OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Parsimonious Higher-Order Hidden Markov Models for Improved Array-CGH Analysis with Applications to *Arabidopsis thaliana*

Michael Seifert[1]*, André Gohr[2], Marc Strickert[3], Ivo Grosse[2]

1 Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, **2** Institute of Computer Science, Martin Luther University Halle, Halle (Saale), Germany, **3** Science and Technology, University of Siegen, Siegen, Germany

Abstract

Array-based comparative genomic hybridization (Array-CGH) is an important technology in molecular biology for the detection of DNA copy number polymorphisms between closely related genomes. Hidden Markov Models (HMMs) are popular tools for the analysis of Array-CGH data, but current methods are only based on first-order HMMs having constrained abilities to model spatial dependencies between measurements of closely adjacent chromosomal regions. Here, we develop parsimonious higher-order HMMs enabling the interpolation between a mixture model ignoring spatial dependencies and a higher-order HMM exhaustively modeling spatial dependencies. We apply parsimonious higher-order HMMs to the analysis of Array-CGH data of the accessions C24 and Col-0 of the model plant *Arabidopsis thaliana*. We compare these models against first-order HMMs and other existing methods using a reference of known deletions and sequence deviations. We find that parsimonious higher-order HMMs clearly improve the identification of these polymorphisms. Moreover, we perform a functional analysis of identified polymorphisms revealing novel details of genomic differences between C24 and Col-0. Additional model evaluations are done on widely considered Array-CGH data of human cell lines indicating that parsimonious HMMs are also well-suited for the analysis of non-plant specific data. All these results indicate that parsimonious higher-order HMMs are useful for Array-CGH analyses. An implementation of parsimonious higher-order HMMs is available as part of the open source Java library Jstacs (www.jstacs.de/index.php/PHHMM).

Citation: Seifert M, Gohr A, Strickert M, Grosse I (2012) Parsimonious Higher-Order Hidden Markov Models for Improved Array-CGH Analysis with Applications to *Arabidopsis thaliana*. PLoS Comput Biol 8(1): e1002286. doi:10.1371/journal.pcbi.1002286

Editor: William Stafford Noble, University of Washington, United States of America

Received May 5, 2011; **Accepted** October 11, 2011; **Published** January 12, 2012

Funding: MiS and MaS were supported by grant XP3624HP/0606T of the Ministry of Culture of Saxony-Anhalt. MiS was supported by the DAAD PROCOPE grant 50748812. MaS was supported by the DFG graduate school 1546. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: seifert@ipk-gatersleben.de

Introduction

In recent years, the method of array-based comparative genomic hybridization (Array-CGH) [1–5] has been widely applied for the detection of DNA copy number polymorphisms between closely related genomes. Most Array-CGH studies have their focus in cancer research for the genome-wide identification of deletions and amplifications of genomic regions in tumor compared to healthy tissue [6–10]. With the availability of the genome sequence of the accession Columbia (Col-0) of the model plant *Arabidopsis thaliana* [11], studies comparing the genomes of different accessions have been performed using the Array-CGH approach to analyze evolutionary processes and phenotypic features at a molecular level [12–17]. All these studies require efficient bioinformatics methods for the precise identification of copy number polymorphisms from Array-CGH data.

Over the last years, a large number of different methods for the identification of copy number polymorphisms from Array-CGH data have been developed including approaches based on Gaussian mixture models [18], circular binary segmentation [19–21], genetic local search algorithms [22,23], dynamic programming [24–26], hierarchical clustering [27], sparse Bayesian learning [28], variational methods [29,30], smoothing techniques [31–34], regression models [35,36], or wavelets [37,38]. In-depth contributions to the comparison of different methods have been made by two studies [39,40]. Selected well-performing methods have been made publicly available by webservers [41–44].

Despite these different methods, the identification of copy number polymorphisms by methods based on Hidden Markov Models (HMMs) is very popular [45–61] providing a natural way for modeling genomic spatial dependencies present in Array-CGH data. Most of these HMM -based methods use three up to six states with specific Gaussian emission densities for the modeling of Array-CGH measurements. Greater differences exist in learning principles used for adapting models to data. The Baum-Welch algorithm [62–65] has been used in [47,48,56,59,61] for estimating the parameters of the HMM by maximizing the likelihood without integrating prior knowledge on the distribution of Array-CGH measurements. Due to specific model extensions, numerical estimations of the likelihood have been considered in [50,51]. Bayesian approaches using Markov Chain Monte Carlo simulations have been developed in [52–55,58], a numerical Bayesian estimation has been applied in [57], and a Bayesian

Author Summary

Array-based comparative genomics is a standard approach for the identification of DNA copy number polymorphisms between closely related genomes. The huge amounts of data produced by these experiments require efficient and accurate bioinformatics tools for the identification of copy number polymorphisms. Hidden Markov Models (HMMs) are frequently used for analyzing such data sets, but current models are based on first-order HMMs only having limited capabilities to model spatial dependencies between measurements of closely adjacent chromosomal regions. We develop parsimonious higher-order HMMs enabling the interpolation between a mixture model ignoring spatial dependencies and a higher-order HMM exhaustively modeling these dependencies to overcome this limitation. In an in-depth case study with *Arabidopsis thaliana*, we find that parsimonious higher-order HMMs clearly improve the identification of copy number polymorphisms in comparison to standard first-order HMMs and other frequently used methods. Functional analysis of identified polymorphisms revealed details of genomic differences between the accessions C24 and Col-0 of *Arabidopsis thaliana*. An additional study on human cell lines further indicates that parsimonious HMMs are well-suited for the analysis of Array-CGH data.

Baum-Welch algorithm has been utilized in [60]. All these Bayesian approaches enable the integration of prior knowledge on the distribution of Array-CGH measurements for improving the identification of copy number polymorphisms.

A characteristic of all these HMMs is that they are based on the mathematical theory of standard first-order HMMs [65,66]. This leads to a common limitation that all these HMMs can only model dependencies between Array-CGH measurements of two directly adjacent chromosomal regions. Yet, no attention has been paid to higher-order HMMs enabling the modeling of dependencies between a chromosomal region and its most recent predecessors that are clearly present in Array-CGH data (e.g. Figure 1).

In contrast to the broad usage of first-order HMMs in applied sciences [66–68], published applications of higher-order HMMs are relatively rare, but they have been demonstrated to be powerful extensions of first-order HMMs for several applications including speech recognition [69–76], image segmentation [77–79], robotic [80], handwriting recognition [81], or DNA and protein sequence analysis [82–85]. Extensions of the mathematical theory of first-order HMMs to higher-order HMMs are comprehensively described in [86–89]. The improved modeling of spatio-temporal dependencies by higher-order HMMs is realized by a more complex state-transition process defined on the basis of a higher-order Markov model reviewed in [90]. A limitation of this improved modeling is the exponential increase of transition parameters with increasing model order requiring growing amounts of data and computational resources for model training and evaluation. This has generally limited the usage of large model orders. Consequently, most existing studies have only focused on second-order HMMs [69–73,78,80,82,84].

To enable the usage of improved modeling characteristics of greater model orders by simultaneously overcoming the exponential increase of transition parameters, a fast incremental training has been developed in the domain of speech recognition [87,91]. This heuristic algorithm iteratively increases the model order by only including transition parameters that are required for the representation of the training data. That has led to higher-order HMMs with reduced model complexities [87,91,92] and to mixed-order HMMs [93–95] reaching improved results in speech recognition in comparison to first-order HMMs and standard higher-order HMMs. In addition, a variable-length HMM has been developed to improve the modeling of motion capture data [96,97]. The state-transition process of this model is defined by a variable memory Markov chain for which the transition parameters are determined by a minimum entropy criterion integrated into an extended Baum-Welch training. However, since implementations of both approaches for reducing the number of transition parameters are not publicly available and since algorithmic extensions would be necessary to enable the integration of prior knowledge, these models cannot directly be utilized for the analysis of Array-CGH data.

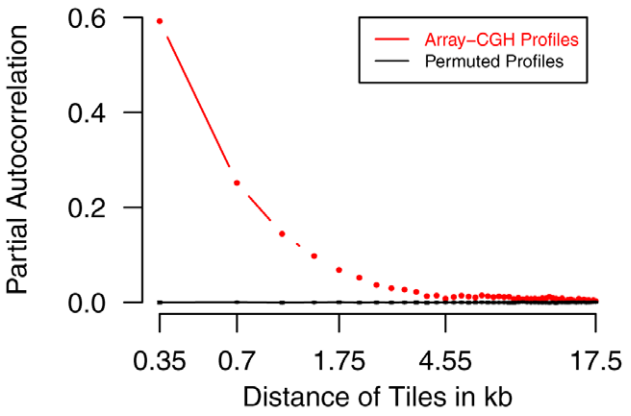

Figure 1. Spatial dependencies of measurements in Array-CGH profiles of *Arabidopsis thaliana*. The partial autocorrelation function characterizes spatial dependencies between measurements of adjacent chromosomal regions (tiles) in Array-CGH profiles. This function has been computed for the five chromosome-specific Array-CGH profiles by [103] comparing the genomes of the *Arabidopsis thaliana* accessions C24 and Col-0. The red curve represents the weighted mean partial autocorrelation function of the original Array-CGH profiles for increasing chromosomal distance of adjacent tiles in steps of 0.35 kb. The black curve represents mean values and standard deviations (both close to zero) of the mean weighted partial autocorrelation function for randomly permuted measurements in each of the five original Array-CGH profiles across 100 repeats. The significant presence of spatial dependencies of measurements in the Array-CGH profiles (red) compared to permuted profiles (black) motivates the modeling of such dependencies for the analysis of Array-CGH data.
doi:10.1371/journal.pcbi.1002286.g001

Here, we develop the novel model class of parsimonious higher-order HMMs enabling the interpolation between a mixture model ignoring spatial dependencies and a higher-order HMM exhaustively modeling spatial dependencies between measurements of closely adjacent chromosomal regions. This interpolation is realized by incorporating a dynamic programming approach [98,99] into a specifically developed Bayesian Baum-Welch training algorithm enabling the integration of prior knowledge and a data-dependent reduction of transition parameters. Based on that interpolation, a parsimonious higher-order HMM can effectively model spatial dependencies between measurements of closely adjacent chromosomal regions.

In an in-depth case study with the model plant *Arabidopsis thaliana*, we apply parsimonious higher-order HMMs to compare the genomes of the accessions C24 and Col-0 based on a publicly available Array-CGH data set. This enables the identification of DNA polymorphisms (deletions or sequence deviations, amplifications) in C24 with respect to the reference genome of Col-0 [11]. We evaluate and compare parsimonious higher-order HMMs

against standard first-order HMMs and other existing methods by making use of deletions or sequence deviations identified in an independent array-based resequencing experiment of C24 [100,101]. Moreover, we perform a functional analysis of identified genomic differences revealing novel details of differences between C24 and Col-0, and we also consider widely used human cell lines [102] for additional model comparisons.

Materials and Methods

In the materials part of this section, the Arabidopsis Array-CGH data set comparing the genomes of C24 and Col-0 is introduced and candidate regions of deletions or sequence deviations for model evaluation determined by an independent public resequencing experiment are considered. The model class of parsimonious higher-order HMMs is developed in the methods part of this section.

Materials

In this section, the Arabidopsis Array-CGH data set is introduced and candidate regions of deletions or sequence deviations for model evaluation identified in resequencing data are considered.

Arabidopsis Array-CGH data. An Array-CGH data set by [103] (GEO accession: GSM611097) comparing the genomes of the accessions C24 and Col-0 of the model plant *A. thaliana* is used to investigate the identification of DNA polymorphisms (deletions or sequences deviations, amplifications) by different methods. This data set was measured on a NimbleGen tiling array representing the five chromosomes of the Col-0 reference genome [11] by 364,339 genomic regions (tiles). The length of each tile is about 60 bp. All tiles on the array are spaced nearly equidistantly along the chromosomes with a mean distance of about 350 bp between two adjacent tiles. Lengths of single-stranded DNA segments hybridized to this array were in the range of 300 bp up to 900 bp. The tiling array was processed using the NimbleScan software resulting in normalized measurements.

