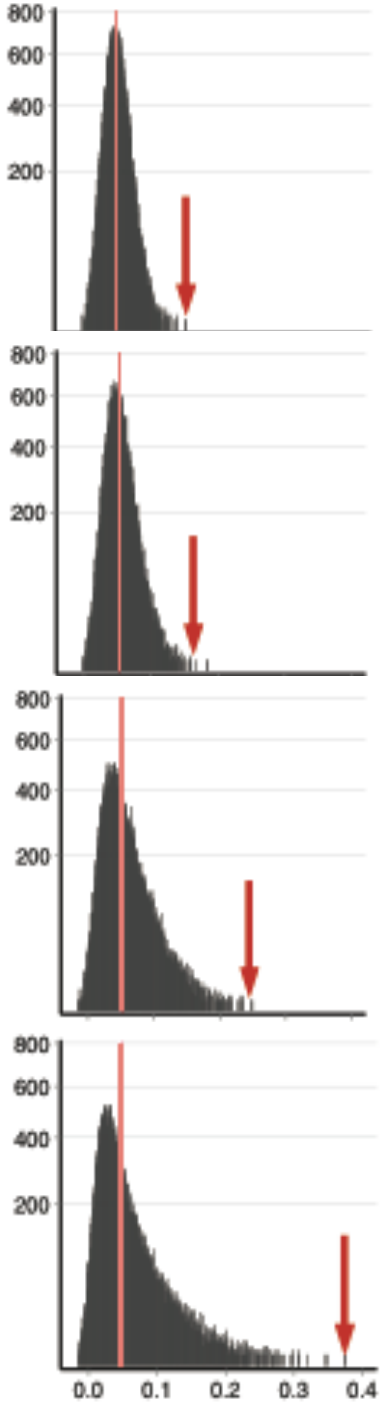


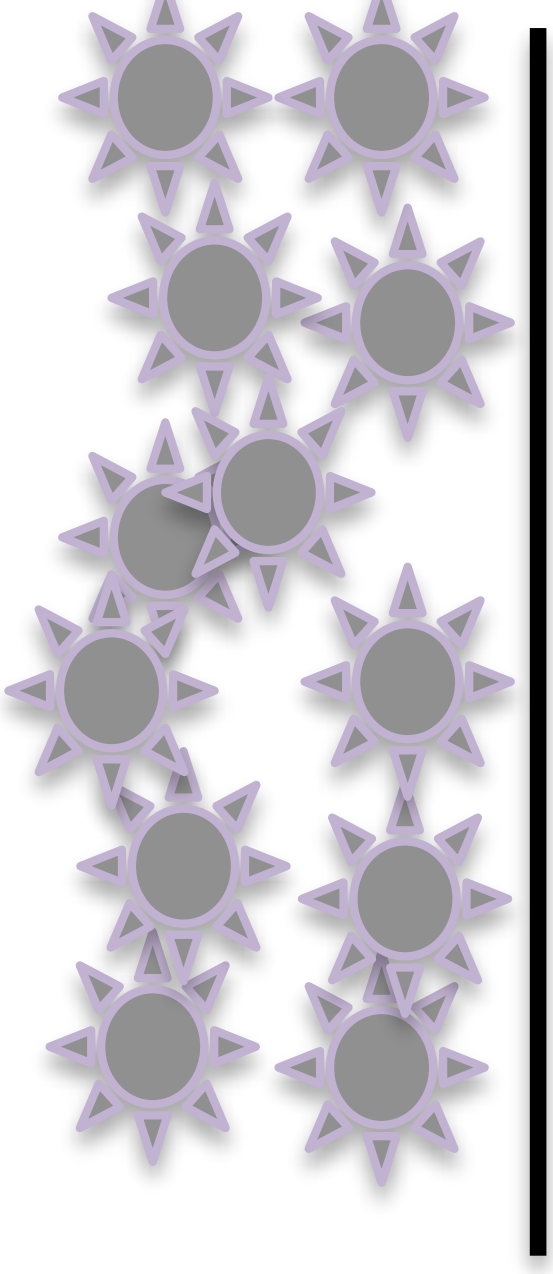
Reducing false positives in F_{ST} outlier tests with OutFLANK

Katie E. Lotterhos[1,2]
Michael C. Whitlock[1]

[1]University of British Columbia
[2]Wake Forest University

Question:

What is the genomic basis of local adaptation in heterogeneous environments?



N **Cold**

S **Warm**

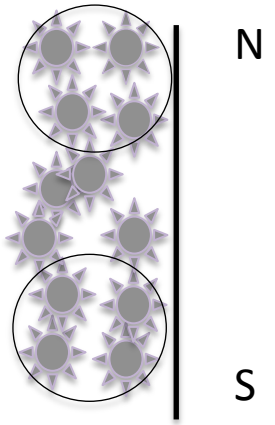
F_{ST} measures how different allele frequencies are between populations

No difference

Complete difference

$$F_{ST} = \frac{var(p)}{\bar{p}(1-\bar{p})}$$

F_{ST} measures how different allele frequencies are between populations



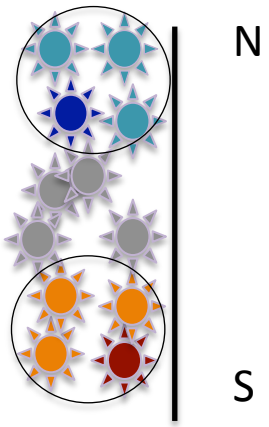
$$F_{ST} = \frac{var(p)}{\bar{p}(1-\bar{p})}$$

Non-adapted locus

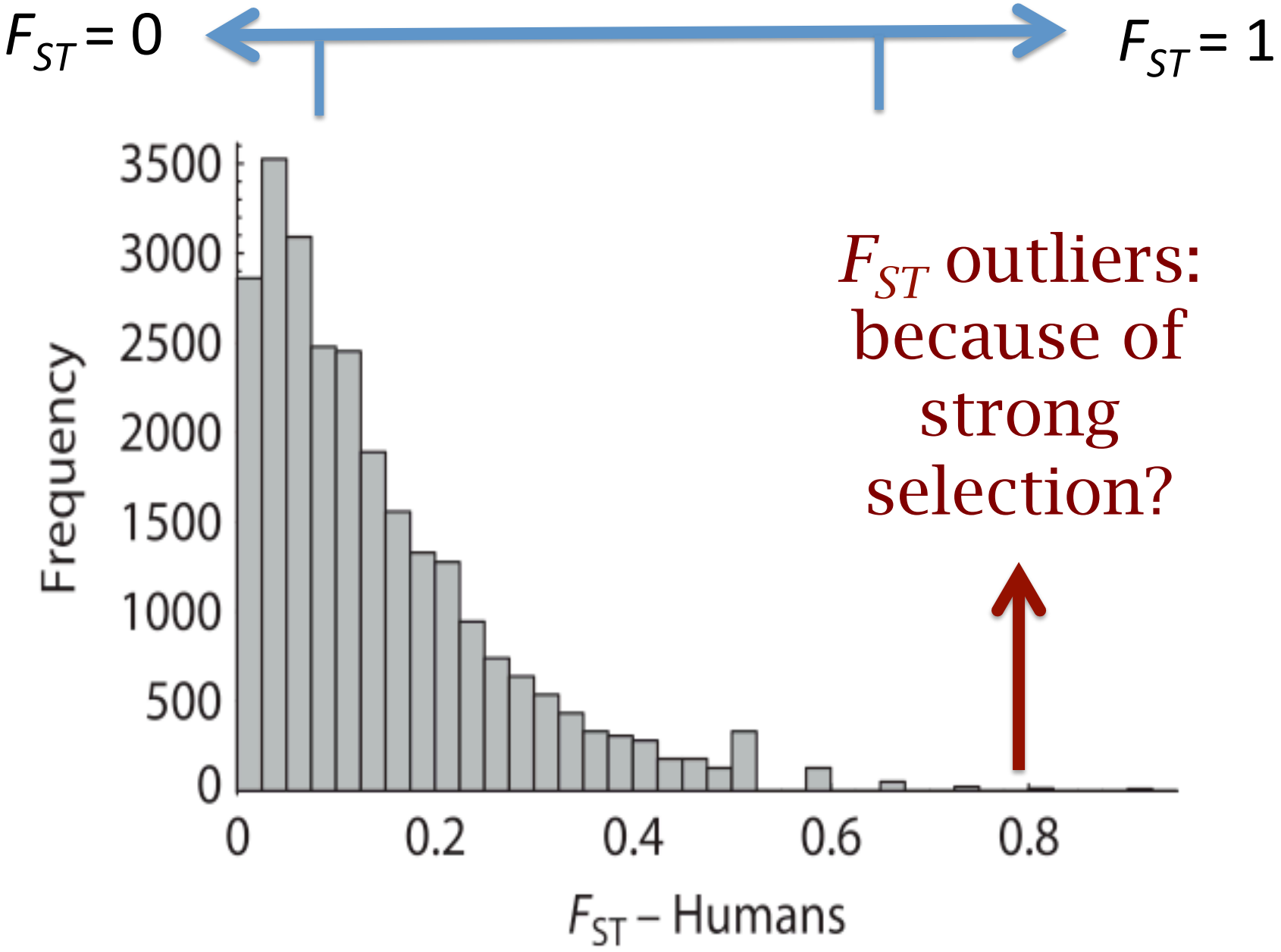
F_{ST} measures how different allele frequencies are between populations

$F_{ST} = 0$ ⟵ ⟶ $F_{ST} = 1$

$$F_{ST} = \frac{var(p)}{\bar{p}(1-\bar{p})}$$



Adapted locus


F_{ST} = 0
F_{ST} = 1
3500
3000
2500
2000
1500
1000
500
0
Frequency
0
0.2
0.4
0.6
0.8
F_{ST} – Humans
F_{ST} outliers: because of strong selection?

Problem:

How to separate the effects of selection from effects of an unknown demographic history?

