Hidden Markov Models for Genomic Segmentation and Annotation

INAUGURAL-DISSERTATION

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Rafael Campos Martín

aus Ciudad Real, Spanien

Köln 2021

Berichterstatter: Prof. Dr. Achim Tresch
Prof. Dr. Andreas Beyer
Tag der letzten mündlichen Prüfung: 03.April.2019

Acknowledgments

Many people have been by my side along this path that started four years ago. Although it would take a whole thesis to acknowledge all of them one by one, I would like to make an emphasis to those persons that have had a major impact:

First and foremost, I want to thank Achim Tresch for trusting me in the first instance and his patience during this time. He has been an excellent mentor in the lab and a friend outside.

I am really thankful to the members of my thesis advisory committee (TAC) Andreas Beyer and Michael Nothnagel as well as Stanislav Kopriva as part of my thesis defense committee.

This work could have not been completed without the contribution of our collaborations. I would like to thank Patrick Cramer for his ChIP and RNA-seq data in the yeast genome. Especially, I am really grateful to Michael Lidschreiber, whose help, dedication and kindness have pushed this project forward. Also, many thanks to Dirk Isbrandt and his group for the single-cell data and such fruitful collaboration. Even though the collaboration with Korbinian Schneeberger came on a really last minute, I am really happy for it. It has been a really interesting project, which has brought fresh ideas and horizons.

Special gratitude goes out to all the members in the Tresch group. They have provided me with an excellent work environment. I would like to thank especially Till Baar. He had the most difficult tasks of all, to share the office with me. Not only has he been able to put up with me all this time but also he has been of great help to deal with all the German bureaucracy. Also, I would like to thank Katharina Moss and Zahra Sadat Nasrollah for all the laughs and good moments. Sometimes it can be taxing to deal daily with me, but you have always been there with a smile.

I send my gratitude to all my flatmates in PS08. They have created a great atmosphere and coming home was always a place to feel welcomed. I would like to remark Gianna Noeckel, for being there since the beginning, to stand on my side when I felt the most lonely. Thanks to Jana Fisher, for carrying always a smile and the long weekends working together in our thesis, we made it! To Claudia, whose coolness and good mood have been always a plus. And thanks to Pablo, it has been an honor running by your side and discussing with you so many scientific topics with a glass of wine.

Rodrigo "Toti", thanks a lot for standing on my side so many times, for your good humor and your way of enjoying life. I have learned so many things from you that I cannot even mention them one by one. I owe so much to Daniel for all these years sharing passionate conversations about science at 5 am. Your company has been extremely joyful. All these years would have been much more different without the friendship of David and Raul. Thanks a lot for proving that it does not matter how far or how long takes us to meet again, things remain always as we left them before. I would like to thank Carmen all the nice and relaxing

times during these years. I would like to give my more sincere thanks to Henrik “Enrique”. Everyone should have a “Doktorbruder” like you. Thanks for showing me around the Ph.D. student life, your patience and support.

I always bring on my heart my siblings. Thank you, Samantha, for always carrying me with you, for teaching me even before I was able to walk. My love for nature and science comes from you. Moreover, you have brought into my life two amazing persons, Jose and Gonzalo. Thanks Jose for making my sister happy day by day, only for that, you have already won my respect. But on top, you have always welcomed me with open arms and love, and I am really thankful. To my little brother Antonio, whom I thought I was taking care of but has taught me more than I ever could have imagined. Having you near has been a gift all these years.

Por último, pero no menos importante. Mi más eterna gratitud a mi madre, Manuela Martín Prieto. Gracias por luchar por nosotros incluso a "navaja" si la situación lo ha requerido. Gracias por no dejar que nadie decida si puedes comprarnos unas zapatillas o no. Por enseñarnos que la familia unida jamás será vencida y que en la mesa y en las discusiones se conocen a los señores. Son lecciones que siempre llevaré conmigo. Además, gracias por retarme siempre, a amar las matemáticas, a ser mejor, a dar lo máximo y no dejar de luchar. Esta tesis te la dedico a ti.

"The scientist has a lot of experience with ignorance and doubt and uncertainty, and this experience is of very great importance, I think. When a scientist does not know the answer to a problem, he is ignorant. When he has a hunch as to what the result is, he is uncertain. And when he is pretty darn sure of what the result is going to be, he is in some doubt. We have found it of paramount importance that in order to progress we must recognize the ignorance and leave room for doubt. Scientific knowledge is a body of statements of varying degrees of certainty some most unsure, some nearly sure, none absolutely certain."

Richard Feynman

Summary

There are many processes in the genome and epigenome level that still remain elusive. Recent developments in high-throughput in sequencing have increased the amount of data exponentially. In order to analyze and obtain meaningful information and find natural patterns or clusters within these data sets, many bioinformatics laboratories are developing new algorithms and pipelines to face these challenges. On the genome, genomic and epigenomic data points within the same loci are generally more similar than distant data. Nevertheless, regions with similar functions that are scattered through the genome will as well produce similar data points. In order to model this linear dependency in the locus and find similar clusters throughout the genomic position, Hidden Markov models (HMMs) have been widely used.

In this thesis, we will introduce the theory of HMMs and the extended case of bidirectional HMMs (bdHMMs). In the genome context, many processes take place in a specific direction, e.g. DNA repair or DNA transcription. bdHMMs were developed to model processes that have some intrinsic directionality by defining conjugate, or twin, states.

In addition, we show a method by which any HMM model can be transformed into a clustering model and vice versa. Thus, the learning algorithm for HMM can be used to learn and fit the parameters for a clustering method. Moreover, the same procedure can be used for bdHMM and what we have named as bidirectional clustering.

In the second chapter, we have applied the bdHMM algorithm to study the tri-methylation status of histone three (H3) in three different lysines of its tail (H3K4me3, H3K36me3 and H3K79me3) together with their putative methyltransferase proteins (Set1, Set2 and Dot1 respectively) and new possible candidates that might use these modifications as a signaling mark to carry out their function (Asr1, Ioc4, Nto1, Pdp3, Rad9 and, Set4). Transcriptomics data was used to evaluate more closely the relationship of the marks and the gene expression in a metagene analysis.

Finally, we have worked out a new class of HMMs in which an extra hidden layer is added to infer haplotypes in populations of recombinant parents using low coverage sequencing. This extra layer models the high variability in SNP detection and provides a probability of how good is the specific marker based on the information provided by all the samples being analyzed. We use this method to study the effect of three proteins (RECQ4A, RECQ4B, and FIGL1) with known roles in resolving recombinant events during meiosis I.

Taken together, we have extended the theory of Hidden Markov models. First by introducing bidirectionality and second by a transformation of HMMs into clustering models. Using these new models in real data, we investigated actual questions that molecular biology is facing.

Zusammenfassung

Viele Prozesse auf der Genom- und Epigenomebene sind immer noch nicht vollständig erforscht. Die jüngsten Entwicklungen im Bereich der Hochdurchsatz-Sequenzierung haben zu einer exponentiellen Vergrößerung der Datenmenge geführt. Um aussagekräftige Informationen aus diesen Daten zu erhalten und natürlich vorkommende Muster oder Cluster innerhalb der Datensätze zu finden, entwickeln viele Bioinformatiklabore neue Algorithmen und Pipelines. In genomischen und epigenomischen Daten sind Datenpunkte innerhalb derselben Loci ähnlicher als Datenpunkte, die eine große Distanz auf dem Genom aufweisen. Allerdings führen Regionen mit ähnlichen Funktionen, die über das Genom verstreut sind, ebenfalls zu ähnlichen Datenpunkten. Um diese lineare Abhängigkeit im Locus zu modellieren und ähnliche Cluster im Genom zu finden, werden häufig Hidden Markov-Modelle (HMMs) verwendet.

Thema der vorliegenden Doktorarbeit sind die theoretischen Grundlagen von HMMs und die Erweiterung zu bidirektionalen HMMs (bdHMMs). Im Genom kann bei einer Vielzahl von Prozessen wie DNA- Reparatur oder Transkription eine definierte Richtung beobachtet werden.

bdHMMs wurden entwickelt, um Prozesse mit intrinsischer Direktionalität zu modellieren. Dies geschieht durch die Definition von Konjugaten, oder Twin, Zuständen.

Darüber hinaus wurde eine Methode entwickelt, mit der jedes HMM in ein Clustering-Modell umgewandelt werden kann und umgekehrt. Mit dem Lernalgorithmus für HMMs können so die Parameter für eine Clustering-Methode erlernt und angepasst werden. Das gleiche Verfahren kann für bdHMMs und das Verfahren des bidirektionalen Clusterings verwendet werden.

Im zweiten Kapitel der Arbeit findet der bdHMM-Algorithmus seine Anwendung: Der Tri-Methylierungsstatus von Histon drei (H3) wird in drei verschiedenen Lysinen seines Poly(A)-Schwanzes (H3K4me3, H3K36me3 und H3K79me3) zusammen mit seinen mutmaßlichen Methyltransferase-Proteinen (Set1, Set2 und Dot1) und weiteren möglichen Genen untersucht, die diese Modifikationen als Signalzeichen verwenden könnten (Asr1, Ioc4, Nto1, Pdp3, Rad9 und, Set4). Mithilfe einer Metagenanalyze von Transkriptomikdaten konnte die Beziehung zwischen den Marken und der Genexpression genauer bewertet werden.

Als weiterer Schritt wurde eine neue Klasse von HMMs mit einer zusätzlichen versteckten Schicht entwickelt, um Haplotypen in Populationen rekombinanter Eltern mittels Low-Coverage-Sequenzierung zu ermitteln. Die zusätzliche Schicht modelliert die hohe Variabilität in der SNP-Erkennung und ermöglicht Aussagen über die Zuverlässigkeit des spezifischen Marker basierend auf den Informationen aller analysierten Proben. Die verwendete Methode untersucht die Wirkung von drei verschiedenen Proteinen (RECQ4A, RECQ4B und FIGL1), von denen bekannt ist, dass sie eine Rolle bei der Lösung rekombinanter Ereignisse während der Meiose I spielen.

Insgesamt konnte so die Theorie über HMM in zweifacher Hinsicht erweitert werden: Erstens durch die Einführung von Bidirektionalität und zweitens durch die Überführung von HMMs in Clustering-Modelle. Die entwickelten Modelle wurden auf reale Daten angewendet, um aktuelle Fragen der Molekularbiologie zu untersuchen.

Contents

List of Figures

List of Tables

Part I
(Bidirectional) Hidden Markov Models and (Bidirectional) Clustering Method

1 Introduction

The development of the new omic technologies is helping molecular biologists to unravel the complexity of the genome [1]. Nevertheless, these technologies are futile without appropriate unsupervised algorithms, which are able to detect patterns within a large, high-dimensional dataset. Historically, different clustering methods have been applied to define distinct clusters, e.g. in cancer research [2, 3], gene expression analysis [4], prediction of gene function [5], etc. An important characteristic of these methods is that they explicitly or implicitly rely on distributional assumptions, and the soundness of these assumptions will affect the quality of the outcome.

At the genome level, most observations have a spatial dependency, e.g. adjacent loci tend to remain in the same cluster [6, 7]. Hidden Models (HMM) have been widely used for the clustering of genomic data since HMMs model dependency of consecutive observations. HMMs have proved to be extremely useful in various fields in Biology, such as gene prediction [8], protein structure [9], and many others [8, 10].

Consortia like ENCODE [11], pENCODE [12], ROADMAP epigenomics [13], etc. are gathering huge amounts of genomic and epigenomic data. In order to cope with this load of data, algorithms such as Segway [14] or EpiCSeg [15], among others [16], have been developed. Nevertheless, most of these algorithms do not model strand-specific (RNA-Seq) data [17], nor do they take into account the intrinsic directionality of many biochemical DNA-associated processes. During transcription, e.g., polymerases synthesize RNA using the DNA template from the 3' to 5' end [6, 7]. To this end, HMMs have been extended to the double-stranded HMM (dsHMM) [6], and the bidirectional HMM (bdHMM) [7]. While the former models the forward and reverse DNA strands using two different Markov chains in opposite direction, the latter assigns to each state a directionality flag, indicating whether a state operates in forward (respectively reverse) direction, or is undirected.

In this chapter, we introduce the STAN software, an R/Biocondcutor [7] package that implements algorithms for mixture clustering, direction-aware clustering (bdClustering), HMM and bdHMM learning. First, the HMM theory and the algorithms to update the parameters will be described. Second, it will be intro-

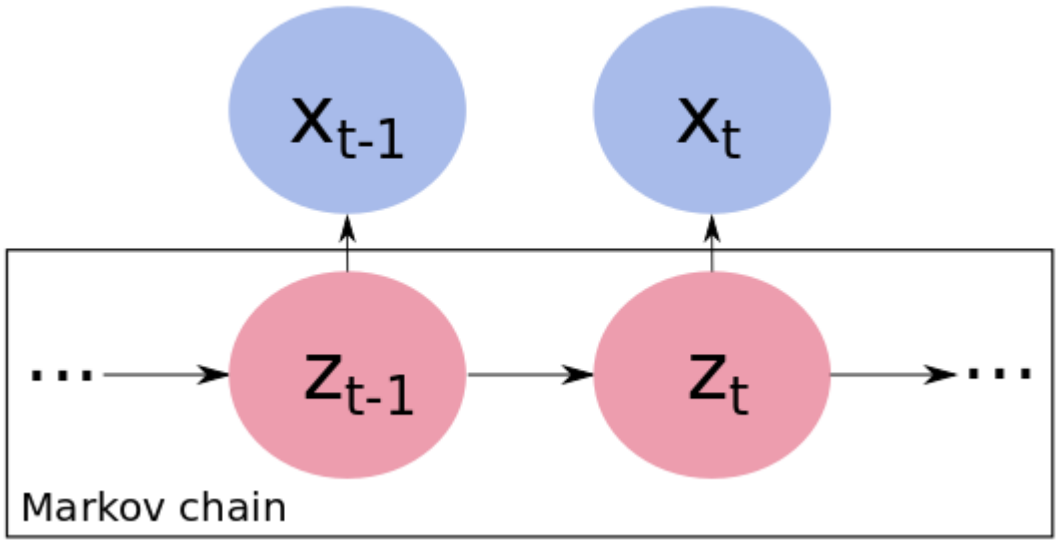


Figure 1: **Graphical structure HMMs**. The latent variable Z is described by a Markov chain. Therefore, the state of the variable z_t depends exclusively on the state of the variable z_{t-1}. Moreover, the observation x_t is conditioned in the value of z_t.

duced the bdHMMs and their semantic with some biological examples. Finally, based on the HMM theory it will be shown how to derive the likelihood function of clustering. Furthermore, it will be shown that a similar approach can be taken to obtain a direction-aware clustering method similar to the bdHMMs.

2 Hidden Markov Models

2.1 Model Statement

Hidden Markov Models (HMM) are a powerful tool to analyze data points which are not independent [18]. In other words, if the observation at a certain genomic position depends on an observation at the previous position, HMM can capture and model this dependency. Time series experiments, e.g. currency exchange rate, speech recognition or online handwriting recognition generate data sets which the time point of an observation is informative about future observations. Biological sequences such proteins or DNA sequences also generate data with this kind of dependency structure [8, 9]. For convenience, I will use the terms "past" and "future" observations in a sequence synonymous to "previous" and "subsequent" observations, respectively.

HMMs are composed by a latent variable which is a discrete and correspond to a (time-independent) Markov chain, here the name of hidden. It assumes that ever observation x_t is *emitted* by the corresponent hidden variable z_t1. Therefore,

the value of z_t defines the probability of observing x_t by $P(x_t|z_t) = \psi_{z_t}(x_t)$. From 1, it is clear thet the conditional proability distribution of the hidden variable z_t is dependent only on the value of z_{n-1}, also known as markov property.

More formally, an HMM is deifned by a tuple $\theta = (\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ such that:

1. $\mathcal{K}$ is a finite set, which elements are called states.
2. The initial probability vector $\pi = \pi_{i \in \mathcal{K}}$ is a row vector with $0 \leq \pi_i \leq 1$, $i \in \mathcal{K}$, and $\sum_{i \in \mathcal{K}} \pi_i = 1$.
3. The matrix A is a stochastic matrix of dimension $\mathcal{K} \times \mathcal{K}$ and defines the transition probabilities a_{ij}. The row vectors are probability vectors with $0 \leq a_{ij} \leq 1,\ i \in \mathcal{K}$, and $\sum_{j \in \mathcal{K}} a_{ij} = 1$.
4. The *emission distribution* $\Psi = \{\psi_i, i \in \mathcal{K}\}$ form a set of probability distributions on a space $\mathcal{D}$,the *observation space.*

Given a sequence of observations $X = (x_{0,}, ..., x_T)$ generated by a sequence of hidden observations $Z = (z_0, ..., z_T)$ from an HMM, we can compute the likelihood:

$$\begin{aligned} P(X, Z; \theta) &= P(X|Z; \theta) \cdot P(S; \theta) \\ &= \prod_{t=0}^{T} P(x_t|z_t; \Psi) \cdot \prod_{t=1}^{T} P(z_t|z_{t-1}; A) \cdot P(z_0; \pi) \\ &= \prod_{t=0}^{T} \psi_{z_t}(x_t) \cdot \prod_{t=1}^{T} a_{z_{t-1}, z_t} \cdot \pi_{z_0} \end{aligned} \tag{1}$$

2.2 Parameter Inference

To estimate the model parameters θ, one must maximize the marginal likelihood $P(O; \theta)$. This marginal probability can be obtained by summing the joint probability (1) over all possible state sequences:

$$P(O; \theta) = \sum_{all\ Z} P(X|Z; \theta) \cdot P(S; \theta) \tag{2}$$

This calculation increase in complexity as the time points T and number of states $\mathcal{K}$ increases. The number of calculations needed to compute the marginal likelihood augment in order $2T \cdot \mathcal{K}^T$. If one considers that the smallest known genome is 220 base pairs (bp), the calculation of the marginal likelihood is already unfeasible for any modern computer. The solution to this problem was given in the Baum-Welch algorithm, an instance of the expectation-maximization (EM) algorithm. The EM algorithm is an iterative method to estimate the parameters

of models that contain latent variables. It iterates over 2 steps: the Expectations (E) and the Maximization (M) step. The underlying idea of this algorithm is that instead of working with the marginal likelihood (2) a new target function $\mathcal{Q}(\theta, \theta^{old})$ is maximized with respect to the parameters θ, given a previous parameter guess θ^{old}. The algorithm will converge to a local maximum of the marginal likelihood $P(X;\theta)$. Therefore, the selection of the initial parameters is crucial to obtain good results.

The target function, denoted as $\mathcal{Q}(\theta, \theta^{old})$, is given by:

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{all\ Z} P(Z|X, \theta^{old}) \cdot logP(X, Z; \theta) \tag{3}$$

Nevertheless, the summation over all possible sequence of states Z needs to be computed. An efficient procedure to solve this problem in HMMs is the forward-backward algorithm, developed in the early seventies by Leonard E. Baum and Lloyd R. Welch [18, 19]. This method introduces two new probabilities:

$$\alpha_t(k) = P(z_t = k, x_0, ..., x_t; \theta^{old}) \tag{4}$$

$$\beta_t(k) = P(z_t = k|x_{t+1}, ..., x_T, z = k; \theta^{old}) \tag{5}$$

The $\alpha_t(k)$ is the joint probability of observing all data points up to time t and the value of z_t is k, whereas the conditional probability $\beta_t(k)$ defines the conditional probability of the data points from $t+1$ to T given the state z_t is equal to k. More precisely, we can compute this probabilities inductively, as follows:

1. Forward/backward initiation:

$$\alpha_1(k) = \pi_k^{old} \psi_k^{old}(x_1) \tag{6}$$

$$\beta_T(k) = 1 \tag{7}$$

2. Forward/backward induction:

$$\alpha_t(k) = \psi_k^{old}(x_t) \sum_{i \in \mathcal{K}} a_{ij}^{old} \alpha_{t-1}(i) \tag{8}$$

$$\beta_t(k) = \psi_k^{old}(x_t) \sum_{i \in \mathcal{K}} a_{ji}^{old} \beta_{t+1}(i) \tag{9}$$

We can now use these terms to expresses two new posterior probabilities. The posterior probability $\gamma_t(k)$, defined as the conditional probability of being in state k at time t, given the observation sequence X. And the value $\xi(k,l)$, defined as the probability of being in state k at time t and in state l at time $t+1$, given the full observation X:

$$\begin{aligned} \gamma_t(k) &= P(z_t = k|X, \theta^{old}) \\ &= \frac{\alpha_t(k)\beta_t(k)}{\sum_{i\in\mathcal{K}} \alpha_t(i)\beta_t(i)} \end{aligned} \tag{10}$$

$$\begin{aligned} \xi(j,k) &= P(z_t = k, z_{t+1} = l|X, \theta^{old}) \\ &= \frac{\alpha_t(k)a_{kl}^{old}\beta_{t+1}(l)\psi_l^{old}(o_{t+1})}{\sum_{i\in\mathcal{K}} \alpha_t(i)\beta_t(i)} \end{aligned} \tag{11}$$

The target function $\mathcal{Q}(\theta, \theta^{old})$ can now be written using equation (10) and (11):

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{k,l}\sum_{t=1}^{T} \xi(k,l) log\, a_{kl} + \sum_{k} \gamma_0(k) log\, \pi_k + \sum_{k}\sum_{t=1}^{T} \gamma_t(k) log\, \psi_k(x_t) \tag{12}$$

The calculation of $\mathcal{Q}(\theta, \theta^{old})$ in terms of (12) is now computationally feasible, its time complexity is $T \cdot |\mathcal{K}|$. To perfrom the M-step of the algorithm, i.e., to optimize Q with respect to θ, we can differentiate and use appropiate Lagrange multipliers to satisfy the constrains $\sum_{i\in\mathcal{K}} \pi_i = 1$ and $\sum_{j\in\mathcal{K}} a_{ij} = 1$. Setting the derivate to zero and solving for each parameter we obtain:

$$\pi_i = \gamma_0(i) \tag{13}$$

and,

$$a_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{14}$$

The update of the Ψ parameters is specific for each distribution family belonging to the observations. For an update of the parameters of the used emission distributions (Bernoulli, Negative Multinomial, (multivariate) Gaussian, Negative Binomial, Poisson and Poisson Log-normal in our case), refer to *Zacher et. al.* [7]. Furthermore, we worked out the update formulas for the Beta-Binomial distribution family in Chapter 4 of this thesis.

2.3 Viterbi Algorithm

Given an HMM model θ and a sequence of observations X, the most common problem is to find the most probable sequence of states Z that explain the observations X. In other words, we want to find the sequence of states Z such $P(Z, X|\theta)$ is maximized. The solution to this problem is given by a recursive method called the Viterbi algorithm:

$$\begin{aligned} \delta_0(i) &= \pi_i \psi_i(o_o) \\ \delta_t(i) &= \max_{z_0, z_1, ..., z_{t-1}} P(z_0, z_1, ..., z_t = i|\theta) \\ \delta_{t+1}(j) &= \max_i \delta_t(i) a_{ij} \psi_j(o_{t+1}) \end{aligned}$$

The optimal hidden state path, or Viterbi path can be obtained applying backtracking [19].

3 Bidirectional Hidden Markov Models (bdHMM)

The bdHMM theory was developed by Achim Tresch and Benedikt Zacher. In this section, I recover the theory explained in *Zacher et al.* [7] to introduce the bdHMMs and then explain how to obtain the bidirectional clustering algorithm.

A bdHMM is a special instance of HMM that satisfies three additional conditions. Two conditions define the structure of the hidden variable Z, and the other deals with the observations. These three conditions are not defined *ad hoc* as it will be demonstrated in the semantic of bdHMMs.

Definition.

A bdHMM is defined as a tuple $\theta = ((\mathcal{K}, i_\mathcal{K}), \pi, A, (\mathcal{D}, i_\mathcal{D}), \Psi)$. The parameter tuple is composed by the parameters of an HMM $\theta_{HMM} = (\mathcal{K}, \pi, A, \Psi)$, and the involutions $i_\mathcal{K} : \mathcal{K} \to \mathcal{K},\ k \mapsto \bar{k}$ and $i_\mathcal{D} : \mathcal{D} \to \mathcal{D},\ o \mapsto \bar{o}$; (a map i is called an involution if $i^2 = id$). Morevoer, the following symmetry conditions hold:

1. *Generalized detailed balance:* The first symmetry condition satisfies condtions in the initial probability and the transition probability

$$\pi_i a_{ij} = \pi_{\bar{j}} a_{\bar{j}\bar{i}}\ ,\ i, j \in \mathcal{K} \tag{15}$$

2. *initiation symmetry:* The initial probability π satisfies

$$\pi_i = \pi_{\bar{i}}\ ,\ i \in \mathcal{K} \tag{16}$$

3. *Observation symmetry:* The emission distribution Ψ satisfies

$$\psi_i(o) = \psi_{\bar{i}}(\bar{o})\ ,\ i \in \mathcal{K},\ o \in \mathcal{D} \tag{17}$$

3.1 The Semantic of bdHMMs

In the following subsection we will motivated the selection of the above mentioned conditions: (15), (16) and (17), let $\theta = ((\mathcal{K}, i_\mathcal{K}), \pi, A, (\mathcal{D}, i_\mathcal{D}), \Psi)$ be a bdHMM. By (15) and (16)

$$(\pi A)_j \quad = \quad \sum_{i \in \mathcal{K}} \pi_i a_{ij} = \sum_{i \in \mathcal{K}} \pi_{\bar{j}} a_{\bar{j}\bar{i}} = \pi_{\bar{j}} = \pi_j \ , \ j \in \mathcal{K},$$

which proves $\pi A = \pi$, so π is a (left) eigenvector of A. If π is an eigenvector of the stochastic matrix A, then the initial probability π of a bdHMM is always a stationary state distribution of A. Consequently, the previously mentioned assumptions imply that

$$\begin{aligned} P(z_{t-1} = i, z_t = j) &= P(z_{t-1} = i) \cdot P(z_t = j | z_{t-1} = i) = \pi_i a_{ij} \\ &= \pi_{\bar{j}} a_{\bar{j}\bar{i}} = P(z_{t-1} = \bar{j}) \cdot P(z_t = \bar{i} | z_{t-1} = \bar{j}) \\ &= P(z_{t-1} = \bar{j}, z_t = \bar{i}) \end{aligned} \tag{18}$$

holds for all $i, j \in \mathcal{K}$ and $t = 1, ..., T$. The consequence of this is that at any location in the sequence, the probability of observing the state i preceding state j is equal to the probability of observing the respective conjugates in reversed order. Hence, equation (18) defines the *Generalized detailed balance* and the *Initiation symmetry*. The reason to chose two different conditions (condition (1) and (2)) over a simpler one (equation 18) is exclusively because conditions (1) and (2) are explained based on model parameters, whereas equation (18) has a more abstract definition.

To understand the involution $i_\mathcal{K}$ one must understand that a bdHMM models hidden processes with a directionality. These processes can occur in forward or reverse direction, but as well undirected processes can be expected. The involution $i_\mathcal{K}$ splits the sets of states $\mathcal{K}$ in directed states, defined as pairs $(k, \bar{k})$, $k \neq \bar{k}$ of conjugate states (or twin states). One of the conjugate, or twin, state define processes in forward and the other in reverse direction. The appointment of which state $(k, \bar{k})$ models forward or reverse is assigned in a post-processing step. Those states that model forward processes are called forward states $\mathcal{K}^+$ and the ones modeling reverse processes are called reverse states $\mathcal{K}^-$. Moreover, there is a third set of states that describe undirected processes named undirected states $\mathcal{K}^0$, and the conjugate states are equal $k = \bar{k}$.

The third condition is related to the observations. The bdHMMs can model observations generated by forward, reverse, and undirected underlying processes. The observations may explain the directionality of such process that produced it. Therefore, the involution $i_\mathcal{D}$ maps the observation $o \in \mathcal{D}$ to its conjugate observation $\bar{o}$, which denotes the corresponding observation if the directionality of the

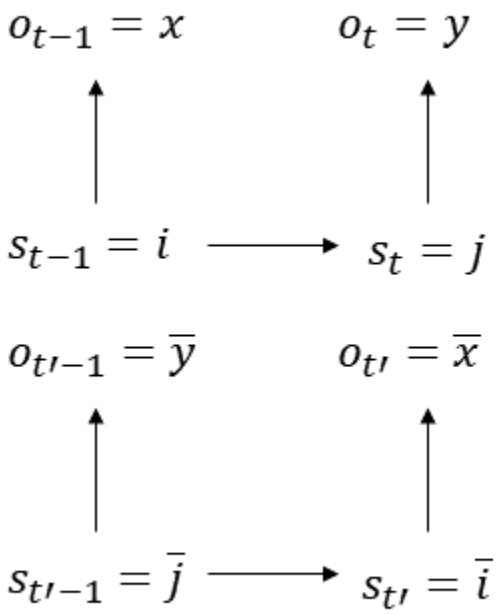


Figure 2: **Defining Property of bdHMMs.** The marginal likelihood $P(o_{t-1}, o_t, s_{t-1}, s_t; \theta)$ for the hidden and observable variables in two consecutive positions in the sequence is invariant under simultaneous reversal of state directionality and direction information in the observations. In simpler terms, the model should be invariant under reversal of the annotation of the two DNA strand as "forward" respectively "reverse" strand. This means that $P(o_{t-1} = x, o_t = x, s_{t-1} = i, s_t = j; \theta) = P(o_{t-1} = \bar{y}, o_t = \bar{x}, s_{t-1} = \bar{j}, s_t = \bar{x}; \theta)$ for any admissible values of x, y, i, j, t, t'.

process was reversed. A clear example of such observations is motivated by the advent of strand-specific sequencing methods in genomics. In the case that one wanted to include strand-specific data to ChIP-seq experiments, $\mathcal{D}$ is modeled as the Cartesian product of $\mathcal{D}^0$, for the ChIP measurements of protein binding, a space $\mathcal{D}^+$ generated by the transcription of genes in the forward or plus strand and captured by the strand-specific RNA-seq protocol and the corresponding observations by the reverse or minus strand $\mathcal{D}^-$. Therefore, the involution $i_{\mathcal{D}}$ acts on the observations, $i_{\mathcal{D}} : o = (o^0, o^+, o^-) \mapsto \bar{o} = (o^0, o^-, o^+)$. The observation symmetry is ligated to the involution $i_{\mathcal{K}}$, such that twin states have the same probability distribution, up to involution $i_{\mathcal{D}}$ of the observations (figure 2).

Note that if there is no underlying directed process, meaning that the involution $i_{\mathcal{K}}$ is the identity map, the condition (16) is void and the detailed balance

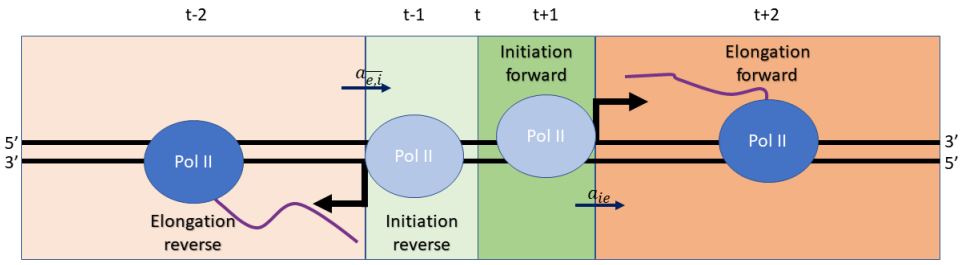


Figure 3: **Bidirectiona promoter example.** Bidirectional promoters have been widely described in literature. Eukaryotic cells use these functional region for initiation of transcription of one gene in the Watson and another in the Crick strand. From a sequential point of view $(t-2, ..., t+2)$, if taken the Watson strand as reference, the transcription of those two genes occur in opposite directions. Still, the underlying process that explain the observations are the same. Therefore, the transition from the elongation state of Pol II to initiation stage in the gene transcribed in the Crick strand, also reverse strand $(a_{\bar{e},\bar{i}})$ is equal to the transition from initiation to elongation in the Watson, or forward strand $(a_{i,e})$. Hence, the elongation state in reverse $(\bar{e})$ is the conjugate state of the elongation state in forward (e), and so are the initiation state in forward and reverse $(i, \bar{i})$.

condition reduces to a reversible HMM. Additionally, if there is no information about the directionality on the observations, such that the involution $i_{\mathcal{D}}$ is the identity map, the observation symmetry is void as well. Thus, our bdHMM is a reversible HMM which satisfies the detailed balance relation $\pi_i a_{ij} = \pi_j a_{ji},\ i, j \in \mathcal{K}$. Consequently, our algorithm can model reversible HMMs observations.

Example: in many eukaryotic cells, some promoter regions can start transcription in both strands, e.g. Watson and Crick strands. This family of promoters are known in the literature as bidirectional promoters. After Pol II is bound to the promoter region (initiation), it will start to transcribe the template DNA into RNA (elongation). Depending in which direction this process takes place, the Watson or Crick strand will be used as the template for transcription. For an external observer, these processes will take place in opposite directions. One taking place in forward (Watson) and the other in reverse (Crick) direction. Nevertheless, (bd)HMM algorithm analyzes every genome position as time points using as reference the Watson strand. This means that the transition from an elongation stage to initiation stage in the Crick strand will have the same probability as the transition from initiation to an elongation stage in the Watson 3. Therefore, the elongation state in reverse is the conjugate of the elongation in forward and the same for the initiation state.

Given an observation $X = (x)_{t=0,...,t=T}$, we can define the reversed observations

by applying the involution $i_{\mathcal{D}}$ and reverse the order, $X^{rev} = (x_t^{rev} = \bar{x}_{T-t})_{t=0,\ldots,t=T}$. The same can be applied to the hidden variable $Z = (z_t)_{t=0,\ldots,T}$ to obtain the reversed $Z^{rev} = (z_t^{rev} = \bar{z}_{T-t})_{t=0,\ldots,t=T}$. We verify that bdHMM are symmetric with respect to the "reversal" of the observation sequence:

$$\begin{aligned} P(Z;\theta) &= \pi_{z_o} \prod_{t=1}^{T} a_{z_{t-1}z_t} \overset{(15)}{=} \pi_{z_T} \prod_{t=1}^{T} a_{\bar{z}_t \bar{z}_{t-1}} \\ &\overset{(15,16)}{=} \pi_{\bar{z}_T} \prod_{t=1}^{T} a_{\bar{z}_{T-(t-1)} z_{T-t}^{-}} = \pi_{z_0^{rev}} \prod_{t=1}^{T} a_{z_{t-1}^{rev} z_t^{rev}} \\ &= P(Z^{rev};\theta) \end{aligned} \tag{19}$$

Moreover,

$$\begin{aligned} P(X|Z;\theta) &= \prod_{t=0}^{T} \psi_{z_t}(x_t) \overset{(17)}{=} \prod_{t=0}^{T} \psi_{\bar{z}_t}(\bar{x}_t) \\ &= \prod_{t=0}^{T} \psi_{\bar{z}_{T-t}}(\bar{x}_{T-t}) = \prod_{t=0}^{T} \psi_{z_t^{rev}}(o_t^{rev}) \\ &= P(X^{rev}|Z^{rev};\theta) \end{aligned} \tag{20}$$

The equalitites (19) and (20) imply

$$\begin{aligned} P(X,Z;\theta) &= P(X|Z;\theta) \cdot P(Z;\theta) \\ &= P(X^{rev}|Z^{rev};\theta) \cdot P(Z^{rev};\theta) \\ &= P(X^{rev}, Z^{rev};\theta) \end{aligned} \tag{21}$$

also

$$P(Z|X;\theta) = P(Z^{rev}|X^{rev};\theta) \tag{22}$$

Therefore, a bdHMM is reversible in the generalized sense

$$P(O;\theta) = P(O^{rev};\theta)$$

3.2 Baum-Welch Algorithm for bdHMMs

Given a parameter set $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A, (\mathcal{D}, i_{\mathcal{D}}), \Psi)$ of a bdHMM, we can define a new target function such

$$\hat{\mathcal{Q}}(\theta, \theta^{old}) \;=\; \mathcal{Q}(\theta, \theta^{old}) + \mathcal{Q}(\theta, \bar{\theta}^{old}) \tag{23}$$

where $\bar{\theta}^{old}$ is the bdHMM with parameter set $\bar{\theta} = ((\mathcal{K}, i_{\mathcal{K}}), \bar{\pi}, \bar{A}, (\mathcal{D}, i_{\mathcal{D}}), \bar{\Psi})$. The parameters are defined based on the bdHMM parameter as $\bar{\pi}_i = \pi_{\bar{i}},\ \bar{a}_{ij} = a_{\bar{i}\bar{j}},\ \bar{\psi}_i(x) = \psi_{\bar{i}}(x),\ i, j \in \mathcal{K},\ o \in \mathcal{D}$.

The full derivation of the function $\hat{\mathcal{Q}}(\theta, \theta^{old})$ and the updated formulas for the emission distributions can be found in *Zacher et al.* Here, we will only introduce the update formulas for the parameters a_{ij} and π_i:

$$a_{ij} \;=\; \frac{\sum_{t=1}^{T}(\xi_t(i,j) + \xi(\bar{j},\bar{i}))}{\sum_{t=1}^{T}(\gamma_{t-1}(i) + \gamma_t(\bar{i}))}\ ,\ i, j \in \mathcal{K} \tag{24}$$

$$\pi_i \;=\; \frac{1}{2T}\sum_{t=1}^{T}(\gamma_{t-1}(i) + \gamma_t(\bar{i}))\ ,\ i \in \mathcal{K} \tag{25}$$

Although this method has not been proved to converge, we have not found any case in which that was not the case in practice. Moreover, the algorithm is significantly faster than other numerical approaches.

3.3 Directionality Score

The selection of the number of states is one of the major concerns in HMMs. Some solutions to this problem have been proposed as using Bayesian Information Criterion (BIC), Akaike information criterion (AIC) or minimum description length (MLD). These methods try to balance the number of states and the precision of the data fit.

Since our bdHMM have two different sets of states, directed and undirected, the solutions mentioned above might not be suitable. The goal is to find the right number of directed and undirected states. To this end, we have defined a directionality score, which will help us to find the most appropriate number of directed states.

The directionality score uses the posterior probability $\gamma_t(k)$ for the conjugate states $(k, \bar{k})$ over all the position to determine if, on average, one of there is no ambiguity among the two states. Formally, we can define our directionality score as

$$DirScore \;=\; \frac{\sum_{t=0}^{T}|\gamma_t(k) - \gamma_t(\bar{k})|}{\sum_{t=0}^{T}(\gamma_t(k) + \gamma_t(\bar{k}))} \tag{26}$$

This score can be interpreted as a probability measure as well since $0 \leq DirScore \leq 1$. After computing the directionality score for all the pairs of conjugate states a

threshold q must be selected and remove all states under this threshold q. There is no proper way to determine a statistical value for q but we recommend using $q = 0.5$. When a pair of twin states $(k, \bar{k})$ has a directionality score bigger than 0.5, means that more than half of the time on of the conjugate states is suited than the other. Thus, it is more likely that it belongs to a directed process in which the conjugate state is not convenient.

One might notice that the directionality score might be affected by the number of undirected states. Thus, as the method to find the correct number of directed and undirected states consists in compute the dirScore for different combinations of directed and undirected states and chose the model which all directed states have a dirScore higher than 0.5 and has the lower number of total states.

4 Bidirectional Clustering

We have introduced a solution to model observations with spatial or temporal dependence. But, when the data points are independent of each other and the order does not provide any information the mixture model is the most suitable one. The goal of mixture models is to find the clusters $\mathcal{K}$, or subpopulations, in a dataset. It is similar to the HMM in that a new latent variable C is defined and is responsible for each observation.

More formally, a mixture model is defined by a tuple $\theta = (\mathcal{K}, \phi, \mathcal{D}, \Psi)$such that:

1. $\mathcal{K}$ is a finite set, which elements are called clusters.
2. The mixture weight vector $\phi = \phi_{i \in \mathcal{K}}$ is a row vector with $0 \leq \phi_i \leq 1$, $i \in \mathcal{K}$, and $\sum_{i \in \mathcal{K}} \phi_i = 1$.
3. The emission distribution $\Psi = \{\psi_i, i \in \mathcal{K}\}$ form a set probability distribution on a space $\mathcal{D}$, the *observation space.*

Given a sequence of observations $X = (x_t)_{t=0,\ldots,t=T}$, note that here the subscript t does not have a temporal connotation. The observations x_t were drawn from the distribution ψ_{c_t} where $c_t \in \mathcal{K}$ is a the latent class or cluster variable. Let $C = (c_t)_{t=0,\ldots,t=T}$. The likelihood takes the form:

$$\begin{aligned} P(X, C; \theta) &= P(X|C; \theta) \cdot P(C; \theta) \\ &= \prod_{t=0}^{T} P(x_t|c_t; \Psi) \cdot \prod_{t=0}^{T} P(c_t; \phi) \\ &= \prod_{t=0}^{T} \psi_{c_t}(x_t) \cdot \prod_{t=0}^{T} \phi_i \end{aligned} \tag{27}$$

Thus, we verify that the likelihood of an HMM with parameter set $\theta_{HMM} = (\mathcal{K}, \phi, A, \mathcal{D}, \Psi)$ specializes to (27) if we set $a_{ij} = \phi_j,\ i, j \in \mathcal{K}$ in equation (1):

$$
\begin{aligned}
P(X, C; \theta_{HMM}) &= P(X|C; \theta_{HMM}) \cdot P(C; \theta_{HMM}) \\
&= \prod_{t=0}^{T} P(x_t|c_t; \Psi) \cdot \prod_{t=1}^{T} P(c_t|c_{t-1}; A) \cdot P(z_0; \phi) \\
&= \prod_{t=0}^{T} \psi_{c_t}(x_t) \cdot \prod_{t=1}^{T} a_{c_{t-1}, c_t} \cdot \phi_{z_0} \\
&= \prod_{t=0}^{T} \psi_{c_t}(x_t) \cdot \prod_{t=1}^{T} \phi_{c_t} \cdot \phi_{z_0} \\
&= P(X, C; \theta) \qquad (28)
\end{aligned}
$$

4.1 Update Formula for bdClustering

For the learning of ϕ, we will plug the addictional constraints into the equation (3) and add the Lagrange multipliers. This leads to an EM update algorithm for (bd)clustering:

$$
\begin{aligned}
\mathcal{Q}(\theta, \theta^{old}) &= \sum_{k,l} \sum_{t=1}^{T} \xi(k, l) log\ \phi_l + \sum_{k} \gamma_0(k) log\ \phi_k + \sum_{k} \sum_{t=0}^{T} \gamma_t(k) log\ \psi_k(x_t) \\
&\quad + \lambda(1 - \sum_{k} \phi_k) \\
&= \sum_{k} \sum_{t=1}^{T} \gamma_t(k) log\ \phi_l + \sum_{k} \sum_{t=0}^{T} \gamma_t(k) log\ \psi_k(x_t) \qquad (29) \\
&\quad + \lambda(1 - \sum_{k} \phi_k)
\end{aligned}
$$

Since the observations are independent, the computation of the responsabilities $\gamma_t(k)$ can be performed in parallel:

$$
\begin{aligned}
\gamma_t(k) &= P(c_t = k|X; \theta^{old}) = \frac{P(x_t|c_t = k; \theta^{old}) P(c_t = k; \theta^{old}}{P(x_t; \theta^{old})} \\
&= \frac{\phi_k \psi_k(x_t; \theta^{old})}{\sum_{j \in \mathcal{K}} \phi_j \psi_j(x_t; \theta^{old})},\ t = 0, ..., T \qquad (30)
\end{aligned}
$$

The partial derivates of $\mathcal{Q}(\theta, \theta^{old})$ with respect to ϕ_k are:

$$
\frac{\partial \mathcal{Q}(\theta, \theta^{old})}{\partial \phi_k} = \begin{cases} \sum_{t=0}^{T} \frac{\gamma_y(k)}{\phi_k} - \lambda & if\ k = \bar{k} \\ \sum_{t=0}^{T} \frac{\gamma_t(k) + \gamma_t(\bar{k})}{\phi_k} - 2\lambda & if\ k \neq \bar{k} \end{cases} \qquad (31)
$$

Setting the partial derivatives to zero and solving for λ in the usual fashion leads to

$$\lambda = \sum_{k} \sum_{t=0}^{T} \gamma_t(k)$$

and the updated formula results in

$$\phi_k = \begin{cases} \frac{\sum_{t=0}^{T} \gamma_t(k)}{\sum_{k} \sum_{t=0}^{T} \gamma_t(k)} & if\ k = \bar{k} \\ \frac{\sum_{t=0}^{T} \gamma_t(k) + \gamma_t(\bar{k})}{2 \cdot \sum_{k} \sum_{t=0}^{T} \gamma_t(k)} & if\ k \neq \bar{k} \end{cases} \tag{32}$$

The learning of the emission distributions in bidirectional clustering can be done exactly in the same way as for bdHMMs.

4.2 Semantics of bdClustering

We have defined the bdHMM by three main properties: (1) Generalized detailed balance, (2) initiation symmetry, and (3) the observation symmetry. But, do they hold for bdClustering? Since the hidden latent variables are not dependent, there is no underlying process that has directionality. Therefore, condition (1) and (2) are void in the case of bdClustering. condition(3) is a constraint of the bdClustering model: Let $i_{\mathcal{K}} : k \mapsto \bar{k}$ be an involution on $\mathcal{K}$ mapping twin clusters onto each other. Then,

$$\psi_k(o) = \psi_{\bar{k}}(\bar{o}) \tag{33}$$

Thus, bidirectional clustering will be only possible if the observations contain information about the underlying direction of the process.

5 bdHMM vs bdClustering Example

To illustrate the differences between bdHMM and bdClustering, we applied our method implemented in the R/Bioconductar package STAN to a set of ChIP-chip experiments and RNA expression measures performed with tiling array in the yeast genome.

For this experiment we used the signal of 9 ChIP-chip experiments on different proteins that regulate or are involve during transcription: the RNA polymerase II (Pol II) subunit Rpb3, the CTD terminal of Pol II phosphorylated in Ser 5 and 2 (5SP, 2SP), the transcription factor II B (tfIIb), Ctk1 protein that phosphorylates Pol II, the Paf1, Spt5 and Spt16 proteins that regulates elongation, and the 3' end processing factor that cleaves and polyadenylates the pre-mRNA, the Pdf11 protein. To infer directionality, strand-specific expression signal is added

to the data set. These tracks will inform whether the Watson or Crick strand is transcribed in the region.

The bdHMM and bdClustering algorithms are applied to the data set using 4 directed and 1 undirected states. Viterbi decoding in the case of bdHMM, and posterior decoding for bdClustering, is then used to infer the latent variable Z *and* C respectively. In figure 4 we show a segment of chromosome IV with the data and results. As it has been mentioned in the introduction, genomic observations have a linear dependency. So close loci have similar properties. Therefore, the bdHMM is the optimal model to decipher the underlying process. Moreover, if compared with the SGD gene annotation, the bdHMM can differ regions that are expressed in the Watson (F states) or in the Crick (R states) strand, as well as the undireceted (U) processes. bdClustering is more prone to jump to different states. This is due that every single data point is taken independently of the surroundings, producing a discontinuous signal (figure 4).

Although we could not find any practical application of our bdClustering algorithm, the upcoming of single-cell technologies might be a field in which the application of such algorithm could be helpful in the discovery of new cell clusters using strand-specific sequencing methods. Nevertheless, bdClustering can be used for the initialization of the parameters of a bdHMM.

6 Conclusion

In this chapter, we have introduced the concept of HMMs and their theory. The Baum-Welch algorithm is able to estimate the HMM parameters through an iterative optimization process of the target function $\mathcal{Q}(\theta, \theta^{old})$. Thereafter, we have explained the properties that define a bdHMM and we derived a modified target function to estimate the parameters of the bdHMM.

Finally, we introduced a simple method to transform a (bd)clustering model into a (bd)HMM model such that we can re-use the EM algorithm implemented in the STAN package to fit the parameters of a (bd)clustering. The STAN package has implemented a vast set of emission distribution families that can be used to model genomic and other data: (multivariate) Gaussian, negative binomial, zero-inflated negative binomial, Poisson log-normal, Bernoulli, mulginomial and negative multinomial. Therefore, our new method to transform mixture models into HMM gives the opportunity to fit mixture models with all the previous mentioned distributions. Moreover, if the data set contains observations that may have some directionality information, we can find twin clusters independently of the source of the observation. In summary, here we have presented different clustering methods. We show that a mixture model can be expanded to HMM when there is a linear dependence on the hidden variables. Moreover, if the underlying process does is ruled by some directionality constraints we can generalized both clustering methods to a whole new system of directionality, see figure5.

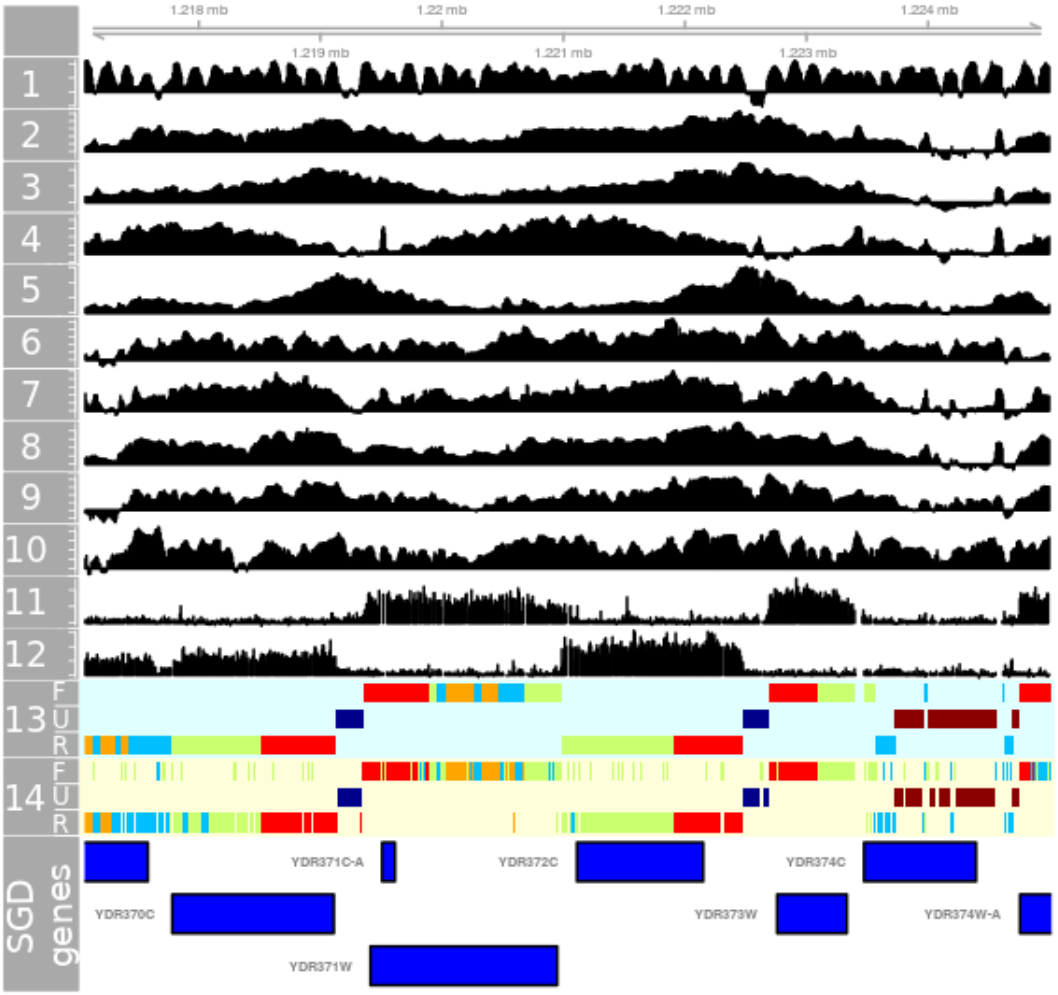


Figure 4: **Example**. (1) Nucleosome, (2) Rpb3, (3) S5P, (4) S2P, (5) TFIIB, (6) Ctk1, (7) Paf1, (8) Spt5, (9) Spt16, (10) Pcf11, (11) RNA Watson strand, (12) RNA Crick strand, (13) bdHMM Viterbi decoding; F- Forward states; U- Unidirected states; R- Reverse states, (14) bdClustering posterior decoding; F- Forward clusters; U- Undirected clusters; R- Reverse clusters. We applied bdHMM and bdClustering algorithm to segment a small fragment from Chromosome 4 of the Saccharomyces cerevisiae genome. We used as undirected observations the ChIP-chip signal of 9 proteins involved in the transcription machinery and, as directed observations, the strand-specific expression signal. After Viterbi decoding for the bdHMM and the posterior decoding for bdClustering, we compare the state annotation for both algorithms together with the known SGD gene annotation.

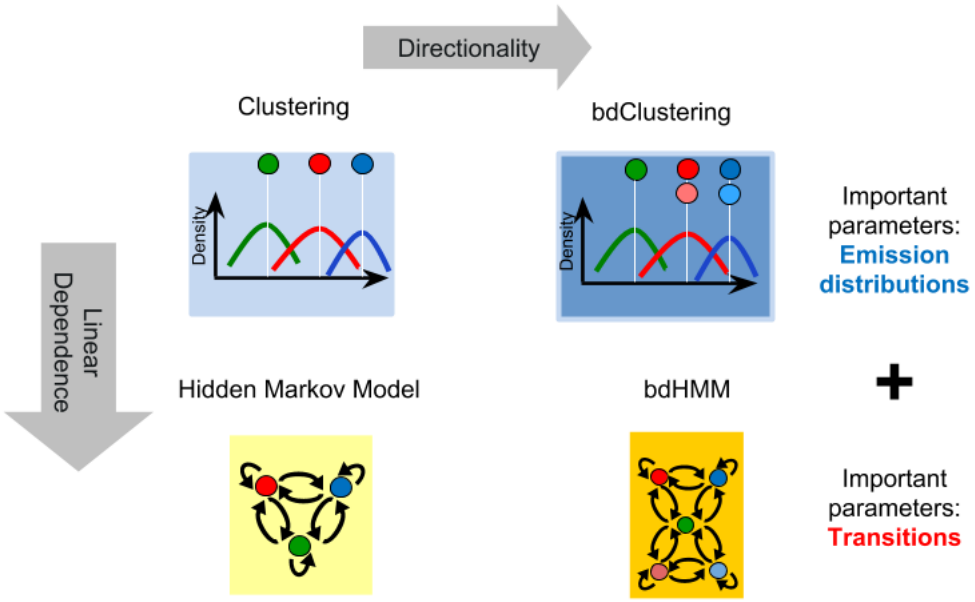


Figure 5: **Summary.** Clustering methods have been widely studied in the literature. Here, we show that when adding linear dependence to a standard mixture model, we obtain an HMM. Moreover, we have added another layer of complexity or information in which the underlying directionality of the process generating the data can be estimated. Thus, mixture models and HMM can be expanded to the bdClustering and bdHMM respectively.

Part II
Histone binding proteins in the regulation of Gene Expression

7 Introduction

Eukaryotic DNA is packaged inside the cell in the form of chromatin. Chromatin is the total of all protein-DNA complexes found in the nucleus [20, 21]. One of the challenges in the chromatin field is to understand how the transcription machinery interacts with the packaged DNA [22, 23]. Several studies [24] have demonstrated the involvement of different factors such as DNA methylation, histone modifications or even small nuclear RNA in the regulation of transcription. The interaction of these factors has been called "epigenetic regulation" in the literature [25].

The fundamental chromatin unit is the nucleosome. Nucleosomes are formed by histone octamers (two copies of each histone H2a, H2b, H3, and H4) and 147 bp of DNA that coil 1.7 times around the histone octamer. Nucleosome movement, and remodeling (binding/unbinding to DNA) are related to chemical modifications of the histone proteins. The relationship between histone modifications and transcription has been studied in depth [25, 26, 27, 28]. Specific histone modifications or combinations thereof have been found to activate or repress gene expression [26].

7.1 RNA Synthesis by RNA Polymerase

Eukaryotic organisms are equipped with a complex machinery for the transcription of DNA into RNA. The most important protein are the RNA-polymerases, which can transcribe a template DNA sequence into an RNA sequence always from the 5' end of the gene to the 3'end [22, 29, 30]. Polymerases are assisted and regulated by numerous enzymes during transcription. Several RNA polymerases are known to act in the nucleus of the eukaryotic cell. These polymerases differ in their promoter regions and the chromatin structure that they can recognize. Therefore, they target different DNA regions: Pol I transcribes the 35S subunit of the ribosomal precursor RNA (rRNA) [24], Pol III synthesizes the transfer RNA (tRNA) and the small 5S subunit of the ribosomal RNA (rRNA) [24]. Pol II performs the transcription of protein-coding genes (mRNAs), non-coding RNAs, small nucleolar RNAs (snoRNAs) and cryptic unstable transcripts (CUTs) [24, 30].

There are also RNA Polymerases found in other DNA-containing cell organelles. Mitochondria and chloroplasts have their own polymerases that catalyze the transcription of their own genes (mitoPol and PEP, respectively) [31]. Both

have a structure similar to the RNA polymerase of prokaryotes. In addition, two new polymerases have been identified in plants: PolIV and Pol V that have a specific function in DNA methylation.

7.2 Transcription Cycle of Pol II

The transcription process is tightly regulated by many different factors that interact with the Pol II and maintain the chromatin structure. The factors that interfere at the start of transcription and the end are significantly different (figure 6). Based on the factors required, the transcription process can be divided into three essential phases: initiation, elongation, and termination. During transcription, Pol II undergoes several post-translational changes, the best known being the phosphorylation of the C-terminal domain (CTD) of the Rpb1 subunit [30, 24]. The CTD consists of a repetition of the heptamer Tyr-Ser-Pro-Thr-Ser-Pro-Ser that undergoes specific chemical modifications throughout the transcription [32].

Initiation is the most crucial step for successful transcription of a gene [33]. In this step, Pol II is recruited to the promoter region of the gene by a protein complex that forms upstream of the transcription start site (TSS) and forms the preinitiation complex (PIC) [34]. The PIC is complete by the binding of TFIIH, which unwinds the DNA and facilitates the binding of Pol II to the template strand. Once the Pol II-DNA complex is established, TFIIH catalyzes the phosphorylation at the Ser5 residues in the heptamer repeats of the CTD and releases the PIC complex. The COMPASS complex recognizes the Ser5 phosphorylation of the CTD and links to the initiating Pol II. The methyltransferase Set1 belongs to the COMPASS complex and is responsible to methylate the Lys4 of the histone H3 at the promoter region [35, 36].

Around 150 nucleotides downstream from the TSS, Pol II enters in the elongation state. At this state Pol II is tightly regulated by many factors that help Pol II to transcribe and move along the chromatin. After phosphorylation of the Ser2 of the CTD by the Ctk kinase, the methyltransferase Set2 is recruited to the complex and methylates histone H3 at Lys 36 at the gene body region [24, 34].

For the termination of transcription, Pol II is released from the template DNA strand through a process of endonucleolytic excision of RNA followed by synthesis of the poly-A tail. To carry out this excision, termination factors (e.g. Pcf11 and Rtt103) have to be recruited to the transcription machinery by de-phosphorylation of Tyr 1 at the CTD, while some elongation factors (e.g. Nrd1) have to be released [34].

Finally, Pol II can be recruited again through the complex mediator at the promoter regions and facilitates the restart of the transcription. In yeast, gene loops have been found, in which the promoter and terminator ends of a gene are in physical proximity to each other, and assist in iterating transcription initiation [37, 38, 39, 40].

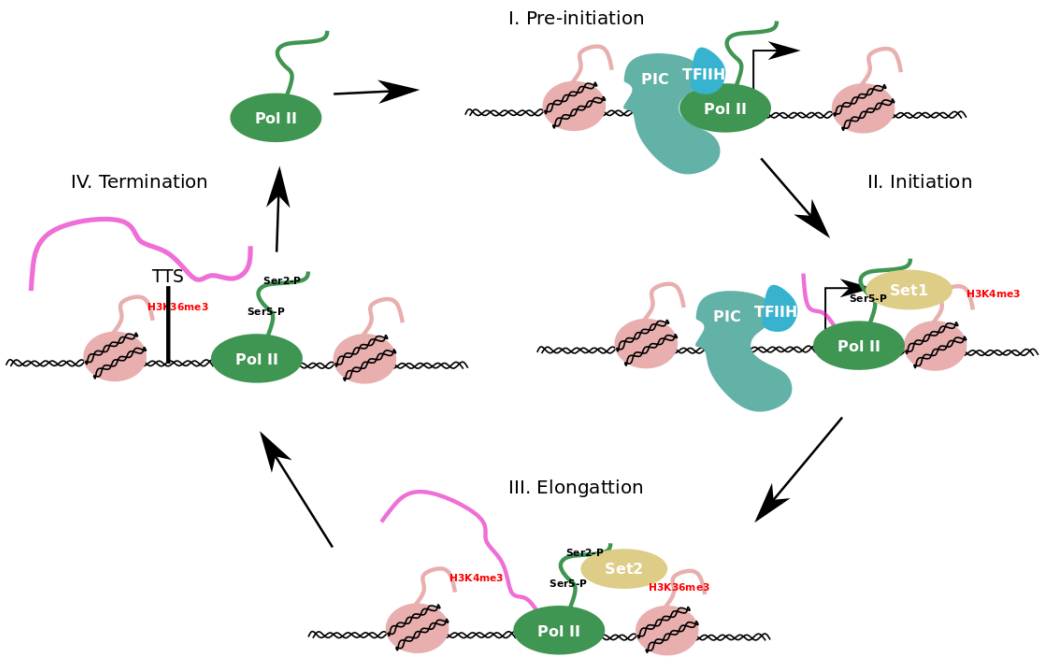


Figure 6: **Essential steps of the Pol II transcription Cycle.** In this figure, are highlighted those factors important for the histone methylation marks along the chromatin. The histones are drawn as red circles. Only the N-terminal part of histone H3 is shown in detail. **I. Pre-initiation**. During pre-initiation, the PIC complex binds to the DNA at the nucleosome free region (NFR). This complex is needed for Pol II to bind to the DNA. TFIIH phosphorylates the Ser5 of the CTD heptamer of Pol II and releases it. **II. initiation.** Pol II synthesizes around 150 nucleotides of RNA (violet). Set 1, a methyltransferase protein, binds to Pol II through the phosphorylated Ser 5 and introduces the methyl mark at Lys 4 of the H3 N-terminal. **III. Elongation.** The Ser2 amino acid of the CTD repeats are phosphorylated by the Ctk protein. This chemical modification is recognized by the Set2 protein, which in turn methylates the histone 3 at lysine 36. This mark is set along the ORF body of the gene. **IV. Termination.** After the transcription termination site (TTS), the RNA is cleaved from the Pol II and released from the DNA.

7.3 Histone Methylation and Gene Regulation

There is a wide literature about post-translational modifications (PTMs) of histones. Including: Acetylation, methylation, Phosphorylation etc [41, 25, 42]. They all occur at the N-terminal tail of histone proteins. PTMs are found throughout the genome and many have been characterized to define different regions (e.g. heterochromatin, telomeric, etc). Some modifications, like acetylation, change the total charge of the histone altering the strength by which the DNA-histone bind [41]. Therefore, changing the fluidity of the DNA and accessibility of other factors to the DNA [43]. Methylation does not change the total charge of the histones but rather works as a signaling mark that is recognized by other proteins [21, 28, 44].

Methylation at lysines in positions 4, 36, and 79 of histone H3 (H3K4me, H3K36me, and H3K79me) are markers for transcribed genes and their methylation state has been reported to correlate with active gene transcription [25, 29, 45, 46]. Moreover, Lys can carry up to three methylations on the $\epsilon-$nitrogen and the methylation state have different locations and patterns along the genes (figure 7).

Set1, is the methyltransferase that travels together with Pol II and methylates monomethyl H3K4 (H3K4me1) [47]into the di- and tri- methyl state (H3K4me2 and H3K4me3). The H3K4me3 is correlated with the transcription frequency of a gene and serves as the defining mark for the start of the ORF [46, 48]. The function of this mark is not completely understood yet but it has been shown that it serves as signaling mark for other histone modifiers and chromatin remodelers, which can recognize it and bind to it through their PHD domain [43, 20].

Tri-methylation of histone H3 at Lys 36 (H3K36me3) is associated with the body of transcribed genes [49]. This mark is catalyzed by the enzyme Set2 that is attached to Pol II during the elongation stage. The level of H3K36me3 has been correlated to the transcription rate of the gene and recruits the Rpd3S histone deacetylase complex to the gene body. H3K36me3 can be read by several proteins that contain the domain PWWP, named after its central core Pro-Trp-Trp-Pro (PWWP) [43, 46].

Lys 79 belong to the globular domain in H3 and it is methylated by the Dot1 protein [50, 36], which does not belong the Set family. H3K79me is a genome-wide mark that can be found over 90% of the genome. Therefore, its function is not well understood but it is associated with actively transcribed genes. It is also the only methyl PTM that has no demethyltransferase known at the moment [25]. Proteins containing the Tudor domain are able to read and bind to H3K79me3 [43, 20].

Thus, the histone PTMs described above have an important role in gene transcription and it has been hypothesized the "short term memory" model of transcription based on these marks [25, 39]. When a gene is marked with H3K4me3 and H3K36me3, it informs the cell of its transcription status. Depletion or errors of those marks have been linked to numerous cancers in humans [51, 29, 52, 50].

In this chapter, we are going to try to identify new histone protein readers

	Methyl-transferase	Domain
H3K4me3	Set1	PHD
H3K36me3	Set2	PWWP
H3K79me3	Dot1	Tudor

Table 1: **Histone marks, methyltransferases and binding domains.** The methylation at different lysines is carried out by different methyltransferases enzymes. These methyltransferases contain specific domains that allow the recognition of the specific chemical modifications. There are two main classes of methyltransferases for H3: those with the SET domain and the ones that do not contain another domain (Dot1).

that can bind these different marks and localize them inside the transcription cycle. For that purpose, we used HHPred software to find possible candidates that bind to those markers and ChIP-seq them. After pre-processing of the data we applied our bdHMM model to identify possible candidate readers with the different histone marks.

7.4 Chromatin Immunoprecipitation Sequencing Assay

ChIP-seq is the reference method to localize protein-DNA interactions throughout the genome by applying chromatin immunoprecipitations (ChIP) and DNA sequencing[53]. The first step consists in the cross-linking of the DNA with the proteins that are in contact in vivo. Next, the cells are lysed and the chromatin is isolated. The DNA is fragmented and the protein-DNA complexes of interest are precipitated with a specific antibody for the protein to be studied. The DNA is released from the protein, labeled with an oligonucleotide adaptor and sequenced [24, 30].

Generally, parallel to ChIP-seq a control experiment is run to identify non-specific background called input signal. The experimental design of the control is similar to the ChIP-seq but instead of a specific antibody, a mock immunoprecipitation is used before sequencing. This way, enrichment between the ChIP-seq signal and the input can be computed controlling for potential biases.

8 Material and Methods

Our analysis workflow consists of several steps: The candidate screening, the raw data acquisition by ChIP-Seq, their mapping, and processing into a smooth, genome-wide protein occupancy profile, and the joint analysis of all these profiles by a bidirecitonal HMM (see figure 8).

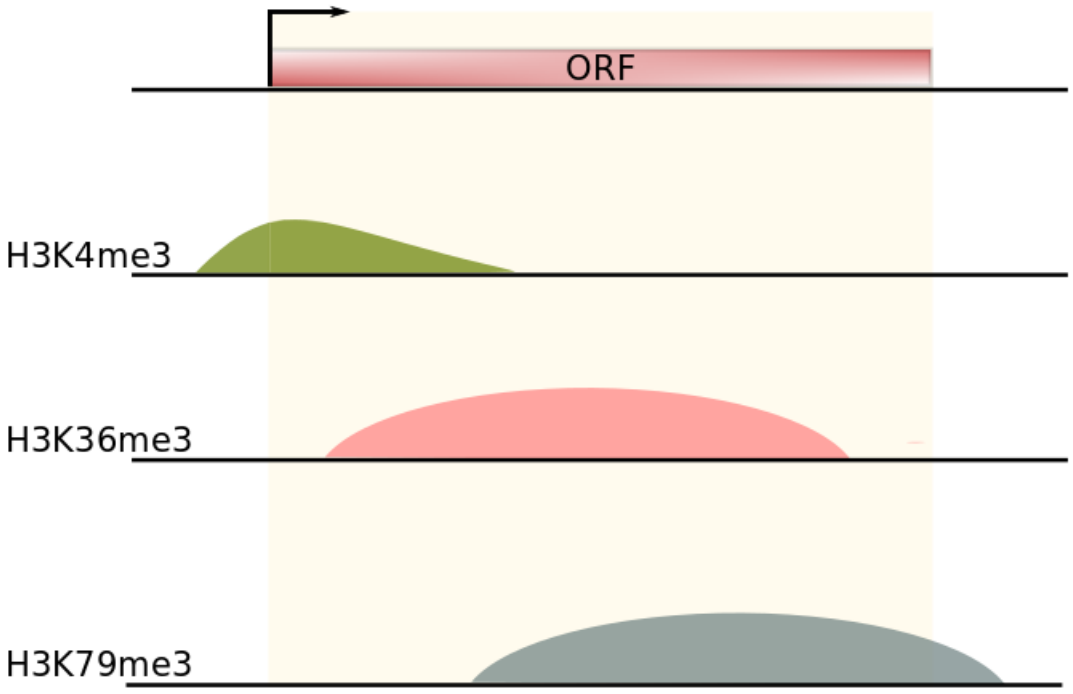


Figure 7: **Histone methylation pattern at transcribed ORF**. The distribution of histone H3 methylation along the average ORF can be characterized depending on the Lys and methylation state. H3K4me3 mark is found at the 5' end of the ORF. The H3K36me3 is a known mark for the gene body and highly correlates with gene expression status in a cell. The H3K79me3 is well known and has several functions in the cell, e.g. cell cycle regulation and DNA repair. It has also been linked to transcribed genes. The methyltransferase Dot1 is thought to interact with Pol II due that the H3K79me3 signal is found enriched at the end of the gene body of transcribed genes. Misplacement of these patterns or completely loss have been linked to cancer in humans.

8.1 Finding Candidate Readers

In order to obtain possible readers of the histone PTMs, we used the software HHPred [54] to identify new proteins that might contain the domains necessary to recognize them. We took the PHD (H3K4me3), PWWP (H3K36me3) and Tudor (H3K79me3) domains to compare them against the protein database of the Saccharomyces cerevisiae.

As a result we found 6 candidates that can bind to either of the marks:

- **Asr1** is a ubiquitin ligase that interacts with Pol II during transcription. It has not been reported to bind any of the histone marks but it contains the PHD finger that could bind the H3K4me3 and potentially the H3K36me3 mark [55, 56].
- **Ioc4** belongs to the chromatin remodeler complex Isw1b. It contains the PWWP domain that binds to H3K36me3 but the results of HHPred shows that it contains a domain similar to the PHD [46].
- **Nto1** is a subunit of the histone acetyltransferase complex NuA3. It has been reported that binds to the H3K4me3 through its PHD domain and potentially could bind H3K36me3 [57, 58, 45].
- **Pdp3** as Nto1, Pdp3 also belongs to the complex NuA3 and relocalize to the cytosol in response to hypoxia. Regulates the interaction of the histone mark H3K36 and NuA3b subunit. It contains the PHD finger that could potentially bind to H3K4me3 [58, 45].
- **Set4** is not well described in the literature. The function is nowadays still unknown. It belongs to the methyltransferase family SET. The HHPred results show that it has a domain that could bind to H3K4me3 or H3K36me3 [59, 49, 48].
- **Rad9** is a DNA damage-dependent checkpoint protein. It is the only protein found to have a Tudor domain able to bind H3K79me3 [60, 61, 62].

The candidate proteins and their known or putative binding preferences are listed in table 2.

To investigate the link of these proteins with the histone marks ChIP-seq experiments were carried out on the 3 histone marks (H3K4me3, H3K36me3, and H3K79me3), 3 known methyltransferases to these marks (Dot1, Set1 and Set2) and the 6 candidates (Asr1, Ioc4, Nto1, Pdp3, Rad9 and, Set4).

8.2 ChIP-seq

Yeast strains were grown in 600 mL YPD medium to mid-log phase and cross-linked with formaldehyde. DNA was extracted and whole-genome libraries for IP

	Asr1	Ioc4	Nto1	Pdp3	Set4	Rad9
H3K4me3						
H3K36me3						
H3K79me3						

Table 2: **Putative chromatin readers and their binding preferences.** After using HHPred to find candidates proteins in the *Saccharomyces cerevisiae* genome, 6 proteins arose as possible binders to the marks H3K4me3, H3K36me3, and H3K79me3. Some of the proteins were known to bind some histone modifications but potentially could bind different ones. In this table are highlighted binding that has previously been reported in green, binding predicted by HHPred algorithm in orange and binding that can be excluded is marked in red.

and input were prepared using the ThruPLEX DNA-seq Kit. The libraries were sequenced with an Illumina HiSeq 1500 sequencer in 50bp paired-end read mode.

Pre-processing of the data was carried out by Michael Lidschreiber. The paired-end reads were aligned to the reference genome of *Saccharomyces cerevisiae* (sacCer3, version 64.2.1) using Bowtie [63]. The reads were filtered based on standard quality control settings. The mid-points of the two ends of each read were kept to obtain the coverage tracks (counts per genomic position). The data was imported into R for further analysis, see table 3.

8.3 GenoGAM

The coverage tracks for IP and input samples of each protein were transformed into protein occupancy tracks using the R/Bioconductor package GenoGAM. Each chromosome was split in tiles of thousand nucleotides, such that consecutive tiles had an overlap of 300nt. A generalized additive model (GAM) was fitted on each tile using parallel processing. The result of this step was combined into a single smooth signal representing the (log) occupancy of each protein of interest along the genome. This signal was sampled every 50 nucleotides (nt) and served as input to the next step, the bdHMM annotation.

8.4 RNA Sequencing

RNA sequencing (RNA-seq) is an experimental procedure to determine the amount of RNA present in an organism at a given moment. Through this method, one can capture the different species, e.g., mRNA, tRNA, etc. For this work, the strand-specific RNA-seq data set was prepared as in *Battaglia et al. [24]* in two different replicates.

The bam files were loaded to R and the strand-specific read count for each count was performed using the GenomicRanges package. The annotation of the

features was taken from [24].

In order to normalize the read count of each gene by its length and library size, we used the transcripts per million (TPM) measurement. TPM is a similar measurement as reads per kilobase million (RPKM) but with some modifications:

Given an RNA-seq experiment, we can compute the reads per kilobase (RPK) for gene i as:

$$RPK_i = \frac{c_i}{l_i} \tag{34}$$

where c_i and l_i are the read count and length for gene i. Next step is to normalize for sequencing depth. To this end, the RPK_i for all the transcritps T are summed up:

$$N = \sum_{i=1}^{T} RPK_i \tag{35}$$

then, we can use equation 34 and 35 to obtain the TPM for gene i:

$$TPM_i = \frac{RPK_i}{N} \tag{36}$$

The average TPM of both replicates per gene was used to estimate the level of expression of each gene i.

8.5 Bidirectional HMM

We used different combinations of directed and undirected states to fit a bdHMM to the data. The directionality in all combinations was evaluated and the bdHMM with the lowest number of directed states that had a directionality score higher than 0.5 was taken for further analysis. For a more detailed understanding of how the directionality score works refer to chapter 1. We chose multivariate Gaussian distributions as emission distributions in the bdHMM algorithm. The fitted model was applied to segment the genome in the different states using the Viterbi algorithm.

9 Results

We applied the bdHMM to ChIP-seq data in *S. cerevisiae,* including three histone marks (H3K4me3, H4K36me3 and H3K79me3), three known methyltransferases to these marks (Dot1, Set1 and Set2) and another six proteins which contain a domain able to bind to either of the histone marks (Asr1, Ioc4, Nto1, Pdp3, Rad9 and, Set4).

ChIP-seq experiments are performed with an specific antibody to precipitate the protein of interest and a mock antibody to identify nonspecific background.

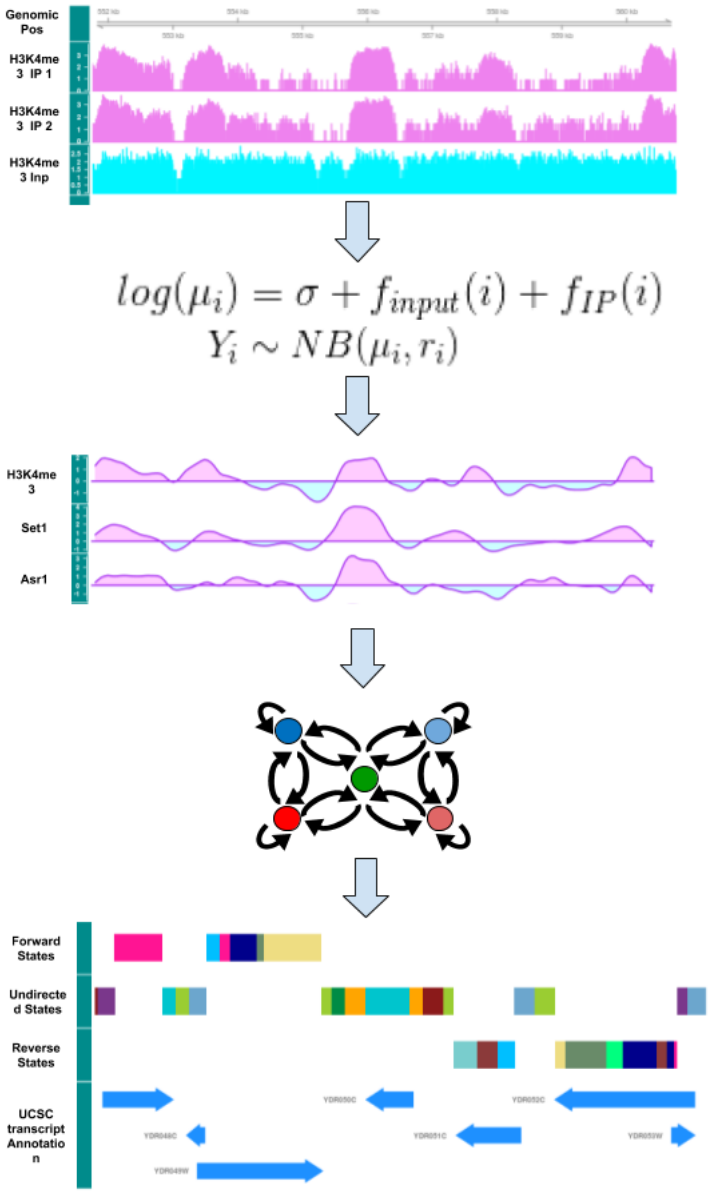


Figure 8: **Method workflow**. The ChIP-seq IP and input reads are aligned against the SacCer3 genome to find the physical location. The midpoint of the paired-end reads is taken to compute the coverage of each track. Each protein was immunoprecipitated using a different number of replicates and inputs. Several methods have been discussed in order to processed ChIP-seq signal along the genome and estimate the real coverage. For our study, we used generalized additive models (GAM), which are implemented in the bioconductor package GenoGAM. After obtaining a smooth signal for each protein, we learned a bdHMM model and calculated its Viterbi path on the signal to segment the genome in 9 directed and 12 undirected genomic states.

Protein	# IP	# Inp
H3K4me3	2	2
H3K36me3	2	1
H3K79me3	2	1
Dot1	1	1
Set1	1	1
Set2	1	1
Asr1	2	2
Ioc4	2	2
Nto1	1	1
Pdp3	2	2
Rad9	2	2
Set4	2	2

Table 3: **IP and Input replicate**. The genome-wide coverage of the histone marks, methyltransferases and candidates was obtained using ChIP-seq experiments. This table shows the number of replicates that were carried out for each protein mark.

The raw reads from the specific antibody are comonly referred as IP signal and the data from the mock experiment as input (Inp). Many heuristic methods have been used to correct the bias using those two signals. Usually, the trivial way has been to divide the IP counts by the control counts over some sliding window of width w *[64, 65]*. This methodology carries some limitations, i.e., dividing count data can lead to some problems (division by zero) or that there is no flexible way to integrate multiple IPs and Inps. Moreover, there is no statistical solution to estimate the correct coverage at each position [66, 67, 53].

Generalized additive models applied to ChIP-seq data have been recently developed by *Stricker et al. [68]*. This statistical framework is able to overcome the problems mentioned above and provides a robust method to estimate the correct coverage of each protein along the genome. We used the R/Bioconductor package genoGAM [68] for each protein and used the provided IP and Inp raw data (see table 3) signals to estimate the occupancy of each signal along the genome.

The smoothed signal for the 12 protein marks was then used as the training data for the bdHMM using multivariate Gaussian distributions. The number of directed and undirected states need to be chosen. Although standard model selection criteria as BIC, AIC have been previously used to estimate the number of states in HMM, this method does not apply well in our case. Fitting ChIP-seq data with HMMs have been reported to be problematic due to the high dimension and the high variance along the genome [7]. We tried different directed and undirected state numbers and computed the directionality score of the directed states in each case. After analyzing the directionality score of all models we continued with a model using nine directed states and twelve undirected states. The criteria to

select this model was that it was the state with the lower number of states for which all the directed states had a directionality score higher than 0.5. This way, we aim to have the model that is capable to capture the real directionality of the data while using the minimum number of states. This will help to interpret of each state into a biological context.

The mean profile of a state identifies co-occurrences of histone readers, modifiers and histone modifications. Each state represents a multivariate Gaussian distribution. The mean of this distribution indicates the composition of proteins in that state (figure 9). For instance, state F/R0-8 is highly enriched in all measured proteins (note that by definition, forward and reverse twin states share identical emission distributions). In contrast, state U-2 is completely void of any protein binding signal. State U-10 is characterized by its high Rad9 levels. Another state worth noting is U5, which is undirected yet shows high occupancies for most proteins. This is, at first sight, surprising for undirected states, which are most likely found in intergenic regions. Vice versa, the absence of a strong signal in the directed state F/R-7 is also unexpected.

Some states are enriched at functionally relevant genomic features. It is natural to ask whether a certain state is enriched or depleted in specific genomic regions. To this end, we used the annotation provided by *Battaglia et al.* to define genomic features, i.e., regions in the genome that share a common function. The set of genomic features that we have looked at are: promoter regions, open reading frames (ORF), 200 nucleotides downstream from the TTS, SUT and CUT transcripts, tRNA genes, snoRNAs, and snRNAs. We then computed the fold enrichment of each state in the different genomic features, relative to its mean frequency along in the whole genome (figure 10). As observed in figure figure 10, state U9 is highly enriched in the promoter regions, as well as state U-4 and U-1 but with less enrichment. Most of the directed states (e.g. F/R-3, F/R-4, F/R-5, F/R-6, and F/R-8) are enriched in ORFs together with the state U-5. State U-8 and U-11 are mostly found at the 3' end of the genes. tRNA genes are decoded by the state U-10 exclusively. The predominant mark in the state U-10 is Rad9 as previous studies by *Clelland et al. [69]* have shown the link between tRNA genes and Rad9. Finally, snoRNA and snRNA are mostly integrated by states U-11 and U-5.

To investigate the state frequencies along ORF we performed a metagene analysis. Since the histone marks that we are analyzing have been reported to expressed genes we cluster our genes in low expressed, medium expressed and highly expressed based on RNA-seq data. To avoid problems when re-sizing the genes we removed the short genes, leaving out 1234 genes (see table table 4). In our metagene analysis, we looked in the state frequency 200 nucleotides upstream the promoter start site, along with the gene body and 200 nucleotides downstream

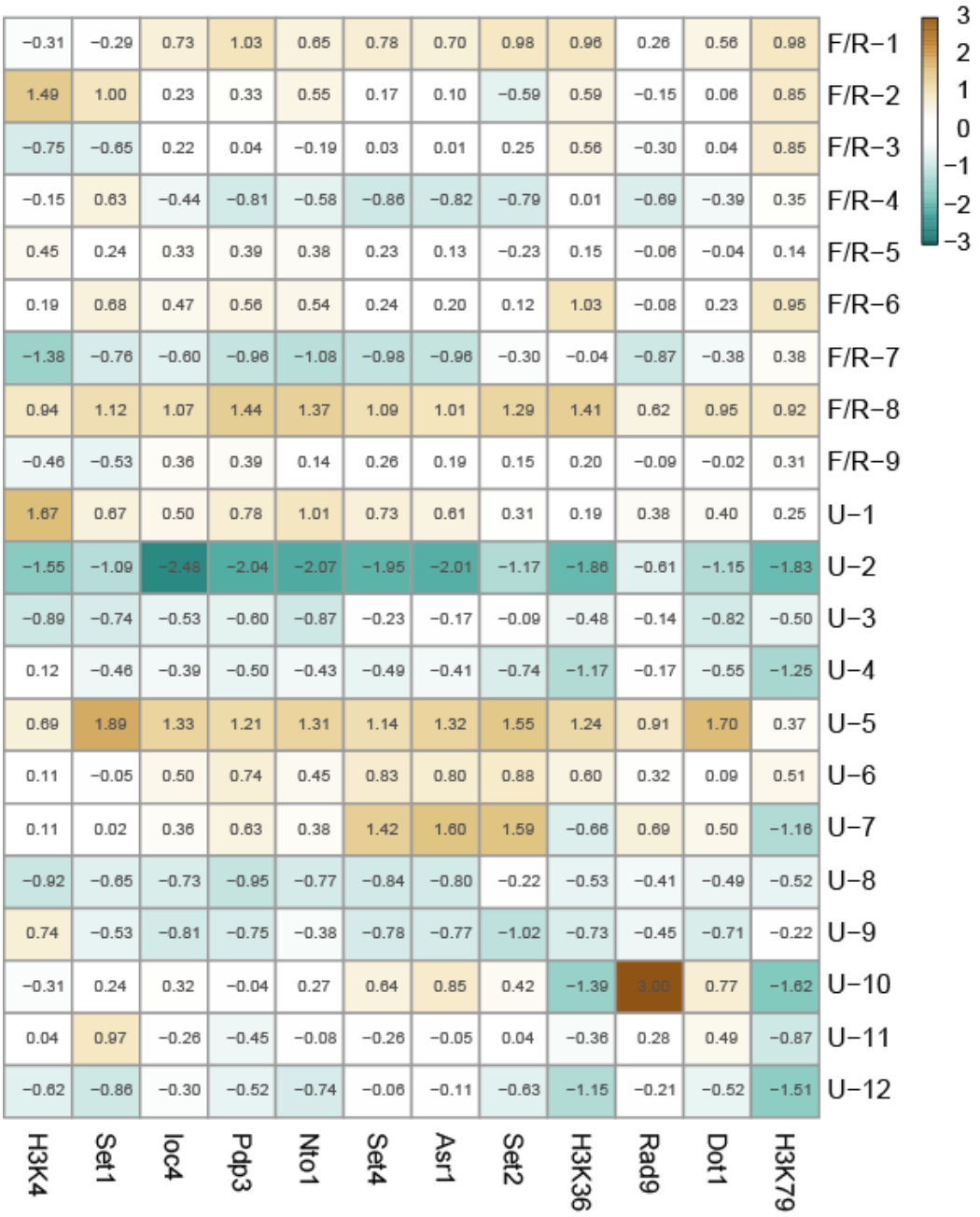


Figure 9: **Mean signal of states.** After fitting the bdHMM using the multivariate Gaussian distribution, each state is defined by the mean vector of the state and the covariance matrix. Twin states Forward-Reverse have the same mean vector since they describe the same combination of marks but in different underlying direction.

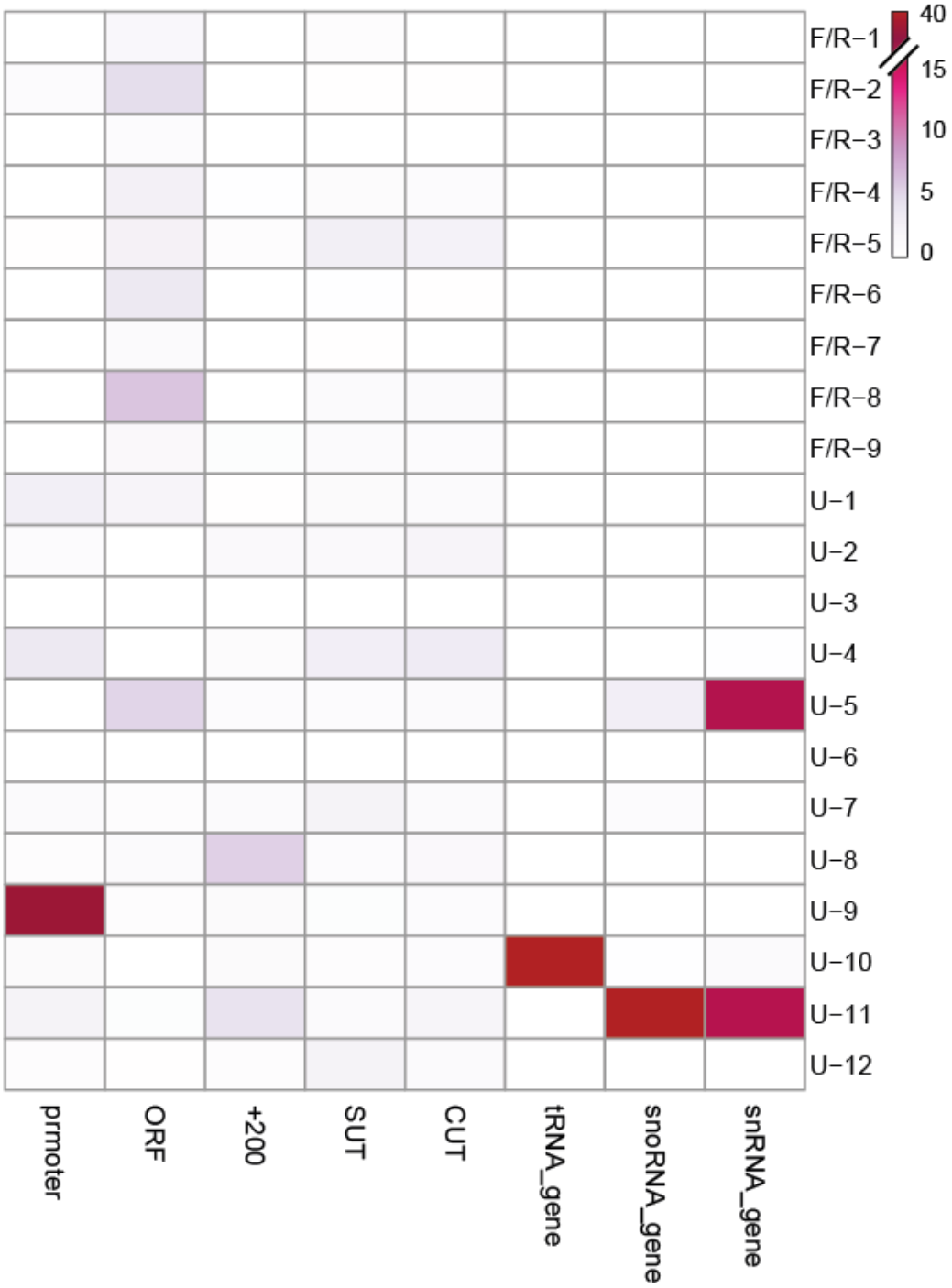


Figure 10: **State fold enrichment.** We looked for the enrichment of each state at different genome feature. The x-axis shows the 8 features selected to test the enrichment of each state (y-axis).

	Short	Medium-Large
Low	444	819
Medium	554	1884
High	236	991
Total	1234	3694

Table 4: **Number of genes per group.** The set of total genes were divided by length and expression level. The Low/Short and High genes were selected as the lower and upper quantile respectively. Since the final goal is to performa a metagene analysis of the state sequence along the ORF, all the short genes (1234 genes) were removed.

the transcription termination site (TTS) for the low, medium and high expressed remaining genes.

9.1 Low Abundance Genes State Frequency

Low expressed genes (figure 11) have a peak in the promoter region of state U-4 and U-9. State U-4 contains a low signal of all tracks and U-9 is rich in H3K4me3 (figure 9). After this first states, it jumps to state F/R-5 with is mildly enriched in H3K4me3, Ioc4, Pdp3, and Nto1 and in really low quantity H3K36 and H3k79 trimethylated. They reach the end of transcription by jumping to state F/R-9 which the mean tracks are Ioc4 and Pdp3 to end with the state U-8. State U-8 is also depleted in all tracks what indicates that it is a termination state before entering intergenic regions.

9.2 Medium Expression Genes State Frequency

Medium expressed genes (figure 12) have a similar starting and ending of the genes as the low expressed with the difference that the state enriched in H3K4me3, state U-9, is much more frequent. Therefore, state U-9 can be interpreted as an initiation state for transcription. The main difference between the low expressed genes comes from the gene body. First, we have a transition to the state F/R-2 which is highly enriched in the histone marks H3K4me3, H3K36me3, and H3K79me3, as well as the methyltransferase Set1 in higher amount followed by the Pdp3 and Nto1. State U-1, which is present in low quantity, is enriched in most of the marks except Set2, H3K36me3, and H3K79me3. It is arguably why an undirected state is enriched in the gene body. This observation could be explained by the high density of genes in the yeast genome. There are many genes that overlap, hence, two directed mechanisms in opposite direction would be seen as an undirected observation overall. After state F/R-2, the most common state is the state F/R-6

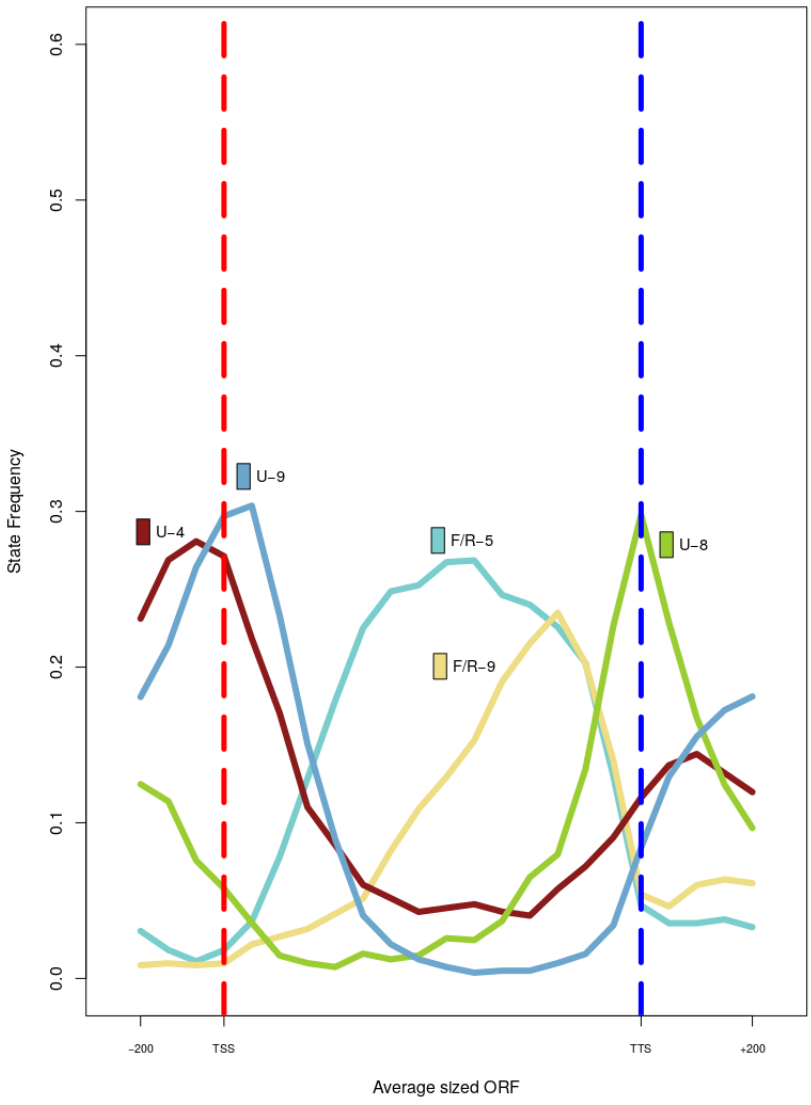


Figure 11: **Spatial state frequency of low abundance genes.** A meta-gene analysis of medium-large size genes with low expression levels (819 genes). The y-axis shows the frequency of each state on an average gene. All genes were normalized to a similar length and 200 nt upstream of the TSS and 200 nt downstream of the TTS was taken as well.

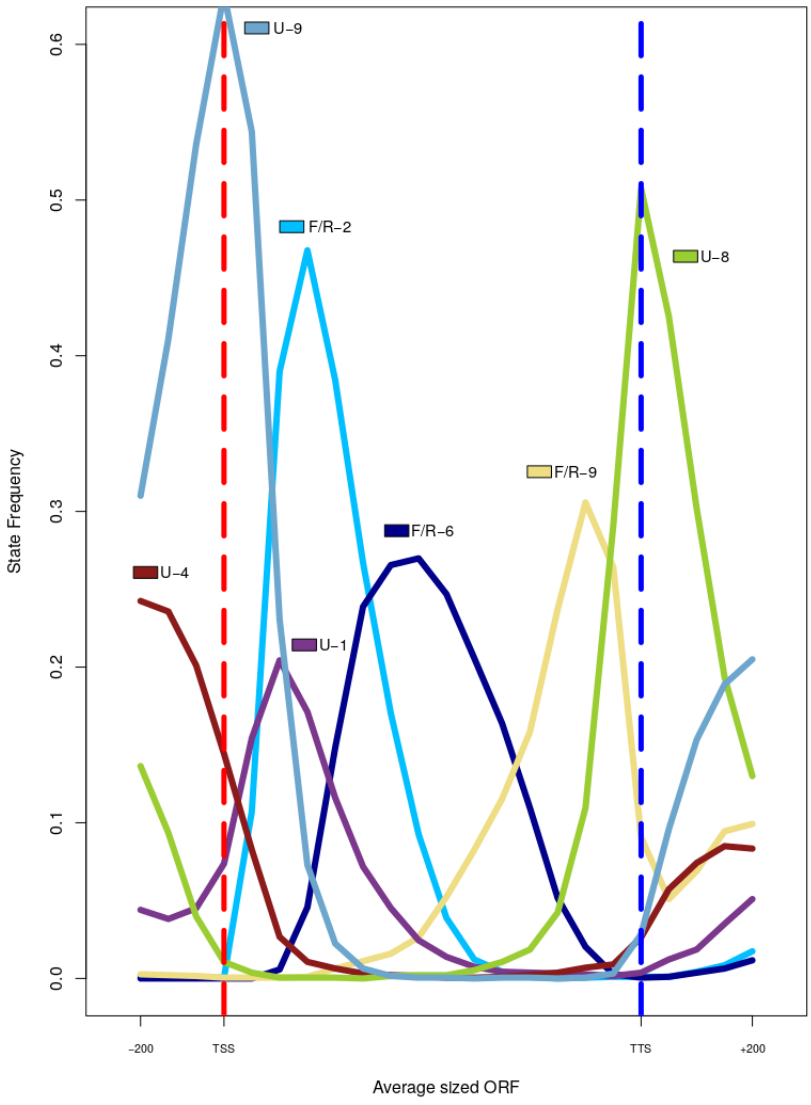


Figure 12: **Spatial state frequency of medium expression genes.** A meta-gene analysis of medium-large size genes with medium expression level (1884 genes). The y-axis shows the frequency of each state on an average gene. All genes were normalized to a similar length and 200 nt upstream of the TSS and 200 nt downstream of the TTS was taken as well.

which is predominated by the histone marks H3K36me3 and H3K79me3 and in lower intensity by Set1, Pdp3, and Nto1.

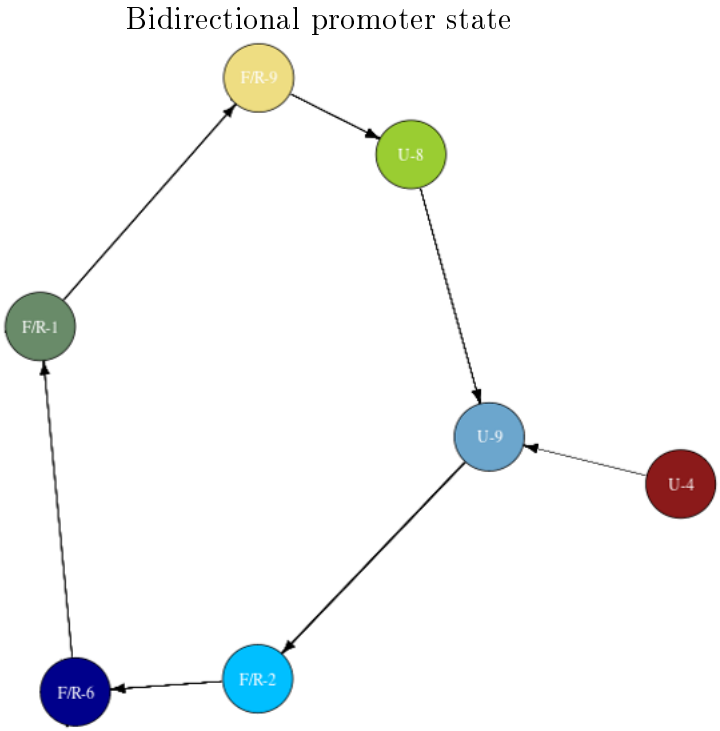
The flux diagram of states at genes with medium expression is illustrated in (figure 13). This graphic shows the most probable transitions in this set of genes (arrows). As mentioned in the paragraph above, from an intergenic state U-4 we enter to the promoter region at state U-9 which will go to state F/R-2 then to F/R-6 in early elongation to terminate with states F/R-1 and F/R-9 and at the TTS the state U-8. As a consequence of the close proximity of the genes in the yeast genome, termination state U-8 can jump directly to a new promoter state U-9 and start the transcription of a new gene.

9.3 Highly Expressed Genes State Frequency

The highly expressed genes (figure 14) start in a similar way with state U-9 being the most predominant one, then elongation is started with state F/R-2 to transit to the state F/R-6. Together with F/R-6, it is equally abundant the state F/R-8 which is enriched in all marks led by H3K36me3, its methyltransferase Set2 and, the Pdp3 and Nto1 proteins. Elongation is terminated with the state F/R-1, which is enriched with all marks except H3K4me3 and its methyltransferase Set1.

Not highly enriched, but two new undirected states appear to be more enriched in this set of genes. Those states are U-11 and U-5. State U-11 has a peak at the TSS and at the TTS while the state U-5 is highly enriched at the gene body. To investigate this further, we used hierarchical clustering of the state sequences of all genes using the Hamming distance (figure 14). Hamming distance measures the similarity of two strings of equal length by comparing the symbols at each position. One cluster of genes popped out having the state U-11 as initiation and termination and the state U-5 at the gene bodies. This observation explains why we observe undirected states in a directed process as gene transcription. Since the initiation and termination of the states are equal, there is a symmetry along the gene. Therefore, our symmetry assumptions for the bdHMM cannot distinguish the directionality.

State U-11 is enriched by the methyltransferases Dot1 and Set1, which methylate H3K79me3 and H3K4me3, respectively. State U-5 is enriched in all marks but specifically in the methyltransferases. Contradictory, the histone marks are not enriched, with H3K79 being the lowest one. This results might suggest some special mechanism involved on those genes. It has been reported previously [37, 38], that some genes in the yeast genome might take a loop conformation so that the polymerase can re-attach to the TSS in order to start another round of transcription. The state sequence on these genes might point to this specific path.



fre

Figure 13: **State flow for a medium expressed gene**. This graph diagram is obtained after computing the most common transitions during transcription of medium expressed genes. The standard flow of states starts in state U-4 and jumps to the promoter state U-9. After state U-9, transcription of the ORF starts by transitioning to state F/R-2, then to state F/R6 and so on. Termination occurs after state F/R-9 by state U-8. A new cycle of transcription can follow this sequence by jumping to the initiation state U-9 after U-8. Note that genes in the reverse strand had to be reversed in order to obtain the same flow. 7

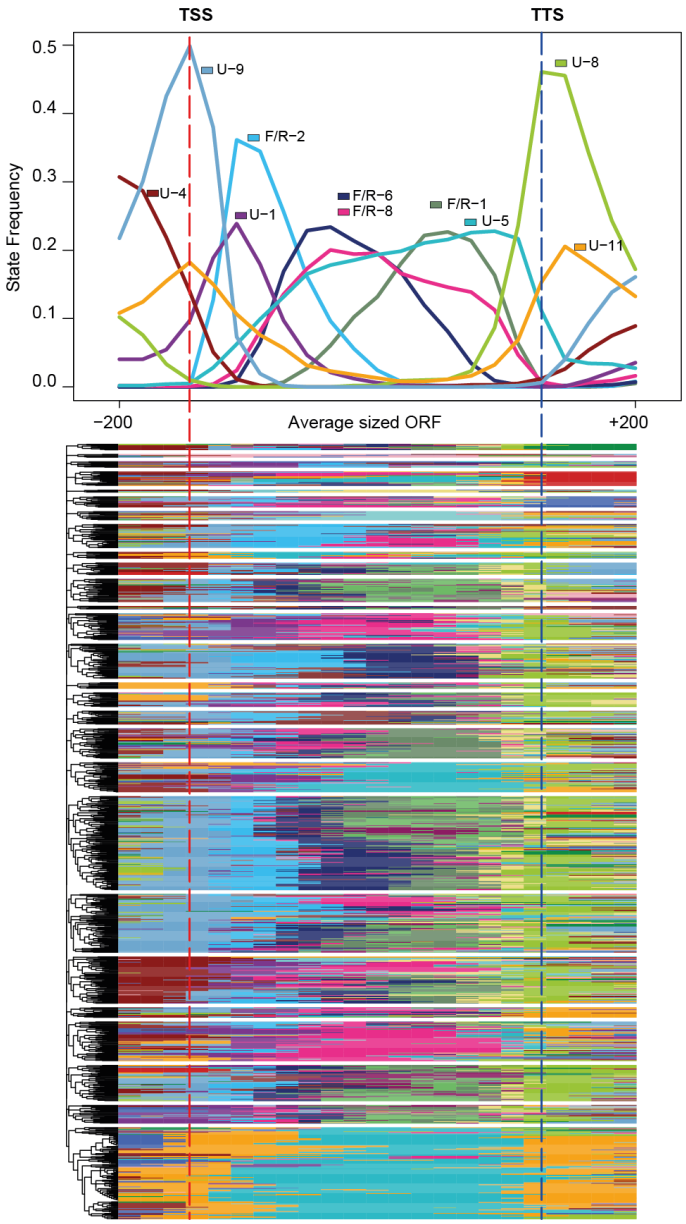


Figure 14: **Spatial state frequency of highly expressed genes.** A meta-gene analysis of medium-large size genes with high expression levels (991 genes). The y-axis shows the frequency of each state on an average gene. All genes were normalized to a similar length and 200 nt upstream of the TSS and 200 nt downstream of the TTS was taken as well. On the bottom, the genomic state sequence of those genes is shown. The sequences are clustered using the Hamming distance into 27 different clusters.

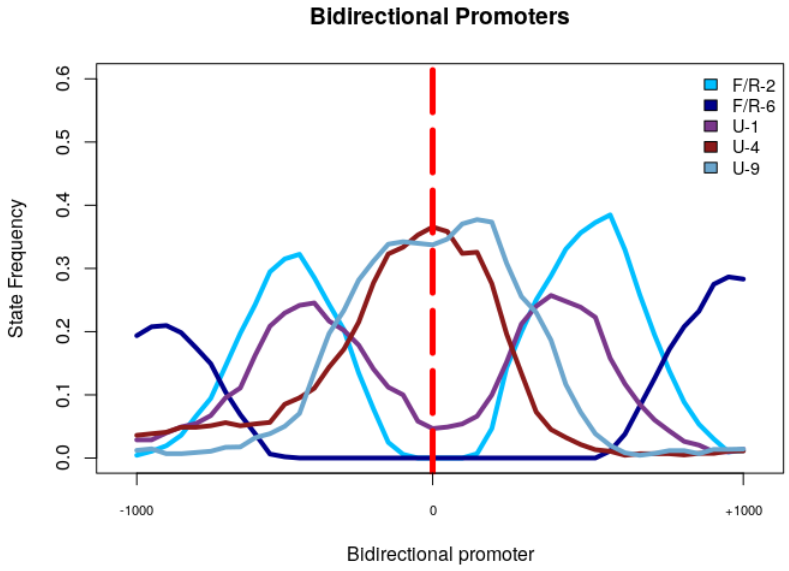


Figure 15: **Bidirectional promoter state frequency**. A meta-gene analysis of bidirectional pair of genes. The y-axis shows the frequency of each state on an average gene. From the pair of genes, the gene with higher expression level is on the right-hand side of the bidiretional promoter region.

9.4 State Frequency in Bidirectional Promoters

As the yeast genome is highly compacted, there are many genes that share the same promoter region. One gene is encoded on the minus strand and the other on the plus strand. We further looked into these pairs of genes and analyzed their shared promoter region.

On the left hand, we aligned the genes with the lower expression of the pair and on the right the ones with higher expression and analyzed the state frequency in both. In the central region, the state U-4 and U-9 compete with similar frequencies. State U-4 is a state depleted from any protein mark, and in this case, the frequency in the region between the TSS of both genes is increased. In bidirectional promoters, these regions will be occupied by the PIC proteins and no nucleosome could bind. This region is known as the nucleotide free region (NFR). Besides these observations, the flow of states along the genes are similar to those of medium transcribed genes. With transitions from state U-9 to state F/R-2. Although almost symmetric, states with higher expression (to the right of the red dashed line) (figure 15) have increased frequency of state F/R-2 and F/R-6.

10 Discussion

In this section, we have analyzed a set of ChIP-seq experiments composed of three histone methylation marks (H3K4me3, H3K36me3 and H3K79me3), their respective histone methyltransferases (Set1, Set2 and Dot1) and six candidate proteins that contain a potential binding domain for either of the 3 histone marks (Asr1, Ioc4, Nto1, Pdp3, Rad9 and, Set4). Our goal was to find evidence of the association of the candidate binders with the histone marks using clustering techniques. To this end, we have used generalized additive models to estimate the coverage of each track using multiple replicates of the IP and input. The smooth signal was then analyzed by our bdHMM algorithm in order to find different clusters which contain possible combinations of the tracks that suggest any correlation of the histone marks with the candidates. We further performed metagene analysis to provide a biological interpretation of the states in our bdHMM.

The metagene analysis based on expression (low abundance, medium, and highly expressed genes) shows differences in the transcription state flow. These differences are more obvious along the gene bodies, with genes highly expressed being enriched by states with a higher signal in most of the tracks. Low abundance genes show states almost depleted of any ChIP signal.

Despite the differences along the gene body, state U-9 is the promoter state in all cases. State U-9 is enriched in the H3K4me3 signal, which agrees with previous studies that show experimentally that H3K4me3 is a mark characteristic of promoter regions [45, 57]. As termination state, the state U-8, which is depleted from all signals, is the unique state at the TTS.

Although we could verify already published results about the location of the histone methylations along the gene body during transcription, and their relationship with known binders, our approach does not provide information regarding which histone modification (if any) might be bound by the new candidates. Recently, however, it has been shown that Pdp3 binds to H3K36me3 [58]. We provide here ChIP-seq data confirming their *in vitro* binding result. The reason of such negative results might be caused by the close spatial co-occurrence of these histone marks along the genome and inside the genes. Also, many of the candidates belong to protein complexes that are known to bind to other histone marks [58, 46, 48] and participate in the epigenetic signal pathway [44, 43]. The high combinatorial complexity of our analysis makes the discovery of new interactions difficult. Our model often cannot discern binding patterns with sufficient resolution.

Interestingly, we found a small set of genes with a symmetric flow of three states. initiation and termination of those genes are annotated by the state U-11, whereas the gene body is covered by the state U-5. These results may be explained by a physical conformation of a loop. By this conformation the 3' and 5' ends of the gene are in close proximity which has been previously reported to enhance transcription [37, 38, 39, 70]. This model would explain the observation of state U-11 at initiation and termination.

Hi-C or 3-C experiments have been developed in the last decades that can study the physical conformation of the genome [71, 72]. Although gene looping is a dynamic mechanism that the cell uses depending on the environment and necessities, these experiments could be carried out to investigate this further.

Part III
HMM for Genotyping

11 Introduction

Somatic cells of diploid organisms contain two copies of each chromosome (homologous chromosomes) and hence every individual has two possible alleles for a gene. Different alleles can be manifested in specific traits in the phenotype or even be defective [73, 74]. The extra copy lends robustness to the organism when one of the copies carries a non-functional mutation. Therefore, a recessive mutation does not necessarily result in a reduced fitness [75]. If both alleles at a specific locus are identical, the organism is homozygous with respect to that allele, otherwise, the organism is heterozygous with respect to the same allele [73, 76](figure 16).

Meiotic recombination is an essential mechanism in the sexual reproduction of diploid eukaryotic organisms that consists of the overcrossing (chiasmata) between pairs of homologous chromosomes [77]. This process ensures equitable segregation of the genomic material in each gamete. Due to this process, new combinations of alleles are transmitted to new generations allowing genetic variability among individuals and benefits the adaptation of populations. Recombination occurs in the so-called meiosis I and is triggered by double strand breaks (DBSs). The DBSs can be repaired by crossover with the homologous chromosome (CO) or by the same chromosome (NCO). The CO/NCO ratio is controlled in each chromosome and taxa and the mechanism by which it is determined the proportion of DSBs is resolved as CO is unknown [78, 77, 79].

Genotyping is the process by which the different alleles in an organism are characterized [80, 81]. Given some phenotype which is measured on a continuous scale (a quantitative trait), one seeks to find those genetic loci whose alleles explain the phenotypic variation in the quantitative trait. Such a locus is called a quantitative trait locus (QTL)[82]. In QTL studies, many individuals are genotyped and phenotyped, in order to find a statistical dependency between the alleles of a certain locus and the phenotype [73]. Most quantitative traits are complex traits, which result from the interaction of several weak QTLs [82]. Therefore, traditional genome-wide association approaches, which start perform the genotyping and phenotyping of naturally occurring populations, often lack the statistical power to detect such QTLs.

In plants, the distribution of CO along the chromosome is largely random, with some locations such as the centromere being disfavored, and some CO hotspots [83]. A better understanding of the crossing over process is essential for various branches of molecular biology such as medicine, agriculture, etc. [77]. The aim of this chapter is to offer a possible solution for the genotyping of recombinant populations, generated from ancestors with known genotype, by sparse sequencing

of their genomes.

11.1 Genotyping using NGS (GBS)

A detailed QTL analysis may still require the genotyping of hundreds or even thousands of individuals [81, 84, 85, 73]. The genotyping of individuals by partial sequencing of genomes by next generation sequencing (NGS) has become a fundamental tool in molecular biology as it allows the genotyping of thousands of markers in one batch, at an affordable cost.

Thanks to the development of NGS, the study of the genome of thousands of organisms has accelerated dramatically in the last decades [86, 87]. Illumina short read sequencing technology is an NGS that consists of random fragmentation of the genome into small fragments, called reads, which are anchored to adapters at both ends of these to bind them on the surface of a plate called flowcell[84, 88]. The flowcell contains the complementary oligos to the adapter fixed on the surface to which the reads will be attached. After bonding the adapters to the surface, several amplification cycles are carried out to create identical clusters of the same sequence. Once these clusters have formed, sequencing begins with the help of fluorescence-labeled terminator nucleotides and DNA polymerase. After the addition of each nucleotide, a laser scans the plate to excite the fluorophores that will emit a specific light pulse for each nucleotide. The light emission is stored in the form of a photograph, the terminator is enzymatically removed and the cycle is repeated. The images taken after each cycle are converted into nucleotide sequences thanks to a "base-calling" software [88, 89].

Today this technique is so efficient that it allows the sequencing of multiple samples. To this end, in addition to the adapter, a specific barcode is added to all the reads of each sample. These barcodes are short sequences, usually 6-mers [85, 88], of nucleotides that allow to map back each read to the sample from which it was obtained. This method supposes a great advantage for the laboratories since it allows to sequence multiple samples to a very reduced and affordable price but it is a great challenge for the bioinformatics analysis since new algorithms capable of working with low coverage are needed.

Once the sequences of the reads have been obtained, they have to be aligned to a reference genome to locate their position in the genome. The reference genome is a representative sequence example of a species that is constructed by the sequencing of different individuals. Differences in nucleotides between the sample genotype and the reference genome are marked as a mismatch and form the allelic position marks. GBS uses this difference between the reference sequence and the sequence on the sample reads to reconstruct the complete genome [73].

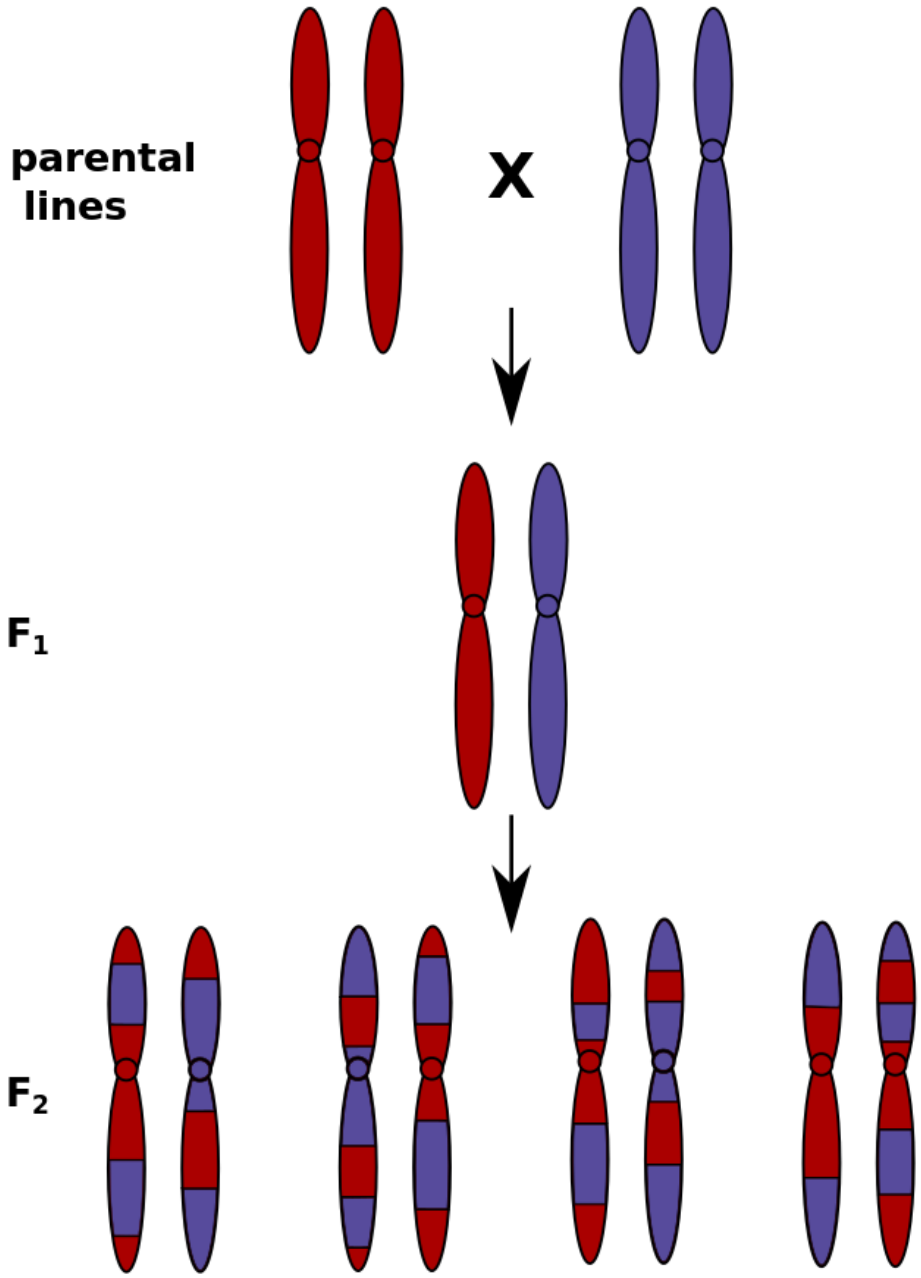


Figure 16: **Crossover events.** The founding generation consists of two homozygous, genetically diverse parental lines. Those are obtained by repeated inbreeding. After sexual reproduction between two homozygous organisms, the offspring F_1 has one copy from each parent. To produce gametes, during meiosis I, crossover events (CO) will occur among homologous chromosomes. After CO, new recombinant chromosomes are formed and transferred to the new generation F_2. The pair of chromosomes in the F_2 will have loci were both chromosomes have the same allelic information as one of the ancestral parental lines (homozygous) and locus were each chromosome has the allele information from a different ancestral line (heterozygous).

11.2 Genotyping Based on Sparse Sequencing Data

Multiplexed sequencing is a great method to sequence multiple samples in one batch in a cost-effective manner. However, it goes along with a lower read coverage compared with deep sequencing methods. However, the low number of reads will make it difficult to distinguish SNPs from sequencing errors and not all markers will be sequenced in different samples [81] (figure 17).

Low coverage is an obstacle when estimating the allelic frequency of SNPs. Besides that, the low number of reads in each allele can be responsible for high variability in the measurements ("shot noise") [90, 91]. Some other alleles might even be missed entirely. To overcome this bias, different bioinformatics methods have been developed [73, 85, 84, 79]. These methods take advantage of the non-random association of alleles in neighboring loci in recombinant genomes. This means that it is very likely that consecutive alleles in the sequence come from the same parent.

We have developed a method which integrates information from all sequenced samples in the study, thus avoiding the loss of alleles due to low sequencing and increasing the power to uncover mapping or other errors.

Here, we present a bioinformatics strategy to genotype many individuals with very low sequencing coverage. We present a pipeline that can genotype and detect CO positions with low coverage data. We apply the pipeline to investigate the role of RECQ4A, REQ4B and fidgetin-like 1 (FIGL1) proteins in CO during Meiosis I in *Arabidopsis thaliana*. RECQ4A is a helicase protein that acts upon DNA during replication, recombination, and repair. [92, 93] Loss of function in *Arabidopsis thaliana* has been proved to increase the number of CO events [94, 79]. FIGL1 is an AAA-ATPase protein that acts as a negative regulator of COs. Hence, loss of the function has also been shown to increase CO frequency up to 72% compared with wild type[95]. Our analysis confirms these results quantitatively.

12 A HMM for Genotyping by Sequencing

12.1 Model Statement

Hidden Markov models are widely used to assign ancestry of parental lines to chromosomal segments [73, 85]. The reason for using this statistical framework is that SNPs have a strong spatial dependency along the genome. We consider the situation in which each SNP locus can carry one of two alleles, originating either from parent one or parent two. Merely for ease of presentation, we will call the two parental lines "paternal" and "maternal", without assuming any specific mode of reproduction. At each genomic position, one then has to identify one out of three different haplotypes inside each chromosome: Homozygous paternal (p), homozygous maternal (m), or heterozygous (h). Additionally, we address the problem of "bad" SNP positions. Due to sequencing errors, alignment errors,

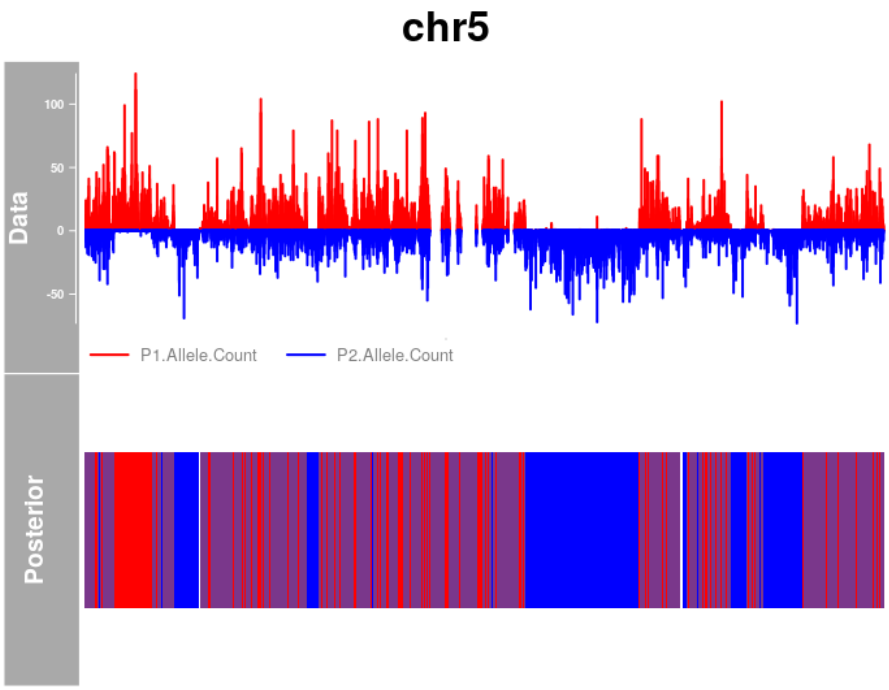


Figure 17: **Raw data and decoding by 3-betabinomial-mixture clustering.** SNPs are distributed randomly along the chromosomes. In re-sequencing experiments one can count the number of reads suporting the allele from parent one (red) or parent two (blue) at each SNP position. Upper panel: The number of counts for each parental line at each SNP position. Lower panel: A simple mixture of 3 beta-binomail distributions is fitted to the raw data to estimate which SNPs are homozygous in homologous chromosomes for parent one (red bars), for parent two (blue bars) or heterozygous (violet bars). If the spatial dependency structure of the genome (linkage) is neglected, the assigned homozygozity/heterozygozity calls are erratic and do not reveal the expected haplotypes.

parental allele bias etc., a certain SNP position may be covered by DNA sequences originating from a different region of the genome. This will lead to a distorted overrepresentation of one allele at this locus, across all samples. In our method, we use the information from the combined samples to infer the quality of each marker. The expected allele frequency in the pooled data should be around 0.5. If a marker has a ratio closer to one means that most of the samples have a higher allele frequency towards the parental line one and this marker should be considered a bad marker (b_p). If on the contrary, it is closer to 0, the marker shows a higher frequency towards the other parental line. Hence, it should be called a bad marker as well (b_m).

The specific hidden Markov Model presented here was developed by professor Achim Tresch. Given a set of T conesecutive markers, which were measured in J different samples, a single observation $O_t^j = (k_t^j, n_t^j)$ consists of the number n_t^j of reads that were mapped to position t in sample j, and the number k_t^j of which that mapped to the allele from the paternal line. The complete data is $O = (O_t^j; t = 1, ..., T, j = 1, ..., J)$. The variable $M_t \in \{+, b_m, b_p\}$ tells whether the marker at position t is a good marker (+), or whether it is a bad marker based on the observations in all the samples (b_p most of the samples contain the allele belonging to the parental line one / b_m most of the samples contain the allele belonging to the parental line two). The memory variables $V_t^j \in \mathcal{S} = \{m, p, h\}$, record the marker state of the most recent *good* marker position s, $s = \max\{s' \leq t, M_{s'} = +\}$. Here, p, m, and h denote respectively parental line one, parental line two and heterozygous states. We will adopt the convention that $O^j = (O_t^j, t = 1, ..., T)$, and $V^j = (V_t^j, t = 1, ..., T)$, $j = 1, ..., J$. According to its graphical representation (figure 18), the model factors into

$$P(O, V, M) = P(M) \cdot \prod_{j=1}^{J} P(V^j \mid M) \cdot \prod_{j=1}^{J} P(O^j \mid V^j, M) \tag{37}$$

An obsevation o_t^j can be drawn from one of five distributions $\Psi = (\psi_e, e \in \mathcal{E} = \{m, p, h, b_m, b_p\})$. In addition to the p(parental line one), m(parental line two), and h(eterozygous) distributions, b_m and b_p denote the distributions for bad markers. The index e of the emission distribution ψ_e from which the observations at position t in sample j were drawn is a function of M_t and V_t^j,

$$e(V_t^j, M_t) = \begin{cases} V_t^j & \text{if } M_t = + \\ M_t & \text{if } M_t \neq + \end{cases} \tag{38}$$

The model is fully specified by

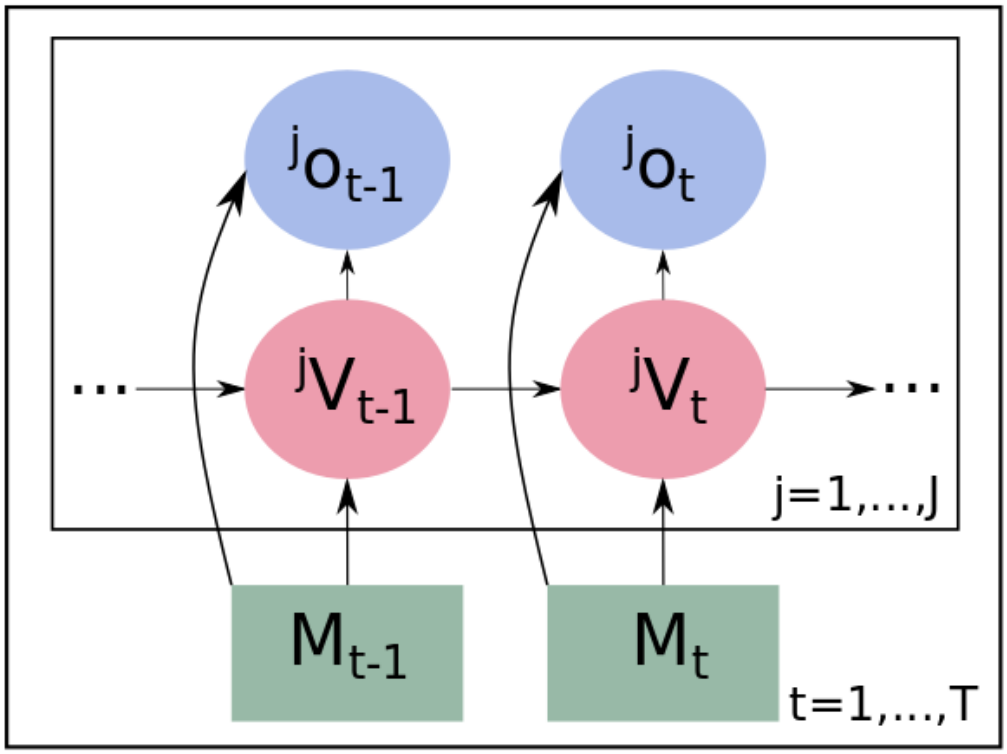


Figure 18: **Graphical representation of the HMM model.** Our statistical framework for analyzing mapping recombinant population is based on a Hidden Markov model. The allele count of each sample at every position t is assumed to be produced by an underlying Markov chain of three states $V_t \in \{m, p, h\}$. Furthermore, we define a new hidden layer of $M \in \{+, b_m, b_p\}$ which is the hidden layer given the observations of all the samples J. This new layer discerns each position as a good $(+)$ or bad marker(b_m, b_p). The bad marker b_p (higher than expected) and b_m (lower than expected) are defined by the observed frequencies of the allele counts.

$$P(M) = \prod_{t=1}^{T} \pi^M_{M_t} \tag{39}$$

$$P(V^j \mid M) = \pi^V_{V^j_1} \cdot \prod_{t=2}^{T} P(V^j_t \mid V^j_{t-1}, M_t) \tag{40}$$

$$P(V^j_t \mid V^j_{t-1}, M_t) = \begin{cases} a_{V^j_{t-1}V^j_t} & \text{if } M_t = + \\ \delta(V^j_t = V^j_{t-1}) & \text{if } M_t \neq + \end{cases}, \quad t = 2, ..., T \tag{41}$$

$$P(O^j \mid V^j, M) = \prod_{t=1}^{T} \psi_{e(V^j_t, M_t)}(O^j_t) \tag{42}$$

Its parameters are $\Theta = (\pi^M, \pi^V, A = (a_{rs})_{r,s\in\mathcal{S}}, \Psi = (\psi_s)_{s\in\mathcal{E}})$. Here, π^M is the prior for the markers, π^V are the initial probabilities for the hidden states, A is a transition matrix (i.e., it has non-negative entries, and its row sums are 1), and Ψ are the state-specific emission probabilities.

12.2 The Forward-Backward Algorithm

For the moment, assume Θ to be known. For each sample j, the variables (O^j, V^j) form an HMM with transition probabilities $B = (b_{rs})_{r,s\in\mathcal{S}}$, initial state probabilities π^V, and emission probabilities $\Phi = (\phi_s)_{s\in\mathcal{S}}$. Here,

$$\begin{aligned} b_{rs} &= P(V^j_t = s \mid V^j_{t-1} = r) \\ &= \sum_{M_t} P(V^j_t = s, M_t \mid V^j_{t-1} = r) \\ &= \sum_{M_t} P(V^j_t = S \mid V^j_{t-1} = r, M_t)\pi^M_{M_t} \\ &= (\pi^M_{b_p} + \pi^M_{b_m})\delta(r = s) + \pi^M_+ a_{rs} \end{aligned} \tag{43}$$

and hence $B = \pi^M_+ A + (\pi^M_{b_p} + \pi^M_{b_m})E$. Further,

$$\begin{aligned} \phi_s(O^j_t) &:= P(O^j_t \mid V^j_t = s) \\ &= \sum_{M_t} \pi^M_{M_t} P(O^j_t \mid V^j_t = s, M_t) \\ &= \pi^M_{b_p}\psi_{b_p}(O^j_t) + \pi^M_{b_m}\psi_{b_m}(O^j_t) + \pi^M_+\psi_s(O^j_t) \end{aligned} \tag{44}$$

and hence $\phi_s = \pi^M_+\psi_s + \pi^M_{b_p}\psi_{b_p} + \pi^M_{b_m}\psi_{b_m}$.

We apply the standard forward-backward algorithm with the parameter set (B, π^V, Φ) in order to calculate the sample-specific forward- and backward probabilities

$$\alpha_t^j(V_t^j) := P(V_t^j, O_1^j, ..., O_t^j) \ , \quad t = 1, ..., TThe \tag{45}$$
$$\beta_t^j(V_t^j) := P(O_{t+1}^j, ..., O_T^j \mid V_t^j) \ , \quad t = 1, ..., T-1; \quad \beta_T(V_T^j) = 1 \tag{46}$$

12.3 Bad Marker Detection

The final goal of this new model is to find the bad markers. To do so we calculate the posterior probabilites for the marker states, $\mu_t(s) := P(M_t = s \mid O)$. To this end, we will define some auxiliary quantities:

$$P(V_1^j, O_1^j \mid M_1) = \psi_{e(V_1^j, M_1)}(O_1)\pi^V_{V_1^j} \tag{47}$$

$$\begin{aligned} P(V_t^j, , O_1^j, ..., O_t^j \mid M_t) &= \sum_{V_{t-1}^j} P(V_{t-1}^j, V_t^j, O_1^j, ..., O_t^j \mid M_t) \quad , \ t = 2, ..., T \\ &= \sum_{V_{t-1}^j} P(O_t^j, V_t^j \mid V_{t-1}^j, M_t) P(V_{t-1}^j, O_1^j, ..., O_{t-1}^j) \\ &= \psi_{e(V_t^j, M_t)}(O_t^j) \sum_{V_{t-1}^j} P(V_t^j \mid V_{t-1}, M_t) \alpha_{t-1}^j(V_{t-1}^j) \\ &= \psi_{e(V_t^j, M_t)}(O_t^j) \begin{cases} \sum_{V_{t-1}^j} a_{V_{t-1}^j V_t^j} \alpha_{t-1}^j(V_{t-1}^j) & \text{if } M_t = + \\ \alpha_{t-1}^j(V_t^j) & \text{if } M_t \in \{b_p, b_m\} \end{cases} \end{aligned} \tag{48}$$

With these auxilar terms now we can calculate

$$\begin{aligned} P(O^j \mid M_t) &= \sum_{V_t^j} P(V_t^j, O^j \mid M_t) \\ &= \sum_{V_t^j} P(V_t^j, O_1^j, ..., O_t^j \mid M_t) P(O_{t+1}^j, ..., O_T^j \mid V_t^j) \\ &= \sum_{V_t^j} P(V_t^j, O_1^j, ..., O_t^j \mid M_t) \beta_t^j(V_t^j) \\ &\overset{(47,48)}{=} \sum_{V_t^j} \beta_t^j(V_t^j) \psi_{e(V_t^j, M_t)}(O_t^j) \cdot \begin{cases} \pi^V_{V_t^j} & \text{if } t = 1 \\ \sum_{V_{t-1}^j} a_{V_{t-1}^j V_t^j} \alpha_{t-1}^j(V_{t-1}^j) & \text{if } t > 1, M_t = + \\ \alpha_{t-1}^j(V_t^j) & \text{if } t > 1, M_t \in \{b_p, b_m\} \end{cases} \end{aligned} \tag{49}$$

In the next line, we use a naive Bayes approximation:

$$P(M_t, O) = \pi^M_{M_T} \cdot P(O \mid M_t) \approx \pi^M_{M_T} \cdot \prod_j P(O^j \mid M_t) \tag{50}$$

From equation (50), we finally obtain

$$\begin{aligned}\mu_t(s) &= P(M_t = s \mid O) \qquad (51)\\ &= P(M_t = s, O)/P(O) = P(M_T = s, O)/\sum_{s' \in \{+, b_m, b_p\}} P(M_t = s', O)\end{aligned}$$

12.4 Marginalization with respect to Hidden Genetic States

Following, we can compute the posterior probabilities for the hidden states, $\gamma_t^j(s) := P(V_t^j = s \mid O)$. For $t = 1$, this is obtained from

$$\begin{aligned}P(V_1^j, M_1, O^j) &= P(V_1^j, M_1, O_1^j) \cdot P(O_2^j, ..., O_T^j \mid V_1^j) \qquad (52)\\ &\overset{(47)}{=} \psi_{e(V_1^j, M_1)}(O_1^j)\pi_{V_1^j}^V \pi_{M_1}^M \cdot \beta_1^j(V_1^j)\end{aligned}$$

$$P(M_1^j, O^j) = \sum_{V_1^j} P(V_1^j, M_1, O^j) = \pi_{M_1}^M \sum_{V_1^j} \psi_{e(V_1^j, M_1)}(O_1^j)\pi_{V_1^j}^V \beta_1^j(V_1^j) \qquad (53)$$

$$\begin{aligned}\gamma_1(s) &= P(V_1^j = s \mid O) = \sum_{M_1} P(V_1^j = s, M_1 \mid O) \qquad (54)\\ &= \sum_{M_1} P(V_1^j = s \mid M_1, O) P(M_1 \mid O)\\ &= \sum_{M_1} P(V_1^j = s \mid M_1, O^j) \mu_1(M_1)\\ &= \sum_{M_1} \frac{P(V_1^j = s, M_1, O^j)}{P(M_1, O^j)} \mu_1(M_1) \qquad (55)\\ &= \sum_{M_1} \frac{\psi_{e(s, M_1)}(O_1^j)\pi_s^V \beta_1^j(s)}{\sum_r \psi_{e(r, M_1)}(O_1^j)\pi_r^V \beta_1^j(r)} \mu_1(M_1)\end{aligned}$$

For $t > 1$, $\gamma_t(s)$ can be calculated from

$$P(V_t^j, M_t, O^j) = P(V_t^j, M_t, O_1^j, ...O_t^j) \cdot P(O_{t+1}^j, ..., O_T^j \mid V_t^j)$$

$$\stackrel{(47)}{=} \psi_{e(V_1^j,M_1)}(O_1^j)\pi^V_{V_1^j}\pi^M_{M_1} \cdot \beta_t^j(V_t^j) \tag{56}$$

$$P(M_1^j, O^j) = \sum_{V_1^j} P(V_1^j, M_1, O^j) = \pi^M_{M_1} \sum_{V_1^j} \psi_{e(V_1^j,M_1)}(O_1^j)\pi^V_{V_1^j}\beta_1^j(V_1^j) \tag{57}$$

$$\begin{aligned}\gamma_t(s) = P(V_t^j = s \mid O) &= \sum_{M_t} P(V_t^j = s, M_t \mid O) = \sum_{M_t} P(V_t^j = s \mid M_t, O)P(M_t \mid O) \qquad (58)\\ &= \sum_{M_t} P(V_t^j = s \mid M_t, O^j)\mu_t(M_t) = \sum_{M_t} \frac{P(V_t^j = s, M_t, O^j)}{P(M_1, O^j)}\mu_t(M_t)\\ &= \sum_{M_1} \frac{\psi_{e(s,M_1)}(O_1^j)\pi_s^V\beta_1^j(s)}{\sum_r \psi_{e(r,M_1)}(O_1^j)\pi_r^V\beta_1^j(r)}\mu_1(M_1)\end{aligned}$$

Note that $\gamma_t^j(s) = P(V_t^j = s \mid O)$ differs from the conventional posterior probability $P(V_t^j = s \mid O^j) = \frac{\alpha_t^j(s)\beta_t^j(s)}{\sum_r \alpha_t^j(r)\beta_t^j(r)}$. Nevertheless, this difference is small though for positions t with "good" markers, for which $\mu_t(M_t = +) \approx 1$.

After calculating the posterior probabilities for the markers, we remove all markers for which $P(M_t = + \mid O) < c$ for some threshold c, which we set to $c = 0.99$ (it is more relevant not to include bad markers than to include all good markers). Based on the remaining good marker positions, we then calculate a (robust) Viterbi path for each sample j using a standard HMM with the parameters (π^V, A, Ψ).

12.5 Parameter Learning

Parameter estimation for this model can be carried out using the Baum-Welch algorithm, the EM algorithm [19]. Given a previous parameter guess Θ', we have to optimize a target function $Q(\Theta; \Theta')$ with respect to Θ, replace Θ' by this Θ, and iterate until convergence to a local minimum. The target function is

$$\begin{aligned}Q(\Theta; \Theta') &= \mathbb{E}_{P(V,M|O;\Theta')} \log P(O, V, M; \Theta) \qquad (59)\\ &= \sum_{V,M} P(M, V \mid O; \Theta') \log P(O, V, M; \Theta)\\ &= \sum_{V,M} P(V, M \mid O; \Theta') \cdot \left[\log P(O \mid V, M; \Psi) + \log P(V \mid M; A, \pi^V) + \log P(M; \pi^M)\right]\end{aligned}$$

12.5.1 Update of the Transition Probability

Let us start with the transition probability matrix A.

$$
\begin{aligned}
\frac{\partial}{\partial a_{rs}} Q(\Theta;\Theta') &= \sum_{V,M} P(V,M \mid O;\Theta') \frac{\partial}{\partial a_{rs}} \log P(M,V,H,O;\Theta) \qquad (60)\\
&= \sum_{V,M} P(V,M \mid O;\Theta') \frac{\partial}{\partial a_{rs}} \log P(V \mid M;A,\pi^V)\\
&= \sum_{V,M} P(V,M \mid O;\Theta') \left(\sum_j \sum_{t=2}^{T} \frac{\partial}{\partial a_{rs}} \log P(V_t^j \mid V_{t-1}^j, M_t;A) \right)\\
&= \sum_j \sum_{t=2}^{T} \left(\sum_{V,M} P(V,M \mid O;\Theta') \frac{\partial}{\partial a_{rs}} \log P(V_t^j \mid V_{t-1}^j, M_t;A) \right)\\
&\overset{(41)}{=} \sum_j \sum_{t=2}^{T} \left(\sum_{V_{t-1}^j, V_t^j} P(M_t=+, V_{t-1}^j, V_t^j \mid O;\Theta') \frac{\partial}{\partial a_{rs}} \log a_{V_{t-1}^j V_t^j} \right)\\
&= \frac{1}{a_{rs}} \sum_j \sum_{t=2}^{T} P(M_t=+ \mid O;\Theta') \cdot \underbrace{P(V_{t-1}^j = r, V_t^j = s \mid M_t=+, O;\Theta')}_{=:\zeta_t^j(r,s)}\\
&= \frac{1}{a_{rs}} \sum_j \sum_{t=2}^{T} \mu_t(+) \zeta_t^j(r,s)
\end{aligned}
$$

We point out that the terms $\mu_t(+)$ and $\zeta_t^j(r,s)$ are both calculated with respect to the known parameter set Θ'. The terms $\zeta_t^j(r,s)$ can be obtained from the forward/backward probabilities via

$$
\begin{aligned}
\zeta_t^j(r,s) &= P(V_{t-1}^j = r, V_t^j = s \mid M_t=+, O) \qquad (61)\\
&= P(V_{t-1}^j = r, V_t^j = s \mid M_t=+, O^j)\\
&\propto P(V_{t-1}^j = r, V_t^j = s, M_t=+, O^j)\\
&= P(O_1^j,...,O_{t-1}^j, V_{t-1}^j = r) P(V_t^j = s \mid V_{t-1}^j = r, M_t=+)\\
&\quad \cdot P(O_t^j \mid V_t^j = s, M_t=+) P(O_{t+1}^j,...,O_T^j \mid V_t^j = s)\\
&= \alpha_{t-1}^j(r) \cdot a_{rs} \cdot \psi_s(O_t^j) \cdot \beta_t^j(s)
\end{aligned}
$$

Again, we emphasize that the forward and backward probabilities $\alpha_{t-1}^j(r)$ and $\beta_t^j(s)$ are calculated with respect to the known parameter set Θ'. After calculating the terms in (61) for all r,s, we exploit that $\sum_{r',s'} \zeta_t^j(r',s') = 1$ in order to find the missing normalization factor:

$$
\zeta_t^j(r,s) = \frac{\alpha_{t-1}^j(r) \cdot a_{rs} \cdot \psi_s(O_t^j) \cdot \beta_t^j(s)}{\sum_{r',s'} \alpha_{t-1}^j(r') \cdot a_{r's'} \cdot \psi_{s'}(O_t^j) \cdot \beta_t^j(s')} \qquad (62)
$$

Introducing Lagrange multipliers λ_r, which account for the contraints $\sum_s a_{rs} = 1$, $r \in \{m, p, h\}$, and solving for $\frac{\partial}{\partial a_{rs}} Q(\Theta; \Theta') = 0$ in equation (60) yields

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_{rs}} Q(\Theta; \Theta') + \frac{\partial}{\partial a_{rs}} \sum_r \lambda (1 - \sum_s a_{rs}) \\ 0 &= \frac{1}{a_{rs}} \sum \sum_{t=2}^{T} \mu_t(+) \zeta_t^j(r, s) - \lambda \\ a_{rs}\lambda &= \sum_j \sum_{t=2}^{T} \mu_t(+) \zeta_t^j(r, s) \end{aligned} \tag{63}$$

We can solve λ by summing over s:

$$\lambda = \sum_s a_{rs}\lambda = \sum_s \sum_j \sum_{t=2}^{T} \mu_t(+) \zeta_t^j(r, s) \tag{64}$$

and thus, we can substitute λ and obtaine the estimate for a_{rs}:

$$a_{rs} = \frac{\sum_{t=2}^{T} \mu_t(+) \left(\sum_j \zeta_t^j(r, s) \right)}{\sum_s \sum_{t=2}^{T} \mu_t(+) \left(\sum_j \zeta_t^j(r, s) \right)} \tag{65}$$

12.5.2 Parameter Estimation of the Beta-Binomial Distribution

Allele counts in each marker can be taken in account as a binomial trial if the allele count of one of the parental lines is interpreted as a succes. Hence, one could assume that the distribution in each marker follows a binomial distribution with probability $p_e, e \in \mathcal{E} = \{m, p, h, b_m, b_p\}$, of success. Fitting a binomial distribution have undesired consequences. For instance, the parameter p_e provides a rigidity to the model that makes it unable to capture the overdispersion created during sequencing methods [96]. For that reason, a beta-binomial distribution is more appropriate to capture the overdispersion and allow a more flexible estimation of the states [97]. It can be assumed that the probability of success p_e is randomly drawn from a beta distribution $B(\alpha_e, \beta_e)$, $e \in \mathcal{E} = \{m, p, h, b_m, b_p\}$.

Therefore, we learn the parameter α_e and β_e for the beta-binomial emission probabilities ψ_e for each state s. We assume that $\psi_e(k; n, \alpha_e, \beta_e)$ is a Beta-Binomial distribution with parameters (α_e, β_e) (the parameter n is determined by the observations). Let $O_t^j = (k_t^j, n_t^j)$, where n_t^j is the number of mapped reads at position t in sample j, and k_t^j is the number of those reads that support the alleles from one of the parental genome. In the following, "$\propto$ " means that we left out additive or multiplicative terms that are irrelevant for the optimization with respect to (α_e, β_e).

$$
\begin{aligned}
Q(\Theta;\Theta') &= \sum_{V,M} P(V,M \mid O;\Theta') \log P(O,V,M;\Theta) \qquad (66)\\
&\propto \sum_{V,M} P(V,M \mid O;\Theta') \log P(O \mid V,M;\Psi)\\
&= \sum_{V,M} P(V,M \mid O;\Theta') \left(\sum_j \sum_{t=1}^{T} \log \psi_{e(V_t^j,M_t)}(k_t^j; n_t^j, \alpha_{H_t^j}, \beta_{H_t^j}) \right)\\
&= \sum_j \sum_{t=1}^{T} \left(\sum_{V_t^j, M_t} P(V_t^j, M_t \mid O) \log \psi_{e(V_t^j,M_t)}(k_t^j; n_t^j, \alpha_s, \beta_s) \right)\\
&\propto \sum_j \sum_{t=2}^{T} \underbrace{\sum_{V_t^j, M_t; e(V_t^j,M_t)=s} P(V_t^j, M_t \mid O)}_{=:\tau_t^j(s)} \cdot \log \psi_s(k_t^j; n_t^j, \alpha_s, \beta_s)
\end{aligned}
$$

where the quantity $\tau_t^j(s)$ is given as

$$
\begin{aligned}
\tau_t^j(s) &= P(M_t = s \mid O) = \mu_t(s) \quad , \text{ if } s \in \{b_p, b_m\}, t = 1, ..., T \qquad (67)\\
\tau_t^j(s) &= P(V_t^j = s, M_t = + \mid O) \quad , \text{ if } s \in \{m, p, h\}, t = 2, ..., T \qquad (68)\\
&= P(V_t^j = s \mid M_t = +, O) P(M_t = + \mid O)\\
&= \sum_r P(V_{t-1}^j = r, V_t^j = s \mid M_t = +, O) \mu_t(+) = \mu_t(+) \sum_r \zeta_t^j(r,s)
\end{aligned}
$$

In order to obtain $\tau_1(s)$, $s \in \{m, p, h\}$, we first calculate

$$
\begin{aligned}
P(V_1^j = s, M_1 = +, O^j) &= P(V_1^j = s, M_1 = +, O_1^j) \cdot P(O_2^j, ..., O_T^j \mid V_1^j) \qquad (69)\\
&\overset{(47)}{=} \psi_s(O_1^j) \pi_s^V \pi_+^M \cdot \beta_1^j(s)\\
P(M_1 = +, O^j) &= \sum_{s' \in \{m,p,h\}} P(V_1^j = s, M_1 = +, O^j) \qquad (70)\\
&= \sum_{s' \in \{m,p,h\}} \psi_{s'}(O_1^j) \pi_{s'}^V \pi_+^M \cdot \beta_1^j(s')
\end{aligned}
$$

This yields

$$
\begin{aligned}
\tau_1^j(s) &= P(V_1^j = s, M_1 = + \mid O) = P(M_1 = + \mid O) \cdot P(V_1^j = s \mid M_1 = +, O) \qquad (71)\\
&= \mu_1(+) \cdot P(V_1^j = s \mid M_1 = +, O^j) = \mu_1(+) \cdot P(V_1^j = s, M_1 = +, O^j) / P(M_1 = +, O^j)\\
&= \mu_1(+) \cdot \frac{\psi_s(O_1^j) \pi_s^V \beta_1^j(s)}{\sum_{s' \in \{m,p,h\}} \psi_{s'}(O_1^j) \pi_{s'}^V \beta_1^j(s')}
\end{aligned}
$$

The optimization of (66) is analytically intractable. We suggest to use a method of moments: the expectation of ψ_s according to the τ_t^j-weighted empirical mean of the observations,

$$\frac{\alpha_s}{\alpha_s + \beta_s} = \frac{1}{n}\mathbb{E}\left(\psi_s(.; n, \alpha_s, \beta_s)\right) = \frac{\sum_{j,t} k_t^j \tau_t^j(s)}{\sum_{j,t} n_t^j \tau_t^j(s)} \tag{72}$$

and then perform a line search along $\alpha_s + \beta_s$.

12.5.3 Initial Probabilitites π^V and Marker Frequencies π^M

We can use a similar strategy as in the previous section to learn π^M and π^V.

$$\begin{aligned}
\frac{\partial}{\partial \pi_s^M} Q(\Theta; \Theta') &= \sum_{V,M} P(V, M \mid O; \Theta') \frac{\partial}{\partial \pi_s^M} \log P(O, V, M; \Theta) - \lambda \frac{\partial (1 - \sum_S \pi_S^M)}{\partial \pi_S^M} \qquad (73) \\
&\propto \sum_{V,M} P(V, M \mid O; \Theta') \frac{\partial}{\partial \pi_s^M} \log P(M; \pi^M) - \lambda \frac{\partial (1 - \sum_S \pi_S^M)}{\partial \pi_S^M} \\
&= \sum_{M} P(M \mid O; \Theta') \left(\sum_{t=1}^{T} \frac{\partial}{\partial \pi_s^M} \log \pi_{M_t}^M \right) - \lambda \frac{\partial (1 - \sum_S \pi_S^M)}{\partial \pi_S^M} \\
&= \sum_{t=1}^{T} \sum_{M_t} P(, M_t \mid O) \frac{\partial}{\partial \pi_s^M} \log \pi_{M_t}^M - \lambda_S \\
&= \frac{1}{\pi_s^M} \sum_{t=1}^{T} \mu_t(s) - \lambda
\end{aligned}$$

Using the same strategy as in equation (64):

$$\lambda = \sum_{m' \in \{+, b_m, b_p\}} \lambda \pi_S^M = \sum_{m' \in \{+, b_m, b_p\}} \sum_{t=1}^{T} \mu_t(s) \tag{74}$$

$$\frac{\partial}{\partial \pi_s^M} Q(\Theta; \Theta') = \frac{1}{\pi_s^M} \sum_{t=1}^{T} \mu_t(s) - \sum_{m' \in \{+, b_m, b_p\}} \sum_{t=1}^{T} \mu_t(s) = 0 \tag{75}$$

$$\pi_s^M = \frac{\sum_t \mu_t(s)}{\sum_{m' \in \{+, b_m, b_p\}} \sum_t \mu_t(m')} = \frac{\sum_t \mu_t(s)}{T} \tag{76}$$

Similar calculations for π^V lead to

$$\begin{aligned}
\frac{\partial}{\partial \pi_s^V} Q(\Theta;\Theta') &= \sum_{V,M} P(V,M \mid O;\Theta') \frac{\partial}{\partial \pi_s^V} \log P(O,V,M;\Theta) \qquad (77)\\
&\propto \sum_{V,M} P(V,M \mid O;\Theta') \frac{\partial}{\partial \pi_s^V} \log \prod_j P(V^j \mid M;\pi^M)\\
&\propto \sum_{V} P(V \mid O;\Theta') \left(\sum_j \frac{\partial}{\partial \pi_s^V} \log \pi^V_{V_1^j} \right)\\
&= \sum_j P(V_1^j = s \mid O;\Theta')/\pi_s^V\\
&= \frac{1}{\pi_s^V} \sum_{j=1}^{J} \nu_1^j(s)
\end{aligned}$$

Using the Lagrange multiplier $\lambda(1 - \sum_s \pi_s^V)$, and setting the above derivatives to zero yields

$$\pi_s^V = \frac{\sum_j \nu_1^j(s)}{\sum_{s'} \sum_j \nu_1^j(s')} = \frac{\sum_j \nu_1^j(s)}{J} \qquad (78)$$

Generally, it needs to be taken into account that each sample consists of several parts (e.g., chromosomes), which have to be treated as independent observations respectively Markov chains. One has to introduce another index to the formulas above, but this leaves the above formulas essentially unchanged (mostly, the summation over the samples j has to be replaced by summation over j and the parts).

13 Localization of Cross Over Events

The crossover points represent a change in the ancestral genome of the analyzed sample. Being able to locate and estimate them with some precision is very important and challenging for genotyping methods. In HMMs, Viterbi path will provide a good inference method to call genomic breakpoints. However, the very location where a state change of the Viterbi sequence takes place is itself a random variable and cannot be safely limited to one single SNP position. In order to address this problem, we define "interval transitions" and "wiggly interval transitions", which are events describing a transition from one state to another within a given interval. We calculate the probability of these events, and we derive an efficient method to screen for the occurrence of these events along the whole genome. Our calculations apply to any HMM and therefore serve the general purpose of localizing state transitions in HMMs. Our method consist of three steps:

1. Screening for valid interval transitions: Before computing the score for a transition in every position, one must do a screening along the genome in

order to find possible transitions. Our method uses the posterior probability $\gamma_t(s)$ to identify potential locations where transition may arise.

2. Calculation of the probability of an interval transition / wiggly interval transition. Given a list of candidates for transitions, we apply our score method in order to give a numerical value to the probability of a transition occurring.

3. Filtering non-valid transitions. We remove all transitions with a score lower than a threshold c. If a transition is predicted inside a specific interval, but, there is no change of states in the Viterbi chain, this interval transition will be discarded as well.

13.1 Probability of an Interval Transition

We say that the hidden state chain $s = (s_1, s_2, ..., s_T)$, $s_t \in \{1, ..., S\}$, transitions from state j to state k within the interval $[a, b]$, $a, b \in \{1, ..., T\}$, $a < b$, if

$$s_t = f_m(t) := \begin{cases} j & \text{for } a \leq t \leq m \\ k & \text{for } m < t \leq b \end{cases} \tag{79}$$

for some $m \in [a, ..., b-1]$. Let $T_{jk}[a, b]$ denote the corresponding event, which we call interval transition. Assume that the hidden state sequence belongs to a hidden Markov model with parameters $\Theta = (\pi, A, \Psi)$. Here, $\pi = (\pi_s)$ is the vector of initiation probabilities, $A = (a_{jk})$ is the transition matrix with $a_{jk} = P(s_{t+1} = k \mid s_t = j)$, and $\Psi = (\psi_s)$ are the emission distributions. We keep these parameters fixed and omit their explicit mention in the following. Let $\mathcal{O} = (o_1, ..., o_T)$ be a sequence of observations generated by this HMM. We are interested in calculating the posterior probability

$$P(T_{jk}[a, b] \mid \mathcal{O}) = P(T_{jk}[a, b], \mathcal{O})/P(\mathcal{O}) \tag{80}$$

Let $\alpha_t(s) = P(s_t = s, o_1, ..., o_t)$, $t = 1, ..., T$, respectively $\beta_t(s) = P(o_{t+1}, ..., o_T \mid s_t = s)$, $t = 1, ..., T$-1, $s = 1, ..., S$, denote the well-known forward and backward probabilities, which can be calculated efficiently using the forward-backward algorithm. For convenience, let $\beta_T(s) = 1$, $s = 1, ..., S$. The denominator in (80) can be obtained easily, given the forward/backward probabilities. For any $t = 1, ..., T - 1$,

$$
\begin{aligned}
P(\mathcal{O}) &= \sum_{s=1}^{S} P(\mathcal{O}, s_t) \\
&= \sum_{s_t=1}^{S} P(s_t = s, o_1, ..., o_t) \cdot P(o_{t+1}, ..., o_T \mid s_t = s) \qquad (81) \\
&= \sum_{s=1}^{S} \alpha_t(s)\beta_t(s)
\end{aligned}
$$

Note that this formula also holds for $t = T$. Following, the joint probability can be calculated

$$
\begin{aligned}
P(s_{a+1}, ..., s_b, o_{a+1}, ..., o_b \mid s_a) &= \prod_{t=a}^{b-1} P(s_{t+1} \mid s_t) P(o_{t+1} \mid s_{t+1}) \qquad (82) \\
&= \prod_{t=a}^{b-1} a_{s_t s_{t+1}} \cdot \psi_{s_{t+1}}(o_{t+1})
\end{aligned}
$$

$$
\begin{aligned}
P(s_a = j, s_{a+1}, ..., s_b = k, \mathcal{O}) &= P(s_{a+1}, ..., s_b = k, o_{a+1}, ..., o_b \mid s_a = j) \qquad (83) \\
&\quad \cdot \underbrace{P(s_a = j, o_1, ..., o_a)}_{=\alpha_a(j)} \\
&\quad \underbrace{\cdot P(o_{b+1}, ..., o_T \mid s_b = k)}_{=\beta_b(k)} \\
&\overset{(82)}{=} \alpha_a(j) \cdot \beta_b(k) \cdot \prod_{t=a}^{b-1} a_{s_t s_{t+1}} \cdot \psi_{s_{t+1}}(o_{t+1})
\end{aligned}
$$

This allows us to calculate

$$
\begin{aligned}
P(T_{jk}[a, b], \mathcal{O}) &= \sum_{m=a}^{b-1} P(s_t = f_m(t),\, t = a, ..., b, \mathcal{O}) \qquad (84) \\
&\overset{(83)}{=} \alpha_a(j)\beta_b(k) \cdot \sum_{m=a}^{b-1} P(s_t = f_m(t), t = a+1, ..., b, o_{a+1}, ..., o_b \mid s_a = j) \\
&= \alpha_a(j)\beta_b(k) \cdot \sum_{m=a}^{b-1} \prod_{t=a}^{b-1} a_{f_m(t) f_m(t+1)} \cdot \psi_{f_m(t+1)}(o_{t+1})
\end{aligned}
$$

13.2 Numerically Stable Calculation of $P(T_{jk}[a, b] \mid \mathcal{O})$

If $a \ll b$, equation (84) contains large products, which may become very small and prone to numerical underflow. In order to avoid numerical instability, we make use of the normalized forward probabilities, which are calculated via the recursion

$$\tilde{\alpha}_1(s) := \pi_s \psi_s(o_1)/N_1 \ , \quad N_1 = \sum_{r=1}^{S} \pi_s \psi_s(o_1) \tag{85}$$

and for $t = 1, ..., T-1$:

$$\tilde{\alpha}_{t+1}(s) := \sum_{r=1}^{S} \tilde{\alpha}_t(r) \cdot a_{rs} \cdot \psi_s(o_{t+1})/N_{t+1} \ , \quad N_{t+1} = \sum_{r=1}^{S} \tilde{\alpha}_t(r) \cdot a_{rs} \cdot \psi_s(o_{t+1}) \tag{86}$$

Note that the original forward probabilities can be obtained from their normalized counterparts via

$$\alpha_1(s) := \pi_s \psi(o_1) = N_1 \tilde{\alpha}_1(s) \tag{87}$$

$$\begin{aligned} \alpha_{t+1}(s) &:= \sum_{r=1}^{S} \alpha_t(r) \cdot a_{rs} \cdot \psi_s(o_{t+1}) \\ &= \left(\prod_{u=1}^{t} N_u \right) \sum_{r=1}^{S} \tilde{\alpha}_t(r) \cdot a_{rs} \cdot \psi_s(o_{t+1}) \\ &= \left(\prod_{u=1}^{t+1} N_u \right) \tilde{\alpha}_{t+1}(s) \end{aligned} \tag{88}$$

As opposed to the calculation of the $\alpha_t(s)$, the calculation of their normalized counterparts is numerically stable. Let $\tilde{\beta}_t(s)$ be the normalized backward probabilities, obtained in an analogous way. In particular, $\frac{\beta_{b+1}(k)}{\beta_{b+1}(s)} = \frac{\tilde{\beta}_{b+1}(k)}{\tilde{\beta}_{b+1}(s)}$ holds. We can therefore write

$$\begin{aligned} P(T_{jk}[a,b] \mid \mathcal{O}) &= \frac{P(T_{jk}[a,b], \mathcal{O})}{P(\mathcal{O})} \\ &= \frac{\left(\prod_{u=1}^{a} N_u\right) \tilde{\alpha}_a(j) \beta_b(k) \cdot \sum_{m=a}^{b-1} \prod_{t=a}^{b-1} \left(a_{f_m(t) f_m(t+1)} \cdot \psi_{f_m(t+1)}(o_{t+1}) \right)}{\left(\prod_{u=1}^{b} N_u \right) \cdot \sum_{s=1}^{S} \tilde{\alpha}_b(s) \beta_b(s)} \\ &= \frac{\tilde{\alpha}_a(j) \tilde{\beta}_b(k) \cdot \sum_{m=a}^{b-1} \prod_{t=a}^{b-1} \left(a_{f_m(t) f_m(t+1)} \cdot \psi_{f_m(t+1)}(o_{t+1})/N_{t+1} \right)}{\sum_{s=1}^{S} \tilde{\alpha}_b(s) \tilde{\beta}_b(s)} \end{aligned} \tag{89}$$

Be aware that the product in the nominator will not suffer from numerical underflow in case the transition from j to k in $[a,b]$ is a likely event.

13.3 Probability of a Wiggly Interval Transition

We say that the hidden state chain $s = (s_1, s_2, ..., s_T)$, $s_t \in \{1, ..., S\}$, transitions *wiggly* from state j to state k within the interval $[a,b]$, $a, b \in \{1, ..., T\}$, $a < b$, if

$$s_a = j,\ s_b = k,\ s_t \in \{j,k\} \text{ for } a < t < b \tag{90}$$

Let $W_{jk}[a,b]$ denote the corresponding event, which we call *wiggly* interval transition. Note that the event $T_{jk}[a,b]$ implies $W_{jk}[a,b]$. Similar to equation (84), we calculate

$$\begin{aligned} P(W_{jk}[a,b],\mathcal{O}) &= P(s_b = k, s_t \in \{j,k\}, a < t < b, \mathcal{O}) \\ &= \alpha_a(j)\beta_b(k) \cdot P(s_b = k, s_t \in \{j,k\}, a < t < b, o_{a+1}, ..., o_b \mid s_a = j) \\ &\overset{(83)}{=} \alpha_a(j)\beta_b(k) \cdot \underbrace{\sum_{s_{a+1}\in\{j,k\}} ... \sum_{s_{b-1}\in\{j,k\}} \prod_{t=a}^{b-1} a_{s_t s_{t+1}} \cdot \psi_{s_{t+1}}(o_{t+1})}_{=:g_b(k)} \\ &= \alpha_a(j)\beta_b(k) \cdot g_b(k) \end{aligned} \tag{91}$$

with $s_a = j$ and $s_b = k$ in the last line. The term $g_{b-1}(k)$ is calculated recursively and in a numerically stable way by letting

$$\tilde{g}_a(j) = 1,\ \tilde{g}_a(k) = 0 \tag{92}$$

$$\tilde{g}_{t+1}(s) = (\tilde{g}_t(j)a_{js} + \tilde{g}_t(k)a_{ks})\,\psi_s(o_{t+1})/N_{t+1}\ , \quad t = a, ..., b-1,\ s \in \{j,k\}$$

The quantities N_t in the equation above are defined in (85) and (86). Then

$$g_b(k) = \tilde{g}_b(k) \cdot \prod_{t=a+1}^{b} N_t$$

Then,

$$\begin{aligned} P(W_{jk}[a,b] \mid \mathcal{O}) &= \frac{P(W_{jk}[a,b],\mathcal{O})}{P(\mathcal{O})} = \frac{\alpha_a(j)\beta_b(k)\cdot g_b(k)}{\sum_{s=1}^{S}\alpha_b(s)\beta_b(s)} \\ &= \frac{(\prod_{t=1}^{a} N_t)\,\tilde{\alpha}_a(j)\tilde{\beta}_b(k)\cdot\left(\prod_{t=a+1}^{b} N_t\right)\tilde{g}_{b-1}(k)}{\left(\prod_{t=1}^{b} N_t\right)\sum_{s=1}^{S}\tilde{\alpha}_b(s)\tilde{\beta}_b(s)} \\ &= \frac{\tilde{\alpha}_a(j)\tilde{\beta}_b(k)\cdot\tilde{g}_{b-1}(k)}{\sum_{s=1}^{S}\tilde{\alpha}_b(s)\tilde{\beta}_b(s)} \end{aligned} \tag{93}$$

13.4 Efficient Screening for Valid Interval Transitions

Suppose for a given threshold probability $q > 0.5$ and some maximum interval length $d \geq 1$, suppose we want to find all tuples (j,k,a,b), $j,k = 1,...,S$, $a,b =$

$1, ..., T$, $b-a \leq d$, such that the interval transition $T_{jk}[a,b]$ or the wiggly transition $W_{jk}[a,b]$ has posterior probability greater than q. We call these interval transitions *valid*.