The measurement of tile t on chromosome k is given by the log-ratio $o_t(k) := \log_2(C24_{k,t}/Col-0_{k,t})$ in dependency of the corresponding measured accession-specific fluorescent intensities $C24_{k,t}$ and $Col-0_{k,t}$. All log-ratios belonging to a chromosome $k \in \{1,\ldots,K=5\}$ are summarized in an Array-CGH profile $\vec{o}(k)=(o_1(k),\ldots,o_{T_k}(k))$ with T_k log-ratios represented in increasing order of the chromosomal locations of tiles.

Spatial dependencies between log-ratios on chromosomes are characterized in Figure 1. Tiles in close chromosomal proximity are highly correlated indicating that they have very similar measurements. These spatial dependencies between measurements of tiles in close chromosomal proximity (less than 5 kb) are most likely caused due to the lengths of single-stranded DNA fragments hybridized to the tiling array. Since the spacing between directly adjacent tiles on a chromosome is about 350 bp and because typically hybridized DNA fragments are having lengths up to 900 bp, it is expected that tiles in close chromosomal proximity are having very similar measurements.

The distribution of log-ratios in the Array-CGH data set is shown in Figure 2a. Most of the tiles have log-ratios close to zero as expected for unchanged genomic regions between C24 and Col-0. A smaller proportion of tiles has log-ratios much smaller than zero as expected for deletions or sequence deviations for genomic regions in C24 compared to the corresponding regions in Col-0. Only a very small proportion of tiles has log-ratios much greater than zero as expected for amplifications of genomic regions in C24 in comparison to Col-0. The asymmetry of the log-ratio distribution is caused by the design of the tiling array exclusively representing genomic regions of the reference genome of Col-0 [11].

Arabidopsis resequencing data. An array-based Affymetrix resequencing experiment of C24 was performed in [100] for identifying single nucleotide polymorphisms and long stretches of deletions or sequence deviations. This experiment was further processed in [101] by the developed mPPR algorithm resulting in candidate regions of deletions or sequence deviations in C24 with respect to the reference genome sequence of Col-0 [11]. The identified candidate regions of deletions or sequence deviations have additionally been evaluated in [101] by comparisons against available sequence data and known deletions. This clearly indicated that these candidate regions are also present in other data sets. Thus, this data set provides a useful resource for the evaluation of deletions or sequence deviations identified by different models in the Array-CGH data set.

We used the determined candidate regions of deletions or sequence deviations from the resequencing experiment to identify each tile in the Array-CGH data set for which at least 75% of its nucleotides ($\geq$45 bp of 60 bp) are covered by candidate regions. This results in 11,025 tiles labeled as candidates for deletions or sequence deviations among the 364,339 tiles in the Array-CGH data set. As expected for potential deletions or sequence deviations in C24, most of these labeled tiles have log-ratios much less than zero in the Arabidopsis Array-CGH data set (Figure 2b). This indicates that deletions or sequence deviations determined in [101] are clearly present in the Arabidopsis Array-CGH data set and suggests that these candidate regions are useful for model evaluations.

Methods

This section provides the basics of parsimonious higher-order HMMs. In the following, these models are introduced, a prior distribution for integrating prior knowledge into the training is specified, a model-specific Bayesian Baum-Welch training algorithm is developed, and details to the parameter initialization are given. Finally, a link to related work is given.

Parsimonious higher-order Hidden Markov Models. A parsimonious higher-order HMM with three states $S := \{-,=,+\}$ and Gaussian emissions is used for the analysis of Array-CGH profiles. Under consideration of the distribution of log-ratios in Array-CGH data (e.g. Figure 2a), the three states are defined to represent the following DNA polymorphisms. State '$-$' models deletions or sequence deviations with log-ratios much smaller than zero, state '$=$' models unchanged regions with log-ratios close to zero, and amplifications with log-ratios much greater than zero are modeled by state '$+$'. In contrast to other HMM -based methods like [48,50,52,55], the states of the parsimonious higher-order HMM are not explicitly modeling specific genomic copy numbers, but the states are covering a broad range of state-specific log-ratios by making use of flexible Gaussian emission densities.

Each state $i \in S$ is characterized by a Gaussian emission density $b_i(o_t) := 1/(\sqrt{2\pi}\sigma_i)\exp(-0.5(o_t-\mu_i)^2/\sigma_i^2)$ with state-specific mean $\mu_i \in \mathbb{R}$ and standard deviation $\sigma_i \in \mathbb{R}^+$ for modeling a log-ratio $o_t \in \mathbb{R}$. All emission parameters are summarized in the matrix $B := (\mu_i,\sigma_i)_{i \in S}$.

The state underlying a chromosomal region t with corresponding log-ratio o_t is denoted by $q_t \in S$. A state sequence $\vec{q} := (q_1,\ldots,q_T)$ belonging to an Array-CGH profile $\vec{o} := (o_1,\ldots,o_T)$ is assumed to be modeled by a parsimonious Markov model of order L [98,99]. This Markov model realizes the state-transition processes of the parsimonious higher-order HMM. The state-

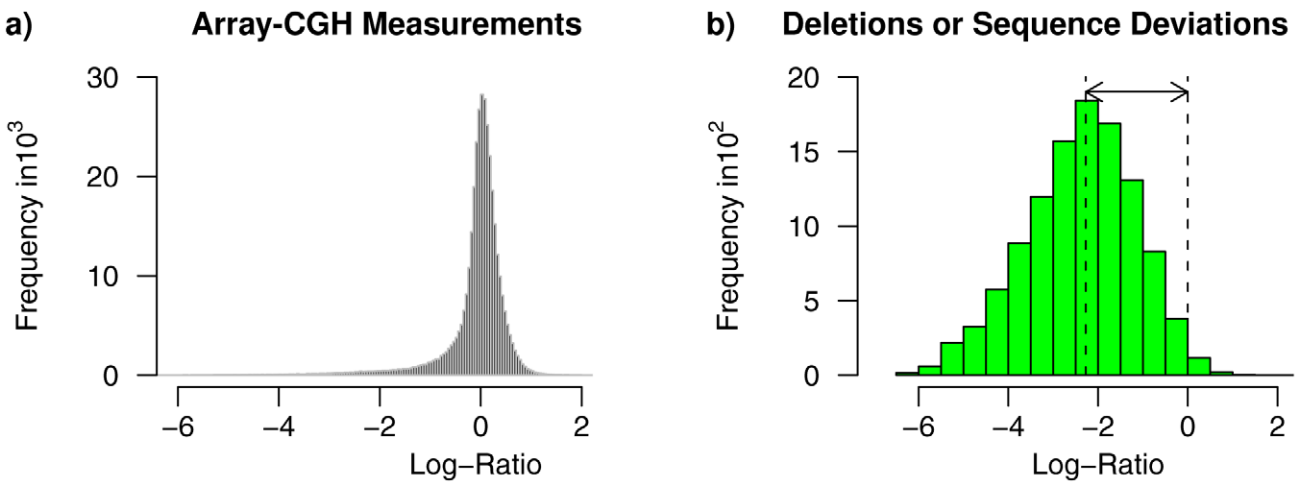

Figure 2. Characteristics of the *Arabidopsis thaliana* Array-CGH data set. a) Distribution of log-ratios measured for genomic regions in the Array-CGH data set by [103] comparing the genomes of the *Arabidopsis thaliana* accessions C24 and Col-0. The log-ratio of a genomic region characterizes changes in copy numbers (deletions or amplifications) or sequence deviations of this region in C24 in comparison to Col-0. Unchanged genomic regions between C24 and Col-0 have log-ratios close to zero. Deletions or sequence deviations of genomic regions in C24 have log-ratios much smaller than zero. Amplifications of genomic regions in C24 have log-ratios much greater than zero. **b**) Distribution of log-ratios of genomic regions (tiles) in the Array-CGH data set covered to at least 75% ($\geq$45 bp of 60 bp) by candidate regions of deletions or sequence deviations identified in [101] based on Affymetrix array-based resequencing data [100]. The large proportion of highly negative log-ratios indicates that these candidate regions are also present in the Array-CGH data set. Tiles covered by such candidate regions provide a useful resource for evaluating the identification of deletions or sequence deviations in the Array-CGH data set by different methods.
doi:10.1371/journal.pcbi.1002286.g002

transition process is similar to that of a higher-order HMM [89] additionally enabling a data-dependent sharing of transition parameters for state-transitions from specific state-contexts.

In more detail, the state-transition process of a parsimonious HMM of order $L \geq 1$ is defined by an initial state distribution $\vec{\pi} := (\pi_i)_{i \in S}$ with initial state probability $\pi_i \in (0,1)$ fulfilling $\sum_{i \in S} \pi_i = 1$ and a set of L transition matrices $A := \{A_{\tau_1}, \ldots, A_{\tau_L}\}$. Each transition matrix $A_{\tau_l} \in A$ is defined on the basis of a state-context tree τ_l subdividing the product set of state-contexts $S^l := \{(s_1, \ldots, s_l) : s_1 \in S, \ldots, s_l \in S\}$ into disjoint sets of equivalent state-contexts. A specific set of equivalent state-contexts of τ_l is denoted by ξ. All state-contexts $i \in \xi$ are assumed to share the identical transition probability $a_{\xi j}$ for a transition from each state-context in ξ to a next state $j \in S$. Thus, the parsimonious representation of state-contexts by sets of disjoint equivalent state-contexts reduces the total number of transition parameters of the model. Hence, the transition matrix $A_{\tau_l} := (a_{\xi j})_{\xi \in \tau_l, j \in S}$ is defined by corresponding transition probabilities $a_{\xi j} \in (0,1)$ fulfilling $\sum_{j \in S} a_{\xi j} = 1$.

Generally, the transition matrix A_{τ_l} with $l \in \{1, \ldots, L-1\}$ is used for the transition from the current state q_l to the next state q_{l+1} in dependency of the predecessor states $q_1, \ldots, q_{l-1}$, while the transition matrix A_{τ_L} is used for the transition from q_t to q_{t+1} under consideration of the predecessor states $q_{t-L+1}, \ldots, q_{t-1}$ for all $t \geq L$.

Exemplarily, three different types of state-context trees underlying a transition matrix A_{τ_2} are illustrated in Figure 3. The completely fused tree (Figure 3a) assigns all state-contexts to one leaf node, the complete tree (Figure 3c) represents each state-context in a separate leaf node, and the parsimonious tree (Figure 3b) groups selected state-contexts together resulting in less leaf nodes than in a complete tree.

Completely fused trees are the basis for a mixture model of Gaussian densities (HMM of order zero) that does not model spatial dependencies between log-ratios in Array-CGH profiles. Complete trees are underlying a higher-order HMM exhaustively modeling spatial dependencies. Parsimonious trees provide the basis for a parsimonious higher-order HMM interpolating between a mixture model and a higher-order HMM. This interpolation poses the problem of selecting optimal state-context trees for an HMM. For a fixed set of states, the number of different state-context trees grows super-exponentially for increasing model order (Figure S1 in Text S1). Thus, each existing state-context tree cannot be analyzed separately. To overcome this, we compute optimal state-context trees by an efficient dynamic programming approach [98,99] that has been incorporated into the Bayesian Baum-Welch training algorithm of the parsimonious higher-order HMM.

