Creating Standards for Evaluating Tumour Subclonal Reconstruction

Adriana Salcedo[1,2,*], Maxime Tarabichi[3,4,*], Shadrielle Melijah G. Espiritu[1,*], Amit G. Deshwar[5,*], 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], DREAM SMC-Het Participants, Jared T. Simpson[1], Joshua M. Stuart[13], Dimitris Anastassiou[8,9,10,14], Yuanfang Guan[7,15,16], Adam D. Ewing[11,17], Kyle Ellrott[11,12,#], David C. Wedge[18,19,#], Quaid D. Morris[6,#], Peter Van Loo[3,20,#], Paul C. Boutros[1,2,21,#]

[1] Ontario Institute for Cancer Research, Toronto, Canada

[2] Department of Medical Biophysics, University of Toronto, Toronto, Canada

[3] The Francis Crick Institute, London, United Kingdom

[4] Wellcome Trust Sanger Institute, Hinxton, United Kingdom

[5] The Edward S. Rogers Sr. Department of Electrical & Computer Engineering

[6] Department of Computer Science, University of Toronto, Toronto, Canada

[7] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA, 48109

[8] Department of Systems Biology, Columbia University, New York, New York, USA

[9] Center for Cancer Systems Therapeutics, Columbia University, New York, New York, USA

[10] Department of Electrical Engineering, Columbia University, New York, New York, 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] Department of Biomolecular Engineering, Center for Biomolecular Sciences and Engineering, University of California, Santa Cruz; Santa Cruz, CA, USA

[14] Herbert Irving Comprehensive Cancer Center, Columbia University, New York, USA

[15] Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA.

[16] Department of Electronic Engineer and Computer Science, University of Michigan, Ann Arbor, Michigan, USA

[17] Mater Research Institute, University of Queensland, Woolloongabba, Queensland, Australia.

[18] Big Data Institute, University of Oxford, Oxford, United Kingdom

[19] Oxford NIHR Biomedical Research Centre, Oxford, United Kingdom

[20] Department of Human Genetics, University of Leuven, Leuven, Belgium

[21] Department of Pharmacology and Toxicology, University of Toronto, Toronto, Canada

[*] These authors contributed equally

[#] These authors jointly directed the work

Abstract

Tumours evolve through time and space. To infer these evolutionary dynamics for DNA sequencing data, many subclonal reconstruction techniques have been developed and applied to large datasets. Surprisingly, though, there has been no systematic evaluation of these methods, in part due to the complexity of the mathematical and biological questions and the difficulties in creating gold-standards. To fill this gap, we systematically elucidated key algorithmic problems in subclonal reconstruction, and developed mathematically valid quantitative metrics for evaluating them. We then developed approaches to simulate realistic tumour genomes that harbour all known mutation types and processes. Finally, we benchmarked a set of 500 subclonal reconstructions, creating a key resource, and quantified the impact of sequencing read-depth and somatic variant detection strategies on the accuracy of specific subclonal reconstruction approaches. Inference of tumour phylogenies is rapidly becoming standard practice in cancer genome analysis, and this work sets standards for evaluating its accuracy.

Introduction

Most tumours arise from a single ancestral cell, whose genome has accumulated somatic driver mutations[1,2], giving it a fitness advantage over its neighbours by, for example, manifesting some of the hallmark characteristics of cancers[3]. This ancestral cell and its descendants proliferate, ultimately giving rise to all cancerous cells within the tumour. Over time, tumour cells accumulate mutations, in some cases giving rise to further fitness advantage and leading to local clonal expansions that result in subpopulations of tumour cells sharing subsets of mutations, which we term "subclones". As the tumour extends spatially beyond its initial site and around the body, spatio-genomic variability will arise as different regions harbour tumour cells with distinctive genetic characteristics[4–8].

DNA sequencing of tumours allows quantification of the frequency of a specific mutation within a tumour, based on measurements of the fraction of mutant sequencing reads, the copy number state of the locus and the tumour purity[9,10]. By aggregating these noisy frequency measurements across mutations within each subclone, a tumour's sample subclonal architecture can be reconstructed from bulk sequencing data[6,10]. Subclonal reconstruction methods have proliferated rapidly in recent years[11–14], and have already revealed key characteristics of tumour evolution[4,7,15–19], spread[20–22], and response to therapy[23,24]. However, to date, 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 for truth. While single-cell sequencing could theoretically provide ground truth, existing single-cell datasets do not provide sufficient depth and breadth to adequately assess subclonal reconstruction methods. Further, single-cell sequencing has distinctive and pervasive error profiles[25]. Alternative gold-standard datasets may be generated using simulations. However, existing tumour simulation methods like BAMSurgeon[26] neither create representative subclonal populations nor phase simulated variants, as required by some methods[6,9]. 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 in reconstruction is identification of the mutations present in each subclonal lineage. Metrics are needed that penalize errors in both the number of subclonal lineages and the placement of mutations across them. Third, subclonal reconstruction methods are recent developments, and few groups have equal expertise with multiple tools. Rather, algorithm developers themselves are experts in parameterizing their own algorithm. Thus, an unbiased third-party is needed to fairly compare the strengths and weaknesses of different methods.

To fill this gap, we developed a crowd-sourced benchmarking Challenge: The ICGC-TCGA DREAM Somatic Mutation Calling Tumour Heterogeneity Challenge (SMC-Het).

Challenge organisers simulated realistic tumours, developed robust scoring metrics and created a computational framework to facilitate unbiased method evaluation. Challenge participants then created re-distributable software images representing their methods. These images were run in an automated pipeline on a series of test tumours to evaluate the accuracy of their subclonal reconstructions. Here, we describe the creation of quantitative metrics for scoring tumour subclonality reconstructions. We then outline novel tools for constructing simulated tumours with realistic subclonal architecture. Finally, we characterise the sensitivity of subclonal reconstruction methodologies to somatic mutation callers and technical artefacts.

Results

How should subclonal reconstruction methods be evaluated?

Subclonal reconstruction is a complex procedure that involves estimating many attributes of the tumour, including its purity, number of lineages, and the phylogenetic relationships between lineages. We structured our evaluation of these attributes into three categories, which comprise the five sub-challenges of SMC-Het (**Figure 1**). Sub-challenges 1 a, b, and c (SC1) quantify the ability of an algorithm to reconstruct global characteristics of tumour composition. Specifically, they evaluate each algorithm's predictions of the total fraction of tumour cells or purity of the sample (SC1a), the number of subclonal lineages (SC1b), and the fraction of tumour cells or cellular prevalence and number of mutations associated with each subclone (SC1c). By contrast, sub-challenge 2 (SC2) 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 (SC2a) and soft assignments represented through co-clustering frequencies (*i.e.* the probability that two SNVs are in the same lineage) (SC2b). Finally, sub-challenge 3 (SC3) evaluates the ability of algorithms to recover the phylogenetic relationships between subclonal lineages, again both from a single best hard assignment (SC3a) and based on soft assignments (SC3b). Taken together, these define seven specific outputs based on which subclonal reconstruction methods can be benchmarked.

