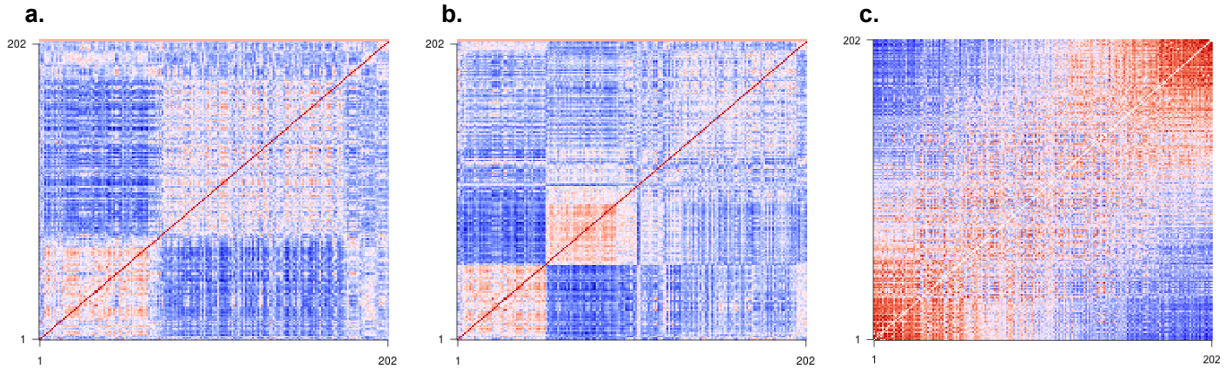
Supplementary materials for

Critical limitations of consensus clustering in class discovery

Yasin Şenbabaoğlu, George Michailidis, Jun Z. Li

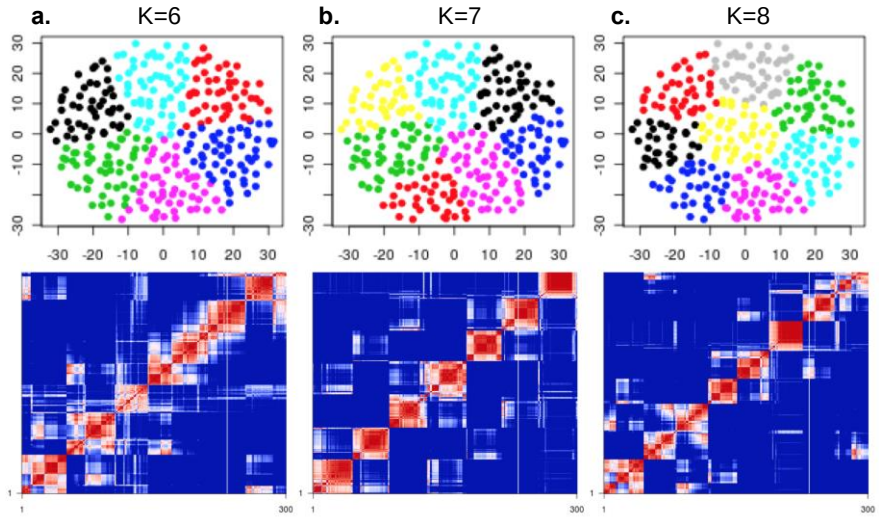
This PDF file includes:

Supplementary Figure 1



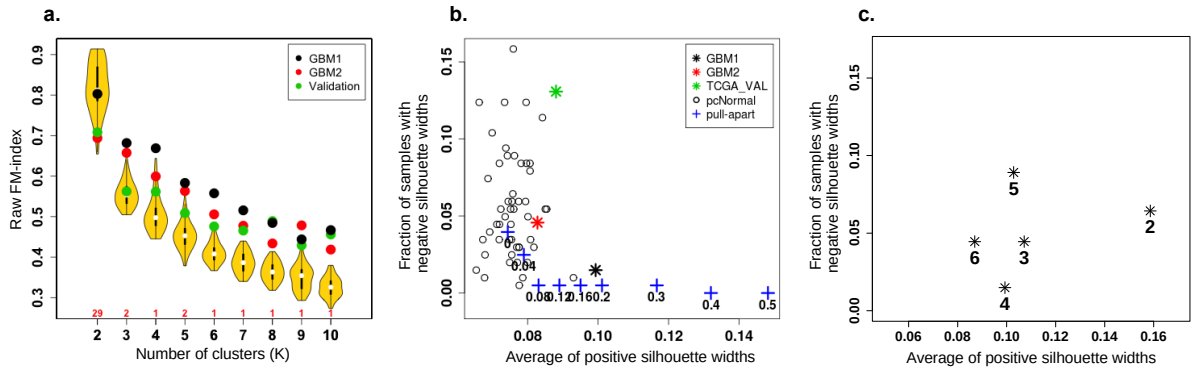
Supplementary Figure 1. Spearman's correlation coefficient matrix among GBM1 samples displayed in three different orders. The same matrix is shown, with the samples ordered by (a) average-linkage HCLUST on the K = 2 gene subsampling consensus matrix (the consensus matrix is not shown but just used to derive the sample order for the correlation matrix), (b) average linkage HCLUST on the K = 3 gene subsampling consensus matrix, and (c) decreasing PC1 scores. It is possible to re-order the same sample-sample matrix in different ways to support different impressions or conclusions of sample structure. For example, plot (c) supports a gradual transition, or the grades-of-membership model. See **Supplementary Note 1** for further discussion.

