nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-019-0364-z

A community effort to create standards for evaluating tumor subclonal reconstruction

Adriana Salcedo[1,2,29], Maxime Tarabichi[3,4,29], Shadrielle Melijah G. Espiritu[1,29], Amit G. Deshwar[5,29], Matei David[1], Nathan M. Wilson[1], Stefan Dentro[3,4], Jeff A. Wintersinger[6], Lydia Y. Liu[1], Minjeong Ko[1], Srinivasan Sivanandan[1], Hongjiu Zhang[7], Kaiyi Zhu[8,9,10], Tai-Hsien Ou Yang[8,9,10], John M. Chilton[11], Alex Buchanan[12], Christopher M. Lalansingh[1], Christine P'ng[1], Catalina V. Anghel[1], Imaad Umar[1], Bryan Lo[1], William Zou[1], DREAM SMC-Het Participants[13], Jared T. Simpson[1], Joshua M. Stuart[14], Dimitris Anastassiou[8,9,10,15], Yuanfang Guan[7,16,17], Adam D. Ewing[18], Kyle Ellrott[11,12,30], David C. Wedge[19,20,30], Quaid Morris[1,6,21,22,30], Peter Van Loo[3,23,30] and Paul C. Boutros[2,24,25,26,27,28,30]*

Tumor DNA sequencing data can be interpreted by computational methods that analyze genomic heterogeneity to infer evolutionary dynamics. A growing number of studies have used these approaches to link cancer evolution with clinical progression and response to therapy. Although the inference of tumor phylogenies is rapidly becoming standard practice in cancer genome analyses, standards for evaluating them are lacking. To address this need, we systematically assess methods for reconstructing tumor subclonality. First, we elucidate the main algorithmic problems in subclonal reconstruction and develop quantitative metrics for evaluating them. Then we simulate realistic tumor genomes that harbor all known clonal and subclonal mutation types and processes. Finally, we benchmark 580 tumor reconstructions, varying tumor read depth, tumor type and somatic variant detection. Our analysis provides a baseline for the establishment of gold-standard methods to analyze tumor heterogeneity.

Most tumors arise from a single ancestral cell, whose genome acquires one or more somatic driver mutations[1,2], which give it a fitness advantage over its neighbors by manifesting hallmark characteristics of cancers[3]. This ancestral cell and its descendants proliferate, ultimately giving rise to all cancerous cells within the tumor. Over time, they accumulate mutations, some leading to further fitness advantages. Eventually local clonal expansions can create subpopulations of tumor cells sharing subsets of mutations, termed subclones. As the tumor extends spatially beyond its initial location, spatial variability can arise as different regions harbor independently evolving tumor cells with distinctive genetic and nongenetic characteristics[4–9].

DNA sequencing of tumors allows quantification of the frequency of specific mutations based on measurements of the fraction of mutant sequencing reads, the copy number state of the locus and the tumor purity[10,11]. By aggregating these noisy frequency measurements across mutations, a tumor sample's subclonal architecture can be reconstructed from bulk sequencing data[6,11]. Subclonal reconstruction methods have proliferated rapidly in recent years[12–15], and have revealed key characteristics of tumor evolution[4,7,16–20], spread[21–23] and response to therapy[24,25]. Nevertheless, there has been no rigorous benchmarking of the relative or absolute accuracy of approaches for subclonal reconstruction.

There are several reasons why such benchmarking has not yet been performed. First, it is difficult to identify a gold-standard truth for subclonal reconstruction. While single-cell sequencing could provide ground truth, it has pervasive errors[26], and existing DNA-based datasets do not have sufficient depth and breadth to

[1]Ontario Institute for Cancer Research, Toronto, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Canada. [3]The Francis Crick Institute, London, UK. [4]Wellcome Trust Sanger Institute, Hinxton, UK. [5]The Edward S. Rogers Senior Department of Electrical & Computer Engineering, Toronto, Canada. [6]Donnelly Centre, University of Toronto, Toronto, Canada. [7]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [8]Department of Systems Biology, Columbia University, New York, NY, USA. [9]Center for Cancer Systems Therapeutics, Columbia University, New York, NY, USA. [10]Department of Electrical Engineering, Columbia University, New York, NY, USA. [11]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA. [12]Oregon Health & Sciences University, Portland, OR, USA. [13]A full list of authors and affiliations appears at the end of the paper. [14]Department of Biomolecular Engineering, Center for Biomolecular Sciences and Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA. [15]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, USA. [16]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [17]Department of Electronic Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [18]Mater Research Institute, University of Queensland, Woolloongabba, Queensland, Australia. [19]Big Data Institute, University of Oxford, Oxford, UK. [20]Oxford NIHR Biomedical Research Centre, Oxford, UK. [21]Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [22]Vector Institute for Artificial Intelligence, Toronto, Canada. [23]Department of Human Genetics, University of Leuven, Leuven, Belgium. [24]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada. [25]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [26]Department of Urology, University of California, Los Angeles, Los Angeles, CA, USA. [27]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA, USA. [28]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA. [29]These authors contributed equally: Adriana Salcedo, Maxime Tarabichi, Shadrielle Melijah G. Espiritu, Amit G. Deshwar. [30]These authors jointly supervised this work: Kyle Ellrott, David C. Wedge, Quaid Morris, Peter Van Loo, Paul C. Boutros. *e-mail: pboutros@mednet.ucla.edu

adequately assess subclonal reconstruction methods. Alternatively, gold-standard datasets may be generated using simulations, but existing tumor simulation methods such as BAMSurgeon[27] neither create representative subclonal populations nor phase simulated variants, which can be exploited in subclonal reconstruction[6,10]. Second, it is unclear how subclonal reconstruction methods should be scored, even in the presence of a suitable gold-standard. For example, one key goal of reconstruction is identification of the mutations present in each subclonal lineage. Metrics are needed that penalize errors both in the number of subclonal lineages and in the placement of mutations across them. Third, subclonal reconstruction methods have only been developed in recent years; few groups have equal expertise with multiple tools. Algorithm developers themselves are typically experts in parameterizing their own algorithms; an unbiased third party is needed to compare different methods, each run with expert parameterization.

To fill this gap, we developed a crowd-sourced benchmarking challenge: the ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity Challenge (SMC-Het). Challenge organizers simulated realistic tumors, developed robust scoring metrics and created a computational framework to facilitate unbiased method evaluation. Challenge participants then created re-distributable software containers representing their methods. These containers were run by the challenge organizers in an automated pipeline on a series of test tumors never seen by the challenge participants.

Here, we report the creation of quantitative metrics for scoring tumor subclonality reconstructions and describe tools for simulating tumors with realistic subclonal architecture. We apply these tools and metrics to characterize the sensitivity of subclonal reconstruction methodologies to somatic mutation detection algorithms and technical artifacts.

Results

How can subclonal reconstruction methods be evaluated?. Subclonal reconstruction is a complex procedure that involves estimating many attributes of the tumor including its purity, number of lineages, lineage genotypes and the phylogenetic relationships among lineages. We structured our evaluation of these attributes into three categories (Fig. 1). Subchallenge 1 quantifies the ability of an algorithm to reconstruct global characteristics of tumor composition. Specifically, it evaluates each algorithm's predictions of the total fraction of cells that are cancerous (tumor purity; Subchallenge 1A), the number of subclonal lineages (Subchallenge 1B) and for each subclone the fraction of cells (cellular prevalence, CP) and number of mutations associated with it (Subchallenge 1C). Subchallenge 2 evaluates how accurately each algorithm assigns individual single nucleotide variants (SNVs) to each subclonal lineage. It evaluates both their single best guess at a hard assignment of SNVs to lineages (Subchallenge 2A) and soft assignments represented through coclustering probabilities (that is, the probability that two SNVs are in the same lineage; Subchallenge 2B). Finally, Subchallenge 3 evaluates the ability of algorithms to recover the phylogenetic relationships between subclonal lineages, again both as a single hard assignment (Subchallenge 3A) and as a soft assignment (Subchallenge 3B). Taken together, SMC-Het comprises seven specific subchallenges, each corresponding to specific outputs on which subclonal reconstruction methods can be benchmarked (Methods).

To quantify the accuracy of these seven outputs, we considered several candidate scoring metrics, all bound between zero (very poor performance) and one (perfect performance). Appropriate metrics for Subchallenge 1 were trivially identified (Methods and Supplementary Note 1), but Subchallenges 2 and 3 required us to modify existing metrics and develop new ones. Specifically, because Subchallenges 2B and 3B are based on pairwise probabilities of coclustering, we were unable to use either clustering quality metrics designed for hard clustering or those that require explicit estimation of the number of clusters, such as normalized mutual information (also known as the V measure[28]).

As Subchallenges 2 and 3 involve assigning mutations to subclonal lineages, we required candidate metrics to satisfy three conditions:[28]

1. The score decreases as the predicted number of subclonal lineages diverges from the true number of subclonal lineages.
2. The score decreases as the proportion of mutations assigned to incorrect subclonal lineages (predicted subclonal lineages that do not correspond to the true subclonal lineage) increases.
3. The score decreases as the proportion of mutations assigned to noise subclonal lineages (predicted subclonal lineages that do not correspond to any true subclonal lineage) increases.

Moreover, metrics for evaluating cluster assignments have a number of desirable properties[28]. We identified a set of these applicable to each task (Supplementary Note 1), used a simulation framework to assess how well a candidate metric satisfies them. We identified four complementary metrics that satisfy all three properties: Matthew's correlation coefficient (MCC), Pearson's correlation coefficient (PCC), area under the precision-recall (AUPR) curve and average Jensen–Shannon divergence (AJSD) (Supplementary Fig. 1).

To further refine this set, we tested their behavior relative to subclonal reconstruction errors related to parent versus child and parent versus cousin relationships, and splitting or merging of individual nodes (Supplementary Note 1). Nine experts ranked the overall severity of up to eight error cases for each of 30 tree topologies, providing 2,088 total expert rankings. We then simulated each error case and scored it with all candidate metrics (Fig. 2a–d). For subchallenge 3, we added one metric, the clonal fraction (CF), which scores the accuracy of the predicted fraction of mutations assigned to the clonal peak. Unlike Subchallenge 2, which scores mutation assignment, that is, genotyping of the (sub)clones, subchallenge 3 scores tree topology, which indicates an ordering of events. The CF was designed to capture expert knowledge that emerged from the expert ranking: experts tended to favor the merging of two subclonal clusters over merging of the clonal cluster with a subclonal cluster, which was not captured by other metrics. The fraction of (sub)clonal mutations is indeed a biologically relevant measure that varies widely across cancer types[29]. Given that our metric rankings are based on subjective expert viewpoints, we have made our ranking system available online to allow others to create their own rankings and compare them to ours or use them to fine-tune scoring metrics for their own applications (https://mtarabichi.shinyapps.io/SMCHET).

Between-expert agreement, measured as pairwise rank correlations (0.52 ± 0.22), were much higher than metrics-expert agreement (for Subchallenge 2B, mean: 0.14 ± 0.12 s.e.m., $n = 270$; for Subchallenge 3B, mean: 0.12 ± 0.15 s.e.m., $n = 270$; Fig. 2d). Subsets of metrics were highly correlated (Jensen–Shannon (JS), PCC and MCC; range: 0.97–0.99, $n = 464$), whereas others were less correlated (AUPR, JS/PCC/MCC and CF; range: 0.47–0.78, $n = 464$). We reasoned that less-correlated metrics might capture complementary aspects of the reconstructions and derived additional metrics combining the best of them, as well as an average of all (Fig. 2d). For Subchallenge 2, the average of two metrics $\left(\frac{\text{AUPR}+\text{JS}}{2}\right)$ and AUPR was significantly better correlated to experts than any individual metric ($\bar{\rho}_{\text{Spearman}} = 0.21$, $n = 30$; Fig. 2c,d). For Subchallenge 3, AUPR, MCC, PCC and JS were comparable and significantly better than the other metrics ($\bar{\rho}_{\text{Spearman}} \in [0.19, 0.23]$, $n = 30$). We chose the PCC for subsequent analysis as it allows for assessment with a nonbinary truth. The resulting expert rankings and quantitative comparisons provide a basis for future development of improved scoring metrics.

