Whole-genome doubling-aware copy number phylogenies for cancer evolution with MEDICC2

Marina Petkovic[1,2,*], Thomas BK Watkins[3,4,*], Emma C Colliver[3,4], Sofya Laskina[5], Charles Swanton[3,4,6], Kerstin Haase[7,8], Roland F Schwarz[1,9,#]

1 Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Robert-Rössle-Str. 10, 13125 Berlin, Germany.
2 Department of Biology, Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany.
3 Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK.
4 Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK.
5 Department of Mathematics and Computer Science, Free University of Berlin, Germany.
6 Department of Medical Oncology, University College London Hospitals, London, UK.
7 Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Department of Pediatrics, Division of Oncology and Hematology, Augustenburger Platz 1, 13353 Berlin, Germany.
8 German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
9 BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
* both authors contributed equally
# contact: roland.schwarz@mdc-berlin.de

Abstract

Chromosomal instability (CIN) and somatic copy number alterations (SCNA) play a key role in the evolutionary process that shapes cancer genomes. SCNAs comprise many classes of clinically relevant events, such as localised amplifications, gains, losses, loss-of-heterozygosity (LOH) events, and recently discovered parallel evolutionary events revealed by multi-sample phasing. These events frequently appear jointly with whole genome doubling (WGD), a transformative event in tumour evolution, which generates tetraploid or near-tetraploid cells. WGD events are often clonal, occuring before the emergence of the most recent common ancestor, and have been associated with increased CIN, poor patient outcome and are currently being investigated as potential therapeutic targets.

While SCNAs can provide a rich source of phylogenetic information, so far no method exists for phylogenetic inference from SCNAs that includes WGD events. Here we present MEDICC2, a new phylogenetic algorithm for allele-specific SCNA data based on a minimum-evolution criterion that explicitly models clonal and subclonal WGD events and that

takes parallel evolutionary events into account. MEDICC2 can identify WGD events and quantify SCNA burden in single-sample studies and infer phylogenetic trees and ancestral genomes in multi-sample scenarios. In this scenario, it accurately locates clonal and subclonal WGD events as well as parallel evolutionary events in the evolutionary history of the tumour, timing SCNAs relative to each other.

We use MEDICC2 to detect WGD events in 2778 tumours with 98.8% accuracy and show its ability to correctly place subclonal WGD events in simulated and real-world multi-sample tumours, while accurately inferring its phylogeny and parallel SCNA events. MEDICC2 is implemented in Python 3 and freely available under GPLv3 at https://bitbucket.org/schwarzlab/medicc2.

Introduction

Aneuploidy and chromosomal instability (CIN) are hallmarks of many tumours and contribute to genome plasticity and intra-tumour heterogeneity (ITH) [1,2]. CIN enables the tumour to generate somatic copy number alterations (SCNAs) which we have demonstrated to contribute to continuous evolution and selection across human tumour types [3]. Besides its immediate clinical relevance in carcinogenesis and resistance development [4], SCNAs also form a rich source of genetic variation that can be exploited for reconstructing tumour evolution in the patient in retrospective studies [5,6].

We originally developed MEDICC (Minimum-Event Distance for Intra-tumour Copy-number Comparisons) [5] for reconstructing cancer evolution from SCNAs and for phylogenetic quantification of ITH in a multi-sample sequencing scenario [5] and used it to link SCNA heterogeneity to clinical outcome in high-grade serous ovarian cancer [6]. MEDICC solved the problem of phylogenetic inference from SCNA data with horizontal dependencies between genomic loci and enabled tree inference and reconstruction of ancestral SCNA states in the history of a tumour. MEDICC has since been widely adopted in many studies, including some of the largest multi-sample cancer cohorts available [3,7,8] and its underlying principle has been theoretically studied [9] and extended to total copy numbers in single cells [10].

Recently, whole-genome doubling (WGD) events [11–14], which have long been known to be linked to tumorigenesis [15–20], have received renewed interest as key contributors to CIN [3,21,22] and as potentially therapeutically targetable characteristics of cancer cells

[23–25]. WGD events describe tetraploidization of genomes, frequently followed by immediate loss of individual chromosomal copies, leading to an average tumour ploidy of between 3 and 4 copies [13,21]. WGD events have been shown to buffer cancer genomes against the accumulation of deleterious mutations and homozygous deletions [22] and have been identified as an important substrate for further genomic diversification [3,22]. Indeed, whole genome doubling events have also been shown to facilitate the propagation of aneuploid genomes in cell models [21]. It is therefore becoming increasingly clear that understanding cancer evolution requires detection of the presence of WGD events in both single-sample and multi-sample studies.

Despite a clear correlation with the ploidy of a tumour, the detection of WGD events remains challenging. General indicators are a major copy number > 2 in allele-specific SCNA profiles [13], and current best estimates rely on a combination of average tumour ploidy and percentage of the genome affected by loss-of-heterozygosity (LOH) events [26] or on the clone structure in a multi-sample scenario [19]. Despite these efforts, it is frequently impossible to distinguish between WGD events and successive gain/loss events of individual chromosomes, and so far no evolutionary model exists that directly identifies the presence of clonal and subclonal WGD events in the history of a tumour [13]. MEDICC, while accurately modelling gains and losses of arbitrary length, is not able to take WGD events into account and instead treats individual chromosomes independently. In addition, MEDICC is restricted in the length of the copy number profiles (CNP) and in the number of copies per segment it can process.

To overcome these issues we present MEDICC2, a new algorithm for phylogenetic reconstruction from allele-specific SCNA profiles that explicitly models clonal and subclonal WGD events and takes parallel evolutionary events into account [3]. MEDICC2 solves the Minimum-Event Distance (MED) [5] or copy number transformation problem (CNTP) [20] in the presence of WGD and determines the minimum number of gains, losses and whole-genome doublings necessary to transform one SCNA profile into another as an evolutionary metric with biological constraints. MEDICC2 infers ancestral SCNA states, times SCNA events including WGD relative to each other and thereby successfully reveals clonal and subclonal WGD events. To achieve this, MEDICC2 employs a novel evolutionary finite-state transducer (FST) with an increased alphabet, enabling the computation of the MED across several thousand segments genome-wide. MEDICC2 further provides improved heuristics for more accurate ancestral reconstructions and implements a novel evolutionary phasing strategy with improved computational complexity.

We apply MEDICC2 to 2778 whole-genome allele-specific CNPs from the Pan-Cancer Analysis of Whole Genomes (PCAWG), where it accurately identifies WGD events against a "gold standard" [27]. Using a multi-sample prostate cancer case from the same cohort we further demonstrate MEDICC2's ability to detect subclonal whole-genome doubling in the most recent common ancestor (MRCA) of the metastatic samples. There, MEDICC2 also detects and correctly places mirrored subclonal allelic imbalance (MSAI) events, SCNAs affecting different parental alleles at the same genomic locus in separate subclones, and parallel evolutionary events revealed by multi-sample phasing [3,28].

Methods

Method overview

MEDICC2 infers phylogenetic trees from sets of allele-specific copy number profiles (CNP) as follows (Figure 1A). CNPs are either pre-phased using existing phasing methods, such as multi-sample reference phasing (MRP) implemented in Refphase [3], or subjected to evolutionary phasing which chooses a haplotype configuration that minimises the sum of the evolutionary distances between CNPs of both haplotypes and a reference haplotype, typically a diploid normal sample [5]. MEDICC2 then computes pairwise MEDs between all genomes. Next, the pairwise distance matrix is processed with the neighbour joining algorithm [29] to infer the topology of the tree. Pairwise distances are also used for summary statistics and statistical tests as previously described [5]. After inferring the tree topology, ancestral CNPs at the internal nodes of the tree are inferred using an upward and downward pass on the tree, analogous to Felsenstein's Pruning algorithm for maximum-likelihood phylogenetics [30]. After fixing the ancestral CNPs, the final branch lengths are determined in a final pass of the tree by computing MEDs between parent and child clades along each branch of the tree. The result is reported to the user as a patient summary and plot which includes the tree and inferred ancestral and terminal CNPs (Figure 1A).

