PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Refphase: Multi-sample phasing reveals haplotype-specific copy number heterogeneity

**Thomas B. K. Watkins[1,2☯], Emma C. Colliver[2☯], Matthew R. Huska[3☯¤], Tom L. Kaufmann[5,4,3,13], Emilia L. Lim[2,1], Cody B. Duncan[13], Kerstin Haase[6,7,8], Peter Van Loo[2,9,10], Charles Swanton[2,1,11‡], Nicholas McGranahan[1,12‡*], Roland F. Schwarz[13,5,3‡*]**

**1** Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, United Kingdom, **2** The Francis Crick Institute, London, United Kingdom, **3** Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC) Berlin, Germany, **4** Department of Electrical Engineering & Computer Science, Technische Universität Berlin, Berlin, Germany, **5** BIFOLD—Berlin Institute for the Foundations of Learning and Data, Berlin, Germany, **6** German Cancer Consortium (DKTK), partner site Berlin, and German Cancer Research Center (DKFZ), Heidelberg, Germany, **7** Experimental and Clinical Research Center (ECRC) of the MDC and Charité Berlin, Berlin, Germany, **8** Department of Pediatric Oncology and Hematology, Charité–Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany, **9** Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **10** Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **11** Department of Medical Oncology, University College London Hospitals, London, United Kingdom, **12** Cancer Genome Evolution Research Group, University College London Cancer Institute, London, United Kingdom, **13** Institute for Computational Cancer Biology (ICCB), Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

☯ These authors contributed equally to this work.
¤ Current address: Genome Competence Center (MF1), Robert Koch Institute, Berlin, Germany
‡ These authors are joint senior authors on this work.
* nicholas.mcgranahan.10@ucl.ac.uk (NMG); roland.schwarz@uni-koeln.de (RFS)





**Citation:** Watkins TBK, Colliver EC, Huska MR, Kaufmann TL, Lim EL, Duncan CB, et al. (2023) Refphase: Multi-sample phasing reveals haplotype-specific copy number heterogeneity. PLoS Comput Biol 19(10): e1011379. https://doi.org/10.1371/journal.pcbi.1011379

**Editor:** Maxwell Wing Libbrecht, Simon Fraser University, CANADA

**Received:** December 13, 2022

**Accepted:** July 22, 2023

**Published:** October 23, 2023



**Data Availability Statement:** Refphase is available as a user-friendly standalone R package from http://bitbucket.org/schwarzlab/refphase. All code relating to this publication is also available at Zenodo https://doi.org/10.5281/zenodo.7148458. All datasets used in this study can be found in S1 Table. Clinical trial information (if applicable) is available in the associated publications: doi: 10.1056/NEJMoa1616288 doi: 10.1038/nature22364 doi: 10.1038/ng.3214 doi: 10.1158/2159-8290.

## Abstract

Most computational methods that infer somatic copy number alterations (SCNAs) from bulk sequencing of DNA analyse tumour samples individually. However, the sequencing of multiple tumour samples from a patient's disease is an increasingly common practice. We introduce Refphase, an algorithm that leverages this multi-sampling approach to infer haplotype-specific copy numbers through multi-sample phasing. We demonstrate Refphase's ability to infer haplotype-specific SCNAs and characterise their intra-tumour heterogeneity, to uncover previously undetected allelic imbalance in low purity samples, and to identify parallel evolution in the context of whole genome doubling in a pan-cancer cohort of 336 samples from 99 tumours.

## Author summary

In the course of cancer evolution, cancer genomes frequently change the number of copies of their chromosomes or parts thereof. These changes in copy number can be estimated from DNA sequencing data, but current tools can typically not identify which parental

**Funding:** This work was supported by the Francis Crick Institute that receives its core funding from Cancer Research UK (CC2008, CC2041), the UK Medical Research Council (CC2008, CC2041) and the Wellcome Trust (CC2008, CC2041). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission. CS is a Royal Society Napier Research Professor (RSRP\R\210001); CS is funded by Cancer Research UK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; NovoNordisk Foundation (ID16584); Royal Society Professorship Enhancement Award (RP/EA/180007); National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre; the Cancer Research UK–University College London Centre; the Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US); the Mark Foundation for Cancer Research Aspire Award (21-029-ASP); and is in receipt of an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (835297). This work was supported by a Stand Up To Cancer?LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17 to S. M. Dubinett and A. E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. TBKW is funded by the Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025). ECC is funded by Cancer Research UK TRACERx (C11496/A17786). NM is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (211179/Z/18/Z) and also receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. PVL is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. PVL is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). NM is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant Number 211179/Z/18/Z), and also receives funding from Cancer Research UK Lung Cancer Centre of Excellence, Rosetrees, and the NIHR BRC at University College London Hospitals.

copy of a chromosome the copy number change occurred on. We here introduce Refphase, a new algorithm and software tool for assigning copy number changes to the parental chromosomal copy they have occurred on. This assignment allows us to measure differences between cancer genomes in much more detail and allows us to systematically catalogue different types of evolutionary events, including the detection of parallel evolution where highly similar genetic changes occur in different parts of the evolutionary tree simultaneously. As an example, we show how after whole-genome duplication successive losses of chromosomes affect different parental copies in different regions of a tumour, reducing the average copy number from four copies to close to three copies independently.

## Background

As a consequence of genomic instability, cancers accumulate somatic mutations [1]. These include somatic copy number alterations (SCNAs) affecting focal genomic segments, chromosome arms, or entire chromosomes, and whole genome doubling (WGD) events that alter the entire karyotype. SCNAs change the number of physical copies of a given genomic region and often result in aneuploidy, which affects up to 90% of solid tumours [2,3]. These mutational events may be clonal, shared by all cancer cells, or subclonal and thus present only in a subset of cells, resulting in intra-tumour heterogeneity (ITH) [4,5]. It is now clear that single-sample sequencing studies derived from single biopsies are often insufficient to capture the extent of mutational heterogeneity and the field increasingly relies on multi-sample bulk and single-cell analyses from the same tumour to better describe this complexity. Such studies have permitted the mapping of the landscape of clonal and subclonal SCNAs [6–12], and have revealed a relationship between SCNA intra-tumour heterogeneity and poor prognosis in multiple tumour types [6,8,13–16]. Additionally, chromosomal instability and aneuploidy have been associated with and proposed as causes of cancer drug resistance [17] and linked to metastasis [18–21]. Inference of SCNAs and quantification of their intra-tumour heterogeneity is therefore of clinical importance and a prerequisite for understanding tumour evolution.

SCNAs are typically identified in an allele-specific manner from DNA sequencing or single nucleotide polymorphisms (SNP) array data using two measures: the log read-depth ratio (LogR) of a genomic locus between the tumour and a matched normal sample, which informs total copy number estimates; and the B-allele frequency (BAF) at heterozygous SNPs, which informs allelic imbalance estimates [22]. BAF and LogR profiles are sensitive to sequencing noise and sample purity. The difficulty of characterising SCNAs in low tumour purity bulk samples can lead to many being discarded [23]. Imposing such purity thresholds on analyses is likely problematic as sample purity co-segregates with other important clinical covariates and survival [23]. Therefore, approaches that accurately resolve copy number states in low purity samples are of clinical interest.

Additionally, while BAF and LogR allow inference of allele-specific SCNAs, the resulting copy number profiles and underlying SNPs are often not phased, meaning that SCNAs cannot be assigned to the physical haplotypes they reside on. Instead, allele-specific copy number states are typically reported in an unphased major/minor configuration, where major and minor refer to the greater and lesser copy number at a genomic locus respectively. However, phasing of SCNAs can offer additional insights into the clonality of mutational events, giving resolution on whether the same parental allele has been gained or lost across different samples in a single tumour.

RFS is a Professor at the Cancer Research Center Cologne Essen (CCCE) funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia. CBD is also supported by the CCCE. RFS, MRH and TLK thank the Helmholtz Association (Germany) for support. RFS and TLK received funds from the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** I have read the journal's policy and the authors of this manuscript have the following competing interests: CS acknowledges grant support from AstraZeneca, Boehringer-Ingelheim, BMS, Pfizer, Roche-Ventana, Invitae (previously Archer Dx, collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical, and Personalis; is an AstraZeneca advisory board member and chief investigator for the AZ MeRmaiD 1 and 2 clinical trials and is also co-chief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's scientific advisory board; receives consultant fees from Achilles Therapeutics (also a scientific advisory board member), Bicycle Therapeutics (also a scientific advisory board member), Genentech, Medicxi, China Innovation Centre of Roche (CICoR) formerly Roche Innovation Centre – Shanghai, Metabomed (until July 2022) and the Sarah Cannon Research Institute; has received honoraria from Amgen, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, Illumina, and Roche-Ventana; had stock options in Apogen Biotechnologies and GRAIL until June 2021, and currently has stock options in Epic Bioscience, Bicycle Therapeutics, and has stock options and is co-founder of Achilles Therapeutics; is listed as an inventor on a European patent application relating to assay technology to detect tumour recurrence (PCT/GB2017/053289), the patent has been licensed to commercial entities and, under his terms of employment, CS is due a revenue share of any revenue generated from such license(s); holds patents relating to targeting neoantigens (PCT/EP2016/059401), identifying patient 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), methods for lung cancer detection (US20190106751A1) and both a

Statistical phasing utilises large collections of genotypes [24] and local linkage disequilibrium structure to phase SNPs [25,26]. Multiple groups have previously implemented statistical phasing approaches in the context of whole-genome sequencing (WGS) [27], single-sample bulk sequencing studies [27,28] and in single-cell studies, using DNA [29] and RNA [30]. However, while highly accurate locally, statistical phasing accuracy rapidly decreases with increasing genomic distance, limiting the genomic span within which a SNP and the corresponding SCNA can accurately be assigned to its haplotype-of-origin. Therefore, statistical phasing for SCNA detection is mostly restricted to WGS data, ignoring a large proportion of cancer genomics studies based on whole-exome (WES). Additionally, while single-cell DNA sequencing removes the complications of sample purity, and several novel tools for phasing-based estimation of haplotype-specific copy-numbers from single cells have been introduced [29,31,32], the low genomic coverage in single-cell DNA sequencing makes allele-specific or haplotype-specific copy-number calling a difficult task. For this reason and owing to low cell throughput and high costs, single-cell DNA sequencing is so far not routinely used in clinical cancer genomics studies, where the majority of data is still generated from bulk sequencing.

To address these challenges, we present Refphase. Refphase phases the copy number states in samples comprising a multi-sample tumour against an internal reference sample, an approach we refer to as *multi-sample reference phasing* [9]. In doing this, we use Refphase to infer haplotype-specific copy number states, rescue previously undetected SCNAs in low purity samples, and standardise quantification of SCNA-based ITH. Refphase is, to our knowledge, the first long-range phasing algorithm applicable to all of multi-sample WGS, SNP array, and exome. Unlike statistical phasing, Refphase does not require reference haplotype panels or large collections of genotypes. Instead, it leverages the common germline background between multiple samples from the same patient to phase heterozygous SNPs and SCNAs. We have previously used reference phasing to describe mirrored subclonal allelic imbalance (MSAI), independent SCNAs that occur on opposite haplotypes in different samples from the same tumour, and to identify parallel and convergent copy number events [8,9].

Refphase takes as input for a multi-sample tumour the copy number profiles for each sample generated by commonly used copy number tools relying on single-sample approaches, such as ASCAT [22] and others [33–35]. Refphase provides output compatible with MEDICC2 [36]. This end-to-end approach enables the streamlined processing of multi-sample copy number data from raw read counts to event-based SCNA phylogenies.

Here, we demonstrate Refphase's ability to infer SCNAs and characterise their intra-tumour heterogeneity; compare the SCNA landscapes in primary and metastatic tumour samples from a single patient's disease; showcase its ability to uncover previously undetected allelic imbalance in low purity samples; and use it to identify parallel evolution in the context of whole genome doubling in a pan-cancer cohort.

## Results

### Refphase algorithm

Refphase takes as input a copy number segmentation for each of $\boldsymbol{N}$ tumour samples from the same patient which it processes in four discrete steps or modules (Fig 1 and Methods).

First, a minimum consistent segmentation is created from the union of all segment start and end positions of the input copy number segmentations of each tumour sample (Fig 1A and Methods). Breakpoints separated by a distance smaller than a user-defined maximum gap size (default = 100kbp) are subsequently merged, unless these breakpoints belong to the same sample-of-origin. This exception is invoked to preserve focal gains and losses present in the original samples. This minimum consistent segmentation step yields a final set of $\mathbf{m}$ bins.

European and US patent related to identifying insertion/deletion mutation targets (PCT/GB2018/051892) and is listed as a co-inventor on a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987) and is a named inventor on a provisional patent protection related to a ctDNA detection algorithm. NM has received consultancy fees and has stock options in Achilles Therapeutics. NM holds European patents relating to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/ EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221). All other authors declare no competing interests.

