SCOPE: a normalization and copy number estimation method for single-cell DNA sequencing

Rujin Wang[1], Dan-Yu Lin[1,2], Yuchao Jiang[1,2,3,*]

[1] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.

[2] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA

[3] Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

* To whom correspondence should be addressed. Tel: (+1) 919-843-3656; Fax: (+1) 919-966-3804; Email: yuchaoj@email.unc.edu.

Abstract

Whole genome single-cell DNA sequencing (scDNA-seq) enables characterization of copy number profiles at the cellular level. This technology circumvents the averaging effects associated with bulk-tissue sequencing and increases resolution while decreasing ambiguity in tracking the evolutionary history of cancer. ScDNA-seq data is, however highly sparse and noisy due to the biases and artifacts that are introduced during the library preparation and sequencing procedure. Here, we propose SCOPE, a normalization and copy number estimation method for scDNA-seq data. The main features of SCOPE include: (i) a Poisson latent factor model for normalization, which borrows information across cells and regions to estimate bias, using negative control cells identified by cell-specific Gini coefficients; (ii) modeling of GC content bias using an expectation-maximization algorithm embedded in the normalization step, which accounts for the aberrant copy number changes that deviate from the null distributions; and (iii) a cross-sample segmentation procedure to identify breakpoints that are shared across cells from the same subclone. We evaluate SCOPE on a diverse set of scDNA-seq data in cancer genomics, using array-based calls of purified bulk samples as gold standards and whole-exome sequencing and single-cell RNA sequencing as orthogonal validations; we find that, compared to existing methods, SCOPE offers more accurate copy number estimates. Further, we demonstrate SCOPE on three recently released scDNA-seq datasets by 10X Genomics: we show that it can reliably recover 1% cancer cell spike-ins from a background of normal cells and that it successfully reconstructs cancer subclonal structure from ~10,000 breast cancer cells.

Key words: single-cell DNA sequencing, copy number variation, normalization, tumor heterogeneity, cancer genomics.

Introduction

Copy number variations (CNV) refer to duplications and deletions that lead to gains or losses of large segments of the chromosomes. CNVs are an abundant source of variations[1] and have been associated with diseases,[2] such as HIV acquisition and progression,[3] autism,[4] schizophrenia,[5] and systemic autoimmune diseases.[6,7] In cancer, somatic CNVs, also referred to as copy number aberrations (CNAs), are prevailing, as shown by the Cancer Genome Atlas[8] and the International Cancer Genome Consortium;[9] these CNAs have been associated with cancer progression and metastases.[10,11] It is, therefore, important to detect CNVs with high sensitivity and specificity.

In order to retain statistical power for association testing, CNV profiles across a large population of samples are required.[12] Recent advances in next-generation sequencing enable genome-wide CNV detection in a high-throughput manner. However, biases and artifacts are introduced during the library preparation and sequencing step, making data normalization crucial for accurate CNV profiling. Several algorithms have been developed to remove experimental noise from GC content bias,[13] sequencing depth,[14] mappability,[15] amplification efficiency,[16] and other latent systematic artifacts.[17,18] Yet profiling of somatic CNVs in cancer remains challenging, due to the heterogeneous nature of the tumor samples—the observed copy number signals from bulk DNA sequencing (DNA-seq) are averaged and attenuated across multiple genotypically distinct cancer subpopulations.[19-21] Tumor purity further dampens the CNV signals, and overall tumor ploidy leads to genome-wide gains or losses of chromosomal copies, all of which make statistical estimation and inference less tractable.[19,20]

Whole genome single-cell DNA-seq (scDNA-seq) enables the characterization of copy number profiles at the cellular level without the cell subpopulation confounding. This circumvents the averaging effects associated with bulk-tissue DNA-seq, increasing resolution while decreasing ambiguity in deconvolving cancer subclones, thus elucidating cancer evolutionary history (Supplementary Table S1).[22-24] Conventional whole-genome amplification methods for scDNA-seq include degenerate oligonucleotide primed PCR (DOP-PCR), multiple displacement amplification (MDA), and multiple annealing and looping-based amplification cycles (MALBAC).[25] MDA[26] lacks uniformity, yet has a low sequencing error rate, and it is thus better suited for calling point mutations; DOP-PCR[27]

57 and MALBAC,[28] on the other hand, have higher uniformity and have been successfully
58 used for calling CNVs;[19-21] see Supplementary Table S2 on details of the perspective
59 amplification methods. Recently, 10X Genomics released the Single-Cell CNV Solution
60 (https://www.10xgenomics.com/solutions/single-cell-cnv), which automatically uses
61 microfluidic droplets to barcode cells and performs library construction. In the first step,
62 the single cells are encapsulated into hydrogel cell beads and then lysed. This is followed
63 by a second step of amplification and cell bead partitioning with 10X barcoded gel beads.
64 The amplified and barcoded fragments are then pooled and converted to sequencing
65 libraries. This significantly increases throughput and offers great potential for profiling
66 CNVs across a large population of cells.

67 However, scDNA-seq data across all existing platforms and technologies is sparse,
68 noisy, and highly variable, even within a homogeneous cell population. The extremely
69 shallow and highly non-uniform depth of coverage, which is caused by the non-linear
70 amplification and significant dropout events during the library preparation and sequencing
71 step,[25,29] makes detecting CNVs by scDNA-seq challenging. Furthermore, the cancer
72 genomes undergo large chromosome or chromosome-arm level deletions or duplications,
73 as well as changes in cellular ploidy, which lead to recurrent and frequent CNAs that
74 disrupt a large proportion of the genome across multiple samples.[30] Existing methods for
75 scDNA-seq either build an optimized normal/reference set for normalization[31] or adopt a
76 cell-specific normalization procedure for removing systematic biases (due to, e.g., GC
77 content), followed by a *post hoc* procedure ploidy estimation and adjustment.[32] These
78 methods, which are adapted from those that are developed for bulk DNA-seq data, do not
79 address the challenges and complexities caused by aberrant copy number changes or
80 the complicating factor of tumor ploidy, and they cannot adequately and correctly remove
81 the artifacts, for they make the invalid assumption that all read counts are sampled from
82 diploid regions under the null when estimating the noise terms. The recurrent CNV signals
83 can either be accidentally removed during the normalization step or bias the correction
84 step.

85 To meet the widespread demand for CNV detection with single-cell resolution, we
86 propose a new statistical and computational framework, SCOPE, for *S*ingle-cell *COP*y
87 number *E*stimation. The distinguishing features of SCOPE include: (i) utilization of cell-

specific Gini coefficients for quality control and for identification of normal/diploid cells, which are then used as negative control samples in a Poisson latent factor model for read depth normalization; (ii) modeling of GC content bias using an expectation-maximization (EM) algorithm embedded in the Poisson regression models to account for the different copy number states along the genome; and (iii) a cross-sample iterative segmentation procedure to identify breakpoints that are shared across cells with the same genetic background. We evaluate the performance of SCOPE on real scDNA-seq datasets from several cancer genomic studies, as summarized in Supplementary Table S1. Compared to existing methods, SCOPE is shown to more accurately estimate subclonal copy number aberrations and to have higher correlation with array-based copy number profiles of purified bulk samples from the same breast cancer patient. We further show that the copy number profiles by scDNA-seq are also be well recapitulated, although at low resolution, by whole-exome sequencing (WES) and single-cell RNA-sequencing (scRNA-seq). We finally demonstrate SCOPE on the recently released scDNA-seq data that was produced using the 10X Genomics single-cell CNV pipeline, showing that it can reliably recover 1% of the cancer cell spike-ins from a background of normal cells and successfully reconstruct cancer subclonal structure across 10,000 breast cancer cells. SCOPE is compiled as an open-source R package available at https://github.com/rujinwang/SCOPE.

Material and Methods

Methods overview

An overview of the SCOPE workflow is shown in Figure 1. SCOPE takes as input the mapped reads from assembled BAM files,[33] which are pre-processed using the same bioinformatic pipeline (see Supplementary Note S1 for details). SCOPE then generates consecutive bins along the genome and computes the cell-by-bin read depth matrix, as well as the mappability and GC content for each bin. For data normalization, SCOPE adopts a Poisson latent factor model with an embedded EM algorithm to capture both cell- and bin-specific systemic biases and artifacts, as well as GC content bias. SCOPE also incorporates a cross-sample segmentation procedure, enabling shared breakpoints

across cells from the same cancer subclone. It then outputs integer-valued copy numbers, allowing direct ploidy estimation without the need for *post hoc* adjustment.

In the following sections, we start with an overview of the sample-specific data normalization and segmentation procedure by Ginkgo[32] and SCNV[31] and the cross-sample Poisson latent factor model by CODEX[16] and CODEX2.[18] We then discuss the limitations of the existing models and the reason that existing methods cannot adequately capture the biases and artifacts, explaining why careful statistical modeling with further adaptations is warranted for detecting CNVs by scDNA-seq data. Finally, we describe the model for SCOPE, leaving algorithmic details to the supplements.

Review of existing methods

Several methods have been developed for CNV detection by scDNA-seq, including Ginkgo[32] and SCNV.[31] To compute depth of coverage, both methods aggregate read-depth information into variable-length bins. To generate boundaries for the bins, *in silico* generated reads at every position of the genome are mapped back to the genome, and the bin sizes are chosen such that each bin has the same number of uniquely mapped reads. This step ensures that the bins have approximately the same number of mapped reads on average. To correct for GC content bias, both Ginkgo[32] and SCNV[31] perform normalization within each individual cell separately. Ginkgo[32] employs a locally weighted linear regression (LOWESS curve) to fit the relationship between GC content and log-normalized bin counts;[32] similarly, SCNV adjusts for GC content with LOWESS regression.[31] Both methods assume that all reads are from the diploid regions, and both control for the GC biases by modeling the unimodal relationship between GC content and log-scaled read counts. This ignores the multiplicative effects contributed by the different sequential copy-number states in each single cell.

In addition, because depth of coverage within each cell is normalized against its own coverage baseline, each cell has a normalized value with mean zero and thus completely masks the cellular ploidy. To solve this issue, both Ginkgo and SCNV adopt a *post hoc* ploidy estimation procedure and further scale the normalized results by the estimated ploidy to reflect the true copy numbers. In more details, they estimate the ploidy by minimizing a cost function that measures the difference between the scaled copy

number (normalized copy number multiplied by the ploidy) and its integer rounded copy number. However, the copy numbers and the ploidy, which is the genome-wide average of the copy numbers, are interrelated—the normalized copy number depends on the true underlying ploidy, and the two-step approach estimates ploidy based on the normalization results. This leads to a chicken-and-egg problem that is cyclic.