Let $\gamma_t(s) = P(s_t = s \mid O)$ is the marginal posterior probability for state s_t. These quantities are obtained from the normalized forward / backward probabilities,

$$\gamma_t(s) = \alpha_t(s)\beta_{t+1}(s)/P(\mathcal{O}) = \frac{\tilde{\alpha}_t(s)\tilde{\beta}_t(s)}{\sum_j \tilde{\alpha}_t(j)\tilde{\beta}_t(j)} \tag{94}$$

Since the event $T_{jk}[a,b]$ implies $W_{jk}[a,b]$, which in turn implies $s_a = j$ and $s_b = k$, it follows that

$$P(T_{jk}[a,b] \mid \mathcal{O}) \leq P(W_{jk}[a,b]) \leq \min(P(s_a = j \mid \mathcal{O}), P(s_b = k \mid \mathcal{O})) = \min(\gamma_a(j), \gamma_b(k)) \tag{95}$$

Since $q > 0.5$, it follows that at every position t, there is at most one state which may serve as an endpoint of a valid transition interval. Hence, the sequence $r_t = \begin{cases} s & \text{if } \gamma_t(s) > q \text{ for some } s = 1, ..., S \\ 0 & \text{else} \end{cases}$ is well-defined. A valid interval transition $T_{jk}[a,b]$ therefore needs to satisfy the conditions

1. $$b - a \leq d$$

2. $$r_a \neq r_b \ , \ r_a, r_b \neq 0$$

This gives rise to a very simple and efficient pre-screening algorithm, which returns a list including all valid interval transitions (and, possibly some more invalid candidates).

14 HMM Augmentation and Rigid Viterbi Recoding

HMM have been widely used for genome segmentation into haplotypes [73, 85]. Yet, when using the Viterbi decoding the haplotype segments might result shorter than expected, or non-natural transitions of states might be estimated. Many different heurisitic strategies have been applied to match the real haplotype distribution with the estimated one. These errors usually arise from the intrinsic variability and low coverage of the data, and the innate characteristics of the HMM.

In diploid organisms, when attempting to infer the ancestral genome at different chromosome loci, the expected sequence is the succession of one of the parental genomes to a heterogeneous state and then to the genome of the other parental

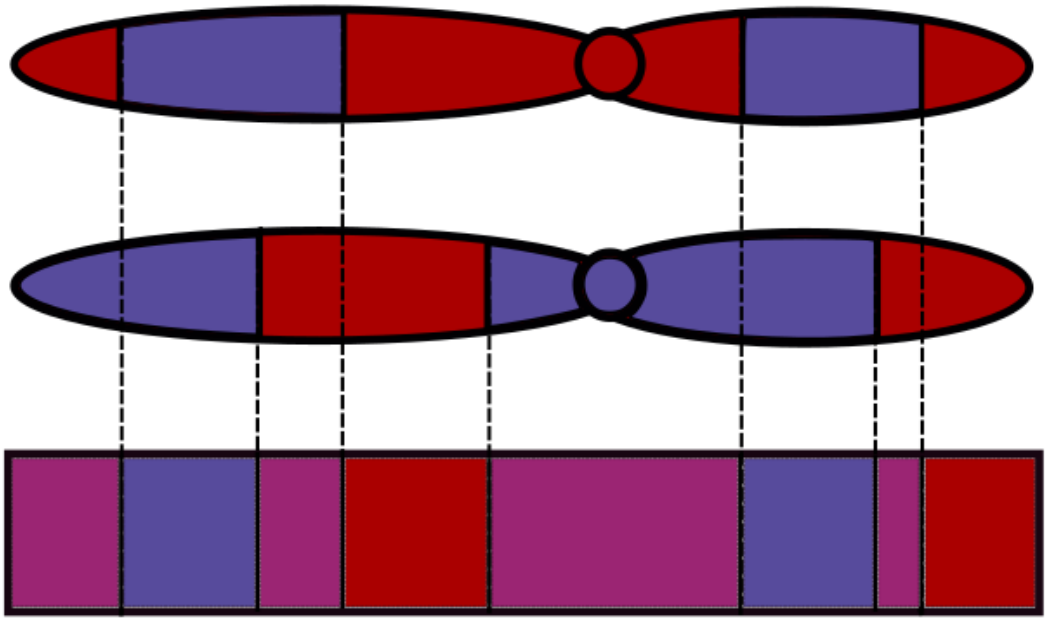


Figure 19: **Haplotype representation.** When two recombinant homologous chromosomes are genotyped, every locus can have 3 possible outcomes: parental line one, parental line two or heterozygous. Moreover, the transition between these outcomes must be: From one of the parental genotypes to a heterozygous state to the other parental line. The reason for such observation is that CO events happen randomly along the chromosomes and independently in each parental gametogenesis. Therefore, the probability of one parental line block to lie inside a bigger block of the other parental line is negligible, as well as, blocks of heterozygous states inisde a bigger block of one of the parental lines.

line (figure 19). Our preliminary analyzes showed small regions coding for one of the parental lines embedded within a larger block coding for the other parental line. This type of error was already studied by Patel *et al.* and they call them "islands" [98, 83]. This error arises due to HMM have no “memory”. After the first transition from one of the ancestor genotype to heterozygous state, it will “forget” whether the last non-heterozygous state was from one parental line or the other. Therefore, the HMM will allow a transition to any of the two homozygous states.

Further, there is no restriction on the minimum times that a state must be called before changing to another. Crossover events are rare in the chromosome. Therefore, genomic states are hundreds of thousands of nucleotides long. Standard Viterbi decoding produces state chains which can have haplotypes shorter than expected due to local random fluctuations in the data.

In order to account for these two drawbacks, we augment the HMM in two steps. First, we duplicate the heterozygous state depending on whether the previous state was the parental line one (hetp) or parental line two (hetm). Second, each state must be repeated at least a number of times R until it can transition to another. The allowed transitions are from parental state one to heterogeneous state "hetp", from "hetp" state to paternal state two, from this to "hetm" state and from "hetm" to parental state one.

Therefore, this method only requires to extend the matrix of transitions by substituting the fitted transition probabilities as shown in figure 20 and setting all other transition probabilities to 0. Once the observations have been Viterbi decoded by the augmented HMM, it is possible to map the three initial states again, thus removing the "island" regions.

The value for the rigidity parameter must be chosen manually. The rigidity depends on the quality of the reference genome, and the evolutionary distance of the two parental genomes. Consequently, the value R depends on the experiment in question.

15 Experimental Methods

Six F2 seeds from *Arabidopsis thaliana* Columbia-0 (Col-0) x Landsberg erecta (Ler-0), as well as 13 offspring genomes from a cross of recq4a/b and figl1 mutants of the same backgrounds, were stratified for 7 days at 4°C, sown on soil, and grown under normal greenhouse conditions. DNA was extracted and whole-genome libraries were prepared using the Illumina DNA TruSeq protocol. These Illumina libraries were then sequenced by the Max Planck Genome Center using a HiSeq 3000 machine with 151 bp paired-end reads targeting 2x genome coverage per sample. The short reads of each genome were aligned to the Arabidopsis reference TAIR10 using bowtie2 with default parameters. The initial marker set was composed of 519.215 SNPs which includes the actual position and knowl-

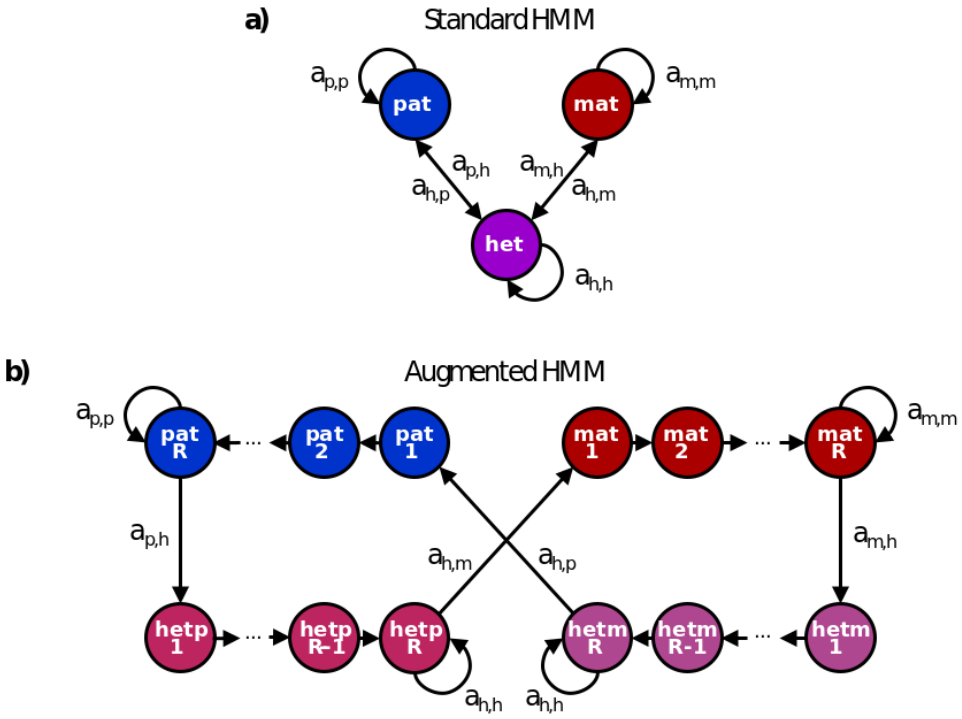


Figure 20: **Rigid Viterbi decoding.** The raw data contain many sources of variability that arise from many sources of errors and biases introduced during the sequencing and mapping of the reads. This leads to underperformance of the Viterbi algorithm. Standard Viterbi decoding will predict much more transitions among states than what one may expect from real haplotypes. Moreover, the transition probabilities are learned without any restriction. This results in the estimation of virtual transitions which have not been reported and are unlikely to happen. The crossover events are rather low and the expected transition of states in an organism would be: One of the parental lines to a heterozygous state and jump to the other parental line. For that reason, we have implemented an intermediate step where we duplicate the heterozygous state depending on whether the anterior state is of parental line one (hp) or the parental line two (hm). We force to stay in the same state a minimum of times R before jumping to another state to ensure a minimum length of the haplotypes.

edge of the genotypes of the two parental genomes for each SNP marker. These markers were selected from comparing Col-0 and Ler-0 assemblies [99]. For each sequenced sample and marker position we recorded the number of aligned bases to either of the parental alleles. All those SNPs with total allele count equal to 0 were removed from our initial set. These step discarded 172.499 SNPs.

To increase the coverage, the genome was partitioned in bins of thousand nucleotides and the number of bases in each SNP was summed up for both parental alleles in each bin for all samples. While some bins had allele count for all 19 samples, some other had for few samples or even none. In order to set a lower threshold, we filtered out all those bins that did not have allele counts in at least 5 samples. After the preprocessing, 65.623 bins per sample were kept to further train the HMM model.

The allele counts from Col-0 accession were used as the "success" counts and the sum of both alleles counts as the number of trials. Our model was fitted to the data and the marker annotation was obtained. All those bins that were classified as bad markers were removed and the rigid Viterbi algorithm was applied.

Finally, we screened all samples for potential transition regions and applied our transition estimation score. Only those transitions that had a higher score than 0.65 and agreed with a transition in the rigid Viterbi path were taken as valid transitions.

16 Results

16.1 In Silico Validation

The validation of our method was comparing the underlying genotype generated from 19 models, which parameters values were chosen randomly, with the predicted genotype after applying our learning method. Each sample contains 5 chromosomes and 5 thousand bins per chromosome. The coverage in each genomic position was generated from the empirical count distribution of the actual data. We also introduced missing values at some marker positions reflecting the missing information in our marker set.

The true genomic state was generated using an HMM and the marker states were set with π_m distribution. The data was generated from the Viterbi chain and marker states using different emission probabilities for each.

The generated data was used to fit a new model. The bad markers were removed and the markers were labeled using the Viterbi algorithm. Finally, the transition regions were pinpointed and compared against the transitions of the underlying genotype producing the observations. The same procedure was repeated ten more times to evaluate the accuracy of our method.

To asses the performance of our method to infer bad markers and state chains we compared our results with the model real states generating the observations.

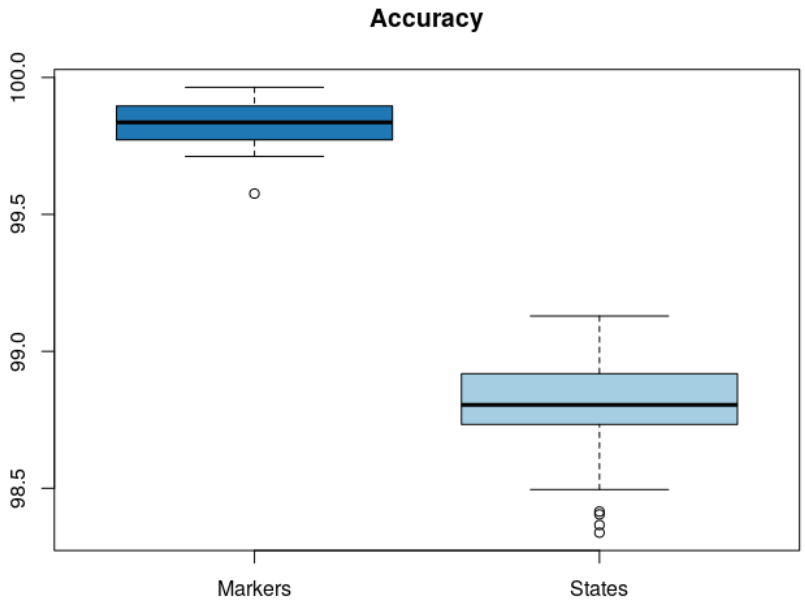


Figure 21: **Marker and State Accuracy.** Marker and State Accuracy in a simulation of 10 different experiments with 19 samples each. After fitting the model, the accuracy in recalling good and bad markers is 99.82%. The Viterbi path was applied to the data after removing bad markers. The agreement with the real underlying genotype is 98.1%.

With our method, we are able to correctly predict 99.82% of the markers as good or bad. Moreover, the predicted Viterbi chain agrees with the original one with an accuracy of 98.81% (figure 21).

The estimated bad markers were removed from the simulated data and the transitions regions were estimated with three different thresholds for our wiggly transitions scores (0.65, 0.75 and 0.95).

To evaluate the precision of our method we computed the percentage of our estimated transition regions contain a transition of genotypes in the real underlying Viterbi path, denoted as true positives (TP), percentage of regions that do not contain a real transition, denoted as false positive (FP) and the percentage of real transitions that our method missed, denoted as false negative (FN) (figures **??** and **??**).

The results of the analysis show that in general, our method performs well in predicting the transition regions. Our accuracy calling transition regions is on average higher than 80% and false negative rate lower than 20% with a mild threshold (0.65). Although, no big significant difference in true positive and false positive rates are found among different thresholds the false negative rate increases

Figure 22: **Graphical Representation FP, FN, TP.** First, given a state sequence, we generated a set of data points. Then, the data was used to fit an HMM and estimate the intervals within a change of states was given. If in the interval that we estimated there was a real switch of states in the real Viterbi path, we call it a true positive (TP, black box), if on the contrary there is no real switch of states, we call it false positive (FP, red box). If our analysis missed a transition, then this is a false negative (FN, dashed box).

considerably when the threshold is too high. Hence, many true transitions will be missed out (figure 23).

If the aim is to find the maximum true transitions one must use a mild threshold.

The results show that using a stringent threshold for the score of transitions boost the false negative rate significantly without an increase in the same magnitude of the true positive rate. Therefore, many transitions are missed. In consequence, for an exploratory analysis in which the aim is to find the maximum transitions, it will be preferred not to use a so stringent analysis. Thus, a threshold of 0.65 will be an optimum solution to obtain a good ratio of false negative and false positive maintaining a high true positive rate.

16.2 F2 Mapping Recombinant Population of *Arabidopsis thaliana*

Firstly, the genome was tiled in bins of thousand nucleotides and the number of SNPs per bin is presented in figure 24. On average, every bin contained 4.1 SNPs (figure 24). Secondly, the count numbers for each allele were summed up increasing the coverage of each position. After preprocessing the data the mean coverage of our data sets was $\sim 10x$ which guarantees a rise in the statistical power for our model (figure 25).

The count data at each bin was used to fit our model. The state path was

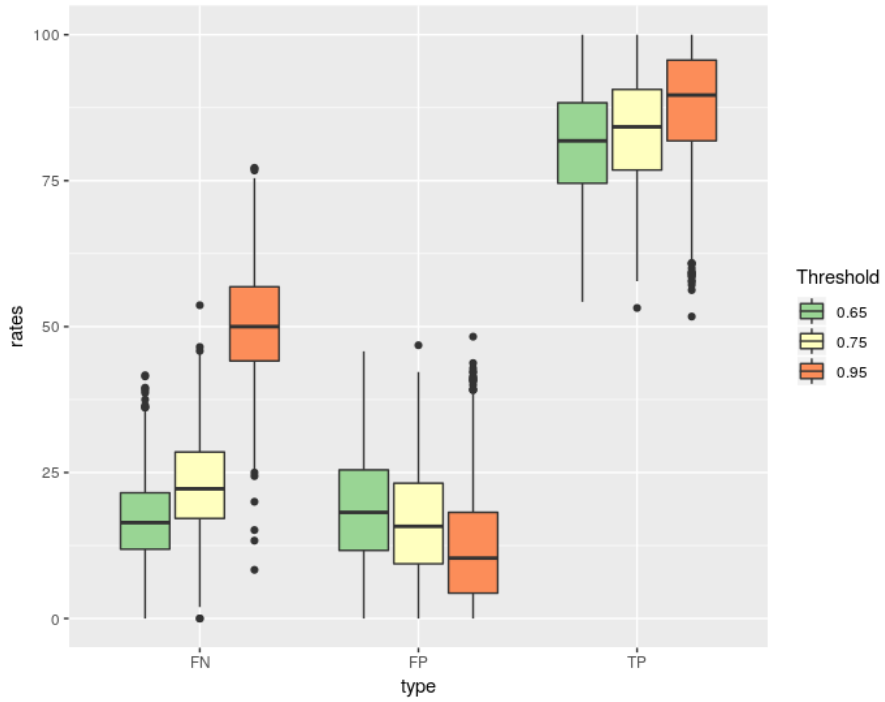


Figure 23: **Transition score threshold.** In order to determine the most adequate transition threshold, we computed the false negative (FN), false positive (FP) and true positive (TP) rates for the transition regions (y-axis). These values were computed as explained in figures **??** and **??**. The thresholds used for the analysis were 0.65 (green boxplot), 0.75 (yellow boxplot) and 0.95 (orange boxplot).

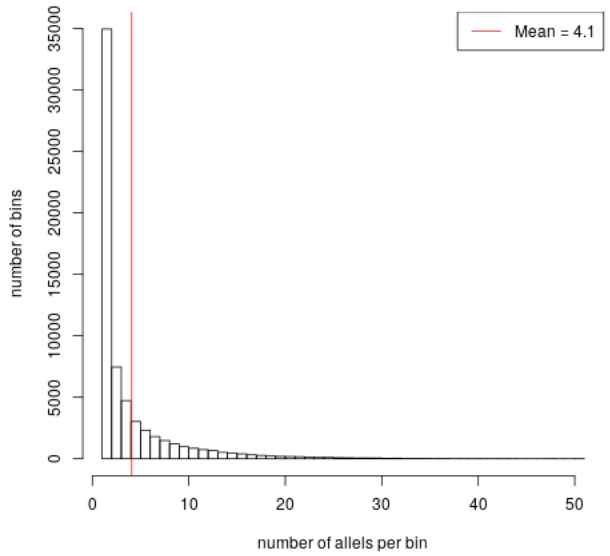


Figure 24: **SNP density distribution.** Distribution of number of SNPs per thousand nucleotides in *Arabidopsis thaliana* genome.

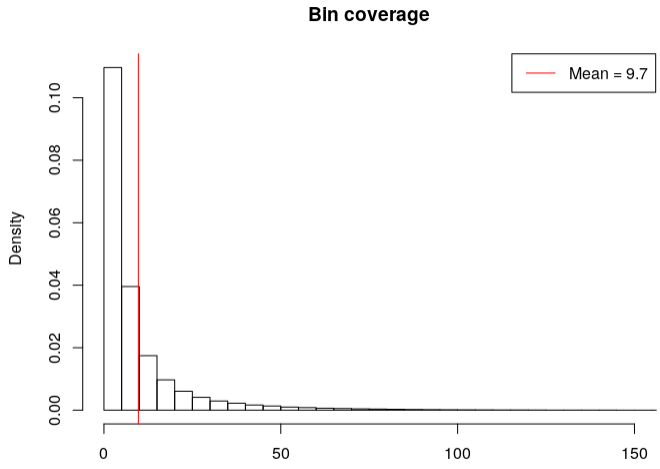


Figure 25: **Bin coverage distribution.** Total allele count distribution per bin after summing up the allele counts of the SNPs in each bin.

decoded using the Viterbi algorithm even at those positions which had no allele count information. We used the state blocks length distribution to estimate the minimum block length and filter out short haplotypes. All blocks shorter than 27 thousand nucleotides were removed using our rigid Viterbi decoding method. As observed in figure 26 our method smooths the Viterbi sequence and is less exposed to variability in the data.

To study the effect of the triple mutant in Arabidopsis, we estimated the intervals where a valid transition in wild-type and mutant samples. To this end, we have set the threshold to 0.65 for the wiggly transition score and ensure that exist a transition by comparing with the R-Viterbi path. Our method is able to map 50 % of the CO breakpoints within intervals of four thousand nucleotides wether is a wild-type or mutant (figure 27).

The ratio of COs per Mb was computed genome-wide and compared between wild-type and mutant samples. Mutant samples have a 3.86 fold increase in the number of estimated CO per Mb compared to wild-type (Wilcoxon test p.value < 0.01) (figure 28).

To determine whether the increase in CO breakpoints in mutants is consistent within all chromosomes, we computed the number of meiotic CO and compared in each chromosome to wild type. All chromosomes have an increased number of CO events in mutant and are in agreement with previous results from *Serra et al*[79] (figure 29).

Further analysis to investigate the CO frequency along the chromosome was carried out. All chromosomes show a depletion in CO events at the centromeres positions and increasing CO frequency along the chromosomal arms. This observations, agree with previous studies on the CO landscape in *Arabidsopsis* [100, 76, 83]. The increased CO frequency can also be observed in the arms indistinctly of the location.

17 Discussion

The construction of genetic maps is of great importance for the study of the relation between genotype and phenotype. GBS is a major technique for obtaining genetic information. We have developed a bioinformatics method capable of obtaining the ancestor genotype from mapping populations with low sequencing coverage. In this way, GBS has become an effective and efficient tool in genetic association studies.

HMMs are the method of choice for GBS [85, 84]. We have introduced several novel twists to the design of the HMM in order to boost its performance. First, our method uses Beta-Binomial distributions as emission distributions instead of Binomial distributions, to account for the overdispersion of the of the low sequencing data.

jointly exploits the information provided by all samples at a given genomic

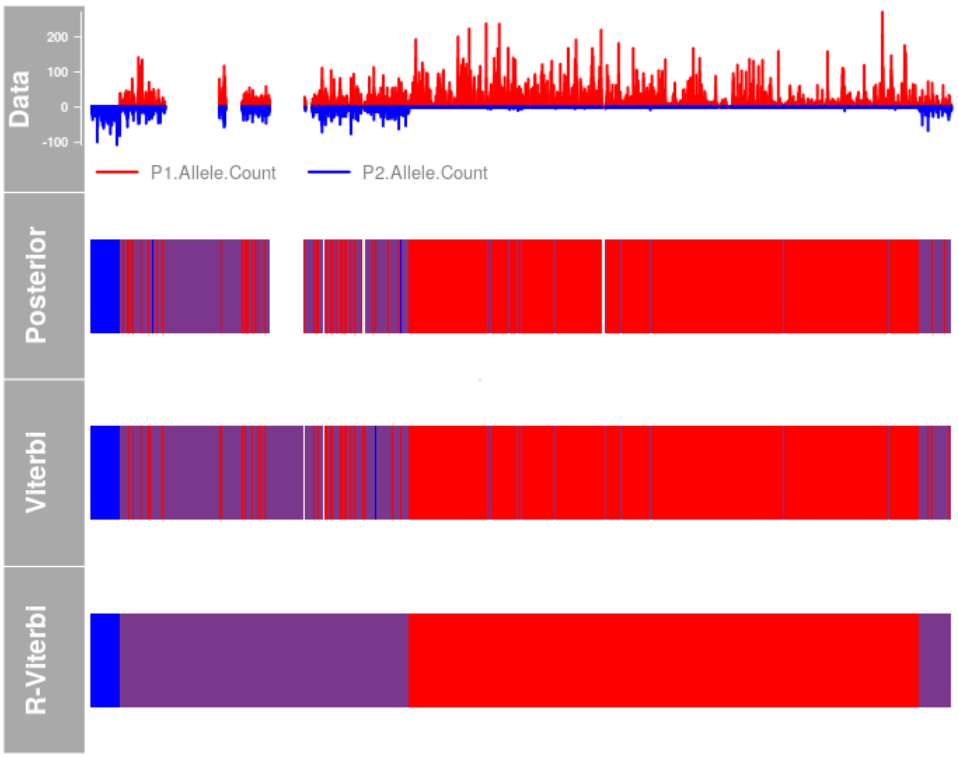


Figure 26: **Genome segmentation with different methods.** After the model is fitted there are several strategies to segment the genome to estimate ancestral parental genome. Here is shown the raw data from chromosome 4 in a wild-type sample and 3 different methods to estimate the underlying state responsible for the observable raw counts. (1) Posterior: The posterior decoding assumes that the state that generated the observable data point is the state k with higher posterior probability $P(s_t|o_t)$. (2) The Viterbi algorithm finds the most likely sequence of hidden states given the sequence of observed data points. (3) Our R-Viterbi method exploits the dynamic programming algorithm of the Viterbi algorithm but the transition matrix is transformed to ensure specific transitions. As a result the Viterbi path is smoother and fits better previous knowledge about recombinant chromosomes.

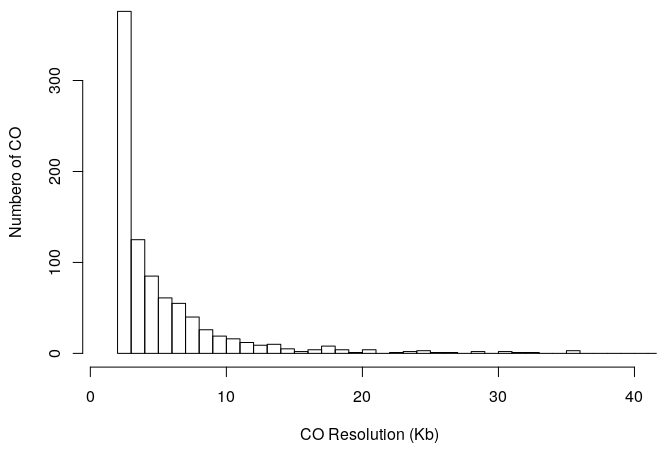


Figure 27: **CO resolution.** After screening for (wiggly) interval transitions, we obtain a set of regions of putative CO regions. The length of these regions is an objective measure of the precision by which CO breakpoints can be mapped.

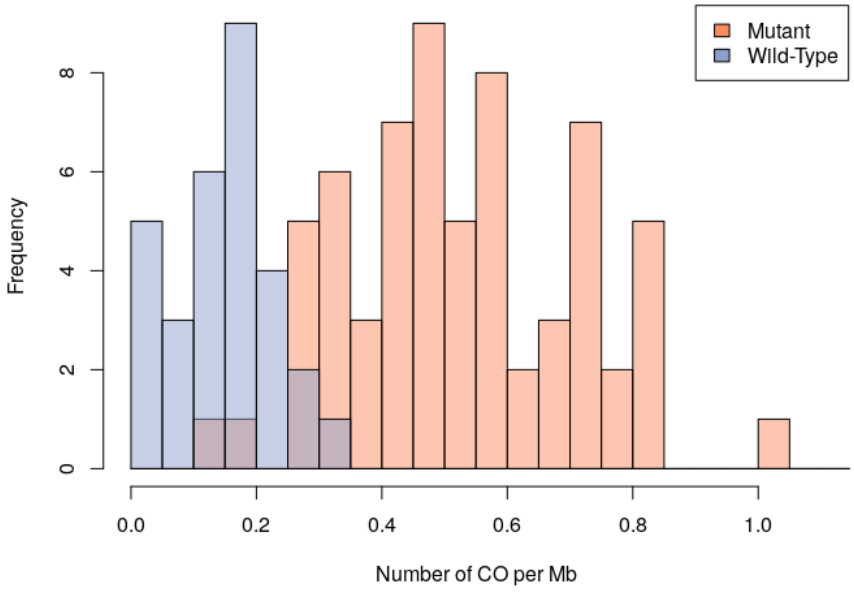


Figure 28: **CO events per Mb.** Histogram of the ratio of CO events per Mb estimated by our algorithm. The blue histogram shows the CO ratio in the wild-type samples. The triple mutant Recq4a/4b and Figl1 histogram is shown in orange.

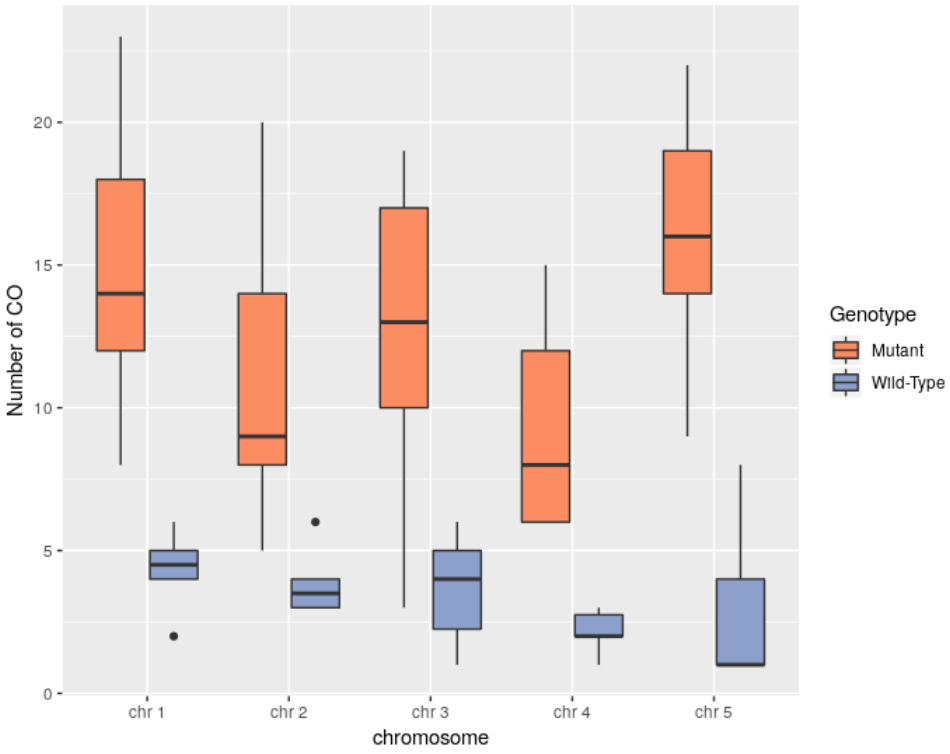


Figure 29: **CO events per chromosome.** Boxplot of the number of CO events predicted by our algorithm per chromosome (x-axis). The blue boxplots show the CO number in the wild-type chromosomes. The triple mutant Recq4a/4b and Figl1 boxplots are shown in orange.

break 102

Figure 30: **CO frequency along chromosomes.** The x-axis describes the genomic position at each chromosome, and the y-axis describes the ratio of CO events per 500 KB per wild-type samples (blue) respectively mutants (orange).

position to spot invalid respectively uninformative marker positions. As a second novelty, we have introduced an algorithm which augments the standard HMM used for genotype analysis. The augmented HMM is more accurate in the sense that it forbids fast (and presumably erroneous) forth-and-back state changes. Finally, we have defined the notion of a (wiggly) interval transition, and we provide an efficient formula for calculating the probability of such an event. We can use this new defined score in further steps to define regions that guarantees a transition from one haplotype to another base on the underlying Viterbi path. This score is not exclusive of our method and can be further used in all kinds of HMMs methods.

We have applied our pipeline to simulated data, where we have shown that we are able to reconstruct the sample genotypes with superior precision.

Further, we have applied our method in real data to investigate the effect of the triple mutant recq4a/4b and figl1 on the resolution of CO events during meiosis I. As demonstrated above, our results show a significant increase in the frequency of CO in the triple mutant as well as in the number of breakpoints. This increase is constant along the chromosome except for the centromeres where recombination is null in both cases [101].

To narrow down the length of our estimated transition regions one might decompose the bins into the original SNPs markers and apply again the HMM method. The additional step would focus on the markers and estimate new transition regions to place the CO breakpoints.

The use of third generation sequencing methods will improve also the analysis since the mapping errors would decrease. This might be an experimental solution to the islands [98].

Furthermore, remains to wrap up the source scripts into an R package to ease the analysis by the scientific community.

References

[1] Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. Transcriptome analysis by strand-specific sequencing of complementary {DNA}. *Nucleic Acids Research*, 37(18):e123, 2009.

[2] Marcilio C P de Souto, Ivan G Costa, Daniel S A de Araujo, Teresa B Ludermir, and Alexander Schliep. Clustering cancer gene expression data: a comparative study. 9:497.

[3] Michael Seifert, Marc Strickert, Alexander Schliep, and Ivo Grosse. Exploiting prior knowledge and gene distances in the analysis of tumor expression profiles with extended Hidden Markov Models. *Bioinformatics*, 27(12):1645–1652, jun 2011.

[4] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, nov 2009.

[5] Xiao-Li Li, Yin-Chet Tan, and See-Kiong Ng. Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method. *BMC Bioinformatics*, 7(S4):S23, dec 2006.

[6] Julia Glas, Sebastian Dümcke, Benedikt Zacher, Don Poron, Julien Gagneur, and Achim Tresch. Simultaneous characterization of sense and antisense genomic processes by the double-stranded hidden Markov model. *Nucleic Acids Research*, 44(5):e44–e44, mar 2016.

[7] Benedikt Zacher. Genomic data integration with hidden Markov models to understand transcription regulation. may 2016.

[8] Srabanti Maji and Deepak Garg. Gene Finding Using Hidden Markov Model. *Journal of Applied Sciences*, 12(15):1518–1525, dec 2012.

[9] Christos Lampros, Costas Papaloukas, Themis Exarchos, and Dimitrios I. Fotiadis. HMMs in Protein Fold Classification. pages 13–27. Humana Press, New York, NY, 2017.

[10] Byung-Jun Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, 10(6):402–415, sep 2009.

[11] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012.

[12] Amanda K. Lane, Chad E. Niederhuth, Lexiang Ji, and Robert J. Schmitz. pENCODE: A Plant Encyclopedia of DNA Elements. *Annual Review of Genetics*, 48(1):49–70, nov 2014.

[13] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, Manolis Kellis, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, feb 2015.

[14] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, may 2012.

[15] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):151, dec 2015.

[16] Jason Ernst and Manolis Kellis. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 12(12):2478–2492, nov 2017.

[17] Shanrong Zhao, Ying Zhang, William Gordon, Jie Quan, Hualin Xi, Sarah Du, David von Schack, and Baohong Zhang. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16(1):675, dec 2015.

[18] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[19] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[20] Bing Li, Michael Carey, and Jerry L Workman. The role of chromatin during transcription. *Cell*, 128(4):707–19, feb 2007.

[21] Chih Long Liu, Tommy Kaplan, Minkyu Kim, Stephen Buratowski, Stuart L Schreiber, Nir Friedman, and Oliver J Rando. Single-Nucleosome Mapping of Histone Modifications in S. cerevisiae. *PLoS Biology*, 3(10):e328, aug 2005.

[22] Daniel Holoch and Danesh Moazed. RNA-mediated epigenetic regulation of gene expression. *Nature Reviews Genetics*, 16(2):71–84, feb 2015.

[23] Travis N Mavrich, Ilya P Ioshikhes, Bryan J Venters, Cizhong Jiang, Lynn P Tomsho, Ji Qi, Stephan C Schuster, Istvan Albert, and B Franklin Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome.

[24] Sofia Luciana Battaglia. RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases. may 2017.

[25] Ali Shilatifard. Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression. *Annu. Rev. Biochem*, 75:243–69, 2006.

[26] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2926–31, feb 2010.

[27] Ronen Sadeh, Roee Launer-Wachs, Hava Wandel, Ayelet Rahat, Nir Friedman Correspondence, and Nir Friedman. Elucidating Combinatorial Chromatin States at Single-Nucleosome Resolution. *Molecular Cell*, 63, 2016.

[28] Brian D Strahl and David Allis. The language of covalent histone modi®cations. *NATURE www.nature.com*, 403(6), 2000.

[29] Mark Gerber and Ali Shilatifard. Transcriptional Elongation by RNA Polymerase II and Histone Methylation*. 2003.

[30] Michael Maximilian Lidschreiber. Genome-wide occupancy profiling of the RNA polymerase II transcription machinery in S. cerevisiae. sep 2013.

[31] Thomas Börner, Anastasia Yu. Aleynikova, Yan O. Zubo, and Victor V. Kusnetsov. Chloroplast RNA polymerases: Role in chloroplast biogenesis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1847(9):761–769, sep 2015.

[32] Stephen Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular cell*, 36(4):541–6, nov 2009.

[33] Steven Hahn and Elton T Young. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189(3):705–36, nov 2011.

[34] Geoffrey M Cooper. Eukaryotic RNA Polymerases and General Transcription Factors. 2000.

[35] Bing Zhu, Subhrangsu S Mandal, Anh-Dung Pham, Yong Zheng, Hediye Erdjument-Bromage, Surinder K Batra, Paul Tempst, and Danny Reinberg. The human PAF complex coordinates transcription with events downstream of RNA synthesis. *Genes & development*, 19(14):1668–73, jul 2005.

[36] Nevan J. Krogan, Jim Dover, Adam Wood, Jessica Schneider, Jonathan Heidt, Marry Ann Boateng, Kimberly Dean, Owen W. Ryan, Ashkan Golshani, Mark Johnston, Jack F. Greenblatt, and Ali Shilatifard. The Paf1 Complex Is Required for Histone H3 Methylation by COMPASS and Dot1p: Linking Transcriptional Elongation to Histone Methylation. *Molecular Cell*, 11(3):721–729, mar 2003.

[37] Jason H Brickner. Transcriptional Memory: Staying inÂ the Loop. *Current Biology*, 20:R20–R21.

[38] Pawel Grzechnik, Sue Mei Tan-Wong, and Nick J Proudfoot. Terminate and make a loop: regulation of transcriptional directionality. 2014.

[39] Sue Mei Tan-Wong, Hashanthi D Wijayatilake, and Nick J Proudfoot. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex.

[40] Belal El Kaderi, Scott Medler, Sarita Raghunayakula, and Athar Ansari. Gene Looping Is Conferred by Activator-dependent Interaction of Transcription Initiation and Termination Machineries * â¡ S. 2009.

[41] Dmitry K. Pokholok, Christopher T. Harbison, Stuart Levine, Megan Cole, Nancy M. Hannett, Tong Ihn Lee, George W. Bell, Kimberly Walker, P. Alex Rolfe, Elizabeth Herbolsheimer, Julia Zeitlinger, Fran Lewitter, David K. Gifford, and Richard A. Young. Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell*, 122(4):517–527, aug 2005.

[42] Siavash K. Kurdistani and Michael Grunstein. Histone acetylation and deacetylation in yeast. *Nature Reviews Molecular Cell Biology*, 2003.

[43] Miyong Yun, Jun Wu, Jerry L Workman, and Bing Li. Readers of histone modifications. *Nature Publishing Group*, 21(21), 2011.

[44] Stuart L Schreiber and Bradley E Bernstein. Signaling network model of chromatin. *Cell*, 111(6):771–8, dec 2002.

[45] Benjamin J E Martin, Kristina L Mcburney, Vicki E Maltby, Kristoffer N Jensen, Julie Brind 'amour, and Leann J Howe. Histone H3K4 and H3K36 Methylation Independently Recruit the NuA3 Histone Acetyltransferase in Saccharomyces cerevisiae.

[46] Michaela Smolle, Swaminathan Venkatesh, Madelaine M Gogol, Hua Li, Ying Zhang, Laurence Florens, Michael P Washburn, and Jerry L Workman. Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. 2012.

[47] Stephen M. Fuchs, R. Nicholas Laribee, and Brian D. Strahl. Protein modifications in transcription elongation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1):26–36, jan 2009.

[48] Michaela Smolle and Jerry L Workman. Transcription-associated histone modifications and cryptic transcription â. 2013.

[49] Swaminathan Venkatesh, Hua Li, Madelaine M Gogol, and Jerry L Workman. ARTICLE Selective suppression of antisense transcription by Set2-mediated H3K36 methylation. 2016.

[50] Anh Tram Nguyen and Yi Zhang. The diverse functions of Dot1 and H3K79 methylation. *Genes & development*, 25(13):1345–58, jul 2011.

[51] Zhi Han, Lu Tian, Thierry Pécot, Tim Huang, Raghu Machiraju, and Kun Huang. A signal processing approach for enriched region detection in RNA polymerase II ChIP-seq data. *BMC Bioinformatics*, 13(Suppl 2):S2, mar 2012.

[52] Huck Hui Ng, Franç Ois Robert, Richard A Young, and Kevin Struhl. Targeted Recruitment of Set1 Histone Methylase by Elongating Pol II Provides a Localized Mark and Memory of Recent Transcriptional Activity. *Molecular Cell*, 11:709–719, 2003.

[53] Anaïs F Bardet, Qiye He, Julia Zeitlinger, and Alexander Stark. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, 7(1):45–61, jan 2012.

[54] J. Soding, A. Biegert, and A. N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33(Web Server):W244–W248, jul 2005.

[55] Christian Betz, Gabriel Schlenstedt, and Susanne M Bailer. Asr1p, a novel yeast ring/PHD finger protein, signals alcohol stress to the nucleus. *The Journal of biological chemistry*, 279(27):28174–81, jul 2004.

[56] Anne Daulny, Fuqiang Geng, Masafumi Muratani, Jonathan M Geisinger, Simone E Salghetti, and William P Tansey. Modulation of RNA polymerase II subunit composition by ubiquitylation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(50):19649–54, dec 2008.

[57] Sean D Taverna, Serge Ilin, Richard S Rogers, Jason C Tanny, Heather Lavender, Haitao Li, Lindsey Baker, John Boyle, Lauren P Blair, Brian T Chait, Dinshaw J Patel, John D Aitchison, Alan J Tackett, and C David Allis. Yng1 PHD Finger Binding to H3 Trimethylated at K4 Promotes NuA3 HAT Activity at K14 of H3 and Transcription at a Subset of Targeted ORFs. *Molecular Cell*, 24:785–796, 2006.

[58] Tonya M Gilbert, Stephen L Mcdaniel, Stephanie D ByrumÊ, Jessica A Cades, Blair C R Dancy, Herschel Wade, Alan J TackettÊ, Brian D Strahl, and Sean D Taverna. A PWWP Domain-Containing Protein Targets the NuA3 Acetyltransferase Complex via Histone H3 Lysine 36 trimethylation to Coordinate Transcriptional Elongation at Coding Regions* â¡ S.

[59] W W Pijnappel, D Schaft, A Roguev, A Shevchenko, H Tekotte, M Wilm, G Rigaut, B Séraphin, R Aasland, and A F Stewart. The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes & development*, 15(22):2991–3004, nov 2001.

[60] TA Weinert and LH Hartwell. The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae. *Science*, 241(4863), 1988.

[61] Ted A Weinert and Leland H Hartwell. The RAD9 Gene Controls the Cell Cycle Response to DNA Damage in Saccharomyces cerevisiae.

[62] Muriel Grenon, Thomas Costelloe, Sonia Jimeno, Aisling O'Shaughnessy, Jennifer FitzGerald, Omar Zgheib, Linda Degerth, and Noel F. Lowndes. Docking onto chromatin via theSaccharomyces cerevisiae Rad9 Tudor domain. *Yeast*, 24(2):105–119, feb 2007.

[63] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, mar 2009.

[64] Daechan Park, Yaelim Lee, Gurvani Bhupindersingh, Vishwanath R. Iyer, TS Furey, PJ Park, J Rougemont, F Naef, MB Gerstein, ZJ Lu, EL Van Nostrand, C Cheng, BI Arshinoff, S Roy, J Ernst, PV Kharchenko, P Kheradpour, N Negre, LH Chadwick, B-K Lee, AA Bhinge, A Battenhouse, RM McDaniell, Z Liu, CT Workman, HC Mak, S McCuine, J-B Tagne, M Agarwal, CT Harbison, DB Gordon, TI Lee, NJ Rinaldi, KD Macisaac, BJ Venters, S Wachi, TN Mavrich, BE Andersen, P Jena, HS Rhee, BF Pugh, P Lefrançois, GM Euskirchen, RK Auerbach, J Rozowsky, T Gibson, K Yen, V Vinayachandran, K Batta, RT Koerber, BF Pugh, S Ghaemmaghami, W-K Huh, K Bower, RW Howson, A Belle, EG Wilbanks, MT Facciotti, EA Winzeler, DD Shoemaker, A Astromoff, H Liang, K Anderson, V Iyer, K Struhl, H Li, R Durbin, WJ Kent, CW Sugnet, TS Furey, KM Roskin, TH Pringle, JD Hunter, Y Zhang, T Liu, CA Meyer, J Eeckhoute, DS Johnson, M Preti, C Ribeyre, C Pascali, MC Bosio, B Cortelazzi, L Zhang, H Ma, BF Pugh, Z Shao, Y Zhang, G-C Yuan, SH Orkin, DJ Waxman, SC Tippmann, R Ivanek, D Gaidatzis, A Schöler, L Hoerner, S Shivaswamy, VR Iyer, JS Hardwick, FG Kuruvilla, JK Tong, AF Shamji, SL Schreiber, X Li, M Cai, CE Horak, NM Luscombe, J Qian, P Bertone, S Piccirrillo, VR Iyer, CE Horak, CS Scafe, D Botstein, M Snyder, RL Smith, AD Johnson, GE Zentner, T Tsukiyama, S Henikoff, J-S Hahn, Z Hu, DJ Thiele, VR Iyer, BL Kidder, G Hu, K Zhao, X Fan, K Struhl, X Zhu, M Wirén, I Sinha, NN Rasmussen, T Linder, PG Giresi, J Kim, RM McDaniell, VR Iyer, JD Lieb, RK Auerbach, G Euskirchen, J Rozowsky, N Lamarre-Vincent, Z Moqtaderi, RE Thurman, E Rynes, R Humbert, J Vierstra, MT Maurano, C Moorman, LV Sun, J Wang, E de Wit, W Talhout, XY Li, S MacArthur, R Bourgon, D Nix, DA Pollard, K Zaret, EL Van Dijk, CL Chen, Y D'Aubenton-Carafa, S Gourvennec, M Kwapisz, FC Holstege, EG Jennings, JJ Wyrick, TI Lee, and CJ Hengartner. Widespread Misinterpretable ChIP-seq Bias in Yeast. *PLoS ONE*, 8(12):e83506, dec 2013.

[65] Vibhor Kumar, Masafumi Muratani, Nirmala Arul Rayan, Petra Kraus, Thomas Lufkin, Huck Hui Ng, and Shyam Prabhakar. Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology*, 31(7):615–622, jul 2013.

[66] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, sep 2008.

[67] Elizabeth G. Wilbanks and Marc T. Facciotti. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS ONE*, 5(7):e11471, jul 2010.

[68] Georg Stricker, Alexander Engelhardt, Daniel Schulz, Matthias Schmid, Achim Tresch, and Julien Gagneur. Genome-wide generalized additive models.

[69] Brett W. Clelland and Michael C. Schultz. Genome stability control by checkpoint regulation of tRNA gene transcription. *Transcription*, 1(3):115–125, nov 2010.

[70] Jean-Philippe Lainé, Badri Nath Singh, Shankarling Krishnamurthy, and Michael Hampsey. A physiological role for gene loops in yeast.

[71] Jon-Matthew Belton, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, and Ye Zhan. HiâC: A comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276, nov 2012.

[72] Jinlei Han, Zhiliang Zhang, and Kai Wang. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, 11(1):21, dec 2018.

[73] Beth A. Rowan, Vipul Patel, Detlef Weigel, and Korbinian Schneeberger. Rapid and Inexpensive Whole-Genome Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping. *G3: Genes, Genomes, Genetics*, 5(3):385–398, mar 2015.

[74] Antoni Rafalski. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, 5(2):94–100, apr 2002.

[75] N H Barton and B Charlesworth. Why sex and recombination? *Science (New York, N.Y.)*, 281(5385):1986–90, sep 1998.

[76] Laurène Giraut, Matthieu Falque, Jan Drouaud, Lucie Pereira, Olivier C. Martin, and Christine Mézard. Genome-Wide Crossover Distribution in Arabidopsis thaliana Meiosis Reveals Sex-Specific Patterns along Chromosomes. *PLoS Genetics*, 7(11):e1002354, nov 2011.

[77] M C Whitby. Making crossovers during meiosis. *Biochemical Society transactions*, 33(Pt 6):1451–5, dec 2005.

[78] Ian R Henderson. Control of meiotic recombination frequency in plant genomes. *Current Opinion in Plant Biology*, 15(5):556–561, nov 2012.

[79] Heïdi Serra, Christophe Lambing, Catherine H. Griffin, Stephanie D. Topp, Divyashree C. Nageswaran, Charles J. Underwood, Piotr A. Ziolkowski, Mathilde Séguéla-Arnaud, Joiselle B. Fernandes, Raphaël Mercier, and

Ian R. Henderson. Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis. *Proceedings of the National Academy of Sciences*, 115(10):2437–2442, mar 2018.

[80] Robert J. Elshire, Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S. Buckler, and Sharon E. Mitchell. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, 6(5):e19379, may 2011.

[81] Stephan Ossowski, Korbinian Schneeberger, Richard M Clark, Christa Lanz, Norman Warthmann, and Detlef Weigel. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome research*, 18(12):2024–33, dec 2008.

[82] Jacob J Michaelson, Rudi Alberts, Klaus Schughart, and Andreas Beyer. Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11(1):502, sep 2010.

[83] P A Salomé, K Bomblies, J Fitz, R A E Laitinen, N Warthmann, L Yant, and D Weigel. The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*, 108(4):447–455, apr 2012.

[84] Weibo Xie, Qi Feng, Huihui Yu, Xuehui Huang, Qiang Zhao, Yongzhong Xing, Sibin Yu, Bin Han, and Qifa Zhang. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(23):10578–83, jun 2010.

[85] Peter Andolfatto, Dan Davison, Deniz Erezyilmaz, Tina T Hu, Joshua Mast, Tomoko Sunayama-Morita, and David L Stern. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome research*, 21(4):610–7, apr 2011.

[86] Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, Joanna L. Kelley, Michael Kertesz, Dmitry Pushkarev, Dmitri A. Petrov, and Anna-Sophie Fiston-Lavier. Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLoS ONE*, 9(9):e106689, sep 2014.

[87] J. W. Davey and M. L. Blaxter. RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 9(5-6):416–423, dec 2010.

[88] Koon Ho Wong, Yi Jin, and Zarmik Moqtaderi. Multiplex Illumina Sequencing Using DNA Barcoding. In *Current Protocols in Molecular Biology*, volume 101, pages 7.11.1–7.11.11. John Wiley & Sons, Inc., Hoboken, NJ, USA, jan 2013.

[89] Michael Quail, Miriam E Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, jul 2012.

[90] Todd J. Treangen and Steven L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, jan 2012.

[91] Jacob F. Degner, John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, dec 2009.

[92] A. Knoll and H. Puchta. The role of DNA helicases and their interaction partners in genome stability and meiotic recombination in plants. *Journal of Experimental Botany*, 62(5):1565–1579, mar 2011.

[93] James D. Higgins, Maheen Ferdous, Kim Osman, and F. Christopher H. Franklin. The RecQ helicase AtRECQ4A is required to remove inter-chromosomal telomeric connections that arise during meiotic recombination in Arabidopsis. *The Plant Journal*, 65(3):492–502, feb 2011.

[94] Frank Hartung, Stefanie Suer, and Holger Puchta. Two closely related RecQ helicases have antagonistic roles in homologous recombination and DNA repair in Arabidopsis thaliana. *PNAS*, 104:18836–18841, 2007.

[95] Chloe Girard, Liudmila Chelysheva, Sandrine Choinard, Nicole Froger, Nicolas Macaisne, Afef Lehmemdi, Julien Mazel, Wayne Crismani, and Raphael Mercier. AAA-ATPase FIDGETIN-LIKE 1 and Helicase FANCM Antagonize Meiotic Crossovers by Distinct Mechanisms. *PLOS Genetics*, 11(7):e1005369, jul 2015.

[96] Yan Zhou, Xiang Wan, Baoxue Zhang, and Tiejun Tong. Classifying next-generation sequencing data using a zero-inflated Poisson model. *Bioinformatics*, 34(8):1329–1335, apr 2018.

[97] Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, may 2005.

[98] Vipul Kumar Patel. *Genotyping by sequencing from sparse sequenced genomes representations from bi- and multi- parental mapping population using a HMM approach*. PhD thesis, mar 2016.

[99] Luis Zapata, Jia Ding, Eva-Maria Willing, Benjamin Hartwig, Daniela Bezdan, Wen-Biao Jiao, Vipul Patel, Geo Velikkakam James, Maarten Koornneef, Stephan Ossowski, and Korbinian Schneeberger. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):E4052–60, jul 2016.

[100] Jan Drouaud, Raphaël Mercier, Liudmila Chelysheva, Aurélie Bérard, Matthieu Falque, Olivier Martin, Vanessa Zanni, Dominique Brunel, and Christine Mézard. Sex-Specific Crossover Distributions and Variations in Interference Level along Arabidopsis thaliana Chromosome 4. *PLoS Genetics*, 3(6):e106, 2007.

[101] C. Somerville, S Somerville, R. W. Davis, C. Dean, N. M. Crawford, Xiaoying Lin, Michael Bevan, George Murphy, Barbara Harris, Laurence D. Parnell, W. Richard McCombie, Robert A. Martienssen, Marco Marra, and Daphne Preuss. Genetic Definition and Sequence Analysis of Arabidopsis Centromeres. *Science*, 285(5426):380–383, jul 1999.

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Achim Tresch betreut worden.

Köln, 11^{th} of March 2021, Rafael Campos Martín