OXFORD

Briefings in Bioinformatics, 22(6), 2021, 1–16

https://doi.org/10.1093/bib/bbab292
Case Study

Tracing the evolution of aneuploid cancers by multiregional sequencing with CRUST

Subhayan Chattopadhyay, Jenny Karlsson, Anders Valind, Natalie Andersson and David Gisselsson

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

Abstract

Clonal deconvolution of mutational landscapes is crucial to understand the evolutionary dynamics of cancer. Two limiting factors for clonal deconvolution that have remained unresolved are variation in purity and chromosomal copy number across different samples of the same tumor. We developed a semi-supervised algorithm that tracks variant calls through multi-sample spatiotemporal tumor data. While normalizing allele frequencies based on purity, it also adjusts for copy number changes at clonal deconvolution. Absent à priori copy number data, it renders *in silico* copy number estimations from bulk sequences. Using published and simulated tumor sequences, we reliably segregated clonal/subclonal variants even at a low sequencing depth (∼50×). Given at least one pure tumor sample (>70% purity), we could normalize and deconvolve paired samples down to a purity of 40%. This renders a reliable clonal reconstruction well adapted to multi-regionally sampled solid tumors, which are often aneuploid and contaminated by non-cancer cells.

Key words: clonal deconvolution; subclonal reconstruction; multiregional sequencing; phylogeny

Introduction

Genetic diversification during tumorigenesis and disease progression is governed by Darwinian principles acting on the level of single cells. A concerted effort has been dispensed in recent years to unravel the mechanisms of evolutionary dynamics in cancer. 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 genetic heterogeneity can result as a product of different evolutionary trajectories specific to the spatiotemporal localization of cells residing in a tumor [4]. Although all such cells are popularly believed to be neutrally evolved progenies of a common ancestor, depending on oncogenicity of the mutations acquired, some daughter cells observe a greater fitness advantage than those with neutral mutations. This pattern of divergent evolution can be observed by interrogating bulk sequencing data from tumors. As the genetic landscape in solid tumors often varies geographically within the same cancer, comprehensive reconstruction of tumor phylogeny requires multi-regional analysis and subclonal deconvolution [5, 6]. Such

S. Chattopadhyay, PhD is a postdoctoral researcher at Lund University. His research focuses on mathematical modeling of evolutionary dynamics in cancer.
J. Karlsson, PhD is a research scientist at Lund University. She is researching genetic signatures and evolutionary trajectories in pediatric tumors
A. Valind, MD, PhD is a postdoctoral researcher at Lund University focusing in tumor cell heterogeneity and evolution within individual tumors, both with a focus on biological issues but also on the clinical implications of these processes
N. Andersson, MS is a doctoral student at Lund University. Her research interest lies in modeling *in silico* processes to analyze tumor heterogeneity and reconstruction of phylogenies.
D. Gisselsson, MD, PhD is a professor and Vice Dean at Lund University heading the research group pathways in cancer cell evolution. His team combines high-resolution genomics on clinical samples with mathematical modeling and methods from species evolution to answer questions in cancer evolution.

Submitted: 15 March 2021; **Received (in revised form):** 16 June 2021

a deconvolution process leverages the relative abundance of a mutation across samples represented by variant allele frequencies (VAFs) [7]. Besides being very loosely defined, designation of clonality of a mutant variant depends on when in tumor development it emerges, its effect on cellular fitness, and on the spatial architecture of the tumor. A subclonal mutation is suggestively defined as any lesion that emerges out of a clone and does not observe extensive positive selection [8]. However, subclonal populations do accumulate mutations in known 'driver' genes, sometimes even emerging as events of convergent evolution [9, 10]. A subclonal lesion thus may represent itself in a fraction of the tumor cells, and its relative abundance often vary among samples even if these are acquired from a tumor at the same time. Hence subclonal mutations may become regionally fixed (present in all tumor cells) and thus appear clonal. Systematic bioinformatic approaches are critical to resolve such complex scenarios.

Clonal deconvolution is generally attempted using unsupervised clustering. It determines subclonal populations with distributional assumption on the variant read count [7]. Most methods assume that in a repertoire of clones and subclones, the relative abundances of variants resemble a binomial or beta-binomial admixture [11–13]. Thus, a Dirichlet finite mixture can segregate several clonal populations with distinct shapes and scales. Following this dogma, one can expect the clonal mutations to be distributed with markedly higher mean relative frequency than subclonal progenies. However, even with a modest rate of silent substitutions, many passenger mutations accumulate in subclonal populations within a few generations presenting a gradually regressing heavy tail of private mutations [13]. Furthermore, samples collected from a tumor at different locations or stages of progression can contain a remarkably varied proportion of normal cells from adjacent non-cancerous tissue. As a result, the VAF distribution of clonal mutations of one sample can mimic that of subclonal mutations of another, purer sample. Variants can also end up with higher (or lower) than expected relative abundance if residing in chromosomal regions affected by copy number changes. As copy number alterations can appear as both clonal and subclonal lesions, they can significantly complicate clonal deconvolution based on VAF distributions [14].

Here, we intended to solve the problem of how VAF values are influenced by purity and copy number with an analysis suite for clonal deconvolution named CRUST (Clonal Reconstruction of tUmors with Spatio-Temporal sampling). It can classify clonal and subclonal mutations from bulk sequencing data of multi-sampled tumor tissue. CRUST will probabilistically rescale VAFs from samples with purity contrasting against the sample with the highest purity. The program is additionally built to integrate data from precise copy number estimations such as those from single nucleotide polymorphism (SNP array) data and produce allelic composition specific clonality predictions. In absence of SNP array data, CRUST can estimate copy number profiles given sequencing summaries (allelic frequency data corresponding to alternate and reference allele) from tumor and constitutional genome. While these processes need to have mathematical rigor by parameterizing stochastic assumption on the distribution of clonal variants, we also recognized the need of a semi-supervised algorithm to reconcile a purely traditional data driven approach and a user driven heuristic pattern recognition for clonal deconvolution. The user thus will be able to actively curate the semi-supervised clustering process prior to the deconvolution based on visual input. CRUST also allows sample specific user driven readjustments in deconvolution post analysis. We were able to demonstrate in clinical samples and in simulated tumor biopsies that in presence of at least one relatively pure sample, without user intervention, CRUST can rectify clonality estimates for samples with compromising purity that would otherwise be heavily biased towards a prediction of subclonality. Furthermore, we demonstrated that CRUST increases the resolution of clonal deconvolution for aneuploid tumors by taking the influence of chromosomal copy numbers into account. With curation provided by the user and proper consideration to the temporal fluctuations in appearance of each genetic lesion, CRUST can thus help reconstruct the most likely phylogenetic history of a tumor.

Results

A semi-supervised approach to clonal deconvolution

The primary functionality of CRUST is in clonal deconvolution from a substrate of sequencing summaries of single nucleotide variants. As a first example, we demonstrate this on a hypothetical tetraploid tumor where mutations are present in either one or three out of the four available homologous chromosomes (Figure 1A). The simulated tumor is represented by eight biopsies (inbuilt data *test.dat*). With a set of samples from the same tumor obtained at different locations and/or time points, CRUST deconvolves 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 (Figure 1B and C**; see section 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). With the copy number analysis, sequence variants from a single tumor are 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. The subclonal estimation process is based on 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 1D **I-II)** 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). In this hypothetical case, CRUST thus correctly assigns mutations present in 1/4 and 3/4 alleles to both clonal and subclonal states, while a cluster-based deconvolution without accounting for copy number may assign all mutations present in 1/4 alleles to the subclonal stratum.

Performance testing of scaling with simulation

To assess the accuracy of CRUST-based deconvolution across varied purity and sequencing coverage, we simulated tumor samples under three distinct assumptions (Figure 2A; Supplementary Figure 2). 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 [13], we modeled the reduction in variability biased towards the left tail with a log-exponent function and simulated

