Accucopy: Accurate and Fast Inference of Allele-specific Copy Number Alterations from Low-coverage Low-purity Tumor Sequencing Data

Xinping Fan[1,2], Guanghao Luo[1,2,3], Yu S. Huang[1,2,*]

[1] *State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

[3] *School of Pharmaceutical Sciences, Jilin University, Changchun 130021, China*

* Corresponding author.

E-mail: polyactis@gmail.com (Huang Y S)

Abstract

Background: Copy number alterations (CNAs), due to its large impact on the genome, have been an important contributing factor to oncogenesis and metastasis. Detecting genomic alterations from the shallow-sequencing data of a low-purity tumor sample remains a challenging task.

Results: We introduce Accucopy, a CNA-calling method that improves and adds another layer to our previous Accurity model to predict both total (TCN) and allele-specific copy numbers (ASCN) for the tumor genome. Accucopy adopts a tiered Gaussian mixture model coupled with an innovative autocorrelation-guided EM algorithm to find the optimal solution quickly. The Accucopy model utilizes information from both total sequencing coverage and allelic sequencing coverage.

Accucopy is implemented in C++/Rust, available at http://www.yfish.org/software/.

Conclusions: We describe Accucopy, a method that can predict both TCNs and ASCNs from low-coverage low-purity tumor sequencing data. Through comparative analyses in both simulated and real-sequencing samples, we demonstrate that Accucopy is more accurate than existing methods.

Keywords: Cancer genomics; Copy number alterations; Next-generation sequencing

Background

Genomic alterations discovered in large-scale cancer genomic projects [1, 2], have therapeutic implications in being an important source of drug development [3, 4]. Copy number alterations (CNAs), due to its large impact on the genome, have been an important contributing factor to oncogenesis and metastasis [5]. Different approaches have been applied to infer CNAs from genomic sequencing data [6-10], however, detecting genomic alterations from a cancer sample mixed with normal cells remains a challenging task, esp. in low-coverage and low-purity samples. Our previous tumor purity inference method, Accurity [11], leveraging the periodic patterns among the clonal CNAs to infer the tumor purity and ploidy, can work in challenging low-purity and low-coverage settings. In this paper, we introduce Accucopy, a CNA-calling method that extends the Accurity model to predict both total (TCN) and allele-specific copy numbers (ASCN) for the tumor genome. Accucopy adopts a tiered Gaussian mixture model coupled with an innovative autocorrelation-guided EM algorithm to find the optimal solution quickly. The Accucopy model utilizes information from both total sequencing coverage and allelic sequencing coverage. Through comparative analyses in both simulated and real-sequencing samples, we demonstrate that Accucopy is more accurate than existing methods: Sclust, Sequenza, and ABSOLUTE.

Next, we first describe the data and the Accucopy model. Then, we evaluate Accucopy on numerous simulated and real-sequencing samples and compare Accucopy with Sclust [9], Sequenza [10] and ABSOLUTE [7]. We end the paper with discussions on the strength and weakness of Accucopy.

Methods

Simulated tumor and matching normal sequencing data

We generated *in silico* tumor and matching-normal WGS data using an EAGLE-based workflow at three coverage settings: 2X, 5X, and 10X. EAGLE is a software developed by Illumina to mimic their own high-throughput DNA sequencers and the simulated reads bear characteristics that are close to real-sequencing reads. We introduced twenty-one somatic copy number alterations (SCNAs), with length ranging from 5MB to 135MB and copy number from 0 to 8, affecting about 28% of the genome, to each simulated tumor genome. The entire genome of its matching normal sample is of copy number two. Over one million (=1.8 million) heterozygous single-nucleotide loci (HGSNVs) were introduced to each normal and its matching tumor sample. For each coverage setting, we first generated a pure tumor sample (purity=1.0) and its matching normal sample. We then generated nine different impure tumor samples (purity from 0.1 to 0.9) by mixing the pure tumor sample sequencing with its matching normal data proportionately. The mixing proportion determines the tumor sample's true purity.

HCC1187 cancer cell line dataset

The genome-wide CNA profile of HCC1187 has been widely studied via the spectral karyotyping (SKY) tool, which is one of the most accurate tools for characterizing and visualizing genome wide changes in ploidy [12, 13]. We used the SKY result from [14] as the ground truth for CNA comparison. SKY does not reveal a genome-wide ASCN profile and only identifies the LOH (Loss-of-Heterozygosity) regions. For these LOH regions, about 60% of the HCC1187 genome, we inferred the ASCN based on their LOH and CNA states. The whole genome sequencing data of pure HCC1187 cancer cells and its matched normal HCC1187BL cell lines was downloaded from Illumina BaseSpace. The sequencing coverage for HCC1187 and HCC1187BL is 104X and 54X respectively. Based on the pair of pure-tumor and normal real sequencing data, we generated eight impure tumor samples with purity from 0.1 to 0.9 (exclude 0.5) by proportionately mixing the HCC1187 reads with its matching normal reads.

TCGA samples

One hundred sixty-six random pairs of TCGA tumor-normal samples, downloaded from TCGA to provide a more comprehensive evaluation of Accucopy on real-world samples. The cancer types include breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), and Prostate adenocarcinoma (PRAD). The TCGA database also contains CNA profiles that are derived from Affymetrix SNP 6.0 array data. The TCGA CNA calling pipeline is built onto the existing TCGA level 2 data generated by Birdsuite [15] and uses the DNAcopy R-package to perform a circular binary segmentation (CBS) analysis [16], which translates noisy intensity measurements into chromosomal regions of equal copy number. These TCGA database CNA profiles are only for TCN (Total Copy Number). During comparative analysis, we compared Accucopy TCN estimates with these TCGA database TCN estimates.

Evaluation metrics

Define T and P as the truth and the predicted sets of copy-number segments of a sample respectively.

$$T = \{(T_i, TC_i), i \in [1, m]\} \tag{1}$$

$$P = \{(P_j, PC_j), j \in [1, n]\} \tag{2}$$

In equations above, m and n are the number of the segments in the truth and predicted sets respectively, T_i or P_j is the coordinate interval of a segment in the form of (chromosome, start, stop), and TC_i or PC_j is the copy number (float type) of this segment. Segments with no copy-number assigned by a method are excluded from *T* and *P* because any normal or abnormal assumption regarding their copy number status is hard to justify.

To evaluate the performance of a method, we defined two metrics. The first metric, CallF, is the fraction of the genome whose copy number status is identified by a method.

$$\mathrm{CallF} = \frac{\sum_{j=1}^{\mathrm{n}} Length(P_j)}{GenomeLength} \tag{3}$$

The second metric, FullC, is a correlation-like metric that measures how the predicted CNAs are concordant with the truth set. It is the fraction of matching segments, treating the copy-number difference as a continuous outcome between 0 and 1 by applying an exponential function. Segments that are normal (copy-number=2) in both *T* and *P* are excluded because copy number 2 is sometimes the most ubiquitous state of a cancer genome and a simple method calling the entire genome as copy number 2 will still perform OK if normal segments are included in the comparison.

$$\mathrm{FullC} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(T_i \cap P_j\right) \cdot e^{-|TC_i - PC_j|}}{\sum_{i=1}^{m}\sum_{j=1}^{n} T_i \cap P_j} \tag{4}$$

The value of FullC is between 0 and 1, with 0 being completely discordant and 1 being completely concordant and can be applied to both TCN and ASCN performance evaluation. In the ASCN case, we calculate FullC for the Major Allele Copy Number (MACN) as FullC is the same for the minor allele copy number. The evaluation metric used by early publications that ignores the mismatch of copy numbers and only considers the concordance of the coordinates of non-normal segments, can be misleading. For example, a duplication could be considered a match to a deletion. FullC remedies this issue by taking the copy number difference into account.

For the simulation data and the HCC1187 dataset, the truth set is known. For the TCGA samples, we use the TCN profiles in the TCGA database as a proxy for the truth set. They are derived from the Affymetrix SNP 6.0 array, not strictly a truth set, but helpful in our comparison analyses.

Summary for the Accucopy model

The Accucopy model is a probabilistic model that infers the TCN and ASCN from two types of input: the sequencing coverage information summarized by Tumor Read Enrichment (TRE) and the allele-specific coverage information summarized by Log ratio of Allelic-coverage Ratios (LAR) at HGSNVs. Definitions of TRE and LAR are in Additional File 1. TREs are samples from a multi-component Gaussian mixture model with the tumor purity, the tumor ploidy, and the total copy number of each genomic region as parameters. LARs are samples from a two-component Gaussian mixture model with the allele-specific copy numbers of genomic regions as additional parameters. The entire genome is segmented by an enhanced version of the public GADA [17] (unpublished). We developed an autocorrelation-guided EM algorithm to find optimal parameters. Bayesian Information Criterion is adopted to avoid model overfitting. More details of the Accucopy model are in the Additional File 1.

