CNETML: Maximum likelihood inference of tumour phylogeny from copy number profiles of spatio-temporal samples

Bingxin Lu[1,2], Kit Curtius[3], Trevor Graham[3], Ziheng Yang[4], and Chris P. Barnes*[1,2]

[1]Department of Cell and Developmental Biology, University College London, UK
[2]UCL Genetics Institute, University College London, UK
[3]Barts cancer institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[4]Department of Genetics, Evolution and Environment, University College London, London, UK

Abstract

Phylogenetic trees based on copy number alterations (CNAs) for multi-region samples of a single cancer patient are helpful to understand the spatio-temporal evolution of cancers, especially in tumours driven by chromosomal instability. Due to the high cost of deep sequencing data, low-coverage data are more accessible in practice, which only allow the calling of (relative) total copy numbers due to the lower resolution. However, methods to reconstruct sample phylogenies from CNAs often use allele-specific copy numbers and those using total copy number are mostly distance matrix or maximum parsimony methods which do not handle temporal data or estimate mutation rates. In this work, we developed a new maximum likelihood method based on a novel evolutionary model of CNAs, CNETML, to infer phylogenies from spatio-temporal samples taken within a single patient. CNETML is the first program to jointly infer tree topologies, node ages, mutation rates, and ancestral states from total copy numbers when samples were taken at different time points. Our extensive simulations suggest CNETML performed well even on relative copy numbers with subclonal whole genome doubling events and under slight violation of model assumption. The application of CNETML to real data from Barrett's esophagus patients also generated consistent results with previous discoveries and novel early CNAs for further investigations.

Keywords

tumour phylogeny; copy number alteration; maximum likelihood; model of evolution; low-coverage sequencing

1 Introduction

Phylogenetic trees have been widely used in the study of cancer, providing important insights into tumour evolution [1]. Various markers have been used for phylogeny inference, including data derived from comparative genomic hybridisation (CGH), single nucleotide polymorphism (SNP) array, fluorescence in situ hybridization (FISH), and next-generation sequencing (NGS) technologies. The rapid advances of NGS, such as whole genome sequencing (WGS) and whole exome sequencing (WES), allow the generation of huge amounts of genomic data from patient samples. NGS-derived somatic variants, mainly single nucleotide variants (SNVs) and copy number alterations (CNAs), have become common markers for phylogeny inference. CNAs are more complex than SNVs and often related to chromosomal instability (CIN) which may generate different types of structural variations (SVs) or

*christopher.barnes@ucl.ac.uk

aneuploidy [2]. Although most phylogeny inference approaches use SNVs, a number of methods are based solely on CNAs [1, 3–10]. One reason is that it is hard to detect point mutations for some cancers mainly driven by SV or CIN [11], such as high grade serous ovarian cancer [12], oesophageal cancer [13], and osetosarcoma [14]. Another reason is that it is difficult to detect SNVs from low-coverage data whereas the larger sizes of CNAs provide more signal for reliable detection.

Given different input data and aims, the trees reconstructed from CNAs called from a single patient are of four major types: 1) mutation tree when the order and evolutionary history of mutational events are of interest [15], as in SCICoNE [5] and CONET [7], where each tip represents copy number events and cells are attached to each node; 2) clone tree when clonal deconvolution is feasible, as in CNT-MD [4] and DEVOLUTION [8], where each tip represents a clone; 3) cell tree when CNAs can be called for each cell, as in FISHtree [16, 17], sitka [6], and NestedBD [10], where each tip represents a cell; 4) sample tree when each sample is assumed to be homogeneous, as in MEDICC [18], MEDICC2 [9] and PISCA [3], where each tip represents a sample.

Intra-tumour heterogeneity (ITH) causes difficulty in analysing bulk DNA sequencing (bulk-seq) data, where only the aggregated signals can be observed. Therefore, phylogeny inference from bulk-seq data is often coupled with clonal deconvolution that determines the number and fraction of clones in a sample [1]. Reliable quantification of subclonal CNAs and ITH requires deep sequencing on samples of good quality and is expensive. Single cell DNA sequencing (sc-seq) circumvents the need to infer clone structure, but the data are still very noisy [15, 19]. Low-coverage bulk-seq, such as shallow WGS (sWGS), are instead more cost-effective and accessible, especially for SV-driven cancers [20]. They have been widely applied to detect CNAs, particularly on formalin-fixed paraffin-embedded (FFPE) samples, which are commonly available for diagnostics but have low DNA quality [11, 13, 21–23], and cell-free DNA in plasma [24, 25]. There have been sWGS data for patient samples taken over time and space during a long time period, such as in the surveillance of Barrett's esophagus (BE) [13]. These longitudinal samples also provide temporal information to estimate node ages and mutation rates, which are important parameters in cancer evolution. However, only a few reliable methods exist to detect CNAs from sWGS data, especially absolute copy numbers [26]. Most of the previous sample phylogeny inference methods are designed for absolute allele-specific integer copy numbers which are often called from SNP arrays and high-coverage NGS data, such as MEDICC [18], MEDICC2 [9], and PISCA [3]. To better understand cancer progression from these sWGS data, it is important to have methods that can build sample trees based solely on (relative) total copy numbers, which will be addressed in this paper.

The model of CNA evolution is critical for phylogeny inference, but it is challenging to propose a model which maintains a good trade-off between biological relevance and complexity [19]. The underlying mechanisms of CNAs are often very complicated, such as chromothripsis, breakage fusion bridges, and failure in cell cycle control [22]. As a result, CNAs vary from small focal duplication/deletion to chromosome-level gain/loss and whole genome doubling (WGD) at different rates [27], which creates complex dependencies across the genome, such as overlaps, back mutations, convergent and parallel evolution [28]. Therefore, the infinite sites or perfect phylogeny assumption, which is commonly used in inferring tumour phylogeny from SNVs, is often violated, as is the infinite alleles or multi-state perfect phylogeny assumption [19]. The models for genome rearrangement, microsatellite, and multigene families seem relevant yet hard to transfer to CNAs [19, 29].

Some methods transform original copy number calls into presence or absence of changes (breakpoints) [6, 30], which are less likely to overlap, so that the infinite sites assumption is well approximated. Although this representation simplifies the complex spatial correlations across sites, it does not use the full copy number data. Other methods represent the genome as a vector of copy number values, often called copy number profile (CNP) [4]. Based on CNPs, some methods build trees without a model, such as the maximum parsimony (MP) method with the Fitch algorithm [23, 31] and distance matrix methods based on Euclidean [32] or Manhattan [33] distance, and hence they may underestimate the true evolutionary distance as no correction of hidden changes is applied [9, 30]. Other methods use copy number transformation models that allow the computation of minimum evolutionary distance

between CNPs, which is the shortest sequence of events that transform one CNP to the other. One such model was implemented in FISHtree [16, 17], which assumes each event (single gene gain/loss, chromosome gain/loss, or WGD) affects a single unit (gene or chromosome or genome) independently, with or without weights for different types of events. Another well-studied model, within MEDICC [18], assumes an event (segment duplication/deletion) may affect contiguous segments of variable size. This model deals with horizontal dependencies caused by overlapping CNAs and hence is less likely affected by convergent evolution. It has been extended to allow weights on CNAs of different position, size, and type (duplication/deletion) [34] and WGD [9, 35]. The weighted versions of both models allow the estimation of CNA rates in term of event probabilities [17, 34], but mutation rates by calendar time cannot be estimated. A few CNP-based methods use the finite sites models, or continuous-time Markov chains, which have good theoretical properties and are frequently used to model nucleotide changes [36]. Although Markov model often assumes independent sites to simplify computation, which is violated by overlapping CNAs, it corrects multiple hits at the same site and serves as a workable model of CNAs. For example, SCONCE used a Markovian approximation that combines the temporal Markov process with a spatial hidden Markov model (HMM) to detect CNAs in sc-seq data [37]; Elizalde et al. used the product of 23 Markov chains to model numerical CIN of individual chromosomes in clonally expanding populations [38]. Markov model makes it possible to use statistical methods to infer CNA-based trees, mutation rates by time, and ancestral genomes, such as maximum likelihood (ML) method and Bayesian method. PISCA used such a Markov chain to model gain, loss, and conversion of haplotype-specific copy numbers called from SNP array data [3]. NestedBD used a birth-death model, a special type of Markov chain where transitions from state i can only go to state $i+1$ or $i-1$, for total copy numbers called from sc-seq data, where a birth (death) event corresponds to copy number amplification (deletion) [10]. Both PISCA and NestedBD are implemented as packages in the popular Bayesian evolutionary analysis platform BEAST [39, 40], and hence are not easily adapted for more bespoke mutation models that will be required for understanding tumour evolution. In addition, most phylogeny inference methods based on CNPs cannot handle multiple scales of chromosomal changes due to the inherent complexity. Notable exceptions are FISHtree, which was designed for FISH data and is not scalable for longer CNPs [16, 17], and MEDICC2 which only considered segment duplication/deletion and WGD but excluded chromosome or arm level gain/loss [9].

In this paper, we developed an approach based on a novel Markov model of duplication and deletion, CNETML, to do maximum likelihood inference of single patient tumour phylogeny from total copy numbers of spatio-temporal samples. To the best of our knowledge, this is the first method to jointly infer tree topology, node ages, and mutation rates of temporal patient samples from total CNPs. CNETML is applicable to haplotype-specific CNPs as well, which is the basis of our model and considered as missing information when total CNPs are taken as input. We also developed a program to simulate CNAs from patient samples, CNETS (Copy Number Evolutionary Tree Simulation), which was used to validate sample phylogeny inference methods. Our programs are mainly implemented in C++, available at `https://github.com/ucl-cssb/cneta`. The results on extensive simulations suggest that CNETML accurately recovered tree topology, node ages, mutation rates, and ancestral CNPs when there were sufficient CNAs present in the data and large time differences among samples. CNETML on total CNPs performed as well as haplotype-specific CNPs when less than 10% of copy-neutral CNAs existed in the simulated data. CNETML also had good performance when applied to relative CNPs from simulated data with subclonal WGDs, which is desirable for applications to sWGS data. Moreover, the simulations suggest CNETML was robust to slight violations of model assumption and that it obtained reasonable inferences on data of typical focal CNA size. We applied CNETML on relative CNPs called from two BE patients in existing literature and obtained results consistent with previous findings and novel early CNAs from reconstructed ancestral CNPs which worth further validations, suggesting the practicability of CNETML.

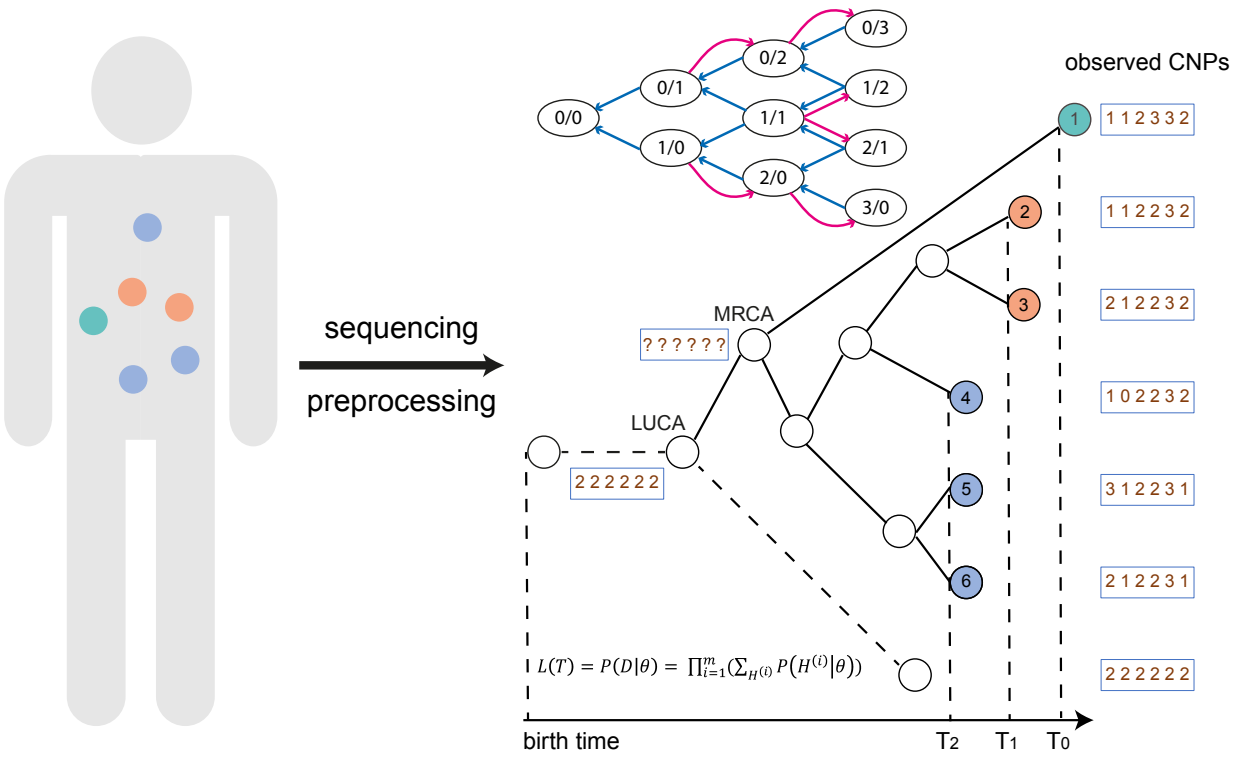

Figure 1: The schematic overview of CNETML. Given the CNPs and/or sampling times of all patient samples, CNETML aims to infer a sample tree in which tips correspond to observed CNPs in samples and internal nodes correspond to ancestral CNPs. From the root, which represents the last unaltered common ancestor with normal copy number state (LUCA), there is a branch of length zero (dashed line), which leads to a tip representing the normal CNP for illustration purpose. LUCA is connected to the most recent common ancestor (MRCA) of the patient samples. We added an additional node before LUCA to show the CNP at the birth time. The state transition diagram of the Markov chain shows the duplication (red arrow) and deletion (blue arrow) of haplotype-specific copy number, with the maximum total copy number being 3. The Markov model allows the computation of tree likelihood by taking the product over all sites along the genome. The CNPs of internal nodes (including MRCA) are unknown and inferred with ancestral reconstruction algorithms. Samples taken at different time points are denoted by different colours.

