Brief Communication

Tracing the evolution of aneuploid cancers from multiregional sequencing

Subhayan Chattopadhyay[1], Jenny Karlsson[1], Anders Valind[1,2], Natalie Andersson[1], David Gisselsson[1,3,4]

[1]Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden

[2]Department of Pediatrics, Skåne University Hospital, Lund, Sweden.

[3]Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden

[4]Clinical Genetics and Pathology, Laboratory Medicine, Lund University Hospital, Skåne Healthcare Region, Lund, Sweden

Corresponding Author Subhayan Chattopadhyay, Lund University, SE22184 Lund, Sweden. Phone: 467-369-78242; E-mail: subhayan.chattopadhyay@med.lu.se

Abstract

To understand the evolutionary dynamics of cancer, clonal deconvolution of mutational landscapes across multiple biopsies from the same patient is crucial. However, the frequencies of mutated alleles are often distorted by variation in copy number of mutated loci as well as the purity across samples. We present a semi-supervised algorithm that normalizes for purity and incorporates allelic composition with bulk sequencing to reliably segregate clonal/subclonal variants even at low sequencing depth (~50x). In presence of at least one tumor sample with >70% purity, it deconvolves samples down to ~40% purity, allowing robust tracking of mutated cell populations through cancer evolution.

Genetic diversification during tumorigenesis and disease progression is governed by Darwinian principles. Next generation sequencing across cancer types has confirmed that intratumor heterogeneity through phylogenetic branching is a common scenario[1], although the relative contributions from clonal selection versus neutral evolution in this process remain a matter of debate[2,3]. We recently demonstrated that intratumor heterogeneity can result as a product of different evolutionary trajectories specific to the spatiotemporal localization of resident tumor cells[4]. As a cancer's clonal landscape thus often varies with tumor geography, comprehensive reconstruction of tumor phylogeny requires multi-regional analysis and subclonal deconvolution[5,6]. Established bioinformatic tools for deconvolution are typically based on unsupervised clustering of the relative abundance of mutations across samples, represented by variant allele frequencies (VAFs)[7], determining subclonal populations with a distributional assumption on the variant read counts[8]. However, VAF is influenced by the purity of the specific sample and if variants reside in chromosomal regions affected by copy number changes. As copy number alterations can appear both clonally and subclonally, and may vary in kind for the same chromosome within a tumor, they can significantly complicate clonal deconvolution[9]. To mitigate this, we developed CRUST (**C**lonal **R**econstruction of t**U**mors with **S**patio-**T**emporal sampling), an analysis suite that parameterizes stochastic assumptions on the distribution of variants across samples, suggesting the most statistically probable clustering of mutations into clones and subclones, while integrating observed biological inference, on a case by case basis.

From a set of samples from the same tumor obtained at different locations and/or time points, CRUST deconvolves each sample separately by assigning each variant to a predicted clonal or sub-clonal status, calibrating clonality assignment against given parameters on allele-

specific copy number status and sample purity **(**see **Supplementary Methods Quick user guide)**. It realigns the frequency distribution across samples with probabilistic quotient normalization. Hereafter the distribution is queried to fit into an optimum number of clusters based on statistics comparing loss of information (**Supplementary Figure 1**).

