OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Estimation of Parent Specific DNA Copy Number in Tumors using High-Density Genotyping Arrays

Hao Chen[1], Haipeng Xing[2], Nancy R. Zhang[3]*

1 Department of Statistics, Stanford University, Stanford, California, United States of America, **2** Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, New York, United States of America, **3** Department of Statistics, Stanford University, Stanford, California, United States of America

Abstract

Chromosomal gains and losses comprise an important type of genetic change in tumors, and can now be assayed using microarray hybridization-based experiments. Most current statistical models for DNA copy number estimate total copy number, which do not distinguish between the underlying quantities of the two inherited chromosomes. This latter information, sometimes called *parent specific copy number*, is important for identifying allele-specific amplifications and deletions, for quantifying normal cell contamination, and for giving a more complete molecular portrait of the tumor. We propose a stochastic segmentation model for parent-specific DNA copy number in tumor samples, and give an estimation procedure that is computationally efficient and can be applied to data from the current high density genotyping platforms. The proposed method does not require matched normal samples, and can estimate the unknown genotypes simultaneously with the parent specific copy number. The new method is used to analyze 223 glioblastoma samples from the Cancer Genome Atlas (TCGA) project, giving a more comprehensive summary of the copy number events in these samples. Detailed case studies on these samples reveal the additional insights that can be gained from an allele-specific copy number analysis, such as the quantification of fractional gains and losses, the identification of copy neutral loss of heterozygosity, and the characterization of regions of simultaneous changes of both inherited chromosomes.

Citation: Chen H, Xing H, Zhang NR (2011) Estimation of Parent Specific DNA Copy Number in Tumors using High-Density Genotyping Arrays. PLoS Comput Biol 7(1): e1001060. doi:10.1371/journal.pcbi.1001060

Editor: Scott Markel, Accelrys, United States of America

Received January 7, 2010; **Accepted** December 17, 2010; **Published** January 27, 2011

Funding: Hao Chen's research was supported in part by NIH Merit Award R37EB02784. Haipeng Xing's research was supported by the National Science Foundation grant DMS-0906593. Nancy R. Zhang's research was supported in part by the National Science Foundation grant DMS-0906394. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nzhang@stanford.edu

Introduction

DNA copy number aberration (CNA), defined as gains or losses of specific chromosomal segments, are an important type of genetic change in tumors. Various microarray based experimental platforms [1–7] have made possible the fine scale measurement of CNAs. Whereas the earlier platforms such as comparative genome hybridization arrays were designed to measure the total copy number of both inherited chromosomes, other platforms such as high density genotyping microarrays [6–8] can measure allele specific DNA quantity. For alleles that represent known variants of genes, it would be of biological interest to know which allele has undergone copy number change [9]. Also, some genetic mechanisms, such as gene conversion, mitotic recombination, and uniparental disomy, cause loss of heterozygosity (LOH) without change in total DNA copy number, and thus can not be detected through conventional analysis methods relying only on total copy number. Even in the case where the total DNA copy number changes, it would be informative to know whether one or both of the inherited parental chromosomes are involved. Thus, to construct a more detailed molecular portrait of tumors, we need to distinguish between the underlying quantities of the two inherited chromosomes, which we call the *parent specific copy numbers*.

This paper addresses the problem of parent specific copy number estimation using allele-specific raw copy number data from high-density genotyping arrays. We will describe the data in more detail in the next section. Here, we clarify the differences between total copy number analysis and parent specific copy number analysis, and review the background of the computational treatment of this problem.

The genome of each somatic human cell normally contains two copies of each of the 22 autosomes, one inherited from each biological parent. At any genome location, one or both of these two chromosomes may gain or lose copies, thus creating a change in total copy number at that location. Microarray experiments for measuring total copy number produce a sequence of continuous valued measurements mapping to ordered locations along the chromosomes. Computational methods can be applied to segment this noisy sequence of measurements into regions of homogeneous copy number [10–21], see Lai and Park [22] and Willenbrock and Fridlyand [23] for a review. Since chromosomes are gained and lost in contiguous segments, the true total copy number should be piecewise continuous. This is why change-point models and hidden Markov models have been very useful for total copy number estimation.

Total copy number estimates do not reveal which (or both) of the two inherited chromosomes have been gained or lost, and if a locus is polymorphic, which (or both) of the alleles have been affected. This information is now available in data produced by high density genotyping platforms, which give, at selected single nucleotide polymorphisms (SNPs), a bivariate measurement quantifying the two alleles which we arbitrarily label A and B,

Author Summary

Many genetic diseases are related to copy number aberrations of some regions of the genome. As we know, each chromosome normally has two copies. However, under some circumstances, for some regions, either one or both of the chromosomes change. Genotyping microarray data provides the copy number of the two alleles of polymorphic sites along the chromosomes, which make the inference of the copy number aberrations of the chromosome feasible. One difficulty is that genotyping microarray data cannot provide the haplotype of the two copies of a chromosome. In this paper, we model the copy number along the chromosome as a two-dimensional Markov Chain. Using the observed copy number of both alleles of all the sites, we can determine the parent specific copy number along the chromosome as well as infer the haplotypes of the two copies of the inherited chromosomes in regions where there is allelic imbalance. Simulation results show high sensitivity and specificity of the method. Applying this method to glioblastoma samples from the Cancer Genome Atlas data illustrate the insights gained from allele-specific copy number analysis.

as shown in the left panel of Figure 1. Some platforms output the total raw copy number (R), which is the sum of A and B, and the B-allele frequency (BAF), which is the percentage of B allele raw copy number among the total allele raw copy number, i.e., $B/(A+B)$. The logR quantifies the total copy number, while the BAF quantifies the imbalance between the two alleles. The right panel of Figure 1 shows R, the sum of A and B allele intensities, and BAF. Unlike the total copy number, the allele-specific measurements are mixtures that depend on the unknown genotype at each location. For this reason, conventional change-point models can not be applied to allele specific copy number estimation.

This problem can be formulated statistically as follows: The observed A and B intensities form a bivariate sequence whose underlying distribution undergoes abrupt changes. The distributions at each location are mixtures. Both the change-points, the mixture components, and the cluster memberships at each data point are unknown and must be estimated from the data.

There have been much effort extending existing genotyping and total copy number segmentation procedures to analyze allele-specific data. At the probe level, CNAT [24], CN5 [24], CRMA [25], dChipSNP [26,27], PLASQ [28], and PICR [29] can be applied to Affymetrix data to produce allele-specific probe-set summaries at each SNP location. However, just as in the estimation of total copy number, the allele-specific intensities for adjacent SNPs should be smoothed to infer the underlying parent-specific copy numbers. LaFramboise et al. [28] first segmented the total copy number using Circular Binary Segmentation [30], and then estimated the parent-specific copy numbers for each segment. This early approach misses copy neutral loss-of -heterozygosity (LOH) events, defined as the simultaneous gain of one chromosome and balanced loss of the other chromosome resulting in loss of heterozygosity but no change in total copy number. Many other existing approaches rely on discrete-state hidden Markov models [27,31–34], which are hidden Markov models assuming a pre-specified finite set of underlying states. For example, PennCNV [32] and QuantiSNP [33] assume that the underlying copy numbers belong to the integer classes $\{0,1,\ldots,6\}$, and that the allele-specific copy numbers can be described by "generalized genotypes" AA, AB, BB, A-, B-, AAB, ABB, etc. While these types of models are very useful for detecting germline copy number variants in normal tissue, they do not generalize well to genetically heterogeneous samples. This is because by requiring a fixed set of pre-defined discrete states, they do not account for the heterogeneity of cells within the sample, which produces data with apparently fractional copy number changes rather than the idealized unit-copy changes. This is especially problematic for tumor samples, which are usually heterogeneous mixtures of cells with different genetic profiles. Through titration studies, Staaf et al. [35] showed that methods relying on idealized genotype states lose sensitivity when tumors are diluted with normal cells.

The fractional changes in tumors inspired recent approaches [35,36] that segment both the logR and BAF simultaneously. Since BAF is a mixture of homozygous and heterozygous SNPs, it cannot be processed using existing segmentation procedures. Current methods solve this problem through a pre-processing step that gets rid of the homozygous SNPs. However, identifying the "homozygous SNPs" is nontrivial when the regions of CNA are unknown, and a segmentation procedure that simultaneously genotype each SNP while inferring the underlying parental copy numbers is desirable, unless a matched normal is available.

In light of these recent developments, we need a systematic stochastic model for parent specific copy number which can accommodate fractional copy number changes. We propose a general two-chromosome hidden Markov model for this problem. The hidden states of the model represent the copy numbers of each of the two inherited chromosomes, and take value in the continuous space of real numbers. Thus, unlike discrete state space HMMs, this model is not limited to idealized unit-copy changes. Computationally efficient fitting algorithms are given that scale well to data obtained from the current high density genotyping arrays. The estimation procedure based on the two chromosome model, which we call Parent-Specific-Copy-Number (PSCN), extends the framework developed in Lai et al. [37] for total copy number analysis.

After segmenting the genome into regions of constant parent-specific copy number, we identify, for each region, whether both or only one of the parental chromosomes have changed copies. We also determine, in regions containing simultaneous gain of one chromosome and loss of the other, whether the changes are balanced. Thus, we classify the regions into six different types of aberrations depending on the status of the two parental chromosomes: gain of both chromosomes (gain/gain), gain of only one chromosome (gain/normal), gain of one chromosome and balanced loss of the other chromosome (balanced gain/loss), gain of one chromosome and unbalanced loss of the other chromosome (unbalanced gain/loss), loss of only one chromosome (normal/loss) and loss of both chromosomes (loss/loss). To our knowledge, this is the most detailed classification available among methods for allele-specific analysis. The PSCN method outputs the copy number for both chromosomes in each segment.

We evaluate the accuracy of the proposed procedure on a series of simulated tumor titration data provided by Staaf et al. [35], as well as a new set of simulation data containing a larger variety of chromosomal aberrations. We then apply the new approach to 223 glioblastoma samples from the Cancer Genome Atlas project [38], and illustrate through case studies some of the insights gained from an analysis of allele-specific data.

