GIMscan: A New Statistical Method for AnalyzingWhole-Genome Array CGH Data

Yanxin Shi, Fan Guo, Wei Wu, Eric P. Xing

Nov 2006
CMU-ML-06-115

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon

GIMscan: A New Statistical Method for Analyzing Whole-Genome Array CGH Data

Yanxin Shi[1,2] **Fan Guo**[3] **Wei Wu**[4]
Eric P. Xing[1,2,3]

Nov 2006
CMU-ML-06-115

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[1]Language Technology Institute, School of Computer Science, Carnegie Mellon University.
[2]Machine Learning Department, School of Computer Science, Carnegie Mellon University.
[3]Computer Science Department, School of Computer Science, Carnegie Mellon University.
[4]Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh.

Correspondence should be addressed to Eric P. Xing. Email: epxing@cs.cmu.edu.

Keywords: Array comparative genome hybridization, switching Kalman filters, whole-genome analysis, microarray

Abstract

Array comparative genome hybridization (aCGH) data are often seriously confounded by exogenous (e.g., experimental conditions) and endogenous (i.e., DNA contents) noises, and variations of hybridization signal intensities even within each gene-dosage state. We propose a new statistical method, Genome Imbalance Scanner (GIMscan), for automatically decoding the underlying DNA dosage-states from aCGH data. GIMscan fits a hidden hybridization trajectory to each state. It employs a hidden switching process to stochastically select for each examined clone on the genome its underlying trajectory (therefore the dosage state) that has generated the observed CGH data. Our method captures both the intrinsic (nonrandom) spatial change of genome hybridization intensities, and the prevalent (random) measurement noise during data acquisition; and it simultaneously segments the chromosome and assigns different states to the segmented DNA. We tested the proposed method on both simulated data and real data measured from a colorectal cancer population, and the results demonstrated superior performance of GIMscan in comparison with popular extant methods. We applied GIMscan to a whole-genome aCGH assay of 125 primary colorectal tumors reported in (1), and we report a high-quality genome-level gene dosage alteration map for colon cancer. A software implementation of GIMscan is available from the authors.

1 Introduction

A hallmark of the defective cells in precancerous lesions, transformed tumors, and metastatic tissues, is the abnormality of gene dosage caused by regional or whole chromosomal amplification and deletion in these cells (2). Cytogenetic and molecular analysis of a wide range of cancers have suggested that amplifications of proto-oncogenes and deletions or loss of heterozygosity (LOH) of tumor suppressor genes can seriously compromise key growth-limiting functions (e.g., cell-cycle checkpoints), cell-death programs (e.g., apoptotic pathways), and self-repair abilities (e.g., DNA repair systems) of injured or transformed cells that are potentially tumorigenetic (3). Thus DNA copy number aberrations are crucial biological markers for cancer and possibly other diseases. The development of fast and reliable technology for detecting (the presence of) and pinpointing (the location of) such aberrations has become an important subject in biomedical research, with important applications to cancer diagnosis, drug development and molecular therapy.

Array comparative genomic hybridization (array CGH, or, aCGH) assay offers a high-throughput approach to measure the DNA copy numbers across the whole genome (4). In an array CGH assay, cDNAs from the test sample (e.g., tumor cells) and the control sample (e.g., normal cells) are labelled with different fluorescence, then they are hybridized to microchips coated with probes each of which corresponding to a location-specific clone of the genome and overall covering uniformly the entire genome. The outcome of an array CGH assay is a collection of log-ratio (LR) values reflecting the relative DNA copy number of test versus control samples at all examined locations in the genome. Ideally, for diploid cells, assuming no copy-number aberration in the control and perfect measurement in the assay, the LRs of clones with k copies in the test sample can be exactly computed. For example, when k equals to some integer such as one (reflecting an LOH), two (normal copy number), three (reflecting a haploid duplication) and four (reflecting a diploid duplication), in principle we expect to observe their corresponding LRs to be -1 ($= log_2(\frac{1}{2})$), 0 ($= log_2(\frac{2}{2})$), 0.58 ($= log_2(\frac{3}{2})$), and 1 ($= log_2(\frac{4}{2})$), respectively. A naive approach can directly use these relations to infer the actual copy numbers along the genome from the measured LRs. It is noteworthy that among all possible magnitudes of k, usually only a few need to be distinguished, such as 0, 1, 2, 3, and collectively all integers that are greater (often significantly greater) than 3.

These are the typical copy numbers that reflect distinct cytogenetic mechanisms of chromosome alteration and rearrangements, and hence they are commonly referred to as *gene dosage states*, namely, deletion (D), loss (L), normal (N), gain (G), and amplification (A).

Although by definition the gene dosage states are deterministically related to the LR values, and appear to be trivial to infer, in practice the LR measurements from a real aCGH assay can exhibit severe deviations from their theoretical values due to various reasons, such as impurity of the test sample (e.g., mixture of normal and cancer cells), intrinsic inhomogeneity of copy numbers among defective cells, variations of hybridization efficiency, and measurement noises arising from the high-throughput method (5). These interferences can significantly complicate the estimation of true gene dosage along the genome from the empirical LR measurements, making manual annotation of gene dosage tedious and inaccurate. Numerous computational methods have been developed for efficient and automated interpretation of array CGH data. Earlier methods used value-windows defined by hard thresholds (e.g., according to the deterministic relationship between copy number and LR value discussed above) to determine gene dosage state for each clone based on noisy LR measurement (e.g. (1), see Figure 1). However, these methods suffer from high false positive rate and low coverage because they often ignore, among many factors, the actual value ranges of LR values in an aCGH assay, which can deviate significantly from their theoretical values, as well as variations of signal intensities among clones, chromosomes, and individuals due to reasons other than DNA copy numbers. (We will discuss these issues in detail in the sequel.) Recent developments resort to more sophisticated statistical modeling and inference techniques to interpret aCGH data. Based on the underlying statistical assumptions on signal distribution (either explicitly or implicitly) adopted by these methods, they largely fall into the following four categories.

Mixture models: This class of methods assume that the LR measurements of all the clones in an aCGH assay are INDEPENDENT samples from an underlying distribution consisting of multiple components, each corresponding to a specific gene dosage state. In the statistical literature, such distributions are referred to as mixture models. Standard algorithms are available for estimating the parameters (e.g., the mean and variance of each component) of a mixture model from the

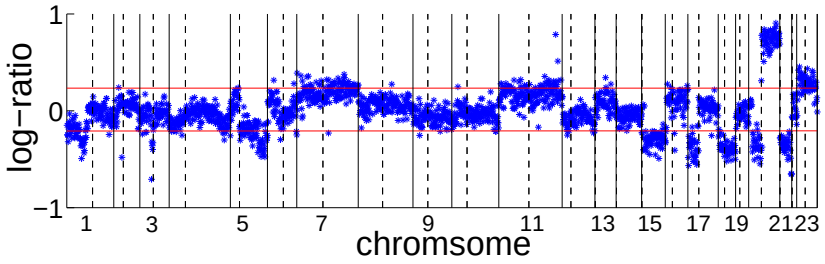


Figure 1: The LR values (blue dots) of genome X77 from Nakao *et al.* (1). The solid vertical lines delineate the boundaries of each chromosome; and the dashed vertical lines indicate the positions of the centromere of each chromosome. The red horizontal lines indicate the thresholds used by Nakao *et al.* (1) to classify clones into dosage states. Clones within these two lines are predicted to be normal state.

observed LR measurements; and the estimated parameters can be used to define either thresholds or classification boundaries for the LRs. For example, Hodgson *et al.* (6) used a Gaussian mixture to fit the LRs measured over the entire genome of an individual, and then determined the gene dosage state based on the upper and lower thresholds of the Gaussian component corresponding to the normal copy number, which were set to be the $\pm$ 3 standard deviation (s.d.) with respect to the mean of this component. This model can be easily extended to a Bayesian classifier of LR values over multiple gene dosage states.

Regression models: Instead of resorting to a mixture model and estimating the plausible dosage-state underlying the LR signal of each clone, regression models try to fit the noisy LRs with a smooth intensity curve over the chromosome to facilitate detection of gene dosage change via visual inspection. For example, Eilers and De Menezes (7) proposed a quantile regression method that employs an L_1 error for both of fitness measure and roughness penalty. Hsu *et al.* (8) used wavelet regression to fit the data. Unlike the mixture models, these nonparametric techniques impose no parametric assumption on the structure of the data. They can effectively capture the spatial continuity of the LR signals along the chromosome; thereby effectively smooth highly noisy LR signals and reveal the overall trend of the data. However, due to the continuous nature of the regression approach, such methods are only suitable for data denoising and visualization, rather than explicitly predicting the discrete dosage state underlying the LR signals.

Segmentation models: A number of recent models attempt to explicitly capture the property that, clones that are consecutive on a chromosome tend to have similar dosage state, and changes of dosage state along a sequence of clones tend to segment the sequence into internally uniform stretches. Rather than focusing on dosage-state estimation, these approaches directly search for breakpoints in sequentially ordered LR signals so that the resulting LR segments have the minimum within-segment signal variations. Examples of this class of methods include ChARM (9), DNACopy (10), aCGH-Smooth (11), CGHseg (12), GLAD (13), CLAC (14), and the algorithm proposed in (15). With such segmentations, the problem of inferring the dosage state of each independent clone is replaced by an arguably more informative task of determining the dosage state of the "coupled clones" in each segments. A number of heuristic annotation algorithms have been developed for this purpose, some of which appear to suffer from state "over-representation", in which numerous spurious states without apparent biological meanings are uncovered from the segments. Various states-merging methods, such as GLADmerge (13) and MergedLevels (16), have been developed as heuristic remedies.

Spatial dynamic models: Willenbrock and Fridlyand (16) pointed out that a more accurate and robust interpretation of aCGH profiles can be expected when the problems of dosage-state annotation and clone-sequence segmentation are solved *jointly* under a unified model for array CGH data, rather than being treated as two separate problems as described above, because the quality of the answer to one of the problems can critically affect the answer to the other problem. They proposed a spatial dynamic framework that models the LR sequence as the output of a hidden Markov model (HMM) that governs the distribution of the dosage-states along the chromosome (5). Using standard inference algorithms for HMMs, their methods can simultaneously learn the LR distributions underlying each dosage-state (i.e., the emission models) and the coupling coefficients of spatially adjacent clones (i.e., the transition model), according to which the hidden dosage-state of each clone can be inferred based on a *maximum a posteriori* (MAP) principle from the observed LRs. Later developments by, e.g., Marioni *et al.* (17) improved this model by considering the distances between adjacent clones when modeling the transition matrix in HMM. Broet and Richardson (18) developed a Bayesian HMM which takes into account the spatial dependence between genomic

sequences by allowing the weights of the dosage-states to be correlated for neighboring genomic sequences on a chromosome. More recently, Shah *et al.* (19) proposed a new Bayesian HMM model that integrates prior knowledge of DNA copy number polymorphisms (CNPs). On simulated data, their method was shown to be robust to outliers by leveraging prior knowledge about the CNPs obtained from high quality cytogenetic experiments.

This progress notwithstanding, the computational methods for aCGH analysis developed so far are still limited in their accuracy, robustness and flexibility for handling complex aCGH data, and are inadequate for addressing some of the deep biological and experimental issues underlying aCGH assay. Take the whole-genome aCGH data displayed in Figure 1 as an example. Overall, the LR signals are highly fluctuating, but exhibit visible spatial auto-correlation patterns within the chromosomes. A caveat of the mixture-model-based or threshold-based methods is that they are very sensitive to such random fluctuations of the LR signals because they treat each measurement as an independent sample and ignore spatial relationships among clones. This could lead to highly frequent dosage-state switching (e.g., alternating back and forth between gain and loss, as we will show in our results) within short genetic distances, which is biologically implausible as we discuss below. Furthermore, due to variations of the length, base content, and other biophysical properties of the clones, which can lead to different abilities of the aCGH probes in matching with their targets on the chromosome, each clone may show a different relationship between the LR and the actual copy number, and such variations of the LR values among clones with the same copy number can become more serious when the copy number depart farther from the genome average (20). These elevated variances could make threshold- or window-based state predictions for clones with high copy numbers even more error-prone.

According to known cytogenetic mechanisms of chromosomal deletion and amplification, such as the onion-skin model (21), chromosomal end breakage (22), or break-fusion-bridge cycle (23), the copy numbers of sequence clones along a chromosome are often spatially correlated rather than being independent. In particular, clones that are consecutive on chromosomes tend to have similar copy numbers, and changes of the copy number along the chromosomes are usually interspersed with sizable regions of stable copy numbers rather than being densely clustered within a short re-

gion, as would be predicted by the simple threshold-based methods. A number of recent methods, particularly the spatial dynamic models based on HMM, have offered various ways to address this issue, which have significantly improved the performance of computational array CGH analysis. Nevertheless, a key limitation of the HMM-based methods is that they all assume invariance of the true hybridization signal intensity alone chromosome for each dosage state, which is not always satisfied in real data. As shown in Figure 1, an outstanding feature of the spatial pattern of the LR signals is that, within each chromosome, there exist both *segmental patterns* that are likely due to change of the copy number of the corresponding region, and *spatial drift* of the overall trend of the LR intensities along the chromosome. For example, in chromosome 4, the LR signals along the sequence of clones are not fluctuating around a baseline (presumably corresponding to a certain gene dosage state) that is invariant along the chromosome; instead, it is apparent that the baseline itself first has an increasing trend from left to right on 4p and into 4q, and then turns to a decreasing trend along the rest of 4q. Visually, there is not many abrupt breakage points that would signal a dosage-state alteration alone this continuously evolving sequence of LRs. But an HMM approach, which models spatially-dependent choices among different copy-number state, each associated with an *invariant* distribution of LR values, can fail to capture the spatial drift of LRs over chromosome region with the same copy number. It would have to make a trade-off between allowing a high-variance dosage-specific LR distribution to accommodate the spatial drift of the LR values in chromosome 4 whilst compromising the stringency of state-prediction along other chromosomes with less LR drift, or maintaining low-variance dosage-specific LR distributions that would do well on other chromosomes but possibly fragment chromosome 4 with many likely-faulty breakage points, mistakenly interpreting the spatial drift as copy-number changes.

Rather than reflecting the discrete change of copy numbers of the clones, the non-random spatial trend of LR signals possibly reflects a continuous change of the biophysical properties and hybridization quality along the chromosome. As discussed in (3), the intensity of the hybridization signal of each clone is affected by a number of factors. First of all, the base compositions of different probes can affect the signal intensity. For example, GC rich probes usually yield higher intensities because of the increased binding affinity (three hydrogen bonds as opposed to

2 hydrogen bonds in A and T). Secondly, the proportion of repetitive content in sequence also exerts an influence. The distribution of such intrinsic sequence patterns maybe nonrandom along the chromosome, but bears a regular spatial trend. Although several methods (e.g., solution denaturation before hybridization and the use of repeat-blocking nucleic acids such as $Cot-1$ DNA) have been demonstrated to suppress the cross-hybridization of repetitive sequences in genomic and expression arrays, the use of these methods will bring other side effects. For instance, when added to target DNA, $Cot-1$ enhances hybridization (2.2- to 3-fold) to genomic probes containing conserved repetitive elements (24). Other factors that may affect the signal intensity include, but not limited to, the saturation of array, divergent sequence lengths of the clones, reassociation of double-stranded nucleic acids during hybridization, and the amount of DNA in the array element available for hybridization. These factors may further contribute random or correlated stochasticity of the LR values on top of the content-derived spatial drift. Pinkel and Albertson (3) reported that signal intensity may vary by a factor of 30 or more among array elements even if there are no copy-number changes. These complexities present in real aCGH data render extant models based on fixed state-specific LR distributions, such as an HMM, incapable of making accurate and robust state prediction.

