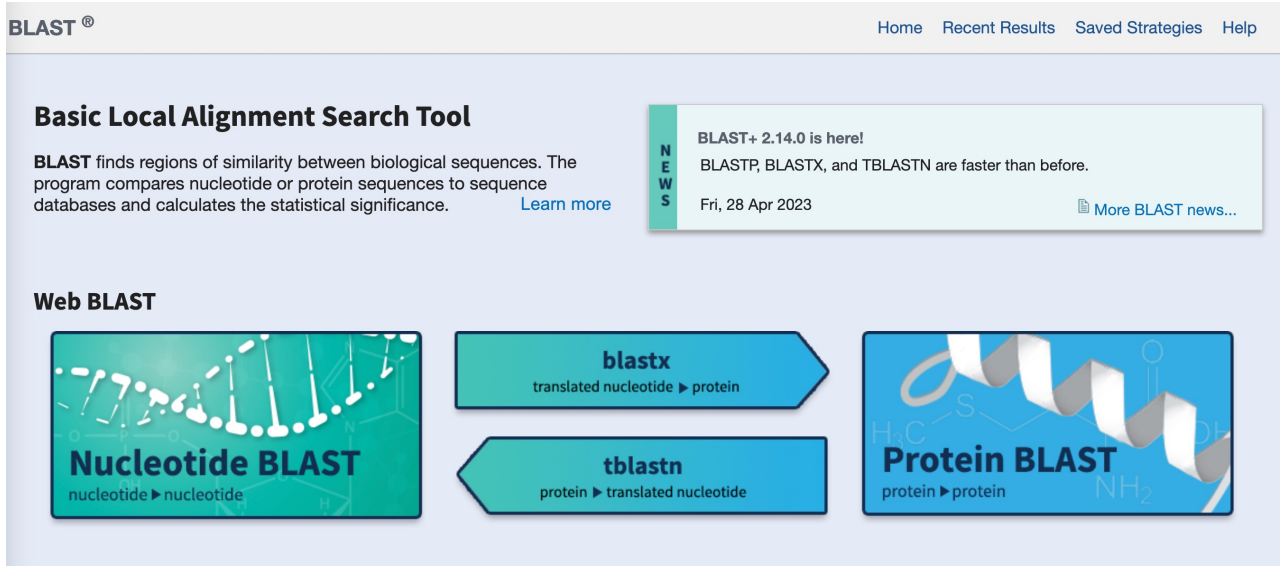
Short lectures

Short Lecture:
Alignment, reference genomes, denovo assemblies, stacks

Blast



- Finds conserved sequence (seed), then extends.
  - Mostly looking for homologous sequences, not exact matches
- Good for low #'s
- Slow!
  - Need to align hundreds of millions of reads

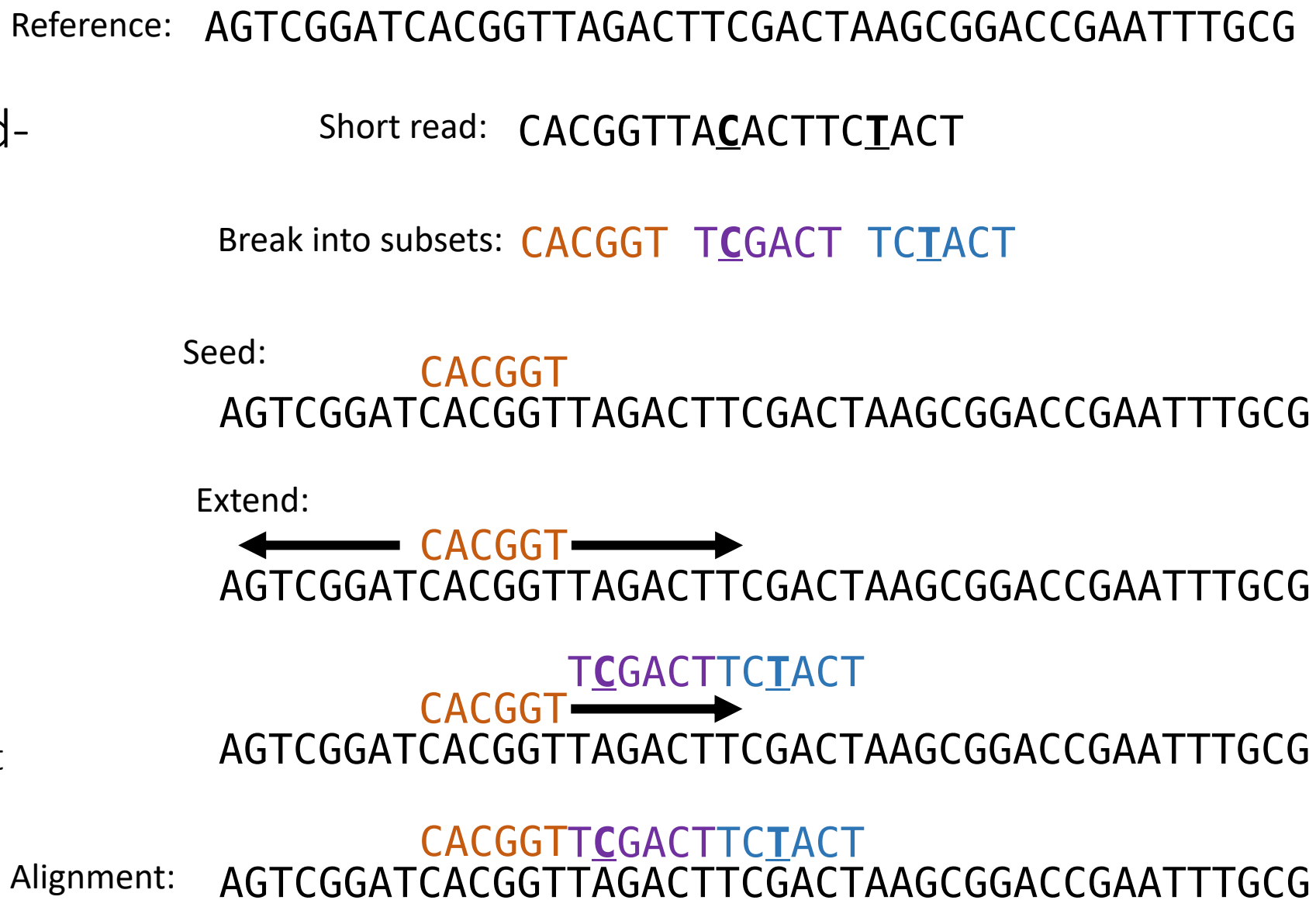
Short read aligners

- We want to find near exact matches within a genome for millions of reads
  - Usually working with one species
- Needs to be
  - Fast
  - Accurate

- Reference genomes are big
- We have MANY reads to align
- We don’t want exact matches→ why?

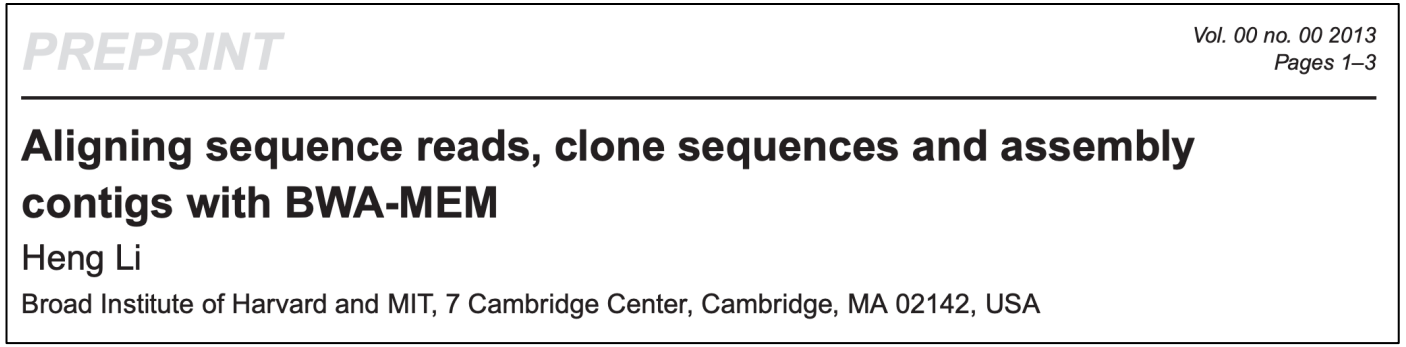
Short read aligners

- Basic approach: seed-and-extend
  - Take a subset of your sequence (k; usually ~19)
  - Find the match in the reference genome (seeding)
  - Try to extend these matches (extending)
  - How might we alter k?
    - shorter = more thorough but slower
    - longer = less thorough but faster.



Short read aligners

- BWA
- Bowtie2
- STAR
- Pseudo aligners (for RNAseq)
  - Kallisto
  - Salmon
- And many others…
  - Depends on what you need
    - Think read length, speed, accuracy, RNA vs DNA, etc.


PREPRINT

Vol. 00 no. 00 2013
Pages 1–3

Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

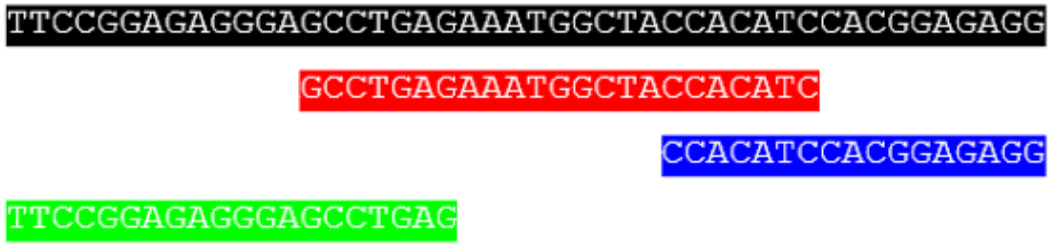
Heng Li

Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

- BWA: Burrow-Wheeler Aligner for short-read alignment
  - Burrows-Wheeler Alignment → a compression method
  - Maximal Exact Match→ longest match that can't be extended further
  - Most short read aligners use similar approaches

What if I don’t have a genome?

