Markov models for inferring Copy Number Variations from genotype data

Hui Wang
Jan Veldink
Roel Opoff
et al.

Publication Date
2008

Markov models for inferring Copy Number Variations from genotype data on Illumina platforms

Hui Wang[1], Jan Veldink[2], Roel Opoff[2,3], and Chiara Sabatti[3,4]

1 Perlegen Sciences, Mountain View, CA

2 Departments of Neurology and Medical Genetics, University of Utrecht, The Netherlands

3 Department of Human Genetics, UCLA, Los Angeles CA

4 Department of Statistics, UCLA, Los Angeles, CA.

UCLA Statistics Department Preprint # 533

January 2008

Abstract

We develop an algorithm to analyze data from Illumina genotyping arrays for the detection of copy number variations in a single individual or in a random sample of individuals. We use a Hidden Markov Model framework, appropriately extended to take into account linkage disequilibrium between nearby loci. We describe a multisample approach to estimate the frequency of copy number variants in the population. With appropriate dataset, our methodology simultaneously analyzes the data for copy-number variation and tests for association between this and a disease trait of interest.

1 Copy number variation and genotyping arrays

Two important surveys appeared in 2004 [5, 14] documented the presence of copy number variation in the genome in unsuspected high frequencies, both in the form of deletion of small genomic segments as well as duplication. 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, [17]) as well as duplications, on the other hand, patterns of Mendelian inconsistencies can also provide 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, and should, be used to detect copy number variations. Linkage disequilibrium by itself explain the presence of stretches of homozygous markers in the genome. Modeling linkage disequilibrium helps to narrow the number of genomic region that could harbor copy number variation on the basis of genotype information alone [18], however, it is not sufficient. Only by looking at variation on the intensity levels associated to DNA amounts at each marker, one can reliably distinguish deletions, duplications or homozygosity corresponding to a regular diploid genome segment. 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 [4], while [12] 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 the specific problem we would like to tackle, which is in large part still open.

1. Our main goal is to develop an algorithm to analyze data from Illumina genotyping arrays for the detection of copy number variations in a single individual or in a random sample of individuals.

2. In the present effort, we use a Hidden Markov Model framework, appropriately extended to take into account linkage disequilibrium between nearby loci;

3. We adopt a Bayesian approach with which information on copy number variation detected in previous studies informs the analysis;

4. With appropriate dataset, our methodology simultaneously analyzes the data for copy-number variation and tests for association between this and a disease trait of interest.

2 Illumina genotype data

We have become interested in studying the data obtained with the Illumina genotyping arrays for a number of reasons. On the one hand, this is the platform of choice for a number of studies we

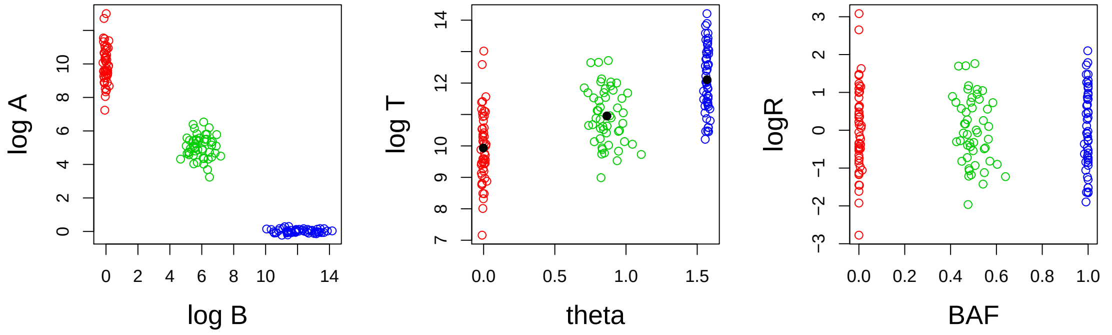


Figure 1: (a) Scatterplot of the logarithm of intensity values for A and B allele in one SNP for 150 individuals. (b) Scatterplot of the same points of (a) in the changed coordinate system. (c) standardization and re-scaling of the clusters to define LogR and BAF.

are involved in (2000 dutch individuals (case control sample), 5000 Finnish individuals (population sample), and a smaller family sample from Costa Rica). On the other hand, our analysis of Illumina data, as well as the experience of other researchers (Feingold, personal communications), indicates that this platform provides particularly clean signal with considerable potential for copy number variation detection. Finally, this is a relatively unexplored area, with perhaps [2, 19] the only contributions to date.

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 Log R. 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. 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 1(a) present 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 1(b) illustrate 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 1(a). Finally, the centers of the three clusters are compared to reference 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. [13].

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 hemizygos 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 position 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

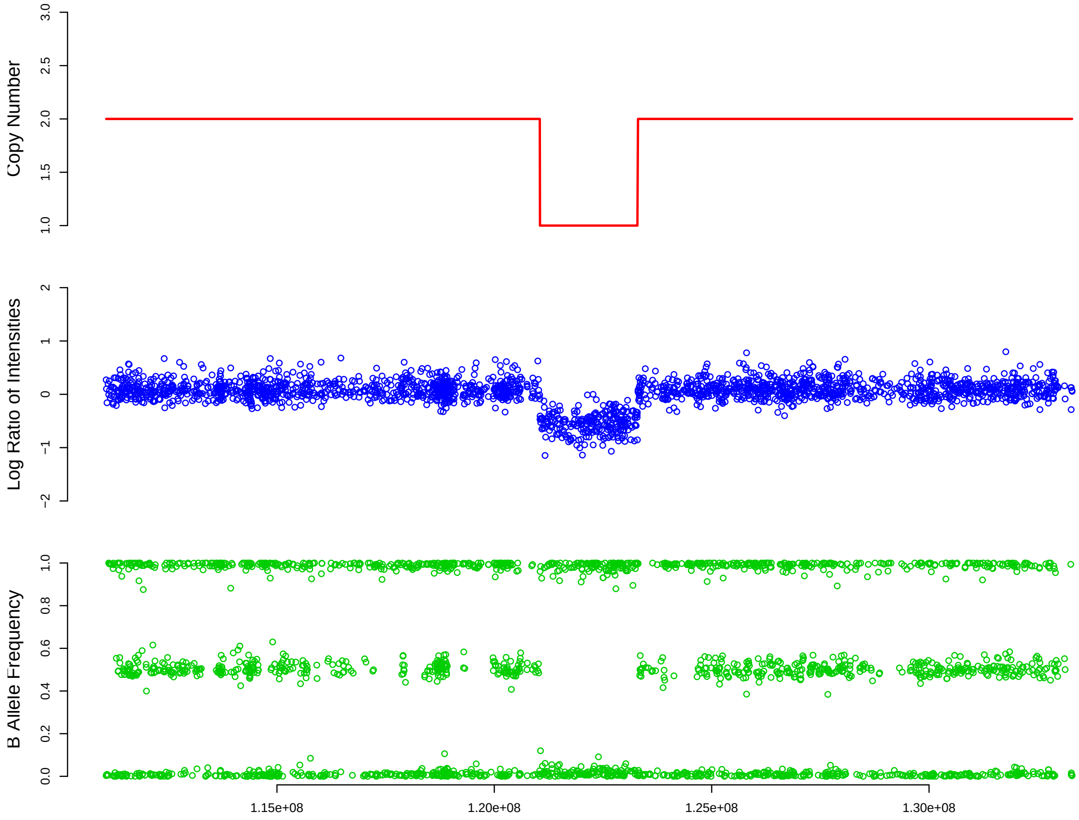


Figure 2: A deletion encompassing 245 SNPs on Chromosome 4. Data for additional 1000 SNPs flanking the deletion is also presented. On the x-axis, we report the positions of queried SNPs in base pairs. The top plot displays the copy number values; the central plot presents the y LogR values associated with each SNP; and the bottom plot displays the x B allele frequency values.