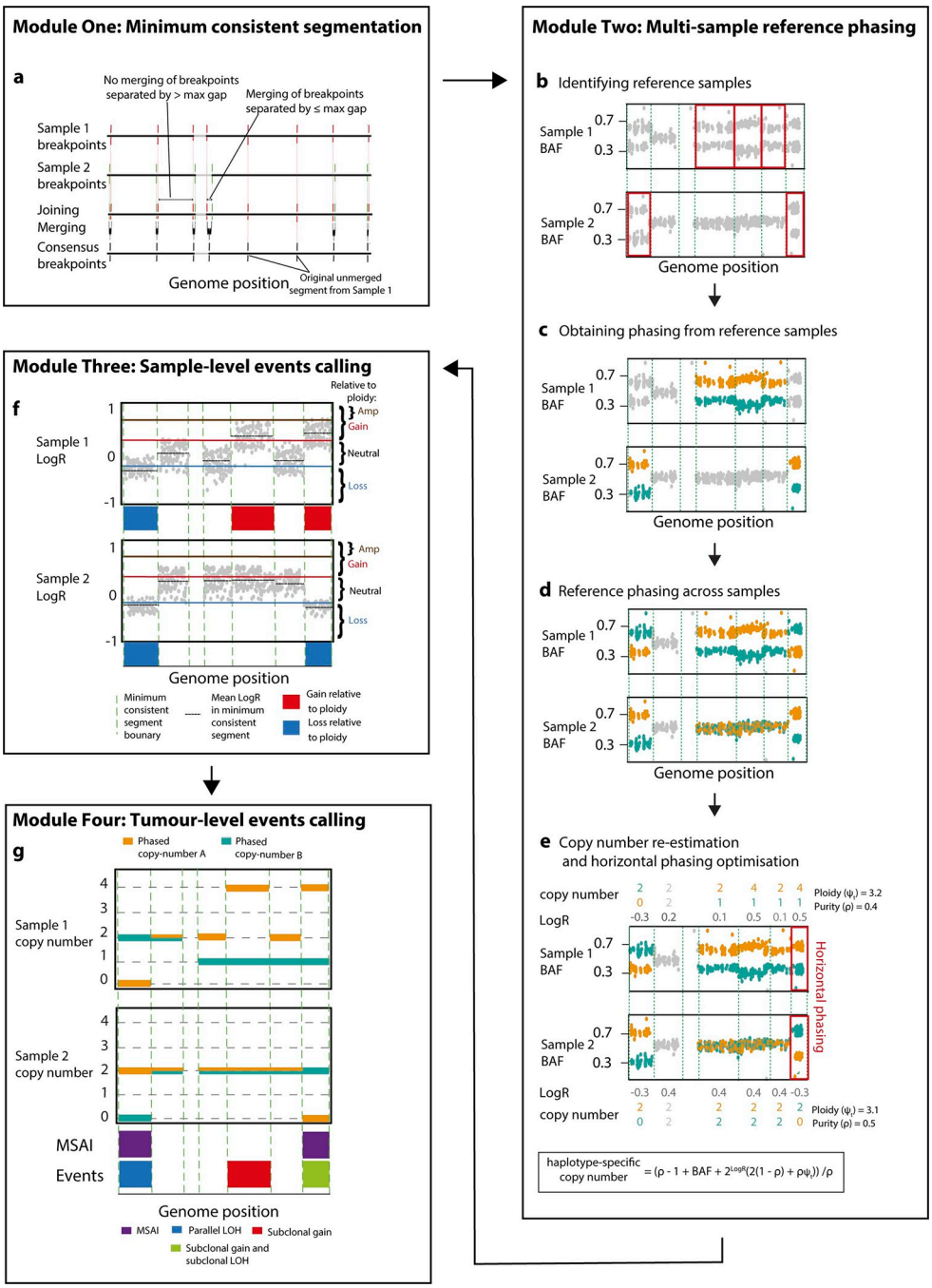


**Fig 1. Overview of Refphase algorithm. a)** Refphase creates a minimum consistent segmentation across the single-sample segmentations input for each tumour. **b)** In each segment in which at least one sample had allelic imbalance in the tumour input, an optimal reference sample for phasing is determined. **c)** The phasing of each reference sample is derived from its BAF. **d)** Phasing is then applied to the BAFs in all other samples which are not the reference. **e)** Allele-specific copy numbers are re-estimated for each sample utilising the reference phasing, and the most parsimonious phasing solution along each chromosome is then chosen in horizontal phasing optimization. **f)** In each segment, event categories relative to the input ploidy of the corresponding sample are defined using LogR values. **g)** Tumour-level events are called and intra-tumour heterogeneity metrics calculated.

https://doi.org/10.1371/journal.pcbi.1011379.g001

Next, using this set of **m** bins, multi-sample reference phasing is performed (Fig 1B and 1E). Heterozygous SNPs are either defined by the user or identified from the BAF values of the normal sample (Methods). For each bin $\mathbf{m_i}$, the sample with the highest degree of allelic imbalance is then identified, which will act as a reference $\boldsymbol{n}_{\mathbf{ref}}$ for $\mathbf{m_i}$ (Fig 1B and Methods). This reference sample $\boldsymbol{n}_{\mathbf{ref}}$ is then used to assign alleles of all heterozygous SNPs within $\mathbf{m_i}$ to the "A" or "B" haplotype based on their BAFs being greater than or less than a threshold of 0.5 (Fig 1C). This haplotype assignment in turn is then applied to all other tumour samples for the same bin $\mathbf{m_i}$ (Fig 1D) and each tumour sample is assessed for the presence of any additional previously undetected allelic imbalance using an effect size threshold based on Cohen's d (Methods). Haplotype-specific integer copy number states are then re-estimated for all reference phased segments (Fig 1E and Methods).

Reference phasing as described above accurately phases the haplotypes of all samples for each bin relative to one another, but independently of other bins. To phase bins along the genome "horizontally", we next estimate this horizontal phasing by using an evolutionary criterion. Briefly, the assignment of heterozygous SNPs to "A" and "B" haplotypes for all bins within a single chromosome is chosen to minimise the number of copy number breakpoints across all tumour samples [37] (Figs 1E and S1 and Methods). The relative phasing between samples from the previous step therefore remains unchanged.

The third module then uses the re-estimated copy number and updated allelic imbalance states to determine whether a segment from a tumour sample may be categorised as an SCNA relative to the ploidy of that specific tumour sample, in keeping with both previous genomics studies [8,9,38] and clinical practice [39]. This relative-to-ploidy classification enables the comparison of the same area of the genome in bin $\mathbf{m_i}$ between samples with differing ploidies and classifies segments as either an amplification, gain, neutral, or loss relative to sample ploidy (Fig 1F and Methods). Loss of heterozygosity (LOH), copy-neutral loss of heterozygosity (CNLOH) and homozygous deletion events are also classified (Methods).

The fourth and final module integrates the multi-sample reference phasing and relative-to-ploidy classification to produce a tumour-level estimate of SCNA event clonality across all samples and to infer the presence of MSAI (Fig 1G). All events are then summarised as clonal (present in all tumour samples without MSAI) or subclonal (present in only a subset of tumour samples, or in all samples but with MSAI detected) (Methods).

## Refphase identifies mirrored subclonal allelic imbalance and parallel evolution

We first demonstrated the functionality of Refphase by analysing WES data from three spatially separated primary tumour samples and a matched normal from a non-small cell lung cancer patient CRUK0034 from the TRACERx 100 cohort [8]. All three tumour samples were pre-segmented with ASCAT [22] and subjected to reference phasing. Dividing the genome into 178 bins, Refphase identified either clonal or subclonal allelic imbalance in 88.6% of the genome, with the remaining 11.4% of the genome allelically balanced. Fig 2 shows parts of the Refphase output for this tumour. A more complete output example is available in S2 Fig. For this example and the remainder of the text, we refer to unphased copy number states in major/minor configuration as e.g. 4/1 and to phased copy number states as e.g. 1|4, in line with genotype notation conventions.

While some SCNAs (e.g. gain on 1q) are clonal and present in all samples from the tumour, Refphase also identifies substantial inter-sample heterogeneity (Figs 2A, 2B, and S2). For example, 3p demonstrates allelic imbalance only in R2, where even after reference phasing, no allelic imbalance was detectable in the other samples, as can be seen from the well mixed BAF

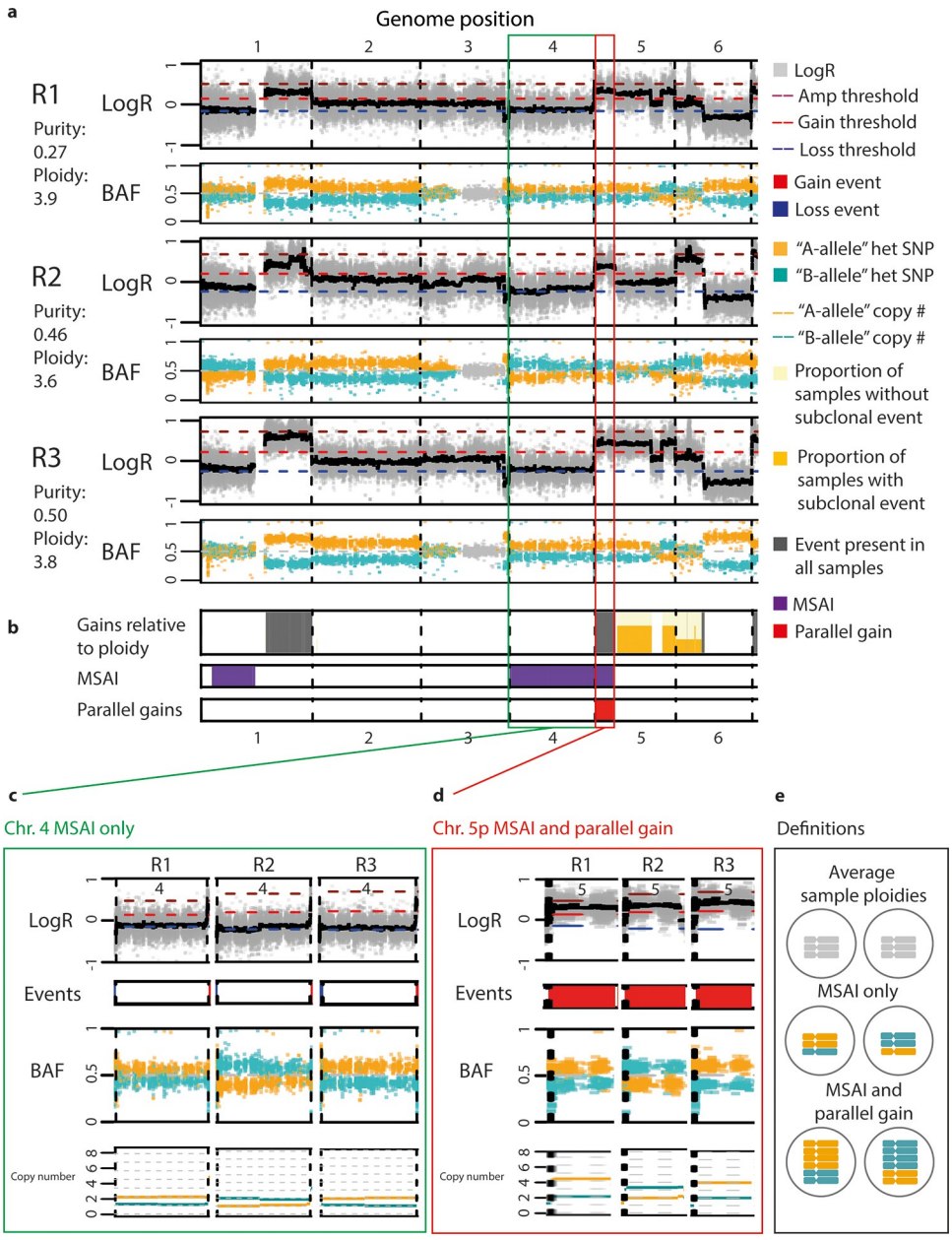


**Fig 2. Detection of mirrored subclonal allelic imbalance and parallel evolution. a)** LogR and BAF tracks in chromosomes 1 to 6 from tumour CRUK0034. LogR tracks show LogR values in light grey points. The black line shows the median LogR within a minimum consistent segment. BAF tracks show phased BAF as either orange "A" haplotype points or blue "B" haplotype points. Unphased BAF values are shown as light grey points. **b)** SCNA summary tracks showing (top) gains relative to ploidy. The full height grey bar indicates that a gain is identified in every sample from tumour CRUK0034. A light yellow background indicates the presence of a subclonal gain, and the height of the stacked darker yellow bar indicates the proportion of samples in which a subclonal gain is present. (middle) Track indicates the presence of MSAI between at least two samples, shown by purple fill. (bottom) Track indicates the presence of parallel gains, shown by red fill. **c)** MSAI detected from tumour CRUK0034 affecting chromosome 4. **d)** Parallel evolution of chromosome arm 5p gain from tumour CRUK0034. **e)** Schematic of the copy number states related to MSAI and parallel evolution.

https://doi.org/10.1371/journal.pcbi.1011379.g002

values for both haplotypes (orange and blue) on 3p in R1 and R3. Refphase identifies subclonal SCNA events including subclonal gain of 5q affecting R1 and R3, and a subclonal gain of 6p affecting R2.

Reference phasing also permitted the identification of both MSAI and parallel evolution in this tumour. One additional copy of chromosome 4 relative to diploid (Fig 2C) (copy number state 2/1) is present in all tumour samples. Reference phasing revealed that this additional copy was derived from the "A" haplotype (orange) in samples R1 and R3, and from the "B" haplotype (blue) in sample R2, an example of MSAI [8]. A second instance of MSAI on chromosome 5p co-occurs with relative-to-ploidy gains in all three samples with copy number states of 4|2 in R1 and R3 but of 2|3 in R2 (Fig 2D). A schematic outlining the distinction between the classifications applied to the MSAI affecting chromosome 4 and the parallel gains affecting 5p can be seen in Fig 2E. In addition to the detection of MSAI on chromosome 4 and parallel evolution of 5p gain, Refphase also identified an additional instance of MSAI affecting 1p (Fig 2B).

Since MSAI and parallel events are not detectable from the analysis of single tumour samples nor from unphased data, Refphase's ability to identify haplotype-specific copy number states and its event classifications provide a refined view of the evolution of this tumour.

## Refphase quantifies SCNA heterogeneity in complex multi-sample cases

The increasing availability of DNA sequencing data from multiple tumour samples from the same patient has begun to address complex questions regarding SCNA intra-tumour heterogeneity and changes in SCNAs during metastatic dissemination. However, there are no standardised frameworks in which such questions may be addressed and compared between studies. Refphase not only supports the quantification of SCNA clonality across all samples from the same patient but also allows standardised analysis and comparisons of user-defined subgroups of samples within a single patient's disease. In addition, Refphase produces correctly formatted input for the state-of-the-art SCNA phylogenetic reconstruction algorithm MEDICC2 [36], as well as the option to output its own naive haplotype-specific clustering of minimum consistent segments.

Patient CRUK0063, previously analysed in work by Abbosh *et al*. in 2017 [40], was examined through the PEACE post-mortem study 24 hours after death. WES data from five post-mortem tumour samples (paravertebral and lung metastases) and five primary tumour samples were pre-segmented using ASCAT and subjected to reference phasing using Refphase. We investigated differences in SCNA profiles between primary and metastatic samples (Fig 3). MEDICC2 [36] was run on the Refphase output and grouped the metastasis samples together in one clade and the primary samples in a separate clade (Fig 3A). Refphase identified SCNA events present in both primary and metastatic samples (Fig 3B), present in primary samples only (Fig 3C), and present in metastatic samples only (Fig 3D).

