INFERRING ADAPTIVE INTROGRESSION USING HIDDEN MARKOV MODELS

Jesper Svedberg[1], Vladimir Shchur[2], Paloma Medina[1], Rasmus Nielsen[2], Russell Corbett-Detig[1]

1. Department of Biomolecular Engineering, Genomics Institute, UC Santa Cruz, CA.
2. Department of Integrative Biology, UC Berkeley, CA.

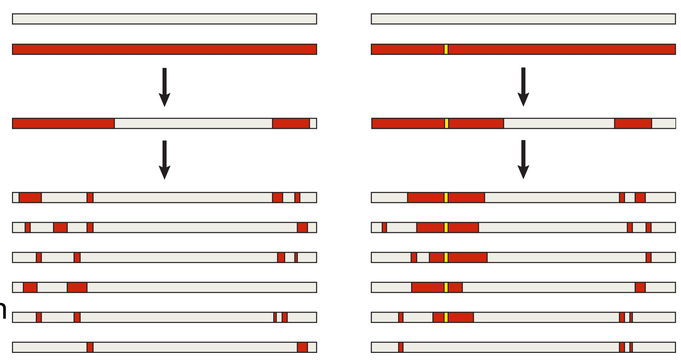
• **Adaptive introgression** are attracting more and more scientific interest and it has been implicated in a number of cases of adaptation, from pesticide resistance to immunity and local adaptation.
• Methods for either **identifying introgression** or for **inferring and quantifying** selection are available, but they generally use similar signals, making it difficult to untangle the contribution of each phenomenon on patterns of variability across the genome.
• This has made it difficult to evaluate the role of selection on introgressed tracts on a genome-wide scale and it is currently not known how common adaptive introgression is or to what extent it drives introgression in general.
• We are **developing a method** which uses a **Hidden Markov Model framework** to infer both introgression and selection at the same time, thus solving previous issues with separating the to factors.
• Given a population genomic dataset of a population with a known level of admixture and reference panels of each parental population, **we can identify adaptively introgressed regions in the genome and estimate the strength of selection that has acted upon them.**
• A version of this method has been **implemented in C++** and using simulated data we can identify both the genomic location and the selective coefficient with reasonable precision. The method is fast enough for genome wide scans of real population genomic datasets, suggesting that it can be of great use for the population genomics community.

ADAPTIVE INTROGRESSION

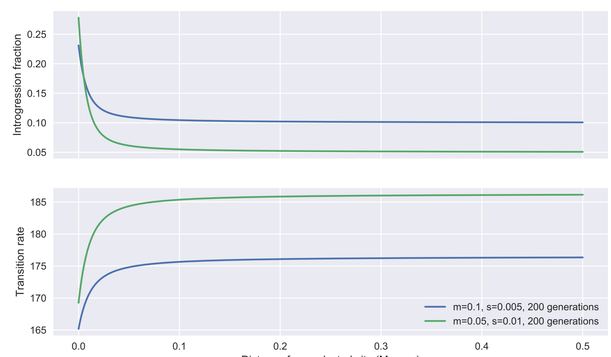
Adaptive introgression is the phenomenon where adaptive alleles are introduced into a species or population through interspecies or interpopulation gene flow. With the increasing amount of population genomic data, it has also become increasingly clear that interspecies gene flow in the shape of hybridization and introgression/admixture is a common phenomenon in nature. Some genetic material can spread in a population following an admixture event because it confers a selective advantage, so called “adaptive introgression”, and this is thought to be one of the three major sources of adaptive genetic variation, next to mutations and standing genetic variation.

INFERRING ADAPTIVE INTROGRESSION

Following an admixture event, recombination will break up introgressed haplotypes over time. In the absence of selection, the frequency of the introgressed genotype is expected to be low and haplotype lengths short (left figure). If positive selection is acting upon an introgressed locus, the genotype frequency is expected to be higher, and the haplotype lengths larger (right figure).