Another problem that affects all the approaches discussed above lies in the calibration of the signals across chromosomes and across individuals. As observed from Figure 1, the mean and the variance of the LRs, and their spatial trends vary significantly from chromosome to chromosome, and more so from individual to individual (not displayed in the figure), due to reasons possibly beyond (whole chromosome) copy number differences. This makes measurements from different individuals and/or for different chromosomes difficult to compare. Engler *et al.* (25) recently proposed a parameter sharing scheme for a Gaussian mixture model for genetic variability between chromosomes and within chromosomes. In the new statistical model for aCGH data we present below, we introduce more careful treatments of this issue, which employ different parameter sharing scheme for effects shared among different chromosomes in the same individual (e.g., state baselines) and effects common to the same chromosomes in different individuals (e.g., signal dynamics).

In this paper, we introduce a new method—*Genome Imbalance scanner*, or GIMscan—for computational analysis of aCGH data. GIMscan employs a more powerful spatial dynamic model, known as switching Kalman filters (SKFs) (26), to jointly capture the spatial-trends of evolving LR signals along chromosomes, and spatially dependent configuration of gene dosage states along chromosomes. Unlike an HMM, which captures all the stochasticities in LRs with invariant dosage-specific distributions, an SKF breaks the accumulation of the stochasticities into two stages: 1) the *hybridization stage*, which involves physical sensory of clone-copies from the digested chromosomes, during which the spatial trend of DNA content and its biophysical properties, saturation effects, etc., can cause stochastic spatial drift of the mass of the hybridized material; 2) the *measurement stage*, which involves acquisition of the readings of fluorescence intensity of each clone, during which errors from reagents, instruments, environment, personal effects, etc., can cause another layer of random noises on top of the hybridization signal. Under the SKF, we model the variations in the hybridization stage using dosage-state-specific continuous dynamic processes, akin to the regression approach discussed above. Specifically, for each specific gene dosage, its resulting hybridization intensities along a chromosome are modeled by a unique linear dynamic process that allows the intensities to change according to a spatially continuous trend, with zero mean Gaussian noise. These hybridization intensities can be understood as the "true" *sensory signals* in an aCGH assay, which are unobservable to the examiner. We refer the sequence of hybridization intensities following such a linear dynamic model as a *hybridization trajectory*. Given the hybridization trajectory, we model the random noise from the measurement stage by a conditional Gaussian distribution whose mean is set by the sensory signal which evolves over each clone according to the trajectory. Overall, for each dosage state, we have a unique linear dynamic model for the sensory signals and a Gaussian emission model for their corresponding noising measurements. This model is known as a Kalman filter (or in the statistical community, a state-space model). To model changes of dosage-state along the chromosome, we follow the HMM idea to set up a hidden Markov state-transition process, but in our case not over state-specific distributions of LRs with fixed means, but over state-specific Kalman filters over both the observed LR measurements and the unobserved hybridization trajectories of the clones. For this reason the model

is known as switching Kalman filter, which has been previously used successfully in predicting moving objects and content from video streams or radar records (27, 28).

On both simulated and experimental aCGH data, GIMscan has shown superior performance over other approaches such as HMM or mixture-model-based threshold methods, being able to handle a number of complex LR patterns beyond the recognition power of the reference models. We applied our methods to a whole-genome aCGH assay of 125 primary colorectal tumors reported in (1), and constructed a high-quality genome-level gene dosage alteration map for colon cancer. On average, 19% of the genome region of each individual have altered copy number. There is a significant difference between the frequencies of loss in stage 1 and stage 4 tumors with p-value 0.033, and between the frequencies of amplification of stage 2 and stage 3 tumors with p-value 0.025. A number of significant dosage alteration hotspot were identified, in particular, the loss/deletion rate in 8p, 17p and 18p and 18q are close to or above 50%, and the gain/amplification rates were found to range from 41.6% to 65.65% in 8q and 20q. Many of these hotspot regions coincide with oncogenes and tumor repressor genes, including BRCA1 and p53, which are previously known to be highly relevant to tumorigenesis, and there is an apparent correlation between the dosage-state with genes involved, in that the frequently gained regions tend to harbor oncogenes, whereas the frequently lost regions tend to harbor tumor repressors.

2 Results

We first present results of GIMscan's performance on simulated aCGH data and small-scale real data with complex aCGH patterns to demonstrate the working principle and general trends of our method in gene dosage prediction, and to evaluate our prediction quality under nontrivial genome imbalance and hybridization scenarios. Then we describe results of a populational whole-genome genetic instability analysis of 125 primary colorectal tumors published previously in (1) using GIMscan, and discuss some new implications revealed in our results on the molecular alteration spectrum of this cancer. The benefit of applying a sophisticated probabilistic model as in GIMscan to capture both discrete changes of gene dosage state (i.e., DNA copy number) along the chromosomes, and continuous spatially-correlated variations of the hybridization trajectory underlying

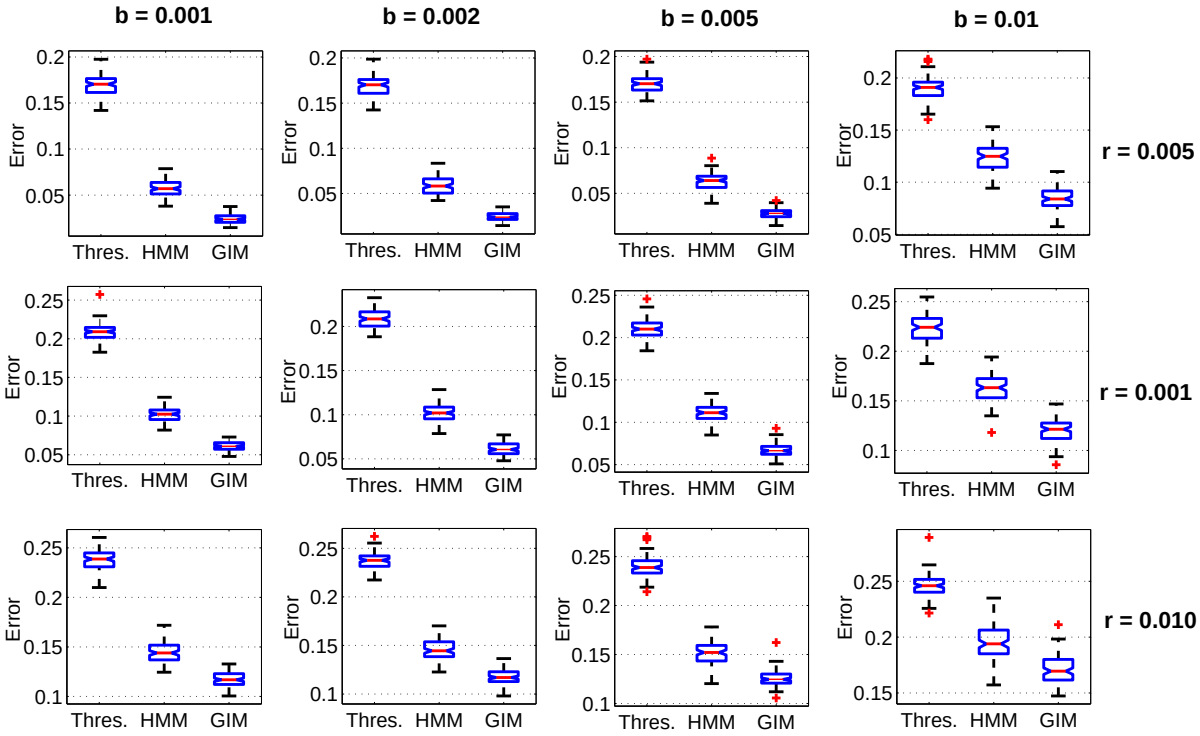


Figure 2: Performance of DNA dosage-state prediction on simulated aCGH datasets. Each row corresponds to a specific level of simulated measurement noise, and each column corresponds to specific level of simulated variability of the hybridization signal. In each case, we report the results of threshold method (Thres.), HMM and GIMscan (GIM). The red line represents the median, and the blue box indicates upper and lower quantiles. The black bars are the range of the error rate. Outliers are plotted by "+".

aCGH measurements along chromosomal regions with the same gene dosage, is evidenced in each level of genetic scales (e.g., regional, whole-genome, and populational levels) we have analyzed.

2.1 Simulated aCGH Data

We first validate GIMscan on simulated aCGH datasets, which mimic typical spatial patterns of LR sequences in real aCGH assays. This experiment allows a quantitative assessment of model performance based on known underlying gene dosage states in the simulation.

In our simulation experiments, three methods—threshold, HMM as in (5), and GIMscan—were tested on 12 datasets simulated with different settings of two parameters: the Gaussian emission variance r, and the KF transitional variance b (see §4.1 for a rigorous definition and description

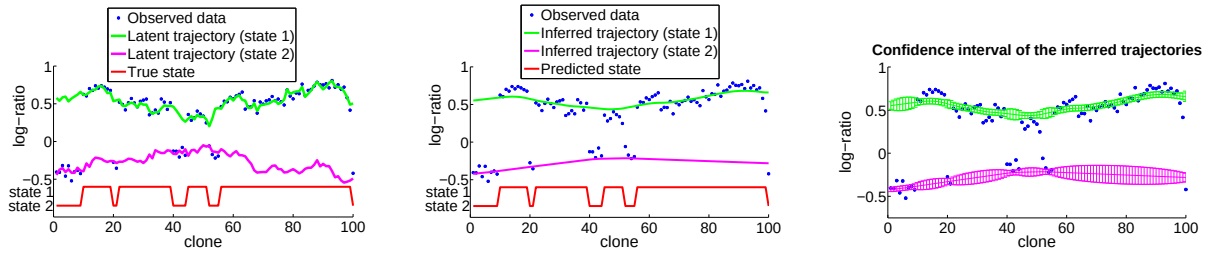


Figure 3: Performance of GIMscan on simulated "high-quality" aCGH data. (a) The simulated data (blue dots), from two latent trajectories (green and pink) and a process (red) switching between them. The length of the simulated data is 100 clones. (b) The inferred trajectory (green and pink) and switching process (red). (c) The confidence intervals of the inferred trajectories.

of these parameters). Statistically, these two parameters represent the two sources of the overall noise in the data: r reflects the quality the LR measurements in an aCGH experiment, whereas b reflects the variability of the hybridization signal intensity along the chromosome. Our datasets correspond to three different values of r, ranging from low, to medium, high; and four values of b also spanning a significant range (Figure 2). For each combination of r and b, a total of 100 LR sequences, each containing 100 clones, were generated. For each sequence, we simulated a random 5×5 stochastic matrix, T, for modeling transitions between gene dosage states, and T was set to allow both short and long stretch of gene dosage alterations, but not high-frequency oscillations between different states [1]. All three methods were applied to each dataset to infer the gene dosage-states underlying the simulated LRs, which was then compared against the true dosage-states recorded during the sumilation. The experiments were repeated 100 times. Figure 2 summarizes the medians, quantiles and ranges of the prediction error rates by different methods under various parameter settings. Consistently, GIMscan outperformed the other two methods by a significant margin.

As an illustration of the advantage offered by the SKF model adopted by GIMscan, and the effectiveness of our inference algorithm, Figure 3 and 4 show two examples of GIMscan's performance in the simulated datasets. The first example concerns "high-quality" aCGH records simu-

[1] Such presumably realistic genome imbalance patterns can be enforced by allowing only low transition probabilities between the normal and aberrant states and/or between different aberrant states, and defining high self- transiting probabilities for the normal states and/or the long-duration aberrant states.

lated with low measurement noise ($r = 0.001$) over 100 clones switching between two gene dosage states both with low spatial drift in their corresponding true hybridization intensities ($b = 0.001$) (Figure 3a). Figure 3b presents the inferred gene dosage state and the inferred dosage-state-specific "trajectories" (i.e., the latent dynamical trend captured by each KF) of the latent true hybridization intensities underlying the observed LR sequence. As shown in this illustration, each inferred latent trajectory indeed represents a smoothed and spatially changing baseline of the LR signals corresponding to a particular dosage state. Both of the inferred trajectories agree well, in positions where corresponding LRs are present, with the true trajectories (shown in Figure 3a) of hybridization intensities used for simulating the observed LR signals. As a result, the inferred switching process over these trajectories gives a highly accurate prediction of the gene dosage-states underlying the LR sequence. GIMscan can also estimate the confidence intervals (i.e., standard deviation) of the inferred hybridization trajectories, as shown in Figure 3c. As one can see, only in positions where corresponding LRs are sparse or absent (e.g., for state 2 at position 60-100, where all but one LR are belonging to the other state), the conference interval over the estimated trajectories become large; otherwise, we can obtain a quite reliable estimate of the latent hybridization trajectory that have generated the LR signals.

The second example shown in Figure 4 concerns low-quality, arguably more realistic aCGH records simulated with high measurement noise ($r = 0.01$) and severer spatial change ($b = 0.01$) in the true hybridization trajectories. The combined effects of high measurement noise and high spatial variance of the hybridization trajectories are excepted to lead to misassignment of gene dosage state due to inaccurate estimation of the dosage-state-specific hybridization intensities when spatial trajectory of the hybridization intensities is ignored. Note that the "true" trajectories (shown in Figure 4a) of both dosage states are not flat, which reflect severe spatial drift of hybridization signal intensities within each state. When assuming spatial invariance of dosage-state-specific signal distribution, the unflatness of both trajectories can cause the estimated mean of LR signals for each dosage-state to be highly biased (e.g., higher for state 1, and lower for state 2), and their variances to be significantly greater than the actual fluctuation over the underlying trajectory. Consequently, the estimated dosage-state-specific signal distributions can be seriously overlapping,

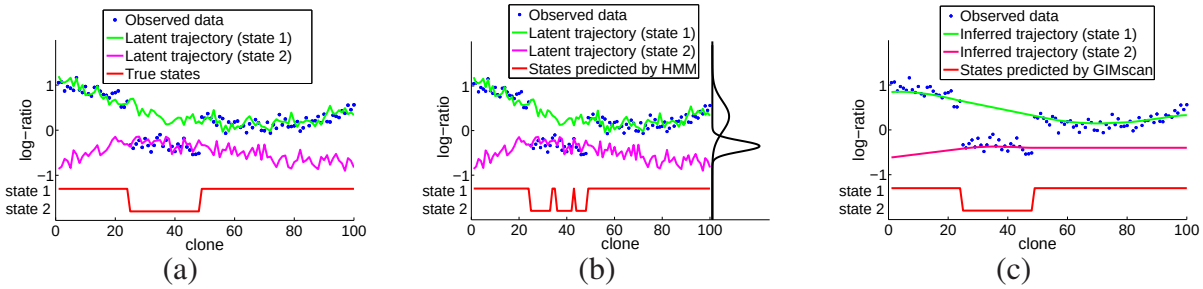


Figure 4: A comparison of performances of GIMscan and HMM on simulated "low-quality" aCGH data. (a) The simulated data. (b) The dosage-states (red line) predicted by an HMM; and the Gaussian densities of LRs corresponding to the two dosage-states (black curves left to the simulated data), as estimated by HMM. (c) The trajectories estimated by GIMscan (green and pink curves) and the predicted gene dosage states (red).

causing the LR signals from two states hard to distinguish. Figure 4b shows exactly this effect, on the quality of state estimation by an HMM model. in contrast, the SFK model underlying GIMscan can readily mitigate this effect, and produces much more accurate estimation of the latent trajectory and measurement noise (Figure 4c). [2].

2.2 Interpreting Real aCGH data with Diverse Spatial Patterns

Now we present case studies of selected real aCGH data with a diverse spectrum of spatial patterns. Our dataset was obtained from an online repository of whole-genome aCGH profiles of 125 colorectal tumors originally studied in (1). This dataset was found to contain highly stochastic LR measurements with severe spatial variance and drifts along the chromosomes, and bear rich cohorts of genome imbalance patterns (as shown in the sequel). Such complications present a great challenge to naive algorithms for gene dosage inference, and are thus particularly suitable for evaluating our newly proposed method.