Results

Evaluation of Accucopy on the simulated data

We evaluated the FullC and CallF of Accucopy using simulated tumor and normal data under three coverage settings: 2X, 5X, and 10X, and nine different purity settings, 0.1-0.9, (Fig. 1). For the TCN inference, Accucopy achieved high FullC and CallF, mostly >0.95, regardless of the tumor purity level, even if the coverage is only 2X (Fig. 1A, 1C and 1E). In the low-purity (0.1-0.4) cases of the 2X coverage, the TCN FullC deteriorates only slightly to about 0.9. For the ASCN inference, Accucopy achieved robust FullC and CallF, >0.8, when the sample purity is equal to or above 0.2 (Fig. 1B and 1D) for the 5X and 10X coverage. In the low coverage settings (2X), Accucopy requires the purity of sample to be at least 0.6 to achieve good performance in ASCN inference, but the total copy number (TCN) estimates are still >90% correct (Fig. 1E). We think the extremely low tumor content (<0.4), less than 1.2X (=0.6*2X) coverage on average for the tumor cells, renders the ASCN inference quite challenging.

We compared Accucopy with Sclust, Sequenza, and ABSOLUTE. All three methods can infer the tumor purity, SCNAs, and ASCNs. Sclust performs well in high-coverage (>=10X) and high-purity settings (purity >=0.6) (Fig. 1A and 1B) and performs reasonably well in medium coverage (5X) and high purity settings (purity ≥ 0.6), but performs poorly in low-purity (<=0.5) or low-coverage (2X) settings. Sequenza performs similarly to Sclust in 5X and 10X but outperforms Sclust in 2X and low-purity conditions. Sequenza has the strange phenomenon that it performs better on 5X samples than 10X samples. We found that Sequenza over-segments the genome in 10X samples and calls many of these small segments with the wrong copy number and thus has lower performance in 10X than 5X. The case of ABSOLUTE is curious. It achieves good TCN performance on par with Accucopy in some conditions but performs quite poorly in other conditions. We found that its TCN performance is dependent on its ability to estimate the tumor purity correctly, (Table 1). Across the coverage and purity level, Accucopy is the top performer or a very close second.

To illustrate the performance difference, we plotted the estimated TCN and ASCN of chromosome 1 by Accucopy and Sclust, with the true purity being 0.4 (low) or 0.8 (high) and coverage being 2X (low) and 10X (high) (Supplementary Fig. 1). In the 10X-coverage high-purity sample, the output by Accucopy and Sclust are very close to the truth (Supplementary Fig. 1B and 1C). If the purity decreases 0.4, the TCN and ASCN estimates of Accucopy are still very close to the truth while Sclust underestimates TCN and ASCN (Supplementary Fig. 1D and 1E). If the sequencing coverage decreases to 2X, Accucopy can still infer the true TCN in both high and low purity settings but its ASCN inference deteriorates in the low-purity low-coverage sample (Supplementary Fig. 1F and 1H). Sclust overestimated both TCN and ASCN in the 2X high-purity setting (Supplementary Fig. 1G) and failed completely in the 2X low-purity setting (Supplementary Fig 1I). This detailed comparison confirmed conclusions drawn from the summary evaluation plot (Fig. 1). The main strength of Accucopy, compared to Sclust and other methods, is that it can perform in low-coverage and/or low-purity settings while others are unstable.

The simulation analysis suggests: a) Accucopy can accurately estimate TCN in a wide range of purity (0.1-0.9) and coverage (2X and above) settings. b) Accucopy can robustly infer ASCN as long as the purity is above 0.1 in moderate or high coverage (>=5X) settings; c) In low-coverage (2X) settings, the ASCN inference by Accucopy requires the purity to be above 0.5, which suggests that a minimal 1X tumor content (=total-coverage*purity), i.e. 10X*0.1, 5X*0.2, 2X*0.5, in a sequenced sample is required for an accurate ASCN inference by Accucopy.

Evaluation of Accucopy on the HCC1187 dataset

The prior simulation study has shown the solid performance of Accucopy in low (5X) and medium (10X) coverage settings. In this section, we run Accucopy on the HCC1187 dataset to validate Accucopy on a real sequencing dataset. We know the true purity of the eight impure HCC1187 tumor samples because we designed the mixing of HCC1187 and its corresponding normal cells. The Accucopy performance in TCN inference is on par with that of the simulation study (Fig. 2A). The MACN inference is better than that of the simulation study under similar conditions (Fig. 2B), because all the true MACNs of HCC1187 are effectively LOHs (Loss-Of-Heterozygosity). The non-LOHs of HCC1187 have unknown MACN state and are excluded in comparison.

The statistical power to infer MACNs is highest for LOHs than non-LOHs because the difference between the major and the minor allele copy number is at its widest gap. This exercise indicates for real-sequencing samples, Accucopy can still achieve solid performance, comparable to its performance in the simulated data.

Inferring SCNAs for TCGA samples

We ran Accucopy and Sclust on 166 pairs of TCGA tumor-normal samples that have corresponding TCN profiles in the TCGA database. Accucopy succeeded for 110 samples. Accucopy failed on 56 samples due to noisy TRE data, which is caused by high level of intra-tumor heterogeneity and/or genomic alterations. Sclust succeeded for 57 samples. We compared the TCN output by either method against the corresponding TCGA TCN profiles.

The Accucopy FullC metric is strongly correlated with the tumor purity (Fig. 3A), and is independent of CallF (Fig. 3C). The average Accucopy CallF is about 95%, regardless of the tumor purity (Fig. 3B), which indicates Accucopy predicts TCNs for almost the entire genome of all analyzed samples. The Sclust FullC is also correlated with the tumor purity, but only among samples with purity above 0.5 and coverage above 10X (Fig. 4A). These samples tend to have high CallF (Fig. 4B and 4C). The decline of FullC with the decreasing tumor purity observed in both Accucopy and Sclust are quite interesting. The prior simulation and HCC1187 studies indicate that Accucopy performs well in predicting TCNs for samples with coverage 2-10X and purity 0.1-0.9 and Sclust performs well in purity>0.5 and coverage>=10X samples.

We carefully compared the Accucopy TCN prediction for samples in the high-

FullC-high-purity top-right part of Fig. 3A vs samples in the low-FullC-low-purity lower-left part of Fig. 3A and found that the decline of FullC with the decreasing tumor purity is primarily caused by the diminishing statistical power of the TCGA pipeline as the tumor purity declines (Fig. 4). The TCGA CNA pipeline (Birdsuite + CBS) assumes a tumor sample consisting of 100% tumor cells. Thus the copy number of a genomic segment predicted by the TCGA pipeline is a weighted average of its respective copy numbers in the tumor and normal cells. As the purity of a tumor sample declines, the increasing fraction of normal cells, whose genomic copy number is two, will move the predicted average copy number closer to two. Accucopy and Sclust explicitly model the tumor purity and do not suffer from this issue. This is exactly what we observe in detailed TCGA vs. Accucopy comparisons (Fig. 5). In both samples, the segmentations of the genome by the TCGA pipeline and Accucopy are highly similar. In addition, the copy number qualitative predictions (duplication or deletion) for individual segments are highly similar too. Were it not for FullC to consider copy number differences, both samples would have shown near perfect concordance between the TCGA profile and the Accucopy output. In the low-purity sample (Fig. 5A), the copy number quantitative predictions of abnormal segments are closer to two and are numerically less concordant with those by Accucopy, manifested by a lower FullC than the high-purity sample (Fig. 5B).

It is also clear from Fig. 4 that Sclust works in more limited conditions than Accucopy. Sclust requires the tumor purity above 0.5 and the sequencing coverage above or near 10X. For samples with sequencing coverage below 10X, Sclust may predict copy numbers for only a fraction of the genome (Fig. 4B). For samples with coverage lower than 5X, (Fig. 4D), Sclust failed completely. This is consistent with the simulation finding that Sclust has lower power in detecting CNAs from the low-purity (<0.5) and/or low-coverage (<=5X) samples.

The TCGA study indicates Accucopy is capable of identifying copy number alterations in complex real-world samples, some of which may have very low sequencing coverage and are of low tumor purity.

Implementation and performance

Accucopy is implemented in vanilla C++ and Rust and is released for Ubuntu 18.04 in a docker. In theory, it can be built for Windows or MacOS but we have not tested it. Average runtime of Accucopy is about 45 minutes for a 5X tumor/normal matched pair; about three hours for a 30X tumor/normal matched pair on a single core of Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz; with the peak RAM consumption under 4GB.

We provided all methods with the same input bam files and ran all programs with default parameters under the same computational environment as stated above.

Discussion

Through extensive simulated and real-sequencing data analyses, we have demonstrated that Accucopy is a fast, accurate, and fully automated method that infers TCN and ASCN of somatic CNAs from tumor-normal high-throughput sequencing data. The strength of Accucopy, relative to other methods, lies particularly in its performance in low-coverage and low-purity samples. This makes Accucopy an excellent choice in first-round low-coverage screening type of analysis. It can offer crucial insight regarding the tumor purity, ploidy, TCNs, and ASCNs before an expensive in-depth high-coverage analysis is started.

