A novel signal processing approach for the detection of copy number variations in the human genome

Catherine Starnoulis\(^1\),\(^2\),\(^3\),\(^4\),\(^*\) and Rebecca A. Betensky\(^5\)

\(^1\)Department of Radiology, \(^2\)Department of Neurology, \(^3\)Clinical Research Program, Children’s Hospital Boston, \(^4\)Harvard Medical School and \(^5\)Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

ABSTRACT

Motivation: Human genomic variability occurs at different scales, from single nucleotide polymorphisms (SNPs) to large DNA segments. Copy number variations (CNVs) represent a significant part of our genetic heterogeneity and have also been associated with many diseases and disorders. Short, localized CNVs, which may play an important role in human disease, may be undetectable in noisy genomic data. Therefore, robust methodologies are needed for their detection. Furthermore, for meaningful identification of pathological CNVs, estimation of normal allelic aberrations is necessary.

Results: We developed a signal processing-based methodology for sequence denoising followed by pattern matching, to increase SNR in genomic data and improve CNV detection. We applied this signal-decomposition-matched filtering (SDMF) methodology to 429 normal genomic sequences, and compared detected CNVs to those in the Database of Genomic Variants. SDMF successfully detected a significant number of previously identified CNVs with frequencies of occurrence \(\geq 10\%\), as well as unreported short CNVs. Its performance was also compared to circular binary segmentation (CBS). Through simulations, SDMF had a significantly lower false detection rate and was significantly faster than CBS, an important advantage for handling large datasets generated with high-resolution arrays. By focusing on improving SNR (instead of the robustness of the detection algorithm), SDMF is a very promising methodology for identifying CNVs at all genomic spatial scales.

Availability: The data are available at http://tcga-data.nci.nih.gov/tcga/

The software and list of analyzed sequence IDs are available at http://www.clear.rice.edu/elec301/Projects02/empiricalMode/code.html

Contact: caterina@mit.edu

Received on November 24, 2010; revised on May 2, 2011; accepted on June 8, 2011

1 INTRODUCTION

Genome-wide DNA aberrations (duplications, deletions and rearrangements) are part of the normal human genetic variability. Copy number variations (CNVs) occur at thousands of loci across the genome (Carter, 2007; Iafrate et al., 2004; Redon et al., 2006). To date, over 60000 DNA aberrations at \(\geq 100\) loci have been identified [Database for Genomic Variants (DGV)]. CNVs may occur in \(\geq 10\%\) of the genome (Beckmann et al., 2007; Zhang et al., 2009), and pairwise comparisons of human genomes suggest that they may differ by at least \(1\%\), a 10-fold higher percentage than the previously estimated \(0.1\%\) based on single nucleotide polymorphism (SNP) incidence (Beckmann et al., 2007). CNVs have also been associated with many diseases and disorders (Kallioniemi et al., 1992; Sebat et al., 2007), and may specifically play an important role in sporadic diseases (Lupski, 2007).

Array comparative genomic hybridization (aCGH) is a high-resolution technology that enables simultaneous interrogation of the entire genome. It involves hybridization of differentially fluorescent dye-labeled reference and test sequences on a microarray, and estimation of relative allelic changes as the log2-ratio of the two fluorescence intensities. CNV detection is often limited by the resolution of the microarray and the data signal-to-noise ratio (SNR). Current platforms with \(\geq 1\) million probes enable the detection of shorter CNVs (10–25 kb), and consequently capture the finer scale of genomic variability (Lupski, 2007; Perry et al., 2008). In practice, detection of short CNVs is still limited by high noise levels in the data. Consequently, methodologies that explicitly increase SNR to improve CNV detection are needed.

Several detection methods have been proposed and applied to genomic data with variable success (Barros et al., 2007; Fiegler et al., 2006; Wineinger et al., 2008). They include smoothing procedures (Hupe et al., 2004), Hidden Markov Models (HMM) (Fridlyand et al., 2004; Snijders et al., 2001), Circular Binary Segmentation (CBS) (Olshen et al., 2004), a change-point method that segments a chromosome into regions of constant copy number, and the Genome Alteration Detection Algorithm (GADA) (Pique-Regi et al., 2008), based on Bayesian learning. These methods aim to improve the true positive detection rate, not to increase SNR. Their performance also depends on the selection of appropriate thresholds for calling CNVs. The false discovery rate (Benjamini et al., 1995) or the Genomic Identification of Significant Targets in Cancer (GISTIC) approach are often used to establish these thresholds (Beroukhim et al., 2007).