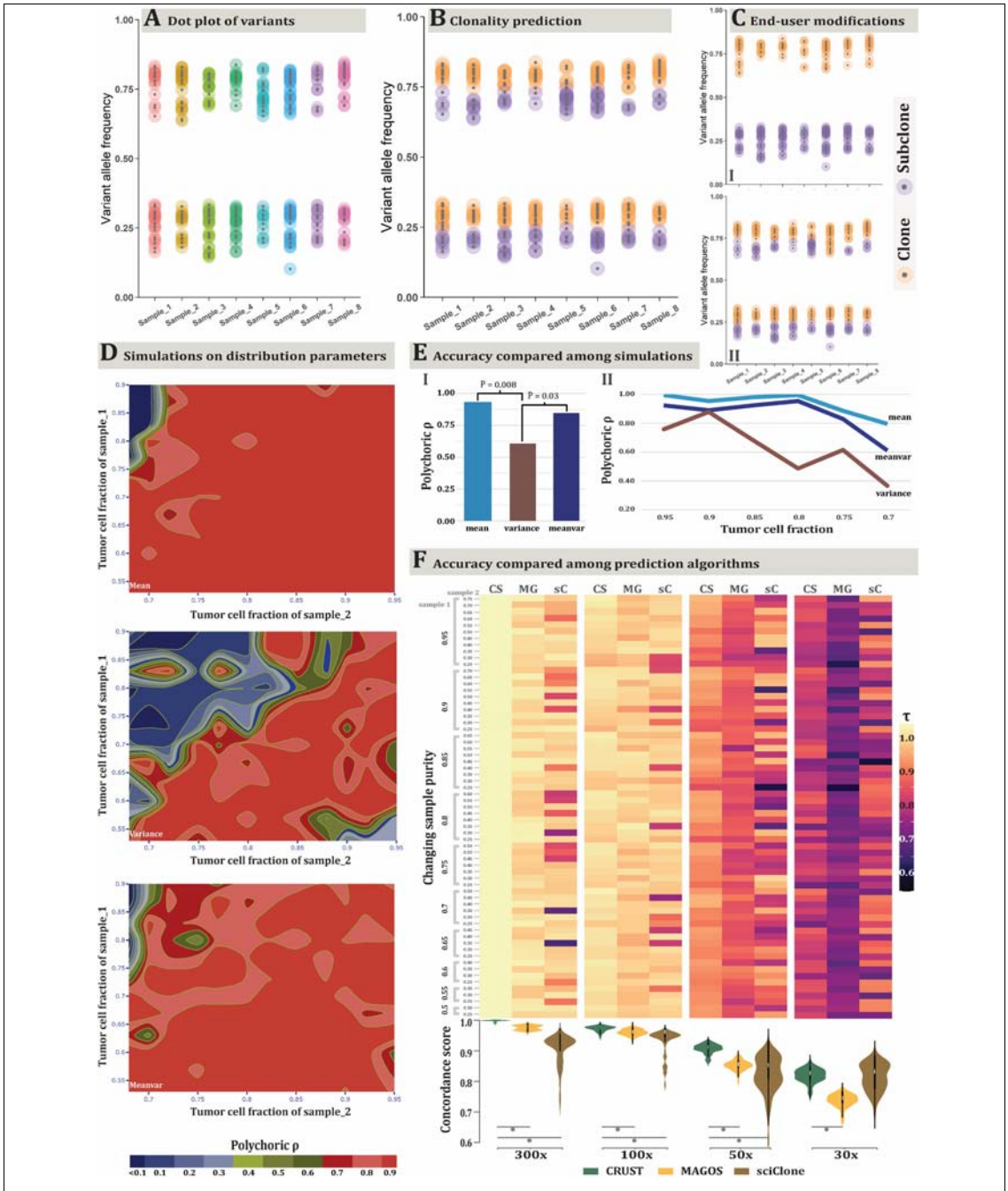

Figure 1. Clonal deconvolution of a simulated tumor genome. There are eight samples representing different biopsies (**A**). All samples here adhere to an allelic composition of 1+3. CRUST first displays a dot plot of the VAFs pertaining to all samples. Given provision for a purely estimation driven approach, it predicts clonality from the optimum number of clusters determined without supervision. This results in a deconvolution independent of the user suggested input (**B**). A

