Inference of Genomic Landscapes using Ordered Hidden Markov Models with Emission Densities (oHMMed)

Claus Vogl[a,b], Mariia Karapetiants[a], Burçin Yıldırım[a,b], Hrönn Kjartansdóttir[a], Carolin Kosiol[c], Juraj Bergman[d], Michal Majka[e], Lynette Caitlin Mikula[c,]

[a] *Department of Biomedical Sciences, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*
[b] *Vienna Graduate School of Population Genetics, A-1210 Vienna, Austria*
[c] *Centre for Biological Diversity, School of Biology, Univeristy of St Andrews, United Kingdom*
[d] *Centre for Biodiversity Dynamics in a Changing World (BIOCHANGE) & Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, DK-8000 Aarhus C, Denmark*
[e] *Erste Group Bank AG, 1100 Vienna, Austria*

Structured Abstract

Motivation. Genomes are inherently inhomogeneous, with features such as base composition and gene density varying along chromosomes. Biological analyses often aim to quantify this variation, account for it during inference procedures, and ultimately determine the causal processes behind it. Since sequential observations along chromosomes are not independent, it is unsurprising that autocorrelation patterns have been observed *e.g.,* in the base composition of hominids. In this article, we develop a class of Hidden Markov models (HMMs) called oHMMed (ordered HMM with emission densities): They identify the number of comparably homogeneous regions within observed sequences with autocorrelation. These are modelled as discrete hidden states; the observed data points are then realisations of continuous probability distributions with state-specific means that enable ordering of these distributions. The observed sequence is labelled according to the hidden states, permitting only neighbouring states that are also neighbours within the ordering of their associated distributions. The parameters that characterise these state-specific distributions are then inferred.

Results. We apply oHMMed to the proportion of G and C bases (modelled as a mixture of normal distributions) and the number of genes (modelled as a mixture of poisson-gamma distributions) in windows along the human, mouse, and fruit fly genomes. This results in a partitioning into statistically distinguishable regions according to these two features for each species, and facilitates the quantification of the positive correlation between them.

Availability and Implementation. The algorithms are implemented in the R package oHMMed, and details are available on GitHub.

Contact. lcm29@st-andrews.ac.uk (Lynette Caitlin Mikula)

Supplementary Information. Online-only supplementary data is available at the journal's web site.

1. Introduction

Hidden Markov models (HMMs) are often described as the workhorse of modern biological sequence analysis [*e.g.,* 11]. While originally used for speech recognition [*e.g.,* 26, 16], they are now central to any field that utilises advanced statistical methods. The first biological application of HMMs was to genome segmentation, in particular to segmentation of genomes according to the level of the bases guanine and cytosine (G+C) vs adenine and thymine (A+T), *i.e.,* according to GC-rich and GC-poor regions [5, 25].

In general, such genome segmentation algorithms infer the marginal probability that each position along a chromosome or stretch of DNA belongs to one of a moderately sized number K of hidden states. As the genome is traversed, the transition matrix $\mathbf{T}_{K\times K}$ describes the probabilities of remaining in the same hidden state or switching to another based solely on the current state, making the model a Markov Chain. The assignment of each genomic region to a state is known as "annotation". The observed data point per region is assumed to be emitted conditional on the assigned hidden state. These data points are classically assumed to be drawn from a discrete R dimensional alphabet according to the matrix of emission probabilities $\mathbf{E}_{K\times R}$. However, they can also be modelled as realisations of a continuous distribution with state-specific parameters such that *e.g.,* the overall distribution of the emitted data conform to a Gaussian mixture model [*e.g.,* 16, 27, 14]. Conditional on both $\mathbf{T}$ and either $\mathbf{E}$ or the parameters of the emission densities, the likelihood of the observed data can then be calculated together with the marginal probability of the state at each position using dynamic programming (recursive forward and backward passes of an algorithm) [5, 11]. The Baum-Welch algorithm [1, 11], a variant of the expectation-maximisation algorithm, can often be employed to find a local maximum of the likelihood, corresponding estimates $\hat{\mathbf{T}}$, $\hat{\mathbf{E}}$, and the initial probability of states $\hat{\pi}$. Alternatively, Bayesian approaches typically use Markov Chain Monte Carlo (MCMC) methods, *e.g.,* the Gibbs sampler, to obtain a sample from the joint posterior distribution, from which all marginal posterior distributions follow [*e.g.,* 2, 28, 13, 14].

Our method - oHMMed (ordered HMM with emission densities) - assumes continuous emissions. In one case, the emission is a normal mixture that corresponds to the observed density of the data points. In the other, the emission density is a gamma mixture initially; however, rate parameters of poisson distributions are subsequently drawn from the individual gamma distributions, yielding an observed density of gamma-poisson mixtures (where the data points are discrete counts). Our core assumption for both variants of emissions is that the observed sequence data exhibit appreciable autocorrelation. In order to model this pattern within our HMM framework, the emission densities are parameterised so that they become convex functions within their natural range. This is done by first assuming one shared parameter among the hidden states (the standard deviation for normal distributions, and the shape parameter for gamma distributions), while the other varies between the states (the mean for the normal distributions, and the rate parameter for the gamma distributions).

The state-specific parameters can then be used to sort the states by increasing mean of their emitted distributions. Restricting transitions to neighbouring states within the thereby imposed order induces a tridiagonal transition matrix $\mathbf{T}$ that governs the autocorrelation pattern (and makes the Markov Chain of hidden states reversible). Utilising a Markov Chain Monte Carlo (MCMC) algorithm, oHMMed provides a best-fit annotation of the observed sequence, corresponding estimates of the transition rate matrix, and estimates of the state-specific and shared parameters of the emitted distributions.

The inherent ordering of hidden states by a single parameter bestows oHMMed with several distinguishing properties: Firstly, it avoids the problem of "label switching" that plagues most MCMC methods [20]. Even comparison of output of the same algorithm run with different starting conditions is not straightforward when this occurs; with oHMMed, however, the labels of the states relative to each other are clearly defined and facilitate "label matching" between runs or algorithms. Secondly, the number of estimable parameters is reduced. We will show that this improves algorithm behaviour and guards against over-fitting. Thirdly, we are able to propose intuitive diagnostic criteria for selecting the appropriate number of hidden states (which is typically assumed as given in classic forward-backward and Baum-Welch algorithm). This is noteworthy because development of model selection criteria for HMMs is complex and no consensus criteria exist [4, 7, 34].