Examining all CRUK0063 samples together revealed clonal SCNAs (SCNAs here encompassing relative-to-ploidy gains and losses, and LOH events) affecting 25% of the genome, subclonal SCNAs affecting 69% of the genome, MSAI affecting 26% of the genome and parallel evolution evident in 8% of the genome. Refphase also provided the opportunity to analyse the five primary and five metastatic samples separately, while using the phasing derived from all samples. Differences between the two groups emerged and enabled us to distinguish between SCNAs that are clonal and subclonal in the primary samples (primary-clonal/primary-subclonal) and those that are clonal and subclonal in the metastatic samples (metastasis-clonal/metastasis-subclonal). Metastasis-clonal SCNAs were far more prevalent than primary-clonal SCNAs, both in terms of proportion of the genome affected (55% *vs*. 36%), and of proportion

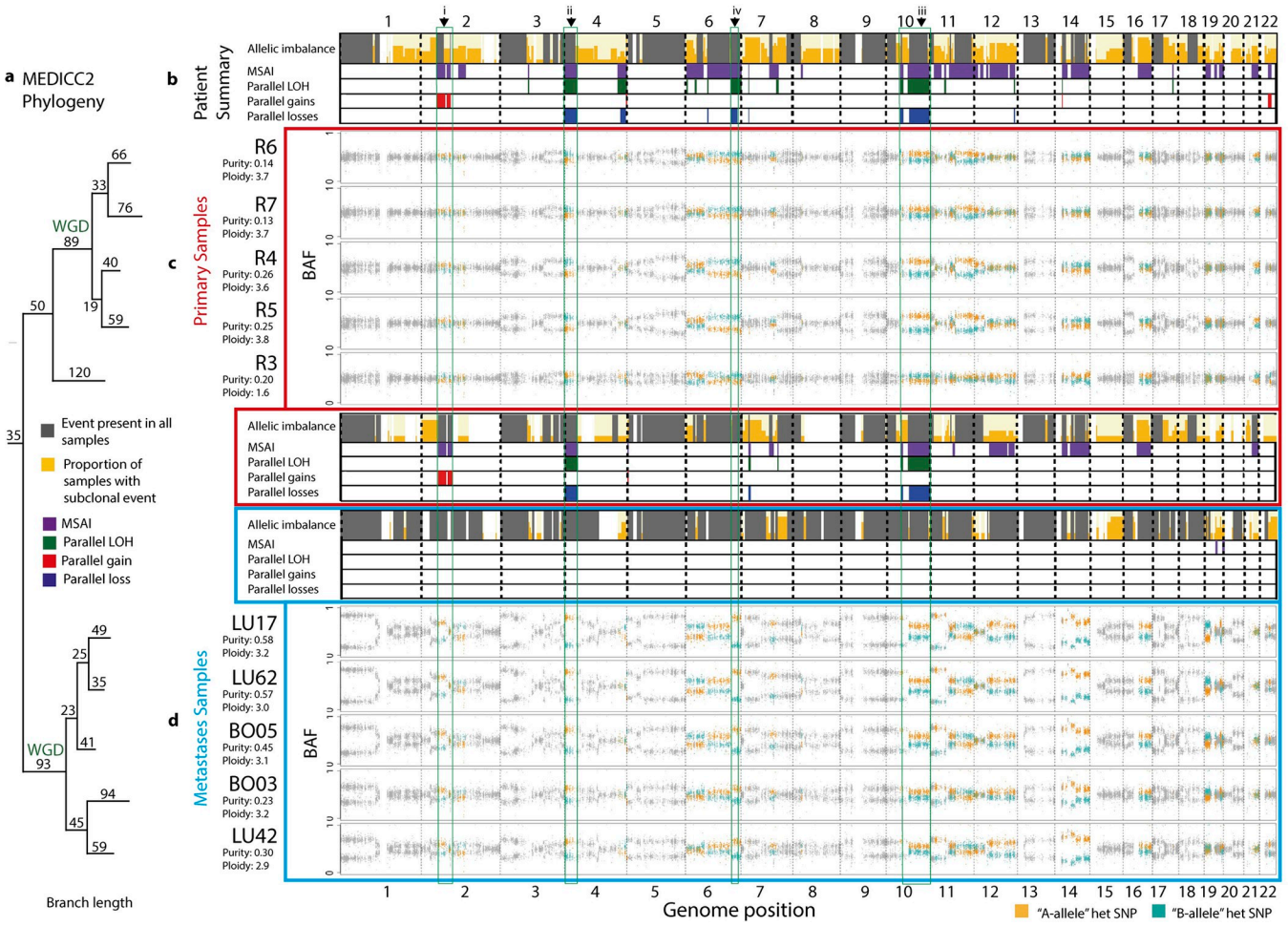


**Fig 3. Analysis of CRUK0063 multi-sample and multi-time point NSCLC case. a)** MEDICC2 phylogeny. Multi-sample reference phased allele-specific copy number output from Refphase can be passed directly to MEDICC2 to produce a phylogenetic reconstruction. **b)** SCNA summary tracks for all samples—primary and metastases—from patient CRUK0063. **c)** BAF profiles and SCNA summary tracks for the primary samples from CRUK0063 are indicated by a red border. **d)** SCNA summary tracks and BAF profiles from five post-mortem metastatic samples with a blue border. Sample BAF tracks are ordered by their position in the MEDICC2 phylogenetic reconstruction. Green boxes, arrowheads and associated Roman numerals highlight selected examples of MSAI on chromosomes 2p, 6q, 4p and 10q, described in the main text. WGD indicates whole genome doubling events inferred by MEDICC2.

https://doi.org/10.1371/journal.pcbi.1011379.g003

of segments with any SCNA event (relative-to-ploidy gain or loss, or LOH) for which a clonal SCNA was identified, quantified within the respective group of samples (55% *vs*. 33%). Metastasis-clonal allelic imbalance and LOH were particularly prevalent, affecting 72% and 34% of the genome respectively, while primary-clonal allelic imbalance and LOH affected only 39% and 27% of the genome. The majority of LOH was found to be shared between primary and metastatic samples.

In line with the high levels of metastasis-clonal SCNAs in CRUK0063, the metastatic samples were also characterised by a relative absence of both MSAI and parallel evolution compared to the primary samples (proportion of genome: MSAI, 0.3% *vs*. 13%; parallel events, encompassing parallel gains, parallel losses and parallel LOH events, 0% *vs*. 5%). Examples of specific events observed when analysing the primary samples alone are visible in Fig 3 and include parallel gain of 2p (i), and multiple parallel LOH events including on 4p (ii) and 10q (iii). Specifically, primary tumour sample R3 had a different major haplotype to other primary samples at these loci. The divergence of sample R3 from other primary samples is reflected in

it branching earlier than the remaining four primary samples in the accompanying MEDICC2 phylogeny. In the Abbosh et al. study [40], mutational phylogenetic analysis suggested that all metastatic samples arose from a single ancestral subclone; similar results are observed using the MEDICC2 phylogenies with all metastatic samples belonging to a single clade. There is also MSAI affecting 13% of the genome and an instance of parallel evolution of LOH affecting 6q between the primary and metastatic samples that are not identified within either the primary samples or metastatic samples alone (Fig 3C and 3D (iv)). Collectively, these MSAI and parallel evolution results suggest that the metastatic samples demonstrate less inter-sample heterogeneity than the primary samples in CRUK0063 and the presence of MSAI and parallel evolution between the primary samples and metastatic samples suggests continued copy number evolution. One area of the genome subject to parallel gain in CRUK0063 is a region of chromosome arm 2p (i), encompassing 2p16, that overlaps the second most commonly amplified locus in lung squamous cell carcinoma revealed by TCGA through GISTIC2 analysis [41,42]. This locus contains the transcription factor *BCL11A*, a known oncogene in triple-negative breast cancer [43] and B-cell lymphoma [44], that has been described as integral to the pathology of lung squamous carcinoma through its interaction with *SOX2* to control the expression of epigenetic regulators [45].

In summary, Refphase revealed novel insights into the evolution of this tumour. The heterogeneity of the primary samples and continued and parallel evolutionary events would have remained hidden without reference phasing and the results from Refphase.

## Refphase reveals previously undetected events in low purity samples

To quantify the extent to which Refphase reveals previously undetected SCNAs, we applied Refphase to multi-sample cohorts from two highly-cited studies: (1) a multi-sample investigation of primary colorectal adenocarcinoma and adenoma [46] (15 tumours, 140 samples), and (2) a matched primary sample and metastatic sample cohort of various primary cancer types and their brain metastases [47] (84 tumours, 196 samples) (Fig 4). These datasets demonstrate a range of copy number landscapes, purity levels, and data types (SNP array and WES) found in both research and clinical studies.

As already described, Refphase's multi-sample phasing permits the detection of MSAI not otherwise discernible with single-sample data or unphased copy number profiles. Examining our cohort utilising the multi-sample reference phasing of all samples of each tumour, we detected MSAI in 65% of tumours (64/99), affecting up to 34% of the genome (IQR across all tumours [0%,3%]; Fig 4A and Methods). In addition to revealing MSAI, multi-sample phasing can be used to identify previously undetected allelic imbalance. This newly identified allelic imbalance is often not MSAI but simply allelic imbalance with the same major haplotype as the other samples from the same tumour previously identified as demonstrating allelic imbalance. Refphase detected new allelic imbalance not previously detected by ASCAT in 77% of tumour samples (256/334). The greatest degree of new allelic imbalance identified was in sample MET019 in which 66% of the genome was identified to be allelically imbalanced by Refphase which had not been identified as such by ASCAT (Fig 4B).

Whilst we identified a significant increase in total allelic imbalance with increasing sample purity (Fig 4C; LME ANOVA $p = 0.03$, adjusted for patient and cohort—the latter defined by tumour type and profiling platform—as random effects, S3A Fig), we observed a significant negative association between sample purity and newly identified allelic imbalance (Fig 4B and 4C; LME coefficient = -0.28 (2.s.f), LME ANOVA $p<0.0001$, adjusted, S1 and S3C Fig). This result is consistent with Refphase's ability to rescue low purity samples and is in accordance with previous work demonstrating that SCNA detection using methods designed for single

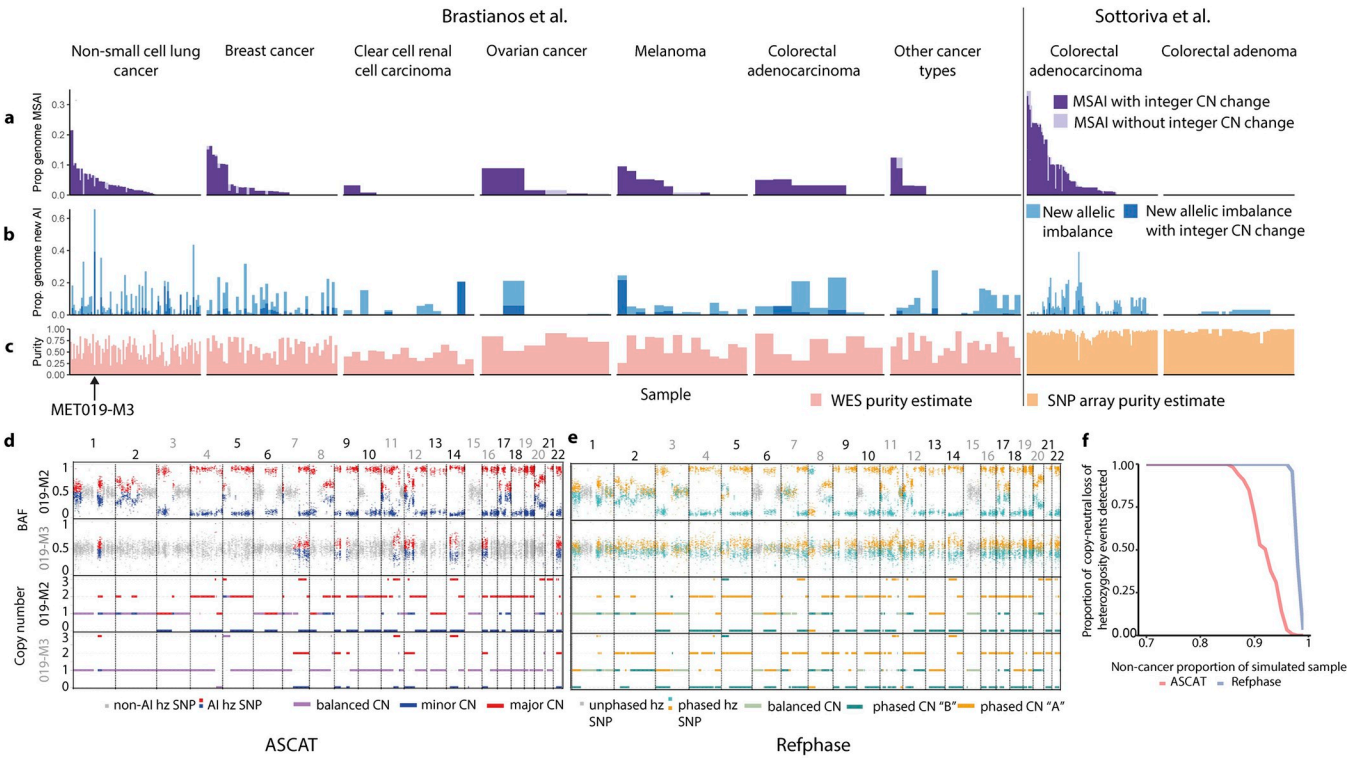


**Fig 4. Low cancer cell fraction SCNA detection and cohort-level analysis for the Sottoriva et al. [46] and Brastianos et al. [47] cohorts. a)** Barplots showing the proportion of the genome affected by MSAI in each tumour sample in the pan-cancer cohort grouped by tumour type. b) Barplots showing the proportion of the genome with allelic imbalance that was identified using multi-sample reference phasing and that was previously undetected using ASCAT. Each tumour sample in the pan-cancer cohort is arranged by tumour type. Light blue bars represent the proportion of the genome in each sample affected by previously undetected allelic imbalance that did not result in an alteration in the previously estimated integer allele-specific copy number. Dark blue bars represent instances in which newly detected allelic imbalance that resulted in new integer copy number state being estimated. c) Barplots representing the estimated cancer cell fraction of each sample in the pan-cancer cohort grouped by cancer type. d) Unphased BAF and integer copy number states across the genome from two samples analysed using ASCAT. e) Phased BAF and haplotype-specific copy number states across the genome from two samples analysed using Refphase. f) Line plot showing the proportion of simulated copy-neutral loss of heterozygosity events identified at differing non-cancer proportions of a sample using ASCAT (red) or Refphase (blue).

https://doi.org/10.1371/journal.pcbi.1011379.g004

samples is drastically impaired by low tumour purities when sequencing coverage remains unchanged [48]. No statistically significant associations were observed between purity and MSAI detection at sample or tumour-level (sample-level: LME ANOVA p = 0.5, adjusted, S3D Fig; tumour-level: p = 0.07, adjusted for cohort, S3E Fig). Schematic examples of the effect of tumour purity and copy numbers states on BAF and LogR profiles are shown in S4 Fig.