Results

The Two Chromosome Hidden Markov Model

Let $\mathbf{y}=\left\{\mathbf{y}_t=\left(y_t^A, y_t^B\right)^T:\ t=1,\ldots,n\right\}$ be the allele-specific signals for alleles A and B at n SNPs ordered by their locations in a

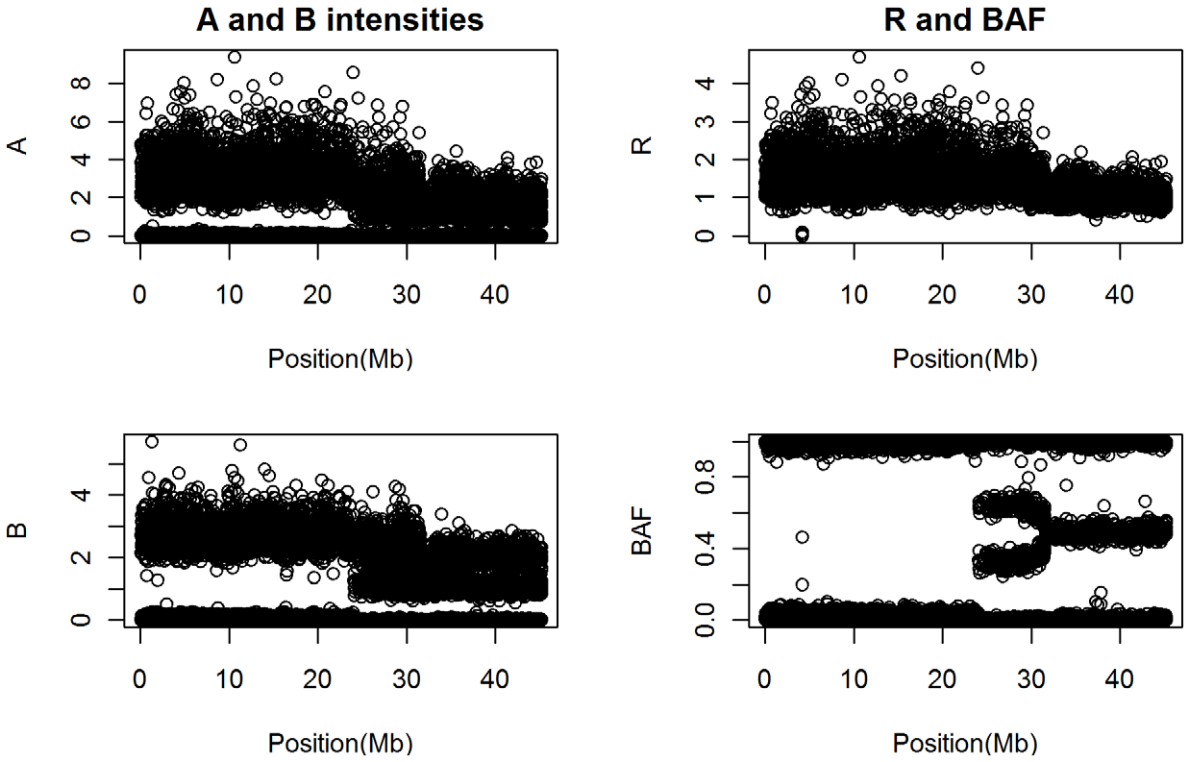

Figure 1. An example data sequence taken from a stretch of a TCGA glioblastoma sample (first 10000 SNPs of TCGA sample 02-0258 chromosome 2) assayed using the Illumina HumanHap 550k SNP array. The left panel shows the A and B allele intensities. The right panel shows the R and BAF. All x-axes are in mega base pairs.
doi:10.1371/journal.pcbi.1001060.g001

reference genome. The way of obtaining $\mathbf{y}$ depends on the experimental platform (see "Data Transformation" in Methods). Our goal is to infer the quantities of the parent specific copy numbers, which we denote by $\boldsymbol{\theta}=\left\{\boldsymbol{\theta}_t=\left(\theta_t^1,\,\theta_t^2\right)^T\colon t=1,\ldots,n\right\}$. By *parent-specific*, we distinguish between the chromosomes inherited from the two parents, which we treat as exchangeable and do not label as maternal or paternal. Let $s_t\in S=\{AA,\,AB,\,BA,\,BB\}$ be the configuration at SNP t specifying the alleles carried by the inherited chromosomes. Let $\mathbf{x}_t=\left(x_t^A,\,x_t^B\right)^T$ be the true copy numbers of alleles A and B at SNP t. The relationship between $\boldsymbol{\theta}_t$, s_t, and $\mathbf{x}_t$ is shown in Table 1.

Note that when a somatic event causes a change in copy number of one or both parental chromosomes at SNP t, the allele-specific copy numbers $\mathbf{x}_t$ change, but s_t remains fixed. For example, if the inherited genotype is AB, and if θ_t^1 is amplified two-fold, then the true copy number of allele A would also be amplified two-fold, but s_t would still be AB. The *observed* allele specific signals $\mathbf{y}_t$ are assumed to be equal to the true allele specific quantities plus an independent measurement error,

$$\mathbf{y}_t=\mathbf{x}_t+\epsilon_t, \tag{1}$$

where $\epsilon_t\sim N(0,\,\Sigma_{s_t})$ and Σ_{s_t} are state specific error covariance matrices. The model that relates $\mathbf{y}_t$ to $\mathbf{x}_t$, $\boldsymbol{\theta}_t$ and s_t is illustrated in Figure 2.

To model the gains and losses of the two inherited chromosomes, we assume that $\boldsymbol{\theta}$ is a Markov jump process with state space $\Re^2$. Conceptually, each time $\boldsymbol{\theta}$ jumps, it can choose between two states: The *normal* state (one copy each of maternal and paternal chromosome), where $\boldsymbol{\theta}$ must assume a known baseline value $\boldsymbol{\mu}_0$, or the *variant* state, where $\boldsymbol{\theta}$ picks a new random value from the bivariate Gaussian $N(\boldsymbol{\mu},V)$. The prior mean $\boldsymbol{\mu}$ and prior covariance V, along with the other hyperparameters of the prior, will be estimated by maximum likelihood. To allow the possibility of the copy number changing from a variant state to a different variant state, for example, $(2,1)^T$ to $(3,1)^T$, we technically need two identically distributed variant states in our formulation of the Markov chain. Hence we let the states be $\{Normal,\,Variant_1,\,Variant_2\}$. Then, the dynamics of the Markov model can be described by the transition matrix

$$P=\begin{pmatrix}1-p & \frac{1}{2}p & \frac{1}{2}p\\ c & a & b\\ c & b & a\end{pmatrix}. \tag{2}$$

The matrix P specifies that if $\boldsymbol{\theta}$ is in the normal state at SNP t, then at SNP $t+1$, $\boldsymbol{\theta}$ stays in the normal state with probability $1-p$, or jumps

Table 1. Relationship between the inherited allele configuration s_t and the true allele specific copy numbers x_t.

s_t	x_t^A	x_t^B
AA	$\theta_t^1+\theta_t^2$	0
AB	θ_t^1	θ_t^2
BA	θ_t^2	θ_t^1
BB	0	$\theta_t^1+\theta_t^2$

doi:10.1371/journal.pcbi.1001060.t001

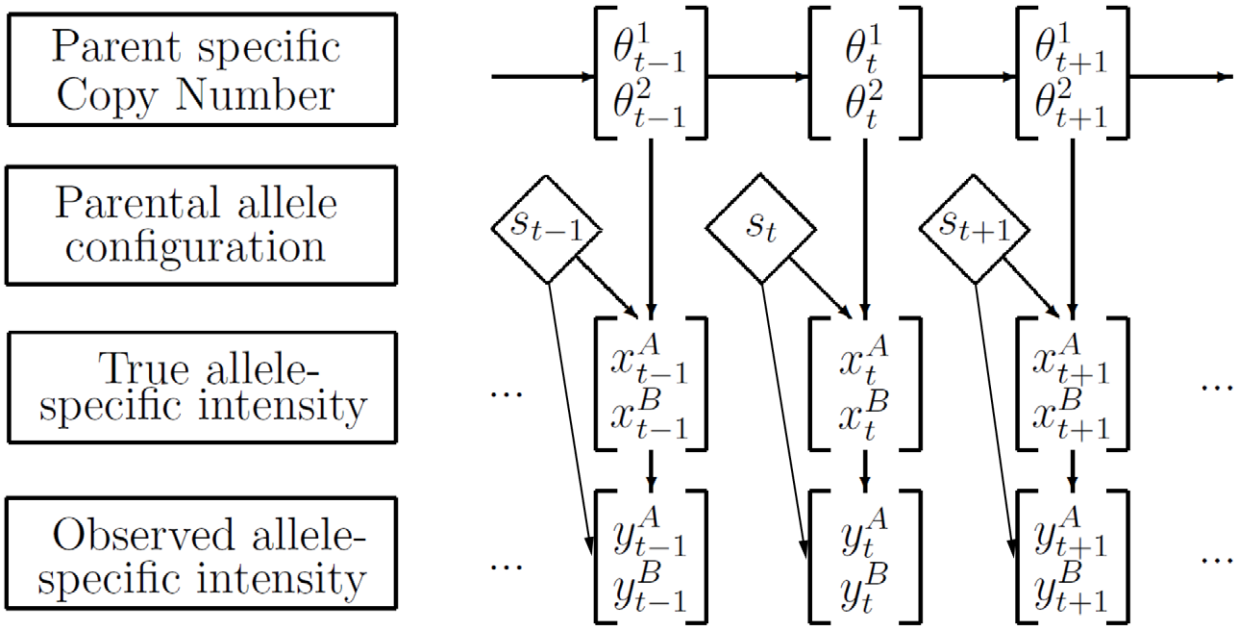

Figure 2. Overview of the stochastic segmentation model. The Markov sequence $\{\mathbf{q}_t\}$ represent the parent-specific copy number, i.e. the underlying copy numbers of the two inherited chromosomes. For each SNP t, the *allele*-specific copy numbers $\mathbf{x}_t$ depend on both $\mathbf{q}_t$ and the inherited allele configuration s_t. The observed allele-specific signals, $\{\mathbf{y}_t\}$, are $\{\mathbf{x}_t\}$ overlayed with Gaussian noise. s_t affects $\{\mathbf{y}_t\}$ in the way that different type of s_t can have different covariance structure for the Gaussion noise.
doi:10.1371/journal.pcbi.1001060.g002

to a variant state with probability p. If $\boldsymbol{\theta}_t$ is in a variant state, then at SNP $t+1$, it would stay at the *same* variant state with probability a, or jump to a *different* variant state with probability b, or jump back to the normal state with probability c. One can verify that this formulation of the Markov chain, with one baseline state and two variant states, allows for a model with a baseline state and generic "variant" states as desired. This model extends the one used for the analysis of total copy number in Lai et al. [37]. This Markov chain has the stationary distribution $\left(c/(p+c), \frac{1}{2}p\Big/(p+c), \frac{1}{2}p\Big/(p+c)\right)$. The three-state Markov chain with transition probability matrix $\boldsymbol{P}$ and initialized at the stationary distribution is reversible, which provides substantial simplification for the estimation of $\boldsymbol{\theta}$. Practically, the reversibility of the Markov model implies that we would obtain the same segmentation going from right to left as we do going from left to right. Biologically, this seems logical, as there is no known directionality of copy number aberration events.

