Markov Models for Inferring Copy Number Variations from Genotype Data on Illumina Platforms

Hui Wang^a Jan H. Veldink^e Hylke Blauw^e Leonard H. van den Berg^e
Roel A. Ophoff^{b, c, f} Chiara Sabatti^{b, d}

^aDepartment of Biostatistics, University of California at Berkeley, Berkeley, ^bDepartment of Human Genetics, ^cSemel Institute, and ^dDepartment of Statistics, UCLA, Los Angeles, Calif., USA; Departments of ^eNeurology and ^fMedical Genetics, University of Utrecht, Utrecht, The Netherlands

Key Words

Linkage • Disequilibrium association

Abstract

Background/Aims: Illumina genotyping arrays provide information on DNA copy number. Current methodology for their analysis assumes linkage equilibrium across adjacent markers. This is unrealistic, given the markers high density, and can result in reduced specificity. Another limitation of current methods is that they cannot be directly applied to the analysis of multiple samples with the goal of detecting copy number polymorphisms and their association with traits of interest. **Methods:** We propose a new Hidden Markov Model for Illumina genotype data, that takes into account linkage disequilibrium between adjacent loci. Our framework also allows for location specific deletion/duplication rates. When multiple samples are available, we describe a methodology for their analysis that simultaneously reconstructs the copy number states in each sample and identifies genomic locations with increased variability in copy number in the population. This approach can be extended to test association between copy number variants and a disease trait. **Results and Conclusions:** We show that taking into account linkage disequilibrium between adjacent markers can in-

crease the specificity of a HMM in reconstructing copy number variants, especially single copy deletions. Our multisample approach is computationally practical and can increase the power of association studies.

Copyright © 2009 S. Karger AG, Basel

1 Copy Number Variation and Genotyping Arrays

Two important surveys appeared in 2004 [6, 16] documenting the presence of copy number variation (CNV) in the genome in unsuspected high frequencies, both in the form of deletion of small genomic segments as well as duplications. Since then, considerable attention has been paid to the study of copy number polymorphisms and their relevance in disease traits. It has become apparent that high resolution genotype data offers an important source of information for detection of copy number polymorphisms: on the one hand, stretches of homozygous markers can indicate deletions [see, for example, 22] as well as duplications, on the other hand, patterns of Mendelian inconsistencies can also provide an indication of the presence of copy number variations [3]. Moreover, the technology used to obtain high density genotyping provides quantitative information on DNA amounts that can

be used to detect copy number variations. A number of recent studies have indeed explored the effectiveness of high density genotyping arrays to detect copy number variations in relation to other experimental methodologies: a careful comparison of different technologies can be found, for example in [5], while [14] represents one of the most recent and comprehensive screens.

This interest in the detection of copy number variation has spurred the development of statistical methods to analyze the experimental data: while a number of algorithms and software are available, both from commercial and academic sources (see, for example, the list provided by affymetrix http://www.affymetrix.com/products/application/cna_dataanalysis.affx), this area of research is still in its infancy. With new contributions appearing in almost any issue of relevant journals, it is hard to give a complete account of the current literature. Instead, we focus on two specific problems we would like to tackle: accounting for linkage disequilibrium between markers and multiple samples analysis.

We are interested in data obtained with Illumina genotyping arrays. A couple of contributions to date have underscored the power of a Hidden Markov Model framework for this data [2, 24]. It has also become apparent that including in the analysis all the available information [for example, marker specific allele frequencies as in 24] leads to increased sensitivity and specificity. As we become more knowledgeable about human variation, there is more relevant information we can rely upon. In particular, both [23] and [18] have noted the importance of accounting for linkage disequilibrium between adjacent SNPs in HMM models for high density genotyping. We here address this need.

We further tackle one of the open problems in the study of copy number variation, irrespectively of the technology used: how to jointly analyze multiple samples to discover CN polymorphisms, with possible effects on phenotypes. One of the first efforts in the systematic description of statistical tests for this goal is described in [4]. In this approach, however, the researcher assumes that copy number variations have been previously identified. A method that simultaneously identifies CN changes in multiple individuals and tests for their association with a trait has both the potential of detecting CNV more powerfully and the advantage to result in more accurate p values.

In summary, we extend HMM models [2, 24] to incorporate information on linkage disequilibrium between markers and varying rates of CNV across the genome. Furthermore, we show how this flexible framework can be adapted to jointly analyze multiple samples, with the

goal of increasing the power to detect CN polymorphisms and testing their association with traits of interest.

After a description of the characteristics of Illumina data, in section 3 we present a HMM model for the analysis of one sample. We include information on linkage disequilibrium and provide a new MM algorithm for the estimation of the parameters in a Bayesian framework. In section 4 we extend the HMM model to account for the presence in the genome of locations with higher rates of copy number variation. We also show how data on multiple samples can be used to detect copy number polymorphism and test for their association with relevant phenotypes. Section 5 illustrates the potential of our algorithm on real and simulated data sets: it is clear that accounting for linkage disequilibrium increases specificity; our multiple sample analysis facilitates the comparison of the reconstruction of common deletions from noisy data.

2 Illumina Genotype Data

We have become interested in studying the data obtained with the Illumina genotyping arrays because it is the platform of choice for a number of genome-wide studies we are involved in. Its use for CN detection has been explored in [2, 12, 24]. We briefly summarize here the characteristics of the raw data we will use.

In the following, adopting Illumina convention, we will indicate the two alleles at each SNP with A and B. We are going to base our analysis on two measurements that Illumina software generates for each SNP: the 'B allele Frequency' (BAF) and LogR. These capture, respectively, the estimated proportion of allele B and the log ratio of the overall level of DNA in the sample over a reference value.

For specific definitions of these values, we refer to Illumina documentation [12]. We here limit ourselves to an approximate description of the data transformation that lead to BAF and LogR. For each SNP, the Illumina genotyping platform leads to quantitative measurements of the amount of A and B alleles. Figure 1a presents (with simulated data) the scatterplot of these measurements (on the log scale) on the same SNP on a large number of individuals: clearly three clusters can be identified, corresponding to the AA, AB, and BB genotype. Illumina's software further transforms the data before proceeding to a genotype call. Figure 1b illustrates the change in coordinate system: θ is the angle between the vector defined by each of the points in 1(a) and the y axis, while $\log T$ is the sum of the log-intensities plotted in figure 1a. Finally, the centers of the three clusters are compared to reference

Fig. 1. Simulated data that illustrates the definition of LogR and BAF. (a) Scatterplot of the logarithm of intensity values for A and B allele in one SNP for 150 individuals. (b) Scatterplot of the same points of (a) in the changed coordinate system. (c) Standardization and re-scaling of the clusters to define LogR and BAF.

values and standardized, so that the center of the clusters on the x axis is either 0, 0.5 or 1, and it is 0 on the y axis (note that x values are truncated at 0 and 1). These final values are referred to as BAF and LogR.

Few remarks are in order. It is important to note that BAF and LogR represent an intermediate step between raw intensity values and final genotype calls. Typically, BAF and LogR are available for each SNP, while anomalous values of these will result in a 'no call' genotype. While sufficiently close to the raw data to carry relevant information for copy number quantification, LogR and BAF values are obtained through a standardization process that generates a great deal of homogeneity across SNPs. Our previous experience in low level analysis of intensity levels for allele probes in Affymetrix technology underscore the importance of this normalization process [15].

In the rest of the paper we will refer to the BAF value as x and to the LogR value as y . By definition, x ranges between 0 and 1, while y is standardized to have mean zero. In normal diploid state, x takes on values close to 0, 1/2 or 1, corresponding to the three possible genotypes, while y has zero mean. In presence of a hemizygous deletion, x takes on only values close to 0 or 1, and y tends to have negative values. In presence of a duplication, x can assume values close to 0, 1/3, 2/3 and 1 – corresponding to the 4 possible genotypes – and y is increasingly positive.

The preprocessing steps adopted by Illumina to define the x and y values are such that these carry almost independent information. Figures 2 and 3 illustrate how they are both important to detect CNV. Plotted against genomic positions are the x and y signal for about 2000

SNPs in the proximity of a deleted (2) and duplicated (3) region. It can be noted how a lower value of y allows to separate homozygous signals corresponding to a 1 or 2 copy number, and values of x close to 1/3 or 2/3 clearly mark a duplication, even when the corresponding y values are not elevated. Mindful of this observation, we want to develop an algorithm that detects CNV using both these signals. A Hidden Markov model is particularly useful in this setting, and, indeed, recent literature contributions [2, 24] document its effectiveness.

However, other applications of HMM to the analysis of high density genotype data as [18, 23] underscored the importance of accounting for linkage disequilibrium between adjacent markers. This should be important even when studying variations in copy number. Indeed, higher LD induces longer stretches of homozygous markers, which is one of the signatures of deletion. Unless LD information is appropriately incorporated in the analysis, a minor, random decrease of the intensity signal would often be interpreted as a deletion in these long homozygous stretches. In the following section we describe a new HMM that incorporates information on linkage disequilibrium between markers.

