in network alignment) are
400 either merged to the closest larger subpopulation or establish their own sample-specific
401 subpopulation by expression alignment; small subpopulations were clamped with larger clades
402 by grouping the subpopulations into 5 clusters per sample based on the means (of marker signal).
403 The representative subpopulations (90 total) follow the approximate distribution of the cell
404 events on the t-sne plots and the aggregating effect cleans up the heterogeneities due to over-
405 partitioning in the PAC step.

406

407 **Fig 12. Visualization of PAC-MAN results for Blood, Bone Marrow, Colon, Inguinal**
408 **Lymph Node, and Liver samples.** Each t-sne plot was generated using 10,000 randomly drawn
409 cell events from each mouse tissue sample. The results from PAC (top panel) and MAN (bottom
410 panel) steps are presented as a pair. Initial PAC discovery was set to 35 subpopulations without
411 advanced knowledge of the number of subpopulations in each sample. In MAN, 80 network
412 clades were outputted, and the cellular states are defined by expression (marker signal), network
413 structure, and dataset-level variation. This composite definition naturally aggregates the initial 35
414 subpopulations to yield smaller number of subpopulations in less variable samples.

415

416 **Fig 13. Visualization of PAC-MAN results for Lung, Mesenteric Lymph Node, Spleen,**
417 **Thymus, and Small Intestine samples.** The settings and descriptions are the same as those in
418 Fig 12. Continuation of visualization of PAC-MAN results for the mouse tissue data.

419

420 The cell type clades are the representative subpopulations for the entire dataset, and they
421 could either be present across samples or in one sample alone. Their distribution is visualized by
422 a heatmap (Fig 14). While the bone marrow sample contains many cell types, only a subset of
423 them are directly aligned to cell types in other samples, which means using the bone marrow data
424 as the reference point leaves much information unlocked in the dataset. Therefore, the data
425 suggests that the bone marrow cell types are not adequate in representing all cell types in the
426 immune system. The cell types in the blood and spleen samples have more alignments with cell
427 types in other samples. The lymph node samples share many clades; the small intestine and colon

428 samples also share many clades, probably due to closeness in biological function. The thymus
429 sample has few clades shared with other samples, which may be due to its functional specificity.

430

431 **Fig 14: Heatmap of clade proportions across the tissue samples.** Sample-specific clades have
432 a value of 1, while shared clades have proportions spread across different samples.
433 Physiologically similar samples share more clades.

434

435 PAC-MAN style analysis can be applied to align the tissue subpopulations by their means
436 instead of network similarities (S5 Fig). As done previously, representative clades (88 total) were
437 outputted. The same aggregating effect is observed (S5a Fig), and this is due to the organization
438 from dataset-level variation in the means. Comparing to the network alignment, the means
439 linkage approach has slightly more subpopulations per sample; the subpopulation proportion
440 heatmap (S5b Fig) shows more linking. Although the bone marrow sample subpopulations co-
441 occur in the same clades slightly more with other sample subpopulations, this sample does not
442 co-occur with many clades in the dataset. Thus, a PAC-MAN style analysis with means linkage
443 also harvests additional information from the entire dataset.

444 To compare the network and means approaches with PAC-MAN, we study the F-measure
445 and p-measure results with 88 total clades from each approach. The overall F-measure with all
446 cell events is 0.7969 and the overall F-measure with clades assignments of PAC-discovered
447 subpopulations is 0.3143. The two F-measure values suggest that the assignment of PAC-
448 discovered subpopulations is more consistent for larger subpopulations.

449 To illustrate the assignment purities, the p-measures are computed for the following two
450 cases. 1) Network clade assignment is the basis (network-justified), similar to the ground truth in
451 the clustering comparisons previously; or 2) means clade assignment is the basis (means-
452 justified) (S4 Table). P-measure cutoff is set at 0.3 (to remove unreliable comparisons) to obtain
453 purer clade assignments. In the network-justified case, PAC subpopulations with more than 0.3
454 in p-measure constitute 93.44 % of all cell events. In the means-justified case, PAC
455 subpopulations with more than 0.3 in p-measure constitute 92.67 % of all cell events.
456 Furthermore, if the p-measure cutoff were to increase to 0.5, the percentages of cells left for the
457 network-justified and mean-justified cases are 6.25% and 75.16%, respectively. The network-
458 justified case yields drastically lower numbers of cell events in the purer PAC subpopulations
459 because the means approach has more heterogeneity in the linkages (defined as PAC-
460 subpopulation participants in each shared clade with size of at least 2). In fact, the network
461 approach has 100 linkages while the means approach has 209 linkages. Therefore, the extra
462 linkages in the means approach would yield greater impurities in the network-justified case. The
463 linkage plot (S6a Fig) shows that the low linkages occur slightly more frequently for the network
464 approach. One consequence is that the network approach aggregates PAC subpopulations within

465 sample more frequently; for instance, in the thymus sample, the network approach yields 14
466 clades while the means approach yields 21 clades.

467 After aggregating, the clade sizes (with unique participants per sample) are plotted (S6b
468 Fig). The network approach tends to find fewer linkages, as more clades have sizes of less than
469 4, while the means approach has more clades than the network approach with clade sizes greater
470 than 4. The network approach is more conservative due to the additional constraints from
471 network structures. Conventionally, in the cytometry field, only the means are considered in the
472 definition of cellular states. Assuming the absence of batch and dynamic effects, the researcher
473 could view the purer shared clade assignments in the network-justified case (general agreement
474 between constrained network approach and means approach) as more reliable candidates of
475 cross-sample relationships to investigate in future experiments (S6c Fig).

476 Hence, the network alignment approach is in agreement that of the means approach, with
477 network alignment being more stringent in the establishment of linkages. The network PAC-
478 MAN approach defines cellular states with the additional information from network structures,
479 and it has the effect of constraining the number of linkages between samples while finding
480 linkages for subpopulations that are distant in their means.

481

482 **Network hubs provide natural annotations**

483 To further characterize the cell types, we annotate the clades within each sample using
484 the top network hub markers, which constrain the cellular states. The full annotation, along with
485 mean average expression profiles, is presented in S3 Table. The clade information is presented in
486 the ClusterID column. The annotations for cells across different samples but within the same
487 clades share hub markers. For example, in clade 1 for the blood and bone marrow samples, the
488 cells share the hub markers Ly6C and CD11b. In the bone marrow sample, one important set of
489 subpopulations is the hematopoietic stem cell subpopulations. One such subpopulation is present
490 as clade 18 with the annotation CD34-CD27-cKit-Sca1 and is about 1.87 percent in the bone
491 marrow sample. Clade 18 is only present in the bone marrow sample, indicating that the PAC-
492 MAN pipeline defines this as a sample-specific and coherent subpopulation using dataset-level
493 variation. The thymus contains a large subpopulation (84.07 percent) that is characterized as
494 CD5-CD4-CD43-CD3, suggesting it to be the maturing T-cell subpopulation.

495

496

497

498

499 **Conclusion**

500

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

513

514 **Materials and Methods**

515

516 **Partition-assisted clustering has two parts**

517 1) Partitioning: a partition method (BSP[5] or DSP[7]) is used to learn N initial cluster centers
518 from the original data.

519 2) Post-processing: A small number (m) of k-mean iterations is applied to the rectangle-based
520 clusters from the partitioning, where m is a user-specified number. We used m=50 in our
521 examples. After this k-means refinement, we merge the N clusters hierarchically until the desired
522 number of clusters (this number is user-specified) is reached. The merging is based on a given
523 distance metric for clusters. In the current implementation, we use the same distant metric as in
524 flowMeans[1]. That is, for two clusters X and Y, their distance D(X, Y) is defined as:

525
$$D(X, Y) = \min \{(\bar{x} - \bar{y})^T S_x^{-1}(\bar{x} - \bar{y}), (\bar{x} - \bar{y})^T S_y^{-1}(\bar{x} - \bar{y})\} \quad (1)$$

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

530

531 **Partition Methods**

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

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

541 BSP + LL is an approximation inference algorithm for Bayesian sequential partitioning density
542 estimation (BSP)[5]. It borrows ideas from Limited-Look-ahead Optional Pólya Tree (LL-OPT),
543 an approximate inference algorithm for Optional Pólya Tree[8]. The original inference algorithm
544 for BSP looks at one level ahead (i.e. looking at the possible cut points one level deeper) when
545 computing the sampling probability for the next partition. It then uses resampling to prune away
546 bad samples. Instead of looking at one level ahead, BSP + LL looks at h levels ahead ($h > 1$)
547 when computing the sampling probabilities for the next partition and does not do resampling (Fig
548 1b). In other words, it compensates the loss from not performing resampling with more accurate
549 sampling probabilities. For simplicity, 'BSP + LL' is shortened to 'BSP' in the rest of the article.

550

551 **F-measure**

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

$$559 P = \frac{TP}{TP+FP} \quad (2)$$

$$560 R = \frac{TP}{TP+FN} \quad (3)$$

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

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

567

568 **Purity-measure (p-measure)**

569 Most of the existing measurements for clustering accuracy aim at measuring the overall
570 accuracy of the entire datasets, i.e. comparing with the ground truth over all clusters. However,
571 we are also interested in analyzing how well a clustering result matches the ground truth within a
572 certain class. Specifically, consider a dataset D with K classes: $\{C_1, C_2, \dots, C_K\}$ and a given
573 ground truth cluster labels g , we construct an index called the purity measure, or p-measure for
574 short, to measure how well our clustering result matches g for each class C_i . This index is
575 computed as follows:

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

578 2) Define

$$579 S_1 = \frac{|C_k \cap L_{i_k}|}{|L_{i_k}|}, S_2 = \frac{|C_k \cap L_{i_k}|}{|C_k|} \quad (4)$$

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

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

$$582 P = \frac{2S_1S_2}{S_1+S_2} \quad (5)$$

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

587

588 **Network construction and comparison**

589 After PAC, the discovered subpopulations typically have enough cells for the estimation
590 of mutual information. This enables the construction of networks as the basis for cell type
591 characterization. Computationally, it is not good to directly use the mutual information networks
592 constructed this way to organize the subpopulations downstream. The distance measure used to
593 characterize the networks could potentially give the same score for different network structures.