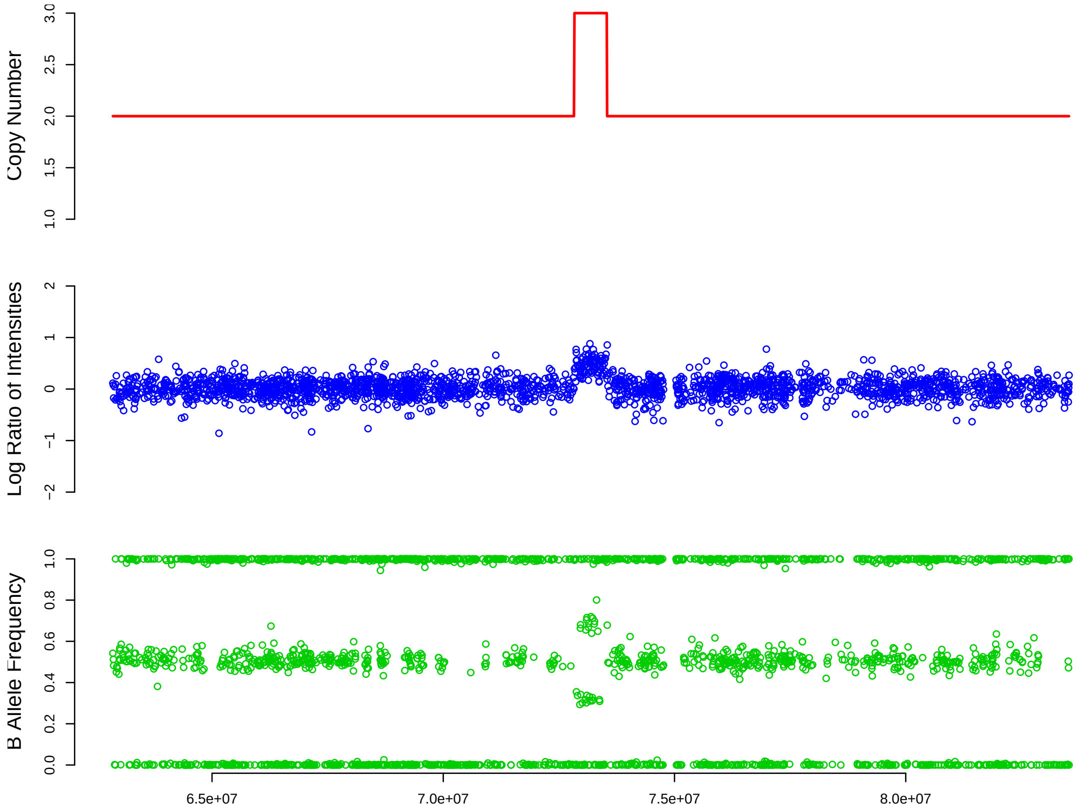


Figure 3: A duplication encompassing 82 SNPs on Chromosome 12. Data for additional 1000 SNPs flanking the duplication is also presented. On the x-axis, we report the positions of queried SNPs in base pairs. The top plot displays the copy number values; the central plot presents the y LogR values associated with each SNP; and the bottom plot displays the x B allele frequency values.

y allows to separate homozygous signal corresponding to a 1 or 2 copy numbers, 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, a recent literature contribution [2] documents a HMM for this problem. In the following section we describe our own model that we believe has more realistic features.

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 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 Newton [8, 9, 10] 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.

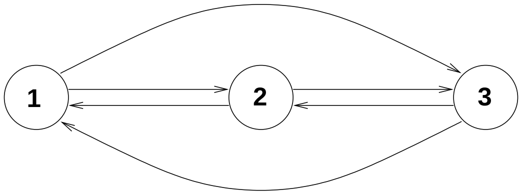


Figure 4: Generic infinitesimal rates for the homogeneous copy number process with three states.

3.1 A homogeneous Markov model for Copy Number States

Let $\{\pi_t\}$ be the continuous process that describes copy number status of one individual across the genome. We model $\{\pi_t\}$ 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. However, we ask the reader to bear with us for a little. 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.

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

$$\mathrm{p}(\pi_t = 1) \;=\; \frac{\phi}{\phi+\lambda+\xi}, \quad \mathrm{p}(\pi_t = 2) \;=\; \frac{\lambda}{\phi+\lambda+\xi}, \quad \mathrm{p}(\pi_t = 3) \;=\; \frac{\xi}{\phi+\lambda+\xi}. \qquad (1)$$

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 $i = 1, \ldots n$, corresponding to the genotyped SNPs. Using standard continuous Markov chain theory [6], for model (a), we can obtain a following discrete time transition matrix between points at distance d:

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

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

$$\begin{aligned} t^b_{11}(d) &= \frac{\gamma}{\gamma+\delta}e^{-(1-\gamma-\delta)\eta d} + \delta + \frac{\delta}{\gamma+\delta}(1-\delta-\gamma)e^{-\eta d} \\ t^b_{13}(d) &= -\frac{\gamma}{\gamma+\delta}e^{-(1-\gamma-\delta)\eta d} + \gamma + \frac{\gamma}{\gamma+\delta}(1-\delta-\gamma)e^{-\eta d}, \end{aligned}$$

where $t^b_{31}(d)$ and $t^b_{33}(d)$, 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 to be rather unreasonable.

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 a part 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.

3.2 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.

Figures 5 presents summary statistics of the distribution of x and y in the training dataset that guided our modelling choices (for details on this dataset, 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

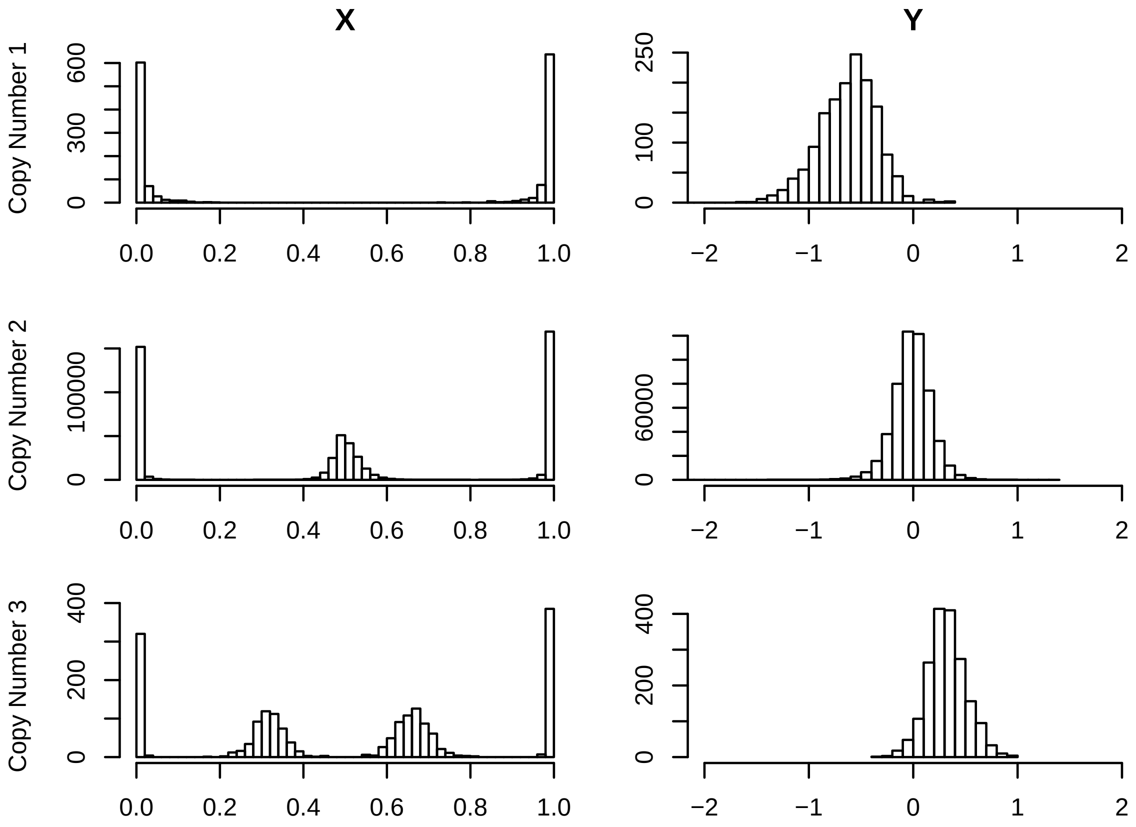


Figure 5: Histogram of the values of the emitted signals x and y for approximately 500,000 SNPs genotyped in a set of individuals including known deletions and duplications. Each row corresponds to one copy number. (See section 5 for more detail)

previous work on detection of deletions and estimation of inbreeding coefficients [18] 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 [18] we describe an extension of the typical HMM framework to account for this dependency.