Overall, oHMMed is specifically designed to segment autocorrelated sequences into states that have *statistically significant* differences in mean emissions; these can then be compared meaningfully since they differ only in this metric. This simple and otherwise assumption-free approach is generally applicable whilst remaining agnostic to any causal biological forces; in fact, these can be studied further and without bias after oHMMed segmentation. Previously, general biological autocorrelation patterns have been modelled by incorporating HMM components into more complex econometric and socioeconomic time series models [14]; these describe the observed pattern of stochastic variation (as a "random walk") with recurring sections where the fluctuations are different means (termed "regime changes") (see *e.g.,* Markov Switching Models [19] or Changepoint Models, reviewed in [31]).

Recall that the first biological application of HMMs was to the variation of GC proportion along genomes [5, 25]. Even before this, mammalian chromosomes had been described as a "mosaic" of long chromosomal regions of relatively homogeneous GC-content termed "isochores" [10]. Traditionally, five states of increasing average GC proportion falling within fixed (predetermined) ranges are assumed as part of "isochore theory" [*e.g.,* 6, their Fig 1], and these five states have been delineated in genomes of various species, also including invertebrates [3]. Note that the variance in the distribution of GC proportion is generally considered to differ between states, *e.g.,* in humans the states of higher average GC proportion are more variable [6]. While the validity of "isochore theory" has been fiercely debated [22, 23], varying and considerably autocorrelated genomic GC proportions are evident in mammalian sequence data: Transitions from regions with high GC content to regions with low GC content

generally proceed through a sequence of regions with intermediate GC content *i.e.,* transitions seem to occur predominantly between neighbouring states (see Fig. (1b) in [12] and Fig. (4) in [8]). Therefore, we apply oHMMed with normal emission densities to the variation of GC proportion along the genomes of humans, mice, and fruit flies. Further, we employ oHMMed with gamma-poisson emission densities to the gene content of the genomes of these species, as it is known to correlate with GC content [32], and are thus able to further assess this correlation.

2. System and Methods

2.1. System: Genomic Sequences

2.1.1. Human

The *Homo sapiens* GRCh38 reference genome was obtained from the UCSC Genome Browser [21]. The autosomes were subdivided into non-overlapping $100\,kb$ windows. For each window with at least 90% successfully sequenced bases (*i.e.,* $> 90\,kb$ non-N bases), the GC proportion was calculated. The number of protein coding genes was determined for the retained windows using GENCODEv21 [15] via the UCSC Genome Browser [21]. Genes were included in a window if the gene coordinates partially or completely overlap with the window coordinates, *i.e.,* genes that overlap into neighbouring windows were not excluded.

2.1.2. Mouse

The *Mus musculus* GRCm39 reference genome was obtained from Ensembl108 [9]. The autosomes were subdivided into non-overlapping $100\,kb$ windows. For each window with at least 60% successfully sequenced bases (i.e., $> 60\,kb$ non-N bases), the GC proportion was calculated. Additionally, the number of protein coding genes was counted for each window as before.

2.1.3. Fruit Fly

The reference genome of *Drosophila melanogaster* (version r5.57) was obtained from FlyBase(2022-05) [17]. Since the genome is much shorter than that of the mammals above, the autosomes were subdivided into non-overlapping $10\,kb$ windows. For each window comprising more than 90% successfully sequenced bases (*i.e.,* $> 90\,kb$ non-N bases), the GC proportion was calculated. Again, the number of protein coding genes was determined for each window as before.

2.1.4. Data Availability Statement

All data, augmented with the corresponding oHMMed-inferred hidden states, are available on GitHub.

2.2. Methods: Hidden Markov Models with Constrained Transition Probabilities and Emission Densities

2.2.1. General Notation

Assume a vector $\boldsymbol{\theta}$ of random variables, where each entry θ_l represents an unknown or hidden state at position l along the genome, where $1 \leq l \leq L$. Each θ_l may assume one of K states indexed by i, with $1 \leq i \leq K$. The random vector is a Markovian sequence with an invariant $K \times K$ transition matrix $\mathbf{T}$, where elements $\Pr(\theta_{l+1} \,|\, \theta_l)$ govern the step-by-step probabilities of going from a specific state to the next as the genome is traversed. Depending on the state probabilities at every position along the genome $1 \leq l \leq L$, a realisation (data point) y_l is sampled from a (continuous) probability density function. These are collected in the vector of realised emissions $\boldsymbol{y}$.

2.2.2. Convex Emission Densities

They key assumption behind oHMMed is that the hidden states can be ordered in a way that is reflected in their emissions. To illustrate: assume two states θ_i, θ_j with $(i, j) \,\epsilon\, [1, .., K]$ and $i < j$, and two emissions $(y_i, y_j) \,\epsilon\, \boldsymbol{y}$ with $y_i < y_j$. Then the relationship:

$$\frac{\Pr(y_i \,|\, \theta_i)}{\Pr(y_i \,|\, \theta_j)} \leq \frac{\Pr(y_j \,|\, \theta_j)}{\Pr(y_j \,|\, \theta_i)}\,, \tag{1}$$

should hold, and hold with equality only on a null set. In other words, the ratio $\frac{\Pr(\boldsymbol{y} \,|\, \theta_i)}{\Pr(\boldsymbol{y} \,|\, \theta_j)}$ should be convex. For strictly positive densities, we can take the logarithm here and define a convexity condition that must be fulfilled for an ordering of states to be possible:

$$\frac{d}{d\boldsymbol{y}}\big(\log(\Pr(\boldsymbol{y} \,|\, \theta_i)) - \log(\Pr(\boldsymbol{y} \,|\, \theta_j)\big) \leq 0\,, \text{ for } i < j. \tag{2}$$

Normal Emissions. Consider any two of K total states that each emit normal distributions; in particular y_i is drawn from $N(\mu_i, \sigma^2)$ and y_j is drawn from $N(\mu_j, \sigma^2)$ where $\mu_i \leq \mu_j$. The above convexity condition holds:

$$\begin{aligned} \frac{d}{d\boldsymbol{y}}\big(\log(\Pr(\boldsymbol{y} \,|\, \mu_i, \sigma)) - \log(\Pr(\boldsymbol{y} \,|\, \mu_j, \sigma)\big) &\leq 0 \\ -(\boldsymbol{y} - \mu_i)/\sigma^2 &\leq -(\boldsymbol{y} - \mu_j)/\sigma^2 \\ \mu_i &\leq \mu_j\,. \end{aligned} \tag{3}$$

Note that the shared standard deviation σ between states is a necessary assumption here.

