A fully automated approach for quality control of cancer mutations in the era of high-resolution whole genome sequencing

Jacob Househam, *Barts Cancer Institute, Queen Mary University of London, UK*
William CH Cross [(*)], *UCL Cancer Institute, University College London, UK*
Giulio Caravagna [(*)], *Department of Mathematics and Geosciences, University of Trieste, Italy*

[(*)] 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 point mutations, copy numbers, and other metrics into a single quantitative score. We demonstrate its power on 1,065 whole-genomes from a large-scale pan-cancer cohort, and on multi-region data of two colorectal cancer patients. We highlight how our approach significantly improves the generation of cancer mutation data, providing visualisations for cross-referencing with other analyses. Our approach is fully automated, designed to work downstream of 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 for both (Barnell et al. 2019). Using bioinformatic tools to cross reference the normal genome against the 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 (Gerstung et al. 2020; Nik-Zainal et al. 2012; Caravagna et al. 2020), which informs us how a particular tumour developed from normal cells.

There are several types of mutations that we can retrieve from DNA sequencing data (Campbell et al. 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). All types of mutations can drive tumour progression, and are therefore important entities to study (Kent and Green 2017-4; Levine, Jenkins, and Copeland 2019). Luckily, the steady drop in sequencing costs is fueling the creation of large amounts of data, which are becoming increasingly available 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 major improvements over whole-exome 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 from the data. 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. 11 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% - if the SNV occurred after the amplification - or 66% - if the SNV is on the amplified chromosome and occurred before the amplification. The theoretical

frequencies are observed with a Binomial noise model that depends on the depth of sequencing and the actual VAF (Nik-Zainal et al. 2012; Caravagna et al. 2020). 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 (Nik-Zainal et al. 2012).

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 in parallel (Boeva et al. 2011; Cmero et al. 2020; Zaccaria and Raphael 2020; Van Loo et al. 2010). 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., mutation allele ratios, 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 datasets with millions of data points (Campbell et al. 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 (Data Availability), a computational framework with a de novo statistical model to assess the conformance of expected SNVs, CNAs, and purity estimates. We strived to make the tool as simple to implement as possible, maximising compatibility across differing pipelines. CNAqc computes a quantitative quality check (QC) score for the overall agreement of the calls, which can be used to tune the parameters of callers (e.g., decrease purity or increase ploidy), or select among multiple CNA profiles (e.g., tetraploid versus diploid tumours) until a fit is achieved. In CNAqc we also integrate these measures to determine CCF values (Dentro, Wedge, and Van Loo 2017).

CNAqc is implemented as a highly optimised R package that can be used downstream of any cancer mutation calling pipeline. It can be run on WGS data, and can automatically compute a QC score in a matter of seconds, which is an extremely useful

feature for large-scale genomics consortia that analyse many samples per day. To demonstrate the tool we analysed 11 bulk WGS datasets from two multi-region colorectal cancers, and analysed 1065 high-quality whole-genomes from the Pan Cancer Analysis of Whole Genomes (PCAWG) cohort (Campbell et al. 2020).

Results

The CNAqc framework

CNAqc can perform different types of operations on CNAs and somatic mutation calls obtained from bulk WGS. In what follows, we will refer explicitly to SNVs as the main type of mutation used, but in principle other types of substitutions such as insertions or deletions also apply. The package supports the most common CNA copy types found in cancers: 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. The tool also works with exome data, but the reduced mutational burden can, in general, lower the reliability of the QC score (Supplementary Figure S1).

Many metrics output by CNAqc are derived from the link between copy-state profiles (i.e., the copies of the major and minor alleles, which sum up to the ploidy of a segment) and allele frequencies that are explicit from read counts. Combinatorial equations and frequency spectrum analysis can quantitatively determine if CNAs and purity are consistent with the VAF distribution (Online methods). This score also suggests “corrections” to automatically fine-tune and repeat CNA calling runs. This works for tools that use either Bayesian priors or point estimates of the parameters.

The key equations for a somatic mutation link its VAF v and CCF c, to sample purity π, tumour ploidy p, and $m \in \{1, 2\}$, the number of copies of a mutation (Figure 1a). Effectively, for complex 2:0, 2:1 and 2:2 copy states, m phases mutations that were acquired before or after the copy number event (Figure 1b). We remark that we observe v, and infer π, p and m, finally deriving c, which is difficult to estimate (Figure 1c).

In CNAqc we use the following formula for VAF (Figure 1d)

$$v = \frac{\pi}{2(1-\pi) + \pi p}$$

and CCF

$$c = \frac{v[(p-2)\pi + 2]}{m\pi}\,.$$

These formulas lead to other interesting quantities (Online methods). For instance, if we know tumour purity and the ploidy of a CNA segment, then the VAF mutations mapped to the segment must peak at a known location x. The value for x follows from combinatorial arguments relating all other variables (Nik-Zainal et al., 2012). From a QC perspective, the euclidean distance between the theoretical expectation x and the peaks observed from data is an error score that approaches 0 for perfect calls, and grows otherwise. CNAqc can visualise the input segments (Figure 2a) and read counts (Figure 2b-d). Other analysis such as CCFs computation and genome fragmentation analysis are also available, and have other visualisations (Figure 2e).

The scores of CNAqc can be used to determine a QC PASS or FAIL status for every copy state within a tumour genome, weighting different evidence from the data. One score is for the quality of CNA segmentation and tumour purity, and one for CCF values. The former is based on a density-based analysis of the VAF distribution, and uses both a non-parametric kernel density and a univariate Binomial mixture to match peaks in the VAF data (Figure 3a-d). The latter is based on the entropy of the latent variables in a Binomial mixture model, whose components are peaked at the expected VAF. From this density we identify VAF ranges for which it is hard to estimate the mutation multiplicity, and therefore the CCF of the mutation (Figure 3e-h). To the best of our understanding, this is the only framework providing quantitative metrics for all the most widespread types of tumour mutations.

Multi-region colorectal cancer data

We have run CNAqc on previously published WGS multi-region data (Cross et al. 10 2018; Caravagna et al. 2020), which was collected from multiple regions of primary colorectal adenocarcinomas across two distinct patients. For all these samples we have high quality somatic mutation calls (Cross et al. 10 2018) that were obtained using CloneHD (Fischer et al. 2014). We have re-called CNAs with the Sequenza CNA caller (Favero et al. 2015), and sought out to check the inferred copy states and tumour purity with CNAqc, along with SNVs generated by Mutect2 (Benjamin et al. 2019).

Sequenza was run using distinct parameterizations. We begun with the default range proposals for purity and ploidy[1], which we then improved in a final run following CNAqc analysis. We also forced a Sequenza fit with constrained tetraploid genome (ploidy equal 4), and one with low purity. All these steps could have been easily automatised in a procedure that runs the caller, obtains score metrics for the solution from CNAqc, and re-run the fits with adjusted parameters if required. The results for one sample of patient Set7 - Cancer 7 in the original manuscript (Cross et al. 10 2018) - are in Figure 4; the other samples for patient Set7 are in Supplementary Figures S2-S4. All samples for patient Set6 are in Supplementary Figures S5-S10.

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 just default parameters (Figure 4b). The whole-genome CNA profile for this sample shows some degree of aneuploidy (Figure 4c), and it is easy with CNAqc to assess miscalled CNA segments ahead of the VAF data (Figure 4d). The analysis of all the samples available for Set7 shows an overall CNA profile with many diploid regions and mild aneuploidy (Figure 4e), consistent with a microsatellite stable colorectal cancer (Cross et al. 10 2018).

Large-scale pan cancer PCAWG calls

We have run CNAqc on a subset of the full PCAWG cohort, which contains thousands of samples from multiple tumour types (Campbell et al. 2020). The median coverage of this cohort is 45x, with purity ~65% (Caravagna et al. 2020); a much lower resolution than the data available for the multi-region samples discussed in the previous section.

Because of this, peak detection from the VAF distribution across some of the samples would be challenged by signal quality; in practice, for genomes with complex aneuploidy and massive drops in purity and coverage the VAF distribution is unsuitable for peak-detection, leading to false-positives in the QC process. To avoid this and work with suitable samples, we identified $n = 1065$ cases adopting the following conditions: (i) the tumour type contains >20 samples, (ii) the tumour genome used for QC contains >30% of the overall SNVs in the tumour - so a substantial part of the overall mutational burden - and (iii) the purity of the sample is >60% - so the signal is suitable for peak detection. On a standard cluster CNAqc ran in less than 1 hour for these samples; notably the

[1] Technically the default Sequenza values for ploidy reach maximum at 7; being unrealistic for our cases we limited the maximum ploidy to be 5.

completion time (per sample) on a laptop is less than 1 minute, meaning that preliminary analysis can be carried out very quickly and without large computing infrastructures.

The calls in PCAWG were obtained by consensus with multiple bioinformatics tools, and for this reason we expected them to be reliable. Manual inspections of some patient data showed indeed many high-quality calls, but also highlighted a variety of interesting cases. For instance, tumours with extremely low mutational burden but high quality calls still yielded a useful report, suggesting that CNAqc can work also with mutational burden from whole-exome sequencing (Supplementary Figure S1). For other tumours, we found high purity levels >90%, which are probably overestimated (Supplementary Figure S11) compared to others where purity is genuinely very high (Supplementary Figure S12). Overall, the scores from peak detection are reliable for the majority of the analysed samples (Figure 5a) - the diploid 85% purity tumour in Figures 2 and 3 is taken from this list - with only a few cases requiring further checks (Figure 5b). The peak detection by CNAqc therefore confirms the calls reliability in terms of breakpoints, segments ploidy and tumour purity.

CCF computations showed a higher rate of failures with CNAqc analysis (Figure 5a). This is inevitably due to the lack of signal separability stemming from low coverage of these samples, even for high-quality genomes. Therefore while peaks could be determined for these data, mutation multiplicity assessment would have required higher coverage than what was found available.

In summary, from these analyses we revealed that the problem of validating CNA calls, compared to determining CCF estimates, can be approached with lower coverage and purity values using CNAqc.

Discussion

WGS is a powerful approach to detect extensive mutations that drive human cancers. Many large-scale initiatives such as PCAWG (Campbell et al. 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 offers the first principle framework to control the quality of tumour mutation calls. The tool can analyse SNVs and more general types of nucleotide substitutions; SNVs are more reliable and depend less on alignment quality than other mutations, and therefore should be checked first. CNAqc uses a peak-detection analysis to validate CNA segments and purity, exploiting a combinatorial model for cancer alleles. Within the same framework, CNAqc also computes CCF values, highlighting mutations for which such values are uncertain. CNAqc features can be used to clean up data, automatising parameter choice for virtually 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, e.g., 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 fast way to integrate quality metrics in standard pipelines.

Generating high quality calls is just a prelude 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. 07 12, 2016; Jamal-Hanjani et al. 2017; Turajlic et al. 2018; Caravagna et al. 09 2018). 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 sequencing efforts, 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 WGS assays. We believe CNAqc can help reduce the burden of manual quality checking and parameter tuning.

References

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.

Benjamin, David, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. 2019. "Calling Somatic SNVs and Indels with Mutect2." *bioRxiv*, December, 861054. https://doi.org/10.1101/861054.

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.

