scientific reports

OPEN

Inference of chromosome selection parameters and missegregation rate in cancer from DNA-sequencing data

Zijin Xiang, Zhihan Liu & Khanh N. Dinh

Aneuploidy is frequently observed in cancers and has been linked to poor patient outcome. Analysis of aneuploidy in DNA-sequencing (DNA-seq) data necessitates untangling the effects of the Copy Number Aberration (CNA) occurrence rates and the selection coefficients that act upon the resulting karyotypes. We introduce a parameter inference algorithm that takes advantage of both bulk and single-cell DNA-seq cohorts. The method is based on Approximate Bayesian Computation (ABC) and utilizes CINner, our recently introduced simulation algorithm of chromosomal instability in cancer. We examine three groups of statistics to summarize the data in the ABC routine: (A) Copy Number-based measures, (B) phylogeny tip statistics, and (C) phylogeny balance indices. Using these statistics, our method can recover both the CNA probabilities and selection parameters from ground truth data, and performs well even for data cohorts of relatively small sizes. We find that only statistics in groups A and C are well-suited for identifying CNA probabilities, and only group A carries the signals for estimating selection parameters. Moreover, the low number of CNA events at large scale compared to cell counts in single-cell samples means that statistics in group B cannot be estimated accurately using phylogeny reconstruction algorithms at the chromosome level. As data from both bulk and single-cell DNA-sequencing techniques becomes increasingly available, our inference framework promises to facilitate the analysis of distinct cancer types, differentiation between selection and neutral drift, and prediction of cancer clonal dynamics.

Current advancements in genomics technologies have enabled researchers to examine the extent and patterns of tumor chromosomal instability. Over the past two decades, there has been great technological and computational progress in bulk DNA-sequencing (bulk DNA-seq) methods, resulting in more uniform coverage and deeper sequencing depth at a lower cost. This paved the way for large pan-cancer genomic studies, such as The Cancer Genome Atlas (TCGA)[21] and Pan-Cancer Analysis of Whole Genomes (PCAWG)[20]. The enhanced statistical power resulting from the large sample sizes has enabled identification of cancer drivers, classification of tumor subtypes, and subsequently better diagnosis and treatment decisions based on genetic biomarkers[47,50,56]. More recently, single-cell DNA-sequencing (scDNA-seq) technologies have emerged as a powerful method to uncover the genomic heterogeneity in individual tumors[28,30,35,40]. The DNA profiles of individual cells and their inferred phylogenies also enable the analysis of how the cancers evolved over time, and which genetic features are associated with tumor expansions, metastasis and relapse[18,40].

The application of both bulk and scDNA-seq in cancer research has led to an increased understanding of the selective role of Copy Number Aberrations (CNAs)[4,7,53]. Defined as deletions or amplifications of large genomic regions, CNAs have been observed to enrich oncogenes and inactivate tumor suppressor genes, contributing to uncontrolled proliferation and apoptosis evasion in cancer[3]. Successful CNA detection has resulted in better patient outcome prediction[5] and personalized treatment[13,16].

We have recently introduced CINner, an efficient algorithm for simulating chromosomal instability during tumorigenesis[9]. It allows for flexible characterization of copy numbers in individual cells, and considers both the generation and selection of diverse karyotypes. When limited to whole chromosomes, CINner extends the approach by Lynch et al.[32] to simultaneously consider missegregation rate and tissue-specific selection parameters.

In this paper, we examine the problem of inferring the missegregation rate and selection coefficients from DNA-seq data with the CINner model. We construct the parameter inference method based on the Approximate

Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY, USA. ✉email: knd2127@columbia.edu

Bayesian Computation (ABC) framework[19], utilizing statistics that characterize the observed copy number profiles and cell phylogeny from a combination of bulk and scDNA-seq data. We first investigate which statistics are most informative in capturing the signals of CNA heterogeneity and selection, and analyze the accuracy of the inferred parameters when applied to simulated DNA-seq data. We then investigate the dependence of inference accuracy on the data sample sizes. Finally, we assess the important open question of the fidelity of inferred statistics, and how it might impact the accuracy of our method. Specifically, the analysis of scDNA-seq data includes the inference of cell-specific copy numbers from readcount data, and deduction of a phylogeny tree that is most compatible of the inferred genomes. As the sequencing technology and computational analysis might be prone to noise, the resulting statistics might not be accurate and can impact further parameter estimation. We test the reliability of the statistics by comparing between CINner simulations and inferences from MEDICC2[23], a phylogeny algorithm that has shown great applicability for scDNA-seq analysis.

Results

Inference of copy number aberration rates and selection parameters in the CINner framework

In order to fully characterize the genomic evolution from DNA-seq data, it is important to quantify the rates at which CNAs arise and the selection forces that act upon them. Valind et al. constructed a discrete in silico model to analyze the increased prevalence of aneuploidy in cancer[51]. Their model assumes strong negative selection on aneuploid cells in normal tissues that is relaxed in tumors. While CNAs are indeed more tolerated in cancer, sequencing data has shown elevated frequencies of certain aneuploidies, either across tumors or specifically to some cancer subtypes[38], indicating that they are under positive selection. Elizalse et al. adapted a Markov-chain model to estimate the karyotypic distributions[12], building upon a stochastic chromosome copy number evolution model[29]. This approach derived optimal chromosome missegregation probabilities for maximal karyotypic heterogeneity while optimizing the computational efficiency. Salehi et al. employed a Bayesian fitness model grounded in the Wright-Fisher diffusion to infer clonal fitness coefficients from their growth trajectories in time-series scDNA-seq data[40]. Since the model does not consider new CNAs, it is applicable for the analysis of short-term dynamics and might be inefficient for studying the entire tumor history. Lynch et al. developed a model to infer both selection and CNA rates from scDNA-seq[32]. The framework models fitness as a scaling factor multiplied by the OG-TSG score from Davoli et al.[7], which quantifies the count and potency of oncogenes and tumor suppressor genes, then infers the scaling factor and missegregation rate. As the OG-TSG score is computed for all genes across pan-cancer TCGA samples, it might be challenging to adapt this framework to study individual cancer types, where identification of driver genes and their mutation frequencies is difficult due to low sample counts.

We recently introduced CINner[9], an efficient algorithm to simulate the evolution of chromosomal instability (CIN) during tumorigenesis. CINner uses a birth-death process to model tumor cells[25], where a cell's fitness and probability of division depend on its karyotype (Figure 1a). In this paper, we focus on the problem of inferring the parameters governing whole-chromosome missegregations, which have been frequently observed across different cancer types[28,38]. The parameters of interest include p_{misseg}, the probability that a missegregation event occurs in a cell division, and selection parameters $\{s_i\}$, where s_i quantifies the change in cellular fitness as chromosome i is amplified or deleted (Figure 1b). The parameters p_{misseg} and $\{s_i\}$ are assumed to be constant among cells and samples. Each CINner simulation starts from a diploid population at time 0 and a sample is taken at 80 years, when a hypothetical patient is diagnosed (see Methods).

In traditional Bayesian inference, the posterior distributions for p_{misseg} and $\{s_i\}$ are proportional to the data's likelihood and prior probabilities[27]. However, numerical computation of the likelihood requires many simulations for each parameter set, rendering this approach too computationally expensive for our problem. Therefore, we implement Approximate Bayesian Computation (ABC), a Bayesian inference approach that replaces the likelihood by a distance function between statistics from the data and those simulated from a model. Simulation has been used to approximate the likelihood[8], and it was applied to estimate the posterior distributions of coalescence times and mutation rates from DNA-seq data in population genetics[48].