We assume that the inherited allele configurations s_t are independent multinomial with prior parameters

$$\left(p_t^{AA}, p_t^{AB}, p_t^{BA}, p_t^{BB}\right),$$

which can be obtained from the genotyping data of a set of normal control samples. Note that AB and BA cannot be distinguished in normal samples, so we can set p_t^{AB} and p_t^{BA} to one-half of the proportion of heterozygotes for SNP t. When these figures are not available, we have found that a uniform prior usually works reasonably well. This is because the main purpose of the model is to estimate the parent-specific copy numbers, with s_t as surrogate information. With the large number of data points obtained from the high density arrays, the posterior for the parent-specific copy numbers is usually quite insensitive to the prior on s_t. Note that for platforms, such as the Affymetrix 6.0 array, have non-polymorphic copy number markers rather than SNP markers. For those markers, the prior for s_t can be set to (1, 0, 0, 0). In this way, the posterior will always remain at (1, 0, 0, 0) and only the total copy number information at these markers would contribute to the overall segmentation.

Note that this model contains many assumptions, including Gaussianity of the allele specific intensities and Markovicity of the underlying copy number states. These assumptions allow fast and explicit analytic formulas to be derived, thus avoiding the need for Monte Carlo based estimates. For most platforms, the allele-specific intensities deviate from Gaussianity, despite careful normalization. Also, there has never been proof that chromosomal breakages are Markovian. These assumptions are made for modeling convenience, just as in the total-copy number estimation problem [11,16,30,37]. It is reassuring that the estimation method is robust to deviations from both the Gaussian and Markov assumptions, as we show using the titration data from Staaf et al. [35] and through our own spike-in studies.

Our primary objective is to estimate the parent specific copy numbers $\boldsymbol{\theta}$, which depend on the observed signals through the unobserved inherited allele configurations $\boldsymbol{s}=\{s_t : t=1,\ldots,n\}$. Let $\mathcal{S}=S^n$ and $\Theta=(\Re^2)^n$ be the set of all possible realizations for $\boldsymbol{s}$ and $\boldsymbol{\theta}$, respectively. We describe below an iterative algorithm to estimate s and $\boldsymbol{\theta}$.

Allele-specific iterative smoothing. Fix stopping threshold δ. Initialize $i=0$ and $\boldsymbol{s}=\boldsymbol{s}^{(0)}=\left(s_1^{(0)},\ldots,s_n^{(0)}\right)$ through an initial 4-group clustering of $\{\mathbf{y}_t : t=1,\ldots,n\}$. Repeat:

1. Expectation step: Given $\boldsymbol{s}^{(i)}$, set $\boldsymbol{\theta}^{(i+1)}$ to its posterior mean

$$\boldsymbol{\theta}_t^{(i+1)}=E[\boldsymbol{\theta}_t|\boldsymbol{s}^{(i)},\mathbf{y}],\quad t=1,\ldots,n. \tag{3}$$

Computationally efficient formulas for (3) are given in Methods.

2. Maximization step: Given $\boldsymbol{\theta}^{(i+1)}$, set $\boldsymbol{s}^{(i+1)}$ to its maximum a posterior value

$$\boldsymbol{s}^{(i+1)} = \text{argmax}_{\boldsymbol{s}\in\mathcal{S}} P(\boldsymbol{s}|\boldsymbol{\theta}^{(i+1)}, \mathbf{y}). \quad (4)$$

This can be done easily because given $\boldsymbol{\theta}^{(i+1)}$, $\mathbf{y}_t$ is a four-component mixture of Gaussians at each t, and $s_t^{(i+1)}$ is simply the identifier for each mixture component. The exact formula for (4) is given in Methods.

3. If $\left\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}\right\| < \delta$, stop and report $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i+1)}$, $\hat{\boldsymbol{s}} = \boldsymbol{s}^{(i+1)}$. Otherwise, set $i \leftarrow i+1$ and go back to step 1.

In each iteration of the above algorithm, the expectation step estimates $\boldsymbol{\theta}_t$ by its posterior mean given the data and the current estimate of the configuration states s_t. Then, s_t is set to its posterior mode given the data and the current estimate of $\boldsymbol{\theta}_t$. Computationally efficient forward-backward equations for (3) and formulas for (4) are given in Methods, where we also describe an expectation maximization procedure for estimating the hyper-parameters P, $\boldsymbol{\mu}$, V, and $\{\Sigma_j : j = AA, AB, BA, BB\}$ from the data, so that they do not need to be specified a priori.

The above algorithm returns a soft segmentation of $\mathbf{y}$ in the form of a Bayesian estimate $\hat{\boldsymbol{\theta}}$ for the parent specific copy numbers at each location. A hard segmentation is sometimes desirable, for example, to give a sparse representation of the data. A hard segmentation can be obtained from the soft segmentation as follows: Compute for each t the one-step Euclidean distance $\Delta_t = \left\|\hat{\boldsymbol{\theta}}_{t+1} - \hat{\boldsymbol{\theta}}_t\right\|$. Estimate the change-points to be the locations where Δ_t are larger than the threshold, with the constraint that they must be separated by a pre-chosen minimum number of SNPs (e.g. 20). The segmentation algorithm starts with the set $\hat{\tau} = \{0, n\}$ containing only the end points of the sequence. Change-points are added recursively to the set by maximizing Δ_t under the separation constraint, until no more change-point can be added. We start with a low threshold for $\Delta_t(0.01)$ allowing some false positives, with most of the false positives eliminated by a subsequent Wilcoxon Rank-Sum test (p-value threshold of 0.05) that combines adjacent segments with no significant difference in mean. We found this to be more accurate than a one-step procedure using a more stringent threshold on Δ_t.

Identifying the Type of Aberration

The segmentation divides the genome into regions where the copy numbers of the two inherited chromosomes are constant. It is often useful to know, for each region, whether the copy numbers of one or both parental chromosomes deviate from the normal level. This involves classifying each region into one of the following six types of chromosomal change: gain/gain, gain/normal, balanced gain/loss, unbalanced gain/loss, normal/loss and loss/loss.

For each segmented region, we define the major copy number to be the normalized raw copy number of the more abundant chromosome, and the minor copy number to be the normalized raw copy number of the less abundant chromosome. If the two chromosomes have equal copy numbers, then the major and minor chromosome labels are assigned arbitrarily. The major and minor copy numbers are estimated after the hard-segmentation using a mixture model on the heterozygous SNPs in each region (which can be identified using $\hat{s}$). Then, a t-test is used to compare the estimated major and minor copy numbers of each region to the estimated allele copy number of the normal level in the unchanged segments. The Bonferroni correction is used to adjust for multiple testing. The technical details are given in Methods. This procedure allows us to discover and distinguish all of the six types of CNVs.

An additional caveat is that when both parental chromosomes carry the same haplotype, a balanced gain/loss would be called if the region were long enough. Without matched data from normal tissue, it is impossible to distinguish with certainty between inherited and somatic LOH. However, we rely on the fact that long regions of LOH are infrequent, and thus the minor allele frequency of SNPs and the linkage disequilibrium between them can be used to conduct a test for the probability that an inherited LOH appears by chance. This haplotype correction only takes care of the unique common haplotypes, i.e., when a region is dominated by one haplotype. If a haplotype is not common in that region, or if there are several haplotypes in that region, this test loses sensitivity. In this case, paired normal cell information would be useful. More details are given in Methods.

Results on Simulated Dilution Data from Staaf et al. [35]

Staaf et al. [35] performed a systematic comparison of existing methods for allele-specific copy number estimation. They created a simulated dilution data set based on experimental 550k Illumina data for HapMap sample NA06991. To the diploid HapMap sample, ten regions of aberrant copy number were added at increasing fractions to mimic a tumor sample that is contaminated with normal cells. Here, u% normal cell contamination means u part normal cells are mixed with $100-u$ part tumor cells. The aberrant regions vary by type and length, and represent regions of hemizygous gains and losses and copy neutral LOH. Since the locations of the true aberrant regions are known, the specificity and sensitivity of the detection methods can be evaluated.

We applied PSCN, the R package we developed based on our method, to this dilution data set and compared it with existing approaches in an analysis that parallels the insightful analysis in Staaf et al. [35]. The sensitivity and specificity of results from PSCN at varying contamination ratios is shown in Figures 3 and 4 overlayed onto plots reproduced from Staaf et al. [35]. In order to compare with the sensitivity analysis of other models done in the paper by Staaf et al. [35], we define a "correct detection" to mean that a true CNA region has been called, but do not require that the type of CNA (e.g. gain/loss, normal/loss) has been correctly identified. All the other current procedures only categorize the CNAs into Gain, Loss and LOH, which are the three types of CNAs used in the Dilution data in Staaf et al. [35]. We assess the accuracy of PSCN in a more detailed classification of identified CNAs based on the six types of chromosomal change in a separate data set that contains a wider diversity of chromosomal events (see next section). In the simulated dilution data, the regions vary in length, magnitude, and type of aberration, with some regions harder to detect than the others. There is a separate sensitivity plot for each of the 10 aberrant regions created by [35]. As expected, for all regions, sensitivity is maintained at a high level up to a certain contamination ratio, then drops sharply. Since Staaf et al. and we used very stringent detection thresholds, the specificity is maintained near 1 for all contamination ratios, as shown in Figure 4. The sensitivity of PSCN is comparable to SOMATICs [36], but the latter method has much lower specificity, as shown in the analysis of Staaf et al., see Figure 4. PSCN achieves good accuracy compared to the other existing methods, especially methods based on discrete-state hidden Markov models for high levels of contamination. The discrepancy between the two specificity plots in Figure 4 are due to the fact that when an aberration is called, it may be labeled as an incorrect type (for example, a copy neutral LOH may be labeled as single copy gain). When the correct calling of aberration type is required, the specificity of PSCN is maintained through a higher level of contamination as compared to existing models. The new model

can identify the correct aberration type if the normal cell contamination is below 80%. Above 80%, PSCN gains significantly in sensitivity compared to existing methods but also sacrifices slightly in specificity.

