Cancer Informatics

ORIGINAL RESEARCH

Model-Integrated Estimation of Normal Tissue Contamination for Cancer SNP Allelic Copy Number Data

Susann Stjernqvist, Tobias Rydén and Chris D. Greenman

Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden, [1]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK. [2]The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK. Corresponding author email: tryd@math.kth.se

Abstract: SNP allelic copy number data provides intensity measurements for the two different alleles separately. We present a method that estimates the number of copies of each allele at each SNP position, using a continuous-index hidden Markov model. The method is especially suited for cancer data, since it includes the fraction of normal tissue contamination, often present when studying data from cancer tumors, into the model. The continuous-index structure takes into account the distances between the SNPs, and is thereby appropriate also when SNPs are unequally spaced. In a simulation study we show that the method performs favorably compared to previous methods even with as much as 70% normal contamination. We also provide results from applications to clinical data produced using the Affymetrix genome-wide SNP 6.0 platform.

Keywords: allelic copy number, hidden Markov model, cancer, normal cell contamination

Cancer Informatics 2011:10 159–173

doi: 10.4137/CIN.S6873

This article is available from http://www.la-press.com.

1. Introduction

DNA in tumor cells can contain abnormalities in the form of copy number aberrations such as segments with losses or gains of one or several copies of either allele. The lengths of such aberrations can vary between short segments up to an entire chromosome, and their positions are essential both for detecting and for improving knowledge of various sorts of cancer. Therefore, methods that localize copy number aberrations are of great importance. In addition to changes in the total copy number of both alleles together, changes in the *allelic copy numbers*, ie, the number of copies of each allele, are also important. We denote the two different alleles at a given genomic location by A and B, so that for normal cells the possible genotypes are AA, AB and BB. One example of a genotype aberration is loss of heterozygosity (LOH), for which the only attainable genotypes are AA and BB.

Different techniques to measure DNA copy numbers have been developed, as have methods to evaluate the measurement data. One technique is array comparative genomic hybridization (aCGH), which provides ratios of the copy numbers of a sample DNA, compared to those of some reference DNA. Several different statistical methods have been applied to this kind of data, including different segmentation methods,[4,20,21] smoothing[6,12] and hidden Markov models.[1,7,9,16,19,24,27,28,29] aCGH data provides information only about the total copy number and gives no information about the amount of each allele. Another drawback with such data is the limited number of probes on the arrays. For this reason there is an increased use of single nucleotide polymorphism (SNP) data, which offers denser measurements and provides intensities for the two alleles separately. Using SNP data it is possible not only to estimate copy number changes, but also to find allelic changes such as LOH. Indeed, a copy number amplification may be caused by different allelic changes. For example, a copy number of four could correspond either to {AAAA, AAAB, ABBB, BBBB}, to {AAAA, AABB, BBBB} or to {AAAA, BBBB}, depending on which allele that has gained extra copies.

SNP data has previously been analyzed using various sorts of methods, such as smoothing[11,15] and pattern recognition.[22] The most frequently used methods are however based on hidden Markov models (HMMs).[3,8,13,17,18,26,30,31] A brief introduction to Markov chains and HMMs is found in Appendix 1. HMMs suit SNP data well since genomic alterations often appear in longer or shorter segments, implying that copy numbers across probes in a small genomic region are correlated. For example, Wang et al[31] and Colella et al[3] model SNP data from the Illumina array, which provides log R ratio data ($\log_2$-ratio of total observed intensities to total expected intensities) and BAF data (normalized measure of the relative intensities of the two alleles), using an HMM with six states, while Sun et al[30] apply a more comprehensive model with nine states. Korn et al[13] combine an HMM to model copy number variants with a clustering algorithm to detect genotypes. Li et al[18] also model the proportion of the major allele while Lamy et al[17] use both allelic intensities provided by the Affymetrix array and model them using bivariate Normal distributions.

Several of the methods above assume that the ploidy, ie, mean copy number, of a chromosome is two. This holds for normal cells, but cancer cells are anueploid, ie, their ploidy may differ from two. The necessity for considering ploidy when modeling cancer data is well described by Greenman et al,[8] but in brief one can say that the measured normalized intensity for a probe in a diploid chromosome is twice as large as for a probe with the same copy number in a quadroploid chromosome. Two methods that include ploidy are those of Attiyeh et al[2] and Greenman et al,[8] which both contain a pre-processing step in which the ploidy is estimated. Greenman et al then continue by using an HMM while Attiyeh et al apply a window-based model.

Another feature common in tumor samples, arising from the difficulty to dissect tumor cells only from a tissue sample, is contamination of the tumor cell sample by normal cells. As a result the measured allelic intensities are mixtures of intensities from tumor and normal cells, thus yielding non-integer DNA copy numbers. One way to incorporate such contamination is to model total copy numbers of the mixed sample in a non-parametric way,[2,29] but this provides limited information about the copy numbers of the cancer cells. Sun et al[30] estimate the fraction of normal tissue contamination using an empirical method and Colella et al[3] write that it is possible to extend their method to handle contamination, but without being more specific. Li et al[18] show that their method can handle a fraction of normal tissue contamination up to 30%, while Lamy et al[17] report a simulation study with slightly better results. Some tumors however form in a manner such that even with microdisection, a significant proportion of normal cells (say 50% or more) can arise in the sample, and none of the above methods provide results that are satisfactory for such high fractions of normal tissue contamination.

The purpose of the present paper is to devise a method to estimate allelic copy numbers, with ploidy and fraction of normal tissue contamination integrated in the model. Indeed, in all of the above papers, ploidy and/or normal fraction are estimated by adding more or less ad-hoc steps to a model that does not account for these parameters in itself. The model reported here is thus particularly suited for cancer data, for which both of these features are common. By including these parameters in the model they can be estimated alongside the other parameters using all data, rather than adding a pre-processing step or empirical methods using only a small subset of the data. In the simulation study presented below, samples with 30%, 50% and 70% normal contamination are simulated and even for the largest amount of contamination, 97% of the probes are reconstructed to the correct copy number state.

An additional feature of our model is that it is based on a continuous-index Markov chain, which accounts for the fact that the SNP probes are often unevenly spread over the genome. The relevance of a continuous-index model was highlighted by Gupta and Mitra[10] (Section 5.3) for the different but related problem of classifying regions of DNA as nucleosome free regions (NFR) or non-NFR using a two-state HMM. Indeed, they showed that with irregularly spaced probes, a continuous-index model can provide substantially better results than a discrete-index model; 99% vs. 85% or 68% correct classifications in simulations for two different arrangements of the probes. Also the methods by Wang et al,[31] Colella et al[3] and Li et al,[18] who apply discrete-index HMMs to SNP data, aim to take distances between probes into account by letting the Markov transition probabilities depend on these distances in different ways. Common to all of these methods is however that the stipulated transition probabilities violate the Chapman-Kolmogorov equation of Markov chains. That is, letting $P(t)$ be the matrix of transition probabilities over a distance t between two probes, the equality $P(t_1 + t_2) = P(t_1)P(t_2)$ does not hold. In essence this means that there is in fact no Markov chain with the given transition probabilities.

The paper is organized as follows. The model is described in in Section 3. Section 4 provides results from a simulation study as well as from an application to clinical data. Concluding remarks are given in Section 5.