Over the last decades, ABC has been used in different fields of study to estimate parameters for complex models, especially in biology[2]. Many algorithms have been developed to improve ABC's performance for different task requirements[44]. For our problem, we consider a wide range of statistics for DNA-seq observations, to be discussed in the next section. We utilize ABC-random forest (ABC-rf), as the algorithm is less sensitive to noise impacted by poor choices of summary statistics[39] (Figure 1c). For each parameter, ABC-rf builds a random forest from a training set consisting of sampled values from the prior distribution and corresponding simulated statistics. It then predicts the posterior distribution with regression from the random forest conditional on observed statistics from data, without requiring a metric on the statistic space.

Statistics for copy number profiles and cell phylogeny from DNA-sequencing data

The accuracy and robustness of ABC's application in parameter inference depend heavily on the choice of summary statistics[39]. This is especially true for methods depending on a metric to compare statistics from data and model[44]. The large sample sizes of recent bulk DNA-seq studies enable accurate depiction of the selection landscape. However, the method offers only limited information about cancer clonality, which carries the signals for CNA rates. On the other hand, scDNA-seq captures the heterogeneity in individual tumors, but current technical and financial limits result in modest datasets that are prone to over-fitting. By combining the two data sources, we can accurately recover both missegregation rate and selection parameters. In this section, we describe several statistics for Copy Number profiles and cell phylogeny that can be measured from bulk and scDNA-seq data. We then analyze their performance in indicating the signals of CNA rates and selection parameters. The statistics are categorized into three groups, depending on the target aspect of the data.

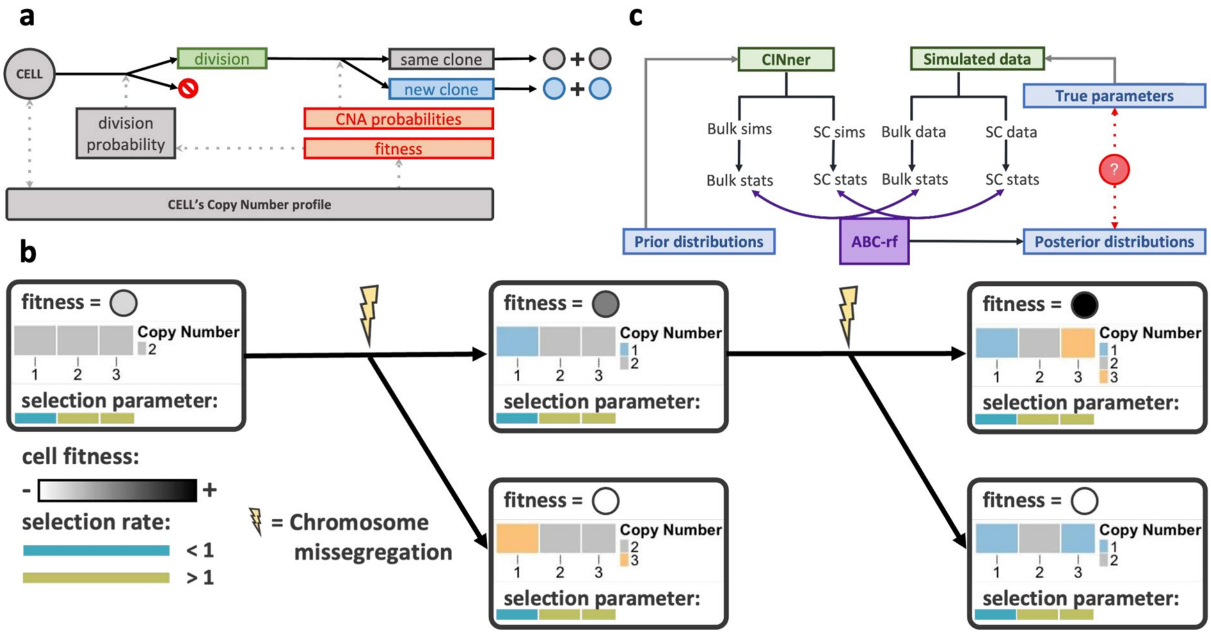

Figure 1. Overview of the methodology. (**a**) Schematic of CINner (adapted from[9]). Cells follow a birth-death process. The probability of division depends on the cell's fitness, determined based on its copy number profile. When cells divide, new clones may arise according to established CNA probabilities. (**b**) Selection model (adapted from[9]). A cell's fitness depends on its copy numbers and chromosome selection parameters. The fitness increases after a missegregation if the cell gains a chromosome with selection parameter > 1, or loses one with selection parameter < 1. (**c**) Application of Approximate Bayesian Computation (ABC) in parameter inference. Parameters are drawn from prior distributions, then statistics are computed from bulk and scDNA-seq samples simulated with CINner. ABC-rf then determines the parameter posterior distributions that can be compared against true values.

Several statistics estimate the heterogeneity and CNA burden from CN profiles, which we categorize as "CN statistics". Shannon diversity index[43] measures species diversity in a given cohort based on species count and abundance, and has been widely used in genetic population studies. Given a scDNA-seq data set, individual cells can be clustered into subclones based on their CN profiles. The Shannon diversity index can then be computed from the subclone count and the sizes of each subclonal population. A more heterogeneous sample would result in higher Shannon index, and vice versa. The clonal and subclonal CNA event counts can also be calculated from CN profiles (Figure 2a). Clonal CNAs are exhibited in all cells, and therefore assumed to have occurred before the Most Recent Common Ancestor (MRCA) in the cell phylogeny. Subclonal CNAs occur after the MRCA and are carried only by a subgroup of cells. We define the subclonal CNA count as $\sum_{\gamma} \frac{N_{\gamma}}{N}$, in which each CNA γ is weighted by N_{γ}, the number of cells carrying it, against the total cell count N. Finally, we directly compute the distances between CINner simulations and DNA-seq data cohorts. For a bulk DNA-seq set containing n samples, we produce n simulations. The distance matrix $A \in \mathbb{R}^{n \times n}$ indicates the precision of the simulation cohort, where $A(i, j)$ is the Euclidean distance between CN profiles of simulation i and DNA-seq sample j. The optimal transport algorithm[10,41] then calculates the minimal distance from the simulation cohort to the data cohort, which we refer to as the bulk DNA distance (Figure 2b). A similar strategy is used to compute the scDNA distance (Figure 2c). Each data sample j can be represented as $\{c_l^j\}$, where c_l^j is the CN profile of cell l. Similarly, the CN profiles from simulation i are defined as $\{\bar{c}_l^i\}$. We define the distance matrix $B^{i,j} \in \mathbb{R}^{N_i \times N_j}$, where $B^{i,j}(l_1, l_2)$ is the Euclidean distance between $\bar{c}_{l_1}^i$ and $c_{l_2}^j$. The entry $A(i, j)$ in the distance matrix A is then the optimal transport distance measured from $B^{i,j}$, and a final application of optimal transport on A produces the scDNA-seq distance. This approach can be computationally prohibitive, as recent scDNA-seq cohorts contain up to several thousand cells. To decrease the runtime, we define $B^{i,j}$ for each pair of subclones in simulation i and data sample j. The value for $A(i, j)$ then results from the optimal transport where the probability distributions for simulation i and data sample j are weighted for the subclonal cell counts.

We divide the cell phylogeny statistics into two groups, the "tip statistics" and the "balance statistics" (Figure 2a). The tip statistics are associated with the leaves in the phylogeny tree. A cherry is defined as two leaves that merge directly with each other, and a pitchfork is a group of three leaves merging into one internal node. Cherry and pitchfork counts are the normalized numbers of these structures in the entire tree[6]. A ladder is defined as a sequence of internal nodes where each node has exactly one direct descendant. IL number and average ladder are the count of ladders and the average ladder length in the phylogeny tree, respectively[26]. In contrast to the tip statistics, the balance statistics quantify whether the tree is balanced or imbalanced. Max depth is the height of the phylogeny tree when branch lengths are normalized, which is smaller for a more balanced tree[15]. Stairs measures the proportion of subtrees that are imbalanced[26]. Colless is the sum of balance values among all internal nodes, where the value for each internal node is the absolute difference between sizes of clades stemming from it[26]. Both Sackin and Colless indices measure the imbalance extent of trees. In contrast, B2 is a balance index.