Campbell, Peter J., Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, et al. 2020. "Pan-Cancer Analysis of Whole Genomes." *Nature* 578 (7793): 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Caravagna, Giulio, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A. Graham, Guido Sanguinetti, and Andrea Sottoriva. 09 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. 07 12, 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.

Cmero, Marek, Ke Yuan, Cheng Soon Ong, Jan Schröder, Niall M. Corcoran, Tony Papenfuss, Christopher M. Hovens, Florian Markowetz, and Geoff Macintyre. 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. 10 2018. "The Evolutionary Landscape of Colorectal Tumorigenesis." *Nature Ecology & Evolution* 2 (10): 1661–72. https://doi.org/10.1038/s41559-018-0642-z.

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.

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.

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.

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.

Kent, David G., and Anthony R. Green. 2017-4. “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.

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. “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.

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.

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.

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.

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.

Watkins, Thomas B. K., Emilia L. Lim, Marina Petkovic, Sergi Elizalde, Nicolai J. Birkbak, Gareth A. Wilson, David A. Moore, et al. 11 2020. "Pervasive Chromosomal Instability and Karyotype Order in Tumour Evolution." *Nature* 587 (7832): 126–32. https://doi.org/10.1038/s41586-020-2698-6.

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.

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.

CNAqc is implemented as an open source R package that is hosted at the GitHub space of the Caravagna Lab

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

The tool webpage contains RMarkdown tutorial vignettes to run CNAqc analysis of a generic dataset, as well as documents that explain visualisation and parameterizations of the execution. All analyses in this paper can be replicated following the vignettes.

Authors contribution

All authors conceived the method, which GC formalised and implemented. All authors analysed the data and wrote the manuscript.

Competing interests.

The authors declare no competing interests.

Online methods

CNAqc supports two human genome references (GRCh38 and hg19), and the most common CNA profiles found in cancers:

- heterozygous diploid states (1:1)[2];
- 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.

We make a simplifying assumption, whereby CNAs have been acquired in one step, starting from a simple heterozygous diploid state (the germline). For this reason, for tetraploid segments we only consider copy state 2:2, instead of 3:1 or 4:0. This allows us to make simpler computations. In practice, we avoid working with copy states for which the computation of CCFs is very difficult, and that are quite unlikely to be observed in real data. Also, we consider only clonal CNA segments. While subclonal CNA segments are certainly important for cancer genomics, the calls that we seek to quality check regard just clonal CNA events; being the one most prevalent in the majority of cancer cells, they have to be prioritised, with subclonal CNAs being only reliable for tumours with good clonal CNA calls.

CNAqc works primarily with Whole-Genome Sequencing (WGS) data. For exome data, the reduced exonic mutation burden can make it more difficult to work with the spectrum of the VAF distribution. In general, the key determinant to detect peaks in the VAF, is the number of mutations per copy state. For tumours with strong endogenous mutant factors (e.g., smoking) or very high mutation rate (e.g., microsatellite unstable tumours), the number of exonic mutations could be high enough to use CNAqc.

Peak-detection QC

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

We consider a somatic mutation present in m copies of the tumour genome, when the sample purity is π and the segment ploidy is p. Note that p can be computed summing the total number of copies of the minor and major allele at the mutation locus (Figure 1).

The key equations for the expected VAF of a clonal mutation and its CCF are presented in the Main Text. Here we discuss how peaks can be used to QC both tumour purity and CNA segments and, consequently, overall tumour ploidy. From a QC perspective, if we solve for m and π the equations, we can get v as

$$v = \frac{v[(p-2)\pi + 2]}{m\pi}$$

which means that if we know tumour purity and CNA, we expect a peak at VAF v, for a given value of m, in the data distribution (Figure 1a and 1b). For instance, for a 1:1 segment ($p = 2$), the expected VAF for a heterozygous clonal ($m = 1$) mutation is 25% for a 50%-purity tumour, and 50% for a 100%-purity tumour. Similarly, for a 2:2 genome ($p = 4$) of a tumour with 75% purity, the expected VAF for clonal mutations accrued before genome doubling and therefore visible in two copies ($m = 2$) is ~54%, while for those accrued after genome doubling, and therefore present in single copy ($m = 1$), we expect a ~21% VAF (Dentro, Wedge, and Van Loo 2017). CNAqc checks the data for peaks at these VAFs, with a tolerance $\epsilon > 0$. From the distance between the theoretical expectation and the estimator derived from data, we obtain an error metric for the calls.

CNAqc first performs peak detection from the input VAF with two, separate, methods:

1. Via a kernel density estimation with fixed bandwidth, which is used to determine a smooth density profile. Peaks are then estimated from the discretized smooth, using specialised R packages for peak-detection and removing peaks with density below a parameterized cutoff.
2. Via Binomial mixture from the BMix (Caravagna et al. 2020) package (https://caravagn.github.io/BMix/), a peak is associated with each Binomial probability, for all mixture components .

Peaks are matched to the expected theoretical values based on their euclidean distance. A theoretical peak can be matched to the closest peak in the data, or the one to the most right side of the frequency spectrum. This latter strategy works only if there are no miscalled CNAs. The first strategy (closest match), is the default CNAqc choice.

For every peak a QC value (PASS or FAIL) is determined based on some tolerance $\varepsilon > 0$. The overall QC status of copy states with multiple peaks is the QC of the peak with most mutations underneath. The overall QC status for a sample with many copy states is determined by summing up the QC status of individual copy states, and weighting them by the number of mutations associated (majority rule).

CCF estimation

CNAqc can compute CCFs in two ways. One of the two uses the idea of the mixture highlighted in Figure 1c, the other is simpler and works better when data resolution is low, and the entropy of the mixture model would leave too many mutations unassigned.

For the mixture approach, we build a 2-components Binomial mixture from the theoretical expectations and the data. This implicitly assumes that peaks have been QCed first. We constraint the success parameters to match the expected VAF, and use the proportion of mutations that appear underneath a peak as mixing proportions π. Then, from the latent variables z of the model we compute the probability of assigning a mutation with VAF x_n to cluster c,

$$p(z_{n,k} = c \mid \theta, \pi).$$

From this information we obtain the entropy $H(z)$ of z, which is low for values that are assignable to only one cluster. Recall in this respect that the maximum entropy distribution is the uniform one, which is when a mutation can be equally likely in 1 or 2 copies, based on VAF.

We use a simple peak detection heuristic to find 2 points of changes in $H(z)$; in between those values we cannot reliably assess m, i.e. assess if the mutation is in single or double copy. For these CNAqc leaves the CCF value as NA.

The alternative approach uses a simpler idea, still working on the expected theoretical VAF. Here instead of fitting a mixture we determine the midpoint o, between the two expected theoretical VAF peaks. The midpoint is computed by weighting each of the two peaks proportionally to the number of mutations that appear underneath each peak. The midpoint is a cut: values below o are in single copy, values above in two. This procedure requires data with good sequencing coverage, and a good general quality.

When mutation multiplicities have been determined, CCF computation is trivial, and follows the formula presented in the Main Text. A QC PASS status is assigned to the

CCF values for a copy state, if less than 10% (or any custom threshold) are unassigned. The overall sample is given a QC status based on a majority policy.

Genome fragmentation

Some recently identified patterns of somatic CNA changes 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 bioinformatics tools, CNAqc offers a simple statistical test to detect the presence of over-fragmentation in a chromosome arm, a prerequisite that could point to the presence of such patterns.

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., $\mu = 0.2$). Then, a null hypothesis is used to compute a p-value. That is defined using a Binomial test based on k, the number of trials given by the total segment counts 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, a one-tailed test for whether the observations are biased towards shorter 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 obtain that two contiguous segments are merged if they have exactly the same ploidy profile (i.e. same numbers for the major and minor

alleles), and if they are a maximum distance apart (e.g. 1 megabase). 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.

Main Text Figures

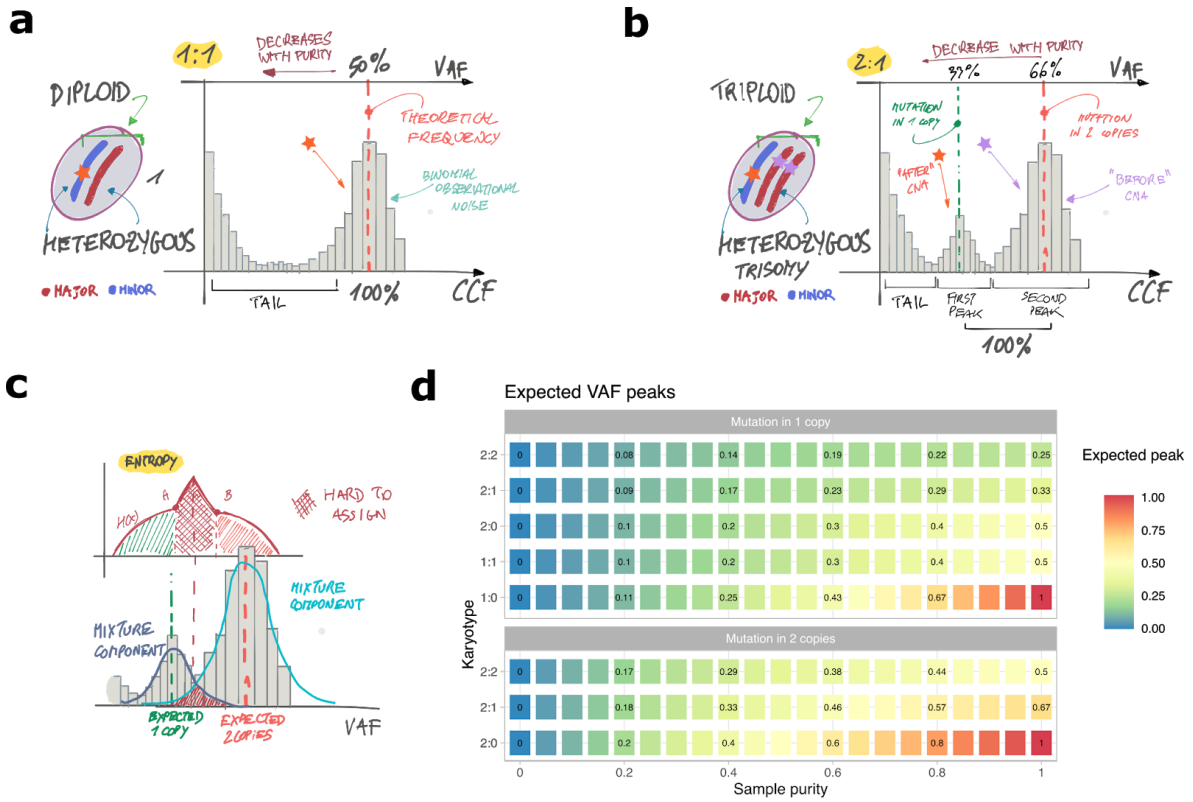


Figure 1. a. Theoretical VAF histogram for diploid 1:1 mutations in a tumour. A clonal heterozygous mutation has 50% VAF; all mutations are observed with some Binomial sequencing noise. The clonal mutations form a peak at 100% CCF, plus other features that characterise the tumour clonal composition (e.g., the tail). The expected theoretical VAF decreases if sample purity reduces. **b.** The case of a 2:1 tumour genome, where we expect 2 peaks in the VAF originating from mutations present in one (orange) or two copies (purple). 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 clonal mutations (100% CCF). **c.** Computing CCFs requires caution for mutations with different multiplicities; we support 2:0, 2:1 and 2:2 copy states in CNAqc, and offer two methods to compute CCFs. The one depicted is based on the entropy of a Binomial mixture. From the expected VAF peaks we construct a mixture density and use the entropy of its latent variables to capture uncertainty in the multiplicities. At the crossing of the components we cannot easily assign multiplicities, and therefore CCFs; the entropy peaks at the top of the uncertainty by definition. **d.** Heatmap expressing the relationship between copy states, mutation multiplicity and sample purity. The color reflects the expected VAF for the corresponding mutations, and can be used to QC both CNAs and purity estimates.

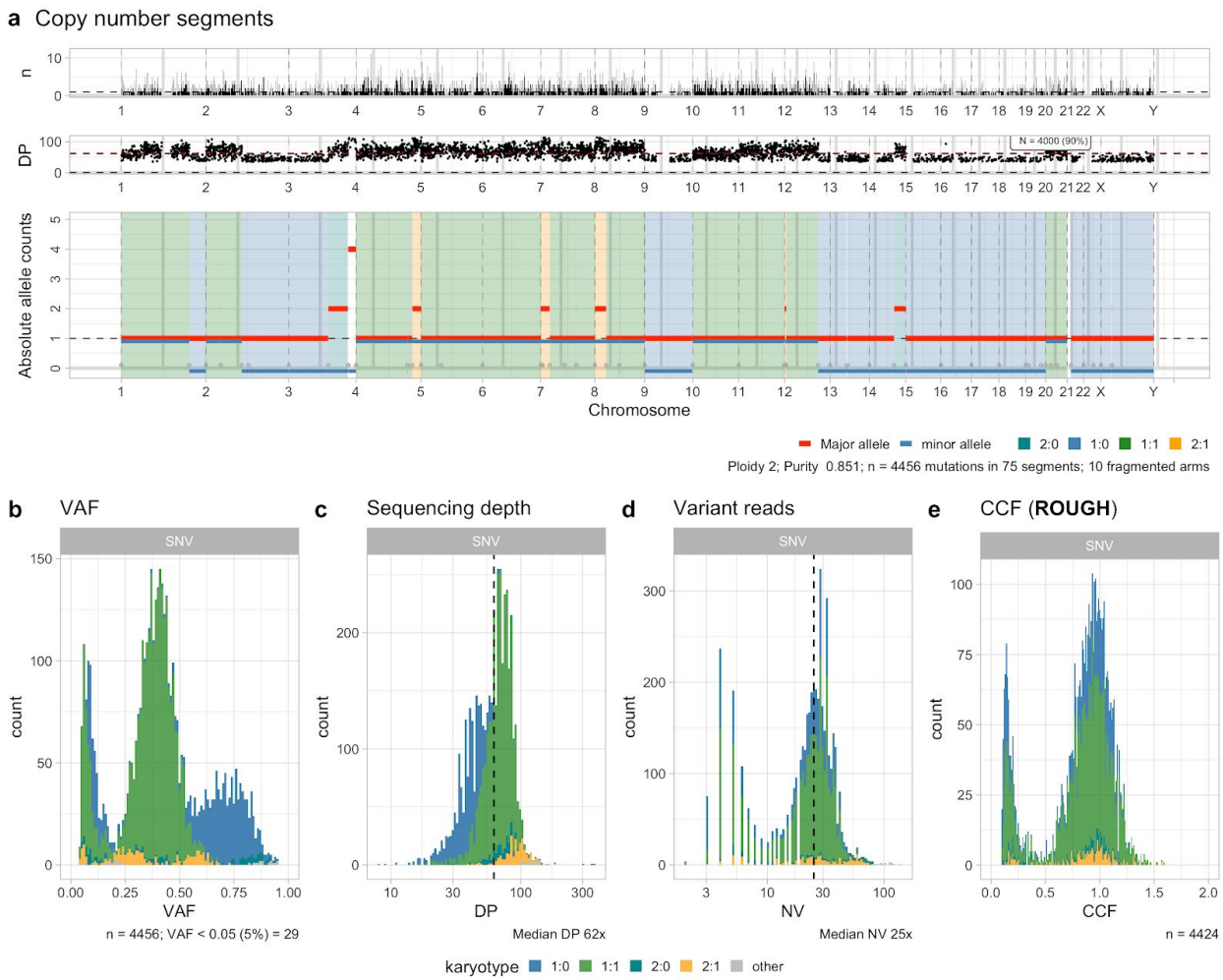


Figure 2. a. Genome-wide total clonal copy number segments for a PCAWG cancer sample with overall ploidy 2, and sample purity ~85%. The panel is composed of three illustrations. The bottom plot reports the copies of the major and minor alleles in each segment, and some genome areas are shaded. The central plot shows genome-wide somatic mutations with their depth of sequencing, and the top plot shows the total number of mappable mutations binned every megabase. **b.** Variant Allele Frequencies (VAFs) for the mutations that map to the input segments (note that these are all SNVs). **c.** Depth of sequencing (DP) for every SNV. **d.** Number of reads (NV) with the variant allele for every SNV. **e.** Cancer Cell Fractions (CCF) estimation for this sample, obtained from CNAqc.

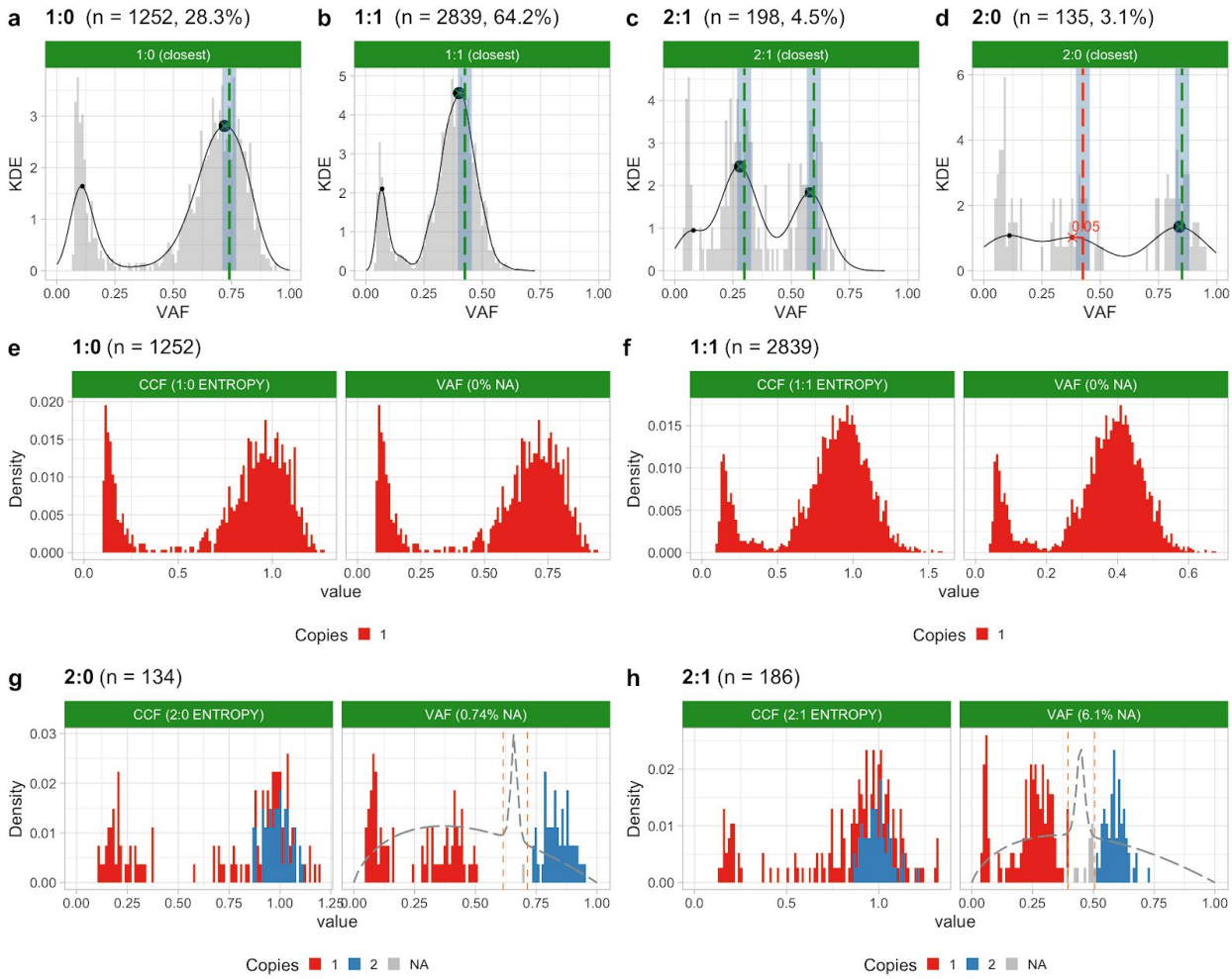


Figure 3. a-d. Peak detection analysis assessing the quality of CNA segments (split by copy state), and tumour purity. The shaded gray area are input mutations, and the thin black profile is its kernel density estimation (KDE). The black circles represent the peaks detected from the KDE, and the vertical dashed lines are the expected peaks, given the tumour purity. If the data peaks fall within the shaded area surrounding the vertical line, the estimates are consistent and the plot is therefore green (QC pass). For copy states with total copy number >2, multiple peaks are checked independently. In that case the overall QC status for the copy state is a linear combination of the results, weighted by the number of mutations assignable to each peak. **e-h.** Cancer Cell Fractions (CCF) estimation for each tumour genome, using the entropy method. Each plot shows both CCF, and the VAF from which mutation multiplicities are computed. In the rightmost panel we overlay the entropy profile computed by a 2-dimensional Binomial mixture. Areas within the red vertical dashed lines are those for which CNAqc cannot assign a confident CCF value. For copy states 1:0 and 1:1 the mutation multiplicity is fixed to 1 by definition.

