Ancestry HMM with Gene Conversion

Author: Trevor Ridgley

Advisor: Dr. Russell Corbett-Detig

Department of Biomolecular Engineering, UC Santa Cruz

Abstract

Admixture, the result of interbreeding between genetically distinct populations, can give an early glimpse into reproductive incompatibility and speciation processes (Corbett-Detig et al., 2017; Medina et al., 2018). When recombination introduces breaks in ancestry along a chromosome, the resulting ancestry tracts yield valuable information about the admixture process. For example, in more ancient admixture, there is more time for recombination, and the resulting tracts should be shorter. In particular these signatures of recombination can be inferred and studied using probabilistic models, but this picture is incomplete because double strand breaks during meiotic recombination can induce gene conversions as well as meiotic crossovers (Korunes & Noor, 2016). Gene conversion is a strong evolutionary force, especially over short genetic distances, that is thought to have shaped Eukaryotic genomes and our meiotic machinery (Burt & Trivers, 2008). Hence, we have updated the inference model of Ancestry HMM to include the gene conversion process in a comprehensive recombination model. For diploid individuals sampled from our simulated admixed population, over 85% of individual gene conversion segments are detected down to single variant resolution. We demonstrate that the gene conversion process is modelled well by our HMM across various simulated data parameters using both curated genotype and read pile-up data. Finally, we employ a Nelder-Mead parameter optimization algorithm for learning the unknown parameters that generated the data. This will enable researchers to identify and study gene conversion across species with the genetic map also serving as an effective scaffold in de novo genome assembly applications.

1. Introduction

Hilliker et al defined meiotic gene conversion as a non-reciprocal exchange of genetic information between one homologous chromosome and another [1]. Unlike meiotic crossover which involves reciprocal exchange of genetic material, meiotic gene conversion can change the frequencies of genetic variants, or alleles, within a population [2]. At the cellular level, gene conversion is not only thought to have played an important role in shaping our genomes, but the meiotic machinery as well [3]. Consequently, gene conversion is a strong evolutionary force that should be incorporated into biological recombination models if we wish to study biological processes such as genetic mutation, evolution, and population structure as accurately as possible. Here, we apply a Sequential Markovian Coalescent ("SMC"') model to the problem of studying admixture [4][5].

As species sub-populations become established, their genomes diverge from those of other sub-populations, eventually leading to reproductive incompatibility and speciation. But on rare occasions, these sub-populations can interbreed after significant time apart, and admixed individuals can exhibit interesting phenotypes. By detecting local ancestry (or haplotype blocks) in these admixed individuals, we can learn about the genotypes that underlie traits of interest in a population genetics context. Traditionally, these haplotype blocks have been uncovered via breeding experiments in model organisms, but learning haplotypes in this way poses practical and ethical challenges in many species, including humans. For these reasons, we would like to infer the haplotype blocks through detection of local ancestry, but existing methods only capture the fraction of the genetic mapping that results from meiotic crossover. Therefore, the scientific problem that this revised method will address is inferring complete genetic maps that consist of both meiotic crossover and meiotic gene conversion.

The development of genomic modeling methods is challenging due to the complexity of underlying biological systems, but discoveries in the field of population genetics have accelerated with the advent of Coalescent Theory. Proposed by Kingman, Tavaré, and Griffiths, Coalescent Theory makes surprisingly accurate predictions by looking into the ancestry of a sample. The effectiveness of Coalescent Theory is rooted in the notion of looking back in time rather than forward, restricting much of the forward-in-time stochasticity that gives rise to an enormous number of possibilities. By modeling backwards as genetic lineages coalesce, there are only 2 possible outcomes for a pair of sequences: genetic sequences in a sample are related by a common ancestor in the previous generation, or they are not. This logically leads to a bifurcating tree structure of underlying genealogy upon which sequence variation can be modelled [6]. Coalescent models continue to evolve which allows the interplay between an increasing number of biological mechanisms to be understood; for instance, gene conversion and recombination [7]. In concert with a Wright-Fisher model that assumes a large, fixed population size, constant mutation rate, and selectively neutral mutation, Coalescent Theory has been used for modeling genealogical variation, single nucleotide polymorphisms, and recombination [6]. With contributions by Hilliker and Wiuf, similar modeling approaches are being applied to another type of variation: Gene Conversion [1][7].

1.1 Recombination and Gene Conversion

Modeling gene conversion presents unique challenges. First, gene conversion tract lengths are extremely short (on the order of 500 nucleotides), involving just tens or thousands of base pairs (bp) while meiotic crossovers spanning hundreds of thousands or millions of bp [2]. Consequently, genetic variants identified in high throughput sequencing experiments must overlap GC tracts for their detection to be possible. Sufficient divergence in base sequences between individuals of a subpopulation is another necessary criterion for GC detection. In other words, even if a genetic variant overlaps a GC tract, the GC may go undetected if most or all of the individuals in a genetic sample possess the same variant. Experimental studies of model organisms including fungi, yeast, and *Drosophila melanogaster* have revealed that non-Mendelian segregation occurs at some infrequent rate as a result of gene conversion [3]. As more about particular genes in model organisms became known, some of their individual peculiarities could be exploited using sophisticated breeding crosses and Biotechnology to suss out better estimates of gene conversion [1]. But this methodology does not scale genome-wide within well-studied species, nor does it extend to novel species for which there is little-to-no prior knowledge of their Biochemistry. With sufficient data collection from thousands of experiments, it became possible to fit a statistical distribution to the observations and validate them in the lab.

Developing mathematical models that are consistent with the true biological data is necessary for making substantially faster progress in our understanding of population genetics mechanisms. Based on existing genotype data, previous work estimated the mean length of gene conversions at roughly 1300 base pairs [1]. However, there is bias with these figures because shorter gene conversion events are less likely to overlap with a mutation signature that can be detected by traditional experimental approaches. Two fascinating results from the Hilliker and Chovnik experiments in *D. melanogaster* were the relationship between gene conversion and recombination, and the GC sequence mean length using a Maximum-Likelihood Estimator of ~ 342 base pairs, a result that was significantly shorter than previous estimates. Apparently, the longer gene conversions were more likely to become recombinant crossovers and helped to derive the prevailing recombination rate of 3 crossovers per 10^8 base pairs per meiosis on average.

1.2 Probabilistic Modeling

Coalescent Theory provided the foundation for many useful mathematical properties derived from genetic sequence variation. One of the most recognizable results of population genetics is the heterozygosity parameter $\theta = 4N\mu$ [8]. It says that, for a population of size N the rate that variation is being lost due to stochastic genetic drift is in equilibrium with the rate that new variation is being added to the population by spontaneous mutation with rate μ. The term $4N$ represents diploid organisms like humans and most other eukaryotes containing 2 pairs of chromosomes with 2 genealogical branches separating them. And when there is variation in the population, we can ask questions about how two sequences relate and how far back they share a common ancestor.

