Integrated quality control of allele-specific copy numbers, mutations and tumour purity from cancer whole genome sequencing assays

Jacob Househam, *Evolution and Cancer Lab, Centre for Genomics and Computational Biology, Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK*
Riccardo Bergamin, *Department of Mathematics and Geosciences, University of Trieste, Italy*
Salvatore Milite, *Centre for Computational Biology, Human Technopole, Italy*
William CH Cross $^{(\star)}$, *UCL Cancer Institute, University College London, UK*
Giulio Caravagna $^{(\star)}$, *Department of Mathematics and Geosciences, University of Trieste, Italy*

$^{(\star)}$Joint last authors.

Corresponding: (GC) gcaravagna@units.it.

Abstract. Cancer is a global health issue that places enormous demands on healthcare systems. Basic research, the development of targeted treatments, and the utility of DNA sequencing in clinical settings, have been significantly improved with the introduction of whole genome sequencing. However the broad applications of this technology come with complications. To date there has been very little standardisation in how data quality is assessed, leading to inconsistencies in analyses and disparate conclusions. Manual checking and complex consensus calling strategies often do not scale to large sample numbers, which leads to procedural bottlenecks. To address this issue, we present a quality control method that integrates somatic point mutations, allele-specific copy numbers, and tumour purity into a single quantitative score. We demonstrate its power via simulations, and on $n = 2778$ whole-genomes from PCAWG, on $n = 10$ multi-region whole-genomes of two colorectal cancers and on $n = 48$ whole-exomes from TCGA. Our approach significantly improves the generation of cancer mutation data, providing visualisations for cross-referencing with other analyses. The method is fully automated and designed to be compatible with any bioinformatic pipeline, and can automatise tool parameterization paving the way for fast computational assessment of data quality in the era of whole genome sequencing.

Introduction

Cancer remains an unsolved problem, and a key factor is that tumours develop as heterogeneous cellular populations (Greaves and Maley 2012; McGranahan and

Swanton 2017, 2015). Cancer genomes can harbour multiple types of mutations compared to healthy cells (Macintyre et al. 2018; Martincorena et al. 2018, 2015; Nik-Zainal et al. 2012), and many of these events contribute to the pathogenesis of the disease, and therapeutic resistance. A popular design of studies intending to understand tumour development involves collecting tumour and matched-normal biopsies, and generating so-called “bulk” DNA sequencing data to identify both germline and tumour somatic mutations (Barnell et al. 2019). Using bioinformatic tools to cross reference the normal genome against a paired aberrant one, the mutations and heterogeneity thereof found in the tumour sample can be derived and used in other analyses. These analyses include, but are not limited to, driver mutation identification (Bailey et al. 2018; Gonzalez-Perez et al. 2013), which aims to discern the key aberrations that cause a tumour to grow, patient clustering, which aims to identify treatment groups with similar biological characteristics, and evolutionary inference (Ding et al. 2012; Landau et al. 2013; Caravagna et al. 2016; Jamal-Hanjani et al. 2017; Turajlic et al. 2018; Caravagna et al. 2018; Roth et al. 2014; Miller et al. 2014; Cross et al. 2018; Gerstung et al. 2020; Deshwar et al. 2015; Strino et al. 2013), which unravels how a particular tumour developed from normal cells.

There are several types of mutations that we can retrieve from DNA sequencing (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Broadly these can be categorized as single nucleotide variants (SNVs), copy number alterations (CNAs) and other more complex changes such as structural variants (Li et al. 2020; Zack et al. 2013). All types of mutations can drive tumour progression, and are therefore important entities to study (Kent and Green 2017; Levine, Jenkins, and Copeland 2019). Luckily, the steady drop in sequencing costs is fueling the creation of large datasets for researchers to access through public databases. Notably, we are entering the era of high-resolution whole-genome sequencing (WGS), a technology that can read out the majority of a tumour genome, providing significant improvements over whole-exome or targeted counterparts. Generating some of these data, however, poses challenges. While SNVs are the simplest type of mutations to detect using bioinformatic analysis and perhaps have the most well established supporting tools (Li et al. 2020), CNAs are particularly difficult to call since the baseline ploidy of the tumour (i.e., the number of chromosome copies) is usually unknown and has to be inferred (Van Loo et al. 2010; Favero et al. 2015; Boeva et al. 2011; Poell et al. 2019; Cun et al. 2018; Fischer et al. 2014). CNAs are important types of cancer mutations; large-scale gain and loss of chromosome arms or sections of arms can confer tumour cells with large-scale phenotypic changes, and are often important clinical targets (Gerstung et al. 2020; Watkins et al. 2020).

SNVs and CNAs are intertwined mutation groups. They can overlap within a tumour cell's genome, meaning the number of copies of an SNV can be amplified or indeed reduced by CNAs. This depends on the ploidy of the genome regions overlapping with the variants. For instance, for a *clonal* - meaning present in every cell of the tumour sample - heterozygous SNV in a diploid tumour genome the expected variant allele frequency (VAF) is 50% (i.e., half of the reads from tumour cells will harbour the SNV). Alternatively, if each chromosome is present in three copies (triploid), the expected VAF is 33%, for SNVs occurring after amplification (or on the non-amplified chromosome), or 66%, for SNVs on the amplified chromosome. The theoretical frequencies are observed with a Binomial noise model that depends on sequencing depth and VAF (Nik-Zainal et al. 2012; Caravagna, Heide, et al. 2020; Roth et al. 2014; Miller et al. 2014; Strino et al. 2013; Tarabichi et al. 2021; Yuan et al. 2018). We note that these VAFs hold for pure bulk tumour samples (100% tumour cells). Realistically, most bulk samples contain normal cells, the percentage of which shifts these theoretical frequencies towards lower values. These ideas are leveraged by methods that seek to compute the Cancer Cell Fractions (CCFs) of the tumour, i.e., a normalisation of the observed tumour VAF for the CNA, the number of copies of a mutation (mutation multiplicity) and tumour purity (Dentro, Wedge, and Van Loo 2017).

Many bioinformatics pipelines are designed to start from a BAM formatted input file and, following variant calling, extract the VAF of mutations while calling CNAs (Boeva et al. 2011; Cmero et al. 2020; Zaccaria and Raphael 2020; Van Loo et al. 2010; Fischer et al. 2014; Carter et al. 2012). These analyses are nearly always decoupled, and can return inconsistent variant calls; i.e., CNAs and purity that mismatch the empirical VAF from the BAMs. Since CNAs and purity are inferred through various measurements that are subject to noise - i.e., tumour-normal depth ratios and B-allele frequencies are prime examples - they are the most likely cause of error. While in some cases these errors can be spotted and fixed by manual intervention, this process is also subject to inconsistencies in the absence of a proper statistical framework, and does not scale in studies seeking to generate very large datasets (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Priestley et al. 2019; Turnbull et al. 2018). The intrinsic performance of a variant caller and sequencing noise therefore massively impacts CNA calling and purity inferences, propagating errors in downstream analysis that eventually lead to incorrect biological conclusions, becoming a crucial computational bottleneck in the era of high-resolution whole-genome sequencing.

To solve these problems we developed CNAqc, a computational framework with a *de novo* statistical model to assess the conformance of SNVs, CNAs, and purity estimates. We strived to make the tool as simple as possible, maximising compatibility across

differing pipelines. CNAqc computes a quantitative quality control (QC) score for the overall agreement of the calls, which can be used to tune the parameters of callers (e.g., decrease or increase purity), or select among multiple profiles (e.g., tetraploid versus diploid tumours) until a good fit is achieved. In CNAqc we also integrate these measures to determine Cancer Cell Fractions (CCF) after phasing mutation multiplicity from VAFs (Dentro, Wedge, and Van Loo 2017).

CNAqc is implemented as a highly optimised R package which can be used between somatic calling and downstream analyses (Figure 1a). CNAqc has a small computational overhead compared to typical downstream analyses, e.g., subclonal deconvolution, which are much more complicated because they interpret the clonal and subclonal VAF spectrum (Gerstung et al. 2020; Nik-Zainal et al. 2012; Caravagna, Heide, et al. 2020; Roth et al. 2014; Miller et al. 2014; Jamal-Hanjani et al. 2017). The tool can process both WGS and WES data, and can automatically compute a QC score in a matter of seconds, making it extremely useful for large-scale genomics consortia or retrospective analyses of public datasets. To demonstrate the tool we analysed $n = 2723$ high-quality whole-genomes from the Pan Cancer Analysis of Whole Genomes (PCAWG) cohort (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), $n = 10$ high-resolution bulk whole-genomes datasets from two multi-region colorectal cancers, and $n = 48$ whole-exomes from The Cancer Genome Atlas (TCGA) cohort (Cancer Genome Atlas Research Network 2014).

Results

The CNAqc framework

CNAqc integrates clonal CNAs, tumour purity and somatic mutation calls obtained from bulk sequencing (Figure 1). The tool is intended to be used after variant calling, and before downstream analysis (Figure 1a), to compute a quality control score for allele-specific CNAs and purity based on mutation VAFs, determining a PASS or FAIL status for each segment type and the overall sample. CNAqc can also be used to select among alternative genome segmentations and purity/ploidy estimates available from a caller (e.g., a 100% pure diploid tumour versus a 50% pure tetraploid). The score also suggests corrections for tumour purity to fine-tune tools that use Bayesian priors or point parameters. Lastly, CNAqc can determine Cancer Cell Fractions (CCFs) for input mutations, together with PASS or FAIL status; mathematical details are available in the Online Methods.

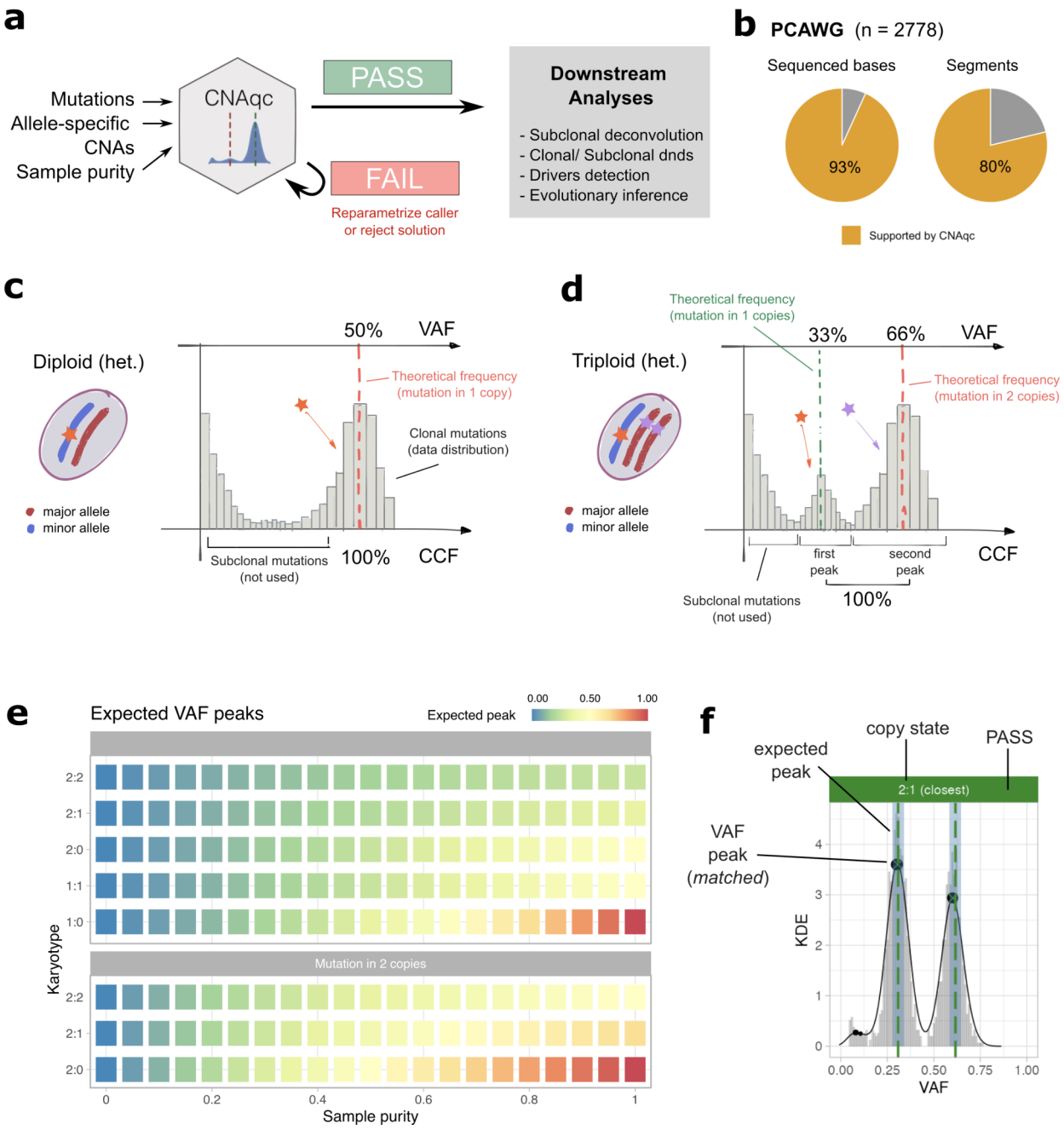


Figure 1. a. CNAqc integrates somatic mutations, allele-specific copy number segments and tumour purity estimates to determine a sample-level PASS or FAIL status that reflects the consistency between the inputs. The status depends on a quantitative score assigned to the calls, and a used-defined error tolerance. When a sample passes quality control, the calls can be safely used for downstream analysis (e.g., subclonal deconvolution, etc.). Otherwise, CNAqc suggests adjustments to the current purity estimates, which can be used to re-parametrise the copy number caller. CNAqc can also be used to select among multiple fits to the input data that can be returned by a caller (e.g., a diploid solution with twice the purity of an equivalent tetraploid). **b.** CNAqc supports copy states 1:0 (LOH), 1:1 (heterozygous diploid), 2:0 (copy neutral LOH), 2:1 (triploid) and 2:2 (tetraploid). These span ~93% of bases sequenced and ~80% of the 600,000 segments available in $n = 2778$ PCAWG whole-genome samples. **c.** Theoretical VAF histogram for diploid 1:1 mutations in a tumour. A clonal heterozygous mutation has 50% VAF; all mutations are observed with Binomial sequencing noise. The clonal mutations form a peak at 100% CCF, plus other

features that characterise the tumour clonal composition (e.g., subclonal mutations, which are not required to quality control clonal CNAs). **d.** The analogous case of a 2:1 tumour genome segment, where we expect 2 peaks in the VAF originating from mutations present in one or two copies. The multiplicity of a mutation can phase whether it happened before or after the CNA. For 2:1 we expect peaks at 66% and 33% VAF: both represent clonal mutations with 100% CCF. **e.** Heatmap expressing equation (1), the relationship between allele-specific copy number segments, mutation multiplicity and sample purity. The color reflects the expected VAF for the corresponding clonal mutations - i.e., if CNAs and purity are correct, we expect a peak of clonal mutations at the corresponding matrix value. The distance between the expected VAF and the empirical peaks in the VAF data gives the score metric used to quality control a sample. **f.** Example quality control for SNVs mapping in triploid 2:1 heterozygous segments with 2 copies of the major allele, and 1 copy of the minor allele, in a WGS assay reporting ~90% purity. The horizontal dashed lines are the expected VAF peaks $v_1$and v_2, determined from the table in panel (e) for the mutations in single or double copy. The black dots on the data are the peaks estimated by CNAqc, which determines that 2:1 segments and tumour purity match the VAF data distribution since peaks fall within the area shaded around $v_1$and v_2. The PASS status of the sample is reflected by the green bar.

In what follows, we will refer explicitly to SNVs as the main type of mutation used by CNAqc, but in principle other types of substitutions such as insertions or deletions also apply. The method supports clonal heterozygous normal states (1:1 chromosome complement), loss of heterozygosity (LOH) in monosomy (1:0) and copy-neutral (2:0) form, trisomy (2:1) or tetrasomy (2:2) gains. According to data (Figure 1b) available in $n = 2778$ PCAWG samples (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), the most common segments among human cancers are supported by CNAqc (>80% of ~600.000 total PCAWG segments, 93% of sequenced bases, much more prevalent than subclonal CNAs, Supplementary Figure S1). Therefore, besides rare exceptions, CNAqc can analyse any cancer sample.

Many output metrics are derived from the link between “copy states” (i.e., the copies of the major and minor alleles, which sum up to the ploidy of a segment) and allele frequencies that we find in mutation calls. Combinatorial equations inspired from ASCAT (Van Loo et al. 2010) are used to determine if CNAs and purity are consistent with observed VAFs (Online methods). The key equation links our belief about the *expected* VAF of a clonal mutation (i.e., the clonal VAF peak), if the input CNA segment and tumour purity were correct (Figure 1c, 1d). We consider a tumour sample with purity $\pi \in [0, 1]$, and all CNA segments with allele-specific copy numbers n_A and n_B for the major and minor alleles. If we consider mutations with multiplicity m and mapping to any of the segments with copy state $n_A : n_B$ (e.g., all 1:1, diploid heterozygous segments), we expect a VAF peak (Figure 1e) at

$$v_m = \frac{m\pi}{2(1-\pi) + \pi(n_A + n_B)} \quad (1)$$

Notation v_m explicits that the expected VAF, which is observed with Binomial noise, depends on the multiplicity of the mutation (Dentro, Wedge, and Van Loo 2017). For copy states 2:0 ($n_A = 2$, $n_B = 0$), 2:1 and 2:2, m phases mutations acquired before or after the copy number event (Figure 1d). CNAqc supports simple copy states (1:0, 1:1, 2:0, 2:1, 2:2) and restricts $m \in \{1, 2\}$, assuming the CNAs are acquired directly from diploid heterozygous normal states (1:1).

For each copy state, VAF peaks are detected via fast heuristics based on kernel density estimation and maximum likelihood Binomial mixtures (Supplementary Figure S2). An optimal peak is then selected and matched to v_m, measuring the VAF distance between which is then converted into units of purity (Online Methods). The CNAqc sample score $\lambda \in \Re$ (e.g., $+3\%$, -7%) is based on a linear combination of the distances, representing an *error* that approaches 0 for perfect calls, reflecting corrections to the input π. The cut to determine PASS or FAIL is the error $\epsilon > 0$ we can tolerate in the purity estimate: e.g., for heterozygous diploid mutations with $\epsilon = 0.025$ (2.5% maximum error) and real purity 60%, CNAqc will PASS a tumour purity estimate in $[55; 65\%]$, corresponding to VAF range $[27.50\%; 32.5]$. To normalise this error against aneuploidy and contamination, ϵ is adjusted for copy state, multiplicity and tumour purity (Online methods, Figure 1f).

CNAqc can also determine and score CCFs which are used for downstream subclonal deconvolution (Van Loo et al. 2010; Nik-Zainal et al. 2012). Assuming input CNAs and purity π are validated by peak detection, CNAqc normalises the VAF v of a mutation that sits on segment n_A: n_B in a tumour with the formula

$$c = \frac{v[(n_A + n_B - 2)\pi + 2]}{m\pi}. \quad (2)$$

This equation applies to clonal and subclonal mutations, and the main difficulty in obtaining the correct value for c is determining if the mutation is in single or double copy (multiplicity $m = 1$ or $m = 2$); we term this phasing m from the VAF spectrum.

CNAqc uses a heuristic based on a two-component Binomial mixture to compute multiplicities by clustering. The default method identifies a VAF range at the crossing of the mixture components, where m cannot be unequivocally phased. The phasing uncertainty is estimated from the entropy $H(z)$ of the mixture latent variables z. For every copy state, depending on the maximum proportion of unassigned mutations that

we decide to tolerate (e.g., 10% of total), a CCFs PASS or FAIL status is determined. An alternative method is available, which can force a value to m through a hard split on the VAF, regardless of entropy values (Online Methods).