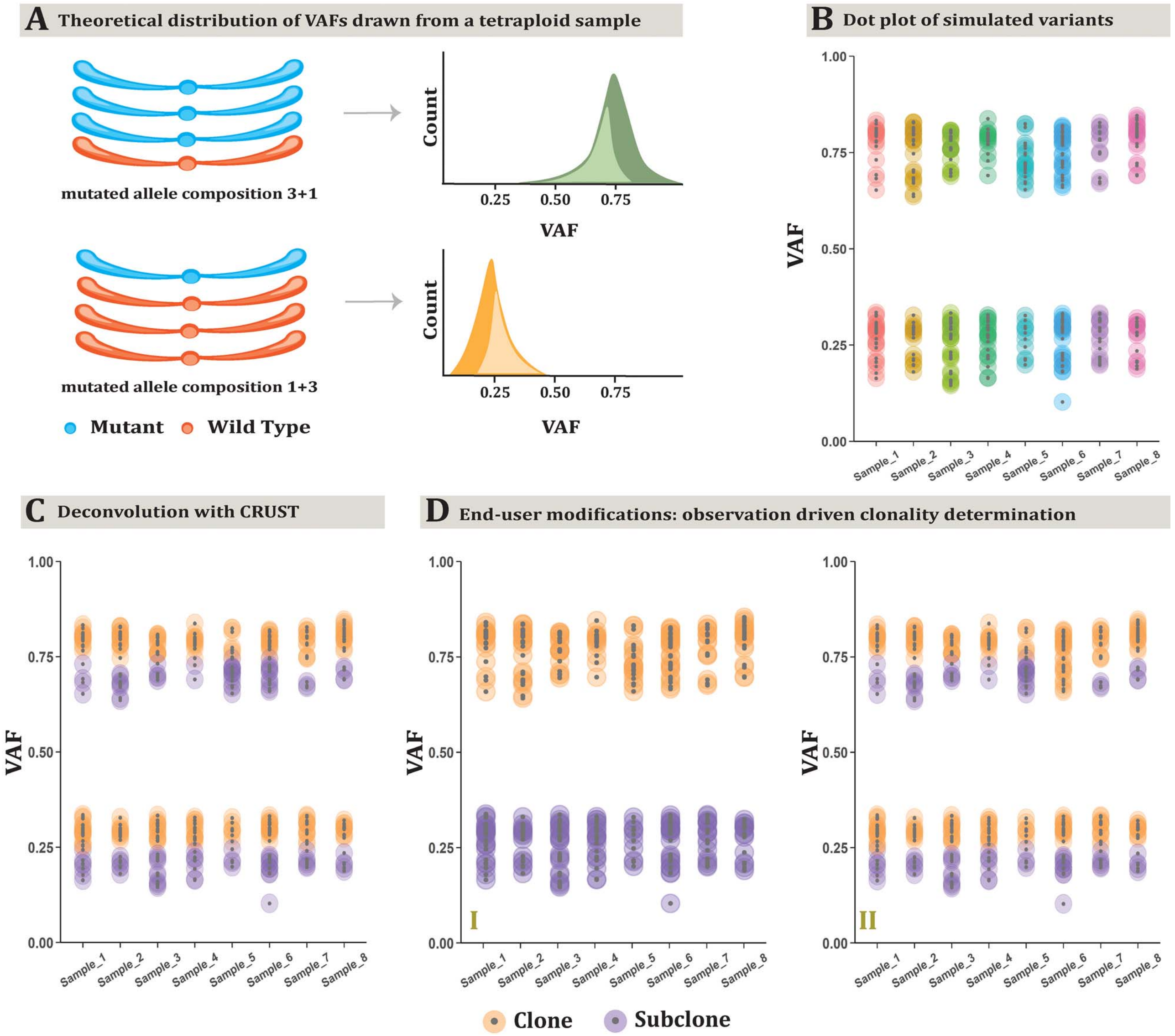

Figure 1. Clonal deconvolution of a simulated tumor genome. A tetraploid tumor is simulated where all samples adhere to an allelic composition of 3 + 1 (or 1 + 3) (A). This makes the expected VAF distribution bimodal with corresponding peaks at frequencies 0.25 (i.e. 1/4) and 0.75 (i.e. 3/4). There are eight samples representing different biopsies. CRUST first displays a dot plot of the VAFs pertaining to all samples (B). 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 (C). 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 four clusters as seen above (two clonal and two subclonal). In (D 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 (D II). Over the default optimum prediction, for *Sample_6* a user has here chosen to fit a 3-cluster deconvolution that picks up two clusters attributed to the clonal population (at allelic compositions 1 + 3 and 3 + 1) and one subclonal. It is to highlight that first; the initial estimate of clonal clustering is important and second; given enough noise in a real sample one can envisage that the unsupervised prediction may result in over clustering. In such cases one still can change clonality assignment post-hoc and assess the feasibility of the deconvolution.

an admixture of clonal and subclonal variants assuming a balanced background copy number. The model assumptions indicated what the varying purities could affect; first, the mean i.e. the cluster centroid of the VAF distribution; second, variance i.e. the scale of the distribution and lastly, both mean and variance. Accuracy of prediction was measured by comparing expected clonality based on the simulation assumption and the predicted clonality inferred by CRUST with the Jaccard index. For all three-model assumptions, 1500 simulations were performed, respectively, to generate two samples in each iteration. Purities were sequentially modified for each iteration to have produced 100 variants for each sample resembling an impure pre-treatment biopsy and a much purer metastatic sample. When only the variance of the distributions were varied, the deconvolution and prediction accuracy broke down fastest and the departure was statistically significant (Figure 2B) suggesting subclonal clusters can retain the same centroids while increasing in variance but the VAFs overlap between clusters making it virtually impossible to segregate them. CRUST scaling maintained a concordance of $\rho > 0.7$ in presence of at least one representative sample (purity >0.7) given that the difference in purity between samples was less than 30% of the maximum. Hence, theoretically a sample with 40% purity can be confidently rescaled given another available sample that is 70% pure. This simulation draws sequencing coverage randomly for any given variant between 50× and 300×. Hence if we would like to analyze a sample with low purity (<40%) but sequenced with consistently high coverage (>100×), CRUST is still able to rescale the data given a corresponding purer sample (tested with AML samples later). In simulations with a larger than 30% departure in purity between samples, the concordance decreased drastically ($\rho < 0.4$).

Next, we compared CRUST against some of the contemporary and frequently used clustering algorithms (PyClone, sciClone, MAGOS) [11, 12, 15]. First, we compare against MAGOS and sciClone with varying sequencing coverage as the authors of the former paper claim that both outperform PyClone in moderate to low covered sample [11]. All tumor biopsies were simulated as containing two samples with varied PURITIEs drawn from a lognormal-binomial mixture. At sequencing coverage of 300× all three algorithms were able to maintain a $\boldsymbol{\tau}_{\mathbf{median}}$ (median Jaccard index that varies between 0 and 1) of 0.95 with CRUST leading ($\boldsymbol{\tau}_{\mathbf{median}} = 0.99$, Figure 2C). This pattern remained consistent throughout 100× and 50× simulations and all comparisons with CRUST were statistically significant at the 0.001 level. At 30×, CRUST had a lowered $\boldsymbol{\tau}_{\mathbf{median}}$ of 0.83. The interquartile distances between the $\boldsymbol{\tau}_{\mathbf{median}}$ estimates also markedly increased at 30× (range: 0.037–0.11). At lower coverages (≤50×), sciClone failed to predict the correct number of clusters often predicting as many as six clusters instead of four although CRUST and MAGOS predicted the simulated clonality status more accurately (statistically significant with two-tailed *P*-value <0.05, Mann–Whitney test) with at least 50× coverage (Figure 2C) [11, 12]. Additionally, CRUST was compared to PyClone on the inbuilt simulated tetraploid tumor described in Figure 1. As PyClone requires all variants to be present across all samples to not get filtered out, we declared negligible alternate allele count for those absent. PyClone determined 21 different clusters over all 116 distinct variants Supplementary Figure 3A. This apparent over-clustering could have been remedied by looking into the similarity indices of the clusters, represented in Supplementary Figure 3B. Four main subclusters are apparent in the heatmap, which showed a near perfect overlap ($\rho = 0.92$) with the two clonal and two subclonal clusters determined previously (Figure 1).

Illustration of the impact of scaling for correct deconvolution

As an example of the importance of scaling, we extracted from a published dataset on childhood cancer [4], three neuroblastoma tumor tissue samples with varied purity (55–90% tumor cells) from a patient, two from the primary tumor (NB12, P1 and P2) and one from metastatic relapse (R) (Figure 3A). 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 100×, resulting in 32 variants. 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 3B and 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; ***Scaling***). For example, the unscaled data predicted that an *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 3C **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).

Illustration of the importance of accounting for allelic copy numbers

As an example of how CRUST improves deconvolution by accounting for copy number variations, we then analyzed four different patient samples (NB22; Figure 3A) comprising of two primary (P1, P2) and two metastatic tumor samples (M1, M2), [4]. CRUST is inherently dependent upon variant specific allelic composition information. Best practice is to assimilate sequencing summaries with separately obtained SNP array for a precise allelic copy number estimate. CRUST also contains a function to approximate copy numbers from tumor sequencing compared to germline variants (polymorphisms) called from the constitutional genome. To identify distinct aneuploidies across samples it graphically presents segmental allelic imbalance and average log-relative coverage as done elsewhere [16]. We compared the segmental plots generated from SNP array data and that estimated by CRUST (Figure 4). The allelic imbalances estimated from exome sequences closely resembled those obtained from the SNP array, but with slightly less fidelity (Supplementary Table 1; ***phs000159_seq***). The 1q gain, 6p gain, whole chromosome 7 gain and, 17q11 loss and distal 17q gain were clearly identified with the estimates (Supplementary Figure 5, Supplementary Table 1; ***phs000159_iontorrent***). Overall, the CRUST-based copy number estimation resulted in only a small number of discrepancies (2.7%) in the estimated allelic compositions compared to the available array-based estimates (Supplementary Table 1). [4] These were removed prior to further deconvolution (Supplementary Table 1; ***NB22_copynumber***).

To elucidate the geographical makeup and temporal evolution across NB22 biopsies, VAFs were then scaled against a diploid background and purities 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 3D-G) [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 3F). By integrating copy number and sequencing data, CRUST thus revealed details in the evolution of this tumor that would have had passed unnoticed if copy number data were not considered (Figure 3F-G).

CRUST resolves clone topographies in published datasets

We extracted publicly available whole exome sequencing (WES) data on 20 multi-regionally sampled local primary non-small cell lung cancer (NSCLC, adenocarcinoma) from the initial release of the TRACERx project [17]. Deconvolution with CRUST, including copy numbers of mutated alleles in clone size estimates, made it possible to infer clonal topographies for all included samples at

