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

Jacob Househam [1,•], Riccardo Bergamin [2,•], Salvatore Milite [2,3,•], Nicola Calonaci [2,•], Alice Antonello [2,•], Marc J Williams [4], William CH Cross [5,⋆], Giulio Caravagna [2,⋆]

1) 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, UK
2) Cancer Data Science Laboratory, Department of Mathematics and Geosciences, University of Trieste, Italy.
3) Centre for Computational Biology, Human Technopole, Italy.
4) Department of Computational Oncology, Memorial Sloan Kettering, USA.
5) Department of Research Pathology, UCL Cancer Institute, University College London, UK

[•] The authors contributed equally.
[⋆]Joint last authors.

Corresponding: (GC) gcaravagna@units.it.

Abstract. Cancer genomes contain thousands of somatic point mutations, chromosome copy alterations and more complex structural variants, which contribute to tumour growth and therapy response. Whole genome sequencing is a well established approach for somatic variant identification, but its broad application comes with complications, particularly in how proposed calls are quality assessed. To address this issue, we present CNAqc, a quantitative framework to quality control somatic mutations and allele-specific copy numbers, both in clonal and subclonal settings while accounting for variations in tumour purity, as commonly seen in bulk sampling. We test the model via extensive simulations, validate it using low-pass single-cell data, and apply it to 2778 single-sample PCAWG whole-genomes, 10 in-house multi-region whole-genomes and 48 TCGA whole-exomes. CNAqc is compatible with common bioinformatic pipelines and designed to support automated parameterization processes that are crucial in the era of large-scale whole genome sequencing.

Introduction

Cancer genomes can harbour multiple types of somatic mutations compared to healthy cells and many of these events contribute to disease onset and therapeutic resistance

(Greaves and Maley, 2012; McGranahan and Swanton, 2015, 2017). The most popular experimental design to generate cancer somatic variants matches a tumour to a normal biopsy sample (Barnell *et al.*, 2019). After bulk DNA sequencing, bioinformatic tools cross reference the normal and tumour genomes to identify germline and tumour mutations (Xu, 2018), before isolating driver mutations (Gonzalez-Perez *et al.*, 2013; Bailey *et al.*, 2018). Commonly, researchers working in the field of cancer evolution also perform tumour subclonal deconvolution from somatic data (Ding *et al.*, 2012; Nik-Zainal *et al.*, 2012; Miller *et al.*, 2014; Roth *et al.*, 2014).

Several types of cancer mutations can be retrieved from DNA sequencing (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), the most common being single nucleotide variants (SNVs) and copy number alterations (CNAs) (Zack *et al.*, 2013; Li *et al.*, 2020). All types of mutations can drive tumour progression (Kent and Green, 2017; Levine, Jenkins and Copeland, 2019), and the steady drop in sequencing costs has fueled the creation of large public databases of cancer variants. Moreover, the era of high-resolution whole-genome sequencing (WGS), a technology that can read out the full tumour genome, has provided significant improvements over whole-exome or targeted counterparts. Generating some of these data, however, poses challenges. While SNVs have the most well-established detection tools (Li *et al.*, 2020), CNAs are particularly difficult to detect since the baseline ploidy of the tumour (i.e., the number of chromosome copies) has to be inferred (Van Loo *et al.*, 2010; Boeva *et al.*, 2011; Fischer *et al.*, 2014; Favero *et al.*, 2015; Cun *et al.*, 2018; Poell *et al.*, 2019). CNAs are important types of cancer mutations since gain and loss of chromosome arms can confer tumour subclones with large-scale phenotypic changes, and are therefore important clinical targets (Gerstung *et al.*, 2020; Watkins *et al.*, 2020).

Mutations and CNAs overlap within the tumour genome, with the number of copies of a mutation amplified or reduced by copy number aberrations. For instance, for a mutation present in every tumour cell and sitting in a heterozygous diploid segment, the variant allele frequency (VAF) should peak at 50% because half the tumour reads harbour the mutation. Alternatively, if the mutation sits on a triploid chromosome, the VAF should peak at 33% or at 66% depending on whether the variant sits on the non-amplified or amplified chromosome. These intuitions extend to CNAs which affect only a portion of the tumour cells, with subclonal VAF peaks depending on the tumour architecture and the abundance of each subclone (Dentro, Wedge and Van Loo, 2017). In any case, theoretical VAF peaks are affected by Binomial noise and tumour purity, i.e., the proportion of normal cells in the bulk assay, another parameter inferred together with CNAs. These ideas are also leveraged by methods that normalise VAFs by allele-specific CNAs and purity estimates in order to compute Cancer Cell Fractions

(CCFs), the fraction of cancer cells with a mutation, a popular measure for cancer evolutionary inference (Dentro, Wedge and Van Loo, 2017).

Despite a clear connection between copy numbers and VAFs, CNA and mutation calling are decoupled analyses that can return inconsistent results. Since CNAs (and tumour purity) are inferred from sequencing measurements that are subject to noise, miscalled purity and ploidy estimates are the most likely cause of error. While sometimes errors can be fixed by manual curation or consensus calling, in the era of high-resolution WGS a quality control (QC) framework for somatic mutations, clonal and subclonal CNAs and tumour purity estimates is highly desired. For this reason we have developed CNAqc, the first quantitative framework to QC allele-specific CNAs, tumour purity and somatic mutations from bulk sequencing. In particular, using fast peak-detection algorithms, i) CNAqc can QC simple clonal segments and estimate a purity error to either fine-tune copy number calling or select among multiple copy number profiles (e.g., tetraploid versus diploid tumour). Moreover, CNAqc can also QC ii) complex clonal CNAs and iii) subclonal copy numbers by determining their linear or branching evolutionary relationship. Finally, CNAqc can also determine CCFs from VAFs and implement a QC procedure that flags mutations for which CCFs cannot be estimated with confidence, as well as determine patterns of fragmentation in the data and other utilities. CNAqc can process both WGS and whole-exome sequencing (WES) in a matter of seconds, making it extremely useful for large-scale genomics consortia or retrospective analyses (Figure 1a). The tool was validated using single-cell copy number data and calibrated through extensive synthetic simulations. In this paper we show its application to 2788 WGS samples from the Pan Cancer Analysis of Whole Genomes (PCAWG) cohort (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), 10 WGS samples from 2 multi-region colorectal cancers, and 48 WES samples from The Cancer Genome Atlas (TCGA) cohort (Cancer Genome Atlas Research Network, 2014).

