Abstract

1 Introduction

2 Results

2.1 Inference method

We have recently introduced CINner, an algorithm capable of simulating tumorigenesis driven by different mechanisms of chromosomal instability (CIN), including whole-genome duplications, whole-chromosome missegregations and chromosome-arm missegregations [Dinh et al., 2024a]. We further evaluated the signals of different statistics measurable from bulk and single-cell DNA data in a synthetic setting, toward inferring both occurrence rates of different CIN mechanisms and chromosome-specific selection parameters that together shape the selection landscape of chromosomally instable tumors [Xiang et al., 2024]. Finally, we developed Approximate Bayesian Computation Sequential Monte Carlo via random forests (ABC-SMC-RF) [Dinh et al., 2024b], a likelihood-free inference method [Diggle and Gratton, 1984] that is robust to statistical noise and hyper-parameters due to the use of RF [Raynal et al., 2019] and converges quickly to the posterior distribution thanks to the incorporation within the SMC framework [Beaumont et al., 2009]. Here, we take advantage of these previous works, toward constructing a numerical method to quantify parameters governing CIN events and impact to cell fitness from single-cell genomic data.

ZIJIN: Details about inference algorithm for chrom level

Building on this framework, we now apply the model to investigate the tumorigenesis process in high-grade serous ovarian cancer (HGSOC), particularly in cases characterized by homologous recombination deficiency and enriched with small tandem duplications (HRD-Dup).

In the first iteration, ABC-SMC-RF performs original ABC-rf with our chosen summary statistics. In further iterations, parameters are sampled from previous posterior distributions and then perturbed with Markov kernels. These updated parameters are subsequently used in ABC-rf to infer the posterior distributions.

In current framework, the effects of whole-genome duplication (WGD) are not considered in our inference. Summary statistics are computed from copy number (CN) profiles, CNA event lists, and phylogenetic trees derived from preprocessed patient data containing only non-WGD cells. These statistics build upon our previous framework [Xiang et al., 2024] and have been refined to fit the specific requirements to our study. As before, we divide the summary statistics into three key groups (**Fig. 1B**).

The scDNA CN group focuses on capturing variation and overall copy number changes derived from CN profiles. We retain the previous methods for calculating both the Wasserstein distance [Kantorovich, 1939, Vaserstein, 1969] and Shannon diversity [Shannon, 1948] within scDNA data [Xiang et al., 2024]. To measure subclonal divergence, we construct a nested distance matrix by comparing CN profiles of individual cells across simulated and experimental data. Optimal transport is applied to this matrix to quantify pairwise subclone-to-subclone distances, and a final optimal transport step aggregates these distances into a cohort-level scDNA-seq divergence metric. To improve scalability for large scDNA-seq datasets, we optimized the computational pipeline by focusing on the subclonal architecture. The Shannon diversity [Shannon, 1948] provides a robust measure for genetic complexity in scDNA sequencing. It is calculated using the total number of subclones and proportion of cells assigned to each subclone. Higher Shannon index values reflect greater clonal heterogeneity, while lower values indicate reduced diversity. We classify all CNA events into gains and losses of copy number and compute the