One under-appreciated factor contributing to the excellent performance of Accucopy is the large amount of simulation and real-sequencing samples with known truth (or near truth for TCGA samples). This trove of data leads us to adjust many aspects of the Accucopy model during development. Here are a few notable adjustments. A coverage smoothing step greatly reduced the random noise in sequencing coverage. Adoption of Strelka2 [18] dramatically reduced the number of false positives in calling heterozygous SNPs, compared to other variant callers we tried. Extensive in-depth analyses uncovered that the expectation of Log ratio of Allelic Ratios (LAR) needed to be adjusted due to the exclusion of zero-allele-coverage SNPs, which improved the Accucopy performance in the ASCN inference by an order of magnitude. These adjustments may look trivial but cumulatively are very effective in improving the overall performance of Accucopy.

The requirement of a periodic TRE pattern arising from varying copy numbers means that Accucopy is not suitable for tumor samples with little or no copy number alterations. An excessive amount of point mutations in a tumor, relative to its matching normal, resulting in many wrong alignments, will also render Accucopy unable to confidently discover a period from the TRE pattern. Another case that could weaken Accucopy is the presence of copy number variations (CNVs) in healthy normal individuals. At these genomic regions, the TCN and ASCN predictions by Accucopy will be inaccurate as Accucopy assumes the entire genome of a normal sample to be of copy number two.

For regions where different tumor subclones may harbor different SCNAs, the current way of outputting averaged TCNs and MACNs is less than satisfactory from the standpoint of tumor clonal evolution. The next iteration of Accucopy will try to address this.

Conclusions

Through extensive simulated and real-sequencing data analyses, we have demonstrated that Accucopy is a fast, accurate, and fully automated method that infers TCN and ASCN of somatic CNAs from tumor-normal high-throughput sequencing data. The strength of Accucopy, relative to other methods, lies particularly in its performance in low-coverage low-purity samples. This makes Accucopy an excellent choice in first-round low-coverage screening type of analysis. It can offer crucial insight regarding the tumor purity, ploidy, TCNs, and ASCNs before an expensive in-depth high-coverage analysis starts.

Availability and Requirements

Project name: Accucopy

Project home page: https://github.com/polyactis/Accucopy

Operating system(s): Linux

Programming language: C++/Rust/Python

License: SIMM Institute License, free for non-commercial use.

Any restrictions to use by non-academics: license needed.

Abbreviations

CNAs: Copy number alterations

TCN: total copy number

ASCN: allele-specific copy numbers

SCNAs: somatic copy number alterations

HGSNVs: heterozygous single-nucleotide loci

SKY: spectral karyotyping

LOH: Loss-of-Heterozygosity

BRCA: breast invasive carcinoma

COAD: Colon adenocarcinoma

GBM: Glioblastoma multiforme

HNSC: Head and Neck squamous cell carcinoma

PRAD: Prostate adenocarcinoma

CBS: circular binary segmentation

MACN: Major Allele Copy Number

TRE: Tumor Read Enrichment

LAR: Log ratio of Allelic-coverage Ratios

CNVs: copy number variations

Declarations

Ethics approval and consent to participate

This publication only uses public data only. Therefore, we do not need ethics approval or consent to participate.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the TCGA. Barcode and UUID of each TCGA sample file can be found in Additional File 3.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Chinese Academy of Sciences Hundred-Talent program, "Personalized Medicines -- Molecular Signature-based Drug Discovery and Development" Strategic Priority Research Program of the Chinese Academy of Sciences XDA12050202, and grant Y5G203F018 from State Key Laboratory of Drug Research of Shanghai Institute of Materia Medica, awarded to YSH.

Authors' contributions

YSH conceived the idea and supervised the study. XF and YSH designed and implemented the method. XF and GL performed the simulated data analysis. XF performed the other data analyses. XF and YSH wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We wish to acknowledge the guidance provided by the authors of Sclust on how to

configure their software.

References

1. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN *et al*: The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 2013, 45(10):1113-1120.
2. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS *et al*: International network of cancer genome projects. *Nature* 2010, 464(7291):993-998.
3. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J *et al*: Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012, 483(7391):570-575.
4. Simon R, Roychowdhury S: Implementing personalized cancer genomics in clinical trials. *Nature Reviews Drug Discovery* 2013, 12(5):358-369.
5. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W *et al*: Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 2019, 575(7781):210-216.
6. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E: Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)* 2012, 28(3):423-425.
7. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW,

Onofrio RC, Winckler W, Weir BA *et al*: Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* 2012, 30(5):413-421.

8. Chen H, Bell JM, Zavala NA, Ji HP, Zhang NR: Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic acids research* 2015, 43(4):e23.

9. Cun Y, Yang TP, Achter V, Lang U, Peifer M: Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nat Protoc* 2018, 13(6):1488-1501.

10. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, Szallasi Z, Eklund AC: Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 2015, 26(1):64-70.

11. Luo Z, Fan X, Su Y, Huang YS: Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. *Bioinformatics (Oxford, England)* 2018, 34(12):2004-2011.

12. Schröck E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D *et al*: Multicolor spectral karyotyping of human chromosomes. *Science (New York, NY)* 1996, 273(5274):494-497.

13. Sirivatanauksorn V, Sirivatanauksorn Y, Gorman PA, Davidson JM, Sheer D, Moore PS, Scarpa A, Edwards PA, Lemoine NR: Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping.

International journal of cancer 2001, 91(3):350-358.

14. Chen W, Robertson AJ, Ganesamoorthy D, Coin LJM: sCNAphase: using haplotype resolved read depth to genotype somatic copy number alterations from low cellularity aneuploid tumors. *Nucleic acids research* 2017, 45(5):e34.

15. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K *et al*: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008, 40(10):1253-1260.

16. Olshen AB, Venkatraman ES, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics (Oxford, England)* 2004, 5(4):557-572.

17. Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S: Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics (Oxford, England)* 2008, 24(3):309-318.

18. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P *et al*: Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods* 2018, 15(8):591-594.

Table 1 Purity estimates by all methods

Coverage	True purity	Accucopy	Sequenza	ABSOLUTE	Sclust
2X	0.1	0.1047	1.0	0.26	-
2X	0.2	0.2069	1.0	0.21*	-

2X	0.3	0.31214	0.98	0.32*	-
2X	0.4	0.41966	0.47	0.42*	-
2X	0.5	0.5199	0.56	0.53*	0.59
2X	0.6	0.62211	0.66	0.31	0.61
2X	0.7	0.73768	0.86	0.37	0.74
2X	0.8	0.83951	0.91	0.72	0.63
2X	0.9	0.94051	0.95	0.61	0.99
5X	0.1	0.09751	0.13	0.22	-
5X	0.2	0.20108	0.22	0.2*	-
5X	0.3	0.31011	0.38	0.36	0.24
5X	0.4	0.40559	0.44	0.41*	0.32
5X	0.5	0.51208	0.55	0.69	0.37
5X	0.6	0.6125	0.65	0.47	0.64
5X	0.7	0.71576	0.76	0.36	0.72
5X	0.8	0.816	0.84	0.81*	0.81
5X	0.9	0.91884	0.94	0.46	0.98
10X	0.1	0.098563	0.29	0.22	0.27
10X	0.2	0.20085	0.31	0.25	-
10X	0.3	0.30411	0.36	0.27	-
10X	0.4	0.40553	0.42	0.51	0.3
10X	0.5	0.5066	0.52	0.51*	-
10X	0.6	0.60569	0.63	0.3	0.65
10X	0.7	0.70754	0.74	0.53	0.7
10X	0.8	0.80739	0.83	0.8*	0.83
10X	0.9	0.90612	0.94	0.45	0.92

Note: Asterisk (*) in the ABSOLUTE column indicates ABSOLUTE performed well on the sample. Dash (-) in the Sclust column indicates Sclust failed on the sample.