Accuracy of Aberration Type Identification

The dilution data set from Staaf et al. [35] contains only three types of aberrations: hemizygous loss (normal/loss), single copy gain (gain/normal), and copy neutral LOH (balanced gain/loss). We created a simulated data set containing all six types of aberrations: gain/gain, gain/normal, balanced gain/loss, unbalanced gain/loss, normal/loss and loss/loss. To make the simulation resemble real data, we started with the 550k Illumina data for chromosome 1 of HapMap sample NA06991. To this normal sequence we imposed six different signal types on six regions. The positions and magnitudes of the added signals are shown in Table 2. The top panel of Figure 5 (first row) shows the R and BAF before the signals are imposed. The middle and bottom panels show the R and BAF after the signals have been imposed, at 0% and 80% contamination respectively, with true signals indicated by black lines. Signal becomes weaker when normal cell contamination increases, and thus are harder to detect. The estimated parent-specific copy numbers are shown in Figure 6. We can see from the plots that the estimated parent-specific copy numbers are very close to the true allele copy numbers. Table 3 shows the largest normal cell contamination under which the

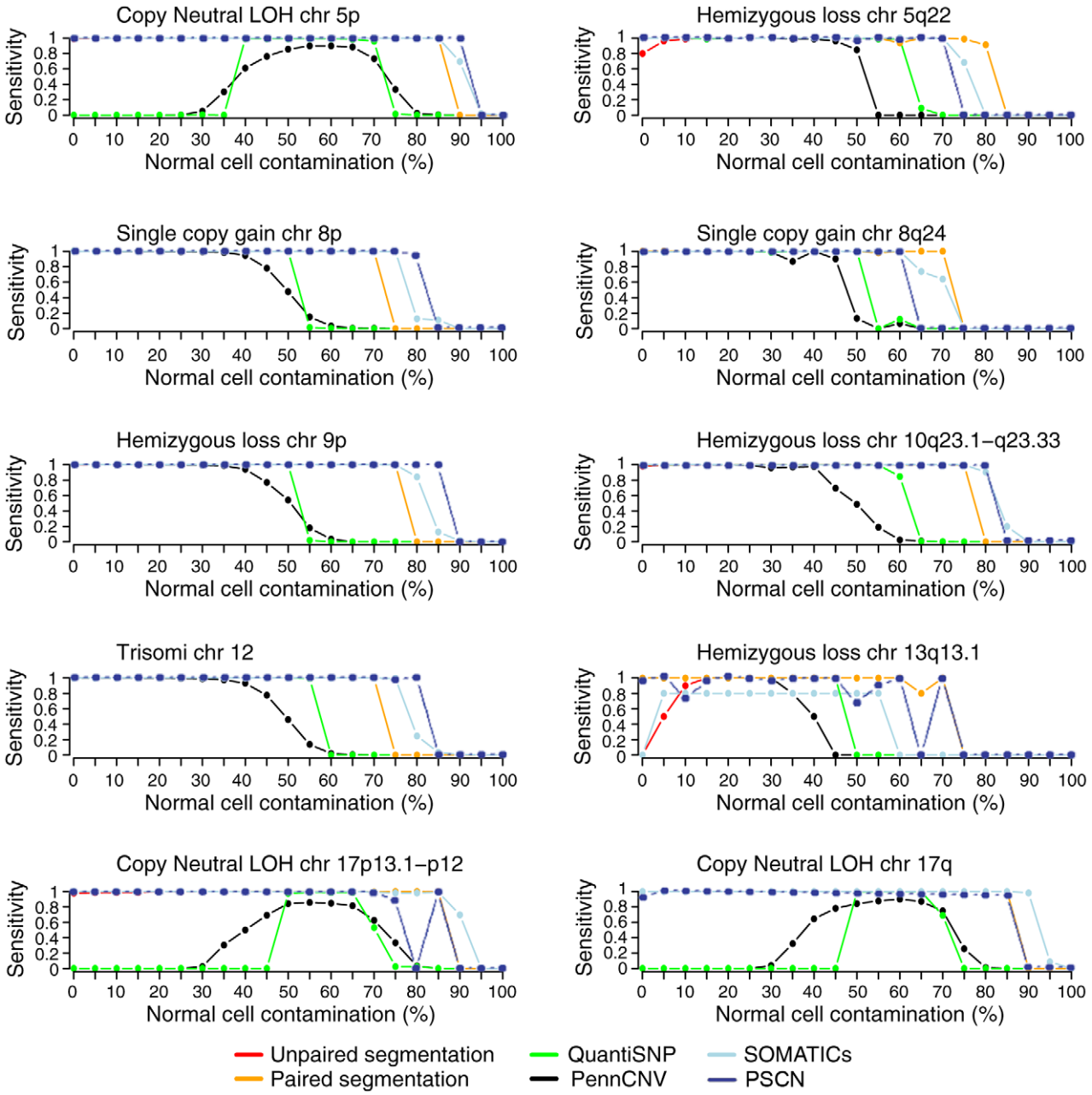

Figure 3. Sensitivity versus normal cell contamination for 10 regions in the dilution data set of Staaf et al. [35]. We overlayed our results on top of plots reproduced from [35].
doi:10.1371/journal.pcbi.1001060.g003

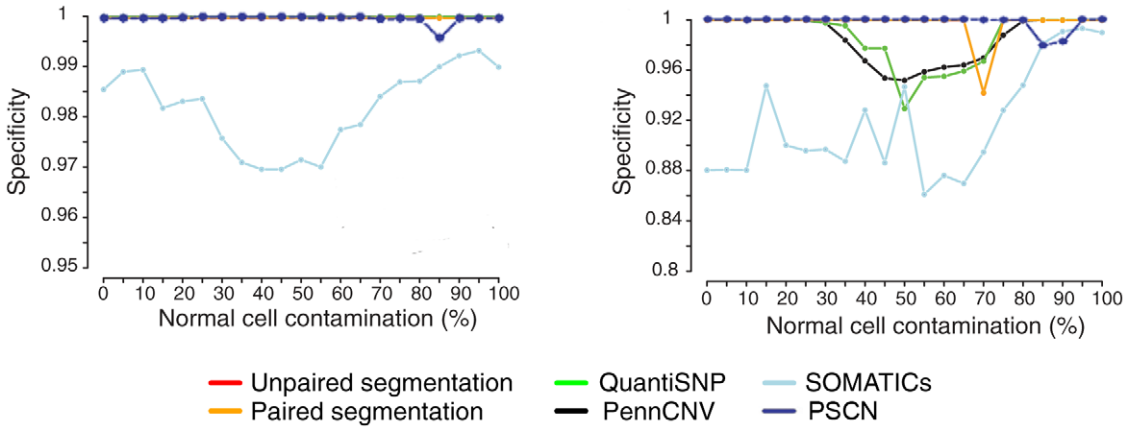

Figure 4. Specificity versus normal cell contamination in the dilution data set of Staaf et al. [35]. We overlayed our results on top of plots reproduced from [35]. Left panel shows the overall specificity, which is the fraction of SNPs outside of all simulated alletic imbalances that are not called. The right panel shows the specificity of correct calling of the type of allelic imbalance, i.e., if a truly aberrant region is identified as aberrant, but of an incorrect type, then it is also considered as a wrong call. Note that the scale of y-axis is different for the two plots.
doi:10.1371/journal.pcbi.1001060.g004

signals are detectable by PSCN. When normal cell contamination is less than 80%, our model can detect most of the signals with both alleles assigned to the correct type. When the normal cell contamination rises to 90%, our model can still detect three out of the six CNA regions, but assigns the correct type to only one of the two alleles. For example, at a high contamination level of 90%, there is a tendency for a fractional loss of both chromosomes to be mistaken for a fractional loss of only one of the two chromosomes. From this study, we see that the correct type of aberration can be identified robustly for all but the highest levels of normal cell contamination.

Accuracy of Estimation of Genotype States

Using the dilution data set created from HapMap sample NA06991, we can also assess the accuracy of PSCN in identifying the genotype states $\{s_t\}$. Since the genotypes for the SNPs on this sample are known, we simply compared the estimated $\{\hat{s}_t\}$ with the true values.

Table 4 shows the percent of homozygous SNPs that are misclassified as heterozygous, and vice versa. When the SNP is classified as homozygous, the determination between the states AA and BB is trivial, and no errors are made. When normal cell contamination is extremely low, less than 10%, genotyping errors are common in regions of loss of heterozygosity (either normal/loss or gain/loss). This is expected, since in a region with complete LOH and zero contamination, only one of the two parental alleles is left, and thus it would be impossible to distinguish between the homozygous configurations $\{AA, BB\}$ and the heteryzogous configurations $\{AB, BA\}$. Fortunately, these types of genotyping errors would not affect the accurate estimation of $\boldsymbol{\theta}_t$, since the mean levels for the heterozygous and homozygous tracks merge for LOH regions under zero contamination. It is slightly unintuitive that the correct estimation of s_t depends on the fact that there is normal cell contamination! This is reflected in Table 4, where accuracy quickly improves as normal cell contamination increases, with a total misclassification rate of .54% at 10% normal cell contamination.

A complete analysis of the misclassification rates of $\{s_t\}$ are given in the Supporting Information file (Text S1).

Analysis of TCGA Glioblastoma Samples

We applied PSCN to 223 glioblastoma samples from the TCGA project [38]. These samples were assayed using Illumina HumanHap 550k SNP arrays.

Almost all of the 223 samples analyzed contain substantial copy number aberrations. Table 5 shows the distribution of the types of copy number events found in the samples. Of the gain/loss events, which comprise 45.4% of all of the events, 22.8% are copy neutral LOH and 22.5% are unbalanced gain/loss. We see from this table that, among these glioblastoma samples, single chromosome losses

Table 2. Signals imposed on to Chromosome 1.

	SNP begin	SNP end	Major copy number	Minor copy number
Gain/Gain	2000	5000	3	2
Gain/Normal	9000	12000	2	1
Balanced Gain/Loss	16000	19000	2	0
Unbalanced Gain/Loss	23000	26000	3	0
Normal/Loss	30000	33000	1	0
Loss/Loss	37000	40000	0	0

"SNP begin" and "SNP end" are the indices of the SNP where the added signal begins and ends, respectively. "Major" and "minor" copy numbers are the intensities of the signal in the two alleles.
doi:10.1371/journal.pcbi.1001060.t002