Given an aCGH profile, GIMscan first employs a k-nearest neighbor regression procedure (e.g., $k = 3$) to impute the missing values in the LR records. Then it fits the processed data with a Gaussian mixture based on maximum likelihood estimation, and performs model selection based on

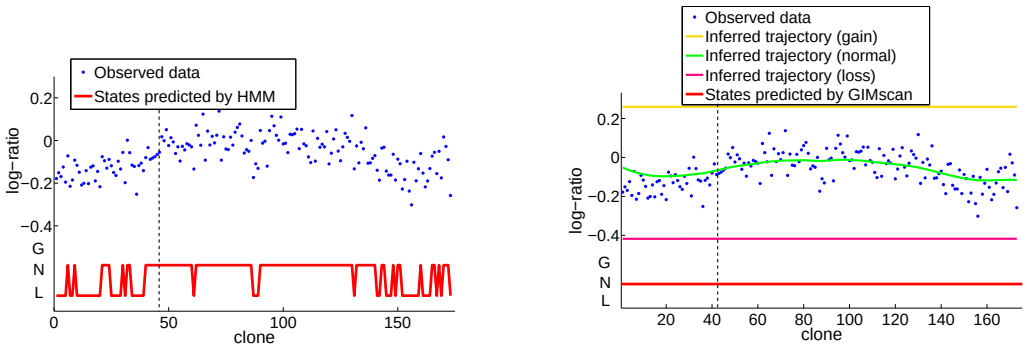
[2]Note that as shown in Figure 4c, although the latent trajectory of the second state is not inferred accurately in GIMscan (i.e., it exhibit a flat rather than a curved trajectory) due to insufficient supporting signals from the data (i.e., no LR signals were generated from this state at the left and right ends of the sequence), it nevertheless still enables a reasonable resolution of the two states.

AIC (29) to determine the total number of gene dosage states, M (which is constrained between 1 to 5), for each individual. Afterwards, the number of component KFs (i.e., dosage-state-specific hybridization trajectories) in GIMscan is set to be M, and the mean of the starting clone of each KF takes on the mean of a component in the estimated Gaussian mixture as initial value. Note that with this setup, we still need to establish the exact mapping between the KFs inferred by GIMscan and the possible gene dosage states, namely deletion, loss, normal, gain, and amplification. Since GIMscan provides estimations of the hybridization trajectories of each KF, we follow a straightforward statistical and biological argument and determine the corresponding dosage-state of each trajectory based on the relative mean-values of the estimated true hybridization intensities of all clones of each trajectory. (E.g., the trajectory having near-zero hybridization intensity at the majority of the clones is deemed to be corresponding to a normal state; other trajectories are then labeled accordingly.)

For comparison, we re-implemented the HMM methods according to Fridlyand *et al.* (5), with modest extension (i.e., parameter sharing) so that it can be applied to whole genome CGH profiles covering multiple chromosomes. Following (5), AIC is also used for model selection for this HMM.

The dataset we studied contains a total of $\sim 2.75 \times 10^5$ LR measurements from 23×125 chromosomes (i.e., 125 human genomes). In this subsection we present a small-scale case study of four representative chromosomes, each containing a typical spatial pattern for the LR sequence that was found to be difficult to analyze by conventional methods. For convenience, we refer to these patterns as, flat-arch, shallow-step, deep-step, and spike, respectively, according to their shapes in the LR intensity plots (Figure 5 - 8).

Pattern I: Flat-Arch As shown in Figure 5, which displays the LR measurements from chromosome 4 of individual X77, this pattern is marked by lower magnitudes of LRs at the two telomere regions of the chromosome and elevated magnitudes in the central region. Locally (i.e., along the plotted chromosomal region), there is a continuous trend of spatially evolving hybridization intensity along the chromosome (but within a relatively-small intensity range well bellow the theoretical LR difference due to dosage-state change), and there are few abrupt breakage points that



(a) (b)

Figure 5: The Flat-Arch pattern, from chromosome 4 of individual X77. (a) The state sequence predicted by HMM (red). (b) The inferred trajectories for loss (pick), normal (green) and gain (yellow) states, respectively, by GIMscan. The Red solid line indicates states predicted by GIMscan. The centromere position is indicated by dashed vertical line in these two plots.

would signal a dosage-state alteration. But due to the high dispersion of LR values as a result of such a spatial drift, methods based on invariant state-specific hybridization intensity, such as the HMM, would either fit the observed LR values with a single biased and high-variance Gaussian distribution, or split the LRs with two highly overlapping Gaussians. These caveats could seriously compromise the quality of gene dosage state estimation. Figure 5a shows the dosage estimation by an HMM fitted on this chromosome. The outcome suggests heavy oscillations between two dosage states throughout the chromosome, which is biologically implausible.

Figure 5b shows the dosage state sequence and dosage-state-specific trajectories underlying chromosome 4 of individual X77 inferred by GIMscan. A whole-genome fitting resulted in three estimated dosage-states. On this particular chromosome, the trajectories of the loss and gain states (the pink and yellow curves, respectively) were not matched to any observations, and the entire region is determined to be corresponding to a normal state whose hybridization intensity varies along the chromosome (the green curve). Indeed, a more global visual inspection of these Flat-Arch patterns in the context of whole aCGH profile revealed that the flat-arch shapes in the LR-plots often merely reflect modest (but spatially correlated) change of the LR magnitude most likely within a single dosage state.

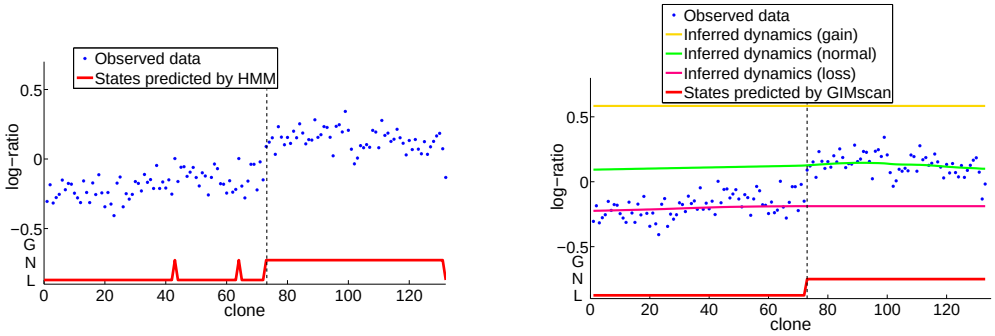


Figure 6: The Short-Step pattern, from chromosome 1 of individual X232. (a) The state sequence predicted by HMM (red). (b) The inferred trajectories for loss (pick), normal (green) and gain (yellow) states, respectively, by GIMscan; and the state sequence predicted by GIMscan (red).

Pattern II: Short-Step As shown in Figure 6, this pattern is typical when there appears to be a quantum change of LR magnitudes from one to the other end of the chromosome, but the boundary of the change is not sharp and the overall sequence is moderately noisy. As shown in Figure 6a, HMM interpreted this pattern reasonably well, predicting the state-switching position at a visually apparent site, where the $\delta(LR)$, the difference of LRs of the clones immediately proximal and distal to the site, is maximum. However, it also predicted additional point-gains at two other sites with big $\delta(LR)$.

GIMscan predicted a gene dosage state-switching at the same site as the HMM did, but classified the two other sites with high $\delta(LR)$ as mere local fluctuation rather than dosage-gain (Figure 6b). A close inspection of the overall local distribution of LRs neighboring this two sites, which display notable fluctuations, appears to support this conclusion. In particular, in both cases, the $\delta(LR)$'s between the "gained clone" and the one immediately distal to it exhibit a much smaller value, suggesting that they are unlikely to be a point gain between two normal clones. It would be valuable to go back to the original sample to check the GIM status cytologically to verify our results, but unfortunately at this stage this remains technically difficult for us.

Pattern III: Long-Step This pattern is marked by a prolonged region in which the LR signals display a gradual change in their magnitude, and an overall large difference of overall LR mag-

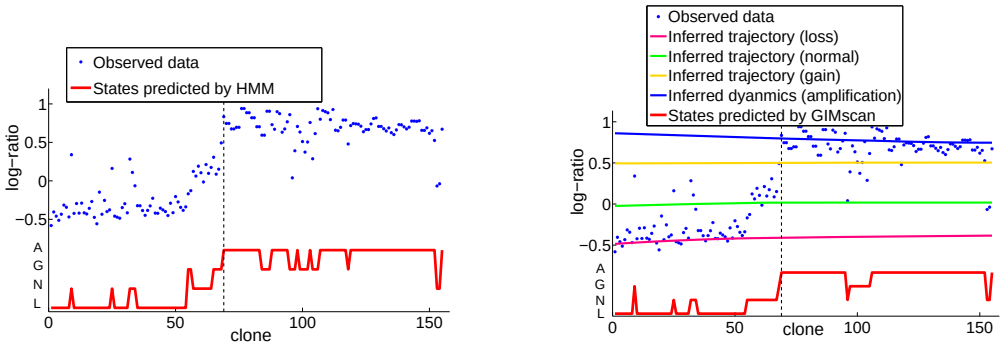


Figure 7: The Long-Step pattern, from chromosome 8 of individual X265. (a) The state sequence predicted by HMM (red). (b) The inferred trajectories for loss (pick), normal (green), gain (yellow), and amplification (blue) states, respectively, by GIMscan; and the state sequence predicted by GIMscan (red).

nitudes between the two sides of the step. Figure 7 shows such an example, which is taken from chromosome 8 from individual X265. In addition to the long step, this sample also harbors a number of local spikes and short regions potentially implying dosage-state alterations. Via AIC model-selection, the HMM adopted four dosage states when processing this data. The states predicted by HMM are shown in Figure 7a. As can be seen, the results are reasonable, expect that several positions near clone 100 contains highly frequent switching between states.

The dosage-state sequence and dosage-state-specific trajectories inferred by GIMscan are shown in Figure 7b. Note that there is a slightly decreasing trend in the hybridization trajectory corresponding to the amplification state. While the trajectories of the gain and normal states correspond to only a few clones on this chromosome, a genome-level parameter sharing scheme adopted by GIMscan enables them to be reliably estimated, and thereby leads to plausible prediction of point changes on isolated clones (e.g., clone 97 and 152).

Pattern IV: Spikes Spikes are a typical pattern often accompanying other patterns, such as steps. It is marked by short sequences, sometimes singletons, of elevated or attenuated LR measurements along the chromosomes. Figure 8 shows such an example from chromosome 8 of individual X318. In this chromosome, the copy-number loss was apparent on 8p arm, while three spikes (around clone 75, 110 and 140) were visible on 8q arm. These spikes correspond to the gain state with a large measurement variance.

Figure 8a shows the states predicted by HMM. Although HMM correctly predicted the states

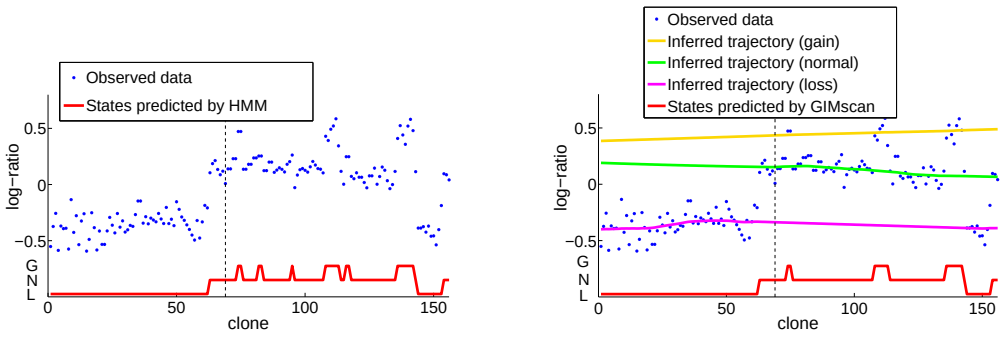


Figure 8: The Spikes pattern: chromosome 8 from individual X318. (a) The state sequence predicted by HMM (red). (b) The inferred trajectories for loss (pick), normal (green) and gain (yellow) states, respectively, by GIMscan; and the state sequence predicted by GIMscan (red).

on 8p, it predicted more clones on 8q to be of gain state. However, by our visual check, some clones (e.g. around clone 79, 96) are more likely to be classified as of normal state. The possibly faulty predictions of gain states were resulted from the large variance of the spikes estimated by the HMM.

GIMscan gives a more plausible interpretation of the spikes, as well as giving convincing predictions on other clones (Figure 8b). Compared to the case for the same chromosome (i.e., no. 8) from another individual (X265) shown in Figure 7, where four dosage-state-specific trajectories were determined, here we uncovered only three states for chromosome 8. This is because model selection for SKF in this individual based on the whole-genome aCGH only identifies three states—normal, loss and gain. Parameter-sharing was adopted by GIMscan for all chromosomes in this individual, and leads to three common trajectories. Comparing Figure 7b with Figure 8b, one can notice that the elevates of the trajectories corresponding to the same dosage state (e.g., normal) can be quite different across individual, which is likely due to some unidentified systematic error or difference in hybridization-efficiency across individuals. The parameter-sharing scheme adopted by GIMscan (i.e., sharing dosage-state-specific trajectories across chromosomes within individual, but not across individual) provides a reasonable strategy to tackle such variations.

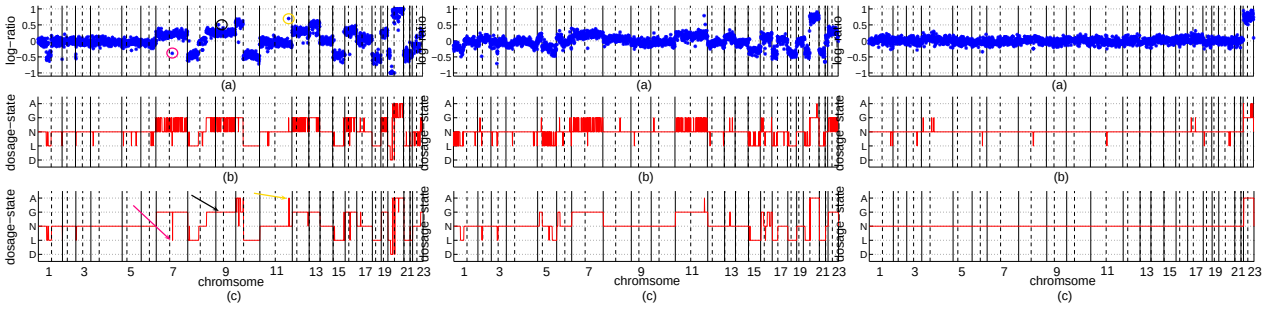


Figure 9: Whole-genome GIM patterns of three individuals: X31, X77 and X40. In each panel, row (a) shows the LR values for the entire genome; row (b) shows the states assigned by threshold method; and row (c) shows the states assigned by GIMscan. The solid vertical lines show the boundaries between chromosomes, and the dashed vertical lines indicate the centromere of chromosomes. In the first panel, we highlight exemplary spike LR signals that were determined to be true local change of dosage state, such as point loss (red arrow) or point amp (yellow arrow); or merely an deviations within the same dosage state (black arrow).

2.3 Whole-Genome Analysis of Genetic Instability in Colorectal Tumors

In this section, we present a population analysis of whole-genome genetic imbalance in colorectal tumors using GIMscan. Here we analyze the complete aCGH dataset by Nakao *et al.* (1) mentioned before, which contains 125 primary colorectal tumors assayed by tiling arrays with 2463 bacterial artificial chromosome (BAC) clones provided by the UCSF Cancer Center Array Core. Roughly, these clones are 1.5 Mb apart from each other in the genome. In an early analysis of this data based on the threshold method, Nakao *et al.* (1) reported presence of genetic imbalance in an average of 17.3% of the genome (8.5% gain and 8.8% loss) in this population.