3 One Sample Analysis

The main interest of this paper is in the analysis of genotype data from normal individuals to detect presence of copy number polymorphisms. By definition, these consist in specific regions in the genome that are

Fig. 2. A deletion encompassing 245 SNPs on Chromosome 4. Data for additional 1000 SNPs flanking the deletion is also presented. On the x -axis, we report the positions of queried SNPs in base pairs. The top plot displays the copy number values; the central plot presents the y LogR values associated with each SNP; and the bottom plot displays the x B allele frequency values. This deletion was observed in a sample from the ALS study described in section 5.

lost or duplicated with non-negligible frequency in the population. A first goal consists, then, in the identification of such regions. Secondly, one may be interested in testing for association between the CN polymorphism and traits of interest.

While we do not analyze tumor data in the present work, similar goals can be stated for cancer-related research. When studying CN variation in tumor cells, one is especially interested in the identification of regions that appear preferentially lost or duplicated in cancer, suggesting that they contain tumor related genes.

In order to carry out either of these plans, one needs data from multiple individuals. The model we adopt for CN variation, however, is best described first specifying the probability distribution for data pertinent to a single sample and then illustrating how it can be modified to best exploit the information contained in multiple samples. A similar framework was first considered by New-

ton [9–11] in the context of copy number variation in tumors. We refer the interested reader to that work which documents the biological basis of this approach in the context of cancer studies.

A Homogeneous Markov Model for Copy Number States

Let $\{\pi_i\}$ be the continuous process that describes copy number status of one individual across the genome. We model $\{\pi_i\}$ as a homogeneous continuous-time Markov chain moving along each chromosome. For ease of exposition, we start considering three possible states, corresponding to 1, 2, or 3 DNA copies. In light of our goal of identifying copy number polymorphisms, our homogeneity assumption may appear rather futile. We will relax this assumption when analyzing data from multiple samples, but we rely on it as a regularization device when dealing with data from one individual alone.

Fig. 3. A duplication encompassing 82 SNPs on Chromosome 12. Data for additional 1000 SNPs flanking the duplication is also presented. On the x -axis, we report the positions of queried SNPs in base pairs. The top plot displays the copy number values; the central plot presents the y LogR values associated with each SNP; and the bottom plot displays the x B allele frequency values. This duplication was observed in a sample from the ALS study described in section 5.

Figure 4 illustrates the generic infinitesimal transition rates one can choose for this process. In an effort to build a parsimonious model, we initially considered fixing three rates: ϕ is the transition rate for deletions, ξ for duplications, and λ for restoration to diploidy. We considered two models: (a) each state can lead to any state; (b) states 1 and 3 can be reached only from state 2. Under both models, the equilibrium distribution for the chain is

$$\begin{aligned} p(\pi_i = 1) &= \frac{\phi}{\phi + \lambda + \xi}, \\ p(\pi_i = 2) &= \frac{\lambda}{\phi + \lambda + \xi}, \\ p(\pi_i = 3) &= \frac{\xi}{\phi + \lambda + \xi}. \end{aligned}$$

Under model (a), the expected length of a deletion is $(\xi + \lambda)^{-1}$, the expected length of a duplication is $(\phi + \lambda)^{-1}$,

and the expected length of a diploid stretch is $(\phi + \xi)^{-1}$. Under model (b) the expected values are, respectively $(\lambda)^{-1}$, $(\lambda)^{-1}$, and $(\phi + \xi)^{-1}$.

In effect, this continuous process is observed (partially) only at a set of discrete time points $i = 1, \dots, n$, corresponding to the genotyped SNPs. Using standard continuous Markov chain theory [7], for model (a), we can obtain a following discrete time transition matrix between points at distance d :

$$t^a(d) = \begin{pmatrix} 1 - (1 - \delta)(1 - e^{-\eta d}) & (1 - \delta - \gamma)(1 - e^{-\eta d}) & \gamma(1 - e^{-\eta d}) \\ \delta(1 - e^{-\eta d}) & 1 - (\delta + \gamma)(1 - e^{-\eta d}) & \gamma(1 - e^{-\eta d}) \\ \delta(1 - e^{-\eta d}) & (1 - \delta - \gamma)(1 - e^{-\eta d}) & 1 - (1 - \gamma)(1 - e^{-\eta d}) \end{pmatrix}, \quad (2)$$

where $\delta = \pi_1$, $\gamma = \pi_3$, and $\eta = \phi + \lambda + \xi$. Model (b) leads to a similar finite time transition matrix, with second row

Color version available online

Fig. 4. Generic infinitesimal rates for the homogeneous copy number process with three states. Edges labelled with the same greek letter corresponds to rates that are identical in the models we analyze. Edges depicted with a lighter color are not present in model (b).

and second column identical to (2) and other terms as follows:

$$t_{11}^b(d) = \frac{\gamma}{\gamma + \delta} e^{-(1-\gamma-\delta)\eta d} + \delta + \frac{\delta}{\gamma + \delta} (1 - \delta - \gamma) e^{-\eta d}$$

$$t_{13}^b(d) = -\frac{\gamma}{\gamma + \delta} e^{-(1-\gamma-\delta)\eta d} + \gamma + \frac{\gamma}{\gamma + \delta} (1 - \delta - \gamma) e^{-\eta d},$$

where $t_{31}^b(d)$ and $t_{33}^b(d)$ can be obtained appropriately exchanging the values of γ and δ .

In the following, we will use model (a). The practical differences between model (a) and (b) are not substantial and model (a) leads to a transition matrix that can be more directly compared with others used in hidden Markov models for data of our type. For example, a similar model has been proposed in [2]: these authors, however, assume each copy number state to be equally probable, which appears unwarranted and unnecessary. Model (a) is adopted also in [24].

Each of the three parameters in (2) can be easily related to observable quantities: δ and γ represent respectively the population frequency of deletions and duplications. In model (a), the chance that a change of state occurs between two locations d apart depends only on the parameter η , which captures the level of dependency of the process and can be interpreted as a ‘smoothing’ parameter. Note that expected lengths of deletions and duplications can be expressed in terms of these parameters: easy algebra leads to obtain that the expected length of a deletion is $1/(\eta(1 - \delta))$ and the expected length of a duplication is $1/(\eta(1 - \gamma))$.

While we select model (a) for definiteness, it is important to note, that the specific form of the homogeneous

Markov model for CN is not crucial for the remainder of the analysis. One can easily decide to adopt another model, with a higher number of parameters, for example. Adapting the algorithm to consider this will not entail very substantial work. Indeed, one of our current goals for further investigation is to consider a variety of different models.

Emission Probabilities and Expanded State Space

The states of the Markov model described in the previous section are not observed; rather, for each genotyped SNP i , we observe values $o_i = (x_i, y_i)$. Emission probabilities link these to the underlying states of the Markov process. We assume that conditional on the unobserved copy number, x and y are independent.

Figure 5 presents summary statistics of the distribution of x and y in the training data set that guided our modelling choices (for details on this data set, please see section 5).

We propose modeling the distribution of y_i using Gaussian distributions. A close inspection of the histograms in figure 5 reveals that the histograms of the y values corresponding to one and three DNA copies are asymmetric. However, we attribute this to a selection effect: the CNV used to compile those histograms have been identified due to their ‘anomalous’ y values. This results in the following distribution for the y values:

$$\begin{aligned} y | \text{aploid} &\sim \mathcal{N}(\mu_1, \sigma_1) \\ y | \text{diploid} &\sim \mathcal{N}(0, \sigma_2) \\ y | \text{triploid} &\sim \mathcal{N}(\mu_3, \sigma_3) \end{aligned}$$

Let us now consider models for the distribution of x . By looking at the data presented in figure 5, one immediately notices that values of x depend on the underlying genotypes. In particular, the theoretical fraction of B alleles over total corresponding to different genotypes appears important; in our model we have five such ratios: 0, 1/3, 1/2, 2/3, 1. The emission probabilities we choose for x need to model the empirical distribution of x conditional on the true underlying allelic ratio. Before we specify such distributions, however, another consideration is in order. Our previous work on detection of deletions and estimation of inbreeding coefficients [23] underscores the importance of accounting for linkage disequilibrium when interpreting the signal from consecutive homozygous markers. In other words, the distribution of x_i and x_{i+1} , the B allele frequencies at two consecutive markers, cannot be considered independent given the copy number status at loci i and $i + 1$. In [23] we describe an extension of the typical HMM framework to account for this

Fig. 5. Histogram of the values of the emitted signals x and y for approximately 500,000 SNPs genotyped in a set of individuals including known deletions and duplications. Each row corresponds to one copy number. (See section 5 for more detail).

dependency. The same model has been proposed by [18] in the context of ancestry reconstruction. We here adopt a more direct strategy.