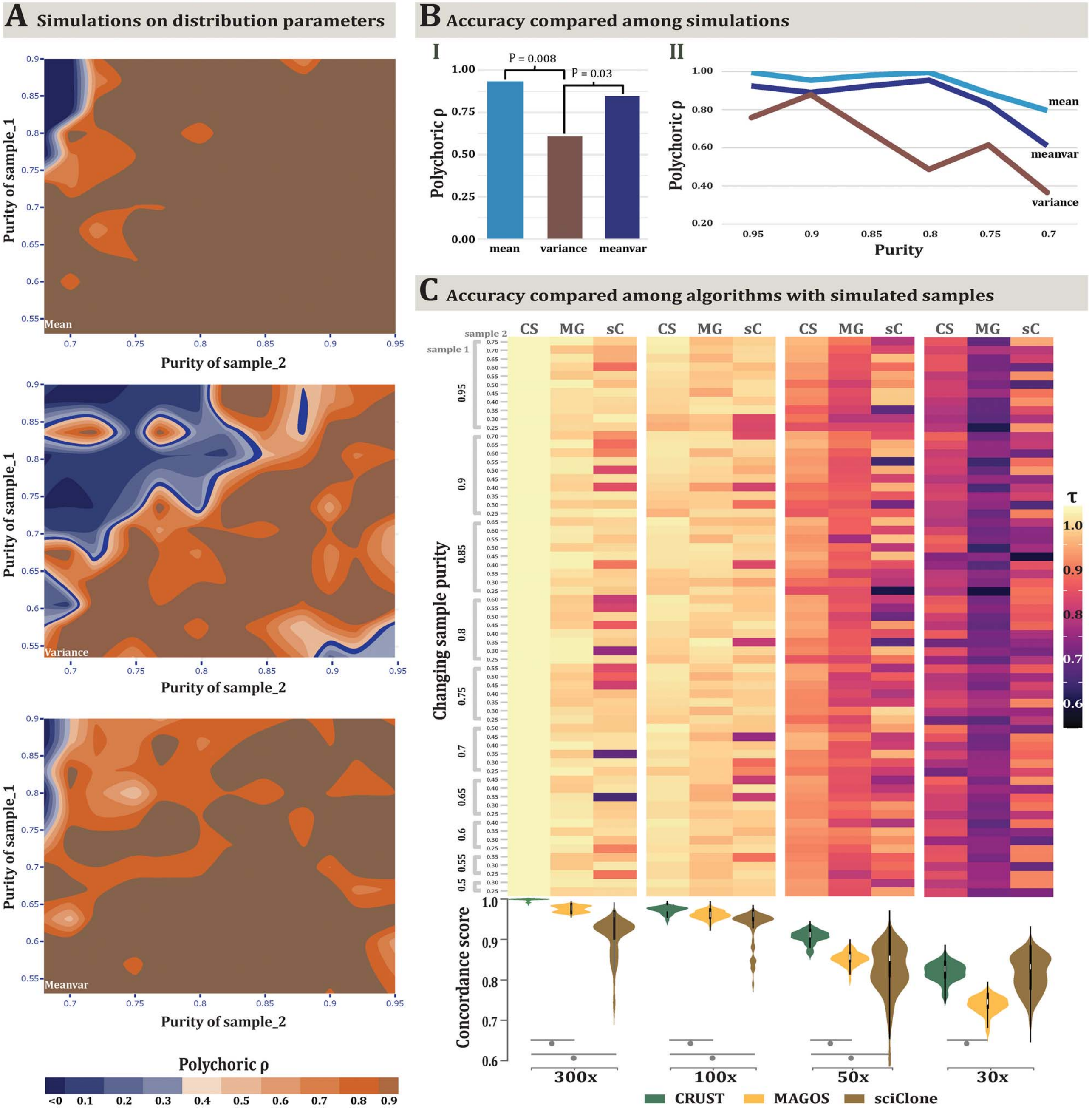

Figure 2. Evaluation of efficacy of CRUST with simulation. In (A) simulations of scaling with varying sample composition are shown. Each iteration generates two samples, say X and Y with purity Tx and Ty, respectively. Assuming Tx > Ty, CRUST rescales the variants in Y based on those in X. Simulations 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 purity that only affects the mean of the VAF distribution. The middle shows changes in purity 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 (B-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 (B-II) the trend of change in average concordance with varying levels of purity between the three algorithms is depicted. A comparison across deconvolution methods was done with simulation of varying sequencing coverage (C). Samples are drawn with varying purity for four sets of coverage at 300×, 100×, 50× and 30×. Ordinal cluster similarities were assessed between 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 purities. Each cell in the heatmap reflects that obtained from a paired simulated sample denoted in the joint y axis purity. The leftmost y axis annotation denotes purity 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.

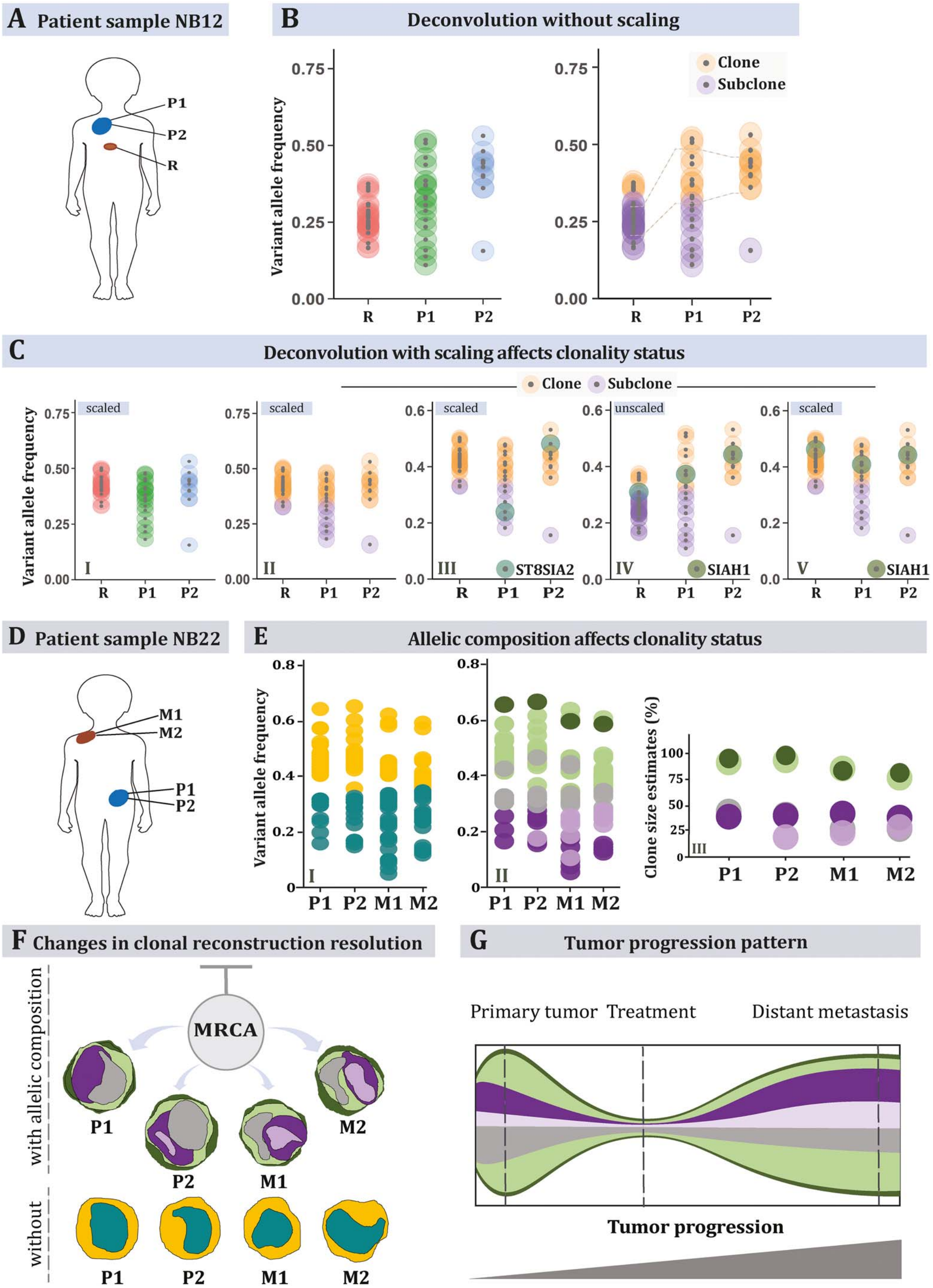

Figure 3. Tracing clonality trajectories across samples. Multiple samples with varied purity from each of two neuroblastomas (NB) were used as examples. NB12 is represented with three samples (A), two primary tumor biopsies (P1, P2) and one relapse (R). The primary sample was ~90% pure whereas the relapse 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 (original sample specific VAFs are on the left, clonality predictions are on the right, B). 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 (E I), whereas considering allelic composition results in five clone/subclone clusters across all four samples (E II). This deconvolution would not have been possible without copy number data taken into account. Estimated clone sizes are depicted below with purities of each cluster (E III). Inferring tumor evolution from deconvolution (F-G) shows how starting from an unknown most recent common ancestor (MRCA) one of the primary clones (in grey) shrinks whilst another subclone (in light purple) expanded at metastatic sites. Clone sizes estimated from the 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 an MRCA. The bottom panel in (F) devoid of copy number data lacks resolution to detect any such change.