594 Thus, it is necessary to threshold the network edges based on the strength of mutual information
595 to filter out the noisy and miscellaneous edges. In this work, these subpopulation-specific
596 networks are constructed using the MRNET network inference algorithm in the Parmigene [13]
597 R package. The algorithm is based on mutual information ranking, and outputs significant edges
598 connecting the markers. The top d edges (d is set to be 1x the number of markers in all examples)
599 are used to define a network for the subpopulation. This process enables a careful calculation of
600 the distance measure.

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

604
$$\text{Similarity}(G_1, G_2) = \frac{|S|}{|A|} \quad (6)$$

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

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

610

611 **Cross-sample linkage of subpopulations**

612 We perform agglomerative clustering of the pool of subpopulations from all samples.
613 This clustering procedure greedily links networks that are the closest in Jaccard dissimilarity, and
614 yields a dendrogram describing the distance relationship between all the subpopulations. We cut
615 the dendrogram to obtain the k clades of subpopulations. Subpopulations from the same sample
616 and falling into the same clade are then merged into a single subpopulation (Fig 5). This merging
617 step has the effect of consolidating the over-partitioning in the PAC step. No merging is
618 performed for subpopulations from different samples sharing the same clade. In this way, we
619 obtain k clades of subpopulations, with each clade containing no more than one subpopulation
620 from each sample. We regard the subpopulations within each clade as being linked across
621 samples.

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

627

628 **Annotation of Subpopulations**

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

636

637 **Running Published Methods**

638 To run t-SNE [14] a dimensionality reduction visualization tool, we utilized the scripts
639 published here (<https://lvdmaaten.github.io/tsne/>). Default settings were used.

640 To run SPADE, we first converted the simulated data to fcs format using Broad
641 Institute's free CSVtoFCS online tool in GenePattern[15]
642 (<http://www.broadinstitute.org/cancer/software/genepattern#>).

643 Next, we carried out the tests using the SPADE package in Bioconductor R[16]
644 (<https://bioconductor.org/packages/release/bioc/html/spade.html>).

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

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

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

651 To run network inference, we use the mrnet algorithm in the parmigene R package [13].
652 (<https://cran.r-project.org/web/packages/parmigene/index.html>).

653

654 **Code Availability**

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

656

657

658 **Simulated Data for Clustering Analysis**

659 To compare the clustering methods, we generated simulated data from Gaussian Mixture
660 Model varying dimension, the number of mixture components, mean, and covariance. The
661 dimensions range from 5 to 39. The number of mixture components is varied along each
662 dimension. The mean of each component was generated uniformly from a d-dimensional
663 hypercube; we generated datasets using hypercube of different sizes, but kept all the other
664 attributes the same. The covariance matrices were generated as AA^T , where A is a random matrix
665 whose elements were independently drawn from the standard normal distribution. The sizes of
666 the simulated dataset range from 100k to 200k.

667 The simulated data are provided as (Datasets 1-6). Datasets 1-4 are for the PAC part.
668 Dataset 1 contains data with 5 dimensions; Dataset 2 contains data with 10 dimensions; Datasets
669 3a and 3b contain data with 20 dimensions; and Datasets 4a and 4b contain data with 35
670 dimensions. The ground truth labels are included as separate sheets in each dataset.

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

673

674 **Gated Flow Cytometry Data**

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

684

685 **Gated Mass Cytometry Data**

686 Human gated mass cytometry data was obtained by gating for the conventional
687 immunology cell types using the mouse bone marrow data recently published[11]. The expert
688 gating strategy is provided as Fig S1. The gated sample subset contains 64,639 cell events with
689 39 markers and 24 subpopulations and it is provided as Dataset 7.

690 To test the performance of different analysis methods, the data was first transformed
691 using the $\text{asinh}(x/5)$ function, which is the transformation used prior to hand-gating analysis; For

692 SPADE analysis, we utilize the `asinh(x/5)` option in the SPADE commands. The post-clustering
693 results from flowMeans, SPADE, b-PAC, and d-PAC were then subsetted using the indexes of
694 gated cell events. These subsetted results are compared to the hand-gated results.

695

696 **Simulated Data for MAN Analysis**

697 To test the linking of subpopulations, we generated simulated data from multivariate
698 Gaussian with preset signal levels and randomly generated positive definite covariance matrices.
699 There are two cases, batch effect and dynamic. Each simulated sample file has five dimensions,
700 with two of these varying in levels; these are the dimensions that are visualized. Dataset 5
701 contains the data for general batch effects case and Dataset 6 contains the data for dynamic
702 effects case. The ground truth labels are included as separate sheets in each dataset.

703

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

714 **Dynamic scenario.** Dynamic scenario models the treatment-control and perturbation studies. In
715 the simulation, we have generated two subpopulations that nearly converge over the time course
716 (Fig 9). The researcher could lose the dynamic information if they were to combine the samples
717 for clustering analysis. The related subpopulations were generated from multivariate Gaussian
718 distributions with changing means with fixed covariance structure.

719

720 **Raw CyTOF Data Processing**

721 The researcher preprocesses the data to 1) normalize the values to normalization bead
722 signals, 2) de-barcode the samples if multiple barcoded samples were stained and ran together,
723 and 3) pre-gate to remove irrelevant cells and debris to clean up the data[10,17]. Gene
724 expressions look like log-normal distributions[18]; given the lognormal nature of the values, the
725 hyperbolic arcsine transform is applied to the data matrix to bring the measured marker levels
726 (estimation of expression values) close to normality, while preserving all data points. Often,

727 researchers use the $\text{asinh}(x/5)$ transformation, and we use the same transformation for the
728 CyTOF datasets analyzed in this study.

729

730 **Mouse Tissue Data**

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

736 First, all individual samples were filtered by taking the top 95% of cells based on DNA
737 content and then the top 95% of cells based on cisplatin: DNA content allows the extraction of
738 good-quality cells and cisplatin level (low) allows the extraction of live cells. Overall, the top
739 90% of cell events were extracted. The filtered samples were then transformed by the hyperbolic
740 arcsine ($x/5$) function, and merged as a single file, which contains 13,236,927 cell events and 39
741 markers per event (S2 Table).

742 Using PAC-MAN, we obtained 35 subpopulations in each sample then 80 clades for the
743 entire dataset. The 80 clades account for the traditional immune subpopulations and sample-
744 specific subpopulations. Small subpopulations not used in alignment are later merged into the
745 closest clades; this is done by performing hierarchical clustering with the marker signals to
746 obtain 5 “expression” subclades per sample. Subsequently, any clade with less than 100 cells is
747 discarded. Subpopulation proportion heatmap was plotted to visualize the subpopulation-
748 specificities and relationships across the samples. Finally, annotation was performed using the
749 hub markers of each representative subpopulation in each sample.

750

751 **Acknowledgements**

752 We thank the members of Wong Lab, in particular Tung-yu Wu, Chen-yu Tseng and Kun
753 Yang, for critical feedback.

754

755

756

757

758

759 **References**

760

- 761 1. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in
762 flow cytometry data. *Cytometry A*. 2011 Jan 1;79A(1):6–13.
- 763 2. Aghaeepour N, Finak G, Consortium TF, Consortium TD, Hoos H, Mosmann TR, et al.
764 Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*.
765 2013 Mar;10(3):228–38.
- 766 3. Qiu P, Simonds EF, Bendall SC, Gibbs Jr KD, Bruggner RV, Linderman MD, et al.
767 Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat*
768 *Biotechnol*. 2011 Oct;29(10):886–91.
- 769 4. Wong WH, Ma L. Optional Pólya tree and Bayesian inference. *Ann Stat*. 2010
770 Jun;38(3):1433–59.
- 771 5. Lu L, Jiang H, Wong WH. Multivariate Density Estimation by Bayesian Sequential
772 Partitioning. *J Am Stat Assoc*. 2013 Dec 1;108(504):1402–10.
- 773 6. Yang K, Wong WH. Discovering and Visualizing Hierarchy in the Data. *ArXiv14034370*
774 *Stat* [Internet]. 2014 Mar 18 [cited 2015 Nov 27]; Available from:
775 <http://arxiv.org/abs/1403.4370>
- 776 7. Yang K, Wong WH. Density Estimation via Adaptive Partition and Discrepancy Control.
777 *ArXiv14041425 Stat* [Internet]. 2014 Apr 4 [cited 2015 Nov 27]; Available from:
778 <http://arxiv.org/abs/1404.1425>
- 779 8. Jiang H, Mu JC, Yang K, Du C, Lu L, Wong WH. Computational Aspects of Optional
780 Pólya Tree. *J Comput Graph Stat*. 2015 Feb 13;0(ja):00–00.
- 781 9. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype
782 space with single-cell data. *Nat Methods*. 2016 Jun;13(6):493–6.
- 783 10. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, et al. Normalization of
784 mass cytometry data with bead standards. *Cytometry A*. 2013 May 1;83A(5):483–94.
- 785 11. Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, et al.
786 An interactive reference framework for modeling a dynamic immune system. *Science*. 2015
787 Jul 10;349(6244):1259425.
- 788 12. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype
789 space with single-cell data. *Nat Methods*. 2016 Jun;13(6):493–6.
- 790 13. Sales G, Romualdi C. *parmigene*—a parallel R package for mutual information estimation
791 and gene network reconstruction. *Bioinformatics*. 2011 Jul 1;27(13):1876–7.

792 14. Maaten L van der, Hinton G. Visualizing Data using t-SNE. *J Mach Learn Res.*
793 2008;9(Nov):2579–605.

794 15. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. *GenePattern 2.0. Nat Genet.*
795 2006 May;38(5):500–1.

796 16. Linderman MD, Bjornson Z, Simonds EF, Qiu P, Bruggner RV, Sheode K, et al.
797 *CytoSPADE: high-performance analysis and visualization of high-dimensional cytometry*
798 *data. Bioinformatics.* 2012 Sep 15;28(18):2400–1.

799 17. Zunder ER, Finck R, Behbehani GK, Amir ED, Krishnaswamy S, Gonzalez VD, et al.
800 *Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell*
801 *deconvolution algorithm. Nat Protoc.* 2015 Feb;10(2):316–33.

802 18. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. *Gene expression profiling in single cells*
803 *from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels.*
804 *Genome Res.* 2005 Oct 1;15(10):1388–92.