- Option 1: Denovo assembly
  - Best option
  - Challenging:
    - Genomes are big
    - Repetitive
    - Inversions
  - Long reads address these issues
    - PacBio, Nanopore
  - Can also link scaffolds with other methods:
    - Hi-C, optical mapping, HiRise, etc.



Repeat regions:

AACCAACCAACCAACCAACCAACCAACC

What if I don’t have a genome?

- Option 2: Reduced representation
  - Most commonly RAD-seq/GBS or similar
  - Still must generate a pseudo reference genome
    - Stacks, ipyrad, ddocent, others

***Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences**

Julian M. Catchen,* Angel Amores,† Paul Hohenlohe,* William Cresko,* and John H. Postlethwait†,1
*Center for Ecology and Evolutionary Biology and †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403

Sequence analysis

ipyrad: Interactive assembly and analysis of RADseq datasets

Deren A.R. Eaton[1,*] and Isaac Overcast[2]

[1]Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY 10027, USA and [2]Department of Biology, Graduate School, University Center of the City University of New York, New York, NY 10016, USA

***dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms**

Jonathan B. Puritz, Christopher M. Hollenbeck and John R. Gold

Marine Genomics Laboratory, Harte Research Institute, Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

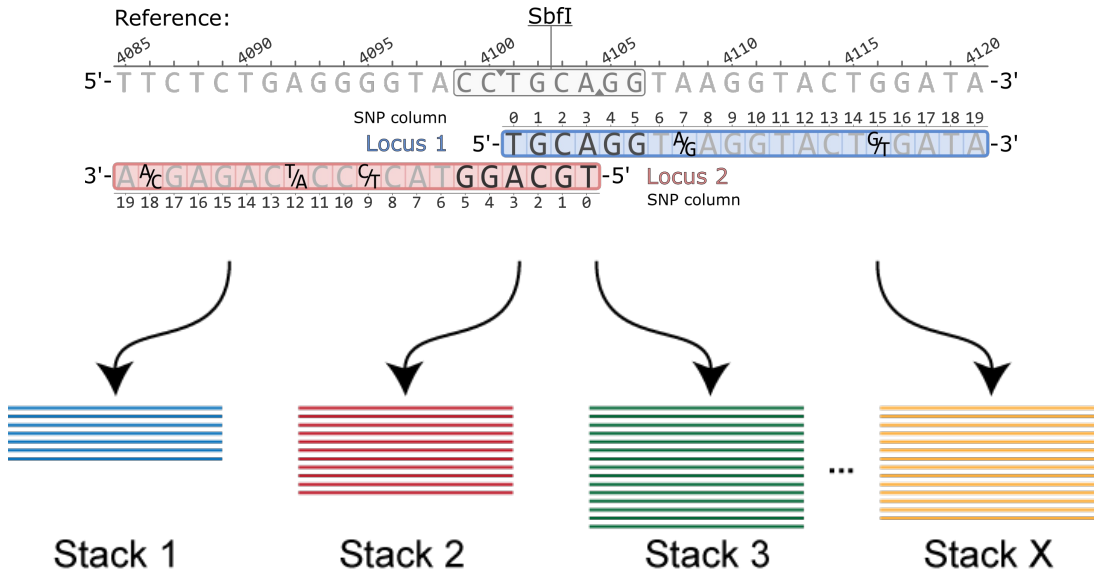
Stacks

pipeline for de novo rad assembly

Stacks Manual

Julian Catchen[1], Nicolas Rochette[1], Angel Rivera-Colón[1], William A. Cresko[2], Paul A. Hohenlohe[3], Angel Amores[4], Susan Bassham[2], John Postlethwait[4]

- Most popular method, good for within species comparisons
- Clustering by similarity:
  - Need to decide:
    - Minimum depth for a stack
    - # of mismatches to allow when making a stack
      - Allow for genetic variation
      - But not merge different genomic regions
- Harder than it seems…



https://catchenlab.life.illinois.edu/stacks/manual/

Short Lecture:
Mapping results, PCR duplicates
Properly paired
multi-mapping

Alignment types

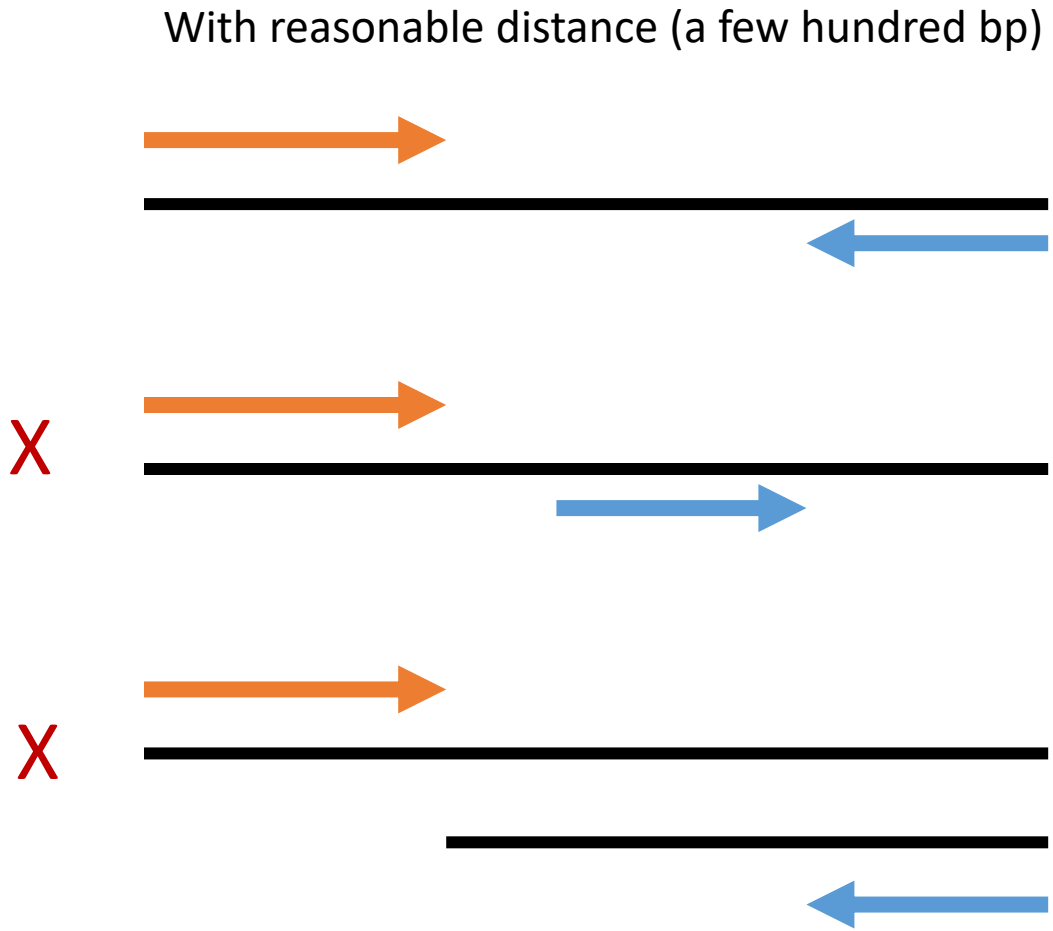
- Primary
- Secondary
- Supplementary
- Duplicates
- Properly paired

Alignment types

- **Primary** → When a read maps to multiple locations, the best alignment
- Secondary → When a read maps to multiple locations, the alternative, alignments (bwa mem doesn't use this flag, gives quality score of 0 instead)
- Supplementary
- Duplicates
- Properly paired

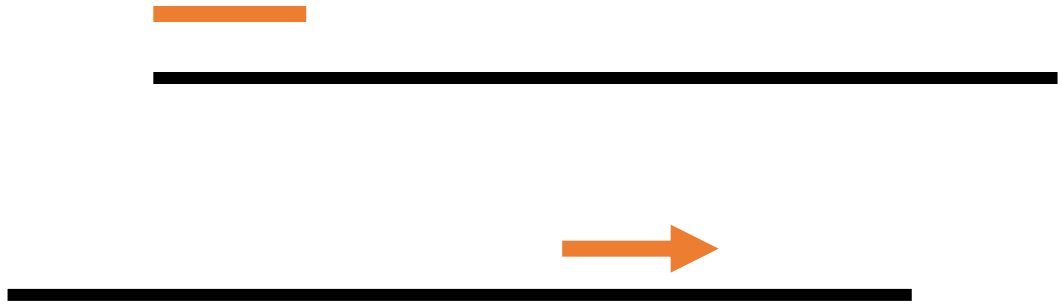
Alignment types

- Primary
- Secondary
- Supplementary
- Duplicates
- **Properly paired**



Alignment types

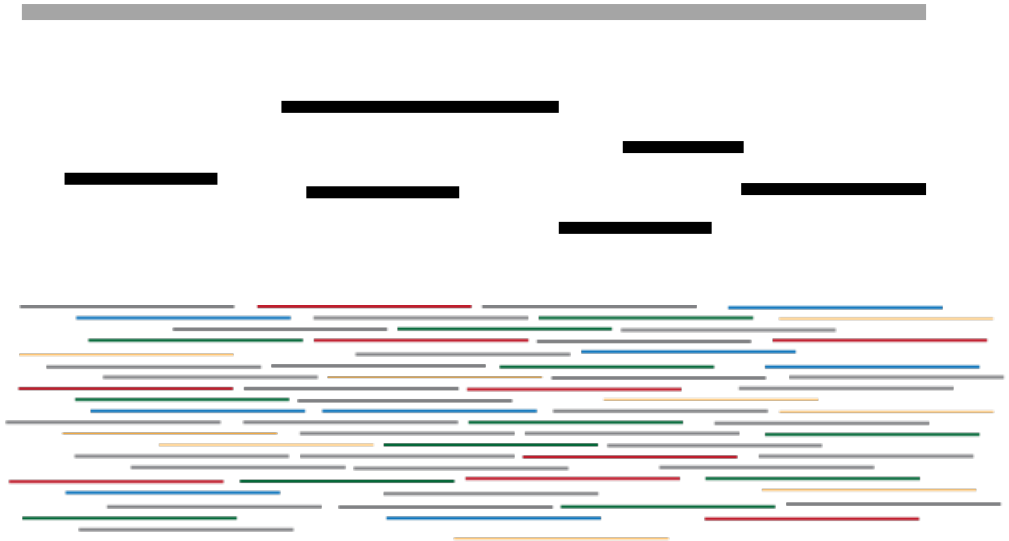
- Primary
- Secondary
- Supplementary → Split reads
- Duplicates
- Properly paired