To quantify the accuracy of these seven outputs, we evaluated several candidate scoring metrics. We required each candidate metric to range from zero (very poor performance) to one (perfect performance). Appropriate metrics for SC1 were trivially identified (**Online Methods**), but SC2 and SC3 required us to test and modify existing metrics, and to develop new ones. As SC2 and SC3 involve assigning mutations to subclonal lineages, we required candidate metrics to satisfy three conditions[20]:

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.

Note further that because SC2b and SC3b are based on pairwise probabilities of co-clustering, we were unable to use clustering quality metrics, such as normalised mutual information (NMI, also known as the V-measure[27]), that require hard clustering or explicit estimation of the number of clusters.

Metrics for evaluating cluster assignments have a number of desirable properties[27]. We identified a set of these properties applicable to our task (**Online methods**), used a simulation framework to assess how well each metric satisfies these properties and identified four complementary metrics that satisfy all three properties: Matthew's Correlation Coefficient (MCC), Pearson's Correlation Coefficient (PCC), area under the precision recall curve (AUPR) and average Jensen-Shannon divergence (AJSD; **Supplementary Figure 1**). To further refine this set, we tested their behaviour relative to different types of subclonal reconstruction errors, such as inversion or merging of individual nodes. We assessed six error cases for SC2 and 27 for SC3 (**Supplementary Table 1**), simulating each and scoring them with each candidate metric (**Figure 2**). For SC2, no individual metric ranked the errors in the expected order. To address this, we defined a composite metric, ψ, as the arithmetic mean of the AJSD, MCC and PCC, as these three metrics complemented one another (**Figure 2**), and their mean had both a near-optimal ranking and satisfied our three main requirements (**Supplementary Figure 1**). For SC3, we calculated the Spearman's correlation between the ideal ranking and the metric ranking, and identified AJSD as the best approach (mean Spearman's ρ = 94.11 95% CI: 93.9-94.3; **Supplementary Table 2**).

Finally, for both SC2 and SC3, we scaled individual scoring metrics to [0,1] by computing an affine transform (*i.e.* a scale and an offset) so that the highest possible value receives a score of one. To set a baseline score of zero, we created two naive reconstructions: 1) One-cluster**,** all mutations are in a single subclone (*i.e.* cluster) and 2) N-clusters in a star phylogeny, *i.e.* each mutation is its own cluster and the clusters are all mutually exclusive one of each other. The worst-scoring of these two possibilities was set as the baseline score of zero for SC2 and 0.5 for SC3. We set a higher baseline for SC3 as SC3 penalises phylogenetic errors as well as co-clustering errors leading to uniformly low scores that would hinder interpretation and downstream analysis. Any negative scores achieved by contestants after scaling are also set to zero.

Simulating realistic subclonal tumour genomes

We elected to use simulated tumour data to run SMC-Het. The key reasons were the unavailability of deep single-cell sequencing data as a gold-standard dataset, the lack of single-cell sequencing data that match arbitrary tree structures and characteristics, the ability to simulate a large number of tumours at low-cost, and the demonstrated ability of tumour simulations to recapitulate sequencing error profiles. We elected to use the BAMSurgeon tool created for the SMC-DNA Challenges[26,28], which creates tumours with accurate SNVs, indels and small genomic rearrangements at varying allelic fractions. However, this tool lacked a number of key features associated with tumour evolution, and therefore five major features were added: (1) the phasing of variants, (2) whole-chromosome and whole-genome copy number changes, (3) translocations, (4) trinucleotide signature injection, and (5) simulation of replication-timing effects (**Figures 3-4**). We describe each of these briefly below.

Phasing of mutations. To properly simulate a tumour it is critical that genetic variation of all types - both somatic and germline - are fully phased, as they are in real tumours. Because it relied solely on short-read sequencing, BAMSurgeon was unable to do this, so reconstruction of subclonal events did not yield the original tree structure (data not shown). To correctly phase all mutations, it is necessary to phase each read, *i.e.* determine which of the two homologous copies of each autosome it derives from. To achieve this, we leveraged NGS data from a trio of individuals from the Genome-in-a-Bottle consortium (**Supplementary Figure 2)**. First, we constructed an unphased set of variants using GATK-based germline SNP prediction, identifying 2,559,193 diploid heterozygous short insertions, deletions, and single nucleotide variants in the child sample. Next, we created the PhaseTools package to accurately phase these heterozygous variants. This phasing prioritised connections between alleles that were directly supported by NGS data. Due to the availability of both paired-end and 6 kbp mate-pair Illumina sequencing data for this sample, we were able to construct initial per-chromosome phase sets (*i.e.* sets of heterozygous variants phased together) at a rate of 1 phase set per ~12 kbp. 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 1 per ~76 kbp. The phasing was extended once more by incorporating phasing information produced by Beagle, reaching an ultimate rate of 1 phase set per ~86 kbp. 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 in order to phase the fragment. Fragments not spanning any heterozygous variants were phased randomly. The final result of this process is two BAM files per chromosome, each representing a single phase. In male patients, we phased only the two well-known pseudo-autosomal regions (PAR1 and PAR2) that are homologous between chromosomes X and Y – one phase set was kept in chromosome X while the other was re-mapped onto chromosome Y.

Whole arm & whole genome copy number changes. Once we had fully phased the SNPs in the genome, the next step to create accurate simulations was to allow changes in copy number of entire chromosomes and whole-genome ploidy changes (*e.g.* whole genome duplications present in 30-50% of human cancers[29–31]). To accomplish this, we developed a method to account for copy number changes, both gains and losses, for each chromosome, including sex chromosomes (**Figure 3**). Given a tumour design structure, the phased genomes were split further into individual subpopulations (leaf nodes) that make up the tumour population. We assigned a virtual number of reads, which we term pseudoreads, to each node based on the cellular prevalence of the node, gaining and losing reads as necessary to represent the copy number gains and losses. The proportions of reads were normalised based on the total number of pseudoreads. If at a leaf node, multiple DNA copies of a given genomic region existed, the reads were split evenly among the copies, and BAMSurgeon was used to spike mutations into each leaf node. The extracted reads were merged to generate the final tumour BAM file that had a logR profile consistent with the design (**Figure 4**).

Translocations. Translocations are a critical type of oncogenic mutation, which was not included in the SMC-DNA simulated data challenges[28]. To address this gap, we developed a new approach. For two regions (named A and B), an unbalanced translocation is simulated by selecting reads aligned to region A and reads aligned to region B and assembling contigs for each set of reads. To control for contig mis-assembly, each contig is aligned to the reference genome using exonerate[32], any unaligned portion at the ends is trimmed, and reads corresponding to the trimmed portion(s) of the contigs are de-selected. The contig break-ends are then fused either head-to-tail or head-to-head depending on user specification. Read coverage is generated over the fused contigs using *wgsim*[33]. Finally, altered reads are re-aligned to the reference genome and used to replace reads in the original BAM file based on read name, creating a simulated translocation that accurately reflects the expected pattern of discordant read pair mappings and split reads.