Figure 9 shows three examples (individual X31, X77 and X40) of whole-genome GIM patterns inferred by GIMscan, in comparison with the GIM patterns predicted by simple thresholding method as reported in (1). Comparing to the thresholding results, for each individual, GIMscan gives a more smoothed, arguably biologically more plausible prediction of the gene dosage aberration spectrum of the genome. We found that the total number of copy-number breakage points predicted by GIMscan in each genome is significantly reduced. The lengths of the chromosome segments with abnormal copy numbers are much less dominated by point (i.e., single-clone) alterations as predicted by thresholding (Fig 10), and exhibit a much higher frequency of long-range

alterations. In particular, GIMscan detected a significantly greater number of occurrences of very long-range alterations (i.e., longer than 50 clone-coverage) than the threshold method. The number of loss/deletion whose length is greater than 50 clones is 146 for GIMscan and only 21 for threshold method; and the number of gain/amplification whose length is greater than 50 clones is 182 for GIMscan and 33 for threshold method. Interestingly, we found that the number of amplification/gain and that of deletion/losse are roughly comparable for short-range (e.g., < 30 clones) alterations, whereas for long-range alterations, amplifications and gain appear to be slightly more abundant. The longest chromosomal segments we found to be altered encamps 183, 290, 127, and 265 clones, respectively, for amplifications, gain, deletions, losses. (The longest altered segments detected by threshold methods are significantly shorter, covering 54, 88, 89, and 131 clones, respectively, for amplifications, gain, deletions, losses).

It is noteworthy that in different individuals, there may exist different number of detected gene-dosage states, which is statistically determined by AIC-based model-selection. For example, X33 contains all five states (i.e., A, G, N, L, D); whereas only fours states (i.e., A, G, N, L) are present in X77. For individual 40, GIMscan only finds two possible states by a genome-wide maximum likelihood mixture model fitting, via AIC model-selection.

Furthermore, the baseline values (the mean trajectory) and the spreads (i.e., the variance) of each state-specific intensity trajectory, according to which all clones are classified into different gene-dosage states, can exhibit significant difference in different individuals. This suggests the presence of categorical influences from possibly sample, environmental, and experimental conditions on the LR measurements. For example, we found that the initial mean $\mu_0^{(N)}$ of the normal trajectory and the variance $b^{(N)}$ along the trajectory is -0.0397 and 0.0023, respectively, for individual 77, whereas the value of this same pair of parameters in X40 is 0.0063 and 0.0016, respectively. For the amplification state, we have 0.8792 and 0.0013 for $(\mu_0^{(A)}, b^{(A)})$ in individual 77, and 0.7934 and 0.0009 in individual 40. These systematic differences indicate the importance of proper normalization and parameter sharing scheme during population-wide GIM analysis. The inter- and intra-gnomic coupling scheme of different SKF parameters adopted by GIMscan offers a reasonable empirical solution to this issue.

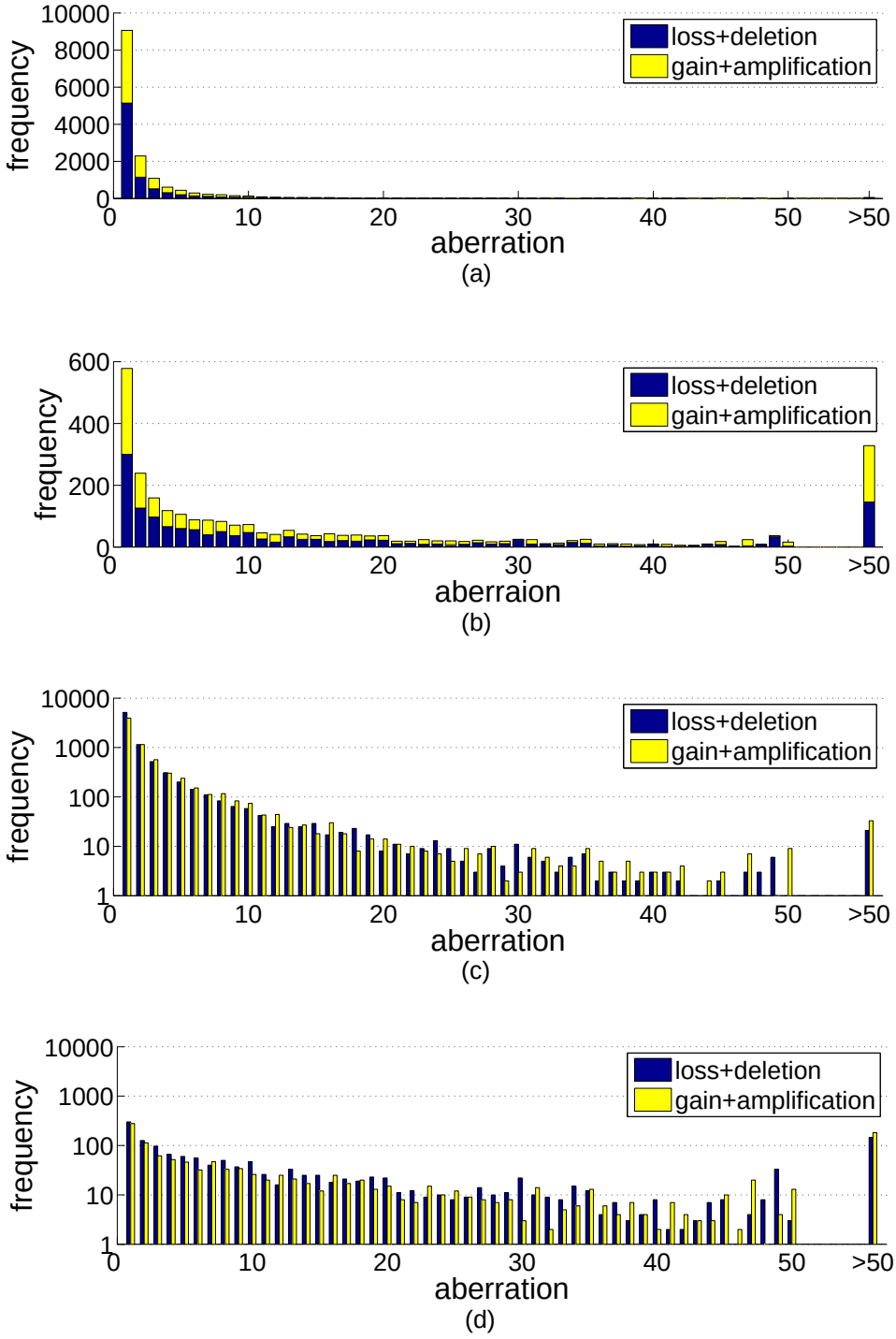


Figure 10: (a) Stacked bar graph of the number of consecutive aberrations versus the log value of the number of their appearances according to Nakao *et al.* (1). (b) Stacked bar graph of the number of consecutive aberrations versus the log value of number of their appearances according to GIMscan. (c) same as (a), but frequency is in a log scale. (d) same as (b), but frequency is in a log scale.

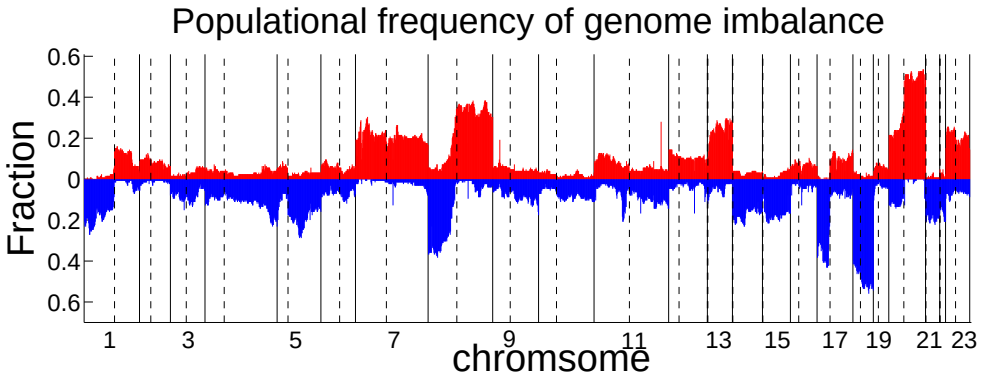


Figure 11: Overall frequency of DNA dosage-state alteration over the entire genome of 125 individuals. The blue bars represent the clones with DNA copy number loss or deletion. The red bars represent the clones with DNA copy number gain or amplification. The solid vertical lines show the boundary between chromosomes. The dashed vertical lines indicate the centromeres of chromosomes.

One of the most important information coming out of our GIM analysis is the *spatial spectrum* of *GIM occurrences* of each individual and the spatial spectrum of *GIM rates* over the study population. As shown in Figure 9, the spatial spectrum of GIM occurrences can be very different from individual to individual, even though they all represent samples of colorectal cancer. In particular, some individual (e.g., X31) exhibit highly prevalent GIM affecting many regions in the genome, others (e.g., X40) merely have GIM in very restricted genomic regions. This phenomena suggests that the molecular mechanisms underlying colorectal cancer can be highly heterogeneous across the patient population.

Overall, over the 125 genomes each examined at the 2463 clones roughly uniformly distributed across the genome, on average each genome have 19.18% (or 407) of the clones suffered either gain or loss (9.25% and 9.94%, respectively), and another 1.33% of the clones were hit by amplification or deletion (0.93% and 0.4%, respectively). The whole-genome spatial spectrum of GIM rates over the entire study population is displayed in Figure 11. As can be seen, the population rates of gain and amplification of clones in chromosome 7, 8q, 13q, 20q and 23 were significantly higher than that of the other regions, suggesting possible presence of proto-oncogenes in these regions. We refer to these regions as ”GA-hotspots”. Likewise, the population rates of loss and deletion in chromosome 1p, the distal-end of 4q, 5q, 8p, 14, 15, 17p, 18, and 21, were significantly

higher than that of the other regions, suggesting possible presence of tumor suppressor genes in these regions. These regions are referred to as "LD-hotspots". Interestingly, the GA-hotspots and the LD-hotspots are spatially complementary in the genome, meaning that a regions can not have frequent gain/amplification (in a subset of the individuals) and at the same time suffer frequent loss/deletion (in the remaining subset of the individuals) in the colorectal cancer population. Among the GIM hotspots, no clone exhibited amplification or deletion in more than 35% of the study population, whereas 18 clones (covering 0.85% of the genome) were gained and 74 clones (covering 3.48% of the genome) were lost in more than 35% of the study population.

2.4 Analysis of selected GIM hotspots in Colorectal Tumor

In this section we take a close look of 4 of the GA/LD hotspot regions identified by GIMscan in the Colorectal patient population, and explore their implication to the carcinogenic mechanisms of this disease. We first examined the GIM patterns on chromosome 8 in the study population. As shown in the high-resolution result reported in the appendix (Figure 14h), frequent losses occurred on 8p arm, whereas frequent gains were detected on 8q arm. Overall, 52 individuals (41.6%) have gains on their 8q arm, and 54 individuals (43.2%) have losses on their 8p arm. Interestingly, the transition between the loss and the gain does not occur at the centromere (48Mb). Instead, the frequency of loss began to drop and the frequency of gain began to rise around 41Mb (8p11-8p12). The spatial distribution of clones subject to amplification and deletion has the same pattern, although their coverage and frequency are much lower than gain and loss. To reveal the molecular implications of this GIM hotspot region, in Figure 12a we show a literature-based rough annotation of the cancer related genes on chromosome 8. DLC1, PDGFRL, LZTS1, MSR1 and TNFRSF10B were found in 8p arm, where highly frequent loss occur. Although we did not find any minimal loss region that can pinpoint exactly a particular target among these genes, they may be candidates of tumor suppressors. Among these genes, DLC1 is Rho-GTPase-activating protein coding gene. It was found to be frequently deleted in human liver cancer (30), and its expression is with rather lower levels in prostate, testis, ovary, small intestine and colon. Indeed it has been validated as a tumor suppressor gene in many cancers (31). RB1CC1 and TNFRSF11B was found in the GA-

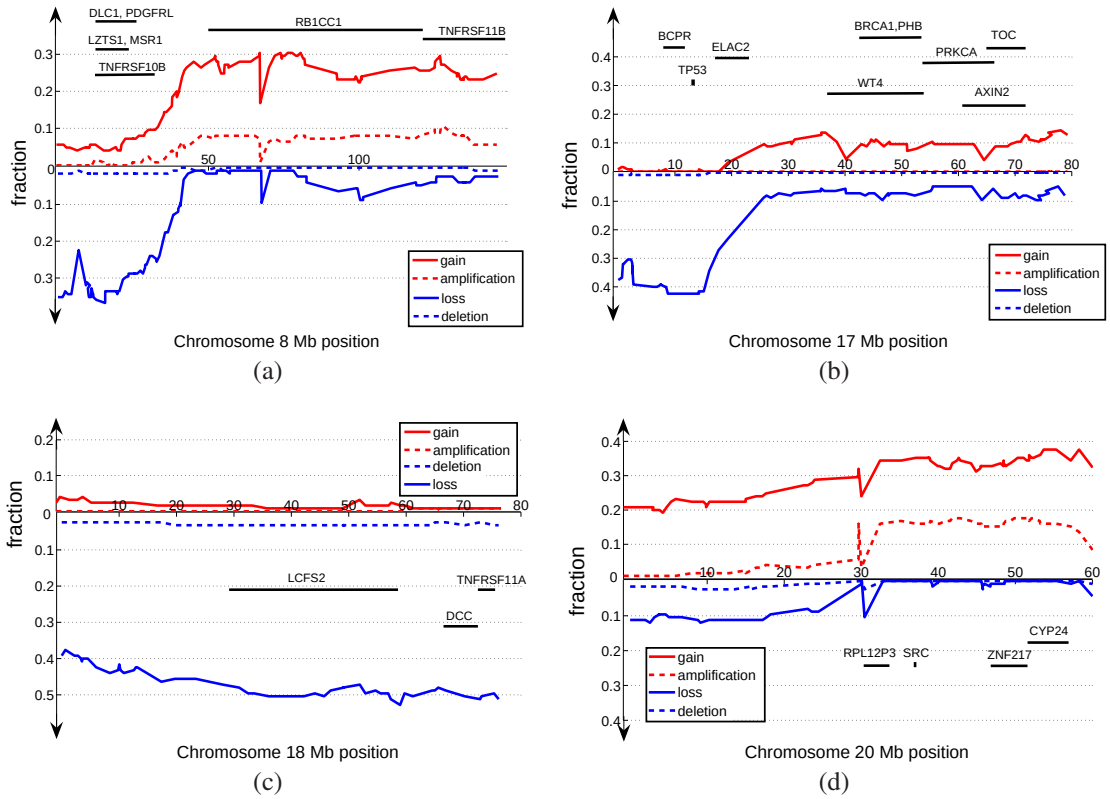


Figure 12: Detailed spectrum of GIM rates over 125 Colorectal cancer patients in 4 hotspot region. In each panels, the upper half shows gain and amplification rates, and the lower half shows loss and deletion rates. The cancer related genes in the vicinity of the study regions are annotated on all the graphs.

hotspot on 8q, which suggests that these two genes might be oncogenes related to cancer growth. Indeed RB1CC1 is a DNA-binding transcription factor. It has been identified as a key regulator of the tumor-suppressor gene RB1 (32), and its sequence spans a wide region on chromosome 8 (from 8p11 to 8q22), which begins at the loss-to-gain transition sites. This correspondence offers a possible explanation to the fact that the loss-to-gain transition on chromosome 8 does not happen at the centromere.

