

Copy number variant detection

August 19, 2015

Sample to Insight -





Abstract

In this white paper, we present the Copy Number Variant (CNV) detection tool, available in the Biomedical Genomics Workbench from version 2.1 and above. We outline the algorithm implemented in the tool, and present accuracy benchmarks comparing the tool to state-of-the-art methods. Our results show that the CNV detection tool is capable of highly accurate identification of copy number variations in a broad range of next-generation sequencing data, achieving a gene-level detection sensitivity of 100% in nearly all our benchmarks. An added advantage of the CNV detection tool is that it is an in-built functionality of the Biomedical Genomics Workbench, so the results can be visualized and interpreted together with the read mappings and the results of other tools. We conclude that the CNV detection tool is a versatile and accurate method for the identification of copy number variants in NGS data, and it has the potential to provide valuable insights in basic medical research, as well as translational clinical settings.

Contents

1	Intro	Introduction								
	1.1	Aims for the CNV detection tool	4							
	1.2	Requirements for the CNV detection tool	4							
2	A br	ief overview of the statistical model	4							
3	Deta	ailed Methods	6							
	3.1	Computing base-level coverages and generation of a robust baseline	6							
	3.2	Normalization of coverages and chromosome-level analysis	6							
	3.3	Log-ratios and their adjustment	7							
	3.4	Chromosome segmentation	8							
	3.5	Estimation of σ	8							
	3.6	Statistical testing	9							
4	Acc	uracy benchmarks	9							
	4.1	Data and parameters used in benchmarking	9							
	4.2	Benchmarking approach	11							
	4.3	Results and Discussion	12							
		4.3.1 Benchmark 1	12							
		4.3.2 Benchmark 2	12							
		4.3.3 Benchmark 3	13							
5	Con	clusions	14							



6	Ack	nowledgements	14
4	Deta	ailed sample-level results	16
	A.1	Benchmark 1: Deep sequencing of inherited disease samples after PCR-based target-enrichment	16
	A.2	Benchmark 2: Whole-exome sequencing of melanoma	18
	A.3	Benchmark 3: Clinical cancer genomic profiling	19



1 Introduction

Copy number variations (CNVs) are a common class of structural alterations in the genome, where sections of the genome are deleted or duplicated compared to a reference set. CNVs can range from about 50 bases to several megabases in size, and CNVs greater than 1 kilobase account for most bases that vary among human genomes [1]. Some of this variation is benign, but CNVs have also been implicated in many diseases, including cancer, inherited disorders, cardiovascular disease, autism and schizophrenia [2–4].

Traditionally, CNVs have been identified using array technologies, such as comparative genomic hybridization (CGH) [5]. With the rapidly increasing adoption of next-generation sequencing (NGS) technologies for the detection of genomic variants, there is an emerging need to identify CNVs from NGS-based data. Several approaches have focused on whole-genome sequencing (WGS) data (see [6] for a review), but targeted resequencing (TR) remains the most cost-effective strategy to identify disease-causing variants.

The identification of CNVs from TR data carries a unique set of challenges. The targeted regions are small, typically only 100-300 bp. The genomic coverage is sparse and non-contiguous, making it highly unlikely that the breakpoints are covered by reads. Approaches developed for WGS data are therefore unsuitable for detecting CNVs from TR data, and utilizing comparative depth-of-coverage remains the most widely used strategy in currently existing CNV detection tools.

Depth-of-coverage methods are based on counting the number of reads that cover each targeted region. Due to the variability of target coverages, affected by issues such as GC content bias or the mappability of reads, most state-of-the-art methods require a set of control samples. After normalization to correct for varying library sizes, statistical inferences are made on the ratio of the coverages in the case sample and the control samples.

It is difficult to evaluate the performance of CNV tools, particularly because a true "gold standard" reference list does not exist for CNVs [8]. The high number of predicted CNVs and the high rate of false positive calls by state-of-theart methods also makes the biological experimental validation of CNVs unrealistic [7]. In this benchmarking effort, we used a range of biological datasets with known copy number variations to evaluate the sensitivity and specificity of the CNV detection tool in the Biomedical Genomics Workbench.

1.1 Aims for the CNV detection tool

The purpose of the CNV detection tool in the Biomedical Genomics Workbench is to identify copy number variants in the following types of targeted resequencing data:

- 1. Whole-exome sequencing
- 2. NGS data from hybridization-capture and amplicon-based gene panels

1.2 Requirements for the CNV detection tool

The CNV detection tool has the following requirements:

- 1. Mapped NGS reads both for the case sample and one or more control samples
- 2. The list of non-overlapping targeted regions

The statistical models are most accurate when a large number of targets are available.

2 A brief overview of the statistical model

The CNV detection tool is based on an analysis of the coverage depth in the case sample in comparison to a baseline generated from the control samples. The generation of the baseline is described in Section 3.1. The base-level coverages of the case and baseline are first normalized to account for varying library sizes (see



Section 3.2). Every target is then characterized by the average log-coverage in the case sample and in the baseline. The log-ratio of these coverages is the *non-adjusted* target-level log-ratio, which is subsequently corrected for coverage bias (see Section 3.3). The resulting quantity is the *adjusted* target-level log-ratio, which we simply refer to as the log-ratio, X_i , for target i.

The log-ratios vary from target to target, caused both by statistical noise and variations due to "true" CNVs. In a simple approach, we can model the log-ratio for target i with a normal distribution:

$$X_i \sim N(0, \sigma_c^2) \tag{1}$$

where the variance, σ_c^2 (a measure for the statistical noise) is a function of coverage c. To estimate σ_c , a 'binning' approach can be used, as in the CONTRA algorithm [9]. Briefly, targets are binned by their log-coverage, and the standard deviation is calculated for each bin. A curve is fitted, which is used to produce a value for the expected variation, $\hat{\sigma}_c$, given any log-coverage value (more details are provided in Section 3.5.) A p-value is then computed for each target using $\hat{\sigma}_c$ for the target's log-coverage. CNVs are identified as the targets whose p-values indicate that they are statistical outliers compared to the rest of the dataset.

The above procedure works well if many regions are targeted, and only few of them are expected to be affected by CNVs (as in the case of many inherited diseases, for example). But if many targets are affected by CNVs (such as in Figure 1(a) and (b)), then true copy number variations will be incorrectly attributed to statistical noise. This leads to a decreased sensitivity to large CNVs, as has also been noted by Tan et al. for the CONTRA algorithm [7]. It is therefore necessary to extend the model to account for potential large-scale changes in log-ratios caused by changes in copy number. In the CNV detection tool, we do this by allowing for a non-zero mean in constant copy-number regions during the estimation of the statistical noise. Combining Equation 1 with the sparse normal mean model described in [10], we model the log-ratios X_i with a normal distribution, where

the mean μ_R only depends on the copy number of the region, and the variance σ_c^2 is a function of target coverage c:

$$X_i \sim N(\mu_R, \sigma_c^2)$$
 (2)

Here, R is a constant copy-number region with $i \in R$. The main goal of the algorithm in the CNV detection tool is to estimate μ_R and σ_c accurately, and thus separate the large-scale copy-number changes from the smaller-scale statistical noise in each copy-number-constant region.

The estimation of these parameters and the statistical testing using the combined model is carried out in a stepwise fashion.

- 1. We segment each chromosome into constant copy-number regions, using the Screening and Ranking Algorithm (SaRA) described in [10]. Here, we use a constant estimator $\hat{\sigma}$ for the intra-region variability, as we assume that the intra-region variability is much smaller than the interregion variability. Thus, the coverage-dependence of the intra-region variance is ignored for the purposes of this segmentation step only (Section 3.4).
- 2. For each region, we compute a medianbased robust estimate $\hat{\mu}_R$ for μ_R . Subtracting this estimate from each X_i , we expect $X_i - \hat{\mu}_R \sim N(0, \sigma_c^2)$, where $i \in R$. This is identical in form to the simple model described in Equation 1, and the binning approach of CONTRA can therefore be used to compute a copy numberadjusted coverage-dependent model for σ_c (Section 3.5).
- 3. For each target, we obtain (one-sided) p-values for amplification and deletion against a null hypothesis of no change, using the copy-number adjusted estimator $\hat{\sigma}_c$. Finally, we use Fisher's method for combining the target-level p-values to obtain p-values for each region, against a hypothesis of no change (Section 3.6).



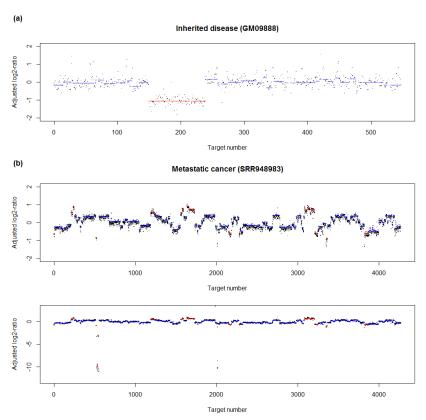


Figure 1: The adjusted log2-ratios in two different types of samples. The black dots represent the adjusted log-ratios calculated at each targeted region. The blue and red plateaus are the regions produced by the CNV detection tool after segmentation. The red regions mark the targets that were predicted to be part of a CNV at a significance level of 0.05 and fold-change cutoff of 1.4. The blue regions were not called by the CNV detection tool. The datasets are described in more detail in Section 4.1. (a) Sample M62-4 from Benchmark 1, inherited large-scale genomic alterations. A large part of chromosome 8 (chr8:107120037-119294603) is affected by a heterozygous deletion. The sample (NA09888, Corriel Institute) was probed with QIAGEN amplicon-based CNV panel CNA902Y. (b) Sample from Benchmark 3, metastatic cancer. The top and bottom plots are identical, except the scaling on the y-axis. The sample is affected by a large number of CNVs. The sequencing data are accessible under accession number SRR948983, originally published in [13]. The target coordinates were obtained from the authors of [13] (personal communication).

3 Detailed Methods

In this section, we describe the steps of the algorithm implemented in the CNV detection tool in more details.

3.1 Computing base-level coverages and generation of a robust baseline

For each case and control sample, base-level coverages are computed by counting the number of reads aligned to every targeted base. Then, a robust baseline for base-level coverages is generated from all control samples by

computing the trimmed mean of the coverages at each base in the control samples, as described in [9]. In this baseline, each targeted nucleotide is associated with a coverage value, which quantifies the overall depth-of-coverage of the nucleotide in the control samples.

3.2 Normalization of coverages and chromosome-level analysis

Both the library size of the case sample, $L_{\rm case}$, and library size of the baseline, $L_{\rm baseline}$, are defined as the sum of the coverages at all positions. Clearly, the library sizes depend on



the sequencing depths of the different samples, and normalization must be done before computing the target-level log-ratios. A simple approach is to scale both the case and the baseline library sizes to their geometric mean, as is done in [9]. Thus, a single scaling value, s, can be used to normalize every raw coverage in the case-baseline pair:

$$s = \frac{\sqrt{L_{\text{case}}L_{\text{baseline}}}}{L_{\text{case}}} \tag{3}$$

The nucleotide-level normalized coverage d at each targeted position in the case and baseline is computed from the un-normalized coverage c:

$$d_{\mathsf{case}} = c_{\mathsf{case}} \cdot s + \delta \tag{4}$$

$$d_{\text{baseline}} = \frac{c_{\text{baseline}}}{s} + \delta \tag{5}$$

 δ is a small offset (equal to 0.5 in the CNV detection tool) used to prevent zero coverages in either the case or the baseline.

However, when large parts of the genome are affected by CNVs, as is the case in Figure 1(a) and (b), then the library size of the case sample will be significantly affected by these CNVs, and the above approach will incorrectly attribute this effect to a difference in sequencing depth. Therefore, in the CNV detection tool, we first detect chromosomal coverage outliers (chromosomes possibly containing large-scale CNVs), and compute $L_{\rm case}$ and $L_{\rm baseline}$ only using the chromosomes that are found to have "typical" coverages.

We detect chromosomal coverage outliers using linear regression analysis, in the following steps.

1. A linear model is used to model the chromosomal coverages:

$$C_{\mathsf{case}} = mC_{\mathsf{baseline}} + \epsilon$$
 (6)

where:

 C_{case} is a random variable representing the total coverage at targeted positions on a chromosome in the case sample

- C_{baseline} is a random variable representing the total coverage at targeted positions on the same chromosome in the baseline
- m is a constant multiplier, related to the different sequencing depth in the case sample compared to the baseline
- ullet is the error, assumed to be Gaussian
- 2. The parameter m is estimated from the data by linear regression analysis.
- 3. For each chromosome k, the studentized residual s_k is computed using the observed chromosomal coverages $c_{\mathsf{baseline},k}$ and $c_{\mathsf{case},k}$:

$$s_k = \frac{e_k}{\sqrt{\mathsf{MSE} \cdot (1 - h_k)}} \tag{7}$$

where $e_k = mc_{\mathsf{baseline},k} - c_{\mathsf{case},k}$ is the observed residual, h_k is the leverage (equal to the kth diagonal entry in the hat matrix), and the MSE is the mean squared error.

4. The studentized residuals s_k are T-distributed with 1 degree of freedom. We use this distribution to identify outliers at 5% significance. "Normal" chromosomes are defined as the non-significant chromosomes under this test. The remaining chromosomes are outliers.

In the following sections, we use "coverage" to refer to normalized coverages computed using Equations 4 and 5, but only using the chromosomes that are not detected to be outliers.

3.3 Log-ratios and their adjustment

In the CNV detection tool, we define the baseline log-coverage of a target i with length N as:

$$\phi_i = \log_2 \frac{1}{N} \sum_p d_{\mathsf{baseline},p} \tag{8}$$

and the non-adjusted log-ratio for the same target i as:

$$r_i = \log_2 \frac{\frac{1}{N} \sum_p d_{\mathsf{case},p}}{\frac{1}{N} \sum_p d_{\mathsf{baseline},p}} \tag{9}$$



where the sums are over all positions p in the target.

Li et al. [9] observed a linear variation of logratios with log-coverage, depending on the differences in library sizes between case and controls. They also proposed a linear correction model, which has been implemented in a similar way in the CNV detection tool. In the CNV detection tool, a straight line is fitted between the non-adjusted log-ratios and baseline log-coverages, and this fitted line is subtracted from each non-adjusted log-ratio to produce the adjusted log-ratio x_i .

$$x_i = r_i - a \cdot \phi_i - b \tag{10}$$

where a and b are the parameters of the fitted line. In the following sections, by "log-ratios" we refer to the adjusted log-ratios.

3.4 **Chromosome segmentation**

The chromosomes are segmented using the multi-bandwidth SaRA algorithm [10]. A bandwidth is an integer value that corresponds to a window in which the algorithm searches for breakpoints. In the CNV detection tool, the bandwidths are determined by the "graining level" parameter, and do not depend on the chromosome lengths. Here we give a brief description of the SaRA approach.

1. Taking one bandwidth value h, the diagnostic function is computed for every target $i \in [1, n]$ on a chromosome with length n. The diagnostic function is the difference of the averages of the adjusted logratios near the target i, within a window of size h:

$$D(i,h) = \frac{1}{h} \sum_{m=1}^{n} w_m(i) x_m$$
 (11)

where $w_m(i) = \operatorname{sgn}(m + \frac{1}{2} - i)$ for i - h < i $m \leq i + h$ and $w_m(i) = 0$ otherwise.

2. The set of local maximizers corresponding to bandwidth h is identified. The target iis an h-local maximizer if

$$\forall i' \in]i-h, i+h] : D(i,h) > D(i',h)$$
 (12)

- 3. Steps 1 and 2 are repeated with the other bandwidth values. A common set of local maximizers is computed as the union of the sets obtained for the different bandwidths. This set contains the potential breakpoints to define the segments.
- 4. Local maximizers are removed from this set one-by-one, until the Bayesian Information Criterion (BIC) for the remaining local maximizers is minimized. The BIC for a set \tilde{J} of breakpoints is:

$$\mathsf{BIC}(\tilde{J}) = \frac{n}{2} \cdot \log(\sigma^2) + \tilde{J}\log(n) \quad \text{ (13)}$$

where n is the total number of targets on the chromosome, and σ is the weighted average of the variance within all regions produced by the breakpoints. The value of the BIC will, in general, decrease if a breakpoint is removed. In each round of the optimization, we remove the local maximiser whose removal leads to the least decrease in variance. We continue removing local maximizers until the BIC no longer decreases.

5. Once the final set of local maximizers is obtained, the targets in between the local maximizers are joined into regions.

3.5 Estimation of σ

Once the constant copy-number regions are identified, we compute the median log-ratio in each region R and use it as an estimate $\hat{\mu}_R$ for μ_R . The median is used instead of the average, because it is a more robust estimator of the mean in the cases where some observations are incorrect. This can for example happen in amplicon-based data, where a single SNP in a primer region can reduce the coverage of the entire amplicon to zero, despite no true differences in copy number.

Subtracting this estimate from each log-ratio X_i inside the region, we define $Y_i = X_i - \hat{\mu}_R$, which we expect is normally distributed: $Y_i \sim$ $N(0, \sigma_c^2)$, where $i \in R$.

The remaining task is to estimate σ_c as a $\forall i' \in [i-h, i+h]: D(i,h) \geq D(i',h)$ (12) function of coverage. Similarly to CONTRA [9],



we classify the observations into approximately equal-sized bins on the basis of their logcoverage in the baseline. In the CNV detection tool, the number of bins is approximately 10, but the precise number of bins is determined dynamically on the basis of the number of targets in the data. For each bin, we compute the median log-coverage, as well as the standard deviation of the adjusted log-ratios of the targets in the bin. We estimate the standard deviation robustly, based on the interquartile range. The result of the binning step is that we have a set of log-coverage-log-ratio value pairs. An exponentially decreasing function $f(d_{\text{baseline}}) = \alpha \cdot \exp(-\beta \cdot d_{\text{baseline}}), \beta > 0$, is fitted to these points, to produce a continuous model for the expected variation in the log-ratios for any given log-coverage value (see illustration in Figure 2).

3.6 Statistical testing

The null hypothesis under the combined model for each target is

$$H_0: X_i \sim N(\mu, \sigma^2)$$
 (14) where: $\mu = 0, \sigma = f(d_{\mathsf{baseline},i})$

Thus, we compute one-sided p-values to test the alternative hypotheses $H_1: \mu > 0$ (amplification case) and $H_1: \mu < 0$ (deletion case).

Finally, to test the statistical significance of each region, we combine these target-level p-values using Fisher's method:

$$\chi_{2k}^2 \sim -2\sum_{i=1}^k \ln(p_i)$$
 (15)

(16)

where p_i is the p-value for the i^{th} hypothesis test. If k tests are combined, the test statistic χ_{2k} has a chi-squared distribution with 2k degrees of freedom.

4 Accuracy benchmarks

4.1 Data and parameters used in benchmarking

To assess the accuracy of the CNV detection tool, we used three markedly different datasets, as detailed below. The datasets were chosen to cover both inherited diseases and cancer, one-copy as well as multiple-copy number variations, a broad size range for genomic alterations, varying sample purities (from 20% to 100%), different target enrichment technologies (PCR-based as well as whole exome sequencing), and a wide range of sequencing depths (from 30x up to 1500x). All three benchmark datasets have also been published previously, enabling the direct comparison of our method with alternative methods for CNV predictions.

1. Benchmark 1: Multiplex-PCR enrichment and deep sequencing of inherited chromosomal disease. This PCR-enriched custom QIAGEN GeneRead gene panel dataset was originally described in [11], where it was used to benchmark the quandico tool for copy number analysis. The sequencing reads were mapped to the hg19 assembly and used directly in the CNV detection tool, without any additional steps. As in [11], the samples were grouped into three sets: (a) M62 "high coverage dataset", with a read depth of approximately 1500x, (b) M63 "mediumcoverage" dataset, with a read depth of approximately 650x, and (c) M117 "validation" dataset, with a read depth of approximately 1000x. In our case, the algorithm did not require a training step, so all three datasets were used for validation. The M62 and M63 datasets were generated using the same custom-made gene panel (CNA902Y), and M117 was generated using a second custom panel (NGHS-991Y). The following samples were used as controls for the M62 and M63 datasets: NA12878 and NA19219. The following samples were used as controls for the M117 dataset: NA12878 and



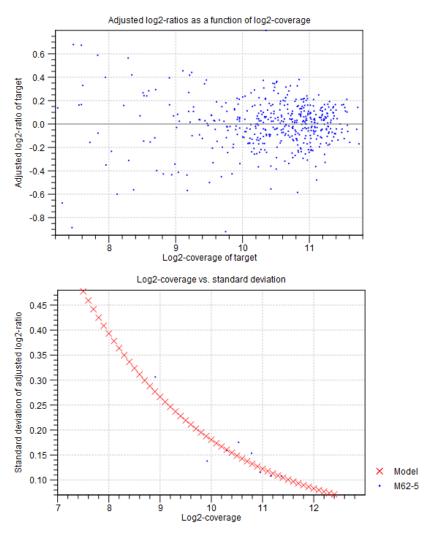


Figure 2: Illustration of the binning procedure to model variation as a function of coverage. The figures were produced in the Biomedical Genomics Workbench using the CNV detection tool. In the top figure, the adjusted log-ratio of each target is plotted against its log-coverage in the baseline. The log-coverages are centered around 0 for all coverages, with the greatest variation at the lowest coverages. In the bottom figure, each blue dot is derived from a "bin" containing a set of targets. For each bin, the mean of the log-ratios is plotted against the mean of the log-coverages in the baseline. A curve is then fitted (red crosses), to give a continuous model for the expected variation in log-ratios for any given log-coverage.

NA19240. As only female controls were available, targets on chromosomes X and Y were ignored in both the prediction and in the benchmarking counts. The CNV detection tool was run with default parameters, except the "Low coverage cutoff" parameter, which was set to 150 for the M62 dataset, 65 for the M63 dataset, and 100 for the M117 dataset, corresponding to approximately a tenth of the average read depth in each case.

2. Benchmark 2: Whole-exome sequencing

of melanoma. This exome-sequencing dataset was originally described in [12], where it was used to benchmark the *Excavator* tool for copy number analysis. In that study, the samples were also independently profiled for CNVs on the Affymetrix 250K SNP Array platform. The sequencing reads were downloaded from the Sequence Read Archive under accession ERP001844, mapped to the hg19 assembly, and used directly in the CNV detection tool, without any additional



steps. The following samples were used as the set of controls for all predictions on this dataset: ERR174237, ERR174238, ERR174239, ERR174240, ERR174241, and ERR174242. As the target file, we used the Agilent SureSelect S0274956 definitions. The CNV detection tool was run with default parameters, except the "Low coverage cutoff" parameter, which was set to 10 due to the low coverage (approximately 30x) of the samples, and the "Minimum fold-change cutoff" parameter, where runs were made both with a value of 1.4 and 1.2.

3. Benchmark 3: Clinical cancer genomic This dataset was originally described in [13], where it was used to benchmark a proprietary pipeline for copy number detection. The sequencing reads were downloaded from the Sequence Read Archive under accession SRP028580. The target file was obtained from the authors of [13] (personal communication). The sequencing reads were mapped to the hg19 assembly and used directly in the CNV detection tool, without any additional steps. The following samples were used as the set of controls for all predictions on this dataset: SRR948995, SRR948996 and SRR948997. As only female controls were available, targets on chromosomes X and Y were ignored in both the prediction and in the benchmarking counts. The CNV detection toolwas run with default parameters, except the "Low coverage cutoff" parameter, which was set to 150 for all the samples, and the "Graining level" parameter, which was set to "Fine" (as CNVs spanning only a few targets were expected).

4.2 Benchmarking approach

We used the CNV detection tool to produce a "region-level" CNV track for each sample. This track contains the CNV regions that passed the p-value and fold-change cutoff filtering criteria.

To evaluate the accuracy of each "region-level" CNV prediction result on the target-level, we used the original target track that was used to produce the prediction, and classified each target into one of the following categories: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). A TP target was defined as one that fulfilled all of the following criteria:

- The target overlaps with an expected CNV region.
- The target overlaps with a CNV region produced by the CNV detection tool.
- The direction of change (gain or loss) predicted by the CNV detection tool is the same as the expected direction of change.
 The magnitude of the predicted change was not considered in these benchmarks.

A FP target overlaps with a predicted CNV region, but not with an expected CNV region. A FN target overlaps with an expected CNV region, but not with a predicted CNV region. Lastly, a TN target is not included either in the expected or in the predicted CNV regions.

We used the following measures to evaluate the accuracy of our predictions:

- Sensitivity: Sensitivity is defined as the TPR (true positive rate), i.e. the fraction of CNVs that were correctly called: TPR = TP/(TP + FN)
- Specificity: Specificity is defined as the TNR (true negative rate), i.e. the fraction of non-CNVs that were correctly not called: TNR = TN/(TN + FP)
- NPV: The NPV (negative predictive value) is the fraction of the predicted negative (non-CNV) calls that were correct, and is thus another measure of sensitivity: NPV = TN/(TN + FN)
- PPV: The PPV (positive predictive value) is the fraction of the predicted positive (CNV) calls that were correct, and is thus another measure of specificity: PPV = TP/(TP + FP)



To evaluate the gene-level accuracy, similar classifications were used, but the counting was done on the basis of a gene track, where the target regions aimed at the same gene were joined. This reflects the currently most common use case, where the gene-level predictions are more relevant, and the exact breakpoints are of secondary importance. Furthermore, it also eliminates possible artifacts in the results resulting from "edge effects" due to uncertain CNV breakpoints. To date, no absolute goldstandard dataset exists for benchmarking CNV tools, and there is usually some uncertainty associated with the precise CNV breakpoint locations even in the reference datasets. We note also that the CNV detection tool was not designed to pinpoint breakpoints inside target regions.

All benchmarks were automated. Target-level comparisons were done in Benchmarks 1 and 2, and gene-level comparisons were done in Benchmarks 1 and 3. The detailed sample-level results are provided in Appendix A.

4.3 Results and Discussion

4.3.1 Benchmark 1

The samples in Benchmark 1 were enriched using a multiplex PCR strategy, where sequence variations can lead to different PCR enrichment efficiencies, and variants (particularly short insertions or deletions) may lead to a reduced sequencing depth without the presence of CNVs [11]. Our results can be found in Section A.1. On the gene-level, again we observed a sensitivity of 100% (all affected genes were detected), and a specificity of 99%, when measured using the true negative rate. The targetlevel sensitivity was also high at 95%, with a specificity of 100% (rounded to the nearest percent). We conclude, therefore, that accuracy of the CNV detection tool was very high on this dataset, with both sensitivities and specificities close to 100%.

Both the gene-level and the target-level PPVs were high in this benchmark (86% and 96%, respectively), as the segmentation was carried

out on a large scale, due to the "Graining level" parameter being set to "Coarse". The PPV can be increased significantly by filtering the resulting CNV calls by the number of targets included in the CNV region.

Interestingly, the FN and FP calls clustered in a few samples. On closer inspection (data available on request), we found that several of the same samples were also affected by FN and FP calls in the independent method, *quandico* [11], which was benchmarked using the same data. This indicates the presence of experimental errors, or that the reference annotations may have been slightly inaccurate for these regions.

Neither the sensitivity of the predictions nor the specificity was correlated with sequencing depth in these benchmarks. This is expected, because all samples were deeply sequenced, and the most important errors affecting accuracy were most likely to be systematic (rather than random) errors, which cannot be reduced by increasing sequencing depth.

4.3.2 Benchmark 2

Benchmark 2 was a whole-exome sequencing dataset, targeting a very large number of regions (over 170,000 targets) at a low depth of coverage (approximately 30x). Furthermore, the samples were derived from melanoma cell lines affected by large-scale genomic alterations. Due to the whole-exome nature of this dataset, we did not carry out gene-level benchmarks on this data. Our results can be found in Section A.2.

Even though this dataset has been previously used in benchmarking the *Excavator* tool [12], no known "true positive" list of CNVs exists for it. In [12], genomic SNP array profiling was performed on the same samples, producing an independent list of CNVs based on non-NGS technology. In the same work, it was found that the NGS-based analysis was significantly more sensitive than the SNP-array technology, particularly for the detection of small CNVs. Our results for this dataset generally support the same conclusion.



We compared our prediction results both with the predictions of Excavator, and the SNP array profiling results reported in [12]. As Excavator reports only 5 copy number states, whereas our tool reports a continuum of fold-changes compared to the normal, we ran the CNV detection tool both with two fold-change cutoff values: 1.4, to increase specificity, and 1.2, to increase sensitivity. As expected, we found that there was a very good correlation between the predictions of our tool and Excavator. Our tool called 95% of the targets that were called by Excavator at a fold-change cutoff of 1.4, and this increased to 99% when the fold-change cutoff was reduced to 1.2. Furthermore, 98% of the targets that were called using the SNP array technology were also called by the CNV detection tool, confirming the high sensitivities we have observed with the other benchmarking datasets.

The specificity of the calls is more difficult to evaluate, particularly because it is not practically feasible to verify the many calls made by either Excavator or the CNV detection tool. However, we found a great degree of overlap between the predictions of the two tools: with the fold-change cutoff set to 1.4, 92% of the targets called by our tool were also called by Excavator, and 98% of the targets not called by our tool were not called by Excavator, either. When the fold-change cutoff was reduced to 1.2, the specificities compared to the Excavator calls reduced (PPV of 47%). This was expected, because Excavator only reports heterozygous (one-copy) or homozygous (multiplecopy) changes, corresponding to a minimum fold-change of 1.5 in the case of amplifications. and it is thus not sensitive to fold-changes of smaller magnitudes. The CNV detection tool, in principle, is capable of predicting significant fold-changes of any magnitude, which may occur in large numbers in cancer-affected samples due to tumor heterogeneity.

As was also observed for *Excavator* [12], the specificity of the CNV detection tool was not high when evaluated against the SNP array results. With the fold-change cutoff set to 1.4,

only 34% of the targets predicted to be affected by the CNV detection tool were also called in the SNP array approach. In comparison, 35% of the targets predicted by *Excavator* were also called in the SNP array approach (data not shown). With the fold-change cutoff reduced to 1.2, the agreement reduced even further to 17%, indicating that fewer than 1 in 5 targets called by the CNV detection tool were also among the SNP array calls. However, as discussed in [12], the low degree of specificity of the NGS-based methods compared to the SNP array-based method is more likely to be a result of higher sensitivity rather than a sign of lower specificity in the NGS-based methods.

Overall, as *Excavator* and the CNV detection tool are based on very different algorithms, such a degree of agreement between these two tools greatly increases our confidence that the sensitivity and the specificity of both tools are indeed very high for whole-exome sequencing data.

4.3.3 Benchmark 3

This clinical sequencing dataset was generated by Foundation Medicine, Inc., and recently published as part of a larger study describing a clinical pipeline based on NGS [13]. This titration dataset enables the evaluation of the sensitivity of CNV prediction in cancer samples at tumor purities of less than 100%. As only gene-level results were provided in that study, and target-level resolution were not available, we also only carried out gene-level analysis in this case. Our results can be found in Section A.3.

The evaluation of the sensitivity and specificity of predictions on this dataset was challenging due to several factors. Firstly, the CNVs were not verified using a non-NGS technology, and the published study [13] only reported homozygous deletions or amplifications of \geq 6 copies. (This corresponds to an observed fold-change of 3 for a case sample with 100% sample purity, or a fold-change of just 1.4 in a sample with 20% sample purity). Furthermore, the specificity of the predictions was not evaluated at all.

Nevertheless, the CNV detection tool was capa-



ble of identifying every affected gene reported by [13] when the sample purity was 50% or greater, and the sensitivity of detection was only reduced below 96% when the sample purity was lower than 30%. This suggests that the CNV detection tool is highly sensitive in the gene-level detection of CNVs in this clinical sequencing dataset, even in the case of low-purity samples. (In a separate experiment, when the fold-change cutoff was reduced to 1.2, a sensitivity of 100% was obtained by the CNV detection tool even in the lowest-purity sample - data not shown.)

The fraction of called targets also called by the cited study is generally rather low, particularly when evaluated using the PPV metric. However, as in the case of Benchmark 2, we believe this is due to an increased sensitivity in our method, especially at higher sample purities: indeed, at a sample purity of 100%, the CNV detection tool with a fold-change cutoff of 1.4 is sensitive enough to call a heterozygous amplification (three copies), in comparison to the six or more copies that are required in the published study [13].

5 Conclusions

The CNV detection tool in the Biomedical Genomics Workbench is capable of accurate identification of copy number variations in a broad range of next-generation sequencing data. In nearly all of our benchmarks, genes which were affected by copy number variations were detected with a 100% sensitivity, even at reduced sample purities. The target-level sensitivities were similarly over 95% under most conditions. Our benchmarks using samples affected by inherited diseases indicate that the specificity of the CNV detection tool also approaches 100%. The specificity of the tool is more challenging to evaluate on cancer datasets, particularly due to the lack of gold standard datasets. However, an analysis of the agreement between our tool and other methods shows that the specificity of the CNV detection tool is comparable to other state-of-the-art tools, such as Excavator [12].

The CNV detection tool has several advantages over other tools currently available for the prediction of CNVs. Firstly, it reports fold-changes on a continuum, enabling the prediction of CNVs at less than 100% sample purity, and places no fundamental restriction on the predictable effect size. Secondly, it includes an in-built segmentation step, enabling the determination of a minimum size for the detected CNVs. Importantly, as the CNV detection tool is an in-built functionality of the Biomedical Genomics Workbench, the results of the predictions can be easily visualized alongside the read mappings and the results of other tools in the Genome Browser View.

The main limitation of the CNV detection tool is that it requires the presence of unaffected targets for optimal performance, as it's based on a statistical method where an equal sequencing depth is not required in the case and control samples. The normalization procedure models the differences in sequencing depth between the case and control samples, and the CNV detection tool depends on this model to recognize biological CNVs as coverage outliers. This method works most effectively when a large number of unaffected targets is available, so an accurate model can be computed for the sequencing depth. For the same reason, panel design is of high importance, and the tool is not suitable for the analysis of datasets where only a single gene was sequenced.

In conclusion, the CNV detection tool is a versatile and accurate method for the identification of copy number variants in NGS data, and it has the potential to provide valuable insights in basic medical research, as well as translational clinical settings.

6 Acknowledgements

We would like to thank R. Yelensky from (Foundation Medicine, Inc.) for sharing the target file for Benchmark 3 with us, and our QIA-GEN colleague F. Reinecke for the data used in Benchmark 2 and for useful conversations.



References

- [1] Conrad DF et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010, **464**(7289):704-12.
- [2] Ni X et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proceedings of the National Academy of Sciences of the United States of America. 2013, 110(52):21083-8.
- [3] Girirajan S, Campbell CD, and Eichler EE. Human copy number variation and complex genetic disease. Annual Review of Genetics. 2011, 45:203-226.
- [4] Zack TI et al. Pan-cancer patterns of somatic copy number alteration. Nature Genetics. 2013, **45**(10):1134-1140.
- [5] Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nature Genetics. 2007, **39**:S16-S21
- [6] Zhao M et al. Computational tools for copy number variation (CNV) detection using nextgeneration sequencing data: features and perspectives. BMC Bioinformatics. 2013, 14(Suppl 11):S1.
- [7] Tan R et al. An evaluation of copy number variation detection tools from whole-exome sequenc-

- ing data. Human Mutation. 2014, **35**(7):899-907.
- [8] Teo SM, et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012, **28**(21):2711-2718.
- [9] Li et al. CONTRA: copy number analysis for targeted resequencing, Bioinformatics. 2012, **28**(10):1307-1313.
- [10] Niu YS and Zhang H. The screening and ranking algorithm to detect DNA copy number variations., Annals of Applied Statistics. 2012, **6**(3): 1306-1326.
- [11] Reinecke F, Satya RV and DiCarlo J. Quantitative analysis of differences in copy numbers using read depth obtained from PCR-enriched samples and controls., BMC Bioinformatics. 2015, **16**:17.
- [12] Magi A et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data., Genome Biology. 2013, **14**(10):R120.
- [13] Frampton GM et al., Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing., Nature Biotechnology. 2013, **31**:1023-1031.



A Detailed sample-level results

.

A.1 Benchmark 1: Deep sequencing of inherited disease samples after PCR-based target-enrichment

Benchmark 1: gene-level results											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
M62-1 (NA01201)	4	36	0	0	1.00	1.00	1.00	1.00			
M62-3 (NA05067)	2	38	0	0	1.00	1.00	1.00	1.00			
M62-4 (NA09888)	5	33	2	0	1.00	0.94	1.00	0.71			
M62-5 (NA11672)	3	37	0	0	1.00	1.00	1.00	1.00			
M62-6 (NA12606)	3	36	1	0	1.00	0.97	1.00	0.75			
M62-8 (NA13783)	4	36	0	0	1.00	1.00	1.00	1.00			
M62-9 (NA14164)	2	37	1	0	1.00	0.97	1.00	0.67			
M62-10 (NA20022)	2	38	0	0	1.00	1.00	1.00	1.00			
Total, M62	25	291	4	0	1.00	0.99	1.00	0.86			
M63-1 (NA01201)	4	36	0	0	1.00	1.00	1.00	1.00			
M63-3 (NA05067)	2	36	2	0	1.00	0.95	1.00	0.50			
M63-4 (NA09888)	5	35	0	0	1.00	1.00	1.00	1.00			
M63-5 (NA11672)	3	36	1	0	1.00	0.97	1.00	0.75			
M63-6 (NA12606)	3	36	1	0	1.00	0.97	1.00	0.75			
M63-8 (NA13783)	4	35	1	0	1.00	0.97	1.00	0.80			
M63-9 (NA14164)	2	37	1	0	1.00	0.97	1.00	0.67			
M63-10 (NA20022)	2	37	1	0	1.00	0.97	1.00	0.67			
Total, M63	25	288	7	0	1.00	0.98	1.00	0.78			
M117-3 (NA11213)	4	54	2	0	1.00	0.96	1.00	0.67			
M117-4 (NA09367)	9	51	0	0	1.00	1.00	1.00	1.00			
M117-5 (NA14485)	7	53	0	0	1.00	1.00	1.00	1.00			
M117-6 (NA06226)	7	53	0	0	1.00	1.00	1.00	1.00			
M117-7 (NA16595)	2	57	1	0	1.00	0.98	1.00	0.67			
M117-8 (NA09216)	2	58	0	0	1.00	1.00	1.00	1.00			
M117-9 (NA10925)	4	56	0	0	1.00	1.00	1.00	1.00			
M117-10 (NA10985)	1	59	0	0	1.00	1.00	1.00	1.00			
Total, M117	36	441	3	0	1.00	0.99	1.00	0.92			
All samples	85	961	14	0	1.00	0.99	1.00	0.86			



Benchmark 1: target-level results											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
M62-1 (NA01201)	71	567	0	0	1.00	1.00	1.00	1.00			
M62-3 (NA05067)	36	602	0	0	1.00	1.00	1.00	1.00			
M62-4 (NA09888)	88	546	4	0	1.00	0.99	1.00	0.96			
M62-5 (NA11672)	53	584	0	1	0.98	1.00	1.00	1.00			
M62-6 (NA12606)	42	564	8	24	0.64	0.99	0.96	0.84			
M62-8 (NA13783)	71	567	0	0	1.00	1.00	1.00	1.00			
M62-9 (NA14164)	39	587	6	6	0.87	0.99	0.99	0.87			
M62-10 (NA20022)	37	601	0	0	1.00	1.00	1.00	1.00			
Total, M62	437	4618	18	31	0.93	1.00	0.99	0.96			
M63-1 (NA01201)	71	567	0	0	1.00	1.00	1.00	1.00			
M63-3 (NA05067)	23	577	25	13	0.64	0.96	0.98	0.48			
M63-4 (NA09888)	88	550	0	0	1.00	1.00	1.00	1.00			
M63-5 (NA11672)	53	582	2	1	0.98	1.00	1.00	0.96			
M63-6 (NA12606)	63	570	2	3	0.95	1.00	0.99	0.97			
M63-8 (NA13783)	48	562	5	23	0.68	0.99	0.96	0.91			
M63-9 (NA14164)	44	592	1	1	0.98	1.00	1.00	0.98			
M63-10 (NA20022)	34	594	7	3	0.92	0.99	0.99	0.83			
Total, M63	424	4594	42	44	0.91	0.99	0.99	0.91			
M117-3 (NA11213)	96	824	18	0	1.00	0.98	1.00	0.84			
M117-4 (NA09367)	175	762	0	1	0.99	1.00	1.00	1.00			
M117-5 (NA14485)	119	818	0	1	0.99	1.00	1.00	1.00			
M117-6 (NA06226)	101	837	0	0	1.00	1.00	1.00	1.00			
M117-7 (NA16595)	59	877	2	0	1.00	1.00	1.00	0.97			
M117-8 (NA09216)	57	881	0	0	1.00	1.00	1.00	1.00			
M117-9 (NA10925)	35	902	0	1	0.97	1.00	1.00	1.00			
M117-10 (NA10985)	46	890	0	2	0.96	1.00	1.00	1.00			
Total, M117	688	6791	20	5	0.99	1.00	1.00	0.97			
All samples	1549	16003	80	80	0.95	1.00	1.00	0.96			



A.2 Benchmark 2: Whole-exome sequencing of melanoma.

Benchmark 2: target-level results													
Fold-change cutoff: 1.4, overlap with Excavator results													
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV					
Me01 (ERR174231)	9268	154120	7707	503	0.95	0.95	1.00	0.55					
Me02 (ERR174232)	48275	119777	788	2758	0.95	0.99	0.98	0.98					
Me04 (ERR174233)	36752	130974	2484	1388	0.96	0.98	0.99	0.94					
Me05 (ERR174234)	56278	109576	1099	4645	0.92	0.99	0.96	0.98					
Me08 (ERR174235)	49382	118002	932	3282	0.94	0.99	0.97	0.98					
Me12 (ERR174236)	22901	141723	6606	368	0.98	0.96	1.00	0.78					
Total	222856	774172	19616	12944	0.95	0.98	0.98	0.92					
Fold-change cutoff: 1.4, overlap with SNP array results													
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV					
Me01 (ERR174231)	8506	153868	8469	755	0.92	0.95	1.00	0.50					
Me02 (ERR174232)	181	122365	48882	170	0.52	0.71	1.00	0.00					
Me04 (ERR174233)	10729	132352	28507	10	1.00	0.82	1.00	0.27					
Me05 (ERR174234)	28122	113386	29255	835	0.97	0.79	0.99	0.49					
Me08 (ERR174235)	12884	121252	37430	32	1.00	0.76	1.00	0.26					
Me12 (ERR174236)	21190	141853	8317	238	0.99	0.94	1.00	0.72					
Total	81612	785076	160860	2040	0.98	0.83	1.00	0.34					
Fold-change cutoff: 1	L.2, overlap	with Exca	vator result	s									
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV					
Me01 (ERR174231)	9639	117028	44799	132	0.99	0.72	1.00	0.18					
Me02 (ERR174232)	50608	93873	26692	425	0.99	0.78	1.00	0.65					
Me04 (ERR174233)	38056	59300	74158	84	1.00	0.44	1.00	0.34					
Me05 (ERR174234)	59120	103837	6838	1803	0.97	0.94	0.98	0.90					
Me08 (ERR174235)	51856	110736	8198	808	0.98	0.93	0.99	0.86					
Me12 (ERR174236)	23256	45988	102341	13	1.00	0.31	1.00	0.19					
Total	232535	530762	263026	3265	0.99	0.67	0.99	0.47					
Fold-change cutoff: 1	L.2, overlap	with SNP a	array result										
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV					
Me01 (ERR174231)	8878	116777	45560	383	0.96	0.72	1.00	0.16					
Me02 (ERR174232)	319	94266	76981	32	0.91	0.55	1.00	0.00					
Me04 (ERR174233)	10734	59379	101480	5	1.00	0.37	1.00	0.10					
Me05 (ERR174234)	28138	104821	37820	819	0.97	0.73	0.99	0.43					
Me08 (ERR174235)	12896	111524	47158	20	1.00	0.70	1.00	0.21					
Me12 (ERR174236)	21418	45991	104179	10	1.00	0.31	1.00	0.17					
Total	82383	532758	413178	1269	0.98	0.56	1.00	0.17					



A.3 Benchmark 3: Clinical cancer genomic profiling.

Benchmark 3: gene-level results, ≥30% tumor purity											
100% tumor purity											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
SRR948953	9	189	74	0	1.00	0.72	1.00	0.11			
SRR948959	2	187	83	0	1.00	0.69	1.00	0.02			
SRR948965	2	209	61	0	1.00	0.77	1.00	0.03			
SRR948971	1	215	56	0	1.00	0.79	1.00	0.02			
SRR948977	5	225	42	0	1.00	0.84	1.00	0.11			
SRR948983	4	237	31	0	1.00	0.88	1.00	0.11			
SRR948989	5	187	80	0	1.00	0.70	1.00	0.06			
Total, 100% tumor	28	1449	427	0	1.00	0.77	1.00	0.06			
75% tumor purity											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
SRR948954	9	198	65	0	1.00	0.75	1.00	0.12			
SRR948960	2	198	72	0	1.00	0.73	1.00	0.03			
SRR948966	2	245	25	0	1.00	0.91	1.00	0.07			
SRR948972	1	260	11	0	1.00	0.96	1.00	0.08			
SRR948978	5	238	29	0	1.00	0.89	1.00	0.15			
SRR948984	4	213	55	0	1.00	0.79	1.00	0.07			
SRR948990	5	225	42	0	1.00	0.84	1.00	0.11			
Total, 75% tumor	28	1577	299	0	1.00	0.84	1.00	0.09			
50% tumor purity											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
SRR948955	9	209	54	0	1.00	0.79	1.00	0.14			
SRR948961	2	247	23	0	1.00	0.91	1.00	0.08			
SRR948967	2	255	15	0	1.00	0.94	1.00	0.12			
SRR948973	1	270	1	0	1.00	1.00	1.00	0.50			
SRR948979	5	250	17	0	1.00	0.94	1.00	0.23			
SRR948985	4	247	21	0	1.00	0.92	1.00	0.16			
SRR948991	5	254	13	0	1.00	0.95	1.00	0.28			
Total, 50% tumor	28	1732	144	0	1.00	0.92	1.00	0.16			
40% tumor purity											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
SRR948956	8	248	15	1	0.89	0.94	1.00	0.35			
SRR948962	2	233	37	0	1.00	0.86	1.00	0.05			
SRR948968	2	261	9	0	1.00	0.97	1.00	0.18			
SRR948974	1	271	0	0	1.00	1.00	1.00	1.00			
SRR948980	5	256	11	0	1.00	0.96	1.00	0.31			
SRR948986	4	249	19	0	1.00	0.93	1.00	0.17			
SRR948992	5	257	10	0	1.00	0.96	1.00	0.33			
Total, 40% tumor	27	1775	101	1	0.96	0.95	1.00	0.21			
30% tumor purity											
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV			
SRR948957	8	256	7	1	0.89	0.97	1.00	0.53			
SRR948963	2	257	13	0	1.00	0.95	1.00	0.13			
SRR948969	2	262	8	0	1.00	0.97	1.00	0.20			
SRR948975	1	271	0	0	1.00	1.00	1.00	1.00			
SRR948981	5	264	3	0	1.00	0.99	1.00	0.62			
SRR948987	4	227	41	0	1.00	0.85	1.00	0.09			
SRR948993	5	264	3	0	1.00	0.99	1.00	0.62			
Total, 30% tumor	27	1801	75	1	0.96	0.96	1.00	0.26			
Total, \geq 30% tumor	138	8334	1046	2	0.99	0.89	1.00	0.12			



Benchmark 3: gene-level results, <30% tumor purity												
20% tumor purity												
Sample	TP	TN	FP	FN	Sensitivity	Specificity	NPV	PPV				
SRR948958	8	256	7	1	0.89	0.97	1.00	0.53				
SRR948964	1	263	7	1	0.50	0.97	1.00	0.12				
SRR948970	2	269	1	0	1.00	1.00	1.00	0.67				
SRR948976	1	270	1	0	1.00	1.00	1.00	0.50				
SRR948982	2	266	1	3	0.40	1.00	0.99	0.67				
SRR948988	4	254	14	0	1.00	0.95	1.00	0.22				
SRR948994	0	267	0	5	0.00	1.00	0.98	NA				
Total, 20% tumor	18	1845	31	10	0.64	0.98	0.99	0.37				
Total, <30% tumor	18	1845	31	10	0.64	0.98	0.99	0.37				