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.

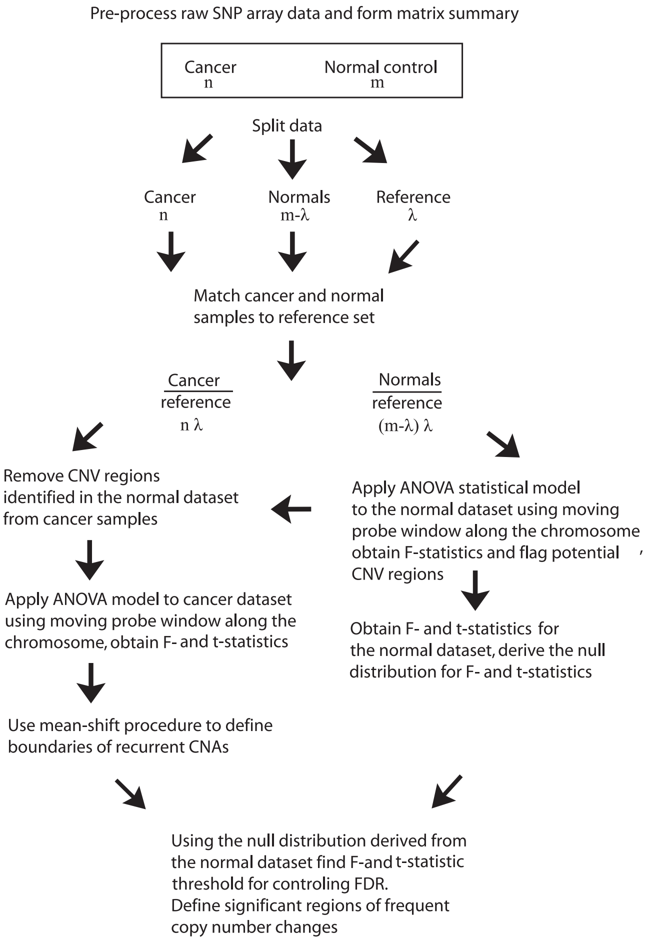
Pre-process raw SNP array data and form matrix summary
Cancer n
Normal control m
Split data
Cancer n
Normals m-λ
Reference λ
Match cancer and normal samples to reference set
Cancer / reference n λ
Normals / reference (m-λ) λ
Remove CNV regions identified in the normal dataset from cancer samples
Apply ANOVA statistical model to the normal dataset using moving probe window along the chromosome obtain F-statistics and flag potential CNV regions
Apply ANOVA model to cancer dataset using moving probe window along the chromosome, obtain F- and t-statistics
Obtain F- and t-statistics for the normal dataset, derive the null distribution for F- and t-statistics
Use mean-shift procedure to define boundaries of recurrent CNAs
Using the null distribution derived from the normal dataset find F-and t-statistic threshold for controling FDR.
Define significant regions of frequent copy number changes

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 λ, 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 λ 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 λ 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 ρ 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/ρ windows.

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

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

where

- n is the number of cancer samples;
- $i=0$ corresponds to the ANOVA reference sample and $i=1,\dots,n$ 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;
- μ 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;
- ε_{ik} are the random error terms such that ε_{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 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 τ_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/ρ and a P-value matrix with dimensions t/ρ and n, which help to identify both the boundaries and sample content of recurrent CNAs.