805

806

807 **Supporting Information**

808

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

815

816 **S2 Fig. Subpopulation purity of simulated and real CyTOF data.** (a) Subpopulation-specific
817 purity plot of 35-dimensional simulated data with 10 subpopulations. The blue points denote the
818 differences between the p-measures of the partition-based method (either d-PAC or b-PAC) and
819 flowMeans, while the red points denote the p-measure differences between the partition methods
820 and SPADE. The horizontal line at 0 means no difference between the methods. Most of the blue
821 and red points are above 0, indicating that the PAC generates purer subpopulations compared to
822 the ground truth. The two subplots are very similar, which means that d-PAC and b-PAC give
823 very similar p-measures. More precisely, the sum of differences between d-PAC and flowMeans
824 and d-PAC and SPADE are 0.85 and 1.09, respectively; and the overall difference between b-
825 PAC and flowMeans and b-PAC and SPADE are 0.84 and 1.08, respectively.

826 (b) Subpopulation-specific purity plot of the hand-gated CyTOF data. The same convention is
827 used as in (S2a Fig). Again, more blue and red points are above 0, indicating that the partition-
828 based methods generate purer subpopulations compared to the ground truth. There is a cluster of
829 points below 0 occurring in the middle of the plot, suggesting that flowMeans and SPADE
830 capture the mid-size subpopulations more similar to hand-gating than the partition-based
831 methods. More specifically, flowMeans does better (p-measure difference of 0.1 or better;
832 difference of less 0.1 is considered practically no difference) with finding subpopulations of
833 GMP, CD8 T cells, MEP, CD4 T cells (compared to d-PAC), and Plasma cells, while SPADE
834 does better with CD19+IgM- B cells, NK cells (compared to d-PAC), CD8 T cells, NKT cells,
835 Basophils, Short-Term HSC, and Plasma cells. However, overall, PAC has a much better
836 performance, as the absolute sum of points above 0 is higher than that of points below 0. More
837 precisely, the sum of differences between d-PAC and flowMeans and d-PAC and SPADE are
838 1.21 and 1.45, respectively; and the overall difference between b-PAC and flowMeans and b-
839 PAC and SPADE are 2.06 and 2.31, respectively. The difference table is provided in S1 Table.

840

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

845

846 **S4 Fig. Comparison between aligning cross-sample subpopulations by network, expression**

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

856

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

860

861 **S6 Fig. Comparison between network and means PAC-MAN.** (a) PAC-discovered
862 subpopulations are aggregated by MAN into clades; the number of PAC subpopulations/clades
863 for the network and means PAC-MAN approaches are plotted. (b) After aggregating shared
864 clades within samples, the number of shared clades for the entire dataset is plotted for the two
865 PAC-MAN approaches. c) Using the network approach results as basis, the clades with strong
866 agreement (high p-measures) with the means PAC-MAN approach are given. The shared clades
867 (present in more than one sample) are reliable candidates for future experiment to find cross-
868 sample relationships.

869

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

874

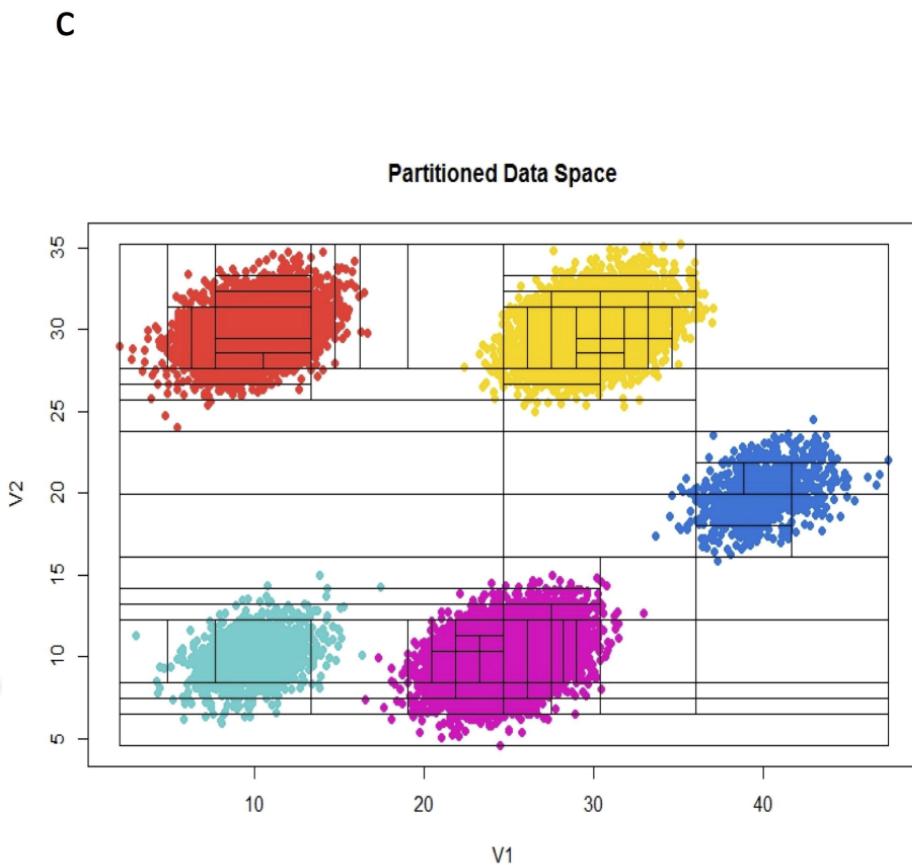
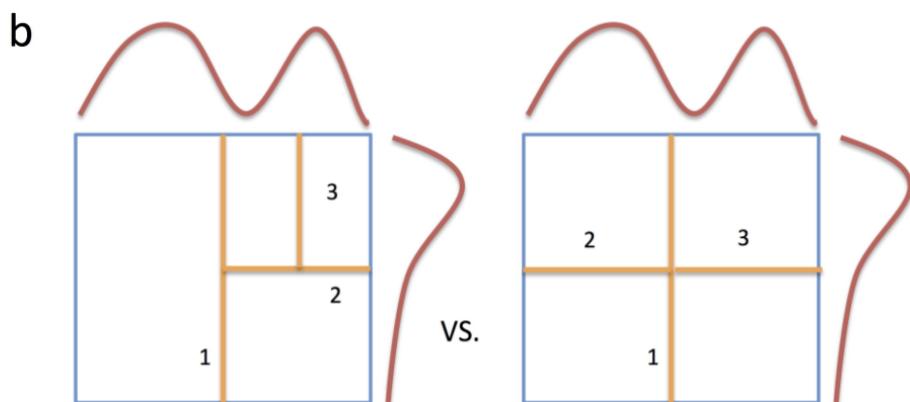
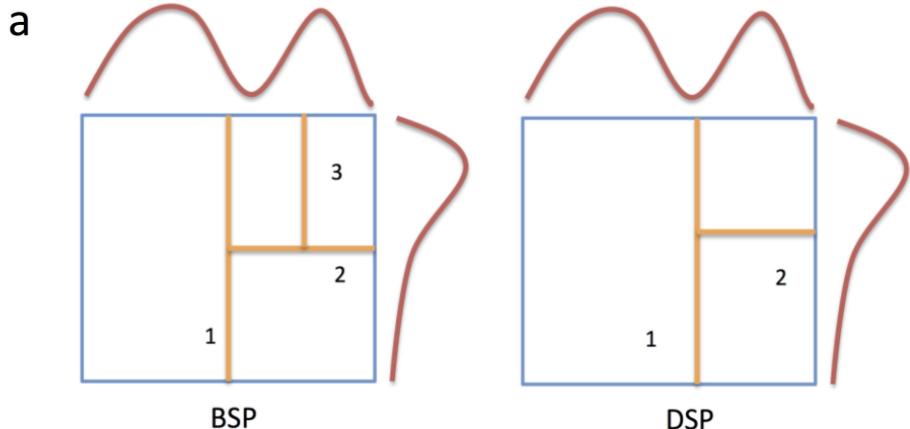
875 **S2 Table. Sample Sizes in Mouse Tissue CyTOF Dataset.** The numbers of cells in the samples
876 of Spitzer et al., 2015 CyTOF dataset. The data is from the C57BL6 mouse strain and a total of
877 ten tissue samples are present. The raw column shows the number of cells prior to filtering by
878 DNA and cisplatin values. The final cell counts are shown in the filtered file (3rd) column.

