Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach

Meijun Gao

Michigan State University
Dept. of Computer Science and Engineering
East Lansing, Michigan, U.S.A.
gaomeiju@msu.edu

Kevin J. Liu

Michigan State University
Dept. of Computer Science and Engineering
East Lansing, Michigan, U.S.A.
kjl@msu.edu

ABSTRACT

Genetic recombination in eukaryotes can occur with or without crossover, where the latter event is referred to as gene conversion. New discoveries in the genomic and post-genomic era have shed new light into the complex interplay between recombination and other evolutionary processes such as point mutations. In particular, G/C content of genomic regions can increase over evolutionary time due to recombination in the form of gene conversion – a phenomenon known as GC-biased gene conversion (gBGC) – and gBGC is increasingly appreciated as serving an important role in genome evolution throughout the eukaryotic Tree of life. These findings have largely relied on computational advances for analyzing recombinant sequences for indirect signatures of gBGC. However, deeper insights into the functional and evolutionary significance of gBGC require a unified framework that accounts for variable-across-sites recombination and point mutation processes.

In this study, we introduce PHYNCH (or “PHYlogeNetIC-HMM for analyzing gBGC and recombination hotspots”). PHYNCH utilizes a statistical model that combines a hidden Markov model to capture local genealogical variation due to recombination and gene conversion with a finite-sites model of sequence evolution along a local genealogy. Inference and learning under the new model is used to detect and analyze local patterns of gBGC and recombination hotspots within genomic sequences. We validate the performance of PHYNCH using simulated benchmarking data. Furthermore, we use PHYNCH to create a new genomic map of gBGC and recombination in rice.

CCS CONCEPTS

• **Applied computing** → *Computational genomics; Computational biology; Molecular sequence analysis; Molecular evolution; Computational genomics; Bioinformatics; Population genetics.*

KEYWORDS

recombination, GC-biased gene conversion, phylogeny, phylogenetic, hidden Markov model, rice

ACM Reference Format:

Meijun Gao and Kevin J. Liu. 2021. Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21)*, August 1–4, 2021, Gainesville, FL, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3459930.3469509>

INTRODUCTION

In eukaryotes, meiotic recombination can occur either with or without crossover between recombining chromosomes. Following the terminology used by [15], we refer to the former as a recombination event and the latter as a gene conversion event; both are understood to have played central roles in the evolution of eukaryotic genomes.

The spatial distribution of recombination and gene conversion events varies locally within genomes and can be concentrated in genomic regions with up to multiple orders of magnitude higher recombination rate compared to surrounding regions [19]. In humans, for example, around 80% of all recombination breakpoints are concentrated in less than 15% of the genome [4].

Recombination can also strongly affect base composition and substitution processes, as in the case of GC-biased gene conversion (gBGC) where G/C-content of recombining DNA regions increases over evolutionary time due to gene conversion during recombination [9, 12]. Prior systematic studies have provided evidence of gBGC in animals [22, 23, 29] and plants [8]. For example, Muyle et al. [28] found that gBGC affects the GC content of third codon positions and intronic regions in grass genomes. The studies indicate that the interplay of gBGC with other evolutionary forces has shaped important eukaryotic genomic features including local recombination breakpoint distributions, the ancestral tract landscape, and local base composition heterogeneity. The studies also provide mechanistic insights into the functional consequences of gBGC on key cellular processes including transcription and translation. For example, evidence suggests that gBGC biases tRNA abundance [13] and can play functional role in elevating point mutation rates [20].

Advances in high-throughput biomolecular sequencing and the resulting explosion of large-scale genomic and other -omics data [26, 34] have enhanced efforts to comprehensively study gBGC across a diverse range of organisms. These studies have also relied on computational and statistical approaches to detect and analyze genomic signatures left by the evolutionary processes under study. Many of these methods were originally developed for analyzing recombinant sequences and related biomolecular sequence analysis

This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.

BCB '21, August 1–4, 2021, Gainesville, FL, USA

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8450-6/21/08.

<https://doi.org/10.1145/3459930.3469509>

tasks without necessarily focusing on gBGC [6, 7, 17, 18, 24, 25]. But the complex interplay of recombination and mutation processes is inherent to gBGC, and joint analysis within a single unified framework may yield insights that may be inaccessible when reasoning about one in isolation of the other. A well-known example is the confounding effect that gBGC has upon traditional approaches for inferring bio-molecular signatures of adaptive evolution (e.g., substitution rate-based analysis) [3]. One of the few methods available for model based analysis of gBGC in genomic sequences is the method of Capra et al. [3]. The method has the advantage of also analyzing evolutionary conservation alongside gBGC, but is defined for a fixed 4-taxon phylogeny.

MATERIALS AND METHODS

To address this gap, we introduce PHYNCH (or “PHYlogeNetiC-HMM for analyzing gBGC and recombination hotspots”), a phylogenetic hidden Markov model-based framework for detecting and analyzing genomic patterns of gBGC and recombination hotspots in eukaryotic genomes. The framework utilizes a combined statistical model for analyzing recombination and mutation processes that jointly vary across sites. The framework is well-suited to fine-scale mapping and is not restricted to a specific number of taxa (i.e., groups of organisms under study) or fixed phylogenetic hypothesis.

The contributions of our study are summarized as follows. First, the new PHYNCH framework accounts for linked variation across sites for recombination and mutation processes. The framework adopts an explicit phylogenetic model to compare and analyze biomolecular sequences. Second, PHYNCH is a general phylogenetic-HMM framework that is well-suited to fine-scale mapping. In theory, the framework will support analysis of inputs with an arbitrary number of taxa (although see Discussion for additional practical considerations). Finally, we apply the PHYNCH framework to an empirical rice genomic sequence dataset. The analysis provides a new high-resolution map of gBGC and recombination hotspots in the rice genome.

Problem definition

For the purposes of clarity, we begin by defining a special case of the computational problem addressed in our study. (The full computational problem is defined in a subsequent subsection.) Let G be a set of aligned genomes $g_1, g_2, \dots, g_n$, each of length k , and the multiple sequence alignment A with dimension $n \times k$ has rows consisting of the genomes and columns consisting of aligned sites. Let A_i be the i^{th} site in the alignment, which corresponds to the i^{th} column in the matrix. Every site A_i has evolved under a local genealogy (i.e., evolutionary history of a site). Note that recombination can cause local genealogies to vary across different sites – both in terms of topology and branch length. Hotspots evolved under recombination and point mutation processes with higher evolutionary rates and/or different base frequency distributions compared to background regions, and local genealogies in the former can be expected to exhibit greater local variation compared to the latter.

Let $\Delta(n)$ be the set of all unrooted binary tree topologies on n leaves. Let $T_b(n)$ be the set of $|\Delta(n)|$ unrooted binary trees on n leaves, each of which a distinct topology from $\Delta(n)$ and branch

lengths ℓ_b . The set of trees $T_b(n)$ are used to model local genealogies of sites in “background” regions of A that evolved under a baseline evolutionary model. Note that any tree $t \in T_b(n)$ could be the local genealogy of a given site A_i – i.e., $P(A_i|t, \theta_b) \geq 0$ where θ_b corresponds to the parameters of a finite-sites substitution model in background regions (e.g., the general-time reversible model (GTR) [33] or nested models such as the HKY85 model [14]). Local genealogies can change along the genome with different rates due to the existence of hotspots. Let $T_h(n)$ be the set of $|\Delta(n)|$ trees, each with a distinct topology from $\Delta(n)$ but with different branch lengths ℓ_h . Together with a different substitution model θ_h , the set of trees $T_h(n)$ are used to model local genealogies in “hotspot” regions of A that evolved under a non-baseline evolutionary model; the hotspot model captures local genealogical discordance and local base composition bias that are generated by GC-biased gene conversion and recombination hotspots. Given a genomic sequence alignment A , local trees $T_b(n) \cup T_h(n)$, and the substitution model instances θ_b and θ_h , we define a sequence of n random variable q_i , each of which takes on an ordered pair of values $(x, y) \in (T_b(n) \times \{\theta_b\}) \cup (T_h(n) \times \{\theta_h\})$. We then define the computational problem as follows.