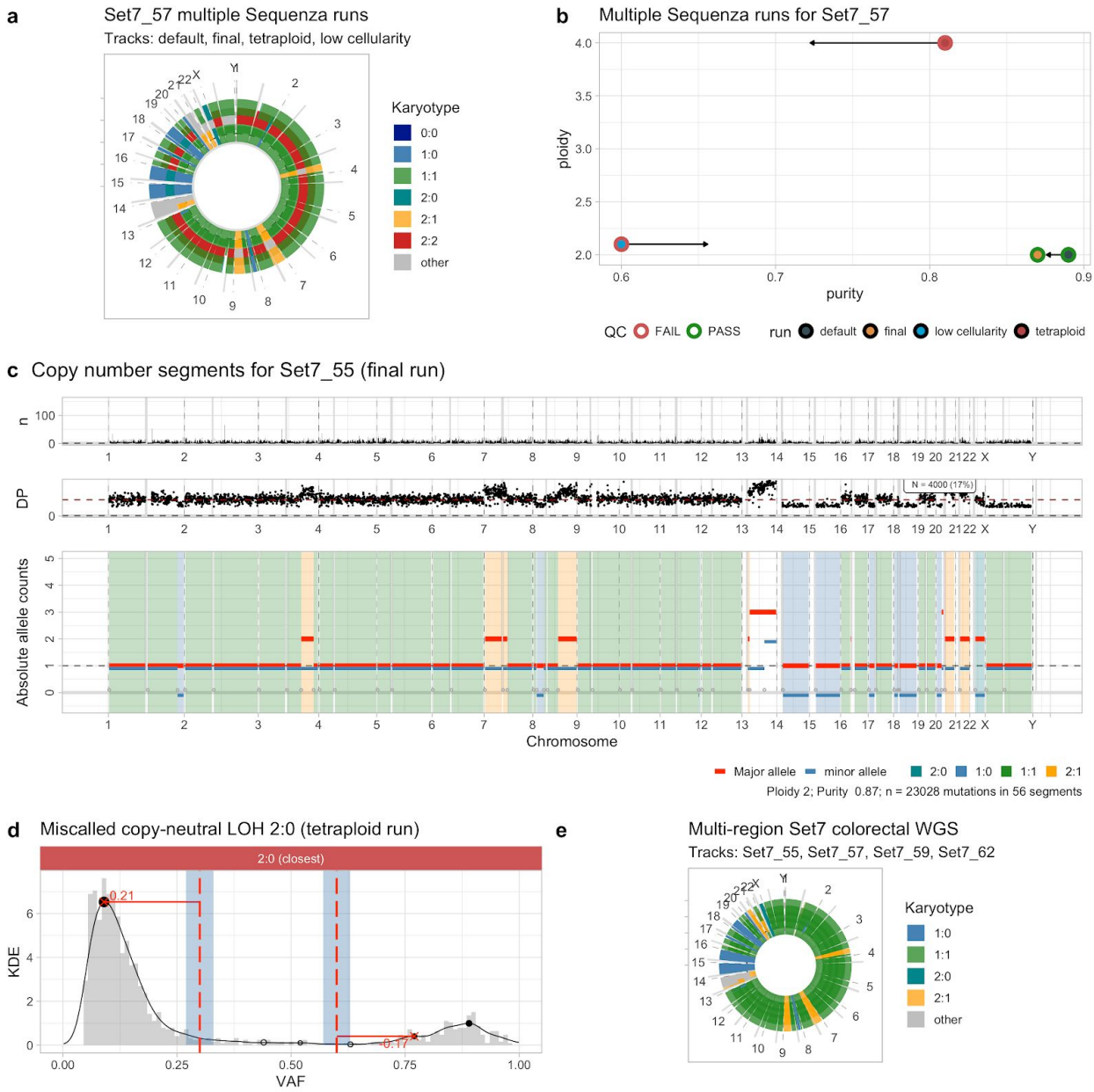


Figure 4. a. Circos plot for four possible whole-genome CNA segmentations determined by Sequenza with WGS data (~80x median coverage, purity 87%). The input sample is Set7_57, one of four multi-region biopsies for colorectal cancer patient Set7. The first run is with default Sequenza parameters. With CNAqc, we slightly adjust purity estimation and obtain a final run of the tool. We also one run forcing overall tumour ploidy to 4 (tetraploid), 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 QC. **c.** Final run with perfect results for Set7_57: copy number segments, depth of coverage per mutation and mutation density per megabase. **d.** Miscalled copy-neutral LOH segment, obtained by forcing a tetraploid solution in Sequenza. For a 2:0 segment with the estimated Sequenza

purity we expected peaks at ~60% and ~30% VAF, which cannot be matched. **e.** CNA calling with CNAqc and Sequenza for 4 WGS biopsies of the primary colorectal cancer Set7.

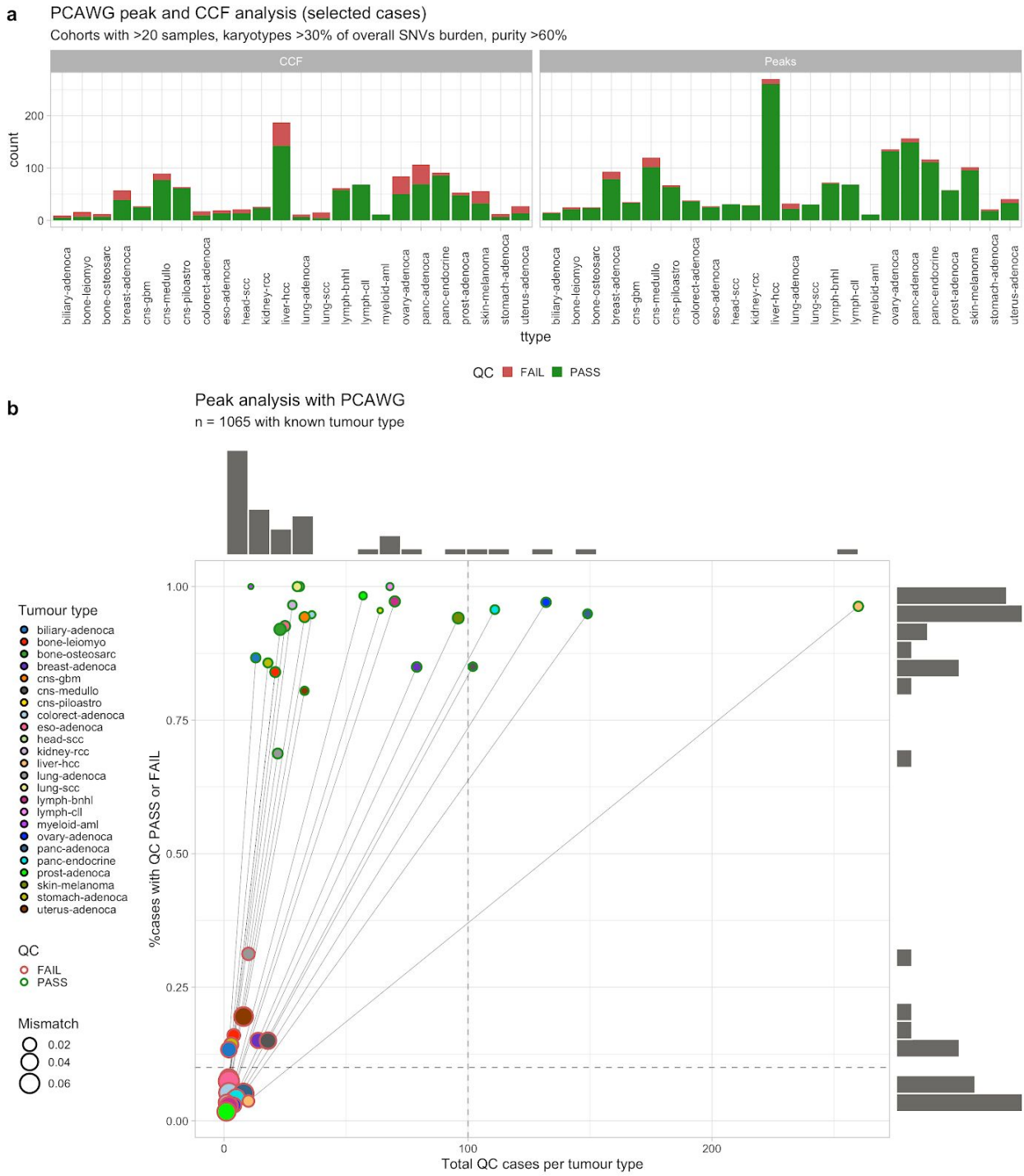
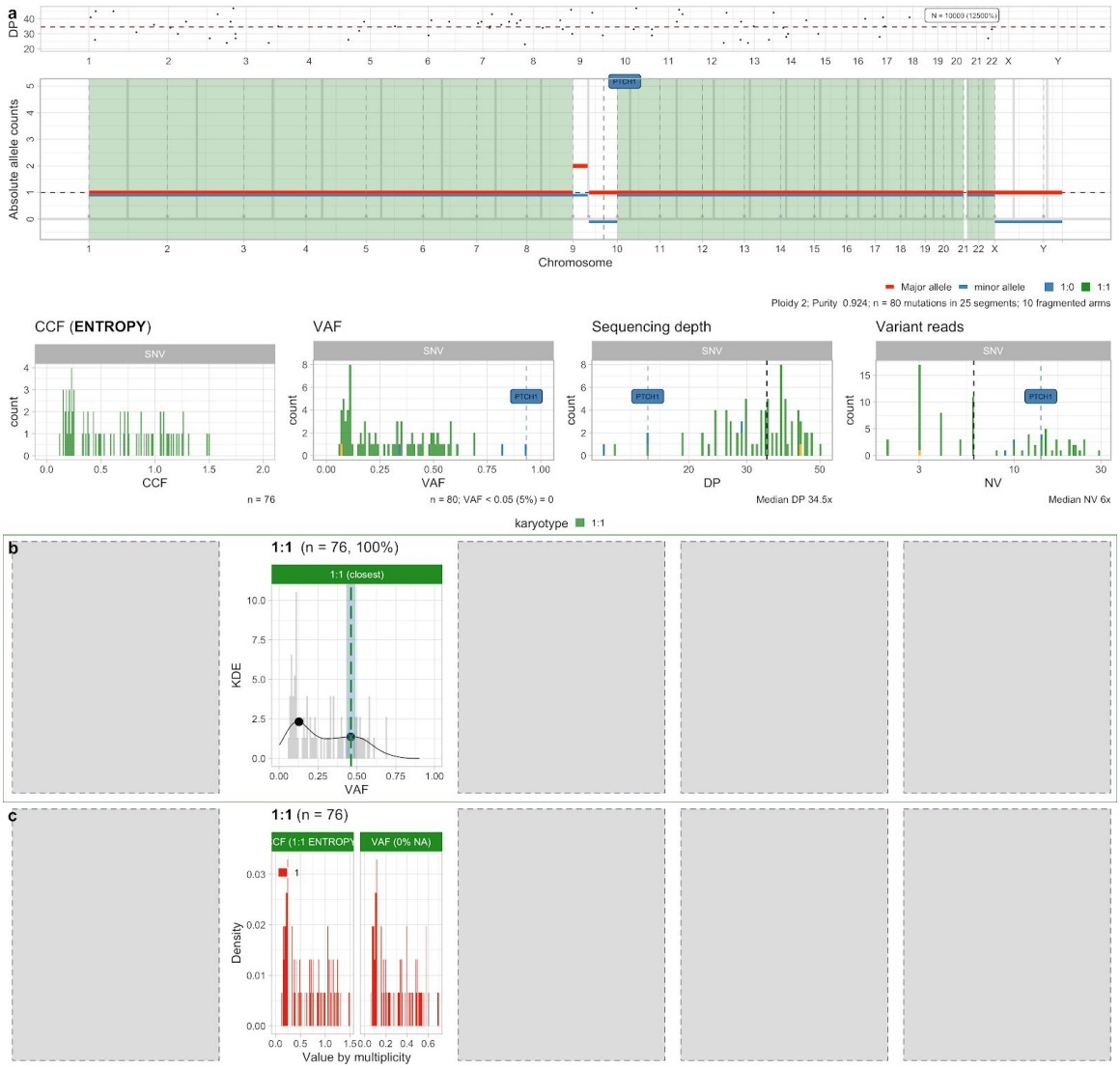