Using the coalescent, the probability that any pair of genetic sequences coalesce in the previous generation is $1/2N$, where N is the effective population size. When a difference between actual and effective population size is observed, this indicates that the heterozygosity of the population is not in equilibrium [8]. Next, given a sample of n genetic sequences in a sample, the probability that any pair of those sequences coalesce in the next generation is $nC2/2N$ (read: n choose 2, or the possible sets of size 2). In the context of gene conversion, we are asking a slightly different question about the lineage of the segment. Instead of asking the probability that a sample of sequences coalesce, we want to know how much time we must go back in time before reaching the sequence that gave rise to the gene conversion between 2 adjacent sites [7]. But first, we have to understand the length distribution for gene conversions and how the coalescent relates to waiting times.

According to Wiuf and Hein, genetic sequences can be assigned the length variable L, and some event occurring within this genetic sequence length (whether a sequencing read, mutation, or gene conversion) can be called A. If event A occurs with probability g, and gene conversions are significantly shorter than recombination with length Z, then we can model this length using the Geometric distribution from Hilliker [1][7], note also that "memoryless" distributions such as the geometric are convenient for hidden Markov modelling frameworks, see below. According to the probability mass function, $P(Z) = q(1-q)^L$ which says that the probability of our event is the rate of failure over L trials before the successful trial occurs with rate q. The Geometric Distribution with parameter q has mean $q/(1-q)$, so the parameter gives some sense for the mean length of a Gene Conversion tract which Hilliker described empirically for *Drosophila* as ~ 342bp [1].

With genome sizes being incredibly large and the rate of nucleotide variation incredibly small, a common mathematical simplification that is popular in genetics allowed Wiuf and Hein to convert the Hilliker discrete Geometric distribution into a continuous Exponential distribution and derive a convenient new distribution for the GC length [1][7]. In this case, supposing that genome sequence has extreme length L, and the rate of our biological event A has a vanishingly small probability q, then the new rate parameter $Q = qL$ can be used in the Exponential distribution. Because q only tells us the probability of event A occurring, and we would like to model the possible length of event Q with respect to the total genome sequence length L, $z = Z/L$ can be defined as the relative length of GC with respect to the full sequence of length L. Hence, the GC length distribution becomes:
$P(Z=z) \sim Expo(qL) = qLe^{-qLt} = Qe^{-Qt}$ [7].

Hudson (1983) described the coalescent with recombination as an exponential distribution with parameter $R = 4rNL$, where r is the probability of recombination (ie, recombination rate in Centimorgans per millions bases), N and L are the population and sequence length as defined previously. Wiuf and Hein claim that this same distribution, except with parameter $G = 4gNL$ applies to gene conversion as well, and suggests that combining this idea with the length distribution described previously, then the amount of waiting time until a gene conversion event occurs in a particular sequence can be expressed using the exponential distribution, as well [7].

2. Materials and Methods

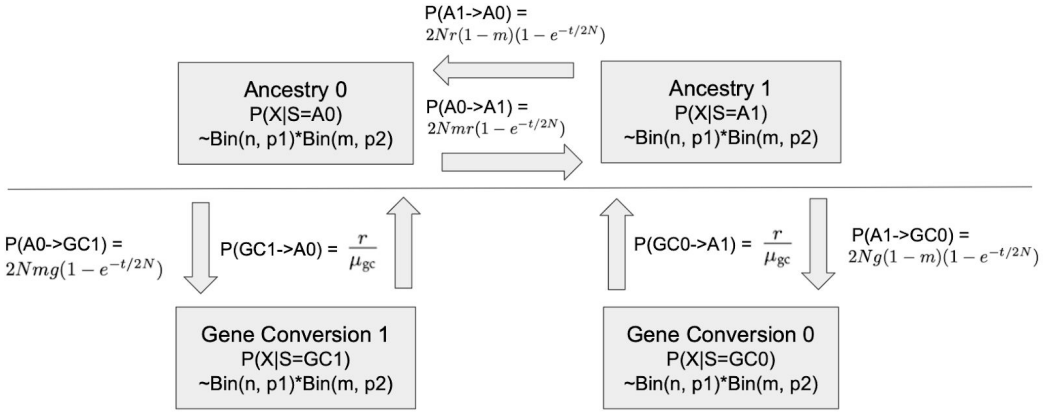
Ancestry HMM [4][5] is available at the following github repository:
https://github.com/russcd/Ancestry_HMM

The code changes enabling Gene Conversion were made to a forked repository at my github:
https://github.com/tridgley/Ancestry_HMM

2.1 Mathematical Model

The existing Ancestry HMM state is a tuple of length |S|, where S is the set of ancestral populations. Each admixture event is known as an admixture pulse, so there are |S|-1 admixture pulses [5]. Each element of the state tuple can take values from 0 to n, where n is the number of chromosomes in a genotype or read pileup sample [4]. For simplicity, a diploid sample will be assumed along with 1 ancestral pulse between 2 ancestral populations such that the state space along a single chromosome transition process is S = {anc0, anc1} and possible states within a diploid sample are (2, 0), (1, 1), and (0, 2). In words, these states represent both chromosomes from the sample having Ancestry Type 0 (Anc0) at a particular locus, one chromosome from Ancestry Type 0 and the second chromosome from Ancestry Type 1 (Anc1) at a particular locus, or both chromosomes from Ancestry Type 1 at a particular locus, respectively.

Figure 1: Ancestry HMM state diagram
Valid Markovian paths indicated by arrows between gray boxes representing the hidden ancestral states.



The state space was expanded to include a GC state for each ancestral state (GC0 and GC1), allowing us to specify local ancestry parameters that extend its detection to include gene conversions. In this example, the state space A_gc becomes {Anc0, Anc1, GC0, GC1} and possible states are a multichoice of 2 chromosomes among 4 categories: (2, 0, 0, 0), (1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 2, 0, 0), (0, 1, 1, 0), (0, 1, 0, 1), (0, 0, 2, 0), (0, 0, 1, 1), (0, 0, 0, 2). The underlined states are analogous to the original Ancestry HMM ancestral states, and the updated transition table in Figure 2 reflects this new state space.

Figure 2: Ancestry HMM transition matrix

Matrix of transition probabilities from states in left margin to states in top row. Elements with 0 rate deemed biologically infeasible or undetectable based on tag SNP data alone.

$$
\begin{array}{c} \\ Anc0 \\ Anc1 \\ Anc0gc \\ Anc1gc \end{array}
\begin{array}{c}
\begin{array}{cccc} Anc0 & Anc1 & Anc0gc & Anc1gc \end{array} \\
\begin{pmatrix}
1-\sum_{i\neq j}^{n} x_{1j} & 2Nmr(1-e^{-t/2N}) & 0 & 2Nmg(1-e^{-t/2N}) \\
2Nr(1-m)(1-e^{-t/2N}) & 1-\sum_{i\neq j}^{n} x_{2j} & 2Ng(1-m)(1-e^{-t/2N}) & 0 \\
0 & \frac{r}{\mu_{gc}}+2Nmr(1-e^{-t/2N}) & 1-\sum_{i\neq j}^{n} x_{3j} & 0 \\
\frac{r}{\mu_{gc}}+2Nr(1-m)(1-e^{-t/2N}) & 0 & 0 & 1-\sum_{i\neq j}^{n} x_{4j}
\end{pmatrix}
\end{array}
$$

The upper-left quadrant represents the original Ancestry HMM transitions from Anc0 → Anc0, Anc0 → Anc1, Anc1 → Anc0, and Anc1 → Anc1. N is the number of individuals in the admixed population, m is the fraction of migrants from Anc1 that mixed with individuals from Anc0, and r is the recombination rate. The exponential function of t is a coalescent term representing the cumulative probability of two sites sharing a common ancestor by the admixture t generations in the past. The coalescent idea behind these transition rates is that, over longer periods of time t since the admixture pulse, there are more recombinations that break down ancestry tracks, resulting in their greater frequency and shorter length distribution. A smaller migration rate m or lower recombination rate r each slows the breakdown of ancestry tracts, reducing the probability of a transition from Anc0 → Anc1 and vice-versa.