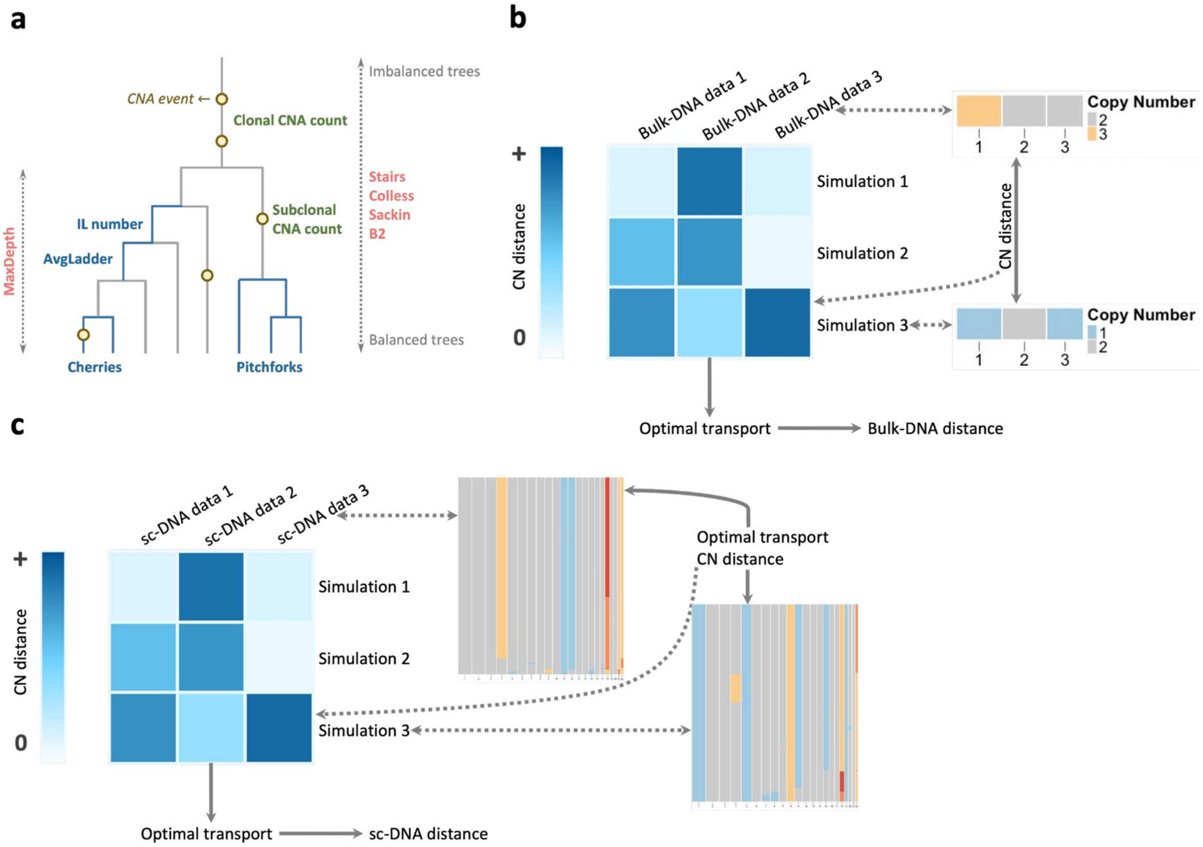

Figure 2. Some statistics quantifying the similarity between DNA-seq data and simulations. (**a**) Statistics from CN data and cell phylogeny are grouped into CN statistics (green), phylogeny tip statistics (blue) and phylogeny balance statistics (red), depicted on a representative phylogeny tree. (**b**) Bulk DNA distance is defined as the optimal transport cost from simulations to data samples. The distance matrix consists of Euclidean distances between each pair of samples. (**c**) The scDNA distance is similarly based on the optimal transport from simulations to data samples. The distance between each pair of samples requires first finding the optimal transport between the sampled cells.

B2 for tree N is defined as $B_2(N) = -\sum_{\ell \in L} p_\ell \log_2 p_\ell$, where p_ℓ is the probability of ending at leaf ℓ conditioned on traveling from the root with equal probabilities at each merge[17].

In order to evaluate the effectiveness of each statistic in capturing the signals of CNA rates and selection parameters, we compute the statistics from CINner simulations with varying parameters. The chromosome selection parameters for each simulation are sampled from Uniform(1/1.2, 1.2), and the missegregation probability is sampled as $\log_{10}(p_{misseg}) \sim \text{Uniform}(-5, -3)$. For each parameter set, we simulate 100 bulk samples and 50 scDNA samples, then compute the statistics as described above. Afterward, we compute the correlation between each parameter and the mean and variance of each statistic from 100,000 simulations. If a statistic strongly correlates with a parameter, it is potentially a good candidate for ABC parameter inference. We note that, as the selection parameters only affect the chromosomes that they are assigned to, they have minor impact on genome-wide statistics. Therefore, the CN statistics (e.g. Shannon diversity index, clonal and subclonal CNA counts, and CN distances) are computed based only on chromosome i when compared agaisnt selection parameter s_i. In contrast, CNA rates affect all chromosomes, therefore they are compared against CN statistics computed for the entire genome. The other statistics are based on cell phylogeny, which cannot be segregated for individual chromosomes. Therefore, the same values are compared with both selection parameters and missegregation probability. Finally, the CN distances require direct comparison to the data samples. Therefore, we simulate 100 bulk samples and 50 scDNA samples to serve as the DNA data, with ground-truth parameters $p_{misseg} = 2 \times 10^{-4}$ and $s_i \sim \text{Uniform}(1/1.15, 1.15)$. Because the CN distances measure the proximity between entire cohorts of data samples and simulations, they can be used at once without other summary statistics such as mean or variance.

The correlations between the DNA-seq statistics and CINner parameters are presented in Figure 3. All of the CN statistics are strong signals for the missegregation probability. As the probability increases, the samples become more heterogeneous and contain more aneuploidies both at clonal and subclonal levels, Therefore, the Shannon index and missegregation counts are positively correlated with the missegregation probability. Compared to the CN statistics, the correlations between the phylogeny tip statistics and the missegregation probability are extremely weak. This indicates that these statistics are more representative of the sample size than of the heterogeneity associated with the scale of CNA rates. Specifically, if the subclone count is small compared to the sample size, then the cherry count in the phylogeny tree is approximately the sum of cherry counts in distinct