user can decide to opt for a semi-supervised approach instead if the optimum number of clusters predicted is dissimilar to a biologically expected deconvolution, for example prior knowledge from single cell karyotyping or sequencing. In this example the default optimization is given with 4 clusters as seen above (two clonal and two subclonal). In (**C I)** however, the user chooses to fit a 2-cluster deconvolution resulting in a prediction of one clonal and one subclonal cluster. The predictions can also be modulated post-hoc for individual samples **(C II)**. Over the default optimum prediction, for *Sample_6* a user has here chosen to fit a 3-cluster deconvolution that picks up two clones (at allelic compositions 1+3 and 3+1) and one sub-clone. In (**D**) simulations of scaling with varying sample composition are shown. Each iteration generates two samples, say X and Y with tumor cell fractions (TCF) Tx and Ty, respectively. Assuming Tx > Ty, CRUST rescales the variants in Y based on those in X. Simulations are performed to see how well the scaling works when Tx and Ty are varied. Three parametric beta-log normal models are in effect to generate simulated samples. The top panel shows changes in TCF that only affects the mean of the VAF distribution. The middle shows changes in TCF affecting the variance (ergo spread) of the VAF distribution and the lower most panel shows when it dynamically affects both mean and variance (referred as *Meanvar*). The measured statistic is polychoric correlation among predictions and its scale for all three simulations is the same, as is indicated at the bottom. In (**E I**), average marginal concordance is estimated with geometric mean for all three methods and tests are performed between pairs. Only significant deviations are marked with corresponding P values. In (**E II**) the trend of change in average concordance with varying levels of TCF between the three algorithms is depicted. A comparison across deconvolution methods was done with simulation of varying sequencing coverage **(F)**. Samples are drawn with varying TCF for four sets of coverage at 300x, 100x, 50x and 30x. Ordinal cluster similarities were assessed for CRUST (CS), MAGOS (MG) and sciClone (sC) with Jaccard coefficient (τ). The four combined heatmap and violin plots correspond to four coverage settings denoted in the x axis. Each combination represents summary statistics obtained as median τ for paired TCFs. Each cell in the heatmap reflects that obtained from a paired simulated sample denoted in the joint y axis TCF. The leftmost y axis annotation denotes TCF for sample 1 (Tx) and the inner annotation denotes that of the second sample (Ty). The highest Tx was 0.95 and the lowest was set at 0.5. For Ty, the highest by default was chosen to be 0.2 lower than that of the highest Tx, hence 0.75 and the lowest was set at 0.25. The violin plots are drawn correspondingly under the heatmaps on the lower panel denoting the dispersion and central tendency of the estimates with significant p values of the paired association tests marked by grey points.

In absence of available copy number data, CRUST can also assess allelic composition based on sequencing summaries from the constitutional genome (**Supplementary Figure 2)**. After copy number analysis, sequence variants from a single tumor are then analyzed separately for each allelic configuration (1+1, 1+2, 2+0 etc.), where CRUST visualizes the predicted clonal/subclonal assignments for all spatiotemporal samples (**Figure 1A, B)**. The subclonal estimation process is based on a semi-supervised cluster determination. It verifies the optimal solution first without user input; next the user is given opportunity to override the unsupervised solution after visual inspection of the expected subclonality (**Figure 1C)** to retain provision for a biologically derived deconvolution assessment, if needed. In addition, subclonality assignment can be altered for specific samples post-prediction, a feature useful in presence of compromising purity or inter-sample heterogeneity with respect to the complexity of chromosomal alterations (e.g. chromothripsis and whole genome doubling).

To assess the accuracy of CRUST-based deconvolution across varied purity and sequencing coverage, we simulated tumor samples (**Figure 1D, E; Supplementary Figure 3, Supplementary Methods**). Here, the frequency distribution of variants queried from low depth calls were left-tail heavy although the pure distribution is expected to follow a beta-binomial distribution. Extending from a one-parametric power law function[10], we modelled the reduction in variability biased towards the left tail with a log-exponent function, a lognormal prior. Samples were drawn from the closed form cumulative function upon which accurate predictions were observed for those with purity down to 40%, scaled against a complementary sample with purity of at least 70%. Additionally, CRUST predicted the simulated clonality status more accurately (statistically significant with two-tailed P value < 0.05, Mann-Whitney test) compared to contemporary algorithms for sequences with coverage of at least 50x **(Figure 1F)**[11,12].

