Published in final edited form as:
Int J Comput Biol Drug Des. 2008 January 1; 1(4): 368–395. doi:10.1504/IJCBDD.2008.022208.

Statistical issues in the analysis of DNA Copy Number Variations

Nathan E. Wineinger,
Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA, Fax: 205-975-2540, E-mail: nwineing@uab.edu

Richard E. Kennedy,
Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA, Fax: 205-975-2540, E-mail: rkennedy@ms.soph.uab.edu

Stephen W. Erickson,
Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA, Fax: 205-975-2540, E-mail: serickson@ms.soph.uab.edu

Mary K. Wojczynski,
Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA, Fax: 205-975-2540, E-mail: mwojczynski@ms.soph.uab.edu

Carl E. Bruder, and
Viral Biochemistry, Division of Drug Discovery, Southern Research Institute, Birmingham, Alabama 35205, USA, Fax: (205) 581-2097, E-mail: bruder@southernresearch.org

Hemant K. Tiwari
Section on Statistical Genetics, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA, Fax: 205-975-2541, E-mail: htiwari@uab.edu

Abstract

Approaches to assess copy number variation have advanced rapidly and are being incorporated into genetic studies. While the technology exists for CNV genotyping, a further understanding and discussion of how to use the CNV data for association analyses is warranted. We present the options available for processing and analysing CNV data. We break these steps down into choice of genotyping platform, normalisation of the array data, calling algorithm, and statistical analysis.

Keywords

DNA; copy number; normalization; array CGH; calling algorithm; CBS; circular binary segmentation; hidden Markov model; whole genome amplification; microarray; GWAS; genome-wide association study; complex disorders

Correspondence to: Hemant K. Tiwari.

1 Introduction

Biological knowledge of genetic variants has grown as the technology to measure various forms of genetic variation continues to develop, which is evident through the increasing resolution and coverage of the genome through available genotyping platforms. Initially biomarkers and early genetic markers such as blood types, restriction length polymorphisms, and microsatellites were used to map simple Mendelian disorders. Now high-density genome-wide platforms are used to characterise the extent of phenotypic variation in common and complex diseases. As private and commercial entities compete for the most thorough and accurate genetic maps available, investigators are entrusted to determine the effect these new genetic variants have on phenotypic variation. Only recently has the abundance and frequency of DNA copy number aberrations been detected (Freeman et al., 2006). Termed Copy Number Variation (CNV), these structural variants have been classified as being genomic segments larger than 1 kb that are present in variable copy number (Feuk et al., 2006). CNVs can be simply thought of as segmental deletions or tandem duplications compared to a reference genome, which are referred to as a loss or gain in copy number, respectively. However, more complex gains and losses also fall into this category.

Early CNV discoveries can be credited with laying the methodological foundation to identify and measure this relatively new form of genetic variation, while more recent studies have attempted to characterise the extent of this distribution throughout the genome. Redon et al. (2006) constructed a first-generation map of CNV in the human genome from 270 individuals in the HapMap collection (International HapMap Consortium, 2005). Their results documented the prevalence of CNVs in these representative populations. They found over 1400 Copy Number Variable Regions (CNVR) covering roughly 12% of the genome. These regions overlapped with over half the known reference sequence genes which, in turn, aided in bringing questions forward as to what degree CNVs may play in phenotypic variation. Since then a number of studies have continued to map CNVs onto the genome, usually using samples from different populations. A few such studies include Pinto et al. (2007) who narrowed their study of the genome-wide distribution of CNVs to unrelated healthy individuals from Northern Germany and de Smith et al. (2007) who examined Caucasian male controls from Northern France.

In anticipation of the growing list of copy number variable loci and regions, Iafrate et al. (2004) developed the Database of Genomic Variants (DGV) (http://projects.tcag.ca/variation/). As of July 2008, 49 papers have been cited in the DGV, including those mentioned above. This collection of CNVRs covers 29% of the human genome, with coverage on particular chromosomes ranging from 20% on chromosome 18–43% on chromosome 19. Many copy number variants have no impact on phenotypic variation (Lupski, 2007) and are biased towards non-coding regions (Conrad et al., 2006). However, 39% of "Online Mendelian Inheritance in Man (OMIM)" genes overlap CNVRs in the DGV. A number of studies have already discovered associations between common copy number variants, referred to as Copy Number Polymorphism (CNP), and common diseases such as pancreatitis, Alzheimer's disease, HIV progression, Schizophrenia, Autism, and Crohn's Disease (Gonzalez et al., 2005; Le Maréchal et al., 2006; Lee and Lupski, 2006; Stefansson et al., 2008; Walsh et al., 2008; Sebat et al., 2007; Marshall et al., 2008; McCarroll et al., 2008). Non-inherited, *de novo* CNVs have also been reported to be associated with complex diseases such as autism Sebat et al., 2007). Recently the Wellcome Trust Sanger Institute announced that they are expanding their original genome-wide association study (Wellcome Trust Case Control Consortium, 2007) to include a comprehensive set of copy number variants (http://www.sanger.ac.uk/Info/Press/2008/080414.shtml). We expect this trend of CNV inclusion to continue until genome-wide coverage of copy number variants is integrated into all such studies.

Investigators exploring CNV in genome-wide studies have a number of options available in terms of processing and analysing DNA samples. We break these options into the following choices:

- technological platform from which raw data will be obtained
- removal of non-biological variation through array normalisation and/or smoothing algorithm
- calling algorithm to obtain biological meaning of the measurements
- statistical method used to analyse the processed data.

Here we present these options as well as provide an overview of the statistical issues imbedded in each step.

2 CNVs genotype technology

Large-scale copy number changes in the genomes of humans and other vertebrates have been overlooked partly due to technical difficulties in scoring these changes genome wide with satisfactory resolution. The human genome project enabled for the development of Comparative Genomic Hybridisation (CGH) to microarrays (array CGH), which allowed for higher resolution detection of DNA copy number aberrations (Pinkel et al., 1998; Solinas-Toldo et al., 1997). Bacterial Artificial Chromosome (BAC) arrays are a particular type of array CGH that is comprised of large BAC insert clones. Snijders et al. (2003) described a genome-wide DNA microarray that consisted of 2400 BACs distributed across the genome, and the first BAC array contiguously covering a human chromosome appeared in 2002 (Buckley et al., 2005). Current platforms cover the human genome entirely, using approximately 26,000–30,000 overlapping BAC clones (Redon et al., 2006; de Ståhl et al., 2008). The array's detection technique is based on the assessment of fluorescence ratios between differentially labelled test and reference DNA; using unlabeled Cot-1 DNA to suppress the large amounts of repeats in the genome that otherwise would compromise the integrity of the analysis. BAC arrays have high signal-to-noise ratios, and produce data with low standard deviation. We have recently shown that the 32K platform is capable of identifying copy number changes occurring in only a fraction of the analysed cell population (Bruder et al., 2008). The relatively low effective resolution (max. 15–35 kb), and the large amount of DNA needed for the analysis has, however, prompted further platform development.

There exist several commercial alternatives to BAC arrays. These platforms either contain oligonucleotides targeting different genomic sequences (Agilent or NimbleGen) or single nucleotide polymorphisms (Affymetrix or Illumina). The oligonucleoetide arrays were initially constrained by printing, as spotting hundreds of thousands of synthesised DNA features on a single microscope slide is technically very challenging. However, new technologies making DNA synthesis directly on the array surface possible, have allowed as many as 2.1 million oligonucleotides to be accessible on the same array (Walsh et al., 2008). Oligonucleotide arrays offer the highest theoretical resolution for CGH, as the hybridisation probes can be designed to virtually cover the entire genome. The current oligonucleotide design is either based on isothermic oligos of varying length (Nimblegen) or hybridisation optimised 65-mers (Agilent). However, the signal-to-noise ratios and variation in measurements from these arrays are not as good as the clone based assays (Carter, 2007). In addition, none of these platforms allow for accurate detection of copy number aberrations using only one probe, and usually a sliding average of three or more adjacent probes are used to reliably call a loss or a gain (Greshock et al., 2007).

While BAC and oligonucleotide arrays typically compare a test and reference sample in the same array (two-channel array), assays based on differences in Single Nucleotide

Polymorphisms (SNPs) offers yet another way to identify copy number changes. The main distinction between SNP arrays and other array CGH is that the former are single-channel assays, based on hybridisation of a single sample to each chip. Affymetrix introduced the Whole-Genome Sampling Analysis (WGSA) technique in 2003 (Kennedy et al., 2003). WGSA use a complexity reduction PCR reaction step prior to hybridisation to an array with SNPs. In essence, total genomic DNA is digested by restriction enzymes and ligated to adaptors that recognise specific overhangs introduced by the restriction endonucleases. PCR using an adaptor sequence primer then amplifies the DNA fragments, which later are fragmented, labelled, and hybridised to the array. Recent genomic arrays from Affymetrix contain over 1.8 million probes for SNP genotyping and CNV analysis.