Figure 12b and 12c show the annotations of the cancer related genes on chromosome 17 and chromosome 18, respectively, along with their GIM-rate spectrum. As reported in the previous section, the 17p region was found to suffer high-frequency (i.e., (54.4%) lost/deletion. It is noteworthy that this LD-hotspot coincides with the P53 and BCPR genes. P53 is a well-known tumor suppressor related to many types of cancers, such as cancers of lung, breast, esophagus, and colon

(33). BCPR is known as a breast cancer-related regulator of P53 (34). The loss of these two genes as detected by GIMscan from aCGH profiles provide additional validation of their involvement colorectal cancer. Chromosome 18 harbors large LD-hotspot regions spanning both 18p and 18q. Loss or deletion were found in 70 out of 125 individuals on 18p, and 73 out of 125 individuals on 18q. We also found three cancer-related genes mapping to this region (Figure **??**), including DCC, LCFS2 and TNFRSF11A. Among these genes, DCC was found to be deleted in colorectal carcinoma, and it acts as a colorectal cancer suppressor (35). LCFS2 and TNFRSF11A were also found to be deleted in gastric and esophageal adenocarcinoma (36).

Another high-frequency GIM hotspot we identified resides in chromosome 20, where dosage-state alterations were found in 88 individuals representing 70.40% of all the study population. Highly frequent gains (48.0%) or amplifications (17.6%) was observed on 20q. The regions on 20q that were gained or amplified were relatively contiguous except for one high magnitude amplification region centered at 30Mb (12c). The contiguous region from 35Mb to the telomere harbors three cancer-related genes, SRC, ZNF217 and CYP24, which are candidate oncogenes. SRC encodes pp60c-Src (or c-Src), which is a 60 kDa nonreceptor tyrosine kinase and is the cellular homologue to the potent transforming v-Src viral oncogene (37). ZNF217 encodes a Kruppel-like transcription factor that has been demonstrated to promote the immortalization of human mammary cells and to play a role in the suppression of apoptosis (38). Over-expression of CYP24 may abrogate growth control mediated by vitamin D or calcium (39). Centered at 30Mb, high-magnitude amplification was observed with high frequency. This region harbors RPL12P3 (30Mb - 33Mb), a ribosomal protein L12 pseudogene. Although the function of this gene is still unclear, our high resolution analysis suggests that it may be over-expressed in tumor cell and is related with cancer growth.

Table 1 summarizes the aforementioned candidate cancer-related genes and the average length-fraction of their spanning regions on the genome that exhibits GIM.

To explore possible correlations between GIM and some phenotypic characteristics of the study patient population, we test the extent of GIM (i.e., the genotype), as measured by the proportion of genomic regions harboring GIM over the total length of the genome, against empirical frequencies

Gene name	Chr	Location (Mb) Start / End	Fraction of gain	Fraction of loss	Fraction of amplification	Fraction of deletion	Related diseases
LZTS1	8	13.74 / 22.23	0.044	0.330	0.002	0.016	Esophageal squamous cell carcinoma
MSR1	8	13.74 / 22.23	0.044	0.330	0.002	0.016	Prostate cancer
DLC1	8	13.74 / 24.79	0.055	0.315	0.001	0.016	Colorectal cancer
PDGFRL	8	13.74 / 24.79	0.055	0.315	0.001	0.016	Hepatocellular cancer Colorectal cancer
TNFRSF10B	8	13.74 / 30.68	0.065	0.300	0.005	0.016	Squamous cell carcinoma
RB1CC1	8	46.13 / 113.58	0.265	0.035	0.076	0.001	Breast cancer
TNFRSF11B	8	113.58 / 140.93	0.257	0.034	0.082	0.001	Paget disease
BCPR	17	8.92 / 12.15	0.000	0.424	0.000	0.008	Breast cancer
ELAC2	17	17.35 / 23.25	0.096	0.088	0.000	0.000	Prostate cancer
WT4	17	37.15 / 53.31	0.089	0.070	0.000	0.000	Wilms tumor
BRCA1	17	42.07 / 53.31	0.098	0.073	0.000	0.000	Breast cancer-1 Ovarian cancer
PHB	17	42.07 / 53.31	0.098	0.073	0.000	0.000	Breast cancer
PRKCA	17	53.31 / 65.16	0.070	0.071	0.000	0.000	Pituitary tumor
AXIN2	17	60.33 / 71.43	0.091	0.086	0.000	0.000	Colorectal cancer Oligodontia-colorectal cancer
TOC	17	64.55 / 71.43	0.104	0.084	0.000	0.000	Tylosis with esophageal cancer
LCFS2	18	29.16 / 58.58	0.013	0.498	0.000	0.032	Lynch cancer family syndrome II
DCC	18	66.51 / 72.49	0.008	0.504	0.008	0.032	colorectal carcinoma
TNFRSF11A	18	72.49 / 75.50	0.008	0.510	0.008	0.030	Osteolysis
RPL12P3	20	30.81 / 33.50	0.334	0.000	0.168	0.000	Unknown
ZNF217	20	47.75 / 52.06	0.348	0.000	0.172	0.000	Breast Cancer
CYP24	20	52.06 / 56.90	0.361	0.001	0.160	0.000	Breast Neoplasms

Table 1: Summary of cancer related genes on chromosome 8, 17, 18 and 20. The fraction of aberrations were averaged over the region the gene spans.

Phenotype		Total number of individuals	% Genome gained	% Genome lost	% Genome amplified	% Genome deleted
Stage	1	11	9.89	3.45	0.51	0.00
	2	37	7.83	10.55	1.41	0.80
	3	35	9.23	9.92	0.42	0.00
	4	38	9.54	12.48	0.92	0.46
Location	Right	38	8.14	8.50	0.66	0.47
	Left	85	9.11	10.91	0.97	0.38
Age	$<$ 50 years	10	3.58	9.39	1.93	1.21
	50-70 years	47	9.29	10.48	0.92	0.44
	$>$ 70 years	66	9.26	10.06	0.68	0.26
Sex	Male	64	10.18	10.03	1.03	0.59
	Female	59	7.33	10.31	0.71	0.20
Bat26	Stable	102	9.22	10.54	0.99	0.49
	Unstable	7	3.59	6.34	0.00	0.00

Table 2: Fractions of the genome altered by clinical phenotype.

Phenotype			p-value (gain)	p-value (loss)	p-value (amplification)	p-value (delete)
Stage	1	2	0.539	**0.061**	0.204	0.251
		3	0.848	**0.033**	0.841	0.662
		4	0.922	**0.042**	0.565	0.697
	2	3	0.539	0.812	**0.014**	0.180
		4	0.515	0.538	0.381	0.703
	3	4	0.917	0.381	0.247	0.593
Tumor location	Right	Left	0.640	0.270	0.418	0.744
Age	$<$ 50 years	50-70 years	**0.059**	0.812	0.194	0.266
		$>$ 70 years	0.111	0.837	**0.044**	0.143
	50-70 years	$>$ 70 years	0.988	0.818	0.487	0.681
Sex	Male	Female	0.143	0.900	0.3740	0.302
Bat26	Stable	Unstable	0.128	0.368	0.123	0.169

Table 3: P-values reflecting statistical significance of possible correlations between the phenotype and GIM. P-values less than 0.1 are indicated by bold.

of phenotypes of interest in the study population. We examined 5 phenotypes, cancer state, location, patient age, sex, and Bat26. Table 2 summarizes the empirical frequencies of each phenotype, and the average extent of GIM of the sub-population bearing the phenotype. To compute the statistical significance of possible correlations between the phenotype and the genotype, we conducted an unpaired student t-test followed by a MaxT test using permutation analysis to control family-wise false positive error rates (40). The test scores were shown in Table 3. Overall, in most cases there is no significant correlation between the phenotypes we tested and the GIM genotypes, which is not surprising because the phenotypes available for our test are very generic and not particularly informative from a pathological stand point. But there are some interesting exceptions that worth further investigation. For example, while the extent of genome loss in cancer stage 2, 3 and 4 did not exhibit significant difference, genome loss is significantly less severe in stage 1 than in other states. Genome amplification is significantly more severs in stage 2, than in stage 3. There is not a significant different in the GIM genotypes (i.e. the extend of all types of GIM) between stage 3 and 4, suggesting that chromosomal instability is an early event in colorectal carcinogenesis (41). It is noteworthy that significantly more fraction of the genome has amplification in patients whose ages are below 50 than those older than 70.

3 Discussion

Genetic instability represents an important type of biological markers for cancer and many other diseases. However, experimental array CGH data currently available for examining genetic instability are often corrupted by severe noises resulted from both exogenous (e.g., experimental conditions) and endogenous (i.e., DNA contents) origin. In particular the variations of hybridization signal intensities even within one specific GIM state can seriously confound data interpretation. Extant computer-aided methods for determining GIMs from raw aCGH data are limited in their ability to interpret aCGH data compromised such noises and measured over the whole genome and over different individuals. In this paper, we present a new statistical method, Genome Imbalance Scanner, for automatically decoding the underlying DNA dosage-states from aCGH data. GIMscan fits a hidden hybridization trajectory to each state. It employs a hidden switching process to

stochastically select for each examined clone on the genome its underlying trajectory (therefore the dosage state) that have generated the observed CGH data. Our method captures both the intrinsic (nonrandom) spatial change of genome hybridization intensities, and the prevalent (random) measurement noise during data acquisition; and it simultaneously segments the chromosome and assigns different states to the segmented DNA. We tested the proposed method on both simulated data and real data measured from a colorectal cancer population, and we report competitive or superior performance of GIMscan in comparison with popular extant methods.

Genome alterations can be classified into long range aberration in which consecutive clones have same altered underlying state and short range aberration in which only single clone changes. Empirically, we noticed that GIMscan yields much fewer predictions on point alteration of gene dosage-state (i.e., copy number change in a single clone out of a sequence) compared to other methods. As discussed in the introduction, while single clone copy number changes is non uncommon in aberrant genomes, common genome rearrangement mechanisms tend to produce genome imbalances over certain span on the chromosome. Therefore a reasonable balance between sensitivity to these two types of abbreviations is crucial to the accuracy of GIM detection. Extend algorithms, especially the threshold-based methods are very sensitive to local fluctuations of LR signals, and tend to predict a large number of spurious point GIMs. GIMscan makes use of the spatial dependency between adjacent clones, which smooths the spatial variations in the long-range aberrations, nevertheless it remains sensitive to significant short range aberrations. For example, in Figure 9a, the single amplified clone near the telomere of 11q (yellow circle) was significantly far from its context clones. Therefore, GIMscan classified this clone into amplification state (yellow arrow). Similarly, the clone near the centromere of chromosome 7 (purple circle) was also detected by GIMscan to be loss state (purple circle), different from its context clones. In contrary, the clone on chromosome 9 (black circle) was classified to be within the gain state, though they tended to be higher than their neighboring clones. GIMscan determined these two clones as noises of the gain state because their LR values are not significantly different from their context. Such a decision is very hard to reach when using methods (such as an HMM) that model each dosage-state with spatially invariant state-specific LR distributions.

Specifying the underlying state (loss, gain, normal) for one clone is not a straightforward task, not to mention discovering the exact dosage state for each clone. Although many method such as HMM used "gain state" or "loss state" to refer to the different states, the true DNA dosage states, strictly speaking, are unknown, a problem shared by all models simultaneously segmenting the genome and annotating the DNA dosage state. This is because for the selected states we do not have a "reference state", and the assigned states only have relative meaning (e.g. we can only say state 1 has more copy number than state 2). In order to specify the exact DNA copy number to each clone, further biological experiments, such as PCR, is necessary. In GIMscan, we require the chromosomes of one individual to share the same number of trajectories and the means and variances of the starting values of these trajectories because the normal cell contamination effect within each experiment (individual) is expected to be similar and the observed LR values for the same state in one individual should have the similar expected value. According to the learned means and variances of the trajectories, we manually assign states to the trajectories for each individual by human expertise.

We will make GIMscan available to interested users via a web-interface and maintain and upgrade it regularly. We expect that with improved GIM screening of aberrant genomes related to various disease of interest, it is possible to perform more comprehensive and reliable statistical analysis of associations between dosage-state alterations and phenotype abnormalities. In particular, array CGH provides a powerful tool to analyze the DNA dosage-state of transcriptional regulation regions, which make it possible to conjoin microarray-based high-throughput gene expression profiling with genome imbalance screening. Association studies of these two patterns can offer a more complete picture of the (alteration of) gene regulation mechanisms underlying interested biological processes such as carcinogenesis.

4 Materials and Methods

In this section we present the modeling, algorithmic and implementational choices that we made in GIMscan.

4.1 Switching Kalman Filters: Model and Adaptation to aCGH Analysis

For each specific gene dosage state, we model the spatial drift of its hybridization signal intensities using a hidden trajectory and model the uncertainty in LR measurements using a zero-mean Gaussian noise. This corresponds to a standard dynamic model named Kalman filter (KF), or state-space model (SSM). Observed LR values arise as a mixture of the outputs of state-specific Kalman filters. The mixing proportion, modeled as latent variables indicating gene dosage states, is also spatially dependent as captured by the HMM-based methods using a hidden Markov state-transition process (or switching process). Now we have multiple linear Kalman filters controlled by a dynamic switching process, which can be formulated as a factored switching Kalman filters (SKF), also known as a switching state-space model (SSSM).

SKFs generalize HMMs and KFs by relaxing inherent assumptions made individually in both models and allowing continuous, non-Gaussian posterior distributions over latent trajectory. However, exact inference of the posterior distribution over the hidden states in an SKF is no longer tractable. Ghahramani and Hinton (42) approximated this intractable posterior by a tractable "variational" distribution over decoupled hidden trajectory processes and switching process, which is parameterized by an additional set of "variational parameters". Iterative updates of the variational parameters minimize the KL divergence between variational and true posterior distributions. Dosage-state annotation and clone-sequence segmentation can be solved jointly by computing expectations of switching-state variables under the approximate posterior. Parameter estimation can be derived from the EM algorithm which iteratively computes some expected values from the approximate posterior (E steps) and re-estimates the parameters of the model (M step).

SKFs have been applied to model regime-switching sequential data in multiple contexts: physiological time series such as the chest volume of a patient with sleeping apnea characterized by periods of normal breathing, gasping, and no-breathing (42), financial time series such as currency exchange rate data for which the model not only uncovers underlying regimes but also provides online prediction with confidence interval to facilitate risk estimation in financial engineering (43), and time series in signal processing such as acoustic segmentation in speech recognition (44), mobile robots localization using state and sensory data (45), on-line inference of hand-kinematics

from the firing rates of a population of motor cortical neurons (27).

Our proposed method, GIMscan (Genome IMbalance SCANner), adapts the SKF model to whole-genome analysis of aCGH data by allowing a parameter sharing scheme among multiple chromosomes and multiple individuals which makes best use of data. Model selection and further extension of the model are also summarized or discussed briefly at the end of this section.

