BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 5 2011, pages 678–685
doi:10.1093/bioinformatics/btq717

Genetics and population analysis Advance Access publication December 23, 2010

DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors

Vonn Walter[1,2,*], Andrew B. Nobel[1,2,3] and Fred A. Wright[1,2,*]

[1]Department of Biostatistics, [2]Lineberger Comprehensive Cancer Center and [3]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: DNA copy number gains and losses are commonly found in tumor tissue, and some of these aberrations play a role in tumor genesis and development. Although high resolution DNA copy number data can be obtained using array-based techniques, no single method is widely used to distinguish between recurrent and sporadic copy number aberrations.

Results: Here we introduce Discovering Copy Number Aberrations Manifested In Cancer (DiNAMIC), a novel method for assessing the statistical significance of recurrent copy number aberrations. In contrast to competing procedures, the testing procedure underlying DiNAMIC is carefully motivated, and employs a novel cyclic permutation scheme. Extensive simulation studies show that DiNAMIC controls false positive discoveries in a variety of realistic scenarios. We use DiNAMIC to analyze two publicly available tumor datasets, and our results show that DiNAMIC detects multiple loci that have biological relevance.

Availability: Source code implemented in R, as well as text files containing examples and sample datasets are available at http://www.bios.unc.edu/research/genomic_software/DiNAMIC.

Contact: vwalter@email.unc.edu; fwright@bios.unc.edu

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

Received on September 30, 2010; revised on December 6, 2010; accepted on December 21, 2010

1 INTRODUCTION

DNA copy number aberrations (CNAs) are commonly found in tumor tissue, and can range from losses (deletions) of one or both copies of chromosomal regions to gains of numerous additional copies (amplifications). The size of these aberrations can range from entire chromosome arms to less than 100 kb (Myllykangas and Knuutila, 2006). A variety of platforms are used to detect CNAs, and provide quantitative signals that reflect the underlying discrete copy number (Coe *et al*., 2007; Davies *et al*., 2005; Zhao *et al*., 2004). Much of the statistical effort in analyzing CNAs has focused on discerning copy number at each location within individual tumors (Hupe *et al*., 2004; Olshen *et al*., 2004; Venkatraman and Olshen, 2007), and in handling the potential contamination of normal tissue in tumor samples (Sun *et al*., 2009).

In contrast to heritable copy number variation, CNAs are the result of genomic instability in somatic tumor tissue (Albertson *et al*., 2003). From the earliest days of modern cancer genetics, it was recognized that such instability could unmask or promote the effects of tumor suppressors and oncogenes (Knudsen, 1971; Stratchan and Read, 1999). However, surveys of a number of tumor types (e.g. Miller *et al*., 2003) demonstrate that sporadic gains and losses can also occur throughout the genome, likely representing generic genomic instability, with little effect on tumor progression. The phenomenon of *recurrent* CNAs, which affect the same region in multiple tumors, is of great interest, as such CNAs may highlight genes or regions that are directly involved in tumor progression. Past studies have detected recurrent CNAs in a wide range of tumor types, with an extensive catalog of these findings in the Mitelman Database (Mitelman *et al*., 2010) and the Genetic Alterations in Cancer (GAC) database (Jackson *et al*., 2006).

Despite the apparent successes in the field, there is no clear basis for a general approach for sensitive detection of recurrent CNAs, as many regions important for tumor progression may affect only a minority of tumors. The task of distinguishing between sporadic and recurrent CNAs is thus largely a statistical issue. The instability-selection model introduced by Newton *et al.* (1998) provides a statistical framework specific to loss of heterozygosity (LOH) data, but even for this specific data type difficulties remain in assessing significance over multiple markers (Sterrett and Wright, 2007). The problem of assessing significance for general copy number data has received relatively little attention until recently (Shah, 2008). Few of the existing methods (reviewed below) provide an explicit description of the null hypothesis being tested, or fully acknowledge the inherent correlation structure of copy number data. For these reasons, it has been difficult to place the techniques in a traditional statistical framework or to understand error rates on a genome-wide scale. The purpose of this article is to introduce an explicit testing scheme for recurrent CNAs that preserves correlations inherent to the data.

Before proceeding to our testing framework, we review the current methods for copy number calling/segmentation, which can serve as a useful intermediary to the detection of recurrent CNAs. Numerous technologies are available to measure DNA copy number, ranging from array comparative genomic hybridization (Coe *et al*., 2007; Davies *et al*., 2005) at tens of thousands of probes, to high-density SNP platforms (up to 1 million probes or more, Zhao *et al*., 2004). Reviews of the technologies are provided elsewhere (Davies *et al*., 2005; Zhao *et al*., 2004), but a common feature is that a quantitative signal is extracted at each probe that reflects underlying

*To whom correspondence should be addressed.

copy number, with additional noise and potentially probe-specific bias inherent to the platform.

Regional losses and gains within a single tumor typically cover contiguous sets of numerous probes (Myllykangas and Knuutila, 2006), and so segmentation approaches (Hupe *et al*., 2004; Olshen *et al*., 2004; Venkatraman and Olshen, 2007) are popular as a means to estimate the underlying copy number state at each position per tumor. Here we distinguish between *discrete* segmentation, where the copy number is constrained to the non-negative integers, and *continuous* segmentation, where the segmented values need not be integers. Examples of continuous segmentation include methods that essentially average over quantitative probe values within a genomic region determined by the algorithm to be a copy number segment. Regardless of the segmentation procedure, technical artifacts and differences in probe characteristics can lead to probe-specific bias, potentially reducing the accuracy of segmentation. A number of authors have established the presence of probe bias. Marioni *et al.* (2007) show that aCGH data exhibits serial autocorrelation, a phenomenon they term 'genomic waves', and Komura *et al.* (2006) note a correlation between apparent DNA copy number and GC content. Marioni *et al.* (2007) and van de Wiel *et al.* (2009) present procedures for 'smoothing' genomic waves. Both procedures can be applied to copy number data from normal samples, and the method of van de Wiel *et al.* (2009) can be applied to tumor samples as long as normal samples are present. For sufficient sample sizes, we describe further below an approach to correct the bias by comparing intensities of individual probes using data from surrounding probes (via segmentation), without the need to model or otherwise consider the sequence context.