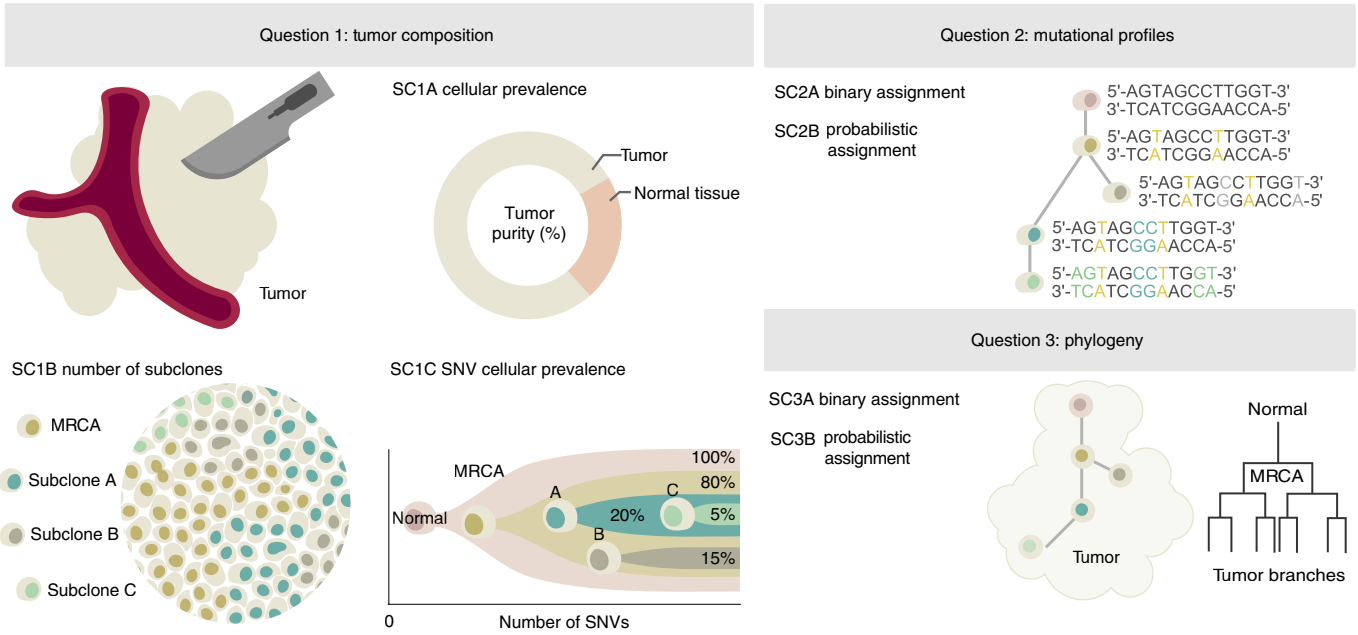

Fig. 1 | Features of tumor subclonal reconstruction. Overview of the key performance aspects of subclonal reconstruction algorithms, grouped into three broad areas covered by three key questions: Subchallenge 1 (SC1) 'What is the composition of the tumor?' This involves quantifying its purity, the number of subclones and their prevalence and mutation load. Subchallenge 2 (SC2) 'What are the mutational characteristics of each subclone?' This can be answered both with a point-estimate and a probability profile, that is, hard or probabilistic assignments of mutations to subclones, respectively. Subchallenge 3 (SC3) 'What is the evolutionary relationships among tumor subclones?' This again can be answered with both a point-estimate and a probability profile. MRCA, most recent common ancestor.

Simulating accurate subclonal tumor genomes. We elected to use simulated tumor data to run SMC-Het. The key reasons were the unavailability of deep single-cell DNA sequencing data as a gold-standard, the lack of single-cell sequencing data that match arbitrary tree structures and characteristics, the ability to simulate a large number of tumors at low-cost and the demonstrated ability of tumor simulations to recapitulate complex sequencing error profiles[27]. We elected to use the BAMSurgeon tool created for the earlier SMC-DNA Challenges[27,30], which creates tumors with accurate SNVs, indels and small genomic rearrangements at varying allelic fractions. However, that version of BAMSurgeon lacked a number of key features for our purpose. We added five main features: (1) phasing of variants, (2) large-scale allele-specific copy number changes (including whole-genome duplications), (3) translocations, (4) trinucleotide SNV signatures and (5) replication-timing effects (Figs. 3 and 4). We describe each of these briefly.

Phasing of mutations. To correctly simulate a tumor, it is critical that genetic variants—both somatic and germline—are fully phased, as they are in real genomes. Without phasing, allele-specific copy number changes cannot be simulated correctly and will lead to incorrect B-allele frequencies (BAF) and allele-specific copy number calls, among other errors. To achieve correct and complete phasing, we leveraged Next-Generation Sequencing (NGS) data from a trio of individuals from the Genome-in-a-Bottle consortium (Supplementary Fig. 2a–e) and created the PhaseTools package to accurately phase heterozygous variants identified in these data (Methods and Supplementary Note 2). The final result of this process is two BAM files per chromosome, each representing a single parental copy.

Simulation of a tumor BAM with underlying tree topology (Fig. 3a). To simulate a tumor BAM starting from the fully phased genome, we assigned subsets of the reads to each tree node, generating down-sampled BAM files. To simulate whole-chromosome copy number events, we adjusted the proportion of reads assigned to each node of the tree (Fig. 3b; see below). Then, BAMSurgeon was used on each subBAM to simulate mutations, including SNVs, indels and structural variants (SVs) (Fig. 3c). This strategy allowed us to efficiently and reliably simulate copy number changes of arbitrary size and add specific mutations on each allelic copy. Finally, these subBAMs were merged to produce the final BAM. By contrast, when we used the subclonally naive BAMSurgeon, copy number inference was incorrect (Supplementary Fig. 2f,g). After adding subclonal mutations only by specifying the variant allele frequency (VAF; that is, without phasing or subsampling BAM files) SNVs that occurred after duplications or deletions often appeared at the wrong frequency (Supplementary Fig. 2h).

Whole arm and whole-genome copy number changes. To allow changes in copy number of entire chromosomes and whole-genome ploidy changes (for example, whole-genome duplications, present in 30–50% of human cancers[31–33]), we developed a method to account for gains or losses of any chromosome, including sex chromosomes based on bookkeeping of reads assigned to each node. Given a tumor design structure (Fig. 3b), reads from the phased genomes were further split into individual subpopulations (subBAMs for leaf nodes) that make up the tumor in proportion to the copy number state of the region they aligned to and the CP of their node. The extracted and modified reads were merged to generate a final BAM file (Fig. 3c).

Translocations and large-scale SVs. As the prior BAMSurgeon functionality could not reliably simulate SVs larger than 30 kilobase pairs or any translocations due to its use of assembly, we extended it to simulate translocations, inversions, deletions and duplications of arbitrary size. This required a new approach of creating a simulated translocation that accurately reflects the expected pattern of discordant read pair mappings and split reads (Supplementary Note 2). This also allows us to simulate translocations, which were not included in the SMC-DNA simulated data challenges[30].

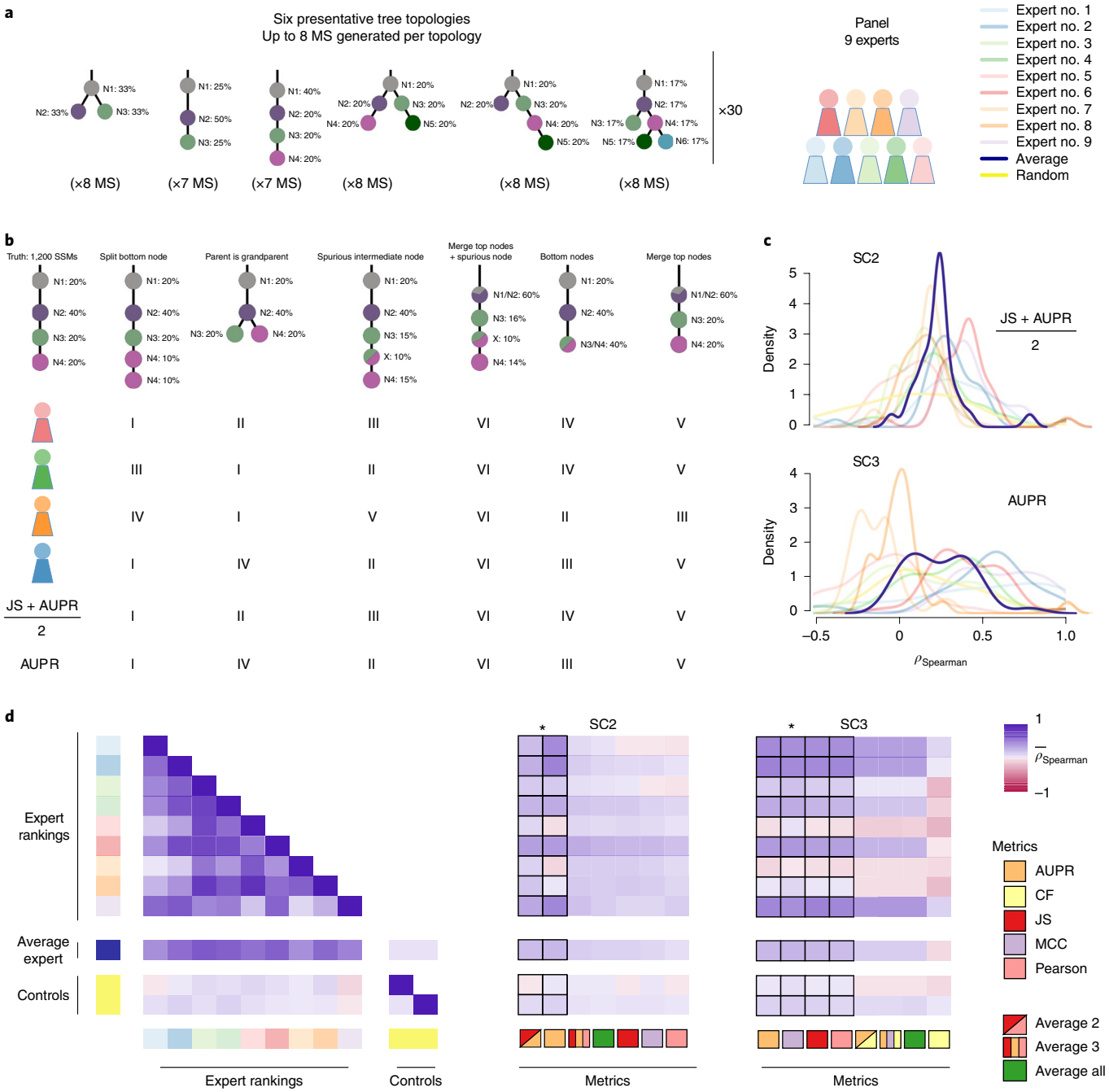

Fig. 2 | Quantifying performance of subclonal reconstruction algorithms. a, Tree topologies and mistake scenarios. For each of 30 tree topologies with varying number of clusters and ancestral relationships, 7–8 mistake scenarios (MS) were derived and scored using the identified metrics for Subchallenges 2 and 3. For each tree topology, a panel of nine experts independently ranked the mistake scenarios from best to worse. **b**, Expert ranking. One tree topology is shown with six of the seven mistake scenarios together with the ranks of four experts and two of the metrics. The trivial all-in-one case, that is, identifying only one cluster, is not shown and ranked last by all metrics and experts. **c**, Density distributions of Spearman's correlations between metrics and experts across tree topologies. For Subchallenges 2 and 3, we show the Spearman's correlations between JS and AUPR/2 and the experts, and AUPR and the experts, respectively. **d**, All average correlations between experts and metrics for Subchallenges 2 and 3. Heatmaps of average Spearman's correlations across tree topologies between experts and metrics for Subchallenges 2 and 3. Controls are randomized ranks. Asterisks show equivalent metrics (nonsignificantly better or worse according to a Wilcoxon rank-sum test $P > 0.05$ but better than the others $P < 0.01$; $n = 270$; range of median increase in correlation coefficients: Subchallenge 2, 0.018–0.23 and subchallenge 3, 0.024–0.36).