In summary, in existing methods, the recurrent CNV signals, as well as the complicating factor such as ploidy, can either be accidentally removed during the normalization step, which needs a second-step recovery, or bias the correction for known biases, such as those due to different GC content. For illustration, Figure 2 shows fitting of GC content bias in two ways: (i) assuming all bins are from the null region and fitting a non-parametric function using read depths across all bins, a method that, to our best knowledge, is adopted by all existing methods; and (ii) adopting an EM algorithm with the missing data being the carrier status for each bin, which is a simplified implementation of SCOPE. Three cells are chosen as examples, including a diploid, a hypodiploid, and a hyperdiploid cell. For the diploid cell, the fitting by SCOPE is the same as that by the all-null fitting, as expected. For hypodiploid and hyperdiploid cells, however, sequential multiplicative increments in the observed $f(GC)$—the GC content biases—are observed due to prominent copy number changes along the genome. The “all-null” fit is biased by the global genomics structures due to deletions and duplications, but with the embedded EM algorithm, SCOPE is able to successfully identify the true null regions and correctly estimate the GC bias term (Figure 2, Supplementary Figure S1).

After proper data normalization, segmentation is performed to return regions with homogeneous copy number profiles. Existing methods adopt segmentation procedures based on either circular binary segmentation (CBS) or hidden Markov model (HMM),[11] and they segment each cell separately.[31,32] Specifically, Ginkgo[32], which adopts the CBS[34] algorithm, provides either independent segmentation, where cell-specific normalized read counts are segmented, or segmentation with reference samples, which calibrate segmentation against the cells with the most uniform coverage and the highest quality. SCNV[31] relies on a set of normal cells to serve as the composite control and adapts the model proposed in Shen and Zhang[35] designed for bulk DNA-seq for bin-free segmentation. Both of these methods lack the ability to construct CNV profiles by

integrating shared cellular breakpoints from the same genetic background. SCOPE adopts a cross-sample segmentation procedure that simultaneously estimates changepoints that are shared across cells. This is extremely important in the single-cell settings, where multiple cells from the same subclone share the same genetic background and thus the same breakpoints for CNVs. The lower panels in Figure 2 show the cross-sample segmentation results by SCOPE across three cells, where the predicted ploidy by SCOPE is in concordance with that from single-cell flow sorting.

SCOPE model for data normalization

The extremely low depth of sequencing coverage, the nonlinear cell-specific biases due to amplification and sequencing, as well as other low-rank systemic artifacts that cannot be directly measured or quantified,[36] make the *sample-specific* bias correction procedure inadequate to successfully and unbiasedly capture all noise terms. *Cross-sample* normalization, on the other hand, relies on multiple samples processed in the same experiment run, and it borrows information both across regions and across samples to estimate the bias terms. This normalization strategy based on matrix factorization has been applied in the bulk setting for different types of omics data[16,18,37,38] to adjust for GC content bias and other latent artifacts. Furthermore, RUV[37] and CODEX2[18] extensively utilize negative control genes and negative control samples to estimate latent factors. For CNV detection, it has been shown that this increases CNV signal-to-noise ratio and achieves high sensitivity for both rare and common variants. The experience we have gained over a decade of next-generation sequencing data analysis collectively as a community motivates many of the ingredients of our new proposed approach for single-cell DNA sequencing.

SCOPE is based on a Poisson latent factor model[16,18] for count-based read depth normalization but is completely standalone and specifically adapted for the single-cell setting. Furthermore, to estimate CNVs by bulk DNA-seq across samples and to estimate CNVs by scDNA-seq across cells from the same sample are two different problems – in cancer genomics, the former profiles inter-tumor heterogeneity across patients, while the latter profiles intra-tumor heterogeneity looking at single cells within a patient. The key innovation in SCOPE is the way it integrates both null and non-null regions, in a genome-

wide fashion, for unbiased estimation of both GC content bias and latent factors for cell- and position-specific background correction.

Specifically, let $Y = \{Y_{ij}; i = 1, \dots m; j = 1, \dots n\}$ be the raw read count matrix, where Y_{ij} is the read depth for cell $j \in \{1, \dots, n\}$ and bin $i \in \{1, \dots, m\}$. SCOPE assumes that

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}),$$

$$\lambda_{ij} = N_j \beta_i f_j(GC_i) \alpha_{ij} \exp\left(\sum_{k=1}^{K} g_{ik} h_{jk}\right).$$

N_j is a cell-specific library size factor, which can be globally estimated as the median ratio;[39] β_i reflects bias due to bin-specific length, capture and amplification efficiency; $f_j(GC_i)$ is a sample-specific non-parametric function to capture the GC content bias, g_{ik} and h_{jk} $(1 \le k \le K)$ are the kth bin- and cell-specific latent factors with orthogonality restraints, which force identifiability; and α_{ij} specifies the sequential multiplicative increment due to different copy number states. α_{ij} is discretized to fit the single-cell setting with integer-valued copy numbers and to ensure identifiability with T_j different copy number states within a specific cell j $(1 \le j \le n)$:

$$\alpha_{ij} = \begin{cases} 1/2 & \text{with probability } \pi_1^{(j)} \\ 2/2 & \text{with probability } \pi_2^{(j)} \\ \vdots & \vdots \\ T_j/2 & \text{with probability } \pi_{T_j}^{(j)} \end{cases},$$

where $\left\{\frac{1}{2}, \frac{2}{2}, \dots, \frac{T_j}{2}\right\}$ corresponds to the linear increments in read depths across all copy number states in cell j, with corresponding incidence rates $\left\{\pi_1^{(j)}, \pi_2^{(j)}, \dots, \pi_{T_j}^{(j)}\right\}$ and $\sum_{t=1}^{T_j} \pi_t^{(j)} = 1$. We denote $\pi_1^{(j)}$ as the incident rate for heterozygous deletion, $\pi_2^{(j)}$ as the probability of a bin residing in a null region, and $\pi_3^{(j)}, \cdots, \pi_{T_j}^{(j)}$ as the incident rates for amplifications in cell j. The optimal number of copy number groups T_j is determined by the Bayesian information criterion (BIC) for each cell j separately (Supplementary Figure S1). By introducing the α_{ij} term, we aim to unmask biases based on a mixture of Poisson distributions with observed-data likelihood:

$$\phi(Y_{ij}) = \sum_{t=1}^{T_j} \pi_t^{(j)} \phi_t(Y_{ij}) = \sum_{t=1}^{T_j} \pi_t^{(j)} \text{pPoisson}\left(Y_{ij}; N_j \beta_i f_j(GC_i) \frac{t}{2} \exp\left(\sum_{k=1}^{K} g_{ik} h_{jk}\right)\right).$$

For each individual cell j, we adopt an EM algorithm within the iterative estimation procedure, where the missing data indicates carrier status:

$$Z_{it}^{(j)} = \begin{cases} 1 & \text{if bin } i \text{ of cell } j \text{ is from copy number group } t, \\ 0 & \text{otherwise.} \end{cases}$$

For parameter estimation, SCOPE adopts an iterative procedure, where EM is embedded as one step to estimate the GC content bias. For EM initializations, SCOPE utilizes ploidy estimates from a first-pass normalization run to ensure fast convergence and to avoid local optima (Supplementary Figure S2). See algorithmic details in the Supplementary Note S2. For estimation of K, the number of latent factors, SCOPE, by its default, adopts BIC as a model selection metric. Notably, we demonstrate that with the EM algorithm to account for the local genomic contexts and the use of normal cells to estimate bin-specific noise terms, the procedure outlined above by SCOPE is robust to the different choice of K (see under Discussion for more details).

The histograms in Figure 3 show the normalization results for a breast cancer dataset across four methods – CODEX2, Ginkgo, Ginkgo with *post hoc* adjustment by the estimated ploidy, and SCOPE. We observe that neither CODEX2 nor Ginkgo attains unbiased separation of different copy number states. To solve this issue and to obtain integer-valued copy number estimates, Ginkgo[32] adopts a *post hoc* ploidy estimation procedure, followed by a cell-specific scaling step; these ensure that the average copy number across all bins matches the estimated ploidy, an adjustment that allows Ginkgo to achieve better separation of CNV signals. SCOPE, on the other hand, retains precise segregation of CNV signals from the biological and technical noise, and distributions of the normalized z-score are centered at the expected integer copy-number values. Furthermore, SCOPE offers direct ploidy estimates, based on the estimated copy numbers along the genome, completely off the shelf.

Identification of negative control cells

In most, if not all, single-cell cancer genomics studies, diploid cells are inevitably picked up for sequencing from adjacent normal tissues and they can thus serve as normal controls for read depth normalization. However, not all platforms/experiments allow or adopt flow-sorting based techniques before scDNA-seq, and thus cell ploidy and case-control labeling are not always readily available. To solve this issue, SCOPE opts to *in*

silico identify normal cells as controls based on the scDNA-seq data. Specifically, for each cell, we calculate its Gini coefficient as two times the area between the diagonal line and the Lorenz curve, shown in Supplementary Figure S3. Equivalently, the Gini coefficient of cell j can be calculated as

$$Gini_j = \frac{\sum_{i=1}^{m}\sum_{k=1}^{m}\left|Y_{ij} - Y_{kj}\right|}{2m\sum_{i=1}^{m} Y_{ij}},$$

which serves as a robust and scale-independent measurement of coverage uniformity. As such, cell-specific Gini coefficients can be used as good proxies for identifying cell outliers, which have extreme coverage distribution due to failed library preparation, and for indexing normal cells out of the entire cell population.

The utility of Gini coefficients is supported by empirical evidence showing that diploid, hyperdiploid, and hypodiploid cells, categorized by single-cell flow sorting, can be perfectly classified using the estimated Gini indices (Supplementary Figure S4). This suggests that the Gini coefficient is an effective metric, with which to identify negative control samples. In practice, SCOPE does not require identification of all diploid cells from the cell population but requires around 10 to 20 cells to serve as normal controls; see under Discussion for details on choice of Gini coefficient threshold. The identified normal cells are then used as normal controls in the cross-sample normalization step to estimate bin-specific noise terms that are free of being biased by the CNV signals. Refer to Supplementary Note S2 for details.