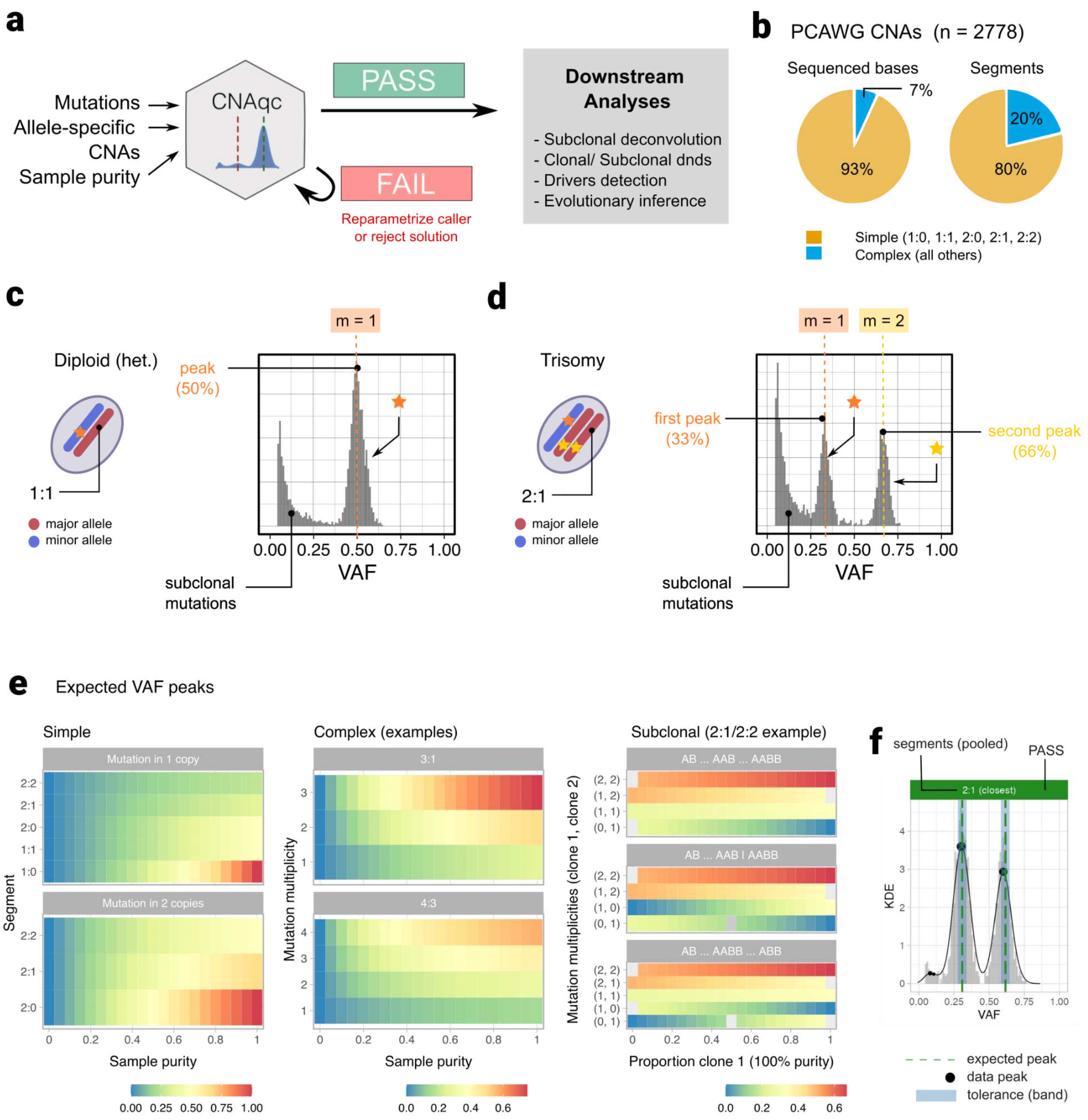

Figure 1. a. CNAqc provides quality control (QC) algorithms to integrate somatic mutations, allele-specific CNAs and tumour purity π. The tool can determine a sample-level PASS or FAIL status based on $\epsilon > 0$, a purity error tolerance parameter, returning a score $\lambda \in \Re$. Calls for a sample passing QC can be used for downstream analysis (e.g., subclonal deconvolution). Otherwise, λ can be used to correct π and re-parametrise the copy number caller, or to select among multiple segmentations/purity estimates returned by a caller (e.g., a diploid versus a tetraploid solution). **b.** CNAqc splits clonal CNAs into simple - 1:0 (LOH), 1:1 (heterozygous diploid), 2:0 (copy neutral LOH), 2:1 (triploid) and 2:2 (tetraploid) - versus complex ones. Simple CNAs comprise ~80% of ~600,000 segments and span ~93% of the bases sequenced in 2778 PCAWG samples **c.** Theoretical VAF histogram for diploid 1:1 mutations: clonal mutations peak at 50% VAF and are observed with Binomial sequencing noise, and peak at 100% CCF. **d.** The analogous case of a 2:1 segment, where we expect two VAF peaks from mutations present in one or two copies at 33% and 66% VAF respectively; both represent clonal mutations with 100% CCF. Note that the multiplicity of a

mutation phases whether it happened before or after amplification. **e.** Expected VAF peaks by theory, for simple CNAs (left), example complex CNAs (centre) and subclonal CNAs (right). Each peak originates from the combinatorial relationship between sample purity, mutation multiplicity and allele-specific CNA segments; the distance between data peaks and expected peaks laids the foundations for the QC approach in CNAqc. Left: all simple CNAs segments; peaks computed spanning tumour purity. Centre: example cases with 3:1 and 4:3 segments (mutation multiplicities range from 1 to the copies of each allele). Right: example 2:1-2:2 subclone assuming 100% purity; peaks computed for variable sizes of the first subclone (clone 1). Distinct evolutionary models are possible (linear or branching); multiplicities are relative to each of the clones involved. **f.** Example QC for mutations in 2:1 segments for a bulk assay with ~90% purity. The horizontal dashed lines are the expected VAF peaks $v_1$and v_2, determined from the table in panel (e), and with matching bandwidth (shaded area) given by a purity error tolerance $\epsilon > 0$, adjusted for the segment, multiplicity and purity. Black dots are the peaks estimated from data; since they fall within the bandwidths around $v_1$and v_2 the CNAqc QC result is PASS (green status bar).

Results

The CNAqc framework

CNAqc is implemented as an open source R package (Software Availability) to be used after variant calling from bulk sequencing, offering different ways to integrate allele-specific CNAs, tumour purity estimates and somatic mutations (Figure 1a). CNAs are classified as simple or complex, the former including heterozygous normal states (1:1 chromosome complement, AB in genotype notation), configurations of loss of heterozygosity (LOH) in monosomy (1:0, or A) and copy-neutral form (2:0, or AA), trisomy (2:1 or AAB genotype) or tetrasomy (2:2 or AABB genotype) gains. Complex CNAs are all the other combinations of alleles (e.g, 4:2 or 5:0). Simple CNAs can be generated with few (≤ 2) copy number events from a starting normal state (1:1) and are also very common in $n = 2778$ PCAWG samples, being ~80% of ~600,000 total segments and covering ~93% of sequenced bases (Figure 1b; Supplementary Figure S1). CNAqc supports subclonal segments with 2 subclones and simple CNAs involved, which represent ~70% of the Battenberg calls in PCAWG (Nik-Zainal *et al.*, 2012).

CNAqc uses the fact that the allelic configuration of a copy number segment and tumour purity determine expected peaks in VAFs. All QC procedures use peak-detection algorithms to check whether data peaks are consistent with expectations. For clonal CNAs all the mutations mapping onto a certain segment are also pooled together to increase signal strength (e.g., all 2:1). QC analyses can also be run on individual chromosomes to facilitate the identification of miscalled segments in isolated portions of the tumour genome (Supplementary Figure S2). Specifically for simple CNAs, the tool also builds a quantitative score to finally determine a pass or fail status per sample. This score suggests tumour purity corrections to adjust Bayesian priors or point parameters of a copy number caller. The same metric can also be used to select among alternative genome segmentations and purity estimates of a caller (e.g., a 100% pure diploid tumour versus a 50% pure tetraploid). Lastly, through similar ideas CNAqc also determines Cancer Cell Fractions (CCFs) for input somatic mutations, and uses a QC procedure to determine a pass or fail status based on the number of mutations with uncertain CCF estiamtes.

We briefly introduce the main ideas of the model; all mathematical details are available in the Online Methods, and inspired from earlier works (Van Loo *et al.*, 2010; Dentro, Wedge and Van Loo, 2017). The key equation models the expected VAF of a mutation that appears in a proportion $0 < c \leq 1$ of tumour cells, if the input clonal CNA segment and tumour purity were correct (Figure 1c, 1d). Note that the case of a clonal mutation is defined by $c = 1$. We consider a tumour sample with purity $0 < \pi \leq 1$, and all clonal CNA segments of the form $n_A : n_B$ which have copy numbers n_A and n_B for the major and minor alleles. If we consider mutations with multiplicity m and mapped onto $n_A : n_B$ segments, we expect a VAF peak (Figure 1e) at

$$(1) \quad v_m(c) = m\pi c / [2(1 - \pi) + \pi(n_A + n_B)] \,.$$

Notation $v_m(c)$ explicits the dependency on mutation multiplicity, and we remark that VAFs are observed with Binomial noise. Clonal mutations mapped to clonal copy numbers are a proxy for tumour purity, which we obtain in the special case of $c = 1$. For 2:0 ($n_A = 2$, $n_B = 0$), 2:1 and 2:2 segments, m phases mutations acquired before or after the copy number event (Figure 1d). Note that one advantage of simple CNAs is that $m \in \{1, 2\}$, being acquired directly from diploid heterozygous normal states (1:1). This equation can be generalised for subclonal segments by considering the ordering with which subclones emerged, i.e., their evolutionary relationship (Figure 1e). Routines to QC subclonal calls require a degree of interpretation of the tumour evolutionary trajectory, and CNAqc implements linear and branching models of subclonal growth in order to capture multiple types of allelic imbalance (Supplementary Figure S3).