2. Data

The data used in this study are the cancer samples in Greenman et al,[8] produced using the Affymetrix genome-wide SNP 6.0 platform. We applied the algorithm to about 15 different cell line and primary tumor samples, representing various cancer forms including breast, lung and renal cancer. The primary tumor sample PD1753a for which results are reported in Section 4.2 are from a clear cell renal cell carcinoma sample.[32]

For probes at SNPs the intensities of the two different alleles are provided, while at other positions only a single total copy number intensity is available. Following Greenman et al,[8] the intensities are normalized by first dividing each measurement by the total intensity of the sample (ie, the sum of all probe intensities over the entire genome), to remove chip-to-chip variation. The mean signals for each allele (or probe at non-SNP positions) are then transformed into a copy number intensity and a genotype intensity that are indicators of total copy number and allelic ratio dosages. The model presented below incorporates intensities for SNP probes only, but is easily extendable to include also probes measuring total copy number only; we elaborate on this further in the Discussion. The cancer data is available from the Cancer Genome Project, subject to a manual transfer agreement, and our Matlab code is available on the WWW.[33]

3. The Model

3.1. Basic structure

Let there be N_c probes on chromosome c, and denote these probes as probe (k, c), $k = 1, 2, \ldots, N_c$. The genomic location of probe (k, c) is denoted by t_{kc}, measured in the unit base pairs (bp) starting from the beginning of the chromosome. We denote the two different alleles at any genomic location by A and B. We will write $g = (g_A, g_B)$ for the allelic copy numbers, ie, g_A and g_B are the number of copies of the A and B allele respectively. For example, the genotype AAB corresponds to $g = (2, 1)$. Obviously the genotype and the allelic copy numbers are in a one-to-one correspondence to each other, and at times we will make no real distinction between the two. The allelic intensities are modeled using an HMM for which each state i corresponds to one *genotype set* G_i as specified in Table 1. The Markov chain can be extended to include more states with copy numbers above six, but the model as stated here has proved to be enough for the studied samples. To explain the genotype sets in Table 1, we note that through cancer development any region in the genome starts with one parental copy of each region and ends up with m copies of one allele and n copies of the other. If the genotype was originally AA or BB then the genotype will be $(m + n)$A or $(m + n)$B, respectively. If the SNP was heterozygous then we must end up with either mA and nB, or mB and nA. These are the genotypes indicated in Table 1. We refer to state 4, with genotype set {AA, AB, BB} as the *normal state*, and by an *abnormal state* we mean any other state.

For each chromosome c the sequence of copy number states, according to Table 1, is modeled by a continuous-index Markov chain $(X_c(t))_{t_{1c} \le t \le T_c}$, where t and T_c are respectively the genomic location (in bp) within the chromosome and the length (in bp) of the chromosome. The Markov chains for different chromosomes are assumed independent. The genomic location (in bp) is, strictly speaking, a discrete variable, but since the number of bp's within a chromosome is much larger than the number of jumps of the Markov chain, the error caused by using a continuous approximation is negligible. With a discrete-index model the Markov transition probabilities would either be very close to unity (for staying in the same state from one bp to another) or close to zero (for changing state), and dealing with such probabilities is unstable numerically. For a continuous-index model, using transition rates rather than probabilities, this problem does not exist.

Table 1. Genotype sets for the different states of the Markov chain, sorted in the order given by the total copy number and copy number of the minor allele.

State i	(Total CN, minor CN)	Genotype set G_i
1	(0,0)	{ }
2	(1,0)	{A, B}
3	(2,0)	{AA, BB}
4	(2,1)	{AA, AB, BB}
5	(3,0)	{AAA, BBB}
6	(3,1)	{AAA, AAB, ABB, BBB}
7	(4,0)	{4A, 4B}
8	(4,1)	{4A, 3AB, A3B, 4B}
9	(4,2)	{4A, 2A2B, 4B}
10	(5,0)	{5A, 5B}
11	(5,1)	{5A, 4AB, A4B, 5B}
12	(5,2)	{5A, 3A2B, 2A3B, 5B}
13	(6,0)	{6A, 6B}
14	(6,1)	{6A, 5AB, A5B, 6B}
15	(6,2)	{6A, 4A2B, 2A4B, 6B}
16	(6,3)	{6A, 3A3B, 6B}

With 16 different states there are 240 different types of jumps and equally many transition rates (per chromosome). It is infeasible to estimate such many rates, and to make the model more parsimonious we assume a large number of them to agree. Specifically we assume, for chromosome c, a common rate λ_c for jumps from any state (normal or abnormal) to the group of abnormal states, with each such state, except for the current one in case the chain resides in an abnormal state, being equally likely, and another common rate η_c for jumps to the normal state from any abnormal state. The total rate out of any abnormal state, for chromosome c, is thus $\lambda_c + \eta_c$. This dynamic provides Markov chains whose stationary versions are time-reversible.[29] Finally we let $\delta_{ic} = P(X_c(t_{1c}) = i)$ denote the initial probability for Markov state i in chromosome c.

Write $y_{kc} = (y_{\mathrm{A}kc}, y_{\mathrm{B}kc})$ for the measured allelic intensities at probe (k, c). Greenman et al[8] studied the correlation between the allele A and B intensities, for each probe, using 460 wild-type samples. For probe (k, c), plotting the two allele intensities for all wild-type samples against each other reveals three clusters (see,[8] Figure 1, for an example). These clusters correspond to the genotypes AA, AB and BB, with the coordinates of the cluster centers written as $(A_{0kc} + 2A_{1kc}, B_{0kc})$, $(A_{0kc} + A_{1kc}, B_{0kc} + B_{1kc})$ and $(A_{0kc}, B_{0kc} + 2B_{1kc})$ respectively for suitable parameters A_{0kc}, B_{0kc}, A_{1kc} and B_{1kc}. These parameters were all estimated by Greenman et al[8] using the wild-type samples. Their interpretation is that A_{0kc} is the background intensity of the A allele (at diploid probes BB), and A_{1kc} is the increase in A allele intensity from BB to AB and from AB to AA; B_{0kc} and B_{1kc} have analogous interpretations.

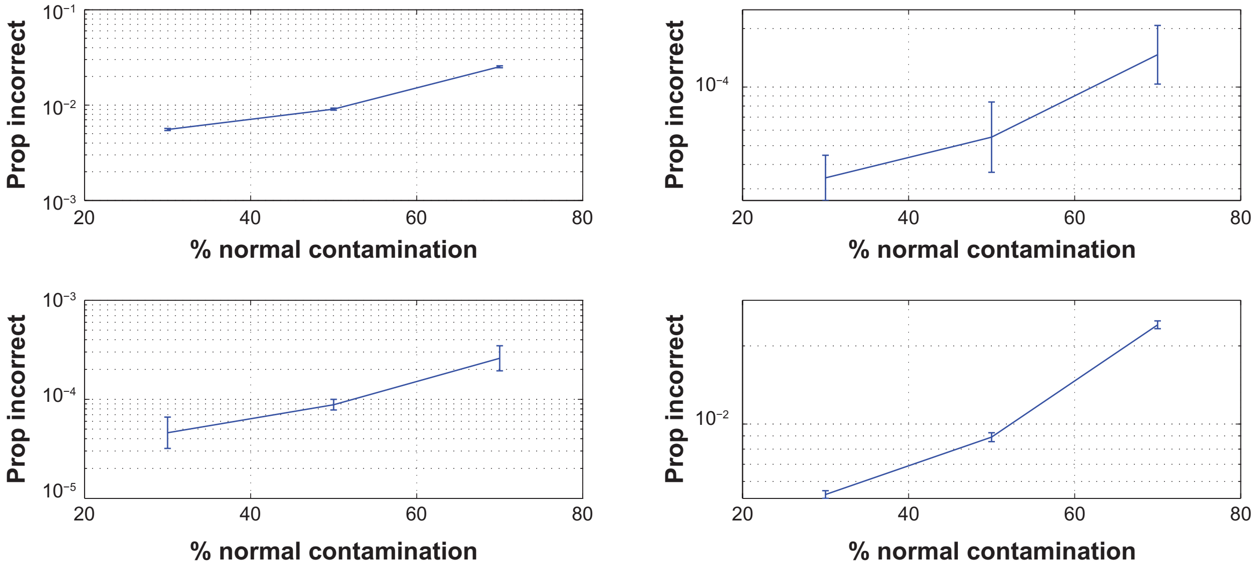

Figure 1. Proportions of probes at which the Markov state was incorrectly reconstructed by the Viterbi algorithm with MAP parameter estimates computed by the EM algorithm. Markov transition rates were $\lambda_c = \eta_c = 10^{-7}$ (top left), $\lambda_c = 10^{-7}$, $\eta_c = 10^{-9}$ (top right), $\lambda_c = 10^{-9}$, $\eta_c = 10^{-7}$ (bottom left), $\lambda_c = \eta_c = 10^{-9}$ (bottom right) (unit: bp^{-1}). Confidence intervals were obtained by exponentiating two-sided 95% student-*t* confidence limits based on the log-proportions for 10 genome replicates.