Figures

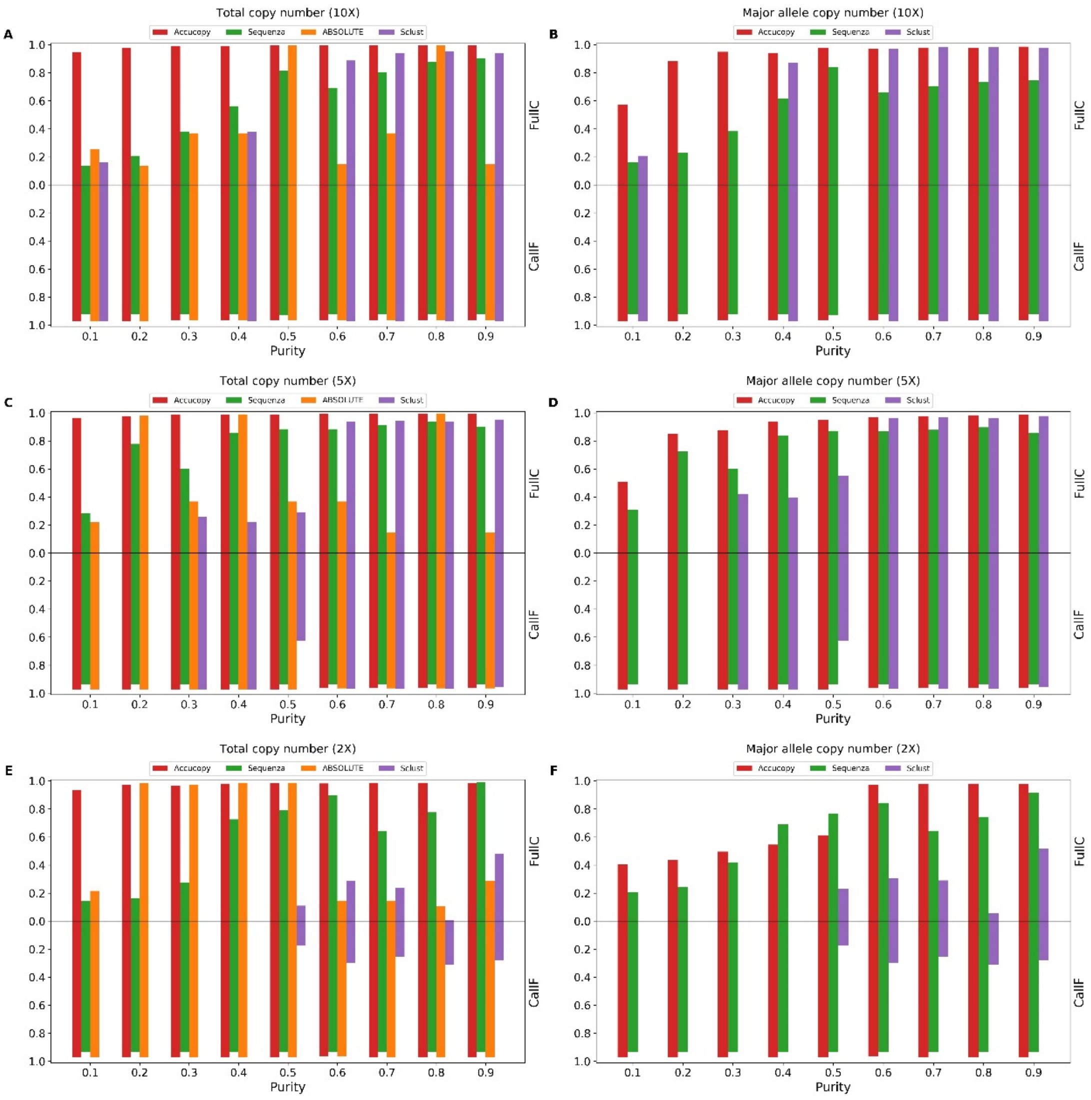


Fig. 1 Evaluation of Accucopy and Sclust on simulation data.

FullC and CallF of total copy number on low-coverage 10X (**A**), low-coverage 5X (**C**) and low-coverage 2X (**E**). FullC and CallF of major allele copy number on low-coverage 10X (**B**), low-coverage 5X (**D**) and low-coverage 2X (**F**). The blank space in the figure indicates Sclust failed on this sample. The red, green, orange and purple bar represent Accucopy, Sequenza, ABSOLUTE and Sclust respectively.

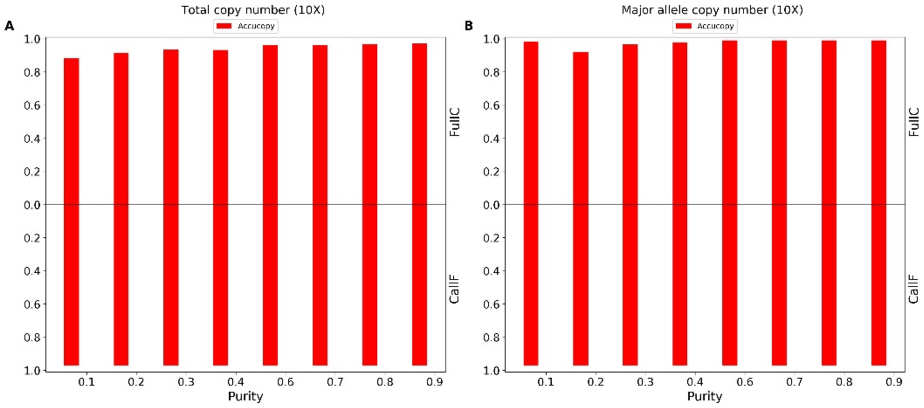


Fig. 2 Accucopy performance on the HCC1187 dataset.

The sequencing coverage for all samples is 10X. The tumor purity varies from 0.1 to 0.9. **A.** The TCN FullC plot shows the Accucopy TCN calls are at least 90% concordant with the true TCN calls. The TCN CallF indicates Accucopy predicts TCN for close to 100% of the genome. **B.** The MACN FullC plot indicates the Accucopy MACN calls are close to 95% concordant with the true MACN calls. The MACN inference is better than that of the simulation study under similar conditions because all the true MACNs of HCC1187 are effectively LOHs (Loss-Of-Heterozygosity). The non-LOHs of HCC1187 have unknown MACN state and are excluded in comparison. The statistical power to infer MACNs is highest for LOHs than non-LOHs because the difference between the major and the minor allele copy number is at its widest gap.

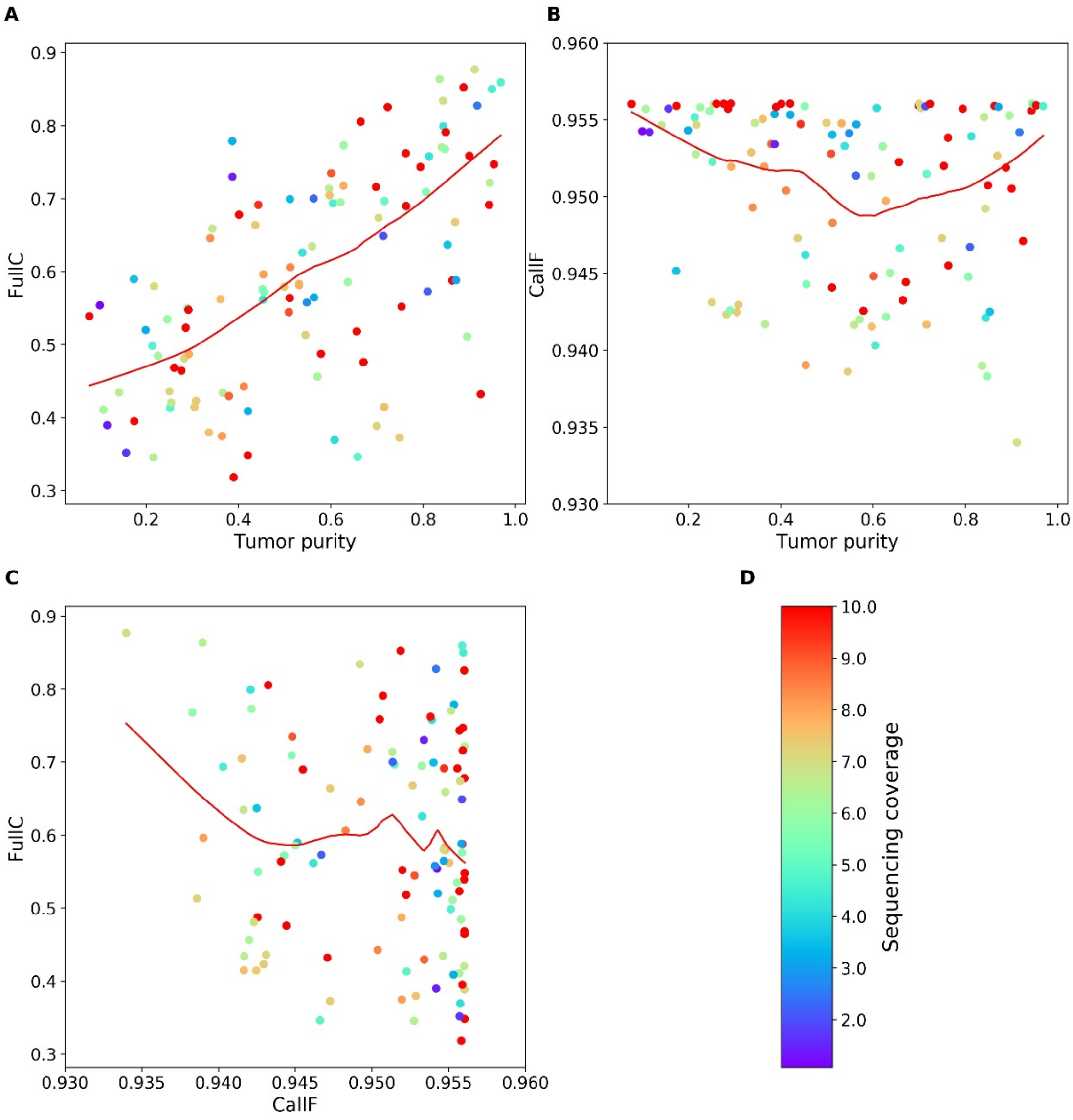


Fig. 3 The performance of Accucopy in predicting TCNs for TCGA samples.