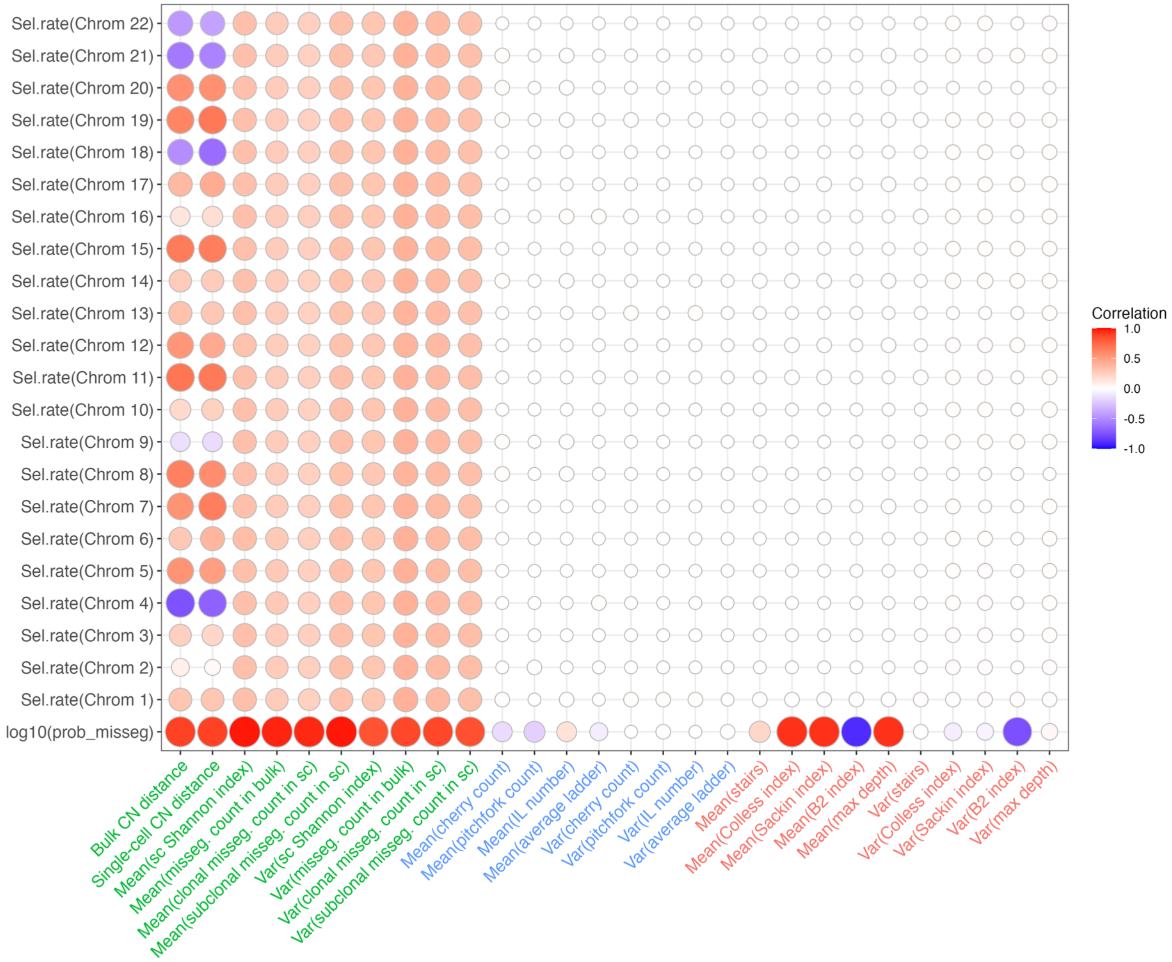

Figure 3. Correlations between sample statistics and model parameters. Statistics are categorized into CN statistics (green), phylogeny tip statistics (blue) and phylogeny balance statistics (red). Size and color of each circle indicate the correlation.

subclones. Because the cells in each subclone are identical, the count only depends on the subclonal population size. The same is true for pitchforks and ladders, rendering these statistics ineffective of capturing the heterogeneity and selection from genomic data. Finally, the mean phylogeny balance indices correlate strongly with the missegregation probability. As the missegregation probability is elevated, there are increasingly more subclones with distinct fitness competing to expand, resulting in higher imbalance in the cell phylogeny. Therefore, the correlation is negative if the index increases when the trees are more balanced (B2) and positive if the index rises for more imbalanced phylogony (Colless and Sackin indices, stairs, and max depth). Moreover, the variance of B2 also correlates with the missegregation probability.

In contrast, only CN statistics strongly correlate with the chromosome selection parameters. Nevertheless, the correlations are weaker than between CN statistics and missegregation probability. This is because the aneuploidy level in each chromosome is significantly lower than in the entire genome. Therefore, the CN statistics for specific chromosomes are sparser than the genome-wide equivalences. As selection parameters increase, subclonal competition is intensified, and only the cells harboring the most favorable karyotypes can expand, resulting in higher missegregation counts both clonally and subclonally. The correlation signs of the CN distances depend on the individual chromosomes. If a chromosome i has ground-truth selection parameter $s_i \gg 1$ (e.g., chromosomes 4, 18, 21 and 22, Figure 4a), it is frequently amplified in the observed DNA samples. As the selection parameter increases in CINner, the simulated samples also regularly exhibit gains of this chromosome, reducing the CN distances, therefore the correlation is negative. On the other hand, if the ground-truth selection parameter $s_i \ll 1$ (e.g. chromosomes 11, 15 and 19), the observed DNA samples frequently display lower copy numbers for chromosome i. If the CINner simulations have higher selection parameters, the increased gain counts of chromosome i expand the CN distances and lead to positive correlations. The phylogeny-based statistics, including the tip statistics and even the phylogeny balance indices, have no correlation with individual selection parameters. This is because the selection parameter of one chromosome has minimal impact on the entire cell phylogeny tree. As a result, we expect that the application of these statistics is inefficient in uncovering the selection landscape from DNA data.