Genomic and measurement noise, as well as impurities and artifacts may contaminate aCGH data, decrease their SNR and introduce spurious spatial correlations (Fridlyand et al., 2004; Marioni et al., 2007). Dimensionality reduction methods, including principal component analysis (PCA), independent component analysis (ICA) and wavelet decomposition (Alter et al., 2000; Hsu et al., 2005), have been applied for denoising and artifact introduction detection.
suppression in genomic data. The lack of physical interpretation of some components, or lack of a priori knowledge of an appropriate mother wavelet, are respective drawbacks of these methods.

Signal processing methods for locally increasing SNR are widely used in the physical sciences and engineering, where measured signals are often severely contaminated by different types of noise. When a waveform of interest is known or can be robustly estimated, it may be recovered from a noisy signal via pattern matching methods, e.g. matched filtering, which increases SNR locally, in regions where the known signal (the template) matches the observed signal, and reduces SNR elsewhere (Turin, 1960; Willet et al., 1991). The application of signal processing methods to genomic data has been very limited, since these data are not truly continuous (Lai et al., 2005). However, such methods may be very useful for improving CNV detection. In this article, data-driven signal decomposition (for denoising), followed by matched filtering, hereafter referred to as Sequence-Decomposition-Matched Filtering (SDMF), is proposed to locally improve SNR, and consequently restrict the genomic regions of interest and increase the specificity of CNV detection. SDMF is applied in normal genomic sequences from The Cancer Genome Atlas (TCGA). The performance of SDMF is compared to CBS through simulations. Detected CNVs are also compared to identified CNVs in the DGV.

2 APPROACH

A two-step methodology is proposed, summarized in Figure 1. Signal decomposition is performed to increase the overall data SNR and suppress artifacts, followed by matched filtering, where pairs of decompositions is performed to increase the overall data SNR and

3 METHODS

3.1 Sequence decomposition for artifact elimination

3.1.1 Data ACGH sequences (log$_2$ intensity ratios) from 429 normal samples were obtained from the Cancer Genome Atlas (TCGA). These included all 322 normals matched to glioblastoma samples, from all available data batches (1-4, 6-10, 16, 20, 26), and all 107 normals matched to ovarian cancer samples, from all available data batches (9, 11, 15). The Agilent Human Genome CGH Microarray 244A was used for hybridization in both sets. Cytoband information was obtained from Agilent and TCGA (Gene list, http://chem.agilent.com). Detected CNVs were compared to those catalogued in the DGV. All sequences were segmented into individual chromosomes, which were then analyzed separately.

3.1.2 Mode estimation and signal decomposition A wave-like genomic artifact has been identified in array CGH, SNP arrays and whole-genome tiling arrays (Diskin et al., 2008; Komura et al., 2006), and appears to be highly correlated with the guanine-cytosine (GC) content in DNA (Diskin et al., 2008). This artifact introduces spurious spatial correlations in the data, which makes the detection of CNV breakpoints difficult. Previous studies have applied different methods to suppress the artifact (Diskin et al., 2008; Marioni et al., 2007; Van de Wiel et al., 2009). Each has advantages and shortcomings. For example, smoothing methods may eliminate small amplitude CNVs in addition to artifacts.

The structure of typically non-stationary genomic log$_2$ sequences may be investigated using an appropriate decomposition method. There are very few truly data-driven methods for non-stationary signals, that do not make a priori assumptions on the shape of the unknown signal components (referred to as modes). The most widely used method is the Empirical Mode Decomposition (EMD) (Huang et al., 1998), which recursively decomposes a non-stationary signal into modes with significant amplitude contributions. These modes must satisfy two conditions: (i) they must have zero mean and (ii) they must contain a single extremum between zero crossings. In addition, they must be such that the signal reconstruction error between the original signal $x(t)$ and mode-based estimated signal $\tilde{x}(t)$ [see Equation (2)], defined as $\sum_{k=1}^{K} (x(k) - \tilde{x}(k))^2$, is minimized. The algorithm first identifies one set of maxima and one set of minima in a signal, and fits respective cubic spline functions through them to obtain the upper and lower signal envelopes. The mean of the two envelopes $m_{1,1}$ is then subtracted from the original signal. If the mean of the residual $h_{1,1}$ is zero, then $h_{1,1}$ is the first mode. In practice, multiple iterations are necessary, at which new means are estimated and removed from successive residual signals. This procedure continues until convergence, i.e. until the first $q$ that satisfies:

$$\sum_{k=1}^{K} (h_{q,k} - \hat{h}_{q,k})^2 < \epsilon,$$

which implicitly ensures that $h_{1,1}$ is zero mean. In Equation (1), $k$ denotes probe and $K$ is the total number of probes in a chromosome sequence. The first mode is then defined to be $d_1(t) = h_{1,1}(t)$, and is subtracted from $x(t)$ to obtain the residual $r_1(t)$, which is then treated as the new signal. The sifting process is repeated for $r_1(t),...,r_{M-1}(t)$, to obtain modes $d_2(t),...,d_{M}(t)$. The decomposition stops once the difference between successive reconstruction errors, based on sets with progressively higher number of modes, is negligible. Additional criteria have been used to stop the mode estimation (Huang et al., 1998), e.g. (i) the root mean-squared residual $r_{M,\text{rms}} = \sqrt{\frac{\sum_{k=1}^{K} r_{M,k}^2}{K}} \leq 1$, and/or (ii) a depletion of extrema through which to fit the envelopes, and/or (iii) a pre-defined upper bound on the number of modes is reached. Here we set an upper bound ($M_{\text{up}} = 30$), which was, however, never reached. The signal reconstruction error was typically minimized when $M \leq 15$ modes were estimated. Thus, EMD is also a dimensionality reduction approach. The reconstructed signal based on $M$ modes is given by:

$$\hat{x}_{M}(t) = \sum_{k=1}^{M} d_{k}(t)$$

for $k=1,...,K$. Some of these $M$ modes may still correspond to high-amplitude artifacts and/or noise. Thus, a subset $M' \leq M$ is ultimately selected, as described below.

Although a powerful method, there are sources of variability in the EMD estimation. First, there are theoretically infinite combinations of modes with different amplitude distributions that can be estimated from a given signal (Huang et al., 1998). There is currently no consensus on what constitutes an optimal basis (these modes form an approximately orthogonal basis) (Huang et al., 1998), and thus the requirement of basis uniqueness is not satisfied. Second, the stoppage criteria for mode estimation are often arbitrarily set. Thus, a criterion for the selection of an optimum set of modes is desirable. Thresholding approaches, model selection criteria and penalized least-squares approaches can all be used to estimate a subset of modes (Donoho et al., 1994), since the stoppage criterion may not penalize...
higher order (redundant) decompositions. However, these methods do not directly provide a means for selecting an optimum mode subset. A different upper bound may result in a different distribution of modes, and a higher number of modes is not necessarily optimal. An example of a sequence and reconstructed signals using upper bounds of 20, 15 and 10 modes, respectively, are shown in Figure 2. The reconstruction error based on 15 modes is two orders of magnitude higher than the error corresponding to 10 modes.

To assess the effect of individual modes, signal $x(k)$ may be decomposed into a set of modes that are sequentially eliminated to obtain the $m$-mode reduced $\hat{x}_m(k)$. This is then compared to $x(k)$:

$$\hat{x}_m(k) = x(k) + \hat{e}_m(k)$$

where $\hat{e}_m(k)$ is the noiseless signal, $\hat{x}_m(k)$ is the estimate of this signal and $e \sim N(0, \sigma_e^2)$ and $\hat{e}_m \sim N(0, \sigma_{\hat{e}_m}^2)$ with $\sigma_{\hat{e}_m}^2 < \sigma_e^2$. The additive Gaussian noise of the original and reconstructed signals. The requirement for denoising may be expressed as $\sigma_{\hat{e}_m}^2 < \sigma_e^2$, i.e. the variance of the estimated signal is less than the variance of the noise in the measured signal. When additional modes are eliminated, the noise variance decreases further, i.e. $\sigma_e > \sigma_{\hat{e}_m} > \sigma_{\hat{e}_{m-1}} > \ldots > \sigma_{\hat{e}_1}$. However, sequential mode reduction may also eliminate significant components, resulting in large signal reconstruction errors. Thus, we need to bound the number of eliminated modes. We define the risk function as:

$$R_m(\hat{x}_m, x) = \frac{1}{K} \sum_{k=1}^{K} (x(k) - \hat{x}_m(k))^2 + \sigma_e^2$$

