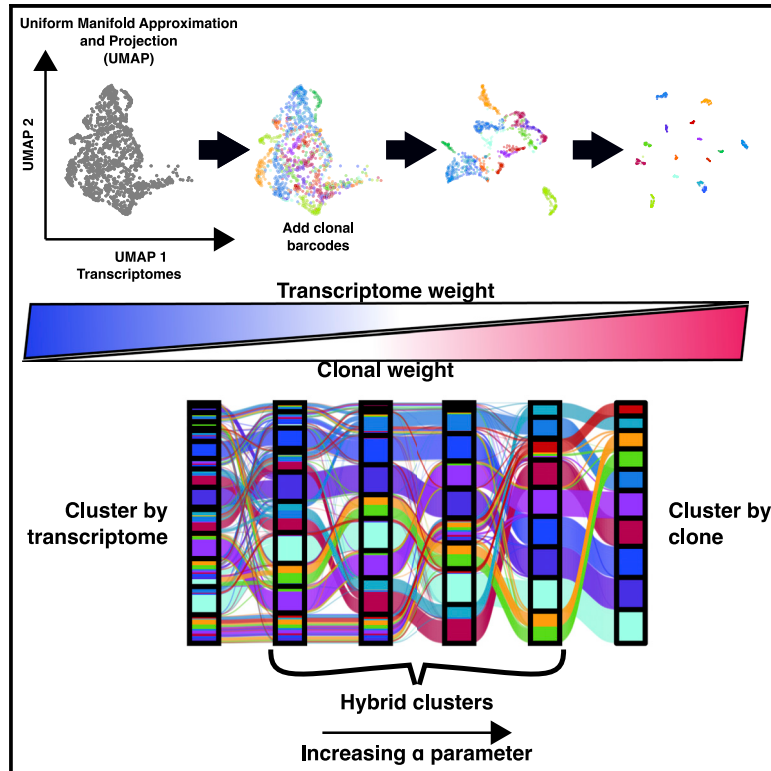


ClonoCluster: A method for using clonal origin to inform transcriptome clustering

Graphical abstract



Authors

Lee P. Richman, Yogesh Goyal, Connie L. Jiang, Arjun Raj

Correspondence

arjunrajlab@gmail.com

In brief

Novel cellular barcoding methods in single-cell RNA sequencing provide simultaneous clustering substrates: transcriptomic and clonal information. Richman et al. developed *ClonoCluster* to combine these approaches to produce hybrid clusters. Hybrid clusters are separated by biologically significant markers and reduce entropy of labeled cell types across clusters.

Highlights

- Clonal information represents an alternative clustering approach to transcriptomes
- *ClonoCluster* combines transcriptomic and clonal data to generate hybrid clusters
- Incorporating clonal data rearranges clusters via ECM and ribosomal gene expression
- Hybrid clusters identify biologically relevant markers and reduce cell type entropy



Technology

ClonoCluster: A method for using clonal origin to inform transcriptome clustering

Lee P. Richman,^{1,2} Yogesh Goyal,^{3,4,5} Connie L. Jiang,⁶ and Arjun Raj^{2,3,7,*}¹Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA²Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA³Department of Bioengineering, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, PA, USA⁴Department of Cell and Developmental Biology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA⁵Center for Synthetic Biology, Northwestern University, Chicago, IL, USA⁶Genetics and Epigenetics, Cell and Molecular Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA⁷Lead contact*Correspondence: arjunrajlab@gmail.com<https://doi.org/10.1016/j.xgen.2022.100247>

SUMMARY

Clustering cells based on their high-dimensional profiles is an important data reduction process by which researchers infer distinct cellular states. The advent of cellular barcoding, however, provides an alternative means by which to group cells: by their clonal origin. We developed ClonoCluster, a computational method that combines both clone and transcriptome information to create hybrid clusters that weight both kinds of data with a tunable parameter. We generated hybrid clusters across six independent datasets and found that ClonoCluster generated qualitatively different clusters in all cases. The markers of these hybrid clusters were different but had equivalent fidelity to transcriptome-only clusters. The genes most strongly associated with the rearrangements in hybrid clusters were ribosomal function and extracellular matrix genes. We also developed the complementary tool Warp Factor that incorporates clone information in popular 2D visualization techniques like UMAP. Integrating ClonoCluster and Warp Factor revealed biologically relevant markers of cell identity.

INTRODUCTION

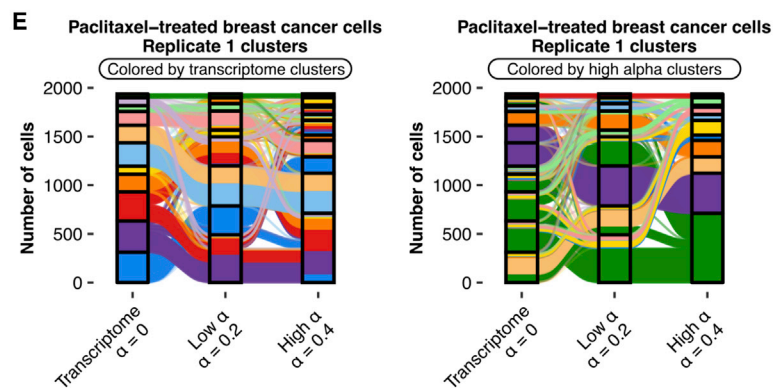
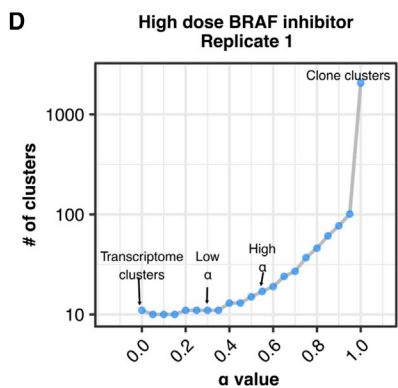
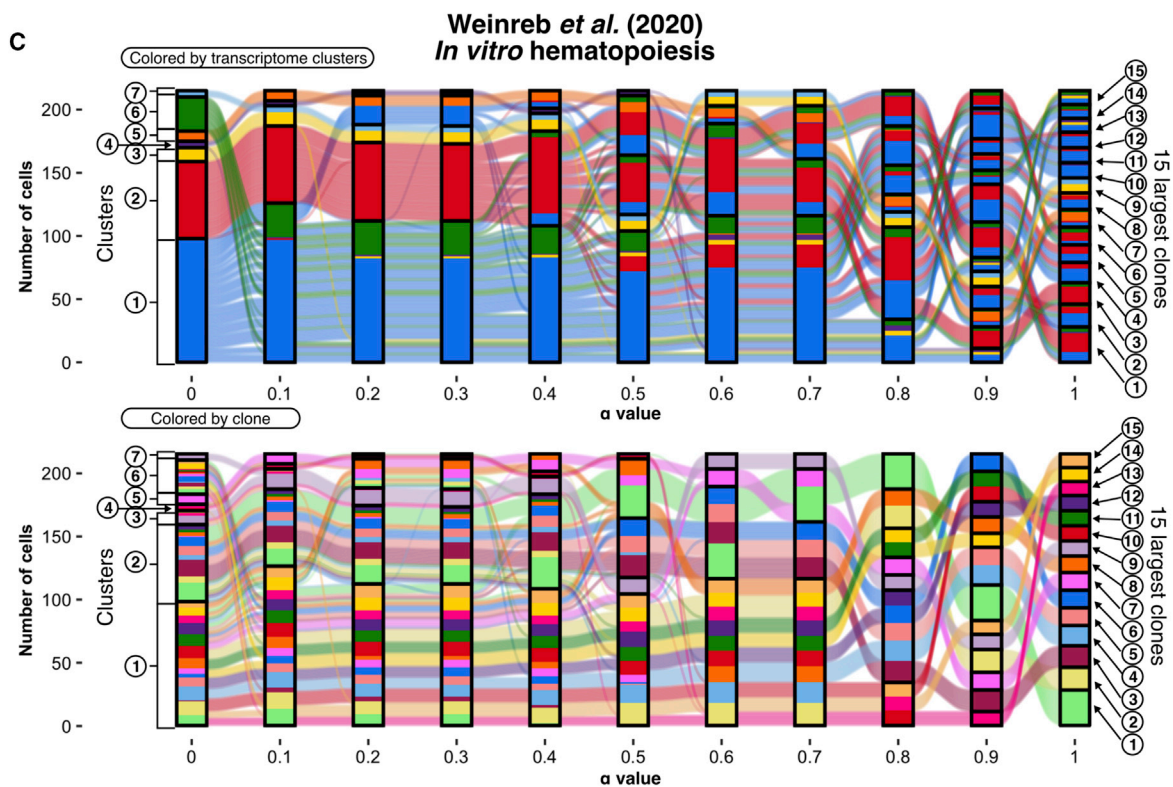
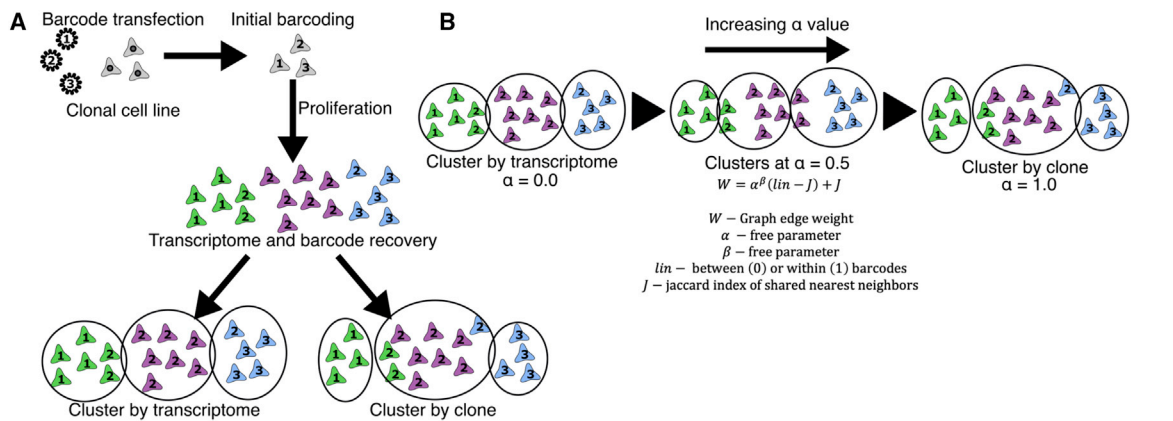
Since the advent of high-dimensional molecular profiling, clustering has been the most common form of data analysis and visualization applied, allowing one to form groups out of entities with similar profiles.^{1,2} Clustering has enabled the detection of networks of disease genes and loss of function patterns across cancer cell lines.^{3,4} More recently, the development of single-cell measurement technologies has allowed the high-dimensional profiling of individual cells. In this context, clustering has been used to categorize cells into discrete molecular states, often termed “fates” or “types,” that can be associated with distinct cellular functions.^{5,6} At the same time, in many biological contexts, single cells also have lineage relationships; i.e., they may arise from a common progenitor. This information in principle provides a complementary way to cluster profiles of single cells, but it has not been incorporated into the systematics of single-cell profiles.⁷

Clustering purely based on molecular profiling has been successful in many instances but relies on a number of assumptions that have been difficult to evaluate rigorously. Briefly, most clustering approaches start with some form of feature selection or dimensionality reduction. Data points are then grouped to minimize the distances between elements of the same groups, by k-means clustering, hierarchical clustering, or graph-based community detection methods that use connectivity to inform

group membership.^{5,6,8} Intrinsic to all these methods is the use of some sort of distance between cells measured by some metric of distance between their molecular profiles, but in principle many other types of information could be used to inform or modify these distances.

Recently, the development of cellular barcoding systems has provided an alternative means by which to group cells. Primarily, these systems have been used for the longitudinal tracking of molecular profiles through some sort of biological process, such as differentiation or therapy resistance.^{9–15} Strikingly, these studies have concluded, in some instances, that single-cell differences in the expression of few or even just one gene can predict the ultimate fate of the cell, differences that would not have been detected by standard clustering algorithms. For example, this approach identified the expression of *Mettl7a1*, a methyltransferase, as a driver of successful stem cell reprogramming,¹³ identified TCF15 as necessary and sufficient for hematopoietic stem cell self-renewal, and identified *Pou2f2* expression in progenitor cells as predicting DC-like versus neutrophil-like monocyte fates.^{12,16,17} Such results suggest that the incorporation of clone information could be very useful in determining the expression differences between cells that have distinct functional outcomes. Being able to weigh both transcriptome and clone information in clustering algorithms would potentially enable the identification of such “hidden” factors.





(legend on next page)

We thus developed an algorithm that we call “ClonoCluster” that integrates transcriptome and clone barcode information, allowing one to cluster cells using a continuous parameter (α) that adjusts the relative weight of transcriptome versus clone information. We applied ClonoCluster to six previously published independent single-cell RNA sequencing datasets, including *in vitro* hematopoiesis, directed stem cell differentiation, and drug treatment of tumor cell lines.^{11,12,18} We found massive rearrangements of the assignments of cells to clusters as α was shifted to weigh clonal origin more heavily. These rearrangements had novel, potentially more biologically interpretable, cluster markers, and were associated with expression of genes involved in extracellular matrix production and translation. These results held across datasets where clone fate was determined intrinsically, and the effects were considerably less strong in the dataset where cell fate was determined extrinsically. Inspired by this clone-weighted network graph clustering approach, we developed a tunable parameter (the Warp Factor, ranging from 0 to 10), that incorporates clonality information into the dimensionality reduction step prior to the commonly used Uniform Manifold Approximation and Projection (UMAP) algorithm for visualizing high-dimensional datasets. We included ClonoCluster and Warp Factor into an open-source R package, *ClonoCluster* (<https://github.com/leeprichman/ClonoCluster>). As barcoding data become more prevalent, ClonoCluster can provide a means to evaluate the degree to which clustering can be altered by factoring in clonal origin.