We also clarify that we are interested in somatic copy number changes in tumors, rather than heritable copy number variants (CNVs). As the resolution of typing technologies increases, it is possible that CNVs, which are rarely larger than 1 Mb (Itsara *et al.*, 2009) and thus considerably shorter than the aberrations found in solid tumors (Albertson *et al.*, 2003), can be mistaken for recurrent CNAs. The distinction can be clarified by comparisons of matched tumor and normal tissue. Researchers using tumor-only datasets should be alert to the possible presence of common copy number polymorphisms when interpreting DiNAMIC output (Redon *et al.*, 2006).

Over a dozen software packages for analyzing DNA copy number data are discussed by Rueda and Diaz-Uriate (2008), Baross *et al.* (2007) and Shah (2008). We focus here on the approaches that attempt to identify recurrent copy number changes, highlighting the input formats and a few relevant similarities and differences.

- STAC (Diskin *et al.*, 2006) and CGHregions (van de Wiel and van Wieringen, 2007) require discrete segmented input data, i.e. categorical values such as aberrant/normal, gain/normal/loss or some numerical equivalent.
- GISTIC (Beroukhim *et al.*, 2007) requires continuous segmented input data, such as one might obtain from a segmentation program such as GLAD (Hupe *et al.*, 2004) or DNAcopy (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007).
- KC-SMART (Klijn *et al.*, 2008) and MSA (Guttman *et al.*, 2007) accept *continuous* input data, such as $\log_2$ intensity ratios, although MSA performs discrete segmentation internally and then makes multiple calls to the STAC algorithm.
- GISTIC, KC-SMART, STAC and MSA assess the statistical significance of the most striking marker or region using permutation-based null distributions, while adjusting for multiple comparisons. However, the resulting output differs among the methods. GISTIC produces false discovery rate (FDR) q-values for the 'significant' regions. STAC and MSA control the family-wise error rate (FWER) by using the max-T procedure of Westfall and Young (1993), while KC-SMART controls FWER by using a Bonferroni adjustment.
- GISTIC and KC-SMART analyze genome-wide data, whereas STAC and MSA analyze data at the level of the chromosome or chromosome arm.

Here we introduce Discovering Copy Number Aberrations Manifested In Cancer (DiNAMIC), a new procedure to map recurrent CNAs and assess their statistical significance. DiNAMIC can be applied in the analysis of data from individual chromosomes or genome wide. The input can consist of segmented data, either discrete or continuous. Alternately, quantitative probe measurements may be used directly, although the reader is advised to read the material on Probe Bias in DNA Copy Number Data in Section 2.4 before analyzing individual probe-level data. DiNAMIC is computationally fast, statistically robust and requires no specialized software. We believe that DiNAMIC is a valuable addition to the methods available to search for recurrent CNAs.

2 METHODS

2.1 Data format and definitions

The data are contained in a numeric $n \times m$ matrix X. Each entry x_{ij} represents DNA copy number (or LOH data) for subject i at marker j. In other words, each row $X_{i\cdot}$ of X corresponds to copy number for one subject at m markers, while each column $X_{\cdot j}$ corresponds to data at a single marker for n subjects. Markers that exhibit high or low average copy number are of interest, so it is natural to examine summary statistics for each marker. We define S_j to be the sum of the entries in the j-th column of X, leading to the local summary statistics $S_1, S_2, \ldots, S_m$.

Copy number gains and losses are analyzed separately, and for either type of analysis we want a global summary statistic that is sensitive to the presence of the corresponding CNA. We will restrict our attention to

$$T_{\text{gain}}(X) = \max(S_1, S_2, \ldots, S_m),$$

when copy number gains are of interest and

$$T_{\text{loss}}(X) = \min(S_1, S_2, \ldots, S_m)$$

when the focus is on copy number losses. For brevity, we restrict all subsequent discussion to $T_{\text{gain}}(X)$ and make comments regarding $T_{\text{loss}}(X)$ when necessary.

2.2 Permutation, cyclic shift and assessing statistical significance

Sporadic variation in DNA copy number often occurs throughout the genome, so it important to determine the statistical significance of recurrent CNAs. This can be done if we have a null distribution for $T_{\text{gain}}(X)$ under the hypothesis that no recurrent CNAs are present, i.e. that all CNAs are sporadic. We prefer to not make assumptions about the entries in X—i.e. they may be discrete segmented, continuous segmented or continuous values. Thus, the estimation of a null distribution for $T_{\text{gain}}(X)$ using permutation is attractive. This approach is also broadly taken by GISTIC, STAC, MSA and KC-SMART with their corresponding statistics.

A variety of permutation schemes are possible. Entries in different rows come from different subjects, and may have different rates of CNV, contamination of normal tissue in tumor samples, etc. Thus, permutation of entries across rows should be avoided. KC-SMART randomly permutes the DNA copy number values within a given row, and thus breaks up the genomic positional relationship of the entries. The null distribution assumed by GISTIC is based on a convolution of histograms. Each histogram is created from the entries in a given row, but the serial structure of the entries is not preserved. In contrast, STAC and MSA perform random rearrangements of 'aberrant' regions within a given row. These permutations maintain the serial structure within aberrant regions, but require a clear and prior distinction between aberrant and non-aberrant regions.