The simplest way to take linkage disequilibrium into account is to augment the hidden states of the HMM described above to include underlying genotypes. Thus, we propose a model based on the nine possible hidden states $\mathcal{S} = \{A, B, AA, AB, BB, AAA, AAB, ABB, BBB\}$. The transition probabilities between these states are defined by the deletion-duplication mechanism described in models (a) (equation 2) and the frequencies of the two marker haplotypes they imply. As an illustration, the transition between states A and AB , respectively at marker i and j , can be calculated as

$$P_{ij}(AB | A) = t_{12}(d_{ij})[q_{ij}(A | A)p_j(B) + q_{ij}(B | A)p_j(A)], \quad (3)$$

where $t_{12}(d_{ij})$ is the probability of going from 1 to 2 copies over a distance d_{ij} , $p_j(A)$ is the probability of allele A at marker j and $q_{ij}(A | A)$ and $q_{ij}(B | A)$ are the probabilities of allele A and B at marker j conditionally on allele A at marker i . That is, the probability of the transition between states A and AB , is decomposed in the probability of transition from copy number 1 to copy number 2 and the probability of observing the genotype AB at marker j given that at marker i we have one allele A . The latter is calculated considering that AB at marker j is obtained by the pairing of either an haplotype (A, A) at markers i and

j and an allele B at marker j or an haplotype (A, B) at markers i and j and an allele A at marker j .

In general, we can write the transition probability between two of the nine possible states as the product of a component that depends only on the copy number associated to each state π_i and a component that depends also on the genotype value for the state π_i : $t(\pi_i, \pi_{i+1}) = t_{\text{cn}(\pi_i), \text{cn}(\pi_{i+1})}(d_{i,i+1})G(\pi_i, \pi_{i+1})$. We are going to assume that the $G(\pi_i, \pi_{i+1})$ component of the transition probability is known. Indeed, it is determined by two-marker haplotype frequencies that can be estimated easily from association genome scan data. So that the unknown parameters of the transition probability remain $\mathcal{T} = (\delta, \gamma, \eta)$.

Given that we are going to work with a HMM with nine states, corresponding to the nine genotypes, it is convenient to specify the emission probabilities for x conditional on each of the nine genotypes (the emission probabilities for y still depend only on the underlying copy number). Since there is no prior distinction between the A and B allele, we enforce appropriate symmetries in the distributions. Figure 6 presents the histograms of x conditional on known underlying genotypes for a collection of about 500,000 SNPs that form one of our training data sets (see section 5 for more specific information).

One detail that is not immediately evident from figure 6, is that the distribution of x for homozygous genotypes is not simply continuous. The algorithm used by Illumina

Fig. 6. Histograms of the value of the emitted x signals, conditional on underlying genotype, for approximately 500,000 SNPs including deletions and duplications. Each row corresponds to one copy number, each column to one of the five relevant allele ratios: 0, 1/3, 1/2, 2/3, 1. The specific genotype is indicated in the main title of each plot.

sets equal to 0 (1) extremely small (large) x values. Our emission probabilities need to incorporate these discrete component. The continuous components appear to be adequately modeled with exponential distribution. A double exponential distribution provides a good fit for the emission of the AB genotype, while the x values corresponding to ABB and AAB are better modeled as Gaussian variates. Obviously, both exponential and Gaussian distribution need to be truncated as to assign positive probability only to the $[0,1]$ interval.

To start with, we enforce a certain amount of symmetry and consider then the following emission probabilities for x :

$$f(x|A) = f(x|AA) = f(x|AAA) = \begin{cases} \omega & x = 0 \\ (1 - \omega)\lambda e^{-\lambda x} & x \neq 0 \end{cases}$$

$$f(x|B) = f(x|BB) = f(x|BBB) = \begin{cases} \omega & x = 1 \\ (1 - \omega)\lambda e^{-\lambda(1-x)} & x \neq 1 \end{cases}$$

$$f(x|AB) = \lambda_2/2 \exp\{-\lambda_2|x - 0.5|\}$$

$$f(x|AAB) = \mathcal{N}(1/3, \nu)$$

$$f(x|ABB) = \mathcal{N}(2/3, \nu)$$

In summary, the emission probabilities we described are defined by the following parameters $\mathcal{E} = (\mu_1, \sigma_1, \sigma_2, \mu_3, \sigma_3, \omega, \lambda, \lambda_2, \nu)$.

Algorithms for Evaluating Conditional Probabilities and Parameter Estimation

To evaluate the probability of the data $O = \{o_{ij}\}_{i=1}^n$ for one individual, we resort to the standard hidden Markov

model machinery. Rather than attempting to calculate directly the sum

$$\Pr(O|\mathcal{T}, \mathcal{E}) = \sum_{\Pi} p(\pi_1|\mathcal{T}) e(o_1|\pi_1, \mathcal{E}) \prod_{i=2}^n t(\pi_i|\pi_{i-1}, \mathcal{T}) e(o_i|\pi_i, \mathcal{E}),$$

we define the forward and backward probabilities $\alpha(\pi_i) = \Pr(o_1, \dots, o_i, \pi_i)$, and $\beta(\pi_i) = \Pr(o_{i+1}, \dots, o_n | \pi_i)$ that can be evaluated with recursions:

$$\alpha(\pi_i) = \sum_{\pi_{i-1} \in \mathcal{S}} \alpha(\pi_{i-1}) t(\pi_i|\pi_{i-1}) e(o_i|\pi_i)$$

$$\beta(\pi_i) = \sum_{\pi_{i+1} \in \mathcal{S}} \beta(\pi_{i+1}) t(\pi_{i+1}|\pi_i) e(o_{i+1}|\pi_{i+1}).$$

The sample probability can then be obtained as

$$\Pr(O) = \sum_{\pi_n \in \mathcal{S}} \alpha(\pi_n).$$

The sequences of α and β can also be used to evaluate the distribution of π_i conditional on the observed data and the conditional probability of a specific transition. Let $\text{Prob}(\pi_i = s_j, \pi_{i+1} = s_k | O) = \xi_{i,i+1}(s_j, s_k)$ and $\text{Prob}(\pi_i = s_k | O) = \rho_i(s_k)$, then

$$\xi_{i,i+1}(s_j, s_k) = \frac{\alpha(\pi_i) t(\pi_i|\pi_{i+1}) e(o_{i+1}|\pi_{i+1}) \beta(\pi_{i+1})}{\Pr(O)} \quad (4)$$

$$\rho_i(s_k) = \sum_{\pi_{i+1} \in \mathcal{S}} \xi(\pi_i, \pi_{i+1}). \quad (5)$$

The hidden Markov model machinery can also be used to estimate the parameters $\mathcal{T}, \mathcal{E}$. We will discuss in a later session the precise strategy with which we choose the parameter values in data analysis. We here present the algorithm that can be used given availability of appropriate data. We start considering a maximum likelihood approach which we will extend to maximum a posteriori. Consider the maximization problem

$$\max_{\mathcal{T}, \mathcal{E}} \mathcal{L}(\mathcal{T}, \mathcal{E} | O) =$$

$$\max_{\mathcal{T}, \mathcal{E}} \sum_{\Pi} p(\pi_1|\mathcal{T}) e(o_1|\pi_1, \mathcal{E}) \prod_{i=2}^n t(\pi_i|\pi_{i-1}, \mathcal{T}) e(o_i|\pi_i, \mathcal{E}).$$

We solve this using an iterative algorithm in the MM framework [8]. That is, given a current value for the parameters $(\mathcal{T}^\ell, \mathcal{E}^\ell)$, we need to find a minorizing function $q(\mathcal{T}, \mathcal{E} | \mathcal{T}^\ell, \mathcal{E}^\ell)$ such that $q(\mathcal{T}, \mathcal{E} | \mathcal{T}^\ell, \mathcal{E}^\ell) \leq \mathcal{L}(\mathcal{T}, \mathcal{E} | O)$ for all $(\mathcal{T}, \mathcal{E})$, with equality holding for $(\mathcal{T}, \mathcal{E}) = (\mathcal{T}^\ell, \mathcal{E}^\ell)$. At each iteration, new values $(\mathcal{T}^{\ell+1}, \mathcal{E}^{\ell+1})$ will be obtained maximizing the minorizing function with respect to $(\mathcal{T}, \mathcal{E})$. The sequence of parameter values $\{\mathcal{T}^\ell, \mathcal{E}^\ell\}$ so defined is guaranteed to lead to increasing values of the likelihood (see [8]). The theory of the EM algorithm offers a recipe for identifying a minorizing function for the loga-

rithm of the likelihood. Indicating with $\Pi = \{\pi_i\}_{i=1}^n$ the missing data corresponding to the unobserved copy-number states, we can use $q(\mathcal{T}, \mathcal{E} | \mathcal{T}^\ell, \mathcal{E}^\ell) = E(\log \mathcal{L}(\mathcal{T}, \mathcal{E} | O, \Pi))$, the expected value of the logarithm of the complete data log-likelihood. Following this suggestion, we obtain a minorizing function where the emission parameters $\mathcal{E}$ and the transition parameters $\mathcal{T}$ are separated:

$$E(\log \mathcal{L}(\mathcal{T}, \mathcal{E} | O, \Pi)) =$$

$$E\left(\log p(\pi_1|\mathcal{T}) + \sum_{i=2}^n \tau(\pi_i|\pi_{i-1}, \mathcal{T})\right) + E\left(\sum_{i=1}^n \epsilon(o_i|\pi_i, \mathcal{E})\right),$$

where τ and ϵ indicate the logarithms of t and e .