- **Input:** A genomic sequence alignment A consisting of n aligned sequences and k sites.
- **Output:** For each site $1 \leq i \leq k$, the per-site hotspot probability

$$P(q_i = (x, y)|A, \theta)$$

where $(x, y) \in (T_b(n) \times \{\theta_b\}) \cup (T_h(n) \times \{\theta_h\})$ and the PHYNCH model instance θ includes a set of local tree models $T_b(n)$ and $T_h(n)$, and substitution model instances θ_b and θ_h .

The solution to this problem includes statistical inference of hotspot regions within the input genomic sequences. The site A_i is located in a hotspot region when $q_i = (x, y)$ where $x \in T_h(n)$ and $y = \theta_h$ and in a background region otherwise. The problem outputs also provide fine-scale annotation of recombination breakpoints and local recombination and substitution model variation (including GC-content variation) down to single-site resolution. Note that the problem addresses soft inference since q_i is a random variable and every ordered pair (x, y) has (possibly non-zero) probability $P(q_i = (x, y)|A)$.

PHYNCH model

PHYNCH utilizes a phylogenetic hidden Markov model (HMM) to analyze genomic signatures of GC-biased gene conversion and recombination hotspots within the set of input sequences. To facilitate discussion, we begin by considering a special case where the input involves $n = 4$ taxa and one genomic sequence is sampled per taxon (i.e., group of organisms under study); otherwise, no restrictions are placed on sample relatedness and an out-group is not needed. Furthermore, we restrict our discussion to the above two-class problem where genomic regions fall into either background or hotspot categories. (See the following subsection for a more general formulation of the PHYNCH model.)

In this case, the number of all possible unrooted binary tree topologies for genomes is $|\Delta(n)| = 3$, and the PHYNCH model would include a total of $1 + 2|\Delta(n)| = 7$ states: a start state s_0 , and six additional states. The latter consists of background states b_i

for $1 \leq i \leq 3$ – each corresponding to one of the three possible local gene tree topologies that can appear in a background genomic region – and hotspot states h_i for $1 \leq i \leq 3$ similarly for hotspot regions. Let $g(b_i)$ be the distinct gene tree topology associated with background state b_i ; the state b_i includes a local gene tree model that consists of both the topology $g(b_i)$ and a set of branch lengths $\ell(b_i)$. Each hotspot state h_i includes a local gene tree model with topology $g(h_i)$ and branch lengths $\ell(h_i)$ similarly. The resulting set of states is shown in Figure 1. Note that each distinct topology appears twice in the model – once in a background state and once in a hotspot state – but possibly with differing branch lengths in the two different states in which it appears.

Figure 1: Illustration of PHYNCH model states. The top row consists of states b_i for $1 \leq i \leq 3$ that correspond to local genealogies – each having a distinct unrooted topology – that are observed in background genomic regions. The bottom row consists of states h_i that similarly correspond to hotspot regions. The model also includes a start state s_0 . To simplify the illustration, transitions and model parameters are not shown (see Methods for more details).

Based on the row/column arrangement of states in Figure 1’s illustration, “within-row” transitions between different b states correspond to topological discordance of local genealogies across recombination breakpoints within a background genomic region; within-row transitions between different h states correspond to recombination breakpoints within a hotspot region similarly. “Across-row” transitions from a b state to an h state or vice versa correspond to background and hotspot region boundaries.

The PHYNCH model includes the following transition, initial state, and emission probabilities and parameters. Similar to the HMM proposed by Westesson and Holmes [36], the within-row transition probability parameter ϵ_b accounts for the level of local topological discordance due to recombination in background regions; higher parameter values are correlated with increased recombination breakpoint frequency. Transitions from a given state to a different state in the background row of states occur with equal probability. Within-row transition probabilities for hotspot regions and the probability parameter ϵ_h are defined similarly. The within-row transition probability parameters are specified such that $1 - \epsilon_h \geq 1 - \epsilon_b$ WLOG. Across-row transition probabilities are parameterized by a switching parameter γ . The switching parameter

accounts for local variation in substitution and recombination processes and is correlated with the frequency of background/hotspot-delineating breakpoints. Transitions from a given state in one row to any other state in a different row occur with equal probability. Transitions from the start state s_0 to any other state occur with equal probability as well. For convenience, we define the following probabilities where $1 \leq i \leq 3$ and $1 \leq j \leq 3$:

$$\begin{aligned} f_{s_0} &= \frac{1}{2|\Delta(n)|} \\ f_{b_i} &= f_{h_i} = \frac{1}{|\Delta(n)|} \\ \delta(b_i, b_j) &= \begin{cases} \frac{1-\epsilon_b}{|\Delta(n)|-1} & \text{if } i \neq j \\ \epsilon_b & \text{otherwise} \end{cases} \\ \delta(h_i, h_j) &= \begin{cases} \frac{1-\epsilon_h}{|\Delta(n)|-1} & \text{if } i \neq j \\ \epsilon_h & \text{otherwise} \end{cases} \\ t_{b_i, b_j} &= (1 - \gamma)\delta(b_i, b_j) \\ t_{h_i, h_j} &= (1 - \gamma)\delta(h_i, h_j) \end{aligned}$$

The transition probability matrix is then specified as follows, with rows labeled by the states $s_0, b_1, b_2, b_3, h_1, h_2, h_3$ from top to bottom, columns labeled from left to right similarly, and each cell containing the probability of transitioning from a row’s corresponding state to a column’s corresponding state:

$$\begin{bmatrix} 0 & f_{s_0} & f_{s_0} & f_{s_0} & f_{s_0} & f_{s_0} & f_{s_0} \\ 0 & t_{b_1, b_1} & t_{b_1, b_2} & t_{b_1, b_3} & \gamma f_{h_1} & \gamma f_{h_2} & \gamma f_{h_3} \\ 0 & t_{b_2, b_1} & t_{b_2, b_2} & t_{b_2, b_3} & \gamma f_{h_1} & \gamma f_{h_2} & \gamma f_{h_3} \\ 0 & t_{b_3, b_1} & t_{b_3, b_2} & t_{b_3, b_3} & \gamma f_{h_1} & \gamma f_{h_2} & \gamma f_{h_3} \\ 0 & \gamma f_{b_1} & \gamma f_{b_2} & \gamma f_{b_3} & t_{h_1, h_1} & t_{h_1, h_2} & t_{h_1, h_3} \\ 0 & \gamma f_{b_1} & \gamma f_{b_2} & \gamma f_{b_3} & t_{h_2, h_1} & t_{h_2, h_2} & t_{h_2, h_3} \\ 0 & \gamma f_{b_1} & \gamma f_{b_2} & \gamma f_{b_3} & t_{h_3, h_1} & t_{h_3, h_2} & t_{h_3, h_3} \end{bmatrix} \quad (1)$$

The initial state of the PHYNCH model is always the start state s_0 .