For each QC task, VAF peaks are detected either via fast non-parametric kernel density or slower parametric Binomial mixtures (Supplementary Figure S4); the approaches are joined only for simple CNAs, when the tool computes purity adjustments for caller recalibration. This is to optimise speed and dedicate more resources to QC clonal simple CNAs; being most prevalent, clonal simple CNAs determine the final tumour genome segmentation and sample purity, while more complex segments or subclonal segments usually follow directly from clonal ones. The CNAqc sample score $\lambda \in \Re$ (e.g., $+3\%$, -7%) is based on a linear combination of the distances between the expected peaks and the data peaks. Therefore the score represents an error reflecting corrections to the input purity π. The threshold to determine PASS or FAIL is the error $\epsilon > 0$ in the purity estimate the user decides to tolerate: 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, ϵ is adjusted for CNA segments, mutation multiplicity and tumour purity (Figure 1f).

CNAqc can also determine CCFs for mutations sitting on clonal simple CNAs; the tool normalises the VAF v of a mutation that sits on segment n_A: n_B by solving formula (1) for c, leading to

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

Notice that here we denote by v a generic VAF, omitting its dependency from m and c. This equation applies to clonal and subclonal mutations, and its main challenge is phasing m from VAFs, i.e., determining if a mutation is in single ($m = 1$) or double ($m = 2$) copy. Phasing is certainly easier with simple CNAs because $m \in \{1, 2\}$. CNAqc uses a two-component Binomial mixture to phase multiplicities and determine phasing uncertainty; to the best of our knowledge, no other CCF-based tool introduced this or similar notions of uncertainty. In particular, our method identifies a VAF range at the crossing of the mixture components, where m cannot be clearly phased and CCFs are unassigned; here uncertainty is estimated from the entropy $H(z)$ of the mixture latent variables z. For every CNA, a CCFs's PASS or FAIL status is determined from the maximum proportion of unassigned mutations that one wants to tolerate. An alternative method is available to force m a value through a hard split on the VAF, regardless of entropy (Online Methods).

CNAqc provides several functions to visualise segments (Figure 2a), read counts (Figure 2b,c), CCFs (Figure 2d) and, after QC, peak-based visualisations for both VAFs (Figure 2e) and CCFs (Figure 2f). The tool can also visualise subclonal CNAs (Figure 3a) or process hypermutant tumour with hundreds of thousands somatic mutations (Figure 3b). In Figure 3c we show an example QC of complex CNAs from 3:0, 3:1, 3:2, 3:3, 4:0 and 4:2 segments. For this example sample, CNAqc also validates a large subclonal CNA on chromosome 11, assigned to a 2:1 subclone (21% of cells) and a 2:2 subclone (79% of cells). For this complex sample, most peaks are matched and found to be compatible with both linear and branching models of evolution (Figure 3d). Finally, CNAqc contains extra utilities to smooth copy number segments and detect patterns of over-fragmentation from breakpoint frequencies, which help to determine events of chromothripsis, kataegis or chromoplexy (Zack *et al.*, 2013; Gerstung *et al.*, 2020).

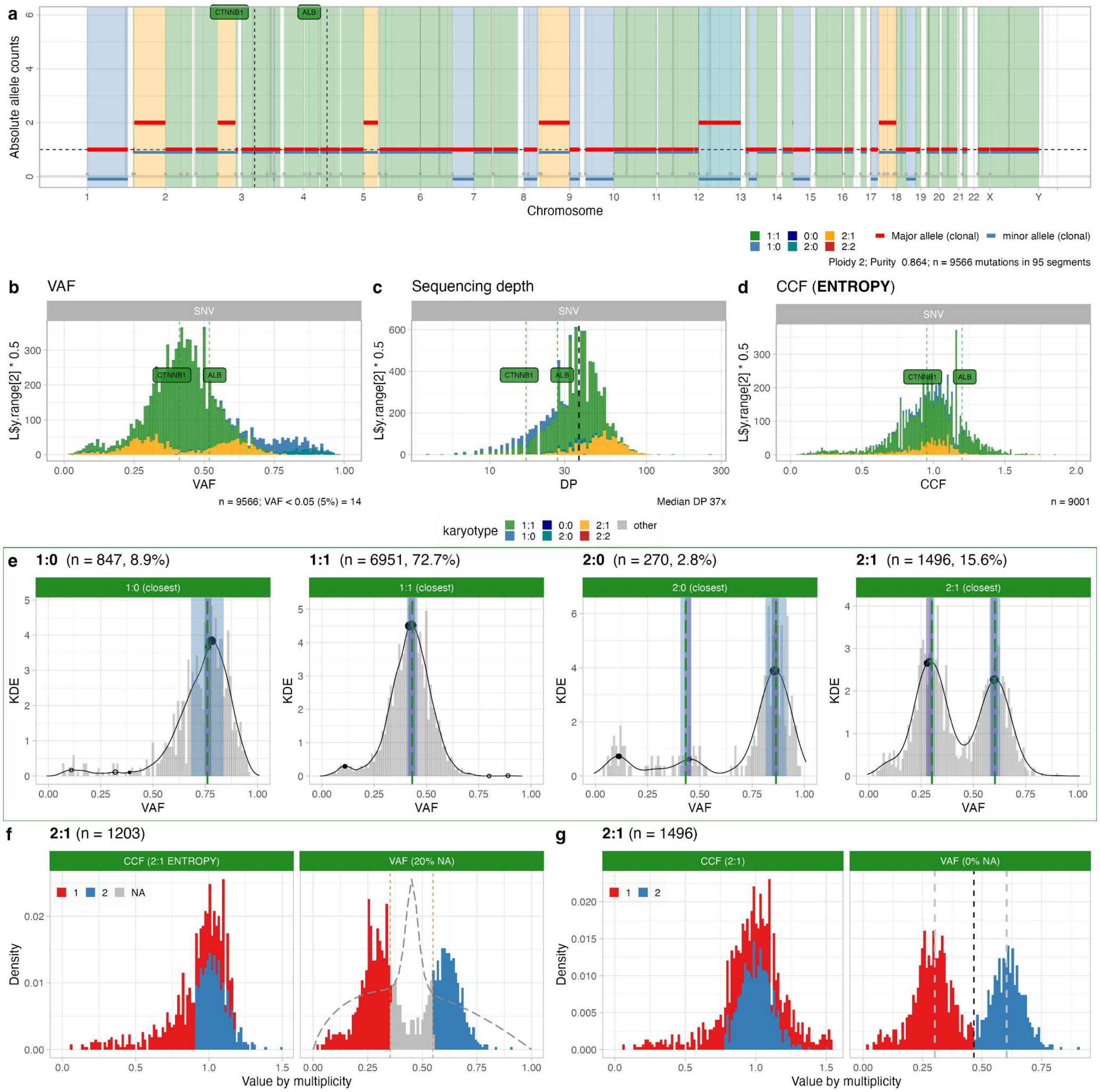

Figure 2. a. CNAqc visualisation of PCAWG sample ca5ded1c-c622-11e3-bf01-24c6515278c0 (DCC project code LIRI-JP, hepatocellular carcinoma). Genome-wide allele-specific consensus CNAs (ploidy 2, purity ~85%). The plot shows major and minor allele counts in each segment. This sample harbours two driver SNVs hitting genes CTNNB1 and ALB, which are shown annotated in diploid heterozygous segments (1:1). **b,c.** Read count data for SNVs visualised as Variant allele frequencies (VAFs) and depth of sequencing (DP). **d.** Cancer Cell Fractions (CCF) obtained by CNAqc show that the two CTNNB1 and ALB drivers are clonal. **e.** Peak detection QC for simple clonal

CNAs and tumour purity. Multiple peaks are checked independently for every copy state with total copy number >2 (2:1 and 2:0); final QC depends on the number of mutations assignable to each peak, and whether the peak is matched. The sample-level QC is a linear combination of results from each copy state. Here calls are assigned a PASS status (green box surrounding plot). **f,g.** CCF estimation for mutations mapping to triploid 2:1 segments, obtained using the entropy-based method as in panel (d), and the rough method. CCF values of clonal mutations spread around 1 as expected. The panels show CCFs and VAFs, coloured by mutation multiplicity phased from VAFs. The entropy profile is the dashed line, and the grey area delineates crossings of Binomial densities where CNAqc detects multiplicities uncertainty; the entropy method detects uncertainty in the phasing of 20% of the SNVs. The rough method in panel (g) assigns multiplicities regardless of uncertainty. In both cases the CCF estimates PASS quality control.

