Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2016 *Paper* 29

FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool Estimates for High-Throughput DNA Sequencing

Ronglai Shen* Venkatraman Seshan†

*Memorial Sloan-Kettering Cancer Center, shenr@mskcc.org

†Memorial Sloan-Kettering Cancer Center, seshanv@mskcc.org

http://biostats.bepress.com/mskccbiostat/paper29

FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool Estimates for High-Throughput DNA Sequencing

Ronglai Shen and Venkatraman Seshan

Abstract

Allele-specific copy number analysis (ASCN) from next generation sequenc- ing (NGS) data can greatly extend the utility of NGS beyond the iden- tification of mutations to precisely annotate the genome for the detection of homozygous/heterozygous deletions, copy-neutral loss-of-heterozygosity (LOH), allele-specific gains/amplifications. In addition, as targeted gene panels are increasingly used in clinical sequencing studies for the detection of "actionable" mutations and copy number alterations to guide treatment decisions, accurate, tumor purity-, ploidy-, and clonal heterogeneity-adjusted integer copy number calls are greatly needed to more reliably interpret NGS- based cancer gene copy number data in the context of clinical sequencing. We developed FACETS, an ASCN tool and open-source software with a broad application to whole genome, whole-exome, as well as targeted panel sequencing platforms. It is a fully integrated stand-alone pipeline that in- cludes sequencing BAM file post-processing, joint segmentation of total- and allele-specific read counts, and integer copy number calls corrected for tumor purity, ploidy and clonal heterogeneity, with comprehensive output and integrated visualization. We demonstrate the application of FACETS using the Cancer Genome Atlas (TCGA) whole-exome sequencing of lung adenocarci- noma samples. We also demonstrate its application to a clinical sequencing platform based on a targeted gene panel.

FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool for High-throughput DNA Sequencing

Ronglai Shen and Venkatraman Seshan
Memorial Sloan-Kettering Cancer Center

1 Abstract

Allele-specific copy number analysis (ASCN) from next generation sequencing (NGS) data can greatly extend the utility of NGS beyond the identification of mutations to precisely annotate the genome for the detection of homozygous/heterozygous deletions, copy-neutral loss-of-heterozygosity (LOH), allele-specific gains/amplifications. In addition, as targeted gene panels are increasingly used in clinical sequencing studies for the detection of "actionable" mutations and copy number alterations to guide treatment decisions, accurate, tumor purity-, ploidy-, and clonal heterogeneity-adjusted integer copy number calls are greatly needed to more reliably interpret NGS-based cancer gene copy number data in the context of clinical sequencing. We developed FACETS, an ASCN tool and open-source software with a broad application to whole genome, whole-exome, as well as targeted panel sequencing platforms. It is a fully integrated stand-alone pipeline that includes sequencing BAM file post-processing, joint segmentation of total- and allele-specific read counts, and integer copy number calls corrected for tumor purity, ploidy and clonal heterogeneity, with comprehensive output and integrated visualization. We demonstrate the application of FACETS using the Cancer Genome Atlas (TCGA) whole-exome sequencing of lung adenocarcinoma samples. We also demonstrate its application to a clinical sequencing platform based on a targeted gene panel.

2 Introduction

Large-scale sequencing studies including the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) projects have generated tens of thousands whole-genomes (WGS) and whole-exomes (WES) of tumor-normal sample pairs. Allele-specific copy number analysis can greatly extend the utility of sequencing data beyond the identification of mutations. We present FACETS (which stands for Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing), an ASCN analysis pipeline and open-source software for next generation sequencing (NGS) data.

ASCN analysis has several major advantages over conventional total copy number analysis. First, it provides a much more comprehensive identification of copy number aberrations including copy-neutral LOH events not detectable by analyzing total copy number alone. Thus genome-wide LOH pattern can be systematically evaluated. In addition, while conventional analysis typically converts total copy number ratio into qualitative copy number states (high versus low level gains, shallow versus deep losses, normal), ASCN analysis can be used to precisely annotate the genome for the detection of homozygous deletions, heterozygous deletions, copy-neutral LOH, allele-specific gains and amplifications with corresponding integer copy number. Furthermore, ASCN analysis provides more accurate estimates of tumor purity and ploidy. The output can be used for enhanced clonal heterogeneity analyses of somatic point mutations.

Early ASCN methods were primarily developed for copy number array platforms (Rasmussen et al., 2011, Sun et al., 2009, Van Loo et al., 2010, Yau et al., 2010). More recently, a number of ASCN methods have been developed for next generation sequencing data, building on different analytical strategies. Patchwork (Mayrhofer et al., 2013) segments the genome based on total read count and then estimates the allele-specific copy number within each segment. The limitation lies in that segmenting total read count alone does not provide the complete picture and will inevitably miss certain events such as copy neutral LOH (Figure 1). Falcon (Chen et al., 2014) provides a joint segmentation procedure using a Binomial process for the allelic read count from heterozygous SNP loci. Several other methods including TITEN (Ha et al., 2014) further considered tumor purity and clonal heterogeneity to enhance the accuracy of copy number analysis by using various probabilistic

modeling approaches including Bayesian mixture model (Chen et al., 2013), Hidden Markov Model (Ha et al., 2014) or other maximum likelihood methods (Li and Xie, 2014, Oesper et al., 2013).

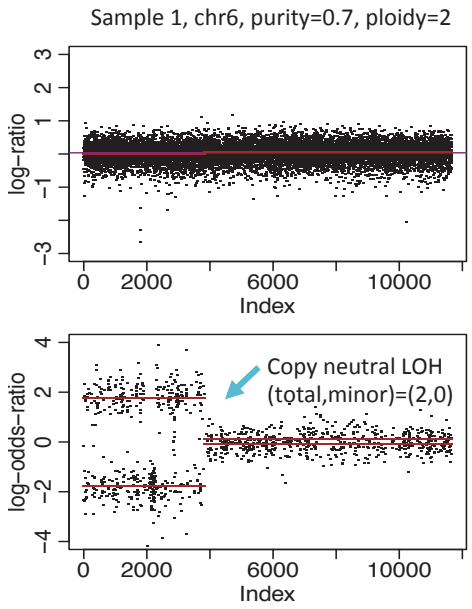

Figure 1: Joint segmentation identifies copy number neutral loss-of-heterozygosity (LOH) event. Top panel shows copy number log-ratio of total sequence read count in the tumor to that in the normal along genomic positions on chromosome 6 from a whole-exome sequencing of a lung cancer patient sample. Second panel shows the allelic log-odds-ratio of the variant allele read counts in the tumor/normal pair revealing a copy-neutral LOH event on 6p.

FACETS provides several unique contributions over existing methods. For one, we employ a nonparametric joint segmentation approach based on a Hotelling T^2 statistic by directly combining the total and allele-specific read counts which does not depend on any model assumption and provides a fast implementation to search for change points in the genome.