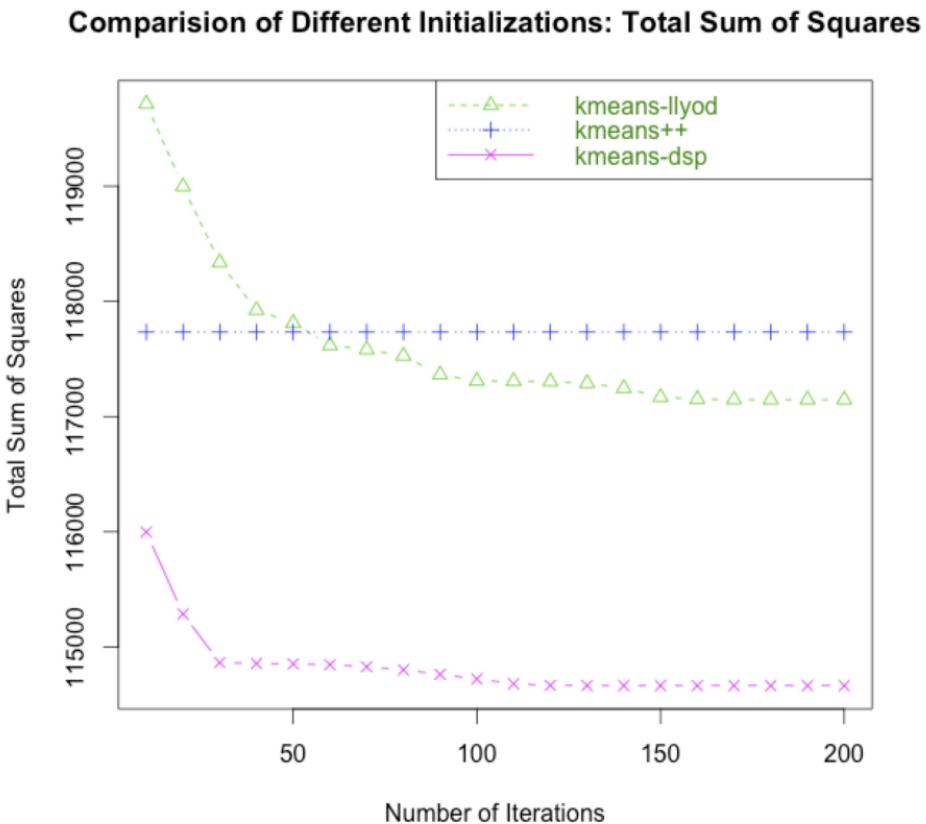
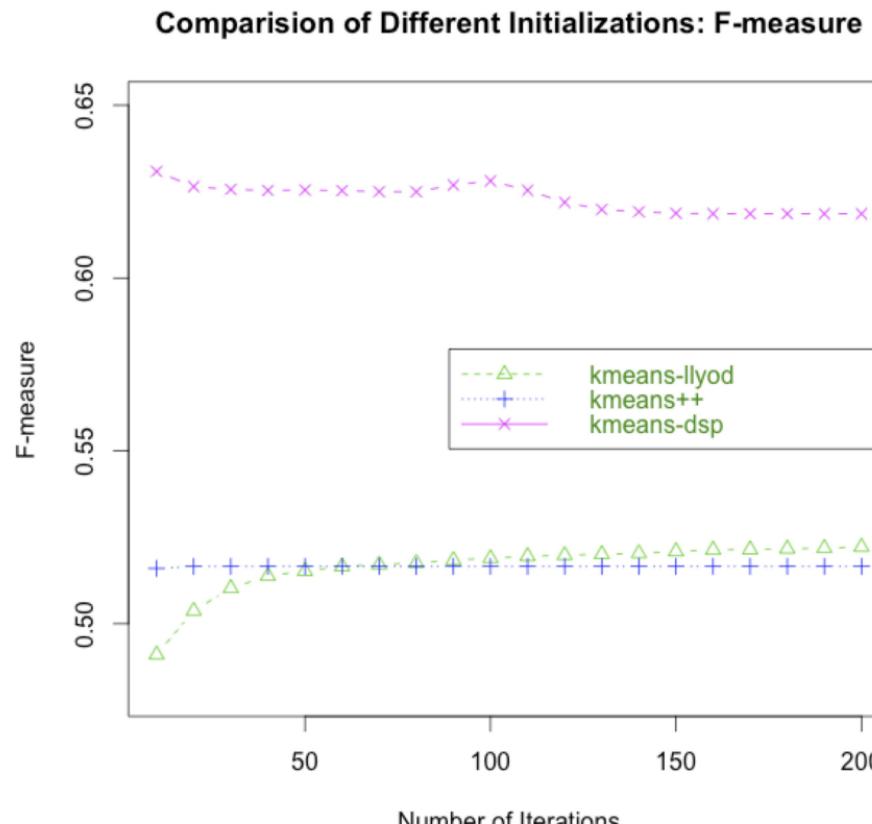
879

880 **S3 Table. PAC-MAN Subpopulation Characterization Output for Mouse Tissue CyTOF**
881 **Dataset.** The full set of annotated results, along with mean expressions, subpopulation
882 proportion and counts, are reported.

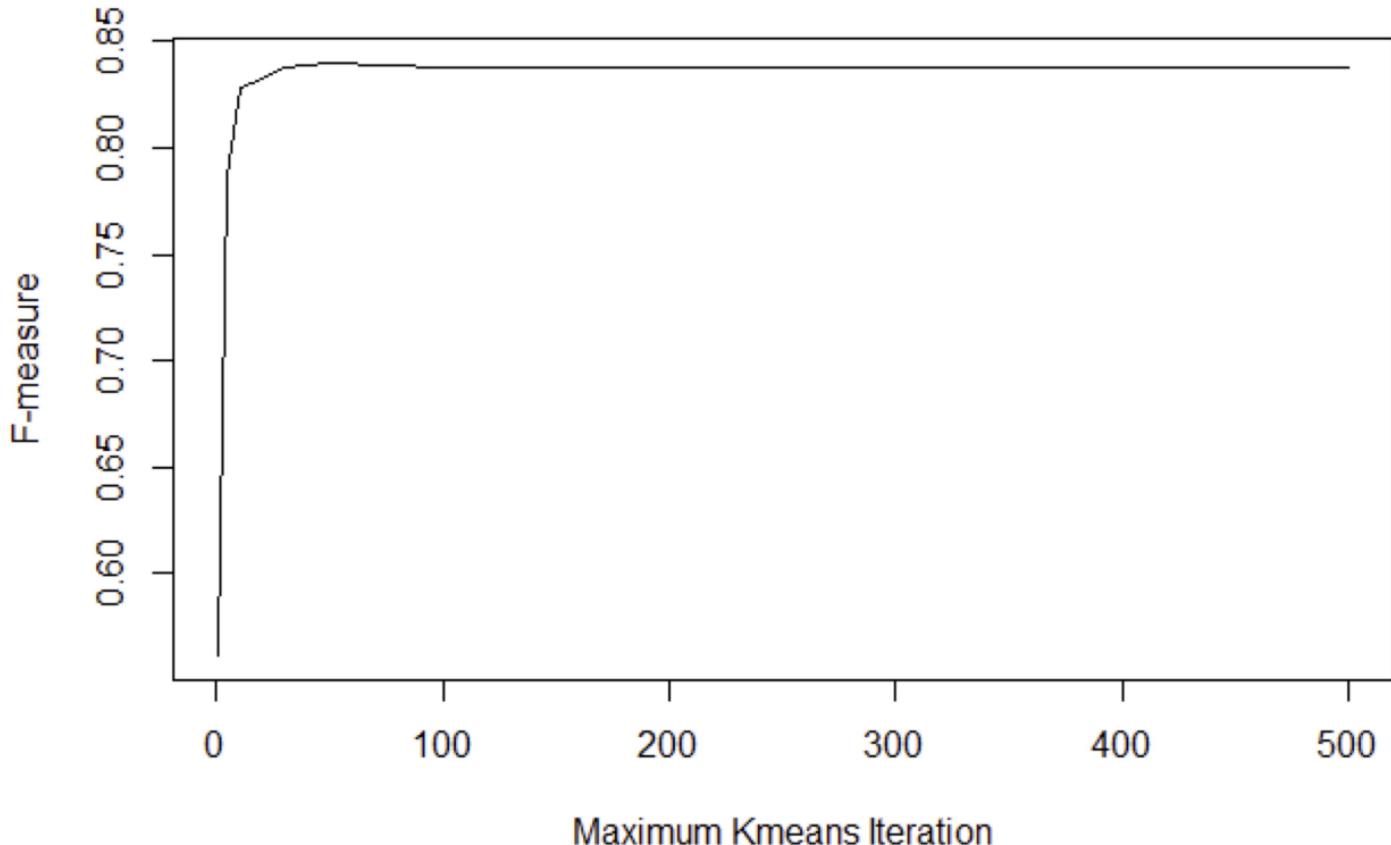
883

884 **S4 Table. Network-justified and means-justified p-measures for Alignments of PAC-**
885 **discovered Subpopulations.** The PAC-discovered subpopulations were mapped as clades in
886 both the network and means PAC-MAN approaches. The p-measures were calculated for the
887 cases 1) network approach mapping as the basis and 2) means approach mapping as the basis.
888 The comparison is the same in principle to the comparison of labels for clustering methods. The
889 results are ordered by p-measures.

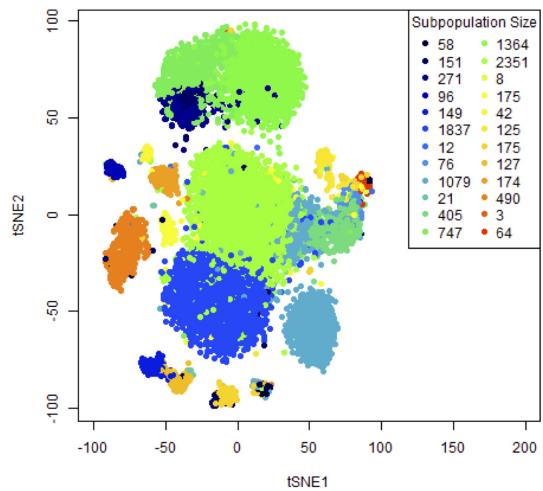


a**b**

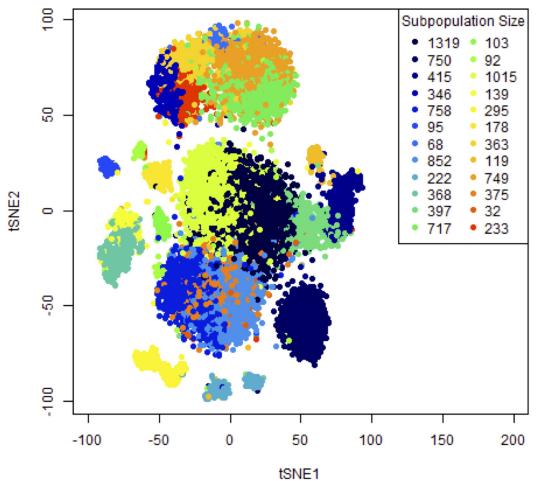
Convergence of Clustering Results to Handgate Labels