Let's consider first the updates relative to the emission parameters. These can be obtained maximizing

$$E\left(\sum_{i=1}^n \epsilon(o_i|\pi_i, \mathcal{E})\right) = \sum_{k=1}^9 \sum_{i=1}^n \epsilon(o_i|\pi_i, \mathcal{E}) \rho_i^\ell(s_k).$$

In this sum, emission parameters relative to different states are separated. Simple calculations show that the maximum value of the surrogate function is obtained in correspondance of the classical MLE estimators for the Gaussian, binomial, and exponential parameters that constitute $\mathcal{E}$, with observations appropriately weighted for their probability of deriving from the relevant unobserved state. For example, consider the parameters (μ_1, σ_1^2) relative to the distribution of y , given copy number 1. Let $w_i^\ell(1) = \rho_i^\ell(s_1) + \rho_i^\ell(s_2)$ be the current probability that observation i derives from a hidden state with copy number 1. Then,

$$\mu_1^{(\ell+1)} = \frac{\sum_{i=1}^n y_i w_i^\ell(1)}{\sum_{i=1}^n w_i^\ell(1)}$$

$$\sigma_1^{(\ell+1)} = \sqrt{\frac{\sum_{i=1}^n (y_i - \mu_1^{(\ell+1)})^2 w_i^\ell(1)}{\sum_{i=1}^n w_i^\ell(1)}}$$

Let us now consider the transition parameters. For ease of exposition, let us omit consideration of $p(\pi_1)$. Then, the relevant portion of the logarithm of the complete data likelihood is

$$\sum_{i=2}^n \tau(\pi_i|\pi_{i-1}, \mathcal{T}) = \sum_{i=2}^n \log t_{\text{cn}(\pi_i), \text{cn}(\pi_{i-1})}(d_{i,i+1}) + \sum_{i=2}^n \log G(\pi_i, \pi_{i+1}),$$

where the parameter $\mathcal{T}$ depends on π_i only through the copy number and not the genotype values. To carry out

traditional EM, at each iteration, we would need to maximize

$$\sum_{j,k=1}^9 \sum_{i=2}^n \xi_{i-1,i}^{\ell} (s_j, s_k) \log t_{\text{cn}(s_j), \text{cn}(s_k)}(d_{i,i+1}), \quad (6)$$

however, the three parameters δ , γ , and η are not separated. Referring back to the transition matrix in (2), it is easy to see that η and the frequency parameters are separated, when taking the logarithms, in all matrix entries, with the exception of the main diagonal. To circumvent this difficulty, we seek a further minorization. Let's consider the first entry of the finite time transition matrix: $\log(1 - (1 - \delta)(1 - e^{-\eta d})) = \log(\delta(1 - e^{-\eta d}) + e^{-\eta d})$. Setting $a = \delta(1 - e^{-\eta d})$ and $b = e^{-\eta d}$, we can use the concavity of the logarithm function to find a minorization:

$$\log(a + b) \geq \frac{a_0}{a_0 + b_0} \log\left(\frac{a_0 + b_0}{a_0} a\right) + \frac{b_0}{a_0 + b_0} \log\left(\frac{a_0 + b_0}{b_0} b\right),$$

with equality holding for $a = a_0$, $b = b_0$. This says that for optimization purposes, we can concentrate on the function

$$\frac{a_0}{a_0 + b_0} \log(a) + \frac{b_0}{a_0 + b_0} \log(b),$$

which separates the δ and η parameters. Precisely, let

$$\begin{aligned} \alpha_i^{\ell} &= \frac{\delta^{\ell} (1 - e^{-\eta^{\ell} d_{i,i+1}})}{\delta^{\ell} (1 - e^{-\eta^{\ell} d_{i,i+1}}) + e^{-\eta^{\ell} d_{i,i+1}}} \\ \beta_i^{\ell} &= \frac{\gamma^{\ell} (1 - e^{-\eta^{\ell} d_{i,i+1}})}{\gamma^{\ell} (1 - e^{-\eta^{\ell} d_{i,i+1}}) + e^{-\eta^{\ell} d_{i,i+1}}} \\ \zeta_i^{\ell} &= \frac{(1 - \gamma^{\ell} - \delta^{\ell}) (1 - e^{-\eta^{\ell} d_{i,i+1}})}{(1 - \gamma^{\ell} - \delta^{\ell}) (1 - e^{-\eta^{\ell} d_{i,i+1}}) + e^{-\eta^{\ell} d_{i,i+1}}}. \end{aligned}$$

We can minorize each of the logarithms of the terms on the main diagonal of (2) as follows:

$$\begin{aligned} \log t_{11}^{\ell}(d_{i,i+1}) &\geq \alpha_i^{\ell} \log \delta + \alpha_i^{\ell} \log(1 - e^{-\eta^{\ell} d_{i,i+1}}) - (1 - \alpha_i^{\ell}) \eta^{\ell} d_{i,i+1} \\ \log t_{22}^{\ell}(d_{i,i+1}) &\geq \zeta_i^{\ell} \log(1 - \delta - \gamma) + \zeta_i^{\ell} \log(1 - e^{-\eta^{\ell} d_{i,i+1}}) - (1 - \zeta_i^{\ell}) \eta^{\ell} d_{i,i+1} \\ \log t_{33}^{\ell}(d_{i,i+1}) &\geq \beta_i^{\ell} \log \gamma + \beta_i^{\ell} \log(1 - e^{-\eta^{\ell} d_{i,i+1}}) - (1 - \beta_i^{\ell}) \eta^{\ell} d_{i,i+1}. \end{aligned}$$

Plugging these expressions in (6), we obtain a simple function of δ and γ , separated from η . Let $\kappa_{i,i+1}^{\ell}(j, k) = \sum_{l: \text{cn}(l) = j, \text{cn}(l) = k} \xi_{i,i+1}^{\ell}(s_j, s_k)$. Then, $N_{j,k}^{\ell} = \sum_i \kappa_{i,i+1}^{\ell}(j, k)$ is the expected number of transitions from state j to state $k \neq j$. And let $N_{1,1}^{\ell} = \sum_i \kappa_{i,i+1}^{\ell}(1,1) \alpha_i^{\ell}$, $N_{2,2}^{\ell} =$

$\sum_i \kappa_{i,i+1}^{\ell}(2,2) \zeta_i^{\ell}$, $N_{3,3}^{\ell} = \sum_i \kappa_{i,i+1}^{\ell}(3,3) \beta_i^{\ell}$. Then, to obtain the up-dates of δ and γ we need to maximize the following function:

$$\log(\delta)(N_{1,1} + N_{2,1} + N_{3,1}) + \log(\gamma)(N_{1,3} + N_{2,3} + N_{3,3}) + \log(1 - \delta - \gamma)(N_{1,2} + N_{2,2} + N_{3,2}).$$

This leads to the obvious updates

$$\begin{aligned} \delta^{\ell+1} &= \frac{N_{1,1} + N_{2,1} + N_{3,1}}{N_{1,1} + N_{2,1} + N_{3,1} + N_{1,2} + N_{2,2} + N_{3,2} + N_{1,3} + N_{2,3} + N_{3,3}} \\ \gamma^{\ell+1} &= \frac{N_{1,3} + N_{2,3} + N_{3,3}}{N_{1,1} + N_{2,1} + N_{3,1} + N_{1,2} + N_{2,2} + N_{3,2} + N_{1,3} + N_{2,3} + N_{3,3}} \end{aligned}$$

The corresponding function of η is slightly more complicated. Let

$$\begin{aligned} M_i^{\ell} &= \sum_{j \neq k} \kappa_{i,i+1}^{\ell}(j, k) \\ S_i^{\ell} &= \sum_j \kappa_{i,i+1}^{\ell}(j, j) \\ W_i^{\ell} &= \kappa_{i,i+1}^{\ell}(1,1) \alpha_i^{\ell} + \kappa_{i,i+1}^{\ell}(2,2) \zeta_i^{\ell} + \kappa_{i,i+1}^{\ell}(3,3) \beta_i^{\ell} \end{aligned}$$

The minorizing function that we have to maximize at iteration $\ell + 1$ is then

$$g(\eta | \mathcal{T}^{\ell}, \mathcal{E}^{\ell}) = \sum_i \log(1 - e^{-\eta^{\ell} d_{i,i+1}}) (O_i^{\ell} + W_i^{\ell}) - \eta^{\ell} d_{i,i+1} (S_i^{\ell} + W_i^{\ell}).$$

