MaCroDNA: Accurate integration of single-cell DNA and RNA data for a deeper understanding of tumor heterogeneity

Mohammadamin Edrisi[1,*] Xiru Huang[1] Huw A. Ogilvie[1,*]
Luay Nakhleh[1,*]
[1]Department of Computer Science, Rice University, Houston, Texas, USA.

Abstract

Cancers develop and progress as mutations accumulate, and with the advent of single-cell DNA and RNA sequencing, researchers can observe these mutations, their transcriptomic effects, and predict proteomic changes with remarkable temporal and spatial precision. However, to connect genomic mutations with their transcriptomic and proteomic consequences, cells with either only DNA data or only RNA data must be mapped to a common domain. For this purpose, we present MaCroDNA, a novel method which uses maximum weighted bipartite matching of per-gene read counts from single-cell DNA and RNA-seq data. Using ground truth information from colorectal cancer data, we demonstrate the overwhelming advantage of MaCroDNA over existing methods in accuracy and speed. Exemplifying the utility of single-cell data integration in cancer research, we propose, based on results derived using MaCroDNA, that genomic mutations of large effect size increasingly contribute to differential expression between cells as Barrett's esophagus progresses to esophageal cancer.

Over the last decade, single-cell sequencing (SCS) has grown dramatically. SCS can now provide data with much higher throughput and quality with a diverse range of protocols that can assay single cells' genomic, transcriptomic, epigenomic, and proteomic contents [1, 2, 3, 4, 5, 6]. These technologies have changed our understanding in different fields of biology, including developmental biology, immunology, microbiology, and cancer biology.

*Corresponding authors: {edrisi,huw.a.ogilvie,nakhleh}@rice.edu

SCS technologies can shed light on our understanding of the genotype-phenotype relationship. In particular, single-cell multi-omic assays sequencing RNA and DNA from the same single cells are of great potential to unravel the underlying mechanisms by which genomic alterations, including single-nucleotide variations, copy number aberrations (CNAs), and other structural variations, might impact transcriptomic programs at the single-cell level. This would lead to further understanding of clonal development in healthy and diseased tissues, especially in cancer therapies where drug responses are affected in part by phenotypes induced by genomic mutations [7].

Despite being ideal for such studies, SCS technologies that measure both DNA and RNA from the same cells, such as G&T-seq [8], DR-seq [9], and scTrio-seq [10] are of low throughput and face scalability issues. On the other hand, high-throughput sequencing technologies such as 10X genomics single-cell RNA sequencing (scRNA-seq) [11] or Direct Library Preparation (DLP) [12] can measure either genomic or transcriptomic content of single cells imposing the computational challenge of associating cells across different modalities.

In this work, we are interested in tackling the cell association problem in high-throughput SCS data where for each single cell, either CNA measurements or gene expression values are provided. There is a plethora of methods developed for the integrative analysis on SCS multi-omic data which mostly integrate the single-cell chromatin accessibility—scATAC-seq—or DNA methylation measurements with the gene expression data from scRNA-seq such as [13, 14, 15, 16, 17, 18, 19, 20], to name a few. However, there are, yet, a handful of methods that specifically address the cell association problem between scDNA-seq and scRNA-seq data such as clonealign [21], Cardelino [22], and CCNMF [23] among which clonealign and CCNMF are developed for the integration of copy number and gene expression data. In the following, we review clonealign, CCNMF, and Seurat [19]. Although Seurat was introduced for the data integration between different scRNA-seq assays or between scRNA-seq and scATAC-seq data sets, we decided to incorporate it into this study as the exemplar of the commonly-used methods in SCS multi-omic analyses.

clonealign [21] is a statistical method aiming to assign single cells whose gene expression values are provided by scRNA-seq to cancer clones inferred from low-coverage single-cell CNA data. Here, the clones are the *clades* obtained by defining a cut-off on a phylogenetic tree reconstructed from

CNA data. clonealign's dosage compensation function is predicated on the assumption that an increase (decrease) in the copy number value of a gene would yield a higher (lower) gene expression value at that gene. This assumption was established based on the previously observed relationship between the genes' expressions and their corresponding copy number values in bulk and single-cell assays [24, 9, 8, 25]. Given the clone-specific copy number profiles, clonealign predicts the assignment of the cells measured using scRNA-seq to cancer clones measured using scDNA-seq by approximating the posterior probability distribution of the data given the method parameters using variational inference [26].

Seurat's integration method [19] is a manifold alignment algorithm to solve the integration of multiple modalities. In manifold alignment methods, it is assumed that the *data points*—cells in this context—from the two modalities share a low-dimensional space where each cell can be represented by a low-dimensional vector—interpreted as the *underlying biological state* of the cell. The cells from the two domains whose vector representations are either identical or adjacent to each other in the manifold are identified as the associated cells. To establish such shared manifold between two different modalities, Seurat adopts canonical correlation analysis (see, e.g., [27]). This dimensionality reduction step is followed by the identification of mutual nearest neighbors—or *"anchors"*—across the two data sets. Anchors are the pairs of single cells, each from a different data set, that neighbor each other in the shared low-dimensional space. The anchors are filtered, scored, and finally utilized for batch correction and further downstream analyses.

Both clonealign and Seurat are map-to-reference inference methods, meaning that the data from one domain is mapped to the clones inferred from another domain. The reference may differ in such methods. For example, in clonealign, the copy number profile of each clone is the reference, while in Seurat, the scRNA-seq data on the latent manifold is the reference. Since the choice of the reference in these methods might introduce *"systematic bias to the method"* [23], a reference-free method, CCNMF [23] was put forward. CCNMF applies non-negative matrix factorization technique (see [28, 29] for details about this technique) to the scRNA-seq and scDNA-seq data. Specifically, for each of the two SCS data sets, CCNMF infers a pair of matrices including the cluster (clonal) centroids in a low-dimensional space as well as the *soft* clonal assignment of the single cells

to their clones. The low-dimensional representations of the DNA and RNA clones are coupled by incorporating a prior information about the relationship between CNAs and gene expression values, such as the information acquired by the linear-regression modeling of CNAs and gene expression values measured from the paired RNA and DNA bulk sequencing data. Performing this coupling technique on the cluster centroids of the two data sets yields a co-clustering on both data sets. CCNMF's unique co-clustering technique differentiates it from the other existing methods such as clonealign and Seurat; as CCNMF can infer clones by itself, it does not accept the predefined clones as input.

Here we report on a new method for the inference of cell association in SCS multi-omic data with CNA and gene expression measurements. Inspired by clonealign's underlying assumption that the gene expression value of a gene is proportional to the corresponding CNA value at that gene, we hypothesized that the Pearson correlation coefficients between the gene expression measurements of scRNA-seq cells and the CNA profiles of scDNA-seq cells could be an effective metric for the cell association inference. Moreover, we realized that a strict one-to-one mapping between scRNA-seq cells and scDNA-seq cells—or as close to one-to-one as possible given unequal numbers of cells—provides an unprecedented accuracy by implicitly respecting the assumption that the underlying clonal prevalences in the two modalities are almost identical.

We applied our method, MaCroDNA (Mapping Cross Domain Nucleic Acid) to a scTrio-seq2 data set [30]. We exploited the partially-known ground truth information in this data set for our evaluation purposes. It is noteworthy that our study is the first to evaluate the performance of the methods for integration of independent single-cell gene expression and copy number data directly on empirical data. By taking advantage of the available ground-truth information and using a variety of clustering techniques, we rigorously measured the accuracy of clonealign, Seurat, and MaCroDNA in assigning the scRNA-seq cells to the scDNA-seq clones as well as predicting the clonal prevalences of the scRNA-seq cells. Since CCNMF did not fit into our evaluation scheme (due to not accepting the predefined clones as its input), we decided to assess the performance of CCNMF separately. Even on this basis its performance was poor, therefore we excluded its results from our main analysis and included them in the supplementary materials.

To demonstrate the insight that may be gained from integrated DNA and RNA single-cell data, we applied MaCroDNA to a previously published data set from a study of Barrett's esophagus (BE), where biopsies were taken from multiple patients and individual cells from those biopsies sequenced for either DNA or RNA [31]. BE is a metaplasia, presumed to be caused by gastroesophageal reflux disease, where squamous epithelium of the esophagus is replaced by intestinal epithelium [32]. Non-dysplastic BE (NDBE) may progress to low-grade dysplasia (LGD), high-grade dysplasia (HDG), or esophageal adenocarcinoma (EAC), although the exact relationship of this progression in terms of cellular descent remains uncharacterized [33]. While a diagnosis of EAC is devastating with dismal prospects for survival, most BE does not progress to EAC and predicting when EAC will develop is very difficult [34]. To make useful predictions and improve patient outcomes, we must first understand why progenitor cells evolve into EAC or dysplastic BE instead of NDBE. Our MaCroDNA-based results suggest that in comparison to NDBE and LGD, HGD and EAC are associated with genomic mutations of large effect size which are heterogeneously present within individual biopsies.

1 Results

1.1 Overview of MaCroDNA

The inputs of MaCroDNA include the scRNA-seq gene expression and the scDNA-seq copy number values. These measurements are provided to the method in the form of counts matrices of the single cells sequenced from the same tissue (see Fig. 1a and also Section MaCroDNA). Given these inputs, for each cell in the scRNA-seq data, MaCroDNA identifies the *best* corresponding cell in the scDNA-seq data. Here, our main criterion for the best correspondence is the Pearson correlation coefficient between the paired cells. More specifically, MaCroDNA seeks an assignment that maximizes the sum of the Pearson correlation coefficients between the paired cells (see Section MaCroDNA for $N_{\mathbf{G}} \leq N_{\mathbf{C}}$).

Supposing the single cells are sampled from the same tissue, we assume that they share the same underlying clones, and therefore, the clonal prevalences in the two data sets are nearly identical.

To respect such an assumption, MaCroDNA avoids assigning any two scRNA-seq cells to the same scDNA-seq cell *as much as possible*. To elaborate more on how MaCroDNA accounts for this constraint, let us consider the two following cases:

- When the number of the scRNA-seq cells is smaller than the number of the scDNA-seq cells, MaCroDNA identifies a unique scDNA-seq cell for each scRNA-seq cell by viewing the correspondence problem as an instance of *maximum weighted bipartite matching* problem. This is a well-known problem that is polynomially solvable [35] (see Section MaCroDNA for $N_{\mathbf{G}} \leq N_{\mathbf{C}}$ for more details).

- When the number of the scRNA-seq cells is larger than the number of the scDNA-seq cells, some of the scRNA-seq cells would inevitably be matched to the same scDNA-seq cells. For this case, we devised a heuristic algorithm that breaks down the problem into multiple maximum weighted bipartite matching problems and solves them, one by one, in a sequential manner (Figs.1b,c demonstrate an example of this case. Also see Section MaCroDNA for $N_{\mathbf{G}} > N_{\mathbf{C}}$).