An individual state b_i for $1 \leq i \leq 3$ in a background region emits a site pattern A_z for $1 \leq z \leq k$ according to an emission probability $P(A_i | g(b_i), \ell(b_i), \theta_b)$ under the site’s local genealogy model with tree topology $g(b_i)$, branch lengths $\ell(b_i)$, and substitution model θ_b . Emission probabilities are defined similarly for each hotspot state h_i . Under the finite-sites substitution models in this study, the emission probability of a site pattern at a given state can be efficiently calculated using dynamic programming, as in the peeling algorithm used in traditional phylogenetic MLE [10, 11]. Our experiments utilize the HKY85 model of nucleotide substitution [14], and the GTR model [33] and other nested models are readily substituted.

We also adopt a modeling choice that is intended to reduce model complexity, improve learning, and mitigate overfitting. Branch lengths in background state b_i for $1 \leq i \leq 3$ are shared with hotspot state h_i and scaled by scaling factor parameter β such that $\ell(h_i) = \beta \ell(b_i)$. The scaling parameter β therefore controls the relative evolutionary divergence of background versus hotspot regions.

Likelihood calculations under a fixed PHYNCH model instance and posterior decoding can be performed efficiently using dynamic programming in the form of the peeling algorithm for emission

probability calculations and the forward and backward algorithms [30]. PHYNCH model parameters are learned using statistical optimization under the maximum likelihood criterion. A variety of heuristics are commonly used to address this computationally difficult problem [30]. We utilize a hill-climbing algorithm coupled with Brent’s method for univariate optimization [2] in our experiments. Statistical inference is addressed using a modified posterior decoding algorithm. The algorithm annotates each site A_i with the probability $\sum_{(x,y) \in (T_h(n) \times \{\theta_h\})} P(q_i = (x, y) | A, \theta)$ that it falls within a hotspot region.

Finally, our simulation study included model selection experiments that coupled PHYNCH inference and learning with a likelihood ratio test (LRT). Two nested models were evaluated in each test: an alternative model that consisted of a standard PHYNCH model (i.e., a PHYNCH model that includes both background and hotspot states) and a null model that consisted only of background states. The alternative model includes eight additional parameters compared to the null model: ϵ_h , γ , β , and θ_h , where the HKY model used for the latter contributes four base frequency parameters and one substitution rate parameter. Both models were fitted using MLE in a manner identical to the other PHYNCH analyses in our study.

General problem formulation and model

The PHYNCH model and framework is readily extended to more general inputs. First, inputs with n sequences require $\Delta(n)$ to consist of the set of all binary tree topologies on n taxa, and the number of states in a row will equal $|\Delta(n)|$. Furthermore, the PHYNCH model can be extended to handle $r > 2$ classes of recombination and substitution models beyond the background/hotspot models considered in our study; the PHYNCH model extensions will create a corresponding row of states for each of the r rate classes, along with “within-row” and “between-row” transitions and associated parameters analogous to the simpler formulation above. In this study, we focus on two-class background/hotspot model where $r = 2$.

Simulation study

We used msHOT [16] to simulate local coalescent histories under an extended multi-species coalescent model that incorporated hotspots, where the latter exhibited elevated recombination, mutations, and GC content bias relative to background regions. Note that PHYNCH’s phylogenetic HMM makes use of a sequentially Markovian approximation for both the coalescent-with-recombination model and related gene conversion models [15, 37, 38], and the msHOT simulations utilize the latter full models. Each simulation sampled 4 or 5 individuals from a panmictic population. Either 0, 1 or 2 hotspots were simulated per dataset, where the one-hotspot simulations utilized fixed placement between 2000 and 4000 bp and the two-hotspot simulations utilized fixed placements between 1000 and 3000 bp and between 4000 and 4500 bp. The scaled recombination rate ρ in background and hotspot regions was set to 5.0 and 50.0, respectively. The scaled mutation rate θ in background and hotspot regions was set to 1.0 and 10.0, respectively. Coalescent times were converted into branch lengths under finite-sites substitution model (see equation 3.1 from [15]).

Then, seq-gen [31] was used to simulate sequence evolution on the local gene tree corresponding to each local coalescent history; sequence evolution proceeded under finite-sites substitution model. Our study utilized the HKY85 model [14] for the latter. The substitution model parameter values for background and GC-biased hotspot regions were based on empirical analyses of the rice genomic sequence dataset in our study (see below for dataset details). First, the GC content histogram for annotated genes was used to assess the bimodal nature of base composition bias in the rice genome. The assessment was used to partition genes into two GC content categories – either high or low – based on a visually assessed fixed threshold. For each of the two sets of genes – low and high – RAxML [35] was used to perform concatenated phylogenetic MLE under the HKY model; the estimated substitution rates and base frequencies were used in the seq-gen simulations of background and hotspot regions, respectively. The resulting background model instance consisted of base frequencies $\pi_A = 0.267$, $\pi_G = 0.213$, $\pi_C = 0.200$, $\pi_T = 0.320$ and a transition/transversion rate of 1.855; the hotspot model instance consisted of base frequencies $\pi_A = 0.162$, $\pi_G = 0.341$, $\pi_C = 0.347$, $\pi_T = 0.150$ and a transition/transversion rate of 2.058. The total simulated sequence length of each 4-taxon and 5-taxon dataset were 5 kb and 2 kb, respectively.

For each model condition, the simulation procedure was repeated to obtain 20 replicate datasets. Model condition parameters and summary statistics for simulated datasets are shown in Table 1.

PHYNCH’s performance on the simulated benchmarking data was assessed using multiple performance criteria. A modified posterior decoding was used to perform statistical inference (see above). We evaluated type I and type II error by comparing PHYNCH’s soft inference that a site is located within a hotspot region versus ground truth (i.e., the true hotspot location(s)): the type I and type II error comparisons were based on receiver operating character (ROC) curves, precision-recall (PR) curves, area under ROC curve (ROC-AUC), and area under PR curve (PR-AUC). A true positive is a site that has PHYNCH posterior decoding probability above a fixed threshold and is actually located within a true hotspot region, a true negative is a site that has PHYNCH posterior decoding probability below a fixed threshold and is located outside of any true hotspot region, a false positive is a site that has PHYNCH posterior decoding probability above a fixed threshold but is located outside of any true hotspot region, and a false negative is a site that has PHYNCH posterior decoding probability below a fixed threshold but is located within a true hotspot region; varying the threshold yields different points along the ROC and PR curves. We also assessed PHYNCH’s computational runtime and peak main memory usage.

Rice genomic sequence dataset analysis

We used PHYNCH to analyze an empirical dataset that consisted of whole genome sequence data for two rice subspecies – *Oryza sativa japonica* and *O. s. indica* – and two sister species – *O. rufipogon* and *O. nivara*. We downloaded whole-genome sequences and gene annotations from Ensembl Plants [5]. The accessions for the *O. s. japonica*

Table 1: Simulation study: model condition parameters and summary statistics for simulations involving recombination. The model condition parameters consists of the number of taxa (“Num taxa”), the simulated sequence length (bp), the scaled recombination rate for background and hotspot regions (“Scaled recomb rate bkgd:hot”), the scaled mutation rate for background and hotspot regions (“Scaled mut rate bkgd:hot”), and the number of hotspot regions and their location(s) (“Num hotspot(s)” and “Hotspot location(s)”, respectively). We also report the average normalized Hamming distance of the simulated MSA for each replicate dataset (“ANHD”) and the average recombination breakpoint frequency normalized by sequence length (“Recomb bkpt freq”) ($n = 20$).

