Genome Biology

METHOD

Open Access

CNETML: maximum likelihood inference of phylogeny from copy number profiles of multiple samples

Bingxin Lu[1,2*], Kit Curtius[3,4], Trevor A. Graham[3,5], Ziheng Yang[6] and Chris P. Barnes[1,2*]

*Correspondence: b.lu@ucl.ac.uk; christopher.barnes@ucl.ac.uk

[1] Department of Cell and Developmental Biology, University College London, London, UK
[2] UCL Genetics Institute, University College London, London, UK
[3] Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[4] Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, CA, USA
[5] Centre for Evolution and Cancer, Institute of Cancer Research, London, UK
[6] Department of Genetics, Evolution and Environment, University College London, London, UK

Abstract

Phylogenetic trees based on copy number profiles from multiple samples of a patient are helpful to understand cancer evolution. Here, we develop a new maximum likelihood method, CNETML, to infer phylogenies from such data. CNETML is the first program to jointly infer the tree topology, node ages, and mutation rates from total copy numbers of longitudinal samples. Our extensive simulations suggest CNETML performs well on copy numbers relative to ploidy and under slight violation of model assumptions. The application of CNETML to real data generates results consistent with previous discoveries and provides novel early copy number events for further investigation.

Keywords: Phylogeny inference, Copy number alteration, Maximum likelihood, Model of evolution, Low-coverage sequencing

Background

Phylogenetic trees have been widely used in the study of cancer, providing important insights into carcinogenesis [1]. Various markers have been used for phylogeny inference, including data derived from comparative genomic hybridization (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 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 SVs or CIN [11], such as high grade serous ovarian cancer [12], oesophageal cancer

[13], and osteosarcoma [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, 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) single 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) bulk sample tree when each bulk sample is assumed to be homogeneous, as in MEDICC [18] and PISCA [3], where each tip represents a bulk sample. Here, we distinguish single cell tree and bulk sample tree mainly because of intra-tumour heterogeneity (ITH) in bulk samples and scalability in phylogeny inference of hundreds to thousands of single cells which are typical in practice, although some methods can reconstruct both types of trees, such as MEDICC2 [9].

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 and more expensive than bulk-seq [15, 19]. Low-coverage bulk-seq, such as shallow WGS (sWGS), are instead more cost-effective and accessible, especially for SV-driven cancers [11]. They have been widely applied to detect CNAs, particularly on formalin-fixed paraffin-embedded samples, which are commonly available for diagnostics but have low DNA quality [13, 20–23], and cell-free DNA in plasma [24, 25]. In addition to multi-regional sWGS samples at one time point, there have been sWGS data for patient samples taken over time and space during a longer time period, such as in the surveillance of Barrett's oesophagus (BE) [13] and inflammatory bowel disease [23]. There are also longitudinal sWGS data from experimental evolution studies that use organoids to study the process of tumorigenesis [26], which seem promising to track cancer evolution [27]. The recent development of non-invasive liquid biopsy approaches allows the generation of sWGS data from cell-free DNA samples at multiple time points throughout the disease progression [25, 28]. Potentially, more longitudinal samples will be sequenced by the human tumour atlas network that aims to obtain multiparametric spatio-temporal data of cancers during their evolution from precancerous lesions to advanced disease [29]. These longitudinal samples contain temporal information that can be used to estimate the timing and rates of CNAs, which are important parameters in carcinogenesis, yet rarely studied. However, only a few reliable methods exist to detect CNAs from sWGS data and the detected copy numbers are often relative to the ploidy of the genome, called relative copy numbers [30]. 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 bulk 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 realism 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 [31], which creates complex dependencies across the genome, such as overlaps, back mutations, convergent and parallel evolution [32]. Therefore, the infinite sites or perfect phylogeny assumption, which is commonly used in inferring 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, 33].

Some methods transform original copy number calls into presence or absence of changes (breakpoints) [6, 34], 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 method with the Fitch algorithm [23, 35] and distance matrix methods based on Euclidean [36] or Manhattan [37] distances, and hence they may underestimate the true evolutionary distance as no correction of hidden changes is applied [9, 34]. Other methods use copy number transformation (CNT) 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) [38] and WGD [9, 39]. The weighted versions of both models allow the estimation of CNA rates in term of event probabilities [17, 38], 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 [40]. Although Markov chains often assume 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 to detect CNAs in sc-seq data [41]; Elizalde et al. used the product of 23 Markov chains to model numerical CIN of individual chromosomes in clonally expanding populations [42]. 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 [43, 44], and hence are not easily adapted for more bespoke mutation models that will be required for understanding carcinogenesis. 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 incorporates WGD into the previous model of segment duplication/deletion and implicitly accounts for chromosome or arm level gain/loss by grouping segments on the same chromosome or arm together [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 phylogeny from total copy numbers of multiple samples. To the best of our knowledge, this is the first method to jointly infer the tree topology, node ages, and mutation rates of temporal patient samples from (relative) total CNPs called from sWGS data. 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. The results on extensive simulations suggest that CNETML accurately recovered the tree topology, node ages, mutation rates, and ancestral CNPs when there were sufficient CNAs present in the data with at least two sampling time points whose time difference may be smaller than three years. 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 accuracy 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 assumptions 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 are worth further validations, suggesting the utility of CNETML.

Results

Overview of CNETML

The input of CNETML (Fig. 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 the 'Methods' section 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

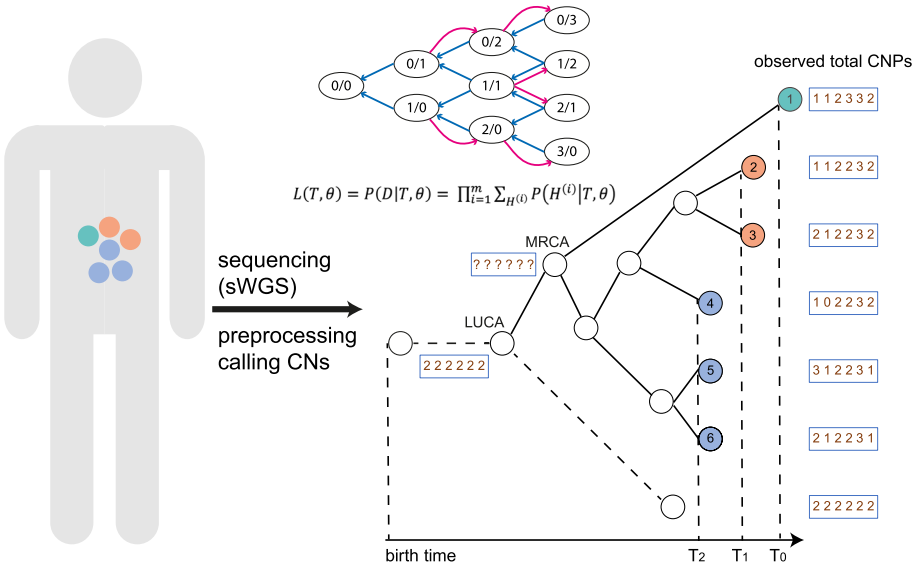

Fig. 1 The schematic overview of CNETML. Samples may be taken from a patient at different locations and times during surveillance and get sequenced (with sWGS) and analysed to generate copy numbers (CNs). Given the CNPs and/or sampling times of all the patient samples, CNETML aims to infer a sample tree 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 to get a binary tree. 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, which was used to constrain the age of LUCA in inference. 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 and the value in each oval representing a possible combination of copy number for haplotype A and B, denoted by c_A/c_B. 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 (T_0, T_1, and T_2) are denoted by different colours

obtained by joint segmentation across samples, namely merging consecutive bins with the same copy number in a sample with change points aligned across all the samples. In CNA detection, the general steps include binning, bias removal, segmentation, and copy number assignment [19, 45]. 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 where the site is natually the individual nucleotide, it is feasible to consider a bin or a segment as a site. We will discuss the ramifications of using bin or segment as a site in Results.

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 diploid 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 Additional file 1: Table S1 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. Note that (c_A, c_B) and (c_B, c_A) are distinguishable in the data when haplotype-specific copy numbers are provided. 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 by $(c_{1j}, c_{2j}, ..., c_{mj})$. The CNPs for all the n sampled genomes form a data matrix of n rows and m columns, denoted by D. 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. We call variant sites with unique site patterns as unique variant sites.