DNA copy number data is inherently correlated, due to the underlying loss or gain of chromosomal segments, even if the CNAs are sporadic. It is desirable to maintain this correlation under permutation in order to maximize the retained information and to provide proper control of error rates. These considerations provide the motivation for DiNAMIC's permutation scheme.

Let $X_{i\cdot} = x_{i1}x_{i2}\ldots x_{im}$ be the i-th row of X, which corresponds to the data from the i-th subject. For $1 \le k \le m$, we define a *cyclic shift* of $X_{i\cdot}$ of index k to be

$$\sigma_k(X_{i\cdot}) = x_{ik}x_{i(k+1)}\ldots x_{im}x_{i1}\ldots x_{i(k-1)}.$$

More generally, a *cyclic shift* $\sigma(X)$ of X is obtained by applying cyclic shifts σ_k to each row of X, where the shift index k can vary from one row to the next. This yields a total of $m^n - 1$ distinct possible cyclic shifts other than the original state. The biological motivation for cyclic shifts is clear for organisms such as bacteria that have circular chromosomes, for the serial structure of the copy number data from a given row is completely preserved under cyclic shifts. Thus, if the observed copy numbers for each row mimic a circular stationary process (Anderson, 1960), the correlation structure is not changed by the cyclic shifts. Human chromosomes are of course not circular, but the cyclic shift $\sigma_k(X_{i\cdot})$ maintains all of the serial structure between the markers, except at the breakpoint $x_{i(k-1)}$. As the number of markers m is much larger than the total number of breakpoints, we conclude that the correlation structure is approximately maintained. Another way of motivating the cyclic shift is to consider that the hallmark of a recurrent CNA is that gains or losses tend to 'line up' at similar positions across multiple tumors. In contrast, the null hypothesis maintains that sporadic aberrations may occur anywhere on the genome, but that no region is 'special'. Thus, the cyclic shift assesses significance by shifting each row of X, so that the rarity of apparently recurrent events (multiple tumors showing aberration at a marker) is easily assessed. These assumptions are directly tested in simulations described further below.

We now describe our method for assessing the statistical significance of $T_{\text{gain}}(X) = \max(S_1, S_2, \ldots, S_m)$.

Algorithm 1. Assessing Statistical Significance of $T_{\text{gain}}(X)$

(1) (Optional) Set random seed r,

(2) Perform N random cyclic shifts $\sigma^1(X), \sigma^2(X), \ldots, \sigma^N(X)$ of the data matrix X,

(3) Compute the value of the summary statistic $T_{\text{gain}}(\sigma^l(X))$ for each shifted dataset $l = 1, 2, \ldots, N$,

(4) Define the quantile-based P-value

$$p(T_{\text{gain}}(X)) = \min\left(\frac{1 + f_{\text{gain}}(N, X)}{N}, 1\right),$$

where $f_{\text{gain}}(N, X) = \sum_{l=1}^{N} I(T_{\text{gain}}(\sigma^l(X)) \ge T_{\text{gain}}(X))$ and $I(.)$ is the indicator function.

The empirical P-value of $T_{\text{loss}}(X)$ is computed by replacing 'gain' with 'loss' and reversing the inequality in Step 4. Both definitions yield P-values that are easy to interpret and are automatically adjusted for multiple comparisons across the markers.

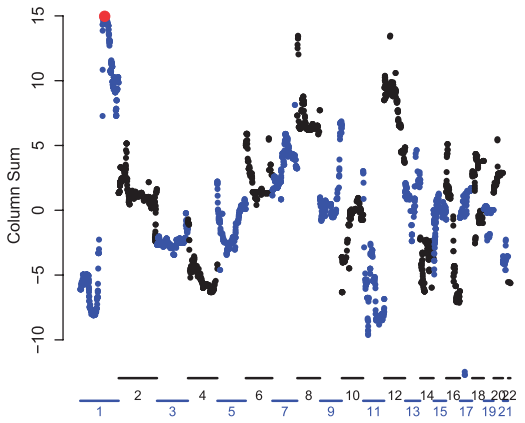

Fig. 1. A plot of the column sums of the Wilms' tumor data of Natrajan *et al.* (2006). The point corresponding to probe RP11-393K10, which corresponds to $T_{\text{gain}}(X)$, is shown in red. Using $N = 1000$ cyclic shifts, we obtain $p(T_{\text{gain}}(X)) = 0.001$.

Natrajan *et al.* (2006) obtained genome-wide aCGH copy number data using 3288 probes on 97 Wilms' tumor samples. By analyzing this data in conjunction with data on tumor relapse, these authors concluded that copy number gains in chr1q are associated with increased risk of tumor relapse. The original data consists of $\log_2$ quantitative copy number values, for which we perform segmentation and bias correction (see the material on Probe Bias in DNA Copy Number Data in the Section 2.4). The column sums of the resultant X matrix are plotted in Figure 1. Marker 196 on chr1q with probe name RP11-393K10 corresponds to the location of the maximum column sum of X, and applying Algorithm 1 with a random seed $r = 12345$ and $N = 1000$ cyclic shifts yields $p_{\text{gain}}(T(X)) = 0.001$.

2.3 Peeling

Figure 1 shows that a number of markers have large column sums, both in the region surrounding probe RP11-393K10 and elsewhere in the genome. In fact, Natrajan *et al.* (2006) detected frequent gains on chromosomes 8 and 12. If multiple highly significant recurrent gains are present in the genome, they presumably contribute high copy number values to the overall null distribution, which may reduce power for detection of less-extreme loci. We use these observations as motivation to extend DiNAMIC so that the significance of multiple regions can be assessed in a straightforward manner. Our approach is similar to the one employed by the GISTIC procedure of Beroukhim *et al.* (2007). GISTIC assesses the significance of a 'new' region conditional on having found the previously most significant region(s) using a procedure the authors term the 'peel off' algorithm. Similarly, we successively correct for each significant region in order to better detect and dissect the recurrent CNAs. However, our peeling method is tailored to the cyclic shift procedure used by DiNAMIC.