The ability to simulate translocations combined with adjustments to read coverage makes the simulation of arbitrarily large and complex SVs possible.

Trinucleotide mutation profile and replication timing. Single nucleotide mutations are not uniformly distributed throughout cancer genomes. They are biased both regionally and locally[34]. Mutations result from specific mutagenic stresses, which can induce biased rates of occurrence at specific trinucleotide contexts[35]. Replication-timing bias refers to the increase in the mutation rate of regions of the genome that replicate late in the cell cycle[34]. To resolve this issue, we created an extensible approach as part of BAMSurgeon. Each

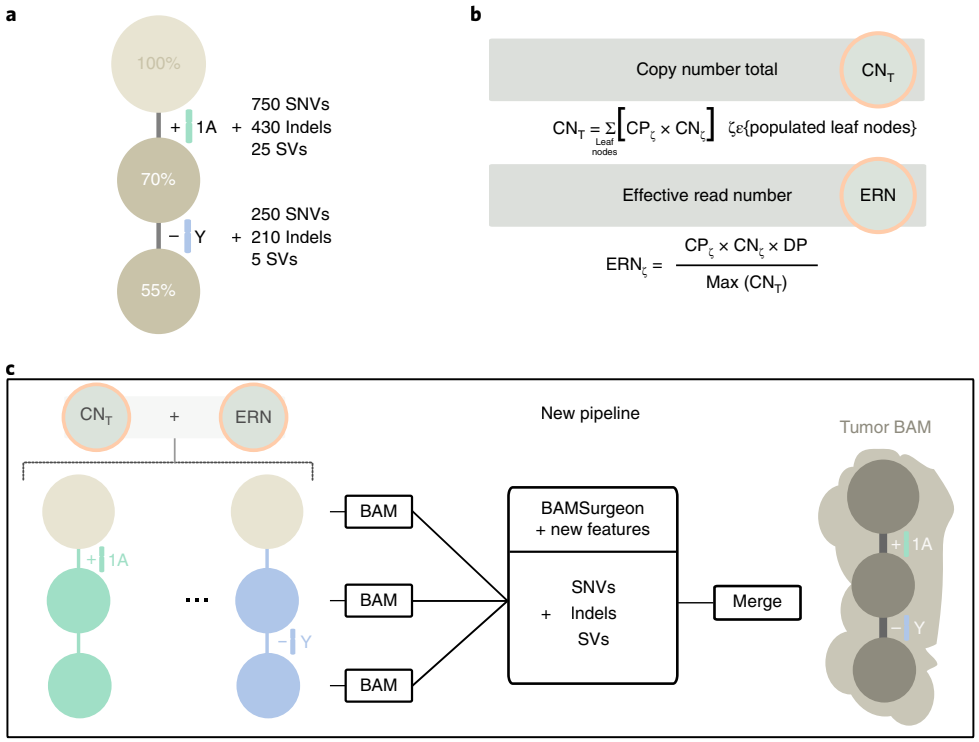

Fig. 3 | Simulating subclonal CNAs in tumor BAM files and spiking somatic mutations. Example case of read number adjustment to simulate subclonal CNAs. **a**, Desired structure of the tumor being simulated. **b**, Read number adjustment calculations. The copy number total (CN_T) for each chromosome is its copy number by adjusted by node CP summed across all nodes. The maximum CN_T across the genome is retained to normalize copy number for all chromosomes. The number of reads assigned to each chromosome at each node (the chromosome's effective read number) is then computed as the product of the node's CP, the chromosome's copy number and the total tumor depth normalized by the maximum CN_T. DP, read depth. **c**, Separation per-chromosome phase and per node with the new pipeline to simulate tumor BAM files with underlying intra-tumor heterogeneity. The first tumor clone (70% CP) has a gain in one copy (referred to as copy A) of chromosome 1 and one of its descendant subclones (55% CP) bears a loss of the Y chromosome. After adjusting read number for CNAs in each BAM corresponding to a node, BAMSurgeon spikes in additional mutations including the new features (complex structural variants, SNVs with trinucleotide contexts and replication-timing effects and so on), and then merges the extracted reads into a final tumor BAM file.

nucleotide in the genome is weighted according to its trinucleotide context, replication timing and the set of mutational signatures. Bases are then sampled from the genome until the expected trinucleotide spectrum is reached (Supplementary Note 2). BAMSurgeon can handle arbitrary mutational signatures, replication-timing data at any resolution and any arbitrary type of locational bias in mutational profiles.

Selection. Our framework for picking selecting point mutations can easily be extended to incorporate other biases in mutation frequency or location such as selection. Although explicit tumor growth models remain an area of active development[36–38] and discussion[39,40] we sought to illustrate this functionality using a recent model of 3D tumor growth that shows selection is reflected in VAF distributions across 3D tumor subvolumes[37]. We obtained VAFs from this simulator at five different levels of selection. For each level of selection, we simulated one 3D tumor and the resection of three tumor subregions. These were taken as basis for our simulator to generate 15 tumor BAM files in which the spiked-in SNVs and their VAF were directly derived from the tumor growth models. The VAFs of the genotyped SNVs allowed accurate inference of the selection input parameter (Supplementary Fig. 2i and Supplementary Note 2), while also incorporating trinucleotide signatures and replication-timing effects. By contrast, we were unable to recover the signature of selection with MuTect SNV calls, suggesting that more than three tumor regions might be needed to detect selection through this method when substantial variant detection errors are present, emphasizing the use of simulated tumor BAMs in algorithm and model assessment (Supplementary Note 2).

Each of the simulated features was verified by comparing simulated to designed values: observed to expected measurements in the BAMs (Methods and Supplementary Fig. 3). Starting from a tumor design (Fig. 4a) we systematically and quantitatively compared observed and expected trinucleotide context (Fig. 4b), cancer cell fraction (CCF) (Fig. 4c) and copy number segment logR ratios and BAF (Fig. 4d,e). These were reviewed across all simulations to verify simulated data. These results also confirmed that BAMSurgeon can now generate complex subchromosomal events, including large deletions or duplications (Fig. 4f).

General features of subclonal reconstruction. We next sought to quantify how different factors affect subclonal reconstruction. We therefore simulated five tumors derived from different tissue types (prostate, lung, chronic lymphocytic leukemia, breast and colon) from published subclonal structures (Supplementary Fig. 3). We also analyzed a real tumor (PD4120) sequenced at 188× coverage with a high-quality consensus subclonal reconstruction based on the full-depth tumor[41] as the gold standard.

For each of these six tumors, we then down-sampled each tumor BAM to create a titration series in raw read depth of 8×, 16×, 32×, 64× and 128× coverage. For each of the 30 resulting tumor-depth combinations, we identified subclonal copy-number aberrations (CNAs) using Battenberg[6], both with down-sampled tumors and with tumors at the highest possible depth to assess the influence

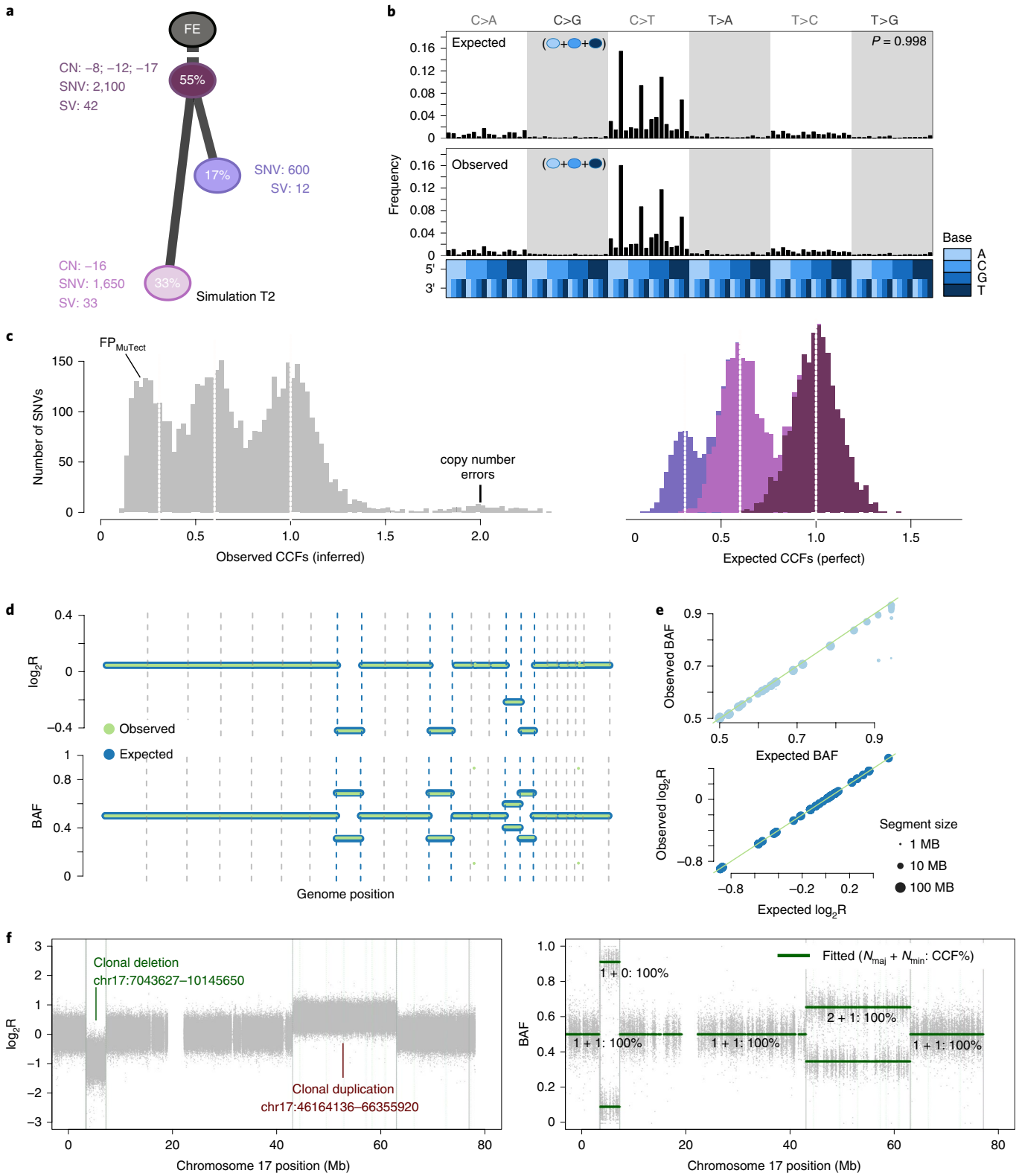

Fig. 4 | Simulated realistic tumor genomes. a, Tumor design. Simulation T2 with 55% purity (fraction of cancer cells) and two subclones. Whole-chromosome copy number events (for example, clonal loss of chromosomes 8, 12 and 17), number of SNVs and SVs are shown for each node. **b**, SNV trinucleotide contexts. Observed versus expected frequencies of trinucleotide contexts in the SNVs. **c**, Population frequency (CCF) of the variants for T2. Observed versus expected CCF distributions; false positive (FP) SNVs due to mutation calling as well as copy number errors lead to errors in the inferred CCFs. **d**, Observed (green) versus expected (blue) logged coverage ratio (logR) and BAF of copy number segments along the genome for T2. **e**, Observed versus expected BAF and logR across all segments and across all simulations. **f**, Simulation of subchromosomal copy number events and rearrangements. The logR and BAF tracks show how one large deletion and one large duplication simulated on chromosome 17 are correctly being called. Structural variants as called by Manta (Methods) are shown as vertical lines, true positives are at the breakpoints defining the copy number events.