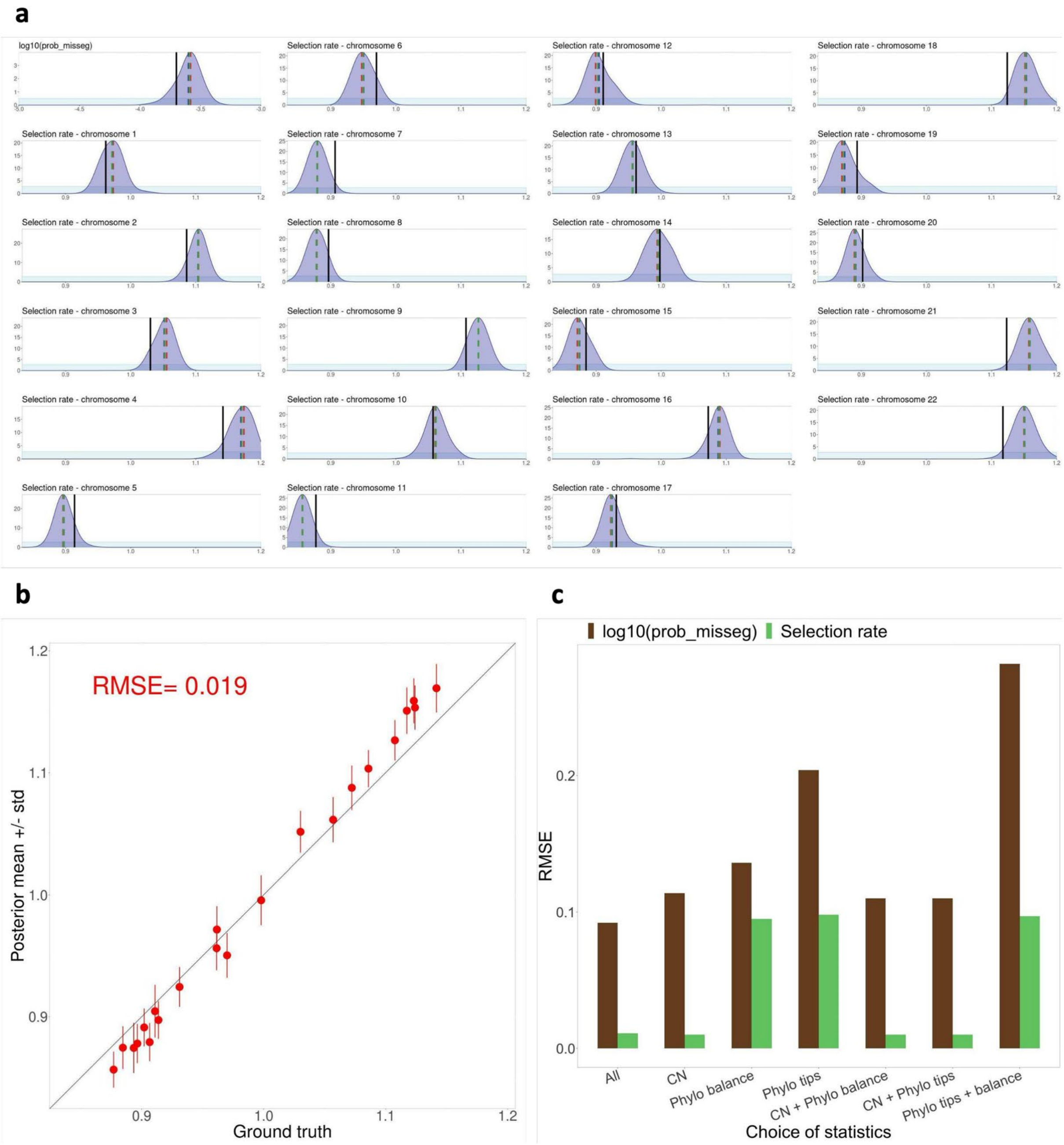

Figure 4. Parameter inference results with ABC-rf. (**a**) Prior (light blue) and ABC-rf posterior (dark blue) distributions for each parameter, with posterior mean (blue line), mode (red line), median (green line) compared against ground truth value (black line). Inference for p_{misseg} utilizes all genome-wide statistics, inference for selection parameters uses only chromosome-specific CN statistics. Simulated data = 100 bulk and 50 scDNA-seq samples, ABC-rf training set from 100,000 CINner simulations. (**b**) Correlation between ground truth and posterior mean values for all chromosome selection parameters (+/- standard deviation). Root mean square error (RMSE) computed for all selection parameters. (**c**) RMSE of the posterior mean values, depending on statistic groups used in inference.

ABC-based parameter inference accurately recovers selection and CNA rates

We construct an ABC-based parameter inference method to recover the missegregation rate and selection parameters from a mixture of bulk and scDNA-seq data cohorts. To test the algorithm, we use CINner to create a simulated dataset by combining $N_{bulk} = 100$ bulk and $N_{sc} = 50$ scDNA samples, using ground truth parameters as described previously ($p_{misseg} = 2 \times 10^{-4}$ and $s_i \sim \text{Uniform}(1/1.15, 1.15)$). A training set is built from 100,000 CINner simulations, with prior distributions $\log_{10}(p_{misseg}) \sim \text{Uniform}(-5, -3)$ and $s_i \sim \text{Uniform}(1/1.2, 1.2)$. Each CINner simulation consists of creating $N_{bulk} + N_{sc}$ independent samples with the selected parameters, then computing the statistics as described above. Finally, we train ABC-rf[39] on the library and use the random forest to infer the posterior distribution for each parameter. To evaluate the quality of the inference, we compute the root mean square error (RMSE)[55]. Let $\{\alpha_1, \ldots, \alpha_n\}$ and $\{\beta_1, \ldots, \beta_n\}$ be the ground truth and inferred values for n parameters. The RMSE is then defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \beta_i)^2}$$

As described in the previous section, the genome-wide impact of the CNA rates means that their signals are contained in all CN and phylogeny-based statistics (with the exception of statistics based on scDNA-seq phylogeny tips). In contrast, the local effect of selection parameters renders only CN-based statistics appropriate for inference (Figure 3). Therefore, our inference for p_{misseg} employs all statistics, and the inference for each s_i uses only the CN statistics. The posterior distributions (Figure 4a) of either $\log_{10}(p_{misseg})$ or each s_i exhibit almost identical mode, mean and median. Moreover, the posterior peaks are centered close to the ground truth values. Specifically, the mode, mean, and median of $\log_{10}(p_{misseg})$ are -3.58, -3.6, -3.59, which are slightly higher but close to the ground truth value of $\log_{10}(2 \times 10^{-4}) \approx -3.69$. Similarly, the inferred s_i's are close to the ground truth, with the RMSE of posterior means being 0.019 (Figure 4b). The inference tends to be slightly more extreme than the true values: if true $s_i > 1$, then inferred $\bar{s}_i$ is often greater than s_i, and vice versa. Combined, these results indicate that our inference method, based on ABC and employing appropriate statistics for each parameter class, can recover the missegregation rate and the selection landscape driving cancer evolution from DNA-seq data.

We further examine the performance of our inference method when the statistics are chosen differently (Figure 4c). As expected, for fitting p_{misseg}, ABC-rf using only CN statistics performs the best, followed by phylogeny balance statistics and much higher RMSE when using only phylogeny tips measures. Intriguingly, the inference with the combination of phylogeny tips and balance statistics fares even worse than ABC-rf using either group individually. One possible explanation is that there are too few signals contained in the balance statistics to offset the increased noise introduced by the tip statistics. Note that ABC-rf is relatively insensitive to noisy statistics, compared to other ABC-based methods[39]. Algorithms such as sequential Monte Carlo (SMC)[49] or Markov Chain Monte Carlo (MCMC)[33], which use a metric to compare statistics between the model and data, would likely suffer more from the increased noise level of these measures. Finally, our chosen statistics set, consisting of all three groups, results in the lowest RMSE in inferred p_{misseg}.

Similarly, the inferred selection parameters incur the lowest RMSE when CN statistics are employed, either unaccompanied or combined with phylogeny-based statistics. Because the latter is essentially noise, the inference using all statistics is similar to only using CN statistics. ABC-rf's analysis of variable importance (Figures S1, S2 and S3) confirms that the CN statistics and the mean phylogeny balance indices are most important in inferring p_{misseg} and s_i's.

Sensitivity analysis

We showed that our inference method can recover the missegregation probability and selection parameters from a cohort of 100 bulk DNA-seq and 50 scDNA-seq samples. However, due to many constraints, the available bulk DNA-seq samples for specific cancer types can be of much smaller sizes[21]. Additionally, despite recent advances, single-cell DNA sequencing remains expensive and technologically challenging, and existing data cohorts rarely consist of more than a few samples[18,28,40]. Therefore, we examine the impact of the sample sizes on the accuracy of the inferred selection landscape. For a given $\{N_{bulk}, N_{sc}\}$, we use CINner to simulate N_{bulk} bulk DNA-seq samples and N_{sc} scDNA-seq samples, using similar ground truth parameters as in previous sections. We then infer p_{misseg} and selection parameters s_i from these samples, and compute the RMSE of the inferred s_i's, as well as the standard deviation in their posterior distributions. Lower RMSE indicates that the inferred selection parameters are accurate, and lower standard deviation implies that there is less uncertainty in the inference.

We first examine the inference accuracy as dependent on the scDNA-seq cohort size. We fix $N_{bulk} = 100$ and vary $N_{sc} = 5, 10, \ldots, 50$. Unsurprisingly, both RMSE and standard deviation decrease as N_{sc} increases (Figure 5a). This is likely because many of the statistics used in our method are based on CN profiles and phylogenies from the scDNA cohort. We note that although having very low N_{sc} incurs higher standard deviation in the selection

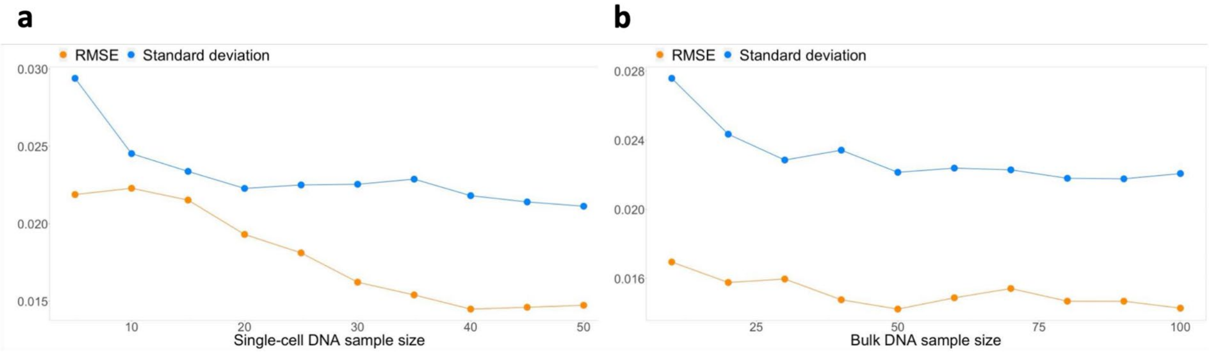

Figure 5. Dependence of inference accuracy on sample sizes. RMSE and standard deviation of selection parameter posterior means, as a function of scDNA-seq (**a**) and bulk DNA-seq (**b**) sample sizes.

parameters, the RMSE is still acceptable. This is probably because the signals from bulk samples (given that there are a sufficient number of them) can compensate for the low scDNA count in capturing the selection landscape, albeit with higher uncertainty. The standard deviation decreases as the scDNA sample size increases and plateaus at $N_{sc} \approx 20$, and the RMSE continues decreasing up to $N_{sc} \approx 40$.