An alternative genome-wide SNP based CNV approach has been developed by Illumina. The concept of their Infinium Whole Genome Genotyping (WGG) assay is based on direct hybridisation of Whole Genome-Amplified (WGA) genomic DNA to an array of locus-specific 50 base-pairs long oligonucleotide probes. After hybridisation, SNP determination is achieved through an allele specific enzymatic primer extension, incorporating labelled nucleotides. After extension, these labels are visualised by staining with a sandwich-based immunohistochemistry assay that increases the overall sensitivity of the assay (Steemers and Gunderson, 2007). In collaboration with deCODE, Iceland, Illumina has recently developed an array that, in addition to assessing roughly a million SNPs, also targets nearly 9,000 CNV regions throughout the genome not currently available in any public database. Several recent studies comparing CNV platforms indicate that baseline variation is higher and dynamics of copy-number ratios is lower for SNP based arrays (Greshock et al., 2007; Gunnarsson et al., 2008; Hehir-Kwa et al., 2007). The higher variance within these platforms is, however, compensated by the high density of probes.

3 Normalisation and smoothing

Normalisation is an essential part of preprocessing for microarray data, which is intended to adjust the signal intensities to make measurements between different arrays comparable (Smyth and Speed, 2003). Normalisation has been extensively studied for gene expression microarrays, and similar methods are applied to CNV data, although differences do exist between the two (Staaf et al., 2007). The goal of most CNV experiments is to determine if significant copy number differences exist between conditions (for example, affected vs. unaffected; treatment vs. placebo; tumour vs. normal tissue). This is accomplished by comparing the variation between conditions to the variation within conditions. For microarray data, including array CGH and SNP arrays, the sources of this variation may be broadly divided into two categories. These are often referred to as ‘biological’ and ‘technical’ variation (Parrish and Delongchamp, 2006). Bolstad et al. (2005) aptly call this ‘interesting and obscuring’ variation, respectively. The first source is biological differences between conditions, which are usually the quantities of interest to the experimenter. The second source is technical differences between the chips, which may include differences in sample preparation, hybridisation and scanning (Zakharkin et al., 2005). This additional source of variation alters the distribution of the signal intensities, making comparisons across conditions difficult. The purpose of normalisation is to alter the distribution to minimise technical variation, so that valid statistical tests of the remaining biological variation may be performed.

As with expression microarrays, array CGH and SNP arrays may be categorised as single-channel and two-channel as discussed above. The former requires comparison between arrays to determine changes between conditions. This introduces several potential sources of variation between chips, both systematic and random, which must be addressed in the normalisation. The two-channel platform is inherently comparative. This eliminates some of the obscuring variation that is present in the single-channel platform, since the experimental and the control

sample are hybridised to the same array. However, quality control for two-channel arrays that are produced by noncommercial sources may not be as rigorous as commercial one-channel arrays. This may introduce additional sources of variation to the two-channel platform. For example, the quality of array spots often varies by the print tip used in producing the array, which, in turn, requires correction in the normalisation step.

3.1 Normalisation methods

As previously noted, CNV normalisation algorithms for both array CGH and SNP arrays have been derived from the corresponding algorithms used in expression microarrays. These approaches may generally be categorised as global (or scaling) and intensity dependent (or local). Although both can achieve the goal of normalisation – to make arrays similar to each other by eliminating obscuring variation – there are significant differences between the two. This section will review the most commonly used normalisation methods for CNV data and highlight the advantages and disadvantages of each.

3.1.1 Global normalisation—Global normalisation transforms the distribution of the intensities for a target array using an affine function to more closely resemble the distribution for a reference array. Thus, under global normalisation, a constant may be subtracted from all intensities to shift the location of the distribution, and a multiplicative factor is used to adjust the scale of the distribution. If used, the constant that is subtracted is the mean or median intensity for the array so that the intensities are located about a measure of central tendency. Next, a linear regression is fitted for the intensities, and the estimated slope is used as the multiplicative factor for the normalisation of raw intensities.

Let **x** represent a vector of intensities for the reference chip and let **y** represent a vector of intensities for the chip to be normalised. According to standard linear model theory, the least squares estimate for the slope β is given by

$$\beta = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}},$$

where the intercept term has been omitted from the model, as the intensities are assumed to be centred. The normalised intensities $\mathbf{x}_{\text{norm}}$ are equal to the fitted values from the model:

$$\mathbf{x}_{\text{norm}} = \widehat{\beta}\mathbf{x}.$$

The approach used in global normalisation is straightforward and achieves the goal of making the chips more comparable. However, some of the assumptions underlying this method are problematic. As global normalisation is based on a linear transformation, it cannot account for the biological observation that the signal intensities are non-linear with intensity-dependent biases (Yang et al., 2002a, 2002b). It also assumes that DNA from both channels (for two-channel platforms) or both chips (for single-channel platforms) has hybridised equally across the arrays, or that spatial bias is negligible. As noted above, spatial bias appears to be a significant concern in many CGH studies. Because of these disadvantages, global normalisation would not be suitable for most CGH analyses, although it is available as an option in some packages such as CGHPRO (Chen et al., 2005). Global normalisation is also used in the ITALICS package for the analysis of SNP array data (Rigaill et al., 2008).

3.1.2 Loess normalisation—The loess (or lowess) normalisation algorithm utilises the robust locally weighted regression described by Cleveland (1979). This form of regression is

a general method for analysing nonlinear data that relies on the data to specify the form of the model. A *loess curve* is constructed locally so that, for any given point, the fitted value of the function depends on that point and its closest neighbouring values using a specified distance metric. Fitting the curve using only local values serves to reduce the variability present in the data and create a smooth function. As applied to CNV data, the reduction in variability would correspond to removing obscuring variation. The implementation of loess regression is described in Algorithm 1 (Appendix A).

The loess algorithm may be applied to all intensities on an array, as in the CGHPRO software. However, the more common form is to divide the intensities into groups based on a specified characteristic such as print tip or spatial location. Once appropriate groups are defined, a separate normalisation step is performed for each of the groups. This allows for correction of biases that may vary regionally across a chip.

Several software packages implement variants of the loess algorithm. As noted, the CGHPRO program implements global loess normalisation; local loess normalisation may also be applied to user-defined subsets of the intensities. The MANOR package (Neuvial et al., 2006) performs spatial loess normalisation across the two dimensions of the microarray chip, with the neighbourhood corresponding to 3% of the intensities surrounding the probe to be normalised. The Rob-HMM software (Shah et al., 2006) implements a stepwise procedure using loess normalisation, with groupings based on probe intensity (Khojasteh et al., 2005). The snapCGH package (Marioni et al., 2007) from Bioconductor implements local loess normalisation with groups based on print tip; a global loess option is also available.

3.1.3 Invariant set normalisation—A third approach to normalisation is the invariant set method, which attempts to base normalisation only on those probes with normal copy number across chips (Li and Wong, 2001). Such probes, called the invariant set, would be expected to have similar intensities after obscuring variation is removed. Unfortunately, the probes with normal copy number cannot be identified prior to the experiment, so heuristics must be employed to determine the invariant set. Probes in the invariant set would be expected to have similar (but not necessarily identical) intensity-based ranks between two chips, with one chip identified as the reference and the other as the chip to be normalised. A nonlinear function is then fitted between the probes in the invariant set for the two chips. This function is used to map the array to be normalised onto the reference array. A procedure to identify the invariant set is detailed in Algorithm 2 (Appendix A). Invariant set normalisation, as given in this algorithm, is implemented in the dChip software for analysis of CGH data (Zhao et al., 2004).

3.1.4 Quantile normalisation—Another approach to normalisation is quantile normalisation (Bolstad et al., 2003). It is an aggressive form of normalisation that ensures each chip has the same distribution of intensities. Since the arrays within a condition are postulated to have the same distribution if only biological variation is present, the quantile-quantile plot of pairs of chips would be expected to show a diagonal line through the origin with slope 1. Thus a common distribution can be enforced by projecting the pairwise plots onto this 45° line, which effectively removes the obscuring variation. The procedure for performing quantile normalisation is shown in Algorithm 3 (Appendix A). The elements of $\mathbf{Y}_{\text{norm}}$ represent the quantile-normalised intensities that are used in subsequent computations. Quantile normalisation is implemented in the CNAT algorithm within the Affymetrix Genotyping Console Software (Huang et al., 2004; Affymetrix, 2008). It is also available in the oligo package from Bioconductor for SNP data used for the detection of CNVs (Carvalho et al., 2007).

3.2 Smoothing methods

In array CGH, smoothing utilises spatial information of probes along the chromosome to reduce technical variation for individual arrays, especially for the purpose of visualisation. This reduction in technical variation would facilitate comparisons among different arrays. However, unlike normalisation, smoothing does not explicitly make chips comparable and does not address technical variation occurring between arrays. Some authors group smoothing and normalisation methods together, as both achieve a reduction in the variability of the data (Lai et al., 2005). Still, the inability of the former to account for variation between chips makes a distinction between the two useful. Smoothing has been implemented for the analysis of array CGH data but not for the analysis of SNP data.