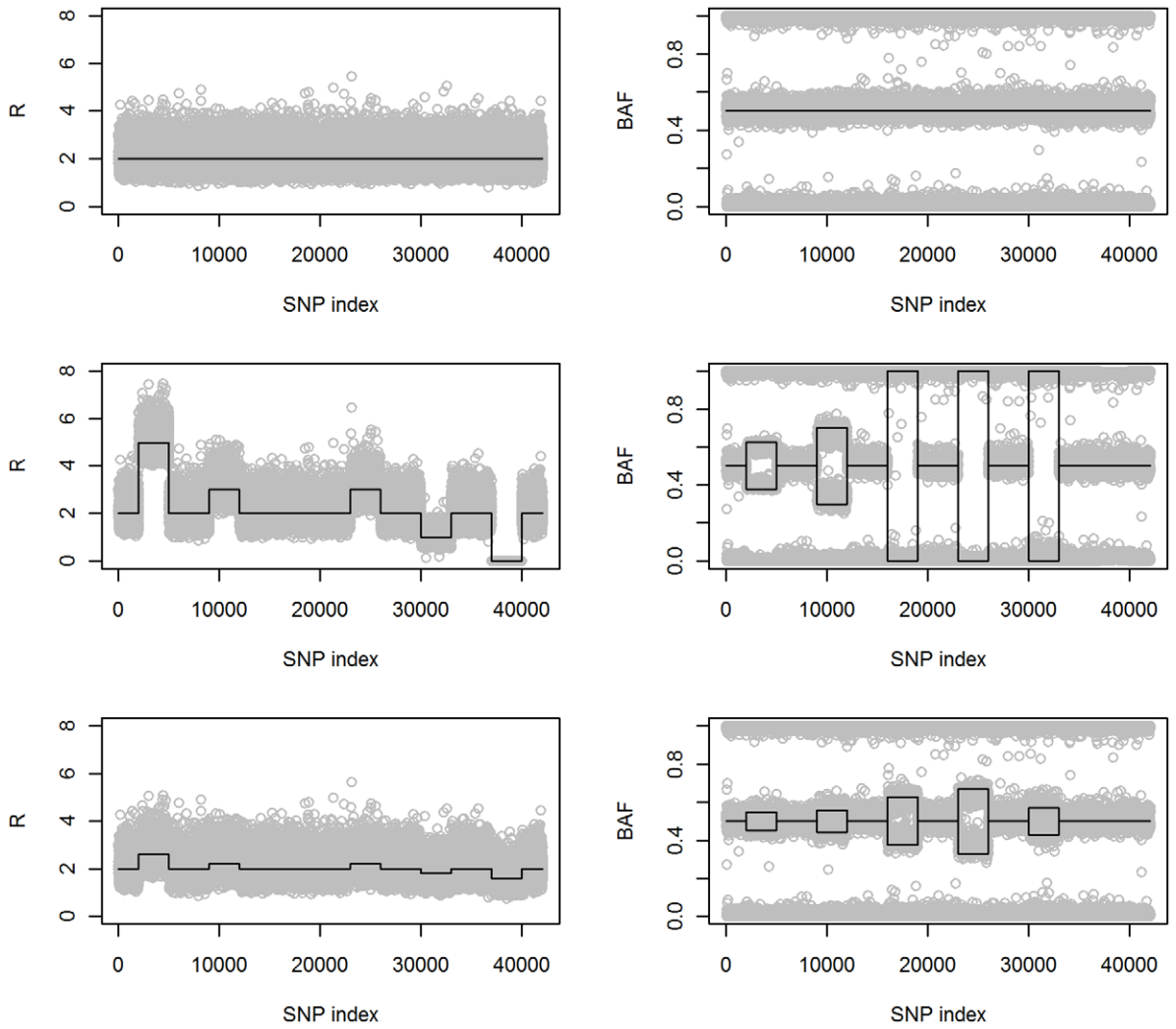

Figure 5. Signal of the simulated data by imposing six types of aberrations on chromosome 1 of HapMap sample NA06991. The first row shows *R* and BAF before the signals are imposed. The second row shows *R* and BAF after the signals are imposed, under normal cell contamination 0%. True signals are indicated by black lines. The third row shows *R* and BAF after the signals are imposed, under 80% normal cell contamination.
doi:10.1371/journal.pcbi.1001060.g005

or single chromosome gains comprise 49.6% of all the events, which means that more than half of the events involve change of both inherited chromosomes.

We now zoom in on two example regions to illustrate the additional insights gained from parent-specific copy number analysis. These regions are shown in Figure 7. The figures in the left panel correspond to the entire chromosome 3 of TCGA glioblastoma sample 02-0332, while those on the right panel correspond to the first 10000 SNPs on chromosome 2 of TCGA glioblastoma sample 02-0258. The top two plots in each panel show the *R* and BAF values. The color scheme for these plots show the segmentation obtained using PSCN. We transformed the *R* and BAF values back to the (*A*, *B*) raw copy number values, and fitted two dimensional densities separately to each region in the segmentation. The contours of the two dimensional density estimates, delineating the locations of the clusters, are shown in the third plot from the top in each panel. The color scheme for the contours is the same as the color scheme for the *R* and BAF plots. Finally, the bottom plot of each panel shows the estimated major and minor copy numbers for each region (we will call this type of plot the *mm*-plot). The color scheme of the *mm*-plot reflects the gain/loss status of each region, where red represents gain, blue represents loss, and green represents normal. It is usually difficult to discern the relative magnitudes of gains and losses from the *R* and BAF plots, especially when both inherited chromosomes have undergone copy number changes. Such relative changes in parent specific copy numbers can be quantified more easily by examining the (*A*, *B*) contour and *mm*-plots.

Copy neutral LOH (Balanced Gain/Loss). First, consider the example region from TCGA sample 02-0332 on the left panel. There are three instances of copy neutral LOH, colored in purple. Based on the BAF plot, the loss seems to be complete, that is, it is carried by almost all of the cells in the sample. The *mm*-plot also gives this information, as the estimated major copy number (red line) is close to 2, and the estimated minor copy number (blue line) is close to 0. These LOH regions do not change the total copy number, and thus would not have been detected if the segmentation were based on the *R* profile. On the other hand, an analysis based only on the BAF plot would not have revealed that the LOH is copy neutral; e.g. in the TCGA sample 02-0258,

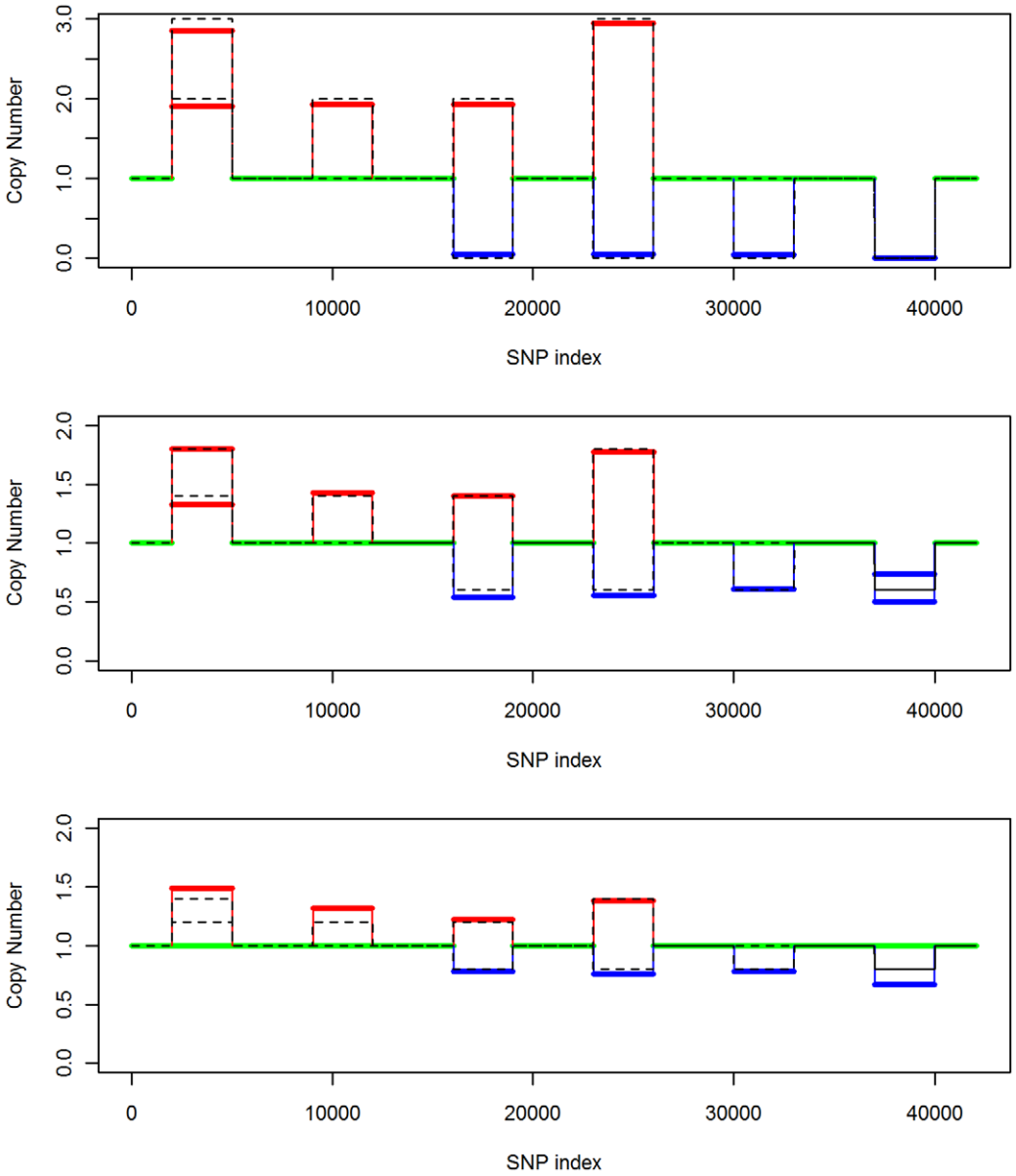

Figure 6. Copy number estimation of PSCN on the simulated data by imposing six types of aberrations on chromosome 1 of HapMap sample NA06991. Top panel: no normal cell contamination. Middle panel: normal cell contamination 60%. Bottom panel: normal cell contamination 80%. In all panels, solid lines denote estimated allele copy numbers and dashed lines denote true copy numbers.
doi:10.1371/journal.pcbi.1001060.g006

the LOH region (purple) with similar pattern in BAF is not copy neutral. The estimates in the *mm*-plot can only be obtained through a joint analysis of both the R and the BAF profiles.

Fractional single chromosome gains and losses. Following the copy neutral LOH regions in chromosome 3 of sample 02-0332, there is a stretch of alternating gains and losses, colored respectively in red and blue. The copy of the other parental chromosome in these regions is one. As seen from the *mm*-plot, all of these regions contain changes that affect only one of the two inherited chromosomes. The changed chromosome may differ across segments. For example, the paternal chromosome may have been differed in one segment, and the maternal chromosome in the next. The copy number of the other chromosome in these regions remain at the normal level. This fact can not be deduced from total copy number analysis, as an increase in R can be due to gains of both inherited chromosomes, or an unbalanced gain of one chromosome and loss of the other; see the next example (TCGA 02-0258). The (A, B) contour plot discriminates between these two possible cases. If we examine the cluster centers corresponding to the heterozygotes in the red and blue segments we see that for any one cluster, only one of the A and B coordinates is significantly shifted from the corresponding coordinate of the normal AB cluster (coded in gray). This is evidence that the copy number of only one of the chromosomes has changed in these regions. The positions of the heterozygote cluster centers of the red and blue regions indicate only a partial gain and loss, as their shifts from normal are only a fraction of that expected in a complete event. The estimated major and minor copy numbers in the *mm*-plot quantifies the partial change explicitly, with the major copy numbers at around 1.5 for the gain and the minor copy numbers at around 0.5 for the loss. Assuming a

Table 3. The largest tolerable percentage for normal cell contamination under which the type of aberration can be correctly detected (left column), and under which the type of aberration can be correctly identified for one of the two alleles when both alleles are different from normal (eg. Gain/Gain identified as Gain/Normal) (right column).