Supplementary Figure 2



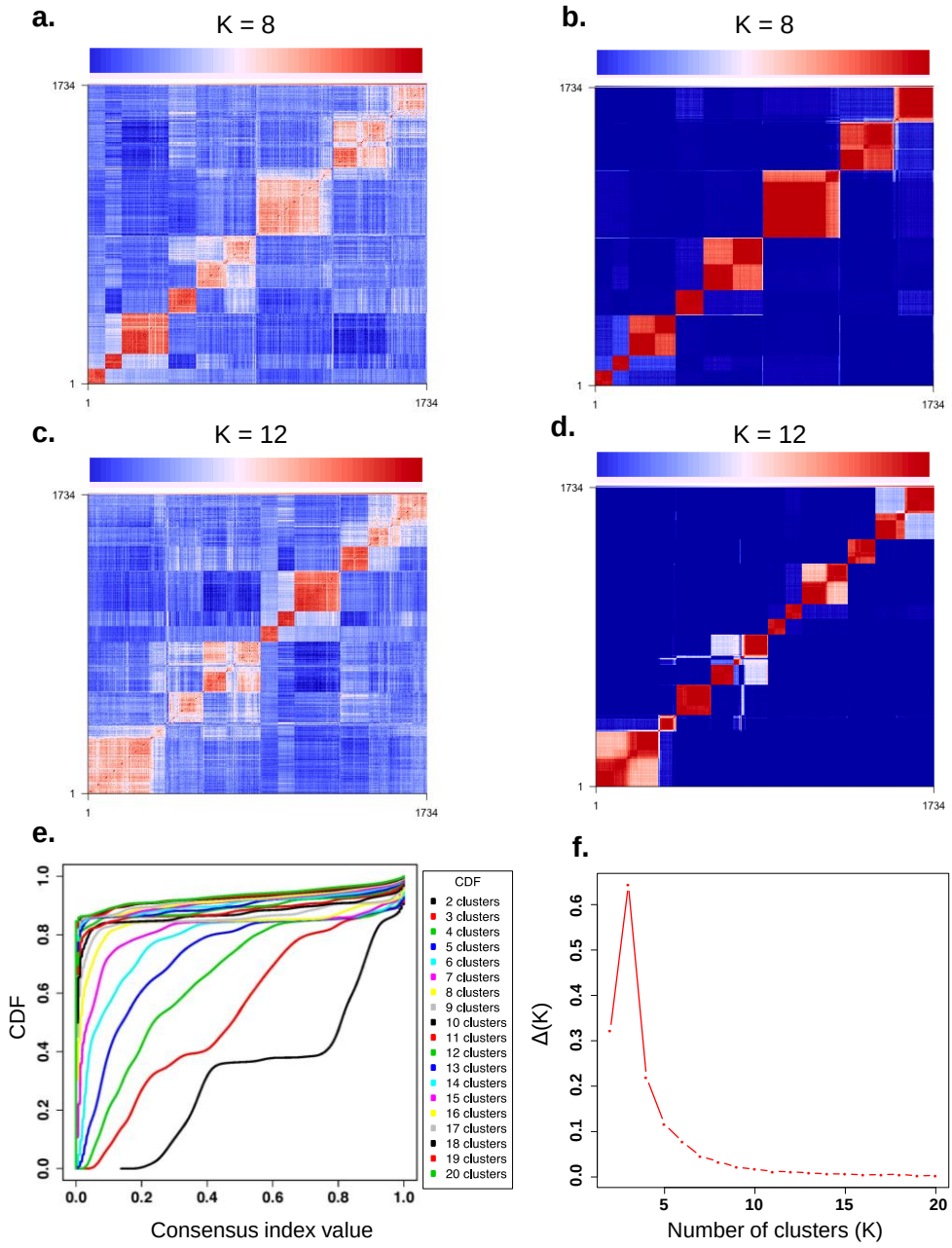
Supplementary Figure 2. The stability of *Circle1* CC heatmaps improves for larger K. Top panels in (a-c) show k-means partitioning of *Circle1* samples for K = 6 – 8, displayed with PC1 (17.7% variance explained) on the x-axis and PC2 (15.1%) on the y-axis. **Bottom** panels show consensus heatmaps for K= 6 - 8 with 80% sample subsampling and k-means as the base method. The stability of the heatmaps for 7 or 8 makes it tempting to conclude that the original data contain 7 or 8 clusters.