3.2.1 Moving window approaches—The simplest type of smoothing technique is based on a moving window about the data point of interest, so that the data point is not considered individually but as part of a neighbourhood of values. The size of the window can be varied so that greater or lesser degrees of smoothing are achieved, with larger window sizes giving greater smoothing but less accurate estimates at each individual data point. Once the window size has been selected, the smoothed estimate for each data point is computed as a measure of central tendency (generally the mean or median) of the window about that point. An example of the moving window technique, which uses a symmetric window about a data point, is given in Algorithm 4 (Appendix A).

In general, data points near the beginning and the end of the sequence will have a lesser degree of smoothing than data points in the interior, as there will be fewer points included in the window. A moving window smoother based on means is implemented in the CGH-Explorer program (Lingjaerde et al., 2005) and in the CGH-Miner based on Clustering Along Chromosomes method (Wang et al., 2005), while one based on medians is implemented in the ChARM package (Myers et al., 2004).

3.2.2 Penalised least squares—Another method of smoothing is based on the least squares approximation to a function with appropriate penalties. One such method is the Potts filter (Winkler and Liebscher, 2002) contained in the CGH-Explorer program. Let $Y_i = 1, 2, \ldots, n$ be the n intensity measurements such that $Y_1 \leq Y_2 \leq \ldots \leq Y_n$, and let $\hat{Y}_i$ be the estimate corresponding to Y_i. Then the values of are found by minimising

$$Q_2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 + \lambda \cdot \#(\widehat{Y}_i \neq \widehat{Y}_{i+1})$$

where $\#(\cdot)$ is the count function. The left-hand term in this equation is the familiar least squares criterion, which measures the goodness of fit for the estimate. The right-hand term is a penalty, weighted by λ, for changes in the estimation function. Although the penalised least squares approaches do achieve smoothing of the intensity data, such methods are generally not the best choice for CNV data as they are not well suited to data with discontinuities and abrupt changes (Eilers and de Menezes, 2005).

3.2.3 Quantile smoothing—Quantile smoothing, which is distinct from quantile normalisation, reduces variation in the intensity data by minimising the distance from the data and a fitted function. In contrast to least squares, the distance metric is the absolute value of the difference between the data points and their estimated values, rather than the squared value of the difference. Thus the function to be minimised is

$$Q_1 = \sum_{i=1}^{m} y_i - x_i + \lambda \sum_{i=2}^{m} y_i - y_{i-1},$$

where x_i, $i = 1, 2, \ldots, m$ are the data points and y_i are the fitted values. The first term of the above equation is the distance metric, while the second term is a penalty against changes in the fitted values y. The parameter λ controls the degree of smoothing, with larger values resulting in more severe penalties and a smoother function. This equation does not have a closed form solution. However, it may be solved using quantile regression (Koenker and Bassett, 1978; Portnoy and Koenker, 1997), which employs linear programming techniques to estimate the quantiles of a distribution, similar to the estimation of the mean in ordinary least squares regression. The use of the absolute value of the differences, rather than the square, results in a function with sudden jumps and flat plateaus, which gives a better fit to the theoretical shape of intensities for CNV data. Quantile smoothing is implemented in the R package *quantreg* (Eilers and de Menezes, 2005) available from Bioconductor.

3.2.4 Adaptive weights smoothing—The Adaptive Weights Smoothing algorithm for CNV data (Hupé et al., 2004) is a smoothing algorithm based on local-likelihood modelling (Polzehl and Spokoiny, 2003). It is a modification of the general-purpose Adaptive Weights Smoothing estimation algorithm proposed by Polzehl and Spokoiny (2000). In the version specific to array CGH, the ordered set of n chromosomal locations is defined as $X = \{X_1, X_2, \ldots, X_n\}$ such that $X_1 \leq X_2 \leq \ldots \leq X_n$ with the corresponding intensity measurements $Y = \{Y_1, Y_2, \ldots, Y_n\}$. It is assumed that each Y_i, $i = 1, 2, \ldots, n$ is a function of X_i through the parameter θ such that

$$Y_i = \theta(X_i) + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. The set of chromosomal locations X is partitioned into m distinct subsets $X_1, X_2, \ldots, X_m$ such that $\cup_{i=1}^{m} \mathrm{X}_i = \mathrm{X}$ and $\mathrm{X}_i \cap \mathrm{X}_j = \varphi$ for all $i \neq j$. The function θ is assumed to be of the form $\theta(x) = \sum_{i=1}^{m} a_m 1(x \in \mathrm{X}_m)$, where $(1(\cdot)$ is the indicator function. Thus all intensities in the subset X_i are smoothed with the estimated value a_i. The subsets $X_1, X_2, \ldots, X_m$, the smoothed values $a_1, a_2, \ldots, a_m$, and the number of subsets m are unknown but estimated from the data.

The modified Adaptive Weights Smoothing algorithm for CNV data is used to estimate the parameters. For each point X_i, an iterative procedure is used to find the largest neighbourhood about X_i for which the parameter θ can be reasonably fitted as a constant. Details of this algorithm are presented in Algorithm 5 (Appendix A). The Adaptive Weights Smoothing algorithm is implemented in the GLAD (Gain and Loss Analysis of DNA) package available from Bioconductor.

3.2.5 Wavelet smoothing—Wavelet smoothing (Hsu et al., 2005) is based on the theory of wavelets, which are a class of functions in mathematics that can be used as the basis of transformations of other functions. In this sense, wavelets share similarities to the familiar Fourier transform, but the former are more general in that wavelets incorporate two domains, scale and time. Wavelets also offer advantages over traditional Fourier methods in the representation of functions with discontinuities and sharp spikes, which are characteristic of CNV data. Wavelet analysis first defines a prototype function $\psi(x)$, called the *mother wavelet*. Next, a series of basis functions for wavelet analysis are defined as dilatations and

translations of the mother wavelet. Different values for the dilatation and translation among the basis functions determine whether the function emphasises overall patterns or fine details of the data. The target function can be represented as a linear combination of the wavelet basis functions. Data manipulation may then be performed using only the wavelet coefficients, which can simplify computations considerably. Back-transformations are used to recover the results of the data manipulations on the original function.

In wavelet smoothing, exemplified by the *waveslim* package from Bioconductor, the two domains used are the location and the scale of the change in copy number. The Haar step function is chosen as the mother wavelet function; it is defined by

$$\psi(u)=\begin{cases} -\frac{1}{\sqrt{2}}, & -1<u\le 0,\\ \frac{1}{\sqrt{2}}, & 0<u\le 1,\\ 0, & \text{elsewhere}. \end{cases}$$

Any continuous function can be approximated by Haar functions. The wavelet basis functions are given by the Maximal Overlap Discrete Wavelet Transform (MODWT), with dilatations and translations of the form $\psi_{j,t}(u) = 2^{j/2}\psi(2^j u - t)$, where $j \in Z^+$ is the scale index and $t \in R$ is the location index. The target function f can then be represented as a linear combination of Haar wavelets with the wavelet coefficients α given by

$$\alpha_{j,t}=\int \psi_{j,t}(u) f(u) \mathrm{d}u.$$

The wavelet coefficients α can be conceptualised as average signal intensity at a location. Regions with no CNV would be expected to have a coefficient of zero, while a breakpoint would lead to regions with a nonzero coefficient. Smoothing is accomplished by setting thresholds for the wavelet coefficients, with values of α below the threshold assumed to be noise and set to 0, while values of α greater than the threshold assumed to be signal and retained. Thus, the smoothing process under the wavelet transformation becomes similar to the variable selection process in multiple regression. Finally, back-transformation yields a smoothed set of data on the scale of the original intensity measurements.

4 Calling algorithm

A CNV calling algorithm is designed to take normalised intensity measurements from array CGH, oligonucleotide array, or SNP array data and infer the copy number state at each locus; essentially transforming noisy, continuous data into discrete calls of normal, gain, or loss of copy number. CNVs typically span multiple measured loci. Thus, a calling algorithm should take into account the spatial dependence of copy number data, as adjacent loci will tend to share the same copy number state. Over the past five years or so, there has grown an extensive literature on CNV calling. Among the multitude of algorithms that have been published, however, most can be placed in one of two broad categories: change-point methods and Hidden Markov Models (HMMs).

4.1 Change-point methods

Given an ordered sequence of random data $Y_1, \ldots, Y_n$, a change-point is formally defined as a location i for which the marginal probability distribution of Y_i differs from that of Y_{i+1}. One is typically interested in changes in mean, although there may also be changes in variance or other features of the distribution. In the context of CNV calling, the data $Y_1, \ldots, Y_n$ are

normalised intensity measurements, ordered by physical location along a chromosome. A change-point represents a locus at which the underlying copy number changes; in other words, the beginning or end of a segmental deletion or duplication. Note that when we refer to physical location, it is in comparison to a reference genome map. Mapping the true physical location of copy number gains is beyond the capabilities of CNV-measuring arrays.