When 0 modes are eliminated from the original signal, $\hat{x}_0(k) = x(k)$ and $R_0 = 0$. When the reconstructed signal is the optimum estimate of the $x(k)$, then $\hat{x}_m(k) = x(k)$ and Equation (5) becomes:

$$R_{m, opt}(\hat{x}_m, x) = \frac{1}{K} \sum_{k=1}^{K} (x(k) - \hat{x}_m(k))^2 =$$

$$\frac{1}{K} \sum_{k=1}^{K} (x(k) - \hat{x}_m(k))^2 =$$

$$\frac{1}{K} \sum_{k=1}^{K} w_m(k)^2$$

It is assumed that $\hat{e}_m(k)$ and $x(k)$ are independent and thus $\sigma_{\hat{e}_m}^2 = \sigma_e^2$. Then, if we set the variance of the mode-reduced signal $\hat{x}$ equal to the risk function, we have (from Equation (6))

$$\sigma_{\hat{e}_m}^2 + \sigma_{\hat{e}_m}^2 = 2\mathrm{Cov}(e, x)$$

which ensures the desired $\sigma_{\hat{e}_m}^2 < \sigma_e^2$ (assuming $\sigma_e$ and $e$ are positively correlated), with equality when $x(k)$ or $\hat{e}_m(k)$ is zero, i.e. a noiseless measurement $x(k) = \hat{x}(k)$ or noiseless estimate $\hat{x}(k) = \hat{x}_m(k)$ and asymptotic equality when $K \rightarrow \infty$.

There are several threshold estimation approaches for denoising, e.g. wave-shrinkage (Donoho et al., 1997). Although perhaps not optimal, our proposed approach has the advantage of not requiring detailed estimation of the threshold. Setting the variance of the mode-reduced signal equal to the risk function ensures that mode reduction increases SNR. An example is shown in Figure 3, using a subset of the data. Typically, the variance and risk function intersect at mode 1 or 2, depending on the chromosome, and thus 1–2 modes were eliminated.

We compared the proposed approach to loess curve fitting (Marioni et al., 2007), PCA and ICA. Corresponding changes in SNR (averaged over all sequences) are shown in Figure 4. SNR was computed as the $\log_2$ of the ratio of the signal to noise (SNR).

3.2 Effect of normalization by a common reference

In addition to the wave-like artifact, a common reference for normalization may also introduce spurious spatial correlations in genomic data. To explore this, the normalization was reversed by adding the log$_2$ reference to each observed sequence, to obtain non-normalized data:

$$\log_2(x_{\text{batch}}(k)) = \log_2(x_{\text{obs}}(k)) + \log_2(x_{\text{batch}}(k))$$

where $x_{\text{batch}}$ is the non-normalized sequence and $x_{\text{obs}}$ the batch-normalized sequence. Each non-normalized sequence was decomposed. Contributions of low-amplitude modes were compared in raw and normalized sequences. The wave-like artifact was not suppressed in non-normalized data, which implies that it is unrelated to the normalization, as also noted in Diskin et al. (2008).

We assessed the effect of renormalization (measured by the variance of the sequence) by renormalizing each sequence by the mean of all others, as shown in Figure 5 for (20 sequences). The variance of renormalized raw and mode-reduced data were compared. The diagonal entries correspond to the variance of the non-normalized sequence. The effect of normalization was in some sequences significantly higher (~10x) than in others, e.g. the variance increase in sequence 1 was negligible when the sequence was normalized by 2–4, 8–14, 18–20, but substantial when normalized by sequences 15–17. In contrast, renormalization of mode-reduced sequences resulted in small variance changes. Thus, denoising via mode elimination reduces the uncertainty associated with normalization by different sequences.

3.3 Matched filtering

Matched filtering is a waveform-matching method that is widely used in pattern recognition, sonar and communications, to detect a desired waveform.
c(k) in a measured signal x_{obs}(k). The matched filter is a quasi-optimum filter that locally maximizes the output SNR by decreasing SNR in regions of mismatch between observed and desired (template) signals and increasing SNR in matched regions. It improves SNR by reducing the noise spectral bandwidth to that of the desired waveform (Turin, 1960). Theoretically, the optimum filter that maximizes SNR is the time-reversed signal itself, i.e. c(k) = x_{obs}(-k) under the assumption of Gaussian noise, and matched filtering is a convolution operation:

\[ h(k) = c_{template}(k) \ast x_{obs}(k) \]  