The likelihood for a tree T of n samples with parameters θ, $L(T, \theta)$, is the probability of observing D at the tips of T given θ. The Markov model specified by Q allows the computation of $L(T, \theta)$ by taking the products of probabilities at individual sites:

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

where $D^{(i)}$ is the i_{th} column of D. When taking total CNPs as input, we revise $L(T, \theta)$ to incorporate haplotype-specific copy numbers as missing information, which is similar to the handling of ambiguities in a nucleotide substitution model [40]:

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

where H is a data matrix of unknown haplotype-specific copy number states that are compatible with 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). In other words, the haplotype-specific copy numbers are latent variables, and the likelihood is an average over them.

We computed $L(T, \theta)$ with Felsenstein's pruning algorithm [46] with a few adaptations described in Methods. $L(T, \theta)$ was maximized by minimizing its negative logarithm function with L-BFGS-B algorithm [47], a numerical iterative method with

bound constraints. 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, which is efficient for trees of no more than seven samples. The other is heuristic search, which is adapted from the approach in IQ-TREE [48], a popular ML phylogeny inference program, and efficient for larger trees.

When there are at least three samples taken at no less than two time points, it is feasible to estimate mutation rates according to the differences of CNPs and sampling times, similar to the dating of virus divergences [40]. Although mutation rates during neoplastic progression are likely to change over time due to CIN [3], there are few studies on the rates of CNAs and how they change. Given insufficient information in the data, it is unlikely that reliable estimates of parameters resulting from rate changes can be obtained in the current maximum likelihood framework. Therefore, we assumed constant mutation rates under a global clock as a reasonable trade-off, which helps to get approximate early CNA timing information that is indicative of the potential disease onset time [49]. We jointly estimated the tree topology, mutation rates, and node ages (starting from 0 at birth time) with the following constraints in optimization: (1) The age of each internal node must be smaller than all its descents; (2) The age of root node must be smaller 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 Methods for more details). When all the samples are taken at the same time, the model is unidentifiable as there is no information to estimate mutation rates and node ages separately [40]. Therefore, $\theta = (l_1, l_2, ..., l_{2n-1})$, 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. Here, we separate $(u_0 + e_0)$ and t_i for the convenience of implementation within CNETML.

Ancestral reconstruction may suggest early CNAs that are likely cancer driver events and useful for early diagnostics. Therefore, we reconstructed ancestral states at unique variant sites based on the obtained ML tree using classical methods, including both marginal reconstruction of the most recent common ancestor (MRCA) node and joint reconstruction of all ancestral nodes [50, 51]. For marginal reconstruction, we computed the posterior probability of each possible copy number state for MRCA and assigned the state with highest probability to each site. For joint reconstruction, we assumed the best reconstruction is obtained when the root has normal diploid copy number states.

We used bootstrapping to measure the uncertainties of an estimated ML tree T_m [40]. 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` [52].

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 (Fig. 2, see the 'Methods' section 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 to obtain a binary tree. 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 [40]. When CNAs of multiple scales are considered, events were simulated by generating exponential waiting times with each site being a bin of fixed size (500 Kbp by default), which allows more complex models of evolution and the recording of more detailed information for each event. CNETS generates files 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 the 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

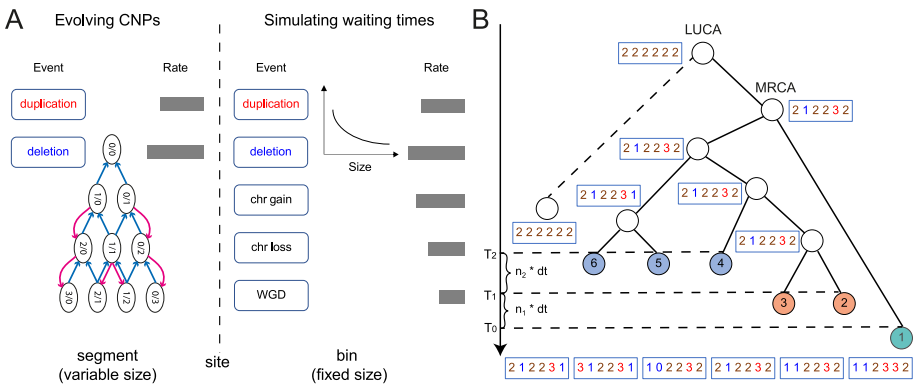

Fig. 2 The schematic overview of CNETS. **A** Two modes of simulation implemented in CNETS. One is simulating CNPs directly for site duplication/deletion based on segments of variable size, which follows exactly the Markov model used for phylogeny inference. The other is simulating waiting times for 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 the number 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 $T_0, T_1 = T_0 + n_1 * dt$, and $T_2 = T_1 + n_2 * dt$, with dt, n_1, and n_2 being integers and $1 <= n_1, n_2 <= 6$ (the number of samples)

present (Additional file 1: Table S2). 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) (Additional file 1: Fig. S1). 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.

The accuracy of tree inference was measured by normalized Robinson–Foulds (RF) distance [53] and branch score distance [54]. The normalized RF distance is commonly used to quantify the topological differences between trees due to easy computation. Each branch in the tree divides the tips into two sets, called a partition. RF distance is simply the number of partitions present in one tree but not the other, whose value ranges from 0 (complete agreement) to twice the number of internal branches ($2n - 4$ for a rooted tree with n tips). The normalized RF distance is RF distance divided by the maximum possible value. The branch score difference is the square root of squared differences between branch lengths in the two trees, which can evaluate the accuracy of branch length (divergence time). The smaller values of the normalized RF distance and branch score difference indicate more accurate estimation. These distances were computed with function `treedist` in R library `phangorn` [55]. The branch length was measured by time in year in the computation of branch score distance. When the mutation rates were not estimated, u_0 and e_0 were set to be real values used for simulation. We also computed the differences between the estimated and true values of duplication/deletion rates and LUCA age to check the accuracy of their estimation. To measure the accuracy of ancestral reconstruction, we computed the fraction of correctly recovered states over the number of unique variant sites 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 statistically consistent, which means that the ML tree will converge to the true tree when the size of the data (the number of sites) increases [40]. 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 Fig. 3A, all the simulated trees were better recovered with more sites and higher mutation rates, which confirms the statistical consistency of CNETML. Because what is informative for inference is the number of (unique) variant sites, we also counted the number of (unique) variant sites in the simulated data, which suggested that topologies of more than 80% of simulated trees were correctly reconstructed with between 90 and 180 variant sites (between 16 and 32 unique variant sites) when $m = 1000$ and $u = e = 0.001$ (per haplotype per site per year) (Additional file 1: Fig. S2, Table S3). In the subsequent simulations, we fixed the number of sites $m = 1000$ when not stated.

We tested the consistency of the exhaustive and heuristic tree search algorithm on simulated data with 5, 6, and 7 samples under different mutation rates. Although the heuristic search had decreased performance with increasing number of samples, as reflected by the increase of minimal negative log likelihood values compared with those obtained from the exhaustive search (Additional file 1: Fig. S3), it reconstructed the same tree topologies as the exhaustive search on data with 5 samples and around 70% of the same tree topologies on data with 7 samples (Additional file 1: Table S4), with slightly larger errors on the branch length estimation (Additional file 1: Fig. S4). To check how the heuristic tree search algorithm performed on data with a larger number of samples, we applied it on simulated data with 10, 20, 30, 50, 100, and 200 samples under different mutation rates (Additional file 1: Fig. S5). In general, the reconstructed trees were more similar to the ground truth with more mutations and the distances of reconstructed trees to simulated true trees increased almost linearly with the number of samples, with the coefficient being around 0.01 for normalized RF distance and around 0.36 (0.8) for branch score distance when mutation rate was high (low). Therefore, the topologies of reconstructed trees were less affected by the increasing number of samples, whereas the branch lengths were more affected especially when there were fewer observed mutations.

We also checked the performance of CNETML on reconstructing ancestral states on the simulated data with 1000 sites under different mutation rates by supplying 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 (Fig. 3B). The fraction of accurately reconstructed sites decreased with larger mutation rate due to the presence of more variant unique sites. Joint reconstruction appeared more accurate, probably because it computes the joint probability of all the internal nodes [40].

