oHMMed

Alkouri Wassim and Godin Maximilien
Université Libre de Bruxelles

May 2025

Abstract

Motivation. Genomic signals such as base composition or mutation burden vary smoothly along chromosomes. Classical Hidden Markov Models (HMMs) can detect these domains but suffer from label-switching and over-parameterised transition matrices that allow biologically unrealistic long-range jumps.

Methods. *oHMMed* is an ordered HMM whose hidden states are ranked by their emission means, restricted to neighbour-only transitions, and share a common dispersion parameter. These constraints eliminate label-ambiguity, cut the free transition parameters from K^2 to $2K-2$, and preserve biological realism.

Results. Applied to (i) GC content in *Arabidopsis thaliana* and (ii) somatic SNV counts in a TCGA breast-cancer genome, a five-state oHMMed model is able to recover the main AT/GC domains and mutation hotspots, respectively, using far fewer parameters than an unconstrained HMM.

Conclusions. oHMMed delivers compact, stable and interpretable segmentations for any autocorrelated genomic track and is available as an open-source `R` package.

Code. The code we used to get these results can be found on this github repository: https://github.com/Wassim-AlKhouri/bioinfo.git

1 Introduction

Genomes are spatially heterogeneous: base composition, recombination rate, gene density and expression all fluctuate along chromosomes. Dividing the sequence into homogeneous blocks such as "GC-rich domains", "gene deserts" or "recombination hot zones" helps biologists interpret these patterns.

Hidden Markov Models (HMMs) are the work-horse for genome segmentation, but standard HMMs have three well-known problems:

1. **Over-parameterised transitions.** A fully connected $K \times K$ matrix permits biologically implausible long-range jumps and introduces K^2 free parameters.

2. **Label switching.** The K hidden states are *a priori* unlabeled, so Markov-chain Monte Carlo (MCMC) samplers can swap labels between iterations, rendering posterior summaries meaningless.

3. **Choosing K.** Selecting an appropriate number of states is difficult because likelihood typically increases with K while interpretability decreases.

We address these issues with **oHMMed**: an ordered HMM that (i) ranks states by their emission means, (ii) allows transitions only between neighbouring states, and (iii) shares one dispersion parameter across states. We demonstrate oHMMed on *A. thaliana* GC content and human breast-cancer mutation counts.

2 Materials and Methods

2.1 Data

- **Human breast cancer (TCGA-BRCA).** Masked somatic-mutation MAF files aligned to the GRCh38 reference were downloaded from the

NCI Genomic Data Commons. Using `maftools`, single-nucleotide variants (SNVs) were extracted for sample `TCGA-BH-A201-01A-11D-A14K-09`. The autosomes (chr1–22) were divided into non-overlapping 100 kb windows with `GenomicRanges`, retaining windows with $\geq 90\%$ high-quality bases, and SNVs were counted per window.

- ***Arabidopsis thaliana* (TAIR10).** The TAIR10 reference genome was downloaded in FASTA format. Nuclear chromosomes 1–5 were chopped into non-overlapping 100 kb windows (Biostrings + IRanges) and windows with $\geq 90\%$ sequenced bases were kept. GC proportion was computed for each window.

2.2 Method

2.2.1 Conceptual idea

oHMMed models an autocorrelated numeric sequence with an HMM in which the hidden states can be placed in a natural order.

The model has two key properties:

1. **Convex emission densities.** All states share one scale parameter (σ for normal, α for gamma), so the log-ratio of any two state densities is convex, allowing the states to be linearly ordered by their means or rates.
2. **Neighbour-only transitions.** Because states are ordered, biological changes are assumed to be gradual: the $K \times K$ transition matrix T is tridiagonal, making the hidden chain reversible with stationary distribution π.

2.3 Model Overview

2.3.1 General notation

Let $y_{1:L} = (y_1, \ldots, y_L)$ be the observed sequence of continuous or count data along the genome, and let $\theta_{1:L} = (\theta_1, \ldots, \theta_L)$ be the corresponding hidden states, where $\theta_l \in \{1, \ldots, K\}$. The hidden states form a reversible Markov chain with transition matrix $T = (T_{ij})_{1 \leq i,j \leq K}$, so that

$$\Pr(\theta_{l+1} = j \mid \theta_l = i) = T_{ij},$$

and the stationary distribution π satisfies $\pi T = \pi$ and detailed balance $\pi_i T_{ij} = \pi_j T_{ji}$, ensuring reversibility.

2.3.2 Convex emission densities

Each state θ_i emits observations according to a probability $Pr(y \mid \theta_i)$, where

- **Normal emissions:** The emissions are drawn from a normal distribution $N(\mu_i, \sigma)$, with common σ and ordered means $\mu_1 \leq \cdots \leq \mu_K$.
- **Gamma–Poisson emissions:** The emissions are drawn from $G(\alpha, \beta_i)$, with common shape α and ordered rate $\beta_1 \geq \cdots \geq \beta_K$.