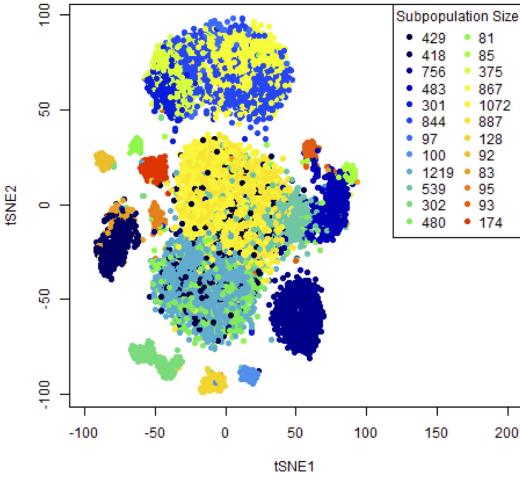
Handgate



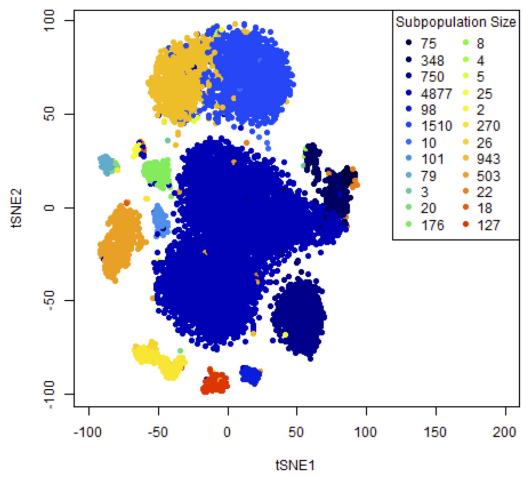
kmeans 5000 iterations



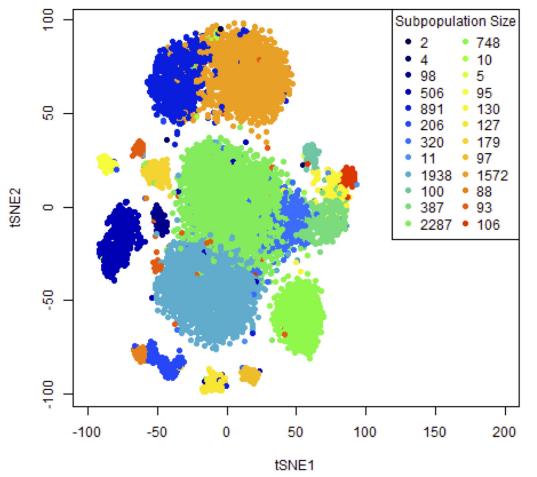
SPADE



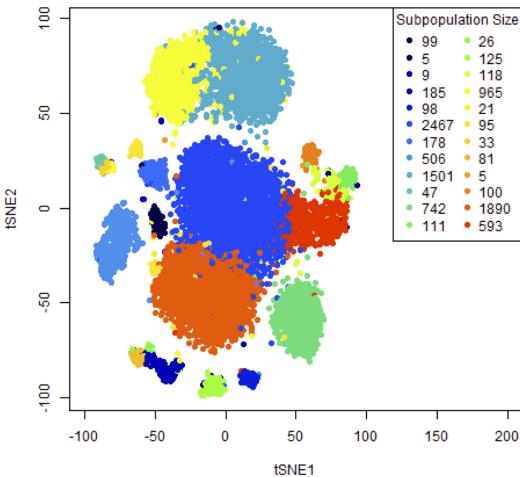
flowMeans

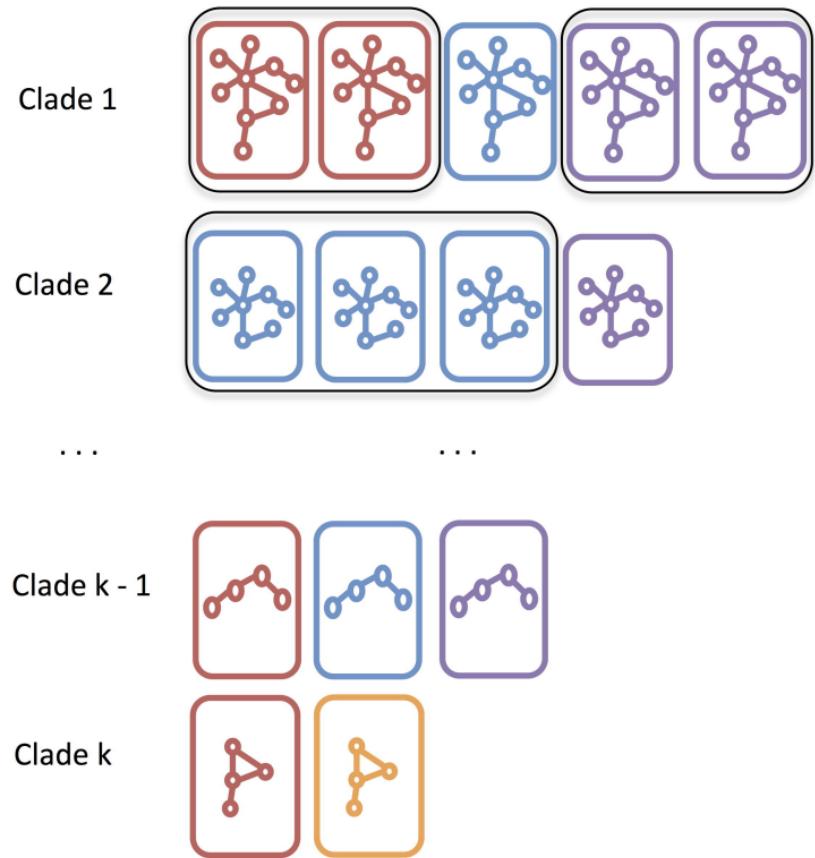
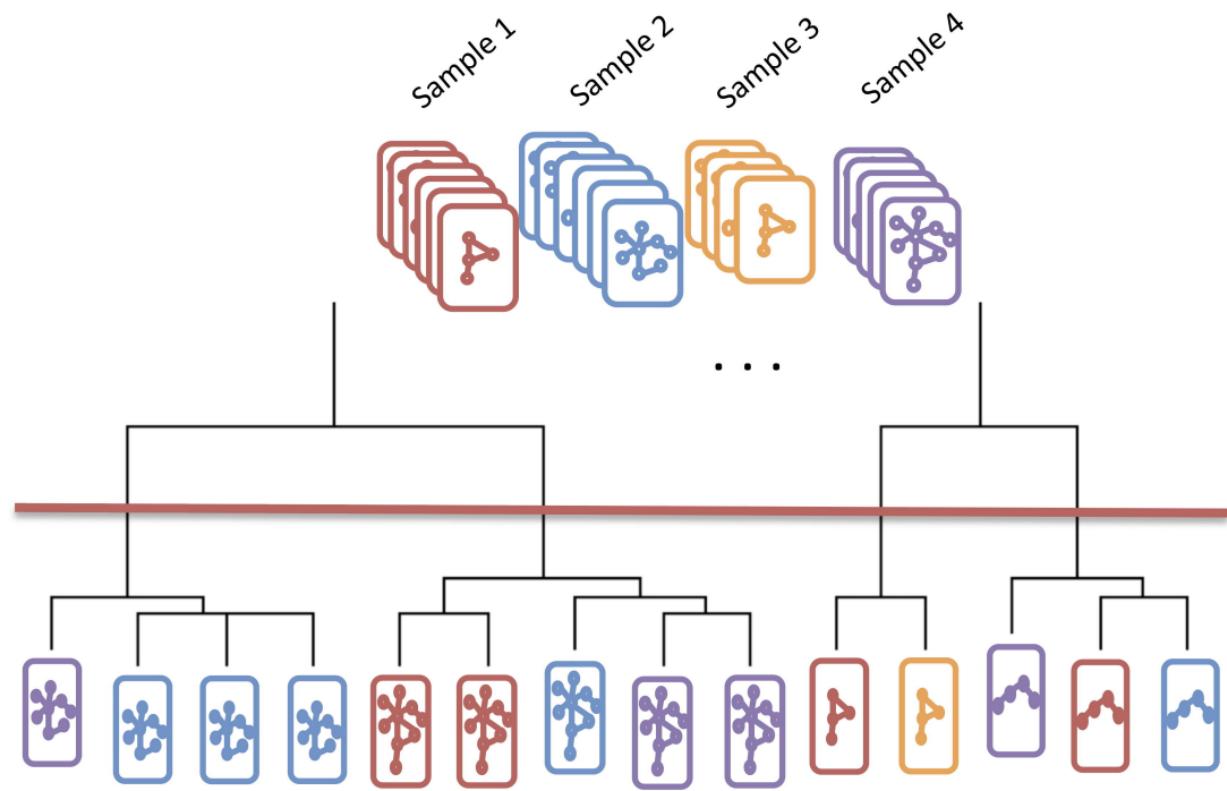


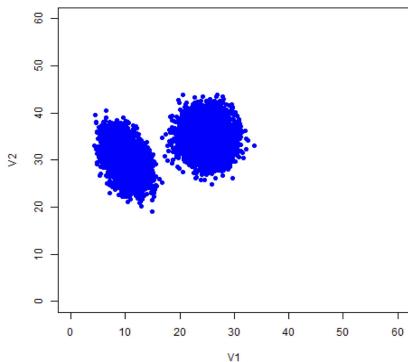
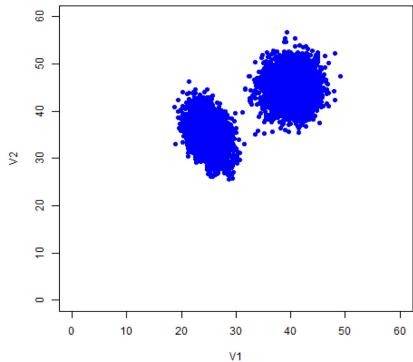
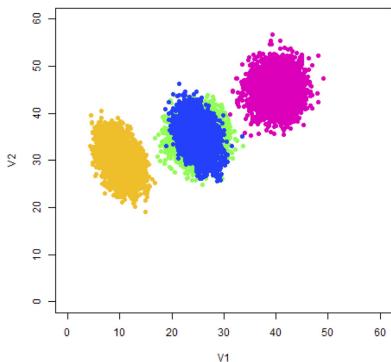
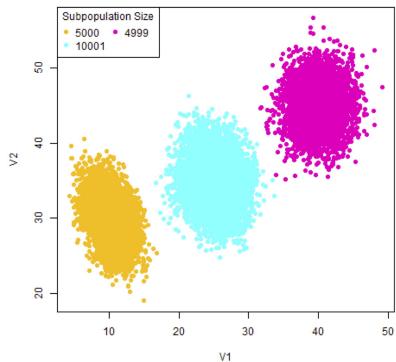
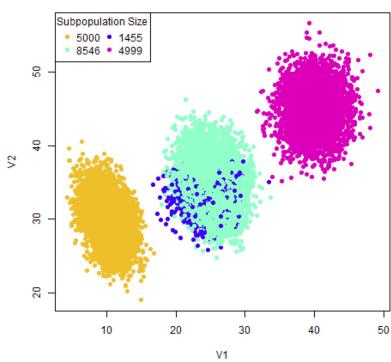
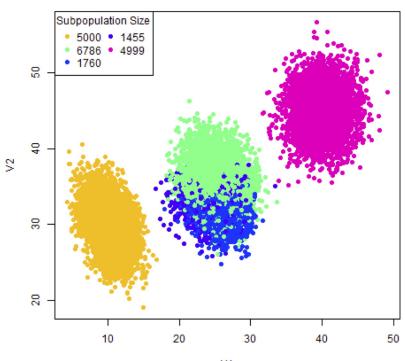
b-PAC

