TITLE: A cluster-aware, weighted ensemble clustering method for cell-type detection

Daphne Tsoucas[1,2,*], Guo-Cheng Yuan[1,2]

1. Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA

2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

*corresponding author

Emails: dtsoucas@g.harvard.edu; gcyuan@jimmy.harvard.edu

ABSTRACT

Single-cell analysis is a powerful tool for dissecting the cellular composition within a tissue or organ. However, it remains difficult to detect rare and common cell types at the same time. Here we present a new computational method, called GiniClust2, to overcome this challenge. GiniClust2 combines the strengths of two complementary approaches, using the Gini index and Fano factor, respectively, through a cluster-aware, weighted ensemble clustering technique. GiniClust2 successfully identifies both common and rare cell types in diverse datasets, outperforming existing methods. GiniClust2 is scalable to very large datasets.

Keywords

Clustering; Consensus clustering; Ensemble clustering; Single-cell; scRNA-seq; Gini index; Rare cell type

INTRODUCTION

Genome-wide transcriptomic profiling has served as a paradigm for the systematic characterization of molecular signatures associated with biological functions and disease-related alterations, but traditionally this could only be done using bulk samples that often contain significant cellular heterogeneity. The recent development of single-cell technologies has enabled biologists to dissect cellular heterogeneity within a cell population. Such efforts have led to an increased understanding of cell-type composition, lineage relationships, and

mechanisms underlying cell-fate transitions. As the throughput of single-cell technology increases dramatically, it has become feasible not only to characterize major cell types, but also to detect cells that are present at low frequencies, including those that are known to play an important role in development and disease, such as stem and progenitor cells, cancer-initiating cells, and drug-resistant cells [1, 2].

On the other hand, it remains a computational challenge to fully dissect the cellular heterogeneity within a large cell population. Despite the intensive effort in method development [3-8], significant limitations remain. Most methods are effective only for detecting common cell populations, but are not sensitive enough to detect rare cells. On the other hand, a number of methods have been developed to specifically detect rare cells [9-12], but the features used in these methods are distinct from those distinguishing major populations. Existing methods cannot satisfactorily detect both large and rare cell populations. A naïve approach combining features that are either associated with common or rare cell populations fails to characterize either type correctly, as a mixed feature space will dilute both common and rare cell-type-specific biological signals, an unsatisfactory compromise.

To overcome this challenge, we have developed a new method, GiniClust2, to integrate information from complementary clustering methods using a novel ensemble approach. Instead of averaging results from individual clustering

methods, as is traditionally done, GiniClust2 selectively weighs the outcomes of each model to maximize the methods' respective strengths. We show that this cluster-aware weighted ensemble approach can accurately identify both common and rare cell types and is scalable to large datasets.

RESULTS

Overview of the GiniClust2 method

An overview of the GiniClust2 pipeline is shown in **Fig. 1**. We begin by independently running both a rare cell-type detection method and a common cell-type detection method on the same data set (**Fig. 1a**). In a previous study [11], we showed that different strategies are optimal for identifying genes associated with rare cell types than for common ones. Whereas the Fano factor is a valuable metric for capturing differentially expressed genes specific to common cell types, the Gini index is much more effective for identifying genes that are associated with rare cells [11]. Therefore, we were motivated to develop a new method that combines the strengths of these two approaches. To facilitate a concrete discussion, here we choose GiniClust as the Gini-index based method and k-means as the Fano-factor based method. However, the same approach can be used to combine any other clustering methods with similar properties. We call this new method GiniClust2.

Our goal is to consolidate these two differing clustering results into one consensus grouping. The output from each initial clustering method can be

93 represented as a binary-valued connectivity matrix, M_{ij}, where a value of one
94 indicates cells i and j belong to the same cluster (**Fig. 1b**). Given each method's
95 distinct feature space, we find that GiniClust and k-means tend to emphasize the
96 accurate clustering of rare and common cell types, respectively, at the expense
97 of their complements. To optimally combine these methods, a consensus matrix
98 is calculated as a cluster-aware, weighted sum of the connectivity matrices (**Fig.**
99 **1b**). Higher weights are assigned to rare clusters over common clusters for
100 GiniClust, whereas higher weights are assigned to common clusters over rare
101 clusters for k-means (**Fig. 1c**). For simplicity, the weighting functions are
102 modeled as logistic functions which can be specified by three tunable
103 parameters: μ is the cluster size at which GiniClust and Fano-factor-based
104 clustering methods have the same detection precision, s represents how quickly
105 GiniClust loses its ability to detect rare cell types, and f represents the
106 importance of the Fano cluster membership in determining the larger context of
107 the membership of each cell. The values of these parameters are estimated
108 from the data (**Supplemental Information**). An additional round of clustering is
109 then applied to the consensus matrix to identify both common and rare cell
110 clusters. The mathematical details are described in the **Materials and Methods**
111 section.
112
113 **Accurate detection of both common and rare cell types in a simulated**
114 **dataset**

115 We started by evaluating the performance of GiniClust2 using a simulated
116 scRNA-seq dataset, which contains two common clusters (of 2000 and 1000
117 cells, respectively) and four rare clusters (of 10, 6, 4, and 3 cells, respectively)
118 (**Materials and Methods, Fig. 2a**). We first applied GiniClust and k-means
119 independently to cluster the cells. As expected, GiniClust correctly identifies all 4
120 rare cell clusters, but merges the two common clusters into a single large cluster
121 (**Fig. 2b**). In contrast, k-means (with k=2) accurately separates the two common
122 clusters, while lumping together all four rare cell clusters into the largest group
123 (**Fig. 2b**). Increasing k does not improve the ability to detect rare cells, but rather
124 divides each common cluster into smaller clusters, indicating an intrinsic
125 limitation of selecting gene features using the Fano factor (**Supplemental Fig.**
126 **1a**).
127
128 We next used the GiniClust2 weighted ensemble step to combine the results
129 from GiniClust and k-means. Of note, all six cell clusters are perfectly
130 recapitulated by GiniClust2 (**Fig. 2b**), suggesting that GiniClust2 is indeed
131 effective for detecting both common and rare cell clusters. To aid visualization,
132 we created a composite tSNE plot, projecting the cells into a three-dimensional
133 space based on a combination of a two-dimensional Fano-based tSNE map and
134 a one-dimensional Gini-based tSNE map (**Fig. 2c**). A three-dimensional space is
135 required because, although the Fano-based dimensions are able to clearly
136 separate the two common clusters, the rare clusters are overlapping and cannot

be fully discerned. The third (Gini) dimension results in complete separation of the rare clusters.

Since the number of common clusters is unknown in advance, we also tested the robustness of GiniClust2 with respect to other choices of k. We found that setting k=3 provides the same final clustering, while further increase results in poorer performance by splitting of the larger clusters (**Supplemental Fig. 1b**). By default, the value of k was chosen using the gap statistic [13]. However, this may not be optimal in various cases [14], therefore additional exploration is often needed to select the optimal value for k.

For comparison, we evaluated the performance of two unweighted ensemble clustering methods. First, we used the cluster-based similarity partitioning algorithm (CSPA) [15] to combine the GiniClust and k-means (k=2) clustering results. The consensus clustering splits the common clusters into six subgroups, whereas cells in the four rare clusters are assigned to one of two clusters shared with the largest common cell group (**Fig. 2b**). Without guidance, the consensus clustering treats all clustering results equally and attempts to resolve any inconsistency via suboptimal compromise. The second method we considered, known as SC3 [4], is specifically designed for single-cell analysis. This method performs an unweighted ensemble of k-means clusterings for various parameter choices without specifically targeting rare cell detection. Regardless of the specific parameter choices, k-means cannot resolve the rarest clusters, and the

160 final ensemble clustering splits the largest group into three and differentiates only
161 one of the four rare clusters (**Fig. 2b**). These analyses suggest that our cluster-
162 aware, weighted ensemble approach is important for optimally combining the
163 strengths of different methods.
164
165 We also compared the performance of GiniClust2 with other rare cell type
166 detection methods. In particular, we compared with RaceID2 [10], which is an
167 improved version of RaceID [9] developed by the same group. For fair
168 comparison, we considered k=2, the exact number of common cell clusters.
169 RaceID2 over-estimated the number of clusters, and split both common and rare
170 cells clusters into smaller subclusters (**Fig. 2b**). Using other choices of k led to
171 even poorer outcomes. This tendency of over-clustering was consistent with our
172 previous observations [11].
173
174 **Robust identification of rare cell types over a wide range of proportions**
175 In order to evaluate the performance of GiniClust2 on analyzing real
176 scRNA-seq datasets, we focused on one of the largest public scRNA-seq
177 datasets generated by 10X Genomics [16]. The dataset consists of
178 transcriptomic profiles of about 68,000 peripheral blood mononuclear cells
179 (PBMCs) [16], which were classified into 11 subpopulations based on
180 transcriptomic similarity with purified cell-types. It was noted that the
181 transcriptomic profiles of several subpopulations are nearly indistinguishable [16].
182

To reduce the effects of stochastic variation and technical artifacts, we started by considering only a subset of cell types that are distinct from one another. In particular, we focused on three large subpopulations: CD56+ natural killer (NK) cells, CD14+ Monocytes, and CD19+ B cells (**Fig. 3a**). To systematically compare the ability of different methods in detecting both common and rare cell types, we created a total of 120 random subsamples that mix different cell types at various proportions (**Supplementary Table 1**), with the rare cell type (NK cells) proportions ranging from 0.4% to 11.6% (see **Materials and Methods** for details).