Model condition	Num taxa	Sequence length (bp)	Scaled recomb rate bkgd:hot	Scaled mut rate bkgd:hot	Num hotspot(s)	Hotspot location(s)	ANHD	Recomb bkpt freq
4.R.0	4	5000	5:NA	1:NA	0	NA	0.6502	0.0047
4.R.1	4	5000	5:50	1:10	1	[2000,4000]	0.6760	0.0219
4.R.2	4	5000	5:50	1:10	2	[1000,3000],[4000,4500]	0.6782	0.0248
5.R.0	5	2000	5:NA	1:NA	0	NA	0.6303	0.0053
5.R.1	5	2000	5:50	1:10	1	[2000,4000]	0.6654	0.0226
5.R.2	5	2000	5:50	1:10	2	[1000,3000],[4000,4500]	0.6727	0.0275

Table 2: Simulation study: model condition parameters for simulations involving gene conversion. Figure layout and description are identical to Table 1, other than an additional parameter contributed by the model of gene conversion: mean gene conversion tract length (“Mean GC tract length (bp)”).

Model condition	Num taxa	Sequence length (bp)	Scaled recomb rate bkgd:hot	Scaled mut rate bkgd:hot	Num hotspot(s)	Hotspot location(s)	Mean GC tract length (bp)
4.G.0	4	5000	5:NA	1:NA	0	NA	50
4.G.1	4	5000	5:50	1:10	1	[2000,4000]	50
4.G.2	4	5000	5:50	1:10	2	[1000,3000], [4000,4500]	50

IRGSP-1.0 assembly, the *O. s. indica* ASM465v1 assembly, the *O. rufipogon* OR_W1943 assembly, and the *O. nivara* Oryza_nivara_v1.0 assembly are GCA_001433935.1, GCA_000004655.2, GCA_000817225.1, and GCA_000576065.1, respectively. MAFFT with default settings was used to align gene sequences for each gene. For each chromosome, SNP positions in gene MSAs were concatenated and analyzed using PHYNCH. Summary statistics for the empirical dataset are shown in Table 3. Model condition parameters were estimated from the SNP MSAs using the same optimization procedures that were used elsewhere in our study. Soft inference was performed using modified posterior decoding in a manner identical to the simulation study.

Software and data

An open-source software implementation of the PHYNCH framework and the data used in our study are publicly available under permissive copyleft open licenses at https://gitlab.msu.edu/liulab/gBGC_Phylo_HMM.

RESULTS AND DISCUSSION

Simulation study

We began by assessing PHYNCH’s performance in terms of type I and II error of its modified posterior decoding-based inference (i.e., the probability that a site is located within a hotspot region).

Across all of the model conditions in our study, PHYNCH’s inference resulted in receiver operating characteristic (ROC) curves and precision-recall (PR) curves with greater than 0.999 area-under-curve (AUC) on average (Table 4). PHYNCH yielded average accuracy greater than 0.99 on the one- and two-hotspot models conditions with recombination, based on a stringent classification threshold of posterior decoding probability greater than 0.95; accuracy on the model conditions with gene conversion were greater than 0.945 on average. Base frequencies for the background and hotspot substitution models θ_b and θ_h were estimated with error less than 0.005 (Supplementary Tables S1 and S2).

In all of the non-hotspot model conditions in our study, LRT-based model selection consistently avoided rejecting the non-hotspot null model when compared against the PHYNCH alternative model. Median and minimum corrected q-values are listed in Supplementary Table S3 ($n = 20$), and none were statistically significant ($\alpha = 0.05$). LRT-based model selection was similarly effective on the one- and two-hotspot model conditions in our study. All tests returned statistically significant q-values ($\alpha = 0.05$) and the non-hotspot null model was rejected in favor of the PHYNCH alternative model in all cases.

Runtime and main memory usage for PHYNCH analyses are shown in Figure 2. On average, PHYNCH required at most 3 hours and just over 200 MiB to complete analysis of each four-taxon

Table 3: Empirical study: rice genomic sequence dataset statistics. Rice whole-genome sequence (WGS) data was downloaded from Ensembl Plants [5]; total WGS length for each chromosome is shown (“Length (bp)”). Annotated genes were then aligned and concatenated; the number of SNPs, average normalized Hamming distance, and percentage of cells consisting of indels in the concatenated alignment matrix (“Concatenated MSA”) is reported (“Num SNPs”, “ANHD”, and “Gappiness”, respectively).

Chromosome	Length (bp)	Concatenated MSA		
		Num SNPs	ANHD	Gappiness
1	13401096	122210	0.285	0.281
2	10307950	60632	0.280	0.272
3	10712212	70737	0.263	0.255
4	7941634	73186	0.298	0.288
5	6952511	46830	0.260	0.252
6	6540066	41423	0.265	0.259
7	6412536	49462	0.289	0.291
8	5943838	42428	0.282	0.275
9	4249722	29209	0.270	0.267
10	4456037	65809	0.308	0.294
11	4031078	48072	0.325	0.331
12	3504895	26023	0.279	0.275

dataset. PHYNCH’s computational runtime requirements greatly increased on the five-taxon datasets, however – a difference of nearly two orders of magnitude – although its memory usage remained the same. We attribute the increased runtime to the computational difficulty of learning PHYNCH’s model. Learning is addressed using model likelihood maximization, which is already known to be computationally difficult for the statistical models that are integrated into PHYNCH’s phylogenetic HMM [30, 32]. Furthermore, the number of states required for the general formulation of PHYNCH’s model will rapidly grow as the number of input sequences increases, and runtime and main memory usage will increase as well. (Below, we discuss a promising algorithmic approach for addressing anticipated scalability challenges.)

Rice genomic sequence dataset analysis

PHYNCH was used to detect genomic signatures of GC-biased gene conversion and recombination hotspots in rice and two sister species. The resulting fine-scale genomic map is shown in Figure 3. Putative genomic regions exhibiting GC-biased gene conversion and recombination were detected in all 12 chromosomes in the rice genome. Longest regions with high posterior decoding probability appeared around coordinates ~27 Mb to ~28 Mb in chromosome 7 and around coordinates ~17 Mb to ~18 Mb in chromosome 12. Some chromosomes contained more such regions relative to other chromosomes, with the largest number appearing in chromosomes 3, 7, 1, 2, 4, 6, and 12. Genes within PHYNCH-inferred posterior decoding probability above a high threshold are listed in Supplementary Table S4. Panther [27] analysis identified subsets of genes in the gene list that had Gene Ontology (GO) term enrichment (Table 5). Further investigation is needed to test specific molecular hypotheses that have been generated by PHYNCH analysis.

We note that there are several important differences between the empirical study and simulation study. First, the sequence lengths in the empirical study datasets are greater than in the simulation study datasets by 1 to 2 orders of magnitude. PHYNCH inference and

Figure 2: Simulation study: runtime (h) and main memory usage (GiB) of PHYNCH analyses. Average runtime and peak memory usage are reported for simulation conditions with recombination; standard error bars are also shown ($n = 20$).

learning remained tractable on the larger datasets in our empirical

Table 4: Simulation study: type I and type II error of PHYNCH inference. A modified posterior decoding calculation was used to perform soft inference (see Methods section for details). Type I and type II error was assessed based on three calculations: area under receiver operating characteristic curve (ROC-AUC), area under precision-recall curve (PR-AUC), and accuracy based on hard classification with a fixed probability threshold of 0.95; for each calculation, an average across all replicates in a model condition is shown ($n = 20$). Only accuracy is reported for the zero-hotspot model conditions since all sites are background and multiple classes are needed for ROC and PR curves to be well-defined.