In both cases, MaCroDNA returns as output a binary correspondence matrix whose entries indicate whether two cells (one from scRNA-seq and the other from scDNA-seq data) are matched or not.

Using the above correspondence matrix, we assign the scRNA-seq cells to the clones identified from scDNA-seq data. First, we perform a clustering technique of choice on the scDNA-seq cells to obtain the clonal assignments (Fig. 1d). Given the clonal assignment of scDNA-seq cells, the clonal assignment of a scRNA-seq cell will be that of its corresponding scDNA-seq cell (Fig. 1e). In the following, we will describe our scheme for evaluating the performance of MaCroDNA, Seurat, and clonealign when applying them to an empirical data set.

1.2 Evaluating MaCroDNA on a colorectal cancer scTrio-seq data set

To assess the performance of MaCroDNA and the other methods, we applied them to a colorectal cancer (CRC) data set originally studied by Bian *et al.* [30]. The CRC data set consists of the scRNA-seq and scDNA-seq data of several patients. We decided to use three of these patients,

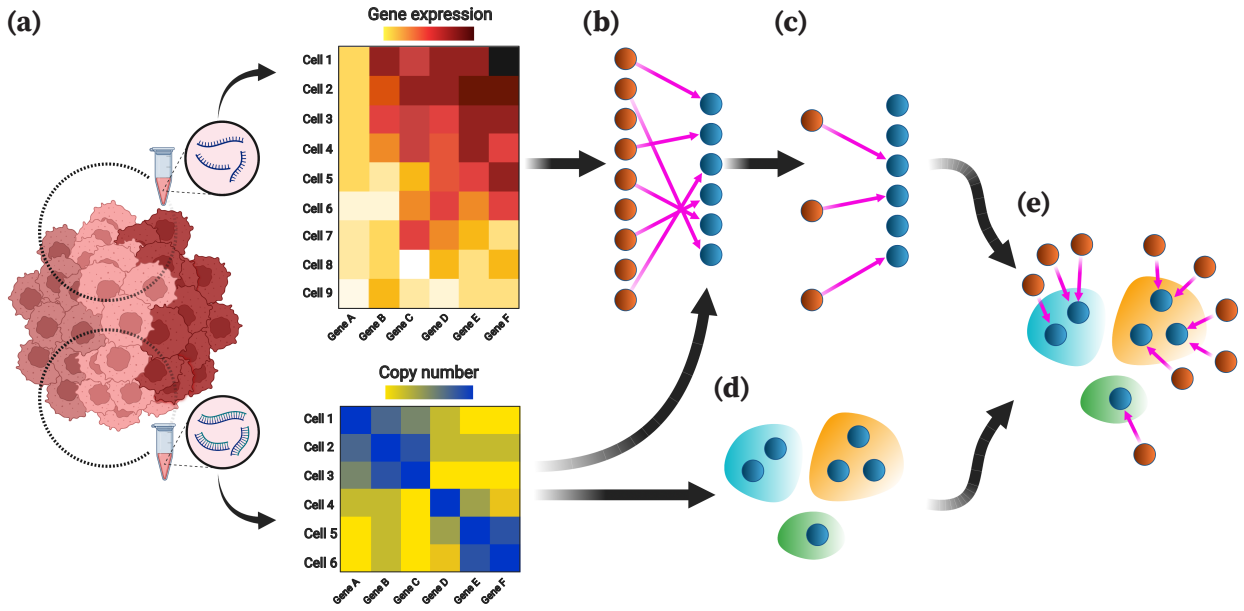

Figure 1: **Overview of MaCroDNA. (a)** The input of MaCroDNA consists of the scRNA-seq gene expression and the scDNA-seq copy number values that are supposedly obtained from the same tissue. **(b)** Given the gene expression and copy number matrices, MaCroDNA identifies the assignment of the scRNA-seq cells with gene expression values to the scDNA-seq cells with copy number values. When the number of the scRNA-seq cells is higher than the number of the scDNA-seq cells (as in the above example), MaCroDNA infers the correspondence between the cells by solving a series of maximum weighted bipartite matching problems (in this example, two steps are required). In the first step, only six scRNA-seq cells (equal to the number of the scDNA-seq cells) are assigned to the scDNA-seq cells such that no two scRNA-seq cells are paired with the same scDNA-seq cell, and the sum of the Pearson correlation coefficients between the pairs is maximized. **(c)** The scRNA-seq cells whose correspondences were identified in the last step are removed. Next, the remaining three scRNA-seq cells are assigned to the best scDNA-seq cells in a one-to-one fashion according to the same correlation-based criterion. By the end of these two consecutive steps, each scRNA-seq cell is associated to a scDNA-seq cell. **(d)** The clones are inferred from the scDNA-seq data using an algorithm of choice such as hierarchical clustering. **(e)** Given the cell-to-cell correspondences inferred by MaCroDNA along with the clonal assignment of the scDNA-seq cells, we can assign the scRNA-seq cells to the scDNA-seq clones. The clonal assignment of a scRNA-seq cell is the clonal assignment of its corresponding scDNA-seq cell. This figure was created with BioRender.com.

namely CRC04, CRC10, and CRC11 [30]. The remaining patients were excluded from our analysis because their transcriptomic data were obtained using a different library preparation protocol or their cell amount was insufficient (see Description of CRC data set).

The CRC data set is particularly helpful for our evaluation purposes as it contains some cells for which both the RNA and DNA measurements are provided. These cells can readily be used in our experiments as the ground truth data. In all our experiments, we exploited this ground truth information for evaluating the performance of the methods (see below). Of note, all the methods (including MaCroDNA) were oblivious to this information at the time of inference, and we used it only for the assessment. To our knowledge, this is the first time that the accuracy of methods for integrating DNA and RNA data have been measured using an empirical data set for which a degree of ground truth is available. While simulation-based analyses are widely used and useful, they cannot match the realism of this approach, and results from empirical data sets without any ground truth can only be described rather than measured for accuracy.

Table 1 provides the information on the number of cells that are provided with RNA data, DNA data, and both RNA and DNA data for each patient.

Table 1: Cell counts from colorectal cancer data set

Patient ID	RNA[1]	DNA[1]	Both
CRC04	93	93	57
CRC10	85	123	69
CRC11	192	249	174

[1] Inclusive of cells with both RNA and DNA

In the following, we will describe the experiments we conducted on MaCroDNA, Seurat, and clonealign. Through these experiments, we sought to answer the two following questions:

1. **Cell-to-clone assignment accuracy**: Given a clonal assignment identified on the scDNA-seq cells using a clustering algorithm, what is the accuracy of a method in assigning the scRNA-seq cells to the scDNA-seq clones? We measured the accuracy only for the scRNA-seq cells whose joint DNA measurement is also provided. For these scRNA-seq cells we know their true corresponding scDNA-seq cell. Following this idea, the assignment of such scRNA-seq cell to a scDNA-seq cell is considered correct if the predicted scDNA-seq cell is the true

corresponding scDNA-seq cell or if both the predicted and true scDNA-seq cells belong to the same clone according to the results of the clustering algorithm.

2. **Predictive accuracy for clonal prevalences**: According to the ground truth information and the clonal assignment of the scDNA-seq cells (again, provided by a clustering algorithm of choice), one can count the number of the scRNA-seq cells that are assigned to each of the scDNA-seq clones. While such counts can serve as the *true* clonal prevalences in the scRNA-seq data, we can obtain the *predicted* clonal prevalences by counting the number of the scRNA-seq cells that are assigned to the scDNA-seq clones by a method. In particular, we are interested in understanding the correlation between the true and predicted clonal prevalences for each clone.

It is to be noted that answering the above questions reveals two distinct characteristics of a method's performance. A method could be very inaccurate in mapping the scRNA-seq cells to the scDNA-seq clones while being *"faithful"* to the proportion of clones in the scRNA-seq data or vice versa.

We first inferred the scDNA-seq clones to measure both the cell-to-clone assignment accuracy and the predictive accuracy for clonal prevalences. Here, we chose two different clustering techniques, namely intNMF [36] and agglomerative clustering. Testing the integration methods under different clustering results leads to a better understanding of the robustness of the methods' performance to the choice of clustering algorithm.

In addition to the clustering algorithm, data transformation is another contributing factor in clonal inference. The original copy number values in the scDNA-seq data sets are in the form of non-negative integers. Along with the original copy number values, we used the log-transformed copy number values for the clonal inference as well. Log transformation is a commonly-used transformation for the genomic data [37, 38, 39, 40, 41]. We transformed the data using $\log(x + 1)$, where x is the original copy number value, and $\log(\cdot)$ denotes the natural logarithm.

Given the original and log-transformed data, we performed intNMF and agglomerative clustering techniques on each patient's scDNA-seq data separately. intNMF's clustering algorithm provides the optimal number of clusters along with the clonal assignments for the single cells (see Clonal inference by intNMF). intNMF identified two clusters in CRC04, three clusters in CRC10,

and two clusters in CRC11 using the original copy number values as input. Here, in both patients, CRC10 and CRC11, the largest clone contained more than half of the patient's cells, while CRC04 had relatively balanced clusters (see Figure S16 in the Supplementary materials). Using the log-transformed data, intNMF inferred two clusters in CRC04, three clusters in CRC10, and three clusters in CRC11. Again, while CRC04 had two balanced clones, CRC10 and CRC11 had one or two dominant clones.

Unlike intNMF, agglomerative clustering does not infer the optimal number of clones. Instead, it takes a user-specified value for the number of clusters as input. We ran agglomerative clustering to identify four clusters in each patient independently. This resulted in 12 clusters in total across all patients. Next, we merged the clusters one by one until reaching two clusters in each patient (see Clonal inference by agglomerative clustering for more details on the merging procedure). By doing so, we obtained the clonal assignments at different resolutions, ranging from six to 12 clusters in total.

After inferring the clusters under different configurations of the clustering method and data transformation, we assessed the performance of MaCroDNA, clonealign, and Seurat for the cell-to-clone assignment and predicting the clonal prevalences. To better understand the complexity of the problem and assess the methods' performance, we devised a random baseline that assigns an scRNA-seq cell to a scDNA-seq cell randomly sampled (with replacement) from the pool of scDNA-seq cells in the same patient (see Random baseline for the details).

Prior to running each method for integrating the scDNA-seq and scRNA-seq data, we pre-processed the CRC patients' data according to the requirements of each method as well as the preprocessing strategies that we considered (for the details on the specific preprocessing strategy applied to each method, see Preprocessing on CRC data set). Since clonealign requires the clone-specific copy number profiles as input, we provided the mean of all copy number profiles in a cluster as the clone-specific profile to clonealign. We ran Seurat and clonealign with their default parameters. Note that MaCroDNA does not require any user-specified parameters as the input.