of CNA detection accuracy, yielding 60 tumor-depth-CNA combinations. For each of these combinations, we identified somatic SNVs using four algorithms (MuTect[42], SomaticSniper[43], Strelka[44] and MutationSeq[45]), as well as the perfect somatic SNV calls for the simulated tumors, yielding 290 synthetic tumor-depth-CNA–SNV combinations. We also applied these pipelines to the real PD4120 BAM (except those involving of perfect SNV calls) resulting in 40 additional depth-CNA–SNV combinations based on a real tumor, for a total of 290 combinations. The somatic SNV detection algorithms were selected to span a range of variant calling approaches: SomaticSniper uses a Bayesian approach, MuTect and Strelka model allele frequencies while MutationSeq predicts somatic SNVs with an ensemble of four classifiers trained on a gold-standard dataset. Finally, subclonal reconstruction was then carried out on each of these using two algorithms (PhyloWGS[13] and DPClust[6]), to give a final set of 580 tumor-depth-CNA–SNV-subclonal reconstruction algorithm combinations (see Supplementary Note 3 for algorithm descriptions). Each combination was evaluated using the scoring framework outlined above (Fig. 5, Supplementary Fig. 4 and Supplementary Tables 1 and 2). In general, MuTect and SomaticSniper are more sensitive to low frequency variants and potentially preferable for subclonal reconstruction[46,47]. MuTect achieved the highest SNV detection sensitivity in our synthetic tumors (mean sensitivity 0.65 ± 0.037 s.e.m., $n = 25$), followed by Strelka (0.59 ± 0.032), SomaticSniper (0.50 ± 0.031, $n = 25$) and finally MutationSeq (0.46 ± 0.045, n = 25).

This large-scale benchmarking of 580 simulated tumors reveals general features of subclonal reconstruction accuracy. For example, consider Subchallenge 1C: estimation of SNV CP. All reconstruction and SNV detection algorithms showed a consistent increase in accuracy with increasing sequencing depth for Subchallenge 1C (Fig. 5a,b). No somatic SNV detection algorithm matched the performance of perfect SNV calls ($\beta = 0.22$, $P = 0.0011$, generalized linear model, $n = 500$, d.f. = 29). By contrast, the use of high- versus low-depth sequencing for subclonal detection of CNAs had no detectable influence on reconstruction accuracy in either real or simulated tumors ($P > 0.05$, generalized linear model, $n = 500$, d.f. = 29; Supplementary Table 2), likely due to the copy number changes being mostly whole-chromosome aberrations. In Subchallenge 1C, neither the use of low- versus high-depth tumors for CNA detection nor the specific subclonal reconstruction algorithm used had a statistically significant influence on the accuracy of subclonal reconstruction. Both PhyloWGS and DPClust performed interchangeably on this question in the simulated tumors ($P = 0.14$, $t = -1.47$, $n = 290$ Supplementary Fig. 5g–l and Supplementary Table 2).

A different story emerged for Subchallenge 2A: identifying the mutational profiles of individual subclones (Fig. 5c,d). All algorithms performed relatively poorly, with major intertumor differences in performance. Tumor T2 was systematically the most challenging to reconstruct and T6 the easiest (Fig. 5c and Supplementary Table 5). This in part reflects the higher purity of T6, and indeed we see a strong association between effective read depth and reconstruction accuracy in both the simulated and real tumors, with each additional doubling in read depth increasing reconstruction score by about 0.1 (Fig. 5d). At effective read depths above 60×, the performance of all tumor-CNA–SNV-subclonal reconstruction combinations seemed to plateau, suggesting that a broad range of approaches can be effective for detection of subclonal mutational profiles at sufficient read depth. Again, the use of high- versus low-depth sequencing for subclonal CNA detection had no discernible influence (and this held true for all subchallenges, see Supplementary Table 2). By contrast, Subchallenge 2A scores were strongly dependent on the SNV detection pipeline, with perfect calls out-performing the best individual algorithm (MuTect) by ~0.05 at any given read depth. Differences in SNV detection algorithm sensitivity largely accounted for performance differences among algorithms ($\beta_{\text{sensitivity}} = 0.30$, $P = 8.92 \times 10^{-13}$, generalized linear models, $n = 500$, d.f. = 30; Supplementary Table 3). MuTect, the most sensitive SNV detection algorithm, had the best performance and MutationSeq, the least sensitive, had the poorest. Broadly, SomaticSniper and Strelka showed similar performance, but showed significant tumor-by-algorithm interactions for several subchallenges (Supplementary Fig. 5a–f), which may reflect tumor-specific variability in their error profiles. Notably, MutationSeq performed much better on the real tumor than on simulated tumors (Supplementary Fig. 5a–f).

In general, DPClust and PhyloWGS showed very similar performance, but with exceptions that reflect their underlying algorithmic features. First, in Subchallenge 1A DPClust, which uses purity measures derived from CNA reconstructions, showed a significant and systematic advantage over PhyloWGS ($\beta_{\text{PhyloWGS}} = -0.42$, $P = 1.5 \times 10^{-7}$, generalized linear model, $n = 500$, d.f. = 13), which uses purity measures partially dependent on SNV clustering. The last measures are more sensitive to errors in VAF due to low sequencing depth and this is reflected in the pattern of Subchallenge 1A scores. Second, in Subchallenge 2B PhyloWGS, which uses a phylogenetically aware clustering model, had significantly better performance than DPClust, which uses a flat clustering model (Supplementary Fig. 5g and Supplementary Table 2). Thus, our metrics are sensitive to differences in modeling approaches, which manifest in variability in performance on different aspects of subclonal reconstruction. Validating these results, for the real high-depth tumor, DPClust significantly outperformed PhyloWGS in Subchallenge 1, while PhyloWGS was superior in Subchallenge 2 (Supplementary Table 4).

Robustness of subclonal reconstruction. Surprised by the insensitivity of scores to the use of high- or low-depth sequencing data for subclonal CNA assessment, we sought to characterize the sensitivity of subclonal reconstruction to errors in CNA detection. We repeated the analyses described above using five types of CNA input: original (untouched), CNAs with doubled ploidy, CNA calls with a random portion of existing calls wrongly assigned (scramble) and CNAs with additional gains (scramble gains) or with additional losses (scramble loss). The last three error types were titrated in intensity, scrambling 10, 20, 30, 40 and 50% of all CNAs, gains and losses, respectively.

The resulting 4,250 tumor-depth-CNA–SNV-reconstruction combinations were each assessed using our scoring metrics (Supplementary Table 1). For Subchallenge 1 and Subchallenge 2, incorrect ploidy impaired reconstruction accuracy overall (Fig. 6a,c). As expected, scores decreased as the proportion of incorrectly assigned CNAs increased (Supplementary Fig. 6a,b). The effect of incorrect calls on Subchallenge 2A accuracy was only apparent at >32× coverage and was strongest with perfect and MuTect SNVs (Fig. 6b,d), suggesting the relative impact of CNA errors increases with reconstruction quality. PhyloWGS had significantly better performance for all subchallenges than DPClust when CNA errors were introduced (Subchallenge 1C: $\beta_{\text{PhyloWGS}} = 0.042$, $P = 6.06 \times 10^{-10}$; Subchallenge 2A: $\beta_{\text{PhyloWGS}} = 0.066$, $P = 1.85 \times 10^{-10}$ generalized linear models, $n = 4{,}250$, d.f. = 21 and d.f. = 33; Supplementary Table 5). These results indicate that PhyloWGS's strategy of incorporating CNAs in the allele count model may be more robust to errors in CNA detection than only using them to initially correct SNV VAFs (Supplementary Fig. 5g and Supplementary Note 3). As CNA handling in the presence of errors distinguishes algorithms with otherwise comparable performance, increasing robustness to errors in CNA calls may be a promising avenue for improvement of subclonal reconstruction algorithms.

Taken together, these results suggest that subclonal reconstruction accuracy is highly sensitive both to SNV and CNA detection, with interactions between specific pairs of variant detection and subclonal reconstruction algorithms (Methods and Supplementary Fig. 6c,d). There is significant room for algorithmic improvements