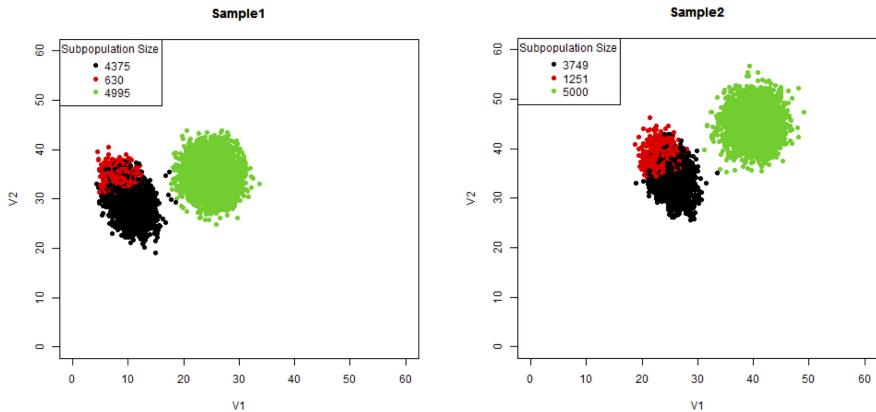
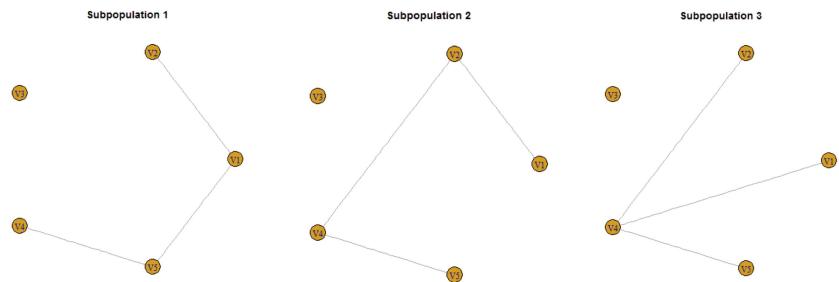
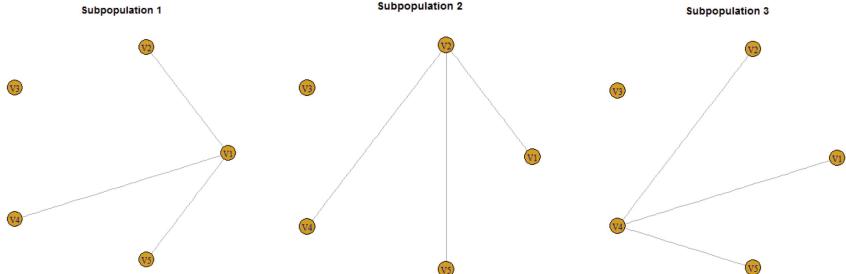


d-PAC

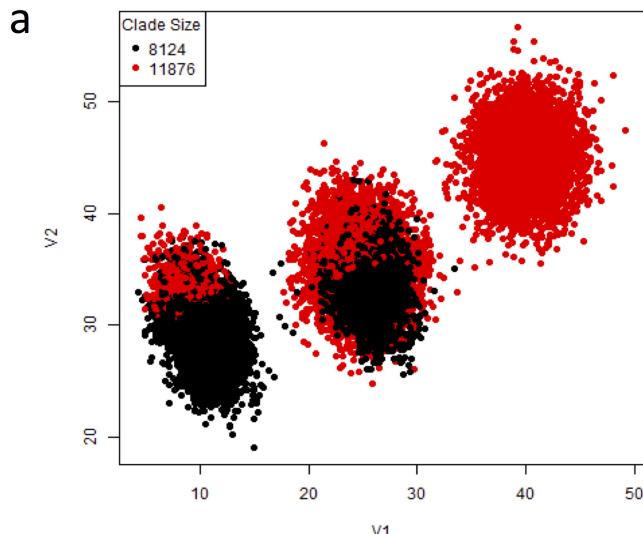




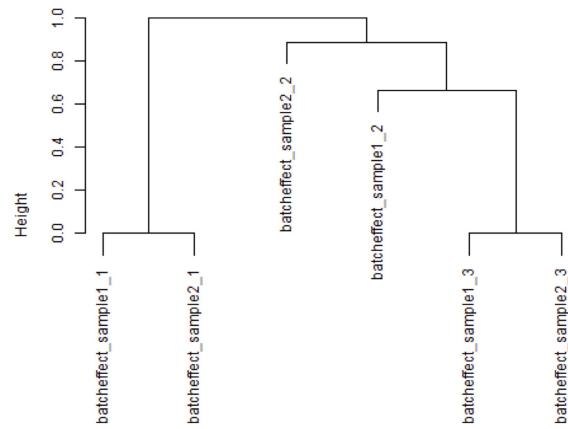
a**Sample1****Sample2****b****All Samples Merged****Merged Samples****c****Merged Samples****Merged Samples**

a**b****c**

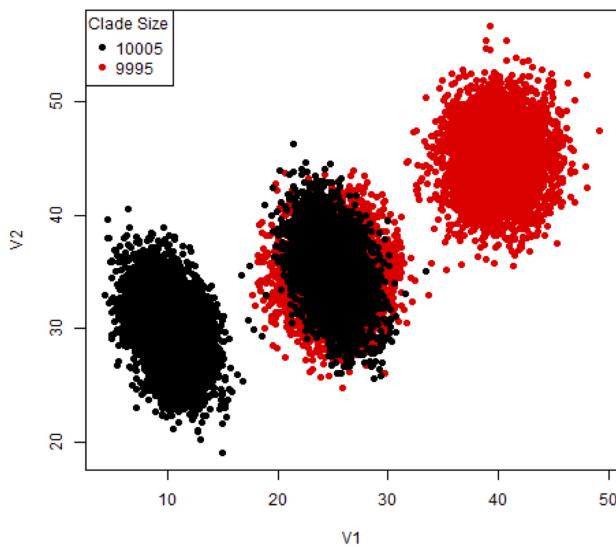
All Samples, Rep Clades by Networks



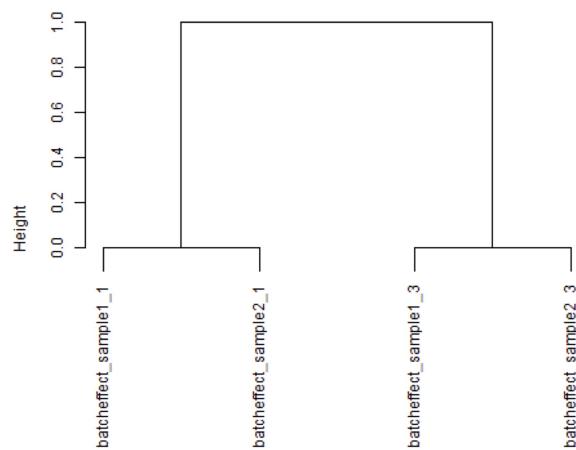
Cluster Dendrogram



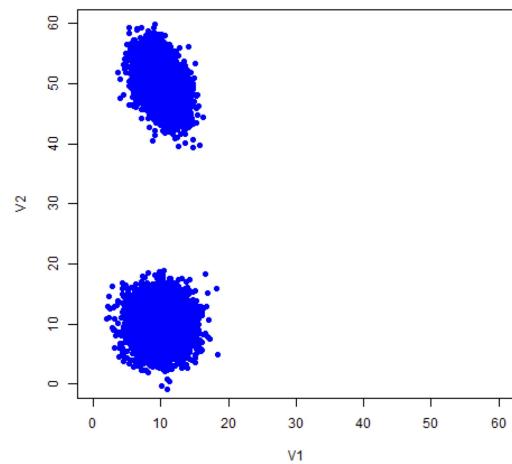
All Samples, Rep Clades by GeneExp and Networks