We applied GiniClust2 and comparable methods to the down-sampled datasets generated above. Each method was evaluated based on its ability to detect each cell type using three Matthews correlation coefficients (MCC) [17] (**Fig. 3b**). The MCC is a metric that quantifies the overall agreement between two binary classifications, taking into account both true and false positives and negatives. The MCC value ranges from -1 to 1, where 1 means a perfect agreement between a clustering and the reference, 0 means the clustering is as good as a random guess, and -1 means a total disagreement between a clustering and the reference.

RaceID2 is the best method for detecting the rare NK cell type at a frequency of 1.6% or lower, and GiniClust2 is the next best method. While the performance of GiniClust degrades as the “rare” cell type becomes more abundant, k-means

206 becomes more powerful in such cases, and results in the near-optimality of
207 GiniClust2 over a wide range of rare-cell proportions. For the common groups,
208 k-means tends to perform better, but only if the parameter is chosen correctly.
209 For example, k-means with k=4 systematically splits the largest Monocyte group
210 and leads to a relatively low MCC value. The low specificity of RaceID2 leads to
211 false rare cell types within the common groups, and consistently low MCC
212 values. On the other hand, GiniClust2 consistently ranks at or near the top
213 among all methods for these common groups. Taken together, these
214 comparative results suggest that GiniClust2 reaches a good balance for
215 detecting both common and rare clusters.
216
217 **Detection of rare cell types in differentiating mouse embryonic stem cells**
218 To test if GiniClust2 is useful for detecting previously unknown, biologically
219 relevant cell types, we analyzed a published dataset associated with leukemia
220 inhibitory factor (LIF) withdrawal induced mouse embryonic stem cell (mESC)
221 differentiation [18]. Previously, we applied GiniClust to analyze a subset
222 containing undifferentiated mESCs, and identified a rare group of Zscan4-
223 enriched cells [11]. As expected, these rare cells were rediscovered using
224 GiniClust2.
225
226 In this study, we focused on the cells assayed on Day 4 post LIF withdrawal, and
227 tested if GiniClust2 might uncover greater cellular heterogeneity than previously
228 recognized. GiniClust2 identified two rare clusters consisting of 5 and 4 cells

229 respectively, corresponding to 1.80% and 1.44% of the entire cell population.
230 The first group contains 25 differentially expressed genes when compared to the
231 rest of the cell population (MAST likelihood ratio test p-value<1e-5, fold
232 change>2), including known primitive endoderm (PrEn) markers such as *Col4a1*,
233 *Col4a2*, *Lama1*, *Lama2*, and *Ctsl*. These genes are also associated with high
234 Gini index values. Overall there is a highly significant overlap between
235 differentially expressed and high Gini genes (Fisher exact test p-value<1e-18).
236 The second group contains 10 differentially expressed genes (MAST likelihood
237 ratio test p-value<1e-5, fold change>2), including maternally imprinted genes
238 *Rhox6*, *Rhox9*, and *Sct*, all of which are also high Gini genes. Once again there
239 is a significant overlap between differentially expressed and high Gini genes
240 (Fisher exact test p-value<1e-12). Although these clusters were detected in the
241 original publication [18], this was achieved based on *a priori* knowledge of
242 relevant markers. Here, the strength of GiniClust2 is that it can identify these
243 clusters without previous knowledge.
244
245 In addition, GiniClust2 identified 2 common clusters. The first group specifically
246 expresses a number of genes related to cell growth and embryonic development,
247 including *Pim2*, *Tdgf1*, and *Tcf15* (MAST likelihood ratio test p-value<1e-5, fold
248 change>2), indicating it corresponds to undifferentiated stem cells. The second
249 group is strongly associated with a number of genes related to epiblast cells,
250 including *Krt8*, *Krt18*, *S100a6*, *Tagln*, *Actg1*, *Anxa2*, and *Flnc* (MAST likelihood
251 ratio test p-value<1e-5, fold change>2), suggesting this group corresponds to an

252 epiblast-like state. Of note, 114 of the 128 genes (Fisher exact test p-value<1e-
253 88) specifically expressed in this group were selected as high Fano-factor genes,
254 confirming the utility of Fano factor in detecting common cell-types. Both
255 populations were discovered in the original publication [18]. The dissimilarity
256 between these cell types is evident in the heatmap (**Fig. 4a**) and composite tSNE
257 plot (**Fig. 4b**).
258
259 For comparison, we applied RaceID2 to analyze the same dataset. Unlike
260 GiniClust2, RaceID2 broke each cluster into multiple subclusters, and failed to
261 identify the rare cell clusters (**Fig. 4c**). With k=2, RaceID2 found a total of 11
262 clusters. Clusters 1, 2, 4, and 9 display an epiblast-like signature, clusters 7, and
263 10 overexpress genes relating to maternal imprinting, and clusters 8 and 11
264 correspond to PrEn cells. From these results it appears that RaceID2 has
265 difficulty in differentiating rare, biologically meaningful cell types from outliers.
266
267 **Scalability to large data sets**
268 With the rapid development of single-cell technologies, it has become
269 feasible to profile thousands or even millions of transcriptomes at single-cell
270 resolution. Thus, it is desirable to develop scalable computational methods for
271 single-cell data analysis. As a benchmark, we applied GiniClust2 to analyze the
272 entire 68k PBMC data set [16] described above to uncover hidden cell types. The
273 complete analysis took 2.3 hours on a computing cluster with 56 CPUs, 28 cores,
274 and 660 GB memory. For comparison, RaceID2 analysis could not be completed

275 for this large dataset. One possible explanation is this method may be limited to
276 handling data sets with less than 65,536 data points due to an intrinsic restriction
277 in R.
278
279 Our analysis identified 9 common clusters and two rare clusters (**Fig. 5a**), In
280 general, the results of GiniClust2 and k-means are similar; both agree well with
281 the reference cell types (**Fig. 5b,** NMI = 0.540 for GiniClust2, NMI = 0.553 for k-
282 means). Most of the discrepancy between the clustering results and reference
283 labels are associated with T-cell subtypes. As noted by the original authors [16],
284 these subtypes are difficult to separate because they share similar gene
285 expression patterns and biological functions. The common clusters detected by
286 GiniClust2 and k-means express marker genes known to be specific to the cell
287 types represented in the reference [19] (**Fig. 5c**).
288
289 With respect to rare cell types, our first group contains a homogeneous and
290 visually distinct subset of 171 of 262 total CD34+ cells (Cluster 2, **Fig. 5a**). This
291 cluster was partially detectable using k-means, although it was partially mixed
292 with major clusters. The second rare cell cluster is previously unrecognized
293 (Cluster 3, **Fig. 5a**). This cluster contains 118 cells (0.17%) within a large set of
294 5433 immune cells with similar gene expression patterns. Among these 118
295 cells, 101 cells are classified as monocytes, whereas 16 are classified as
296 dendritic cells, and 1 is classified as a CD34+ cell. Differential expression
297 analysis (MAST likelihood ratio test p-value<1e-5, fold change>2) identified 187

genes that are specifically expressed in this cell cluster, including a number of genes associated with tolerogenic properties, such as *Ftl*, *Fth1*, and *Cst3* [20], suggesting these cells may be associated with elevated immune response and metabolism. Taken together, these results strongly indicate the utility of GiniClust2 in analyzing large single-cell datasets.

DISCUSSION AND CONCLUSIONS

According to the "no free lunch" theorems [21], an algorithm that performs well on a certain class of optimization problems is typically associated with degraded performance for other problems. Therefore it is expected that clustering algorithms optimized for detecting common cell clusters are unable to detect rare cell clusters, and vice versa. While ensemble clustering is a promising strategy to combine the strengths of multiple methods [15, 4, 5], our analysis shows that the traditional, unweighted approach does not perform well.