Detecting simultaneous changepoints across samples

For segmentation, SCOPE adopts a cross-sample Poisson likelihood-based recursive segmentation procedure to identify changepoints/breakpoints that are shared across cells from the same subclone. Specifically, for cell j, let $Y_{sj}, \dots, Y_{tj}$ and $\lambda_{sj}^0, \dots, \lambda_{tj}^0$ be the observed and normalized read depth from a region spanning bin s to bin t, where $\lambda_{ij}^0 = N_j\beta_i f_j(GC_i)\exp\left(\sum_{k=1}^{K} g_{ik}h_{jk}\right)$ is the "control" depth of coverage under the null, i.e., the coverage we expect to see if there is no CNV. We further denote $Y_{s:t,j} = \sum_{i=s}^{t} Y_{ij}$ and $\lambda_{s:t,j}^0 = \sum_{i=s}^{t}\lambda_{ij}^0$. The scan statistic for one cell is $\max_{s,t} U_j(s,t)$, where

$$U_j(s,t) = \sup_{\mu_{s:t,j}} \left(\log \left(\frac{\mu_{s:t,j}^{Y_{s:t,j}} \exp(-\mu_{s:t,j})}{\lambda_{s:t,j}^{0\,Y_{s:t,j}} \exp(-\lambda_{s:t,j}^0)} \right) \right) = Y_{s:t,j} \log \left(\frac{Y_{s:t,j}}{\lambda_{s:t,j}^0} \right) - \left(Y_{s:t,j} - \lambda_{s:t,j}^0 \right).$$

$U_j(s,t)$ is a generalized likelihood ratio test statistic, derived from the null model $Y_{s:t,j} \sim \mathrm{Poisson}(\lambda_{s:t,j}^0)$ against the alternative $Y_{s:t,j} \sim \mathrm{Poisson}(\mu_{s:t,j})$. [40] For simultaneous changepoints detection across all cells, the scan statistic is $\max_{s,t} Z_{s,t}$, where

$$Z_{s,t} = \sum_{j=1}^{n} U_j(s,t) = \sum_{j=1}^{n} \left[Y_{s:t,j} \log \left(\frac{Y_{s:t,j}}{\lambda_{s:t,j}^0} \right) - \left(Y_{s:t,j} - \lambda_{s:t,j}^0 \right) \right].$$

SCOPE performs CBS[34] given $Z_{s,t}$ as the scan statistic and uses a cross-sample modified BIC (mBIC)[41] as the stopping rule:

$$\mathrm{mBIC}(p) = \log \left(\frac{L_\tau}{L_0} \right) - \frac{P}{2} \log \frac{2 \log(L_\tau / L_0)}{P} - \frac{P}{2} - \log \binom{m}{p} - p(\kappa_1 - \kappa_2)$$
$$- \sum_{\rho=1}^{p} \log \left(\sum_{j\text{th carrier}} \hat{\delta}_{\rho,j}^2 \right) + P \log \pi + (np - P) \log(1 - \pi),$$

where $1 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_p < \tau_{p+1} = m$ denote the changepoints that are shared across cells; $\log(L_\tau / L_0)$ is the generalized log-likelihood ratio for the alternative model, which has p changepoints, against the null model, which has no changepoints; P denotes the total number of shift parameters (δ), where $\delta_{\rho,j} = \lambda_{\rho,j}^0 - \lambda_{\rho-1,j}^0$ for cell j with changepoint ρ $(1 \le \rho \le p)$; and κ_1 and κ_2 are pre-defined numerical constants. When the carrier probability π is not known a priori, it is estimated empirically by

$$\hat{\pi} = P/(np).$$

The optimal number of changepoints is determined via $\max_p \mathrm{mBIC}(p)$. For more details on the interpretation of the terms in mBIC, see Zhang and Siegmund.[41]

After segmentation, SCOPE reports integer-valued copy numbers and allows direct ploidy estimation, which is calculated as the weighted average of the estimated copy number across the genome. For further downstream analysis, SCOPE also includes the option to cluster cells based on the matrix of normalized z-scores, estimated copy numbers, or estimated changepoints, which returns clusters of cells from the same subclones with the same mutational profiles. As a final step, cells from the same subclone can be combined together to generate pseudo-bulk samples that are presumably homogeneous. Another normalization on the *in silico* generated pseudo-bulk samples

returns copy number profiles with higher resolution. More details are described in the Methods section. The pseudo-bulk samples also have the potential for somatic point mutation profiling, with the sequencing sparsity averaged out. Refer to the Discussion section for more details.

Availability of data and material

SCOPE is an open-source R package available at https://github.com/rujinwang/SCOPE with license GPL-2.0. Breast cancer scDNA-seq data are available from the NCBI Sequence Read Archive with accession number SRA018951 and SRP114962; breast cancer aCGH benchmark data are available from the NCBI Gene Expression Omnibus with accession number GSE16607. ScDNA-seq data from 10X Genomics are available at https://support.10xgenomics.com/single-cell-dna/datasets. A detailed data summary is provided in Supplementary Table S1.

Results

We first set out to compare the biases and differences in depth of coverage and coverage uniformity between the conventional whole-genome amplification methods and the 10X Genomics pipeline. Compared to protocols that use conventional methods for cell isolation and whole-genome amplification (WGA), 10X Genomics produces data with both higher throughput (by the number of cells captured and sequenced) and yet significantly lower sequencing depth (Supplementary Table S1). Conventional WGA data achieves a mean total read count of 5.3 million per cell, while 10X Genomics data achieves a mean total read count of 0.4 million per cell (Supplementary Figure S5A). To assess coverage uniformity, we compared the cell-specific median absolute deviation (MAD) of all pairwise differences in read counts between neighboring bins, with a default bin size of 500kb for all platforms and datasets (refer to the Discussion section on choice of bin size). We pre-normalized the read count data by a cell-specific factor for library size difference and a bin-specific factor for baseline coverage difference, then calculated the MADs across all cells; the MADs are resilient to outliers, as the transitions between copy number states are relatively infrequent along the genome. Strikingly, our results show that, even though cells from the 10X Genomics pipeline are sequenced at a much shallower depth, the

351 coverage uniformity metrics show trivial differences between data from the 10X Genomics
352 and the conventional methods, except that the former has a higher number of cell outliers
353 (Supplementary Figure S5B).
354
355 ***Analysis of scDNA-seq data of breast cancer patients with aCGH for validation***
356 We first demonstrated SCOPE on the scDNA-seq data of two breast cancer patients, T10
357 and T16, from Navin et al.,[22] who sequenced 100 flow-sorted single cells from each
358 patient. Fluorescence activated cell sorting (FACS) analysis of the single cells showed
359 different distributions of ploidy: hyperdiploid, hypodiploid, and diploid.[22] Furthermore,
360 array comparative genomic hybridization (aCGH) was applied to the same tumor sectors
361 from both patients, and this returned array-based copy number profiles of flow-sorted cell
362 populations.[42] Cell ploidy by FACS, copy number profiles of purified bulk samples by
363 aCGH, and initially reported copy number profiles of single cells estimated by scDNA-seq
364 all suggested that patient T10 had three distinct clonal subpopulations, indicating a
365 polygenomic tumor. In patient T16, a relatively homogeneous cancer cell population was
366 identified in both primary tumor and metastasis of patient T16, indicating a monogenomic
367 tumor.[22] We sought to apply SCOPE to this dataset to replicate previous findings, and,
368 further, used copy number calls by aCGH as orthogonal validations.
369 Figure 4 gives heatmaps of genome-wide estimated copy numbers across all cells
370 from T10 and T16. SCOPE identified two subpopulations of hyperdiploid cancer cells,
371 one subpopulation of hypodiploid cancer cells, and a normal cell subpopulation within
372 T10, which is consistent with the previous report. For T16, SCOPE returned two cancer
373 cell subclones, one for the hyperdiploid subpopulation from the primary tumor and the
374 other for the hyperdiploid subpopulation from the metastasis. Upon careful inspection of
375 the consensus copy number profiles of these two subpopulations, we find them highly
376 similar, indicating that the same subclone of origin from the primary tumor survived the
377 selection bottleneck and led to relapse. Notably, SCOPE also identified three
378 pseudodiploid cells (red annotation bars in Figure 4B), in concordance with the previous
379 report. Overall, we showed that SCOPE was able to reproduce subclonal structures that
380 are consistent with previous results. The normalization results by SCOPE attain better
381 separation of signal from noise than other existing methods, as can be seen from the

histograms of the normalized z-scores shown in Figure 3 and Supplementary Figure S6 (for T10 and T16, respectively). If we only focus on normal cells, there is an even clearer separation of normalized copy numbers, as shown by the disappearance of the spike at two from the normal cells (Supplementary Figure S7). In addition, the ploidy estimates by SCOPE are highly concordant with those from previous reports based on single-cell flow sorting (Supplementary Figure S8).

To further assess the performance of SCOPE and to benchmark against existing methods, we adopted CNV calls from aCGH of purified bulk samples[42] as gold standards (Supplementary Table S3). FACS revealed three cancer subpopulations (aneuploid 1, aneuploid 2, and hypodiploid), which is, again, in concordance with SCOPE's results. This was followed by aCGH for copy number profiling of the purified bulk samples. Relative copy numbers from aCGH, scDNA-seq by SCOPE, and scDNA-seq by Ginkgo are shown in Figure 5A-C across the three cancer subpopulations: for aCGH, the subpopulations were inferred by the initial step of flow sorting; for SCOPE and Ginkgo, the subpopulations were inferred and aggregated based on the estimated single-cell copy number profiles. Notably, array-based intensity measurements were normalized against a sample-specific baseline, producing copy number signals that have a mean of one for all three subpopulations. Therefore, relative copy numbers, which are absolute copy numbers adjusted by ploidy (shown as horizontal solid lines in Figure 5A-C for diploid regions) were used by Ginkgo and SCOPE to mask ploidy for comparison against aCGH calls. Using spearman correlation and root mean squared error (RMSE) as performance metrics, SCOPE outperforms Ginkgo by the aCGH gold standards.

Analysis of scDNA-seq data of triple-negative breast cancer patients with paired WES and scRNA-seq

We further applied SCOPE to scDNA-seq data of temporally separated tumor resections from triple negative breast cancer patients with chemotherapy treatment.[24] Specifically, scDNA-seq was performed on matched pre-treatment, mid-treatment, and post-treatment tumor samples from the same patient. We applied SCOPE to four "clonal extinction" patients, where tumor cells were previously reported to exist only in the pre-treatment samples (Supplementary Table S1). For example, 92 cells from patient KTN302 were

413 sequenced at two timepoints, pre- and mid-treatment. In this case, SCOPE successfully
414 detected two subclones of aneuploid cells in the pre-treatment tumors and found that all
415 the cells from the mid-treatment group were normal, as shown in Figure 6. The consensus
416 copy number profiles of the two subpopulations indicate a large number of shared CNAs
417 between the two inferred subclones; this is, likely due to a punctuated copy number
418 evolution,[43] which produces the majority of CNAs in the early stages of tumor evolution.
419 Results for patients KTN126 and KTN129 are included in Supplementary Figure S10.