Next, we fix $N_{sc} = 50$ and vary $N_{bulk} = 10, 20, \ldots, 100$. The selection parameter standard deviation decreases up to $N_{bulk} \approx 30$, but the RMSE only decreases slightly (Figure 5b). That the inference method depends less on N_{bulk} than on N_{sc} could be because there are only a few statistics in our method that are based on bulk DNA-seq data, namely the total missegregation counts and the distance between CN profiles in the data cohort and the CINner samples (Figure 2b). Another possible interpretation is that for adequately large N_{bulk}, the bulk DNA samples can already capture the chromosome-specific CNA signals required to construct the CN distance. As a result of either or both explanations, the RMSE and standard deviation stabilize for $N_{bulk} \geq 50$.

Overall, the performance of our inference method in uncovering the selection landscape reaches its peak for data cohorts consisting of as low as 50 bulk DNA-seq and 40 scDNA-seq samples. There already exists data of these scales for certain cancer types[20,54]. More importantly, the performance is only moderately worse for much smaller data cohorts. As both sequencing technologies become more widely available, our method and its insights could become instrumental in understanding the selective forces and CNA mechanisms that drive tumorigenesis.

Impact of phylogeny inference accuracy on single-cell DNA statistics

The scDNA-seq statistics employed in our method are based on CN profiles and cell phylogeny inferred from sequencing readcount data. The phylogeny inference is particularly challenging, as the observed CN profiles can typically be explained by different evolution models[11]. In this section, we investigate the impact of phylogeny inference error on scDNA-seq summary statistics.

We create 1,000 CINner simulations, each with $\log_{10}\left(p_{misseg}\right)$ sampled from Uniform$(-5, -3)$ and chromosome selection parameters from Uniform$(1/1.15, 1.15)$. We then apply MEDICC2 on the simulated single-cell CN profiles (Figure 6a). MEDICC2 reconstructs the phylogeny and infers the ancestral genomes from somatic CNAs, by computing the minimum-event distance between each pair of cells using a weighted finite-state transducer framework[23]. MEDICC2 produces a phylogeny tree rooted in the diploid genome, with CNAs occurring on specific branches such that the tips recover the observed CN profiles in the sample.

We first compare the true phylogeny against the tree inferred by MEDICC2 by using the generalized Robinson-Foulds distance[45] (Figure 6b). Unsurprisingly, the MEDICC2 inferred tree becomes more accurate as p_{misseg} increases. This is because MEDICC2 cannot stratify a group of cells if they have the same CN profiles. As p_{misseg} increases, there are more CNAs segregating distinct cells, and the MEDICC2 phylogeny becomes more resolved and closer to the ground truth.

We then analyze the summary statistics computed from MEDICC2 phylogeny (Figure 6c). The counts of clonal and subclonal missegregations, which require assigning each event on the phylogeny tree, are largely in agreement with the true values. As expected, increasing p_{misseg} results in higher missegregation counts. However, the accuracy of MEDICC2 inferences does not depend on the value of p_{misseg}. This suggests that these statistics can reliably indicate the level of aneuploidy observed in the sequencing data, which we have shown to be a strong signal for CNA probabilities and selection parameters.

In contrast, the phylogeny tip statistics differ significantly between MEDICC2 and ground truth. These statistics require accurately segregating individual tips from the remaining of the sample. For instance, locating the two tips in a cherry requires at least one CNA that differentiates them from the other cells. A pitchfork likewise requires one CNA to distinguish a group of three tips. Similarly, identifying a ladder depends on a sequence of tips that are uniquely related to each internal node. Even for high levels of p_{misseg}, the number of missegregations observed in a sample is typically not high enough to separate distinct phylogeny tips, resulting in the discrepency between MEDICC2 results and the true values. We note that there is evidence for cell-specific CNAs in recent scDNA data[18,28,40]. Most of these events are focal amplifications and deletions, and only a few are large events such as whole-chromosome or chromosome-arm missegregations. The cell-specific events can help identifying single tips in the phylogeny, refining the phylogeny tip statistics. Another potential approach is to utilize unique mutations in single cells. Because mutations occur at a much higher frequency than CNAs, they can be used to differentiate among distinct cells. However, due to low coverage, it is challenging to reliably detect unique mutations in scDNA-seq data. Future improvements in sequencing technologies and developements of phylogeny inference from both CNA and mutational data could increase the accuracy of phylogeny tip statistics.

Finally, we analyze the MEDICC2-based phylogeny balance statistics. Compared to the true values, the balance indices inferred from MEDICC2 phylogeny are mostly accurate, with the exception of the stairs index. In our correlation study, the stairs index has the weakest correlation with p_{misseg} among balance indices (Figure 3). The ABC-rf variable importance analysis also finds it to be a limited indicator for the missegregation probability (Figure S1).

In short, we find that all statistics based on CN profiles and most phylogeny balance indices in our study can be estimated reliably from MEDICC2, across different values of missegregation probabilities. In previous sections, we also found that these statistics are valuable in inferring CNA probabilities and selection parameters. On the other hand, phylogeny tip statistics cannot be reliably estimated from CN profiles, and do not carry a strong signal for the inference problem either.

Conclusion

We have recently introduced CINner[9], a simulation framework that models the impact of CNA probabilities and tissue-specific selection coefficients on the observed karyotypes. In this paper, we investigate the problem of inferring these parameters, which together shape the aneuploidy patterns in cancers. Accurate parameter