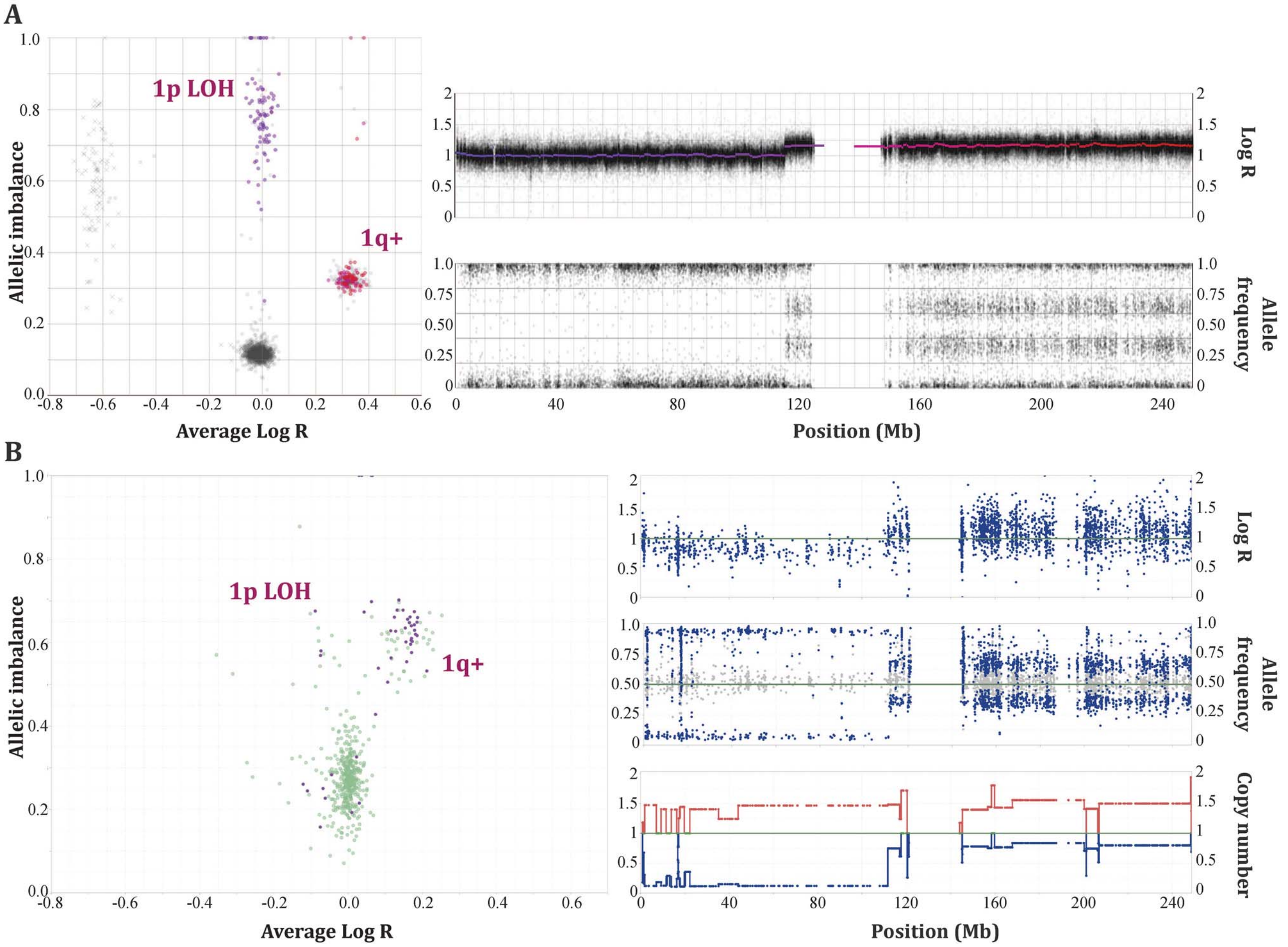

Figure 4. Estimation of copy number with CRUST. The segmental copy number estimates generated with CRUST are compared with SNP array profiles from the same tumor (NB22). We generated chromosome-wise plots of estimated segmental allelic imbalances against corresponding average log transformed relative coverages. In the top panels (A) the plots are generated with well-established analysis tool TAPS from SNP array data [16]. In the lower panel (B) the same plot is generated with data estimated with CRUST. The down-right panel includes three figures, from top to bottom representing the estimated relative coverage, allelic frequencies and of segmental copy number. As demonstrated by TAPS, clones with different allelic compositions would show up at unit-separated distinct clusters along a fixed tangent in the allelic imbalance plots and the corresponding subclones would appear with a slight departure in the y- and x-axes. The allelic imbalance plots generated with CRUST retained the copy number specific cluster structures. We recommend the users to consult these plots to verify the CRUST estimated allelic compositions.

a level of detail not provided in the original publication (Supplementary Figure 6). Subclones could be distinguished from clones in all tumors and about 15% (median) of all variants in each tumor were predicted to be subclonal. All 20 tumors presented evidence of topological genetic diversity; i.e. variants were predicted to have changed clonality status (clonal crossover, clonal to subclonal and vice versa) between different samples of a tumor given that the allelic composition remained unchanged, indicative of clonal sweeps across the primary tumor space. Some tumors were found to have an exceptionally high number of subclonal and crossover mutations, in particular CRUK003 (71%), CRUK004 (73%) and CRUK0018 (55%). This finding was well in accordance with the high proportion of branch mutations previously reported in these tumors (71, 76 and 45%, respectively) [17].

The relative proportion of genes predicted to be globally clonal or subclonal were sometimes different between CRUST and the original analysis (Figure 5A) [17]. For each patient, approximately 3% (median) of the quality-controlled variants (i.e. 1% of all variants) were responsible for clonal crossover. These belonged to about 5.6% of all annotated genes retained post quality control. Most of these crossover genes detected by CRUST (91%) were found to be absent from the previously published phylogenetic analyses [17]. A majority (18 of 31) of those retained in the original analysis were in fact predicted to concur between CRUST and the original analysis as subclonal driver mutations (e.g. *TP53*, *KRAS*, *PIK3CA*, *CDKN2A*, *ATM*, etc.), and the proportion of variants discarded from the original study resembled the proportion of crossover variants detected by CRUST (Supplementary Table 2). This indicates that CRUST adds value by resolving shifts in clonality status between samples.

We then compared the proportion of driver mutations predicted by CRUST to be truncal with that suggested previously [17]. The original analysis included 100 patients whereupon inferences were drawn with the constructed phylogenies. The CRUST analysis with 20 individuals retained approximately a third of the original set of annotated genes. When only this subset of 20 patients were considered, 63 genes in the published study were annotated as driver of which 31 (49%) were annotated as 'subclonal driver' in at least one patient. Post quality control, CRUST retained 56 of the said 63 driver genes among which 21 (38%) were subclonal in all samples along with 18 (32%) having undergone clonal crossover bringing the totality of subclonal drivers to 70%. Of these 18 crossover genes,

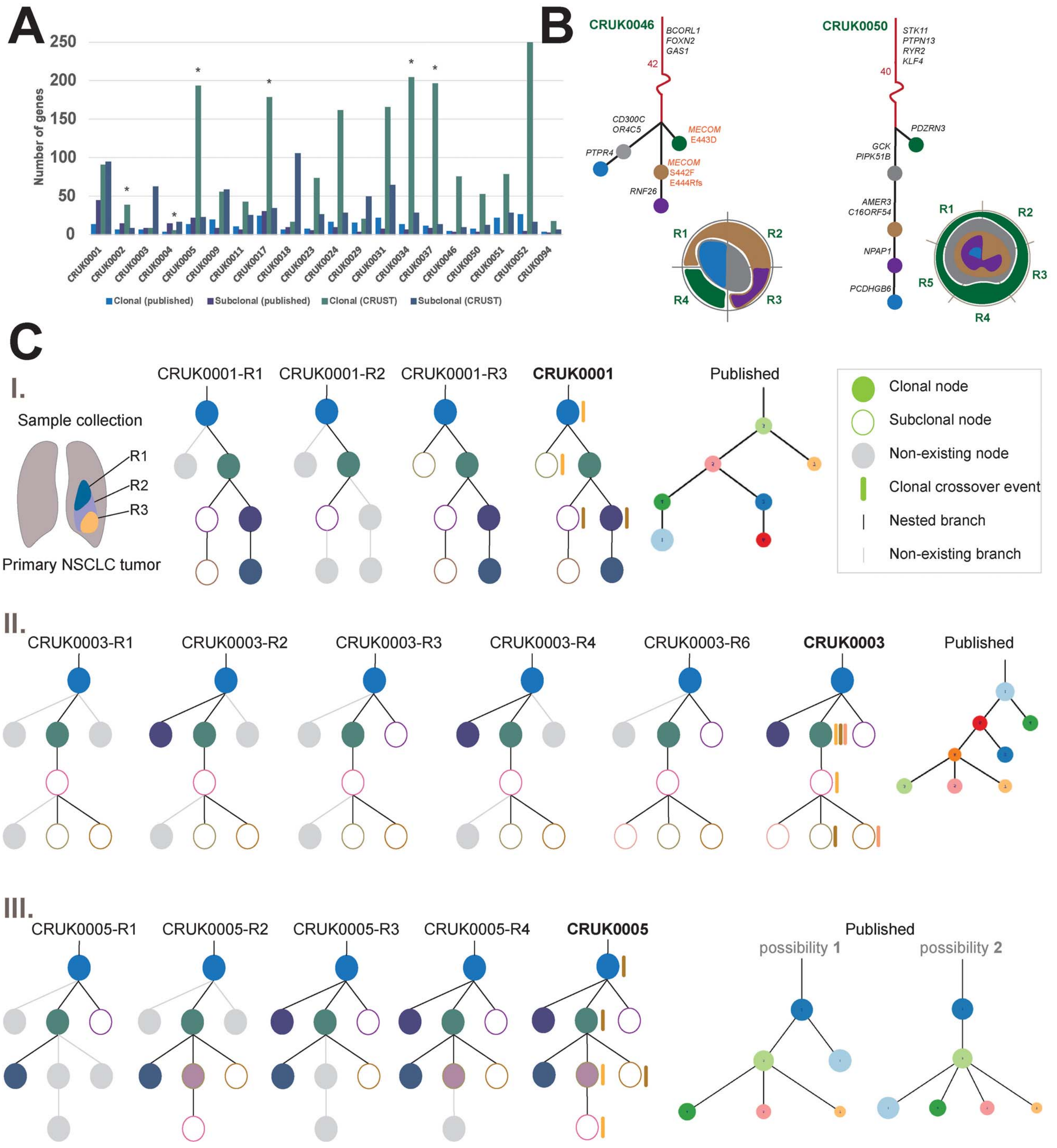

Figure 5. TRACERx Clonal geography. A comparison is made between clonality prediction from CRUST and that previously published (A). The bar chart shows the number of globally clonal and subclonal genes identified across all 20 samples in the two respective analyses. A non-parametric two tailed z-test is performed to denote significance difference ($P < 0.0025$) between them, highlighted by star. In B phylogenetic trees of the two tumors, CRUK0046 and CRUK0050 are depicted along with the tentative sample specific clonal geography in the corresponding radial diagrams. The stems are in red with some of the noteworthy genes accompanying in black. Each subsequent branch is in black and the corresponding branch length is in proportion to the number of genes in each branch. The three variants corresponding the gene *MECOM* are highlighted for CRUK0046 as it depicts convergent evolution. The clonal / subclonal population nodes in the tree carry forward their colors in the respective clonal geography diagram. C depicts comparison of nested phylogenies between CRUST and the published analysis for three tumors, CRUK0001 (C-I), CRUK0003 (C-II), and CRUK0005 (C-III). Each sample has its own phylogenetic tree, and all samples from a tumor are combined to build a consensus tree. If nodes and branches that are present in the consensus tree are absent from a sample-specific tree, they are greyed-out in the sample tree. Accompanied by the CRUST-predicted consensus trees, the corresponding published tree (s) is provided alongside. The clonal and subclonal populations are represented by filled and hollow circles, respectively; clonal crossover events are denoted only in the consensus trees with colored bars alongside the clone (s) and the subclone (s) where the variants appeared. Each color of the crossover bars represents a unique cross over pair between the corresponding clone and subclone i.e. a variant that belongs to different clonal/subclonal clusters in different samples. The largest clonal population in each sample is assumed to represent the stem of the phylogeny. Please refer to the original publication for color codes of published trees.

12 were present subclonally at least in one of the other 80 patients analyzed in the previous study. This was in line with the original finding, as 75% of the tumors were observed to have acquired driver mutations late indicative of temporal fluctuations in evolution. The lacking spatial resolution of the previous study meant only nine driver genes were inferred to have both clonal and subclonal status in different patients. One such gene was *PIK3CA*, for which crossover was confirmed by CRUST. This led us to speculate that crossover variants if seen in genes of lesser biological impetus had been misclassified as ambiguous observation previously. Omitting these might deprive the analyses of spatial heterogeneity as shown in several of the tumors (CRUK0001, CRUK0005, CRUK0051), and miss the presence of regional clonal sweeps (CRUK0003, CRUK0029, CRUK0050, CRUK0094).