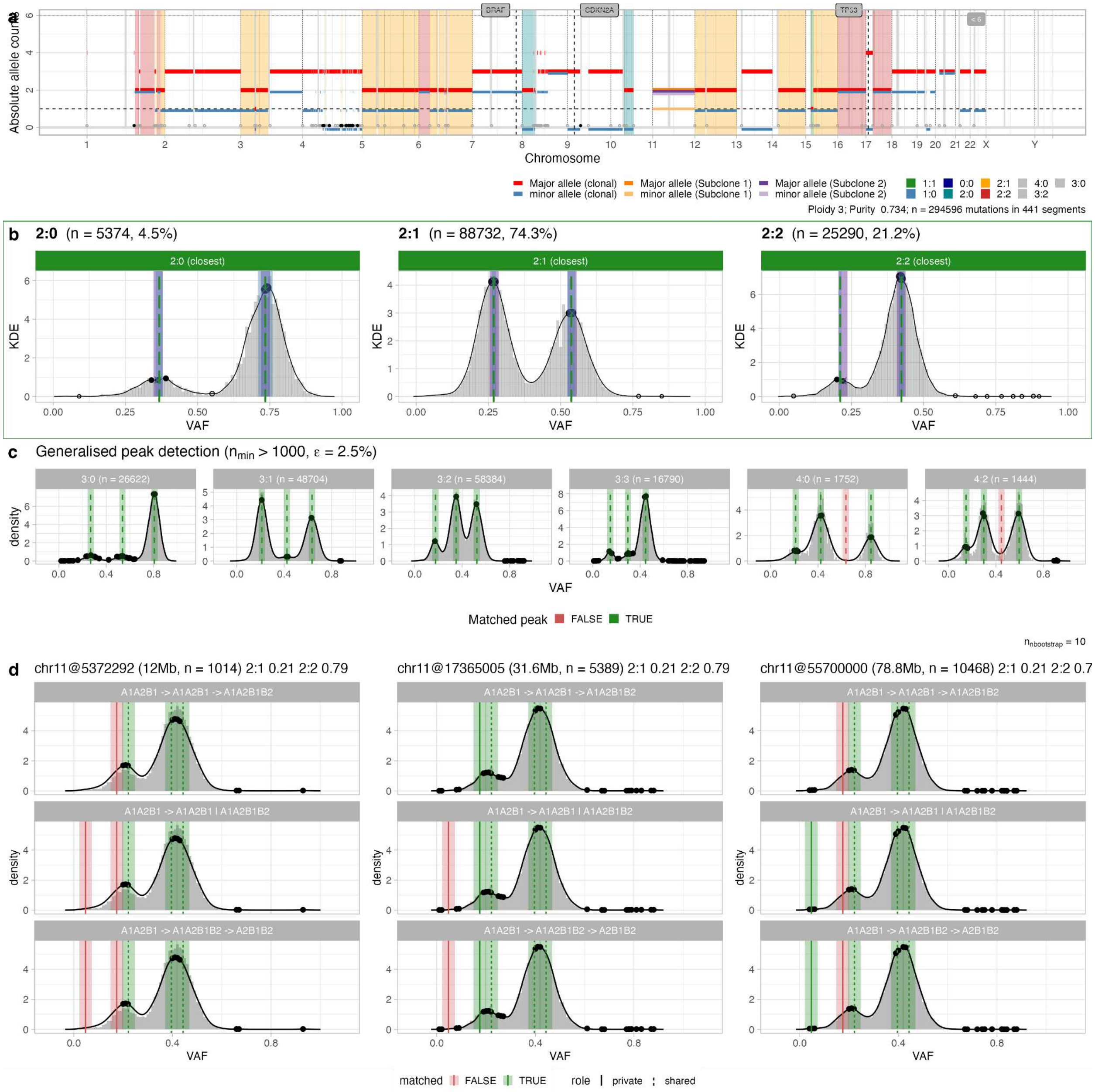

Figure 3. a. CNAqc visualisation of PCAWG sample 51893d3f-e7f3-43f9-9fd0-c0f25ae96804 (DCC project code MELA-AU, skin melanoma), which presents high levels of aneuploidy (mean ploidy 3.69) and most genome covered

by 2:1 segments. The sample has a very large mutational burden (~294,000 somatic mutations), and large subclonal CNAs on chromosome 11. **e.** CNAqc validates the calls by peak detection; most of the signal is due to ~80,000 mutations mapping to 2:1 segments. **c.** CNAqc also validates 18 out of 20 expected peaks in more complex copy number segments, which in this sample are 3:0, 3:1, 3:2, 3:3, 4:0 and 4:2. **d.** CNAqc also validates subclonal CNAs on chromosome 11 obtained by Battenberg (Nik-Zainal *et al.*, 2012), which calls two subclones with 2:1 genome (21% of cells) and 2:2 genome (79% of cells). In all segments almost all peaks are matched, and the peaks are compatible with both linear and branching models of evolution starting from an ancestral 2:1 state.

Model simulations, single-cell validation and parameters calibration

First, 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), using simple CNAs (Online methods). 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. 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 how VAF quality impacts performance, with low coverage/ purity making peak detection harder (Supplementary Figure S5). 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 role of data resolution to assess reliable CCFs (Supplementary Figures S6).

Second, we validated CNAqc with low-pass single-cell copy number data from ovarian cancer cell lines generated using the Direct Library Preparation (DLP+) protocol (Laks *et al.*, 2019), an amplification-free whole-genome sequencing method suitable for calling single-cell CNAs and somatic mutations (Figure 4a) (Williams *et al.*, 2021). First we focused on 3 tumour subclones which represented 100%-pure monoclonal tumour populations with different CNAs (Supplementary Figure S7). From read count pile-ups we generated mutation data, and gathered CNAs from consensus clone-level copy number calls (Online Methods). The CNAqc algorithms passed the true 100% purity (Figure 4b) using both simple and more complex clonal CNAs (Figure 4c, Supplementary Figure S8). We then pooled clonal segments across the subclones to create a larger clonal population, and tested the quality of purity-adjustments proposed by CNAqc. The test showed that, if we input a purity value of the form $1 - \epsilon$ with $\epsilon \geq 0$, the adjustment proposed by CNAqc follows ϵ linearly ($0.88 < R^2 < 0.99$; $p < 10^{-16}$) (Figure 4d, Supplementary Figure S9). Finally, we assessed subclonal CNA QC upon introduction of subclonal population mixtures with both triploid (2:1) and tetraploid genome-doubled (2:2) subclones; CNAqc could match VAF peaks from subclonal CNAs (Supplementary Figure S10). Moreover, using analogous 10x data from (Zaccaria and Raphael, 2021) and a similar procedure, we could create a mixture of 2 subclones with subclonal trisomy (2:1) characterised by mirrored allelic imbalance (i.e., the joint presence of AAB and ABB genotypes, Supplementary Figure S11). Notably, CNAqc could validate these subclonal CNAs and was also able to identify, from the pseudo-bulk, the correct AB → AAB | ABB evolutionary model that generated this mixture (Figure 4e, Supplementary Figure S12). This shows that, besides performing QC, CNAqc can give insights into the underlying evolutionary process that generates CNA-associated tumour sub-populations.

Third, we performed further tests (Online Methods) to support users in choosing the best value for the parameter which modulates QC (ϵ), as a function of data coverage and input purity. To do this, we generated >350,000 synthetic tumours spanning plausible coverage and purity ranges, and then performed a test to measure the False Positive Rate (FPR) of the tool – i.e. the probability of passing a sample that should be failed, for a certain value of ϵ. Then, for every coverage and purity configuration we regressed the FPR against ϵ, so that the tool can suggest, for a desired upper bound on FPR, the best value of ϵ for based on coverage and purity of the input dataset (Supplementary Figure S13).