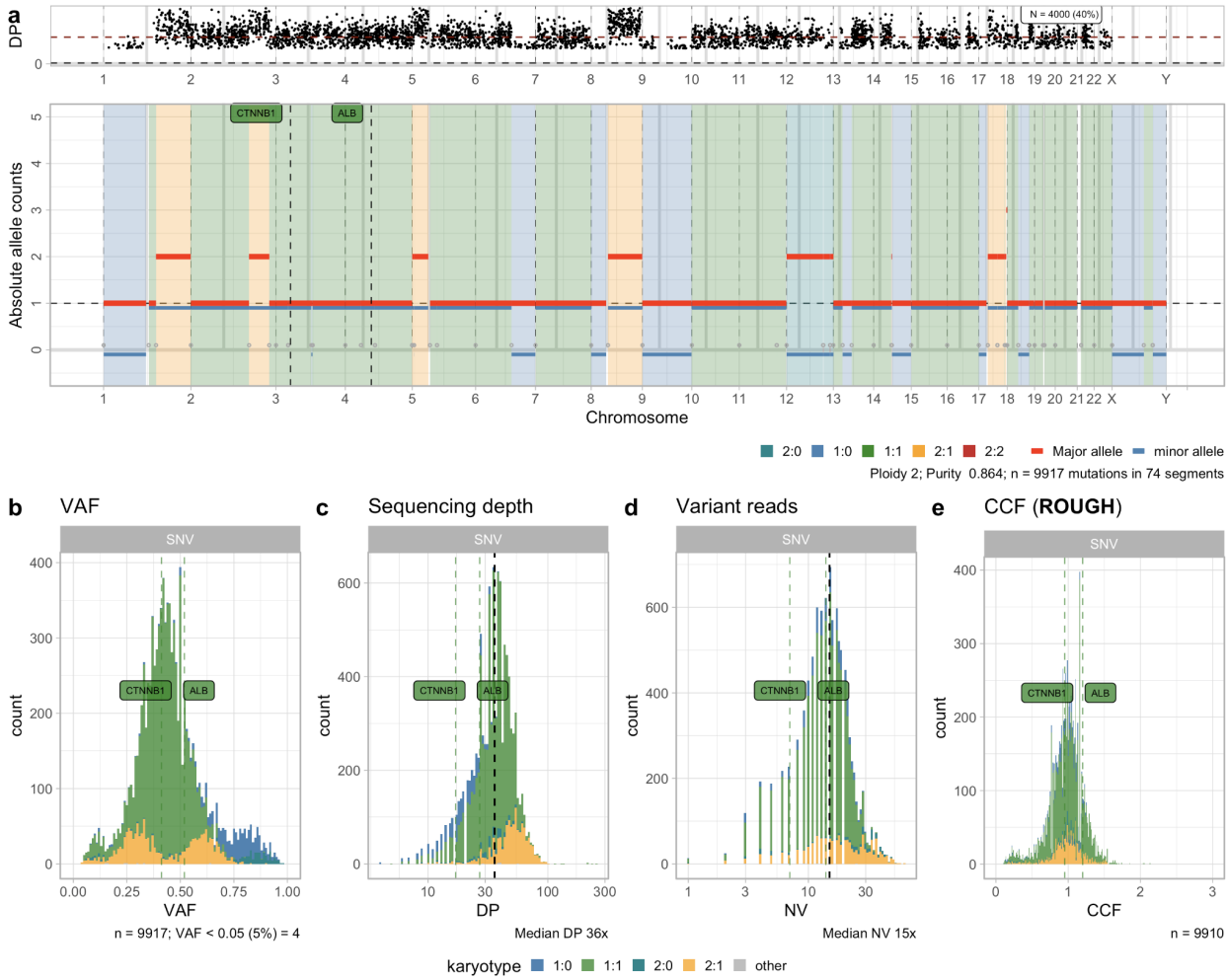


Figure 2. a. CNAqc visualisation of PCAWG sample ca5ded1c-c622-11e3-bf01-24c6515278c0 (DCC project code LIRI-JP, hepatocellular carcinoma). Genome-wide allele-specific consensus CNAs (ploidy 2, purity ~85%). The bottom plot reports the copies of the major and minor alleles in each segment, highlighting segments supported by CNAqc. PCAWG reported two driver SNVs hitting genes CTNNB1 and ALB, which appear annotated in diploid heterozygous segments (1:1). The top plot shows genome-wide somatic mutations with their depth of sequencing. **b,c,d.** Read count data for the input SNVs: Variant allele frequencies (VAFs), depth of sequencing (DP) and number of reads supporting the variant allele (NV). **e.** Cancer Cell Fractions (CCF) obtained from CNAqc for this sample show that the two CTNNB1 and ALB drivers are clonal.

CNAqc provides several functions to visualise segments and read count data (Figure 2), peak detection and CCFs (Figure 3), and utilities to smooth segments and detect patterns of over-fragmentation (Online Methods). This information can be used to

augment and prioritise downstream analysis that seeks to determine patterns of chromothripsis, kataegis or chromoplexy from mutation and copy number data (Zack et al. 2013; Gerstung et al. 2020).

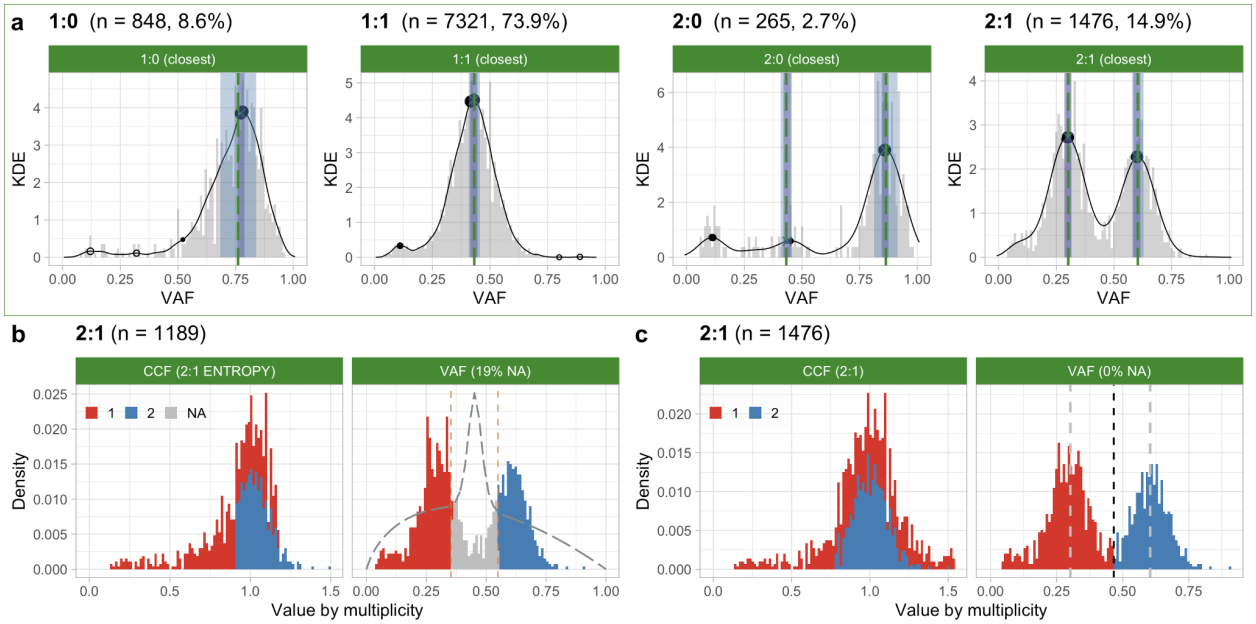


Figure 3. a. Peak detection analysis assessing the quality of CNA segments and tumour purity, split by copy state (see Figure 1f). Note that for copy states with total copy number >2 (2:1 and 2:0 here), multiple peaks are checked independently; the quality control status of a copy state depends on the number of mutations assignable to each peak, and whether the peak is matched or not. The sample-level status is a linear combination of results from each copy state (PASS here, surrounding green rectangle). **b.** Cancer Cell Fractions (CCF) estimation for mutations mapping to triploid (2:1) segments, obtained using the entropy-based method. The panels show CCFs and VAFs, coloured by phased mutation multiplicity (single or double copy). The entropy profile is the dashed line above VAF; areas in gray are crossings of Binomial densities where CNAqc cannot assign phaset multiplicity confidently from VAF (19% of mutations). The CCFs for this copy state are PASS when we accept 20% of maximum unknown multiplicities. **c.** Alternative CCF values computed by the CNAqc method that hard splits density peaks, and phases mutations regardless of uncertainty. In this case the split is at the mid point where we expect.

Simulations

We tested CNAqc on ~20,000 synthetic VAF distributions obtained for different values of coverage (30x, 60x, 90x, 120x) and purity (0.4, 0.6, 0.8, 0.95). For each dataset, we run CNAqc with the input purity corrupted by a variable error factor ϵ_{err}, and scan multiple levels of tolerance ϵ to match peaks.

We observed that the proportion of rejected samples approaches 100% when the purity error exceeds tolerance ($\epsilon_{err} > \epsilon$), suggesting that the model in CNAqc works as

expected, i.e., we *detect errors as big as tolerance*. From simulations, we could observe that VAF quality impacts performance, and that low coverage or purity make peak detection harder (Supplementary Figure S3).

For the same batch of tumours we computed CCFs to measure their uncertainty - i.e., the number of mutations that CNAqc cannot phase from VAFs. Low coverage and low purity generate VAF peaks that overlap, where exact multiplicity phasing becomes unachievable. The performance gradient highlights the importance of data quality to assess reliable CCFs (Supplementary Figures S4).

Large-scale pan cancer PCAWG calls

We have run CNAqc on the full PCAWG cohort, for which we gathered consensus calls from SNVs, allele-specific CNAs and purity ($n = 2778$ samples, 40 tumour types). Excluding samples with unsuitable data, we ran $n = 2723$ cases on a single multi-core machine in <1 hour (Figure 4).

Median depth of sequencing and purity are 45x and ~65% (Caravagna, Heide, et al. 2020), therefore the PCAWG resolution is comparable to the mid and low range of parameters adopted in our simulations (Supplementary Figure S3). As expected, peak detection passed $2425/2723$ samples using $\epsilon = 0.03$ error purity tolerance, confirming that PCAWG consensus calls are top quality (Figure 4a). As in our simulations, the acceptance rate was determined by tumour purity and coverage (Figure 4b), with purity adjustments distributed around 0 for PASS samples, spreading towards left or right for FAIL cases (Figure 4c).

Manual inspections of some samples presented some interesting cases. For instance, tumours with low burden but high quality calls still yielded a useful report (Supplementary Figure S6). Tumours with estimates of 100% purity which are at odds with VAF peaks might suggest purity over-estimation (Supplementary Figure S7), while other cases did possess genuinely very high purity (>95%, Supplementary Figure S8).

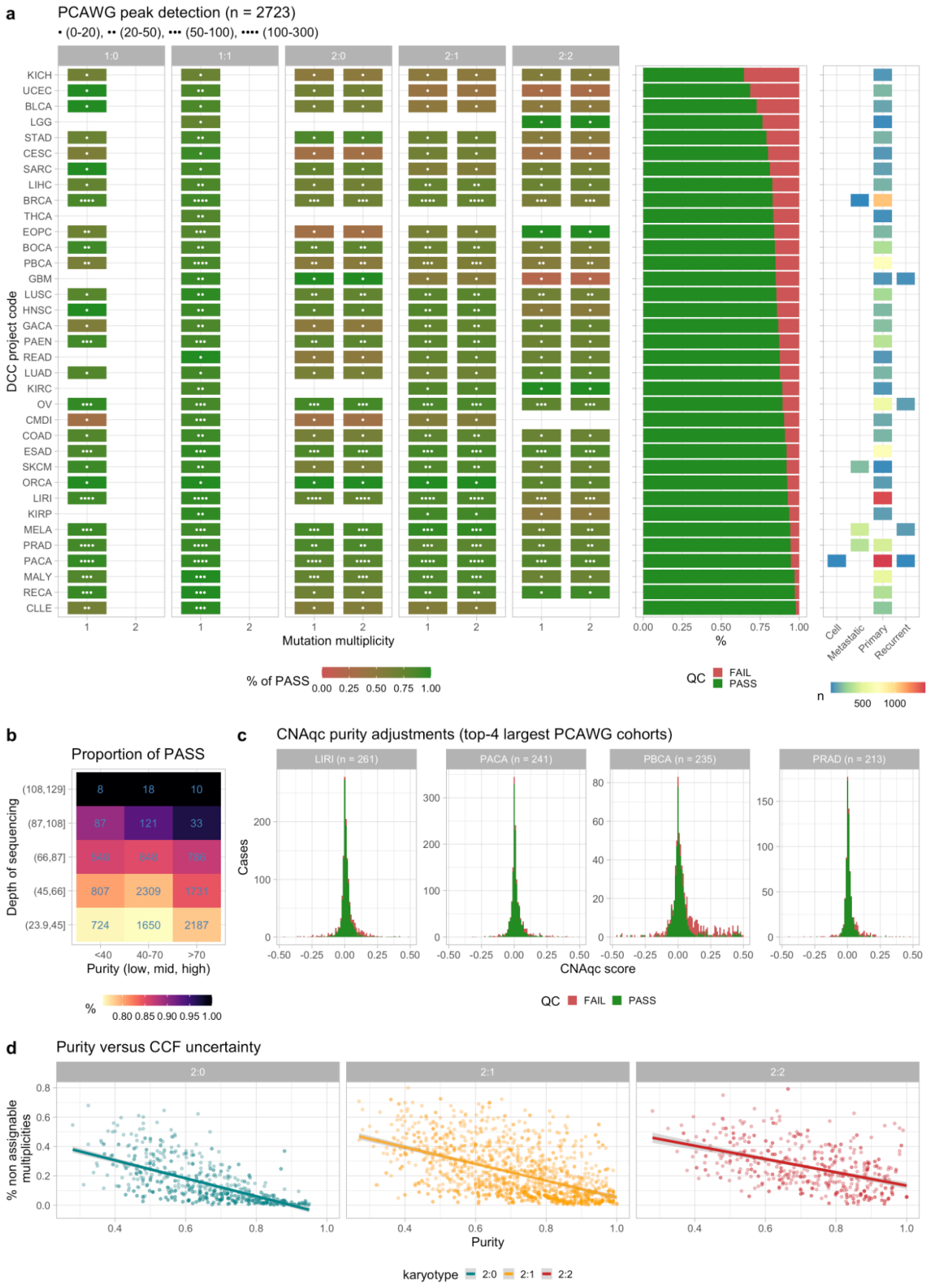


Figure 4. a. Peak detection quality control for $n = 2723$ WGS samples available in PCAWG. The plot shows the percentage of cases with PASS status, split by copy state, multiplicity and tumour type. The dots annotated report the number of cases, the barplot the tumour types sorted by percentage of sample-level FAIL cases and the coloured heatmap the sample classification (primary, metastatic etc.). **b.** Proportion of PASS cases split by purity (low, $< 40\%$, high, $> 70\%$, and mid-level) and median depth of sequencing (DP), after removing two samples with DP > 150. **c.** Histogram of peak distances (expected versus observed) for clonal CNA segments in the 4 tumour subtypes with most samples. The reported values are split by CNAqc PASS or FAIL status. **d.** Regression of tumour sample purity against the proportion of CCF values that cannot be confidently assessed by CNAqc, split by copy state.

CCFs were computed for the whole PCAWG cohort. Consistently with simulated data, the percentage of mutations for which CCF cannot be computed negatively correlated with sample purity (Figure 4d and Supplementary Figure S4). We found the CCFs produced by CNAqc (Supplementary Figure S9) are comparable to those computed by Ccube (Yuan et al. 2018) across the whole cohort, but also found cases where CNAqc helped to detect spurious subclonal clusters, which we could explain by miscalled mutation multiplicities (Supplementary Figure S10).

Summarising, while peaks could be determined for almost all PCAWG samples, mutation multiplicity assessment would have required higher coverage and purity. Our analyses reveal that every type of computation - peak detection or CCF - has different data quality requirements, and should therefore be quality controlled with specific methods like the ones available in CNAqc.

Multi-region colorectal cancer data

We have run CNAqc on previously published WGS multi-region data (Cross et al. 2018; Caravagna, Heide, et al. 2020), which was collected from multiple regions of primary colorectal adenocarcinomas (10 samples, 2 patients, median coverage ~80x, purity ~80%, Figure 5). We augmented somatic mutations called by Platypus (Cross et al. 2018) with allele-specific CNAs and purity from Sequenza (Favero et al. 2015), and used CNAqc to rank segments and purity obtained by multiple parameterizations of the tool, which were defined considering also the alternative fitting solutions proposed during the fit.

Sequenza was first run with the default range proposals for purity and ploidy, which we then improved in a final run following CNAqc. From the default Sequenza runs, we collected the proposed alternative solution, which was tetraploid 2:2 with halved purity. We used these parameters to compute a *de novo* Sequenza fit with ploidy ranging 3.8-4.2, together with a run constrained with low purity. Runs for sample Set7_57 (patient Set7) highlighted that both Sequenza (not shown) and CNAqc are strongly

confident about the diploid solution with the correct purity (Figure 5a). The peak detection scores produced by CNAqc invariably fail both the tetraploid and low purity solutions, passing the others; the little adjustment suggested to the default parameters slightly improves the purity, but the overall quality is high even with default parameters (Figure 5b) and the final segments for Set7_57 show mild aneuploidy (Figure 5c).

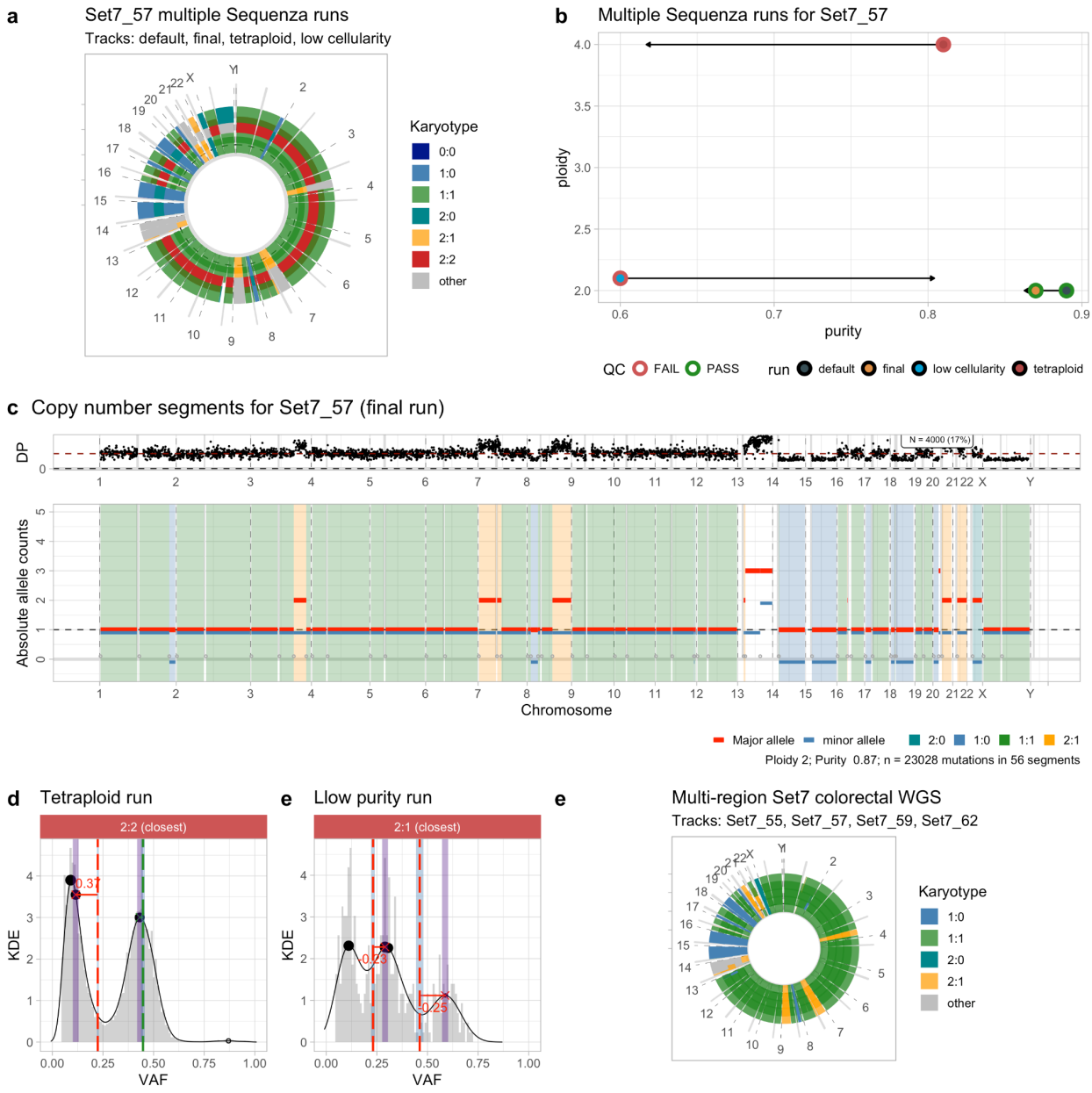


Figure 5. a. Circos plot for four possible whole-genome CNA segmentations determined by Sequenza with WGS data (~80x median coverage, purity 87%). The input sample is Set7_57, one of four multi-region biopsies for colorectal cancer patient Set7. The first run is with default Sequenza parameters. With CNAqc, we slightly adjust

purity estimation and obtain a final run of the tool. Following Sequenza alternative solutions, we also fit a tetraploidy solution to data, and one with maximum tumour purity 60%. **b.** Purity and ploidy estimation for the four Sequenza runs. Arrows show the adjustment proposed by CNAqc, the default and final runs are the only ones to pass quality control. **c.** Final run with Set7_57: allele-specific copy number segments and depth of coverage per mutation. **d,e.** Miscalled tetraploid and triploid segments in the tetraploid and low purity Sequenza solutions, identified by CNAqc. **e.** CNA calling with CNAqc and Sequenza for 4 WGS biopsies of the primary colorectal cancer Set7.