Trinucleotide mutation profile and replication timing. Single nucleotide mutations in cancer are not uniformly distributed throughout the genome. Rather, they are biased both regionally and locally. We have added the capability to BAMSurgeon to simulate the most common mutational biases of each type: trinucleotide signatures and replication-timing bias. Mutations result from specific mutagenic stresses, each of which leads to particular mutation types that occur at specific trinucleotide contexts, and may have a different mutation rate to other mutational processes[34]. Replication-timing bias refers to the increase in the mutation rate of regions of the genome that replicate late in the cell cycle[35]. We generated an extensible approach (**Online Methods**, **Figure 4**) that weights each nucleotide in the genome according to its trinucleotide context, replication timing, and the set of mutational signatures and then samples bases from the genome until the expected trinucleotide spectrum is reached. BAMSurgeon can handle arbitrary mutational signatures and arbitrary replication timing data at any resolution, and indeed this can be generalised to any type of location bias in mutational profiles.

Legacy and reproducibility

To maximise reproducibility, for each SMC-Het entry made, participants were required to submit a working copy of their approach, which could then be applied to new data. This was accomplished using two technologies: Docker[36] and Galaxy[37]. Docker is a technology for packaging a piece of software and all of its dependencies into a single container that can easily be moved and redeployed on new systems. To describe how the program should be invoked, and how the different steps of computation fit within the evaluation framework, the participants also included Galaxy tool wrappers and workflows. To enable development, participants were given a pre-built virtual machine image that could be deployed in their own Google project space. Once they were able to run a workflow on the test data, they could run a submission script that packaged the Docker image, Galaxy tool wrappers and workflow, and uploaded the package to Synapse for evaluation[38]. Workflows were then run on a set of held out samples.

General features of subclonal reconstruction

To confirm that the simulated tumours accurately reflected real tumours and to demonstrate that our scoring framework could identify factors known to impact subclonal reconstruction, we simulated, reconstructed and evaluated five tumours with a range of depths and somatic mutation callers. These five tumours were derived from different tissue types (prostate, lung, CLL, breast and colon) and all had previously described subclonal structures (**Supplementary Figure 3)**. We then explored the sensitivity of subclonal reconstruction to both sequencing coverage and to the variant-calling pipeline used. For sequencing coverage, we downsampled each tumour to create a titration series in raw read-depth of 8x, 16x, 32x, 64x and 128x coverage - this resulted in 25 tumour-depth combinations. For each of these, we then identified subclonal copy number aberrations (CNAs) using Battenberg[6], both with downsampled tumours and at the highest possible depth, yielding 50 tumour-depth-CNA combinations. We identified somatic SNVs using four detection tools (Mutect[39], SomaticSniper[40], Strelka[41], and MutationSeq[42]) as well as the perfect (spiked-in) somatic SNV calls, yielding 250 tumour-depth-CNA-SNV combinations. Subclonal reconstruction was then carried out on each of these using two algorithms (PhyloWGS and DPClust), to give a final set of 500 tumour-depth-CNA-SNV-subclonal reconstruction algorithm combinations, which were evaluated using the scoring framework described above (**Supplementary Figure 4**, **Supplementary Table 3**).

Figure 5 shows the results of this large-scale benchmarking on SC1c (cellular prevalence of subclones) and SC2a (mutational profile of subclones). The top heatmap shows scores for high-depth CNA detection and the bottom scatterplot shows the score as a function of effective read-depth (*i.e.* number of reads per tumour chromosome, after adjusting for purity and ploidy). For SC1c, all algorithms showed a consistent decrease with depth (**Figure 5a, c**). As expected, there was significant sensitivity to the somatic SNV caller, with the perfect calls outperforming those from any algorithm by a significant margin (β = 0.29, P = 0.013, generalised linear models). By contrast, the use of high-depth *vs.* low-depth sequencing for subclonal CNA detection had almost no influence on the reconstruction accuracy (P> 0.05 for all sub-challenges, generalised linear model; **Supplementary Tables 4-10**). Similarly, both PhyloWGS and DPClust performed very well, and essentially interchangeably on this question (**Supplementary Figure 5, Supplementary Tables 4-10**).

This general profile of algorithm performance was mirrored for all sub-challenges with two exceptions, which outline differences between DPClust and PhyloWGS. In SC1a, DPClust, which uses purity measures derived from CNA reconstructions, showed a significant advantage over PhyloWGS, which uses purity measures partially dependent on SNV clustering. The latter are more sensitive to errors in VAF due to low sequencing depth and this is reflected in the pattern of SC1a scores. In SC2B, PhyloWGS, which uses a phylogenetically-aware clustering model, had significantly better performance than DPClust, which uses a flat clustering model (**Supplementary Figure 5**). Thus, our metrics are sensitive to differences in modelling approaches, which manifest in variability in performance on different aspects of subclonal reconstruction.

All methods seemed to perform poorly on SC2a - identifying the mutational profiles of individual subclones (**Figure 5b,d**). Here, we saw major inter-tumour differences in performance, with tumour T2 having the least accurate reconstructions and T4 the most (**Supplementary Figure 6**). This in part reflects the higher purity of T4, and indeed we see a strong association between effective read-depth and reconstruction accuracy, with each doubling in read-depth increasing reconstruction score by about 0.1. At effective read-depths above 60x, all tumour-CNA-SNV-subclonal reconstruction combinations performed well, suggesting that a range of approaches can be effective for detection of subclonal mutational profiles. There remained a strong dependence of accuracy on the SNV detection pipeline, with perfect calls out-performing the best individual caller (MuTect) by ~0.05 at any given read-depth. Broadly, SomaticSniper and Strelka showed similar performance, but interestingly showed significant tumour-by-caller interactions in generalised linear modelling for several sub-challenges (**Supplementary Figure 5**). This may reflect tumour-specific variability in their error profiles. As in SC1c, neither the use of low- *vs.* high-depth tumours for CNA detection nor the specific subclonal reconstruction algorithm used significantly influenced the accuracy of subclonal reconstruction. Taken together, these data suggest both that subclonal reconstruction accuracy is highly sensitive to upstream SNV detection approaches, and that there is significant room for algorithmic improvements that capture inter-tumour differences, build on prior distributions of phylogenies and better model the error characteristics of upstream feature-detection pipelines.

Discussion

Increasingly large numbers of tumours receive genomic interrogation each year, and the extent of this interrogation grows as DNA sequencing costs diminish. While panel sequencing is ubiquitous today, whole genome sequencing will eventually achieve similar penetration. Nevertheless, it remains most common for just a single, spatially distinct region of a cancer to be sequenced in any such study. The reasons for this are many: the increase in costs with the number of tumour regions sequenced, the need to preserve tumour tissue for future clinical use and the increasing use of scarce biopsy-derived specimens for sequencing in the diagnostic and metastatic settings. While robust subclonal reconstruction from multi-region sequencing is well-known[5–8], the ability to accurately reconstruct evolutionary properties of tumours from single-region would open major new avenues for linking these to clinical features of tumours.

