Delimitation of contiguous regions of differentiation using Hidden Markov Models

https://visoca.github.io/popgenomworkshop-hmm

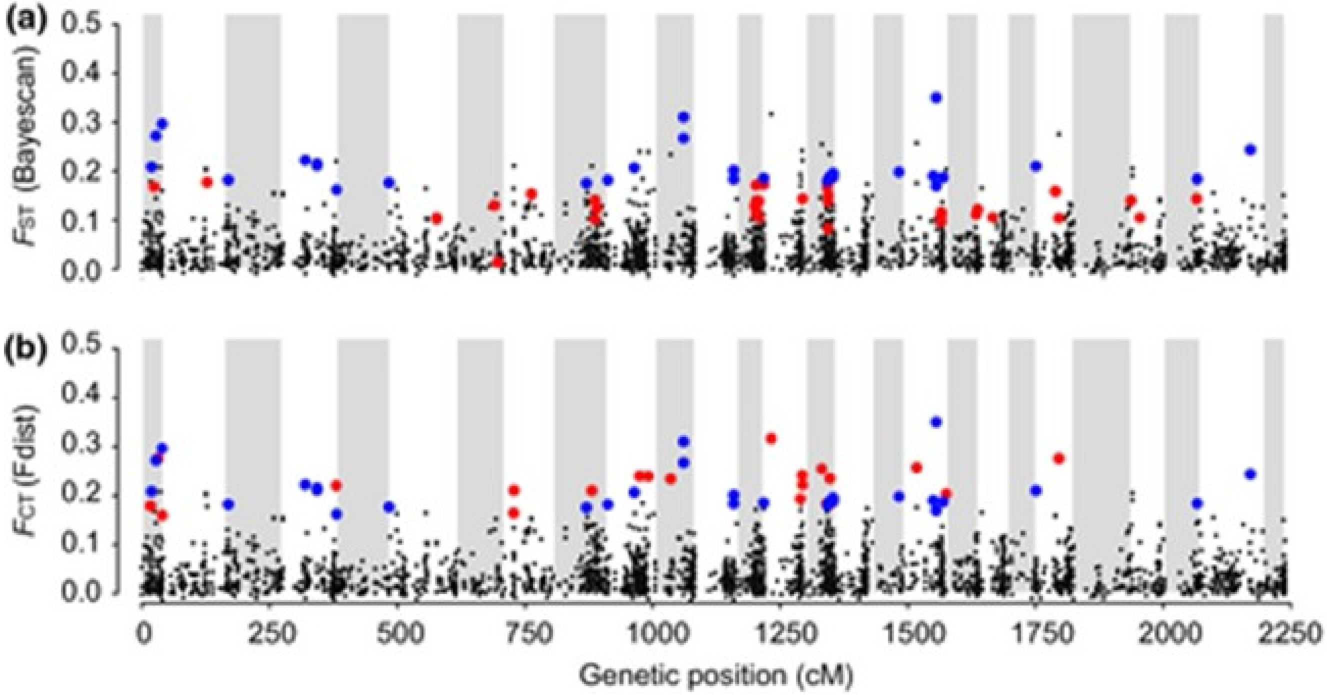
Víctor Soria-Carrasco
v.soria-carrasco@sheffield.ac.uk
Leverhulme Early Career Fellow

The Leverhulme Trust
The University Of Sheffield.
nbaf

Detecting selection

Locus by locus genome scans - outlier tests

- Generally using some sort of population genetics statistic to measure divergence: allele frequency, F_{ST}, D_{xy}, Π
- Some simple, some involving models and simulation
- Usually involving multiple population comparisons