For identifying DNA polymorphisms in an Array-CGH profile, an extension of the standard state-posterior decoding algorithm [65] is used to compute the state-posterior probability $\gamma_t(i) := P[q_t = i | \vec{o}, \lambda]$ for quantifying the potential of a chromosomal region t to be represented by a state $i \in S$. Details to the state-posterior decoding and the computation of state-posterior probabilities for a parsimonious HMM are given in [89]. The state-posterior probabilities are used to rank log-ratios according to their tendency of being modeled by a specific state of the model (e.g. state '$-$' with respect to known deletions or sequence deviations from independent validation data). Additionally, these state-posterior probabilities can also be used to perform a decoding of individual measurements in an Array-CGH profile into the discrete states of the model by assigning the most likely state to each chromosomal region in an Array-CGH profile.

In summary, the parameters of the parsimonious higher-order HMM are denoted by $\lambda := (\vec{\pi}, A, B)$ and the three-state architecture of this model is illustrated in Figure S2 in Text S1.

Prior distribution. A problem-specific characterization of the parameters of a parsimonious higher-order HMM λ is achieved by integrating prior knowledge about Array-CGH profiles into the training. This is realized by specifying a prior

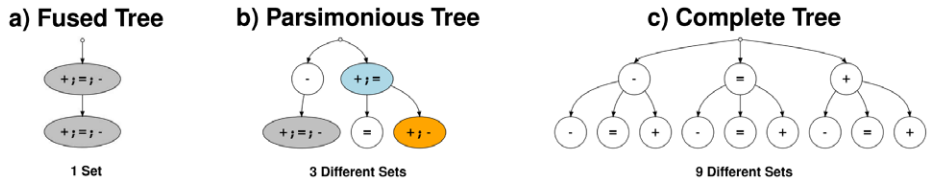

Figure 3. Examples of state-context trees. Selected state-context trees of height two representing different sets of disjoint sets of equivalent state-contexts of length two. The fused tree (**a)**) and the complete tree (**c)**) define marginal cases of state-context trees underlying a parsimonious higher-order HMM. Fused trees are underlying the mixture model, while complete trees are the basis of a higher-order HMM. The fused tree has the most parsimonious structure representing all state-contexts in one set of equivalent state-contexts, while the complete tree represents each state-context of length two by an individual set. The parsimonious tree (**b)**) with three disjoint sets of equivalent state-context has a complexity between the fused and the complete tree. More formally, each path from the root node at the top of a tree to a leaf node at the bottom of a tree represents a set of state-contexts defined to share common transition probabilities. The nodes directly under the root node of a tree represent possible current states, and the nodes under these nodes represent the corresponding predecessor states of the current state. Predecessor states have a specific influence on the state-transition from the current state to the next state depending on the type of the node. Exemplarily, some different types of nodes are highlighted in color. White nodes represent unfused nodes characterizing important states for a state-transition. Blue and orange nodes represent partially fused states of equal importance for a state-transition. Grey nodes represent completely fused nodes defining that the corresponding position in a state-context has no influence on a state-transition.
doi:10.1371/journal.pcbi.1002286.g003

distribution

$$P[\lambda|\Theta] := D_1(\vec{\pi}|\Theta_1)\cdot D_2(A|\Theta_2)\cdot D_3(B|\Theta_3) \quad (1)$$

for the parameters of the HMM $\lambda := (\vec{\pi},A,B)$ in dependency of the hyper-parameters $\Theta := (\Theta_1,\Theta_2,\Theta_3)$. This prior is defined to be a product of independent priors for the initial state distribution $\vec{\pi}$, the set of transition matrices A, and the emission parameters B. A conjugate prior distribution is chosen for each class of model parameters enabling the analytical parameter estimation during the training of a parsimonious higher-order HMM.

The prior distribution $D_1(\vec{\pi}|\Theta_1)$ of the initial state distribution is defined to be a transformed Dirichlet distribution [104], and the prior distribution $D_3(B|\Theta_3)$ of the state-specific Gaussian emission densities is defined to be a product of Gaussian-Inverted-Gamma distributions [105]. These two prior distributions are the usual ones applied for HMMs (e.g. [66,106]). Details to the prior of the initial state distribution and to the prior of the emission parameters are given in the section Prior distribution in Text S1.

In the following, the central transition prior $D_2(A|\Theta_2)$ is specified in detail to provide the basics for computing the optimal state-context trees and corresponding transition parameters during the training. Since each transition matrix $A_{\tau_l}\in A$ is defined by an underlying state-context tree τ_l that represents different classes of equivalent state-contexts that share their transition parameters, the typically used Dirichlet prior for transition parameters of a fixed state-context must be re-defined to enable the evaluation of different structures of the underlying state-context tree. This is realized as follows.

The transition prior for the set of transition matrices A is defined by

$$D_2(A|\Theta_2) := \prod_{l=1}^{L} D_2^l(A_{\tau_l}|\Theta_2^l)\cdot D_2^l(\tau_l|\varphi)$$

consisting of a product of transformed Dirichlet distributions $D_2^l(A_{\tau_l}|\Theta_2^l)$ in combination with a tree structure prior $D_2^l(\tau_l|\varphi)$ for each transition matrix $A_{\tau_l}\in A$. The corresponding hyper-parameters $\Theta_2 := (\Theta_2^1,\ldots,\Theta_2^L)$ are specified with respect to each hyper-parameter matrix $\Theta_2^l := (\vartheta_{ij})$ defining the pseudocounts $\vartheta_{ij}\in\mathbb{R}^+$ for a transition from a state-contexts $i\in S^l$ to a next state $j\in S$.

The transformed Dirichlet distributions

$$D_2^l(A_{\tau_l}|\Theta_2^l) := \prod_{\xi\in\tau_l} Z(\Theta_{2,\xi}^l)\prod_{j\in S}\exp(\Lambda_{a_{\xi j}}\cdot\vartheta_{\xi j}) \quad (2)$$

define the prior for the transition parameters of the transition matrix A_{τ_l} in dependency of the corresponding state-context tree τ_l. For each class of equivalent state-contexts ξ of the state-context tree τ_l underlying the transition matrix A_{τ_l}, a transformed Dirichlet distribution is specified. Each transition probability $a_{\xi j}$ of A_{τ_l} is parameterized in the log-space by $\Lambda_{a_{\xi j}} := \log(a_{\xi j})$. The corresponding hyper-parameter vector $\Theta_{2,\xi}^l := (\vartheta_{\xi j})_{j\in S}$ with $\vartheta_{\xi j} := \sum_{i\in\xi}\vartheta_{ij}$ is defined with respect to Θ_2^l, and the normalization constant is specified by $Z(\Theta_{2,\xi}^l) := \Gamma(\sum_{j\in S}\vartheta_{\xi j})/\prod_{j\in S}\Gamma(\vartheta_{\xi j})$ in dependency of the Gamma function $\Gamma(x)$ defined for all $x\in\mathbb{R}^+$.

The tree structure prior

$$D_2^l(\tau_l|\varphi)\propto\prod_{\xi\in\tau_l}\varphi \quad (3)$$

is defined for rating the state-context tree τ_l by its number of disjoint sets of equivalent state-contexts. During the training of a parsimonious higher-order HMM, the tree structure hyper-parameter $\varphi\in\mathbb{R}^+$ enables the regulation of the number of leaf nodes of a state-context tree influencing the tree structure of τ_l. A fixed value of $\varphi\in(0,1)$ leads to a decreased value of the tree structure prior for an increasing number of leaf nodes, whereas a fixed value of $\varphi>1$ leads to a greater value of the tree structure prior for an increasing number of leaf nodes.

The choice of hyper-parameter values for the prior distribution of a parsimonious higher-order HMM should provide the basics for distinguishing between DNA polymorphisms and unchanged chromosomal regions in Array-CGH profiles. A histogram of log-ratios (e.g. Figure 2a) helps to characterize the states of the model. Different values of the hyper-parameter of the tree structure prior are chosen to enable the interpolation of the parsimonious higher-order HMM between a mixture model and a higher-order HMM. The interval of tree structure hyper-parameter values that has to be considered for this interpolation is depending on the size of the Array-CGH data set. Details to the chosen hyper-parameter values of the prior distribution are given in the section Prior distribution in Text S1.

Bayesian Baum-Welch training. A Bayesian Baum-Welch algorithm is developed to adapt the initial parameters of a parsimonious higher-order HMM to Array-CGH profiles. This algorithm extends the commonly used Baum-Welch algorithm [62–65] by integrating prior knowledge into the parameter estimation. The Bayesian Baum-Welch algorithm is an iterative training procedure belonging to the class of Expectation Maximization (EM) algorithms [107] for maximizing the log-posterior density of the parameters of a parsimonious higher-order HMM for a given data set. This is done by iteratively computing new parameters of the parsimonious higher-order HMM

$$\lambda(h+1) = \underset{\lambda}{argmax}(Q(\lambda|\lambda(h)) + \log(P[\lambda|\Theta]))$$

under consideration of its parameters $\lambda(h)$ of the current iteration step h starting with initial parameters $\lambda(1)$. The parameter estimation is done based on Baum's auxiliary function $Q(\lambda|\lambda(h))$ in combination with the logarithm of the prior distribution $P[\lambda|\Theta]$ defined in (1).

Baum's auxiliary function is specified in [65] for a standard first-order HMM. Specific modifications are required for a parsimonious higher-order HMM due to the realization of the state-transition process by a parsimonious higher-order Markov model. In analogy to [65], Baum's auxiliary function is defined by

$$Q(\lambda|\lambda(h)) := Q_1(\vec{\pi}|\lambda(h)) + Q_2(A|\lambda(h)) + Q_3(B|\lambda(h))$$

consisting of an auxiliary function for each class of model parameters. No modifications are required for the auxiliary function $Q_1(\vec{\pi}|\lambda(h))$ of the initial state distribution $\vec{\pi}$ and for the auxiliary function $Q_3(B|\lambda(h))$ of the emission parameters B. Details to these two functions and the corresponding parameter estimation are given in the section Bayesian Baum-Welch algorithm in Text S1.

The auxiliary function for the set of transition matrices A is given by

$$Q_2(A|\lambda(h)) := \sum_{l=1}^{L} Q_2^l(A_{\tau_l}|\lambda(h))$$

providing the basis for the computation of each state-context tree τ_l representing optimal disjoint sets of equivalent state-contexts of length l and corresponding transition probabilities of the transition matrix $A_{\tau_l} \in A$. This requires the auxiliary function for each transition matrix A_{τ_l} given by

$$Q_2^l(A_{\tau_l}|\lambda(h)) := \begin{cases} \sum\limits_{\xi\in\tau_l}\sum\limits_{j\in S} \Lambda_{a_{\xi j}} \sum\limits_{k=1}^{K}\sum\limits_{i\in\xi} \varepsilon_l^k(i,j) & 1 \le l < L \\ \sum\limits_{\xi\in\tau_L}\sum\limits_{j\in S} \Lambda_{a_{\xi j}} \sum\limits_{k=1}^{K}\sum\limits_{t=L}^{T_k-1}\sum\limits_{i\in\xi} \varepsilon_t^k(i,j) & l = L \end{cases} \quad (4)$$

under consideration of the log-transition probability $\Lambda_{a_{\xi j}} := \log(a_{\xi j})$ and the probability $\varepsilon_t^k(i,j) := P[\vec{q}_{max(1,t-L+1)\ldots t} = i, q_{t+1} = j|\vec{o}(k), \lambda(h)]$ for a transition from state-context i to next state j given the Array-CGH profile $\vec{o}(k)$ and the current parameters of the parsimonious higher-order HMM. The log-transition probability $\Lambda_{a_{\xi j}}$ has to be estimated for the next parsimonious higher-order HMM $\lambda(h+1)$. Each probability $\varepsilon_t^k(i,j)$ is computed under the parsimonious higher-order HMM $\lambda(h)$ using extended versions of the standard Forward-Backward algorithm [65] as developed in [89]. Details for deriving $Q_2^l(A_{\tau_l}|\lambda(h))$ in (4) are provided in the section Bayesian Baum-Welch algorithm in Text S1.