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 deletion-duplication mechanism still operates, but the genotype at marker $j = i + 1$ no longer occurs independently of the genotype at the previous marker i. As an illustration, the transition between states A and AB can be calculated as

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

where $t_{12}(d_{ij})$ is the probability of going from 1 to 2 copies over a distance d_{ij} and $q_{ij}(A \mid A)$ and $q_{ij}(B \mid A)$ are conditional haplotype frequencies. 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 $\mathcal{T} = (\delta, \gamma, \eta)$.

Note that our previous work [18] suggests that the simple order one Markov model that we use here to describe linkage disequilibrium may be inadequate for some genomic regions. Nevertheless, it is reasonable to believe that this should not pose a serious difficulty in the present context, where inference will not depend only on x values, but also on the y values that are unaffected by

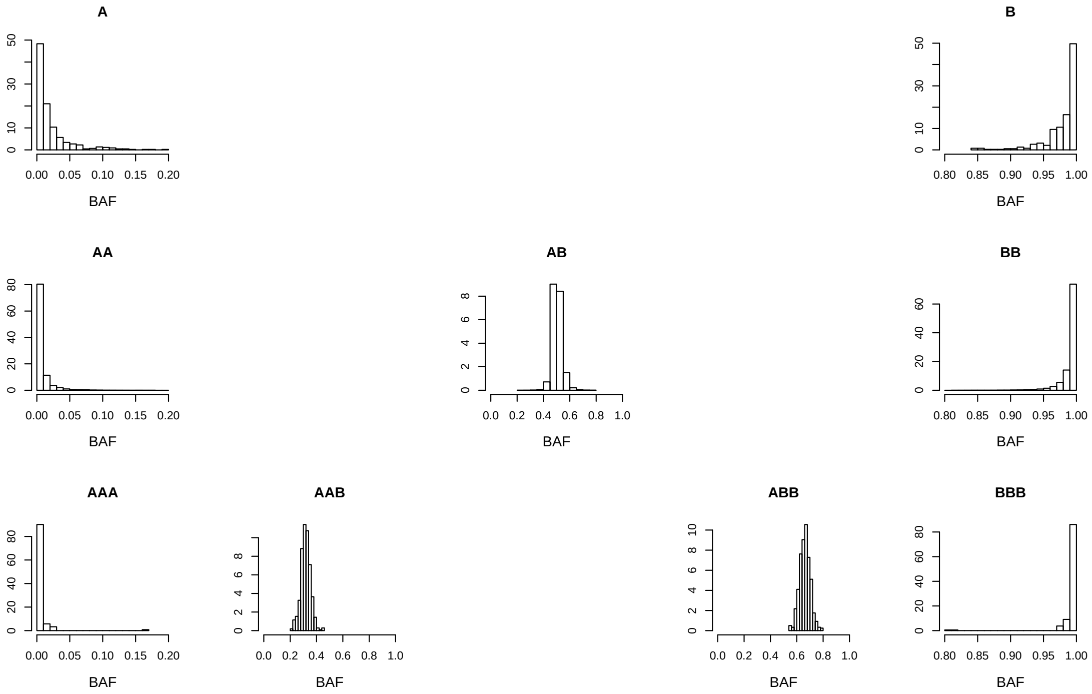


Figure 6: Histograms of the value of the emitted x signals, conditional on underlying genotype, for approximately 500,000 SNPS including deletions and duplications. Each row corresponds to one copy number, each column to one of the five relevant allele ratios: 0, 1/3, 1/2, 2/3, 1. The specific genotype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

linkage disequilibrium.

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 datasets (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 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:

$$
\begin{array}{rcl}
f(x|A) = f(x|AA) = f(x|AAA) & = & \begin{cases} \omega & x = 0 \\ (1-\omega)\lambda e^{-\lambda x} & x \neq 0 \end{cases} \\
f(x|B) = f(x|BB) = f(x|BBB) & = & \begin{cases} \omega & x = 1 \\ (1-\omega)\lambda e^{-\lambda(1-x)} & x \neq 1 \end{cases} \\
f(x|AB) & = & \lambda_2/2 \exp\{-\lambda_2|x-0.5|\} \\
f(x|AAB) & = & \mathcal{N}(1/3, \nu) \\
f(x|ABB) & & \mathcal{N}(2/3, \nu)
\end{array}
$$

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

3.3 Algorithms for evaluating conditional probabilities and parameter estimation

To evaluate the probability of the data $O = \{o_i\}_{i=1}^n$ for one individual, we resort to the standard hidden Markov model machinery. Rather then attempting to calculate directly the sum

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

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

$$\begin{aligned} \alpha(\pi_i) &= \sum_{\pi_{i-1} \in \mathcal{S}} \alpha(\pi_{i-1}) t(\pi_i|\pi_{i-1}) e(o_i|\pi_i) \\ \beta(\pi_i) &= \sum_{\pi_{i+1} \in \mathcal{S}} \beta(\pi_{i+1}) t(\pi_{i+1}|\pi_i) e(o_{i+1}|\pi_{i+1}). \end{aligned}$$

The sample probability can then be obtained as $\text{Pr}(O) = \sum_{\pi_n \in \mathcal{S}} \alpha(\pi_m)$. 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

$$\begin{aligned} \xi_{i,i+1}(s_j, s_k) &= \frac{\alpha(\pi_i) t(\pi_i|\pi_{i+1}) e(o_{i+1}|\pi_{i+1}) \beta(\pi_{i+1})}{\text{Pr}(O)} \\ \rho_i(s_k) &= \sum_{\pi_{i+1} \in S} \xi(\pi_i, \pi_{i+1}). \end{aligned}$$

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

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

We tackle this using an iterative algorithm in the MM framework [7]. 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 [7]). The theory of the EM algorithm offers a recipe for identifying a minorizing function for the logarithm of the likelihood. Indicating with $\Pi = \{\pi_i\}_{i=1}^n$ the missing data corresponding to the unobserved copy-number states, we can use $q(\mathcal{T}, \mathcal{E}|\mathcal{T}^\ell, \mathcal{E}^\ell) = \mathrm{E}(\log \mathcal{L}(\mathcal{T}, \mathcal{E}|O, \Pi))$, 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:

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

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

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

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,

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

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

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

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

$$\sum_{j,k=1}^{9} \sum_{i=2}^{n} \xi_{i-1,i}^{\ell}(s_j, s_k) \log t_{\text{cn}(s_j),\text{cn}(s_k)}(d_{i,i+1}), \tag{3}$$

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:

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

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:

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

Plugging these expressions in (3), we obtain a simple function of δ and γ, separated from η. Let $\kappa_{i.i+1}^{\ell}(j,k) = \sum_{l,t:\text{cn}(l)=j,\text{cn}(t)=k} \xi_{i,i+1}^{\ell}(s_j, s_k)$. 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 $N^\ell_{1,1} = \sum_i \kappa^\ell_{i,i+1}(1,1)\alpha^\ell_i$, $N^\ell_{2,2} = \sum_i \kappa^\ell_{i,i+1}(2,2)\zeta^\ell_i$, $N^\ell_{3,3} = \sum_i \kappa^\ell_{i,i+1}(3,3)\beta^\ell_i$. 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

$$\begin{aligned} M^\ell_i &= \sum_{j \neq k} \kappa^\ell_{i,i+1}(j,k) \\ S^\ell_i &= \sum_j \kappa^\ell_{i,i+1}(j,j) \\ W^\ell_i &= \kappa^\ell_{i,i+1}(1,1)\alpha^\ell_i + \kappa^\ell_{i,i+1}(2,2)\zeta^\ell_i + \kappa^\ell_{i,i+1}(3,3)\gamma^\ell_i \end{aligned}$$