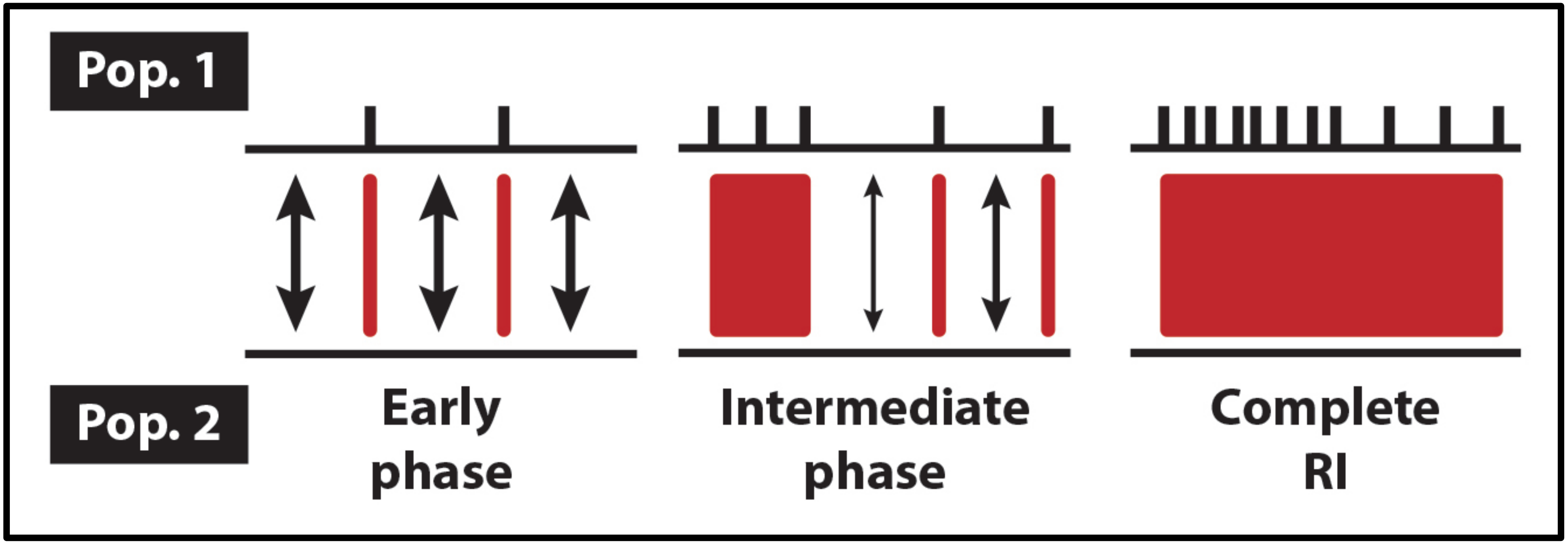
Moore et al. 2014 Molecular Ecology
North American Atlantic salmon - Genome scan among regions

Genomic islands

Genic model of speciation with gene flow

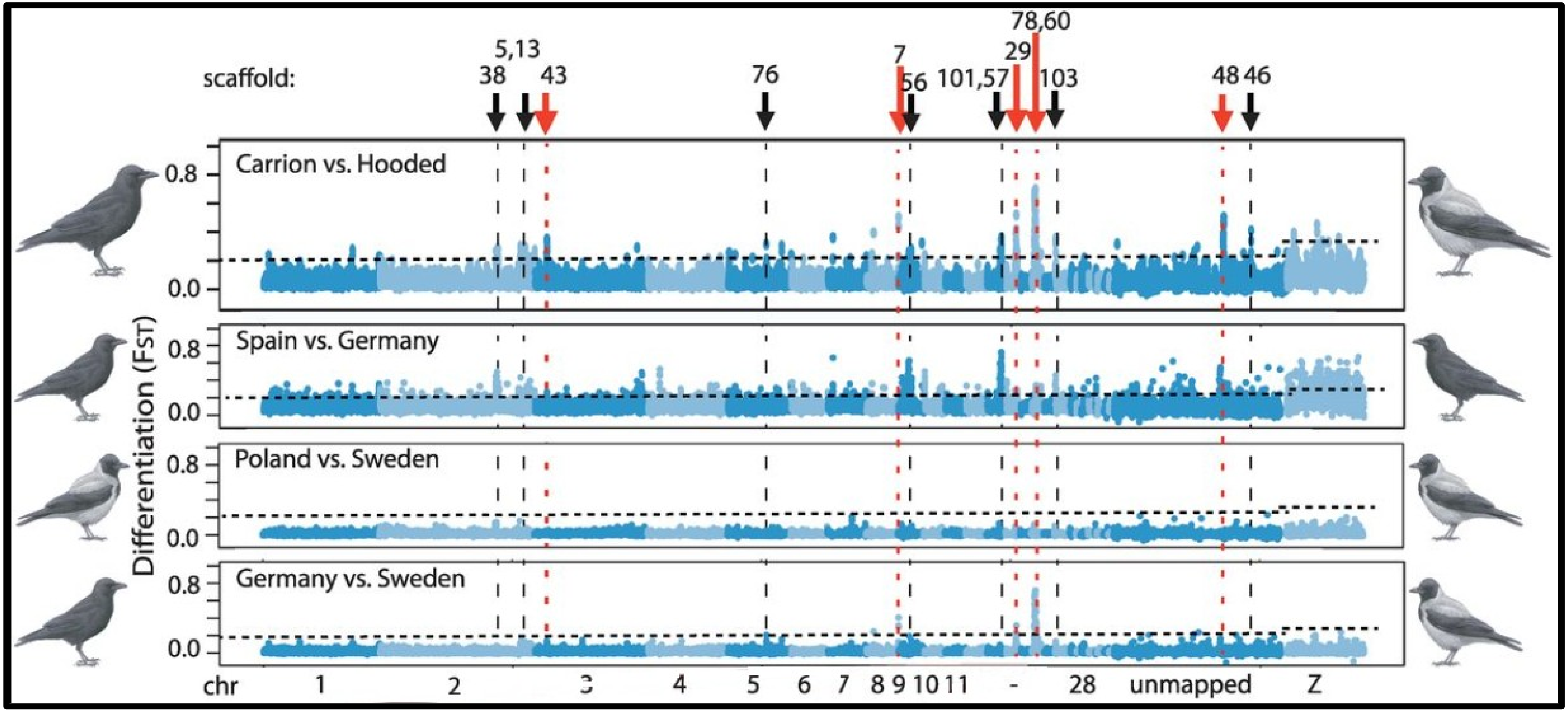
1) Early genic phase: Few localized regions of accentuated differentiation (‘genomic islands’)