420 We used bulk-tissue WES and single-cell RNA-seq data to validate the copy
421 number profiles returned by SCOPE. For the data from patient KTN302, WES and
422 scRNA-seq were performed on the matched normal (blood), pre-treatment, mid-treatment,
423 and post-treatment bulk samples (Supplementary Table S1) and were used to profile
424 CNAs separately for validation. For bulk WES, a modified CBS algorithm based on
425 seqCBS[35] was utilized for the paired tumor-normal experimental design, which returns
426 relative copy numbers. Despite signal attenuation due to intra-tumor heterogeneity and
427 normal cell contamination, we observe large deletions on chromosome 5, 10, and 18, and
428 a distinct copy-number amplification on chromosome 8, in accordance with the profiles
429 returned by scDNA-seq (Figure 6, Supplementary Figure S10). For scRNA-seq, we used
430 the InferCNV toolkit[44] and employed a sliding-window approach of 50 genes[45] to infer
431 CNVs, using as input the transcript per million (TPM) matrix, returned by SALMON
432 (Supplementary Figure S9).[46] In spite of the low resolution, we observed deletions on
433 chromosome 5, 10, and 18, as well as duplications on chromosome 8, in accordance with
434 WES and scDNA-seq calls (Figure 6, Supplementary Figure S10).

435 To quantitatively assay the quality of the call set produced by SCOPE, we plotted
436 the relative copy number estimates by WES and scDNA-seq for the inferred amplified,
437 deleted, and copy-number-neutral regions by SCOPE. Although the copy number signals
438 from WES and scRNA-seq are low-resolution and are attenuated due to normal cell
439 contaminations, the relative copy numbers from both WES and scRNA-seq are less than
440 one for deletion regions, but greater than one for putative amplifications, in agreement
441 with results from SCOPE (Figure 7, Supplementary Figure S11). Notably, in the WES and
442 scRNA-seq data, the relative copy numbers in the null regions are less than one due to

443 over-normalization of ploidy in hyperdiploid samples (Figure 7, Supplementary Figure
444 S11).
445
446 ***Analysis of scDNA-seq data of gastric cancer spike-ins and breast cancer***
447 ***dissections from the 10X Genomics***
448 We finally demonstrated SCOPE on three recently released scDNA-seq datasets from
449 the 10X Genomics Single-Cell CNV Solution pipeline. We first adopted two publicly
450 available spike-in datasets from the 10X Genomics website, where 1% and 10% MKN-45
451 gastric cancer cell lines are mixed with ~500 and ~1000 normal BJ fibroblast cells,
452 respectively (Supplementary Table S1). The second dataset consists of ~10,000 nuclei
453 extracted and sequenced from five different dissections of frozen breast tumor tissue from
454 a triple negative ductal carcinoma (Supplementary Table S1).
455 As a proof of concept, we started by applying SCOPE to the two 10X Genomics
456 spike-in datasets—1,055 single cells with a 1% spike-in of cancer cells and 462 single
457 cells with a 10% spike-in of cancer cells were sequenced. We adopted stringent quality
458 control procedures to remove cells with extreme Gini coefficients and to remove bins that:
459 (i) have low mappability; (ii) reside in "blacklist" regions; and (iii) have extremely low read
460 depth across cells. For the two spike-in datasets, this resulted in 5,055 and 5,064 bins for
461 the copy number estimation, respectively. Despite the higher sparsity and lower
462 sequencing depth of this data, SCOPE successfully identified the cancer cell cluster from
463 the cluster of normal diploid cells, with heatmaps shown in Supplementary Figure S12.
464 Specifically, 11 and 36 cancer cells were identified from the 1% and 10% spike-in datasets
465 respectively, with estimated proportions as 1% and 8%. For data visualization, we
466 performed t-SNE projections[47] on the normalized z-scores, revealing that SCOPE
467 successfully recovers 1% and 10% non-diploid cancer cells from a mixture of diploid cells
468 in the background (Figure 8A-B). The colors in Figure 8A-B indicate the classification by
469 a threshold of 0.12 for the Gini coefficients. We see that while not all diploid cells are used
470 as negative control cells, there is a perfect separation between normal cells and cancer
471 cells; see under Discussion for details on setting the threshold of Gini coefficients.
472 We further applied SCOPE to a 10X Genomics scDNA-seq dataset of ~10,000
473 cells from five adjacent tumor dissections of a breast cancer patient. We followed the

474 same quality control procedure as previously described, which resulted in 4864, 5016,
475 5032, 5023, and 5026 bins from each dissection. SCOPE is then separately applied to
476 each dissection for identification of normal cells, read depth normalization, and copy
477 number estimation. The proportions of normal cells vary across the five dissections, with
478 mean cellular ploidy of sections A to E being 2.12, 3.13, 3.29, 3.36, and 3.53, respectively.
479 This indicates a gradient of normal cells contaminating the tumor (Supplementary Figure
480 S13). We further integrated the inferred copy number profiles by SCOPE across all tumor
481 cells from the five dissections and demonstrated that SCOPE was able to identify
482 subclonal structures. For example, the distinct duplication events on chromosome 3 and
483 chromosome 4 are mutually exclusive and mark a split in the tumor evolutionary history.
484 Hierarchical clustering based on the normalization result suggests that the different
485 cancer subclones consist of cancer cells from all sections. This indicates an early
486 branching evolutionary model, where copy number aberrations happen at a potentially
487 early stage in the disease advancement (Figure 8C).
488

489 **Discussion**

490 Here we propose SCOPE, a new method to remove technical noise and improve CNV
491 signal-to-noise ratio for scDNA-seq data. SCOPE includes a Poisson latent factor model
492 for data normalization and a non-parametric functional term for adjusting GC content bias
493 in a context-specific manner. The EM algorithm embedded in the Poisson log linear model
494 accounts for the different copy number states as it adjusts for GC content bias—a step
495 we show to be crucial for cancer genomic studies, where recurrent copy number changes
496 across the genome lead to significant deviations from the null. Capacity is increasing in
497 single-cell isolation and single-cell whole-genome sequencing, and there is an increasing
498 need to profile CNVs at the single-cell level as a non-negligible source of genetic variation.
499 We believe that SCOPE can be a useful tool for the genetics and genomics community,
500 for it lays the statistical foundation that will make robust and accurate single-cell CNV
501 profiling possible.

502 We demonstrated SCOPE on two breast cancer scDNA-seq datasets, where
503 SCOPE successfully parsed normal cells, exposed a mixture of tumor cells based on their
504 inferred copy number profiles, and unmasked the ploidy. We showed that this enabled

505 characterization of genetically distinct populations of tumor cells that were not evident in
506 bulk-tissue analysis, shedding light upon tumor heterogeneity at a much finer resolution.
507 We benchmarked SCOPE against other existing methods using array-based CNV calls
508 from purified bulk samples as the gold standard, and we showed that SCOPE
509 outperformed those methods. We further demonstrated SCOPE on scDNA-seq data from
510 the recently released Single Cell CNV Solution by 10X Genomics and showed that
511 SCOPE could accurately detect 1% of the gastric cancer cells from the background of
512 normal fibroblast cells and that it successfully reconstructed the subclonal structure
513 across 10,000 breast cancer cells.

514 To systematically investigate how varying parameters (e.g., bin size, threshold of
515 cell-specific Gini coefficients to identify normal controls, and the number of Poisson latent
516 factors) influence performance, we carry out additional evaluation and benchmark
517 analysis on the scDNA-seq data from the breast cancer patient T10, with the aCGH calls
518 as gold standard. Our results indicate that a bin size of 3Mb, compared to a bin size of
519 200Kb, leads to smoother normalization results yet meanwhile misses smaller CNVs due
520 to low resolution (Supplementary Figure S14). Meanwhile, bin sizes between 200Kb to
521 1Mb do not affect the performance of SCOPE, especially for chromosome-arm level
522 CNVs (Supplementary Figure S15). For an optimal choice of the Gini coefficient threshold,
523 we provided empirical evidence that SCOPE does not need to include all normal cells as
524 negative controls – 10 to 20 normal cells suffice to achieve accurate estimation
525 (Supplementary Figure S16). In addition, we show that SCOPE's performance is invariant
526 to the choice of K, the number of Poisson latent factors (Supplementary Figure S17),
527 since SCOPE adopts the EM algorithm to account for genomic contexts when estimating
528 GC content biases and uses only the normal cells to estimate the bin-specific noise terms.
529 As such, we show that SCOPE is robust to the choice of Gini coefficient threshold and
530 the number of Poisson latent factors. In summary, compared to existing methods, SCOPE
531 directly provides copy number and ploidy estimates, completely off the shelf and without
532 much need for manual tuning by the users.

533 Correct inference of tumor evolution history depends on accurate copy number
534 estimates, which in turn depend on effective normalization. While we show that data
535 normalization is a first-order concern, the cross-sample segmentation procedure by

536 SCOPE also returns changepoints that are shared across cells, a unique feature that is
537 necessary for the single-cell setting, where cells from the same genetic backgrounds
538 share the same breakpoints. SCOPE also enables hierarchical clustering and phylogeny
539 reconstruction, based on the normalized log-ratios, the estimated copy numbers, or the
540 estimated changepoints—data that is useful for further downstream analysis. To further
541 reduce the ambiguity in unmasking intra-tumor heterogeneity, future research may
542 include force-calling somatic point mutations from scDNA-seq and integrating both copy
543 number and point mutation profiles for reconstructing tumor evolutionary history.
544 A few recently developed methods for CNV detection use single-cell RNA
545 sequencing (scRNA-seq).[44,48] While these gene-expression based approaches have
546 been successfully applied to detect chromosome or chromosome-arm level CNAs,
547 multimodal data alignment between scDNA-seq and scRNA-seq has the potential to
548 increase the resolution in detecting CNAs in a larger cell population. This alignment could
549 also shed light upon the interplay between genomic and transcriptomic variations in
550 cancer, for it is still unclear how transcriptomic variation is modulated by genetic variation
551 and phylogenetic evolution in cancer. The joint analysis framework would allow
552 quantification of clonal differences in gene expression while accounting for DNA
553 confounding.
554
555 **Acknowledgements**
556 This work was supported by the National Institutes of Health (NIH) grant P01 CA142538
557 (to DYL and YJ), R35 GM118102 (to YJ), a developmental award from the UNC
558 Lineberger Comprehensive Cancer Center 2017T109 (to YJ), and a pilot award from the
559 UNC Computational Medicine Program (to YJ).
560
561 **Authors' Contributions**
562 YJ initiated and envisioned the study. RW and YJ formulated the model, developed and
563 implemented the algorithm. RW, DYL, and YJ performed data analysis. RW and YJ wrote
564 the manuscript, which was edited by DYL. All authors read and approved the final
565 manuscript.
566

Competing Interests

The authors declare no conflict of interest.