Under these shared–parameter assumptions, the log-ratio

$$\frac{Pr(\mathbf{y} \mid \theta_i)}{Pr(\mathbf{y} \mid \theta_j)}$$

is a convex function of y when $i < j$. And if $(y_i, y_j) \in \mathbf{y}$ such that $y_i < y_j$ then

$$\frac{Pr(y_i \mid \theta_i)}{Pr(y_i \mid \theta_j)} \leq \frac{Pr(y_j \mid \theta_j)}{Pr(y_j \mid \theta_i)}$$

ensuring a natural ordering of states by their emission means or rates.

2.3.3 Neighbour-only transitions

By imposing $T_{ij} = 0$ for $|i - j| > 1$, T becomes tridiagonal with only $2K - 2$ free parameters. This "neighbour–only" constraint encodes the assumption of gradual, locally autocorrelated changes along the genome and, together with detailed balance, guarantees reversibility of the hidden chain.

2.3.4 Inference via MCMC Gibbs sampler

Parameters and hidden states are jointly inferred with a Gibbs sampler that alternates:

1. *Forward–backward sampling:* Given current T and emission parameters, sample the path $\theta_{1:L}$ via the HMM forward–backward algorithm.

2. *Emission updates:* Conditioned on $\theta_{1:L}$, update state-specific parameters $\{\mu_i\}$ (or $\{\beta_i\}$) from their conjugate posteriors, then sort them to enforce $\mu_1 \leq \cdots \leq \mu_K$ (or $\beta_1 \geq \cdots \geq \beta_K$).

3. *Transition updates:* Given $\theta_{1:L}$, count transitions C_{ij} and sample the nonzero entries of T from a symmetric Dirichlet prior plus C_{ij}, ensuring tridiagonality and detailed balance.

This sampler avoids label-switching because the ordering constraint anchors state labels, and yields full posterior distributions for all parameters and hidden paths.

2.3.5 Selecting the number of states

Models are fit for $K = 2, \ldots, 9$. We choose the smallest K that maximizes predictive fit (median log-likelihood) while retaining well-separated state means (assessed via posterior credible intervals and t-tests).

2.3.6 Output

For each position l we report the maximum-a-posteriori state $\hat{\theta}_l$ and its marginal posterior probability. Contiguous runs of identical $\hat{\theta}_l$ define homogeneous segments, with segments of $\min p < 0.6$ flagged as low confidence.

3 Results

3.1 Verifying oHMMed assumptions

To confirm autocorrelation we compared the variance of first-order differences

$$d_i = y_{i+1} - y_i$$

to the same quantity after shuffling the windows.

Data set	$\widehat{\sigma^2}_{\text{original}}$	$\widehat{\sigma^2}_{\text{shuffled}}$
GC proportion	3.77×10^{-4}	34.57
SNV counts	7.91×10^{-4}	48.30

F-tests give $p < 2.2 \times 10^{-16}$ in both cases, confirming strong autocorrelation.

3.2 Segmentation of *Arabidopsis thaliana*

We fitted oHMMed with normal emissions to GC proportion measured in non-overlapping 100 kb windows and compared models with $K = 2, \ldots, 9$ hidden states. A five-state model (8 000 iterations; 2 000 burn-in) maximised the *mean* log-likelihood, whereas the *median* log-likelihood peaked marginally at $K = 6$. Because the gain from $K = 5$ to $K = 6$ was negligible (Fig. 1), and inspection of the component densities showed that the sixth state simply split one of the five existing components into two nearly identical ones (Fig. 2), we selected $K = 5$ for downstream analysis.

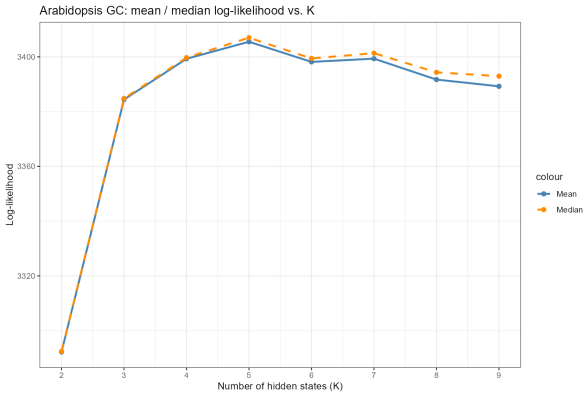
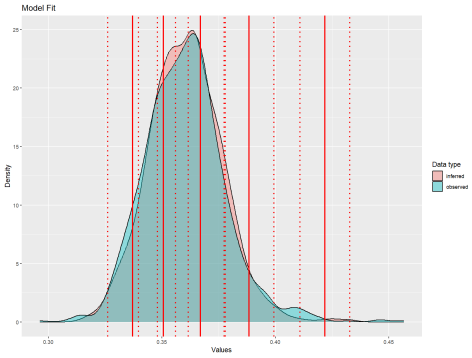


Figure 1: Mean (blue) and median (orange) log-likelihood of converged oHMMed runs as a function of the number of hidden states K.