We describe here a framework for evaluating such subclonal reconstruction methods, comprising a novel way of scoring the accuracy of relevant biological features of their outputs, a technique for robustly phasing short-read sequencing data, an enhanced read-level simulator of tumour genomes with realistic biological properties and a portable software framework in which multiple subclonal reconstruction algorithms can be rapidly executed in a consistent and predictable way. These features, each implemented in open-source software and reusable on their own, form an integrated system that allows identification of key algorithmic features of subclonal reconstruction. We use them to generate a titration-series that will serve as a key community resource for evaluating algorithm sensitivity to specific parameters. From this titration series, we quantify the sensitivity of subclonal reconstruction to both effective read depth and to the characteristics of specific somatic SNV detection pipelines. These data give key guidance for improving cancer genomics for subclonal reconstruction: increasing effective read-depth above 60x, after controlling for tumour purity and ploidy, enables accurate inference of multiple key evolutionary features from a single sample. They also provide new avenues for algorithm developers, highlighting the interactions of variant callers with specific tumour phylogenies, and the association of variant calling accuracy with subclonal reconstruction accuracy.

In many areas of biology, ground-truth is either inaccessible or impractical to measure with precision. In cases like these, simulations are extremely valuable in providing a lower bound on error profiles and an upper bound on the accuracy of methods. By incorporating all currently known features of a phenomenon, simulators codify our understanding and the divergence between simulated results and real ones provides a quantitation of 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.

Moving forward, large-scale benchmarking of multiple subclonal reconstruction methods using this framework on larger numbers of tumours is needed to create a gold-standard. Such a benchmark would not only inform algorithm users, who will benefit from an understanding of the specific error profiles of different methods, but also algorithm developers, who will be able to update and improve methods, while ensuring software portability. Tumour simulation frameworks provide a valuable way for method benchmarking, and can help complement other approaches, like comparison of single-region and multi-region subclonal reconstruction and the use of model organism and sample-mixing experiments.

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. *bioRxiv* 132324 (2017). doi:10.1101/132324
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. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30,** 413–421 (2012).
10. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb. Perspect. Med.* **7,** (2017).

11. 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).
12. Deshwar, A. G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16,** 35 (2015).
13. 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).
14. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11,** 396–398 (2014).
15. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** 751–759 (2015).
16. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346,** 251–256 (2014).
17. Turajlic, S. *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* **173,** 595-610.e11 (2018).
18. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. *Cell* **0,** (2018).
19. Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0086-z
20. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520,** 353–357 (2015).

21. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48,** 758–767 (2016).
22. Turajlic, S. *et al.* Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell* **173,** 581-594.e12 (2018).
23. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5,** 2997 (2014).
24. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526,** 525–530 (2015).
25. Van Loo, P. & Voet, T. Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.* **24,** 82–91 (2014).
26. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12,** 623–630 (2015).
27. Rosenberg, A. & Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. in *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic* 410–420 (2007).
28. Lee, A. Y.-W. *et al.* Combining accurate tumour genome simulation with crowd sourcing to benchmark somatic structural variant detection. *bioRxiv* 224733 (2017). doi:10.1101/224733
29. Cheng, J. *et al.* Pan-cancer analysis of homozygous deletions in primary tumours uncovers rare tumour suppressors. *Nat. Commun.* **8,** 1221 (2017).

30. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2016).

31. Storchova, Z. & Kuffer, C. The consequences of tetraploidy and aneuploidy. *J. Cell Sci.* **121,** 3859–3866 (2008).

32. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).

33. Li, H. *wgsim: Reads simulator*. (2018).

34. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

35. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

36. O'Connor, B. D. *et al.* The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Research* **6,** 52 (2017).

37. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44,** W3–W10 (2016).

38. Omberg, L. *et al.* Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat. Genet.* **45,** 1121–1126 (2013).

39. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

40. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinforma. Oxf. Engl.* **28,** 311–317 (2012).

41. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28,** 1811–1817 (2012).

42. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinforma. Oxf. Engl.* **28,** 167–175 (2012).

Accession Codes

Sequences files are available at EGA under study accession number EGAS00001002092. 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/Sage-Bionetworks/SMC-Het-Challenge.

Acknowledgements

The authors thank the members of their labs for support, and Sage Bionetworks and the DREAM Challenge organization for their ongoing support of the SMC-Het Challenge. In particular, they thank T. Norman, J.C. Bare, S. Friend and G. Stolovitzky for their patience, technical support and scientific insight. 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 #RS2014-01. 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 (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), Genome Canada and the Canada Foundation for Innovation (CFI). This work was supported by an NSERC operating grant grant and a gift from the NVIDIA Foundation, an advised fund of Silicon Valley Community Foundation, to Q.D.M. 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 (FC001202), the UK Medical Research Council (FC001202), and the Wellcome Trust (FC001202). M.T. is a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant Agreement No. 747852-SIOMICS). P.V.L. is a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. 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 NIH Grants U41 HG006620 and R01 AI134384-01 as well as NSF Grant 1661497. The following NIH grants supported this work: R01-CA180778 (J.M.S.) and U24-CA143858 (J.M.S.). The authors thank Google Inc. (in particular N. Deflaux) for their ongoing support of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge.

Author contributions

All authors: Edited & approved final manuscript

A.S. performed statistical analyses, performed bioinformatics analyses, performed data visualisation. M.T. generated tools & reagents, performed statistical analyses, performed bioinformatics analyses, performed data visualisation. S.M.G.E. generated tools & reagents, performed bioinformatics analyses, performed data visualisation. A.G.D. generated tools & reagents, performed bioinformatics analyses. M.D. generated tools & reagents. S.D. generated tools & reagents. L.Y.L. generated tools & reagents. S.S. generated tools & reagents. H.Z. generated tools & reagents. K.Z. generated tools & reagents, performed bioinformatics analyses. T.O.Y. generated tools & reagents, performed bioinformatics analyses. J.M.C. generated tools & reagents. A.B. generated tools & reagents. C.M.L. generated tools & reagents. I.U. generated tools & reagents. B.L. generated tools & reagents. A.D.E. generated tools & reagents, supervised research. NMW performed bioinformatics analyses, performed data visualisation. J.A.W. performed bioinformatics analyses. M.K.H.Z. performed bioinformatics analyses. C.V.A. performed bioinformatics analyses. C.P. performed data visualisation. J.T.S. supervised research. J.M.S. supervised research. D.A. supervised research. Y.G. supervised research. K.E. supervised research. D.C.W. supervised research. Q.D.M. generated tools & reagents, supervised research. P.V.L. supervised research. P.C.B. supervised research.

Figure Legends

Figure 1 | Features of tumour subclonal reconstruction

Overview of the key performance aspects of subclonal reconstruction algorithms, grouped into three broad areas covered by three key questions: (SC1) ‘What is the composition of the tumour?’ This involves quantifying its purity, the number of subclones, and their prevalence and mutation loads; (SC2) ‘What are the mutational characteristics of each subclone?’ This can be answered both with a point-estimate and a probability profile, *i.e.* a hard or probabilistic assignments of mutations to subclones, respectively; (SC3) ‘What is the evolutionary relationships amongst tumour subclones?’ This again can be answered with both a point-estimate and a probability profile.

Figure 2 | Quantifying the performance of subclonal reconstruction algorithms

(a) Eight of the 27 possible metric error cases used to assess how well metrics reflect expert opinion of subclonal reconstruction error ordered from most to least severe. **(b,d)** Scores resulting from the candidate metrics for SC2A (b) and SC2B (d) for error cases observable through the co-clustering of mutations (without yet inferring any phylogenetic relationships). **(c)** Scores resulting from the candidate metrics for SC3A (c) and SC3B (e), regarding the inference of phylogenetic relationships, for each error case shown in (a). All considered metrics converge on the same score for 3A after normalization.