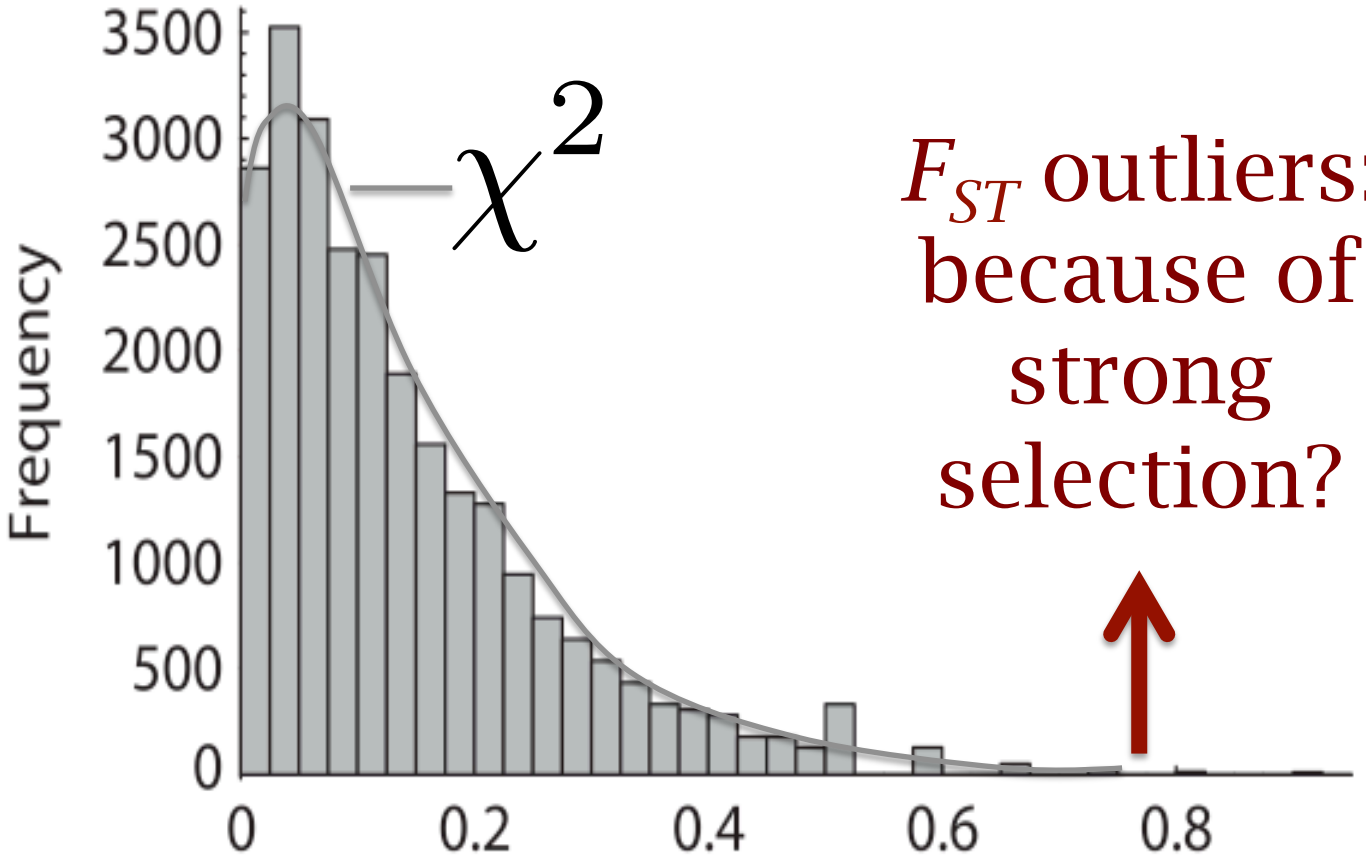
The good news:

The genome is filled with approximately neutral genes

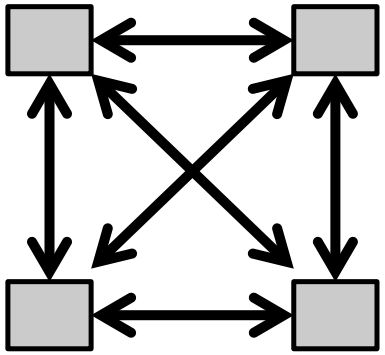
The bad news:

We don't know which genes are neutral, or precisely how they behave

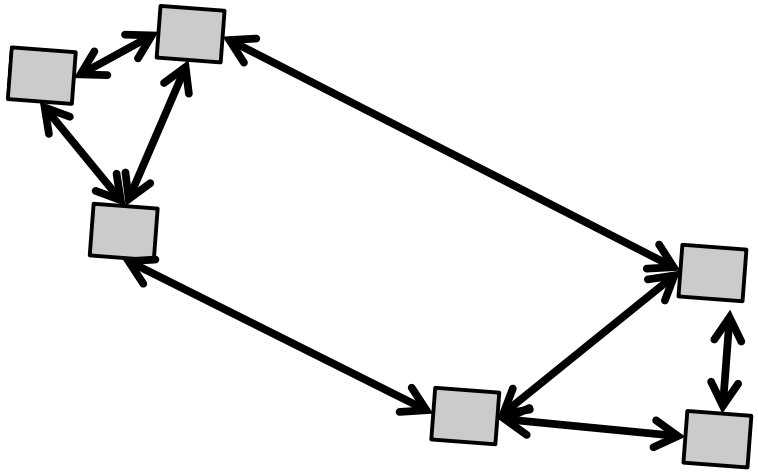
Lewontin-Krakauer Test



Difficulty: LK assumed sampled populations are equally independent of each other



True:
Island Model



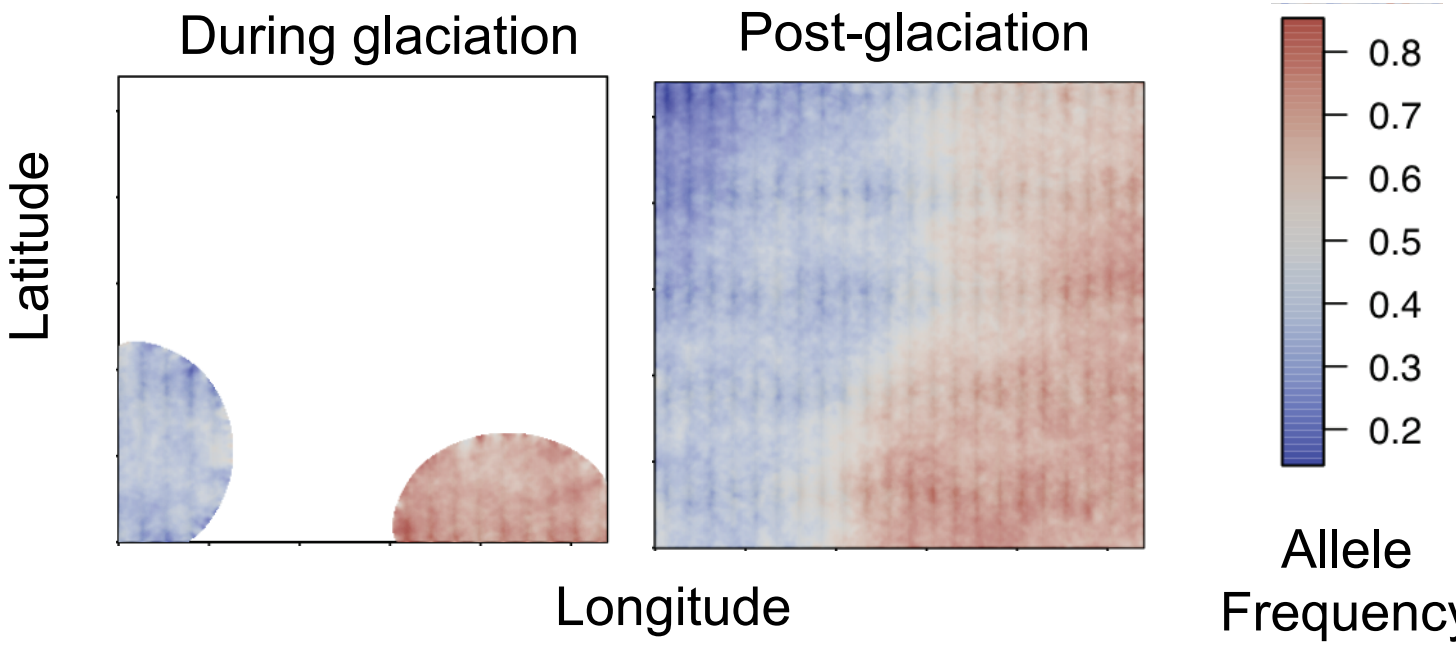
Not True:
Isolation by distance

first noticed by Robertson (1975)

How sensitive are F_{ST} outlier tests to demographic history?

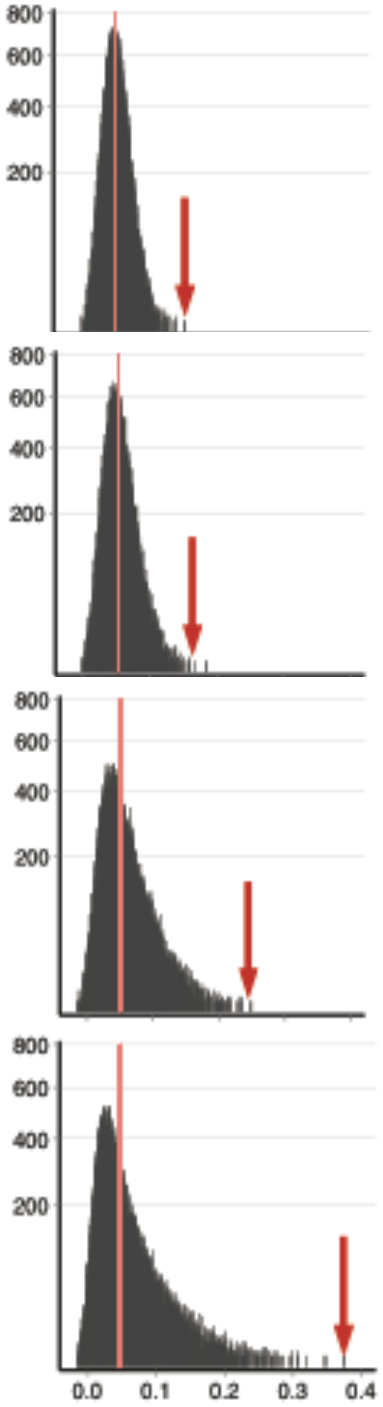
LandSHARC:

Landscape Simulator for Haploid Alleles in Realistic Climates



Lotterhos and Whitlock (2014) *Mol. Ecol*

F_{ST} distributions: neutral loci



IM

Island Model
(at equilibrium)

IBD

Isolation by distance
(at equilibrium)

1R

Expansion from one refuge
(non-equilibrium)

2R

Expansion from two refugia
(non-equilibrium)

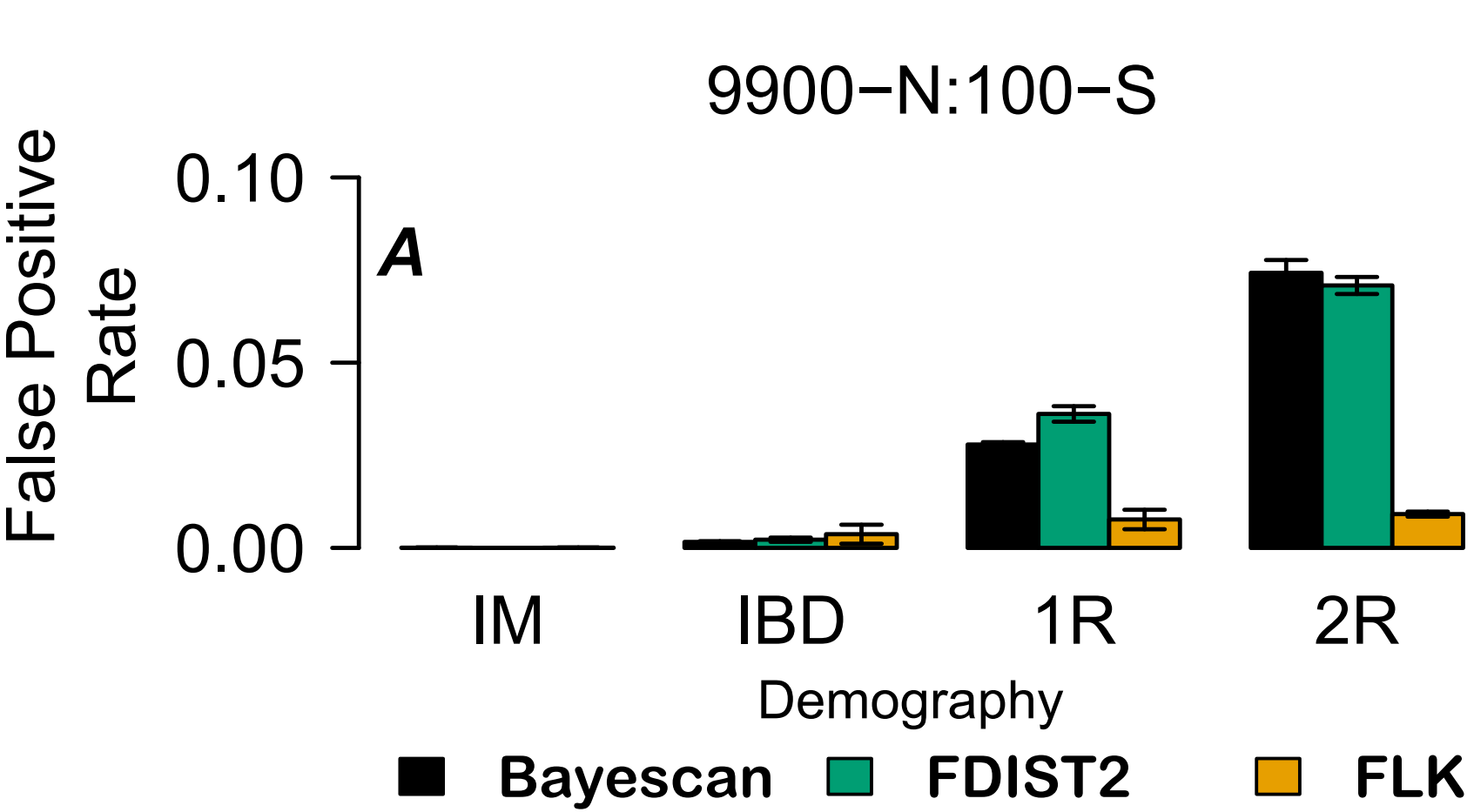
Lotterhos and Whitlock (2014) *Mol. Ecol*

False Positive Rates

(False Positive Neutral)/(Total Neutral)

Want: 1/1000

Lotterhos and Whitlock (2014) *Mol. Ecol*



Lotterhos and Whitlock (2014) *Mol. Ecol*

The bad news:

A false positive rate of 1% is still too high

Error rates depend on neutral parameterization

Lotterhos and Whitlock (2014) *Mol. Ecol*

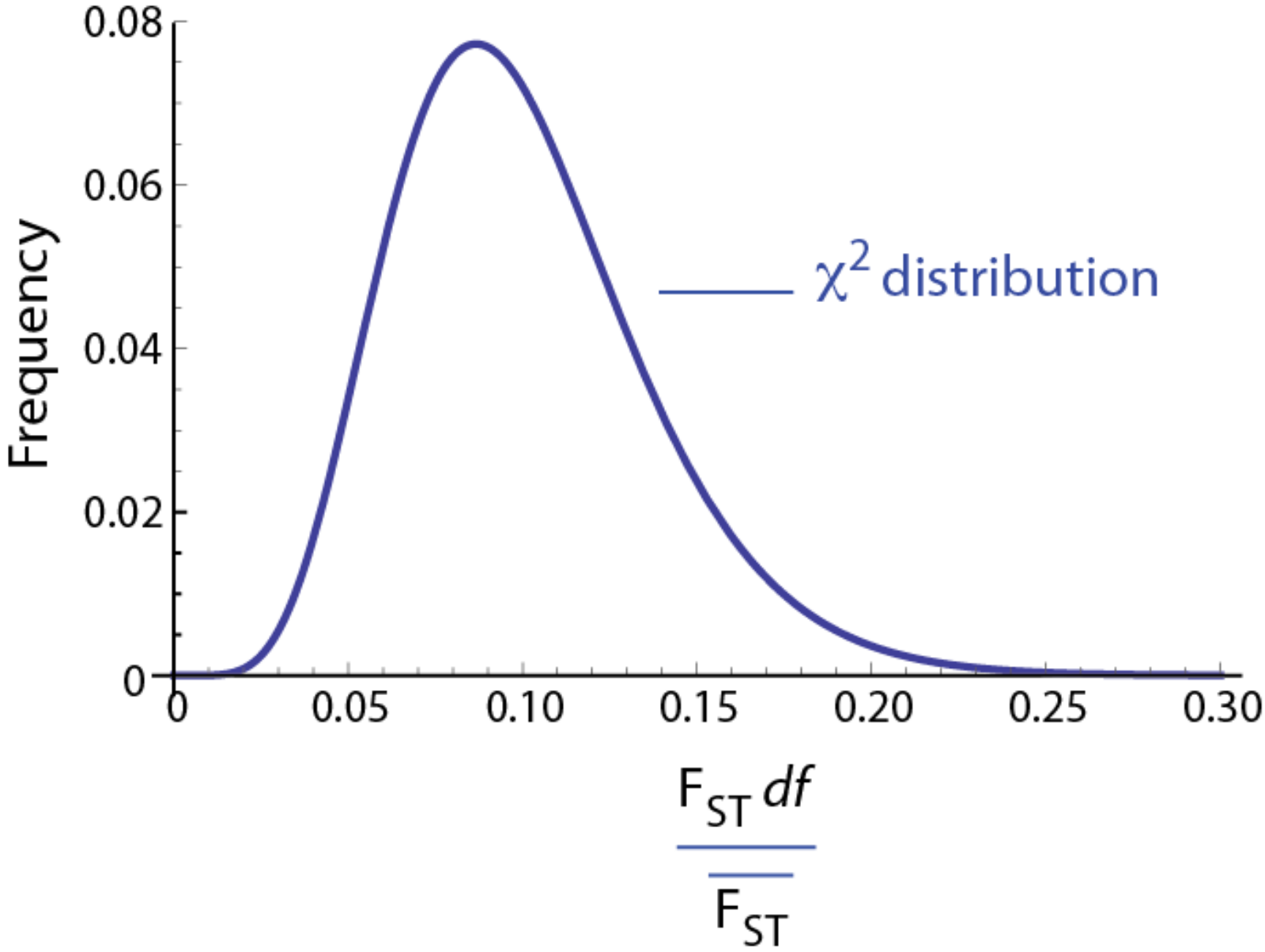
The good news:

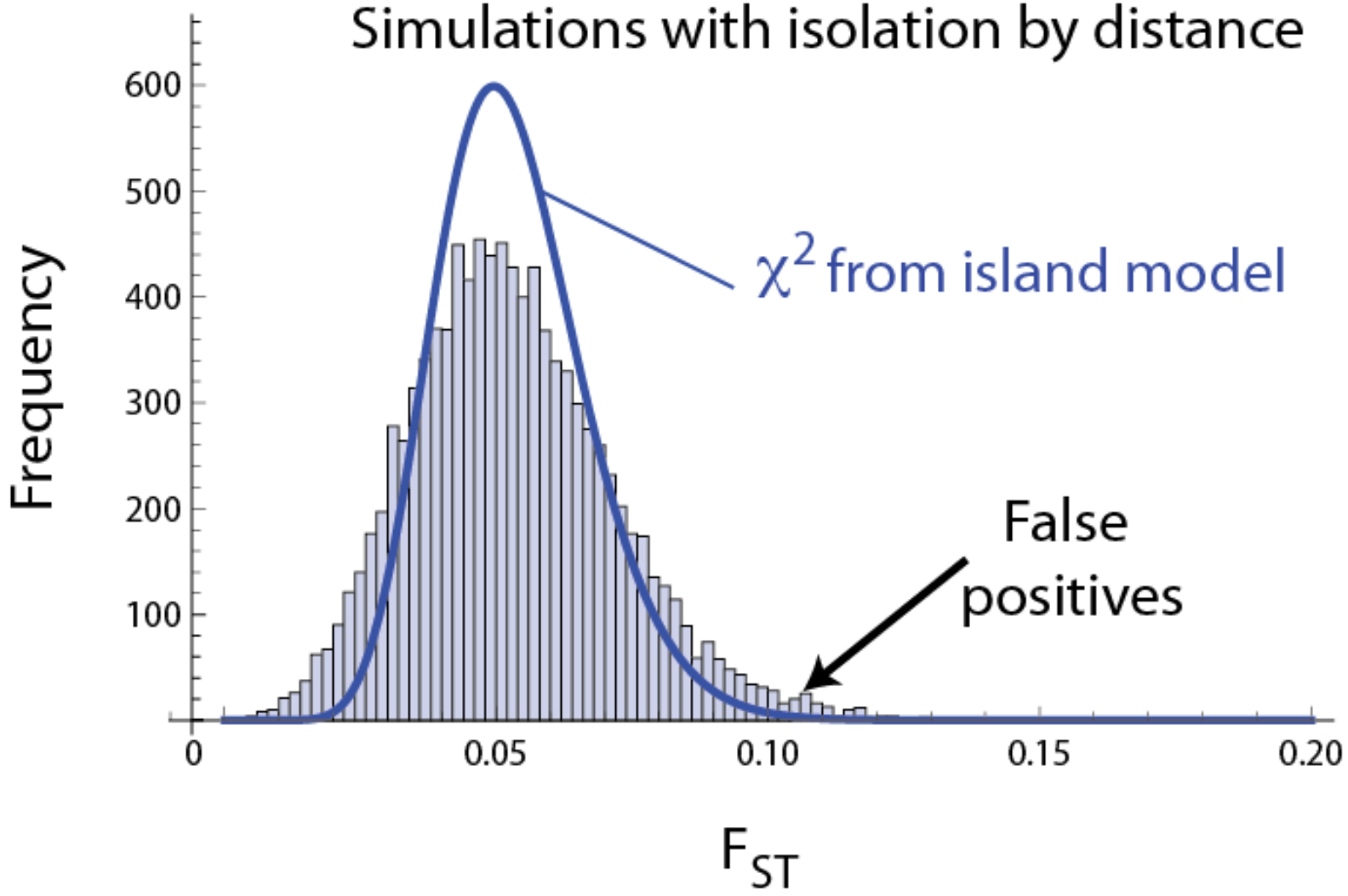
OutFLANK

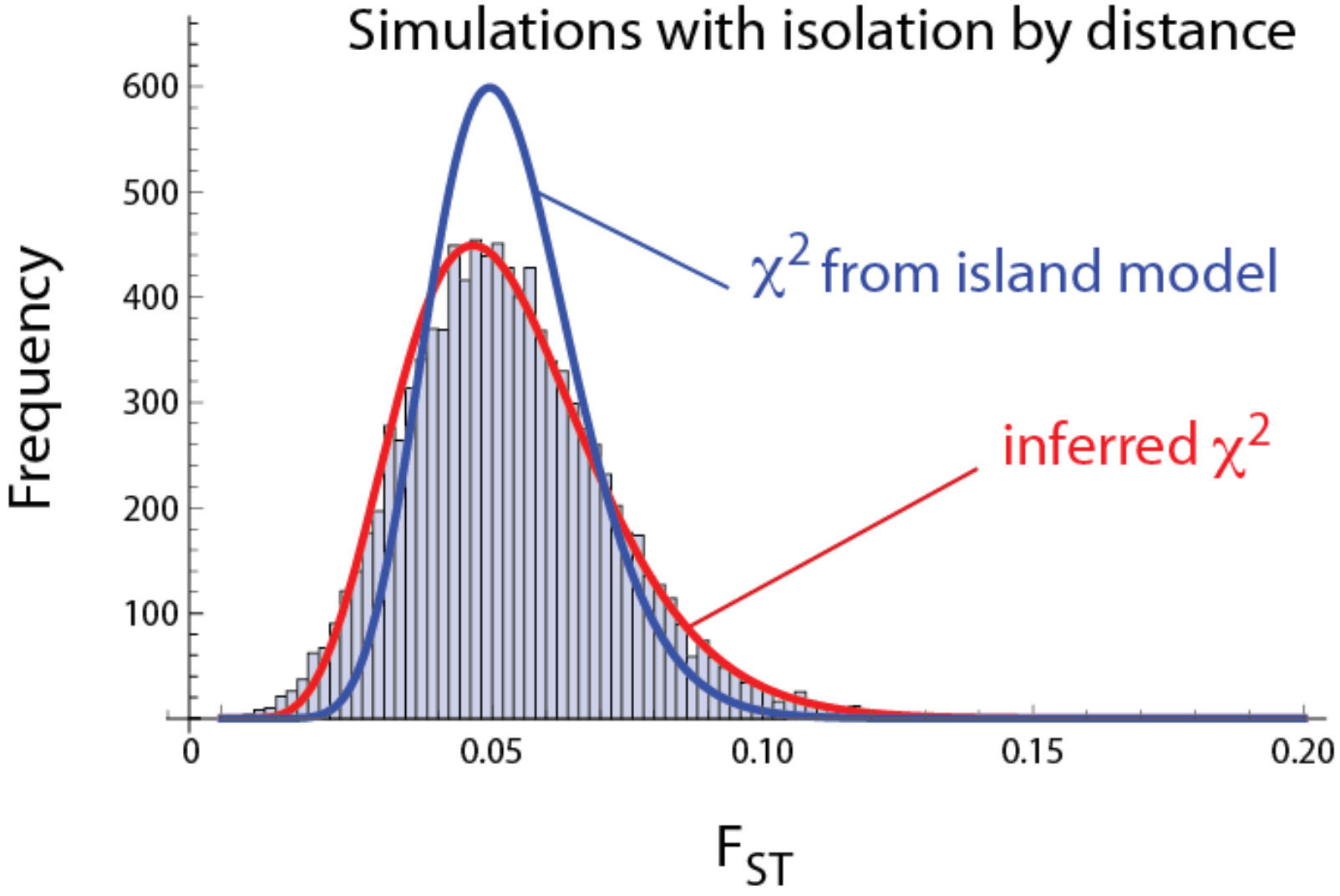
…doesn't rely on having a neutral set

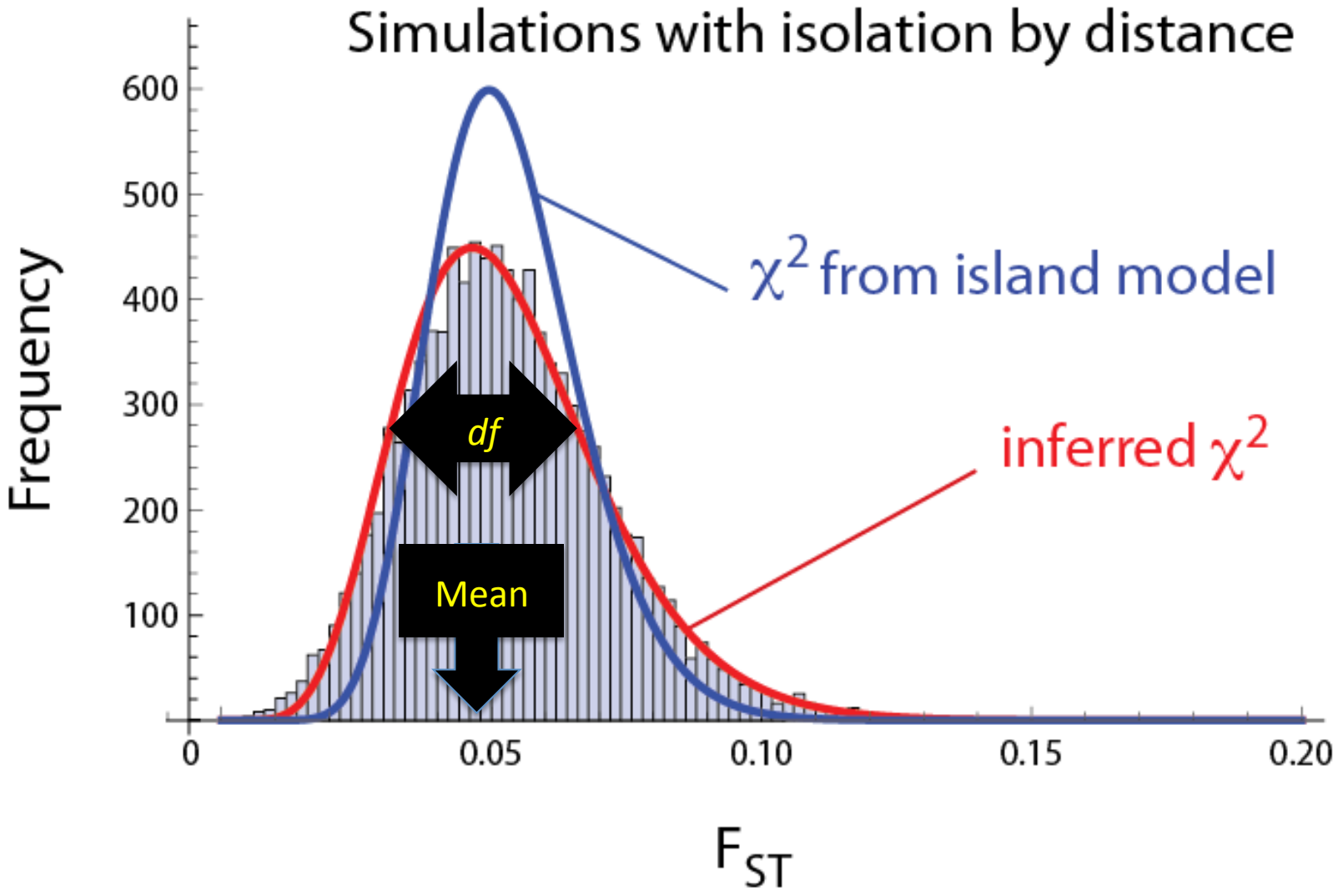
Whitlock and Lotterhos *Am. Nat. Accepted*

Let's revisit the Lewontin-Krakauer test




Simulations with isolation by distance
Frequency
600
500
400
300
200
100
0
0.05
0.10
0.15
0.20
F_{ST}
χ^2 from island model
False positives

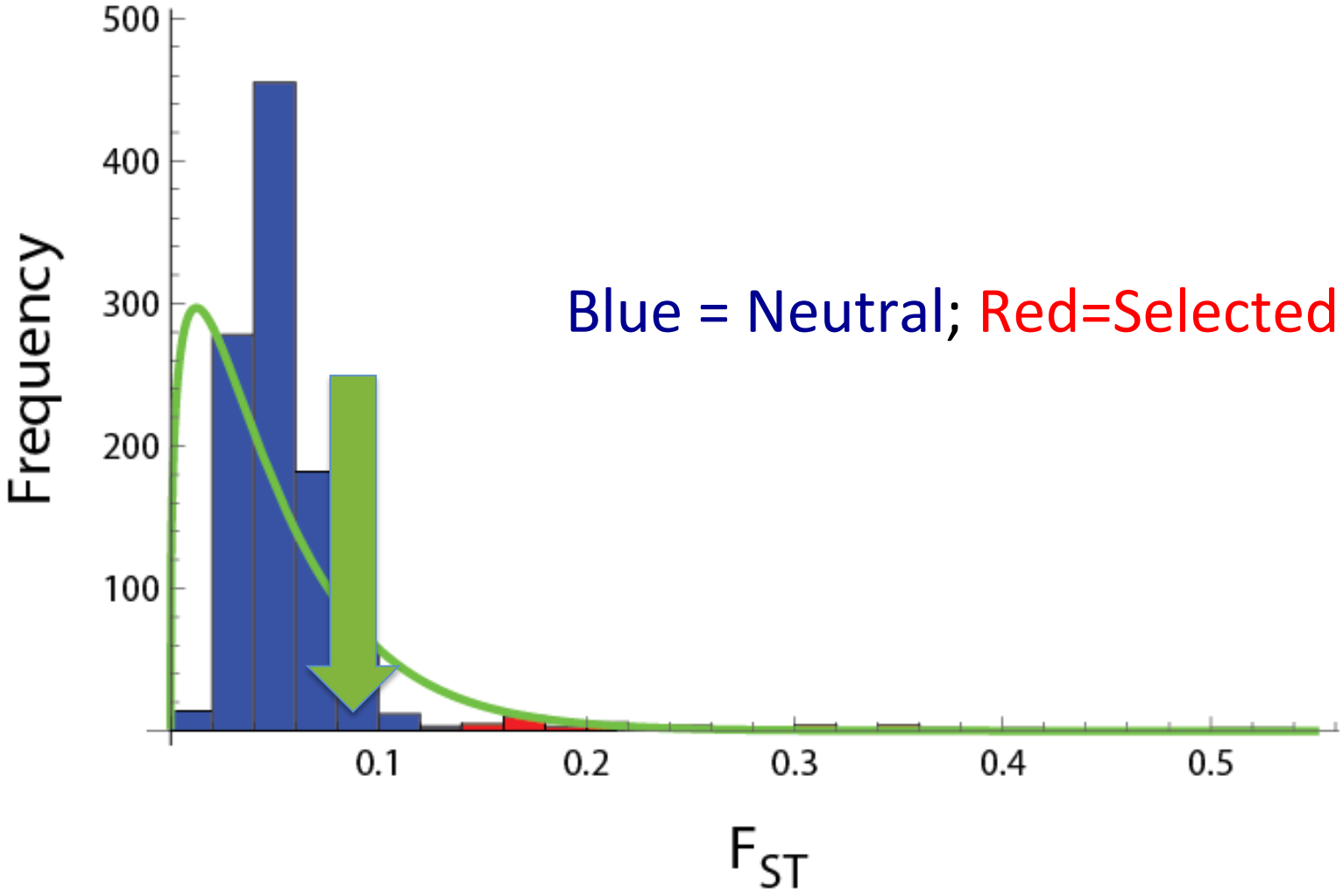

Simulations with isolation by distance
Frequency
600
500
400
300
200
100
0
0.05
0.10
0.15
0.20
F_{ST}
χ^2 from island model
inferred χ^2



Whitlock and Lotterhos *Am. Nat. Accepted*

Finding neutral mean F_{ST} and appropriate *df* for χ^2

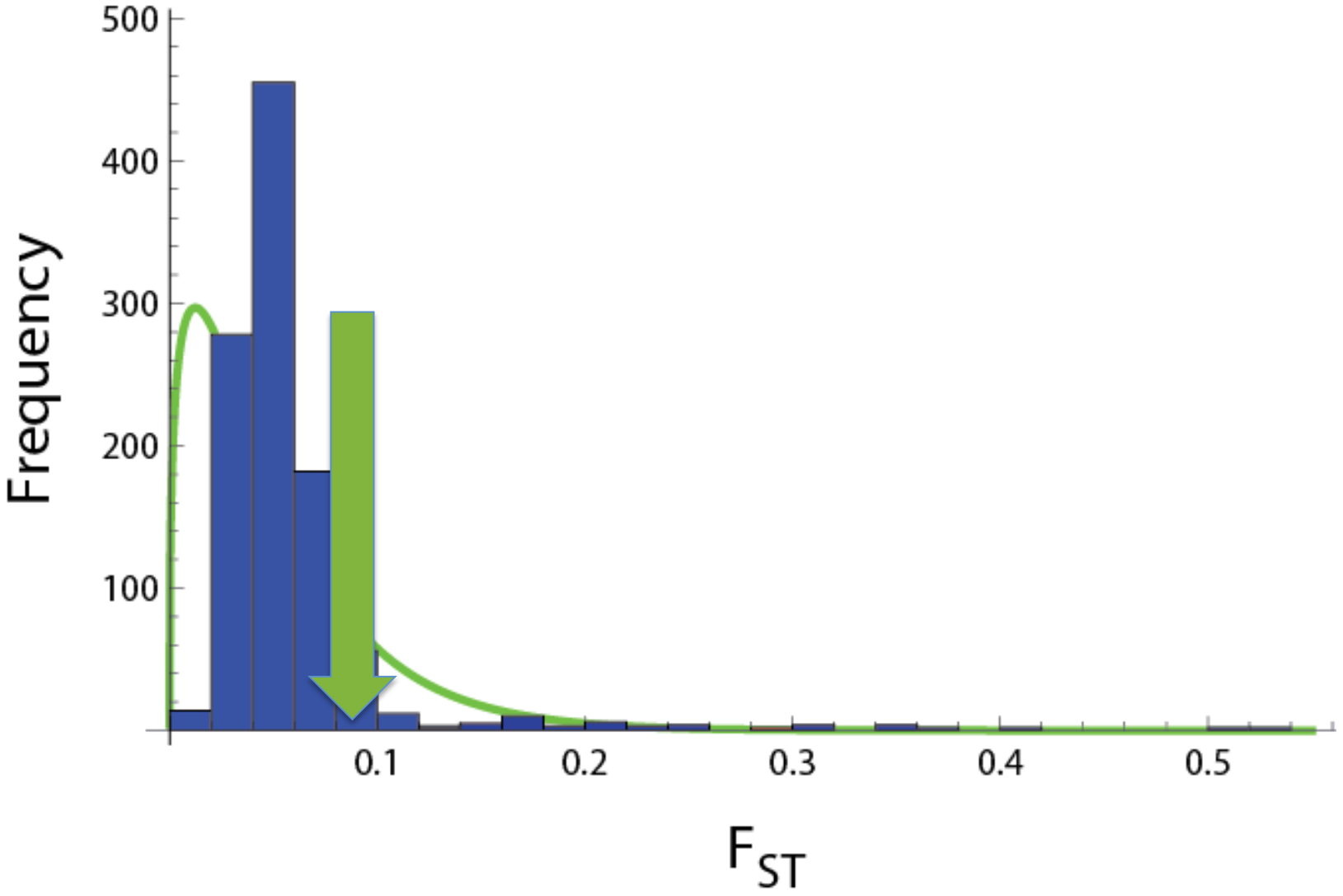
Finding neutral mean F_{ST} and appropriate *df* for χ^2



Blue = Neutral; Red=Selected

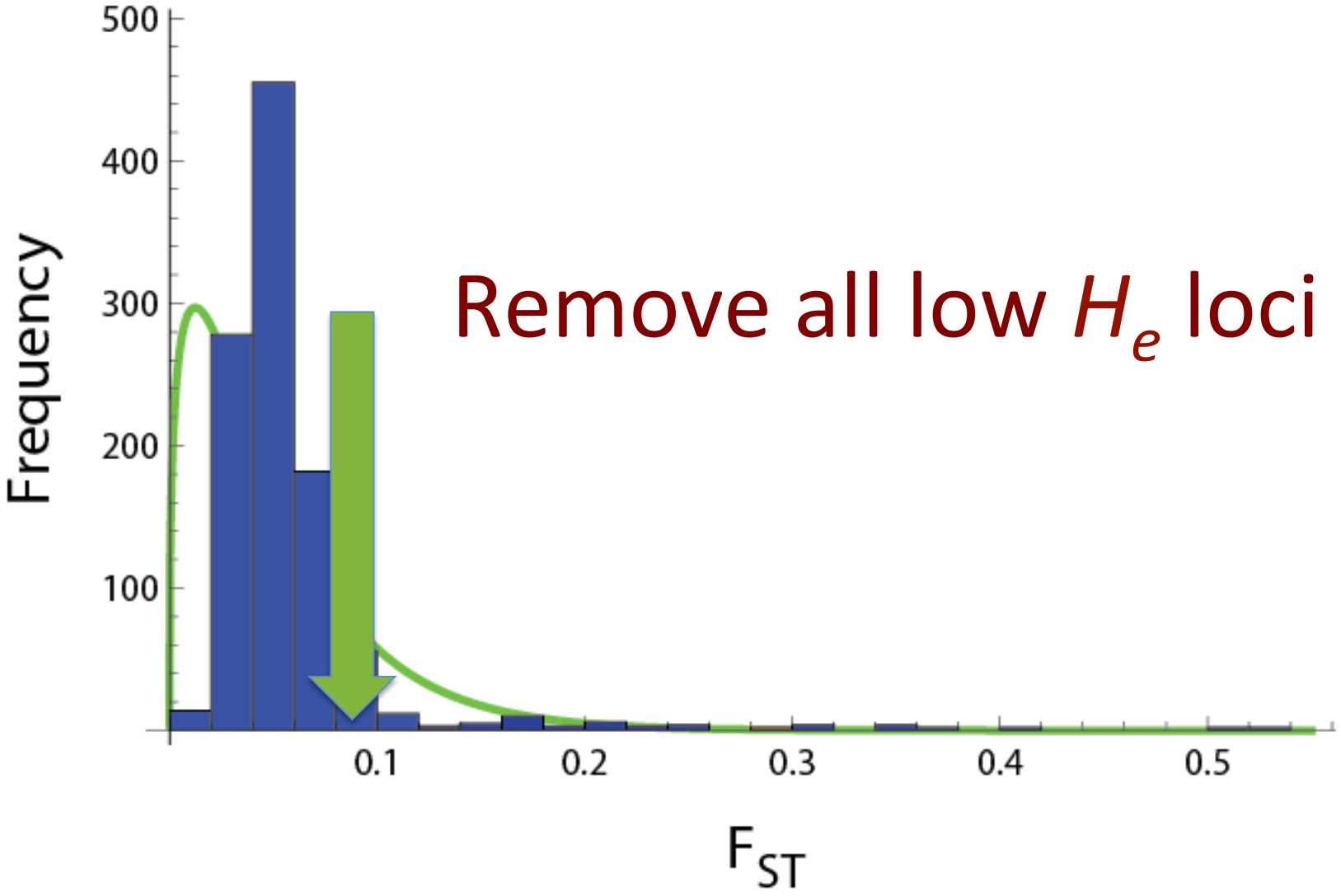
Whitlock and Lotterhos *Am. Nat. Accepted*

Finding neutral mean F_{ST} and appropriate *df* for χ^2



Whitlock and Lotterhos *Am. Nat. Accepted*

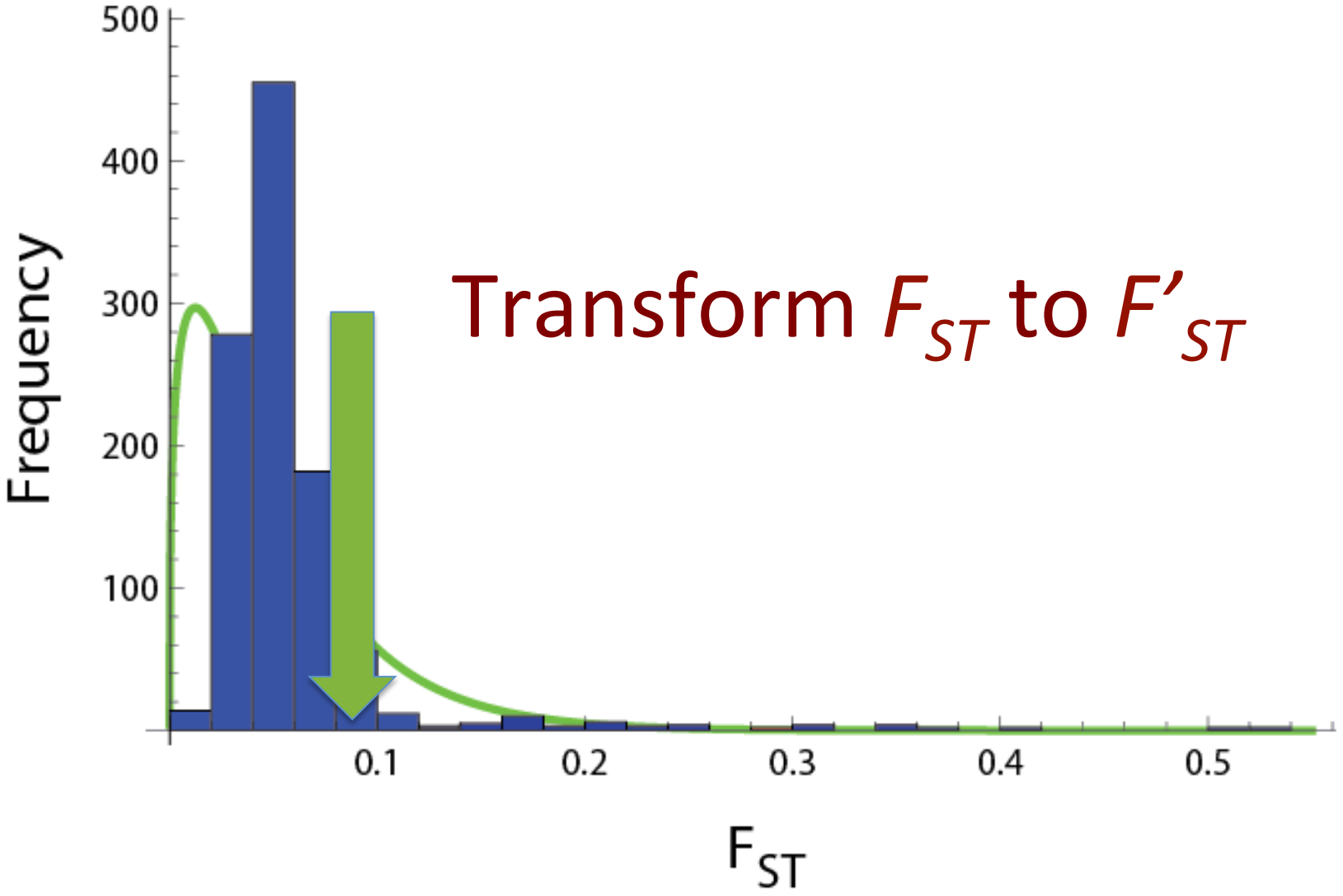
Finding neutral mean F_{ST} and appropriate *df* for χ^2



Remove all low H_e loci

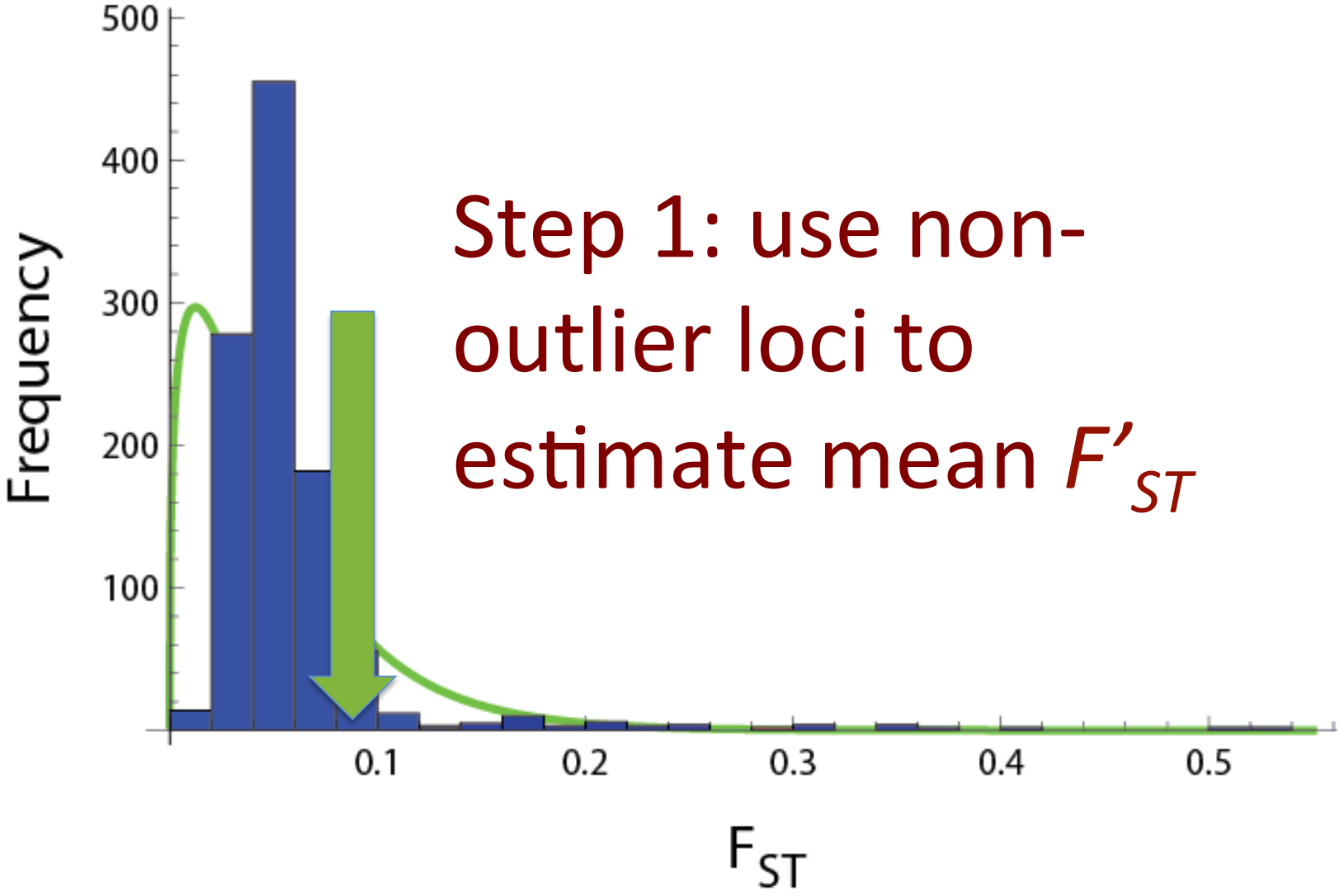
Whitlock and Lotterhos *Am. Nat. Accepted*

Finding neutral mean F_{ST} and appropriate *df* for χ^2



Whitlock and Lotterhos *Am. Nat. Accepted*

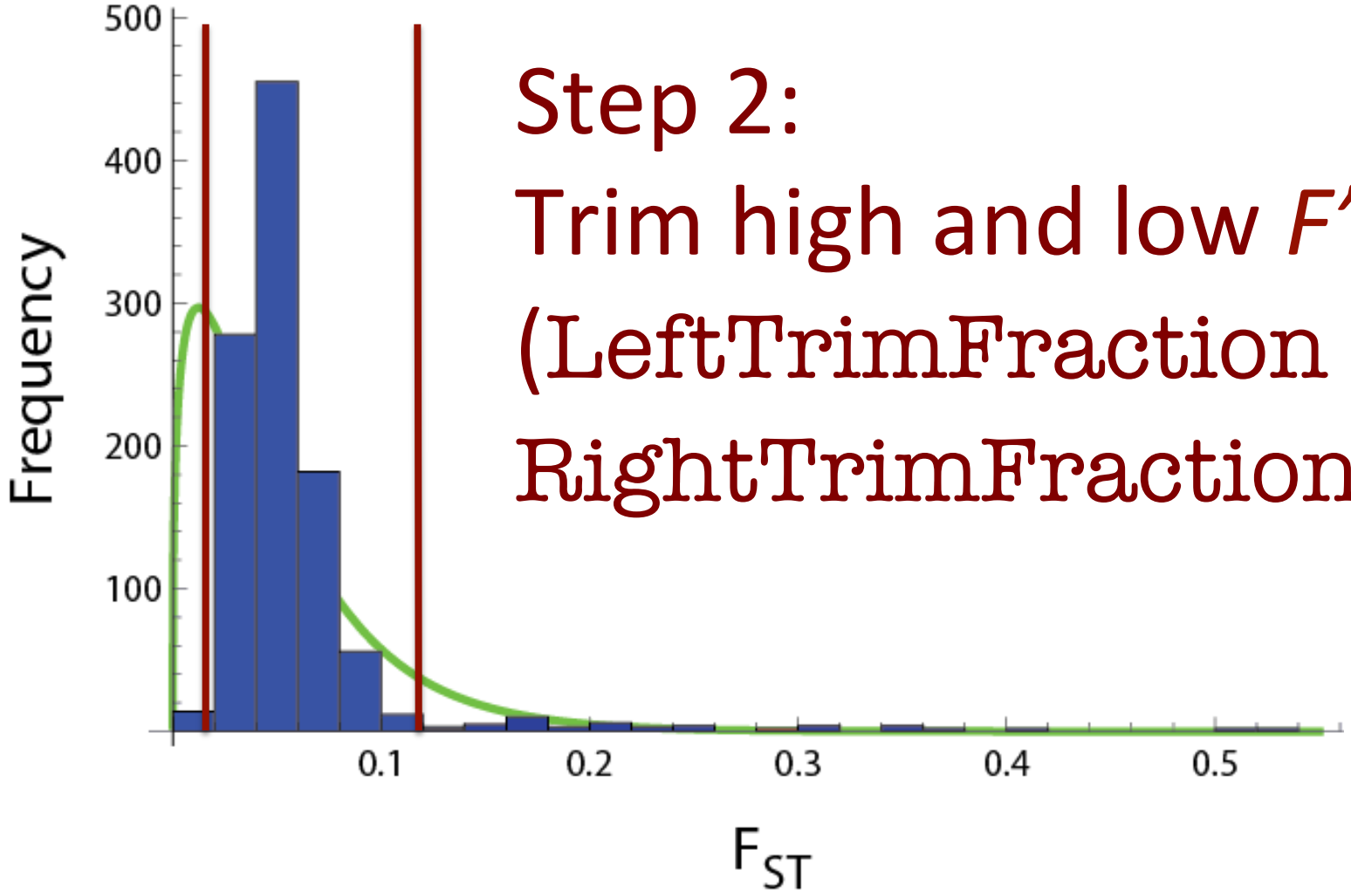
Finding neutral mean F_{ST} and appropriate *df* for χ^2



Step 1: use non-outlier loci to estimate mean F'_{ST}

Whitlock and Lotterhos *Am. Nat. Accepted*

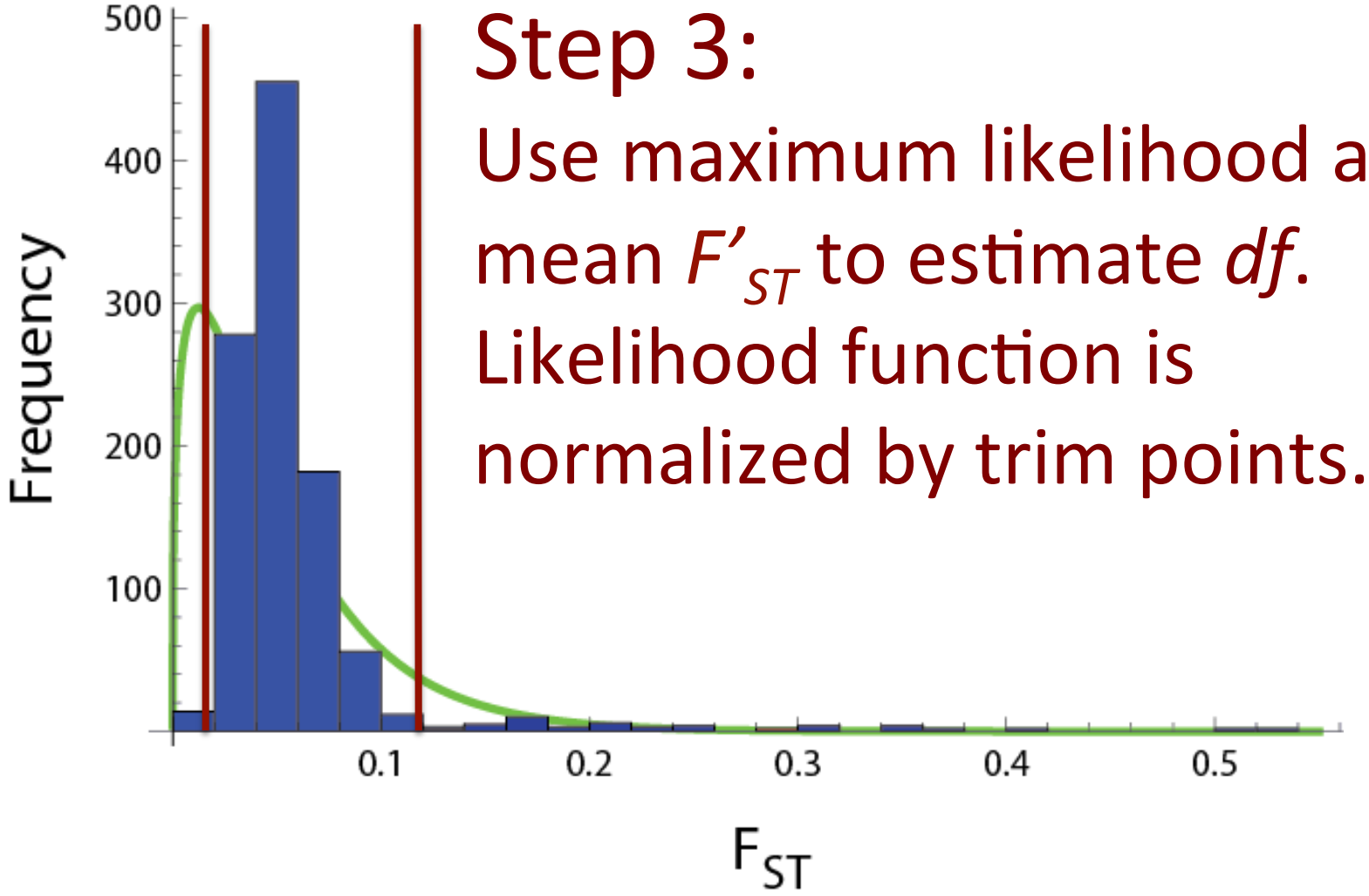
Finding neutral mean F_{ST} and appropriate *df* for χ^2



Step 2:
Trim high and low *F'ST*s
(LeftTrimFraction RightTrimFraction)

Whitlock and Lotterhos *Am. Nat. Accepted*

Finding neutral mean F_{ST} and appropriate *df* for χ^2

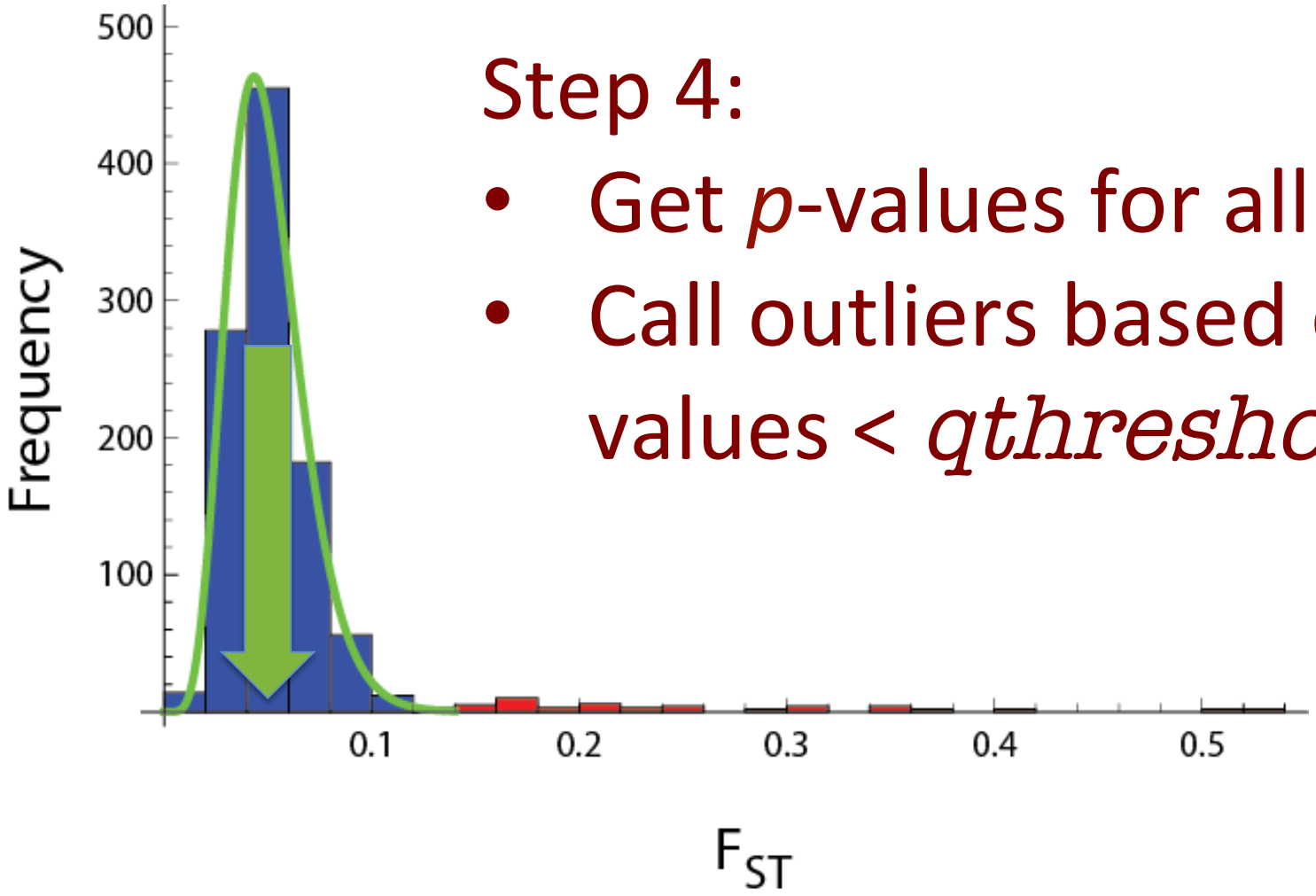


Step 3:

Use maximum likelihood and mean F'_{ST} to estimate *df*. Likelihood function is normalized by trim points.

Whitlock and Lotterhos *Am. Nat. Accepted*

Finding neutral mean F_{ST} and appropriate *df* for χ^2



Step 4:

- Get *p*-values for all loci
- Call outliers based on *q*-values < *qthreshold*

Whitlock and Lotterhos *Am. Nat. Accepted*

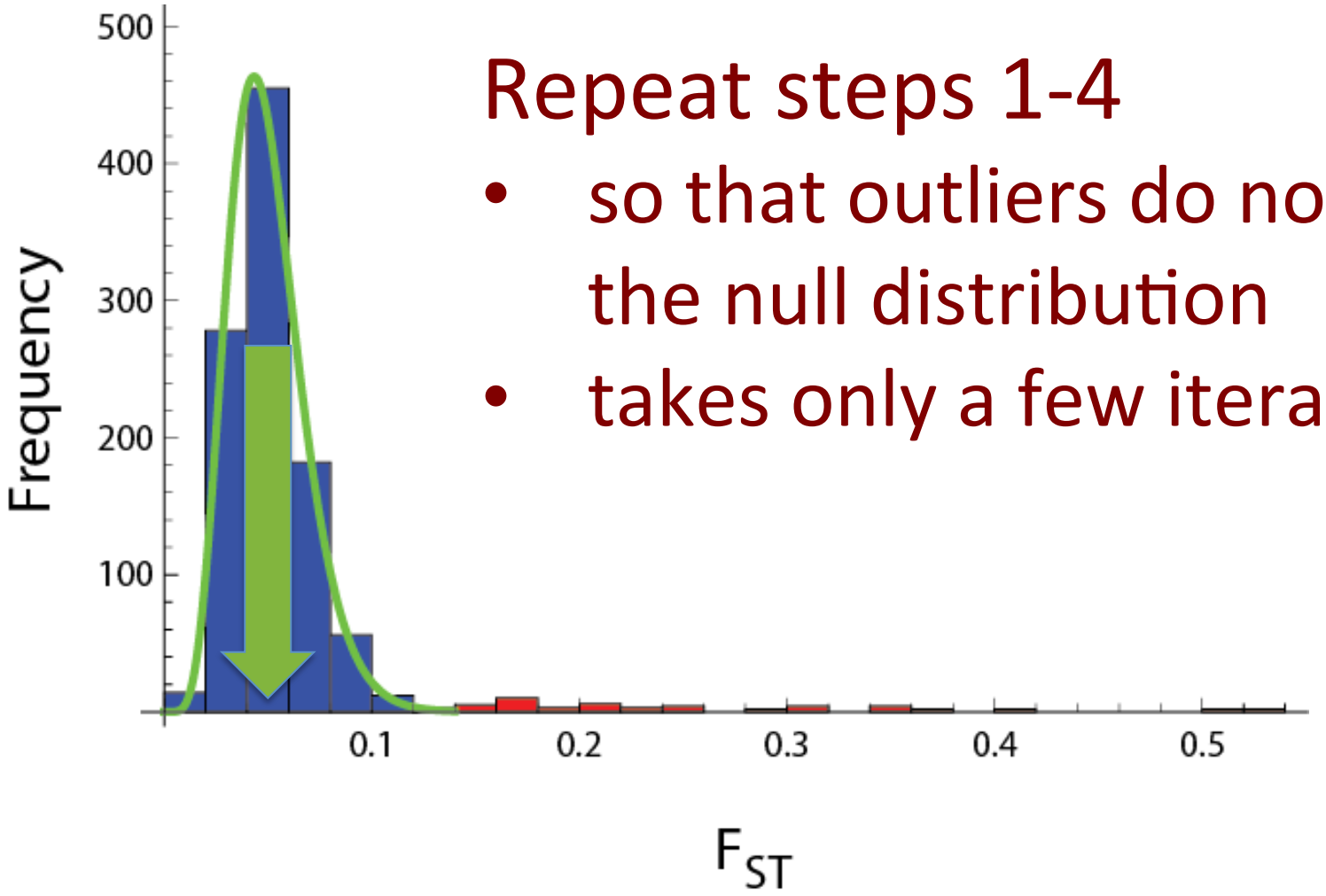
P-value vs. *q*-value

- The *p*-value in OutFLANK is calculated from the fit of the chi-square distribution
- Because there are 1000's of SNPs in the dataset, we need to correct for multiple tests
- OutFLANK controls for False Discovery Rate (FDR) with *q*-values using the method of Storey and Tibsharani (2003)

$$FDR = \frac{Number\ of\ false\ positives}{Total\ number\ of\ positive\ tests}$$

- $q < 0.05$ means that 5% of positive results are expected to be false positives

Finding neutral mean F_{ST} and appropriate *df* for χ^2



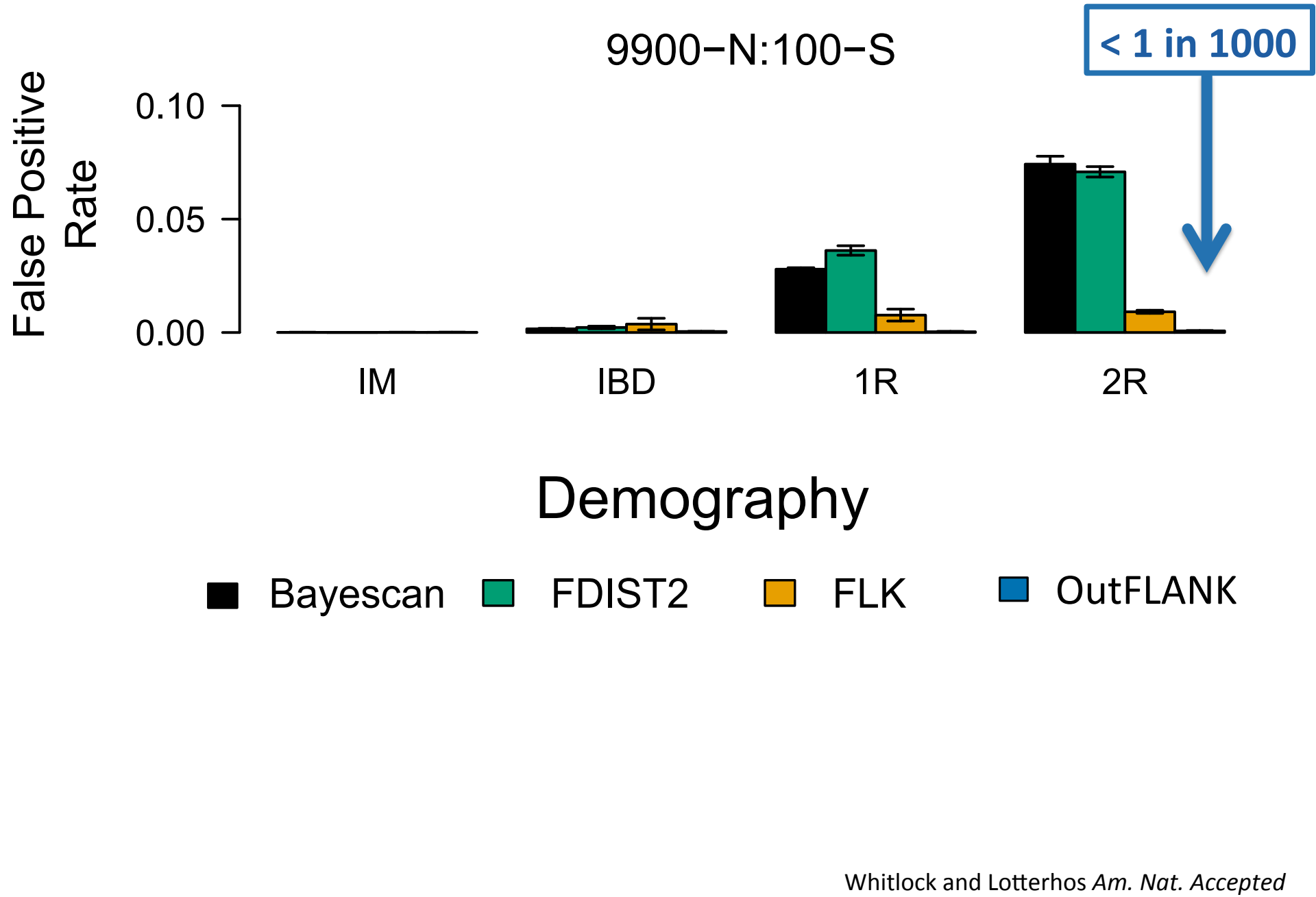
Repeat steps 1-4

- so that outliers do not affect the null distribution
- takes only a few iterations

Whitlock and Lotterhos *Am. Nat. Accepted*

How well does OutFLANK work?