While we cannot find an analytic expression for the max of this function, one can easily calculate its first and second derivative and use them to define a Newton update (it is well known that one can substitute a step of Newton algorithm for the explicit maximization in the MM framework [8]):

$$\begin{aligned} \frac{\partial g(\eta | \mathcal{T}^{\ell}, \mathcal{E}^{\ell})}{\partial \eta} &= \sum_i \frac{d_{i,i+1} e^{-\eta^{\ell} d_{i,i+1}}}{1 - e^{-\eta^{\ell} d_{i,i+1}}} (O_i^{\ell} + W_i^{\ell}) - \sum_i d_{i,i+1} (S_i^{\ell} + W_i^{\ell}) \\ \frac{\partial^2 g(\eta | \mathcal{T}^{\ell}, \mathcal{E}^{\ell})}{\partial \eta^2} &= - \sum_i d_{i,i+1}^2 \frac{e^{-\eta^{\ell} d_{i,i+1}}}{(1 - e^{-\eta^{\ell} d_{i,i+1}})^2}. \end{aligned}$$

Before we conclude this section devoted to the description of computational methods, we want to describe how the MM algorithm above can be modified to obtain maximum a posteriori estimate rather than MLE. A Bayesian framework may be useful when prior data is available to provide the user with reasonable values for the parameter, and current data is to be used simply to adjust these estimates. Conjugate priors allow one to incorporate available information without increasing the computational burden. We assume that the emission $\mathcal{E}$ and transition parameters

$\mathcal{T}$ are independent a priori. Furthermore, we assume that a priori all the emission parameters are independent from each other. We take a Gaussian prior for each of (μ_1, μ_3) , a Gamma prior for each of the precisions $(1/\sigma^2_1, 1/\sigma^2_2, 1/\sigma^2_3, 1/\nu^2)$, a Beta prior for ω , and a Gamma prior for the exponential parameters λ, λ_2 . We use a Dirichlet prior for (δ, γ) and a Gamma prior for η . To obtain maximum a-posteriori estimates, we can use an MM algorithm, adopting the same minorizations described above. The updates will change in a predictable fashion. For example, consider again the parameters (μ_1, σ^2_1) . Say that we take a Gaussian prior on μ_1 with parameters (m_1, p_1) (with p_1 precision of the distribution), and a gamma prior on $1/\sigma^2_1$ with parameters (a_1, b_1) . Then, the updates for the parameters are:

$$\mu_1^{(\ell+1)} = \frac{\sum_{i=1}^n y_i w_i^\ell(1)}{\sum_{i=1}^n w_i^\ell(1)}$$

$$\sigma_1^{(\ell+1)} = \sqrt{\frac{\sum_{i=1}^n (y_i - \mu_1^{(\ell+1)})^2 w_i^\ell(1)}{\sum_{i=1}^n w_i^\ell(1)}}$$

4 Multiple Sample Analysis

When data from multiple samples is available, it becomes possible to identify genomic locations that are lost or duplicated with higher frequency. In the case of DNA from normal individuals, this leads to the identification of copy number polymorphisms. In the case of tumor cells, this localizes tumor-related genes. Furthermore, when analyzing multiple samples together and accounting for differential rates of CNV across the genome, we increase our power to detect copy number variations in individuals with noisier signal. Let us consider, then, how to exploit the information available in multiple samples, and relax our assumption of homogeneity across the genome of the copy number process. We do this by defining the location specific parameters δ_i and γ_i .

Identifying Cancer Related Genes

We start considering the setting studied by Newton [9–11]: the study of cancer cell lines to localize tumor suppressor genes. The biological process that generates the observed data can be grossly summarized as follows. (1) all somatic cells experience occasionally loss or duplication of DNA portions. These are typically rare events, and

cell survival is more likely, the smaller the size of the perturbation; (2) when, because of one such event, the region harboring a tumor-related gene is affected, the cell has an increased chance of becoming cancerous; (3) tumor cells experience, uniformly across the genome, higher rates of loss and duplication, because of the frequency and instability of their division process.

When we observe DNA from a population of cancer cells, then, we expect to see the effects of (a) relatively common deletions and duplications across the genome; (b) and effects of a selection process in which cells that experienced losses or duplications in a specific position became cancerous. With reference to our model, these observations mean that (a) the loss and duplication probabilities δ, γ can be assumed to be constant across ‘most’ of the genome, and to have non-transcurable values. Furthermore, (b) we can identify the location of cancer-related genes by looking for positions that call for a loss/duplication rate different from the background one. Hence, for SNP i , we will want to test $H_0 : \delta_i = \delta$ versus $H_1 : \delta_i > \delta$, indicating that SNP i is lost at a higher rate than the rest of the genome. We may be interested in $H_0 : \gamma_i = \gamma$ versus $H_1 : \gamma_i > \gamma$, or in a combination of these hypothesis. Furthermore, if our sample contains a set of cases (e) and controls (c) for a given trait, we may want to test the $H_0 : \delta_i^c = \delta_i^e$ versus $H_1 : \delta_i^c \neq \delta_i^e$, to investigate possible association between the CN at location i and the trait of interest.

Taking this view point, when analyzing each location i we assume that all the remaining locations on the chromosome are lost or duplicated at the background rates δ, γ . The probability of the data O^k for $k = 1, \dots, m$ individuals is then equal to:

$$P(O^1, \dots, O^m | \mathcal{E}, \mathcal{T}, \delta_i, \gamma_i) = \prod_{k=1}^m (\delta_i P(O^k | \pi_i^k = 1) + (1 - \delta_i - \gamma_i) P(O^k | \pi_i^k = 2) + \gamma_i P(O^k | \pi_i^k = 3)).$$

If we keep the parameters $\mathcal{E}, \mathcal{T}$ fixed, it is easy to estimate the pair (δ_i, γ_i) using a MLE approach. Resorting once again the EM framework, we consider $\pi_i^1, \dots, \pi_{mi}$ as the missing data, and we can iterate on the basis of the following

$$E(\pi_i^k = 1 | O, \delta_i^\ell, \gamma_i^\ell) \propto \delta_i^\ell P(O^k | \pi_i^k = 1) \quad (7)$$

$$\delta_i^{(\ell+1)} = \sum_{k=1}^m E(\pi_i^k = 1 | O, \delta_i^\ell, \gamma_i^\ell) / m, \quad (8)$$

with analogous expressions holding for γ_i .

These maximum likelihood estimates are used to carry out a likelihood ratio test for the hypothesis of interest. Suppose for example that we want to test $H_0^i : \delta_i = \delta$ versus

$H_1 : \delta_i > \delta$, which is the case when we are trying to identify the location of a tumor suppressor gene. Then, we can construct the following lod-score curve:

$$L_i = \begin{cases} \log_{10} \frac{\mathcal{L}(\mathcal{E}, \mathcal{T}, \delta_i^* | O^1, \dots, O^m)}{\prod_{k=1}^m \mathcal{L}(\mathcal{E}, \mathcal{T} | O^k)} & \delta_i^* = \underset{\delta_i}{\operatorname{argmax}} \mathcal{L}(\mathcal{E}, \mathcal{T}, \delta_i^* | O^1, \dots, O^m) > \delta \\ 0 & \delta_i^* < \delta \end{cases}$$

The researcher's interest focuses on locations where the values of L_i are particularly high.

On purely statistical grounds, the determination of an appropriate cut off depends on the distribution of L_i under the null H_0 and on the necessity of taking into account that multiple tests are being performed. Furthermore, notice that the tests L_i and L_j corresponding to two locations on the same chromosome are not independent. To determine a significance cut-off one ideally would like to know the distribution of the entire process $\{L_i\}_k$ under the complete null hypothesis. Unfortunately, this is unknown at this stage. The marginal distribution of L_i , as $m \rightarrow \infty$, can be roughly approximated using the known results for likelihood ratio tests: under H_0 , $2 \ln 10 L_i$ is asymptotically distributed as a 50:50 mixture of a mass at zero and $\chi^2_{(1)}$ (the mass at zero derives from the fact that we place a constraint on the values of $\delta_i > \delta$, and the 0.5 mixing coefficient can be derived from the consistency and Gaussianity of the MLE of δ_i). While this approximation of the distribution of L_i is rather crude, it provides us a guideline of what a reasonable significance cut-off may be. The appropriate cut-off for L_i depends on the distribution of L_i and, roughly speaking, on the number of 'effectively independent' tests, which is determined by the length of the segment of the genome studied and the value of the η parameter. We suggest that once the instability parameters are estimated, a small scale simulation study be conducted where genotype data with the same structure as the real one is generated from the instability model, with no selection effect, and a cut-off for L_i that controls the desired measure of error rate to be determined. It may be of use to refer once again to the analogy with linkage mapping which carries through in terms of distribution for L_i : in these genetic mapping studies, a value of L_i greater than 3, or 3.5 is typically considered strong evidence in favor of H_1^i (Lander and Kruglyak, 1995).

DNA from Normal Cells and Identification of CN Polymorphisms

Let us now consider the case of normal cells where we are interested in discovering CN polymorphisms. The

biological mechanisms behind the observed data are here quite different. It is still true that (1) all somatic cells experience occasionally loss or duplication uniformly across the genome. These are typically very rare events. (2) In addition, there appear to be specific genomic regions that are present with variable copy number in the population.