Figure 3 | Simulating subclonal CNAs in Tumour BAM files

Example case of read number adjustment to simulate subclonal copy number aberrations (CNAs). (a) Desired structure of the tumour being simulated. (b) The first tumour 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. (c) Read number adjustment calculations. The copy number total (CNT) for each chromosome is its copy number by adjusted by node cellular prevalence summed across all nodes. The maximum CNT across the genome is retained to normalise 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 cellular prevalence, the chromosome’s copy number, and the total tumour depth normalised by the maximum CNT. (c) After adjusting read number for CNAs in each node and adding additional mutations, BAMSurgeon merges the extracted reads into a final BAM file.

Figure 4 | Simulating realistic tumour genomes

To create BAM files that accurately mirror those from real human tumours, we expanded the BAMSurgeon framework. We used *Genome-In-A-Bottle* data to provide a high-coverage normal for simulation, and then developed PhaseTools - an approach to phase short-read sequencing data. Panel (1) outlines the increased length of phased contigs from using only NGS data (median ~15 kbp regions) to using the full PhaseTools pipeline (~85 kbp regions). Next, we expanded BAMSurgeon to handle changes in chromosome number, with mutational changes before and after these ploidy changes. Panel (2) gives an exemplar of this behaviour, showing the logR ratio of different tumour subclones as simulated chromosomes are lost and gained. Finally, we enhanced the simulation of SNVs to allow for trinucleotide mutational signatures and replication timing effects. Panel (3) illustrates how the simulated composite trinucleotide signature (bottom) matches the design (top).

Figure 5 | Error profiles of subclonal reconstruction algorithms

To identify general features of subclonal reconstruction algorithms, we created a set of tumour-depth-CNA-SNV-subclonal reconstruction algorithm combinations by using the framework outlined in Figure 3 and 4 to simulate five tumours 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 Figure 2, creating a dataset to explore general features of subclonal reconstruction methods. All scores are normalised to the score of the best performing algorithm when using perfect calls at the full tumour depth. Scores exceeding this baseline likely represent noise or overfitting and were capped at 1. a) For SC1C (identification of the number of subclones and their cellular prevalence), all combinations of methods perform well. b) c) By contrast, for SC2a (detection of the mutational characteristics of individual subclones), there is large inter-tumour variability in performance. (c) Score for SC1C (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 SC2A show significant changes in performance as a function of effective read-depth.

Supplementary Figure Legends

Supplementary Figure 1

(a) The score for each candidate 2A metric considered with an increasing proportion of mutations assigned to the wrong useful clusters. **(b)** The score for each candidate metric considered with an increasing proportion of mutations in noise clusters. **(c)** The score for each candidate metric considered as the number of predicted clusters increases. The true number of clusters (four) is marked by the vertical line. Excess clusters retain correct co-clustering and are subsets of the true clusters. **(d)** For each potential scoring metric, the proportion of simulation runs that satisfied each of the four desirable metric properties for a given simulation parameter setting. Each property is tested by fixing all but one of the simulation parameters and then looking at the effect of changing the fourth parameter on the metric score.

Supplementary Figure 2

Example of the PhaseTools algorithm constructing an extended phase set from four heterozygous sites by leveraging NGS and parent phasing. **(a)** ngs_phasing of 5 heterozygous sites(hets) in the child and the corresponding nsg-phased sites in the mother and father, shown with informative NGS fragments. Hets boxed together represent phase sets. There is not enough information to construct a single phase set. **(b)** parent_base phasing uses parental genotypes to assign parent of origin to the 5 hets in the child. Hets 2 and 5 remain unresolved while hets 1, 3, and 4 show at least one unambiguous parent of origin. **(c)** parent_ngs phasing extends parent_base phasing with parental NGS fragments from ngs_phasing. The linked NGS fragment in sites 2 and 3 (T, T) of the maternal genotype is not informative as site 3 is homozygous, however the linked NGS fragment in sites 2 and 3 (T and A) of the paternal genotype is heterozygous and therefore informative. The phasing proposed by ngs_father of sites 3 and 4 (GG/AC) contradicts parent of origin information in hets 3 and 4 (A and G). This event is recognised as a pre-meiosis recombination event in the child and the ngs_father phasing is ignored. **(d)** ngs+parent_ngs phasing extends ngs_child phasing with parent_ ngs, giving priority to ngs_child phasing. NGS fragments such as hets 2 and 3 (T and T) take precedent over any phasing assigned by parent_ngs phasing see hets 2 and 3 (C and T) and indicate probable recombination events (shown with diagonal lines). Two possible sets of recombination events are shown. The proximity between phased hets determines which recombination events are most probable. Here, the recombination events shown on the right are selected, as recombination between sites 1 and 2 is more likely than recombination between sites 3 and 4, as sites 1 and 2 are further apart. The final phase sets are shown. **(e)** Schematic of phase-set reconstruction. Priority is given to procedures on the left.

Supplementary Figure 3

True subclonal structures of the simulated tumours (**T2**, **T3**, **T4**, **T5**, and **T6**) that were simulated with their desired and observed variant allele frequency histograms and logR profiles. In each panel, we show the phylogenetic tree, inspired by published reconstructed tumours, and the mutations associated with each (sub)clone. The top figures compared expected cancer cell fractions of the SNVs under a diploid setting, against the inferred cancer cell fractions from the simulated data. T5, for which the inferred purity is off due to the limitations of the copy number caller to call subclonal whole genome duplication, shows an observed space that departs from the expected. The bottom figures compare the observed and expected BAF and logR of the genomic segments identified by the copy number caller.

Supplementary Figure 4

Subclonal reconstruction scores based on the five tumours with each variant caller-depth-algorithm combination. All scores are normalised to the score of the best performing algorithm using perfect calls at the full tumour depth. Scores exceeding this baseline likely represent noise or overfitting and were capped at 1**. (a)** Scores for 1A are uniformly high. **(b)** Scores for SC1B improve with depth and but not continuous as the metric reflects a true proportion. **(c,d)** Scores for SC2B (c) and SC3B(d) closely mirror those of SC2A and SC3A, respectively.