For a given data matrix X, our peeling procedure has three components. When analyzing copy number gains we start by identifying the marker k that yields the maximum column sum. Then we find all of the entries in X that contribute to the significance of marker k (i.e. are above their row mean). Algorithm 2, given below, outlines our method for identifying the appropriate entries of X. Next, we multiply these entries of X by a scaling factor τ to create a new data matrix $\hat{X}$ in which the effect of marker k has been removed. We describe the computation of τ and the creation of $\hat{X}$ in the second part of Algorithm 2. Then $\hat{X}$ is subjected to the cyclic shift procedure, with a null distribution conditional on having found marker k in X.

Below we describe the peeling procedure in detail. For convenience, we restrict our attention to copy number gains at marker k, the marker corresponding to the maximum column sum. We write $\bar{x}_{i\cdot}$ and $\bar{x}_{\cdot j}$ for the

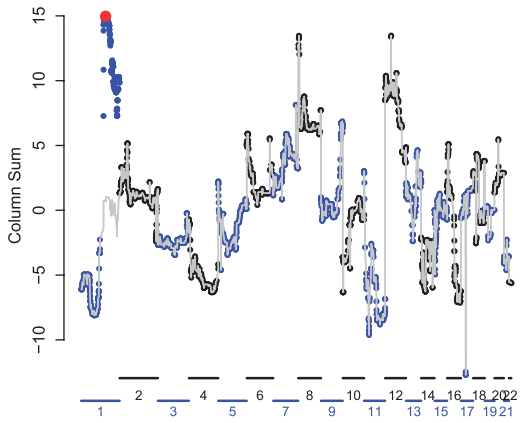

Fig. 2. A plot of the column sums of the Wilms' tumor data of Natrajan *et al.* (2006). The gray lines show the plot of the column sums after applying the peeling procedure to probe RP11-393K10.

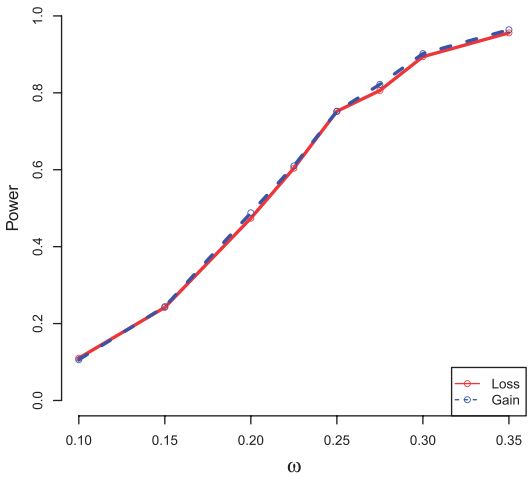

Fig. 3. Power curves for DiNAMIC for simulated datasets containing a single recurrent copy number aberration.

means of the i-th row and j-th column of X, respectively, and $\bar{x}_{..}$ for the mean of X.

Algorithm 2. The Peeling Procedure for Copy Number Gains at Marker k

(1) Find the largest interval $[a,b]$ containing k such that the column means $\bar{x}_{\cdot j} > \bar{x}_{..}$ for all $j \in [a,b]$.

(2) If necessary, reduce $[a,b]$ so that the interval contains only markers from the same chromosome arm as marker k.

(3) Let $I = \{i : x_{ik} > \bar{x}_{i\cdot}\}$ be the set of rows such that the entry x_{ik} exceeds the mean of the i-th row.

(4) For each $i \in I$, find the maximal interval $[a_i,b_i]$ such that (i) $[a_i,b_i] \subseteq [a,b]$, (ii) $k \in [a_i,b_i]$, (iii) $x_{ij} > \bar{x}_{i\cdot}$ for all $j \in [a_i,b_i]$.

We say that $\{x_{ij} : i \in I, j \in [a_i,b_i]\}$ is the set of all matrix entries that contribute to the significance of marker k. Supplementary Figure 1 in the Supplementary Material illustrates the entries of a simulated data matrix X identified by Steps (1)–(4) of the peeling procedure.

We now show how to compute the scaling factor τ and the new data matrix $\hat{X}$.

Algorithm 2 Continued.

(5) Find a constant τ such that $n\bar{x}_{..} =$

$$\sum_{i:\, x_{ik} \le \bar{x}_{i\cdot}} x_{ik} + \sum_{i:\, x_{ik} > \bar{x}_{i\cdot}} \bar{x}_{i\cdot} + \sum_{i:\, x_{ik} > \bar{x}_{i\cdot}} \tau(x_{ik} - \bar{x}_{i\cdot}).$$

(6) Define

$$\hat{x}_{ij} = \begin{cases} \tau x_{ij} & \text{if } i \in I \text{ and } j \in [a_i,b_i]; \\ x_{ij} & \text{otherwise.} \end{cases}$$

(7) Let $\hat{X}$ be an $n \times m$ matrix whose entries are $\hat{x}_{ij}$.

Recall that marker k corresponds to the maximum column sum in X. By construction, the sum of the k-th column of $\hat{X}$ is $n\bar{x}_{..}$, the mean of the column sums in X. Thus, applying the peeling procedure yields a new dataset $\hat{X}$ in which marker k is null. The same constant τ is used to rescale all matrix entries x_{ij} that contribute to the aberration at marker k, so in $\hat{X}$ we expect markers near k to have column sums close to $n\bar{x}_{..}$, and thus also be null. Figure 2 shows a plot of the column sums for the Wilms' tumor data of Natrajan *et al.* (2006) before peeling (black and blue, as in Fig. 1) and after peeling (grey). After peeling, column 196 is no longer significant.