References

1. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81.

2. McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. Nat Genet 39, S37-42.

3. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307, 1434-1440.

4. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466, 368-372.

5. Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet 49, 27-35.

6. Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., et al. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80, 1037-1054.

7. Fanciulli, M., Norsworthy, P.J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J.M., Gough, S.C., de Smith, A., Blakemore, A.I., et al. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39, 721-723.

8. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113-1120.

9. International Cancer Genome, C., Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature 464, 993-998.

10. Jiang, Y., Qiu, Y., Minn, A.J., and Zhang, N.R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A 113, E5528-5537.

11. Urrutia, E., Chen, H., Zhou, Z., Zhang, N.R., and Jiang, Y. (2018). Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny. Bioinformatics 34, 2126-2128.

12. Wellcome Trust Case Control, C., Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713-720.

607 13. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-
608 throughput sequencing. Nucleic Acids Res 40, e72.

609 14. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage:
610 key considerations in genomic analyses. Nat Rev Genet 15, 121-132.

611 15. Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M., and Boeva, V. (2016). SV-Bay: structural
612 variant detection in cancer genomes using a Bayesian approach with correction for GC-content
613 and read mappability. Bioinformatics 32, 984-992.

614 16. Jiang, Y., Oldridge, D.A., Diskin, S.J., and Zhang, N.R. (2015). CODEX: a normalization and copy number
615 variation detection method for whole exome sequencing. Nucleic Acids Res 43, e39.

616 17. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E.,
617 McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping
618 of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91, 597-607.

619 18. Jiang, Y., Wang, R., Urrutia, E., Anastopoulos, I.N., Nathanson, K.L., and Zhang, N.R. (2018). CODEX2:
620 full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol
621 19, 202.

622 19. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C.,
623 Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in
624 human cancer. Nat Biotechnol 30, 413-421.

625 20. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C.
626 (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data.
627 Ann Oncol 26, 64-70.

628 21. Chen, H., Jiang, Y., Maxwell, K.N., Nathanson, K.L., and Zhang, N. (2017). Allele-Specific Copy Number
629 Estimation by Whole Exome Sequencing. Ann Appl Stat 11, 1169-1192.

630 22. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D.,
631 Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature 472, 90-
632 94.

633 23. Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang,
634 H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing.
635 Nature 512, 155-160.

636 24. Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N.E.
637 (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell
638 Sequencing. Cell 173, 879-893 e813.

639 25. Liu, J., Adhav, R., and Xu, X. (2017). Current Progresses of Single Cell DNA Sequencing in Breast Cancer
640 Research. Int J Biol Sci 13, 949-960.

641 26. Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et
642 al. (2002). Comprehensive human genome amplification using multiple displacement
643 amplification. Proc Natl Acad Sci U S A 99, 5261-5266.

644 27. Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D.,
645 Lakshmi, B., et al. (2012). Genome-wide copy number analysis of single cells. Nat Protoc 7, 1024-
646 1041.

647 28. Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and
648 copy-number variations of a single human cell. Science 338, 1622-1626.

649 29. Navin, N.E. (2014). Cancer genomics: one cell at a time. Genome Biol 15, 452.

650 30. Shlien, A., and Malkin, D. (2009). Copy number variations and cancer. Genome Med 1, 62.

651 31. Wang, X., Chen, H., and Zhang, N.R. (2018). DNA copy number profiling using single-cell sequencing.
652 Brief Bioinform 19, 731-736.

653 32. Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M., and Schatz, M.C.
654 (2015). Interactive analysis and assessment of single-cell copy-number variations. Nat Methods
655 12, 1058-1060.

656 33. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.,
657 and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and
658 SAMtools. Bioinformatics 25, 2078-2079.

659 34. Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the
660 analysis of array-based DNA copy number data. Biostatistics 5, 557-572.

661 35. Shen, J.J., and Zhang, N.R. (2012). Change-Point Model on Nonhomogeneous Poisson Processes with
662 Application in Copy Number Profiling by Next-Generation DNA Sequencing. Annals of Applied
663 Statistics 6, 476-496.

664 36. Huang, L., Ma, F., Chapman, A., Lu, S., and Xie, X.S. (2015). Single-Cell Whole-Genome Amplification
665 and Sequencing: Methodology and Applications. Annu Rev Genomics Hum Genet 16, 79-102.

666 37. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor
667 analysis of control genes or samples. Nat Biotechnol 32, 896-902.

668 38. Lee, S., Chugh, P.E., Shen, H., Eberle, R., and Dittmer, D.P. (2013). Poisson factor models with
669 applications to non-normalized microRNA profiling. Bioinformatics 29, 1105-1111.

670 39. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome
671 Biol 11, R106.

672 40. Zhang, N.R., Siegmund, D.O., Ji, H., and Li, J.Z. (2010). Detecting simultaneous changepoints in multiple
673 sequences. Biometrika 97, 631-645.

674 41. Zhang, N.R., and Siegmund, D.O. (2012). Model Selection for High-Dimensional, Multi-Sequence
675 Change-Point Problems. Stat Sinica 22, 1507-1538.

676 42. Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J.,
677 Grubor, V., et al. (2010). Inferring tumor progression from genomic heterogeneity. Genome Res
678 20, 68-80.

679 43. Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.C., Casasent, A., Waters, J., Zhang,
680 H., et al. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast
681 cancer. Nat Genet 48, 1119-1130.

682 44. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed,
683 B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral
684 heterogeneity in primary glioblastoma. Science 344, 1396-1401.

685 45. Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C.,
686 Mount, C., Filbin, M.G., et al. (2016). Single-cell RNA-seq supports a developmental hierarchy in
687 human oligodendroglioma. Nature 539, 309-313.

688 46. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-
689 aware quantification of transcript expression. Nat Methods 14, 417-419.

690 47. Maaten, L.v.d., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning
691 research 9, 2579-2605.

692 48. Fan, J., Lee, H.O., Lee, S., Ryu, D.E., Lee, S., Xue, C., Kim, S.J., Kim, K., Barkas, N., Park, P.J., et al. (2018).
693 Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-
694 seq data. Genome Res 28, 1217-1227.

695

696 **Figure Legends**

697 **Figure 1.** A flowchart outlining the procedures for profiling single-cell CNV by SCOPE.
698 Bioinformatic pre-processing returns assembled BAM files as input for computing depth
699 of coverage. Given the reference genome, mappability and GC content are calculated for
700 each bin. SCOPE then adopts a Poisson latent factor model with an embedded EM
701 algorithm for data normalization to remove biases due to sample-specific sequencing
702 depth, bin-specific amplification/mapping efficiency, GC content, as well as other
703 systemic latent artifacts. A cross-sample Poisson likelihood segmentation is performed to
704 return copy number profiles, which are further used to infer cancer subclones.

705

706 **Figure 2.** EM algorithm for correction of GC content bias. GC content bias $f(GC)$, EM
707 fitting, and cross-sample segmentation results of scDNA-seq data from breast cancer
708 patient T10 are shown across three cells: (A) a diploid cell; (B) a hypoploid cell with four
709 copy number states; (C) a hyperploid cell with five copy number states. Predicted ploidy,
710 taken as genome-wide average, agrees with previous FACS sorting results.

711

712 **Figure 3.** Normalization results for scDNA-seq data of ploygenomic tumor T10.[22]
713 Normalized *z*-scores are shown for (A) CODEX2 as $\log(Y/\hat{\lambda})$, (B) Ginkgo, which centers
714 bin counts of each cell at one and further fits a locally weighted linear regression (LOESS
715 curve) to adjust for GC content bias, (C) Ginkgo with *post hoc* ploidy adjustment, which
716 scales relative copy numbers by the optimal ploidy multiplier, and (D) SCOPE as $\log(Y/\hat{\lambda})$,
717 which automatically takes into account and estimates ploidy.ch

718

719 **Figure 4.** SCOPE successfully detects subclonal structures for triple negative breast
720 cancer samples. (A) Inferred copy number profiles of single cells from polygenomic tumor
721 T10. (B) Inferred copy number profiles of single cells from monogenomic tumor T16.
722 Upper heatmaps show post-segmentation copy number profiles of integer values.
723 Hierarchical clustering separates normal cells from tumor cells and reveals three cancer
724 subclones in T10 and two highly similar subclones in T16. Lower heatmaps and stairstep
725 plots show consensus copy-number profiles of all single cells from the same cancer
726 subclones.

727

728 **Figure 5.** Orthogonal validation of single-cell copy number profiles by aCGH of purified
729 bulk samples by FACS. FACS reveals (A) hyperdiploid 1, (B) hyperdiploid 2, and (C)
730 hypodiploid cancer subclones within breast cancer patient T10. This is followed by aCGH
731 on the sorted bulk samples to profile copy number aberrations, which are used as ground
732 truth. Relative copy number (CN) estimates from the three subpopulations are shown in
733 (A)-(C) for aCGH, SCOPE, and Ginkgo. (D) Spearman correlation and RMSE as
734 performance metrics comparing results by SCOPE and Ginkgo against the gold standard.
735 SCOPE achieves higher correlation and lower estimation errors.

736

737 **Figure 6.** Copy-number profiles of triple-negative breast cancer patients KTN302, inferred
738 by scDNA-seq, bulk WES, and scRNA-seq. Upper three panels show heatmap of inferred
739 copy-number profiles by SCOPE with cells clustered by hierarchical clustering, heatmap
740 and stairstep plot of consensus copy number profiles for each cell subpopulation (N for
741 normal cell, A, B, and C for cancer subclones). Lower two panels show copy number
742 profiles inferred by WES and scRNA-seq, which are further used to validate CNV calls by
743 SCOPE.

744

745 **Figure 7.** Orthogonal validations using copy number profiles by WES and scRNA-seq.
746 Each genomic segment is categorized as amplification (amp), deletion (del), or copy-
747 number neural (null) based on the copy number estimates returned by SCOPE. The
748 boxplot shows the estimated relative copy numbers by bulk WES and scRNA-seq, which
749 are higher in amplified regions and lower in deleted regions, compared to the null regions.
750 Due to low resolution and high signal-to-noise ratio by WES and scRNA-seq, there is no
751 perfect separation between the different copy number states. Furthermore, for scRNA-
752 seq and bulk WES, the relative copy numbers are estimated comparing to a sample-
753 specific baseline. For cases where the tumor cells are hyperdiploid/hypodiploid, the
754 baseline is shifted up/down making the true null regions have relative copy numbers
755 less/greater than one.