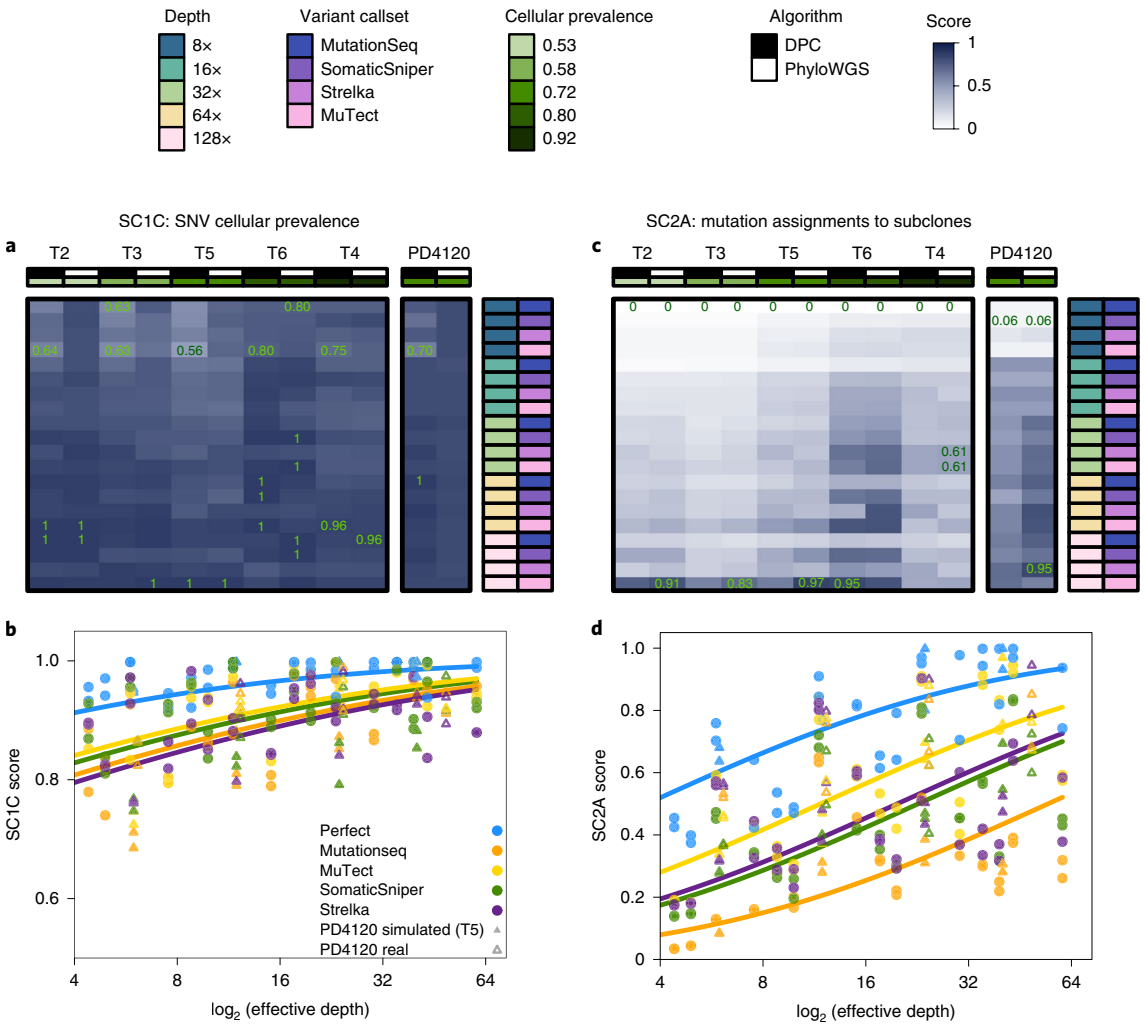

Fig. 5 | Error profiles of subclonal reconstruction algorithms. To identify general features of subclonal reconstruction algorithms, we created a set of tumor-depth-CNA-SNV-subclonal reconstruction algorithm combinations by using the framework outlined in Figs. 3 and 4 to simulate five tumors with known subclonal architecture, followed by evaluation of two CNA detection approaches, five SNV detection methods, five read depths and two subclonal reconstruction methods. The resulting reconstructions were scored using the scoring harness described in Fig. 2, creating a dataset to explore general features of subclonal reconstruction methods. All scores are normalized to the score of the best performing algorithm when using perfect calls at the full tumor depth. Scores might exceed this baseline likely due to noise or overfitting and were capped at 1. Only scores from reconstructions using down-sampled CNAs are shown ($n = 300$ tumor-SNV-depth-subclonal reconstruction algorithm combinations). **a**, For Subchallenge 1C (SC1C) (identification of the number of subclones and their CP), all combinations of methods perform well. **b**, By contrast, for Subchallenge 2A (SC2A) (detection of the mutational characteristics of individual subclones), there is large intertumor variability in performance. **c**, Score for Subchallenge 1C (same as **a**) as a function of effective read depth (depth after adjusting for purity and ploidy) improves with increased read depth, and also changes with the somatic SNV detection method, with MuTect performing best but still lagging perfect SNV calls by a significant margin. **d**, Scores in Subchallenge 2A show significant changes in performance as a function of effective read depth.

that capture intertumor differences and better model the error characteristics of feature-detection pipelines.

Discussion

As DNA sequencing costs diminish and evidence for clinical use accumulates, increasingly large numbers of tumors are sequenced each year. Nevertheless, it remains common practice for only a single spatial region of a cancer to be sequenced. The reasons for this are myriad: costs of multi-region sequencing, needs to preserve tumor tissue for future clinical use and increasing analysis of scarce biopsy-derived specimens in diagnostic and metastatic settings. While robust subclonal reconstruction from multi-region sequencing is well-known[5–8], accurately reconstructing tumor evolutionary properties from single-region sequencing could open new avenues for linking these to clinical phenotypes and outcomes.

We describe a framework for evaluating single-sample subclonal reconstruction methods, comprising a new way of scoring their accuracy, a technique for phasing short-read sequencing data, an enhanced read-level simulator of tumor genomes with realistic biological properties and a portable software framework for rapidly and consistently executing a library of subclonal reconstruction algorithms. These elements, each implemented as open-source software and independently reusable, form an integrated system for quantitation of key parameters of subclonal reconstruction. We generate a 580 tumor titration series for evaluating subclonal reconstruction sensitivity to both effective read depth and specific somatic SNV detection pipelines. These data give guidance for improving subclonal reconstruction: increasing effective read depth above 60×, after controlling for tumor purity and ploidy. They also suggest reconstruction algorithm developers should consider

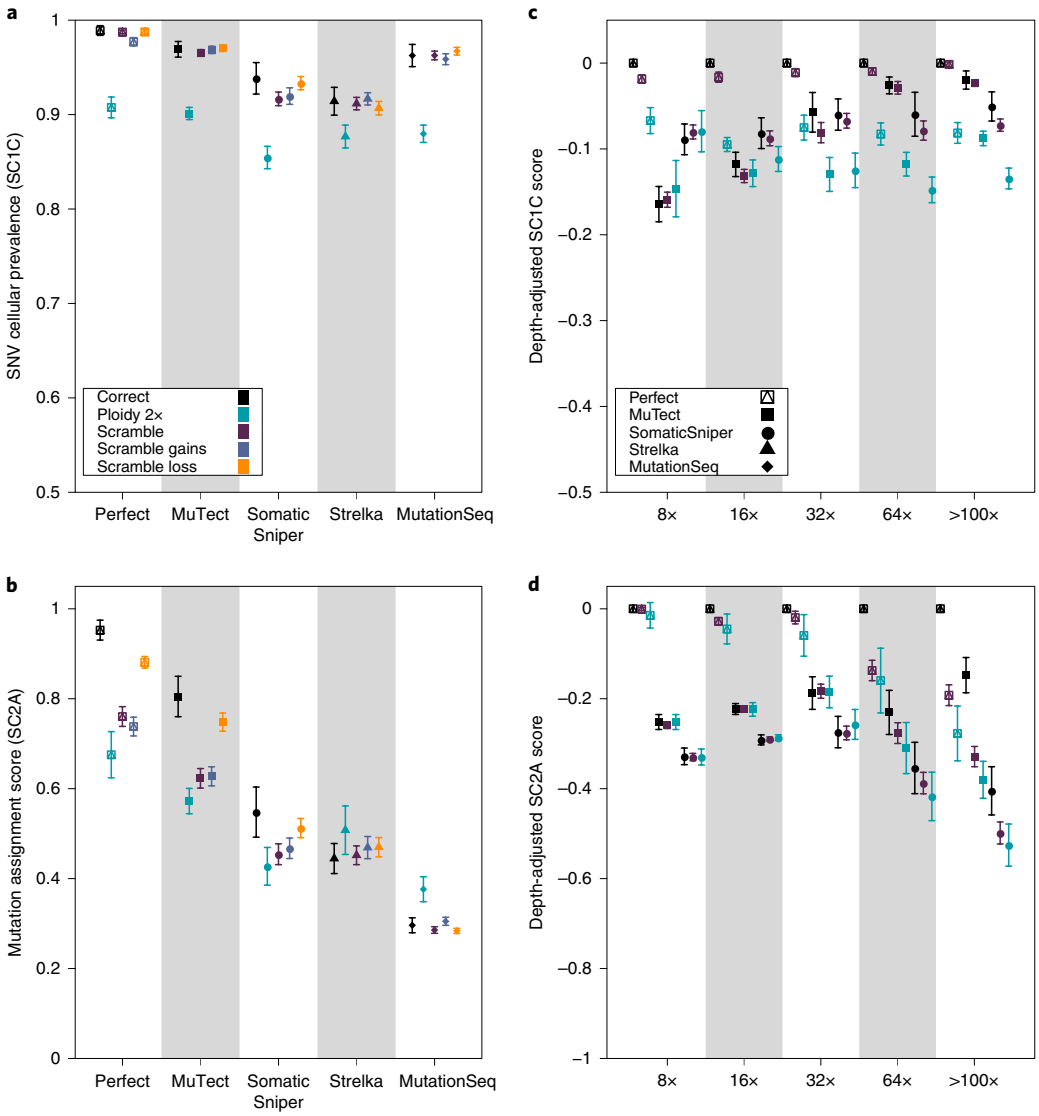

Fig. 6 | Impact of CNA error profiles on subclonal reconstruction. a,**b**, Effect of CNA errors on mean Subchallenge 1C scores (**a**) and Subchallenge 2A (**b**) scores (with standard errors shown) at 100× across somatic SNV detection algorithms ($n=850$). **c**,**d**, Effect of CNA errors on mean Subchallenge 1C (**c**) and Subchallenge 2A (**d**) scores (with standard errors shown, $n=2250$) at various depths when scores for perfect calls are set to zero to yield depth-adjusted scores.

accounting for the error properties of specific somatic variant detection approaches.

Lineage-tracing tools are emerging that will likely revolutionize our understanding of tissue growth and evolution, such as GESTALT[48], ScarTrace[49] and MEMOIR[50]. However, these are not applicable to the study of human cancer tissues in vivo. In many areas of biology, ground truth is still either inaccessible or impractical to measure with precision. In cases such as these, simulations are extremely valuable in providing a lower bound on error profiles and an upper bound on method accuracy. By incorporating all currently known features of a phenomenon, simulators codify our understanding. Divergence between simulated and real results quantitates the gaps in our knowledge. The creation of an open-source, freely available simulator capturing most known features of cancer genomes thus represents one avenue for exploring the boundaries of our knowledge.

Large-scale benchmarking of multiple subclonal reconstruction methods using this framework on larger numbers of tumors is needed to create a gold-standard. Such a benchmark would both inform algorithm users, who will benefit from an understanding of the specific error profiles of different methods, and algorithm developers who will be able to update and improve methods while ensuring software portability. Tumor simulation frameworks provide a valuable way for method benchmarking, and can complement other approaches such as comparison of single-region to multi-region subclonal reconstruction, and the use of model organism and sample-mixing experiments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-019-0364-z.

Received: 27 April 2018; Accepted: 18 November 2019;
Published online: 9 January 2020

References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
4. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
5. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
6. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
7. Cooper, C. S. et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).
8. Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
9. Caiado, F., Silva-Santos, B. & Norell, H. Intra-tumour heterogeneity—going beyond genetics. *FEBS J.* **283**, 2245–2258 (2016).
10. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
11. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
12. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
13. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
14. Fischer, A., Vázquez-García, I., Illingworth, C. J. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).
15. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
16. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
17. de Bruin, E. C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
18. Turajlic, S. et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell* **173**, 595–610.e11 (2018).
19. Espiritu, S. M. G. et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell* **173**, 1003–1013 (2018).
20. Wedge, D. C. et al. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* **50**, 682–692 (2018).
21. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
22. McPherson, A. et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48**, 758–767 (2016).
23. Turajlic, S. et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal. *Cell* **173**, 581–594.e12 (2018).
24. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
25. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).
26. Van Loo, P. & Voet, T. Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 82–91 (2014).
27. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
28. Rosenberg, A. & Hirschberg, J. V-Measure: a conditional entropy-based external cluster evaluation measure. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28–30, 2007, Prague, Czech Republic* (ed Eisner, J.) 410–420 (Association for Computational Linguistics, 2007).
29. Dentro, S. C. et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. Preprint at *bioRxiv* https://doi.org/10.1101/312041(2018).
30. Lee, A. Y.-W. et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19**, 188 (2018).
31. Cheng, J. et al. Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat. Commun.* **8**, 1221 (2017).
32. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
33. Storchova, Z. & Kuffer, C. The consequences of tetraploidy and aneuploidy. *J. Cell Sci.* **121**, 3859–3866 (2008).
34. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
35. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
36. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
37. Sun, R. et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
38. Williams, M. J. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
39. Tarabichi, M. et al. Neutral tumor evolution? *Nat. Genet.* **50**, 1630–1633 (2018).
40. Bozic, I., Paterson, C. & Waclaw, B. On measuring selection in cancer from subclonal mutation frequencies. *PLoS Comput Biol.* **15**, e1007368 (2019).
41. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. Preprint at *bioRxiv* https://doi.org/10.1101/162784 (2017).
42. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotech.* **31**, 213–219 (2013).
43. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
44. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
45. Ding, J. et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).
46. Xu, C. A review of somatic single nucleotide variant calling algorithms for Next-Generation Sequencing data. *Comput. Struct. Biotech. J.* **16**, 15–24 (2018).
47. Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244 (2014).
48. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
49. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
50. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).