2 Results

Overview of CNETML

The input of CNETML (Figure 1) includes a set of integer total/haplotype-specific CNPs for multiple samples of a patient and/or sampling timing information (in year) if available (see section 4 for details on input preparation). The length of each CNP is the number of sites in a sampled genome, which can be either bins or segments, and we assume all genomes have the same sites. Here, a bin is a genomic region of fixed size and a segment is a genomic region of variable size which may be obtained by merging consecutive bins with the same copy number. In CNA detection, the general steps include binning, bias removal, segmentation, and copy number assignment [19, 41]. In binning, the genome is divided into bins of certain size, usually fixed, and reads aligned to each bin are counted. In segmentation, the genome is partitioned into a series of segments whose copy number is different from that of the adjacent segment. Therefore, although a site is not as well defined as when modelling SNVs on individual nucleotides, it is feasible to consider a bin or a segment as a site.

We treat an integer copy number at each site as a discrete trait whose states are dependent on the maximum possible copy number. To maintain model simplicity, we assume copy numbers at the sites of a genome change independently of each other (independent sites assumption) and the change of copy number at each site follows a continuous-time non-reversible Markov chain. The Markov chain naturally starts from the normal copy number and has an absorbing state when no copy remains. Due to the difficulty in incorporating CNAs of different scales, we propose a model of site duplication and deletion at haplotype-specific level, which is similar to that in PISCA [3] yet designed for processing total CNPs. Moreover, we consider CNA rate (or mutation rate) per haplotype per site per year and allow user-specified maximum copy number.

Suppose c_{max} is the maximum total copy number, then each site has S possible states $\{0, 1, 2, ..., S-1\}$, where

$$S = \begin{cases} c_{max} + 1 & \text{total CNP input,} \\ \frac{(c_{max}+1)*(c_{max}+2)}{2} & \text{haplotype-specific CNP input.} \end{cases} \quad (1)$$

The change of haplotype-specific copy numbers on each site via duplication (deletion) at rate u (e) per haplotype per site per year is specified by the rate matrix Q (see Supplementary Table 1 for Q at $c_{max} = 4$). In Q, we list haplotype-specific copy numbers in order of increasing total and haplotype A copy number so that each combination of c_A and c_B, (c_A, c_B), corresponds to a unique state, where c_A and c_B represent the copy numbers for two haplotypes respectively. For example, normal copy number (1, 1) is represented by state 4, and copy number (4, 0) is represented by state 14. Suppose a genome j has m sites and c_{ij} is the copy number state at site i, which is either the total copy number or the state corresponding to the haplotype-specific copy number in Q, then its observed CNP is denoted as $(c_{1j}, c_{2j}, ..., c_{mj})$. The CNPs for all the n sampled genomes were converted into a data matrix $D = \{c_{ij}\}$ of n rows and m columns. The observed copy number states across all samples at a site i is called a site pattern, denoted by $s_i = (c_1^i, c_2^i, ..., c_n^i)$. We say site i is invariant if s_i is composed of normal copy number states only, and variant otherwise.

The likelihood for a tree T of n samples with parameters θ, $L(T)$, is the probability of observing D at the tips of T given θ, ~~whose computation is critical in phylogeny inference~~. The Markov model specified by Q allows the computation of $L(T)$ by taking the products of ~~likelihoods~~ at individual sites:

$$L(T) = P(D|\theta) = \prod_{i=1}^{m} P(D^{(i)}|\theta), \quad (2)$$

where $D^{(i)}$ is the i_{th} column of D. When the input are total CNPs, we revise $L(T)$ to incorporate haplotype-specific copy numbers as missing information, which is similar to the handling of ambiguities in a nucleotide substitution model [36]:

$$L(T) = \prod_{i=1}^{m} (\sum_{H^{(i)}} P(H^{(i)}|\theta)), \quad (3)$$

where H is a data matrix of unknown haplotype-specific copy number states that ~~determine~~ D, $H^{(i)}$ is the i_{th} column of H, and there may be multiple such matrices for D. For example, the probability of observing total copy number 3 is a sum over all compatible haplotype-specific copy numbers (0, 3), (1, 2), (2, 1), and (3, 0).

We computed $L(T)$ with Felsenstein's pruning algorithm [42] and a few adaptations which are described in section 4. $L(T)$ was maximised by minimizing its negative logarithm function with L-BFGS-B algorithm [43], a numerical iterative method with bound constraints, ~~which starts with initial guesses of the parameters and improves the estimations iteratively until convergence or the maximum number of iterations is reached~~. Due to the super-exponentially increasing number of trees with the number of tips, we implemented two approaches to search the tree space and get the ML tree. One is exhaustive search which enumerates all the possible tree topologies for trees of less than seven samples.

The other is stochastic search for larger trees, adapted from the approach in IQ-TREE [44], a popular ML phylogeny inference program.

When there are samples taken at different times, it is feasible to estimate mutation rates according to the differences of CNPs and sampling times, similar to the dating of virus divergences [36]. Although mutation rates during tumour progression are likely to change over time due to CIN [3], it is unlikely that they can be estimated reliably, so we assumed constant mutation rates under a global clock for simplicity. We jointly estimated tree topology, mutation rates, and node ages (starting from 0 at birth time) with the following constraints in optimization: 1) Each internal node must be younger than its youngest descents; 2) The root node must be younger than the patient age at the first sample time or the tree height in year is smaller than the patient age at the last sample time. We transformed node age variables to encode the constraints imposed by patient ages at different sampling times so that $\theta = (x_1, x_2, ..., x_n, u, e)$, where x_i is the transformed variable for age of an internal node i and converted back to branch length in year later (see section 4 for more details). When all the samples are taken at the same time, there is no information to estimate mutation rates and node ages at the same time, so $\theta = (l_1, l_2, ..., l_{2n-1})$, where $l_i = (u_0 + e_0) * t_i$ is the length of branch i measured by expected number of CNAs per site, u_0 (e_0) is the pre-specified duplication (deletion) rate per haplotype per site per year, and t_i is the time in year covered by branch i.

Ancestral reconstruction may suggest early CNAs that are likely cancer driver events and useful for early diagnostics, so we reconstructed ancestral states at variant sites with unique site patterns based on the obtained ML tree using classical methods. Both marginal reconstruction of the most recent common ancestor (MRCA) node and joint reconstruction of all ancestral nodes were implemented, ~~which just differ in the number of ancestral nodes in consideration~~ [36]. For marginal reconstruction, we computed the posterior probability of each possible copy number state for MRCA. For joint reconstruction, we used the dynamic programming algorithm in [45], assuming the likelihood of the best reconstruction is when the root has normal copy number states.

We used bootstrapping to measure the uncertainties of an estimated ML tree T_m [36]. To get a bootstrap tree, we sampled sites from the input data matrix D with replacement to get a pseudo-sample D' with the same dimension as D and built a ML tree from D'. The branch support value in T_m is defined as the percentage of bootstrap trees including this branch (split) and computed with function `pro.clade` in R library `ape` [46].

Validation on simulated data

Data simulation and comparison metrics

To validate CNETML, we developed CNETS to simulate CNAs along a phylogenetic tree of multiple patient samples (Figure 2, see section 4 for more details). In CNETS, we first generated a coalescence tree to represent the genealogical relationships among samples, the subtree starting from MRCA, under either the basic coalescent or an exponential growth model with rate β. We then added another node before MRCA to represent the last unaltered common ancestor with normal copy number state (LUCA) and a branch of length zero from LUCA to a new tip which represents a normal genome. The time from LUCA to MRCA was sampled from an exponential distribution with rate which was either based on the exponential growth rate β or sampled from an uniform distribution $\mathcal{U}(0, 1)$. To get different sampling times, we increased the terminal branch lengths by random integer multiples of dt (in year), with the maximum multiple being the number of samples. We implemented two modes of simulating CNPs which differ in the types of CNAs and recorded details. When only site-level CNAs are considered and the exact mutational events are not of interest, CNPs were simulated directly along each branch of the tree according to the rate matrix Q with each site being a segment of variable size [36]. When CNAs of multiple scales are considered, events were simulated by waiting times with each site being a bin of fixed size (4,401 bins of 500 Kbp by default), which allows more complex models of evolution and the recording of more detailed information for each event. CNETS generates files

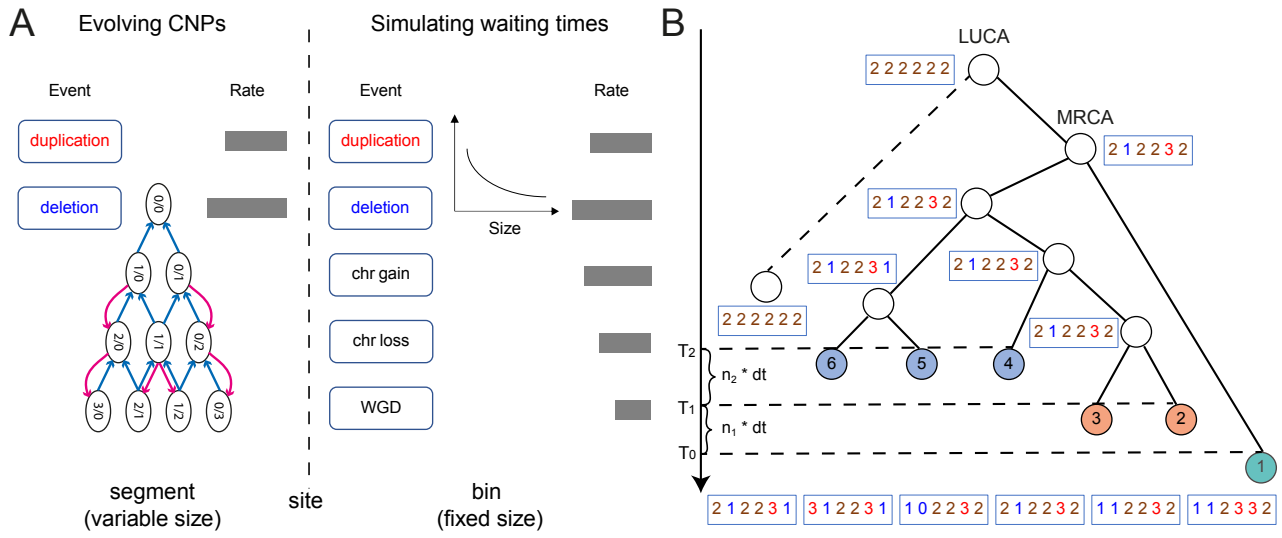

Figure 2: The schematic overview of CNETS. **A**: Two modes of simulation implemented in CNETS. One is simulating CNPs directly for site duplication/deletion based on segments of variable size, which follows exactly the Markov model used for phylogeny inference. The other is simulating waiting times for events of multiple scales based on bins of fixed size, in which at most five types of events (site duplication/deletion, chromosome gain/loss, and WGD) are allowed and the duplication/deletion size in terms of bins is sampled from an exponential distribution of the user-specified mean size. **B**: The simulated tree and CNPs (red: duplication, blue: deletion, brown: normal), where coloured tips represent patient samples taken at different time points.

that record haplotype-specific/total CNPs, sampling times in year, tree topology, and CNAs along the branches respectively. The simulated CNPs at the tips and/or tip timing information serve as input for CNETML.

In tests, we simulated trees with parameters used in [3], which approximate an exponentially expanding haploid cancer cell population with MRCA being 20 years from the present (Supplementary Table 2). To ensure that the model used for simulation and phylogeny inference are the same, we used the simulation mode of evolving CNPs when only site-level mutations were considered. We simulated trees with $n = 5$ samples when not testing the performance of tree searching, as it is fast to enumerate all the possible trees for such small trees. Without loss of generality, we set $c_{max} = 6$ and used the same rates for duplication and deletion. To get a reasonable range of mutation rates suitable for phylogeny inference, we performed tests with $u = e \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ (per haplotype per site per year) (Supplementary Figure 1). This analysis suggests that intermediate rates $\{0.001, 0.01\}$ (per haplotype per site per year) are more informative for phylogeny inference, which were used in our subsequent tests.

To quantify the differences of both topologies and branch lengths between the simulated (true) tree and the inferred tree, we computed normalized Robinson–Foulds (RF) distance [47] and branch score distance [48], with smaller values indicating more accurate estimations. These distances were computed with function `treedist` in R library `phangorn` [49]. The branch length was measured by time in year when computing branch score distance and (u_0, e_0) was set to be real values used for simulation when the mutation rates were not estimated. We also computed the differences between estimated and true rates and LUCA age to check the accuracy of their estimations. To measure the accuracy of ancestral reconstruction, we computed the fraction of correctly recovered states over the number of variant sites with unique site patterns for each internal node and the mean fraction over all internal nodes under joint reconstruction.