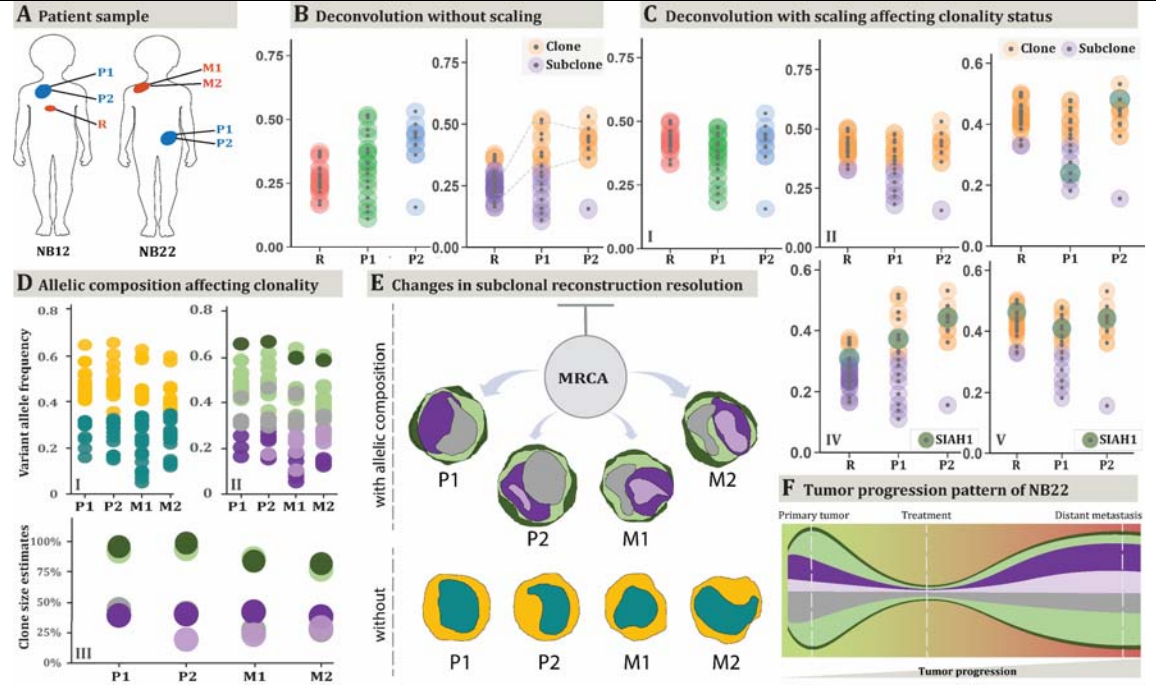

Figure 2 Multiple samples with varied purity from each of two neuroblastomas (NB) were used as examples (**A**). NB12 is represented with three samples, two primary tumor biopsies (P1, P2) and one relapse (R). The primary sample was ~90% pure whereas the relapse sample contains only 55% tumor cells as estimated by previous studies[4]. Hence, a deconvolution without rescaling the variant allele frequencies (VAFs) results in all shared variants (linked by grey lines between R and P1) being classified as subclonal in R (**B**, original sample specific VAFs are on the left, clonality predictions are on the right). Post-scaling (**C**), the relapsed variants re-adjust (**C I**), and the predictions reflect a reasonable nature of the clonality (**C II**). It is worth noting that in both analyses, the optimum cluster number is unchanged. This indicates that a traditional

subclonality reconstruction algorithm would fail to account for the noise in the relapsed sample if analyzed in conjunction with the primary samples. The next three panels demonstrate how scaling impacts the predictions. In panel (**C III**), an *ST8SIA2* mutation changes clonality status between P1 and P2, in concordance with a clonal sweep between these regions (see **Supplementary Figure 4**)[4]. Panels C-IV and V show a *SIAH1* exonic variant that is present in all three samples. In R, it is classified as subclonal while unscaled, but the prediction is overturned to be clonal post scaling. Deconvolution of the copy number aberrant neuroblastoma NB22 **(D)**, based on samples from the primary tumor (P1, P2) and a metastatic lesion (M1, M2). This tumor contained several copy number changes that required consideration for accurate deconvolution. CRUST was used to detect the segmental copy number alterations of all variants which were classified in two allelic composition make-ups, balanced 1+1 segments, and unbalanced 1+2 segments. These were deconvolved separately. Predicting clonality status without consideration of the copy number aberrations results in two predicted clusters **(D I)**, whereas considering allelic composition results in five clone/subclone clusters across all four samples **(D II).** This deconvolution would not have been possible without copy number data considered. Estimated clone sizes are depicted below with tumor cell fractions of each cluster (**D III)**. Inferring tumor evolution from deconvolution **(E-F)** shows how starting from an unknown MRCA (most recent common ancestor) one of the primary clones (in grey) shrinks whilst another subclone (in light purple) expanded at metastatic sites. Clone sizes estimated from set of variants with two different allelic compositions indicated a major clone size (1+1 in dark green and 1+2 in light green) of about 92% (mean) indicating the aberrations carried forward from a most recent common ancestor. The bottom panel in (**E)** devoid of copy number data lacks resolution to detect any such change.

