# EvoGen workshop 2:
# The basics of selection and demographic analysis


Phred Benham

Email: phbenham@gmail.com

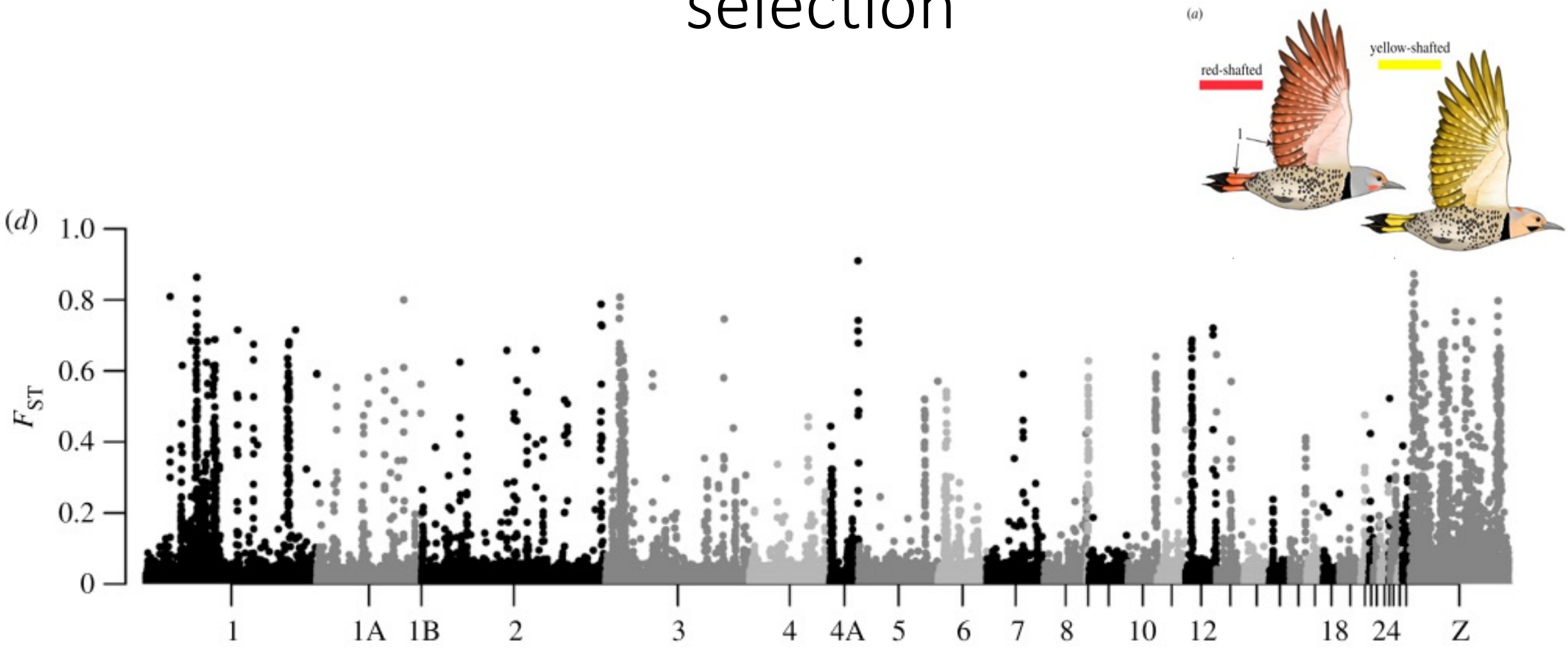

# Outline of workshop2

1. Genome scans to detect loci under selection (60 minutes)

2. Genome associations (60 min.)

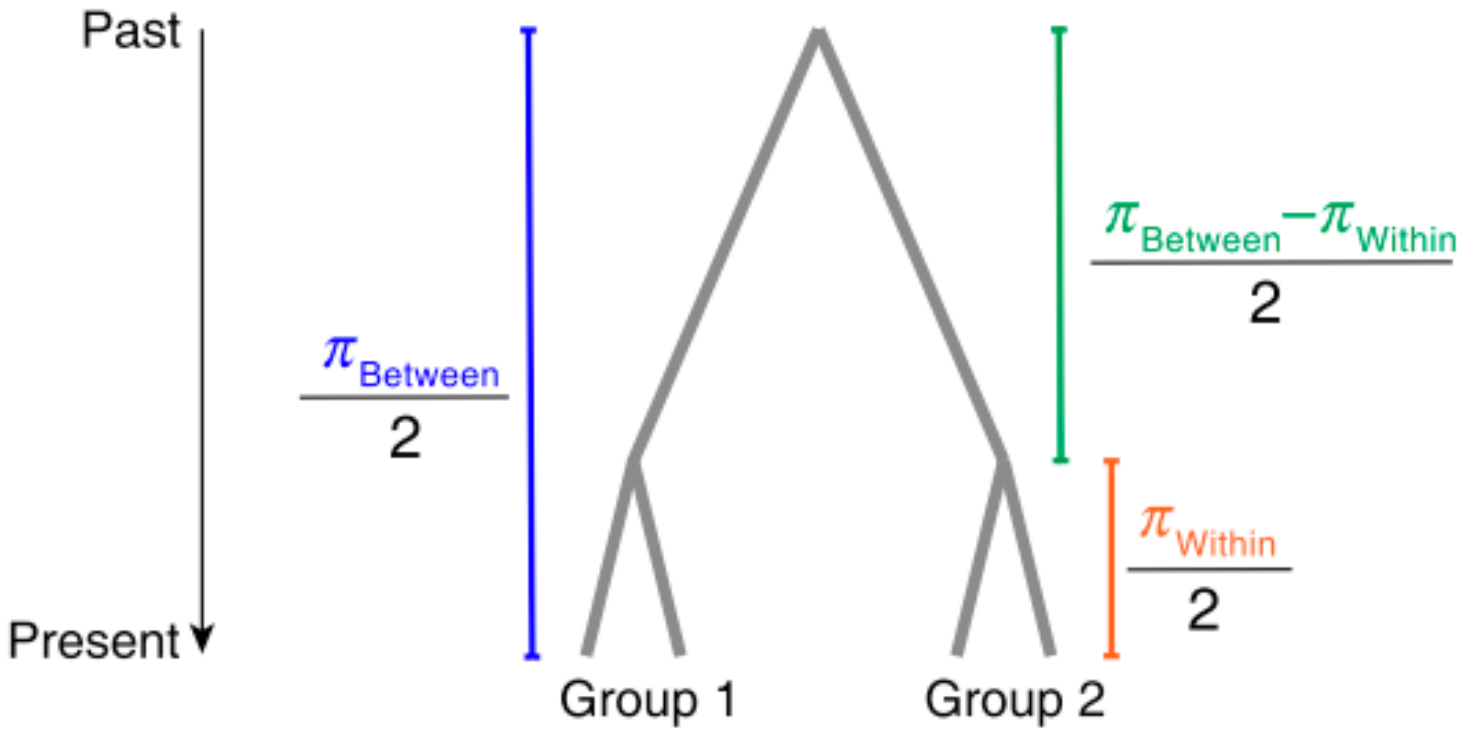
3. The basics of demographic analysis (60 minutes).

# Part 1: Genome scans to detect loci under selection




Aguillon et al. 2021 *Proc. Roy. Soc. B*

# Fst and selection



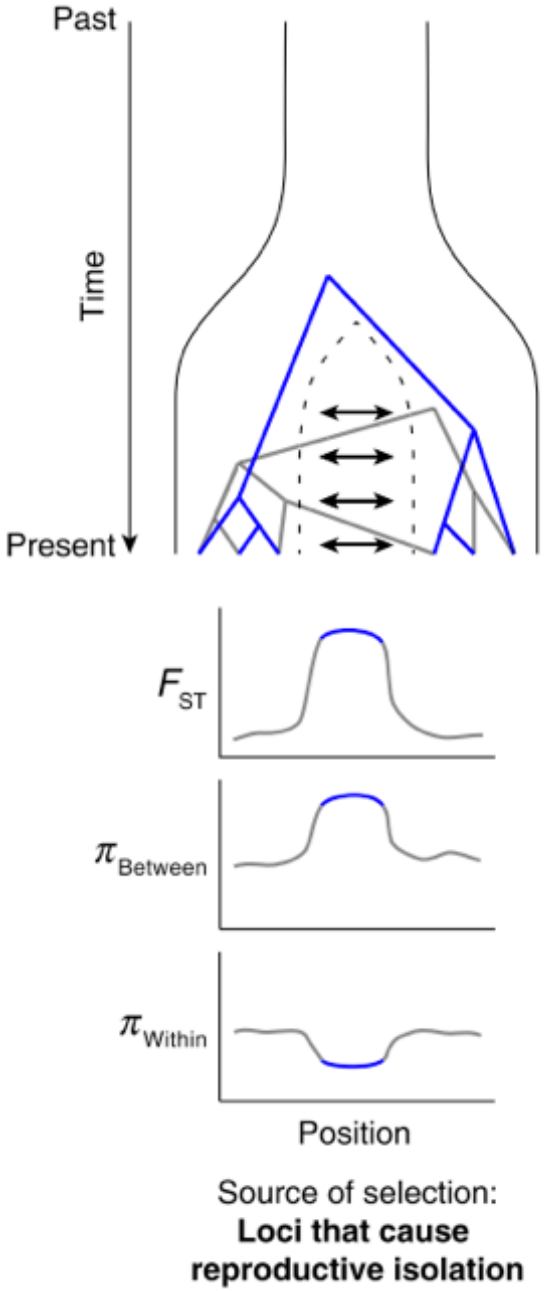

$$F_{ST} = \frac{\pi_{Between}-\pi_{Within}}{\pi_{Between}}$$

$$F_{ST} = 1 - \frac{\pi_{Within}}{\pi_{Between}}$$

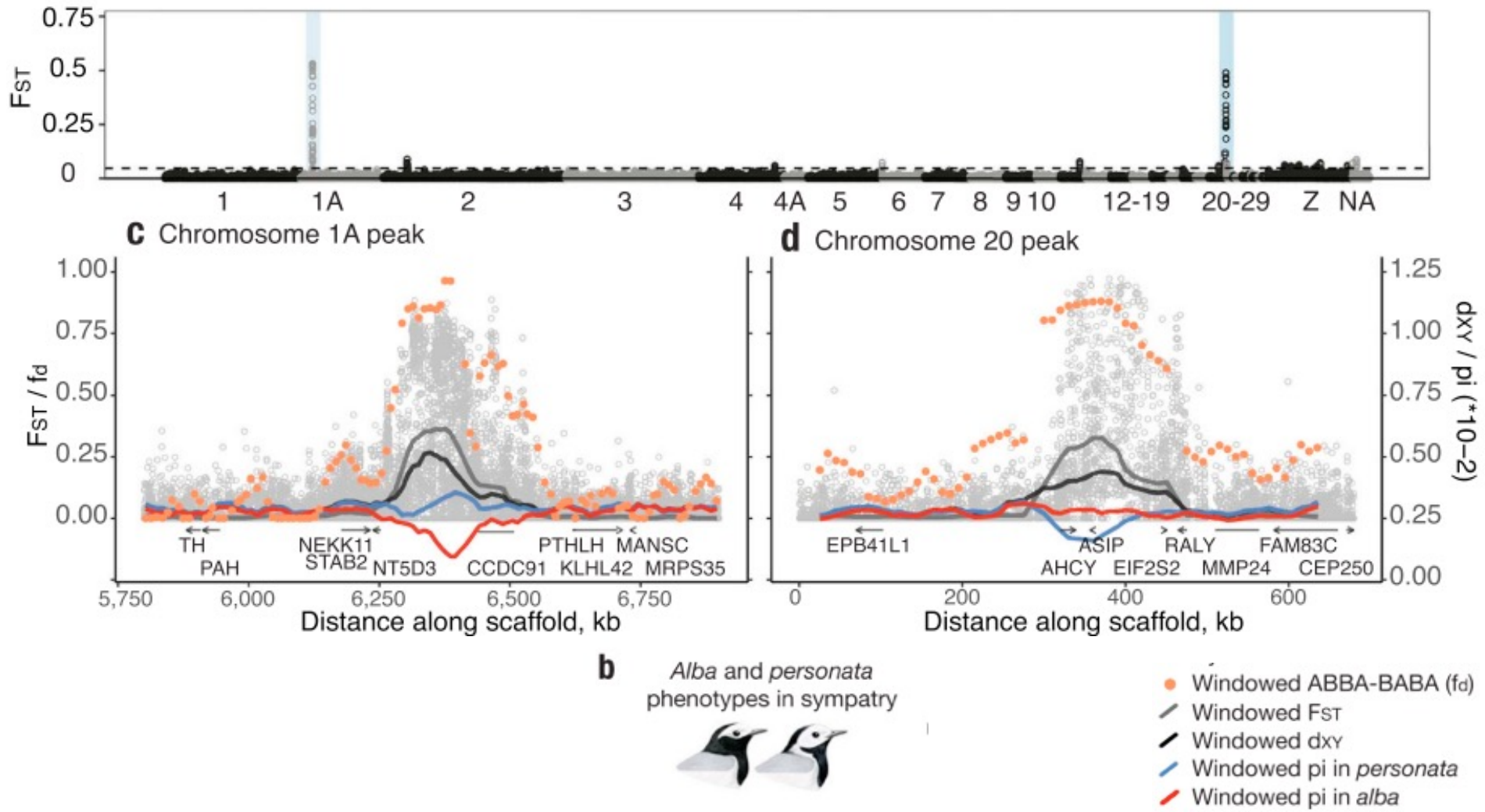
- Both increases in $\pi_{Between}$ and decreases in $\pi_{Within}$ will lead to increased Fst.
- Purifying or background selection leads to elevated Fst by decreasing $\pi_{Within}$.
- Positive selection increases Fst due to increased $\pi_{Between}$ and decreased $\pi_{Within}$.
- Recent divergence, high gene flow decrease $\pi_{Between}$ and decrease Fst.

Irwin et al. 2018 *Molecular Ecology*

# Signatures of positive selection

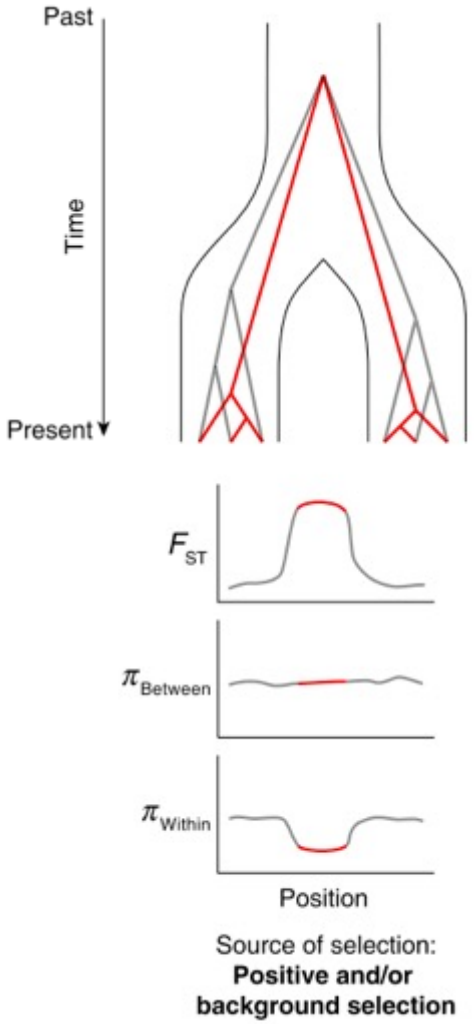



Irwin et al. 2018 *Molecular Ecology*

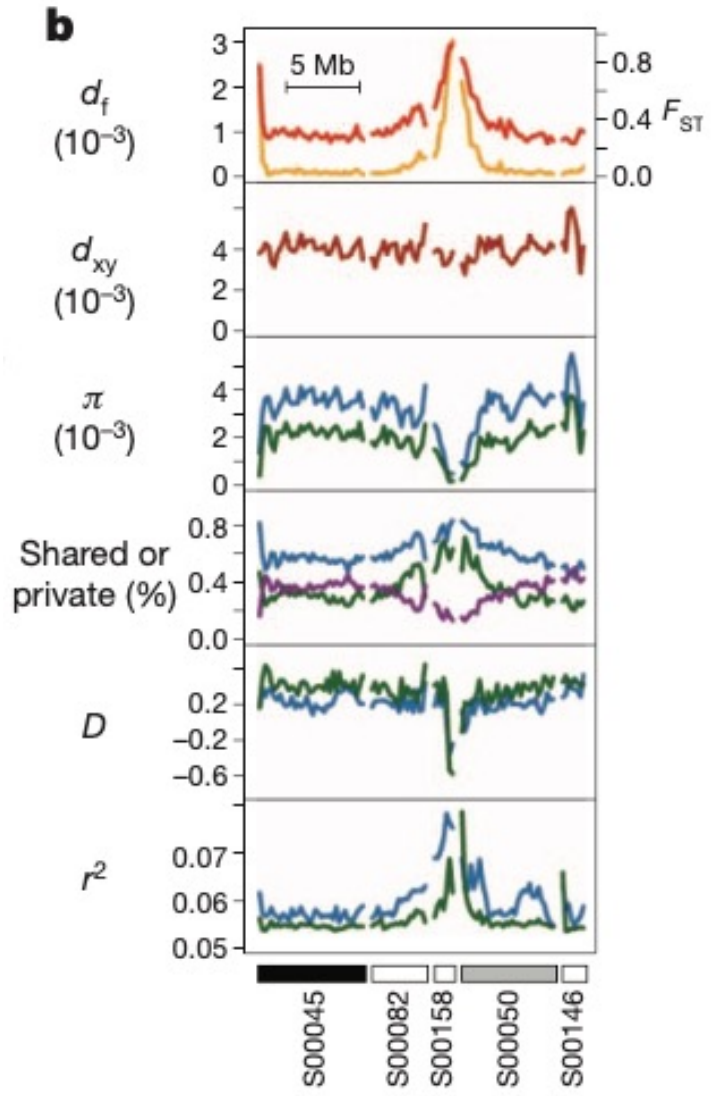



Semenov et al. 2021 *Nature Communications*

# Signatures of purifying or positive selection

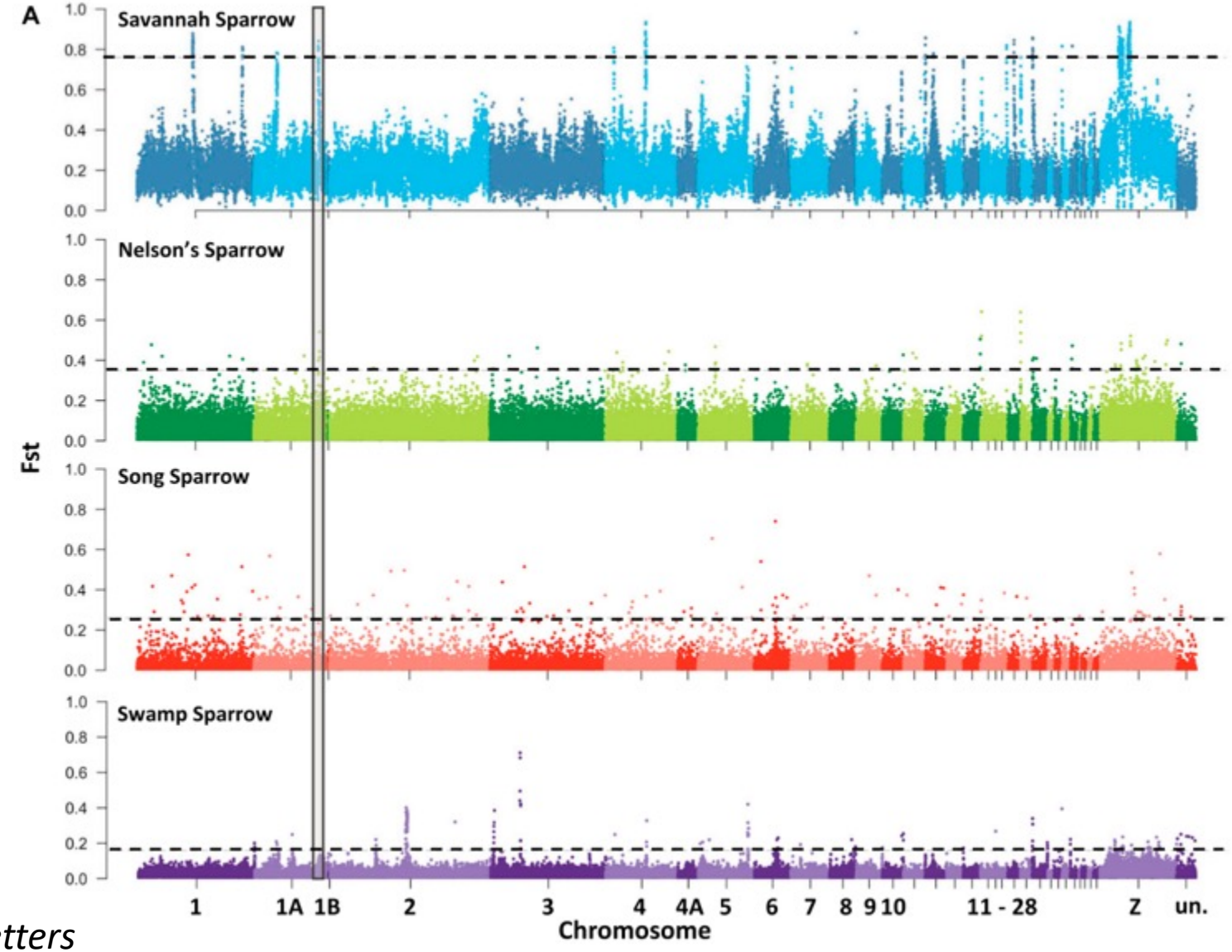



Irwin et al. 2018 *Molecular Ecology;*
Also see Cruickshank & Hahn 2014 *Molecular Ecology*



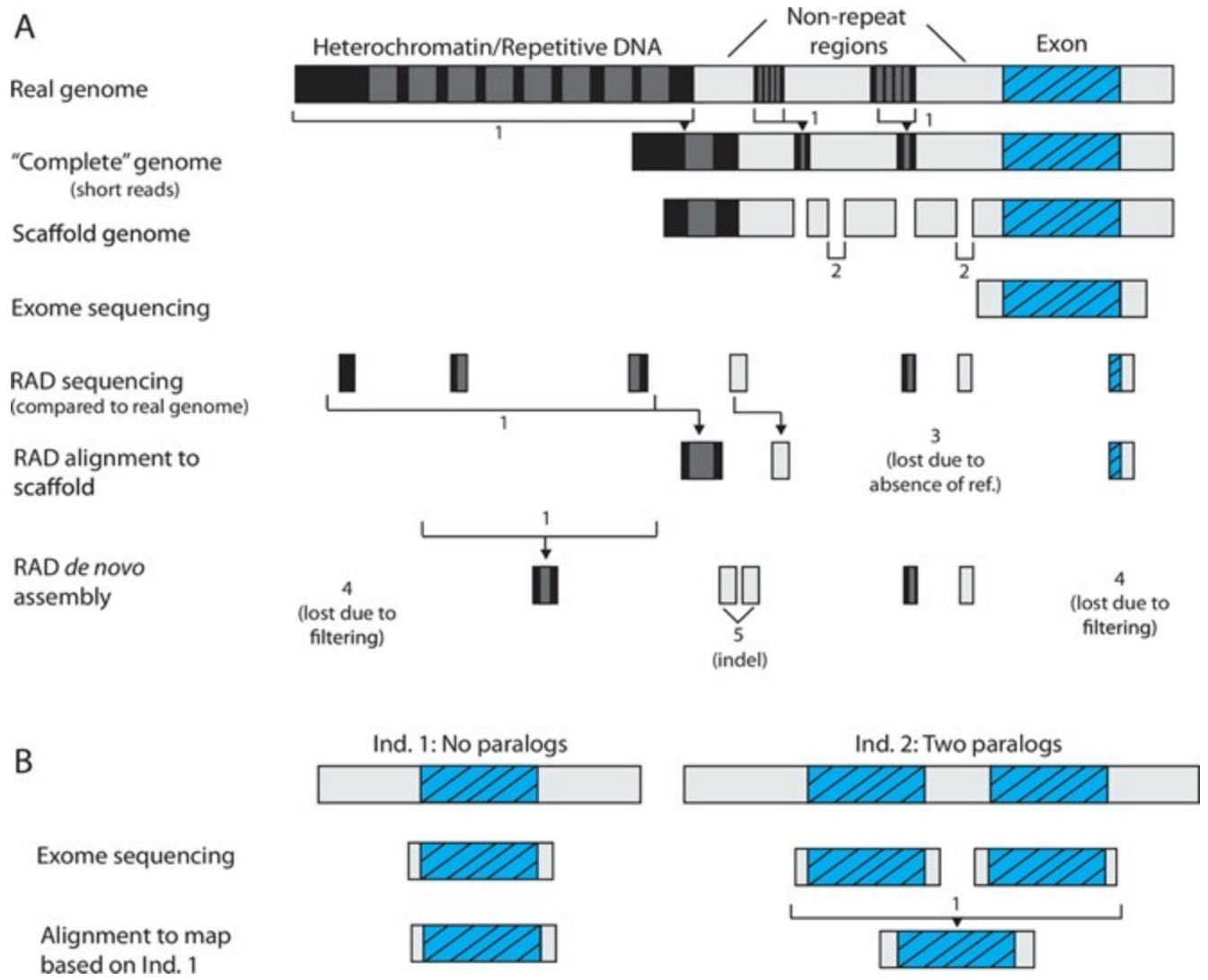

Ellegren et al. 2012 *Nature*

# Practical considerations: sampling design




Walsh et al. 2021 *Evol. Letters*

# Practical considerations: sequencing data




Hoban et al. 2016 *The American Naturalist*

# Practical consideration: Windowed vs per-SNP Fst

**Estimating Fst per SNP**

- May be better for smaller reduced representation datasets.
- Sequencing errors more likely to result in false positives.

**Estimating Fst across windows**

- Accounts for non-independence among SNPs.
- Increase power to ID outliers.
- Reduces sampling noise and false positives.
- Choice of window size will influence results.

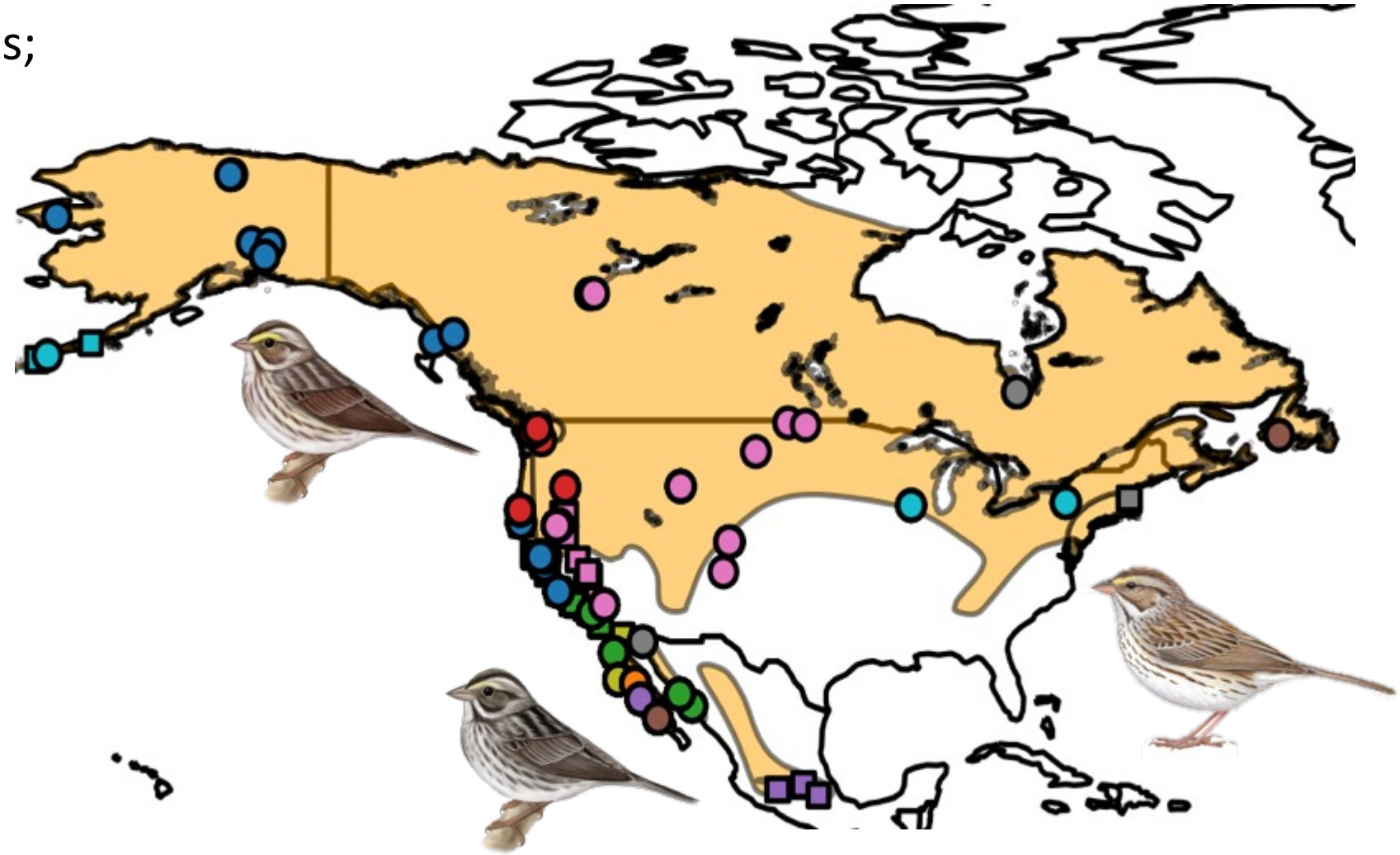
For both approaches a challenge will be determining when an outlier is an outlier

# Example dataset

56,533 exons from 13,813 genes;
3,443 non-coding regions

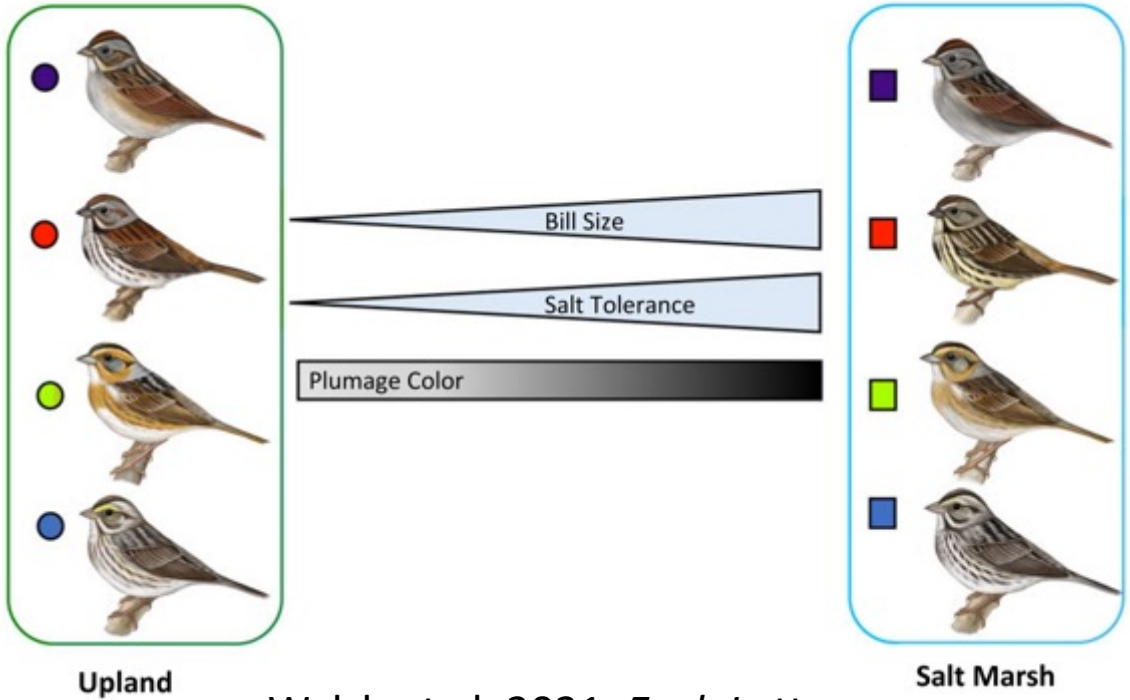
Lots of labwork

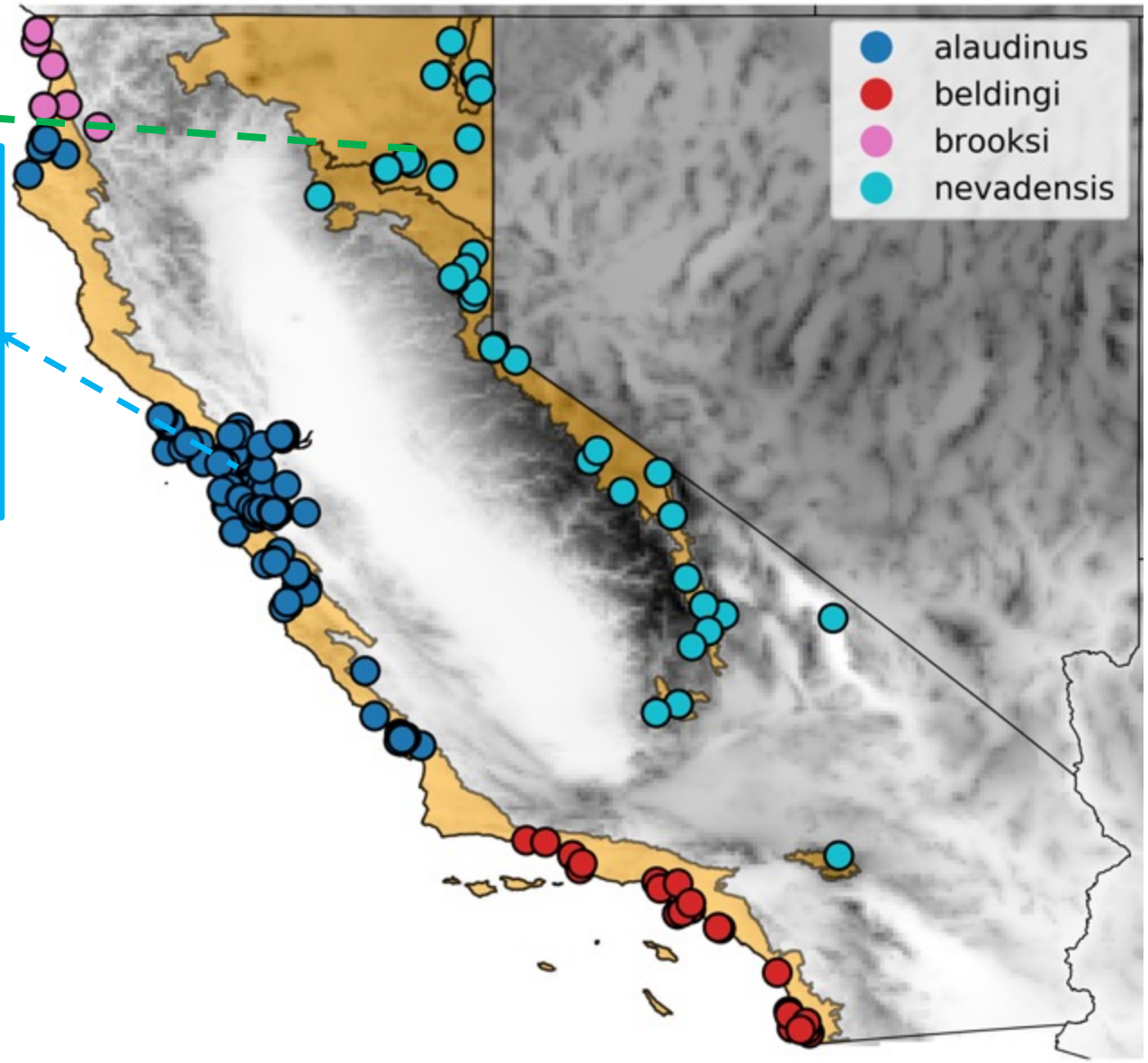
Lots of bioinformatics



Filtered VCF file

https://www.dropbox.com/s/ilzyhv1a4kw3s2k/SAVS_exampleDataset_exercise1.vcf.gz?dl=0




Walsh et al. 2021 *Evol. Letters*

# Genome scans using VCFtools

vcftools –(gz)vcf <input.vcf here> --weir-fst-pop <pop1_list.txt> --weir-fst-pop <pop2_list.txt> \
--fst-window-size <integer window size> --out <output_file_name>

**OUTPUT FST STATISTICS**

**--weir-fst-pop** <filename>

This option is used to calculate an Fst estimate from Weir and Cockerham's 1984 paper. This is the preferred calculation of Fst. The provided file must contain a list of individuals (one individual per line) from the VCF file that correspond to one population. This option can be used multiple times to calculate Fst for more than two populations. These files will also be included as "--keep" options. By default, calculations are done on a per-site basis. The output file has the suffix ".weir.fst".

**--fst-window-size** <integer>
**--fst-window-step** <integer>

These options can be used with "--weir-fst-pop" to do the Fst calculations on a windowed basis instead of a per-site basis. These arguments specify the desired window size and the desired step size between windows.

https://vcftools.github.io/man_latest.html

# Output file from VCFtools

| CHROM | BIN_START | BIN_END | N_VARIANTS | WEIGHTED_FST | MEAN_FST |
|---|---|---|---|---|---|
| 1536 | 1 | 25000 | 22 | 0.00405276 | 0.00588584 |
| 1536 | 50001 | 75000 | 10 | 0.0851365 | 0.0535534 |
| 1536 | 175001 | 200000 | 9 | 0.0133095 | 0.0202818 |
| 1536 | 300001 | 325000 | 20 | -0.00448846 | 0.0020468 |
| 1536 | 375001 | 400000 | 2 | -0.00975324 | 0.0158286 |

# Basic structure of a BED file

| | | | |
|---|---|---|---|
| NW_005081536.1 | 150 | 1137 | 111_NW005081536.1:0-1000_Length=1000 |
| NW_005081536.1 | 50138 | 50479 | 4139_NW005081536.1:50000-51000_Length=356 |
| NW_005081536.1 | 181618 | 182577 | NW_005081536.1:181668-182527:gene=ZNF521 |
| NW_005081536.1 | 315367 | 318850 | NW_005081536.1:315417-318800:gene=ZNF521 |
| NW_005081536.1 | 388400 | 388710 | NW_005081536.1:388450-388660:gene=ZNF521 |

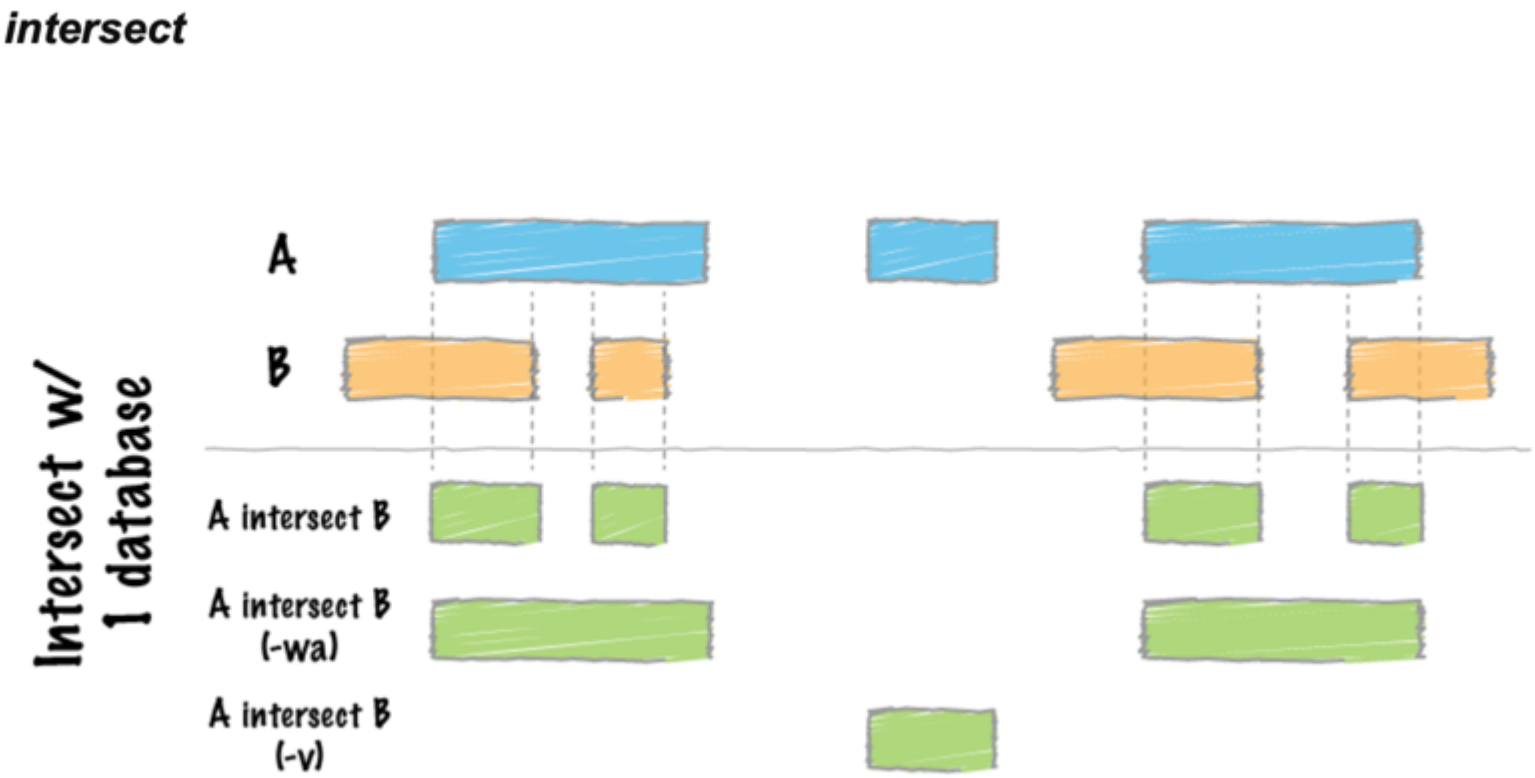
- Many other files have a similar structure including VCF, GFF, GTF and BAM files

# Using bedtools: the swiss-army knife of genomics analysis.

bedtools

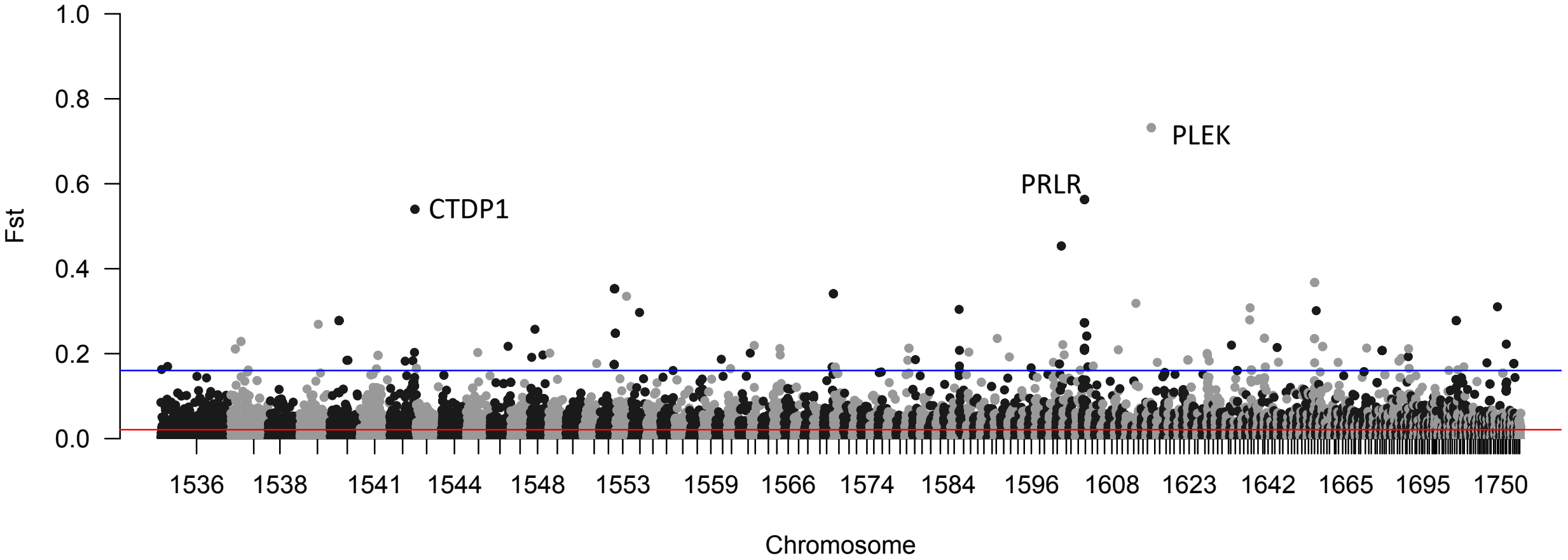
Table of Contents

*intersect*
- Usage and option summary
- Default behavior
- Intersecting against MULTIF
- **-wa** Reporting the original A
- **-wb** Reporting the original B
- **-loj** Left outer join. Report
- **-wo** Write the *amount* of ov




bedtools intersect –a <file1> -b <file2> –wao –new_annotated file.txt

# Making a Manhattan plot



# Final considerations

- A wide range of processes can contribute to false positives in these analyses. Always a good idea to draw inferences from multiple approaches and account for demographic history.

- These kinds of methods work best when applied to two closely related populations still exchanging genes.