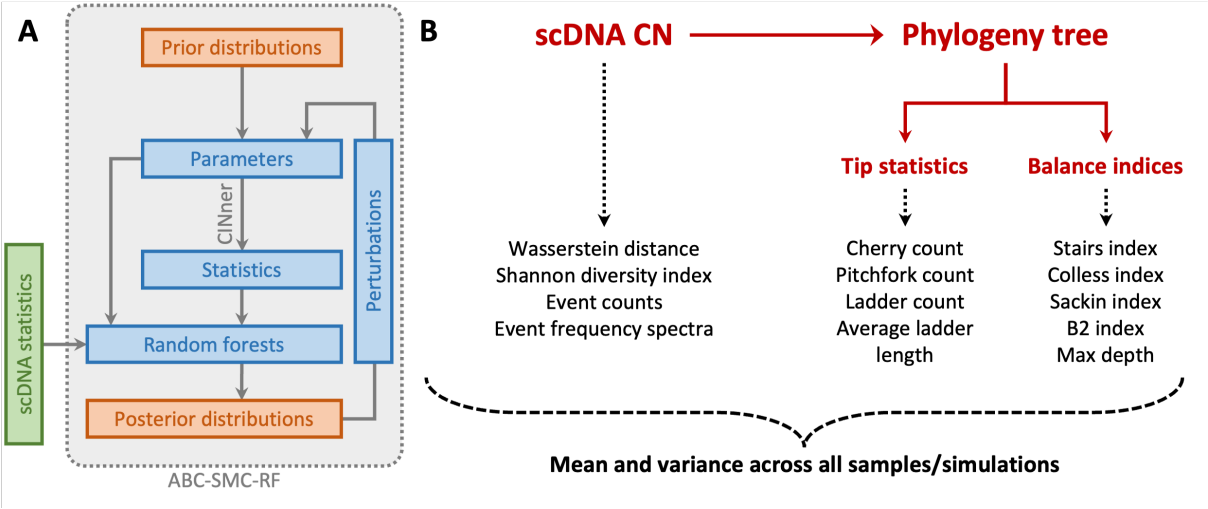

Figure 1: Overview of the computational method to infer chromosomal instability (CIN) probabilities and chromosome selection coefficients from single-cell DNA-sequencing (scDNA) data. **A**: Schematic of the inference method, based on Approximate Bayesian Computation sequential Monte Carlo via random forests (ABC-SMC-RF) and CINner. **B**: Summary statistics employed to compare scDNA data and CINner simulations, based on the copy number (CN) profiles, or tip-related statistics and balance indices of the phylogeny trees. ABC-SMC-RF compares the means and variances of these statistics across simulations and the scDNA cohort to identify likelier parameter sets.

⟨fig_methods⟩

proportion of cells exhibiting each event. By grouping these events based on their frequency and type, we derive the event frequency spectra, indicating whether they occur early or late in tumor evolution. Additionally, we count the total number of each type of CNA events across individual chromosome, offering further insights into chromosomal instability patterns.

As in the previous framework, there are two groups of phylogenetic statistics to compare the phylogeny trees inferred from experimental data using MEDICC2 [Kaufmann et al., 2022] with those simulated by CINner [Dinh et al., 2024a]. The tip statistics focus on tree leaf structures, including cherries counts, pitchforks counts, ladders counts, and the average ladder length, which provides information about the connectivity and branching patterns within the trees. The balance statistics measure the overall structure balance of the tree using metrics such as maximum depth, stairs, Sackin, and Colless indices, alongside the B2 balance index. Together, these metrics capture both localized and global features of tree topology. Detailed descriptions of these statistics can be found in our previous work [Xiang et al., 2024].

2.2 Results for DLP-Signature HRD-Dup (n=6)

ZIJIN: Discuss goodness of fit

Comparison of p_{misseg} and selection coefficients to the literature

Biological validations:

- Ovarian serous cystadenocarcinomas preferentially loses small chromosomes [Duijf et al., 2013]
- `https://bionumbers.hms.harvard.edu/bionumber.aspx?&id=105284&ver=7`
- `https://bionumbers.hms.harvard.edu/bionumber.aspx?s=n&v=11&id=105290`

- https://bionumbers.hms.harvard.edu/bionumber.aspx?id=112099&ver=12&trm=loss+rate&org=
- https://bionumbers.hms.harvard.edu/bionumber.aspx?id=105283&ver=2&trm=missegregation+rate&org=
- https://bionumbers.hms.harvard.edu/bionumber.aspx?id=105319&ver=4&trm=missegregation+rate&org=
- Predicting CIN rates from single-cell whole genome sequencing data using an in silico model: Bjorn Bakker, Michael Schubert, Ana C.F. Bolhaqueiro, Geert J.P.L. Kops, Diana. C.J. Spierings, Floris Foijer bioRxiv 2023.02.14.528596; doi: https://doi.org/10.1101/2023.02.14.528596

  [New]
- https://doi.org/10.1371/journal.pgen.1006707
- https://doi.org/10.1016/S0968-6053(03)00050-4
- https://aacrjournals.org/cancerres/article/63/12/3378/510095/Ovarian-Carcinoma-Develops-through-Multiple-Mod
- https://aacrjournals.org/cancerres/article/69/9/4036/553394/Analysis-of-DNA-Copy-Number-Alterations-in-Ovari
- https://onlinelibrary.wiley.com/doi/full/10.1111/jcmm.17893
- https://aacrjournals.org/cebp/article/28/7/1117/72093/Genome-wide-Analysis-of-Common-Copy-Number
- https://onlinelibrary.wiley.com/doi/10.1002/ijc.32288
- https://www.nature.com/articles/6600896#Fig1

www

3 Discussion

Code availability

Acknowledgments

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

Author contributions

Competing interests

The authors declare no competing interests.