	Correct Type Estimated for both alleles	Correct Type Estimated for one allele
Gain/Gain	70	90
Gain/Normal	85	Not applicable.
Balanced Gain/Loss	80	90
Unbalanced Gain/Loss	85	90
Normal/Loss	85	Not applicable.
Loss/Loss	65	80

All numbers are in percent.
doi:10.1371/journal.pcbi.1001060.t003

linear signal response curve for the Illumina platform in the range between 0 and 3 fold change in DNA quantity, this translates to about 50% of the cells in the tumor sample carrying the aberrations coded in blue and red.

The same reasoning can be applied to the red and pink regions of chromosome 2 of TCGA sample 02-0258 (right panel), which contains a fractional gain. By teasing apart the copy numbers of each inherited chromosome, we are now able to characterize and quantify these fractional changes.

Simultaneous unbalanced gain and loss of both chromosomes (unbalanced gain/loss). Now consider the example region color coded in purple from TCGA sample 02-0258 in the right panel. The *R* plot suggests that there is a gain in total copy number. However, the BAF plot reveals that there seems also to be an almost complete loss of heterozygosity in this region. Loss of one of the inherited chromosomes is necessary for loss of heterozygosity. Thus we conclude that the region colored in purple contains both a gain of one as well as an almost complete loss of the other inherited chromosome. Indeed, as the *mm*-plot shows, the estimated major and minor copy number fold changes for this region have values of 3 and 0, respectively. The gain and loss of the two inherited chromosomes is thus unbalanced, suggesting that this region may have experienced multiple mutations. This region is immediately followed by a gain of only one of the two inherited chromosomes (see the *mm*-plot), of magnitude roughly equal to the difference between the deviations of the major and minor copy numbers from normal. This suggests the hypothesis that this sample first experienced a gain of one of the inherited chromosomes that covered the purple and red regions, then a LOH which caused a gain of the already amplified chromosome and a simultaneous loss of the other inherited chromosome. Our analysis of the TCGA data shows that these types of unbalanced gain and loss events are quite common.

Discussion

We have developed a method for simultaneous estimation of parent-specific DNA copy number and inherited genotypes for tumor samples using allele-specific raw copy number data. The model and estimation procedure start with transforming allele-specific data into A and B intensities, which may vary across experimental platforms. The model assumes that the A and B allele intensities should be roughly symmetric, roughly variance stabilized and have approximately bivariate Gaussian errors. Indeed, the model is quite robust to the violation of the bivariate Gaussian error assumption. The model gives satisfying results even if this assumption is heavily violated. More details are shown in the the Supporting Information file (Text S1). We illustrated the method and evaluated its performance on both published and newly generated dilution data sets on the Illumina platform.

A rigorous assessment using in silico titration data provided by Staaf et al. [35] shows that PSCN has good accuracy. The proposed method does not require paired normal samples. However, if such samples were available, then they can be used to further improve accuracy and to distinguish between inherited LOH and somatic LOH. In such cases, s_t can simply be set to the genotypes inferred from the normal samples.

PSCN is not platform specific, and we have also applied it to data from the Affymetrix Genotyping 6.0 array, with an example analysis given in the Supporting Information file (Text S1). The

Table 4. The number of misclassifications of each type in the identification of s_t on the NA06991 dilution data set, at different levels of normal cell contamination.

Normal Contamination (%)	Homozygous → Heterozygous	Heterozygous → Homozygous	Misclassification Rate (%)
0	1285	2791	9.7
5	986	1	2.3
10	228	0	.54
25	20	0	.048
50	93	0	.22
90	39	0	.093

There are 42037 SNPs total.
doi:10.1371/journal.pcbi.1001060.t004

Table 5. Distribution of types of copy number aberrations across all events found in the 223 glioblastoma samples.

Event type	%	count
gain/gain	3.6	1315
gain/normal	21.0	7773
balanced gain/loss	22.9	8568
unbalanced gain/loss	22.5	8352
normal/loss	28.6	10598
loss/loss	1.4	521

doi:10.1371/journal.pcbi.1001060.t005

segmentation accuracy of PSCN seems to be reasonable for Affymetrix data, but can potentially be improved significantly by better probe-level normalization. This is due to the fact that the BAF of Affymetrix data is much noisier than the BAF of Illumina data, which makes the estimation of $\{s_t\}$ much more difficult. Bengtsson et al. [39] have shown that much of the variation in the BAF of Affymetrix data are due to probe-specific effects that can be removed if a matched normal sample is available. Another promising method for probe-level normalization of Affymetrix data is the probe raw copy number composite representation (PICR) model of Wan et al. [29], which uses probe sequence information and physico-chemical modeling to estimate binding affinity. However, since the PICR model relies on mismatch probes, it is only applicable to Affymetrix platforms prior to the 6.0 array. Thus, better probe-level normalization of Affymetrix 6.0 data for unmatched samples is still an important problem for further investigation.

An overview of an analysis of the TCGA glioblastoma samples reveal that a substantial fraction of copy number changes are copy-neutral loss of heterozygosity events. These events would not have been found using analyses based only on total copy number. Cases of unbalanced simultaneous changes in the copy numbers of both inherited chromosomes were also found. It would be of interest to quantify the frequency of such changes among different cancer subtypes and in other types of tumors.

A final point that we would like to emphasize is the quantification of fractional changes, as exemplified by the two case studies on the TCGA glioblastoma samples. Since this requires teasing apart the quantities of the two inherited chromosomes, it can only be achieved through allele-specific estimates. The fraction of cells that carry each copy number event is important for downstream analyses, such as quantifying normal cell contamination and studying tumor microevolution. The parent-specific copy number estimates obtained from the proposed method provides a starting point for these types of investigations.

The R package for PSCN is registered on R-Forge (http://r-forge.r-project.org/) under project name PSCN.

Methods

Data Transformation

The proposed model is not platform specific, and can theoretically be applied to any type of allele-specific copy number data where the errors on the raw copy number values of the alleles can be normalized to approximately adhere to a bi-variate Gaussian distribution. As we show below, the Gaussian error assumption allows for explicit analytic formulas for the posterior mean of the underlying inherited chromosome copy numbers, thus bypassing the need for computationally intensive Monte Carlo methods. For most platforms, the raw allele-specific raw copy number values must be properly normalized for this error model to be a good approximation. However, as we mentioned in the Discussion section, the model is quite robust to the violation of the Gaussian error assumption.

A unified approach that gives satisfying results for data from both Illumina and Affymetrix platforms is as follows. Since

$$R_t = A_t + B_t, \quad \mathrm{BAF}_t = B_t/(A_t + B_t)$$

we have

$$A_t = R_t(1-\mathrm{BAF}_t), \quad B_t = R_t\mathrm{BAF}_t$$

Note that the "BAF" given by the Illumina platform [6] is not the intuitive quantity $(B/(A+B))$, but the arc-tangent of the ratio of B raw copy number versus A raw copy number scaled to [0,1]. Use BAF^* to denote the so called BAF given by Illumina, then

$$A_t = R_t/\left[1+\tan(\mathrm{BAF}_t^*\pi/2)\right], \quad B_t = R_t - A_t.$$

For PSCN we use $y_A^t = A_t$, $y_B^t = B_t$.

Explicit formulas for θ given $\mathbf{y}$ and s

We give here exact formulas for the conditional expectation (3). Let δ_z denote the probability distribution that assigns probability 1 to the value z. Denote by $\mathcal{Y}_{i,j} = (\mathbf{y}_i, \ldots, \mathbf{y}_j)$, and $\mathcal{S}_{i,j} = (s_i, \ldots, s_j)$. A brief outline of the estimation procedure is as follows: First, conditioned on all data to the left of t, $\boldsymbol{\theta}_t$ is distributed as a mixture of Gaussians:

$$\boldsymbol{\theta}_t|(\mathcal{Y}_{1,t}, \mathcal{S}_{1,t}) \sim p_t\delta_{\boldsymbol{\mu}_0} + \sum_{i=1}^{t} q_{i,t}N(\boldsymbol{\mu}_{i,t}, V_{i,t}), \tag{5}$$

where the formulas for computing the parameters of the mixture p_t, $q_{i,t}$, $\boldsymbol{\mu}_{i,t}$, and $V_{i,t}$ are given below. We call (5) the forward filter. Since by our model $\{\boldsymbol{\theta}_t\}$ is a reversible Markov chain, we can reverse time and obtain a backward filter that is analogous to (5):

$$\boldsymbol{\theta}_{t+1}|(\mathcal{Y}_{t+1,n}, \mathcal{S}_{t+1,n}) \sim \tilde{p}_{t+1}\delta_{\boldsymbol{\mu}_0} + \sum_{j=t+1}^{n} \tilde{q}_{j,t+1}N(\boldsymbol{\mu}_{t+1,j}, V_{t+1,j}), \tag{6}$$

where the parameters $\tilde{p}_{t+1}$, $\tilde{q}_{j,t+1}$, $\boldsymbol{\mu}_{t+1,j}$, and $V_{t+1,j}$, as for the forward filter, are given in explicitly computable form below. The Bayes theorem can then be used to combine the forward filter (5) and backward filter (6) to derive the posterior distribution of $\boldsymbol{\theta}_t$ given the complete sequence $\mathcal{Y}_{1,n}$, which is a mixture of normal distributions

$$\boldsymbol{\theta}_t|(\mathcal{Y}_{1,n}, \mathcal{S}_{1,n}) \sim \alpha_t\delta_{\boldsymbol{\mu}_0} + \sum_{1\le i\le t\le j\le n} \beta_{i,j,t}N(\boldsymbol{\mu}_{i,j}, V_{i,j}) \tag{7}$$

whose parameters can be derived from the forward and backward filters as described below. This forward-backward procedure can be reduced to $O(n)$ computation time by the BCMIX algorithm [40]. From (7), it follows that the conditional expectation in Equation (3) can be computed as

$$E(\boldsymbol{\theta}_t|\mathcal{Y}_{1,n}, \mathcal{S}_{1,n}) = \alpha_t\delta_{\boldsymbol{\mu}_0} + \sum_{1\le i\le t\le j\le n} \beta_{i,j,t}\boldsymbol{\mu}_{i,j}. \tag{8}$$