756

757 **Figure 8.** Visualization by t-SNE projections of normalized scDNA-seq data from the 10X
758 Genomics cancer spike-in datasets. (A) 1% and (B) 10% of gastric cancer cell lines were
759 mixed with normal fibroblast cell lines. After normalization, SCOPE successfully identified
760 the cancer cell population from the background of normal cells. (C) Heatmap of
761 normalization results by SCOPE on the scDNA-seq dataset of 10,000 breast cancer
762 nuclei. Cell clustering reveals cancer subclonal structure, not previously detectable by
763 bulk DNA-seq.

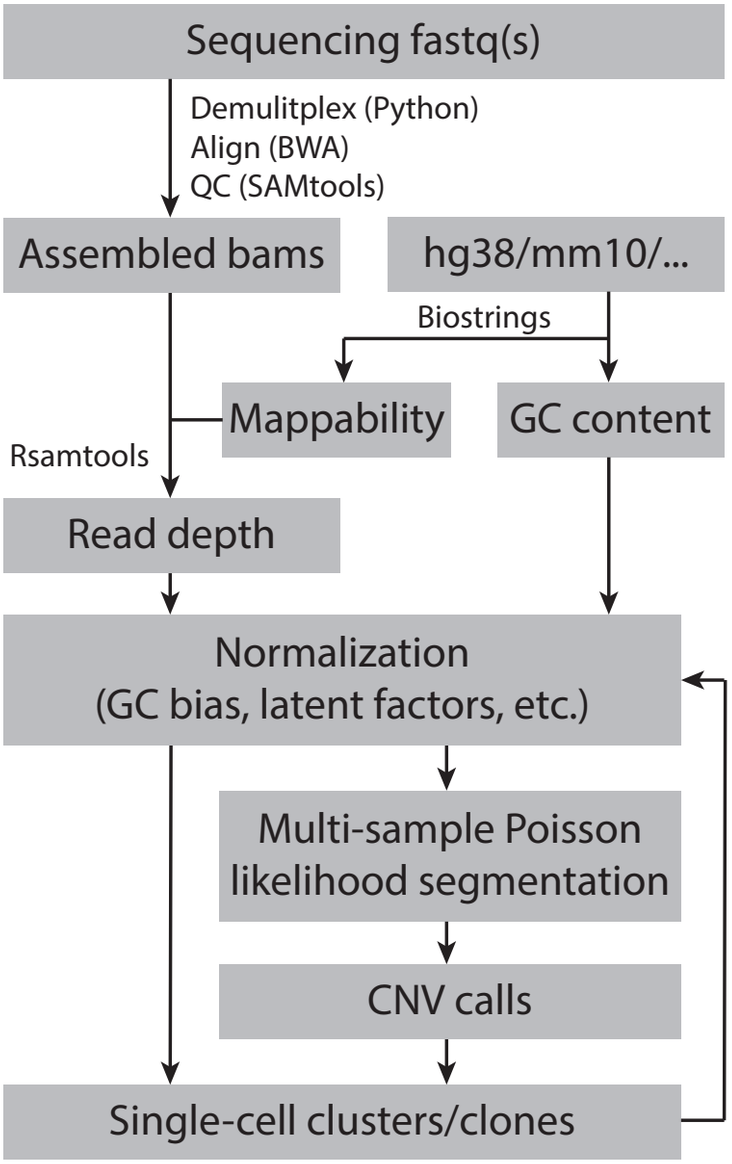
Sequencing fastq(s)
Demulitplex (Python)
Align (BWA)
QC (SAMtools)
Assembled bams
hg38/mm10/...
Biostrings
Mappability
GC content
Rsamtools
Read depth
Normalization
(GC bias, latent factors, etc.)
Multi-sample Poisson
likelihood segmentation
CNV calls
Single-cell clusters/clones

Figure 1

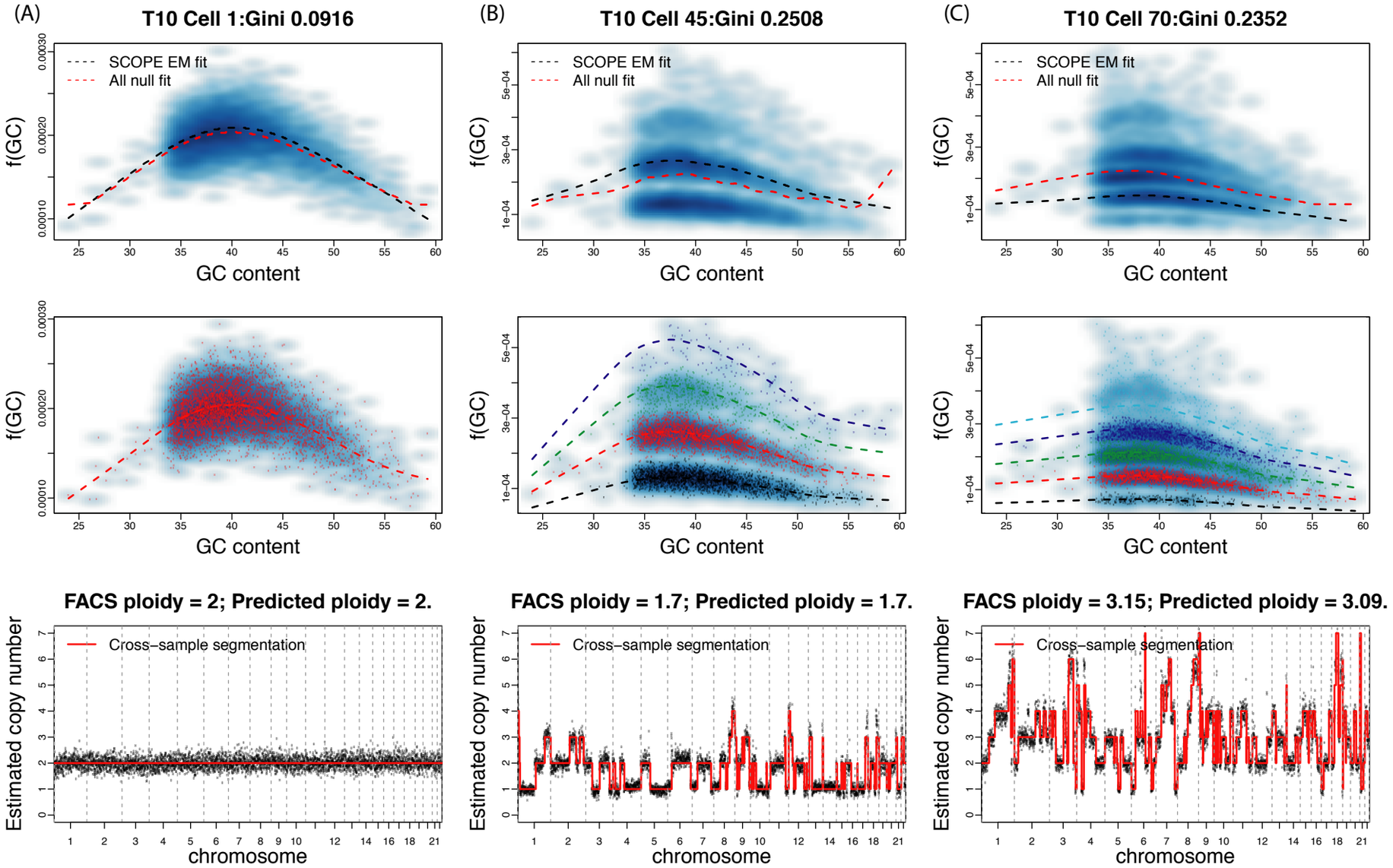
(A)
T10 Cell 1:Gini 0.0916
SCOPE EM fit
All null fit
f(GC)
GC content
(B)
T10 Cell 45:Gini 0.2508
SCOPE EM fit
All null fit
f(GC)
GC content
(C)
T10 Cell 70:Gini 0.2352
SCOPE EM fit
All null fit
f(GC)
GC content
f(GC)
GC content
f(GC)
GC content
f(GC)
GC content
FACS ploidy = 2; Predicted ploidy = 2.
Cross−sample segmentation
Estimated copy number
chromosome
FACS ploidy = 1.7; Predicted ploidy = 1.7.
Cross−sample segmentation
Estimated copy number
chromosome
FACS ploidy = 3.15; Predicted ploidy = 3.09.
Cross−sample segmentation
Estimated copy number
chromosome