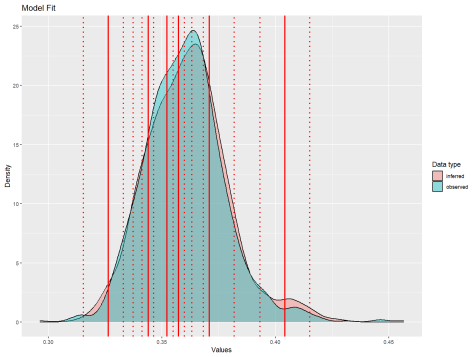
For $K = 5$, oHMMed collapses the windows into five homogeneous domains with mean GC proportions 0.337, 0.350, 0.366, 0.382, and 0.415, respectively, sharing a common standard deviation of 0.0111. The genomic fractions assigned to each state are 5.6 %, 38.5 %, 47.2 %, 6.0 %, and 2.5 %.

Figure 3 shows the state track across the five chromosomes (blue = AT-rich, purple = GC-rich). Most of the genome belongs to state 3, close to the species-wide mean GC content. AT-rich domains (state 1) form longer tracts, whereas strongly GC-rich domains (state 5) are short and focal.

Most features in the five-state track align with the window-less Z-curve map of Zhang & Zhang [2]:



(a) $K = 5$



(b) $K = 6$

Figure 2: Observed GC-proportion density (grey) overlaid with posterior component densities for (a) $K = 5$ and (b) $K = 6$. Vertical lines show posterior means with 68 % credible intervals.

- **Centromeric islands.** All five centromeres appear GC-rich, classified as states 4–5.
- **AT isochores.** Seven broad AT isochores are captured almost entirely by states 1–2.
- **GC isochores.** The three GC isochores on chromosomes 3, 4, and 5 appear as golden blocks (state 5) in our segmentation.

Where the two methods differ, oHMMed resolves finer-scale structure, revealing subtle composition changes that the Z-curve misses. Exact coordinates for all seven AT isochores, three GC isochores, and five centromere isochores are provided in Zhang & Zhang [2].

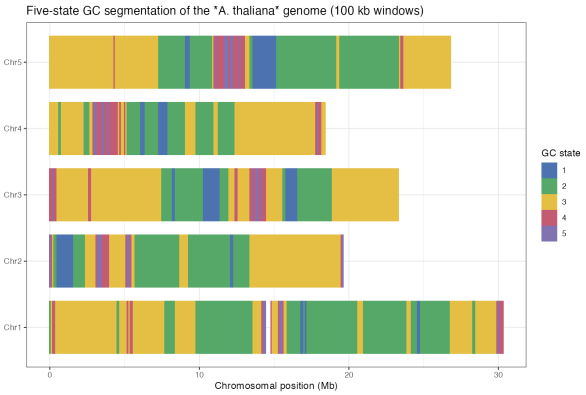


Figure 3: Genome-wide five-state segmentation (100 kb windows) of the *A. thaliana* genome.

3.3 Segmentation of Human breast cancer (TCGA-BRCA)

We applied oHMMed with a gamma–Poisson emission model to the SNV counts in non-overlapping 100 kb windows across chromosomes 1–22 of a single TCGA-BRCA tumor (sample `TCGA-BH-A201-01A-11D-A14K-09`). Models with $K = 2, \ldots, 9$ hidden states were each run for 8 000 iterations (2000 warmup), and predictive fit was assessed by the median log-likelihood across the MCMC samples. As shown in Fig. 4, the gain in log-likelihood plateaus after $K = 5$ and decreases slightly for larger K, indicating that a 5-state model captures genuine structure without over-fitting.

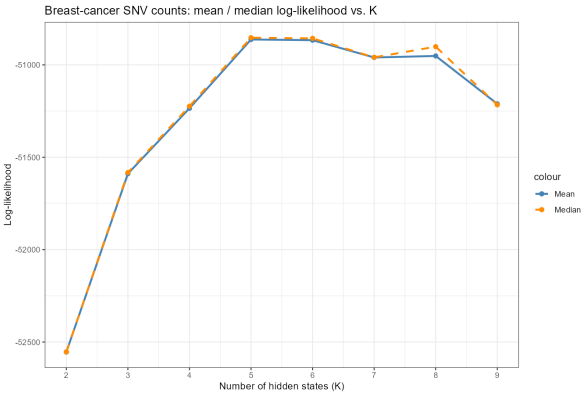


Figure 4: Mean (blue) and median (orange) log-likelihood of the converged oHMMed runs as a function of the number of hidden states K, on breast-cancer SNV counts. The elbow at $K = 5$ indicates the optimal balance between model complexity and fit.