As an example of the importance of scaling for correct deconvolution, we extracted from a published dataset on childhood cancer[4], three neuroblastoma tumor tissue samples from a patient with varied purity (55%-90% tumor cells), two from the primary tumor (NB12, P1 and P2) and one from metastatic relapse (R) (**Figure 2A**). Available copy number data and whole exome sequencing summaries were filtered for variants at a 1+1 allelic composition and sequenced at a depth of at least 100x, resulting in 32 variants (**Supplementary Table 1**). Rescaling the VAFs of two samples (P2, R) against that with the highest purity (P1), had a major impact on the subclonality prediction of the relapsed sample R (**Figure 2B, C I-II**). If unscaled, almost all variants shared among the three samples were predicted to be subclonal in sample R, contradicting their status as clonal (present in all tumor cells) in the other two samples (**Supplementary Table 1**). For example, the unscaled data predicted that a *SIAH1* mutation, was clonal in the primary but subclonal in the relapse which was rectified post scaling resulting in a prediction of clonal mutation across all samples. Only one mutation, in *ST8SIA2,* exhibited changed clonality status between two samples, i.e. the two regions of the primary tumor (**Figure 2C III-V**). This was indicative of a regional clonal sweep at geographic transition between these regions, an event corroborated by copy number profiling, which showed a subclonal copy-number neutral imbalance of chromosome 4 in P1, which transited to clonality in P2 (**Supplementary Figure 4**).

As an example of how CRUST improves deconvolution by accounting for copy number variations, we then analyzed four different patient samples (NB22; **Figure 2A**)[4]. The CRUST-based copy number estimation resulted in a small number of discrepancies (2.7%) in the estimated allelic compositions compared to the available array-based estimates.[4] These were removed prior to analysis (**Supplementary Table 2**). VAFs were scaled against a diploid background and tumor cell fractions were calculated with allelic copy numbers considered. This revealed a varied tumor architecture across samples with evidence of polyclonal seeding of the metastatic sites, well in accordance with previous analysis of this case based on copy number alone (**Figure 2D**)[4]. Disregarding the copy number information, we reanalyzed the data assuming a balanced copy number state (1+1) for all chromosomes. The resulting deconvolution failed to pick up between-sample variations in clonality with considerable loss of resolution at backtracking of clones into geographic domains (**Figure 2E**). CRUST thus enabled consideration of corresponding copy number data revealing the true evolutionary progression **(Figure 2E-F).**

To observe the effectiveness of normalization and inclusion of allelic composition in a large scale mutational landscape interrogated by multiple platforms, we turned to a publicly available case of acute myeloid leukemia, with samples available from presentation and relapse[13]. Post scaling, CRUST was able to identify a single clonal population existing in both the primary and the relapse samples while analyzing whole genome, whole exome, and a custom-made mutation panel (**Supplementary Figure 5A-C**). The predictions concurred with that obtained from sciClone. However, a custom ion torrent assay resulted in a clonal/subclonal separation in disagreement with others (**Supplementary Figure 5D, E**). While investigating the respective total coverage provided by all four technologies, we noted

for the whole '*platinum list*' of SNVs declared by the original authors, that the ion torrent assay had a median coverage of < 50x for both samples (**Supplementary Figure 5F**). To increase robustness, we therefore extracted only SNVs called at a minimum depth of 15x in ion torrent for both samples resulting in 33 SNVs (**Supplementary Table 5**), with increased median coverage of 79x and 88x respectively for the primary and the relapsed sample. After scaling, these SNVs were predicted to belong to a single clonal population, consistent with the other methods (**Supplementary Figure 5G**).