2) Intermediate genomic phase: differentiation become genome-wide



Wu 2001 J Evol Biol; Mallet 1995 TREE

Genomic islands



Poelstra et al. 2014 Science
50Kb sliding windows

Hidden Markov Models

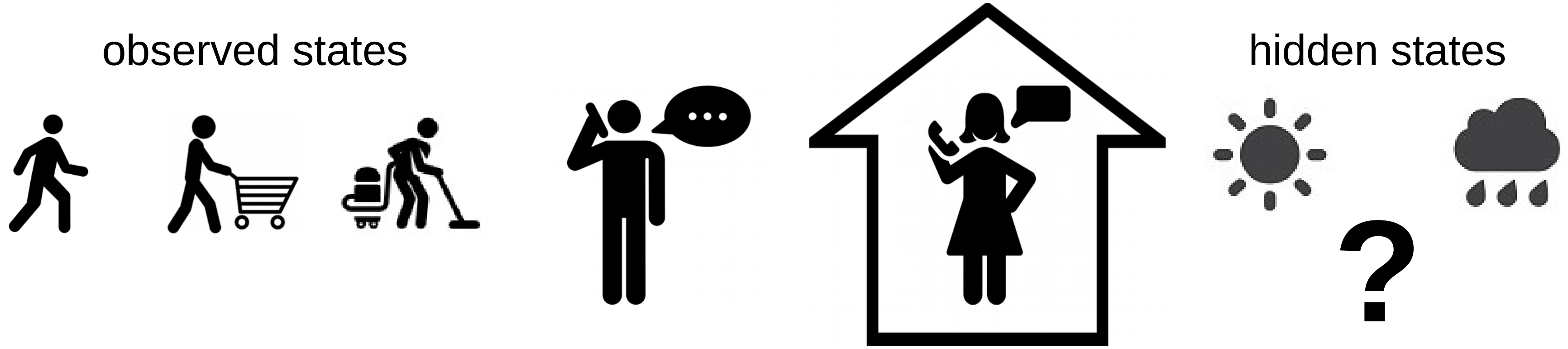
Sliding window approaches issues:

- Choice of window size not trivial
- Window size can have strong impact on number and size of regions
- Random fluctuations of the test statistic in a delimited window might lead to the detection of a cluster when there is none

Hidden Markov Models (HMM)

- Probabilistic models for linear sequence 'labeling'
- Statistical model in which the system is modeled as a Markov process with hidden states (Markov process: probability of subsequent state depends only on previous state)
- Explicitly model dependencies among neighbouring markers