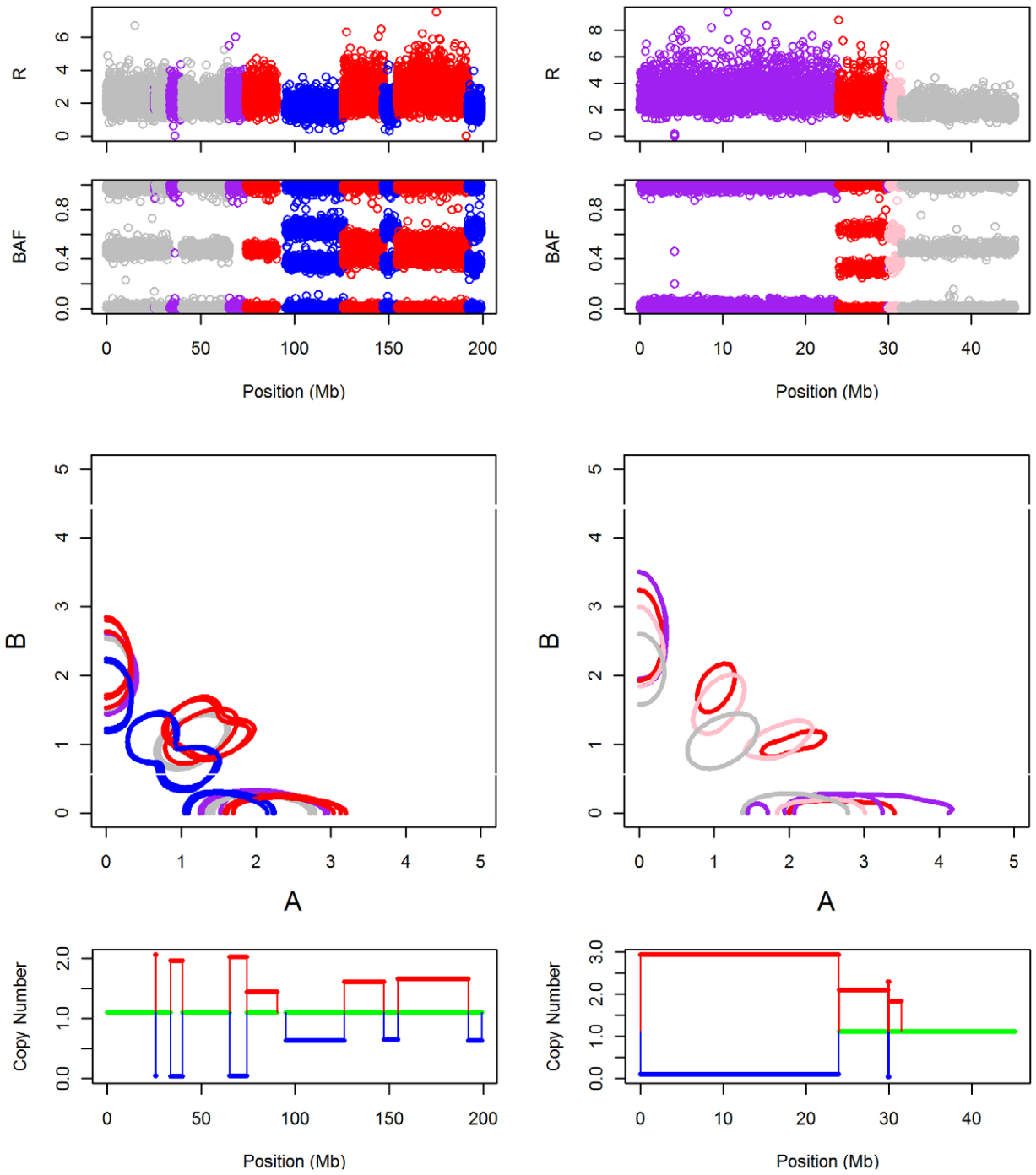

Figure 7. Example regions from TCGA sample 02-0332 chromosome 3 (left) and TCGA sample 02-0258 chromosome 2 (first 10000 SNPs) (right). The plots, in order from the top, show the R values, BAF values, (A, B) contours and estimated major and minor copy numbers. The top three plots are color coded by the segmentation estimated using our procedure. In the color coding of the bottom plot, red represents gain, blue represents loss, and green represents normal.
doi:10.1371/journal.pcbi.1001060.g007

The forward filter. Let X_{s_t} be allele assignment matrices depending on s_t:

$$X_{s_t} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} 1_{\{s_t = AA\}} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} 1_{\{s_t = AB\}} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} 1_{\{s_t = BA\}} + \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} 1_{\{s_t = BB\}}.$$

Let $K_t = \max\{s \le t : \boldsymbol{\theta}_s = \ldots = \boldsymbol{\theta}_t, \boldsymbol{\theta}_{s-1} \ne \boldsymbol{\theta}_s\}$ denote the nearest change-point at a location less than or equal to t. Define

$$p_t = P(\boldsymbol{\theta}_t = \boldsymbol{\mu}_0 | \mathcal{Y}_{1,t}, \mathcal{S}_{1,t}), \qquad q_{i,t} = P(\boldsymbol{\theta}_{K_t} \ne \boldsymbol{\mu}_0, K_t = i | \mathcal{Y}_{1,t}, \mathcal{S}_{1,t})$$

for $1 \le i \le t$. The conditional distribution of $\boldsymbol{\theta}_t$, given $\mathcal{Y}_t$ and the event that $K_t = i$ and $\boldsymbol{\theta}_{K_t} \ne \boldsymbol{\mu}_0$, is $N(\boldsymbol{\mu}_{i,t}, V_{i,t})$, where

$$V_{i,j}=\left(V^{-1}+\sum_{t=i}^{j}X_{s_t}^T\Sigma_{s_t}^{-1}X_{s_t}\right)^{-1},$$

$$\boldsymbol{\mu}_{i,j}=V_{i,j}\left(V^{-1}\boldsymbol{\mu}+\sum_{t=i}^{j}X_{s_t}^T\Sigma_{s_t}^{-1}\mathbf{y}_t\right)$$

for $j\geq i$. It follows that the posterior distribution of $\boldsymbol{\theta}_t$ given $\mathcal{Y}_{1,t}$, $\mathcal{S}_{1,t}$ is the mixture of normal distributions and a point mass at $\boldsymbol{\mu}_0$ given by (5). Let $\phi_{\boldsymbol{\mu},V}$ denote the density function of the $N(\boldsymbol{\mu},V)$ distribution, i.e.,

$$\phi_{\boldsymbol{\mu},V}(\mathbf{y})=(2\pi)^{-1}\ \det(V)^{-1/2}\ \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T V^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}.$$

Making use of $p_t+\sum_{i=1}^{t}q_{i,t}=1$, it is possible to show as in Lai et al. [37] that the conditional probabilities p_t and $q_{i,t}$ can be determined by the recursions

$$p_t\propto p_t^*:=[(1-p)p_{t-1}+cq_{t-1}]l_B\ (\mathbf{y}_t|s_t), \tag{9}$$

$$q_{i,t}\propto q_{i,t}^*:=\begin{cases}(pp_{t-1}+bq_{t-1})\psi/\psi_{t,t}, & i=t,\\ aq_{i,t-1}\psi_{i,t-1}/\psi_{i,t}, & i<t,\end{cases}$$

where $q_t=\sum_{i=1}^{t}q_{i,t}=1-p_t$, $l_B(\mathbf{y}_t|s_t)=\exp\left(\mathbf{y}_t^T\Sigma_{s_t}^{-1}X_{s_t}\boldsymbol{\mu}_0-\boldsymbol{\mu}_0{}^TX_{s_t}^T\Sigma_{s_t}^{-1}X_{s_t}\boldsymbol{\mu}_0/2\right)$, $\psi=\phi_{\boldsymbol{\mu},V}(0)$ and $\psi_{i,j}=\phi_{\boldsymbol{\mu}_{i,j},V_{i,j}}(0)$ for $i\leq j$. Specifically, the mixture probabilities in (5) are $p_t=p_t^*/\left[p_t^*+\sum_{i=1}^{t}q_{i,t}^*\right]$ and $q_{i,t}=q_{i,t}^*/\left[p_t^*+\sum_{i=1}^{t}q_{i,t}^*\right]$.

The smoothing estimate. Since $\{\boldsymbol{\theta}_t\}$ is a reversible Markov chain, we can reverse time and apply the same steps as in the forward equations to obtain (6), in which the weights $\tilde{\boldsymbol{p}}_s,\tilde{\boldsymbol{q}}_{j,s}$ can be obtained by backward induction using the time-reversed counterpart of (9):

$$\tilde{p}_s\propto\tilde{p}_s^*:=\left[(1-p)\tilde{p}_{s+1}+c\tilde{q}_{s+1}\right]l_B\ (\mathbf{y}_s|s_s), \tag{10}$$

$$\tilde{q}_{j,s}\propto\tilde{q}_{j,s}^*:=\begin{cases}(p\tilde{p}_{s+1}+b\tilde{q}_{s+1})\psi/\psi_{s,s} & j=s,\\ a\tilde{q}_{j,s+1}\psi_{s+1,j}/\psi_{s,j} & j>s,\end{cases}$$

where $\tilde{q}_{s+1}=\sum_{j=s+1}^{n}\tilde{q}_{j,s+1}=1-\tilde{p}_{s+1}$. Since for any set A, $P(\boldsymbol{\theta}_t\in A|\mathcal{Y}_{t+1,n})=\int P(\boldsymbol{\theta}_t\in A|\boldsymbol{\theta}_{t+1})dP(\boldsymbol{\theta}_{t+1}|\mathcal{Y}_{t+1,n})$, it follows from (6) and the reversibility of $\{\boldsymbol{\theta}_t\}$ that

$$\boldsymbol{\theta}_t|\mathcal{Y}_{t+1,n}\sim[(1-p)\tilde{p}_{t+1}+c\tilde{q}_{t+1}]\delta_{\boldsymbol{\mu}_0}+(p\tilde{p}_{t+1}+b\tilde{q}_{t+1})N(\boldsymbol{\mu},V)$$
$$+a\sum_{j=t+1}^{n}\tilde{q}_{j,t+1}N(\boldsymbol{\mu}_{t+1,j},V_{t+1,j}).$$

The recursions for deriving the components of the mixture for (7) are exactly the same as those for the earlier model limited to total copy number in Lai et al. [37]:

$$\alpha_t=\alpha_t^*/A_t,\quad \beta_{i,j,t}=\beta_{i,j,t}^*/A_t,\quad A_t=\alpha_t^*+\sum_{1\leq i\leq t\leq j\leq n}\beta_{i,j,t}^*,$$
$$\alpha_t^*=p_t[(1-p)\tilde{p}_{t+1}+c\tilde{q}_{t+1}]/c,$$
$$\beta_{i,j,t}^*=\begin{cases}q_{i,t}(p\tilde{p}_{t+1}+b\tilde{q}_{t+1})/p, & i\leq t=j,\\ aq_{i,t}\tilde{q}_{j,t+1}\psi_{i,t}\psi_{t+1,j}/(p\psi\psi_{i,j}), & i\leq t<j.\end{cases}$$

and we refer the reader to Lai et al. [37] for their derivation.