For estimating the transition probabilities of transition matrix A_{τ_l}, the logarithm of the transition prior $D_2^l(A_{\tau_l}|\Theta_2^l)$ in (2) and the logarithm of the tree structure prior $D_2^l(\tau_l|\varphi)$ in (3) are added to the corresponding auxiliary function $Q_2^l(A_{\tau_l}|\lambda(h))$ in (4). The resulting function is then maximized by a dynamic programming approach [98,99] efficiently evaluating the set of all existing state-context trees τ_l. This results in an optimal state-context tree τ_l and a corresponding transition matrix A_{τ_l} for the next parsimonious higher-order HMM $\lambda(h+1)$. Details to the transition parameter estimation are given in the section Bayesian Baum-Welch algorithm in Text S1.

Generally, the applied dynamic programming approach starts with an initialization step having a computational complexity of $O\left((2^N-1)^L \cdot N^{L+1} \cdot T\right)$ in dependency of the number of hidden states N and the order L of the parsimonious HMM, and the length T of a processed emission sequence. The term $N^{L+1} \cdot T$ is standardly occurring for higher-order HMMs specifying the computational complexity required to compute all weights for the estimation of transition probabilities, and the term $(2^N-1)^L$ is specific for the dynamic programming approach used for the parsimonious higher-order HMMs.

The initialization step is followed by iteration steps that have a total computational complexity of $O\left(((2^N-1)^L-1)/((2^N-1)-1)\cdot(B_N \cdot N + 2^N - 2)\right)$. Here, $((2^N-1)^L-1)/((2^N-1)-1)$ specifies the number of iteration steps, and the Bell number B_N defines the number of partitions existing for N states growing faster than 2^N for $N>4$. Details to the derivation of the computational complexities of the initialization and the iteration steps are given in the section Bayesian Baum-Welch algorithm in Text S1.

The estimation of new parameters $\lambda(h+1)$ is iterated until the log-posterior density increases less than 10^{-9} for two successive iteration steps of the Baysian Baum-Welch algorithm. This iterative scheme reaches at least a local optimum in dependency of the initial parameters $\lambda(1)$ [107].

Model initialization. An initial parsimonious higher-order HMM has to distinguish between deletions or sequence deviations, unchanged chromosomal regions, and amplifications in an Array-CGH data set. A histogram of measured log-ratios (e.g. Figure 2a) assists to choose initial parameters for the state-specific Gaussian emission densities characterizing the three states of the model in Figure S2 in Text S1.

For the Array-CGH data set comparing the genomes of C24 and Col-0, the initial means of the state-specific Gaussian emission densities are set to $\mu_- = -3$, $\mu_= = 0$, and $\mu_+ = 1.5$. The initial standard deviation of the Gaussian emission density of each state $i \in S$ is set to $\sigma_i = 0.67$ according to the standard deviation of log-ratios in the Array-CGH data set.

The initial state distribution $\vec{\pi}$ is sampled from the prior distribution of the initial state distribution. Each initial transition matrix A_{τ_l} is sampled from its corresponding transition prior distribution by assuming an underlying complete state-context tree (e.g. Figure 3c) of a higher-order HMM. That means, a parsimonious higher-order HMM is initially representing a corresponding higher-order HMM.

Parsimonious HMMs of order one up to five have been considered for the analysis of the Array-CGH data set. For each model order, forty different model complexities ranging from the mixture model up to the corresponding higher-order HMM have been evaluated by using forty different values of the hyper-parameter φ of the tree structure prior in Equation (3). Details to

the chosen hyper-parameter values are given in the section Prior distribution in Text S1. For each of these forty different hyper-parameter values, twenty different initial models have been adapted to the Array-CGH data set using the Bayesian Baum-Welch training. Thus, in total 800 different models were computed for each model order. The best performing models with clearly reduced model complexities in comparison to higher-order HMMs were obtained for $\log(\varphi)$ in the range of -100 to 0.

Generally, apart from this in-depth study considering the Arabidopsis Array-CGH data set, a parsimonious higher-order HMM can be specified for the analysis of Array-CGH data by choosing appropriate values for the mean values of the Gaussian emission densities of the states '$-$' and '$+$'. The mean value of the Gaussian emission density of state '$=$' can be assumed to be zero, because unchanged chromosomal regions are expected to have log-ratios of about zero. The standard deviations of the state-specific Gaussian emission densities can be initially set to the standard deviations of the considered Array-CGH data set. Using the pre-defined hyper-parameter values for the prior distributions (see section Prior distribution in Text S1), good-performing models have been obtained on Arabidopsis and human Array-CGH profiles. Especially for model orders greater than one, good-performing models with a clearly reduced model complexity in comparison to the corresponding higher-order HMM have been obtained for choosing the tree structure hyper-parameter value $\log(\varphi)$ in the range of -100 to 0. This initialization concept is realized in the provided software and further specific hints are given in the corresponding documentation.

Related work in other domains: Variable-length Hidden Markov Models. Related to parsimonious higher-order HMMs, a variable-length HMM was developed in [96,97] for the analysis of motion capture data of modern human dance. The state-transition process of the variable-length HMM is defined by a variable memory Markov chain. The transition parameters of this Markov chain are determined by a minimum entropy criterion based on the Kullback-Leibler divergence integrated into an extended Baum-Welch training. The minimum entropy criterion is used for pruning or growing the state-contexts that are underlying the state-transition process. The Baum-Welch algorithm developed for the variable-length HMM does not enable the integration of prior knowledge into the training of model parameters.

In contrast to this, the parsimonious higher-order HMM is trained by a Bayesian Baum-Welch algorithm enabling the integration of prior knowledge. Especially for HMM -based analysis of DNA microarray data, the modeling of prior knowledge can have a substantial impact on the quality of analysis results [106]. Generally, the concept of pruning or growing of state-contexts developed for the variable-length HMM is related to the concept of determining sets of equivalent state-contexts forming the basis of the parsimonious higher-order HMM. The state-transition process of the parsimonious higher-order HMM is more flexible enabling shared transition probabilities due to fusions of nodes in the underlying state-context tree. This allows to model dependencies between non-directly adjacent states for which the intermediate states are not or only partially contributing to these dependencies. That is exemplarily illustrated in Figure 3b in which the right tree branch contains a partially fused node with non-completely fused child nodes. Such dependencies cannot be modeled by a variable-length HMM because pruning or growing only enables to shorten or extend state-contexts but not to fuse states.

Results/Discussion

In this section, first the modeling of spatial dependencies between Arabidopsis Array-CGH measurements is investigated to choose a range of model orders for parsimonious HMMs. Based on this, parsimonious HMMs of different model complexity are compared regarding their ability to identify deletions or sequence deviations in the Arabidopsis Array-CGH data set. Additionally, parsimonious HMMs are compared to existing methods utilizing the Arabidopsis and human cell lines Array-CGH data. Finally, a detailed functional classification of identified copy number polymorphisms or sequence deviations is made to investigate potential functions of genomic regions in which the genomes of C24 and Col-0 differ.

Choice of Model Order

The modeling of the partial autocorrelation function [108] of the Arabidopsis Array-CGH profiles by higher-order HMMs was initially studied to determine a range of model orders for an in-depth analysis by parsimonious HMMs. The partial autocorrelation function quantifies linear dependencies between measurements of chromosomal regions in close chromosomal proximity for an increasing distance of regions. As shown in Figure 1, such dependencies are clearly present in the Arabidopsis Array-CGH profiles motivating the application of HMMs of different model orders for modeling of these dependencies.

Initially, HMMs of order zero up to five were trained on the Arabidopsis Array-CGH profiles using the Bayesian Baum-Welch algorithm. Next, each HMM was used to sample 100 artificial profiles with 10,000 log-ratios. These profiles were used to compute the mean partial autocorrelation function modeled by each HMM.

As expected from theory, the HMM of order zero (mixture model) does not model dependencies between log-ratios in any chromosomal distance. The first-order HMM shows a clear improvement in comparison to the mixture model, but especially HMMs of order three up to five reached the best, nearly identical approximation of the partial autocorrelation function of Array-CGH profiles. A better modeling of the partial autocorrelation function by higher-order HMMs is expected from theory because of their more complex state-transition processes enabling an improved modeling of spatial dependencies compared to HMMs with a smaller model order. Still, none of these HMMs was able to perfectly approximate the partial autocorrelation structure of the Array-CGH profiles. But, despite of that, this study helped to determine a range of model orders for further analyses. The results of this study are summarized in Figure S3 in Text S1.

Based on this initial study with higher-order HMMs, parsimonious HMMs of order one up to five are subsequently investigated in detailed studies to analyze their abilities to identify DNA polymorphisms between C24 and Col-0.

Stringent Identification of Deletions or Sequence Deviations

An Array-CGH data set by [103] comparing the genomes of the accessions C24 and Col-0 of *A. thaliana* is used to identify polymorphic regions between both genomes by parsimonious higher-order HMMs. These models are evaluated based on deletions or sequence deviations determined in [101] for the genome of C24 in comparison to the reference genome of Col-0 using publicly available array-based resequencing data [100]. The mapping of these polymorphic regions to corresponding chromosomal regions in the Array-CGH data set shows an obvious coupling with potential deletions or sequence deviations present in

the Array-CGH data set (Figure 2b). These potential deletions or sequence deviations are used as reference for model comparisons.

Parsimonious higher-order HMMs of different model complexities were adapted to the Array-CGH data using the developed Bayesian Baum-Welch training. For each model, all chromosomal regions in the Array-CGH data set were ranked in decreasing order of their state-posterior probabilities of state '−' modeling deletions or sequence deviations. Using the knowledge about potential deletions or sequence deviations in the Array-CGH data set, the identification of these polymorphic regions was quantified for each model in terms of the true-positive-rate (TPR) at 1% false-positive-rate (FPR). The mean TPRs obtained for twenty different initializations of each model at 1% FPR are shown in Figure 4a (see Figure S4a in Text S1 for standard deviations of TPRs and see Figure S5a in Text S1 for FPRs at fixed TPR). The application of parsimonious higher-order HMMs has clearly improved the identification of deletions or sequence deviations in comparison to the standard first-order HMM. Moreover, parsimonious higher-order HMMs with much smaller model complexities than corresponding higher-order HMMs can also reach a clearly improved accuracy for identifying polymorphic regions in comparison to corresponding higher-order models. The best parsimonious higher-order HMMs have model complexities in the range of 3 up to 9 leaves. This range of model complexities includes parsimonious HMMs of order two up to five that nearly reach the same performance for identifying deletions or sequence deviations. State-context trees underlying well-performing parsimonious HMMs of order three up to five are clearly reduced leading to model complexities comparable with that of parsimonious second-order HMMs. Thus, not all higher-order dependencies are required for reaching a good performance at the stringent level of 1% FPR.

Similar results are shown in Figure S6a in Text S1 using a less restrictive mapping of the independently determined deletions or sequence deviations from [101] to the Array-CGH data set for model comparisons.

Less Stringent Identification of Deletions or Sequence Deviations

Parsimonious higher-order HMMs have initially been compared against the standard first-order HMM and higher-order HMMs at a stringent FPR of 1%. Next, these models are compared at a less stringent FPR of 2.5%. That leads to an identification of deletions or sequence deviations comparable with those obtained by applying the state-posterior decoding algorithm [65,89] that computes for each chromosomal region in the Array-CGH data set the most likely state under the given model. The results are shown in Figure 4b (see Figure S4b in Text S1 for standard deviations of TPRs and see Figure S5b in Text S1 for FPRs at fixed TPR).

Generally, parsimonious higher-order HMMs reach a higher accuracy for the identification of deletions or sequence deviations than the standard first-order HMM. The best parsimonious higher-order HMMs also reach an accuracy that is comparable or slightly better than that of corresponding higher-order HMMs. This accuracy is obtained at much lower model complexities than for higher-order HMMs. That can become particularly useful for avoiding overfitting in small data.

In comparison to the results at 1% FPR, the complexity of the best models is more shifted into the range of 9 to 27 leaves at 2.5% FPR (Figure 4 and Figure S4 in Text S1). This indicates that the identification of polymorphic regions is more complicated. Because at a higher FPR, the Array-CGH measurements of additionally identified polymorphic regions are more similar to that of non-polymorphic regions. These difficulties tend to be managed best by parsimonious higher-order HMMs. The best models in Figure 4b are among the fourth-order parsimonious higher-order HMMs.

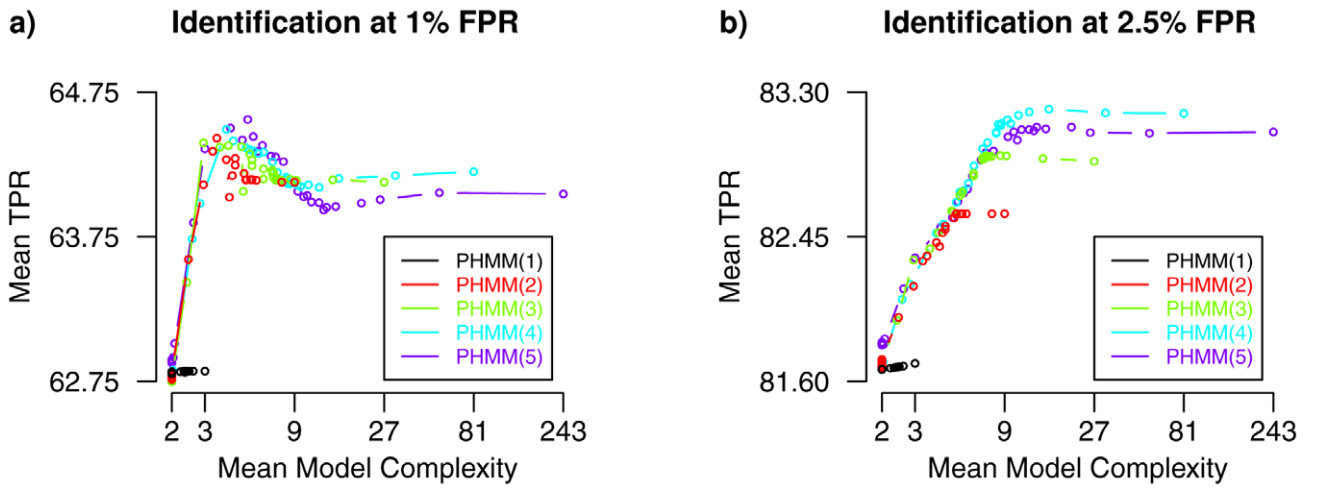

Figure 4. Identification of deletions and sequence deviations in the Arabidopsis Array-CGH data set by parsimonious HMMs. Curves of mean true-positive-rates (TPRs) for the identification of candidate regions of deletions or sequence deviations at a fixed false-positive-rate (FPR) of 1% (**a**)) and of 2.5% (**b**)) obtained by parsimonious HMMs of order $L \in \{1,\ldots,5\}$ of different model complexities across twenty different initializations. The rightmost point of each curve of parsimonious HMMs of order L (PHMM(L)) represents the corresponding higher-order HMM of order L with highest model complexity of 3^L leaf nodes in the state-context tree underlying the transition matrix A_{τ_L}. The rightmost point of the black curve represents the standard first-order HMM. Standard deviations of the mean TPRs are shown in Figure S4 in Text S1. At both levels of FPRs, parsimonious higher-order HMMs are clearly better than parsimonious HMMs of order one including the standard first-order HMM. At the level of 1% FPR, parsimonious higher-order HMMs with a mean model complexity in the range of 3 up to 9 also identify deletions or sequence deviations better than higher-order HMMs. At 2.5% FPR, clearly reduced model complexities are sufficient to reach identifications of deletions or sequence deviations by parsimonious higher-order HMMs comparable or slightly better than corresponding higher-order HMMs.
doi:10.1371/journal.pcbi.1002286.g004

A tree structure of one of the best models is shown in Figure 5. The underlying parsimonious fourth-order HMM has still some specific fourth-order transition probabilities for the states '−' (deletion or sequence deviation) and '=' (non-polymorphic), whereas those of state '+' (amplification) are completely reduced to second-order transition probabilities. This unbalanced reduction of transition parameters tends to be coupled with the asymmetry of the Array-CGH measurement distribution in Figure 2a. Most of the chromosomal regions in the Array-CGH data set are non-polymorphic, a small proportion tends to be deleted or affected by sequence deviations, whereas only a very small proportion of regions tends to be amplified. The tree structure indicates that these tendencies are transferred to the number of transition parameters per state. This parsimonious fourth-order HMM is considered in all further studies with the Arabidopsis Array-CGH data set because of its good performance at the level of 2.5% FPR comparable with the results obtained by applying the state-posterior decoding algorithm enabling an in-depth analyses of genomic differences between C24 and Col-0.

Generally, similar tendencies like shown in Figure 4b are also present in Figure S6b in Text S1 considering a less restrictive mapping of the independently determined deletions or sequence deviations from [101] to the Arabidopsis Array-CGH data set for model comparisons.

Comparison to Existing Methods

Here, the well-performing parsimonious fourth-order HMM is compared against other existing methods on the Arabidopsis data set. Then, another widely considered human cell lines data set by [102] is used for additional model comparisons. Subsequent to this, the focus is on comparative genomics of the accessions C24 and Col-0 of *A. thaliana*.

Comparison on Arabidopsis data. Next, the parsimonious fourth-order HMM with underlying tree structure shown in Figure 5 is compared to other existing methods for analyzing Array-CGH data. The standard method for the analysis of the Array-CGH data set measured on a NimbleGen tiling array is the segMNT algorithm [21]. Additionally, all eight methods provided by the ADaCGH webserver [43] including the best performing methods of two in-depth comparison studies [39,40] were applied to the Arabidopsis Array-CGH data set. From these eight methods, only ACE [33], CBS [20], FHMM [48], and GLAD [32] were able to manage the huge number of Array-CGH measurements. Besides FHMM, also three other methods based on first-order HMMs were considered for the comparison including wuHMM [56] and two Bayesian methods RJaCGH [55] and GHMM [52]. All methods were applied to the Arabidopsis Array-CGH data set using standard settings.

The identification of deletions or sequence deviations by these methods is compared against the predictions of the parsimonious fourth-order HMM with respect to the known potential deletions or sequence deviations characterized in Figure 2b. For this comparison, a receiver operating characteristic (ROC) curve was computed for each method. This was done by ranking all chromosomal regions of the Array-CGH data set according to their method-specific scores enabling the evaluation of identified deletions or sequence deviations under consideration of known potential polymorphic regions.

The ROC curves are shown in Figure 6. The two Bayesian HMMs RJaCGH and GHMM identify deletions or sequence deviations with a nearly identical accuracy and better than wuHMM and all methods provided by the ADaCGH webserver. This is further improved by the parsimonious fourth-order HMM identifying chromosomal regions affected by deletions or sequence deviations with higher accuracy than all other methods. Comparable results were obtained considering a less restrictive mapping of identified deletions or sequence deviations from [101] to the Array-CGH data set (Figure S7 in Text S1). The improved performance of the parsimonious fourth-order HMM for identifying deletions or sequence deviations in comparison to the standard first-order HMM is highlighted in the direct comparison shown in Figure S8 in Text S1. Again all these findings indicate that parsimonious higher-order HMMs are useful for the analysis of Array-CGH data.

Comparing the different HMM-based methods by the number of hidden states required for modeling chromosomal aberrations in the Arabidopsis data set, wuHMM and FHMM both determined seven states, RJaCGH used six, and GHMM and the parsimonious HMM required only three states for reaching the reported

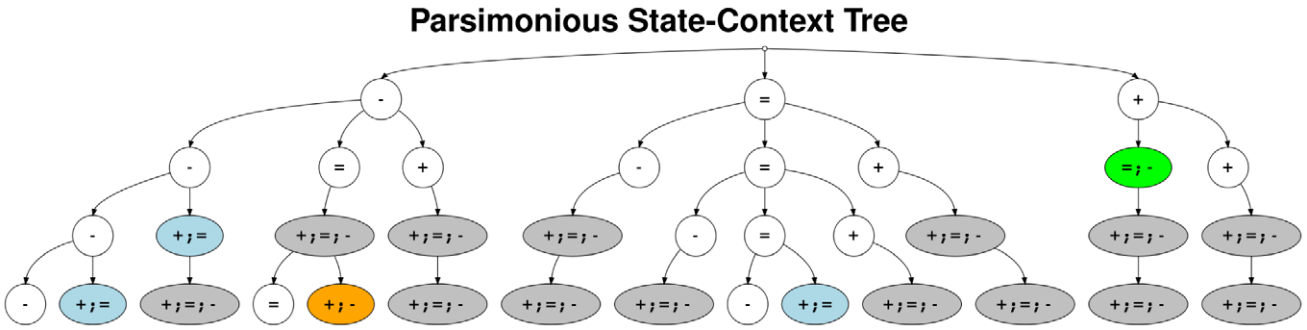

Figure 5. State-context tree of a parsimonious fourth-order HMM. Parsimonious state-context tree selected among the best parsimonious HMMs of order four at a fixed FPR of 2.5% in Figure 4b. Each path from the root node at the top of the tree to a leaf node at the bottom of the tree represents a set of state-contexts defined to share common transition parameters in the transition matrix A_{τ_4} of the selected model. The three nodes directly under the root node represent the possible current states of the selected parsimonious fourth-order HMM, and the subtrees under these three nodes represent the influence of predecessor states on a state-transition from one of these current states to a next state. Fusions of nodes are highlighted in different colors. White nodes represent unfused nodes characterizing important states for a state-transition. Blue, orange, and green nodes represent partially fused states of equal importance for a state-transition. Grey nodes represent completely fused nodes defining that the corresponding position in a state-context has no influence on a state-transition. The states '−' and '=' of the selected model are still representing some fourth-order transition probabilities, whereas only second-order transition probabilities remain for state '+'. The selected parsimonious fourth-order HMM has a model complexity of 14 leaf nodes leading to 42 different transition parameters in $A\tau_4$. This is much less than for a corresponding fourth-order HMM with 81 leaf nodes in a complete state-context tree representing 243 transition parameters.
doi:10.1371/journal.pcbi.1002286.g005

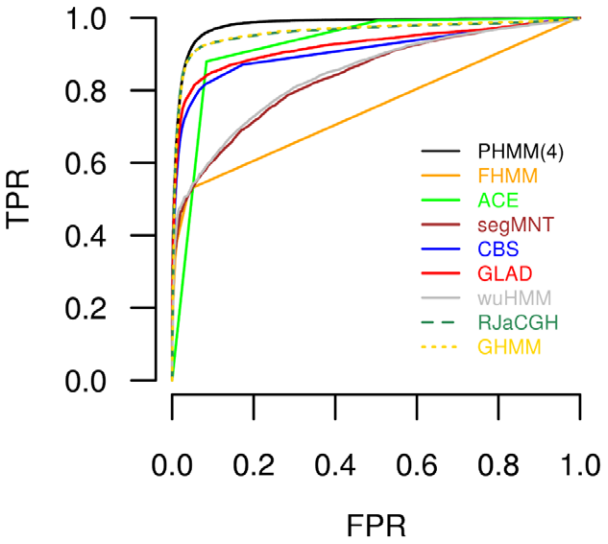

Figure 6. Comparison of a parsimonious fourth-order HMM to existing methods on the Arabidopsis Array-CGH data set. Receiver operating characteristic (ROC) curves for comparing the identification of deletions or sequence deviations in the Array-CGH data set. ROC curves are shown for FHMM, ACE, CBS, and GLAD of the ADaCGH webserver [43], segMNT [21], wuHMM [56], GHMM [52], RJaCGH [55] and the parsimonious fourth-order HMM with underlying state-context tree in Figure 5. The parsimonious fourth-order HMM reaches the best identification of deletions and sequence deviations (black).
doi:10.1371/journal.pcbi.1002286.g006

performance. Supported by the best identification of deletions or sequence deviations, this indicates that the three states of the parsimonious HMM are flexible enough for modeling complex Arabidopsis Array-CGH profiles.

The two good-performing Bayesian HMMs RJaCGH and GHMM had substantially different run-times. RJaCGH required 30 hours and 42 minutes for analyzing the Arabidopsis data set, while GHMM only required about 24 minutes. An overview of run-times of all methods is given in Table 1. The training of a parsimonious first-order HMM on the Arabidopsis data set took about 2 minutes. This time is increased by a factor of three (number of hidden states) for increasing model order leading to a training time of about 54 minutes for the parsimonious fourth-order HMM. Using such a trained parsimonious HMM, analyses of data sets with a similar measurement distribution (e.g. comparisons of other accessions against Col-0) can be obtained in less than five minutes.

Table 1. Method run-times on the Arabidopsis Array-CGH data set.

Shortcut	Method	Reference	Computing time
wuHMM	First-order HMM	[56]	8 min
GHMM	Bayesian first-order HMM	[52]	24 min
PHMM	Parsimonious fourth-order HMM	see Methods	54 min
CBS	Circular Binary Segmentation	[20]	1 h 18 min
ACE	Analysis of Copy Errors	[33]	4 h 14 min
GLAD	Gain and Loss Analysis of DNA	[32]	4 h 19 min
FHMM	First-order HMM	[48]	5 h 04 min
RJaCGH	Bayesian first-order HMM	[55]	30 h 42 min

Run-times in hours/minutes required for the analysis of the Arabidopsis Array-CGH data set by the different methods. All methods except GHMM, PHMM, wuHMM, and RJaCGH were run on the ADaCGH web-server [43] (AMD Opteron 2.2 GHz CPU with 6 GB RAM). The other methods GHMM, PHMM, wuHMM, and RJaCGH were run on a standard desktop computer with Intel CPU T9500 2.6 GHz and 4 GB RAM.
doi:10.1371/journal.pcbi.1002286.t001

In summary, this study further illustrated that parsimonious higher-order HMMs can outperform existing methods and are well-suited for analyzing Arabidopsis Array-CGH data. It should also be noted that experts of specific methods might be able to improve the results of individual methods by fine-tuning of specific parameters. Still, parsimonious higher-order HMMs represent an important contribution to the field of Array-CGH data analysis because they combine improved modeling of spatial dependencies with the integration of prior knowledge and because these models have reached a good performance on the Arabidopsis Array-CGH data.

Comparison on human cell lines. Additional model evaluations were also done on Array-CGH data of human cell lines [102] frequently considered in other model comparison studies like e.g. [20,32,48,52,55]. Details to the cell lines and the study are given in the section Model evaluations on human cell lines in Text S1. Using standard settings, six methods from the ADaCGH webserver [43], wuHMM [56], RJaCGH [55], and GHMM [52] were compared against a parsimonious first-order HMM to evaluate the identification of known trisomies and monosomies in the human cell lines. The resulting ROC curves are shown in Figure S9 in Text S1. The parsimonious first-order HMM, but also both Bayesian HMMs RJaCGH and GHMM reach the best, nearly perfect identification of known chromosomal aberrations in the individual human cell lines.

Considering the run-times on the human data set with about 17.5 times less measurements than in the Arabidopsis data set, RJaCGH required the longest time with about seventy minutes. Both, the GHMM and the parsimonious first-order HMM required only about one minute for analyzing the human cell lines. A summary of run-times of the ten different tested methods is given in Table S1 in Text S1. This additional study indicates that parsimonious HMMs are also useful for the analysis of non-plant-specific Array-CGH data.

Functional Analysis of Genomic Differences between C24 and Col-0

The genome annotation of the reference genome of Col-0 provides the opportunity to investigate what is functionally behind chromosomal regions where the genomes of C24 and Col-0 differ. The parsimonious fourth-order HMM with underlying parsimonious tree structure in Figure 5 was applied to identify polymorphic regions in the Arabidopsis Array-CGH data set. The state-posterior decoding algorithm [65,89] was used to classify each chromosomal region in the Array-CGH data set either as a deletion or sequence deviation, as unchanged, or as an amplification in C24 with respect to the reference genome of Col-0. This algorithm assigns the most likely state of the three-state architecture of the HMM (Figure S2 in Text S1) to each chromosomal region measured in the Array-CGH data set. The identification of deletions or sequence deviations by state-posterior decoding is comparable to that shown in Figure 4b.

In total, about 4.7% (17,306 of 364,339) of all chromosomal regions of the reference genome of Col-0 were identified as being affected by deletions or sequence deviations in the genome of C24, and about 0.2% (855 of 364,339) of all chromosomal regions were

identified as amplified in C24 (Table S1). This asymmetry in predictions is expected from the distribution of measurements in the Array-CGH data set (Figure 2) reflecting the design of the tiling array that only represents chromosomal regions present in the reference genome of Col-0 [11]. Of the 17,306 chromosomal regions identified as being affected by deletions or sequence deviations, 2,647 are singletons consisting of only one tile and 76.5% of these singletons are containing a micro-deletion or sequence deviation in C24 compared to Col-0 that is covering at least 40% of the underlying tile. In all, genomic regions affected by deletions or sequence deviations represent about 5.59 Mb of the Col-0 reference genome. This is in good accordance with the findings in [100,101]. Subsequently, all identified genomic differences are analyzed in detail.

Genome annotation analysis. Chromosomal regions identified as being affected by deletions or sequence deviations and regions identified as being affected by amplifications were analyzed separately using the *Arabidopsis* Information Resource (TAIR8) genome annotation of Col-0 [109]. The results of these functional categorizations are summarized in Figure 7.

By definition, the TAIR8 categories are not completely disjoint meaning that each chromosomal region can have annotations in more than one category (e.g. chromosomal regions within genes). Comparisons of the identified polymorphic regions in C24 to randomly chosen control sets revealed that a significant proportion of chromosomal regions affected by deletions or sequence deviations and also that regions affected by amplifications are caused by transposons. Such mobile genomic elements were also identified to be involved in rearrangements of the genomes of other accessions of *A. thaliana* [17,100,110]. Moreover, genic regions as well as 5′ and 3′ untranslated regions (UTRs) are significantly less affected by amplifications and deletions or sequence variations.

Thus, genomic differences between C24 and Col-0 do not occur randomly because transposons differ more than other parts of the genome. These results are also supported by the finding that transposons change faster than genes [111].

Ontology classification of genes. Ontology classification was performed for genes affected by amplifications and for genes affected by deletions or sequence deviations using the MIPS Functional Catalogue [112] to investigate if specific functional categories of genes are over-represented.

No prevalence of any functional category was found for the 39 genes affected by amplifications. In contrast to this, among the 1,675 genes affected by deletions or sequence deviations, five significantly over-represented functional clusters of genes with p-values less than $5{\cdot}10^{-6}$ were identified. The first cluster comprises 104 genes with functions in ATP-binding, the second cluster contains 109 genes with functions in cellular communication and signal transduction, the third cluster represents 127 genes playing a role in cell rescue, defense and virulence, the fourth cluster contains 5 genes encoding for N-actetylglucosamine deacetylases, and the fifth cluster comprises 541 unclassified proteins.

In coincidence with these findings, over-representations of sequence polymorphisms in defense-related genes or genes involved in signaling were previously identified in different accessions of *A. thaliana* [17,100]. Also the over-representation of deletions or sequence deviations in genes involved in ATP-binding, such as genes encoding for transporters or enzymes, might represent a functional adaptation to specific environmental conditions [113]. Copy number variations in N-actetylglucosamine deacetylases were recently reported for *A. thaliana* grown under different temperature conditions [114].

In summary, the five identified gene clusters with increased rate of deletions or sequence deviations indicate a rapid evolutionary change between C24 and Col-0. All genes affected by deletions or sequence deviations are provided in Table S2, and genes affected by amplifications are provided in Table S3.

Superfamily analysis of transposons. A superfamily classification of transposons affected by deletions or sequence deviations and of transposons affected by amplifications was performed using the TAIR8 transposon annotation of Col-0 to identify under- or over-representations of specific transposon superfamilies in C24. This analysis was done in comparison to randomly sampled control sets of transposons. The results are summarized in Figure 8.

Retrotransposons (LTR/Copia, LINE/L1) moving by a RNA-mediated copy-and-paste mechanism and DNA transposons (DNA, DNA/En-Spm,DNA/Harbinger) moving by a DNA-mediated cut-and-paste mechanism are significantly over-represented among the 2,695 transposons identified as being affected by

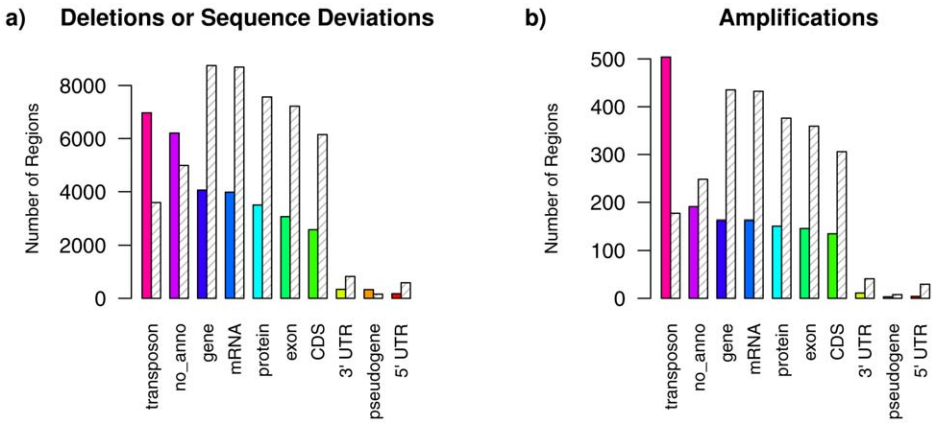

Figure 7. Functional classification of genomic differences in the Arabidopsis Array-CGH data set. Functional classification of the 17,306 tiles identified to be affected by deletions or sequence deviations (**a**)) and of the 855 tiles identified to be affected by amplifications (**b**)) in C24 in comparison to Col-0 according to the categories of the TAIR8 genome annotation. Colored bars show the counts in each category obtained for the Array-CGH data set by using the state-posterior decodings of the parsimonious fourth-order HMM with underlying state-context tree structure in Figure 5. Grey dashed bars represent the mean values of counts in each category obtained by sampling 500 times 17,306 tiles (or 855 tiles) from the total number of tiles in the Array-CGH data set. All counts in the different categories obtained for the Array-CGH data set, except 'pseudogene' for tiles identified as amplified, differ significantly from the random counts with p-values less than 0.01.
doi:10.1371/journal.pcbi.1002286.g007

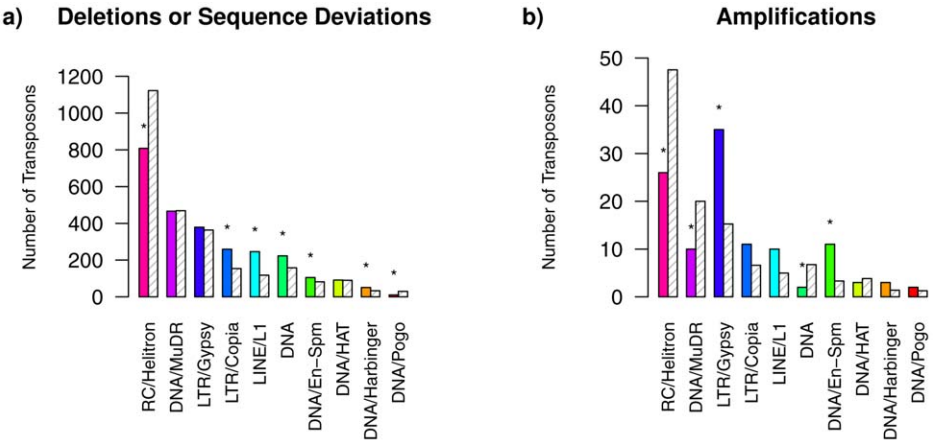

Figure 8. Superfamily classification of transposons in the Arabidopsis Array-CGH data set. Superfamily classification of the 2,695 transposons identified to be affected by deletions or sequence deviations (**a**)) and of the 114 transposons identified to be affected by amplifications (**b**)) under consideration of the TAIR8 transposon annotation. Colored bars show the numbers of affected transposons in each superfamily identified using the state-posterior decoding of the parsimonious fourth-order HMM with underlying state-context tree structure in Figure 5. Grey dashed bars represent the mean number of transposons assigned to these superfamilies for sampling 500 times 2,695 transposons (or 114 transposons) from the total number of transposons of the TAIR8 annotation. Superfamilies highlighted by an asterisk '*' are significantly different (over- or under-represented) with p-values less than 0.01 in comparison to random sampling.
doi:10.1371/journal.pcbi.1002286.g008

deletions or sequence deviations in C24 with respect to the reference genome of Col-0. DNA transposons (RC/Helitron, DNA/Pogo) are significantly under-represented among the 2,695 affected transposons.

For transposons affected by amplifications, retrotransposons (LTR/Gypsy) and DNA transposons (DNA/En-Spm) are significantly over-represented among the 114 transposons identified as being affected by amplifications in C24. DNA transposons (RC/Helitron, DNA/MuDR, DNA) are significantly under-represented among these 114 transposons.

Thus, these results indicate that some transposon superfamilies tend to play a more prevalent role for driving the evolution of genomic differences between C24 and Col-0. All these transposons represent fundamental components of *A. thaliana* genomes contributing to size, structure, and variation of genomes [115,116]. Table S4 provides all transposons identified to be affected by deletions or sequence deviations, and transposons affected by amplifications are contained in Table S5.

Conclusions

The development of parsimonious higher-order HMMs for the analysis of Array-CGH data has been motivated by the observation of strong spatial dependencies between measurements in close chromosomal proximity. A parsimonious higher-order HMM represents an interpolation between a mixture model ignoring spatial dependencies and a higher-order HMM exhaustively modeling spatial dependencies. To enable this interpolation, the mathematical theory of widely used first-order HMMs has been extended. A central point is the extension of the Bayesian Baum-Welch training by incorporating a dynamic programming approach [98,99] enabling a data-dependent modeling of spatial dependencies.

In a detailed study based on Array-CGH data for comparing the genomes of the *Arabidopsis thaliana* accessions C24 and Col-0, parsimonious higher-order HMMs clearly improved the identification of deletions or sequence deviations in comparison to typically used first-order HMMs and other existing methods. Especially, parsimonious HMMs of order three up to five with clearly reduced model complexities in comparison to corresponding higher-order HMMs reached the best results.

In-depth functional analyses of identified DNA polymorphisms revealed that most of these genomic differences between C24 and Col-0 are caused by transposons. Genic regions as well as 5′ and 3′ untranslated regions are less affected, but still genes with functions in ATP-binding, cellular signaling, or cell pathogen defense have been found to be specifically affected by deletions or sequence deviations in C24 in comparison to the reference genome of Col-0. These findings are in accordance with other studies [17,100] and might indicate specific environmental adaptations of both accessions. Additionally, a superfamily classification of transposons has revealed that specific retrotransposon and DNA transposon superfamilies tend to be more involved than others in driving the evolution of C24 and Col-0.

Additional model evaluations performed on widely considered human cell lines showed that parsimonious HMMs are also well-suited for the analysis of non-plant-specific Array-CGH data sets.

All these results indicate that parsimonious higher-order HMMs are useful tools for the analysis of Array-CGH data. Potential future applications could include other domains in which standard first-order HMMs are frequently used. This might include the HMM -based analysis of ChIP-chip data [117–120] or the analysis of next-generation sequencing data [121–125].

Supporting Information

Table S1 Arabidopsis Array-CGH data set including detected DNA polymorphisms identified by the parsimonious fourth-order HMM using the state-posterior decoding algorithm.
(TXT)

Table S2 Genes affected by deletions or sequence deviations in C24.
(TXT)

Table S3 Genes affected by amplifications in C24.
(TXT)

Table S4 Transposons affected by deletions or sequence deviations in C24.
(TXT)

Table S5 Transposons affected by amplifications in C24.
(TXT)

Text S1 Mathematical basics of prior distributions for initial state and emission parameters and details to chosen prior parameters are given in the section 'Prior distribution'. A detailed derivation of the Bayesian Baum-Welch algorithm for a parsimonious higher-order HMM is given in the section 'Bayesian Baum-Welch algorithm'. Details to the case study on human cell lines are given in the section 'Model evaluations on human cell lines'. The supporting Figures S1, S2, S3, S4, S5, S6, S7, S8, S9 are provided in the section 'Supporting Figures'.
(PDF)

Acknowledgments

We thank Ali M. Banaei Moghaddam, Andreas Houben and Michael Florian Mette from the IPK Gatersleben, and François Roudier and Vincent Colot from the IBENS Paris for the Arabidopsis Array-CGH data set and valuable discussions. We thank Jan Grau from the MLU Halle and Jens Keilwagen from the IPK Gatersleben for providing basics within Jstacs. We thank the reviewers for their valuable comments helping to improve the manuscript.

Author Contributions

Wrote the paper: M. Seifert. Implemented basic algorithms for parsimonious HMMs: M. Seifert. Implemented dynamic programming approach: A. Gohr. Combined algorithms: M. Seifert, A. Gohr. Designed studies: M. Seifert, M. Strickert, I. Grosse. Performed studies: M. Seifert. Read and commented manuscript: A. Gohr, M. Strickert, I. Grosse.

References

1. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, et al. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Canc 20: 399–407.
2. 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. Nat Genet 20: 207–211.
3. Mantripragada KK, Buckley PG, de Stahl TD, Dumanski JP (2004) Genomic microarrays in the spotlight. Trends Genet 20: 87–94.
4. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. Nat Genet 37: S11–S13.
5. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, et al. (2005) Applications of DNA tiling arrays for whole-genome analysis. Genomics 85: 1–15.
6. Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, et al. (2002) Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. Cancer Res 62: 6240–6245.
7. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, et al. (2002) 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: 12963–12968.
8. Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamäki EH, et al. (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. Oncogene 24: 1794–1801.
9. Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, de Medina SGD, et al. (2006) Regional copy number-independent deregulation of transcription in cancer. Nat Genet 38: 1386–1396.
10. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.
11. The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.
12. Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513–523.
13. Martienssen RA, Doerge RW, Colot V (2005) Epigenomic mapping in Arabidopsis using tiling microarrays. Chromosome Res 13: 299–308.
14. Fan C, Vibranovski MD, Chen Y, Long M (2007) A microarray based genomic hybridization method for identification of new genes in plants: Case analyses of Arabidopsis and Oryza. J Integr Plant Biol 49: 915–926.
15. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, et al. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet 39: 1151–1155.
16. Gregory BD, Yazaki J, Ecker JR (2008) Utilizing tiling microarrays for whole-genome analysis in plants. Plant J 53: 636–644.
17. Childs LH, Witucka-Wall H, Günther T, Sulpice R, Korff MV, et al. (2010) Single feature polymorphism (SFP)-based selective sweep identification and association mapping of growth-related metabolic traits in arabidopsis thaliana. BMC Genomics 11: 188.
18. Hodgson G, Hager JH, 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.
19. Olshen AB, Venkatraman ES (2002) Change-point analysis of array-based comparative genomic hybridization data. Proceedings of the Joint Statistical Meetings American Statistical Association. AlexandriaVA: American Statistical Association. pp 2530–2535.
20. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557–572.
21. Roche NimbleGen, Inc (2008) A Performance Comparison of Two CGH Segmentation Analysis Algorithms: DNACopy and segMNT. Available: http://www.nimblegen.com.
22. Jong K, Marchiori E, Vaart Avd, Ylstra B, Weiss M, et al. (2003) Chromosomal Breakpoint Detection in Human Cancer. Lect Notes Compt Sci 2611: 107–116.
23. Jong K, Marchiori E, Meijer G, Vaar Avd, Ylstra B (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. Bioinformatics 20: 3636–3637.
24. Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, et al. (2004) A versatile statistical analysis algorithm to detect genome copy number variation. Proc Natl Acad Sci USA 101: 16292–16297.
25. Price TS, Regan R, Mott R, Hedman A, Honey B (2005) 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: 3455–3464.
26. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ (2005) A statistical approach for array CGH data analysis. BMC Bioinformatics 6: 27.
27. Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirami R (2005) A method for calling gains and losses in array CGH data. Biostatistics 6: 45–58.
28. Pique-Regi R, Monso-Verona J, Ortega M, Seeger RC, Triche TJ, et al. (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24: 309–318.
29. Nilsson B, Johansson M, Al-Shahrour F, Carpenter AE, Ebert BL (2009) Ultrasome: efficient aberration caller for copy number studies of ultra-high resolution. Bioinformatics 25: 1078–1079.
30. Morganella S, Cerulo L, Viglietto G, Ceccarelli M (2010) VEGA: variational segmentation for copy number detection. Bioinformatics 26: 3020–3027.
31. Myers CL, Dunham MJ, Kung SY, Troyanskaya OG (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. Bioinformatics 20: 3533–3543.
32. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 20: 3413–3422.
33. Lingjaerde OC, Baumbusch LO, Liestol K, Glad IG, Borresen-Dale AL (2005) CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics 21: 821–822.
34. Eilers PHC, de Menezes RX (2005) Quantile smoothing of array CGH data. Bioinformatics 21: 1146–1153.
35. Huang T, Wu B, Lizardi P, Zhao H (2005) Detection of DNA copy number alterations using penalized least squares regression. Bioinformatics 21: 3811–3817.
36. Gao X, Huang J (2010) A robust penalized method for the analysis of noisy DNA copy number data. BMC Bioinformatics 11: 517.
37. Hsu L, Self SG, Grove D, Randolph T, Wang K, et al. (2005) Denoising array-based comparative genomic hybridization data using wavelets. Biostatistics 6: 211–226.
38. Ben-Yaacoc E, Eldar YC (2008) A fast and flexible method for the segmentation of aCGH data. Bioinformatics 24: i139–i145.
39. Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21: 3763–3770.
40. Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics 21: 4084–4091.
41. Liva S, Hupé P, Neuvial P, Brito I, Viara E (2006) CAPweb: a bioinformatics CGH array Analysis Platform. Nucleic Acids Res 34: W477–W481.
42. Conde L, Montaner D, Burguet-Castell J, Tárraga J, Medina I, et al. (2007) ISACGH: a webbased environment for the analysis of Array CGH and gene expression which includes functional profiling. Nucleic Acids Res 35: W81–W85.

43. Diaz-Uriarte R, Rueda OM (2007) ADaCGH: A Parallelized Web-Based Application and R Package for the Analysis of aCGH Data. PLoS ONE 2: e737.
44. Lai W, Choudhary V, Park PJ (2008) CGHweb: a tool for comparing DNA copy number segmentations from multiple algorithms. Bioinformatics 24: 1014–1015.
45. Snijders AM, Fridlyand J, Mans DA, Segraves R, Jain AN, et al. (2003) Shaping of tumor and drug-resistent genomes by instability and selection. Oncogene 22: 4370–4379.
46. Zhao X, Li C, Paez JG, Chin K, Jänne PA, et al. (2004) An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. Cancer Res 64: 3060–3071.
47. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-Scale Copy Number Polymorphisms in the Human Genome. Science 305: 525–528.
48. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN (2004) Hidden Markov models approach to the analysis of array CGH data. J Multivariate Anal 90: 132–153.
49. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al. (2005) A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. Cancer Res 65: 6071–6079.
50. Marioni JC, Thorne NP, Tavaré S (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics 22: 1144–1146.
51. Engler DA, Mohapatra G, Louis DN, Betensky RA (2006) A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. Biostatistics 7: 399–421.
52. Guha S, Li Y, Neuberg D (2008) Bayesian Hidden Markov Modeling of Array CGH Data. J Amer Statist Assoc 103: 485–497.
53. Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, et al. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22: 431–439.
54. Stjernqvist S, Ryden T, Sköld M, Staaf J (2007) Continuous-index hidden Markov modeling of array CGH copy number data. Bioinformatics 23: 1006–1014.
55. Rueda OM, Diaz-Uriarte R (2007) Flexible and accurate detection of genomic copy-number changes from aCGH. PLoS Comput Biol 3: e122.
56. Cahan P, Godfrey LE, Eis PS, Richmond TA, Selzer RR, et al. (2008) wuHMM: a robust algorithm to detect DNA copy number variation using oligonucleotide microarray data. Nucleic Acids Res 36: 1–11.
57. Andersson R, Bruder CEG, Piotrowski A, Menzel U, Nord H, et al. (2008) A segmental maximum a posteriori approach to genome-wide copy number profiling. Bioinformatics 24: 751–758.
58. Rueda OM, Diaz-Uriarte R (2009) RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions. Bioinformatics 25: 1959–1960.
59. Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand F, et al. (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41: 424–429.
60. Seifert M, Banaei A, Keilwagen J, Mette MF, Houben A, et al. (2009) Array-based genome comparison of Arabidopsis ecotypes using Hidden Markov Models. In: Encarnação P, Veloso A, eds. BIOSIGNALS 2009: Proceedings of the International Conference on Bio-inspired Systems and Signal Processing; 14–17 January 2009 INSTICC Press. pp 3–11.
61. Zöllner S (2010) CopyMap: localization and calling of copy number variation by joint analysis of hybridization data from multiple individuals. Bioinformatics 26: 2776–2777.
62. Baum LE, Eagen JA (1967) An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to model for ecology. Bull Amer Math Soc 73: 360–363.
63. Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Statists 41: 164–171.
64. Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3: 1–8.
65. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc IEEE 77: 257–286.
66. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis – Probabilistic models of proteins and nucleic acids Cambridge University Press.
67. Mac Donald IL, Zucchini W (1997) Hidden Markov and Other Models for Discrete-valued Time Series Chapman & Hall.
68. Jelinek F (1998) Statistical Methods for Speech Recognition The MIT Press.
69. Kriouile A, Mari JF, Haton JP (1990) Some improvements in speech recognition based on HMM. In: Internaional Conference on Accoustics, Speech, and Signal Processing, 1990; 3–6 April 1990; Albuquerque, New Mexico, United States. doi: 10.1109/ICASSP.1990.115770.
70. Watson B, Tsoi AC (1992) Second Order Hidden Markov Models for Speech Recognition. In: Proceedings of the 4th Australian International Conference on Speech Science and Technology; December 1992; Brisbane, Australia: ASSTA. pp 146–151.
71. Mari JF, Haton JP (1994) Automatic word recognition based on second-order hidden Markov models. In: Third International Conference on Spoken Language Processing; 18–22 September; Yokohama, Japan. pp 247–250.
72. Mari JF, Fohr D, Junqua JC (1996) A second-order HMM for high-performance word and phoneme-based continuous speech recognition. In: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. doi: 10.1109/ICASSP.1996.541126.
73. Mari JF, Halton JP, Kriouile A (1997) Automatic word recognition based on second-order hidden Markov models. In: IEEE Transactions of Speech and Audio Processing 5: 22–25.
74. de Villiers E, du Preez J (2001) The advantage of using higher order HMM's for segmenting acoustic files. In: 12th Annual Symposium of the Pattern Recognition Association of South Africa. Franschhoek, South Africa. pp 120–122.
75. Lee LM, Lee JC (2006) A Study on High-Order Hidden Markov Models and Applications to Speech Recognition. Lect Notes Compt Sci 4031: 682–690.
76. Engelbrecht HA, du Preez JA (2009) Efficient backward decoding of high-order hidden Markov models. Pattern Recogn 43: 99–112.
77. Derrode S, Carincotte C, Bourennane S (2004) Unsupervised image segmentation based on highorder hidden Markov chains. Markov chains, International Conference on Acoustics, Speech and Signal Processing (ICASSP 04). pp 769–772.
78. Mari JF, Le Ber F (2006) Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. Soft Comput 10: 406–414.
79. Benyoussef L, Carincotte C, Derrode S (2008) Extension of Higher-Order HMC Modeling with Applications to Image Segmentation. Digit Signal Process 18: 849–860.
80. Aycard O, Mari JF, Washington R (2004) Learning to automatically detect features for mobile robots using second-order Hidden Markov Models. Int J Adv Robotic Sy 1: 231–245.
81. Nel EM, du Preez JA, Herbst BM (2005) Estimating the pen trajectories of static signatures using hidden Markov models. IEEE Trans Pattern Anal Mach Intell 27: 1733–1746.
82. Ching WK, Fung ES, Ng MK (2003) Higher-Order Hidden Markov Models with Applications to DNA Sequences. IDEAL, Lect Notes Compt Sci 2690: 535–539.
83. Bouqata B, Carothers CD, Szymanski BK, Zaki MJ (2006) VOGUE: A Novel Variable Order-Gap State Machine for Modeling Sequences. Lect Notes Compt Sci 4213: 42–54.
84. Eng C, Asthana C, Aigle B, Hergalant S, Mari JF, et al. (2009) A New Data Mining Approach for the Detection of Bacterial Promoters Combining Stochastic and Combinatorial Methods. J Comp Biol 16: 1211–1225.
85. Zaki MJ, Carothers CD, Szymanski BK (2010) VOGUE: A Variable Order Hidden Markov Model with Duration based on Frequent Sequence Mining. ACM Trans Knowl Discov Data 4: Article 5.
86. Schimert J (1992) A high order hidden Markov model. PhD Thesis. University of Washington.
87. du Preez JA (1998) Efficient higher-order hidden Markov modeling. PhD Thesis. University of Stellenbosch. Available: www.ussigbase.org/downloads/jadp phd.pdf.
88. Hadar U, Messer H (2009) High-order Hidden Markov Models - estimation and implementation. In: Statistical Signal Processing; 31 August–3 September 2009; Cardiff, UK. doi: 10.1109/SSP.2009.5278591.
89. Seifert M (2010) Extensions of Hidden Markov Modles for the analysis of DNA microarray data. PhD Thesis. University of Halle-Wittenberg. Available: http://nbn-resolving.de/urn:nbn:de:gbv:3:4-4110.
90. Berchtold A, Raftery AE (2002) The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. Stat Sci 17: 328–356.
91. du Preez JA (1998) Efficient training of high-order hidden Markov models using first-order representations. Comput Speech Lang 12: 23–39.
92. du Preez JA, Weber DM (1998) Efficient Higher-Order Hidden Markov Modelling. In: Proceedings of the International Conference on Spoken Language Processing; 30 November–4 December 1998; Sydney, Australia: paper 1073.
93. Schwardt L, du Preez JA (2000) Efficient Mixed-Order Hidden Markov Model Inference. In: Volume 2, Proceedings of the International Conference on Spoken Language Processing-2000; 16–20 October 2000; Beijing, China. pp 238–241.
94. Schwardt L, du Preez JA (2000) Automatic Language Identification Using Mixed-Order HMMs and Untranscribed Corpora. In: Volume 2, Proceedings of the International Conference on Spoken Language Processing-2000; 16–20 October 2000; Beijing, China. pp 254–257.
95. Schwardt L (2007) Efficient Mixed-Order Hidden Markov Model Inference. PhD Thesis. University of Stellenbosch.
96. Wang Y (2006) The Variable-length Hidden Markov Model and Its Applications on Sequential Data Mining. Technical Report, Department of Computer Science, Tsinghua University, Beijing, China.
97. Wang Y, Zhou L, Feng J, Wang J, Liu ZQ (2006) Mining Complex Time-Series Data by Learning Markovian Models. In: Sixth International Conference on Data Mining; 18–22 December 2006; Hong Kong, China. pp 1136–1140.
98. Bourguignon PY, Robelin D (2004) Modèles de Markov parcimonieux: sélection de modèle et estimation. Noûs 48: 1–12.
99. Gohr A (2006) The Idea of Parsimony in Tree Based Statistical Models - Parsimonious Markov Models and Parsimonious Bayesian Networks with

Applications to Classification of DNA Functional Sites. Diploma Thesis. Martin Luther University Halle-Wittenberg.
100. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, et al. (2007) Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. Science 317: 338–342.
101. Zeller G, Clark RM, Schneeberger K, Bohlen A, Weigel D, et al. (2008) Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Res 18: 918–929.
102. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 29: 263–264.
103. Banaei AM, Roudier F, Seifert M, Bérard C, Martin Magniette ML, et al. (2011) Additive inheritance of histone modifications in Arabidopsis thaliana intraspecific hybrids. Plant J 67: 691–700.
104. MacKay DJC (1998) Choice of Basis for Laplace Approximation. Mach Learn 33: 77–86.
105. Evans M, Hastings N, Peacock B (2000) Statistical Distributions. 3rd edition. Wiley Series in Probability and Statistics John Wiley & Sons, Inc.
106. Seifert M, Strickert M, Schliep A, Grosse I (2011) Eploiting prior knowledge and gene distances in the analysis of tumor expression profiles by extended Hidden Markov Models. Bioinformatics 27: 1645–1652.
107. Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J Royal Stat Soc B 39: 1–38.
108. Gottman JM (1981) Time-Series Analysis Cambridge University Press.
109. Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, et al. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 31: 224–228.
110. Feschotte C, Pritham EJ (2007) DNA Transposons and the Evolution of Eukaryotic Genomes. Annu Rev Genet 41: 331–368.
111. Kazazian HH (2004) Mobile elements: Drivers of genome evolution. Science 303: 1626–1632.
112. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32: 5539–5545.
113. Jasinski M, Ducos E, Martinoia E, Boutry M (2003) The ATP-Binding Cassete Transporters: Structure, Function, and Gene Family Comparison between Rice and Arabidopsis. Plant Physiol 131: 1169–1177.
114. de Bolt S (2010) Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. Genome Biol Evol 2: 441–453.
115. Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in Arabidopsis thaliana. Proc Natl Acad Sci U S A 97: 7376–7381.
116. Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetic meets genomics. Nat Rev Genet 3: 329–341.
117. Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 21: i274–i282.
118. Ji H, Wong WH (2005) TileMap: create chromosomal map of tiling array hybridizations. Bioinformatics 21: 3629–3636.
119. Humburg P, Bulger D, Stone G (2008) Parameter estimation for robust HMM analysis of ChIPchip data. BMC Bioinformatics 9: 343.
120. Seifert M, Keilwagen J, Strickert M, Grosse I (2009) Utilizing gene pair orientations for HMMbased analysis of ChIP-chip data. Bioinformatics 25: 2118–2125.
121. Simpson JT, McIntyre RE, Adams DJ, Durbin R (2010) Copy number variant detection in inbred strains from short read sequence data. Bioinformatics 26: 565–567.
122. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, et al. (2010) CNAseg - a novel framework for identification of copy number changes in cancer from second-generation sequenencing data. Bioinformatics 26: 3051–3058.
123. Song Q, Smith AD (2011) Identifiying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics 27: 870–871.
124. Shen Y, Gu Y, Pe'er I (2011) A Hidden Markov Model for Copy Number Variant prediction from whole genome resequencing data. BMC Bioinformatics 12(Suppl 6): S4.
125. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817–825.