This case is instructive of how CNAqc can be used to assess miscalled CNA segments ahead of the VAF data, for both tetraploid and low purity solutions (Figure 5d,e). With CNAqc we obtained, in an completely automated manner, good mutations, copy numbers and purity for all samples in patient Set_7 (Figure 5f and Supplementary Figure S11), profiling a tumour consistent with a microsatellite stable colorectal cancer (Cross et al. 2018). An equivalent result is also obtained for 6 WGS samples of patient Set_6 (Supplementary Figure S12).

Whole exome data

CNAqc is conceptualised and designed to exploit properties of the VAF distribution in high-resolution whole genomes. Lower-resolution whole-exomes can be analysed if the reduced mutational burden does not compromise VAF quality, peak detection or multiplicity estimation.

We tested CNAqc with WES from $n = 48$ TCGA (Cancer Genome Atlas Research Network 2014) lung adenocarcinomas samples available in the LUAD cohort (Online Methods), selecting the lowest-purity and highest-purity cases to capture different levels of data quality, which we could analyse successfully in most cases (Supplementary Figure S13). Interestingly and in line with the multi-region colorectal cohort (Figure 5), CNAqc could rank calls generated by multiple callers even with WES data. For instance, for sample TCGA-53-7624-01A (Supplementary Figure S14), the TCGA consensus measurement of purity estimations (CPE) obtained by running ESTIMATE (Yoshihara et al. 2013), IHC, LUMP (Aran, Sirota, and Butte 2016) and ABSOLUTE (Carter et al. 2012) is ~80%. CNAqc showed that the CPE consensus is likely wrong, and that the correct purity was estimated only by ABSOLUTE (69%).

Discussion

WGS is a powerful approach to detect extensive mutations that drive human cancers. Many large-scale initiatives such as PCAWG (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), the Hartwig Medical Foundation (Priestley et al. 2019) and Genomics England (Turnbull et al. 2018) have already generated WGS data for thousands of cancer patients, with many cancer institutes converging towards these

efforts. Calling mutations from WGS data requires complex bioinformatics pipelines (Barnell et al. 2019; Cmero et al. 2020; Li et al. 2020) and any downstream analysis relies upon these calls, putting the quality of the generated data under the spotlight.

CNAqc leverages on statistical properties of VAF distributions from WGS, offering the first principle framework to quality control the tumour mutation calls, allele-specific clonal copy number segments and tumour purity. The tool can analyse SNVs and more general types of nucleotide substitutions; SNVs are more reliable and depend less on alignment quality, and should be checked first. CNAqc uses a peak detection analysis to validate segments and purity, exploiting a combinatorial model for somatic alleles applied to the most frequent CNAs found across human cancers. Within the same framework, CNAqc also computes CCF values, highlighting mutations whose multiplicity cannot be phased and are therefore uncertain. This can help to interpret subclonal clusters found by downstream deconvolution tools. We have also shown that CNAqc can process both whole-genome and whole-exome data, across data from different callers. CNAqc features can be used to clean up data, automatising parameter choice for any caller, prioritizing good calls and selecting information for downstream analyses.

The CNAqc framework leverages the relationship between tumour VAF and ploidy. The quality of the control process itself depends on the ability to process the VAF spectrum and detect peaks. Therefore, if the VAF quality is very low because, for example, the sample has low purity or coverage, the overall quality of the check decreases, making it more difficult to completely automate quality checking. However, for the large majority of samples, CNAqc provides a very effective and efficient way to integrate quality metrics in standard pipelines. We note that CNAqc is much faster than quality control by using standard deconvolution tools (Supplementary Figure S15).

Generating high quality calls is a forerunner to more complex analyses that interpret cancer genotypes and their history, with and without therapy (Ding et al. 2012; Landau et al. 2013; Caravagna et al. 2016; Jamal-Hanjani et al. 2017; Turajlic et al. 2018; Caravagna et al. 2018; Roth et al. 2014; Miller et al. 2014; Cross et al. 2018; Gerstung et al. 2020; Deshwar et al. 2015; Strino et al. 2013). CNAqc can pass a sample at an early stage, leaving the possibility of assessing, at a later stage, whether the quality of the data is high enough to approach specific research questions. With the ongoing implementation of large-scale WGS sequencing efforts, and the great amount of WES data already available, CNAqc provides a good solution for modular pipelines that self-tune parameters based on quality scores. To our knowledge, this is the first stand-alone tool which leverages the power of combining the most common types of cancer mutations - SNVs and CNAs - to automatically control the quality of cancer

sequencing assays. We believe CNAqc can help reduce the burden of manual quality checking and parameter tuning. In the future, this tool could be extended to consider other types of CNAs (e.g., extrachromosomal DNA, ecDNA, or subclonal CNAs). ecDNA fragments usually span small genomic regions (Zeng, Wan, and Wu 2020; Verhaak, Bafna, and Mischel 2019) and involve copy states that are not yet supported by CNAqc. Nevertheless, their specific role in amplifying oncogenes and driving tumour evolution and drug resistance (Wu et al. 2019; Kim et al. 2020; Turner et al. 2017) is becoming increasingly important. Adjusting for subclonal CNAs could improve the QC especially at the local level and for those tumors characterized by a strong karyotypic heterogeneity (Ha et al. 2014).

Data Availability

Multiregion colorectal cancer data is deposited in EGA under accession number EGAS00001003066. PCAWG calls are publicly available at (https://dcc.icgc.org/), the ICGC Data Portal. TCGA calls are publicly available at (https://portal.gdc.cancer.gov), the GDC Data Portal.

Software Availability

CNAqc is implemented as an open source R package that is hosted at

https://caravagnalab.github.io/CNAqc/.

The tool webpage contains RMarkdown vignettes to run analyses, visualisation inputs and outputs, and parametrise the tool. All analyses presented in this paper can be replicated following those vignettes; multiregion colorectal cancer data to replicate our analysis is hosted in the GitHub repository.

https://github.com/caravagnalab/CNAqc_datasets.

Authors contribution

All authors conceived the method, which GC formalised and implemented. RB and SM carried out CNAqc simulations and comparisons against other methods. All authors analysed the data and wrote the manuscript.

Competing interests.

The authors declare no competing interests.

Acknowledgments

The research leading to these results has received funding from AIRC under MFAG 2020 - ID. 24913 project – P.I. Caravagna Giulio. Some research was performed using the Cancer Research UK City of London Major Centre High performance computing facility (colcc.ac.uk) and was also funded by a Wellcome Trust grant (ID: 202778/Z/16/Z).

References

Aran, Dvir, Marina Sirota, and Atul J. Butte. 2016. “Corrigendum: Systematic Pan-Cancer Analysis of Tumour Purity.” *Nature Communications* 7 (February): 10707. https://doi.org/10.1038/ncomms10707.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. “Comprehensive Characterization of Cancer Driver Genes and Mutations.” *Cell* 173 (2): 371–85.e18. https://doi.org/10.1016/j.cell.2018.02.060.

Barnell, Erica K., Peter Ronning, Katie M. Campbell, Kilannin Krysiak, Benjamin J. Ainscough, Lana M. Sheta, Shahil P. Pema, et al. 2019. “Standard Operating Procedure for Somatic Variant Refinement of Sequencing Data with Paired Tumor and Normal Samples.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (4): 972–81. https://doi.org/10.1038/s41436-018-0278-z.

Boeva, Valentina, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. 2011. “Control-Free Calling of Copy Number Alterations in Deep-Sequencing Data Using GC-Content Normalization.” *Bioinformatics* 27 (2): 268–69. https://doi.org/10.1093/bioinformatics/btq635.

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Profiling of Lung Adenocarcinoma.” *Nature* 511 (7511): 543–50. https://doi.org/10.1038/nature13385.

Caravagna, Giulio, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A. Graham, Guido Sanguinetti, and Andrea Sottoriva. 2018. “Detecting Repeated Cancer Evolution from Multi-Region Tumor Sequencing Data.” *Nature Methods* 15 (9): 707–14. https://doi.org/10.1038/s41592-018-0108-x.

Caravagna, Giulio, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra. 2016. “Algorithmic Methods to Infer the Evolutionary Trajectories in Cancer Progression.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (28): E4025–34. https://doi.org/10.1073/pnas.1520213113.

Caravagna, Giulio, Timon Heide, Marc J. Williams, Luis Zapata, Daniel Nichol, Ketevan Chkhaidze, William Cross, et al. 2020. “Subclonal Reconstruction of Tumors by Using Machine Learning and Population Genetics.” *Nature Genetics* 52 (9): 898–907. https://doi.org/10.1038/s41588-020-0675-5.

Caravagna, Giulio, Guido Sanguinetti, Trevor A. Graham, and Andrea Sottoriva. 2020. “The MOBSTER R Package for Tumour Subclonal Deconvolution from Bulk DNA

Whole-Genome Sequencing Data." *BMC Bioinformatics* 21 (1): 531. https://doi.org/10.1186/s12859-020-03863-1.

Carter, Scott L., Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, et al. 2012. "Absolute Quantification of Somatic DNA Alterations in Human Cancer." *Nature Biotechnology* 30 (5): 413–21. https://doi.org/10.1038/nbt.2203.

Cmero, Marek, Ke Yuan, Cheng Soon Ong, Jan Schröder, PCAWG Evolution and Heterogeneity Working Group, Niall M. Corcoran, Tony Papenfuss, et al. 2020. "Inferring Structural Variant Cancer Cell Fraction." *Nature Communications* 11 (1): 730. https://doi.org/10.1038/s41467-020-14351-8.

Cortés-Ciriano, Isidro, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. "Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing." *Nature Genetics* 52 (3): 331–41. https://doi.org/10.1038/s41588-019-0576-7.

Cross, William, Michal Kovac, Ville Mustonen, Daniel Temko, Hayley Davis, Ann-Marie Baker, Sujata Biswas, et al. 2018. "The Evolutionary Landscape of Colorectal Tumorigenesis." *Nature Ecology & Evolution* 2 (10): 1661–72. https://doi.org/10.1038/s41559-018-0642-z.

Cun, Yupeng, Tsun-Po Yang, Viktor Achter, Ulrich Lang, and Martin Peifer. 2018. "Copy-Number Analysis and Inference of Subclonal Populations in Cancer Genomes Using Sclust." *Nature Protocols* 13 (6): 1488–1501. https://doi.org/10.1038/nprot.2018.033.

Dentro, Stefan C., David C. Wedge, and Peter Van Loo. 2017. "Principles of Reconstructing the Subclonal Architecture of Cancers." *Cold Spring Harbor Perspectives in Medicine* 7 (8). https://doi.org/10.1101/cshperspect.a026625.

Deshwar, Amit G., Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. 2015. "PhyloWGS: Reconstructing Subclonal Composition and Evolution from Whole-Genome Sequencing of Tumors." *Genome Biology* 16 (February): 35. https://doi.org/10.1186/s13059-015-0602-8.

Ding, Li, Timothy J. Ley, David E. Larson, Christopher A. Miller, Daniel C. Koboldt, John S. Welch, Julie K. Ritchey, et al. 2012. "Clonal Evolution in Relapsed Acute Myeloid Leukaemia Revealed by Whole-Genome Sequencing." *Nature* 481 (7382): 506–10. https://doi.org/10.1038/nature10738.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. 2015. "Sequenza: Allele-Specific Copy Number and Mutation Profiles from Tumor Sequencing Data." *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 26 (1): 64–70. https://doi.org/10.1093/annonc/mdu479.

Fischer, Andrej, Ignacio Vázquez-García, Christopher J. R. Illingworth, and Ville Mustonen. 2014. "High-Definition Reconstruction of Clonal Composition in Cancer." *Cell Reports* 7 (5): 1740–52. https://doi.org/10.1016/j.celrep.2014.04.055.

Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, et al. 2020. "The Evolutionary History of 2,658 Cancers." *Nature* 578 (7793): 122–28. https://doi.org/10.1038/s41586-019-1907-7.

Gillis, Sierra, and Andrew Roth. 2020. "PyClone-VI: Scalable Inference of Clonal Population Structures Using Whole Genome Data." *BMC Bioinformatics* 21 (1): 571. https://doi.org/10.1186/s12859-020-03919-2.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P. Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. "IntOGen-Mutations Identifies Cancer Drivers across Tumor Types." *Nature Methods* 10 (11): 1081–82. https://doi.org/10.1038/nmeth.2642.

Greaves, Mel, and Carlo C. Maley. 2012. "Clonal Evolution in Cancer." *Nature* 481 (7381):

306–13. https://doi.org/10.1038/nature10762.

Ha, Gavin, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, et al. 2014. “TITAN: Inference of Copy Number Architectures in Clonal Cell Populations from Tumor Whole-Genome Sequence Data.” *Genome Research* 24 (11): 1881–93. https://doi.org/10.1101/gr.180281.114.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Jamal-Hanjani, Mariam, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B. K. Watkins, Selvaraju Veeriah, Seema Shafi, et al. 2017. “Tracking the Evolution of Non-Small-Cell Lung Cancer.” *The New England Journal of Medicine* 376 (22): 2109–21. https://doi.org/10.1056/NEJMoa1616288.

Jiang, Yuchao, Yu Qiu, Andy J. Minn, and Nancy R. Zhang. 2016. “Assessing Intratumor Heterogeneity and Tracking Longitudinal and Spatial Clonal Evolutionary History by next-Generation Sequencing.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1522203113.

Kent, David G., and Anthony R. Green. 2017. “Order Matters: The Order of Somatic Mutations Influences Cancer Evolution.” *Cold Spring Harbor Perspectives in Medicine* 7 (4). https://doi.org/10.1101/cshperspect.a027060.

Kim, Hoon, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, et al. 2020. “Extrachromosomal DNA Is Associated with Oncogene Amplification and Poor Outcome across Multiple Cancers.” *Nature Genetics* 52 (9): 891–97. https://doi.org/10.1038/s41588-020-0678-2.

Landau, Dan A., Scott L. Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S. Lawrence, Carrie Sougnez, et al. 2013. “Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia.” *Cell* 152 (4): 714–26. https://doi.org/10.1016/j.cell.2013.01.019.

Levine, Arnold J., Nancy A. Jenkins, and Neal G. Copeland. 2019. “The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters.” *Cancer Cell* 35 (1): 10–15. https://doi.org/10.1016/j.ccell.2018.11.009.

Li, Yilong, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, et al. 2020. “Patterns of Somatic Structural Variation in Human Cancer Genomes.” *Nature* 578 (7793): 112–21. https://doi.org/10.1038/s41586-019-1913-9.

Macintyre, Geoff, Teodora E. Goranova, Dilrini De Silva, Darren Ennis, Anna M. Piskorz, Matthew Eldridge, Daoud Sie, et al. 2018. “Copy Number Signatures and Mutational Processes in Ovarian Carcinoma.” *Nature Genetics* 50 (9): 1262–70. https://doi.org/10.1038/s41588-018-0179-8.

Martincorena, Iñigo, Joanna C. Fowler, Agnieszka Wabik, Andrew R. J. Lawson, Federico Abascal, Michael W. J. Hall, Alex Cagan, et al. 2018. “Somatic Mutant Clones Colonize the Human Esophagus with Age.” *Science* 362 (6417): 911–17. https://doi.org/10.1126/science.aau3879.

Martincorena, Iñigo, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C. Wedge, et al. 2015. “Tumor Evolution. High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin.” *Science* 348 (6237): 880–86. https://doi.org/10.1126/science.aaa6806.

McGranahan, Nicholas, and Charles Swanton. 2015. “Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution.” *Cancer Cell* 27 (1): 15–26. https://doi.org/10.1016/j.ccell.2014.12.001.

———. 2017. “Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future.” *Cell* 168 (4): 613–28. https://doi.org/10.1016/j.cell.2017.01.018.

Miller, Christopher A., Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, et al. 2014. “SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution.” *PLoS Computational Biology* 10 (8): e1003665. https://doi.org/10.1371/journal.pcbi.1003665.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell* 149 (5): 994–1007. https://doi.org/10.1016/j.cell.2012.04.023.

Poell, Jos B., Matias Mendeville, Daoud Sie, Arjen Brink, Ruud H. Brakenhoff, and Bauke Ylstra. 2019. “ACE: Absolute Copy Number Estimation from Low-Coverage Whole-Genome Sequencing Data.” *Bioinformatics* 35 (16): 2847–49. https://doi.org/10.1093/bioinformatics/bty1055.

Priestley, Peter, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Charles Shale, Korneel Duyvesteyn, et al. 2019. “Pan-Cancer Whole-Genome Analyses of Metastatic Solid Tumours.” *Nature* 575 (7781): 210–16. https://doi.org/10.1038/s41586-019-1689-y.

Roth, Andrew, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. 2014. “PyClone: Statistical Inference of Clonal Population Structure in Cancer.” *Nature Methods* 11 (4): 396–98. https://doi.org/10.1038/nmeth.2883.

Strino, Francesco, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. 2013. “TrAp: A Tree Approach for Fingerprinting Subclonal Tumor Composition.” *Nucleic Acids Research* 41 (17): e165. https://doi.org/10.1093/nar/gkt641.

Tarabichi, Maxime, Adriana Salcedo, Amit G. Deshwar, Máire Ni Leathlobhair, Jeff Wintersinger, David C. Wedge, Peter Van Loo, Quaid D. Morris, and Paul C. Boutros. 2021. “A Practical Guide to Cancer Subclonal Reconstruction from DNA Sequencing.” *Nature Methods* 18 (2): 144–55. https://doi.org/10.1038/s41592-020-01013-2.

Turajlic, Samra, Hang Xu, Kevin Litchfield, Andrew Rowan, Stuart Horswell, Tim Chambers, Tim O’Brien, et al. 2018. “Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal.” *Cell* 173 (3): 595–610.e11. https://doi.org/10.1016/j.cell.2018.03.043.

Turnbull, Clare, Richard H. Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, et al. 2018. “The 100 000 Genomes Project: Bringing Whole Genome Sequencing to the NHS.” *BMJ* 361 (April): k1687. https://doi.org/10.1136/bmj.k1687.