DREAM SMC-Het Participants

Alokkumar Jha[31], Tanxiao Huang[32], Tsun-Po Yang[33], Martin Peifer[33], Cenk Sahinalp[34], Salem Malikic[35], Ignacio Vázquez-García[36], Ville Mustonen[37], Hsih-Te Yang[38], Ken-Ray Lee[39], Yuan Ji[40], Subhajit Sengupta[41], Justine Rudewicz[42], Macha Nikolski[43], Quentin Schaeverbeke[42], Ke Yuan[44], Florian Markowetz[45], Geoff Macintyre[45], Marek Cmero[46], Belal Chaudhary[45], Ignaty Leshchiner[47], Dimitri Livitz[47], Gad Getz[47], Phillipe Loher[48], Kaixian Yu[49], Wenyi Wang[49] and Hongtu Zhu[50]

[31]Insight Centre for data analytics, NUIG, Dublin, Ireland. [32]Bioinfo,HaploX Biotechnology, Shenzhen, China. [33]University of Cologne, Cologne, Germany. [34]Indiana University, Bloomington, IN, USA. [35]Simon Fraser University, Burnaby, Canada. [36]Memorial Sloan Kettering Cancer Center, Columbia University, Wellcome Sanger Institute, University of Cambridge, Cambridge, UK. [37]Wellcome Sanger Institute, University of Helsinki, Helsinki, Finland. [38]Levine Cancer Institute, Atrium Health, Charlotte, NC, USA. [39]Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan. [40]The University of Chicago, Chicago, IL, USA. [41]NorthShore University HealthSystem, Evanston, IL, USA. [42]Bordeaux University, Bordeaux, France. [43]Bordeaux University/CNRS, Bordeaux, France. [44]School of Computing Science, University of Glasgow, Glasgow, Scotland. [45]CRUK Cambridge Institute, University of Cambridge, Cambridge, UK. [46]University of Melbourne, Melbourne, Australia. [47]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [48]Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA, USA. [49]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [50]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Methods

Subchallenge descriptions. To evaluate subclonal reconstruction algorithms, we posed seven subchallenges and designed associated scoring metrics to evaluate performance in each subchallenge. Subchallenges 1A–1C, collectively called the subclonal architecture challenges, evaluated properties of the subclonal reconstruction without considering the assignment of individual SNVs to subclones. Subchallenges 2A and 2B, the clustering challenges, evaluated the assignments of individual SNVs to subclones. Subchallenges 3A and 3B, the ancestry challenges, evaluated the ancestral relationships of individual SNVs. Each of the subchallenge required submission data in a specific format described below.

Subchallenge 1: subclonal architecture. *Subchallenge 1A: cellularity.* Predict the proportion of cells in the sample that are cancerous (that is, the cellularity of the sample) or CP.

Output data: c is a real number with $0 \le c \le 1$ where c represents the predicted cellularity of the tumor sample.

Subchallenge 1B: lineage count. Predict the number of lineages (either subclonal or clonal) in the sample.

Output data: κ is a positive integer and $\kappa \ge 1$, where κ is the predicted number of lineages in the tumor sample. Note that we do not distinguish between clonal and subclonal lineages here, but it is assumed that each sample has at one (that is, clonal) lineage.

Subchallenge 1C: subclonal architecture. Predict (1) the proportion of the cells in the tumor sample in each of the subclonal lineages (that is, their CPs) and (2) the proportion of SNVs associated with each lineage. Collectively, we call these two predictions the estimated subclonal architecture.

Output data: $\boldsymbol{\varphi}$ is a vector containing κ real numbers, each of which, for example, $\boldsymbol{\varphi}_k$, represents the predicted CP in the associated predicted lineage k. Clearly, $0 \le \boldsymbol{\varphi}_k \le 1$ for all lineages k. Similarly, $\mathbf{N}$ is a vector containing κ positive integers, each of which, for example, $\mathbf{N}_k$, represents the predicted number of mutations in the associated lineage k. We insist that $\mathbf{N}_k \ge 1$.

Subchallenge 2: clustering. Predict the lineage assignment of each SNV.

Subchallenge 2A: single best hard assignment. Predict the assignment of each mutation to each lineage.

Output data: $\boldsymbol{\tau}$ is a vector of n positive integers, where n is the number of SNVs, in which each element $\boldsymbol{\tau}_i$ represents the index of the subclonal lineage to which mutation i is predicted to be assigned. Thus, $1 \le \boldsymbol{\tau}_i \le \kappa$.

Subchallenge 2B: probabilistic coclustering. Predict which pairs of mutations are in the same cluster. Note that this challenge differs from the previous one because the coclustering predictions can be probabilistic.

Output data: the predicted coclustering matrix (*CCM*), which is an $n \times n$ matrix of real numbers, where CCM_{ij} is the probability that mutation i is in the same subclone as mutation j, and $0 \le CCM_{ij} \le 1$. Note that a single best assignment can be represented by setting $CCM_{ij} = 1$ when mutation i and mutation j are assigned to the same lineage, and $CCM_{ij} = 0$ otherwise. Every mutation is assigned to the same lineage as itself, so we require that all the values on the diagonal of the *CCM* matrix be 1.

Subchallenge 3: ancestry. Predict the ancestral relationships between the SNVs.

Subchallenge 3A: single best ancestry. Predict the ancestral relationships among the predicted lineages.

Output data: $\mathbf{p}$ is a vector of κ positive integers, each one, for example, $\mathbf{p}_k$, is the index of the predicted parental lineage for lineage k where $\mathbf{p}_k = 0$ indicates that lineage k has no parent, that is, that it descends from the normal lineage. In other words, lineage k is a clonal lineage. Thus, $0 \le \mathbf{p}_k \le \kappa$ and $\mathbf{p}_k \ne k$.

Subchallenge 3B: probabilistic ancestor-descendant matrix (ADM). Predict the ancestral relationships among pairs of SNVs. Note that this challenge differs from the previous one because these predictions can be probabilistic.

Output data: the predicted *CCM*, as defined in Subchallenge 2B, and a predicted *ADM*, which is an $n \times n$ matrix where ADM_{ij} is the probability that mutation i is assigned to a subclonal lineage that is ancestral to the subclonal lineage the mutation j is assigned to, and $0 \le ADM_{ij} \le 1$. As in Subchallenge 2B, above, a single best ancestry can be represented by the *ADM* by setting ADM_{ij} if and only if mutation i is assigned to a lineage ancestral to that of mutation j. Elements on the diagonal of the *ADM* matrix required to all be 0.

Scoring metrics. Here, we describe each scoring metric used to evaluate the subclonal reconstruction algorithms.

Subchallenge 1A metric. The Subchallenge 1A score is

$$1 - |\rho - c|$$

where ρ is the true cellularity, c is the predicted cellularity and $|x|$ is the absolute value of x. Note that we require that $0 \le \rho \le 1$ and $0 \le c \le 1$.

Subchallenge 1B metric. The Subchallenge 1B score is $[L - d + 1]/(L + 1)$, where $L \ge 1$ is the true number of subclonal lineages, d is the absolute difference between the predicted and actual number of lineages, $d = \min(|\kappa - L|, L + 1)$. We do not allow d to be higher than $L + 1$ so that the Subchallenge 1B score is always ≥ 0.

Subchallenge 1C metric. Scoring Subchallenge 1C is challenging because the number of subclonal lineages can differ between the truth and the prediction, as can their size and CP. As such, we adopted metric based on the earth-mover distance (EMD) between the true and predicted architectures. First, we note that the subclonal architectures can be viewed as a clustering of data points in one dimension. In this view, each data point is a SNV, and they are clustered on the basis of their predicted CP into clusters corresponding to each lineage.

If we were considering individual SNVs in this metric, we could compute a distance between the real and the predicted clustering of those data points by computing the average value of $|\boldsymbol{\varphi}_k - \boldsymbol{\delta}_l|$ where $\boldsymbol{\varphi}_k$ is the CP of the lineage, k, that mutation i is assigned to in the predicted clustering and $\boldsymbol{\delta}_l$ is the CP of the lineage, l, that mutation i is assigned to in the true clustering. However, since we are not considering individual SNVs, we define the distance between two clusterings as the minimum possible value of this average, given the real and predicted subclonal architectures (that is, the vectors of CPs and counts of number of SNVs assigned to each cluster). This value is the exact (normalized) EMD between the real and predicted clusterings.

The procedure described below computes $1 - \text{EMD}$ given the true and predicted subclonal architectures.

First, we sort both the predicted subclonal lineages from 1 to κ and the true subclonal lineages from 1 to L in ascending order according to their CP. Let αk be the proportion of mutations assigned to predicted subclonal lineage k, for $k = 1 \ldots \kappa$. Similarly, let $\boldsymbol{\beta}_l$ be the proportion of mutations assigned to true subclonal lineage l, for $l = 1 \ldots L$. Let $\boldsymbol{\varphi}_k$ be the predicted CP of predicted subclonal lineage k for $k = 1 \ldots \kappa$ and let $\boldsymbol{\delta}_l$ be the true CP of true subclonal lineage l for $l = 1 \ldots L$.

Let $\boldsymbol{\omega}_p$ be a vector of S predicted real numbers with:
$\omega_{p,i} = \phi_1$ for $\frac{i}{S} \le \alpha_1$, or
$\omega_{p,i} = \phi_k$ for $\sum_{j \in 1 \ldots k-1} \alpha_j < \frac{i}{S} \le \sum_{j \in 1 \ldots k+1} \alpha_j$ or
$\omega_{p,i} = \phi_\kappa$ for $\sum_{j \in 1 \ldots \kappa-1} \alpha_j < \frac{i}{S}$
And let $\boldsymbol{\omega}_t$ be a vector of S true real numbers with:
$\omega_{t,i} = \delta_1$ for $\frac{i}{S} \le \beta_1$, or
$\omega_{t,i} = \delta_k$ for $\sum_{j \in 1 \ldots k-1} \beta_j < \frac{i}{S} \le \sum_{j \in 1 \ldots k+1} \beta_j$ or
$\omega_{t,i} = \delta_L$ for $\sum_{j \in 1 \ldots L-1} \beta_j < \frac{i}{S}$
We set S to 1,000 and the Subchallenge 1C scoring metric is then defined as:

$$\frac{1}{S} \sum_{s=1}^{S} |\omega_{t,s} - \omega_{p,s}|$$

We compute the Subchallenge 1C scoring metric using two different sets of true subclonal lineages. One set contains only the mutations that were spiked into the simulation. The other set of lineages also contains false positive mutations that were not spiked in, but were detected in somatic variant calling. In this set, the lineage containing the false positive mutations is assigned a CP of 0. Contestants receive the higher of the two scores.

Subchallenge 2 metric. Both Subchallenges 2A and 2B use the same scoring metric. This metric is the mean of two different correlation measures between the predicted coclustering matrix (CCM^{Pr}) and the true coclustering matrix (CCM^{Tr}); the AUPR curve and the AJSD of the coassignment probabilities. CCM^{Tr} is computed from the true SNV assignments to lineages using the procedure described in the previous section under the description of Subchallenge 2B. CCM^{Pr} for Subchallenge 2A is also computed using this procedure.