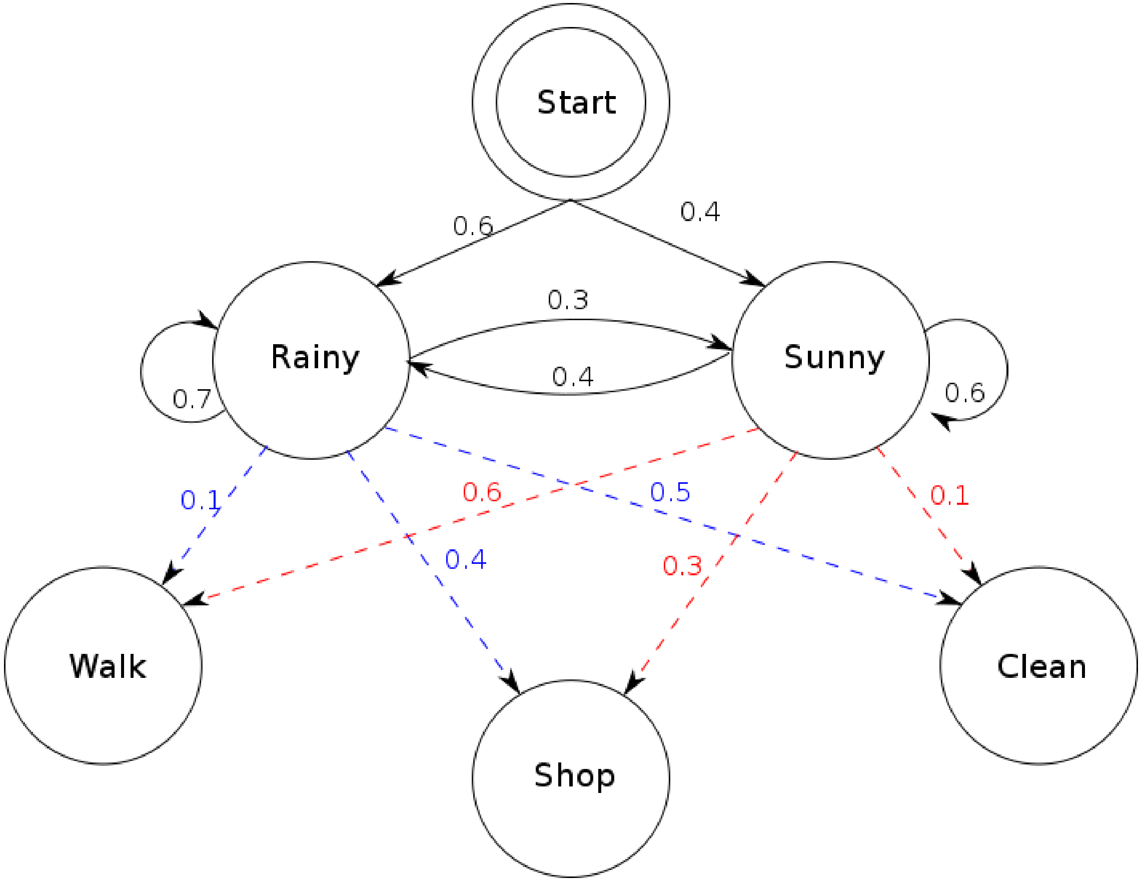
Hidden Markov Models



observed states

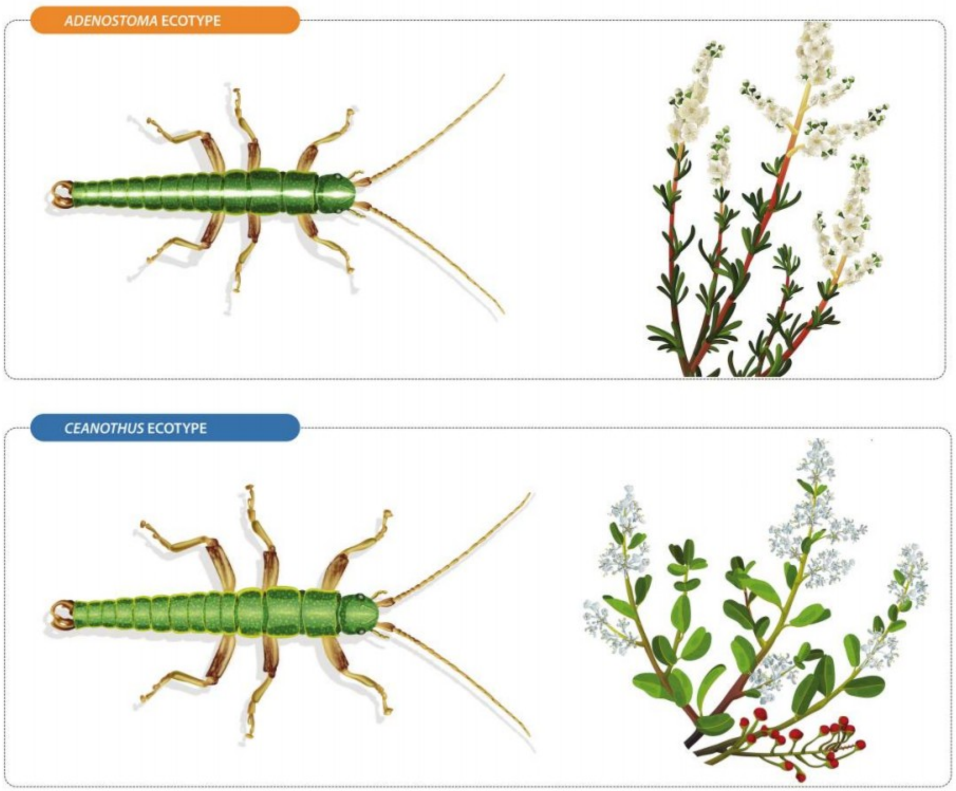
hidden states

?



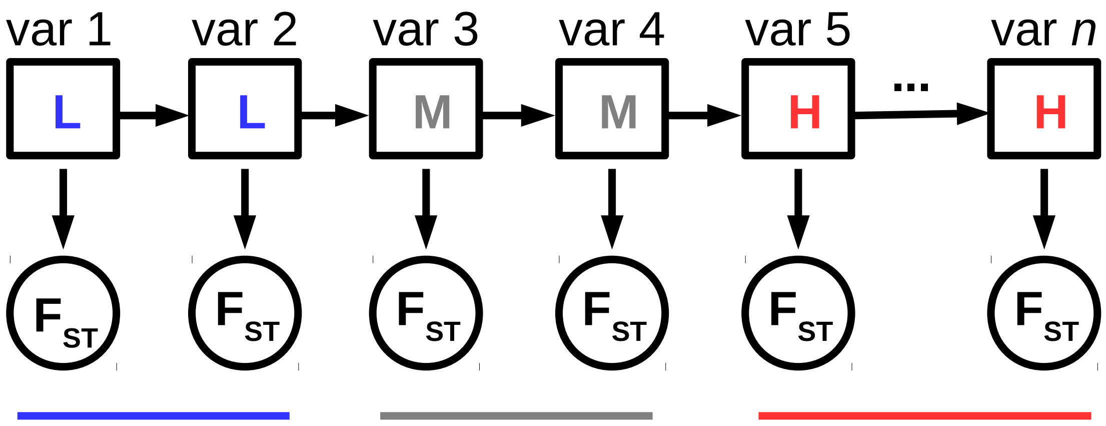