CallF and FullC were calculated to assess the performance of Accucopy. Each dot represents one TCGA sample and is colored according to its sequencing coverage. Each scatterplot is fitted with a redline by loess smoothing. **A.** FullC is between 0.3 and 0.9, strongly dependent on the tumor purity level. **B**. CallF is between 0.93 and 0.96, independent of the tumor purity level, indicating Accucopy predicted copy numbers for almost the entire genome for all analyzed TCGA samples. **C.** FullC is independent of CallF. **D.** The colorbar maps the sequencing coverage of each sample to the color of each dot. The sequencing coverage is set to 10 for samples with sequencing coverage above 10.

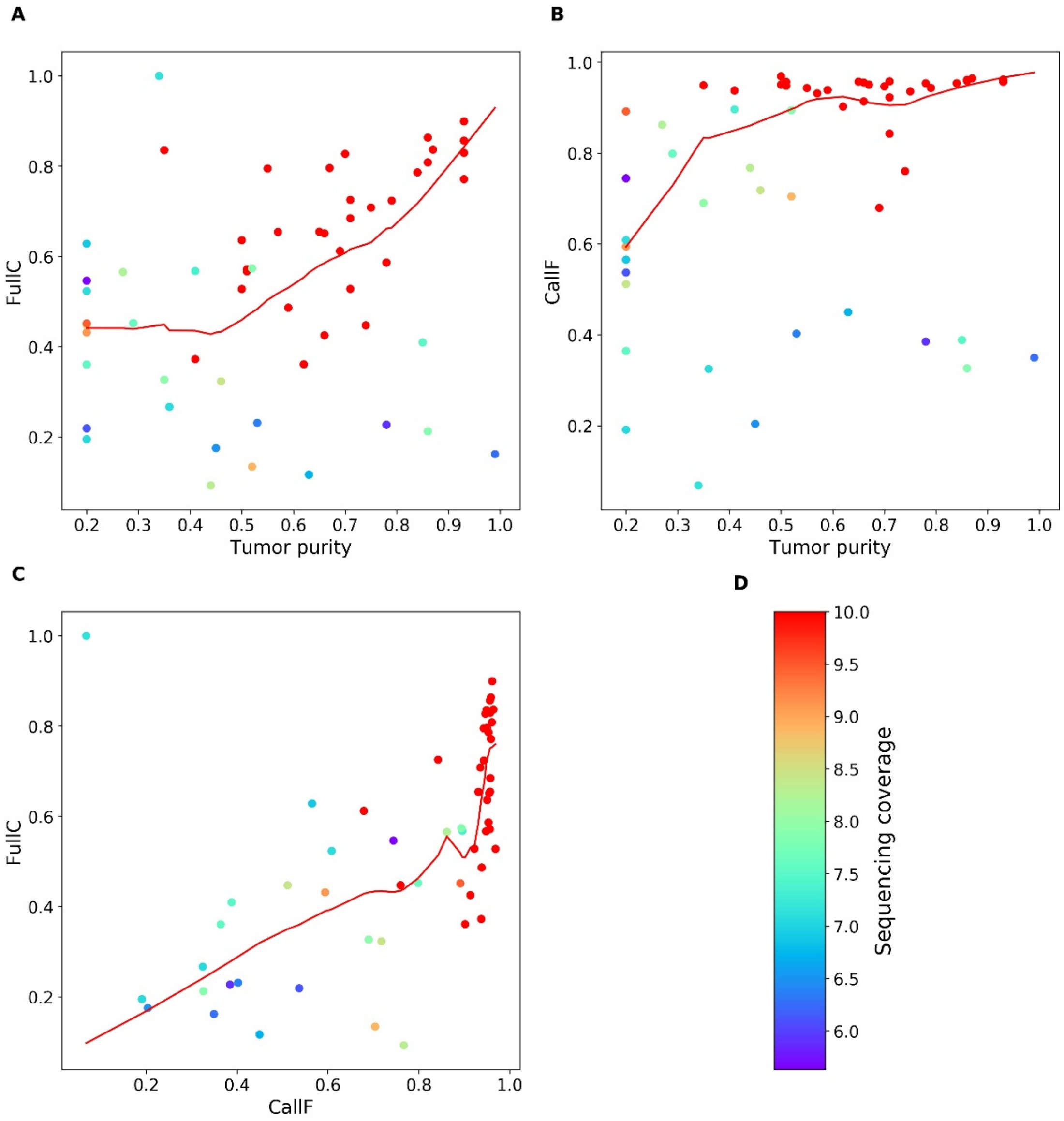


Fig. 4 The performance of Sclust in predicting TCNs for TCGA samples. CallF and FullC were calculated to assess the performance of Sclust. Each dot represents one sample and is colored according to its sequencing coverage. Each scatterplot is fitted with a redline by loess smoothing. **A.** FullC shows a dependency

on both the sequencing coverage and the tumor purity. **B.** CallF is high (~0.9) only for samples with sequencing coverage near or above 10. For samples with lower coverage, Sclust may fail to predict copy numbers for significant portions of their genomes. **C.** FullC is highly correlated with CallF. This suggests the more regions that Sclust fails to predict copy numbers, the less concordant its predicted copy numbers are with the TCGA calls. **D.** The colorbar maps the sequencing coverage of each sample to the color of each dot. The sequencing coverage is set to 10 for samples with sequencing coverage above 10. The colorbar scale starts from around 5 because Sclust failed on samples with sequencing coverage below 5.

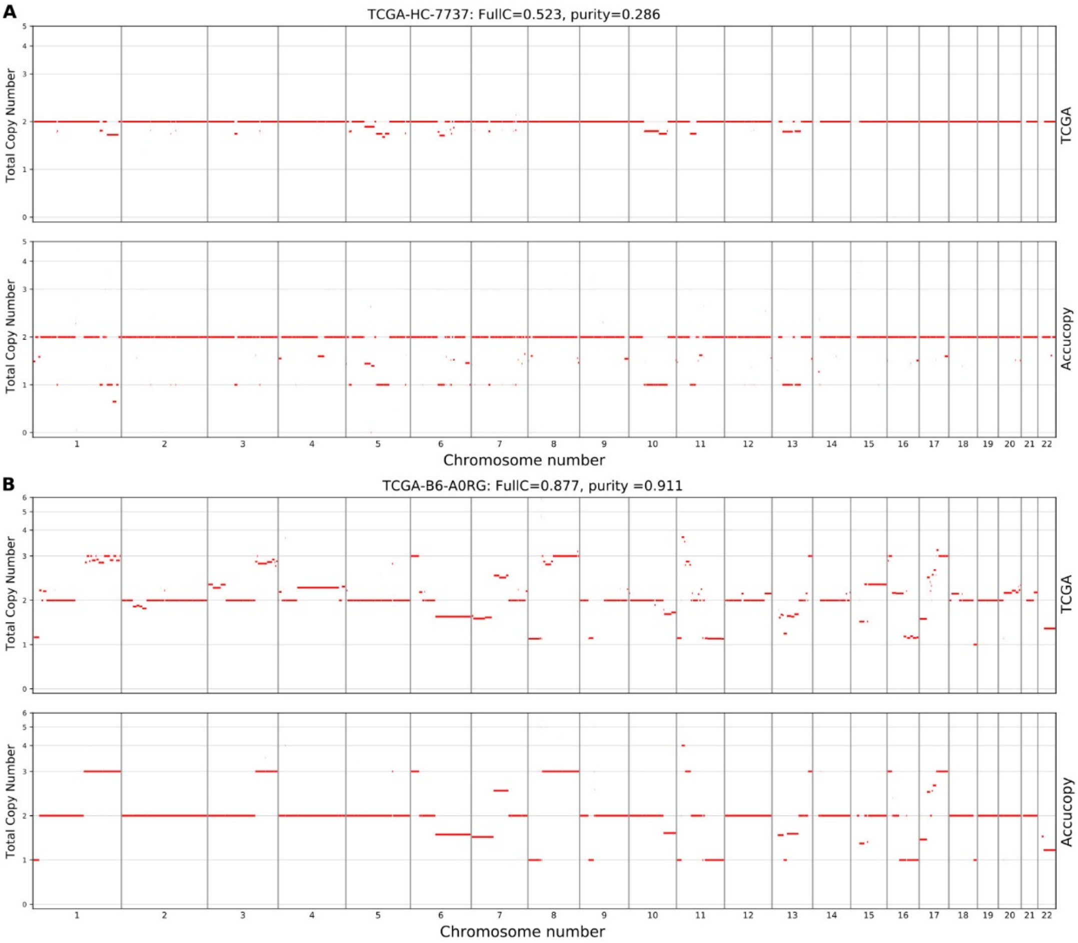


Fig. 5 The loss of power of the TCGA CNA pipeline in low-purity samples.