Further denote by $(\mu_{\mathrm{A}kcg}, \mu_{\mathrm{B}kcg})$ the mean allele A and B intensities at probe (k, c) for allelic copy numbers $g = (g_{\mathrm{A}}, g_{\mathrm{B}})$. The cluster centers above then write

$$\begin{aligned}\mu_{kcg} &= (\mu_{\mathrm{A}kcg}, \mu_{\mathrm{B}kcg}) \\ &= (A_{0kc} + g_{\mathrm{A}}\, A_{1kc}, B_{0kc} + g_{\mathrm{B}} B_{1kc}),\end{aligned} \quad (1)$$

and this model applies for the normal Markov state $i = 4$, ie, for allelic copy numbers such that $g_{\mathrm{A}} + g_{\mathrm{B}} = 2$. Moreover, the clusters in Greenman et al[8] (Fig. 1) are tilted ovals, indicating that the intensities for alleles A and B are correlated and have unequal variances. Greenman et al[8] found that a suitable model for the covariance matrix is

$$\sum_{kcg} = v_{kc} \begin{pmatrix} \mu^2_{\mathrm{A}kcg} & \rho_{kc}\mu_{\mathrm{A}kcg}\mu_{\mathrm{B}kcg} \\ \rho_{kc}\mu_{\mathrm{A}kcg}\mu_{\mathrm{B}kcg} & \mu^2_{\mathrm{B}kcg} \end{pmatrix}; \quad (2)$$

note that the variances are taken proportional to the squared means. The probe-specific variance factors v_{kc} and correlations ρ_{kc}, as well as the means parameters A_{0kc}, B_{0kc}, A_{1kc} and B_{1kc} described above, were all estimated by Greenman et al[8] using the wild-type samples and assuming a bivariate Normal distribution for each cluster.

We now carry this model further by assuming that for each probe, the allele intensities follow the mean-variance model given by Eqs. (1)–(2) also for genotypes $(g_{\mathrm{A}}, g_{\mathrm{B}})$ for which g_{A} and g_{B} do not sum to two, ie, for all pairs $(g_{\mathrm{A}}, g_{\mathrm{B}})$ corresponding to genotypes listed in Table 1. That is, we assume that the response from amount of each allele on the microarray to measured intensity is linear, with the variance also increasing linearly. In reality the allelic intensities have a linear response for lower copy numbers, while at higher copy numbers the intensities start to saturate and our method is approximate. This could be adjusted for by a non-linear transformation, cf. Section 5, but we have not attempted such an adjustment in the analyses presented in this paper.

The above model specifies the conditional density of Y_{kc} given a particular genotype. To specify the conditional density of Y_{kc} given a Markov state, we recall that each Markov state has a genotype set comprising between one and four different genotypes. Thus the conditional density of Y_{kc}, given the state, is a mixture of bivariate Normal distributions for which each mixture component represents a different genotype. The mixture weights were taken as the Hardy-Weinberg weights; for the copy number-aberrated genotypes, Hardy-Weinberg was used to compute the germline genotype proportions. Thus letting p_{kc} be the allele frequency for an A allele at probe (k, c), the probability for the different genotypes, denoted by w_{kcig}, are the binomial probabilities p_{kc} and $1 - p_{kc}$ for states with two genotypes, p_{kc}^2, $2p_{kc}(1 - p_{kc})$ and $(1 - p_{kc})^2$ for states with three genotypes, and p_{kc}^2, $p_{kc}(1 - p_{kc})$, $p_{kc}(1 - p_{kc})$ and $(1 - p_{kc})^2$ respectively for states with four different genotypes. The frequencies p_{kc} were also estimated by Greenman et al,[8] using the wild-type samples. The conditional density for a measurement Y_{kc} given the Markov state, often referred to as the *emission density* of the HMM, thus writes

$$f_{Y_{kc}|X_c(t_{kc})}(y \mid i) = \sum_{g \in G_i} w_{kcig} f_{Y_{kc}|G_{kc}}(y \mid g), \qquad (3)$$

where G_{kc} is the allelic copy numbers for probe (k, c) and $f_{Y_{kc}|G_{kc}}(\cdot|g)$ is the bivariate Normal density with mean and covariance matrix as in Eqs. (1)–(2).

As pointed out in the introduction we include the ploidy K, ie, average copy number over the entire genome, in the model to make it suitable for cancer data. The ploidy is defined genome-wide and not per chromosome, as the probe intensities are normalized per genome. The HMM described above models the normalized intensities, and its parameters were estimated for wild-type samples (ie, diploid samples; $K = 2$). For a sample with $K > 2$ the normalized intensities will thus be smaller by a factor $2/K$ (on average), so that the model for the normalized intensities becomes

$$Y_{kc}|G_{kc} = g \sim N\left(\frac{2}{K}\mu_{kcg}, \frac{4}{K^2}\sum\nolimits_{kcg}\right). \qquad (4)$$

This completes the specification of the basic model. As described above, the parameters A_{0kc}, A_{1kc}, B_{0kc}, B_{1kc}, v_{kc}, ρ_{kc} and p_{kc} were all estimated from the wild type samples, and were thus considered as fixed when the model was applied to cancer cell data. The intensities λ_c and η_c, the initial probabilities δ_c and the ploidy K were on the other hand estimated from the actual cancer data.

3.2. Normal tissue contamination

As mentioned above it is often difficult to dissect cancer cells without including any surrounding normal tissue, ie, diploid tissue. Such contamination implies that the measured allelic intensities correspond to a mixture of cancer and normal cells. We denote the fraction of normal tissue in the sample by γ, and consequently the fraction of tumor tissue is $1 - \gamma$. Then for a given probe with, as above, copy numbers g_A and g_B or alleles A and B in the tumor but also copy numbers g_A^N and g_B^N for the two alleles in the normal tissue, we assumed the same mean-covariance model as in Eqs. (1)–(2), but with (g_A, g_B) replaced by

$$(g_A^\gamma, g_B^\gamma) = ((1-\gamma)g_A + \gamma g_A^N, (1-\gamma)g_B + \gamma g_B^N). \quad (5)$$

Similarly, the conditional distribution of Y_{kc} given Markov state i is a mixture of bivariate Normals, but now each four-tuple (g_A, g_B, g_A^N, g_B^N) contributes to a component of that mixture. Thus, the number of mixture components will for some Markov states be larger than without normal tissue contamination (see Table 2).

The weights for the combined genotypes are Hardy-Weinberg weights as in the model without normal contamination. For example, for a state in Table 2 with three combined genotypes, the weights are p_{kc}^2, $2p_{kc}(1 - p_{kc})$ and $(1 - p_{kc})^2$ respectively.

3.3. Estimation of parameters and the Markov path

The parameters estimated from a tumor sample are the transition rates λ_c and η_c, the initial probabilities δ_c, the ploidy K and also the fraction γ of normal tissue contamination.

Table 2. Combined genotype sets for the different states of the Markov chain, in a model with normal contamination γ. The weights for the respective combined genotypes are the Hardy-Weinberg weights as in the model without normal tissue contamination, and the total and minor copy numbers for the abberated components are as in Table 1.

State *i*	**Combined genotype set** G_i
1	{2γA, γAγB, 2γB}
2	{(1 + γ)A, AγB, γAB, (1 + γ)B}
3	{2A, (2 – γ)AγB, γA(2 – γ)B, 2B}
4	{AA, AB, BB}
5	{(3 – γ)A, (3 – 2γ)AγB, γA(3 – 2γ)B, (3 – γ)B}
6	{(3 – γ)A, (2 – γ)AB, A(2 – γ)B, (3 – γ)B}
7	{(4 – 2γ)A, (4 – 3γ)AγB, γA(4 – 3γ)B, (4 – 2γ)B}
8	{(4 – 2γ)A, (3 – 2γ)AB, A(2 – γ)B, (4 – 2γ)B}
9	{(4 – 2γ)A, (2 – γ)A(2 – γ)B, (4 – 2γ)B}
10	{(5 – 3γ)A, (5 – 4γ)AγB, γA(5 – 4γ)B, (5 – 3γ)B}
11	{(5 – 3γ)A, (4 – 3γ)AB, A(4 – 3γ)B, (5 – 3γ)B}
12	{(5 – 3γ)A, (3 – 2γ)A(2 – γ)B, (2 – γ)A(3 – 2γ)B, (5 – 3γ)B}
13	{(6 – 4γ)A, (6 – 5γ)AγB, γA(6 – 5γ)B, (6 – 4γ)B}
14	{(6 – 4γ)A, (5 – 4γ)AB, A(5 – 4γ)B, (6 – 4γ)B}
15	{(6 – 4γ)A, (4 – 3γ)A(2 – γ)B, (2 – γ)A(4 – 3γ)B, (6 – 4γ)B}
16	{(6 – 4γ)A, (3 – 2γ)A(3 – 2γ)B, (6 – 4γ)B}