https://en.wikipedia.org/wiki/Hidden_Markov_model

Example: genetic differentiation in *Timema* stick insect ecotypes



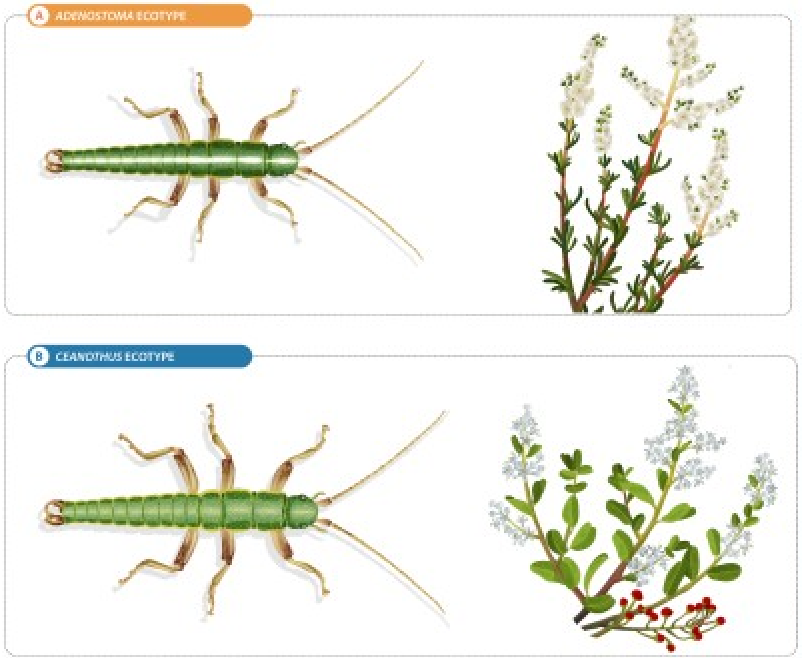
Observed: F_{ST} across genome (SNPs)