To understand the impact of copy number consideration in clonal deconvolution, tentative clonal geographies were constructed for CRUK0046 and CRUK0050 (Figure 5**, B**). These tumors were selected as they had adequate number of samples (4 and 5) coupled with relatively simple ploidy profiles (up to tetraploid). In CRUK0046, we found several genes (*FOXN2*, *GAS1*, etc.) in the stem of the phylogeny that are of suggested importance for lung cancer initiation, resistance, and metastasis [18, 19]. Even more interestingly, we noticed convergent evolution in different sites on the gene *MECOM*, commonly shown to be harboring structural aberrations in NSCLC [20, 21]. CRUK0050 on the other hand had mostly undergone linear evolution except for one monogenic branching. Here, the tree stem encompassed several well-known mutations in lung cancer such as *STK11* (prognostic marker, aids drug resistance with fusion partner *LKB1*, [22, 23], *PTPN13* (tumor suppressor in lung adenocarcinoma) [24], *NOVA1* (promotes telomerase activity in NSCLC) [25], *ANXA2* (influential on lung cancer cell survival, apoptosis, as well as a construct of EGFR- fusion gene) [26–28], *RYR2* (associated with high NSCLC mutation burden in conjunction with exposure to high air pollution) and *KLF4* (regulates lung cancer initiation) [29]. In both tumors, the stem harboring most influential aberrations allude to these being early events in initiation.

To further understand the prevalence of crossover variants and their subsequent effect on the tumor clonal landscape, we finally constructed cross-sample phylogenies of the tumors and compared them with the previously published phylogenies. Three interesting cases (CRUK0001, 0003 and 0005) are highlighted in Figure 5C. For CRUK0001, the published and CRUST-inferred phylogenies were essentially the same, despite crossover variants being present across two different clone–subclone pairs. In CRUK0003, the crossover events resulted in subclonal diversification. The published results identified the distinct clones and subclones correctly, but the nesting structure was quite possibly different lacking the reinforcement from the three pairs of crossover events. CRUK0005 demonstrated how crossover events may aid in untangling ambiguity in phylogeny reconstruction. The published results included two possible tree structures for this tumor as the nested clones could be placed in different hierarchies. There were three distinct sets of crossover events detected by CRUST, which again pointed towards a greater degree of subclonal diversification than previously detected. This was consistent with one of the two phylogenies originally suggested. We conclude that CRUST adds significant detail by adding a component of cross-sample spatial resolution to solid tumor evolution, but that sample specific purity estimates were unavailable in this (as in most) public datasets, possibly hampering the detection of false positive crossovers.

CRUST deconvolution across platforms is stable but coverage dependent

We finally turned to a case of acute myeloid leukemia (AML), with samples available from presentation and relapse [30]. We selected a case (AML31) that had two biopsies sampled (primary and relapse) that were sequenced on several different platforms. The VAFs needed to be normalized as the samples varied greatly in purity (90.7 and 36.2%, respectively). 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 7A-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). 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 <50× for both samples (Supplementary Figure 7E). To increase robustness, we therefore extracted only SNVs called at a minimum depth of 15× in ion torrent for both samples resulting in 33 SNVs (Supplementary Table 1**;** ***phs000159_iontorrent***), with increased median coverage of 79× and 88×, respectively, for the primary and the relapsed sample. Thereafter, the remaining variants in the primary sample attained a similar distribution as that observed by the other three technologies and scaling resulted in variants in the relapsed sample undergoing a similar magnitude of displacement, whereupon the centroid of the relapse sample VAFs realigns itself with that of the primary sample. This inferred a single clonal population, consistent with assessment by the other techniques (Supplementary Figure 7F). In all, this cross-platform analysis confirmed the result from simulations, indicating that sequencing depth is a limiting factor for deconvolution with CRUST.

Discussion

Parameters based on clonal deconvolution and tumor cell phylogeny have been shown to carry prognostically essential information in a range of cancer types [31–33]. Such phylogenetic reconstruction of tumors has mostly relied on bulk sequencing data, although this is about to change with the advent of single cell analysis. However, because single cell DNA sequencing data are still limited by a high cost and a relatively low resolution, we anticipate more studies will take place where bulk sequencing is used to investigate cancer cell evolution based on multiple samples from the same tumor. Distinguishing clonal from subclonal mutations is a critical step in all studies where tumor phylogenies are deduced from such data.

Here we presented a parametric semi-supervised method of clonal deconvolution developed to interrogate variant clonality in multi-regional/temporal samples of a tumor. In comparison with most available tools for clonal deconvolution, CRUST has several major features including a robust normalization for purity, an inbuilt assessment and integration of copy number alterations, and a possibility to take à priori biological knowledge into account through user supervision. 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. Nevertheless, even at 50× coverage with at least one sample with 70% purity, the clonality determination was accurate in simulations free of artifacts.

CRUST builds upon unsupervised hierarchical clustering with an admixture model or purely bootstrapped agglomerative clustering with optimal cluster determination. This is in significant contrast to its predecessors such as sciClone that uses Bayesian mixture models to determine prior probabilities of clusters assuming a copy neutral background. Unfortunately, this means sciClone is not optimal for multiregionally or temporally sequenced samples where the tumor genome evolves over time [12]. MAGOS has a similar limitation where it removes all variants that do not belong to the same copy number segments [11]. PyClone on the other hand clusters the probability distribution of estimated cancer cell fraction taking into account the copy number profiles but tend to over cluster in certain situations [15]. There are other tools that performs clonal deconvolution such as EXPANDS that does not deconvolve variants independently across samples or PhyloWGS that uses Monte Carlo Markov Chain to create decision tree [34, 35]. However, none of these algorithms allow adjustments based on user input based on available prior knowledge. Depending on departure of allelic composition on log transformed relative coverage one can determine the possible admixture of clonal and subclonal clusters. With this CRUST adds value to the deconvolution by independently analyzing variants across samples with respect to their corresponding allelic compositions. The result of this was delineated in the TRACERx analysis where CRUST was able to unravel previously unresolved ambiguity in the phylogenetic structures as well as underline new events that strongly corroborates the original conclusion. In addition, it was able to identify further evidence of varying evolutionary trajectories. In neuroblastomas, CRUST identified regional clonal sweeps that was previously demonstrated as typical of high-risk variants of the disease [4]. While with the AML samples sequenced more than half a decade apart with varied technologies with varied coverage, CRUST was able to build a consensus deconvolution from each data set individually.

Through CRUST, we have demonstrated how a semi-supervised model can yield biological insight with a purely mathematical framework to determine clonality of mutations. CRUST not only delivers a high-resolution spatio-temporal clonal deconvolution of multi-sampled tumors, but also provides users a much-needed means for manual curation. The sequence of somatic mutation can thus be traced more accurately across multiple samples from a tumor especially in cases with a large burden of chromosomal gains and losses.

Methods

Quick user guide

Clonal deconvolution can be performed with CRUST following a minimal number of steps. The argument data needs to contain at least two columns specifying sample ID and the VAFs. Annotation for the variants are optional but is required to be able to create variant specific plots of the deconvolution. If sample (s) need to be scaled, an additional column should be provided with the purities for each sample. The allelic composition of each variant corresponding to each sample should be available to the user. Here we will assume that SNP array data is not available to the user. Hence, it needs to be estimated and we will provide a protocol for a sample analysis:

- First make sure the.vcf file contains calls from the constitutional DNA that enables CRUST to estimate copy number profiles. Each.vcf file should only have summaries for one tumor sample and one corresponding normal tissue sample. For each tumor sample, this analysis needs to be performed separately.
  a) Sample specific allelic composition can be estimated with the function *AlleleComp*. This function estimates allele specific copy numbers using sequencing inputs from constitutional genome. It is usual practice for variant callers to remove all SNVs with the 'REJECT' flag (that includes all inherited variants) to produce the final .vcf file. CRUST requires these variants to estimate copy number. Input for this function requires the allelic depth (*AD*) identifier usually present in the *FORMAT* argument of the .vcf file. There are two compatible methods for this estimation that can be selected at user's discretion.
  b) A summary of the estimates can be visualized with allelic imbalance and copy number summary statistics if the estimation was performed with the 'naive' method. The *View_summary* function can generate this.
- The data thus obtained can be merged with the original data file containing sequencing summaries. As each distinct allelic composition have a distinct VAF distribution, the data needs to be analyzed separately for each allelic make up.
  a) Before deconvolving, the VAFs need to be scaled according to the corresponding purities with the *seqn.scale* function.
  b) The scaled VAFs now can be used for deconvolution with *cluster.doc*
     I. User is here asked to provide a postulated clonal/-subclonal composition of the variants from visual inspection of the dot plot
     II. A measure of clustering accuracy in terms of Bayesian information criteria or Smin statistics is provided to the user along with the best (statistically) suggested clonal deconvolution, which the user is free to retain or override (Supplementary Figure 1).
- Depending on biological agreement and objective plausibility, the user can choose to fit different deconvolution structures with *cluster.doubt*
- Variant specific plots of the deconvolution can be obtained with the function *variant.plot*. It is also possible to automate this process and obtain plots for all variants with *variant.auto.plot*
- The clonality estimates thus obtained can be used to estimate clone sizes and to formulate phylogenies. We estimated clone sizes for the example dataset named *Neuroblastoma* [36] provided along with CRUST.

Requirements: CRUST was built with R 4.0.0 and depends on several bioinformatic packages as well as mathematical and data processing packages (please refer to dependencies in package description). For successful installation, Rtools40 as well as devtools (≥2.4.0) is also required.

Phylogenetic analysis: We recommend performing phylogenetic analysis with the CRUST output using DEVOLUTION (https://github.com/NatalieKAndersson/DEVOLUTION), a phylogenetic reconstruction algorithm specifically built to handle multiple samples. The clone size estimates along with the progenial nesting inferred by DEVOLUTION can be used to create a tumor evolution map with tools like clonevol or Evofreq.

Study design and data preparation

Clinical samples included here were fresh frozen tumor biopsies analyzed as a part of a larger study [4]. The DNA was extracted with the AllPrep DNA/RNA/Protein Mini kit (Qiagen); segmental aberrations were analyzed using the Cytoscan HD platform (Thermo Fisher Scientific/Affymetrix). Whole exome sequencing of neuroblastoma samples were performed by SciLife lab (Stockholm, Sweden; Illumina) on NB12 (two primary and one metastatic relapse sample) and NB22 (two primary and two metastatic relapse samples). Subsequently a targeted resequencing on all NB22 samples and the two NB12 primary samples were performed based on the variants identified by exome sequencing at SciLife lab (Uppsala, Sweden) using the AmpliSeq technology for design and Ion Torrent for sequencing (both from Thermo Fisher Scientific). Further details on human sample collection, preparation, relevant technologies, quality control measures and basic bioinformatic analyses are described elsewhere [4].