The peeling procedure immediately enables the researcher to filter out minor variations in the vicinity of major peaks, and to potentially distinguish among multiple major peaks. Without any further computation (other than the trivial demands of the peeling procedure itself), the peeling procedure may be applied sequentially while comparing all peaks to the original $T_{\text{gain}}(\sigma(\hat{X}))$ for significance testing. We note that the number of iterations is an input parameter, and thus the user may choose an appropriate balance between sequential significance of CNAs and computational cost. Such an approach is conservative, as the original X contains the extreme values of true recurrent CNAs, but the resulting P-values are corrected for multiple comparisons. A more powerful but computationally demanding approach is to repeat the cyclic shift assessment of statistical significance using the post-peeled matrix $\hat{X}$. Accordingly, DiNAMIC provides two options, *Quick Look* and *Detailed Look*, and the flow chart in Supplementary Figure 2 illustrates the differences between the two procedures. In Quick Look, the original distribution of $T_{\text{gain}}(X)$ is used for significance testing of the most extreme markers, whereas in Detailed Look the null distribution of $T_{\text{gain}}(X)$ is recomputed after each peeling. It is natural to wonder if additional power to detect aberrant markers can be gained by recomputing the null distribution of $T_{\text{gain}}(X)$ after each peeling, and Supplementary Figure 3 indicates that this is the case. A comparison of computation times for Quick Look and Detailed Look can be found in the Supplementary Materials.

2.4 Probe bias in DNA copy number data

As noted in the Section 1, probe-specific variations in hybridization affinity can lead to corresponding variations in array intensity. These in turn can result in biased estimates of DNA copy number. To get some sense of the potential magnitude of the bias, suppose Z is the chromosome 2 data from the glioma dataset of Kotliarov *et al.* (2006), and let $\text{Seg}(Z)$ be a continuous segmented version of Z obtained using DNA copy. Segmentation algorithms use the existing data to model the true underlying copy number as a piece-wise constant function, so the expected column mean of $\text{Resid}(Z) = Z - \text{Seg}(Z)$ should be zero, with variation that should reflect random error. This assumption can be tested for each marker using a t-test for the mean of the entries of each column of $\text{Resid}(Z)$ differing from zero. The histograms in Supplementary Figure 4 of the Supplementary Material show that the t-statistics are markedly overdispersed, which provides clear evidence that probe bias is widespread.

Probe bias can lead to matrices with statistically significant column sums, even in the absence of recurrent CNAs. Failure to correct for it can result in increased type I error. Nevertheless, to the best of our knowledge none of the currently available methods for analyzing DNA copy number data have addressed this issue. One possible method of obtaining a bias-corrected version of the data is to perform continuous segmentation as a preprocessing

step, and then to analyze the segmented data. [GISTIC takes this approach, although Beroukhim *et al.* (2007) make no mention of probe bias.] In the Supplementary Material, we discuss simulations that show that probe bias may still affect segmented data.

We recommend the following bias-correction procedure when working with data matrices X that contain quantitative probe-level data:

Algorithm 3. Removing Probe Bias

(1) Use a segmentation algorithm to get $\text{Seg}(X)$, a segmented version of X.

(2) Compute $\text{Resid}(X) = X - \text{Seg}(X)$.

(3) Let $\mathbf{b}$ be a $1 \times m$ (row) vector whose j-th entry is the mean of the entries in the j-th column of $\text{Resid}(X)$.

(4) Let B be an $n \times m$ matrix with each row equal to $\mathbf{b}$.

(5) Define $\tilde{X} = X - B$.

(6) Use a segmentation algorithm to obtain $\text{Seg}(\tilde{X})$, a segmented version of $\tilde{X}$.

The vector $\mathbf{b}$ is an estimate of the probe bias in X, and this estimated bias is removed when we compute $\tilde{X} = X - B$. However, it is not appropriate to use DiNAMIC to analyze $\tilde{X}$ directly. To see why this is the case, we first note that the column sums of $\tilde{X}$ are identical to the column sums of $\text{Seg}(X)$. On the other hand, the entries of the two matrices do not have the same form, because the rows of $\text{Seg}(X)$ are piece-wise constant, whereas the rows of $\tilde{X}$ are not. As a result, the entries in $\tilde{X}$ are likely to be more variable than those of $\text{Seg}(X)$, and clearly the same statement can be made when comparing the entries of $\sigma(\tilde{X})$ and $\sigma(\text{Seg}(X))$. This observation becomes important when we examine the column sums of $\sigma(\tilde{X})$ and $\sigma(\text{Seg}(X))$, which need not be the same, even if the same cyclic shift σ is applied to both matrices. Since the entries of $\sigma(\tilde{X})$ tend to be more variable than those of $\sigma(\text{Seg}(X))$, the column sums of $\sigma(\tilde{X})$ tend to be more variable than those of $\sigma(\text{Seg}(X))$ as well. However, the column sums of $\sigma(\tilde{X})$ and $\sigma(\text{Seg}(X))$ should have approximately the same mean value. As a result, $T_{\text{gain}}(\sigma(\tilde{X}))$ is likely to assume larger values than $T_{\text{gain}}(\sigma(\text{Seg}(X))$. Therefore, we expect to observe conservative behavior if we use the empirical null distribution $\{T_{\text{gain}}(\sigma^l(\tilde{X}))\}_{l=1}^N$ to assess the significance of $T_{\text{gain}}(\text{Seg}(X))$. Since $T_{\text{gain}}(\tilde{X}) = T_{\text{gain}}(\text{Seg}(X))$, it follows that the same conclusion holds when we use cyclic shifts to assess the statistical significance of $T_{\text{gain}}(\tilde{X})$, which is why it is not appropriate to use DiNAMIC to analyze $\tilde{X}$. Simulation studies of the effectiveness of the peeling procedure are discussed in the Supplementary Material.