Estimation of s_t

The variables s_t are assumed to be i.i.d., with

$$s_t\sim\text{Multinomial}(p_t^{AA},\ p_t^{BA},\ p_t^{AB},\ p_t^{BB}).$$

The inherited allele configurations s is assumed to be independent of $\boldsymbol{\theta}$, so

$$P(\mathrm{s}\mid\theta,\mathbf{y})\propto P(\mathbf{y}\mid\mathrm{s},\theta)P(\mathrm{s}\mid\theta)=P(\mathbf{y}\mid\mathrm{s},\theta)P(\mathrm{s}),$$

where

$$\log P(\mathbf{y}|\mathrm{s},\theta)=\frac{1}{2}\sum_{t=1}^{n}\Big[-(\mathbf{y}_t-X_{s_t}\theta_t)^T\Sigma_{s_t}^{-1}(\mathbf{y}_t-X_{s_t}\theta_t)- \tag{11}$$
$$\log|\Sigma_{s_t}|+2\log p_t^{s_t}\Big]+C,$$

where C is a constant. Each component of the above sum can be maximized separately to give, for each t,

$$\hat{s}_t=\text{argmax}_{c\in S}\left[-(\mathbf{y}_t-X_c\boldsymbol{\theta}_t)^T\Sigma_c^{-1}(\mathbf{y}_t-X_c\boldsymbol{\theta}_t)-\log|\Sigma_c|+2\log p_t^c\right].$$

Region Characterization

Let $\{w_i,\ i=1,\ldots,m\}$ be A and B intensities of heterozygous SNPs for segments at normal state and $\{v_i,\ i=1,\ldots,n\}$ be A and B intensities of heterozygous SNPs for the segment being tested. Then, w_i, v_i follow the model:

$$\begin{array}{ll}\text{Normal State :} & w_i\sim N(\mu_0,\sigma_0^2),\ i=1,\ldots,m,\\ \text{Changed State :} & v_i\sim\pi_iN(\mu_1,\sigma_1^2)+(1-\pi_i)N(\mu_2,\sigma_2^2),\\ & i=1,\ldots,n;\ \ \pi_i\sim Bernoulli(p_i).\end{array}$$

For the normal state, we can estimated the parameters easily as

$$\hat{\mu}_0=\bar{w}=\sum_{i=1}^{m}w_i/m;$$
$$\hat{\sigma}_0^2=\sum_{i=1}^{m}(w_i-\bar{w})^2/(m-1).$$

For the target segment, μ_1, μ_2, σ_1^2, σ_2^2 can be estimated by EM algorithm:

Step 1: Initialize: $\mu_1^{(0)}=0.9$, $\mu_2^{(0)}=1.1$, $\sigma_1^{2\,(0)}=1$, $\sigma_2^{2\,(0)}=1$, $\sigma_2^{2\,(0)}=1$, $\pi_i^{(0)}=0.5$, $i=1,\ldots,n$

Step 2: Set

$$\pi_i^{(1)}=\pi_i^{(0)}\phi_{\mu_1^{(0)},\sigma_1^{2(0)}}(v_i)\Big/\left[\pi_i^{(0)}\phi_{\mu_1^{(0)},\sigma_1^{2(0)}}(v_i)+(1-\pi_i^{(0)})\phi_{\mu_2^{(0)},\sigma_2^{2(0)}}(v_i)\right],$$

where

$$\phi_{\mu,\sigma^2}(v)=\exp\left\{-\frac{(v-\mu)^2}{2\sigma^2}\right\}(\sqrt{2\pi}\sigma)^{-1}.$$

Step 3: Set

$$\mu_1^{(1)} = \sum_{i=1}^{n} \pi_i^{(1)} v_i \Big/ \sum_{i=1}^{n} \pi_i^{(1)},$$

$$\mu_2^{(1)} = \sum_{i=1}^{n} \left(1-\pi_i^{(1)}\right) v_i \Big/ \sum_{i=1}^{n} (1-\pi_i^{(1)}),$$

$$\sigma_1^{2(1)} = \sum_{i=1}^{n} \pi_i^{(1)} \left(v_i-\mu_1^{(1)}\right)^2 \Big/ \sum_{i=1}^{n} \pi_i^{(1)},$$

$$\sigma_2^{2(1)} = \sum_{i=1}^{n} \left(1-\pi_i^{(1)}\right) \left(v_i-\mu_2^{(1)}\right)^2 \Big/ \sum_{i=1}^{n} (1-\pi_i^{(1)}).$$

Step 4: Stop if $\left(\mu_1^{(1)}-\mu_1^{(0)}\right)^2+\left(\mu_2^{(1)}-\mu_2^{(0)}\right)^2+\left(\sigma_1^{2(1)}-\sigma_1^{2(0)}\right)^2+\left(\sigma_2^{2(1)}-\sigma_2^{2(0)}\right)^2<\delta_0$, where δ_0 is a pre-chosen threshold (PSCN has default value 10^{-7}). Otherwise, set $\mu_1^{(0)}=\mu_1^{(1)}$, $\mu_2^{(0)}=\mu_2^{(1)}$, $\sigma_1^{2(0)}=\sigma_1^{2(1)}$, $\sigma_2^{2(1)}=\sigma_2^{2(0)}$, and go back to step 2.

The motivation of the initial and default settings are as follows. For segment with changed states, the goal is to estimate minor and major copy number. It is expected that the minor copy number would be less than or equal to 1 and the major copy number would be larger than or equal to 1, so the initial values for μ_1 and μ_2 are set to 0.9 and 1.1 respectively. Although it is possible that both chromosomes in a segment are gained or lost, a small discrepancy of the initial values of μ_1 and μ_2 will also be a good start. Also, it is expected that the numbers of AB and BA states in a segment is similar, so the initial value of π is set to 0.5. The initial values for σ_1 and σ_2 can be quite arbitrary, with 1 being a reasonable value to use. δ_0 is set to be 10^{-7}, which is small enough to indicate a convergence of the iterative algorithm.

Denote the estimated parameters by $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\pi}_i$. To test the hypothesis $H_0 : \mu_1=\mu_0$, the standard t-statistic is

$$T_1 = \frac{\hat{\mu}_0-\hat{\mu}_1}{\sqrt{\frac{\hat{\sigma}_0^2}{m}+\frac{\hat{\sigma}_1^2}{\sum_{i=1}^{n}\hat{\pi}_i}}}$$

Under H_0, the distribution of T_1 is t with degree of freedom $m+n-2$, so p-value can be calculated and compared with the level of the test. The null hypothesis that $\mu_2=\mu_0$ needs also be tested, by replacing $\hat{\mu}_1$ with $\hat{\mu}_2$ in the above equation.

Supporting Information

Text S1 Supporting materials for PSCN.
Found at: doi:10.1371/journal.pcbi.1001060.s001 (0.28 MB PDF)

Acknowledgments

We thank Pierre Neuvial and Henrik Bengtsson for helpful discussions, and an anonymous reviewer for the many helpful comments during the paper revision stage.

Author Contributions

Analyzed the data: HC NRZ. Wrote the paper: HC HX NRZ. Designed methodology: NRZ HC HX.

References

1. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–11.
2. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29: 263–264.
3. Pollack J, Perou C, Alizadeh A, Eisen M, Pergamenschikov A, et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23: 41–46.
4. Matsuzaki H, Dong S, Loi H, Di X, Liu G, et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 1: 109–111.
5. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37: 549–554.
6. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Res 16: 1136–1148.
7. Wang Y, Moorhead M, Karlin-Neumann G, Falkowski M, Chen C, et al. (2005) Allele quantification using molecular inversion probes (MIP). Nucleic Acids Res 33: e183.
8. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 14: 287–295.
9. Hanahan D, Weinberg R (2000) The hallmarks of cancer. Cell 100: 57–70.
10. Broët P, Richardson S (2006) Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. Bioinformatics 22: 911–918.
11. Fridlyand J, Snijders A, Pinkel D, Albertson DG, Jain A (2004) Application of hidden Markov models to the analysis of the array-CGH data. J Multivar Anal 90: 132–153.
12. Hsu L, Self S, Grove D, Randolph T, Wang K, et al. (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6: 211–226.
13. Venkatraman E, Olshen A (2007) A faster circular binary segmentation algorithm for the analysis of array cgh data. Bioinformatics 23: 657–663.
14. Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R (2005) A method for calling gains and losses in array-CGH data. Biostatistics 6: 45–58.
15. Tibshirani R, Wang P (2008) Spatial smoothing and hot spot detection for CGH data using the fused LASSO. Biostatistics 9: 18–29.
16. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20: 3413–3422.
17. Picard F, Robin S, Lavielle M, Vaisse C, Daudin J (2005) A statistical approach for array CGH data analysis. BMC Bioinformatics 6: 27.
18. Engler D, Mohapatra G, Louis D, Betensky R (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridications. Biostatistics 7: 399–421.
19. Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, et al. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. Proc Natl Acad Sci U S A 101: 16292–16297.
20. Wen CC, Wu YJ, Huang YH, Chen WC, Liu SC, et al. A Bayes regression approach to array-CGH data. Stat Appl Genet Mol Biol 5: 3.
21. Xing B, Greenwood CMTM, Bull SBB (2007) A hierarchical clustering method for estimating copy number variation. Biostatistics 8: 632–653.
22. Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763–3770.
23. Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation to arrayCGH data for downstream analyses. Bioinformatics 21: 4084–4091.
24. Affymetrix (2006) Copy number and loss of heterozygosity estimation algorithms for the genechip human mapping array sets. Whitepaper, http://www.affymetrix.com.
25. Bengtsson H, Irizarry R, Carvalho B, Speed T (2008) Estimation and assessment of raw copy numbers at the single locus level. Bioinformatics 24: 759–767.
26. Li C, Wong W (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98: 31?–36.
27. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, et al. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20: 1233–1240.
28. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, et al. (2005) Allele-specific amplification in cancer revealed by SNP array analysis. PLoS Comput Bio 1l: e65.
29. Wan L, Sun K, Ding Q, Cui Y, Li M, et al. (2009) Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Res 37: e117.
30. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics 5: 557–572.

31. Beroukhim R, Lin M, Park Y, Hao K, Zhao X, et al. (2006) Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. PLoS Comput Biol 2: e41.
32. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665–1674.
33. Colella S, Yau C, Taylor JM, Mirza G, Butler H, et al. (2007) QuantiSNP: an objective Bayes hidden Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013–2025.
34. Li C, Beroukhim R, Weir BA, Winckler W, Garraway LA, et al. (2008) Major copy proportion analysis of tumor samples using SNP arrays. BMC bioinformatics 9: 204+.
35. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, et al. (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol 9: R136+.
36. Assié G, LaFramboise T, Platzer P, Bertherat J, Stratakis C, et al. (2008) SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. Am J Hum Genet 82: 903–915.
37. Lai TL, Xing H, Zhang NR (2008) Stochastic segmentation models for array-based comparative genomic hybridization data analysis. Biostatistics 9: 290–307.
38. The Cancer Genome Atlas (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.
39. Bengtsson H, Neuvial P, Speed T (2010) TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC Bioinformatics 11: 245+.
40. Lai T, Liu H, Xing H (2005) Autoregressive models with piecewise constant volatility and regression parameters. Stat Sin 15: 279–301.