(9)

**Template signal:** typically, the template c(k) is either precisely known or robustly estimated. An observed signal is then matched-filtered with this template. However, in this study there was no unique or known template. Every normal chromosome sequence was treated both as a template and an observation. To identify CNVs, regions of signal mismatch rather than regions of match were examined. Each of N_c mode-reduced sequences was segmented according to the filter length described below, and each of Q segments was filtered with N_c - 1 corresponding segments (the templates), to obtain N_c - 1 matched-filtered segments. The ratio of pre- over post-filtering SNR \( \Delta SNR \) was computed for each segment, and averaged over N_c - 1 SNRs. A matrix of \( \Delta SNRs \) was ultimately obtained of size N_c x Q. A segment was further examined for potential CNVs if \( \Delta SNR \), averaged over N_c was less than an estimated threshold, described below.

**Filter length:** the choice of filter length is important for achieving optimal SNR. We examined the change in SNR following matched filtering as a function of filter length. We used the typical definition of SNR \( 20 \log_{10} \frac{\text{rms}}{\text{rms}_{\text{in}}} \) in dB, estimated the change in SNR as \( \Delta SNR = 20 \log_{10} \frac{\text{rms}_{\text{in}}}{\text{rms}} \), where \( < >_{\text{rms}} \) denotes root-mean squared. Figure 6 shows the average (over all sequences and chromosomes) SNR change as a function of filter length. Since the first change in slope of average SNR change occurs at a length of \( \sim 500 \) probes, which was determined empirically to be sufficiently long for matched filtering, we selected this as the segment/fILTER length, though it may not be optimal.

**Threshold for estimating regions of mismatch:** contrary to the traditional application of matched filtering, for CNV detection we are interested in regions of mismatch rather than regions of maximum match. This is because we presume that CNVs occur with relatively low frequency. Regions of mismatch not only include both allelic variability between two sequences but also potentially distinct noise and artifacts. Thus, the first step of mode decomposition for denoising is necessary. Furthermore, a threshold or random mismatch due to noise rather than true waveform mismatch must also be estimated, instead of assuming that any decrease in SNR, i.e. \( \Delta SNR < 0 \), following matched filtering reflects waveform mismatch. We simulated and matched-filtered 500 pairs of identical, and thus theoretically perfectly matched signals, but corrupted with distinct additive Gaussian noise sequences, while maintaining high SNR. We estimated that noise-induced mismatch resulted in a maximum decrease in SNR of 1.5 dB. Therefore, regions where mean SNR in the filtered sequence decreased by more than \( \sim 1.5 \) dB were identified as regions of potential true waveform mismatch.

**CNV call:** matched filtering increases/decreases SNR according to match/mismatch, and thus the resulting signal amplitude in the filtered data may be different from the log₂ ratios in the original (but denoised) data. Thus, once potential segments of mismatch have been identified, we need to examine the corresponding segments in the denoised data for determining the actual type of aberration at the probe level. For simplicity, we set a detection threshold of one loss or one gain, i.e. log₂ \( \frac{1}{2} \) for gain and log₂ \( \frac{1}{2} \) for loss.

Fig. 3. Effects of sequential mode elimination on the risk function (black) and variance (red) of the mode-reduced sequence. (A) The inter sequence parameter variability for one chromosome (dashed lines). Mean parameter values for several chromosomes (B).

Fig. 4. Comparison of artifact elimination via PCA (A), ICA (B), loess curve fitting (C) and mode elimination (D). Each plot is for SNR before (x-axis) versus SNR after application of the respective method (y-axis), averaged over all normal sequences.
probability of a CNV at probe \( k \) was defined as the union of the probabilities of (mutually exclusive) gain and loss. This is practically equivalent to the maximum of these probabilities since one of the probabilities is typically negligible relative to the other:

\[
P_r(CNV_k) = \frac{\sum_{i \leq k} \log_2(1 + \frac{\text{SNR}_i}{\text{SNR}_k})}{N_s} + \frac{\sum_{i > k} \log_2(1 + \frac{\text{SNR}_i}{\text{SNR}_k})}{N_s} \tag{10}
\]

where \( N_s \) is the total number of sequences, and \( \log_2(1 + \frac{\text{SNR}_i}{\text{SNR}_k}) \) corresponds to the mode-reduced sequence \( j \) at probe \( k \). A probability of 0.1 (10% frequency) was chosen as the threshold for a CNV call. Although this choice is somewhat arbitrary, low CNV frequencies (\( \leq 10\% \)) have been reported in the normal genome (Ionita-Laza et al., 2008; Jakobsson et al., 2008).