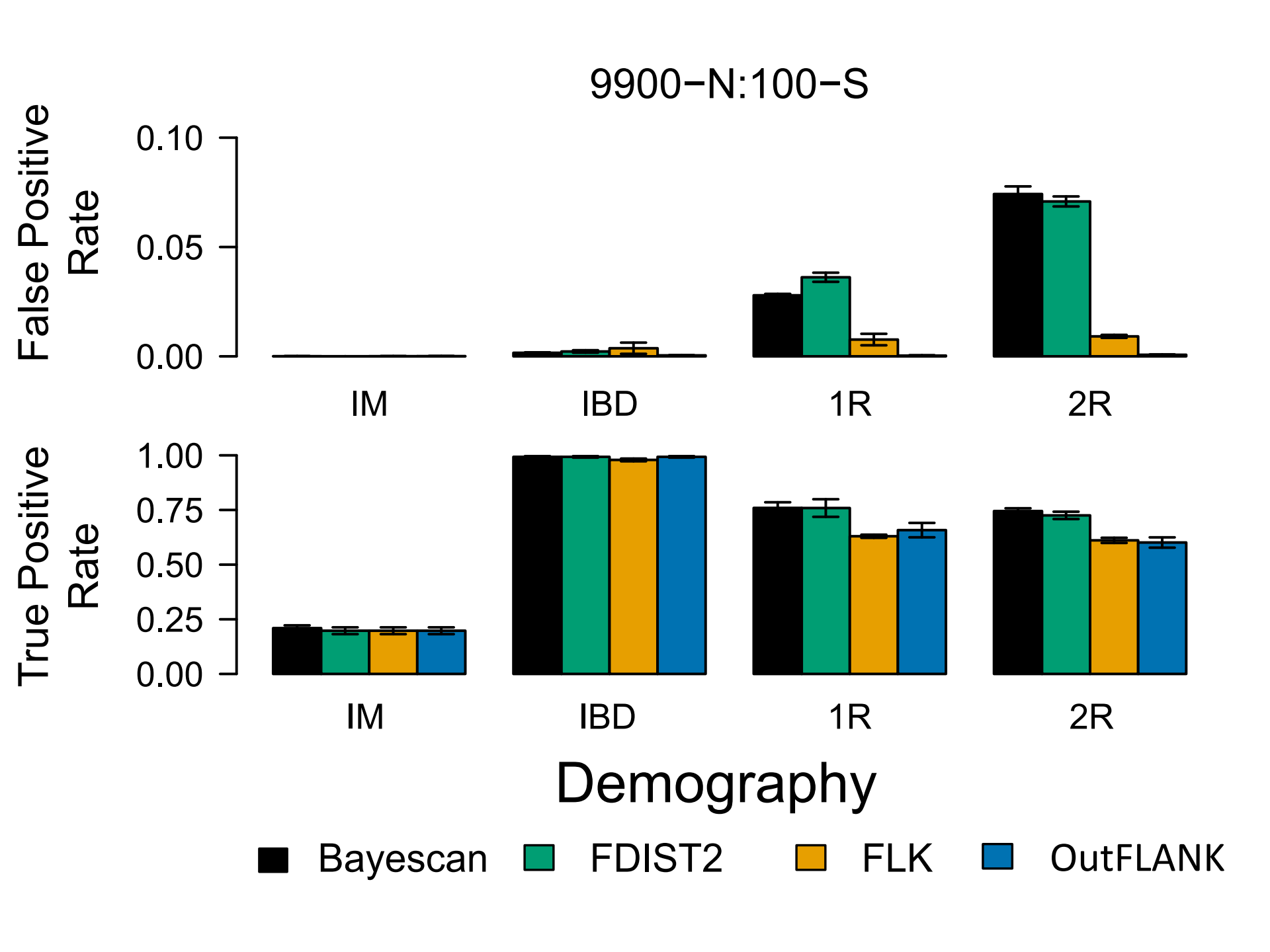
Comparison to current methods



Whitlock and Lotterhos *Am. Nat. Accepted*

Does the new method have power?

(True positive)/(Total number under Selection)


9900−N:100−S
False Positive Rate
0.10
0.05
0.00
IM
IBD
1R
2R
True Positive Rate
1.00
0.75
0.50
0.25
0.00
IM
IBD
1R
2R
Demography
Bayescan
FDIST2
FLK
OutFLANK

Caution

- Power of OutFLANK increases when more populations are sampled and more individuals per population are sampled
- OutFLANK also needs a large number of loci (>1000 SNPs, see discussion in manuscript)

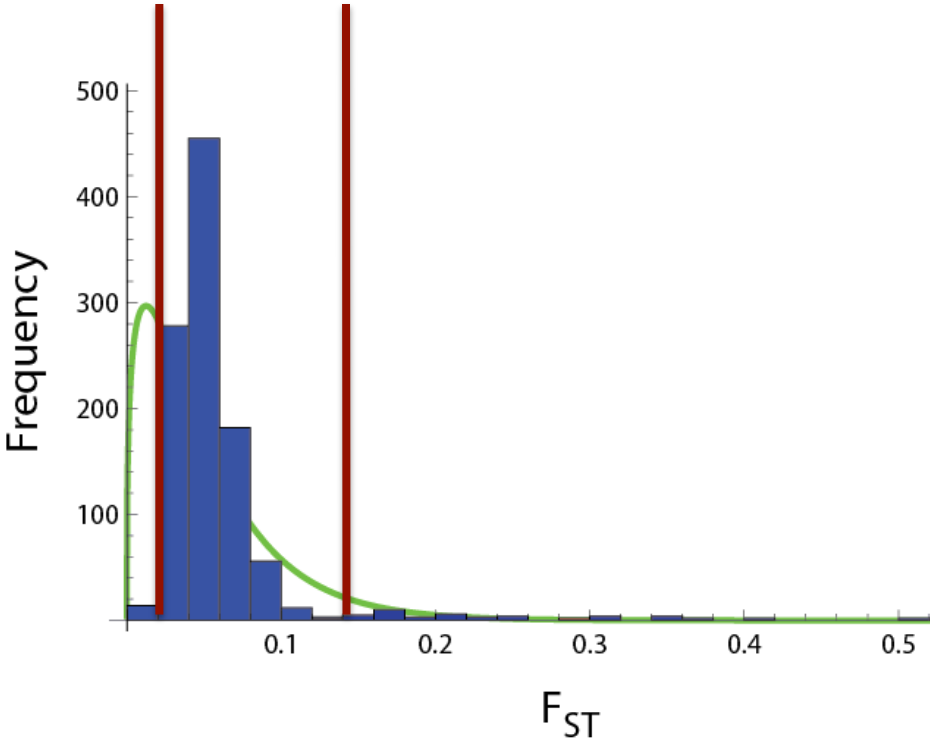
No. individuals per population	5 populations	10 populations	20 populations	40 populations
5	0	.09	.52	.84
10	.10	.56	.82	.94
20	.37	.75	.90	.95
40	.55	.81	.94	.97

Steps to running OutFLANK

1. Prepare a dataframe for the OutFLANK() function
   - From a formatted SNP dataset using MakeDiploidFSTMat()
   - (From your own data using functions provided with the package (see Section “For Advanced Users”))
2. Check for SNPs of low sample size or that have F_{ST} values differentially affected by sample size correction
3. Run OutFLANK
4. Plot results

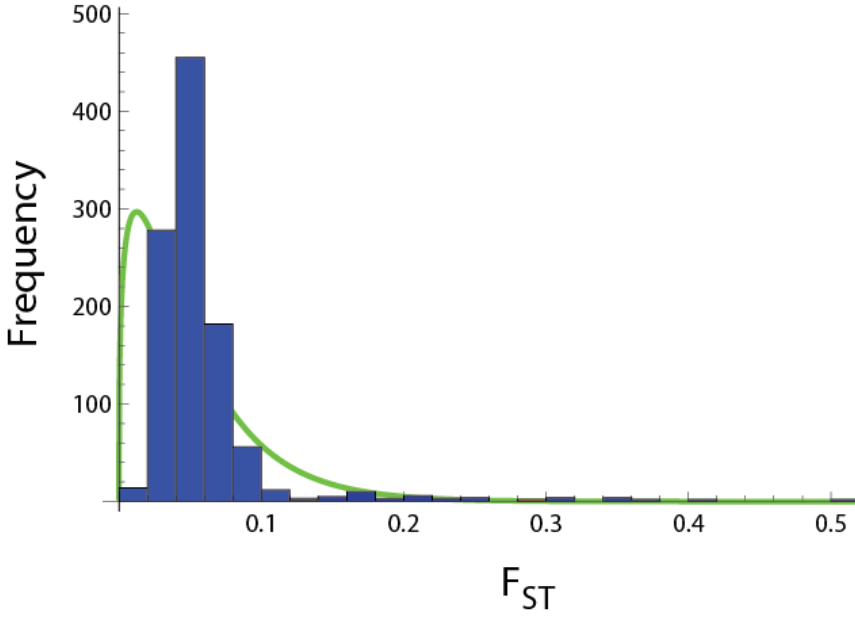
The OutFLANK() function

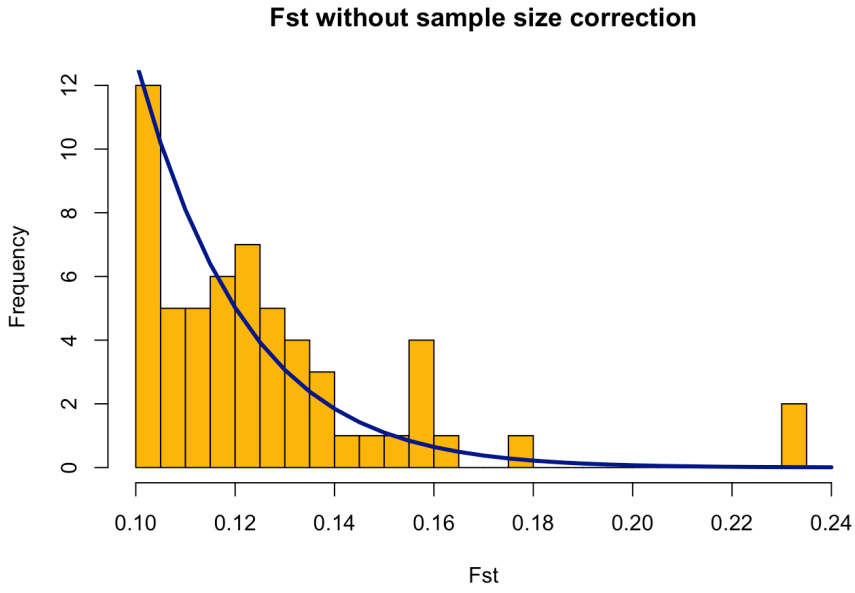
- **FstDataFrame**
- **LeftTrimFraction**=0.05
- **RightTrimFraction**=0.05
- **Hmin**=0.1 (loci with low H_e do not follow chi-square assumption)
- **NumberOfSamples** (Number of populations)
- **qthreshold**=0.05 (desired false discovery rate)



Hints

- **qthreshold** is used to call outliers. The trim points must be located within any potential outliers. If **qthreshold** is set too high, OutFLANK will return an error.
- The algorithm may fail if **RightTrimFraction** is set too high
- Use the plotting function to check the fit of the inferred distribution to the right tail
- OutFLANK will not fit the left tail of the F_{ST} distribution well





Other questions

- Can I use OutFLANK on pool-seq data? Not Recommended
- Does pool-seq data fit a chi-square distribution (the big assumption of the method)?
  - We understand how sampling variance from finite populations and finite individuals per population affect the variance of the Fst distribution, and we know it still fits a chi square for the cases we've simulated.
  - However, we are uncertain how the random sampling of chromosomes for pool-seq data would affect the Fst distribution.
  - Other issues arise because we are not using a sample size correction in outflank (so we assume loci have roughly equal sample sizes), and if individuals contribute unequally to pools then this could be a huge violation of this assumption.
- So, in short, if you wanted to do it I would suggest the following steps:
  - 1) Use the haploid Fst estimator that comes with the distribution instead of the default in preparing the matrix for the outflank function
  - 2) Show that your Fst distribution fits the chi-square output by outflank
  - 3) Do some simulations that show that outflank performs well for your study system and sampling design

Questions

- What if some populations have fewer individuals than others? THAT'S OK – only a problem if some *loci* have lower sample sizes

Notes based on student questions

- Confusion about q_threshold and the qvalue used for decision making
- a need to illustrated how increasing q_threshold can make small changes in the tail and the resulting q-values of the outliers
- a need to plot p-value histograms and q-q plots