References

adaptive_2009 M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96:983–990, 2009.

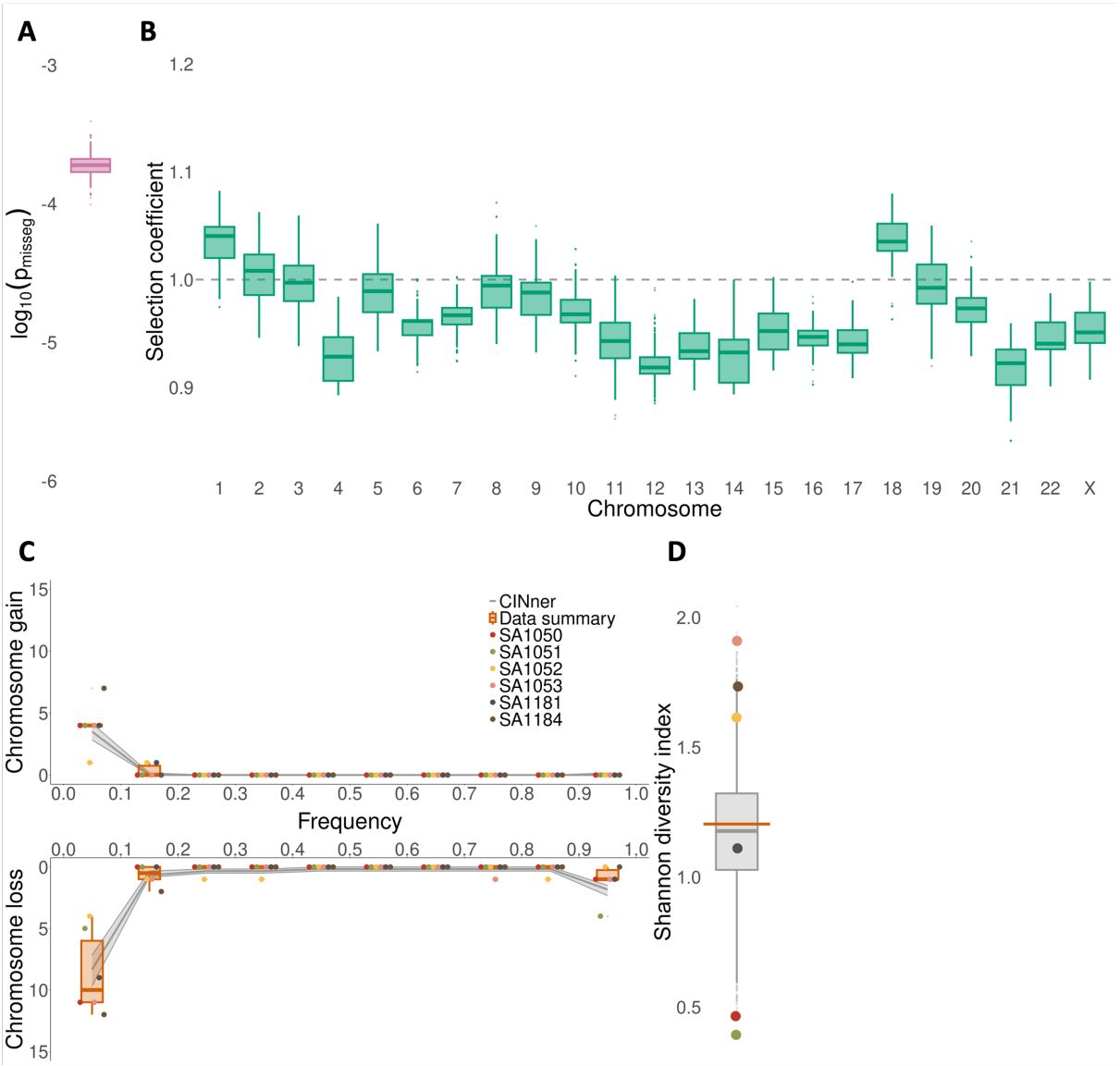

Figure 2: Inference of chromosomal instability in HGSOC HRD-Dup [Funnell et al., 2022] ($n = 6$). **A-B**: Posterior distributions of chromosome missegregation probability (**A**) and chromosome-specific selection coefficients (**B**). **C**: Spectra of chromosome gain and loss counts, with frequency bin = 0.1. Median, 25% and 75% quantiles of CINner simulations from the posterior distribution are plotted in gray. Distribution from data is plotted as orange boxplots at each bin, with sample-specific counts as dots. **D**: Shannon diversity index (**D**) from the posterior distribution (gray boxplots), against measurement from [Funnell et al., 2022] (orange circles).