### 4 RESULTS

#### 4.1 CNV detection in the normal genome

We applied SDMF to 429 normal sequences from the TCGA database, and compared them to CNVs in the DGV detected in datasets with least 420 samples, irrespective of their frequency of occurrence, as well as CNVs in two studies with 270 samples and high frequency (\( \geq 40\% \)) (McCarroll et al., 2008; Pinto et al., 2007). For the Y chromosome, the maximum sample size in any study was 270. For two CNVs to be comparable, we assumed a very conservative 50% spatial overlap of the length of the detected CNV with respect to the length of the CNV in the DGV. In cases where the same CNV was reported by multiple studies, but with slightly different lengths, the intersection of these regions was assumed as the CNV length. Finally, when a detected aberrations was entirely contained within the genomic coordinates of the corresponding CNV in the DGV, it was called a separate CNV irrespective of its overlap (this occurred in only very few instances). On average, filtering restricted the regions of mismatch for further examination to 19–34% of the total number of segments, depending on the chromosome. This is expected as these were all normal sequences, and thus with high waveform similarities and low mismatch. Table 1 summarizes the number of CNVs in DGV, in each chromosome, irrespective of frequency (column 2), the overall sensitivity of SDMF with respect to DGV, calculated as the ratio of the number of CNVs in the intersection of the two sets over the total number of CNVs in DGV (column 3); Cohen’s kappa statistic (Cohen, 1960), as a measure of agreement between SDMF and DGV that accounts for overlap between the two occurring by chance (column 4); CNVs detected by SDMF with at least 10% frequency (column 5); CNVs in DGV with at least 10% frequency (column 6); sensitivity of SDMF with respect to CNVs in DGV with frequencies at least 10% (column 7); and the corresponding kappa (column 8). Note that the kappa statistic is calculated at the probe level, by assuming that the maximum number of CNVs is the number of probes in each chromosome. However, CNVs in DGV have been detected with a wide range of arrays of different resolution. To account for this, we obtained the genomic coordinates for each identified CNV in the DGV and calculated the corresponding number of probes in our data, so that the two sets are comparable for the estimation of kappa.

A significant number of detected CNVs have also been reported in DGV with frequencies \( \geq 10\% \). For this subset, the sensitivity of SDMF with respect to DGV was \( \geq 83\% \). The overall sensitivity of SDMF was lower (17–60%). Furthermore, based on the accepted interpretations of levels of kappa (Landis et al., 1977), the agreement between DGV and SDMF was slight (0 < \( \kappa < 0.2 \)) to moderate (0.4 < \( \kappa < 0.6 \)), depending on the chromosome. This may be due to: (i) differences in the populations of DGV and TCGA normals; (ii) novel CNVs detected by SDMF; and (iii) the use of distinct thresholds for CNV calls in different studies. We can only assess the false positive rate of SDMF through simulations.

#### 4.2 Comparison of SDMF with CBS via simulation

SDMF may be applied to identify regions of mismatch, and thus to detect CNVs occurring at low frequencies in a relatively large set of genomes, or to identify regions of significant waveform match between sequences, and thus to detect common CNVs occurring at high frequencies. We compared SDMF to CBS. Although other
Table 1. CNVs in DGV (column 2), sensitivity of SDMF with respect to DGV (column 3), $\kappa$ statistic (column 4), CNVs in DGV and in this study, with at least 10% frequency (columns 5 and 6), sensitivity of SDMF with respect to CNVs in DGV with at least 10% frequency (column 7) and the corresponding $\kappa$ (column 8).

| Chr | DGV     | Sens. | SDMF | $\kappa$ | SDMF | DGV | Sens. | SDMF | $\kappa$ |
|-----|---------|-------|------|________|------|-----|-------|------|________|
|     | w.r.t. DGV |       | $\geq$10% |       |       | w.r.t. DGV | $\geq$10% |       |       |
| 1   | 1235    | 0.44  | 0.41  | 1178    | 115  | 0.92 | 0.41  | 1178    | 115  | 0.92 |
| 2   | 833     | 0.46  | 0.46  | 586     | 83   | 0.89 | 0.46  | 586     | 83   | 0.89 |
| 3   | 726     | 0.42  | 0.5   | 456     | 82   | 0.96 | 0.5   | 456     | 82   | 0.96 |
| 4   | 727     | 0.34  | 0.38  | 478     | 108  | 0.91 | 0.38  | 478     | 108  | 0.91 |
| 5   | 652     | 0.31  | 0.32  | 495     | 80   | 0.95 | 0.32  | 495     | 80   | 0.95 |
| 6   | 586     | 0.43  | 0.4   | 550     | 100  | 0.9  | 0.4   | 550     | 100  | 0.9  |
| 7   | 642     | 0.48  | 0.56  | 423     | 73   | 0.94 | 0.56  | 423     | 73   | 0.94 |
| 8   | 638     | 0.4   | 0.5   | 530     | 76   | 0.92 | 0.5   | 530     | 76   | 0.92 |
| 9   | 585     | 0.49  | 0.54  | 428     | 37   | 0.94 | 0.54  | 428     | 37   | 0.94 |
| 10  | 677     | 0.37  | 0.47  | 488     | 55   | 0.94 | 0.47  | 488     | 55   | 0.94 |
| 11  | 607     | 0.23  | 0.28  | 313     | 73   | 0.91 | 0.28  | 313     | 73   | 0.91 |
| 12  | 577     | 0.41  | 0.42  | 473     | 74   | 0.94 | 0.42  | 473     | 74   | 0.94 |
| 13  | 347     | 0.37  | 0.4   | 256     | 12   | 0.92 | 0.4   | 256     | 12   | 0.92 |
| 14  | 382     | 0.28  | 0.33  | 225     | 31   | 0.94 | 0.33  | 225     | 31   | 0.94 |
| 15  | 441     | 0.44  | 0.54  | 253     | 37   | 0.95 | 0.54  | 253     | 37   | 0.95 |
| 16  | 411     | 0.39  | 0.42  | 296     | 45   | 0.91 | 0.42  | 296     | 45   | 0.91 |
| 17  | 365     | 0.51  | 0.57  | 269     | 52   | 0.91 | 0.57  | 269     | 52   | 0.91 |
| 18  | 213     | 0.17  | 0.25  | 65      | 12   | 0.83 | 0.25  | 65      | 12   | 0.83 |
| 19  | 466     | 0.28  | 0.32  | 241     | 36   | 0.89 | 0.32  | 241     | 36   | 0.89 |
| 20  | 255     | 0.54  | 0.58  | 203     | 24   | 0.88 | 0.58  | 203     | 24   | 0.88 |
| 21  | 159     | 0.57  | 0.67  | 106     | 15   | 0.93 | 0.67  | 106     | 15   | 0.93 |
| 22  | 287     | 0.23  | 0.35  | 72      | 17   | 1    | 0.35  | 72      | 17   | 1    |
| X   | 309     | 0.59  | 0.62  | 267     | 22   | 0.91 | 0.62  | 267     | 22   | 0.91 |
| Y   | 27      | 0.6   | 0.52  | 34      | 2    | 1    | 0.52  | 34      | 2    | 1    |

Methods have been shown to be computationally faster, e.g. Pique-Regi et al. (2008), CBS is widely used for CNV detection (Olshen et al., 2004), by simulating both cases (match/mismatch). The respective performance of the two methods was measured by the true and false detection rates, as a function of SNR and inter-CNV distance. Sequences were corrupted with additive, multiplicative or periodic (sine wave) Gaussian noise. SNR was measured in decibel. We simulated 1000 sequences (100 per each of 10 noise levels). In this set, CNV frequency was 100% and matched filtering was applied in the traditional sense, to identify regions of maximum match between template and test sequences, which corresponded to regions containing CNVs. To simulate low-frequency occurring CNVs, all but 10–25% of sequences in this set were randomly selected and replaced by pure Gaussian noise. An example of a sequence containing 1 short CNV (300 markers long), contaminated with progressively higher additive noise levels (decreasing SNR) is shown in Figure 7. Note that SDMF was applied to simulated data in the same way as in real data, using noise-contaminated sequences as both the template and test signals. All sequences were first mode-decomposed for denoising.