Baseline normalization of VAFs

Human tumor samples are often contaminated with non-neoplastic cells, e.g. local epithelium, fibroblasts, endothelium, pericytes and immune cells. Even though these cells are expected not to carry the clonal somatic mutations present in the tumor cells, they can affect the VAFs of such variants by a dilution effect. A quantitative measure of sample purity is given by the tumor cell fraction. It can be estimated from either SNP arrays, sequencing or methylation data [37–39]. Given that purities may vary among samples of a tumor procured at different sites or time points, variants may be present with altered VAFs even though the relative abundance of them among neoplastic cells are identical in the respective samples. On the contrary, they can also appear to be identical despite being influenced by selection or genetic drift that significantly changed their relative frequency among tumor cells. A normalization strategy can realign the VAFs of a variant from several different samples, given a common baseline such as the purity.

It is customary to avail an integral normalization in presence of a k-dimensional reference space to transform the n-k dimensional scalable space, assuming the total integral uniformly influences dilution of all signals in the space. In its stead, we incorporated probabilistic quotient normalization of VAFs based on sample specific purities, as quotient normalization assumes that changes in unique elemental signal dilution affects only the signal of that element in the complete spectrum and dilution in global signal of a spectrum influences that of the overall spectrum [40]. This normalization algorithm is carried out between spectra where,

a. Integral normalization is performed on the scalable spectra.
b. Median spectra of the reference sample are calculated; lacking a reference spectrum, a control sample can be obtained from the scalable spectra but is advised against.
c. Quotients are calculated for the scalable spectra scaled against that of the reference.
d. Scalable spectra are normalized with the median of the quotients.

As the reference spectrum is same as the purity and VAF combination for the highest quality sample, it is assumed to be devoid of (detectable) signal dilution. Hence, we realign VAFs of all other samples against their corresponding purity scaled with the departure in purity between that sample and the purest one.

Cluster detection and assignment

We use VAFs as the metric for clustering as is popularly used in the clonal deconvolution literature. VAF is defined as the relative frequency of read count pertaining to a single variant and is calculated by

$$\upsilon = {}^{f_{alternate}}\!\Big/\!{}_{\left(f_{reference} + f_{alternate}\right)} \tag{1}$$

where f denotes read depth of an allele, alternate allele is the mutate allele of a presumed biallelic variant, and the reference allele is the wild-type allele.

We take a set of N variants that may be observed in S samples procured from a tumor. Given the set of S samples, VAF of the i^{th} variant can be represented by υ_i for a set of k VAF values where $i, k \in \mathbb{N}^+$ and $\upsilon_i \in [0, 1]$ obtained from Equation (1). Hence, we find a K dimensional matrix $V = [\widetilde{\upsilon_1}, \widetilde{\upsilon_2}, \ldots, \widetilde{\upsilon_k}]$. To determine clonality of the variant space V, we employ two different techniques: first, parametric stochastic modeling with Gaussian finite mixture modeling [41] and second, bootstrapping estimated cluster stability modeling [42]. Here, V can be interpreted to be a random sample of K independent identically distributed random variables. A joint probability function defined with a finite mixture model of C components takes the form

$$f\left(\upsilon_i, \theta\right) = \sum_{j=1}^{C} \alpha_j f_j\left(\upsilon_j, \varphi_j\right) \tag{2}$$

Where the parameter space of Equation (2) is defined by $\theta = (\alpha_1, \alpha_2, \ldots, \alpha_{C-1}, \varphi_1, \varphi_2, \ldots, \varphi_C), \alpha_1, \alpha_2, \ldots, \alpha_{C-1}$ are the weights of the C mixture components given $\sum_{j=1}^{C} \alpha_j = 1, \alpha_j \in \mathbb{R}^+$; and marginal density function of υ_j is $f_j(\upsilon_j, \varphi_j)$ with parameter φ_j. Applying the assumptions of a gaussian mixture model, the marginal density of each mixture component follows a multivariate normal distribution $N(\mu_j, \Sigma_j)$. As each of the mixture component represents an ellipsoidal mutually exclusive cluster, these are centered at μ_j and the shapes are defined by the variance–covariance matrices Σ_j, respectively. Henceforth, Bayesian information criteria is computed with penalized log likelihood (loglikelihood at maximum likelihood estimate—penalty) so that with increasing likelihood in proportion to increasing number of mixture components, information loss is incorporated with the logarithm of the number of estimates. Thus, CRUST determines the optimum number of mixtures. Decided the ideal clustering parameter, the algorithm performs hierarchical multifactor analysis to obtain cluster centroids [43]. With the centroids determined, sample points are assigned to clusters with supervised k-means clustering [44]. To elaborate on the features of the data we include a provision for multivariate mixture component. This method can be utilized on such data with dimensionality reduced sample space, although for univariate samples of VAFs the first principal component and the original vector are identical.

In addition, the stability of the clusters can be assessed with strength of the clustering method [42]. This technique relies on bootstrap resampling and iterative clustering. As the resampling is performed *B* times, the minimum mean Jaccard index is observed and scaled over the *B* resamplings. With an increase in the number of predicted clusters over the number of true clusters, the stability (defined by S_{min}) decreases affecting the Jaccard index as a given independent cluster is split in random sub-clusters. We leverage S_{min} to obtain the number of clusters devising partitions closest to true separation in VAFs. The optimum number of clusters thus observed is used to fit a *k*-means clustering algorithm with the centroids obtained from the *B* resamplings.

Clonality determination

In absence of clone size estimates corresponding to a specific variant, a cluster-wise detection of subclonality by available algorithms often lacks prediction of clonality. If the longitudinal sequence in which a variant appears in the tumor is not apparent, then we cannot distinguish a clonal event from a subclonal one. We conceived a metaheuristic process to determine the clonality of identified clusters. First, an iterative unsupervised density-based neighbor joining algorithm is used to cluster the VAFs by varying epsilon boundaries from 0.1 to 0.3. The epsilon boundaries define the minimum margin around the cluster centroid, which passively characterizes the distance between the centroids at initiation. In a diploid, balanced copy number state, the separation between centroids of a clonal and a subclonal cloud is expected to lie in this range [13]. An emergent feature of this unsupervised clustering with the prediction obtained from the algorithm above is that a pairwise contingency comparison of the two produces a matrix (*M*) spanned with either mutually orthogonal vectors or that belonging to same subspace i.e.If

$$M^{n,k} = \left[\widetilde{m}_1, \widetilde{m}_2, \widetilde{m}_3, \ldots, \tilde{m}_k\right],$$

where *n* is the optimal number of predicted (or user defined) clusters and *k* is that obtained from the density-based algorithm; then,

either $\tilde{m}_i \perp \tilde{m}_j$ or $\tilde{m}_i \ni \tilde{m}_j \ni basis\ S_l \forall i, j; 0 \leq \mid S_l \mid \leq n$.

This property ensures that a set of predicted clusters is reduced to a set of vectors belonging to the basis of a *d* dimensional space that spans the vector space defined by the *n* clusters. The orthogonal vectors and corresponding clusters are linearly independent as the linearly dependent clusters are merged to reduce the matrix to have *d* rows. Hence, we obtain a set of clusters with cardinality $d \leq n$. The clonality is assigned sequentially thereafter with a reductionist assumption that the subclones in a sample are represented in the left heavy tail of the VAF distribution if noise corrected [13]. This is justified by largely late-arriving subclonal events and stronger signal dampening experienced by rarer variants [45].

Estimation of allelic composition

Somatic copy number alterations are notoriously capable of modifying the relative abundance of a variant from sample to sample. Since subclonal copy number events can often occur during cancer progression, stratification of VAFs by allelic composition produces a better clonality prediction. CRUST stratifies variants according to baseline copy number at start and performs separate analyses for each stratum. In the absence of SNP array or other specific copy number data, the allelic composition is estimated from two separate standalone processes that enquires somatic tumor variants along with polymorphisms in the constitutional genome [46, 47]. These predict variant specific copy number by modeling mean relative depth ratio and segment specific B allele frequency for each estimated segmental copy number alteration, adjusting for cellularity, and overall tumor ploidy. Estimation of allelic segmentation from bulk sequencing data is known to contain several sources of bias. GC content is one such major source of bias, which induces an undulating pattern in allele frequencies [48]. As a measure of quality control, prior to copy number estimation, CRUST normalizes the B allele absolute frequencies against small segmental GC content with penalized lasso regression [49]. Additionally, it rescales the same against aggregate allele frequency to account for technical artifacts. Subsequently, segmental copy numbers are estimated with *copynumber* or *falcon* [47, 50].

CRUST visualizes segmental copy numbers by estimating allelic imbalance ratio and average log relative coverage ratio. To this end, both are estimated over genome-wide dynamically created chromosomal segments. A principal component analysis is then leveraged for distance-based pruning to discard outlying variants in each segment. Allelic imbalance ratios are calculated with mirrored B allele frequencies as described elsewhere [16]. As median log ratio is assumed to reflect segmental copy number and allelic imbalance ratio dictates the relative state of segmental zygosity, the unknown states of segmental copy number are reflected in the cluster positions of the segments. Presence of a subclonal copy number event is thus indicated by departure of a smaller segmental cluster from its parental cluster.

Calibration on simulated data

We presume the apparent alteration in representation of a variant in samples from the same tumor, given the same copy number state, is due to variation in the purities in each sample. The following stochastic modeling of the variability in variant abundance generated synthetic sample data closely resembling that of a solid tumor. Leveraging the heavy tailedness of log-normal distribution, a random variable with varying parameters drew sample points representing departure in VAFs from the true distributions due to fluctuating purities. We varied the mean of the distribution according to the logarithm of the re-centered true mean and re-scaled empirical variance in three separate set ups. Assuming a 100% purity and copy-neutral (disomic) ploidy status, true mean of VAFs for the clonal variants was assumed to be at 0.5 and that of the subclonal variants at 0.25 [51]. First only mean then only variance and subsequently both mean and variance dynamically were used to affect the simulations by changing the nature of variability in the data (Supplementary Figure 2A). Changing purity of the samples would always affect the mean of the VAFs as allele frequencies would converge towards the left tail. Whereas sequencing artifacts may not change the centroid of the clonal or subclonal clusters but introduce intra-cluster dispersion, which makes the distribution heavy tailed. To mimic this, we altered the variance keeping mean unchanged. The samples were drawn with these three different strategies to compare how restructuring the clonality of the variants may affect the deconvolution efficacy. For all iterations, the purities were sequentially varied from 0.95 to 0.5. We used polychoric correlation estimates computed between scaled and unscaled predictions to measure level of concordance [52].