The accuracy results of the methods (including the baseline) for the cell-to-clone assignment are shown in Fig. 2. In the panels for agglomerative clustering (Fig. 2a,b), we have demonstrated

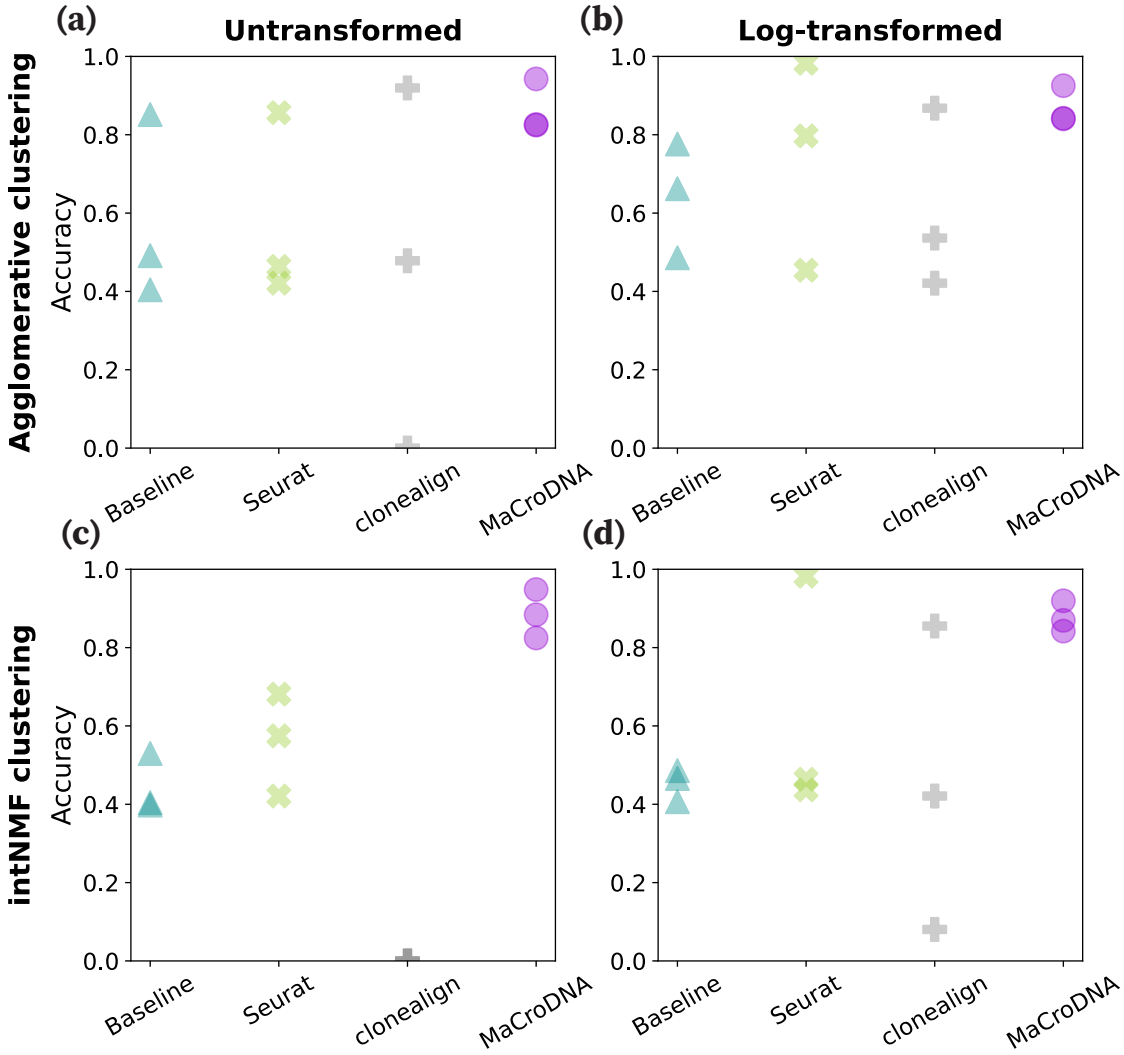

Figure 2: **Accuracy of the methods for the cell-to-clone assignment.** The accuracy of a method was calculated as the percentage of correct predictions (regardless of which clone an scRNA-seq cell was assigned to among all predictions made by the method in each patient's data. Each panel displays the methods' accuracy across the three patients when the clustering algorithm and the input data to the clustering algorithm were **(a)** the agglomerative clustering, and the original data, **(b)** the agglomerative clustering and the log-transformed data, **(c)** the intNMF clustering and the original data, and **(d)** intNMF and the log-transformed data.

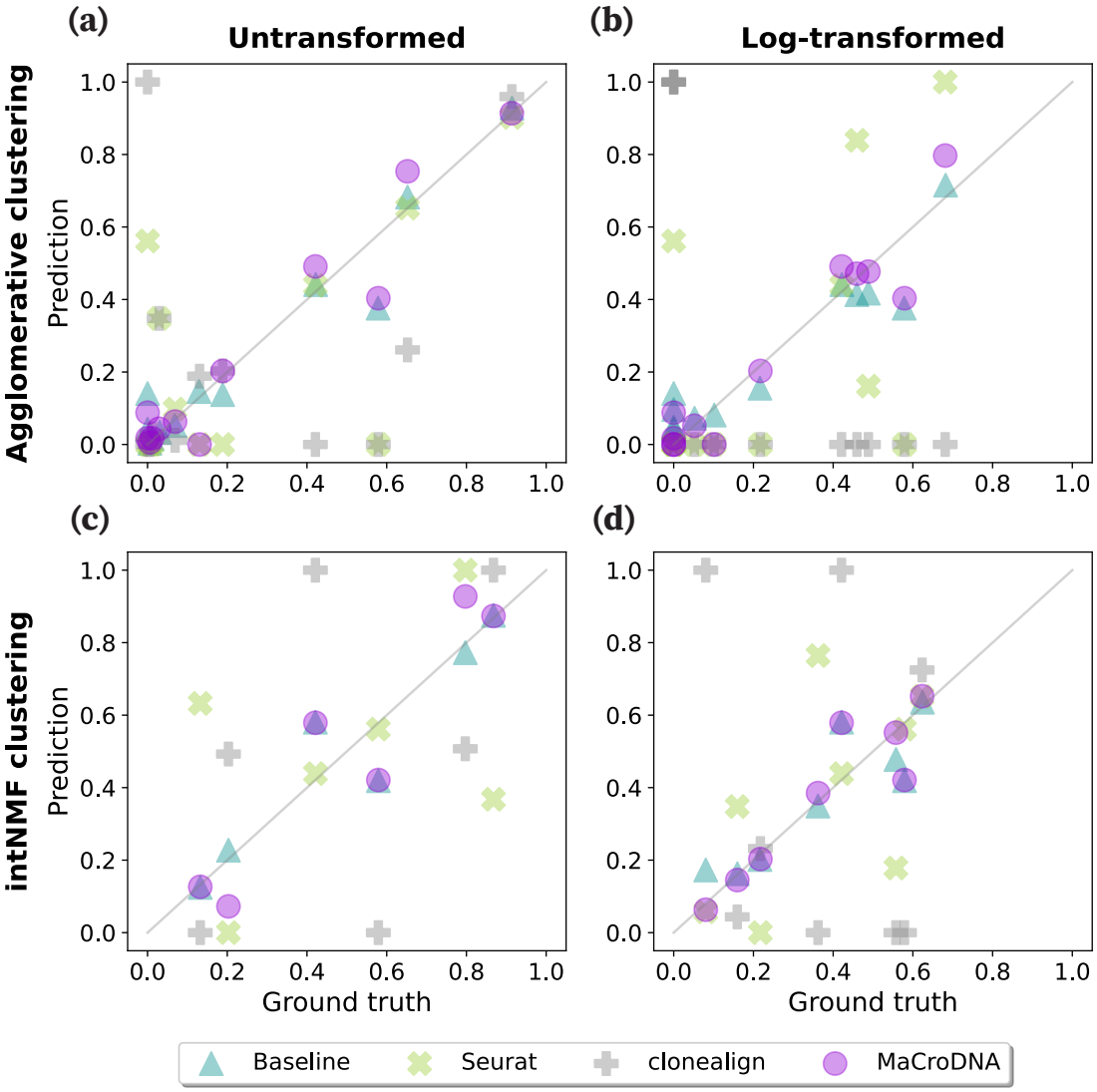

Figure 3: **Correlation between the true and predicted clonal prevalences from different methods.** In each plot, the x and y-axis indicate the true and predicted clonal prevalences, respectively. Each dot shows the true and predicted clonal prevalence of a patient-specific clone obtained from a method when when the clustering algorithm and the input data to the clustering algorithm were **(a)** the agglomerative clustering and the original data, **(b)** the agglomerative clustering and the log-transformed data, **(c)** intNMF and the original data, and **(d)** intNMF and the log-transformed data.

the results of agglomerative clustering with the highest resolution, i.e., four clusters per patient. As illustrated in Fig. 2, MaCroDNA had the lowest variance and highest median value among all methods in all four different configurations. Not only did MaCroDNA achieve the highest accuracy but it also was robust to the choice of clustering method and data transformation. In contrast, clonealign's performance was highly affected by these two factors. The median of clonealign's accuracy was lower than that of the baseline except for the combination of agglomerative clustering and original data where it is comparable to that of the baseline (Fig. 2a). Compared to clonealign, Seurat showed a relatively better performance. However, its median was not much higher than that of the random baseline for the choices of agglomerative clustering and original data (Fig. 2a) and intNMF and log-transformed data (Fig. 2d).

Next, we assessed the performance of the methods in predicting the clonal prevalences. Fig. 3 shows the true and predicted clonal prevalences. Each data point in Fig. 3 shows the proportion of a particular clone according to the ground truth on the x-axis and the proportion of the same clone predicted by a method on the y-axis. Here, again, we have used the results of agglomerative clustering at the highest-resolution setting, i.e., four clusters per patient (12 clusters in total). In such a plot, the less deviated a method's results are from the diagonal line, the more correlated the true and predicted clonal prevalences are. As shown in Fig. 3, MaCroDNA's results were more aligned with the diagonal line compared to Seurat's and clonealign's in all cases. The most diverged cases were observed in clonealign's results where either a relatively large clone in the ground truth was assigned almost no scRNA-seq cell, or a small clone was assigned most of the scRNA-seq cells. It is worth noting that the random baseline was as accurate as MaCroDNA in all cases.