3 IMPLEMENTATION

A number of simulated null datasets were created and subsequently analyzed with DiNAMIC in order to study its behavior under the null hypothesis that no recurrent CNAs are present, and the results of these analyses are presented in Table 1. Various marker spacing and correlation schemes were considered in an effort to show that DiNAMIC is robust to the type of deviation from stationarity that can be found in real datasets. A full description of the simulated datasets appears in the Null Simulation Studies section of the Supplementary material. In each case, the observed type I error was computed as follows.

(1) Create a data matrix X^l using the appropriate simulation scheme.

(2) Compute $\hat{p}(T_{\text{gain}}(X^l))$ using $N = 1000$ cyclic shifts of X^l.

(3) Determine whether $T_{\text{gain}}(X^l)$ is significant at the $\alpha = 0.05$ level.

Table 1. Observed type I error for datasets simulated under the null hypothesis

Null simulation model	Type I error
Copy number data	0.0424
Segmented copy number data	0.0429
Serially correlated normal	0.0466
Clumped copy number data (25%)	0.0473
Clumped copy number data (50%)	0.0450
Clumped copy number data (75%)	0.0456
Clumped copy number data (100%)	0.0410

Steps (1)–(3) were repeated 10 000 times, and the observed type I error was defined to be the proportion of $T_{\text{gain}}(X^l)$ that was significant at the $\alpha = 0.05$ level.

The values of the observed type I error given in Table 1 suggest that DiNAMIC is slightly conservative, which seems reasonable in light of the effect of the cyclic shift procedure on the underlying correlation of the markers. Markers on either side of a breakpoint will be essentially independent, and hence they are more likely to exhibit greater variability than neighboring markers in the original data. As a result, the distribution of the maximum column sum after cyclic shift should yield larger values than the corresponding distribution for the original data, and similarly for the minimum column sums. Because the values in Table 1 are quite close to 0.05, any difference in the distributions appears to be very minor.

Additional simulations were performed under the alternative hypothesis that a recurrent CNA is present, and a detailed discussion of these simulations can be found in the Power Simulations and Peeling Accuracy section of the Supplementary Material. Briefly, we note that these simulations show that DiNAMIC has equal power to detect gains and losses. Moreover, DiNAMIC's power to detect CNAs increases with the effect size of the aberration. Both of these properties are illustrated by the power curves in Figure 3.

Next we present the results of the analysis of two publicly available tumor datasets. The dataset of Natrajan *et al.* (2006) contains a number of copy number gain and loss loci that are potentially statistically significant. Using both GISTIC and DiNAMIC's Detailed Look, we analyzed a segmented version of this dataset after applying the bias correction scheme described in the Section 2. Because no normal tissue reference set was available, the thresholds for amplification and deletion, which are required input parameters for GISTIC, were set to the default values of ± 0.1. Table 2 shows all markers that were peeled by DiNAMIC and have either $p(T_{\text{gain}}(X)) < 0.025$ or $p(T_{\text{loss}}(X)) < 0.025$ (marked by 'X'), thereby controlling the overall genome-wide false positive rate (FWER) at $\alpha = 0.05$. For comparison, we also show all regions detected by GISTIC. By default, GISTIC uses an FDR threshold of $q = 0.25$, and in order to facilitate comparison with DiNAMIC, we distinguish between GISTIC findings with $q < 0.05$ (marked by 'X') from those with $0.05 \leq q < 0.25$ (marked by 'O'). Note that there are fewer regions declared significant by GISTIC than by DiNAMIC at the respective 0.05 level. For a given error threshold, the FWER is more conservative than the FDR, so this comparison is meaningful.

Natrajan *et al.* (2006) noted that the most common copy number gains were found in 1q, 8 and 12, with focal gains located at 1q22-25, 8p21-12 and 12p13. Both DiNAMIC and GISTIC detected markers

Table 2. Markers in the Glioma dataset of Natrajan *et al.* (2006) discovered by DiNAMIC's detailed look and GISTIC

Gain marker	DiNAMIC	GISTIC	Loss marker	DiNAMIC	GISTIC
1q23	X	X	1p36		X
2p16	X	X	1p31	X	
2q32	X		2p14	X	
2q37	X		2q37	X	
5p15	X		3p21	X	
6p25		X	3q13	X	
6p24	X		4p15	X	
6q24	X		4q22		X
7p11	X		4q31	X	
7q21		X	5p15	X	
7q34	X		5q11	X	
8p23	X	X	6p12	X	
8q24	X		6q12	X	
9q34	X	X	9p24		O
11p15	X	X	9p21	X	
12p13	X	X	9q21	X	
12q12	X		10p15	X	X
13q12	X		10q11	X	
13q31	X		11p15	X	
13q32		O	11p13	X	X
15q11		O	11q23	X	
16p13	X	X	11q24	X	X
17q25	X		13q21	X	X
18p11	X		14q12	X	
18q11	X		14q21	X	X
18q12		X	15q12	X	X
20p11	X		16p11	X	
20q13	X		16q21		X
			16q23	X	
			17p13	X	
			17q12	X	
			17q21	X	
			18q11	X	
			18q21	X	X
			19p12	X	
			19q12	X	
			21q21	X	X
			22q12	X	
			22q13		X

Markers denoted with an 'X' are significant at the 0.05 level (DiNAMIC genome-wide *P*-value and GISTIC *q*-value). For GISTIC, markers with $0.05 \leq q < 0.25$ are denoted by 'O.'

corresponding to these gains. DiNAMIC and GISTIC detected markers at 9q34, the site of the *SET* oncogene, which is supported by SET protein amplification findings by Carlson *et al*. (1998) in Wilms' tumor. Natrajan *et al.* (2006) also found that gains at 13q31 and 16p13 were associated with tumor relapse. Both methods detected 16p13. DiNAMIC's *P*-value for the locus in 13q31 is significant at the 0.05 level, whereas GISTIC's *q*-value for the locus in the neighboring cytoband 13q32 is not. DiNAMIC's detection of 7q34 and 8q24 is noteworthy because the oncogenes *BRAF* and *c-Myc* lie in these regions, respectively. Neither of these regions were detected by GISTIC.