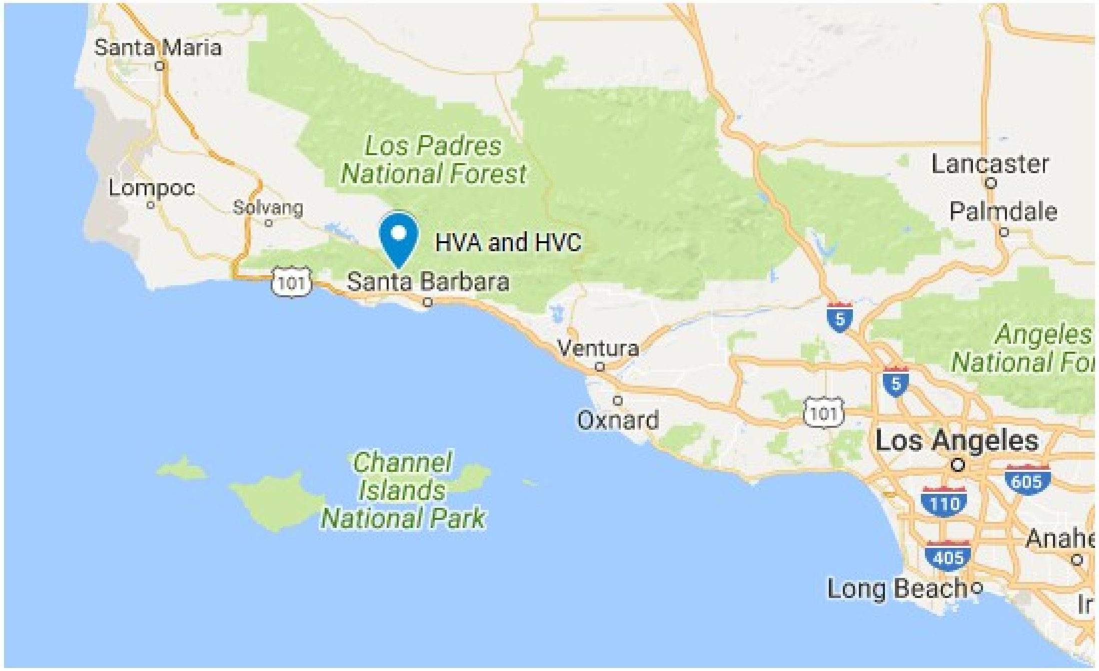
Hidden: 3 differentiation states – low (L), medium (M), and high (H)



delimitation contiguous genetic regions
number, size, distribution

Practical





Whole genome sequence data

20 individuals HVA

20 individuals HVC

Procedure

Infer allele frequencies from genotype likelihoods

using an implementation of the iterative soft expectation-maximization algorithm (EM) described in Li 2011 (code kindly provided by Zach Gompert, Utah State University).

Estimate F_{ST} from allele frequencies

using the F_{ST} Hudson's estimator (as described in Bhatia *et al.* 2013, SI):

$$F_{ST}^{Hudson} = 1 - \frac{Hw}{Hb} = \frac{p_1(1-p_1) + p_2(1-p_2)}{p_1(1-p_2) + p_2(1-p_1)}$$

where *Hw* is the within-population heterozygosity, *Hb* is the between-population heterozygosity, and p_1 and p_2 are the allele frequencies in each population.

- Bhatia et al. 2013 Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Research*, 23(9), pp.1514–1521. http://goo.gl/TqWnur
- Li 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), pp.2987–2993. http://goo.gl/yTSYjn

Procedure

Fit a 3-state discrete homogenous Hidden Markov Model (HMM)

using the R package HiddenMarkov to classify the genome in regions of high, medium, and low differentiation (hidden states) from F_{ST} (observed states)

assuming distribution of F_{ST} for each hidden state follow a normal distribution with standard deviation fixed to the genome-wide standard deviation

disallowing direct transitions between extremes (from low to high or high to low)

Go!

https://visoca.github.io/popgenomworkshop-hmm/