Design

ClonoCluster integrates clone barcode and transcriptome information

Clone barcode assignment and transcriptome-level data represent two different modalities of data that can be used to cluster single-cell RNA sequencing. In a prototypic clonal barcoding experiment, a population of cells is transfected with random transcribed barcodes such that each initial clone is likely to express a unique barcode. After proliferation, experimenters apply some additional experimental conditions, such as drug treatment or differentiation.^{11–14,18} At the chosen endpoint, one can perform single-cell RNA sequencing on a pool of individual cells, which is itself composed of some number of clones marked by barcodes. The barcodes themselves can be determined by

various side reactions and subsequent sequencing, thus adding a clone identifier to each cell’s transcriptome. (In practice, technical constraints on clone identification and sampling for single-cell RNA sequencing mean that only some subsets of sequenced cells will have an identifiable barcode.)

Once cells have both transcriptome and clone information attached to them, one can compare methods of classification. Two popular software packages for classifying cells by transcriptome information alone are *Seurat* and *scanpy*,^{19,20} both of which apply community detection algorithms to network graphs to identify the most interconnected cell clusters. We can then directly compare the classification of cells by transcriptome clusters versus clone barcodes (Figure 1A). In principle, these two classification schemes could be virtually identical, or they could be completely uncorrelated with each other.

The tunable parameter α yields hybrid clone-transcriptome defined clusters

We wondered whether there was some way to incorporate both clone and transcriptome information to generate “hybrid” clusters that group cells that balance transcriptomic similarity and clonal relationships. In order to generate such hybrid clusters, we developed the ClonoCluster model for measuring similarity between cells. This model includes a tunable parameter, α , that shifts weight between clustering by cell transcriptomes alone ($\alpha = 0$) and “clustering” by clone barcode alone ($\alpha = 1$) (Figure 1B). In well-established single-cell RNA sequencing analysis packages like *scanpy* and *Seurat*, the algorithms build a network graph of cells (nodes) connected by edges that are weighted by transcriptional similarity (“transcriptome weight”), determined by the number of shared nearest neighbors in principle component space.^{19,20} The clustering itself is then determined by community detection within this graph, returning the most highly interconnected groupings of cells as the assigned clusters. In ClonoCluster, we retained this overall structure, incorporating clone information by modifying the weights as follows. For each edge between cells, we also created a “clone weight” of 1 or 0 based on whether the cells have the same or different barcodes. We then normalized the “clone weight” by the number of cells with that barcode to ensure that the attractive “force” of barcodes did not scale with the number of cells. We linearly combined transcriptome and clone weights using α such that it returned purely transcriptome weights for $\alpha = 0$ and

Figure 1. The ClonoCluster method integrates transcriptome and clone clustering modalities in a tunable manner using the α parameter

- (A) Schematic depicting a generic approach to single-cell barcoding that yields output clusterable by two modalities of data, transcriptome clustering, and clone clustering by recovered barcode.
- (B) Schematic depicting the integration of these clone and transcriptome clustering modalities using the ClonoCluster method, in which the transcriptome nearest neighbor network graph edge weights are modified to incorporate clone clusters with a tunable free parameter, α . At $\alpha = 0$, clustering is identical to traditional transcriptome clusters. At $\alpha = 1$, clusters are consistent with clone barcode assignments.
- (C) Sankey diagram depicting the reorganization of cell clusters from the 15 largest clone clusters present at day 2 in an *in vitro* hematopoiesis assay¹² with increasing α value colored by initial transcriptome cluster (top) and clone (bottom). Nodes/boxes represent clusters and ribbons depict the flow of cells between clusters.
- (D) Representative plot of high-dose BRAF inhibitor-treated melanoma clonal cell line (clonal WM989 cells treated with 1 μ M of the BRAF inhibitor vemurafenib¹¹) showing that cluster number approaches unique clone barcode number with increasing α value at fixed community detection resolution. “High α ” values were chosen based on the approximate shoulder of the curve after which clusters rapidly break into individual clone clusters with increasing α . “Low α ” values were determined by the half “high α ” value rounded to the nearest 10th value. At $\alpha = 0$, clustering is identical to traditional transcriptome clusters. At $\alpha = 1$, clusters are consistent with clone barcode assignments.
- (E) Representative Sankey diagram of the 15 largest clone clusters from the high-dose BRAF inhibitor dataset treated melanoma clonal cell line depicting rearrangements of clusters at transcriptome, low α , and high α levels, colored by initial transcriptome cluster assignment (left) and high α cluster assignment (right).

Table 1. Dataset descriptions and sources

Dataset name	Source	Description	No. of Replicates	Fate determination
Low-dose BRAF inhibitor	Goyal et al. <i>bioRxiv</i> (2021) ¹¹	WM989 A6-G3 clonal melanoma cells treated with 100 nM vemurafenib for 3–4 weeks	1	Intrinsic
High-dose BRAF inhibitor	Goyal et al. <i>bioRxiv</i> (2021) ¹¹	WM989 A6-G3 clonal melanoma cells treated with 1 μ M vemurafenib 3–4 weeks	2	Intrinsic
Cario-directed iPSCs	Jiang et al. <i>Genome Biol.</i> (2022) ¹⁸	PENN123i-SV20 human iPSC line directed toward cardiomyocyte fate sequenced on day 14	1	Extrinsic
<i>In vitro</i> hematopoiesis	Weinreb et al. <i>Science</i> (2020) ¹²	Hematopoietic progenitor cells derived from murine bone marrow differentiating in culture for 2 days	1	Intrinsic
WM983B BRAF inhibitor	Goyal et al. <i>bioRxiv</i> (2021) ¹¹	WM983B E6-C6 clonal melanoma cells treated with 100 nM vemurafenib for 3–4 weeks	2	Intrinsic
Paclitaxel-treated breast cancer cells	Goyal et al. <i>bioRxiv</i> (2021) ¹¹	MDA-MB-231D4 clonal breast cancer cells treated with 1 nM paclitaxel for 3–4 weeks	2	Intrinsic

purely clone weights for $\alpha = 1$ (Figure S1A). We could then use this graph for cluster assignment just as is done by the conventional algorithms (Figure S1B). Thus, for values of α in between 1 and 0, ClonoCluster provides hybrid clusters that weigh both clone and transcriptome information.

RESULTS

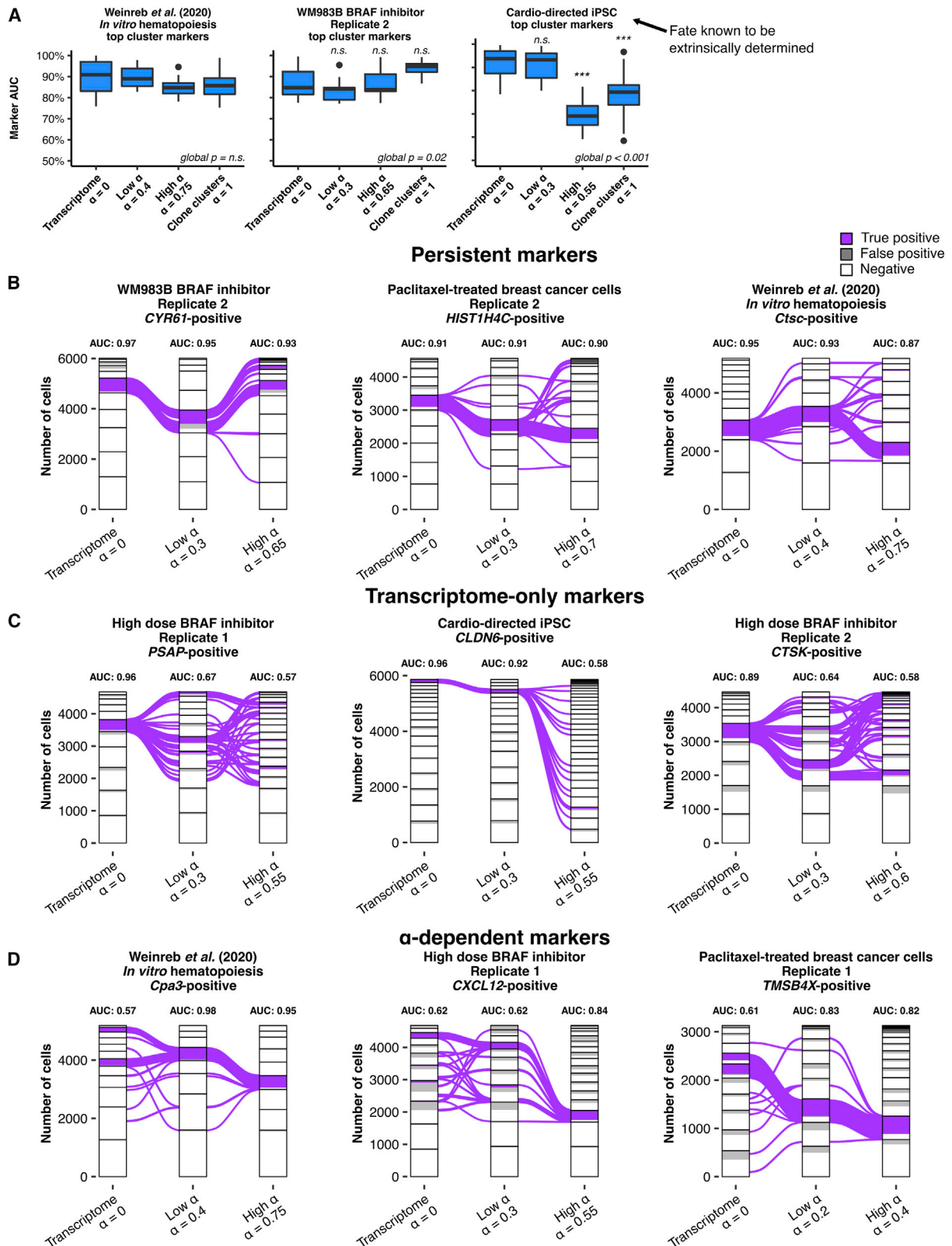
Clone-to-cluster concordance variably increases with α

We used ClonoCluster to generate hybrid clusters from six different clone barcoded single-cell RNA sequencing datasets from our lab and others^{11,12,18} (see STAR methods and Table 1). We first performed hybrid clustering at stepwise values of α from traditional transcriptome-only clusters ($\alpha = 0$) to the clone groupings ($\alpha = 1$). We wanted to determine how distinct hybrid clusters were from transcriptome or clone clustering alone as well as how increasing α reorganized clusters from initial transcriptome clusters to individual clone clusters. In order to visualize the flow of cells through these progressive hybrid clusters, we constructed Sankey diagrams²¹ of the datasets using the 15 largest clone clusters per sample (Figures 1C and S2A). There was visible reorganization of cells throughout clusters among the top 15 clone clusters as α increased, with differences in cluster number, composition, and size throughout much of the range of α in all samples. As expected, when α approached 1, the concordance between hybrid clusters and clone clusters increased, meaning that hybrid clusters were largely composed of individual clones or combinations thereof, with the frequency of clones being split across clusters becoming less. (At $\alpha = 1$, clusters were equivalent to clone groupings.) Within each dataset, cells with the same clonal origin (i.e., sharing a barcode) demonstrated variable degrees of reorganization into concordant clusters at given levels of α , with some requiring higher α values to completely unify into a single hybrid cluster than others.

The patterns of rearrangement also varied between datasets. For example, the cardio-directed iPSC dataset showed less

concordance between clone clusters and transcriptome clusters among the top 15 largest clone clusters compared with WM989 A6-G3 melanoma cells treated with high-dose vemurafenib. In the cardio-directed iPSC dataset, each of the top clone clusters segregated to a separate cluster by $\alpha = 0.6$, compared with $\alpha = 0.8$ or 0.9 in the two melanoma replicates (Figure S2A). We formally evaluated clone-to-cluster concordance using Cohen's κ , a measure of agreement between classifications (grouping by clone versus grouping by assigned cluster) computed from the 2×2 confusion matrix of clone barcode and cluster membership. When clonal cells sharing a barcode are more frequently assigned to a single cluster, Cohen's κ is high. (At $\alpha = 1$ where hybrid clusters and clonal clusters are equivalent, Cohen's κ must also equal the maximum value of 1.) For most datasets, we observed a gradual transition of Cohen's κ from 0 to 1 as α increased, suggesting that there was a graded conglomeration of clone barcodes within hybrid clusters as clonality was increasingly factored into the clustering algorithm. The cardiomyocyte-directed iPSCs, however, displayed a very sudden switch between low κ and high κ (Figure S2B), suggesting that there are few meaningful intermediate hybrid clusters between purely transcriptome clustering and purely clone clustering. Sankey visualization matched this interpretation by showing that clones were quite intermixed between the transcriptome-only clusters, with little obvious areas of partial conglomeration. In the cardiomyocyte-directed iPSC case, biologically, our prior findings suggested that cell fate was extrinsically determined and did not correlate well with clonality; hence, hybrid clusters would increasingly be composed of cells of essentially random types, explaining these observations for this dataset.

As α approaches 1, the number of hybrid clusters identified approaches the large number of individual clone clusters in the dataset, which are generally more numerous than the standard transcriptome clusters returned using common methods. We wondered if we could identify the maximum α value that



(legend on next page)

generates a number of hybrid clusters similar to the number of transcriptome clusters that also showed significant reorganization of cells to be distinct from transcriptome-only clustering. We therefore computed the number of clusters returned in each sample across α values to identify the maximum α value that still returned a number of clusters in this range; the final cluster composition and cluster number was determined by the Louvain algorithm as implemented in the *Seurat* package.^{19,22} The Louvain algorithm iteratively severs network graph edges to achieve optimal internal interconnectedness of community members (modularity) and returns a greater number of communities at higher values for the input resolution parameter. We performed all clustering at fixed resolution, therefore the rising number of clusters with increasing α value can be attributed to improved modularity of the clusters as networks break apart into isolated networks of like-barcoded clones. “High α ” values were chosen based on the maximum value before the inflection point where cluster numbers dramatically increase, with “low α ” values being chosen at half this value (Figures 1D and S3). The high α value varied between samples, from 0.4 to 0.7. Even within this restricted range that limited the number of clusters returned, we observed reorganization at low and high α for the top 15 clone clusters (Figure 1E).