Gamma Emissions. Consider any two of K total states that each emit a gamma distribution; in particular y_i is drawn from $G(\alpha, \beta_i)$ and y_j is drawn from

$G(\alpha, \beta_j)$ where $\beta_i \leq \beta_j$. The above convexity condition holds:

$$\begin{aligned} \frac{d}{d\boldsymbol{y}}\big(\log(\Pr(\boldsymbol{y} \,|\, \alpha, \beta_i)) - \log(\Pr(\boldsymbol{y} \,|\, \alpha, \beta_j)\big) \leq 0 \\ \frac{\alpha - 1}{\boldsymbol{y}} - \beta_i \boldsymbol{y} \leq \frac{\alpha - 1}{\boldsymbol{y}} - \beta_j \boldsymbol{y} \\ \beta_i \geq \beta_j \,. \end{aligned} \tag{4}$$

Here, the shared shape parameter α between states is a necessary assumption.

2.2.3. Transition Probabilities

In addition to assuming that the hidden states can be ordered, we restrict transitions to neighbouring states within this ordering. This results in a tridiagonal KxK transition matrix $\mathbf{T}$, since $\Pr(\theta_i \,|\, \theta_j) > 0$ only for j in $(i-1, i, i+1)$, while $\Pr(\theta_i \,|\, \theta_j) = 0$ otherwise. Note that the transition matrix thus has $2K - 2$ estimable parameters. Further, we let the prior probabilities of the hidden states correspond to the stationary distribution of the transition matrix $\mathbf{T}$, which we denote as the row vector $\boldsymbol{\pi}$ with entries $\pi_i = \Pr(\theta_{l-1} = i)$. It follows that the system is in detailed balance, i.e fulfills the equations:

$$\begin{aligned} \Pr(\theta_{l-1} = i) \Pr(\theta_l = i \,|\, \theta_{l-1} = j) = \ \Pr(\theta_{l-1} = j) \Pr(\theta_l = j \,|\, \theta_{l-1} = i) \\ \Pr(\theta_l = i, \theta_{l-1} = j) = \Pr(\theta_{l-1} = j, \theta_l = i) \,. \end{aligned} \tag{5}$$

Note that this corresponds to the structure of double-stranded DNA sequences, where the 5' end of one strand corresponds to the 3' end of the other.

3. Algorithm

The assignment of each position along the genome to a hidden state is determined by the forward-backward passes of a HMM algorithm (see Supplementary Data A1, Section "Forward backward HMM algorithm"). After each pass, the transition rates and parameters of the emission densities per state must be estimated and their fit evaluated. Baum-Welch expectation-maximisation algorithms, which are often employed for parameter estimation with HMMs [*e.g.,* 11], require independent maximisation of $\boldsymbol{\pi}$ and $\mathbf{T}$. Therefore, we develop a Markov Chain Monte Carlo algorithm, in particular a Gibbs sampler, that estimates the posterior distributions of the transition rates and parameters of the emitted distributions given the current annotation and the observed data. The samplers are fully characterised in Supplementary Data A1 for normal and Supplementary Data A2 for gamma-poisson emissions.

4. Implementation

Our algorithms are available as the R package oHMMed, with the source files deposited on GitHub. Explicit usage recommendations can be found there, which include pointers on setting partially informative priors and initial values

for the estimable parameters. We also describe the accompanying suite of diagnostics for assessing convergence and model fit. Note that oHMMed performs 10 iterations of the Gibbs sampler in roughly 1.06 seconds on a sequence of length 2^{11}, and that the speed decreases linearly with sequence length (details in the usage recommendations on GitHub).

5. Results

5.1. Empirical Transition Rates

To ensure our data satisfy oHMMed assumptions, we compare the mean differences in average GC proportion and in protein coding gene content of consecutive windows along the genomes of our study species to the same differences in random permutations of these windows (see Fig. 1). The variance of differences for the genomic GC proportion in humans, mice, and fruit flies respectively is $\hat{\sigma}_n^2 = (0.000673, 0.00055, 0.00103)$ compared to $\hat{\sigma}_r^2 = (0.00469,\ 0.00552,\ 0.0021)$ for the permuted genome sequence, with the ratios being highly significant in all cases (F-test with p-val $p < 2.2e^{-16}$). Similarly, we report variance differences of $\hat{\sigma}_n^2 = (1.14, 1.73, 2.54)$ in the number of protein coding genes in consecutive genomic windows of the three species compared to $\hat{\sigma}_r^2 = (2.43, 5.01, 4.29)$ in the random sequence, attaining the same level of significance for these ratios. Overall, we find evidence for an underlying autocorrelation structure in the observed sequences of both base composition and gene content in all species. Note, however, that this pattern is particularly pronounced in the genomic GC proportion of humans and mice (see Fig. 1A, B).