The copy number of a genomic segment predicted by the TCGA pipeline, which does not model the tumor purity, is a weighted average of its respective copy numbers in the tumor and normal cells. As the purity of a tumor sample declines, the increasing fraction of normal cells will move the predicted average copy number closer to two. Accucopy treats the copy numbers of the tumor and normal cells within one tumor sample as two separate parameters in its model. Panel (**A**) exhibits the copy number profile of a low-purity sample (purity=0.286) predicted by the TCGA pipeline, the upper panel, versus that predicted by Accucopy, the lower panel. The FullC between the two profiles is 0.523. Panel (**B**) is a similar plot to panel A, for a high-purity sample (purity=0.911). The FullC between the TCGA-pipeline and Accucopy predicted CNA profiles is 0.877. In both samples, the segmentations of the genome by the TCGA pipeline and Accucopy are highly similar. In addition, the copy number qualitative predictions (duplication or deletion) for individual segments are also highly similar. However, in the high-purity sample (B), the copy number quantitative predictions of abnormal segments are further away from two and are numerically more concordant with those by Accucopy, manifested by a higher FullC than the low-purity sample (A).

Supplementary information

Additional File 1 Method details of Accucopy

1 Basic definitions

We define the fraction of cancer cells in a tumor sample as the tumor purity γ and the fraction of normal cells is $1-\gamma$. We assume that the ploidy of a normal cell is 2 and denote the average copy number of a cancer cell as the tumor cell ploidy: κ. The tumor sample ploidy ω is a weighted average of the ploidy of normal and cancer cells, expressed in γ and κ as follows:

$$\omega = (1 - \gamma) \times 2 + \gamma \times \kappa \qquad (5)$$

We denote the total copy number (TCN) of a chromosomal segment s of all tumor cells as C_s. Then, the TCN of the same segment for the tumor sample, C_t, is the average TCN of tumor and normal cells in the tumor sample:

$$\mathrm{C_t} = (1 - \gamma) \times 2 + \gamma \times C_s \qquad (6)$$

Note the difference between the tumor cell ploidy and the tumor sample ploidy. The latter includes ploidy contribution from normal cells in a tumor sample while the former is only about tumor cells. The two are identical for a 100% pure tumor. Similarly, the tumor cell TCN of a segment is different from the tumor sample TCN of the same segment. The observed sequencing coverage of a tumor sample should be proportional to the tumor sample ploidy, thus dependent on the tumor purity and tumor cell ploidy, γ, κ.

2 Tumor Read Enrichment (TRE) for a chromosomal segment

Denote the number of reads covering a genomic segment s for a tumor sample and its matching normal sample as $\mathrm{n_t^s}$ and $\mathrm{n_n^s}$, respectively, and a total number of $\mathrm{N_t}$ and $\mathrm{N_n}$ reads for a tumor sample and its matching normal sample. The Tumor Read Enrichment (TRE) for segment bin s, $\mathrm{e_s}$, is defined as follows:

$$\mathrm{e_s} = \frac{n_t^s}{N_t} \Big/ \frac{n_n^s}{N_n} \qquad (7)$$

TRE is a normalized read enrichment of a chromosomal segment in a tumor sample relative to its matching normal sample. Factors that influence both tumor and normal samples, such as the read mappability and the GC biases, are canceled out. To have a better statistical representation, TRE is calculated for each 500bp (roughly the sequencing fragment length) bin throughout the whole genome. A fast version of GADA is applied to segment the entire genome based on calculated TREs.

3 The TRE expectation and the TCN Gaussian mixture model

For a chromosomal segment bin s, assuming independence between the local and global coverage, the expected TRE of a segment bin s can be approximated as follows:

$$E_s = E(e_s) = E\left(\frac{n_t^s}{N_t}\Big/\frac{n_n^s}{N_n}\right) \approx \frac{E(n_t^s)}{E(n_n^s)} \times \frac{E(N_n)}{E(N_t)} \tag{8}$$

We define a few nuisance parameters to help to further derive $\mathrm{E_s}$. The length of segment bin s is $\mathrm{L_s}$. The length of the reference genome, about three billions, is $\mathrm{L_{gw}}$. The genome-wide average sequencing coverage is $\mathrm{V_{gw}^T}$ for the tumor sample and $\mathrm{V_{gw}^N}$ for its matching normal sample. The average sequencing coverage for segment bin s from a tumor sample is $\lambda_s \times \mathrm{V_{gw}^T}$, which multiplies a sequence-specific factor λ_s to the genome-wide sequencing coverage. The average sequencing coverage for segment bin s from the matching normal sample is $\lambda_s \times \mathrm{V_{gw}^N}$. With all these definitions, we can derive the expected TRE, $\mathrm{E_s}$, as a statistic only dependent on tumor purity, γ, tumor cell ploidy, κ, and the TCN of the segment bin in cancer cell, $\mathrm{C_s}$:

$$\begin{aligned}\mathrm{E_s} &= \frac{\mathrm{E(n_t^s)}}{\mathrm{E(n_n^s)}} \times \frac{\mathrm{E(N_n)}}{\mathrm{E(N_t)}} = \frac{\mathrm{C_t} \times \mathrm{L_s} \times \lambda_s \times \mathrm{V_{gw}^T}}{2 \times \mathrm{L_s} \times \lambda_s \times \mathrm{V_{gw}^N}} \times \frac{2 \times \mathrm{L_{gw}} \times \mathrm{V_{gw}^N}}{\omega \times \mathrm{L_{gw}} \times \mathrm{V_{gw}^T}} \\ &= \frac{\mathrm{C_t}}{\omega} = \frac{(1-\gamma) \times 2 + \gamma \times \mathrm{C_s}}{(1-\gamma) \times 2 + \gamma \times \kappa}\end{aligned} \tag{9}$$

The entire segment, s, is assumed to be of the same TCN and thus all observed $\mathrm{e_s}$ should have the same expectation, $\mathrm{E_s}$. Thus, we drop the subscript s of $\mathrm{E_s}$ and add the superscript i to denote the expected TRE for all segments with $\mathrm{TCN} = i$ as $\mathrm{E^i}$:

$$E^i = \frac{(1-\gamma) \times 2 + \gamma \times i}{(1-\gamma) \times 2 + \gamma \times \kappa} \tag{10}$$

For all segments with TCN=$\mathrm{i} + 1$, the corresponding $\mathrm{E^{i+1}}$ is

$$E^{i+1} = \frac{(1-\gamma) \times 2 + \gamma \times (i+1)}{(1-\gamma) \times 2 + \gamma \times \kappa} \tag{11}$$

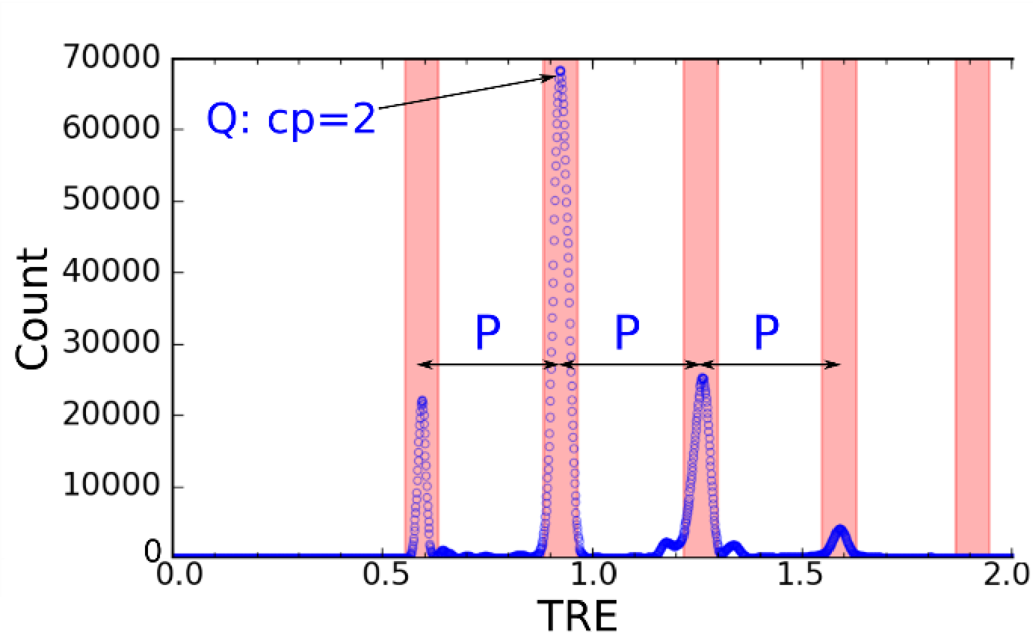


Figure 1 A typical Tumor Read Enrichment (TRE) histogram shows a periodic pattern

Auto-correlation analysis can identify the period of the histogram, P, as the interval between major peaks. Q, one major peak that corresponds to copy-number-two segments, is identified through the Accucopy probabilistic model. Minor peaks between the major ones consist of subclonal segments.