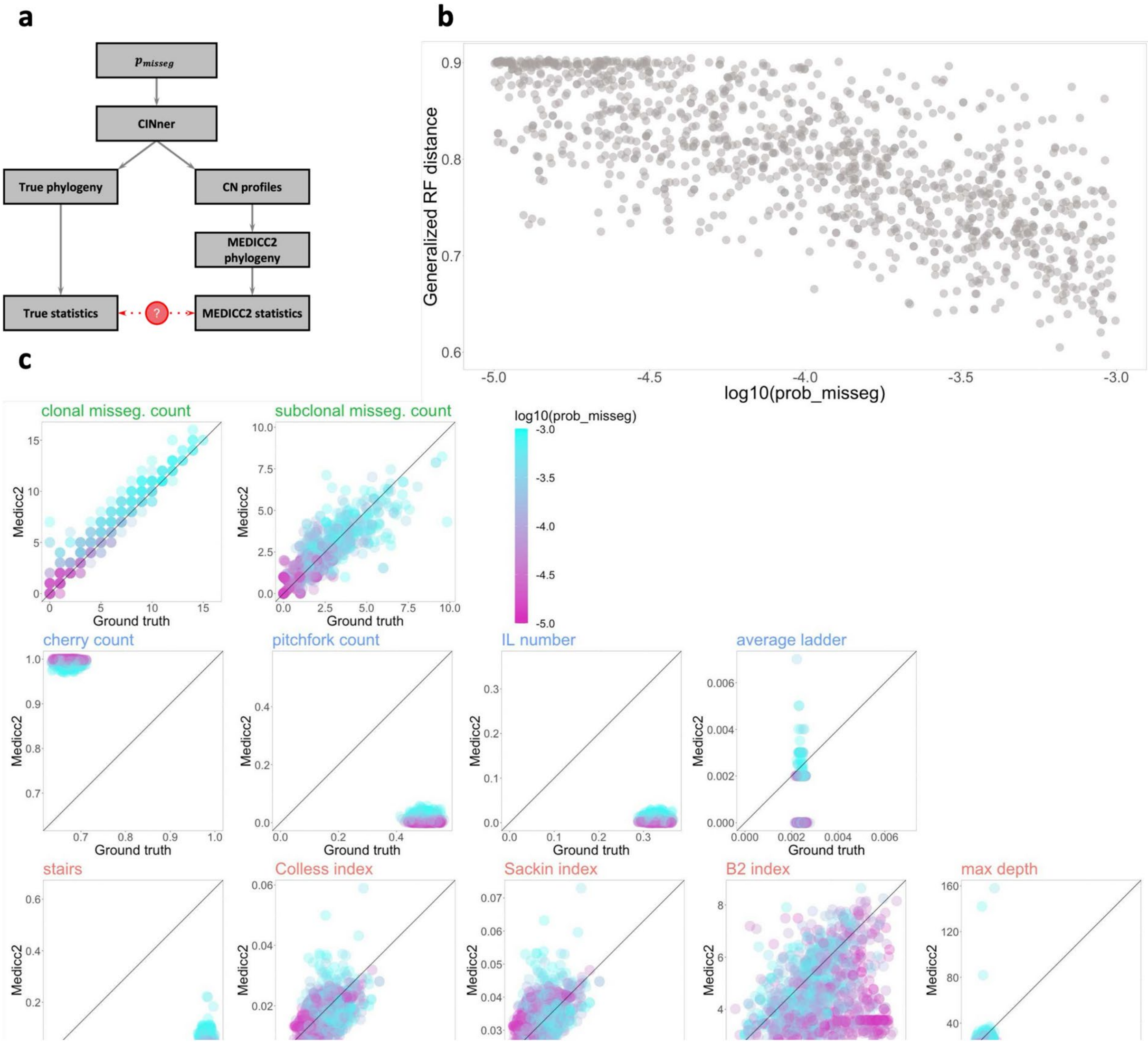

Figure 6. Accuracy of phylogeny statistics inferred from MEDICC2. (**a**) Study schematic. Statistics computed from the true phylogeny from each CINner simulation are compared against statistics estimated from the MEDICC2-inferred tree. 1,000 simulations are performed for this study, each with $\log_{10}\left(p_{misseg}\right) \sim \text{Uniform}(-5,-3)$. (**b**) Generalized Robinson-Foulds distances between true and MEDICC2 phylogeny, against corresponding p_{misseg}. (**c**) Comparison between statistics computed from true and MEDICC2 phylogeny. Color of each point denotes p_{misseg}.

inference promises to aid in stratifying patient subtype and predict tumor progression, since heterogeneity has been linked to poor patient outcome[1]. To solve the nonidentifiability issue in simultaneous inference of both CNA rates and selection parameters, we construct an inference method that employs ABC and CINner, which takes as input a mixture of bulk and single-cell DNA-sequencing data.

We find that statistics depicting clonal heterogeneity and CNA levels in sample CN profiles, together with indices for the degree of balance in scDNA-seq cell phylogenies, are important in estimating CNA probabilities. In contrast, only statistics for CN profiles are effective for inferring selection parameters. This is likely due to the fact that one chromosome's selection parameter has negligible impact on the whole sample phylogeny. In both cases, there is little signal in the statistics quantifying local features of the phylogeny.

Our algorithm can accurately recover both missegregation probability and chromosome-specific selection parameters from DNA-seq data. Peak performance requires at least 50 bulk DNA samples, which already exists for many cancer types[20,21]. However, most available scDNA studies contain less than the minimum of 20 or 40 samples necessary to minimize the uncertainty and error in the inference, respectively[18,28,40]. Nevertheless, the higher errors associated with smaller scDNA-seq cohorts still appear adequately modest, therefore biologically meaningful interpretations can still be gained. Furthermore, our numerical experiment with MEDICC2 shows that the statistics most important in inferring missegregation probability and the selection forces can be estimated accurately from real DNA-seq data.

In this work, we consider only whole-chromosome missegregations in a synthetic test study. However, it is straightforward to expand the framework to consider other CNA mechanisms simultaneously, such as chromosome-arm missegregations and whole-genome duplication[3,31]. In such cases, more detailed copy number segmentation and phylogeny reconstruction are necessary to distinguish between different CNA events. The approach outlined in this work can be applied to bulk[20,21] and single-cell DNA-sequencing data[18,28,30,35,40] for different cancers, toward uncovering tissue-specific selection coefficients and rates of chromosomal instability. However, cancer evolution is known to be highly dependent on underlying mutational processes[18]. The data cohort for inference should therefore be limited to genetically similar samples, such that the selection forces driving each tumor can be reasonably assumed to be similar. Furthermore, the sample size should be large enough to prevent overfitting. As both bulk and single-cell DNA-sequencing become more readily available, the combination of CINner and the inference framework described in this paper provides a promising approach to disentangle the effects of heterogeneity and selection that drive specific cancer types.

Data availability

Synthetic data serving as targets for the inference problems are included in the code, available at https://github.com/dinhngockhanh/CINner_missegregation_inference.

Code availability

All code for parameter inference, data analysis and sensitivity studies is available at https://github.com/dinhngockhanh/CINner_missegregation_inference.