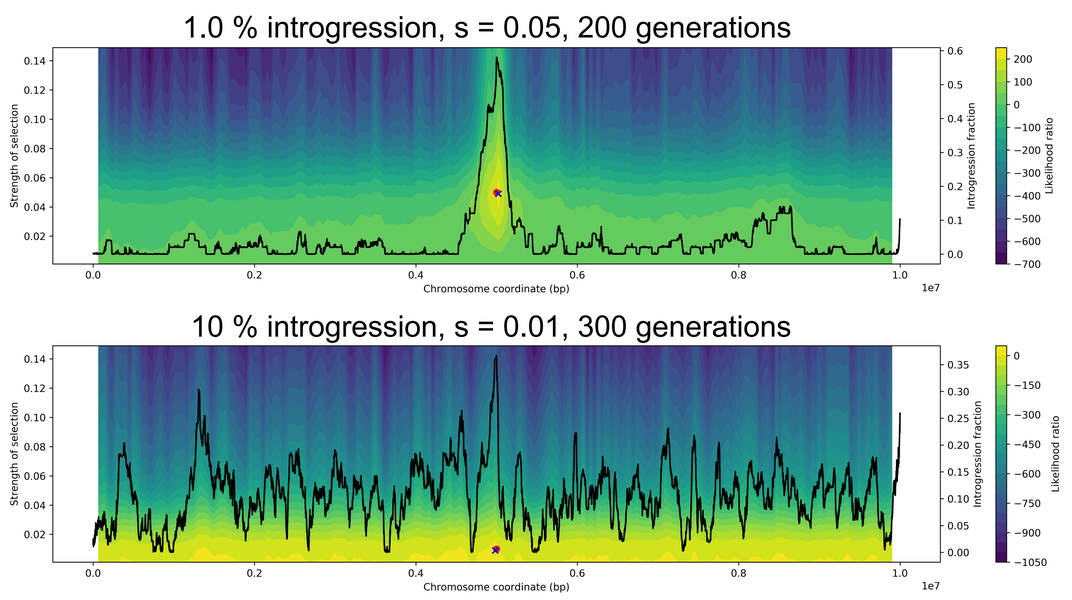


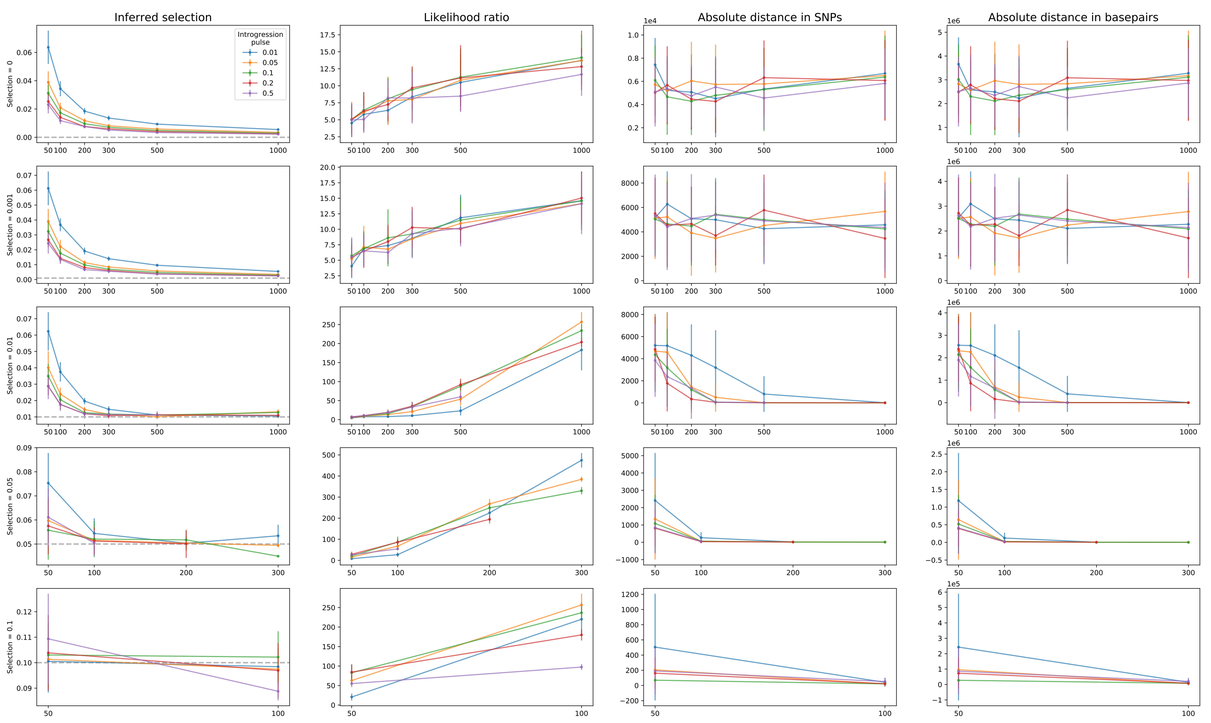
There are methods available that can identify such selected loci and quantify the selective coefficient based on haplotype lengths and frequencies, but they are not adapted for handling the signal of introgression, making them less precise.