Each correlation measure, calculated by comparing CCM^{Pr} to CCM^{Tr}, and is normalized, by subtracting a constant value and linearly scaling, to be between 0 and 1. This normalization is computed so that 1 corresponds to a 'perfect score' that is, when $CCM^{\text{Pr}} = CCM^{\text{Tr}}$ and 0 corresponds to the smaller of the scores achieved by two 'bad scenarios': $CCM^{\text{Pr}} = I_{n \times n}$ or $CCM^{\text{Pr}} = I_{n \times n}$. If a method achieves a score <0 after normalization, then the score is set to zero. The overall Subchallenge 2 score is calculated as the mean of the two individual normalized correlation measures:

1. **AUPR**: The area under the receiver operating characteristic curve, also known as the precision-recall curve, which plots the false positive rate against the true positive rate across all possible thresholds for classifying matrix entries as true or false (for Subchallenges 2 and 3, all real values $r \in [0,1]$). To calculate the AUPR we create the precision-recall curve using the matrix values and then estimate the area under this curve using point estimators.
2. **AJSD of coassignment**: To define this correlation measure, we transform each *CCM* matrix so that each row could be interpreted as a discrete probability distribution. Then, for each row in the predicted *CCM*, we compute the Jensen–Shannon divergence between it and the corresponding row in the true *CCM* matrix. Our measure, the AJSD is the average of these divergences.

Specifically, for the predicted CCM matrix, C, for the i-th row, we define a real valued vector, $\mathbf{p}^i$, for each mutation i, whose j-th element, $\mathbf{p}^i_j = \frac{C_{ij}}{\sum_{k \ne i} C_{ik}}$ for

$i \neq j$ and $\mathbf{p}_i^i = 0$. Because of how $\mathbf{p}^i$ is defined, it can be interpreted as a discrete probability distribution over all of the SNVs in the sample. Similarly, for the actual CCM matrix, K, for the i-th row, we define $\mathbf{q}^i$, by setting $\mathbf{q}_j^i = \frac{K_{ij}}{\sum_{k \neq i} K_{ik}}$ for $i \neq j$, and $\mathbf{q}_j^i = 0$ otherwise. Then AJSD is the average across all rows i of the Jensen–Shannon divergence (JSD) between $\mathbf{p}^i$ and $\mathbf{q}^i$. To compute the JSD, to avoid taking the log of 0, we define $\mathbf{p}^{i*}$ as $\mathbf{p}_j^{i*} = \frac{(1-\alpha)\mathbf{p}_j^i + \alpha}{(1-\alpha)+N\alpha}$ for a small value $\alpha = 0.01$ and we define $\mathbf{m}_j^{i*}$ similarly and set $\mathbf{m}_j^{i*} = \frac{\mathbf{p}_j^{i*} + \mathbf{q}_j^{i*}}{2}$. And JSD is: $\mathrm{JSD}(\mathbf{p}^i, \mathbf{q}^i, \alpha) = KL(\mathbf{p}^{i*} || \mathbf{m}^{i*})/2 + KL(\mathbf{q}^{i*} || \mathbf{m}^{i*})/2$

Subchallenge 3 metric. To compute the Subchallenge 3 scoring measure, we require the *CCM* and *ADM* matrices as defined above and we must compute the cousin matrix (*CM*). The *CCM* and *ADM* matrices are either provided by the user or constructed from the true ancestral relationships. To construct the *CM*, we note that each mutation pair (*i,j*) must have one of four relationships: *i* is clustered with *j*, *i* is the ancestor of *j* (or vice versa), or *i* and *j* are in branching lineages (in other words, they are cousins). As such, given *ADM* and *CCM*, we compute the *CM* by setting $CM_{ij} = 1 - CCM_{ij} - ADM_{ij} - ADM_{ji}$.

Then, to compute the subchallenge score, we horizontally append the *CCM*, the *ADM*, the transpose of the *ADM*, and the *CM* for the true and predicted versions of these matrices, making two matrices of size *n* by 4*n*. In other words, one of these matrices is constructed from all of the true matrices and the other from all of the predicted ones.

We then compute the Pearson correlation coefficient (PCC) between these two rectangular matrices.

The PCC between two matrices *C* and *K* is defined as:

$$\mathrm{PCC} = \frac{\mathrm{Cov}(C, K)}{\sigma_C \sigma_K}$$

where Cov(*C*,*K*) is the covariance of the vectorized versions of *C* and *K*, σ_C is the standard deviation of vectorized *C*, and σ_K is the standard deviation of vectorized *K*.

Data preparation. To create our phase-separated mapping set, we used public data from the Genome-in-a-Bottle consortium obtained from sequencing the trio of individuals with Coriell IDs: GM24385 (son), GM24149 (father) and GM24143 (mother). We used both the high-coverage (300×) paired-end Illumina data and the low coverage (16×) 6 kb mate-pair Illumina data.

For the paired-end datasets, we downloaded the publicly available FASTQ files, and mapped them locally with bwa v.0.7.10 using the flag -M and otherwise default settings, against the *hs37d5* human reference with decoys. We marked duplicates with Picard (v.1.121). For the mate-pair datasets, we downloaded and used the publicly available mappings.

To identify variants, we used only the paired-end data for each sample, and a standard variant calling pipeline with GATK (v.2.4.9). The BAM files were realigned and calibrated using GATK's RealignerTargetCreator command, followed by IndelRealigner. Bases were recalibrated using the BaseRecalibrator and PrintReads commands. Germline calling was performed using UnifiedGenotyper and variant calls without the 'PASS' field were filtered out. Short indels and SNVs that were present in both maternal and paternal BAMs were used for phasing.

Phasing. First, we constructed an unphased set of variants using GATK-based germline single-nucleotide polymorphism prediction, identifying 2,559,193 diploid heterozygous short insertions, deletions and SNVs in the child sample. Next, we created the PhaseTools package to accurately phase heterozygous variants identified in these data (Supplementary Fig. 2 and Supplementary Note 2). This phasing prioritized connections between alleles that were directly supported by NGS data. Due to the availability of both paired-end and 6 kilobases (kb) mate-pair Illumina sequencing data for this sample, we were able to construct initial per-chromosome phase sets (that is, sets of heterozygous variants phased together) at a rate of one phase set per ~12 kb. The phasing was then extended by connecting phase sets using parent-of-origin information, in cases where this information could be computed by inspecting parental genotypes or parental NGS phasing. This increased the extent of our phase sets, decreasing their rate to one per ~76 kb. The phasing was extended once more by incorporating phasing information produced by Beagle, reaching an ultimate rate of one phase set per ~86 kb. We note that this long-range phasing could be obtained even without leveraging any long-read data. Remaining phase sets were then randomly rotated and collapsed to obtain a final complete phasing of all heterozygous variants in the child. Given the complete phasing of the variants described above, we used the BAM-phase-split program, also part of PhaseTools, to phase each fragment in an NGS dataset of the child sample. The program inspected the reads in each fragment, collecting information for which alleles that fragment supported at each heterozygous variant, and combined that information to phase the fragment. Fragments not spanning any heterozygous variants were phased randomly.

At the end of the process, while the median length of phased contigs from using only NGS data was ~15 kb regions, it increased to ~85 kb regions using the full PhaseTools pipeline.

Splitting BAM reads into subclones and spiking-in mutations. Read splitting at nodes occurs in a pseudo-random manner using a windowed approach. For each node, let *w* be every window of reads (set to 1,000) and *p* be the proportions of reads to extract. BAM files are sorted by coordinate using SAMtools sort. For every *w* paired reads ordered by first read-pair coordinate, exactly floor($w \times p$) paired reads are chosen at random and retained. As compared to a global resampling to the target coverage per node (that is, setting the window size to the total number of reads aligning to the chromosome), this local sampling accomplishes a less variable coverage across the final chromosome. All extracted reads are merged together using Picard tools, first by phase, then by chromosome, and finally into the tumor BAM. The merged BAM file is then sorted by coordinates, avoiding any possibility to identify from which subBAM reads originate.

To complete the final tumor BAM, we further normalize the phases of chromosomes relative to all the phases, based on their individual total fractional copies. For each phase of each chromosome, let $\mathbf{p}^i$ be the CP and $\mathbf{c}_i$ the number of copies at the *i*th leaf node. Then $C_{\mathrm{chr,phase}} = \mathrm{sum}_i\,(\mathbf{p}^i \times \mathbf{c}_i)$ represent the total fractional copies. Take *M* to be the maximum of all CNAs, including tandem duplications, across chromosomes and set this value as the 100% copy proportion. Leaf nodes are down-sampled by taking $C_{\mathrm{chr,phase}}/M$ of the read pool assigned to it. Read pools are adjusted using a bottom-up approach. At each internal node, the cellular copies of its children are summed and the read pool proportions are adjusted (Fig. 3).

```
designatePortions {
if leaf node:
return p_i * c_i / C_chr,phase
else:
quantities = []
quantity_sum = 0
for each child:
quantity[child] = designatePortions{config->child}
quantity_sum += quantity[child]
for each child:
config->child->read_proportion = quantity[child] / quantity_sum
}
```

If tandem duplications are present, reads that are not incorporated in a node (surplus reads) are down-sampled similarly to provide donor BAMs at the right depth. Surplus reads are down-sampled in proportion to their depth-adjusted copy number for a given node, starting with the highest copy number duplications for each node to yield the maximum depth donor bam for each node. If lower copy number duplications exist, these donor BAMs are subsequently down-sampled again in proportion to copy number to yield the lower copy number donor BAMs.

After calculating the per-phase-per-chromosome read pools, BAMSurgeon spikes in mutations given a set number of SNVs, Indels and SVs into the appropriate read pool before merging them into the final BAM. In Supplementary Note 2 we describe how we spike in mutations compatible with replicating timing, pre-defined trinucleotide context spectra and selection.

Altogether, using this approach we achieved a median accuracy of 90.6%, with a median false positive rate of 4.5% and a median false negative rate of 5.92% for the five tumors reported after calling SNVs with MuTect before down-sampling.

Large-scale SV simulations. We extended BAMSurgeon to simulate large SVs by simulating two SV breakpoints with local alignment and contig assembly. We used a two-pronged approach to simulate copy number changes as the existing BAMSurgeon functionality could not reliably simulate SVs larger than 30 kb (Supplementary Fig. 2f–h). To simulate smaller scale copy number changes (>10 kb) we extended the BAMSurgeon SV framework to simulate translocations, inversions, deletions, and duplications of arbitrary size (Supplementary Note 2). To simulate chromosome-level CNAs, we locally down-sampled reads.

Chromosome-level copy number simulations. A gain of N_a chromosomes from a given node a is simulated by first splitting the reads in a evenly into $N_a + N_b$ (where N_b is the number of chromosomes in the parent of *a*) while down-sampling the reads in all other nodes by $N_a + N_b$. Since each node is handled individually, a deletion of a copy is simulated by elimination of a node. Before any node split or phase gain, intermediate BAM files are sorted by read name using SAMtools sort -n. And before any spike-in mutations, intermediate BAM files are sorted by coordinate using SAMtools sort. After deriving the BAMs for each copy of that chromosome, BAMsurgeon is used to spike in all SNVs, Indels and SVs into both copies (simulating that these mutations precede the copy number event).

Subclonal copy number calling. We used Battenberg[6] based on allele-specific copy number analysis of tumors equations[51] to call subclonal copy number and validated the calls by comparing observed and expected logR and BAF of the identified segments as well as inferred versus expected CCF of the mutations (Fig. 4 and Supplementary Fig. 2).