Model condition	ROC-AUC	PR-AUC	Accuracy
4.R.0			0.9501
4.R.1	0.9999	0.9999	0.9971
4.R.2	0.9999	0.9999	0.9954
4.G.0			1.000
4.G.1	0.9996	0.9995	0.9461
4.G.2	0.9998	0.9998	0.9486
5.R.0			0.9682
5.R.1	0.9999	0.9999	0.9968
5.R.2	0.9999	0.9999	0.9935

study. Second, the genomic sequences in our simulation study arose through the complex interplay of different evolutionary processes including point mutations, recombination and gene conversion, gBGC, and others. The simulation conditions in our study capture a subset of these processes. Third, empirical genomes include structural and functional features that are not directly captured in our simulations. Finally, the empirical data in our study were obtained using next-generation sequencing. Real-world data acquisition and sequencing can introduce non-negligible error upstream of biomolecular sequence analysis.

CONCLUSION

In this study, we introduced PHYNCH, a new phylogenetic HMM framework for analyzing genomic patterns of gBGC and recombination hotspots. We conducted simulation experiments to evaluate PHYNCH's performance. PHYNCH returned type I and type II error that was largely robust to the range of evolutionary scenarios explored in our simulations, but a primary bottleneck was scalability based on the number of input sequences. We anticipate that larger and more divergent datasets may reveal further performance bottlenecks; future algorithmic enhancements can boost PHYNCH's scalability (see below).

We conclude with directions for future research. First, other evolutionary processes can also cause local genealogical discordance, particularly genetic drift and incomplete lineage sorting. As in related HMM frameworks [7, 17, 24, 25], the PHYNCH model can be augmented with coalescent model-based extensions to account for the latter. Second, we note that not all local genealogies are equally likely, and some may have low or trivial probability. This insight can be exploited to perform model reduction as an approximation technique [39]. We hypothesize that adapting these techniques to the PHYNCH framework will improve computational runtime and main memory usage by orders of magnitude. Finally, appropriate extensions to the PHYNCH model would allow inference that distinguishes between crossover events, gene conversion events, and combinations of both. Our future work aims to explore alternate

transition models for distinguishing between recombination outcomes – either with or without crossover. As a proxy, convolutional neural networks [21] may be able to distinguish between local tract length distributions left by the two processes as different classes of local motifs.

ACKNOWLEDGMENTS

This work was supported in part by the NSF (grant no. 1565719, 1714417, and 1737898), the BEACON Center for the Study of Evolution in Action (NSF STC Cooperative Agreement 093954), and the MSU High Performance Computing Center (HPCC).

REFERENCES

- [1] Y. Benjamini and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57, 1 (1995), 289–300.
- [2] RP Brent and P Richard. 1973. Algorithms for minimization without derivatives. Mineola.
- [3] John A Capra, Melissa J Hubisz, Dennis Kostka, Katherine S Pollard, and Adam Siepel. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet* 9, 8 (2013), e1003684.
- [4] International HapMap Consortium et al. 2005. A haplotype map of the human genome. *Nature* 437, 7063 (2005), 1299.
- [5] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. 2019. Ensembl 2019. *Nucleic acids research* 47, D1 (2019), D745–D751.
- [6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological Sequence Analysis*. Cambridge University Press.
- [7] Julien Y. Duthiel, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K. Uyenoyama, and Mikkel H. Schierup. 2009. Ancestral Population Genomics: The Coalescent Hidden Markov Model Approach. *Genetics* 183, 1 (2009), 259–274. [arXiv:http://www.genetics.org/content/183/1/259.full.pdf+html](http://www.genetics.org/content/183/1/259.full.pdf+html) <http://www.genetics.org/content/183/1/259.abstract>
- [8] Juan S Escobar, Sylvain Glémin, and Nicolas Galtier. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Molecular Biology and Evolution* 28, 9 (2011), 2561–2575.
- [9] Adam Eyre-Walker. 1993. Recombination and mammalian genome evolution. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 252, 1335 (1993), 237–243.
- [10] Joseph Felsenstein. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology* 22, 3 (1973), 240–249.
- [11] J. Felsenstein. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17, 6 (1981), 368–376.

Table 5: Functional categorization of PHYNCH-flagged genes in the empirical rice genomic sequence dataset. Panther [27] was used to analyze the list of genes for Gene Ontology (GO) process category enrichment. The first column contains the name of the GO biological process category; the second contains the number of genes in the reference *O. s. japonica* genome that map to the GO category; the third contains the number of PHYNCH-flagged genes that map to the GO category; the fourth contains the expected number of genes in the gene list for the category; the fifth shows the fold enrichment of the PHYNCH-flagged genes over expectation; the sixth column is the uncorrected p-value as determined by Fisher's exact test; and the last column is the corrected q-value as calculated by Benjamini and Hochberg [1]'s multiple test correction. The table lists results with statistically significant corrected q-value ($\alpha = 0.05$) in order of highest fold enrichment first.

GO Biological Process Category	Reference Genome Num. Genes	PHYNCH-flagged Genes				
		Num. Genes	Expected Num. Genes	Fold Enrich.	Uncorrected p-value	Corrected q-value
Xenobiotic transport	50	4	0.19	21.56	4.93E-3	2.36E-2
RNA metabolic process	1387	16	5.15	3.11	7.01E-5	2.80E-2
Nucleic acid metabolic process	1818	19	6.75	2.82	5.32E-5	2.32E-2
Macromolecule metabolic process	6456	49	23.96	2.05	6.05E-7	4.84E-4
Cellular macromolecule metabolic process	5009	36	18.59	1.94	1.00E-4	3.70E-2
Cellular metabolic process	8398	58	31.16	1.86	1.07E-6	7.30E-4
Organic substance metabolic process	8869	61	32.91	1.85	3.92E-7	4.70E-4
Primary metabolic process	8347	57	30.97	1.84	1.72E-6	1.03E-3
Metabolic process	9757	66	36.20	1.82	2.04E-7	3.26E-4
Nitrogen compound metabolic process	7038	47	26.12	1.80	3.82E-5	2.03E-2
Cellular process	12114	75	44.95	1.67	5.98E-7	5.74E-4
Biological_process	16117	95	59.80	1.59	2.82E-8	6.77E-5
Unclassified	27542	67	102.20	0.66	2.82E-8	1.35E-4