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

$$g(\eta|\mathcal{T}^\ell, \mathcal{E}^\ell) = \sum_i \log(1 - e^{-\eta d_{i,i+1}})(O^\ell_i + W^\ell_i) - \eta d_{i,i+1}(S^\ell_i - W^\ell_i).$$

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 [7]):

$$\begin{aligned} \frac{\partial g(\eta|\mathcal{T}^\ell, \mathcal{E}^\ell)}{\partial \eta} &= \sum_i \frac{d_{i,i+1} e^{-\eta d_{i,i+1}}}{1 - e^{-\eta d_{i,i+1}}}(O^\ell_i + W^\ell_i) - \sum_i d_{i,i+1}(S^\ell_i - W^\ell_i) \\ \frac{\partial^2 g(\eta|\mathcal{T}^\ell, \mathcal{E}^\ell)}{\partial \eta^2} &= -\sum_i d^2_{i,i+1} \frac{e^{-\eta d_{i,i+1}}}{(1 - e^{-\eta d_{i,i+1}})^2}. \end{aligned}$$

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 then 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 $(1/\sigma_1^2, 1/\sigma_2^2, 1/\sigma_3^2, 1/\nu^2)$, 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_1^2). 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 $1/\sigma_1^2$ with parameters (a_1, b_1). Then, the updates for the parameters are:

$$\mu_1^{(\ell+1)} = \frac{\frac{1}{\sigma_1^2}\sum_{i=1}^{n} y_i w_i^{\ell}(1) + p_1 m_1}{\frac{1}{\sigma_1^2}\sum_{i=1}^{n} w_i^{\ell}(1) + p_1}$$

$$\sigma_1^{(\ell+1)} = \sqrt{\frac{\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_1^{(\ell+1)}) w_i^{\ell}(1) + b_1}{\frac{1}{2}\sum_{i=1}^{n} w_i^{\ell}(1) + a_1 - 1}}$$

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.

4.1 Identifying cancer related genes

We start considering the setting studied by Newton [8, 9, 10]: 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 $k = 1, \ldots, m$ individuals is then equal to:

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

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 an EM framework, we consider $\pi_i^1, \ldots, \pi_i^m$ as the missing data, and we can iterate on the basis of the following

$$\begin{aligned} \mathrm{E}(\pi_i^k = 1 | O, \delta_i^\ell, \gamma_i^\ell) &\propto \delta_i^\ell P(O^k | \pi_i^k = 1) \\ \delta_i^{(\ell+1)} &= \sum_{k=1}^{m} \mathrm{E}(\pi_i^k = 1 | O, \delta_i^\ell, \gamma_i^\ell) / m, \end{aligned} \tag{4}$$

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:

$$L_i = \begin{cases} \log_{10} \dfrac{\mathcal{L}(\mathcal{E}, \mathcal{T}, \delta_i^* \,|\, O^1, \ldots, O^m)}{\prod_{k=1}^m \mathcal{L}(\mathcal{E}, \mathcal{T} \mid O^k)} & \delta_i^* = \mathrm{argmax} \mathcal{L}(\mathcal{E}, \mathcal{T}, \delta_i^* \,|\, O^1, \ldots, O^m) > \delta \\ 0 & \delta_i^* < \delta \end{cases} .$$

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^i 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 $m \to \infty$, can be roughly approximated using the known results for likelihood ratio tests: under H_0^i, $2 \ln 10 L_i$ 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).

4.2 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 copy number variations derived from 1). CN polymorphisms, instead, call for a non homogeneus component in our process. The approach described in section 4.1, however, is 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.

We distinguish here two problems: a) how to take into account known CNPs; b) how to discover new CNPs. a) The HMM framework can be easily adapted to account for known CNPs. To avoid unnecessary complications, let's focus on modifications of the transition matrix defined on copy number states (2). To incorporate knowledge of known CNPs, one can consider two additional states: $1P$ and $3P$, corresponding to copy number equal to 1 and 3 in a CN polymorphic region. These states are visited with positive probability only in portions of the genome that exhibit CN polymorphisms. Once in states $1P$ or $3P$, the process has to remain there till the end of the CNP region. To fix ideas, suppose that the only known CNP is covered by SNPs i through $i+m$ with copy number equal to 1 (3) in a portion $d(g)$ of the population. Then we will have three transition matrices

$$T_{j-1,j} = \begin{array}{c|ccccc} & 1P & 1 & 2 & 3 & 3P \\ \hline 1P & 0 & \delta & (1-\delta-\gamma) & \gamma & 0 \\ 1 & 0 & t_{11}(d_{i,i+1}) & t_{12}(d_{i,i+1}) & t_{13}(d_{i,i+1}) & 0 \\ 2 & 0 & t_{21}(d_{i,i+1}) & t_{22}(d_{i,i+1}) & t_{23}(d_{i,i+1}) & 0 \\ 3 & 0 & t_{31}(d_{i,i+1}) & t_{32}(d_{i,i+1}) & t_{33}(d_{i,i+1}) & 0 \\ 3P & 0 & \delta & (1-\delta-\gamma) & \gamma & 0 \end{array} \qquad j \neq i, \ldots, i+m$$

$$T_{i-1,i} = \begin{array}{c|ccccc} & 1P & 1 & 2 & 3 & 3P \\ \hline 1P & d & \delta(1-d-g) & (1-\delta-\gamma)(1-d-g) & \gamma(1-d-g) & g \\ 1 & d & t_{11}(d_{i,i+1})(1-d-g) & t_{12}(d_{i,i+1})(1-d-g) & t_{13}(d_{i,i+1})(1-d-g) & g \\ 2 & d & t_{21}(d_{i,i+1})(1-d-g) & t_{22}(d_{i,i+1})(1-d-g) & t_{23}(d_{i,i+1})(1-d-g) & g \\ 3 & d & t_{31}(d_{i,i+1})(1-d-g) & t_{32}(d_{i,i+1})(1-d-g) & t_{33}(d_{i,i+1})(1-d-g) & g \\ 3P & d & \delta(1-d-g) & (1-\delta-\gamma)(1-d-g) & \gamma(1-d-g) & g \end{array}$$

$$T_{k-1,k} = \begin{array}{c|ccccc} & 1P & 1 & 2 & 3 & 3P \\ \hline 1P & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & t_{11}(d_{i,i+1}) & t_{12}(d_{i,i+1}) & t_{13}(d_{i,i+1}) & 0 \\ 2 & 0 & t_{21}(d_{i,i+1}) & t_{22}(d_{i,i+1}) & t_{23}(d_{i,i+1}) & 0 \\ 3 & 0 & t_{31}(d_{i,i+1}) & t_{32}(d_{i,i+1}) & t_{33}(d_{i,i+1}) & 0 \\ 3P & 0 & 0 & 0 & 0 & 1 \end{array} \qquad k = i+1, \ldots, i+m$$

This described modification assumes that one knows exactly the frequency of copy number variants in the genome. More realistically, we may simply be aware of the presence of a copy number variant, without knowing exactly its frequency. This situation can be dealt with using the framework we are about to describe and the use of appropriate priors.

To partially account for the presence of CN polymorphisms, one can modify the homogeneous model in 2, introducing 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. 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 $\{\gamma_i\}_{i=1}^n$: $\Pr(\{\delta_i\}) \propto \sum_{i=2}^n |\delta_i - \delta_{i-1}|$ (procedures that use such penalty have been recently suggested for the analysis of CGH data under the name of "fused lasso").