Figure 5. a. Summary CNAqc pass or fail barplot for $n = 1065$ top-quality PCAWG samples across distinct tumour types. Failures for peaks are with a 3% error tolerance, and CCFs with 10% of SNVs not assignable, per copy state. **b.** Zoom peak analysis with a scatter showing, for

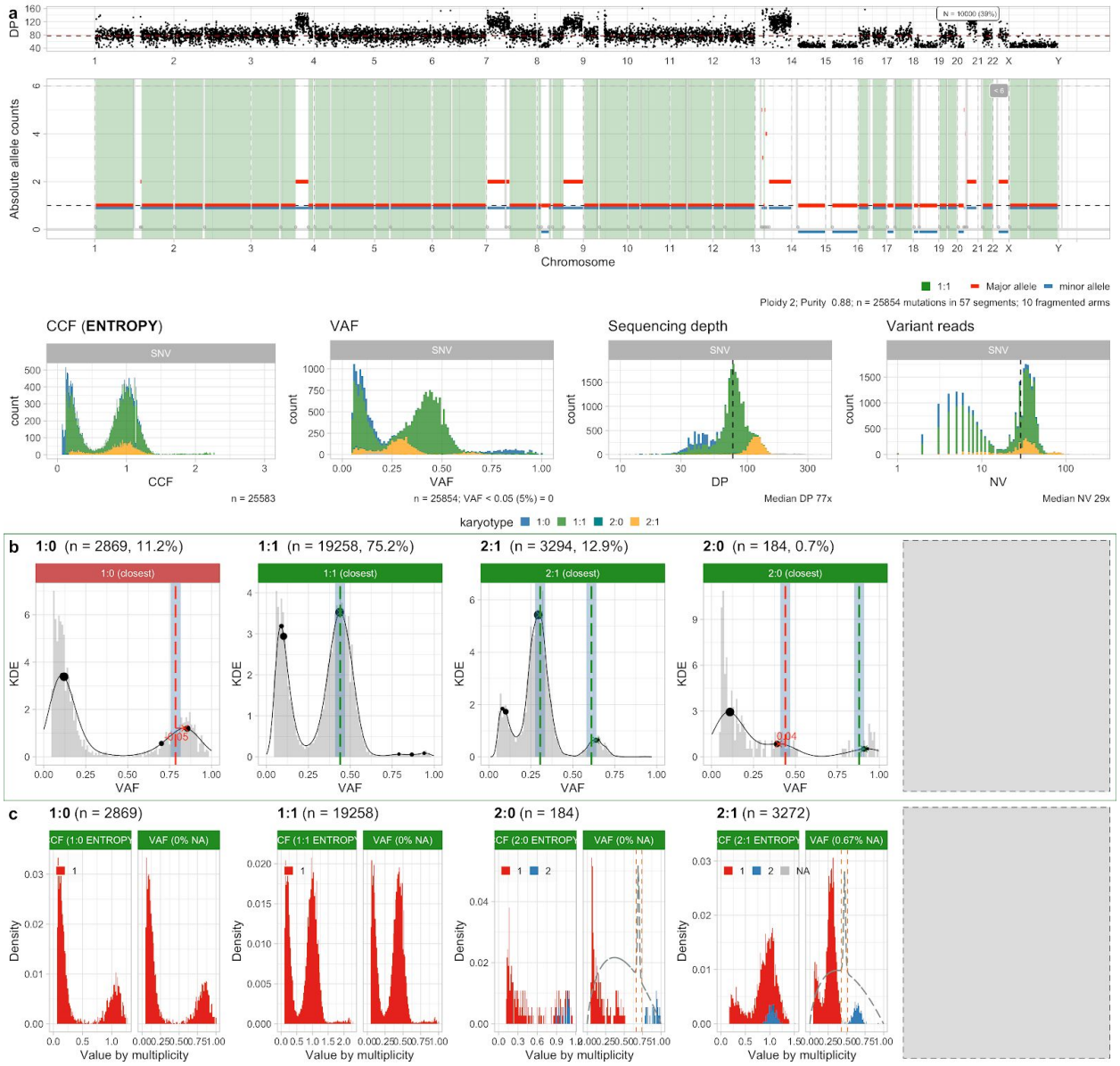
every tumour type, the total cases per tumour against the proportion of pass or fails; each dot size is proportional to the error measure from mismatched peaks.

Supplementary Figures

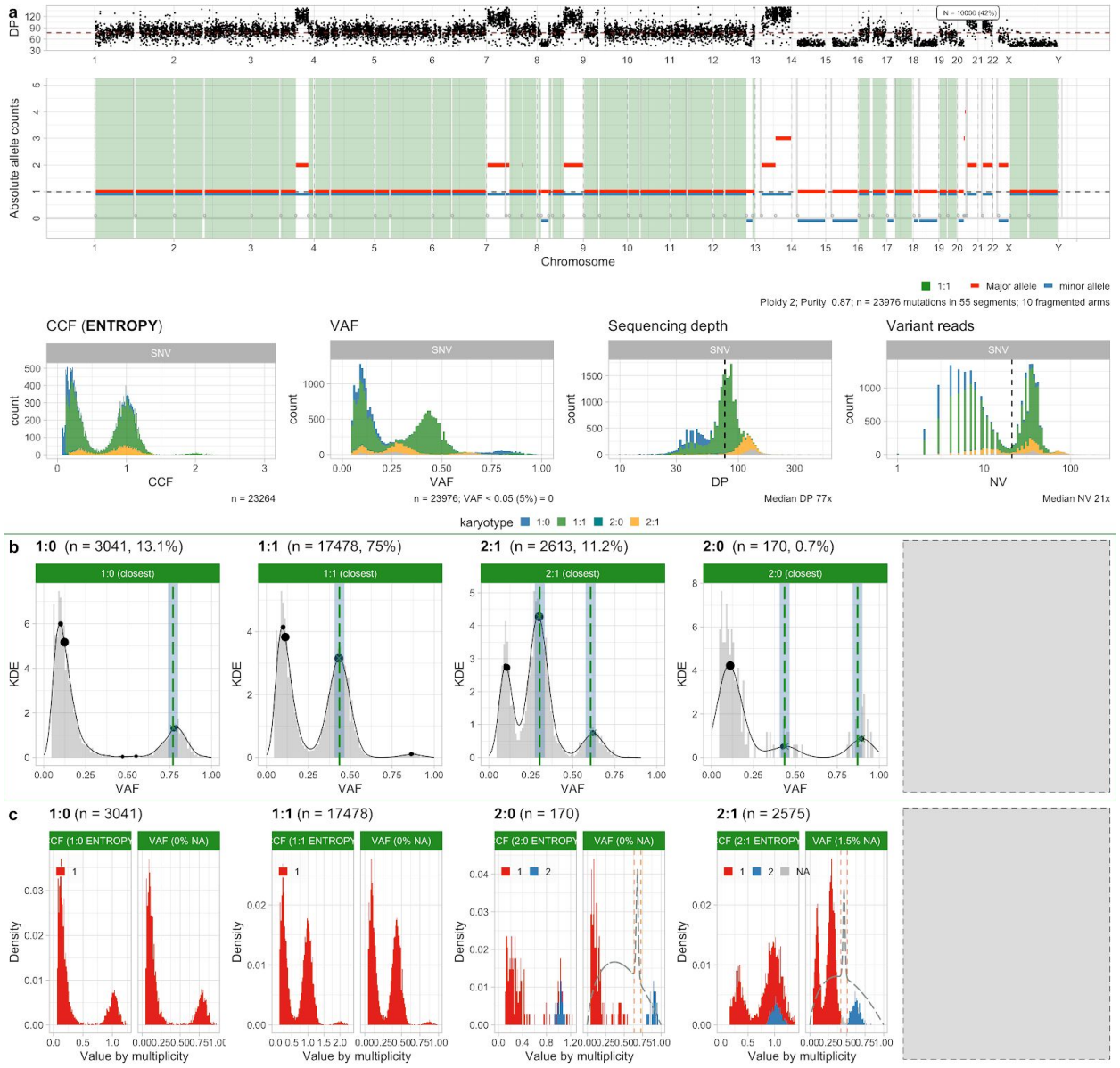
Supplementary Figure S1. PCAWG sample with low mutational burden.
Supplementary Figure S2. Sample Set7_55 (multi-region).
Supplementary Figure S3. Sample Set7_59 (multi-region).
Supplementary Figure S4. Sample Set7_62 (multi-region).
Supplementary Figure S5. Sample Set6_42 (multi-region).
Supplementary Figure S6. Sample Set6_44 (multi-region).
Supplementary Figure S7. Sample Set6_45 (multi-region).
Supplementary Figure S8. Sample Set6_46 (multi-region).
Supplementary Figure S9. Sample Set6_47 (multi-region).
Supplementary Figure S10. Sample Set6_48 (multi-region).
Supplementary Figure S11. PCAWG sample with overstimated 100% purity.
Supplementary Figure S12. PCAWG sample with true 99% purity.



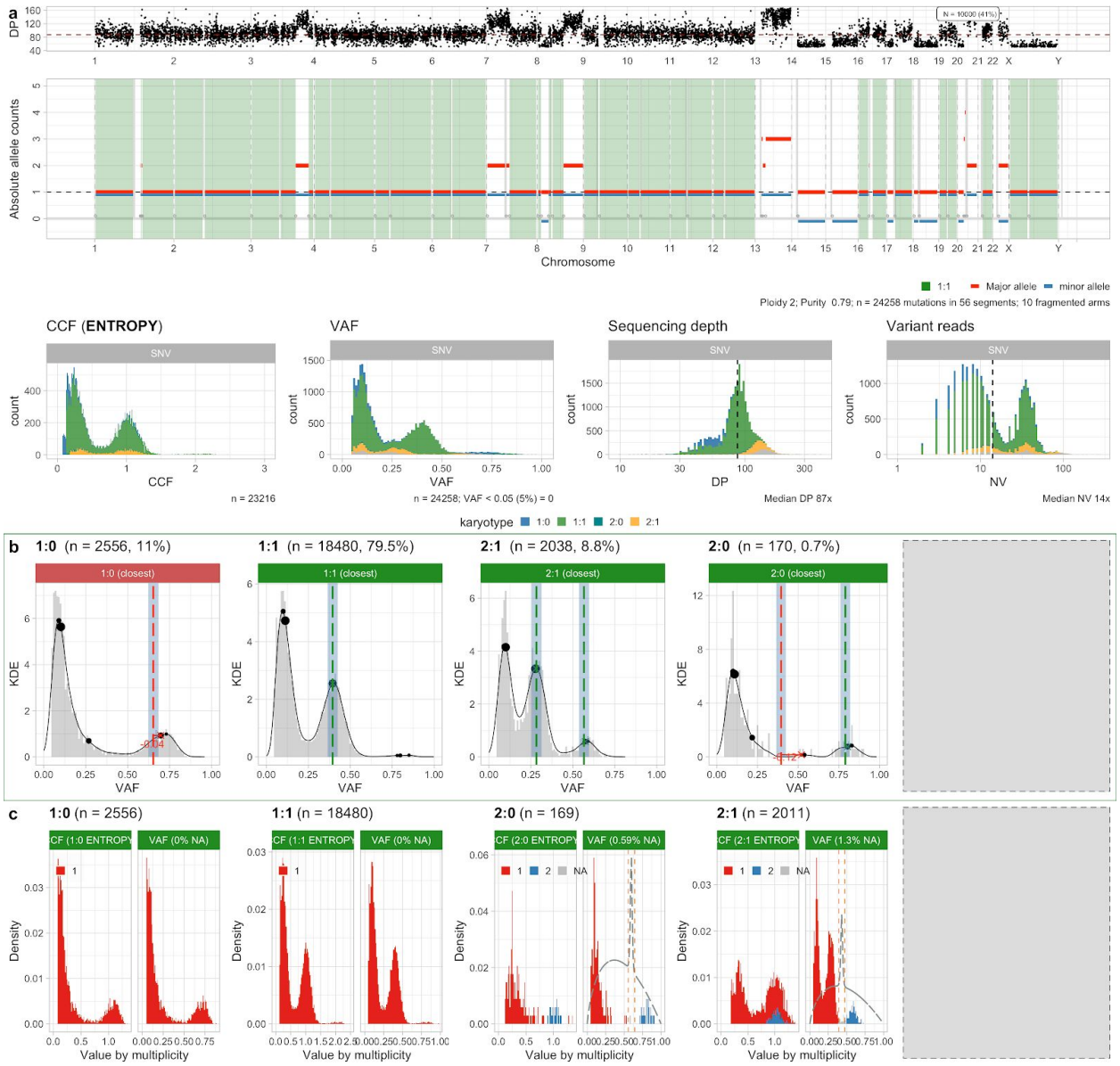
Supplementary Figure S1. Example PCAWG medulloblastoma sample with low-mutational burden, which passes data QC with CNAqc. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). Note that this sample has only 76 SNVs in diploid tumour regions, like we observe in whole-exome assays. **b,c.** Peak analysis and CCF computation for diploid SNVs.