- [12] Nicolas Galtier, Gwenaél Piganeau, Dominique Mouchiroud, and Laurent Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 2 (2001), 907–911.
- [13] Richard J Harrison and Brian Charlesworth. 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu stricto* group of yeasts. *Molecular biology and evolution* 28, 1 (2011), 117–129.
- [14] M. Hasegawa, K. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22 (1985), 160–174.
- [15] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. 2004. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*. Oxford University Press, Oxford.
- [16] Garrett Hellenthal and Matthew Stephens. 2007. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* 23, 4 (2007), 520–521.
- [17] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics* 3, 2 (2007), e7.
- [18] Dirk Husmeier and Frank Wright. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology* 8, 4 (2001), 401–427.
- [19] Alec J Jeffreys, Rita Neumann, Maria Panayi, Simon Myers, and Peter Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nature genetics* 37, 6 (2005), 601–606.
- [20] Dennis Kostka, Melissa J Hubisz, Adam Siepel, and Katherine S Pollard. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Molecular biology and evolution* 29, 3 (2012), 1047–1057.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25 (2012), 1097–1105.
- [22] Florent Lassalle, Séverine Périan, Thomas Bataillon, Xavier Nesme, Laurent Duret, and Vincent Daubin. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet* 11, 2 (2015), e1004941.
- [23] Yann Lezecque, Dominique Mouchiroud, and Laurent Duret. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Molecular biology and evolution* 30, 6 (2013), 1409–1419.
- [24] Thomas Mailund, Julien Y. Dutheil, Asger Hobolth, Gerton Lunter, and Mikkel H. Schierup. 2011. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genetics* 7, 3 (03 2011), e1001319.
- [25] Thomas Mailund, Anders E. Halager, Michael Westergaard, Julien Y. Dutheil, Kasper Munch, Lars N. Andersen, Gerton Lunter, Kay Prüfer, Aylwyn Scally, Asger Hobolth, and Mikkel H. Schierup. 2012. A New Isolation with Migration Model along Complete Genomes Infers Very Different Divergence Processes among Closely Related Great Ape Species. *PLoS Genet* 8, 12 (12 2012), e1003125.
- [26] Michael L Metzker. 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics* 11, 1 (2010), 31–46.
- [27] Huaiyu Mi, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D Thomas. 2021. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* 49, D1 (2021), D394–D403.
- [28] Aline Muyle, Laurana Serres-Giardi, Adrienne Ressayre, Juan Escobar, and Sylvain Glémin. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular biology and evolution* 28, 9 (2011), 2695–2706.
- [29] Eugénie Pessia, Alexandra Popa, Sylvain Mousset, Clément Rezvoy, Laurent Duret, and Gabriel AB Marais. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution* 4, 7 (2012), 675–682.
- [30] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.
- [31] A. Rambaut and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13 (1997), 235–238.
- [32] Sebastien Roch. 2006. A Short Proof That Phylogenetic Tree Reconstruction by Maximum Likelihood Is Hard. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 3, 1 (Jan. 2006), 92–.
- [33] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* 142 (1990), 485–501.
- [34] Jay Shendure and Erez Lieberman Aiden. 2012. The expanding scope of DNA sequencing. *Nature Biotechnology* 30, 11 (2012), 1084.
- [35] Alexandros Stamatakis. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 9 (2014), 1312–1313.
- [36] Oscar Westesson and Ian Holmes. 2009. Accurate Detection of Recombinant Breakpoints in Whole-Genome Alignments. *PLoS Comput Biol* 5, 3 (03 2009),

e1000318. <https://doi.org/10.1371/journal.pcbi.1000318>

- [37] Carsten Wiuf. 2000. A coalescence approach to gene conversion. *Theoretical Population Biology* 57, 4 (2000), 357–367.
- [38] Carsten Wiuf and Jotun Hein. 2000. The coalescent with gene conversion. *Genetics* 155, 1 (2000), 451–462.

- [39] Qiqige Wuyun, Nicholas W VanKuren, Marcus Kronforst, Sean P Mullen, and Kevin J Liu. 2019. Scalable Statistical Introgression Mapping Using Approximate Coalescent-Based Inference. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 504–513.

Figure 3: Empirical study: a PHYNCH-inferred genomic map of gBGC and recombination hotspots in rice. A modified posterior decoding was used to infer the probability that a site was contained within a gBGC and recombination hotspot (see Methods for details). Genome coordinates are based on the IRGSP-1.0 *O. s. japonica* reference genome.

Appendix – Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach

Meijun Gao

Michigan State University
Dept. of Computer Science and Engineering
East Lansing, Michigan, U.S.A.
gaomeiju@msu.edu

Kevin J. Liu

Michigan State University
Dept. of Computer Science and Engineering
East Lansing, Michigan, U.S.A.
kjl@msu.edu

ACM Reference Format:

Meijun Gao and Kevin J. Liu. 2021. Appendix – Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach. In *ACM-BCB '21: ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, August 1–4, 2021, Virtual*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/1122445.1122456>

1 SUPPLEMENTARY METHODS

There are four different software commands were used in simulation performance study. The following commands were used to generate the simulation data:

- (1) r8s version 1.7 [4] was used to generate random trees for the model trees in the simulations:

```
simulate diversemodel=yule_c T=1.0 seed=1023492  
charevol=yes ntaxa=<num-of-taxa> infinite=yes  
nreps=<num-of-replicates>
```

- (2) msHOT [2] was used to generate evolutionary sequences with recombination hotspots. The command is:

```
msHOT <num-of-taxa> 1 -r < $\rho$ > <num-of-sites> -v n  $a_1$   
 $b_1$   $\lambda_1$   $\dots$   $a_n$   $b_n$   $\lambda_n$  -I <num-of-taxa> <1 1  $\dots$  1> <model  
tree coalescent history> -T > <output file>
```

where the ρ is the background recombination rate; n is the number of recombination hotspots; the first hotspot region having intensity λ_1 ($=0, \dots, \infty$) between basepairs a_1 and b_1 (site location, with $1 \leq a_i \leq b_i \leq \text{num-of-sites}$) and the n th one having intensity λ_n between basepairs a_n and b_n . The intensity λ_i refers to the multiple of the background recombination rate by which the recombination rate is increased (or decreased) in the hotspot region. In our simulation, all λ_i were set to 10. The second parameter <1 1

$\dots$ 1> after -I is the sample configuration, which refers to a list of integers ($n_1 n_2 \dots$) indicating the number of chromosomes sampled from each subpopulation. In our simulation, we sampled one individual per population, that means $n_1 = n_2 = \dots = n_{\text{num-of-taxa}} = 1$. The parameter of <model tree coalescent history> was based on the model trees generated by r8s. The output file of msHOT is the local gene trees.

- (3) When we want to investigate the gene conversion and its hotspot but no recombination, the following command can be used. The -c specifies the gene conversion rate and conversion tract length. The -V switch specifies the parameters of gene conversion, where n , a_n , and b_n have the same meaning in -v switch. The intensity λ_n has the similar definition as before, here referring to the multiple of the background gene conversion rate in hotspot region.

```
msHOT <num-of-taxa> 1 -c <conversion rate> <con-  
version tract length> -r 0 <num-of-sites> -V n  $a_1$   $b_1$   $\lambda_1$   
 $\dots$   $a_n$   $b_n$   $\lambda_n$  -I <num-of-taxa> <1 1  $\dots$  1> <model tree  
history> -T > <output file>
```

- (4) seq-gen [3] was used to simulate the evolution of nucleotide sequences along local gene trees, using HKY85 models of the substitution process.

```
seq-gen -m HKY -f <equilibrium frequencies of the  
four nucleotides> -t <transition transversion ratio> -  
of -s <mutation rate> -l <num-of-sites> -p <num-of-  
partitions> <<gene trees>> <seqfile>
```

where -m specifies the substitution model of HKY model, and its parameters were specified by -f and -t. The -l parameter specifies sequence length of 5000 bp and 2000 bp for 4 and 5 taxa, respectively, the -p parameter is the number of local gene trees, the -s parameter specifies the mutation rate (1 for background regions, 10 for hotspot regions), the -z option specifies the pseudo-random number generator seed, and the <gene trees> argument is the list of true gene trees that were output by msHOT.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM-BCB '21, August 1–4, 2021, Virtual

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-9999-9/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

2 SUPPLEMENTARY RESULTS AND DISCUSSION

PHYNCH model parameters are estimated using maximum likelihood optimization (see Methods section for details). Table. S1 and S2 show the estimation results of four base frequencies of the two HKY85 substitution models, M_b and M_h , respectively. We use Pearson correlation coefficient (PCC) between the ground truth and estimated values of the substitution model parameters to evaluate the learning performance. When PCC is closer to 1, the learning performance is better. Regardless of the simulation conditions (with or without hotspots), the base frequencies of M_b and M_h were accurately estimated at the ground truth, except for the results of model condition 4.A and 5.A for M_h in Table. S2. The exceptional ones are the estimated result of M_h under the simulations without hotspot. They are not failed estimation results, since there is no hotspots under this simulation condition and the ground truths are the base frequencies of hotspot substitution model, M_h .