Supplementary Figure 3



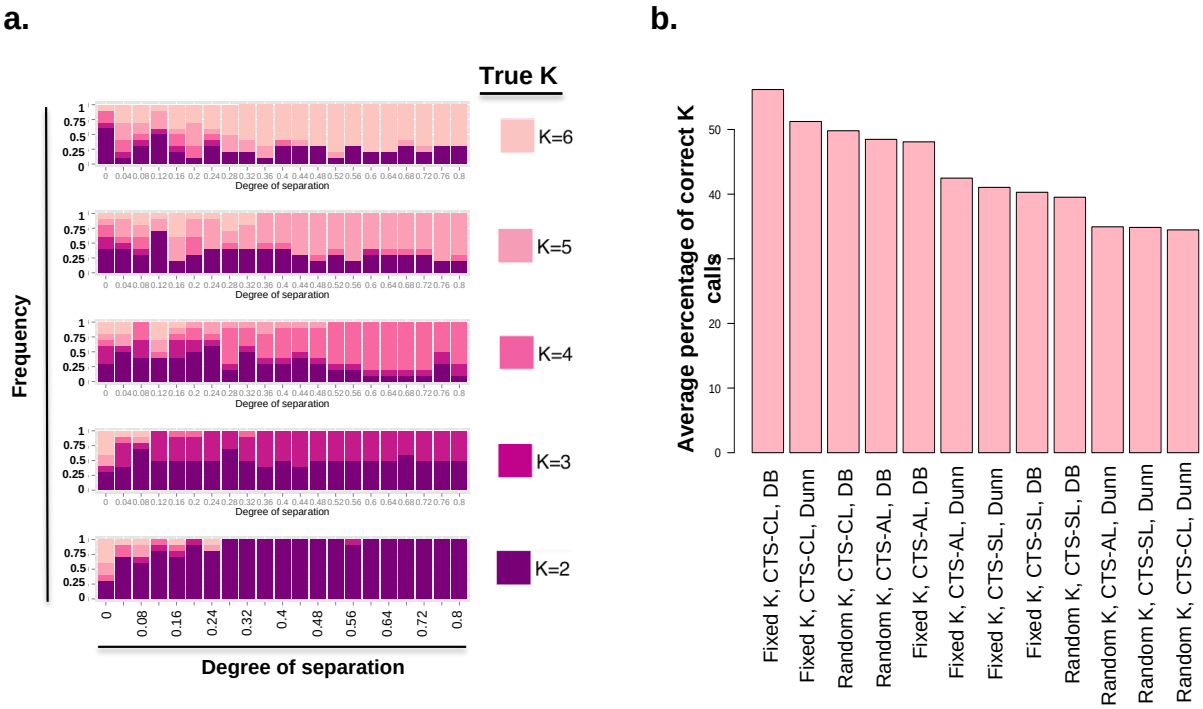
Supplementary Figure 3. CLEST and silhouette width comparisons of cluster strength between GBM and null datasets. (a) CLEST results (as FM-scores) for GBM1, GBM2, Validation, and the distribution of pcNormal null datasets in the range K = 2 - 10. (b) K=4 silhouette width analysis for GBM1, GBM2, Validation, pcNormal null datasets, and pull-apart positive datasets, with the x-axis showing the average of positive silhouette widths, and the y-axis showing the fraction of negative silhouette widths. GBM1 is within the range of 50 pcNormal simulations (shown with hollow circles) along the y-axis. But it appears as an outlier along the x-axis when compared with the null datasets. The pull-apart degree for positive datasets ranges from 0 to 0.5 (shown with blue plus symbols). Along the x-axis, GBM1 is close to the positive dataset with pull-apart degree 0.2. (c) Silhouette width analysis for GBM1 in the range K = 2 - 6. It is not obvious that K=4 as optimal. See **Supplementary Note 2** for further discussion.