The minimum-event distance with whole-genome doubling

We here define the minimum-event distance with whole-genome doublings (MED-WGD) problem for CNPs. The original MED problem was formulated by us in 2014 [5] and recently studied by Zeira et al. under the term *copy number transformation problem* [9]. Briefly, CNPs are represented as vectors of positive integer copy numbers $(k)_{1..n}$, with $k \geq 0$, where each integer copy number represents a genomic segment i. The size of the genomic segment

does not affect the MED. The MED between a pair of CNPs is then defined as the minimum number of gains (+1 operations) and losses (-1 operations) of arbitrary length needed to transform the first CNP into the second. To ensure biological constraints, gains of segments with copy number zero are not permitted and zero segments are "ignored" by subsequent operations, mimicking the absence of that segment of genomic DNA (Figure 1B). In MEDDIC2 we model whole-haplotype CNPs by concatenating individual chromosomes and marking chromosomal boundaries with chromosome separation symbol 'X'. Standard gain and loss events must terminate when they reach 'X'.

We model WGD events per haplotype: such events begin on the first non-zero segment of a haplotype and end at the last non-zero segment, applying a +1 operation to all non-zero segments in between. WGD events do not terminate at 'X' and leave these elements unchanged. These events correspond to the typically observed tetraploidization followed by rapid chromosomal losses to create a near-triploid intermediate [13,21], a phenomenon directly observed, for example, in the genome evolution in many plant species [31]. To reach the less likely state of a fully tetraploid tumour from a diploid ancestor without subsequent losses, this +1 operation is applied separately to both haplotypes of the tumour.

It has been shown previously, that the standard MED can be solved by considering losses separately before any gains [9]. Indeed, only loss-of-heterozygosity (LOH) events, i.e. losses which reduce haplotype-specific copy numbers to zero, must be considered first, as subsequent gain and loss events must ignore the positions with copy number zero. The MED however is oblivious to the ordering of any subsequent gains and losses. When including WGD events (MED-WGD), LOH events must again be dealt with before any other event. In addition, WGD events must come before any segmental losses and gains, for example to allow for the deletion of segments previously gained during a WGD event (Figure 1B). The inclusion of WGD events further introduces non-determinism into the problem as locally WGD events cannot be distinguished from segmental gains before taking the full sequence into consideration.

Due to the biological constraint of preventing gains of fully deleted genomic regions, both MED and MED-WGD are asymmetric distances (asymmetric MED) and the final symmetric distance (symmetric MED) is computed from a CNP to an arbitrary ancestral CNP and from the ancestral CNP back to the second CNP, thereby minimising over all possible ancestral CNPs [5] (Figure 1C). We further assume that both parental haplotypes evolve independently. Genomes are thus represented as pairs of CNPs, one for each parental

haplotype, and the total pairwise MED matrix D_T between a set genomes is the sum of the pairwise MED matrices for both parental haplotypes A and B: $D_T = D_A + D_B$.

Computational framework

To solve the MED problem we employ a finite-state transducer (FST) framework as previously described [5], following the notation of Mohri [32]. We model allele-specific copy number profiles as unweighted finite-state acceptors (FSA) $A = (\Sigma, Q, E, i, F)$ and evolutionary events as weighted FSTs $T = (\Sigma, Q, E, i, F, \lambda, \rho)$ with (input and output) copy number alphabet $\Sigma = \{0, .., 8\}$ (per allele), a finite set of states Q, a finite set of transitions E, an initial state $i \in Q$, a set of final states $F \subseteq Q$, an initial weight λ and a final weight ρ. Transitions between states are equipped with an input symbol $l_i \in \Sigma$ (input copy number) and an output symbol $l_o \in \Sigma$ (output copy number) and a weight w. All weights λ, ρ, w are taken from the positive integers including zero and calculations are carried out over the tropical semiring, i.e. weights are summed along the path of a FST and the final weight between a pair of sequences is the minimum over all possible paths

$$T[x,y] = \min_{\pi \in P} \sum_i w[\pi]_i \qquad \text{(Eq. 1)}$$

(see [33]), where P is the set of all possible paths transforming x to y.

FSTs and FSAs can be subjected to a variety of operations, of which *composition* ("$\circ$") is of particular importance. During composition a new FST is constructed in which the set of states is the cartesian product of the set of states of the two input FSTs. The composition S of two FSTs T_1 and T_2 then assigns a weight to any pair of input and output sequences by chaining their transduction

$$S[x,y] = (T_1 \circ T_2)[x,y] = \min_z (T_1[x,z] + T_2[z,y]) \qquad \text{(Eq. 2)}$$

via intermediary sequence z [32]. Composition is also used to effectively compute the score or total weight $T[x,y]$ that a FST T assigns to a pair of sequences x and y (Eq. 1) by representing x and y as two unweighted acceptors and running a single-source shortest distance algorithm (SD) over the composition $x \circ T \circ y$ [32].

$$T[x,y] = \min_{\pi} \sum_i w[\pi]_i = SD(x \circ T \circ y) \ . \qquad \text{(Eq. 3)}$$

Composition enables us to combine multiple evolutionary event FSTs into a final FST in which the individual events are carried out successively in order of composition, and to transform the asymmetric MED into a symmetric MED for calculation of the pairwise distance matrix [5].

Solving the MED-WGD over an extended alphabet

We first extended the alphabet to a maximum copy number of 8 per allele (total maximum copy number 16) to make use of the resolution of whole-exome and whole-genome sequencing data. We then defined four one-step FSTs which model one of four different evolutionary events considered: (i) LOH events ($T^1{}_{LOH}$), (ii) segmental (+1) gains ($T^1{}_G$), (iii) segmental (-1) losses ($T^1{}_L$) without LOH and (iv) WGD (+1 for all non-zero segments) events ($T^1{}_{WGD}$). In the one-step FSTs, each sequence position can only be affected by a single event. For example, the one-step FST for segmental gains $T^1{}_G$ only allows copy number changes of arbitrary length from 1 to 2, 2 to 3, and so on, but not, for example, from 1 to 3. To span the full range of possible events, the one-step FSTs are each composed n times with themselves and the maximum copy number dictates the number of compositions necessary: $n = |\Sigma| - 1$ for LOH events and $n = |\Sigma| - 2$ for segmental losses and gains and WGD events [5]. The resulting event FSTs T_{LOH}, T_G, T_L, and T_{WGD} are then chained (composed) into the asymmetric MED-WGD FST)

$$T = T_{LOH} \circ T_{WGD} \circ T_L \circ T_G \qquad \text{(Eq. 4)}$$

The final MED-WGD between CNPs x and y is then computed following Eq. 3:

$$T[x, y] = SD(x \circ T \circ y) = SD(x \circ T_{LOH} \circ T_{WGD} \circ T_L \circ T_G \circ y) \qquad \text{(Eq. 5)}$$

Analogously, the simple MED is built via composition as in

$$T = T_{LOH} \circ T_L \circ T_G \qquad \text{(Eq. 6)}$$

and distance calculation is carried out as in Eq. 5.

As noted previously the MED and MED-WGD are asymmetric. To compute symmetric distances $S[x,y]$ between pairs of CNPs connected in a phylogenetic tree we compute the score between x and y via its common ancestor using the kernel composition of T with its inverse T^{-1} [34,35] (Figure 1C):

$$S[x, y] = SD(x \circ T^{-1} \circ T \circ y) \qquad \text{(Eq. 7)}$$

As the number of states in a composed FST is the product of the states of the input FSTs, explicit computation of the composition in Eq. 7 is computationally expensive. We therefore employ a new computation strategy based on lazy (on-demand) composition followed by shortest path computation using a shortest-first queue [36]. Lazy composition prevents full expansion of the composed FST before determining the shortest path and instead expands the FST only along the path visited [36]. This substantially decreases compute time over previous implementations. We here provide two versions for reference and comparison: a "lazy composition strategy" where $K = T^{-1} \circ T$ is computed explicitly followed by lazy

composition of the input acceptor x with K and the output acceptor y, and a "lazy kernel composition" strategy, where all composition steps are carried out on-demand.

Fast evolutionary phasing in linear time

Traditionally, allele-specific SCNAs are reported in major and minor copy number, as the relative phasing of copy number segments to each other is unknown. We introduced the MRP implementation *refphase* [3,37] to leverage relative phasing information in a multi-sample sequencing scenario and used it to identify MSAI events across human cancers [3,28,38]. In situations where MRP is not feasible, e.g. in single-sample scenarios, we developed evolutionary phasing [5], where the assignment of major and minor copy numbers to parental haplotypes is chosen to minimize the sum of MEDs over both parental haplotypes (minimum evolution criterion). In its original form, evolutionary phasing was achieved through the use of a weighted context-free grammar in concert with our original MED [5], a computationally costly solution. To enable phasing for a large number of segments and genomes, we here provide a novel phasing strategy which solves the evolutionary phasing problem exactly, but at a fraction of the original runtime, by staying within the realm of regular grammars and FSAs.