4.2 The dosage-state-specific Kalman filter

The dosage-state-specific Kalman filter is a linear chain graphical model with a backbone of hidden real-valued variables (trajectory) emitting a series of real-valued observation (Figure 13(a)). The transition model is linear and subject to Gaussian noise which reflect the evolving hybridization densities. The observation model imposes a Gaussian noise arising in the measurement stage to generate the LR ratio at each position (clone). This model for dosage state m can be formulated as:

$$P(X_t^{(m)}|X_{t-1}^{(m)}) \sim \mathcal{N}(a^{(m)}X_{t-1}^{(m)}, b^{(m)})$$
$$P(Y_t^{(m)}|X_t^{(m)}) \sim \mathcal{N}(X_t^{(m)}, r)$$

where $X_t^{(m)}$ is the hidden real-valued variable at position t (t^{th} clone) on the trajectory, $Y_t^{(m)}$ is the observed real-valued variable emitted by $X_t^{(m)}$. The parameters $a^{(m)}, b^{(m)}, r$ are all position-invariant; r determines the degree of uncertainty in observation measurements. We also assume the initial value of the hidden trajectory, $X_1^{(m)}$, has a Gaussian distribution $P(X_1^{(m)}) \sim \mathcal{N}(\mu^{(m)}, \sigma^{(m)})$. All the variables and parameters are univariate. Computing the online (filtering) and offline (smoothing) versions of the posterior probabilities of the hidden state variable at position t, $P(X_t^{(m)}|Y_{1:t}^{(m)})$ and $P(X_t^{(m)}|Y_{1:T}^{(m)})$, where $Y_{1:t}^{(m)}$ denotes the observed variables up to position t and $Y_{1:T}^{(m)}$ denotes all the observed variables, are the most important operations associated with the model. The probabilities are all normal-distributed and the computation is tractable because of the conjugacy of the normal distribution to itself. Computation of the offline probabilities for all possible t will be part of the variational inference discussed in Section 4.4 in which we decouple the model to a number

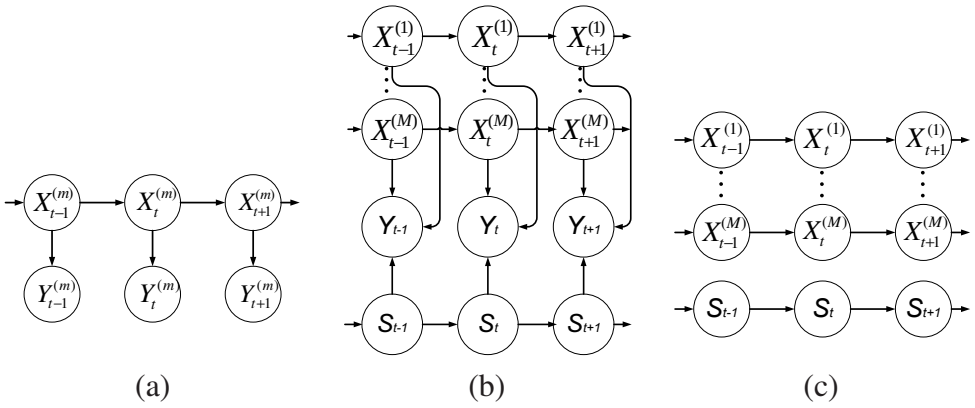


Figure 13: (a) Graphical structure of dosage-state-specific Kalman filter for dosage state m. $X_t^{(m)}$ is the hidden variable at clone t on the trajectory, and $Y_t^{(m)}$ representing the corresponding observed variable of the Kalman filter. (b) Graphical structure of the switching Kalman filter (SKF) model. The model consists of M linear chains of Kalman filters ($X_{1:T}^{(1:M)}$), a Markov chain of switching processes ($S_{1:T}$) and a series of observed variables ($Y_{1:T}$). (c) Graphical structure of the uncoupled model which represents the tractable subfamily of distributions to approximate the posterior distribution of the SKF model.

of tractable linear chains.

4.3 SKF formulation for aCGH analysis

Given the dosage-state-specific Kalman filter for M dosage states, a switching Kalman filter generates the LR value at each position from one of the outputs. The choices are spatial-dependent and are determined by a discrete switching process which evolves according to Markov dynamics. The generation of the observed LR values in SKF can be formulated as:

$$P(S_t^{(m)} = 1 | S_{t-1}^{(n)} = 1) = \phi_{mn}$$

$$Y_t = \sum_{m=1}^{M} Y_t^{(m)} S_t^{(m)}$$

where S_t is the M-dimensional multinomial switching variables for clone t following $1 \times M$ binary coding scheme, Y_t is the observed variable representing LR value of clone t. $\Phi = \{\phi_{mn}\}$ is the transition matrix for the hidden Markov dynamics. The initial switching state follows a multinomial distribution parameterized by π: $S_1 \sim \text{Multinomial}(1, \pi)$. The observation Y_t is a mixture

of individual KF outputs $Y_t^{(1:M)}$, and is deterministic given $Y_t^{(1:M)}$ and the mixing proportion S_t. Alternatively, we could save the variables $Y_t^{(1:M)}$ and generate the observation from the M hidden linear Gaussian trajectories as:

$$P(Y_t|X_t^{(1:M)}, S_t) \sim \mathcal{N}(\sum_{m=1}^{M} X_t^{(m)} S_t^{(m)}, r)$$

The graphical structure of the SKF model is shown in Figure 13 (b). The joint probability of SKF can be written as:

$$P(Y_{1:T}, X_{1:T}^{(1:M)}, S_{1:T}) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{m=1}^{M} P(X_1^{(m)}) \prod_{t=2}^{T} P(X_t^{(m)}|X_{t-1}^{(m)}) \prod_{t=1}^{T} P(Y_t|X_t^{(1:M)}, S_t)$$

4.4 Variational inference and Parameter Estimation

To facilitate dosage state annotation and clone-sequence segmentation, the posterior probabilities $P(S_t|Y_{1:T})$ need to be computed for $t = 1, \ldots, T$. However, exact inference of this posterior probability is intractable (42). We employed a variational algorithm (42) which approximates the posterior probabilities $\mathcal{P}$ with a parameterized distribution $\mathcal{Q}(\mathbf{v})$ from some tractable subfamily of distributions. It iteratively updates the values of the variational parameters $\mathbf{v}$ to minimize a measure of "distance" between the approximate posterior distribution and the true posterior distribution: the Kullback-Leibler divergence $KL(\mathcal{Q} \parallel \mathcal{P})$. The choice of the tractable subfamily for the SKF model is a discrete Markov chain and M uncoupled KFs (Figure 13 (c)). Two sets of variational parameters are introduced for the Markov chain and KFs respectively. Their updates can be carried out using fix-point equations (42), which include terms of expectations under the approximate posterior. They can be computed using the forward-backward algorithm (for the Markov chain) and Kalman filtering and smoothing algorithms (for the M KFs) in polynomial time. These updates maintain or increase a lower bound of the log likelihood of the model and converge in a few iterations empirically. The fast rate of convergence is mainly due to low data dimension: LR ratios are univariate.

Parameter estimation is performed under the EM framework. The E step employs the varia-

tional inference algorithm to find the best approximate posterior via iterative updates of the variational parameters. The M step reestimates the model parameters Θ to maximize the same lower bound of log-likelihood in variational inference. This reestimation can be performed exactly by zeroing the derivatives with respect to the model parameters, which also involve computing expectations under the approximate posterior as we discussed above. Parameter estimation is implemented by a coordinate ascent procedure which maximize the lower bound $\mathcal{B}(\mathbf{v}, \mathbf{\Theta})$ by iteratively updating variational parameters $\mathbf{v}$ (E steps) and model parameters Θ (M steps).

4.5 Parameter sharing: adapting SKF to whole-genome analysis

We introduce the model framework of GIMscan (Genome IMbalance SCANner). We first describe the settings for a whole-genome analysis. The aCGH dataset are collected from experimental data of J individuals, the genome of which consists of K chromosomes, and chromosome k contains T_k clones. The LR values $Y_{1:T,j,k}$ on individual j, chromosome k are generated by a SKF model with hidden trajectory $X_{1:T,j,k}$ and switching states $S_{1:T,j,k}$.

As we described in Section 4.3, the model parameters of SKF include $\mu^{(m)}, \sigma^{(m)}, a^{(m)}, b^{(m)}, r, \pi$ and ϕ. Each of these parameters in the SKF model delineates one property behind the aCGH data. $\mu^{(m)}$ and $\sigma^{(m)}$ are the mean and variance of Gaussian distribution of the starting clone on hidden trajectory for dosage state m. $a^{(m)}$ and $b^{(m)}$ determine the transition model of that trajectory which dictate the spatial drift of the signal intensities. r is the variance of the Gaussian accounting for the noise introduced in the experiment stage, and is independent of the hidden dosage-state. Lastly, π and Φ are initial state parameters and transition matrix for the discrete switching process between different dosage states.

We are now ready to describe the parameter sharing scheme used in GIMscan for the analysis of whole genome aCGH data. We consider two groups of parameters. Firstly, we let $\mu^{(m)}$, $\sigma^{(m)}$, r, π and Φ be shared across all chromosomes of one particular individuals. Mainly due to the normal cell contamination, the magnitude of starting value for the trajectory of one particular state varies across different individuals. Different $\mu^{(m)}$ and $\sigma^{(m)}$ for different individuals can account for this "un-normalized" starting value of the trajectory of one state. r is also shared by chromo-

somes from one individual because one individual corresponds to one experiment, and different experiments may have different noise levels. π and Φ are shared by one individual because the number of dosage states are individual-specific: different individuals may have different number of dosage states. The second group of parameters, $a^{(1:M)}$ and $b^{(1:M)}$, is assumed to be shared by one particular chromosome of all individuals, because the physical-chemical properties (e.g. the base composition) of one particular chromosome of different individuals (the same tumor cell line of the same species) are very similar. This similarity leads to similar hybridization signal intensity across one chromosome. The parameter sharing scheme we introduced does not change the approximate inference algorithm which only updates the variational parameters. Reestimate of the model parameters under the EM framework is similar and it is straightforward to derive the update equations.

The maximum number of hidden trajectories or dosage states (M in the model specification) one individual can have remains to be determined. we employ Gaussian mixture model using penalized likelihood criteria such as AIC to select the number of states for each individual.

The SKF model can be naturally extended by considering multiple switching chains which simultaneously determine the functioning parents for the observed variables. We can also add more variables in each slice in order to make use of the information of the distance between adjacent clones and the sequence length of each clone. For simplicity, we do not consider these extensions in this paper.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. DBI-0546594.

References

[1] Nakao K, Mehta K, Fridlyand J, Moore D, Jain A, et al. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic

hybridization. In: Carcinogenesis. volume 25(8).

[2] Diep C, Teixeira M, Thorstensen L, Wiig J, Eknas M, et al. (2004) Genome characteristics of primary carcinomas, local recurrences, carcinomatoses, and liver metastases from colorectal cancer patients. Molecular Cancer 3:6.

[3] Pinkel D, Albertson D (2005) Array comparative genomic hybridization and its applications in cancer. Nat Genet 37:S11–S17.

[4] 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 20:207–211.

[5] Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A (2004) Hidden markov models approach to the analysis of array cgh data. J Multivariate Anal 90(1):132–153.

[6] Hodgson G, Hager J, Volik S, Hariono S, Wernick M, et al. (2001) Genome scanning with array cgh delineates regional alterations in mouse islet carcinomas. Nat Genet 29:459–464.

[7] Eilers P, De Menezes R (2005) Quantile smoothing of array cgh data. Bioinformatics 21(7):1146–1153.

[8] Hsu L, Self S, Grove D, Randolph T, Wang K, et al. (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6(2):211–226.

[9] Myers C, Dunham M, Kung S, Troyanskaya O (2004) Accurate detection of aneuploidies in array cgh and gene expression microarray data. Bioinformatics 20(18):3533–3543.

[10] Olshen A, Venkatraman E, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array based dna copy number data. Biostatistics 5(4):557–572.

[11] Jong K, Marchiori E, Meijer G, Vaart A, Ylstra B (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. Bioinformatics 20(18):3636–3637.

[12] Picard F, Robin S, Lavielle M, Vaisse C, Daudin J (2005) A statistical approach for array cgh data analysis. BMC Bioinformatics 6(1):27.

[13] Hupe P, Stransky N, Thiery J, Radvanyi F, Barillot E (2004) Analysis of array cgh data: from signal ratio to gain and loss of dna regions. Bioinformatics 20(18):3413–3422.

[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(1):45–58.

[15] Daruwala R, Rudra A, Ostrer H, Lucito R, Wigler M, et al. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. P Natl Acad Sci USA 101(46):16292–16297.

[16] Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation to array cgh data for downstream analyses. Bioinformatics 21(22):4084–4091.

[17] Marioni J, Thorne N, Tavare S (2006) Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. Bioinformatics 22(9):1144–1146.

[18] Broet P, Richardson S (2006) Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. Bioinformatics 22(8):911–918.

[19] Shah S, Xuan X, De Leeuw R, Khojasteh M, Lam W, et al. (2006) Integrating copy number polymorphisms into array cgh analysis using a robust hmm. Bioinformatics In press.

[20] Snijders A, 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.

[21] Schimke R, Sherwood S, Hill A, Johnston R (1986) Overreplication and recombination of dna in higher eukaryotes: Potential consequences and biological implications. PNAS 83:2157–2161.

[22] Kraus E, Leung W, Haber J (2001) Break-induced replication: A review and an example in budding yeast. PNAS 98:8255–8262.

[23] Toledo F, Roscouet D, Buttin G, Debatisse M (1992) Co-amplified markers alternate in megabase long chromosomal inverted repeats and cluster independently in interphase nuclei at early steps of mammalian gene amplification. The EMBO Journal 11:2665–2673.

[24] Newkirk H, Knoll J, Rogan1 P (2005) Distortion of quantitative genomic and expression hybridization by cot-1 dna: mitigation of this effect. Nucleic Acids Res 33(22):e191.

[25] Engler D, Mohapatra G, Louis D, Betensky R (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. Biostatistics 7(3):399–421.

[26] Murphy K (1998) Learning switching kalman filter models. Compaq Cambridge Research Lab Tech Report 98-10 .

[27] Wu W, Black M, Mumford D, Gao Y, Bienenstock E, et al. (2003) A switching kalman filter model for the motor cortical coding of hand motion. Proceedings of the 25th Annual International Conference of the IEEE EMBS 3:2083–2086.

[28] Manfredi V, Mahadevan S, Kurose J (2005) Switching kalman filters for prediction and tracking in an adaptive meteorological sensing network. Sensor and Ad Hoc Communications and Networks :197–206.

[29] Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716C723.

[30] Yuan B, Miller M, Keck C, Zimonjic D, Thorgeirsson S, et al. (1998) Cloning, characterization, and chromosomal localization of a gene frequently deleted in human liver cancer (DLC-1) homologous to rat RhoGAP. Cancer Res 58(10):2196–2199.

[31] Yuan B, Jefferson A, Baldwin K, Thorgeirsson S, Popescu N, et al. (2004) Dlc-1 operates as a tumor suppressor gene in human non-small cell lung carcinomas. Oncogene 23(7):1405–1411.

[32] Chano T, Ikegawa S, Kontani K, Okabe H, Baldini N, et al. (2002) Identification of rb1cc1, a novel human gene that can induce rb1 in various human cells. Oncogene 21(8):1295–1298.

[33] Harris C (1996) p53 tumor suppressor gene: at the crossroads of molecular carcinogenesis, molecular epidemiology, and cancer risk assessment. Environ Health Perspect 104(Suppl 3):435–439.

[34] Coles C, Thompson AM, Elder PA, Cohen BB, Mackenzie IM, et al. (1996) Evidence implicating at least two genes on chromosome 17p in breast carcinogenesis. Lancet 336:761–763.

[35] Mehlen P, Rabizadeh S, Snipas S, Assa-Munt N, Salvesen G, et al. (1998) The DCC gene product induces apoptosis by a mechanism requiring receptor proteolysis. Nature 395(6704):801–804.

[36] Stocks S, Pratt N, Sales M, Johnston D, Thompson A, et al. (2001) Chromosomal imbalances in gastric and esophageal adenocarcinoma: specific comparative genomic hybridization-detected abnormalities segregate with junctional adenocarcinomas. Genes Chromosomes Cancer 32(1):50–58.

[37] Dehm S, Bonham K (2004) SRC gene expression in human cancer: the role of transcriptional activation. Biochem Cell Biol 82(2):263–274.

[38] Collins C, Volik S, Kowbel D (2001) Comprehensive genome sequence analysis of a breast cancer amplicon. Genome Res 11:1034–1042.

[39] Albertson D, Ylstra B, Segraves R, Collins C, Dairkee S, et al. (2000) Quantitative mapping of amplicon structure by array cgh identifies cyp24 as a candidate oncogene. Nat Genet 25:144–146.