Alignment types

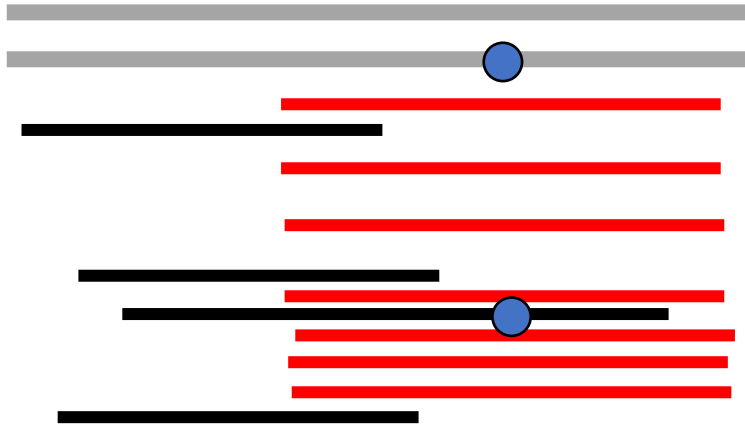
- Primary
- Secondary
- Supplementary
- Duplicates
- Properly paired

Full genome sequencing



Randomly fragment DNA

PCR amplify



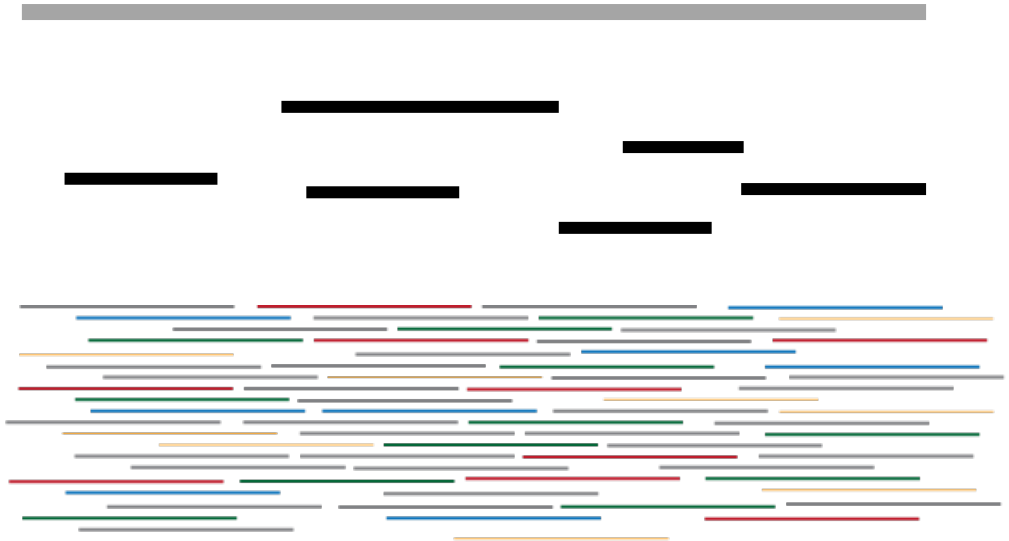
Map

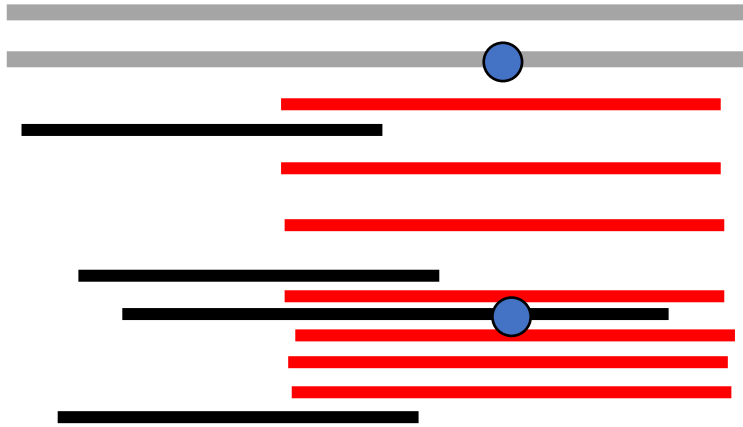
Each DNA fragment should be:

- Different length
- Different position

False homozygote

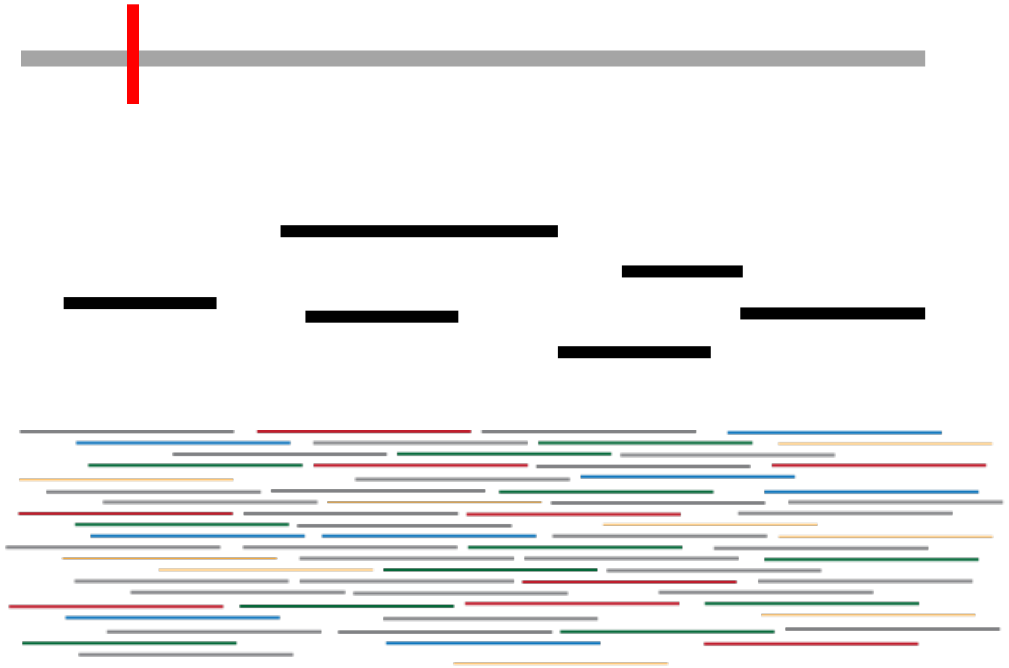
Full genome sequencing

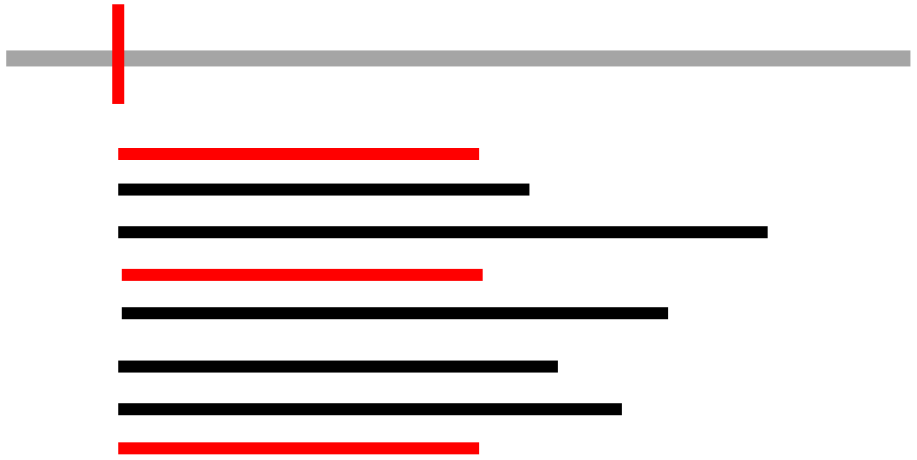




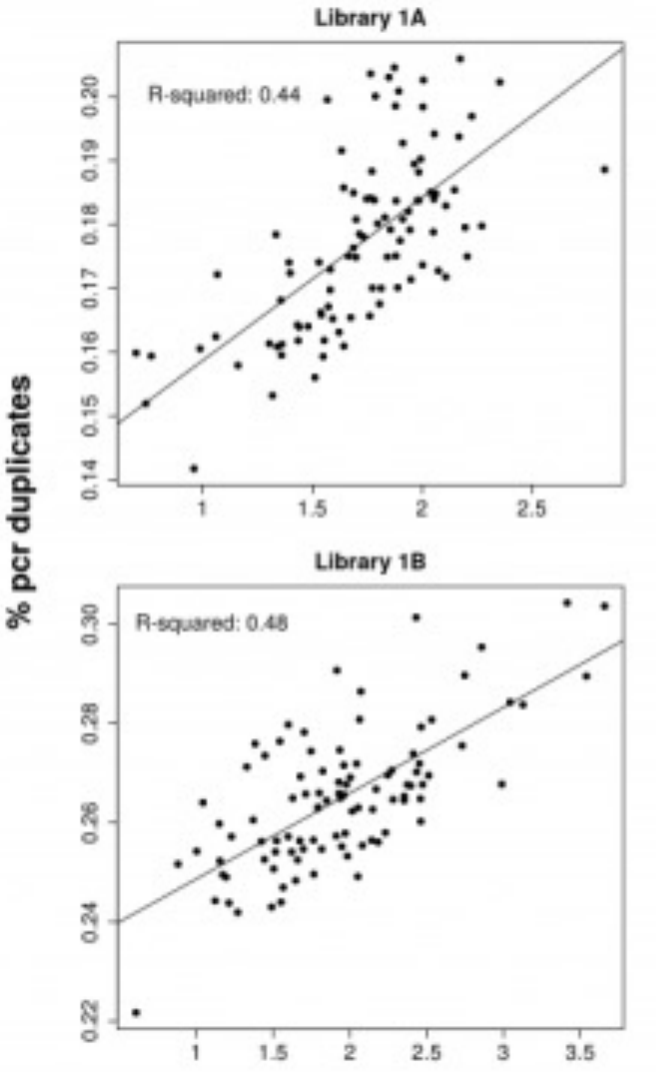
False homozygote

Rad seq

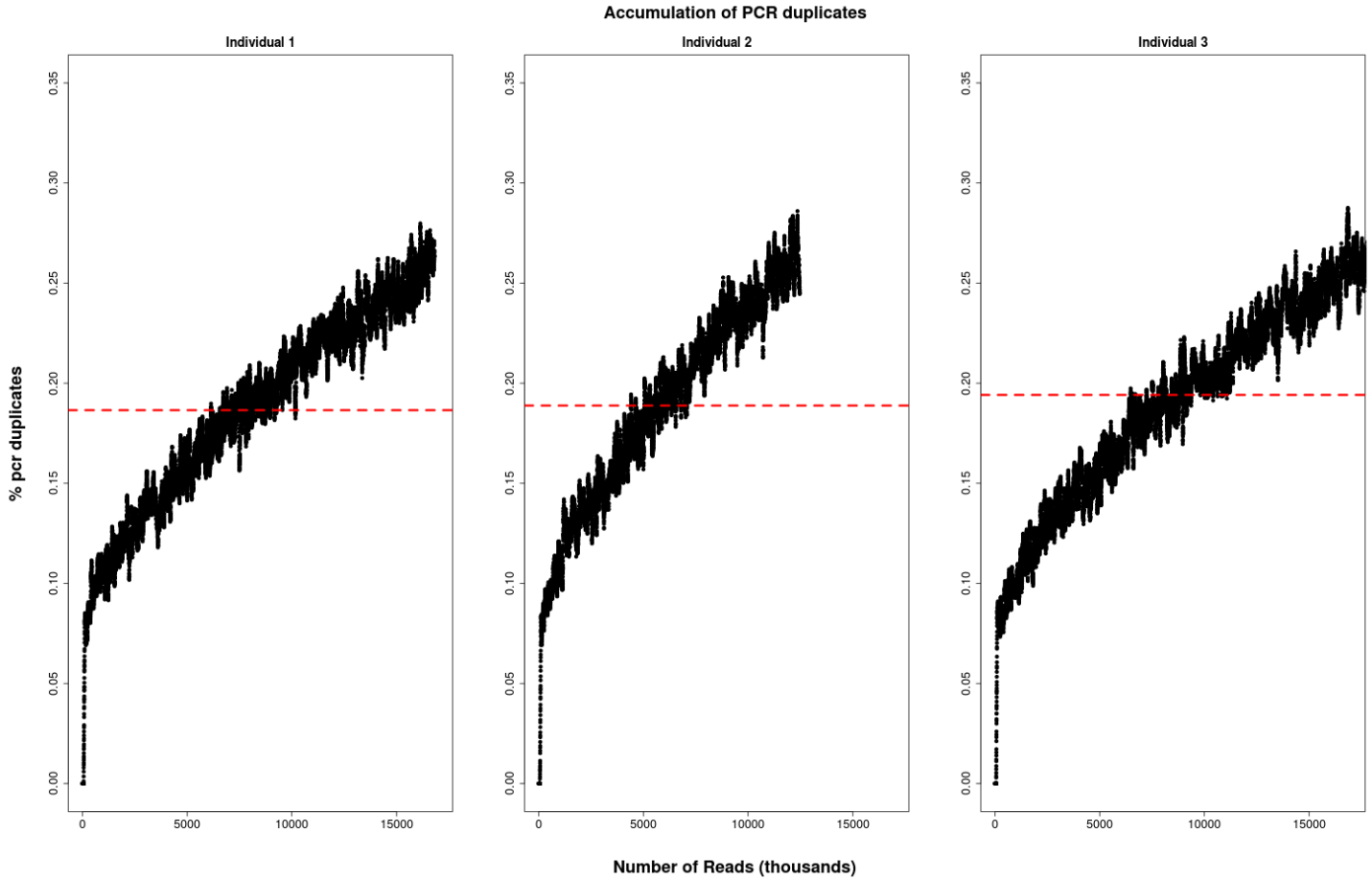




More depth = more duplicates



PCR duplicates accumulate as you process reads



For more info: https://www.molecularecologist.com/2016/08/25/the-trouble-with-pcr-duplicates/

Short Lecture: Variant calling

Simple in theory:

AGCTAGCCTAGTGACTAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

But in practice:

AGCTAGCCTAGTGACTAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

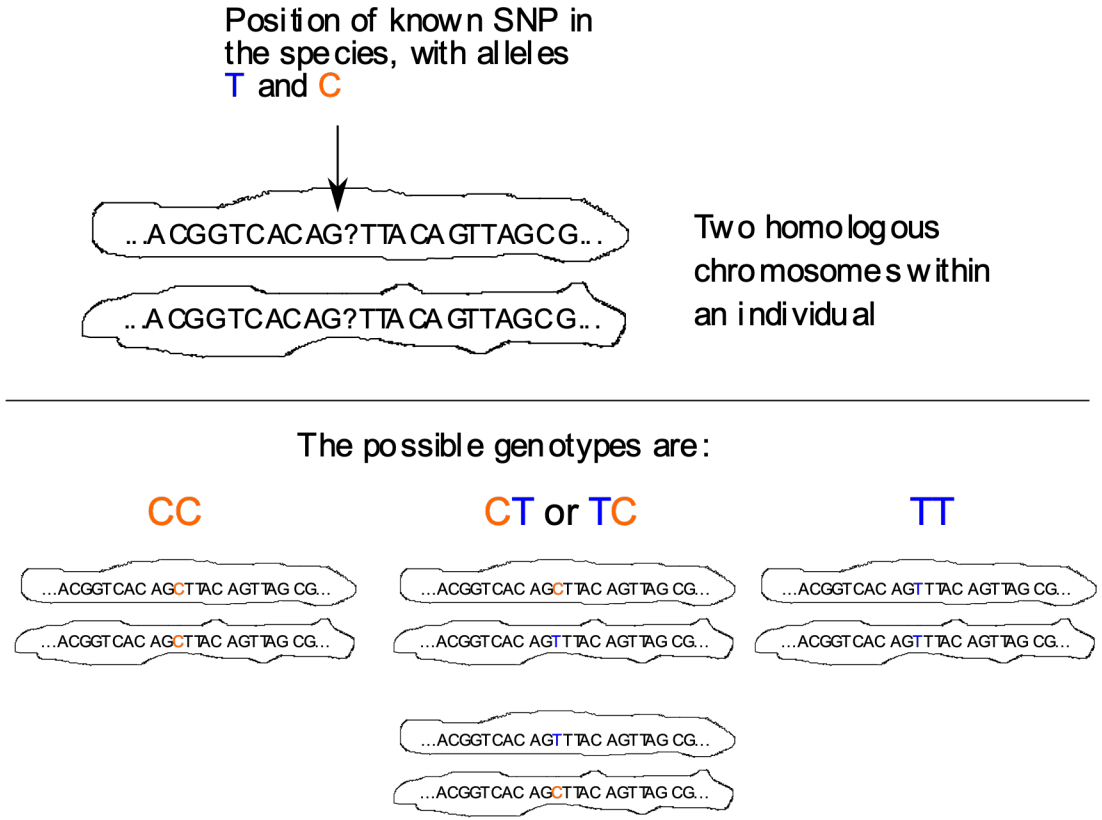
AGCTAGCCTAGTGACGAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

AGCTAGCCTAGTGACGAAGCCTGA

What is genetic variation, what is an error?

Genotype likelihood



The data are: 4 reads covering that site,
and
the associated base quality scores

Read #	*Read*	*Observed Base*	*PHRED-scaled base quality score*
1	CAGCTTACA	C	32 (A)
2	ACAGCT	C	37 (F)
3	GTTTA	T	35 (D)
4	AGCTTACAG	C	33 (B)

https://eriqande.github.io/eca-bioinf-handbook/variant-calling.html

Read #	Read	Observed Base	PHRED-scaled base quality score
1	CAGCTTACA	C	32 (A)
2	ACAGCT	C	37 (F)
3	GTTTA	T	35 (D)
4	AGCTTACAG	C	33 (B)

3 hypotheses:

- H1: CC
- H2: CT
- H3: TT

What is the evidence in our data for each hypothesis?

$$L(H_i \mid D) = P(D|H)$$

$$L(H_i \mid D) = L(CC \mid D)$$

To calculate likelihood → the probability of observing our data (C,C,T,C) given our hypothesis (CC) is true

We assume reads are randomly sampled from each chromosome and there is an error rate

If e = 0.00063

Genotype likelihoods if we have only one read that is C

$P\ (R1 = C \mid G = CC)$	1-e	0.99937
$P\ (R1 = C \mid G = CT)$	$\frac{1}{2}\ (1 - e) + \frac{1}{2}(e) = \frac{1}{2}$	0.5
$P\ (R1 = C \mid G = TT)$	e	0.00063

https://eriqande.github.io/eca-bioinf-handbook/variant-calling.html

Read #	Read	Observed Base	PHRED-scaled base quality score
1	CAGCTTACA	C	32 (A)
2	ACAGCT	C	37 (F)
3	GTTTA	T	35 (D)
4	AGCTTACAG	C	33 (B)

3 hypotheses:

- H1: CC
- H2: CT
- H3: TT

If we have only read 3 (T), basically the opposite results:

If e = 0.00063

$P\ (R3 = T \mid G = CC)$	e	0.00063
$P\ (R3 = T \mid G = CT)$	$\frac{1}{2}\ (1 - e) + \frac{1}{2}(e) = \frac{1}{2}$	0.5
$P\ (R3 = T \mid G = TT)$	$1 - e$	0.99937

What if we have one additional C?

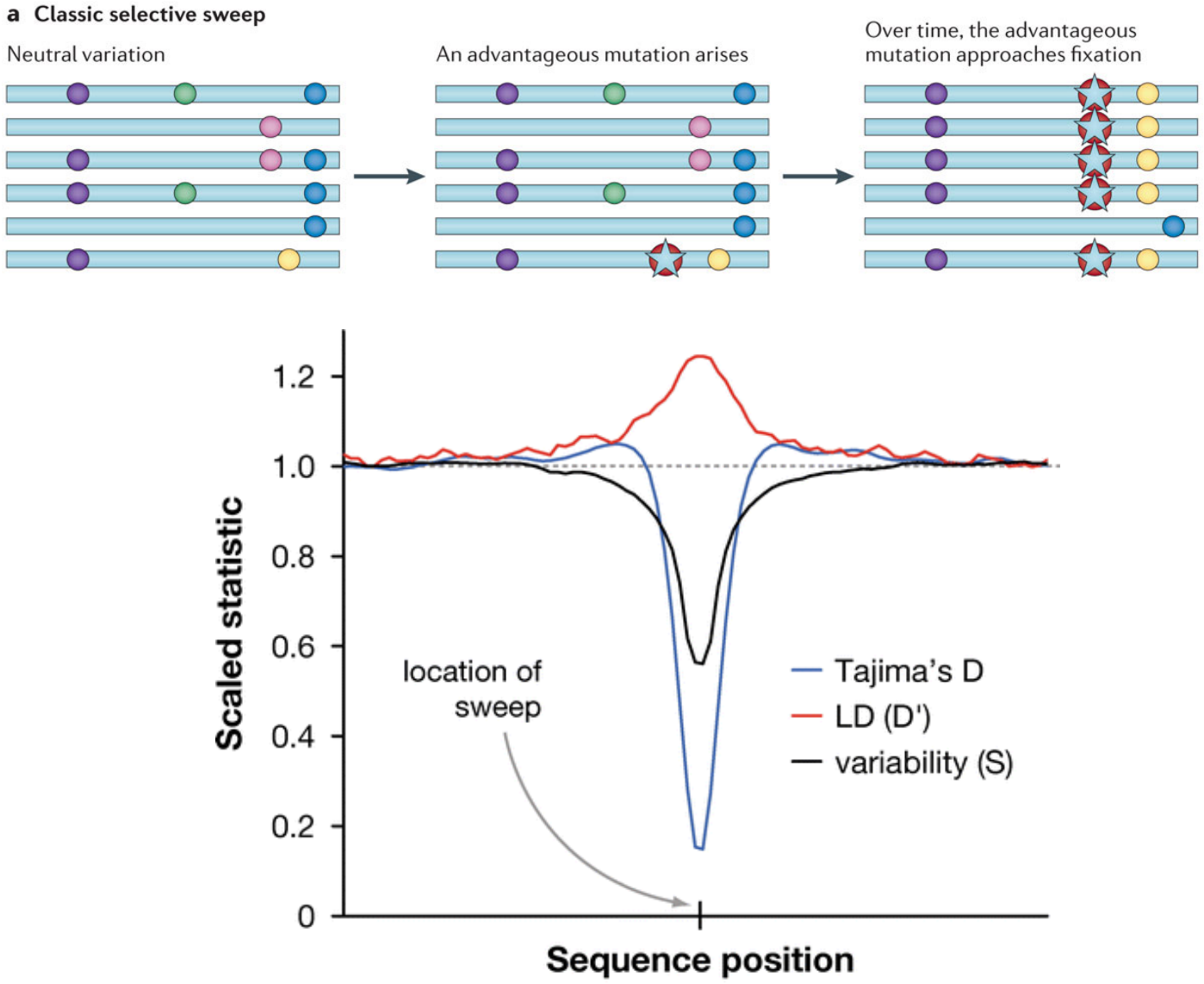
Multiply the probabilities together to get total likelihood:

L(G=CC|B1=C,B3=T)= 0.99937*0.00063 = 0.0006
L(G=CT|B1=C,B3=T)= 0.5*0.5 = 0.25
L(G=TT|B1=C,B3=T)= 0.99937*0.00063 = 0.0006

https://eriqande.github.io/eca-bioinf-handbook/variant-calling.html

Short Lecture: Selection Scans

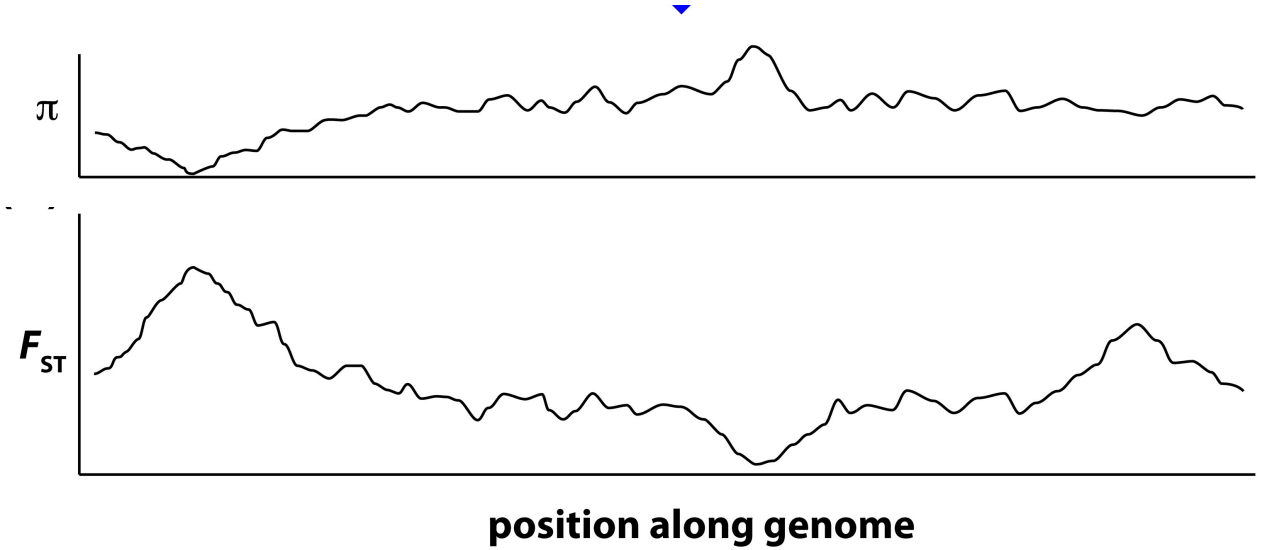
Types of selective sweeps

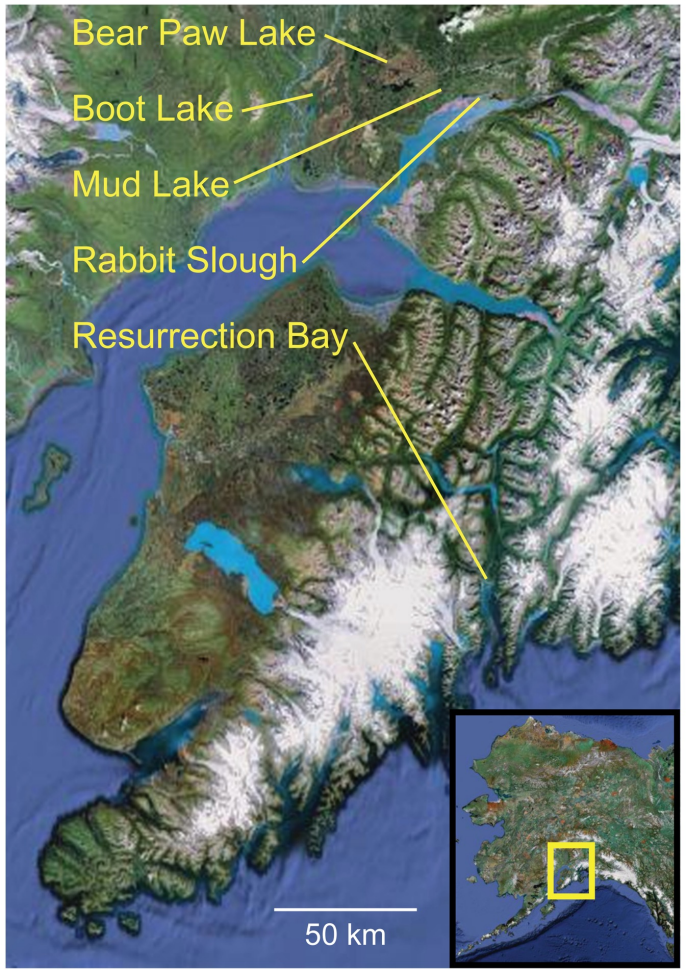


Nielsen R. 2005.
Annu. Rev. Genet. 39:197–218

Elevated: Fst, LD

Decreased: Tajima's D, genetic diversity





OPEN ACCESS Freely available online

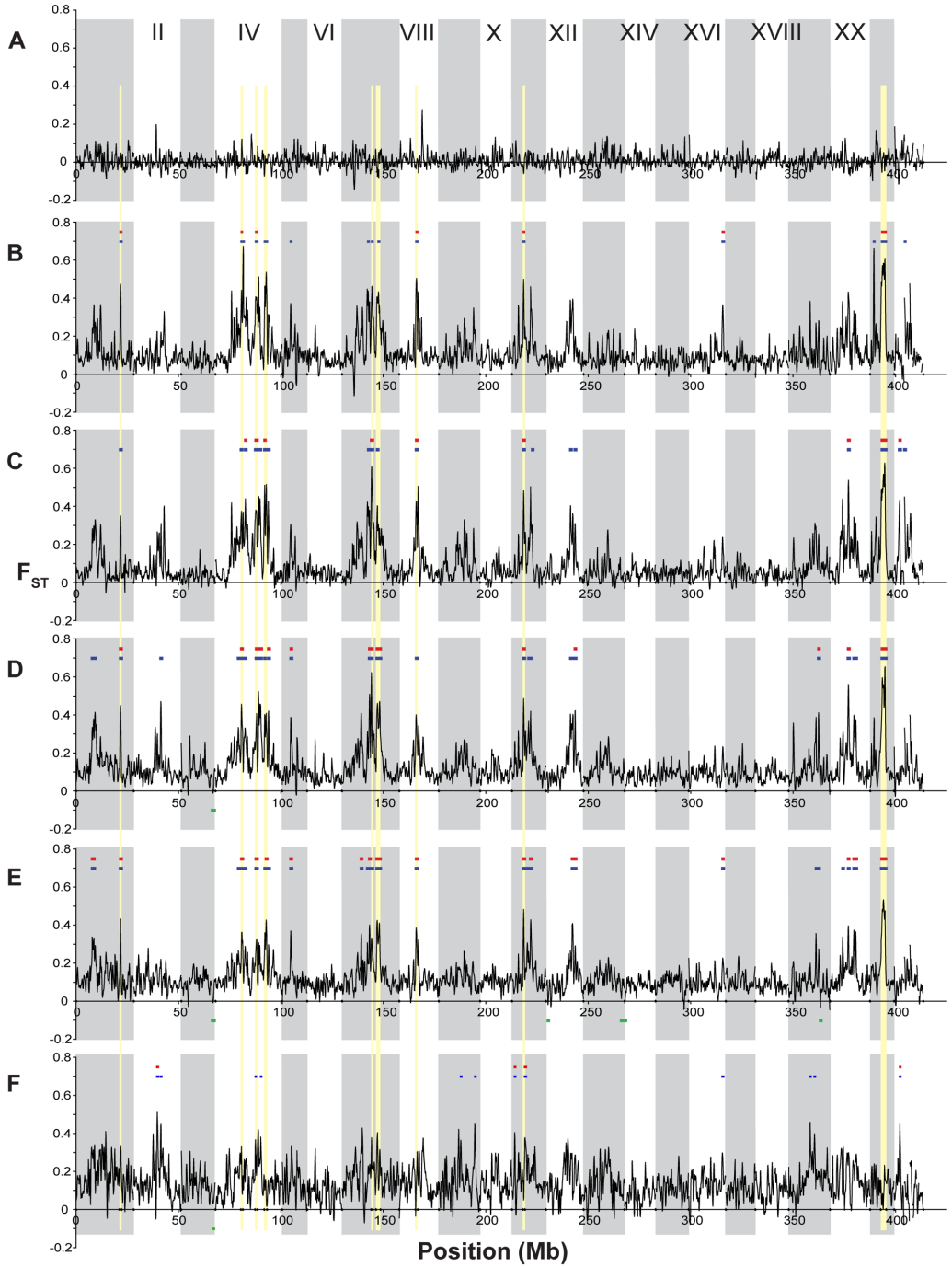
PLoS GENETICS

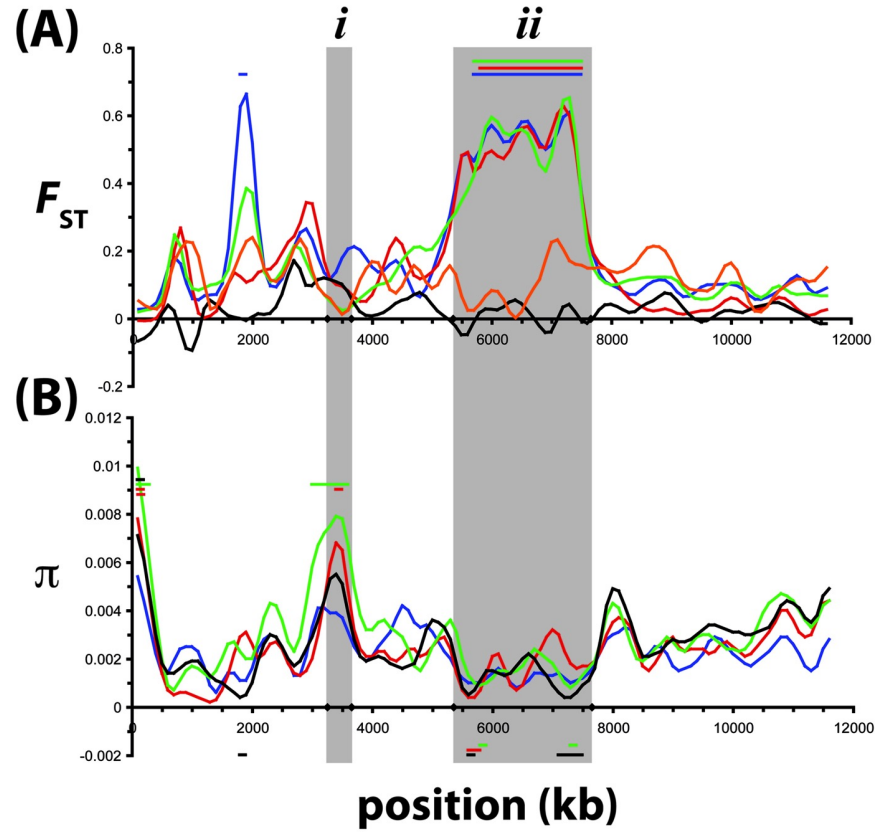
Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags

Paul A. Hohenlohe[1☯], Susan Bassham[1☯], Paul D. Etter[2], Nicholas Stiffler[3], Eric A. Johnson[2], William A. Cresko[1*]

1 Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon, United States of America, **2** Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America, **3** Genomics Core Facility, University of Oregon, Eugene, Oregon, United States of America

Hohenlohe, Phillips, and Cresko 2010

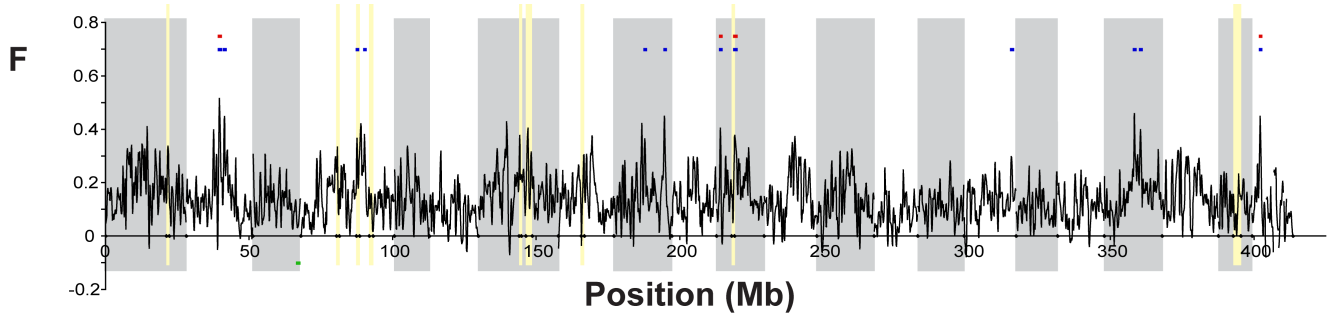


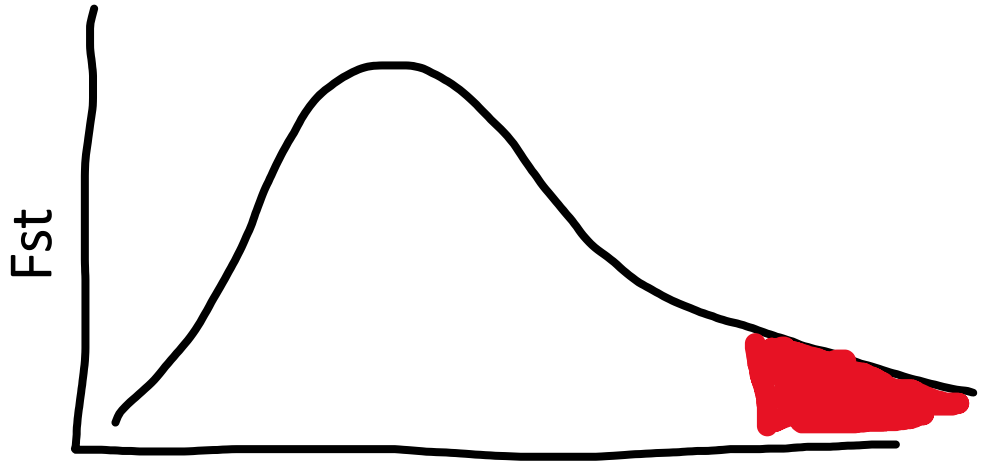


Hohenlohe, Phillips, and Cresko 2010

Detecting outlier regions

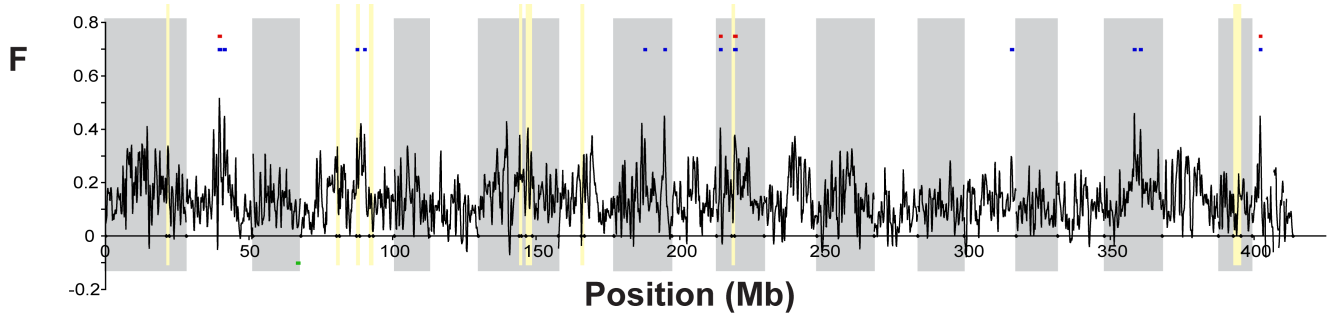
- What is the null distribution?
  - How to define “outliers”
  - Empirical: Tail of the distribution

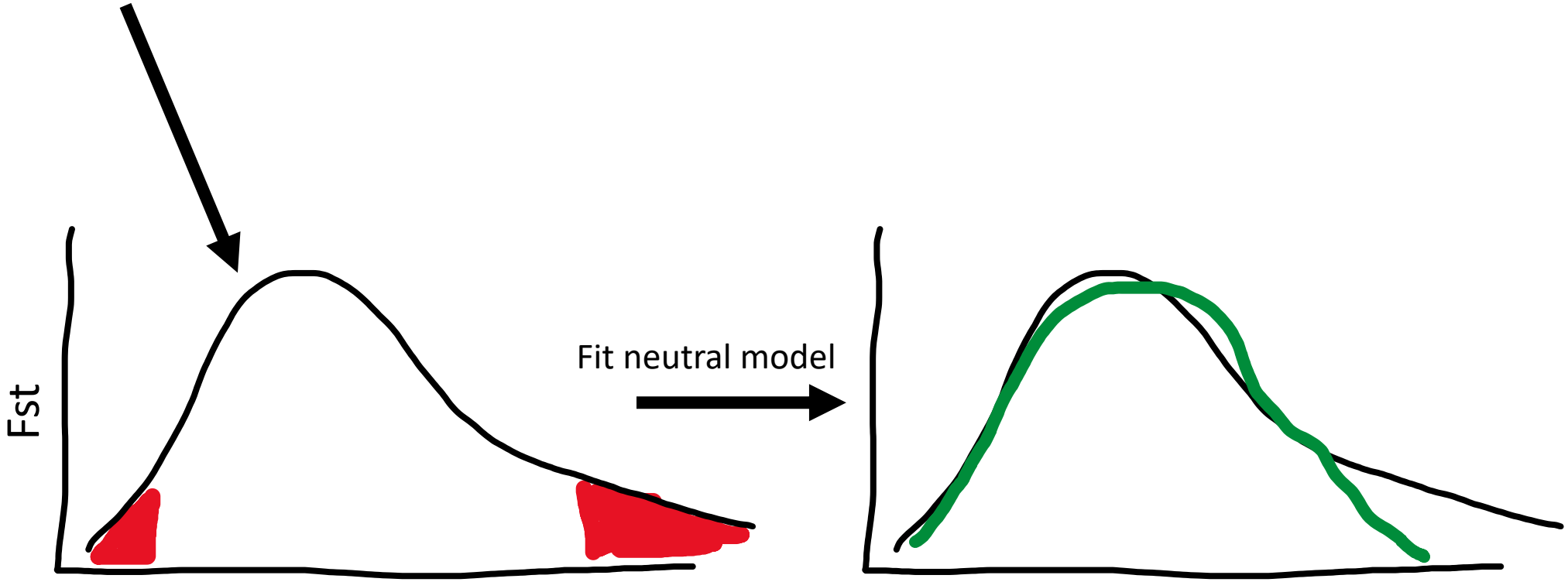




Detecting outlier regions

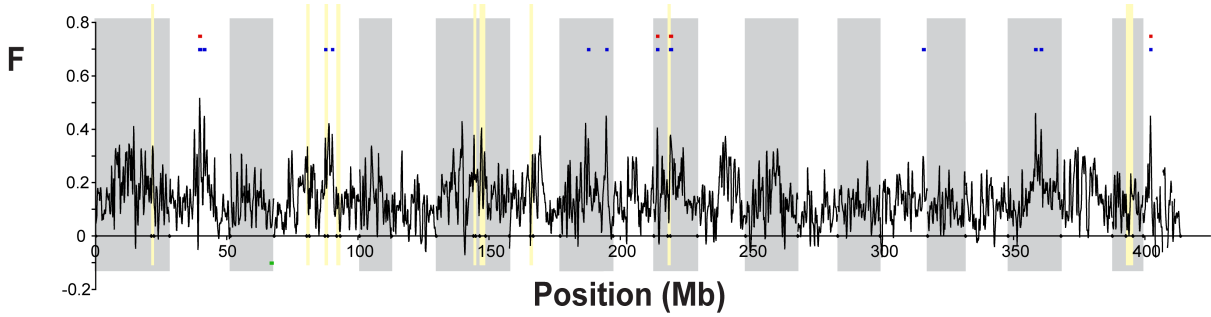
- What is the null distribution?
  - How to define “outliers”
  - Empirical: Tail of the distribution
  - OutFLANK

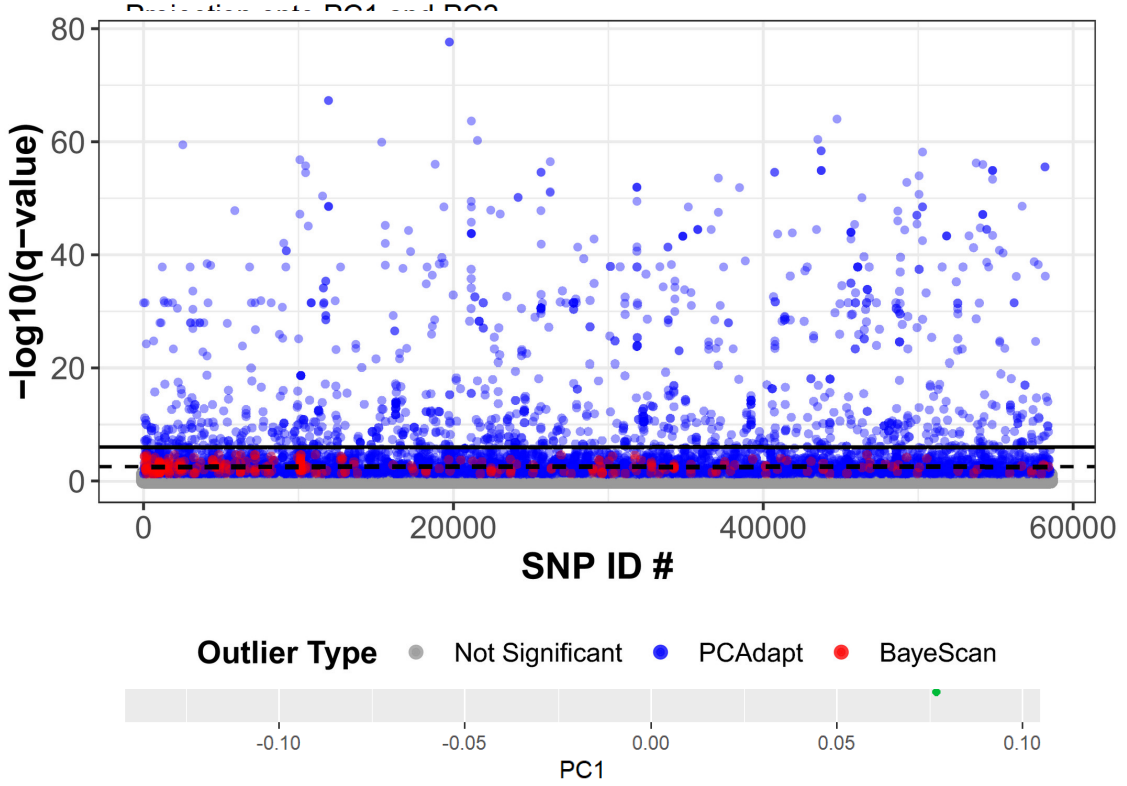




Detecting outlier regions

- Estimate covariance/structure among all populations, look for outliers after correction
- Baypass, bayenv2
- PCAdapt
  - PCA based
  - Don’t need population assignments
  - markers excessively related to population structure are candidates for local adaptation