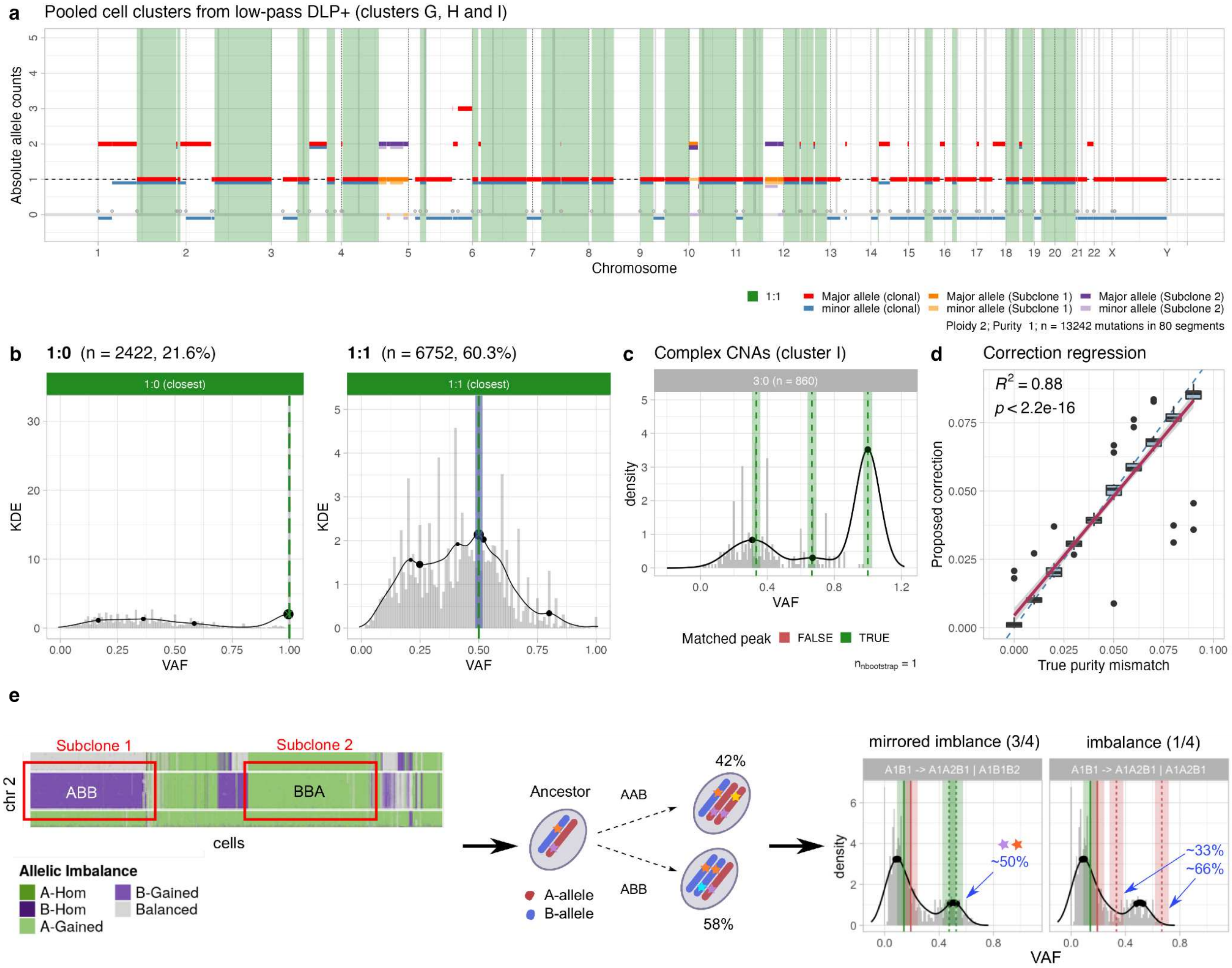

Figure 4. **a.** Pseudobulk CNA segments from single-cell low-pass DLP+ data of an ovarian cell line (Laks *et al.*, 2019) used to validate CNAqc. This copy number profile is obtained from pooling simple CNAs across several diploid tumour subclones (Supplementary Figure S7). **b.** Quality control of LOH and diploid heterozygous segments from panel (a) with CNAqc, using true tumour purity (100%, monoclonal tumour). **c.** Quality control of more complex 3:0 segments identified in cluster I from Supplementary Figure S7. **d.** Correlation between purity-adjustments suggested by CNAqc and input errors imputed to data from panel (a). By construction the desired correction needs to sit on the plot diagonal; the tool achieves $R^2 = 0.88$, p-value $p < 10^{-16}$. **e.** Quality control of subclonal CNAs from admixing of tumour subclones with mirrored allelic imbalance. Here, using data from (Zaccaria and Raphael, 2021) we created two subclones with 2:1 genotypes and mirrored alleles (AAB versus ABB). CNAqc can validate these calls via subclonal peak detection, and identify the true branching patterns of evolution from the input data (AB → AAB | ABB), which is characterised by a peak of shared mutations at around 50% VAF.

Large-scale pan cancer PCAWG calls

We have run CNAqc on consensus calls (single nucleotide variants, allele-specific CNAs and purity) from the PCAWG cohort ($n = 2778$ samples, 40 tumour types).

Excluding samples with unsuitable data, we ran 2589 samples on a 36-core machine in <1 hour. Median depth of sequencing and purity of PCAWG are 45x and ~65% (Caravagna *et al.*, 2020), a resolution comparable to our simulations (Supplementary Figure S1). PCAWG consensus calls are obtained from several callers and selected with a multi-tier classification system; simple CNAs were passed by CNAqc in 2339 out of 2589 samples (~90%) with 3% error purity tolerance ($\epsilon = 0.03$), confirming the quality of the calls (Figure 5a). As with simulations, the QC pass rate was determined by tumour purity and coverage (Figure 5b). Manual inspections of some samples presented interesting cases. For instance, tumours with low burden but high quality calls still yielded a useful report (Supplementary Figure S14). Other cases instead with 100% consensus PCAWG purity were at odds with VAF peaks and, upon re-analysis, were found to be compatible with low purity solutions (Supplementary Figure S15). Other cases did possess genuinely very high purity (>95%, Supplementary Figure S16), instead.

CNAqc performed QC of 610 PCAWG samples carrying complex CNAs in segments harbouring >150 mutations (Figure 5c). The most prevalent segments were 3:1, 3:2 and 3:0 (15%, 13% and 10%), for which we matched >60% of peaks on average (Online Methods). Tumour types with higher prevalence of complex CNAs were esophageal, liver, melanoma, ovarian, pancreatic and breast cancers, each with >100 patients carried complex CNAs. Finally, we applied CNAqc to subclonal CNAs called by Battenberg (Figure 5d). Overall, the tumour types carrying most subclonal CNAs were esophageal, liver, melanoma, pancreatic and gastric cancers. Interestingly, some of these tumour types also carried complex CNAs, suggesting the existence of some biological mechanism driving instability in these tumours. The most frequent combinations of CNAs analysed were those that could develop through deletion or amplification of a single allele – 1:0-1:1 (~33%) and 1:1-2:1 (~31%), followed by 2:1-2:2 (~19%), 2:0-2:1 (~9%) and 1:0-2:0 (~7%). For each segment, we determined the best fitting model by assessing the percentage of matched peaks; we did not assign any model if less than 50% of the expected peaks were matched and, overall, we assigned a model to ~87% of the subclonal segments (Supplementary Figure S18 and S19). Interestingly, subclones 1:0-2:0, 1:1-2:1 and 2:1-2:2 were generally better explained by linear evolution implying a temporal ordering among the subclones. This was the best model in 39%, 48% and 52% of cases, including for subclones 2:1-2:2 in triploid 2:1 (30%) or tetraploid 2:2 (33%) tumours. Conversely, subclones 2:0-2:1 were better explained by branching models (38%), implying the independent formation from a common ancestor. In subclones 1:0-1:1, instead, while we validated >50% peaks in >73% samples, the linear and branching models were found indistinguishable (by construction). Finally, at the level of tumour types, 2:0-2:1 subclones in esophageal adenocarcinomas were well explained by both models, while 2:1-2:2 subclones were better explained by linear models, as in the case of liver and pancreatic cancers, and melanoma.

