Genome Biology

SOFTWARE

Open Access

Tumor ploidy determination in low-pass whole genome sequencing and allelic copy number visualization using the Constellation Plot

Sarah H. Johnson[1,2], James B. Smadbeck[1,2], Roman M. Zenka[3], Michael T. Barrett[4], Athanasios Gaitatzes[1,9], Arnav Solanki[5], Angela B. Florio[1,6], Mitesh J. Borad[4], John C. Cheville[1,7] and George Vasmatzis[1,8*]

*Correspondence:
Vasmatzis.George@mayo.edu

[1] Biomarker Discovery Program, Mayo Clinic, Rochester, MN 55905, USA
[2] Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
[3] Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA
[4] Hematology/Oncology, Mayo Clinic, Scottsdale, AZ 85259, USA
[5] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA
[6] Mayo Clinic Graduate School of Biomedical Sciences, Rochester, MN 55905, USA
[7] Anatomic Pathology, Mayo Clinic, Rochester, MN 55905, USA
[8] Department of Molecular Medicine, Mayo Clinic, Rochester, MN 55905, USA
[9] Center for Digital Health, Mayo Clinic, Rochester, MN 55905, USA

Abstract

Ploidy determination across the genome has been challenging for low-pass-WGS tumor-only samples. We present BACDAC**,** a method that calculates tumor ploidy down to 1.2X effective tumor coverage. Allele fraction patterns displayed in the Constellation Plot verify tumor ploidy and reveal subclonal populations. BACDAC outputs a metric, $2N^{+}LOH$, that when combined with ploidy better distinguishes near-diploid from high-ploidy tumors. Validated using TCGA, BACDAC had good agreement with other methods and 88% agreement with experimental methods. Discrepancies occur mainly when BACDAC predicts diploidy with subclones rather than high-ploidy. Applied to 653 low-pass-WGS samples spanning 12 cancer subtypes, BACDAC calls 40% as high-ploidy.

Keywords: Next generation sequencing, Ploidy, Whole genome doubling, Loss of heterozygosity

Background

Aneuploid tumor genomes, a hallmark of cancer, typically contain numerous large deletions and gains and are often associated with whole genome doubling (WGD). This imbalance of genomic material impacts both diagnosis and treatment planning, as therapeutic options depend on correct identification and assessment of ploidy-associated phenomena including biallelic gene inactivation, gene amplification, subclonality, and chromosome instability.

Tumor ploidy is traditionally assessed through experimental methods such as karyotyping, a literal count of chromosomes in actively dividing cells, or DNA content flow cytometry. Computational methods to assess ploidy from single nucleotide polymorphisms (SNPs), aCGH arrays, and deep/high-coverage whole genome

sequencing (WGS) have evolved over the past decade, examples include ASCAT [1], ABSOLUTE [2], FACETS [3], HATCHet2 [4], and CLONETv2 [5]. However, ploidy analysis does not yet exist for low-pass (low-coverage, shallow) tumor WGS (lpWGS) with no patient matched normal sample. Moreover, there is no low-pass option that provides a visual representation of allele-specific copy number segmentation, which is adequate for assessment of the ploidy status and its implications for clinical diagnosis and treatment.

Many low-pass and/or low tumor WGS applications exist because until recently, high-coverage sequencing has been cost-prohibitive. Long insert libraries such as mate-pair sequencing (MPseq) have been widely used by us and others to identify somatic structural variants, including copy number and chromosomal rearrangements. Although MPseq achieved adequate bridged coverage for these purposes (>30X), it lacked the (base) coverage needed to detect mutations and their allelic fractions. Newer WGS applications include combining large high-coverage gene panels with lpWGS. WGS of cell-free DNA from liquid biopsies is an emerging technique for deciphering tumor alterations in the blood of patients with solid tumors. However, tumor purity is typically limited making ploidy and direct allelic content determination impossible.

The prediction of ploidy and copy number in complex tumor genomes is complicated by WGD, loss of heterozygosity (LOH), and subclonal cell populations (subsets of cells with additional genomic variants compared to the main clone). Whole genome copy number variant (CNV) prediction algorithms typically use window-based read depth distributions from tumor sequencing [6]. Then, the dominant read depth may be assigned as diploid with copy number extrapolated to the remaining read depth peaks. Although this approach is suitable for low-complexity genomes, these methods fail in complex aneuploid tumors with many CNVs and potential WGD. Advanced methods have been developed to incorporate allelic content with CNV prediction. However, these methods fail when tumor purity and sequencing coverage are low, due to reduced sensitivity and accuracy. We hypothesize that the allelic content could be assessed even via low-pass genome sequencing, improving CNV prediction and ploidy calculations for these samples.

To address this unmet need, we developed BACDAC a workflow based on Binomial distribution statistics of common SNPs to calculate Allelic Content, a Discretization Algorithm to translate read depth to allele-specific copy number and a Constellation Plot to visualize the data. BACDAC is a process that combines novel algorithmic methods with a graphical interpretation to report tumor ploidy relative to the haploid genome and tumor purity from lpWGS. First, allele-specific copy number is determined using an approximation of allele heterozygosity, defined here as the Heterozygosity Score (hetScore). The hetScore is based on biallelic single nucleotide polymorphism (SNP) content across large regions, similar to B-allele frequency, but is computationally valid and meaningful in a lpWGS tumor sample without the need for a matched normal sample. Second, the accuracy of the resulting tumor ploidy and purity was assessed via the Constellation Plot. This two-dimensional plot, of hetScore versus copy number for all genomic segments is an approachable and intuitive method for visualizing allele-specific copy number and patterns of aneuploidy. Similar figures have been published [3, 4, 7] and used in previous studies to decipher subclonal populations.

BACDAC was compared and validated with published methods. We also present a two-dimensional threshold, $2N^{+}$LOH, based on LOH content and ploidy, to distinguish diploid (including near-diploid) tumors from high-ploidy tumors. This approach was applied to multiple cancer types with variable ploidies, and the fraction of tumors designated high-ploidy is presented.

Results

From our internal collection of 885 lpWGS tumor samples, BACDAC was successfully applied to 73% of the samples (653/885) (Fig. S1A). The coverage and tumor purity for the 653 samples is shown in Additional file 1: Fig. S1B. BACDAC's ability to predict ploidy depends on both coverage and purity. To determine the minimum input requirements for coverage and tumor purity, we evaluated the "effective tumor coverage" (ETC) of each sample, which is the product of coverage multiplied with tumor fraction. A histogram of ETC for the 653 cases is shown in Additional file 1: Fig. S1C. The lowest ETC for reliable ploidy prediction was 1.2X. Of the 885 samples analyzed, 232 samples did not pass minimum requirements for ploidy prediction: 120 where below 1.2X ETC, 101 did not have any CNVs detected, possibly because they were driven by mutations and balanced rearrangements or because the tumor percentage was too low and in the remaining 11 samples, the read depth was too noisy due to inadequate library preparation methods or degraded tissue.

To demonstrate the philosophy of BACDAC, we present several representative examples, starting with a low-complexity tumor, as shown in Fig. 1. This diploid, single-clone tumor was sequenced at 5X coverage with 92 million fragments. The read depth, the number of reads per 30 kb window, was plotted linearly by chromosome (Fig. 1C). While most of the genome is diploid, numerous whole and large chromosomal gains and a large deletion on chromosome 9 are evident.

The read depth distribution, Fig. 1A, shows that the distances between successive clonal peaks are almost equal. BACDAC's main assumption is that if read-depths (NRD) can be discretized in equal intervals, then they are clonal. Therefore, we developed a discretization grid-based algorithm (see Methods) to align peaks to an equally spaced (regular) one-dimensional linear grid. When applied to this sample, the peak-to-grid algorithm assigned all three peaks to the grid. Then, based on allelic content, the algorithm assigned the dominant peak (gray) as 2N, then the upper peak was incrementally assigned as 3N (blue) and the lower peak was assigned as 1N (red). Tumor ploidy and purity were then calculated from the average copy number assigned to each genomic segment (see Methods), resulting in an overall tumor ploidy of 2.3N and a tumor purity of 74%.

The hetScore for the same sample is shown linearly by chromosome in Fig. 1D. As described in the Methods, hetScore is not true allelic content but an approximate measure that correlates with allele fraction. Note that (1) hetScore = 1 for all balanced heterozygous regions, 2N(1:1), 4N(2:2), etc., denoted as "copy number N (major:minor allele)," where major is the more prevalent allele and minor is the less prevalent allele, and (2) the hetScore corresponding to a nonheterozygous allele fraction decreases from 1, at a rate dependent on tumor purity and depth of coverage. The hetScore for a majority of the diploid regions in this sample hovered near 1, indicating a balanced

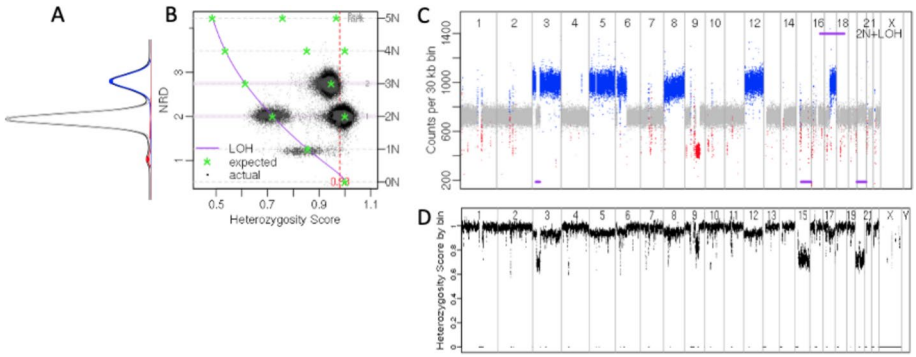

Fig. 1 Interpretive plots of a low-complexity single-clone tumor (coverage 5X, ploidy 2.3N, tumor purity 74%): read depth with read depth distribution, Heterozygosity Score (hetScore), and Constellation Plot. **A** Read depth distribution. **B** The Constellation Plot combines allele fraction, read depth, copy number, and tumor purity assessment in one view. The normalized read depth (NRD), where the read depth of the diploid peak is NRD = 2, is indicated on the left *y*-axis, and the corresponding copy number, as determined by BACDAC's grid-to-peak and ploidy algorithm, is indicated on the right *y*-axis. Cloud clusters are composed of (1) black dots for each 30 kb segment spanning the genome, and (2) grey symbols for each segment of similar read depth greater than a specified length (usually 5 Mb, not less than 3.5 Mb). Stars (green asterisks) are located at the expected theoretical hetScore for each allele fraction for a given tumor ploidy and purity solution. For reference, a Constellation Plot with labeled allele fractions for each star is shown in Additional file 1: Fig. S2A. Allele fractions with LOH (minor allele = 0) are the left-most stars for each copy number and are connected by a purple line. The heterozygous allele fractions (1:1, 2:2, etc.) are the right-most stars for each even-numbered copy number and have a hetScore = 1. The cloud at position hetScore (*x*-axis) = 0.72, and NRD (*y*-axis) = 2 revealed diploid segments of chromosomes 3, 15, and 20 with 2N copy-neutral loss of heterozygosity. **C** Read depth, the number of reads per 30 kb window, for chromosomes 1–22, X, and Y. Gain = blue, diploid = gray, loss = red. The purple line marks areas of $2N^{+}$LOH. **D** The hetScore, the measure of biallelic SNP content, for chromosomes 1–22, X, and Y; hetScore 1 for balanced allele fractions, and decreases with decreasing heterozygosity, which includes unbalanced allele fractions

heterozygous allele fraction of 2N(1:1). The slight decrease in hetScore for the 3N and 1N segments reflects a decrease in heterozygosity, due to unbalanced allele fractions of 3N(2:1) and 1N(1:0) respectively. The large decrease in hetScore in the diploid regions of chromosomes 3, 15, and 20 corresponds to LOH and an allele fraction of 2N(2:0).

Interestingly, when plotted together in a two-dimensional plot, hetScore vs copy number, named here the "Constellation Plot" (Fig. 1B), four clusters, or clouds, form, in this sample each cloud represents an allele fraction that corresponds to an allele-specific copy number state. The three most prominent clouds are the 2N(1:1), 3N(2:1), and 2N(2:0) clouds, and the fourth, fainter cloud represents the 1N(1:0) deletions. To facilitate interpretation of the Constellation Plot, green stars (asterisks) mark the hetScore for each expected allele-specific copy number, calculated as a function of tumor purity and coverage. While the linear genome-wide read depth view (Fig. 1C) shows the existence of a 3N and 1N population, it is not until hetScore is combined with read depth in the Constellation Plot that the 2N copy-neutral LOH (cnLOH) population in this sample is clearly revealed.

The constellation plot is particularly useful for visualizing aneuploid tumors and identifying subclonal cell populations

Advanced tumors with high aneuploidy often have many (> 4) copy number populations and subclonal cell populations, complicating the determination of the copy number

state. This complexity will be reflected in the cloud patterns in their Constellation Plots. To demonstrate the variety of cloud patterns that may emerge, three higher aneuploidy and higher ploidy tumors are shown in Fig. 2 and Additional file 1: Fig. S3. The samples PT58184 and PT58197 had 9X and 5X coverage, respectively. Both had ploidy = 3.1N but distinct variation in the amount of heterozygosity and copy number states within each tumor, as demonstrated by the numerous clouds representing individual allele fractions. PT58184 has only five allele fractions spanning 2N to 6N; note the presence of the 2N(2:0) allele fraction but not an 2N(1:1) fraction (Fig. 2A). PT58197 has 11 allele fractions spanning 1N to 7N. Of these 11 allele fractions, LOH was present in four, while only two were completely heterozygous: 2N(1:1) and 4N(2:2) (Fig. 2C). AG74002, the third sample, which was sequenced at 7X coverage, had a ploidy of 4.7N. The Constellation Plot revealed 15 allele fractions and a strong right-sided orientation of the clouds, including heterozygous allele fractions of 1:1, 2:2, 3:3, 4:4 and 5:5 respectively for the even-numbered copy numbers 2N-10N (Fig. 2E). Additional file 1: Fig. S4 shows additional Constellation Plots illustrating the variety of allele fraction patterns in tumor genomes.

In addition to highlighting the presence or absence of LOH in a genome, the Constellation Plot highlights the presence or absence of subclonal cell populations. Subclonality

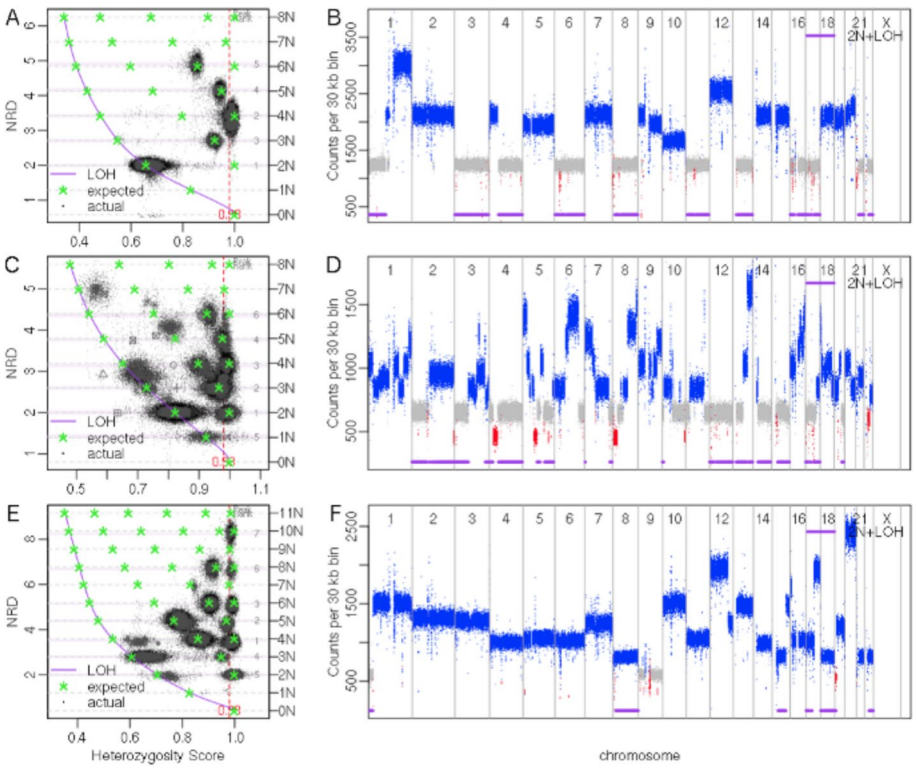

Fig. 2 Constellation Plot and read depth for three high complexity tumor samples. Read depth plots: gain = blue, diploid = gray, loss = red, purple line marks areas of LOH. Coverage, ploidy, tumor purity, and key genomic features are listed for each tumor sample: **A**, **B** PT58184—9X, 3.1N, 71%; 5 allele fraction clouds, all 2N segments are 2NcnLOH, 2N(2:0) with no 2N(1:1) present. **C**, **D** PT58197—5X, 3.1N, 59%; 11 allele fraction clouds, chromosome 2p is 2N cnLOH, while 2q is a subclonal 3N LOH gain. **E**, **F** AG74002—7X, 4.7N, 79%; 15 allele fraction clouds, all clonal, multiple allele fractions present in each of six copy number states

is indicated when clouds fall between the stars of the Constellation Plot. The samples in Fig. 1A and Fig. 2E have no subclonal cell populations, as all the clouds overlap a star, but Fig. 2C shows a subclonal 3N gain of chromosome 2q, indicated by the cloud falling between the 3N and 4N LOH stars. Sample TCGA-14–1402-02A (Fig. 3) has a subclonal gain on chromosome 9 with multiple subclonal losses on chromosome 3 and chromosome 1. Additional examples of subclonal identification using the Constellation Plot are shown in Additional file 1: Fig. S5- S6.

Calculation of valid ploidy solutions

Calculation of the average ploidy requires proper assignment of the copy number state across the genome. As described by Tarabichi et.al. [8], a major limitation of CNV calling is the assumption of clonality. BACDAC's main assumption is that if peaks from the read depth distributions fall on an equidistant grid, that is, if read depth deviations are spaced at equal intervals in linear space, they are likely clonal. In samples like Fig. 2E, clonality is the likely scenario, as the clouds align well on the equidistant grid on the *y*-axis. Once the grid is set by the algorithm, in the discretization step, the copy number state is derived by choosing the simplest solution (the solution with lower average ploidy) where clouds align best with stars in both dimensions. The Constellation Plot can confirm the validity of a tumor ploidy and purity solution and provide confidence and context to the solution. A solution is confirmed when a majority of the clouds are centered over a star and no cloud is positioned left of the purple LOH line. This last condition can be violated in the case of runs of homozygosity (see Limitations section). A Constellation Plot showing a valid and nonvalid solution for the same sample is shown in Additional file 1: Fig. S6. Both solutions had a ploidy of 1.9N but different tumor purities of 59% and 40%. The nonvalid solution was the result of incorrectly assigning a subclonal population as a clonal population, which altered the peak-to-grid assignment.

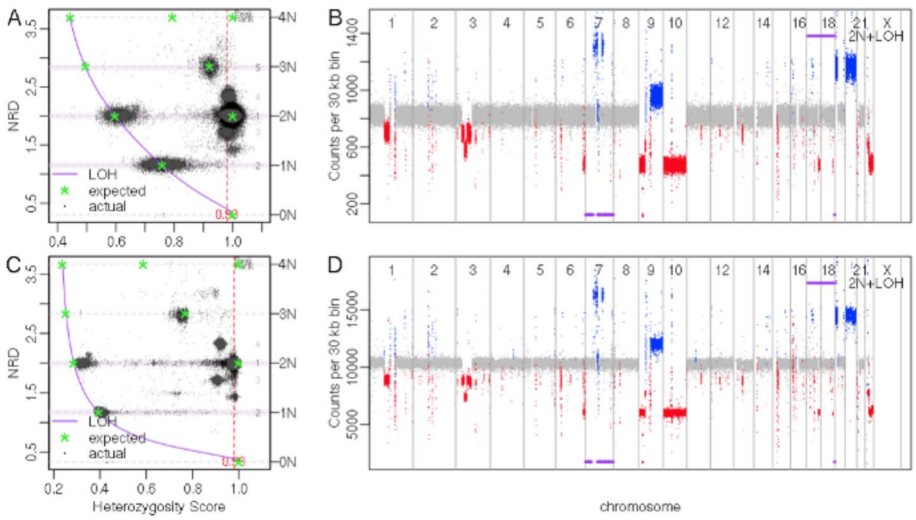

Fig. 3 A representative TCGA sample, TCGA-14–1402-02A, was executed at low (5X) and high (68X) coverage. Genomic highlights observed from the Constellation Plots: chromosome 7—2NcnLOH; chromosome 9—subclonal 2N gain; chromosome 3—multiple subclonal 2N losses; chromosome 1—subclonal 2N loss. **A**, **B** Low-coverage (5X) Constellation Plot and read depth plot. **C**, **D** High-coverage (68X) Constellation Plot and read depth plot. Read depth plots: gain = blue, diploid = gray, loss = red, purple line marks areas of LOH

The BACDAC predictions agree with the results from published methods and gold standard experimental data

BACDAC's accuracy was validated on 63 TCGA (The Cancer Genome Atlas) high-coverage WGS datasets. The influence of sequencing coverage on ploidy and purity predictions was also tested by down-sampling the TCGA datasets (see Methods) to simulate low-coverage. The Constellation Plot and read depth plot for a representative sample, for both high-coverage (68X) and low-coverage (5X) executions, are shown in Fig. 3. Despite the extreme difference in coverage, BACDAC reports the same tumor ploidy and purity (±2%). Both Constellation Plots reveal subclonal gains and losses, as well as the 2N LOH segments on chromosome 7. As expected in low-coverage, the noise increases and is reflected in the larger diameter clouds of the low-coverage data. Additional file 1: Fig. S5 shows additional Constellation Plots of TCGA samples at both high and low-coverage. BACDAC's ploidy predictions between high- and low-coverage were consistent with near-perfect concordance (cor coef = 0.98) for the 63 samples (Fig. 4A).

BACDAC's predictions of tumor ploidy and purity on the TCGA samples were then compared with published results. Ploidy prediction for 24 of these samples were available from a consensus strategy of multiple NGS-based methods that used the high-coverage data, [8], here after referred to as "Dentro." BACDAC ploidy prediction had near-perfect concordance (cor coef = 0.98) with these Dentro results (Fig. 5A). Additionally, ploidy predictions from corresponding SNP-arrays were available from ASCAT2 for all 63 samples (downloaded from the Genomic Data Commons Portal, https://portal.gdc.cancer.gov/) and for 52 samples from ABSOLUTE [9]. BACDAC's predictions had better concordance with ABSOLUTE than with ASCAT2 (Fig. 5B, C). Discrepancies mainly occurred when BACDAC predicted a near-diploid solution rather than a high-ploidy solution. For all such cases, BACDAC detected subclonal populations, as shown in several example samples in Additional file 1: Fig. S5.

We then compared BACDAC with more recently published predictive methods that use WGS, such as FACETS, ASCAT3, and HATCHet2. The comparisons included both the original high-coverage (Fig. 5D–F) and the simulated low-coverage TCGA datasets (Fig. 5H, I). Since these methods require normal data for genotyping, we included the corresponding normal datasets without down-sampling. While concordance of BACDAC ploidy predictions to ASCAT3 and FACETS was good for high-coverage, concordance dropped in low-coverage executions. This drop in concordance is because ploidy predictions by ASCAT3, FACETs, and HATCHet2 are inconsistent

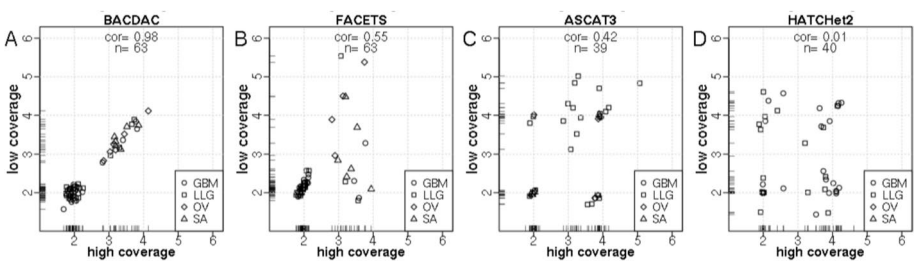

Fig. 4 Intra-method comparison of ploidy prediction for high-coverage and low-coverage sequencing. Ploidy results for 4 methods **A** BACDAC, **B** FACETS, **C** ASCAT3, and **D** HATCHet2

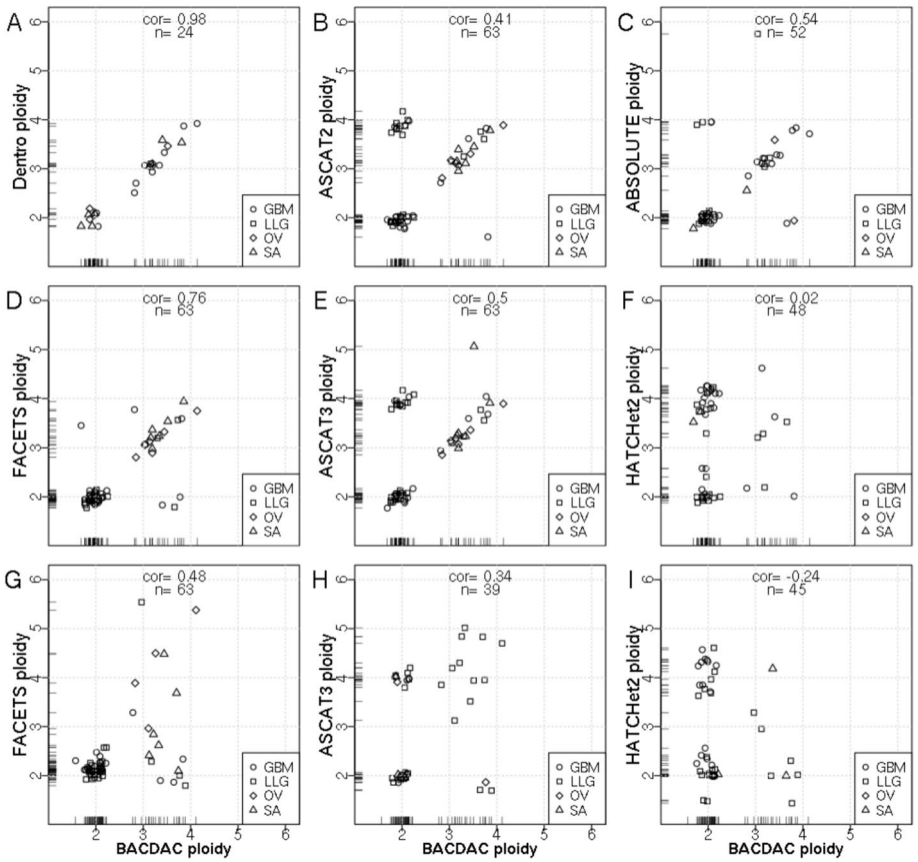

Fig. 5 Inter-method comparison of ploidy prediction for 63 TCGA samples. Comparison of BACDAC ploidy to published results: **A** Dentro consensus, **B** ASCAT2, and **C** ABSOLUTE; BACDAC and high-coverage WGS (~ 50X) executions of **D** FACETS, **E** ASCAT3, and **F** HATCHet2. BACDAC and simulated low-coverage WGS (~ 5X) executions of **G** FACETS, **H** ASCAT3, and **I** HATCHet2

between high- versus low-coverage executions (Fig. 4). As mentioned previously, BACDAC had near-perfect concordance (cor coef. = 0.98) between high- and low-pass coverage. However ASCAT3 and FACETS had low intra-method concordance (cor coef. = 0.42 and 0.55 respectively), and HATCHet2 had no concordance (cor coef. = 0.01) (Fig. 4B–D).

The inter-method discrepancies mainly occurred when BACDAC predicated a near-diploid solution, contrary to the other methods. While BACDAC could be at risk of undercalling higher ploidy (see note in Additional file 1: Fig. S7), BACDAC appropriately predicted subclonal populations which resulted in the lower ploidy solution, where as ASCAT3 ($n = 14$) and FACETs ($n = 2$) predicted ploidy > 3.5N not identifying subclonal populations. Detecting and reporting subclonal cell populations is not a trivial task. It is possible that the subclonal populations were misinterpreted as clonal in SNP-array data and by the other WGS methods, thus resulting in the higher ploidy call. Furthermore, ASCAT2 and ASCAT3 predicted ploidy > 2.7N in a greater percentage of brain samples (20/51, 39% and 22/51, 43% respectively) than expected for WGD in GBM and LGG (GBM, 11.4% [10]; GBM, ~ 25% [2]; GBM, 4%; and LGG, 3% [11]). Thus, it is reasonable to suggest that the discrepancies are more likely the result of ASCAT over-calling higher ploidy rather than BACDAC undercalling higher ploidy.

The pairwise comparisons reported above, while useful to understand discrepancies, do not provide an overall measure for accuracy such as a comparison to a gold standard. Since ground-truth experimental data are not available for these samples, we defined a reference value for tumor ploidy and purity, to approximate a gold-standard value. This reference value was established as the consensus value published in Dentro [9] for the 24 available samples. For the remaining 39, we used the median of the available high-coverage results from the five WGS methods: BACDAC, ABSOLUTE, ASCAT3, FACETS, and HATCHet2. Results are summarized by color with grey representing minimal difference between a method's result and the reference value (Fig. 6). Compared to the other methods, BACDAC had the smallest average difference for both ploidy and purity (Table 1) and agreed with the reference value in the majority of samples. FACETS was quite good in predicting ploidy but suffered in purity predictions (Table 1).

We also tested BACDAC using HCC1395, a cell-line with known ploidy (ploidy = 2.8N), and is available for the purpose of validating new methods and facilitating method comparisons in a hyper-diploid cancer genome [12–14]. Available sequencing included pooled samples of HCC1395 and HCC1395BL at ratios to simulate a tumor purity of 100%, 75%, 50%, 25%, and 10%. We down-sampled this data to simulate low-coverage sequencing data, then ran BACDAC without a normal, and ASCAT3, FACETS, and HATCHet2 with the matched normal (13 × coverage). All methods were run with the same read-pair input, mapping and coverage. The predicted ploidy for each method and sample is shown in Additional file 1: Table S1. BACDAC correctly predicted ploidy for 4/5 samples (ploidy ± 0.1). ASCAT3 failed to run, presumably because of the low coverage of both tumor and normal samples. FACETS correctly predicted ploidy for only 1/5 samples (SPP_GT_3-1_1, expected purity 75%). HATCHet2 was not able to predict correctly for any of the samples.

We further explored ploidy prediction discrepancies by analyzing the F1 scores for gains and losses (Additional file 1: Table S2-S3). Understandably, all algorithms did worse in losses where the coverage was lower. In agreement with previous publications [14], HATCHet2 demonstrated the lowest concordance for losses.

There is a general agreement in the literature that it is important not to overstate the predictive value of in silico methods and that verification with experimental methods is warranted. To validate BACDAC predictions with experimental methods, we compared BACDAC results with additional lpWGS cases which had ploidy measurement performed by either karyotyping and/or FISH (28 B-ALL cases) or nuclei flow sorting followed by DNA content analysis (15 patient-derived Cholangiocarcinoma xenograph (PDX) models). For the B-ALL samples, BACDAC agreed in 25/28 cases. For the 15 PDX models, we analyzed the DNA content histograms by first determining if they contained a prominent non-diploid peak (those were the samples with high ploidy) and calculated ploidy by association to the diploid peak. The samples with no prominent non-diploid peak were deemed near-diploid. BACDAC predicted 13/15 correctly. Overall, BACDAC's agreement with experimental methods was 88%.

High-ploidy can be more accurately predicted when considering allelic content

We applied BACDAC to 653 lpWGS tumor-only samples processed as described in the Methods section. The ploidy for these 653 samples showed a bimodal distribution

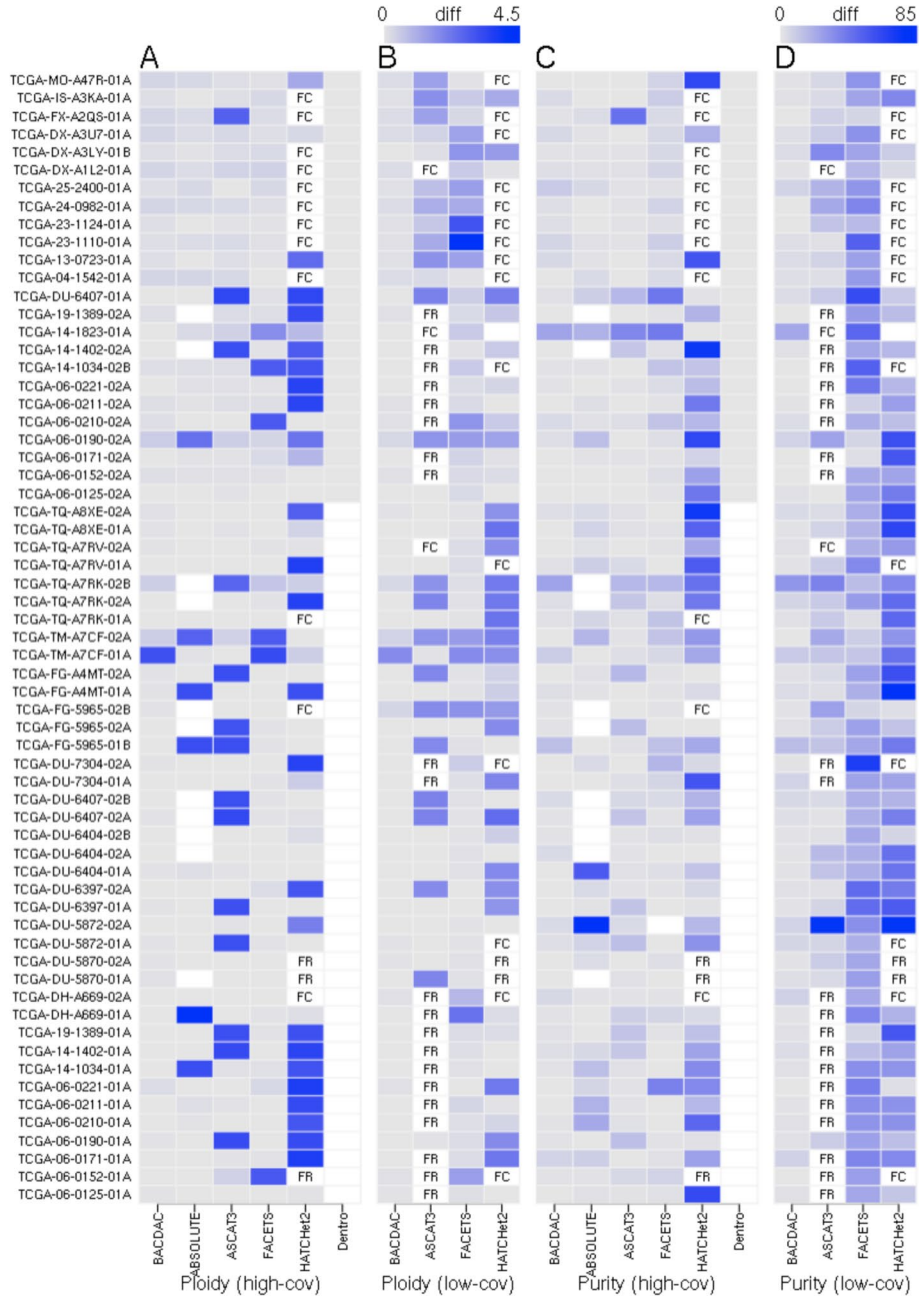

Fig. 6 Comparison of **A**, **B** ploidy and **C**, **D** purity results from multiple methods for 63 TCGA tumors (GBM = 21, LLG = 30, OV = 6, SARC = 6). Results are summarized as the difference from the reference value, **A**, **B** ploidy at high- and low-coverage WGS and **C**, **D** purity at high- and low-coverage WGS. This reference value, a best approximation of a gold standard, was established as the consensus value published in Dentro [9] when available, otherwise as the median of the high-coverage results from the five methods: BACDAC, ABSOLUTE, ASCAT3, FACETS, and HATCHet2. Gray boxes indicate agreement with the reference value while increasingly dark blue indicates an increased difference from the reference. White indicates no data available. FR, failed to run; FC, failed to converge. HATCHet2 was run with the cbc solver and could not install gurobi license. ASCAT3 and HATCHet2 unable to produce a result in 24/63 and 18/63 low-coverage samples respectively

Table 1 Summary of difference (diff) from the reference value for ploidy and purity for each method, as executed on 63 TCGA low-coverage samples. (*n* number of samples completed for each method)

Metric	BACDAC $n=63$	ASCAT3 $n=39$	FACETS $n=63$	HATCHet2 $n=45$
Ploidy diff < 0.3	59	18	40	15
Purity diff < 20%	61	32	14	14
Avg. Ploidy diff	0.1	0.9	0.6	1.1
Avg. Purity diff	3.8	12. 8	26.3	34.3

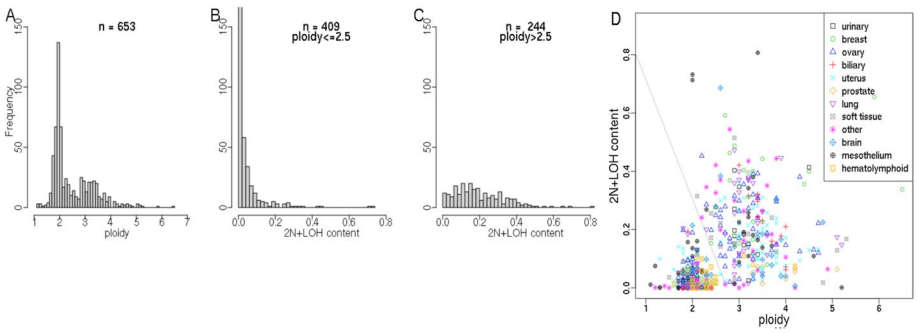

Fig. 7 Distributions of ploidy and 2N$^+$LOH content for 653 low-pass samples. The 2N$^+$LOH content is the proportion of LOH present at 2N and higher (allele fractions major ≥ 2: minor = 0; 2:0, 3:0, etc.). **A** Histogram of ploidy shows a bimodal distribution with a vertex minimum at ploidy = 2.5N, 63% (409/653) were diploid (ploidy ≤ 2.5N), and the remaining 37% (244/653) were ploidy > 2.5N. 1% (8/603) had ploidy < 1.5N. **B** Histogram of 2N$^+$LOH content for samples with ploidy ≤ 2.5N. **C** Histogram of 2N$^+$LOH content for samples with ploidy > 2.5N. **D** 2N$^+$LOH content varies by ploidy. The 2N$^+$LOH content mean increased significantly, from 0.12 to 0.203 for samples with ploidy ≤ 2.5N and ploidy > 2.5N, respectively (p-value = 9.70e-11). Ploidy correlates positively with increased 2N$^+$LOH content. The tumor samples separate into two groups: diploid and high-ploidy, which were divided according to line (Eq. 1): $y = (-0.411)\,x + 1.126$, where x = ploidy and y = 2N$^+$LOH content

(Fig. 7A), with a sharp narrow peak at ploidy = 2N, and a broad peak at ploidy = 3.3N, similar to the findings of other reports [15]. Although it may be a natural choice to define the cutoff between diploid and high-ploidy tumors as the minimum between the peaks of the bimodal ploidy distribution, ploidy = 2.5N, this choice would ignore the substantial overlap between the two distributions. Thus, we investigated whether an additional factor, allelic content, could provide better accuracy to identify diploid from high-ploidy samples, especially near a ploidy of 2.5N.

We explored the relationship between allelic content and ploidy by analyzing the 2N$^+$LOH content, defined as the proportion of LOH present at 2N and higher (allele fractions major ≥ 2: minor = 0; 2:0, 3:0, etc.). The distributions of 2N$^+$LOH content for samples with ploidy ≤ 2.5N and > 2.5N are shown in Fig. 7B and C, respectively. Most diploid samples had zero to very little 2N$^+$LOH content, while samples with ploidy > 2.5N had a much greater and wider distribution. In fact, the 2N$^+$LOH content increased significantly (Fig. 7D) from a mean of 0.12 to 0.203 for samples with ploidy ≤ 2N and > 2.5N, respectively (p = 9.7e-11). Thus, increased ploidy is positively correlated with increased 2N$^+$LOH content.

Interestingly, the tumors formed two major clusters with clear separation (Fig. 7D). We hypothesized that this separation may represent two distinct tumor profiles, the

left lower cluster consisting of diploid or hypodiploid tumors and the right more diffuse cluster consisting of "high-ploidy" tumors, which show evidence of either WGD or successive gains. We used the DBSCAN (density-based spatial clustering of applications with noise) R library to identify the two distinct clusters. Then we trained a Support Vector Machine (SVM) to calculate the following line as the decision boundary that best separates the two clusters (Eq. 1):

$$y = -.411x + 1.26 \quad (1)$$

where x = ploidy and $y = 2N^{+}$LOH content. Others have proposed strategies to identify WGD tumors, for example, using a threshold based on the percentage of the genome with LOH [9, 16, 17] or the percentage of the genome with a major copy number of two or more [15]. These methods were applied to the 653 samples; plots illustrating the difference in calls by each method are shown in Additional file 1: Fig. S8. Compared to the number of samples identified as WGD, more samples were identified by BACDAC as high-ploidy. But BACDAC avoided calling a sample with ploidy < 2N as high-ploidy, while the Dentro method did identify this sample as WGD (Additional file 1: Fig. S8D).

It is important to clarify that BACDAC does not determine WGD status, a task that challenges computational methods in general. BACDAC can, however, predict high-ploidy from near-diploid status as the $2N^{+}$LOH metric combined with average ploidy classifies samples into two distinct clusters, one containing the hypo and near-diploid samples and the other one containing the high-ploidy samples. Also, it is difficult to determine if there is a mixture of populations of cells with both near-diploid and WGD status, simultaneously. This phenomenon has been observed by Steele et al. [18]. BACDAC would not be able to distinguish between the two cell populations as the 2D pattern of the clouds would be identical.

BACDAC results parsed by tissue type revealed that high-ploidy could be underreported

The 653 tumor samples spanned more than 12 primary tissues (Additional file 1: Fig. S1); the tumor ploidy distribution for each tissue group is shown in Additional file 1: Fig. S9A and listed in Additional file 1: Table S4. The percentage of samples with high-ploidy is listed based on (1) ploidy alone and (2) ploidy and $2N^{+}$LOH content (Eq. 1). Slightly more samples were identified as high-ploidy based on the second method (38% vs 40% respectively). Breast and ovarian tumor samples showed the highest percentage (for tissue groups with $n \geq 20$) of high-ploidy samples (60% and 61% of samples, respectively), while hematolymphoid samples, composed primarily of multiple myeloma, showed the lowest incidence of high-ploidy (4% of samples), similar to previous reports [2, 19]. Hypodiploid tumors with low ploidy (ploidy < 1.5N) are common in some diseases, such as B-cell lymphoma [20] and mesothelioma [16]; our nine low ploidy samples included four mesothelioma, two pancreatic, one serous endometrial, one soft tissue Sarcoma, and one Hurthle cell cancer samples. One of the four mesothelioma cases was confirmed as hypodiploid experimentally by nuclei flow sorting. Similar to previous reports considering LOH [11], the $2N^{+}$LOH content increased for all high-ploidy samples compared to that of diploid samples across all tissue types (Additional file 1: Fig. S9B).

Discussion

Tools that both calculate and visualize ploidy and identify high-ploidy or WGD samples from WGS data are limited. As shown in the results, distinguishing subclones from high-ploidy and therefore accurate tumor ploidy determination requires the integration of allelic content across the genome. Due to the previously high costs of WGS, many WGS datasets are tumor-only with varying coverage and tumor purity, complicating the calculation of allelic content, particularly for low-pass sequencing. Even as WGS costs decrease allowing for higher coverage, low tumor purity could still present a challenge for algorithms that require direct measurement of allelic fractions. Therefore, we devised an algorithm that approximates allelic content based on the binomial distribution of common SNPs, that is valid for low-pass sequencing and does not require a matched normal sample. Our analysis has shown that accurate ploidy results can be achieved for tumor-only samples with an effective tumor coverage of 1.2x, such as a sample with 3X coverage and 40% tumor purity. Ploidy results are still possible for samples with tumor purity as low as 20%, if coverage is increased to 6X to compensate for the lack of tumor-specific data.

We developed the Constellation Plot to provide a simple method for viewing ploidy and allele-specific copy number content together. The Constellation Plot is fundamentally biallelic SNP content vs read depth, which in concept is also provided by other methods, for example, FACETS [3], HATCHet2 [4], and CELLULOID [7]; however, the Constellation Plot provides several elements that are unique and valuable for assisting interpretation. Primarily, the locations of the expected allelic fraction for each copy number state, e.g., stars at 4:0, 3:1, 2:2, for copy number 4N, are automatically annotated. These stars are valuable for identifying allele fraction content, and for accessing alignment of the clouds to the stars, a step in verifying accuracy of the ploidy and purity prediction. Another unique element, the purple LOH line, identifies and labels LOH to provide context and extent of LOH within the tumor. Also, annotation of the read depth and copy number are both provided, on the left and right *y*-axis, respectively, in linear scale, for easier interpretation. Furthermore, the Constellation Plot can be plotted in alignment with the linear genome-wide copy number plot to visualize assignment of chromosomal segments to their respective allelic designation. Users have the option to add tumor ploidy and purity annotations directly on the plot.

Interpreting the effect of a variant at the gene level is complicated by complex allele-specific copy number populations. The Constellation Plot simplifies this complexity by revealing patterns of allele fractions and subclones. Pattern components, occurring individually or concurrently, include monoallelic populations, allele fraction combinations, and gains; these patterns may allude to tumor evolution and progression. For example, monoallelic populations are diploid populations that have lost one allele and then either remain in the 1N(1:0) state or progress to be copied one or more times, e.g., 2N((2:0), 3N(3:0), etc. These populations are easily identified as clouds that fall along the purple line in the Constellation Plot. Identification of these populations is important for understanding gene behavior in the presence of a mutation.

We have demonstrated how the Constellation Plot also serves as a tool for verifying the tumor ploidy and purity solution. This need is important for two reasons. First, as algorithms can solve for multiple tumor ploidy and purity combinations, there is a need

to determine which solutions are valid. The clouds in the Constellation Plot provide this feedback; invalid solutions are those with clouds centered left of the purple LOH line or have low overlap between the clouds and the stars. If an invalid solution occurs, BACDAC allows manual configuration of the input settings to induce and test another solution. When multiple valid options for tumor ploidy and purity are possible, a known issue [2, 21], BACDAC defaults to the simplest solution (example in Additional file 1: Fig. S7). Second, the Constellation Plot allows the user to understand the framework of the solution and how other solutions might not be plausible. By showing which read depth distribution peaks are aligned to which copy number, the user can see how shifting the peak-to-grid alignment could result, for example, in large (> 5 Mb) 0N segments, an unstable and unlikely scenario, or could result in more clouds positioned between the stars and thus indicating more subclonal cell populations.

WGD is one of the most common genetic events in cancer [22]. Because WGD is associated with poorer outcomes [15, 23], identification of WGD is important for prognosis and treatment strategies. Many groups have reported strategies for identifying WGD from sequencing data using allele fractions (Additional file 1: Fig. S8B,C). Calling WGD status is challenging with computational methods because it is difficult to discriminate between successive gains and true WGD. BACDAC does not predict WGD status but rather separates high-ploidy from near-diploidy. The $2N^{+}$ LOH metric combined with average ploidy classifies samples to two distinct clusters, one containing the low and near-diploid samples and the other one containing the high-ploidy samples. The cluster separation appears more pronounced than other metrics such as pLOH (1N + LOH). This is because the CN1 state is more often present in the near-diploid samples than the WGD samples, as these CN1 regions became 2NcnLOH regions after doubling. The $2N^{+}$ LOH fraction is capturing the copy-neutral events and subsequent gains that are highly associated with WGD and successive-gains. Thus, strategies that include the 1N allele fraction do not result in as much of a separation as our analysis based ploidy with $2N^{+}$LOH content. These other strategies to distinguish WGD from non-WGD tumors are very sensitive to the slope of the threshold line chosen. As mentioned in MEDICC2 by Kaufmann et al. [24], the high-ploidy cases are likely a mixture of WGD and successive-gains. MEDICC2 may be the most parsimonious method to call WGD but MEDICC2 did not confirm WGD status experimentally.

Limitations

Our ploidy algorithm was optimized to analyze low-pass WGS tumor data without a patient matched normal sample; there is currently no option for including a matched normal. While including a low-pass normal sample would not be beneficial, the inclusion of a high-coverage normal sample would provide high-confidence biallelic positions. These positions could reduce noise and increase resolution, especially in high tumor purity samples. When tumor purity is high (> 95%), SNPs needed for the hetScore calculation are not present, and noise, such as miss-mappings, becomes more dominant in monoallelic fractions. As a result, the clouds in the Constellation Plot associated with monoallelic fractions shift to the right of their expected position and no longer align with the stars or the purple LOH line (example in Additional file 1: Fig. S3D). Using

tumor-specific germline biallelic positions would resolve this issue while also reducing noise.

Computational methods for ploidy fitting can be mathematically ambiguous and, as explained in Tarabichi [8] "true underlying ploidy can only be obtained from experimental ploidy validation and in silico estimates will always be mathematically ambiguous." In cases with ambiguity, BACDAC chooses the simpler solution (the solution with lower average ploidy). Therefore, samples with WGD and no subsequent clonal CNVs to justify a more refined grid would be falsely called near-diploid. The only discrepancy in these situations would be between the algorithmically derived tumor purity and the tumor purity from pathology examination. Therefore, an accurate determination of tumor purity using digital pathology could resolve such ambiguities. We believe that these ambiguous situations are infrequent as most WGD clones become more stable with subsequent deletions, which, in turn, help refine the grid. This assertion is corroborated in the literature as mentioned by Bielski et al. [13] "pure tetraploidization is rare" and is also mentioned in the MECICC2 paper [11] "Tetraploidization followed by rapid chromosomal loss to reach a near-triploid state has been described in many tumor types and is naturally contained in our model in the form of a WGD event followed by multiple losses of individual chromosomes."

Another limitation is identifying runs of homozygosity (ROH) in patients with closely related parents, though this is quite infrequent. BACDAC currently does not predict diploid ROH. However, the constellation plot could be very useful in this matter as ROH regions fall left of the purple LOH line.

Conclusions

Quantifying the allelic content of a tumor genome, allele-specific copy number, is important for understanding tumor progression and treatment options. With allele-specific copy number, genomic regions of LOH can be identified and proper ploidy assessment can be performed. We have presented a method that can report allele-specific copy number, tumor ploidy and purity in samples with as low as $3\times$ coverage, without the need for a matched normal sample. We have presented an alternative method to separate high-ploidy samples from diploid samples using $2N^{+}$LOH content and ploidy. Finally, we have presented a visualization method for allele-specific copy number, especially useful in low-pass data, to identify key allele-specific copy number events.

Methods

Data sets

All samples followed our standard pipeline, BIMA [25] to map to the GRCh38 reference genome and SVAtools [26] to call structural variants 27–35], followed by BACDAC, the ploidy algorithm presented here. For inclusion in this report, samples had to have at least 1.2X ETC, which for example, is equivalent to 20% tumor purity and $6\times$ coverage, or 40% tumor purity and 3X coverage. Thus, lower tumor purity was allowed with increased coverage. Samples with no CNVs were excluded. From our internal Mayo Clinic database, 653 tumor samples originating from more than 12 tissues met these criteria (Additional file 1: Fig. S1). The comparison of ploidy prediction methods was performed on 63 TCGA samples from 36 patients (primary and

recurrent samples). These samples were a mix of TCGA-GBM (glioblastoma $n=11$) and TCGA-LGG (low grade glioma $n=13$), TCGA-OV (ovary $n=6$), and TCGA-SARC (sarcoma $n=6$) samples. To mimic low-pass coverage, the TCGA samples were down-sampled to 5X coverage using seqtk-1.3 (https://github.com/lh3/seqtk). The sequencing statistics, coverage, and BACDAC results for the 63 TCGA and 653 Mayo Clinic samples are provided in Additional file 2: Tables S5 and S6, respectively.

Heterozygosity score

The hetScore depends on the tumor purity, read depth, and copy number. Thus hetScore cannot be compared across samples of varying tumor purity and read depth, and furthermore, no single threshold for hetScore signals LOH, as described below. HetScore is calculated from a binomial distribution of common SNPs within a 1 Mb window, an area large enough to account for low-coverage and underlying sampling probability distributions. Common SNPs are defined as those observed at a frequency of $\geq 1\%$ in the 1000 Genomes Project and AWS, RRID:SCR_008801, approximately 12.6 million SNPs total.

To calculate hetScore, let P_x be every common SNP within 500 kb of position p (Eq. 2):

$$p - 500Kb < P_x < p + 500Kb \tag{2}$$

which satisfy the following condition for germline heterozygosity (Eq. 3):

$$\begin{gathered} 1 + \left\lfloor 0.1 n_x^T \right\rfloor < n_x^A < n_x^T - \left\lfloor 0.1 n_x^T \right\rfloor \\ n_x^T \geq 4 \end{gathered} \tag{3}$$

where n_x^R is the number of bases matching the reference allele for each SNP, and n_x^A is the number of bases matching the known alternate allele for each SNP, such that the total number of bases at position x (Eq. 4) is:

$$n_x^T = n_x^R + n_x^A \tag{4}$$

This condition is necessary because the method is designed for use in the tumor-only setting, and this accounts for homozygous SNPs and possible sequencing and alignment errors.

We can then calculate the actual minor allele value for each 1 Mb region around position p (Eq. 5):

$$A_p = \sum_x \min\left(n_x^R, n_x^A\right) \tag{5}$$

This value is compared against the expected minor allele value, which accounts for the coverage at each position and assumes a binomial distribution for sequencing the reference or alternate allele in a given read at a given position. For this calculation, the binomial probability mass function (PMF) is renormalized to account for the removal of SNPs by the germline heterozygosity condition (Eq. 3) to produce a modified binomial PMF (Eq. 6):

$$Pr(n;k;p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6}$$

$$Pr'(n;k;p) = \frac{Pr(n;k;p)}{\sum_{k'>1+\lfloor 0.1n \rfloor}^{k'<n-\lfloor 0.1n \rfloor} Pr(n;k';p)}$$

Using this modified PMF, the expected minor allele value is calculated (Eq. 7):

$$E_p = \sum_x^{k'<n-\lfloor 0.1n \rfloor} \sum_{k'>1+\lfloor 0.1n \rfloor} Pr'\left(n_x^T; k'; 0.5\right) * min\left(n_x^T - k', k'\right) \tag{7}$$

The hetScore is then calculated as a ratio of actual vs. expected (Eq. 8):

$$\text{hetScore} = \frac{A_p}{E_p} \tag{8}$$

such that regions where the major and minor alleles are equal (i.e., heterozygous regions of the genome) will have hetScore values of approximately 1, while any deviation from heterozygosity (i.e., LOH, allelic gain/loss, etc.) will produce hetScore values < 1. The magnitude of the deviation from 1 for a given allele-specific copy number is determined by the coverage in the region and the tumor fraction. The calculations for the expected deviation in hetScore for other allele-specific copy number values for a given sample are described in the Methods for the Constellation Plot.

Segmentation

The read depth per 30 kb windows were measured as described previously [6]. Read depth normalization included corrections for GC and SINE content. Genomic segments with similar read depth and longer than 3.5 Mb were identified as test segments. The size limit was necessary to avoid segments with homozygous deletions. Large test segments (> 7 Mb) were split into smaller segments of at least 3 Mb. All samples had between 500 and 1000 test segments. The hetScore for each test segment was the median hetScore of the 30 kb segments within the test segment.

Ploidy algorithm

The ploidy algorithm is implemented in R. The inputs include read depth and hetScore of the test segments. The read depth frequency data were tabulated as the number of reads per 100 kb window, for the autosomal chromosomes. Chromosomes X and Y were not included for consistency between male and female samples. A probability density function of the frequency data produced the read depth distribution. The probability density function was performed using the R function *density* from the Stats library via kernel density estimation. Peaks of the probability density function were detected by the R function *turnpoints* from the pastecs library. The largest peak was identified first, peak identification continued by then removing frequency data within a given percentage of the current peak, estimating a new probability density function, and then again identifying the largest remaining peak. This process continued until all peaks greater than 2.5% of the first peak were identified from the read depth distribution (Additional file 1: Fig. S10). Configurable settings allow for the optimization of peak identification. Adjusting

these settings may be necessary for samples when inadequate separation of the peaks occurs. This occurs for samples near the tumor purity or coverage limits or for samples with poor read depth normalization. The default settings were used for 51/63 TCGA samples.

The peaks are then aligned to a one-dimensional linear grid of equally spaced intervals. The peak-to-grid alignment is optimized by testing a range of grid intervals and then applying bonuses for each peak aligned to a grid line and penalties for a too fine-grained interval. The outcome results in a grid interval that aligns all major peaks and as many minor peaks as possible to the grid, but not necessarily all.

The copy number is assigned to each peak aligned to the grid. First, the peak with the lowest read depth is assigned 1N or 2N based on the hetScore of the test segments within that peak. The copy number of the remaining grid-aligned peaks are incremented sequentially by one. A series of checks is performed to confirm the 1N or 2N choice, with adjustments made if needed.

Average ploidy calculation

Ploidy is the average copy number of each 30 kb segment for chromosomes 1–22, not including masked areas. Masked areas include (1) the read depth mask as described previously [6], which removes outliers from the frequency distribution and (2) segments where the hetScore of more than half the samples from a panel of 23 normal samples dropped below the hetScore threshold.

Tumor purity calculation

Tumor purity was calculated by choosing any two peaks from the read depth distribution (rd_1 and rd_2) and their corresponding copy numbers (cn_1 and cn_2). First, we calculate the distance between two successive peaks aligned to the grid (Eq. 9):

$$D = \frac{abs(rd_1 - rd_2)}{abs(cn_1 - cn_2)} \tag{9}$$

Then, the tumor purity is calculated as shown where the denominator reduces to the read depth of the diploid peak (Eq. 10):

$$\text{tumor purity} = \left(\frac{2 * D}{rd_1 + D * (2 - cn_1)} \right) \tag{10}$$

Constellation plot

The components of the Constellation Plot are described in Additional file 1: Fig. S2. To produce the Constellation Plot, the expected deviation in hetScore for all relevant allele-specific copy number values must be calculated given the characteristics of the individual sample, coverage, and tumor fraction. The tumor fraction is provided as an output from the ploidy algorithm. The coverage for a given allele-specific copy number can be extrapolated from the data using the idealized assumption that the coverage in the genome fits a Poisson distribution.

After peak determination, the mode was selected, and the coverage for all common SNPs found within 2.5% normalized read depth (NRD) of this peak was fit to

a Poisson distribution, where NRD is the read depth converted such that the diploid peak is NRD = 2. This produces a normalized read depth for the peak, NRD_M, associated with the expected value of the fit Poisson, λ_M. From these values the expected value, λ_{CN}, for any copy number value, CN, can be calculated given the tumor fraction, τ, determined by the ploidy algorithm using the following scaling equation (Eq. 11):

$$\lambda_{CN} = \frac{\lambda_M(2 + \tau(CN - 2))}{NRD_M} \tag{11}$$

With an estimation of coverage in hand for any desired copy number state, we can calculate the expected hetScore for every allele-specific copy number state defined by the total copy number, CN, and the minor copy number, n_m.

First, as with the binomial distribution, we normalize the Poisson probability mass function to account for the positions removed by the germline heterozygosity requirements (Eq. 12):

$$Pr_P(j;\lambda) = \frac{\lambda^j e^{-\lambda}}{j!} \tag{12}$$

$$Pr_{P'}(j;\lambda) = \frac{Pr_P(j;\lambda)}{\sum_{j' \geq 4} Pr_P(j';\lambda)}$$

Second, we calculate the probability of sequencing the alternate allele for a given CN and n_m value. This probability also depends on the tumor fraction τ (Eq. 13)

$$p^m_{CN} = \frac{n_m\tau + (1 - \tau)}{CN\tau + 2(1 - \tau)} \tag{13}$$

These values can be used to calculate the expected minor allele value for the given CN and n_m values (Eq. 14).

$$E_{CN,n_m} = \sum_{j \geq 4} Pr'_p(j, \lambda_{CN}) \sum_{k > 1 + \lfloor 0.1j \rfloor}^{k < j - \lfloor 0.1j \rfloor} Pr'(j; k; p^m_{CN}) * min(j - k, k) \tag{14}$$

This value is then compared against the expected minor allele value for perfect heterozygosity for the same expected value, λ_{CN} (Eq. 15).

$$E_{CN,\text{net}} = \sum_{j \geq 4} Pr'_p(j, \lambda_{CN}) \sum_{k > 1 + \lfloor 0.1j \rfloor}^{k < j - \lfloor 0.1j \rfloor} Pr'(j; k; 0.5) * min(j - k, k) \tag{15}$$

To produce the expected hetScore for the given CN and n_m values (Eq. 16).

$$\text{hetScore}_{CN,n_m} = \frac{E_{CN,n_m}}{E_{CN,het}} \tag{16}$$

This calculation is used to calculate the expected shift for all stars and the LOH curve on the Constellation Plot.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03599-2.

Additional file 1. Figures S1 to S10, Tables S1 to S4.

Additional file 2. Tables S5 to S6.

Acknowledgements
The following contributed samples to the 653 sample meta-analysis: Drs. Marie Christine Aubry, Linda Baughn, Terry Burns, Joaquin Garcia, Caterina Giannini, Mark Jentoft, Bradley C. Leibovich, Vanda Lennon, Minetta Liu, Aaron Mansfield, Andrea Mariani, Robert McWilliams, Julian Molina, Michael Rivera, Aubrey Thompson, and Yanyan Lou.
This research was partially funded by the Center for Individualized Medicine, Mayo Clinic, Rochester MN and the Center for Digital Health, Mayo Clinic, Rochester MN. PhD training was supported by the Virology and Gene Therapy Grant, NIH National Institute of Allergy and Infectious Diseases (2 T32 AI132165).

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
SJ algorithm conception, design; data interpretation; manuscript writing and editing. JS algorithm conception, design; data interpretation; manuscript writing and editing. RZ algorithm conception, design; data interpretation; manuscript writing and editing. MTB data interpretation; manuscript writing and editing. AG data analysis, and source code repository development. AS algorithm design. AF data interpretation; manuscript writing and editing. MJB data interpretation; manuscript writing and editing. JC data interpretation. GV algorithm conception, design; data interpretation; manuscript writing and editing. All authors read and approved the final manuscript.

Funding
This research was partially funded by the Center for Individualized Medicine, Mayo Clinic, Rochester MN and Center for Digital Health, Mayo Clinic, Rochester MN. PhD training was supported by the Virology and Gene Therapy Grant, NIH National Institute of Allergy and Infectious Diseases (2 T32 AI132165).