Supplementary Table S1: The performance of the estimated parameters of substitution model M_b . The last column is the Pearson correlation coefficient (PCC) between the ground truth and the estimated value of the four base frequencies.

Model condition	The estimated base frequency of M_b				PCC
	A	C	G	T	
Ground truth	0.267	0.200	0.213	0.320	
4.R.0	0.257	0.210	0.199	0.333	0.976
4.R.1	0.268	0.197	0.214	0.320	0.999
4.R.2	0.272	0.196	0.213	0.318	0.998
5.R.0	0.275	0.194	0.216	0.314	0.991
5.R.1	0.261	0.201	0.212	0.324	0.997
5.R.2	0.264	0.202	0.209	0.323	0.998

In order to explore the empirical data further, the fine-grained posterior decoding plots of the 163 hotspot genes were shown in Fig. S1, S2, S3, S4, and S5. The y-axis shows the probability of hotspot. The x-axis shows the physical position along the chromosome of the given gene (beginning and ending at the start and end location of the given gene). Furthermore, we display the gene's name in the upper right corner. In each sub-figure, we only plot the probability of hotspots in the MSAs' SNP sites since only the SNP sequences were the input of the PHYNCH (in order to relieve the computational load). Hence, some of the probability curves do not cover the whole x-axis.

REFERENCES

- [1] Y. Benjamini and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57, 1 (1995), 289–300.
- [2] Garrett Hellenthal and Matthew Stephens. 2007. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* 23, 4 (2007), 520–521.

Supplementary Table S2: The performance of the estimated parameters of substitution model M_h . The last column is the Pearson correlation coefficient (PCC) between the ground truth and the estimated value of the four base frequencies.

Model condition	The estimated base frequency of M_h				PCC
	A	C	G	T	
Ground truth	0.162	0.347	0.341	0.150	
4.R.0	0.282	0.201	0.231	0.285	−0.963
4.1	0.162	0.344	0.343	0.151	0.999
4.2	0.161	0.342	0.343	0.152	0.999
5.0	0.308	0.172	0.264	0.256	−0.650
5.1	0.159	0.348	0.342	0.150	0.999
5.2	0.162	0.351	0.339	0.147	0.999

Supplementary Table S3: Simulation study: PHYNCH model selection experiments. We used a likelihood ratio test (LRT) to select between two nested models: an alternative model consisting of the PHYNCH model and a background-only null model (see Methods for details). Multiple test correction was performed using the method of [1]. Averages ("Avg") and standard errors ("SE") of uncorrected p-values and corrected q-values are reported ($n = 20$).

Model condition	Corrected q-value	
4.R.0	0.750	
4.R.1	$< 10^{-10}$	$< 10^{-10}$
4.R.2	$< 10^{-10}$	$< 10^{-10}$
5.R.0	0.600	
5.R.1	$< 10^{-10}$	$< 10^{-10}$
5.R.2	$< 10^{-10}$	$< 10^{-10}$

- [3] A. Rambaut and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13 (1997), 235–238.
- [4] Michael J Sanderson. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 2 (2003), 301–302.

Supplementary Figure S1

Supplementary Figure S2

Supplementary Figure S3

Supplementary Figure S4

Supplementary Figure S5

Supplementary Table S4: The summary of the 163 hotspot genes, I. After getting the decoded hotspot sites, the corresponding genes of these hotspot sites were obtained according to the genome annotation. Hotspot genes are those whose hotspot sites exceed 95% of its SNP length. The HSP means the percentage of hotspot sites in gene's SNP sequences

RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID	
Os01g0124900	LOC_Os01g03400.1 LOC_Os01g03410.1	Os01g0205100	LOC_Os01g10790.1	Os01g0550000	LOC_Os01g36920.2	Os01g0566500	LOC_Os01g38580.1 LOC_Os01g38580.2
Os01g0574500	LOC_Os01g39260.1	Os01g0595900	LOC_Os01g41270.1	Os01g0618500	LOC_Os01g43130.1	Os01g0721000	LOC_Os01g52290.1
Os01g0721100	LOC_Os01g52304.1	Os01g0721200	LOC_Os01g52320.1	Os01g0781100	LOC_Os01g57270.1 LOC_Os01g57270.2	Os01g0798500	LOC_Os01g58470.1
Os01g0827500	LOC_Os01g61180.1	Os01g0897200	LOC_Os01g67180.1	Os01g0897300	LOC_Os01g67190.1	Os01g0927000	LOC_Os01g70220.1
Os02g0102300	LOC_Os02g01220.1	Os02g0137400	LOC_Os02g04480.1	Os02g0137500	LOC_Os02g04490.1	Os02g0137600	LOC_Os02g04500.1 LOC_Os02g04500.2
Os02g0137700	LOC_Os02g04510.1	Os02g0137800	LOC_Os02g04520.1 LOC_Os02g04520.2	Os02g0143200	LOC_Os02g05040.1	Os02g0144800	LOC_Os02g05190.1
Os02g0146400	LOC_Os02g05310.1 LOC_Os02g05310.2 LOC_Os02g05310.3 LOC_Os02g05310.4	Os02g0248800	LOC_Os02g15070.1	Os02g0288700	LOC_Os02g18700.1	Os02g0288925	LOC_Os02g18720.1
Os02g0588500	LOC_Os02g37590.1	Os02g0588600	LOC_Os02g37600.1 LOC_Os02g37600.2	Os02g0733300	LOC_Os02g50040.1	Os02g0796000	LOC_Os02g55270.1
Os02g0807900	LOC_Os02g56380.1	Os03g0150700	LOC_Os03g05630.1	Os03g0151400	–	Os03g0151700	LOC_Os03g05720.1
Os03g0151800	LOC_Os03g05730.1	Os03g0151900	LOC_Os03g05740.1	Os03g0152000	LOC_Os03g05750.1	Os03g0152100	LOC_Os03g05760.1
Os03g0173900	LOC_Os03g07790.1	Os03g0174100	LOC_Os03g07800.1	Os03g0174200	LOC_Os03g07810.1	Os03g0193400	LOC_Os03g09810.1
Os03g0193600	LOC_Os03g09820.1	Os03g0194100	LOC_Os03g09840.1 LOC_Os03g09840.2	Os03g0194300	LOC_Os03g09850.1	Os03g0194900	LOC_Os03g09900.1
Os03g0199000	LOC_Os03g10230.1	Os03g0199100	LOC_Os03g10240.1 LOC_Os03g10240.2	Os03g0199500	LOC_Os03g10260.1	Os03g0200000	LOC_Os03g10300.1
Os03g0200200	LOC_Os03g10320.1 LOC_Os03g10320.2	Os03g0201700	LOC_Os03g10460.1	Os03g0201901	LOC_Os03g10478.1	Os03g0202001	LOC_Os03g10500.1
Os03g0362500	LOC_Os03g24760.1	Os03g0393700	LOC_Os03g27590.1 LOC_Os03g27590.2 LOC_Os03g27590.3	Os03g0571700	LOC_Os03g37470.1	Os03g0571900	LOC_Os03g37490.1
Os03g0572250	LOC_Os03g37570.1	Os03g0572900	LOC_Os03g37640.1 LOC_Os03g37640.2	Os03g0629900	LOC_Os03g43114.1 LOC_Os03g43114.2	Os03g0797000	LOC_Os03g58260.1 LOC_Os03g58260.2
Os03g0828100	LOC_Os03g61260.1 LOC_Os03g61260.3 LOC_Os03g61260.4	Os03g0828500	LOC_Os03g61280.1 LOC_Os03g61280.2	Os03g0828600	LOC_Os03g61290.1	Os03g0828800	LOC_Os03g61310.1
Os03g0829000	LOC_Os03g61330.1 LOC_Os03g61330.2 LOC_Os03g61330.3 LOC_Os03g61330.4	Os03g0829100	LOC_Os03g61340.1	Os03g0829200	LOC_Os03g61360.1	Os03g0838400	LOC_Os03g62200.1
Os03g0839100	LOC_Os03g62250.1 LOC_Os03g62260.1	Os03g0839200	LOC_Os03g62270.1	Os04g0107300	LOC_Os04g01660.1	Os04g0289500	LOC_Os04g22210.1
Os04g0289800	LOC_Os04g22240.1	Os04g0360500	LOC_Os04g29090.1	Os04g0367400	LOC_Os04g29920.1	Os04g0371000	LOC_Os04g30250.1 LOC_Os04g30250.2 LOC_Os04g30250.3
Os04g0479000	LOC_Os04g40290.1	Os04g0479100	LOC_Os04g40300.1	Os04g0479200	LOC_Os04g40310.1 LOC_Os04g40310.2	Os04g0479300	LOC_Os04g40320.1
Os04g0480900	LOC_Os04g40490.1	Os04g0539300	LOC_Os04g45630.1	Os04g0593700	LOC_Os04g50730.1	Os04g0613200	LOC_Os04g52320.1
Os04g0659300	LOC_Os04g56430.1	Os04g0659400	LOC_Os04g56440.1 LOC_Os04g56440.2	Os04g0659500	LOC_Os04g56450.1 LOC_Os04g56450.2	Os05g0458500	LOC_Os05g38410.1
Os05g0458600	LOC_Os05g38420.1	Os05g0508300	LOC_Os05g43230.1	Os05g0545300	LOC_Os05g43230.1	Os06g0108500	LOC_Os06g01890.1
Os06g0115400	LOC_Os06g02500.1	Os06g0115500	LOC_Os06g02510.1 LOC_Os06g02510.2 LOC_Os06g02510.3 LOC_Os06g02510.4 LOC_Os06g02510.5	Os06g0115600	LOC_Os06g02520.1	Os06g0118800	LOC_Os06g02800.1

Supplementary Table S5: The summary of the hotspot genes, II.

RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID		RAP Gene ID <-> MSU Gene ID	
Os06g0119100	LOC_Os06g02830.1	Os06g0133000	LOC_Os06g04200.1 LOC_Os06g04200.2 LOC_Os06g04200.3 LOC_Os06g04200.4	Os06g0186400	LOC_Os06g08720.1 LOC_Os06g08720.2 LOC_Os06g08720.3	Os06g0186500	LOC_Os06g08730.1
Os06g0186600	LOC_Os06g08740.1 LOC_Os06g08740.2	Os06g0186700	LOC_Os06g08750.1	Os06g0186900	LOC_Os06g08770.1	Os06g0253600	LOC_Os06g14280.1
Os06g0687800	LOC_Os06g47330.1 LOC_Os06g47330.2	Os06g0687900	LOC_Os06g47340.1	Os06g0688100	LOC_Os06g47350.1 LOC_Os06g47350.2	Os06g0688200	LOC_Os06g47360.1
Os06g0688400	LOC_Os06g47400.1	Os06g0689600	LOC_Os06g47470.1	Os07g0213300	LOC_Os07g11280.1 LOC_Os07g11280.2	Os07g0213400	LOC_Os07g11290.1
Os07g0213800	LOC_Os07g11330.1	Os07g0238600	LOC_Os07g13450.1	Os07g0238700	LOC_Os07g13460.1	Os07g0412100	LOC_Os07g22930.2 LOC_Os07g22930.3 LOC_Os07g22930.4
Os07g0462500	LOC_Os07g27850.1	Os07g0521600	LOC_Os07g33730.1 LOC_Os07g33730.2	Os07g0599100	LOC_Os07g40810.1	Os07g0628900	LOC_Os07g43570.1
Os07g0645500	LOC_Os07g45100.1	Os07g0653000	LOC_Os07g45950.1	Os07g0657600	LOC_Os07g46400.1	Os07g0657900	LOC_Os07g46410.1
Os08g0161750	LOC_Os08g06440.1	Os08g0171000	LOC_Os08g07430.1	Os08g0190100	LOC_Os08g09080.1 LOC_Os08g09080.2	Os08g0190700	LOC_Os08g09170.1
Os08g0191000	LOC_Os08g09190.1	Os08g0191200	LOC_Os08g09210.1 LOC_Os08g09210.2	Os08g0364900	LOC_Os08g27720.1 LOC_Os08g27720.2	Os08g0503200	LOC_Os08g39350.1
Os08g0517600	LOC_Os08g40580.1 LOC_Os08g40580.2 LOC_Os08g40580.3 LOC_Os08g40580.4	Os08g0540500	LOC_Os08g42760.1	Os08g0558000	LOC_Os08g44380.1 LOC_Os08g44380.2	Os08g0558800	LOC_Os08g44450.1 LOC_Os08g44450.2 LOC_Os08g44450.3
Os09g0350900	LOC_Os09g18159.1	Os09g0356000	LOC_Os09g19140.1	Os09g0408900	LOC_Os09g24330.1	Os09g0463100	LOC_Os09g28840.1
Os09g0545700	LOC_Os09g37369.1	Os09g0549400	LOC_Os09g37700.1	Os09g0549450	LOC_Os09g37710.1 LOC_Os09g37710.2	Os09g0549500	LOC_Os09g37720.1
Os09g0549600	LOC_Os09g37730.1	Os09g0559800	LOC_Os09g38690.1 LOC_Os09g38690.2 LOC_Os09g38690.3 LOC_Os09g38690.4	Os10g0443200	LOC_Os10g30640.1	Os10g0497000	LOC_Os10g35380.1
Os10g0520100	LOC_Os10g37570.1	Os11g0485900	LOC_Os11g29520.1	Os12g0166000	LOC_Os12g06890.1	Os12g0284700	LOC_Os12g18690.1
Os12g0285100	LOC_Os12g18729.1 LOC_Os12g18729.2 LOC_Os12g18729.3 LOC_Os12g18729.4 LOC_Os12g18729.5	Os12g0285500	LOC_Os12g18760.2	Os12g0468600	LOC_Os12g28270.1 LOC_Os12g28270.2	Os12g0472800	LOC_Os12g28770.1
Os12g0473900	LOC_Os12g28880.1	Os12g0612300	LOC_Os12g41810.1	Os12g0612400	LOC_Os12g41820.1		