This experiment emphasized the importance of respecting the same underlying clonal abundance in the two modalities as a determining factor in the methods' performance. Moreover, the performance of the random baseline in the two above experiments showed that a method can be highly accurate in predicting the clonal prevalence while inaccurate in the cell-to-clone assignment.

1.3 Genomic heterogeneity drives differential expression of cancer-related genes in high-grade Barrett's esophagus and esophageal adenocarcinoma

To demonstrate the utility of MaCroDNA within the field of cancer biology, we applied it to previously published BE data [31]. In the original study, scDNA-seq and scRNA-seq experiments were performed on multiple biopsies to identify distinctive genomic and transcriptomic—including copy number and gene expression—signatures of NDBE, dysplastic BE, and EAC. Within individual cells, changes to gene expression can result from *cis* or *trans* genomic mutations. This, however, is just one contribution among a multitude of factors. Genetically identical cells which have differentiated into distinct cell types will exhibit completely different gene expression profiles [42]. Transcriptomes undergo massive changes as cells proceed through their cell cycles [43]. Cells also respond to transient environmental signals such as hormones or the availability of nutrients [44].

Table 2: Cells and genes from esophageal biopsies after post-processing and analysis

Biopsy		Cell counts		All overlapping genes		COSMIC genes $p < 0.05$ and $K^* > 1$	
PID[1]	Histology	RNA	DNA	No. any p and $K*$	No. $p < 0.05$ and $K^* > 1$	No.	Max K* (gene)
20	CARD[2]	249	342	6671	98	7	1.12 (*MUC1*)
20	ESO[3]	181	339	7581	2	0	—
9	NDBE[4]	160	362	3783	35	4	1.01 (*ERBB2*)
14	NDBE	259	200	9213	8	0	—
16	NDBE	237	227	7217	0	0	—
6	LGD[5]	221	353	7818	58	0	—
19	LGD	354	317	8566	5	0	—
6	HGD[6]	189	312	8308	485	35	1.54 (*ERBB2*)
14	HGD	172	117	7282	2612	119	1.27 (*CXCR4*)
20	HGD[7]	233	339	7148	416	39	1.33 (*PTPRC*)
16	EAC[8]	187	274	7713	244	20	1.46 (*GNAS*)

[1] Patient identification number from the original study
[2] Healthy gastric cardia
[3] Healthy esophagus
[4] Non-dysplastic Barrett's esophagus
[5] Low-grade dysplastic Barrett's esophagus
[6] High-grade dysplastic Barrett's esophagus
[7] The first of two HGD biopsies from this patient separated by at least 5cm
[8] Esophageal adenocarcinoma

We hypothesized that within healthy tissue, differential expression would mostly be driven

by factors other than mutations, but as BE develops and progresses to dysplasia and EAC, the accumulation of mutations would cause genomic mutations to become an increasing driver of transcriptomic heterogeneity. To test our hypothesis, we first selected the biopsies for which both the scDNA-seq and scRNA-seq were provided. This includes one EAC biopsy, eight BE biopsies, and one biopsy each from healthy esophageal (ESO) and gastric cardia (CARD) tissues. After filtering and normalizing DNA and RNA data sets, we created separate DNA and RNA read count tables for genes with mapped RNA reads and which also overlapped bins with mapped DNA reads (Table 2). We then used MaCroDNA to assemble a set of virtual cells; each virtual cell merged one cell from the scDNA-seq data set, and one cell from the scRNA-seq data set, based on the one-to-one correspondences identified by MaCroDNA. Since no ground truth correspondence information was available for these data sets, MaCroDNA was an indispensable part of our analysis.

To measure the contribution of genomic mutations to transcriptomic changes, we used the K^* statistic of phylogenetic signal, which quantifies correlations between a continuous trait and a phylogenetic tree [45]. For our analysis, the trait was the normalized read count for a particular gene. The phylogenetic tree, which was estimated from normalized binned DNA read counts as a proxy for copy number profiles, encoded cell lineage information. Using a phylogenetic tree to model cell lineages is natural, given the bifurcating nature shared by such trees and by the cell cycle, and correlations between cell transcriptome and lineage will result from mutations or environmental factors which are both specific to sub-lineages of cells within a biopsy and have some effect on the transcriptome. K^* has an expected value of 1 when traits evolve along the tree following a Brownian motion (BM) model, where the variance of a trait increases linearly with time, so a K^* of 1 would represent a very strong correlation between cell transcriptome and lineage. $K^* < 1$ or $K^* > 1$ when the correlation is weaker or stronger respectively than expected under BM [45].

Initial inspection of the distribution of K^* values for all post-filtering genes within each biopsy revealed that the bulk of gene expression patterns exhibit correlations with the cell lineage phylogeny below what would be expected under BM, regardless of tissue type or BE staging (Figure 4). Since a calculation for the expected value of K^* when traits are completely independent of the phylogeny has yet to be derived, especially when considering measurement error of gene expression and

estimation error of the cell lineage phylogeny, it is not possible to appraise the weakness of this correlation. Further inspection uncovered a substantial number of genes had expression patterns inferred as being more strongly correlated with the cell lineage phylogeny than expected under BM in HGD and EAC biopsies (Figure 4h-k), but not in healthy tissues or less advanced stages of BE (Figure 4a-g).

In simulation studies, phylogenetic signal stronger than expected under BM can be induced by increasing the amount of change in a trait that occurs deeper in the tree (e.g. along internal branches) relative to the amount of change towards the tips [46]. Following from that, there are complementary plausible explanations of greater than expected signal in the BE data, both adaptive and neutral. Sometimes large changes in gene expression may confer a selective advantage, ensuring such changes are preserved within sub-lineages of cells. For example, amplification of the HER2-encoding gene *ERBB2* results in HER2 overexpression which drives proliferation and survival, and HER2-positive cells are dependent on this overexpression to avoid growth inhibition and apoptosis [47]. While mutations that induce large changes in gene expression are rare [48], as cancer develops and progresses mutation rates rise, increasing the chance that such mutations will occur.

After restricting our analysis to gene expression patterns with statistically significant phylogenetic signal based on the K^* test [45], we identified hundreds of genes with inferred $K^* > 1$ in every HGD and EAC biopsy. In biopsies of lower grade or healthy tissue, far fewer genes were identified following the same criteria (Table 2). Many of the identified genes are present in the Cancer Gene Census [49] of genes causally implicated in cancer, including *ERBB2*, which had the highest inferred K^* of any statistically significant COSMIC genes in two biopsies (Table 2, Supplementary Table 1).

1.4 Computational cost

We measured the computational cost, including the runtime and peak memory usage of each method with the clusters inferred by intNMF. To study the effect of input size on computational cost of the methods, we increased the size of each patient's data by sampling the cells with replacement prior to running a method on each patient. We randomly sampled a proportion of cells λ in increments of 0.1 from the original set of cells from each patient, and replicated this sampling 10 times per

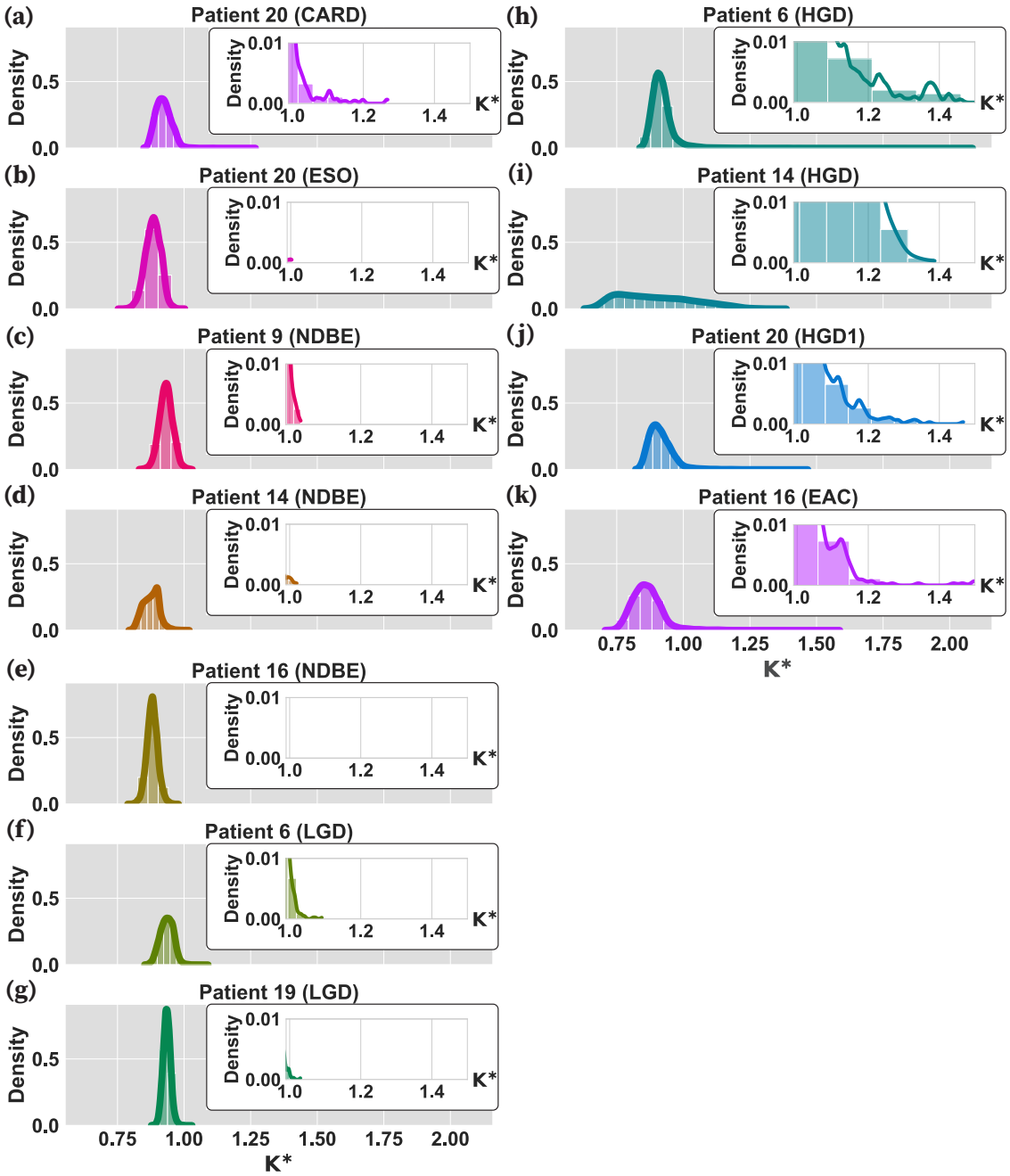

Figure 4: **Distribution of the K^* indices for all genes in each biopsy.** Each panel illustrates the distribution curve of the $K*$ indices calculated by phylosignal for all the genes in a biopsy. In each panel, the x-axis and y-axis are the $K*$ index values and their density, respectively. The left panels are the distribution curves of the healthy, NDBE, and LGD biopsies **(a-g)**. The distribution curves of the HGD and EAC biopsies are shown on the right panels **(h-k)**. Inserts within each panel magnify the bars and curve for $1 < K^* < 1.5$.

method and λ value. Because Seurat encountered errors when the number of cells is too small, the lower bound on λ was 0.5. Since the amount of genes can also have a significant impact on computational cost, both `2000_genes_log` and `all_genes_log` (see Preprocessing on CRC data set) were used as the preprocessed data for MaCroDNA

The simulation was run on Ubuntu 20.04.4 using one core of an Intel Core i9-10900K CPU. To measure the runtime of a method, first, we calculated the method's average runtime for each patient across 10 replications, next, we reported the sum of the average running times across all patients. As shown in Figs. 5a,b, MaCroDNA had the lowest runtime when input was small (e.g., for 418 and 501 cells). We observed that the runtime of clonealign and MaCroDNA increased with the size of data, while that of Seurat was relatively unchanged since its runtime depends mostly on the dimension of the latent space in canonical correlation analysis (which was fixed at 30), rather than the size of the input.