Performance on reconstructing trees and ancestral states

In principle, ML phylogeny inference is consistent, which means that it is more likely to recover the true tree given more sites [36]. To check the consistency of CNETML, we applied it on data simulated with different number of sites and mutation rates. To reduce confounding effects, we simulated trees with $n = 5$ samples at the same time and did not infer mutation rates. As shown in Figure 3A, all the simulated trees were better recovered with more sites and higher mutation rates, which confirms the consistency of CNETML. In the subsequent simulations, we fixed the number of sites $m = 1000$.

We showed that the stochastic tree search algorithm performed well on simulated data with 10, 20, and 30 samples under different mutation rates (Supplementary Figure 2). Given more samples, the estimated tree topologies did not deteriorate much, whereas the branch score distances to the true trees were slightly increased. As expected, the reconstructed trees were more similar to the ground truth with more mutations.

We also checked the performance of CNETML on reconstructing ancestral states on the simulated data with 1000 sites under different mutation rates. To reduce the effect of incorrect tree topologies, we supplied the simulated true tree and real mutation rates as input. The results suggest that more than 90% of the unique variant sites were accurately reconstructed, except when doing marginal reconstruction (Figure 3B). Joint reconstruction appeared more accurate, probably because it computes the joint probability of all the internal nodes [36]. The fraction of accurately reconstructed sites decreased with larger mutation rate due to the presence of more variant unique sites.

With total copy numbers, copy-neutral loss of heterogeneity (cn-LOH) and mirrored subclonal allelic imbalance (MSAI) events (CNAs affecting different alleles of the same sites in different samples) cannot be detected. To see how total CNPs deteriorates the inference, we applied CNETML on haplotype-specific CNPs and found that the results were not largely different except when there were more than 10% sites with cn-LOH or MSAI events (Figure 3A,C)). However, the accuracies of reconstructing ancestral states were better with haplotype-specific CNPs (Figure 3B). The analysis on PCAWG dataset [50] shows that around 80% samples have no more than 10% of the genome with cn-LOH (Supplementary Figure 3), and hence these results suggest that total CNPs can provide good approximations in practice despite information loss.

Performance on jointly estimating the tree and mutation rates

One major utility of CNETML is to jointly estimate tree topology, node ages, and mutation rates when the samples were taken at different time points. The reliability of estimations depends on the extent of time differences at the tips, with larger differences providing more information for inference [51]. Since the sampling time differences for a patient may range from one year to 15 years [13], we simulated data under different mutation rates and temporal signal strengths ($dt = 1$ year and $dt = 5$ years), where the range of simulated sampling times approximated real data and samples simulated under a larger dt generally had larger time differences (Supplementary Figure 4). Because the L-BFGS-B optimization algorithm is iterative, the initial values of parameters $\theta^0 = (x_1^0, x_2^0, ..., x_n^0, u^0, e^0)$ are required, where $(x_1^0, x_2^0, ..., x_n^0)$ is derived from the initial tree (see section 4 on how to get initial trees)) and (u^0, e^0) has to be specified manually. To test the sensitivity of inference to initial mutation rates, we tried four initial values, $u^0 = e^0 \in \{0.0005, 0.001, 0.005, 0.01\}$ (per haplotype per site per year), and found that CNETML was robust, except when the real mutation rate was low (0.001) and a high initial mutation rate (0.005 or 0.01) was supplied (Supplementary Figure 5). Therefore, we recommend using smaller initial mutation rates in real data when the range of rates is unknown and reported the results with $u^0 = e^0 = 0.0005$ (per haplotype per site per year) in Figure 4.

As shown in Figure 4A-B, CNETML accurately estimated both tree topology and mutation rates when there was sufficient information in the data, with larger mutation rates or sampling time differences leading to higher accuracies. The estimated median LUCA ages were closer to real values when $dt = 5$ years despite larger variances which is because the variances of simulated LUCA ages were about four times larger than those when $dt = 1$ year. We also ran CNETML on haplotype-specific

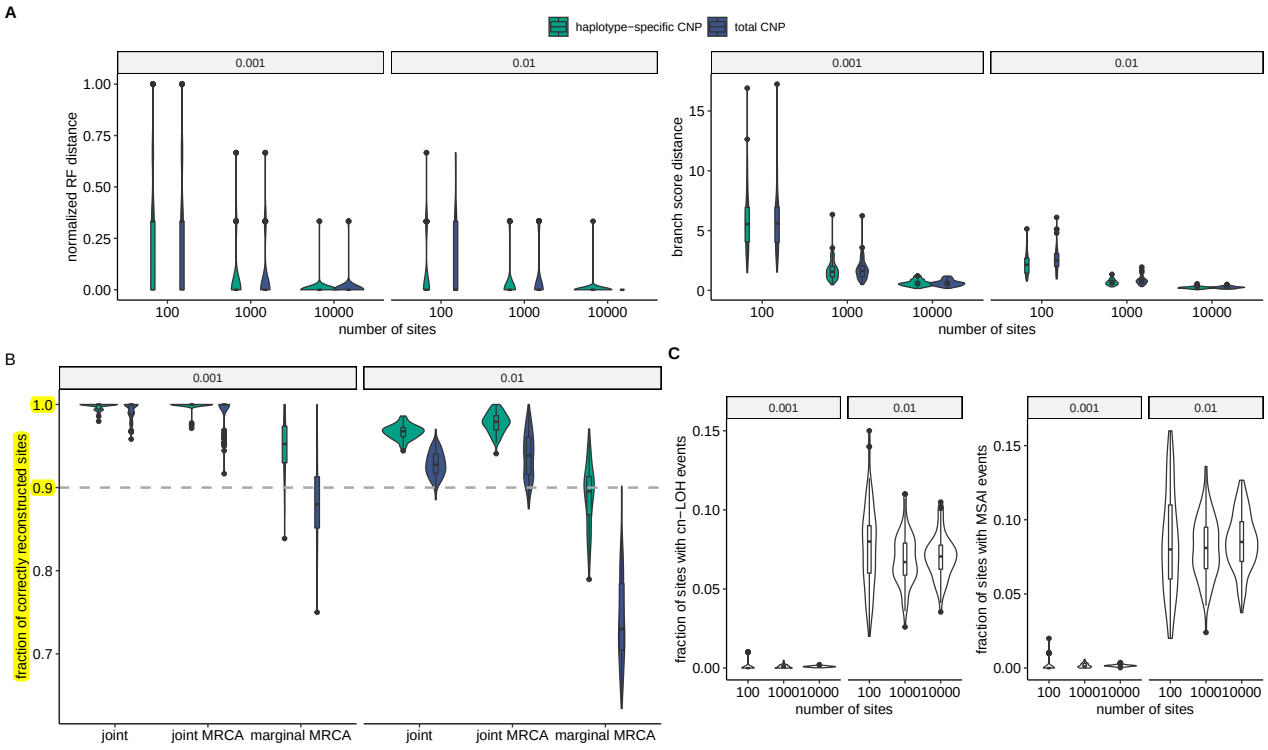

Figure 3: The performance of CNETML on reconstructing trees and ancestral states with total or haplotype-specific CNPs when all the samples are taken at the same time. **A**: The accuracy of phylogeny inference on data simulated with different number of sites and mutation rates. **B**: The accuracy of ancestral state reconstruction on simulated data with 1000 sites under different mutation rates. **C**: The fraction of cn-LOH and MSAI events in the simulated data. There are five samples at the same time in each simulated tree and 100 datasets for each parameter setting. The plots are grouped by mutation rates. The box plots show the median (centre), 1st (lower hinge), and 3rd (upper hinge) quartiles of the data; the whiskers extend to 1.5× of the interquartile range (distance between the 1st and 3rd quartiles); data beyond the interquartile range are plotted individually.

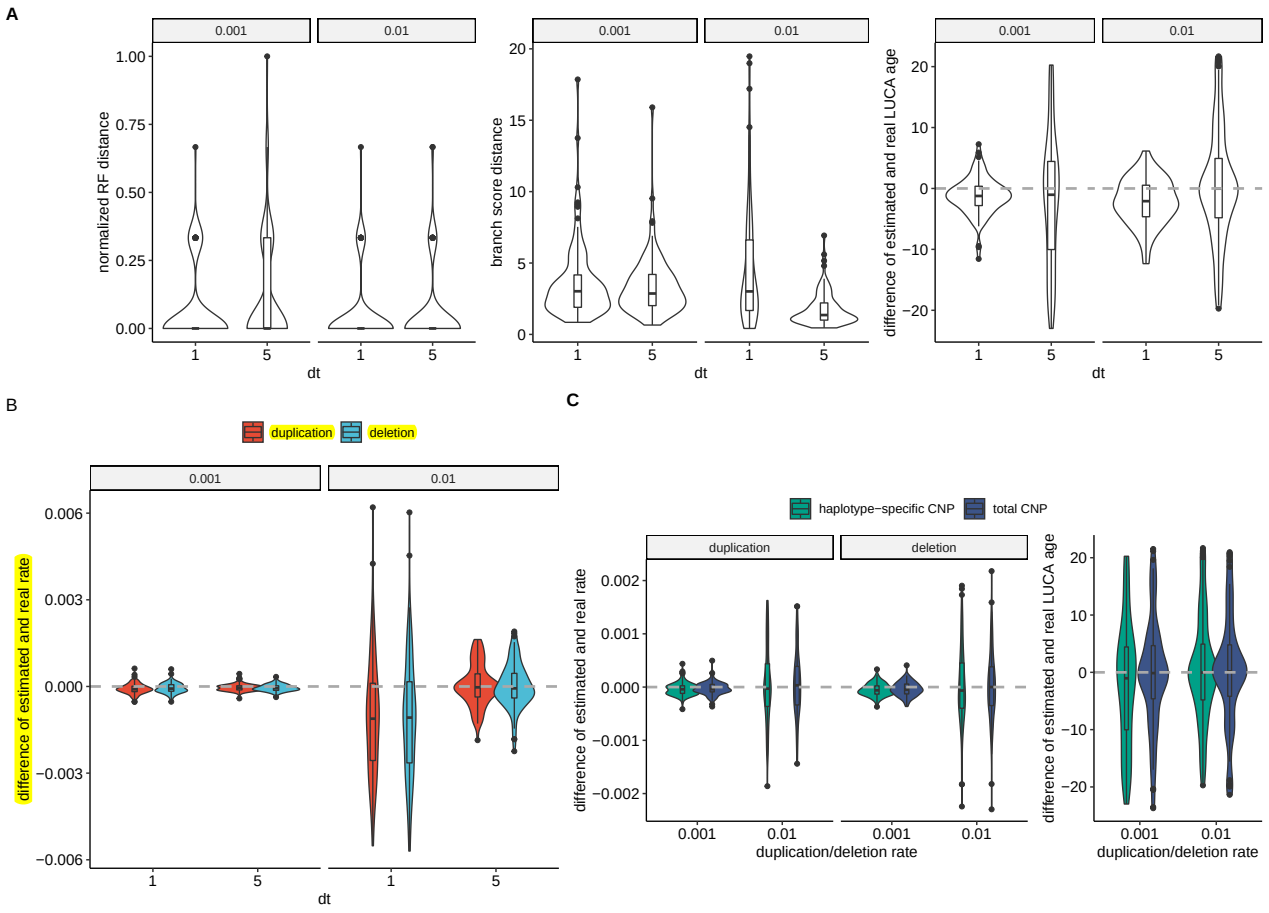

Figure 4: The performance of CNETML on jointly estimating tree topology, node ages, and mutation rates on data simulated with different values of dt and mutation rates. **A**: The accuracy of phylogeny inference. **B**: The accuracy of mutation rate estimation. **C**: The accuracy of the estimations of mutation rate and LUCA age with total or haplotype-specific CNPs on simulated data with $dt = 5$. There are five samples in each simulated tree and 100 datasets for each parameter setting. Grey dashed line: real values. Box plots as those in Figure 3.

CNPs of data simulated with $dt = 5$ years to see the extent of underestimation when using total copy numbers (Figure 4C), but similar to our previous results in Figure 3, we did not observe large differences.

Performance on relative copy numbers

CNAs called from sWGS data with common tools, such as QDNAseq [21], are often relative values, which are hard to interpret, but they provide a way to mitigate the effect of WGD in phylogeny inference. For example, PISCA used a baseline strategy to convert absolute haplotype-specific copy numbers to relative values, which lead to better phylogeny inference and more accurate rate estimation on simulated data with WGD [3]. The basic idea is to divide the observed copy numbers by an estimated baseline (rounded mean copy number) for each haplotype and then round the values up or down randomly to reduce bias when the remainder is not zero. This is a simple strategy to process the absolute CNPs for reasonable phylogeny inference when WGD is present, as it is just one event changing ploidy, and the normalization by baseline copy number may cancel its effect. We adopted a similar strategy in CNETS to simulate relative haplotype-specific copy numbers by using the known ploidy as the baseline and relative total copy numbers by using $2^{N_{WGD}}$ as the baseline. CNETS

output simulated relative total CNPs by reducing the normalized copy numbers with the ploidy of the genome, with values smaller than -2 and larger than 2 set it to be -2 and 2 respectively for consistency with QDNAseq output.

To see how CNETML performs on relative copy numbers, we ran CNETML on relative total and haplotype-specific CNPs simulated with $c_{max} = 8$, $dt = 1$ year, $u = e = 0.001$ per haplotype per site per year and WGD rate 0.05 per year. When running CNETML on relative total CNPs, we added the copy numbers with normal ploidy so that all values are positive. As a comparison, we also ran MEDICC2 [9], the only method to infer CNA-based phylogenies from NGS data at the presence of WGDs, on allele-specific CNPs which were converted from haplotype-specific CNPs by custom R scripts and CNETML on total CNPs respectively.

The results are grouped into four types by the distribution of WGD among samples: clonal WGD where WGD appears in all samples, multiple WGDs where there are more than one WGD across the tree but each sample has at most one WGD, single WGD where there is only one WGD across the tree, and no WGD. As expected, CNETML on absolute total CNPs reconstructed inaccurate phylogenies and missestimated mutation rates in most cases whenever WGD was present, with duplication rates largely overestimated especially on data with clonal WGD and deletion rates slightly underestimated. On data with single or multiple subclonal WGD(s), CNETML on relative CNPs can achieve similar performance in phylogeny inference to MEDICC2 on absolute allele-specific CNPs and the mutation rates were also accurately estimated with slight underestimation of deletion rates on relative haplotype-specific CNPs (Figure 5). On data without WGD, the performances of CNETML were similar on all types of data, which suggests using relative copy numbers still conserves the information for phylogeny inference and rate estimation. On data with clonal WGD, CNETML on relative CNPs reconstructed phylogenies less similar to the truth and underestimated mutation rates, particularly deletion rates on relative haplotype-specific CNPs, which is probably due to greater signal loss when converting copy numbers relative to a doubled ploidy. In summary, it seems entirely feasible to get good phylogeny inference directly from relative copy numbers, such as those from QDNAseq, when WGD is not clonal. For further validation of the inference, empirical information or methods to detect WGD [52] or call absolute copy numbers [26] from sWGS data may be used to estimate the presence of clonal WGD.

Performance under violation of independent sites assumption

Classically, the ML approach was shown to be highly robust to assumption violations [36]. As our model of CNA evolution strongly depends on the independent sites assumption, we ran CNETS using the waiting time approach to generate duplications and deletions of different sizes to examine how overlapping CNAs affect the performance of CNETML. We simulated trees with $dt = 1$ year so that rate estimation is feasible and introduced duplications/deletions along the tree with rate $u = e = 0.001$ per haplotype per site per year and mean size being 1, 10, and 100 bins (500 Kbp, 5 Mbp, and 50 Mbp), respectively. These sizes were chosen because focal CNAs are typically defined as CNAs of size no larger than 3 Mbp [53] and 50 Mbp is larger than p-arm size of 15 autosomes and q-arm size of 4 autosomes to include arm-level CNAs. We built trees with CNETML using original bin-level data (site as bin) and post-processed segment-level data (site as segment, see section 4 for details of the post-processing).

As expected, the inferences were more dissimilar to the ground truth with larger CNA sizes due to information loss as a result of overlaps, with more overestimated branch lengths and mutation rates and slightly more underestimated LUCA age (Figure 6A). However, when the mean duplication/deletion size was 5 Mbp, shorter than the typical focal CNA size, the extent of misestimation was not very large, and the tree topologies were still recovered well. In addition, when we scaled the estimated rates by the mean duplication/deletion size, the estimation errors were much smaller (Figure 6B). These results suggest that slight violation of independent sites assumption seems acceptable in phylogeny inference, and the estimated mutation rates may be scaled to account for the size of CNAs. On the other hand, the differences of using bin-level and segment-level data were slight because the site patterns in the

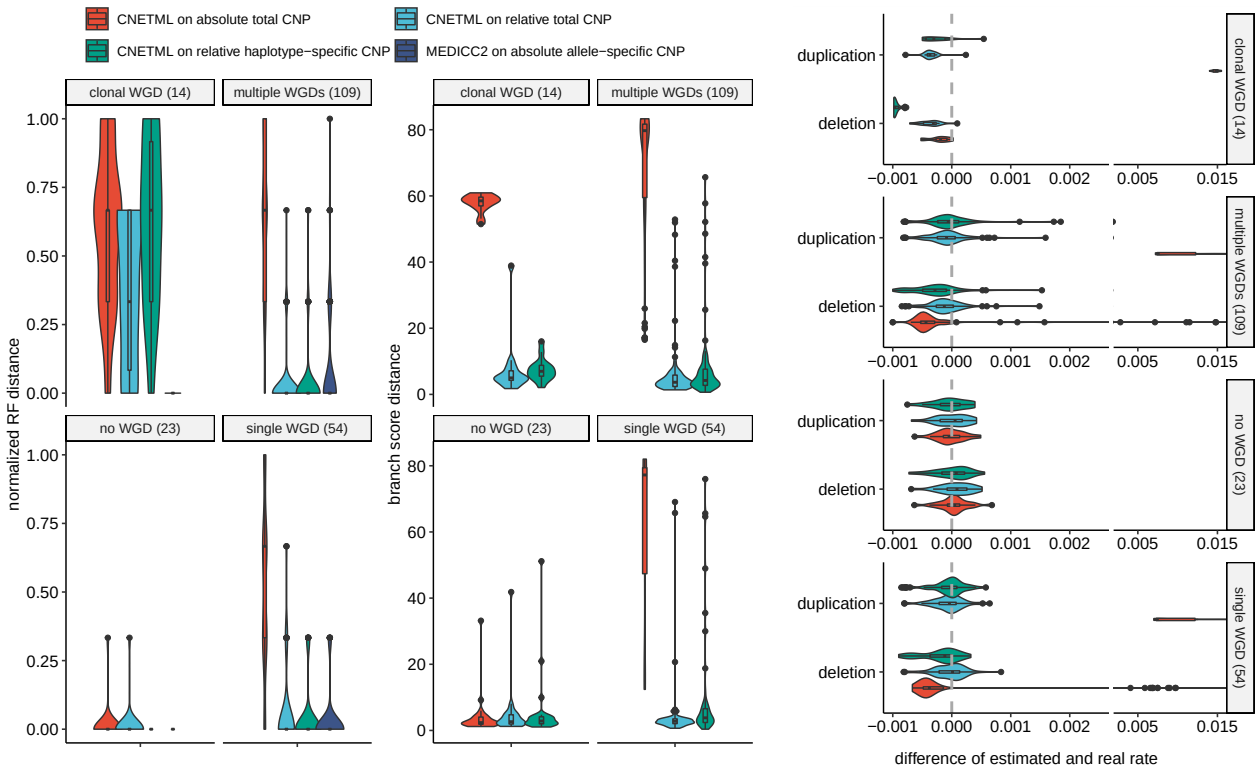

Figure 5: The performance of CNETML on relative copy number data. There are 200 simulated datasets in total, which are divided into four groups by the types of WGDs. The number of datasets in each group is shown in brackets. MEDICC2 was excluded when comparing branch score distance because the branch length in a tree built by it represents the number of events between genomes of two nodes. Outlier values larger than 0.02 on datasets with subclonal WGDs are excluded in the plot of mutation rates. Box plots as those in Figure 3.

input data were consistent in both cases, except that bin-level data might contain more sites and a larger number of the same pattern which lead to larger likelihoods and longer computing time (Figure 6C). The rates and LUCA age were slightly overestimated with bin-level data on larger CNAs, which is probably due to larger overdispersion caused by dependencies between adjacent bins with the same copy numbers [54]. Therefore, we recommend using segment-level data for faster computation, less correlation between segments, and better interpretability in practice.

Application to Barrett's esophagus patients

To demonstrate the applicability of CNETML on real data, we applied it to data for two BE patients in Figure 1 of [13], where CIN was used to predict risk progression (Figure 7). QDNAseq was applied on sWGS data to get relative CNPs in 589 bins of fixed size (about 5 Mbp) for each patient, which were normalized across the cohort of 777 endoscopy samples from 88 patients. One nonprogressor patient, 51, has 15 samples taken from 2006 to 2011, which shows similar CNPs across samples. The other progressor patient, 20, has 12 samples taken from 1998 to 2008, which shows more copy number variation across samples. Although WGD was shown to be prevalent in BE patients [3], it seems less likely to have clonal WGDs for these two patients given the large span of sampling times and diverse sampling locations. We rounded the provided fractional copy numbers to the nearest integers and merged consecutive bins with the same copy numbers across all samples into segments. Since the exact patient ages were not provided, the patient age at the first sampling time was set to be 60 (the mean age of all nonprogressors in the cohort) for patient 51 and 62 (the mean age of all progressors in the cohort) for patient 20, respectively, which provides good approximations of the upper bounds of the tree heights during optimization.

We first ran CNETML 100 times on the input data, selected the tree with largest likelihood, T_b, and did 100 bootstraps to get branch support values for T_b. Then we fixed the tree topology to be the same as that of T_b and did optimization of node ages and mutation rates to get T_b', with the initial mutation rates set to be the estimated rates on T_b. We ran another 100 bootstraps with the topology of T_b' to get the confidence intervals (2.5th and 97.5th percentile) of node ages and mutation rates in T_b'. Lastly, we reconstructed ancestral CNPs based on T_b' and checked the biological significance by computing their overlap with cancer-related genes from COSMIC Cancer Mutation Census (CMC) with keyword "oesophag" in the description of disease [55] and 75 regions selected by the elastic-net regression model as being predictive of BE progression (predictive regions) in [13].

The tree topology for patient 20 had bootstrap support values of more than 80% except for two branches. Although the branch connecting the samples taken at 12 months location 2 and 108 months location 1 (times before final endoscopy) had the lowest support value, they shared a loss on gene SMAD4, which was shown to promote tumorigenesis from BE toward esophageal cancer [56]. In contrast, the tree topology for patient 51 had much poorer support due to the lack of changes in copy numbers. The estimated mutation rate of patient 20 was slightly higher than that of patient 51, around 0.006 and 0.004 per haplotype per site per year respectively, which is as excepted because progressors tend to have higher mutation rates, and both estimations seem consistent with previous results for BE patients [3]. The LUCA age for patient 20 was about 40 years before the first sample, about 10 years earlier than patient 51. From the reconstructed CNPs of the MRCA of both patients (node 25 for patient 20 and node 31 for patient 51 in Figure 7), we found gene LRP1B included in a region on chr 2q with copy number gain (see Supplementary Table 2 for the complete list of overlaps). The original average relative copy number for patient 20 (1.4) across all samples is about twice as that for patient 51 (0.6), suggesting more gain in patient 20. Although most common alterations involving LRP1B are simple somatic mutations or copy number losses, 4.89% cases have copy number gains in the TCGA-ESCA cohort [57]. The MRCA CNP of patient 20 also had a region of gain on chr 4, which overlapped with the predictive region whose associated coefficient of variation for the relative risk (CV) is 1.018 (ranked 15th among 75 regions) [13]. The MRCA CNPs of patient 51 and the top and bottom lineages of patient 20 (node 20 and 22 in Figure 7), all overlapped with the

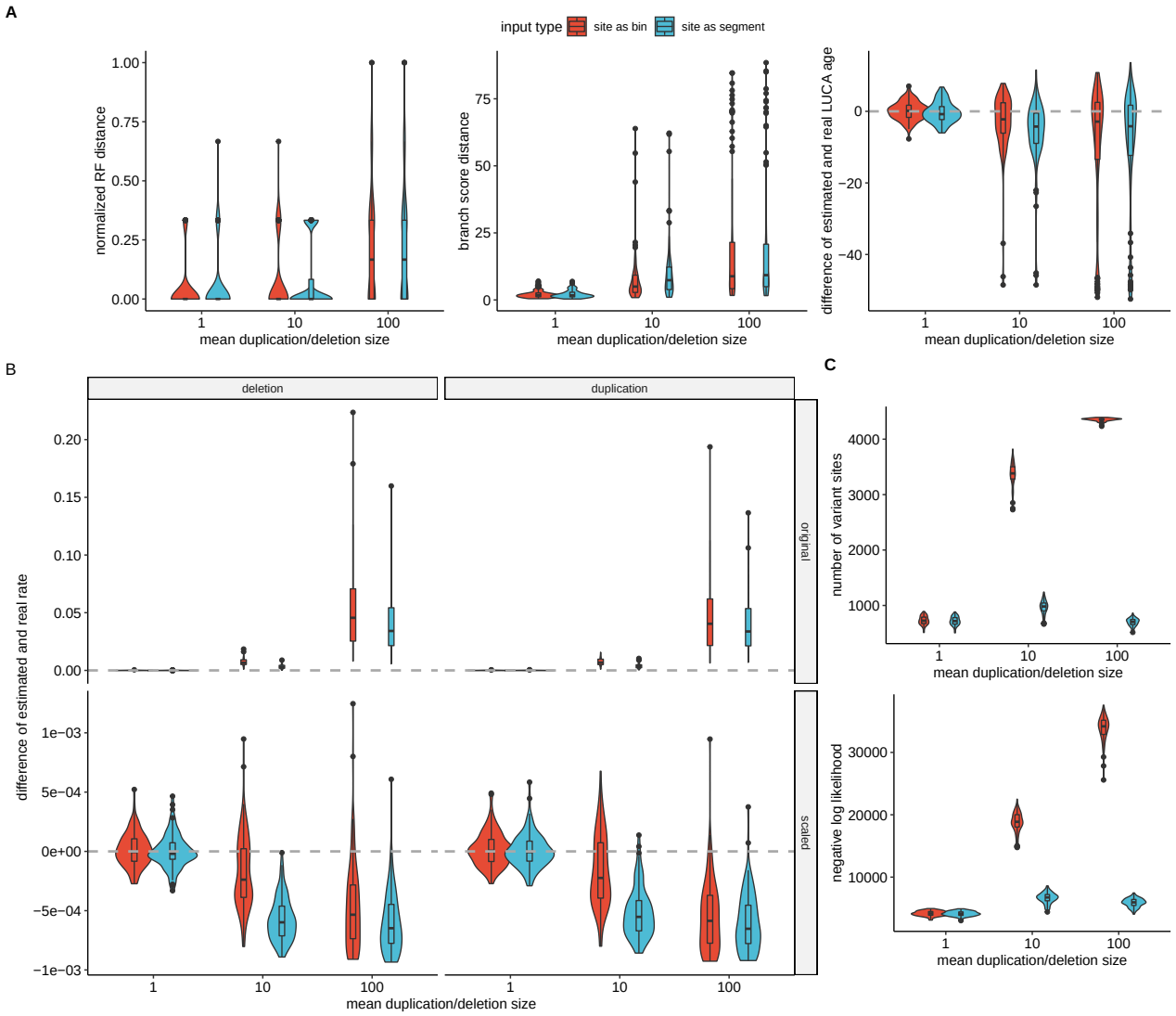

Figure 6: The performance of CNETML with different types of sites on data simulated with mean duplication/deletion of different sizes. **A**: The accuracy of phylogeny inference. **B**: The accuracy of mutation rate estimation before and after scaling. **C**: The number of variant sites used for likelihood computation and the negative log likelihoods for the ML trees. There are five samples in each simulated tree and 100 datasets for each parameter setting. Box plots as those in Figure 3.

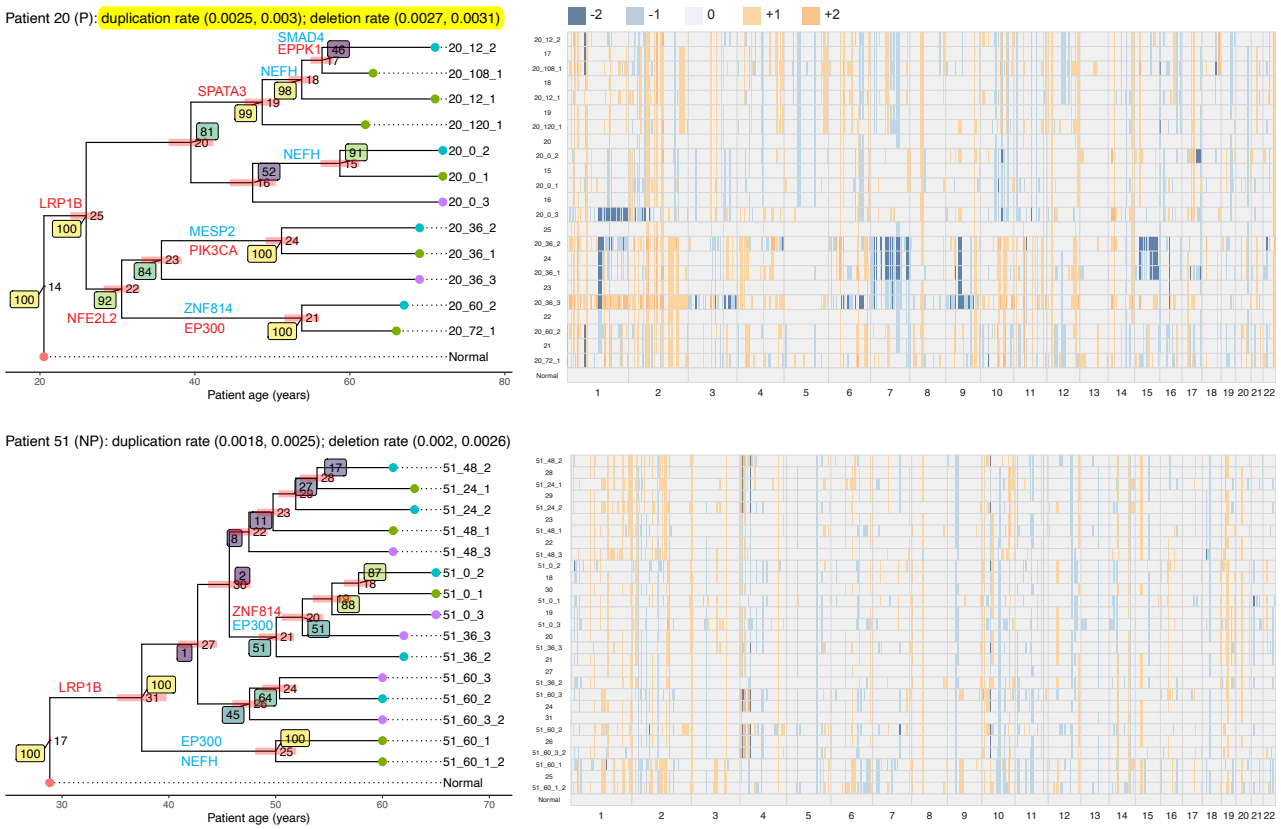

Figure 7: The ML trees and ancestral states reconstructed by CNETML for patient 20 and 51. The bootstrap support values are shown in coloured rectangles with lighter colours suggesting stronger support. The light red bars at the internal nodes show the confidence intervals of node ages. Each sample is denoted by "patientID_timeID_locationID", where timeID is the month before final endoscopy and locationID is the relative esophageal sample location. The cancer-related genes overlapping with the reconstructed CNPs are shown on the branches (red: copy number gain, blue: copy number loss). The confidence intervals of duplication/deletion rates are shown in round brackets in the title of the plot for each patient.

predictive region whose associated CV is 5.090 (ranked 6th among 75 regions). [13]. The MRCA CNP of the bottom lineages of patient 20 also overlapped with gene NFE2L2 on chr 2q, which has about 11.41% cases with copy number gains in TCGA-ESCA cohort [57]. For patient 51, the lineage starting from node 21 had a unique gain overlapping with gene ZNF814, which has 13.04% cases in TCGA-ESCA cohort [57]. All these findings suggest that the phylogenies and mutation rates inferred by CNETML are biologically meaningful and additional insights into tumour evolution can be gained from the reconstructed ancestral CNPs.

3 Discussion

In summary, we developed CNETML, a new ML method to reconstruct the evolutionary history of multiple samples taken from a single patient at different locations and/or times, which can take as input (relative) total integer copy numbers called from sWGS data. CNETML is capable of jointly estimating node ages and mutation rates by year when patient samples were sampled at different times. This capability is derived from a novel Markov model of CNA evolution, which assumes the sites (bins or segments) in a CNP are independent and hence allows the usage of classical methods for phylogeny

inference and ancestral reconstruction. We evaluated CNETML on data simulated with CNETS, our novel program of general utility to simulate CNAs along a phylogenetic tree. The simulations suggest that CNETML performed well when there were sufficient CNAs and/or timing information in the data, even on relative CNPs with subclonal WGDs. The ability to work on relative CNPs makes CNETML applicable to a wide range of sWGS data obtained from cancer patients, which was demonstrated by its application on two BE patients, where we inferred sample phylogenies along with ancestral CNPs which suggest the time MRCA arose and early CNAs driving the malignancy. Although caution is still required when interpreting the inference on relative CNPs without knowing the exact presence of clonal WGDs, the inference on relative copy numbers provide a reference for further improvement. CNETML is also applicable to allele-specific CNPs if they are phased, and the performances were similar to those on total CNPs when there were less than 10% of copy-neutral CNA events (such as cn-LOH and MSAI) across all sites. Despite the independent sites assumption, CNETML was robust to considerable amounts of overlaps among simulated focal CNAs.

Although CNETML aims to build a sample tree where each tip is a CNP from a patient sample, it can be used to build trees of tumour clones or single cells, since the main input is simply an integer copy number matrix for any taxa of interest. The input CNPs are assumed to be called from sWGS data and hence cover the whole genome, but it may be applicable to SNP array or WES data if the gaps between segments with atypical copy numbers are filled to avoid acquisition bias [58].

In principle, the likelihood-based approach adopted in CNETML is more sophisticated than distance matrix and MP methods. To allow for more flexible evolutionary models specific for tumour evolution, we implemented a novel tool rather than using existing frameworks designed for traditional phylogenetic inferences, such as BEAST [39, 40]. Future development of the model could include Markov chains at different scales to incorporate chromosomal and/or arm level gain/loss and WGD, and the use of regularization in the optimization to get better inference when there are insufficient information in the data [59]. Another development would be the estimation of varying mutation rates in different lineages under a relaxed local clock [60]. Finally we can extend to a fully Bayesian approach, which can impose informative prior distribution and naturally provide a measure of uncertainty of the inference (posterior probabilities of sampled trees).

The inference of tumour phylogeny from CNAs called from sWGS data is a very challenging problem. Although CNETML makes progresses in tackling some issues, it still has a few limitations in data handling. First, we assumed the input CNPs are complete and accurate, which is often violated in reality. Errors in copy number calls, which may arise from poor calling or missing data, directly affect the inference, since just one wrong copy number called at a site of a sample may lead to a unique site pattern and bias likelihood computation. Methods developed for sc-seq data often incorporate approaches to deal with data noise [19], which may be extended to CNAs called from sWGS data, such as combining CNA calling from raw read counts with phylogeny inference [5, 7] and incorporating false positives and false negatives into the model directly [6]. Moreover, we assume each sample is homogeneous with only one clone and do not deal with clonal deconvolution. This is reasonable to some extent as CNAs detected from sWGS data typically represent the dominant clone in a sample, which is different from sample trees built from SNVs that often represent highly admixed cell lineages [61]. However, given data of higher resolution, it would be helpful to quantify ITH and how it affects the monoclonal assumption.

In summary, we have provided a tool that can enhance the use of sWGS and allow for spatio-temporal inferences of tumour evolution in patients. Due to its relatively low cost we believe this approach will have increasing impact in understanding the biology of tumour evolution and will underlie future clinical applications.

4 Methods

Preprocessing of input data

The input CNPs for CNETML are mainly obtained from common CNA calling methods for sWGS data. For example, QDNAseq [21] is often used to get relative copy numbers by computing read counts in fixed-sized bins, doing segmentation, and calling copy numbers with CGHcall [62] which classifies copy numbers into: double deletion (-2), single deletion (-1), normal (0), gain (1), and amplification (2). To get the data matrix D, we assume the same binning or segmentation across all samples to get consistent sites. When raw copy number calls are at bin level, segments can be obtained by merging consecutive variant bins on the same chromosome with the same copy number across all samples.

The input sampling dates are converted to years (divided by 365). For convenience, the time for the first sample is set to 0 and the time for other samples is then counted as the number of years starting from the first sample.

The computation of likelihood

According to Felsenstein's pruning algorithm [42], a vector of conditional probabilities was computed for each node i at each possible haplotype-specific copy number state d_i on each site. We define $L_i(d_i)$, ~~partial likelihood,~~ as the conditional probability of observing data at tip(s) ~~below~~ a node i ~~which has copy number state d_i and an incoming branch of length t_i~~. When node i is an internal node with children node j and k,

$$L_i(d_i) = (\sum_{d_j} p_{d_i d_j}(t_j) L_j(d_j))(\sum_{d_k} p_{d_i d_k}(t_k) L_k(d_k)), \tag{4}$$

where $p_{d_i d_j}(t_j)$ represents the transition probability of d_i becoming d_j after time t_j. When the input are total copy numbers, there may be multiple states corresponding to one observed copy number and all the relevant states are assigned to one when initializing the likelihood values for tip nodes. Suppose S_t is a set of haplotype-specific copy number states corresponding to the observed total or haplotype-specific copy number at a tip node t, then $|S_t| = 1$ when the input are haplotype-specific copy numbers and $|S_t| >= 1$ when the input are total copy numbers. The partial likelihood for tip t is then

$$L_t(d_t) = \begin{cases} 1 & d_t \in S_t, \\ 0 & otherwise. \end{cases} \tag{5}$$

The root can only be in the normal state, so there is no need to make a weighted average of all possible states as the original algorithm. Suppose the root (LUCA node) is r with $d_r = 4$ and an outgoing branch of length t_m leading to MRCA node m, then the conditional probability of observed CNPs at a site p is:

$$P(D^{(p)}|\theta) = L_r(d_r) = p_{d_r d_m}(t_m) L_m(d_m). \tag{6}$$

To improve efficiency in likelihood computation, transition probability matrix, $P(t_j) = \{p_{d_i d_j}(t_j)\} = e^{Qt_j}$, was computed once with scaling and squaring method [63] for each branch of length t_j and used for all sites. $L_r(d_r)$ for two sites with the same site patterns were computed once too, as they have the same probability of being observed. In addition, the invariant sites were excluded when parsing the input and later corrected by a straightforward reconstituted method, which multiplies the number of invariant sites by the ~~likelihood~~ of observing an invariant site [58].

Statistical phylogeny inference with maximum likelihood method

An important aspect in optimization of likelihood function $L(T)$ is the incorporation of bound constraints in L-BFGS-B algorithm. To avoid negative branch length, we define a minimal branch length

l_m (1e-3 by default). To encode the constraints imposed by patient ages at different sampling times, we define a new variable x_i for an internal node i with child j on a tree T of n samples:

$$x_i = \begin{cases} t_1 & i = 1, \\ \frac{t_j - t_j^m - l_m}{t_i - t_i^m - 2l_m} & 1 < i <= n-1, \end{cases} \tag{7}$$

where i is from 1 (the root) to $n-1$, t_i is the age of node i, and t_i^m is the maximum age of the tips below node i. Because the parent age should always be smaller than those of the children nodes, x_i has bounds as below:

$$\begin{aligned} d + nl_m <= x_1 <= A_0 + d, \\ 0.01 <= x_i <= 0.9, \quad 1 < i <= n-1, \end{aligned} \tag{8}$$

where A_0 is the patient age at the first sample and d is the time difference between the last and first sample.

For exhaustive tree search, we simulated all the possible tree topologies for the given number of samples, computed their maximum likelihoods, and ordered optimized trees by likelihoods to get the ML tree. ~~To get all the possible tree topologies for n samples, we generated a random coalescence tree, represented it with a string, ordered the string in a specific way to represent the topology only, and repeated this procedure until the number of simulated unique tree topologies equals to the number of rooted trees for n tips.~~ For stochastic tree search, we started with a number of initial trees (100 by default), selected those with unique topologies, and computed their approximate likelihoods. Then we selected the top n_1 (20 by default) trees ordered by decreasing likelihoods to do hill-climbing nearest neighbor interchanges (NNIs) [44] and kept the top n_2 (5 by default) trees with largest likelihoods for further optimization to get the ML tree. To avoid local optima, we built parsimony-based stepwise addition trees as initial trees, which were obtained by using function `random.addition` in R library `ape` [46] and transformed into the formats acceptable by CNETML.

Data simulation

The overall procedure of simulations in CNETS is as follows:

1. Generate a random coalescence tree of n samples. Available trees can also be given as input. Optionally simulate temporal samples from a patient of specified age with two steps below.
   (a) Assign random times (in year) to the tips by changing terminal branch lengths.
   (b) Rescale the tree so that its height is no larger than the patient age at the last sampling time.
2. Simulate CNPs on the tree with the Markov model of CNAs.
   (a) Generate the CNP for the root (normal diploid genome).
   (b) Simulate CNPs directly at the end of each branch according to the transition probability matrix or simulate mutational events along each branch by using exponential waiting times.
3. Output result files.

When simulating CNPs directly given the total number of sites (segments), we distributed the sites roughly according to the size of each chromosome with Dirichlet distribution. Each genome with m sites was represented by its CNP $(c_1, c_2, ..., c_m)$ whose initial values at all sites are 2 for total copy number data or 4 for haplotype-specific copy number data. For a site i with state c_i, we sampled its target state from the discrete distribution specified by row i of transition probability matrix $P(l)$ for a branch of length l.

When simulating events of multiple scales by waiting times, we pre-specified the number of sites (bins) on each autosome of the reference genome with an array [367, 385, 335, 316, 299, 277, 251, 243,

184, 210, 215, 213, 166, 150, 134, 118, 121, 127, 79, 106, 51, 54], which were extracted from QDNAseq output on real data with bins of 500 Kbp. Each genome with m sites was initially represented by the set of sites, denoted as $G = (l_1, l_2, ..., l_m)$. To simulate a diploid genome, we copied another set of sites after the first set to represent the other haplotype, namely, $G_d = [G, G]$. The final CNP of the genome $(c_1, c_2, ..., c_m)$ was computed by adding up the number of copies of the same site across all the haplotypes when considering total copy number or the specific haplotype when considering haplotype-specific copy number. Some constraints were imposed to get more realistic data: 1) Chromosomal gain and WGD were only possible when the resultant maximum copy number is smaller than the specified c_{max}; 2) The duplication/deletion stopped at the end of a chromosome. For the simulation of specific mutational events along a branch of length l from initial time $t = 0$, we used the following steps:

1. Generate a random waiting time e from the exponential distribution with rate r, where r is the total mutation rate across the genome, obtained by adding up the duplication and deletion rates across all sites along the genome, chromosomal gain and loss rates across all chromosomes, and WGD rate.

2. Generate a mutation, whose type is randomly chosen based on the relative rates of different event types ~~which are summarized at different scales~~.

   (a) For segment duplication/deletion, randomly choose the start bin based on the rates across sites, the haplotype, and the size in terms of bins, where a duplication can be either tandem (duplicated at the end of the current location) or interspersed (inserted in any position across the genome) with equal possibilities.

   (b) For chromosome gain/loss, randomly select the chromosome according to the rates across chromosomes and the haplotype.

3. $t = t + e$.

4. Stop when $t >= l$.

5 Acknowledgements

CPB and BL acknowledge funding from the Wellcome Trust (209409/Z/17/Z). The authors acknowledge the use of the UCL Myriad High Throughput Computing Facility (Myriad@UCL), and associated support services, in the completion of this work. We thank Simone De Angelis, Rachel Muir, and Christos Magkos for testing our programs. We thank William Cross for helpful suggestions on the manuscript.

References

[1] Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017.

[2] Samuel F Bakhoum and Lewis C Cantley. The multifaceted role of chromosomal instability in cancer and its microenvironment. *Cell*, 174(6):1347–1360, 2018.

[3] Pierre Martinez, Diego Mallo, Thomas G Paulson, Xiaohong Li, Carissa A Sanchez, Brian J Reid, Trevor A Graham, Mary K Kuhner, and Carlo C Maley. Evolution of barrett's esophagus through space and time at single-crypt and whole-biopsy levels. *Nature communications*, 9(1):794, 2018.

[4] Simone Zaccaria, Mohammed El-Kebir, Gunnar W Klau, and Benjamin J Raphael. Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology*, 25(7):689–708, 2018.

[5] Jack Kuipers, Mustafa Anıl Tuncel, Pedro Ferreira, Katharina Jahn, and Niko Beerenwinkel. Single-cell copy number calling and event history reconstruction. *bioRxiv*, 2020.

[6] Sohrab Salehi, Fatemeh Dorri, Kevin Chern, Farhia Kabeer, Nicole Rusk, Tyler Funnell, Marc J Williams, Daniel Lai, Mirela Andronescu, Kieran R. Campbell, Andrew McPherson, Samuel Aparicio, Andrew Roth, Sohrab Shah, and Alexandre Bouchard-Côté. Cancer phylogenetic tree inference at scale from 1000s of single cell genomes. *bioRxiv*, 2021.

[7] Magda Markowska, Tomasz Cakala, Blazej Miasojedow, Dilafruz Juraeva, Johanna Mazur, Edith Ross, Eike Staub, and Ewa Szczurek. Conet: Copy number event tree model of evolutionary tumor history for single-cell data. *bioRxiv*, 2021.

[8] Natalie Andersson, Subhayan Chattopadhyay, Anders Valind, Jenny Karlsson, and David Gisselsson. Devolution—a method for phylogenetic reconstruction of aneuploid cancers based on multiregional genotyping data. *Communications biology*, 4(1103), 2021.

[9] Tom L Kaufmann, Marina Petkovic, Thomas BK Watkins, Emma C Colliver, Sofya Laskina, Nisha Thapa, Darlan C Minussi, Nicholas Navin, Charles Swanton, Peter Van Loo, Kerstin Haase, Maxime Tarabichi, and Roland F Schwarz. Medicc2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *bioRxiv*, 2021.

[10] Yushu Liu, Mohammadamin Edrisi, Huw A. Ogilvie, and Luay Nakhleh. Nestedbd: Bayesian inference of phylogenetic trees from single-cell dna copy number profile data under a birth-death model. *bioRxiv*, 2022.

[11] A.M. Piskorz, D. Ennis, G. Macintyre, T.E. Goranova, M. Eldridge, N. Segui-Gracia, M. Valganon, A. Hoyle, C. Orange, L. Moore, M. Jimenez-Linan, D. Millan, I.A. McNeish, and J.D. Brenton. Methanol-based fixation is superior to buffered formalin for next-generation sequencing of dna from clinical cancer samples. *Annals of Oncology*, 27(3):532–539, 2016.

[12] Giovanni Ciriello, Martin L. Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, 2013.

[13] Sarah Killcoyne, Eleanor Gregson, David C Wedge, Dan J Woodcock, Matthew D Eldridge, Rachel De La Rue, Ahmad Miremadi, Sujath Abbas, Adrienn Blasko, Cassandra Kosmidou, et al. Genomic copy number predicts esophageal cancer years before transformation. *Nature medicine*, 26:1726–1732, 2020.

[14] Chia-Chin Wu, Hannah C. Beird, J. Andrew Livingston, Shailesh Advani, Akash Mitra, Shaolong Cao, Alexandre Reuben, Davis Ingram, Wei-Lien Wang, Zhenlin Ju, Cheuk Hong Leung, Heather Lin, Youyun Zheng, Jason Roszik, Wenyi Wang, Shreyaskumar Patel, Robert S. Benjamin, Neeta Somaiah, Anthony P. Conley, Gordon B. Mills, Patrick Hwu, Richard Gorlick, Alexander Lazar, Najat C. Daw, Valerae Lewis, and P. Andrew Futreal. Immuno-genomic landscape of osteosarcoma. *Nature Communications*, 11(1):1008, 2020.

[15] Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1867(2):127–138, 2017.

[16] Salim Akhter Chowdhury, Stanley E Shackney, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A Schäffer, and Russell Schwartz. Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS computational biology*, 10(7):e1003740, 2014.

[17] Salim Akhter Chowdhury, E. Michael Gertz, Darawalee Wangsa, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A. Schäffer, and Russell Schwartz. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics*, 31(12):i258–i267, 2015.

[18] Roland F Schwarz, Anne Trinh, Botond Sipos, James D Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology*, 10(4):e1003535, 2014.

[19] Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biology*, 21:208, 2020.

[20] Geoff Macintyre, Bauke Ylstra, and James D. Brenton. Sequencing structural variants in cancer for precision therapeutics. *Trends in Genetics*, 32(9):530–542, 2016.

[21] Ilari Scheinin, Daoud Sie, Henrik Bengtsson, Mark A Van De Wiel, Adam B Olshen, Hinke F Van Thuijl, Hendrik F Van Essen, Paul P Eijk, François Rustenburg, Gerrit A Meijer, et al. Dna copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome research*, 24:2022–2032, 2014.

[22] Geoff Macintyre, Teodora E. Goranova, Dilrini De Silva, Darren Ennis, Anna M. Piskorz, Matthew Eldridge, Daoud Sie, Liz-Anne Lewsley, Aishah Hanif, Cheryl Wilson, Suzanne Dowson, Rosalind M. Glasspool, Michelle Lockley, Elly Brockbank, Ana Montes, Axel Walther, Sudha Sundar, Richard Edmondson, Geoff D. Hall, Andrew Clamp, Charlie Gourley, Marcia Hall, Christina Fotopoulou, Hani Gabra, James Paul, Anna Supernat, David Millan, Aoisha Hoyle, Gareth Bryson, Craig Nourse, Laura Mincarelli, Luis Navarro Sanchez, Bauke Ylstra, Mercedes Jimenez-Linan, Luiza Moore, Oliver Hofmann, Florian Markowetz, Iain A. McNeish, and James D. Brenton. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, 50(9):1262–1270, 2018.

[23] Ann-Marie Baker, William Cross, Kit Curtius, Ibrahim Al Bakir, Chang-Ho Ryan Choi, Hayley Louise Davis, Daniel Temko, Sujata Biswas, Pierre Martinez, Marc J Williams, et al. Evolutionary history of human colitis-associated colorectal cancer. *Gut*, 68(6):985–995, 2019.

[24] Samuel D. Abbou, David S. Shulman, Steven G. DuBois, and Brian D. Crompton. Assessment of circulating tumor DNA in pediatric solid tumors: The promise of liquid biopsies. *Pediatric Blood Cancer*, 66(5):e27595, 2019.

[25] Gitta Boons, Timon Vandamme, Laura Mariën, Willem Lybaert, Geert Roeyen, Tim Rondou, Konstantinos Papadimitriou, Katrien Janssens, Bart Op de Beeck, Marc Simoens, et al. Longitudinal copy-number alteration analysis in plasma cell-free dna of neuroendocrine neoplasms is a novel specific biomarker for diagnosis, prognosis, and follow-up. *Clinical Cancer Research*, 28(2):338–349, 2022.

[26] Carolin M Sauer, Matthew D Eldridge, Maria Vias, James A Hall, Samantha Boyle, Geoff Macintyre, Thomas Bradley, Florian Markowetz, and James D Brenton. Absolute copy number fitting from shallow whole genome sequencing data. *bioRxiv*, 2021.

[27] Travis I. Zack, Steven E. Schumacher, Scott L. Carter, Andrew D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H. Mermel, Carrie Sougnez, Stacey B. Gabriel, Bryan Hernandez, Hui Shen, Peter W. Laird, Gad Getz, Matthew Meyerson, and Rameen Beroukhim. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140, 2013.