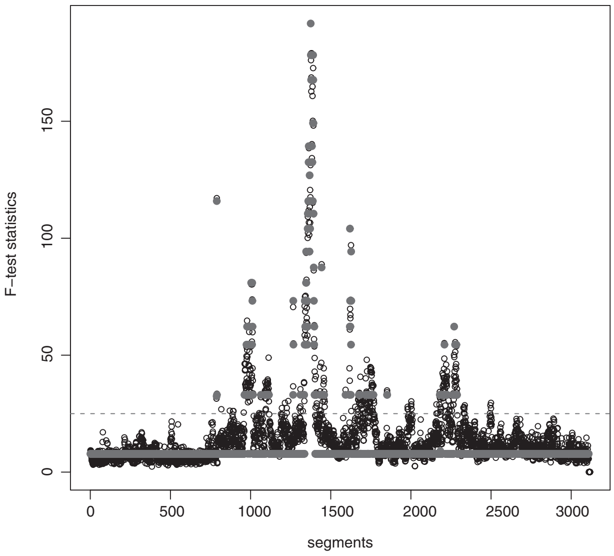

Fig. 2. Application of mean-shift procedure to the distribution of F-statistics along the chromosome. Solid grey lines indicate the location of mean-shift minima and dotted line shows the threshold for calling significant F-test scores. The mean-shift algorithm allows smoothing of the non-significant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

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 σ 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 Φ_r be the set of F-statistics in the reference dataset and Φ_c in the corresponding cancer dataset. For each window, remove that F-statistic from Φ_r if the maximal value of the median log-ratios across all samples in that window is more than the threshold θ (here set to 0.5);

(2) Sort the remaining F-statistics in Φ_r, and in Φ_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:

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

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;
- μ is the mean of the probe log-ratios for the normal reference sample in the window, and $\tau_0=0$;
- τ_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;
- ε_{iuk} are the random error terms such that ε_{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

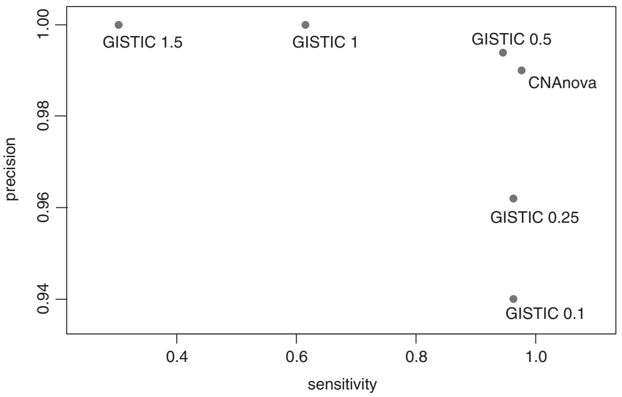

Fig. 3. Sensitivity versus precision curve comparing CNAnova and GISTIC in ability to find recurrent copy number alterations in the simulated data.

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 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).

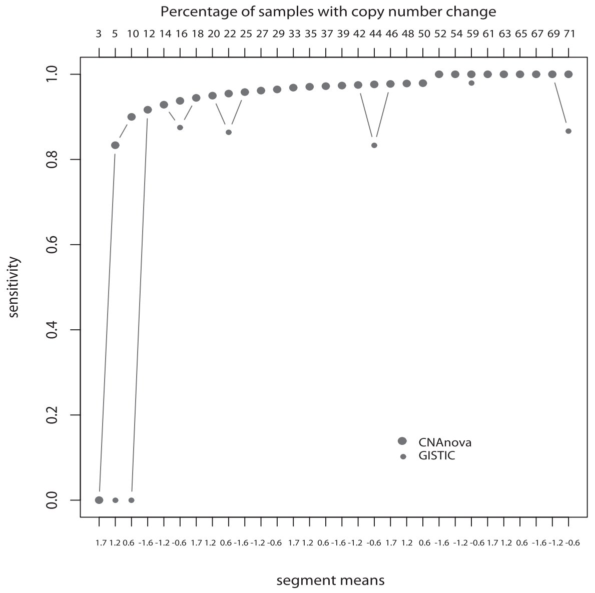

Fig. 4. Sensitivity of CNAnova and GISTIC, run with 0.5 threshold, in identifying recurrent copy number changes of different frequency and magnitude. Top *x*-axis label gives the percentage of samples containing a given recurrent CNA; bottom *x*-axis label gives the mean log-ratio value used in simulation.