In comparison to most available tools for clonal deconvolution, CRUST thus has several major features including a robust normalization for purity, an inbuilt assessment and integration of copy number alterations, and a possibility for user supervision to take à priori biological knowledge into account. While it determines clonality with stochastic algorithms, depending on sequence quality variation between samples or technical artifacts, sometimes no mathematical model can adequately harmonize spurious signals. As the variance of each clonal subcluster inflates with compromised quality of sampling/sequencing, CRUST expands on the prediction with a non-parametric test indicating the probability of a variant belonging to a certain cluster that compensates for hard thresholding. Because copy number profiles are not always available by a dedicated method such as SNP array for sequenced tumors, CRUST can estimate copy numbers from sequencing datasets. However, there remain risks of detecting spurious signals if copy numbers are solely estimated by this approach. Hence, a dedicated estimation should always take priority and we would recommend strict monitoring of sample quality, purity, sequencing technology variation, variable coverage across chromosomes, unstable GC content scaling and other factors. Another quality issue arises from low sequencing depth, leading to allele frequencies unsuitable for scaling resulting in false positive signals. Nevertheless, even at 50x coverage

with at least one sample with 70% purity, the clonality determination was accurate in simulations free of artifacts. CRUST thus not only delivers an accurate spatio-temporal clonal deconvolution of multi-sampled tumors, but also provides users a much-needed means for manual curation.

Software availability

CRUST depends on R (>3.5.0) and is available for download from GitHub repository https://github.com/Subhayan18/CRUST

Conflict of interest

None.

Authors' contribution

Conception and design: S. Chattopadhyay, D. Gisselsson

Development of methodology: S. Chattopadhyay, D. Gisselsson

Data acquisition: S. Chattopadhyay, A. Valind, J. Karlsson, D. Gisselsson

Analysis and interpretation of data: S. Chattopadhyay**,** J. Karlsson, A. Valind, D. Gisselsson

Technical support: J. Karlsson, A. Valind

Contribution towards manuscript: S. Chattopadhyay**,** J. Karlsson, A. Valind, N. Andersson, D. Gisselsson

Acknowledgments

The authors acknowledge technical support from the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure founded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced

Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. We would also like to thank the Swegene Centre for Integrative Biology at Lund University (SCIBLU) for assistance. This study was supported by grants to D. Gisselsson from the Swedish Research Foundation (2016-01022), the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the the Royal Physiographic Society, and the LMK Foundation. A special thanks to Chris Miller, Malachi Griffith for their assistance in procuring data from community resources (dbGaP: phs000159).

References

1. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366**, 883-892 (2012).
2. Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48**, 238-244 (2016).
3. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nature Genetics* **47**, 209-216 (2015).
4. Karlsson, J. *et al.* Four evolutionary trajectories underlie genetic intratumoral variation in childhood cancer. *Nature Genetics* **50**, 944-950 (2018).
5. Shibata, D. Heterogeneity and Tumor History. *Science* **336**, 304 (2012).
6. Longo, D.L. Tumor Heterogeneity and Personalized Medicine. *New England Journal of Medicine* **366**, 956-957 (2012).
7. Schwartz, R. & Schäffer, A.A. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics* **18**, 213-229 (2017).
8. Dentro, S.C., Wedge, D.C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor perspectives in medicine* **7**, a026625 (2017).
9. Andersson, N. *et al.* Extensive clonal branching shapes the evolutionary history of high-risk pediatric cancers. *Cancer Research*, canres.3468.2019 (2020).
10. Caravagna, G. *et al.* Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics* **52**, 898-907 (2020).
11. Miller, C.A. *et al.* SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLOS Computational Biology* **10**, e1003665 (2014).
12. Ahmadinejad, N., Troftgruben, S., Maley, C., Wang, J. & Liu, L. MAGOS: Discovering Subclones in Tumors Sequenced at Standard Depths. *bioRxiv*, 790386 (2019).
13. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems* **1**, 210-223 (2015).