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 impact 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 (Fig. 3A, C). However, the accuracy of reconstructing ancestral states was better with haplotype-specific CNPs (Fig. 3B). The analysis on PCAWG dataset [56] shows that around 80% samples have no more than 10% of the genome with cn-LOH (Additional file 1: Fig. S6), 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 the tree topology, node ages, and mutation rates when the samples were taken at different time points. The reliability of rate estimation depends on the extent of time differences at the tips, with larger differences providing more information for inference [57]. Since the sampling time differences for a patient may range from one year to 15 years as in a BE dataset [13], we simulated data under different temporal signal strengths, $dt \in \{1, 3, 5\}$ (years), where the range of

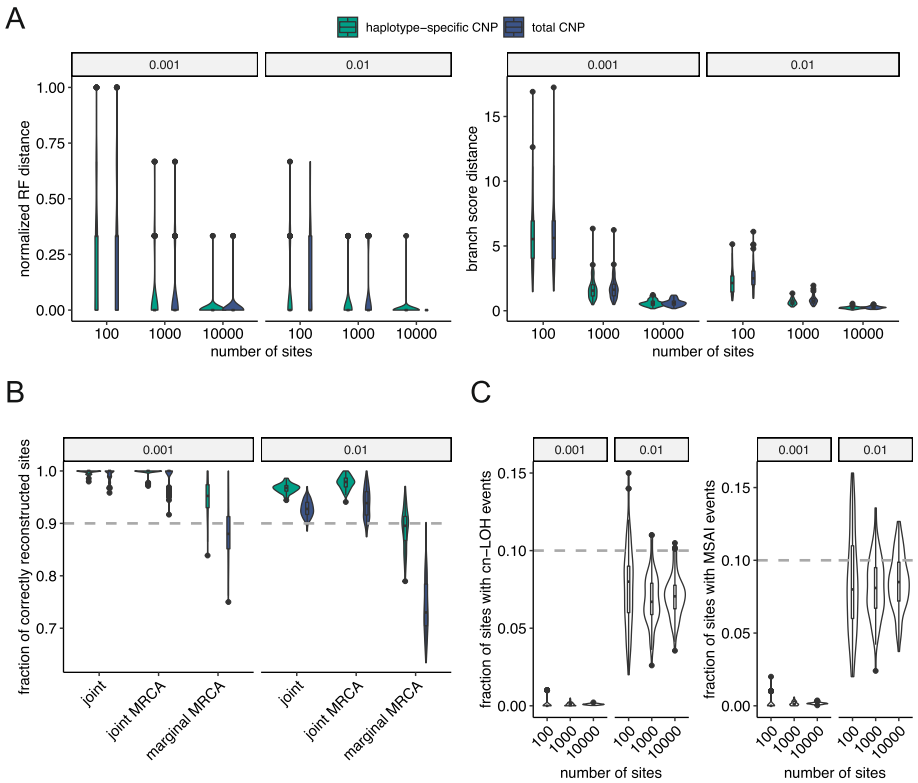

Fig. 3 The performance of CNETML on reconstructing trees and ancestral states with total or haplotype-specific CNPs when samples are taken at the same time. **A** The accuracy of phylogeny inference on data simulated with different number of sites and mutation rates, measured by the normalized RF distance and branch score distance, respectively. **B** The accuracy of ancestral state reconstruction on simulated data with 1000 sites under different mutation rates, measured by the fraction of correctly reconstructed unique variant sites. **C** The fraction of cn-LOH and MSAI events in the simulated data. There are five samples 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

simulated sampling times approximated real data and samples simulated under a larger dt generally had larger time differences (Additional file 1: Fig. S7). We also simulated data with 10,000 sites to check the consistency of CNETML during joint estimation. We grouped the simulated data by the mean pairwise absolute difference of tip relative times (denoted by T_m) into three groups: "small" difference when $T_m < 3$, "intermediate" difference when $3 <= T_m < 7$, and "large" difference when $T_m >= 7$. The numbers of samples in each group are shown in Additional file 1: Table S5. 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 Methods on how to get initial trees)) and (u^0, e^0) has to be specified manually. Since the L-BFGS-B algorithm may converge to a local peak on the likelihood surface of a tree, we tried different 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 (Additional file 1: Fig. S8).

Therefore, we recommend starting from 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 Fig. 4.

The joint estimation was generally better with higher mutation rates and a larger number of sites (Fig. 4A, B). As shown in Fig. 4A, the sampling time differences did not affect much the inference of tree topologies but yielded better branch length estimation when being larger, and the estimated median LUCA ages were closer to real values when the sampling time differences were not small despite larger variances, which was probably caused by a wider range of the simulated LUCA ages (Additional file 1: Fig. S9). The mutation rates (Fig. 4B) were slightly underestimated with fewer mutations and small sampling time differences and more accurately estimated otherwise. In summary, CNETML inferred phylogenies well when there was sufficient information in the data, with larger mutation rates or sampling time differences leading to higher accuracy. We also ran CNETML on haplotype-specific CNPs of data simulated with $dt = 5$ years and 1000 sites to compare with the results when using total CNPs (Fig. 4C), but similar to our previous results in Fig. 3, we did not observe large differences.

To see how the joint estimation affects the inference of parameters, we also ran CNETML on the data with tree topology fixed. The results (Additional file 1: Fig. S10) suggest that joint estimation performed almost as well as when the tree topology was fixed.

Performance on relative copy numbers

CNAs called from sWGS data with common tools, such as QDNAseq [20], are often values relative to ploidy, 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 some similar strategies in CNETS to simulate relative copy numbers by using baseline and rounding after scaling with baseline. We tested using either $2^{N_{WGD}}$ or rounded mean copy number as baseline, where N_{WGD} is the number of WGD events in a genome. For haplotype-specific copy numbers, we used the rounded mean copy number for each haplotype as baseline. For total copy numbers, we used half the rounded mean copy number as baseline. We tested both direct rounding to the nearest integer and random rounding as in [3]. CNETS output simulated relative total CNPs by reducing the normalized copy numbers by the normal ploidy, with values smaller than -2 and larger than 2 set to -2 and 2 respectively for consistency with QDNAseq output.

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, which generated data of four types according to the distribution of WGD among samples: clonal WGD where WGD appears in all samples, multiple WGDs where

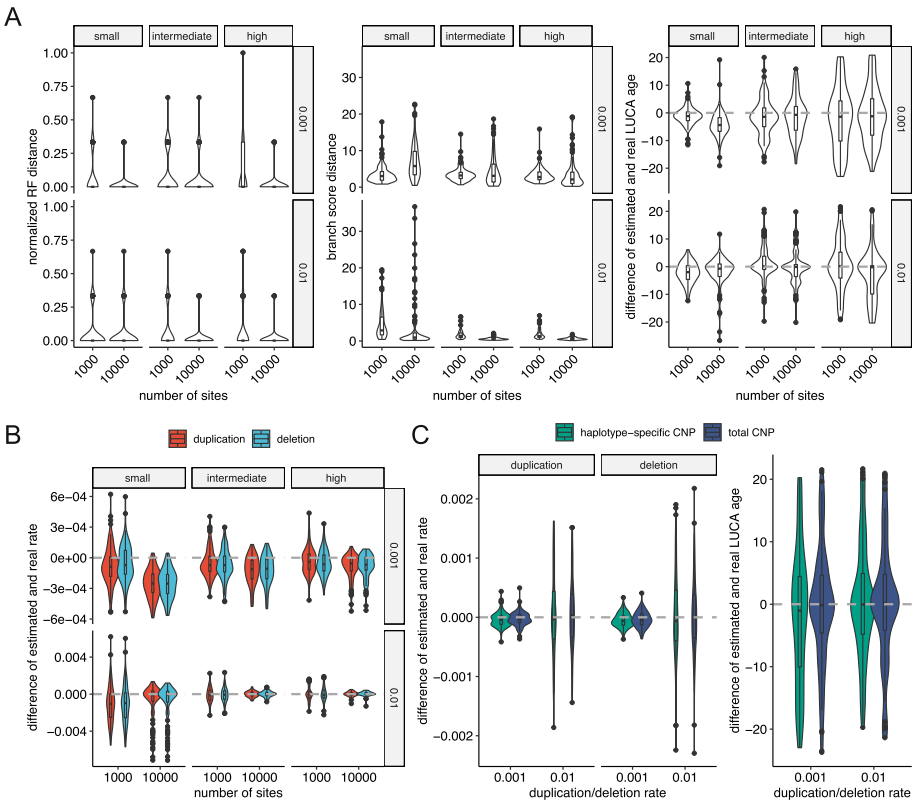

Fig. 4 The performance of CNETML on jointly estimating the tree topology, node ages, and mutation rates on data simulated with different sampling times and mutation rates. **A** The accuracy of phylogeny inference, measured by the normalized RF distance, branch score distance, and difference of estimated and real LUCA age, respectively. **B** The accuracy of mutation rate estimation, measured by the difference of estimated and real rate. **C** The accuracy in estimation of the mutation rate and LUCA age with total or haplotype-specific CNPs on simulated data with $dt = 5$ years, measured by the difference of estimated and real rate and difference of estimated and real LUCA age, respectively. There are five samples in each simulated tree and 100 datasets for each parameter setting. Grey dashed line: real values. Box plots have the same interpretations as those in Fig. 3