For a model like the present one, the maximum-likelihood estimator (MLE) typically overestimates the transition rates λ_c and η_c[25] (Section 4.3), thereby letting an aposteriori reconstruction of the Markov chain trajectory capture also very short transients of the observed data. When using the EM algorithm to compute the MLE, this becomes visible as an overestimated number of jumps of the Markov chain. In order to control the jumps and make their number biologically plausible, we take a Bayesian approach and penalize overly large transition rates by placing Gamma distribution priors on each λ_c and η_c. Other parameters are assigned uniform (flat) priors. All parameters are apriori independent. We then compute the maximum aposteriori (MAP) parameter estimate using the EM algorithm, by incorporating the priors into the M-step[5] (p. 6). Otherwise this algorithm is a variant of the EM algorithm described by Roberts and Ephraim,[23] designed to estimate parameters of a continuous-index HMM observed at discrete positions. The method is detailed in Appendix 2.1.

Finally, to construct an estimate of the trajectory of the hidden Markov chain we use a Viterbi algorithm adapted to continuous-index Markov chains (see Appendix 2.2).

4. Results

4.1 Application to simulated data

To evaluate our method's ability of making correct reconstructions for different amounts of normal contamination, we simulated data from the assumed model, computed MAP parameter estimates using the EM algorithm, reconstructed the hidden Markov chain using the Viterbi algorithm, and finally computed the proportion of probes at which the Markov state was correctly reconstructed. For each simulated dataset we first simulated the Markov chain and the genotypes for each probe position, then computed μ_{kcg} and Σ_{kcg} using Eqs. (1)–(2), Eq. (5) and the fixed A_0, A_1, B_0, B_1, ρ and v (estimated from the wild-type samples), and finally simulated data from the bivariate Normal distributions of Eq. (4) with $K = 2$. Note that the actual value of K is irrelevant for these simulations, since the model given by Eqs. (1)–(2) describes the data after normalization.

The simulations were carried out for 30%, 50% and 70% normal contamination, and transition rates $\lambda_c = \eta_c = 10^{-7}$, $\lambda_c = 10^{-7}$ and $\eta_c = 10^{-9}$, $\lambda_c = 10^{-9}$ and $\eta_c = 10^{-7}$, and $\lambda_c = \eta_c = 10^{-9}$ (in units of bp^{-1}) respectively. For each combination of contamination and rates, 10 replicates were simulated. For the Gamma priors of λ_c and η_c we chose shape parameter 2 and means equal to the true transition rates. These choices yield priors that are not overly informative, but which are concentrated enough on small values to prevent the Markov chain from jumping too frequently in our samples.

To verify the convergence of the EM algorithm we present the EM iterations for three different simulated replicates in Figure 2. The proportions of incorrectly reconstructed probes are plotted in Figure 1.

These results can be compared to those from the simulation study by Lamy et al.[17] For a normal contamination of 30% the results are similar, but for 45%, which is the largest fraction studied by Lamy et al, their method provides 8%–18% incorrectly estimated probes while at 50% contamination our model provides an error rate below 1%. In addition, the present model performs well even at such a high amount of normal contamination as 70%, when the Markov state is correctly reconstructed at more than 97% of the probes. Obviously the differences between our results and those of Lamy et al depend not only on the different estimation algorithms but also on differences between the number and location of the probes, and on the model for the observed allele intensities and its parameters. However, given the magnitude of the performance improvement, a significant part of it must be attributed to the estimation algorithm as such.

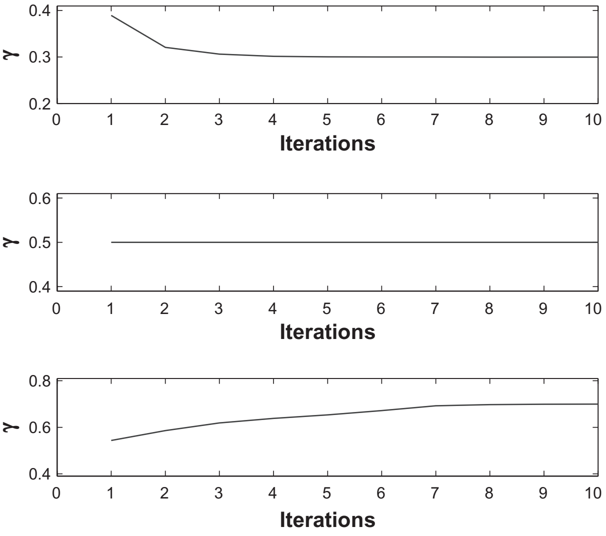

Figure 2. Estimates of normal contamination γ for iterations 1–10 of the EM algorithm and three simulated replicates with different values of γ: γ = 0.3 (top), γ = 0.5 (middle), and γ = 0.7 (bottom). The initial value for γ was 0.5 in all simulations.

4.2. Application to clinical data

We applied our method to a number of samples from the data described in Section 2. An example is displayed in Figure 3, which shows the Viterbi reconstruction of the Markov chain as well as the corresponding copy numbers compared to the data, for chromosome 3 in primary sample PD1753. For the Gamma priors for λ_c and η_c we chose shape parameters 2 and means 10^{-15}.

The reconstruction divides the chromosome into two regions, reconstructed to state 2 ({A, B}) and state 4 ({AA, AB, BB}) respectively. As a simple check of this reconstruction we plotted the standardized allele intensities against each other for all probes in the respective region (Figs. 4–5). Figure 5, corresponding to the normal state, shows three clusters representing the three genotypes AA, AB and BB, while Figure 4 shows four clusters. In Table 1 state 2 is associated to two genotypes, A and B, but with normal contamination this state comprises four combined genotypes (1 + γ)A, AγB, γAB and (1 + γ)B (Table 2). Here γ is estimated at 0.53.

For some of the genomes the values of A_{0kc}, A_{1kc}, B_{0kc} and B_{1kc} needed small adjustments before applying our model; without it, the model did not produce a reasonable fit. A possible explanation for this adjustment being required is a drift in the measured intensities from when data from the wild-type samples, used to estimate most model parameters, was collected, to when the tumor samples were analyzed. A suitable construction of the adjustment was as a common, ie, genome-wide multiplier c_0 for all A_{0kc} and B_{0kc}, and another common multiplier c_1 for all A_{1kc} and B_{1kc}. The multipliers c_0 and c_1 were estimated using data from a chromosome segment known to belong to the normal state. The data within this segment was clustered into three parts using the *k*-means algorithm, and then c_0 and c_1 were estimated by a least squares fit.

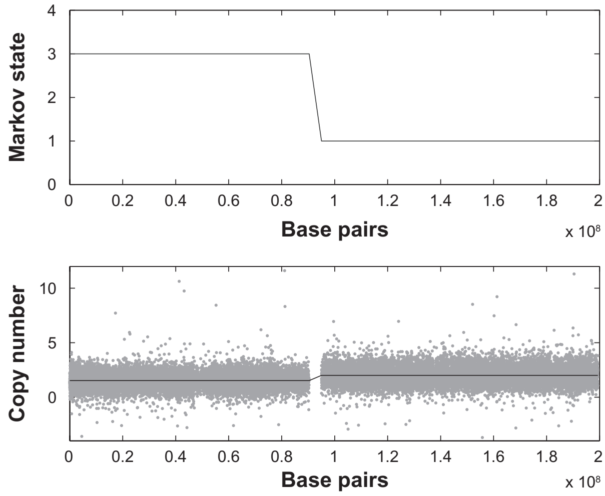

Figure 3. Top: Viterbi reconstruction of the Markov path for chromosome 3 in PD1753. Bottom: sum of (standardized) allele intensities for probes within the same chromosome (grey dots), and the copy number of the corresponding state (black solid line).

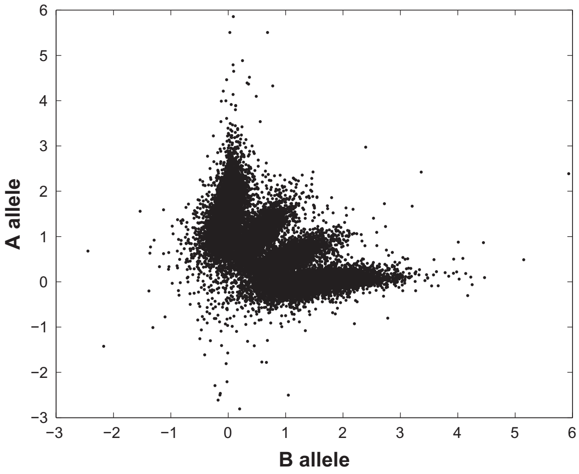

Figure 4. Scatter plot of standardized measured allele intensities in the segment reconstructed to Markov state 2 in Figure 3. The fraction of normal contamination was estimated at 0.53.

5. Discussion

We have presented a method to estimate the number of copies of each of the two alleles in SNP data, taking three features common in cancer data into account; unequally spaced probes, aneuploidy, and normal contamination. Unequally spaced probes are modeled

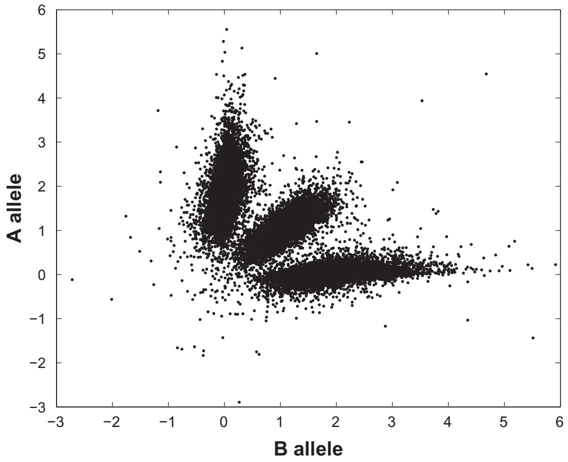

Figure 5. Scatter plot of standardized measured allele intensities in the segment reconstructed to Markov state 4 in Figure 3.

using a continuous-index Markov chain instead of a discrete-index one, which is the usual choice in the literature. The ploidy and fraction of normal contamination are both included as parameters in the model, which allows us to estimate them along with other variables and using all the data, rather than estimating them separately in a pre-processing step. This set-up also allows us to retain the integer structure of the allele copy numbers. The model's ability to estimate the fraction of normal contamination has been demonstrated in a simulation study, with the results being far better than for previous methods and excellent even with as much as 70% normal contamination.

Above we denoted Markov state 4, ie, the state with genotypes {AA, AB, BB}, the *normal state*, irrespective of the ploidy of the chromosome. The reason for singling out this particular state is that it is often particularly interesting whether the Markov chain is in this state or not, at any given probe. One could argue that if the ploidy differs from two this is not 'normal', but it is straightforward to select a different state as 'normal' and then modify the transition rate structure and estimation algorithm accordingly.

The emission model, ie, Eqs. (1)–(4), assume that the means and variances of the measured intensities are both linear in the amount of each allele. In practice this assumption may fail, eg, because for large copy numbers the response is nonlinear. One could then include such a non-linearity in the model, and model the mean intensities as $\mu_{kcg} = h_{kc}(g;\theta_{kc})$ where h is some function and θ_{kc} parameters of this function. Ideally the functional form h as well as all its probe-speficic parameters θ_{kc} should be well estimated beforehand, so that they are essientially known when evaluating an unknown sample. Similar comments apply to the variance of the measured intensities.

In this paper we have only considered probes that provide allele-specific intensity measurements, but, as mentioned in Section 2, microarrays often also contain probes that measure the total copy number only, ie, the sum of the number of alleles. Such probes can easily be included in our model by speficying a corresponding suitable emission density, ie, a density corresponding to Eq. (3). For instance, this could be a univariate Normal density with mean $\mu_{kcg} = C_{0kc} + C_{1kc}(g_A + g_B)$ and variance $\sigma^2_{kcg} = \nu_{kc}\,\mu^2_{kcg}$ for parameters C_{0kc}, C_{1kc} and ν_{kc} that again need to be estimated prior to analyzing an unknown sample. Should the response function from total copy number to intensity not be linear for large copy numbers, this could be handled similarly to what can be done for SNP probes; cf. the previous paragraph.

Finally we mention some possible limitations of our method. Firstly, the accuracy of the method is likely to be reduced in regions of very high copy number where signal saturation occurs, such as in amplicons, and bespoke non-linear adjustments may be required (as discused above). Secondly, we have ignored copy number polymorhisms. These will produce non-integer copy numbers in the cancer sample due to the skewed ratio between the cancer and the contaminating normal. If copy number data is available for the normal, it may be possible to generalise these methods to make such an adjustment, however, such regions are generally a lot smaller in scale than the somatic copy number changes seen in cancer and were not considered further. Lastly, we have assumed that the sample in question is derived from a homogeneous collection of cells. However, cell-to-cell variation is quite possibly going to produce a lot of different clones with differing copy numbers, and more general methods will be required to deal with such complexities.

To sum up this paper, copy number variations in cancer are common and their accurate determination is important for determining homozygous deletion, amplifications and breakpoints, all of which can be functionally implicated in cancer. This problem is compounded by normal contamination, making the accurate estimation of integer copy numbers in cancer samples with normal contamination difficult. Here we have introduced a method that addresses this problem.

Acknowledgments

CDG was supported by the Wellcome Trust at the Sanger Centre. The authors would like to thank the anonymous reviewers for their constructive comments and suggestions that improved the presentation of this paper.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Andersson R, Bruder CEG, Piotrowski A, et al. A segmental maximum a posteriori approach to genome-wide Copy Number profiling. *Bioinformatics*. 2008;24:751–8.
2. Attiyeh EF, Diskin SJ, Attiyeh MA, et al. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res*. 2009;19:276–83.
3. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an objective Bayes hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35:2013–25.
4. Daruwala R, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Nat Acad Sci*. 2004;101:16292–7.
5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc B*. 1977;39: 1–38.
6. Eilers PHC, de Menezes RX. Quantile smoothing of array CGH data. *Bioinformatics*. 2005;21:1146–53.
7. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN. Hidden Markov models approach to the analysis of array CGH data. *J Multivar Anal*. 2004;90:132–53.
8. Greenman CD, Bignell G, Butler A, et al. PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatist*. 2010;11:164–75.
9. Guha S, Li Y, Neuberg D. Bayesian hidden Markov modeling of array CGH data. *J Amer Statist Assoc*. 2008;103:485–97.
10. Mitra R, Gupta M. A continuous-index Bayesian hidden Markov model for prediction of nucleosome positioning in genomic DNA. *Biostatist*. to appear.
11. Huang J, Wei W, Chen J, et al. CARAT: A novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*. 2006;7:83.
12. Hupé P, Stransky N, Thiery J, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*. 2004;20:3413–22.
13. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs common copy number polymorphisms and rare CNVs. *Nature Genetics*. 2008;40:1253–60.
14. Koski T. *Hidden Markov Models for Bioinformatics*. Dordrecht: Kluwer Academic Publishers; 2001.
15. Laframboise T, Harrington D, Weir BA. PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatist*. 2007;8:323–36.
16. Lai TL, Xing H, Zhang N. Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatist*. 2008;9: 290–307.
17. Lamy P, Andersen CL, Dyrskjot L, Torring N, Wiuf C. A hidden Markov model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. *BMC Bioinformatics*. 2007;8:434.
18. Li C, Beroukhim R, Weir BA, Winckler W, Garraway LA, Sellers WT, et al. Major copy proportion analysis of tumor smples using SNP arrays. *BMC Bioinformatics*. 2008;9:204.
19. Marioni JC, Thorne NP, Tavaré S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*. 2006;22: 1144–6.
20. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatist*. 2004;5:557–72.
21. Picard F, Robin S, Lavielle M, Vaisse C, Daudin J. A statistical approach for array CGH data analysis. *BMC Bioinformatics*. 2005;6:27.
22. Popova T, Mani´e E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biology*. 2009;10:R128.
23. Roberts WJJ, Ephraim Y. An EM Algorithm for ion-channel current estimation. *IEEE Trans Signal Proc*. 2008;56:26–33.
24. Rueda OM, Días R. Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput Biol*. 2007;3:1115–22.
25. Rydén T. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective (with discussion). *Bayesian Anal*. 2008;3,659–88.
26. Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I. Hidden Markov models for the assesment of chromosomal alterations using high-throughput SNP arrays. *Ann Appl Statist*. 2008;2:687–713.
27. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, et al. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*. 2006;22:e431–9.
28. Stjernqvist S, Rydén T, Sköld M, Staaf J. Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*. 2007;23:1006–14.
29. Stjernqvist S, Rydén T. A continuous-index hidden Markov jump process for modelling DNA copy number data. *Biostatist*. 2009;10:773–8.
30. Sun W, Wright FA, Tang Z, et al. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res*. 2009;37:5365–77.
31. Wang K, Li M, Hadley D. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17:1665–74.
32. www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=sample&id=919182.
33. www.maths.lth.se/matstat/staff/susann/.

Appendix

1. A Primer on Markov Chains and Hidden Markov Model

The purpose of this section is to provide a brief and rather elementary introduction to Markov chains with discrete and continuous index, and to hidden Markov models. A monograph entirely devoted to bioinformatics applications of hidden Markov models is the text by Koski.[14]

Consider a sequence $t_1, t_2, \ldots, t_N$ of locations (in our case these will be probe locations), and a set $\{1, 2, \ldots, r\}$ of states (which will in our case be as in Tables 1 or 2). At any location t_k there is an actual state $x(t_k)$ (ie, a true copy number state), which we think of as the realization of a random variable $X(t_k)$. These random variables are dependent, since copy number states at nearby probes are correlated. To model this dependence, we use Markov chains.

A discrete-index Markov chain (we use the term *index* rather than the more common 'time', since bp location is not a temporal variable) is specified by *transition probabilities* $p_{ij}(t_{k-1}, t_k)$, giving the (conditional) probability that if the chain happens to be in state i at location t_{k-1}, it will move to state j at location t_k. For $j = i$, the probability concerns the event that the chain will stay in the same state, ie, not move at all. Implicit in this characterization is also the fact that if the states $x(t_1), x(t_2), \ldots, x(t_{k-1})$ at all foregoing locations $t_1, t_2, \ldots, t_{k-1}$ are known, this does not affect the conditional probability, which only depends on the state $x(t_{k-1})$ at the closest location t_{k-1}; this is the *Markov property*. To complete the specification of the Markov chain, we must also provide the *initial probabilities*, ie, the probabilities that at the first location t_1, the chain starts in state i for each respective state.

In our model, the probe locations t_k are separated by different distances $t_k - t_{k-1}$, ie, these distances are not equal. We wish to incorporate this feature into the Markov model, so that the transition probabilities $p_{ij}(t_{k-1}, t_k)$ do not only depend on the states i and j that the chain moves from and to respectively, but also on the distance $h_k = t_k - t_{k-1}$ between the probes. One way to accomplish this is to think of the base pair location along a chromosome, which we denote by t, as a *continuous* variable rather than as a discrete one, and to model the state changes of the Markov chain using this continuous variable, or index. In contrast to a discrete-index Markov chain, a *continuous-index* Markov chain is specified in terms of *transition rates*. For any state i and any other state j, ie, different from i, there is a transition rate q_{ij} from state i state j. For any state i we also define the *total rate out of* i, q_i, as the sum of all transition rates out of this state, ie, $q_i = \sum_{j\neq i} q_{ij}$. One way to interpret these rates is in terms of sojourn lengths and jump probabilities. Given that the chain has entered state i, it will stay there for a sojourn whose length is random and follows an exponential distribution with rate q_i (mean $1/q_\mathrm{i}$); the probability that this sojourn exceeds length s is thus the exponential $\exp(-q_i s)$. When then the chain eventually leaves state i, the probability that it jumps to state j is given by q_{ij}/q_i.

For a continuous-index Markov chain it is also possible to compute the probability that for two locations t_{k-1} and t_k separated by distance h_k, if the chain is in state i at location t_{k-1}, it will be in state j at location t_k. Denoting these probabilities by $p_{ij}(h_k)$ and collecting then into an $r \times r$ matrix $P(h_k)$ (thus $p_{ij}(h_k)$ is the row i column j element of this matrix), it holds that $P(h_k) = \exp(Qh_k)$, where Q is the $r \times r$ rate matrix (or *intensity matrix*, or *generator*) with off-diagonal elements given by the transition rates q_{ij} and diagonal elements $q_{ii} = -q_i$, ie, the negative of the total rates out of the respective states. Moreover, $\exp(Qh_k)$ is the *matrix-exponential function*, defined by the power series $\exp(A) = I + A + A^2/2! + A^3/3! + \ldots$ for any square matrix A, where I is the identity matrix, ie, a matrix of the same size as A and with diagonal elements equal to one and all off-diagonal elements being zero, and $k!$ is the factorial $1 \times 2 \times \ldots \times k$. This definition is a direct generalization of the power series for the ordinary (real-valued) exponential function.

In a *hidden Markov model* (HMM), the Markov chain is not directly observable, but only as disturbed by noise. In the present setting the copy number state cannot be observed with certainty, but for any probe the intensity measurements, for each allele, provide partial information about the copy number state. In an HMM, the link between the state $X(t_k)$ at some location t_k and the corresponding measurement Y_k (here, intensities) is specified through an *emission density* $f_{Y_k|X(t_k)}(y\,|\,i)$, which is the conditional density of Y_k given that $X(t_k) = i$. In the present context the emission density is thus the density of the measured intensities given a certain copy number state. Since there are two intensities available, one for each allele, the density is a bivariate one. Furthermore, since each copy number state (Markov state) contains several genotypes, the emission density for a copy number state is a mixture (weighted average) of densities corresponding to each of these genotypes; this is Eq. (3).

Specifying the HMM thus amounts to specifiying the structure and parameters of the Markov chain, and those of the emission densities. When this has been done, typical tasks are to i) estimate parameters from data, and ii) find the most likely realization of the Markov chain, given data. The first task, parameter estimation, is commonly carried out using the so-called EM (expectation maximization) algorithm, which is an iterative procedure that in each iteration increases the likelihood of the model parameters. The purpose is thus to iterate until convergence, and then to report the resulting parameters as the MLE (maximum likelihood estimate); convergence to the MLE is not guaranteed, however. For our HMM, the algorithm is outlined in Appendix 2.1. The second problem above can be viewed as that of *reconstructing* the Markov trajectory, given data (and model parameters, usually estimated ones). This problem is solved using the so-called *Viterbi algorithm,* which is a dynamic programming algorithm that recursively finds the most likely path. This algorithm, for our HMM, is described in Appendix 2.2.

2. Methods

2.1 The EM algorithm

The parameters to estimate are the ploidy K, the fraction γ of normal tissue, and, for each chromosome c, the two transition rates λ_c and η_c and the initial distribution δ_c. Our starting point is the EM algorithm for continuous-index hidden Markov chains by Roberts and Ephraim.[23] As latent (unobserved) data we take the whole Markov trajectory $(X_c(t))_{t_{1c} \le t \le T_c}$ for each chromosome c, but the complete likelihood involves only the sufficient statistics consisting of the initial state $X_c(t_{1c})$, the total lengths T_{nc} and T_{ac} of sojourns in the normal state and in abnormal states respectively, and the numbers $m_{\cdot ac}$ and m_{anc} of jumps to abnormal states, and from abnormal states to the normal state respectively, for chromosome c.