To estimate the runtime of MaCroDNA for larger samples outside of the observed range, we fitted a linear regression between MaCroDNA's running time measures and the input sizes for the original and log-transformed data, separately (Figs. 5a,b). The fitted R^2 values in both cases were greater than 0.96. Based on the fitted line, we can extrapolate the runtime of MaCroDNA to be around 10 minutes for 20,000 cells.

For memory usage of each method we reported its peak of memory usage across all patients and all replications per sampling rate. The peak memory usage of each method is presented in Figs. 5c,d. MaCroDNA outperformed the other methods with only 0.18 GB for 835 cells. Seurat, on the other hand, required 1.15 GB of memory for the smallest input with 418 cells, which is 6 times the highest peak memory usage of MaCroDNA. Similarly, the lowest peak memory usage of clonealign was 10 times the highest peak memory usage of MaCroDNA with 1.8 GB for 418 cells. The above experiments showed that MaCroDNA was very efficient in both time and memory consumption compared to the other two methods.

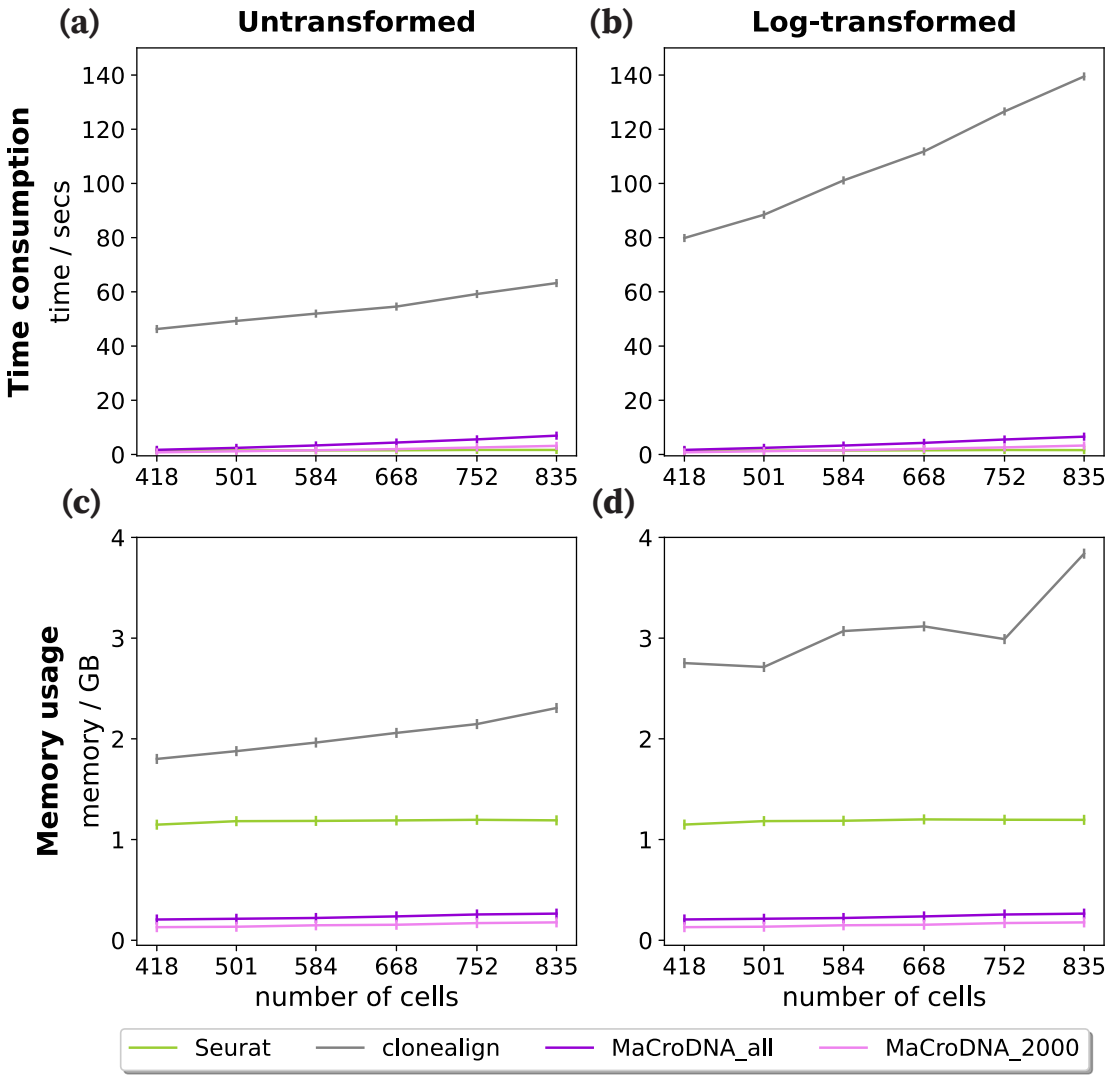

Figure 5: **Time consumption and memory usage of clonealign, Seurat and MaCroDNA using different amount of cells.** The data was sampled separately in each patient at the same sampling rate. For MaCroDNA, both `2000_genes_log` and `all_genes_log` are used as the preprocessed data. **(a)** is the time consumption using the untransformed clones. **(b)** is the time consumption using the log-transformed clones. **(c)** is the peak memory usage using the untransformed clones. **(d)** is the peak memory usage using the log-transformed clones.

2 Discussion

High-throughput scRNA-seq and scDNA-seq can measure either the genomic or transcriptomic contents of the cells. Such abundant scDNA-seq and scRNA-seq data sets, if integrated precisely, would provide an unprecedented opportunity for studying the transcriptomic effects of the genomic alterations. Ensuring this *precise integration* is not an easy task as the integration of the single-cell data across different measurements and samples—in a broader perspective—has been identified as one of the *"eleven grand challenges in single-cell data science"* [50]. In this work, we developed a new method for mapping the single cells with only gene expression measurements to single cells with only copy number values from the high-throughput scRNA-seq and scDNA-seq data. Our method, named MaCroDNA, finds the correspondence between the scRNA-seq and scDNA-seq cells by maximizing the sum of the Pearson correlation coefficients between the paired cells. We also pose a restriction on this correspondence: The number of scRNA-seq cells assigned to a scDNA-seq cell should be limited as much as possible to avoid the trivial solutions (see MaCroDNA for more details).

We assessed MaCroDNA's performance against clonealign [21]—as the state-of-the-art method for integrating the scRNA-seq and scDNA-seq data—and Seurat [19]—as the representative of the commonly-used methods in the single-cell multi-omic studies—through their application to a CRC data set [30]. This data set contains the cells measured for both gene expression and copy number data. We exploited such cells' information as the ground truth. To the best of our knowledge, this is the first study that uses such data for the evaluation purposes. We evaluated the accuracy of the methods for two inference tasks, namely the cell-to-clone assignment and prediction of the clonal prevalences. MaCroDNA outperformed clonealign and Seurat in both tasks by a large margin. We hypothesize that the key to the excellent performance of MaCroDNA is restricting the number of the scRNA-seq cells being assigned to each scDNA-seq cell. This restriction helped MaCroDNA implicitly assume the same underlying clonal prevalences in the two modalities, whereas clonealign and Seurat do not—either explicitly or implicitly—account for such an assumption.

Among the three methods, clonealign showed the poorest performance. It is worth noticing that the only *clonal* inputs to clonealign are the clone-specific copy numbers (a CNA profile for

each clone). Therefore, clonealign is oblivious to the distribution of the clones in the tumor. Another contributing factor that may have degraded the performance of clonealign could be its *"dosage function"*. Although MaCroDNA and clonealign are both based on the assumption that an increase (decrease) in the copy number of a gene can increase (decrease) the corresponding gene expression value at that gene, the linear dosage function of clonealign that maps the copy number of a gene to its expression is *"capped at an upper limit"* [21]. Since the Pearson correlation—which we used as our dosage function—was not restricted or capped, we speculate that limiting the linearity between copy number and gene expression values in clonealign have had rather a negative impact on its performance.

Although our experiments for evaluation of MaCroDNA's accuracy showed the robustness of our method to the choice of clustering technique, we are interested in a simultaneous inference of the scDNA-seq clones and cell-to-clone assignments for future research. We expect this approach to result in an even better accuracy. Furthermore, unlike statistical methods such as clonealign, MaCroDNA does not provide the probabilistic confidence measures for the assignments. As a future direction we will explore adding uncertainty to MaCroDNA's framework. MaCroDNA is a simplistic approach—in both the assumptions and implementation—that provides highly accurate results.

As a reliable and easy-to-run tool, MaCroDNA would be a suitable method of choice for biological scientists interested in single-cell multi-omic (especially cancer) studies. Our re-analysis of the BE data set demonstrates its suitability by identifying strong phylogenetic signal of the HER2-encoding gene *ERBB2* in HGD and NDBE biopsies. This identification supports previous findings of intratumoral heterogeneity of HER2 expression in EAC, but also suggests that this heterogeneity can arise in earlier stages. Those previous findings were based on in situ hybridization to assess *ERBB2* copy number amplification [51] and immunohistochemistry to assess HER2 protein overexpression [52]. Unlike in situ hybridization or immunohistochemistry, which must be tailored to individual genes in advance, MaCroDNA unlocks the entire genome and transcriptome for interrogation and further study.