results_sig⟩?

de2023stage K. De Decker, H. H. Wenzel, J. Bart, M. A. van der Aa, R. F. Kruitwagen, H. W. Nijman, and A.-J. Kruse. Stage, treatment and survival of low-grade serous ovarian carcinoma in the netherlands: A nationwide study. *Acta Obstetricia et Gynecologica Scandinavica*, 102(3):246–256, 2023.

haracterizing S. C. Dentro, I. Leshchiner, K. Haase, M. Tarabichi, J. Wintersinger, A. G. Deshwar, K. Yu, Y. Rubanova, G. Macintyre, J. Demeulemeester, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, 184(8):2239–2254, 2021.

ggle1984monte P. J. Diggle and R. J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(2):193–212, 1984.

CINner K. N. Dinh, I. Vázquez-García, A. Chan, R. Malhotra, A. Weiner, A. W. McPherson, and S. Tavaré. Cinner: modeling and simulation of chromosomal instability in cancer at single-cell resolution. *bioRxiv*, 2024a. doi: 10.1101/2024.04.03.587939. URL https://www.biorxiv.org/content/early/2024/04/03/2024.04.03.587939.

abcsmcrf K. N. Dinh, Z. Xiang, Z. Liu, and S. Tavaré. Approximate bayesian computation sequential monte carlo via random forests, 2024b. URL https://arxiv.org/abs/2406.15865.

ijf2013cancer P. H. Duijf, N. Schultz, and R. Benezra. Cancer cells preferentially lose small chromosomes. *International journal of cancer*, 132(10):2316–2326, 2013.

ell2022single T. Funnell, C. H. O'Flanagan, M. J. Williams, A. McPherson, S. McKinney, F. Kabeer, H. Lee, S. Salehi, I. Vázquez-García, H. Shi, et al. Single-cell genomic variation induced by mutational processes in cancer. *Nature*, 612(7938):106–115, 2022.

9mathematical L. V. Kantorovich. The mathematical method of production planning and organization. *Management Science*, 6 (4):363–422, 1939.

nn2022medicc2 T. L. Kaufmann, M. Petkovic, T. B. Watkins, E. C. Colliver, S. Laskina, N. Thapa, D. C. Minussi, N. Navin, C. Swanton, P. Van Loo, et al. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome biology*, 23(1):241, 2022.

sler2015notch M. Kessler, K. Hoffmann, V. Brinkmann, O. Thieck, S. Jackisch, B. Toelle, H. Berger, H.-J. Mollenkopf, M. Mangler, J. Sehouli, et al. The notch and wnt pathways regulate stemness and differentiation in human fallopian tube organoids. *Nature communications*, 6(1):8989, 2015.

kim2018cell J. Kim, E. Y. Park, O. Kim, J. M. Schilder, D. M. Coffey, C.-H. Cho, and R. C. Bast Jr. Cell origins of high-grade serous ovarian cancer. *Cancers*, 10(11):433, 2018.

raynal2019abc L. Raynal, J.-M. Marin, P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728, 2019.

shannon C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

ein1969markov L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

2024inference Z. Xiang, Z. Liu, and K. N. Dinh. Inference of chromosome selection parameters and missegregation rate in cancer from dna-sequencing data. *Scientific Reports*, 14(1):17699, 2024.

A Methods

A.1 Analysis of single-cell DNA data

Description of DLP+ signature data

We analyze single-cell DNA data produced from Direct Library Preparation+ (DLP+), previously reported in [Funnell et al., 2022] and is publicly available at `https://zenodo.org/record/6998936`. We study a cohort consisting of high-grade serous ovarian cancer (HGSOC) samples classified with homologous recombination deficiency and enriched in small tandem duplications (HRD-Dup; $n = 6$; ID: SA1050, SA1051, SA1052, SA1053, SA1181, SA1184).