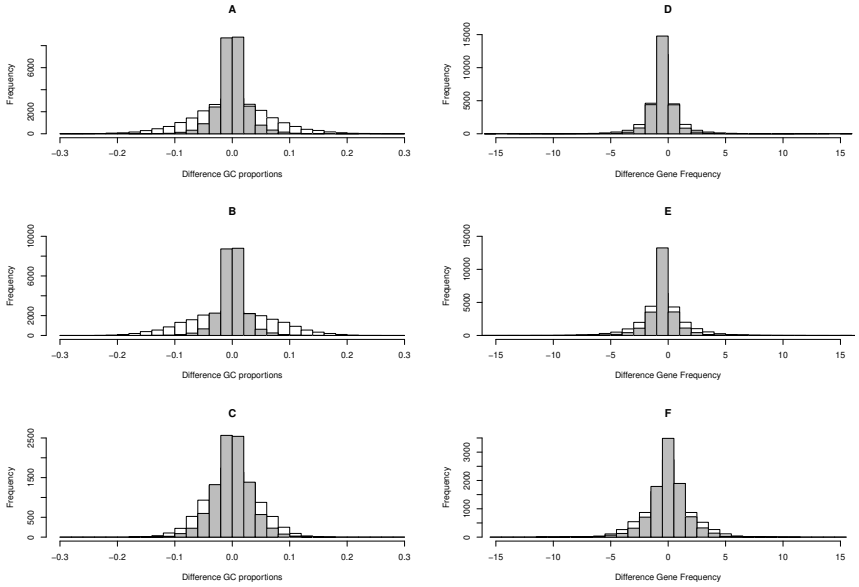

Figure 1: Histograms of pairwise differences in average GC proportion (left column) and average gene content (right column) for neighbouring windows along the observed genomic sequences (grey histograms) and along random permutations of the genomic sequences (white histograms) for humans (A), mice (B), and fruit flies (C) respectively.

5.2. Segmentation of the Human Genome

GC Proportion. After running oHMMed with normal emission densities on the human genomic GC proportion several times assuming $K = (2, .., 8)$ hidden states, we chose to segment the GC content into $K = 5$ states (all runs with 1500 iterations and a 20% burn-in). This is a compromise between two aspects of the inference procedure that we have chosen as diagnostics: The first is model fit. Adding more states one by one starting from $K = 2$ will - at least initially - lead to an increasing posterior log-likelihood as the overall emitted distribution becomes more flexible. However, there should be a number of states after which the increase in log-likelihood plateaus, pointing to a candidate number of hidden states. The second aspect of the decision concerns whether the difference between means of the emissions generated by neighbouring states are statistically different. For the human GC proportion, the posterior log-likelihoods start to plateau after about $K = 5$ (Fig. 2A) and the means are well separated, *i.e.,* the 68% confidence intervals (means plus/minus one standard deviation) barely overlap (see Fig. 2C). In fact, only first and second states clearly overlap at this level, and the second and third do so marginally; however, this interpretation of the plots is conservative and the means are still statistically distinguishable by a standard (one-sided) t-test at the 95% confidence level (part of standard oHMMed output). From Fig. 2C, the continuous variation of observed GC content among windows is also evident: Discretisation into $K = 5$ states intro-

duces discontinuities, in particular between the wider-spaced means. Increasing K would have resulted in an increasing overall likelihood (Fig. 2A) and better fitting overall distribution of emissions compared to the distribution of observed GC proportion along the genome (see Figs. 2C, D) at the cost of increasing the overlap between neighbouring states to the point of non-significant t-test outcomes. Decreasing K would have the opposite result.

For the selected five states, we inferred means of 0.360,0.391,0.430,0.4834,0.567 respectively, and a shared standard deviation of 0.174. The proportion of genomic windows assigned to each state splits to 0.261 ,0.320, 0.256, 0.125, 0.038. Generally, segmentation of the human genome into 5 states aligns with past literature [6, 8], with "isochore theory" dictating a split of comparatively homogeneous DNA regions $\geq 300kb$ into classes with mean GC proportions of $< 0.38, 0.38-0.42, 0.42-0.47, 0.47-0.52, > 0.52$ based on both the variation of mean and standard deviation of G+C alleles. Note that each of these categories contains one of our inferred means. Relative proportions of human DNA in these classes are cited as $0.19, 0.37, 0.31, 0.11, 0.03$, with the lowest two classes often merged; this differs somewhat from our inference.

Our method additionally infers a transition rate matrix of

$$\begin{pmatrix} 0.961 & 0.039 & 0 & 0 & 0 \\ 0.032 & 0.883 & 0.085 & 0 & 0 \\ 0 & 0.108 & 0.777 & 0.115 & 0 \\ 0 & 0 & 0.232 & 0.682 & 0.086 \\ 0 & 0 & 0 & 0.286 & 0.714 \end{pmatrix}$$

between the hidden states. This, together with our inferred means, reflects the often referenced mosaic structure of hominid genome sequences: Broad troughs of regions with low GC content that transition into regions with means that are almost comparable, and narrow rugged peaks of regions with high GC with more frequent transitions into neighbouring states with increasingly differentiated means (Fig. 2B, Fig. 6A). More specifically, an average of $29, 9, 5, 3$ and 4 successive windows fall within the same hidden state for increasing GC proportions respectively. An exception to this landscape is the left arm of chromosome 1 that exhibits a wide region of high GC content (see Fig. 6A).