3 Methods

3.1 Description of CRC data set

To assess the performance of the integration methods, we used a colorectal cancer (CRC) data set introduced by Bian *et al.* [30]. To generate this data set, the authors used single-cell triple omics sequencing2 (scTrio-seq2) sequencing technique. This sequencing technology provides the joint measurement of somatic copy number variations, DNA methylation, and gene expression read counts from the same cell. In the original study, for six patients, namely CRC01, CRC02, CRC04, CRC09, CRC10, CRC11, the joint copy number and gene expression measurements were provided[1]. The sequenced cells with joint copy numbers and gene expression values from these patients are highly valuable for evaluation purposes. However, we did not include all patients' data sets in our analysis.

The authors used Tang protocol [53] for transcriptome sequencing library preparation of CRC01 and CRC02, while for the other patients, multiplexed scRNA-seq method [54] was used. To maintain consistency across our experiments, we did not consider CRC01 and CRC02 in our analysis. Besides, the data of the patient CRC09 contained only 13 cells with both scRNA-seq and scDNA-seq data, which was insufficient for a statistical method such as clonealign. Finally, only the data from patients CRC04, CRC10, and CRC11 were used to evaluate MaCroDNA and the other integration methods.

3.2 Preprocessing on CRC data set

First, we filtered out the non-informative genes from the scRNA-seq gene expression data by keeping only genes that were expressed as non-zero in more than 1% of cells. The genes obtained from filtering are used for the further preprocessing in the integration methods and the clustering techniques. In the following, we will describe our data preparation procedures for the integration methods and the clustering methods. Specifically, data preparation for each of the existing methods is based on the instructions in their original studies.

[1]Note that not all the cells in these patients have both scRNA-seq and scDNA-seq data.

clonealign requires removing the X-chromosome genes prior to the analysis since the presence of the X-chromosome in the data will violate the method's assumption about the relationship between the copy number and gene expression values. After removing X-chromosome, we removed the genes with zero variance from the input data as is required by the clonealign's original study.

Seurat requires different preprocessing procedures. We followed the same procedures in the original paper. First, all cells were log-transformed using a size factor of 10,000 molecules for each cell. Then, z-score normalization was applied to the log-transformed data. Next, a feature selection was done to scRNA-seq data and scDNA-seq data, separately, using `FindVariableFeatures` function in Seurat v3. Following the feature selection on the single measurements, feature selection for integrated analysis was applied to scRNA-seq data and scDNA-seq data together using `SelectIntegrationFeatures` function in Seurat v3. Finally, the top 2,000 genes were selected.

In addition to the Seurat's and clonealign's preprocessing procedures, we considered two other preprocessing strategies for MaCroDNA, resulting in four strategies in total. The first strategy is Seurat's preprocessing, named as `2000_genes_log` in the figure legends in the Supplementary materials. This includes log transformation of the data, z-score normalization with 10,000 factor size, feature selection, and the selection of the top 2,000 genes. All these steps were applied sequentially to the input. The second one includes the log transformation and z-score normalization (with 10,000 factor size), but all genes were kept. This preprocessing is named as `all_genes_log` in the Supplementary figures. The third one is the clonealign's preprocessing, tagged with `noX_genes_raw`. Finally, the last strategy is `all_genes_raw`, where all genes were used, and the data were kept in the original scale. In Figs. 2 and 3, we used `2000_genes_log` preprocessing for Seurat (as it is recommended by the authors) and `all_genes_log` for MaCroDNA. For clonealign, we applied `noX_genes_raw`. The full comparison of these preprocessing strategies and their effect on the accuracy of the methods can be found in the Supplementary Figures S1–S18.

Apart from data preprocessing for the integration methods analyses, the CNA data used for obtaining clusters was also preprocessed. Here, we use two strategies: clustering on the original CNA data and clustering on the log-transformed CNAs. The CNA data can have an extremely right-skewed distribution. Thus, $\log(x + 1)$ was applied, where x is the original copy number.

Through all experiments, both data scaling strategies were used for clustering and clonal inference.

3.3 Preprocessing on BE data set

For scRNA-seq data of each biopsy, first, we filtered out the single cells with less than 3000 transcripts from the data, and next, we kept the genes that were expressed by at least three transcripts in at least one cell, according to the filtering procedure in [31]. For numerical stability of our calculations, we added a pseudocount of 1 to all entries of the scRNA-seq count tables which is a common practice in gene expression analysis [55]. To circumvent the effect of unequal sequencing depth, we performed RPM (reads per million mapped reads) normalization by first dividing each gene's expression value by the total number of reads in the cell, and then scaling this ratio by one million. Finally, we log-normalized the gene expression values by applying $\log(x+1)$ where x is the gene expression value and $\log(\cdot)$ is the natural logarithm.

The scDNA-seq data was provided to us as the binned genomes containing the number of mapped reads in each copy number bin. Similar to scRNA-seq data, we filtered out the scDNA-seq cells with less than 3000 mapped reads. Next, we added the pseudocount of 1 to the number of mapped reads in each bin. Following the instructions from the original study [31], we divided the copy number bin counts by their median and then multiplied by 2 to obtain a rough estimation of the copy number values. In the next step, we log-normalized the copy number values using $\log(x+1)$.

Running MaCroDNA on the scDNA-seq and scRNA-seq data requires the count tables from the two modality with the same set of genes. As mentioned above, the given data was the binned copy number data. Therefore, we first annotated the copy number bins by searching for the genes whose genomic coordinates lie within the bins. We downloaded the GENCODE [56] annotation file (current release for GRCh37 assembly) and extracted the names and coordinates of all known protein-coding genes. Next, for each copy number bin, we annotated it with all the genes whose starting and ending points were within the bin. For each gene, we considered the copy number of the corresponding bin as that gene's copy number value. Lastly, for each biopsy, we found the genes shared between the scRNA-seq and scDNA-seq data sets and performed MaCroDNA on this

set of genes.

3.4 Clonal inference by intNMF

intNMF is a clustering method based on non-negative matrix factorization [57, 28, 29]. intNMF is able to identify the optimal number of clusters for each data type in a multi-omic data set. Its clustering algorithm can also be applied to data sets with single data types (such as scDNA-seq data in CRC data set). In intNMF, the cluster prediction index (CPI) can evaluate the goodness of a particular number of clusters by measuring the adjusted rand indices [58] among the clustering assignments (for more details, see [36]). The optimal number of clusters is the one that yields the maximum CPI value.

We applied intNMF to the scDNA-seq cells of each patient to infer the clusters. First we ran the function, `nmf.opt.k` from the R package of `IntNMF` to identify the optimal number of clusters in each patient. Next, to perform clustering on the scDNA-seq cells and acquire the clustering membership assignments, we ran the function, `nmf.mnnals` with the optimal number of clusters obtained from the previous step as the input. Given the original data, `IntNMF` detected two clusters in each of the patients, CRC04, CRC10 and CRC11. These patient-specific clusters are denoted by crc04_clone0, crc04_clone1, crc10_clone0, crc10_clone1, crc11_clone0, and crc11_clone1 in Figure S16. Using the log-transformed data as input, `IntNMF` identified two clusters in patient CRC04 denoted by crc04_clone0, crc04_clone1 and crc04_clone3 in Figure S18. Three clusters were identified in each of the patients, CRC10 and CRC11. These clusters are denoted by crc10_clone0, crc10_clone1, crc10_clone2, crc11_clone0, crc11_clone1, crc11_clone2, in Figure S18.

3.5 Clonal inference by agglomerative clustering

In addition to intNMF, we applied agglomerative clustering to obtain scDNA-seq clones from each patient's cells. The Python implementation of agglomerative clustering provided in `sklearn` package was directly applied for obtaining 4 clusters in each patient. We used Euclidean distance as the distance metric and *average* as the linkage criterion. By obtaining 4 clusters for each patient, we inferred 12 clusters in total across all patients. Since we sought to investigate the effect of varying

clustering resolution on the methods' accuracy, we applied the following procedure to reduce the number of clusters incrementally: to select 11 clusters out of 12 clusters, the two clusters with least Euclidean distance between their averaged copy number profiles are merged into a new cluster. We required these two clusters to be from the same patient. This merging procedure was repeated until each patient only had two remaining clusters. This way, we varied the resolution of the clusters obtained from all patients, ranging from 6 clusters to 12 clusters in total. The patient-specific clones for different resolutions are tagged in Figures S2–S7 and Figures S9–S14 of the Supplementary materials.

3.6 Random baseline

To evaluate the performance of MaCroDNA and other methods, we established a random baseline. In this random baseline, each scRNA-seq cell is assigned to a scDNA-seq cell. This scDNA-seq cell is selected by random sampling from all scDNA-seq cells in the same patient with replacement. Following this random method for assignment, we can calculate for each patient i, the expected number of scRNA-seq cells (whose true clone id is c_{true} according to the ground truth information and the current clustering assignment) that are being assigned to the scDNA-seq cells with clone id $c_{\text{predicted}}$. Let $n_{\text{true,predicted}}$ denote the above expected number. Then we have:

$$n_{\text{true,predicted}} = n^{\text{RNA}}_{c_{\text{true}}} \cdot n^{\text{DNA}}_{c_{\text{predicted}}} / N^{\text{DNA}}_{i}, \tag{1}$$

where $n^{\text{RNA}}_{c_{\text{true}}}$ is the number of scRNA-seq cells that belong to clone c_{true} in patient i, $n^{\text{DNA}}_{c_{\text{predicted}}}$ is the number of scDNA-seq cells in clone $c_{\text{predicted}}$ in patient i, and N^{DNA}_{i} is the total number of scDNA-seq cells in patient i. This expected number gives us a measurement for the accuracy of the baseline method.

3.7 Phylogenetic signal analysis on the BE data set

To prepare the inputs to phylosignal, we first inferred the single-cell CNA phylogenetic tree of each biopsy using the UPGMA method. We computed the Euclidean distances between all pairs of the cells and passed the pairwise distance matrix as the input to `upgma` function in phangorn `R`

package [59] to infer the UPGMA tree. As mentioned earlier, we used the copy number values of the genes shared between the two modalities to perform MaCroDNA. Here, however, for phylogenetic inference, we used the complete set of copy number bin counts in order to take advantage of all the available evolutionary information from scDNA-seq data for the phylogenetic tree reconstruction.

In addition to the phylogenetic tree, phylosignal requires a table of traits for each cell. Given the cell-to-cell assignments inferred by MaCroDNA, for each scDNA-seq cell, we provided the vector of gene expression values of the paired scRNA-seq cell as the cell's traits to phylosignal. Of note, some of the scDNA-seq cells were assigned to multiple scRNA-seq cells by MaCroDNA. In such cases, for one particular scDNA-seq cell, we ranked the assigned scRNA-seq cells according to the iteration number they were assigned to the scDNA-seq cell (see MaCroDNA for $N_{\mathbf{G}} > N_{\mathbf{C}}$ for more details). An earlier iteration in this case means a better correspondence score. Therefore, we selected the scRNA-seq cell that was assigned to the scDNA-seq in the earliest iteration. Here, for each paired scRNA-seq, we passed the gene expression of all genes (rather than only the shared genes with the DNA data) as the traits to phylosignal.