Figure 2

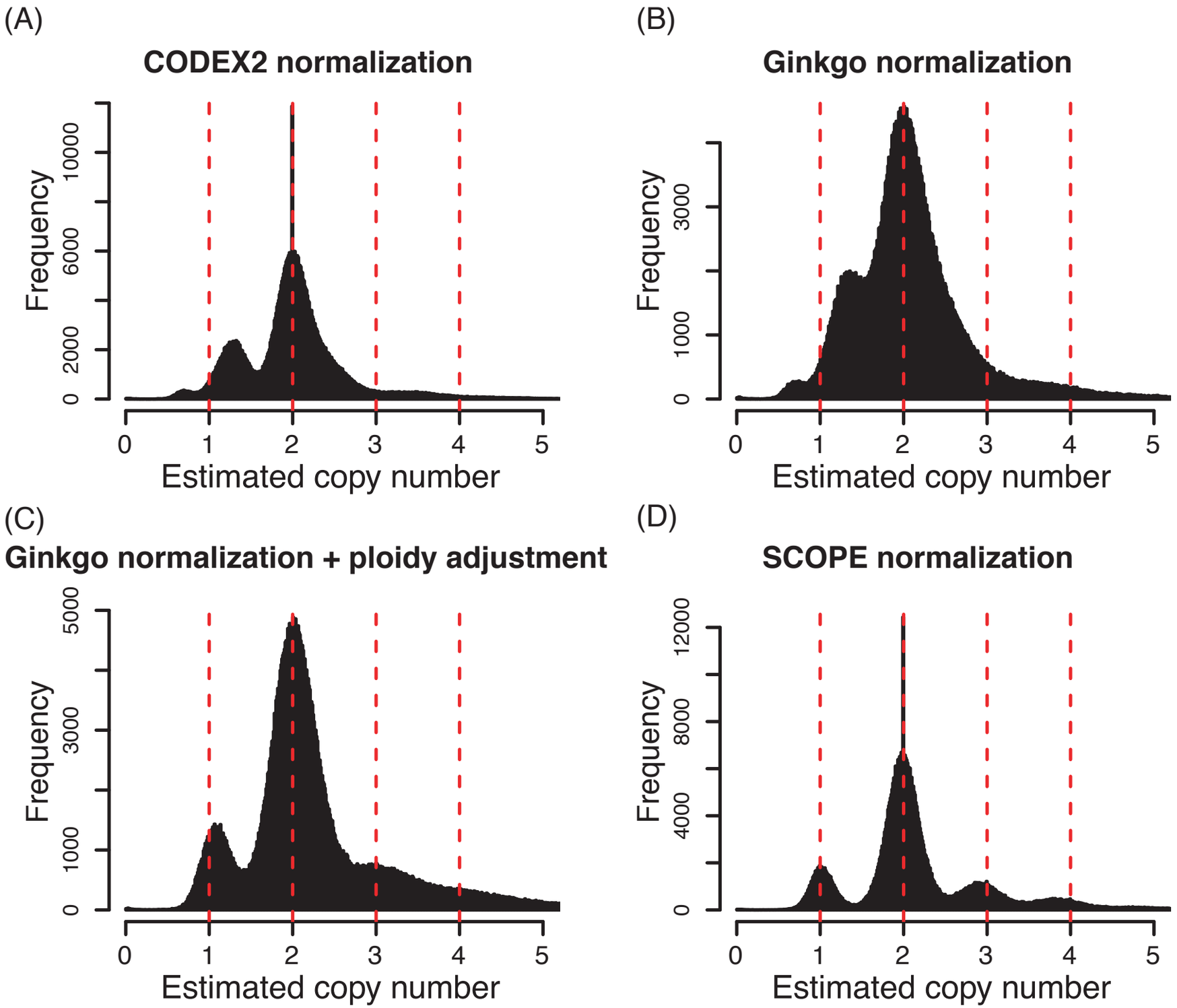
(A)
CODEX2 normalization
Frequency
0
2000
6000
10000
0
1
2
3
4
5
Estimated copy number
(B)
Ginkgo normalization
Frequency
0
1000
3000
0
1
2
3
4
5
Estimated copy number
(C)
Ginkgo normalization + ploidy adjustment
Frequency
0
1000
3000
5000
0
1
2
3
4
5
Estimated copy number
(D)
SCOPE normalization
Frequency
0
4000
8000
12000
0
1
2
3
4
5
Estimated copy number

Figure 3

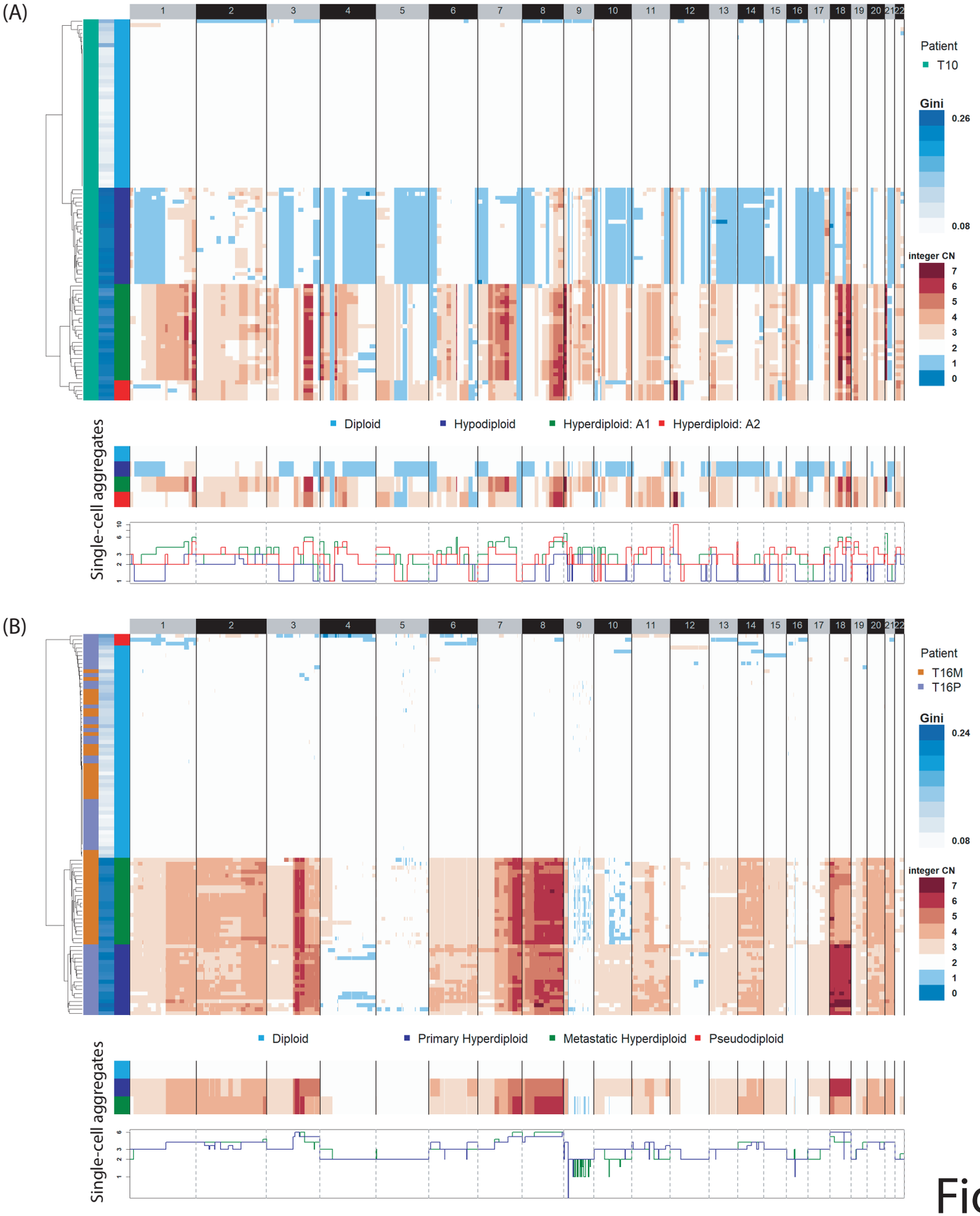
(A)
Patient
T10
Gini
0.26
0.08
integer CN
7
6
5
4
3
2
1
0
Diploid
Hypodiploid
Hyperdiploid: A1
Hyperdiploid: A2
Single-cell aggregates
(B)
Patient
T16M
T16P
Gini
0.24
0.08
integer CN
Diploid
Primary Hyperdiploid
Metastatic Hyperdiploid
Pseudodiploid
Single-cell aggregates

Figure 4

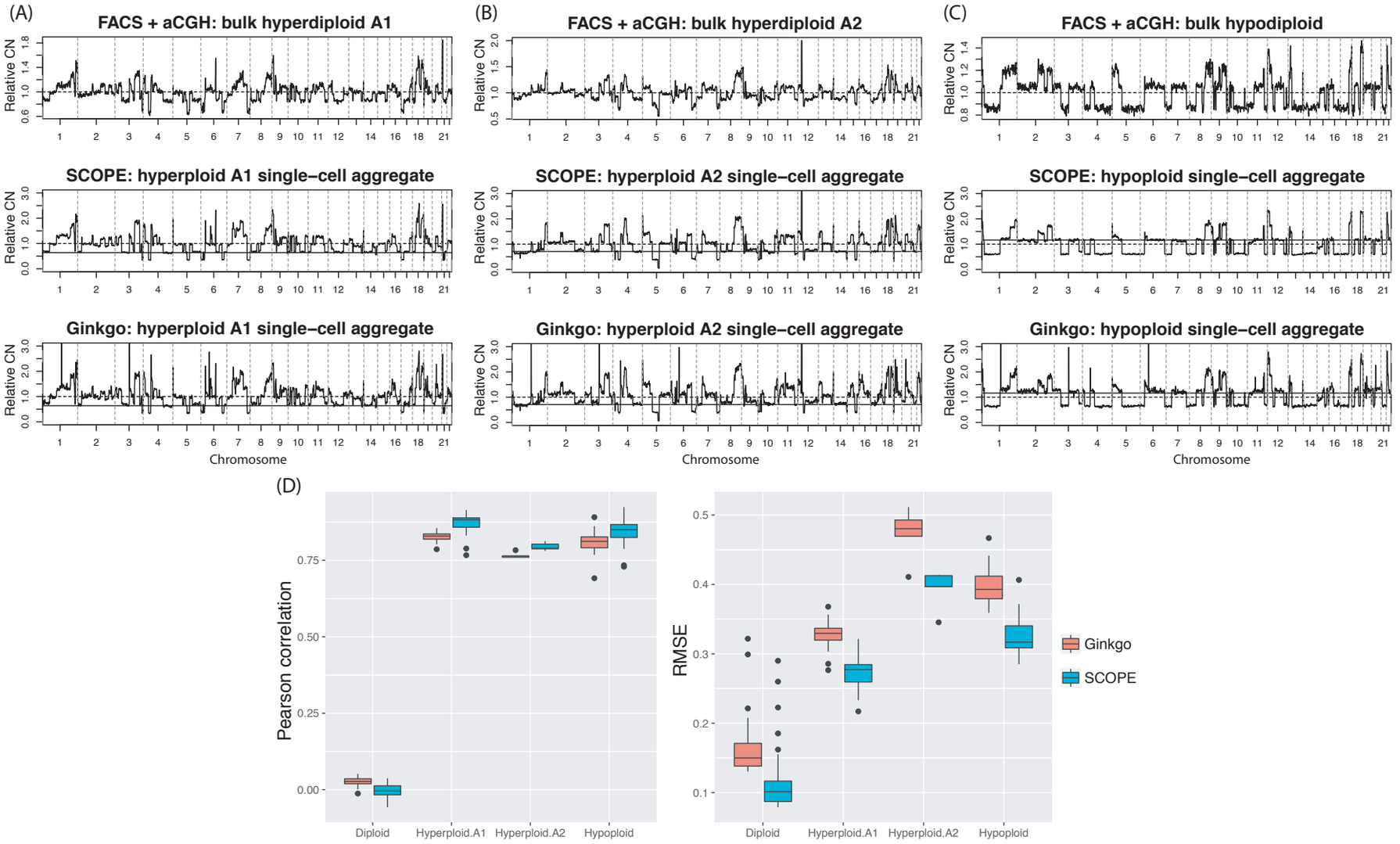
(A)
FACS + aCGH: bulk hyperdiploid A1
SCOPE: hyperploid A1 single–cell aggregate
Ginkgo: hyperploid A1 single–cell aggregate
(B)
FACS + aCGH: bulk hyperdiploid A2
SCOPE: hyperploid A2 single–cell aggregate
Ginkgo: hyperploid A2 single–cell aggregate
(C)
FACS + aCGH: bulk hypodiploid
SCOPE: hypoploid single–cell aggregate
Ginkgo: hypoploid single–cell aggregate
Relative CN
Chromosome
(D)
Pearson correlation
RMSE
Diploid
Hyperploid.A1
Hyperploid.A2
Hypoploid
Ginkgo
SCOPE

Figure 5

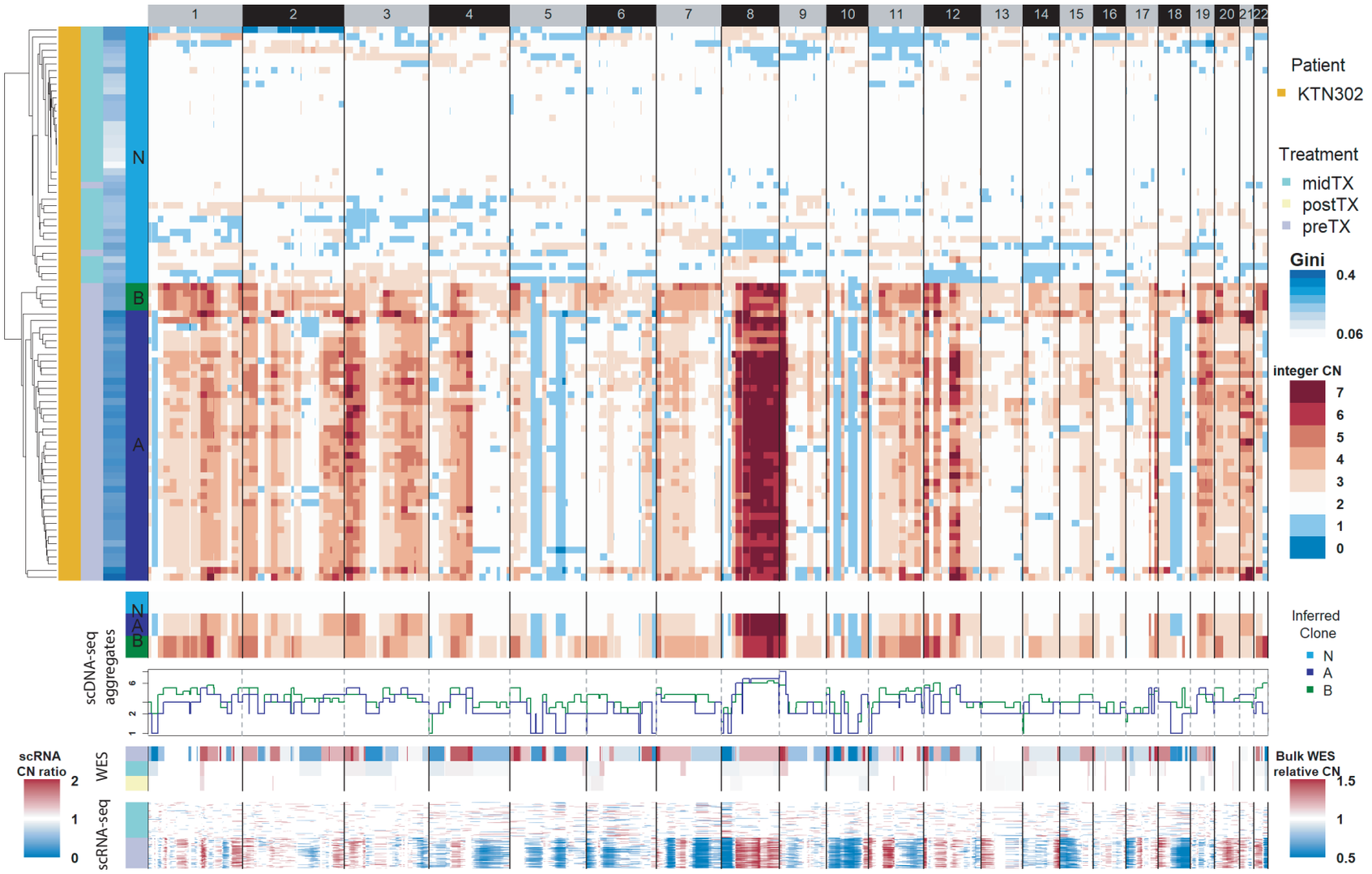
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
N
B
A
Patient
KTN302
Treatment
midTX
postTX
preTX
Gini
0.4
0.06
integer CN
7
6
5
4
3
2
1
0
scDNA-seq aggregates
Inferred Clone
N
A
B
WES
scRNA-seq
scRNA CN ratio
2
1
0
Bulk WES relative CN
1.5
1
0.5

Figure 6

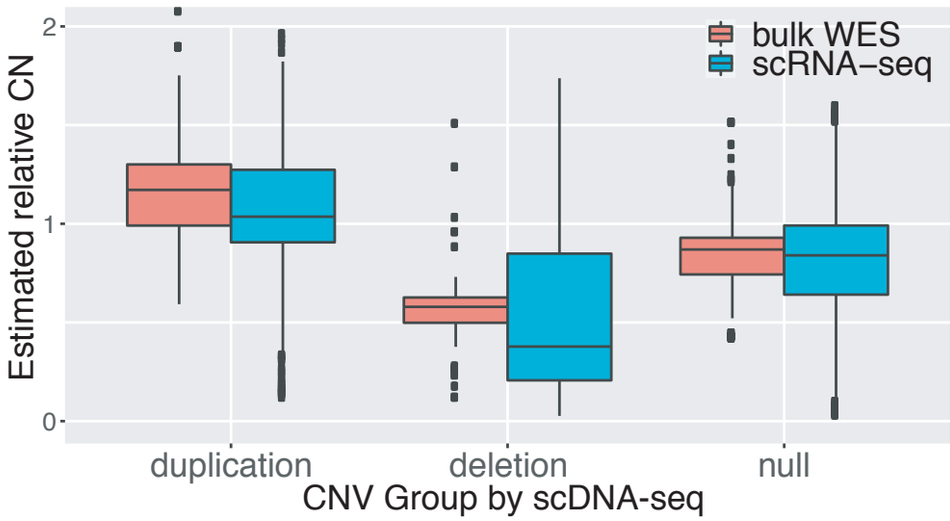
bulk WES
scRNA−seq
Estimated relative CN
2
1
0
duplication
deletion
null
CNV Group by scDNA-seq

Figure 7

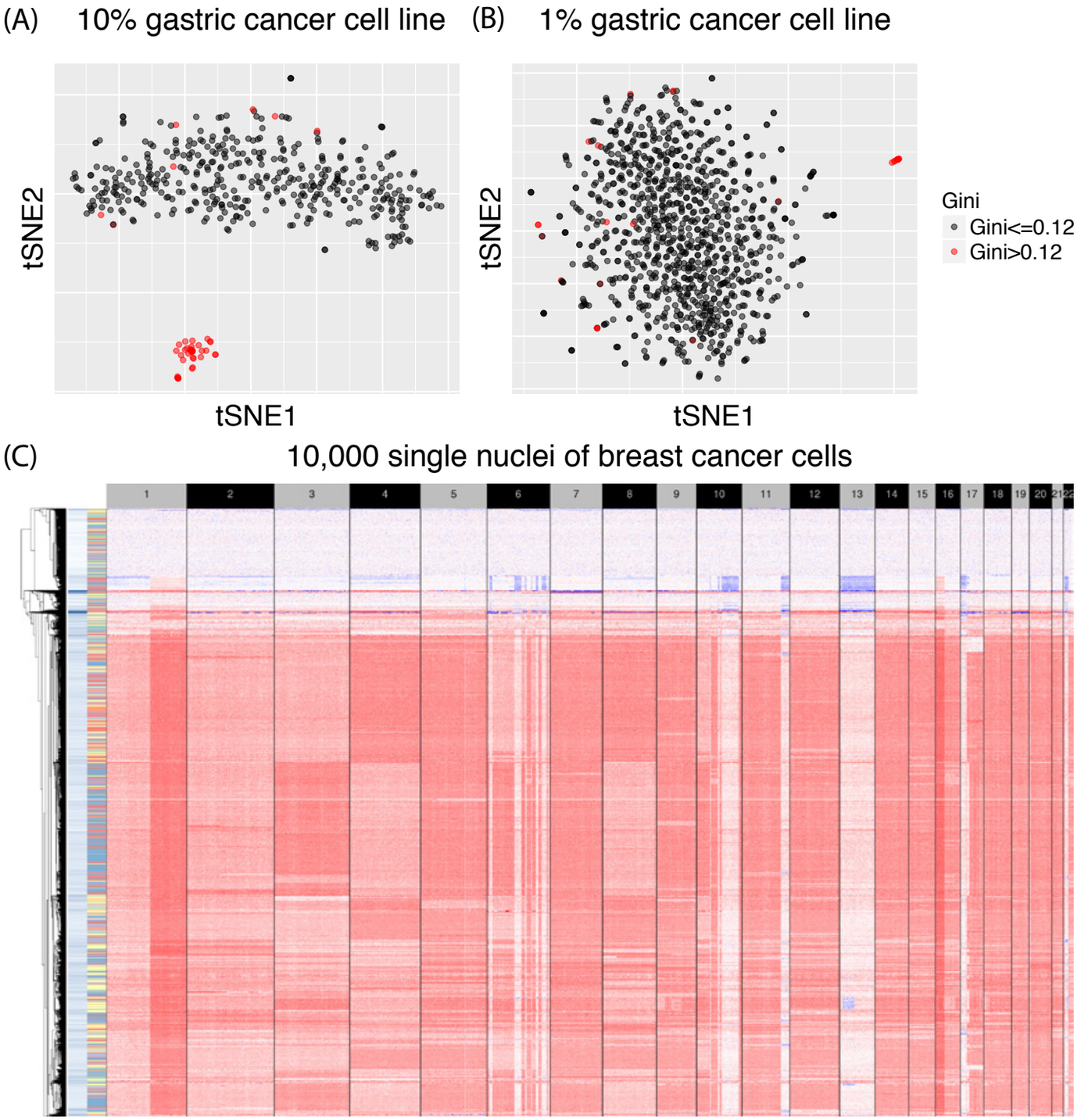
(A) 10% gastric cancer cell line
(B) 1% gastric cancer cell line
tSNE2
tSNE1
Gini
Gini<=0.12
Gini>0.12
(C) 10,000 single nuclei of breast cancer cells

Figure 8