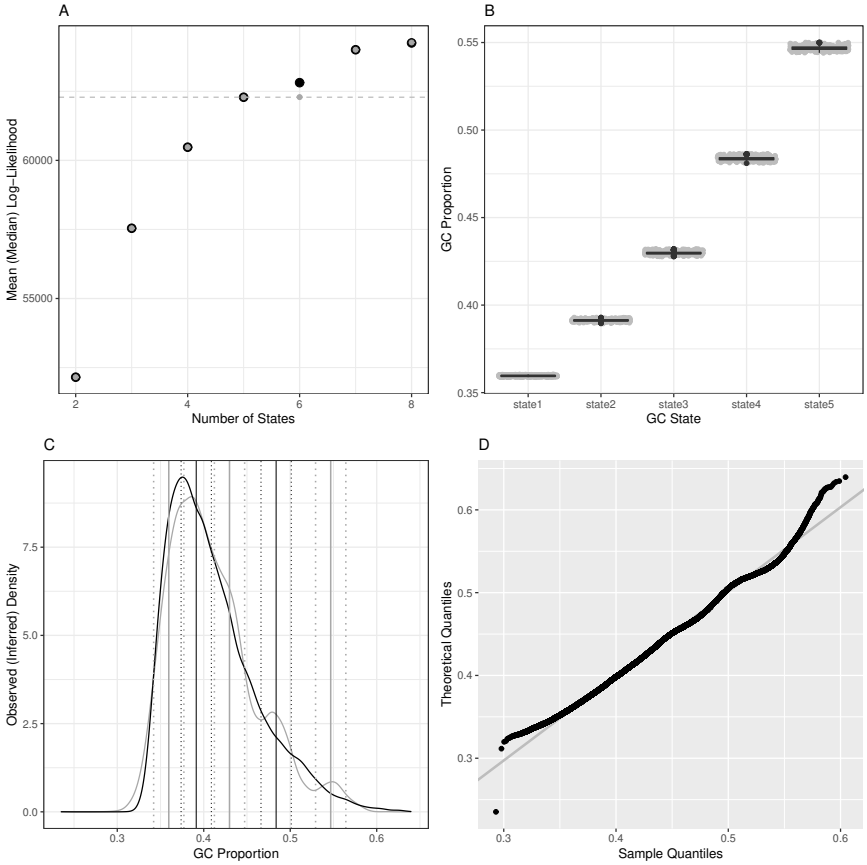

Figure 2: Summarised diagnostics for annotation of the human genome by average GC proportion using oHMMed with normal emission densities. Panel A shows the mean (black) and median (grey) log-likelihood of fully converged runs of the algorithm with different numbers of hidden states, with the dashed horizontal line marking the selected number of hidden states, which is five. The difference between the mean and median for six hidden states is the effect of autocorrelation in the traces of the estimated parameters. In panel B, boxplots of the posterior (*i.e.,* inferred) mean GC proportion of the run with the five hidden states are presented. Panel C shows the observed overall density (black) of the GC proportion superimposed on the posterior (inferred) density, with the inferred means per chosen number of states plus the 68% confidence intervals drawn in vertical lines. The final panel D shows the QQ-plot of the observed density vs. the posterior density (here termed the theoretical distribution). Full descriptions of the diagnostics available for oHMMed can be found in our usage recommendations on GitHub, and the code for this visual summary is available at as an R script.

Gene Content. We used oHMMed with gamma-poisson emission densities to segment the human genome according to the number of protein coding genes per consecutive window several times assuming $K = (2, .., 5)$ hidden states (all runs with 40000 iterations and a 12.5% burn-in). Our diagnostics suggest $K = 3$ hidden states, since this is where the increase in log-likelihood compared between the runs begins to taper off and discrimination between the state-specific means is statistically possible (see Fig. 3). A reasonable fit to the observed histogram of overall counts is achieved by the (smoothed) theoretical curve inferred by this model (see Fig. 4); there is some overestimation of the occurrence of zero counts and underestimation of the single counts.

The landscape of protein coding genes is not as varied as that of the GC proportion: Longer regions of both low and high gene counts occur with a slight abundance of the former; all show similar propensity to transitioning to neighbouring states (see Fig. 6B), as is further evidenced by the inferred transition rate matrix:

$$\begin{pmatrix} 0.960 & 0.040 & 0 \\ 0.048 & 0.938 & 0.014 \\ 0 & 0.062 & 0.938 \end{pmatrix}.$$

Note that this translates to an average of $31, 20$ and 19 subsequent windows being assigned the same state in order of increasing gene content. The inferred means of $0.094, 0.751, 2.80$ per state with the proportions of windows assigned to them being $0.499, 0.412, 0.089$ respectively indicate that genes are generally sparse with mostly only one or none at all per window.

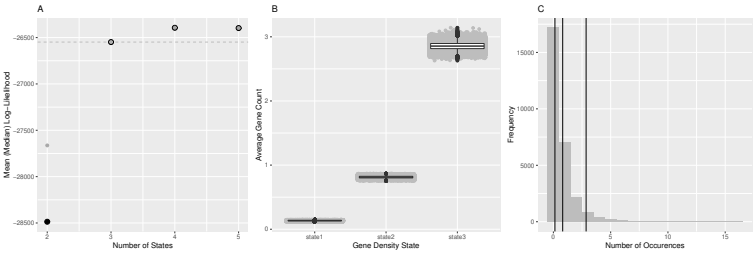

Figure 3: Here we show the first part of the summarised diagnostics for oHHMed with gamma-poisson emission densities as employed on counts of the average number of protein coding genes along the human genome. Panel A shows the mean (black) and median (grey) log-likelihood of fully converged runs of the algorithm with different numbers of hidden states, with the dashed horizontal line marking the chosen number - which is three. In panel B, boxplots of the posterior (i.e. inferred) mean gene densities of the inference run with three hidden states are presented. Panel C shows the observed distribution of gene counts with the inferred means superimposed as vertical lines. These are significantly different on the 95% confidence level as per one-sided poisson rate test (part of standard oHMMed output). Once again, full descriptions of the diagnostics available for oHMMed can be found in our usage recommendations on GitHub, and the code for this visual summary plus the corresponding rootogram below is available at as an R script.

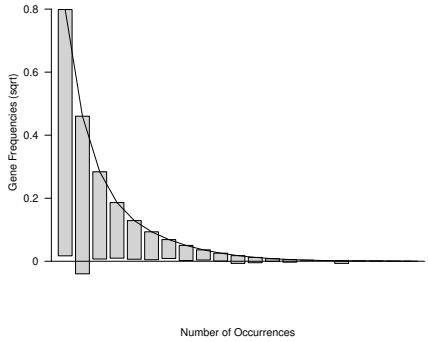

Figure 4: As the final part of the summarised diagnostics for oHMMed with gamma-poisson emission densities and three hidden states as applied to the protein coding genes in humans, we present the above rootogram: The bars represent the observed frequency of counts (square root transformed), and they have been shifted so that the top of each bar aligns with the (smoothed) distribution inferred by oHMMed. Deviations can therefore be assessed by checking the distance of the lower end of each bar to the x-axis.

Correlation. Although the GC proportion and the density of protein coding genes appear to vary on a different scale, counting the number of occurrences of each of the 5 GC states within the 3 states for gene density reveals a clear positive correlation (see Fig. 5A), and compare [3, Fig. 7]. In fact, the position-specific posterior means of the GC content and the gene density, *i.e.,* taking the sum over the inferred means times the probabilities of being assigned to the respective state at each position along the genome, have a highly significant ($p < 2.2e^{-16}$) positive Spearman correlation coefficient of 0.608. Running versions of this correlation can be applied to screen for anomalous regions (see Fig. 6C, and note the negative correlations around most telomeres).