To optimally combine the strengths of different clustering methods, we have developed GiniClust2, which is a cluster-aware, weighted ensemble clustering method. GiniClust2 effectively combines the strengths of Gini-index- and Fano-factor-based clustering methods for detecting rare and common cell clusters, respectively, by assigning higher weights to the more reliable clusters for each method. By analyzing a number of simulated and real scRNA-seq datasets, we find that GiniClust2 consistently performs better than other methods in

320 maintaining the overall balance of detecting both rare and common cell types.
321 This weighted approach is generally applicable to a wide range of problems.
322
323 GiniClust2 is currently the only rare-cell-specific detection method equipped to
324 handle such large data sets, as demonstrated by our analysis of the 68k PBMC
325 dataset from 10X Genomics. This property is important for detecting hidden cell
326 types from large datasets, and may be particularly useful for annotating the
327 Human Cell Atlas [22].
328

329 **MATERIALS AND METHODS**

330 **Data Source and Preprocessing**

331 Mouse ESC scRNA-seq data was obtained from Gene Expression
332 Omnibus (GEO) with the accession number GSE65525 [18]. This processed data
333 is represented as UMI filtered-mapped counts. Removing genes expressed in
334 fewer than 3 cells, and cells expressing fewer than 2000 genes, we were left with
335 a total of 8055 genes and 278 cells.
336
337 The 68k PBMC dataset was obtained from the 10X Genomics website:
338 http://support.10xgenomics.com/single-cell/datasets. The processed data
339 (represented as UMI counts) was filtered and normalized using the code
340 provided by 10X Genomics (https://github.com/10XGenomics/single-cell-3prime-
341 paper). The resulting data consists of a total of 20387 genes and 68579 cells.
342 Cell-type labels were assigned based on the maximum correlation between the

343 gene expression profile of each single cell to 11 purified cell populations, using
344 the code provided by 10X Genomics.
345
346 **GiniClust2 method details**
347 GiniClust2 has been implemented in R and deposited at Github with URL
348 https://github.com/dtsoucas/GiniClust2. The GiniClust2 pipeline contains the
349 following steps.
350
351 *Step 1: Clustering cells using Gini-index based features*
352
353 The Gini index for each gene is calculated and normalized as described
354 before [11]. Briefly, the raw Gini index is calculated as twice the area between
355 the diagonal and the Lorenz curve, taking a range of values between 0 and 1.
356 Raw Gini index values are normalized by removing the trend with maximum
357 expression levels using a two-step LOESS regression procedure as described in
358 [11]. Genes whose normalized Gini index is significantly above zero (p-value <
359 0.0001) are labeled high Gini genes and selected for further analysis.
360
361 A high-Gini-gene-based distance is calculated between each pair of cells using
362 the Jaccard distance metric. This is used as input into DBSCAN [23], which is
363 implemented using the dbscan function in the fpc R package, with method=
364 "dist". Parameter choices for eps and MinPts are discussed in the
365 **Supplemental Information**.

366

367 *Step 2: Clustering cells using Fano-factor based features*

368 The Fano factor is defined as the variance over mean expression value for
369 each gene. The top 1000 genes are chosen for further analysis. Principal
370 component analysis (PCA) is applied to the gene expression matrix for
371 dimensionality reduction, using the svd function in R. The first 50 principal
372 components are reserved for clustering analysis. Cell clusters are identified by k-
373 means clustering, using the kmeans function in R with default parameters.
374 Optimal choice of k is discussed in the **Supplemental Information**. To improve
375 robustness, 20 independent runs of k-means clustering with different random
376 initializations are applied to each dataset, and the optimal clustering result is
377 selected.

378

379 *Step 3. Combining the results from Steps 1 and 2 via a cluster-aware, weighted*
380 *ensemble approach.*

381 We adapted the weighted consensus clustering algorithm developed by Li
382 and Ding [24] by further considering cluster-specific weighting. For GiniClust,
383 higher weights are assigned to the rare cell clusters and lower weights to
384 common clusters, whereas the opposite scheme is used to weight the outcome
385 from k-means clustering. This allows us to combine the strength of each
386 clustering method. The mathematical details are described as follows, and
387 visualized in **Fig. 1b**.

388

Let P^G be the partitioning provided by GiniClust, and P^F the partitioning provided by Fano-factor-based clustering. Each partition consists of a set of clusters: $C^G = C_1^G, C_2^G, \ldots, C_{k_G}^G$, and $C^F = C_1^F, C_2^F, \ldots, C_{k_F}^F$. Define the connectivity matrices as:

$$M_{ij}(P^G) = \begin{cases} 1, (i,j) \in C_k(P^G) \\ 0, otherwise \end{cases}, \text{ and } M_{ij}(P^F) = \begin{cases} 1, (i,j) \in C_k(P^F) \\ 0, otherwise \end{cases}$$

If two cells are clustered together in the same group, their connectivity is 1, while if they are clustered separately, their connectivity is 0. Define the weighted consensus association as:

$$\overline{M_{ij}} = w_{ij}^G M_{ij}(P^G) + w_{ij}^F M_{ij}(P^F)$$

where $w_{ij}^G + w_{ij}^F = 1$, $w_{ij}^G, w_{ij}^G \geq 0 \; \forall \; i,j \in [1,n]$, *n* represents the number of cells. Weights w_{ij}^G and w_{ij}^F are specific to each pair of cells, and are determined based on w_i^G and w_i^F, weights that are specific to each cell. These cell-specific weights are determined based on the rarity of the cell cluster that the cell is assigned to, the specifics in which are discussed in the next section. The cell-pair-specific weights are translated from the cell-specific weights as follows:

$$\widetilde{w}_{ij}^G = \begin{cases} w_i^G, w_i^G = w_j^G \\ \max(w_i^G, w_j^G), \; w_i^G \neq w_j^G \end{cases} \text{ and } \widetilde{w}_{ij}^F = \begin{cases} w_i^F, w_i^F = w_j^F \\ \max(w_i^F, w_j^F), \; w_i^F \neq w_j^F \end{cases}$$

Then, weights are normalized to 1:

$$w_{ij}^{G} = \frac{\widetilde{w}_{ij}^{G}}{\widetilde{w}_{ij}^{G} + \widetilde{w}_{ij}^{F}} \text{ and } w_{ij}^{F} = \frac{\widetilde{w}_{ij}^{F}}{\widetilde{w}_{ij}^{G} + \widetilde{w}_{ij}^{F}}$$

Each cell-cell pair will thus be assigned a weighted consensus association between 0 and 1, which is a cell-type-proportion-based weighted average of both GiniClust and Fano-factor-based clustering associations.

At this point, the weighted consensus association matrix provides a probabilistic clustering for each cell, where each entry represents the probability that cell i and cell j reside in the same cluster. To transform this into a final deterministic clustering assignment, we optimize the following:

$$min_{U} \left|\left|\overline{M} - U\right|\right|^{2},$$

where U is any possible connectivity matrix. In Li and Ding [24], this optimization problem is solved via symmetric non-negative matrix factorization (NMF) to yield a soft clustering. To obtain a hard clustering we add an orthogonality constraint, leading to orthogonal NMF, or k-means clustering [25], implemented once again using the kmeans R function.

432 **Weighting scheme parameters**.

433 Cell-specific GiniClust (w_{ij}^{G}) and Fano-factor-based (w_{ij}^{F}) weights are
434 determined as a function of the size of the cluster containing the particular cell.
435 To illustrate the origins of the weighting scheme, we turn to a conceptual
436 example.

437

438 As the proportion of the rare cell type increases, the utility of GiniClust begins to
439 decline. In a complementary manner, the ability of Fano-factor-based k-means
440 clustering to detect the rare cell type is expected to rise. We find this
441 phenomenon to hold true in all the datasets we analyzed, and to be best
442 demonstrated in the subsampled PBMC data results. **Supplemental Fig. 4a**
443 shows a schematic of the respective abilities of GiniClust and Fano-factor-based
444 k-means clustering to detect a rare cell type over various rare cell type
445 proportions, created by applying smoothing to the subsampled PBMC results.
446 We next calculate the ratio between this GiniClust detection ability and the sum
447 of both GiniClust and k-means detection abilities (**Supplemental Fig. 4b**). This
448 is a measure of the power of GiniClust over k-means in detecting the rare cell
449 type. When this ratio is 1, GiniClust is solely able to detect the rare cell type,
450 while at zero k-means alone is useful. This curve closely follows a logistic shape,
451 and thus the following equation is used to approximate it:

452

$$w^{G}(x) = 1 - \frac{1}{1 + e^{-\frac{x-\mu\prime}{s\prime}}}$$

where μ' is the rare cell type proportion at which GiniClust and Fano-factor-based clustering methods have the same ability to detect rare cell types, and s' represents how quickly GiniClust loses its ability to detect rare cell types above μ' (see **Supplemental Information** for more details). **Supplemental Fig. 4b** also contains this fitted curve, which is then directly used to determine the GiniClust weights (w_{ij}^{G}). Fano-factor-based clustering weights are determined by recognizing that while Fano-factor-based clusters may not contain rare cell clusters, they are still indicative of where these rare cells lie in the context of the larger cell population. As such, weights are set such that $w_{ij}^{F} = f'$, where f' is a constant corresponding to the importance of larger Fano cluster membership for rare cells (see **Supplemental Information** for details on the choice of f'). Clustering results tend to be robust to the specific choice of μ', s', and f'. **Fig. 1c** shows the result of subsequent scaling of Gini and Fano weights to sum to one, and their associated scaled parameters μ, s, and f.

tSNE visualization

Dimension reduction by tSNE [26] is performed using the Rtsne R package. The tSNE algorithm is first run using the Gini-based distance to obtain a one-dimensional projection of each cell. For large data sets, tSNE is run on the first 50 principal components of the Gini-based distance to prevent tSNE from becoming prohibitively slow. Then, the tSNE algorithm is run using the first 50 principal components of our Fano-based Euclidean distance to obtain a separate