The matched-filtered sequences corresponding to Figure 7 are shown in Figure 8. Matched filtering successfully suppressed noise in regions of mismatch, even in very low SNR ($-7$ to $-10$ dB), locally increased SNR in the neighborhood of the CNV by at least 5 dB and decreased SNR elsewhere by at least 3 dB. This example demonstrates the ability of SDMF to detect even short CNVs in very high noise levels. Note that when a template with high SNR is precisely known, the first step of signal denoising by mode-reduction results only in modest additional increases in SNR ($<$2 dB). However, when a robust template is not precisely known and/or is noisy, mode decomposition is a necessary pre-processing step.
Fi.9. Original (black) and CNV detections based on CBS (green).

In the examples shown in Figure 9, CBS performed very well for SNR ≥ −6dB, but detected spurious CNVs at lower SNRs. In contrast, SDMF correctly identified one CNV in all sequences. In cases of pure noise (no CNVs), neither SDMF nor CBS identified any CNVs. Also, the additional simulations for mismatch did not change the results of CBS, which processes each sequence individually. True and false detection rates of SDMF and CBS as a function of SNR, noise type and distance between CNVs are summarized in Table 2. SDMF correctly detected at least 70% of CNVs irrespective of SNR and noise type, compared with at least 15% detected by CBS. Multiplicative noise decreased the true detection rate in both methods, but periodic noise did not significantly affect detections by SDMF, possibly due to the initial denoising. With regard to inter-CNv distance, for fewer than 1000 markers, SDMF detected 84% of CNVs, versus 40% by CBS, at low SNR. CBS sometimes detected closely spaced CNVs as a single CNV. Also, the estimated mean gain/loss, which depended on the accuracy of the CNV breakpoints, was more precisely estimated by SDMF, particularly for SNR < −6 dB. Finally, false detection rate, estimated as the ratio of falsely identified CNV probes over the number of true non-CNv probes, increased significantly for multiplicative and periodic noise for CBS but not for SDMF. This calculation was based on 8000 probes and a varying number of CNVs of different lengths. In simulations for mismatch, the true detection rate was higher for SDMF than CBS and the false detection rate was lower. As expected, the false detection rate of SDMF in these simulations was slightly higher than the corresponding rate in simulations for waveform match, since it is calculated at the probe level and random mismatch at low SNR may affect the estimated CNV length. In general, although SDMF substantially outperformed CBS at low SNR, CBS performed well for higher SNR. However, the computational cost of CBS was significantly higher (O(5) in simulations and O(8) in real data). Next-generation CGH arrays with very high resolution will require computationally efficient analysis methods. SDMF can handle large genomic datasets, whereas CBS is limited in that respect.

5 DISCUSSION

We have developed SDMF, a computationally efficient methodology that combines the signal processing tools of signal decomposition and matched filtering for CNV detection in array CGH data. The latter are inherently noisy and robust CNV detection is difficult at low SNR. SDMF treats genomic sequences as spatially continuous, non-stationary signals, which are denoised through elimination of noise-related components. Spatially localized matched filtering is then applied to identify regions of pairwise match/mismatch. Through these two processes, SDMF increases SNR and improves the specificity of CNV detection. Note that SDMF requires the selection of the matched-filter length and thresholds for CNV detection. These were not explicitly optimized in this study. The choices of filter length and match/mismatch thresholds are dataset dependent. The log2 ratio threshold depends on the goal of the study, e.g. for detection of pathological CNVs it may be based on the log2 ratios of normal CNVs. SDMF was applied to 429 normal sequences from the...
TCGA, and detected CNVs were compared to those in the DGV. Very high sensitivity (≥83%) relative to the DGV for CNVs with frequency ≥10% was estimated. In simulations, SDMF had a higher detection rate and significantly lower false detection rate and computational cost than the CBS method, an important advantage when analyzing large datasets from very high-resolution arrays.

In summary, SDMF is a promising approach for CNV detection even in very noisy genomic data. Microarray platforms are rapidly improving and remain the most appropriate tool for interrogating the entire genome at a reasonable cost. Although next-generation sequencing technologies are very promising, their cost currently prohibits the analysis of the entire genome, and may require significant amounts of DNA not always available. Furthermore, there is much information to be gained from existing well-annotated array-based studies through the application of highly sensitive and computationally efficient methods such as SDMF.

ACKNOWLEDGEMENT

We thank Dr Oliver Hoffmann for the aCGH probe mapping, and Dr Yiping Shen for discussions on the future of microarrays. The results published here are based upon data generated by the Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information from the Database of Genomic Variants has been used in this study.

Funding: National Institutes of Health (Grant R03 CA121884).

Conflict of interest: None declared.

REFERENCES