Under the 5-state model, posterior emission means (average SNV counts per 100 kb

window) increase monotonically from state 1 (lowest burden) to state 5 (highest burden). The genome-wide distribution of the most-likely state $\hat{z}_t$ is plotted in Fig. 5. Key observations:

- **Background vs. hotspots.** States 1–3 (low to moderate burden) cover the vast majority of windows (85%), whereas states 4–5 (high/ultra-high burden) occupy only 15%, highlighting focal clusters of hypermutation.
- **Notable focal peaks.** Dark maroon state 5 windows form extended tracts on chromosome 17 and specially on chromosome 19 where most of it has been mutated.
- **Contiguous blocks.** Thanks to the tridiagonal transition prior, each state forms long, contiguous segments rather than isolated single-window calls, consistent with the strong autocorrelation in adjacent-window SNV counts.

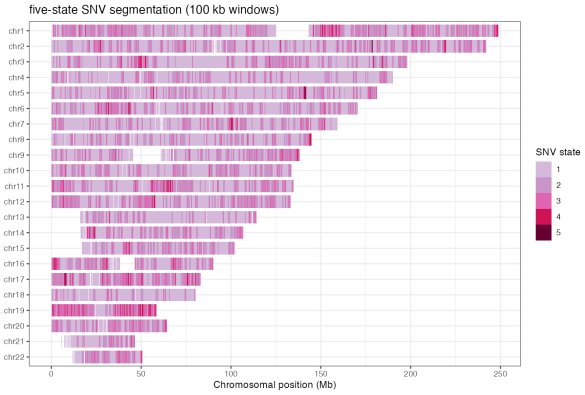


Figure 5: Five-state segmentation of breast-cancer SNV counts in 100 kb windows. Each horizontal line is one chromosome (1–22), and windows are colored by their most-likely oHMMed state (1 = pale pink, ..., 5 = dark maroon).

Overall, oHMMed produces a concise "mutation-landscape" map of the tumor genome, objectively partitioning it into low, moderate, and high mutation-density domains.

4 Conclusion

This study applied **oHMMed** to two contrasting genomic data sets, GC-content variation across the *Arabidopsis thaliana* genome and somatic SNV counts in a human breast-cancer tumour, and showed that its three structural constraints (ordered states, neighbour-only transitions, and a shared dispersion parameter) yield interpretable, biologically coherent segmentations.

- **Interpretable domains.** On *A. thaliana*, oHMMed condensed the genome into five GC classes that recover the classical AT- and GC-isochores of Zhang & Zhang [2]. In the tumour genome, the same five-state model separated background mutation burden from focal hyper-mutated tracts on chromosomes 17 and 19 (Fig. 5), providing a compact "mutation landscape" overview.
- **Stable inference with few parameters.** Ordering the states eliminates label-switching, and the tridiagonal transition matrix reduces the number of free transition parameters from K^2 to $2K-2$. As a result, the Gibbs sampler converged smoothly even on the long breast-cancer sequence ($> 50{,}000$ windows) without the instability typically seen in unconstrained HMMs.
- **Objective choice of K.** For both data sets the median log-likelihood increased sharply up to $K = 5$ and then plateaued (Figs. 1, 4). Because oHMMed suppresses the splitting of nearly identical components, this "elbow" provided a clear, data-driven criterion for selecting the smallest K that still distinguishes biologically meaningful states.
- **Versatility.** Without changing the core algorithm we switched from normal emissions (GC proportions) to gamma–Poisson emissions (SNV counts), underscoring oHMMed's potential as a general tool for any autocorrelated genomic track.

Limitations and future work. The shared-dispersion assumption that guarantees convex

ordering may be violated when true biological variance differs markedly between states.

Outlook. Across two very different applications, oHMMed delivered concise, stable, and biologically interpretable segmentations without the over-parameterisation or label ambiguity that often hinder classical HMMs. We therefore recommend it as a robust default for genome-wide block finding and anticipate its adoption for high-resolution epigenomic and single-cell assays where ordered, neighbour-aware models are essential.

References

[1] C. Vogl, M. Karapetiants, B. Yıldırım, *et al.*, "Inference of genomic landscapes using ordered Hidden Markov Models with emission densities (oHMMed)," *BMC Bioinformatics*, vol. 25, p. 151, 2024, doi:10.1186/s12859-024-05751-4.

[2] Zhang, Ren, and Chun-Ting Zhang. *Isochore structures in the genome of the plant* Arabidopsis thaliana. *Journal of Molecular Evolution* **59**(2):227–238, 2004. doi:10.1007/s00239-004-2617-8.