The homogeneous Markov model we presented for the analysis of a single cell characterizes well copy number variations derived from sporadic events. CN polymorphisms, instead, call for a non homogeneous component in our process. The approach described in section 4.1 is somewhat inadequate to characterize this inhomogeneity as, unlike what happens for tumor cells in regions surrounding tumor suppressor genes, all cell lines that exhibit CN variations are expected to share the same boundaries. Nevertheless, it provides a useful tool for analysis.

We distinguish here two problems: (a) how to take into account known CNPs; (b) how to discover new CNPs. The HMM framework can be easily adapted to account for known CNPs, modifying the homogeneous model in 2, with location specific parameters δ_i, γ_i . All the SNPs in a CNP region will have the same values δ_i, γ_i , different from the background ones. One could imagine using the same framework to discover novel CN polymorphisms. The algorithm we described for the estimation of δ, γ can be easily adapted for δ_i, γ_i in presence of multiple samples. To model the known structure of the CN polymorphisms, we can use a prior that enforces small total variation for $\{\delta_i\}_{i=1}^n$ and $\{\gamma_i\}_{i=1}^n : \Pr(\{\delta_i\}) \sim \sum_{i=2}^n |\delta_i - \delta_{i-1}|$ (procedures that use such penalty have been recently suggested for the analysis of CGH data under the name of 'fused lasso' [19]).

However, one has to be careful not to underestimate the computational costs involved in such procedures. For each iteration, one needs to recompute all the forward and backward probabilities for each of the individual sequences. This is in stark contrast with the estimation described in the previous section, which uses one set of forward and backward probabilities. For this reason, we suggest estimating the location specific parameters as in (8): this approach allows to borrow strength across samples to identify copy number variants that may have lower signal strength, while not increasing substantially the computational costs. A version of the likelihood ratio statistics described for the analysis of tumor samples, can also be used to detect differences in CNV frequencies in a population of cases and controls. We will illustrate this point in the data analysis section.

5 Data Analysis and Examples

As any researcher knows, the analysis of real data sets always poses multiple unexpected challenges. Our goal in this section is simply to exemplify the use of the model we presented, illustrating its potential. In particular, we want to investigate the benefit of taking into account linkage disequilibrium and the advantages of the joint analysis of multiple samples.

We have currently implemented all the algorithms described in sections 3 and 4 in R and most of them in the much faster C++. We are in the process of automating file handling tasks and will be working on the release of a self standing code in the near future.

The performance of any HMM algorithm designed to study copy number variation, crucially depends on the selected values for the parameters. Their estimates are only as good as the training set they are based on. Unfortunately, a large curated data set, including a number of experimentally validated copy number variants, is not available for Illumina data. Both to develop our model and to explore the effectiveness of our algorithm, we have relied on a data set obtained from a genome-wide association study of sporadic amyotrophic lateral sclerosis that comprises 472 patients and 450 matched controls of Dutch origin [20, 21]. Genotyping was carried out with Illumina Hap300 Bead Chips. We used this large genotype collection to define five data sets to use for the purposes detailed below. We will complete fine tuning of parameter values and algorithm implementation once a larger data set becomes available.

A. Curated Set of Copy Number Variants

We compiled a collection of 255 curated copy number variants. A simple script was run through the ALS genome data to identify putative CNV on the basis of consecutive homozygous genotypes and logR values outside a defined window. The results from 70 samples were visually scored by two raters to confirm putative copy number variations on the bases of LogR and BAF values [1]. We further matched these putative variations in copy number with copy number polymorphisms documented in the on-line database of genomic variants (<http://projects.tcag.ca/variation/>).

We considered as true copy number variations the subset of those visually scored that we were able to match with an existing variant in the database, either validated with a PCR study, or independently documented by two different high throughput assays. This led us to a total of 10 homozygous deletions, 105 hemizygous deletions and

140 duplications. Ten of these variants were confirmed with experimental methods. We extracted BAF and LogR information for these copy number variable sites, with a context of 1000 SNPs proximal and 1000 SNPs distal to the CNV (with the exception of the CNVs located at the telomeres).

We use this data set to estimate the parameters of the emission probabilities and for model development and comparisons. This is the most ‘interesting’ of our data sets, as it contains both deletions and duplications in their naturally occurring form and distributed across the entire genome.

B. Signal from the Entire Genome of One Individual

To assess the computational burden of our algorithm with a measure that might be relevant for researchers in the area, we analyze the data for the entire genome of one individual.

C. Simulated Set of Deletions Using Chromosome X Data

We reclustered male X chromosome data to female-only samples, to obtain Log R and BAF signals corresponding to autosomal hemizygous deletions. A set of 270 diploid segments, each 1500 snps long, was obtained by random selection of standard female X chromosome data. In each of these segments we simulated deletions, with positions and lengths generated using the described Markov model with deletion rate $\delta = 0.0035$, and $\eta = 7 \times 10^{-6}$. The BAF and LogR signals for the deleted regions were obtained from the corresponding SNPs in the re-clustered male data.

We use this data set to provide a minimal illustration of how the parameters estimated on the basis of A can be successfully used to analyze an independent data set and to benchmark the performance of our algorithm to other software.

D. Simulated Deletion of a Tumor Suppressor Gene

To generate a challenging data set that mimics the setting of a tumor suppressor gene deletion, we used again true dizygous and hemizygous signals obtained from X chromosome data. Haploid segments comprising the same 1000 SNPs were obtained from 45 different females. A tumor suppressor gene is simulated in the central position with a deletion rate of 0.4, a background deletion rate $\delta = 0.002$, and $\eta = 7 \times 10^{-6}$. The signal for the deleted SNPs is a weighed average of the respective male and female signals. We are using three different weighted averages of hemizygous and haploid signals to

Fig. 7. The top row contains histograms of the lengths in basepairs of all duplications and deletions detected in HapMap individuals in [14]. The bottom row presents the histograms of the proportion of all SNPs that appeared to be in a duplicated and deleted region in each individual in the study.

mimic the noise levels that can characterize real data sets. In the case of tumor data, in particular, since it is often impossible to perfectly separate cancer and normal cells, this mixture model captures the true inhomogeneity of the sample.

We use this simulation to investigate the potential advantage of joint multisample analysis over single sample reconstruction.

E. Common Deletion

We collected signal from 3000 SNPs on chromosome 8 for the entire sample of 922 individuals. This region comprises a known hemizygous deletion of approximately 160 kb, deposited by 11 independent sources in the Database of Genomic Variation (chr8:137585971–138040412).

This data set provides us with the possibility to illustrate our multisample method on real data.

Finally, the entire ALS data set (with the exclusion of the samples in dataset A and B) was used to estimate two-marker haplotype frequencies across the genome, which we use to model LD in our analysis.

Before presenting the results of our data analysis, we need to clarify that, for this purpose, we have actually extended the model described in section 3 in a number of ways. For ease of exposition, we have deliberately omitted these details from the previous sections, but they are important for a successful data analysis. The model we described so far allows only for three possible copy number states. Our training data, however, contains a small number of homozygous deletions. To accommodate for this possibility, we have extended the HMM to include a zero copy number state. Indeed, for the purpose of more general data analysis, one may want to consider other copy number states – this can be done quite easily, as in [2, 24], without substantial modifications of our algorithm. Another modification of the algorithm we described in the previous sections, has to do with the emission probabilities for y : while for ease of exposition, we considered a simple Gaussian distribution, here we use a mixture distribution as in [2] to model outliers.

For the purpose of data analysis, we fixed the parameters of the transition matrix resorting to literature information. We focused on the results reported in [14] with

Table 1. Reconstructed copy number states, per SNP, on data set A, both accounting for LD and not

Accounting for LD					Not Accounting for LD				
True CN	Reconstructed CN				True CN	Reconstructed CN			
	0	1	2	3		0	1	2	3
0	30	0	3	0	0	30	0	3	0
1	0	1,464	40	0	1	0	1,465	39	0
2	3	699	513,451	578	2	3	1,599	512,543	589
3	0	1	31	1,805	3	0	0	31	1,806

HMM parameters are estimated with full knowledge of copy number states.

regard to deletions and duplications scored with Affymetrix genotypes. We obtained the distribution of the lengths of deletions and duplications, as well as of the percentage of SNPs in deleted or duplicated regions for 270 HapMap individuals. These results are presented in figure 7. The median lengths of deletions and duplications are 48,734 and 131,571 base pairs respectively. The median proportions of SNPs in a deletion or duplication are 0.00027 and 0.00049. Using the method of moments, we fixed $\delta = 0.001$, $\gamma = 0.0005$, the marginal probability of the zero copy number state is set to 0.00005, and $\eta = 7 \times 10^{-6}$, with distance measured in number of base pairs separating two loci.

A. Curated Set of Copy Number Variants

We use this data set to estimate the emission parameters for our HMM model. We consider two different approaches in the analysis.