[40] Olshen A, Jain A (2002) Deriving quantitative conclusions from microarray expression data. Bioinformatics 18:961–970.

[41] Shih I, Zhou W, Goodman S, Lengauer C, Kinzler K, et al. (2001) Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. Cancer Res 61:818–822.

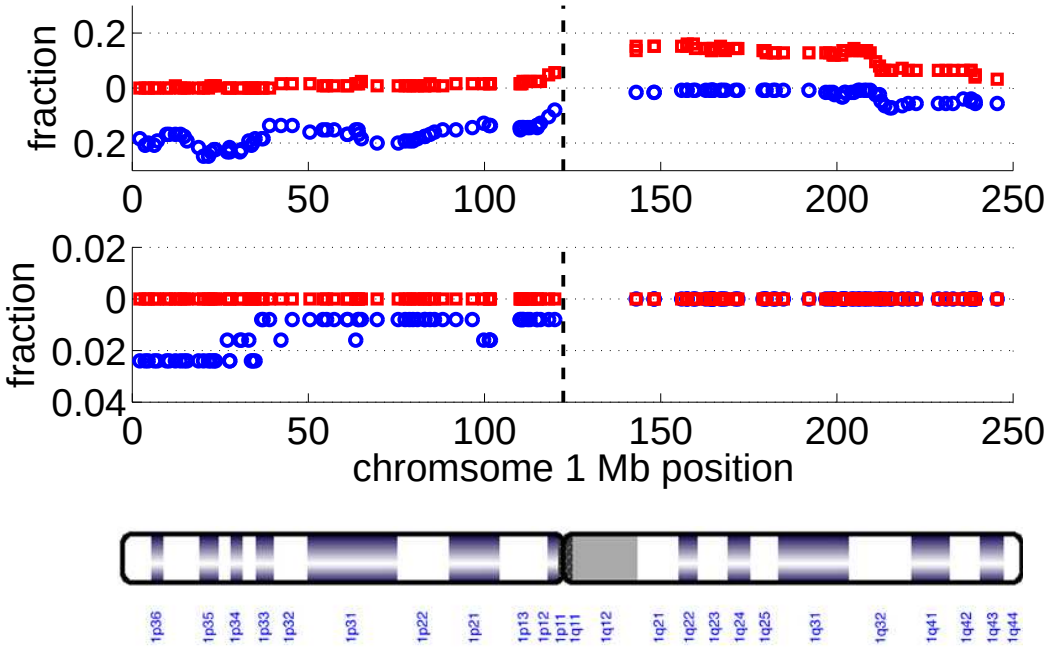
[42] Ghahramani Z, Hinton G (1998) Variational learning for switching state-space models. Neural Comput 12(4):963–996.

[43] Azzouzi M, Nabney I (1999) Modelling financial time series with switching state space models. Proceedings on IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering :240–249.

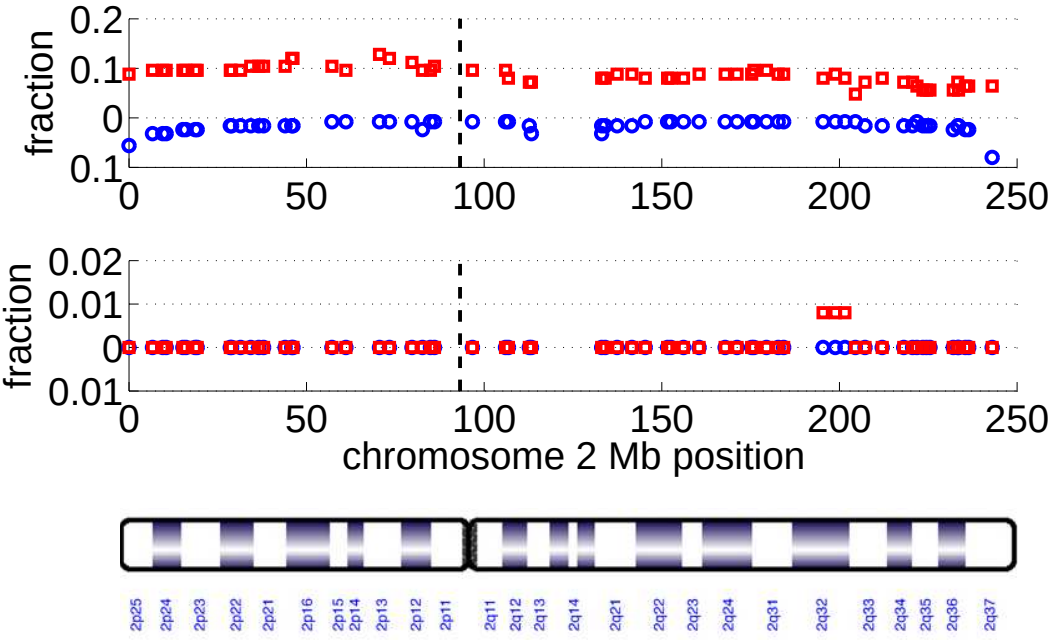
[44] Zheng Y, Hasegawa-Johnson M (2003) Acoustic segmentation using switching state kalman filter. Proceedings of ICASSP'03 1:752–755.

[45] Baltzakis H, Trahanias P (2003) A hybrid framework for mobile robot localization formulation using switching state-space models. Autonomous Robots 15:169–191.

A Genetic Imbalance Patterns of the Genome

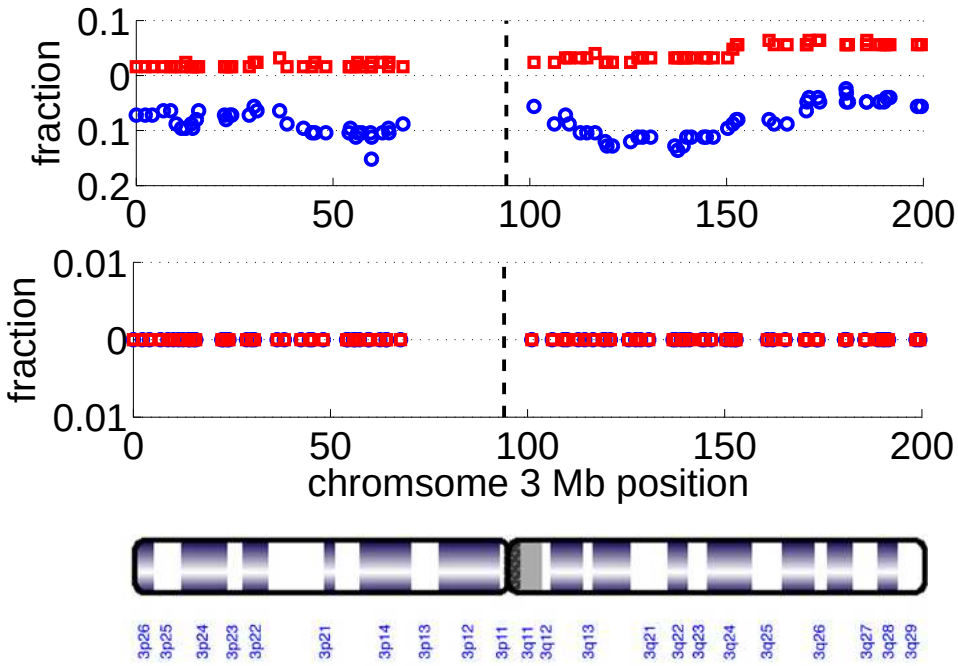


(a)

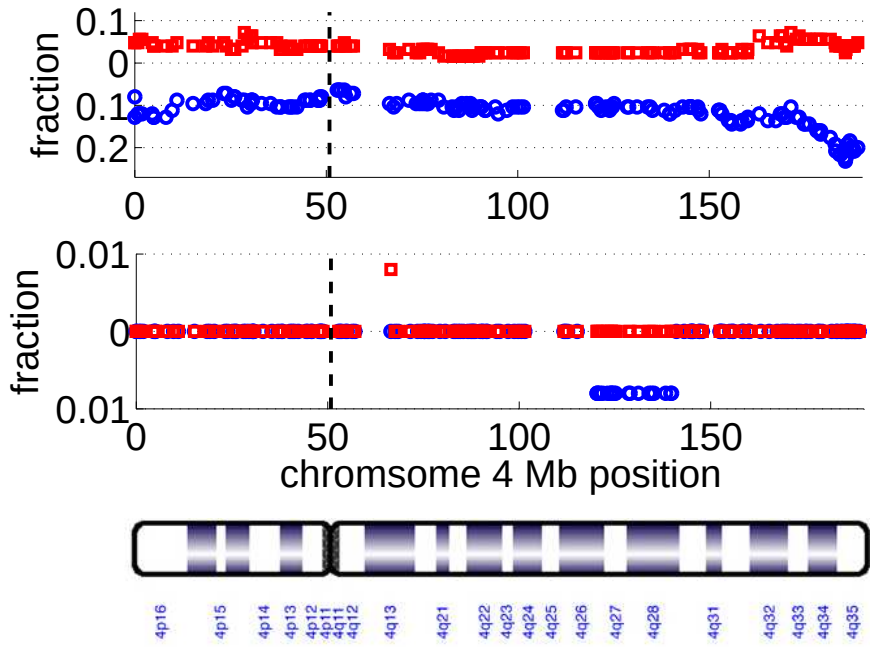


(b)

Figure 14: Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

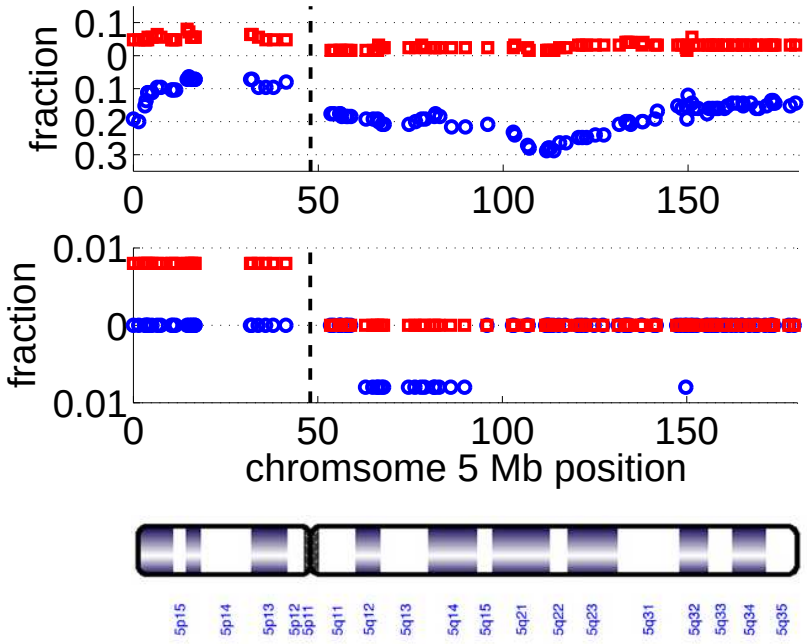


(c)

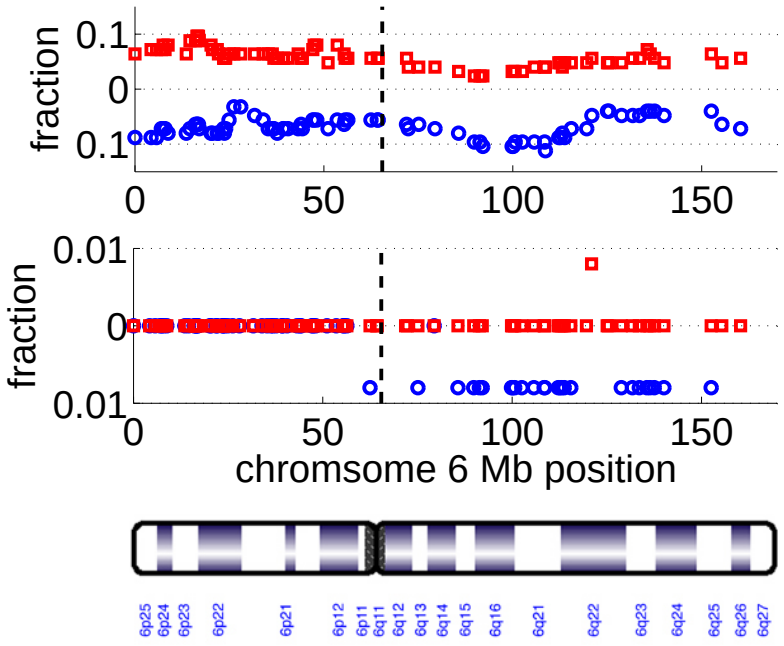


(d)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

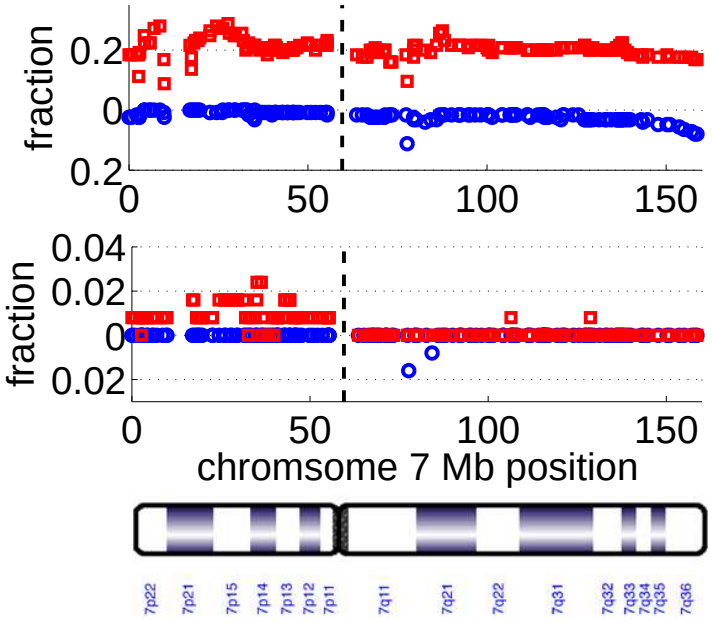


(e)

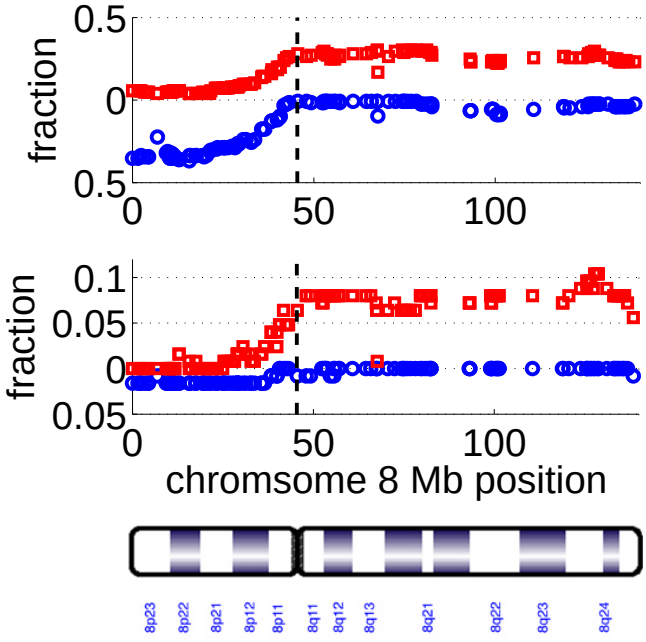


(f)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

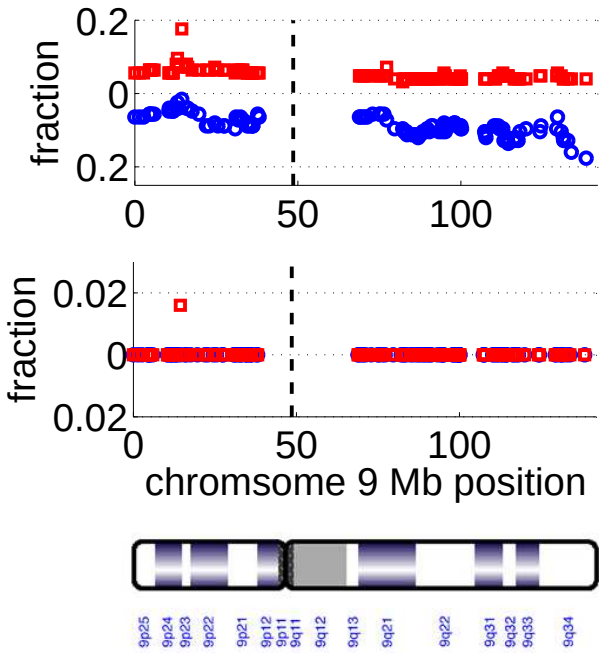


(g)