ASCN analysis typically uses a SNP-based approach as allelic imbalance can only be measured at heterozygous sites. Nearly all ASCN methods for sequencing data uses read count information from heterozygous sites only.

However, heterozygous sites are subject specific and sparse which leads to information loss on total copy number. Thus a systematic enumeration of allele specific read counts from all SNPs, be it heterozygous or homozygous, provides full information on both total and allele specific copy numbers. Furthermore we also use read counts from a set of pseudo-SNPs (non polymorphic loci) along the target intervals so that regions with large gaps between consecutive SNPs are represented in total copy number analysis. In total copy number analysis, a moving window approach in which the read depths are averaged over all the loci within the window is used commonly. However, since the independent units of measurement are DNA fragments this leads to serial correlation as the same fragment contributes to read depth at several loci. Our approach of using read counts at SNPs that are sufficiently spaced from one another provide a way of obtaining information that have negligible serial correlation since each fragment is usually mapped to only one SNP locus. To address the imbalance in the number of loci used for total and allele specific copy numbers we introduce a weighting scheme that is inversely proportional to the overall heterozygous rate in the patient's genome which further enhances the detection of allele-specific alterations.

In addition, the current sequencing analysis methods for allelic imbalances based on B-allele frequency (BAF) has some inherent biases due to differential mapping affinity between the reference and the variant allele. To address this issue, we show that the allelic log-odds-ratio (logOR) metric provides an unbiased estimate of the allelic ratio by leveraging the paired tumor-normal sequencing design that cancels out the mapping bias. To obtain allele-specific copy number calls, we devised a Gaussian-non-central χ^2 mixture model. Tumor purity, ploidy, and clonal heterogeneity are factored in the model to obtain accurate ASCN output and facilitates the identification of subclonal events.

FACETS provides a complete analysis pipeline that include BAM file post-processing steps including library size and GC-normalization, joint segmentation of total and allele-specific signals, and integer copy number calls taking into account of tumor purity, ploidy, and clonal heterogeneity, all seamlessly integrated in a single workflow with comprehensive output, integrated visualization, with fast computation to facilitate large-scale application. Figure 2 shows FACETS analysis of a TCGA chromophobe renal cell carcinoma (chRCC) sample (TCGA-KL-8331), revealing multiple chro-

mosomal losses including chromosomes 1, 2, 6, 10, 13, 17, 21 which are signatures of chRCC genome alteration as characterized in the TCGA chRCC study (Davis et al., 2014). In addition, two major subclonal clusters of losses unique in this tumor sample were further identified that included chr 11, 18, and 22 representing events occurring later in time.

Most existing methods are designed for WGS or WES. As targeted panel sequencing is increasingly used in clinical settings to detect "actionable" mutations and copy number alterations toward precision medicine, robust copy number and clonal heterogeneity analysis tools such as FACETS for targeted panel sequencing are needed to further increase the clinical utility of NGS. The software is available at `https://sites.google.com/site/mskfacets/`.

In this paper, we benchmark our tumor purity and ploidy estimates using the TCGA whole-exome sequencing data in 286 lung adenocarcinoma samples and compared with the estimates from the ABSOLUTE algorithm (Carter et al., 2012). We show that FACETS can enhance the sensitivity of identifying aneuploid tumors by joint modeling of total and allele-specific pattern. In addition, as shown in Figure 2, FACETS facilitates systematic identification of clonal and subclonal copy number events through a cellular fraction feature in the model. Moreover, accurate, purity-, ploidy-, and clonal heterogeneity-adjusted, integer copy number calls will be essential to reliably interpret NGS-based gene copy number calls in clinical sequencing panels. We will demonstrate that using a clinical sequencing sample profiled by the MSK-IMPACT platform (Cheng et al., 2015).

3 MATERIALS AND METHODS

In the next sections, we discuss our approach for sequencing bias corrections, joint segmentation of total and allelic copy ratio, and methods for integer copy number calls correcting for tumor purity, ploidy, and intratumor heterogeneity.

3.1 Total copy number log-ratio (logR).

Sequence read count information are first parsed form paired tumor-normal BAM files (Figure 3A). A normalizing constant is calculated for each tu-

mor/normal pair to correct for total library size. Subsampling within 150-250bp intervals is applied to reduce hypersegmentation in SNP-dense regions of the genome (Figure 3B). logR is then computed from the total read count in the tumor versus normal for all SNPs that have a minimum depth of cov-

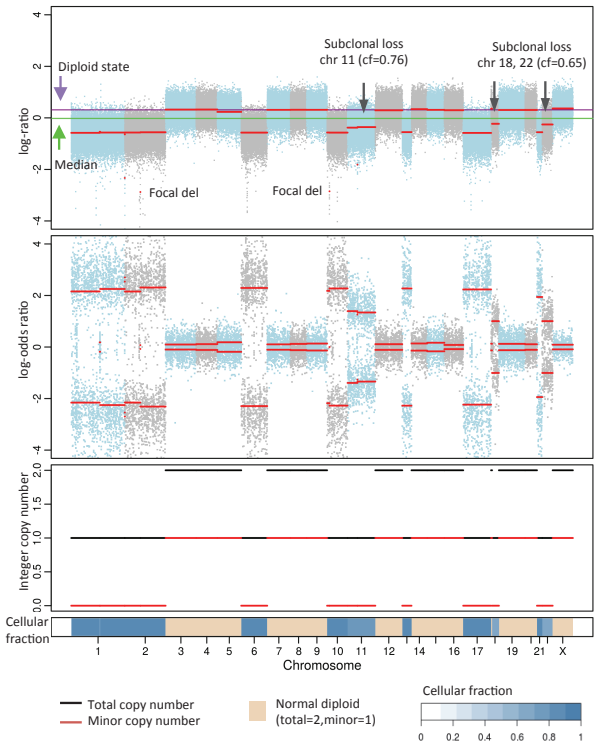

Figure 2: Integrated visualization of FACETS analysis of whole-exome sequencing data from a TCGA chromophobe renal cell carcinoma sample (TCGA-KL-8331). The top panel displays total copy number log-ratio (logR), and the second panel displays allele-specific log-odds-ratio data (logOR) with chromosomes alternating in blue and gray. The third panel plots the corresponding integer (total, minor) copy number calls. The overall tumor ploidy is estimated to be 1.6, revealing a hypodiploid tumor genome due to the whole-chromosomal losses of multiple chromosomes. The tumor sample purity is estimated to be 0.89. The estimated cellular fraction (cf) profile is plotted at the bottom, revealing both clonal and subclonal copy number events.

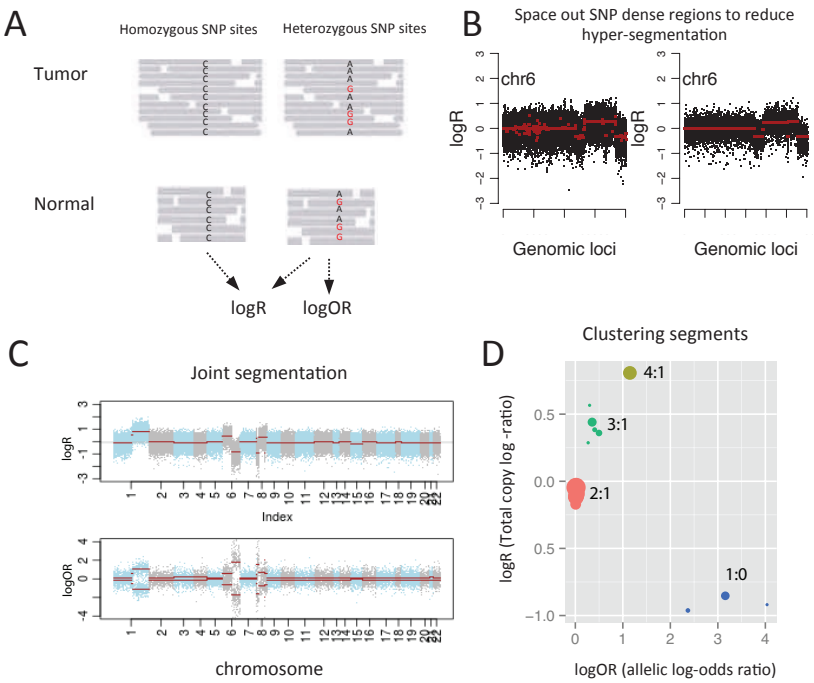

Figure 3: Pre-processing and joint segmentation. A. Parsing reference and variant allele count for SNP sites from tumor-nomal sequencing BAM files. All SNP sites contribute to total copy log-ratio (logR), and heterozygous sites contribute to allelic log-odds ratio (logOR). B. Interval-sampling to reduce local serial dependencies in SNP-dense regions. C. Joint segmentation logR and logOR and the detection of copy number aberrant regions of the genome. D. Segment clustering to form groups with the same latent copy number states.

erage in the normal. logR provides information on total copy number ratio. Specifically, the expected value of logR can be expressed as

$$E[\text{logR}] = log\{(m^* + p^*)/2\} + w(\cdot) + \lambda,$$

where $m^* = m\phi + (1-\phi)$ and $p^* = p\phi + (1-\phi)$ are parental copy number in the tumor sample rising from a mixed normal (1,1) and aberrant (m,p) copy number genotype with mixing proportion ϕ. We term ϕ as the cellular fraction associated with the aberrant genotype, which is a function of tumor purity and clonal frequency (for subclonal alterations). The term $w(\cdot)$ denotes systematic bias. Here we explicitly consider GC-content and use loess regression of logR over GC in 1kb windows along the genome to estimate the GC-effect on read counts and subtract it from logR. In addition, we note

that logR quantifies relative copy number, hence we introduce a constant λ for absolute copy number conversion which will be described in detail later.

3.2 Allelic copy number log odds-ratio (logOR).

Allelic imbalance analysis has been typically based on B-allele (or variant allele) frequency (BAF) in the tumor which informs $m^*/(m^* + p^*)$. In sequencing data, it has been observed that there is a significant bias toward higher mapping rates for the reference allele compared to those for the variant allele at heterozygous loci (Degner et al., 2009). Such bias can significantly impact allele-specific copy number inference if not corrected. To illustrate, let r denote the relative mapping affinity of the variant allele to the reference allele, and typically $r < 1$ (mapping biased in favor of the reference allele). As a result, the normal genotype becomes (1,r) instead of (1,1) and the aberrant genotype becomes (m^*, rp^*) or (p^*, rm^*) (Table 1). Therefore it is easy to see that in sequencing data, BAF in fact informs $m^*/(m^* + rp^*)$, which is a biased estimate of B-allele frequency when $r \neq 1$. To address this issue, we propose to use the log-odds ratio (logOR) of the variant-allele count in tumor versus normal, which is an unbiased estimate of allelic copy ratio. In particular,

$$E[\text{logOR}] = \log(m^*/p^*) \text{ or } \log(p^*/m^*),$$

depending on which parental copy the variant allele resides on. Since we do not have phased data, squared logOR is used to infer $\log^2(m^*/p^*)$.

Table 1: Illustration of how differential mapping bias affects copy number inference.

	Reference allele on maternal copy		Reference allele on paternal copy	
	Reference	Variant	Reference	Variant
Normal	1	r	1	r
Tumor	m^*	rp^*	p^*	rm^*

3.3 Joint segmentation

Segmentation analysis identifies regions of the genome that have constant copy number using change point detection methods. Conventional methods

(e.g. BIC-seq (Xi et al., 2010), ExomeCNV (Sathirapongsasuti et al., 2011)) typically perform one-dimensional segmentation using logR alone, or separate application of one-dimensional segmentation to logR and BAF. Yet a truly joint segmentation can significantly improve the precision for change point detection and downstream analysis for estimating tumor purity, ploidy and allele-specific calls.

To address this challenge, we extended the Circular Binary Segmentation (CBS) algorithm (Olshen et al., 2004, Venkatraman and Olshen, 2007) to a joint segmentation of logR and logOR based on a bivariate Hotelling T^2 statistic:

$$T^2 = \max_{1 \le i < j \le n} T_{1ij}^2 + cT_{2ij}^2$$

where T_{1ij} is the Mann-Whitney statistic comparing the set of observed logR denoted as $\{X_{1k} : i < k \le j\}$ and its complement $\{X_{1k} : 1 < k \le i \text{ or } j < k \le n\}$. and similarly T_{2ij} is the Mann-Whitney statistic comparing the set of observed logOR denoted as $\{X_{2k} : i < k \le j\}$ and its complement $\{X_{2k} : 1 < k \le i \text{ or } j < k \le n\}$. In the above, c is a scaling factor that is inversely proportional to the heterozygous rate which will be discussed shortly.

Here if the maximal statistic is greater than a pre-determined critical value, we declare that a change exists and estimate the change-points as i, j that maximize the statistic. The algorithm iteratively searches for change points between any possible pair of breakpoints and its complement to identify regions of the genome that have constant allele-specific copy number. For each segment, the logR data is summarized using the median of the logR values $\tilde{x}_1$ and the logOR data are summarized by $\tilde{x}_2^2$ which takes the form $\sum\{x_2^2 - s^2)/s^2\}/\sum\{1/s^2\}$ where s^2 is the estimated variance of logOR.

We point out that while logR is defined for all SNPs (both homozygous and heterozygous loci), logOR is only defined for heterozygous loci (het-loci or het-SNPs). This creates a large imbalance between the two in the combined statistic. To address this issue, we introduce a weight that is inversely proportional to the heterozygous rate to increase the het-SNP contributions in subsequent segmentation analysis. Specifically, a scaling factor c is introduced in the T^2 statistic. This is empirically set at $1/\sqrt{4\gamma}$ where γ is the proportion of het-SNPs in the patient sample. Up-weighing the contribution of logOR for het-SNPs increases the power of detecting allelic imbalances

for regions with low frequency of het-SNPs. We denote this a "full model" approach.

Our "full model" approach is distinct from the conventional method in which both logR and BAF are computed for het-loci only. For the whole-exome data we have analyzed, the genome-wide heterozygous rate typically ranges from 10-15%. As such, the het-SNP approach can lead to substantial information loss and reduced power for detecting alterations across the genome. To illustrate, we conducted a down-sampling experiment using two whole-exome samples with high and low tumor purity to assess the sensitivity of detecting genome alterations between the full model and the typical het-SNP approach (Supplementary Figure 1). For low purity tumor samples, the het-SNP approach shows reduced sensitivity at the outset. As genome coverage decreases by down-sampling, the sensitivity of het-SNP only approach for detecting all the altered copy number segments quickly deteriorates while the full model holds up substantially better.

After segmentation, we cluster the segments into groups of the same underlying genotype. Figure 3D shows an example in which a total of 27 segments resulting from the joint segmentation were clustered into four distinct genotype groups. Such clustering reduces the number of latent copy number and cellular fraction states needed in subsequent modeling.

3.4 Determine the 2-copy state.

As mentioned earlier, logR estimates are proportional to the absolute total copy number up to a location constant λ. For diploid genome, logR = 0 (library size normalized logR) is the location for the 2-copy state . However, aneuploidy can lead to a location shift in the tumor. Therefore we need to first determine the 2-copy state in a tumor genome and quantify the location shift λ. Without adjusting for the location shift, absolute copy number calls are not possible.

Let us denote the copy number states using total and minor integer copy number (e.g., 1-0 denotes monosomy with total copy number 1 and minor copy number 0). The estimate of λ should correspond to the logR level at which the segments are in 2-1 (normal diploid) or 2-0 (copy-neutral LOH) state. In order to estimate λ, we first note that normal diploid segments

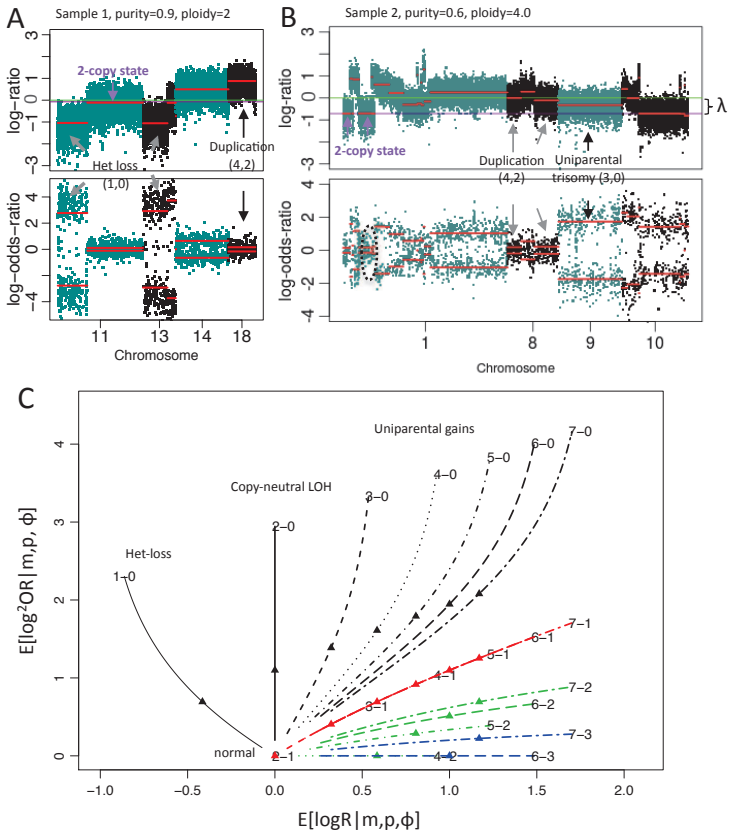

Figure 4: Joint analysis of total and allelic copy number pattern to more accurately estimate tumor purity, ploidy and the precise genotypes of the copy number alterations. Two examples (A, B) are presented here to illustrate the use of allelicly balanced segments (log-odds-ratio close to zero) to determine the 2-copy state (purple line) and location shift λ in total copy number log-ratio (logR) due to aneuploidy of the tumor. The expected value of logR and logOR as a function of total and minor copy number and cellular fraction ϕ are plotted to show the degree of separability among different copy number genotype and cellular fraction (C). Each line traces the cellular fraction from low (0.1) at the original point close to (0,0) to high (0.9) on the other end of the line. Triangles mark the cellular fraction of 0.5 on each line. The colors represent the minor copy number: 0 is black, 1 is red, 2 is green and 3 is blue. Line types change by total copy number.

should be allelically balanced. Thus candidate value for λ (referred to as λ_c) will be obtained from $\tilde{x}_1$ for segment clusters that have $\tilde{x}_2$ values close to zero.

However, note that homozygous deletions (0-0) and balanced gains (4-2,

6-3 etc.) are also allelically balanced and hence will have small $\tilde{x}_2$. Since large scale homozygous deletions of multiple genes will not be conducive to cell survival we can eliminate non-focal segments with small $\tilde{x}_2$ as being homozygous deletions. In addition, for the sake of simplicity we do not consider higher order balanced gains states (6-3, 8-4 etc.) spanning a large part of the genome. Finally, samples in which segments with allelic balance are a small fraction of targeted regions will be flagged and will require a manual review for their λ estimates.

In samples that have large allelically balanced segments, there can be several $\tilde{x}_1$ values from which λ_c can be chosen. The samples in Figures 2 and 3C have several balanced segments with $\tilde{x}_1$ values with small variation around a single level. The samples in Figures 4A and 4B have segments with allelic balance at two distinct $\tilde{x}_1$ levels (chr11q and chr18 in 4A and parts of chr1 and chr8 in 4B). We group the balanced segments into either one or two distinct levels. For the single level scenario the choice of λ_c is obvious whereas in the two distinct levels scenario the higher level cannot be normal diploid (since it would imply the lower level is large scale homozygous deletion) and thus λ_c should the lower one.

We proceed by evaluating whether λ_c represents normal diploid state or balanced 4 copy state using all segments that are reltive losses *i.e.* segments with $\tilde{x}_1$ smaller than the candidate level. If it represents the 2-1 state then the losses are at 1-0 state whereas if it represents the 4-2 state then the losses can be any of 3-0, 3-1, 2-0 or 1-0. We find the best m, p, ϕ that fits $\lambda_c - \tilde{x}_1 = \log(2 + 2 * \phi) - \log\{(m + p - 2)\phi + 2\}$ and $\tilde{x}_2^2 = \log\{[(m-1)\phi + 1]/[(p-1)\phi + 1]\}^2$. Segments at 3-1 and 2-0 states in relation to 4-2 level with a clonal ϕ is indistinguishabe from segments at 1-0 in relation to 2-1 level with different ϕ. On the other hand single copy loss from 2-1 cannot mimic the relationship between 4-2 and 3-0 or 1-0 states. Thus λ_c will be considered to represent 4-2 state if best fit copy numbers for some segments are at 3-0 or 1-0. If all the segments are assigned 3-1 or 2-0 then it will be considered to represent 4-2 if a clonal fit with single ϕ fits as well as subclonal 1 copy loss from diploid with a single subclone fraction. If λ_c represents 2-1 state then we set $\lambda = \lambda_c$ and if it represents 4-2 state then λ is estimated as the $\tilde{x}_1$ value corresponding to the 2-0 state.

In Figure 4A the balanced segments at chr11q and chr18 represent the 2-1

and 4-2 states and although there are several losses and gains the average copy number of the sample is 2 and thus λ is estimated close to zero. In Figure 4B however 2-1 state in parts of chromosome 1 is a small fraction of the genome and chr8 at 4-2 is the dominant location of allelic balance. Even ignoring the 2-1 segments in chr1 the procedure can estimate λ at the 2-0 state represented by chr10q since chr9 is at 3-0 state compare to 4-2 level in chr8. The 2-copy state for this sample is significantly shifted below zero due high average copy number of the tumor.

3.5 Integer copy number call.

In the next step, we obtain integer copy number (major and minor) and the associated cellular fraction estimates for each segment cluster by modeling the expected values of logR and logOR given total (t), and each parental (m,p) copy as a function of a cellular fraction (cf) parameter ϕ, using a combination of parametric and nonparametric methods. This allows us to model both clonal and subclonal events. Figure 4C demonstrates the expected value of logR and logOR as a function of (m,p) and ϕ. Note that the curves for most combinations of m and p are distinct and well separated indicating that they can be estimated well provided the cellular fraction is high.

The procedure starts by first obtaining a moment estimate of $\hat{t}_i$, the total copy number for segment cluster i, by $\lceil 2^{(1+\tilde{x}_{1i})} \rceil$, where $\tilde{x}_{1i}$ denote the median logR for segment cluster i corrected for sequence bias and tumor ploidy (λ-normalized). Once the total number is obtained we calculate the allele specific copy numbers m and p and the cellular fraction ϕ using the fact that the logOR summary measure $\tilde{x}^2$ is a moment estimate of μ^2 which equals $\log^2(\{m\phi + (1-\phi)\}/\{p\phi + (1-\phi)\})$.

To further refine the initial estimates, we employed a Gaussian-noncentral χ^2 model with error terms to account for the noise with a clonal structure imposed on the cellular fraction ϕ. Specifically, let X_{1ij} denote the logR for SNP loci j in segment cluster i (corrected for sequence bias and location shift) and follow a normal distribution:

$$X_{1ij} \sim N(\nu_{ig}, \tau_i^2),$$

where ν_{ig} is the expected value of logR given the underlying copy number

state g taking the form

$$\nu_{ig} = \log_2(2(1-\phi_k) + t_g\phi_k)/2,$$

where $t_g = m_g + p_g$ denotes the total copy number (sum of the two parental copy number) given the underlying copy number state g, ϕ_k denotes the cellular fraction for clonal cluster k, and τ_i^2 is an independent variance parameter. In practice, it is quite reasonable to assume homoscedasticity and set $\tau_i^2 = \tau^2 \ \forall i$.

Furthermore, let X_{2ij} denote the logOR for SNP loci j in segment cluster i and $(X_{2ij}/\sigma_{ij})^2$ follow a non-central chi-squared distribution:

$$(X_{2ij}/\sigma_{ij})^2 \sim \chi^2(\delta_{ijg}),$$

where σ_{ij}^2 is the variance parameter for logOR and $\delta_{ijg} = \mu_{ig}^2/\sigma_{ij}^2$ is the non-centrality parameter in which

$$\mu_{ig}^2 = \log^2 \frac{m_g\phi_k + (1-\phi_k)}{p_g\phi_k + (1-\phi_k)}.$$

Assuming X_{1ij} and X_{2ij} are independent random variables given the underlying copy number state g, the joint data likelihood can then be written as

$$\ell = \sum_i \sum_j \sum_g f(x_{1ij}|\nu_{ig}, \tau_i^2, g) f(x_{2ij}|\delta_{ijg}, g) P(g)$$

where $P(g)$ is the prior probability of the latent copy number state g.

We apply an expectation-maximization (EM) algorithm to maximize the joint data likelihood. It can be viewed as an estimation problem with the latent copy number states as "missing" data. In the E-step of the EM procedure, Bayes theorem is used to compute the posterior probability of segment cluster i being assigned copy number state g given the parameter estimates at the tth iteration:

$$\hat{p}_{ijg}^{(t)} = \frac{f(x_{1ij}|\hat{\nu}_{ig}^{(t)}, \hat{\tau}_i^{2^{(t)}}, g) f(x_{2ij}|\hat{\delta}_{ijg}^{(t)}) P(g)}{\sum_g f(x_{1ij}|\hat{\nu}_{ig}, \hat{\tau}_i^{2^{(t)}}, g) f(x_{2ij}|\hat{\delta}_{ijg}^{(t)}) P(g)}.$$

In the M-step, we first update the normal and non-central Chi-square distribution parameters

$$\hat{\nu}_{ig}^{(t+1)} = \frac{\sum_j \hat{p}_{ijg}^{(t)} \cdot x_{1ij}}{\sum_j \hat{p}_{ijg}^{(t)}}, \quad \hat{\tau}_i^{2^{(t)}} = \frac{\sum_j \hat{p}_{ijg}^{(t)} (x_{1ij} - \hat{\nu}_{ig}^{(t)})^2}{\sum_j \hat{p}_{ijg}^{(t)}},$$

$$\hat{\mu}_{ig}^{2^{(t+1)}} = \frac{\sum_j \hat{p}_{ijg}^{(t)} \cdot (x_{2ij}^2 - s^2)/s^2}{\sum_j \hat{p}_{ijg}^{(t)}/s^2},$$

where s^2 is the sample variance estimate of logOR. After obtaining the estimates of ν and then update the cellular fraction parameter $\phi_k^{(t+1)}$ given

$$\hat{\nu}_{ig^*}^{(t+1)} = \log \frac{2(1-\phi_k) + t_g^* \phi_k}{2}, \ \hat{\mu}_{ig^*}^{(t+1)} = \log \frac{m_g^* \phi_i + (1-\phi_k)}{n_g^* \phi_k + (1-\phi_k)},$$

where g^* is the most likely genotype (with highest posterior probability) given the data and current parameter estimates in the tth iteration. The E-step and M-step are iterated until convergence.

A clonal structure is imposed on the cellular fraction ϕ_k. This is done in a sequential approach where the algorithm starts with a single clonal cluster (k=1) with cellular fraction parameter ϕ_1. We then identify segment clusters for which segment cluster-specific estimates is non-trivially lower (at least by 0.05) from the clonally constrained estimates that result in a suboptimal fit under $k = 1$. These segment clusters with discordant cellular fraction estimates then form a candidate subclonal cluster of events at a lower cellular fraction ϕ_2, and a model is fitted with the joint likelihood optimized under $k = 2$. This procedure is iterated until no additional discordance in cellular fraction estimates are found, or a specified maximum k is reached. In the default parameter setting, a maximum $k = 5$ is allowed although user can change it to a higher number if greater intratumor heterogeneity is expected. In the output, $\hat{\phi}_1$ is the cellular fraction eatimate for the clonal events and also the tumor purity by definition, and $\hat{\phi}_k, k > 1$ for any subclonal clusters identified in the tumor sample.

Figure 5 plots the kernel density of the FACETS estimates of cellular fraction for the copy number alterations detected in the chRCC sample TCGA-KL-8831, revealing three major subclonal clusters. In this tumor

sample, $\hat{\phi}_1 = 0.89$ capturing the clonal alterations (losses of chromosomes 1,2,6,10,13,17 and 21). A subclonal cluster captured the subsequent loss of chromosomes 11 at $\hat{\phi}_2 = 0.76$, followed by additional losses of 18 and 22 at $\hat{\phi}_3 = 0.65$.

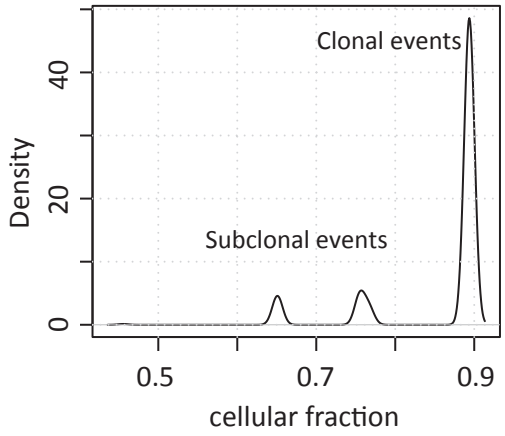

Figure 5: Kernel density plot of estimated cellular fraction reveals clonal and subclonal events.

4 Results

4.1 Sequencing data source.

We applied FACETS to 268 TCGA lung adenocarcinoma whole-exomes. The sequencing bam files were downloaded from the Cancer Genomics Hub `https://cghub.ucsc.edu/`. Each bam file is about 15 GB in size. A pre-processing module that generates sequence count matrix from the sequencing bam file uses samtools/perl/c++ scripts to ensure scaleable and parallelizable implementation. Model fitting, analysis and visualization is done in R statistical programming language which provides a unified front end for analysis and visualization. The ABSOLUTE calls from SNP6.0 array profiling data for the same set of tumor samples published in Zack et al. (2013) (Zack et al., 2013) were obtained from Synapse `https://www.synapse.org/\#!Synapse:syn1703335`. The MSK-IMPACT targeted panel sequencing data are obtained from Paik et al. (2015) (Paik et al., 2015).

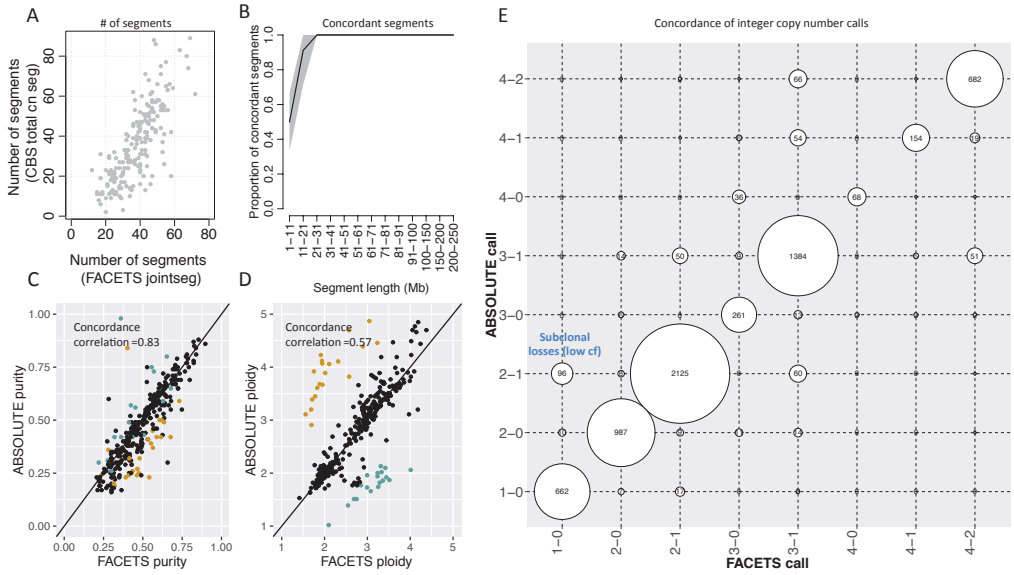

Figure 6: FACETS analysis of whole-exome sequencing of 286 TCGA lung adenocarcinoma samples. A. total number of segments per sample from standard CBS segmentation of total copy number versus FACETS joint segmentation of total and allele-specific copy ratios. B. Proportion of concordantly detected segments between two methods. C. Comparing FACETS and ABSOLUTE tumor purity estimates. D. Comapring FACETS and ABSOLUTE ploidy estimates. E. Bubble plot of FACETS and ABSOLUTE integer copy number calls. The number of concordant (diagonal) and discordant (off diagonal) alterations called are indicated inside each bubble.

4.2 Data pre-processing

The input data for FACETS analysis pipeline are aligned sequence bam file with standard base and mapping quality filter. Reference and variant allele read counts were extracted from the bam file for germline polymorphic sites catalogued in the dbSNP and 1000genome database ($\sim$ 1.9 million polymorphic positions). For whole-exome seq, we include SNPs in target intervals expanded 50-bases on each side (target overhang). Positions with total read count below a lower depth threshold (e.g., < 25 in $50\times$ coverage experiment) or exceed an upper threshold (> 1000) (excessive coverage) in the matched normal were removed.

Analysis of the data from HapMap project has revealed that SNPs are

not distributed at random across the human genome, but are clustered. Regions with increased local variability and SNP clustering has been associated with recombination hotspots. In high-throughput genotyping arrays, such variation has been correlated with elevated rates of genotype failure and allele dropout (Koboldt et al., 2006). In high-throughput sequencing, we show that SNP-dense regions in the genome can cause strong local dependencies in read counts and lead to hyper-segmentation of the genome (Figure 1). To address this issue, we scan all positions by 150-250 bp interval to space out SNP-dense regions and effectively avoid local patterns of strong dependencies. This serial correlation in read counts can cause hyper-segmentation in the downstream steps if not removed.

Read depth ratio between tumor and normal gives information on total copy number. The variant (non-reference) allele frequency at heterozygous loci (germline variant allele frequency greater than 0.25 or less than 0.75) contain information on allelic imbalance. This pre-processing procedure on average yields $\sim 350,000$ SNP loci that pass these quality filters, and $\sim 10-15\%$ of them are heterozygous. Homozygous positions will be kept in our analysis to inform total copy number which increases the precision for genotype calls. The MSK-IMPACT platform target all exons and selected introns of 410 cancer genes (< 1 million bases) with high uniformity of coverage across targets. The pre-processing procedure yields on average $\sim 15,000$ SNP loci with a similar $\sim 10-15\%$ heterozygous rate.

4.3 Application to TCGA whole-exome sequencing data.

Previous TCGA projects have utilized the ABSOLUTE algorithm (Carter et al., 2012) to determine tumor ploidy and purity. This paradigm works by combining segmented copy number output, together with pre-computed models of recurrent cancer karyotypes, and allelic fraction values for somatic point mutations. We compared FACETS output with the ABSOLUTE output reported in the original TCGA studies (Zack et al., 2013).

We first looked at the concordance of the segmentation analysis. Here platform and method differences need to be taken into consideration. First, SNP6.0 array has more even coverage across the genome while whole-exome sequencing may be more sensitive for detecting intragenic changes. The coverage differences have the most effect on the detection of focal changes.

Therefore in this analysis we excluded segments less than 1Mb. Secondly, CBS segmentation which segments total copy number was applied in the Zack et al. study (Zack et al., 2013) for ABSOLUTE input, whereas FACETS implements a joint segmentation of total and allele-specific copy ratios. Bivariate segmentation is more comprehensive and can detect events such as partial chromosomal cn-neutral LOH events that may be missed by a total copy number segmentation approach.

Figure 6A shows the number of segments per tumor sample is relatively comparable between the two methods. Figure 6B further shows the segments are over 90% concordant for segments over 10 Mb in length and less so for smaller segments due to platform and method differences as discussed earlier. In this analysis, we define a segment is concordantly detected by both methods if there is more than 70% overlap between the stat and end positions of two segments.

Figure 6C and 6D show that purity and ploidy estimates are highly concordant between the two methods. FACETS identified additional cases of aneuploidy in about 6% of the tumors (green) by incorporating LOH pattern in determining ploidy. Figure 4B is one of such cases where the total and allelic copy ratio together provide evidence for an aneuploidy tumor that was not identified in the original study based on total copy ratio alone. For a small fraction of tumors that FACETS called lower ploidy than that called by ABSOLUTE (orange), they tend to be lower purity samples.

To compare the integer copy number calls, we focused on samples with concordant ploidy calls (difference in ploidy estimates less than 0.5), tumor purity greater than 30%, and segments length greater than 10 Mb. Figure 6E shows a high concordance of the integer copy number calls.

4.4 Application to targeted cancer gene panel sequencing.

Figure 7 shows a FACETS application to the MSK-IMPACT clinical sequencing platform, a hybridization capture-based next-generation sequencing assay for targeted deep sequencing of all exons and selected introns of 410 key cancer genes in FFPE tumor samples (Cheng et al., 2015). This is a stage IV

lung squamous cell carcinoma (LUSC) patient sample. This patient genome is highly altered. Some key events include homozygous deletion of *CDKN2A*, copy-neutral LOH of chromosomes 9, 11 and 17p. Notably from the FACETS output, high level amplification of known oncogenes including *CCND1* and *PPM1D*, both are druggable targets, are annotated with estimated integer copy number. This tumor also showed aneuploidy with an average ploidy estimated at 3.0.

The FACETS estimate of integer copy number (purity-, ploidy-corrected) for *PPM1D* is 10. By contrast, a conventional *PPM1D* copy number call based on logR ratio (in this case logR=1.3) without adjusting for purity and ploidy would be around 5. This difference is potentially clinically significant as to unambiguously identify amplified cancer genes to guide treatment decisions.

5 DISCUSSION

Comprehensive identification of allele-specific copy number alterations will be invaluable in the search for genomic correlates of clinical outcome and therapeutic targets. In this study, we present FACETS, a unified analysis pipeline and software for joint segmentation and allele-specific copy number analysis with broad applications to NGS platforms. Our method has a number of unique features. We point out that the conventional B-allele-frequency based on sequencing read counts has inherent bias due to mapping affinity toward reference allele. We propose the logOR metric which overcomes such reference bias to provide unbiased estimates of the allelic ratio. The joint segmentation of logR and logOR we developed allows more accurate identification of change points in the genome by directly combing the total and allele-specific read counts. Existing methods use read counts information from heterozygous SNP sites only. We included all SNPs sites. with a weighting scheme that is inversely proportional to the overall heterozygous rate in the patient genome. The combined approach increases the sensitivity and precision for detecting copy number aberrations in the genome especially in low purity samples. Clonal heterogeneity is explicitly considered in our method by introducing a cellular fraction feature associated with segment clusters to allow more accurate inference of ASCNs and facilitate the identification of subclonal events. A normal-non-central χ^2 mixture model is used

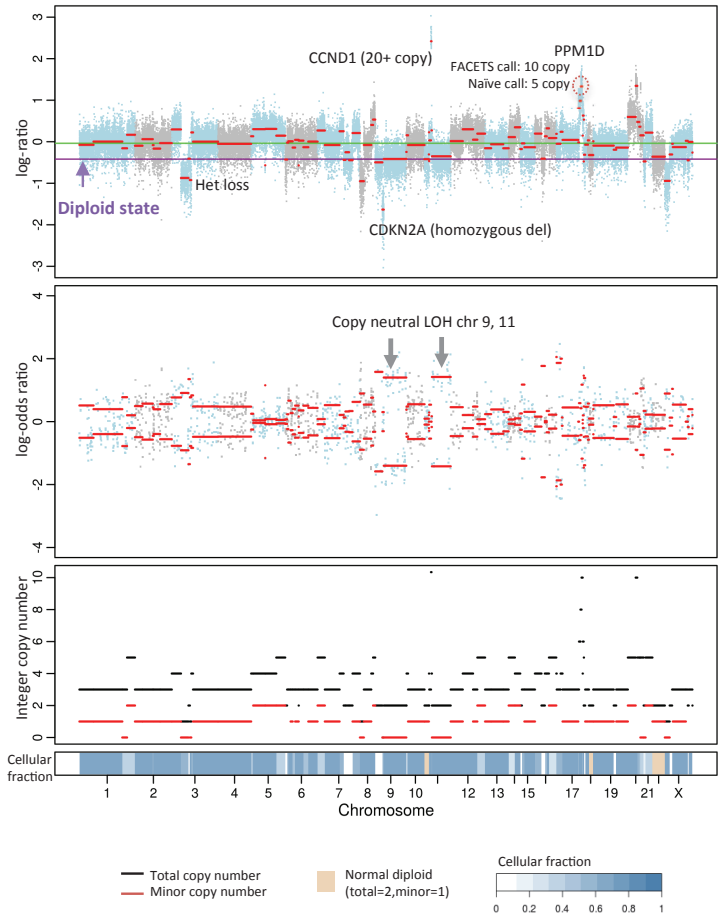

Figure 7: FACETS analysis of a lung squamous cell carcinoma from MSKCC profiled by MSK-IMPACT targeted cancer gene panel sequencing revealed several putative oncogenic drivers and druggable targets. Tumor purity-, ploidy-corrected FACETS analysis provides more accurate integer copy number calls for the driver genes. Integer copy number above 10 are plotted in log10 scale.

to jointly model the total and allelic copy ratio that iterates between imputing the underlying copy number genotype for each segment clusters and updating the model parameters.

FACETS provides a complete ASCN analysis pipeline. This is distinct from most existing methods which often require separate software packages for GC-normalization, sequencing bias adjustment, and/or segmenta-

tion analysis. An integrated analysis pipeline from start to finish will provide more consistent results.

Supplementary Table 1 provides a feature-by-feature comparison between FACETS and other ASCN methods for sequencing data including TITAN and FALCON. Here we highlight several important differences. First, TITAN and FALCON are both based on heterozygous SNP loci which can lead to more rapid loss of sensitivity for detecting copy number alterations when applied to low resolution data (e.g., targeted panel sequencing) and/or low purity tumor samples as we demonstrated in Supplementary Figure 2 using down-sampling approach of whole-exome samples. The output of TITAN and FALCON are presented in Supplementary Figure 1, along with the FACETS output for the chromophobe sample (TCGA-KL-8331) whole-exome.

Average FACETS running time for a whole-exome sample takes $\sim$ 20 minutes for parsing read counts from each pair of tumor-normal BAM files, and 1-3 minutes for subsequent steps including GC-normalzation, joint segmentation and ASCN analysis on a single Intel Xeon E5-2640 core processor. The fast computation facilitates large-scale application. Finally, an application to targeted panel sequencing of clinical samples is also demonstrated. Accurate, purity- and ploidy-corrected, integer copy number calls provided by FACETS will be essential to more reliably interpret NGS-based cancer gene copy number data in the context of clinical sequencing. This may pave the way for the incorporation of NGS-based copy number calls into future updates of these clinical guidelines.

6 ACKNOWLEDGEMENTS

We thank Drs. Nicholas Socci, Barry Taylor, Charlotte Ng for their valuable input. This work is supported in part by funds from the P30-CA008748 Cancer Center Support Grant from the National Cancer Institute to Memorial Sloan Kettering Cancer Center.

6.0.1 Conflict of interest statement.

None declared.

References

Carter, S. L., K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, et al. (2012): "Absolute quantification of somatic dna alterations in human cancer," *Nature biotechnology*, 30, 413–421.

Chen, H., J. M. Bell, N. A. Zavala, H. P. Ji, and N. R. Zhang (2014): "Allele-specific copy number profiling by next-generation dna sequencing," *Nucleic acids research*, gku1252.

Chen, M., M. Gunel, and H. Zhao (2013): "Somatica: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data," *PloS one*, 8, e78143.

Cheng, D. T., T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, et al. (2015): "Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology," *The Journal of Molecular Diagnostics*, 17, 251–264.

Davis, C. F., C. J. Ricketts, M. Wang, L. Yang, A. D. Cherniack, H. Shen, C. Buhay, H. Kang, S. C. Kim, C. C. Fahey, et al. (2014): "The somatic genomic landscape of chromophobe renal cell carcinoma," *Cancer Cell*, 26, 319–330.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard (2009): "Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data," *Bioinformatics*, 25, 3207–3212.

Ha, G., A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, et al. (2014): "Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data," *Genome research*, 24, 1881–1893.

Koboldt, D. C., R. D. Miller, and P.-Y. Kwok (2006): "Distribution of human snps and its effect on high-throughput genotyping," *Human mutation*, 27, 249–254.

Li, Y. and X. Xie (2014): "Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity," *Bioinformatics*, btu174.

Mayrhofer, M., S. DiLorenzo, and A. Isaksson (2013): "Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue," *Genome Biol*, 14, R24.

Oesper, L., A. Mahmoody, and B. J. Raphael (2013): "Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data," *Genome Biol*, 14, R80.

Olshen, A., E. Venkatraman, R. Lucito, and M. Wigler (2004): "Circular binary segmentation for the analysis of array-based dna copy number data." *Biostatistics*, 5, 657–72.

Paik, P. K., R. Shen, H. Won, N. Rekhtman, L. Wang, C. S. Sima, A. Arora, S. Venkatraman, M. Ladanyi, M. F. Berger, and M. G. Kris (2015): "Next generation sequencing of stage iv squamous cell lung cancers reveals an association of pi3k aberrations and evidence of clonal evolution in patients with brain metastases," *Cancer Discovery*, 5, 610–21.

Rasmussen, M., M. Sundstrom, H. Goransson Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, and A. Isaksson (2011): "Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity," *Genome Biol*, 12, R108–R108.

Sathirapongsasuti, J. F., H. Lee, B. A. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson (2011): "Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv," *Bioinformatics*, 27, 2648–2654.

Sun, W., F. A. Wright, Z. Tang, S. H. Nordgard, P. Van Loo, T. Yu, V. N. Kristensen, and C. M. Perou (2009): "Integrated study of copy number states and genotype calls using high-density snp arrays," *Nucleic acids research*, gkp493.

Van Loo, P., S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, et al. (2010): "Allele-specific copy number analysis of tumors," *Proceedings of the National Academy of Sciences*, 107, 16910–16915.

Venkatraman, E. and A. Olshen (2007): "A faster circular binary segmentation algorithm for the analysis of array cgh data." *Bioinformatics*, 23, 657–63.

Xi, R., J. Luquette, A. Hadjipanayis, T.-M. Kim, and P. J. Park (2010): "BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data," *Genome biology*, 11, O10.

Yau, C., D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber, C. C. Holmes, et al. (2010): "A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data," *Genome Biol*, 11, R92.

Zack, T. I., S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, et al. (2013): "Pan-cancer patterns of somatic copy number alteration," *Nature genetics*, 45, 1134–1140.