Received: 18 April 2024; Accepted: 16 July 2024
Published online: 31 July 2024

References

1. Bakhoum, Samuel F. & Cantley, Lewis C. The multifaceted role of chromosomal instability in cancer and its microenvironment. *Cell* **174**(6), 1347–1360 (2018).
2. Beaumont, Mark A., Zhang, Wenyang & Balding, David J. Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035 (2002).
3. Beroukhim, Rameen *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**(7283), 899–905 (2010).
4. Bielski, Craig M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**(8), 1189–1195 (2018).
5. Chin, Koei *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**(6), 529–541 (2006).
6. Choi, Kwok Pui, Kaur, Gursharn & Wu, Taoyang. On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees. *J. Math. Biol.* **83**(4), 40 (2021).
7. Davoli, Teresa *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**(4), 948–962 (2013).
8. Diggle, Peter J. & Gratton, Richard J. Monte Carlo methods of inference for implicit statistical models. *J. R. Stat. Soc. Ser. B Stat Methodol.* **46**(2), 193–212 (1984).
9. Dinh, Khanh N., Vázquez-García, Ignacio, Chan, Andrew, Malhotra, Rhea, Weiner, Adam, Mcpherson, Andrew, & Tavaré, Simon. CINner: modeling and simulation of chromosomal instability in cancer at single-cell resolution. *bioRxiv*, (2024).
10. Dobrushin, Roland L. Prescribing a system of random variables by conditional distributions. *Theory Prob. Appl.* **15**(3), 458–486 (1970).
11. El-Kebir, Mohammed *et al.* Complexity and algorithms for copy-number evolution problems. *Algorithms Mol. Biol.* **12**, 1–11 (2017).
12. Elizalde, Sergi, Laughney, Ashley M. & Bakhoum, Samuel F. A Markov chain for numerical chromosomal instability in clonally expanding populations. *PLoS Comput. Biol.* **14**(9), e1006447 (2018).
13. Esteves, Luísa., Caramelo, Francisco, Ribeiro, Ilda Patrícia, Carreira, Isabel M. & de Melo, Joana Barbosa. Probability distribution of copy number alterations along the genome: an algorithm to distinguish different tumour profiles. *Sci. Rep.* **10**(1), 14868 (2020).
14. Fischer, Mareike. Extremal values of the Sackin tree balance index. *Ann. Comb.* **25**(2), 515–541 (2021).
15. Fischer, Mareike, Herbst, Lina, Kersting, Sophie, Kühn, Annemarie Luise & Wicke, Kristina. *Tree Balance Indices: A Comprehensive Surv.* (Springer Nature, 2023).
16. Fittall, Matthew W. & Van Loo, Peter. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome medicine* **11**(1), 1–14 (2019).
17. François, Bienvenu, Cardona, Gabriel & Celine, Scornavacca. Revisiting Shao and Sokal's B2 index of phylogenetic balance. *J. Math. Biol.* **83**(5), 1–43 (2021).
18. Funnell, Tyler *et al.* Single-cell genomic variation induced by mutational processes in cancer. *Nature* **612**(7938), 106–115 (2022).
19. Griffiths, Robert C. & Tavaré, Simon. The age of a mutation in a general coalescent tree. *Stoch. Model.* **14**(1–2), 273–295 (1998).
20. Hu, Taobo, Kumar, Yogesh, Ma, Eric Z., Wu, Zhenggang, Xue, Hong, *et al.* Pan-cancer analysis of whole genomes. *Nature*, (2020).
21. Kandoth, Cyriac *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**(7471), 333–339 (2013).
22. Kantorovich, Leonid V. The mathematical method of production planning and organization. *Manage. Sci.* **6**(4), 363–422 (1939).
23. Kaufmann, Tom L. *et al.* MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome Biol.* **23**(1), 241 (2022).
24. Kayondo, Hassan W. *et al.* Employing phylogenetic tree shape statistics to resolve the underlying host population structure. *BMC Bioinform.* **22**, 1–20 (2021).
25. Kendall, David G. On the generalized "birth-and-death" process. *Ann. Math. Stat.* **19**(1), 1–15 (1948).
26. Kendall, Michelle, Boyd, Michael, & Colijn, Caroline. *phyloTop: Calculating Topological Properties of Phylogenies*, (2023). R package version 2.1.2.
27. Kruschke, John K. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* **1**(5), 658–676 (2010).
28. Laks, Emma *et al.* Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**(5), 1207–1221 (2019).
29. Laughney, Ashley M., Elizalde, Sergi, Genovese, Giulio & Bakhoum, Samuel F. Dynamics of tumor heterogeneity derived from clonal karyotypic evolution. *Cell Rep.* **12**(5), 809–820 (2015).
30. Leung, Marco L. *et al.* Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* **27**(8), 1287–1299 (2017).
31. López, Saioa *et al.* Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**(3), 283–293 (2020).

32. Lynch, Andrew R., Arp, Nicholas L., Zhou, Amber S., Weaver, Beth A. & Burkard, Mark E. Quantifying chromosomal instability from intratumoral karyotype diversity using agent-based modeling and Bayesian inference. *Elife* **11**, e69799 (2022).
33. Marjoram, Paul, Molitor, John, Plagnol, Vincent & Tavaré, Simon. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.* **100**(26), 15324–15328 (2003).
34. McKenzie, Andy & Steel, Mike. Distributions of cherries for two models of trees. *Math. Biosci.* **164**(1), 81–92 (2000).
35. Minussi, Darlan C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**(7853), 302–308 (2021).
36. Norström, Melissa M., Prosperi, Mattia C.F., Gray, Rebecca R., Karlsson, Annika C. & Salemi, Marco. PhyloTempo: a set of R scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences. *Evolutionary Bioinformatics*, **8**, EBO–S9738, (2012).
37. Oksanen, Jari, Simpson, Gavin L., Blanchet, F. Guillaume, Kindt, Roeland, Legendre, Pierre, Minchin, Peter R., O'Hara, R. B., Solymos, Peter, Stevens, M. Henry H., Szoecs, Eduard, Wagner, Helene, Barbour, Matt, Bedward, Michael, Bolker, Ben, Borcard, Daniel, Carvalho, Gustavo, Chirico, Michael, De Caceres, Miquel, Durand, Sebastien, Evangelista, Heloisa Beatriz Antoniazi, FitzJohn, Rich, Friendly, Michael, Furneaux, Brendan, Hannigan, Geoffrey, Hill, Mark O., Lahti, Leo, McGlinn, Dan, Ouellette, Marie-Helene, Cunha, Eduardo Ribeiro, Smith, Tyler, Stier, Adrian, Braak, Cajo J. F. Ter & Weedon, James. *vegan: Community Ecology Package*, (2022). R package version 2.6-4.
38. Prasad, Kavya *et al.* Whole-genome duplication shapes the aneuploidy landscape of human cancers. *Can. Res.* **82**(9), 1736–1752 (2022).
39. Raynal, Louis *et al.* ABC random forests for Bayesian parameter inference. *Bioinformatics* **35**(10), 1720–1728 (2019).
40. Salehi, Sohrab *et al.* Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature* **595**(7868), 585–590 (2021).
41. Schuhmacher, Dominic, Bähre, Björn, Gottschlich, Carsten, Hartmann, Valentin, Heinemann, Florian & Schmitzer, Bernhard. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, (2023). R package version 0.14-6.
42. Schwarz, Roland F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**(4), e1003535 (2014).
43. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948).
44. Sisson, Scott A., Fan, Yanan & Beaumont, Mark. *Handbook of approximate Bayesian computation* (CRC Press, 2018).
45. Smith, Martin R. Information theoretic Generalized Robinson-Foulds metrics for comparing phylogenetic trees. *Bioinformatics* **36**(20), 5007–5013 (2020).
46. Smith, Martin R. *TreeDist: Distances between Phylogenetic Trees*, (2020). R package version 2.7.0.
47. Tarabichi, Maxime *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**(2), 144–155 (2021).
48. Tavaré, Simon, Balding, David J., Griffiths, Robert C. & Donnelly, Peter. Inferring coalescence times from DNA sequence data. *Genetics* **145**(2), 505–518 (1997).
49. Toni, Tina & Stumpf, Michael PH. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**(1), 104–110 (2010).
50. Tourdot, Richard W., Brunette, Gregory J., Pinto, Ricardo A. & Zhang, Cheng-Zhong. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**(1), 1–31 (2021).
51. Valind, Anders, Jin, Yuesheng & Gisselsson, David. Elevated tolerance to aneuploidy in cancer cells: estimating the fitness effects of chromosome number alterations by in silico modelling of somatic genome evolution. *PLoS ONE* **8**(7), e70445 (2013).
52. Vaserstein, Leonid Nisonovich. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii* **5**(3), 64–72 (1969).
53. Vasudevan, Anand *et al.* Aneuploidy as a promoter and suppressor of malignant growth. *Nat. Rev. Cancer* **21**(2), 89–103 (2021).
54. Vázquez-García, Ignacio *et al.* Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* **612**(7941), 778–786 (2022).
55. Willmott, Cort J. *et al.* Statistics for the evaluation and comparison of models. *J. Geophys. Res. Oceans* **90**(C5), 8995–9005 (1985).
56. Zack, Travis I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**(10), 1134–1140 (2013).

Acknowledgements

The authors acknowledge the support from the Herbert and Florence Irving Institute for Cancer Dynamics and Department of Statistics at Columbia University.

Author contributions

K.D. conceived the presented concept and methodology; Z.X., Z.L. and K.D. wrote the code; Z.X. and Z.L. carried out the analysis and wrote the original manuscript draft; K.D. supervised, reviewed and edited the final manuscript draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67842-9.

Correspondence and requests for materials should be addressed to K.N.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.