Somatic mutation variant calling. To assess the SNVs spiked into the simulated tumor, we used four commonly used somatic SNV detection pipelines, as well as perfect calls. We first obtained perfect calls from BAMSurgeon as a gold-standard. We retained all SNVs with at least one alternate read, one reference read, and a minimum of three total reads covering the site to maximize sensitivity while excluding zero or near-zero depth SNVs. We then executed SomaticSniper (v.1.0.5), Strelka (v.1.0.17)

with the default settings. We executed MutationSeq (v.4.3.8) with a SNV threshold of 0.5, indel threshold of 0.1 and divided chromosomes into three intervals of at least 100 megabase pairs (Mb) and otherwise used the entire chromosome. We retained MutationSeq SNVs with a somatic mutation probability > 0.8, which passed all filters. Last, we used MuTect to call variants using the protocol described above. Similarly, we verified structural variants were present using Manta (v.0.29.5)

Subclonal reconstruction and scoring using PhyloWGS and DPClust. We used PhyloWGS (https://github.com/morrislab/phylowgs commit 3e21cec) with default settings (except for including all SNVs), and converted the output to an SMC-Het compatible format using a custom script (https://github.com/morrislab/smchet-challenge/tree/master/create-smchet-report commit 06a1f1f). We used DPClust (https://github.com/Wedge-Oxford/dpclust_smchet_docker commit a1ef254) with default settings, but added functions to parse SNVs from unsupported somatic SNV detection algorithms (https://github.com/Wedge-Oxford/dpclust_smchet_docker/blob/design_paper/dpc.R commit 1d8c2e7). For all somatic SNV detection algorithms we set the allele with the highest read count in the normal as the reference. We removed the sex chromosomes from both SNV and CNA inputs before running PhyloWGS and DPClust.

We then scored results from both algorithms using the scoring framework described above (https://github.com/asalcedo31/SMC-Het_Scoring/smc_het_eval commit 8b072a2). As the scale of scores for Subchallenges 1C, 2A, 2B, 3A and 3B depend on the mutation set used, solutions across depths and somatic SNV detection algorithms for a given tumor needed to be based on a common set of mutations to be comparable. We added all false and true SNVs called by all other somatic SNV detection algorithms for that tumor to each solution as a single zero cellularity cluster so that all solutions for that tumor contained the union of all SNVs. Additionally, to ensure scores among tumors were comparable, we scaled all scores to the highest scoring 128× perfect SNV call solution for that tumor and capped at 1. We then analyzed the Subchallenge 1A, 1C, 2A and 2B scores using β-regressions with the betareg R package[52]. As 1B scores represent true proportions, we analyzed them using a generalized linear model with a binomial link function. All models used T2, 128×, perfect, DPC, full depth as a reference. Interaction terms were retained for a given model if they reduced its Akaike information criterion and significantly increased log-likelihood of the model in a log-likelihood test comparing models with and without an interaction. See the attached Nature Research Reporting Summary for further information on the statistical analysis.

Effect of copy number calling accuracy on the reconstruction. We also assessed the effect of different copy number calling errors on the reconstruction scores (Fig. 6). To this end, we randomly selected copy number segments from the profiles and changed the copy number states to reflect different types of error (additional gains, losses and a mix of the two).

For gains, for each selected segment the number of copies of the major allele N_{maj} was added {0,1,2,3,4,5} with probabilities {0.01, 0.15, 0.40, 0.25, 0.15, 0.04}, respectively. The minor allele was randomly assigned a state between 0 and N_{maj}. For losses, for each selected segment N_{maj} was subtracted {0,1,2} with probabilities {0.06, 0.63, 0.31}, respectively. N_{min} was randomly selected between 0 and N_{maj} then the ceiling was taken. For the mix scenario, for each selected segment, N_{maj} is replaced by {0,1,2,3,4,5} with probabilities {0.01, 0.15, 0.40, 0.25, 0.15, 0.04}, respectively. N_{min} is randomly and uniformly selected between 0 and N_{maj}.

In each scenario, we increased the proportion of selected segments from 10 to 50% of all segments by 10% increments. We then executed DPClust and PhyloWGS with these copy number call errors and correct copy number calls on the five synthetic tumors for the depth-SNV somatic SNV detection algorithms combinations described above (4,250 combinations total). To reduce computation time, we down-sampled each input file in the Variant Call Format to 5,000 SNVs. We then carried out scoring and analysis for each reconstruction as described above.

Data visualization. Figures were generated using R (v.3.5.3), BPG (v.5.9.8)[53], lattice (v.0.20–38), latticeExtra (v.0.6–28), gridExtra (v.2.3), gtable (v.0.2.0) and Inkscape (v.0.91). Color palettes were generated using the RColorBrewer (v.1.1–2) and BPG packages.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequences files are available at EGA under study accession no. EGAD00001003971.

Code availability

BAMSurgeon is available at: https://github.com/adamewing/bamsurgeon. The framework for subclonal mutation simulation is available at http://search.cpan.org/~boutroslb/NGS-Tools-BAMSurgeon-v1.0.0/. The PhaseTools BAM phasing toolkit is available at https://github.com/mateidavid/phase-tools. Scripts providing the complete scoring harness are available at: https://github.com/asalcedo31/SMC-Het_Scoring/smc_het_eval.

References

51. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *PNAS* **107**, 16910–16915 (2010).
52. Cribari-Neto, F. & Zeileis, A. Beta regression in R. *J. Stat. Soft.* **34**, 1–24 (2010).
53. P'ng, C. et al. BPG: seamless, automated and interactive visualization of scientific data. *BMC Bioinform.* **20**, 42 (2019).

Acknowledgements

We thank the members of their laboratories for support, and Sage Bionetworks and the DREAM Challenge organization for their ongoing support of the SMC-Het Challenge. In particular, we thank T. Norman, J.C. Bare, S. Friend and G. Stolovitzky for their patience, technical support and scientific insight. We also thank R. Sun and C. Curtis for kindly sharing code for calculating the intra-tumor heterogeneity metrics and building the support vector machine predictor in multi-region sequencing simulations. This study was conducted with the support of the Ontario Institute for Cancer Research to P.C.B. and J.T.S. through funding provided by the Government of Ontario. This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation (Grant no. RS2014-01 to P.C.B.). This study was conducted with the support of Movember funds through Prostate Cancer Canada and with the additional support of the Ontario Institute for Cancer Research, funded by the Government of Ontario. This project was supported by Genome Canada through a Large-Scale Applied Project contract to P.C.B., S.P. Shah and R.D. Morin. This work was supported by the Discovery Frontiers: Advancing Big Data Science in Genomics Research program, which is jointly funded by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research (CIHR), Genome Canada and the Canada Foundation for Innovation (CFI). Q.M. is a Canada CIFAR AI chair and is supported by an Associate Investigator award from OICR. This research is part of the University of Toronto's Medicine by Design initiative, which receives funding from the Canada First Research Excellence Fund (CFREF). J.A.W. was partially supported by an Ontario Graduate Scholarship. This work was supported by The Francis Crick Institute, which receives its core funding from Cancer Research UK (grant no. FC001202), the UK Medical Research Council (grant no. FC001202), and the Wellcome Trust (grant no. FC001202). M.T. is a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Sklodowska-Curie Grant Agreement no. 747852-SIOMICS). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support toward the establishment of The Francis Crick Institute. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the UK Medical Research Council (grant no. MR/L016311/1 to M.T. and P.V.L.). A.S. was partly supported by a CIHR CGS-doctoral award. P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. D.C.W. is supported by the Li Ka Shing foundation. The Galaxy portions of the evaluation system were supported by National Institutes of Health (NIH) grant nos. U41 HG006620 and R01 AI134384-01 as well as NSF grant no. 1661497. The following NIH grants supported this work: no. R01-CA180778 (to J.M.S.), no. U24-CA143858 (to J.M.S.) and no. P30-CA008748 (to Thompson, subgrant to Q.M.). We thank Google Inc. (in particular N. Deflaux) for their ongoing support of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge. This work was supported by the NIH/NCI under award no. P30CA016042.

Author contributions

All authors edited and approved the final manuscript. A.S. wrote the first draft of the paper, designed experiments, performed statistical analyses, performed bioinformatics analyses and performed data visualization. M.T. wrote the first draft of the paper, designed experiments, generated tools and reagents, performed statistical analyses, performed bioinformatics analyses and performed data visualization. S.M.G.E. wrote the first draft of the paper, generated tools and reagents, performed bioinformatics analyses and performed data visualization. A.G.D. wrote the first draft of the paper, designed experiments, generated tools and reagents and performed bioinformatics analyses. M.D., S.D., L.Y.L., S.S., H.Z. J.M.C., A.B., C.M.L., I.U. and B.L. generated tools and reagents. K.Z. and T.-H.O.Y. generated tools and reagents and performed bioinformatics analyses. A.D.E. generated tools and reagents and supervised research. N.M.W. performed bioinformatics analyses and performed data visualization. J.A.W., M.K., H.Z. and C.V.A. performed bioinformatics analyses. C.P. performed data visualization. J.T.S., J.M.S., D.A. and Y.G. supervised research. K.E. wrote the first draft of the paper and supervised research. D.C.W. designed experiments and supervised research. Q.M. wrote the first draft of the paper, designed experiments, generated tools and reagents and supervised research. P.V.L. wrote the first draft of the paper, designed experiments and supervised research. P.C.B. wrote the first draft of the paper, designed experiments and supervised research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0364-z.

Correspondence and requests for materials should be addressed to P.C.B.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): PBoutros@mednet.ucla.edu

Last updated by author(s): Oct 16, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- [ ] [x] The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
- [ ] [x] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [ ] [x] The statistical test(s) used AND whether they are one- or two-sided
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- [ ] [x] A description of all covariates tested
- [ ] [x] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [ ] [x] A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [ ] [x] For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
  Give P values as exact values whenever suitable.
- [x] [ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [x] [ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [ ] [x] Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to download the data used in this publication.
Data analysis	All our custom code were deposited in public repositories. BAMSurgeon is available at: https://github.com/adamewing/bamsurgeon. The framework for subclonal mutation simulation is available at: http://search.cpan.org/~boutroslb/NGS-Tools-BAMSurgeonv1.0.0/. The PhaseTools BAM phasing toolkit is available at https://github.com/mateidavid/phase-tools. Scripts providing the complete scoring harness are available at: https://github.com/asalcedo31/SMC-Het_Scoring. R, lattice, latticeExtra, gridExtra, gtable and betareg are available through https://cran.r-project.org. BPG is available at https://labs.oicr.on.ca/boutros-lab/software/bpg .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequences files are publicly available at EGA under study accession number EGAS00001002092. No raw data is shown, Figure 4 and Figure 5 show data analyses based on simulated BAM files.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For our initial simulated tumors, we opted for a grid design covering five tumour types , five sequencing depths, 2 copy number runs (at low and high depths), five sets of variant calls including variant calls from four published callers, and two well-known clustering algorithms, for a total of 500 combinations and reconstructed subclonal structures. We varied these parameters on one real tumour yielding an additional 80 combinations. For our analysis examining the effects of copy number on algorithm performance we expanded our grid search to include 17 copy number error levels that span a range of error classes and intensities for a total 4250 CNA-variant caller-depth-tumour-algorithm combinations. Thus, the sample size was chosen to cover a carefully selected set of parameters and answer bespoke questions (e.g. the effect of depths on reconstruction accuracy across mutation callers and copy number errors) rather than based on predetermined expected effect sizes.
Data exclusions	Since these were simulation-based analyses, no data were excluded, as we ensured all of them passed our quality checks.
Replication	Due to the nature of simulation-based studies, for which the truth is known, we implemented quality checks and systematic comparisons to the truth to assess the reproducibility and validity of the results. We compared results from synthetic tumours to a real tumour which replicated our findings.
Randomization	We opted for a grid design in our simulations, hence did not randomize the parameters or simulations but rather chose to cover every combination of the pre-defined grid. We opted to cover tumour types presenting with a wide range of copy number alterations and mutational spectrum. In the simulations presented here, we covered five of the cancer types with highest incidence in the US, with low to high number of single nucleotide variants and number of copy number events.
Blinding	All variant callers and clustering algorithms were blind to the truth and did not train on the simulated data. Only the copy number caller served as a basis to assess the quality and accuracy of the simulations and was tuned to be able to run on low-depth simulations (8X).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging