BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1056–1063
doi:10.1093/bioinformatics/btt759

Genome analysis

Advance Access publication January 2, 2014

AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data

Lei Bao, Minya Pu and Karen Messer*

Division of Biostatistics, Moores Cancer Center, University of California-San Diego, La Jolla, CA 92093, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Detection and quantification of the absolute DNA copy number alterations in tumor cells is challenging because the DNA specimen is extracted from a mixture of tumor and normal stromal cells. Estimates of tumor purity and ploidy are necessary to correctly infer copy number, and ploidy may itself be a prognostic factor in cancer progression. As deep sequencing of the exome or genome has become routine for characterization of tumor samples, in this work, we aim to develop a simple and robust algorithm to infer purity, ploidy and absolute copy numbers in whole numbers for tumor cells from sequencing data.

Results: A simulation study shows that estimates have reasonable accuracy, and that the algorithm is robust against the presence of segmentation errors and subclonal populations. We validated our algorithm against a panel of cell lines with experimentally determined ploidy. We also compared our algorithm with the well-established single-nucleotide polymorphism array-based method called ABSOLUTE on three sets of tumors of different types. Our method had good performance on these four benchmark datasets for both purity and ploidy estimates, and may offer a simple solution to copy number alteration quantification for cancer sequencing projects.

Availability and implementation: The *R* package absCNseq is available from http://biostats.mcc.ucsd.edu/files/absCNseq_1.0.tar.gz.

Contact: kmesser@ucsd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 29, 2013; revised on October 31, 2013; accepted on December 23, 2013

1 INTRODUCTION

Copy number alterations (CNAs) are an important class of somatic aberrations in tumor genomes and have been extensively studied in the hope of understanding the mutational landscape and clonal evolution of cancer (Beroukhim *et al.*, 2010). High-throughput platforms such as array comparative genomic hybridization (Rueda *et al.*, 2007), single-nucleotide polymorphism (SNP) arrays (LaFramboise *et al.*, 2005) and next-generation sequencing (NGS) (Koboldt *et al.*, 2012; Xie *et al.*, 2009) have been widely used to measure CNAs. The direct readout from these platforms is a measure of the relative DNA content between a tumor sample and its normal counterpart, over a set of genomic regions. However, at the cellular level, CNAs are integers: for example, a region of DNA with normal copy number 2, corresponding to the two chromosomes, may be duplicated in one chromosome, resulting in copy number 3, or deleted, resulting in copy number 1. To estimate absolute copy numbers from the observed ratio of tumor to normal DNA, it is necessary to know two additional parameters, tumor *purity*, which is the ratio of tumor cells to total cells in the sample, and tumor *ploidy*, which is the average copy number of the entire tumor genome and can be used to account for whole-genome duplication events in the tumor (Carter *et al.*, 2012). Tumor ploidy may itself be a prognostic factor in cancer (Zanetti *et al.*, 2012). A related issue occurs with detection of somatic single nucleotide variants (SNV) (DePristo *et al.*, 2011; Koboldt *et al.*, 2012; Li, 2011). As with CNAs, it is more informative to know the integer multiplicity of each gained somatic SNV, rather than the percentage of the alternate allele in the tumor/stromal mixture. Incorporating purity, ploidy and absolute copy number information may also help popular SNV callers like samtools (Li, 2011) and GATK (DePristo *et al.*, 2011) to reduce both false-positive and false-negative calls. All this highlights the importance of accurately estimating the absolute copy number information from noisy genomic data.

Recent work addressing this issue has aimed to control for the effects of either tumor purity or ploidy or both in CNA detection (Attiyeh *et al.*, 2009; Bengtsson *et al.*, 2010; Carter *et al.*, 2012; Greenman *et al.*, 2010; Gusnanto *et al.*, 2012; Van Loo *et al.*, 2010; Yau *et al.*, 2010; Yu *et al.*, 2011; Oesper *et al.*, 2013). ABSOLUTE (Carter *et al.*, 2012) uses allele-specific copy ratios from SNP array data to infer purity, ploidy and absolute copy numbers and has demonstrated high accuracy. In principle, ABSOLUTE can also be extended to sequencing data, although the working pipeline and performance on sequencing data are not yet published. As more cancer samples are sequenced using NGS technology and as whole-exome sequencing (WES) becomes a routine platform for characterization of tumor samples, there is an urgent need to develop a similar tool for NGS. Su *et al.* (2012) developed a method to estimate tumor purity from the deep whole-genome sequencing data; however, their method does not offer the estimation of ploidy and absolute copy numbers, and has not been shown to work equally well on WES data.

In this work, we develop a simple and robust statistical method to estimate tumor purity, ploidy and absolute CNAs from WES data. We validated our method by comparing our purity and ploidy estimates against independent gold standard

*To whom correspondence should be addressed.

values using recently published exome data from 31 cancer cell lines (Abaan *et al.*, 2013), 16 breast (Banerji *et al.*, 2012), 16 prostate (The Cancer Genome Atlas Research Network, 2008) and 17 glioblastoma cancer patients (The Cancer Genome Atlas Research Network, 2008). We also study the effects of the existence of subclonal populations or segmentation errors on the accuracy of the purity estimation. To the best of our knowledge, this is the first work that has provided a working pipeline to estimate purity, ploidy and absolute copy numbers from raw WES data, which has been validated on several independently published WES datasets. We implemented the algorithm in an *R* package called absCNseq that is freely available from our Web site (http://biostats.mcc.ucsd.edu/files/absCNseq_1.0.tar.gz).

2 METHODS

2.1 Exome sequencing datasets and post-alignment processing pipeline

2.1.1 NCI60 cell line dataset The NCI60 tumor cell lines from multiple tissues of origin have long been used for *in vitro* anticancer drug screening, and their molecular and cellular characteristics have been well studied by a variety of biological assays. In particular, they have been individually karyotyped by the spectral karyotyping (SKY) assay (Roschke *et al.*, 2003). Just recently, the majority of these cell lines also had their exomes sequenced (Abaan *et al.*, 2013); hence, this represents an ideal validation dataset for our ploidy inference. We used the SKY-determined modal ploidy as the reference ploidy value and compared our ploidy estimates with these reference values.

2.1.2 dbGAP breast cancer dataset We obtained published data from paired breast tumor-normal samples, which had both SNP array data (Affymetrix Genome-Wide Human SNP Array 6.0) and WES data (Illumina GA II, average coverage 100×) (Banerji *et al.*, 2012). Published estimates of purity and ploidy were available for these samples using ABSOLUTE on the SNP array data. We downloaded the aligned read files (BAM files) from dbGAP (Mailman *et al.*, 2007) under the study number phs000369.v1.p1 for the first 23 Mexican subjects (BR-M-005 through BR-M-085) in the published Supplementary Table S2 (Banerji *et al.*, 2012). We extracted the experimental information from the BAM file headers. We excluded seven subjects from further analysis because the sequencing date for the matched germ line and tumor samples were in different weeks, and thus must have been in different sequencing runs. The remaining 16 subjects all had the matched samples sequenced in the same week. Interestingly, we found that the mean coverage ratio across the genome recapitulated our inclusion/exclusion rule almost perfectly: the mean coverage ratios are all close to 1.0 for the samples we included, and obviously further away from 1.0 for the excluded samples (Supplementary Table S2). This supports our selection of samples for this analysis, as it implies that there may exist systematic bias between different sequencing batches.

2.1.3 TCGA tumor dataset We selected two representative tumor types (prostate and glioblastoma) from the TCGA database (The Cancer Genome Atlas Research Network, 2008). We downloaded the WES data for 20 random samples per each tumor type. We compared the purity and ploidy estimates by our method with those by the ABSOLUTE method.

2.1.4 Processing pipeline The flowchart of absCN-seq is shown in Supplementary Figure S1. For all the WES data, the BAM files (Li *et al.*, 2009) have been realigned, deduplicated and recalibrated. We used samtools (version 0.1.18) (Li *et al.*, 2009) to count the number of total reads that were properly paired and aligned in each BAM file, and used this as a normalization constant to compare the coverage depth between the paired tumor and germ line DNA samples. We then used varscan2 (version 2.3.2) (Koboldt *et al.*, 2012) to compute the ratio of coverage between the paired tumor/normal DNA samples for each bin, with bins determined by varscan2. The bin size was $\leq$100 bp. We used the DNAcopy package (Olshen *et al.*, 2004) to perform copy number segmentation. Segments <200 bp in length were excluded because short segments are likely caused by technical artifacts. In addition to calling copy number ratios, the varscan2 software also outputs a set of high-confidence somatic SNVs in the tumor exome. We computed the observed allele frequencies for these somatic SNVs, which provide orthogonal purity information to copy number variation. We included both the segmented copy ratio information and the SNV allele frequencies in our objective function to be optimized, as described later in the text.

2.2 Statistical framework

2.2.1 Basic notation and description of the sequencing data The input data for AbsCN-seq are a set of N segmented genomic intervals and their accompanying observed read depths, for matched germ line and tumor samples. If the segmentation algorithm is correct, then the tumor cells have constant copy number on each segment. A segment may be *clonal*, in that all tumor cells share the same copy number for the segment, or *subclonal*, in that there are two or more subpopulations of tumor cells with different, but constant, copy numbers on the segment. In addition, the tumor sample is composed of mixed cancer and stromal cells, where the stromal cells contain germ line DNA and the cancer cells contain DNA with somatic mutations, including the copy number aberrations considered here.

Let the ith genomic segment be w_i base pairs in length, and let l be the read length. On the ith segment, the tumor sample has total aligned read count $t_{x,i}$ and corresponding read depth $x_i = lt_{x,i}/w_i$, with total aligned reads in the sample given by $T_t = \sum_{i=1}^{N} t_{x,i}$. Similarly, let the germ line read depth be denoted by $y_i = lt_{y,i}/w_i$, with total read count in the germ line sample T_g. Let $\eta_i = E[y_i]/2$ be the haploid mean read depth for segment i for the germ line sample, given that total read depth is T_g; thus, η_i is the mean read depth expected from each homologous chromosome in the germ line sample for this segment. Note that the η_i capture variation in aligned read depth at different locations in the genome due to sequencing bias, which may arise, for example, from varying GC content in the genomic segment, differences in mappability for repetitive regions, differences in capture efficiency of probes in the case of targeted sequencing, or PCR bias in the amplification step. If there is no sequencing bias, so that coverage is expected to be uniform, then $\eta_i \equiv \eta$ for all segments i. We assume that for every segment i expected read depth is strictly proportional to total read count, so that $2\eta_i T_t/Tg$ is the mean read depth that would be expected from the germ line sample at total read count T_t rather than T_g.

2.2.2 Purity (α) and ploidy (τ') Let α be the purity of the tumor sample, that is, the proportion of cancer cells in the tumor sample. Let τ' represent the average genomic ploidy for the tumor sample, that is, twice the ratio of the average amount of genomic DNA per cancer cell relative to the total DNA within a stromal cell, so that $\tau' = 2$ for a tumor with no copy number aberrations. Then in the tumor sample, the ratio of cancer to stromal DNA will be $\alpha\tau' : 2(1-\alpha)$, and the proportion of DNA in the sample from the cancer cells, ρ', will be

$$\rho' \equiv \alpha\tau'/(\alpha\tau' + 2(1-\alpha)).$$

2.2.3 Segment copy number, q_i First, consider a clonal segment, so that all the cancer cells have identical genomes for this segment. The biologic model is that for each of the two homologous chromosomes, maternal and paternal, the DNA sequence represented by interval i

appears once in each germ line cell (copy number 2) and an integral number of times in each cancer cell, $q_{m,i}$ times for the maternal copy and $q_{p,i}$ times for the paternal copy. Then the copy number in the tumor for the segment is $q_i = q_{m,i} + q_{p,i}$, which is an integer. For a subclonal segment, q_i is the average copy number over the differing cancer cell populations that form the tumor subclones, and so may not be an integer. It is important to note that segmentation errors can also cause q_i to not be an integer, with q_i then defined to be the average copy number over the w_i bases comprising genomic segment i.

To relate copy number to genomic ploidy, note that $q_i w_i$ is the total number of base pairs of DNA from genomic segment i per tumor cell and $2w_i$ the corresponding number for a germ line cell, so that the relative copy number $q_i/2$ gives the ratio of total DNA from a cancer cell relative to a normal cell for this segment. Then total genomic DNA in an average tumor cell is given by $\sum_i q_i w_i$, and so twice the ratio of tumor to germ line DNA is given by

$$\tau' = \frac{\sum_i^N q_i w_i}{\sum_i^N w_i}.$$

However, τ' may not reflect the proportion of *aligned reads* captured from tumor relative to germ line cells because the capture efficiency of the sequencing may differ between genomic segments, as represented by the parameters η_i. Hence, we define the sequencing ploidy τ as

$$\tau = \frac{\sum_i^N q_i w_i \eta_i}{\sum_i^N w_i \eta_i}. \quad (1)$$

Of the T_t total aligned reads from the tumor sample, ρT_t are expected to be from the tumor and $(1-\rho)T_t$ from stroma, with

$$\rho = \alpha\tau/(\alpha\tau + 2(1-\alpha)). \quad (2)$$

2.2.4 A copy-number–based ratio estimator Expected read depth for segment i in the tumor sample, $E[x_i]$, can be expressed in terms of purity α, copy number q_i and expected germ line depth $E[y_i] = \eta_i$. First, the total read count in the tumor sample, overall and in segment i, is the sum of reads from stromal cells, denoted with subscript s, and from cancer cells, denoted by c, that is,

$$T_t = T_c + T_s$$

$$t_{x,i} = t_{c,i} + t_{s,i}.$$

The total stromal reads T_s are not observed, but are equal to $(1-\rho)T_t$. They will distribute across the genomic segments in expected proportion to $2\eta_i w_i/T_g$, so that for each segment the expected number of stromal reads in the tumor sample is given by the following equation:

$$E[t_{s,i}] = 2\eta_i w_i (T_t/T_g)(1-\rho).$$

On segment i, the cancer to stromal DNA is in the ratio $\alpha q_i : (2(1-\alpha))$, so that, assuming equal capture efficiency for the cancer and stromal DNA within the segment,

$$E[t_{c,i}] = E[t_{s,i}]\alpha q_i/(2(1-\alpha)).$$

Combining these expressions,

$$E[t_{x,i}] = 2\eta_i w_i (T_t/T_g)(1-\rho)(1+\alpha q_i/(2(1-\alpha))).$$

$$= 2\eta_i w_i (T_t/T_g)\frac{\alpha q_i + 2(1-\alpha)}{\alpha\tau + 2(1-\alpha)}$$

Let $R = (T_t/T_g)$. Then,

$$R^{-1}E[t_{x,i}]/E[t_{y,i}] = R^{-1}E[x_i]/E[y_i] = \frac{\alpha q_i + 2(1-\alpha)}{\alpha\tau + 2(1-\alpha)}. \quad (3)$$

Equation (3) suggests use of a ratio estimator based on the normalized ratio of means $r_i = R^{-1}x_i/y_i$ to estimate the parameters of interest α and q_i, and thus τ.

2.2.5 An SNV-frequency–based estimator Suppose segment i contains one or more loci with an SNV, and let $m(i)$ be the total number of SNVs on the segment, and $f_j, j = 1 \ldots m(i)$, be the corresponding observed variant allele proportions out of the total aligned reads from the tumor sample, for each of the SNVs. By similar arguments as above, the expected SNV allele frequency may be estimated by the following:

$$E[f_j] \approx \frac{\alpha s_j}{\alpha q_i + 2(1-\alpha)} \quad (4)$$

where s_j is the absolute copy number of the somatic allele at locus j. Because up to q_i copies of tumor DNA may carry the mutation, $s_j \le q_i$. Therefore, the observed allele frequencies f_j can also be used to estimate α and q_i.

2.3 The least squares objective function

The least squares objective function is given by the sum of two terms, one containing information about read depth, and the other about variant allele frequency,

$$\lambda \sum_i \left(r_i - \frac{\alpha q_i + 2(1-\alpha)}{\alpha\tau + 2(1-\alpha)} \right)^2$$

$$+(1-\lambda) \sum_i \sum_{j=1}^{m(i)} \left(f_j - \frac{\alpha s_j}{\alpha q_i + 2(1-\alpha)} \right)^2, \quad (5)$$

and is to be minimized over the non-negative integers $s_j \le q_i$ and $\alpha \in [0,1]$. As in Section 2.2 above, on segment i, r_i is the observed normalized copy number ratio and $r_i = R^{-1}x_i/y_i$, and f_j is the observed variant allele frequency for each of the $m(i)$ SNVs residing in segment i. If there is no SNV in that segment, then $f_j = s_j = 0$. The relative importance of these two parts is given by the weight parameter $\lambda \in (0,1)$. When $\lambda = 1$, the objective function reduces to the simple case where only the copy number information is used. In this work, we assessed $\lambda = 1$ and $\lambda = 0.5$ (equal weights).

2.4 Estimation algorithm

We propose an iterative estimation scheme in which we maximize each parameter in turn, cycling through the parameters. Initially, we assume $q_i \in \{0, 1, \ldots, Q\}$, so that all segments are clonal with maximum copy number Q. The purity α is constrained to lie in the interval [0,1]. Step 0 is an initialization step; steps 1 and 2 are iterated to convergence (relative changes in objective function <1e-4).

Step 0: Set $\hat{\eta}_i = y_i$. Initialize $\alpha^{(0)}$ and $\tau^{(0)}$.

Step 1: Assume all intervals are clonal. For each i, set $q_i^{(k)}$ to solve

$$q_i^{(k)} = \arg\min_{q \in Q} |q - (\alpha^{(k)})^{-1}\{[\alpha^{(k)}\tau^{(k)} + 2(1-\alpha^{(k)})]r_i - 2(1-\alpha^{(k)})\}|$$

$$s_j^{(k)} = \arg\min_{s \in \{0,\ldots,q_i\}} |s - (\alpha^{(k)})^{-1}[\alpha^{(k)}q_i^{(k)} + 2(1-\alpha^{(k)})]f_j|$$

$$\tau^{(k)} = \frac{\sum_i^N q_i^{(k)} w_i \hat{\eta}_i}{\sum_i^N w_i \hat{\eta}_i}$$

Note that the q's and s's are available in closed form, as this step only requires rounding the right-hand term inside the absolute value sign to the nearest integer. Here we have dropped the dependence of s_j on i for convenience.

Step 2: Given the estimates $\{q_i^{(k)}\}, \{s_j^{(k)}\}$ and $\tau^{(k)}$, use the non-linear least square method to find $\alpha^{(k+1)}$ that minimizes the objective function (5).

2.4.1 Grid search over initial values and 'no estimate' results This integer optimization problem converges quickly. As with many integer valued maximization problems, the algorithm may find a local optimum, depending on the initial values of α and τ. Hence, we use a grid search strategy over the starting values $\alpha^{(0)}$ and $\tau^{(0)}$ to find the global optimum of the objective function.

We initiate the core algorithm described above over a grid of starting values of purity (ranging from 0.20 to 0.95 with an increment of 0.05) and ploidy (ranging from 1.5 to 5.0 with an increment of 0.05). The set of local maxima found at convergence is inspected and filtered. Solutions that occur only from a single starting value (i.e. solutions that do not recur from multiple initialization points) are removed, as well as biologically implausible solutions (i.e. solutions outside the search range, e.g. with purity 0 or 100%). From the remaining set of stable solutions, the parameter set $\{q_i, \alpha, \tau\}$ with the minimum mean-squared error is returned as the optimal estimate of copy number, purity and ploidy. Our method returns no estimate if the filtration results in no solution. The stable solution set can be retained for further inspection, as there may be more than one solution with nearly equal values of the objective function.

3 RESULTS

3.1 Performance on simulated data

We used two simulation studies, the first to assess the accuracy and the second to assess robustness of our estimates of purity. We based our simulated data on the empirically observed data in our samples. We randomly chose sample BR-M-028 as a model.

3.1.1 Accuracy For the first study, to generate the data, we obtained empirical values for the N segment lengths w_i from sample BR-M-028. We used the estimated segment integer copy numbers q_i. We then computed the true ploidy using Equation (1). We generated r_i's from $N(\mu_i, \sigma^2)$ where μ_i was computed using Equation (3) with α fixed at 0.3, 0.5 or 0.7, and $\sigma^2 = 0.1$ fixed at the value obtained empirically from cancer sample BR-M-028. Thus, these data model a tumor sample with true known values of α, τ and q_i. We used a simulation size of 50, to be consistent with the more extensive simulation of robustness reported later in the text. Notably, to assess worst-case performance of the algorithm, we did not apply the filtering, which yields a 'no call' result, as described in Section 2.4.1, but instead report the values found by the optimization routine. Thus, most of the extreme values reported here would instead be 'no call' results.

Table 1 shows that the method appears to be unbiased; at all purity levels, the median of the estimates are near the true values. When the tumor purity is moderate to high (50–70%), the variability appears to be low, with interquartile range of 0.03 at 50% purity and 0.02 at 70% purity. However, when tumor purity is low (30%), while the median estimate is still close to the true value, the 95% percentile estimate is much larger, demonstrating increased variability. This is expected because in low purity samples, the signal (tumor cell) to noise (stromal contamination) ratio is much lower, and finding the correct unique solution becomes more challenging.

Table 1. Variation in the purity estimate when all segments are clonal

True purity	Median estimate	Quantile			
		5%	25%	75%	95%
0.30	0.31	0.28	0.30	0.34	0.68
0.50	0.50	0.45	0.47	0.50	0.53
0.70	0.69	0.66	0.68	0.70	0.71

3.1.2 Robustness For the second study, we used the same data as above, but allowed the q_i to be non-integer. Recall that q_i is the true average copy number for the ith genomic segment; q_i will fall between integer values if there is an error in the segmentation algorithm, so that copy number is not constant across the segment, or if the tumor contains two or more subclonal mixtures of cells with different copy numbers, so that read depth is averaged from the two subclonal populations. Each of these circumstances is likely to hold for at least a fraction of the genomic segments that are fed into AbsCN-Seq.

To model non-integer copy numbers, we used the same basic simulation setup as above. However, here we designated a fraction n/N of segments to be non-integer, with n varying from 0 to $N/2$. We chose $N/2$ as the upper limit because we thought it unlikely that more than half of segments would be impacted by a copy number event, which was either poorly estimated by the segmentation algorithm or was impacted by a difference between subclones. For each simulated dataset, we then randomly selected a subset of n segments, and for these, we set the true copy number to be a random draw from the uniform distribution $U(q_i - 0.5, q_i + 0.5)$. For each subclonal fraction n/N and purity α, we simulated 50 different r_i's. We then applied AbsCN-seq to the simulated r_i's and assessed the deviation from the true underlying α as the subclonal fraction increases. Again, we did not filter out any 'no call' solutions, and so we forced the method to produce an estimate in all 50 cases.

Figure 1 depicts the effects of the presence of subclonal segments or incorrect segmentation (i.e. non-integer q_i) on the estimated purity. The X-axis is the proportion of segments that are non-integer, and varies from 0, which is the case reported in Table 1, to 50%. The boxplots on the Y-axis show the estimated purity for each of 50 simulated datasets. True purity is shown by the horizonal line. We did not filter out 'no calls'. In practice, many of the extreme estimates of purity shown here would be replaced by 'no estimate'. At least some of the outliers seen in Figure 1 (and in particular, all of the estimates with $\alpha = 1$) were due to this 'no filtration' strategy. It can be seen that the estimated purity is most accurate when the percentage of subclonal segments is <30%. As the proportion of segments with non-integer q_i's increases (e.g. >30%), our method tends to underestimate tumor purity, and the bias is larger when the true purity is higher. However, even in the 'worst' case, at 70% purity and with 50% of segments impacted by segmentation errors or subclonal copy number events, the median estimated purity is ~66% and almost all estimates (90%) are between 58 and 73%. Hence, even in this extreme case, our method appears to give reasonably good and practically useful estimates, although they are somewhat

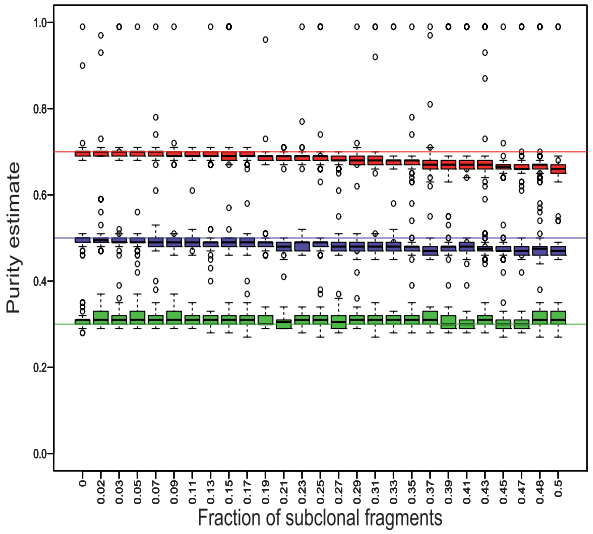

Fig. 1. Simulation showing effects of subclonal fraction on estimated purity, at three purity levels (30, 50 and 70%; horizontal lines). Data modeled after tumor sample BR-M-028. X-axis: proportion of segments set to be subclonal, for which the copy number was perturbed by a uniform random shift away from its integer value. Y-axis: estimated purity for $n = 50$ simulations. Solutions which would normally be 'no call' are included for completeness. A subclonal fraction >30% is seen to introduce some bias, but estimated purity remains reasonable

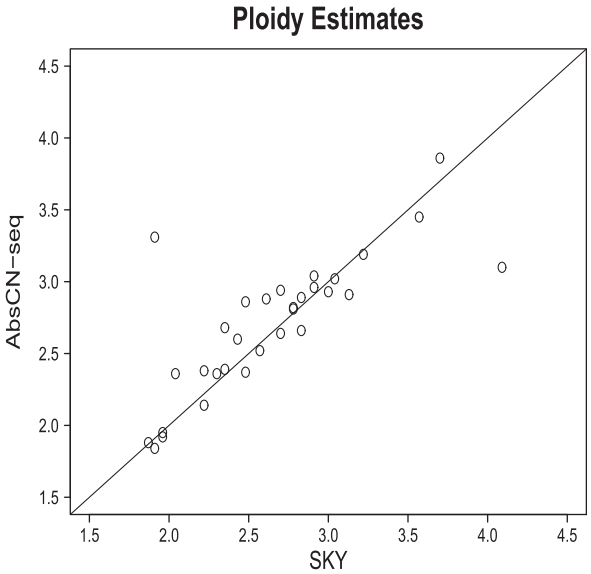

Fig. 2. Ploidy estimates of Abs-CNseq versus SKY for 31 NCI cell lines. Spearman rank correlation coefficient = 0.80, $P < 0.001$; RMSE = 0.34

biased toward lower purity in the presence of segmentation errors.

3.2 Performance on real data

3.2.1 NCI60 cell line dataset We downloaded WES data for 38 cell lines with the first letter from A to P. Unlike the patient tumor samples, the cell lines do not have their paired germ line exomes. Therefore, we used the nearly diploid hematopoietic cell line SR as a common reference for all other cell lines. Seven cell lines with unreasonably small or large mean coverage ratios were excluded. We ran our method on the remaining 31 cell lines and compared our ploidy estimates to the SKY-determined values (Fig. 2 and Supplementary Table S1). Throughout the article, we recorded the top two solutions in terms of minimum fitting errors and listed in the tables the one that better matches the benchmark value. Using this approach, we obtained close ploidy estimates from the WES data as compared with the SKY benchmark values for 29 of 31 (94%) cell lines. The top first solutions and second solutions matched the SKY values for 22 and 7 cell lines, respectively. Among the two outliers, HOP-92 is known for presence of a large number of structural rearrangements (Roschke *et al.*, 2003) that potentially complicated the analysis. The reason why HCC2998 was the other outlier was unclear to us. If these two outliers are excluded, the Spearman correlation coefficient is 0.93 ($P < 0.0001$), the root mean squared error (RMSE) is 0.16 and the concordance correlation coefficient (Lin, 1989) is 0.94, indicating high concordance between our ploidy estimates and the SKY values for most cell lines. If using only the top ranked solution for all 31 cell lines (as compared with using 24 first and 7 second solutions as above), the RMSE is 0.55, similar to the performance of published SNP-based methods on similar cell lines (Carter *et al.*, 2012).

3.2.2 dbGAP breast cancer dataset We used a publicly available exome sequencing dataset from 16 breast cancer patients (Banerji *et al.*, 2012), which has published estimates of purity and ploidy from SNP array data using ABSOLUTE (Carter *et al.*, 2012). ABSOLUTE is a well-established method to estimate purity and ploidy from SNP array data, and we used its estimates as a gold standard.

We applied AbsCN-seq, both with and without SNV information, to the exome sequencing data and compared the resulting estimates of purity and ploidy with those of ABSOLUTE (Table 2). The number of SNVs called in these samples ranges from 17 to 154. As expected, ploidy estimates are almost the same for AbsCN-Seq with and without SNV information unless distinct solutions are selected, as Equation (4) shows that SNV frequency does not provide direct information about ploidy. For estimated purity, AbsCN-seq incorporating SNV information has a higher concordance with ABSOLUTE than AbsCN-seq without this information. Hence, we will take AbsCN-seq incorporating SNV information as our default algorithm.

Overall, the estimates of purity and ploidy from AbsCN-seq with SNV information have good concordance with ABSOLUTE (Fig. 3 and Table 2). ABSOLUTE gave no estimate for three subjects (18.8% no call rate), and AbsCN-seq gave no estimate for one subject (6.3% no call rate). AbsCN-seq failed to converge to a biologically plausible solution for subject 27, leaving 12 subjects remaining for comparison. The top ranked first solutions and second solutions matched the ABSOLUTE estimates for 9 and 1 tumors, respectively. Subject BR-M-074 is the only subject with a large discrepancy in both estimated purity and ploidy (Fig. 3). ABSOLUTE estimates low purity (0.25) and high ploidy (3.88), assigning this sample the extreme estimate in both measures. By comparison, AbsCN-seq approximately

Table 2. Purity $\hat{\alpha}$ and ploidy $\hat{\tau}$ for breast tumor samples comparing published estimates from ABSOLUTE using SNP array data and AbsCN-seq using WES, with and without SNV data

Absolute			AbsCN-seq				Solution rank
			With SNVs		Without SNVs		
PID	$\hat{\alpha}$	$\hat{\tau}$	$\hat{\alpha}$	$\hat{\tau}$	$\hat{\alpha}$	$\hat{\tau}$	
5	0.59	2.08	0.53	2.11	0.6	2.08	1
26	0.44	2.01	0.49	1.99	0.53	1.99	1
27	0.31	3.12	–	–	–	–	–
28	0.58	2.18	0.58	2.18	0.64	2.18	1
30	0.50	2.04	0.47	1.98	0.49	2.00	1
34	–	–	0.49	2.32	0.56	2.29	1
37	–	–	0.49	2.00	0.55	2.00	1
38	0.42	1.90	0.52	2.90	0.50	2.91	*
41	0.34	2.03	0.46	1.99	0.47	2.01	1
45	0.56	2.67	0.62	2.64	0.53	1.66	1
50	0.63	2.13	0.53	2.13	0.58	2.13	1
55	0.62	3.44	0.66	3.35	0.54	2.31	2
74	0.25	3.88	0.50	1.93	0.56	1.94	[a]
76	0.35	1.79	0.46	1.74	0.50	1.74	1
80	–	–	0.49	2.02	0.51	2.02	1
83	0.30	2.00	0.45	2.00	0.48	2.00	1

PID: Patient ID; '–' estimates not available. Solution rank '[a]' indicates neither the first nor the second ranked solution seems to match the benchmark value, and first rank solution is listed.

doubles the estimated purity (0.51) and halves the ploidy (1.95), giving estimates for this sample near the middle of the group. Thus, ABSOLUTE would appear to infer a whole-genome doubling event, which AbsCN-seq does not. Notably, the ABSOLUTE solution fits the exome sequencing data poorly and does not belong to the stable solution set obtained by AbsCN-seq. Thus, there appears to be a true discrepancy in estimated purity and ploidy for subject BR-M-074, with ABSOLUTE providing extreme, and AbsCN-seq mid-range, estimates. In particular, the exome sequencing data do not appear to support a whole-genome doubling event. Hence, we will take the AbsCN-seq estimate as reasonable for this sample, and omit it from further comparisons. A second subject, BR-M-038, has an alternate AbsCN-seq estimate in the stable solution set that closely matches the ABSOLUTE estimates (purity 0.49 and ploidy 1.93, shown as an open triangle on Fig. 3b), although this alternate estimate did not belong to the top two solutions as measured by the AbsCN-seq objective function. It is plausible that, for subject BR-M-038, the discrepancy between the two methods might result from which ABSOLUTE uses an external database with expected frequency of karyotype information to adjust the rank of the observed solutions.

Comparing the two series of estimates across all 12 subjects, the RMSE is 0.20 for purity and 0.35 for ploidy, respectively. Again, the moderate discrepancy of ploidy estimates is driven by the outlier subject BR-M-074 discussed earlier in the text. When this outlier subject is disregarded, the RMSE becomes 0.13 for purity and 0.22 for ploidy, respectively. Averaged over the 11 samples, the estimated purity for ABSOLUTE is 0.48 and for AbsCN-seq is 0.52, suggesting that on these samples, the

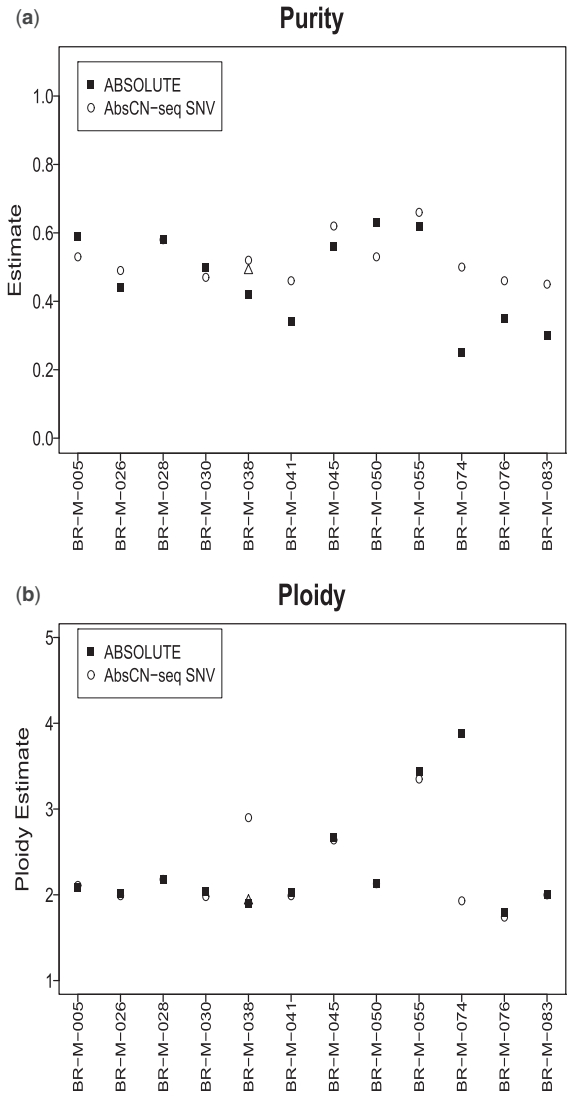

Fig. 3. Comparison of purity (**a**) and ploidy (**b**) estimates from AbsCN-seq (exome data) versus ABSOLUTE (SNP array data). The alternate estimate for BR-M-038 mentioned in the main text is shown as an open triangle

SNP-array–based ABSOLUTE estimates tend to give somewhat lower estimated purity on average than AbsCN-seq. Estimated ploidy is similar between the two algorithms, generally within 3%. Recall that the two algorithms are using independent sets of data, exome sequencing data for AbsCN-seq and SNP array data for ABSOLUTE, and so this performance appears to be reasonable.

After determining the most likely purity and ploidy, AbsCN-seq computed integer copy numbers q_i's and s_j's for each segment. A typical graphical representation of the CNAs on each chromosome is shown in Figure 4, for subject BR-M-030 and chromosome 16.

3.2.3 TCGA tumor dataset To evaluate whether our method can be applied equally well to different tumor types other than cell lines and breast tumors, we assessed our method on ~20 prostate and 20 GBM tumor samples from the TCGA database.

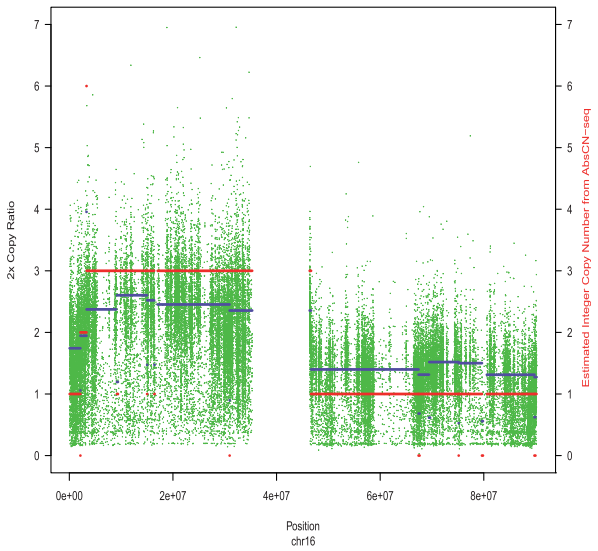

Fig. 4. Assignment of integer copy numbers to the segments on chromosome 16 for subject BR-M-030. Raw copy ratios and segmented copy ratios between the tumor DNA mixture and the match germ line DNA are in green and blue, respectively. The estimated absolute copy numbers in pure tumor cells are in red. As a copy ratio of 1.0 corresponds to a normal diploid segment, we multiply the copy ratios by two to bring the copy ratio and the absolute copy number to the same scale

We compared the purity and ploidy estimates by our method with those from ABSOLUTE, applied to the same data. We excluded a few samples that ABSOLUTE output marked as either 'high entropy' or 'low purity', with 16 prostate and 17 GBM samples remaining for comparison (Fig. 5 and Supplementary Table S3). Figure 5 shows that the two estimates are highly concordant over a broad purity and ploidy range. The Spearman correlation for purity and ploidy is 0.80 ($P<0.0001$) and 0.91 ($P<0.0001$), respectively, and the concordance correlation coefficients are 0.78 and 0.89, respectively. Averaged over the 33 samples, the estimated purity for ABSOLUTE is 0.581 and for AbsCN-seq is 0.589. We no longer observed the minor discrepancy we saw in the breast tumor samples, implying that the minor discrepancy might be induced by the difference between platforms. Interestingly, the distribution of the ploidy estimates for each tumor type (Fig. 5b) recapitulated what was observed in SNP array data for a larger sample size [c.f. Fig. 3c in (Carter *et al.*, 2012)]. For example, prostate tumor ploidy shows a bimodal distribution with peaks at 2.0 and >3.5, and few intermediate cases. In contrast, the GBMs tend to have continuous ploidy in the range from 2.0 to 4.0. This cross-platform consistent pattern not only substantiates that our random samples are representative of the larger population, but also provides additional evidence that our estimates as a whole are likely reasonable.

4 DISCUSSION

We have developed a statistical approach, AbsCN-seq, which can estimate tumor purity and ploidy from WES data, and then use these estimates to infer absolute copy number and absolute multiplicity of somatic SNVs in the tumor. The ploidy

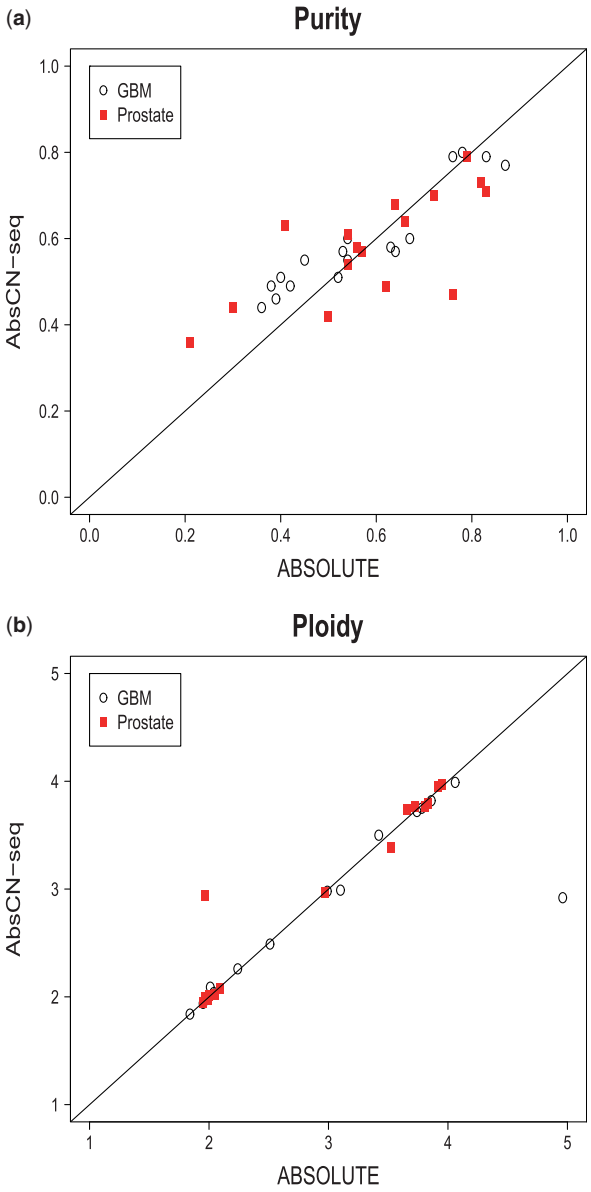

Fig. 5. Comparison of purity (**a**) and ploidy (**b**) estimates from AbsCN-seq versus ABSOLUTE for 17 GBM samples and 16 prostate samples

estimates of our method fit the SKY values well. Our method also has good concordance with ABSOLUTE, although the two differ in several aspects. First, the input data are generated from different high-throughput platforms, ABSOLUTE, in it standard form, using SNP array data while our method takes NGS data. Second, the statistical framework and implementation differ. ABSOLUTE uses a Bayesian framework, whereas our method uses a simpler frequentist approach. Third, ABSOLUTE uses additional independent information to infer which solution is most likely, such as empirical karyotype models, while our solution refers only to the data from the sample and chooses the estimate that gives the best least-squares fit to the data. Given these differences, it is encouraging to see the high concordance between these two methods on our validation datasets. For the breast cancer patients, ABSOLUTE did not make calls on three

subjects, and our method made reasonable calls on all three. We are unclear whether the failure of ABSOLUTE to produce an estimate for these subjects was due to a failure in SNP array experiment or to the algorithm itself. Thus, it seems that AbsCN-seq can work for certain samples where ABSOLUTE cannot make a call. On the other hand, our method failed to produce estimates in one case where ABSOLUTE could make an estimate, thus providing complementary approaches. As has been seen by ABSOLUTE, the best fitting solution is not always the correct solution. The validation studies here show that the top two solutions of our method usually cover the correct answer, but exceptions do exist. We recommend the user to manually inspect each solution whenever possible. Although our method does not yet incorporate subclonal tumor heterogeneity, it has been shown to be robust to a considerable amount of such heterogeneity, as well as to segmentation errors, in the estimates of purity and ploidy.

Our method is flexible in the sense that it can also be applied to shallow whole-genome sequencing data (Li *et al.*, 2011), an emerging NGS application to identification of copy number variants genome-wide. However, presence of somatic mutations cannot be reliably called in these type of data, let alone the corresponding variant allele frequency. For these type of data, we can set the algorithm to only use copy number information, by setting $\lambda = 1$. The reduced objective function (first term of Equation 5) in step 2 now has a closed form solution. Define a new parameter

$$\gamma = \alpha/(\alpha\tau + 2(1-\alpha)).$$

It is easy to see that when q_i's are given, Equation (3) reduces to a simple linear regression on q_i's. Then α can be obtained from the γ estimate. A caveat arises when we apply this form of AbsCN-seq to the WES data as seen in Figure 3a. We observed a systematic upward bias in the purity estimates as compared with the SNP-based ABSOLUTE estimates. However, by incorporating the SNV information, such bias was greatly diminished, and this did not appear to affect the ploidy estimates. This might not be a concern in practice, as for WES data, we will always use the default form with SNV information. At this moment, it is not clear to us what is the major cause for the small remaining observed bias. Platform-specific biases in the raw copy ratio data could be one possible reason. For instance, as shown by Carter *et al.* (2012), the algorithm relying on the Illumina SNP array data (Van Loo *et al.*, 2010) systematically underestimate tumor purity, as compared with the ABSOLUTE method relying on the Affymetrix SNP array data. Likewise, the apparent small overestimation of purity by our method as compared with ABSOLUTE might also be attributed to the characteristics of WES data compared with SNP array data. A supporting evidence is that we did not observe such minor biases when the platform is fixed. We have not tested our method on shallow WGS data with reference to purity estimates and we do not know yet if such overestimation also exists in the WGS data. Future work will examine the performance of the method on shallow WGS data, when a benchmark dataset becomes available.

Funding: This work was supported by [NIH U54HL108460] and [R01CA166293-01A1] as well as a grant from the Breast Cancer Research Foundation (Dr Barbara Parker, PI). The authors thank Viswanath Nandigam at the Moores Cancer Center for configuring the Apache. They also thank dbGAP team to grant them access to the data.

Conflict of Interest: none declared.

REFERENCES

Abaan,O.D. *et al.* (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.*, **73**, 4372–4382.

Attiyeh,E.F. *et al.* (2009) Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res.*, **19**, 276–283.

Banerji,S. *et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405–409.

Bengtsson,H. *et al.* (2010) TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics*, **11**, 245.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Greenman,C.D. *et al.* (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Gusnanto,A. *et al.* (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**, 40–47.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

LaFramboise,T. *et al.* (2005) Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput. Biol.*, **1**, e65.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li,Y. *et al.* (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.

Lin,L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.

Mailman,M.D. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.

Oesper,L. *et al.* (2013) Inferring Intra-tumor Heterogeneity from High-Throughput DNA Sequencing Data. *Res. Comput. Mol. Biol.*, **7821**, 171–172.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Roschke,A.V. *et al.* (2003) Karyotypic complexity of the NCI-60 drug-screening panel. *Cancer Res.*, **63**, 8634–8647.

Rueda,O.M. and Diaz-Uriarte,R. (2007) Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput. Biol.*, **3**, e122.

Su,X. *et al.* (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, **28**, 2265–2266.

The Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Van Loo,P. *et al.* (2010) Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA*, **107**, 16910–16915.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Yau,C. *et al.* (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.*, **11**, R92.

Yu,G. *et al.* (2011) BACOM: in silico detection of genomic deletion types and correction of normal cell contamination in copy number data. *Bioinformatics*, **27**, 1473–1480.

Zanetti,M. *et al.* (2012) Immune surveillance from Chormosomal Chaos? *Science*, **337**, 1616–1617.