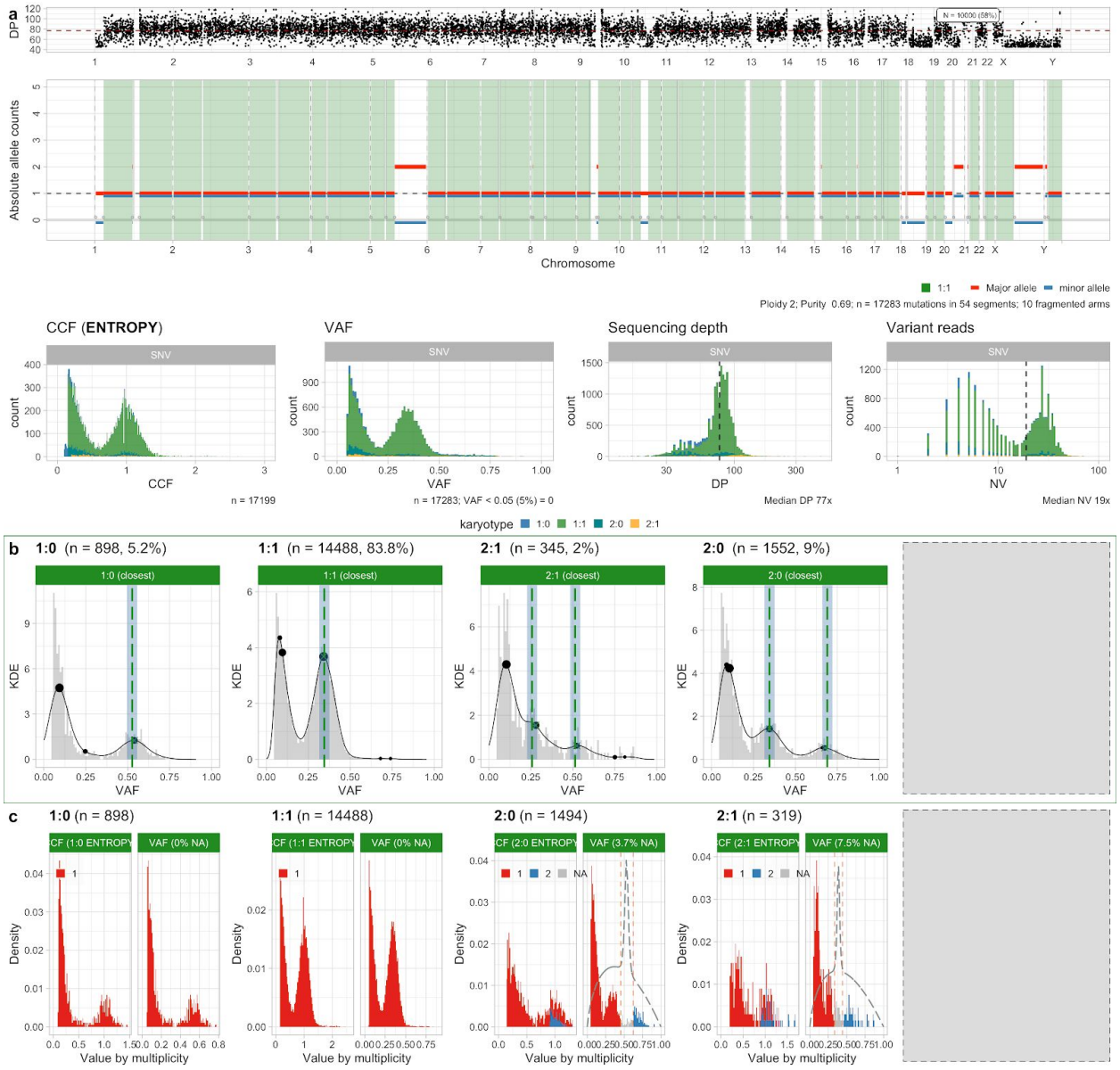
Supplementary Figure S2. Colorectal multi-region sample Set7_55 for patient Set7 (see also Main Text Figure 4). **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



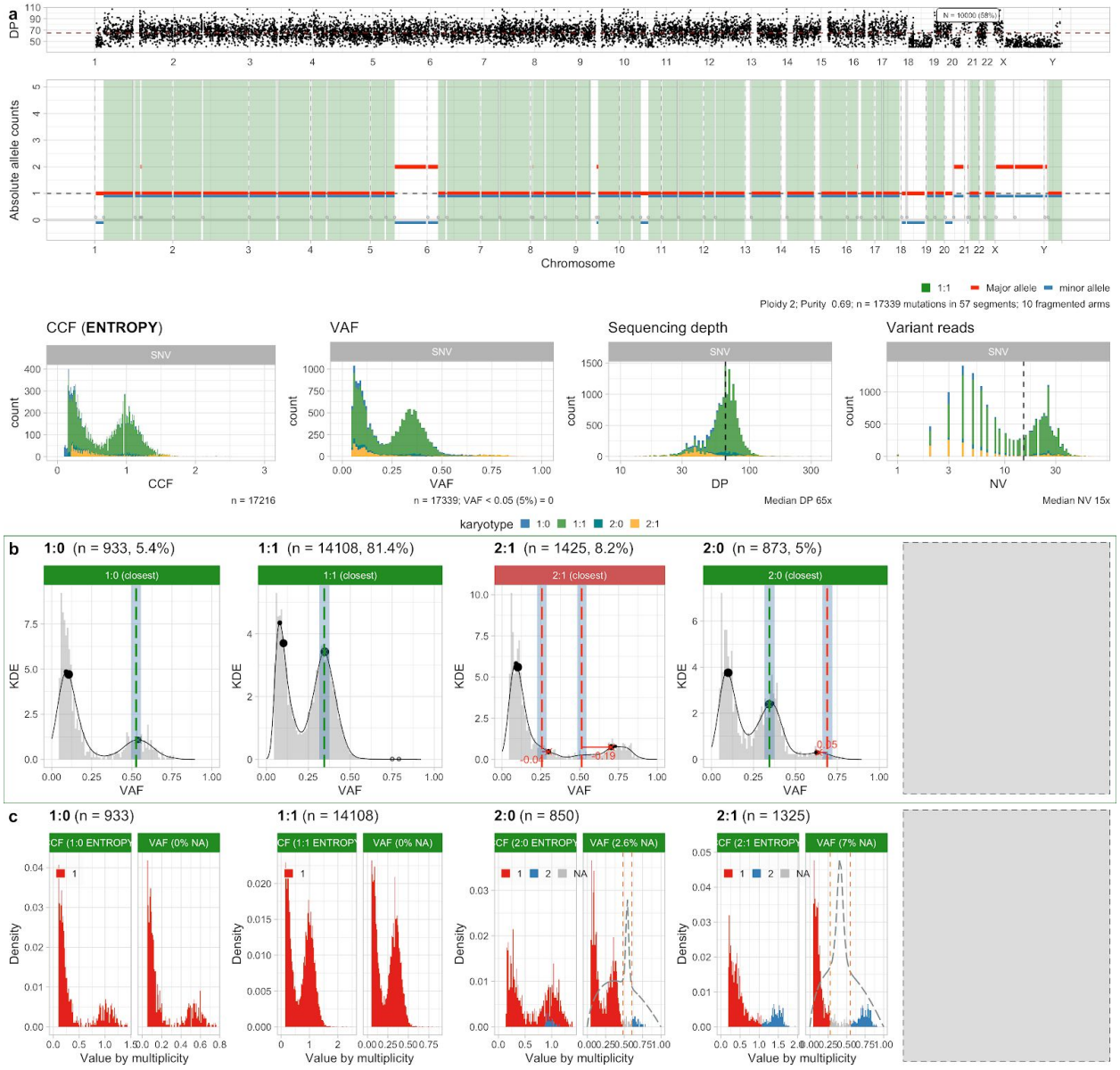
Supplementary Figure S3. Colorectal multi-region sample Set7_59 for patient Set7 (see also Main Text Figure 4). **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



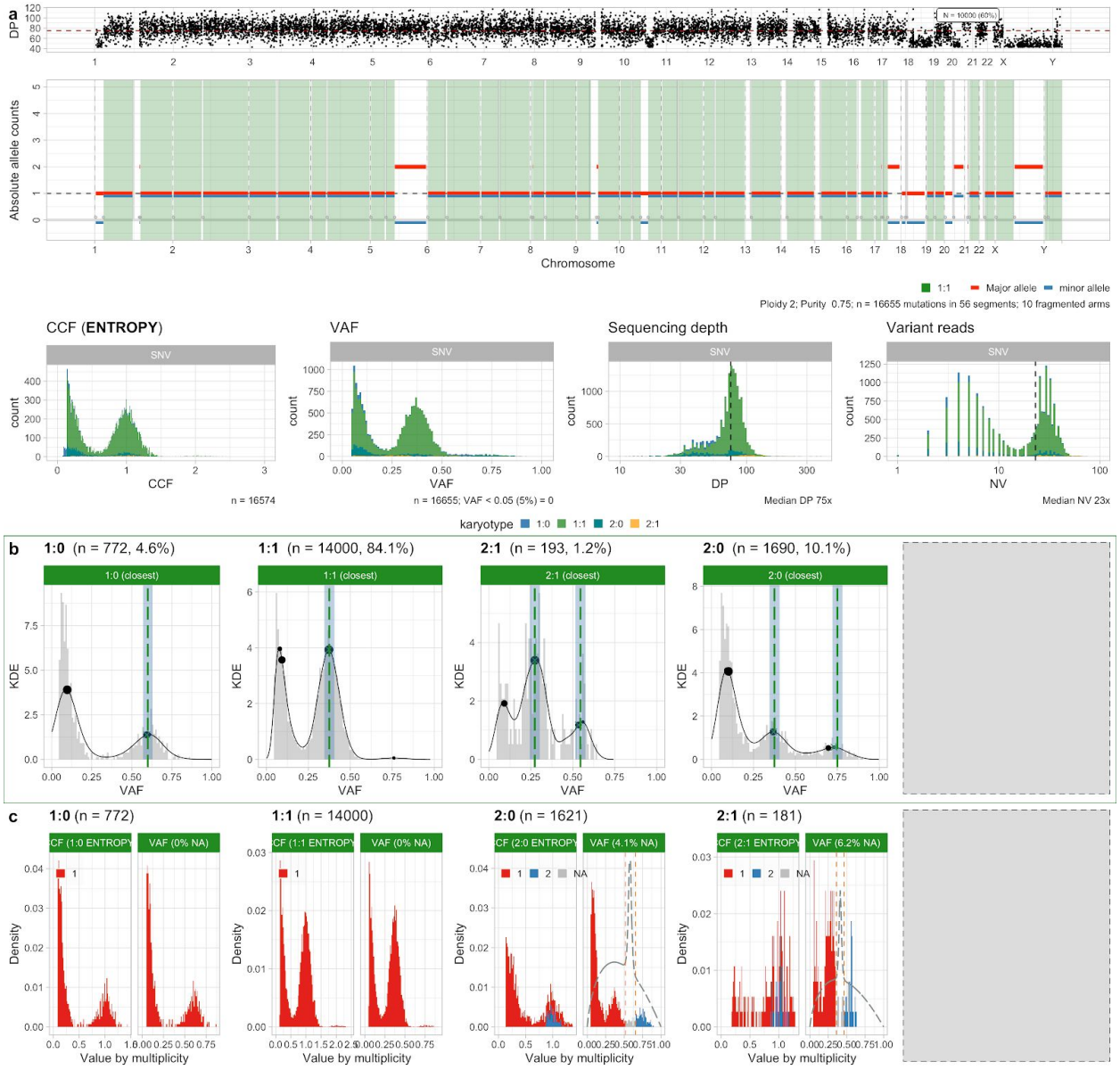
Supplementary Figure S4. Colorectal multi-region sample Set7_62 for patient Set7 (see also Main Text Figure 4). **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



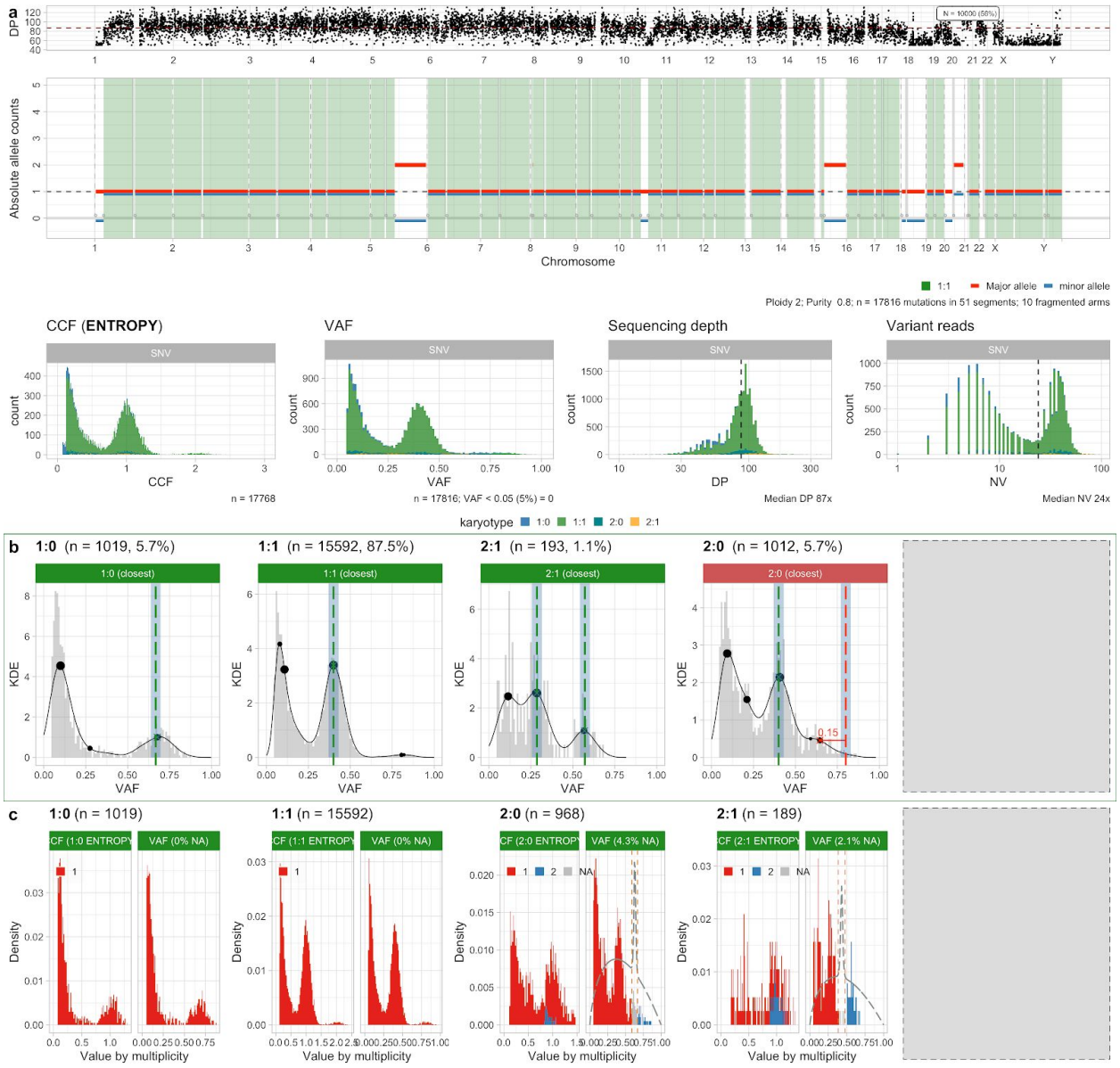
Supplementary Figure S5. Colorectal multi-region sample Set6_42 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



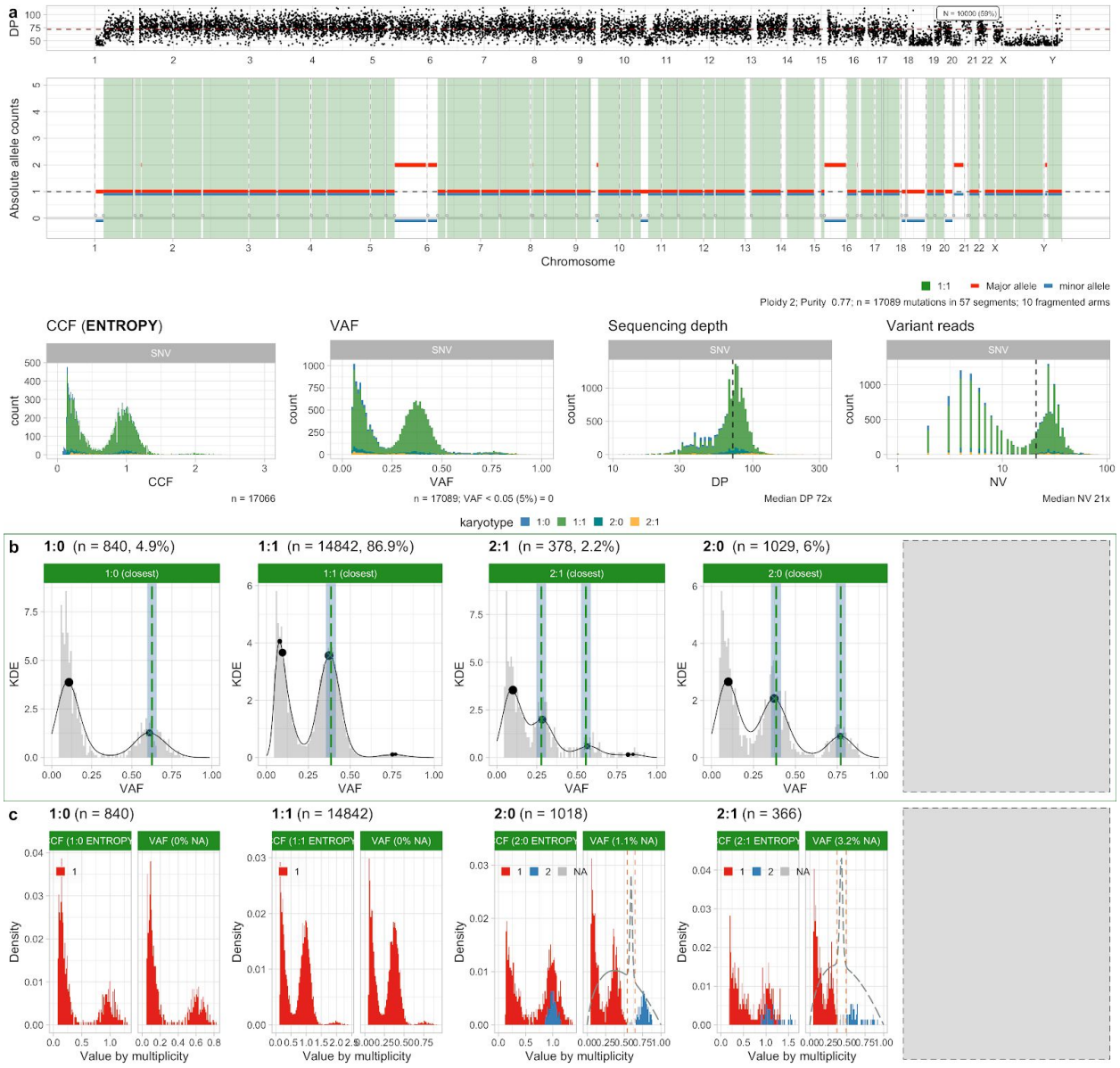
Supplementary Figure S6. Colorectal multi-region sample Set6_44 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



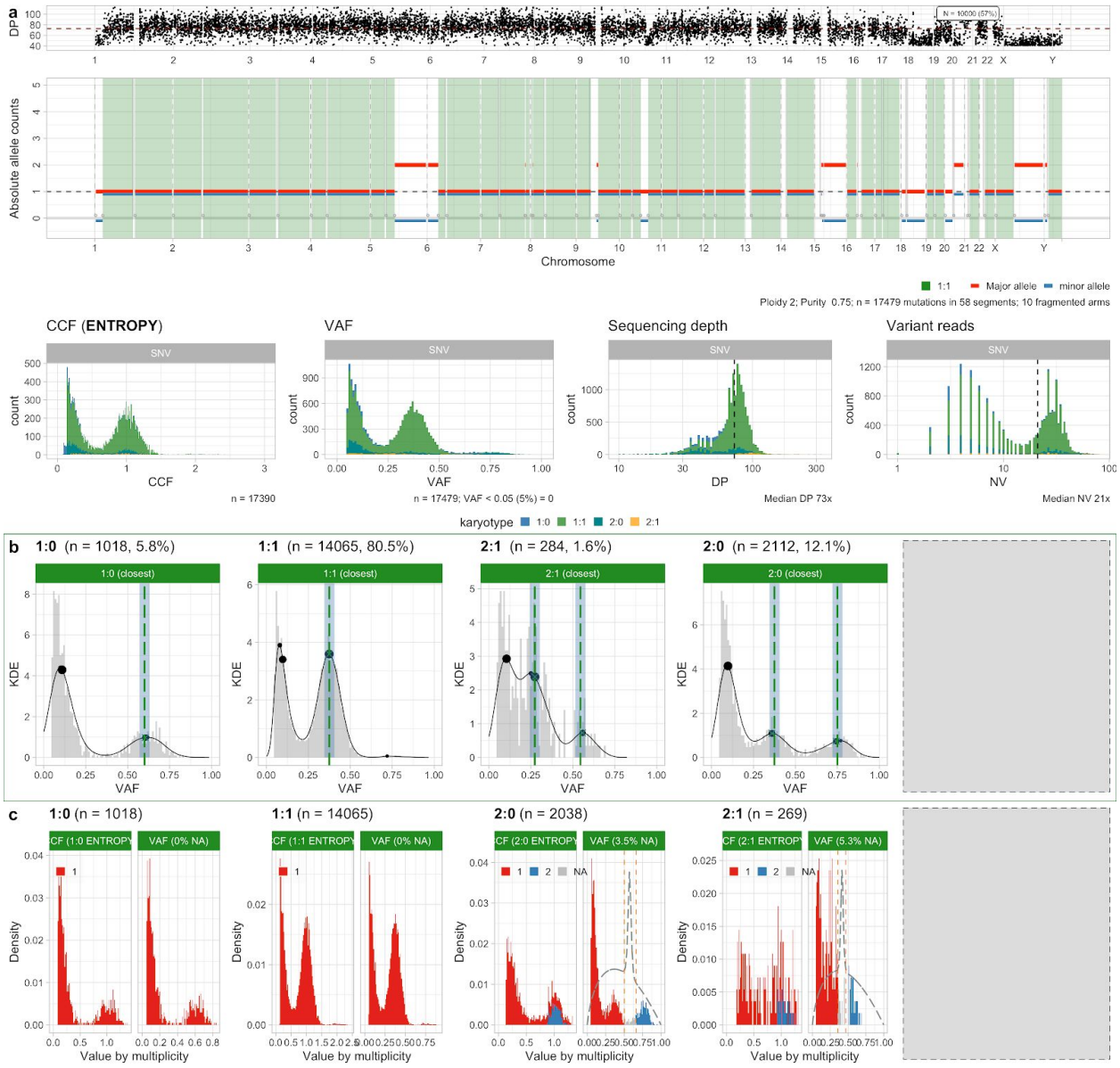
Supplementary Figure S7. Colorectal multi-region sample Set6_45 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



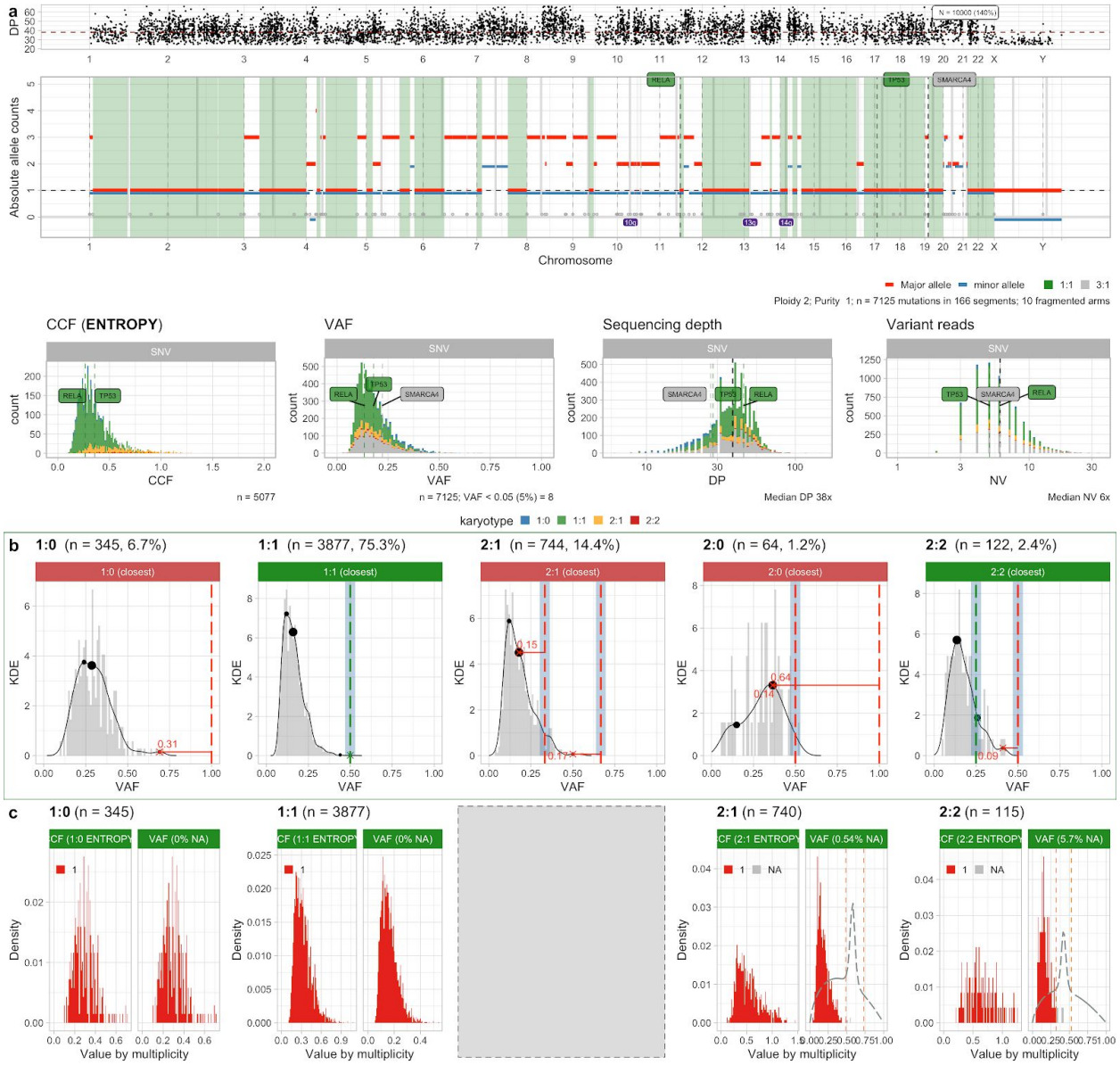
Supplementary Figure S8. Colorectal multi-region sample Set6_46 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



Supplementary Figure S9. Colorectal multi-region sample Set6_47 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.

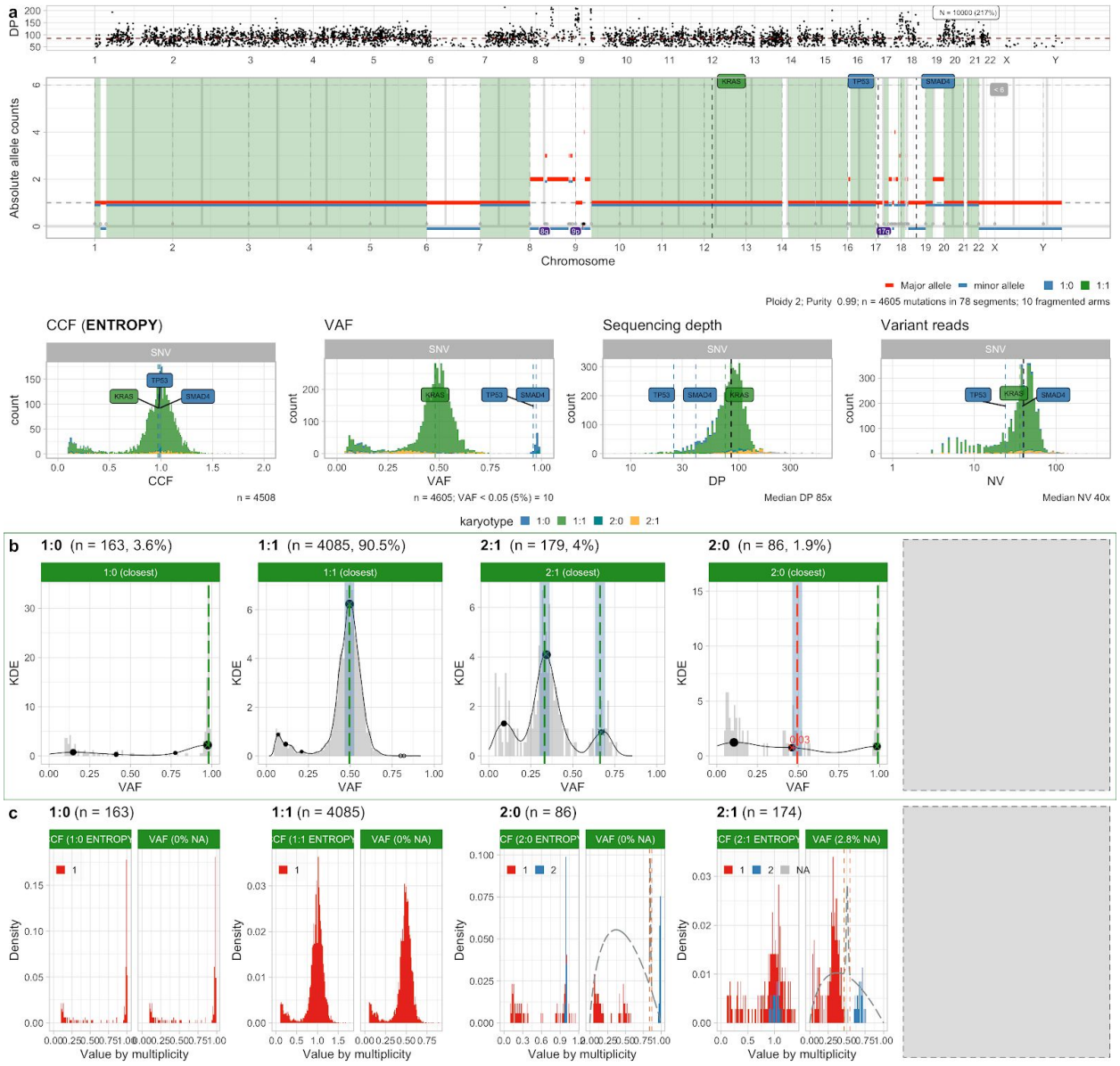


Supplementary Figure S10. Colorectal multi-region sample Set6_48 for patient Set6. **a.** Data for the sample (genome-wide CNA segments, CCF and read counts distribution). **b,c.** Peak analysis and CCF computation for the sample.



Supplementary Figure 11. 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 12. 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.