Forms of E^i and E^{i+1} can explain the periodicity we observed from any TRE histogram. We define the period of a TRE histogram, P, as the interval between two copy numbers (**Figure 1**) and its expected value is

$$P = E^{i+1} - E^{i} = \frac{\gamma}{(1-\gamma) \times 2 + \gamma \times \kappa} \tag{12}$$

In a histogram of TREs (Figure 1), the period P is the interval between two adjacent major peaks. Each major peak in a TRE histogram represents one group of clonal segments with the same integral copy number. Usually the period of a tumor sample decreases with low purity or high ploidy.

Further, we define the Normal TRE (NTRE) Q, as the TRE corresponding to segments of copy number 2, then

$$\begin{aligned} Q &= E^{i}|(i=2) = \frac{(1-\gamma) \times 2 + \gamma \times i}{(1-\gamma) \times 2 + \gamma \times \kappa}|(i=2) \\ &= \frac{2}{(1-\gamma) \times 2 + \gamma \times \kappa} \end{aligned} \tag{13}$$

Solving eq. 8 and 9 produces the tumor sample purity γ and the tumor cell ploidy κ in terms of P and Q.

$$\gamma = \frac{2 \times P}{Q}$$
$$\kappa = 2 + \frac{1 - Q}{P} \qquad (14)$$

We model the observed TREs, e_s, as a Gaussian mixture with each component corresponding to TCN $= i$.

$$\mathrm{e_s} \sim \Sigma_{i=0}^{\mathrm{I}} p_i N\left(E^i, \sigma_i^2\right) \qquad (15)$$

where p_i is the mixing parameter (also prior probability) for component i, with $\Sigma_{i=0}^{\mathrm{I}} p_i = 1$, and σ_i^2 is the component variance. The TRE likelihood for all M segments is:

$$L(e; \gamma, \kappa) = \prod_{s=1}^{M} \sum_{i=0}^{I} p_i P(e_s | E^i, \sigma_i^2) \qquad (16)$$

4 Log ratio of Allelic-coverage Ratios (LAR) of HGSNVs

For an HGSNV, denote n_t^R, n_t^A, n_n^R and n_n^A as the read counts for the reference allele (R) and the alternative allele (A) in a tumor (t) and its matching normal (n) samples. Define r as the log-ratio of allelic-coverage ratios (LAR) for an HGSNV:

$$\mathrm{r} = \log\left(\left(\frac{n_t^R}{n_n^R}\right) \Big/ \left(\frac{n_t^A}{n_n^A}\right)\right) = \log(\frac{n_t^R n_n^A}{n_n^R n_t^A}) \qquad (17)$$

The LAR is defined in the same vein as the TRE. The tumor allelic coverage is normalized by that of the matching normal sample to eliminate various sequencing biases: the GC-bias, the reference mapping bias, etc. However, the definition requires all four read counts to be positive as any zero will render the statistic ill-defined, which necessitates an adjustment in calculating its expectation, detailed in the next section.

The variant calling of HGSNVs is carried out at 44 million SNP loci from the 1000 Genomes project using Strelka2. To improve the quality of the final HGSNVs, the tumor and normal samples were called simultaneously by Strelka2, in so-called multi-sample calling, SNPs must be heterozygous in the normal sample, and the coverage of the SNP must be above two in either sample.

5 The LAR expectation, the ASCN Gaussian mixture model, and the EM algorithm

Denote the ASCNs of a tumor cell and its matching normal cell at an HGSNV as (k, l), lower case of L in it, and $(1,1)$, with k and l denoting the major-allele copy number and the minor-allele copy number in the tumor cell respectively. The alternative allele is less likely to be mapped correctly to the reference genome than the reference allele due to the reference bias. Let ϕ denote the reference mapping bias of the reference allele relative to the alternative allele, and typically $\phi > 1$. Hence, the ASCNs of a pure tumor sample and its matching normal one is either $(\phi\mathrm{k}, \mathrm{l})$ and $(\phi, 1)$, or $(\mathrm{k}, \phi\mathrm{l})$ and $(1, \phi)$, depending on if the major allele is the reference allele or not. Taking the tumor purity γ into account, the ASCNs of a tumor sample is either $(\phi(\gamma k + (1-\gamma)\times 1), \gamma l + (1-\gamma)\times 1)$, or $(\gamma k + (1-\gamma)\times 1, \phi(\gamma l + (1-\gamma)\times 1))$, depending on which allele is the reference allele, with the ASCNs of the normal sample unchanged. The sequencing coverage of a segment is proportional to its copy number. Then take expectation of eq. 13 produces the following naïve expectations of LAR, with the reference bias ϕ cancelled out:

$$E(r) = \mu_1^* \quad or \quad \mu_2^*$$

$$\mu_1^* = \log\frac{\gamma k + (1-\gamma)\times 1}{\gamma l + (1-\gamma)\times 1} \qquad (18)$$

$$\mu_2^* = \log\frac{\gamma l + (1-\gamma)\times 1}{\gamma k + (1-\gamma)\times 1}$$

$$\mu_1^* = -\mu_2^* \qquad (19)$$

However, the definition of LAR (eq. 13) precludes HGSNVs with any zero allelic coverage in either sample, which creates a substantial bias not accounted for in the naïve expectation (eq. 14), we model the allelic sequencing coverage as a Poisson distribution and exclude zero-coverage to derive a better expectation of LAR. Denote $\lambda_{k+l} = \lambda_k + \lambda_l$ as the mean total coverage, with λ_k and λ_l being the mean coverage of the major and minor alleles respectively:

$$\lambda_k = \frac{\gamma k + (1-\gamma)}{\gamma(k+l) + 2\times(1-\gamma)} \times \lambda_{k+l} \qquad (20)$$

$$\lambda_l = \frac{\gamma l + (1-\gamma)}{\gamma(k+l) + 2\times(1-\gamma)} \times \lambda_{k+l} \qquad (21)$$

We estimate λ_{k+l} as the median depth of all HGSNVs within a segment in a tumor sample. Denote d^k and d^l as the observed read counts of the major and the minor alleles respectively and each follows a Poisson distribution:

$$d^k \sim Po(\lambda_k),\ \ d^l \sim Po(\lambda_l) \tag{22}$$

Excluding the zero read counts, the adjusted expectation of LAR, μ_1 and μ_2, are as follows:

$$\mu_1 = \frac{\sum_{d^k=1}^{\infty}\sum_{d^l=1}^{\infty}\log\left(\frac{d^k}{d^l}\right)P(d^k|\lambda_k)P(d^l|\lambda_l)}{\sum_{d^k=1}^{\infty}\sum_{d^l=1}^{\infty}P(d^k|\lambda_k)P(d^l|\lambda_l)} \tag{23}$$
$$\mu_2 = \ -\mu_1$$

Figure 2 shows the adjustment greatly improved the fit between the observed mean LAR and the true mean.

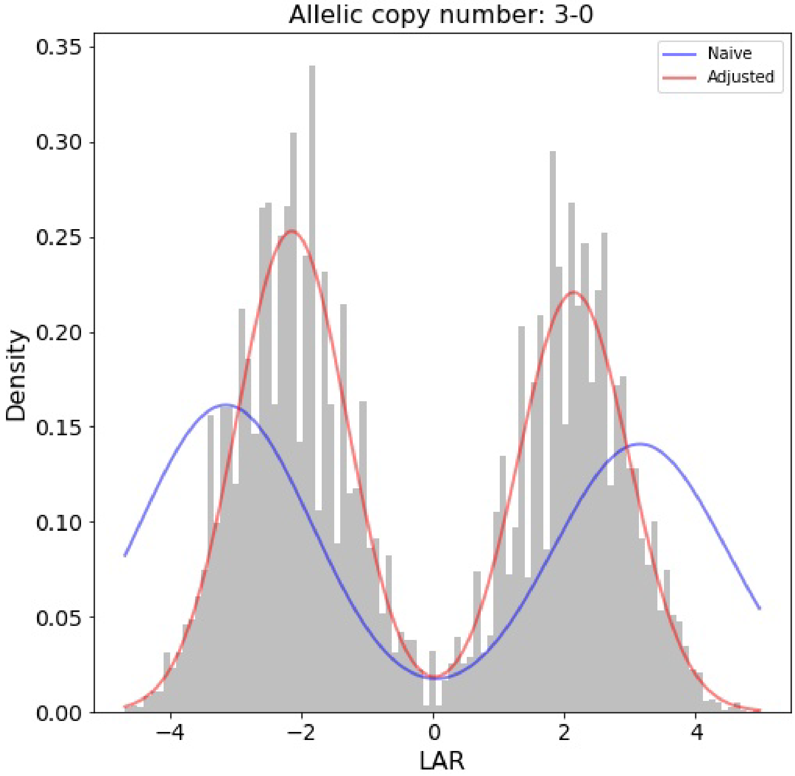


Figure 2 Effect of the adjustment of the expectation of LAR for a segment of HCC1187

This is a histogram of LARs for this segment. The known TCN (Total Copy Number) of this segment is 3. The known ASCN (major allele vs minor allele) is 3-0. The gray bars are the observations. The blue line is the fitted Gaussian mixture distribution based on the naïve expectations of LAR. The red line is the fitted Gaussian mixture distribution based on the adjusted expectations of LAR.