Supplementary Figure 4



Supplementary Figure 4: Consensus clustering diagnostic plots on the Pan-Cancer dataset. Results are from a random 50% subset of the original PANCAN12 dataset (n = 3,468) for 12 tumor types. Correlation heatmaps (**a**, **c**) show weaker structure than consensus heatmaps (**b**, **d)**, but the cluster separation is much more pronounced than in GBM (**Figure 1**). The optimal K according to consensus heatmaps could be any value between K=8 and K=12. The CDF plot shows that, when K increases above 8, the flattening of the middle portion becomes similar, and the increase in the area under the curve becomes negligible (**e**). However, the $\Delta(K)$ elbow occurs at K=4 or K=5 (**f**).

Supplementary Figure 5



Supplementary Figure 5. Performance of LCE on positive datasets. LCE was run in twelve configurations involving two ways to generate the cluster ensemble – fixed or random K, three ways of partitioning the similarity matrix – hierarchical clustering with single linkage (SL), complete linkage (CL) and average linkage (AL), and two internal validation measures – Dunn and Davies-Bouldin (DB) indices. The best among the twelve, FixedK_CTS-CL_DB, is shown in **a**, in a similar layout as in Figure 6. **b**. Comparison across the 12 configurations, where the overall accuracy is defined as the percentage of correct K calls averaged across all (K, *a*) value pairs.

Supplementary Table 1. Optimal K called by LCE for three datasets across 12 parameter combinations.

			Optimal K calls in 3 real datasets		
Scheme	**Consensus function***	**Index**	**GBM**	**Alizadeh**	**Pan-Cancer**
Fixed K	SL	DB	3	3	2
Fixed K	**CL**	**DB**	**4**	**2**	**2**
Fixed K	AL	DB	4	3	2
Random K	SL	DB	2	7	2
Random K	CL	DB	8	2	6
Random K	AL	DB	5	2	6
Fixed K	SL	Dunn	3	2	2
Fixed K	CL	Dunn	2	2	2
Fixed K	AL	Dunn	2	2	2
Random K	SL	Dunn	2	2	2
Random K	CL	Dunn	2	2	6
Random K	AL	Dunn	2	2	6
		True K	4?	2 or 3	>8

* AL, SL and CL refers to hierarchical clustering with average linkage, single linkage, and complete linkage, respectively.

Supplementary Note 1: Robustness of the existence and number of clusters in GBM1 found by consensus clustering (CC)

Glioblastoma multiforme (GBM) was the first cancer type studied by The Cancer Genome Atlas (TCGA) Research Network [1], and was reported to have four molecular subtypes according to gene expression clusters discovered by CC [2]. Here we use the same data in a technical reassessment of CC. We will not discuss the biological implications of GBM subtypes, which have been revised and expanded since the initial study by TCGA [3-5].

We ran CC on the gene expression data for the first GBM cohort (n = 202, referred to as GBM1, see **Methods**), with K = 4 and k-means as the base clustering method. The consensus rate matrix was calculated by 500 repeated clustering runs, taking a random 80% subset of genes (**Fig. 1a**) or samples (**Fig. 1b**) in each run. As originally reported [2], the consensus heatmaps (**Fig. 1a-b**) show four crisp clusters; and it was the crispness of the clusters that was cited as the strong evidence for inherent structure at K = 4 in GBM1. However, the appearance of clusters in the Pearson's correlation coefficient matrix (**Fig. 1c**) is substantially weaker, with many samples having strong correlations with samples in a different cluster. Similarly, principal component analysis (PCA) (**Fig. 1d**) does not show distinct gaps among the four reported clusters, rather they represent contiguous partitions of an unbroken data cloud. These findings raise the question whether CC has over-stated the robustness of clusters.

A related issue is the robustness of the optimal K estimate. We re-ran CC using K = 2 and K = 3, applied average-linkage hierarchical clustering to the consensus matrices, and re-displayed the same correlation matrix in **Figure 1c** with the hierarchical-clustering-based order for K=2 (**Supplementary Fig. 1a**), K=3 (**Supplementary Fig. 1b**) and with the order from the first principal component scores (**Supplementary Fig. 1c**). Each panel in **Supplementary Figure 1** showed interesting structure, suggesting that it is possible to re-order a sample-sample similarity matrix in different ways to support different claims regarding optimal K.

Supplementary Note 2: Quantitative comparisons of cluster strength between GBM and the null datasets

CC heatmaps in **Figure 1** and **3** allow visual comparisons of cluster strength. However, formal inference requires quantitative summaries for the presence of clusters under a range of possible Ks. Two such summaries are CLEST [6] and silhouette width [7]. Briefly, CLEST is a resampling-based method that randomly partitions the original dataset into a learning set and a test set. The former is used in an unsupervised clustering method to build a K-cluster classifier, which is applied to partition the latter (the test set) in supervised assignment. The test set is also partitioned independently using the same unsupervised clustering algorithm as applied for the training set. The concordance between the supervised and unsupervised partitions for the test set is summarized by measures such as the Fowlkes-Mallows (FM) index, for which a higher value indicates stronger clustering signals in the original data. Silhouette width is computed for each sample and each K based on the comparison of its distance to its own cluster and that to other clusters. A dataset with strong clusters tend to show a high average value of positive silhouette width, and fewer samples of negative silhouette width.

We apply these two methods to compare clustering strength between three TCGA datasets for GBM and the 50 pcNormal null datasets. The three datasets are: TCGA's first and second GBM cohort (**GBM1** and **GBM2,** respectively), and the **validation** dataset [2], which is a combination of data from multiple prior studies [8-11]. CLEST results (**Supplementary Fig. 3a**) show that, for all K values except 2, the three real datasets have higher FM values than the null datasets. GBM1, in particular, show the highest FM values, suggesting that GBM1 has more structure than the null datasets. However, K = 4 is not clearly the optimal number of clusters for GBM1, because the differences from the null datasets are comparable across K = 3 - 8.

Supplementary Figure 3b shows the average of positive silhouette widths on the x-axis and the fraction of samples with negative silhouette widths on the y-axis. Datasets with strong clustering signals are expected to appear on the lower-right side of the plot. At K = 4, GBM1 is within the distribution of the 50 pcNormal datasets along the y-axis, but is a positive outlier along the x-axis. We also added to this figure the results from simulated **positive** datasets with four known clusters pulled apart with controlled degrees of separation (as measured by a, explained below and in **Methods**). GBM1 is most similar to the dataset with a = 0.2. This result suggests that GBM1 has a certain clustering structure, however it does not confirm K = 4 as the optimal number of cluster in the range K = 2 - 6 using silhouette width statistics (**Supplementary Fig. 3c**). GBM2 and the Validation dataset are within the range of pcNormal for both axes, reinforcing the results from **Supplementary Figure 3a** that they have weaker structures than GBM1.

The apparent structure of GBM1 can be attributed to the fact that our simulations relied on normally distributed samples in the hyperspace and could not capture all the spatial features of the actual dataset. For example, **Figure 1d** showed "protrusions" of GBM1 samples towards two lower front corners of the PC1-PC2-PC3 cubic space. Such local formations in the hyperspace are difficult to match by simulations, partly explaining the observed difference between GBM1 and the null simulations (**Supplementary Fig. 3b**). In short, while some quantitative measures, such as CLEST and average silhouette width, could bring out the uniqueness of GBM1, other measures, such as the heatmaps in **Figure 1a-1b** and the negative

silhouette width fraction, could not. This underscores the fact that different clustering measures emphasize different features of a given heterogeneous dataset.

References

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
2. Verhaak, R.G.*, et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).
3. Li, B.*, et al.* Genomic estimates of aneuploid content in glioblastoma multiforme and improved classification. *Clin Cancer Res* **18**, 5595-5605 (2012).
4. Sturm, D.*, et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425-437 (2012).
5. Brennan, C.W.*, et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-477 (2013).
6. Dudoit, S. & Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**(2002).
7. Rousseeuw, P.J. Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. *Journal of Computational and Applied Mathematics* **20**, 53-65 (1987).
8. Beroukhim, R.*, et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007-20012 (2007).
9. Murat, A.*, et al.* Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J Clin Oncol* **26**, 3015-3024 (2008).
10. Phillips, H.S.*, et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173 (2006).
11. Sun, L.*, et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9**, 287-300 (2006).