As sequencing depth contributes immensely towards reduction of the baseline noise in signals, the following adaptation describes a probabilistic model conceived for the above-mentioned simulation. VAFs from a well-covered region can be reliably estimated with a beta distribution [12], thus a beta-binomial process with beta priors can be used for simulation [53]. To incorporate a log normal deviation based on purity, we rendered a convolution instead. A joint distribution of beta-log normal was used for this [54], the probability density function of which is thus defined by,

$$f_{(x)} = \exp\left(\frac{-\sigma^2}{2(\log_x - \mu)^2}\right) \Big/ x\sigma\sqrt{2\pi}B\,(a, b)\; \Phi\left(\frac{\log_x - \mu}{\sigma}\right)^{a-1} \tag{3}$$

$$\Phi\left\{1 - \left(\frac{\log_x - \mu}{\sigma}\right)\right\}^{b-1}, x \in \mathbb{R}^+$$

It jointly varies on the aggregates of the parameters of the two marginal distributions (a, b, μ and σ). The re-parameterized cumulative density function given by $F_{(y)}$ was inversed to obtain a canonical population of VAFs with subclonal mutations. A standardizing transformation on x in Equation (3) gets us,

$$F_{(y)} = I_{[\Phi(y)]}\,(a, b)\,, y \in \mathbb{R} \tag{4}$$

Where $y = \frac{\log_x - \mu}{\sigma}$. This method produces the prior distribution of a hypothetical sample. For a true clonal population, we center the marginal beta parameters with mean 0.5 (0.25 for subclonal) and variance 0.001. If these are denoted as a random variable $B \sim \beta$ (a,b) then, a closed form of the beta-log normal variable is given by $X = e^{\sigma\Phi^{-1}(b)+\mu}$. Provided an expected sequencing coverage of λ, if n_{sm}denotes the number of reads for a mutation M of sample S, then $N_{sm} \sim Poisson(\lambda)$. Presuming the variants are all biallelic, we can further estimate the read count r of an allele of the mutation following $r_{sm} \sim Bin(n_{sm}, f_{(x=sm)})$. Hence, the VAF for that allele is $\frac{r_{sm}}{n_{sm}}$. This process progressively aggregates noise as the coverage downsizes and the purity degrades (Supplementary Figure 2B).

Performance testing of scaling with simulation

We simulated an admixture of clonal and subclonal variants assuming a balanced background copy number. A parametric setup for the sample generation was favored under the assumption that sample purity indirectly affects the observed distribution of the relative abundance of variants. Three separate parametric assumptions were tested to obtain mutual concordance in prediction. All formulations are described in Methods and will here be referred to by the *mean*, *variance* and *meanvar*. For all three assumptions, 1500 simulations were performed, respectively, to generate two samples in each iteration. Purities were sequentially modified for each iteration to have produced 100 variants for every single sample. Sequencing coverages were also varied between 50× and 300× for simulating VAFs for each variant and were drawn randomly. As means vary, the clonal distribution and the subclonal tails are linearly translated on the x-axis, which does not affect the clustering if the respective centroids are not too close. We simulated the samples so that the mean and variance parameters (or, for beta distribution: scale and location) are directly affected by the purity and not the outcome sample. Therefore, the changes in purity does not linearly transform to changes in observed VAF distribution; it is instead the population distribution that changes, and samples are drawn from it. When variance was changed, the deconvolution and prediction accuracy broke down fastest as the two VAF distributions could retain the centroids while the range increased making them overlap. Hence the VAF distributions overlap between clusters making them virtually impossible to segregate.

Next, we compared CRUST against some of the contemporary and frequently used clustering algorithms, i.e. MAGOS and sciClone [11, 12]. We refrained from comparison against another popular method PyClone as both MAGOS and sciClone reliably outperforms it [11]. As these methods predominantly are clustering algorithms, there were certain assumptions we had to consider for the sake of contrast. Although CRUST is built to handle non-recurring variants that can appear in a (or disappear from) samples extracted in different stages of progression from a tumor, sciClone and MAGOS operate under the assumption that each ascertained mutation is present in all analyzed samples. Therefore, we used this as a basis for the subsequent comparisons. We also operated with the presumption of samples being copy neutral and clonally consistent; i.e. subclonal populations do not undergo clonal sweep or fixation. For these comparisons we also refrained from the user driven inputs. Each simulation produced two admixture samples for every run of the procedure. Coverages were varied to generate four distinct sets of observations at 300×, 100×, 50× and 30× depth. Among the two samples included in each observation, the first was consistently drawn with a purity higher than that of the second, sequentially reducing both with every iteration. The purities were thus progressively lowered from 0.95 to 0.5 for the first sample whereas the other one was continually initiated with a departure of 0.2 in purity from that of the former and was successively lowered until 0.25. Hence the theoretically lowest quality of sample was restricted with a paired purity of 0.5 and 0.25 with 30× coverage. At each combination we drew 10 observations consisting of a pair of 10 samples with 500 variants each. A sample thus drawn consisted of one clonal and one subclonal population.

We measured the performances of each method with the Jaccard index (τ, also known as Tanimoto similarity index) to quantify cluster agreement by creating a contingency table between the original population assignments and the predictions [55]. This statistic varies between 0 and 1, 1 indicating the highest possible concordance and can be interpreted similar to a correlation coefficient. We discarded CRUST's internal prediction on variant clonality justifiably compromising information in sake of ordinality and generalizability of the classifications. Here it is worth mentioning that sciClone has previously shown to perform reliably given at least a coverage of 300× [30]. Furthermore, MAGOS has been shown to perform almost as good if not better at 300× [11].

Deconvolution on hematological tumor patient samples

We further assessed constraints due to sequencing depth using publicly available data from an acute myeloid leukemia patient (AML31; dbGaP: phs000159) consisting of two temporally separated samples corresponding to the primary (90.7% pure) and relapsed tumor tissue (36.2% pure). These samples were queried with deep whole genome sequencing (up to ∼312×), exome capture (up to ∼433×) and further ultradeep targeted sequencing with custom capture assays (>2000×). To ensure quality, we extracted summary statistics on a putative 'platinum list' of

SNVs (consisting 1343 high-quality validated sites) 8. Subsequent pruning with read count (>10) resulted in a total of 37 variants. Statistics from four assays were used for testing: whole genome sequencing, exome capture, custom capture, and custom ion torrent platform, each providing varied coverage (Supplementary Figure 7). Although the purpose of this analysis was to observe the capability of CRUST to successfully deconvolve the AML31 samples, we performed all the steps using the system inputs and did not specify any user driven subjective input so that one can compare the results with the original deconvolution performed with SciClone.

Deconvolution and phylogenetics of TRACERx tumors

To observe the effect of normalization and inclusion of allelic composition on the predicted clonality status in a large-scale mutational landscape of a typical solid adult tumor, we turned to publicly available data. However, since procuring multiple tissue biopsies from several regions or at separate time points from human tumors are not part of standard care there are only a handful of dedicated studies set up for such investigations. The TRACERx [56] study is one such endeavor where initially upwards of a hundred non-small cell lung cancer (NSCLC) patients were biopsied for tumor tissue, some multi-regionally [17]. We extracted WES data on 20 of these tumors along with their copy number profiles for which at least three viable regional samples were sequenced with at least one mutant variant present in more than one sample. Here, the aim was to select a tangible cohort that reflect a diverse genetic signature without having to re-analyze the whole TRACERx dataset. We extracted sequencing summaries on a total of 21 000 variants (median: 911) from the twenty patients averaging 260 single nucleotide variants (SNVs) per sample. Only 9400 variants passed the quality control.

Copynumber summary data were procured directly from the published study and were linked to each variant to obtain allelic composition. All 20 tumors were subjected clonal deconvolution performed with CRUST where all multiregional samples were analyzed in tandem (Supplementary Figure 8). Each unique allelic composition warranted a distinct run. Due to unavailability of sample purity data, we were unable to normalize the VAFs. The clonality determination was supervised in accordance with the allelic compositions and in case of VAFs clustering too close to be clustered across samples, all variants were presumed clonal. Next, clone sizes corresponding to each variant's VAF were estimated as follows [57]

$$T\hat{C}F = \frac{Base\ ploidy \times VAF}{VAF \times (Base\ ploidy - CN_{mutant} - CN_{wildtype}) + CN_{mutant}} \quad (5)$$

$$Clone\ size = \frac{VAF \times \left((CN_{mutant} + CN_{wildtype}) \times T\hat{C}F + 2\left(1 - T\hat{C}F\right)\right)}{M} \quad (6)$$

The base ploidy was determined as that of the clonal variant with the least total copy number; CN_{mutant} is the copy number of the mutant allele and $CN_{wildtype}$ is that of the wild-type allele, M is alleles harboring the variant. We reconstructed the clonal or subclonal clusters based on these clone sizes. From top down, all clone sizes within 15% of each other were aggregated (separately for clones and subclones). The median clone size of each such aggregate were made to reflect the size of that population. The clone sizes needed to be scaled up as the estimates reflected the effect of purity. For each sample, a scaling constant was determined as follows:

$$Scaling\ constant = \frac{1}{Max.of\ clone\ sizes\ of\ all\ clonal\ variants}$$

Clone sizes of all clonal as well as subclonal variants were multiplied with the scaling constant of the corresponding sample.

Additionally, we created spatially nested phylogeny of some of the tumors as done in the published study [17]. To this end clone sizes (for several clonal/subclonal clusters) corresponding to each allelic composition were aggregated under the assumption that variants with similar VAF and allelic composition belong to the same population. In case more than one such aggregate (with different copy number or from different samples) were seen to have similar clone size (i.e. within 10% of one another), they were inferred to belong to the same population. All such calculations were separately formed for clonal or subclonal clusters (see Supplementary Table 2; ***clonal nesting***). The population nesting pattern thus discovered were used to build phylogenies.