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. 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 WGD distribution in the data (Fig. 5, Additional file 1: Fig. S11 and S12). When using random rounding, CNETML on relative CNPs inferred less accurate tree topologies on datasets with clonal WGDs. When using direct rounding, the results were similar when using either $2^{N_{WGD}}$ or the rounded mean copy number as the base line. Since it is hard to know N_{WGD} in real data, we report the results of using the rounded mean copy number as baseline here. As expected, CNETML on absolute total CNPs reconstructed inaccurate phylogenies and misestimated 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 without WGD, CNETML performed similarly 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 mostly similar to the truth and MEDICC2 on absolute allele-specific CNPs, although the deletion rates were largely underestimated along with slightly overestimated duplication rates, which is probably due to greater signal loss when converting all the copy numbers relative to the baseline. On data with multiple subclonal WGDs, CNETML on relative CNPs achieved slightly better performance than MEDICC2 and the mutation rate estimates were mostly accurate with slight underestimation of deletion rates. On data with single WGD, CNETML on relative CNPs achieved similar performance to MEDICC2 and accurate mutation rate estimation with slightly larger variance compared to cases with no WGD. In summary, it seems entirely feasible to recover the phylogeny directly from relative copy numbers, such as those from QDNAseq, and absolute copy numbers which have been scaled appropriately to mitigate the effects of WGD. For further validation of the inference on relative copy numbers, empirical information or methods to detect WGD [58] or call absolute copy numbers [30] from sWGS data may be used to estimate the presence of WGD and understand its effect on the results.

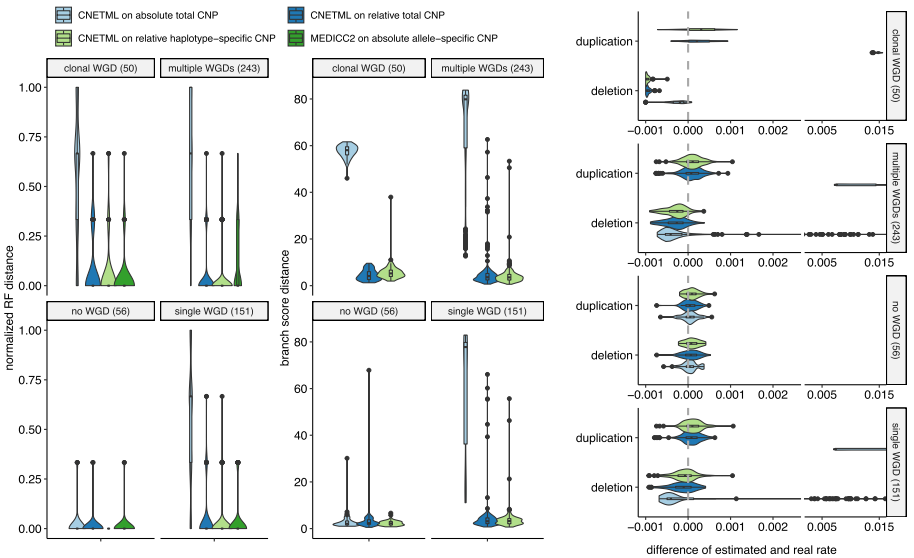

Fig. 5 The performance of CNETML on relative copy number data. The relative copy numbers were obtained by using the rounded mean copy number as baseline and direct rounding. There are 500 simulated datasets in total, which are divided into four groups by the types of WGDs. The number of datasets in each group is shown in brackets. MEDICC2 was excluded when comparing branch score distance because the branch length in a tree built by it has a different meaning (the number of events between CNPs of two nodes based on CNT model) and it is hard to compare fairly. Ninety-one outlier data points with values larger than the maximum of *x*-axis on datasets with subclonal WGDs are excluded in the plot of mutation rates for better visualization of the majority data. Box plots have the same interpretations as those in Fig. 3

Performance under violation of model assumption

ML inference of phylogenetic trees using sequence data was shown to be highly robust to violations of assumptions [40]. 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 [59] 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 Methods for details of the post-processing).

As expected, the inferred phylogenies and mutation rates were more dissimilar to the ground truth with larger CNA sizes due to information loss as a result of overlaps, with overestimated branch lengths and mutation rates and slightly underestimated LUCA age (Fig. 6A). However, even when the mean duplication/deletion size was 5 Mbp, longer than the typical size to define focal CNA (3 Mbp [59]), the bias 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 (Fig. 6B). These results suggest that slight violation of the 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 small because the site patterns in the input data were similar in both cases, except that bin-level data might contain more sites and a larger number of the same pattern which lead to longer computing time. With larger CNAs, the estimated rates obtained using bin-level data generally had slightly higher values and larger variances than those obtained using segment-level data (Fig. 6B, Additional file 1: Fig. S13). This increase in estimator bias and variance is likely due to violation of independent sites assumption which causes more pairwise site correlation [60] and higher among site rate variation [61]. Therefore, we recommend using segment-level data for faster computation, less correlation between segments, and better interpretability in practice.

In addition, we assume 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. To check the sensitivity of CNETML to errors in copy number calling, we simulated data of 5 samples with increasing fraction of sites with error (0, 0.1, 0.3, 0.5, 0.7, see the 'Methods' section for details). The results (Additional file 1: Fig. S14) suggest that more than 70% of the tree topologies were correctly recovered with relatively small branch length estimation errors when the samples were at the same time point and the mutation rate was high, although the performance of CNETML generally decreased with the increasing error rates especially when the mutation rate was low. We further checked how