Losses at 10p15 and 11p13 were found by Natrajan *et al.* (2006) in a number of subjects; these are the sites of *WT1* and *WT2*, genes known to be associated with Wilms' tumor. Both loci were detected

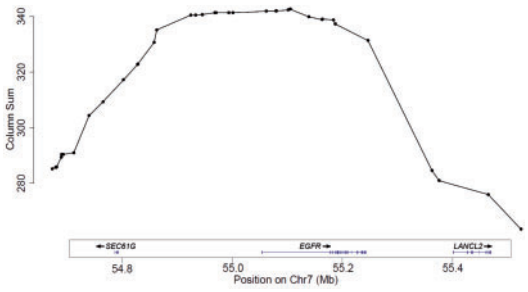

Fig. 4. A plot of the column sums of the glioma dataset of Kotliarov *et al.* (2006) near $T_{\text{gain}}(X)$. This figures was constructed using the UCSC Genome Browser (Kent *et al*., 2002) and LocusZoom (Pruim *et al*., 2010).

by DiNAMIC and GISTIC. The same authors concluded that loss of 21q22 was associated with tumor relapse; both methods detected the nearby locus 21q21. Although the loss sites found by the two methods on 1p, 11q and 16q are not identical, the differences appear to be minor. Using linkage analysis, Rahman *et al.* (1996) discovered *FWT1/WT4*, a familial Wilms' tumor gene located on 17q12. This site was detected by DiNAMIC but not GISTIC. The gene *PDCD6* is located on 5p15, a site that was found by DiNAMIC but not GISTIC. Because *PDCD6* is known to be associated with programmed cell death, detection of this locus may have biological relevance.

GISTIC and DiNAMIC's Detailed Look were also used to analyze the glioma dataset of Kotliarov *et al.* (2006). This dataset contains copy number values from 178 tumors, 82 of which are glioblastomas. As above, GISTIC's amplification and deletion thresholds were set to the default values of ± 0.1; the *q*-value threshold was 0.05. With these settings, GISTIC found 47 significant gain regions and 20 significant loss regions. Using DiNAMIC, over 100 loci for gains and losses were found to be significant at the $\alpha = 0.05$ level. The maximum column sum yielded the most aberrant marker, which is marker 55489 in chr7. Figure 4 shows the column sums near the marker, as well as nearby RefSeq genes (hg18 genomic annotation tracks). The highest peak includes *EGFR* and a region upstream. *EGFR* amplification is a very common genetic mutation in glioblastoma (Heimberger *et al.*, 2005), and the peak finding is a reassuring illustration of the DiNAMIC procedure.

4 DISCUSSION

The analysis of DNA copy number data has proven to be a valuable tool for the study of cancer. Segmentation methods can be used to detect loci where copy number changes occur for a single subject, but different approaches are needed if we wish to assess the statistical significance of DNA copy number changes present in multiple subjects. Here we have introduced DiNAMIC, a new permutation-based method that can be used by researchers to detect statistically significant recurrent CNAs.

When compared to existing methods, DiNAMIC has a number of advantages. First, since DiNAMIC makes no distributional assumptions, it can analyze a variety of input data—continuous, continuous segmented or discrete segmented. In addition, DiNAMIC can analyze genome-wide data, as well as data from a single chromosome or chromosome arm. Finally, it does not require any tuning parameters, such as user-defined thresholds for gain or loss, that are potentially arbitrary.

DiNAMIC is similar to GISTIC, STAC, MSA and KC-SMART in that it assesses statistical significance using a permutation-based null distribution. However, in contrast to the other procedures, the cyclic shifts employed by DiNAMIC preserve essentially the entire serial structure in the data. The serial marker correlation can be very high, especially for segmented data. For example, the average successive marker correlation of the segmented glioma data of Kotliarov *et al.* (2006) was 0.985. Thus, DiNAMIC can preserve the overall false positive rate, without the need to resort to overly conservative procedures, which are sometimes employed by other methods. For example, GISTIC provides multiple comparison control across markers via Benjamini–Hochberg FDR control, which is generally conservative under positive dependence structures (Benjamini and Yekutieli, 2001). KC-SMART uses the Bonferroni method to control for multiple testing, but this is also known to be conservative (Simes, 1986).

Based on our analysis of real datasets, it appears that DiNAMIC performs well when compared to currently available methods. In addition, extensive simulation studies are described in Section 3, and these are performed under a variety of marker correlation schemes. The results of these simulations show that DiNAMIC preserves type I error, and is slightly conservative, even when the markers follow non-stationary correlation structures similar to those found in real data. Probe bias is potentially problematic, however. Even under the null hypothesis that no CNAs are present, probe bias can lead some columns to be more likely to attain the minimum or maximum column sum. We propose a bias correction procedure in Section 2, and use simulations to illustrate its effectiveness, as discussed in the Supplementary Material. Presumably, other methods for detecting recurrent CNAs are also susceptible to probe bias, but we have not seen this issue discussed elsewhere.