Key Points

- CRUST is a new computational tool that simultaneously calculates allelic compositions, predicts clonality of sequenced variants and helps detect underlying evolutionary processes in a bulk sequenced tumor.
- Consideration of changeable allelic composition across samples in a tumor aids proper subclonal reconstruction, which if absent jeopardizes phylogenetic analysis.
- Multiregional sequencing aids in uncovering spatial inclination of solid tumor evolutionary trajectories increasing the genomic and etiological resolution.

Software availability

CRUST depends on R (>4.0.0) and is available for download from GitHub repository https://github.com/Subhayan18/CRUST. We recommend installation from the precompiled repository.

Data availability

WES data on the twenty TRACERx tumors were extracted from cBioPortal (http://www.cbioportal.org/study/summary?id=nsclc_tracerx_2017). The SNP array summary of the TRACERx tumors were available in supplementary tables of the corresponding study [17]. AML samples are part of the *Whole-Genome Sequencing of Acute Myeloid Leukemia* study and is available via dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000159.v8.p4). The neuroblastoma samples are part of a previous study [4]. Simulated data are generated with randomized seed using R and the test data sets are included in the package build.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Authors' contribution

Conception and design and development of methodology: S.C., D.G.; data acquisition: S.C., A.V., D.G.; analysis and interpretation of data: S.C., J.K., A.V., D.G.; technical support: J.K., A.V.; contribution towards manuscript: S.C., J.K., A.V., N.A., D.G.

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.

Funding

The Swedish Research Foundation (2016-01022 to D.G.), the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the Royal Physiographic Society, LMK Foundation and the Gunnar Nilsson Cancer Foundation. A special thanks to Chris Miller and Malachi Griffith for their assistance in procuring data from community resources (dbGaP: phs000159).

References

1. Gerlinger M, Rowan AJ, Horswell S, *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;**366**:883–92.
2. Williams MJ, Werner B, Barnes CP, *et al*. Identification of neutral tumor evolution across cancer types. *Nat Genet* 2016;**48**:238–44.
3. Sottoriva A, Kang H, Ma Z, *et al*. A big bang model of human colorectal tumor growth. *Nat Genet* 2015;**47**:209–16.
4. Karlsson J, Valind A, Holmquist Mengelbier L, *et al*. Four evolutionary trajectories underlie genetic intratumoral variation in childhood cancer. *Nat Genet* 2018;**50**:944–50.
5. Shibata D. Heterogeneity and tumor history. *Science* 2012;**336**:304.
6. Longo DL. Tumor heterogeneity and personalized medicine. *N Engl J Med* 2012;**366**:956–7.
7. Dentro SC, Wedge DC, Van Loo P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb Perspect Med* 2017;**7**:a026625.
8. Swanton C. Intratumor heterogeneity: evolution through space and time. *Cancer Res* 2012;**72**:4875.
9. Yates LR, Campbell PJ. Evolution of the cancer genome. *Nat Rev Genet* 2012;**13**:795–806.
10. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 2017;**168**:613–28.
11. Ahmadinejad N, Troftgruben S, Maley C, *et al*. MAGOS: discovering subclones in Tumors sequenced at standard depths. *bioRxiv* 2019;**790386**. doi: https://doi.org/10.1101/790386.
12. Miller CA, White BS, Dees ND, *et al*. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* 2014;**10**:e1003665.
13. Caravagna G, Heide T, Williams MJ, *et al*. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat Genet* 2020;**52**:898–907.
14. Andersson N, Bakker B, Karlsson J, *et al*. Extensive clonal branching shapes the evolutionary history of high-risk pediatric cancers. *Cancer Res* 2020;**80**:1512–152.
15. Roth A, Khattra J, Yap D, *et al*. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 2014;**11**:396–8.
16. Rasmussen M, Sundström M, Kultima HG, *et al*. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* 2011;**12**:R108.
17. Jamal-Hanjani M, Wilson GA, McGranahan N, *et al*. Tracking the evolution of non–small-cell lung cancer. *N Engl J Med* 2017;**376**:2109–21.
18. Zhang Y-W, Zheng Y, Wang J-Z, *et al*. Integrated analysis of DNA methylation and mRNA expression profiling reveals candidate genes associated with cisplatin resistance in non-small cell lung cancer. *Epigenetics* 2014;**9**:896–909.
19. Ma J, Lu Y, Zhang S, *et al*. β-Trcp ubiquitin ligase and RSK2 kinase-mediated degradation of FOXN2 promotes tumorigenesis and radioresistance in lung cancer. *Cell Death Differ* 2018;**25**:1473–85.
20. Collisson EA, Campbell JD, Brooks AN, *et al*. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;**511**:543–50.
21. Boeckx B, Shahi RB, Smeets D, *et al*. The genomic landscape of nonsmall cell lung carcinoma in never smokers. *Int J Cancer* 2020;**146**:3207–18.
22. Facchinetti F, Bluthgen MV, Tergemina-Clain G, *et al*. <em>LKB1/STK11</em> mutations in non-small cell lung cancer patients: descriptive analysis and prognostic value. *Lung Cancer* 2017;**112**:62–8.
23. Skoulidis F, Goldberg ME, Greenawalt DM, *et al*. &lt;em&gt;STK11/LKB1&lt;/em&gt; mutations and PD-1 inhibitor resistance in &lt;em&gt;KRAS&lt;/em&gt;-mutant lung adenocarcinoma. *Cancer Discov* 2018;**8**:822.
24. Scrima M, De Marco C, De Vita F, *et al*. The nonreceptor-type tyrosine phosphatase PTPN13 is a tumor suppressor gene in non&#x2013;small cell lung cancer. *Am J Pathol* 2012;**180**:1202–14.
25. Sayed ME, Yuan L, Robin JD, *et al*. NOVA1 directs PTBP1 to hTERT pre-mRNA and promotes telomerase activity in cancer cells. *Oncogene* 2019;**38**:2937–52.
26. Feng X, Liu H, Zhang Z, *et al*. Annexin A2 contributes to cisplatin resistance by activation of JNK-p53 pathway in non-small cell lung cancer cells. *J Exp Clin Cancer Res* 2017;**36**:123.
27. Wang C-Y, Chen C-L, Tseng Y-L, *et al*. Annexin A2 silencing induces G<sub>2</sub> arrest of non-small cell lung cancer cells through p53-dependent and -independent mechanisms *. *J Biol Chem* 2012;**287**:32512–24.
28. Zhong R, Li H, Liu Y, *et al*. Chemotherapy combined with bevacizumab for the treatment of advanced lung adenocarcinoma cancer harboring EGFR-ANXA2, EGFR-RAD51, ATR and BRCA2 mutations: a case report. *Thoracic Can* 2020;**11**:456–60.

29. Yu T, Chen X, Zhang W, *et al*. KLF4 regulates adult lung tumor-initiating cells and represses K-Ras-mediated lung cancer. *Cell Death Differ* 2016;**23**:207–15.
30. Griffith M, Miller Christopher A, Griffith Obi L, *et al*. Optimizing cancer genome sequencing and analysis. *Cell Systems* 2015;**1**:210–23.
31. Pang S, Wu L, Shen X, *et al*. The evolutionary landscape of pan-cancer drives clinical aggression. *bioRxiv* 2018;**422667**. doi: https://doi.org/10.1101/422667.
32. Andor N, Graham TA, Jansen M, *et al*. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 2016;**22**:105–13.
33. Gerstung M, Jolly C, Leshchiner I, *et al*. The evolutionary history of 2,658 cancers. *Nature* 2020;**578**:122–8.
34. Andor N, Harness JV, Müller S, *et al*. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* 2014;**30**:50–60.
35. Deshwar AG, Vembu S, Yung CK, *et al*. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* 2015;**16**:35.
36. Chattopadhyay S. CRUST_1.0.0. *bioRxiv* 2020. https://github.com/Subhayan18/CRUST
37. Carter SL, Cibulskis K, Helman E, *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21.
38. Su X, Zhang L, Zhang J, *et al*. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* 2012;**28**:2265–6.
39. Takahashi T, Matsuda Y, Yamashita S, *et al*. Estimation of the fraction of cancer cells in a tumor DNA sample using DNA methylation. *PLoS ONE* 2013;**8**:e82302.
40. Dieterle F, Ross A, Schlotterbeck G, *et al*. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR Metabonomics. *Anal Chem* 2006;**78**:4281–90.
41. Scrucca L, Fop M, Murphy TB, *et al*. Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* 2016;**8**:29.
42. Yu H, Chapman B, Di Florio A, *et al*. Bootstrapping estimates of stability for clusters, observations and model selection. *Comput Stat* 2019;**34**:349–72.
43. Le Dien S, Pagès J. Hierarchical multiple factor analysis: application to the comparison of sensory profiles. *Food Qual Prefer* 2003;**14**:397–403.
44. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw* 2008;**1**(1):2008.
45. Williams MJ, Werner B, Heide T, *et al*. Quantification of subclonal selection in cancer from bulk sequencing data. *Nat Genet* 2018;**50**:895–903.
46. Favero F, Joshi T, Marquard AM, *et al*. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015;**26**:64–70.
47. Chen H, Bell JM, Zavala NA, *et al*. Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* 2014;**43**:e23–3.
48. Ross MG, Russ C, Costello M, *et al*. Characterizing and measuring bias in sequence data. *Genome Biol* 2013;**14**: R51.
49. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;**40**:e72.
50. Nilsen G, Liestøl K, Van Loo P, *et al*. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* 2012;**13**:591.
51. Nik-Zainal S, Van Loo P, Wedge David C, *et al*. The life history of 21 breast cancers. *Cell* 2012;**149**:994–1007.
52. Ekström J. *A Generalized Definition of the Polychoric Correlation Coefficient*. UCLA: Department of Statistics, UCLA, 2011.
53. El-Kebir M, Oesper L, Acheson-Field H, *et al*. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 2015;**31**:i62–70.
54. Castellares F, Montenegro LC, Cordeiro GM. The beta log-normal distribution. *J Stat Comput Simul* 2013;**83**:203–28.
55. Rogers DJ, Tanimoto TT. A computer program for classifying plants. *Science* 1960;**132**:1115.
56. Jamal-Hanjani M, Hackshaw A, Ngai Y, *et al*. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol* 2014;**12**:e1001906.
57. Yasui H, Valind A, Karlsson J, *et al*. A dynamic mutational landscape associated with an inter-regionally diverse immune response in malignant rhabdoid tumour. *J Pathol* 2020;**252**:22–8.