Turner, Kristen M., Viraj Deshpande, Doruk Beyter, Tomoyuki Koga, Jessica Rusert, Catherine Lee, Bin Li, et al. 2017. “Extrachromosomal Oncogene Amplification Drives Tumour Evolution and Genetic Heterogeneity.” *Nature* 543 (7643): 122–25. https://doi.org/10.1038/nature21356.

Van Loo, Peter, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, et al. 2010. “Allele-Specific Copy Number Analysis of Tumors.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (39): 16910–15. https://doi.org/10.1073/pnas.1009843107.

Verhaak, Roel G. W., Vineet Bafna, and Paul S. Mischel. 2019. “Extrachromosomal Oncogene Amplification in Tumour Pathogenesis and Evolution.” *Nature Reviews. Cancer* 19 (5): 283–88. https://doi.org/10.1038/s41568-019-0128-6.

Watkins, Thomas B. K., Emilia L. Lim, Marina Petkovic, Sergi Elizalde, Nicolai J. Birkbak,

Gareth A. Wilson, David A. Moore, et al. 2020. “Pervasive Chromosomal Instability and Karyotype Order in Tumour Evolution.” *Nature* 587 (7832): 126–32. https://doi.org/10.1038/s41586-020-2698-6.

Wu, Sihan, Kristen M. Turner, Nam Nguyen, Ramya Raviram, Marcella Erb, Jennifer Santini, Jens Luebeck, et al. 2019. “Circular ecDNA Promotes Accessible Chromatin and High Oncogene Expression.” *Nature* 575 (7784): 699–703. https://doi.org/10.1038/s41586-019-1763-5.

Yoshihara, Kosuke, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, et al. 2013. “Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data.” *Nature Communications* 4: 2612. https://doi.org/10.1038/ncomms3612.

Yuan, Ke, Geoff Macintyre, Wei Liu, PCAWG-11 working group, and Florian Markowetz. 2018. “Ccube: A Fast and Robust Method for Estimating Cancer Cell Fractions.” *biorXiv*. https://doi.org/10.1101/484402.

Zaccaria, Simone, and Benjamin J. Raphael. 2020. “Accurate Quantification of Copy-Number Aberrations and Whole-Genome Duplications in Multi-Sample Tumor Sequencing Data.” *Nature Communications* 11 (1): 4301. https://doi.org/10.1038/s41467-020-17967-y.

Zack, Travis I., Stephen E. Schumacher, Scott L. Carter, Andre D. Cherniack, Gordon Saksena, Barbara Tabak, Michael S. Lawrence, et al. 2013. “Pan-Cancer Patterns of Somatic Copy Number Alteration.” *Nature Genetics* 45 (10): 1134–40. https://doi.org/10.1038/ng.2760.

Zeng, Xixi, Maoping Wan, and Jianmin Wu. 2020. “ecDNA within Tumors: A New Mechanism That Drives Tumor Heterogeneity and Drug Resistance.” *Signal Transduction and Targeted Therapy*. https://doi.org/10.1038/s41392-020-00403-4.

Online methods

CNAqc supports the most frequent allele-specific clonal copy number profiles found in human cancers **(**Supplementary Figure S1):

- heterozygous diploid states (1:1)[1];
- loss of heterozygosity (LOH) in monosomy (1:0) and copy-neutral (2:0) states;

triploid (AAB or 2:1) or tetraploid (AABB or 2:2) states.

Data supports this design choice. In the PCAWG cohort 36% of allele-specific CNA segments are 1:1, 15% are 2:1, 11% are 1:0, 8% are 2:2 and 8% are 2:0. These are >75% of the whole set of calls (>600,000 segments). Moreover, these segments span 93% of all CNA-covered bases in the whole PCAWG cohort. Moreover, in the same cohort, clonal copy numbers are much more frequent, and span significantly larger portions of the tumour genome, than subclonal counterparts obtained by the Battenberg caller. In ~95% of PCAWG samples, >50% of the genome is covered by clonal CNAs.

[1] The notation 1:1 is sometimes analogously expressed as genotype AB, 1:0 as A, 2:1 as AAB and 2:2 as AABB.

Limiting CNAqc to support simple clonal CNAs comes with the advantage that mutation multiplicities are easier to manage, at least from a computational perspective.

CNAqc supports two human reference genome assemblies (GRCh38 and hg19) and makes the simplifying assumption that CNAs are acquired in a single step from an heterozygous germline diploid state. For this reason, for tetraploid segments we only consider the copy state 2:2, instead of 3:1 or 4:0.

CNAqc is conceptualised to work with high-resolution - i.e., high purity and coverage - whole-genome sequencing (WGS) data, but can also be applied to whole-exome (WES) data. The main challenge with WES or low coverage/purity WGS data is the reduced mutational burden and noise in the VAF, which decreases the signal strength. The key determinant to detect VAF peaks is the number of mutations per copy state, with the idea that thousands from a high-quality WGS assay, are certainly better than hundreds from WES or from low-quality WGS. For tumours which are genomically unstable, or exposed to endogenous mutant factors such as smoking or UV-light, or with high mutation rate like microsatellite unstables, the observable mutational burden in exomes might be suitable for CNAqc analysis. A general main disadvantage of low-quality data is that the automation process available via CNAqc might be more difficult, reporting false positives or negatives. In those cases we suggest performing a manual inspection of the proposed scores to optimise the tool, and check consistency against our intuition.

Expected clonal VAF peaks given CNAs and purity

A bulk is a mixture of tumour and normal cells present in proportion $\pi > 0$ and $(1 - \pi)$, respectively. We derive a simple equation describing our belief about the position of the clonal VAF peak in the data, assuming the input clonal segments and purity are correct. This equation is segment-specific, and links all segments with the same allele-specific copy number profile. In this manuscript, we denote as $n_A : n_B$ allele-specific segments with n_A copies of major allele and and n_B of the minor allele. For instance, with 1:0 we denote $n_A = n_B = 1$, with 1:1 we denote $n_A = 1,\ n_B = 0$, etc.

We introduce the multiplicity $m \geq 1$ (or copies) of a clonal mutation mapping on top of the considered segments. As in ASCAT (Van Loo et al. 2010), the expected proportion of reads that can be attributed to a mutation with multiplicity m is $m\pi$. The difference between ASCAT and CNAqc is that the former considers germline single-nucleotide polymorphisms, while the latter considers somatic mutations (i.e., germline is removed); besides this difference, the conceptualisation is similar. For segments $n_A : n_B$, the

proportion of all reads from the tumour is $\pi(n_A + n_B)$. Here we term $n_A + n_B$ the ploidy of the $n_A : n_B$ segments, remarking that is not the overall tumour ploidy. Similarly, assuming a healthy diploid normal, the proportion of reads that come from normal cells is $2(1 - \pi)$. When we consider clonal mutations with multiplicity m sitting on $n_A : n_B$ segments, we would expect them to peak in the VAF distribution at value

$$v_m = \frac{m\pi}{2(1-\pi) + \pi(n_A + n_B)} \quad (3)$$

This formula describes our belief about the position of the clonal VAF peak in the data, assuming the input segments (determined by segments with n_A and n_B alleles) and purity π is correct.

The formula is intuitive and gives the expected results. Consider $1 = n_A = n_B$ and $\pi = 1$, i.e., heterozygous diploid segments in a pure tumour, since clonal mutations have $m = 1$ the clonal VAF should be ~50%, and in fact $v_m = 0.5$. Instead, for tetraploid segments obtained after whole-genome duplication, where we have $2 = n_A = n_B$ and $\pi = 1$, under the simplifying assumption of CNAqc, clonal mutations could be present in single ($m = 1$) or double ($m = 2$) copy. Evaluating equation (1) we obtain $v_1 = 0.25$ for mutations in a single copy (25%, post-aneuploidy), and $v_{m=2} = 0.5$ (50% VAF, pre-aneuploidy).

Transforming VAFs to CCFs

There are several methods to compute CCFs from VAFs, allele-specific copy number and purity. The equation used by CNAqc is inspired by seminal works (Van Loo et al. 2010; Dentro, Wedge, and Van Loo 2017), and converts the observed VAF $v > 0$ of a mutation with multiplicity m into the CCF c

$$c = \frac{v[(n_A + n_B - 2)\pi + 2]}{m\pi}. \quad (4)$$

Note that all the parameters π, n_A, n_B are as in eq. (1). Given input VAFs, CNAs and purity, the only quantity to be estimated to compute c is m, the multiplicity.

CCFs are proportional to VAFs, as we expect. Consider a heterozygous clonal diploid mutation ($1 = n_A = n_B = m = \pi$ so $p = 2$). Its expected VAF is 50% and $c = 1$ reporting a correct 100% of cells with the mutation, which is clonal. The same formula works for subclonal mutations. As another example, if a single-copy clonal mutation ($m = 1$) sits in an amplified triploid state ($2 = n_A$ and $1 = n_B = \pi$) and has a VAF of 33% - 1 out of 3 copies - we have $c = 1$. The other type of clonal mutation that we can observe in those types of segments has VAF of 66%, with 2 out of 3 copies and then its CCF is again $c = 1$ using equation (2).

Peak detection quality control for allele-specific CNAs and purity

Data peaks can be used to quality control (QC) both tumour purity and CNA segments, following the intuition of equation (1). The procedure is summarised in pseudocode in Supplementary Figure S2, and works by partitioning mutations by the copy states of the segments (after mutation mapping on CNAs), and analysing them independently to determine a PASS or FAIL status per mutation multiplicity. A sample-level PASS or FAIL score is then computed by aggregating all statuses in a majority-based system, where each copy state weights proportionally to the number of mutations it harbours (i.e., the evidence from the data).

In CNAqc there are therefore three levels of PASS or FAIL status: i) for each VAF peak in a given copy state, ii) overall for a copy state and iii) for the whole sample. This allows subsetting of calls according to a fine-grained set of metrics, for instance passing only some variants even if the overall sample fails.

The peak detection strategy take as input $\epsilon > 0$, the *upper bound on the error that we can tolerate on purity*. For example, if $\epsilon = 0.05$, we can accept a 5% error on the purity; if the true purity was 60% and the caller reported a value in range [55%, 65%], CNAqc would pass the sample. The range associated to ϵ is adjusted to account for ploidy and mutation multiplicity, providing a conversion between errors measures in VAF and purity units. The formula that we introduce is presented below.

Peak matching strategies. For every copy state, CNAqc matches either 1 or 2 peaks, depending on the ploidy of the involved segments and multiplicity: one peak is matched for 1:0 and 1:1, two for all others. Here we discuss the strategy to detect a generic peak, assuming to work with copy state $n_A : n_B$ as in equation (1).

The tool implements methods (described below) to detect n peaks $d_1,..., d_n$ in the VAF distribution, and match them against v_m, the expected clonal VAF for a peak with multiplicity m. The matching of v_m, determines the PASS or FAIL status for the associated multiplicity. To compute the match we select one d_* among $d_1,..., d_n$ that is close enough to v_m, where the choice of d_* can be done in two ways:

- *by closest hit match,* where d_* is the data peak that is closest to v_m, i.e. $d_* = \arg_i \min |d_i - v_m|$, where i ranges $1,..., n$..
- *by rightmost hit match*: where d_* is taken from the subset of peaks $D = \{d_i > v_m | \, i = 1,.., n\}$ on the right of the expected peak, and d_* is the largest possible value $d_* = \arg_{i \in D} \max |d_i - v_m|$ (most right apart).

The CNAqc default strategy is the first, which selects d_* as the peak closest to v_m. The second strategy is more stringent, but can help identify miscalled segments. Consider for instance a diploid 1:1 copy state, if in the pool of putative diploid mutations some should have been associated to LOH segments (miscalled), an extra VAF peak is expected on the right of the clonal cluster. The rightmost hit match strategy will associate the theoretical peak v_m to the LOH one, flagging the diploid segment as FAIL because the LOH peak will be far off the clonal peak.

When we match the peaks, the desired purity error ε gets rescaled depending on the copy state $n_A : n_B$, following this general equation

$$\epsilon_m = v_m(\pi + \epsilon) - v_m(\pi) \sim \frac{\partial v_m}{\partial \pi}\epsilon = \frac{2m\epsilon}{[2(1-\pi) + \pi(n_A + n_B)]^2} \quad (5)$$

This means that, in order to match a purity error ε, we create a range of acceptance based on ϵ_m, as we discuss below. In this equation we interpret the VAF as a function of tumour purity and, assuming ϵ to be small, we truncate the Taylor expansion of $\nu_m(\pi+\epsilon)$ at the first order.

Notice that the error on the VAF depends in general on purity and multiplicity. Consider, for instance, a segment with copy state 2:1 for a tumour with purity 90%, $\varepsilon = 0.05$ (5%) corresponds to an error in the VAF of approximately 1% and 2% for the VAF peaks with

multiplicity $m = \{1, 2\}$ respectively. Inverting equation (3), one can express the purity as a function of the VAF, ploidy and multiplicity and derive the error propagation formula from the VAF space to the purity space used in CNAqc. Using the same approach of equation (5), we treat the purity as a function the VAF, shift the VAF by a small error ϵ_m and truncate the Taylor expansion at the first order to get

$$\pi(v_m) = \frac{2v_m}{m + [2 - (n_A + n_B)]v_m}, \quad (6)$$

$$\pi(v_m + \epsilon_m) - \pi(v_m) \sim \frac{\partial \pi}{\partial v_m}\epsilon_m = \frac{2m\epsilon_m}{[m + v_m(2 - (n_A + n_B))]^2}. \quad (7)$$

Peaks are matched by including a *VAF tolerance* $\epsilon_{VAF} > 0$, which helps ameliorate the fact that we do not explicitly model noise affecting peak detection. The intervals

$$I_m^{VAF} = [d_* - \epsilon_{VAF};\ d_* + \epsilon_{VAF}] \quad \text{and} \quad I_m = [v_m - \epsilon_m, v_m + \epsilon_m] \quad (8)$$

are created with centre at d_*, the matched peak in the VAF, with size $2\epsilon_{VAF}$, and tested for overlap with the interval I_m. If I_m^{VAF} overlaps with I_m, i.e., $|I_m^{VAF} \cap I_m| > 0$, the clonal peak for multiplicity m is matched by d_*, and therefore receives a PASS status. Otherwise it receives a FAIL status.

The PASS or FAIL status per copy state with two possible multiplicity values is defined by taking the status of the peak associated with the largest number of mutations n_m. The value of n_m is determined by binning the VAF distribution with 100 bins (from 0 to 1, with size 0.01), and counting the number of mutations that associate to the bin of the matched peak. In this way, we PASS a copy state if the tallest of its peaks is a PASS, and is associated with more mutations than any FAIL peak.

The sample-level QC status is based on an error metric that uses the actual distance between the centres of the intervals, d_* and v_m, which is given by $d_* - v_m$.

CNAqc sample-level error metric. An error metric is assembled across copy states to determine a sample-level PASS or FAIL status. Consider w_k as the normalised number of mutations mapped to copy state k, which we further rescale by 2 if the copy state is supposed to have two peaks. For every copy state and every mutation multiplicity, we have a PASS or FAIL from peak detection.

We split PASS (P_k) from FAIL (F_k) peaks, and define two scores per copy state by linear combination

$$(7)\ \lambda_{k,}^{PASS} = \sum_{d_*^m \in P_k} w_k \left(d_*^m - v_m^k\right)$$

$$(8)\ \lambda_{k,}^{FAIL} = \sum_{d_*^m \in F_k} w_k \left(d_*^m - v_m^k\right)$$

where the subscript k denotes the copy state (i.e., 1:0), and d_*^k denotes the peak matched for multiplicity m in copy state k. We define the overall sample score λ

$$(9)\ \lambda = \sum_{k \in K} \left(\lambda_k^{PASS} + \lambda_k^{FAIL}\right)$$

where K is the set of copy states 1:0, 1:1, 2:0, 2:1 and 2:2 supported by CNAqc. Equation (9) is the linear combination whose terms can be either positive or negative, depending on whether the matched peaks are on the right or left of the expected peaks. The sample score λ is a weighted mean since by construction all the w_k sum to one.

The overall status on the sample is taken by comparing $\lambda_{k,}^{PASS}$ and $\lambda_{k,}^{FAIL}$ and selecting the status corresponding to the largest of the two.

Computing peaks from VAF. CNAqc implements a joint strategy to detect n peaks $d_1,\ldots, d_n$ in the VAF distribution:

1. *Kernel-based*: via kernel density estimation with default adjustment 1 and fixed bandwidth, a smoothed VAF profile is obtained. Peaks are then estimated from the discretized smooth, using specialised R packages for peak detection and removing peaks with density below 1/20 (empirical cut) of the maximum peak.

2. *Mixture-based:* via Binomial mixture from the BMix (Caravagna, Heide, et al. 2020) package (https://caravagn.github.io/BMix/), a peak is associated with each Binomial probability, for all mixture components.

The latter strategy is inspired by subclonal deconvolution methods, and computes the model density for w clusters (default $w < 5$), with model-selection to optimise w using the Integrated Classification Likelihood score (Caravagna, Heide, et al. 2020); the likelihood is

$$f(X \mid \boldsymbol{\pi}, \boldsymbol{p}) = \prod_{x \in X} \sum_{i=1}^{w} \pi_i \text{Bin}(r_x \mid n_x, p_i) \quad (10)$$

where π_i are the mixing proportions of the mixture, not to be confused with sample purity. Here we use a Binomial likelihood for r_x successes determined as the number of reads with the mutant covering mutation x, n_x is the total trials given by the sequencing depth at the locus, and p_i the Binomial success probability Assuming that calls have passed the quality metrics for CNAqc, then p_i is defined as the expected theoretical VAF from equation (1), so it is known. A key advantage of BMix over other deconvolution tools is the fast maximum likelihood implementation, with full access to the model parameters (e.g., latent variables).

CCF estimation

A lot of tools for downstream subclonal deconvolution compute CCFs to normalise mutations, CNAs and purity, and cluster mutations. Some popular tools - e.g, PyClone (Roth et al. 2014) - focus on cluster-level rather than per-mutation CCFs. For this reason, not all deconvolution tools offer the same information accessible from CNAqc, with Bayesian deconvolution algorithms in PyClone or DPclust being computationally much more demanding than CNAqc (Nik-Zainal et al. 2012; Roth et al. 2014).

CNAqc offers a way to estimate CCFs and a PASS or FAIL status which can be used to assess the quality of the estimates.

CCF computation. CNAqc offers two distinct approaches to compute CCFs:

- *Entropy-based computation:* in which a Binomial mixture like in equation (10) is peaked at the VAFs v_m values from equation (1), and input mutations are phased to their multiplicity *only if* the mixture's latent variables are well separated.
- *Rough computation:* in which a Binomial mixture is used and mutations are phased regardless of the latent variables of the mixture

The entropy-based model can fail to compute the multiplicity of a mutation, and return CCF values with NA associated; this is how uncertainty is reported in CNAqc. The latter method, by design, will always assign a multiplicity $m \in \{1, 2\}$.

The final PASS or FAIL status of a copy state is determined from the proportion of mutations with available CCF. Therefore, while the rough computation will always PASS a copy state, this is not the case for the entropy-based method. By default, if more than 10% of the mutations per copy state have no available CCF, a FAIL is raised; the percentage parameter can be set to arbitrary values.

We first detail the rough approach. We describe the case of copy states 2:0, 2:1 and 2:2, the others being trivial. To initialise a mixture analogous to equation (10):

1. we build two Binomial densities from the theoretical expectations of the VAF peaks, i.e., v_1 and v_2, depending on the copy state, as defined in equation (1). This will create, for instance, one Binomial with parameter $p = 0.33$ and one with $p = 0.66$ for a pure ($\pi = 1$) tumour and 2:1 copy state.
2. We fix - in equation (10) - the number of Binomial trials to the median coverage of the considered mutations, and compute the 1% and 99% quantiles of the data distributions to obtain a VAF range around each peak.
3. Finally, we count mutations that, according to VAF, map to either one or the other computed range. The number of mutations n_1 and n_2, associated to multiplicity $m = 1$ and $m = 2$, is then used to obtain the normalised mixing proportions π_1 and π_2 to complete the model in equation (10).

Densities are computed at steps 1 and 2, while mixing proportions are computed at step 3; with these parameters we can compute the mixture likelihood. Akin to mixtures, we introduce the notion of latent variables z as a matrix of mutations by clusters, for which we define, the probability of assigning read counts data for mutation n to component $i \in \{1, 2\}$. With these latents, every row of matrix z is a categorical random variable

reporting the probability of assigning $m = 1$ or $m = 2$ to a mutation, for which we can define the entropy in the standard way.

(10) $H(z_n) = - z_{n,1} log(z_{n,1}) - z_{n,2} log(z_{n,2})$.

The entropy is maximal if $z_{n,1} = z_{n,2}$, and the mutations are equally likely in single and double copies. It is minimal if $z_{n,1} = 1$ and $z_{n,2} = 0$, or vice versa. If the entropy is low, the mutation is often difficult to phase to single or double copy mutations. The shape of the entropy resembles - by construction - a growing curve with a central spike, which we use to create a simple criterion to discriminate high from low entropy. The geometric intuition of this criterion is extremely simple: *at the crossing of Binomial densities peaked at m_1 and at m_2 if the entropy is high we cannot confidently phase mutations to multiplicities*. The amount of Binomial overlap depends on coverage and purity - this is the technical reason CCF is more uncertain for low resolution data.

CNAqc inspects the entropy profile to determine peaks $\{h_1, h_2\}$ around the spike, using the same peak detection tool used for quality control. Every mutation in the range

(11) $I_{NA} = [h_1, h_2]$

cannot be unequivocally assigned multiplicity values, and are therefore undetermined using the entropy-based method.

The rough approach determines the midpoint $o = v_1 + (v_2 - v_1)\pi_1$ between the two expected theoretical VAF peaks v_1 and v_2, given the mixing proportion π_1 of the first mixture component. The midpoint is computed by weighting each of the two peaks proportionally to the number of mutations that appear underneath each peak, which we compute like with the entropy method. The midpoint is a cut: $x < o$ are phased to a single copy, values above to two copies. This procedure requires data with good general quality because it assumes that all mutations can be phased correctly by a hard VAF split, a fact that depends largely on coverage and purity.

When multiplicities have been determined, CCFs are computed with equation (2).

Genome fragmentation

Some recently identified patterns of somatic CNAs can be attributed to the presence of highly fragmented tumour genomes, termed chromothripsis and chromoplexy, or localised hypermutation patterns, termed kataegis (Cortés-Ciriano et al. 2020).

While these can be identified using dedicated tools, CNAqc offers a simple statistical test to detect the presence of potential over-fragmentation in a chromosome arm, a prerequisite that could point to the presence of such patterns. CNAqc analysis does not substitute dedicated tools, but provides preliminary information to determine what parts of the genome might be run with ad hoc methods.

The test works at the level of each chromosome arm (1p, 1q, 2p, 2q, etc.), and uses the length of each input CNA segment to assign a “long segment” or “short segment” status. This is determined by a cut parameter μ that is set, by default, to 20% (i.e., $\mu = 0.2$). Recent evidence from large pan-cancer studies can be used to calibrate this parameter to cancer-specific values (Zack et al. 2013).

Then, a null hypothesis is used to compute a p-value using a Binomial test based on k, the number of trials given by the total segments in the arm, and the observed number of short segments s. The Binomial distribution for H_0 is defined by μ, and the null is the probability of observing at least s short segments, and therefore we defined a one-tailed test for whether the observations are biased towards short segments. The p-value is adjusted for family-wise error rate by Bonferroni, dividing the desired α-value by the number of tests.

This test is applied to a subset of chromosome arms with a minimum number of segments, and that “jump” in ploidy by a minimum amount (empirical default values estimated from trial data). The arm-level jump is determined as the sum of the difference between the ploidy of two consecutive DNA segments. These covariates are similar to those used to infer CNA signatures from single-cell low-pass WGS (Macintyre et al. 2018).

Other features

CNAqc contains multiple functions to subset the data (i.e., select mutations that map only to certain copy states, subset CNAs with a total ploidy, etc.), visualise the data (i.e., plot mutational burden by tumour genome) or smooth the input CNA segments.

Smoothing is an operation that can be carried out before testing for over-fragmentation. In CNAqc, by smoothing we merge two contiguous segments if they have exactly the

same allele-specific profile (i.e. same numbers for the major and minor alleles), and if they are a maximum distance apart (e.g. 1 megabase by default). This operation does not affect the ploidy profile of the calls, but reduces the amount of breakpoints that would inflate the p-value of the Binomial over-fragmentation test.

Peak detection simulations

We tested CNAqc on a synthetic dataset of ~20.000 tumours, generated to mimic data that we observed in real patient tumours.

We first simulated synthetic VAFs from clonal CNA segments generated with breakpoints distributions following Poissons (6 segments per chromosome, on average. We used a Dirichlet copy state concentration 1 for 1:0, 1 for 2:0, 6 for 1:1, 2 for 2:1 and 1 for 2:2). Then we simulated Poisson-distributed coverage with median depth 30x, 60x, 90x and 120x, and set purity to 0.4, 0.6, 0.8 or 0.95. The idea of this test was to simulate a tumour with purity π and run CNAqc with an input purity that contained a positive or negative error ε_{err}, i.e., we imputed CNAqc purity $\pi + \varepsilon_{err}$. Then, for different values of the input tolerance ϵ, i.e., the maximum purity error we want to tolerate in CNAqc, we run the tool with default peak-matching parameters and perform quality control. Ideally, when the input error ε_{err} is lower than tolerance ϵ, $\varepsilon_{err} < \epsilon$, CNAqc should pass the sample.

We performed the quality check applying an error on the purity varying in range [0; 0.2] with intervals of length 0.02, setting a tolerance on the purity error ranging in [0.01; 0.05] with intervals of length 0.004. We tested CNAqc on 100 simulated tumours for any combination of all the considered parameters. We consistently observed that, as the purity error ε_{err} exceeds tolerance ϵ, the proportion of failures approaches 100% (Supplementary Figure S4). For instance, setting a tolerance parameter of 2%, we can accept a purity error of 5% at most. Over this threshold the proportion of FAIL samples increases reaching maximum at ~7%. One can check this behaviour for the samples of purity 0.95 and coverage 90x: for a tolerance of ~2%, the proportion of rejected samples is close to 0% when the purity error is smaller than 5%, it increases to 70-75% for a purity error of ~5/6%, while for a purity error of ~10% the fail proportion is 100%. From the test we also observed that the ability of CNAqc to detect samples with incorrect purity improves consistently as we increase coverage, with this effect more evident for samples with high purity.

For the same tumours we also computed CCFs and the proportion of mutations for which CNAqc could not phase multiplicity (only for copy states 2:0, 2:1, 2:2 since 1:0 and 1:1 have single multiplicity). We plot the percentage of unassignable mutations as a function of purity in Supplementary Figure S5. We can see that the proportion decreases as we increase coverage and purity, meaning that the computation of reliable CCFs can depend largely on data quality. The observed trend was expected, since at low coverage and purity we have the overlaps between clonal clusters which makes it harder to phase multiplicity from VAFs.

Comparison to deconvolution methods

Some of the functioning of CNAqc is inspired by the design of subclonal deconvolution methods (Roth et al. 2014; Nik-Zainal et al. 2012; Dentro, Wedge, and Van Loo 2017; Jamal-Hanjani et al. 2017; Gerstung et al. 2020; Jiang et al. 2016; Caravagna, Heide, et al. 2020; Caravagna, Sanguinetti, et al. 2020). Therefore, we sought to compare CCFs by CNAqc with the one obtained by Ccube (default parameters), a CCF-computation method developed by the PCAWG Evolution and Heterogeneity Working Group (Yuan et al. 2018).

In Supplementary Figure S9 (panel a) we show the correlation among the CCF values computed by Ccube and CNAqc (entropy method) in PCAWG. In the plot we annotate the proportion of cases, split by copy state and mutation multiplicity, where the estimates are different after rounding to the second digit. We observe that the tools report the same CCF for ~99% of the analysed mutations, whenever CNAqc identifies a reliable CCF value. We remark that a feature of CNAqc is reporting the percentage of mutations where the CCF cannot be unequivocally determined. In the above statistics, the CCF values are therefore computed only for mutations where the uncertainty is not present in CNAqc. The information regarding uncertainty is however very helpful to integrate CNAqc with other tools for CCF computations, as we show with two examples from our test.

In Supplementary Figure S9 (panels b-g) we report an example PCAWG case where the CCFs are in perfect agreement (1 out of 307 mutations in 2:2 segments with different CCF). In Supplementary Figure S10, instead, we show a case where CNAqc detects uncertainty in 14% of input triploid mutations, informing of potential challenges in using CCFs for those mutations. In that case the uncertainty is explained by the intermixing between two clonal picks in triploid 2:1 segments. Ccube assigns multiplicity 2 to a group of clonal SNVs at the right tail of the lowest clonal pick. The consequent CCF distribution breaks the expected clonal peak around ~1, alluding to the presence of two close CCF clusters. This is due to Ccube assigning some single-copy mutations

$m = 2$, and vice versa. The entropy-based method by CNAqc highlights 14% of 2:1 mutations as uncertain, including the ones mistaken by Ccube. In turn, CNAqc assigns a FAIL status to these mutations with default values (cutoff >10%). Notably, the CCF distribution returned by CNAqc, which uses 86% of total mutations once the 14% non-assignable are removed, is correctly peaked at ~1.

Errors in CCFs can affect downstream subclonal deconvolution, which in turn inflates evolutionary statistics (e.g., number of subclonal clusters, clonal complexity). In this example, miscalled multiplicities generate a spurious cluster in the CCF distribution fit by Ccube, which leads to subclonal cluster 2 (panel g, Supplementary Figure S10). Even after removal of 14% CCFs flagged as uncertain by CNAqc, Ccube still assigns the wrong mutation multiplicity to a significant number of variants and infers the spurious CCF cluster (panel h, Supplementary Figure S10). For this reason, reporting a FAIL status in CNAqc informs that multiplicity computation in this sample is highly confounded by intermixing of VAFs, cautioning the interpretation of downstream deconvolution analyses.

Whole-exome sequencing data

There is an obvious difference between the richness of information that is available in a whole-genome assay, compared to a whole-exome one. Similarly, there is a difference between samples with high purity and coverage for current standards (e.g., WGS >60x with 70% purity), and those with lower parameters.

We collected whole-exome data from $n = 48$ lung adenocarcinoma samples available in TCGA LUAD (Cancer Genome Atlas Research Network 2014), selecting the 24 ones with top and bottom purity values, as of the consensus purity estimated by TCGA (CPE score). We report example cases in Supplementary Figure 13, where PASS and FAIL values are obtained by using somatic SNVs, CPE purity estimates and default CNAqc parameters.

The case in panel (a), sample TCGA-53-7624-01A, is 84% pure and the inferred ploidy is correct, but purity is slightly overestimated. The case in panel (b) is 82% pure, but with a similar error pattern. The case in panel (c) is PASS with 30% purity; in this case it is difficult to assess if the small peak matched by CNAqc is a noise artifact. This is an example of a VAF distribution that is low resolution. The case in panel (d) is 83% pure tumour, with good calls. The case in panel (e) is 32% pure and passed because most of the tetraploid mutations seem legit, but it contains a poorly-peaked VAF distribution in

triploid states (2:1, 47% of the overall mutational burden). In this case CNAqc struggles to detect peaks from VAF; this is another example of low resolution VAF distribution.

CNAqc can also be used to select among multiple purity estimates provided by different CNA callers, even with WES data. We focus on case (a) from Supplementary Figure S13. In TCGA, we obtain purity estimates from CPE, which is the consensus among ABSOLUTE, ESTIMATE, IHC and LUMP. We used CNAqc to assess the quality of the estimates for the LUAD sample TCGA-53-7624-01A. For this sample, ESTIMATE, IHC and LUMP agree and determine the value for CPE. We found that only ABSOLUTE detected the true tumour ploidy (69%, Supplementary Figure S14), according to CNAqc. This shows that CNAqc can be used to select among multiple purity estimates the value that best integrates mutations and copy number data, even from WES assays, avoiding in principle the need of consensus calling.

From these tests we conclude that CNAqc can also be used on WES data like the data available in TCGA, possibly coupled with manual revision of critical cases.

Wall-time performance

The analysis of PCAWG showed that CNAqc is fast; in order to generalise that assessment and understand how performance scales with sample size, we compared the wall-clock time of CNAqc against common deconvolution tools.

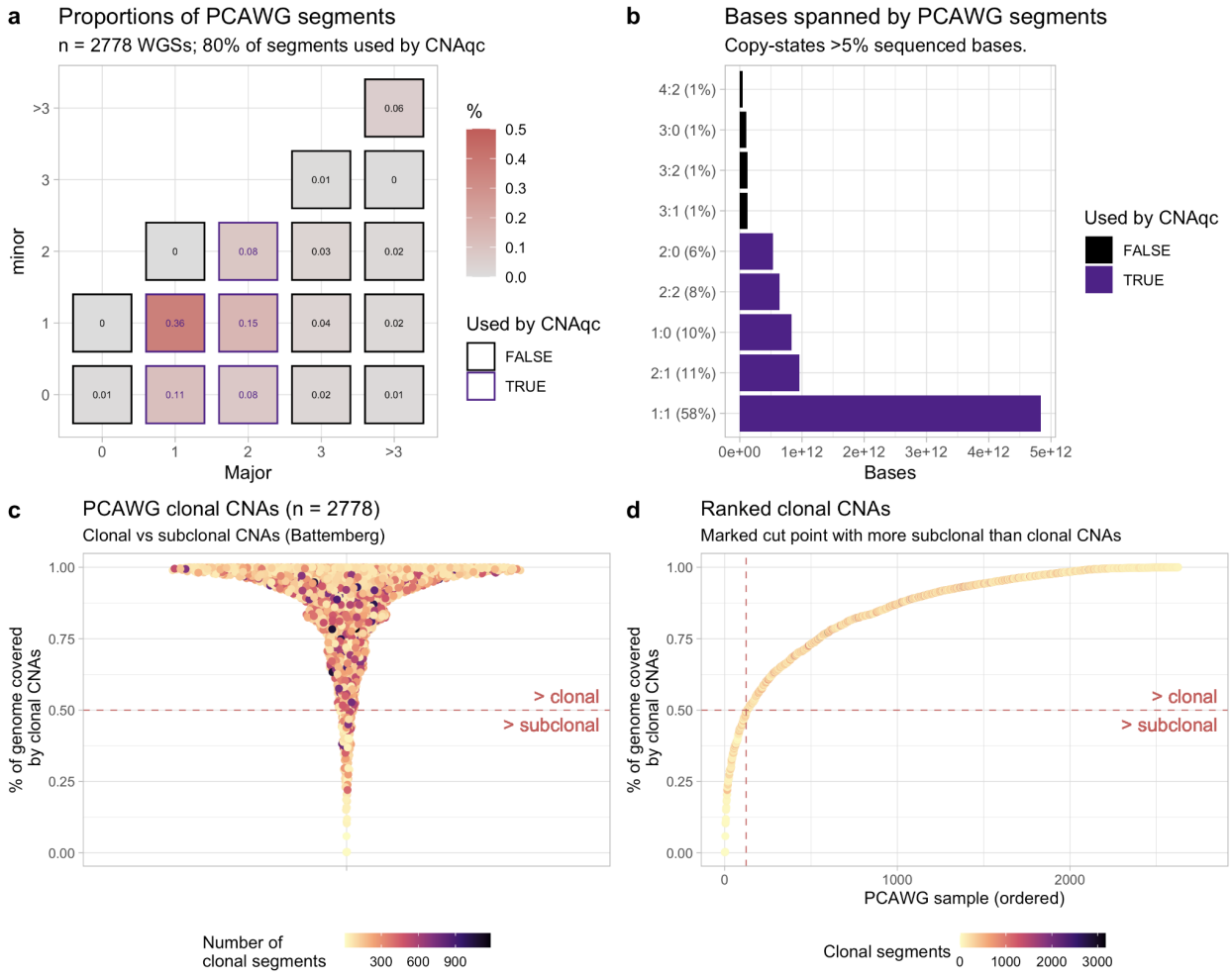
We chose Sciclone (Miller et al. 2014), Ccube (Yuan et al. 2018) and Pyclone-vi (Gillis and Roth 2020) to represent a diverse set of popular algorithms for deconvolution. To build the dataset we subsetted all the mutations in diploid regions from a melanoma sample of the PCAWG cohort (DO220877) leading to a total of 207508 mutations. This is the PCAWG sample with highest mutational burden in the cohort. Then, from those 207508 SNVs we sampled N={500,1000,25000,5000,1000,25000,50000} mutations; this process was repeated 10 times to have 10 replicates for each N. The CNaqc analysis for peak detection was run with default parameters. Similarly, default parameters were also used for Sciclone (default one-dimensional deconvolution) and Ccube (but with *numOfRepeat=1*); Pyclone-vi was run with beta binomial likelihood, number of clusters from 1-10 and 30 repetitions (Supplementary Figure S15).

CNAqc turned out to be the fastest tool, capable of processing up to 500,000 mutations in under one minute. Immediately after, tools based on variational inference were about an order of magnitude slower. The latter two algorithms range from being 4 to 16 times slower than CNAqc for our range of tests (consider the log-scale in the plot y-axis), and the performance gap increased with larger N. Notably, Sciclone took an average of two

hours to process 50,000 mutations, which is 128 times slower than CNAqc as suggested by a log-difference of 5. In all tests, CNAqc, Ccube and Pyclone-vi scaled approximately exponentially, while Sciclone showed a jump from 25,000 to 50,000 mutations. All simulations were performed on a machine with 36 Intel(R) Xeon(R) Gold 6140 CPUs @ 2.30GHz and 220 GB of RAM (Ubuntu 20.04 LTS, Python 3.8.2 and R 4.1.0).

Main Figures

Supplementary Figures



Supplementary Figure S1. a. Proportion of PCAWG CNA segments split by copy state, obtained by consensus calling across multiple callers with $n = 2778$ WGS samples of primary tumours. The matrix reports major and minor alleles, the colour and number reflects the proportion of CNA segments with that copy state across total (e.g., 36% of segments are diploid heterozygous, 1:1). CNA segments used by CNAqc are coloured in purple; in total, 78% of the overall set of segments (>600,000) can be processed by our method (36% of segments are 1:1, 15% are 2:1, 11% are 1:0, 8% are 2:2 and 8% are 2:0). **b.** Number of bases covered, and proportions relative to the total genome spanned by all the PCAWG segments in panel (a). Diploid heterozygous segments cover over a thousand billion bases($> 10^{12}$), accounting for 58% of the genome covered by these segments. The segments supported by CNAqc are

the top-5 most common segments reported across all PCAWG, covering 93% of all bases sequenced in this cohort. **c.** Battemberg clonal and subclonal CNAs available in PCAWG. To simplify the visualisation we remove outliers exceeding the 99-th quantile of the data distribution. Every dot is the percentage of the tumour genome spanned by clonal segments, coloured by the number of segments per sample. So if a sample has >50% of clonal segments it is above the horizontal dashed line. **d.** We rank by sorting the percentages shown in panel (c) to note that only $n = 124$ PCAWG samples (vertical dashed red line) have more subclonal than clonal CNAs.