476 two-dimensional projection. The three resulting dimensions (one for Gini-based
477 distance and two for Fano-based distance) are plotted to visualize cluster
478 separation.
479
480 **Differential expression analysis on resulting clusters**
481 Differentially expressed genes for each cluster are determined by
482 comparing their gene expression levels to all other clusters. This is performed
483 using the zlm.SingleCellAssay function in the R MAST package [27], with
484 method= "glm". P-values for differentially expressed genes are calculated using
485 the lrTest function, with a hurdle model.
486
487 **SC3 analysis**
488 SC3 [4] was accessed through the SC3 Bioconductor R package. SC3
489 was applied to the simulated data set post-filtering using default parameters, with
490 k=6 to match the true number of clusters.
491
492 **CSPA analysis**
493 Matlab code for the CSPA [15] was accessed through
494 http://strehl.com/soft.html, under "ClusterPack_V2.0." CSPA was applied to the
495 Gini and Fano-based clustering results for the simulated data set, using the
496 clusterensemble function, specifying the CSPA option. K was specified as 6 to
497 match the true number of clusters.
498

499 **RaceID2 analysis**

500 RaceID2 [10] R scripts were accessed through
501 https://github.com/dgrun/StemID. RaceID2 was applied to already-filtered data
502 sets as above to make results directly comparable to GiniClust2, with default
503 parameters. k was set to match the corresponding GiniClust2 k parameter
504 specification.

505

506 **Simulation details**

507 We created synthetic data following the same approach as Jiang et al.
508 [11], specifying one large 2000 cell cluster, one large 1000 cell cluster, and four
509 rare clusters of 10, 6, 4 and 3 cells, respectively. Gene expression levels are
510 modeled using a negative binomial distribution, and distribution parameters are
511 estimated using an intestinal scRNA-seq data set using a background noise
512 model as in Grün et al. [9]. To create clusters with distinct gene expression
513 patterns, we permute 100 lowly (mean<10 counts) and 100 highly (mean>10
514 counts) expressed gene labels for each cluster (see Jiang et al. [11] for more
515 details). This results in a 23,538 gene by 3023 cell data set. After filtering (as
516 above) we are left with 3708 genes and 3023 cells.

517

518 **10X Genomics Data Subsampling**

519 The full 68k 10X Genomics PBMC dataset is down-sampled for model
520 evaluation. We created 6 sets of 20 subsampled data sets each for a total of 120
521 data sets in the following manner: five cells were randomly sampled from the

522 CD56+ NK cell population to form a “rare” cell group for all 120 datasets. Then,
523 for each set of 20 data sets, cells were randomly sampled from the CD14+
524 Monocyte and CD19+ B cells in specified numbers to form “common” cell
525 clusters, the details of which are listed in **Supplemental Table 1**.
526

527 **ABBREVIATIONS**

528 tSNE, t-distributed stochastic neighbor embedding; PCA, principal component
529 analysis; CSPA, cluster-based similarity partitioning algorithm; RaceID, rare cell
530 type identification; PBMC, peripheral blood mononuclear cell; NK, natural killer;
531 MCC, Matthews correlation coefficient; LIF, leukemia inhibitory factor; mESC,
532 mouse embryonic stem cell; PrEn, primitive endoderm; NMI, normalized mutual
533 information; MAST, model-based analysis of single-cell transcriptomics; scRNA-
534 seq, single-cell RNA-sequencing; DBSCAN, density-based spatial clustering of
535 applications with noise; NMF, non-negative matrix factorization

536 **DECLARATIONS**

537 **Ethics approval and consent to participate**

538 Not applicable.
539

540 **Consent for publication**

541 Not applicable.
542

543 **Availability of data and materials**

544 GiniClust2 has been implemented in R and deposited at Github with URL
545 https://github.com/dtsoucas/GiniClust2. Datasets analyzed were all from
546 publically available sources, and are described in **Methods and Materials**
547 section.
548
549 **Funding**
550 This work was supported by a Claudia Barr Award and NIH grant R01HL119099
551 to GCY. DT's research was in part supported by an NIH training grant
552 T32GM074897.
553
554 **Competing interests**
555 The authors declare no competing interests.
556
557 **Authors' contributions**
558 DT and GCY conceived of and designed the computational method. DT
559 implemented the method. DT and GCY wrote the manuscript. All authors read
560 and approved the final manuscript.
561
562 **Acknowledgements**
563 We thank Dr. Lan Jiang and members of the Yuan Lab for helpful discussions, as
564 well as Drs. John Quackenbush and Martin Aryee for their support and advice.
565
566 **SUPPLEMENTAL INFORMATION**

567 **Parameter choice for DBSCAN**

568 DBSCAN has two parameters: MinPts and eps. MinPts is specified as 3
569 for all data sets except for the PBMC data, where MinPts is set to 100, in
570 accordance with the larger size of this data set. This corresponds to the
571 minimum cluster size for which we would expect to see a biologically relevant
572 cluster. In general, we find an appropriate MinPts specification to be about 0.1%
573 of the total number of cells.

574

575 The eps parameter is determined by a k-nearest-neighbors (kNN) plot as
576 recommended by the authors of DBSCAN [23]. According to their approach,
577 distance from each point to its kth nearest neighbor is plotted in ascending order,
578 where k=MinPts. This will form a line featuring an inflection point, at which lies
579 the recommended choice for eps [23]. If multiple inflection points exist, this may
580 suggest that multiple values of eps are worth exploring; however, in our case as
581 we are concerned with rare clusters, we only consider the smallest choice of eps,
582 corresponding to the first inflection point. Here, our Gini-based distance metric is
583 particularly low-dimensional due to the use of Jaccard distance and a small
584 number of high Gini genes. This causes cells with similar expression profiles to
585 have pairwise distances of zero, which distorts the traditional kNN-distance curve
586 shape and makes the inflection point harder to visualize (**Supplemental Figure**
587 **3**). We provide an alternate numerical approach to approximating the inflection
588 point: after removing all zero distances, the inflection point roughly corresponds
589 to the kNN-distance of the (0.00125*total number of cells*MinPts)th cell. For the

590 68k data set, the computation of all kNN distances was prohibitive, so we
591 subsampled 2057 cells and computed 3-NN distances to maintain the ratio of
592 MinPts to the total sample size. All analyses of PBMC subsampled data use this
593 same resulting value of eps.
594
595 **Choice of k for k-means clustering**
596 We give the option of automatically determining k using the gap statistic.
597 However, observing differentially expressed genes and visualizing k-means
598 clusters gives the best intuition as to the optimal k. We also do not suggest using
599 the gap statistic for large data sets due to its computational demands. For the
600 synthetic data, we chose k=2, in accordance with the number of large clusters,
601 but show that k=3 will also yield the same result (**Supplemental Fig. 1**). For
602 subsampled PBMC data sets, k was chosen as 2 or 3 depending on the ability of
603 k-means to pick up the rare NK cell group. K was chosen as 2 for the day 4 post-
604 LIF mouse embryonic stem cell data as we found this number to best group
605 biologically meaningful cell types.
606
607 For the 68k PBMC data, we chose k=10 to allow for direct comparison with
608 clustering results from Zheng et al. [16]. For comparison to the k=10 parameter
609 choice, we additionally show results for both choices k=8 and k=12 for the Fano-
610 based clustering step. All parameter choices perform comparably, with NMIs of
611 0.542, 0.541 and 0.498, respectively, when compared to the reference labels.
612 The k=8 clustering results in two fewer clusters within the CD56+ NK, CD8+

613 Cytotoxic T, CD8+/CD45RA+ Naïve Cytotoxic continuum. The k=12 clustering
614 further splits clusters containing overlapping CD8+/CD45RA+ Naïve Cytotoxic,
615 CD4+/CD45RA+/CD25- Naïve T, CD4+/CD25 T Reg, and CD4+/CD45RO+
616 Memory cells, as well as adding another cluster to the aforementioned CD56+
617 NK, CD8+ Cytotoxic T, CD8+/CD45RA+ Naïve Cytotoxic continuum. These
618 changes are minor as they occur predominantly in regions of unclear identity.
619 **Supplemental Fig. 2** shows the Fano-based tSNE dimension reduction labeled
620 with k-means clusters for k=8, 10 and 12.
621
622 **Parameter choices for weighted consensus clustering**
623 Weighting parameters μ', s', and f' are determined by the curve:
624

$$w^G(x) = 1 - \frac{1}{1 + e^{-\frac{x-\mu\prime}{s\prime}}}$$

625
626 and are represented in **Supplemental Fig. 5**. μ' is the rare cell type proportion at
627 which GiniClust and Fano-factor-based clustering methods have the same ability
628 to detect rare cell types. We set this value as 4*(MinPts/total number of cells),
629 where MinPts is the DBSCAN parameter for GiniClust. The smallest possible
630 rare cell type proportion for GiniClust is set as (MinPts/total number of
631 cells). s' represents how quickly GiniClust loses its ability to detect rare cell types
632 above μ'. More quantitatively, we define the 99th percentile of the GiniClust
633 weighting distribution as the point in which GiniClust no longer reliably detects

634 rare cell types. We set s' such that the 99th percentile is reached at
635 6*(MinPts/total number of cells). We expect that at six times the size of smallest
636 possible rare cell type for GiniClust, any cell type found likely consists of a
637 heterogeneous grouping of cells. f' represents the importance of the Fano
638 cluster membership in determining the larger context of the membership of each
639 cell, and by default f' = 0.1. We find that final membership is not sensitive to
640 small changes in f'.