We also tested CCF computations on the whole PCAWG. Consistent with simulated data, the percentage of mutations for which CCF cannot be computed negatively correlated with sample purity (Figure 5e). We found the CCFs produced by CNAqc (Supplementary Figure S20) comparable to those computed by Ccube (Yuan *et al.*, 2018) across the whole cohort, but also found cases where, upon CNAqc analysis, we could detect spurious subclonal clusters explained by miscalled mutation multiplicities (Supplementary Figure S21). This further showed the practical utility of CNAqc, which flags mutations whose CCF values are uncertain and this can be used, post-hoc, to check the plausibility of subclonal deconvolution results. Summarising, from this analyses we could determine that, while peaks could be detected for almost all PCAWG samples, mutation multiplicity phasing would have required higher coverage and purity to reduce uncertainty.

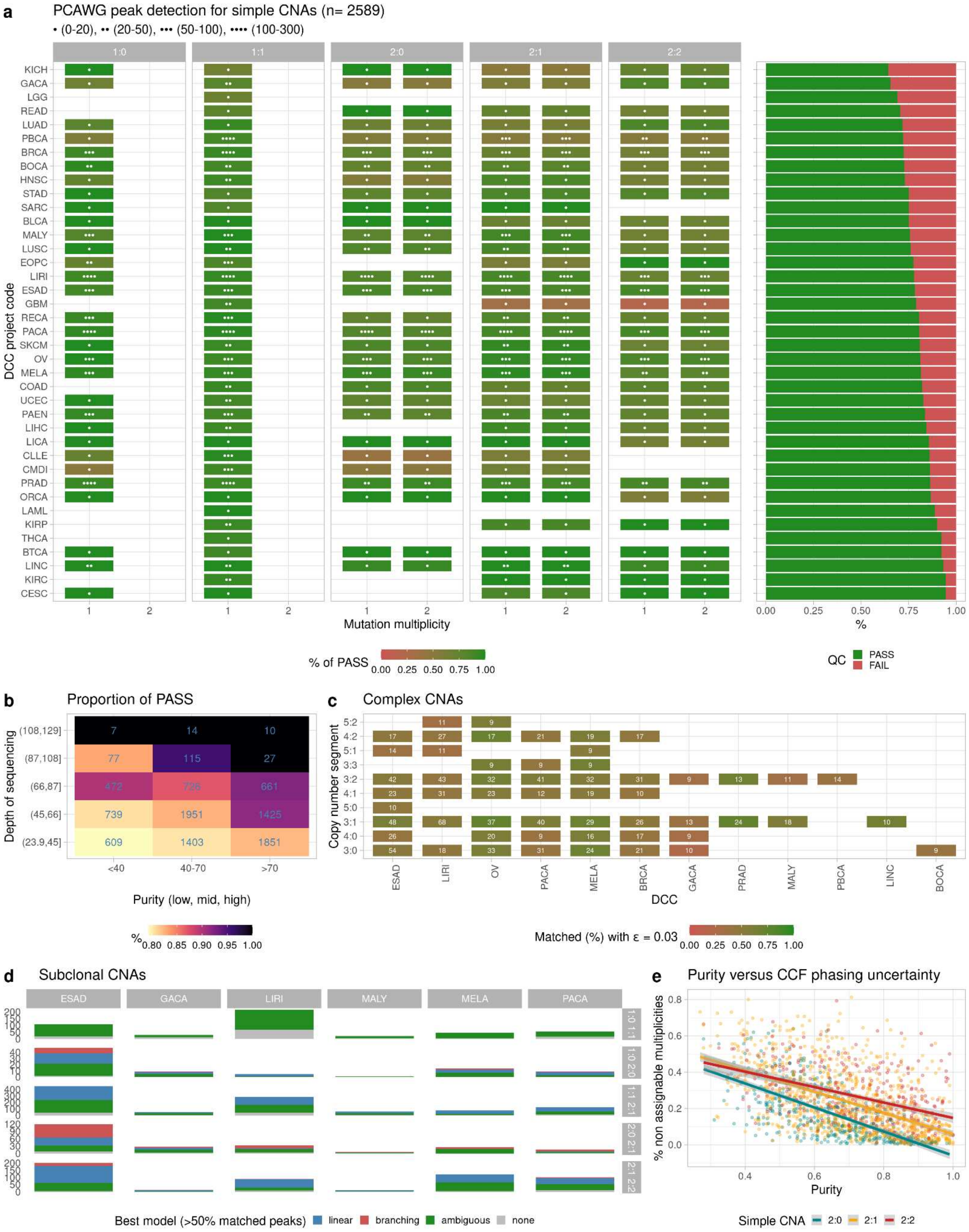

Figure 5. a. Peak detection quality control for $n = 2723$ WGS samples available in PCAWG. The plot shows the percentage of cases with PASS status, split by copy state, multiplicity and tumour type. The dots annotated report the number of cases, the barplot the tumour types sorted by percentage of sample-level 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. **c.** Complex CNAs in PCAWG for segments with >150 somatic mutations (n = 570 samples), the figure shows the top ten tumour types ranked based on the number of samples with complex calls (>9) **e.** Subclonal CNAs in PCAWG for segments with >150 somatic mutations (n = 538 samples), the figure shows the top six tumour types ranked by the number of samples with subclonal CNAs. **e.** Regression of tumour sample purity against the proportion of CCF values that cannot be confidently assessed by CNAqc, split by copy state.

Multi-region colorectal cancer data

We have run CNAqc on previously published WGS multi-region data (Cross *et al.*, 2018; Caravagna *et al.*, 2020), which was collected from multiple regions of primary colorectal adenocarcinomas (10 samples from 2 patients; median coverage ~80x, purity ~80%, Figure 6). We combined 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 select the best tumour segmentation among multiple ones. With this test, we wanted to observe if CNAqc could be used to choose among alternative copy number profiles generated by a copy number caller. As baseline, we compared the Sequenza solutions to published CloneHD (Fischer *et al.*, 2014) calls for this sample.

Sequenza was first run with the default range proposals for purity and ploidy, returning a solution close to the CloneHD one. For every sample, we collected the proposed alternative solution (tetraploid 2:2 with halved purity) and used its parameters to compute a new Sequenza fit with ploidy ranging 3.8-4.2. Finally, we generated a Seqeunza run constrained with low purity; in total we had 3 runs to compare. Runs for sample Set7_57 from patient Set7 (Figure 6a) highlighted that both Sequenza and CNAqc are strongly confident about the correct solution (diploid with 80% purity). Peak detection scores invariably fail both the tetraploid and low purity solution. The default solution was re-fit following small purity adjustments suggested by CNAqc (Figure 6b) and the final Set7_57 segments showed mild aneuploidy (Figure 6c).

Given that copy number calling is an optimization problem with multiple solutions that often have comparable rank (i.e. they are indeed alternative), this case has shown how CNAqc scores can be used to assess miscalled segments ahead of VAFs (Figure 6d,e). With CNAqc we could rank mutations, CNAs and purity for all samples in patient Set_7 (Figure 6f and Supplementary Figure S22), profiling a tumour consistent with a microsatellite stable colorectal cancer (Cross *et al.*, 2018). An equivalent result was also obtained for 6 WGS samples of patient Set_6 (Supplementary Figure S23).