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

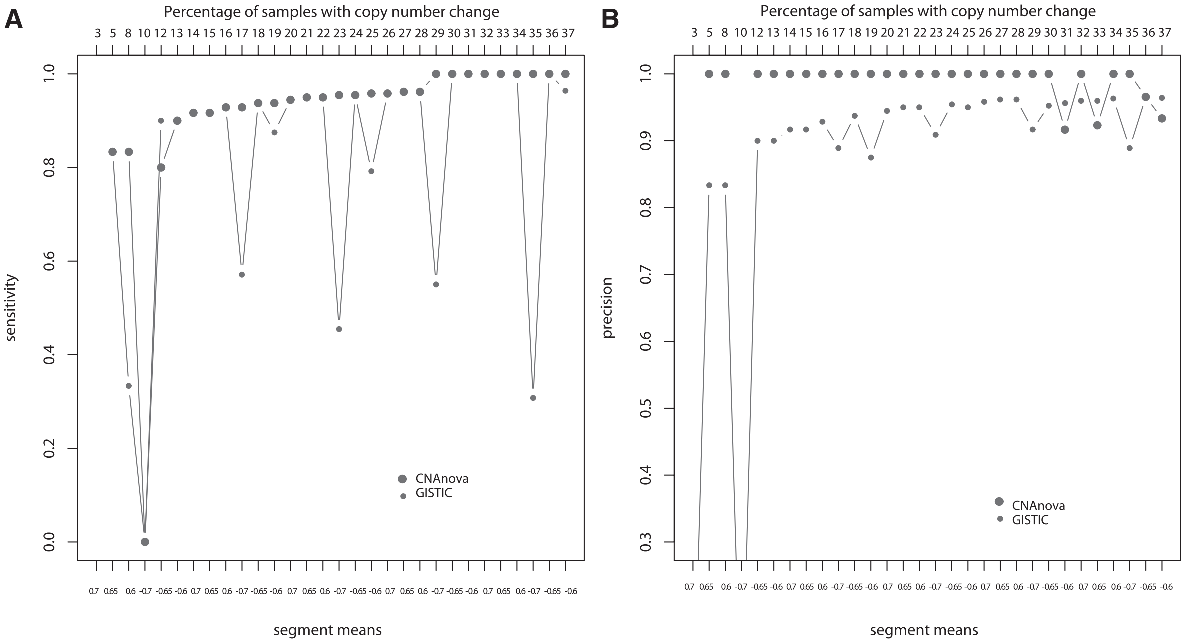

Fig. 5. Sensitivity (**A**) and precision (**B**) of CNAnova and GISTIC, run with the thresholds of 0.5 on log-ratio values, in identifying rare recurrent low-level copy number gains and losses (colour version of the figure is available in Supplementary Material). Top *x*-axis label gives the percentage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

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

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

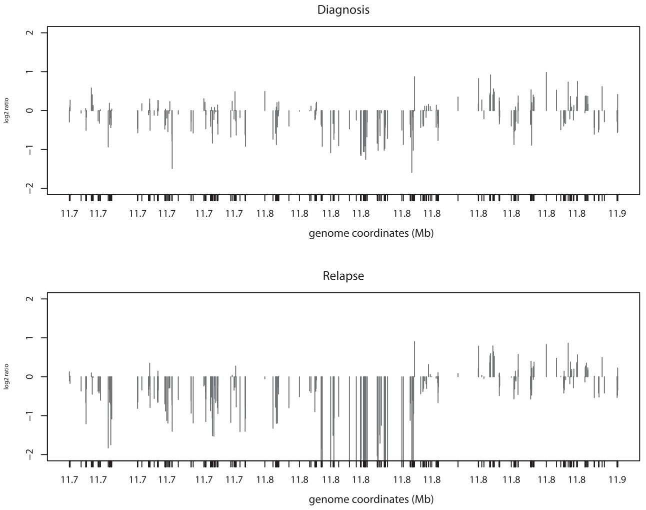

Fig. 6. Example of copy number plots showing the loss of ETV6 in the relapsed ALL samples identified with fixed-factor nested-design CNAnova.

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.