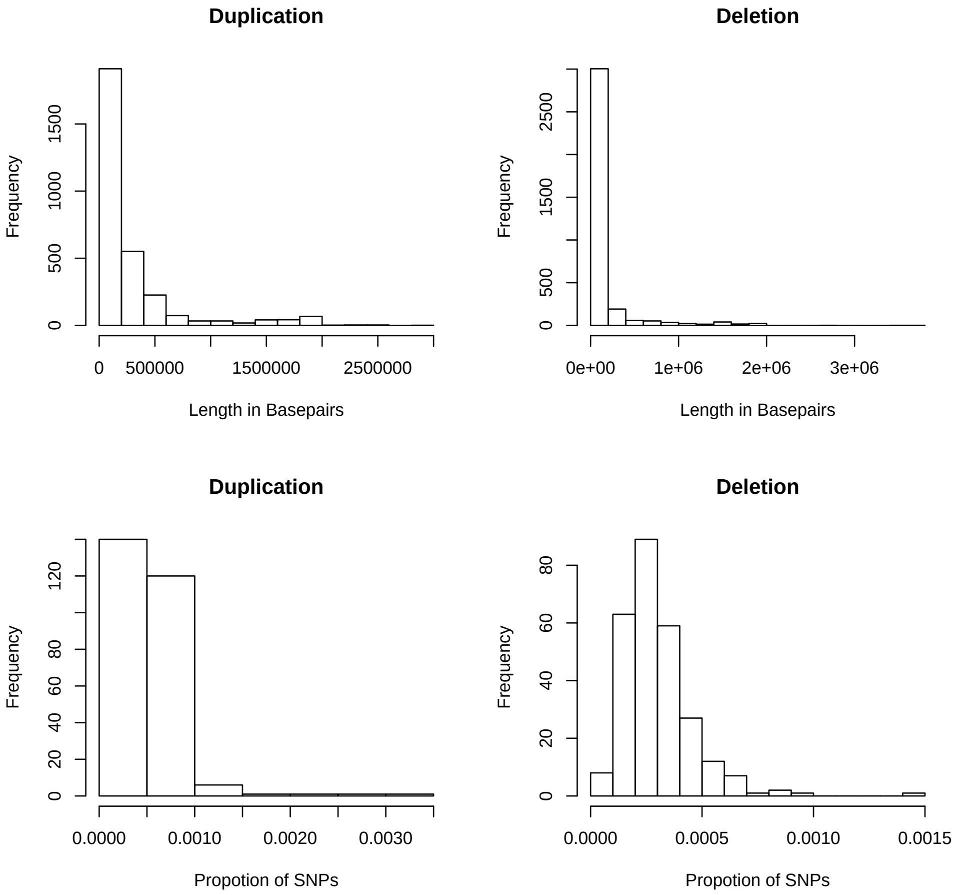
However, one has to be careful not to underestimate the computational costs involved in such procedure. For each iteration, one needs to recompute all the forward and backward probabilities for each of the individual sequences. This is in stark contrast with the estimation described in the previous section, which uses one set of forward and backward probabilities. For this reason, we suggest an heuristic, two stage procedure to detect CN polymorphisms. (I) Estimate the location specific parameters as in (5); (II) regularize these estimates using a fused lasso procedure; (III) finally, focus on regions for which $\delta_i + \gamma_i > C$, with the cut-off C used to define polymorphisms (for example $C = 0.05$). Step (I) increases the chances of detecting CN variations that have small signal and would be undetected with single sample analysis or aggregations of it. Step (II) helps reduce the variance of the estimated CN variant probabilities. Step (III) focuses the researcher attention to CN variants that are of greater biological interests and serves also as a statistical significant threshold (all loci are evaluated with the same number of samples, so that the signal in favor of increased copy number frequency depends only on the number of detected variations.

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

Both to develop our model and to test the effectiveness of our algorithm, we have relied on a dataset collected for genome-wide association study of sporadic amyotrophic lateral sclerosis, and comprising 461 patients and 450 matched controls of dutch origin [16]. Genotyping was carried out with Illumina Hap300 Bead Chips. The entire dataset was used to estimate two-marker haplotype frequencies.

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 [12] with 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 48734 and 131571 base pairs respectively. The median proportions of SNPs in a deletion and duplication are 0.00027 and 0.00049. Using the method of moments, we fixed then $\delta = 0.001$, $\gamma = 0.0005$, and $\eta = 7 \times 10^{-6}$, with distance measured in number of base pairs separating two loci.



Duplication

Deletion

Duplication

Deletion

Figure 7: The top row contains histograms of the lengths in basepairs of all duplications and deletions detected in HapMap individuals in [12]. 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.

We used the ALS dataset to compile a training set informative for emission parameters. A simple script was run through the data to identify putative CNV on the basis of consecutive homozygous genotypes and logR values outside a defined window. The results from 70 were visually scored by two raters to confirm putative copy number variations on the bases of LogR and BAF values. 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, had either been validated with a PCR study, or independently documented with two different high throughput assays. This leads 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). In order to eliminate smooth variation in LogR values (see later portion of this section for more in-depth discussion), we pre-processed the samples, estimating a smooth value for LogR for each sample (we used the Lowess function implemented in R with default values) and subtracting it prior to the analysis with HMM.

The model we described so far allows only for three possible copy number states. Our training data, however, contains a small number of hemizygous deletions. To accommodate for this possibility, we have extended the HMM to include a zero copy number state, with marginal probability 0.00005.

In conducting data analysis, we fixed the transition parameters at the values described above, and estimated the emission parameters with the EM algorithm. However, we felt that the amount of data in our training set was not sufficient to estimate the emission parameters relative to the homozygous deletion states. Instead, we fixed them at the following levels, that appear reasonable

True	Reconstructed			
	0	1	2	3
0	33	0	0	0
1	2	1466	36	0
2	3	189	514298	289
3	0	1	20	1816

Table 1: Comparison of true and reconstructed state values on the training set.

in light of prior evidence and the data in the training set: the mean value for Log R is -5.6 and its standard deviation is 1.133.

Once the parameters of the HMM are estimated, 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 analyzed the same sample with a version of the HMM that did not account for linkage disequilibrium. This leads to a decrease in specificity. Figure 8 illustrates one of the regions in our training set. The deletion here considered spans of 5 SNPs and has been experimentally verified. While, the logR values in correspondence of that SNP are lower then average, identifying the CNV on the basis of this signal alone is not easy. For illustrative purposes we used the Fused Lasso routine by Tibshirani and Wang [15]. This routine correctly identifies the CNV, but in the same region also identifies another putative CNV that is spurious. We will comment further on the behavior of the Fused Lasso on this dataset.

We then proceeded to analyze the entire genome sequence of one individual, keeping the parameters of the HMM fixed at the estimated values. We required that the local posterior probability of copy number other then 2 to be greater then .99 in order to identify a copy number variant. According to this criterion, 20 regions appeared to harbor putative copy number variations, with locations illustrated in Figure 9. Figures 10 and 11 give details on the nature of two of these iden-