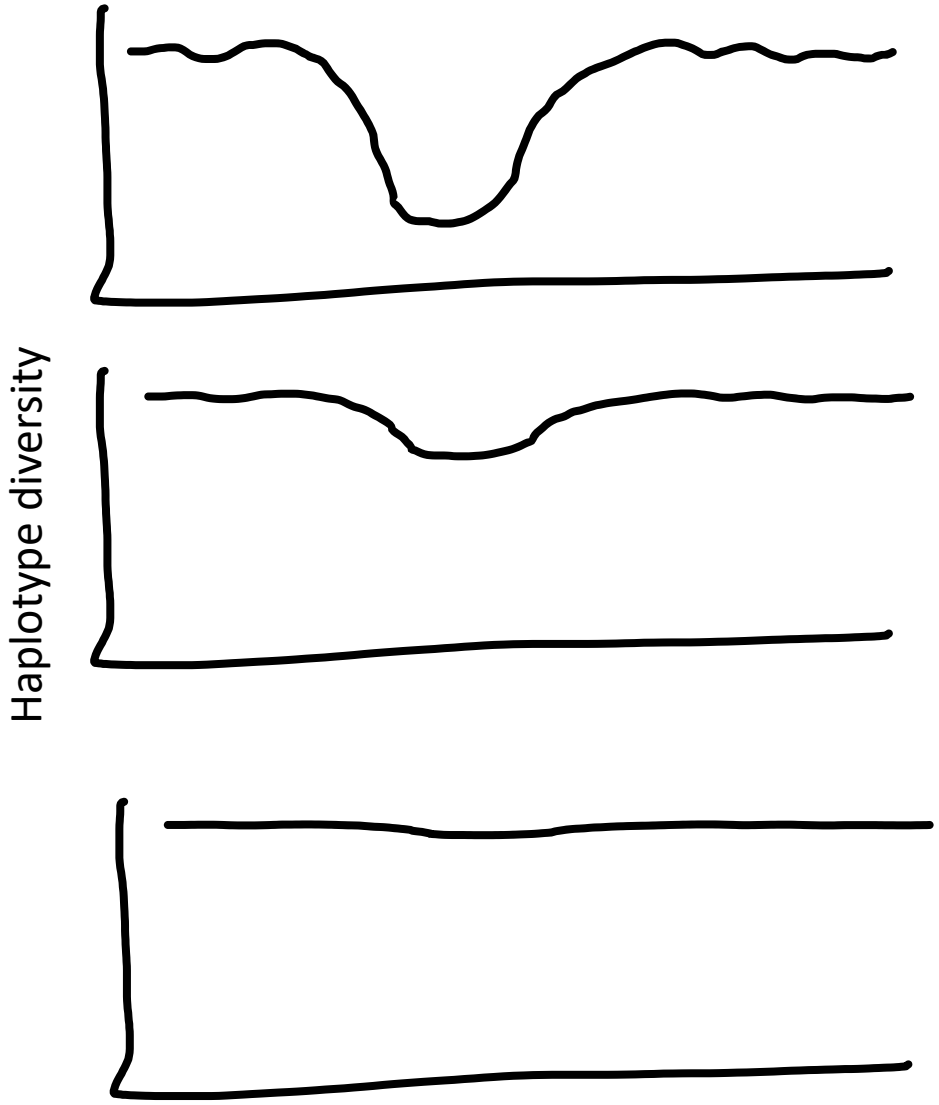




Reich et al., 2021

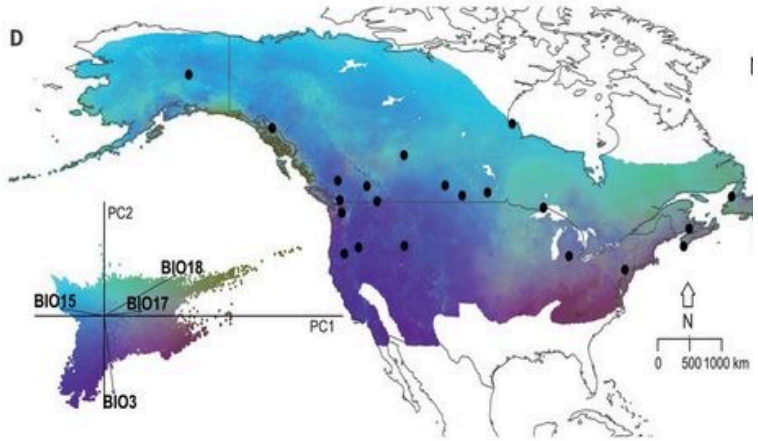
Luu, Bazin, Blum 2016

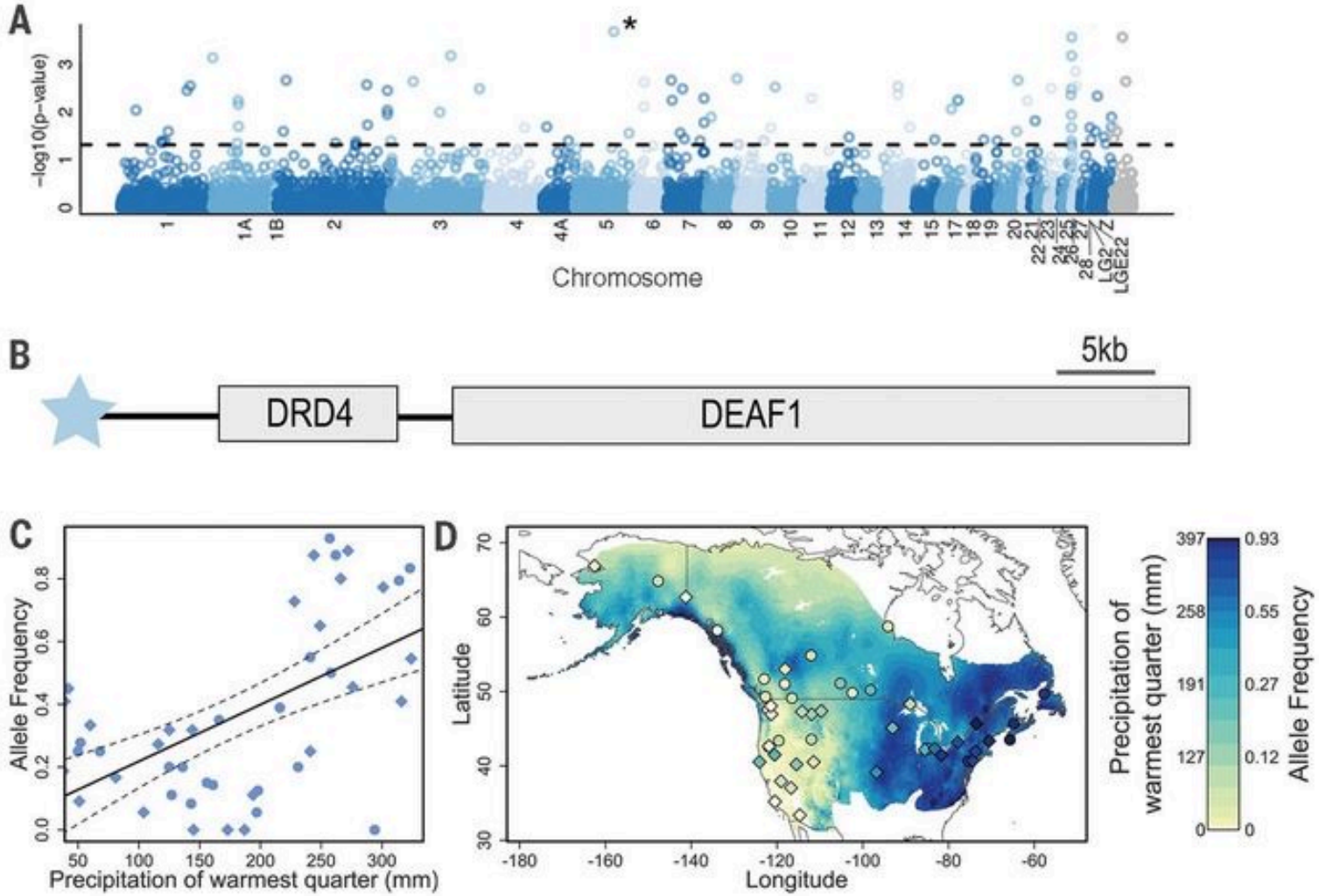
Types of selective sweeps



Gene environment associations

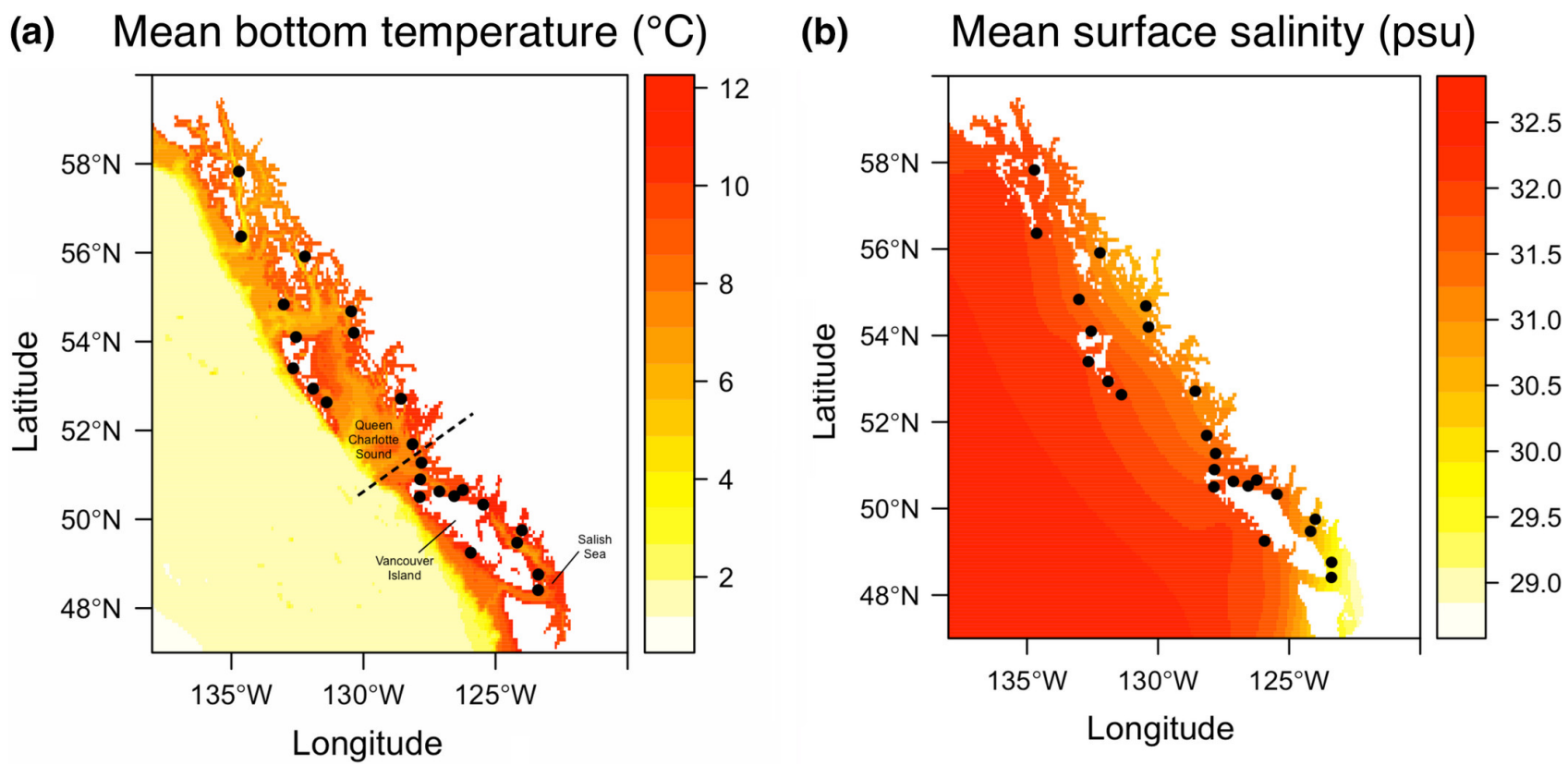
- Where is there a correlation between the environment and genetics?
- Important!
  - Must control for population structure!
- Methods:
  - Bayenv2
  - latent factor mixed models: LFMM2
  - Machine learning
    - Gradient Forests, random forests




A
*
−log10(p−value)
0
1
2
3
1
1A
1B
2
3
4
4A
5
6
7
8
9
10
11
12
13
14
15
17
18
19
20
21
22
23
24
25
26
27
28
LG2
LGE22
Chromosome
B
5kb
DRD4
DEAF1
C
Allele Frequency
0.0
0.2
0.4
0.6
0.8
50
100
150
200
250
300
Precipitation of warmest quarter (mm)
D
Latitude
30
40
50
60
70
-180
-160
-140
-120
-100
-80
-60
Longitude
Precipitation of warmest quarter (mm)
397
258
191
127
0
Allele Frequency
0.93
0.55
0.27
0.12
0

Putatively adaptive genetic variation in the giant California sea cucumber (*Parastichopus californicus*) as revealed by environmental association analysis of restriction-site associated DNA sequencing data

Amanda Xuereb[1] | Christopher M. Kimber[2] | Janelle M. R. Curtis[3] | Louis Bernatchez[4] | Marie-Josée Fortin[1]



Putatively adaptive genetic variation in the giant California sea cucumber (*Parastichopus californicus*) as revealed by environmental association analysis of restriction-site associated DNA sequencing data

Amanda Xuereb[1] | Christopher M. Kimber[2] | Janelle M. R. Curtis[3] | Louis Bernatchez[4] | Marie-Josée Fortin[1]