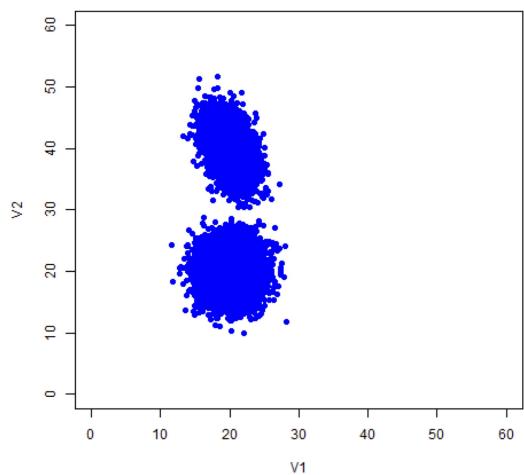
Cluster Dendrogram



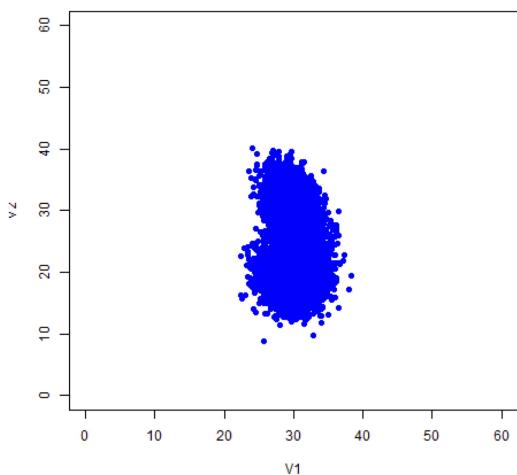
Sample1



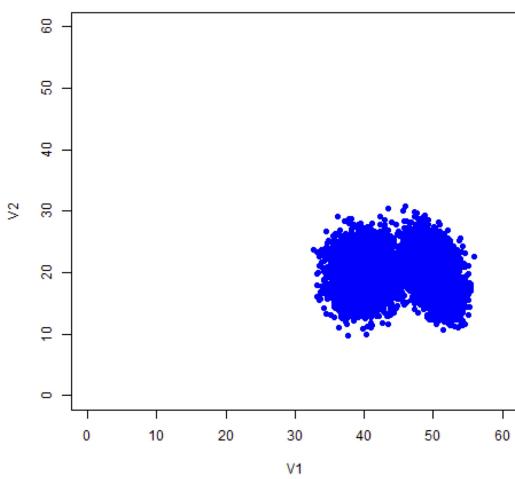
Sample2



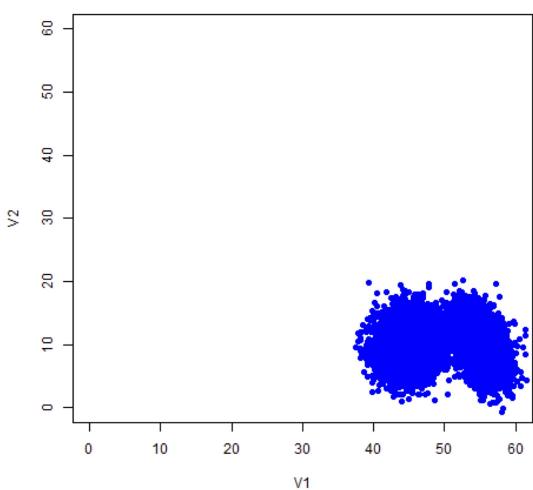
Sample3



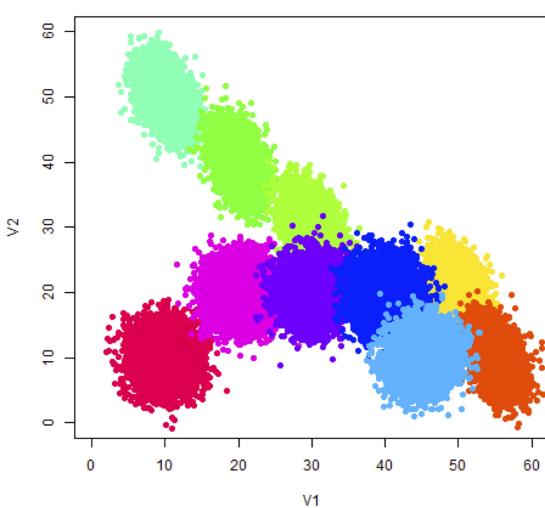
Sample4

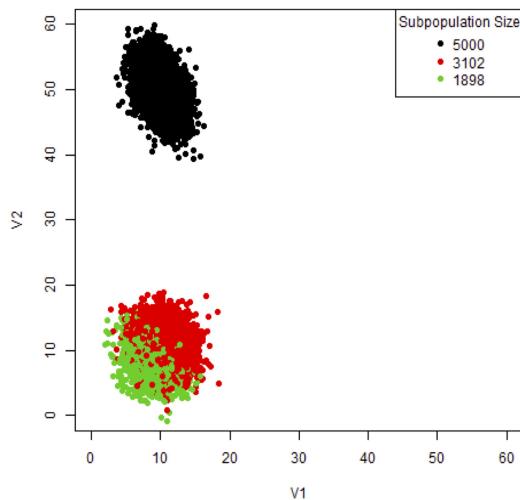
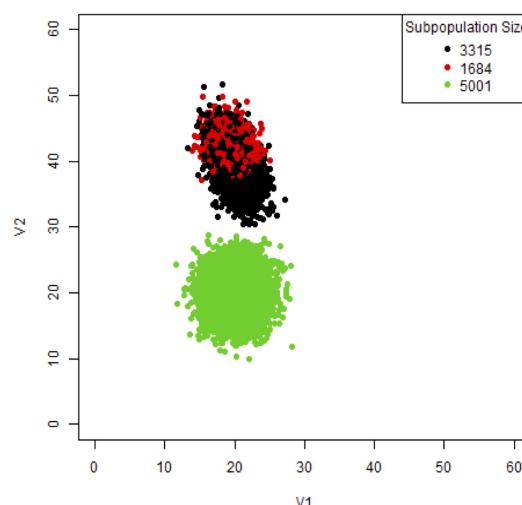
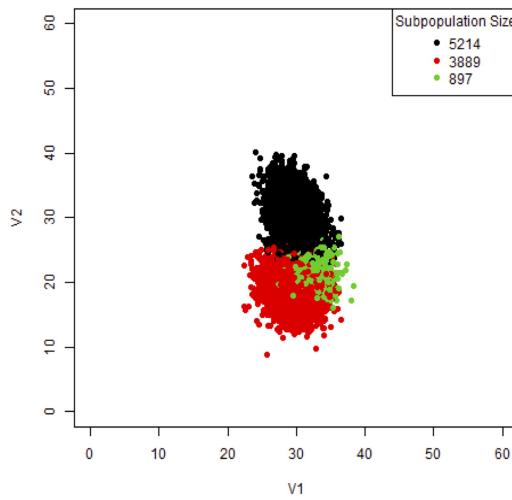
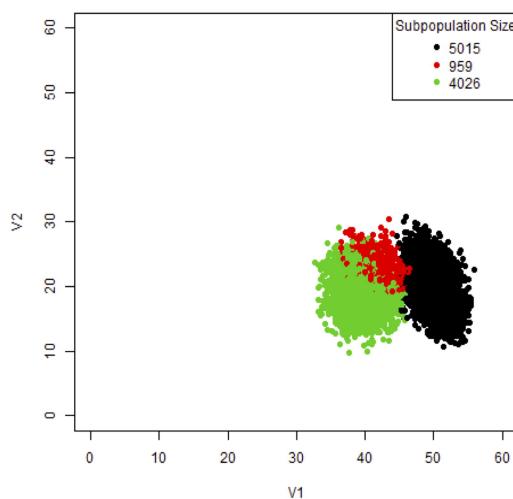
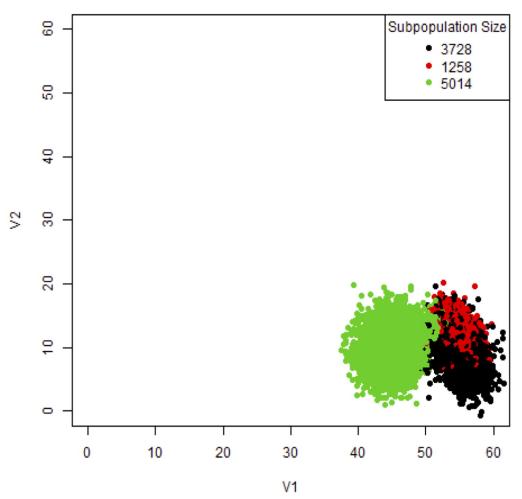


Sample5

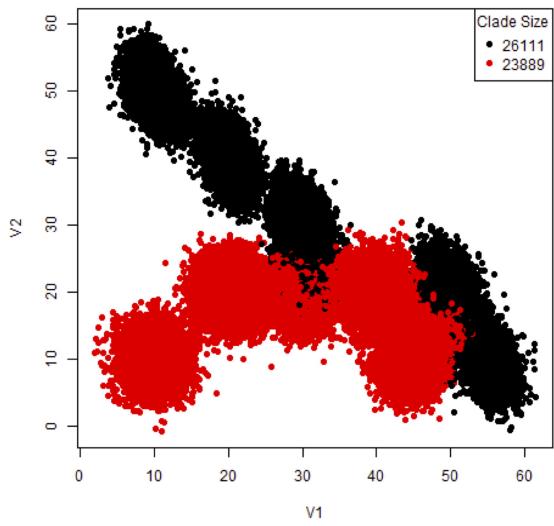


All Samples Merged

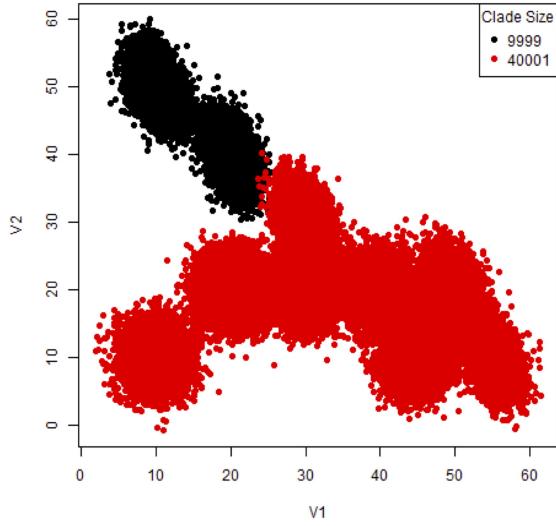


Sample1**Sample2****Sample3****Sample4****Sample5**

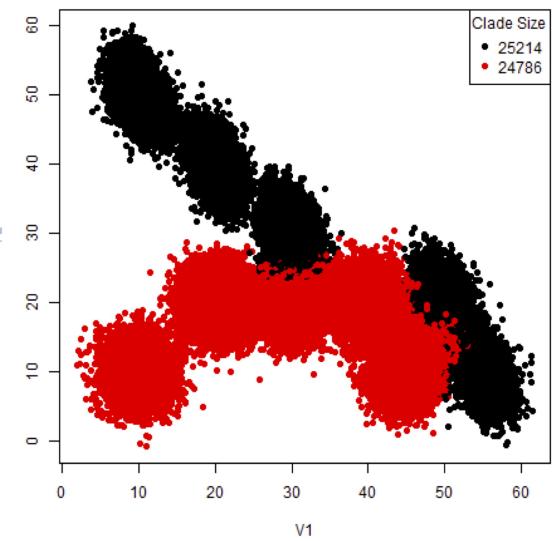
All Samples, Rep Clades by Networks



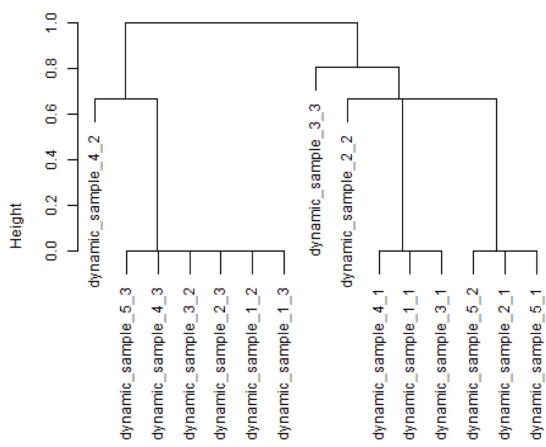
All Samples, Rep Clades by Gene Expression