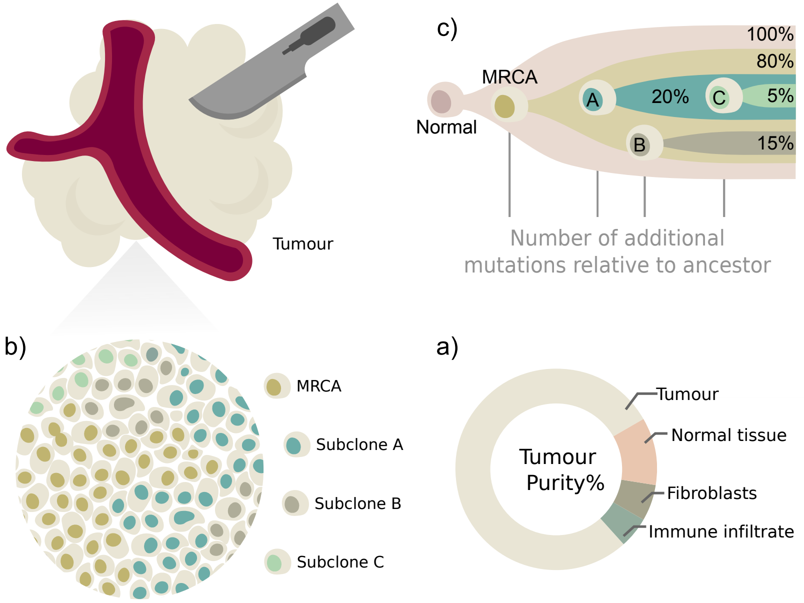
Supplementary Figure 5

Comparison of subclonal reconstruction scores for each sub-challenge using PhyloWGS (x-axis) and DPClust (y-axis). Variant callers are coded by colour and tumours are coded by symbol.

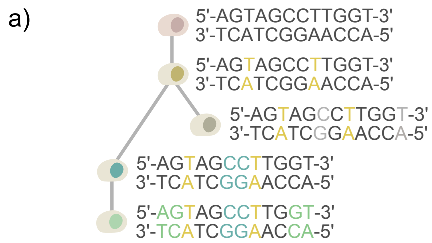
Supplementary Figure 6

SC2A score increases with effective depth for all tumours but the effect of the variant caller depends on the tumour. **(a)** T2 **(b)** T3 **(c)** T4 **(d)** T5 **(e)** T6.

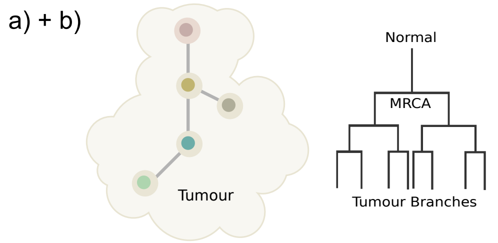
Question 1: Tumour Composition



Question 2: Mutational profiles



Question 3: Phylogeny



Question 1

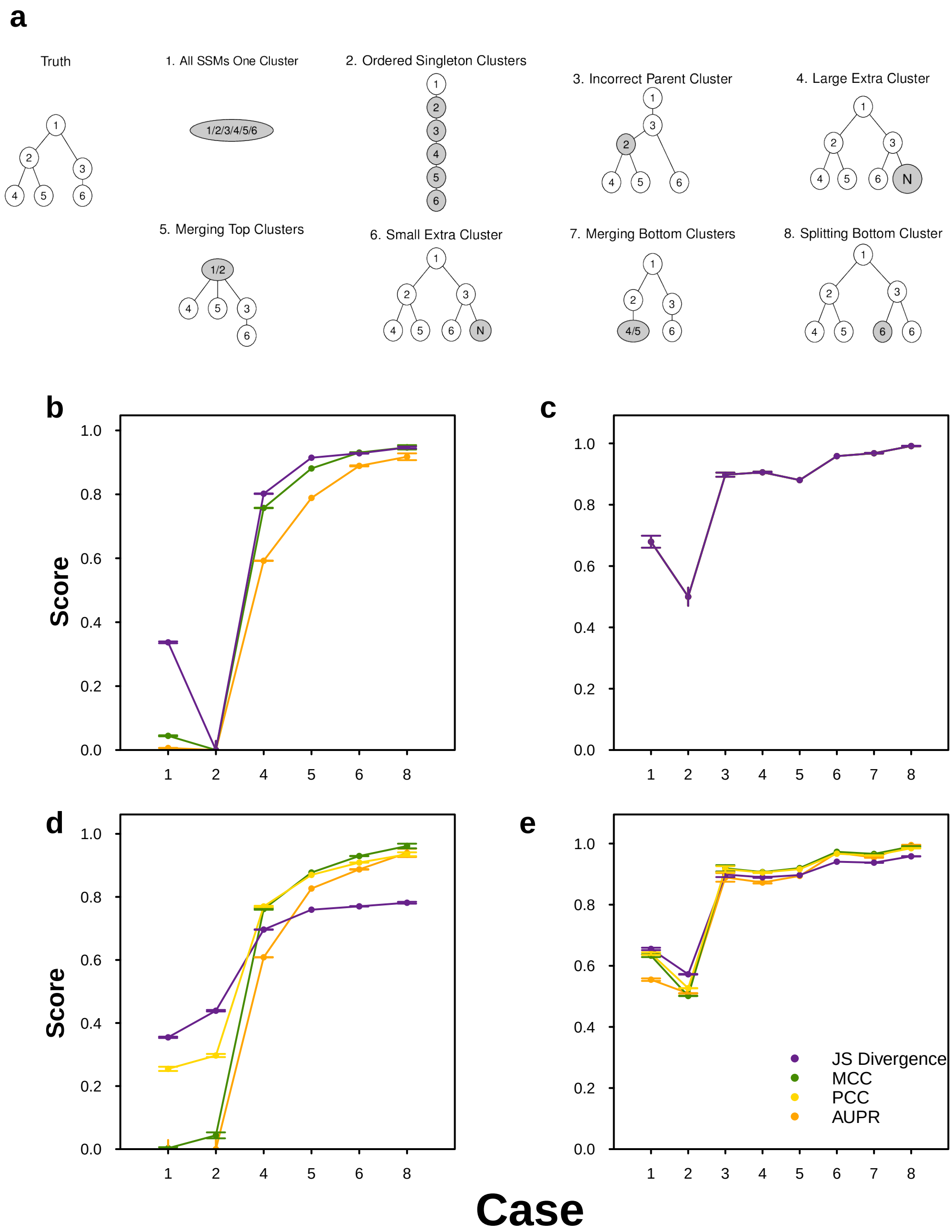
a) Purity of Tumour Sample
b) Number of Subclones
c) Subclonal architecture