Data availability
The TCGA datasets supporting the conclusions of this article have been previously made publicly available. The TCGA datasets include the raw fastq files downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The ASCAT2 tumor ploidy and purity results were also downloaded from the Genomic Data Commons Data Portal. The ABSOLUTE tumor ploidy and purity results were obtained from previously published supplemental data [36]. https://doi.org/10.1016/j.ccell.2018.03.007. The Dentro tumor ploidy and purity results were also obtained from previously published supplemental data [9]. The ASCAT3, FACETS and HATCHet2 results were processed by us using the code downloaded from their respective Github sites. The BACDAC ploidy and tumor purity data generated in this study are available within the article and its additional files. Input data for BACDAC includes read depth counts, ref/alt counts for SNPs analyzed and segmentation. The BACDAC input files to replicate figures 1, 2, 3, 4, 5 and 6 are available on zenodo [37] https://doi.org/10.5281/zenodo.15277102. BACDAC R code repository can be accessed on GitHub [38] (https://github.com/vasmatzis/BACDAC) and zenodo [39] https://doi.org/10.5281/zenodo.15284454. The BACDAC source code repository includes example data, a low coverage TCGA sample, to demonstrate BACDAC. The BACDAC source code repository and the data repository are both available via GNU Affero General Public License version 3.

Declarations

Ethics approval and consent to participate
Results represent a meta-analysis of non-clinical trial data, collected from multiple Mayo Clinic institutional review board studies (IRB) including 12–007850, 13–000942, 15–005545, and 19–005326. All methods were carried out in accordance with relevant guidelines and regulations, and all experimental protocols were approved by the Mayo Clinic IRB studies. Informed consent was obtained from all the participants and/or their legal guardians. The research conformed to the principles of the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
George Vasmatzis is the owner of WholeGenome, LLC.
All other authors have declared no competing interests.

Received: 7 May 2024 Accepted: 29 April 2025
Published online: 20 May 2025

References

1. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107(39):16910–5.

2. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30(5):413–21.
3. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44(16):e131.
4. Myers MA, Arnold BJ, Bansal V, Balaban M, Mullen KM, Zaccaria S, et al. HATCHet2: clone- and haplotype-specific copy number inference from bulk tumor sequencing data. Genome Biol. 2024;25(1):130.
5. Prandi D, Demichelis F. Ploidy- and purity-adjusted allele-specific DNA analysis using CLONETv2. Curr Protoc Bioinformatics. 2019;67(1):e81.
6. Smadbeck JB, Johnson SH, Smoley SA, Gaitatzes A, Drucker TM, Zenka RM, et al. Copy number variant analysis using genome-wide mate-pair sequencing. Genes Chromosomes Cancer. 2018;57(9):459–70.
7. Notta F, Chan-Seng-Yue M, Lemire M, Li Y, Wilson GW, Connor AA, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. Nature. 2016;538(7625):378–82.
8. Tarabichi M, Salcedo A, Deshwar AG, Ni Leathlobhair M, Wintersinger J, Wedge DC, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods. 2021;18(2):144–55.
9. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell. 2021;184(8):2239-54 e39.
10. Whole genome duplication is an early event leading to aneuploidy in IDH -wild type glioblastoma Oncotarget. 2018;9(89):36017–028. https://doi.org/10.18632/oncotarget.v9i89. https://doi.org/10.18632/oncotarget.26330.
11. Ciani Y, Fedrizzi T, Prandi D, Lorenzin F, Locallo A, Gasperini P, et al. Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. Cell Syst. 2022;13(2):183-93 e7.
12. Zhao Y, Fang LT, Shen TW, Choudhari S, Talsania K, Chen X, et al. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. Sci Data. 2021;8(1):296.
13. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol. 2021;39(9):1151–60.
14. Masood D, Ren L, Nguyen C, Brundu FG, Zheng L, Zhao Y, et al. Evaluation of somatic copy number variation detection by NGS technologies and bioinformatics tools on a hyper-diploid cancer genome. Genome Biol. 2024;25(1):163.
15. Bielski CM, Zehir A, Penson AV, Donoghue MTA, Chatila W, Armenia J, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. Nat Genet. 2018;50(8):1189–95.
16. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022;606(7916):984–91.
17. Gonzalez S, Lopez-Bigas N, Gonzalez-Perez A. Copy number footprints of platinum-based anticancer therapies. PLoS Genet. 2023;19(2):e1010634.
18. Steele CD, Tarabichi M, Oukrif D, Webster AP, Ye H, Fittall M, et al. Undifferentiated sarcomas develop through distinct evolutionary pathways. Cancer Cell. 2019;35(3):441-56.e8.
19. Sidana S, Jevremovic D, Ketterling RP, Tandon N, Greipp PT, Baughn LB, et al. Tetraploidy is associated with poor prognosis at diagnosis in multiple myeloma. Am J Hematol. 2019;94(5):E117–20.
20. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45(3):242–52.
21. Li Y, Xie X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. Bioinformatics. 2014;30(15):2121–9.
22. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134–40.
23. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Gronroos E, Endesfelder D, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 2014;4(2):175–85.
24. Kaufmann TL, Petkovic M, Watkins TBK, Colliver EC, Laskina S, Thapa N, et al. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. Genome Biol. 2022;23(1):241.
25. Drucker TM, Johnson SH, Murphy SJ, Cradic KW, Therneau TM, Vasmatzis G. BIMA V3: an aligner customized for mate pair library sequencing. Bioinformatics. 2014;30(11):1627–9.
26. Johnson SH, Smadbeck JB, Smoley SA, Gaitatzes A, Murphy SJ, Harris FR, et al. SVAtools for junction detection of genome-wide chromosomal rearrangements by mate-pair sequencing (MPseq). Cancer Genet. 2018;221:1–18.
27. Kosari F, Disselhorst M, Yin J, Peikert T, Udell J, Johnson S, et al. Tumor junction burden and antigen presentation as predictors of survival in mesothelioma treated with immune checkpoint inhibitors. J Thorac Oncol. 2022;17(3):446–54.
28. Murphy S, Smadbeck J, Eckloff B, Lee Y, Johnson S, Karagouga G, et al. Chromosomal Junction detection from whole-genome sequencing on formalin-fixed, paraffin-embedded tumors. J Mol Diagn. 2021;23(4):375–88.
29. Vasmatzis G, Liu MC, Reganti S, Feathers RW, Smadbeck J, Johnson SH, et al. Integration of comprehensive genomic analysis and functional screening of affected molecular pathways to inform cancer therapy. Mayo Clin Proc. 2020;95(2):306–18.
30. Aypar U, Smoley SA, Pitel BA, Pearce KE, Zenka RM, Vasmatzis G, et al. Mate pair sequencing improves detection of genomic abnormalities in acute myeloid leukemia. Eur J Haematol. 2019;102(1):87–96.
31. Murphy SJ, Harris FR, Smadbeck JB, Serla V, Karagouga G, Johnson SH, et al. Optimizing clinical cytology touch preparations for next generation sequencing. Genomics. 2020;112(6):5313–23.
32. Murphy SJ, Levy MJ, Smadbeck JB, Karagouga G, McCune AF, Harris FR, et al. Theragnostic chromosomal rearrangements in treatment-naive pancreatic ductal adenocarcinomas obtained via endoscopic ultrasound. J Cell Mol Med. 2021;25(8):4110–23.
33. Grassi T, Harris FR, Smadbeck JB, Murphy SJ, Block MS, Multinu F, et al. Personalized tumor-specific DNA junctions to detect circulating tumor in patients with endometrial cancer. PLoS ONE. 2021;16(6):e0252390.

34. Mansfield AS, Peikert T, Smadbeck JB, Udell JBM, Garcia-Rivera E, Elsbernd L, et al. Neoantigenic potential of complex chromosomal rearrangements in mesothelioma. J Thorac Oncol. 2019;14(2):276–87.
35. Vasmatzis G, Kosari F, Murphy SJ, Terra S, Kovtun IV, Harris FR, et al. Large chromosomal rearrangements yield biomarkers to distinguish low-risk from intermediate- and high-risk prostate cancer. Mayo Clin Proc. 2019;94(1):27–36.
36. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell. 2018;33(4):676-89 e3.
37. Johnson SH, Smadbeck JB, Zenka RM. Barrett MT, Gaitatzes A, Solanki A, Florio AB, Borad MJ, Cheville, JC, Vasmatzis G. BACDAC supporting processed data for publication. Datasets.Zenodo. 2025. https://doi.org/10.5281/zenodo.15277102, https://zenodo.org/records/15277102.
38. Johnson SH, Smadbeck JB, Zenka RM. Barrett MT, Gaitatzes A, Solanki A, Florio AB, Borad MJ, Cheville, JC, Vasmatzis G. BACDAC. GitHub. 2024. https://github.com/vasmatzis/BACDAC.
39. Johnson SH, Smadbeck JB, Zenka RM. Barrett MT, Gaitatzes A, Solanki A, Florio AB, Borad MJ, Cheville, JC, Vasmatzis G. . Zenodo. 2025. https://doi.org/10.5281/zenodo.15284454, https://zenodo.org/records/15284454.

Publisher's Note