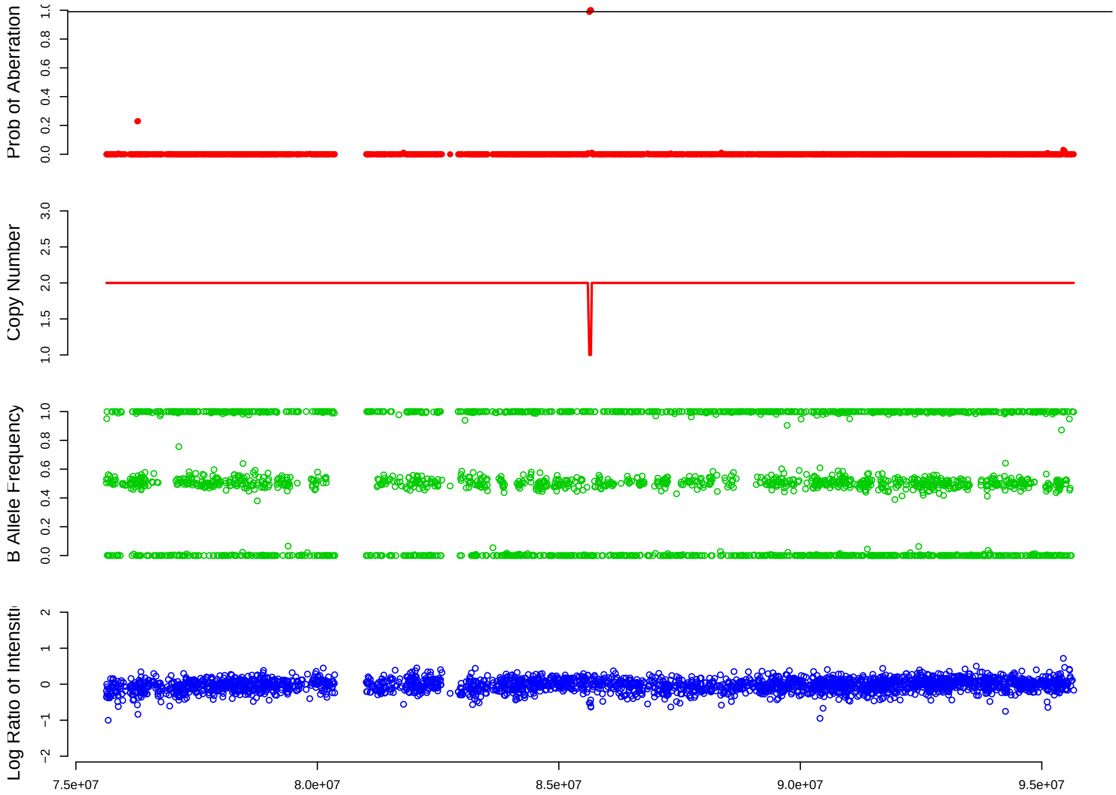


Figure 8: A deletion involving 5 SNPs on chromosome 15. Data for additional 1000 SNPs flanking the deletion is also presented. On the x-axis, we report the positions of queried SNPs in base pairs. The plots, from top to bottom, display the posterior probability of a copy number aberration, the copy number state reconstructed with Viterbi, the values of x (B allele frequency) and y (LogR).

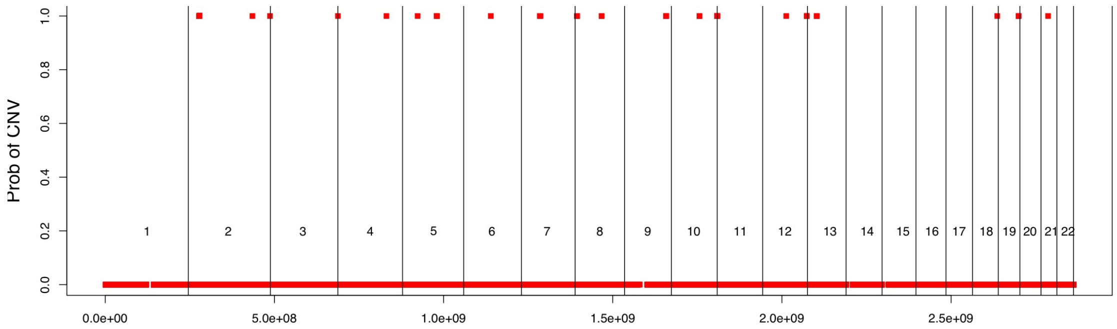


Figure 9: Copy number variants detected in one individual. The posterior probability of a copy number aberration is displayed against the genomic position of analyzed SNPs. Junctures between different chromosomes are identified with vertical lines.

tified variations that appear to be reliable calls. In the case of the CNV on chromosome 19, the consideration of the B allele frequency, again helps to correctly identify the location.

We then considered the analysis of one genomic location that has been documented to contain copy number polymorphisms in the entire sample of 922 individuals. This region is an hemizygous deletion of approximately 160kb on Chromosome 8, deposited by 6 independent sources in the Database of Genomic Variation. Analyzing our dataset with the HMM, we identified 54 individuals carrying the deletion. The averge Viterbi path across all 922 individuals in this region is presented in figure 12. The polymorphic deletion, which spans 17 SNPs is clearly identified. Subsequently, we analyzed 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 wanted 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 turned out to be equivalent, suggesting that we had already captured all deletions with out original scan.

Conducting the multisample analysis, however, we realized one strength of the single sample analysis: its robustness to smooth fluctuations of the logR signal. Consider the situation presented in Figure 13. The logR signal, in the top panel, clearly fluctuates around a trend, captured here

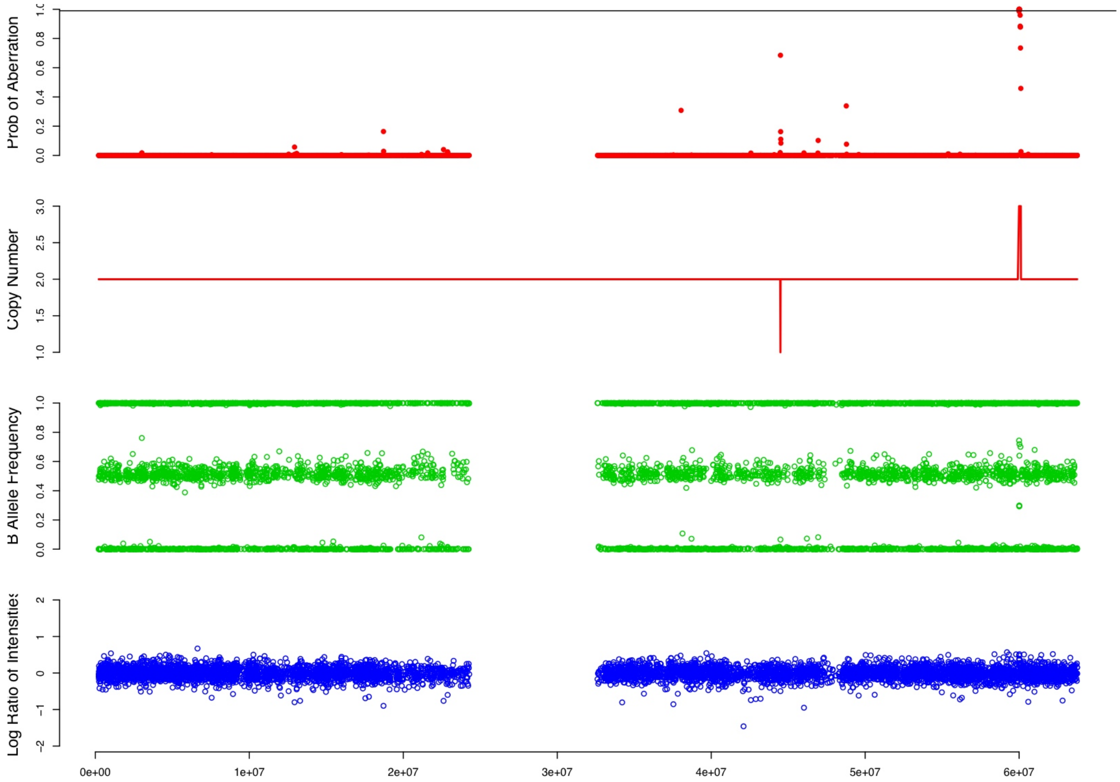


Figure 10: Data relative to SNPs on chromosome 19 for the studied individual. On the x-axis, we report the positions of queried SNPs in base pairs. The plots, from top to bottom, display the posterior probability of a copy number aberration, the copy number state reconstructed with Viterbi, the values of x (B allele frequency) and y (LogR).

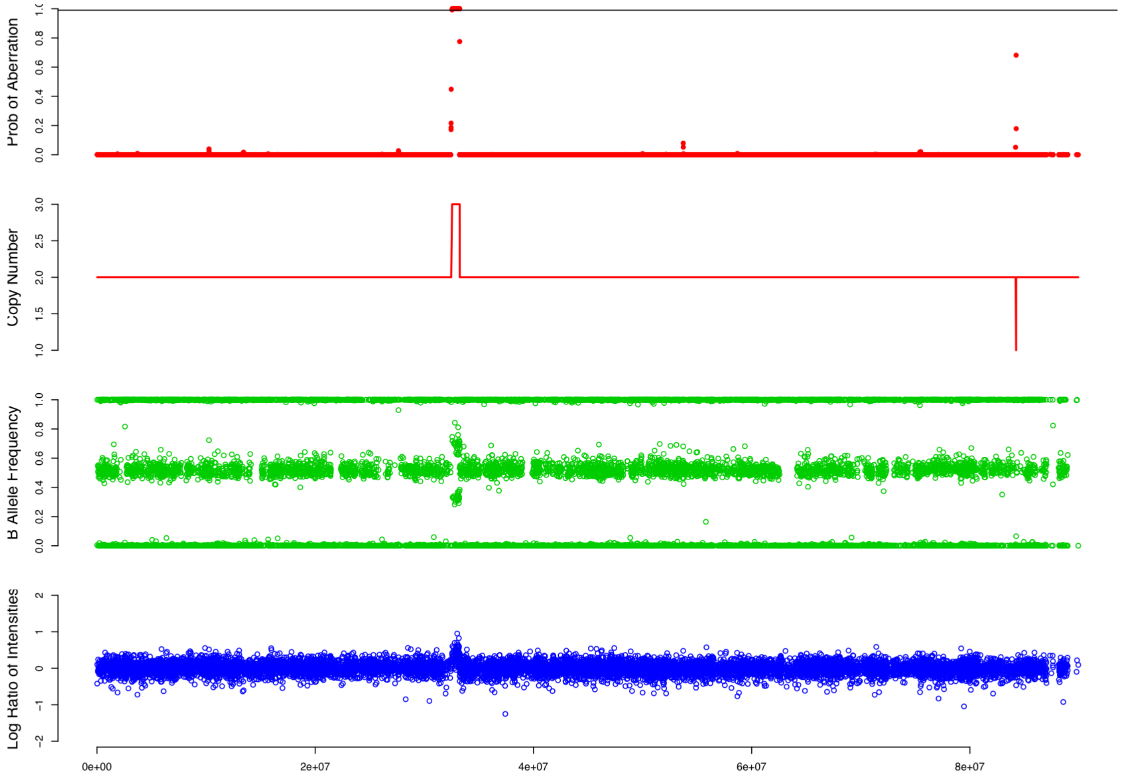


Figure 11: Data relative to the small arm of chromosome 2 for the studied individual. On the x-axis, we report the positions of queried SNPs in base pairs. The plots, from top to bottom, display the posterior probability of a copy number aberration, the copy number state reconstructed with Viterbi, the values of x (B allele frequency) and y (LogR).

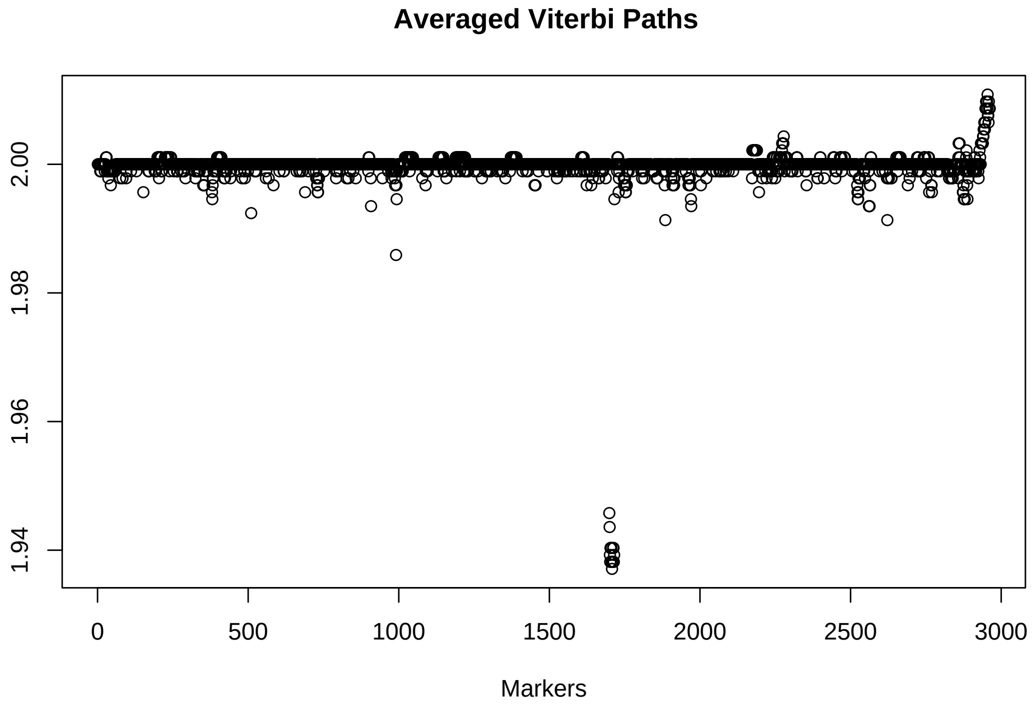


Figure 12: Average of Viterbi paths across 922 individuals in 3000 SNPs surrounding a common deletion, comprising 16-17 SNPs.

with a spline regression. The analysis of this data with Fused lasso leads to results illustrated in the middle panel, while the bottom panel reports the estimated viterbi path. The smooth trend is reflected in the fused lasso results, but it is eliminated from the viterbi path.

To date, we do not have a clear understanding of the experimental effects behind this trend. In our dataset, we were able to identify samples that appear not to have a systematic trend difference from zero, and two groups of samples that shared two different smooth patterns. Multisample analysis is particularly sensitive to these. In order to avoid spurious results, we first pre-processed the row 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 were not substantially different from the ones obtained on the original row data, but the multisample results were now much less noisy. Figure 14 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. In this case there appears to be no differences between cases and controls; while the deletion appears to be longer in some of the controls, it is hard to assess the relevance of this finding.

6 Discussion

We introduced a Hidden Markov Model algorithm to reconstruct variation in copy numbers from Illumina genotype data. Our algorithm differs from others recently appeared in the literature in the following aspects. (1) Like other HMM models proposed for illumina genotype, but differently from adaptations for this purpose of segmentation algorithms, we use the information contained in the B allele frequency signal. (2) We assume that deletion and duplications are “rare” events, and

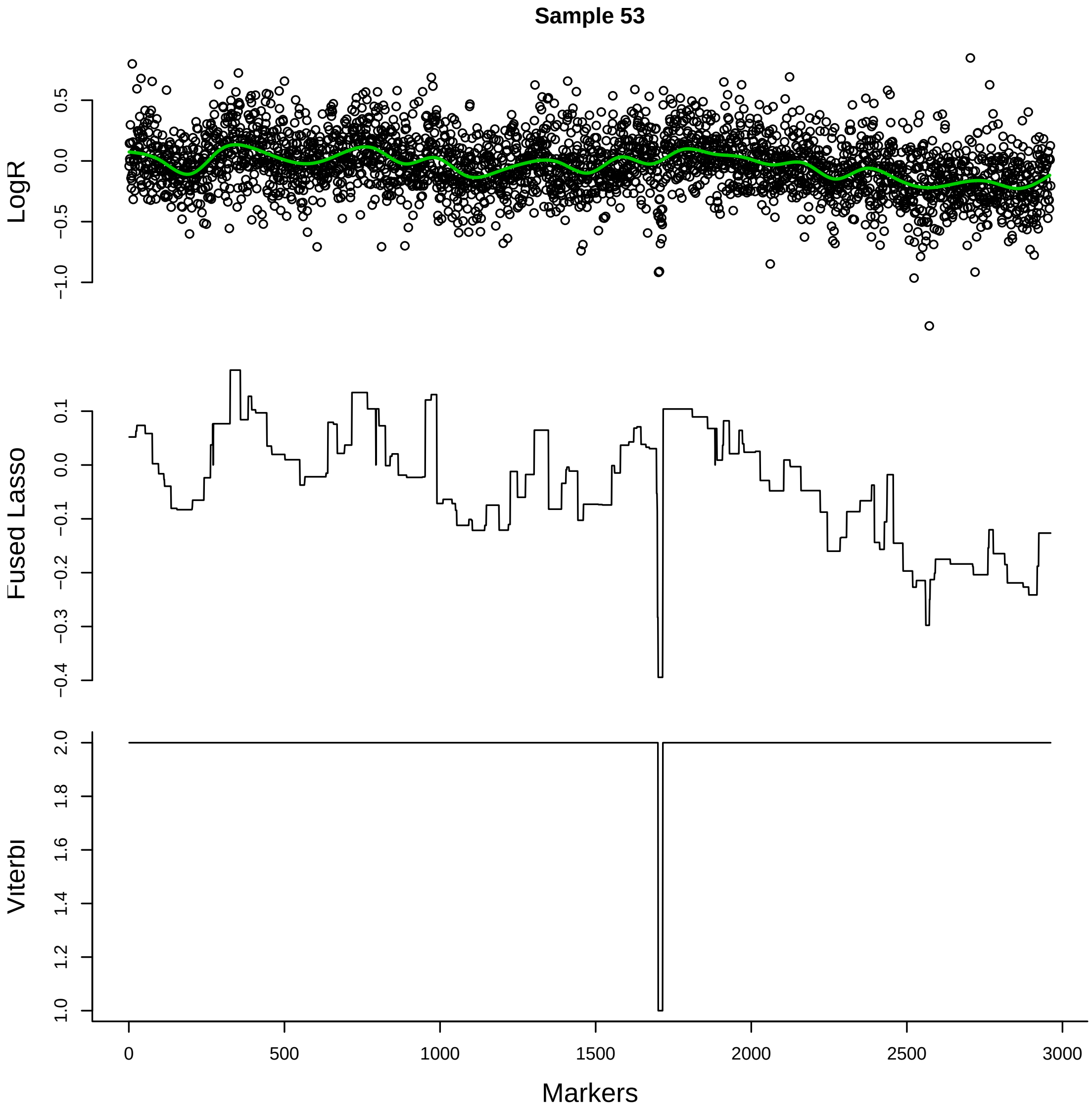


Figure 13: Smooth trends in LogR signal. The top display, provides LogR values for the 3000 analyzed SNPs in one individual. The middle plot presents the result of the cghFlasso algorithm applied on the LogR values. The bottom display reports the reconstructed copy number state by our HMM algorithm that uses both LogR and BAF values.

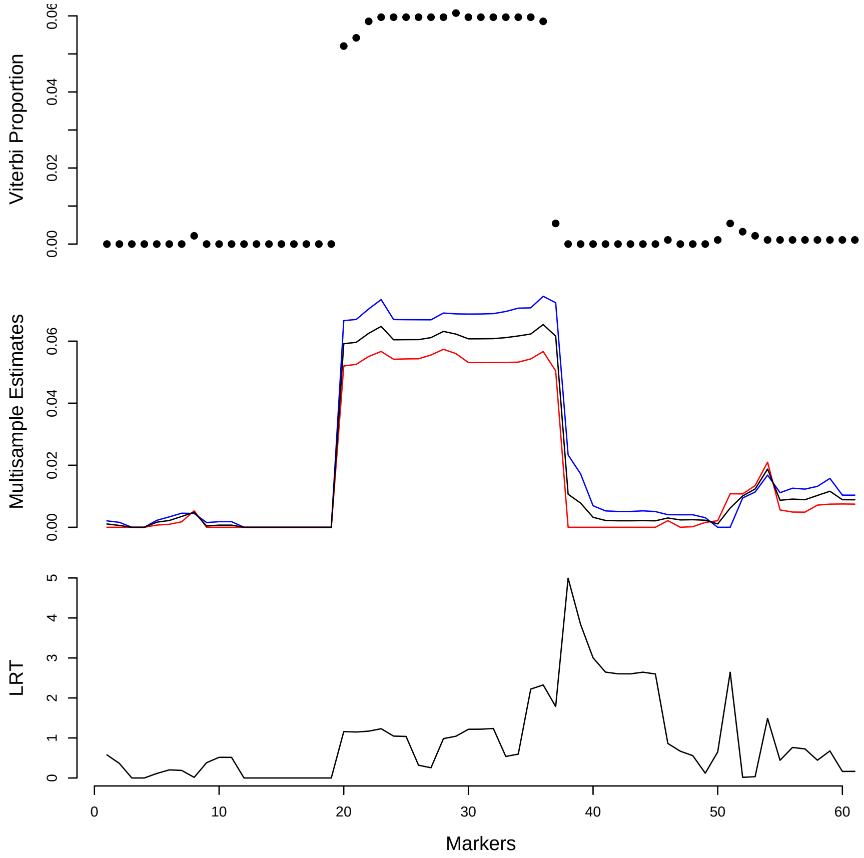


Figure 14: Multisample analysis of 60 SNPs in the proximity of the studied deletion. The top display contains the proportion of individuals whose Viterbi path was equal to 1 among the 922 analyzed. The middle plot presents the location-specific probability of deletion estimated using the algorithm described 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.

model this with a genomewide rate of deletion δ and duplication γ that are much lower then the probability of 2 DNA copies. (3) We explicitly take into account linkage disequilibrium between SNPs. (4) We describe how our model can be adapted to carry out multi-sample analysis.

Aknowledgements

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.

References

[1] Chen Y, Lin C, Sabatti C (2006) Volume measures for linkage disequilibrium, *BMC Genetics* 7:54

[2] Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* 35: 2013–25.

[3] Conrad DF, Andrews TD, Carter NP, Hurles ME and Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome, *Nature Genetics* 38:75–81.

[4] Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, Veltman JA. (2007) 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*, in print.

[5] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. (2004) Detection of large-scale variation in the human genome. *Nat Genet.* 36:949–51.

[6] Lange K (2004) *Applied Probability*, Springer, New York.

[7] Lange K (2004) *Optimization*, Springer, New York.

[8] Newton, M., Gould M., Reznikoff, C., Haag, J. (1998) “On the statistical analysis of allelic-loss data,” *Stat. Med.* 17: 1425–45.

[9] Newton, M., Lee, Y. (2000) “Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data,” *Biometrics* 56: 1088–97.

[10] M.A. Newton (2002) "Discovering combinations of genomic alterations associated with cancer," *Journal of the American Statististical Association* 97: 931–942.

[11] 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. (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci* 103:8006–11.

[12] 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, Hurles ME. (2006) Global variation in copy number in the human genome. *Nature* 444:444–54.

[13] Sabatti C, Lange K (2007) Bayesian Gaussian mixture models for high density genotyping arrays. *JASA* (in press)

[14] 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. (2004) Large-scale copy number polymorphism in the human genome. *Science* 305:525–8.

[15] Tibshirani and Wang (2008) Spatial smoothing and hot spot detection for CGH data using the Fused Lasso, *Biostatistics* 9:18–29.

[16] Van Es M et al. (2007) Genome-wide association study of sporadic amyotrophic lateral sclerosis identifies ITPR2 as a susceptibility gene, *submitted for publication*.

[17] Wang H, Lee Y, Nelson S, and Sabatti C (2005) Inferring genomic loss and location of tumor suppressor genes from high density genotypes, *Journal of the French Statistical Society* 146:153–171.

[18] Wang H, Lin C, Service S, The international collaborative group on isolated populations, Chen Y, Freimer N, Sabatti C (2006) Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density, *Human Heredity* 62:175–189.

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