DiNAMIC currently uses the columns sums S_j to assess the local evidence for excess copy number gains and losses, and global statistics $T_{\text{gain}}(X)$ and $T_{\text{loss}}(X)$ for global testing. These statistics are intuitive and easy to describe, but may potentially ignore additional information, such as the simultaneous evidence provided by multiple markers in a region. In addition, we do not explore the true dissection of multiple regions, which may exhibit correlated gain or loss structure across different tumors. Our intent in developing DiNAMIC is to create a statistically sound testing structure, which has immediate utility and can serve as the basis for further extensions.

We have also performed simulation studies (data not shown) that indicate that DiNAMIC may also be applied to LOH data, provided the data have few missing values. However, in standard SNP-based LOH calling, the presence of heterozygosity produces a larger number of missing values, with missingness rates that vary by marker. Thus, we leave the application of DiNAMIC to LOH data as an extension for future research.

ACKNOWLEDGEMENTS

We wish to thank Dr. Rameen Beroukhim and members of the Broad Institute for providing helpful comments regarding GISTIC.

Funding: This work supported by an award from the National Institutes of Health [grant number P01C142538 to F.A.W.]; a University Cancer Research Fund Award and a Gillings Innovation Award in Statistical Genomics, both from the University of North Carolina at Chapel Hill [to F.A.W]; and a grant from the National Science Foundation [DMS 0907177 to A.B.N.].

Conflict of Interest: none declared.

REFERENCES

Albertson,D.G. *et al.* (2003) Chromosome aberrations in cancer. *Nat. Genet.*, **34**, 369–376.

Anderson,T.W. (1960) Some stochastic process methods for intelligence test scores. In Arrow,K.J. *et al.* (eds), *Mathematical Methods in the Social Sciences, 1959: Proceedings from the First Stanford Symposium*. Stanford Mathematical Studies in the Social Sciences: IV. Stanford University Press, Stanford, CA, pp. 205–220.

Baross,A. *et al.* (2007) Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics*, **8**, 368.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

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.

Carlson,S. *et al.* (1998) Expression of SET, an inhibitor of protein phosphatase 2A, in renal development and Wilms tumor. *J. Am. Soc. Nephrol.*, **9**, 1873–1880.

Coe,B.P. *et al.* (2007) Resolving the resolution of array CGH. *Genomics*, **89**, 647–653.

Davies,J.J. *et al.* (2005) Array CGH technologies and their applications to cancer genomes. *Chromosome Res.*, **13**, 237–248.

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.

Guttman,M. *et al.* (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet.*, **3**, e143.

Harada,T. *et al.* (2008) Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene*, **27**, 1951–1960.

Heimberger,A.M. *et al.* (2005) The natural history of EGFR and EGFRvIII in glioblastoma patients. *J. Trans. Med.*, **3**, 38.

Hupe,P. *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.

Itsara,A. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.*, **84**, 148–161.

Jackson,M.A. *et al.* (2006) Genetic alterations in cancer knowledge system: analysis of gene mutations in mouse and human liver and lung tumors. *Toxicol. Sci.*, **90**, 400–418.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Klijn,C. *et al.* (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, **36**, e13.

Knudsen,A. (1971) Mutations and cancer: a statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA*, **78**, 820–823.

Komura,D. *et al.* (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, **16**, 1575–1584.

Kotliarov,Y. *et al.* (2006) High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res.*, **66**, 9428–9436.

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

Miller,B.J. *et al.* (2003) Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides evidence for multiple tumor suppressors and identifies novel candidate regions. *Am. J. Hum. Genet.*, **73**, 748–767.

Mitelman,F. *et al.* (eds) (2010) Mitelman Database of chromosome aberrations and gene fusions in cancer. Available at http://cgap.nci.nih.gov/Chromosomes/Mitelman (last accessed date January 4, 2011).

Myllykangas,S. and Knuutila,S. (2006) Manifestation, mechanisms and mysteries of gene amplifications. *Cancer Lett.*, **232**, 79–89.

Natrajan,R. *et al.* (2006) Array CGH profiling of favourable histology Wilms tumours reveals novel gains and losses associated with relapse. *J. Pathol.*, **210**, 49–58.

Newton,M. *et al.* (1998) On the statistical analysis of allelic loss data. *Stat. Med.*, **17**, 1425–1445.

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

Pruim,R.J. *et al.* (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

Rahman,N. *et al.* (1996) Evidence for a familial Wilms' tumour gene (FWT1) on 17q12-21. *Nat. Genet.*, **13**, 461–463.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Rueda,O. and Diaz-Uriarte,R. (2008) Finding recurrent regions of copy number variation: a review. *COBRA Preprint Series*: Paper 42. Available at http://biostats.bepress.com/cobra/ps/art42 (last accessed date January 4, 2011).

Shah,S. (2008) Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet. Genome Res.*, **123**, 343–351.

Simes,R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Sterrett,A. and Wright,F.A. (2007) Inferring the location of tumor suppressor genes by modeling the frequency of allelic loss. *Biometrics*, **63**, 33–40.

Stratchan,T. and Read,A.P. (1999) *Human Molecular Genetics*, 2nd edn. Wiley-Liss Publishing, New York, NY.

Sun,W. *et al.* (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.*, **37**, 5365–5377.

van de Wiel,M. and van Wieringen,W. (2007) CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Res.*, **3**, 55–63.

van de Wiel *et al.* (2009) Smoothing waves in array CCH tumor profiles. *Bioinformatics*, **25**, 1099–1104.

Venkatraman,E. and Olshen,A. (2007) A faster circular binary segmentation algorithm for the analysis of aCGH data. *Bioinformatics*, **23**, 657–663.

Westfall,P. and Young,S. (1993) *Resampling-based Multiple Testing*. Wiley-Interscience. New York.

Zhao,X. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.*, **64**, 3060–3071.