We ran the `phyloSignal` function in phylosignal's `R` package with the UPGMA tree and the single-cell gene expression as inputs to compute the $K*$ statistic for each gene. That function computes the p-value of the $K*$ index by testing it against the null hypothesis of the gene expression values being randomly distributed in the phylogeny [60]. Throughout our analysis on the BE data set, we considered the indices with p-values less than 0.05 as the significant indices.

The $K*$ statistic [45], as a measure of phylogenetic signal is calculated for each gene using the equation

$$K^* = \left(\frac{\mathrm{MSE}^*}{\mathrm{MSE}}\right) \Big/ \mathbb{E}_{\mathrm{BM}}\left[\frac{\mathrm{MSE}^*}{\mathrm{MSE}}\right] \tag{2}$$

where MSE^* is the mean squared error between the observed gene expression values at the tips of the phylogenetic tree and their mean. This is calculated as:

$$\mathrm{MSE}^* = \frac{(\mathbf{X} - \bar{a})^{\mathsf{T}}(\mathbf{X} - \bar{a})}{(n-1)} \tag{3}$$

where $\mathbf{X}$ is the vector of observed gene expression values of the scRNA-seq cells for a gene, $\bar{a}$ is the

average gene expression value, and n is the number scRNA-seq cells [45]. MSE is the mean squared error between the gene expression values of cells and their "*phylogenetically correct*" mean [45], calculated using the equation

$$\text{MSE} = \frac{(\mathbf{U} - \hat{a})^{\mathsf{T}}(\mathbf{U} - \hat{a})}{n-1} \tag{4}$$

where $\mathbf{U} = \mathbf{DX}$ is the transformed $\mathbf{X}$ obtained from generalized least-squares [45]. The connection to the phylogenetic structure is made through the matrix $\mathbf{D}$ which satisfies the constraint $\mathbf{DVD}^{\mathsf{T}} = \mathbf{I}$ where $\mathbf{V}$ is the phylogenetic variance-covariance matrix [61, 62, 45] and $\mathbf{I}$ is the identity matrix. The phylogenetic variance-covariance matrix is calculated based on the branch lengths of the phylogeny: the variance of each cell is defined as the sum of the branch lengths starting from the root to the cell, and the covariance between two cells is the sum of the branch lengths from the root to their most recent common ancestor. $\mathbb{E}_{\text{BM}}$ in equation 4 indicates the expected value of the MSE*/MSE ratio under BM [45].

3.8 MaCroDNA

The input of MaCroDNA consists of two single-cell data sets supposedly sampled from the same tissue: one contains CNA values of some single cells at a particular set of genes. The other data set consists of gene expression values belonging to another set of single cells at the same genes as in the CNA data set.

These two data sets are provided as inputs to MaCroDNA in the form of count matrices, $\mathbf{C}$ and $\mathbf{G}$, for CNA and gene expression data, respectively. Formally, let $\mathbf{C} = (c_{ij}) \in \mathbb{N}_0^{N_{\mathbf{C}} \times M}$ be the matrix of single cells' CNA values. Here, $\mathbb{N}_0$ represents the non-negative integers, $N_{\mathbf{C}}$ is the number of single cells, M denotes the number of genes, and c_{ij} is the copy number value in the ith single cell at the jth gene. Similarly, $\mathbf{G} = (g_{ij}) \in \mathbb{Z}^{N_{\mathbf{G}} \times M}$ is the matrix of single cells' gene expression values, where $N_{\mathbf{G}}$ denotes the number of single cells, and g_{ij} is the gene expression value in the ith single cell at the jth gene (the rest of the notations for $\mathbf{G}$ are defined the same as they are for $\mathbf{C}$). Each row in $\mathbf{C}$ or $\mathbf{G}$ represents the CNA profile or gene expression vector of a single cell. We use $\mathbf{c}_i$ and $\mathbf{g}_i$ to denote the ith cell's CNA profile and gene expression vector, respectively.

Given the above matrices, MaCroDNA associates each cell in gene expression data (scRNA-

seq cell) to exactly one cell in CNA data (scDNA-seq cell) according to a criterion described in the following. This correspondence is presented in the form of a binary matrix, $\mathbf{A} = (a_{ij}) \in \{0,1\}^{N_\mathbf{G} \times N_\mathbf{C}}$, where a_{ij} is 1 if the ith scRNA-seq cell is associated with the jth scDNA-seq cell, otherwise a_{ij} is 0. Of note, since each scRNA-seq cell has exactly one corresponding scDNA-seq cell, each row of the correspondence matrix has exactly one entry of value 1.

In the following, we discuss how MaCroDNA infers the above-mentioned correspondence matrix by, first, describing our polynomially solvable algorithm for cases where the number of scRNA-seq cells is less than or equal to the number of scDNA-seq cells ($N_\mathbf{G} \leq N_\mathbf{C}$), and, secondly, our heuristic for data sets where the number of scRNA-seq cells is more than the number of scDNA-seq cells ($N_\mathbf{G} > N_\mathbf{C}$). Finally, we discuss how our solution for the former constitutes the basis for the solution to the latter.

3.8.1 MaCroDNA for $N_\mathbf{G} \leq N_\mathbf{C}$

As mentioned above, we aim at associating each scRNA-seq cell to exactly one scDNA-seq cell. When $N_\mathbf{G} \leq N_\mathbf{C}$, we add another constraint to this problem: each scDNA-seq cell is associated to at most one scRNA-seq cell. The main reason for this constraint is to prevent the situation where a significant number of scRNA-seq cells are associated to a particular scDNA-seq cell—which violates our underlying assumption of the same clonal distribution across the two modalities. An alternative to this constraint is to consider an upper limit on the number of correspondents for each scDNA-seq cell. Pre-specifying or inferring this upper limit for each scDNA-seq cell, however, would be a challenging task that we leave for future exploration.

Considering the above conditions, we require the solution to be of highest similarity score between the corresponding pairs. We formulated this optimization problem as an instance of Mixed Integer Linear Programming (MILP). Specifically, we introduce binary indicator variables $I_{ij} \in \{0,1\}$ which is 1 when the ith scRNA-seq cell is assigned to the jth scDNA-seq cell; otherwise I_{ij} is 0. The objective function to maximize is the sum of the Pearson correlation coefficients (see

Section Calculation of correlation coefficients for more details) of the selected pairs,

$$\max_{I_{ij}} \left\{ \sum_{i=1}^{N_{\mathbf{G}}} \sum_{j=1}^{N_{\mathbf{C}}} \omega_{ij} I_{ij} \right\}, \tag{5}$$

where ω_{ij} is the Pearson correlation coefficient between the ith scRNA-seq cell and the jth scDNA-seq cell. Since each scDNA-seq cell is allowed to be assigned to at most one scRNA-seq cell, at most one of the indicator variables of each scDNA-seq cell can be 1. This following inequality forces this constraint for each scDNA-seq cell:

$$\sum_{i=1}^{N_{\mathbf{G}}} I_{ij} \leq 1, \quad \forall j \in \{1 \leq j \leq N_{\mathbf{C}} | j \in \mathbb{N}_0\}. \tag{6}$$

Another constraint we need to respect is that each scRNA-seq cell must have exactly one pair in scDNA-seq data. Although the constraint $\sum_{j=1}^{N_{\mathbf{C}}} I_{ij} = 1$ for each scRNA-seq cell can satisfy this condition when $N_{\mathbf{G}} \leq N_{\mathbf{C}}$, we propose two alternative constraints that yield the same results and also can be used in the cases where $N_{\mathbf{G}} > N_{\mathbf{C}}$ as well (see Section MaCroDNA for $N_{\mathbf{G}} > N_{\mathbf{C}}$ for more details). First, we force each scRNA-seq cell to have at most one pair in scDNA-seq data by the following inequality:

$$\sum_{j=1}^{N_{\mathbf{C}}} I_{ij} \leq 1, \quad \forall i \in \{1 \leq i \leq N_{\mathbf{G}} | i \in \mathbb{N}_0\}, \tag{7}$$

and then, we restrict the total number of assignments to be exactly equal to the number of scRNA-seq cells by the following equality,

$$\sum_{i=1}^{N_{\mathbf{G}}} \sum_{j=1}^{N_{\mathbf{C}}} I_{ij} = N_{\mathbf{G}}. \tag{8}$$

This is particularly useful when we need to select a certain number of assignments from all the

possible choices in our heuristic algorithm. Thus, the overall optimization problem will be

$$\begin{aligned} &\max_{I_{ij}} \left\{ \sum_{i=1}^{N_{\mathbf{G}}} \sum_{j=1}^{N_{\mathbf{C}}} \omega_{ij} I_{ij} \right\}, \\ \text{subject to} \quad &\sum_{i=1}^{N_{\mathbf{G}}} I_{ij} \leq 1, \forall j, \quad \sum_{j=1}^{N_{\mathbf{C}}} I_{ij} \leq 1, \forall i, \quad \sum_{i=1}^{N_{\mathbf{G}}} \sum_{j=1}^{N_{\mathbf{C}}} I_{ij} = N_{\mathbf{G}}, \end{aligned} \tag{9}$$

which can be solved using well-known MILP-solvers such as Gurobi (see https://www.gurobi.com) and CPLEX (see https://www.ibm.com/analytics/cplex-optimizer). After solving the optimization problem in equation (9), each indicator variable will be an entry of the binary correspondence matrix. Specifically, we have $\mathbf{A} = (a_{ij}) = I_{ij}$.

It is worth mentioning that this problem can be viewed as an instance of *maximum weight perfect matching* on a complete bipartite graph, also known as *assignment problem*. To reformulate as such, we need to introduce a graph $\mathcal{G} = (V, E)$ consisting of V vertices and E edges. Here, the vertices represent single cells from scDNA-seq and scRNA-seq. All possible pairs of single cells in scRNA-seq and scDNA-seq are connected by edges that are weighted with the Pearson correlation coefficient between the pair. The goal of the maximum weight perfect matching problem is to find a one-to-one matching between the nodes in the bipartitions such that the sum of the edge weights connecting the pairs is maximized. The well-known Hungarian algorithm [63] can solve this problem in polynomial time [35]. The time complexity of this algorithm is $\mathcal{O}(|V|^3)$ where V is the set of nodes in the bipartite graph [64, 65].