We further explored the relationship between purity and SCNA event detection for the specific case of tumour MET019 from Brastianos *et al.*[47], a lung adenocarcinoma containing the sample (M3) which demonstrated the highest amount of newly identified allelic imbalance in our cohort when analysed with Refphase (66% of the genome affected). Notably, this M3 sample showed a markedly lower level of inferred tumour purity compared to the tumour's other samples (M3–21% compared to R1–47%, M1–80%, M2–86%). Using the same phasing that revealed the newly identified allelic imbalance, we re-estimated the copy number states for all samples across the genome and compared Refphase results to the previous ASCAT-derived estimates. We focussed analysis on detection of CNLOH. ASCAT identified CNLOH affecting five chromosomes in M3 compared to Refphase finding CNLOH on 15 chromosomes, with an increase of 24% of the genome affected when quantified using Refphase. The same analysis of the purer M2 sample yielded an increase of just 0.09% of the genome affected by CNLOH

using Refphase (Fig 4D and 4E). This result supports the previous observation of a negative association between sample purity and the extent of newly identified allelic imbalance using Refphase.

Finally, to systematically compare ASCAT and Refphase's ability to resolve allelic imbalance as a function of purity, we simulated BAF values for CNLOH events at varying tumour purities in samples from multi-sample NSCLC tumours [8] (125 events; each simulated at tumour purities from 1% CCF to 30% CCF at 1% CCF intervals; 200x sequencing coverage) (Methods and Fig 4F). CNLOH events were investigated to ensure that there were no changes to overall sample ploidy and that LogR changes did not influence event detection. Using these simulation parameters, ASCAT was able to detect allelic imbalance at all simulated CNLOH events at tumour purities of 15% or greater and Refphase at purities of 4% or greater (Fig 4F).

Taken together, multi-sample reference phasing improves the limits of detecting allelic imbalance and offers potentially exciting avenues for improving the sensitivity of SCNA detection at low cancer cell fractions, especially in non-WGS contexts.

## Refphase improves SCNA intra-tumour heterogeneity estimates

Using the same pan-cancer cohort (Fig 5A), we next leveraged Refphase's relative-to-ploidy SCNA event classification and standardised SCNA intra-tumour heterogeneity quantification (Methods) to explore cancer evolution in a range of cancer types.

We first quantified the total proportion of the genome affected by SCNAs (here encompassing relative-to-ploidy gains and losses, and LOH) and the proportion of clonal, early SCNAs, compared with subclonal, late SCNAs (Methods and Fig 5B). We identified clonal SCNAs in every tumour and found that 96% (95/99) of the tumours examined had clonal and 95% (94/99) harboured subclonal SCNAs affecting at least 1% of the genome. A median of 31% of the genome was subject to clonal SCNAs and 25% to subclonal SCNAs meaning that in over half of tumours (52/99), 25% or more of the genome was subject to subclonal SCNAs. The SCNA heterogeneity observed in these 99 tumours of various cancer types supports recent work suggesting that ongoing chromosomal instability is pervasive in cancer [9,10].

Importantly, detection of MSAI at a genomic segment renders any clonal SCNA at that segment subclonal since different haplotypes represent the major allele in different tumour samples. This means that Refphase can resolve additional SCNA heterogeneity which would not be possible with methods relying on unphased data. Notably however, despite the detection of MSAI in a subset of tumours (Fig 4A), the cohort-level effect of multi-sample reference phasing was to decrease the average level of SCNA heterogeneity compared to estimates from ASCAT runs on independently analysed tumour samples (paired Wilcoxon signed rank tests, $p<0.05$, S5 Fig). This suggests that segments previously defined as harbouring subclonal SCNAs are now identified to be clonally affected across all tumour samples. This may in part be due to the increased SCNA detection in low tumour purity samples explored in the previous section.

Overall, these analyses highlight how Refphase can provide potentially more accurate overall estimation of SCNA heterogeneity than approaches relying on unphased data.

## Refphase offers insight into associations between parallel evolution and WGD in a pan-cancer cohort

Whole genome doubling (WGD) is considered a transformative event in tumour evolution [49], and tumours with WGD events show increased prevalence of MSAI [8,9]. In order to explore this relationship in our pan-cancer cohort, we determined WGD status using MEDICC2 run on Refphase output to classify the tumours in our pan-cancer cohort as clonal

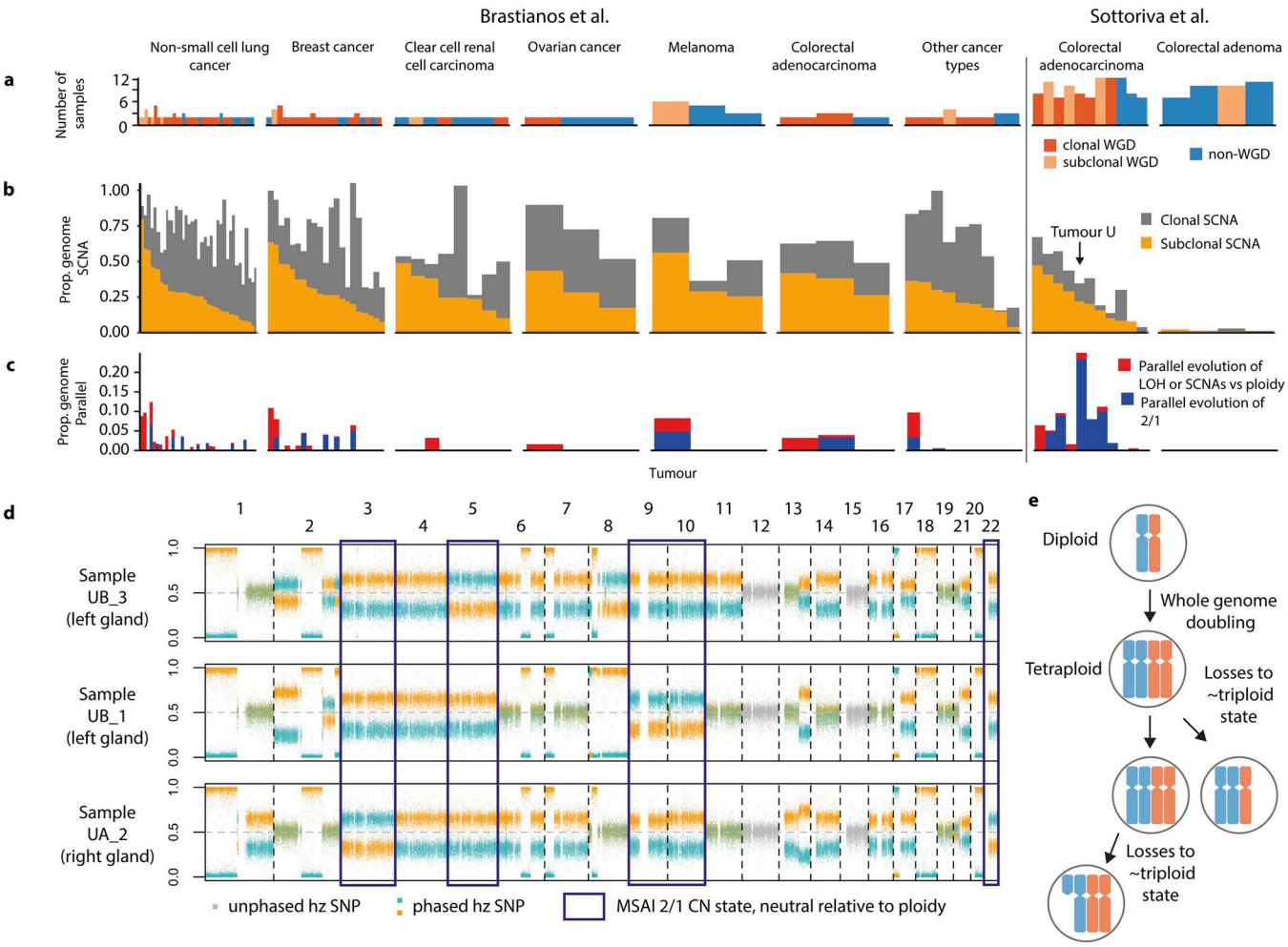


**Fig 5. Cohort-level analysis of cancer evolution for the Sottoriva et al. [46] and Brastianos et al. [47] cohorts. a)** Barplot showing the number of samples per tumour and colored by WGD clonality status with clonal WGD (dark orange), subclonal WGD (light orange), and non-WGD (blue). **b)** Barplot showing proportion of the genome classified as affected by clonal SCNA (grey) and subclonal SCNA (yellow) from the pan-cancer cohort. **c)** Barplot showing proportion of the genome affected by parallel evolution of SCNAs relative to ploidy or LOH (red) and parallel evolution of 2|1 copy number states (dark blue). **d)** BAF of heterozygous SNPs across the genome from tumour samples from colorectal adenocarcinoma U. SNPs are coloured orange and blue according to the phased haplotype they are assigned to while SNPs that could not be phased are coloured grey. Regions of the genome demonstrating parallel evolution of a 2|1 and 1|2 copy number states are highlighted with a dark blue outline. **e)** Schematic demonstrating whole genome doubling and independent subsequent copy number loss events revealed by MSAI.

https://doi.org/10.1371/journal.pcbi.1011379.g005

WGD, subclonal WGD, or non-WGD (Methods and Fig 5A). In keeping with previous results, a higher proportion of the genome was observed to be affected by MSAI in whole genome doubled tumours (clonal and subclonal) than non-WGD tumours (Kruskal-Wallis p = 2e-04, S6 Fig).

We next probed further into the nature of MSAI in our cohort. Specifically, we quantified the extent of parallel evolution of the same type of SCNA event (e.g. gains, losses, LOH) occurring in different samples within the same tumour but affecting different haplotypes, which represents just a subset of MSAI events. Using relative-to-ploidy and LOH definitions of parallel evolution (Methods), we observed parallel evolution in all tumour types in our cohort with the exception of the benign colorectal adenomas (Fig 5C). However, this tumour type also had the lowest levels of SCNAs overall with on average <1% of the genome affected by either clonal or subclonal SCNAs (Fig 5B; median percentage of genome affected: clonal 0.7%, subclonal

0.7%). Notably however, we observed parallel evolution of gains in the Sottoriva *et al.* (malignant) colorectal adenocarcinoma dataset, indicating independent evolution of similar SCNAs in spatially separated areas of tumours, adding additional resolution to the original study [46] (S7 Fig).

Intriguingly, we observed some tumours, such as colorectal adenocarcinoma U from Sottoriva *et al.* [46] (Fig 5D), with greater than 30% of their genomes affected by MSAI but with only a very small subset of this MSAI constituting parallel gains and losses relative to ploidy (2% of the genome, Fig 5C). Upon exploring MSAI in genomic regions matching the overall ploidy of their respective tumour samples (Methods), we observed that MSAI commonly occurred in a copy-neutral context of arm-level and chromosomal triploidy. Tumour U harboured nine chromosome arms affected by this phenomenon and was found to have undergone a clonal and therefore likely relatively early WGD event. This may suggest parallel evolution of losses from a tetraploid to a sub-tetraploid state (Fig 5E). Consistent with this, 2|1 and 1|2 copy number states were simultaneously observed in 22 tumours, of which 21 were determined to have undergone clonal or subclonal WGD using MEDICC2 (Fig 5A and 5C, Methods) [36]. Other groups have observed such parallel losses in *in vitro* models of WGD [50], inferred their presence from single-sample data [51] or characterised them in single cell sequencing [32], but to our knowledge this is the first time that such independent losses have been observed and linked to WGD using phasing in multi-sample bulk data. The parallel evolution of losses from a tetraploid state offers new insights into the potential selective pressure for triploidy and the ability to further determine the timing of such events in tumour evolution.

Together, these results highlight the importance of leveraging information from multiple samples to quantify SCNA heterogeneity during tumour evolution and demonstrate the technical advances offered by Refphase compared to single-sample copy number callers, including its ability to identify MSAI, its increased sensitivity of CNLOH detection, and its potentially more accurate overall estimation of SCNA heterogeneity.

## Benchmarking Refphase

**Comparison to HATCHet and Battenberg.** To further characterise the performance of Refphase we next conducted a systematic comparison of Refphase against two commonly employed copy-number callers, HATCHet [52] and Battenberg [27] to investigate the effect of intra-sample subclonality on Refphase output.

To this end we reanalysed a well-characterised WGS prostate cancer cohort of 44 tumour samples from 9 patients [53]. Both HATCHet and Battenberg are designed to detect subclonal SCNAs owing to within-sample heterogeneity, but do not return phased haplotype-specific copy-numbers (and by extension no MSAI). To facilitate comparison between the three tools, all copy numbers were therefore first brought into a major/minor configuration. We then compared total, major and minor copy numbers of the dominant subclone of each sample detected by HATCHet and Battenberg to the Refphase output for (i) all segments, (ii) clonal segments, and (iii) subclonal segments separately. Subclonal segments here are those where within-sample subclonality was inferred by both HATCHet and Battenberg (15.8% of the genome on average across the cohort).