Hybrid α clusters reveal novel cluster markers without loss of marker fidelity

A common goal of single-cell RNA sequencing analysis is the identification of “marker” genes whose expression is high in cells of a particular cluster (i.e., sensitive for cluster membership) and is low in cells in other clusters (i.e., specific for cluster membership).^{5,6,23} We consider the most useful markers to be those that are specific and sensitive to the cells in a cluster and thus have “high marker fidelity.” Given that increasing α altered cluster number and membership, we wondered whether the genes that served as the best markers changed with α , and whether those markers were able to maintain similar sensitivity and specificity as they did on purely transcriptomically defined clusters. We used the receiver operating characteristic (ROC) to evaluate all expressed genes as markers to classify cells into clusters at all possible marker expression cutoffs. We used area-under-the-curve (AUC) derived from the ROC to summarize the fidelity of a marker by quantifying its sensitivity and specificity as a marker. An AUC of 0.5 indicates a classifier that is no better than random chance at determining cluster membership of a cell, with a value of 1 being perfectly predictive. We identified the cluster markers with the highest AUC for each hybrid cluster at each α and found that the top markers

of the clusters were different for different α values (Figures S4 and S5A).

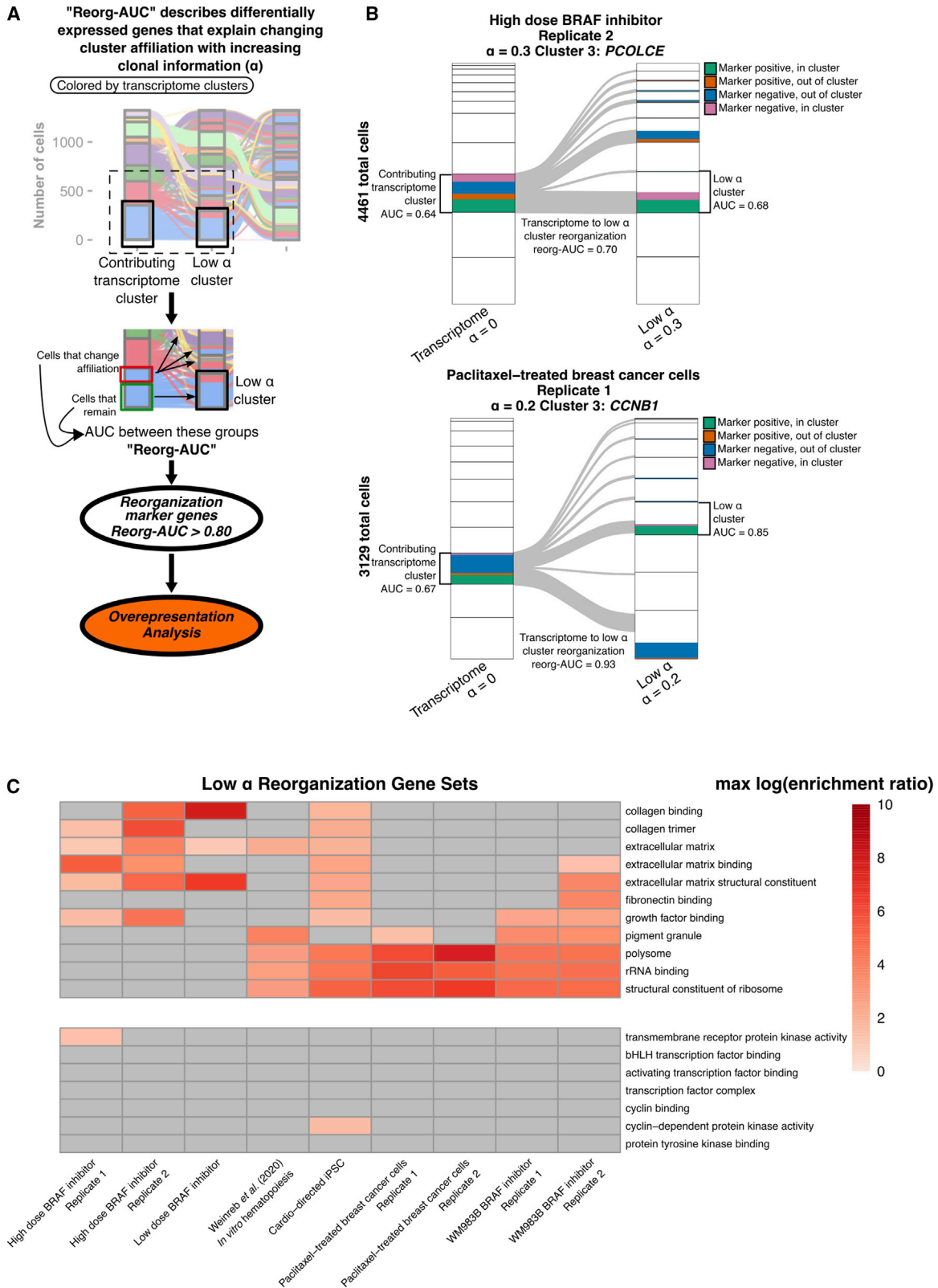
With this observed change in top markers, we wondered whether overall fidelity of the top cluster markers at each α , as reflected by AUC, was preserved with hybrid clustering. We found that across almost all datasets, the median AUC of top cluster markers at each value of α were not significantly different from transcriptome clustering alone (Figures 2A and S5B), suggesting that the rearrangements caused by incorporating clone information did not radically decrease the ability to find a single marker for a cluster, despite the fact that the markers themselves changed. We did see decreased marker fidelity in the cardiomyocyte-directed iPSCs at the high α and clonal cluster levels (0.69 and 0.79 versus 0.94). (Low-dose vemurafenib-treated melanoma samples also had a decrease in marker fidelity, although to a lesser extent, 0.80 and 0.84 versus 0.90.) Again, given the extrinsic determination in cardiomyocyte-directed iPSCs, we would expect that markers for hybrid clusters would be hard to find; indeed, many putative markers had poorer performance at high α ¹⁸ (Figure S5A). Thus, in the datasets analyzed in which cell states are thought to be intrinsically determined (Table 1), hybrid clustering yielded new cluster markers revealed by clonality information, often of equal fidelity to transcriptome clustering. The alternative markers identified by ClonoCluster potentially represent genes whose expression varies more between different clonal populations than within a clonal population.

In addition to top cluster markers, we wondered how marker fidelity of any single marker was likely to change as we increased clonal weight with α . We used Sankey plots to directly visualize the flow of marker-positive cells through hybrid clusters as α changed. We found that the fidelity of transcriptome cluster markers varied as α changed, with some markers persisting across all values of α and others changing their fidelity dramatically. Many markers maintained their fidelity over increasing α values (representative samples shown in Figure 2B). We also identified markers that were strong cluster classifiers at the transcriptome cluster level that lost fidelity in low and high α clusters (Figure 2C), suggesting that the classification properties of these markers based on transcriptomes alone was lessened by the addition of clone information. Conversely, we also observed markers that increased in fidelity with increasing α , meaning that those markers were highly expressed in clones that were brought together into a hybrid cluster (Figure 2D). Analysis of marker fidelity across α therefore identified new sets of markers that balance clonal differences with transcriptome differences in cluster assignment, especially favoring markers that are strongly congruent in both.

Figure 2. Manipulation of α reveals turnover of cluster markers

(A) AUC values for the top marker per cluster at zero, low, high, and maximum ($\alpha = 1$) values of α for multiple samples, an *in vitro* murine hematopoiesis assay at day 2, a BRAF inhibitor-treated clonal melanoma cell line WM989B, and directed differentiation of induced pluripotent stem cells (iPSCs) toward a cardiomyocyte fate, a system where extrinsic determinants of cell fate are expected to be dominant over clone barcodes.¹⁸ “Global p” indicates the p value for the non-parametric Kruskal-Wallis test with Bonferroni correction. When “global p” was less than 0.05, pairwise Wilcoxon tests were performed with Bonferroni correction. Annotations above boxes indicate the corrected Wilcoxon test p value compared with $\alpha = 0$ (n.s., not significant; ***p < 0.001).

(B–D) Sankey diagram including all barcoded cells from various samples (see STAR Methods) depicting transcriptome clusters, low α clusters, and high α clusters. Cells positive for the respective marker that are present in the cluster of interest are marked purple (“True Positive”) while cells positive for the marker but not present in the cluster of interest are marked gray (“False Positive”). Positivity thresholds were determined as described in STAR Methods. Area-under-the-curve (AUC) for the marker is annotated above the cluster nodes. Representative markers are chosen to demonstrate markers whose classifier strength persists across α values (B), markers that are only strong in transcriptome clusters (C), and markers that are stronger at high and low α values (D).



Reorganization of clusters with α is explained by differential expression of extracellular matrix and translation-associated genes

To identify expression of genes associated with reorganization of cells from transcriptome clusters to hybrid clusters, we evaluated genes as classifiers for explaining rearrangement of cells. [Figure 3A](#) demonstrates our approach to comparing reorganizing cells to delineate these differentially expressed genes. For all possible transcriptome cluster and low α cluster pairs, we calculated the AUC for each gene in the dataset as a predictor of whether a cell from the transcriptome cluster would become a member of the low α cluster. Genes with high AUC in this analysis were strong predictors of whether or not a cell would switch affiliation to a particular low α cluster, and thus represent the differentially expressed genes that explain cluster rearrangements that either break apart or aggregate cells from the transcriptome clusters upon the incorporation of clonal information to the clustering algorithm. We have termed this AUC of a differentially expressed gene that explains the sorting of clusters into hybrid α clusters as the reorganization AUC (“reorg-AUC”). Any gene with reorg-AUC greater than 0.80 for any transcriptome to hybrid α cluster was considered a “reorganization marker.” Representative examples of markers associated with contributing cells from a transcriptome cluster and a low α cluster of interest are shown in [Figure 3B](#). Reorganization markers were identified in this way for all datasets. Each set of reorganization markers for a paired hybrid cluster and contributing transcriptome clusters was then used to perform an overrepresentation analysis.²⁴ Multiple gene sets were significantly enriched in reorganization markers from all samples. For visualization, we generated a heatmap of the maximum enrichment ratio for all gene sets significantly overrepresented in three or more samples, as well as several commonly chosen gene sets to serve as negative controls ([Figure 3C](#)). The gene sets shared across the greatest number of samples were associated with translation, “polysome,” “rRNA binding,” and “structural constituent of ribosome,” as well as many samples showing enrichment for genes associated with the extracellular matrix, including “extracellular matrix,” “extracellular matrix structural constituent,” “extracellular matrix binding,” “collagen trimer,” “collagen binding,” and “fibronectin binding.” Enrichment was generally not seen in the negative control gene sets selected. These analyses reveal that the cluster reorganization induced by the incorporation of clonality information, even at low levels of α , are not driven by random gene sets, but rather by specific biological processes, such as those related to the extracellular matrix and translation. Remarkably, these results held across independent datasets taken from very different bio-

logical systems, suggesting that the biological processes associated with extracellular matrix and translation may help define a “fingerprint” for clonal information.