641

642 **Supplemental Table 1**. Cell numbers in three different cell types for each of 120
643 subsampled datasets from 68k PBMCs.

Simulations	CD54+ NK	CD14+ Monocyte	CD19+ B	Rare Cell Type Proportion
1-20.	5	800	400	.004
21-40.	5	400	200	.008
41-60.	5	200	100	.016
61-80.	5	100	50	.032
81-100.	5	50	25	.063
101-120.	5	25	13	.116

644

645 **REFERENCES**

646 1. Tsoucas D, Yuan GC: **Recent progress in single-cell cancer genomics.**
647 *Curr Opin Genet Dev* 2017, **42:**22-32.
648 2. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical**
649 **challenges in single-cell transcriptomics.** *Nat Rev Genet* 2015, **16:**133-
650 145.

651 3. Satija R, Farrell JA, Gennert D, Schier AF, Regev A: **Spatial**
652 **reconstruction of single-cell gene expression data.** *Nat Biotechnol*
653 2015, **33:**495-502.
654 4. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T,
655 Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M: **SC3:**
656 **consensus clustering of single-cell RNA-seq data.** *Nat Methods* 2017,
657 **14:**483-486.
658 5. Giecold G, Marco E, Garcia SP, Trippa L, Yuan GC: **Robust lineage**
659 **reconstruction from high-dimensional single-cell data.** *Nucleic Acids*
660 *Res* 2016, **44:**e122.
661 6. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M,
662 Adiconis X, Levin JZ, Nemesh J, Goldman M, et al: **Comprehensive**
663 **Classification of Retinal Bipolar Neurons by Single-Cell**
664 **Transcriptomics.** *Cell* 2016, **166:**1308-1323.e1330.
665 7. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G,
666 Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al: **Brain**
667 **structure. Cell types in the mouse cortex and hippocampus revealed**
668 **by single-cell RNA-seq.** *Science* 2015, **347:**1138-1142.
669 8. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT,
670 Sorensen SA, Dolbeare T, et al: **Adult mouse cortical cell taxonomy**
671 **revealed by single cell transcriptomics.** *Nat Neurosci* 2016, **19:**335-
672 346.
673 9. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N,

674 Clevers H, van Oudenaarden A: **Single-cell messenger RNA**
675 **sequencing reveals rare intestinal cell types.** *Nature* 2015, **525:**251-
676 255.
677 10. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A,
678 Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al:
679 **De Novo Prediction of Stem Cell Identity using Single-Cell**
680 **Transcriptome Data.** *Cell Stem Cell* 2016, **19:**266-277.
681 11. Jiang L, Chen H, Pinello L, Yuan GC: **GiniClust: detecting rare cell**
682 **types from single-cell gene expression data with Gini index.** *Genome*
683 *Biol* 2016, **17:**144.
684 12. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C,
685 Beqiri M, Sproesser K, Brafford PA, Xiao M, et al: **Rare cell variability**
686 **and drug-induced reprogramming as a mode of cancer drug**
687 **resistance.** *Nature* 2017, **546:**431-435.
688 13. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in**
689 **a data set via the gap statistic.** *Journal of the Royal Statistical Society:*
690 *Series B (Statistical Methodology)* 2001, **63:**411-423.
691 14. Kodinariya T, Makwana P: **Review on determining number of cluster in**
692 **k-means clustering.** *International Journal* 2013**:**90-95.
693 15. Strehl A, Ghosh J: **Cluster ensembles--a knowledge reuse framework**
694 **for combining multiple partitions.** *Journal of Machine Learning*
695 *Research* 2002, **3:**583-617.
696 16. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo

697 SB, Wheeler TD, McDermott GP, Zhu J, et al: **Massively parallel digital**
698 **transcriptional profiling of single cells.** *Nat Commun* 2017, **8:**14049.
699 17. Matthews BW: **Comparison of the predicted and observed secondary**
700 **structure of T4 phage lysozyme.***Biochim Biophys Acta* 1975, **405:**442-
701 451.
702 18. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L,
703 Weitz DA, Kirschner MW: **Droplet barcoding for single-cell**
704 **transcriptomics applied to embryonic stem cells.** *Cell* 2015, **161:**1187-
705 1201.
706 19. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD,
707 Diehn M, Alizadeh AA: **Robust enumeration of cell subsets from tissue**
708 **expression profiles.** *Nat Methods* 2015, **12:**453-457.
709 20. Schinnerling K, García-González P, Aguillón JC: **Gene Expression**
710 **Profiling of Human Monocyte-derived Dendritic Cells - Searching for**
711 **Molecular Regulators of Tolerogenicity.** *Front Immunol* 2015, **6:**528.
712 21. Wolpert DH, Macready WG: **No free lunch theorems for optimization.**
713 *IEEE Transactions on Evolutionary Computation* 1997, **1:**67-82.
714 22. The Human Cell Atlas. https://www.humancellatlas.org. Accessed 12 Dec
715 2017.
716 23. Ester M, Kriegel H-P, Sander J, Xu X: **A Density-Based Algorithm for**
717 **Discovering Clusters in Large Spatial Databases with Noise.** In *2nd*
718 *International Conference on Knowledge Discovery and Data Mining;*
719 *Portland, OR*. AAAI; 1996: 226-231.

720 24. Li T, Ding C: **Weighted consensus clustering.** In *SIAM International*
721 *Conference on Data Mining*. Society for Industrial and Applied
722 Mathematics; 2008
723 25. Ding C, He X, Simon H: **On the equivalence of nonnegative matrix**
724 **factorization and spectral clustering.** In *SIAM International Conference*
725 *on Data Mining*. Society for Industrial and Applied Mathematics; 2005:
726 606-610.
727 26. Maaten Lvd, Hinton G: **Visualizing data using t-SNE.** *Journal of Machine*
728 *Learning Research* 2008, **9:**2579-2605.
729 27. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter
730 CK, Miller HW, McElrath MJ, Prlic M, et al: **MAST: a flexible statistical**
731 **framework for assessing transcriptional changes and characterizing**
732 **heterogeneity in single-cell RNA sequencing data.** *Genome Biol* 2015,
733 **16:**278.

FIGURE LEGENDS

734 **Figure 1**. An overview of the GiniClust2 pipeline. (a) The Gini index and Fano
735 factor are used (left), respectively, to select genes for GiniClust and Fano-based
736 clustering (middle left). A cluster-aware, weighted ensemble method is applied to
737 each of these, where cluster-aware weights w_i are represented by the shading of
738 the cells (middle right), to reach a consensus clustering (right). (b) A schematic
739 of the weighted consensus association calculation, with association matrices in
740 black and white, weighting schemes in red and blue, and final GiniClust2 clusters

741 highlighted in white. (c) GiniClust and Fano-based weights, as a function of cell
742 type proportion.
743
744 **Figure 2**. The application of GiniClust2 and competing methods to simulated
745 data. (a) A heatmap representation of the simulated data with 6 distinct clusters,
746 showing the genes permuted to define each cluster. A zoomed-in view of the rare
747 clusters is shown in the smaller heatmap. (b) A comparison between the true
748 clusters (x-axis) and clustering results from GiniClust2 and comparable methods
749 (y-axis). Each cluster is represented by a distinct color bar. Multiple bars are
750 shown if a true cluster is split into multiple clusters by a clustering method. (c) A
751 three-dimensional visualization of the GiniClust2 clustering results using a
752 composite tSNE plot, combining two Fano-based tSNE dimensions and one Gini-
753 based tSNE dimension. Insets show zoomed-in views of the corresponding
754 regions.
755
756 **Figure 3**. Analysis of the 68k PBMC dataset [16]. (a) A visualization of reference
757 labels for the full data set (left), along with the 3 cell subtypes selected for
758 detailed analysis (right). (b) Comparison of the performance of different clustering
759 methods, quantified by a Matthews correlation coefficient (MCC) [17] for each of
760 the three cell types.
761
762 **Figure 4**. Analysis of the inDrop dataset for day 4 post-LIF mESC differentiation
763 [18]. (a) A heatmap of top differentially expressed genes for each GiniClust2

764 cluster. The colorbar above the heatmap indicates the cluster assignments. (b) A
765 composite tSNE plot, representing the GiniClust2 results. (c) A comparison of
766 GiniClust2 and RaceID2 clustering results, for common (above) and rare (below)
767 cell types. The same color-coding scheme is used in all panels.
768
769 **Figure 5**. Results from the full 68k PBMC data analysis. (a) A composite tSNE
770 plot of the GiniClust2 results; rare cell types are circled. (b) A confusion map
771 showing similarities between GiniClust2 clusters and reference labels. Values
772 represent the proportion of cells per reference label that are in each cluster. (c) A
773 bubble plot showing expression of cluster-specific genes. Size represents the
774 percentage of cells within each cluster with non-zero expression of each gene,
775 while color represents the average normalized UMI counts for each cluster and
776 gene.
777
778 **Supplemental Figure 1**. The effect of various choices of k on (a) the k-means
779 step and (b) the overall clustering of our GiniClust2 method for the simulated
780 data. Each bar represents the contribution of a cluster to the total number of cells
781 in each reference type.
782
783 **Supplemental Figure 2**. The effect of various choices of k on the k-means step
784 for the full 68k PBMC data.
785

786 **Supplemental Figure 3**. An illustration of the eps selection process for
787 DBSCAN. Eps is chosen as the distance at the inflection point in the k-nearest-
788 neighbors distance plot.
789
790 **Supplemental Figure 4**. Weighting scheme details. (a) Rare cell type detection
791 abilities of GiniClust and k-means over a range of rare cell type proportions. (b) A
792 representation of the ability of GiniClust to detect rare cell types over k-means,
793 and its logistic fit. Parameters μ' and s' determine the shape of the curve.

Figure 1.

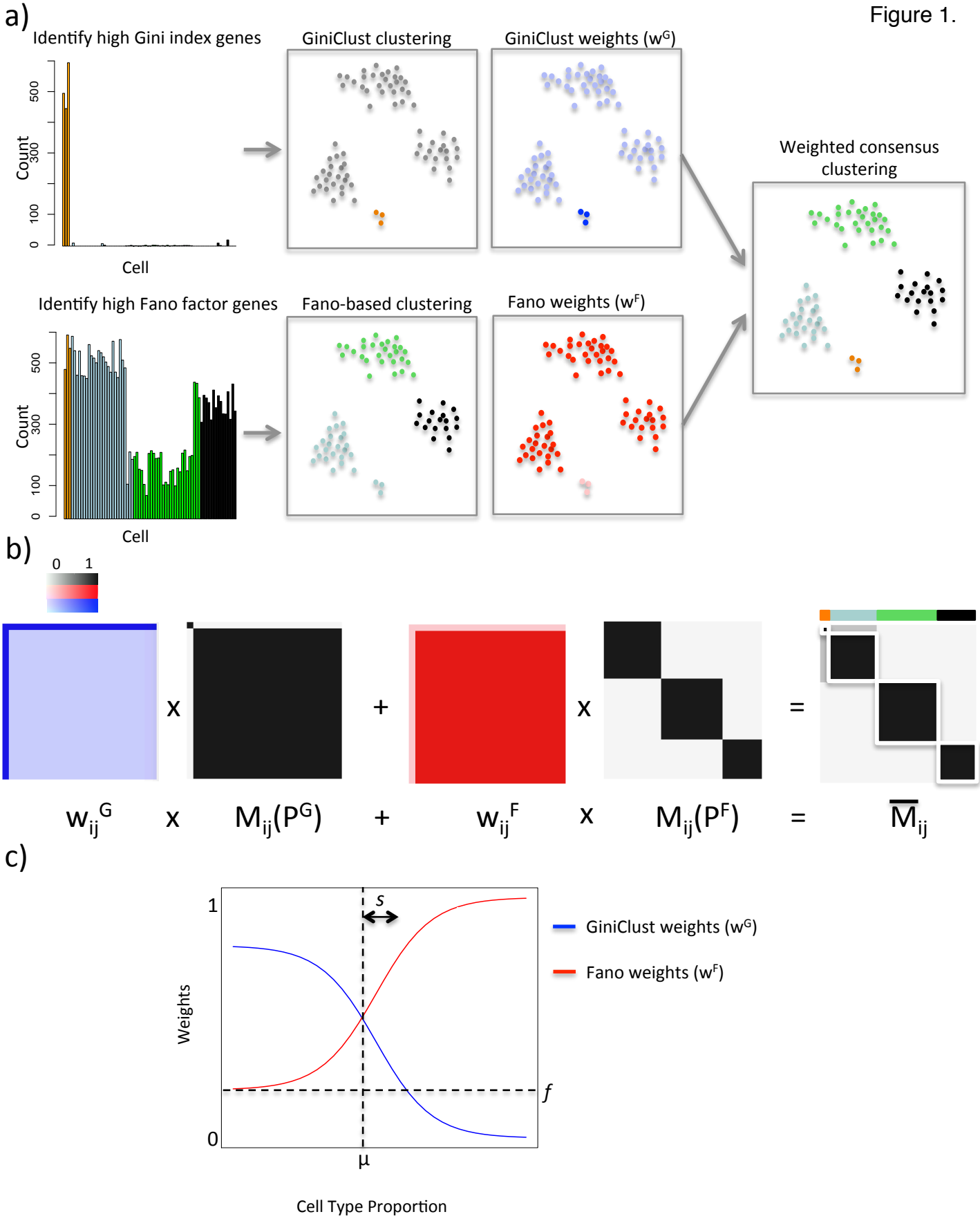

a)
Identify high Gini index genes
Count
Cell
GiniClust clustering
GiniClust weights (w^G)
Weighted consensus clustering
Identify high Fano factor genes
Fano-based clustering
Fano weights (w^F)
b)
0
1
x
+
x
=
w_ij^G x M_ij(P^G) + w_ij^F x M_ij(P^F) = M̄_ij
c)
1
0
Weights
s
f
μ
GiniClust weights (w^G)
Fano weights (w^F)
Cell Type Proportion

Figure 2.

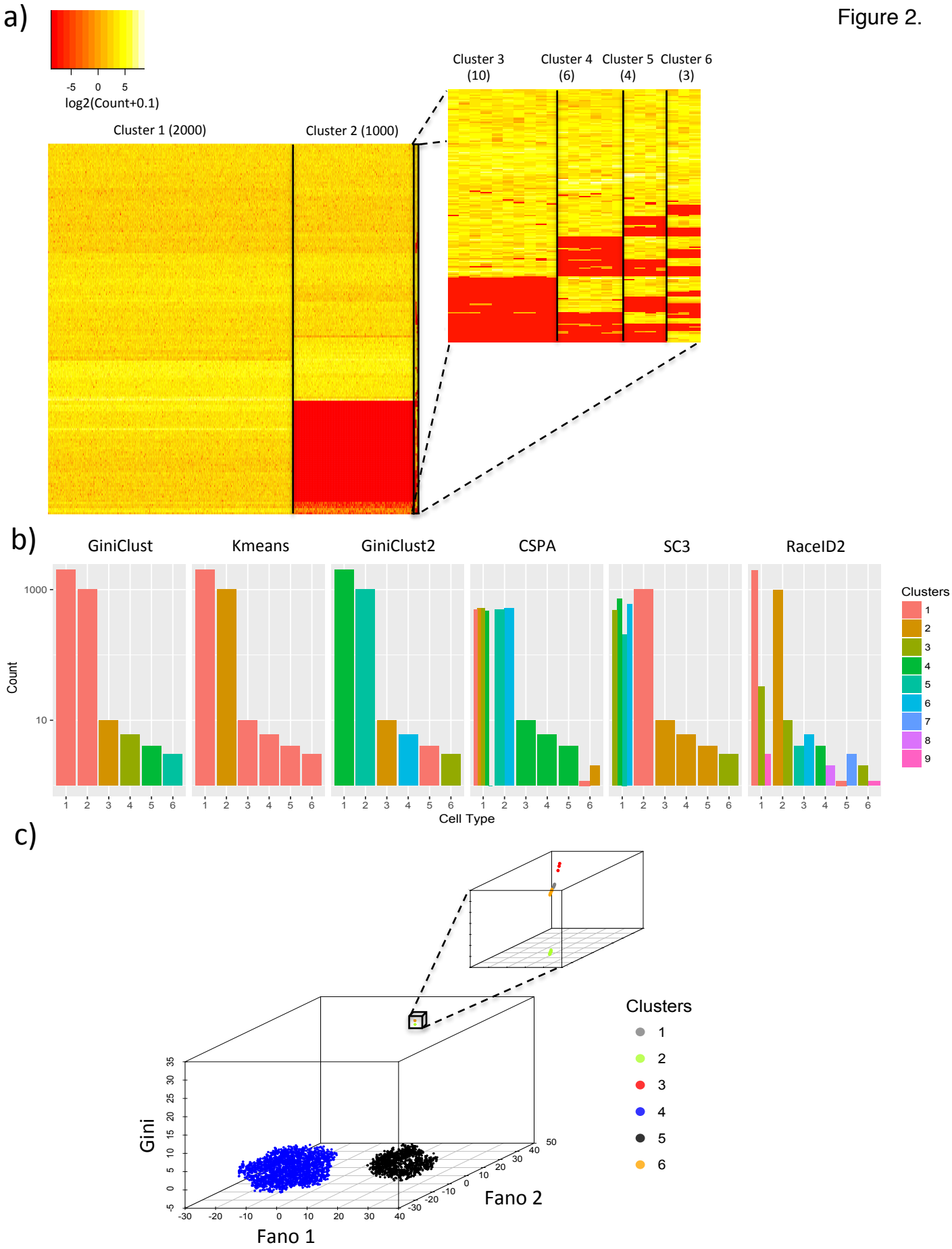

a)
-5 0 5
log2(Count+0.1)
Cluster 1 (2000)
Cluster 2 (1000)
Cluster 3 (10)
Cluster 4 (6)
Cluster 5 (4)
Cluster 6 (3)
b)
GiniClust
Kmeans
GiniClust2
CSPA
SC3
RaceID2
Count
1000
10
Cell Type
Clusters
1
2
3
4
5
6
7
8
9
c)
Gini
Fano 1
Fano 2
Clusters
1
2
3
4
5
6

Figure 3.

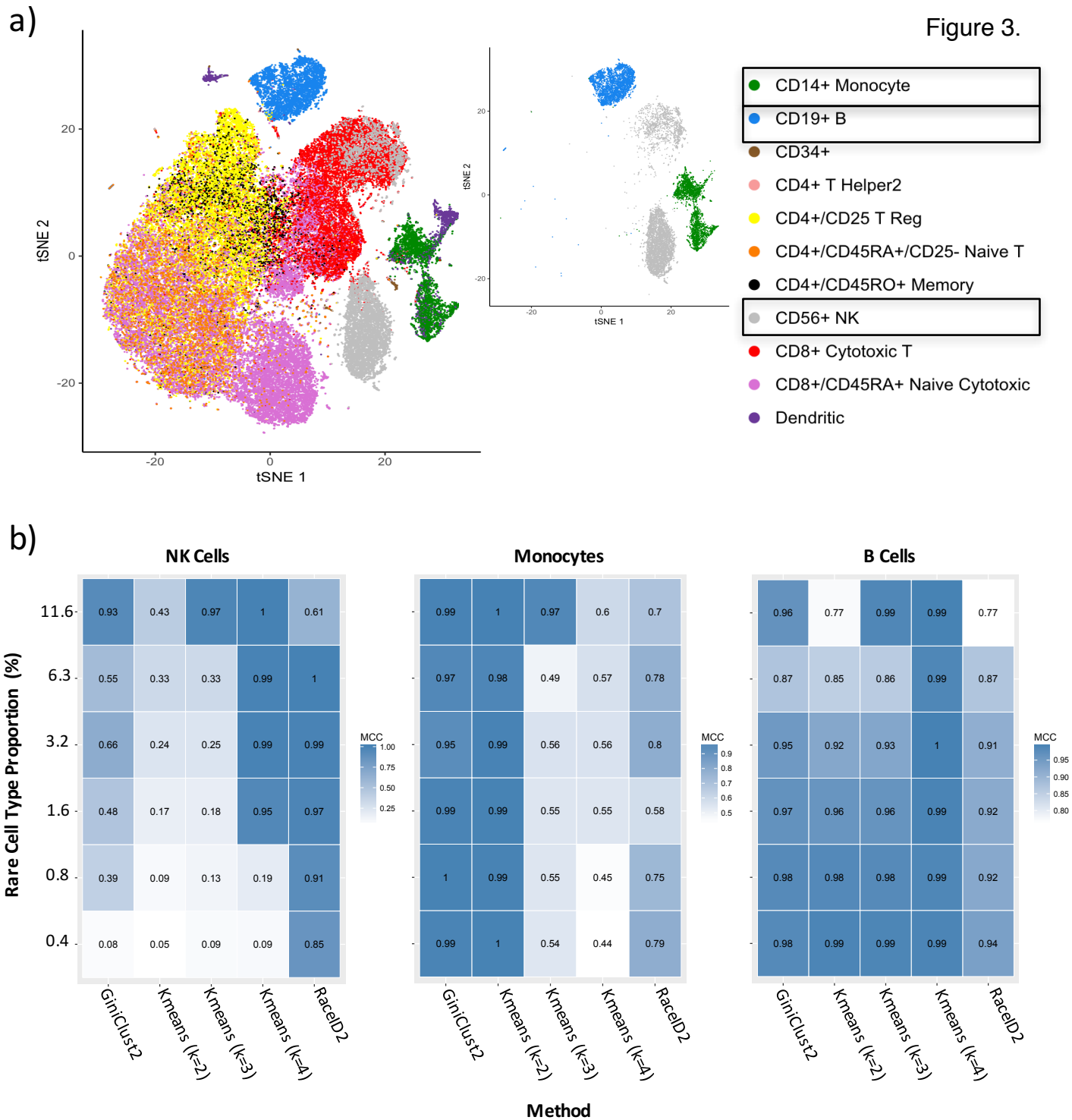
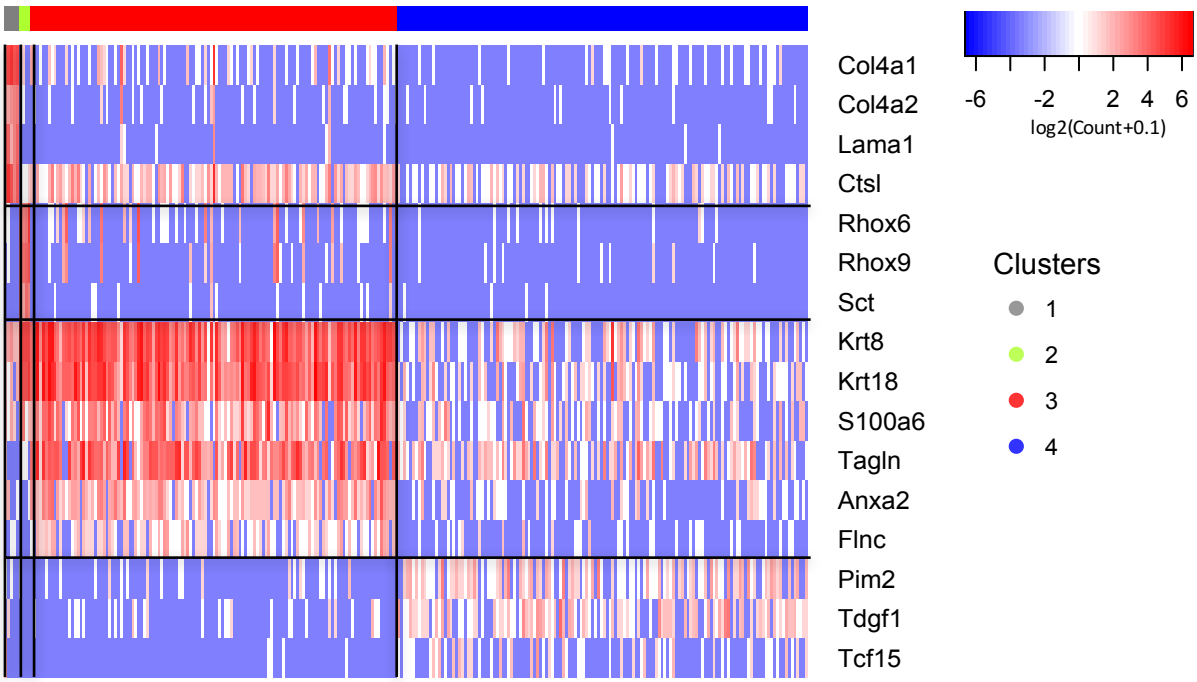
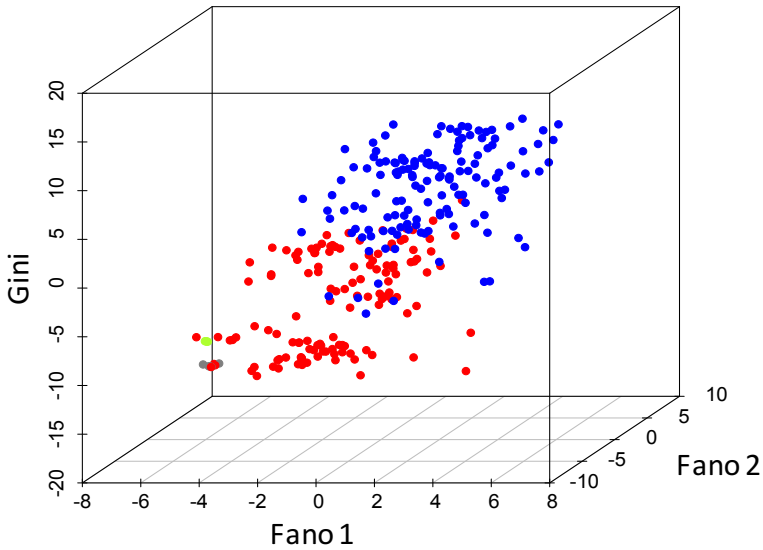

a)
tSNE 2
tSNE 1
20
0
-20
CD14+ Monocyte
CD19+ B
CD34+
CD4+ T Helper2
CD4+/CD25 T Reg
CD4+/CD45RA+/CD25- Naive T
CD4+/CD45RO+ Memory
CD56+ NK
CD8+ Cytotoxic T
CD8+/CD45RA+ Naive Cytotoxic
Dendritic
b)
NK Cells
Monocytes
B Cells
Rare Cell Type Proportion (%)
11.6
6.3
3.2
1.6
0.8
0.4
GiniClust2
Kmeans (k=2)
Kmeans (k=3)
Kmeans (k=4)
RaceID2
Method
MCC
1.00
0.75
0.50
0.25
0.9
0.8
0.7
0.6
0.5
0.95
0.90
0.85
0.80

Figure 4.

a)


Col4a1
Col4a2
Lama1
Ctsl
Rhox6
Rhox9
Sct
Krt8
Krt18
S100a6
Tagln
Anxa2
Flnc
Pim2
Tdgf1
Tcf15
-6
-2
2
4
6
log2(Count+0.1)
Clusters
1
2
3
4

b)


Gini
20
15
10
5
0
-5
-10
-15
-20
-8
-6
-4
-2
0
2
4
6
8
Fano 1
10
5
0
-5
-10
Fano 2

c)

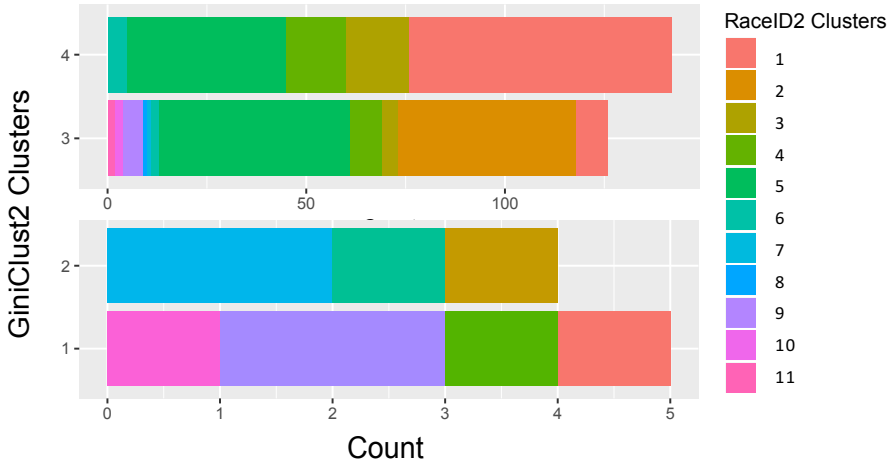
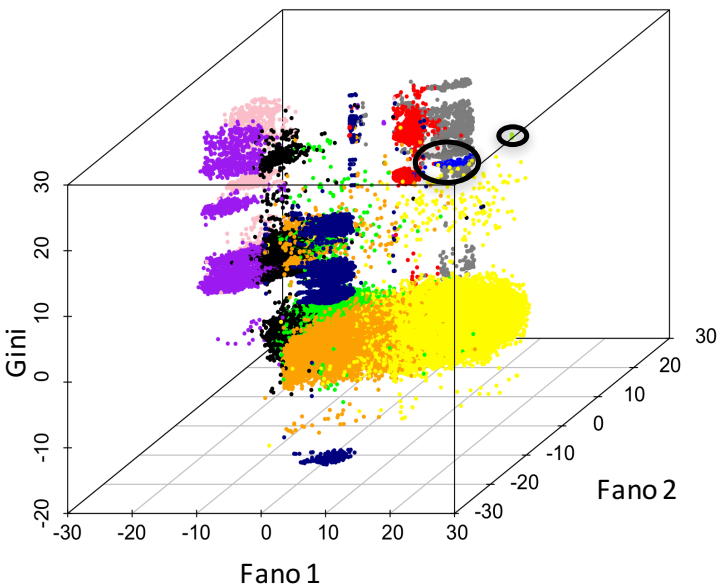
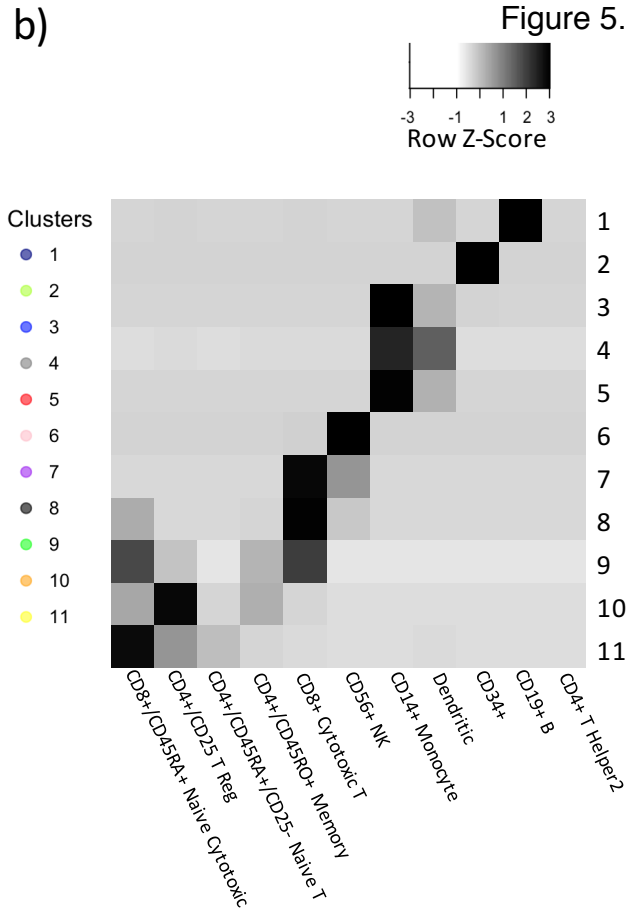

GiniClust2 Clusters
4
3
2
1
0
50
100
0
1
2
3
4
5
Count
RaceID2 Clusters
1
2
3
4
5
6
7
8
9
10
11

Figure 5.

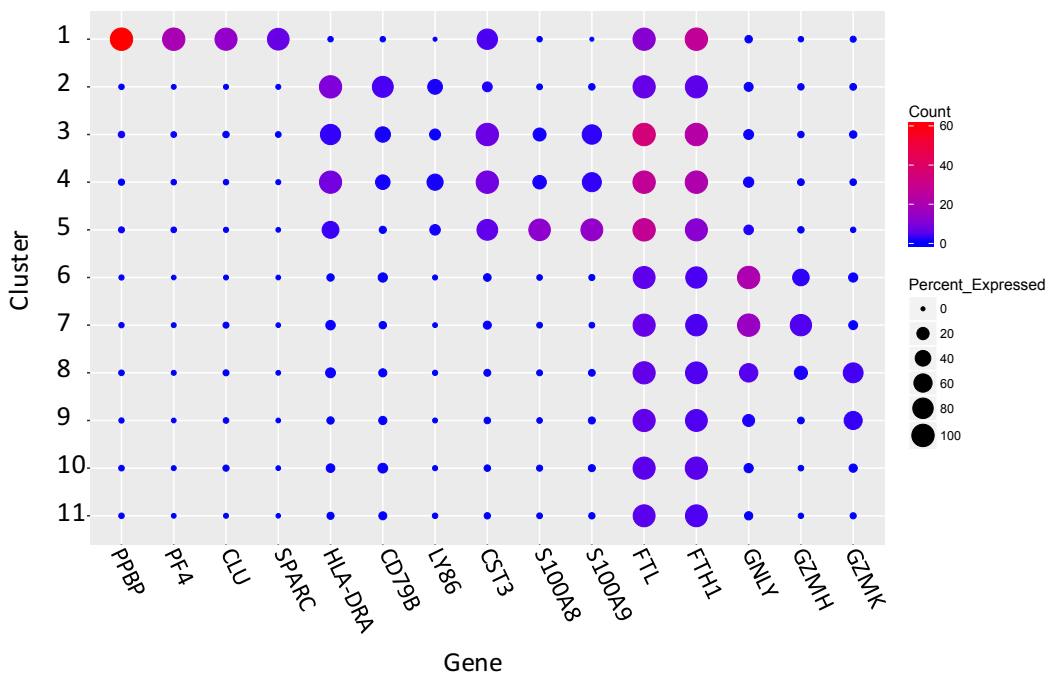
a)


Gini
Fano 1
Fano 2
-30
-20
-10
0
10
20
30

b)


Row Z-Score
-3
-1
1
2
3
Clusters
1
2
3
4
5
6
7
8
9
10
11
CD8+/CD45RA+ Naive Cytotoxic
CD4+/CD25 T Reg
CD4+/CD45RA+/CD25- Naive T
CD4+/CD45RO+ Memory
CD8+ Cytotoxic T
CD56+ NK
CD14+ Monocyte
Dendritic
CD34+
CD19+ B
CD4+ T Helper2

c)


Cluster
1
2
3
4
5
6
7
8
9
10
11
PPBP
PF4
CLU
SPARC
HLA-DRA
CD79B
LY86
CST3
S100A8
S100A9
FTL
FTH1
GNLY
GZMH
GZMK
Gene
Count
60
40
20
0
Percent_Expressed
0
20
40
60
80
100