(1) To make optimal use of the information available, we perform maximum likelihood on data set A assuming full knowledge of the copy number states. These parameter values will be used in the rest of the paper, unless otherwise specified, and, in particular, in the analysis of data sets B–E. Using these parameter values to analyze A can artificially increase the sensitivity and specificity of our algorithm, in comparison with others trained elsewhere. However, it is perfectly legitimate to compare the performance of different versions of our algorithm on data A, since they are all trained on the basis of the same information. We use this approach, then, to develop our data analysis strategy and to illustrate the advantage of accounting for linkage disequilibrium.

(2) To assess the impact of full knowledge of copy number states in the previous analysis, and to test the algorithms developed in section 3.3, we estimate all param-

eters using the MM algorithm and ignoring information on copy number status.

(3) Finally to provide the reader with a benchmark of the ‘difficulty’ of data set A, we analyze it with Quanti-SNP and PennCNV.

Note that in large scale data analysis, one would want to tune the emission parameters using EM for each sample analyzed, as done in [24]. We will take this approach in analyzing data set B. We now report the results of these analyses.

(1) Once the parameters of the HMM are estimated with maximum likelihood, we use the Viterbi algorithm to reconstruct copy number states in the training set, with the performance described in table 1. For comparison purposes, we also analyze the same data set with a version of the HMM that does not account for linkage disequilibrium. This leads to very similar values, but for a substantial increase in the number of false deletions. This is exactly the effect we expected: higher linkage disequilibrium leads to stretches of homozygous markers, which can be mistaken for deletions as soon as the LogR signal is lower than expected. When the HMM is modified to take linkage disequilibrium into account, a part of these spurious reconstructed deletions are eliminated.

We are reporting the values in table 1, purely for comparison purposes. On the one hand, because this is our original training set, the sensitivity and specificity we achieve here cannot be necessarily extended to a generic sample. On the other hand, there are a series of obvious modifications to the analysis strategies that increase performance. For example, from table 1 is evident that, even when linkage disequilibrium is taken into account, we are reconstructing some false positive deletions. This excess of deletions is actually clustered in a subset of our samples, which shows larger variability in the values of y , an ex-

Fig. 8. One sample in the training test with a clear decreasing trend in LogR values, which results in increased deletion calls.

Table 2. Reconstructed copy number states, per SNP, on data set A

True CN	Reconstructed CN			
	0	1	2	3
0	33	0	0	0
1	2	1,466	36	0
2	3	189	514,298	289
3	0	1	20	1,816

Linkage disequilibrium is taken into account and data is pre-processed, reducing smooth trends in LogR.

ample of which is given in figure 8. In the last portion of the analyzed region, the average value of y is lower: when this coincides with a series of homozygous markers, the HMM model can interpret this as a deletion. Reconstructed values can be substantially improved if we estimate in-

dividual-specific values for emission parameters, and if we allow the mean of LogR in diploid state to vary smoothly across different genomic regions. This is achieved, for example, by fitting the data with a smooth Lowess curve for each individual and then analyzing the residuals. Table 2 illustrates these results on the training set. A further increase in specificity can be obtained by basing copy number calls not simply on the Viterbi reconstruction, but on the probability of the various states, conditionally on the observed data at each SNP. By selecting to call as deletion or duplications only those location that has conditional probability higher than a given threshold, one can control directly false positive rates. We give these details to illustrate how the performance of a HMM algorithm depends on a number of implementation choices, beside the specific probabilistic model used. This is why it is important to use exactly the same implementation to evaluate the effect of accounting for LD, as we did in table 1. We have not fully optimized our algorithm, partly

Table 3. Reconstructed copy number states, per SNP, on data set A, on the base of parameters estimated ignoring information on CNV

Accounting for LD					Not Accounting for LD				
True CN	Reconstructed CN				True CN	Reconstructed CN			
	0	1	2	3		0	1	2	3
0	33	0	0	0	0	33	0	0	0
1	0	1,472	32	0	1	0	1,472	32	0
2	3	809	513,447	523	2	0	977	513,277	525
3	0	0	22	1,815	3	0	0	22	1,815

We report both the results obtained accounting for LD and assuming linkage equilibrium.

Table 4. Analysis of data set A with QuantiSNP and PennCNV

QuantiSNP					PennCNV				
True CN	Reconstructed CN				True CN	Reconstructed CN			
	0	1	2	3		0	1	2	3
0	33	0	0	0	0	13	0	0	0
1	0	1,467	37	0	1	0	1,461	43	0
2	4	241	514,192	294	2	9	170	514,352	200
3	0	0	30	1,807	3	0	0	39	1,798

SNPs that were assigned copy number larger than 3 by these programs are re-assigned copy number 3 for comparison purposes.

because data set A is of rather limited size, partly because this is beyond the scope of the present work. We are in the process of acquiring a much larger training data that we are planning to use for the release of our software.

(2) We used the MM algorithm described in section 3.3 to estimate the values of all the parameters of the model using data set A, ignoring information on copy number status. Since the data set contains a very small number of homozygous deletions, however, we fixed the emission parameters of this state to their MLE values, obtained in (1). The reconstructed copy number states, using emission parameters so reconstructed, are presented in table 3. Note that one expects this algorithm to perform significantly worse not only of what reported in table 2, but also than algorithms trained on other data sets with full knowledge of copy number states. Despite this, the results summarized in table 3 are still reasonable, and it is still possible to appreciate an advantage in accounting for linkage disequilibrium.

(3) For benchmarking purposes, we also carried out analysis of data set A with QuantiSNP and PennCNV. While possibly at a disadvantage, as trained on different data sets, these algorithms have been thoroughly optimized both in parameter selection, data quality control, and calling procedures. We want to emphasize that it is not our intention to conduct a full fledged comparison between these two algorithms and the one we present here. All of these algorithms are very close to each other and are expected to perform similarly (PennCNV can also handle family data, but this is not applicable in our context). Table 4 reports the results of these analyses. The performances are substantially identical and comparable to those that we report in table 2, underscoring the similarity of all these algorithms. Differences in achieved sensitivity and specificity are likely to be due to implementation details and calling strategies, rather than to underlying probability assumptions. Since our emphasis is on modifications of the likelihood model, to

Table 5. Reconstructed copy number states, per SNP, on data set C for two versions of our algorithm (top row) not accounting for LD, and QuantiSNP and PennCNV (bottom row)

Viterbi					Posterior Prob				
True CN	Reconstructed CN				True CN	Reconstructed CN			
	0	1	2	3		0	1	2	3
1	0	1,189	135	0	1	0	1,061	61	0
2	1	63	403,478	134	2	0	2	402,088	59

QuantiSNP					PennCNV				
True CN	Reconstructed CN				True CN	Reconstructed CN			
	0	1	2	3		0	1	2	3
1	0	1,189	134	0	1	0	1,132	192	0
2	2	34	403,535	91	2	0	28	402,088	107

account for linkage disequilibrium and to allow for joint multiple sample analysis, it is more meaningful to use the same algorithm and explore the effect of these modifications one at the time: this is the spirit of the comparison presented in tables 1 and 6. The likelihood assumptions underlying QuantiSNP and PennCNV are substantially equivalent to the ones we make in the single sample analysis that does not take into account LD. We hope that our results may convince the curators of these programs to implement some of the modifications we suggest.

B. Signal from the Entire Genome of One Individual

We analyze the entire genome sequence of one individual, keeping the parameters of the HMM fixed at the values estimated on data set A, with the exception of the emission parameters for the diploid state which are estimating using the entire data for this individual. We require that the local posterior probability of copy number other than 2 to be greater than 0.99 in order to identify a copy number variant. According to this criterion, 20 regions appear to harbor putative copy number variations. Figure 9 gives details on the nature of one of these identified variations. Running time for this analysis was 30 s on a laptop with 2.8 GHz single processor, using our C++ implementation.

C. Simulated Set of Deletions Using Chromosome X Data

To validate the applicability of the parameters estimated on A to other data, we consider simulation C. This is also a reasonable data set where to compare the perfor-

Table 6. Reconstructed loss rates for the tumor suppressor gene, as in simulation D

	Proportion of hemizygous signal		
	0.2	0.5	0.8
Single	0.0145 (0.004)	0.1222 (0.011)	0.2871 (0.015)
Joint	0.0486 (0.009)	0.2526 (0.013)	0.3636 (0.016)

We consider three contamination levels for the genotype data, corresponding to proportions of true hemizygous signal of 0.2, 0.5, and 0.8. The true tumor suppressor gene loss rate is 0.4. We present the proportion of samples that are identified as carrying a deletion at the tumor suppressor gene location by single sample analysis and by joint multisample analysis. Respective standard deviations are indicated in parenthesis.

mance of our algorithm with the one of PennCNV and QuantiSNP. In particular, since we have argued that these should be comparable to our algorithm when we do not account for LD, we present the results of these analysis, to validate our claims. As we have remarked, there are variations in performances due to the calling strategies adopted. For our algorithm we present two results: the Viterbi reconstruction, and a setting where SNPs are called only when the posterior probability of one copy number is larger then 0.99. For QuantiSNP we report calls with Bayesfactor larger then 10, as recommended in the manual. Results of our algorithm, together with those of PennCNV and QuantiSNP are presented in table 5. Clearly, the performance of our algorithm remains ac-

Fig. 9. Data relative to the small arm of chromosome 2 for the studied individual. On the x -axis, we report the positions of queried SNPs in base pairs. The plots, from top to bottom, display the posterior probability of a copy number aberration, the copy number state reconstructed with Viterbi, the values of x (B allele frequency) and y (LogR).

ceptable on data outside the training set and is indeed remarkably similar to that of the published software.

D. Simulated Deletion of a Tumor Suppressor Gene

Since our multisample analysis method is in part motivated by the study of cancer cells, we explore its effectiveness in this context, relying on the simulation described in D. We consider three contamination levels for the genotype data, corresponding to a proportion of 0.2, 0.5, and 0.8 of truly hemizygous signal used to generate the deletion signal. For each of these noise levels, we generate 13 data sets according to the parameters described, and we analyze them using the single sample analysis and the joint multisample method. For each of the data sets, we then calculate the proportion of individuals identified as carrying a deletion at the tumor suppressor gene. Results are in table 6: clearly the multisample analysis enables us to correctly reconstruct a larger fraction of the deletions, for all noise levels. False positive rates of

the two methods are entirely comparable (data not shown).

E. Common Deletion

To illustrate the applicability of our multisample method on actual data, we consider data set E. Analyzing firstly one sample at the time, we identify 54 individuals carrying the deletion. The average Viterbi path across all 922 individuals in this region is presented in figure 10. The polymorphic deletion, which spans 17 SNPs is clearly identified. Subsequently, we analyze this region, estimating a location specific loss and duplication rate for SNPs in the neighborhood. Because multisample analysis of this kind is more sensitive, we want to compare the frequency of individuals with a deletion according to the single sample analysis conducted with Viterbi with the location specific estimate of deletion. In this case, the numbers turn out to be equivalent, suggesting that we had already captured all deletions with our original scan.

Conducting the multisample analysis, however, we realized one strength of the single sample analysis: its greater robustness to smooth fluctuations of the logR signal. We have already mentioned these (fig. 8), but it is in the analysis of data set E that their importance becomes more apparent. Consider the situation presented in figure 11. The LogR signal, in the top panel, clearly fluctuates around a trend, captured here with a spline regression. These variations in LogR have little information on copy number and are particularly challenging for algorithms that ignore the BAF signal. To illustrate this, we have analyzed this sample with the Fused lasso method, as implemented in [19]: not surprisingly its reconstruction is much noisier than the HMM one (fig. 11).

To date, we do not have a clear understanding of the experimental effects behind this trend. Other groups have reported association with the probes CG content or

Fig. 10. Average of Viterbi paths across 922 individuals in 3000 SNPs surrounding a common deletion, comprising 16–17 SNPs.

Fig. 11. Smooth trends in LogR signal. The top display provides LogR values for the 3000 analyzed SNPs in one individual. The middle plot presents the result of the cghFlasso algorithm applied on the LogR values. The bottom display reports the reconstructed copy number state by our HMM algorithm that uses both LogR and BAF values.

Fig. 12. Multisample analysis of 60 SNPs in the proximity of the studied deletion. The top display contains the proportion of individuals whose Viterbi path was equal to 1 among the 922 analyzed. The middle plot presents the location-specific probability of deletion estimated using the algorithm described in section 4.1. The black line refers to the entire sample, blue line to ALS cases and red line to controls. The bottom plot presents the LR test statistics for a difference in location-specific propensity of loss between cases and controls.

DNA concentrations [17, 25]. In our data set, we identify samples without apparent trend, and two clusters that share two different smooth patterns. Multisample analysis is particularly sensitive to these, as it tends to interpret them as weak signal. In order to avoid spurious results, we first pre-process the raw LogR values to estimate a smooth trend and remove it. We then run single sample Viterbi and multisample analysis on the residuals of this analysis. The single sample Viterbi are not substantially different from the ones obtained on the original raw data, but the multisample results are much less noisy. Figure 12 illustrates these results. In the top panel, we have the frequency of deletion according to the Viterbi paths. In the middle panel, the estimated frequency of deletions using multisample analysis for the entire sample, cases, and controls. In the bottom panel the value of the log-likelihood ratio statistics for comparing the hypothesis of the same frequency of deletion in cases and controls versus different frequency. Clearly, it appears that the overall sample frequency of deletion is well captured by the single sample Viterbi analysis. There are no

differences between cases and controls; while the deletion is longer in some of the controls, it is hard to assess the relevance of this finding.

6 Discussion

The past year has witnessed the beginning of large scale genome-wide association studies. Their number is constantly increasing and the analysis of the data they generate will allow us to learn a great deal on genetic variation. An important class of variation is represented by copy number changes. Illumina genotyping technology allows to detect variations in DNA amounts, and hidden Markov models have proven to be a useful tool in this process [2, 24]. We have described a novel hidden Markov model that takes into account linkage disequilibrium between neighboring markers. We have further extended the methodology to allow joint analyze multiple samples. These two extensions make our algorithm particularly useful for the analysis of data obtained in genome-wide association

scans. Accounting for LD modestly increases specificity, reducing the number of erroneously called deletions. Importantly, these would tend to cluster in the same genetic locations with high LD, and thus would likely raise the interest of researchers looking for polymorphisms. Our multisample analysis increases the power to detect copy number polymorphisms and provides the researcher with a tool to simultaneously identify copy number variants and test for their association with a trait of interest.

Acknowledgements

We thank Professor Kenneth Lange for stimulating discussions on the topics of this paper. C.S. gratefully acknowledges support of NSF grant DMS0239427 and NIH grants RO1 GM53275, RO1 MG078075-01. J.H.V. was supported by a grant of The Brain Foundation of the Netherlands.

References

- Blauw H, Veldink J, van Es M, van Vught P, Saris C, van der Zwaag B, Frank L, Burbach P, Wokke J, Ophoff R, van den Berg L: Copy number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. *Lancet Neurol* 2008;7:319–326.
- 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;35:2013–2025.
- Conrad DF, Andrews TD, Carter NP, Hurler ME, Pritchard JK: A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006;38:75–81.
- Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, Stoeckert CJ Jr, Weber BL, Maris JM, Grant GR: STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* 2006;16:1149–1158.
- 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;14:1–11.
- 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;36:949–951.
- Lange K: *Applied Probability*. Springer, New York, 2004.
- Lange K: *Optimization*. Springer, New York, 2004.
- Newton M, Gould M, Reznikoff C, Haag J: On the statistical analysis of allelic-loss data. *Stat Med* 1998;17:1425–1445.
- Newton M, Lee Y: Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* 2000;56:1088–1097.
- Newton MA: Discovering combinations of genomic alterations associated with cancer. *J Am Stat Ass* 2002;97:931–942.
- Peiffer D, Le J, Steemers F, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw C, Belmont J, Cheung S, Shen R, Barker D, Gunderson K: High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 2006;16:1136–1148.
- Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C: Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci* 2006;103:8006–8011.
- 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, Gonzalez JR, Gratacos 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, Hurler ME: Global variation in copy number in the human genome. *Nature* 2006;444:444–454.
- Sabatti C, Lange K: Bayesian Gaussian mixture models for high density genotyping arrays. *JASA* 2007;103:89–100.
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: Large-scale copy number polymorphism in the human genome. *Science* 2004;305:525–528.
- Stefansson H, et al: Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;455:232–236.
- Tang H, Coram M, Wang P, Zhu X, Risch N: Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 2006;79:1–12.
- Tibshirani HR, Wang P: Spatial smoothing and hot spot detection for CGH data using the Fused Lasso. *Biostatistics* 2008;9:18–29.
- van Es MA, Van Vught PW, Blauw HM, Franke L, Saris CG, Andersen PM, Van Den Bosch L, de Jong SW, van't Slot R, Birve A, Lemmens R, de Jong V, Baas F, Schelhaas HJ, Sleegers K, Van Broeckhoven C, Wokke JH, Wijmenga C, Robberecht W, Veldink JH, Ophoff RA, van den Berg LH: ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol* 2007;6:869–877.
- van Es MA, van Vught PW, Blauw HM, Franke L, Saris CG, Van den Bosch L, de Jong SW, de Jong V, Baas F, van't Slot R, Lemmens R, Schelhaas HJ, Birve A, Sleegers K, Van Broeckhoven C, Schymick JC, Traynor BJ, Wokke JH, Wijmenga C, Robberecht W, Andersen PM, Veldink JH, Ophoff RA, van den Berg LH: Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nat Genet* 2008;40:29–31.
- Wang H, Lee Y, Nelson S, Sabatti C: Inferring genomic loss and location of tumor suppressor genes from high density genotypes. *J French Stat Soc* 2005;146:153–171.
- Wang H, Lin C, Service S, The international collaborative group on isolated populations, Chen Y, Freimer N, Sabatti C: Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density. *Hum Hered* 2006;62:175–189.
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, 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;17:1665–1674.
- Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res* 2008;36:e126. Epub 2008 Sept 10.