Among change-point methods, perhaps the best-known is the Circular Binary Segmentation (CBS) algorithm of Olshen et al. (2004). The CBS algorithm partitions a chromosome into segments of constant copy number by iteratively searching for splits of chromosomal regions, using a change-point test of Sen and Srivastava (1975). When a change-point is located, the algorithm then searches the two resulting segments for additional change-points. The algorithm stops when no more statistically significant change-points are located, according to a heuristically chosen significance threshold. A potential drawback is that CBS divides a chromosome into segments, but it does not classify the resulting segments into groups of like means; therefore, classifying a segment as normal, deletion, or duplication requires a subsequent analysis step. CBS requires a number of computations that grows quadratically with the number of probes measured. This makes CBS a computationally intensive algorithm when analysing data from popular high-density platforms. Understanding this limitation, Venkatraman and Olshen (2007) revised the algorithm to detect the occurrence of change-points as well as provide a stopping criterion once there is strong evidence of a change in copy number. These modifications drastically reduced the number of computations while having a negligible loss of precision. The algorithm by Venkatraman and Olshen (2007) is implemented in DNAcopy found in Bioconductor.

Another change-point method was developed by Picard et al. (2005) and later given the name CGH segmentation (CGHseg). For a given number of segments K, the algorithm locates the $K-1$ change-points using maximum likelihood, assuming homogeneous Gaussian measurement error of intensity log-ratios. The number of segments is estimated with a penalised log-likelihood, where the penalty constant is chosen using a novel adaptive method. While CBS frames its stopping criterion in terms of hypothesis testing, CGHseg frames its stopping criterion in terms of model selection. CGHseg suffers from the same computational problems as CBS when analysing high-density arrays. However, in a review of early CNV calling algorithms Lai et al. (2005) found both CBS and CGHseg to perform consistently well.

4.2 Hidden Markov Models

As an alternative to change-point methods, HMMs are popular in CNV calling because they model the spatial dependence of array data, lead to computationally efficient implementations based on forward-backward algorithms, and can be approached quite naturally in a Bayesian context leading, in turn, to easily interpretable posterior distributions. In an HMM, the observed data $Y_1, \ldots, Y_n$ are modelled as noisy emissions of true, but unobserved, states $\theta_1, \ldots, \theta_n$. For example, let Y_i be the normalised intensity measurement of the ith probe. Then θ_i may take the value 0 (normal), +1 (duplication), or −1 (deletion). There are two distributional assumptions in an HMM:

- at any locus i, Y_i is conditionally independent of all other data given the underlying state θ_I
- the underlying states $\theta_1, \ldots, \theta_n$ follow a Markov process.

In a Markov process, the probability distribution of the state θ_i is conditionally independent of all other states, given the adjacent states θ_{I-1} and θ_{i+1}. This property is colloquially referred to as the ‘memoryless’ property.

An early application of HMM to copy number assays was brought forth by Fridlyand et al. (2004). Their model is unsupervised in the sense that the number of underlying copy number states, the mean intensity ratios for each state, and the Markov transition probabilities are all estimated from data. For computational purposes, the number of states is capped at five and is chosen to minimise a penalised log-likelihood statistic with measurement errors of log-ratios assumed to be homogeneous Gaussian within a given array. Another approach was developed by Broët and Richardson (2006) using a nearest neighbour Markov random field model. While not explicitly referred to as an HMM, it can be viewed as a hierarchical Bayes extension of HMM because the distribution of the spatial latent variables (i.e., copy number state) depends only on the given locus' neighbours.

The primary difference between HMM approaches and the change-point methods described above is that the former includes an explicitly probabilistic model for copy number transitions. In particular, the Markovian memoryless property implies that copy number change-points follow a homogeneous Poisson process, which almost certainly is an oversimplification of biological reality. In the review paper of Lai et al. (2005), change-point algorithms showed better Receiver Operating Characteristic (ROC) curves than the HMM of Fridlyand et al. (2004) when applied to simulated data under a range of scenarios. There is no consensus, however, that change-point algorithms perform better than all HMMs under all circumstances. More heterogeneous data can be modelled in an HMM framework by increasing the complexity of the underlying state space. Shah et al. (2006), for example, augment a four-state (neutral, loss, single gain, multiple gain) HMM with a fifth state: outlier. In their terminology, an outlier locus is essentially a singleton locus with much higher variance from zero than expected under the null model. Whether the 'outlier' loci represent an underlying biological phenomenon (microdeletions), or are merely an artifact of the measuring device, is beyond the scope of their study. Another extension of the basic discrete homogeneous Markov process is to include the physical proximity of probes in the transition model. The Segmental Maximum a Posteriori (SMAP) approach of Andersson et al. (2008), for example, incorporates both the physical distance between BAC probes, and the overlap between probes where appropriate, into the transition model of a discrete HMM. As in all statistical modelling, the sophistication and complexity of an HMM has to be balanced against identifiability and computational efficiency.

4.3 SNP array methods

Recently there has been increased interest in inferring copy number from SNP arrays due to their widespread use in genome-wide association studies and significantly greater probe density. This technological shift has required the development of algorithms designed for the specific characteristics of SNP array data. As an example, the QuantiSNP algorithm (Colella et al., 2007) is a Bayesian HMM approach specifically geared toward the Illumina BeadArray™ SNP genotyping platform. BeadArray data consists of two statistics associated with each locus represented on an array: a normalised Log R ratio and B allele frequency. The former is a measure of the overall luminosity of a SNPs probe set, and the latter is a ratio of the probes pertaining to one (B) allele in comparison to the alternative (A) allele. These statistics can be combined to infer copy number and detect loss of heterozygosity, and are therefore modelled by QuantiSNP. The algorithm thus uses both genotype and copy number estimates from the SNP arrays to infer copy number state. Scharpf et al. (2008) have introduced a HMM approach that not only uses genotype and copy number estimates, but uncertainty measures of these estimates as well, to infer the underlying state.

The PennCNV (Wang et al., 2007) algorithm follows an approach similar to QuantiSNP, but also allows the inclusion of pedigree information to inform the location of CNVs. A multi-sample approach on unrelated individuals may also help investigators distinguish between heritable CNP and non-heritable de novo variants (see, for example, Wang et al., 2008). Thus,

analysing multiple samples simultaneously can not only improve the accuracy of CNV inference, but can also allow for the detection of additional biological information of interest. We believe that the treatment of CNV data from multiple sources, whether from pedigrees or unrelated individuals, is still in an early stage of development.

5 Association analysis

After normalisation of the raw data, there are two main ways in which CNVs have been used in association studies for disease. These are using discrete calls from the raw data (i.e., gain, loss, neutral) which is performed using any of calling algorithms previously discussed (as in Marshall et al., 2008) or analysing the normalised log R intensity as a continuous variable, similar to Ionita-Laza et al. (2008) and Feuk et al. (2006). Both designations have been used successfully; however there are advantages and disadvantages of both. For the discrete CNV calls, also known as CNV genotypes, an advantage is that existing statistical methods for genetic association studies have already been developed (McCarroll and Altshuler, 2007). One of the major problems with this approach is the potential misclassification and loss of information since the raw data is continuous in nature and the calling algorithms create discrete categories. In response to the potential for misclassification based on CNV calls, investigators have begun using the normalised log R intensities, disregarding the application of a calling algorithm (Ionita-Laza et al., 2008). The drawback of using the log R intensities is that any significant CNV association has to then be re-analysed using a CNV genotype in order to assess the biological effect of the CNV on the phenotype.

5.1 Study design

The majority of genetic association studies to date have been case-control studies for both GWAS (Manolio et al., 2008) and CNV (Lachman et al., 2007; Walsh et al., 2008; Stefansson et al., 2008). In a case-control study, one selects cases based on the presence of the phenotype of interest and controls based on the absence of the phenotype from the same underlying population. Allelic and genotypic differences are then compared between cases and controls using a standard χ^2 test (Pearson and Manolio, 2008). This can also be extended to logistic models where one is able to control for other potential confounding variables (Lewis, 2002). This methodology is applicable if one is dealing with CNV calls or genotypes and has been successfully applied in Bipolar Disorder (Lachman et al., 2007), Schizophrenia (Walsh et al., 2008; Stefansson et al., 2008) and Autism (Sebat et al., 2007). Case-control studies are advantageous because they employ a short time frame, include a large sample size and perform well when studying rare diseases. However, these studies are prone to measurement bias and confounding due to population stratification which can be addressed by using existing methods such as Eigenstrat (Price et al., 2006), genomic control (Devlin and Roeder, 1999), and structured association testing (Structure, Pritchard et al., 2000). Also, typically the most severe or fatal forms of the phenotype of interest are excluded from the study in favour of the more prevalent cases. Investigators should also be cautious when reporting the relative risk for common disease as this statistic is often over-estimated. In this situation an odds ratio is more appropriate (Pearson and Manolio, 2008).

The other main type of association study is a family-based association study. For a family-based study, the initial identified subject displaying the phenotype of interest is considered the proband. Once she agrees to be in your study, the investigators enroll as many family members as possible. Here, each proband is considered a case and other affected family members may or may not be cases. The cases (probands) are then compared to the controls (unaffected family members) similar to the case control study design (Marshall et al., 2008). A family-based study has advantages over a case-control study because there is usually less concern about population stratification and the environment within the family is inherently controlled for, thereby

allowing one to assume that any association detected is due to the effect of the genetic variant. Furthermore, family-based studies allow one to investigate transmission of alleles from parent to offspring, and in the case of CNVs would allow for an assessment of *de novo* or inherited CNV detection as performed by Sebat et al. (2007) for Autism. Of concern for family studies is the difficulty in assembling the families willing to participate. Additionally, family studies are very sensitive to genotyping errors (Pearson and Manolio, 2008).

Another study design, the cohort study, has recently entered the playing field of genetic association studies. Cohort studies tend to require more resources as the prospective studies follow individuals over time and examine the incidence of the phenotype during that time period. Cohort studies offer the advantage of a direct estimate of risk, and usually cover the whole variation of the phenotype. Drawbacks of the cohort study are the need for large sample sizes for rare diseases and the expensive and lengthy follow-up. There are some well-known cohorts such as the Framingham Heart Study that collected DNA for genetic studies and is now being used to perform secondary genetic association studies within the cohort. When using these existing cohorts, secondary consent may be an issue. This may be time consuming and laborious and should be considered when planning the analysis (Pearson and Manolio, 2008). The majority of secondary analyses from existing cohort studies take the form of a nested case-control study.

Disease phenotypes for genetic association studies may take any of the following forms: dichotomous, ordinal, or continuous. The most important aspect of phenotype measurement is that it is measured as precisely as possible with validated, reliable and reproducible instruments (Wojczynski and Tiwari, 2008). Association methods for various definitions of phenotype exist for analysis with SNP data and are easily extended to CNV genotypes (McCarroll and Altshuler, 2007). Some additional methodological work may be necessary for extensions to continuous CNV intensities (Yang et al., 2007), and initial work has been successful for family-based studies of asthma (Ionita-Laza et al., 2008).

5.2 Analysis

CNV Genotypes—Depending on the phenotype/disease of interest and how it is reported in the data (i.e., discrete or continuous), there are two common types of analyses, either an analysis of variance (ANOVA) or a χ^2 test with $n - 1$ degrees of freedom, where n is the number of classifications of CNV genotype (Neter et al., 1996; Yang et al., 2007). More complex linear models that control for confounding variables as covariates may also be considered. These analyses are simple extensions of prior association analysis methodologies that were used for SNP genotypes, and can be used in any study design previously discussed (Hubacek et al., 2008). This approach has already proven successful for CNV association studies using both family and population studies for Autism Spectrum Disorder (Marshall et al., 2008; Sebat et al., 2007), Schizophrenia (Stefansson et al., 2008; Walsh et al., 2008) and Crohn's Disease (McCarroll et al., 2008).

Normalised log R intensities—Investigators may choose to forgo the application of a calling algorithm, and instead use the continuous normalised log R intensity data in their analysis. In this scenario, simple linear regression or a logistic regression is appropriate when the phenotype is continuous or discrete, respectively. Likewise, it is possible to address confounding by adding additional covariates to the model and can be performed using data from any of the study designs mentioned above. The Family-Based Association Test (FBAT) software has already been extended to deal with family-based CNV association studies using continuous log *R* intensities (Ionita-Laza et al., 2008). Summary of analytic methods to detect association between phenotype and CNVs are depicted in Table 1.

Statistical Issues for Genome-Wide Data—Statistical analysis of genome-wide data requires correction for population stratification and multiple testing. Population stratification may occur when analysing data that includes subjects of mixed ancestry. Differences in outcome and allele frequencies among ancestral populations may cause spurious associations; thus must be controlled for. Common approaches to correct for population stratification include genomic control (Devlin and Roeder, 1999), Structure (Pritchard et al., 2000) and Eigenstrat (Price et al., 2006).

Analysis of genome-wide data requires correction for multiple testing to maintain an overall type 1 error rate. An often used, yet naive way to address multiple comparisons is to use the Bonferroni correction (Yang et al., 2005) where a nominal significance level (usually 0.05) is divided by the number of variants tested (k) to obtain a new significance threshold (0.05/k) to which each test is compared (Pearson and Manolio, 2008). The Bonferroni correction is commonly used. However it is generally too conservative because it assumes that all the variants being tested are independent. In typical SNP association studies linkage disequilibrium between SNPs dictates that all tests are not independent. The extent of disequilibrium between common CNVs is not well known. However, more recent methods have been developed to correct for multiple testing that do not assume independence among tests. Such methods include false discovery rate (Hochberg and Benjamini, 1990) and Bayes factor analysis that was used in the latest Wellcome Trust genome-wide association study (2007).

6 Discussion

We have discussed the options available to investigators who wish to study CNV data in the context of an association study. We break these steps down into the choice of genotyping platform, normalisation, calling algorithm, and statistical analysis. There are, however, further aspects to consider.

The limits of detection are naturally of interest in any experiment. For expression microarrays and CNV arrays, both the relative change in DNA concentration between conditions, or fold change and the absolute concentration of DNA to be analysed must exceed certain thresholds. For expression microarrays, a 2-fold change in DNA concentration between conditions is generally seen as the minimum necessary, although lower fold changes may be detected in some circumstances (Sokolov et al., 2006). However, the fold changes expected with CNV arrays may be much lower (Fridlyand et al., 2004; Pinkel and Albertson, 2005). For example, a single copy gain in a homogenous cell population would lead to a 1.5-fold change in the DNA concentration for the gene of interest. These values may be even smaller if the sample also contains cells of normal copy number. Additionally, if the absolute concentration of DNA falls below a certain amount, the signal cannot be distinguished from background noise. Detection limits are obviously important in calling algorithms but also must be considered in normalisation. The normalisation process reduces variation between chips, and doing so may remove interesting variation as well as obscuring variation. This must be considered in the design and evaluation of any normalisation algorithm to avoid excessive false negatives.

The samples used for hybridisation onto CNV arrays may also vary considerably in the degree of CNV. Occasionally, the variation is so extreme that it is difficult to define a ‘normal’ copy number for properly carrying out the normalisation to make chips comparable (Pinkel and Albertson, 2005). A similar problem could occur in expression microarrays, where normalisation would ideally be carried out using an invariant set of genes that show no differential expression between conditions, but does not appear to be a common occurrence (Li and Wong, 2001). In less extreme cases, the normalisation process may still be inaccurate when the majority of probes are targeted to sequences that exhibit copy number differences between conditions. This problem has been noted with published array CGH data but it is

currently unclear how often it arises (Staaf et al., 2007). A similar potential problem has been described for expression microarray data but is thought to happen infrequently (Oshlack et al., 2007).

One of the systematic biases encountered in array CGH data is spatial bias, in which the intensities vary by location on the array. If it occurs, a spatial bias may lead to erroneous declarations of CNV, based on the spatial position of the probes of interest. However, proper normalisation can minimise the bias if it is recognised. Spatial biases have been reported in several (Chen et al., 2005; Neuvial et al., 2006; Marioni et al., 2007), but not all (Fridlyand et al., 2004), studies of array CGH data. They have also been observed in studies utilising SNP microarrays (Rigaill et al., 2008). Spatial biases have been noted in expression array studies (Reimers and Weinstein, 2005), but a specific spatial normalisation is usually not performed. The causes of spatial bias in array CGH and SNP arrays are currently unclear, as is the frequency of its occurrence. Marioni et al. (2007) noted that spatial bias for array CGH is correlated with the GC content of the region, but may not be causative as substitution of dUTP for dCTP in the labelling reaction did not alter the spatial artifact. As GC content is related to a number of genomic characteristics, this correlation may simply be an incidental finding.

A potential source of sampling error in genome-wide association studies is confounding due to population stratification and admixture. A number of different methods have been developed to control for population structure in association studies, including the TDT and its many variations (Spielman et al., 1993; see, Tiwari et al., 2008 for a detailed review) as well as more modern methods such as regional admixture mapping and structured association testing (see, Redden et al., 2006) However, very few methods have been developed that control for population structure in CNV studies. Much of this may be due to the fact that only recently has the extent of CNV between populations been studied. White et al. (2007) examined CNV frequencies in twelve CNVRs, in five distinct populations and found differences in the distribution of CNVs between them. Jakobsson et al. (2008) expanded their study to include 405 individuals from the Human Genome Diversity Project (Cann et al., 2002). These samples represented 29 populations throughout the world from each of the major geographic regions. Jakobsson et al. found the vast majority of CNVs to have low frequencies worldwide. However, higher frequencies of some CNVs were found within geographic regions – particularly in Oceania and the Americas. These findings provide evidence that CNV differs among populations. Therefore, any CNV association study must control for population structure or risk confounding. Most of the methods developed for SNP data to correct for structure should be applicable to CNV data.

As we presented, investigators have the option of invoking a calling algorithm on normalised intensities to gain biological meaning of CNV data, or simply treating the normalised intensity data as the independent variable. SNP genotype data, like CNVs, can be presented as allele calls or as a continuous variable. Advocates for the use of continuous data cite the ambiguity of SNP genotyping calls. We imagine this argument would synonymously carry over to CNV calling. However, to date there has been no published research concerning when it is more appropriate to use one method over the other.

CNV data offers a challenge to investigators as there are a number of analytic steps necessary to achieve reliable copy number measurements. We have presented these steps and the various options available. The inclusion of CNV-targeting probes on high-density SNP arrays allow for the combined examination of these two genetic variants. Going forward, all genome-wide association studies should be prepared for joint analysis of SNPs and CNVs.

Software

There are various software programs applicable for SNP arrays to perform normalisation, smoothing, and calling CNVs. They come in all sorts of flavours from type of statistical model used to type of operating system required to execute the program. Table 2 provides a short list of programs available along with some of their features and URLs (the list is not exhaustive by any means).

Acknowledgments

This study is supported in part by R21LM008791, R01DK074842, R01HL055673, R01DK52431, T32HL079888, T32HL072757.

Biographies

Nathan E. Wineinger received his BS Degree from Grinnell College. He is currently pursuing his PhD Degree in the Department of Biostatistics at the School of Public Health at the University of Alabama at Birmingham. His research focuses on developing statistical methods to model Copy Number Variation data. He is currently funded by NIH T32 Pre-Doctoral Training grant.

Richard E. Kennedy received his MD Degree from the University of Mississippi Medical Centre and his PhD Degree from Virginia Commonwealth University. He is a post-doctoral fellow in the Section on Statistical Genetics at the University of Alabama at Birmingham's Department of Biostatistics. He recently completed an NIH F37 Individual Biomedical Informatics Training grant and is currently funded under an NIH T32 Post-Doctoral Training grant. He interests are in the statistical modelling of microarray data (particularly using multivariate or mixed effects models), quality control in microarray data, statistical computing, and linear models.

Stephen W. Erickson received his PhD in Statistics from the University of California, Los Angeles and is a post-doctoral fellow in the Section on Statistical Genetics at the University of Alabama at Birmingham's Department of Biostatistics. His research has focused on Bayesian analysis of genomic datasets, and he is currently developing a multilevel hidden Markov model of Copy Number Variation. He is currently funded under NIH T32 Post-Doctoral Training grant.

Mary K. Wojczynski received her PhD in Epidemiology from the University of North Carolina at Chapel Hill. She is a post-doctoral fellow in the Section on Statistical Genetics at the University of Alabama at Birmingham's Department of Biostatistics. Her current research focuses upon implementing methods employed in genetic analyses into epidemiological studies, and expanding methodologic research into how to work with missing genetic data and issues related to multiple testing.

Carl E. Bruder received his PhD in Molecular Medicine from the Karolinska Institutet, Sweden. He was a manager of the Microarray Core Facility at the University of Alabama at Birmingham before joining the Southern Research Institute in Birmingham as a project manager, where he assists in planning and executing the development of models for antiviral compound and vaccine testing.

Hemant K. Tiwari received his PhD in Mathematics from the University of Notre Dame at South Bend, Indiana. He is an associate professor in the department of biostatistics at University of Alabama at Birmingham. His research interests include developing methods for genetic linkage/association analysis, disequilibrium mapping, genome-wide association analysis of

SNP data and copy number polymorphisms, admixture mapping, molecular evolution, and bioinformatics.

References

Affymetrix (2008) Affymetrix® Genotyping Console™ Software.

Andersson R, Bruder CE, Piotrowski A, Menzel U, Nord H, Sandgren J, Hvidsten TR, Diaz de Ståhl T, Dumanski JP, Komorowski JA. Segmental maximum a posteriori approach to genome-wide copy number profiling. Bioinformatics 2008 15 March;24(6):751–758. [PubMed: 18204059]

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003 22 January;19(2):185–193. [PubMed: 12538238]

Bolstad, BM.; Irizarry, RA.; Gautie, L.; Wu, Z. Preprocessing high-density oligonucleotide arrays. In: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W., editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer; New York: 2005. p. 13-32.

Broët P, Richardson S. Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. Bioinformatics 2006 15 April;22(8):911–918. [PubMed: 16455750]

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, de Ståhl TD, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommenm GJ, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, Dumanski JP. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 2008 March;82(3):763–771. [PubMed: 18304490]

Buckley PG, Mantripragada KK, Díaz de Ståhl T, Piotrowski A, Hansson CM, Kiss H, Vetrie D, Ernberg IT, Nordenskjöld M, Bolund L, Sainio M, Rouleau GA, Niimura M, Wallace AJ, Evans DG, Grigelionis G, Menzel U, Dumanski JP. Identification of genetic aberrations on chromosome 22 outside the NF2 locus in schwannomatosis and neurofibromatosis type 2. Hum Mutat 2005 December; 26(6):540–549. [PubMed: 16287142]

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science 2002 12 April;296(5566):261, 262. [PubMed: 11954565]

Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 2007 July;39(7 Suppl):S16–21. [PubMed: 17597776]

Carvalho B, Bengtsson H, Speed TP, Irizarry RA. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 2007 April;8(2):485–499. [PubMed: 17189563]

Chen W, Erdogan F, Ropers HH, Lenzner S, Ullmann R. CGHPRO – a comprehensive data analysis tool for array CGH. BMC Bioinformatics 2005 April;5(6):85. [PubMed: 15807904]

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Amer Statist Assoc 1979;74(368):829–836.

Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 1988;83(403):596–610.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35(6):2013–2025. [PubMed: 17341461]

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nature Genet 2006;38:75–81. [PubMed: 16327808]

de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, Tsang P, Ben-Dor A, Yakhini Z, Ellis RJ, Bruhn L, Laderman S, Froguel P, Blakemore AI. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum Mol Genet 2007 1 December;16(23):2783–2794. [PubMed: 17666407]

de Ståhl TD, Sandgren J, Piotrowski A, Nord H, Andersson R, Menzel U, Bogdan A, Thuresson AC, Poplawski A, von Tell D, Hansson CM, Elshafie AI, Elghazali G, Imreh S, Nordenskjöld M, Upadhyaya M, Komorowski J, Bruder CE, Dumanski JP. Profiling of Copy Number Variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. Hum Mutat 2008 March;29(3):398–408. [PubMed: 18058796]

Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999 December;1(4):369–387.

Eilers PH, de Menezes RX. Quantile smoothing of array CGH data. Bioinformatics 2005 1 April;21(7): 1146–1153. [PubMed: 15572474]

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics 2006;7:85–97.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C. Copy number variation: new insights in genome diversity. Genome Res 2006 August;16(8):949–961. [PubMed: 16809666]

Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. J Multivariate Anal 2004;90(1):132–153.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005 4 March;307(5714):1434–1440. [PubMed: 15637236]

Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, Nathanson K, Protopopov A, Weber BL, Chin L. A comparison of DNA copy number profiling platforms. Cancer Res 2007 1 November;67(21): 10173–10180. [PubMed: 17968032]

Gunnarsson R, Staaf J, Jansson M, Ottesen AM, Göransson H, Liljedahl U, Ralfkiaer U, Mansouri M, Buhl AM, Smedby KE, Hjalgrim H, Syvänen AC, Borg A, Isaksson A, Jurlander J, Juliusson G, Rosenquist R. Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukaemia–a comparative study of four differently designed, high resolution microarray platforms. Genes Chromosomes Cancer 2008 August;47(8):697–711. [PubMed: 18484635]

Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA. Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res 2007 28 February;14(1):281–11.

Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9 (7):811–818. [PubMed: 2218183]

Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, Porter P. Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 2005 April;6(2):211–226. [PubMed: 15772101]

Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 2004 May;1(4):287–299. [PubMed: 15588488]

Hubacek JA, Wang WW, Skodová Z, Adámková V, Vráblík M, Horínek A, Stulc T, Ceska R, Talmud PJ. APOA5 Ala315>Val, identified in patients with severe hypertriglyceridemia, is a common mutation with no major effects on plasma lipid levels. Clin Chem Lab Med 2008;46(6):773–777. [PubMed: 18601597]

Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 2004 12 December;20(18):3413–3422. [PubMed: 15381628]

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet 2004 September;36(9):949–951. [PubMed: 15286789]

Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, Laird NM, Weiss ST, Lange C. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol 2008 April;32(3):273–284. [PubMed: 18228561]

International HapMap Consortium. A haplotype map of the human genome. Nature 2005 October;437 (7063):1207–1396.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 2008 21 February;451(7181):998–1003. [PubMed: 18288195]

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SP, Jones KW. Large-scale genotyping of complex DNA. Nat Biotechnol 2003 October;21(10):1233–1237. [PubMed: 12960966]

Khojasteh M, Lam WL, Ward RK, MacAulay C. A stepwise framework for the normalization of array CGH data. BMC Bioinformatics 2005 November;18(6):274. [PubMed: 16297240]

Koenker R, Bassett G Jr. Regression quantiles. Econometrica: Journal of the Econometric Society 1978;46:33–50.

Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, Novak T, Papolos DF, Stopkova P. Increase in *GSK3β* gene copy number variation in bipolar disorder. Am J Med Genet Part B 2007;144B:259–265. [PubMed: 17357145]

Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 2005 1 October;21(19):3763–3770. [PubMed: 16081473]

Le Maréchal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Férec C. Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 2006 December;38(12):1372–1374. [PubMed: 17072318]

Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 2006;52:103–121. [PubMed: 17015230]

Lewis CM. Genetic association studies: design, analysis and interpretation. Briefings in Bioinformatics 2002;3(2):146–153. [PubMed: 12139434]

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2001;2(8):0032.1–0032.11.research

Lingjaerde OC, Baumbusch LO, Liestøl K, Glad IK, Børresen-Dale AL. CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics 2005 March;21(6):821–822. [PubMed: 15531610]

Lupski JR. Structural variation in the human genome. N Engl J Med 2007 15 March;356(11):1169–1171. [PubMed: 17360997]

Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clinical Investigation 2008;118(5):1590–1605.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavaré S, Hurles ME. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Gen Biol 2007;8:R228.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk Lk, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbi V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, Scherer SW. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet 2008;82:477–488. [PubMed: 18252227]

McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nature Genetics 2007;39(7 Suppl):S37–S42. [PubMed: 17597780]

McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ. Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. Nature Genetics 2008;40(9):1107–1112. [PubMed: 19165925]

Myers CL, Dunham MJ, Kung SY, Troyanskaya OG. Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics 2004 12 December;20(18):3533–3543. [PubMed: 15284100]

Neter, J.; Kutner, MH.; Nachtsheim, CJ.; Wasserman, W. Applied Linear Statistical Models. Vol. 4. WCB McGraw-Hill; Boston, MA: 1996.

Neuvial P, Hupé P, Brito I, Liva S, Manié E, Brennetot C, Radvanyi F, Aurias A, Barillot E. Spatial normalization of array-CGH data. BMC Bioinformatics 2006 May;22(7):264. [PubMed: 16716215]

Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004 October;5(4):557–572. [PubMed: 15475419]

Oshlack A, Emslie D, Corcoran LM, Smyth GK. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. Genome Biol 2007;8(1):R2. [PubMed: 17204140]

Parrish, RS.; Delongchamp, RR. Normalization of Microarray Data. Vol. Chapter 2. Chapman & Hall/CRC; Boca Raton, FL: 2006. p. 9-28.

Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008;299(11):1335–1344. [PubMed: 18349094]

Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005 February;11(6):27. [PubMed: 15705208]

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998 October;20(2):207–211. [PubMed: 9771718]

Pinkel D, Albertson DG. Comparative genomic hybridization. Annu Rev Genomics Hum Genet 2005;6:331–354. [PubMed: 16124865]

Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet 2007 15 October;16:R168–73. [PubMed: 17911159]Spec No. 2

Polzehl, J.; Spokoiny, VG. Local Likelihood Modeling by Adaptive Weights Smoothing. 2003. Technical Report, WAIS-Preprint 787

Polzehl J, Spokoiny VG. Adaptive weights smoothing with applications to image restoration. J Roy Statist Soc Ser B 2000;62(2):335–354.

Portnoy S, Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. Statist Sci 1997;12(4):279–300.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000 June;155(2):945–959. [PubMed: 10835412]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–909. [PubMed: 16862161]

Redden DT, Divers J, Vaughan LK, Tiwari HK, Beasley TM, Fernández JR, Kimberly RP, Feng R, Padilla MA, Liu N, Miller MB, Allison DB. Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. PLoS Genet 2006 25 August;2(8):e137. [PubMed: 16934005]

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature 2006 November;23(4447118):444–454. [PubMed: 17122850]

Reimers M, Weinstein JN. Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. BMC Bioinformatics 2005 July;1(6):166. [PubMed: 15992406]

Rigaill G, Hupé P, Almeida A, La Rosa P, Meyniel JP, Decraene C, Barillot E. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics 2008 15 March;24(6):768–774. [PubMed: 18252739]

Scharpf, RB.; Parmigiani, G.; Pevsner, J.; Ruczinski, I. Hidden Markov Models for the Assessment of Chromosomal Alterations using High-Throughput SNP Arrays. 2008. Preprint http://www.biostat.jhsph.edu/~iruczins/software/manuscript.pdf

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong association of de novo copy number mutations with autism. Science 2007 20 April;316(5823):445–449. [PubMed: 17363630]

Sen A, Srivastava MS. On tests for detecting changes in mean. Ann Statist 1975;3:98–108.

Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 2006 15 July;22 (14):e431–e439. [PubMed: 16873504]

Smyth GK, Speed T. Normalization of cDNA microarray data. Methods 2003 December;31(4):265–273. [PubMed: 14597310]

Snijders AM, Nowee ME, Fridlyand J, Piek JM, Dorsman JC, Jain AN, Pinkel D, van Diest PJ, Verheijen RH, Albertson DG. Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma. Oncogene 2003 3 July;22(27):4281–4286. [PubMed: 12833150]

Sokolov MV, Smirnova NA, Camerini-Otero RD, Neumann RD, Panyutin IG. Microarray analysis of differentially expressed genes after exposure of normal human fibroblasts to ionizing radiation from an external source and from DNA-incorporated iodine-125 radionuclide. Gene 2006 1 November; 382:47–56. [PubMed: 16876969]

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 1997 December;20(4):399–407. [PubMed: 9408757]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993 March;52(3):506–516. [PubMed: 8447318]

Staaf J, Jönsson G, Ringnér M, Vallon-Christersson J. Normalization of array-CGH data: influence of copy number imbalances. BMC Genomics 2007 October;22(8):382. [PubMed: 17953745]

Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2007 January;2(1):41–49. [PubMed: 17225249]

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonngvist J, Suvisaari J, Tuulio-Henriksoon A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. Large recurrent microdeletions associated with schizophrenia. Nature 2008;455(7210):232–236. [PubMed: 18668039]

Tiwari HK, Barnholtz-Sloan J, Wineinger NE, Padilla MA, Vaughan LK, Allison DB. Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. Hum Hered 2008;66(2):67–86. [PubMed: 18382087]

Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 2007 15 March;23(6):657–663. [PubMed: 17234643]

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 2008 25 April;320(5875):539–543. [PubMed: 18369103]

Wang, H.; Veldink, J.; Opoff, R.; Sabatti, C. Markov Models for inferring Copy Number Variations from Genotype Data on Illumina Platforms. 2008. UCLA Statistics Department Preprint #533 http://preprints.stat.ucla.edu/

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007 November;17(11):1665–1674. [PubMed: 17921354]

Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R. A method for calling gains and losses in array CGH data. Biostatistics 2005 January;6(1):45–58. [PubMed: 15618527]

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661–678. [PubMed: 17554300]

White SJ, Vissers LE, Geurts van Kessel A, de Menezes RX, Kalay E, Lehesjoki AE, Giordano PC, van de Vosse E, Breuning MH, Brunner HG, den Dunnen JT, Veltman JA. Variation of CNV distribution in five different ethnic populations. Cytogenet Genome Res 2007;118(1):19–30. [PubMed: 17901696]

Winkler G, Liebscher V. Smoothers for discontinuous signals. J Nonpar Stat 2002;14(1–2):203–222.

Wojczynski M, Tiwari HK. Definition of the Phenotype. Adv Genet 2008;60:75–105. [PubMed: 18358317]

Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002a 24 October;3(11):research0062. [PubMed: 12429061]

Yang Q, Cui J, Chazaro I, Cupples LA, Demissie S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. BMC Genet 2005;6(suppl 1):S134. [PubMed: 16451593]

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002b 15 February;30(4):e15. [PubMed: 11842121]

Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, Blanchong CA, McBride KL, Higgins GC, Rennebohm RM, Rice RR, Hackshaw KV, Roubey RA, Grossman JM, Tsao BP, Birmingham DJ, Rovin BH, Hebert LA, Yu CY. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 2007 June;80(6):1037–1054. [PubMed: 17503323]

Zakharkin SO, Kim K, Mehta T, Chen L, Barnes S, Scheirer KE, Parrish RS, Allison DB, Page GP. Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics 2005 29 August; 6:214. [PubMed: 16124883]

Zhao X, Li C, Paez JG, Chin K, Jänne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004 1 May;64(9):3060–3071. [PubMed: 15126342]

Appendix

Appendix A

Algorithm 1

Loess Regression

1 Let $X = \{x_1, x_2, \ldots, x_p\}$ be a set of p points (or observations), and let $Y = \{y_1, y_2, \ldots, y_p\}$ be the corresponding data values.

2 **for** i=1 **to** p **do**

3 Let x_i be the point to be fitted on the loess curve.

4 Let $D(\cdot)$ be an appropriate distance metric. For each point $x_j \in X$, $i \neq j$, define the distance $d_j = D(x_i, xj)$.

5 Let $d_{(1)}, d_{(2)}, \ldots, d_{(p)}$ be the ordered set of distances.

6 Let k be the number of neighbors of x_i to be used in the local regression. The value of k is given as a percentage of the number of observations and is called the *span* of the regression.

7 Define $N(x_i)$, the neighborhood of x_i, as the set of points closest to x_i using the distance metric D:

$x_j \in N(x_i)$ if $d_{(j)} < k$

8 Define Δ, the maximum distance in the span, as

$$\Delta = \max_{x_j \in N(x_i)} | x_i - x_j |$$

9 Assign weights w_j to each of the points x_j

$$w_j = W\left(\frac{| x_i - x_j |}{\Delta}\right)$$

where $W(\cdot)$ is the tri-cube function

$$W(u) = \begin{cases} (1 - u^3)^3, & 0 \le u < 1, \\ 0, & \text{otherwise.} \end{cases}$$

10 Calculate the weighted least squares fit of Y on the neighborhood of X using the weights w.

11 Let $s(\cdot)$, the smooth function, be given by $s(x_i) = \hat{y}_j$ at the point x_i

12 **end for**

Source: Adapted from Cleveland and Devlin (1988)

Algorithm 2
Identification of Invariant Set

1 Initialize the invariant set at iteration zero $IS^{(0)}$ to the set of intensities for the two chips.

2 For iteration i, calculate the proportional rank difference (PRD), defined as the rank difference between the two chips divided by the number of probes per chip, for each of the intensities in the dataset $IS^{(i)}$.

3 If the PRD for a probe is less than the threshold δ, defined as $\delta = 0.003$ for low intensity probes and $\delta = 0.007$ for high intensity probes, then retain the probe for the dataset $IS^{(i+1)}$ of the next iteration. (The different thresholds allow for the sparsity of probes at the upper tail of the intensity distribution.)

4 Repeat steps 2 and 3 until the invariant set does not change between iterations.

5 Fit a piecewise linear running median line using the invariant set and perform normalization by projecting the intensities of the chip to be normalized onto the line.

Source: Adapted from Li and Wong (2001)

Algorithm 3
Quantile Normalization

1 Arrange the m microarrays, each with p probesets, into a $m \times p$ matrix $\mathbf{Y}$.

2 Sort each column of $\mathbf{Y}$ individually to give $\mathbf{Y}_{sort}$.

3 Compute the row means of $\mathbf{Y}_{sort}$.

4 Assign the row mean to each element of the row to give $\mathbf{Y}_{adj}$.

5 Rearrange each column of $\mathbf{Y}_{adj}$. into the original unsorted order in $\mathbf{Y}$ to give $\mathbf{Y}_{norm}$.

Source: Adapted from Bolstad et al. (2003)

Algorithm 4
Moving Window Smoothing

1 Let $x_1 \leq x_2, \leq \ldots \leq x_n$ be the n ordered data points, with corresponding intensity measurements $y_1 \leq y_2 \leq \ldots \leq y_n$.

2 Let $w \geq 0$ be the window size, which defines the neighborhood about the data point.

3 **for** $i = 1$ **to** n **do**

4 Let $l = \max(1, i - w)$ be the left boundary of the window.

5 Let $r = \min(n, i + w)$ be the right boundary of the window.

6 Let $f(\cdot)$ be the smoothing function. For the mean, this would be

$$f(x) = \frac{1}{r - l + 1} \sum_{j=l}^{r} x_j$$

and for the median this would be

$$f(x) = \begin{cases} x_{\lfloor (r-l+1)/2+1 \rfloor}, & (r - l + 1)\,\text{odd}, \\ \frac{1}{2}\left(x_{\lfloor (r-l+1)/2 \rfloor} + x_{\lfloor (r-l+1)/2+1 \rfloor} \right), & (r - l + 1)\,\text{even}, \end{cases}$$

where $[\cdot]$ denotes the floor function.

7 Compute the estimate $\hat{y}_j = f(x_j)$.

8 **end for**

Algorithm 5
Adaptive Weights Smoothing

1 Calculate the global MLE $\hat{\theta}^{(0)}$ of θ as

$$\hat{\theta}^{(0)} = \underset{\theta \in \Theta}{\operatorname{argsup}} \sum_{i=1}^{n} \log p(Y_i, \theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

2 **for** $i = 1$ **to** n **do**

3 Let $\widehat{\theta}_i^{(0)} = \widehat{\theta}^{(0)}$

4 Let $W_i^{(0)} = I_n$

5 **end for**

6 **repeat**

7 **for** $j = 1$ **to** n **do**

8 Calculate the penalties

$$I_{ij}^{(k)} \leftarrow |\rho(X_i, X_j)| / h^{(k)}|^2$$

$$s_{ij}^{(k)} \leftarrow \frac{1}{2\lambda}\left[L\left(W_i^{(k-1)}, \hat{\theta}_i^{(k-1)}, \hat{\theta}_j^{(k-1)}\right) + L\left(W_j^{(k-1)}, \hat{\theta}_j^{(k-1)}, \hat{\theta}_i^{(k-1)}\right)\right]$$

9 Calculate

$$\tilde{w}_{ij}^{(k)} \leftarrow K_I\left(I_{ij}^{(k)}\right) K_s\left(s_{ij}^{(k)}\right)$$

10 Define the weight

$$w_{ij}^{(k)} \leftarrow \eta w_{ij}^{(k-1)} + (1-\eta)\tilde{w}_{ij}^{(k)}$$

end for

11 $W_i^k \leftarrow \text{diag}\left(W_{i1}^{(k)}, W_{i2}^{(k)}, \ldots, W_{in}^{(k)}\right)$13: Calculate the new local MLE of θ_i

$$\hat{\theta}_i^{(k)} \leftarrow \underset{\theta \in \Theta}{\text{argsup}} L\left(W_i^{(k)}, \theta, \theta'\right)$$

12 $k \leftarrow k + 1$

13 $h^{(k)} \leftarrow a h^{(k-1)}$

14 **until** $a h^{(k)} > h^*$

Source: Adapted from Hupé et al. (2004)

Table 1

Suggested analytic methods corresponding to the phenotypic and CNV data

Phenotype/disease	CNV type	Analysis method
Continuous	CNV genotype (calls)	ANOVA
	Log R intensities (continuous)	Linear Regression
Discrete	CNV genotype (calls)	χ^2
	Log R intensities (continuous)	Logistic Regression

Table 2

Selected software for use in copy number data analysis

	Normalisation	Smoothing	Calling CNVs	Web address
aCGH		X	X	http://bioconductor.org/
CGHPRO	X	X		http://www.molgen.mpg.de/~abt_rop/molecular_cytogenetics/
ITALICS	X	X	X	http://bioconductor.org/
MANOR	X			http://bioconductor.org/
Rob-HMM (in CNA-HMMer software)	X		X	http://www.cs.ubc.ca/~sshah/acgh/
snapCGH	X	X	X	http://bioconductor.org/
dChip	X		X	http://biosun1.harvard.edu/complab/dchip/
CNAT	X	X	X	http://www.affymetrix.com/support/technical/software_downloads.affx
CRLMM (in the oligo package)	X	X	X	http://bioconductor.org/
CGH-Explorer		X		http://www.softgenetics.com/CGHExplorer.html
CGH-Minor			X	http://www-stat.stanford.edu/~wp57/CGH-Miner/
GLAD		X		http://bioconductor.org/
DNAcopy			X	http://bioconductor.org/
SMAP			X	http://bioconductor.org/
QuantiSNP			X	http://www.well.ox.ac.uk/QuantiSNP/
VanillaICE			X	http://bioconductor.org/
PennCNV			X	http://www.neurogenome.org/cnv/penncnv/
Illumina BeadStudio			X	http://www.illumina.com/pages.ilmn?ID=169
Birdsuite	X	X	X	http://www.broad.mit.edu/mpg/birdsuite/
CNAM	X		X	http://www.goldenhelix.com/SNP_Variation/CNAM/index.html