To do so we first encode CNPs for both alleles jointly as an unweighted phasing FST P as follows: the FST follows a linear structure with a number of states equal to the number of segments $+1$. Two transitions occur between each neighboring pair of states and the two transitions have as input symbols major copy numbers and as output symbols minor copy numbers and vice versa (Figure 1D). Due to these mirrored input and output symbols every valid path through the phasing FST P thus determines an assignment of copy number alleles to haplotype 1 and haplotype 2. The set of of all 2^n possible paths for a sequence of length n through this FST corresponds to the set of possible phasing choices. To choose the most parsimonious haplotype assignment, this phasing FST is then composed from the left and from the right side with a composed FSA $u = (d \circ T) \downarrow$ of the diploid FSA d (encoding all-1s) with the MED-WGD FST T, projected to its output ($\downarrow$). Shortest-path computation over this composite yields the optimal phase with a total score equal to the sum of MEDs over both parental haplotypes. Separate haplotypes h_a and h_b can be extracted by projection to input and output followed by weight removal:

$$h_a = \uparrow sp(u \circ P \circ u) \qquad \text{(Eq. 8)}$$

$$h_b = sp(u \circ P \circ u) \downarrow$$

Simulating somatic copy number alteration evolution

Random instances of copy number evolution were simulated starting from a haplotype of all 1s (normal genome) as follows: We first generated a random number of evolutionary events using a poisson process with rate parameter $\mu = 10$ (reconstruction accuracy test) and $\mu = 20$ (speed test). In the reconstruction accuracy test each event had a 5% probability to be a WGD event, and a 47.5% probability of being a gain or loss respectively. In the speed test, to prevent too many deletions, the gain probability was set to 80%. The start of gain or loss events in the sequence was selected uniformly at random from the set of remaining available positions (positions with copy number $\neq 0$) and event lengths were drawn from a geometric distribution with success probability parameter $p = 0.2$. Events were applied to the sequence obeying biological constraints, i.e. no gain of segments with copy number zero and forced ending of events at chromosome boundaries, the latter with the exception of WGD. For the reconstruction accuracy test, sequences were fixed at length 50 (five chromosomes of length 10 each). For the speed test, sequence lengths were varied from 20 to 200 segments (Figure 1E).

Implementation and Availability

MEDICC2 is implemented in Python 3 and freely available under GPLv3 at https://bitbucket.org/schwarzlab/medicc2. It employs OpenFST and its Python wrapper pywrapfst [36] for manipulation of finite-state machines. Core algorithms are implemented in C++ as a Cython extension by linking to the OpenFST library.

Results

MEDICC2 infers exact minimum event distance with whole-genome doubling in linear time

We first verified that MEDICC2 computes accurate MEDs by simulating 1000 random instances of copy number evolution starting from a haplotype of all 1s (normal genome), including WGD events with a probability of 5% per event (Methods). We then inferred the MED with and without taking WGD events into consideration. Both non-WGD and WGD MEDs correctly inferred a number of events at no greater than the true distance (minimum event criterion, Figure 2A). Only in cases where CNPs were simulated with WGD, but distances inferred with the non-WGD MED, did MEDICC2 overestimate the true distances as

expected (orange dots, Figure 2B). In many cancer studies, Euclidean distance is still frequently used as an evolutionary metric for CNPs. We have previously demonstrated that the Euclidean metric is inferior to the MED in terms of tree reconstruction accuracy even without taking WGD events into consideration [5]. We here directly compared the true simulated distances to Euclidean distances and found Euclidean distance to perform poorly for genomes both with and without WGD with only 15% of variation in the true distance explained by the Euclidean measurement and vice versa ($R^2 = 0.15$, Figure 2C).

We next investigated the performance of the different composition strategies for the FSTs with and without WGD events to assess their computational complexity. We therefore simulated sequences (Methods) of increasing length from 20 to 200 segments with 10 replicates each and timed MED inference. In all instances computation time of the MEDs was linear with respect to the length of the input sequences. While MED-WGD took significantly longer to compute than the MED without WGD the new lazy and lazy kernel composition strategies (Methods) reduced runtime by a factor of between 5 and 10, keeping runtime manageable even for genomes with several thousand segments (Figure 1E).

MEDICC2 accurately identifies whole-genome doubling in 2778 cancers

We next evaluated the power of MEDICC2 to identify WGD events in a real-world scenario. To this end we obtained 2778 whole-genome allele-specific CNPs from the Pan-cancer Analysis of Whole Genomes (PCAWG) project [27] together with annotations about their WGD status, which serve us as a “gold standard”. As these CNPs are unphased and available in major/minor configuration only, we employed MEDICC2’s new phasing algorithm (Methods) to perform evolutionary phasing on each PCAWG CNP and a diploid normal sample. Recent efforts have demonstrated that average tumour ploidy alone is insufficient to accurately determine the occurrence of WGD events [11,26]. Dentro et al. instead fitted a linear model to a combination of tumour ploidy and the percentage of the genome affected by LOH to determine presence of a WGD in the evolutionary history of the cancer (Supplementary Figure 1A). We hypothesized that identification of WGD events in a cohort-independent manner should be straightforward with MEDICC2 by comparing MED and MED-WGD between each PCAWG tumour and a diploid normal sample. We therefore computed the WGD evidence score s_i as

$s_i = MED(d, t_i) - MED_{WGD}(d, t_i)$, where t_i represents a PCAWG tumour profile and d represents a standard diploid normal sample. Because $MED(x, y) > MED_{WGD}(x, y)$ for any

valid set of CNPs x and y (Methods), the score s_i is always positive ($s_i \geq 0$). For an overview of the calculated distances MED_{WGD} and the WGD evidence scores s_i stratified by cancer type, see Supplementary Figure 1B-C.

We argued that, if the difference between MED and MED-WGD is two or greater (one event for each haplotype), a WGD event is likely to have occurred in the tumour's past.
We found the MEDICC2 WGD score s_i to clearly separate tumours annotated as WGD from those annotated as non-WGD in PCAWG (Figure 2D) [26]. It follows the trajectory identified by Dentro et al. and provides direct evidence for occurrence of WGD events without the need for an additional model fitting step. Our WGD indicator $s_i \geq 2$ agreed with the PCAWG WGD annotation in 2747 out of 2778 cases (98.9%). In 8 cases MEDICC did not call a WGD where the genome was annotated as WGD in PCAWG (false negative) and in 23 cases MEDICC did call a WGD event where the genome was annotated as not having undergone WGD in PCAWG (false positive) (Table 1). An optimal classification rate was achieved by increasing the cutoff to $s_i \geq 4$, leading to 2755 out of 2778 correct classifications (99.2%).

Table 1: Classifier performance of our WGD indicator ($s_i \geq 2$) against PCAWG annotation.

	MEDICC	
PCAWG	**no WGD**	**WGD**
no WGD	1937	23
WGD	8	810

Since PCAWG WGD annotations are also based on noisy data and may contain errors, we next investigated whether the remaining misclassifications were associated with tumours marked as 'WGD uncertain' by the PCAWG heterogeneity and evolution working group. Indeed, genomes where our WGD prediction did not agree with the PCAWG annotation were significantly over-represented among tumours with status 'WGD uncertain' ($p = 0.0002$, chi-square test). We conclude that MEDICC2 accurately infers the presence of WGD events even in single-sample studies, without the need for additional parameter estimation.
To quantify the relationship between our WGD score s_i and the separating line given by Dentro et al. further, we projected each PCAWG tumour orthogonally onto this line and measured its (signed) distance to determine a single PCAWG WGD score. Correlation analysis of this score with s_i showed a strong correlation between both ($r^2 = 0.87$, Figure 2E).

MEDICC2 correctly infers subclonal whole-genome doublings

We next tested MEDICC2's ability to infer subclonal WGD events. For this we reconstructed the evolutionary history of a simple simulated example of two chromosomes and 10 copy number segments (Figure 3). The example was chosen such that a subclonal haplotype doubling event occurred at the second internal node (internal 4) from the root of the tree, leaving one terminal node (sample 1) undoubled. To identify the branch at which an evolutionary event occurs, MEDICC2 computes change events, i.e. the difference in copy number states of each node in the tree relative to its immediate ancestor. It then marks up individual segments in color in the output plot as follows: no change (all samples diploid, dark grey), clonal change (all samples identical but different from diploid, light grey), subclonal gain (red) and subclonal loss (blue) (Figure 3).

MEDICC2 correctly inferred the position of the WGD event at internal node 4, followed by an additional gain event on chromosome 1 segment 5 on the second allele which was shared between sample 2 and sample 5, as well as an additional gain event on chromosome 2 segment 3 which was shared between sample 3 and sample 4. Sample 4 subsequently gained another copy of that locus. MEDICC2 additionally computes a summary of the change events happening in the evolutionary tree by summing events over all samples. This ascertains that each evolutionary event is only counted once and the resulting summary track (Figure 3 bottom) describes the total number of copies gained or lost by each genomic segment across the tree.

We then expanded our analysis to examine an illustrative multi-sample case of a prostate cancer patient "A31", first analysed in Gundem et al. 2015 [39], with five samples subjected to whole-genome sequencing (WGS). A31 consists of one sample ("C") from the primary tumour and four samples ("A", "D", "E", and "F") from distinct metastatic sites. Gundem et al. estimated cancer cell fractions of all single nucleotide variants (SNV) identified in the WGS across samples. They then used an n-dimensional Bayesian Dirichlet process to cluster those and to perform a phylogenetic reconstruction of A31's evolution. This phylogenetic reconstruction is reproduced in Figure 4A.

A31 was later included in and analysed as part of the PCAWG cohort [26] and found to demonstrate a subclonal WGD event with WGD affecting all metastatic samples (A, D, E, and F) but not the primary sample (C). Using multi-sample phased CNPs derived from Battenberg [40], MEDICC2 detected the presence of WGD in the same metastatic samples and its absence in the primary sample from A31 (Figure 4B). The MEDICC2 summary SCNA

distance required to reach the same output state in the absence of WGD was calculated to be 40 compared to 19 with WGD (Figure 4B). Even though SCNA inference from bulk sequencing typically only captures the dominant clone per sample, MEDICC2, without using any SNV data, produced an almost identical phylogeny when compared to the published SNV-based phylogenetic reconstruction, even when adapted to reflect the tumour sample with the greatest contribution to each subclonal cluster previously identified (Figure 4A,B, Methods).

MEDICC2 models parallel evolution

Recently, the application of the MRP algorithm that maintains consistent phased haplotypes across samples from a single patient's disease has revealed additional SCNA heterogeneity [3,28]. This additional heterogeneity results from the detection of MSAI as well as SCNA-mediated parallel evolution where the same SCNA event (e.g. an LOH event) occurs independently affecting distinct haplotypes within an individual patient's disease [3,28,38,41].

Multi-sample phasing analysis of the samples from A31 identified multiple MSAI events as well as parallel LOH on chromosomes 6 and 13 (Figure 4C). MEDICC2 allows the independent origins of these parallel events to be formally assigned to the branch corresponding to the emergence of the dominant clone in the primary sample and to the branch corresponding to emergence of the common ancestor of all the metastatic samples (Figure 4C). MEDICC2's ability to correctly identify and locate these parallel evolutionary events revealed by multi-sample phasing provides additional evidence for a diverging evolutionary trajectory between primary and metastatic samples, absent from its original analysis [39] (Figure 4B).

We next compared relative branch lengths within the SNV-based tree and the SCNA-based MEDICC2 tree. Branch lengths of both trees correspond to the number of SNVs present in each subclonal cluster and to the number of CNA events larger than 1 Mb, respectively. To facilitate comparison between branch lengths in the two trees we computed relative event distances by normalising branch lengths by the maximal root to leaf distance for each respective tree. The relative event distances from the root of each tree to the terminal nodes representing samples A, F and D were conserved in the two phylogenetic reconstructions. However, the SNV-based and SCNA-based trees demonstrated distinctly different relative branch lengths from the diploid normal root leading to the MRCA. This "trunk" was found to be shorter in the SCNA-based MEDICC2 tree (11/54 SCNAs) when compared to the SNV-based tree (2056/2682 SNVs). This suggests that there have been relatively few

founder SCNAs compared to a large number of founder SNVs. In addition, the branch terminating at the dominant clone of metastatic sample E was relatively long in the SCNA-based tree compared to in the SNV tree (24/54 vs 52/2682 (Figure 4C). This relatively large minimum event distance is due to multiple SCNA events on chromosome 12 present only in sample E suggesting a complex set of SCNAs potentially occurring together.

Here we show that construction of an SCNA-based MEDICC2 phylogeny recapitulates the topology from an SNV-based phylogenetic method, identifies and correctly places subclonal WGD events, and informs our understanding of the evolutionary history of a tumour by assigning SCNAs including MSAI events and instances of parallel evolution to the branches of the phylogeny. We also show that the construction of SCNA phylogenies can be used to make informative comparisons between SNV-based and SCNA-based tumour evolution.

Discussion

MEDICC2 retains the key advantages of the original implementation [5] at a fraction of its runtime. It processes output from any allele-specific copy number estimation algorithm analysing any sequencing modality or SNP array. It estimates evolutionary distances and resolves horizontal dependencies between genomic loci as well as multiple overlapping SCNA events. In contrast to approaches based solely on SCNA breakpoints, MEDICC2 models actual genomic events that change copy number, incorporating biological constraints such as LOH. In addition, to our knowledge, MEDICC2 is the first algorithm to be able to infer the presence of clonal and subclonal WGD events and incorporate them into copy number phylogenies. Modelling WGD events is vital to accurately quantify the number of individual SCNAs present in a tumour's evolution and to perform meaningful comparisons of the SCNA burden of non-WGD and WGD tumours.

Our recent work [3,28] and that of others [41] has highlighted the importance of using multi-sample phasing techniques to reveal additional SCNA heterogeneity taking the form of MSAI or parallel evolution of similar SCNAs from distinct haplotypes [3,28]. MEDICC2 accepts phased input and assigns parallel events, alongside all other SCNAs, to branches of copy number phylogenies consistent with the MED criterion. MEDICC2 can reveal additional parallel evolution on the same haplotype in cases where an independent origin of two events leads to a more parsimonious phylogeny than a shared ancestry.

WGD in cancer has previously been proposed to result from endoreduplication [42], mitotic dysfunction [43] and cytokinesis failure events [44], or even potentially cell fusion [45]. Nearly all of these proposed mechanisms suggest a diploid to tetraploid transition (tetraploidization) [46], frequently followed by return to a triploid state through subsequent chromosome losses [21]. MEDICC2 models this by considering both haplotypes in a diploid founder cell independently, thus allowing doubling of a haplotype which directly moves the cell from a diploid to a triploid state. This could be a limitation of the algorithm in that it may prefer to double only a single haplotype rather than both, as would occur during a physiological WGD event without an immediate return to a triploid state. However, this independent modelling of haplotypes seems to fit real-world CNPs extremely well as verified by MEDICC2's ability to identify WGD events on 2778 WGS tumours from PCAWG with very high accuracy. Additionally, misclassifications on that cohort were statistically strongly associated with uncertainty about the true WGD status of the tumours. More detailed investigations will be needed to assess whether MEDICC2's WGD detection might help to re-annotate these tumours with respect to their WGD status.

MEDICC2's ability to identify subclonal WGD events and model parallel evolution were of paramount importance to the analysis of prostate-cancer patient A31 [39] and to the findings of our previous study [3]. In patient A31 MEDICC2 not only recapitulated the WGS SNV-based phylogeny, but also identified a subclonal WGD event at the root of the metastatic samples. MEDICC2 further provided additional support for the divergence between the primary and metastatic samples by detecting parallel evolution of LOH events on chromosomes 6 and 13. Interestingly, other SCNA-mediated parallel evolution events in this tumour were identified at the *AR* gene locus in the original analysis of this tumour through the orthogonal method of exact structural variant breakpoint identification enabled by WGS [39].

Like many computational tools, MEDICC2's performance is limited by the quality of the input data. For example, CNPs are prone to over-segmentation, where local variations in sequencing coverage due to noise or poor parameter choice in the segmentation process can lead to erroneous copy number calls, which in turn can have large effects on the MED. Additionally, and in contrast to SNV data, assessing within-sample subclonality from CNPs from bulk sequencing is notoriously difficult and bulk CNPs thus typically only reflect the major or dominant clone in that sample, as evidenced by our analysis of patient A31. Since MEDICC2 is agnostic of the sequencing modality used to generate its input, future advances that increase CNP resolution will be usable by MEDICC2 without modification. Here, single-cell sequencing technologies will prove very useful and are beginning to enter the

mainstream [47] for both total copy number based [48] and allele-specific CNPs [49]. While tailored adaptations of the MED principle for single-cell data have recently been proposed in the form of MEDALT [10], it does not take allele-specific SCNAs or WGD events into account.

As an additional limitation the original version of MEDICC [5] had a limited alphabet to represent SCNAs, up to a maximum total copy number of 8. This has severely limited its use in some high ploidy tumours and restricted its ability to model high-level amplifications. While we here have doubled the permissible alphabet to a total copy number state of 16 (8 per haplotype), this is still a compromise between computational complexity and realism and might not be sufficient in some cases. Fortunately, extending the alphabet at the cost of runtime is straightforward and MEDICC2 is supplied with tools for users to generate their own MED FSTs with an increased alphabet if desired.

It is further worth noting that MEDICC2 treats SCNAs on the level of copy number changes and does not explicitly model the underlying physical mechanisms, such as whole chromosome missegregation [2], chromothripsis [50], chromoplexy [51], breakage fusion bridges [52], and other more recently defined events [53]. In the future, once the emerging copy number patterns of these events are better characterised, distinguishing between these events might be advantageous.

Despite these limitations, systematically determining the number and order of whole chromosome missegregations, arm-level SCNAs and focal events that have occurred in the evolutionary history of a tumour has not previously been performed on a large scale and has previously been the preserve of theoretical mathematical modelling [54,55]. MEDICC2 enables the reconstruction and timing of the individual SCNA events present in the evolutionary history of a tumour that may overlap and build upon one another. This will allow much more detailed dissection of WGD, aneuploidy and CIN across the genome utilising single sample, multi-sample and single cell approaches, than the measures of the proportion of the genome affected by SCNA that much of the field has previously relied on.

References

1. McGranahan N, Burrell RA, Endesfelder D, Novelli MR, Swanton C. Cancer chromosomal instability: therapeutic and diagnostic challenges. EMBO Rep. 2012;13: 528–538.

2. Sansregret L, Vanhaesebroeck B, Swanton C. Determinants and clinical implications of

chromosomal instability in cancer. Nat Rev Clin Oncol. 2018;15: 139–150.

3. 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. doi:10.1038/s41586-020-2698-6

4. Lee AJX, Endesfelder D, Rowan AJ, Walther A, Birkbak NJ, Futreal PA, et al. Chromosomal instability confers intrinsic multidrug resistance. Cancer Res. 2011;71: 1858–1870.

5. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014;10: e1003535.

6. Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, et al. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. PLoS Med. 2015;12: e1001789.

7. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat Genet. 2015;47: 209–216.

8. Brown D, Smeets D, Székely B, Larsimont D, Szász AM, Adnet P-Y, et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. Nat Commun. 2017;8: 14944.

9. Zeira R, Zehavi M, Shamir R. A Linear-Time Algorithm for the Copy Number Transformation Problem. J Comput Biol. 2017;24: 1179–1194.

10. Wang F, Wang Q, Mohanty V, Liang S, Dou J, Han J, et al. MEDALT: single-cell copy number lineage tracing enabling gene discovery. Genome Biol. 2021;22: 70.

11. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30: 413–421.

12. 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: 1134–1140.

13. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018;50: 1189–1195.

14. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018;33: 676–689.e3.

15. Galipeau PC, Cowan DS, Sanchez CA, Barrett MT, Emond MJ, Levine DS, et al. 17p (p53) allelic losses, 4N (G2/tetraploid) populations, and progression to aneuploidy in Barrett's esophagus. Proc Natl Acad Sci U S A. 1996;93: 7081–7084.

16. Imkie M, Davis MK, Persons DL, Cunningham MT. Biphasic acute myeloid leukemia with near-tetraploidy and immunophenotypic transformation. Arch Pathol Lab Med. 2004;128: 448–451.

17. Watanabe A, Inokuchi K, Yamaguchi H, Mizuki T, Tanosaki S, Shimada T, et al. Near-triploidy and near-tetraploidy in hematological malignancies and mutation of the p53 gene. Clin Lab Haematol. 2004;26: 25–30.

18. Fujiwara T, Bandi M, Nitta M, Ivanova EV, Bronson RT, Pellman D. Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. Nature. 2005;437: 1043–1047.

19. Olaharski AJ, Sotelo R, Solorza-Luna G, Gonsebatt ME, Guzman P, Mohar A, et al. Tetraploidy and chromosomal instability are early events during cervical carcinogenesis. Carcinogenesis. 2006;27: 337–343.

20. Ganem NJ, Godinho SA, Pellman D. A mechanism linking extra centrosomes to chromosomal instability. Nature. 2009;460: 278–282.

21. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Grönroos E, Endesfelder D, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 2014;4: 175–185.

22. López S, Lim EL, Horswell S, Haase K, Huebner A, Dietzen M, et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat Genet. 2020;52: 283–293.

23. Marquis C, Fonseca CL, Queen KA, Wood L, Vandal SE, Malaby HLH, et al. Chromosomally unstable tumor cells specifically require KIF18A for proliferation. bioRxiv. 2020. p. 2020.06.18.159327. doi:10.1101/2020.06.18.159327

24. Quinton RJ, DiDomizio A, Vittoria MA, Kotýnková K, Ticas CJ, Patel S, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. Nature. 2021;590: 492–497.

25. Cohen-Sharir Y, McFarland JM, Abdusamad M, Marquis C, Bernhard SV, Kazachkova M, et al. Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. Nature. 2021;590: 486–491.

26. 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. bioRxiv. 2020. p. 312041. doi:10.1101/312041

27. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.

28. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non–Small-Cell Lung Cancer. N Engl J Med. 2017;376: 2109–2121.

29. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4: 406–425.

30. Felsenstein J. Inferring phylogenies. Sinauer associates Sunderland, MA; 2004.

31. Mandáková T, Gloss AD, Whiteman NK, Lysak MA. How diploidization turned a tetraploid into a pseudotriploid. Am J Bot. 2016;103: 1187–1196.

32. Mohri M. Weighted Automata Algorithms. In: Droste M, Kuich W, Vogler H, editors. Handbook of Weighted Automata. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. pp. 213–254.

33. Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition. Comput Speech Lang. 2002;16: 69–88.

34. Cortes C, Haffner P, Mohri M. Rational Kernels: Theory and Algorithms. J Mach Learn Res. 2004;5: 1035–1062.

35. Schwarz RF, Fletcher W, Förster F, Merget B, Wolf M. Evolutionary Distances in the Twilight Zone—A Rational Kernel Approach. PLoS. 2010. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015788

36. Allauzen C, Riley M, Schalkwyk J, Skut W, Mohri M. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. Implementation and Application of Automata. Springer Berlin Heidelberg; 2007. pp. 11–23.

37. Schwarz RF, Watkins TBK, Van Loo P. Accurate phasing of allele-specific copy-numbers for inferring tumour evolution with probe-level resolution. ISMB Late Breaking Research. 2015. Available: https://www.iscb.org/cms_addon/multimedia/flvmedia.php?i=2089

38. Turajlic S, Xu H, Litchfield K, Rowan A, Horswell S, Chambers T, et al. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. Cell. 2018;173: 595–610.e11.

39. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520: 353–357.

40. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149: 979–993.

41. Zaccaria S, Raphael BJ. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. Nat Commun. 2020;11: 4301.

42. Davoli T, Denchi EL, de Lange T. Persistent telomere damage induces bypass of mitosis and tetraploidy. Cell. 2010;141: 81–93.

43. Elhajouji A, Cunha M, Kirsch-Volders M. Spindle poisons can induce polyploidy by mitotic slippage and micronucleate mononucleates in the cytokinesis-block assay. Mutagenesis. 1998;13: 193–198.

44. Shi Q, King RW. Chromosome nondisjunction yields tetraploid rather than aneuploid cells in human cell lines. Nature. 2005;437: 1038–1042.

45. Miroshnychenko D, Baratchart E, Ferrall-Fairbanks MC, Velde RV, Laurie MA, Bui MM, et al. Spontaneous cell fusions as a mechanism of parasexual recombination in tumour cell populations. Nat Ecol Evol. 2021. doi:10.1038/s41559-020-01367-y

46. Shackney SE, Smith CA, Miller BW, Burholt DR, Murtha K, Giles HR, et al. Model for the genetic evolution of human solid tumors. Cancer Res. 1989;49: 3344–3354.

47. van den Bos H, Bakker B, Spierings DCJ, Lansdorp PM, Foijer F. Single-cell sequencing to quantify genomic integrity in cancer. Int J Biochem Cell Biol. 2018;94: 146–150.

48. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472: 90–94.

49. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in

single cells with CHISEL. Nat Biotechnol. 2021;39: 207–214.

50. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011;144: 27–40.

51. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013;153: 666–677.

52. Umbreit NT, Zhang C-Z, Lynch LD, Blaine LJ, Cheng AM, Tourdot R, et al. Mechanisms generating cancer genome complexity from a single cell division error. Science. 2020;368. doi:10.1126/science.aba0712

53. Hadi K, Yao X, Behr JM, Deshpande A, Xanthopoulakis C, Tian H, et al. Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. Cell. 2020;183: 197–210.e32.

54. Laughney AM, Elizalde S, Genovese G, Bakhoum SF. Dynamics of Tumor Heterogeneity Derived from Clonal Karyotypic Evolution. Cell Rep. 2015;12: 809–820.

55. Elizalde S, Laughney AM, Bakhoum SF. A Markov chain for numerical chromosomal instability in clonally expanding populations. PLoS Comput Biol. 2018;14: e1006447.

Contributions

RFS designed the method; MP, SL and RFS implemented the method; RFS and TBKW analysed the data; KH and EC contributed to data analysis; RFS, TBKW, KH, EC and CS wrote the manuscript.

Acknowledgements

R.F.S. and M.P. thank the Helmholtz Association (Germany) for support. T.B.K.W. was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169) as well as the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), Breast Cancer Research Foundation (BCRF), Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. C.S. is a Royal Society Napier Research Professor. His work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169). C.S. is funded by Cancer Research UK (TRACERx, PEACE and CRUK Cancer Immunotherapy Catalyst Network), Cancer Research UK Lung Cancer Centre of Excellence, the Rosetrees Trust, Butterfield and Stoneygate Trusts, NovoNordisk Foundation (ID16584), Royal Society Research Professorships Enhancement Award (RP/EA/180007), the NIHR BRC at

University College London Hospitals, the CRUK-UCL Centre, Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation (BCRF). This research is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. also receives funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) Consolidator Grant (FP7-THESEUS-617844), European Commission ITN (FP7-PloidyNet 607722), an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (835297) and Chromavision from the European Union's Horizon 2020 research and innovation programme (665233).

Competing interests

C.S. acknowledges grant support from Pfizer, AstraZeneca, Bristol Myers Squibb, Roche-Ventana, Boehringer-Ingelheim, Archer Dx Inc. (collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical; is an AstraZeneca Advisory Board Member and Chief Investigator for the MeRmaiD1 clinical trial; has consulted for Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, Celgene, AstraZeneca, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi and the Sarah Cannon Research Institute; has stock options in Apogen Biotechnologies, Epic Bioscience and GRAIL; and has stock options and is co-founder of Achilles Therapeutics. C.S. holds European patents relating to assay technology to detect tumour recurrence (PCT/GB2017/053289); to targeting neoantigens (PCT/EP2016/059401), identifying patent response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), identifying patients who respond to cancer treatment (PCT/GB2018/051912), a US patent relating to detecting tumour mutations (PCT/US2017/28013) and both a European and US patent related to identifying insertion/deletion mutation targets (PCT/GB2018/051892). N.M. and R.R. have stock options in and has consulted for Achilles Therapeutics and holds a European patent in determining HLA LOH (PCT/GB2018/052004).

Figures

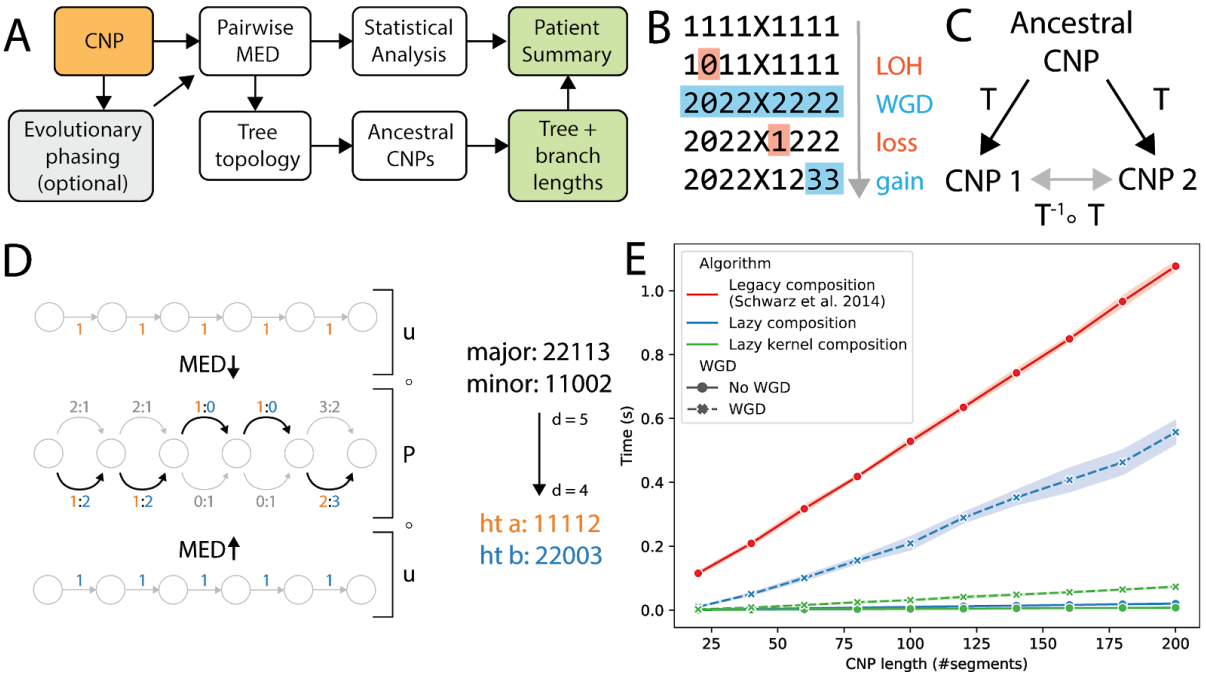


Figure 1: MEDICC2 algorithm and performance. A) Schematic overview of the MEDICC2 workflow: Haplotype-specific CNPs are either pre-phased or undergo evolutionary phasing (see D). Pairwise MEDs are computed between all genomes, followed by tree inference. Ultimately, ancestral CNPs are inferred which determine the final branch lengths of the tree. Results are reported to the user as a patient summary and plot. **B) The copy number transformation problem with WGD:** CNPs are represented as vectors of positive integer copy numbers per chromosome separated by chromosome separator 'X'. To infer the correct MED, LOH events are considered first, as lost segments cannot be re-gained by later events. WGD events span the full CNP, whereas gain and loss events can affect an arbitrary number of segments within a chromosome. **C) Symmetric distance calculation:** MED and MED-WGD are asymmetric distances. The final symmetric distance is computed from a CNP to an arbitrary ancestral CNP and from the ancestral CNP back to the second CNP, thereby minimising over all possible ancestral CNPs. This is achieved by composition of the tree FST T with its inverse T^{-1}. **D) Evolutionary phasing:** CNPs for both alleles are jointly encoded as an unweighted phasing FST where both possible allele configurations are encoded at each position in the sequence (left, middle). Evolutionary phasing then determines the optimal configuration (bold arrows) and extracts final haplotypes (orange and blue) by computation of the MED between the phasing FST and two reference haplotypes (here diploid CNPs of all 1s). An example major/minor and phased configuration is given on the right hand side. **E) Algorithm performance:** Runtime of different composition strategies for the FSTs are shown over CNPs with increasing length from 20 to 200 segments. Computation time of the MEDs is linear with respect to the length of the input sequences. While MED-WGD took significantly longer to compute than the MED without WGD, the new lazy and lazy kernel composition strategies reduced runtime by a factor of between 5 and 10.

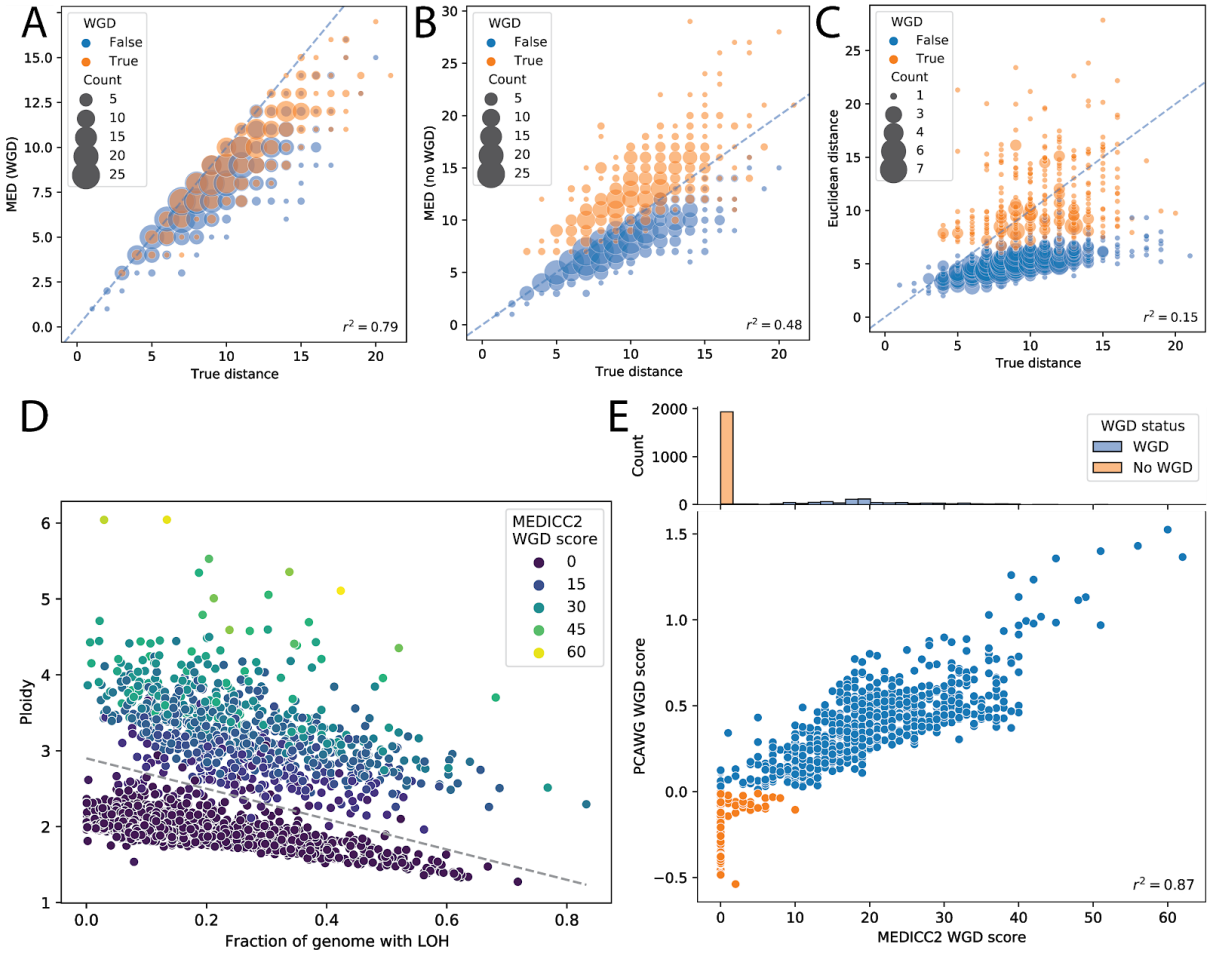


Figure 2: MEDICC2 solves the MED-WGD problem. Using 1000 random instances of copy number evolution starting from a diploid genome, MEDICC2 correctly infers distances no greater than the true simulated distance if using MED-WGD **(A)**, but overestimates distances in sequences where WGD events have occurred if using the MED without taking WGD events into account **(B)**. For comparison, Euclidean distance performs poorly for both genomes with and without WGD (r^2=0.15) **(C)**. **D) The MEDICC2 WGD score for 2778 cancer genomes:** Individual cancers are plotted based on their average ploidy and fraction of genome with LOH. The original separating line between WGD and non-WGD tumours was estimated by Dentro et al. 2020 as y = 2.9 - 2x. Colors indicate the MEDICC2 WGD score and follow the gradient orthogonal to the separating line without further parameter estimation. **E) Comparison between MEDICC2 WGD score and the PCAWG WGD score.** Each PCAWG tumour was projected orthogonally onto the separating line from Dentro et al 2020 and the (signed) distance from the line (PCAWG WGD score) is plotted against the MEDICC2 WGD score, showing a high degree of correlation (r^2=0.87).

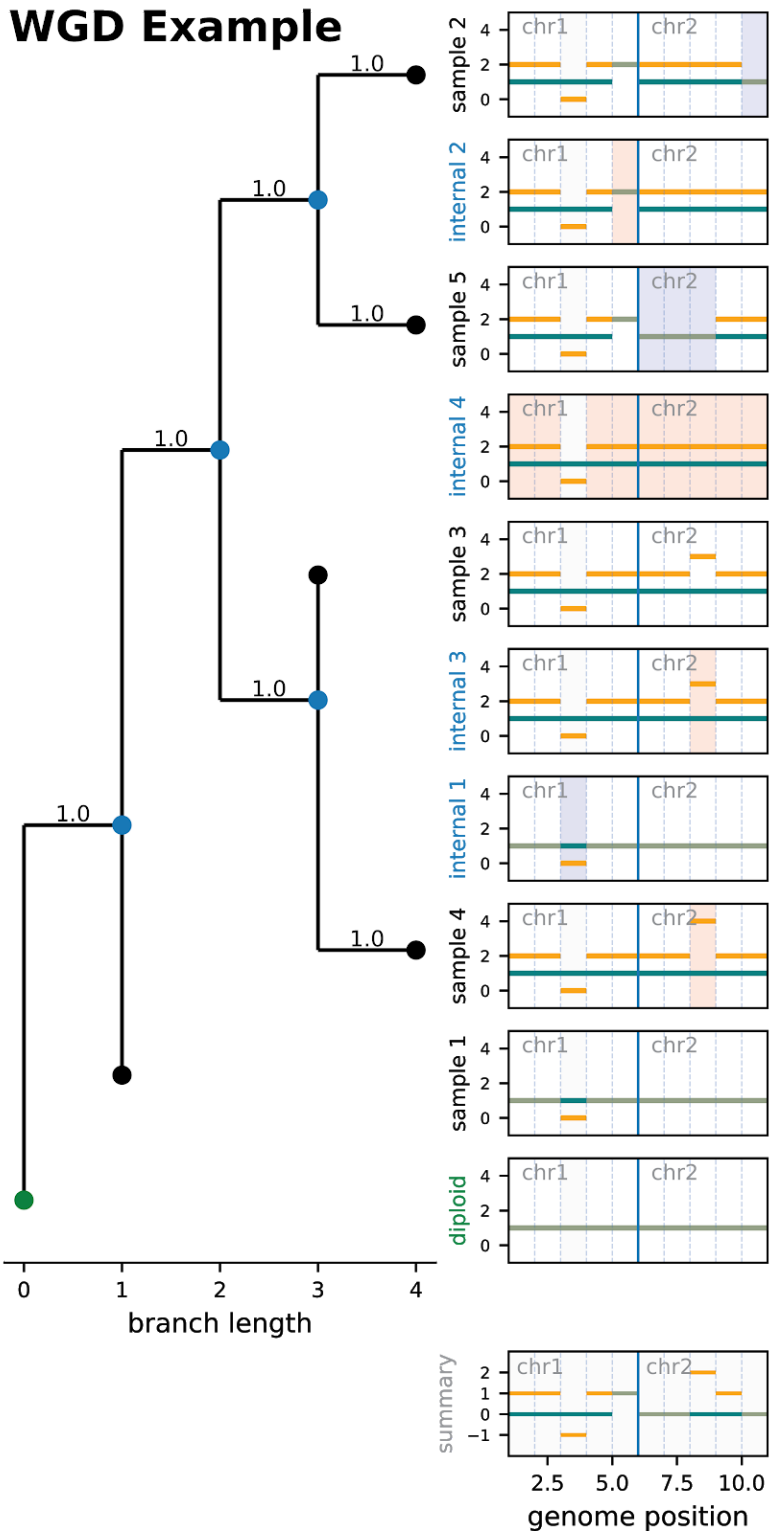


Figure 3: Simulated example of copy number evolution with subclonal WGD. CNPs for both haplotypes (orange and blue) for five samples were simulated over two chromosomes and 10 copy number segments. A subclonal haplotype doubling event (orange allele) occurred at the second internal node (internal 4) from the root of the tree, leaving one terminal node (sample 1) undoubled. MEDICC2 identifies the location of events in the phylogeny and marks up affected segments in the node where the event occurs: gains (red), losses (blue). A summary of all change events (sum over all nodes) is given at the bottom.

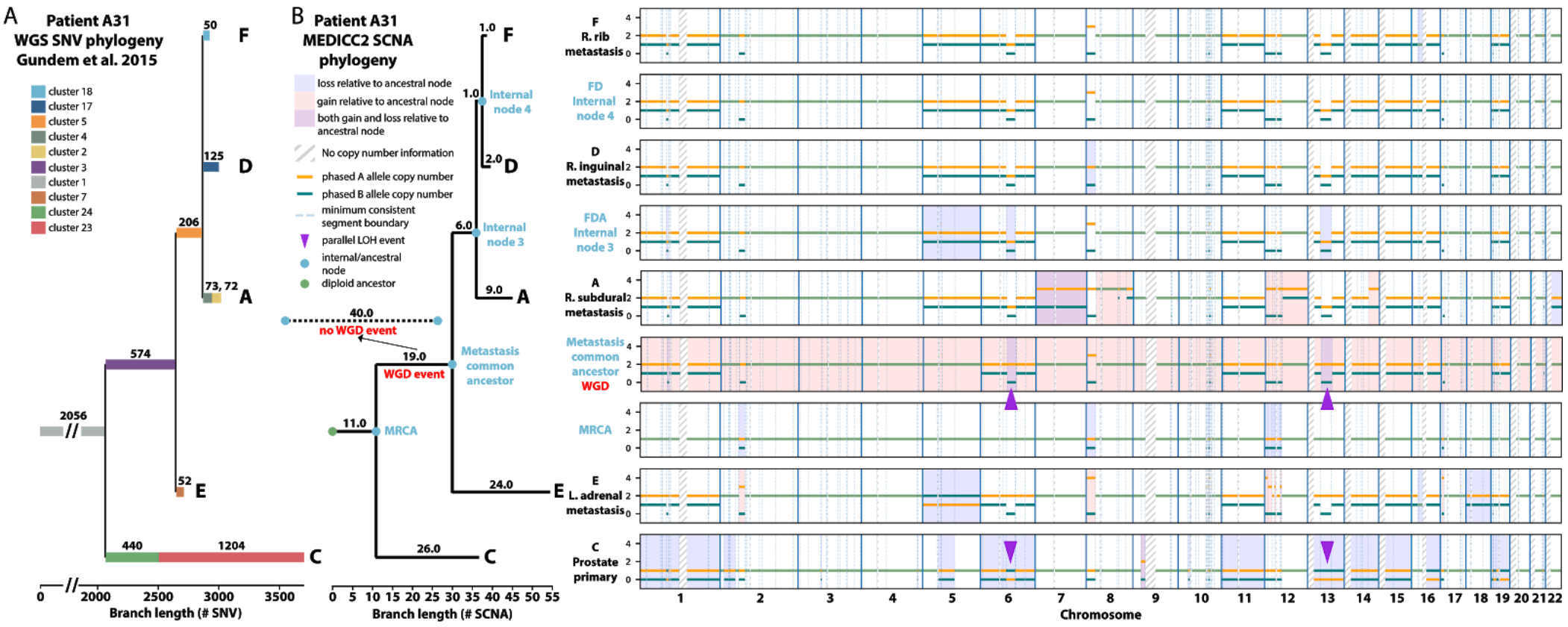
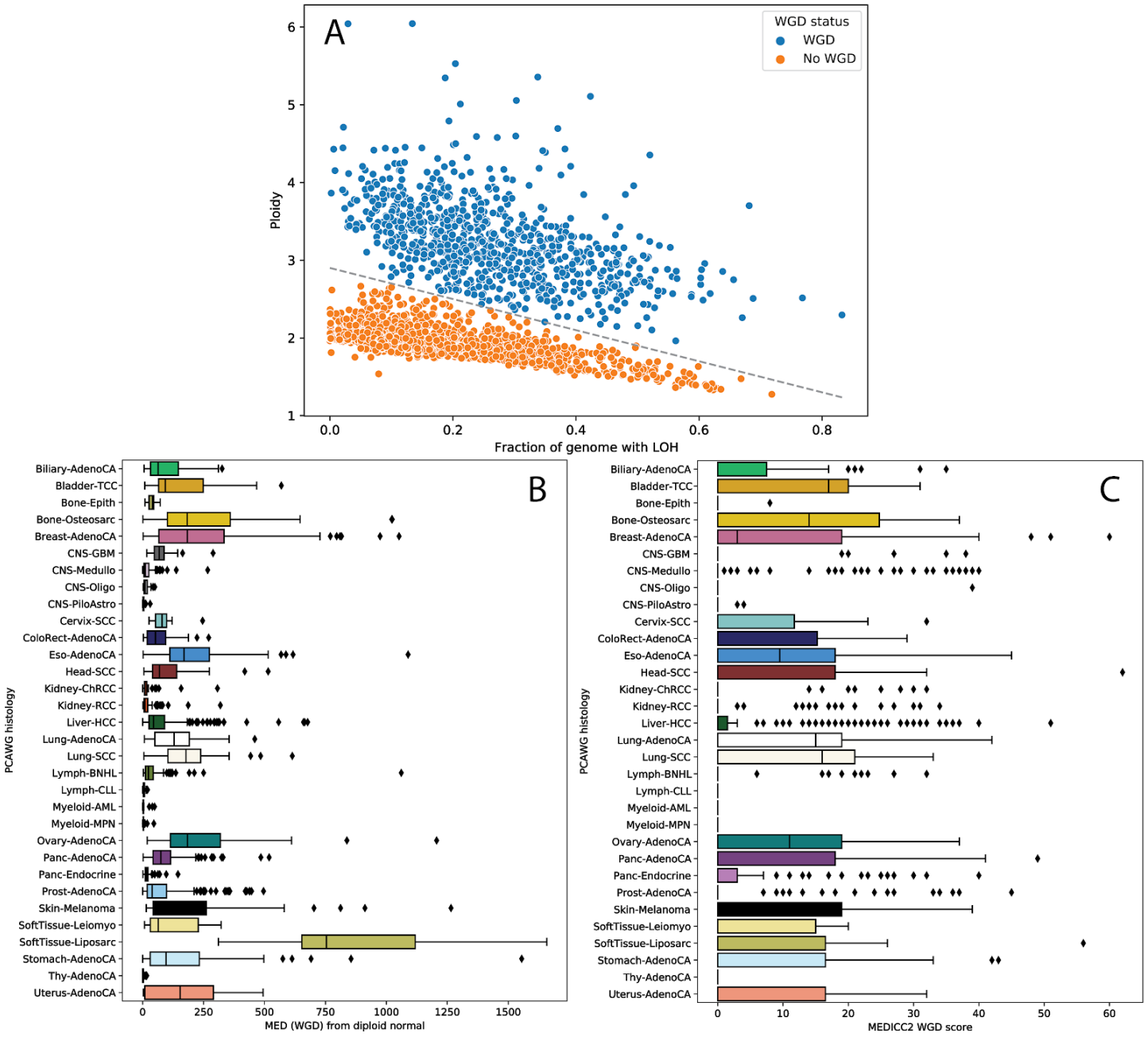


Figure 4: Evolutionary history of tumour subclones from patient "A31".
A) SNV-based phylogeny: Reproduction of the SNV-based phylogeny as described in Gundem et al. 2015 for the multi-sample prostate cancer tumour case with one sample ("C") from the primary tumour and four samples ("A", "D", "E", and "F") from distinct metastatic sites. Original reconstruction was performed using an n-dimensional Bayesian Dirichlet process to cluster estimated cancer cell fractions of the single nucleotide variants (SNV) identified in the WGS across samples. Only the dominant clones of each sample are given.
B) MEDICC2 phylogeny: Using multi-sample phased CNPs, MEDICC2 detected the presence of WGD in the metastatic samples and its absence in the primary sample from A31. The MEDICC2 analysis identifies multiple MSAI events as well as parallel LOH on chromosomes 6 and 13 (purple arrows). Individual events are marked in the copy number track where they occur: gains (red) and losses (blue) (see also Figure 3).



Supplementary Figure 1: Overview of the PCAWG cohort.
A) Original plot reproduced from Dentro et al. 2020. A separating line (dashed gray, y = 2.9 - 2x) separated WGD (blue) from non-WGD (orange) tumours.
B) MEDICC2 MED-WGD for all PCAWG tumours show differing levels of copy number evolution. Soft-tissue Liposarcoma across the board showed the greatest number of copy number events.
C) MEDICC2 WGD score for all PCAWG tumours shows varying degrees of WGD events across the cohort.