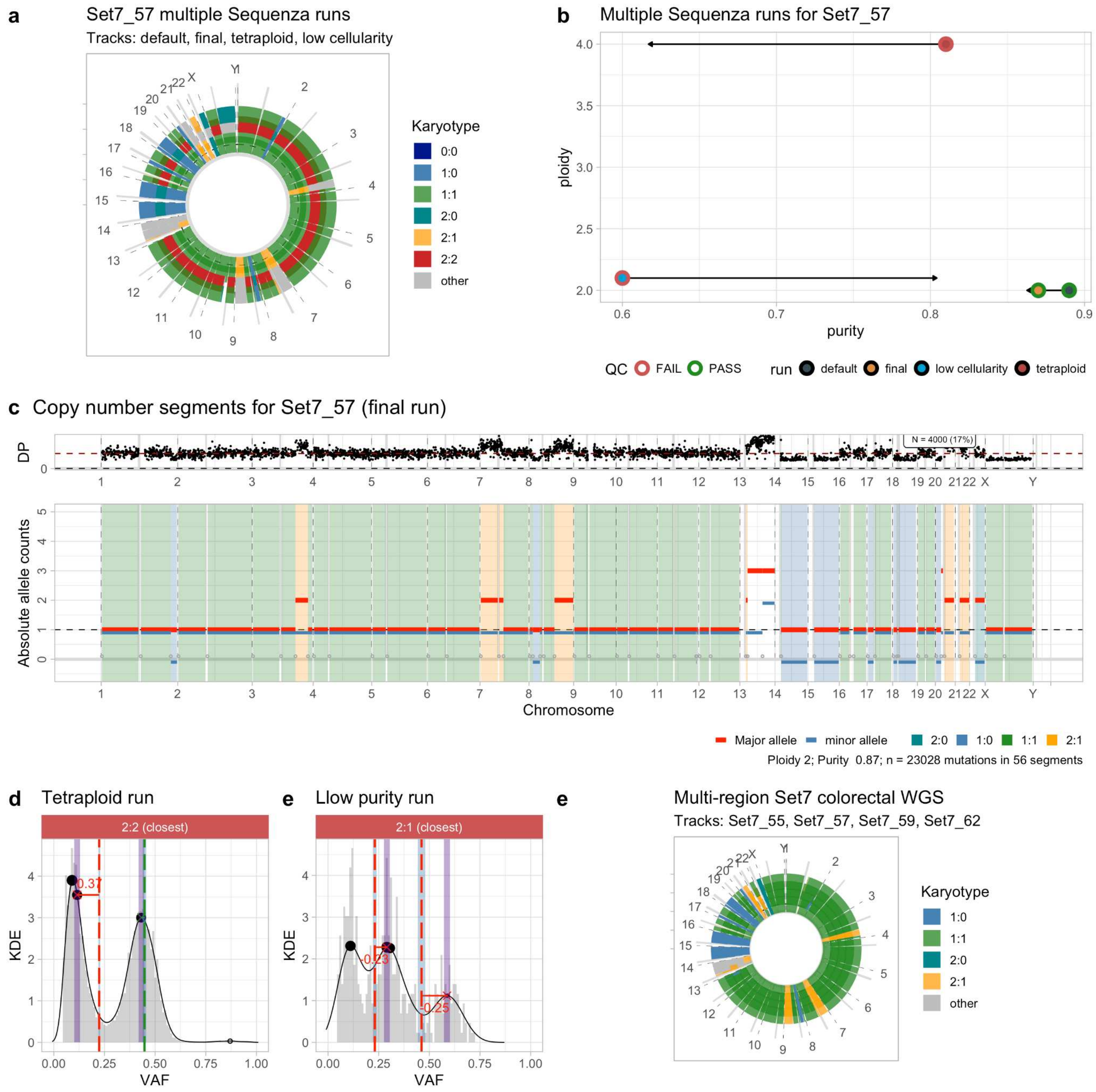

Figure 6. a. Circos plot for four possible whole-genome CNA segmentations determined by Sequenza with WGS data (~80x median coverage, purity 87%). The input sample is Set7_57, one of four multi-region biopsies for colorectal cancer patient Set7. The first run is with default Sequenza parameters. With 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.

Whole exome data

CNAqc is conceptualised to exploit properties of the VAF distribution that are available in high-resolution whole genomes. Lower-resolution whole-exome sequencing (WES) assays produce reduced mutational burden and can be analysed only if VAF quality does not compromise peak detection or multiplicity estimation.

We tested CNAqc with somatic data from WES of $n = 48$ 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; in this test we used data from multiple copy number callers. CNAqc could successfully analyse most of these samples (Supplementary Figure S24); as with the multi-region colorectal cohort (Figure 6), CNAqc could rank calls generated by multiple callers even with WES data. For instance, for sample TCGA-53-7624-01A (Supplementary Figure S25), 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 only ABSOLUTE estimated the correct purity (69%). This shows an interesting scenario of a consensus approach that gets confused by a repeated error across its input callers; in this case our QC tool helps in rescuing the correct solution, despite this being identified by a small subset of the consensus methods.

Discussion

WGS is the most powerful approach to detect 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, and many other institutes are converging towards these efforts. Calling mutations from WGS requires complex bioinformatics pipelines (Barnell *et al.*, 2019; Cmero *et al.*, 2020; Li *et al.*, 2020); generating these data is the prelude to any downstream analysis, putting the quality of the calls under the spotlight at a time in which quality control procedures for somatic data are missing.

CNAqc leverages on statistical properties of VAF distributions to offer the first principled framework to QC somatic mutations, allele-specific CNAs and tumour purity. The tool can analyse nucleotide substitutions for which VAFs can be computed; single-nucleotide variants are more reliable mutation types since they depend less on alignment quality, and should be checked first. CNAqc uses peak detection to validate copy number segments and purity, exploiting a combinatorial model for somatic alleles applied to both clonal and subclonal CNAs. The model distinguishes various types of segments based on clonality (clonal versus subclonal) and complexity (simple versus complex), implementing QC procedures with different features according to the type of segment. All the procedures for QC computation and automatic metric adjustment in CNAqc have been validated using low-pass single-cell copy number data, which we could use to create admixtures of tumour populations with controlled complexity. Moreover, the model has been extensively tested and calibrated with synthetic simulations, and offers the possibility to auto-tune parameters based on coverage and purity of the input sample, and desired performance (from synthetic simulations). Within the same framework, CNAqc also computes CCF values, highlighting mutations whose multiplicity cannot be phased and are therefore uncertain. This can easily be used to create multi-tier CCF calls (certain versus uncertain) for downstream analysis, helping to interpret subclonal clusters found by deconvolution tools.

CNAqc can process both whole-genome and whole-exome data across different callers, and we have identified some example TCGA and PCAWG cases where, with the help of QC metrics, the CNAqc-based analysis improves over consensus-based approaches. While this was based on a few examples, the possibility that QC-based methods could substitute complex consensus approaches would be clearly appealing to make somatic calling pipelines faster and simpler to organise or maintain. CNAqc features can be used to clean up data, automatise parameter choice for any caller, prioritise good calls and select what information should be forwarded for downstream analysis; in this respect, we release with this paper the first automatic copy number calling pipeline that joins Sequenza (Favero *et al.*, 2015) with CNAqc (Supplementary Figure S26). In an effort to integrate QC and copy number calling, this pipeline is designed to iterate calling until a “good” (QC pass) fit is determined, for up to a custom number of attempts (Software Availability). In the future, CNAqc could be extended to other types of CNAs such as extrachromosomal DNA (Verhaak, Bafna and Mischel, 2019; Zeng, Wan and Wu, 2020), whose role in amplifying oncogenes and driving tumour evolution and drug resistance is becoming increasingly important. (Turner *et al.*, 2017; Wu *et al.*, 2019; Kim *et al.*, 2020)

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 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. Similarly, procedures that work at the level of single-segments such as subclonal QC will always depend on the number of mutations mapping on the segment Therefore, for short segments with few mutations, if VAF peaks will be hard to detect, then the QC procedures based on VAF peaks might not be applicable. However, despite obvious data-related limitations, for the large majority of samples CNAqc provides a very effective and efficient way to integrate quality metrics in standard pipelines. Some of the peak-based analysis of CNAqc might be achievable by modifying downstream tools for tumour deconvolution; from our observation however CNAqc is much faster and dedicated to QC, suggesting potentially broader applicability (Supplementary Figure S27).

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; Strino *et al.*, 2013; Miller *et al.*, 2014; Roth *et al.*, 2014; Deshwar *et al.*, 2015; Caravagna *et al.*, 2016, 2018; Jamal-Hanjani *et al.*, 2017; Cross *et al.*, 2018; Turajlic *et al.*, 2018; Gerstung *et al.*, 2020). 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 to automatically control the quality of cancer sequencing assays.

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; we used the following files:

- Somatic consensus SNVs and indels
    - https://dcc.icgc.org/releases/current/Summary#:~:text=simple_somatic_mutation.aggregated.vcf.

- Somatic allele-specific CNAs
    - https://dcc.icgc.org/api/v1/download?fn=/PCAWG/consensus_cnv/consensus.20170119.somatic.cna.annotated.tar.gz

- Purity and ploidy cohort table
    - https://dcc.icgc.org/api/v1/download?fn=/PCAWG/consensus_cnv/consensus.20170217.purity.ploidy.txt.gz

Results from our analysis of PCAWG will be made available at Zenodo before the paper is published.