Algorithm: ***Peak detection in CNAqc***

Input: Mutations, allele-specific CNA segments, purity,
Parameters: purity error tolerance $\varepsilon > 0$ and VAF tolerance $\varepsilon_{VAF} > 0$

- set $K = \{1:0,\ 1:1,\ 2:0,\ 2:1,\ 2:2\}$, where $n_A : n_B$ are the copies of the Major/ minor alleles;

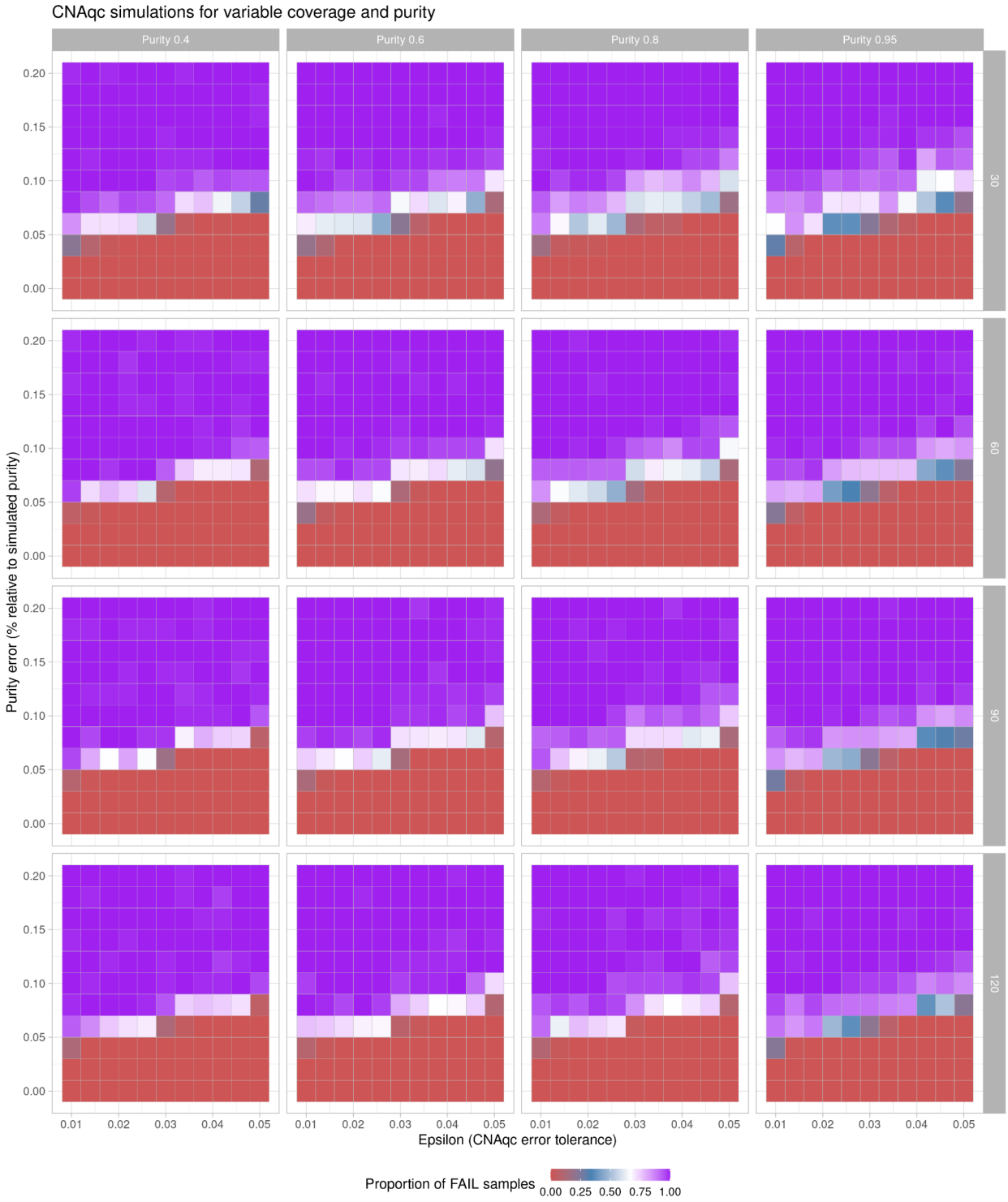
 # Peak detection for every copy state

- for every copy state $k \in K$ and multiplicity $m \in \{1, 2\}$:
  - retain mutations M_k mapping to segments with copy state k;
  - compute v_m with equation (1);
  - compute ε_m with equation (3);
  - determine peaks $d_1, ..., d_n$ from the VAF distribution of M_k;
  - match d_*^m to v_m by either closest or rightmost hit;
  - define interval I_m from v_m and ε_m with equation (6);
  - define I_m^{VAF} from d_*^m and ε_{VAF} with equation (6);
  - define PASS for k and m if $|I_m^{VAF} \cap I_m| > 0$, FAIL otherwise;
  - determine the number n_m of mutations mapping below VAF peak d_*^m;
  - if $k \in \{2:0,\ 2:1,\ 2:2\}$, compare n_m across peaks and define the status for k from the largest n_m, otherwise use the only available peak;

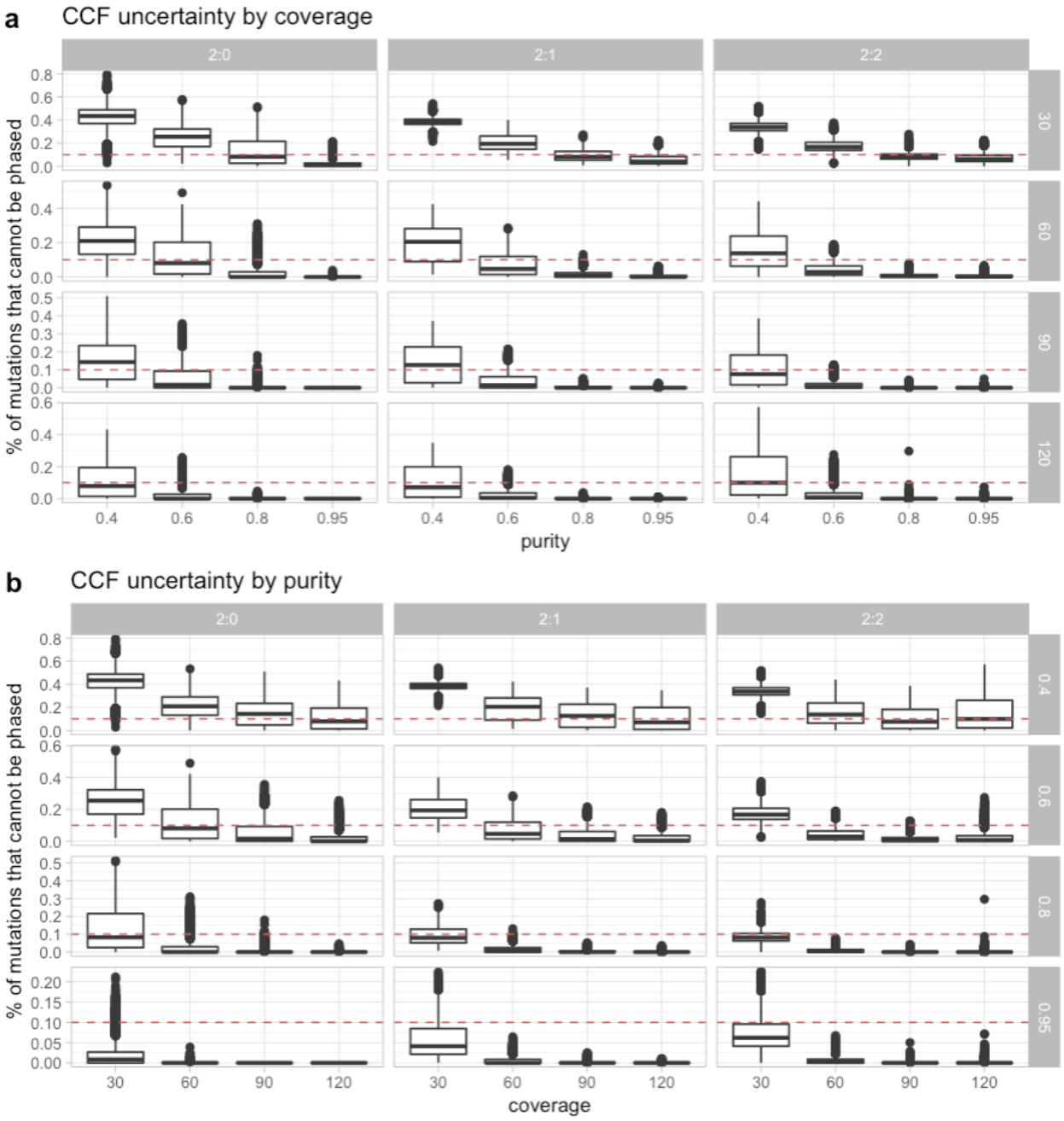
 # Sample level quality control status

- for all copy states $k \in K$, define w_k by normalising the number of mutations mapped to the copy state, and rescale w_k by 2 if $k \in \{2:0,\ 2:1,\ 2:2\}$;
- for every copy state $k \in K$ define $\lambda_{k,}^{PASS}$ and $\lambda_{k,}^{FAIL}$ with equations (7) and (8);
- define the sample score λ with equation (9), and evaluate the sample status from the sum of $\lambda_{k,}^{PASS}$ and $\lambda_{k,}^{FAIL}$ for all $k \in K$, taking the largest.

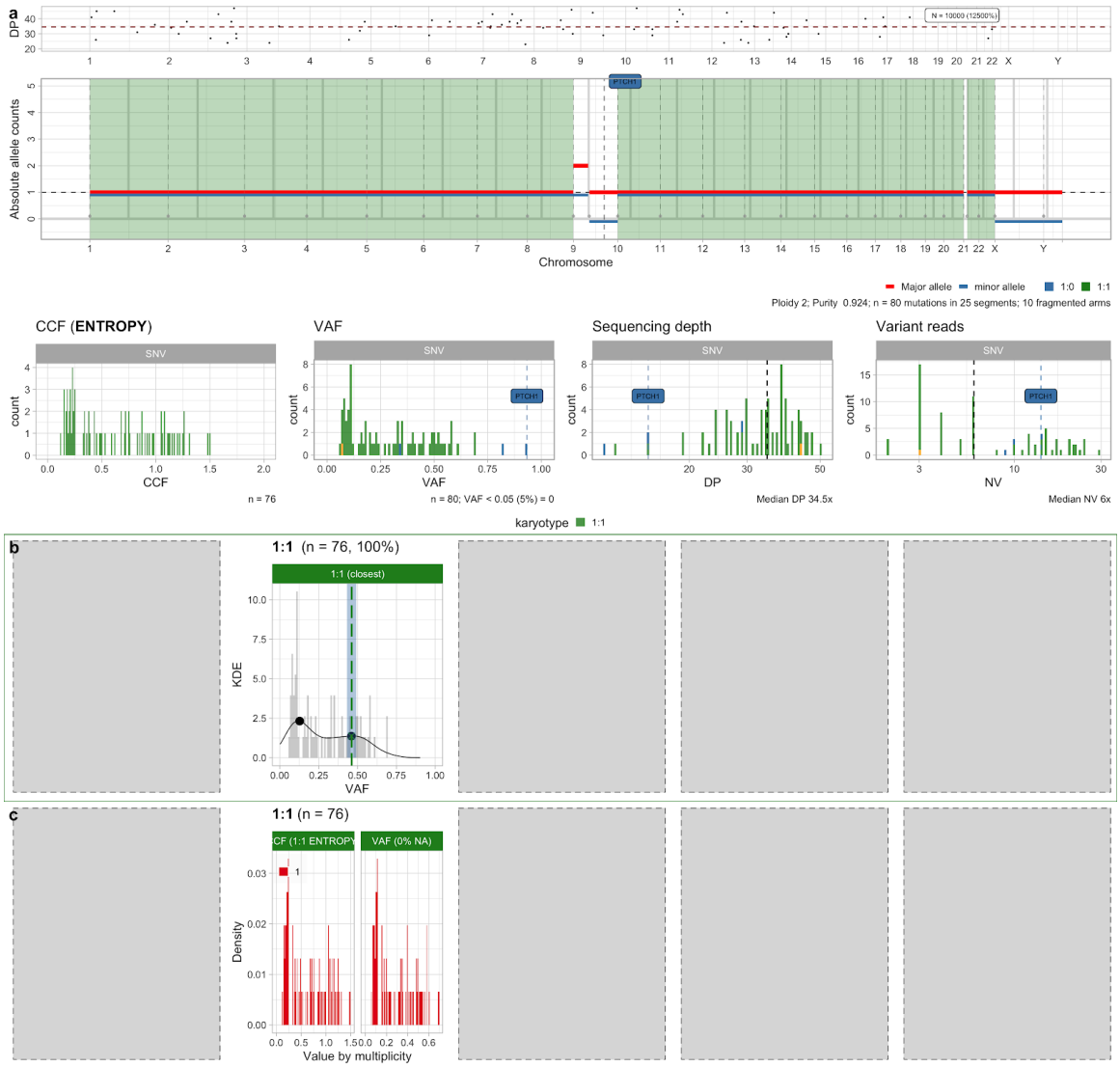
Supplementary Figure S2. Pseudocode of the peak detection algorithm and quality control strategy in CNAqc, described in Online Methods.



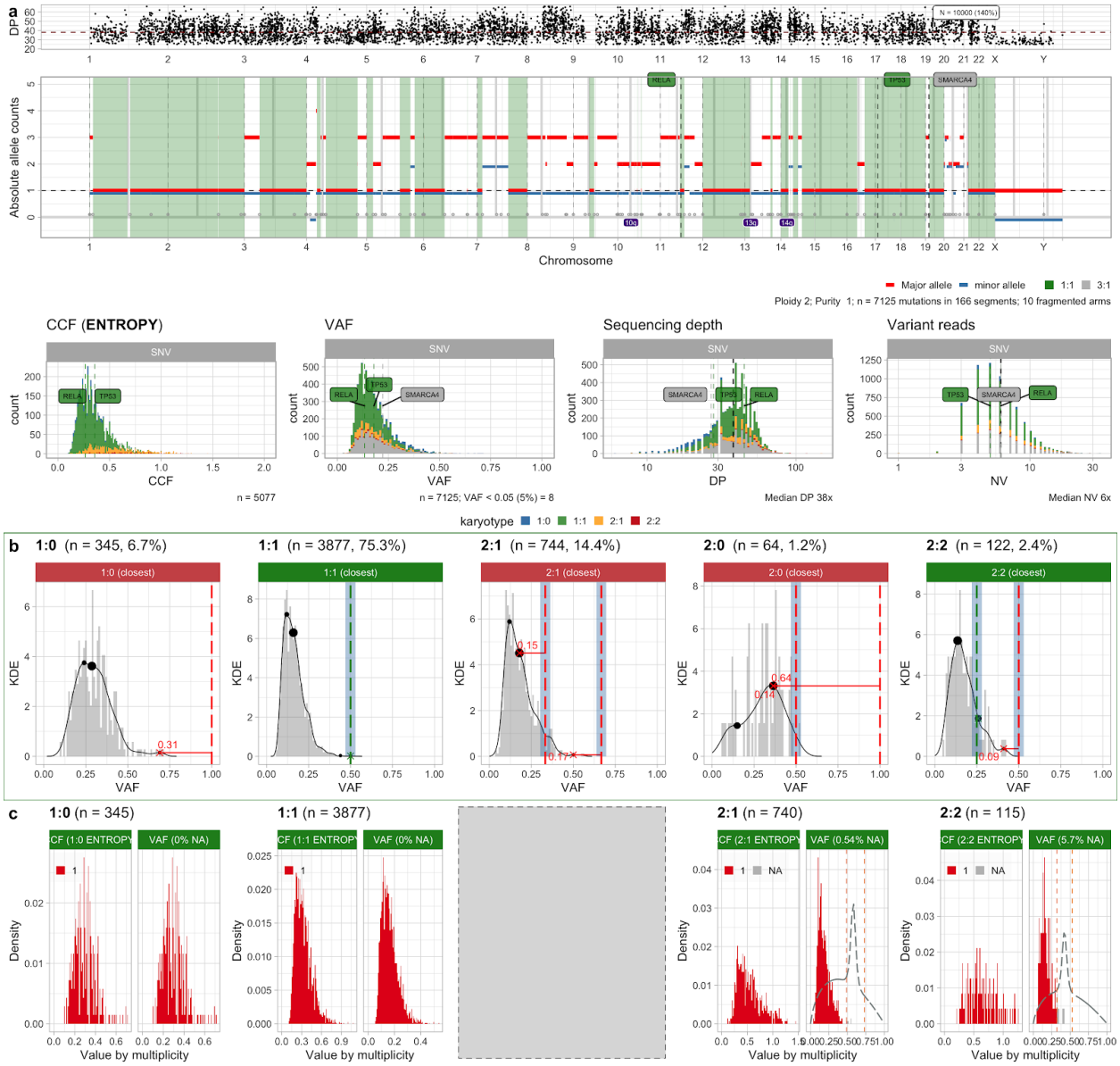
Supplementary Figure S4. CNAqc tests on synthetic tumours generated with different coverage and purity. We report the proportion of rejected samples running the tools with an error on the simulated purity (y-axis), and a tolerance to match peaks (x-axis).



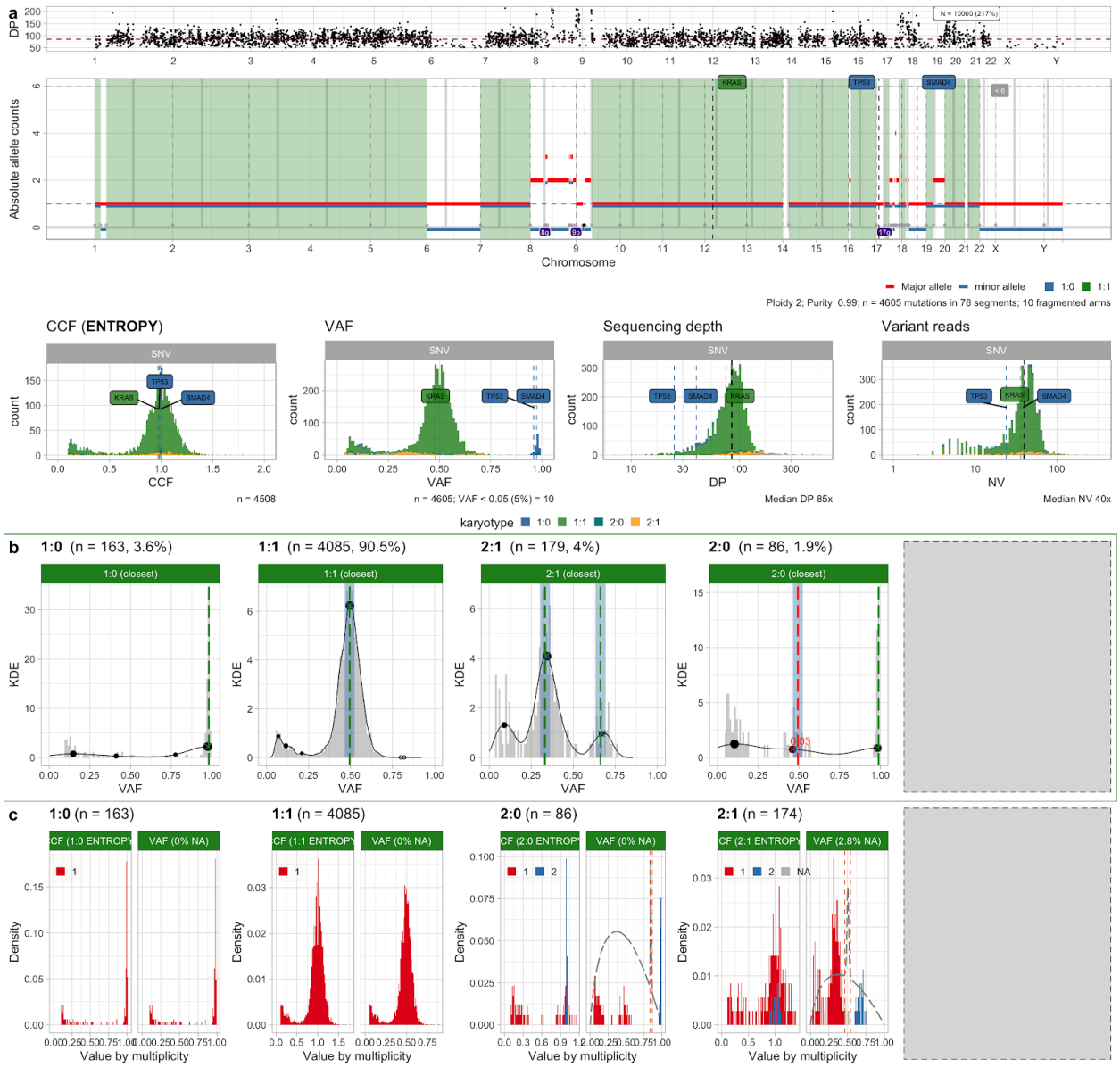
Supplementary Figure S5. a. For the simulated tumours in Supplementary Figure S4, we report the proportion of mutations for which CNAqc does assign a CCF (uncertain in phasing multiplicities), as a function of purity at fixed coverage values. The dashed line at 10% is the default parameter value to determine the final PASS or FAIL status per copy state. **b.** As in panel (a), but fixing purity.