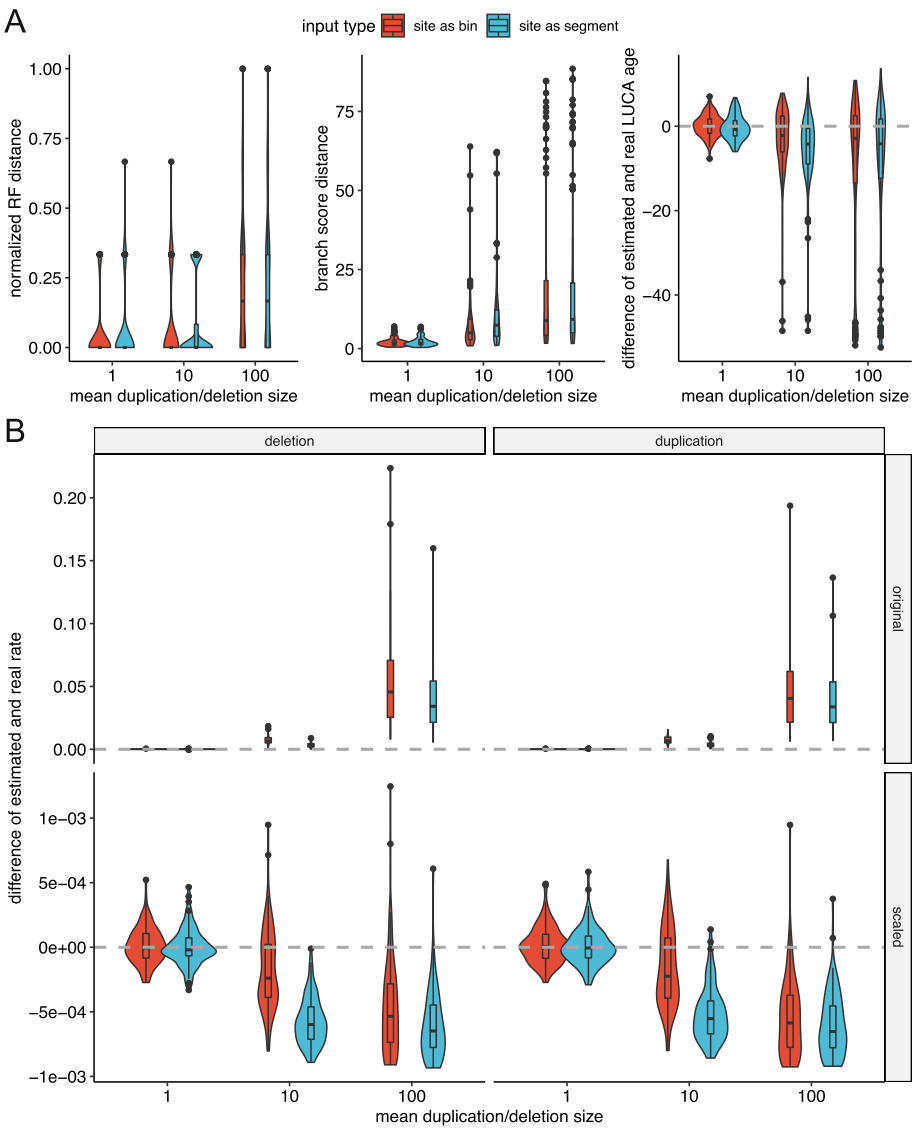

Fig. 6 The performance of CNETML with different types of sites on data simulated with mean duplication/deletion of different sizes. **A** The accuracy of phylogeny inference, measured by the normalized RF distance, branch score distance, and difference of estimated and real LUCA age, respectively. **B** The accuracy of mutation rate estimation before and after scaling, measured by the difference of estimated and real rate. There are five samples in each simulated tree and 100 datasets for each parameter setting. Box plots have the same interpretations as those in Fig. 3

CNETML performed when the samples were at different time points ($dt = 1$ year) under high mutation rates ($u = e = 0.01$ per haplotype per site per year) (Additional file 1: Fig. S15). CNETML still correctly reconstructed no less than 70% of the tree topologies although the branch length estimation errors became larger. With increasing error rates, the estimated duplication rate and LUCA age decreased, whereas the estimated deletion rate slightly increased.

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 [62]. To check the sensitivity of CNETML to the monoclonal assumption, we postprocessed the data simulated by CNETS to get data with subclonal structure (see the 'Methods' section for details). We simulated data with all the samples having the same number of clones $n \in \{3, 4\}$ (including the normal clone). For each sample, we fixed the fraction of the original clone (the clone with the same identifier as the sample) to $f_d \in \{1, 0.8, 0.6, 0.5, 0.4\}$ and sampled the fractions of the other clones from uniform Dirichlet distribution whose sum was $1 - f_d$. The results (Additional file 1: Fig. S16) suggest that CNETML performed well when the original clone was dominant with a larger fraction than the other clones in the sample when the samples were at the same time point, with no less than 70% of the tree topologies correctly recovered. Since CNETML performed similarly with a sample having either 3 or 4 clones, we then checked its performance on simulated samples at different time points ($dt = 1$ year) with each sample having 4 clones (Additional file 1: Fig. S17). More than 55% of the tree topologies were still correctly recovered when the original clone was dominant. When the mutation rate was low, the estimated duplication and deletion rates were close to the true values when the original clone was dominant, whereas the estimated LUCA age slightly increased with decreasing fraction of original clone. When the mutation rate was high, the estimated duplication and deletion rates were more affected by ITH and declined almost linearly with decreasing fraction of original clone, whereas the LUCA age was generally underestimated.

Application to Barrett's oesophagus patients

To demonstrate the applicability of CNETML on real data, we applied it to data for two BE patients in Fig. 1 of [13], where CIN was used to predict risk progression (Fig. 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, set those smaller than -2 to -2 and larger than 2 to 2, 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 for patient 51 and 62 for patient 20, the mean age of all nonprogressors and progressors in the cohort at diagnosis of BE, 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 T_b and optimized node ages and mutation rates to get T_b', with the initial mutation rates set to the estimated rates on T_b. We ran another 100 nonparametric 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

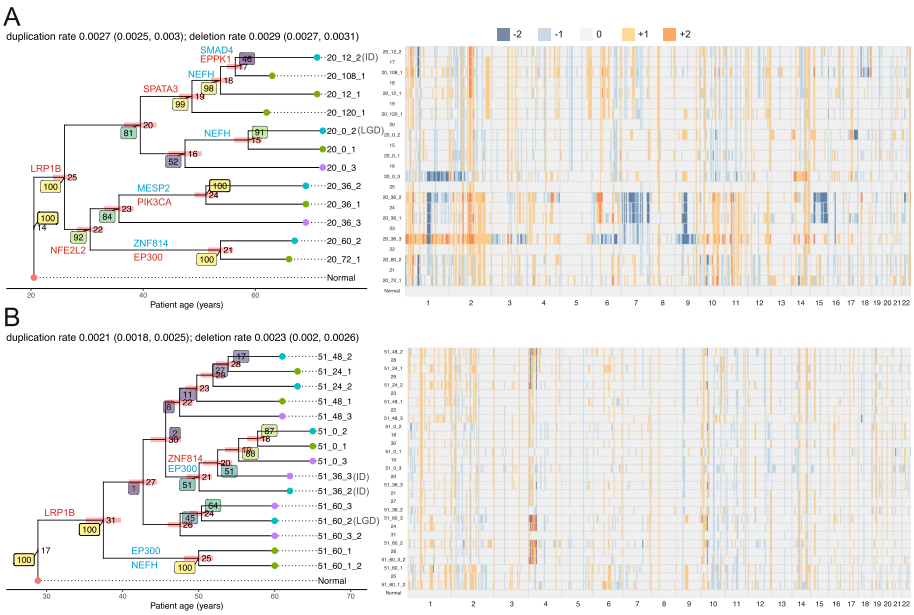

Fig. 7 The ML trees and ancestral states reconstructed by CNETML for patient 20 (**A**) and 51 (**B**). The bootstrap support values are shown in coloured rectangles with lighter colours suggesting stronger support. The coloured bars at the internal nodes show the confidence intervals of node ages. Each sample is denoted by "patientID_timeID_locationID", where timeID is the month before final endoscopy and locationID is the relative esophageal sample location. There are two sets of samples taken at the same time and location for patient 51, which are indicated by "_2" at the end. For patient 20, the samples taken at 12 months before final endoscopy location 2 (20_12_2) is Indeterminate (ID), the final endoscopy at location 2 (20_0_2) is Low-Grade Dysplasia (LGD), and all the other samples are Non-dysplastic BE (NDBE). For patient 51, the samples taken at 36 months before final endoscopy location 2 and 3 (51_36_2 and 51_36_3) are IDs, the sample taken at 60 months before final endoscopy at location 2 (51_60_2) is LGD, and all the other samples are NDBE. 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 parentheses in the title of the plot for each patient

computing their overlap with cancer-related genes from COSMIC Cancer Mutation Census with keyword "oesophag" in the description of disease [63] 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 (Fig. 7A). Although the branch connecting the samples taken at 12 months location 2 and 108 months location 1 (times before final endoscopy) had only 46% of support, they shared a loss of gene SMAD4, which was shown to promote tumorigenesis from BE toward esophageal cancer [64] and hence suggested the reliability of this branch. The tree topology for patient 51 had much poorer support due to the lack of changes in copy numbers (Fig. 7B). 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 seems consistent with previous results for BE patients [3]. The LUCA age approximates the onset time of BE, since CNAs are likely to begin after BE establishes in the oesophagus. The results suggest patient 20 had BE about 40 years before the first sample, about 10 years earlier than patient 51, which also seems consistent with previous results where two progressors