The Hungarian algorithm requires the number of scRNA-seq and scDNA-seq cells to be equal which is not the case necessarily in real applications. However, we can add *"dummy"* scRNA-seq cells that are connected to scDNA-seq cells with edges weighted by negative values with large magnitudes. By doing so, the number of cells in both bipartitions will be the same and the Hungarian algorithm is applicable.

3.8.2 MaCroDNA for $N_{\mathbf{G}} > N_{\mathbf{C}}$

When the number of scRNA-seq cells exceeds the number of scDNA-seq cells, the situation where multiple scRNA-seq cells are assigned to a small group of scDNA-seq cells seems inevitable. Here, we propose a heuristic algorithm to *"regularize"* this assignment. In our scheme, we infer the correspondence of all scRNA-seq cells in multiple steps. Formally, our iterative algorithm consists of the following steps:

- **Initialization**: let $R^{(0)}$ be the set of scRNA-seq cells at the initialization step of the algorithm, and $N_{R^{(0)}}$ be the number of cells in set $R^{(0)}$. Note that at the initialization step, all the scRNA-seq cells are involved, so $N_{R^{(0)}} = N_{\mathbf{G}}$. Perform the optimization described in the previous section in equation (9) with a slight modification to the equality constraint in equation (8): instead of $N_{\mathbf{G}}$ assignments, we need to infer the best $N_{\min} = \min\{N_{\mathbf{G}}, N_{R^{(0)}}\}$ assignments (although in the first step, $N_{\min} = N_{\mathbf{G}}$, we use this general notation for the next iterations). So, the equality constraint of equation (8) becomes:

$$\sum_{i=1}^{N_{\mathbf{G}}} \sum_{j=1}^{N_{\mathbf{C}}} I_{ij} = N_{\min}. \tag{10}$$

  After inference of the best $N_{\min}$ assignments, remove the scRNA-seq cells whose corresponding scDNA-seq cells are inferred from the set $R^{(0)}$. Name the new set, $R^{(1)}$ and proceed to the next step.

- **Iteration**: let $R^{(k)}$ denote the set of scRNA-seq cells at iteration k. Perform the optimization in equation (9) with constraint described in equation (10). Remove the scRNA-seq cells whose assignments are determined from $R^{(k)}$. Name the new set $R^{(k+1)}$. If this set is empty, go to the termination step. Otherwise, proceed to the iteration $k+1$.

- **Termination**: construct the binary correspondence matrix, $\mathbf{A}$, by collecting all the assignments inferred through all the iterations.

Notice that since we remove at most $N_{\mathbf{C}}$ cells from the set of scRNA-seq cells at each iteration, the total number of iterations will be $\lceil \frac{N_{\mathbf{G}}}{N_{\mathbf{C}}} \rceil$.

3.8.3 Calculation of correlation coefficients

To calculate Pearson correlation coefficients between each RNA-DNA pair of cells, we use the raw read counts in the CNA profile and gene expression vector of the cells ($\mathbf{g}_i$, and $\mathbf{c}_j$, respectively). Given two cells' vectors, $\mathbf{c}_i$ and $\mathbf{g}_j$, the Pearson correlation between them is calculated using the following formula

$$\omega_{ij} = \rho_{\mathbf{c}_i,\mathbf{g}_j} = \frac{\sum_{k=1}^{N}(c_{ik} - \mu_{\mathbf{c}_i})(g_{jk} - \mu_{\mathbf{g}_j})}{\sigma_{\mathbf{c}_i}\sigma_{\mathbf{g}_j}}, \tag{11}$$

where $\rho_{\mathbf{c}_i,\mathbf{g}_j}$ denotes the Pearson correlation between $\mathbf{c}_i$ and $\mathbf{g}_j$; c_{ik} and g_{jk} are k^{th} elements in vectors $\mathbf{c}_i$ and $\mathbf{g}_j$, respectively.

4 Data and code availability

All codes including the Python implementation of MaCroDNA and the scripts for data analysis in this study are publicly available at https://github.com/NakhlehLab/MaCroDNA. RNA and DNA read counts for both the CRC and BE were obtained directly from the authors of the original studies.

5 Acknowledgements

We thank Ana C. Afonso Silva for her comments on the implications of strong phylogenetic signal. Our research was supported in part by the National Science Foundation, grants IIS-1812822 and IIS-2106837 (L.N.).

6 Author contributions

M.E., X.H. and H.A.O. designed the study. All authors wrote and approved the manuscript.

7 Competing interests

The authors declare no competing interests.

References

[1] Stuart, T. & Satija, R. Integrative single-cell analysis. *Nature Reviews Genetics* **20**, 257–272 (2019).

[2] Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Molecular Cell* **58**, 598–609 (2015).

[3] Navin, N. E. Cancer genomics: one cell at a time. *Genome Biology* **15**, 1–13 (2014).

[4] Lim, B., Lin, Y. & Navin, N. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* **37**, 456–470 (2020).

[5] Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine* **52**, 1419–1427 (2020).

[6] Tang, X., Huang, Y., Lei, J., Luo, H. & Zhu, X. The single-cell sequencing: new developments and medical applications. *Cell & Bioscience* **9**, 1–9 (2019).

[7] Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer* **22**, 114–126 (2022).

[8] Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods* **12**, 519–522 (2015).

[9] Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology* **33**, 285–289 (2015).

[10] Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research* **26**, 304–319 (2016).

[11] Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 1–12 (2017).

[12] Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods* **14**, 167–173 (2017).

[13] Welch, J. D. *et al.* Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).

[14] Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology* **18**, 1–19 (2017).

[15] Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).

[16] Liu, J., Huang, Y., Singh, R., Vert, J.-P. & Noble, W. S. Jointly embedding multiple single-cell omics measurements. In Huber, K. T. & Gusfield, D. (eds.) *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, vol. 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 10:1–10:13 (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019).

[17] Yang, K. D. *et al.* Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications* **12**, 1–10 (2021).

[18] Amodio, M. & Krishnaswamy, S. MAGAN: Aligning biological manifolds. In Dy, J. & Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, 215–223 (PMLR, 2018).

[19] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

[20] Dou, J., Liang, S. *et al.* Bi-order multimodal integration of single-cell data. *Genome Biology* **23**, 112 (2022).

[21] Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology* **20**, 54 (2019).

[22] McCarthy, D. J. *et al.* Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature Methods* **17**, 414–421 (2020).

[23] Bai, X., Duren, Z., Wan, L. & Xia, L. C. Joint inference of clonal structure using single-cell genome and transcriptome sequencing data. *bioRxiv* (2020). URL https://www.biorxiv.org/content/early/2020/10/07/2020.02.04.934455.

[24] Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

[25] Han, K. Y. *et al.* SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Research* **28**, 75–87 (2018).

[26] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233 (1999).

[27] Yang, X., Liu, W., Liu, W. & Tao, D. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering* **33**, 2349–2368 (2021).

[28] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

[29] Lee, D. & Seung, H. S. Algorithms for non-negative matrix factorization. In Leen, T., Dietterich, T. & Tresp, V. (eds.) *Advances in Neural Information Processing Systems*, vol. 13, 556–562 (MIT Press, 2000).

[30] Bian, S. *et al.* Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).

[31] Busslinger, G. A. *et al.* Molecular characterization of Barrett's esophagus at single-cell resolution. *Proceedings of the National Academy of Sciences* **118**, e2113061118 (2021).

[32] Spechler, S. J. Barrett's esophagus. In Shaker, R., Belafsky, P. C., Postma, G. N. & Easterling, C. (eds.) *Principles of Deglutition: A Multidisciplinary Text for Swallowing and its Disorders*, 723–738 (Springer New York, New York, NY, 2013).

[33] Hayakawa, Y., Sethi, N., Sepulveda, A. R., Bass, A. J. & Wang, T. C. Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? *Nature Reviews Cancer* **16**, 305–318 (2016).

[34] Shaheen, N. J. *et al.* Diagnosis and management of Barrett's esophagus: an updated ACG guideline. *The American Journal of Gastroenterology* **117**, 559–587 (2022).

[35] Munkres, J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics* **5**, 32–38 (1957).

[36] Chalise, P. & Fridley, B. L. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLOS One* **12**, e0176278 (2017).

[37] Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLOS One* **9**, e85150 (2014).

[38] Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biology* **20**, 1–16 (2019).

[39] Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by $\log(x+1)$ or $\log\ (1+x)$. *Bioinformatics* **37**, 2223–2224 (2021).

[40] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 1–5 (2018).

[41] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

[42] Dynlacht, B. D. Regulation of transcription by proteins that control the cell cycle. *Nature* **389**, 149–152 (1997).

[43] He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biology* **21**, 1–34 (2020).

[44] Mullur, R., Liu, Y.-Y. & Brent, G. A. Thyroid hormone regulation of metabolism. *Physiological Reviews* (2014).

[45] Blomberg, S. P., Garland Jr, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).

[46] Felizola Diniz Filho, J. A., Rangel, T. F., Santos, T. & Mauricio Bini, L. Exploring patterns of interspecific variation in quantitative traits using sequential phylogenetic eigenvector regressions. *Evolution* **66**, 1079–1090 (2012).

[47] Moasser, M. M. The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene* **26**, 6469–6487 (2007).

[48] Metzger, B. P. H. *et al.* Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. *Molecular Biology and Evolution* **33**, 1131–1146 (2016).

[49] Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

[50] Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 1–35 (2020).

[51] Yoon, H. *et al.* Adverse prognostic impact of intratumor heterogeneous HER2 gene amplification in patients with esophageal adenocarcinoma. *Journal of Clinical Oncology* **30**, 3932–3938 (2012).

[52] Cappellesso, R. *et al.* HER2 status in gastroesophageal cancer: a tissue microarray study of 1040 cases. *Human Pathology* **46**, 665–672 (2015).

[53] Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).

[54] Dong, J. *et al.* Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biology* **19**, 31 (2018).

[55] Erhard, F. Estimating pseudocounts and fold changes for digital expression measurements. *Bioinformatics* **34**, 4054–4063 (2018).

[56] Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

[57] Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).

[58] Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193–218 (1985).

[59] Schliep, K. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

[60] Keck, F., Rimet, F., Bouchez, A. & Franc, A. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution* **6**, 2774–2780 (2016).

[61] Garland Jr, T., Dickerman, A. W., Janis, C. M. & Jones, J. A. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* **42**, 265–292 (1993).

[62] Garland, T. & Janis, C. M. Does metatarsal/femur ratio predict maximal running speed in cursorial mammals? *Journal of Zoology* **229**, 133–151 (1993).

[63] Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**, 83–97 (1955).

[64] Edmonds, J. & Karp, R. M. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM* **19**, 248–264 (1972).

[65] Tomizawa, N. On some techniques useful for solution of transportation network problems. *Networks* **1**, 173–194 (1971).