TCGA calls are publicly available at the GDC Data Portal (https://portal.gdc.cancer.gov),

Software Availability

CNAqc is implemented as an open source R package 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.

Single-cell data and code are also released alongside the tool at the main webpage.

Authors contribution

GC, WC and JH conceived the method. GC formalised the model and implemented the tool; RB developed the error model; AA and GC developed the model for subclonal copy numbers. MW analysed single-cell data for validation. RB carried out simulations; GC, RB and SM carried out model calibration; NC and GC carried out model validation; SM carried out comparisons against other methods. SM, AA and GC analysed PCAWG data. JH and GC analysed whole-exome data. GC and WC drafted the manuscript, which all authors approved in final form.

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, D., Sirota, M. and Butte, A.J. (2016) ‘Corrigendum: Systematic pan-cancer analysis of tumour purity’, *Nature communications*, 7, p. 10707.

Bailey, M.H. *et al.* (2018) ‘Comprehensive Characterization of Cancer Driver Genes and Mutations’, *Cell*, 173(2), pp. 371–385.e18.

Barnell, E.K. *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), pp. 972–981.

Boeva, V. *et al.* (2011) ‘Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization’, *Bioinformatics* , 27(2), pp. 268–269.

Cancer Genome Atlas Research Network (2014) ‘Comprehensive molecular profiling of lung adenocarcinoma’, *Nature*, 511(7511), pp. 543–550.

Caravagna, G. *et al.* (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), pp. E4025–34.

Caravagna, G. *et al.* (2018) ‘Detecting repeated cancer evolution from multi-region tumor sequencing data’, *Nature methods*, 15(9), pp. 707–714.

Caravagna, G. *et al.* (2020) ‘Subclonal reconstruction of tumors by using machine learning and population genetics’, *Nature genetics*, 52(9), pp. 898–907.

Carter, S.L. *et al.* (2012) ‘Absolute quantification of somatic DNA alterations in human cancer’, *Nature biotechnology*, 30(5), pp. 413–421.

Cmero, M. *et al.* (2020) ‘Inferring structural variant cancer cell fraction’, *Nature communications*, 11(1), p. 730.

Cross, W. *et al.* (2018) ‘The evolutionary landscape of colorectal tumorigenesis’, *Nature ecology*

& evolution, 2(10), pp. 1661–1672.

Cun, Y. *et al.* (2018) ‘Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust’, *Nature protocols*, 13(6), pp. 1488–1501.

Dentro, S.C., Wedge, D.C. and Van Loo, P. (2017) ‘Principles of Reconstructing the Subclonal Architecture of Cancers’, *Cold Spring Harbor perspectives in medicine*, 7(8). doi:10.1101/cshperspect.a026625.

Deshwar, A.G. *et al.* (2015) ‘PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors’, *Genome biology*, 16, p. 35.

Ding, L. *et al.* (2012) ‘Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing’, *Nature*, 481(7382), pp. 506–510.

Favero, F. *et al.* (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), pp. 64–70.

Fischer, A. *et al.* (2014) ‘High-definition reconstruction of clonal composition in cancer’, *Cell reports*, 7(5), pp. 1740–1752.

Gerstung, M. *et al.* (2020) ‘The evolutionary history of 2,658 cancers’, *Nature*, 578(7793), pp. 122–128.

Gonzalez-Perez, A. *et al.* (2013) ‘IntOGen-mutations identifies cancer drivers across tumor types’, *Nature methods*, 10(11), pp. 1081–1082.

Greaves, M. and Maley, C.C. (2012) ‘Clonal evolution in cancer’, *Nature*, 481(7381), pp. 306–313.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) ‘Pan-cancer analysis of whole genomes’, *Nature*, 578(7793), pp. 82–93.

Jamal-Hanjani, M. *et al.* (2017) ‘Tracking the Evolution of Non-Small-Cell Lung Cancer’, *The New England journal of medicine*, 376(22), pp. 2109–2121.

Kent, D.G. and Green, A.R. (2017) ‘Order Matters: The Order of Somatic Mutations Influences Cancer Evolution’, *Cold Spring Harbor perspectives in medicine*, 7(4). doi:10.1101/cshperspect.a027060.

Kim, H. *et al.* (2020) ‘Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers’, *Nature genetics*, 52(9), pp. 891–897.

Laks, E. *et al.* (2019) ‘Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing’, *Cell*, 179(5), pp. 1207–1221.e22.

Landau, D.A. *et al.* (2013) ‘Evolution and impact of subclonal mutations in chronic lymphocytic leukemia’, *Cell*, 152(4), pp. 714–726.

Levine, A.J., Jenkins, N.A. and Copeland, N.G. (2019) ‘The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters', *Cancer cell*, 35(1), pp. 10–15.

Li, Y. *et al.* (2020) 'Patterns of somatic structural variation in human cancer genomes', *Nature*, 578(7793), pp. 112–121.

McGranahan, N. and Swanton, C. (2015) 'Biological and therapeutic impact of intratumor heterogeneity in cancer evolution', *Cancer cell*, 27(1), pp. 15–26.

McGranahan, N. and Swanton, C. (2017) 'Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future', *Cell*, 168(4), pp. 613–628.

Miller, C.A. *et al.* (2014) 'SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution', *PLoS computational biology*, 10(8), p. e1003665.

Nik-Zainal, S. *et al.* (2012) 'The life history of 21 breast cancers', *Cell*, 149(5), pp. 994–1007.

Poell, J.B. *et al.* (2019) 'ACE: absolute copy number estimation from low-coverage whole-genome sequencing data', *Bioinformatics* , 35(16), pp. 2847–2849.

Priestley, P. *et al.* (2019) 'Pan-cancer whole-genome analyses of metastatic solid tumours', *Nature*, 575(7781), pp. 210–216.

Roth, A. *et al.* (2014) 'PyClone: statistical inference of clonal population structure in cancer', *Nature methods*, 11(4), pp. 396–398.

Strino, F. *et al.* (2013) 'TrAp: a tree approach for fingerprinting subclonal tumor composition', *Nucleic acids research*, 41(17), p. e165.

Turajlic, S. *et al.* (2018) 'Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal', *Cell*, 173(3), pp. 595–610.e11.

Turnbull, C. *et al.* (2018) 'The 100 000 Genomes Project: bringing whole genome sequencing to the NHS', *BMJ* , 361, p. k1687.

Turner, K.M. *et al.* (2017) 'Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity', *Nature*, 543(7643), pp. 122–125.

Van Loo, P. *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), pp. 16910–16915.

Verhaak, R.G.W., Bafna, V. and Mischel, P.S. (2019) 'Extrachromosomal oncogene amplification in tumour pathogenesis and evolution', *Nature reviews. Cancer*, 19(5), pp. 283–288.

Watkins, T.B.K. *et al.* (2020) 'Pervasive chromosomal instability and karyotype order in tumour evolution', *Nature*, 587(7832), pp. 126–132.

Williams, M.J. *et al.* (2021) 'Evolutionary tracking of cancer haplotypes at single-cell resolution', *bioRxiv*. doi:10.1101/2021.06.04.447031.

Wu, S. *et al.* (2019) 'Circular ecDNA promotes accessible chromatin and high oncogene expression', *Nature*, 575(7784), pp. 699–703.

Xu, C. (2018) 'A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data', *Computational and structural biotechnology journal*, 16, pp. 15–24.

Yoshihara, K. *et al.* (2013) 'Inferring tumour purity and stromal and immune cell admixture from expression data', *Nature communications*, 4, p. 2612.

Yuan, K. *et al.* (2018) 'Ccube: A fast and robust method for estimating cancer cell fractions', *biorXiv*. doi:10.1101/484402.

Zaccaria, S. and Raphael, B.J. (2021) 'Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL', *Nature biotechnology*, 39(2), pp. 207–214.

Zack, T.I. *et al.* (2013) 'Pan-cancer patterns of somatic copy number alteration', *Nature genetics*, 45(10), pp. 1134–1140.

Zeng, X., Wan, M. and Wu, J. (2020) 'ecDNA within tumors: a new mechanism that drives tumor heterogeneity and drug resistance', *Signal transduction and targeted therapy*, p. 277.