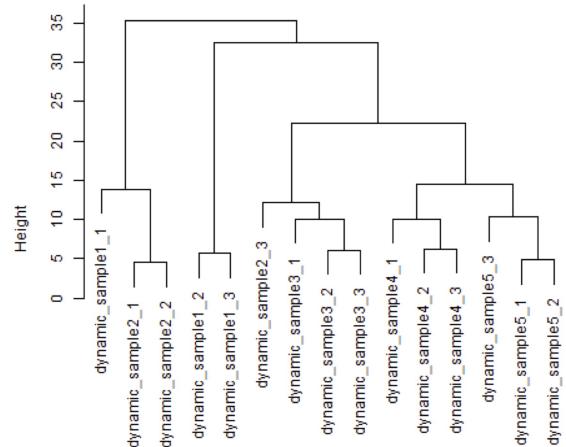
All Samples, Rep Clades by GeneExp and Networks



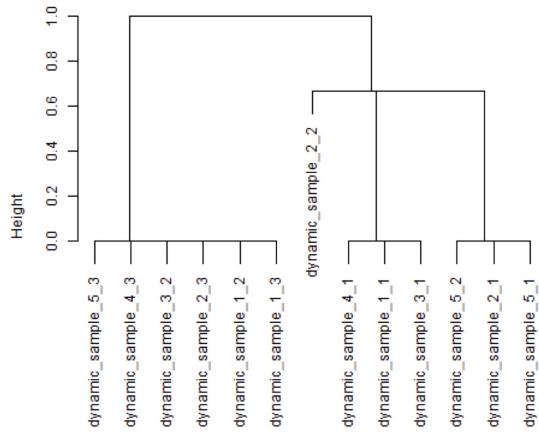
Cluster Dendrogram

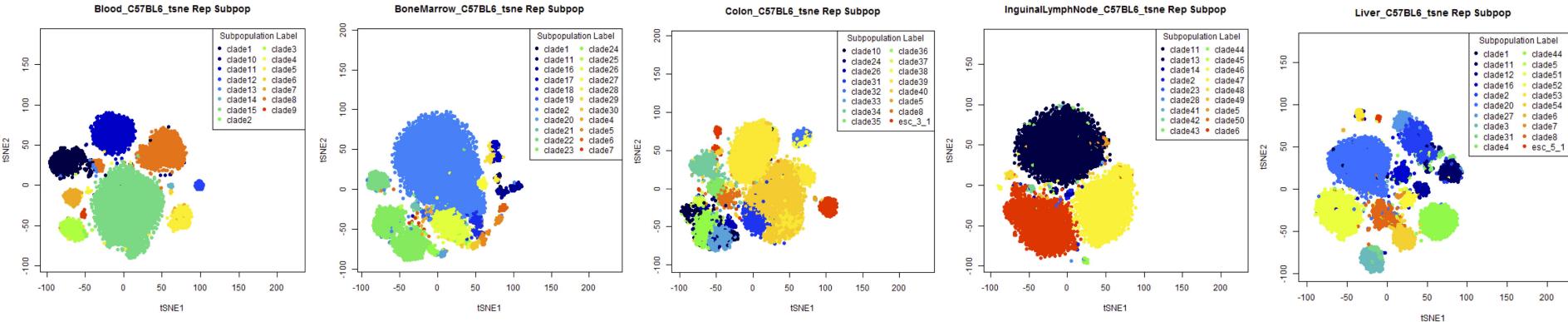
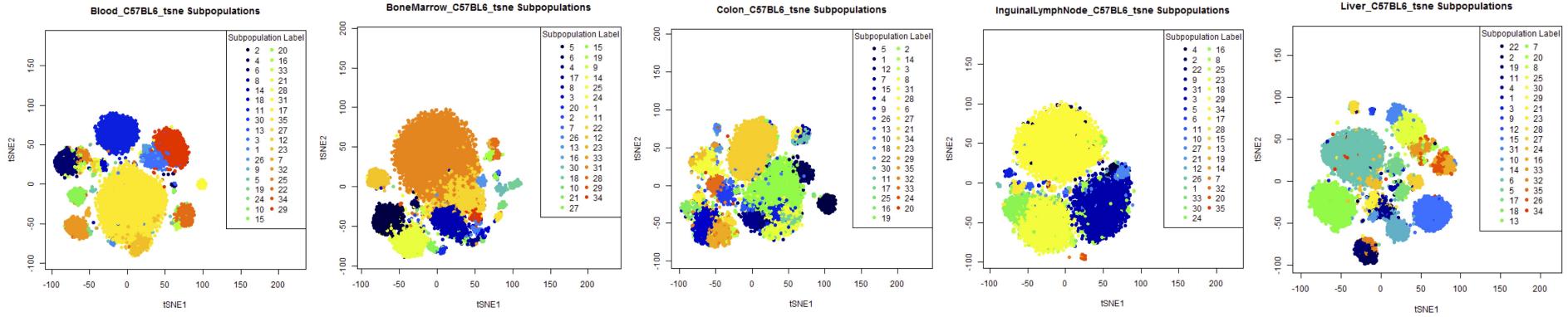


Cluster Dendrogram

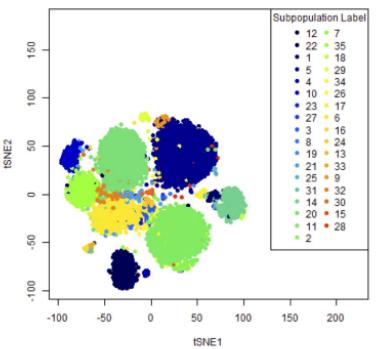


Cluster Dendrogram

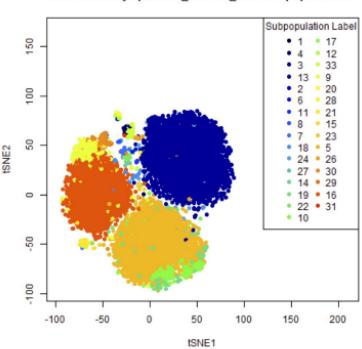




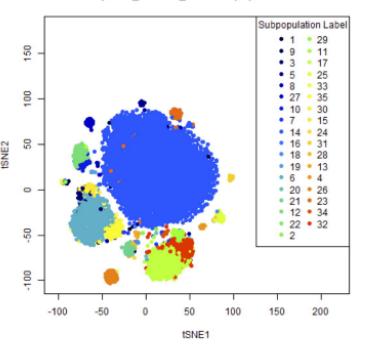
Lung_C57BL6_tsne Subpopulations



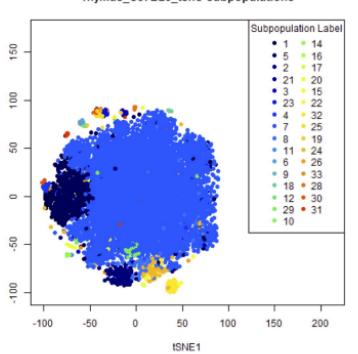
MesentericLymphNode_C57BL6_tsne Subpopulations



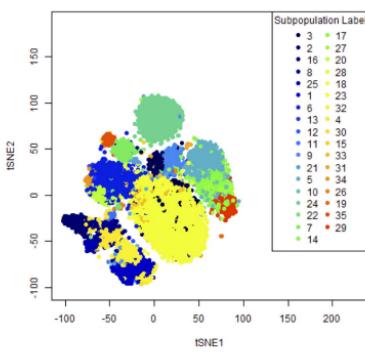
Spleen_C57BL6_tsne Subpopulations



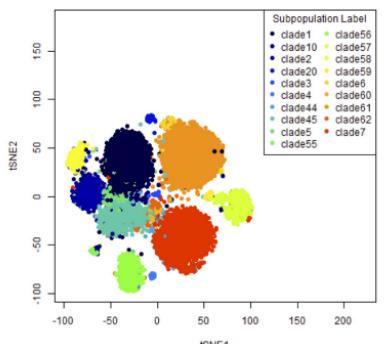
Thymus_C57BL6_tsne Subpopulations



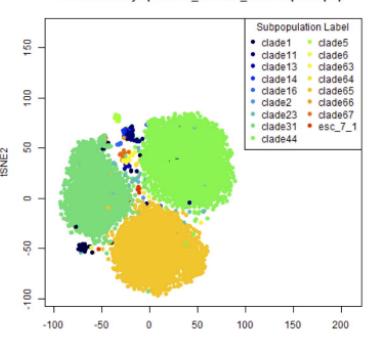
SmallIntestine_C57BL6_tsne Subpopulations



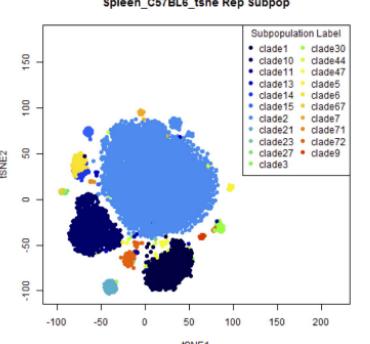
Lung_C57BL6_tsne Rep Subpop



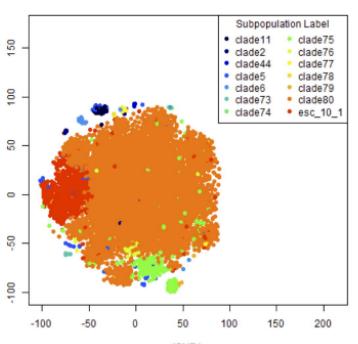
MesentericLymphNode_C57BL6_tsne Rep Subpop



Spleen_C57BL6_tsne Rep Subpop



Thymus_C57BL6_tsne Rep Subpop



SmallIntestine_C57BL6_tsne Rep Subpop

