BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 11 2010, pages 1395–1402
doi:10.1093/bioinformatics/btq145

Genome analysis Advance Access publication April 18, 2010

# CNAnova: a new approach for finding recurrent copy number abnormalities in cancer SNP microarray data

Sergii Ivakhno* and Simon Tavaré

Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Associate Editor: Alex Bateman

**ABSTRACT**

**Motivation:** The current generation of single nucleotide polymorphism (SNP) arrays allows measurement of copy number aberrations (CNAs) in cancer at more than one million locations in the genome in hundreds of tumour samples. Most research has focused on single-sample CNA discovery, the so-called segmentation problem. The availability of high-density, large sample-size SNP array datasets makes the identification of recurrent copy number changes in cancer, an important issue that can be addressed using the cross-sample information.

**Results:** We present a novel approach for finding regions of recurrent copy number aberrations, called CNAnova, from Affymetrix SNP 6.0 array data. The method derives its statistical properties from a control dataset composed of normal samples and, in contrast to previous methods, does not require segmentation and permutation steps. For rigorous testing of the algorithm and comparison to existing methods, we developed a simulation scheme that uses the noise distribution present in Affymetrix arrays. Application of the method to 128 acute lymphoblastic leukaemia samples shows that CNAnova achieves lower error rate than a popular alternative approach. We also describe an extension of the CNAnova framework to identify recurrent CNA regions with intra-tumour heterogeneity, present in either primary or relapsed samples from the same patients.

**Availability:** The CNAnova package and synthetic datasets are available at http://www.compbio.group.cam.ac.uk/software.html

**Contact:** sergii.ivakhno@cancer.org.uk

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

Received on December 9, 2009; revised on February 24, 2010; accepted on April 1, 2010

## 1 INTRODUCTION

Different genetic and epigenetic alterations can lead to the development of cancer by activating oncogenes or inactivating tumour suppressor genes. Copy number changes are one example of such alterations, where amplifications or deletions of genes implicated in cancer progression can cause abnormal cell growth and proliferation (Chin and Gray, 2008). In the past decade, single nucleotide polymorphism (SNP) arrays have become a *de facto* standard for detecting copy number alterations (CNAs) in cancer genomes. The latest versions of SNP arrays, manufactured by Affymetrix and Illumina, have more than a million polymorphic and non-polymorphic (NP) probes. For instance, the Affymetrix SNP 6.0 array has >1.8 million probes with roughly equal proportions of SNP and NP probes.

The high density of SNP arrays and the availability of NP probes has made SNP arrays the technology of choice for identifying CNAs in some recently published large-scale oncogenomic studies (Chin *et al*., 2008; Weir *et al*., 2007). In addition to high-density arrays these studies include a large number of samples. Both the number of probes and samples create novel data analysis challenges. The most common problem is the transformation of normalized log-ratio values into accurate copy number calls at the highest possible resolution. This so-called single sample segmentation problem has received much attention, with numerous methods developed for this task (Colella *et al*., 2009; Greenman *et al*., 2009; Nilsson *et al*., 2009; Wang *et al*., 2007). A discussion of segmentation algorithms for SNP array platforms is available in the Supplementary Material.

Cancer cells usually harbour two types of chromosomal abnormalities: large scale alterations such as gains and losses of whole chromosome arms, and more focal amplifications and deletions. Given the high rate of genomic instability found in cancer cells, large-scale copy number changes usually represent passenger mutations and due to their large size do not facilitate discovery of functional driver events that lead to malignancy (Pinkel and Albertson, 2005). On the other hand, due to their smaller size and recurrence, detection of focal CNAs could lead to the identification of new genes implicated in cancer progression. This can be facilitated by the availability of high-density/large sample-size SNP array datasets, where cross-sample frequency information can be used to identify driver CNAs and distinguish them from random mutations and probe intensity artefacts.

Several methods for finding regions of recurrent CNAs using aCGH and SNP microarray data have been described in the literature. A review and qualitative comparison of different methods can be found in Shah (2008). A common theme in these methods is that they require a preliminary segmentation step to find regions of interest for each sample. For example, significance testing for aberrant copy number (STAC) starts by creating a binary matrix from the normalized log-ratios, with zeros designating no change and ones designating losses and gains (Diskin *et al*., 2006). It then utilizes two complementary statistics, footprint and frequency, to define recurrent CNA regions based on their length and the number of samples they occur in. A potential problem with this approach is the

*To whom correspondence should be addressed.

 

difficulty in selecting the cut-off for defining CNA-spanning probes from segmented data.

Genomic Identification of Significant Targets in Cancer (GISTIC) (Beroukhim *et al*., 2007) is another approach that uses segmentation information. In contrast to STAC, log-ratios in GISTIC are not discretized; rather those log-ratios below a threshold are set to zero. This allows GISTIC to better discriminate between CNA regions of different copy number. However, it still suffers from ambiguities that arise from specifying a threshold for log-ratios. To detect significant CNAs, both GISTIC and STAC use a permutation approach. In this case, the significance cut-off is strongly dependent on the number of copy number changes present in the dataset and, depending on the extent of genomic instability, can increase the number of false positive (FP) or negative hits. The noise level also influences the computation of thresholds.

Here, we present a novel approach, CNAnova, for finding recurrent CNAs, those that appear in multiple samples. The boundaries of such a CNA are determined by the magnitude of probe log-ratios across all samples that contain the CNA. CNAnova uses properties of a control dataset composed of the normal samples to assess statistical significance of identified CNAs and, in contrast to previous methods, does not require the data segmentation and permutation steps. By using the distribution of probe intensities in normal samples, CNAnova can better assess background probe variation present in the dataset. We also describe an extension of the CNAnova framework for identifying recurrent CNAs with intra-sample heterogeneity. Properties of the method are extensively tested and compared using both simulated and real data.

## 2 APPROACH

The CNAnova procedure is composed of several distinct steps designed to preprocess the data (such as removing probes spanning germline CNV regions), transform them into a suitable format for statistical analysis, estimate $F$- and $t$-statistics from the ANOVA model, identify boundaries of the recurrent CNA regions using the gradient kernel density estimation and find significant regions through control of the false discovery rate (FDR). The schematic representation of the method is outlined in Figure 1.

In the following subsections, the implementation of the CNAnova model is described and characteristic features of the method are discussed. The strategy for using reference normal samples to distinguish between somatic copy number changes in cancer, copy number variation in normal individuals and non-biological probe effects such as wave artefacts is described.

## 3 METHODS

### 3.1 Pseudo-replication and creation of reference dataset

CNAnova is based on local decomposition of the variance and takes advantage of the dataset-specific structure generated during the preprocessing step of the method. First, we discuss the general data representation framework and then give examples of its extension. We assume that the dataset is divided into a reference set $\mathcal{N}_r$ of samples representing normal individuals, which serve as a reference for the single-channel array data, and a set $\mathcal{N}_c$ of cancer samples. The normal samples can include matched controls, where tumour and normal tissues are collected from the same individual, or they can be derived from a pooled sample or a subset of individuals. In fact, the algorithm does not impose specific requirements on having a matched reference.

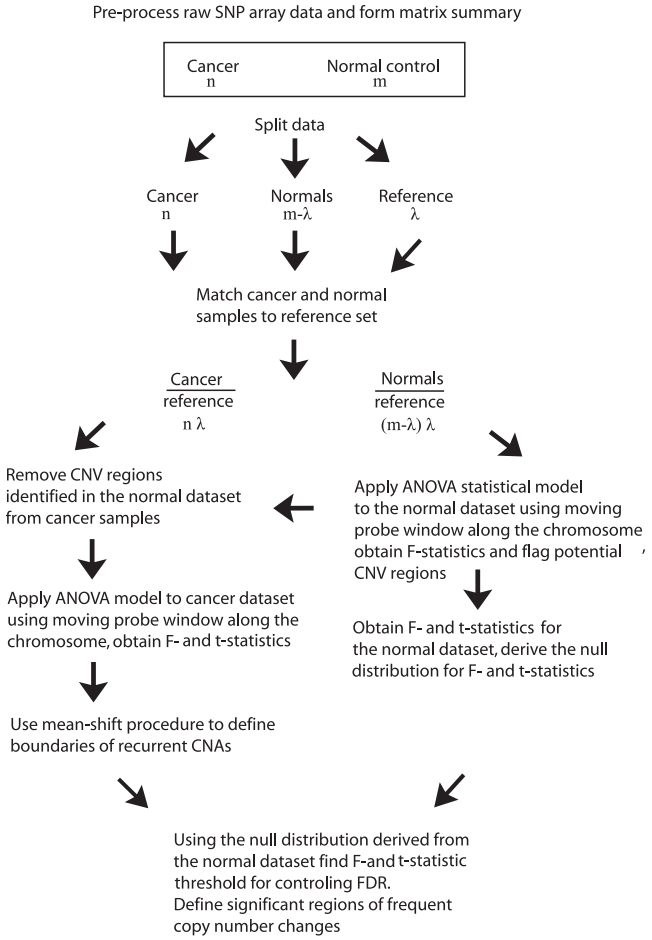


**Fig. 1.** Diagram showing the workflow of the CNAnova procedure. The size of each dataset is shown below its name. Further details are provided in the text.

The derivation of log-ratio values from single channel normal and cancer array samples involves pseudo-replication, performed as follows. Each cancer sample in $\mathcal{N}_c$ is associated with *the same* subset $\mathcal{N}_r' \subset \mathcal{N}_r$, of size $\lambda$, of normal samples. The subset $\mathcal{N}_r'$ is chosen from samples with the most robust quality metric, as discussed in the Supplementary Material. The values in $\mathcal{N}_r'$ are used to calculate log-ratio values, and this make each cancer sample pseudo-replicated $\lambda$ times. These samples will have the same distribution of CNAs as the original cancer sample, but will differ in the background probe variation arising from differences between original normal samples.

A control dataset is then created following a similar procedure that uses the samples in $\mathcal{N}_r'$. The normal samples not in $\mathcal{N}_r'$ are associated with $\mathcal{N}_r'$, thereby insuring that each normal and tumour sample is matched to the same reference samples. These two procedures create pseudo-replicated sets of cancer and normal samples.

Finally, after replication we choose a single normal sample from $\mathcal{N}_r/\mathcal{N}_r'$ to make an ANOVA reference sample of size $\lambda$ using $\mathcal{N}_r'$ that is used to produce treatment contrasts in the analysis of variance below.

### 3.2 CNAnova statistical model

The CNAnova model can be thought of as a one-way analysis of variance. The algorithm first finds the regions of recurrent CNAs in the whole dataset

and then detects samples with copy number changes. Specifically, ANOVA is used to compare the means of probe log-ratios between the pseudo-replicated cancer samples and the reference normal sample, using a sliding widow along each chromosome. The frequency of copy number changes for specific chromosomal regions across the dataset is captured by the magnitude of the $F$-statistics. The algorithm can be decomposed into the following steps:

(1) Let $\rho$ be the shift parameter that defines successive locations at which a window is placed. For a chromosome containing $t$ probes, we apply ANOVA to each of the $t/\rho$ windows.

(2) For each window $j$, with $l$ probes, we fit the model

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}, \quad i = 0, 1, 2, \ldots, n; \quad k = 1, \ldots, l\lambda. \qquad (1)$$

where

- $n$ is the number of cancer samples;
- $i = 0$ corresponds to the ANOVA reference sample and $i = 1, \ldots, n$ to cancer samples;
- $Y_{ik}$ is the probe log-ratio of the $k$-th of the set of $l\lambda$ probes in the cancer samples;
- $\mu$ is the mean of the probes for the normal reference sample in window $j$, and $\tau_0 = 0$;
- For $i \geq 1, \tau_i$ is the fixed effect for a particular cancer sample $i$;
- $\varepsilon_{ik}$ are the random error terms such that $\varepsilon_{ik}$ are independent random variables with $E(\varepsilon_{ik}) = 0$ and $V(\varepsilon_{ik}) = \sigma^2$.

(3) Derive overall $F$- and $t$-statistics for each regression coefficient in the formula.

(4) Apply the same approach to the normal control samples for assessment of statistical significance (Section 3.4).

It is important to ensure robustness of the ANOVA model against outliers and local shifts in the probe distribution, such as those arising from the wave artefact that is prominent in SNP array data (Diskin *et al*., 2008; Marioni *et al*., 2007). One possible solution is to use robust ANOVA methods. However, these methods may remove a significant proportion of probes, which may decrease the number of identifiable CNAs and they add computational overhead. We adopt an alternative two-stage solution to the problem. First, we preprocess the data by removing outliers using modification of the $k$-nearest neighbour smoothing approach described in Olshen *et al*. (2004). Next, we centre the log-ratio distribution on each chromosome using the mean of all log-ratios in the interquartile range. This helps to ensure that in the case of no significant copy number changes, the log-ratios are centred around a zero baseline. In addition to robustness against outliers it is necessary to verify that autocorrelation and heteroscedasticity of the variance do not increase the error rate of the method. In the Supplementary Material, we show that such violation in the distribution of variances is offset by concomitant changes in means, thereby only marginally influencing the performance of CNAnova.

A large proportion of probes tested will be altered in at least one sample, usually as a result of large-scale, single-copy gains or losses arising from genomic instability. This can lead to large and significant $F$-statistics for all the segments spanning the changed regions (even though in most cases they do not reside within regions of recurrent CNAs). To detect truly frequent focal CNA regions, we flag low-level copy number gains and losses in the dataset using smoothing spline normalization. The smoothing parameter is empirically selected to ensure that the transformation preserves focal CNAs; this choice depends on the largest detectable recurrent segment and the number of probes on the chromosome. Although broad regions do not guide the detection of focal recurrent CNAs, they are incorporated into recurrent CNAs when they span the regions of recurrence (Supplementary Fig. S1 and Supplementary section 'Relevance of spline-correction for identification of recurrent CNAs').

Having both spline-corrected and unsmoothed datasets, the estimation of $F$-statistics and $P$-values using CNAnova is carried out as follows. The model in (1) is first applied to the spline-corrected dataset to derive an $F$-statistic

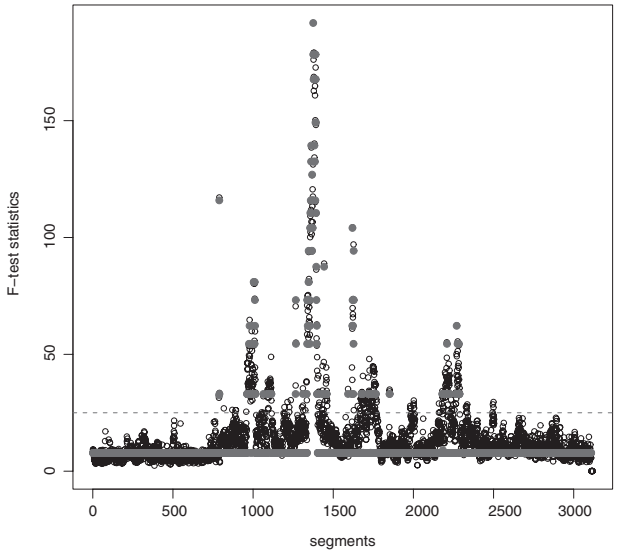


**Fig. 2.** Application of mean-shift procedure to the distribution of $F$-statistics along the chromosome. Solid grey lines indicate the location of mean-shift minima and dotted line shows the threshold for calling significant $F$-test scores. The mean-shift algorithm allows smoothing of the non-significant spikes and detection of significant changes in the $F$-statistics.

and to find regions that harbour recurrent CNAs. This is done with the aid of the mean-shift procedure described in the next section, which delimits CNA boundaries across all samples. The method then uses coordinates of CNA segments with the original unsmoothed dataset and re-applies the model. This time $P$-values for the regression coefficients $\tau_i$ are used to identify *samples* that have copy number changes. Consequently, the two successive applications of the ANOVA model provide a set of $F$-statistics of size $t/\rho$ and a $P$-value matrix with dimensions $t/\rho$ and $n$, which help to identify both the boundaries and sample content of recurrent CNAs.

### 3.3 Finding boundaries of recurrent CNA segments

The distribution of $F$-statistics alone is difficult to use to determine the precise boundaries of CNA regions, due to local discontinuities which might either split a single region into several or lead to FPs due to local outliers among the $F$-statistics. $F$-statistics, therefore, need to be smoothed to remove such aberrant values. The problem with local smoothing methods such as loess is that local least square estimation of means at each point leads to smooth transition between $F$-statistics corresponding to CNA and non-CNA regions, which reduces the precision of correct CNA boundary placement. An alternative local smoothing approach based on modes can provide discontinuous demarcation of boundaries for CNA and non-CNA $F$-statistic regions. This is in part achieved by the mean-shift procedure (cf. Comaniciu *et al*., 2001; Wang *et al*., 2009), which performs discontinuity-preserving smoothing of the $F$-statistic, thereby removing the noise in homogeneous regions of the chromosome and preserving discontinuities at the same time. We extend the mean-shift procedure to find boundaries of recurrent CNA regions by assigning probes with the same mode of the $F$-statistic to separate regions on the chromosome (Fig. 2).

Finding boundaries of recurrent CNAs across samples using the mean-shift procedure requires selection of the appropriate kernel $K$ and bandwidth parameter $h$ that controls the degree of smoothness. We take $K$ to be an univariate Gaussian kernel and use SDs of log-ratios to determine $h$ for each chromosome, as follows:

(1) Using non-overlapping windows of the size 10 000 probes, estimate the SD of log-ratios, and let $\mathcal{S}$ be the set of all such SDs;

(2) Generate $b$ uniformly distributed parameters $\sigma_i, i=1,\ldots,b$ taking values between the minimum value in $\mathcal{S}$ and the first quartile of $\mathcal{S}$;

(3) Apply the mean-shift procedure $b$ times using the Gaussian kernel with $h=\sigma_i, i=1,\ldots,b$;

(4) Estimate the maximal log-likelihood for each model using Gaussian mixtures approximation and their corresponding Bayesian Information Criterion (BIC) values;

(5) Select the model with $\sigma$ that gives the largest BIC value. Use the mean-shift modes identified with this model with bandwidth $h=\sigma$ to find distinct peaks in the distribution of the $F$-statistics. Those exceeding the $F$-statistic threshold are identified as regions of recurrent copy number change.

### 3.4 Defining statistically significant CNA regions

The final step in the process of finding recurrent CNAs includes correction for multiple testing and assignment of statistical significance to the $F$-statistics. For this purpose, CNAnova uses the false discovery rate (Benjamini and Hochberg, 1995) estimated using the distribution of probes in the normal samples matched to the common reference set as described in Section 3.1. Successful estimation of FDR depends on knowledge of the number of FPs in our rejection regions. For the purpose of identifying significant $F$-statistics, we can utilize the distribution of $F$-statistics in the normal dataset to estimate FP on the premise that most of the extreme values of the statistic arising from normal-to-normal comparison will represent local variation in probe intensities (similar arguments for finding significant $P$-values of the $t$-statistics are discussed below) (Supplementary Fig. S7).

Such an approach is based on the assumption that the normal dataset does not have germline CNVs, and therefore any occurrences of large $F$-statistics will be attributed to non-biological variability in probe log-ratio values. To circumvent this problem, all probes with median absolute log-ratios >0.5 are not included in the estimation of the $F$-statistics. The selection of this threshold comes from the fact that median values higher than 0.5 or lower than $-0.5$ could be considered as an indication of potential single copy gains and losses in the region (see Supplementary Material for further discussion of threshold selection). After such adjustment, the following maxF procedure corrects for multiple testing and estimates the threshold for the $F$-statistics:

(1) Let $\Phi_r$ be the set of $F$-statistics in the reference dataset and $\Phi_c$ in the corresponding cancer dataset. For each window, remove that $F$-statistic from $\Phi_r$ if the maximal value of the median log-ratios across all samples in that window is more than the threshold $\theta$ (here set to 0.5);

(2) Sort the remaining $F$-statistics in $\Phi_r$, and in $\Phi_c$, in decreasing order;

(3) For a chosen FDR cut-off $\eta \in (0,1)$, select the smallest value $u$ such that $|\{v \in \Phi_r : v > u\}|/|\{w \in \Phi_c : w > u\}| > \eta$. Use $u$ as the threshold for calling significant $F$-statistics.

The maxF procedure essentially estimates the proportion of $F$-statistics derived from the subset of normal samples that are below the selected $F$-statistic threshold $u$. Since in most cases they represent FP hits, the threshold for controlling FDR can be calculated using the FDR formulation. Given a list of significant regions, CNAnova next finds significant $P$-values in order to determine which samples have CNAs. For this, we use the same steps as in the maxF procedure, but utilize $P$-values of log-ratios for probes that fall within CNA boundaries across all normal samples, leading to the minP procedure.

### 3.5 Identifying CNAs with intra-patient variability

The CNAnova algorithm can be extended to address additional questions that arise from the analysis of SNP array datasets. One important task in cancer research is finding recurrent CNAs that exhibit patterns of intra-patient heterogeneity. For example, researchers might be interested in identifying CNAs that are found in metastasis but not in primary cancer, or CNAs found only in metastasis to specific tissues such as lung. Here, we introduce an extension of the CNAnova framework based on the fixed-factor nested ANOVA model for identification of such recurrent CNAs with intra-patient variability.

The extension is a two-step process. First, samples from the same patient are grouped together and one-way ANOVA is run to identify recurrent regions. To insure that any possible decrease in the values of the nested $F$-statistics due to enlarged and inhomogeneous cancer samples are recreated in the normal samples, a similar transformation of merging normal samples into groups having the same size as the cancer samples is performed. The F-statistics from the one-way ANOVA allow us to identify regions of recurrent CNAs. Next, we apply a nested ANOVA model to each window in the regions previously identified as recurrent using the following model:

$$\begin{aligned} Y_{iuk} &= \mu+\tau_i+\theta_{i(u)}+\varepsilon_{iuk}, \quad i=0,1,2,\ldots,n_c, \\ &\quad u=1,2,\ldots,n_{ci}; \quad k=1,\ldots,l\lambda, \end{aligned} \tag{2}$$

where

- $Y_{iuk}$ is the $k$-th probe log-ratio from the set of $l\lambda$ probes of $u$-th cancer sample from individual $i$;
- $n_c$ is the number of cancer individuals;
- $n_{ci}$ is the number of samples for individual $i$;
- $i=0$ corresponds to the ANOVA-reference sample and $i>0$ to cancer individuals;
- $\mu$ is the mean of the probe log-ratios for the normal reference sample in the window, and $\tau_0=0$;
- $\tau_i$ is the non-random between-sample effect for cancer individual $i$;
- $\theta_{i(u)}$ is the non-random within-sample effect for cancer individual $i$. In our nested CNAnova implementation the reference sample represents a primary non-metastatic tumour;
- $\varepsilon_{iuk}$ are the random error terms such that $\varepsilon_{iuk}$ are independent random variables with $E(\varepsilon_{iuk})=0$ and $V(\varepsilon_{iuk})=\sigma^2$.

We derive overall $F$-statistics and $P$-values of $t$-statistics for each regression coefficient in the formula, using the procedures already described for one-way ANOVA. The focus on only recurrent CNA regions is primarily guided by the longer running time required for fitting complex linear models. The $P$-values for the intra-patient $t$-statistics derived from the nested model can then be used to determine the extent of intra-patient variability for each of the identified recurrent CNAs. The reference group for intra-patient comparisons can be defined by ordering samples within each group; for example, we give below the reference group includes primary tumours at diagnosis that are compared to relapsed samples.

## 4 RESULTS

### 4.1 Simulation data

*4.1.1 Simulation strategy* Due to the absence of an exhaustively validated dataset for benchmarking algorithms, simulated datasets are important. However, when trying to generate a simulated dataset for which the underlying copy number states are known, particular attention should be paid in preserving the noise distribution of the real data (Willenbrock and Fridlyand, 2005). This is usually the hardest part in the simulation process, as the distribution is composed of many components. These include high-affinity probes giving rise to extreme outliers, GC content of the genome producing wave patterns and sample purification steps such as whole genome amplification giving rise to additional probe intensity artefacts. The higher density of SNP arrays and the decrease in the number of probes in the probesets for the latest generation of Affymetrix arrays further complicate this problem. Our approach used normal

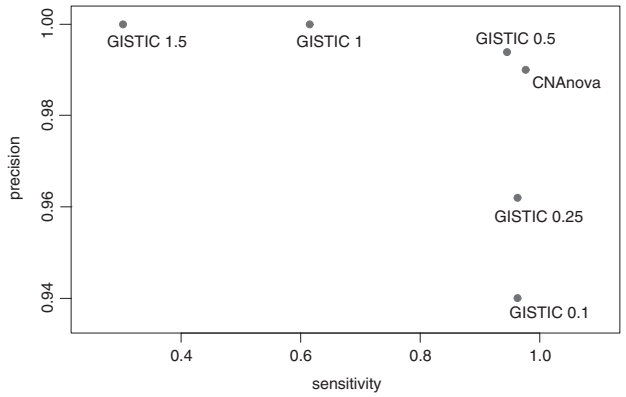


**Fig. 3.** Sensitivity versus precision curve comparing CNAnova and GISTIC in ability to find recurrent copy number alterations in the simulated data.

samples and simulated copy number changes in them. We selected 270 HapMap samples hybridized to Affymetrix SNP 6.0 arrays (McCarroll *et al*., 2008) to derive test data with which we could compare different methods.

The simulation of copy number changes recreates different possibilities for cross-sample occurrence of CNAs. We started with the following simulation scenario. First, the frequency of CNAs and their amplitude were defined by creating CNAs that occur in 1–75% of data and include from 10 to 300 probes (approximately 3–100 kb). The positional effect of CNAs, such as occurrence of deletions in different exons of a gene, was simulated by shifting CNAs around the middle position in some samples and segments. The procedure was applied multiple times to chromosome 17 of the HapMap data and 30 regions of recurrent copy number changes of different magnitude and frequency were simulated (details of the simulation scheme are listed in the Supplementary Material).

*4.1.2 Algorithm comparison* We used CNAnova with a default FDR 0.05 and GISTIC with threshold parameters 0.1 (program default), 0.25, 0.5, 1 and 1.5 to find recurrent CNAs in the simulated dataset. Since all simulated CNAs in our dataset have absolute log-ratio values >0.5, we expect GISTIC with a 0.5 threshold to have the lowest error rate. Before applying GISTIC, circular binary segmentation (CBS) (Olshen *et al*., 2004) was used to segment log-ratios. We verified that CBS found all simulated CNAs, thereby ensuring that segmentation errors did not propagate though GISTIC (Supplementary Fig. S8). True positive rate (TPR; or sensitivity) and positive predictive value (PPV; or precision) were used to assess performance of the algorithms.

Results for the overall dataset (Fig. 3) suggest that although CNAnova and GISTIC share similar precision, CNAnova produces the best results when both precision and sensitivity are taken into account. In addition, we found that GISTIC tended to split single CNA regions into multiple regions, as was observed 1, 3 and 5 times for GISTIC run with threshold parameters 0.5, 1 and 1.5, respectively. In addition to altering the biological significance of results after such an artificial splitting (i.e. by missing some genes or distorting co-occurrence metrics), this may underestimate the actual frequency of CNAs. In contrast, CNAnova detected all simulated CNAs without splitting any regions. Both algorithms have the maximal possible precision of 1 when only CNA regions

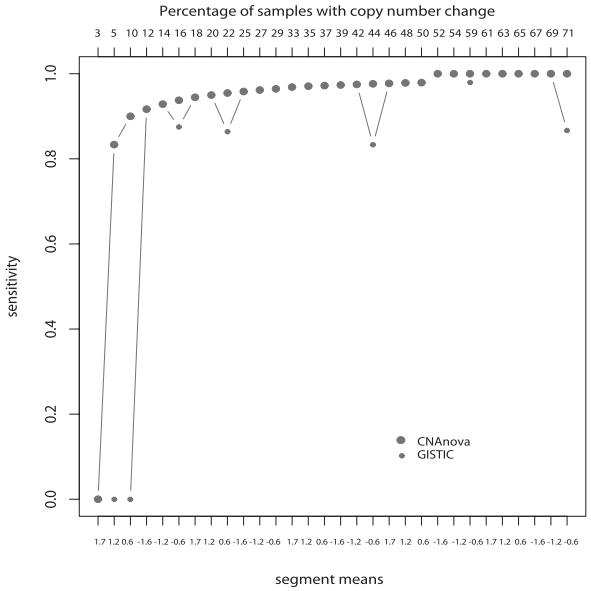


**Fig. 4.** Sensitivity of CNAnova and GISTIC, run with 0.5 threshold, in identifying recurrent copy number changes of different frequency and magnitude. Top *x*-axis label gives the percentage of samples containing a given recurrent CNA; bottom *x*-axis label gives the mean log-ratio value used in simulation.

with >49% of changed samples are taken into account. However, with decreasing CNA frequency CNAnova starts to show higher sensitivity. In particular, GISTIC with the optimal 0.5 threshold cannot detect any samples altered in 5–10% of samples, while CNAnova detects many such samples (Fig. 4).

To confirm this observation using a larger sample size, we created an additional benchmarking dataset where only rare low-frequency CNAs (one copy gain and loss) were simulated. CNAnova outperformed GISTIC with a 0.5 threshold, this time in both sensitivity (0.86 versus 0.98) and precision (0.94 versus 0.99). This trend was observed for rare recurrent CNAs with different mean values and frequency (Fig. 5). The decrease in the size of segments with the lowest simulated mean values after applying the default GISTIC threshold may partly explain GISTIC's fall in sensitivity and precision. This underscores the thresholding limitation of segmentation-based methods and contributes to the jumpy nature of sensitivity plots. The difference in sensitivity and specificity between GISTIC and CNAnova becomes less prominent once the copy number change of rare CNA is increased to include two copy gains and deletions (Supplementary Fig. S11). However, GISTIC had a much higher error rate when assessed on simulated non-significant changes with means (±)0.2/0.3 (29% versus 17% for CNAnova), suggesting that the method is less robust against outliers occurring within recurrent CNA regions. In the Supplementary Material, we compare the performance of CNAnova and GISTIC using a hypothesis testing framework.

## 4.2 Acute lymphoblastic leukaemia (ALL) Affymetrix SNP 6.0 data

To test the performance of CNAnova and GISTIC on real data, we used a recently published study comprising 94 samples of ALL and 36 matched normal samples analysed using Affymetrix SNP

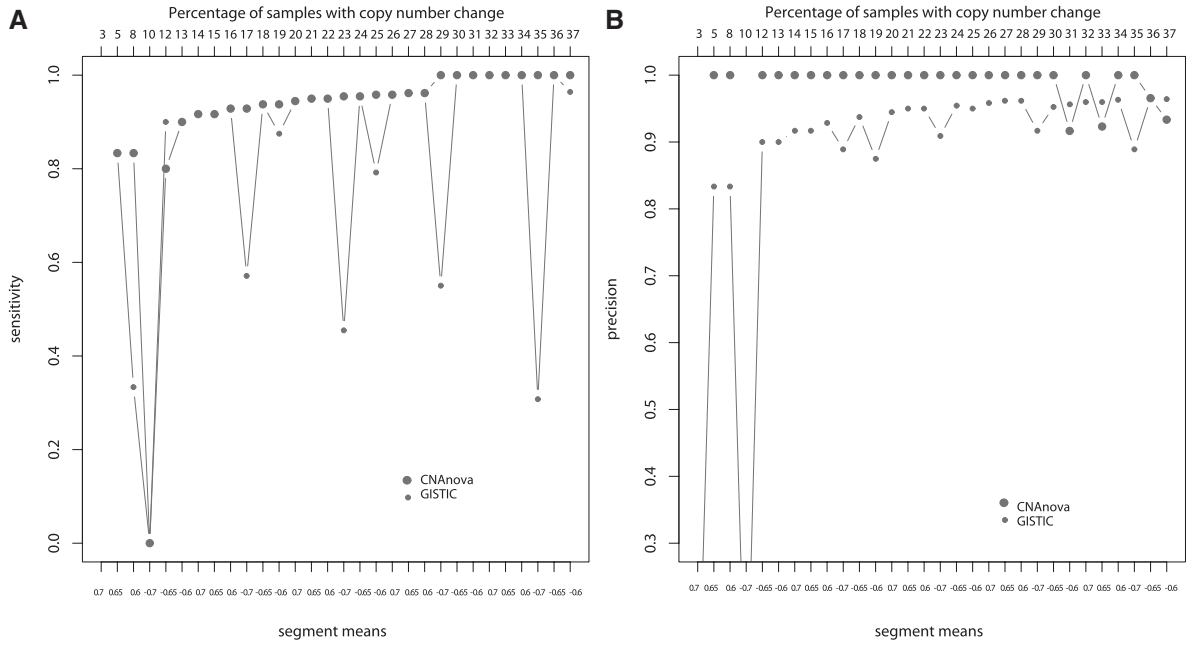


**Fig. 5.** Sensitivity (**A**) and precision (**B**) of CNAnova and GISTIC, run with the thresholds of 0.5 on log-ratio values, in identifying rare recurrent low-level copy number gains and losses (colour version of the figure is available in Supplementary Material). Top *x*-axis label gives the percentage of samples containing a given recurrent CNA; bottom *x*-axis label gives the mean log-ratio value used in simulation (colour version of the figure is available in Supplementary Material).

6.0 arrays (Mullighan *et al*., 2008). The 94 samples were derived from 47 patients and were grouped into diagnosis–relapse pairs. The choice of this particular study was guided by the fact that the authors experimentally confirmed many previously described and newly identified copy number changes with quantitative genomic PCR. We used raw .CEL files and carried out all normalization and summarization steps using *aroma.affymetrix* (Bengtsson *et al*., 2008). The normalized raw probe intensity values were then fed into the CNAnova pipeline. The data put into GISTIC were derived from the same pre-processing steps, except that no pseudo-replication was done. For instance, the same CNV probes were flagged in the GISTIC data. The segmentation was carried out using CBS and GISTIC was run with threshold parameters of 0.1, 0.25, 0.5, 1 and 1.5.

The problem of algorithm comparison using real data usually arises from the fact that many hits are unvalidated. The ALL dataset has two advantages when compared to other similar datasets. First is the fact that many copy number changes were experimentally confirmed. Second, due to the rearrangements of the immunoglobulin genes in the B-cell lineage, the location of deletions can be predicted *a priori* with high accuracy and this can serve as a positive control to assess the error rate of CNAnova and GISTIC. These two sets gave us a reference list of CNA regions for computing sensitivity and precision of the algorithms.

For the segments not overlapping these regions, we used the following strategy to assess the possibility of a segment being a FP. We began by looking at the number of probes that were covered by each segment. Very small segments have a higher chance of being a FP. This chance further increases if segments do not span the known annotated genes. Finally, we looked at the log-ratios.

**Table 1.** Sensitivity and precision for CNAnova and GISTIC on the ALL dataset reference genes list

| | GISTIC0.5 PPV | GISTIC0.5 TPR | GISTIC1 PPV | GISTIC1 TPR | CNAnova PPV | CNAnova TPR |
|---|---|---|---|---|---|---|
| IGK@ | 0.93 | 0.95 | 0.94 | 0.95 | 0.96 | 0.93 |
| IGH@ | 0.93 | 0.96 | 0.91 | 0.96 | 0.95 | 0.92 |
| IGL@ | 0.94 | 0.92 | 0.91 | 0.92 | 0.92 | 0.93 |
| CDKN2A | 0.94 | 0.96 | 0.94 | 0.95 | 0.95 | 0.92 |
| CDKN2B | 0.93 | 0.96 | 0.94 | 0.95 | 0.96 | 0.93 |
| IKZF1 | 0.93 | 0.94 | 0.93 | 0.96 | 0.95 | 0.93 |
| ETV6 | 0.66 | 0.83 | 0.75 | 0.70 | 0.83 | 0.75 |
| PAX5 | 0.71 | 0.76 | 0.77 | 0.71 | 0.80 | 0.76 |

The segments satisfying the two previous conditions and with values <0.5 (more than −0.5 for losses) were deemed as false negative. The rates of sensitivity and precision for CNAnova and GISTIC on the reference list of genes are shown in Table 1.

Both methods performed well in identifying highly recurrent rearrangements of immunoglobulin genes in the B-cell lineage. However, overall GISTIC identified a larger number of gains and losses even at the 'biological' threshold of 0.5 than CNAnova (Supplementary Fig. S12). Many of these CNAs represent very small focal copy number changes. For GISTIC with the threshold of 1, >80% of CNAs encompassed <15 probes and a quarter of all changes were <4 probes (roughly 2 kb in size, based on the median probe spacing for Affymetrix SNP 6.0 arrays) (Supplementary Fig. S13). Most genes residing in these CNA regions have biological functions

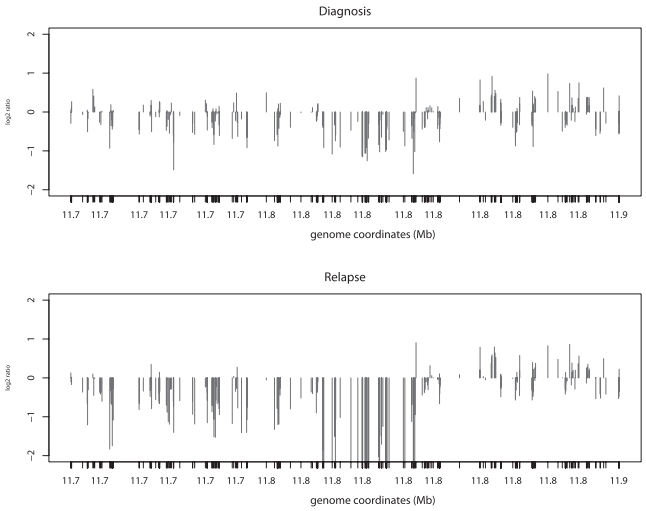


**Fig. 6.** Example of copy number plots showing the loss of ETV6 in the relapsed ALL samples identified with fixed-factor nested-design CNAnova.

that do not suggest involvement in ALL progression (i.e. SIRPB1, PCMTD2 and RBL1 on chromosome 20 and APOBEC3B, VPREB1 and RBL1 on chromosome 22). CNAnova identified 60 recurrent CNAs (Supplementary Table 1). A large number of recurrent CNAs were identified chromosome 9, containing amplicons around the PAX5 and CDKN2A/B genes among others, which confirms the previous observations about high recurrence of CNAs on this chromosome (Mullighan *et al*., 2008).

Compared with simulated data, GISTIC seems to have a higher FP rate on the real dataset, leading to lower precision at a comparatively uniform sensitivity rate. Furthermore, at low thresholds GISTIC splits single continuous regions into several smaller CNAs. For instance, the number of recurrent CNAs reported by GISTIC with a 0.5 threshold was 177 for the ALL dataset, which is much larger than the average number of CNAs reported in similar studies using the same or similar array platforms (Chin *et al*., 2008; Weir *et al*., 2007). These observations allow us to conclude that CNAnova applied to the noisy Affymetrix SNP 6.0 array data produces a much smaller number of FPs and does not require the time-consuming process of selecting an optimal threshold and an appropriate segmentation method. Having been developed and tested on the Affymetrix SNP 250K arrays that have only 250K probes and a more robust probe design, it seems that the noisier and denser Affymetrix SNP 6.0 arrays increase the error rate for GISTIC. This could also explain why in the analysis of the lung cancer data generated using Affymetrix SNP 250K arrays, GISTIC was used with a threshold of 0.1 (Weir *et al*., 2007), while an Affymetrix SNP 6.0 study of human glioblastoma used GISTIC with a threshold of 1 (Chin *et al*., 2008).

*4.2.1 Identification of recurrent CNAs with intra-patient heterogeneity* We applied the generalized CNAnova approach to the ALL dataset to find recurrent copy number changes that were present at diagnosis but not relapse or *vice-versa*. The model was developed and tested for its ability to detect recurrent CNAs with intra-patient heterogeneity. We identified 24 recurrent CNAs that exhibit patterns of intra-patient heterogeneity (Supplementary Table 2). These included the CDKN2-cluster, IFNK, ETV6 and IKZF1, which have been previously reported to have recurrence at relapse, and sometimes diagnosis (Mullighan *et al*., 2008), as well as new genes such as ATF7IP (Fig. 6) on chromosome 12. Deletions of ETV6, which contains ATF7IP, have been reported previously (Montpetit *et al*., 2004); however, we find cases that have an independent focal deletion of ATF7IP.

## 5 DISCUSSION

We have presented a novel framework, CNAnova, for using unsegmented SNP microarray data to identify recurrent copy number aberrations in cancer. It uses pseudo-replication to increase statistical power and, in contrast to segmentation-based methods, does not require permutations to estimate the null distribution of test statistics. Instead, this distribution is approximated using samples from a control set representing normal samples.

CNAnova circumvents the need to specify a threshold on segmented log-ratios for calling probes with copy number changes. This threshold could be sample, region, dataset and platform dependent, which makes it hard to determine an optimal value and can increase the error rate of segmentation-based methods. This is especially true for the current generation of the SNP arrays, such as Affymetrix SNP 6.0, that often exhibit wave patterns and other artefacts.

Another advantage of CNAnova stems from decoupling the process of identifying cross-sample boundaries of the regions of recurrent CNAs and pinpointing the actual samples with copy number changes. This insures that copy number changes with extreme probe intensity values do not influence CNA detectability in other samples. High probe intensity values fed into permutation-based methods such as GISTIC may increase the false negative rate by increasing the G-score (for GISTIC) in the permuted dataset, and therefore the overall threshold for calling significant CNAs. In its later implementation, GISTIC has introduced capping of segmented log-ratios with extreme probe intensities (Chin *et al*., 2008). The comparison with GISTIC suggests that it tends to have higher error rate than CNAnova when using Affymetrix SNP 6.0 arrays. In addition to the dataset-dependent thresholds derived from the permutation-based estimation of the null distribution, the higher error rate could be explained by the fact that GISTIC was originally adapted for use with lower density arrays and a more robust probe-set design.

At the same time, dependence on the control dataset requires noise in the cancer and normal samples to follow approximately identical distributions. In addition, the number of normal samples should be sufficient to capture different probe intensity artefacts and patterns of noise in the (often larger) cancer dataset. Further simulations showed (Supplementary section 'Results') that the algorithm is typically robust against sample size and noise properties of the control dataset. Only in cases of extremely low control sample sizes ($<10\%$ of the size of the cancer dataset) the impact on the error rate becomes significant.

Some conceptual ideas relating to different steps of the CNAnova framework have been discussed in the literature. Hautaniemi *et al*. (2005) derived a parametric mathematical model for protein-based assays, which was then used to simulate additional data points to better capture dependencies between variables within the decision tree. Such an approach resembles our pseudo-replication strategy. The use of normal control samples to control FDR was proposed

by Rozowsky *et al*. (2009) to correct the FP signals in ChIP-seq data arising from the open chromatin conformation at pol-II sites. Decomposition of the variance has been recently proposed in the context of aCGH data analysis by Kim *et al*. (2009), where CNAs were associated with cancer tissue types.

The statistical framework of CNAnova can be extended by incorporating more complex models, as we exemplified with a two-factor nested ANOVA design for identifying CNAs with intra-patient heterogeneity. A useful extension could be developed to identify recurrent CNAs prevalent in a particular tumour subtype or progression stage. In contrast to clustering, such an approach can provide an association of individual CNAs with a specific tumour phenotype. Focus on individual regions rather than clusters could facilitate the discovery and testing of new candidate genes. Similarly, additional data and biological annotation, such as gene expression data or pathway-centred gene sets, could be used to discover functionally related sets of recurrent CNAs.

## ACKNOWLEDGEMENTS

We thank Oscar Rueda, Andy Lynch, Christina Curtis at the Cancer Research UK Cambridge Research Institute and Rameen Beroukhim at the Dana Farber Cancer Institute for helpful discussions. We also acknowledge the assistance of Dr. Charles Mullighan of St. Jude Children's Research Hospital, Memphis in providing the raw ALL data and comments on the predictions from CNAnova.

*Funding*: The authors acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

*Conflict of Interest*: none declared.

## REFERENCES

Bengtsson,H. *et al*. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.

Chin,L. and Gray,J.W. (2008) Translating insights from the cancer genome into clinical practice. *Nature*, **7187**, 553–563.

Chin,L. *et al*. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **7216**, 1061–1068.

Colella,S. *et al*. (2009) QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, **35**, 2013–2025.

Comaniciu,D. *et al*. (2001) Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intel.*, **123**, 343–351.

Diskin,S. *et al*. (2006) STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, **16**, 1149–1158.

Diskin,S. *et al*. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, 1–12.

Greenman,C. *et al*. (2009) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Hautaniemi,S. *et al*. (2005) Modeling of signal-response cascades using decision tree analysis. *Bioinformatics*, **21**, 2027–2035.

Kim,K. *et al*. (2009) Identification of significant regional genetic variations using continuous CNV values in aCGH data. *Genomics*, [doi:10.1016/j.ygeno.2009.08.006].

Marioni,J. *et al*. (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, **8**, R228.

McCarroll,S. *et al*. (2008) A comparison study: applying segmentation to array CGH data for downstream analyses. *Nat. Genet.*, **40**, 1166–1174.

Montpetit,A. *et al*. (2004) Mutational and expression analysis of the chromosome 12p candidate tumor suppressor genes in pre-B acute lymphoblastic leukemia. *Leukemia*, **18**, 1499–1504.

Mullighan,C. *et al*. (2008) Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, **5906**, 1377–1380.

Nilsson,B. *et al*. (2009) Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*, **25**, 1078–1079.

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

Pawitan,Y. *et al*. (2005) Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21**, 3865–3872.

Pinkel,D. and Albertson,D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**(Suppl.1), 11–17.

Rozowsky,J. *et al*. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Shah,S. (2008) Computational methods for identification of recurrent copy number alteration patterns by array CGH, *Cytogenet. Genome Res.*, **25**, 603–619.

Wang,C. *et al*. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.

Wang,L. *et al*. (2009) MSB: A mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res.*, **19**, 106–117.

Weir,B. *et al*. (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **7171**, 893–899.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

# CNAnova: a new approach for finding recurrent copy number abnormalities in cancer SNP microarray data. Supplementary Material

Sergii Ivakhno & Simon Tavaré

February 24, 2010

# Introduction

In this supplementary material we collect together some additional information and analyses that support the results in the main text. At the end we include colour versions of the figures from the text.

## SNP array Segmentation Algorithms

The most common problem in the analysis of SNP CGH data is the transformation of normalized log-ratio values into accurate copy number calls at the highest possible resolution. This so-called *single sample segmentation problem* has received much attention, with numerous methods developed for this task.

- `QuantiSNP` (Colella *et al.*, 2007) Presently, this is Illumina specific, but that is expected to change. Ideally suited for the analysis of germline variants.
- `PennCNV` (Wang *et al.*, 2007) Suitable for either Affymetrix or Illumina data. Can perform trio calling.
- `GADA` (Pique-Regi *et al.*, 2008) Applicable to either Affymetrix or Illumina. Includes cross-sample inference step. Perhaps ideal for detecting germline variants. Matlab-based implementation.
- `SNPchip/VanillaICE` (Scharpf *et al.*, 2007) Bioconductor package, can interface with `oligo`. HMM-based approach that can be applied to Affymetrix or Illumina. Not explicitly suited for cancer samples.
- `PICNIC` (Greenman *et al.*, 2010) Developed for Affymetrix SNP 6.0, particularly for cancer data. HMM-based approach. Quite heavy in terms of compute time. Benefit of ASCN calls and LOH. (Matlab/Java)
- `genoCNA` (Sun *et al.*, 2009) New extension of the PennCNV approach. Simultaneously identifies copy number states and genotype calls. Adapted to detect copy number aberrations in cancer samples.

- `CBS` (Olshen *et al.*, 2004), `GLAD` (Hupé *et al.*, 2004), `HaarSeg` (Ben-Yaacov and Eldar, 2008), and `Ultrasome` (Nilsson *et al.*, 2009). Applicable to either aCGH or SNP-CGH platforms. Perform segmentation, without genotype estimation.

# Methods

### Relevance of spline-correction for identification of recurrent CNAs

Finding recurrent copy number changes in cancer presents an additional challenge of having to deal with widespread genomic instability, which produces many passenger aberrations with gains and losses of large chromosomal segments. These segments may contain a large number of genes; consequently, even if large-scale aberrations are marginally significant for cancer progression, the original molecular driving event is hard to confirm. Therefore, when finding focal recurrent CNAs, CNAnova does not use large-scale copy number changes. It corrects for these large copy number changes by applying smoothing splines before identifying the boundaries of CNA regions.

After inspecting a large number of samples we found that having one degree of freedom (df) in the spline function for every 1500 probes has the desirable effect of flagging aberrations larger than 10 Mb. The degrees of freedom are calculated using this constant and are different for different chromosomes. Although the distribution of CNA sizes may vary between different studies, splines in CNAnova are used only to offset infiltration of potentially causative/driver CNAs with large-scale aberrations caused by genome instability. Exploratory analysis showed that between 1000 and 2500 probes per df produce a similar smoothing, suggesting the robustness of the constant we selected and its transferability between experiments.

However, even large-scale aberrations might involve small regions that harbour a functionally significant oncogene or tumour suppressor gene. Consequently, once CNA regions are identified using spline-corrected data in the first pass of CNAnova (via F-statistics), the original data are used to identify all samples that have copy number change in a selected CNA region. In this way information on the focal recurrent CNAs can be used to find putative driver mutations in regions of large-scale copy number changes. (Figure S1.)

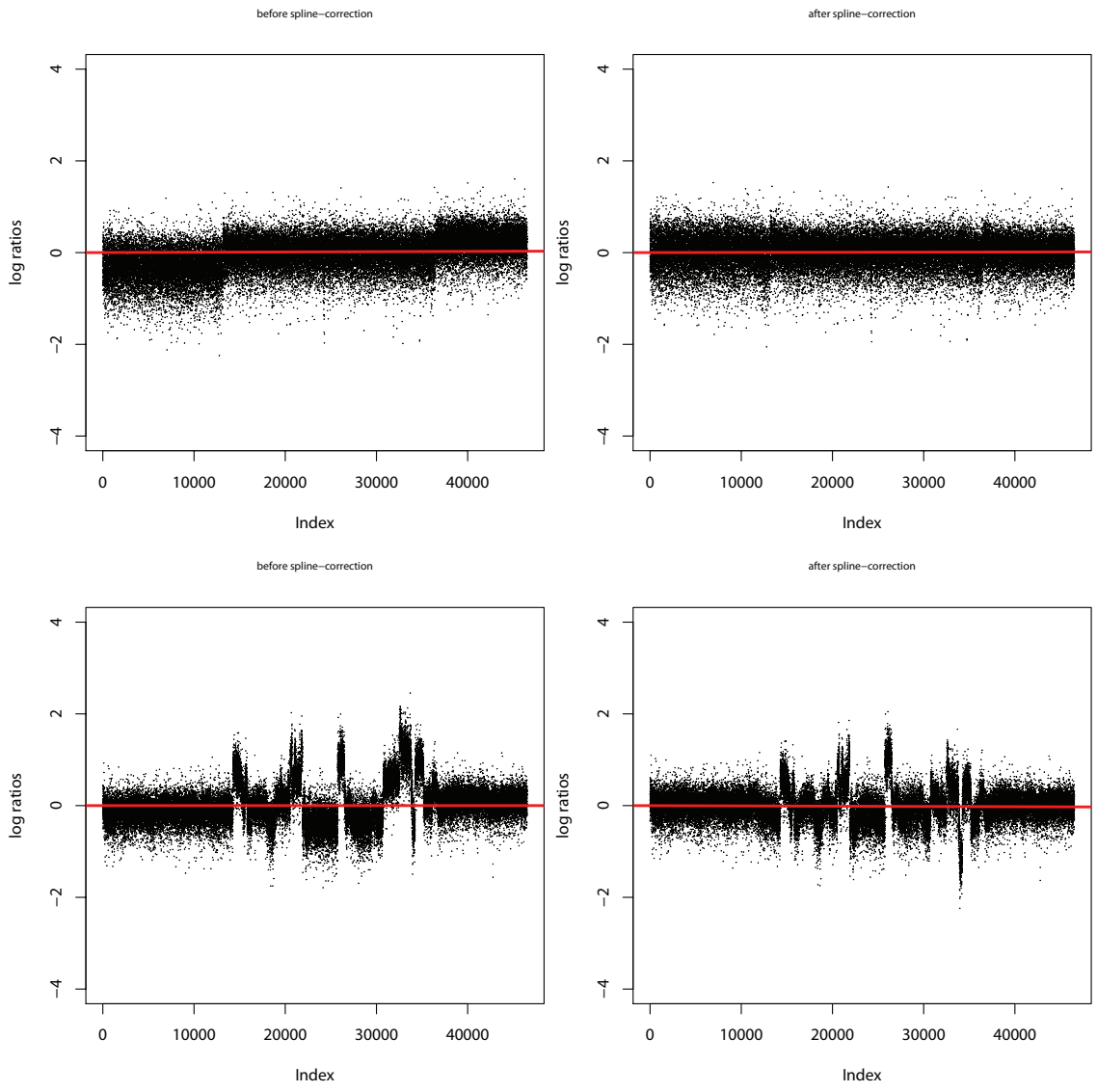


Figure S1: Application of smoothing splines for identifying and correcting for large-scale single copy number gains and losses. Two chromosomes (left) with different distributions of copy number changes were subjected to smoothing spline correction. Results (right) depict the same chromosomes after correction. Although low-level changes were smoothed out, truly focal high-level CNAs remain intact.

### Note on the assumptions of CNAnova model

The applicability of the ANOVA procedure is contingent on satisfying two assumptions about the dataset:

- observations within and between samples are independent;
- variances within and between each of the samples are (approximately) equal (homogeneity of variances requirement).

Considering the positive dependencies (autocorrelation) in the log-ratios of adjacent probes on the chromosome and across samples, the ANOVA model might seem to have severe departures from both assumptions. However, here ANOVA is applied to short probe windows, therefore making any temporal dependencies on the sub-chromosome scale less prominent and influential. In most cases, values in each sample window will represent either the distribution of log-ratios of the non-changed population or a population with some copy-number changes. Although we do observe autocorrelation in many probe windows (p-values $< 0.05$; Durbin-Watson test for autocorrelation), it should be emphasized that CNAnova uses F-statistics and p-values only in relation to normal samples. For example, although autocorrelation will cause shrinkage of the within-sample variance and increase in the F-test statistics (Figure S2), the autocorrelation is equally prominent in the normal samples.

Homoscedasticity of the variance is a more subtle assumption of the CNAnova statistical model. It clearly does not hold for segments that span breakpoint boundaries. However, deleterious effects of violating the homoscedasticity assumption in our model are offset by several factors. First, the large number of groups and equal sample sizes make departures from homoscedasticity less prominent, especially if they occur in only a few samples for a particular segment (and usually exact breakpoint locations vary between samples). The total number of segments exhibiting homoscedasticity depends on the number of recurrent CNAs, which for most datasets does not exceed 0.005% of tested segments. This insures preservation of homoscedasticity for most segments in the dataset. Finally, increase in the variance caused by heteroscedasticity is associated with equally prominent change in segment means, thereby lowering false negative rate for "homoscedastic" segments. We used a simulated dataset (see Results section below) to verify these assertions empirically. Exploratory data analysis using the variance of within-group standard deviations confirmed

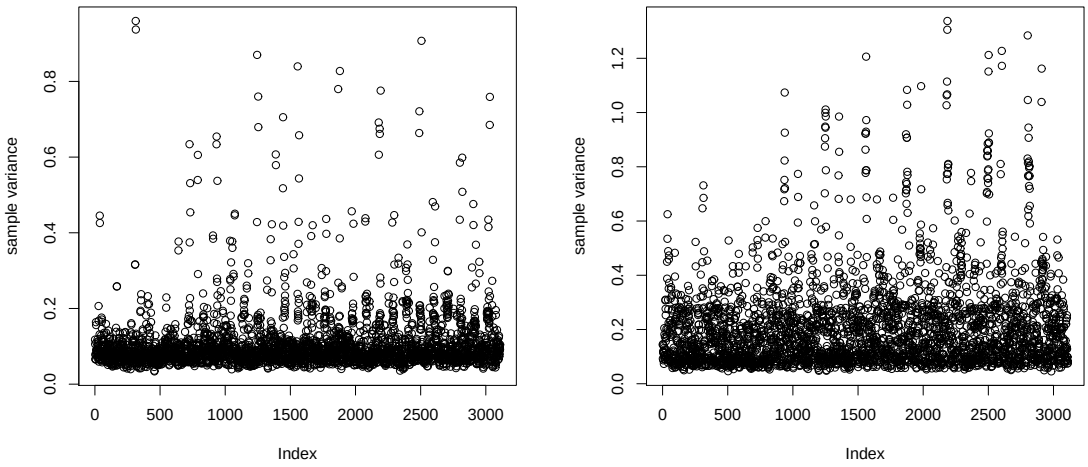


Figure S2: Distribution of the sample variance (window of 150 probes) in the original dataset (left) and from the sample after probe permutation (right). A more prominent autocorrelation in the original dataset causes shrinkage of the sample variance.

that only segments spanning CNA boundaries exhibit heteroscedasticity (Figure S3 and Figure S4).

We have also found correlation between recurrent CNA segments missed by CNAnova (false negatives, FN) and corresponding standard deviations between false negative and true positive segments (Welch two-sample t-test, df = 84.34, p-value = $9.86 \times 10^{-6}$). In contrast, a similar two-sample test comparing absolute mean values of false negative and true positive segments found differences with smaller significance (p-value = $2.26 \times 10^{-3}$), suggesting that change in standard deviations predominate over changes in segment means for the FN segments. We also find higher correlation between segment means and corresponding standard deviations for false positive segments (0.93) than false negative (0.88). These observations suggest that segment means appear to be a more important property of the model that offset negative impacts from heteroscedasticity in the data.

Related to the ANOVA assumptions question is the problem of data transformation. It might appear that a much simpler strategy would be to remove the log-ratio calculation step from CNAnova and work with log-intensities instead. Unfortunately, the distribution of log-intensities is highly skewed, violating an assumption of ANOVA. The mean skewness of

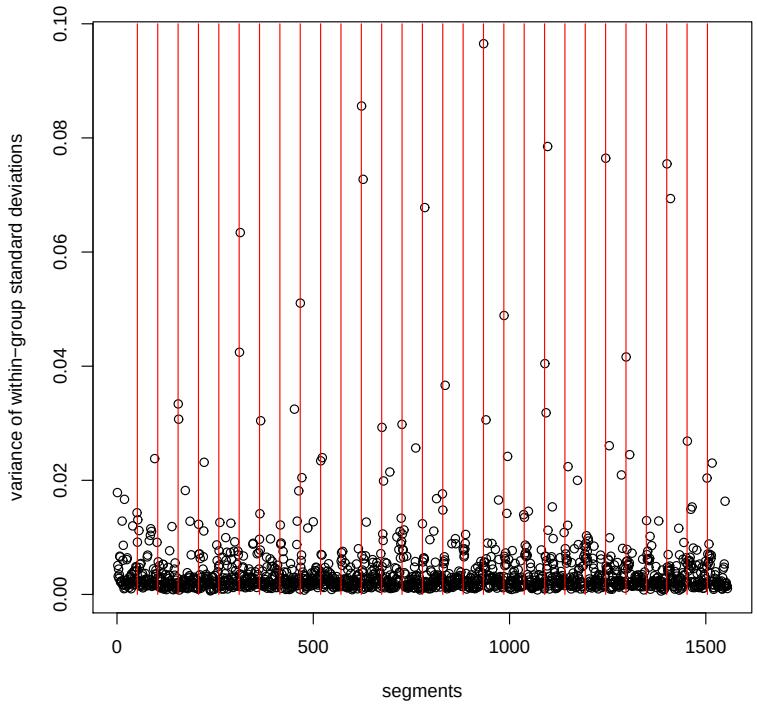


Figure S3: Distribution of variance of within-group standard deviations for all segments in the simulated dataset. The red lines show the location of recurrent copy number changes. Figure shows preferential location of segments exhibiting heteroscedasticity of the variance around recurrent CNAs.

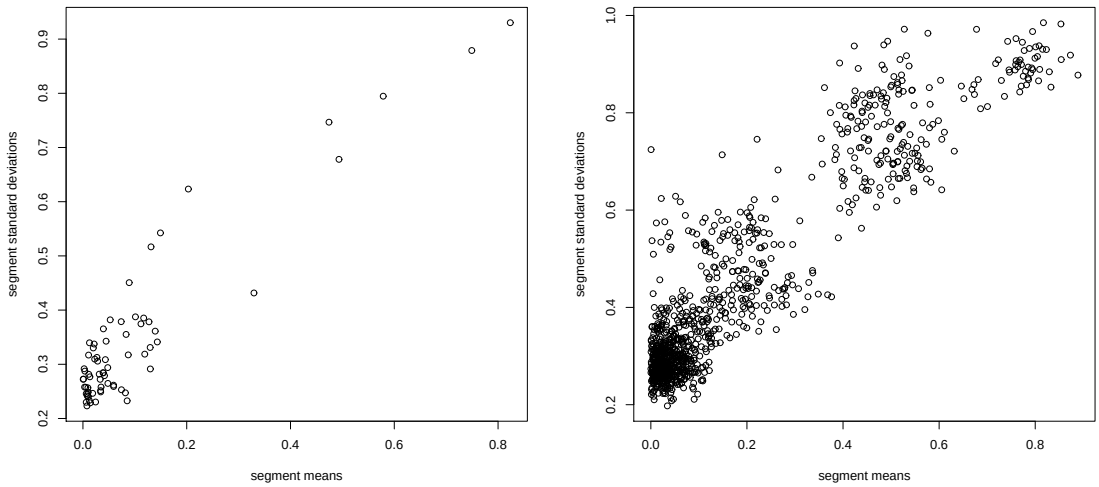


Figure S4: Scatter plot of standard deviations and absolute value of means for all segments in the simulated dataset (left) and for the false positive and negative segments only (right). Most of the segments are correctly classified due to the counterbalancing effect of the increase in mean values, which has a clear positive correlation (= 0.98) with the standard deviation. The misclassified segments are predominantly the ones with very low mean values, which results from the fact that in the misclassified segments spanning the breakpoint, the majority of probes fall into the region with normal copy number.

log-intensities from the normal samples in the AML dataset is -0.43 (2.7 for raw intensities), compared to only 0.09 for log-ratio values for the same samples. This supports the use of log-ratios in CNAnova, although log-intensities might prove easier to work with when non-parametric (non-Gaussian) models are used.

### The mean-shift procedure

Mean-shift density estimation theory was first described by Fukunaga and Hostetler (1975) and expanded in Comaniciu and Meer (2002), who also suggested a versatile application of the method to discontinuity-preserving smoothing of signals in various application domains such as image analysis. Recently, an application of the mean-shift method to the segmentation of aCGH data was described by Wang *et al.* (2009).

Mean-shift uses modes rather than means for data point smoothing, which allows it to demarcate sharply the boundaries of noiseless signal regions. In CNAnova, the mean-shift approach attempts to find maxima in the density (pdf) of the F-statistics; at modes of the pdf, the gradient is zero. The mean-shift vector therefore always points in the direction of the maximum increase in the density. This is an iterative procedure that shifts each data point to the density maxima. Its advantage over model-based methods is that as a non-parametric technique, it does not require prior knowledge of the number of recurrent CNAs or assumptions about the length distribution. By performing a discontinuity-preserving smoothing of the F-test statistic, it removes the noise in homogeneous regions of the chromosome and preserves discontinuities at the same time. For the full derivation of the mean shift procedure and density gradient estimation see Comaniciu and Meer (2002).

From the CNAnova perspective, the key advantage of the mean-shift procedure is the use of local information based on mode estimation, which differentiates it from traditional smoothing methods. Each point is assigned to a significant mode located in its neighborhood, which is a local maximum of the underlying density function.

## Results

### Details of the simulation scheme

We simulated ten instances of chromosome 17 with spiked-in CNAs in different regions. The distribution of log-ratio values in each region was sampled from a normal distribution with means 0.6, 0.65, 0.7, 1.1, 1.15, 1.2 (-0.6, -0.65, -0.7, -1.1, -1.15, -1.2) to recreate one and two copy number gains/losses and a constant variance of 0.31, inferred from the distribution of log-ratios in the real CNVs of the HapMap data. To test algorithms for robustness against outliers present within regions of recurrent CNAs, we selected 6 highly recurrent CNA regions and simulated non-significant log-ratio changes, 3 with mean 0.2 and 3 with mean -0.3, in 4 samples that do not have CNAs in those regions.

To insure an unbiased comparison of CNAnova and GISTIC, particular attention was given to selecting the correct threshold for GISTIC. The selection of the 0.5 threshold was governed by two key observations. First, this threshold was greater than all simulated spiked-in CNAs. Second, by running GISTIC with other thresholds we found that thresholds of 0.25 and 0.1 produce the same number of CNAs (47) as the 0.5 threshold, while increasing the threshold to 1 and 1.5 reduces the number of identified recurrent CNAs to 19 and 7 respectively, suggesting a large number of false negative hits.

### Null hypothesis testing

In part, we attribute the superior performance of CNAnova in detecting rare CNAs to the fundamental difference in how CNAnova and permutation-based algorithms estimate the null distribution. When using permutations (as in GISTIC), the null distribution can be defined in terms of the number of times that copy number changes with a particular cross-sample frequency occur in the dataset, given that CNAs in the genome are distributed randomly. In CNAnova the null distribution is estimated based on the premise that any local variation in probe intensities arises from the non-CNA random effects. The null hypothesis in CNAnova is much more general than that of GISTIC and similar permutation-based algorithms, which allows CNAnova to detect infrequent recurrent CNAs with higher sensitivity. The null hypothesis also does not depend on the overall rate of genomic instability observed in individual samples. For example, a large number of CNAs on a particular chromosome

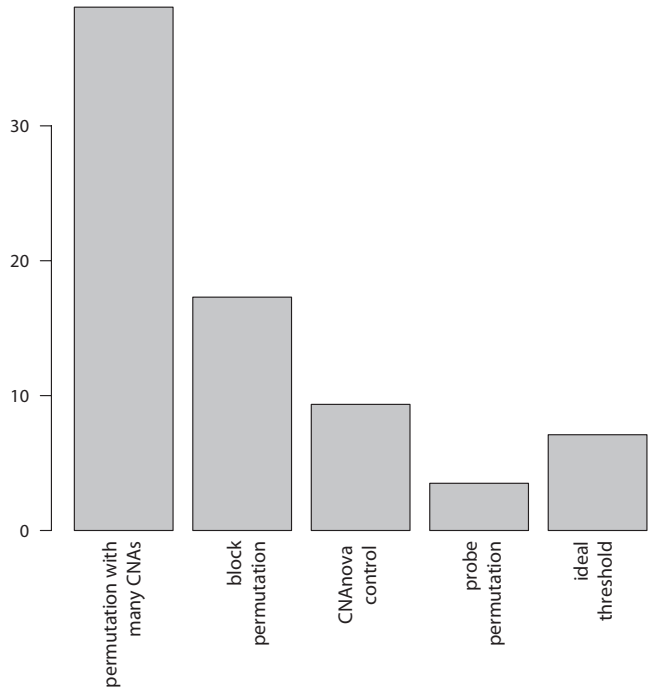


Figure S5: F-test thresholds for the CNAnova algorithm derived using different permutation strategies or probe distribution in normal samples.

will produce too conservative a null distribution that may lead to many false negative hits for a permutation-based method. Alternatively, a small number of CNAs will generate low thresholds and potentially a large number of false positives.

The actual percentage of altered samples within a CNA region provides a more coherent way of defining functionally interesting recurrent CNAs. For example, CNAs with a frequency of 4%–5% could suggest the presence of rare tumour subtypes. To further test this hypothesis we used permutation-based estimation of the null distribution to derive thresholds for the F-statistics and p-values used in CNAnova. The simulated cancer dataset was permuted 100 times, either keeping ties between 5 adjacent probes (the minimal size of a CNA region) or using probe permutation. To assess the effect of large-scale genomic instability, we simulated single copy gains of chromosome arms in 10 samples and used probe permutation on this adjusted dataset. CNAnova was then run to estimate F- and t-statistics for the null distribution and the ideal threshold was defined as the smallest F-statistic that gave the lowest error rate in the simulated dataset (Figure S5). The results suggest that for block-permutation and probe permutation with chromosome arm gains,

both F-test and t-test thresholds were much higher than the ones obtained with the null distribution derived from the control dataset of normal samples. As a result of abnormally high thresholds values, the sensitivity and specificity estimates were much lower and no CNAs with a frequency below 15% were detected (data not shown). We note that the permutation-based null distribution can become very liberal if ties between adjacent probes are not kept. In this case, any correlation of probe intensities along the genome is lost (e.g., autocorrelation of probes in regions displaying the wave artefact) and the estimates of the F-statistics become very low (5.3 in our simulation, compared to 9.35 for CNAnova). For this reason methods such as STAC (Guttman *et al.*, 2007) use permutation of blocks to derive the null distribution.

### Detection of germline CNVs

One of the CNAnova pre-processing steps involves identifying and flagging CNVs in the normal control dataset. Because this stage is crucial for insuring that normal germline variation does not interfere with the detection of somatic cancer mutations, we sought to test the ability of CNAnova to find regions of previously discovered CNVs. It should be emphasized that CNAnova will not find all CNVs in normal population. Since it uses a 30-probe moving window, any CNVs spanning less than 10 probes are unlikely to be detected by the method and therefore should not be a concern when identifying recurrent CNAs.

We utilized the same 90 HapMap samples used to create the control set for the simulated dataset, but this time applied CNAnova to all chromosomes of the normal samples. 505 CNVs detected in more than 3% of those samples were used as a reference (McCarroll *et al.*, 2008). CNAnova was run with segment mean thresholds of 0.3, 0.35, 0.4, 0.45, 0.5 for distinguishing true single copy number changes and random probe variation. We find that CNAnova achieves 41% sensitivity and 63% specificity on the reference set for the 0.4 threshold, which gave the best combination of sensitivity and precision. Although this numbers might appear low, we find that segments missed by the method are either of low frequency or small size (two-sample t-test, p-values 0.09 and 0.18 respectively). For example, the median frequency of false negative CNVs was 13% and the median size 6.2 kb, compared to 37% and 23 kb respectively for true positive CNVs. Although p-values are low, they are still not significant. However, if we use combined minimal p-values from CNV size

and frequency t-tests, the relationship becomes marginally significant (p-value 0.06). It is therefore highly unlikely that any of those missed segments would have been detected as a recurrent CNAs in cancer samples. After removing CNVs below 5% frequency and 3 kb size we observe significant improvement in sensitivity and precision (69% and 78% respectively).

These assertions, however, depend on the assumption that normal and cancer samples come from the same population and have roughly equal sample size. Consequently, the detection of CNVs will be impeded if the size of the normal dataset is much smaller than the corresponding cancer dataset.

## Impact of pseudo-replication on sensitivity of CNAnova and sample matching

Next, we sought to validate empirically the effect of pseudo-replication on the ability of CNAnova to detect recurrent CNAs and to determine the appropriate size of $\lambda$. The dataset was simulated as described above, but without matching tumour samples to multiple normal control samples, or by matching to two, three, four or five samples. We observed a decrease in dataset-wide precision and sensitivity compared to the case with $\lambda = 5$ (0.96 and 0.97 respectively). This decrease was even more pronounced when the recurrence frequency of each CNA was taken into account. In particular, without pseudo-replication or with replication involving two or three samples, CNAnova was not able to detect copy number changes occurring with frequency less than 12%. The ratio of the average F-statistics for those infrequent segments and the threshold parameter was 3.1, 3.3, 3.3 and 3.4 for CNAnova without replication and with $\lambda = 2, 3, 4$ respectively, and 4.4 when $\lambda = 5$. When comparing distributions of the F-statistic from normal and cancer samples for replicated ($\lambda = 5$) and unreplicated data using the Welch two-sample t-test, we observe that the pseudo-replication increases the spread between the two distributions (14.1 versus 13.9 for the two-sample t-test, df = 3119). These results suggest that pseudo-replication increases the power of the method to detect rare recurrent CNAs, substantiating its use during the CNAnova preprocessing phase. In addition, $\lambda = 5$ produces better results compared to lower values of $\lambda$.

We wanted to make the algorithm as flexible as possible regarding the size and the origin of normal samples that it can use, making it applicable to cases when only a small number

of matched normal samples is available. However, we have extended CNAnova to allow it to use matched normal samples when they constitute the majority of cases. CNAnova use the following strategy to derive log-ratios for the cancer dataset. It matches each cancer sample with its paired normal sample when such is available, and selects a random normal sample for matching unpaired cancer samples. The normal dataset is created by matching randomly selected normal samples with all other samples. After this matching all other steps are carried out in exactly the same way as in the standard version of CNAnova. Applying this method to chromosomes 9 to 13 of the ALL dataset identified a similar set of recurrent CNAs although they were composed of a smaller number of samples (data not shown).

### Influence of size and noise level of normal samples on the error rate

Since CNAnova uses the distribution of log-ratios in the normal dataset to control the false positive discovery rate, it is important to check robustness of the method against different patterns of noise that may be present in the normal dataset. We started by studying the influence of the changes in the size of the normal (control) dataset on CNAnova performance. The size of control samples in the simulation dataset was varied from 155 to 20 while holding the number of cancer samples fixed, and the F-statistic threshold was computed for each of the comparisons. CNAnova performs comparably well across all control samples sizes apart from the case when the normal sample constitutes 15% of the dataset (Figure S6a). Simulation of five independent control datasets of each size also suggests that, although variability between different simulations of the F-statistic threshold is greater for smaller sample sizes, it is still adequate for estimation of the false discovery rate (Figure S6b).

In addition to the size, the distribution of log-ratios of normal samples also influences estimation of the F-statistic threshold: the presence of extreme wave artefacts in normal samples might lead to a higher false negative rate, whereas uncharacteristically low signal and amount of noise will increase the number of false positives. We also sought to investigate the influence of QC characteristics in normal samples on error rate. Five different strategies were devised to select 50 normal samples from a total of 180: (1) samples with the highest median absolute deviation (MAD) values (the most noisy); (2) samples with the lowest MAD values; (3) 45% high MAD samples + 5% low MAD samples; (4)

5% high MAD samples + 45% low MAD samples and (5) an equal proportion of noisy and noiseless samples. We found few differences between these selection criteria for normal samples (Figure S6c), indicating that the CNAnova statistical model in combination with different pre-processing steps make the method relatively robust to variations in the variances between cancer and normal datasets. However, the HapMap samples have very good QC characteristics, which may not be the case for other datasets. Consequently, as in the case of other microarray analysis algorithms, CNAnova will still benefit from rigorous pre-processing and normalization pipelines that remove very noisy samples.

Because of normalization and CNAnova's internal post-processing routines, the variation in the mean values of log ratios between normal and cancer datasets has much less significant impact on the algorithm's performance. In particular, zero centering will correct global differences between means during calculation of log-ratios.

## Discussion

### Implications of statistical significance estimation of recurrent CNAs in CNAnova framework

GISTIC permutes the data to estimate the null distribution. It makes an assumption that all aberrations (including driver aberrations) are passenger mutations, thereby generating a conservative, high estimate of the background aberration rate. CNAnova, in contrast, uses the distribution of probe intensities in the control normal dataset (after removing CNVs) to estimate the null distribution. The approach is based on the premise that the control normal dataset has the same platform, dataset, sample and other technical biases that are present in the cancer data. Consequently, it allows estimation of a statistical cut-off that will remove much of the local variation in probe intensities, such as wave artifacts, but will preserve true copy number changes. Such an approach offers greater sensitivity in finding recurrent copy number changes, especially those that have intra-sample frequency less than 10% and which include heterozygous deletions and low-copy number amplifications. However, CNAnova does not assign actual significance to the recurrent copy number changes. The null hypothesis assessed by GISTIC, which assumes that copy number changes in cancer samples are distributed randomly, allows estimation of the false discovery rate. In CNAnova, the

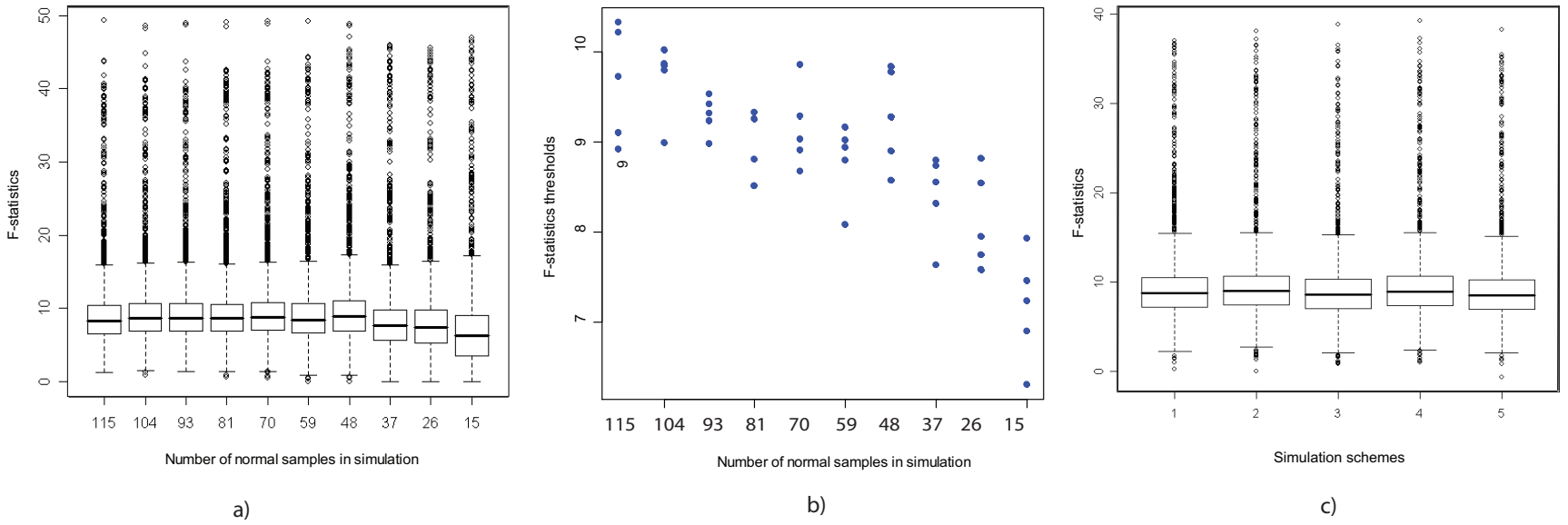


Figure S6: (a) Distribution of normal F-statistics from the simulation of different number of normal samples. (b) F-statistic threshold after comparison with cancer samples for 5 different instances of each simulation. The ideal threshold used in the simulations is the same as in Figure S5. (c) Distribution of normal F-statistics from the simulation of normal samples with different noise characteristics. See text for details of the five simulation schemes.

null hypothesis is that in cancer local changes in probe intensities arise from non-CNA based artefacts. The significance of recurrent copy number changes is therefore assessed by using the percentage of samples exhibiting CNAs. It could be argued that such a representation can provide more informative results for decision making than FDR values. For example, a new CNA present in 5% of 2000 cancer samples could suggest the existence of new subtype-specific genes implicated in cancer progression.

## Supplementary Figures

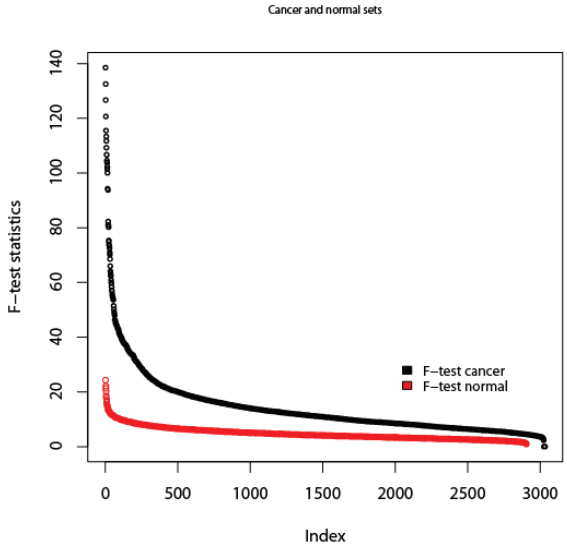


Figure S7: Distributions of F-test statistics from control normal and cancer datasets (sorted in decreasing order). F-test statistics from the normal dataset are used to control False Discovery Rate in CNAnova.

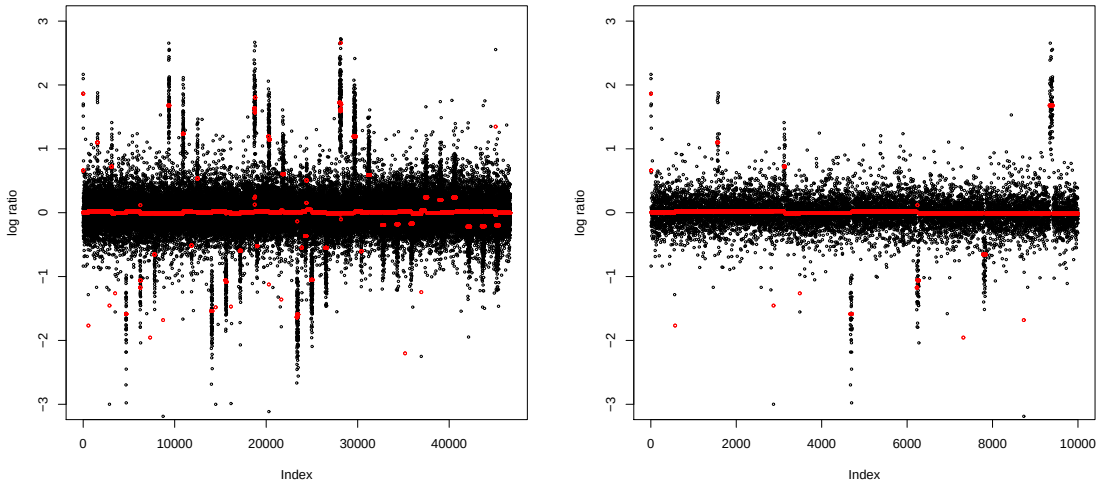


Figure S8: Example of segmentation results after applying circular binary segmentation (CBS) algorithm to the simulated dataset. Whole chromosome (left) and a subset of chromosome (right) are shown, lines in red indicate segment means. CBS is able to identify all CNAs in the sample.

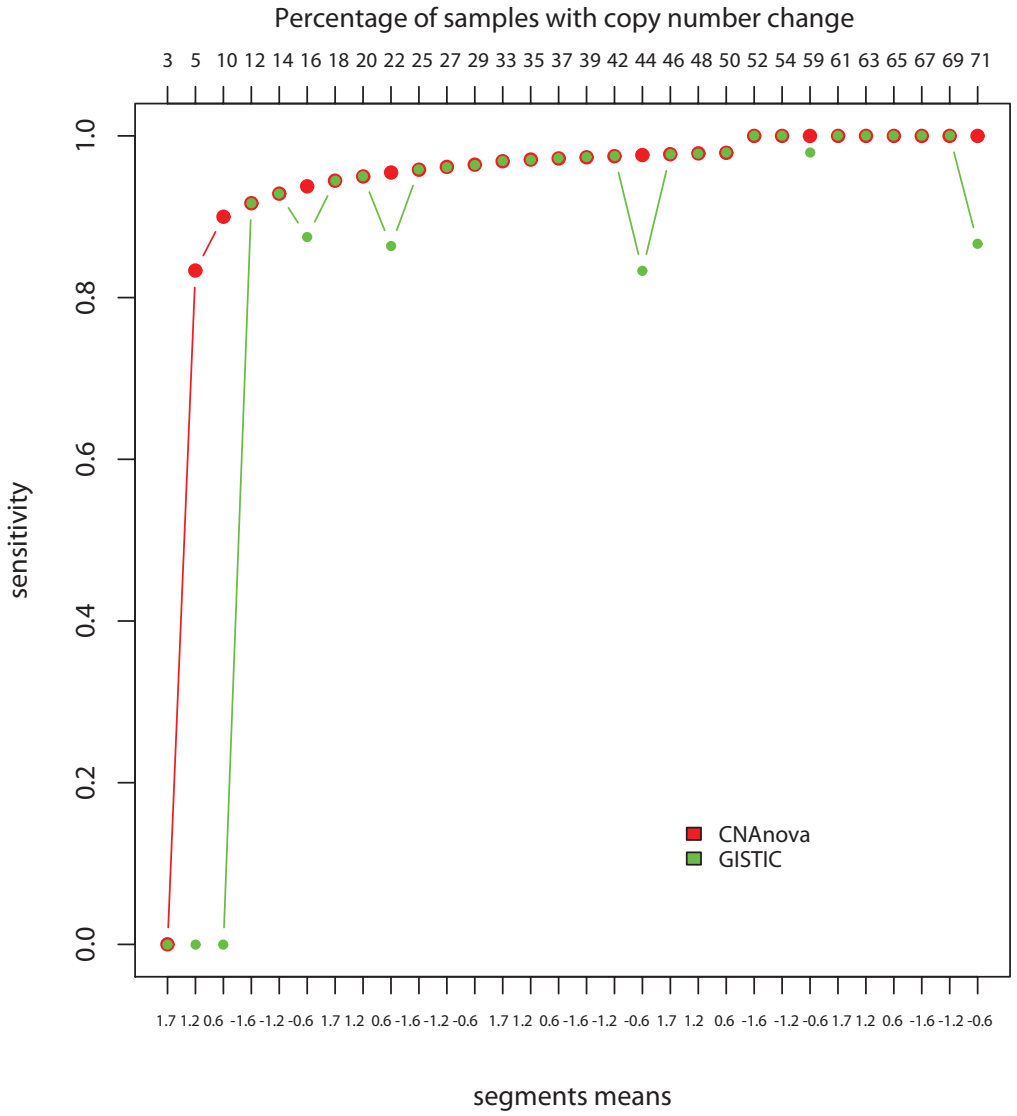


Figure S9: Sensitivity of CNAnova and GISTIC, run with 0.5 threshold, in identifying recurrent copy number changes of different frequency and magnitude. Top x-axis label gives the percentage of samples containing a given recurrent CNA; bottom x-axis label gives the mean log-ratio value used in simulation (colour version of Figure 4 in the main text).

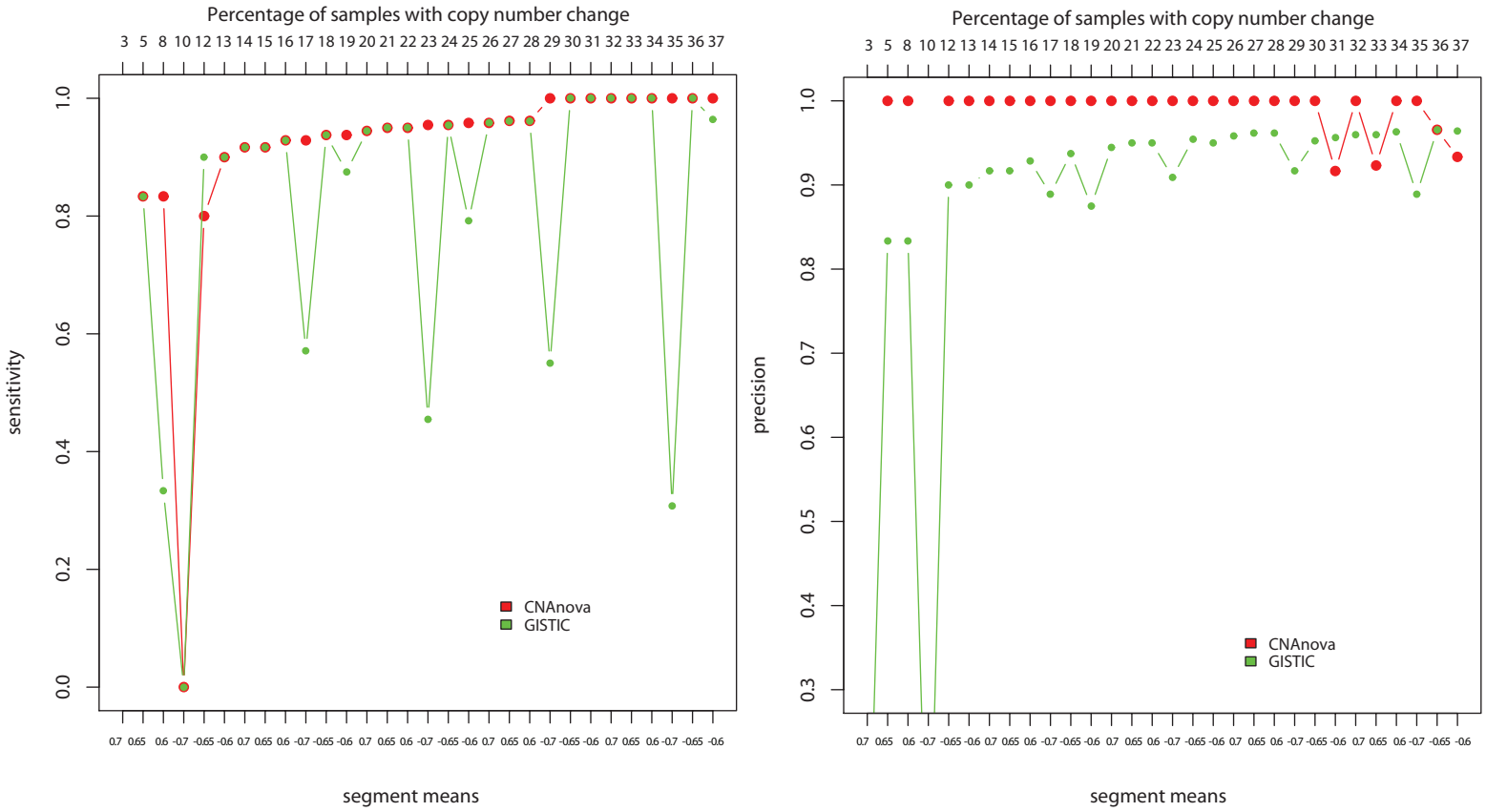


Figure S10: Sensitivity (right) and precision (left) of CNAnova and GISTIC, run with 0.5 threshold, in identifying low-level recurrent number gains and losses. Top x-axis label gives the percentage of samples containing a given recurrent CNA; bottom x-axis label gives the mean log-ratio value used in simulation (colour version of Figure 5 in the main text).

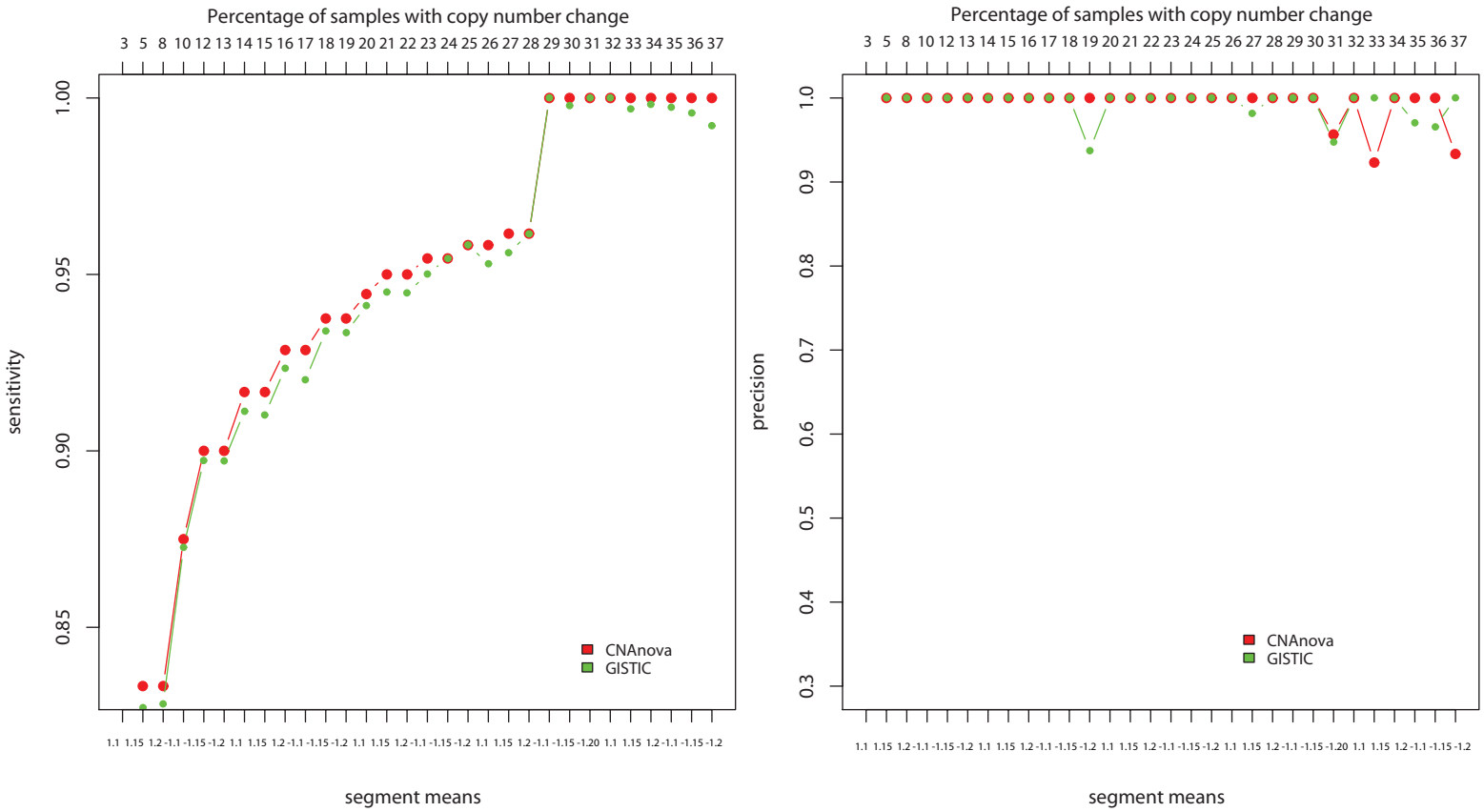


Figure S11: Sensitivity (right) and precision (left) of CNAnova and GISTIC (run with a 0.5 threshold) in identifying rare recurrent two-copy-number gains and deletions. Top x-axis label gives the percentage of samples containing a given recurrent CNA; bottom x-axis label gives the mean log-ratio value used in simulation.

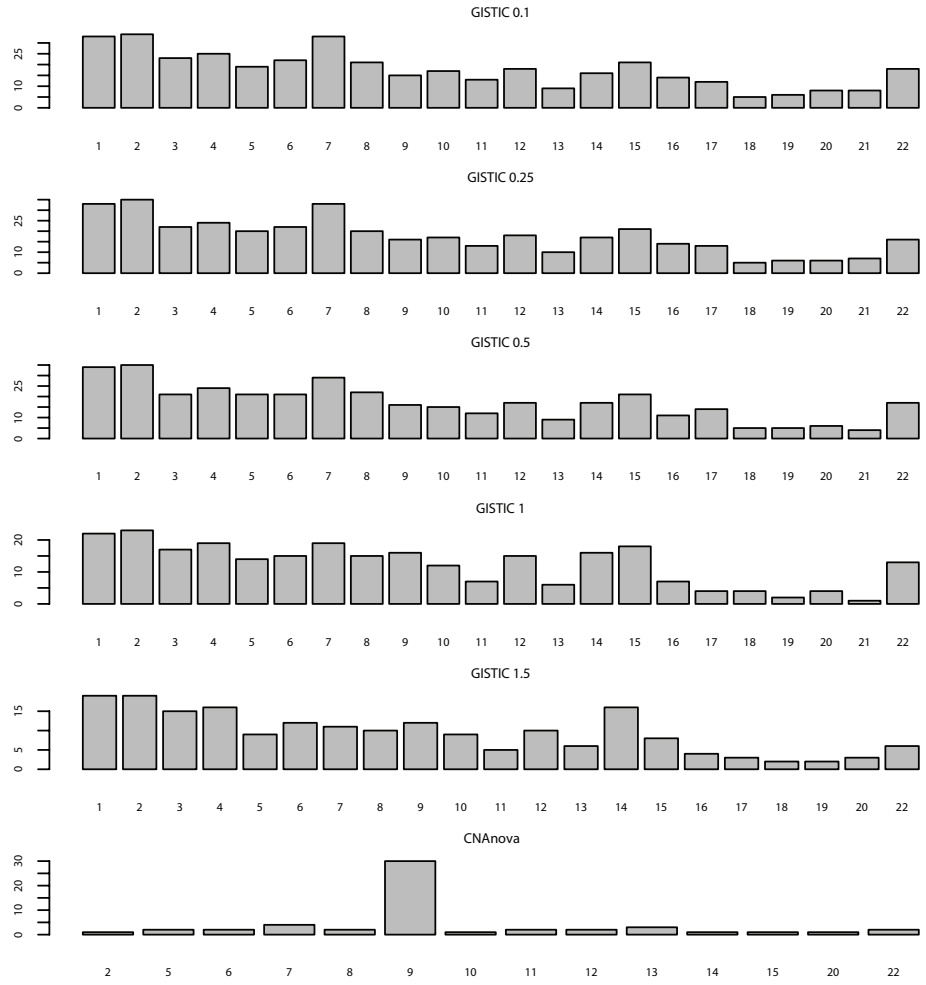


Figure S12: Comparison of the total number of recurrent losses and gains identified by CNAnova and GISTIC in the ALL dataset by chromosome.

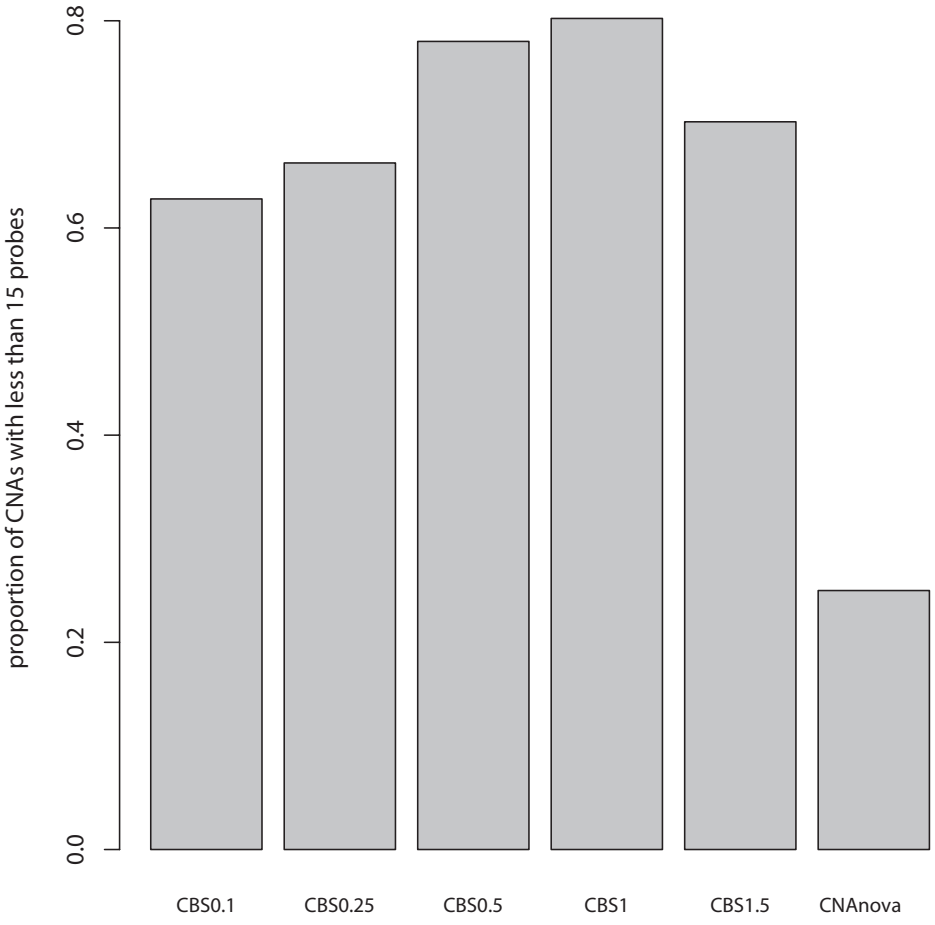


Figure S13: Proportion of recurrent CNA segments spanning less than 15 probes identified by GISTIC (with different thresholds) and CNAnova using CBS-segmented data.

# Bibliography

Ben-Yaacov, E. and Eldar, Y. C. (2008). A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, **24**, 139–145.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, **35**, 2013–2025.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 603–619.

Fukunaga, K. and Hostetler, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, **21**, 32–40.

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., Futreal, P. A., and Stratton, M. R. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S. J., Baldwin, D. A., Stoeckert, C. J., and Grant, G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics*, **3**, 1464–1486.

Hupé, P., Stransky, N., Thiery, J., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B., and Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**, 1166–1174.

Nilsson, B., Johansson, M., Al-Shahrour, F., Carpenter, A. E., and Ebert, B. L. (2009). Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. *Bioinformatics*, **25**(8), 1078–1079.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R. C., Triche, T. J., and Asgharzadeh, S. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, **24**, 309–318.

Scharpf, R. B., Ting, J. C., Pevsner, J., and Ruczinski, I. (2007). SNPchip: R classes and methods for SNP array data. *Bioinformatics*, **23**, 627–628.

Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Loo, P. V., Yu, T., Kristensen, V. N., and Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Research*, **37**(16), 5365–5377.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, **17**, 1665–1674.

Wang, L. Y., Abyzov, A., Korbel, J. O., Snyder, M., and Gerstein, M. (2009). MSB: A mean-shift-based approach for the analysis of structural variation in the genome. *Genome Research*, **19**, 106–117.