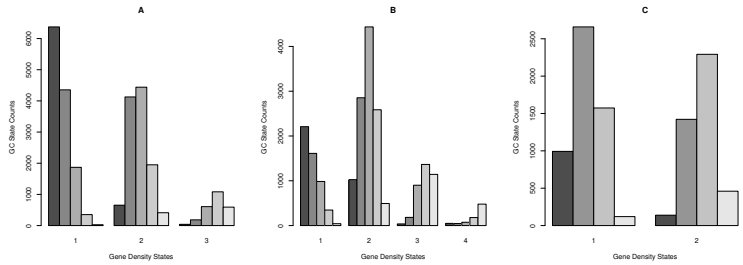

Figure 5: Barplots that visualise the cross-tabulation of the two genome annotation results per species: For the full genome of the human (A), mouse (B), and fruit fly (C) respectively, we count the number of windows assigned to each GC state (y-axis) within regions assigned to each gene density state (x-axis). Essentially, this is a discretised representation of the positive correlation between these features.

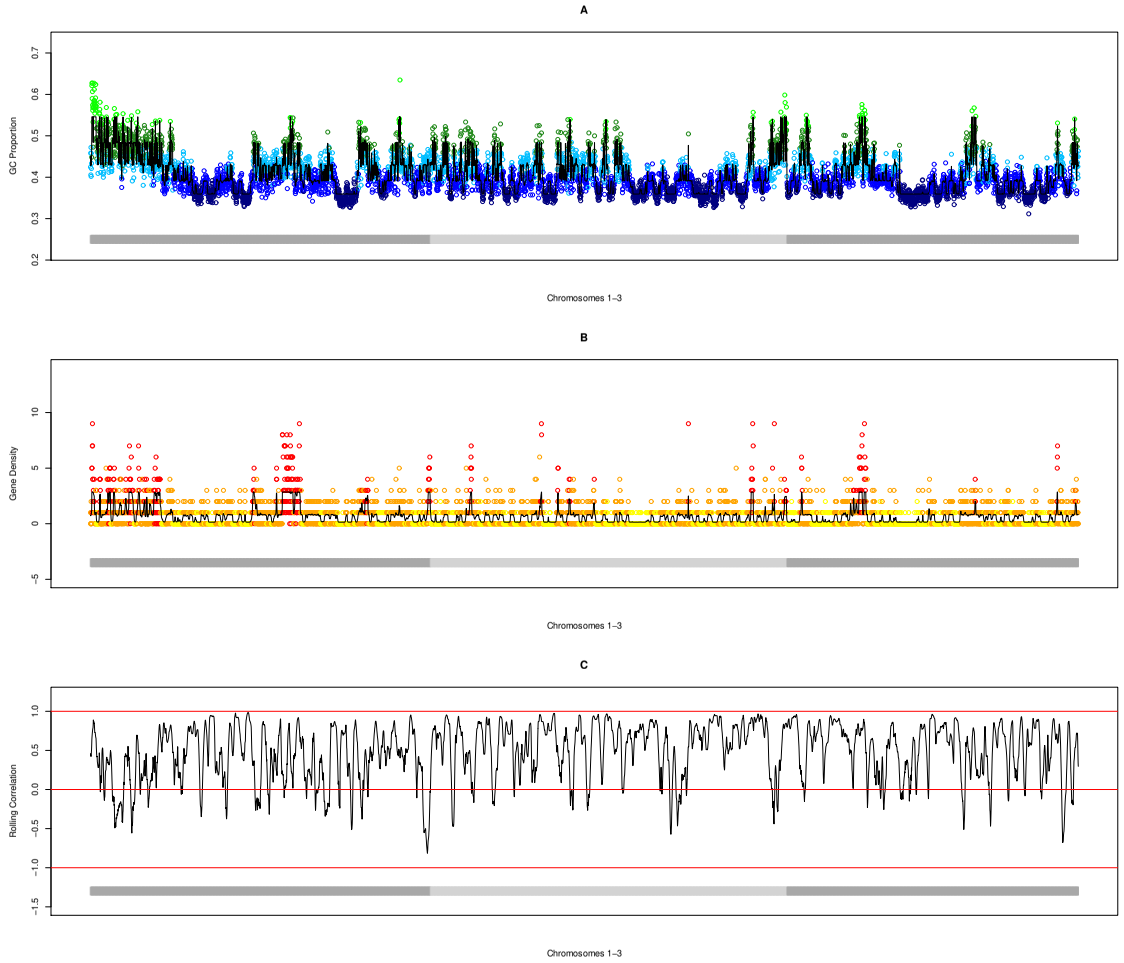

Figure 6: In each of the above panels A-C, the human chromosomes $1-3$ (demarked by alternating dark and light grey horizontal bars) are plotted for different oHMMed analyses: In A, the average GC proportion in every 100kb window is coloured by the oHMMed-inferred GC state. The three lower states are in blue and the two higher ones in green, with the shades lightening with higher GC proportion. In B, the number of protein coding genes for every 100kb window are shown in colours corresponding to the oHMMed-inferred gene density states: yellow, orange, and red mark increasing gene density states. Note that in both A and B, the black lines trace the posterior (inferred) means returned by oHMMed with normal and gamma-poisson emissions respectively. These position-specific posterior means are the sum of estimated means times the respective probabilities of each state, thus combining both estimated mean values and the algorithm's certainty of the assigned state. In C, Spearman's correlation for the two posterior means in shown in rolling windows of 40 collated 100kb windows.

5.3. Segmentation of Mouse and Fly Genomes

The genomic landscape of the mouse is generally more homogeneous and compact than that of humans, with comparatively narrower troughs and less rugged peaks in the variation of GC proportion and a wider range in gene counts per window. Segmentation results in regular, distinct genomic regions ($K = 5$ for the GC proportion and $K = 4$ for the gene content, with the same number of iterations and burn-in percentage as previously; see Supplementary Data B Section "Segmentation of the Mouse Genome"), and a cleaner discretised correlation pattern (see Fig. 5B).

In the much more condensed genome of the fruit fly, clean segmentation based on smaller genomic windows is achieved with fewer hidden states; the results are primarily indicative of chromosomal structure. Specifically, we infer $K = 4$ for the GC proportion and $K = 2$ for the gene content (again with the same number of iterations and burn-in percentage; but see Supplementary Data B Section "Segmentation of the Fruit Fly Genome" for details). Despite comparatively less quantifyable variation in either feature, positive correlation between them is still apparent (see Fig. 5C).

5.4. Comparison to HMM with Unordered States

Recall that, in contrast to HMM algorithms with unordered states, the development of oHMMed has: (i) Reduced the variability of state-specific emitted densities to one of two parameters, (ii) restricted the transitions between hidden states so that the transition rate matrix is tridiagonal, and (iii) used reversibility to determine the prior distributions of states. Thus, we are left with $2K - 2 + (K + 1)$ estimable parameters rather than the full $K(K + 2) - 1$.

Despite greater flexibility, the unordered versions of the oHMMed algorithms perform either no better than oHMMed (in the case of normal emissions) or worse than oHMMed (in the case of poisson-gamma emissions) in terms of the overall log-likelihood and percentage of correctly assigned states when the algorithms are pitted against each other on simulated sequences with inherent autocorrelation (see Supplementary Data C Section "Results: Summary and Interpretation"). While no model selection criterion for comparing HMMs with fixed numbers of states but different numbers of parameters exist, it is readily apparent that classic model selection criteria that penalise overall model fit by model complexity as determined via the (effective) number of parameters (such as AIC, BIC, and DIC) would favour oHMMed to guard against over-fitting.

More importantly, oHMMed is designed to prioritise segmentation into regions with statistically different mean emissions. Indeed, it finds such partitions more readily than its unordered counterpart, particularly on shorter sequences than typically used in full genome analysis (again, see simulations in Supplementary Data C Section "Results: Summary and Interpretation"). For the human genomic sequences analysed in this article, which are very long, both methods yield similar estimates (see Supplementary Data C Section "Results: Hominid Data").

6. Discussion

In this article, we developed algorithms for characterising the large scale variation in genomic data. Part of the inspiration was provided by the clear visual indication that the genomic GC proportion of hominids likens a continuous, reversible random walk with a finite number of re-occurring changes in mean. Furthermore, when partitioning the genome into windows, anecdotal evidence abounds that [8]: "very large GC differences at borders [... are] rare, thus leading to the formation of blocks [...] from closer [... GC levels]". The continuous nature of this observed pattern contrasts with some formulations of the long-standing "isochore theory", which postulates homogeneous stretches of (five or sometimes four) discrete states (aka "isochores") with sharp boundaries between them. Based on "isochore theory", the program IsoFinder [24] and related methods [29] use a binary decision rule to sequentially "slice" genomes into piece-wise constant sections of different means; they have been heavily criticised for finding "isochores" that are not actually there [18]. A closer fit to the observed continuous pattern of variation is provided by Fearnhead and Vasileiou [14], whose Bayesian Online Changepoint model simultaneously infers all the points within the observed sequence at which the underlying assignment of hidden states changes from one to another through direct simulation. The algorithm infers a smoothed mean GC content across the observed sequence by averaging over the likelihood of assignment to "isochore states", and does so in a run-time that scales quadratically with sequence length.

Differently to all the above, the central aspect of oHMMed is definitive sequence segmentation or annotation that is agnostic to the number of hidden states, as well as to the causal forces of the observed sequence pattern. Despite using piece-wise constant approximations to the data purely for methodological reasons, we are able to model autocorrelation patterns by ordering hidden states according to the means of their emission densities and restricting transitions to neighbouring states. This latter assumption is novel compared to other models and seems well-founded empirically. However, it necessitates equal variances between state-specific emitted distributions, which is generally not given. Differences in variability between genomic regions, that oHMMed cannot capture directly, affect the inferred number of hidden states and the estimation of transition rates, conditional on the values of the prior and initial standard deviations: For instance, one state with a large standard deviation may also be modelled as two rapidly alternating states with the same small standard deviation. Therefore, the model has a limited ability to compromise; note that accommodating certain standard deviations may involve deviating from our recommendations for setting priors and initial values, which are set agnostically to benefit convergence stability. Nonetheless, oHMMed specifically captures the pattern of stochastic variation with underlying regime changes observed in genomic sequences of GC proportions that is not fully represented in the harshest formulations of "isochore theory", and provides a descriptive quantification of these patterns. Further, the approach results in a simple model within the traditional HMM framework familiar to most biologists; the accompanying suite of

diagnostics, particularly for finding the appropriate number of hidden states, is straightforward and intuitive. In fact, the lack of complexity makes oHMMed preferable over even a standard unordered HMM with the same underlying MCMC sampler in terms of model fit, particularly when the emission densities are not normally distributed. Since its run-time scales linearly with sequence length, application to long sequences is feasible.

Beyond oHMMed's appeal in methodological tractability and descriptive analyses, we would like to emphasise the interpretability of its output: Since hidden states can be definitively compared by their mean emission densities (as it is the only metric they differ in), segmentation of sequences into regions with statistically significant average patterns of variation is possible. Recall that these states can therefore also be "label matched" between runs, both on the same and on different data sets. Since the causal forces underlying broad scale genomic patterns, as well as the context they provide for finer scale effects, are often still debated (e.g. [30] for the hominid GC proportion), it is crucial to be able to extract distinct genomic regions for subsequent analyses as to these potential driving forces. We propose oHMMed as a powerful, assumption-free tool for this.

Lastly, we would like to stress that oHMMed is a generalised method, which distinguishes it from often more complex and fine-tuned methods developed for specific genomic features (*e.g.,* the GC proportion [24] or the recombination rate [33]). In this article, we developed oHMMed with normal emission densities and applied it to the window-based genomic GC proportion of humans, mice, and fruit flies; however, it could also be employed for e.g. sequences of average recombination rates per genomic window (after normalising transforms), or window-based measures of epigenetic marker density (pending, of course, checks for autocorrelation in these features). Further, oHMMed with gamma-poisson emission densities, which we applied to the protein-coding genes of humans, mice, and fruit flies, could likely be run on many other sequences of window-based count data such as the number of promoters, enhancers, or repressors. Note that altering the size of the genomic windows in all cases could form the basis of comparative studies of genomic variation of the same feature on different spatial scales [30]; the windows in this article were set to best illustrate considerable patterned variation proportional to genome length. Since oHMMed makes no biological assumptions and the inferred hidden states can be ordered, it also facilitates analyses of associations between genome segmentations performed according to different features. In this article, we are able to show the positive correlation between genome annotations by GC proportion and protein-coding gene content. This is done by simply cross-tabulating the assignments of genomic windows to hidden states, as well as by correlating the position-specific posterior means; running versions of the latter can be applied to screen for regions of interest.

Obviously, application of oHMMed is not restricted to genomic data: Any time series data that exhibit the appropriate autocorrelation pattern and conform to the required overall emission distribution can efficiently be segmented into statistically distinct regions using oHMMed; other fields of application may

include ecology or indeed also econometrics.

7. Author Contributions

Concept: CV, CK, JB, LCM;
Planning, Coordination: LCM, CV;
Method - Theory, Usage, Interpretation: LCM, CV;
Method - Implementation, Debugging: CV, LCM, contributions by MK;
Method - Optimisation, Package Development: MM;
Method - Comparisons: LCM, MK, CV;
Data Handling and Analyses: LCM, BY, JB, HK, CV;
Writing: LCM, CV, contributions by MK;
Editing and Manuscript Approval: All authors.

8. Acknowledgements

CV and BY were supported by the the Austrian Science Fund (FWF; W1225-B20); MK and HK were supported by the the Austrian Science Fund (FWF; SFB F6101 and F6106). This work was also partially funded by the Vienna Science and Technology Fund (WWTF) (10.47379/MA16061 to CK). LCM's research was funded by the School of Biology at the University of St.Andrews.

References

[1] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

[2] Boys, R., Henderson, D., and Wilkinson, D. (2000). Detecting homogenous segments in DNA sequences by using hidden Markov models. *Applied Statistics*, **49**, 269–285.

[3] Cammarano, R., Costantini, M., and Bernardi, G. (2009). The isochore patterns of invertebrate genomes. *BMC Genomics*, **10**, 538.

[4] Celeux, G. and Durand, J. (2008). Selecting Hidden Markov Model State Number with Cross- Validated Likelihood. *Computational Statistics*, **23**, 541–564.

[5] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computational Chemistry*, **16**, 107–115.

[6] Cohen, N., Dagan, D., Stone, L., and Graur, D. (2005). GC composition of the human genome: in search of isochores. *Molecular Biology and Evolution*, **22**, 1260–1272.

[7] Costa, M. and DeAngelis, L. (2010). Model selection in hidden Markov models: a simulation study. *Quaderni di Dipartimento, Department of Statistics, University of Bologna*, **7**, ISSN 1973–9346.

[8] Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Research*, **16**, 536–541.

[9] Cunningham, F., Allen, J., Allen, J., Alvarez-Jarreta, J., and Amode, R. e. a. (2022). Ensembl 2022. *Nucleic Acids Research*, **50 (1)**, D988–D995.

[10] Cuny, G., Soriano, P., Macaya, G., and Bernardi, G. (2000). The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Genome Research*, **14**, 886–892.

[11] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis*. Cambridge University Press, Cambridge.

[12] Eyre-Walker, A. and Hurst, L. (2001). The evolution of isochores. *Nature Reviews Genetics*, **2**, 549–555.

[13] Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society B*, **69**, 589—605.

[14] Fearnhead, P. and Vasileiou, D. (2009). Bayesian Analysis of Isochores. *Journal of the American Statistical Association*, **104**, 132–141.

[15] Frankish, A., Diekhans, M., and Jungreis I, e. a. (2021). GENCODE 2021. *Nucleic Acids Research*, **10.1093/nar/gkaa1087**, 49(D1):D916–D923.

[16] Gales, M. and Young, S. (2007). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, **1**, 195––304.

[17] Gramates, L., Agapite, J., Attrill, H., Calvi, B., Crosby, M., dos Santos, G., Goodman, J., Goutte-Gattat, D., Jenkins, V., Kaufman, T., Larkin, A., Matthews, B., Millburn, G., Strelets, V., and Consortium, F. (2022). FlyBase: A guided tour of highlighted features. *Genetics*, **220(4)**, iyac035.

[18] Graur, D. (2009). Slaying (yet again) the brain-eating zombie called the "Isochore Theory": A segmentation algorithm used to "confirm" the existence of isochores creates "isochores" where none exist. *International Journal of Molecular Sciences*, **23**, 6558.

[19] Hamilton, J. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, **57(2)**, 357–384.

[20] Jasra, A., Holmes, C., and Stephens, D. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, **20(1)**, 50–67.

[21] Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and D, H. (2002). The human genome browser at UCSC. *Genome Research*, **12(6)**, 996–1006.

[22] Lander, E. S., Linton, L. M., Birren, B., and Et, A. L. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

[23] Li, W., Bernaola-Galván, P., Carpena, P., and Oliver, J. (2003). Isochores merit the prefix 'iso'. *Computational Biological Chemistry*, **27**, 5–10.

[24] Oliver, J. L., Carpena, P., Hackenberg, M., and Bernaola-Galvan, P. (2004). IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Research*, **32**, 287–292.

[25] Peshkin, L. and Gelfand, M. (1999). Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, **15**, 980–986.

[26] Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP magazine*, **3**, 4–16.

[27] Renals, S. and Hain, T. (2010). *Computational Linguistics and Natural Language Processing Handbook*, chapter Speech Recognition. Blackwell, NY, USA.

[28] Salmenkivi, M., Kere, J., and Mannila, H. (2002). Genome segmentation using piecewise constant intensity models and reversible jump MCMC. *Annals of Mathematical Statistics*, **18**, S211–S218.

[29] Schmidt, T. and Frishman, D. (2008). Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biology*, **9**, R104.

[30] Spencer, C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. *PLOS Genetics*, **2**, e148.

[31] van den Burg, G. J. J. and Williams, C. K. I. (2020). An Evaluation of Change Point Detection Algorithms. *arXiv:2003.06222v3*.

[32] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., and Et, A. L. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351.

[33] Wang, Y. and Rannala, B. (2009). Population genomic inference of recombination rates and hotspots. *Proceedings of the National Academy of Sciences USA*, **106**, 6215–6219.

[34] Yonekura, S., Beskos, A., and Singh, S. (2021). Asymptotic analysis of model selection criteria for general hidden Markov models. *Stochastic Processes and their Applications*, **132**, 164–191.