For on average 82.1% of the genome, all three algorithms agreed in their total copy number estimates with the exception of both samples of tumour A29 and one sample from tumour A31 (A31-C, S10A Fig). In A31-C the HATCHet calls differed most from Battenberg and Refphase, likely due to differences in ploidy estimates owing to a previously identified subclonal WGD event in that tumour [28]. Specifically, since HATCHet cannot assign subclonal

WGDs, it overestimated ploidy in A31-C leading to potentially copy number calls. Similarly for A29 contradictory reports about its WGD status were found in previous analyses [52], which also provided additional evidence of A29 as a complex and difficult to analyse tumour.

In areas of clonal SCNAs, for 90.2% of the genome all three algorithms agreed in their total copy number estimates, closely followed by an agreement between Refphase and Battenberg (but not HATCHet) in 6.9% of the genome across all samples (S10B Fig). Refphase agreed with HATCHet and not with Battenberg in only 1.8% of segments across all samples. These high levels of concordance were also found when investigating major and minor copy numbers separately, both on the level of all segments and clonal segments only (S10D, S10E, S10G and S10H Fig).

In contrast, in segments with subclonal SCNAs all three algorithms agreed on total copy number in only 38.4% of the genome, followed by agreement between only Refphase and HATCHet with a cohort mean of 34.5% (S10C Fig). Investigating major and minor copy number separately we found that while major copy number estimates were highly concordant with full agreement in 68% of the genome (S10F Fig), minor copy number calls demonstrated a marked increase in the mean proportion of genome where Refphase agreed with only HATCHet to 30.1%, and agreement between all algorithms decreasing to a mean of 50.2% (S10I Fig).

In summary, across the majority of the genome in our cohort we see broad concordance between all three algorithms. Excluding sample A31-C, Refphase deviates from estimates produced by either HATCHet or Battenberg in a mean of less than 0.5% of the genome across the cohort for total copy number, major copy number, and minor copy number. In areas of the genome with clonal SCNAs Refphase demonstrates the most concordance with Battenberg and in areas of within-sample subclonality it demonstrates most concordance with HATCHet. This is especially marked in the estimation of the minor copy number in these areas, potentially consistent with Refphase's increased sensitivity for LOH detection. It is also worth noting that, while all comparisons here were performed without phasing information, MSAI has been independently described in the Gundem et al. cohort in a specialised one-off reanalysis for the PCAWG heterogeneity study [28], and which in contrast to Battenberg and HATCHet Refphase is able to detect routinely [53].

**Assessing specificity.** We next systematically assessed Refphase's specificity to identify any potential tendency to overcall allelic imbalance. To this end we generated one additional artificial sample without any allelic imbalance or SCNAs for each of our 84 tumours from the Brastianos *et al*. cohort [47] by combining the germline/normal samples from the Brastianos cohort with normal samples from a different cohort *in-silico* (Methods). Refphase was then rerun on all patients from this cohort and the processed artificial normal samples were investigated for any false positive allelic imbalance calls.

Out of a total of 12778 copy number segments, Refphase detected allelic imbalance in the artificial normal sample in only 71 segments (with one false MSAI call), yielding a segment-level specificity of 99.4%. The 71 segments harboured 153.5 SNPs on average spanning 16.8 Mb, compared to 132.43 SNPs and 17.9 Mb for all segments with allelic imbalance. 21 of these 71 segments originated from the same patient (B_MET065), an ER+ breast cancer. It is conceivable that this allelic imbalance is the result of circulating tumour DNA affecting the artificial normal sample, whose underlying data was derived from blood. This would be consistent with recent results highlighting the use of phasing to identify SCNAs in ctDNA from metastatic samples [54]. A further 24/71 false positive allelic imbalance segments were found to affect chromosome 7 in 23 different tumours, 19 of which overlapped. This suggests a sequencing or pipeline processing issue particular to these tumours as a potential cause.

In summary, across the cohort Refphase showed high specificity with only few false positive calls, of which a genuine biological or technical origin cannot be excluded.

## Discussion

We have shown that Refphase provides phasing of heterozygous SNPs into long-range haplotypes, allows the identification of SCNA-mediated parallel evolution and MSAI [8], and improves the limits of detection of allelic imbalance in low purity samples.

The long-range haplotypes derived by Refphase augment existing haplotype-specific approaches for copy number calling in single-cell DNA [29] and RNA [30,55] sequencing technologies, and are distinct from haploblocks produced by population-level statistical phasing approaches [25,56–58] or those derived from chromatin structure data [59,60] or long read sequencing [61,62]. Reference-phasing haploblocks are not limited in length by recombination rates, read lengths, or structural constraints, and instead stretch the full length of the evolutionary gain or loss event that gave rise to the allelic imbalance, frequently a whole chromosome or chromosome arm. For illustration, benchmarking studies on statistical phasing methods have reported correctly-phased average haploblock lengths from about 450Kbp for datasets of 2,500 samples [63] to 15Mbp for datasets with 400,000 individuals [58]. In contrast, in our cohort with a median of just two samples per patient, the average length of segments with allelic imbalance was 15Mbp covering 63% of the genome. However, due to lack of a reference genome with known haplotype structure and SCNAs, switch error rates for refphase are difficult to estimate, so any direct comparison is difficult. Speculatively, due to the strength of the allelic imbalance signal we expect error rates to be somewhat lower than the 0.5% to 1.5% reported for statistical phasing [58,63]. Refphase haploblocks have the additional advantage that all variant alleles within regions of allelic imbalance with sufficient sequencing depth can be assigned to their haplotype-of-origin. Refphase is therefore able to phase rare and private variants or those from understudied ethnic groups for whom reference sets of haplotypes are unavailable, doing so at a fraction of the computational cost of statistical phasing approaches, and with broad applicability to a variety of different experimental techniques, including WGS, WES, and SNP arrays, and without reliance on external databases.

Besides haplotype reconstruction and phasing of SCNAs, Refphase offers a standardised characterisation and quantification of SCNA intra-tumour heterogeneity from bulk multi-sample tumours, where the field previously relied on simple metrics such as the weighted genome instability index [64] or fraction of genome altered [65]. In the context of multi-sample sequencing, only a few algorithms utilise data from multiple samples for SCNA estimation and either do not produce haplotype-specific copy number estimates [52,66] or use less powerful statistical phasing limited to WGS [67] in concert with reference haplotype databases. Refphase is broadly applicable, supporting multiple formats of user-provided single-sample copy number segmentations as input, including those from commonly used copy number callers such as ASCAT [22], to provide reproducible estimates of SCNA intra-tumour heterogeneity that allow comparisons both between datasets and within grouped sets of samples within a patient's disease, for example to contrast primary samples with metastases. Downstream integration with MEDICC2 [36] also permits standardised detection of WGD.

It is through this joint analysis of samples from a single tumour, utilising WGD detection and Refphase's relative-to-ploidy classifications, that we observe previously undetected copy-neutral MSAI of arm-level and chromosomal triploidies in WGD tumours from multi-sample bulk sequencing. This finding, indicating parallel evolution from a tetraploid to a sub-tetraploid state, may suggest selective pressure for triploidy and offers a new avenue for the exploration of tumour copy number evolution. Additionally, the detection of MSAI and parallel events using reference phasing indicates a hitherto underappreciated number of independently occurring SCNAs. It is worth stressing however, that the presence of parallel evolutionary events does not necessarily imply selective pressures or evolutionary convergence. When

investigating MSAI frequency in our cohort compared to a null model of random events without selection, we only found a few instances of significantly increased MSAI frequency (Methods and S8 Fig).

Despite its many advantages, our method is not without its limitations. Refphase chooses a single best reference sample to inform phasing for any given bin meaning that the same heterozygous SNPs are assayed across multiple samples from the same tumour; however, each sample with allelic imbalance may provide useful phasing information. While Refphase characterises inter-sample heterogeneity by determining the presence and absence of SCNAs in each tumour sample, it does not attempt to identify within-sample subclonal SCNAs present in only a subset of cells in a single sample, in contrast to tools such as TITAN [68] and Battenberg [27] designed for use with WGS. However, in a recent comparison of evolutionary inference from SNV-based subclonal deconvolutions of individual samples to SCNA-based phylogenetic reconstructions without subclonal deconvolutions we demonstrated that phylogenetic reconstructions based only on the dominant clone per sample faithfully recapitulate tumour phylogenies [36]. Refphase also does not attempt to identify subclonal clusters of co-occurring SCNAs present across multiple samples [52]. Additionally, while Refphase updates ploidy estimates of each sample as it performs its phasing, it is reliant on robust initial estimates of purity, ploidy, and input segmentation. Ideally, the three steps of copy-number calling (including purity and ploidy estimation), phasing and tree inference would be done jointly and in an integrated manner for all samples of a tumour. For computational complexity reasons we instead use a more "linear" approach, where only some information, such as about the optimal purity and ploidy values, are fed back into the copy-number calling process in an iterative manner.

The SCNA heterogeneity and parallel evolution revealed and characterised by Refphase is indicative of ongoing chromosomal instability in the tumours examined. However, it should be noted that this is likely still an underestimate of the actual ongoing chromosomal instability present in these tumours, as only a small proportion of each tumour is sequenced [69]. Whilst no significant correlation was observed between the number of samples per tumour and SCNA heterogeneity in this pan-cancer cohort (LME ANOVA $p = 0.1$, adjusted, S9A Fig), further analysis in cohorts with a greater range of per-tumour sample numbers and tumour types is required. Additionally, whilst Refphase's ability to infer SCNA events at low tumour purities exceeds that of single-sample copy number calling methods (Fig 4F), even with Refphase, we observe a moderately significant association between the range of purities within a tumour ('Tumour Purity Difference') and the degree of SCNA subclonality (LME coefficient = 0.28, LME ANOVA $p = 0.04$, adjusted, S9B Fig), indicating that tumour purity may interfere with the estimation of SCNA clonality. Finally, while multi-sample reference phasing reveals instances of parallel evolution from distinct haplotypes, parallel evolution from the same haplotype cannot be detected and as such the amount of parallel evolution found in this cohort may be an underestimate. Detection of such parallel events from the same haplotype instead can be resolved with phylogenetic methods which allow for multiple mutations of the same site, such as MEDICC2 [36].

Despite these limitations, no other tool provides phasing, detection of allelic imbalance in low purity samples, estimation of copy number states and parallel evolution, and systematic characterisation of SCNA heterogeneity.

In the future, combining statistical population-based phasing with multi-sample reference phasing will further strengthen this approach, in particular for genomic regions with weak allelic imbalance. As multi-sample bulk DNA sequencing data of tumours and their metastases become increasingly common, new opportunities for improving our understanding of tumour evolution and how it relates to prognosis and response to treatment, will arise. Algorithms such as Refphase that are able to leverage such data to quantify mutations and their intra-

tumour heterogeneity across a patient's disease will be vital to support new insights to inform care for cancer patients.

## Methods

### Nomenclature

We report integer allele-specific copy numbers from SCNA estimation tools that do not perform either statistical or multi-sample reference phasing in the major/minor configuration. We use the / symbol to separate the most common allele and least common allele for a copy number segment or bin covering a genomic region. For example, in two samples (S1 and S2) from the same tumour in which we observe the same unphased allele-specific integer copy number state of 2 of one allele and 1 of the other allele at the same genomic region, we report the allele-specific integer major and minor allele copy number estimate of this bin as 2/1 in S1 and 2/1 in S2. We would also report the non-allele specific total copy number in sample S1 as 3 and in sample S2 as 3.

In the context of multi-sample reference phasing as performed by Refphase, we report phased haplotype-specific "A" haplotype and "B" haplotype copy number written in the format haplotype "A" | haplotype "B". If Refphase identified MSAI between our example S1 and S2, maintaining the total copy number states of 3 in both, with haplotype "B" being present at 1 copy in S1 but at 2 copies in S2, we would report the copy number in S1 as 2|1 and in S2 as 1|2.

### Refphase input data requirements

To perform multi-sample reference phasing for a tumour with $\boldsymbol{N}$ samples, an initial single-sample copy number segmentation and initial estimates of tumour purity $\rho_j$ and ploidy $\psi_{t,j}$ must be obtained as input for Refphase by applying a single-sample SCNA calling algorithm to each tumour sample $j \in \{1..\boldsymbol{N}\}$ independently.

Single-sample tools typically follow a common approach for the detection of allele-specific SCNAs, as defined in [22]. Briefly, for each tumour sample, sequencing reads are aggregated at heterozygous germline variants in each tumour and paired normal sample over both parental alleles, yielding two readouts: the log ratio $L_i$ of normalised read counts at variant $i$ compared to the matched normal (LogR) and the relative frequency of the minor (B) allele read counts over the total read counts (B-allele frequency, BAF) $b_i$ at variant position $i$. SCNAs are called by segmenting both BAF and LogR tracks into homogeneous segments and determining fractional or integer copy numbers for each segment, jointly inferring the purity $\rho_j$ of sample $j$ (the fraction of cancer cells over total number of cells) and the average ploidy $\psi_{t,j}$ of the tumour sample $j$ as free parameters [22]. Copy numbers are determined per allele, but due to unknown phasing of the underlying germline variants, the values are reported as major (larger) and minor (smaller) copy number instead.

Refphase currently directly supports input derived from the popular segmentation algorithms ASCAT [22] and Sequenza [33], but will run on any user-supplied initial segmentation result which includes estimates of: allele-specific copy number segments (genomic positions and major and minor allele copy number states); sample purity and ploidy estimates; log ratios (LogR) and B-allele frequencies (BAFs) of single nucleotide polymorphisms (SNPs); and—for non-ASCAT input—SNP heterozygosity annotations, for each of the $\boldsymbol{N}$ tumour samples ($\boldsymbol{N} \geq 2$) and a matched normal from the same patient.

### Refphase algorithm overview

Refphase achieves long-range phasing and haplotype-specific estimation of SCNAs through application of the multi-sample reference phasing algorithm. We assume that the input purity

estimates for each tumour sample are correct and Refphase utilises these, alongside the LogR and BAF, to characterise SCNA heterogeneity of the genome in ***m*** bins of variable sizes. These bins are derived from a minimum consistent segmentation (see below) created from the input copy number segments for each tumour sample.

The four outputs from Refphase are: (1) a phasing of heterozygous SNPs, (2) an updated set of phased fractional and integer copy number states across the genome for each sample, (3) a sample-level summary of SCNA events, and (4) a summary of SCNA clonality and intra-tumour heterogeneity, defined either at tumour level or between and across user-defined sub-groups of samples.

For a schematic overview of the Refphase algorithm see Fig 1.

## Minimum consistent segmentation

Having internally preprocessed input data to a standard Refphase format, Refphase combines individual single-sample segmentations to generate a minimum consistent segmentation (MCS) for the set of all ***N*** samples of a tumour (Fig 1A). To do this, Refphase first defines the combined set of breakpoints as the union of the set of individual breakpoints, keeping track of the samples-of-origin. Of the new set of segments defined by this union of breakpoints, only segments present in all samples are retained in the subsequent analysis. An iterative merging strategy is then employed which merges breakpoints that originated from different samples if their pairwise distances are below a user-defined maximum gap threshold (default = 100kbp). These slight variations in breakpoint position typically result from variability in the estimation of breakpoint positions in individual samples even if the true underlying breakpoint is the same. The restriction to only merge breakpoints that originated from different samples of origin preserves focal amplifications and losses present prior to merging in individual samples and yields a final set of **m** bins.

## Reference sample selection

Once the MCS is established, Refphase employs the multi-sample reference phasing algorithm to achieve long-range phasing of germline variants and assignment of SCNAs to haplotypes (Fig 1B and 1E). Multi-sample reference phasing leverages the fact that the phase of the underlying germline genetic variants is constant between samples from the same patient. In a segment affected by allelic imbalance, variants whose alternative alleles are residing on the chromosome with higher copy number will show a theoretical BAF above the segment mean, whereas those residing on the minor copy number chromosome will show BAF values below the segment mean, effectively providing phasing information about the variants contained in the segment.

To leverage this information, for each MCS segment, Refphase first iterates through all segments and samples and, using input major and minor copy numbers, for those segments determined to have a major copy number not equal to half of the total copy number (sum of major and minor copy number) in at least one sample, Refphase assigns the sample with the highest mirrored mean BAF as the reference sample (Fig 1B), where the mirrored mean BAF is defined as:

$$b_{mirr} = \frac{1}{m}\sum_{k=1}^{m} abs(b_k - 0.5) \quad (1)$$

where $b_k$ is the BAF at variant position $k$; $m$ is the total number of variants within the segment indexed from 1; where *abs()* returns the absolute value ($\geq$0) of the expression contained in brackets; and where variant refers to a heterozygous SNP contained within the segment.

Specifically, mirrored mean BAF is calculated directly on all heterozygous SNPs for samples containing fewer than a default of 3 heterozygous SNPs or if the variance of the BAF values for the heterozygous SNPs within a segment is less than a default threshold of 1e-05, and a McClust mixture model applied otherwise.

Segments without copy number imbalance in at least one sample in the input are not considered for reference phasing.

## Haplotype phasing of reference sample segments

Within each reference sample, Refphase then assigns alternative alleles to haplotypes $H_i$ by comparing the empirical BAF of each variant $i$ against the segment mean (Fig 1C):

$$H_i = I\left[b_i > \frac{1}{m}\sum_{k=1}^{m} b_k\right] \quad (2)$$

where $b_i$ is the BAF at variant position $i$; $m$ is the total number of variants within the segment indexed from 1; and $I$ is an operator assigning each variant position to one of two haplotypes based on the evaluation of the logical expression in square brackets.

In this way, the definition of haplotypes is entirely based on germline SNP data and does not take into account somatic mutation data such that exactly two haplotypes are defined at each genome position for which phasing is undertaken.

## Reference phasing across non-reference sample segments

In the next step, Refphase applies this phasing information to the variants in the same segment in all other non-reference samples to determine haplotype-specific BAF values for every other sample (Fig 1D).

## Haplotype-specific copy number quantification

After the variants have been assigned to haplotypes, Refphase uses haplotype-level BAF and LogR values for re-estimation of haplotype-specific copy numbers (Fig 1E).

Here, each sample is tested for allelic imbalance using a Wilcoxon rank-sum test between the BAF values of each haplotype (5% family-wise error rate) and the effect size of allelic imbalance is determined using *Cohen's d* [70]. If additional allelic imbalance is detected compared to the initial copy number states, or if the option is applied universally by the user, copy numbers are re-estimated using either a default parametric or a non-parametric model for each haplotype separately.

The parametric model is consistent with the ASCAT formulation for copy number quantification and can accurately describe theoretical LogR and BAF values from a sample with given purity, ploidy and sequencing depth. The parametric model estimates the new haplotype-specific copy numbers $n_{A,i}$ and $n_{B,i}$ in segment $i$ based on the mean BAF $b_i$ and LogR $L_i$ of that segment as well as sample purity $\rho$ and average tumour ploidy $\psi_t$ obtained from the initial segmentation as follows, as also described in [22]:

$$n_{A,i} = (\rho - 1 + (1 - b_i)2^{L_i}(2(1-\rho) + \rho\psi_t))/\rho \quad (3)$$

$$n_{B,i} = (\rho - 1 + b_i\, 2^{L_i}(2(1-\rho) + \rho\psi_t))/\rho \quad (4)$$

A full derivation of these equations can be found in the Supporting Information of the ASCAT publication [22].

The non-parametric model, in contrast, estimates new copy numbers for a region agnostic of how the LogR and BAF values were calculated by training a naive Bayes classifier on other segments. The non-parametric model was included as an alternative approach to facilitate application of Refphase to the output of copy number calling tools with potentially greatly different mathematical formulations of LogR and BAF. To train the non-parametric model, Refphase uses the mean BAF and LogR values of all other segments of the same sample and from the same segment in all other samples as training data to account for the typical BAF/LogR distributions of the current sample as well as those values from homologous segments in other samples. After re-estimation of all copy number segments, the average tumour ploidy $\psi_t$ of each sample is re-calculated as the sum over the total copy numbers of all segments $i = 1..\boldsymbol{m}$ weighted by their genomic width $w_i$ (in bp) relative to the total width of the genome (in bp).

$$\psi_t = \sum_{i=1}^{m} (n_{A,i} + n_{B,i}) w_i / \sum_{i=1}^{m} w_i \tag{5}$$

Re-estimated fractional and integer copy numbers are available to the user.

## Horizontal phasing

After individual segments have been phased across all samples, there is still no phasing relationship between neighbouring segments 'horizontally' along the genome (Figs 1E and S1). We employ a method for phasing along the genome that uses a parsimony assumption, i.e. the data can be explained with the least amount of copy number changes between neighbouring segments. This is equivalent to the least amount of copy-number breakpoints which we formulate as the total Hamming distance between neighbouring segments

$$D = \sum_{s=2}^{S} D_s = \sum_{s=2}^{S} \sum_{i=1}^{N} H(n^s{}_{A,i}, n^s{}_{A,i-1}) + H(n^s{}_{B,i}, n^s{}_{B,i-1}) \tag{6}$$

where we sum over all $\boldsymbol{N}$ samples and **S** segments and calculate the Hamming distance $H$ between the copy number and its predecessor for both haplotypes $A$ and $B$. The Hamming distance $H(x, y)$ computes to 1 if $x \neq y$ and 0 otherwise. As the reference phasing algorithm only phases segments across samples, we can achieve an optimal horizontal phasing by "flipping" all haplotype assignments for some of the segments. Flipping the haplotypes of segment $i$ amounts to exchanging $n^s{}_{A,i}$ and $n^s{}_{B,i}$ for all samples $s$. As there are a total of $2^S$ possible configurations, the horizontal phasing problem can be formulated as follows: *Choose for each segment s in S its orientation (flipped or unflipped) such that the overall Hamming distance between neighbouring segments (Eq. 6) is minimal.*

To identify the optimal horizontal phasing a dynamic programming approach can be conceived. However, many of the $2^S$ possible configurations are redundant w.r.t. the overall Hamming distance because of the following two equalities
$D_s(i\ unflipped, i-1\ unflipped) = D_s(i\ flipped, i-1\ flipped)$ and
$D_s(i\ unflipped, i-1\ flipped) = D_s(i\ flipped, i-1\ unflipped)$. Therefore, we only have to investigate whether or not to flip a segment w.r.t. its direct predecessor. If flipping the haplotype assignment of segment $i$ does not result in a change of distance $D_s$ (i.e. $D_s = D_s^{flipped}$), the segment $i$ is compared to its next predecessor (segment *i-2*). This is repeated until $D_s \neq D_s^{flipped}$ or the beginning of the current chromosome is reached.

After the optimal flipping assignments of all segments w.r.t. their direct predecessors are found, we translate those to a final flipping assignment of the segments 1-S by keeping the first segment fixed and iteratively (from second to last segment) perform the following action: If the segment-boundary between segment i and i-1 is set to be flipped, flip all segments [i, i+1, . . ., S].

As the assignment to haplotype A and B is arbitrary across chromosome boundaries, for plotting purposes, we flip the A and B allele assignment for all samples on a chromosome-by-chromosome level to ensure that on average haplotype A has a higher copy number than haplotype B.

## Sample-level SCNA calling

Following reference phasing, Refphase uses the re-estimated sample ploidies, input purities and segment-level LogR data to call SCNA events in each sample (Fig 1F).

We consider an SCNA to be a deviation of any length from the diploid major/minor copy number state of (1,1). SCNAs with an unequal number of copies on both parental alleles (corresponding to a significant deviation of the BAF from its balanced value of 0.5) are termed allelic imbalance.

Specifically, Refphase calls segment *amplifications*, *gains* and *losses* relative to ploidy as well as *LOH* events and *homozygous deletions*. Events are called for each segment and sample independently of others by comparing the mean segment LogR distributions to calculated purity-ploidy derived event thresholds (Eqs 7–9).

The >2× ploidy threshold is the same threshold used for clinical decision making in HER2 + breast cancer using fluorescence in situ hybridization samples [39].

Amplifications are called if

$$LogR > log_2\left[\frac{2(1-\rho)+2\rho\psi_t}{2(1-\rho)+\rho\psi_t}\right], \tag{7}$$

gains are called if

$$LogR > log_2\left[\frac{2(1-\rho)+1.25\rho\psi_t}{2(1-\rho)+\rho\psi_t}\right], \tag{8}$$

losses are called if

$$LogR < log_2\left[\frac{2(1-\rho)+0.75\rho\psi_t}{2(1-\rho)+\rho\psi_t}\right]) \tag{9}$$

with sample purity $\rho$ and sample ploidy $\psi_t$.

By default, Refphase also provides additional versions of sample-level SCNA event calls aside from comparing mean segment LogR values to the calculated thresholds (Eqs 7–9). In another version, LogR values within a segment are compared to the calculated thresholds (Eqs 7–9) using a one-tailed Student's t-test, as in [9]. Additionally, a diploid reference can also be used for both mean and t-test approaches rather than purity-ploidy derived event thresholds. All versions of SCNA calls are available to the user in summary Refphase objects after the completion of the core Refphase algorithm, and plotting options are available to visualise the preferred SCNA event output graphically. For Figs 2 and 3 and associated proportion-of-genome metrics, mean segment LogR values were compared to purity-ploidy derived event thresholds.

LOH events are called for segments in which the rounded copy number state of one allele is 0 and the other strictly greater than 0. Homozygous deletion events are called for those segments where both alleles within a sample have rounded copy number state of 0. Copy neutral LOH (CNLOH) is called for segments in which the major rounded integer copy number state equals the rounded tumour sample ploidy.

## Patient-level SCNA calling

Finally, a tumour-level event summary is calculated (Fig 1G).

Inter-sample heterogeneity is quantified. The presence or absence of each class of relative-to-ploidy event, loss of heterozygosity (LOH), and homozygous deletions (HDs) in each tumour sample for each minimum consistent segment is noted. This presence or absence classification is then examined in the context of MSAI detection to determine whether each event affects the same allele in all samples.

For a given segment, an SCNA is considered to be clonal if it is present in every sample and affects the same allele in all samples. An SCNA is assigned as subclonal if it is present in at least one sample but simultaneously absent in at least one other sample (inter-sample subclonality). Crucially, in cases in which the same relative-to-ploidy or LOH event type is determined to occur in every sample of a given tumour but in the context of MSAI such that a different allele is deemed to be the major allele in different samples at the same segment, the event is assigned as subclonal and parallel and not as clonal, since the SCNAs in different samples are deemed to be of different origin and affecting different haplotypes. Specifically, a mirroring of alleles must be observed between the specific samples harbouring the event type of interest for the event to be called parallel. For example, in a tumour composed of three samples, should a gain SCNA be called in Samples 1 and 2 only, the major alleles must differ in Samples 1 and 2 for the gain event—already subclonal—to be called parallel, regardless of allelic arrangement in Sample 3.

Having assigned individual SCNAs as clonal or subclonal, the proportion of the genome affected by each specific event type (e.g. clonal relative-to-ploidy gains, parallel LOH, etc.) is calculated. Proportion of genome measures are calculated over the sum of minimum consistent segment widths for a tumour.

Plotting options are available to visualise tumour-level summary metrics and the degree of reassignment of events from clonal to subclonal based on MSAI context.

## Refphase user-defined grouping functionality

Refphase also offers the option for a user to define subgroups of samples and calculate group-level (instead of whole-tumour-level) summary metrics and create within- and between-group summary plots. Examples of primary-sample-specific and metastases-sample-specific summary tracks in which clonality and heterogeneity analyses are restricted to the respective subgroups of samples are shown in Fig 3.

## Refphase plotting functionality

Figs 2, 3, and S2 showcase selected tracks from Refphase across-genome plots. Refphase provides several optional output plots as standard including across-genome, chromosome-level, user-defined sample subgroup-oriented, and tumour-level event summary plots.

## MEDICC2 implementation and whole genome doubling detection

The phylogeny showcased in Fig 3 was generated by applying MEDICC2 (version 0.6b1) [36] to re-estimated integer copy number states derived from reference phasing of the named 10 input samples for CRUK0063 using default parameters.

In order to derive the WGD status for the Brastianos *et al.* [47] and Sottoriva *et al.* [46] datasets, MEDICC2 bootstrapping WGD detection was used. In short, we create three sets of 100 bootstrapping datasets for each individual tumour sample (three representing three small segment filter sizes - 500kbp, 750kbp and 1Mbp) by resampling Refphase copy-number output

for the sample chromosome-wise (i.e. drawing 22 chromosomes with replacement). MEDICC2 is then run on each bootstrapped dataset, before checking whether MEDICC2 detects a WGD. Specifically, if at least 5% of the bootstrap runs detect a WGD for at least one of the small segment size filters, the sample is labelled as WGD-positive. This low threshold offsets the otherwise conservative WGD detection, as described in Kaufmann et al. 2022 [36]. Subsequently, if all samples for an individual tumour are determined to harbour WGD, the tumour is labelled as having clonal WGD and subclonal WGD if only a subset of samples have WGD detected.

## Definitions of newly identified allelic imbalance and CNLOH (Fig 4)

ASCAT [22] allelic imbalance was defined at the sample-level and assigned for segments where ASCAT non-integer copy number was not equal for the two alleles. Refphase allelic imbalance was assigned for segments at sample-level using the previously described methods (Methods—Haplotype-specific copy number quantification). CNLOH was assigned to segments in which the rounded major copy number state equalled the rounded sample ploidy and the rounded minor copy number state equalled 0, using copy number states and ploidies from ASCAT and Refphase accordingly. Newly identified allelic imbalance and CNLOH specifically referred to scenarios in which the respective event had not been identified in ASCAT and was identified using Refphase. For cases in which multiple ASCAT segments overlapped with the Refphase segment being assessed for allelic imbalance or CNLOH, data for the ASCAT segment most overlapping the Refphase segment under investigation was used.

## Copy-neutral LOH event simulation (Fig 4)

CNLOH events were simulated in chromosomal segments from NSCLC multi-sample bulk sequencing from the TRACERx100 cohort [8].

First, Refphase was run on the TRACERx100 tumour samples. Then, pairs of samples were selected from tumours where at the same genomic region, defined by a Refphase bin, one sample, referred to as "reference", demonstrated allelic imbalance allowing a reference phasing to be obtained and the other sample, referred to as "test" demonstrated total copy number equal to the overall sample ploidy with no allelic imbalance. Specifically, for a genomic bin, the reference sample demonstrated sufficiently clear allelic imbalance in its BAF to be likely to produce a highly accurate phasing if the heuristic condition in Eq 10 was satisfied:

$$\rho * (|n_A - n_B|) \geq 0.5 \quad (10)$$

where ρ is the tumour purity, $n_A$ the copy number state of allele A and $n_B$ the copy number state of allele B.

Additionally, for candidate samples and segments to be chosen, the following conditions must also be satisfied: Genomic bins had to be $\geq$1/5 of the size of the chromosome on which they were located and contain $\geq$1/5 of the heterozygous SNPs present on the same chromosome; the ploidy of the test sample had to be between 1.8 and 2.2 or between 3.8 and 4.2 (lower and upper bounds inclusive); and the total rounded copy number in the genomic bin should equal the total rounded ploidy of the test sample.

This candidate segment selection approach produced 125 segments that were then used to simulate CNLOH at the genomic region in the test sample. CNLOH was simulated at various cancer cell fractions (CCFs) of the sequenced sample, ranging from 1% to 30% in steps of 1%. Specifically, BAF values were assigned at each of the heterozygous SNP positions in candidate segments using a binomial distribution (R rbinom function) with simulated probability equal to the mean BAF which would be predicted for each phased allele at the segment based on

copy number and simulated purity and with the number of trials equalling a simulated sequencing coverage of 200(x).

These simulated BAF profiles are then used as input to ASCAT allele specific piecewise constant fitting segmentation and Refphase, making use of the phasing derived reference sample, and detection of allelic imbalance reported for each method independently for each of the 125 candidate segments at each simulated CCF value. Detection of allelic imbalance was reported as detection of a CNLOH event.

## Initial copy number estimates for data types

**SNP array.** Tumour cellularity and ploidy for each sample assayed with SNP arrays were estimated using the ASCAT algorithm [22]. ASCAT was then used to identify SCNAs that were provided as the initial input for further clonality analysis. One dataset included in our cohort consists of SNP array data [46]. Processed LogR and BAF values as generated in the original papers were obtained from the GEO database. Copy number analysis was performed using ASCAT v2.3 using default parameters set to 1 for sequencing data [22].

**Whole exome sequencing.** All datasets downloaded were processed from FASTQ files using a previously described pipeline [8,9]. Tumour cellularity and ploidy for each sample assayed with exome sequencing were estimated using ASCAT [22] and these estimates as well as the copy number segmentation were taken forward for analysis with our multi-sample SCNA clonality approach.

## Simulations to estimate MSAI selection

To test whether MSAIs were more frequently found than expected by chance, we created a simulation approach (S8 Fig). To this end, we ran MEDICC2 with the "—events" flag which reconstructs the phylogenetic tree and also infers the copy-number events that lead to the observed profiles for every branch of the phylogenetic tree. Events are defined by MEDICC2 as WGDs, chromosomal gains and losses as well as segmental gains and losses. We then created 350 datasets for each of the 99 patients by re-simulating the events such that every branch of the phylogenetic tree maintained the same number and kind of events but the location of the events were randomly chosen. That means that chromosomal events are randomly assigned to chromosomes with equal probability. Similarly, segmental events are uniformly distributed across all possible chromosomes and uniformly along the chromosome while keeping their lengths fixed. This effectively creates a copy-number profile with the exact same copy-number events (number and type) but with permuted event locations. MSAI enrichment was assessed based on the fraction of the total genome which exhibits MSAIs.

## Benchmarking

**Refphase comparisons with HATCHet and Battenberg.** To enable comparison with Battenberg and HATCHet, we removed Refphase's haplotype-specificity by simply designating the major allele copy number and minor allele copy number for each segment in Refphase output as the higher and lower copy number respectively, on a per sample basis. This masks mirrored subclonal allelic imbalance detection but leaves total copy number estimates unchanged.

For the purposes of benchmarking we used only a single full genome-wide copy-number profile from each algorithm for each sample. For Refphase, this was its standard output modified as described above to remove mirrored subclonal allelic imbalance detection. However, for HATCHet and Battenberg, the genome-wide copy-number profile represented only the dominant subclone per sample comprising both clonal SCNAs and those subclonal SCNAs

present in the highest fraction of tumour cells within that sample. We then compared the intersection of copy number segments from all three algorithms.

**Artificial normal samples.** To generate artificial normal samples, we used the BAF values from the matched normal sample of the Brastianos tumour under investigation. The Brastianos samples were sequenced to median average depth of 108.3(x). These BAF values were previously not being used for any calculation, since the matched normal sample is only used as a read depth reference to compute logR values. To generate a new logR track for these samples, we utilised two distinct sequenced normal tissue samples from the same patient from the TRACERx-421 cohort [71]. These two sequenced normal samples were then analysed with the same bioinformatics pipeline as the Brastianos et al. cohort. One of the normal samples was randomly designated as the "normal/germline" sample and the other as a "tumour" sample in order to generate read depth ratio (LogR) values. These LogR values were then assigned to the nearest heterozygous SNP position values from each Brastianos et al. patient's germline sample.

## Supporting information

**S1 Fig. Example for the horizontal phasing algorithm.** The reference phasing algorithm phases individual segments across all samples but not segments with respect to each other. Therefore we employ a horizontal phasing algorithm which minimises the total number of copy-number breakpoints defined as the total Hamming distance between neighbouring segments (Eq 6). To this end we compare the Hamming distance between neighbouring segments with a flipped version of those segments (flipping a segment amounts to exchanging the haplotype assignment of that segment for all samples) to determine which segments need to be flipped.
(PDF)

**S2 Fig. Full Refphase across-genome plotting output for CRUK0034.** The upper three panels show tracks for log read-depth ratio (LogR), B-Allele Frequency (BAF), re-estimated fractional copy number states, and somatic copy number aberration (SCNA) event calling at a sample level. The bottom panel ('Summary') gives a tumour-level summary of SCNA event clonality and detection of mirrored subclonal imbalance (MSAI), loss of heterozygosity (LOH) and parallel events. Events are called relative-to-ploidy using a mean logR threshold.
(PDF)

**S3 Fig. Associations with sample purity.** The association between sample purity and proportion of genome with **(a)** allelic imbalance (LME coefficient = 0.10 (2.s.f.), LME ANOVA p = 0.03), **(b)** newly identified allelic imbalance (LME coefficient = -0.28 (2.s.f.), LME ANOVA $p<0.0001$), **(c)** allelic imbalance, as defined by Refphase and using original estimates from ASCAT ('Original Total'), **(d)** MSAI (LME coefficient = 0.0045 (2.s.f.), LME ANOVA p = 0.5). (e) The association between mean tumour purity and proportion of genome with MSAI (LME coefficient = 0.070 (2.s.f.), LME ANOVA p = 0.07). Proportion of genome data is calculated using Refphase. Analyses are undertaken for the 336 tumour samples from 99 tumours in the pan-cancer cohort described in Fig 4 and summarised in S1 Table. Linear mixed effect (LME) coefficients and ANOVA p-values shown are adjusted for patient and study cohort (defined by tumour type and profiling platform) as random effects for sample-level analyses and study cohort only for tumour-level analysis, calculated using the nlme R package and maximum likelihood method. Best fit lines shown are derived using the LME model coefficient and intercept values.
(PDF)

**S4 Fig. The effects of tumour purity and copy number states on B-Allele Frequency (BAF) and log read-depth ratio (LogR) profiles. a)** An example of the effect of decreasing tumour purity on BAF band separation. **b)** The effect of varying copy number states on BAF band separation for a tumour of fixed purity, here 50%. **c)** Example of typical BAF and LogR profiles for a loss of heterozygosity (LOH) with total copy number loss (left) and copy-neutral LOH (CNLOH) (right). Relative-to-ploidy thresholds are indicated by dashed lines and are derived for a diploid 100% pure tumour using the formulae described in the Methods of this manuscript. Orange and blue points throughout plots represent the phased "A" haplotype and "B" haplotype respectively.
(PDF)

**S5 Fig. Somatic Copy Number Aberration (SCNA) heterogeneity comparisons between ASCAT and Refphase. a)** Proportion of the genome subject to subclonal SCNAs (ASCAT median = 0.28, Refphase median = 0.25, p = 3.2e-08). **b)** Proportion of aberrant genome subject to subclonal SCNAs, where aberrant genome is defined as the total length of genomic segments in which any SCNA event (clonal or subclonal) is called (ASCAT median = 0.48, Refphase median = 0.43, p = 8.5e-08). **c)** Proportion of genome subject to clonal loss of heterozygosity (LOH) (ASCAT median = 0.18, Refphase median = 0.22, p = 5.8e-13). **d)** Proportion of the genome subject to subclonal LOH (ASCAT median = 0.11, Refphase median = 0.07, p = 1.0e-13). All p-values shown are for paired Wilcoxon signed rank tests with continuity correction, across the n = 99 tumours described in Figs 4 and 5. SCNAs in (a) and (b) encompass relative-to-ploidy gains and losses, and LOH events.
(PDF)

**S6 Fig. Association between whole genome doubling (WGD) and mirrored subclonal allelic imbalance (MSAI).** Median proportions by Patient WGD Status (cWGD = 0.02, n = 54 tumours; sWGD = 0.07, n = 13 tumours; nWGD = 0, n = 32 tumours). Kruskal-Wallis p-value is shown (p = 2e-04). Data is shown for n = 99 tumours from the pan-cancer cohort showcased in Fig 5 for which MEDICC2 was used to infer WGD status. Proportion of genome data is assessed by Refphase. cWGD = clonal WGD; sWGD = subclonal WGD; nWGD = non-WGD.
(PDF)

**S7 Fig. Examples of parallel SCNA events in the Sottoriva *et al* colorectal adenocarcinoma cohort. a)** Examples of parallel gain events on chromosome arm 7p, observed between glands on different sides of the tumour. **b)** Example of a parallel gain event on chromosome 2, observed within glands on the same side of the tumour. Tumour IDs shown match those in the original publication [46].
(PDF)

**S8 Fig. Proportion of MSAI segments for the Brastianos and Sottoriva datasets versus simulated data.** We ran MEDICC2 to infer the mutational events that occurred for each sample of the Brastianos and Sottoriva datasets, and used these events as input to a simple karyotype simulator. The simulation uniformly chooses a haplotype to apply the event to, and uniformly samples the starting locus for segmental events. The available phylogeny is then simulated, and the proportion of MSAI segments calculated (sum of MSAI segment lengths divided by sum of all segment lengths). Performing 350 simulations per sample, we generate the above violin plots in the same style as Fig 5C. The star data-point is then the sample's actual/observed MSAI proportion. The corrected p-value of the one-sided right-tailed test is generated, and statistically significant observations are highlighted in blue.
(PDF)

**S9 Fig. The relationship between SCNA intra-tumour heterogeneity and clinical variables.** Association between proportion of aberrant genome with **a)** number of samples (LME coefficient = 0.017, LME ANOVA p = 0.1), and **b)** tumour purity difference—defined as the purity difference between the most and least pure sample within a tumour (LME coefficient = 0.28, LME ANOVA p = 0.04). Proportion of aberrant genome is defined as the proportion of the total length of genomic segments harbouring any relative-to-ploidy SCNA (gain or loss) or loss of heterozygosity (LOH) event which contains a subclonal SCNA or LOH event. Proportion of the aberrant genome is defined at the tumour level. Analyses are carried out for the pan-cancer cohort described in Figs 4 and 5 (n = 99 tumours). Linear mixed effect (LME) ANOVA p-values and LME coefficients calculated using the nlme R package are shown, with analyses adjusted for study cohort (defined by histology and sequencing platform), indicated by the colour legend. Whole-exome sequencing (WES) data is taken from the Brastianos *et al.* study [47]; SNP array data is taken from the Sottoriva *et al.* study [46]. Best fit lines shown have LME slope and intercept values.
(PDF)

**S10 Fig. Refphase, Battenberg, and HATCHet comparisons.** Barcharts for each sample in the Gundem et al. [53] WGS cohort showing concordance and discordance between Refphase derived integer copy number and dominant subclone number from Battenberg and HATCHet showing: **a)** total copy number across all segments **b)** total copy number across all segments without within-sample subclonal copy number called by both Battenberg and HATCHet **c)** total copy number across all segments that demonstrate within-sample subclonal copy number called by both Battenberg and HATCHet.**d)** Major copy number across all segments **e)** Major copy number across all segments without within-sample subclonal copy number called by both Battenberg and HATCHet **f)** Major copy number across all segments that demonstrate within-sample subclonal copy number called by both Battenberg and HATCHet. **g)** Minor copy number across all segments **h)** Minor copy number across all segments without within-sample subclonal copy number called by both Battenberg and HATCHet **i)** Minor copy number across all segments that demonstrate within-sample subclonal copy number called by both Battenberg and HATCHet.
(PDF)

**S1 Table. Cohort Overview for Tumours Analysed in Main Figures.**
(XLSX)

## Acknowledgments

TLK and RFS kindly thank BIFOLD and Klaus-Robert Müller for support. The authors would like to thank Julia Markowski for her expertise and discussions on statistical phasing algorithms.

## Author Contributions

**Conceptualization:** Thomas B. K. Watkins, Peter Van Loo, Nicholas McGranahan, Roland F. Schwarz.

**Data curation:** Thomas B. K. Watkins, Emma C. Colliver, Tom L. Kaufmann, Emilia L. Lim, Cody B. Duncan.

**Formal analysis:** Thomas B. K. Watkins, Tom L. Kaufmann, Emilia L. Lim, Cody B. Duncan.

**Funding acquisition:** Charles Swanton, Nicholas McGranahan, Roland F. Schwarz.

**Methodology:** Thomas B. K. Watkins, Emma C. Colliver, Matthew R. Huska, Tom L. Kaufmann, Cody B. Duncan, Kerstin Haase, Peter Van Loo, Nicholas McGranahan, Roland F. Schwarz.

**Software:** Thomas B. K. Watkins, Emma C. Colliver, Matthew R. Huska, Tom L. Kaufmann, Cody B. Duncan, Nicholas McGranahan, Roland F. Schwarz.

**Supervision:** Charles Swanton, Nicholas McGranahan, Roland F. Schwarz.

**Writing – original draft:** Thomas B. K. Watkins, Emma C. Colliver, Tom L. Kaufmann, Cody B. Duncan, Kerstin Haase, Peter Van Loo, Charles Swanton, Nicholas McGranahan, Roland F. Schwarz.

**Writing – review & editing:** Thomas B. K. Watkins, Emma C. Colliver, Tom L. Kaufmann, Cody B. Duncan, Kerstin Haase, Peter Van Loo, Charles Swanton, Nicholas McGranahan, Roland F. Schwarz.

## References


1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194: 23–28. https://doi.org/10.1126/science.959840 PMID: 959840
2. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nat Rev Genet. 2020; 21: 44–62. https://doi.org/10.1038/s41576-019-0171-x PMID: 31548659
3. Vasudevan A, Schukken KM, Sausville EL, Girish V, Adebambo OA, Sheltzer JM. Aneuploidy as a promoter and suppressor of malignant growth. Nat Rev Cancer. 2021; 21: 89–103. https://doi.org/10.1038/s41568-020-00321-1 PMID: 33432169
4. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366: 883–892. https://doi.org/10.1056/NEJMoa1113205 PMID: 22397650
5. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017; 168: 613–628. https://doi.org/10.1016/j.cell.2017.01.018 PMID: 28187284
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. https://doi.org/10.1371/journal.pmed.1001789 PMID: 25710373
7. Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DCJ, de Jong TV, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 2016; 17: 115. https://doi.org/10.1186/s13059-016-0971-7 PMID: 27246460
8. 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. https://doi.org/10.1056/NEJMoa1616288 PMID: 28445112
9. 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: 126–132. https://doi.org/10.1038/s41586-020-2698-6 PMID: 32879494
10. 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: 302–308. https://doi.org/10.1038/s41586-021-03357-x PMID: 33762732
11. Salehi S, Kabeer F, Ceglia N, Andronescu M, Williams MJ, Campbell KR, et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. Nature. 2021; 595: 585–590. https://doi.org/10.1038/s41586-021-03648-3 PMID: 34163070
12. Schmelz K, Toedling J, Huska M, Cwikla MC, Kruetzfeldt L-M, Proba J, et al. Spatial and temporal intratumour heterogeneity has potential consequences for single biopsy-based neuroblastoma treatment decisions. Nat Commun. 2021; 12: 6804. https://doi.org/10.1038/s41467-021-26870-z PMID: 34815394
13. 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. https://doi.org/10.1016/j.cell.2018.03.043 PMID: 29656894
14. Stopsack KH, Whittaker CA, Gerke TA, Loda M, Kantoff PW, Mucci LA, et al. Aneuploidy drives lethal progression in prostate cancer. Proc Natl Acad Sci U S A. 2019; 116: 11390–11395. https://doi.org/10.1073/pnas.1902645116 PMID: 31085648

15. van Dijk E, van den Bosch T, Lenos KJ, El Makrini K, Nijman LE, van Essen HFB, et al. Chromosomal copy number heterogeneity predicts survival rates across cancers. Nat Commun. 2021; 12: 3188. https://doi.org/10.1038/s41467-021-23384-6 PMID: 34045449

16. Lahoz S, Archilla I, Asensio E, Hernández-Illán E, Ferrer Q, López-Prades S, et al. Copy-number intratumor heterogeneity increases the risk of relapse in chemotherapy-naive stage II colon cancer. J Pathol. 2022. https://doi.org/10.1002/path.5870 PMID: 35066875

17. Lukow DA, Sheltzer JM. Chromosomal instability and aneuploidy as causes of cancer drug resistance. Trends Cancer Res. 2022; 8: 43–53. https://doi.org/10.1016/j.trecan.2021.09.002 PMID: 34593353

18. Bakhoum SF, Ngo B, Laughney AM, Cavallo J-A, Murphy CJ, Ly P, et al. Chromosomal instability drives metastasis through a cytosolic DNA response. Nature. 2018; 553: 467–472. https://doi.org/10.1038/nature25432 PMID: 29342134

19. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019; 575: 210–216. https://doi.org/10.1038/s41586-019-1689-y PMID: 31645765

20. Shih DJH, Nayyar N, Bihun I, Dagogo-Jack I, Gill CM, Aquilanti E, et al. Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. Nat Genet. 2020; 52: 371–377. https://doi.org/10.1038/s41588-020-0592-7 PMID: 32203465

21. Nguyen B, Fong C, Luthra A, Smith SA, DiNatale RG, Nandakumar S, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell. 2022; 185: 563–575.e11. https://doi.org/10.1016/j.cell.2022.01.003 PMID: 35120664

22. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010; 107: 16910–16915. https://doi.org/10.1073/pnas.1009843107 PMID: 20837533

23. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015; 6: 8971. https://doi.org/10.1038/ncomms9971 PMID: 26634437

24. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061–1073. https://doi.org/10.1038/nature09534 PMID: 20981092

25. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81: 1084–1097. https://doi.org/10.1086/521987 PMID: 17924348

26. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011; 9: 179–181. https://doi.org/10.1038/nmeth.1785 PMID: 22138821

27. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012; 149: 994–1007. https://doi.org/10.1016/j.cell.2012.04.023 PMID: 22608083

28. 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: 2239–2254.e39. https://doi.org/10.1016/j.cell.2021.03.009 PMID: 33831375

29. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. Nat Biotechnol. 2021; 39: 207–214. https://doi.org/10.1038/s41587-020-0661-6 PMID: 32879467

30. Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh P-R, et al. Haplotype-enhanced inference of somatic copy number profiles from single-cell transcriptomes. bioRxiv. 2022. p. 2022.02.07.479314. https://doi.org/10.1101/2022.02.07.479314

31. Wu C-Y, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. Nat Biotechnol. 2021; 39: 1259–1269. https://doi.org/10.1038/s41587-021-00911-w PMID: 34017141

32. Funnell T, O'Flanagan CH, Williams MJ, McPherson A, McKinney S, Kabeer F, et al. Single-cell genomic variation induced by mutational processes in cancer. Nature. 2022; 612: 106–115. https://doi.org/10.1038/s41586-022-05249-0 PMID: 36289342

33. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015; 26: 64–70. https://doi.org/10.1093/annonc/mdu479 PMID: 25319062

34. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016; 44: e131. https://doi.org/10.1093/nar/gkw520 PMID: 27270079

35. 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. https://doi.org/10.1038/nbt.2203 PMID: 22544022

36. 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. https://doi.org/10.1186/s13059-022-02794-9 PMID: 36376909

37. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014; 10: e1003535. https://doi.org/10.1371/journal.pcbi.1003535 PMID: 24743184

38. 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. https://doi.org/10.1038/ng.2760 PMID: 24071852

39. Wolff AC, Hammond MEH, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol. 2013; 31: 3997–4013. https://doi.org/10.1200/JCO.2013.50.9984 PMID: 24101045

40. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017; 545: 446–451. https://doi.org/10.1038/nature22364 PMID: 28445469

41. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489: 519–525. https://doi.org/10.1038/nature11404 PMID: 22960745

42. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet. 2016; 48: 607–616. https://doi.org/10.1038/ng.3564 PMID: 27158780

43. Khaled WT, Choon Lee S, Stingl J, Chen X, Raza Ali H, Rueda OM, et al. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nat Commun. 2015; 6: 5987. https://doi.org/10.1038/ncomms6987 PMID: 25574598

44. Weniger MA, Pulford K, Gesk S, Ehrlich S, Banham AH, Lyne L, et al. Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11A(XL) protein are frequent in primary mediastinal B-cell lymphoma. Leukemia. 2006; 20: 1880–1882. https://doi.org/10.1038/sj.leu.2404324 PMID: 16871282

45. Lazarus KA, Hadi F, Zambon E, Bach K, Santolla M-F, Watson JK, et al. BCL11A interacts with SOX2 to control the expression of epigenetic regulators in lung squamous carcinoma. Nat Commun. 2018; 9: 3327. https://doi.org/10.1038/s41467-018-05790-5 PMID: 30127402

46. 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. https://doi.org/10.1038/ng.3214 PMID: 25665006

47. Brastianos PK, Carter SL, Santagata S, Cahill DP, Taylor-Weiner A, Jones RT, et al. Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. Cancer Discov. 2015; 5: 1164–1177. https://doi.org/10.1158/2159-8290.CD-15-0369 PMID: 26410082

48. Chen Y-C, Seifuddin F, Nguyen C, Yang Z, Chen W, Yan C, et al. Comprehensive Assessment of Somatic Copy Number Variation Calling Using Next-Generation Sequencing Data. bioRxiv. 2021. p. 2021.02.18.431906. https://doi.org/10.1101/2021.02.18.431906

49. 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. https://doi.org/10.1038/s41588-020-0584-7 PMID: 32139907

50. 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. https://doi.org/10.1158/2159-8290.CD-13-0285 PMID: 24436049

51. 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. https://doi.org/10.1038/s41588-018-0165-1 PMID: 30013179

52. 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. https://doi.org/10.1038/s41467-020-17967-y PMID: 32879317

53. 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. https://doi.org/10.1038/nature14347 PMID: 25830880

54. Huebner A, Black JRM, Sarno F, Pazo R, Juez I, Medina L, et al. ACT-Discover: identifying karyotype heterogeneity in pancreatic cancer evolution using ctDNA. Genome Med. 2023; 15: 27. https://doi.org/10.1186/s13073-023-01171-w PMID: 37081523

55. Fan J, Lee H-O, Lee S, Ryu D-E, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018; 28: 1217–1227. https://doi.org/10.1101/gr.228080.117 PMID: 29898899

56. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012; 44: 955–959. https://doi.org/10.1038/ng.2354 PMID: 22820512

57. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016; 48: 1443–1448. https://doi.org/10.1038/ng.3679 PMID: 27694958

58. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019; 10: 5436. https://doi.org/10.1038/s41467-019-13225-y PMID: 31780650

59. Selvaraj S, R Dixon J, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013; 31: 1111–1118. https://doi.org/10.1038/nbt.2728 PMID: 24185094

60. Markowski J, Kempfer R, Kukalev A, Irastorza-Azcarate I, Loof G, Kehr B, et al. GAMIBHEAR: whole-genome haplotype reconstruction from Genome Architecture Mapping data. Bioinformatics. 2021. https://doi.org/10.1093/bioinformatics/btab238 PMID: 33830196

61. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J Comput Biol. 2015; 22: 498–509. https://doi.org/10.1089/cmb.2014.0157 PMID: 25658651

62. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017; 27: 801–812. https://doi.org/10.1101/gr.213462.116 PMID: 27940952

63. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. Comparison of phasing strategies for whole human genomes. PLoS Genet. 2018; 14: e1007308. https://doi.org/10.1371/journal.pgen.1007308 PMID: 29621242

64. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013; 501: 338–345. https://doi.org/10.1038/nature12625 PMID: 24048066

65. Hieronymus H, Murali R, Tin A, Yadav K, Abida W, Moller H, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife. 2018; 7. https://doi.org/10.7554/eLife.37294 PMID: 30178746

66. Ross EM, Haase K, Van Loo P, Markowetz F. Allele-specific multi-sample copy number segmentation in ASCAT. Bioinformatics. 2020. https://doi.org/10.1093/bioinformatics/btaa538 PMID: 32449758

67. McPherson AW, Roth A, Ha G, Chauve C, Steif A, de Souza CPE, et al. ReMixT: clone-specific genomic structure estimation in cancer. Genome Biol. 2017; 18: 140. https://doi.org/10.1186/s13059-017-1267-2 PMID: 28750660

68. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014; 24: 1881–1893. https://doi.org/10.1101/gr.180281.114 PMID: 25060187

69. Litchfield K, Stanislaw S, Spain L, Gallegos LL, Rowan A, Schnidrig D, et al. Representative Sequencing: Unbiased Sampling of Solid Tumor Tissue. Cell Rep. 2020; 31: 107550. https://doi.org/10.1016/j.celrep.2020.107550 PMID: 32375028

70. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Routledge; 2013.

71. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023. https://doi.org/10.1038/s41586-023-05783-5 PMID: 37046096