Defining single cells as WGD or non-WGD

For every cell in each DLP+ sample, we compute its ploidy (i.e. average CN) and fraction of the genome with loss of heterozygosity (LOH, i.e. fraction of the genome with Minor CN = 0), then apply the empirical finding in [Dentro et al., 2021] to classify the cell as harboring whole-genome duplication (WGD) if ploidy $> 2.9 - 2\times$LOH fraction, and non-WGD if ploidy $\leq 2.9 - 2\times$LOH fraction (**Fig. 3A**).

Current strategy of finding missegregations and defining clones

We infer the phylogenetic tree for the non-WGD cells in each sample from their allele-specific CN profiles (**Fig. 3B-G**) with MEDICC2 [Kaufmann et al., 2022]. We then find whole-chromosome and chromosome-arm events in the non-WGD cells, following the approach in [Funnell et al., 2022]. For each allele of each chromosome in each cell, we find the difference Δ_b between the cell's allele CN and the normal CN (=1) in each bin b. If a particular offset value $\Delta \neq 0$ covers at least 70% of the chromosome, we classify the chromosome as being missegregated. If not, we further check if Δ covers at least 70% of one chromosome arm, in which case we conclude that the arm was missegregated. The event count is $|\Delta|$, and the event(s) are categorized as loss(es) if $\Delta < 0$ and gain(s) if $\Delta > 0$. A clone at the chromosome level is defined as cells with the same whole-chromosome missegregation events. Arm-level clones consist of cells with the same whole-chromosome and chromosome-arm missegregations.

Eventual strategy of finding missegregations and defining clones

We then infer the phylogenetic tree for the non-WGD cells from their total CN profiles with MEDICC2 [Kaufmann et al., 2022] (**Fig. 3B-G**). MEDICC2 annotates each node in the tree rooted in a diploid profile with copy number aberration (CNA) events, such that the leaves recover the observed CN profiles and the total CNA count in the tree is minimal, under parsimony principle. We classify each CNA as a whole-chromosome missegregation if it covers at least 80% the chromosome's length, or a chromosome-arm missegregation if its width spans at least 80% the length of the arm that the CNA resides on. Other events are classified as focal.

To define clonality in each sample at the chromosome level, we characterize each cell by the whole-chromosome missegregations that occurred during its evolution from the diploid ancestor. Cells in the same clone have the same missegregation events, and therefore the same CN profile at the chromosome level. Clones at the chromosome-arm level are similarly defined, but accounting for both whole-chromosome and chromosome-arm missegregations.

A.2 Simulation of tumor copy number evolution

We simulate the tumor clonal evolution with CINner [Dinh et al., 2024a, Xiang et al., 2024]. Each simulation begins at $t = 0$ with normal diploid female cells (total CN = 2 for chromosomes $1, \ldots, 22$, X). Each cell division can harbor a missegregation of a random chromosome with probability p_{misseg}, or a chromosome arm with probability