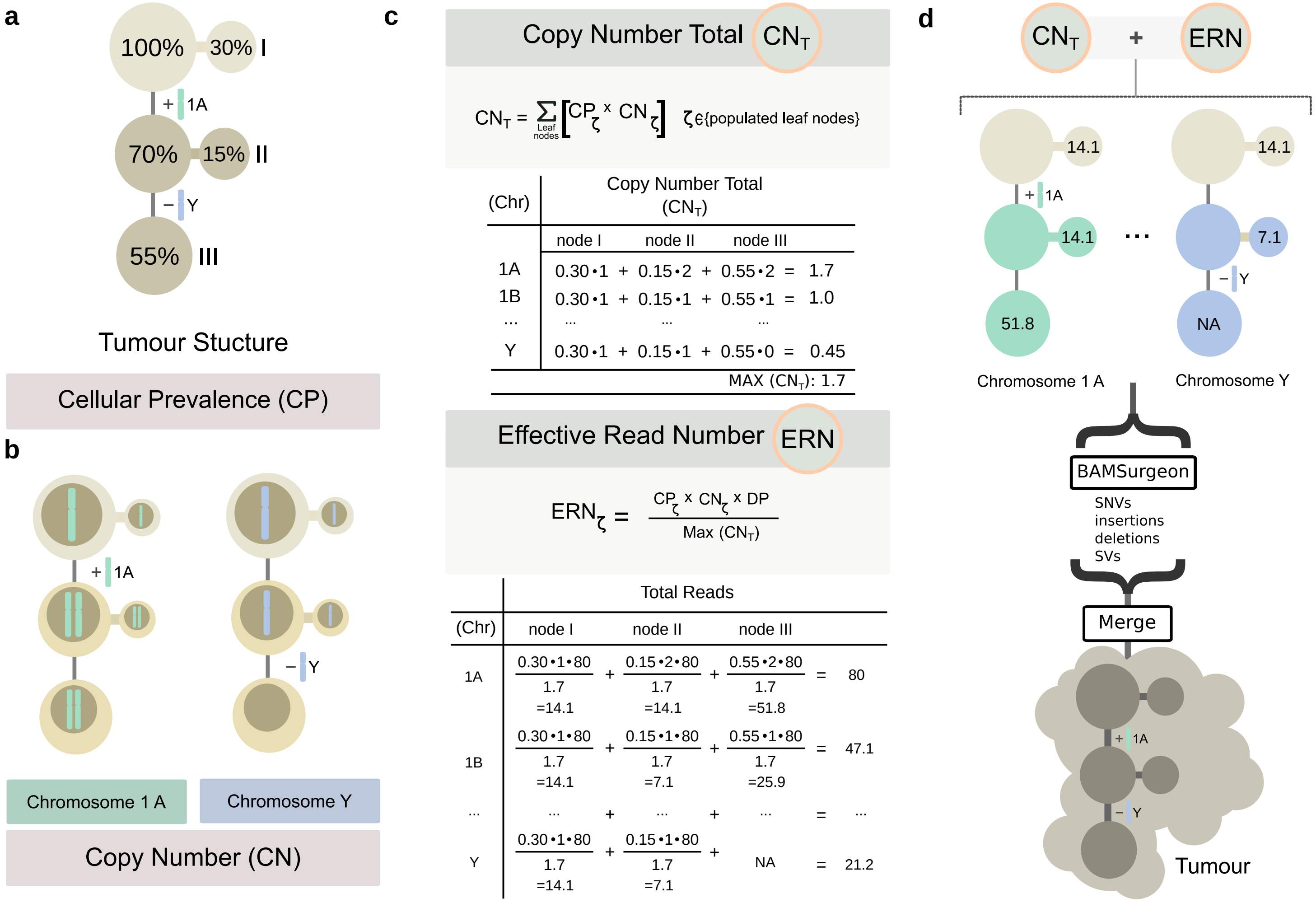
Question 2

a) Mutational Profile of Each Subclone
   i) Point Estimate
   ii) Probability Estimate

Question 3

a) Reconstructing Evolutionary Relationships
b) Ancestor-Descendant Covariance Matrix


a
Truth
1. All SSMs One Cluster
1/2/3/4/5/6
2. Ordered Singleton Clusters
3. Incorrect Parent Cluster
4. Large Extra Cluster
N
5. Merging Top Clusters
1/2
6. Small Extra Cluster
7. Merging Bottom Clusters
4/5
8. Splitting Bottom Cluster
b
c
d
e
Score
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
● JS Divergence
● MCC
● PCC
● AUPR
Case



$$CN_T = \sum_{\substack{\text{Leaf}\\ \text{nodes}}} \left[CP_\zeta \times CN_\zeta \right] \quad \zeta \in \{\text{populated leaf nodes}\}$$

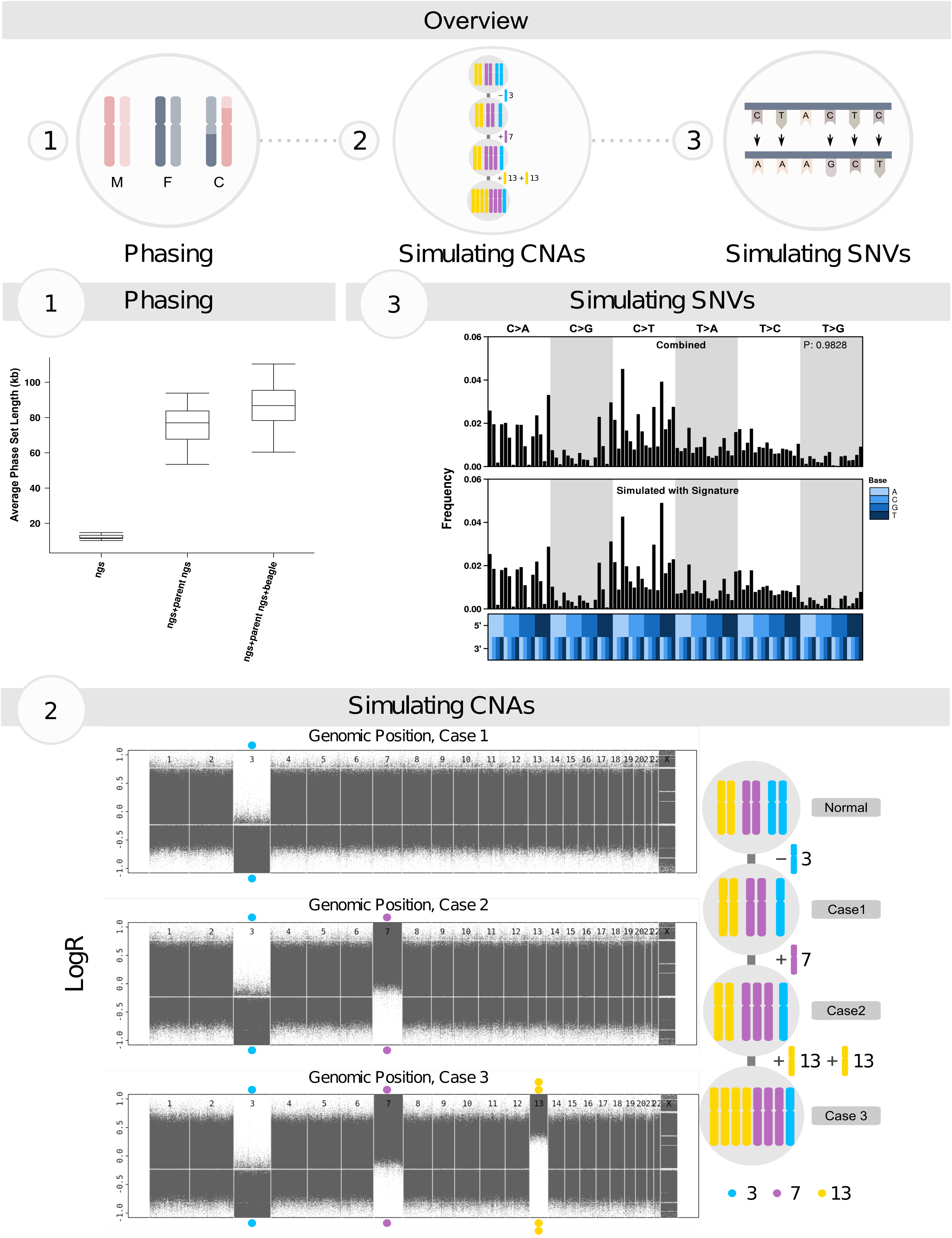
(Chr)	Copy Number Total (CN_T)						
	node I		node II		node III		
1A	0.30•1	+	0.15•2	+	0.55•2	=	1.7
1B	0.30•1	+	0.15•1	+	0.55•1	=	1.0
...	...		...		...		
Y	0.30•1	+	0.15•1	+	0.55•0	=	0.45