Warp Factor, s , modifies UMAP representations to enhance clone separation

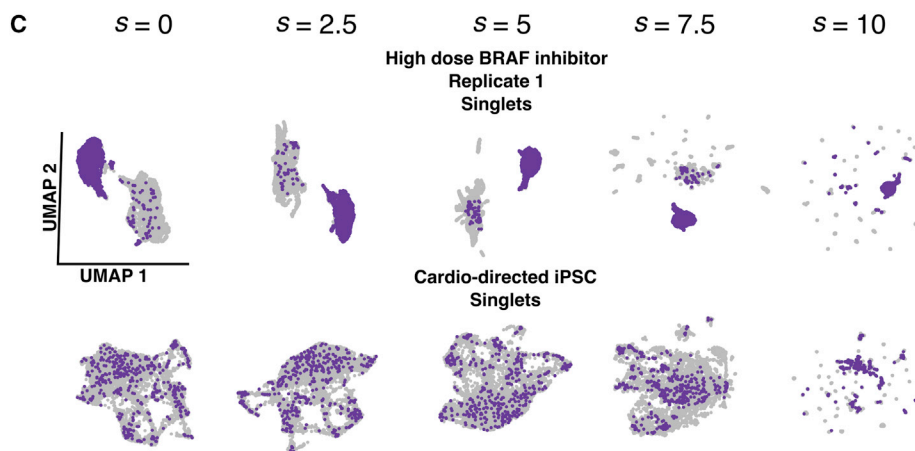
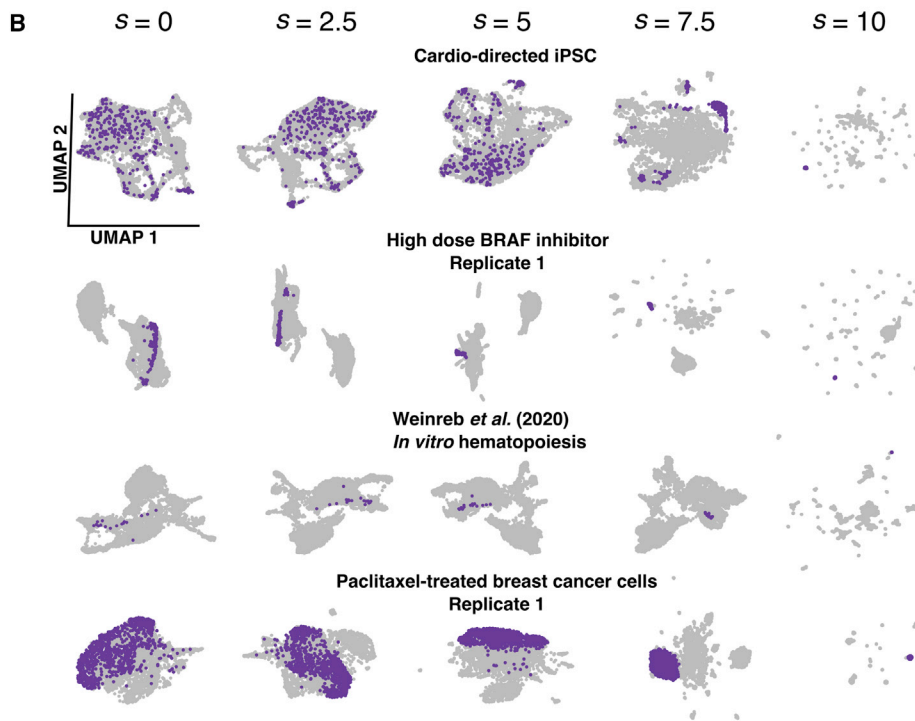
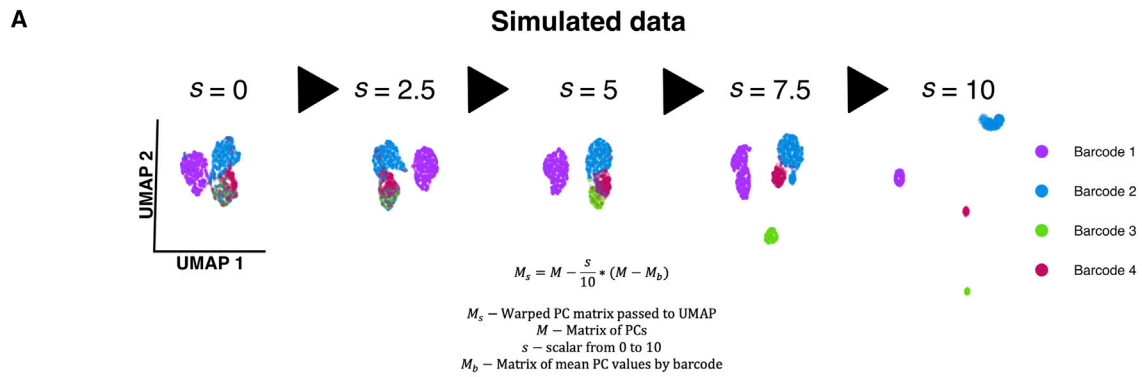
In addition to clustering and marker identification, an independent step in single-cell RNA sequencing analysis is two-dimensional visual representation of transcriptome data using dimensionality reduction techniques such as t-SNE or UMAP.^{6,25} In the same spirit of hybrid clustering, we wondered whether we could apply the principle of adding weighted clone information to these low-dimensional visual embeddings. Whereas α modifies the weight of edges in the network graph input to the clustering algorithm, we sought to construct a method to incorporate clone information with a tunable parameter, the Warp Factor, into the input to the UMAP algorithm. UMAP represents variation in the transcript count matrix-derived principal-component analysis (PCA) in a single manifold using attraction and repulsion components. Two UMAP dimensions are usually projected to visualize the high-dimensional data.^{6,25} We modified the PCA input to UMAP to incorporate clone barcode information alongside transcriptome variation. To test this approach, we simulated data and used a model to create a modified principal component (PC) matrix in which a tunable “Warp Factor” parameter, s , warps the values of PCs for each cell to approach the mean value of the principal component for the clone cluster it belongs to ([Figure S6A](#)); i.e., reducing the variance of cells within the clone. At $s = 0$, the PC matrix is unmodified and at the maximum value, $s = 10$, the only variation present in the data will be between clone clusters ([Figure S6B](#)). The incorporation of the Warp Factor into visualization promoted separation of clone clusters in UMAP space in simulated data ([Figure 4A](#)). We then used the Warp Factor for the visualization of clone clusters in the datasets. As expected, the spread between individual clone clusters reduced across the UMAP with increasing Warp Factor. When we engaged high Warp Factors, individual clone clusters formed distinct spatial clusters in UMAP space. The amount of Warp Factor required to promote distinct separation of individual clone clusters and singlets into isolated spatial groups varied between datasets, likely due to the differences in the size and number of clone clusters as well as initial degree of clone-to-transcriptome concordance ([Figure 4B](#)). Unexpectedly, we also observed that “singlets,” cells that are the only member of their clone cluster (have a unique barcode), conglomerated together into a large group, possibly through repulsion from the non-singlet clone clusters ([Figure 4C](#)). The Warp Factor method was effective at

Figure 3. Reorganization markers are enriched in translation and extracellular matrix-associated gene sets

(A) Schematic depicting approach to differential expression that explains cluster rearrangements with α modulation. For each hybrid α cluster and contributing transcriptome cluster pair, differential expression analysis is performed between cells inside and outside the α cluster, yielding an AUC for reorganization (reorg-AUC). Geneset overrepresentation analysis was then performed on these groups of differentially expressed genes on reorganizing cells with reorg-AUC >0.80 (see [STAR Methods](#)).

(B) Representative Sankey diagrams showing a low α cluster and a contributing transcriptome cluster with colors indicating proportion of cells classified by the marker and the associated cluster AUCs and reorg-AUC.

(C) Heatmap demonstrating maximum log enrichment ratio for gene sets significantly enriched by overrepresentation analysis of reorganization markers across nine different data sources. All gene sets enriched in three or more samples are shown (top) as well as several commonly explored gene sets as negative controls (bottom). Gray tiles indicate no statistically significant enrichment in the sample (false discovery rate >0.05).



(legend on next page)

incorporating clonal information into UMAP visualizations of single-cell RNA sequencing datasets by reducing within-clone cluster variation and increasing between-clone cluster variation in the underlying PC matrix.

Combining hybrid clustering and Warp Factor highlights unique clusters and markers

We sought to demonstrate the potential biological utility of ClonoCluster and Warp Factor when used together, focusing on data from melanoma¹¹ and hematopoiesis.¹² Starting with melanoma, we first identified top marker genes at a high α (α from 0.55 to 0.6); these markers were distinct from those identified by purely transcriptomic clustering (Figure 5A). Among these were *COL6A2* in low-dose and *COL6A1* in high-dose vemurafenib treatment, which was consistent with our findings that collagen-associated gene sets were enriched in reorganization markers (Figure 3C) in the three replicates of low- and high-dose vemurafenib-treated WM989 cells.¹¹ In the standard UMAP, cells that highly expressed these markers were dispersed, but upon engaging a Warp Factor of 5, these cells came together, demonstrating that Warp Factor can visually represent the results of incorporating clonal information.

We also performed a similar evaluation on the day 2 *in vitro* hematopoiesis data from Weinreb et al.¹² The original authors labeled the majority of these cells as “undifferentiated” at day 2, as they did not meet the expression level for conventional markers used for supervised cutoffs to define cell types.¹² Indeed, using conventional transcriptome clustering, many of the top cluster markers identified (*Alox5*, *Plac8*, *S100a9*, *Ptma*, and *Ctsc*) were not ones known to be associated with distinct cell types in this system. However, for high α ($\alpha = 0.75$), two markers emerged as top cluster markers (*Mpo* and *Prtn3*) that are known neutrophil-specific genes (Figure 5B).²⁶ Using a Warp Factor of 9.9 (likely needed due to the small number of cells per clone), we could pull these markers together to some degree.

The fact that ClonoCluster was able to recover more biologically meaningful markers suggests that there may be biological utility in its use. To further assess whether ClonoCluster generated more biologically meaningful groupings, we computed the entropy of distribution of the cells with known types based on cutoffs of marker genes. Lower entropy in this setting indicates that cells of the same type (biologically determined) are distributed across fewer clusters (computationally determined). We found that indeed, at low α , ClonoCluster reduced the entropy for several cell types (basophils, erythroid progenitors, lymphoid progenitors, mast cells, megakaryocytes, migratory dendritic cells, and plasmacytoid dendritic cells) as compared with clustering based on the transcriptome alone. As α was further increased, this decrease in entropy was often reduced or lost, suggesting that both clonal and transcriptomic information are needed to best reflect biologically meaningful clusters (Figure S7). The improvement in entropy for cluster distribution

among cells of the same type alongside the identification of more specific markers is promising evidence that hybrid clonal-transcriptome clusters better reflect known biology.

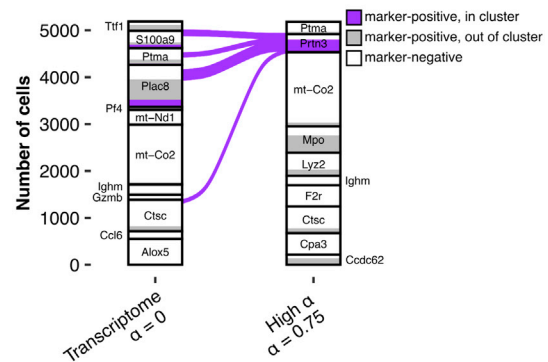
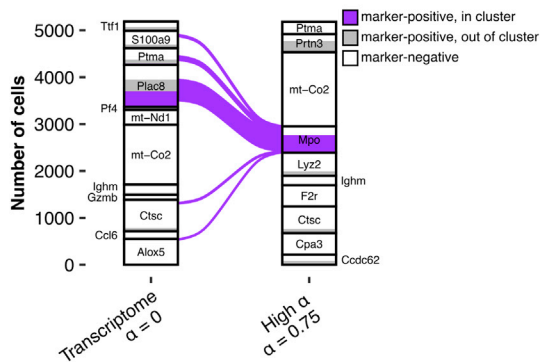
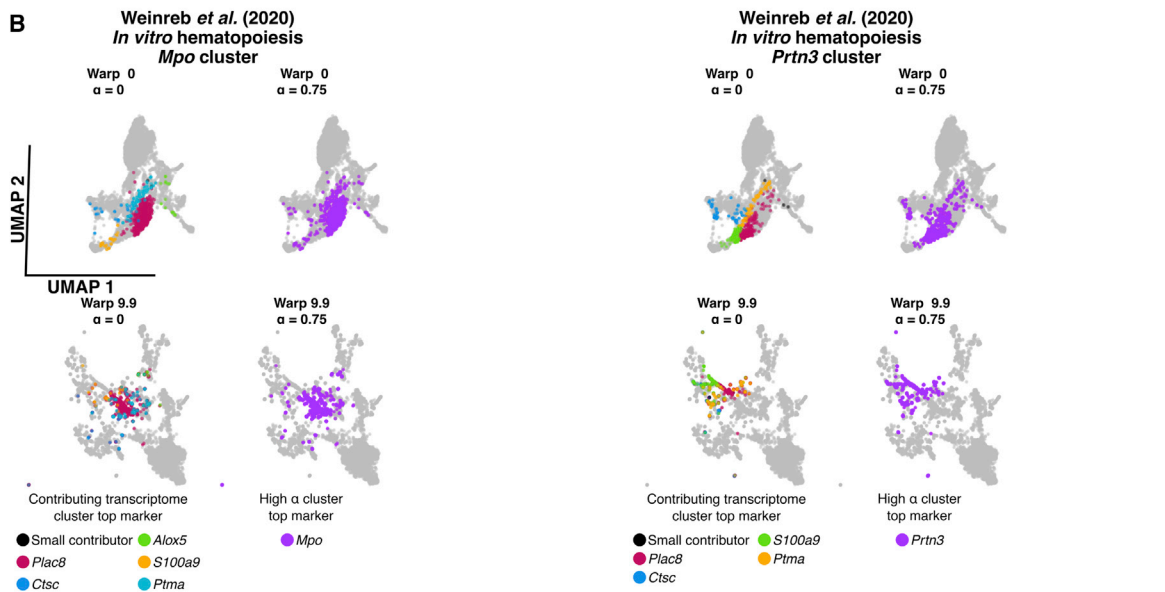
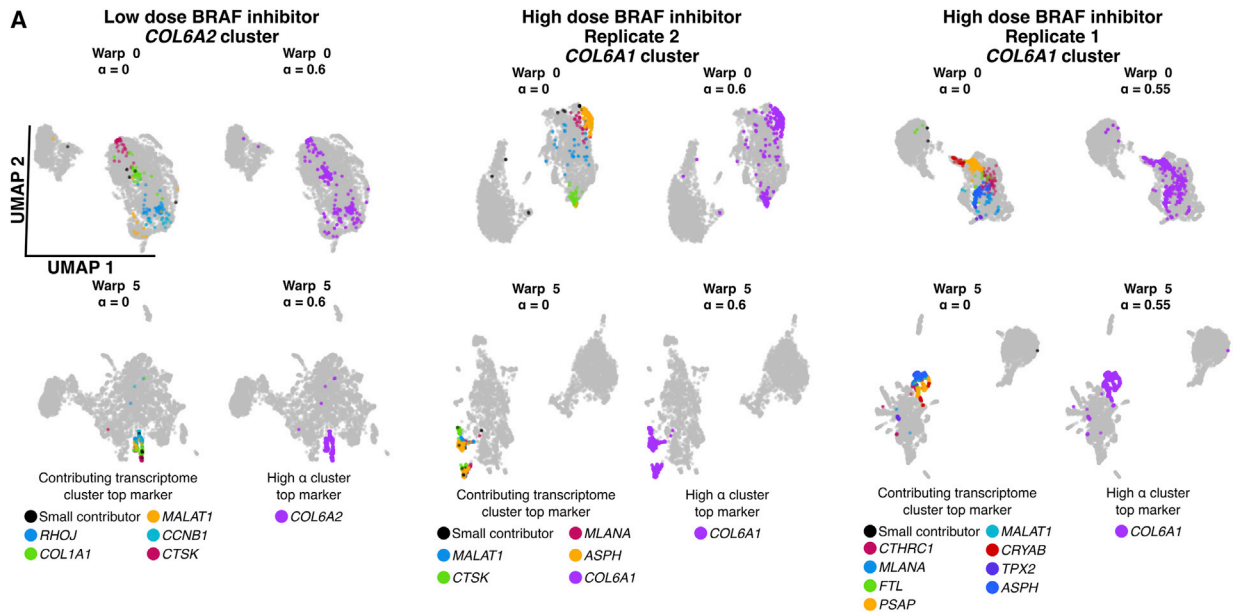
DISCUSSION

Clonal barcoding has provided additional information that could be used for clustering cells based on both clonal origin and transcriptome. ClonoCluster provides a method for weighting clone and transcriptome information—using a tunable parameter α —to generate hybrid clusters. These clusters were distinct from purely transcriptomically defined clusters, with unique sets of marker genes. The reorganization was often accompanied by the alignment of the expression of extracellular matrix proteins, suggesting that category of protein may be important for discriminating clones from each other. Further, we developed Warp Factor, inspired by ClonoCluster, as a way to modify the popular UMAP visualization to incorporate clone information.

A major question is whether these hybrid clusters more accurately reflect biological differences than transcriptomically defined clusters. We propose ClonoCluster as a method to tune the degree to which clonal information is incorporated. There are suggestions that some degree of clone information does reveal biological information. First, the fact that the assignment of cells to clusters changes dramatically as α is changed suggests that there is at least different information captured by clone information. Moreover, the top markers for clusters changed markedly, again suggesting different biological characterizations. It is hard to know what markers are biologically “correct,” but we do point out that in the case of the *in vitro* hematopoiesis dataset, we found that increased α led to the detection of *Mpo* and *Prtn3* as markers, which are well-known neutrophil-specific genes²⁶ that did not appear as a top cluster marker for pure transcriptomic clustering. Furthermore, we found that intermediate levels of α better recapitulated known biology by reducing the entropy of known cell types across computationally generated clusters. Moreover, in the case of cardiomyocyte-directed iPSCs from Jiang et al.,¹⁸ in which we know that cell fate is largely determined by extrinsic factors and thus less by clonal factors, we observed that marker fidelity decreased at the high α levels, providing a negative biological control for ClonoCluster. Our analysis demonstrates that in some systems, the addition of clonal information to clustering generates clusters that reduce the likely erroneous spreading of known cell types across groupings and yields markers that better correlate with known biology. Therefore, a thoughtful interrogation of the experimental system, considering the character of the markers identified in hybrid α clusters alongside the overall marker strength with increasing α , will help to determine if the addition of clonal information to clustering with ClonoCluster provides biologically meaningful insight. We expect that hybrid clustering will reveal important markers and groupings for many systems in

Figure 4. Warp Factor, s , is a tunable parameter to modify UMAP visualization to incorporate clone barcode information

(A) Demonstration of the effect of increasing Warp Factor (s) value on UMAP structure for simulated data of 3,000 cells with four barcodes and six principal components. Each UMAP axis is scaled and centered to allow comparison between facets.
(B) UMAPs with increasing Warp Factor for multiple datasets, highlighting a single large clone cluster in each.
(C) UMAPs with increasing Warp Factors showing singlets in two datasets. Singlets are cells that are the only cells in the sample with their unique barcode.



(legend on next page)

developmental biology and other areas where clonality also seems to correlate with cell fate decisions that would be difficult to obtain using transcriptome-based clustering alone. Recent work by Wang et al.²⁷ developing the Co-Spar algorithm robustly demonstrates the utility of incorporating clonal information to identify early fate biases in progenitor cells. The Co-Spar approach identifies clonal fates over multiple time points to approximate the transition manifold and assign likely progenitor states to even singly observed clones. This approach is powerful for identifying early states that commit to known fates by relying on data from multiple time points. In settings without such time series data, ClonoCluster provides a tunable time-independent approach to incorporating clonal and transcriptomic information into fate identification. Further development of tools will be needed as the number of experiments incorporating lineage tracing in different contexts continues to grow.

A potentially important consideration is whether the system in question is at an endpoint, or is still in the process of transitioning to its final state. It is in principle possible that the appropriate relative weighting between transcriptome and clone information may differ in these two contexts. Most likely, we think scenarios in which cell fate is determined by intrinsic states will benefit the most from ClonoCluster. That property may hold for both transitional or endpoint cell states, and hence will likely need to be evaluated on a case by case basis. Indeed, in our application of ClonoCluster to two transitional cell systems, hematopoiesis and directed cardio-myocyte differentiation, the former was intrinsically determined and ClonoCluster was informative, whereas the latter showed extrinsic determination and ClonoCluster was less informative. Consequently, we think that “twin” analysis (barcoded twin cells subjected to different local environments¹¹) are critical controls that should be performed to determine the relative contributions of intrinsic and extrinsic factors to help judge whether ClonoCluster will be a useful algorithm to apply.

What are the primary transcriptomic correlates with the rearrangements driven by incorporating clone information? We found these correlates were primarily associated with translation, ribosomal activity, and components of the extracellular matrix. These associations were found across diverse, independent datasets, suggesting that it is a common source of variation across the transcriptomes of clones. There is a known association between ribosomal gene expression and overall gene expression level that may be the source of its influence on cell clustering.⁶ Extracellular matrix may indicate some memory of initial microenvironmental differences that differ between clones. This memory would persist independent of subsequent microenvironmental differences, however, because “twin” experiments have shown that the same clone exposed to two

different microenvironments post-drug have virtually identical transcriptomes.¹¹ It is thus also possible that extracellular matrix proteins inherently reflect stable cell type or state identifiers.²⁸ It is also perhaps surprising that clonal information is carried by these sets of factors that are not transcription factors, which are more canonically thought of as the critical regulators of cell type. That difference suggests that extracellular matrix proteins may be more important for these subtle aspects of cell type determination than generally appreciated.

In general, the problem we describe here is in many ways analogous to whether species should be organized by genetic phylogeny or by phenotypic characteristics.²⁹ In some ways, categorizing species by phenotype would be analogous to clustering cells by their phenotype (i.e., transcriptome), whereas categorizing species by genetic phylogeny would be akin to clustering cells by lineage. Early zoology, predating genetic information, categorized species by observable phenotypes, but the addition of genetic information quickly grew to be an important form of categorization; by analogy, using clone information in cell clustering may similarly alter our classifications of cell types.

Limitations of the study

Ultimately, further testing of the homogeneity in the functional properties of cells within each cluster will be required to truly determine what clustering methods most closely match biological distinctions. Our studies did not reveal a reliable method to identify which experimental systems have intrinsic determinants of cell fate and therefore are most likely to yield more biologically meaningful clusters with the addition of clonal information. For now, we recommend an empiric approach to determine if the addition of clonal information is useful by first identifying the resolution value that returns the desired number of clusters from standard transcriptome-only clustering, and then plotting the number of clusters returned with increasing α . As α approaches 1, the number of hybrid clusters approaches the number of unique clone barcodes in the data. This visualization allows the user to determine the maximum α value that the experimental system will tolerate to generate a comparable number of clusters to transcriptome-only clustering while incorporating lineage information. Analysis of marker fidelity or comparison to ground truth data can be used to confirm that the α value produces hybrid clusters that are biologically meaningful. Example code to determine the optimal α value and an example protocol is provided in the *ClonoCluster* GitHub repository (<https://github.com/leeprichman/ClonoCluster/blob/main/Tutorial.Rmd>; <https://doi.org/10.5281/zenodo.3369197>.)

Furthermore, technology for reading out full-lineage data (as opposed to just clone data) has now been developed, often using CRISPR-Cas9 genome editing to make mutations in the genome

Figure 5. Combining hybrid clustering and Warp Factor highlights clusters with distinct markers in UMAP representations

(A) UMAPs highlighting a single high α cluster with the indicated top marker gene (*COL6A1* and *COL6A2*) in different replicates and doses of WM989 cells treated with vemurafenib.¹¹ Representations are stratified by α and Warp Factor. Gray points represent cells outside the high α cluster. Colors reflect the top marker gene of the high α cluster (purple) or the contributing transcriptome clusters. Contributing transcriptome ($\alpha = 0$) clusters were grouped as “small contributors” if fewer than 10 cells were present in the high α cluster of interest.

(B) UMAPs as in (A) for the day 2 *in vitro* hematopoiesis data¹² highlighting high α clusters marked by the granulocyte-enriched markers *Mpo* and *Prtn3* (top) and Sankey diagram depicting marker positivity at both clustering levels, nodes are annotated with top cluster markers. “Small contributors” indicates all contributing transcriptome clusters that contributed fewer than 10 cells to the high α cluster.

that can be read out by sequencing^{30–33} or imaging to generate the full cellular family tree.^{33–36} Our analysis did not address these types of data, although it is relatively straightforward to adapt ClonoCluster to these situations, in which one can weight the distance between cells by the relative phylogenetic distance. We expect that full-lineage tracing with time series information rather than the limited-resolution single time point data we explored will increase in prevalence, and ClonoCluster and like methods will need to adapt to this higher resolution setting.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **METHOD DETAILS**
 - Transcriptome and clone barcode integration with ClonoCluster
 - Cluster assignments
 - Clone barcode to cluster correlation analysis and visualization
 - Cluster marker identification
 - Reorganization marker and overrepresentation analysis
 - Modified principal components and UMAP visualization
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xgen.2022.100247>.

ACKNOWLEDGMENTS

Y.G. acknowledges support from the Burroughs Wellcome Fund Career Awards at the Scientific Interface, the Jane Coffin Childs Memorial Fund, and the Schmidt Science Fellowship. A.R. acknowledges support from NIH Director's Transformative Research Award R01 GM137425, NIH R01 CA238237, NIH R01 CA232256, NIH P30 CA016520, NIH SPORE P50 CA174523, and NIH U01 CA227550.

AUTHOR CONTRIBUTIONS

L.P.R., Y.G., and A.R. conceived and designed the project. L.P.R., Y.G., and C.L.J. compiled and preprocessed data. L.P.R. designed, performed, and analyzed all experiments, supervised by A.R. L.P.R. and A.R. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

A.R. receives royalties related to Stellaris RNA FISH probes.

Received: March 21, 2022

Revised: September 22, 2022

Accepted: December 16, 2022

Published: January 12, 2023

REFERENCES

1. Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* *95*, 14863–14868.
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* *96*, 6745–6750.
3. Tripathi, B., Parthasarathy, S., Sinha, H., Raman, K., and Ravindran, B. (2019). Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Front. Genet.* *10*, 164.
4. Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali, L.D., Gerath, W.F., Pantel, S.E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* *1*, 140035.
5. Peyvandipour, A., Shafi, A., Saberian, N., and Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Sci. Rep.* *10*, 12349.
6. Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* *20*, 273–282.
7. Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* *21*, 410–427.
8. Menon, V. (2018). Clustering single cells: a review of approaches on high- and low-depth single-cell RNA-seq data. *Brief. Funct. Genomics* *17*, 240–245.
9. Fennell, K.A., Vassiliadis, D., Lam, E.Y.N., Martelotto, L.G., Balic, J.J., Holizeck, S., Weber, T.S., Semple, T., Wang, Q., Miles, D.C., et al. (2022). Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature* *601*, 125–131.
10. Oren, Y., Tsabar, M., Cuoco, M.S., Amir-Zilberstein, L., Cabanos, H.F., Hütter, J.C., Hu, B., Thakore, P.I., Tabaka, M., Fulco, C.P., et al. (2021). Cycling cancer persister cells arise from lineages with distinct programs. *Nature* *596*, 576–582.
11. Goyal, Y., Dardani, I.P., Busch, G.T., Emert, B., Fingerma, D., Kaur, A., Jain, N., Mellis, I.A., Li, J., Kiani, K., et al. (2021). Pre-determined diversity in resistant fates emerges from homogenous cells after anti-cancer drug treatment. Preprint at bioRxiv. <https://doi.org/10.1101/2021.12.08.471833>.
12. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* *367*, eaaw3381.
13. Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Wayne, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature* *564*, 219–224.
14. Emert, B.L., Cote, C.J., Torre, E.A., Dardani, I.P., Jiang, C.L., Jain, N., Shaffer, S.M., and Raj, A. (2021). Variability within rare cell states enables multiple paths toward drug resistance. *Nat. Biotechnol.* *39*, 865–876.
15. Umkehrer, C., Holstein, F., Formenti, L., Jude, J., Froussios, K., Neumann, T., Cronin, S.M., Haas, L., Lipp, J.J., Burkard, T.R., et al. (2021). Isolating live cell clones from barcoded populations using CRISPRa-inducible reporters. *Nat. Biotechnol.* *39*, 174–178.
16. Weinreb, C., and Klein, A.M. (2020). Lineage reconstruction from clonal correlations. *Proc. Natl. Acad. Sci. USA* *117*, 17041–17048.
17. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.-W., Migueles, R.P., Janovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* *583*, 585–589.
18. Jiang, C.L., Goyal, Y., Jain, N., Wang, Q., Truitt, R.E., Coté, A.J., Emert, B., Mellis, I.A., Kiani, K., Yang, W., et al. (2022). Cell type determination for cardiac differentiation occurs soon after seeding of human-induced pluripotent stem cells. *Genome Biol.* *23*, 90.

19. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502.
20. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15.
21. Kennedy, A.B.W., and Sankey, H.R. (1898). The thermal efficiency of steam engines. *Minutes of the Proceedings of the Institution of Civil Engineers* **134**, 278–312.
22. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008.
23. Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498.
24. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205.
25. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44. <https://doi.org/10.1038/nbt.4314>.
26. Kessenbrock, K., Krumbholz, M., Schönemarker, U., Back, W., Gross, W.L., Werb, Z., Gröne, H.J., Brinkmann, V., and Jenne, D.E. (2009). Netting neutrophils in autoimmune small-vessel vasculitis. *Nat. Med.* **15**, 623–625.
27. Wang, S.-W., Herriges, M.J., Hurley, K., Kotton, D.N., and Klein, A.M. (2022). CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol.* **40**, 1066–1074.
28. Sacher, F., Feregrino, C., Tschopp, P., and Ewald, C.Y. (2021). Extracellular matrix gene expression signatures as cell type and cell state identifiers. *Matrix Biol.* **10**, 100069.
29. Hull, D.L. (2010). *Science as a Process* (University of Chicago Press).
30. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450.
31. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473.
32. Kalhor, R., Mali, P., and Church, G.M. (2017). Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200.
33. Frieda, K.L., Linton, J.M., Hormoz, S., Choi, J., Chow, K.-H.K., Singer, Z.S., Budde, M.W., Elowitz, M.B., and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111.
34. Chow, K.-H.K., Budde, M.W., Granados, A.A., Cabrera, M., Yoon, S., Cho, S., Huang, T.-H., Koulina, N., Frieda, K.L., Cai, L., et al. (2021). Imaging cell lineage with a synthetic digital recording system. *Science* **372**, eabb3099. <https://doi.org/10.1126/science.abb3099>.
35. Yu, H.-H., Chen, C.-H., Shi, L., Huang, Y., and Lee, T. (2009). Twin-spot MARCM to reveal the developmental origin and identity of neurons. *Nat. Neurosci.* **12**, 947–953.
36. Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971. <https://doi.org/10.1126/science.aax1971>.
37. Shaffer, S.M., Emert, B.L., Reyes Hueros, R.A., Cote, C., Harmange, G., Schaff, D.L., Sizemore, A.E., Gupte, R., Torre, E., Singh, A., et al. (2020). Memory sequencing reveals heritable single-cell gene expression programs associated with distinct cellular behaviors. *Cell* **182**, 947–959.e17.
38. Baglama, J., Reichel, L., and Lewis, B.W. (2021). irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices.
39. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCr: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941.
40. Salgado, J.F. (2018). Transforming the area under the normal curve (AUC) into Cohen's d, pearson's r pb , odds-ratio, and natural log odds-ratio: two conversion tables. *Eur. J. Psychol. Appl. Leg. Context* **10**, 35–47.
41. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.* **25**, 25–29.
42. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1802.03426>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
WM989 melanoma cells treated with vemurafenib, WM983b melanoma cells treated with vemurafenib, MDA breast cancer cells treated with Paclitaxel	Goyal et al., 2021 ¹¹	https://www.biorxiv.org/content/10.1101/2021.12.08.471833v1
Murine hematopoietic stem cells <i>in vitro</i> differentiation	Weinreb et al., 2020 ¹²	https://www.science.org/doi/10.1126/science.aaw3381
Human induced pluripotent stem cell line directed toward cardiomyocyte fate	Jiang et al., 2022 ¹⁸	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02654-6
Software and algorithms		
ClonoCluster (https://doi.org/10.5281/zenodo.3369197)	This paper	https://github.com/leeprichman/ClonoCluster (https://doi.org/10.5281/zenodo.3369197)
Seurat	Satija et al., 2015 ¹⁹	https://cloud.r-project.org/package=Seurat
WebGestaltR	Liao et al., 2019 ²⁴	https://cran.r-project.org/web/packages/WebGestaltR/index.html
<i>data.table</i> , <i>entropy</i> , <i>magrittr</i> , <i>ggplot2</i> , <i>devtools</i>	Comprehensive R Archive Network	https://cran.r-project.org/
Other		
Analysis scripts, count matrices, barcode matrices, and output (https://doi.org/10.5281/zenodo.7105982)	This paper	https://github.com/arjunrajlaboratory/ClonoCluster_paper (https://doi.org/10.5281/zenodo.7105982)

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Arjun Raj (arjunraj@seas.upenn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Single cell RNA sequencing and barcode data for six datasets were compiled from multiple sources. “Low dose BRAF inhibitor” and “high dose BRAF inhibitor” indicate replicates of a monoclonal human melanoma cell line, WM989 A6-G3,¹⁴ transfected with a barcode library and treated with 250 nM and 1 μM vemurafenib respectively. “WM983B BRAF inhibitor” samples indicates replicates of a vemurafenib-resistant monoclonal human melanoma cell line, WM983B E6-C6¹⁴ transfected with a barcode library and treated with 250 nM vemurafenib.¹¹ “Paclitaxel-treated breast cancer cells” indicates replicates of a monoclonal line, MDA-MB-231D4, derived from human breast cancer cell line transfected with a barcode library and treated with 1 nM paclitaxel, as previously described.^{11,37} “Cardio-directed iPSC” indicates a barcode-transfected induced pluripotent stem cell line treated to drive cells toward a cardiomyocyte fate as previously described.¹⁸ Count matrices, barcoding methods, and barcode assignments for these samples were generated as described.^{11,18} “*In vitro* hematopoiesis” samples indicate bar-coded murine hematopoietic stem cells differentiating *in vitro*, collected and sequenced at the day 2 time point when most cells were labeled “undifferentiated” using known cell markers. Count matrices and barcode assignments were deposited by Weinreb et al.¹² and retrieved from the NCBI Gene Expression Omnibus at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140802>. Count matrices and barcodes for all datasets used in this study are available in the analysis repository on GitHub at https://github.com/arjunrajlaboratory/ClonoCluster_paper (<https://doi.org/10.5281/zenodo.7105982>).
- Raw data and scripts for all analyses are available at https://github.com/arjunrajlaboratory/ClonoCluster_paper (<https://doi.org/10.5281/zenodo.7105982>). ClonoCluster is open source and available under the GPL3 license at <https://github.com/leeprichman/ClonoCluster> (<https://doi.org/10.5281/zenodo.3369197>), including worked examples, example data, unit tests, and a Docker image.
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

METHOD DETAILS

Transcriptome and clone barcode integration with ClonoCluster

To integrate the transcriptome and clone barcode information, the shared nearest neighbors network graph is produced, using the single cell RNA sequencing analysis R package *Seurat*.¹⁹ Briefly, gene counts are converted into 100 principal components using the *irlba* package.³⁸ k nearest neighbors in PCA space are identified and a network graph is constructed where cells are nodes and edges between any two cells are given weight equivalent to the jaccard index between shared nearest neighbors (defined below as J). For all analyses, k was set to 20 and the graph was pruned of any edges with weight less than 1/15, consistent with *Seurat* defaults.

A size-normalized clone barcode network graph is also constructed where each cell is represented as a node, with edges drawn between all cells with weights of 0 for cells that are not in the same clone cluster (i.e. do not share a barcode) or $1/n$ for cells that are in the same clone cluster (i.e. do share a barcode), where n is equal to the total number of cells assigned to that barcode (size of the clone cluster.) We found that size normalization is necessary to prevent the highly interconnected larger clone clusters from exerting a dominant effect on graph modularity and final clustering compared to smaller clone clusters.

The transcriptome and clone graph edge weights are integrated using the following equation:

$$W = \alpha^\beta (\text{lin} - J) + J$$

Where W is the final output edge weight between two cells, α is a chosen parameter ranging from 0 to 1, β is a chosen parameter from 0 to 1, lin is the value of the edge in the clone barcode graph $\{0, 1/n\}$ and J represents the edge weight in the transcriptome graph: the jaccard index of shared nearest neighbors in transcriptome space. At $\alpha = 0$, the edge weight between any two cells is therefore equivalent to the transcriptome graph (J), and at $\alpha = 1$ the edge weight is equivalent entirely to the clone barcode graph (lin). β modifies the effect of step sizes in α , lower values of β increase the effect of α on final edge weight W at low α . For all datasets and analyses, β was set to 0.1. The combined graph with modified edge weights is then passed to the Louvain community detection algorithm²² as implemented in *Seurat* to return cluster assignments.

Cluster assignments

Resolution for the community detection algorithm was fixed for each sample, ranging between 0.6 and 1 across all nine replicates, with an initial target of 8–20 clusters at $\alpha = 0$. As α approaches 1, the number of clusters returned at fixed resolution approaches n , the number of unique clone barcodes present. Therefore, the “high α ” level was identified by approximation of the α value at the inflection point where the number of clusters rapidly increases and begins to reflect single clone clusters. The “low α ” level was defined as half the “high α ” value rounded to the nearest tenth. This gave the maximum number of clusters as 32 at the high α level in the Jiang et al.¹⁸ dataset.

Clone barcode to cluster correlation analysis and visualization

Sankey diagrams were generated using the *ggalluvial* R package. For Sankey diagrams inclusive of $\alpha = 1$, the top 15 largest clone clusters per sample are chosen for ease of visualization and limitation of discrete colors used. Unless otherwise specified, the complete dataset is shown for Sankey diagrams. Cohen’s κ was computed at each α level per clone barcode for the hybrid cluster with the largest proportion of the barcoded cells within it.

Cluster marker identification

Cluster markers were identified by receiver operating characteristic (ROC) using the *ROCR* package.³⁹ All genes present in the count matrix were tested for classification of a cell within a cluster versus all other clusters combined. When cells are noted as “marker positive”, the threshold for positivity was identified from the point on the ROC curve with minimum Euclidean distance from 100% TPR and 0% FPR.

Reorganization marker and overrepresentation analysis

For each cluster of interest at a given α value, cells from each contributing cluster at the transcriptome level ($\alpha = 0$) were compared to all other cells from the contributing transcriptome cluster. ROC analysis was performed to determine the strength of genes in the count matrix to classify cells from the contributing transcriptome cluster that reorganize to the α cluster of interest as long as the two comparison groups contained at least 10 cells. The AUC from this analysis was termed the reorganization AUC or reorg-AUC. Reorganization marker genes were defined as genes with reorg-AUC > 80 (corresponding to a “large” effect size.)⁴⁰ This stringent threshold was chosen to eliminate potentially spurious markers and more accurately reflect the true biology, maximizing the usefulness of downstream gene set level analyses. Each set of reorganization markers for the paired hybrid α cluster and contributing transcriptome ($\alpha = 0$) cluster was used for overrepresentation analysis. Overrepresentation analysis is a gene set analysis approach that identifies whether a given list of unranked differentially expressed genes is enriched for members of known gene sets representing biological processes or pathways, such as those from the Gene Ontology Consortium,⁴¹ compared to selection at random. When a transcriptome cluster contributed in its entirety to an α cluster (i.e. the transcriptome cluster was a subset of the α cluster), the AUC for overall cluster markers for the contributing transcriptome cluster was considered the reorg-AUC. Thresholds for marker positivity for visualization were determined from the point on the ROC curve with minimum Euclidean distance from 100% TPR and 0% FPR.

Overrepresentation analysis was performed with the R package *WebGestaltR*.²⁴ The search space included the Gene Ontology Cellular Components and Molecular Function datasets.⁴¹ Significantly overrepresented gene sets were defined as those with FDR < 0.05. For visualization, heatmaps were created of the maximum log base 2 of the enrichment ratio for any cluster reorganization in the sample.

Modified principal components and UMAP visualization

Uni-form manifold approximation and projection (UMAP) is a dimensionality reduction technique frequently used to visualize gross trends in high dimensional single cell RNA sequencing data.^{25,42} Clustering assignments and associated markers computed from principal component space are usually overlaid as well.⁶ To incorporate clone barcode information into the two dimensional UMAP representation of data, we modified the principal components using a tunable parameter s according to the following equation:

$$M_s = M - \frac{s}{10} * (M - M_b)$$

Where M_s is the modified PCA matrix passed to the UMAP algorithm, M is the PCA matrix, s , Warp Factor, is a parameter from 0 to 10, and M_b is a matrix composed of the average value of the PCA matrix per clone. At $s = 0$, the input to the UMAP will be the PCA matrix, M . At $s = 10$, the input to the UMAP will be a matrix with the value of each principal component substituted for the corresponding mean value per clone barcode, M_b , thus the only variation reflected in the UMAP will be between clones (Figure S5B). The parameters for UMAP as implemented in the *uwot* package were identical to the default settings of *Seurat*.^{19,42}

QUANTIFICATION AND STATISTICAL ANALYSIS

Data manipulation and statistical analysis was performed in the R statistical computing language with the *data.table*, *stats*, *ROCR*, *WebGestaltR*, and *magrittr* packages. Entropy was calculated with the *entropy* package using the Laplace method. Transcriptome nearest neighbor graph construction and community detection was performed with the *Seurat* package. Data visualization used the R packages *ggplot2*, *pheatmap*, *ggalluvial*, and *VennDiagram*. Global hypothesis testing was performed using the non-parametric Kruskal-Wallis test with Bonferroni correction for multiple comparisons. Local testing was performed with pairwise Wilcoxon tests with Bonferroni correction.

Cell Genomics, Volume 3

Supplemental information

**ClonoCluster: A method for using clonal
origin to inform transcriptome clustering**

Lee P. Richman, Yogesh Goyal, Connie L. Jiang, and Arjun Raj

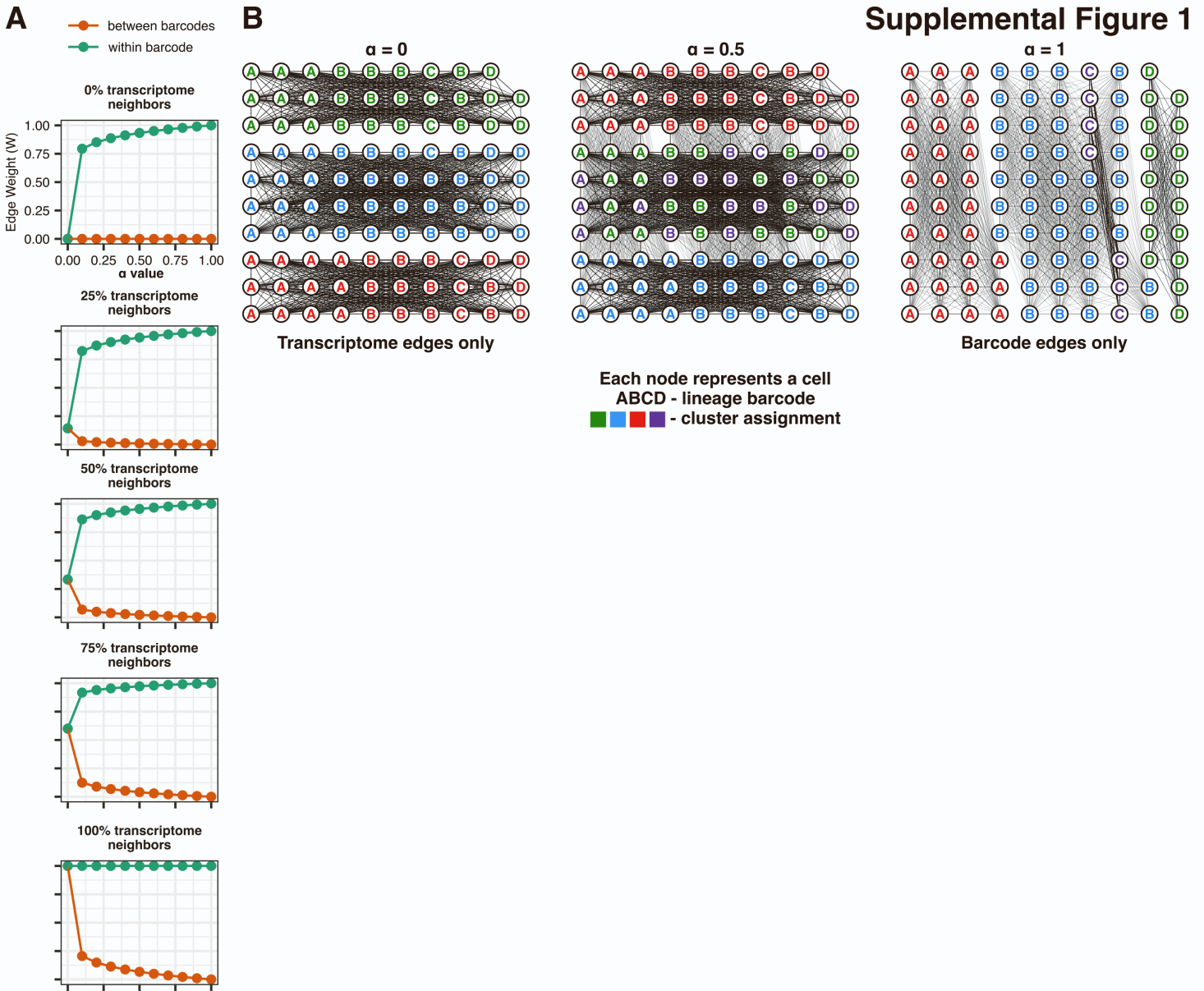


Figure S1, Related to Figure 1: α modifies network graph edge weights to create Louvain clusters that increasingly reflect clone barcode assignments.

A. Output edge weight of the ClonoCluster model for graph edge weights between two cells for different proportions of shared nearest neighbors in transcriptome space with fixed $\beta = 0.1$.

B. Network graphs with modified edge weights and cluster assignments at three α values for simulated data.

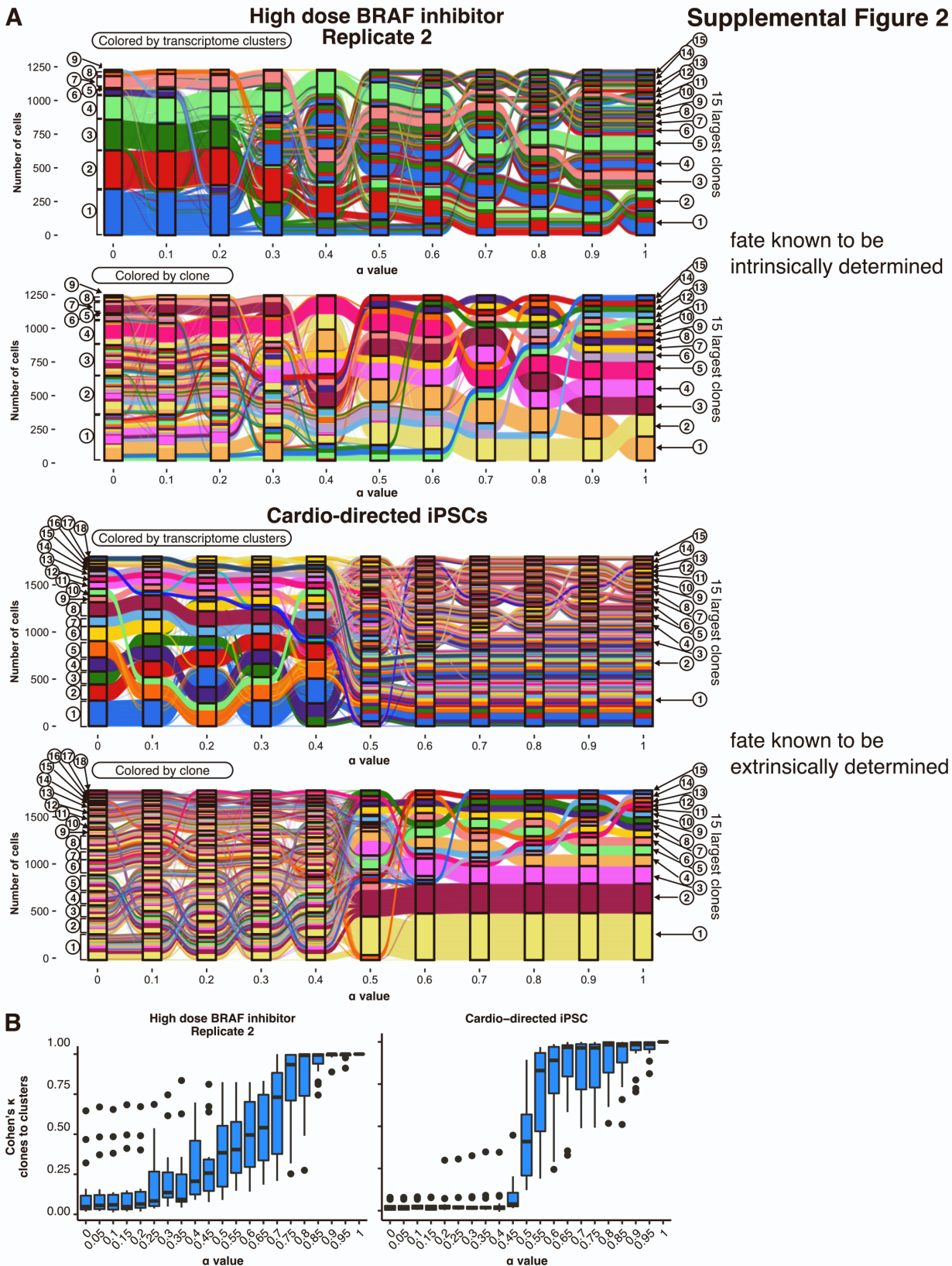


Figure S2, Related to Figure 1: Intrinsic vs extrinsic determinants of cell fate stratify the effect of α on clone-cluster congruence.

A. Sankey diagrams for the 15 largest clone clusters in two samples across α values, a clonal melanoma cell line treated with high dose ($1\mu\text{M}$) vemurafenib¹¹ with known intrinsic determinants of cell fate and directed differentiation of induced pluripotent stem cells (iPSCs) towards a cardiomyocyte fate¹⁸ with previously described dominance of extrinsic determinants of cell fate.

B. Cohen's κ for interrater reliability for classification of a cell by clone barcode and cluster assignment for the top 15 largest clone clusters across α values for the samples in **(A)**.

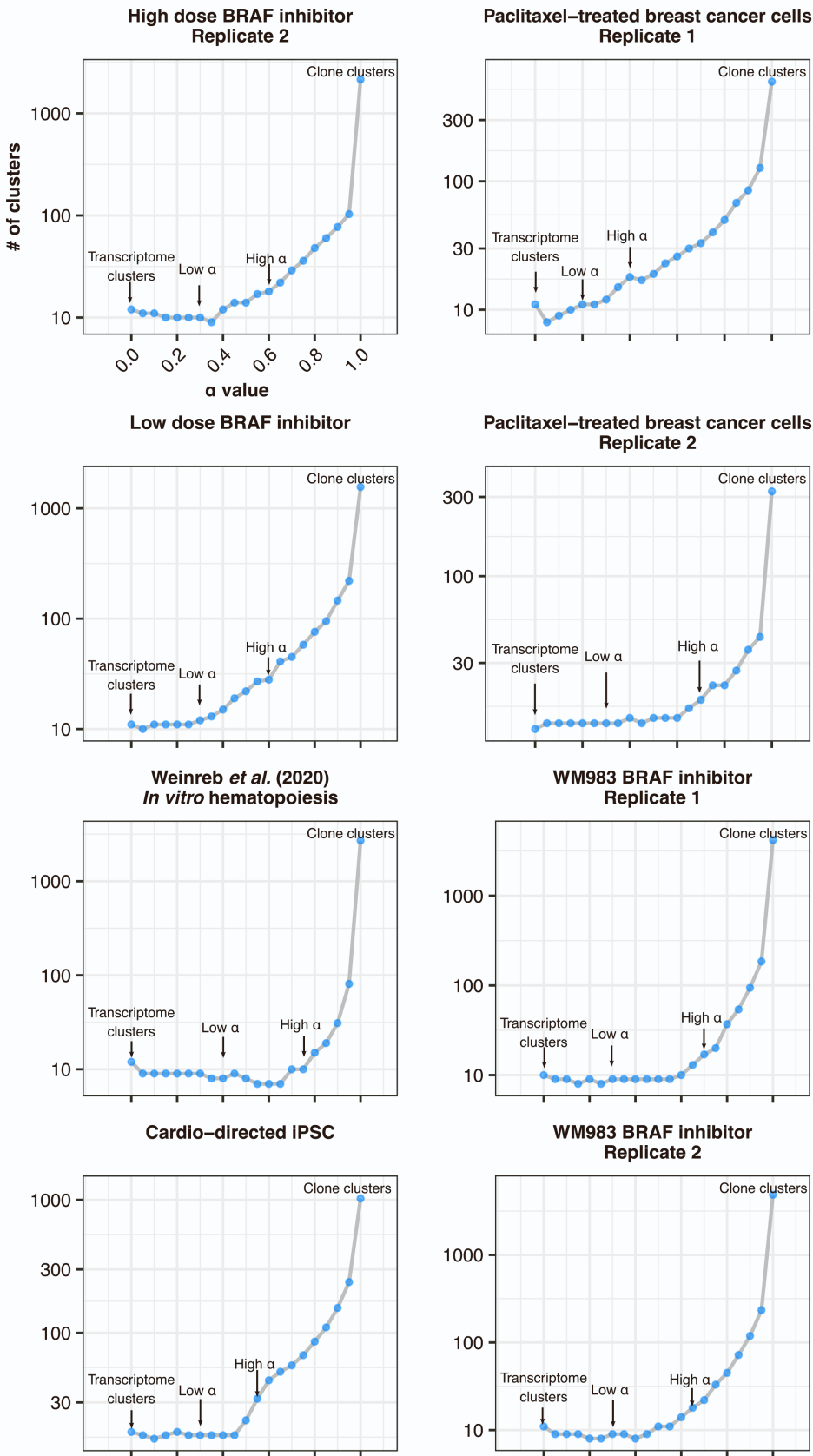


Figure S3, Related to Figure 1: The number of clusters returned by community detection at fixed resolution approaches the number of unique clone barcodes in the data as α approaches 1. Plots showing the number of clusters returned by community detection at constant resolution with increasing α in each dataset. At $\alpha = 1$, the network graph is composed of like-barcoded cells (a clone cluster) connected only to each other and community detection returns the number of unique clone clusters in the data as the number of clusters.

Supplemental Figure 4

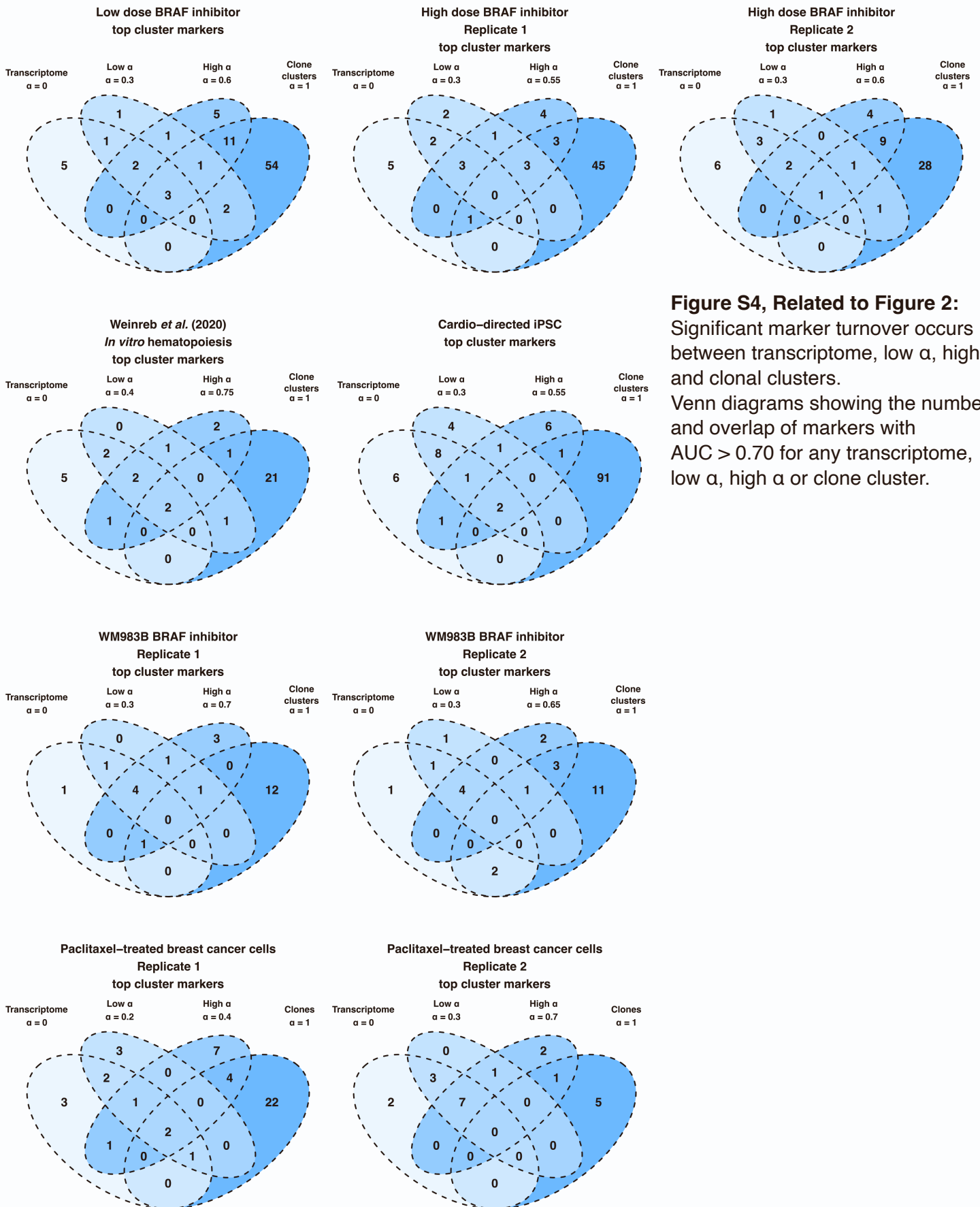
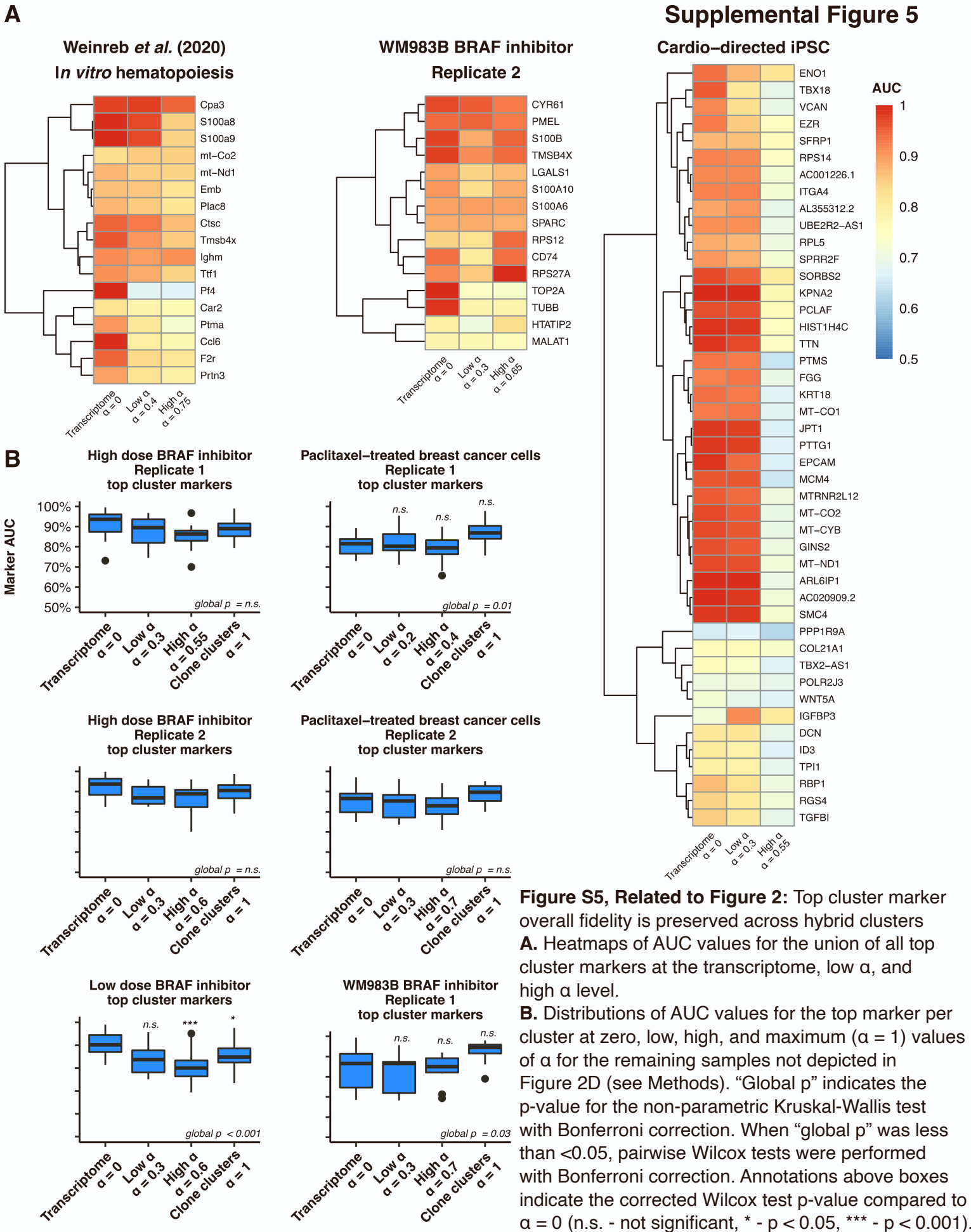


Figure S4, Related to Figure 2: Significant marker turnover occurs between transcriptome, low α , high α , and clonal clusters. Venn diagrams showing the number and overlap of markers with AUC > 0.70 for any transcriptome, low α , high α or clone cluster.

Supplemental Figure 5



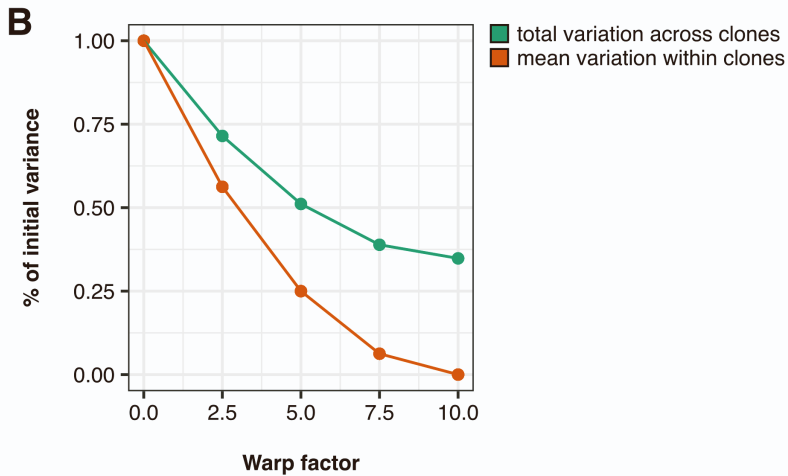
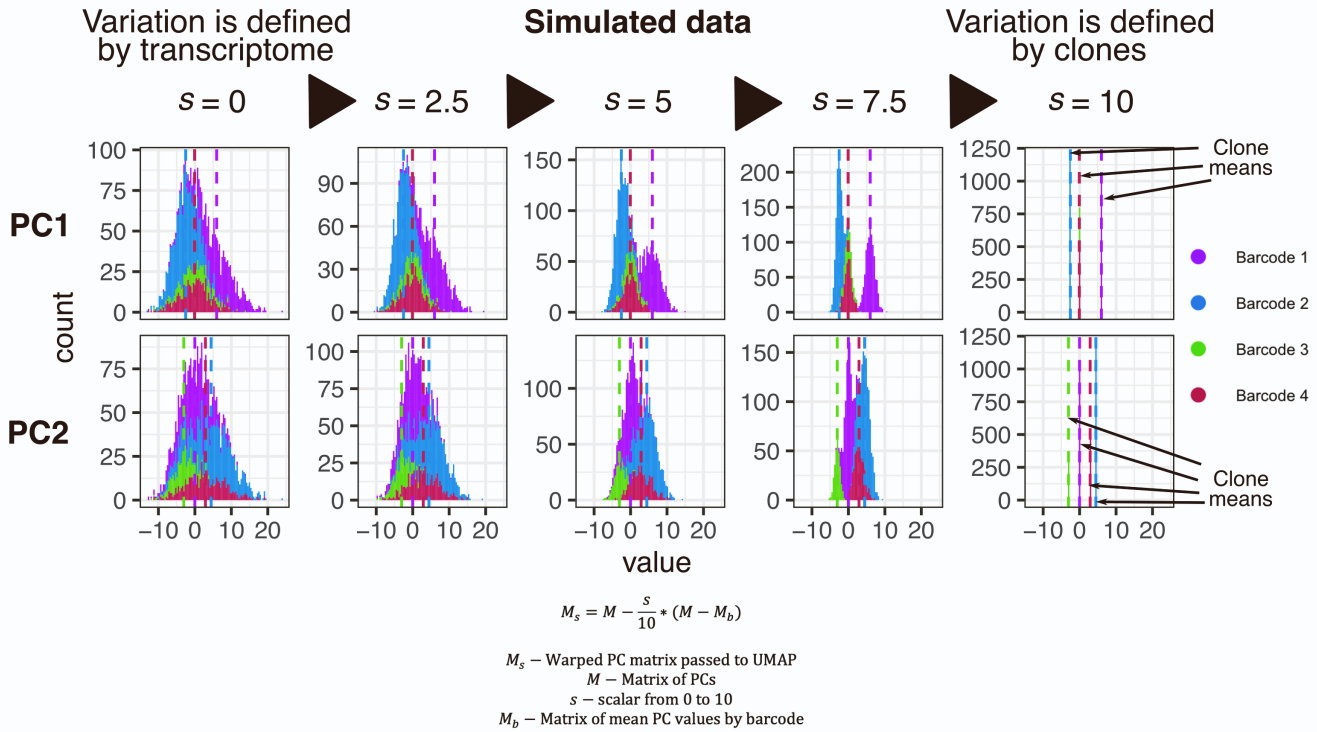


Figure S6, Related to Figure 4: Increasing Warp Factor, s , reduces the variation of principal component values around clone barcode means.

A. Histograms colored by clone barcodes demonstrating the distribution of PC values for two principal components for simulated data for 3,000 cells with four clone barcode assignments. Vertical dashed lines represent the mean values for each clone barcode at $s = 0$, equivalent to the unmodified PC means. At $s = 10$, modified PC values are equivalent to the assigned clone barcode means. Modified PC values are then passed to the UMAP implementation for visualization as in Figure 4.

B. Mean variation of all PCs in simulated data within and between clone barcodes stratified by values of s .

Supplemental Figure 7

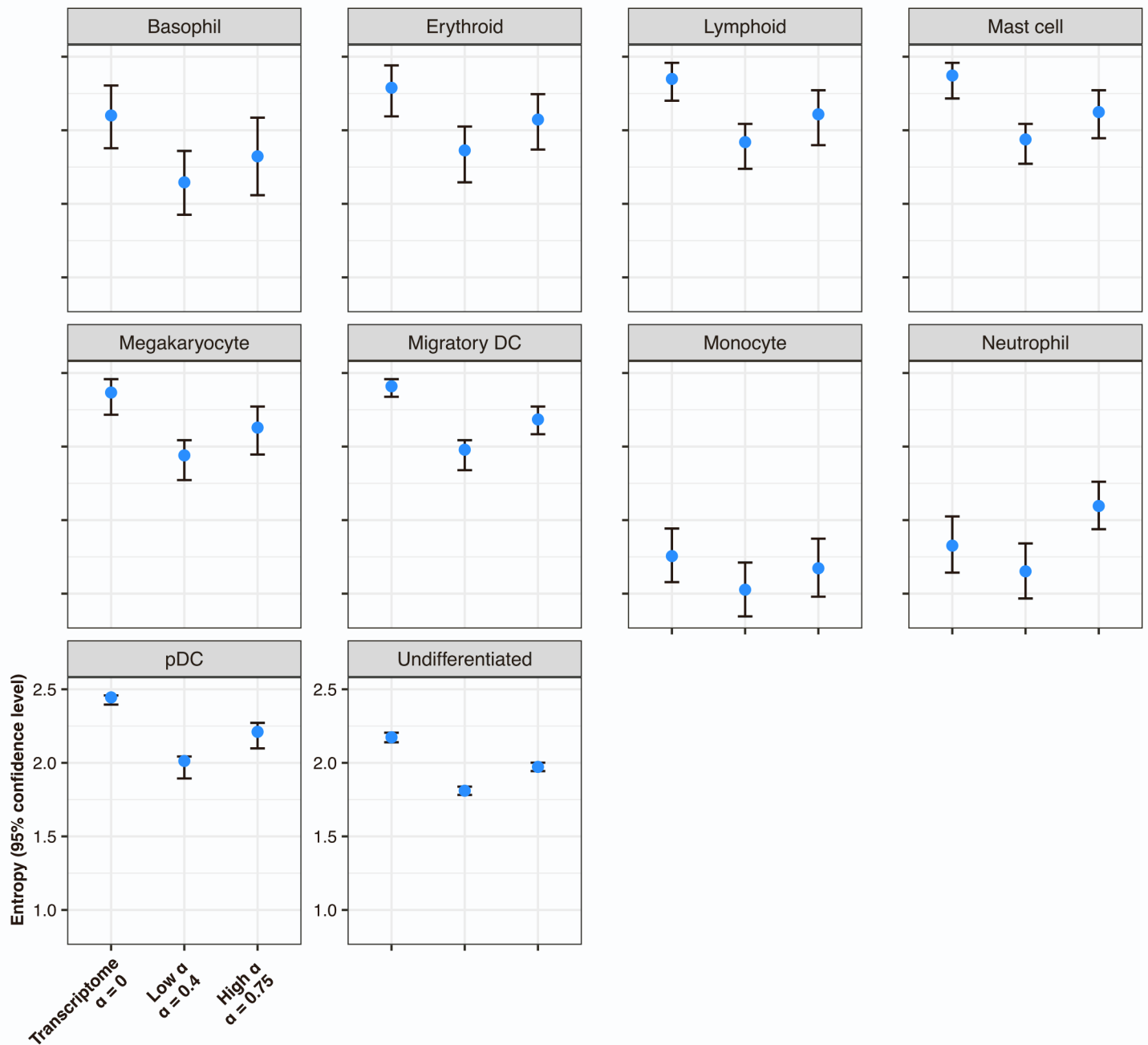


Figure S7, Related to Figure 5: Low α clusters reduce entropy of clustering for multiple hematopoietic cell types. Laplace entropy was calculated for the distribution of known cell types in hematopoiesis data¹² across transcriptome-only clusters, low α , and high α . Entropy is shown in bits, error bars represent 95% confidence intervals using the Monte Carlo method with 1,000 samplings of 33% of the cells.