[28] Thomas B K Watkins, Emilia L Lim, Marina Petkovic, Sergi Elizalde, Nicolai J Birkbak, Gareth A. Wilson, David A. Moore, Eva Grönroos, Andrew Rowan, Sally M Dewhurst, Jonas Demeulemeester, Stefan C Dentro, Stuart Horswell, Lewis Au, Kerstin Haase, Mickael Escudero, Rachel Rosenthal, Maise Al Bakir, Hang Xu, Kevin Litchfield, Wei Ting Lu, Thanos P. Mourikis, Michelle Dietzen, Lavinia Spain, George D. Cresswell, Dhruva Biswas, Philippe Lamy, Iver Nordentoft, Katja Harbst, Francesc Castro-Giner, Lucy R. Yates, Franco Caramia, Fanny Jaulin, Cécile Vicier, Ian P. M. Tomlinson, Priscilla K. Brastianos, Raymond J. Cho, Boris C. Bastian, Lars Dyrskjøt, Göran B. Jönsson, Peter Savas, Sherene Loi, Peter J. Campbell, Fabrice Andre, Nicholas M. Luscombe, Neeltje Steeghs, Vivianne C. G. Tjan-Heijnen, Zoltan Szallasi, Samra Turajlic, Mariam Jamal-Hanjani, Peter Van Loo, Samuel F. Bakhoum, Roland F. Schwarz, Nicholas McGranahan, and Charles Swanton. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, 587:126–132, 2020.

[29] Ron Zeira and Ron Shamir. Genome rearrangement problems with single and multiple gene copies: a review. *Bioinformatics and Phylogenetics*, pages 205–241, 2019.

[30] Eric Letouzé, Yves Allory, Marc A. Bollet, François Radvanyi, and Frédéric Guyon. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology*, 11(7):R76, 2010.

[31] Ruli Gao, Alexander Davis, Thomas O. McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, Funda Meric-Bernstam, Franziska Michor, and Nicholas E. Navin. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*, 48(10):1119–1130, 2016.

[32] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[33] Darlan C Minussi, Michael D Nicholson, Hanghui Ye, Alexander Davis, Kaile Wang, Toby Baker, Maxime Tarabichi, Emi Sei, Haowei Du, Mashiat Rabbani, Cheng Peng, Min Hu, Shanshan Bai, Yu-wei Lin, Aislyn Schalck, Asha Multani, Jin Ma, Thomas O. McDonald, Anna Casasent, Angelica Barrera, Hui Chen, Bora Lim, Banu Arun, Funda Meric-Bernstam, Peter Van Loo, Franziska Michor, and Nicholas E. Navin. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, 592(7853):302–308, 2021.

[34] Ron Zeira and Benjamin J Raphael. Copy number evolution with weighted aberrations in cancer. *Bioinformatics*, 36:i344–i352, 2020.

[35] Ron Zeira, Geoffrey Mon, and Benjamin J. Raphael. Genome Halving and Aliquoting Under the Copy Number Distance. In Alessandra Carbone and Mohammed El-Kebir, editors, *21st International Workshop on Algorithms in Bioinformatics (WABI 2021)*, volume 201 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:25, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[36] Ziheng Yang. *Molecular evolution: a statistical approach*. Oxford University Press, 2014.

[37] Sandra Hui and Rasmus Nielsen. SCONCE: a method for profiling copy number alterations in cancer evolution using single-cell whole genome sequencing. *Bioinformatics*, 01 2022.

[38] Sergi Elizalde, Ashley M. Laughney, and Samuel F. Bakhoum. A Markov chain for numerical chromosomal instability in clonally expanding populations. *PLOS Computational Biology*, 14(9):e1006447, 2018.

[39] Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), 2018.

[40] Remco Bouckaert, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J. Drummond. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650, 2019.

[41] Johannes Smolander, Sofia Khan, Kalaimathy Singaravelu, Leni Kauko, Riikka J. Lund, Asta Laiho, and Laura L. Elo. Evaluation of tools for identifying large copy number variations from ultra-low-coverage whole-genome sequencing data. *BMC Genomics*, 22:357, 2021.

[42] Joseph Felsenstein. *Inferring phylogenies*. Sinauer associates Sunderland, MA, 2004.

[43] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd ed. 2006. edition, 2006.

[44] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2014.

[45] Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896, 2000.

[46] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

[47] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.

[48] Mary K Kuhner and Joseph Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3):459–468, 1994.

[49] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[50] Stefan C. Dentro, Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Yulia Rubanova, Geoff Macintyre, Jonas Demeulemeester, Ignacio Vázquez-García, Kortine Kleinheinz, Dimitri G. Livitz, Salem Malikic, Nilgun Donmez, Subhajit Sengupta, Pavana Anur, Clemency Jolly, Marek Cmero, Daniel Rosebrock, Steven E. Schumacher, Yu Fan, Matthew Fittall, Ruben M. Drews, Xiaotong Yao, Thomas B.K. Watkins, Juhee Lee, Matthias Schlesner, Hongtu Zhu, David J. Adams, Nicholas McGranahan, Charles Swanton, Gad Getz, Paul C. Boutros, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Inigo Martincorena, Florian Markowetz, Ville Mustonen, Ke Yuan, Moritz Gerstung, Paul T. Spellman, Wenyi Wang, Quaid D. Morris, David C. Wedge, Peter Van Loo, Stefan C. Dentro, Ignaty Leshchiner, Moritz Gerstung, Clemency Jolly, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Santiago Gonzalez, Yulia Rubanova, Geoff Macintyre, Jonas Demeulemeester, David J. Adams, Pavana Anur, Rameen Beroukhim, Paul C. Boutros, David D. Bowtell, Peter J. Campbell, Shaolong Cao, Elizabeth L. Christie, Marek Cmero, Yupeng Cun, Kevin J. Dawson, Nilgun Donmez, Ruben M. Drews, Roland Eils, Yu Fan, Matthew Fittall, Dale W. Garsed, Gad Getz, Gavin Ha, Marcin Imielinski, Lara Jerman,

Yuan Ji, Kortine Kleinheinz, Juhee Lee, Henry Lee-Six, Dimitri G. Livitz, Salem Malikic, Florian Markowetz, Inigo Martincorena, Thomas J. Mitchell, Ville Mustonen, Layla Oesper, Martin Peifer, Myron Peto, Benjamin J. Raphael, Daniel Rosebrock, S. Cenk Sahinalp, Adriana Salcedo, Matthias Schlesner, Steven E. Schumacher, Subhajit Sengupta, Ruian Shi, Seung Jun Shin, Lincoln D. Stein, Oliver Spiro, Ignacio Vázquez-García, Shankar Vembu, David A. Wheeler, Tsun-Po Yang, Xiaotong Yao, Ke Yuan, Hongtu Zhu, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, 184(8):2239–2254.e39, 2021.

[51] Adrien Rieux and François Balloux. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular ecology*, 25(9):1911–1924, 2016.

[52] Salpie Nowinski. WGD classifier, 2022. Available at `https://github.com/BCI-EvoCa/CNA_stability/blob/master/WGD_classifier.html`.

[53] Oscar Krijgsman, Beatriz Carvalho, Gerrit A. Meijer, Renske D.M. Steenbergen, and Bauke Ylstra. Focal chromosomal copy number aberrations in cancer—Needles in a genome haystack. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(11):2698–2704, 2014.

[54] Rasmus Nielsen. *Statistical methods in molecular evolution*. Springer, 2006.

[55] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2018.

[56] Jovana R. Gotovac, Tanjina Kader, Julia V. Milne, Kenji M. Fujihara, Luis E. Lara-Gonzalez, Kylie L. Gorringe, Sangeetha N. Kalimuthu, Madawa W. Jayawardana, Cuong P. Duong, Wayne A. Phillips, and Nicholas J. Clemons. Loss of smad4 is sufficient to promote tumorigenesis in a model of dysplastic barrett's esophagus. *Cellular and Molecular Gastroenterology and Hepatology*, 12(2):689–713, 2021.

[57] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[58] Adam D. Leaché, Barbara L. Banbury, Joseph Felsenstein, Adrián nieto-Montes de Oca, and Alexandros Stamatakis. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6):1032–1047, 2015.

[59] Vu Dinh, Lam Si Tung Ho, Marc A. Suchard, and Frederick A. Matsen IV. Consistency and convergence rate of phylogenetic inference via regularization. *The Annals of Statistics*, 46(4):1481–1512, 2018.

[60] Mario dos Reis, Philip C. J. Donoghue, and Ziheng Yang. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80, 2016.

[61] João M Alves, Tamara Prieto, and David Posada. Multiregional tumor trees are not phylogenies. *Trends in cancer*, 3(8):546–550, 2017.

[62] Mark A Van De Wiel, Kyung In Kim, Sjoerd J Vosse, Wessel N Van Wieringen, Saskia M Wilting, and Bauke Ylstra. Cghcall: calling aberrations for array cgh tumor profiles. *Bioinformatics*, 23(7):892–894, 2007.

[63] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45(1):3–49, 2003.

Supplementary document for CNETML: Maximum likelihood inference of tumour phylogeny from copy number profiles of spatio-temporal samples

List of Tables

List of Figures

Table 1: The rate matrix Q when the maximum total copy number $c_{max} = 4$.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	0/0	**0/1**	**1/0**	**0/2**	**1/1**	**2/0**	**0/3**	**1/2**	**2/1**	**3/0**	**0/4**	**1/3**	**2/2**	**3/1**	**4/0**
0/0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0/1	e	$-(u+e)$	0	u	0	0	0	0	0	0	0	0	0	0	0
1/0	e	0	$-(u+e)$	0	0	u	0	0	0	0	0	0	0	0	0
0/2	0	e	0	$-(u+e)$	0	0	u	0	0	0	0	0	0	0	0
1/1	0	e	e	0	$-2(u+e)$	0	0	u	u	0	0	0	0	0	0
2/0	0	0	e	0	0	$-(u+e)$	0	0	0	u	0	0	0	0	0
0/3	0	0	0	e	0	0	$-(u+e)$	0	0	0	u	0	0	0	0
1/2	0	0	0	e	e	0	0	$-2(u+e)$	0	0	0	u	u	0	0
2/1	0	0	0	0	e	e	0	0	$-2(u+e)$	0	0	0	u	u	0
3/0	0	0	0	0	0	e	0	0	0	$-(u+e)$	0	0	0	0	u
0/4	0	0	0	0	0	0	e	0	0	0	$-e$	0	0	0	0
1/3	0	0	0	0	0	0	e	e	0	0	0	$-2e$	0	0	0
2/2	0	0	0	0	0	0	0	e	e	0	0	0	$-2e$	0	0
3/1	0	0	0	0	0	0	0	0	e	e	0	0	0	$-2e$	0
4/0	0	0	0	0	0	0	0	0	0	e	0	0	0	0	$-e$

Table 2: Parameters used for tree generation in CNETS.

effective population size	N_e	90000
generation time in year (365 days)	t	0.002739726
exponential growth rate	β	1.563e-3

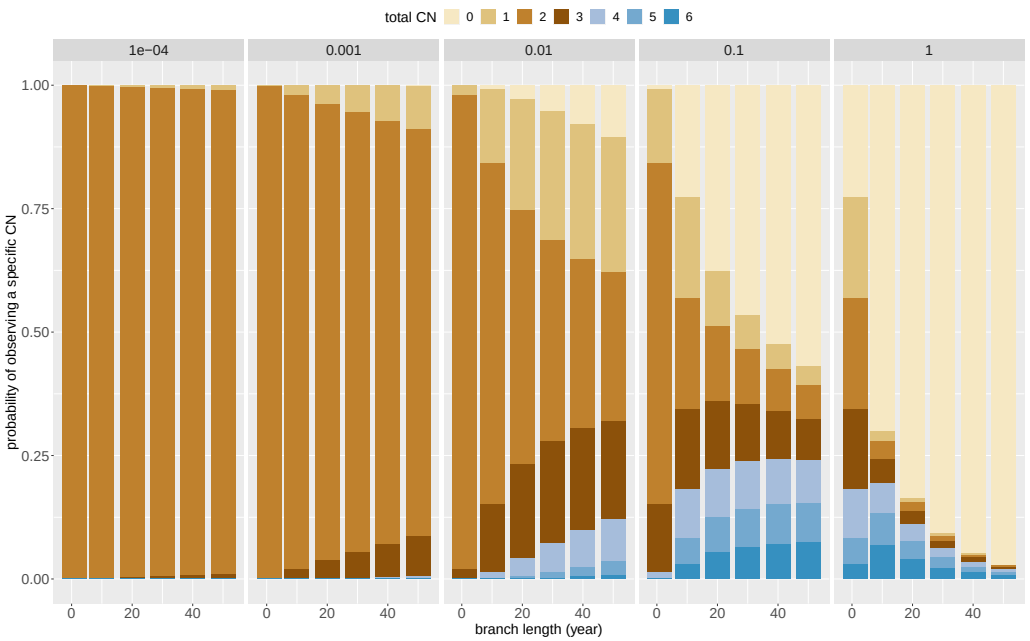

Figure 1: The distribution of possible total copy numbers at the end of one branch under the Markov model. The plots are grouped by mutation rates. In each group, the x-axis shows branch length at size 1, 10, 20, ..., 50. The y-axis shows the probability of changing from normal total copy number (2) to each possible total copy number. We computed the final states of the Markov chain for one branch of varying lengths starting at normal state, copy number (1,1)~~)~~. When the mutation rate is very low (0.0001 per haplotype per site per year), there are only a few mutations and most sites stay normal. When the mutation rate is too high (0.1 per haplotype per site per year), given longer branch lengths, more sites reach absorbing states (copy number 0).

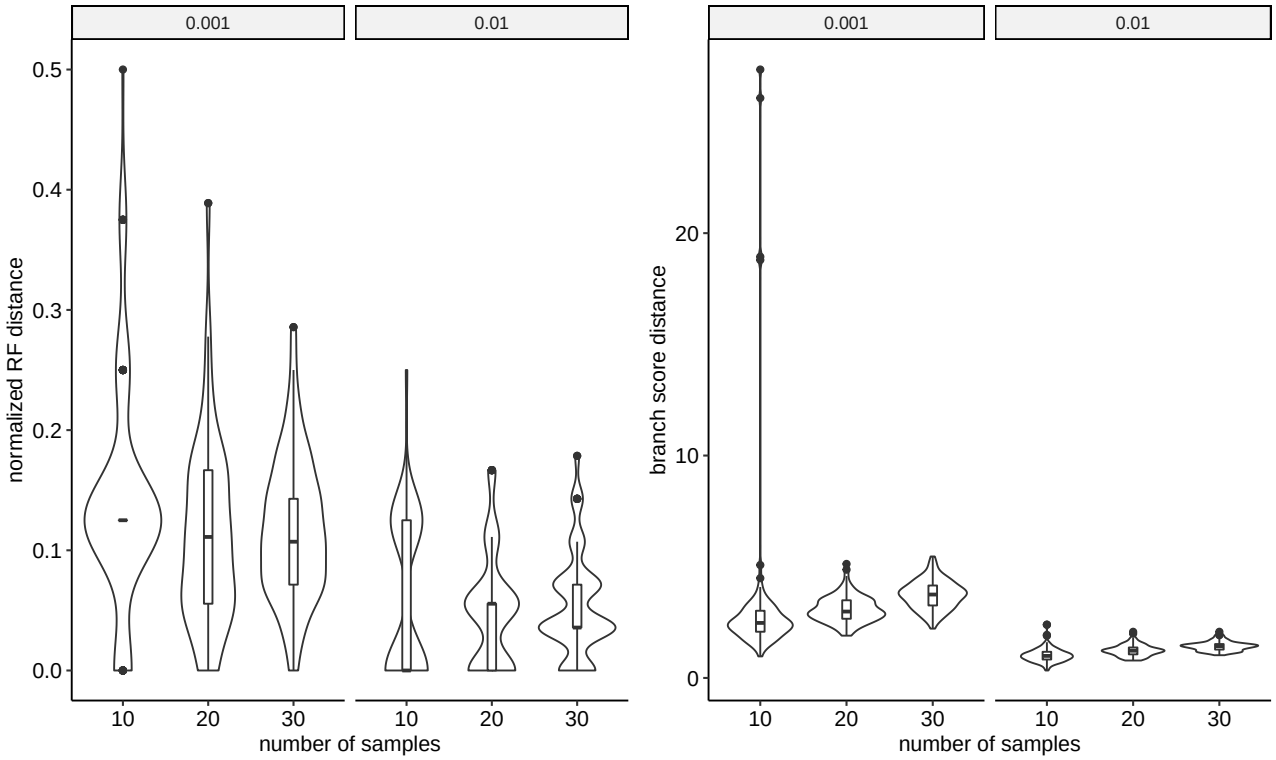

Figure 2: The performance of CNETML (stochastic tree search) on data simulated with different mutation rates and number of samples. All the simulated samples are at the same time. The plots are grouped by mutation rates. There are 100 datasets for each parameter setting.

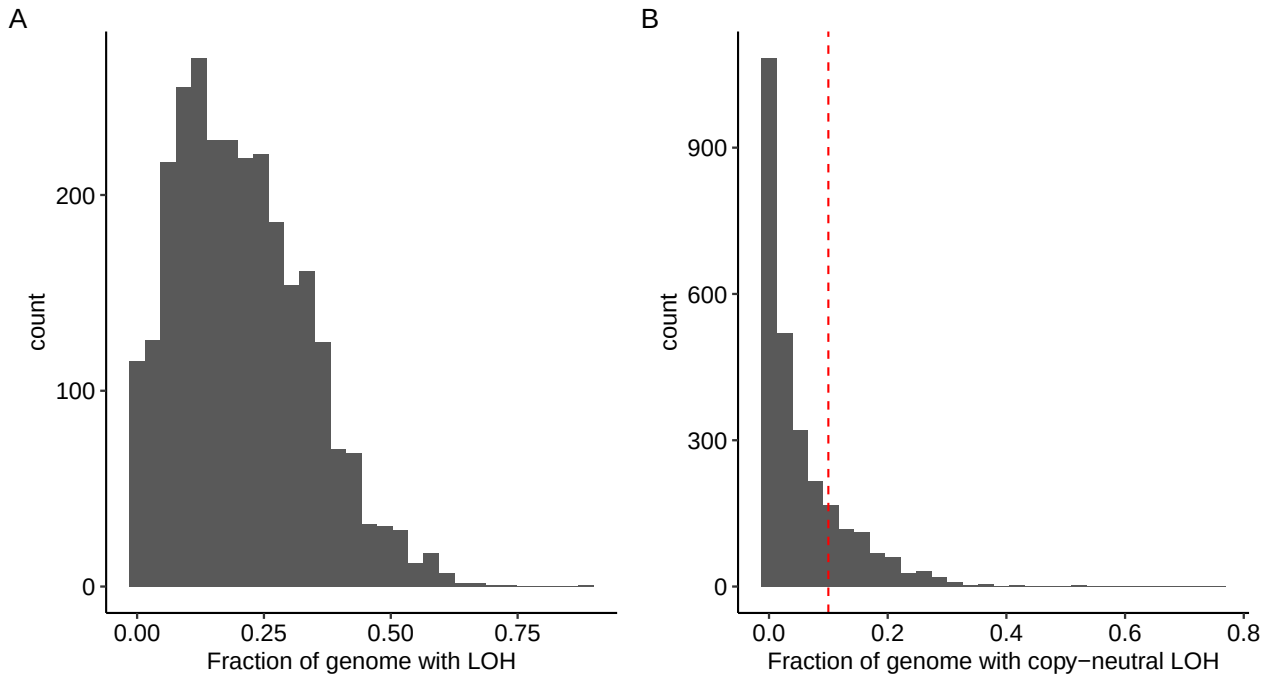

Figure 3: The distribution of the fractions of genome with loss of heterozygosity (LOH) on 2778 samples from PCAWG dataset. **A**: The distribution of the fractions of genome with LOH. **B**: The distribution of the fractions of genome with copy-neutral LOH.

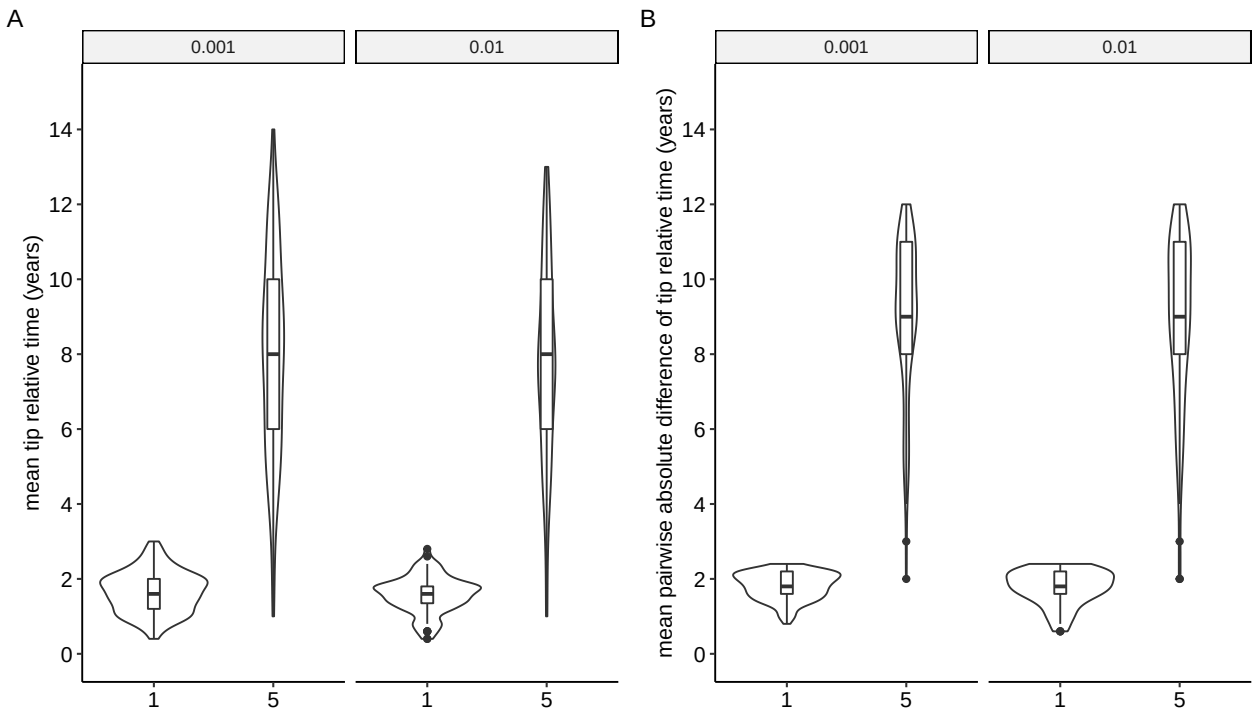

Figure 4: The range of simulated sampling times under different temporal signal strengths and mutation rates. **A**: The average of the relative times at the tips (assuming the first sample is at time 0) in the simulated trees. **B**: The average of pairwise absolute difference of the relative times at the tips in the simulated trees. The plots are grouped by mutation rates. There are 100 datasets for each parameter setting.

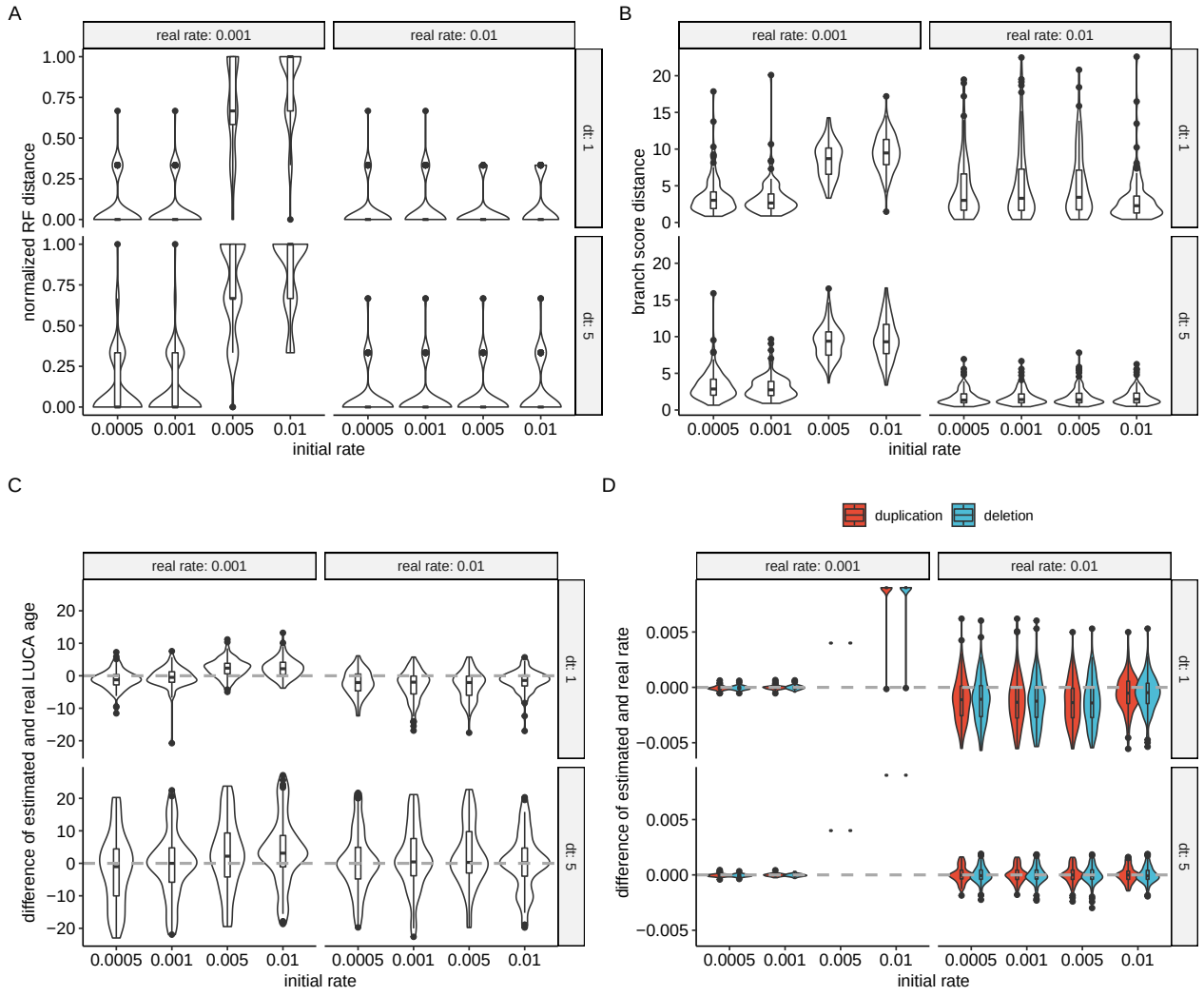

Figure 5: The sensitivity of CNETML to initial mutation rates when jointly estimating tree topology, node ages, and mutation rates on data simulated with different temporal signal strengths and mutation rates. **A-C**: The accuracy of tree inference under different initial mutation rates. **D**: The accuracy of mutation rate estimation under different initial mutation rates.