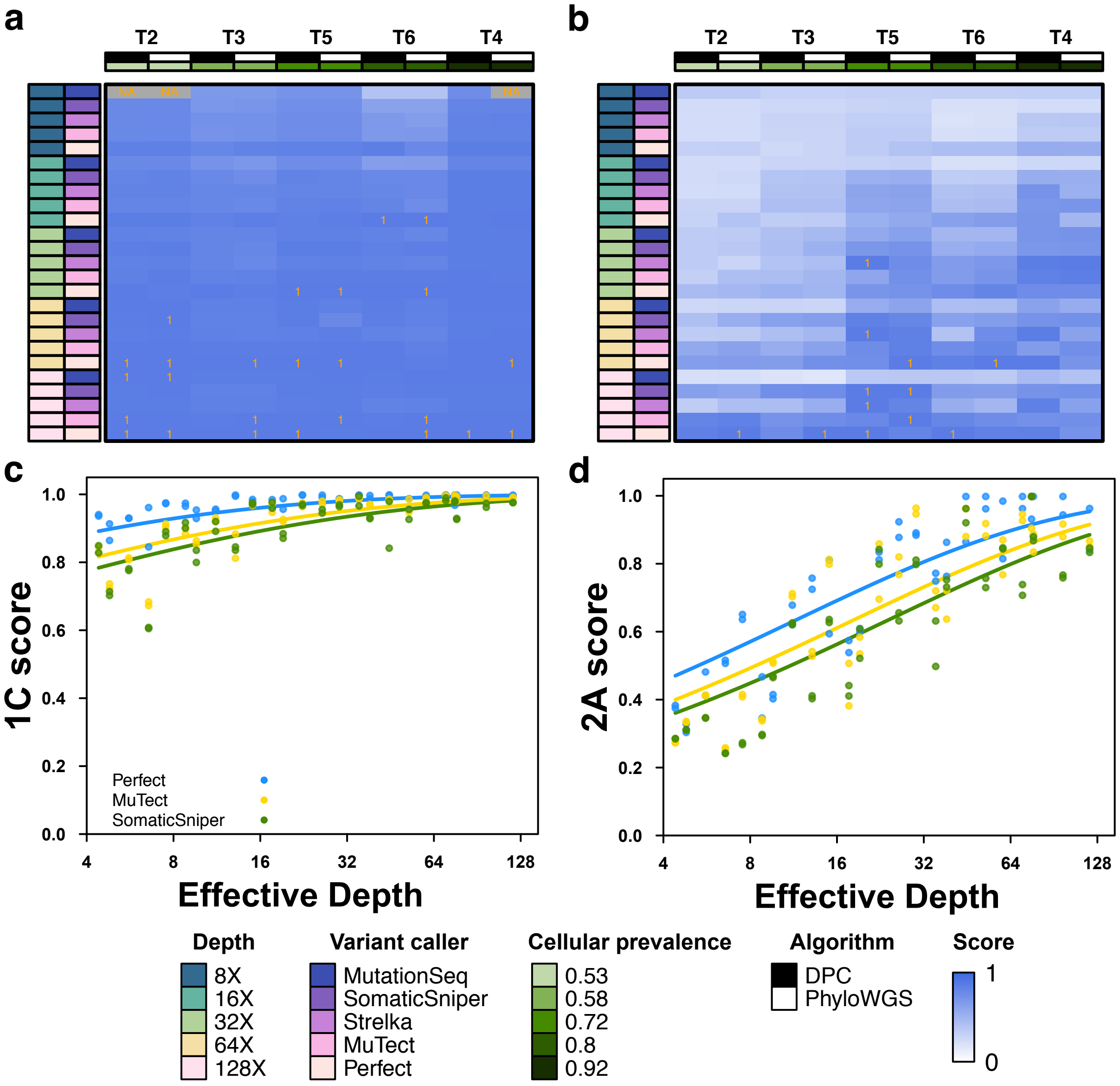
MAX (CN_T): 1.7

$$ERN_\zeta = \frac{CP_\zeta \times CN_\zeta \times DP}{\text{Max}\ (CN_T)}$$

	Total Reads						
(Chr)	node I		node II		node III		
1A	$\frac{0.30 \cdot 1 \cdot 80}{1.7}$ =14.1	+	$\frac{0.15 \cdot 2 \cdot 80}{1.7}$ =14.1	+	$\frac{0.55 \cdot 2 \cdot 80}{1.7}$ =51.8	=	80
1B	$\frac{0.30 \cdot 1 \cdot 80}{1.7}$ =14.1	+	$\frac{0.15 \cdot 1 \cdot 80}{1.7}$ =7.1	+	$\frac{0.55 \cdot 1 \cdot 80}{1.7}$ =25.9	=	47.1
...	...	+	...	+	...	=	...
Y	$\frac{0.30 \cdot 1 \cdot 80}{1.7}$ =14.1	+	$\frac{0.15 \cdot 1 \cdot 80}{1.7}$ =7.1	+	NA	=	21.2

CP = Cellular Prevalence CN = Copy Number CN_T = Copy Number Total ζ = Leaf Node DP = Read Depth ERN = Effective Read Number


Overview
1
M
F
C
Phasing
2
− 3
+ 7
+ 13 + 13
Simulating CNAs
3
C T A C T C
A A A G C T
Simulating SNVs
1
Phasing
Average Phase Set Length (kb)
20
40
60
80
100
ngs
ngs+parent ngs
ngs+parent ngs+beagle
3
Simulating SNVs
C>A
C>G
C>T
T>A
T>C
T>G
Combined
P: 0.9828
Simulated with Signature
Frequency
0.00
0.02
0.04
0.06
Base
A
C
G
T
5'
3'
2
Simulating CNAs
Genomic Position, Case 1
Genomic Position, Case 2
Genomic Position, Case 3
LogR
-1.0
-0.5
0.0
0.5
1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Normal
− 3
Case1
+ 7
Case2
+ 13 + 13
Case 3
3
7
13


a
T2
T3
T5
T6
T4
NA
NA
NA
1
b
T2
T3
T5
T6
T4
1
c
1C score
1.0
0.8
0.6
0.4
0.2
0.0
4
8
16
32
64
128
Effective Depth
Perfect ●
MuTect ●
SomaticSniper ●
d
2A score
1.0
0.8
0.6
0.4
0.2
0.0
4
8
16
32
64
128
Effective Depth
Depth
8X
16X
32X
64X
128X
Variant caller
MutationSeq
SomaticSniper
Strelka
MuTect
Perfect
Cellular prevalence
0.53
0.58
0.72
0.8
0.92
Algorithm
DPC
PhyloWGS
Score
1
0