had younger LUCA ages than two nonprogressors [3]. The estimated phylogenies also show a longer dwell time (the time a patient has lived with the precursor) of BE in the progressor (patient 20) than the non-progressor (patient 51), which is consistent with previous modelling results [49]. From the reconstructed CNPs of the MRCAs of both patients (node 25 for patient 20 in Fig. 7A and node 31 for patient 51 in Fig. 7B), we found gene LRP1B included in a region on chr 2q with copy number gain (see Additional file 2: Table S6 for the complete list of overlaps). The original average relative copy number for patient 20 (1.4) across all samples in this region is about twice that for patient 51 (0.6), suggesting more gains 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 with 185 cases [65]. The CNP of the MRCA 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 CNPs of the MRCA of patient 51 and the ancestors of the top and bottom lineages of patient 20 (node 20 and 22 in Fig. 7A) overlapped with the predictive region whose associated CV is 5.090 (ranked 6th among 75 regions) [13]. The CNP of the ancestor of the bottom lineage of patient 20 also had a region of gain overlapping with gene NFE2L2 on chr 2q, which has about 11.41% cases with copy number gains in TCGA-ESCA cohort [65]. For patient 51, the lineage starting from node 21 had a region of gain overlapping with gene ZNF814, which has 13.04% cases with copy number gains in TCGA-ESCA cohort [65]. All these findings suggest that the phylogenies and mutation rates inferred by CNETML are biologically meaningful and additional insights into carcinogenesis can be gained from the reconstructed ancestral CNPs.

Discussion

In summary, we developed CNETML, a new ML method designed to reconstruct the evolutionary history of multiple samples of a single patient which may be taken 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 the node ages and mutation rates by year when patient samples were taken at two or more time points, which appears to become more available with the development of sequencing techniques. The estimation provides approximate timing of initiating CNAs and hence possible onset age of the disease such as BE in a patient, which cannot be detected clinically due to asymptomaticity but is important in carcinogenesis and may be helpful in cancer screening and surveillance programs [49]. 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 (such as sampling time differences of more than 5 months) 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 LUCA 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 have been phased to distinguish haplotypes, 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 bulk sample tree where each tip is a CNP from a patient bulk sample, it can be used to build trees from CNPs of subclones or single cells, since the main input is simply an integer copy number matrix where the rows can represent subclones or cells. For example, we applied CNETML on total and haplotype-specific copy numbers of subclones detected from single cell DNA sequencing of a breast cancer patient [66] (see the 'Methods' section for details on input preparation). The results (Additional file 1: Fig. S18 and S19) suggest that CNETML reconstructed phylogenies largely consistent with those in [66] and the uncertainties in the inference due to the lack of information were well reflected by the bootstrap support values. Due to the expensive computational cost of applying full model-based phylogenetic inference to thousands of single cells that give rise to a huge number of possible tree topologies [67], CNETML is only practical for hundreds of cells. As a demonstration, we applied CNETML on two single cell datasets from two breast cancer patients [37]. We randomly sampled 100 cells from each dataset and compared the reconstructed trees with those reported in [37] and the trees reconstructed by MEDICC2 [9] (Additional file 1: Fig. S20 and S21, Additional file 3: Table S7). As seen from the alignments of tree topologies and relative total copy number heatmaps, CNETML generated three and five large groups of cells for the two patients respectively, similar to those obtained in [37] and [9]. The relationships among individual cells in each group and the exact tree topologies had large bootstrap uncertainties due to the large sample size and lack of mutational information in the data. 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 [68].

In principle, the likelihood-based approach adopted in CNETML is more sophisticated than distance matrix and maximum parsimony methods. To deal with the specific properties of (relative) CNPs and allow for more flexible evolutionary models specific for carcinogenesis, we implemented a novel tool rather than using existing frameworks designed for traditional phylogenetic inferences, such as BEAST [43, 44]. Future development of the model could include Markov chains at different scales to incorporate chromosomal and/or arm level gain/loss and WGD despite the challenging combinatorial complexity, and the use of penalized maximum likelihood estimation to incorporate prior knowledge on parameters when there are insufficient information in the data [69]. To apply CNETML to data of a much larger scale, such as CNAs from thousands of single cells, further optimizations, especially the acceleration of likelihood computation, are also required. Another development would be the estimation of varying mutation rates in different lineages under a relaxed local clock [70]. Finally we can extend to a fully

Bayesian approach, which can impose informative prior distributions and naturally provide a measure of uncertainty of the inference (posterior probabilities of sampled trees) despite a higher computational cost.

The inference of sample 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. To improve the robustness of CNETML to errors in the input data, we consider incorporating an error model as future work, 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]. Although CNETML performed generally well when there was a dominant clone in each sample, given data of higher resolution, it would be helpful to quantify ITH and build clone trees.

Conclusions

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

Methods

Preprocessing of input data

The input CNPs for CNETML are mainly obtained from common CNA calling methods for sWGS data. For example, QDNAseq [20] is often used to get relative copy numbers by computing read counts in fixed-sized bins, doing segmentation, and calling copy numbers with CGHcall [71] 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 assumed the same binning or segmentation across all samples to get consistent sites. When raw copy number calls were at bin level, segments were obtained by merging consecutive variant bins on the same chromosome with the same copy number across all samples.

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

CNETML can also be applied to copy number data of subclones or single cells. Here, we show how we processed the haplotype-specific copy numbers of suclones detected from single cell data of a patient in [66]. WGD was identified as a clonal event in nearly all tumour cells of the patient. To mitigate the effect of WGD in the inference, we processed the data to get relative haplotype-specific copy numbers by dividing the absolute copy numbers of each haplotype by the rounded mean value for that haplotype and rounding the resultant values to the nearest integer. Then we obtained the relative total copy numbers by adding up the relative copy numbers of each haplotype. In terms of the single cell data from the two patients in [37], we used the allele-specific copy number data derived in [9]. As WGD was clonal for each patient, we computed relative allele-specific copy numbers in a similar way to that for each haplotype and then added them up to get the relative total copy numbers.

The computation of likelihood

The probabilities of observing data at a single site $P(D^{(i)}|T,\theta)$ was computed with Felsenstein's pruning algorithm by post-order traversal of T, in which each node is visited only after all its descendants have been visited [46]. The computation can be expressed as a recursion that computes $L_i(d_i)$ for each node i at each possible haplotype-specific copy number state d_i, the conditional probability of observing data at the descendant tips below i. Let $p_{d_i d_j}(t_j)$ represents the transition probability of d_i becoming d_j after time t_j, where i and j are two nodes in T connected by a branch of length t_j. Suppose the root (LUCA node) is r with state $d_r = 4$ since it is assumed to have normal diploid copy number, then:

$$P(D^{(i)}|T,\theta) = L_r(d_r) = p_{d_r d_m}(t_m) L_m(d_m), \tag{4}$$

where m is the MRCA node connected to r with a branch of length t_m. When node i is an internal node with children node j and k,

$$L_i(d_i) = \sum_{d_j} [p_{d_i d_j}(t_j) L_j(d_j)] \sum_{d_k} [p_{d_i d_k}(t_k) L_k(d_k)], \tag{5}$$

where t_j and t_k are the lengths of the branch from node i to j and k respectively. When node i is a tip with observed haplotype-specific copy number state c_i,

$$L_i(d_i) = \begin{cases} 1 & d_i = c_i, \\ 0 & otherwise. \end{cases} \tag{6}$$

When the input copy numbers are total, there may be multiple haplotype-specific copy number states compatible with the observed value at tip i, denoted by set S_i, and hence

$$L_i(d_i) = \begin{cases} 1 & d_i \in S_i, \\ 0 & otherwise. \end{cases} \tag{7}$$

To improve efficiency in likelihood computation, the transition probability matrix, $P(t_j) = e^{Qt_j}$, was computed once with scaling and squaring method [72] 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.

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 year 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:

$$x_i = \begin{cases} t_1 & i = 1, \\ \frac{t_j - t_j^m - l_m}{t_i - t_i^m - 2l_m} & 1 < i <= n-1, \end{cases} \tag{8}$$

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:

$$\begin{aligned} d + nl_m <= x_1 <= A_0 + d, \\ 0.01 <= x_i <= 0.99, \quad 1 < i <= n-1, \end{aligned} \tag{9}$$

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 enumerated all the possible tree topologies for the given number of samples and then found the ML tree by optimizing the parameters. For heuristic 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 neighbour interchanges (NNIs) [48] 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` [52] and transformed into the formats acceptable by CNETML. When the input data is haplotype-specific, we treat each haplotype as a new site and appended the copy numbers of haplotype B after those of haplotype A to build the stepwise addition tree.

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.
2. Optionally simulate temporal samples for a patient of specified age. One way to generate samples at various time points with just one additional parameter, dt, is as below. Note that this simulation approach destroys the coalescence structure but it is sufficient to generate a sample phylogeny for the purpose of testing phylogeny inference methods like CNETML.
    (a) Assign random times (in year) to the tips by changing terminal branch lengths with multiples of dt.
    (b) Rescale the internal branches of the tree so that the tree height is no larger than the patient age at the last sampling time.
3. 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.
4. 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 4,401 bins of 500 Kbp. Each genome with m sites was initially represented by the set of sites, denoted by $G = (l_1, l_2, ..., l_m)$. We denoted the diploid genome by $G_d = [G, G]$, which was implemented by making a copy of G to represent the other haplotype. The final CNP of the genome $(c_1, c_2, ..., c_m)$ was computed for each site by adding up the number of copies across all the haplotypes when considering total copy number or across 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.
    (a) For segment duplication/deletion, randomly choose the start bin based on the rates across sites, the haplotype, and the size in the number of bins, where a duplication can be either tandem (duplicated at the end of the current location) or interspersed (inserted at a random position of a random chromosome) 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 $t >= l$.

To see how errors in the detected copy numbers affect phylogeny inference, we incorporated the simulation of data with different error rates, where errors may be caused by sequencing noise, sample purity, and ITH. When the fraction of sites with error f_e is specified, we randomly selected $round(m * f_e)$ sites to change their copy numbers. We sampled the new copy number from a Poisson distribution with the mean being

the original copy number at the site. If the new copy number is larger than c_{max}, we set it to c_{max}.

To see how ITH affects phylogeny inference, we added scripts to postprocess the data simulated by the main program to get data with subclonal structure. We assumed the originally simulated data is for clones rather than samples. Then we mixed clones to get the data for each sample by specifying the fraction of the original clone f_d and sampling the fraction of the other clones (including the normal clone) with uniform Dirichlet distribution which is then multiplied by $(1 - f_d)$.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02983-0.

Additional file 1: Supplementary file. Supplementary Tables and Figures.

Additional file 2: Table S6. The complete list of the overlaps of the reconstructed ancestral CNPs with cancer-related genes from COSMIC Cancer Mutation Census and predictive regions for two BE patients.

Additional file 3: Table S7. The complete list of the sampled cells together with their groups and colours as those in the trees reconstructed by CNETML in Additional file 1: Fig. S20 and S21 for two breast cancer patients. The cells which are not assigned to the same group in the trees reconstructed by both the other two methods are highlighted in yellow. The cells which are not assigned to the same group in the tree reconstructed by MEDICC2 alone are highlighted in blue.

Additional file 4: Review history.

Acknowledgements
The authors acknowledge the use of the UCL Myriad High Throughput Computing Facility (Myriad@UCL) and the UCL Department of Computer Science High Performance Computing Cluster, 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.

Review history
The review history is available as Additional file 4.

Peer review information
Tim Sands and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
CB, KC, and TG initialized the study. BL, CB, and ZY developed the model. BL and CB implemented the program. BL tested the program and wrote the initial draft of the manuscript. All authors read, reviewed, and approved the final manuscript.

Authors' Twitter handles
Twitter handles: @lubingxin (Bingxin Lu); @yosoykit (Kit Curtius); @trevoragraham (Trevor A. Graham); @zihengyang (Ziheng Yang); @cssb_lab(Chris P. Barnes).

Funding
CB and BL acknowledge funding from the Wellcome Trust (209409/Z/17/Z). KC received funding from the MRC HDR-UK programme (UKRI Rutherford Fund Fellowship). TG was funded by Cancer Research UK (A19771 and DRCNPG-May21_100001), the Wellcome Trust (202778/Z/16/Z), and the National Institute of Health (NCI U54 CA217376). This study has been supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/T003502/1, BB/R01356X/1) to ZY.

Availability of data and materials
CNETML and CNETS are freely available under GPLv3 license at Github [73] and Zenodo [74]. The simulated and processed real data used for generating results are available at Zenodo [75]. The original real data in [13] are available at https://www.nature.com/articles/s41591-020-1033-y (Source Data Fig. 1). The original real data in [66] are available at https://github.com/raphael-group/chisel-data/tree/master/patientS0/calls [76]. The original real data in [9] are available at https://doi.org/10.5281/zenodo.7300106 [77].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 March 2022 Accepted: 8 June 2023
Published online: 20 June 2023

References

1. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017;18(4):213–29.
2. Bakhoum SF, Cantley LC. The multifaceted role of chromosomal instability in cancer and its microenvironment. Cell. 2018;174(6):1347–60.
3. Martinez P, Mallo D, Paulson TG, Li X, Sanchez CA, Reid BJ, et al. Evolution of Barrett's esophagus through space and time at single-crypt and whole-biopsy levels. Nat Commun. 2018;9(1):794.
4. Zaccaria S, El-Kebir M, Klau GW, Raphael BJ. Phylogenetic copy-number factorization of multiple tumor samples. J Comput Biol. 2018;25(7):689–708.
5. Kuipers J, Tuncel MA, Ferreira P, Jahn K, Beerenwinkel N. Single-cell copy number calling and event history reconstruction. bioRxiv. 2020. https://doi.org/10.1101/2020.04.28.065755.
6. Salehi S, Dorri F, Chern K, Kabeer F, Rusk N, Funnell T, et al. Cancer phylogenetic tree inference at scale from 1000s of single cell genomes. bioRxiv. 2021. https://doi.org/10.1101/2020.05.06.058180.
7. Markowska M, Cakala T, Miasojedow B, Juraeva D, Mazur J, Ross E, et al. CONET: copy number event tree model of evolutionary tumor history for single-cell data. Genome Biol. 2022;23:128.
8. Andersson N, Chattopadhyay S, Valind A, Karlsson J, Gisselsson D. DEVOLUTION-A method for phylogenetic reconstruction of aneuploid cancers based on multiregional genotyping data. Commun Biol. 2021;4:1103. https://doi.org/10.1038/s42003-021-02637-6.
9. Kaufmann TL, Petkovic M, Watkins TBK, Colliver EC, Laskina S, Thapa N, et al. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. Genome Biol. 2022;23:241.
10. Liu Y, Edrisi M, Ogilvie HA, Nakhleh L. NestedBD: Bayesian inference of phylogenetic trees from single-cell DNA copy number profile data under a birth-death model. bioRxiv. 2022. https://doi.org/10.1101/2022.01.16.476510.
11. Macintyre G, Ylstra B, Brenton JD. Sequencing structural variants in cancer for precision therapeutics. Trends Genet. 2016;32(9):530–42.
12. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45(10):1127–33.
13. Killcoyne S, Gregson E, Wedge DC, Woodcock DJ, Eldridge MD, De La Rue R, et al. Genomic copy number predicts esophageal cancer years before transformation. Nat Med. 2020;26:1726–32.
14. Wu CC, Beird HC, Andrew Livingston J, Advani S, Mitra A, Cao S, et al. Immuno-genomic landscape of osteosarcoma. Nat Commun. 2020;11(1):1008.
15. Kuipers J, Jahn K, Beerenwinkel N. Advances in understanding tumour evolution through single-cell sequencing. Biochim Biophys Acta (BBA) - Rev Cancer. 2017;1867(2):127–38.
16. Chowdhury SA, Shackney SE, Heselmeyer-Haddad K, Ried T, Schäffer AA, Schwartz R. Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. PLOS Computational Biology. 2014;10(7):e1003740.
17. Chowdhury SA, Gertz EM, Wangsa D, Heselmeyer-Haddad K, Ried T, Schäffer AA, et al. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. Bioinformatics. 2015;31(12):i258–67.
18. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLOS Comput Biol. 2014;10(4):e1003535.
19. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 2020;21:208.
20. Scheinin I, Sie D, Bengtsson H, Van De Wiel MA, Olshen AB, Van Thuijl HF, 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 Res. 2014;24:2022–32.
21. Piskorz AM, Ennis D, Macintyre G, Goranova TE, Eldridge M, Segui-Gracia N, et al. Methanol-based fixation is superior to buffered formalin for next-generation sequencing of DNA from clinical cancer samples. Ann Oncol. 2016;27(3):532–9.
22. Macintyre G, Goranova TE, De Silva D, Ennis D, Piskorz AM, Eldridge M, et al. Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet. 2018;50(9):1262–70.
23. Baker AM, Cross W, Curtius K, Al Bakir I, Choi CHR, Davis HL, et al. Evolutionary history of human colitis-associated colorectal cancer. Gut. 2019;68(6):985–95.
24. Abbou SD, Shulman DS, DuBois SG, Crompton BD. Assessment of circulating tumor DNA in pediatric solid tumors: the promise of liquid biopsies. Pediatr Blood Cancer. 2019;66(5):e27595.
25. Boons G, Vandamme T, Mariën L, Lybaert W, Roeyen G, Rondou T, 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 Res. 2022;28(2):338–49.
26. Karlsson K, Przybilla M, Xu H, Kotler E, Karagyozova K, Sockell A, et al. Experimental evolution in TP53 deficient gastric organoids recapitulates tumorigenesis. bioRxiv. 2022. https://doi.org/10.1101/2022.04.09.487529.
27. Lu Z, Nie B, Zhai W, Hu Z. Delineating the longitudinal tumor evolution using organoid models. J Genet Genomics. 2021;48(7):560–70.
28. Liu APY, Smith KS, Kumar R, Paul L, Bihannic L, Lin T, et al. Serial assessment of measurable residual disease in medulloblastoma liquid biopsies. Cancer Cell. 2021;39(11):1519-30.e4.

29. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell. 2020;181(2):236–49.
30. Sauer CM, Eldridge MD, Vias M, Hall JA, Boyle S, Macintyre G, et al. Absolute copy number fitting from shallow whole genome sequencing data. bioRxiv. 2021. https://doi.org/10.1101/2021.07.19.452658.
31. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.
32. Watkins TBK, Lim EL, Petkovic M, Elizalde S, Birkbak NJ, Wilson GA, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020;587:126–32.
33. Zeira R, Shamir R. Genome rearrangement problems with single and multiple gene copies: a review. Bioinforma Phylogenet. 2019;29:205–241.
34. Letouzé E, Allory Y, Bollet MA, Radvanyi F, Guyon F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. Genome Biol. 2010;11(7):R76.
35. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48(10):1119–30.
36. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4.
37. Minussi DC, Nicholson MD, Ye H, Davis A, Wang K, Baker T, et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. Nature. 2021;592(7853):302–8.
38. Zeira R, Raphael BJ. Copy number evolution with weighted aberrations in cancer. Bioinformatics. 2020;36:i344–52.
39. Zeira R, Mon G, Raphael BJ. Genome halving and aliquoting under the copy number distance. In: Carbone A, El-Kebir M, editors. 21st International Workshop on Algorithms in Bioinformatics (WABI 2021). Leibniz International Proceedings in Informatics (LIPIcs), vol. 201. Dagstuhl: Schloss Dagstuhl – Leibniz-Zentrum für Informatik; 2021. p. 18:1–18:25.
40. Yang Z. Molecular evolution: a statistical approach. Oxford: Oxford University Press; 2014.
41. Hui S, Nielsen R. SCONCE: a method for profiling copy number alterations in cancer evolution using single-cell whole genome sequencing. Bioinformatics. 2022;38(7):1801–8.
42. Elizalde S, Laughney AM, Bakhoum SF. A Markov chain for numerical chromosomal instability in clonally expanding populations. PLOS Comput Biol. 2018;14(9):e1006447.
43. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution. 2018;4(1). https://doi.org/10.1093/ve/vey016
44. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLOS Comput Biol. 2019;15(4):e1006650.
45. Smolander J, Khan S, Singaravelu K, Kauko L, Lund RJ, Laiho A, et al. Evaluation of tools for identifying large copy number variations from ultra-low-coverage whole-genome sequencing data. BMC Genomics. 2021;22:357.
46. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17:368–76.
47. Nocedal J, Wright S. Numerical optimization. Springer Series in Operations Research and Financial Engineering. New York: Springer; 2006.
48. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2014;32(1):268–74.
49. Curtius K, Wong CJ, Hazelton WD, Kaz AM, Chak A, Willis JE, et al. A molecular clock infers heterogeneous tissue age among patients with Barrett's esophagus. PLOS Comput Biol. 2016;12(5):e1004919.
50. Yang Z, Kumar S, Nei M. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics. 1995;141:1641–50.
51. Pupko T, Pe I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 2000;17(6):890–6.
52. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–8.
53. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53(1–2):131–47.
54. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994;11(3):459–68.
55. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.
56. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell. 2021;184(8):2239-54.e39.
57. Rieux A, Balloux F. Inferences from tip-calibrated phylogenies: a review and a practical guide. Mol Ecol. 2016;25(9):1911–24.
58. Nowinski S. WGD classifier. 2022. https://github.com/BCI-EvoCa/CNA_stability/blob/master/WGD_classifier.html. Accessed 25 Feb 2022.
59. Krijgsman O, Carvalho B, Meijer GA, Steenbergen RDM, Ylstra B. Focal chromosomal copy number aberrations in cancer-Needles in a genome haystack. Biochim Biophys Acta (BBA) - Mol Cell Res. 2014;1843(11):2698–704.
60. Magee AF, Hilton SK, DeWitt WS. Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. Mol Biol Evol. 2021;38(10):4603–15.
61. Yang Z. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol. 1996;11(9):367–72.
62. Alves JM, Prieto T, Posada D. Multiregional tumor trees are not phylogenies. Trends Cancer. 2017;3(8):546–50.
63. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2018;47(D1):D941–7.
64. Gotovac JR, Kader T, Milne JV, Fujihara KM, Lara-Gonzalez LE, Gorringe KL, et al. Loss of SMAD4 is sufficient to promote tumorigenesis in a model of dysplastic Barrett's esophagus. Cell Mol Gastroenterol Hepatol. 2021;12(2):689–713.
65. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.

66. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat Biotechnol. 2021;39(2):207–14.
67. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
68. Leaché AD, Banbury BL, Felsenstein J, nieto-Montes de Oca A, Stamatakis A. Short tree, long tree, right tree, wrong tree: new acquisition bias vorrections for inferring SNP phylogenies. Syst Biol. 2015;64(6):1032–47.
69. Kim J, Sanderson MJ. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. Syst Biol. 2008;57(5):665–74.
70. dos Reis M, Donoghue PCJ, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. Nat Rev Genet. 2016;17(2):71–80.
71. Van De Wiel MA, Kim KI, Vosse SJ, Van Wieringen WN, Wilting SM, Ylstra B. CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics. 2007;23(7):892–4.
72. Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 2003;45(1):3–49.
73. Lu B, Curtius K, Graham TA, Yang Z, Barnes CP. CNETML: maximum likelihood inference of phylogeny from copy number profiles of multiple samples. Github. 2023. https://github.com/ucl-cssb/cneta. Accessed 16 May 2023.
74. Lu B, Curtius K, Graham TA, Yang Z, Barnes CP. CNETML: maximum likelihood inference of phylogeny from copy number profiles of multiple samples. Zenodo. 2023. https://doi.org/10.5281/zenodo.7941806.
75. Lu B, Curtius K, Graham TA, Yang Z, Barnes CP. CNETML: maximum likelihood inference of phylogeny from copy number profiles of multiple samples. Datasets Zenodo. 2023. https://doi.org/10.5281/zenodo.7940187.
76. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Github. 2020. https://github.com/raphael-group/chisel-data. Accessed 10 Sep 2022.
77. Kaufmann TL, Petkovic M, Watkins TBK, Colliver EC, Laskina S, Thapa N, et al. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. Zenodo. 2022. https://doi.org/10.5281/zenodo.7300106.

Publisher's Note