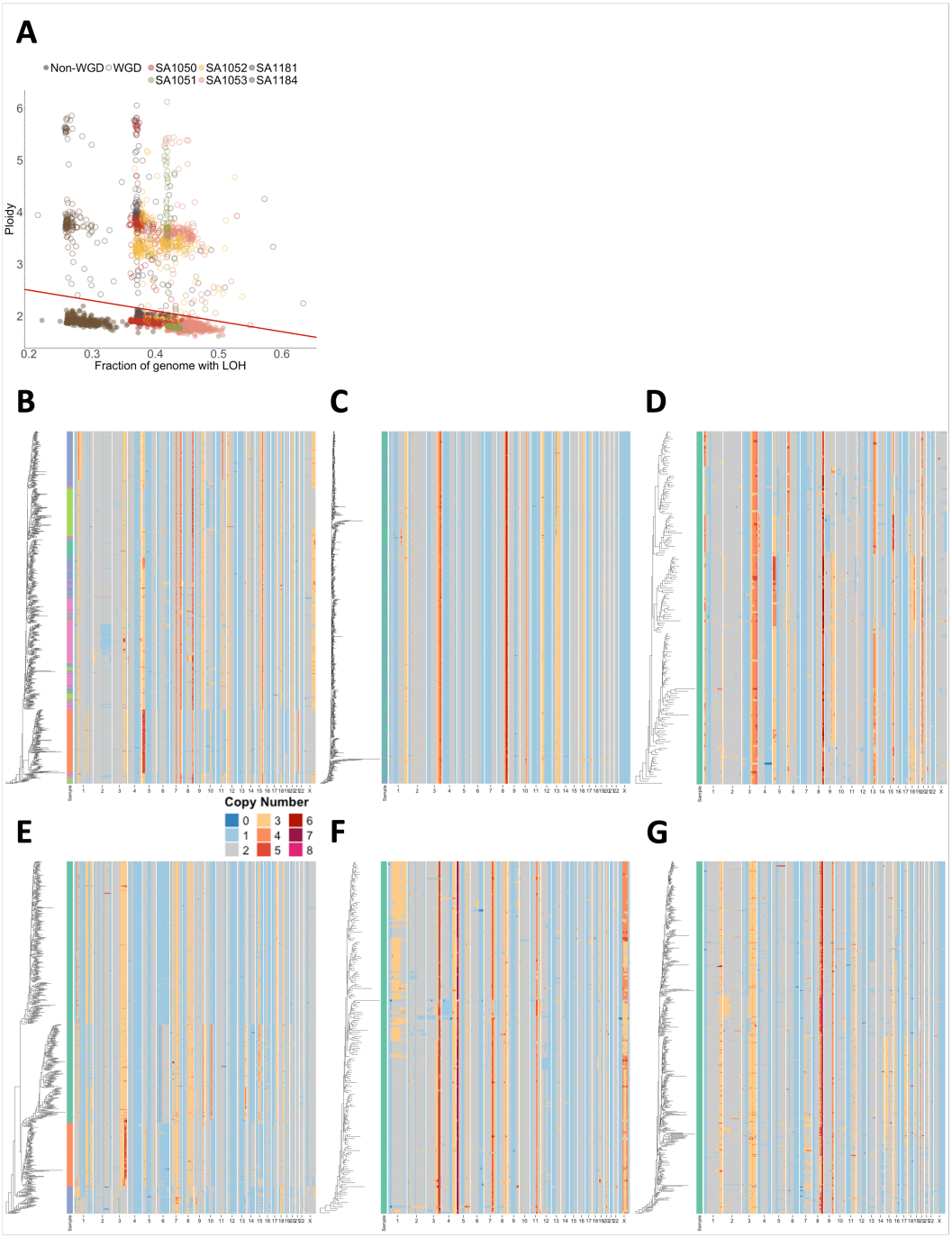

Figure 3: Analysis for 6 samples of HGSOC HRD-Dup [Funnell et al., 2022]. **A**: Distribution of ploidy and fraction of genome with loss of heterozygosity (LOH) in cells from all samples, divided by ploidy $= 2.9 - 2\times$LOH fraction (red line). Cells above and below this line are categorized as WGD and non-WGD, respectively. **B-G**: Copy number heatmaps with phylogeny inferred by MEDICC2 [Kaufmann et al., 2022], for samples SA1050 (**B**), SA1051 (**C**), SA1052 (**D**), SA1053 (**E**), SA1181 (**F**) and SA1184 (**G**).

⟨fig_data_sig⟩

$p_{\text{arm-misseg}}$. A cell's fitness is defined as

$$f = \prod_{\text{arm } i \in \{1\text{p},1\text{q},\ldots,\text{Xp},\text{Xq}\}} s_i^{c_i/2} \quad (1)$$ fitness-arm

where c_i is the CN of arm i in the cell, and s_i is the arm's selection coefficient. As nullisomy is nonviable, $f = 0$ if $c_i = 0$ for any i.

CINner settings for HGSOC

We modify CINner parameters to resemble clinical aspects of HGSOC. Previous works have found that HGSOC originates in the fallopian tube epithelium [Kim et al., 2018], which has the doubling time of 3.5-5 days [Kessler et al., 2015]. Therefore, the lifespans of cells in our simulations are exponentially distributed with mean 5 days, after which they either divide or die with probabilities proportional to their fitness. We assume the total cell population follows logistic growth, with carrying capacity of 10^4 cells, growth rate 0.3 and midpoint at 20 years. The DLP+ data is then compared to the CN profiles of 1,000 randomly selected cells in each simulation at $t = 63$ years, the average age at diagnosis for women with HGSOC [De Decker et al., 2023].

A.3 Inference of chromosome missegregation probability and selection coefficients

We modify our previous inference method in [Xiang et al., 2024] to find $\theta = \{p_{\text{misseg}}, s_1, \ldots, s_{22}, s_{\text{X}}\}$, missegregation probability and chromosome selection coefficients, from the DLP+ data. Because here we focus on whole-chromosome events, the arms of each chromosome share the same copy number, therefore the formula for a cell's fitness can be reduced from Eq. 1 to

$$f = \prod_{\text{chromosome } i \in \{1,\ldots,\text{X}\}} s_i^{c_i/2}$$

ABC-SMC-RF

The inference method for θ is based on Approximate Bayesian Computation sequential Monte Carlo via random forests (ABC-SMC-RF), a likelihood-free inference method that we previously introduced in [Dinh et al., 2024b] (**Fig. 1A**). The first iteration in ABC-SMC-RF samples N_1 parameter sets from the prior distribution, then computes statistics for each parameter set with CINner simulations. It then trains a random forest for each parameter, based on the corresponding sampled values and the resulting statistics. The posterior distribution for the parameter is then predicted by applying DLP+ statistics to the forest and extracting weights for the parameter values. In each successive iteration $t = 2, \ldots, T$, ABC-SMC-RF samples N_t parameter sets from the previous posterior distributions, then perturbs them with adaptable normal distribution kernels based on [Beaumont et al., 2009]. CINner simulations and random forest predictions then follow similarly to iteration 1. In this work, we implement $T = 10$ iterations, each with $N_t = 10^4$ parameter sets and prior distributions:

$$\log_{10}(p_{\text{misseg}}) \sim \text{Uniform}(-6, -3)$$
$$s_1, \ldots, s_{22}, s_{\text{X}} \sim \text{Uniform}\left(\frac{1}{1.2}, 1.2\right)$$

Statistics for inference

We characterize each DLP+ sample and CINner simulation with several statistics (**Fig. 1B**). For each parameter set, ABC-SMC-RF makes 6 CINner simulations, then summarizes them with the means and variances of each statistic. The prediction phase is then performed with the same statistics for the 6 DLP+ samples.

The tip statistics and balance indices for DLP+ samples are measured from the MEDICC2-inferred phylogeny trees, while for CINner we utilize the simulated cell phylogeny. The tip statistics include normalized counts of cherries, pitchforks, ladders, and average ladder lengths. The balance indices consist of maximum phylogeny depth and stairs, Colless, Sackin and B2 indices. See [Xiang et al., 2024] for detailed descriptions.

On the other hand, the CN statistics characterize directly the similarity among the genotypes observed in the DLP+ samples and CINner simulations. The Shannon diversity index and numbers of all/gain/loss missegregations can be measured for each sample. Furthermore, we compute the Wasserstein distance between the DLP+ cohort and CINner simulations under a given parameter set, where the distance between each DLP+ sample and a simulation is measured with the Euclidean metric [Xiang et al., 2024]. Finally, we also compute the missegregation frequency spectrum $\{S(b_{j-1}, b_j]\}_{j=1,\dots,10}$ for each sample, where $S(b_{j-1}, b_j]$ is the number of missegregations whose frequencies are within interval $(b_{j-1}, b_j]$, with $b_j = \frac{j}{10}$. The counts and frequency spectra are measured separately for all missegregations, only gains, and only losses.

The ABC-SMC-RF algorithm iteratively infers the posterior distribution for each parameter [Dinh et al., 2024b]. To limit the effect of noise, we select specific statistics for each parameter:

- For missegregation probability p_{misseg}: phylogeny balance indices, CN-based Wasserstein distance, Shannon diversity index, and missegregation gain/loss frequency spectra are used. All statistics are measured across the entire genome.
- For selection coefficient s_i of chromosome i: genome-wide phylogeny balance indices are used, in addition to CN-based Wasserstein distance, Shannon diversity index and missegregation gain/loss counts measured specifically for chromosome i.

Statistics for validation

- PCA for CINner simulations, DLP+ samples, and TCGA/ICGC samples
- Predict fitness of clones in DLP+ (DLP-signature & DLP-fitness?) $\rightarrow$ compare with inferred selection rates in DLP-fitness or proliferation rates from SPRINTER
- Compare selection coefficients with driver genes (Cancer Gene Consensus, etc)

A.4 Inference of chromosome-arm missegregation probability and selection rates

A.5 Inference of whole-genome duplication probability and associated increase in chromosomal instability