In real data, we have no idea if the expectation of an observed LAR is μ_1 or μ_2 because it is unknown which allele is the major allele. Thus, we adopt a two-component Gaussian mixture model with the two components having an identical variance and their means opposite to each other:

$$\mathrm{r} \sim \Sigma_{m=1}^{2} \alpha_m N(\mu_m, \sigma_s^2) \tag{24}$$

where α_m is the mixing parameter (or prior probability) for component m, with $\alpha_1 + \alpha_2 = 1$, and σ_s^2 is the Gaussian variance of LAR for either component, specific to segment s.

We introduce a missing variable, Δ, that indicates which component an LAR belongs to, and apply the Expectation-Maximization (EM) algorithm to estimate α_m and σ_s^2. Given a segment s with TCN = C_s, which contains N^s LARs, for every possible ASCN combination $(k,l){:}\, k = C_s - l, l = 0,1,\cdots,\lceil C_s/2 \rceil$, μ_m is calculated according to eq. 19, then the E-step computes the conditional probability of one LAR belonging to either component:

$$P\left(\Delta_i | r_i; \sigma_s^{2(g)}, \alpha_1^{(g)}, \alpha_2^{(g)}\right) = \frac{\alpha_m^{(g)} P\left(r_i | \mu_m, \sigma_s^{2(g)}\right)}{\Sigma_{m=1}^{2} \alpha_m^{(g)} P\left(r_i | \mu_m, \sigma_s^{2(g)}\right)} \tag{25}$$
$$\Delta_i = 1,2; \quad i = 1,2,\ldots,N_s$$

The M-step updates these parameters:

$$\begin{aligned} \hat{\sigma}_s^2 &= \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{\Delta_i = 1,2} P\left(\Delta_i | r_i, \sigma_s^{2(g)}, \alpha_1^{(g)}, \alpha_2^{(g)}\right) (r_i - \mu_m)^2 \\ \hat{\alpha}_1 &= \frac{1}{N_s} \sum_{i=1}^{N_s} P\left(\Delta_i = 1 | r_i, \sigma_s^{2(g)}, \alpha_1^{(g)}, \alpha_2^{(g)}\right) \\ \hat{\alpha}_2 &= 1 - \hat{\alpha}_1^{(g)} \end{aligned} \tag{26}$$

The E-step and M-step are iterated until convergence and we calculate the LAR likelihood for ASCN (k,l) as follows:

$$L(r; k, l) = \prod_{i=1}^{N_s} \sum_{m=1,2} \alpha_m P(r_i | \mu_m, \sigma_s^2) \tag{27}$$

The EM algorithm is applied to solve optimal parameters and derive the corresponding likelihood for every possible ASCN combination (k,l). The ASCN estimate $(\hat{k}, \hat{l})$ for segment s is the one with the maximum likelihood.

$$(\hat{k}, \hat{l}) = \arg\max_{k,l} \log L(r; k, l) \tag{28}$$

In the next step, the maximum LAR likelihood of all segments will be combined with the TRE likelihood via the Bayesian Information Criterion (BIC) to determine the most likely tumor purity, tumor ploidy, TCNs and ASCNs of all chromosomal segments.

6 BIC for the combined model and optimization

To avoid model overfitting, we adopt the Bayesian Information Criterion (BIC).

$$\begin{gathered} BIC(e, r; \gamma, \kappa, k, l) = -2\log L(e; \gamma, \kappa) - 2\log L(r; k, l) \\ +I \times \log M + J \times \log N \end{gathered} \tag{29}$$

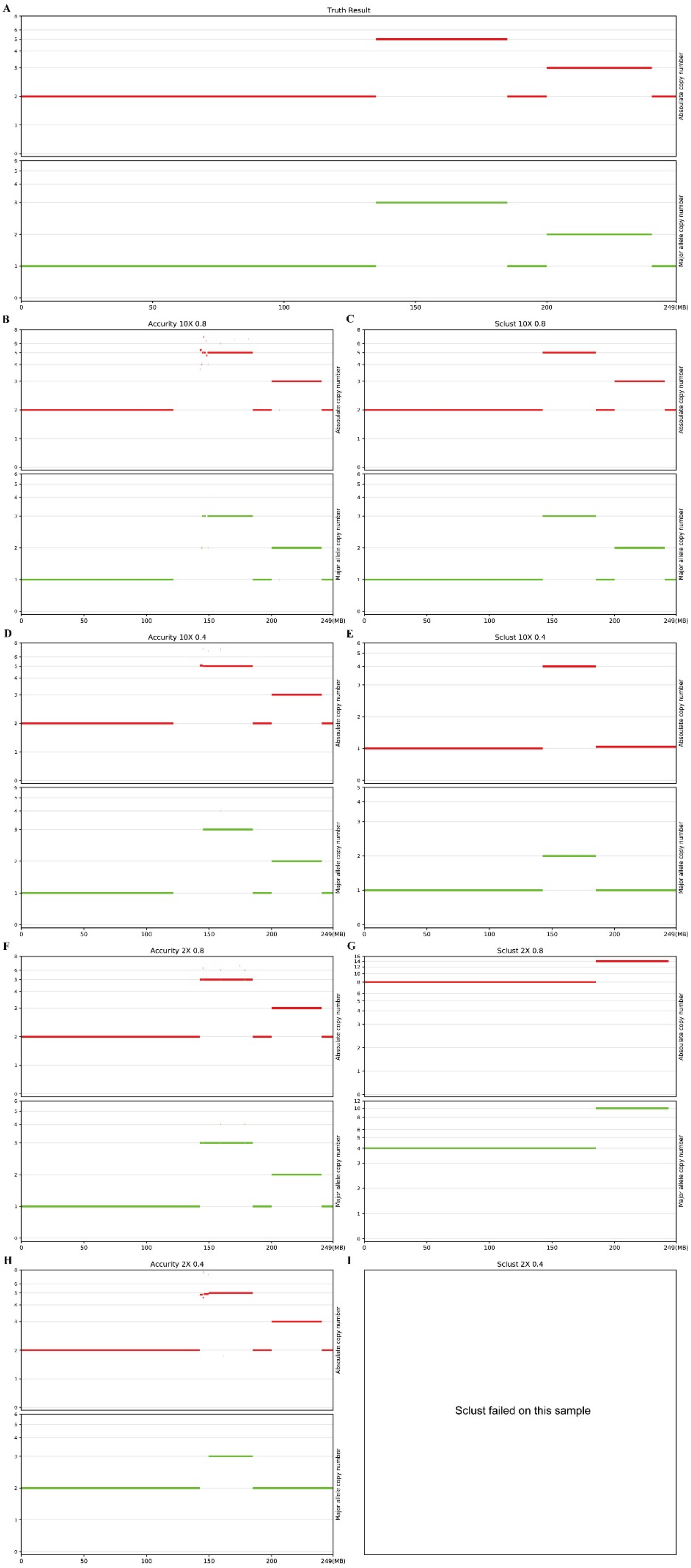
where I is the total number of potential TCNs, M is the total number of observed TREs, J is the total number of potential ASCNs, and N is the total number of HGSNVs.

Instead of searching through the infinite range of tumor purity $\gamma \in [0,1]$, the tumor cell ploidy $\kappa \in [0, \infty]$, all possible TCNs and ASCNs, an optimization scheme that leverages the periodic TRE pattern is adopted. Accucopy first uses an autocorrelation analysis that discovers candidates for P and Q, which are equivalent to γ and κ, as shown in eq. 10, and finds the minimum BIC score among these candidates only.

To reduce the noise in the TRE distribution, Accucopy applies a kernel smoothing (1D Gaussian) before the autocorrelation analysis. The top two lags identified in the auto-correlation analysis form the candidates of P. Given a candidate P, Accucopy further identifies major peaks in the TRE distribution that are P apart, which represent clonal segments of integral copy numbers, and filters out segments that do not belong to any major peak, which are classified as subclonal segments. The TREs of all the major peaks become the candidates for Q. Given a pair of candidate P and Q, the TCN of each clonal segment is determined, the TRE likelihood is computed, and the EM algorithm is carried out to compute the most likely ASCN of this segment and its LAR likelihood. The TRE and LAR likelihoods are then combined into the BIC score. The best estimates of purity and ploidy $(\hat{\gamma}, \hat{\kappa})$, TCNs, and ASCNs $(\hat{k}, \hat{l})$ are obtained by minimizing the BIC score:

$$(\hat{\gamma}, \hat{\kappa}, \hat{k}, \hat{l}) = \arg\min_{\gamma,\kappa,\mu,k,l} BIC(e, r; \gamma, \kappa, k, l) \tag{30}$$

Additional File 2:



Supplementary Figure 1. Copy number profile of chr1 on partial simulation data. X-axis is the chromosomal position and Y-axis is the copy number. Each figure has two panels. The top is the profile of absolute copy number and the bottom is the profile of major allele copy number. **A.** The truth profile of chr1. **B-I.** The copy number profile given by different methods on different samples. The title of each figure has three keys split by space, which indicate the method name, sample coverage and sample purity respectively.