With these sufficient statistics, and recalling that each λ_c has a Gamma prior with shape and intensity parameters say α_c^λ and β_c^λ, and analogously for each η_c, the complete log-posterior, ie, the sum of the complete log-likelihood and the log-prior, is, up to a constant not depending on the parameters,

$$\begin{aligned} L^c(\theta, \boldsymbol{X}, \boldsymbol{y}) = \sum_c \Big\{ & \log \delta_{X(t_1),c} + m_{\cdot\text{ac}} \log \lambda_c + m_{\text{anc}} \log \eta_c \\ & - \lambda_c T_{\text{nc}} - (\lambda_c + \eta_c) T_{\text{ac}} \\ & + (\alpha_c^\lambda - 1) \log \lambda_c - \beta_c^\lambda \lambda_c \\ & + (\alpha_c^\eta - 1) \log \eta_c - \beta_c^\eta \eta_c \\ & + \sum_{k=1}^{N_c} \log f_{Y_{kc}|X_{kc}}(y_{kc}|X_{kc}; \gamma, K) \Big\}, \end{aligned}$$

where $\theta = (\delta_c, \lambda_c, \eta_c, K, \gamma)$. Moreover, $\boldsymbol{y} = \{y_{kc}\}$ is the collection of all data and $\boldsymbol{X} = \{(X_c(t))\}$ is the collection of all (unobserved) Markov chain trajectories. The quantity to maximize in one iteration of the EM algorithm is

$$Q(\theta;\theta') = E_\theta[L^c(\theta';\boldsymbol{X},\boldsymbol{y})|\boldsymbol{y}],$$

where maximization is with respect to θ' and the notation E_θ indicates that the expectation is computed under the current parameter (estimate) θ. Note that $L^c(\theta';\boldsymbol{X},\boldsymbol{y})$ and hence also $Q(\theta;\theta')$ split into two distinct parts, one of which depends on $(\delta_c, \lambda_c, \eta_c)$ only and one of which depends on K and γ only. Maximization with respect to $(\delta_c, \lambda_c, \eta_c)$ and with respect to (K, γ) can thus be carried out separately.

Also note that K and γ are common across the genome, and therefore estimated using the data for all chromosomes. For each iteration of the EM algorithm we compute the forward and backward variables for each chromosome, store them, and then re-estimate K and γ using the information from all chromosomes.

The M-steps for the transition rates read

$$\hat{\lambda}_c = \frac{\alpha_c^\lambda - 1 + \hat{m}_{\cdot\text{ac}}}{\beta_c^\lambda + \hat{T}_{\text{ac}} + \hat{T}_{\text{nc}}}, \quad \hat{\eta}_c = \frac{\alpha_c^\eta - 1 + \hat{m}_{\text{anc}}}{\beta_c^\eta + \hat{T}_{\text{ac}}},$$

where $\hat{m}_{\cdot\text{ac}} = E_\theta[m_{\cdot\text{ac}}|y_{1c}, \cdots, y_{c,N_c}]$ etc. Note that $T_{\text{ac}} + T_{\text{nc}}$ equal the length of the Markov chain trajectory for chromosome c, ie, $T_c - t_{1c}$, so that also

$\hat{T}_{ac} + \hat{T}_{nc} = T_c - t_{1c}$. Moreover, the M-step for the initial distributions is

$$\hat{\delta}_{ic} = P_{\theta}(X(t_1) = i \mid y_{1c}, \ldots, y_{c,N_c}).$$

The M-step for the ploidy is

$$\hat{K} = -\frac{V}{4U} + \sqrt{\frac{V^2}{16U^2} - \frac{\sum_c N_c}{U}},$$

where

$$U = -\sum_{c,k,i,g\in G_i} \frac{1}{8\nu_{kc}(1-\rho_{kc}^2)}\left(\frac{y_{Akc}^2}{\mu_{Akcg}^2} - \frac{2\rho_{kc} y_{Akc} y_{Bkc}}{\mu_{Akcg}\mu_{Bkcg}} + \frac{y_{Bkc}^2}{\mu_{Bkcg}^2}\right)$$
$$\times P_{\theta}(X_c(t_{kc}) = i, G_{kc} = g | y_{1c}, \ldots, y_{c,Nc})$$

and

$$V = -\sum_{c,k,i,g\in G_i} \frac{1}{2v_{kc}(1+\rho_{kc})}\left(\frac{y_{Akc}}{\mu_{Akcg}} + \frac{y_{Bkc}}{\mu_{Bkcg}}\right)$$
$$\times P_{\theta}(X_c(t_{kc}) = i, G_{kc} = g | y_{1c}, \ldots, y_{c,Nc})$$

For the fraction γ of normal contamination there is no closed form expression for the M-step, and to re-estimate γ we maximize $Q(\theta;\cdot)$, as a function of γ, numerically. Note however that $\hat{K}$ above depends on γ, which appears implicitly in the means μ_{Akcg} and μ_{Bkcg} used to compute U and V. Therefore, by maximizing w.r.t. K' (using the current γ) and then w.r.t. γ′ (using the re-estimated $\hat{K}$), as we do, and not w.r.t. K' and γ′ jointly, we in fact obtain a generalized EM algorithm rather than an EM algorithm, in the terminology of Dempster et al[5] (Eq. (3.5)).

The conditional expectations $\hat{m}_{\cdot ac}$ etc. are computed in the E-step, which follows that of Roberts and Ephraim[23] with minor changes. Now write $y_{k:l,c}$ for $\{y_{kc}, y_{k+1,c}, \ldots, y_{lc}\}$, and let m_{ijc} be the number of jumps by the Markov chain from state i to state j, in chromosome c. Then

$$\hat{m}_{ijc} = E_{\theta}[m_{ijc} \mid y_{1:N_c,c}]$$
$$= \int_0^{T_c} P_{\theta}(X_c(t-) = i, X_c(t) = j \mid y_{1:N_c,c})dt$$
$$= \int_0^{T_c} \frac{P_{\theta}(X_c(t-) = i, X_c(t) = j)}{p_{\theta}(y_{1:N_c,c})}$$
$$\times\ p_{\theta}(y_{1:N_c,c} \mid X_{c(t-)} = i, X_c(t) = j)dt$$
$$= \sum_{k=2}^{N_c} \int_{t_{k-1,c}}^{t_{kc}} \frac{P_{\theta}(X_c(t) = j \mid X_c(t-) = i)P_{\theta}(X_c(t-) = i)}{p_{\theta}(y_{1:N_c,c})}$$
$$\times\ p_{\theta}(y_{1:k-1,c}, y_{k:N_c,c} \mid X_c(t-) = i, X_c(t) = j)dt;$$

here the symbol P denotes probabilities as well as densities; note that $P_{\theta}(X_c(t) = j|X_c(t-) = i) = q_{ijc}$, where q_{ijc} is the transition rate from state i to state j in chromosome c. Thus, with r_{abnormal} being the number of abnormal states, q_{ijc} is equal to $\lambda_c/r_{\text{abnormal}}$ if i is the normal state and j is any abnormal state, equal to $\lambda_c/(r_{\text{abnormal}} - 1)$ if i and j are both abnormal states (because the chain cannot jump from a state to itself), and equal to η_c if i is any abnormal state and j is the normal state. Given the HMM structure it follows that $y_{1:k-1,c}$ and $y_{k:N_c,c}$ are conditionally independent given $X_c(t-) = i$ and $X_c(t) = j$, whence

$$\hat{m}_{ijc} = \frac{q_{ijc}}{p_{\theta}(y_{1:N_c,c})} \sum_{\kappa=2}^{N_c} \int_{t_{k-1,c}}^{t_{kc}} P_{\theta}(X_c(t-) = i) p_{\theta}(y_{1:k-1,c}|X_c(t-) = i)$$
$$\times\ P_{\theta}(y_{k:N_c,c}|X_c(t) = j)dt$$
$$= \frac{q_{ijc}}{p_{\theta}(y_{1:N_c,c})} \sum_{\kappa=2}^{N_c} \int_0^{h_{kc}} P_{\theta}(y_{1:k-1,c}, X_c(t_{k-1,c} + t-) = i)$$
$$\times\ P(y_{k:N_c,c}|X_c(t_{k-1,c} + t) = j)dt$$

with $h_{kc} = t_{kc} - t_{k-1,c}$. Here, the two factors in the integrand on the right-hand side are the forward and backward densities respectively.

To compute these factors, and similar ones, we use a forward-backward type algorithm. Let r be the number of Markov states, and let B_{kc} be the $r \times r$ diagonal matrix whose (i, i)-th element is the probability density function of y_{kc} given Markov state i at position t_{kc}, ie, $f_{Y_{kc}|X_c(t_{kc})} = i(y_{kc})$ in Eq. (3). Further let F_{kc} be the $r \times r$ matrix whose (i, j)-th element is

$$[F_{kc}]_{ij} = P_{\theta}(y_{kc}, X_c(t_{kc}) = j \mid X_c(t_{k-1,c}) = i)$$
$$= [\exp(Q_c h_{kc})]_{ij}[B_{kc}]_{jj},$$

where Q_c is the matrix with elements q_{ijc}, $i, j = 1, 2, \ldots, r$ for $i \neq j$, and diagonal elements q_{cii} being the negative of the total rate out of state i for chromosome c (the row sums of Q_c then become zero). We note that the discrete-index process $(X_c(t_{kc}))_{1\le k\le N_c}$, ie, the continuous-index process $(X_c(\cdot))$ sampled at the locations of the probes, is a non-homogeneous discrete-index Markov chain with transition probability matrices, from $t_{k-1,c}$ to t_{kc}, given by $\exp(Q_c h_{kc})$. With this matrix notation we have $F_{kc} = \exp(Q_c h_{kc})B_{kc}$, and the likelihood for chromosome c can be written

$$p_\theta(y_{1:N_c,c}) = \delta_c B_{1c}\left(\prod_{k=2}^{N_c} F_{kc}\right)\mathbf{1}$$

where $\mathbf{1}$ is the $r \times 1$ vector of all ones. The forward densities are

$$\begin{aligned} &P_\theta(y_{1:k-1,c,}X_c(t_{k-1,c}+t-)=i) \\ &\quad= \sum_{s=1}^{r} P_\theta(y_{1:k-1,c}, X_c(t_{k-1,c})=s) \\ &\qquad\times P_\theta(X_c(t_{k-1,c}+t-)=i|X_c(t_{k-1,c})=s) \\ &\quad= \sum_{s=1}^{r}\left(\delta_c B_{1c}\prod_{\kappa=2}^{k-1}F_{\kappa c}\right)\mathbf{1}_s[\exp(Q_c t)]_{si},\end{aligned}$$

where $\mathbf{1}_j$ is the $r \times 1$ vector whose elements are zero except for element j which is one.

The backward densities are

$$\begin{aligned} &p_\theta(y_{k:N_c,c} \mid X_c(t_{k-1,c}+t)=j) \\ &\quad= \sum_{s=1}^{r} p_\theta(y_{k+1:N_c,c} \mid X_c(t_{kc})=s)\,p_\theta(y_{kc} \mid X_c(t_{kc})=s) \\ &\qquad\times P_\theta(X_c(t_{kc})=s \mid X_c(t_{k-1,c}+t)=j) \\ &\quad= \sum_{s=1}^{r}[B_{kc}]_{ss}[\exp(Q_c(h_{kc}-t))]_{js}\,\mathbf{1}_s'\left(\prod_{\kappa=k+1}^{N_c}F_{\kappa c}\right)\mathbf{1}.\end{aligned}$$

The above matrix multiplications are numerically unstable, as the products will either tend to zero or infinity exponentially fast as the number of factors increases. Therefore scaled versions of these recursions are introduced. The scaled forward densities with normalizing constants d_{kc} at probe (k, c) are

$$L_c(k) = \frac{\delta_c B_{1c}}{d_{1c}}\prod_{\kappa=2}^{k}\frac{F_{\kappa c}}{d_{\kappa c}},$$

which we compute recursively as

$$L_c(k) = \frac{L_c(k-1)F_{kc}}{d_{kc}}$$

with $d_{1c} = \delta_c B_{1c}\mathbf{1}$, $L_c(1) = \delta_c B_{1c}/d_{1c}$, $d_{kc} = L_c(k-1)F_{kc}\mathbf{1}$. The scaled backward densities are

$$R_c(k) = \prod_{\kappa=k+1}^{N_c}\frac{F_{\kappa c}}{d_{\kappa c}}\mathbf{1},$$

which we compute as

$$R_c(k) = \frac{F_{k+1}R_c(k+1)}{d_{k+1,c}}$$

with $R_c(N_c) = \mathbf{1}$.

Using these scales quantities, the matrix $\hat{m}_c$ with entries $\hat{m}_{ijc}$ can be expressed as

$$\hat{m}_c = Q_c \odot I'_{k,c},$$

where ⊙ denotes element-wise multiplication and

$$I_{kc} = \int_0^{h_{kc}} \exp(Q_c(h_{kc}-t))V_c\exp(Q_c t)dt$$

with

$$V_c = \sum_{k=2}^{N_c}\frac{B_{kc}R_c(k)L_c(k-1)}{d_{k,c}}.$$

The integrals I_{kc} are evaluated using the matrix

$$D_c = \begin{pmatrix} Q_c & V_c \\ 0 & Q_c \end{pmatrix};$$

I_{kc} is then upper right $r \times r$ block of $\exp(D_c h_{kc})$. Finally, recalling that the normal state is state 4,

$$\hat{m}_{\cdot\mathrm{ac}} = \sum_{1\le i\le r}\ \sum_{1\le j\le r,\ j\ne i, j\ne 4}\hat{m}_{ijc},$$

$$\hat{m}_{\mathrm{anc}} = \sum_{1\le i\le r, i\ne 4}\hat{m}_{i4c}.$$

Using similar types of computations is follows that

$$\hat{T}_{ic} = E[T_{ic} \mid y_{1:N_c,c}] = \hat{m}_{iic}/q_{iic},$$

where T_{ic} is the total length of all sojourns of the Markov chain in state i within chromosome c. Moreover,

$$P_\theta(X(t_{1c}) = i \mid y_{1:N_c,c}) \propto L_c(1)_i R_c(1)_i,$$

and the conditional probabilities in the expressions for U and V are computed using

$$\begin{aligned} &P(X_c(t_{kc}) = i, G_{kc} = g \mid y_{1:N_c}) \\ &\quad = P(G_{kc} = g \mid X_{kc} = i, y_{1:N_c,c}) P(X_{kc} = i \mid y_{1:N_c,c}) \\ &\qquad \propto w_{kcig} f(y_{kc} \mid G_{kc} = g) L_c(k)_i R_c(k)_i, \end{aligned}$$

where the weights and densities on the right-hand side are those in Eq. (3).

2.2 The Viterbi algorithm

We used a Viterbi algorithm, adapted to the continuous-index structure, to find the aposteriori most likely Markov chain trajectory. The algorithm is the usual Viterbi algorithm, but with transition probability matrices $\exp(Q_c h_{kc})$ that vary with probe index (c, k). The algorithm thus finds the most likely sequence at the probe locations only. When the estimated reconstruction of each Markov state $X(t_{kc})$ is available, one may also estimate the corresponding genotype G_{kc} (see below).

For any chromosome c, the Viterbi algorithm is as follows. To ensure numeric stability, it operates on log-scale.

1. Put $\xi_{1c}(i) = \log(\delta_{ic}[B_{1c}]_{ii})$ for $i = 1, \ldots, r$.
2. Iterate for $k = 2, 3, \ldots, N_c$,
$$\xi_{kc}(j) = \max_i \{\xi_{k-1}(i) + \log[\exp(Q_c h_{kc})]_{ij} + \log[B_{kc}]_{jj}\}$$
for $i = 1, 2, \ldots, r$.
3. Put $\hat{x}_c(N_c) = \arg\max_i \xi_{N_c,c}(i)$.
4. Iterate for $k = N_c - 1, N_c - 2, \ldots, 1$,
$$\hat{x}_c(k) = \arg\max_i \left\{\xi_{kc}(i) + \log[\exp(Qh_{k+1})]_{i,\hat{x}_c(k+1)}\right\}$$

Having reconstructed the states $x_c(t_{kc})$, it holds that the corresponding genotypes, given the Markov chain and intensity data, are conditionally independent with

$$\begin{aligned} &P(G_{kc} = g | X_c(t_{kc}) = i, Y_{kc} = y) \\ &\quad = \frac{w_{kcig} f_{Y_{kc}|G_{kc}}(y|g)}{\Sigma_{g' \in G_i} w_{kcig'} f_{Y_{kc}|G_{kc}}(y|g')} \end{aligned}$$

for all $g \in G_i$; here $f_{Y_{kc}|G_{kc}}(y|g)$ is the bivariate Normal density as in Eq. (3). Selecting, for each probe (k, c), G_{kc} as the genotype $g \in G_i$ maximizing this expression thus yields a maximum aposteriori (MAP) reconstruction of the genotypes at all probes.