The upper-right quadrant represents transitions from ancestral states Anc0 or Anc1 to their corresponding gene conversion states GC0 or GC1. The idea here is that gene conversions occur with some rate relative to recombinations [3], and this proportion is captured with the coefficient g. In order to determine g, a command line option for the GC fraction was added to the Ancestry HMM such that $g = gcFraction / (1 - gcFraction)$. For instance, if the $gcFraction = 0.5$, then $g = r$, but if $gcFraction = 0.9$, then $g = 9$ (in units of gc events per Morgan) and the rate of transition from Anc0 → GC1 will be nine times greater than transition from Anc0 → Anc1 since there should be nine times more GC's in the genome than recombinations. The transition states from Anc0 → GC0 and Anc1 → GC1 were set to 0 because, even though these transitions occur in reality, they would not be observable events since the haplotype frequencies are consistent with Anc0 in either case.

The lower-left quadrant represents transition from gene conversion states GC0 and GC1 back to the Anc0 and Anc1 tracts. The GC tract length is another command line parameter that is passed into Ancestry HMM for building the transition rate tables. The main idea here is to capture the reduced length of GC tracts compared to recombinant tract lengths using 1/GC tract length as the rate parameter. The reciprocal of shorter GC tract lengths is a larger rate that will transition back from GC1 to Anc0 much sooner than Anc1 to Anc0, except for extremely large coalescent times. We also include a coalescent term in the transition rates of GC back to Anc states because, as time increases, the probability that long ancestral tracts collide with GC tracts becomes more significant, and without this coalescent parameter the GC lengths were consistently underestimated. Finally, transition rates in the lower-right quadrant consist of

remaining in GC tracts, or inter-converting between GC tracts. The latter possibility is considered to be so remote that we model it as 0. The rate of remaining in a GC tract is 1 minus the probability of leaving, which is the reciprocal of mean length discussed previously.

2.2 Simulating A Ground Truth Dataset

SELAM [9] is available at the following github repository:
https://github.com/russcd/SELAM

In order to validate our revised HMM, recombination and gene conversion simulation was necessary to know precisely where and how frequently each of the biological events takes place, along with their lengths. First, SELAM (Simulation of Epistasis and Local Ancestry during Admixture with Mate Choice with Mating) was updated to accept a gene conversion option with user-specified GC fraction and GC length parameters [9]. The recombination rate was also scaled by the ratio of recombination to GC, i.e. this is simulating the same model as we are estimating from the data. We ran SELAM for an admixed population of 2000 diploid individuals possessing a single autosome 3 morgans in length that could be sampled at various points in time. The parameters used to generate the population data were GC fraction of 0.9 and GC length of 1e-4 centimorgans (equivalent to approximately 500bp). We sampled 50 individuals at time points of $t = 10$ and $t = 100$ (measured in generations since admixture), and each multiple of 100 from 200-1000 generations, with 10 and 100 generations mapping to roughly 200 and 2,000 years for humans. The SELAM logging provided exact GC site coordinates and lengths that were parsed with a custom Python script and referenced as the ground truth dataset.

Next, it was necessary to generate read pile-up or curated genotype data as the direct input to Ancestry HMM; this step effectively abstracts the ground truth data and becomes increasingly convolved with sample size, less sub-population divergence, etc. A custom perl script parses the population_output.txt file from SELAM and generates a track of SNP Tag Markers based on configurable genome size in base pairs, SELAM genome length, recombination rate, interval, and sample size. For the preliminary assessment, we settled on a 10 megabase genome with recombination rate of $3e-8$ and single nucleotide polymorphisms (SNP's) occurring every 100 base pairs. This conveniently gave GC lengths according to $1e-4cM * 1e7bp / 3cM = 333bp$, which is consistent with the GC mean length of 342bp reported by Hilliker et al [1]. The SNP tracks were generated for sample sizes of 4, 10, and 50 individuals.

2.3 Inferring Gene Conversions with Revised HMM

With the sample and tag SNP files as inputs, Ancestry HMM was run with several initial goals in mind: 1) Can recombinant ancestral tracts be detected as before? 2) Can we now detect gene conversions? 3) Is the likelihood maximized with the parameters that generated the data? 4) How finely can the parameters be predicted? 5) Can the parameters be learned? To assess our ability to detect gene conversions, we ran Ancestry HMM on the simulated data both with and without gc command line options, then plotted the calls relative to the ground truth dataset. The results were plotted in terms of genomic location and GC length (see results

section). Next, a custom shell script performed broad grid search with a Python script to parse the Ancestry HMM output files and plot the log-likelihood versus parameter surface. The sets of broad parameters included GC fraction and GC length:

$GC_{frac} \in \{0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.8,\ 0.9,\ 0.95,\ 0.99\}$

$GC_{length} \in \{0.01,\ 0.001,\ 0.0001,\ 0.00001\}$ in units of centimorgans.

Then, fine grid search was performed with the sets of GC fraction and GC length:

$GC_{frac} \in \{0.87,\ 0.88,\ 0.89,\ 0.895,\ 0.899,\ 0.9,\ 0.901,\ 0.905,\ 0.91,\ 0.92,\ 0.93\}$

$GC_{length} \in \{0.009,\ 0.0095,\ 0.0099,\ 0.0001,\ 0.000101,\ 0.000105,\ and\ 0.00011\}$

in units of centimorgans.

Ancestry HMM assigns probabilities for each state at every locus present in the genotype input file. A custom python script called ‘gcReport.py’ parses the Ancestry HMM output and counts the number of contiguous ancestral blocks for each state using $argmax$ for each locus. Using the GC report script along with the custom script ‘SelamGCFinder.py,’ we can compare the detected GC tracts with the ground truth tracts for sensitivity and accuracy assessments.

2.4 Nelder-Mead Parameter Optimization

The utility of Ancestry HMM lies in its ability to make local ancestry inferences from admixed biological sequencing data [6][7]. In most cases, the gene conversion parameters for various species will be unknown and we envision Ancestry HMM with Gene Conversion as a valuable tool in novel species research. Hence, Ancestry HMM should be able to make accurate predictions of the parameters from sequencing data. Based on the log-likelihood of data given the probabilistic model, we employed a Nelder-Mead parameter optimization algorithm using the Scipy Optimize package. In particular, the optimize.minimize() function provides a collection of optimizers. The Nelder-Mead algorithm basically initializes three random values in a simplex set and calculates the next parameters based on an objective function it tries to minimize. In our case, the objective function is the log-likelihood output from Ancestry HMM, except we multiply by -1 since we are trying to maximize the log-likelihood. We wrote a Python script called ‘NelderMead.py’ that uses the Python sub-process and file system interfaces to call Ancestry HMM and parse log files for log-likelihood extraction from the Linux command line. The final simplex should be a highly accurate estimation for the gene conversion fraction and length parameters described previously in our Mathematical Model section.

Our first trial with the Nelder-Mead parameter optimization executed Ancestry HMM with GC parameters of 0.9 and 0.0001 for an admixture event between 2 ancestral populations 100 generations ago. Since the Nelder-Mead algorithm gets initialized with arbitrary parameter values, we used 0.5 as the GC fraction and 0.005 as the GC length for our simulated data. With knowledge of the ground truth parameters that generated the data, we can effectively determine how well the algorithm learns these parameters. We are currently using the default tolerance settings and a hard stop of 100 iterations, although the algorithm consistently converged within 80 iterations. One limitation of the Scipy Nelder-Mead implementation is the inability to constraint input parameters, such as enforcing $0 < GC\ fraction\ < 1$ and $0 < GC\ length$. In cases where the algorithm tries invalid parameters, the log-likelihood returns “NaN” and we excluded these iterations from the simplex set.

2.5 Read Pile-Up Data From *D. melanogaster* Population

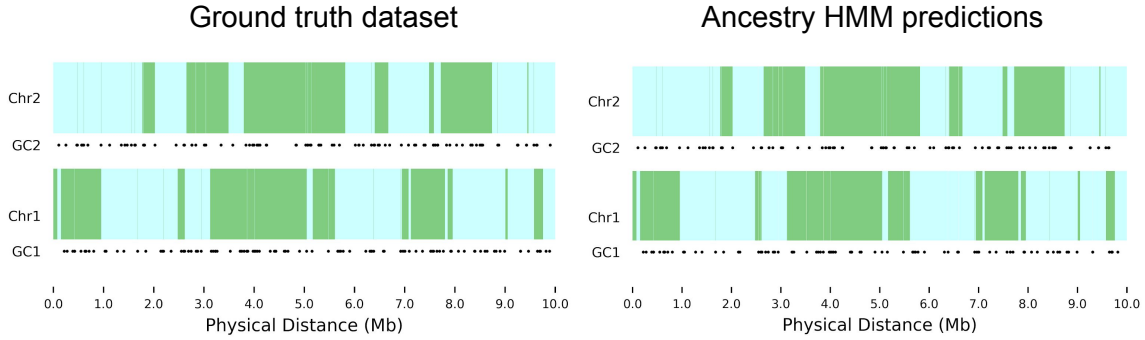
Model performance up to this point was characterized based on an ideal set of genotype data where the ancestral populations were completely divergent. This is unlikely in the real world where subpopulations are only genetically distinct to a limited degree. Hence, we employed read pile-up data from a *Drosophila melanogaster* population from an existing panel of 100 individuals that were sequenced and their SNP's tagged. In contrast to the ideal in-silico genotypes used before, these SNP tags occur in irregularly-spaced random intervals which is also consistent with real genomes. We employed another custom Perl script 'Simulate_reads_drosophila.pl' that takes a population sample file generated using SELAM [9] and the aligns read pileup file to produce a SNP track for Ancestry HMM. We generated custom SNP tracks using samples of 4, 10, 25, and 50 individuals for each of 10, 20, 50, 100, and 150 generations since admixture. Then, we ran broad grid search and Nelder Mead optimization for each of these configurations to characterize the likelihood surface and parameter estimation, respectively.

3. Results

3.1 Local Ancestry Predictions

We first compared Ancestry HMM recombinant tract predictions with the ground truth SELAM predictions. Figure 3 shows both admixed chromosomes from a single individual in the sample. The ground truth chromosomes on the left were created using the meiotic crossover and gene conversion loci results using 'SelamGCFinder.py' , while the predicted chromosomes on the right used the AHMM output parsed using 'gcReport.py'.

Figure 3: Representative Ancestry Switches in an Admixed Individual
Visualization of admixed chromosomes sampled from a single diploid individual after 100 generations. Light blue blocks indicate ancestry type 0 and green blocks indicate ancestry type 1. Transitions between colors result from either meiotic recombination or meiotic gene conversion. With GC tracts being very short, 1D scatterplot below each chromosome indicates presence of GC in the ground truth dataset (left) and successful prediction of the GC (right).



The sizes and locations of the ancestry blocks appear practically identical between both chromosome pairs. GC tracts indicated by a linear series of dots below each chromosome show consistent spacing and quantity. Minor deviations in the GC predictions can be seen, particularly

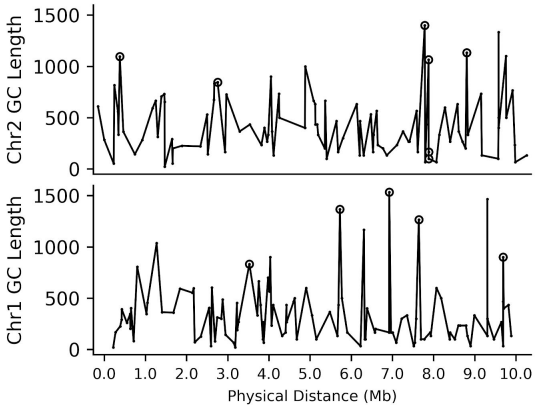
between 9 and 10Mb between ground truth and inferred chromosomes. Based on these observations, the revised AHMM model appears to be making good ancestral tract predictions from meiotic crossover events.

3.2 Gene Conversion Detection

The ground truth dataset permits analysis of gene conversion detection in terms of loci and length, which are plotted in Figure 4 and Figure 5. Prior to the model update, AHMM lacked sensitivity to detect GC's due largely to their much shorter length compared to meiotic crossover tracts that can span tens of thousands or millions of bases. In some cases, such a long admixture pulse time or large mean GC length, the distribution tails can overlap in such a way that AHMM can call longer GC's, but the shorter GC's are still unlikely to be called.

Figure 4: Gene Conversion Detection using old AHMM model

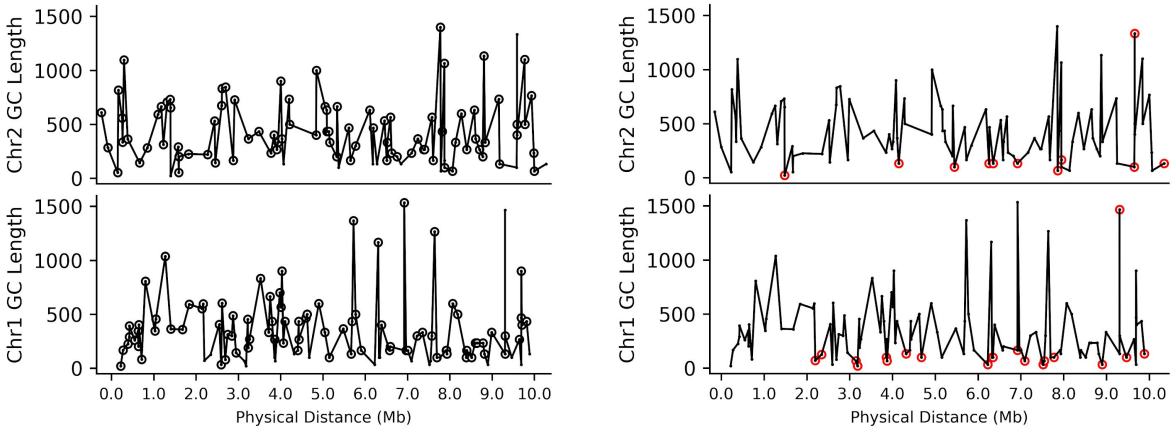
All simulated GC's along a 10Mb simulated genome plotted by their length and position. Circled points indicate successful GC detection by Ancestry HMM without gene conversion.



In Figure 4, we can see that only a handful of the longest GC's from our ground truth dataset were called using the old AHMM model. We used this baseline as a comparison for the revised model and plotted the new run using the same SELAM population, sample time, and parameters to generate the data except now with AHMM set to the appropriate parameters.

Figure 5: Gene Conversion Detection using the revised model AHMM model

All simulated GC's along a 10Mb simulated genome plotted by their length and position. Points circled in black indicate successful GC detection by Ancestry HMM with gene conversion (left panel). Points circled in red indicate undetected GC's by Ancestry HMM with gene conversion (right panel).

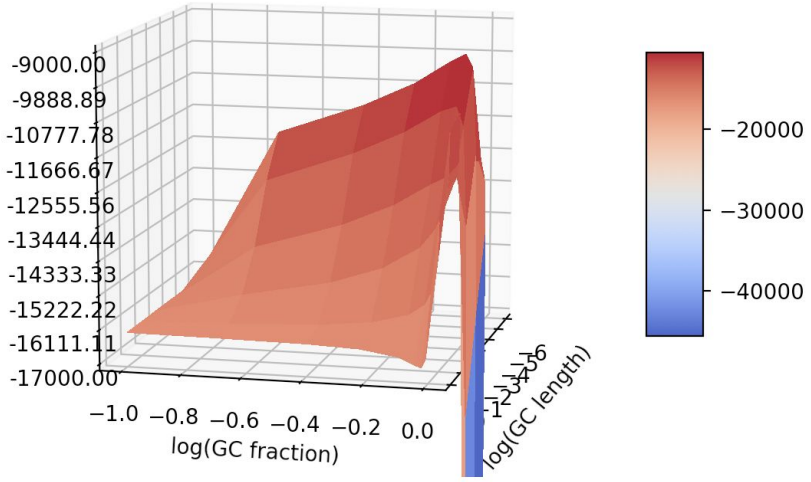


In Figure 5, Ancestry HMM now detects over 85% of gene conversion tracts, but still misses some of the shortest GC's. We attribute undetected GC's to the fact that some of them are lacking an associated Tag SNP because they are too short to span any markers. Also, as will be discussed later, the inferred parameters may differ slightly from the actual parameters which affects the ability to detect GC's at extreme lengths (close to the 100 bp resolution of our simulated Tag SNP) in our genotype file.

3.3 GC Likelihood

Once we could see that the model was working reasonably well in its ability to detect both meiotic recombination and meiotic gene conversion states, the biggest question was whether this new model was a good representation of the biological process. We elected to perform a grid search over a broad parameter space and compare the likelihood for each of the Ancestry HMM inferences based on these parameters. If the model was representative of meiotic processes, then we hypothesized that the likelihood should be maximized when we made predictions from the data using the parameters that generated the data. The log-likelihood surface was plotted as a surface over the GC length and GC fraction parameter space:

Figure 6: Log Likelihood Surface with broad parameters
3D surface plot of GC fraction along Y-axis, GC length along X-axis, with log-likelihood along the Z-axis. The peak of the surface represents the global maximum of log-likelihood.



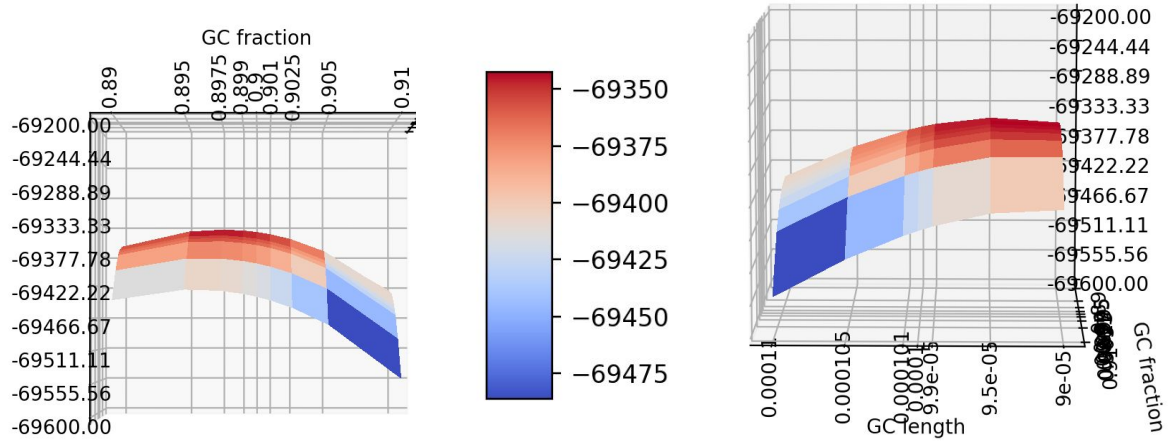
The axes are in log_{10} space where the peak at $10^{-0.045} = 0.9$ which is the GC fraction that was used to simulate this data, and $10^{-4} = 0.0001$ which is precisely the GC length in centimorgans. Also of importance is the well-behaved nature of the log-likelihood surface which will be amenable to any number of parameter optimization approaches down the road.

3.4 Fine Parameter Likelihood

We wanted to further establish the accuracy of the log likelihood for a finer set of parameters. We can see that each of the GC fraction and GC length are under-estimated slightly, hence several more experiments were performed to determine if the under-estimation was due to stochasticity of the simulation or prediction data (Fig 7).

Figure 7: Log Likelihood Surface with fine parameters

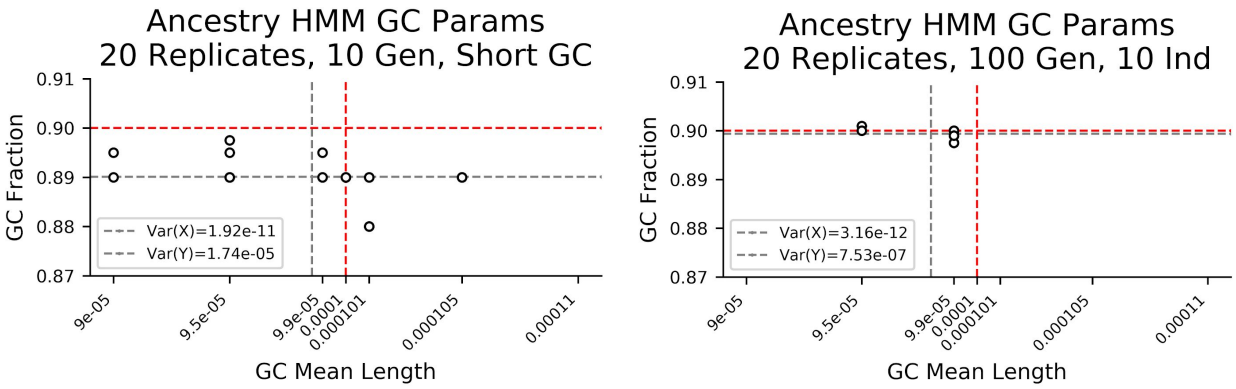
3D surface plot of GC fraction along Y-axis, GC length along X-axis, with log-likelihood along the Z-axis. Compared with Figure 6, GC fraction values were taken from the interval [0.89, 0.91] and GC length values from [1.1e-4, 9e-5).



The under-estimation bias remained despite 20 replicates of SELAM runs and Ancestry HMM inferences:

Figure 8: Parameter Under-Estimation Bias

GC fraction plotted against GC mean length where the dashed red lines indicate the true parameters, and the dashed gray lines indicate the mean for each of the 20 replicates.



We can see that given more time since admixture, the GC fraction bias appears to converge toward the true value but the GC length bias remains (Fig 8). As a result, we realized that meiotic recombinations could break down the GC tracts on occasion, and made the following change to the transition probabilities out of GC tracts:

$$r/\mu_{GC} \quad \Rightarrow \quad r/\mu_{GC} + 2Nmr(1-e^{-t/2N}) \quad and \quad r/\mu_{GC} + 2Nr(1-m)(1-e^{-t/2N})$$

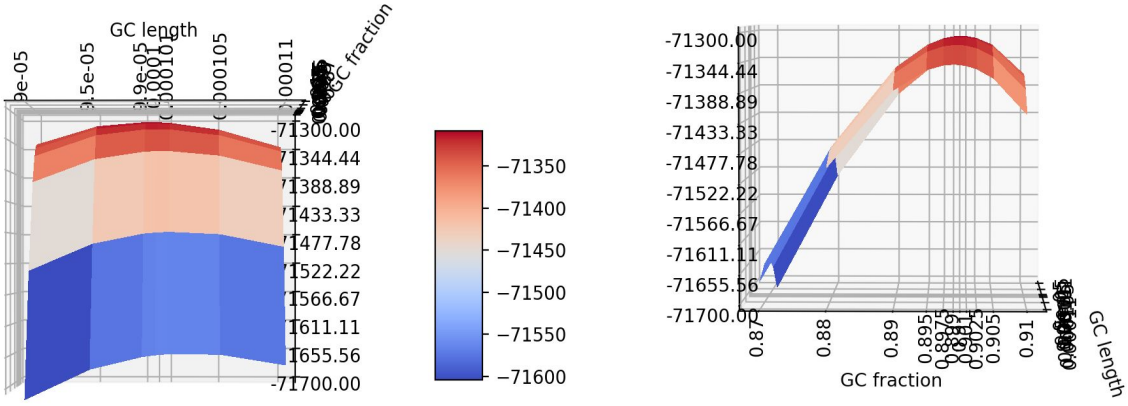
We also included the possibility that GC tracts could recombine, hence we have established the following iteration of our Ancestry HMM transition rate model (Fig 9).

Figure 9:
Matrix of transition probabilities from states in left margin to states in top row. The lower-left quadrant (columns 1, 2 of rows 3, 4) were revised such that GC segments can recombine per migrant, recombination, and coalescent rates.

$$
\begin{array}{c} \\ Anc0 \\ Anc1 \\ Anc0gc \\ Anc1gc \end{array}
\begin{array}{c}
\begin{array}{cccc} Anc0 & Anc1 & Anc0gc & Anc1gc \end{array} \\
\begin{pmatrix}
1-\sum_{i\neq j}^{n} x_{1j} & 2Nmr(1-e^{-t/2N}) & 0 & 2Nmg(1-e^{-t/2N}) \\
2Nr(1-m)(1-e^{-t/2N}) & 1-\sum_{i\neq j}^{n} x_{2j} & 2Ng(1-m)(1-e^{-t/2N}) & 0 \\
2Nr(1-m)(1-e^{-t/2N}) & \frac{r}{\mu_{gc}}+2Nmr(1-e^{-t/2N}) & 1-\sum_{i\neq j}^{n} x_{3j} & 0 \\
\frac{r}{\mu_{gc}}+2Nr(1-m)(1-e^{-t/2N}) & 2Nmr(1-e^{-t/2N}) & 0 & 1-\sum_{i\neq j}^{n} x_{4j}
\end{pmatrix}
\end{array}
$$

Compared to the earlier version of this model (see Methods), GC0 can recombine into Anc0 tract which is independent of the GC length, and vice-versa for GC1 to Anc1. With these refinements, the fine parameter search generated much more accurate predictions (Fig 10).

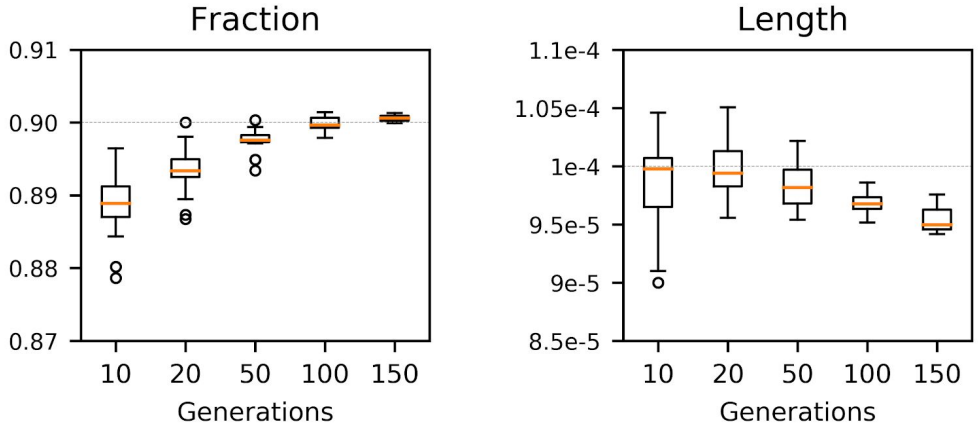
Figure 10: Revised Model GC Fraction
3D surface plot of GC fraction along Y-axis, GC length along X-axis, with log-likelihood along the Z-axis. Left panel along the X-axis with GC fraction values were taken from the interval [0.89, 0.91] with log-likelihood peak around 9.9e-5 to 1e-4. Right panel along the Y-axis with GC length values from [1.1e-4, 9e-5] with log-likelihood peak at 0.9.



We can see that the GC fraction under-estimation bias appears to be practically resolved by these changes, but the GC length bias still appears ever-so-slightly. We increased the admixture time resolution to begin understanding the ideal case for Ancestry HMM to make the most accurate parameter estimations. This can also provide clues for further enhancement of our model. Next, we wanted to gain a better understanding of how the admixture time correlates with parameter estimation bias. Using the same model except with admixed population samples at generational times points of 10, 20, 50, 100, and 150 generations since admixture, we collected 20 Ancestry HMM prediction replicates for each of these samples and plotted the distributions for both GC fraction and GC length parameters.

Figure 11: Box Plot of Parameter Bias versus Admixture Times

Interquartile distributions of GC fraction (left panel) and GC length (right panel) with medians indicated by green horizontal lines. Outliers plotted as circles extend beyond 1.5x of the interquartile range.



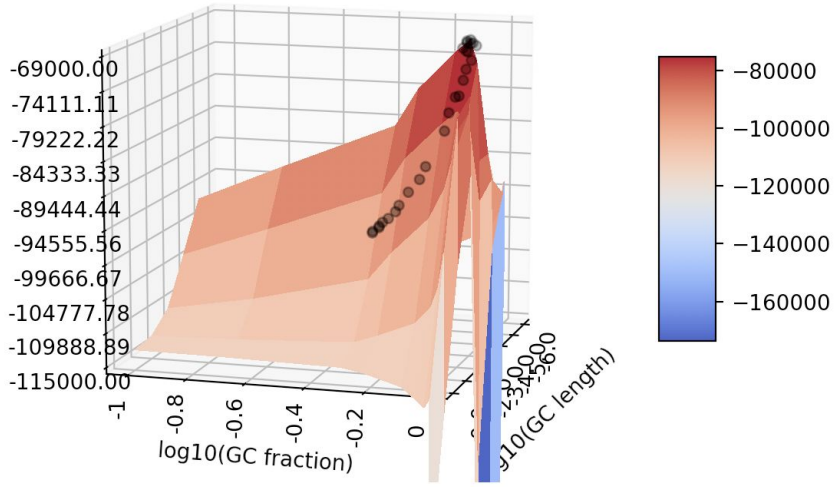
In Figure 11, the nominal GC fraction parameter is estimated between 100 and 150 generations since the admixture pulse in our simulated dataset, while the GC fraction is under-estimated at shorter admixture pulse times. This underestimation bias suggests that the coalescent model has limitations at shorter intervals; Nielsen and Liang stated "...the coalescent approximates the [Wright-Fisher] model well when *T* is large, but not necessarily so for small values of *T*" [10]. In contrast, the nominal GC length parameter is estimated between 10 and 20 generations since admixture pulse. Longer admixture pulse times lead to under-estimation of the GC length, which we attribute to an increased number of GC tracts that recombine, or where two GC tracts intersect, at some place in their interval and effectively reduces their length.

3.5 Nelder-Mead Parameter Optimization

Beginning with a simplex of 0.5 GC fraction and 0.005 GC length, the custom Python script 'NelderMead.py' optimized parameters of Ancestry HMM's log-likelihood function.

Figure 12: Nelder-Mead Learning Path

3D surface plot of GC fraction along Y-axis, GC length along X-axis, with log-likelihood along the Z-axis. Scatterplot overlay shows the path of Nelder Mead parameters starting with the initial guess at 0.5 fraction and 0.005 length, both in log scale. The cluster at the top of the surface peak indicates convergence of the optimal parameters.

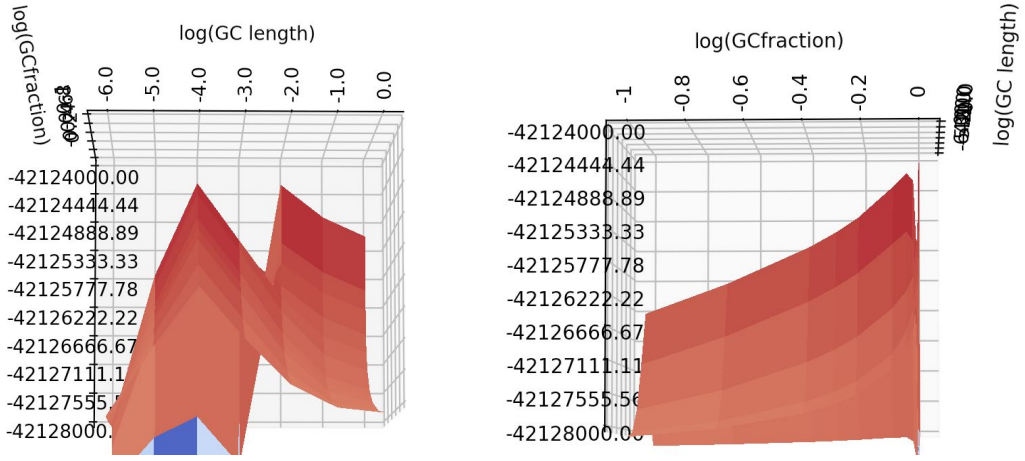


The learned GC fraction and GC length parameters were 0.901 and 9.69e-5 upon convergence, exhibiting relative errors of 0.11% and 3.1%, respectively. We visualized the learning path of the algorithm using a scatterplot of the simplex parameters for all iterations overlaid on the objective function surface plot (Fig 12). This result confirms that the likelihood surface characteristics are favorable for learning the parameters that generated our simulated genomic data. By learning these parameters with little error, we can derive confidence in the GC parameters inferred for unknown genomes.

3.6 Read Pile-Up Data From *D. melanogaster* Population

A critical validation step for assessing the model's ability to make accurate predictions was simulating SNP tracks from noisy sequencing data. We expected the likelihood surface to deviate further from the generating parameters and the surface characteristics to degrade, but perhaps not so much that learning would remain possible. Using the same grid search parameters as before, we first assessed the objective function qualitatively before attempting optimization (Fig 13).

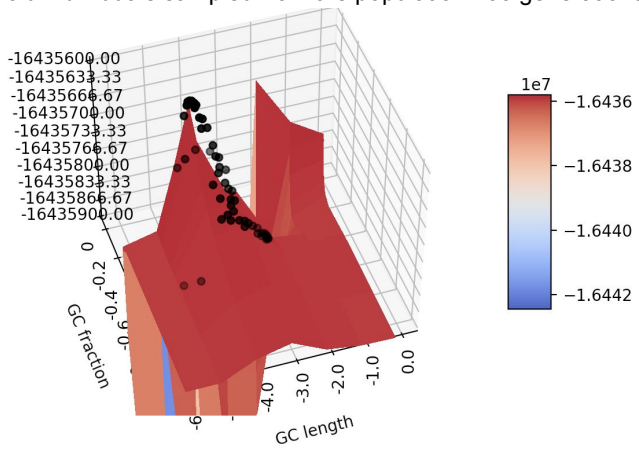
Figure 13: Likelihood Surface Grid Search Using *D. melanogaster* Data
3D surface plot of GC fraction along Y-axis, GC length along X-axis, with log-likelihood along the Z-axis. Left panel along the X-axis shows log-likelihood peaks at -4.0 and -2.0 in log space. Right panel along the Y-axis shows log-likelihood peaks at -0.04 and 0 (thin red sliver) in log space.



Even with the noisy sequencing read data, grid search analysis shows that Ancestry HMM is still able to predict the ideal parameters of log10(1e-4) = -4.0 GC length and log10(0.9) = -0.0458 GC length relatively accurately. The grid search result also demonstrates a similarly well-behaved likelihood surface as before, except with a pronounced bi-modal distribution along the GC length axis. We can see that the peaks differ substantially with regard to the GC fraction, with the 2nd peak representing invalid GC fraction of log10(1) = 0. This introduces the possibility that constrained optimization may be necessary for learning the optimal valid GC parameters, although the initial simplex of 0.5 GC fraction and 0.005 GC length yielded successful convergence of the optimal values for the noisy read dataset (Fig 14). However, given that both

recombination and gene conversion must occur along a genome, GC fraction equal to 1 is not a permissible value and constrained optimization is justifiable.

Figure 14: Nelder Mead Parameter Optimization Using *D. melanogaster* Data
3D surface plot of log-likelihood objective function along Z-axis, GC fraction along Y-axis, and GC length along X-axis. Scatterplot of black markers shows the learning path with a nominal cluster at the surface peak representing parameter convergence. Diploid individuals sampled from the population 100 generations after admixture pulse.



Using the SciPy optimize.minimize function with Nelder Mead learning algorithm, Ancestry HMM with Gene Conversion was able to predict parameters of 0.9047 GC fraction and 1.17e-4 GC length. These results demonstrate relative errors of 0.522% GC fraction and 17% GC length, respectively.

4. Discussion

Our results clearly illustrate that Ancestry HMM predicts meiotic crossover and meiotic gene conversions in simulated chromosomes sampled between 10 and 1000 generations, but there is still a lot of validation and investigation necessary going forward. The Tag SNP track represented an ideal case where the interval was very fine and regularly spaced, so naturally we will adjust these to irregularly spaced intervals closer to 1Kb that is closer to the human rate of heterozygosity. We believe that the GC length distributions are sufficiently different for the model to make good predictions up to 1000 generations, and this can be extended further based on larger genomes in real species with *D. melanogaster* chromosomes spanning 20-35Mb. SELAM can easily simulate more chromosomes in the admixed population, but we will further tune the sub-population divergence to see where the haplotype frequencies are too similar to reliably be detected. The GC model currently assumes a single admixture between 2 ancestral populations, but we need to ensure that it generalizes to more pulses and ancestral populations. Finally, the model can learn the GC fraction and length parameters with a high level of accuracy which theoretically suggests that studies of sufficiently divergent species sub-populations will produce novel scientific findings.

5. Conclusion

Based on the favorable results in our simulated data, particularly the log-likelihood surface characteristics, we are able to employ a Nelder-Mead algorithm that allows our model to learn the parameters by maximizing the likelihood. This is exciting because it will allow us to make predictions using real biological data for which the generating parameters are unknown. We need to be selective with the species choice, though, because our model certainly has detection limitations when sub-population divergence is insufficient, GC and recombination tract lengths have distributions that are too similar, admixture occurred too long ago, etc. The first species we are targeting for genetic mapping is swordfish **(***Xiphias gladius*) because our collaborator at Stanford has evidence that their populations exhibit favorable properties, and this keystone species is poorly understood. Additionally, we hope that Ancestry HMM with Gene Conversion will predict genetic maps that can be used in de novo genome assembly. The ancestral switching points can be predicted from short read sequencing data and used as scaffold markers for long-read assembly of a chromosome level reference genome.

Acknowledgements

I thank Dr. Russell Corbett-Detig for entrusting me with implementation and testing of this Ancestry HMM revision. The project is his brain child. Progress has been exciting with his leadership and guidance, plus the model shows a lot of promise. Thanks to members of the Corbett-Detig Lab whose knowledge sharing and ideas continue to push my scientific understanding forward. We appreciate feedback from our collaborator Dr. Rasmus Nielsen at UC Berkeley who confirmed that GC tracts can recombine into ancestral tracts and should be included in the model. Dr. Richard (Ed) Green's BME-130 Genomes course, particularly the population genetics component, provided a foundation for improving my understanding of Coalescent Theory. Dr. David Bernick's BME-205 course in Bioinformatics Algorithms introduced me to Hidden Markov Models that was critical for my contributions to Ancestry HMM. I would also like to thank fellow UCSC student Derek Quong for his thoughtful edits of my drafts throughout the quarter. Everything reads much better with his input. Finally, thanks to CSE-185 Professor Gerald Moulds and my Teaching Assistant Aaron Hunter for their instructive feedback on my technical writing projects throughout the quarter at UC Santa Cruz.

Citations

[1] Arthur J. Hilliker, George Harauz, Andrew G. Reaume, Mark Gray, Stephen H. Clark, Arthur Chovnik, Meiotic Gene Conversion Tract Length Distribution Within the *rosy* locus of *Drosophila melanogaster*, *Genetics*, **137**, (1019-1026), (1994).
The Genetics Society of America

[2] Katharine Korunes, Mohamed Noor, Gene Conversion and Linkage: Effects on Genome Evolution and Speciation, *Molecular Ecology* **26.1**, (351–364), (2017), Web.
https://onlinelibrary-wiley-com.oca.ucsc.edu/doi/full/10.1111/mec.13736

[3] Austin Burt and Robert Trivers, *Genes in Conflict : the Biology of Selfish Genetic Elements,* Cambridge, MA: Belknap Press of Harvard University Press, Chapter 7, (2006), Print.

[4] Russell Corbett-Detig, Rasmus Nielsen, Hyun Min Kang(editor), A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy, *PLoS Genetics*, **13**, 1, (2017).

[5] Paloma Medina, Bryan Thornlow, Rasmus Nielsen, Russell Corbett-Detig, Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference, *Genetics*, **210,** 3, (1089-1107), (2018).

[6] Michael P. H. Stumpf, Gilean A. T. McVean, Estimating recombination rates from population-genetic data, *Nature Reviews Genetics*, **4**, (959-968), (2003).
Nature

[7] Carsen Wiuf and Jotun Hein, The Coalescent With Gene Conversion, *Genetics*, **155**, 1, (451-462), (2000).
Genetics

[8] John Wakeley, Coalescent Theory: an Introduction, Greenwood Village, Colo: Roberts & Company Publishers, (2008). Print.

[9] Russell Corbett-Detig and Matt Jones, SELAM: Simulation of Epistasis and Local Adaptation During Admixture with Mate Choice, *Bioinformatics*, **32.19**, (3035–3037), (2016), Web.

[10] Mason Liang and Rasmus Nielsen, The Lengths of Admixture Tracts, *Genetics*, **197.3**, (953–67), (2014), Web.