IMPLEMENTATION

Using Hidden Markov Models, we can handle the effects of both introgression and selection at the same time, and both identify loci that have been adaptively introgressed and estimate the selective coeffients. This is done by calculating expected genotype frequencies around the selected site, convert them into Markovian transition rates (figure to the left) and use these rates in a Hidden Markov Model. The HMM is used to compare real data to expected patterns and from that identify loci that are likely to be adaptively introgressed and estimate the strength if selection. This is an extention of the software Ancestry_HMM.





LIKELIHOOD SURFACES OF SIMULATED POPULATIONS

We ran simulations of a population of 10,000 diploid individuals with a single 10 Mbp chromosome, which were introgressed at 1% or 10% from a second population 50-1000 generations ago using SELAM. The introgressed haplotype carries a locus with a selective coefficient s of 0.01 or 0.05 at genomic location 5 Mbp. The HMM was run and likelihood evaluated at every 100 variable sites (~20 kbp) for values of s between 0.001 and 0.15 with a 0.001 step. The likelihood surface is shown in yellow-to-blue. The black line shows the frequency of the introgressed genotype across the chromosome. Estimated values of s and chromosomal location (blue x) correspond well to the actual values (red circle).

VALIDATION

We simulated several adaptive introgression scenarios using SELAM by varying the selective coefficient (s), the size of the introgression pulse (m) and by sampling the simulation at different time points. The selective coefficient was varied between 0 (neutral case) and 0.1. The introgression pulse was varied between 0.01 and 0.5. Twenty simulations consisting of 100,000 individuals for each combination of s and m were sampled at 50, 100, 200, 300, 500 and 1000 generations after the introgression event. We only plot time points where the introgressed genotype did not exceed 99.9%. The plots show results for the SNP with the highest likelihood ratio (i.e. likelihood of selection compared to no selection). We plot the inferred selective coefficient, the likelihood ratio and the distance from the true site in numbers of variable sites and in basepairs. Our method works better for higher strengths of selection (>0.001) and when more time has passed since the introgression event and especially inferring the correct genomic location becomes more difficult when these conditions are not met. The method is reasonably fast and a 10 Mbp chromosome consisting of 20,000 variable sites (modelled after D. melanogaster populations) takes 4-6 hours to run. Further validation work is ongoing.

CONCLUSIONS AND PROSPECTS

• Our Hidden Markov Models can be used to infer adaptive introgression from simulated populations.
• Validation shows that it can both identified adaptively introgressed loci and infer the strength of selection under simulated realistic scenarios.
• Further validation work is ongoing to determine sensitivity to for instance type and amount of data.
• We want to use our method to estimate the strength of selection of known or suspected cases of adaptive introgression.

E-mail: jsvedber@ucsc.edu, Twitter: @jesvedberg