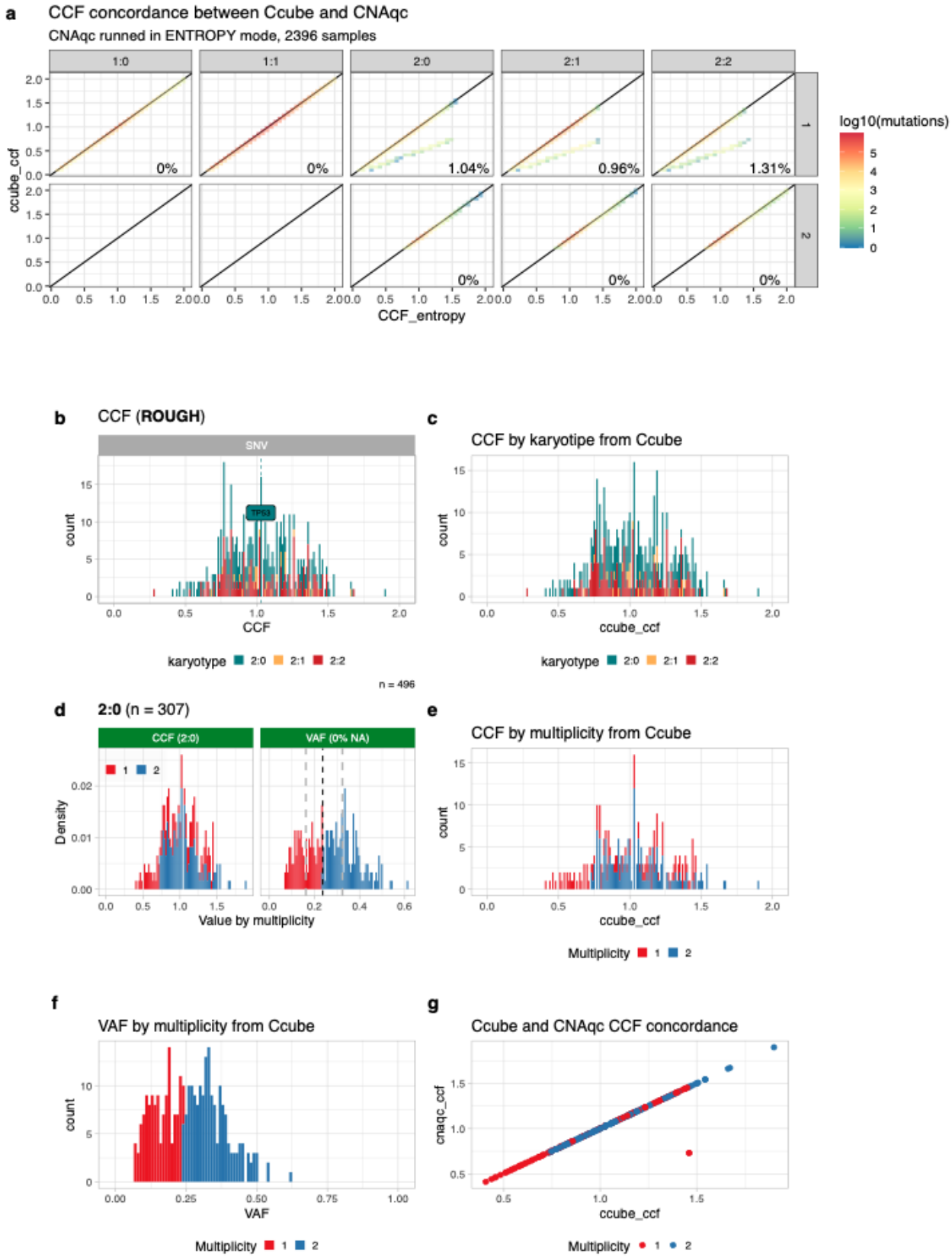
Supplementary Figure S6. Example PCAWG medulloblastoma sample with low-mutational burden, which passes data QC with CNAqc. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). Note that this sample has only 76 SNVs in diploid tumour regions, like we observe in whole-exome assays. **b,c.** Peak analysis and CCF computation for diploid SNVs.



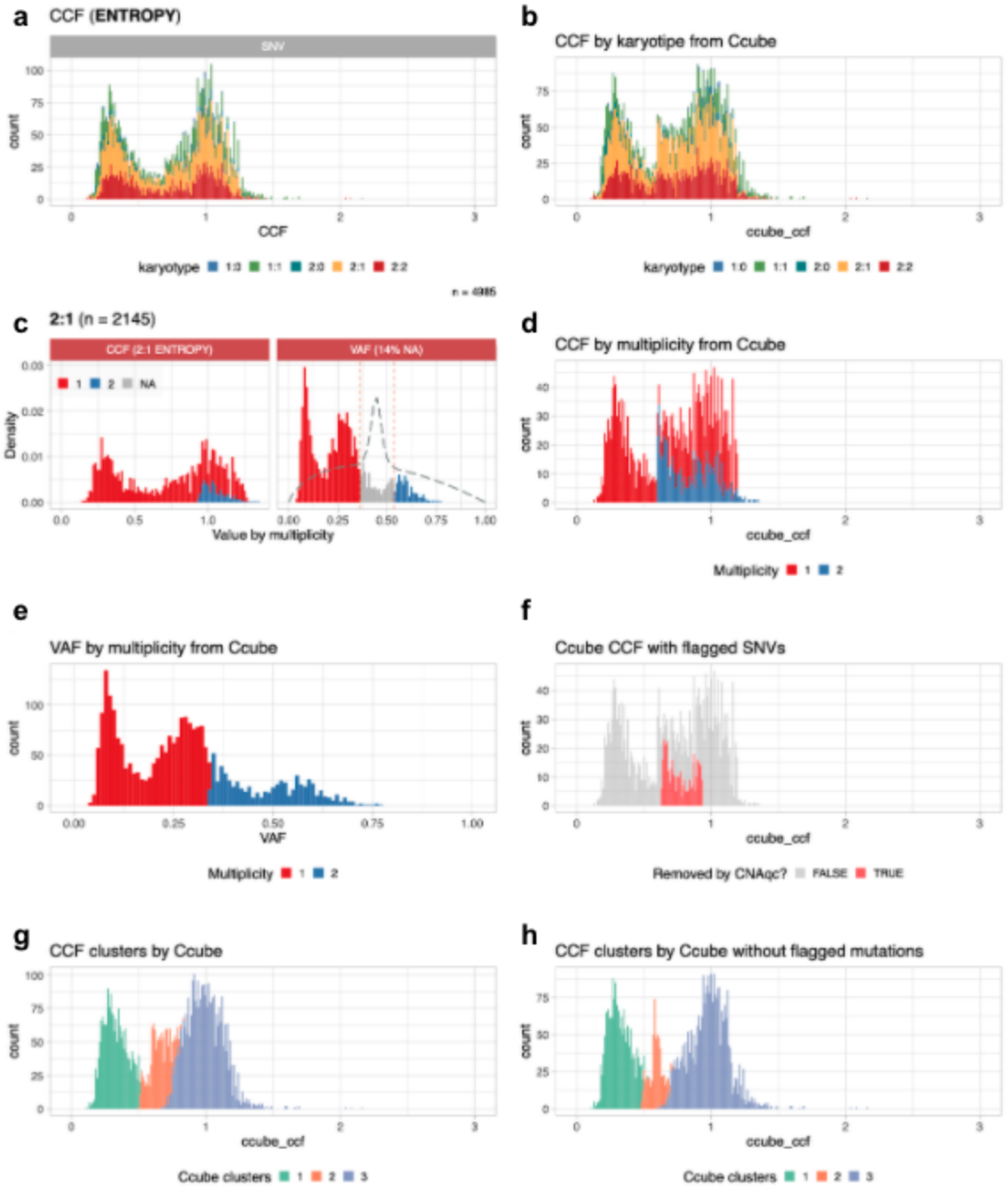
Supplementary Figure S7. Example PCAWG sample with purity of 100%. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b.** This sample has 75% of its SNVs in diploid tumour regions, where a small peak is detectable at the expected purity. The VAF clearly peaks at ~10%, possibly suggesting a purity of 20% or lower, rather than 100%. Further doubts about the current purity come from non-diploid regions, where all peaks are mismatched; for this sample CNAs called with a low-purity solution should be compared to the 100% purity solution. **c.** CCF computation for the sample. Notice that in triploid and tetraploid tumour genomes we do not find mutations present in 2 copies. Was this true then the tumour did not acquire any SNV right before the CNA. Also, here we are not cross-checking QC results from peak detection; for instance we could decide to use only mutations that map to PASS states (1:1, 2:2), and reject all others.



Supplementary Figure S8. Example PCAWG pancreatic adenocarcinoma with 99% purity (and 3 possible driver SNVs, 2 of them involving tumour suppressor genes in LOH regions). **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b.** This sample has 90% of its SNVs in diploid tumour regions, and the others in a variety of distinct CNA segments. From a peak analysis point of view, all the calls are validated. **c.** CCF values for this sample are also good.



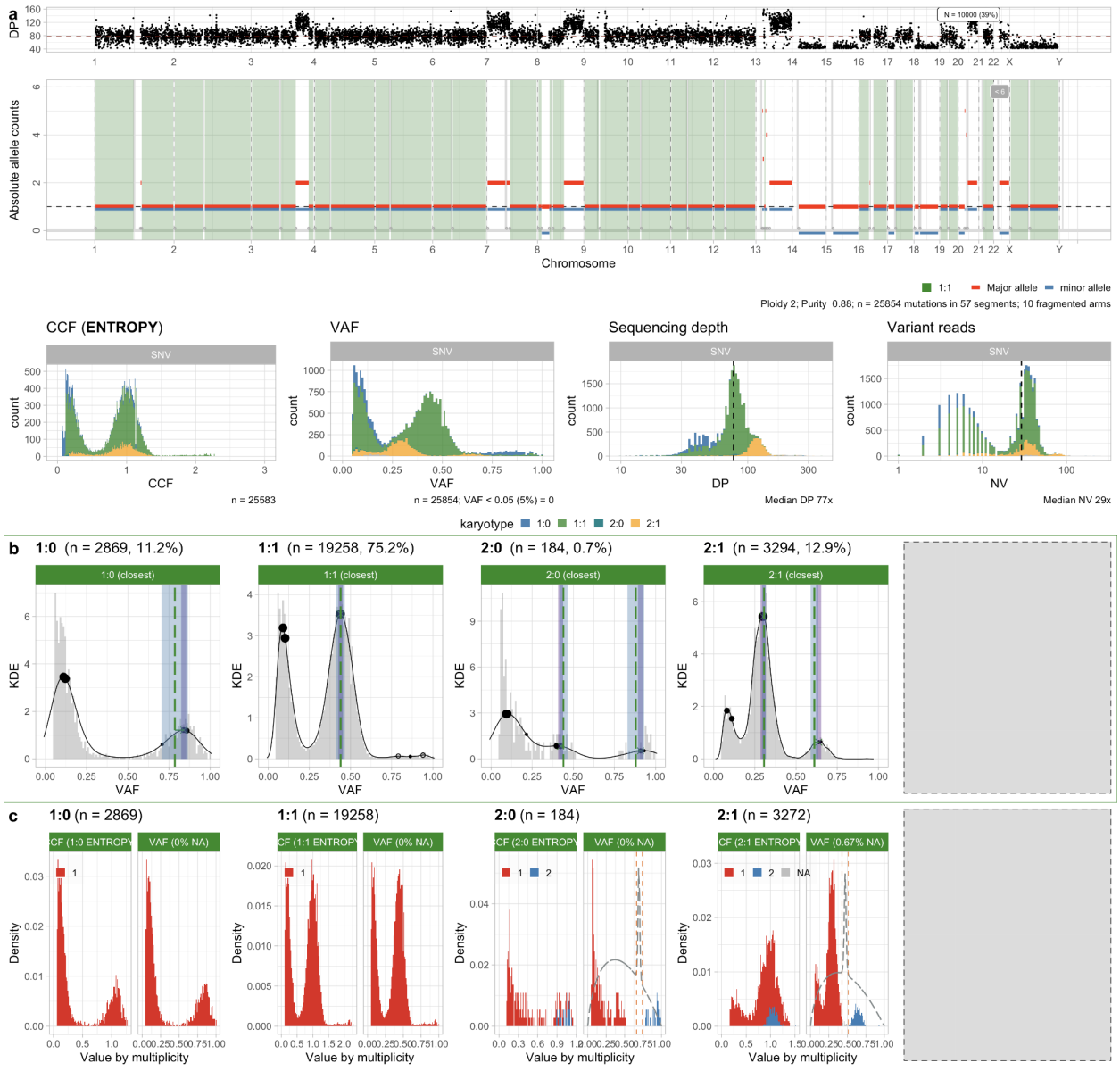
Supplementary Figure S9. a. CCF calculated by CNAqc using the “entropy” method against CCF inferred by Ccube on 2396 samples from the PCAWG cohort. Results are divided by karyotype and mutation multiplicity (taking as a reference the one inferred by CNAqc), mutations off the diagonal are discordant between the two methods. On the bottom left, the percentage of those discordant mutations over the total. **b.** CCF calculated by CNAqc using the “rough” method which assigns the multiplicity by splitting the clonal clusters at VAF level. **c.** CCF inferred by Ccube. **d.** Multiplicity assigned by CNAqc, the dashed black line depicts the splitting point to determine multiplicity. **e.** Multiplicity assigned by Ccube. **f.** VAF split by Ccube for multiplicity assignment. **g.** CCF values between Ccube and CNAqc are in almost perfect agreement (just one sample has different multiplicity). Point color is based on the multiplicity estimated by Ccube.

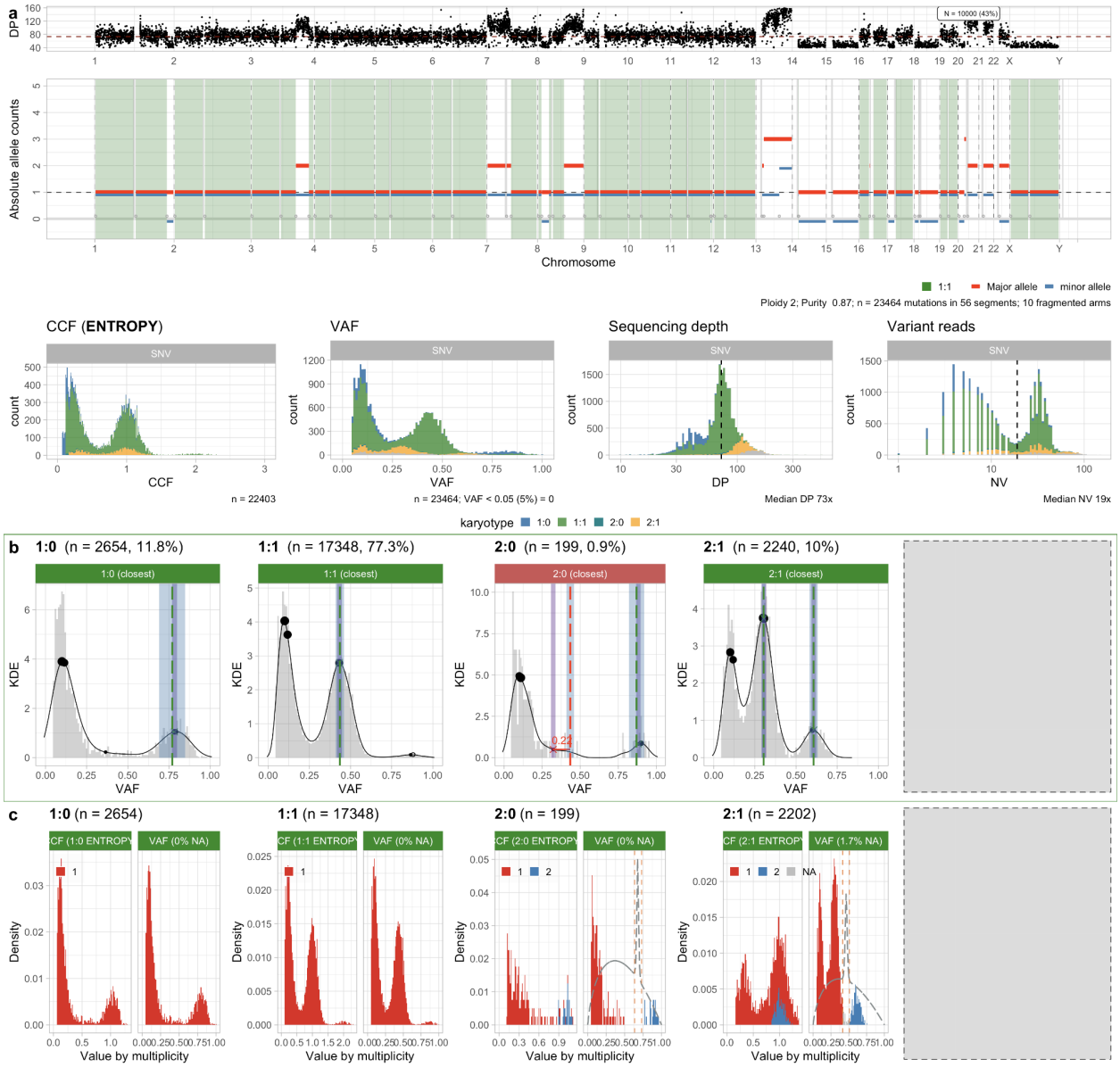


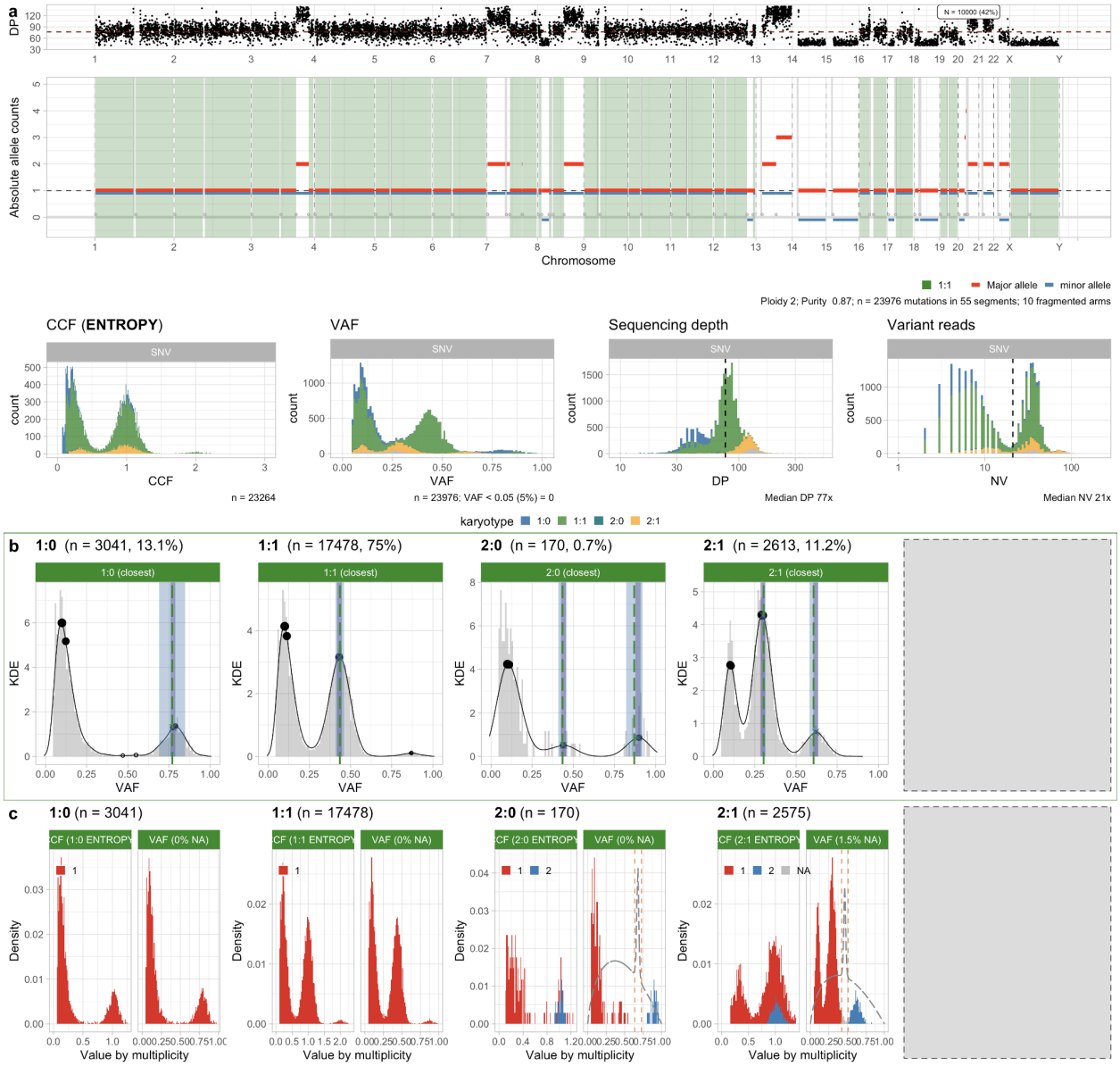
Supplementary Figure S10. a. CCF calculated by CNAqc using the "entropy" method which discards mutations with high multiplicity uncertainty. **b.** CCF inferred by Ccube. The main difference between this profile and the one inferred by CNAqc is a bump around CCF 0.6 **c.** Multiplicity assigned by CNAqc, in grey the mutations with non-estimable multiplicity. **d.** Multiplicity assigned by Ccube, it can be noted how Ccube always assigns a definite multiplicity value to each mutation. **e.** VAF split by Ccube for multiplicity assignment. **f.** High-entropy mutations discarded by CNAqc in the Ccube CCF profile. We clearly see the extra spike in CCF which could confound subclonal deconvolution, splitting the clonal cluster in multiple clones. **g.** Ccube recognizes the spurious peak as a subclonal cluster, as it is not able to accomodate for the overdispersion derived by the errors in multiplicity assignments with just one cluster. **h.** Even after

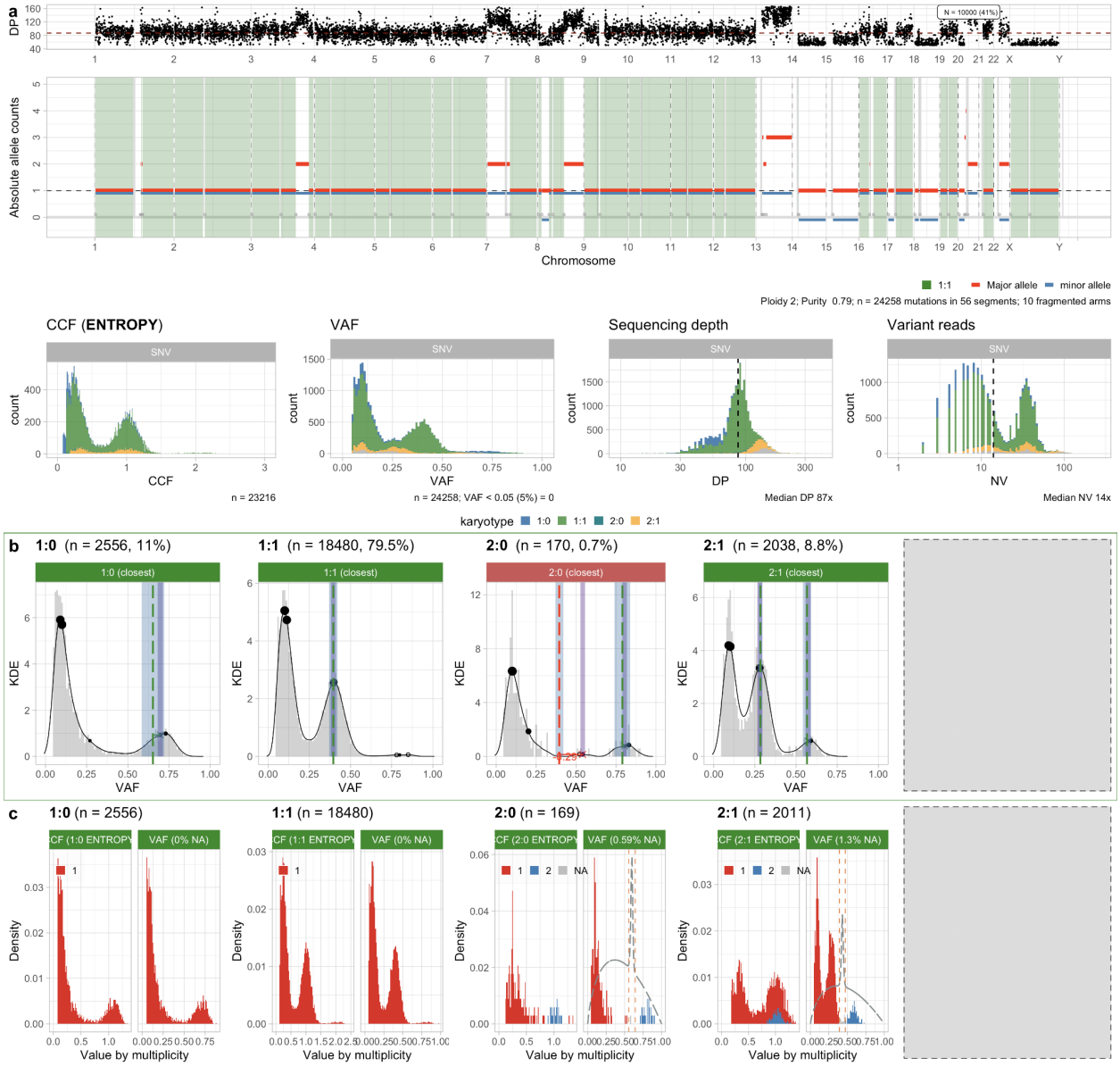
removing the mutations with high entropy from the dataset and rerunning Ccube, we can still see a peak caused by some mutations wrongly assigned to multiplicity 2. This is consistent with the choice of CNAqc to FAIL the available CCFs for this karyotype.

Supplementary Figure S11 (multiple pages). Colorectal multi-region samples (one per page): Set7_55, Set7_57, Set7_59 and Set7_62 for patient Set7. **a.** Allele-specific CNAs, and read data distribution (bottom row). **b,c.** Peak analysis and CCF computation for the sample.



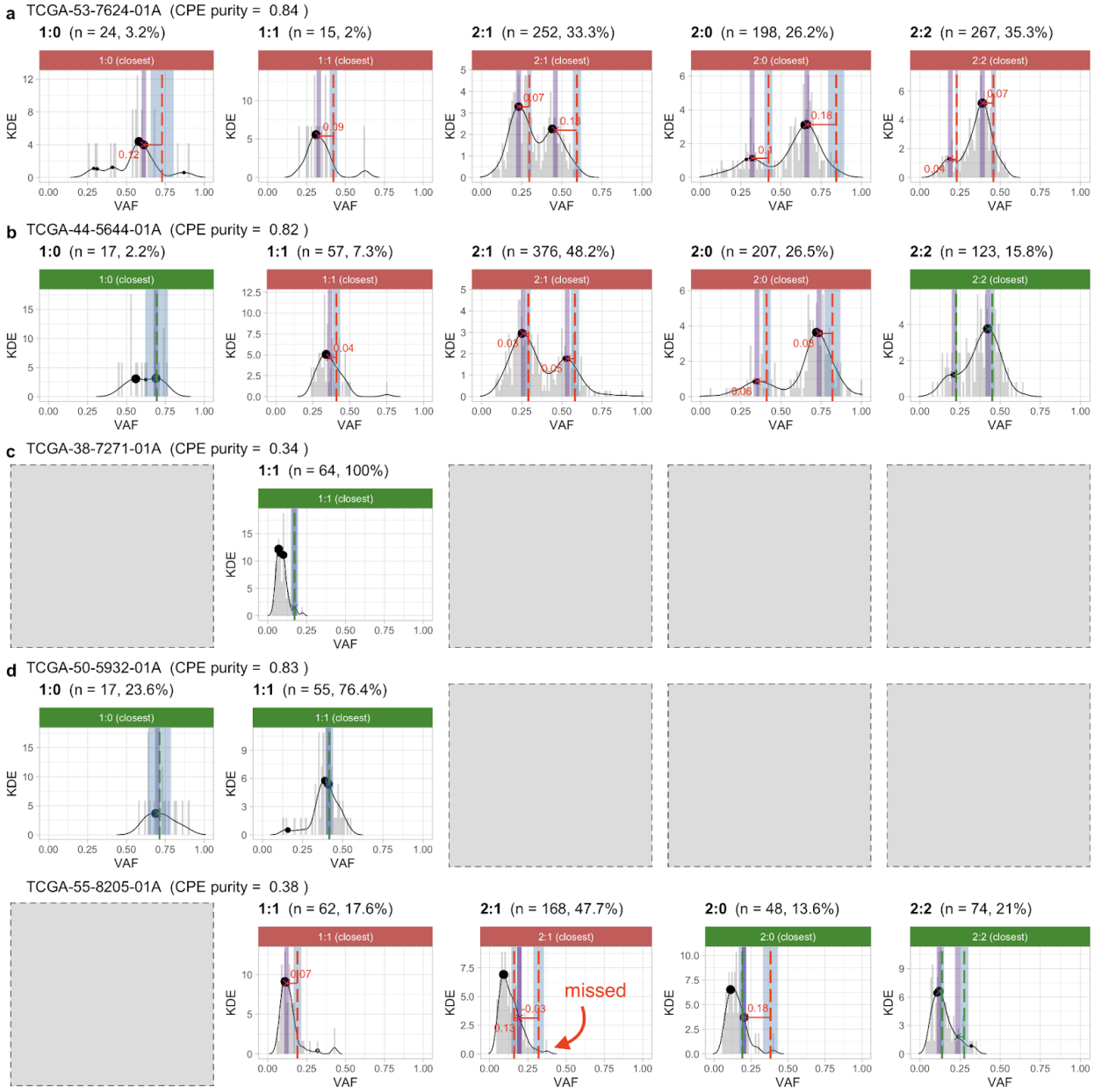




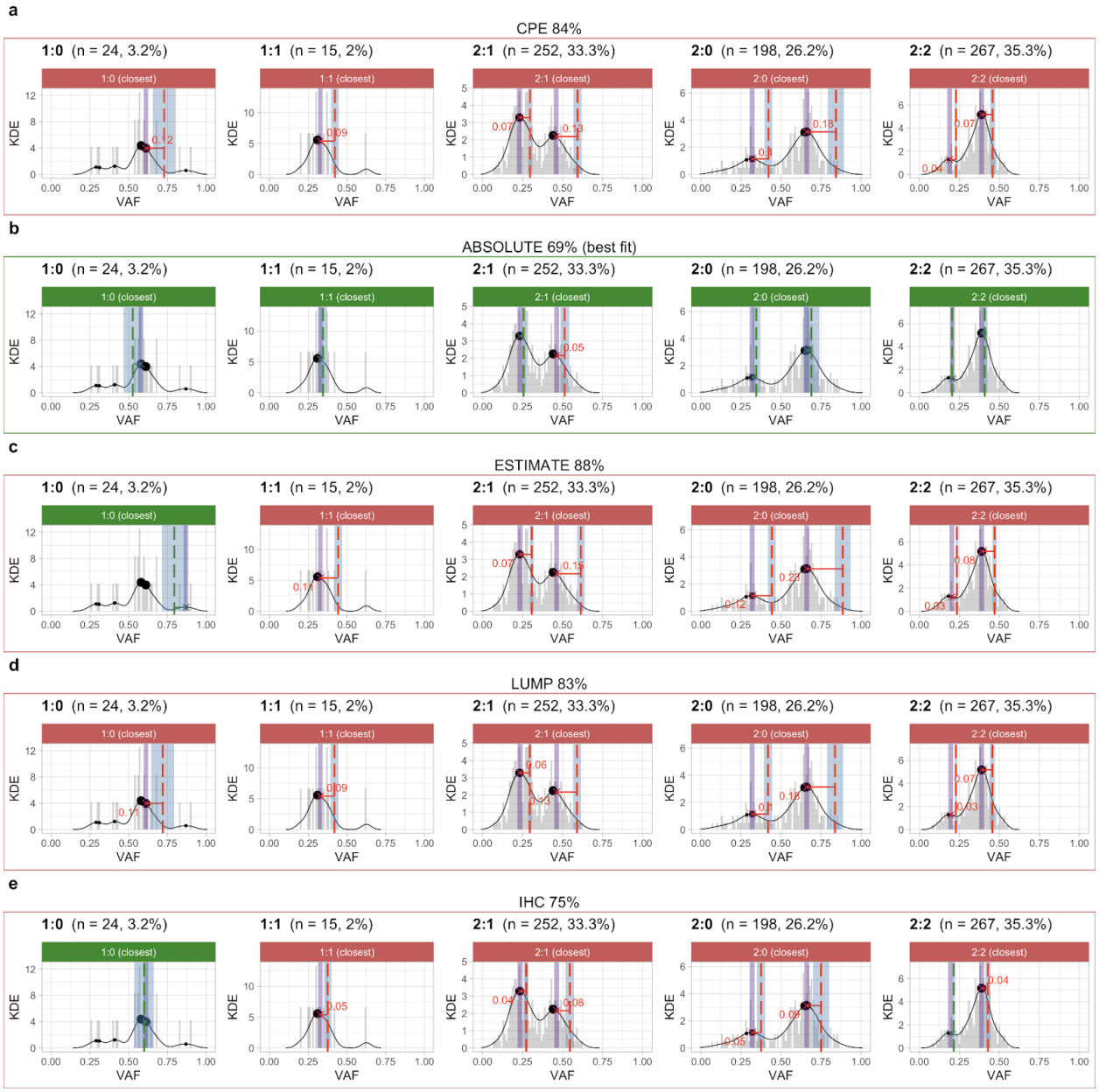


Supplementary Figure S11 ends here.

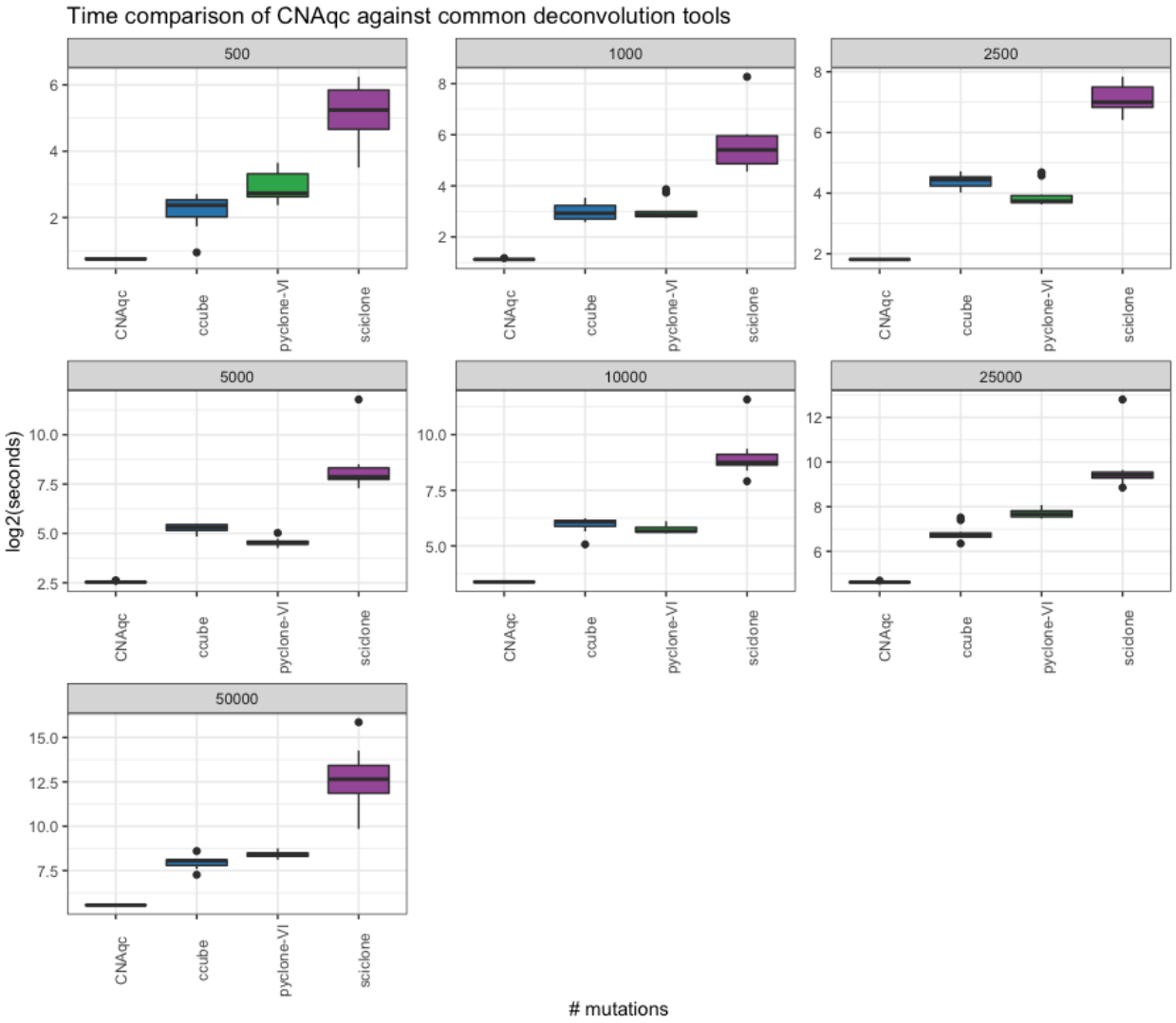
Supplementary Figure S12. a,b,c,d,e. Peak detection quality control with CNAqc, run with default parameters on colorectal multi-region samples available for patient Set_6. All calls are passed (surrounding green rectangles).



Supplementary Figure S13. a-d. CNAqc quality control via peak detection on TCGA whole-exome sequencing data of 5 lung adenocarcinomas (LUAD) with different purity values, selected from a cohort of 48 cases available online.



Supplementary Figure S14. a-e. CNAqc quality control via peak detection for LUAD sample TCGA-53-7624-01A - panel (a) of Supplementary Figure S13 - using purity estimates from CPE (consensus), ABSOLUTE, ESTIMATE, IHC and LUMP. CNAqc determines that, among all callers, only ABSOLUTE detected the true tumour ploidy (69%).



Supplementary Figure S15. Wall-clock time of CNAqc, compared with common subclonal deconvolution tools (Ccube, Pyclone-VI and Sciclone) on datasets with 500, 1000, 2500, 5000, 10000, 25000 rr 50000 mutations. The CNAqc peak detection algorithm is extremely fast and preprocesses even 50000 mutations in less than a minute (~47 seconds). The fastest deconvolution tools are Pyclone-VI and Ccube, both implemented using Variational Inference; Sciclone drops rapidly in performance as the number of SNVs increases. Time is reported in log2(seconds).