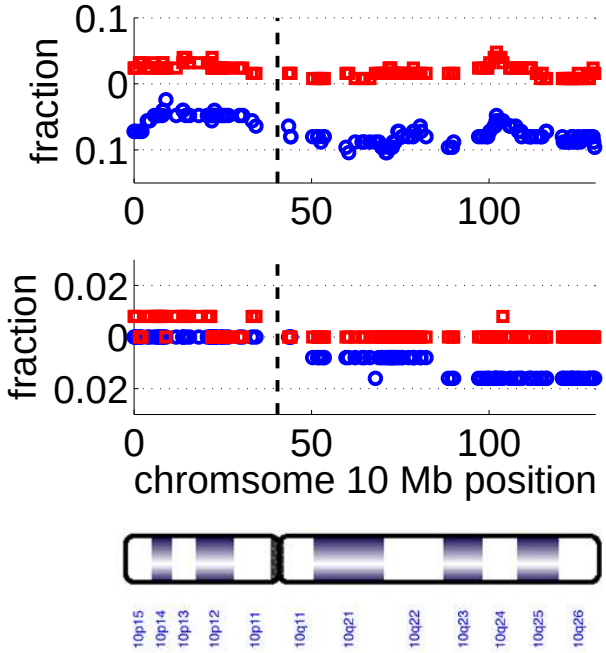


(h)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

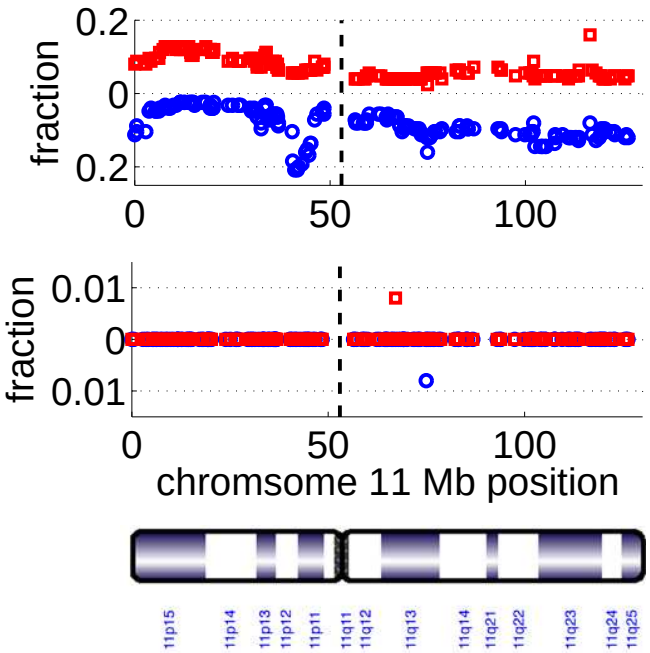


(i)

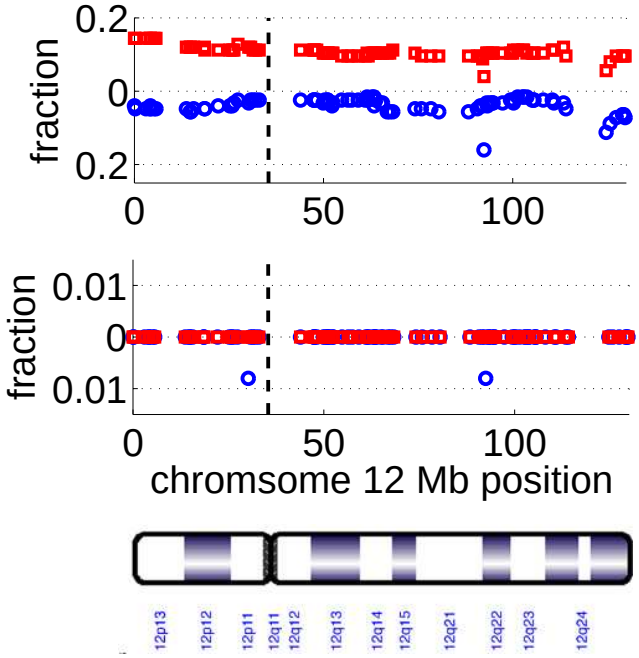


(j)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

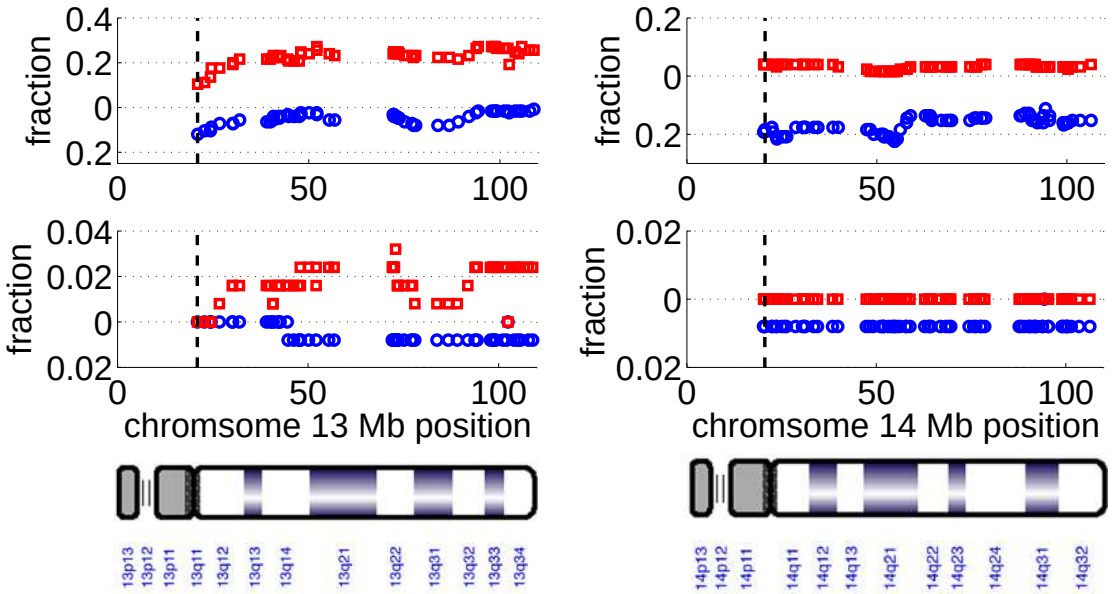


(k)



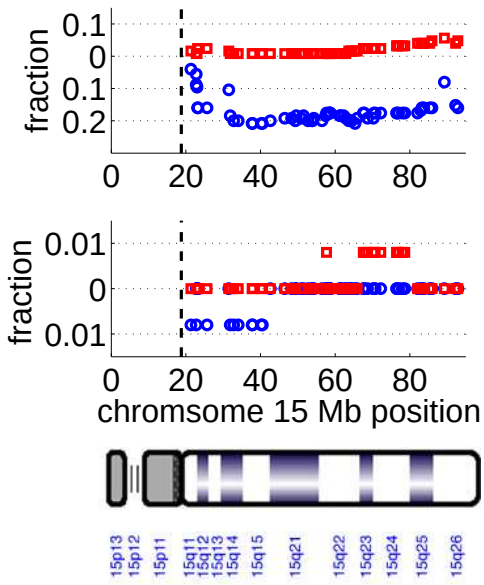
(l)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

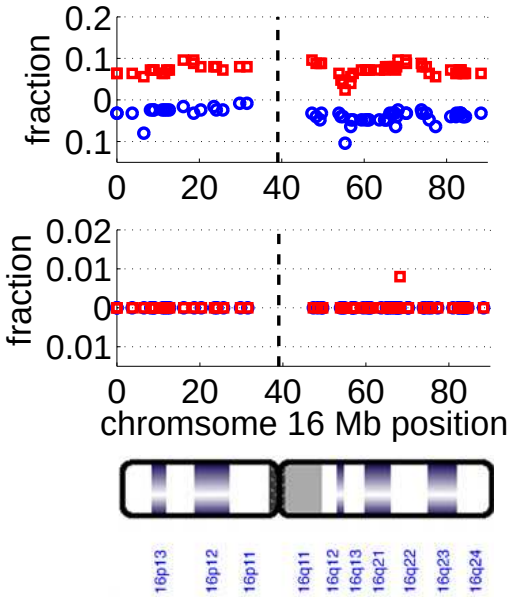


(m)

(n)

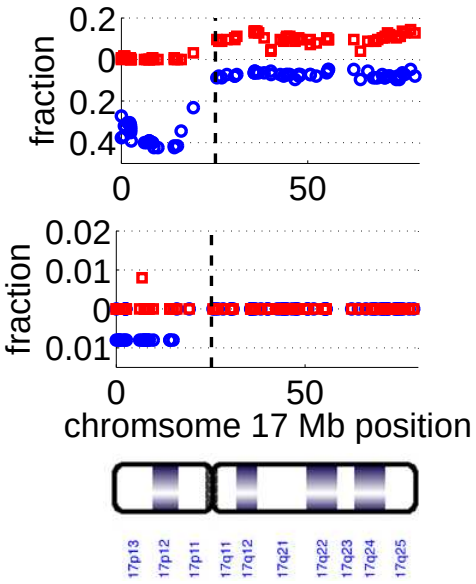


(o)

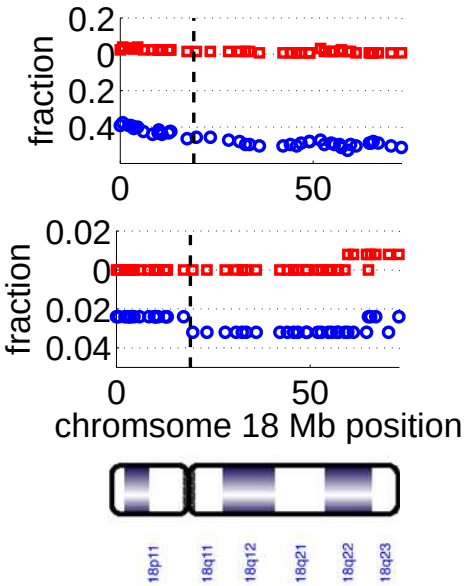


(p)

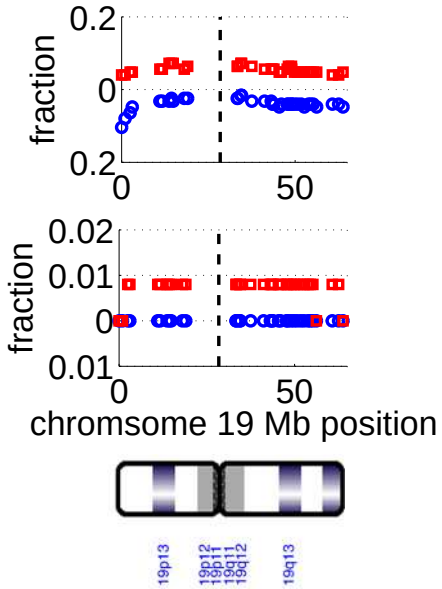
Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.



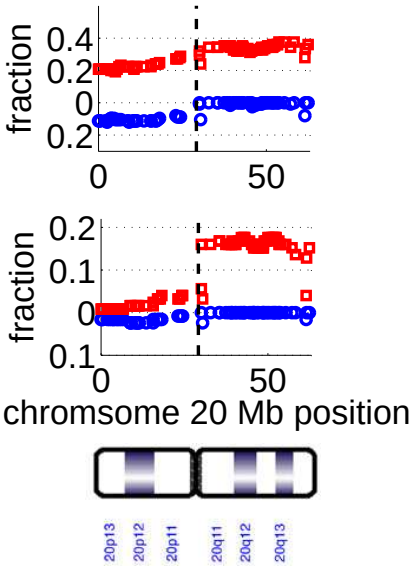
(q)



(r)

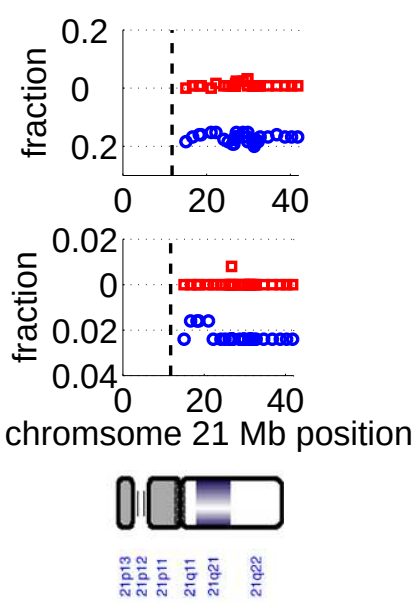


(s)

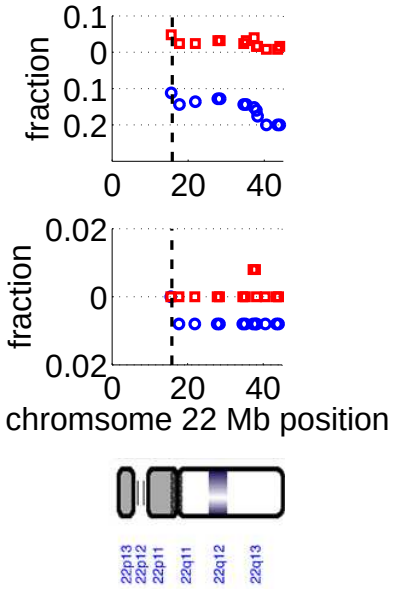


(t)

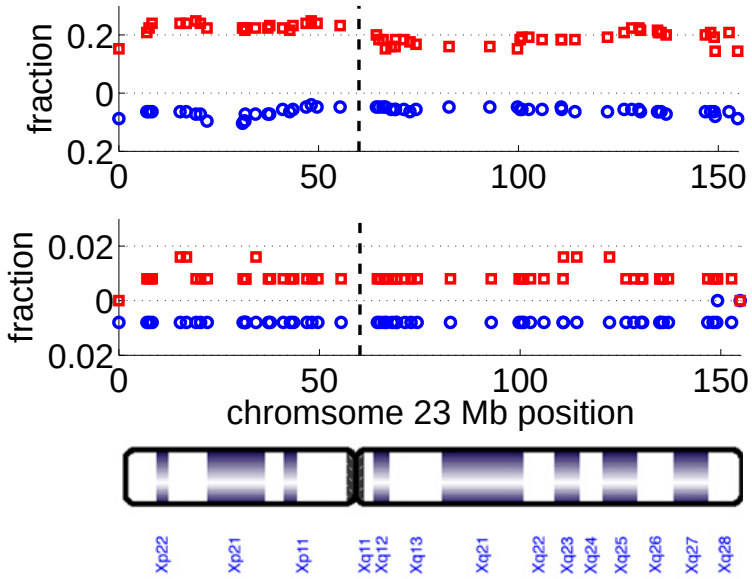
Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.



(u)



(v)



(w)

Figure 14: (cont'd) Fraction of gain (top, red square), loss (top, blue circle), amplification (bottom, red square) and deletion (bottom, blue circle) on human chromosomes.

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon