A flexible, accurate, and extensible statistical method for detecting genomic copy-number changes using Hidden Markov Models with Reversible Jump MCMC

Oscar M. Rueda and Ramón Díaz-Uriarte
Statistical Computing Team,
Structural and Computational Biology Programme,
Spanish National Cancer Centre (CNIO)
Melchor Fernández Almagro 3
28029 Madrid
Spain
E-mail: omrueda@cnio.es, rdiaz02@gmail.com

December 12, 2006

Abstract

Genomic DNA copy number alterations (CNAs) are associated with complex diseases, including cancer: CNAs are indeed related to tumoral grade, metastasis, and patient survival. CNAs discovered from array-based Comparative Genomic Hybridization (aCGH) data have been instrumental for identifying disease-related genes and potential therapeutic targets. To be immediately useful in both clinical and basic research scenarios, aCGH data analysis requires accurate methods that do not impose unrealistic biological assumptions and that provide direct answers to the key question "What is the probability that this gene/region has CNAs?". Current approaches fail, however, to meet these requirements. Here, we introduce a new method for identifying CNAs from aCGH; we use a non-homogeneous Hidden Markov Model fitted via Reversible Jump Markov Chain Monte Carlo, and we incorporate model uncertainty through Bayesian Model Averaging. RJaCGH provides an estimate of the probability that a gene/region has CNAs while incorporating inter-gene distance and the capability to analyze data on a chromosome or genome-wide basis. RJaCGH outperforms alternative methods, and the performance difference is even larger with noisy data and highly-variable inter-gene distance, both commonly found features in aCGH data. Furthermore, our probabilistic method allows identifying minimal common regions of CNAs among samples and could incorporate expression data. In summary, we provide a rigorous statistical framework for locating genes and chromosomal regions with CNAs with potential applications to cancer and other complex human diseases.

1 Introduction

Alterations in the number of copies (gains, losses) of genomic DNA have been associated with several hereditary anomalies and are involved in human cancers (reviews and examples in [PA05, LCCL06, UKS+06, MPN+05a, ABB+04, SLT+04, FMM+00]). For example, amplification of some genes, especially oncogenes, is one well known mechanism for tumor activation [HBC+00, HKS+04] and it is involved in the deregulation of cellular control [VFP+03, VK04]. Copy number alterations have been associated with tumoral grade, metastasis development, and patient survival [PA05, LCCL06, UKS+06, MPN+05a, ABB+04, SLT+04, FMM+00], and studies about copy number changes have been instrumental for identifying relevant genes for cancer development and patient classification [PA05, LCCL06, PSP+02].

A widely used technique to identify copy number changes in genomic DNA is array-based Comparative Genomic Hybridization (aCGH). Two DNA samples (e.g., problem and control) are differentially labeled (often with fluorescent dyes) and competitively hybridized to chromosomal DNA targets. After hybridization, emission from each of the two fluorescent dyes is measured, and the signal intensity ratios are indicative of the relative copy number of the two samples (see reviews in [PA05, LCCL06]). Therefore, a key step in any study of the relationship between altered copy numbers and disease is using the fluorescence

ratio data to identify genes and contiguous chromosomal regions with altered copy numbers.

The main biomedical problem, both for the study of the copy number alterations *per se* and for downstream analysis (e.g., relationship with gene expression changes or patient classification), is the accurate identification of the genes/chromosomal regions that have an altered copy number. Satisfactorily dealing with this problem requires a method that: a) provides direct answers that can be used in different settings (e.g., clinical vs. basic research); b) reflects the underlying biology and accounts for key features of the technological platform; c) can accommodate the different levels of analysis (types of questions) addressed with these data.

First, estimates of the probabilities of alteration (instead of p-values or smoothed means) are the most direct and usable answer to this problem [EMLB06, BR06]. Probabilities can be used in contexts that cover from basic research to clinical applications [PA05, LCCL06] so that, for instance, a clinician might require high certainty of alteration of a specific gene before more invasive procedures, whereas a basic researcher can consider for further study genes that show only moderate probability of alteration (e.g., $probability > 0.5$). Finally, appropriately used, probabilities of alteration can account for uncertainty in model building [SXD+06, HMRV99].

Second, the analysis should incorporate distance between probes [MTT06, FSPA04, LJKP05, BR06, HWLZ05, LCCL06]: widely used aCGH platforms like those based on cDNA microarrays and ROMA lead to variable coverage across chromosomes, with very unequal distances between probes (i.e., some regions have probes that are very close to each other, whereas in other regions probes are very far apart). As copy number changes involve chromosome segments, contiguous loci will have the same copy number, unless there is an abrupt change to another copy number [PA05, DRO+04]: the further apart two loci are, the more likely it is that a copy number event will have taken place in between them. Thus, in densely covered regions the copy number of a probe is a good predictor of the copy number of the neighboring probes. In contrast, in poorly covered regions, contiguous probes or loci might be many thousands of kilobases apart, making it more likely that at least one copy number change has taken place, and consequently a probe provides less information about the likely state of its neighboring probes. Therefore, unless we use a platform where all probes are equally spaced, we need to use the distance between probes (and not just the order), so that the information that consecutive probes provide is adequately accounted for.

Third, depending on the focus of the study, the analysis should be conducted either chromosome by chromosome, or genome-wide [SXD+06, BR06, EMLB06]. Analysis at the chromosome level are appropriate to detect alterations in copy number of loci relative to the rest of the loci in that same chromosome, regardless of that chromosome's ploidy (a trivial example would be detection of copy number changes in loci of the human Y chromosome in an otherwise diploid genome). On the other hand, detection of copy number changes that affect most of a chromosome often require genome-wide analysis (in chromosome-wide

analysis, as the mean or median chromosome level is used as the reference, detection of such changes is virtually impossible). Moreover, the use of genome-wide analysis can offer statistical advantages (e.g., reduced variance of estimation). As both types of analysis offer complementary information, because they focus on different biological phenomena (chromosomal gains/losses vs. gains of loci within chromosome), a suitable method should allow these two approaches.

1.1 Previous approaches

Available methods for the analysis of aCGH fail some or most of these requirements. Smoothing techniques [HST+04, OVLW04, PRM+05, HSG+05, LBL+05, HWLZ05, PRL+05] do not use gene distance nor provide posterior estimates of the likely state of each gene/clone, and data from each chromosome are analyzed independently of each other. Hidden Markov Models (HMMs) and related techniques offer a flexible modeling framework, and can provide probabilities of alteration [EMLB06, SXD+06, BR06]. Some HMM-based methods [FSPA04, SXD+06], however, do not incorporate distance between genes, assuming instead that inter-gene distance is constant. In addition, most of them do not deal satisfactorily with the unknown number of hidden states (the true number of states of copy number). Some methods fix in advance the number of hidden states (three: [BR06,EMLB06]; four: [SXD+06]): pre-specification of the number of states has the consequence of jumbling all changes involving multiple gains into a single state with a common mean, which is biologically questionable [DRO+04], specially as the resolution of the technology improves. A better approach would provide posterior probabilities of the number of states; using such a procedure over many different experiments will tell us whether three- or four-state models are a reasonable simplification. Of those methods that do not assume a fixed number of hidden states [FSPA04, MTT06, DRO+04], one of them [DRO+04] cannot be used for questions about the number of hidden states, or for breaking the data into more categories than gained/lost/no-change, which are increasingly important questions with higher-resolution techniques and are needed for distinguishing regions of moderate copy gains from regions of large copy gains. The remaining two [FSPA04, MTT06] fit HMMs for a range of number of states and then use AIC-based model selection, but AIC-based selection with HMMs has not been theoretically justified [CMR05], does not provide a probability of the likely number of states, and selecting a single model leads to underestimation of the true variability in the data; these two methods, in addition, use a final clustering step of hidden states that introduces several ad-hoc decisions, and do not return probabilities of alteration.

1.2 Statistical model: overview

We have developed a method, RJaCGH, that fulfills the three requirements above, and does not suffer from the limitations discussed for other methods. We start our modeling by noting that, for a given chromosome or genome, the copy numbers of genomic DNA (e.g., 0, 1, 2 copies, ...) of different genes or

segments are an unknown finite number. Thus, genes or segments could be classified into several groups with respect to their (unknown) copy number. In addition, as mentioned above, we expect that the copy number of a gene will be similar to the copy number of its closest neighbors, with that expected similarity decreasing when genes are further apart. Finally, for a given copy number, the aCGH fluorescence ratios should be centered around a $\log_2$ value, with some random noise. We want to use the observed log-ratios to identify regions with altered copy number.

The biological features of this model (a finite number of unknown or hidden states that are indirectly measured, with states of close elements likely to be similar, and variable distances between genes) can be modelled with a non-homogeneous Hidden Markov Model (HMM) [CMR05]. To provide a direct estimate of the probability that a given gene or region has an altered copy number we will use a Bayesian model computed via Markov Chain Monte Carlo (MCMC). Since we do not know the true number of hidden states, we fit models with varying number of hidden states and, to allow for transdimensional moves between models with different numbers of states, we use Reversible Jump [Gre95]. After running a large number of MCMC iterations, we can summarize the posterior probabilities. First, we will obtain posterior probabilities for the number of states. Conditional on a given number of states, each model will provide posterior distributions of the parameters of interest (e.g., means, variances, transition matrices). From the later, we can obtain posterior probabilities that a gene is gained or lost. To obtain our final estimates, we incorporate the uncertainty in model selection by using Bayesian Model Averaging [HMRV99] (estimates are weighted by posterior probability of each number of states), for the probabilities of genes being gained or lost. The complete statistical method we will call RJaCGH (from Reversible Jump-based analysis of aCGH data).

2 Results and discussion

We have applied RJaCGH and several alternative methods (including the best-performing ones [WF05, LJKP05]) to 500 simulated data sets [WF05] (see Supporting Information). These are data "(...) simulated to emulate the complexity of real tumor profiles" and designed to become "(...) a standard for systematic comparisons of computational segmentation approaches" [WF05] and, are not data simulated under our own model. To assess the effect of variable inter-gene distance, we randomly deleted data points (see details in supplementary material) so that each original simulated data set gives rise to another four data sets with (an average of) 10%, 25%, 50% and 65% of observations missing. The length of these gaps is modeled by a Poisson distribution, so larger percentages of missing data correspond to larger variability in inter-gene distances.

Results in Figure 1 (see also Supporting Information Fig. 1) show the excellent performance of RJaCGH, and how it outperforms alternative methods. Moreover, Figure 2 (see also Supporting Information Figs. 2 and 3) shows that

the difference between RJaCGH and alternative approaches is accentuated when we consider jointly the effects of noise and variability in inter-gene distance. Analysis using three other performance statistics (False Discovery Rate, Sensitivity, and Specificity) show the same overall patterns (see Supporting Information, Figs. 2 and 3): for some specific statistics, RJaCGH can be second (but very close) to another approach; this other approach, however, performs poorly with respect to the remaining statistics.

Similar results are obtained when applying these methods to a real data set of nine cell lines [SNS$^+$01], and comparing the predicted ploidy with the known ploidy [SNS$^+$01] (see Supplementary Material, Fig. 4). Overall, therefore, RJaCGH is the best performing method when considering the four available statistics.

The excellent performance of RJaCGH is a result of the statistical method used, which incorporates inter-gene distance, and adapts to variable noise in the data (without the need for fine-tuning of parameters, contrary to some other methods). Moreover, one of the main features of RJaCGH, its returning of posterior probabilities of CNAs, cannot be compared to most alternative methods as they do not provide this type of output. What most alternative approaches return are smoothed means, p-values, or a classification into states without any assessment of the uncertainty of this assignment to states. But a probability of alteration (which RJaCGH returns) is much easier to interpret and to use (with possibly different thresholds depending on the type of research) and is, often, the direct answer to the basic biomedical question. (The few alternative approaches that return probabilities of alteration [EMLB06, SXD$^+$06, BR06] all make the untenable assumption that the true number of biological states of alteration are three [EMLB06, BR06] or four [SXD$^+$06]).

3 Conclusion

We have developed a method to analyze aCGH for copy number changes that incorporates distance between genes, does not fix in advance the number of hidden states, accounts for model selection uncertainty, and allows to analyze one or more chromosomes simultaneously. We have shown that our method performs as well as, or better than, alternative approaches when there is no variation in inter-clone distance, but that it clearly outperforms alternative methods as the variability in inter-gene distance increases and when noise in the data increases. Our method provides clear answers to biological questions using a sound statistical approach, that allows the biologist to answer in an objective way questions about the probability of a gene or region having an altered copy number.

As RJaCGH provides posterior probabilities of alteration of contiguous genes (segments), it is relevant to recent efforts in aCGH methodology [RSH$^+$06, DEG$^+$06]. We can use the probabilities to identify regions with consistent alterations across samples (in a statistically rigorous way, including control of False Discovery Rate), and detect subgroups of samples according to recurrence patterns [MPN$^+$05b]. Likewise, posterior probabilities of being in

a specific state together with the estimated posterior mean of each state can be used as the basis for identifying breakpoints of biological significance. Finally, the model of RJaCGH can be extended to provide a rigorous downstream analysis of aCGH including the integration of gene expression and proteomic data [PSP+02, WF05].

4 Material and methods

4.1 Model

We use a non-homogeneous Hidden Markov Model with Gaussian emissions. We can either fit one model to all the chromosomes of an array or we can fit a different model for each chromosome of an array. Let n be the number of genes, and k the number of different copy numbers in the collection of genes. Let S_i be the true state (copy number) of the gene i: $S_i = \{1, \dots k\}_{i=1,\dots,n}$. Let Y_i be the relative copy number of the gene i, that is the log ratio of fluorescence intensities between tumor and control samples. Let X_i be the distance in bases between gene i and gene $i+1$ (we normalize these distances between 0 and 1 to increase numerical stability). How distance is measured depends on the platform: distance can be the distance from the end of the spot to the start of the next, if the length of the spots is proportional to the length of the gene (so we have the same information for every gene), or the distance between the midpoint of the spots, if the length of the spots is not proportional to the length of the gene.

We assume that $\{S_i\}$ follows a non-homogeneous 1st order Markov process, as: $P(S_i = s_i | S_{i-1} = s_{i-1}, X_{i-1} = x_{i-1}) = Q_{s_{i-1}, s_i, x_{i-1}}$. Biologically, we expect that $Q_{S_{i-1}=r, S_i=r, X_{i-1}}$, the probability of staying in the same hidden state, is a decreasing function of X_{i-1}, so the dependence of the state of a gene onto the next one is lower the further the genes are. We also expect that when the distance between two genes is maximal, the state of a gene should be independent from the state of its predecessor. Thus, we model the transition probabilities as:

$$Q_{i,j,x} = \frac{exp\{-\beta_{i,j} + \beta_{i,j} x\}}{\sum_{p=1}^{k} exp\{-\beta_{i,p} + \beta_{i,p} x\}} \tag{1}$$

Where β has the form:

$$\beta = \begin{pmatrix} 0 & \beta_{1,2} & \dots & \beta_{1,k} \\ \beta_{2,1} & 0 & \dots & \beta_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1} & \beta_{k,2} & \dots & 0 \end{pmatrix} \tag{2}$$

With all $\beta_{i,j} \geq 0 \quad \forall i, j$. Finally, conditioned on $\{S_i\}$, $\{Y_i\}$ follows a Gaussian process: $(Y_i | S_i = s_i) \sim N(\mu_{s_i}, \sigma^2_{s_i})$.

For computational reasons and modeling flexibility, we opted for Bayesian methods using Markov Chain Monte Carlo. To fit models with varying number of hidden states we use Reversible Jump. Suppose that we have a collection of K HMM models, and each of them has a number of k hidden states, from $k = \{1, \ldots, K\}$. Let $\theta(k)$ be the HMM associated to k, that is $\theta(k) = \{\mu(k), \sigma^2(k), \beta(k)\}$. The prior distributions for the model are the usual ones in mixture problems [RG97]: $p(k)$ is the prior for the number of hidden states with $p(k) \sim U(1, k)$, $p(\theta(k)/k)$ is the prior of the HMM conditioned to k, the number of hidden states with $\mu(k) \sim N(\alpha, \varrho^2)$, where α and ϱ are the median and range of Y_i; $\sigma^2(k) \sim IG(ka, g)$, where ka is 2 and g is $\varrho^2(Y_i)/50$; $\beta(k) \sim \Gamma(1, 1)$. The likelihood of the model, $L(y; k, \theta(k))$, can be computed by Forward Filtering [CMR05], so the joint distribution is $p(k)p(\theta(k)/k)L(y; k, \theta(k))$.

4.2 Estimation and fitting

We can draw samples from the posterior distribution through a Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm [Gre95]. In RJMCMC, we explore the posterior distribution of possible models, jumping not only within a model but also between models with a different number of parameters. To match the difference between degrees of freedom, some random numbers u with density $P(u)$ are generated, so if we are in state x, the new one is proposed in a deterministic way $x'(x, u)$. The reverse move is the inverse of that function: $x(x', u')$. This way, the usual Metropolis-Hastings acceptance probability can be computed [RG97]:

$$min\left\{1, \frac{L(y/x)p(x')p(u'/x')}{L(y/x)p(x)p(u/x)}|J|\right\} \tag{3}$$

where $L(y/x)$ is the likelihood, $p(x)$ are the priors, $p(u/x)$ are the densities of the candidates, and $J = |\frac{\partial x'}{\partial(x,u)}|$, the determinant of the Jacobian of the change of variable. We combine several Metropolis steps in a sweep [CMR05, RRT00]:

1. Update HMM of a model using a series of Metropolis-Hastings moves. (We do not use Gibbs Sampler to avoid the hidden state sequence from becoming part of the state space of the sampler, so dimensionality is reduced and reaching convergence is easier).

2. Update model (birth/death). When we have r states, a birth/death move is chosen with probabilities $p_{birth}(r)$ and $p_{death}(r)$ (these are 1/2 except in the cases when no movement of that type can be made, e.g. a death move when there is only one state). If a birth move is selected a new state is created from the prior distributions and accepted with probability:

$$min\{1, p\} \quad \text{where}$$
$$p = \frac{L(y; r+1, \theta(r+1))p(k=r+1)p_{death}(r+1)}{L(y; r, \theta(r))p(k=r)p_{birth}(r)} \times |J_{birth}|$$
$$\text{and} \quad J_{birth} = 1 \tag{4}$$

If a death move is chosen, a random state is deleted with a probability inverse to eq.[4].

3. Update model (split/combine). A split/combine move is attempted with probabilities $p_{split}(r)$ and $p_{combine}(r)$ (again, 1/2 except when a move can not be made). If a split move is selected, an existing state i_0 is split into two, i_1, i_2:

$$\mu_{i_1} = \mu_{i_0} - \epsilon_\mu, \quad \mu_{i_2} = \mu_{i_0} + \epsilon_\mu, \qquad \epsilon_\mu \sim N(0, \tau_\mu) \tag{5}$$
$$\sigma^2_{i_1} = \sigma^2_{i_0}\epsilon_\sigma, \quad \sigma^2_{i_2} = \sigma^2_{i_0}(1-\epsilon_\sigma), \qquad \epsilon_\sigma \sim \beta(2,2) \tag{6}$$

$$\text{Split column}$$
$$i_0: \quad \beta_{i,i_1} = \beta_{i,i_0}\epsilon_\beta, \quad \beta_{i,i_2} = \beta_{i,i_0}/\epsilon_\beta,$$
$$\epsilon_\beta \sim LN(0, \tau_\beta) \quad \text{for} \quad i \neq i_0 \tag{7}$$

$$\text{Split row}$$
$$i_0: \quad \beta_{i_1,j} = \beta_{i_0,j}U_j, \quad \beta_{i_2,j} = \beta_{i_0,j}(1-U_j),$$
$$\text{where} \quad U_j \sim \beta(2,2) \quad \text{for} \quad j \neq i_0$$
$$\beta_{i_1,i_2} \sim \Gamma(1,1) \tag{8}$$

This move is accepted with probability

$$min\{1, p\} \quad \text{where}$$
$$p = \frac{L(y; r+1, \theta(r+1))(r+1)}{L(y; r, \theta(r))}$$
$$\times \frac{P(k=r+1)P(\theta(r+1))P_{combine}(r+1)r}{P(k=r)P(\theta(r))P_{split}(r)(r+1)}$$
$$\times \frac{1}{2P(\epsilon_\mu)P(\epsilon_\sigma)\prod P(\epsilon_\beta)\prod P(U_j)}|J_{split}|$$
$$\text{and} \quad |J_{split}| = |2^r\sigma^2_{i_0}\prod_{j\neq i_0}\beta_{i_0,j}\prod_{i\neq i_0}\frac{\beta_{i,i_0}}{\epsilon_\beta}| \tag{9}$$

The split move must follow the adjacency condition [RG97] (the resulting states must be closer between them than to any other of the existing ones). If a combine step is selected, the symmetric move is performed and the inverse probability of acceptance is computed.

The combination of birth and split moves makes it possible not only to visit models with different number of parameters, but also to explore more thoroughly the posterior probability in the case of a parameter with a multi-modal density.

These moves are common ones [CMR05, RRT00], but we have changed several aspects of their design to improve the probability of acceptance, which is the most difficult step in Reversible Jump [CMR05, Gre95, RRT00]. We constraint the variance of every state so that it can not be greater than the variance of the whole data. Also, we have added the adjacency condition mentioned before, and used centering proposals. To prevent label-switching of states we have ordered the states according to means after every iteration of the sweep [RG97].

4.3 Inference

We run the former algorithm a large number of times (e.g., 50000) and, after discarding the first iterations as burn-in, we keep the last (e.g., 10000) samples as observations from the joint distribution, so we can make inferences from it. For every model that has been visited we obtain the posterior probabilities of the mean copy number of every state, the variance of the copy number of every state, and the function of transitions between hidden states. By counting the number of times that each model has been visited we obtain an estimate of the posterior probability of each model (i.e., we avoid using BIC or AIC). Then, applying the Viterbi algorithm [CMR05] to every sample obtained from the MCMC, and as this sample is a function of the HMM, we can obtain its posterior probability, something that usual Viterbi can not. From the Viterbi paths for all the samples, we can then compute the posterior probability that a gene belongs to every state or the probability that a sequence of genes is in a given state.

When obtaining posterior probabilities of copy number change, we use Bayesian Model Averaging [HMRV99] over all models visited. Let S_i be the lost, gained, no-change status of gene i, K the set of the models considered (in our case, that would be HMMs with $1, \ldots, K$ number of states), M_k the model with k number of states and S_i/M_k the state of gene i according to model k. We compute the unconditional (with respect to model selection) probability for the gene i as:

$$p(S_i = s_i|y) = \sum_{k \in K} p(M_k|y)p(S_i = s_i|M_k, y) \tag{10}$$

When analyzing multiple arrays, it is straightforward to use our approach to identify genes that show consistent copy number alterations across samples as

$$p(S_i = s_i|y) = \sum_{j=1}^{N} \sum_{k \in K} p(M_k|y_j)p(S_i = s_i|M_k, y_j)p(y_j) \tag{11}$$

where y_j are the data from array j and we have N arrays. If we have information about the reliability/representativeness of an array, that can be incorporated via $p(y_j)$; otherwise we set $p(y_j) = 1/N$.

4.4 Checking convergence and influence of priors

As in any MCMC approach, it is crucial to assess convergence of the sampler. We follow common practice [BG98] of running several chains in parallel. The convergence of the sampler depends strongly on the distribution of the candidates in Metropolis-Hastings. That is, every iteration a new value for the parameters is proposed from a distribution centered in their current values. The standard deviation of that distribution must be chosen in a way that samples explore all the parameter space. These standard deviations are not parameters of the model in the sense that different values give different fits, but values that can speed up convergence of the algorithm. The convergence of the posterior probability of the number of hidden states is reached when a large enough number of transdimensional moves is made. This number need not to be large if the likelihood is substantially higher in a particular model and data size is big enough. The birth and death moves only depend on the priors, but the split and combine moves depend also on their own design and the values of τ_μ and τ_β (see eq. [5] and eq. [7]). The priors chosen have been extensibly tested in mixture models [RG97]. In addition, the priors and rest of the parameters have very little effects: even small CGH arrays contain thousands of points so that the likelihood from the data dominates any prior. With the 2500 simulated data sets analyzed, we have only needed to specify the number of burn-in —50000— and to-keep samples —10000— and the number of chains —4— and only in 9 cases was there evidence of non-convergence —which was solved by re-running the samplers again.

4.5 Implementation and analysis

We have implemented RJaCGH using C (for the sweep algorithm) and R [R D06], and all analysis and comparisons have been done in R. See Supplementary Material.

5 Acknowledgments

Funding provided by Fundación de Investigación Médica Mutua Madrileña and Project TIC2003-09331-C02-02 of the Spanish Ministry of Education and Science (MEC). R.D.-U. partially supported by the Ramón y Cajal programme of the Spanish MEC. C. Lázaro-Perea, J. F. Poyatos, and A. Alibés provided comment on the ms.

Author's contributions: Oscar M. Rueda developed the statistical model, did most of the programming and the conducted analysis. Ramon Diaz-Uriarte

conceived the model, participated in model development and programming, and conducted simulations. Both authors wrote the paper.

References

[ABB+04] Andrew J. Aguirre, Cameron Brennan, Gerald Bailey, Raktim Sinha, Bin Feng, Christopher Leo, Yunyu Zhang, Jean Zhang, Joseph D. Gans, Nabeel Bardeesy, Craig Cauwels, Carlos Cordon-Cardo, Mark S. Redston, Ronald A. Depinho, and Lynda Chin. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, 101(24):9067–9072, 2004.

[BG98] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[BR06] P. Broët and S. Richardson. Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics*, 22(8):911–918, April 2006.

[CMR05] Olivier Cappé, Eric Moulines, and Tobias Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer, August 2005.

[DEG+06] Sharon J J. Diskin, Thomas Eck, Joel Greshock, Yael P P. Mosse, Tara Naylor, Christian J J. Stoeckert, Barbara L L. Weber, John M M. Maris, and Gregory R R. Grant. Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Res*, page in press, August 2006.

[DRO+04] R. S. Daruwala, A. Rudra, H. Ostrer, R. Lucito, M. Wigler, and B. Mishra. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc Natl Acad Sci U S A*, 101(46):16292–16297, November 2004.

[EMLB06] D.A. Engler, G. Mohaptra, D.N. Louis, and R. Betensky. A pseudo-likelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, 7(3):399–421, 2006.

[FMM+00] F. Forozan, E. H. Mahlamki, O. Monni, Y. Chen, R. Veldman, Y. Jiang, G. C. Gooden, S. P. Ethier, A. Kallioniemi, and O. P. Kallioniemi. Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary dna microarray data. *Cancer Res*, 60(16):4519–4525, August 2000.

[FSPA04] Jane Fridlyand, Antoine M. Snijders, Dan Pinkel, and Donna G. and Albertson. Hidden markov models approach to the analysis of array cgh data. *Journal of Multivariate Analysis*, 90(1):132–153, July 2004.

[Gre95] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, (82):711–732, 1995.

[HBC+00] M. A. Heiskanen, M. L. Bittner, Y. Chen, J. Khan, K. E. Adler, J. M. Trent, and P. S. Meltzer. Detection of gene amplification by genomic hybridization to cdna microarrays. *Cancer Res*, 60(4):799–802, February 2000.

[HKS+04] K. Holzmann, H. Kohlhammer, C. Schwaenen, S. Wessendorf, H. A. Kestler, A. Schwoerer, B. Rau, B. Radlwimmer, H. Dhner, P. Lichter, T. Gress, and M. Bentz. Genomic dna-chip hybridization reveals a higher incidence of genomic amplifications in pancreatic cancer than conventional comparative genomic hybridization and leads to the identification of novel candidate genes. *Cancer Res*, 64(13):4428–4433, July 2004.

[HMRV99] J.A Hoeting, H. Madigan, A.E. Raftery, and C.T Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

[HSG+05] L. Hsu, S. G. Self, D. Grove, T. Randolph, K. Wang, J. J. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226, April 2005.

[HST+04] P. Hupé, N. Stransky, J. P. Thiery, F. Radvanyi, and E. Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, December 2004.

[HWLZ05] Tao Huang, Baolin Wu, Paul Lizardi, and Hongyu Zhao. Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, September 2005.

[LBL+05] O. C. Lingjaerde, L. O. Baumbusch, K. Liestl, I. K. Glad, and A. L. Borresen-Dale. Cgh-explorer: a program for analysis of array-cgh data. *Bioinformatics*, 21(6):821–822, March 2005.

[LCCL06] William W. Lockwood, Raj Chari, Bryan Chi, and Wan L. and Lam. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *European Journal of Human Genetics*, 14(current):139–148, 2006.

[LJKP05] Weil R R. Lai, Mark D D. Johnson, Raju Kucherlapati, and Peter J J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21:3763–3770, 2005.

[MPN+05a] A. Misra, M. Pellarin, J. Nigro, I. Smirnov, D. Moore, K. R. Lamborn, D. Pinkel, D. G. Albertson, and B. G. Feuerstein. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res*, 11(8):2907–2918, April 2005.

[MPN+05b] A. Misra, M. Pellarin, J. Nigro, I. Smirnov, D. Moore, K. R. Lamborn, D. Pinkel, D. G. Albertson, and B. G. Feuerstein. Array comparative genomic hybridization identifies genetic subgroups in grade 4 human astrocytoma. *Clin Cancer Res*, 11(8):2907–2918, April 2005.

[MTT06] J. C. Marioni, N. P. Thorne, and S. Tavaré. Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, 22(9):1144–1146, May 2006.

[OVLW04] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, October 2004.

[PA05] D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37 Suppl:S11–S17, June 2005.

[PRL+05] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6:27, 2005.

[PRM+05] T. S. Price, R. Regan, R. Mott, A. Hedman, B. Honey, R. J. Daniels, L. Smith, A. Greenfield, A. Tiganescu, V. Buckle, N. Ventress, H. Ayyub, A. Salhan, S. Pedraza-Diaz, J. Broxholme, J. Ragoussis, D. R. Higgs, J. Flint, and S. J. Knight. Sw-array: a dynamic programming solution for the identification of copy-number changes in genomic dna using array comparative genome hybridization data. *Nucleic Acids Res*, 33(11):3455–3464, 2005.

[PSP+02] J. R. Pollack, T. Srlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Brresen-Dale, and P. O. Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20):12963–12968, October 2002.

[R D06] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[RG97] S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the of the*

Royal Statistical Society Series B (Statistical Methodology), 59:731 – 792, 1997.

[RRT00] C. Robert, T. Ryden, and D. Titterington. Bayesian inference in hidden markov models through reversible jump markov chain monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75, 2000.

[RSH+06] C Rouveirol, N Stransky, Ph Hup, Ph La Rosa, E Viara, E Barillot, and F Radvanyi. Computation of recurrent minimal genomic alterations from array-cgh data. *Bioinformatics*, 22:2066–2073, January 2006.

[SLT+04] Jonathan Sebat, B. Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Maner, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, Conrad C. Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, July 2004.

[SNS+01] A. M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J. P. Yue, J. W. Gray, A. N. Jain, D. Pinkel, and D. G. Albertson. Assembly of microarrays for genome-wide measurement of dna copy number. *Nat Genet*, 29(3):263–264, 2001.

[SXD+06] S. P. Shah, X. Xuan, R. J. Deleeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy. Integrating copy number polymorphisms into array cgh analysis using a robust hmm. *Bioinformatics*, 22(14):e431–e439, July 2006.

[UKS+06] A. E. Urban, J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. V. Popescu, J. F. Cubells, R. Green, B. S. Emanuel, M. B. Gerstein, S. M. Weissman, and M. Snyder. High-resolution mapping of dna copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 103(12):4534–4539, March 2006.

[VFP+03] J. A. Veltman, J. Fridlyand, S. Pejavar, A. B. Olshen, J. E. Korkola, S. DeVries, P. Carroll, W. L. Kuo, D. Pinkel, D. Albertson, C. Cordon-Cardo, A. N. Jain, and F. M. Waldman. Array-based comparative genomic hybridization for genome-wide screening of dna copy number in bladder tumors. *Cancer Res*, 63(11):2872–2880, June 2003.

[VK04] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–799, August 2004.

[WF05] Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21:4084–4091, September 2005.

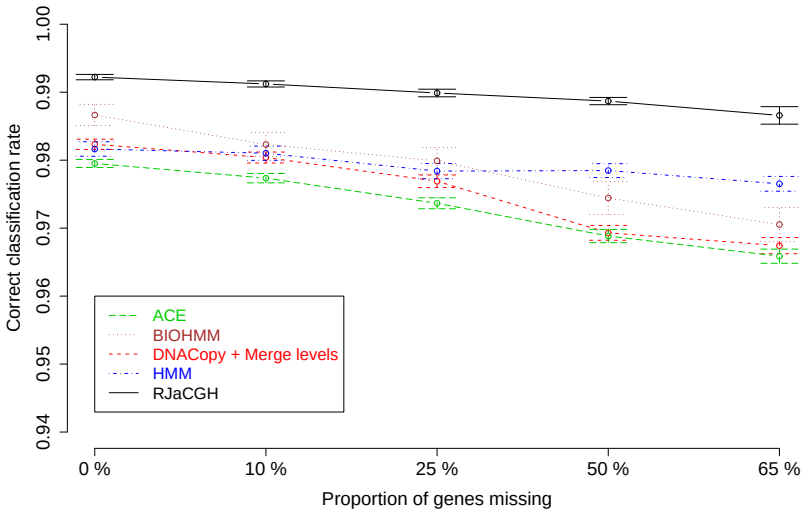

Figure 1: Correct classification: effects of variability in inter-gene distance (percentage of genes missing). Shown are the mean and 95% confidence interval around the mean of the correct classification error rate. Each mean and confidence interval is computed from 500 data sets [WF05] (see text and Supporting Information for generation of inter-gene distance variability).

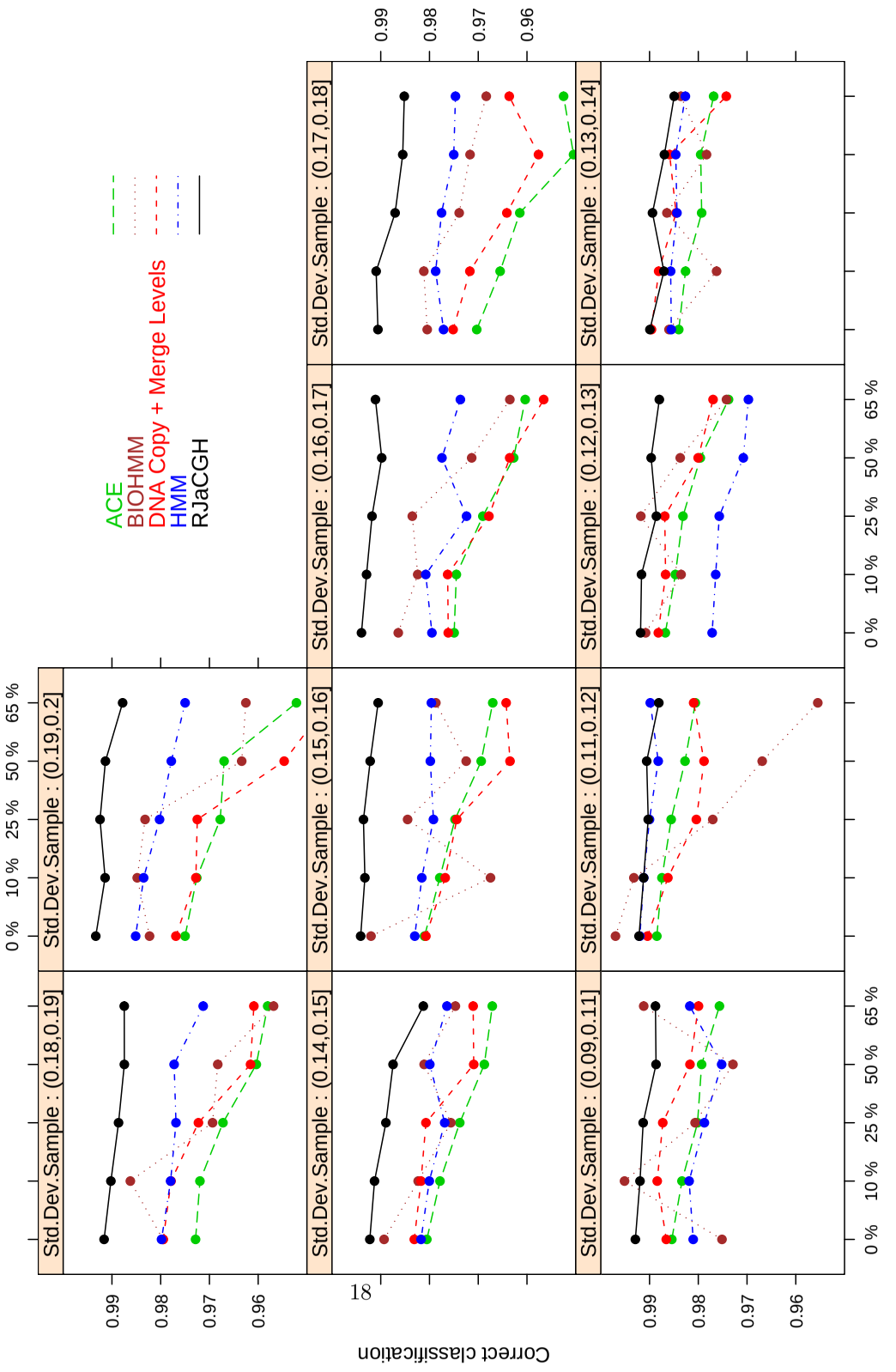

Figure 2: Correct classification: joint effects of noise and variability in inter-gene distance. Same data as in Figure 1. The noise (standard deviation) of each sample is split into ten non-overlapping ranges, and each panel shows the mean correct classification success vs. the proportion of missing genes (i.e., increasing levels of variance in inter-gene distance); each mean is based on approximately 50 samples.

Supporting Information for: "A flexible, accurate, and extensible statistical method for detecting genomic copy-number changes"

Oscar M. Rueda and Ramón Díaz-Uriarte
Statistical Computing Team,
Structural and Computational Biology Programme,
Spanish National Cancer Centre (CNIO)
Melchor Fernández Almagro 3
28029 Madrid
Spain
E-mail: omrueda@cnio.es, rdiaz02@gmail.com

1 Method comparisons: general

1.1 Methods compared

We have examined the performance of our method and compared it two four other methods: DNA copy [1] and ACE [2], two competing approaches that have been shown to be among the best performers in recent reviews [3, 4], and the HMM [5] and non-homogeneous HMM [6] approaches, two methods that share some common features with our method (but see discussion). All of these approaches, except ACE, are available as R/BioConductor packages. ACE is available as a Java program from [2]; however, this Java program is not suitable for batch processing of simulations; thus, we implemented it as a loadable C module, and call it from R. Other promising methods (specially [7]) could not be included in the comparative study because code is not available or directly implementable from the available published descriptions.

1.2 Settings of methods

All methods were run with their default parameters. Details and modifications follow.

For DNA copy, and following the recommendations in [3], we have used the "merge levels" proposal of [3]. The methods of Fridlyand et al. [5] and Marioni et al. [6] include an internal, implicit, merge levels-like algorithm.

For ACE [2] the FDR used is the minimal one of the available (experimenting with the method in these data set showed that other, larger, FDRs lead to much poorer performance).

RJaCGH was run with six parallel chains, each with 60000 iterations of which the first 50000 were discarded as burn-in. For each run, two full chains were discarded by trimming (i.e., eliminating the two most extreme observations, one on each tail, with respect to the average estimated number of states of each chain). The parameters of the distributions of the candidates were selected automatically by a heuristic approach that, within model, leads to an acceptance probability near 0.23 [8]. The parameters of the jumps between models were taken as the mean of the within model parameters.

1.3 Mapping of methods' output to gain/loss/no-change

Only ACE provides, directly, output labels that correspond to "gain/loss/no change" status of the genes. For DNAcopy, and as in [3], we post-processed the merge levels output, so the level with mean closest to zero, which is also the level with the largest number of observations, was assigned to the "no change" class (which is consistent with all assumptions in the normalization step, and most in the analysis step, that most genes/clones are not affected by copy number changes). The remaining levels were assigned to either "gain" or "lost" depending on whether there smoothed value was larger or smaller, respectively, than the "no change" class. Similar procedure was followed with HMM and BIOHMM after these methods returned their output.

Table 1: Confusion matrix

		Predicted		
		Gained	*No change*	*Lost*
True State	*Gained*	TG	G_{nc}	G_l
	No change	NC_g	TNC	NC_l
	Lost	L_g	L_{nc}	TL

For RJaCGH, our method includes a some what similar approach. We consider as "no change" all states whose IQR (interquartile range) includes 0. After this step, we add the groups with posterior mean closes to 0 to the "no change" class until the proportion of observations in the no change class is no less than a pre-specified level (by default 0.65). This procedure is consistent with the assumptions in the normalization step that most genes/clones are not affected by copy number changes.

1.4 Statistics used to evaluate performance

We have evaluated performance of each method using four different statistics. To understand the statistics, it is useful to refer to table 1.

Correct classification rate The percentage of genes that are assigned to the right class. In table 1, the sum of all diagonal terms divided by the total number of clones. This is an overall estimate of how well a method is doing. This is likely to be the most relevant measure in every day usage, as it combines the measures below (and incorporates, for instance, trade-offs between False Discovery Rate and Sensitivity).

False Discovery Rate We define it in here as the number or mistakes made when we call something a gain or a loss: the number of no-changes among the clones Predicted to be gains or losses. In the table above,

$FDR = \frac{NC_g+NC_l}{TG+NC_g+L_g+G_l+NC_l+TL}$

(i.e., the sum of NC_g and NC_l divided by the total number of those predicted to be "gained" or "lost"). (Note that, in our comparisons, there was not a single case, for any method, were a true gain was predicted to be a lost, or vice-versa).

Specificity The probability of predicting no change when the true state is no change. In terms of table 1:

$Specificity = \frac{TNC}{NC_g+TNC+NC_l}$

Sensitivity The probability of predicting a gain (loss) outcome when the true state is gained (lost). Here we sum over both possible deviations from no change:

$Sensitivity = \frac{TG+TL}{TG+G_{nc}+G_l+L_g+L_{nc}+TL}$

It should be noted that there are ways to achieve, e.g., great False Discovery Rate, without being a good overall performer. For instance, by requiring very strong evidence to call something a loss, we can reduce the False Discovery Rate, at the expense of not identifying many changes as such (i.e., at the expense of lowering the sensitivity). Similarly, if a method predicts no change most of the time, the Specificity will be high at the expense of a low sensitivity.

2 Simulations

2.1 Simulation settings

We have used the same simulated data sets as Willenbrock and Fridlyand [3] used in their recent comparison of methods of aCGH analysis [3]. Details of the data are provided in the original paper [3]; briefly, these are data "(...) simulated to emulate the complexity of real tumor profiles" and designed to become "(...) a standard for systematic comparisons of computational segmentation approaches" [3, p. 4]. The authors simulated five hundred data sets based on the profiles of real tumor samples, and a sample-specific variance (between 0.1 and 0.2) was added to each sample. It is unlikely that these data were simulated under a model that is specifically well suited for our method. Other simulated data sets (or simulation approaches) did not seem appropriate to compare alternative approaches; most papers that present simulated data do simulate the data under models that are the same (or very similar to) the model used to analyze the data. The simulations in [1] are useful for examining breakpoint detection, but not for questions related to the recovery of the correct "gained, lost, no change" label, and the simulations in [4] are too simplistic in their settings (only a single type of alteration added) and the number of points generated is too short (100). The 500 data sets of Willenbrock and Fridlyand [3], however, are suitable for examining recovery of true labels, are simulated based on real profiles to which varying levels of noise are added, and provide a sufficiently large and diverse data set to gain valuable information about the relative performance of different methods.

We downloaded the data [3] from `http://www.cbs.dtu.dk/~hanni/aCGH/`, and the actual file used was
`http://www.cbs.dtu.dk/~hanni/aCGH/20chromosome.simulated.data.RData`.
Each of the 500 simulations consisted of 20 chromosomes, with 100 clones in each chromosome. One hundred clones per chromosome are too few points (at least for most aCGH data for human samples) and make it hard to assess the effect of differences in spacing between clones. Thus, instead of using the 2000 clones as if divided in 20 chromosomes, we just regarded all the 2000 clones as if they came from the very same single chromosome which allows us to introduce fairly large numbers of missing data (i.e., variability in spacing).

None of the data sets above included variability in inter-gene distances which, as we argue in the paper, is an important feature of many real aCGH data sets, and a specific problem we try to address with our method. Therefore, to assess if

our method does perform reasonably under varying inter-gene distance (and how it performs compared to other methods) we need to add inter-gene distance to the data set. Instead of modifying the original simulation models of [3], we have instead introduced "holes" (or missings) in the data thus replicating a situation where the data are generated according to the models in [3], but the actual observed data is a sample from the generated data (such as is the case with many aCGH platforms that show unequal coverage of different parts of the genome).

The "holes" or missing fragments in the data have been created with a very simple model: we choose at random 100 locations in the genome, and eliminate a contiguous segment of clones. The length of this segment is modeled with a Poisson distribution (so the actual length of the segment that is missing is drawn, randomly, from a Poisson distribution with parameter λ). This λ parameter determines the average number of missing points; in addition, as this is a Poisson distribution (where the variance is $= \lambda$), increasing λ results in an increase in the variance of the length of the missing fragments. We have used, for the λ parameter, the values 2, 5, 10, or 13. Thus, for each original data set, we obtain another four data sets, with a different number of missing data points. On average, the derived data sets have 10%, 25%, 50% and 65%. In other words, from the 500 data sets, we generate another 2000 data sets. Thus, of the 2500 data sets, each subset of 500 has an average number of missing points of 0% (in this case, 0 is not an average, but the actual number), 10%, 25%, 50% and 65%. To minimize the variability in methods' comparisons, the derived data sets analyzed by all methods were the same.

2.2 Results and discussion

Results are shown in Figures 1, 2, 3.

Overall performance: Correct Classification Rate RJaCGH is better than any of the alternative approaches:

- The difference in performance between RJaCGH and alternative approaches increases as the variability in spacing between clones increases (i.e., as the proportion of missing genes increases). These patterns are seen in Figure 1 (a).
- The difference between RJaCGH and alternative approaches, is accentuated in Figures 2 and Figure 3: contrary to other methods, RJaCGH does not suffer the same decrease in performance as the noise in the data increases.

False Discovery Rate The best performer is DNAcopy, and RJaCGH is the second best; all other methods suffer from much greater False Discovery Rates (Figure 1, (b)). As the noise in the data increases, however, the difference between RJaCGH and DNAcopy becomes smaller with RJaCGH being the method with smallest FDR at the highest noise levels (Figure 3 (b)). For all practical

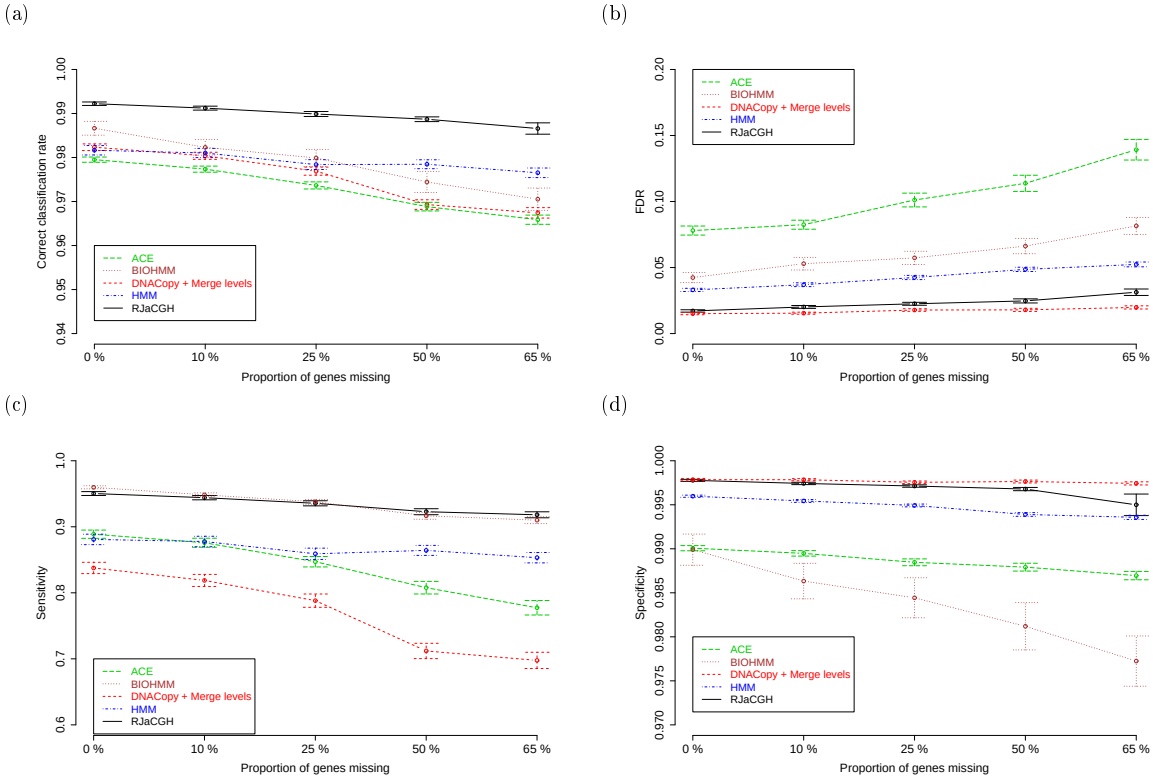

Figure 1: Comparative performance on the simulated data from [3] (see text for details). Relationship between the average value of the statistic and the variability in inter-gene distance (increases in the percentage of genes missing are directly related to increases in the variability in inter-gene distance). Shown are the mean and 95% confidence interval around the mean (based on 500 data sets). In panels (a), (c), (d), higher is better; in panel (b) lower is better.

usages, however, differences between RJaCGH and DNAcopy in terms of FDR are probably negligible.

Note, however, that the good performance of DNAcopy with respect to False Discovery Rate is at the expense of a reduced Sensitivity (see next).

Sensitivity The largest sensitivity is achieved by BIOHMM at small values of noise in the data and by RJaCGH with higher noise levels (see panel (c) in all Figures). Over all levels of noise in the data, however, the performance between RJaCGH and BIOHMM (Figure 1 (c)) is indistinguishable, but clearly superior to other methods. The good performance of BIOHMM with respect to Sensitivity, however, is achieved at the expense of its high False Discovery Rate and low Specificity (see below).

Specificity As could be expected from the definition of Specificity and False Discovery Rate, the patterns of Specificity are similar to those commented above for False Discovery Rate.

In summary, RJaCGH has the largest correct classification. For some specific statistics, RJaCGH can be second (but very close) to some approaches; these other approaches, however, perform poorly in the other performance statistics. Overall, therefore, RJaCGH is the best performing method when considering the four available statistics.

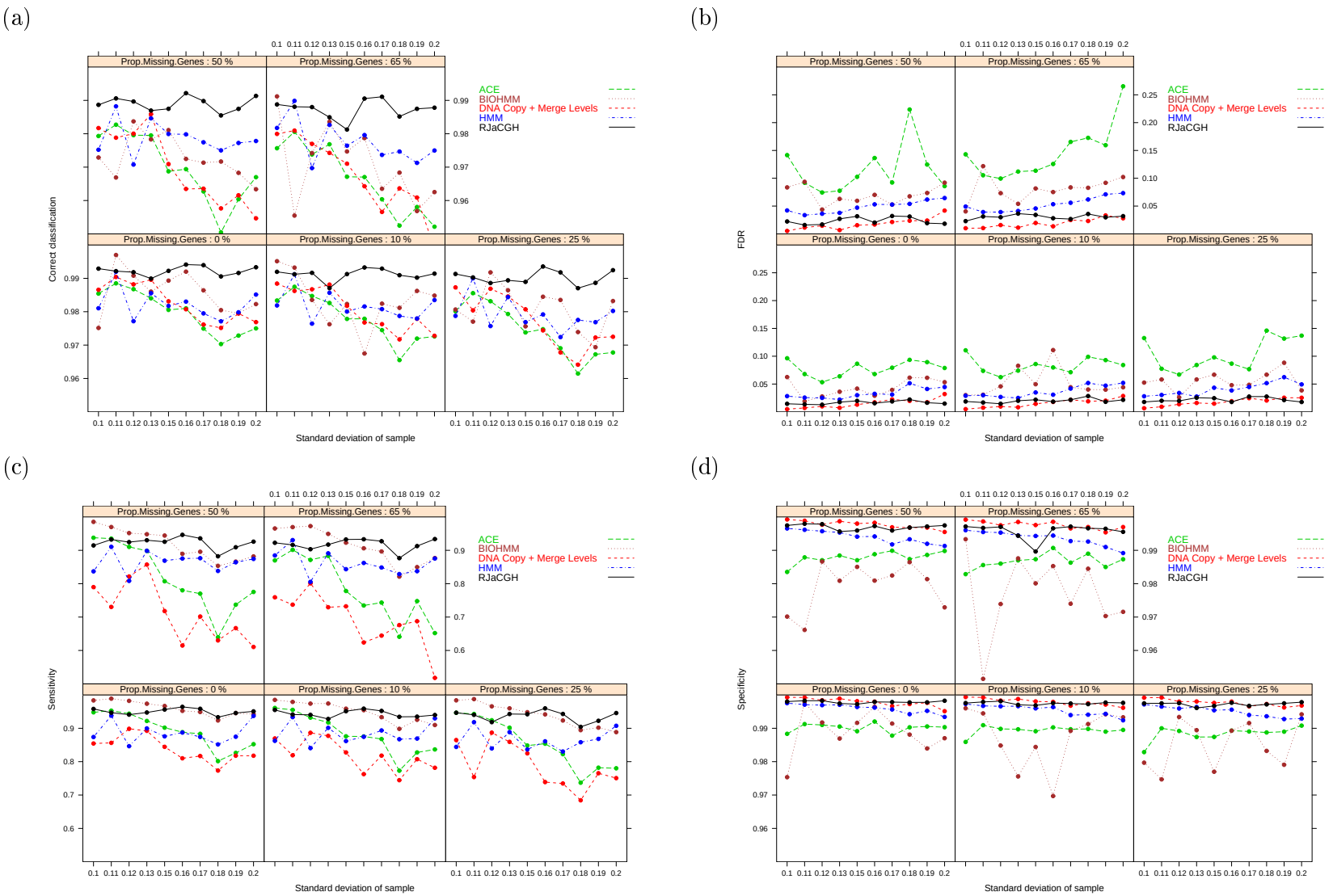

Figure 2: Analysis of simulated data: conditioning on variability of inter-gene distance. Analysis of data from Willenbrock and Fridlyand [3] (see text for details on addition of gaps). For each level of average number of missing genes (0, 10, 25, 50, 65 %) or, equivalently, for increasing levels of variance in the distance between clones, we compute the mean of the statistic at ten equally spaced levels of noise in the data (i.e., the 500 data sets have been divided in 10 groups according to their noise, so that the midpoints of each interval are $0.105, 0.115, 0.125, \ldots, 0.185, 0.195$). Therefore, each point in the figure corresponds to the mean from about 50 samples.

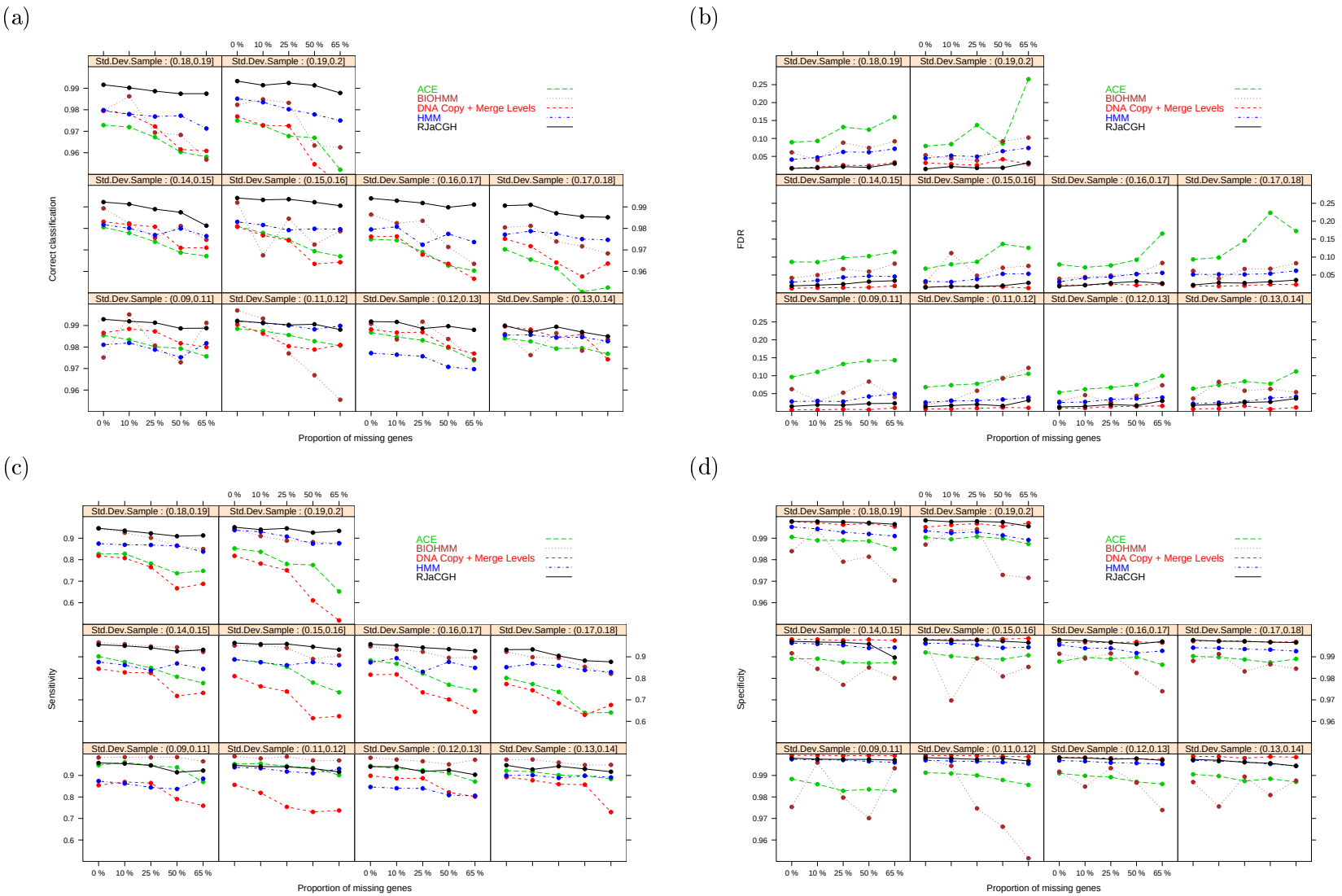

Figure 3: Analysis of simulated data: conditioning on sample noise. Analysis of data from Willenbrock and Fridlyand [3] (see text and Figure 2 for details). The noise (standard deviation) of each sample is split into ten non-overlapping ranges, and each panel shows the average value of the statistic vs. the proportion of missing genes (i.e., increasing levels of variance in inter-gene distance) for a given sample noise.

3 Real data from Snijders et al.

We have also analyzed the well known nine cell lines from Snijders et al. [9] available from
http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html and we have compared the results from our method with the known ploidy, as provided by Snijders et al.

Figure 4 shows the comparative performance of each of the methods. From the figure we see that RJaCGH has performance comparable to that of the best method for each statistic.

As an example of the type of output provided by RJaCGH, Figure 5 shows the results of one analysis for the complete genome of the cell line gm03563. Panel a) indicates a large posterior probability of a model with four hidden states; two of the states of the four-state model, however, are extremely close to each other (panel b) and, because of their posterior means (panel b) and variances (panel c) we consider them to represent the same biological state of no change in copy number. The other two states are well separated, with posterior means clearly negative or positive, so we regard them as biological states of loss and gain of copy number. Note that the component that represents the hidden state of loss is assigned to only two genes

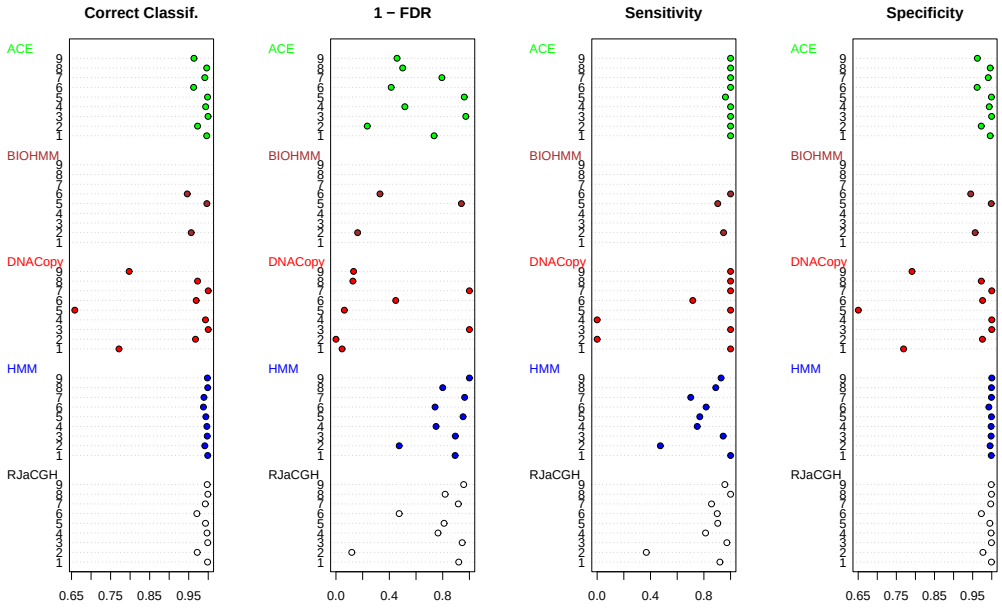

Figure 4: Comparative performance on the nine cell lines from Snijders et al. [9]. We show the value of the performance statistics for each cell line (numbered 1 to 9, which correspond to gm01524, gm01535, gm01750, gm03134, gm03563, gm05296, gm07081, gm13031,gm13330, respectively). In all these figures, "larger is better" (note we use 1-FDR, not FDR). Only three values are shown for BIOHMM, as the rest of data lead to crashes in the program.

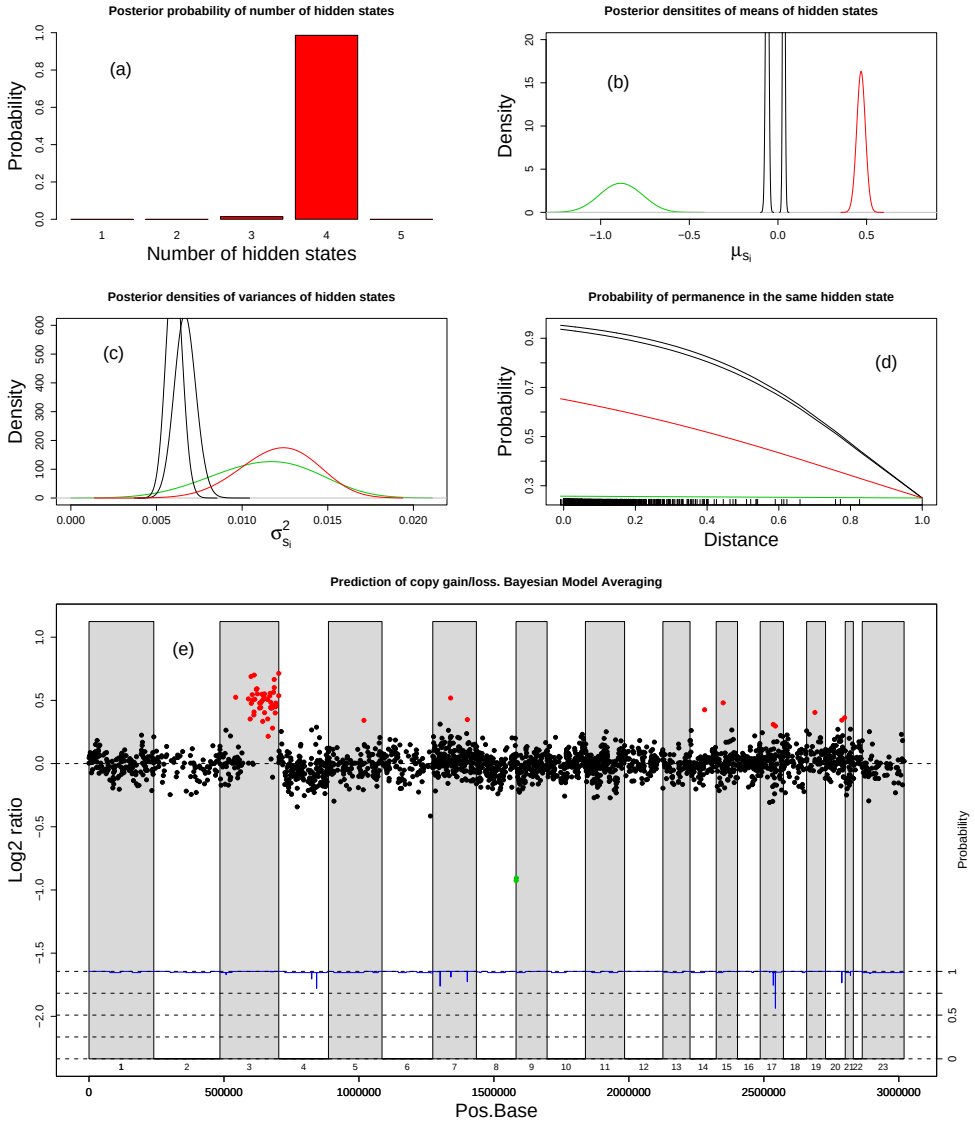

Figure 5: Results of the RJaCGH analysis of gm03563 cell line from Snijders. Results shown are from four parallel chains; see text for details about other parameters. The lower panel shows the results from the Bayesian Model Averaging step (see text); black dots correspond to genes classified as 'normal' or non-changed, red dots to genes classified as 'gained' and green dots to genes classified as 'losses'; the lower blue line shows the posterior probability for every gene of belonging to the predicted state. The vertical alternating white and grey bars denote the different chromosomes with the chromosome number shown at bottom.

(panel e, green dots), exactly the same two genes whose true state is loss [9]. Panel d) shows that the probability of remaining on the same state decreases as distance increases, eventually becoming 0.25(= 1/Number hidden states). Finally, panel e) shows the results from the Bayesian Model Averaging. This is a particularly clear-cut

model, as the posterior probabilities that each gene belongs to the state with highest posterior is very high (the lower blue line is > 0.9 for almost all genes).

4 Implementation and analysis

We have implemented RJaCGH using C (for the sweep algorithm) and R [10]. The code is available from CRAN
(`http://cran.r-project.org/src/contrib/Descriptions/RJaCGH.html`)
and from the Asterias site (`http://www.asterias.info`). All analysis and comparisons have been done in R, using the BioConductor (`http://www.bioconductor.org`) packages DNAcopy by by E. S. Venkatraman and Adam Olshen and aCGH by Jane Fridlyand and Peter Dimitorv, and a version of ACE implemented by O.M.R. in R and C.

References

[1] Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based dna copy number data.** *Biostatistics* 2004, 5:557–572.

[2] Lingjaerde OC, Baumbusch LO, LiestÃ¸l K, Glad IK, Borresen-Dale AL: **Cghexplorer: a program for analysis of array-cgh data.** *Bioinformatics* 2005, 21:821–822.

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

[4] Lai WRR, Johnson MDD, Kucherlapati R, Park PJJ: **Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data.** *Bioinformatics* 2005, 21:3763–3770.

[5] Fridlyand J, Snijders AM, Pinkel D, Albertson DGa: **Hidden markov models approach to the analysis of array cgh data**. *Journal of Multivariate Analysis* 2004, 90:132–153.

[6] Marioni JC, Thorne NP, Tavaré S: **Biohmm: a heterogeneous hidden markov model for segmenting array cgh data.** *Bioinformatics* 2006, 22:1144–1146.

[7] Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B: **A versatile statistical analysis algorithm to detect genome copy number variation.** *Proc Natl Acad Sci U S A* 2004, 101:16292–16297.

[8] Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.

[9] Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of dna copy number**. *Nat Genet* 2001, 29:263–264.

[10] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.