# Genomic analysis in non-model organisms

2013 Workshop on Genomics

**Český Krumlov**

Bill Cresko
Institute of Ecology and Evolution
Department of Biology
University of Oregon

UNIVERSITY OF OREGON

# Outline for today’s lecture

Genomic data and non-model organism research

RAD-seq for ecological & evolutionary genomics

Genomically enabling a non-model organism

*Stacks* software pipeline

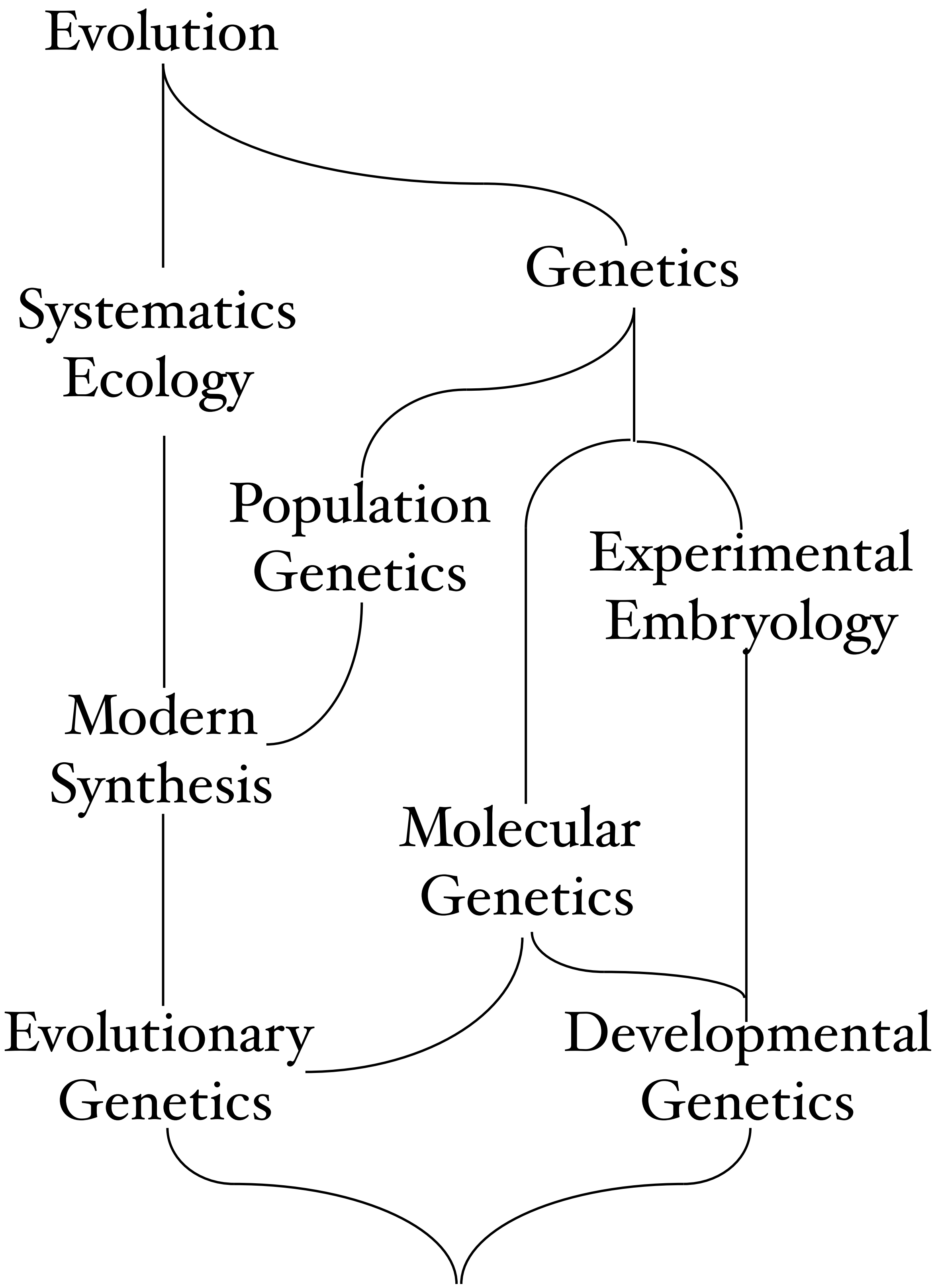
1850

1900

1950

2000

Conditions of Existence



Evolution

Systematics Ecology

Genetics

Population Genetics

Experimental Embryology

Modern Synthesis

Molecular Genetics

Evolutionary Genetics

Developmental Genetics

Unity of Type

functional evolutionary genomics

Modified from Gilbert 1998

# Model organism research has been very important

Vertebrate **zygotes** or embryos

28 day human

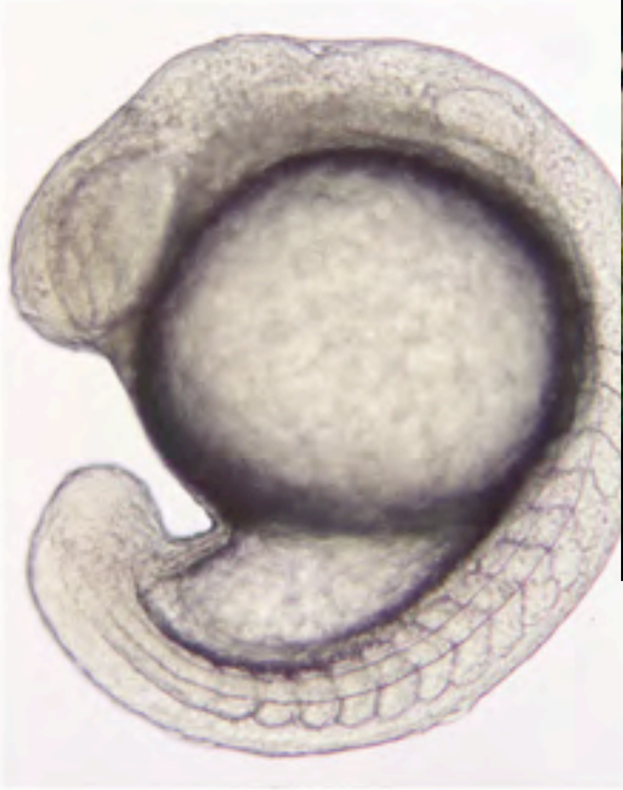
19h zebrafish

Video by Don Kane

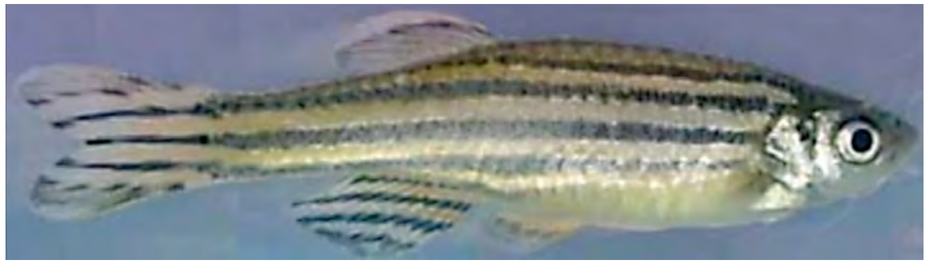
# Model organism research has been very important

Vertebrate **zygotes** or embryos

28 day human

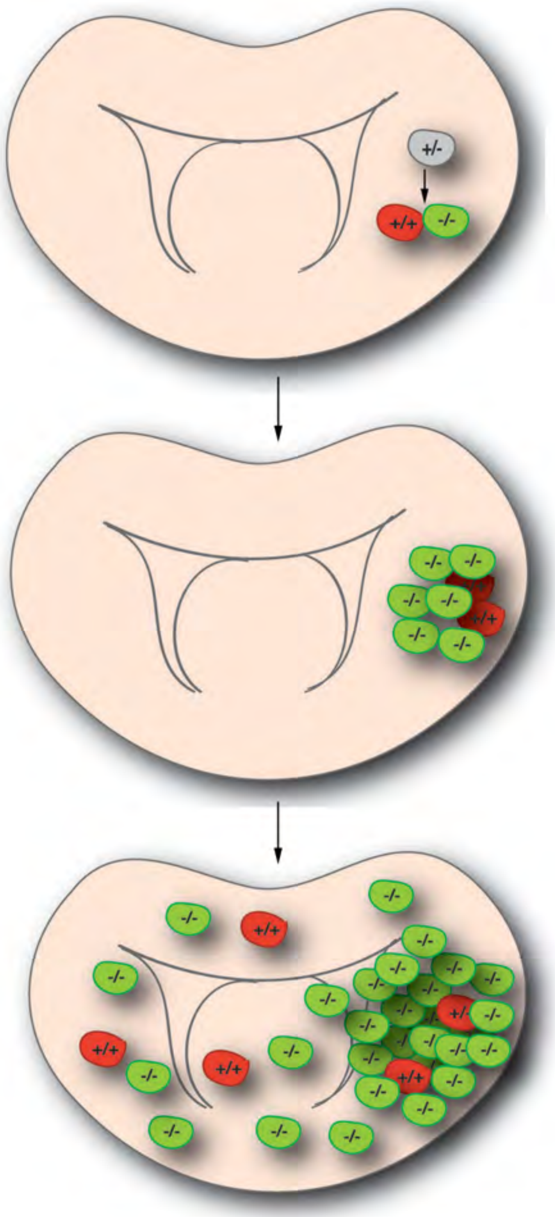


19h zebrafish

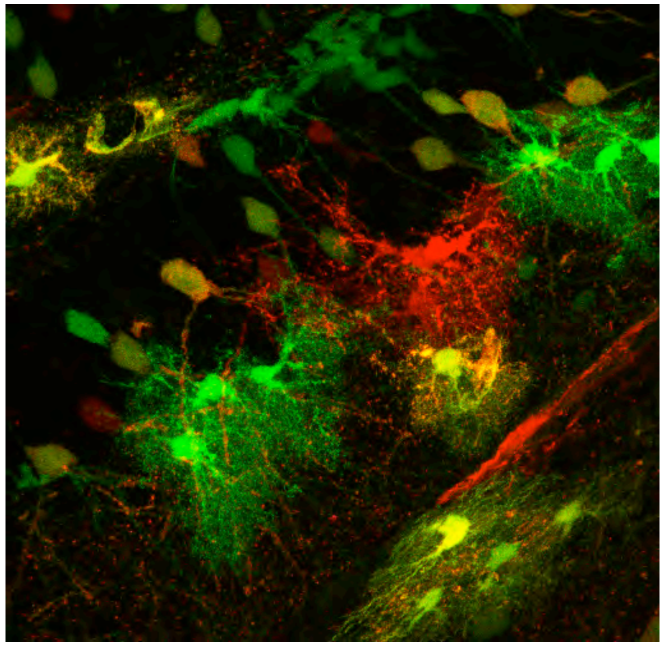


# Studying brain cancer using somatic evolutionary genomics in a model organism

pre-cancerous

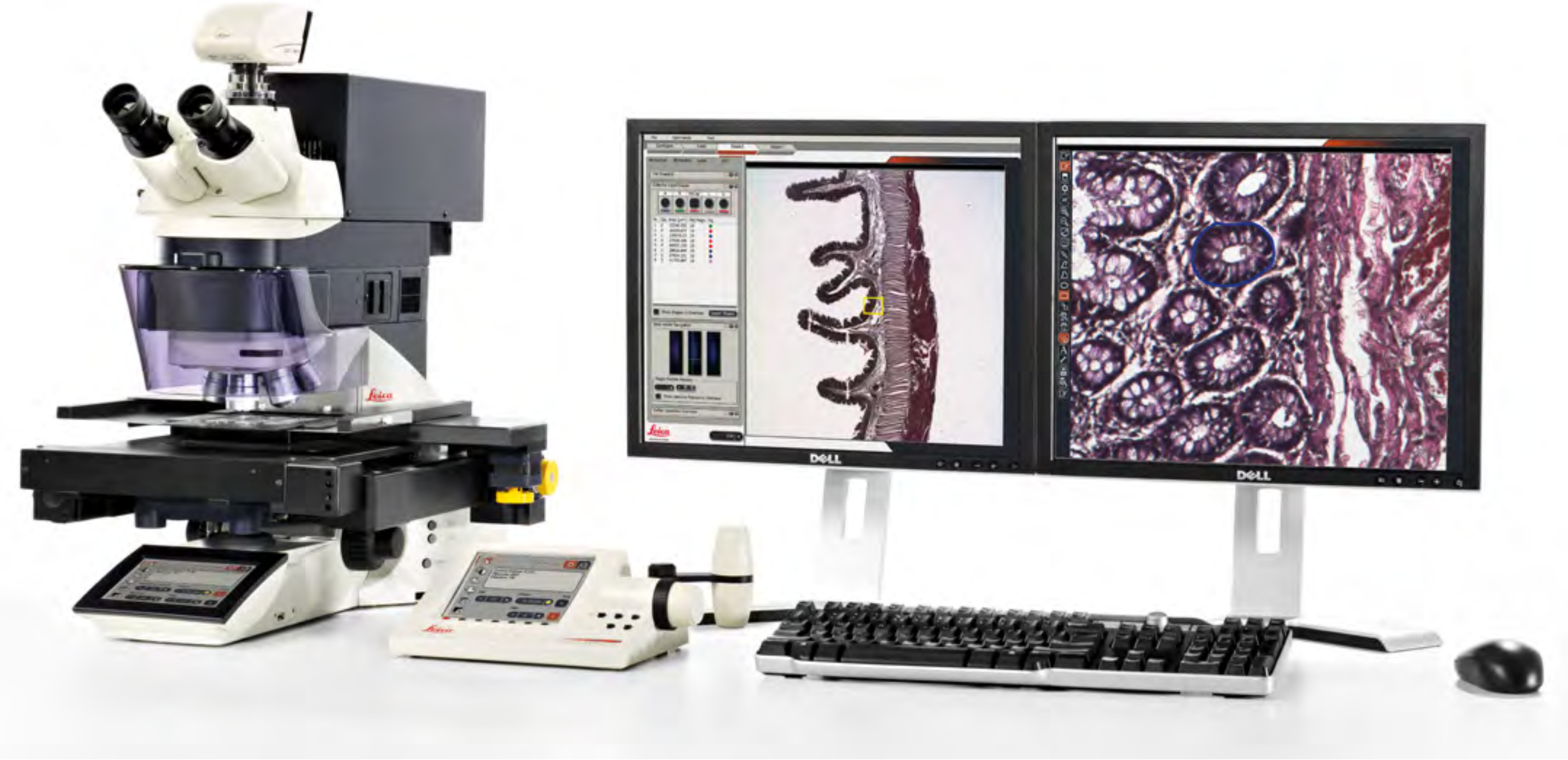


tumor

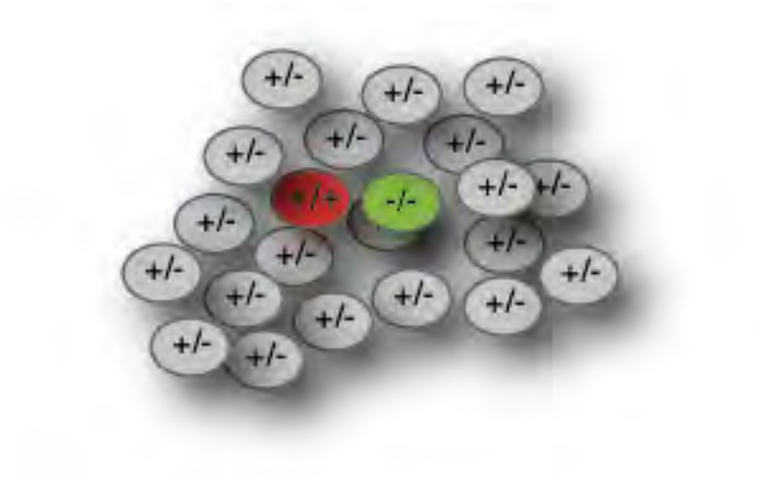


Hui Zong, Rui Galvao, Julian Catchen and Susie Bassham

# Laser Capture Microdissection of cells



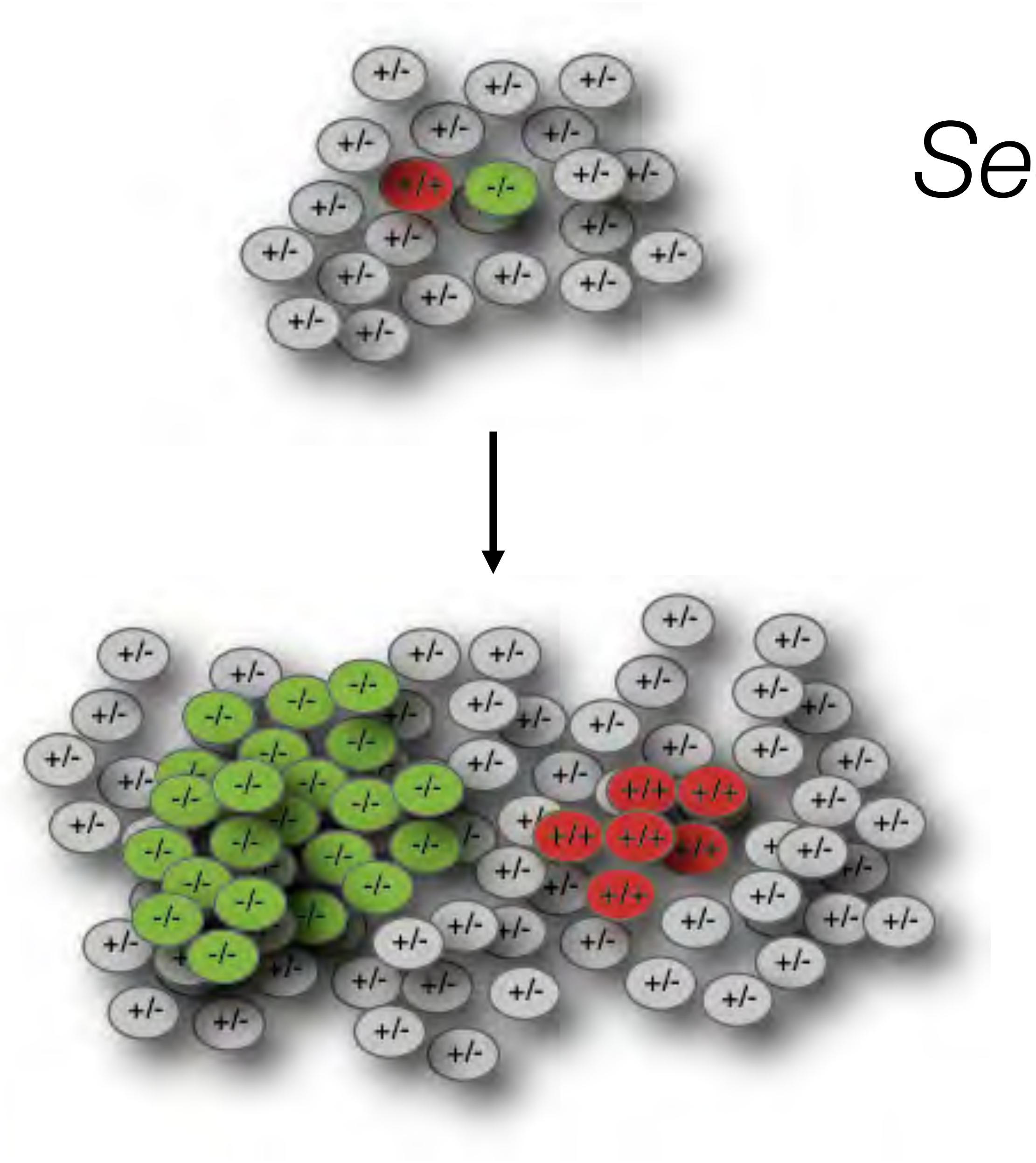
# Transcriptomic and genomic analysis of cells



*Sequence cells here...*

*... and here*

# Transcriptomic and genomic analysis of cells

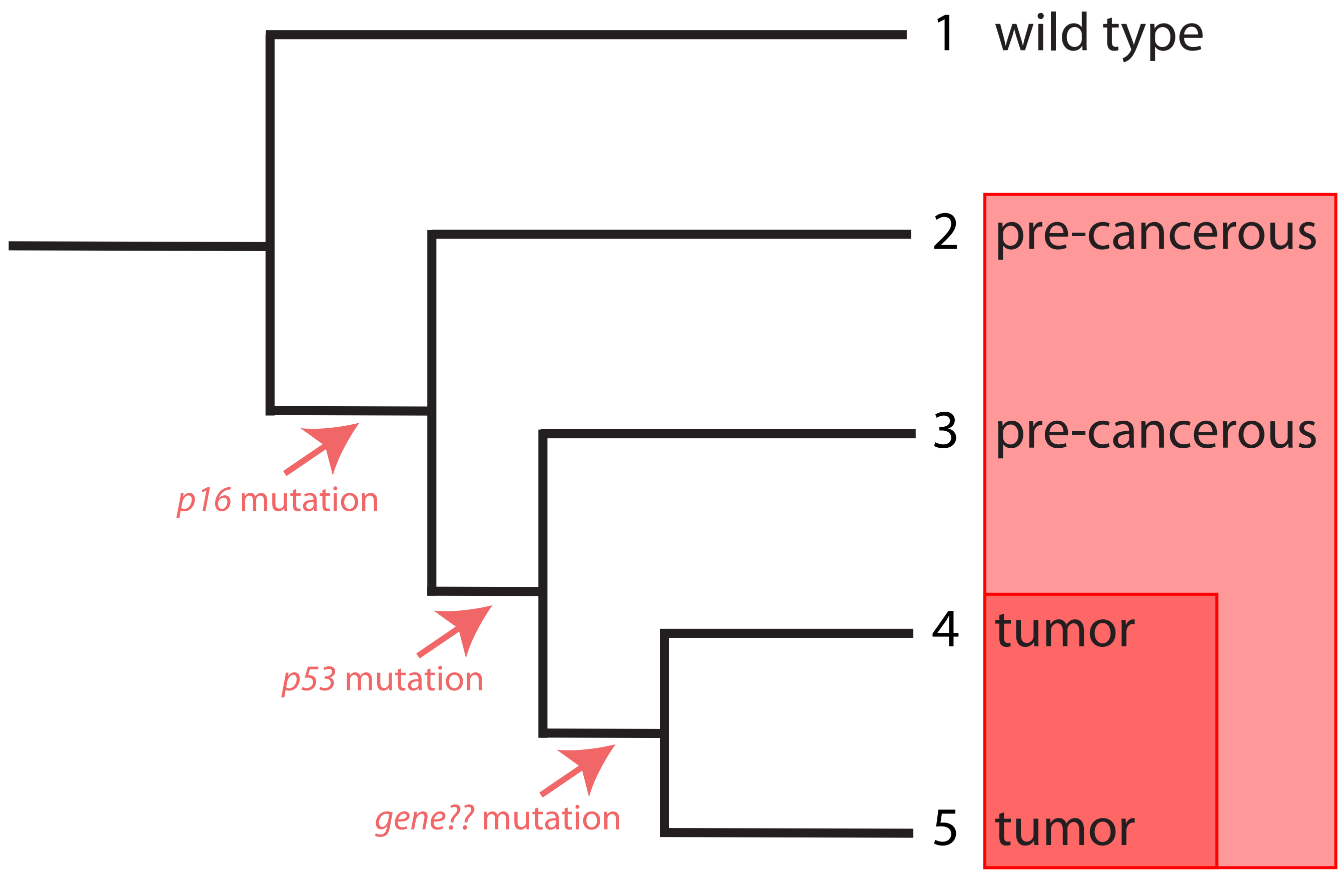


*Sequence cells here...*

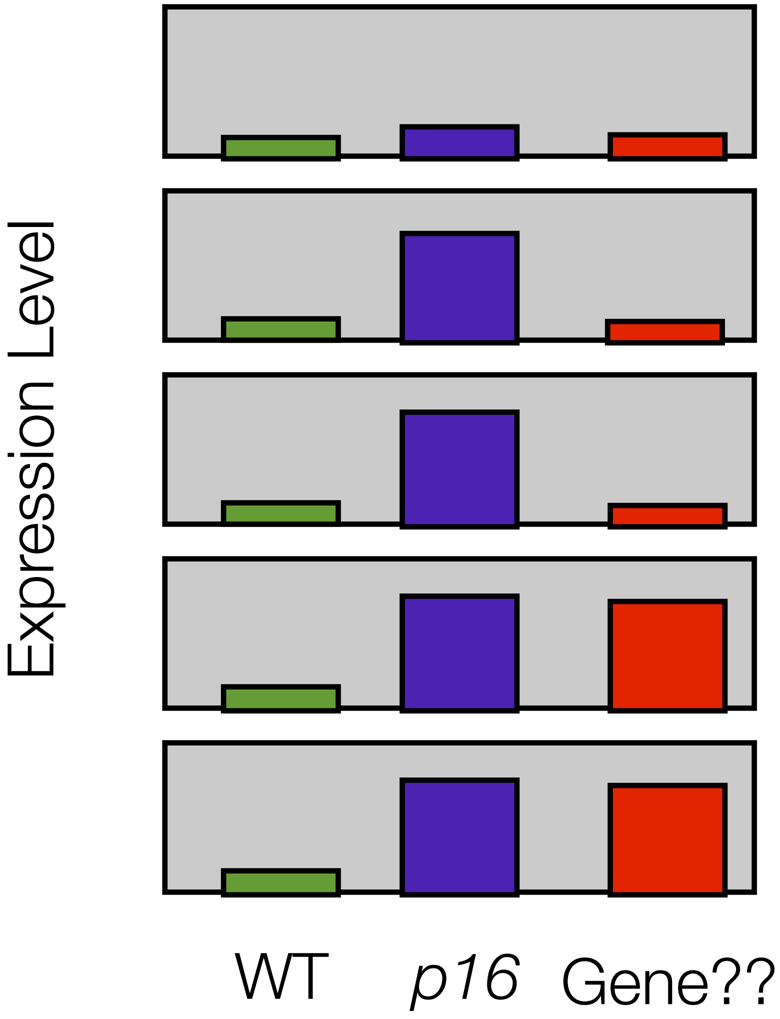
*... and here*

# Multiple lines of genetic evidence for causative mutations

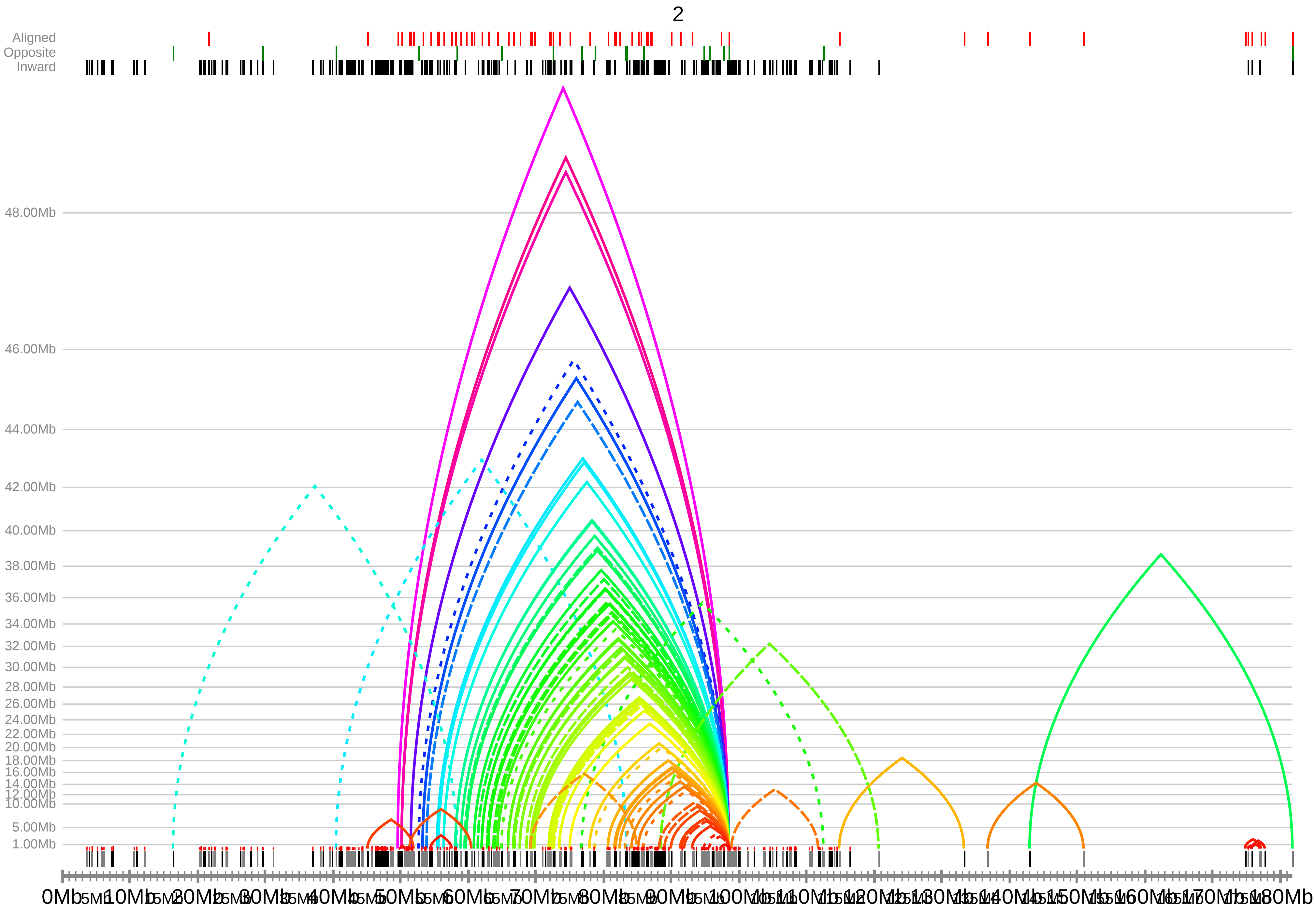
## Gene Mutations



## Gene Expression



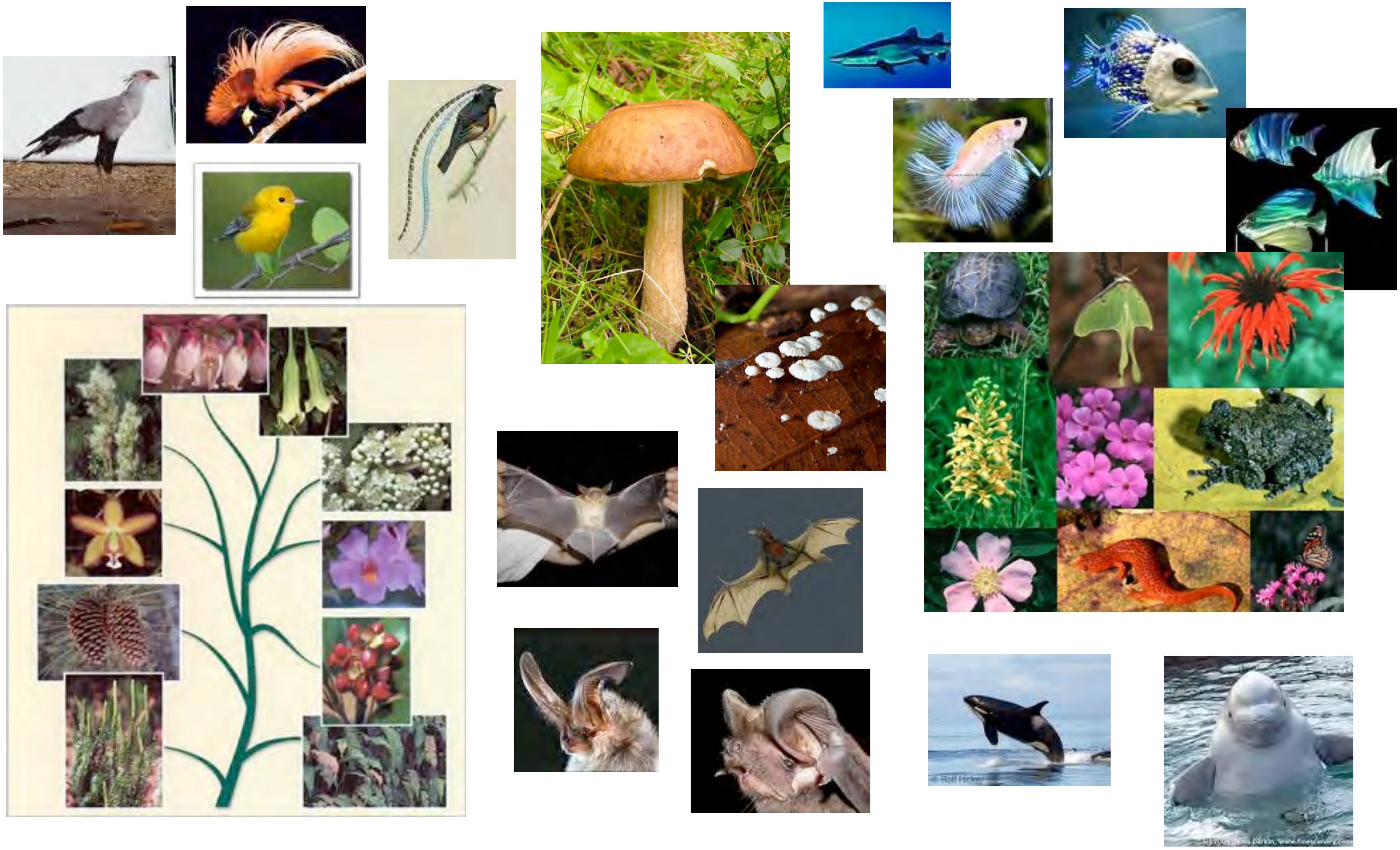
# Genomic rearrangements in cancer cells



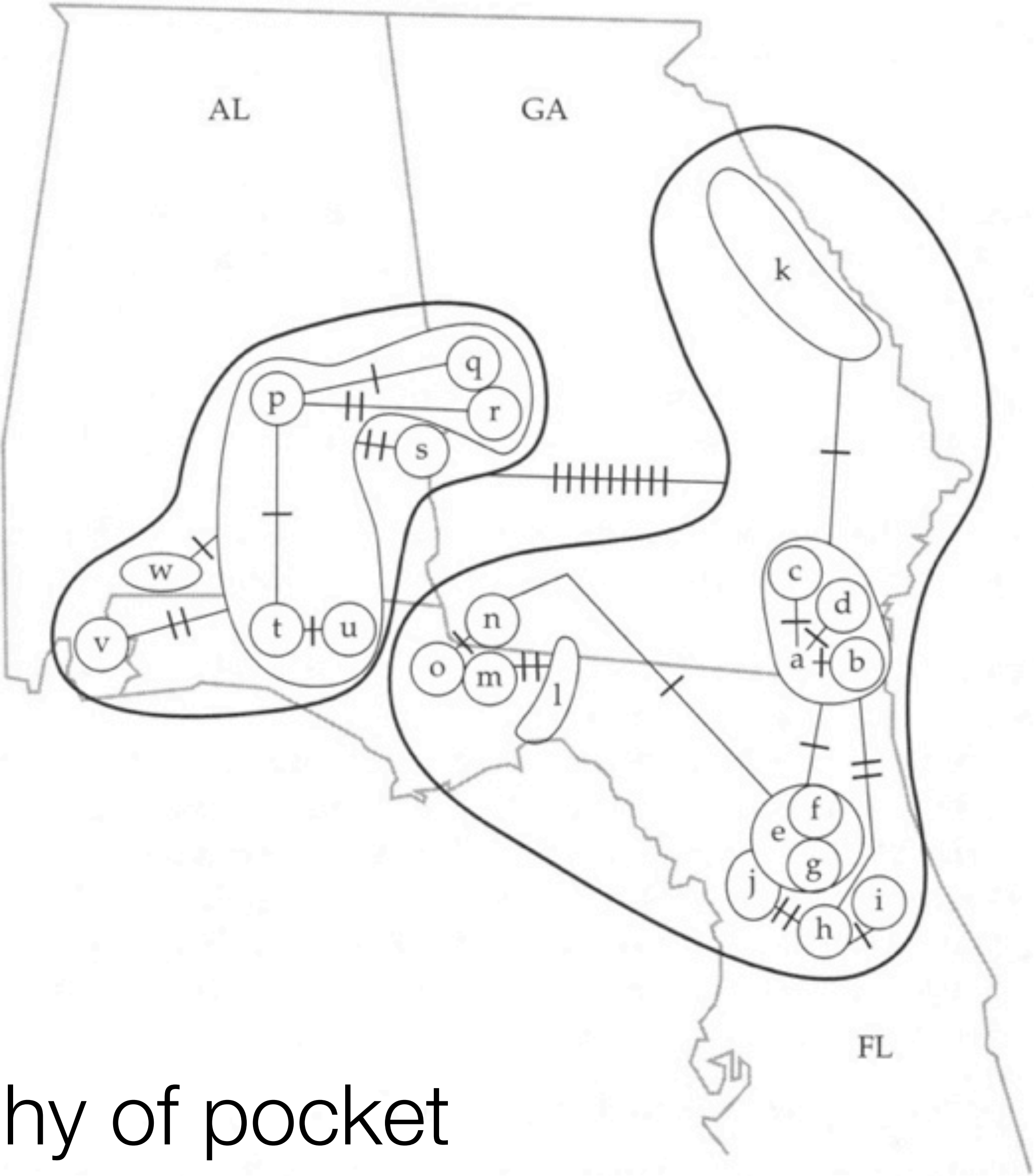
Julian Catchen

How can modern genomics improve studies of non-model organisms??

# How do the major differences among lineages evolve?



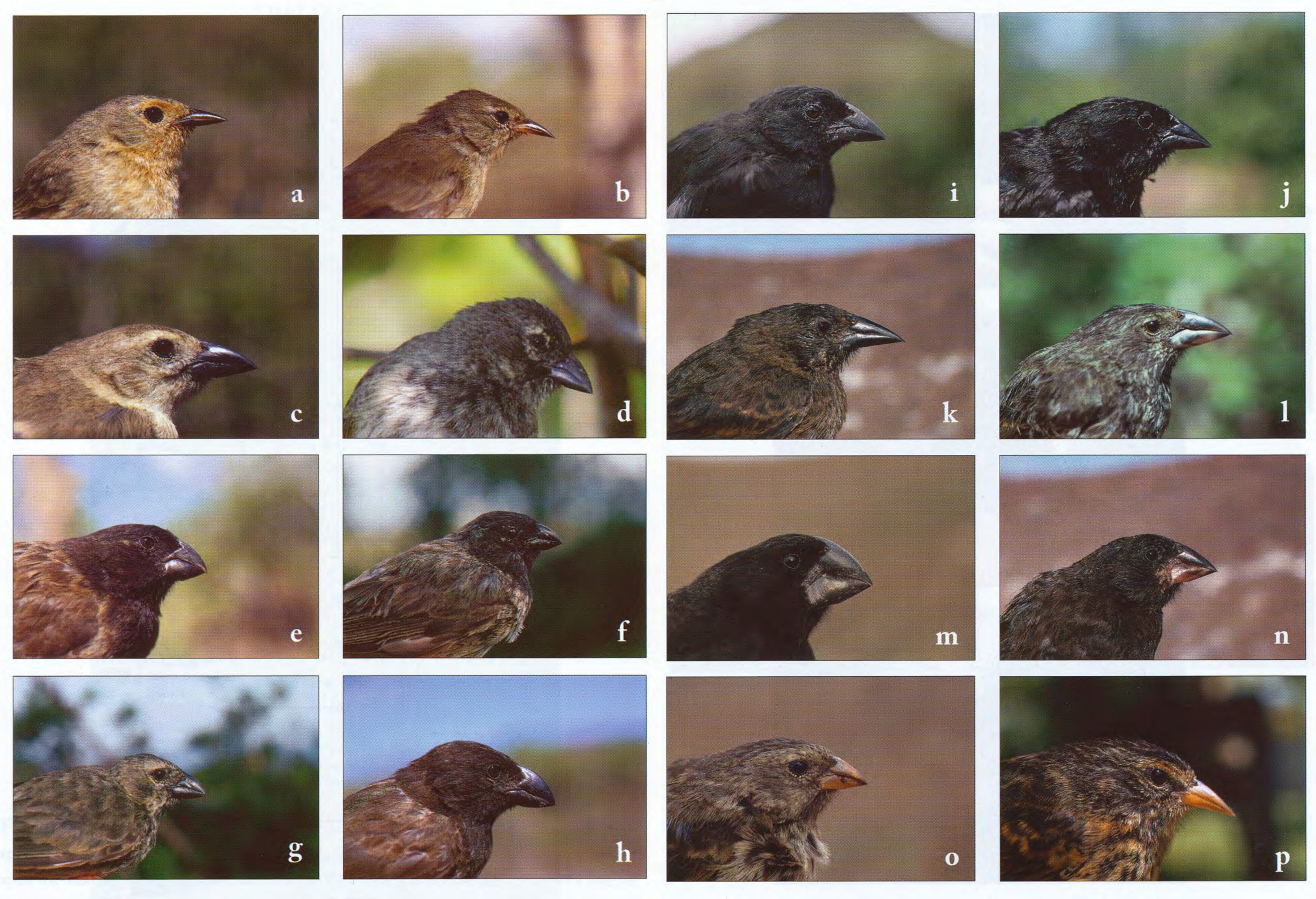
# How are organisms related to one another?



Phylogeography of pocket gophers using mtDNA

Avise, 1979

# How do organisms adapt to novel environments?



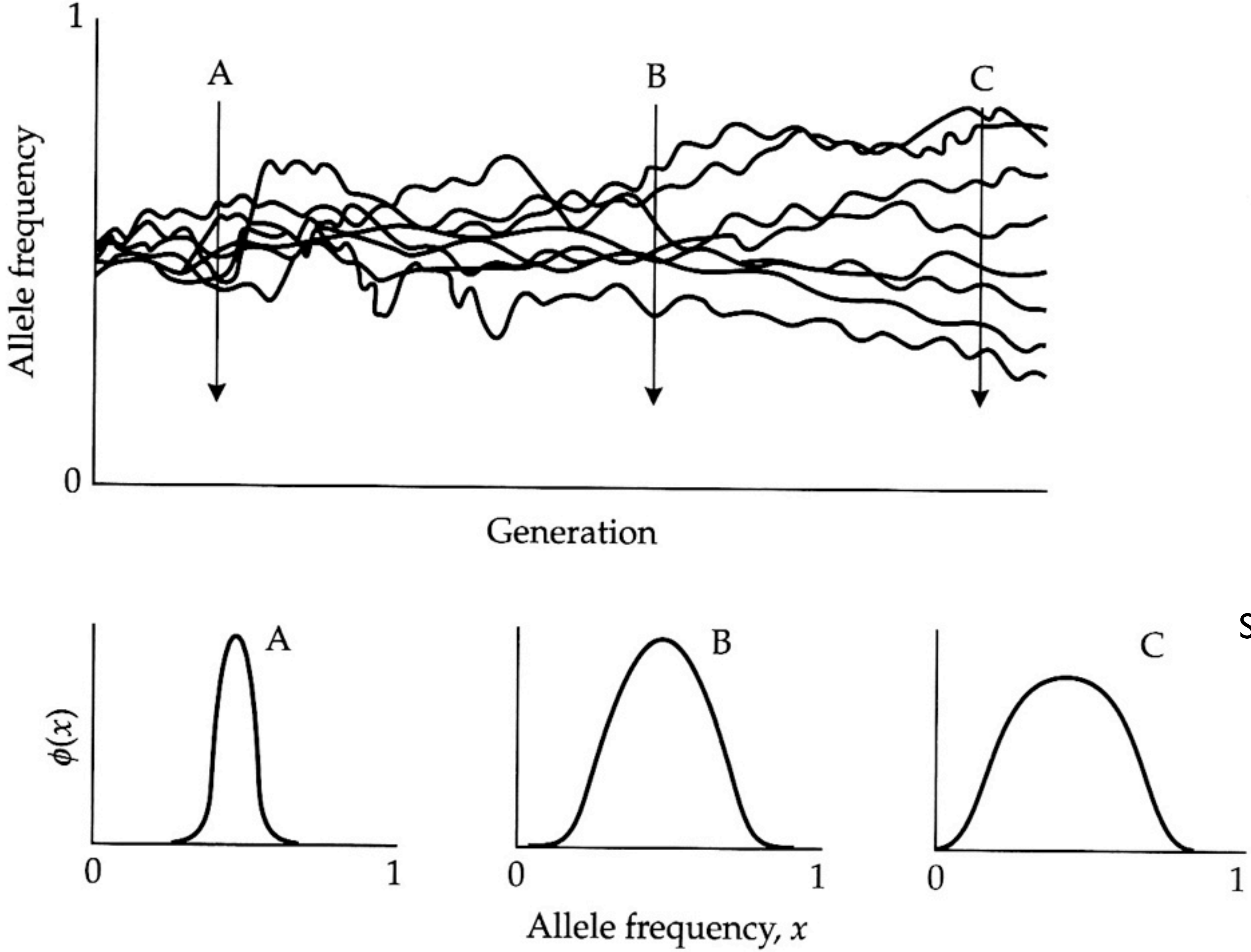
*from* Grant and Grant. 2007. How and why species multiply: The radiation of Darwin's finches. Princeton University Press

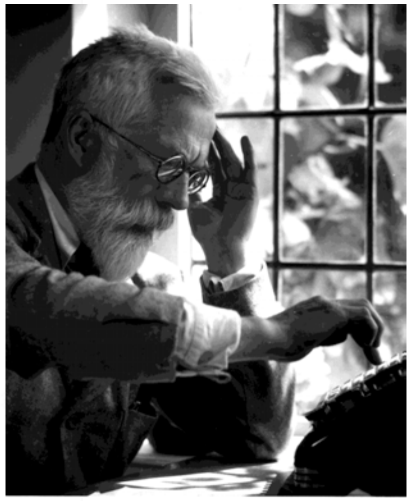
# Four fundamental processes in evolution

Origin of genetic variation;
**mutation**
**migration**

Sorting of variation;
**genetic drift**
**natural selection**

# Genetic drift is a null model

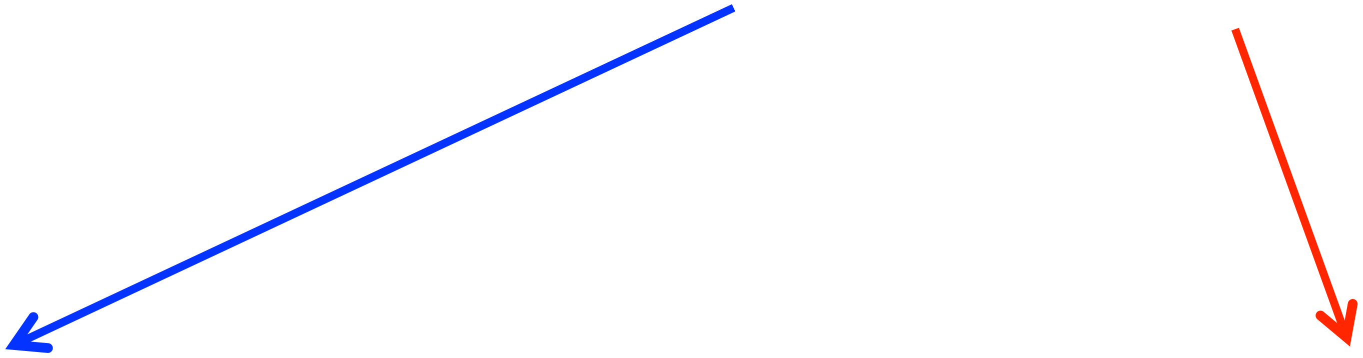



R. A. Fisher

Sewall Wright

# Population genomics

Simultaneous genotyping of **neutral** and **adaptive** loci



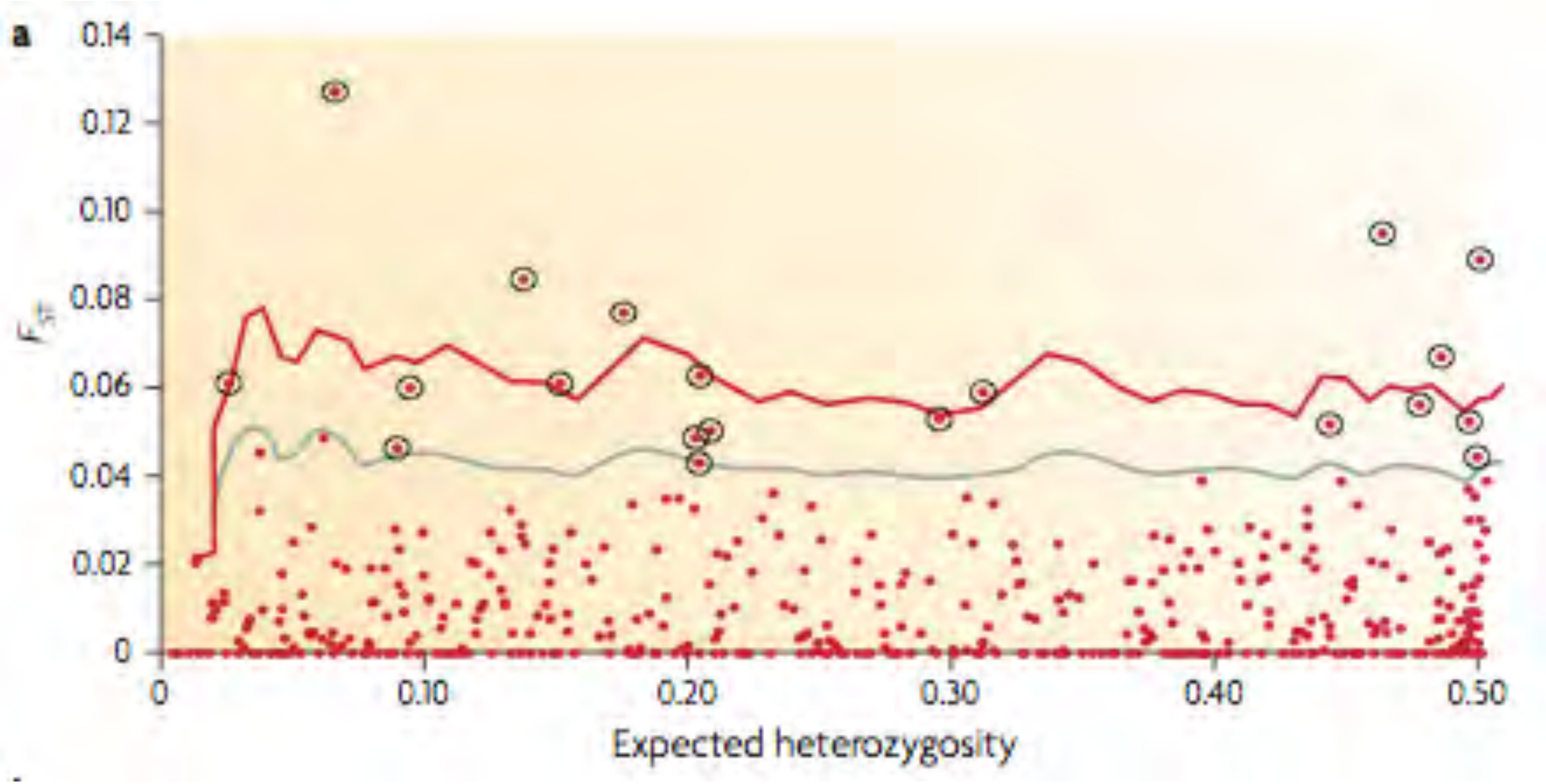
Genome-wide background provides more precise estimates:

- Demographic processes (e.g. $N_e$)
- Phylogeography

Outliers from background indicate:

- Selective sweeps
- Local adaptation

# Population genomics of unordered markers



Namroud et al (2008) *Mol Ecol* 17:3599

# Population genomics of ordered markers

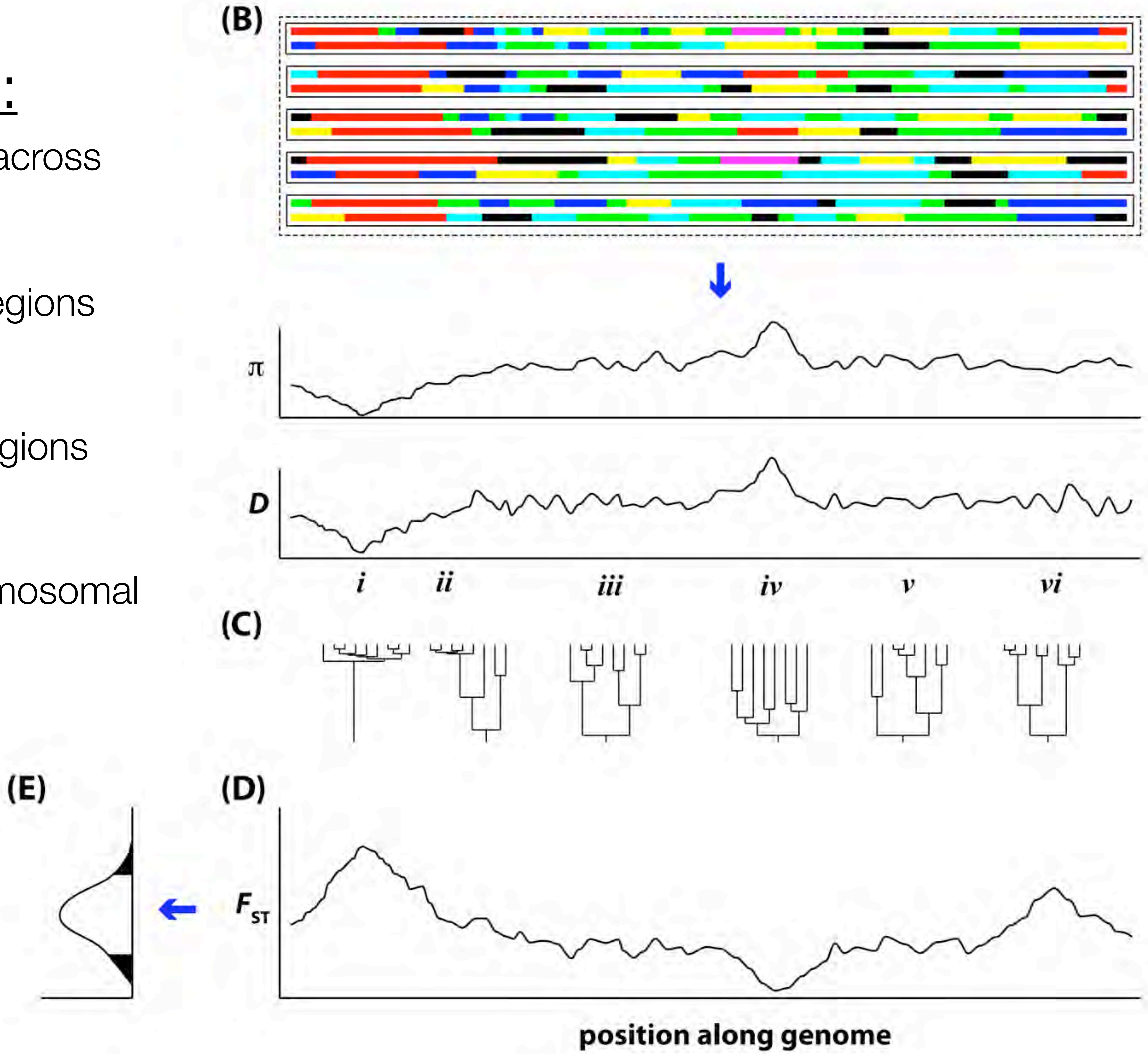
## Genomic architecture:

- Distribution adaptive variation across the genome

- Correlations among genomic regions (linkage disequilibrium)

- Interactions among genomic regions (e.g. epistasis)

- Recombination rates and chromosomal inversions

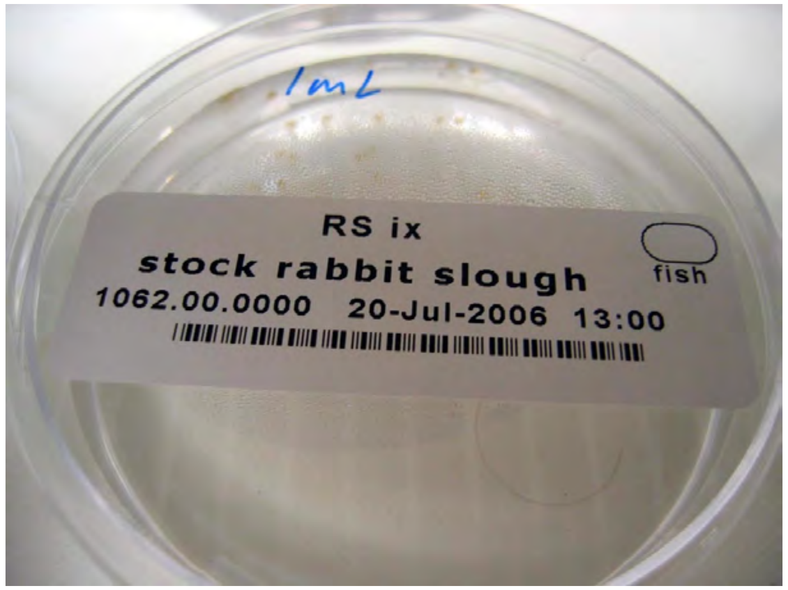


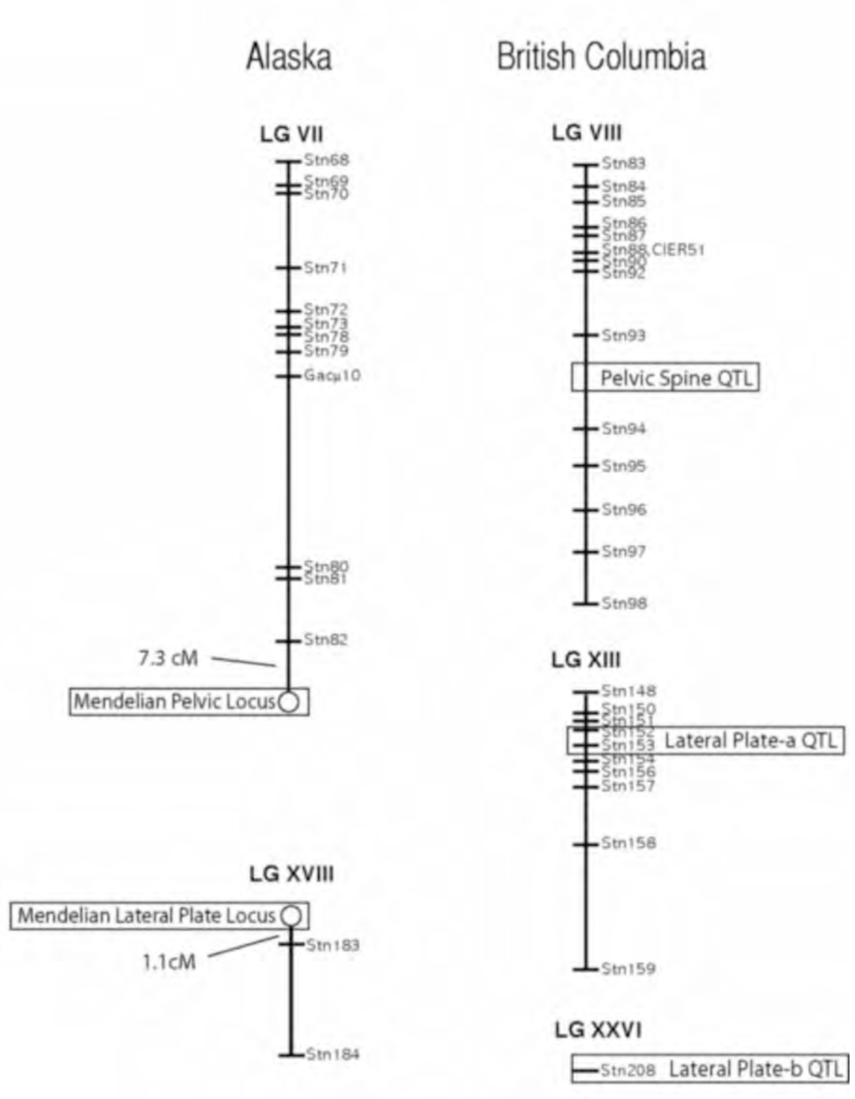
Hohenlohe et al (2010) *Int J Plant Sci 171:1059*

# How do we ‘genomically enable’ research on non-model organisms?

1. Genetic Markers & Maps
2. Physical Maps
3. Transcriptomes
4. Gene Expression Analyses

# In the field and in the lab until a few years ago....





# The open source genomics breakthrough

*Next generation sequencing, high performance computing and new analytical approaches have fundamentally changed the scope of studies of non-model organisms*

# Should we just sequence everything?

# Why not sequence the entire genome??

- Still prohibitively expensive for many studies
    - Human height GWAS; over 15,000 individuals assayed
    - Identified many new regions contributing to the variation
    - Still only identified a fraction of the heritability

- For many studies a full sequence isn’t necessary
    - the genomes of many organisms are organized in linkage blocks
    - well spaced markers will provide the necessary coverage
    - the cost of genotyping will almost always be a fraction of full sequencing

- Genetic maps are very useful in genomic studies
    - a high density genetic map can facilitate genome assembly
    - genomes may be segregating a lot of structural variation

# Alternative approach - Reduced representation NGS for genotyping

- Focus the sequencing on a homologous set of tags spread throughout the genome

- Can lead to the simultaneous identification and typing of single nucleotide polymorphisms (SNPs)

- The cost will always be a fraction of the cost of resequencing the genome
  - i.e. 1% genome coverage will be less than 1% the cost
  - often the coverage is more even than whole genome sequencing

- Can allow thousands of genomes to be assayed in just a few weeks

- WHY NOT - some cases complete genomic sequence is necessary
  - when linkage disequilibrium blocks (LD) are very short
  - Inferring patterns of LD may be easiest with full sequences

# Different flavors of Reduced Representation Library (RRL) Sequencing for genotyping

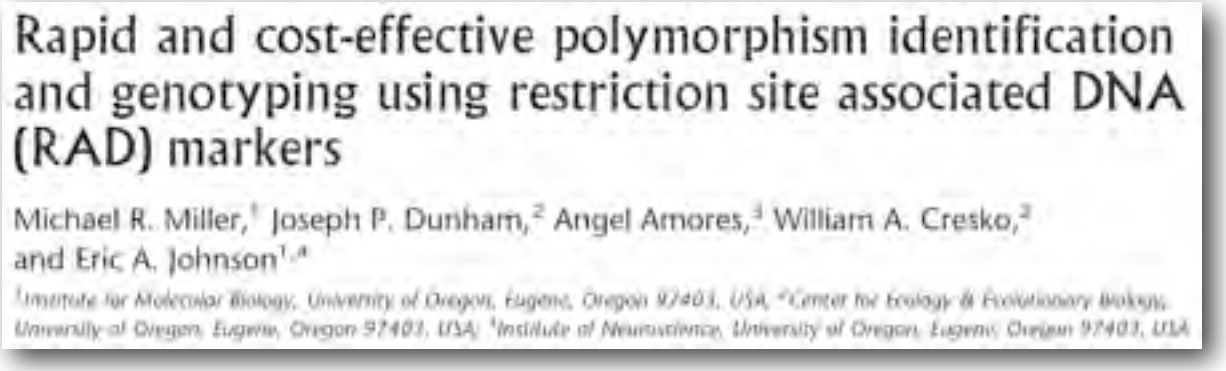
- Common acronyms
    - **RRL** - **R**educed **R**epresentation **L**ibrary
    - **GBS** - **G**enotyping **B**y **S**equencing
    - **CRoPS** - **C**omplexity **R**eduction **o**f **P**olymorphic **S**equences
    - **MSG** - **M**ultiplex **S**hotgun **G**enotyping
    - **RAD** - **R**estriction site **A**ssociated **D**NA
- All rely on restriction enzyme digestion
- RRL, CRoPS, MSG and GBS use one or two restriction enzymes only
- RAD uses an extra shearing step to capture all restriction sites
- Incorporation of barcodes on adaptors for multiplexing
- Aligned against a reference genome or assembled *de novo*
- Statistical issues
    - new level of sampling variation (sequencing in addition to biological)
    - sequencing error and problems for aligning or clustering

# What is RAD-seq?

(Restriction-site Associated DNA)

Illumina

2007

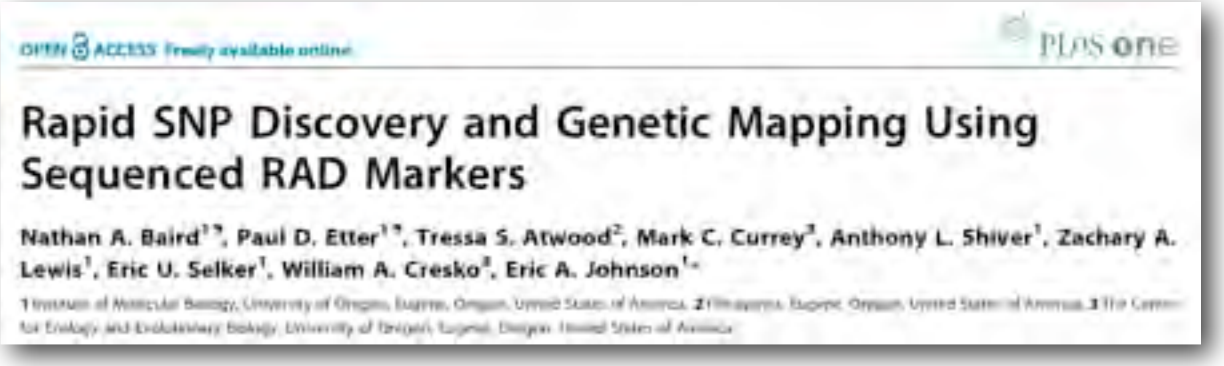


Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers

Michael R. Miller,[1] Joseph P. Dunham,[2] Angel Amores,[3] William A. Cresko,[2] and Eric A. Johnson[1,4]

[1]Institute for Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA; [2]Center for Ecology & Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, USA; [3]Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA

2008



OPEN ACCESS Freely available online

PLoS one

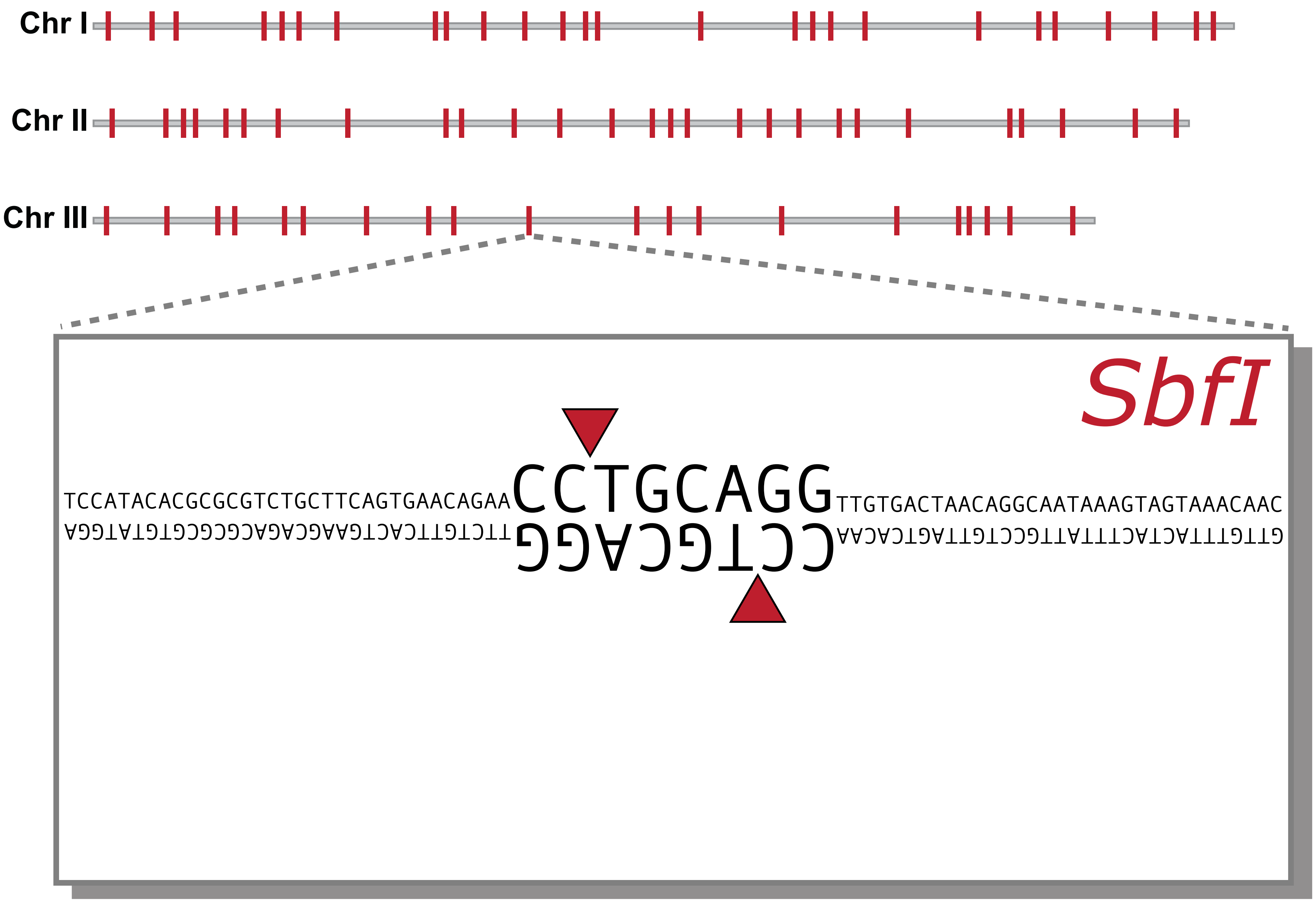
Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

Nathan A. Baird[1*], Paul D. Etter[1*], Tressa S. Atwood[2], Mark C. Currey[3], Anthony L. Shiver[1], Zachary A. Lewis[1], Eric U. Selker[1], William A. Cresko[3], Eric A. Johnson[1*]

1 Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America, 2 Floragenex, Eugene, Oregon, United States of America, 3 The Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon, United States of America

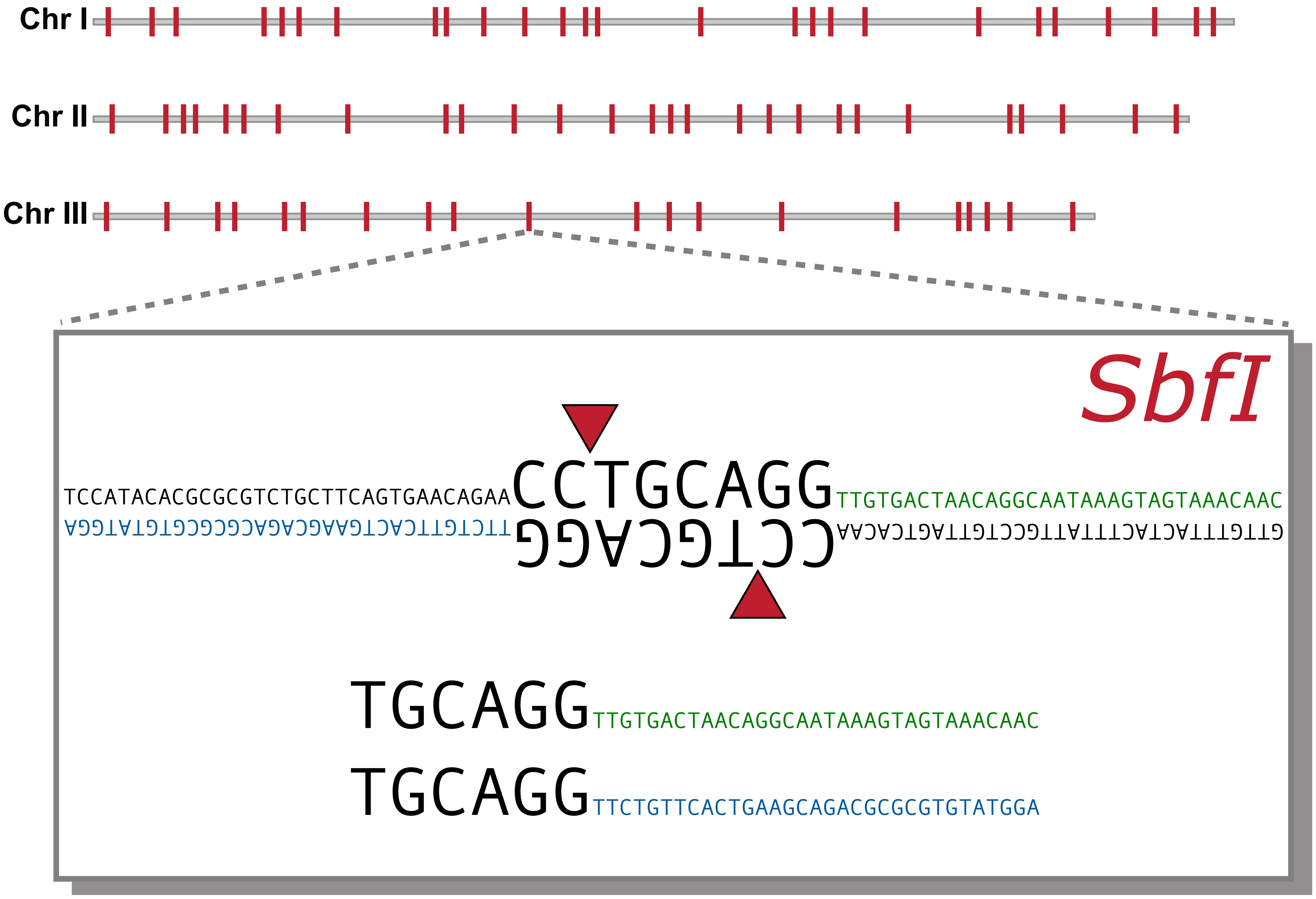
# What is RAD-seq?

(Restriction-site Associated DNA)


Chr I
Chr II
Chr III
SbfI
TCCATACACGCGCGTCTGCTTCAGTGAACAGAA CCTGCAGG TTGTGACTAACAGGCAATAAAGTAGTAAACAAC
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA CCTGCAGG GTTGTTTACTACTTTATTGCCTGTTAGTCACAA

# What is RAD-seq?

(Restriction-site Associated DNA)



TCCATACACGCGCGTCTGCTTCAGTGAACAGAA CCTGCAGG TTGTGACTAACAGGCAATAAAGTAGTAAACAAC

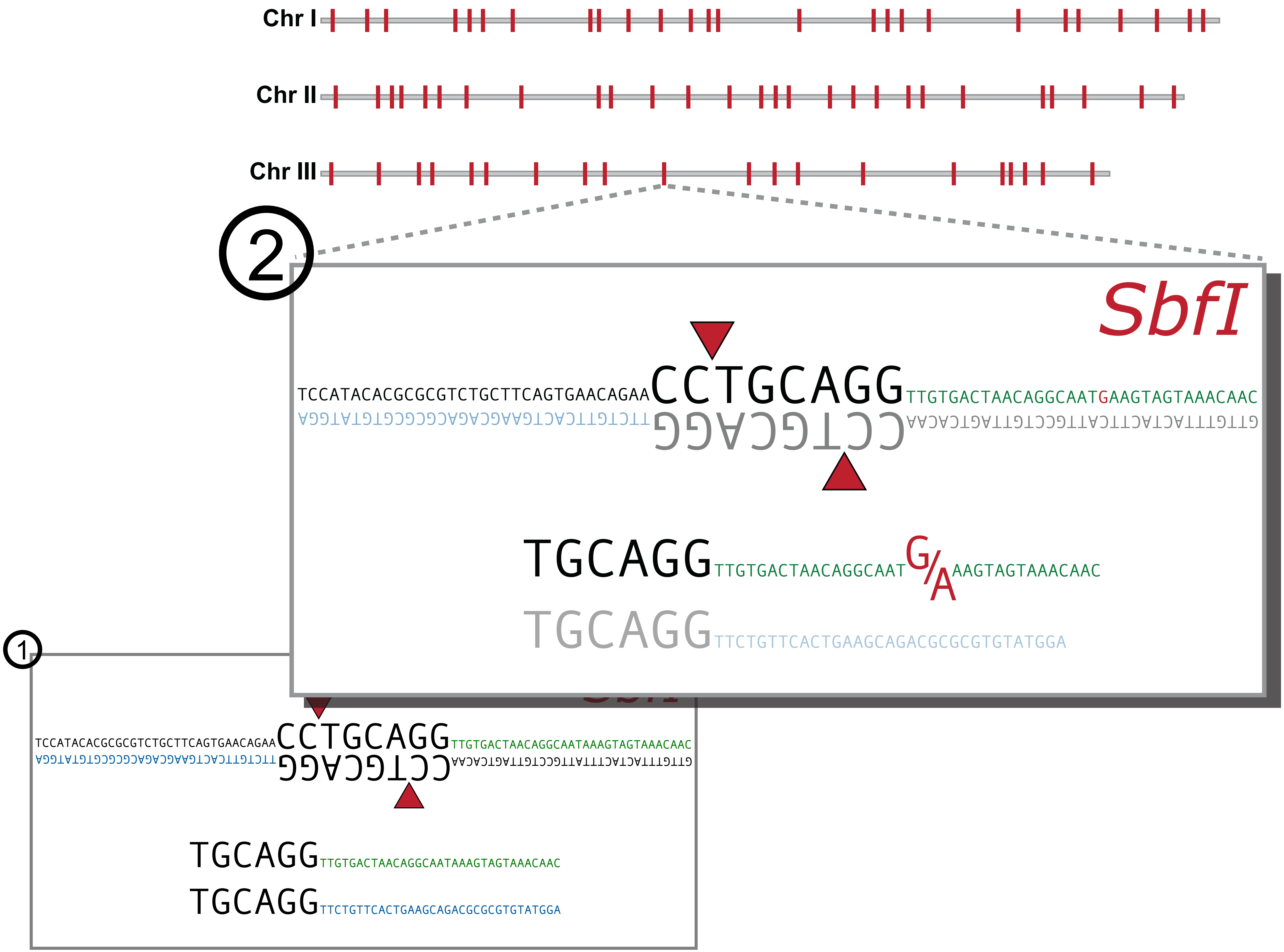
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA CCTGCAGG GTTGTTTACTACTTTATTGCCTGTTAGTCACAA

TGCAGG TTGTGACTAACAGGCAATAAAGTAGTAAACAAC

TGCAGG TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA

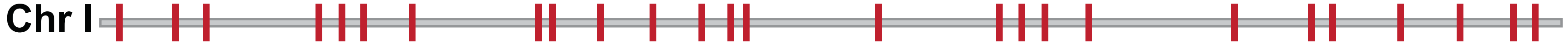
# What is RAD-seq?

(Restriction-site Associated DNA)


Chr I
Chr II
Chr III
2
SbfI
TCCATACACGCGCGTCTGCTTCAGTGAACAGAA
CCTGCAGG
TTGTGACTAACAGGCAATGAAGTAGTAAACAAC
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA
CCTGCAGG
GTTGTTTACTACTTCATTGCCTGTTAGTCACAA
TGCAGG
TTGTGACTAACAGGCAAT
G/A
AAGTAGTAAACAAC
TGCAGG
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA
1
TCCATACACGCGCGTCTGCTTCAGTGAACAGAA
CCTGCAGG
TTGTGACTAACAGGCAATAAAGTAGTAAACAAC
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA
CCTGCAGG
GTTGTTTACTACTTTATTGCCTGTTAGTCACAA
TGCAGG
TTGTGACTAACAGGCAATAAAGTAGTAAACAAC
TGCAGG
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA

# What is RAD-seq?

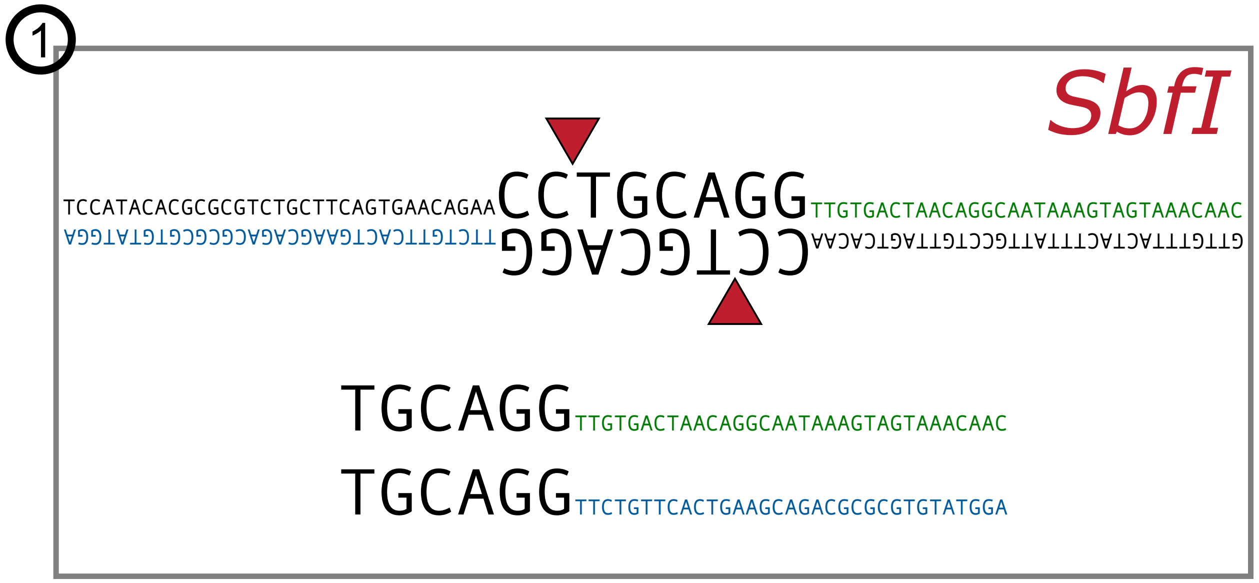
(Restriction-site Associated DNA)

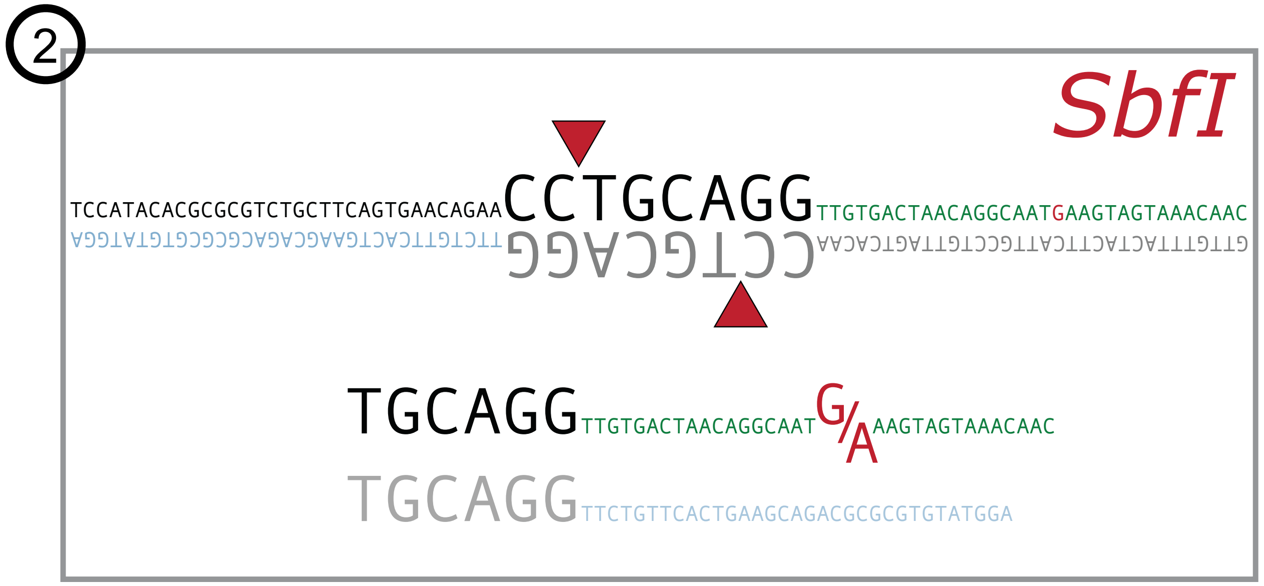


22,830 *SbfI* sites in threespine stickleback

~ 45,000 RAD-Tags

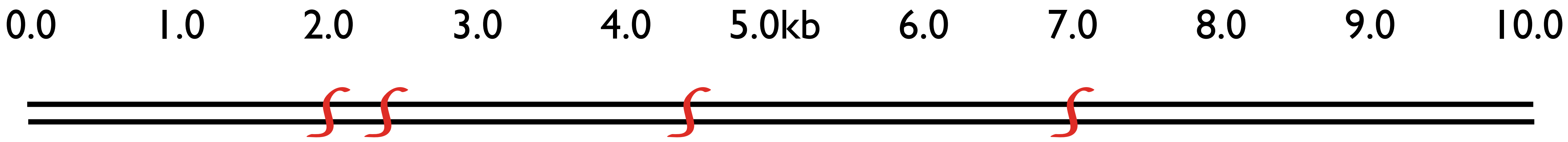
HiSeq Illumina Lane:
160 million reads, 96 barcoded individuals


1
SbfI
TCCATACACGCGCGTCTGCTTCAGTGAACAGAA CCTGCAGG TTGTGACTAACAGGCAATAAAGTAGTAAACAAC
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA CCTGCAGG GTTGTTTACTACTTTATTGCCTGTTAGTCACAA
TGCAGG TTGTGACTAACAGGCAATAAAGTAGTAAACAAC
TGCAGG TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA

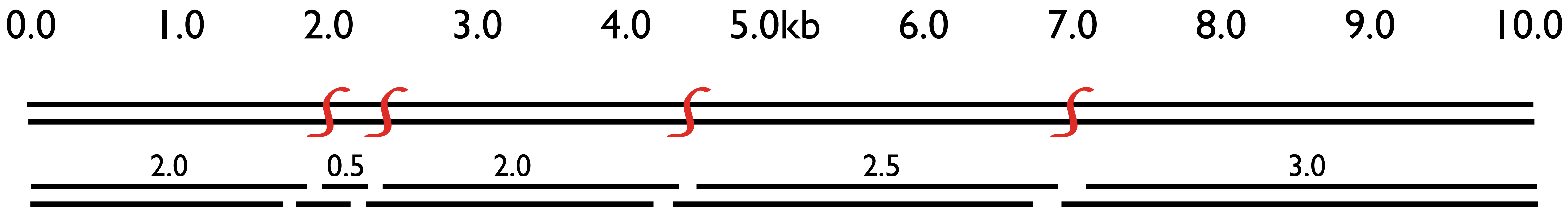

2
SbfI
TCCATACACGCGCGTCTGCTTCAGTGAACAGAA CCTGCAGG TTGTGACTAACAGGCAATGAAGTAGTAAACAAC
TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA CCTGCAGG GTTGTTTACTACTTCATTGCCTGTTAGTCACAA
TGCAGG TTGTGACTAACAGGCAAT G/A AAGTAGTAAACAAC
TGCAGG TTCTGTTCACTGAAGCAGACGCGCGTGTATGGA

# Restriction Enzyme (RE) digestion and first adaptor ligation

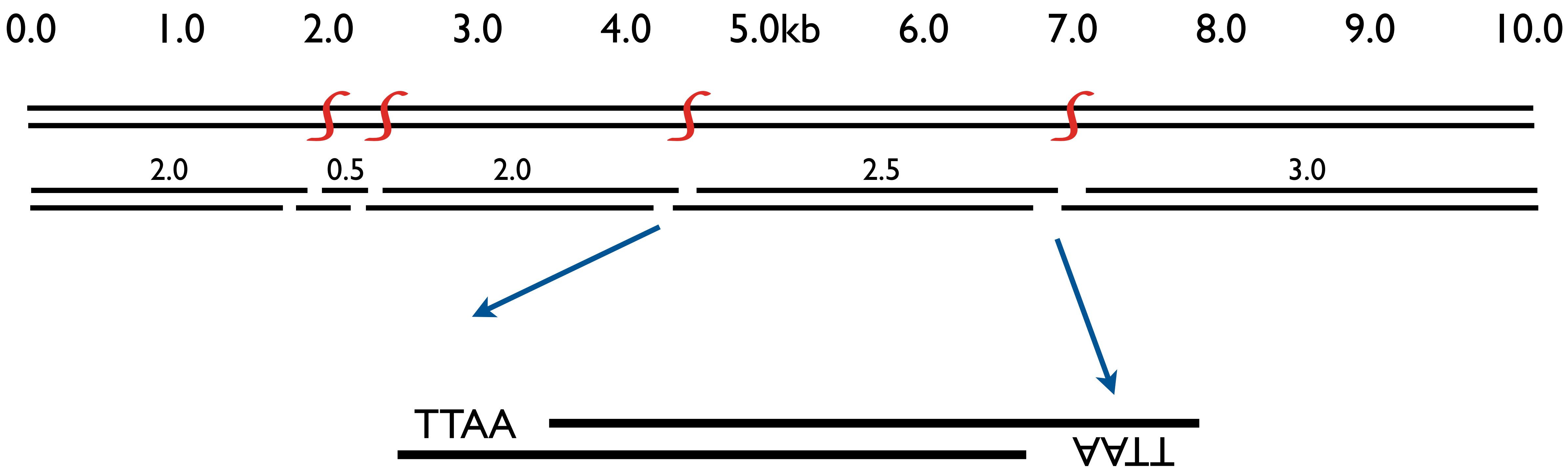
# Restriction Enzyme (RE) digestion and first adaptor ligation



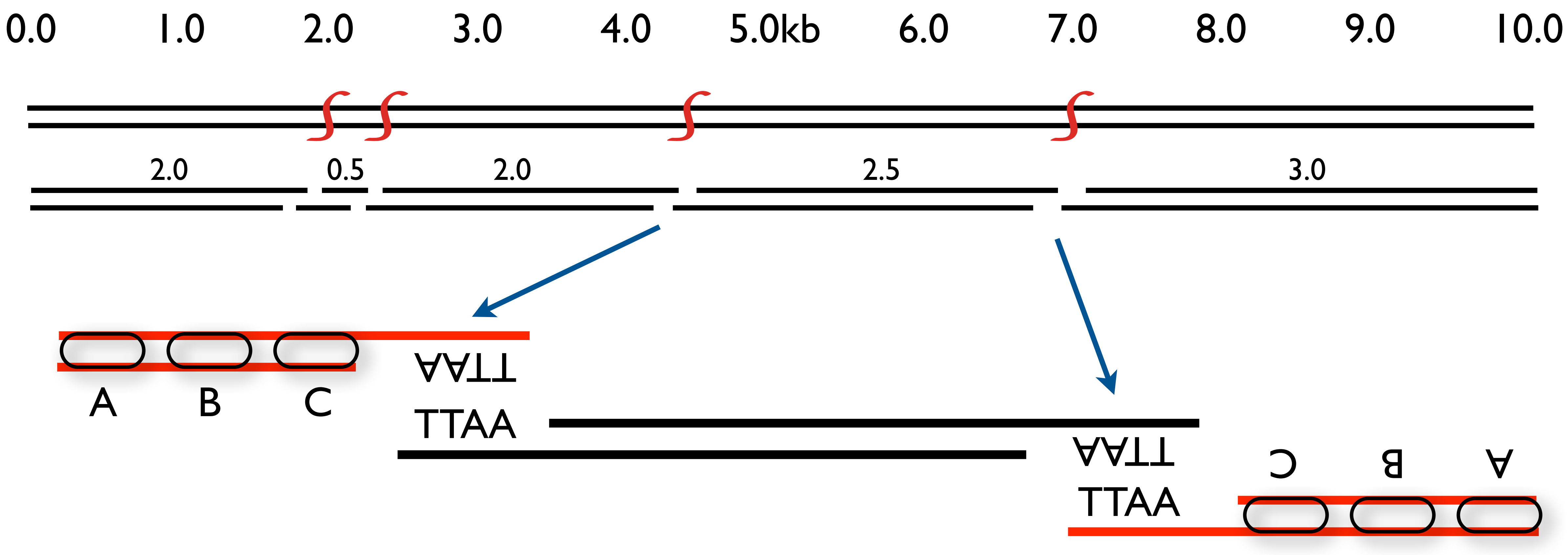
# Restriction Enzyme (RE) digestion and first adaptor ligation



# Restriction Enzyme (RE) digestion and first adaptor ligation

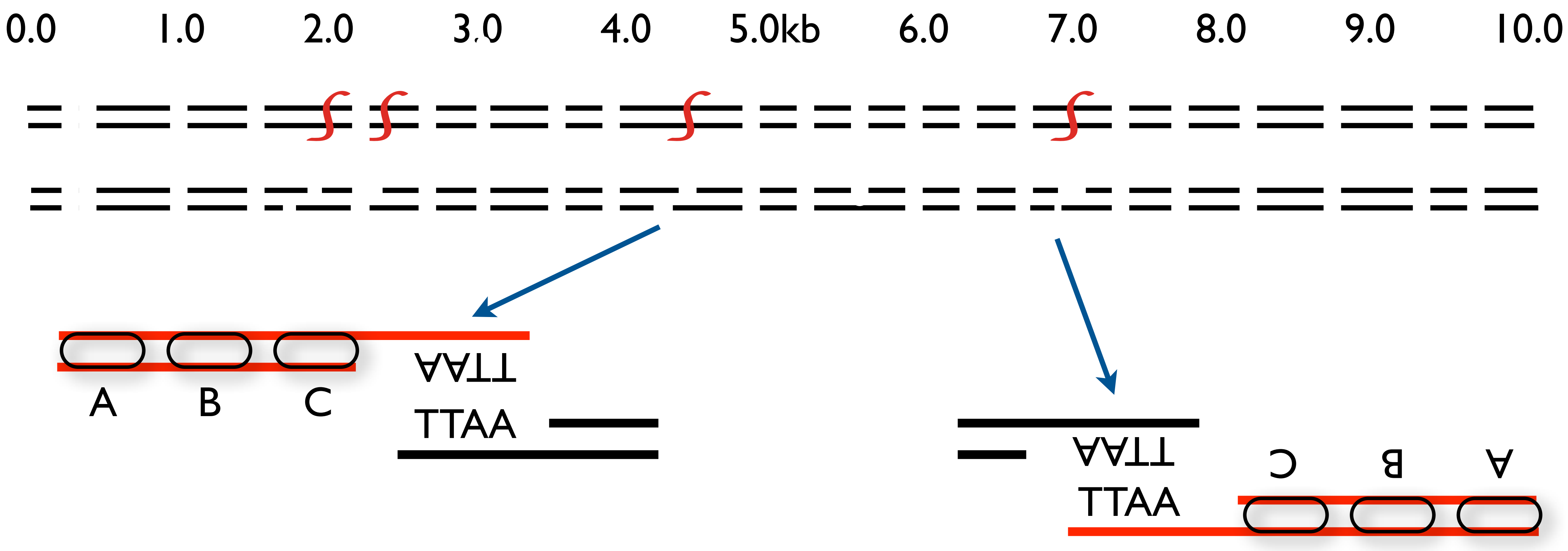


# Restriction Enzyme (RE) digestion and first adaptor ligation



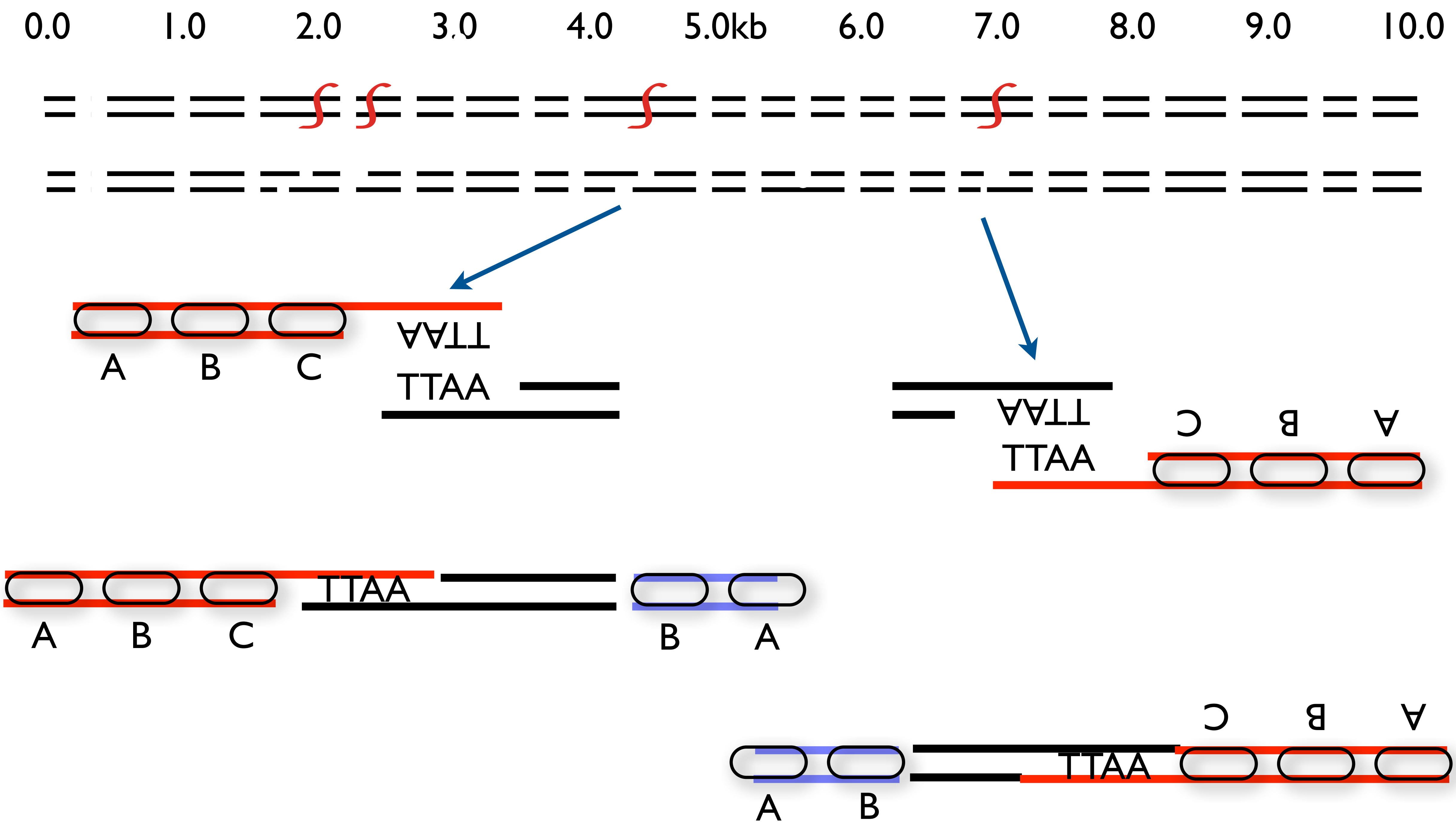
A = Amplification primer
B = Sequencing primer
C = Barcode

# Shearing and second adaptor ligation



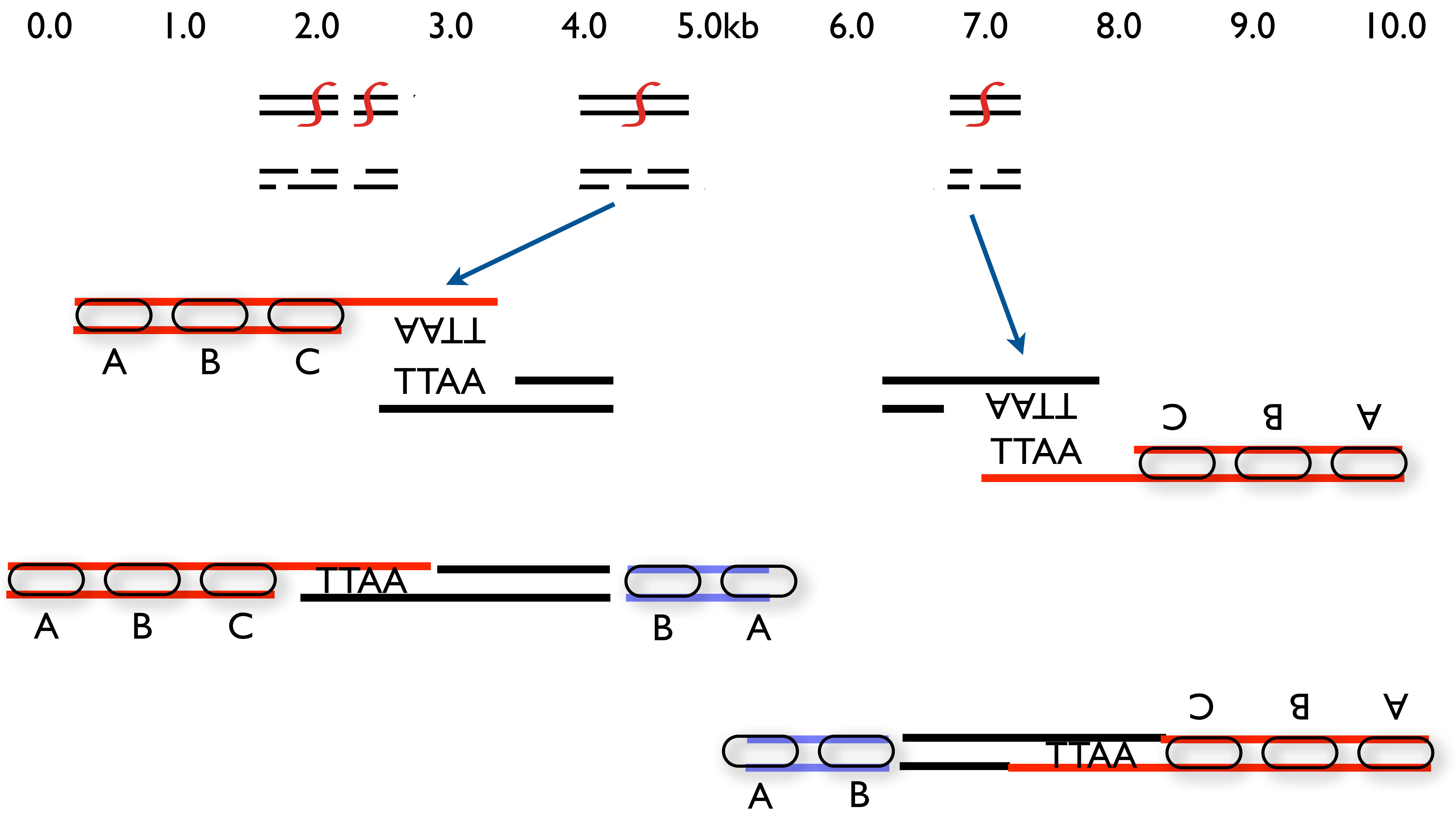
A = Amplification primer
B = Sequencing primer
C = Barcode

# Shearing and second adaptor ligation



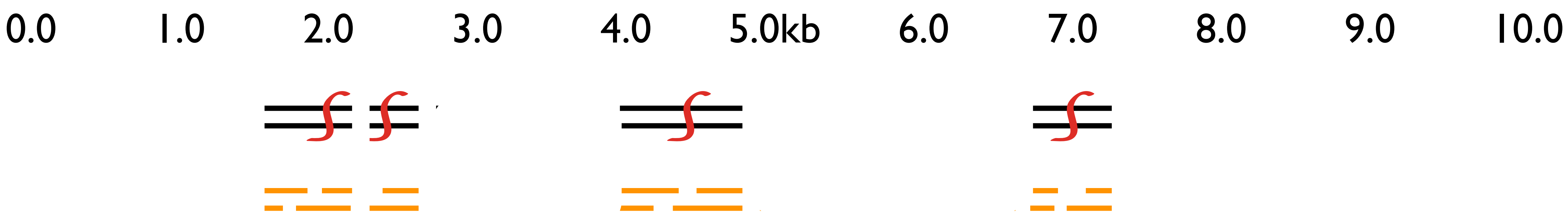
A = Amplification primer
B = Sequencing primer
C = Barcode

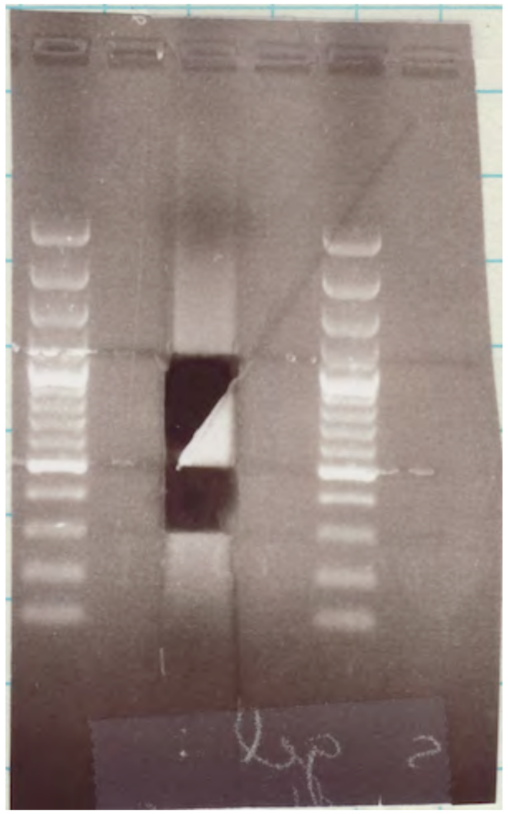
# Shearing and second adaptor ligation

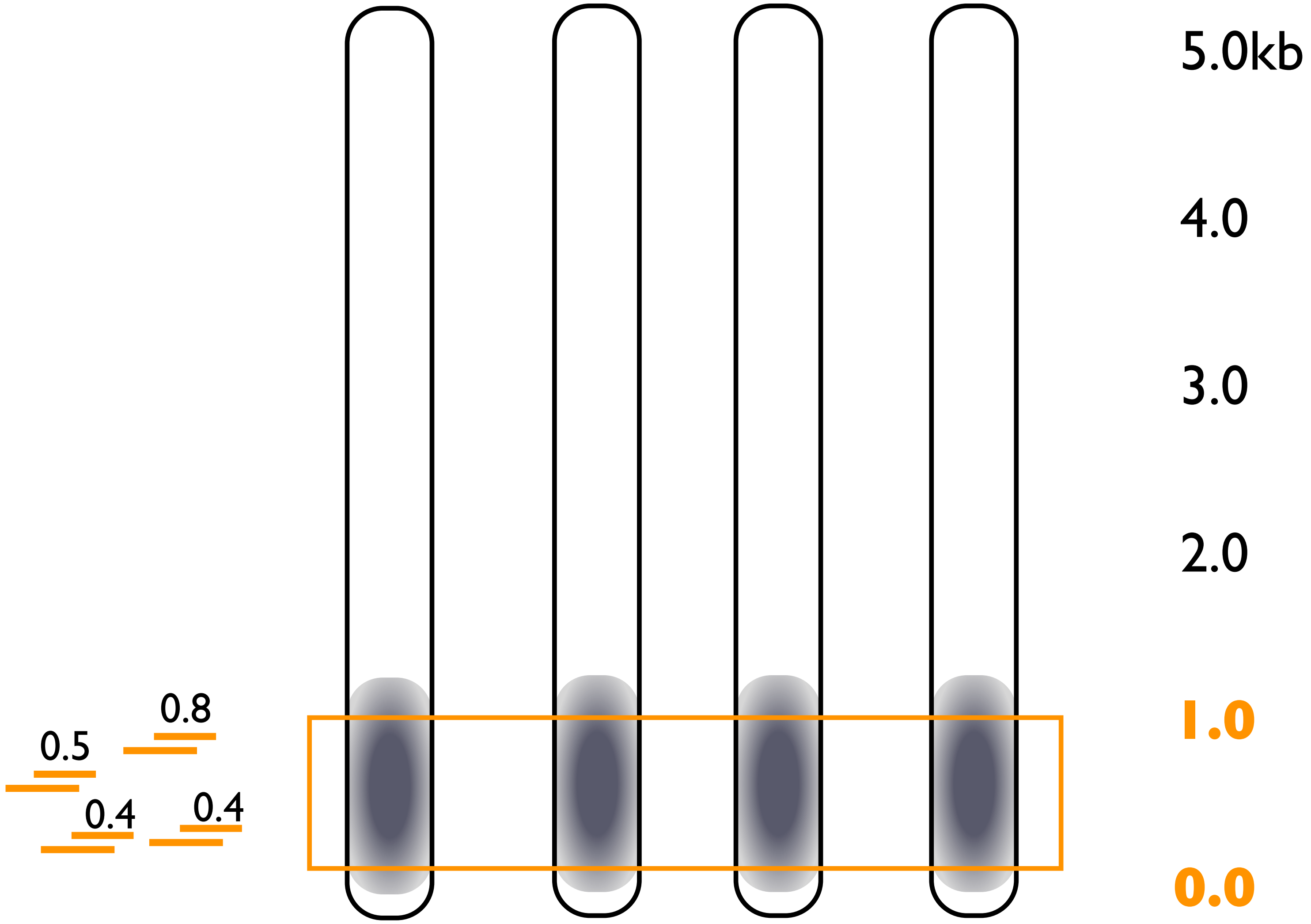


A = Amplification primer
B = Sequencing primer
C = Barcode

# Shearing makes consistent fragments for sequencing

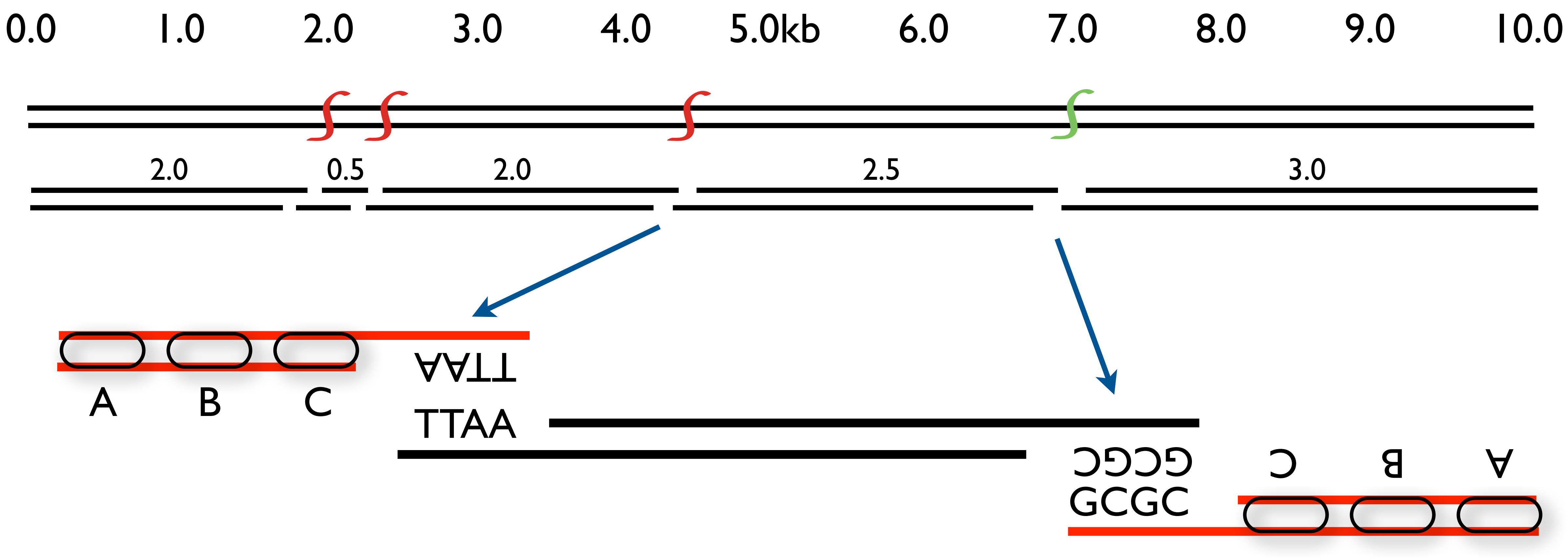






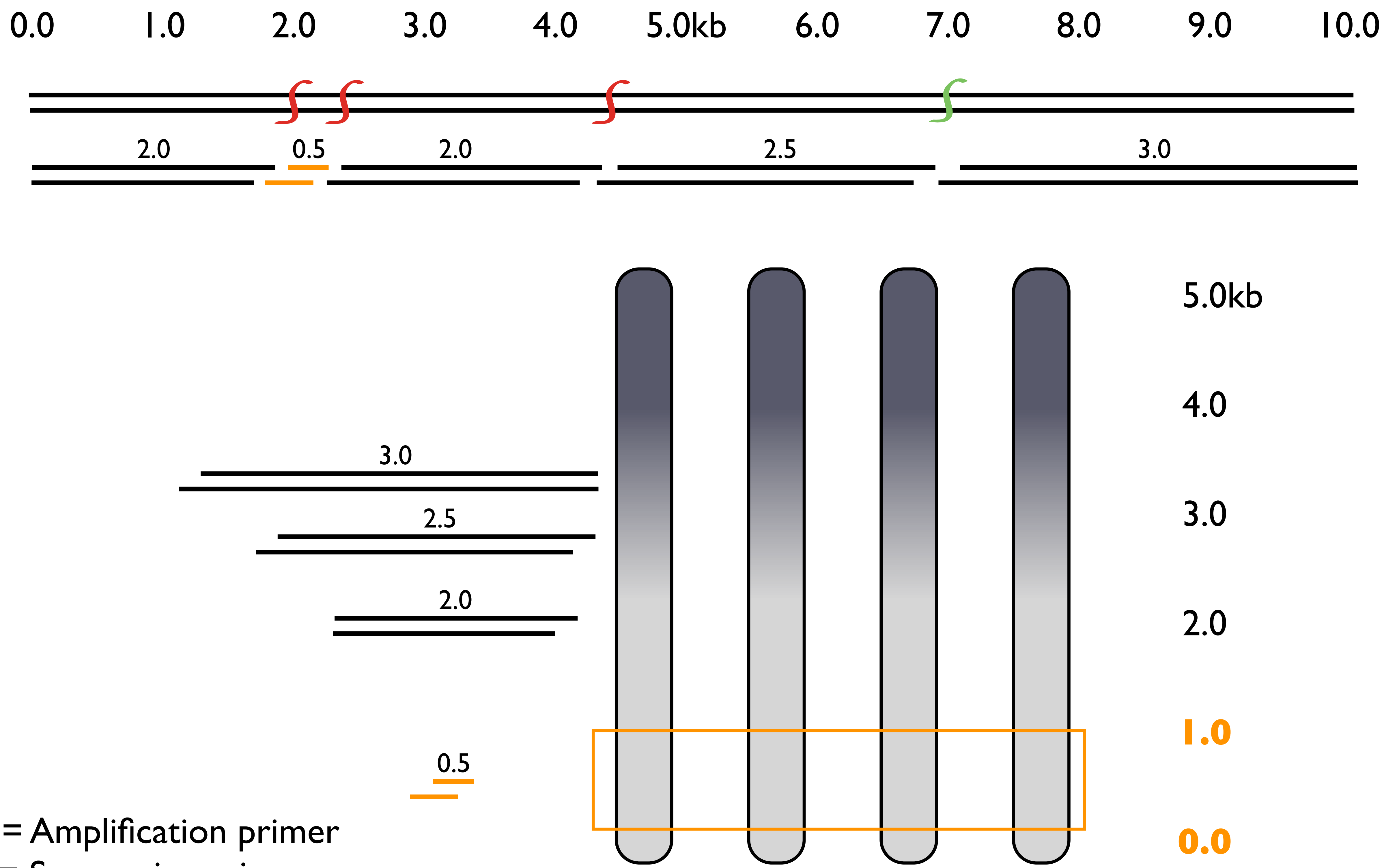
A = Amplification primer
B = Sequencing primer
C = Barcode

# Single (GBS) or Double Digest RAD (ddRAD)



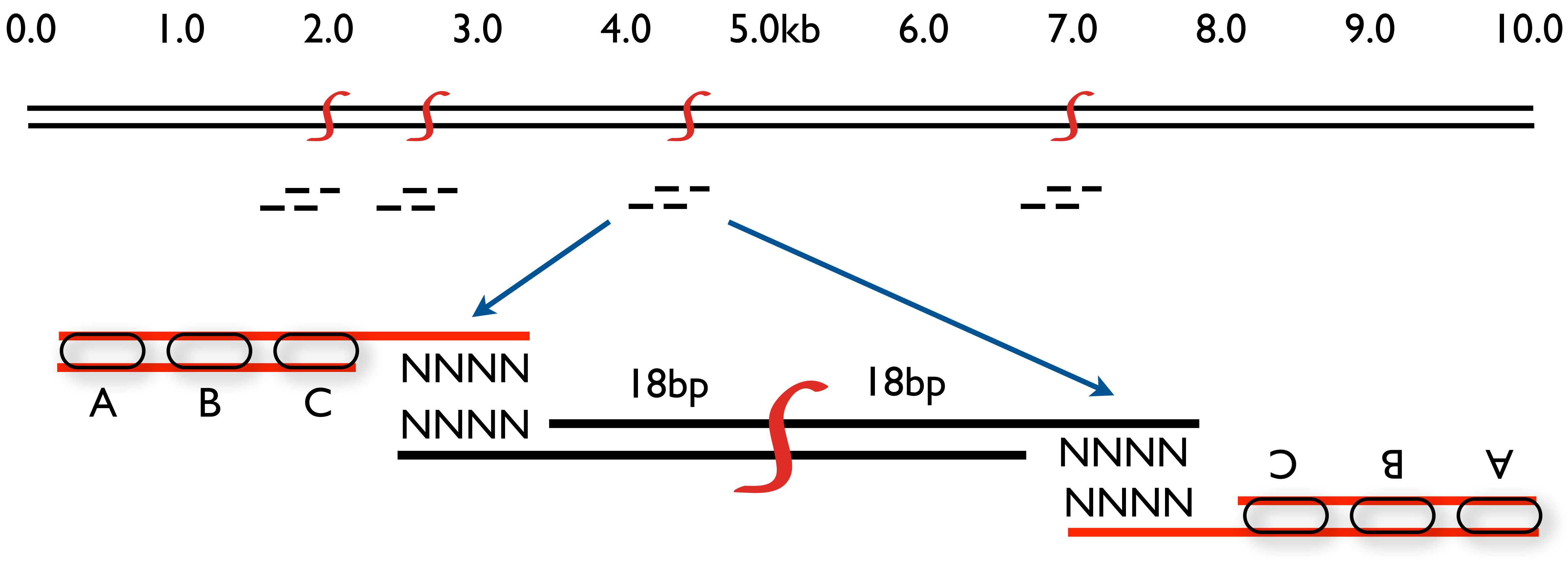
A = Amplification primer
B = Sequencing primer
C = Barcode

# Size selection is more problematic without shearing



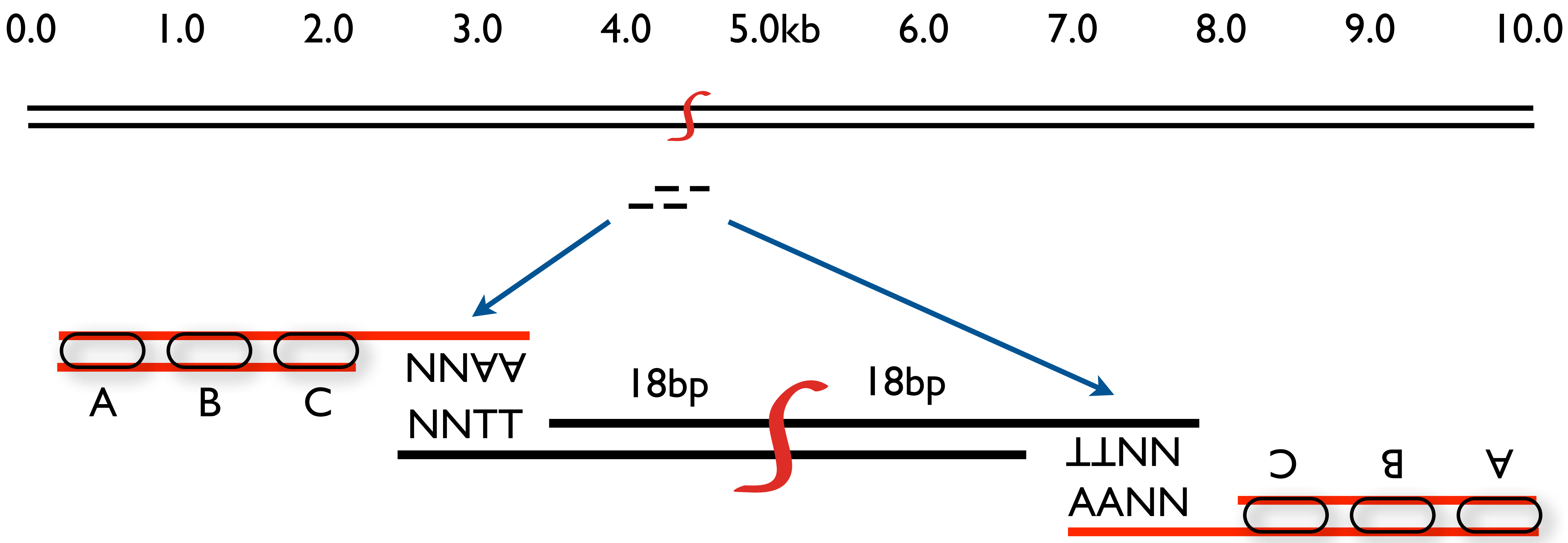
A = Amplification primer
B = Sequencing primer
C = Barcode

# 2bRAD - type 2b restriction enzyme



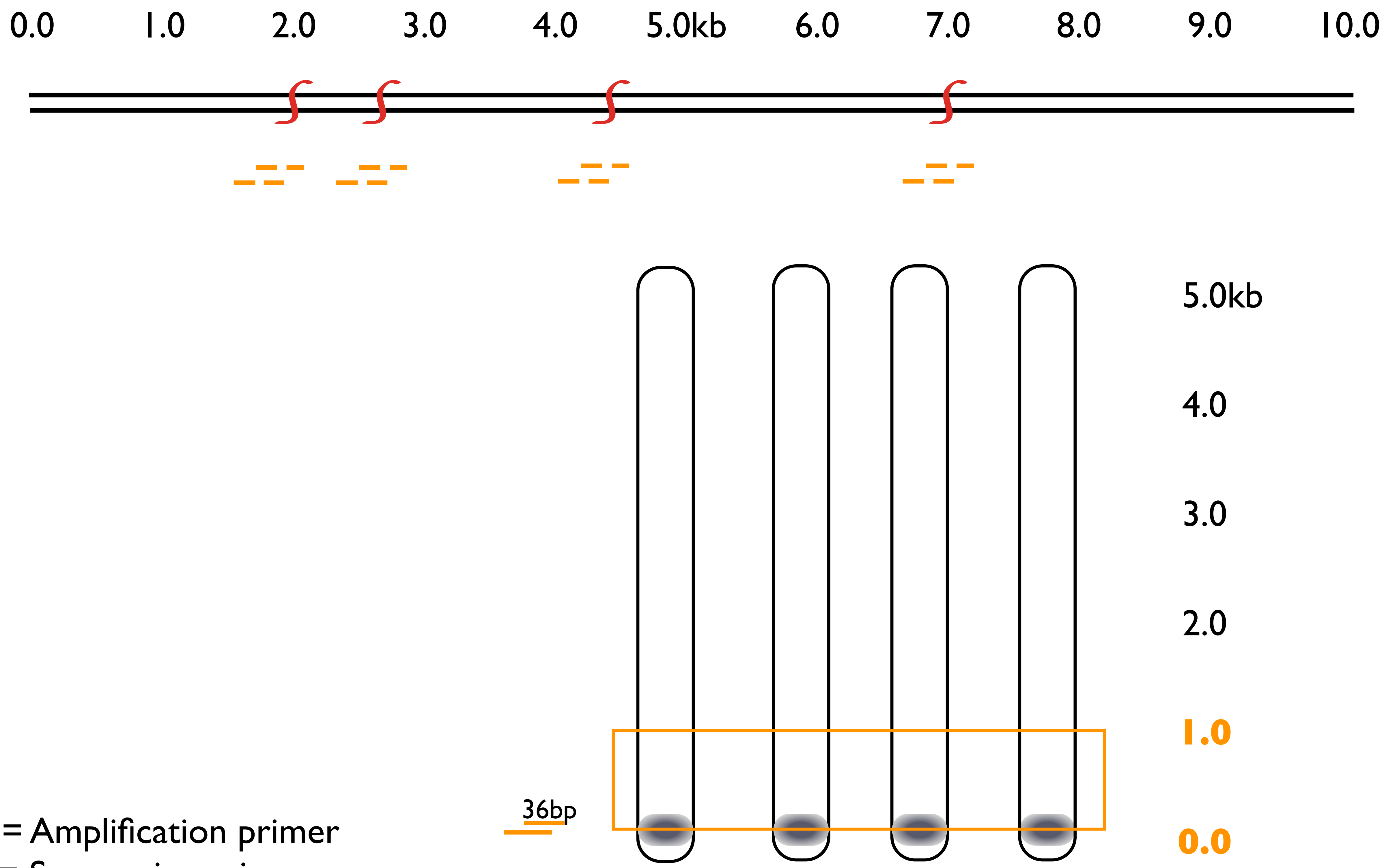
A = Amplification primer
B = Sequencing primer
C = Barcode

# 2bRAD - can scale number of markers easily



A = Amplification primer
B = Sequencing primer
C = Barcode

# 2bRAD - size selection is difficult

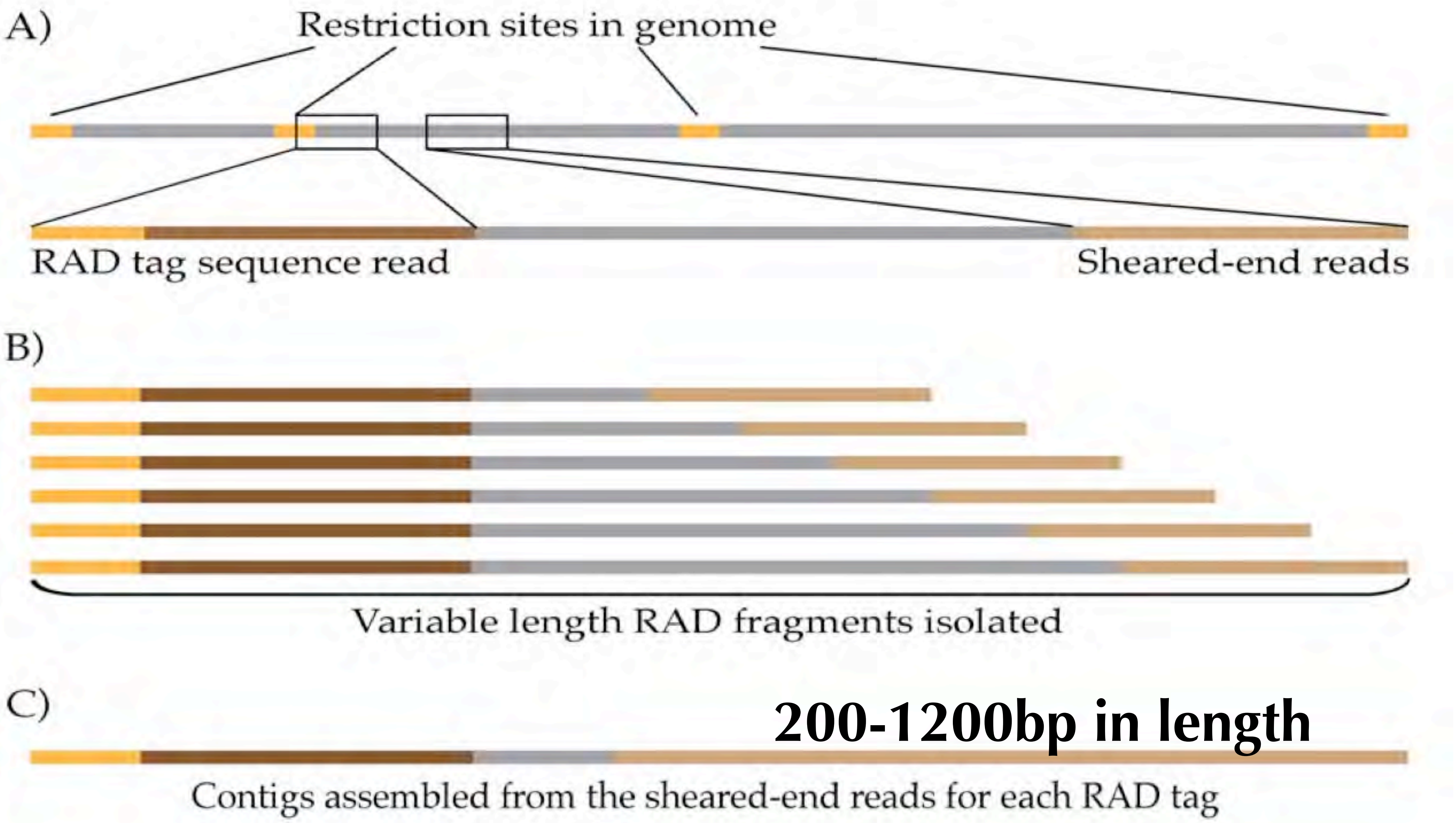


A = Amplification primer
B = Sequencing primer
C = Barcode

# Summary of plusses and minuses of RAD family

| | Sheared RAD | Single or ddRAD | 2b-RAD |
|---|---|---|---|
| plusses | - Consistent reads<br>- Local assemblies<br>- Identify PCR duplicates | - Fewer steps<br>- Easier marker scaling | - Fewest steps<br>- Easiest marker scaling |
| minuses | - Shearing step<br>- Scaling requires different enzymes | - Multiple enzymes<br>- Poor consistency<br>- PCR duplicates | - Very short reads<br>- PCR duplicates |
| | | | |

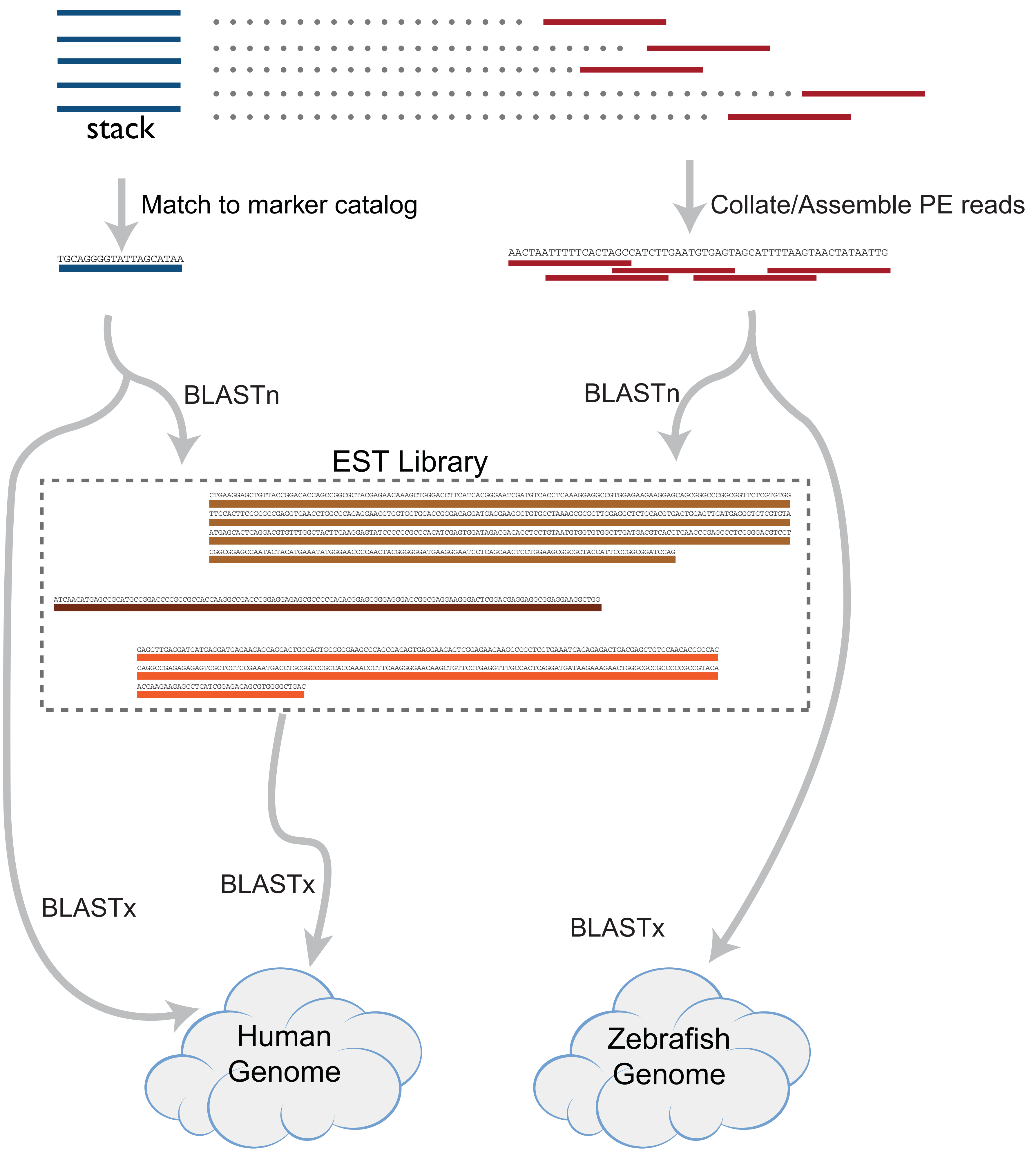
# Benefits of random shearing in RAD


A)
Restriction sites in genome
RAD tag sequence read
Sheared-end reads
B)
Variable length RAD fragments isolated
C)
200-1200bp in length
Contigs assembled from the sheared-end reads for each RAD tag

Acquire paired-end sequence

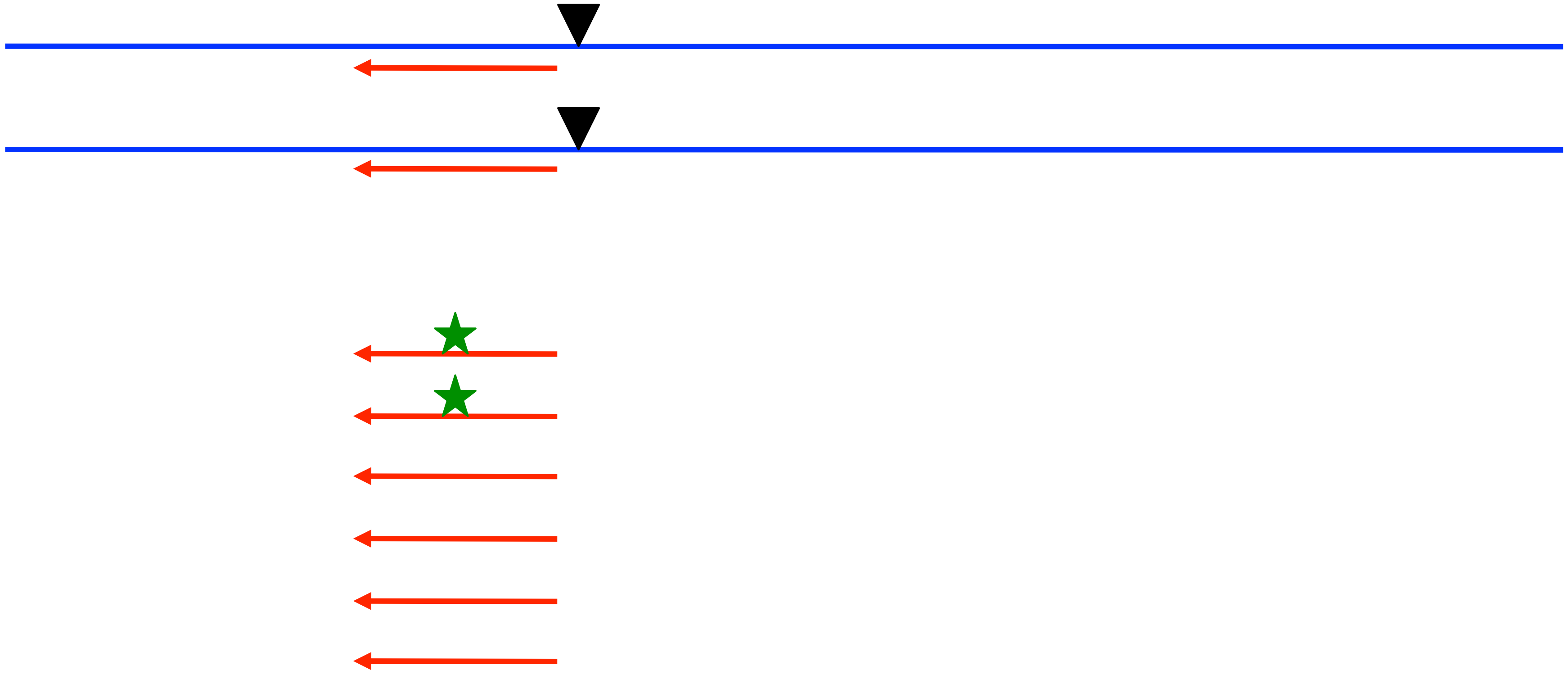
Associate markers / PE contigs with ESTs

Assign orthology to:
markers
PE contigs
ESTs


stack
Match to marker catalog
Collate/Assemble PE reads
TGCAGGGGTATTAGCATAA
AACTAATTTTTCACTAGCCATCTTGAATGTGAGTAGCATTTTAAGTAACTATAATTG
BLASTn
BLASTn
EST Library
CTGAAGGAGCTGTTACCGGACACCAGCCGGCGCTACGAGAACAAAGCTGGGACCTTCATCACGGGAATCGATGTCACCTCAAAGGAGGCCGTGGAGAAGAAGGAGCAGCGGGCCCGGCGGTTCTCGTGTGG
TTCCACTTCCGCGCCGAGGTCAACCTGGCCCAGAGGAACGTGGTGCTGGACCGGGACAGGATGAGGAAGGCTGTGCCTAAAGCGCGCTTGGAGGCTCTGCACGTGACTGGAGTTGATGAGGGTGTCGTGTA
ATGAGCACTCAGGACGTGTTTGGCTACTTCAAGGAGTATCCCCCCGCCCACATCGAGTGGATAGACGACACCTCCTGTAATGTGGTGTGGCTTGATGACGTCACCTCAACCCGAGCCCTCCGGGACGTCCT
CGGCGGAGCCAATACTACATGAAATATGGGAACCCCAACTACGGGGGGATGAAGGGAATCCTCAGCAACTCCTGGAAGCGGCGCTACCATTCCCGGCGGATCCAG
ATCAACATGAGCCGCATGCCGGACCCCGCCGCCACCAAGGCCGACCCGGAGGAGAGCGCCCCCACACGGAGCGGGAGGGACCGGCGAGGAAGGGACTCGGACGAGGAGGCGGAGGAAGGCTGG
GAGGTTGAGGATGATGAGGATGAGAAGAGCAGCACTGGCAGTGCGGGGAAGCCCAGCGACAGTGAGGAAGAGTCGGAGAAGAAGCCCGCTCCTGAAATCACAGAGACTGACGAGCTGTCCAACACCGCCAC
CAGGCCGAGAGAGAGTCGCTCCTCCGAAATGACCTGCGGCCCGCCACCAAACCCTTCAAGGGGAACAAGCTGTTCCTGAGGTTTGCCACTCAGGATGATAAGAAAGAACTGGGCGCCGCCCCCGCCGTACA
ACCAAGAAGAGCCTCATCGGAGACAGCGTGGGGCTGAC
BLASTx
BLASTx
BLASTx
Human Genome
Zebrafish Genome

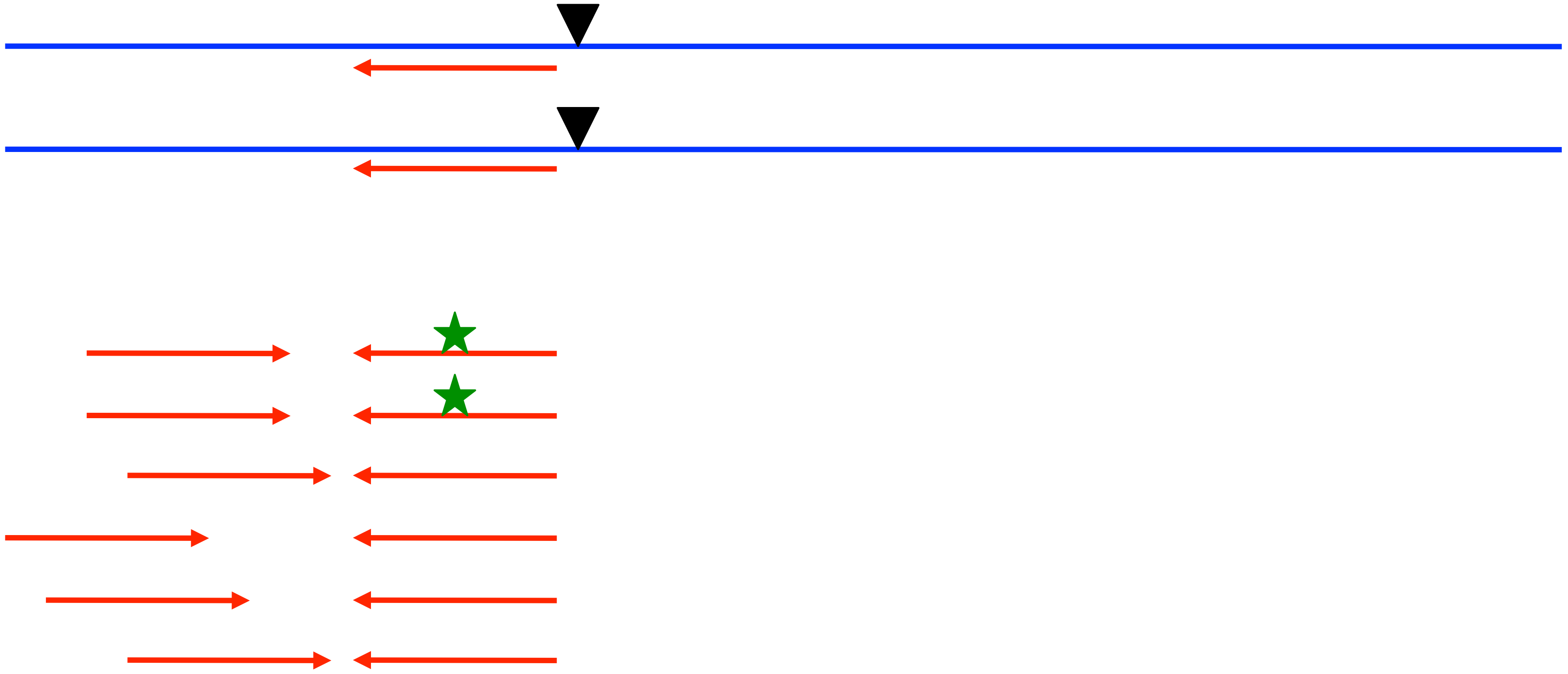
# Random shearing benefits in RAD

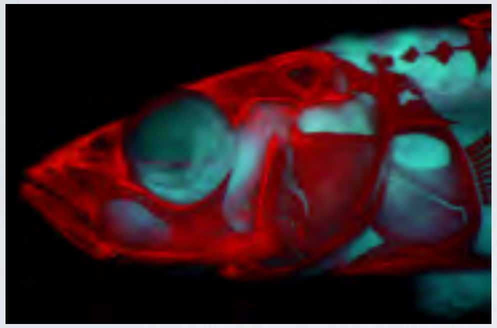
Eliminating PCR duplicates:

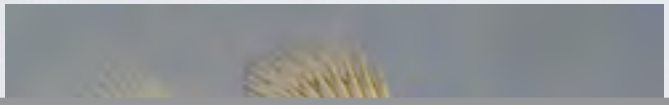


# Random shearing benefits in RAD

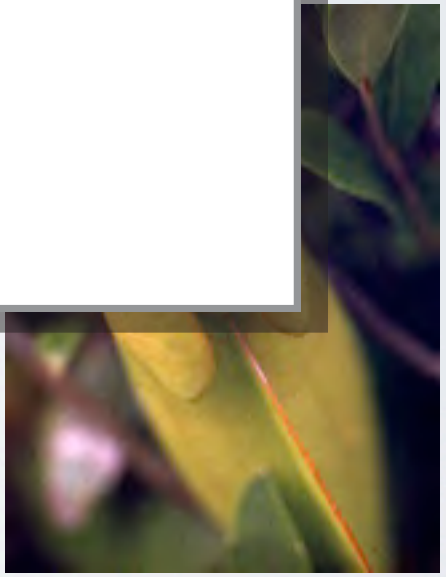
Eliminating PCR duplicates:







# Considerations for RAD-seq studies



# Experimental design considerations for RAD

*Tradeoffs***:**

**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

# Experimental design considerations for RAD

*Tradeoffs***:**

**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

# Experimental design considerations for RAD

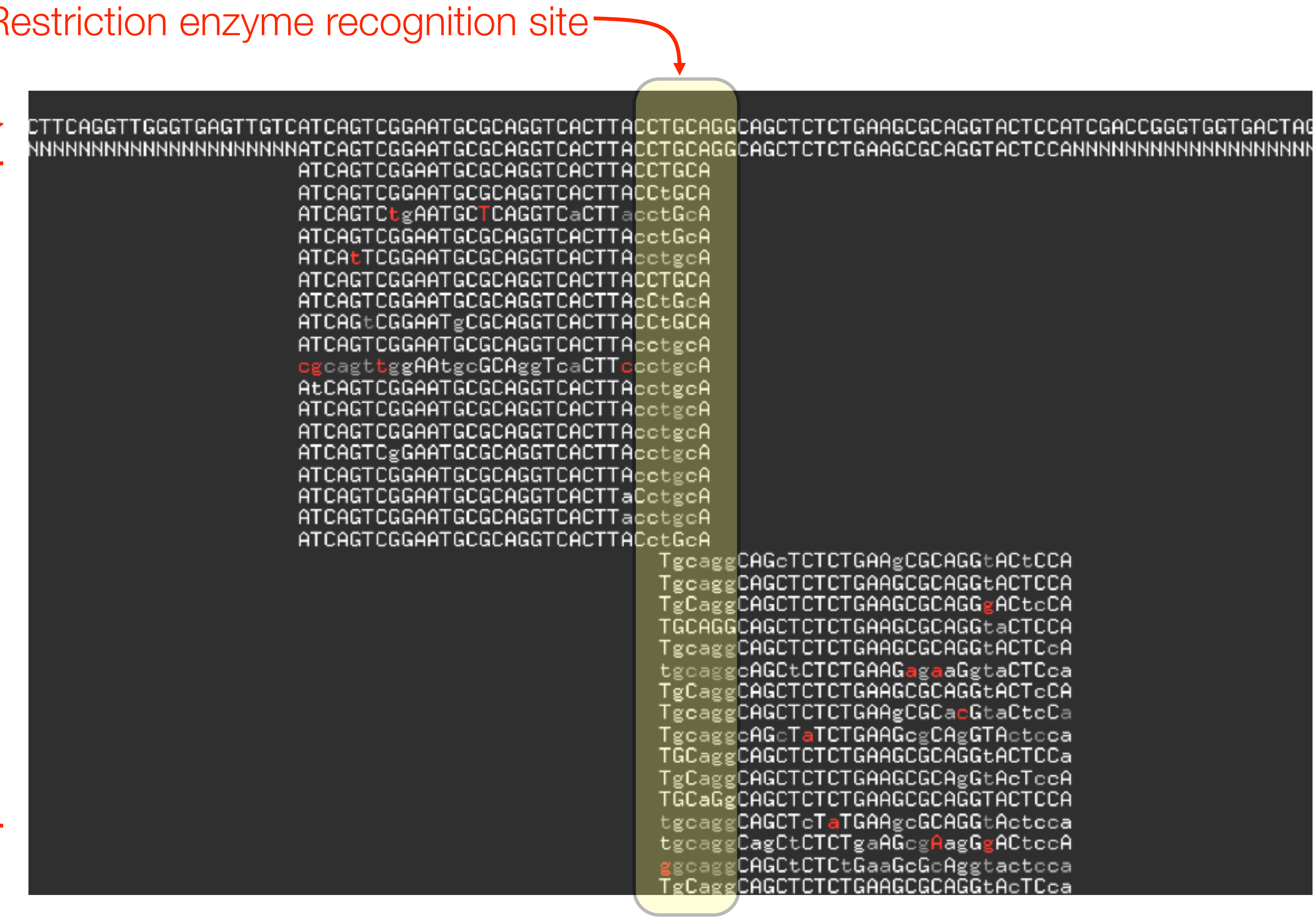
*Tradeoffs*:
**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

raw reads / samples / sites = coverage at each RAD locus

1,000,000 / 100 / 1,000 = 10x coverage

25 to 50x average coverage per RAD locus is a good goal

# Differentiating SNPs from error



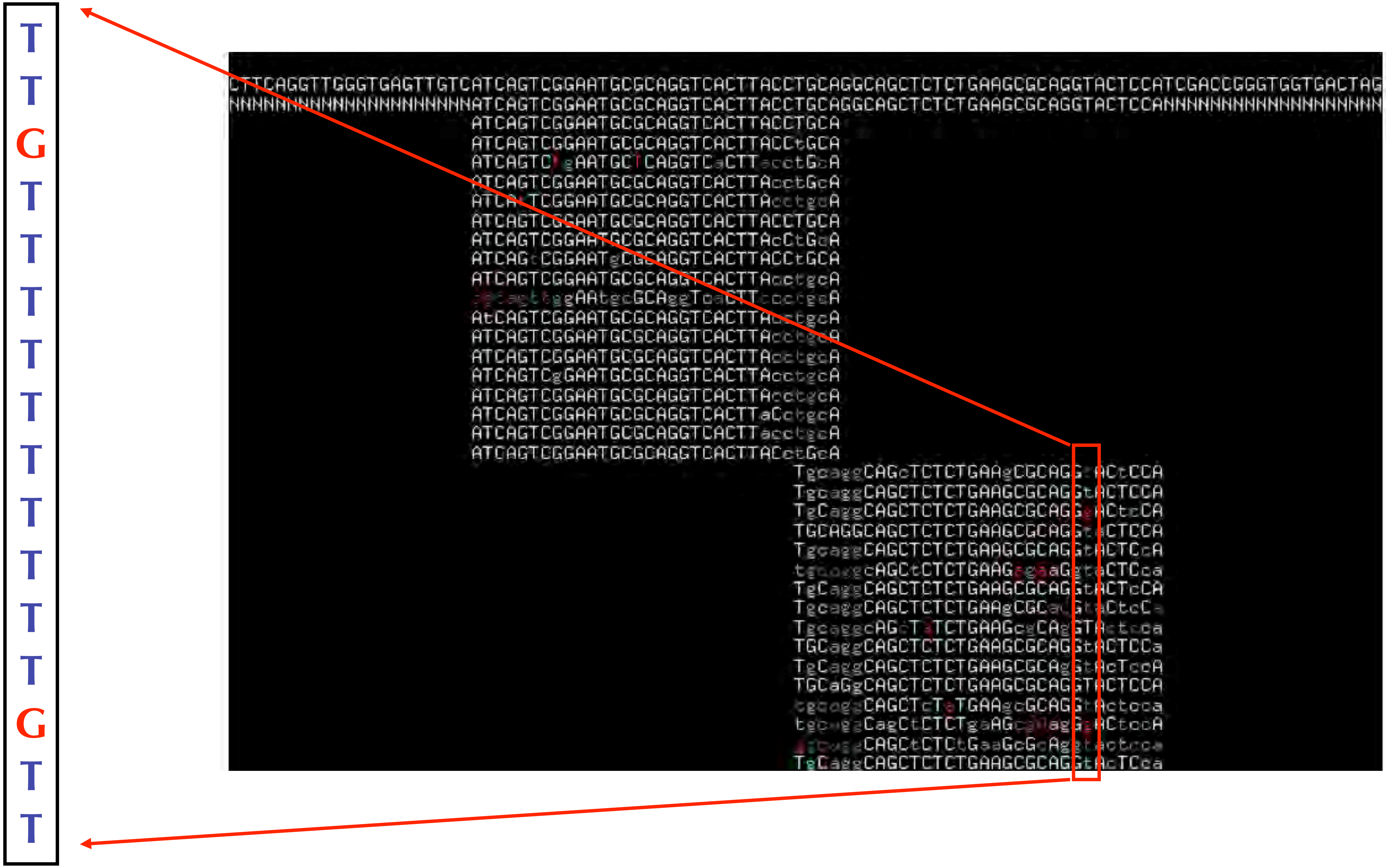
Restriction enzyme recognition site

Reference genome sequence

sequence reads

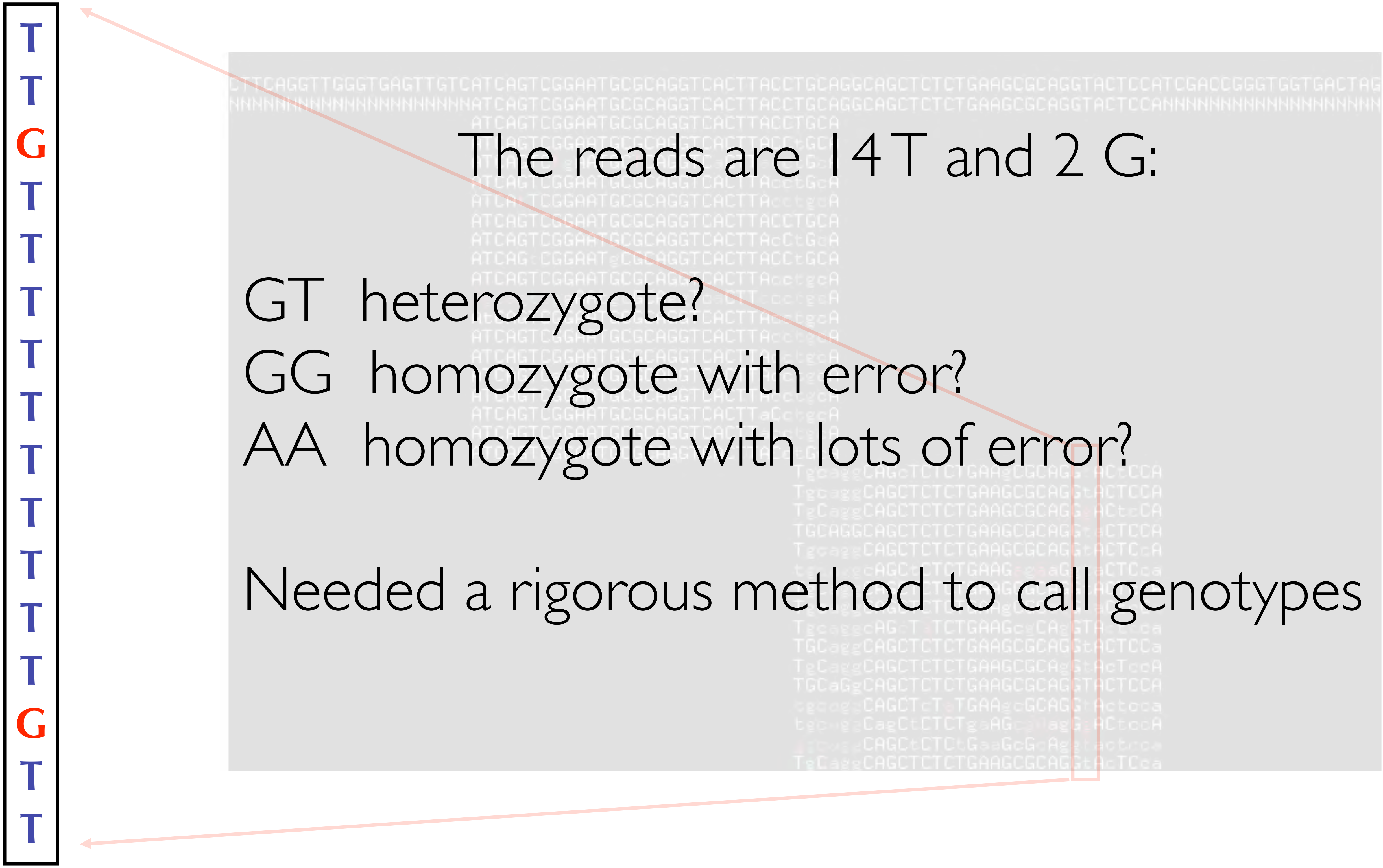
```
CTTCAGGTTGGGTGAGTTGTCATCAGTCGGAATGCGCAGGTCACTTACCTGCAGGCAGCTCTCTGAAGCGCAGGTACTCCATCGACCGGGTGGTGACTAG
NNNNNNNNNNNNNNNNNNNNNATCAGTCGGAATGCGCAGGTCACTTACCTGCAGGCAGCTCTCTGAAGCGCAGGTACTCCANNNNNNNNNNNNNNNNNNN
                     ATCAGTCGGAATGCGCAGGTCACTTACCTGCA
                     ATCAGTCGGAATGCGCAGGTCACTTACCtGCA
                     ATCAGTCtgAATGCTCAGGTCaCTTacctGcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctGcA
                     ATCAtTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTACCTGCA
                     ATCAGTCGGAATGCGCAGGTCACTTAcCtGcA
                     ATCAGtCGGAATgCGCAGGTCACTTACCtGCA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     cgcagttggAAtgcGCAggTcaCTTccctgcA
                     AtCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCgGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTaCctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTacctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTACctGcA
                                                 TgcaggCAGcTCTCTGAAgCGCAGGtACtCCA
                                                 TgcaggCAGCTCTCTGAAGCGCAGGtACTCCA
                                                 TgCaggCAGCTCTCTGAAGCGCAGGgACtcCA
                                                 TGCAGGCAGCTCTCTGAAGCGCAGGtaCTCCA
                                                 TgcaggCAGCTCTCTGAAGCGCAGGtACTCcA
                                                 tgcaggcAGCtCTCTGAAGagaaGgtaCTCca
                                                 TgCaggCAGCTCTCTGAAGCGCAGGtACTcCA
                                                 TgcaggCAGCTCTCTGAAgCGCacGtaCtcCa
                                                 TgcaggcAGcTaTCTGAAGcgCAgGTActcca
                                                 TGCaggCAGCTCTCTGAAGCGCAGGtACTCCa
                                                 TgCaggCAGCTCTCTGAAGCGCAgGtAcTccA
                                                 TGCaGgCAGCTCTCTGAAGCGCAGGTACTCCA
                                                 tgcaggCAGCTcTaTGAAgcGCAGGtActcca
                                                 tgcaggCagCtCTCTgaAGcgAagGgACtccA
                                                 ggcaggCAGCtCTCtGaaGcGcAggtactcca
                                                 TgCaggCAGCTCTCTGAAGCGCAGGtAcTCca
```

# Differentiating SNPs from error



```
CTTCAGGTTGGGTGAGTTGTCATCAGTCGGAATGCGCAGGTCACTTACCTGCAGGCAGCTCTCTGAAGCGCAGGTACTCCATCGACCGGGTGGTGACTAG
NNNNNNNNNNNNNNNNNNNNNATCAGTCGGAATGCGCAGGTCACTTACCTGCAGGCAGCTCTCTGAAGCGCAGGTACTCCANNNNNNNNNNNNNNNNNNN
                     ATCAGTCGGAATGCGCAGGTCACTTACCTGCA
                     ATCAGTCGGAATGCGCAGGTCACTTACCtGCA
                     ATCAGTC gAATGC CAGGTCaCTTacctGcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctGcA
                     ATCA TCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTACCTGCA
                     ATCAGTCGGAATGCGCAGGTCACTTAcCtGcA
                     ATCAGtCGGAATgCGCAGGTCACTTACCtGCA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ggcagttggAAtgcGCAggTcaCTTccctgcA
                     AtCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCgGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTAcctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTaCctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTacctgcA
                     ATCAGTCGGAATGCGCAGGTCACTTACctGcA
                                                 TgcaggCAGcTCTCTGAAgCGCAGG ACtCCA
                                                 TgcaggCAGCTCTCTGAAGCGCAGGtACTCCA
                                                 TgCaggCAGCTCTCTGAAGCGCAGGgACtcCA
                                                 TGCAGGCAGCTCTCTGAAGCGCAGGtaCTCCA
                                                 TgcaggCAGCTCTCTGAAGCGCAGGtACTCcA
                                                 tgcaggcAGCtCTCTGAAGcgcaGgtaCTCca
                                                 TgCaggCAGCTCTCTGAAGCGCAGGtACTcCA
                                                 TgcaggCAGCTCTCTGAAgCGCagGtaCtcC-
                                                 TgcaggcAGcTcTCTGAAGcgCAgGTActcca
                                                 TGCaggCAGCTCTCTGAAGCGCAGGtACTCCa
                                                 TgCaggCAGCTCTCTGAAGCGCAgGtAcTccA
                                                 TGCaGgCAGCTCTCTGAAGCGCAGGTACTCCA
                                                 tgcaggCAGCTcTcTGAAgcGCAGGtActcca
                                                 tgcaggCagCtCTCTgaAGcggagGgACtccA
                                                 tgcaggCAGCtCTCtGaaGcGcAggtactcca
                                                 TgCaggCAGCTCTCTGAAGCGCAGGtAcTCca
```

# Differentiating SNPs from error

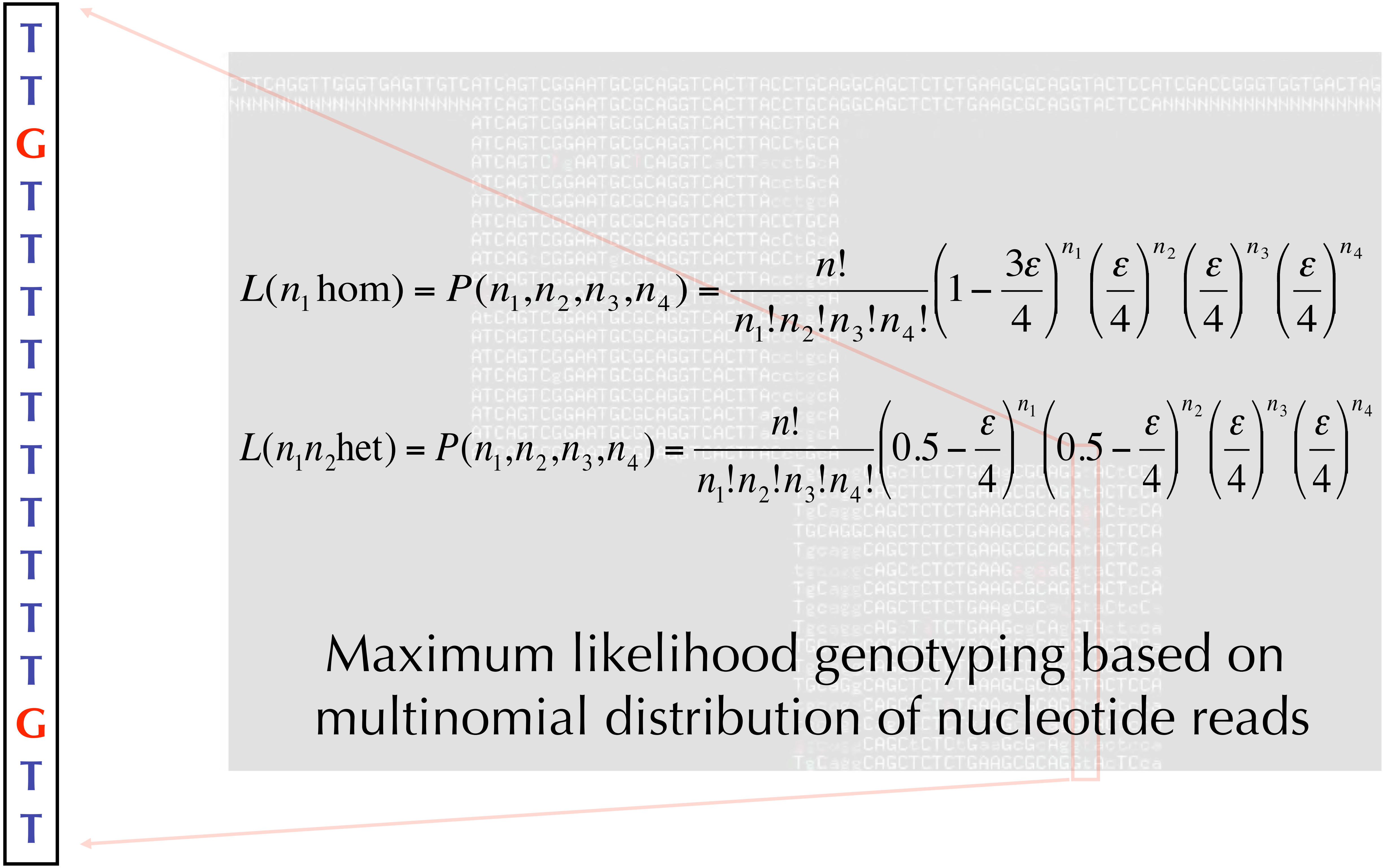


The reads are 14 T and 2 G:

GT heterozygote?
GG homozygote with error?
AA homozygote with lots of error?

Needed a rigorous method to call genotypes

# Differentiating SNPs from error



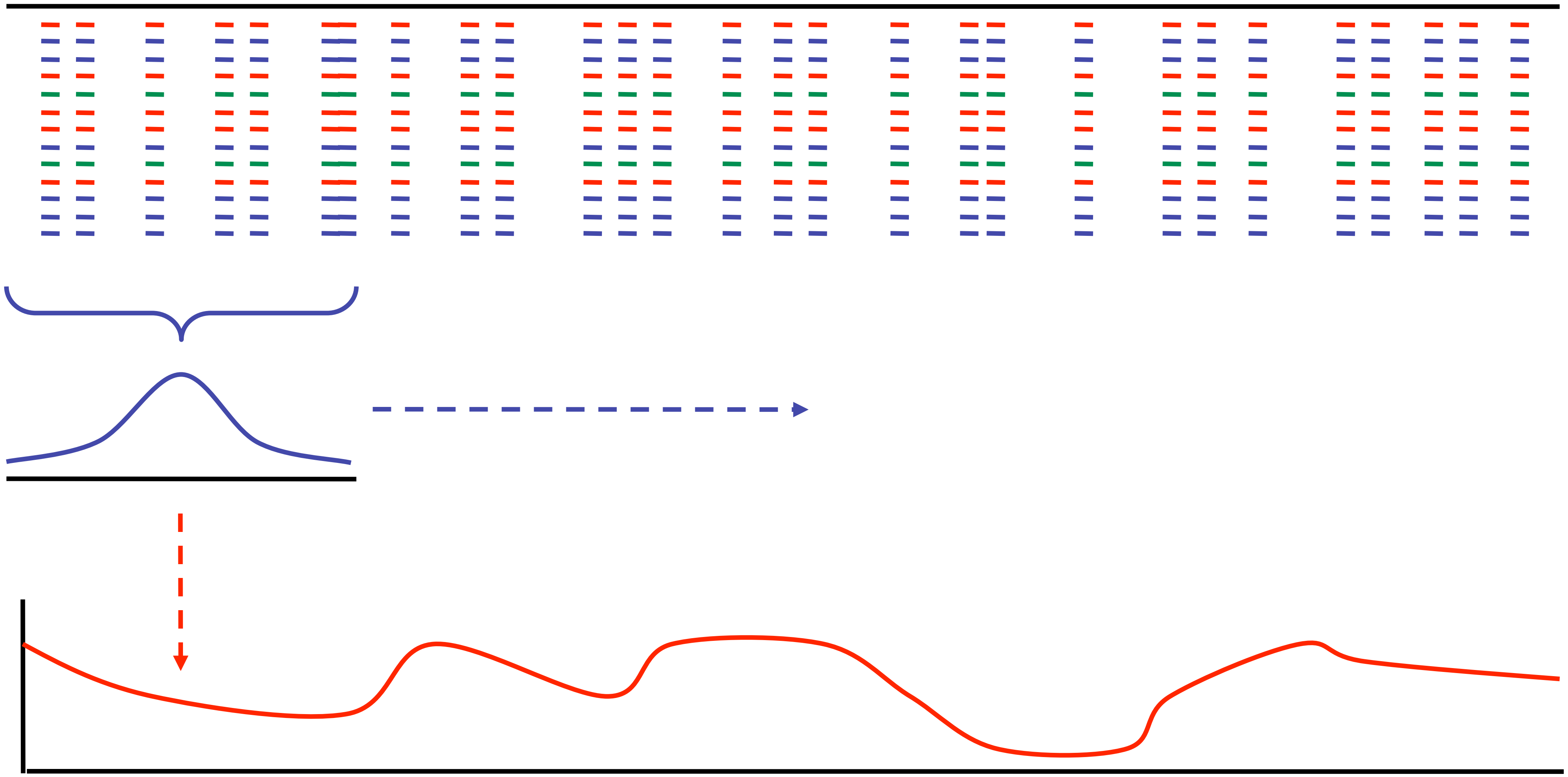
$$L(n_1\,\text{hom}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

$$L(n_1 n_2\text{het}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(0.5-\frac{\varepsilon}{4}\right)^{n_1}\left(0.5-\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

Maximum likelihood genotyping based on multinomial distribution of nucleotide reads

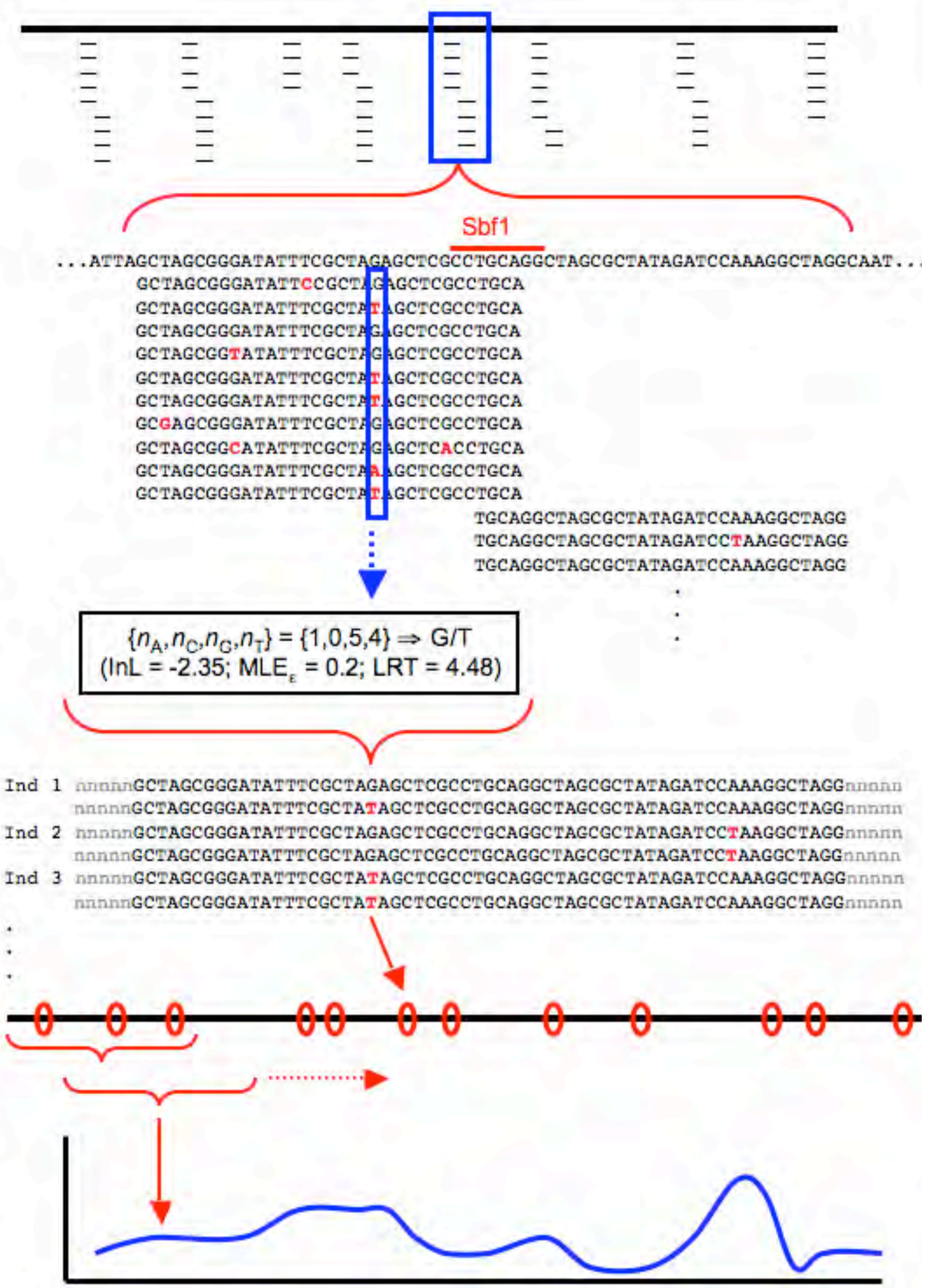
# Making statistics continuous across the genome

Kernel-smoothing average of summary statistics along genome



Bootstrap re-sampling to estimate significance of moving average

# Overall pipeline


Sbf1
...ATTAGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGCAAT...
GCTAGCGGGATATTCCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGTATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCGAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGCATATTTCGCTAGAGCTCACCTGCA
GCTAGCGGGATATTTCGCTAAAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCTAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
$\{n_A, n_C, n_G, n_T\} = \{1,0,5,4\} \Rightarrow$ G/T
(lnL = -2.35; MLE$_\varepsilon$ = 0.2; LRT = 4.48)
Ind 1 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
Ind 2 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
Ind 3 nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn

# Experimental design considerations for RAD

*Tradeoffs***:**
**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

## How many tags do I need?

Things to consider

Choice of enzyme and genome size $(0.25)^n$ × genome size = expected # sites

Genomes are biased:

| | |
|---|---|
| expect 112,300 six-cutter sites in stickleback (460 Mb) | actual ***EcoRI*** sites = 90,000 |
| expect 7000 eight-cutter sites in stickleback | actual ***SbfI*** sites = 22, 800 |
| expect 32,900 six-cutter sites in *C. remanei* (135 Mb) | actual ***EcoRI*** sites = 73,200 |

# Experimental design considerations for RAD

*Tradeoffs***:**
**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

## How many tags do I need?

Things to consider

Choice of enzyme and genome size
Polymorphism and read length

Nucleotide polymorphism rate = 0.01 to 0.001 for most vertebrates

Stickleback populations: 0.01 to 0.02. At least 1 SNP every 100 bp, on average

# Experimental design considerations for RAD

*Tradeoffs***:**
**Number** of sites versus **Depth** of sequencing per site versus **Number of samples**

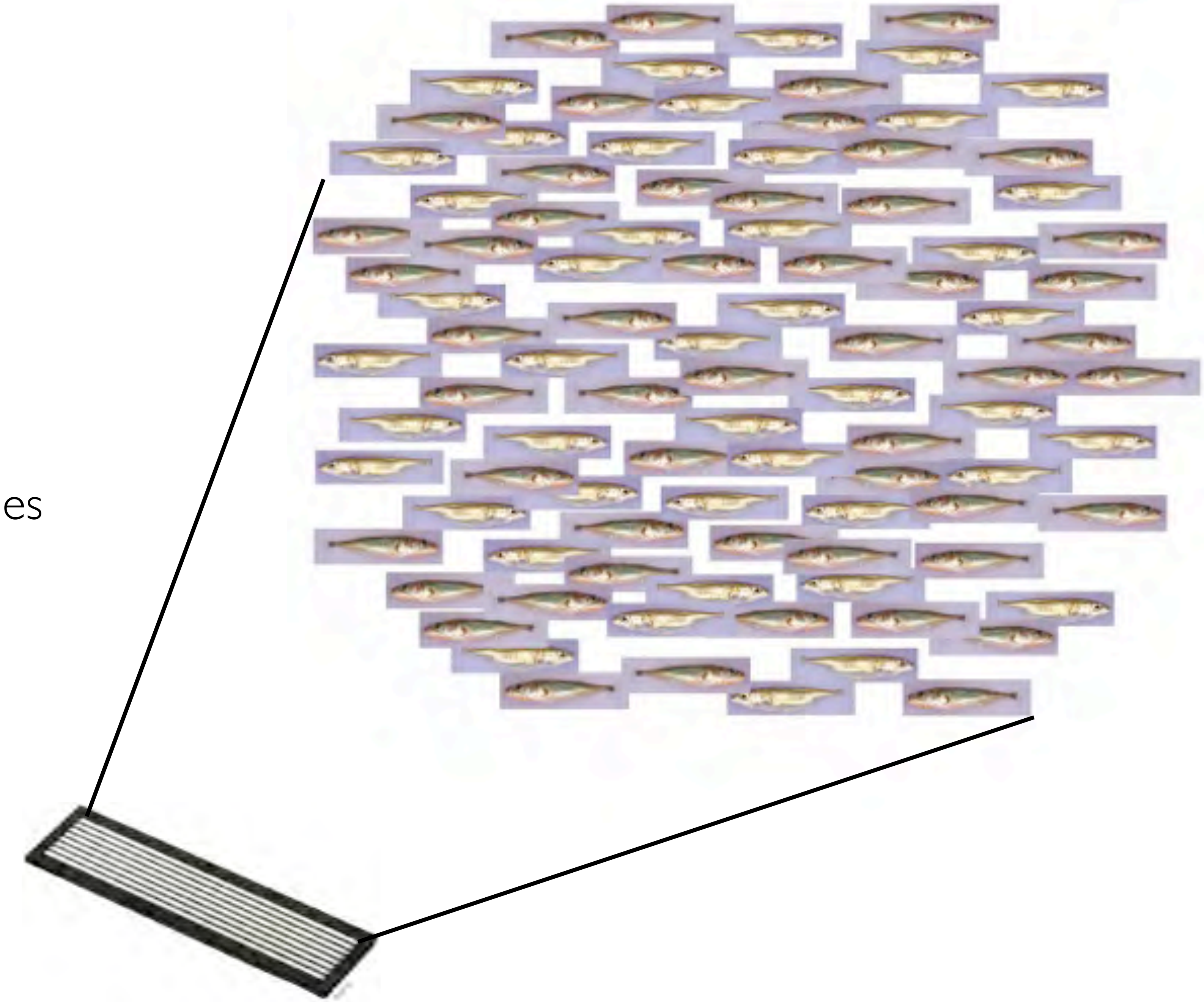
## How many samples should be multiplexed?

*Things to consider*

Barcoded adapters

- 5 to 8nt barcodes
- Variable length barcodes
- Combinatorial barcodes (PE)

Barcode distance - two mismatches



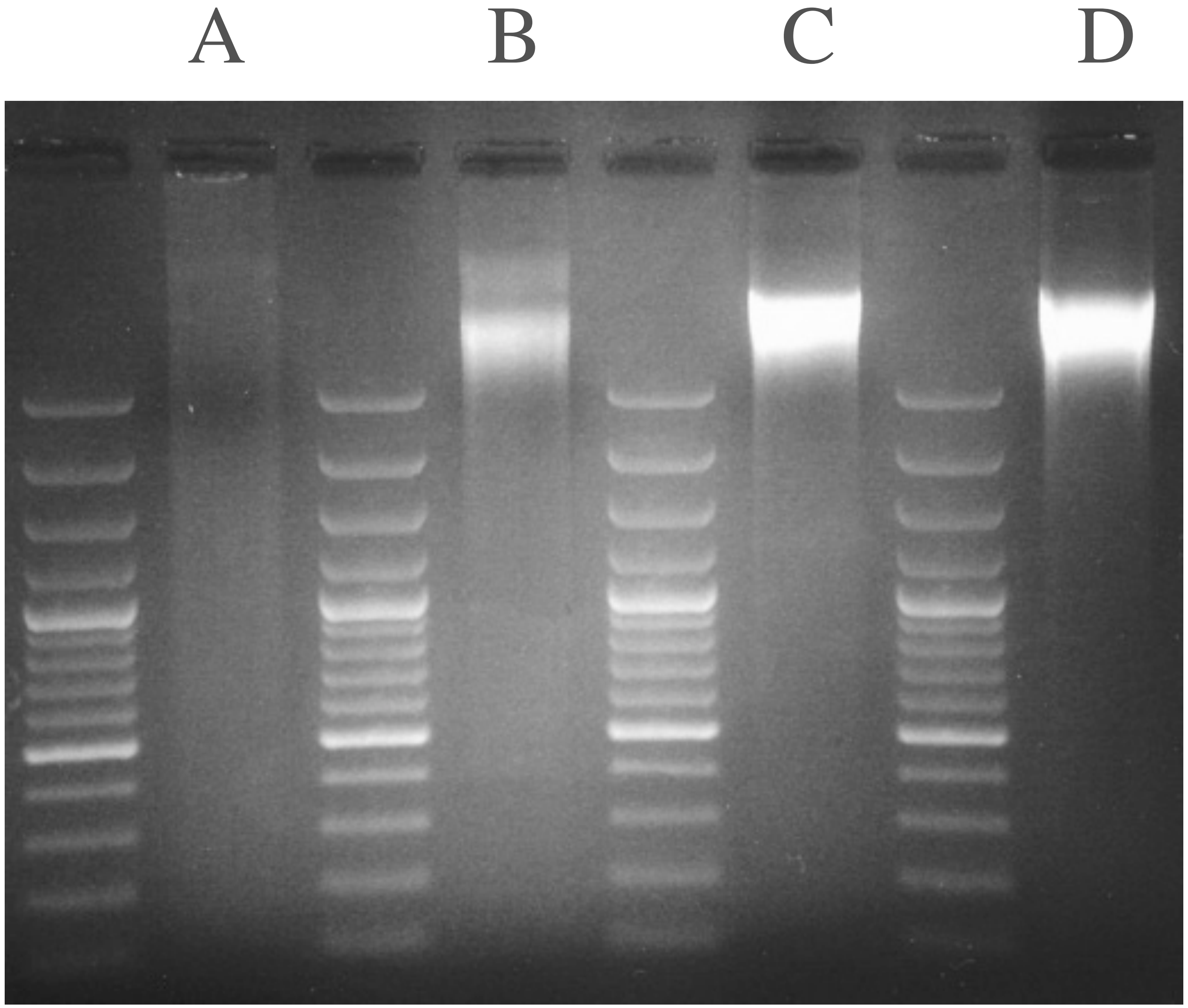
# Molecular considerations in library building

## How many samples should be multiplexed?

Things to consider

DNA Quality

Multiplex only like samples to help equalize representation of poor quality samples



# Molecular considerations in library building

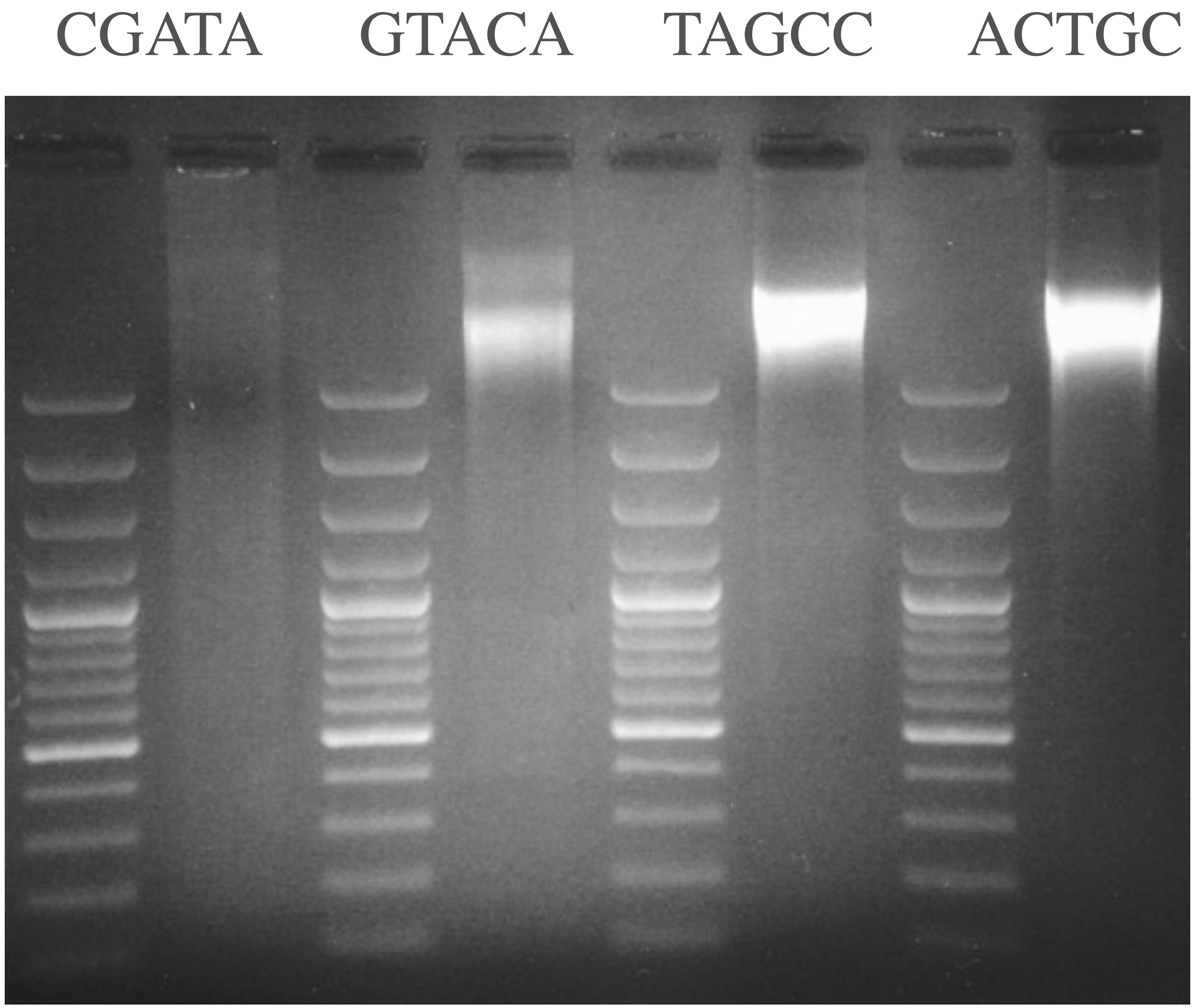
## How many samples should be multiplexed?

Things to consider

DNA Quality

Diversify barcodes

- Illumina cluster calling is confused by repetition in first 4 bases - can offset barcodes



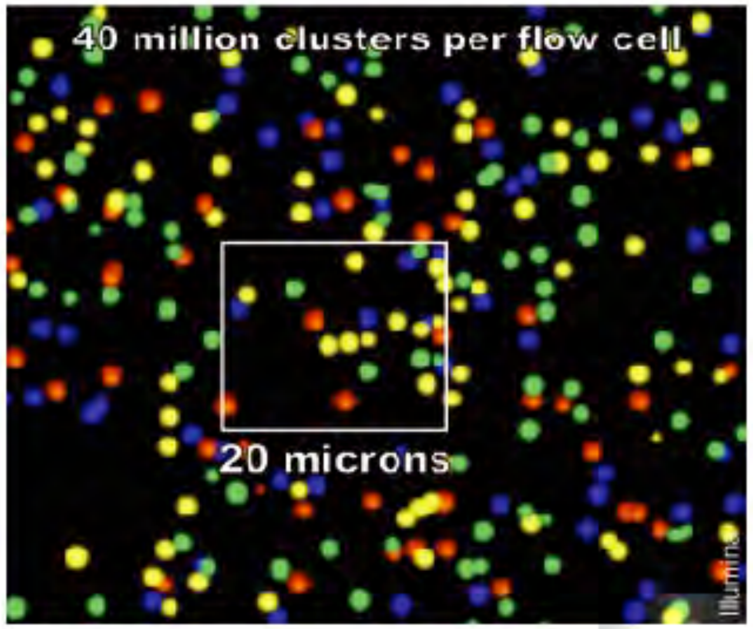
# Molecular considerations in library building

## How can I get the best depth of coverage?

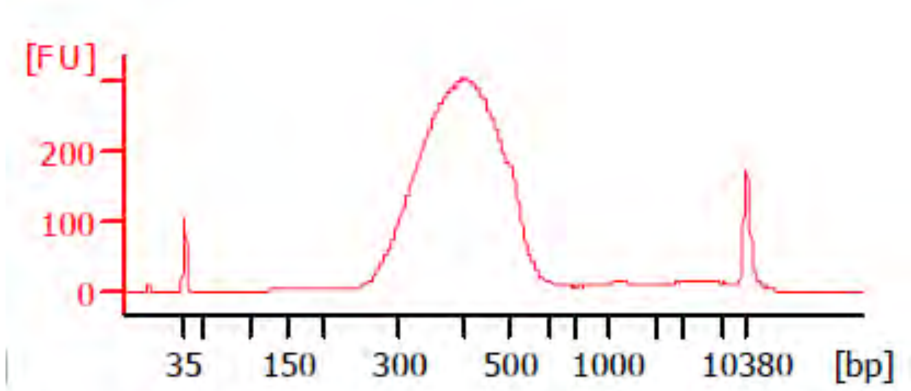
Things to consider

Fragment size

Smaller/tighter is better



Agilent Bioanalyzer



# Molecular considerations in library building

## How can I get the best depth of coverage?

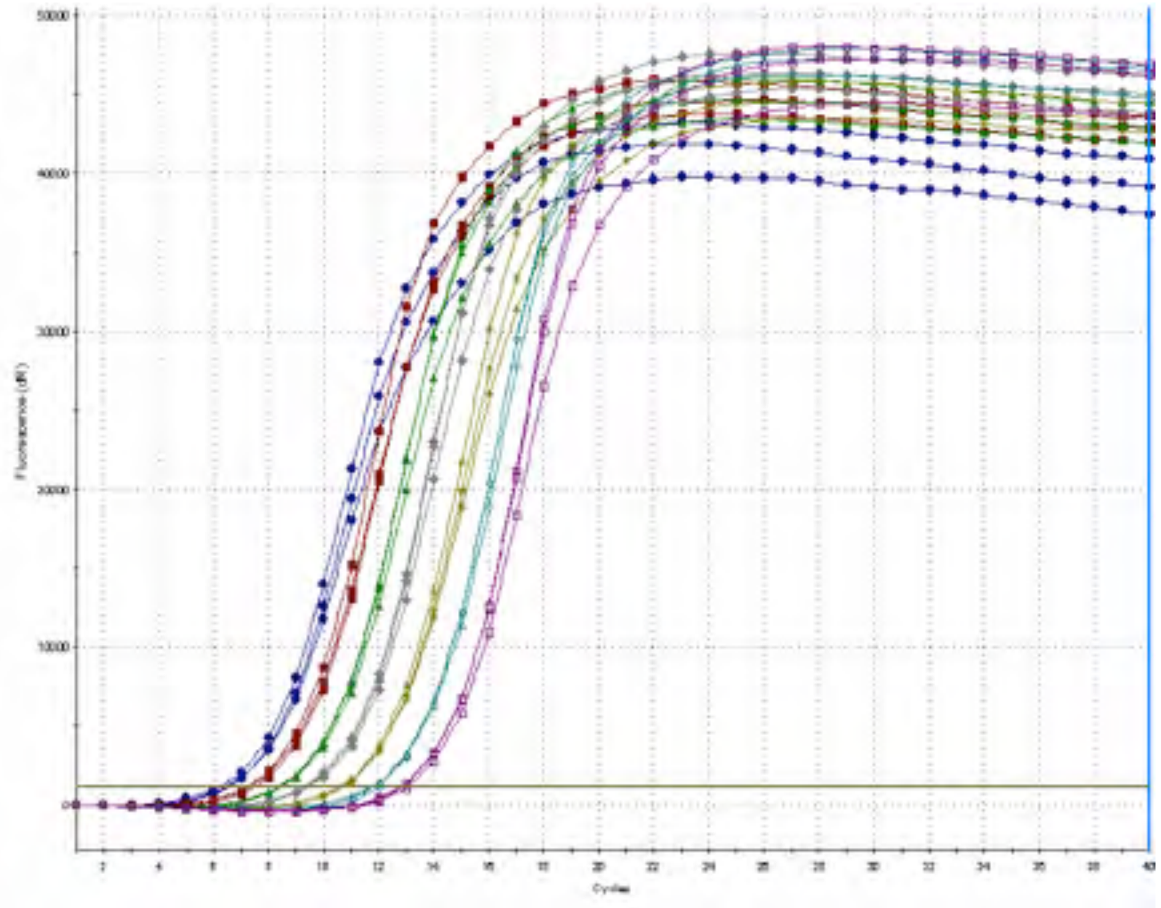
Things to consider

Fragment size

Library quality

qPCR

qPCR control should be similar to measured sample:



# Molecular considerations in library building

## How can I get the best depth of coverage?

Things to consider

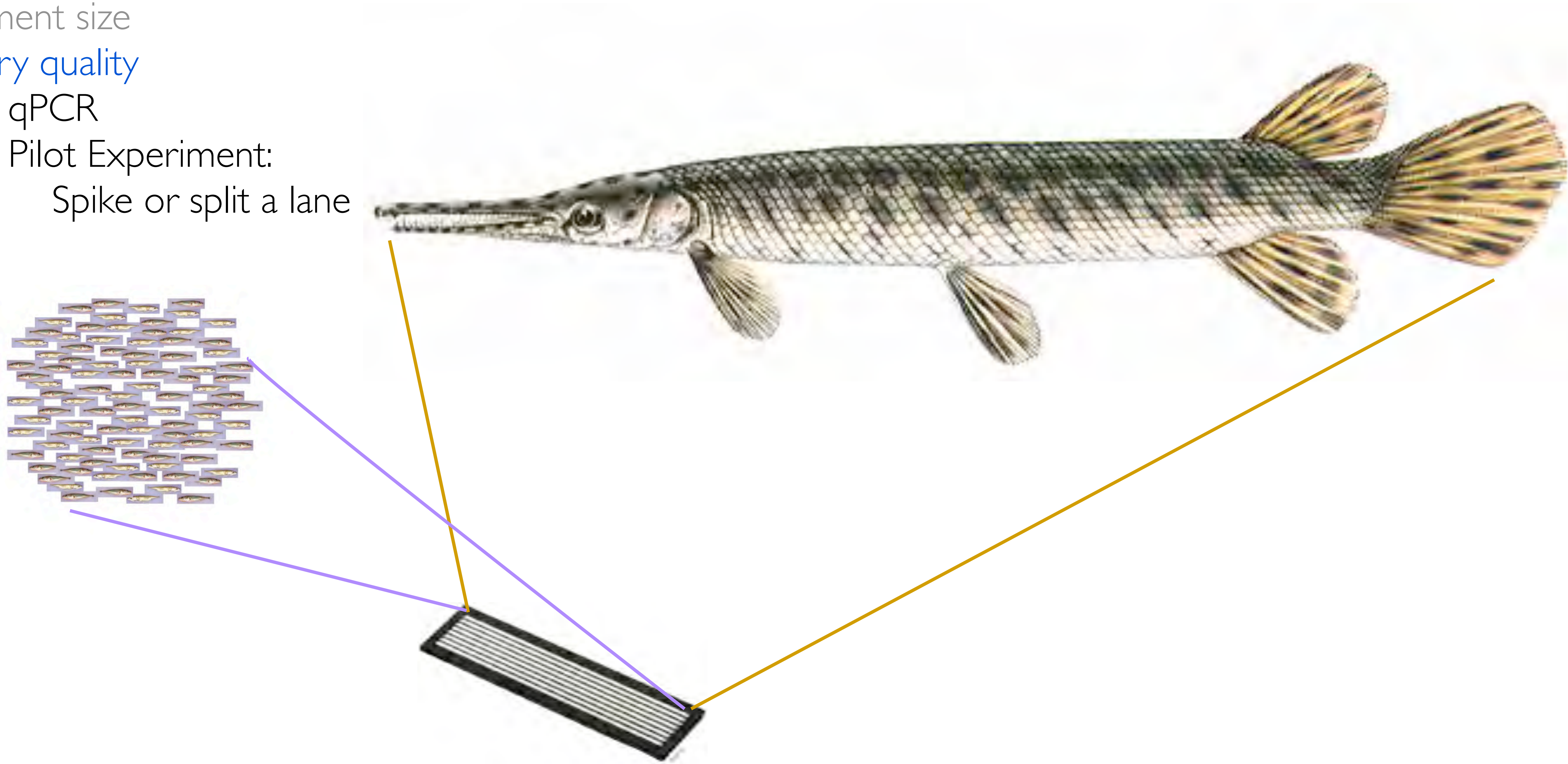
Fragment size

Library quality

qPCR

Pilot Experiment:

Spike or split a lane



# A case study of using RAD for an organism with a reference genome: population genomics of threespine stickleback fish

Susie Bassham, Julian Catchen, Paul Hohenlohe
Emily Lescak and Frank von Hippel

# Threespine stickleback, *Gasterosteus aculeatus*

● *Ancestral Oceanic Populations*

Marine and Anadromous

Old (> 10 million years)

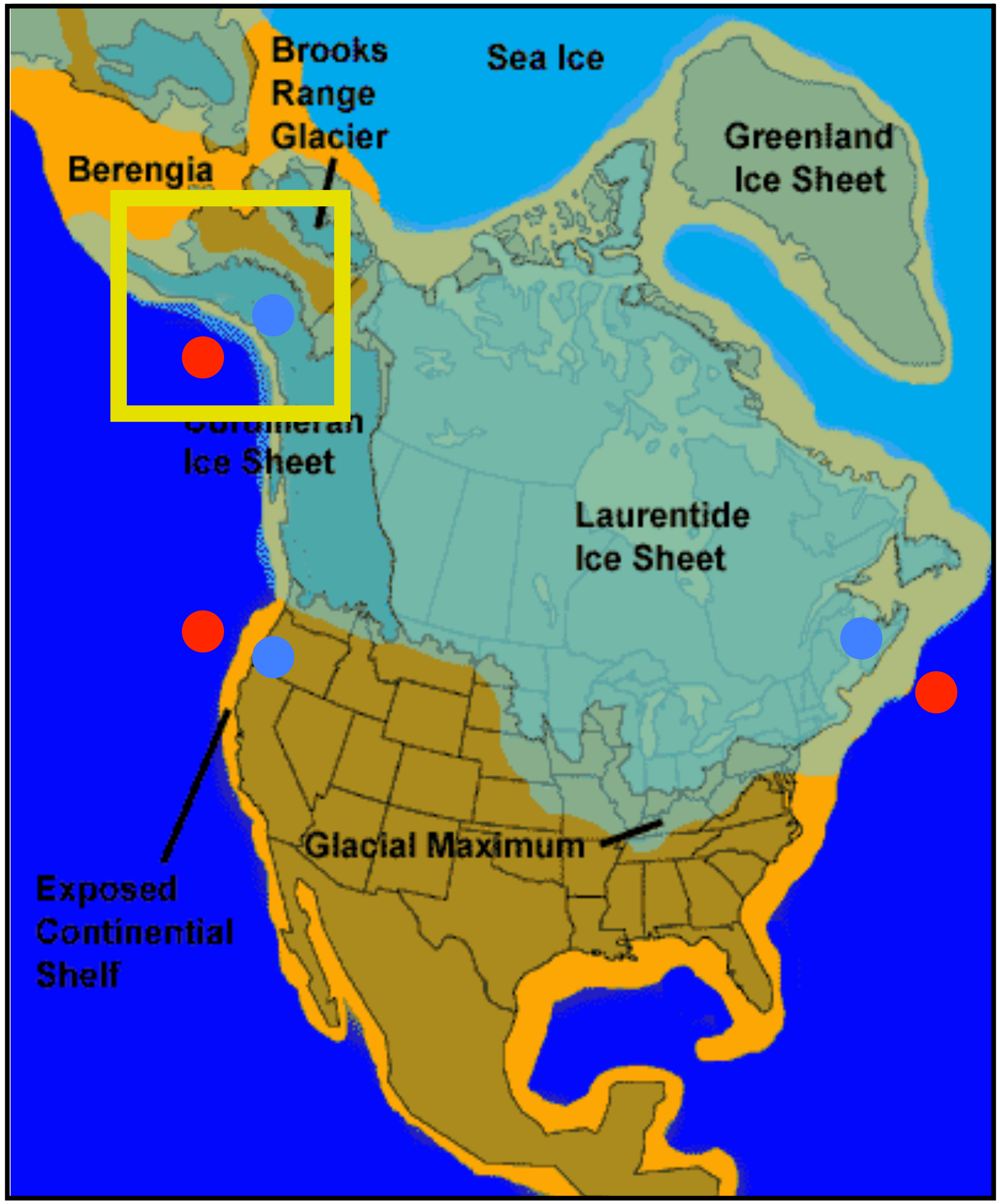
Phenotypically similar

● *Derived Freshwater Populations*

Lake and stream

Young (<15,000 years)

Phenotypically diverse



# Threespine stickleback, *Gasterosteus aculeatus*

🔴 *Ancestral Oceanic Populations*

Marine and Anadromous

Old (> 10 million years)

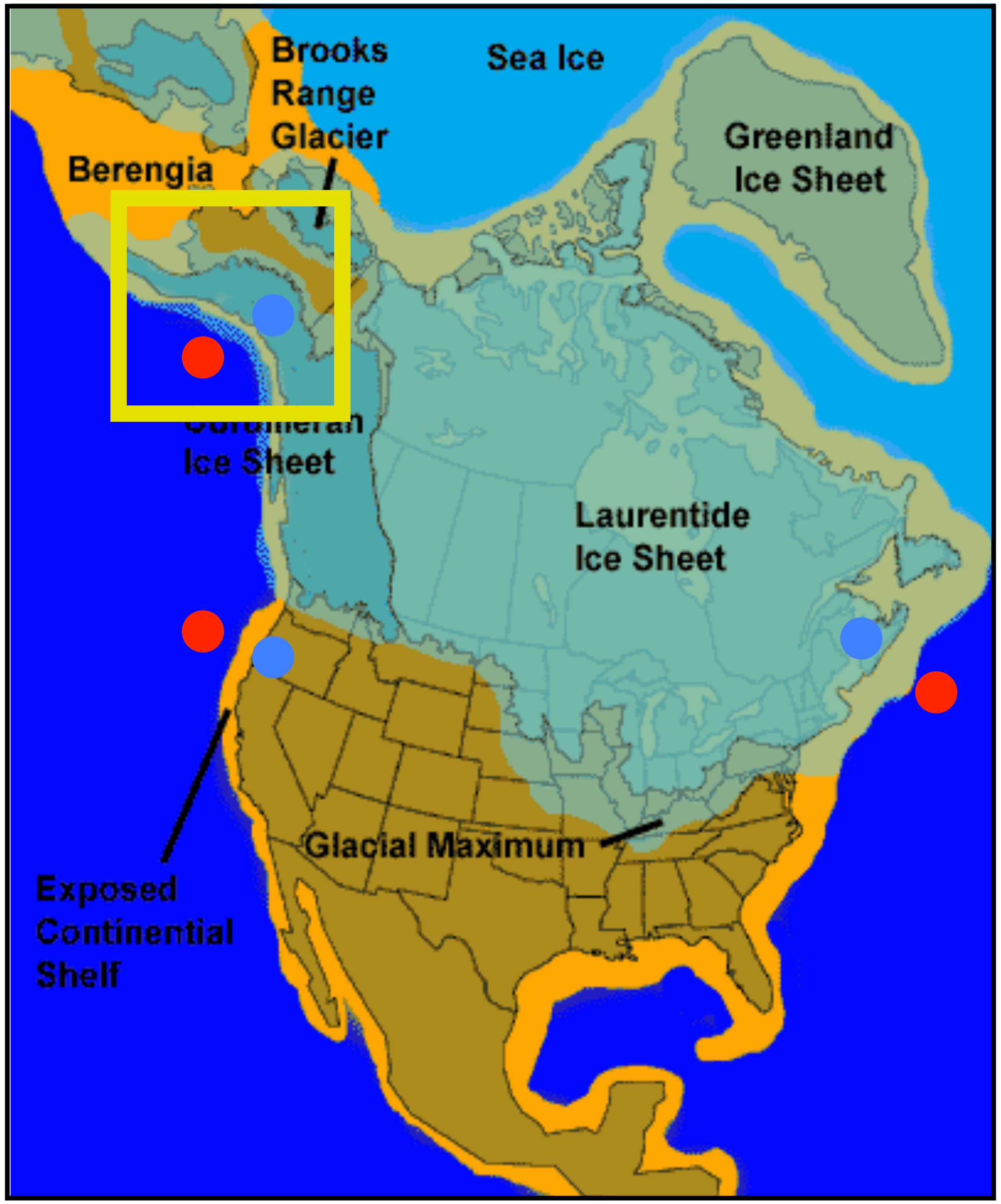
Phenotypically similar

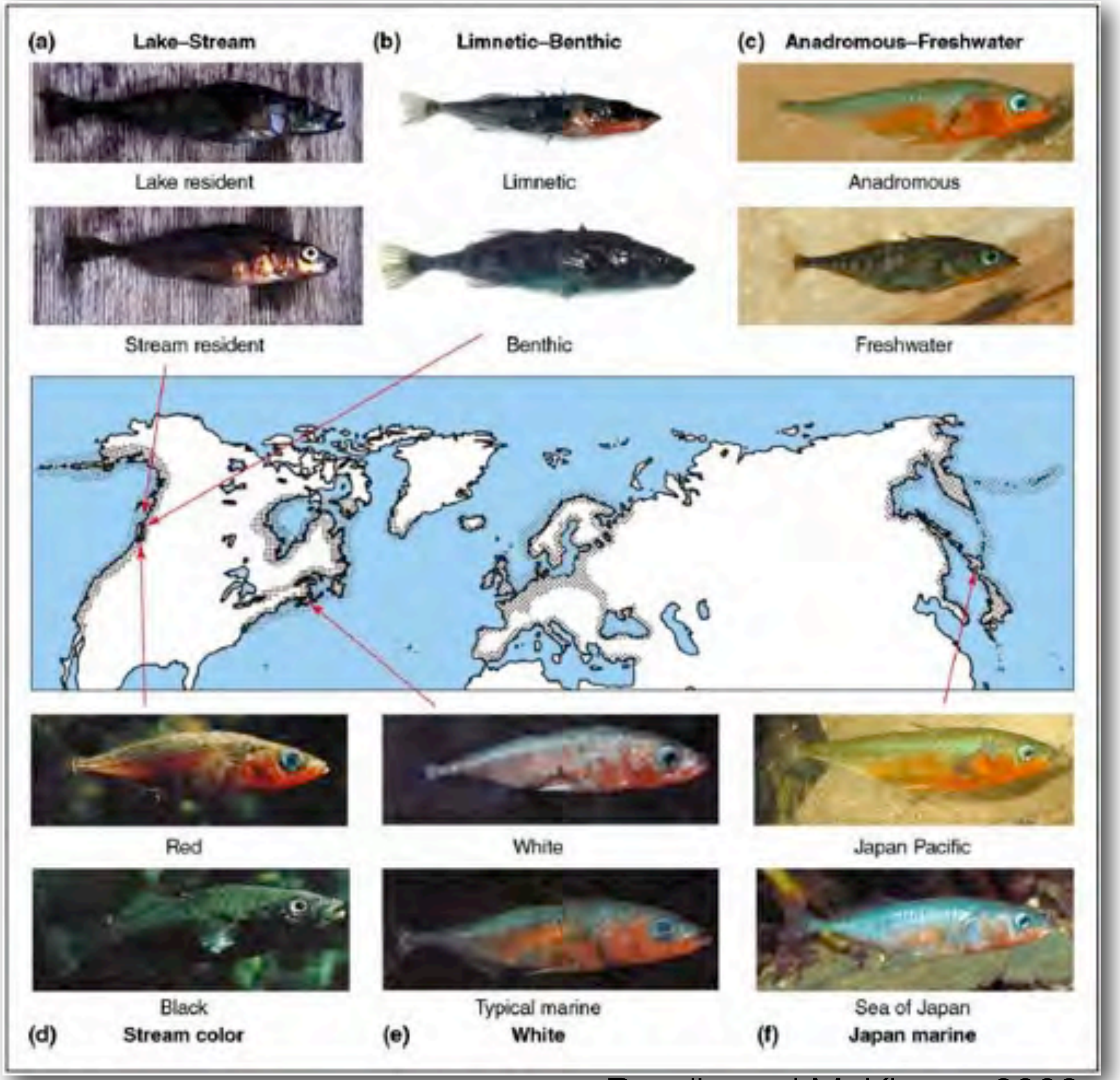
🔵 *Derived Freshwater Populations*

Lake and stream

Young (<15,000 years)

Phenotypically diverse



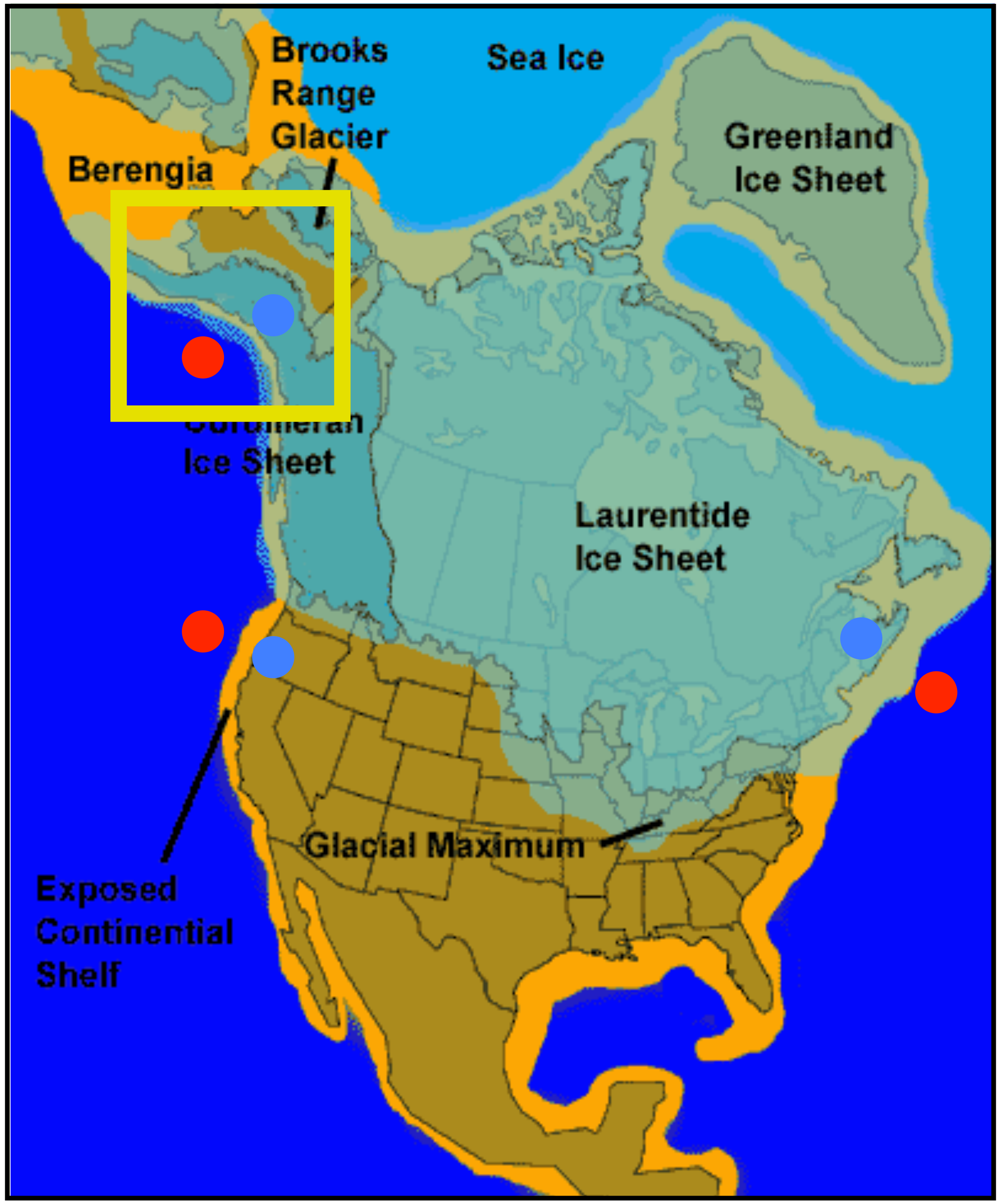


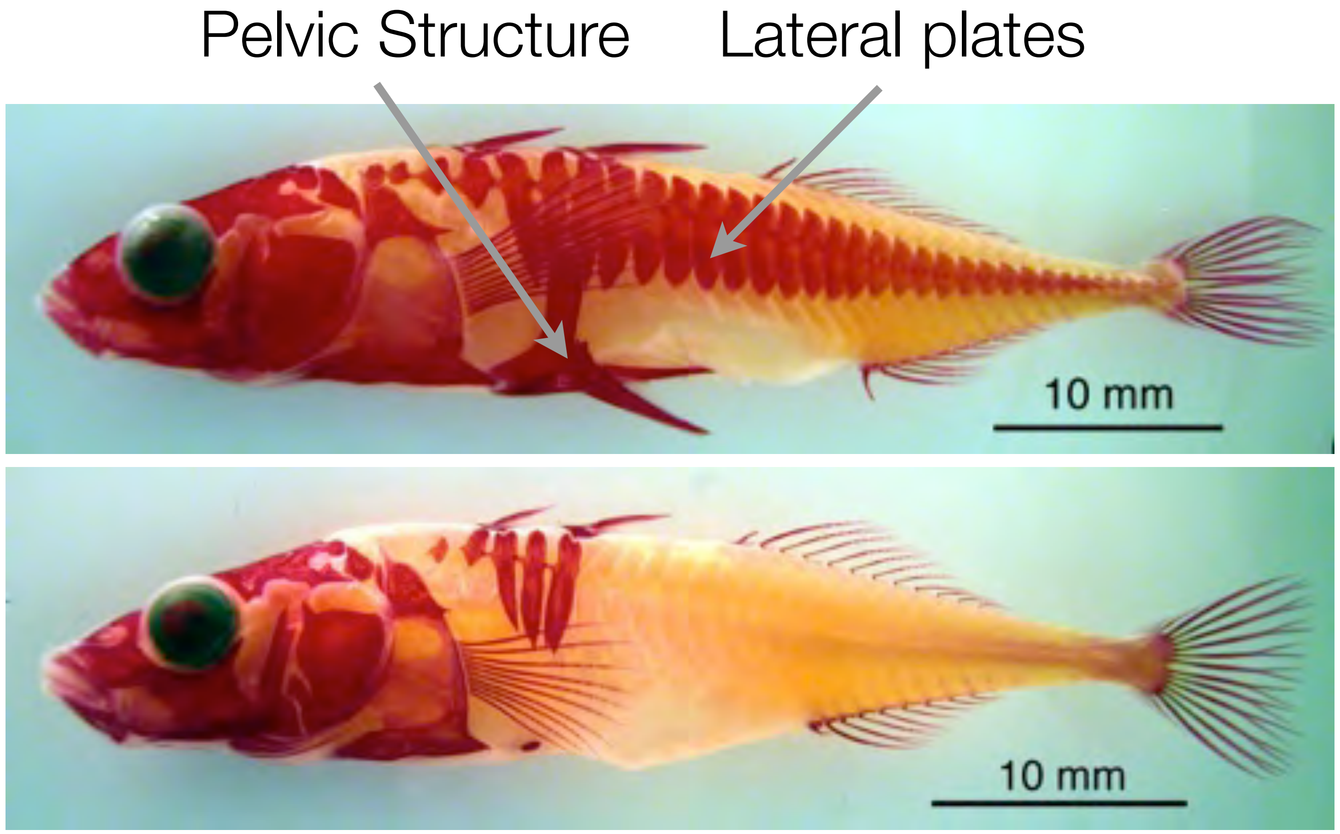
Rundle and McKinnon 2002

# Threespine stickleback, *Gasterosteus aculeatus*

● *Ancestral Oceanic Populations*
Marine and Anadromous
Old (> 10 million years)
Phenotypically similar

● *Derived Freshwater Populations*
Lake and stream
Young (<15,000 years)
Phenotypically diverse

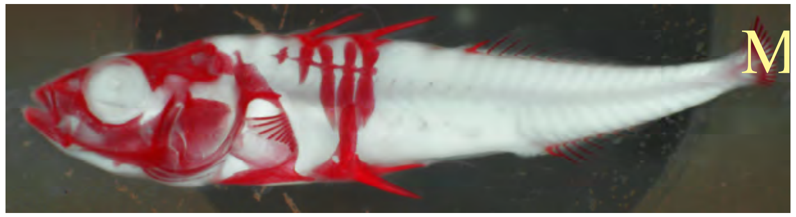


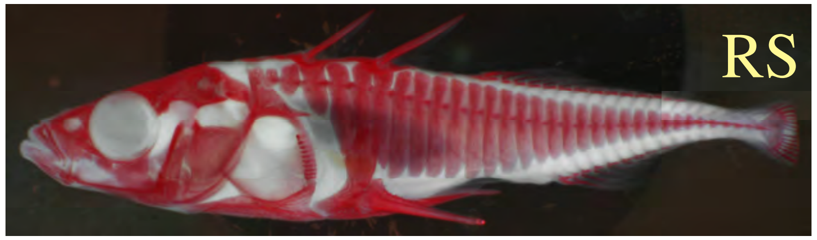


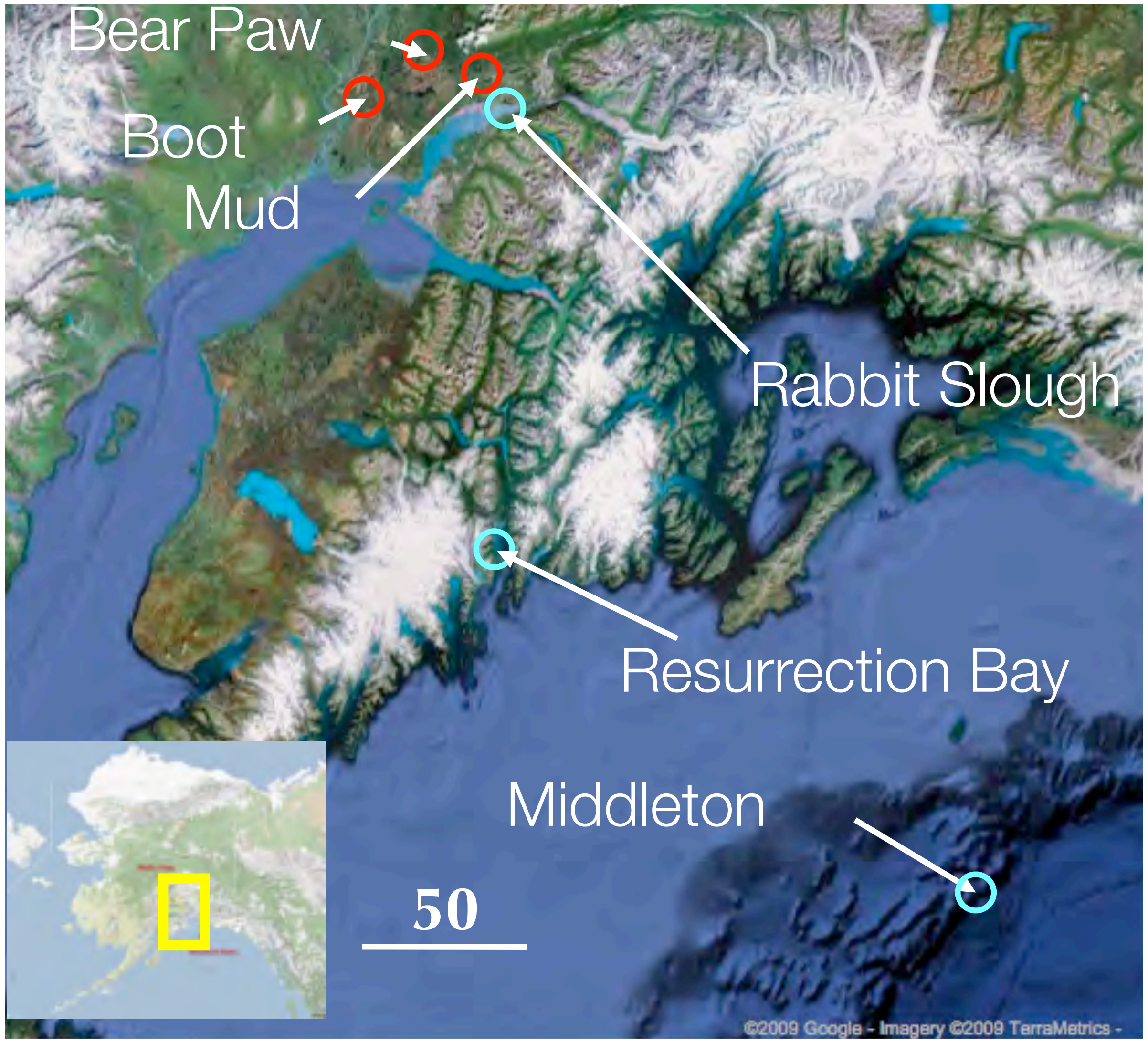
QTL mapping

Cresko et al. 2004. PNAS
Colosimo et al. 2005. Science
Shapiro et al. 2004. Nature
Albert et al. 2008. Evolution
Miller et al. 2007. Cell
Chan et al. 2010. Nature

# Signatures of natural selection across the genome



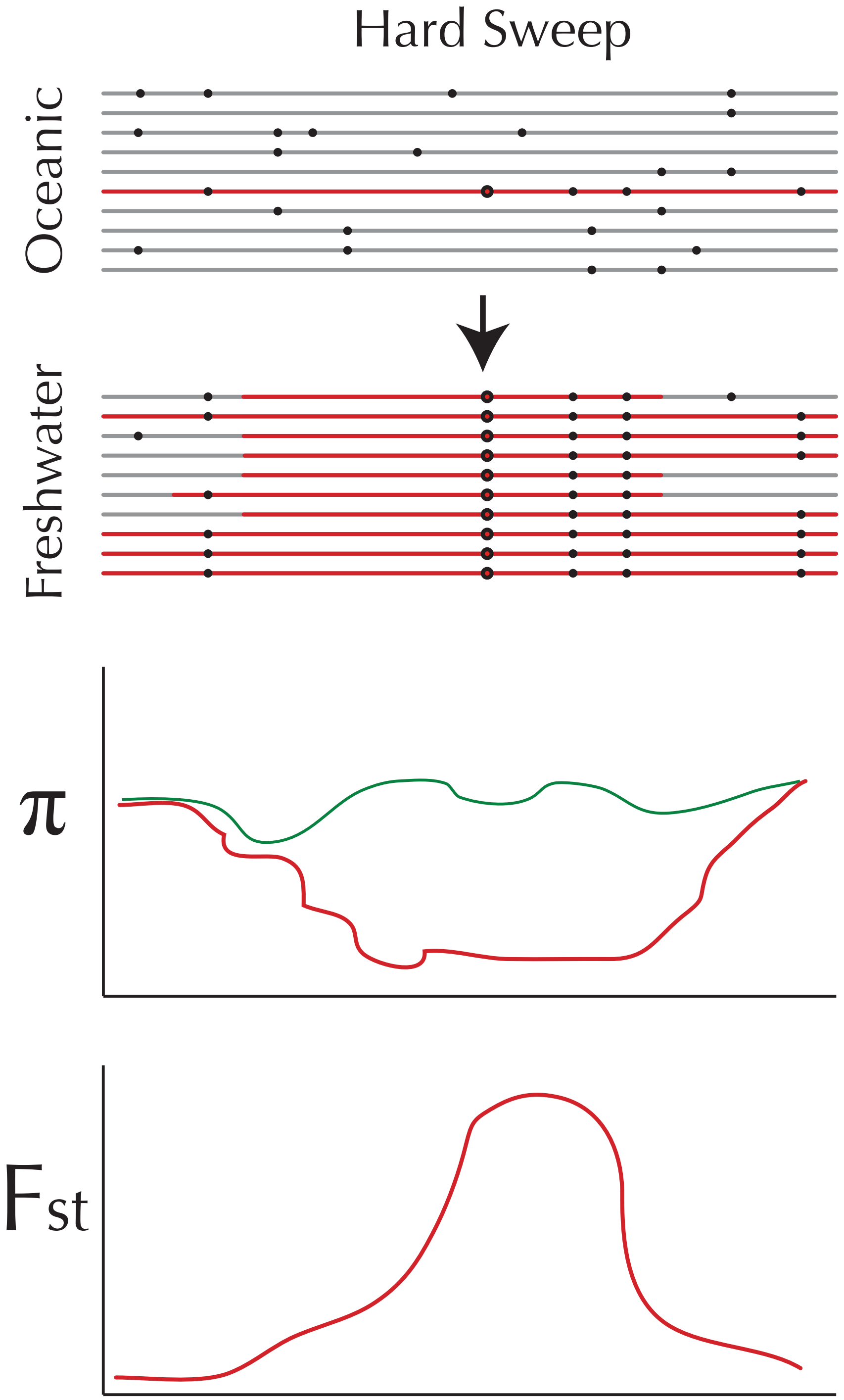




20 individuals in each of 5 popn’s
2 Ocean & 3 Freshwater
45,000 SNPs in each individual

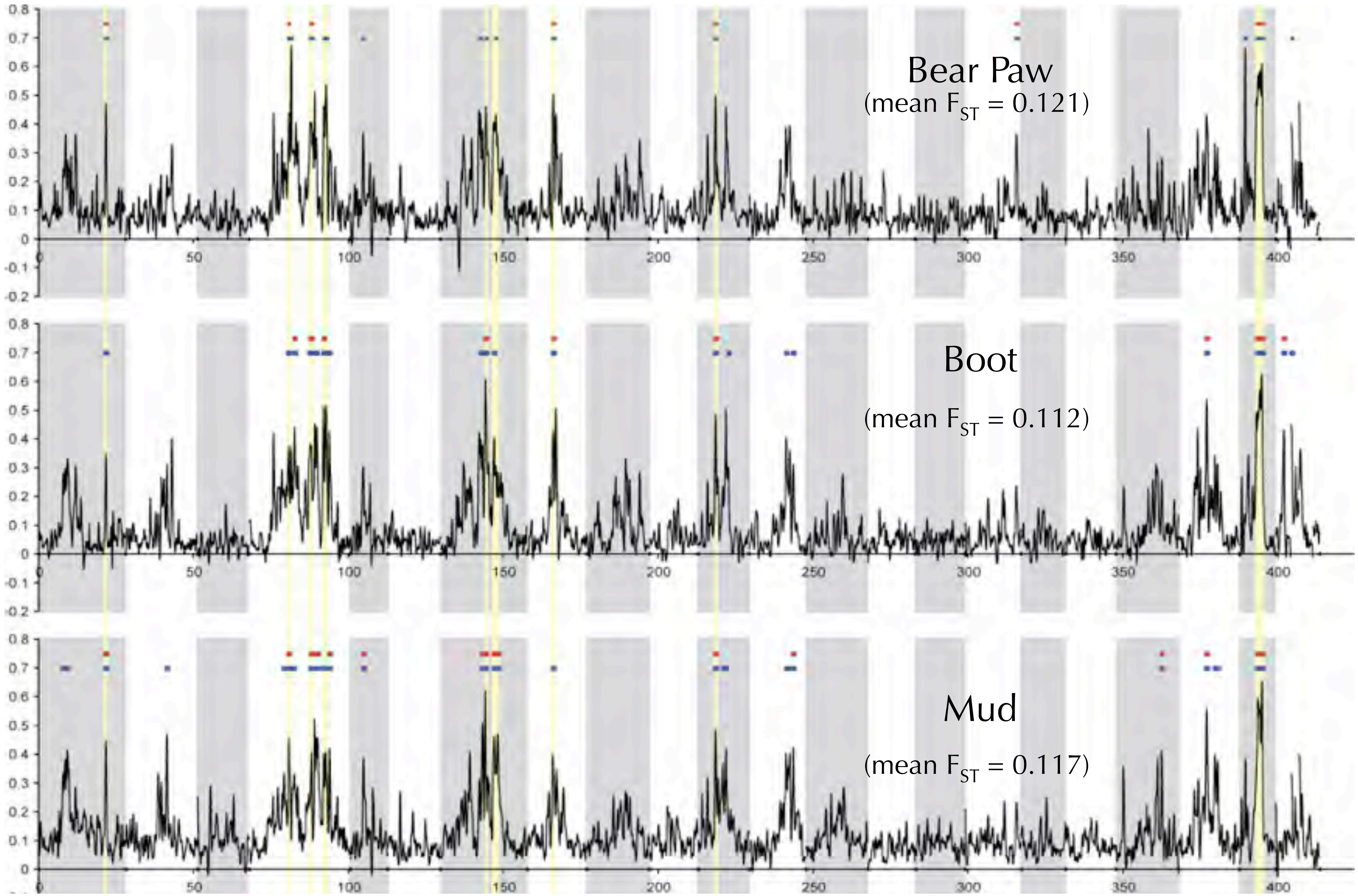
Hohenlohe, Bassham et al. 2010. PLoS Genetics

# What genomic regions are subject to selection during parallel evolution?



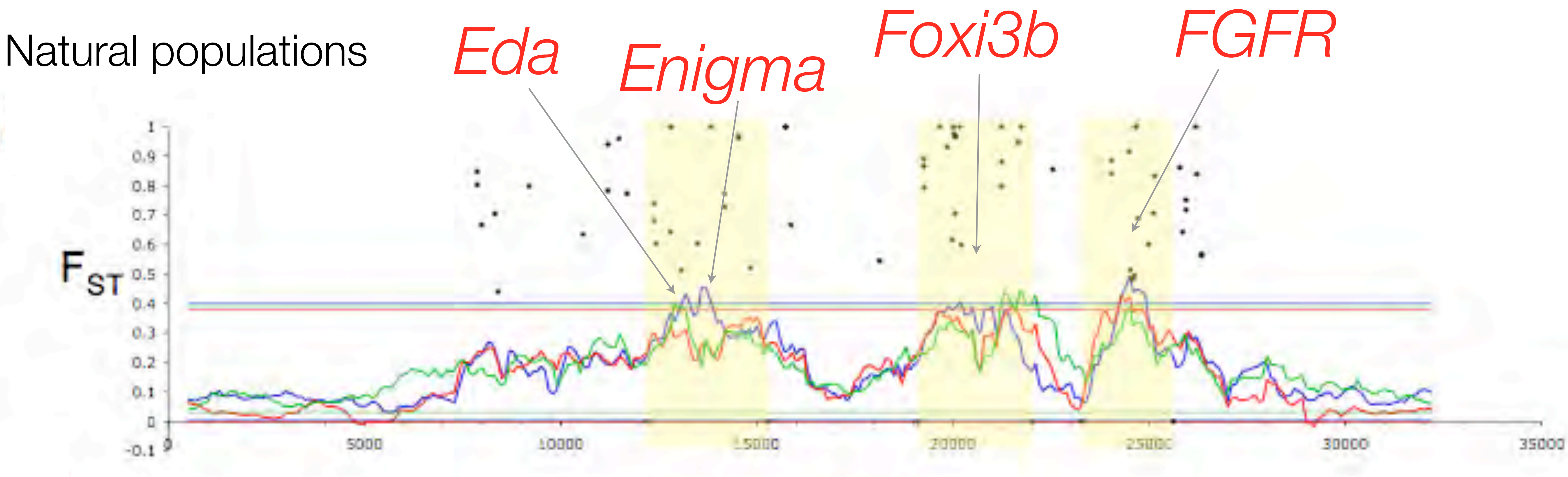
# Parallel signatures of selection across the genomes

$F_{st}$

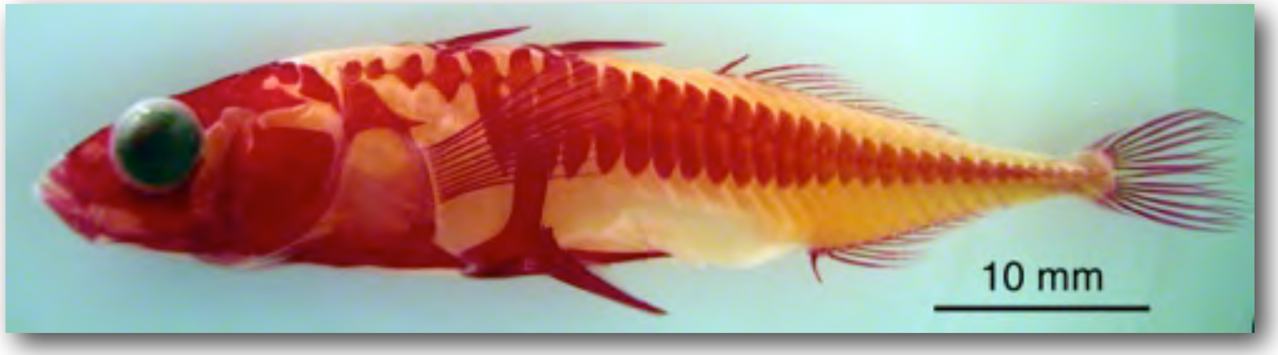


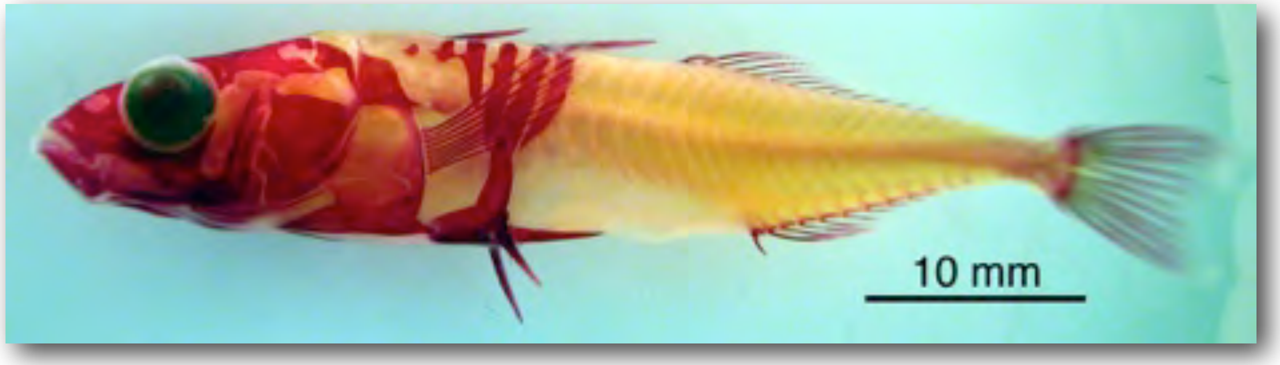
Genomic location (mBases)

# Previously identify quantitative trait loci (QTLs) are under selection

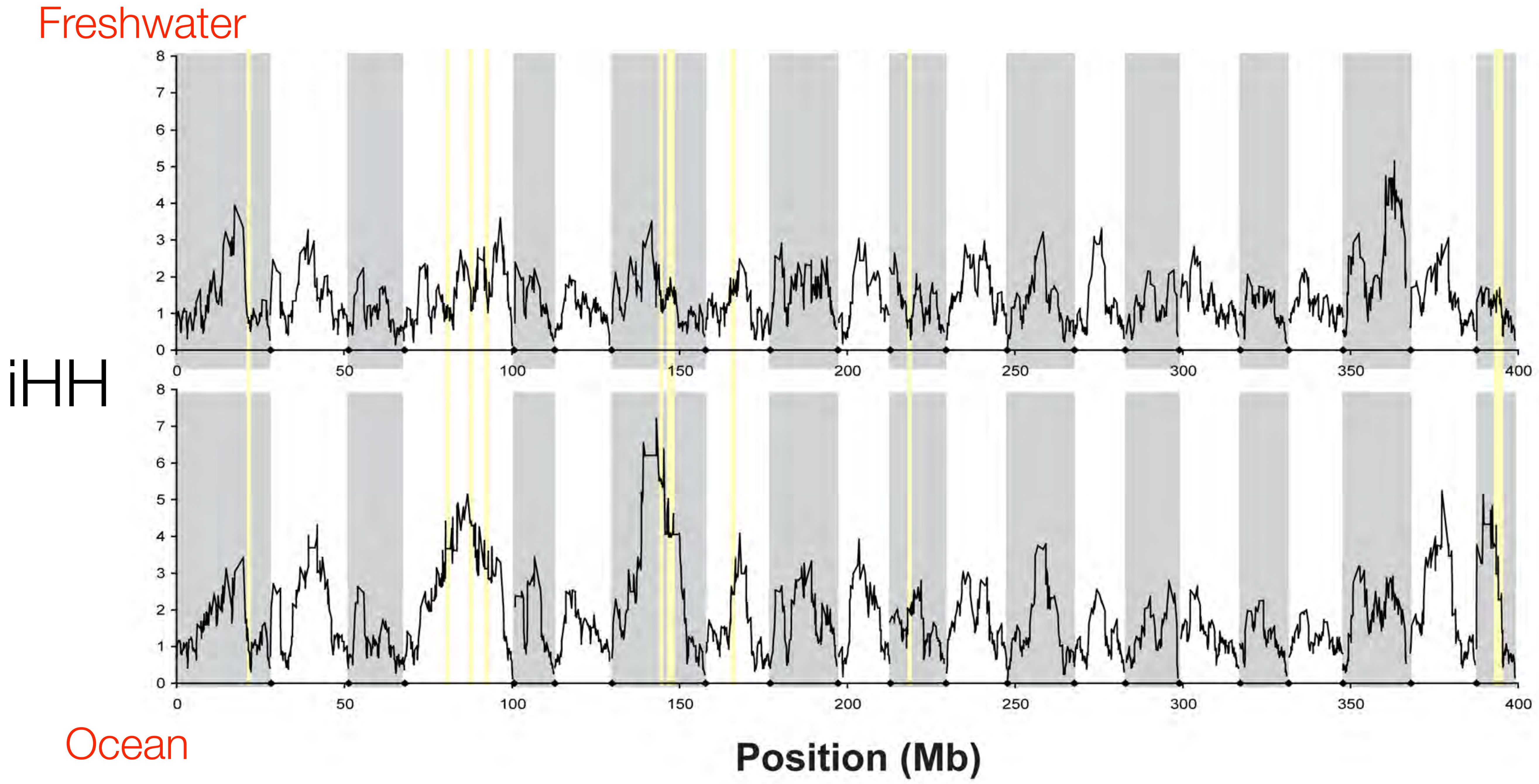


Lateral plate major locus
on LGIV (4000 SNPs)



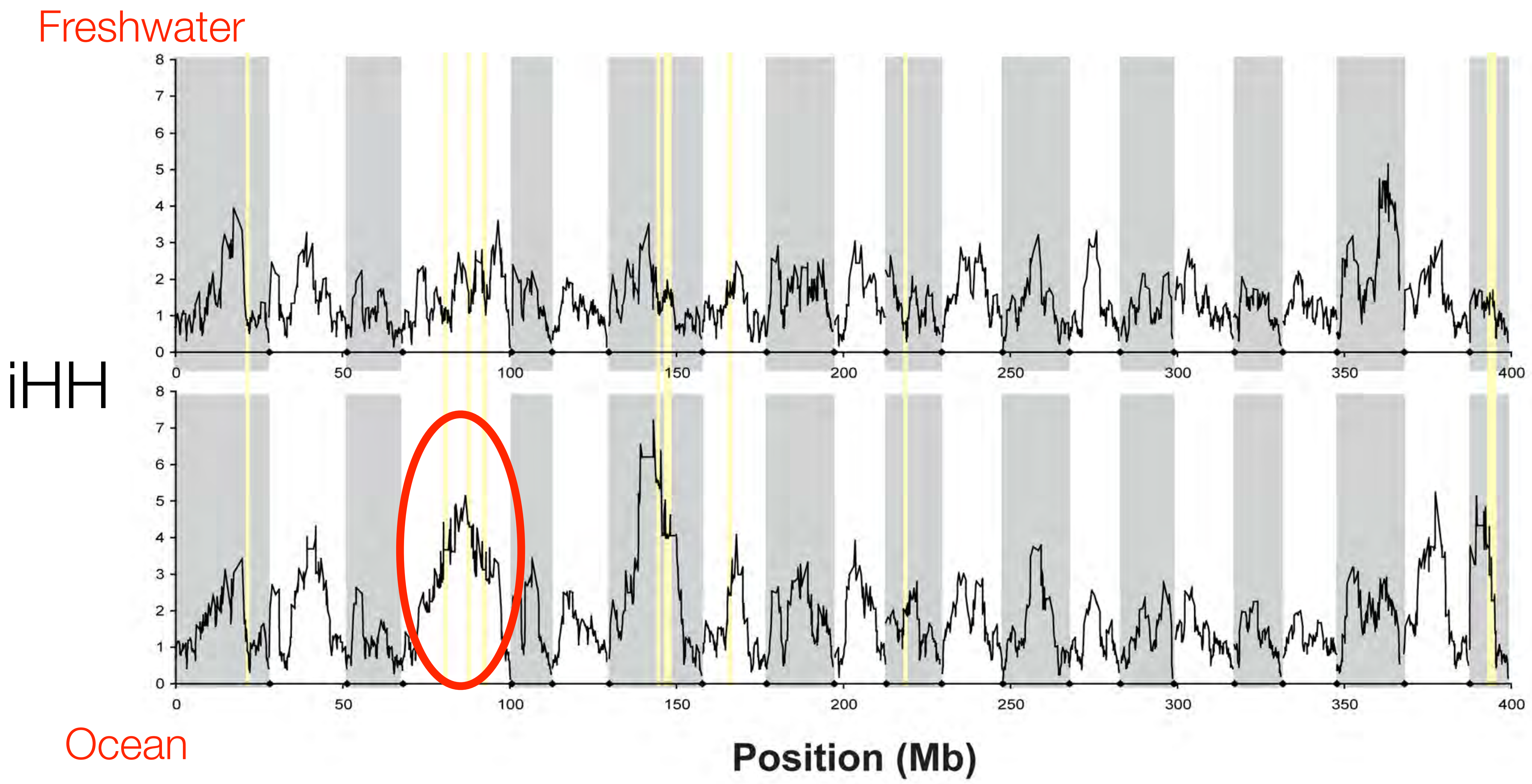


# Extensive LD across the genome



Hohenlohe et al. 2012. Philosophical Transactions of the Royal Society of London

# Extensive LD across the genome
# More in oceanic than in freshwater populations

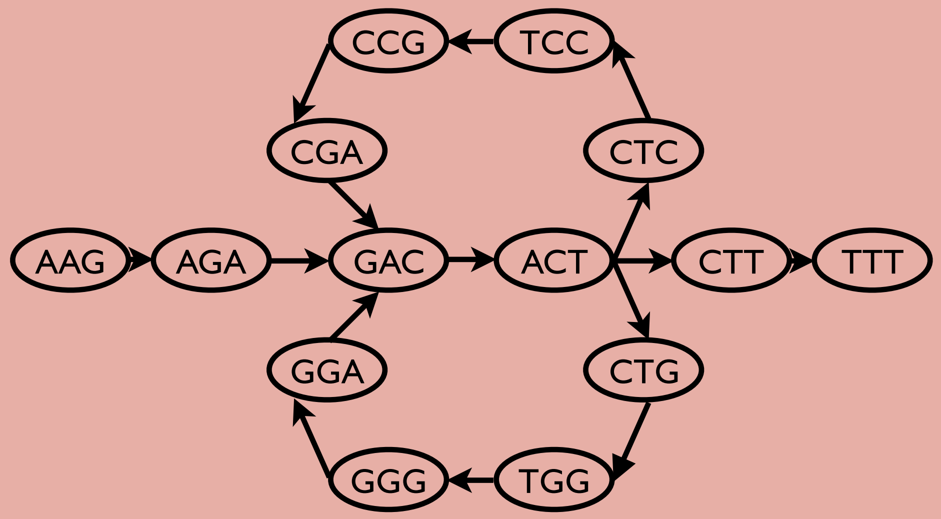


Hohenlohe et al. 2012. Philosophical Transactions of the Royal Society of London

# Could genome rearrangements in the stickleback genome be affecting these patterns?

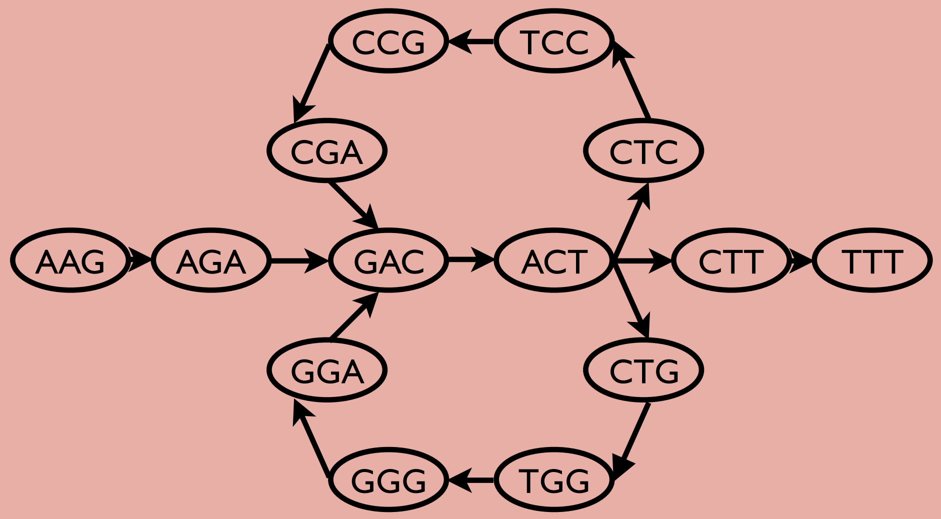
Julian Catchen, Susie Bassham and Kate Ituarte

# Genome Assembly

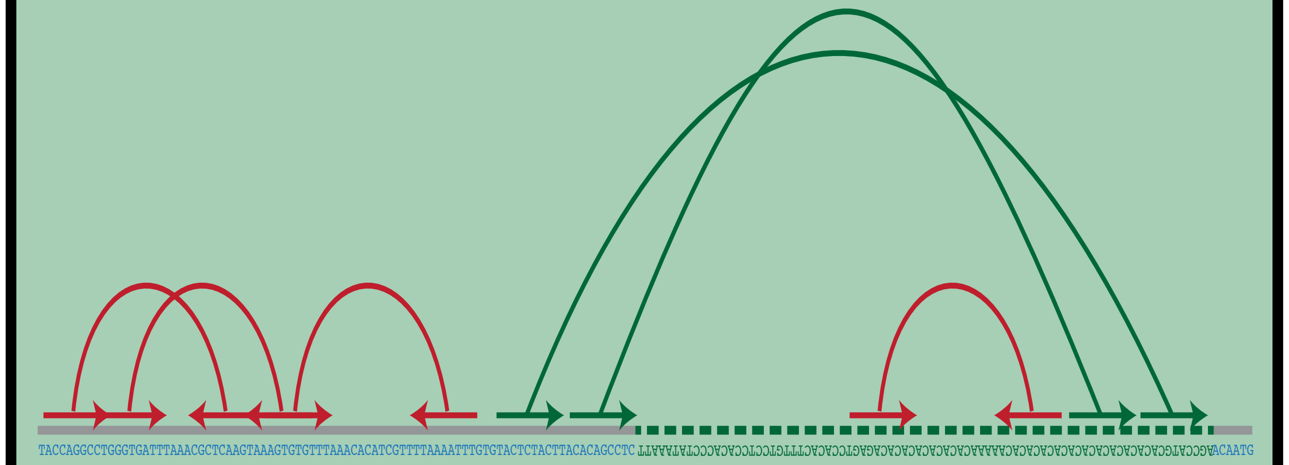


| | ♂ | ♀ | ♂ | ♀ |
|---|---|---|---|---|
| N50 | 17,417 bp | 18,982 bp | 15,555 bp | 15,534 bp |
| Max | 199,905 bp | 192,283 bp | 238,768 bp | 254,734 bp |
| Total | 488.8 Mb | 472.5 Mb | 456.4 Mb | 473.4 Mb |
| Median Coverage | 24.6x | 26.5x | 24.1x | 25.8x |

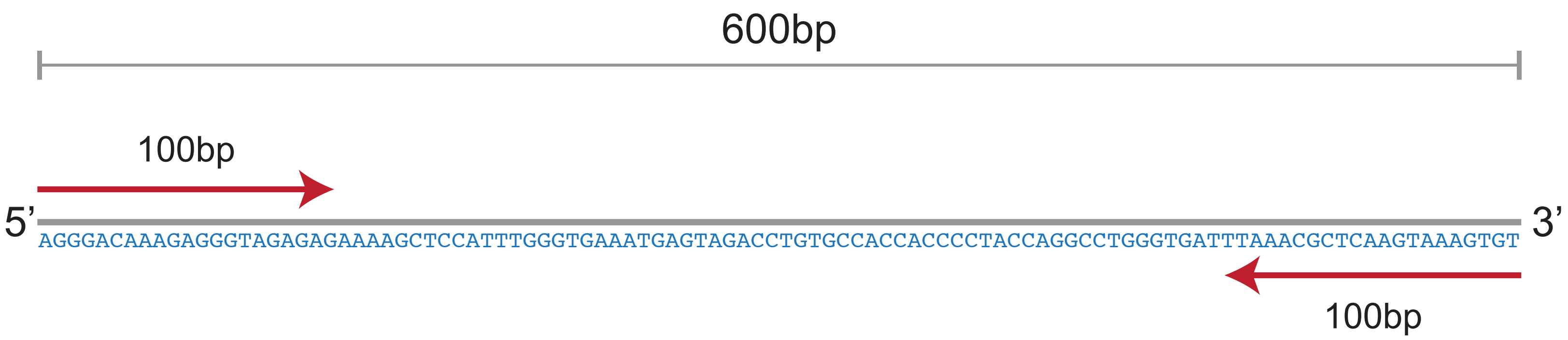
## Genome Assembly



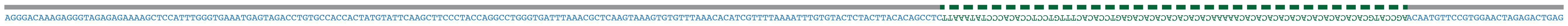
## Paired-end Alignments


TACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTC TTAATATCCCACACCTCCTGTTTCACACCTGAGACAGACACACACACACACAAAACACACACACACACACACACGTACGCATGCCGAACAATG

# Ilumina Paired-end Reads



600bp
100bp
5’
AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGT
3’
100bp

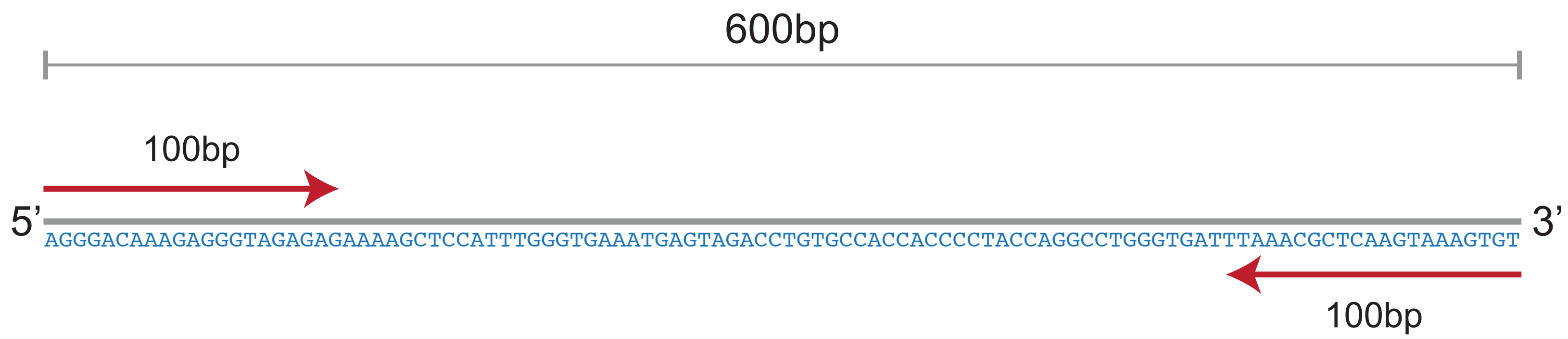


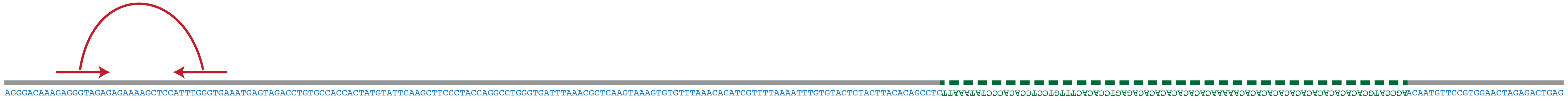
AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACTATGTATTCAAGCTTCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTCTTAATATCCACACCTCCTGTTTCACTCCAGTCCAGCACACACACACACAAACACACACACACACACACACACACACACGCATCCGGAACAATGTTCCGTGGAACTAGAGACTGAG

# Reference Genome

| | | | |
|---|---|---|---|
| 161,305,595 pairs | 144,396,898 pairs | 131,471,548 pairs | 150,786,462 pairs |

# Illumina Paired-end Reads

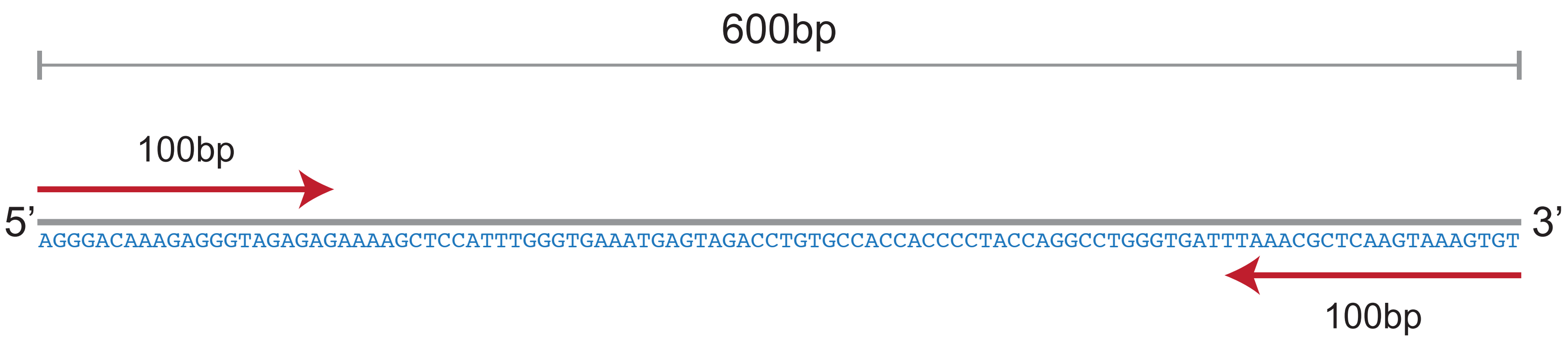

600bp
100bp
5’
3’
AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGT
100bp

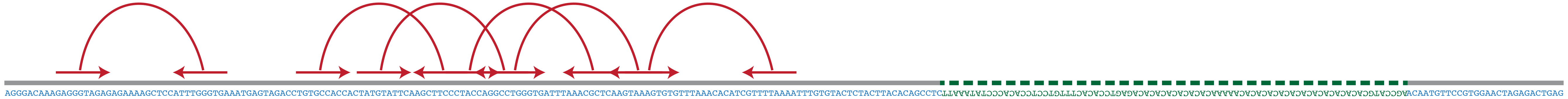

AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACTATGTATTCAAGCTTCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTCTTAAATATCCCACACCTCCTGTTTCACTCCTGAGACACACACACACACAAAACACACACACACACACACACACAGCATGCACCGAACAATGTTCCGTGGAACTAGAGACTGAG

## Reference Genome

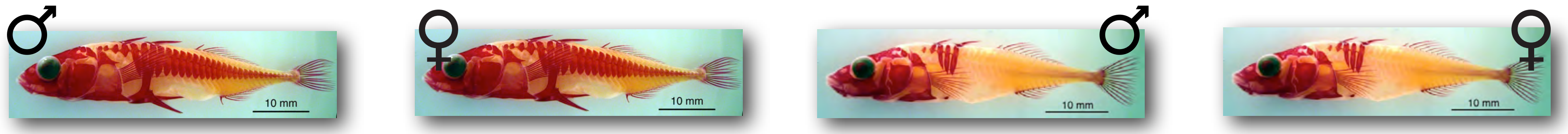
| | | | |
|---|---|---|---|
| 161,305,595 pairs | 144,396,898 pairs | 131,471,548 pairs | 150,786,462 pairs |

# Illumina Paired-end Reads


600bp
100bp
5’
AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGT
3’
100bp

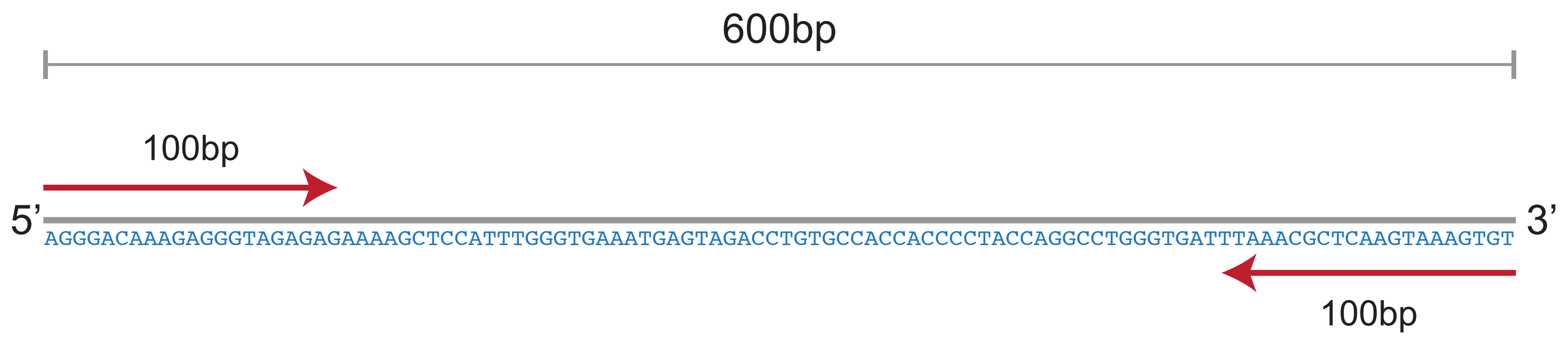

AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACTATGTATTCAAGCTTCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTC
ACAATGTTCCGTGGAACTAGAGACTGAG

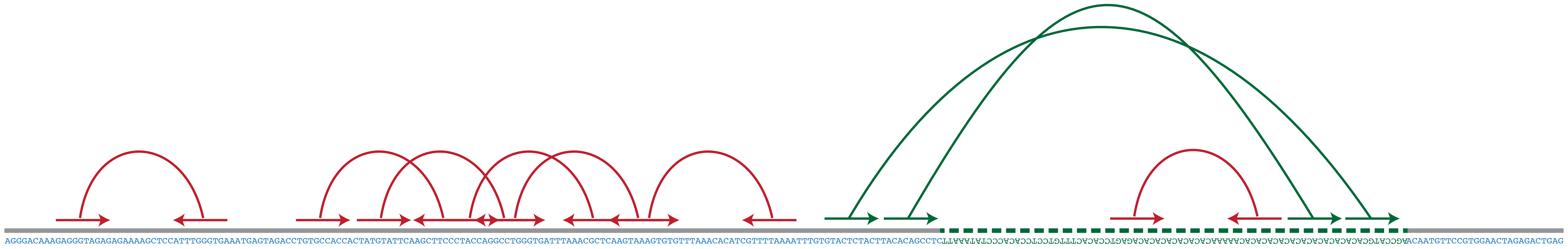
## Reference Genome



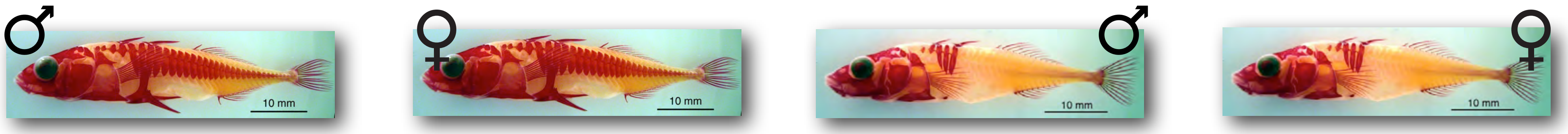
| | | | |
|---|---|---|---|
| 161,305,595 pairs | 144,396,898 pairs | 131,471,548 pairs | 150,786,462 pairs |

# Illumina Paired-end Reads


600bp
100bp
5’
AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGT
3’
100bp

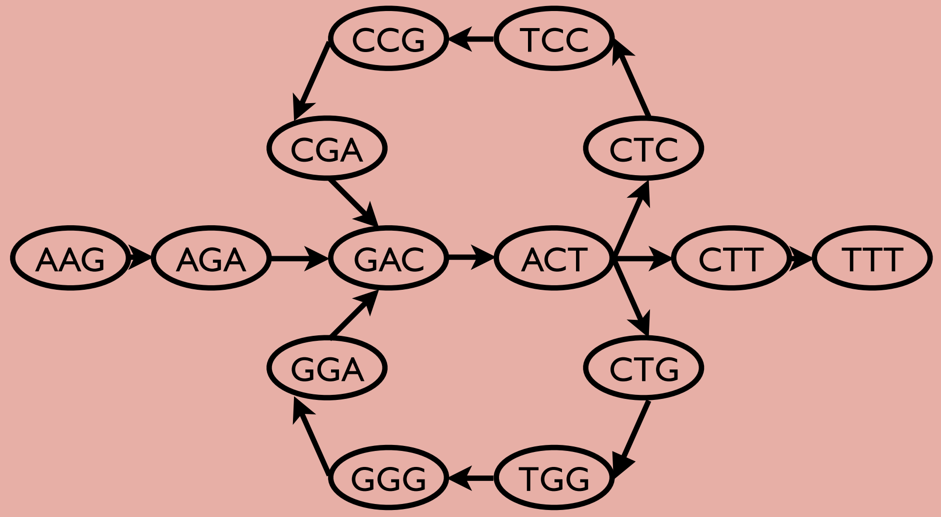

AGGGACAAAGAGGGTAGAGAGAAAAGCTCCATTTGGGTGAAATGAGTAGACCTGTGCCACCACTATGTATTCAAGCTTCCCTACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTC
ACAATGTTCCGTGGAACTAGAGACTGAG

Reference Genome

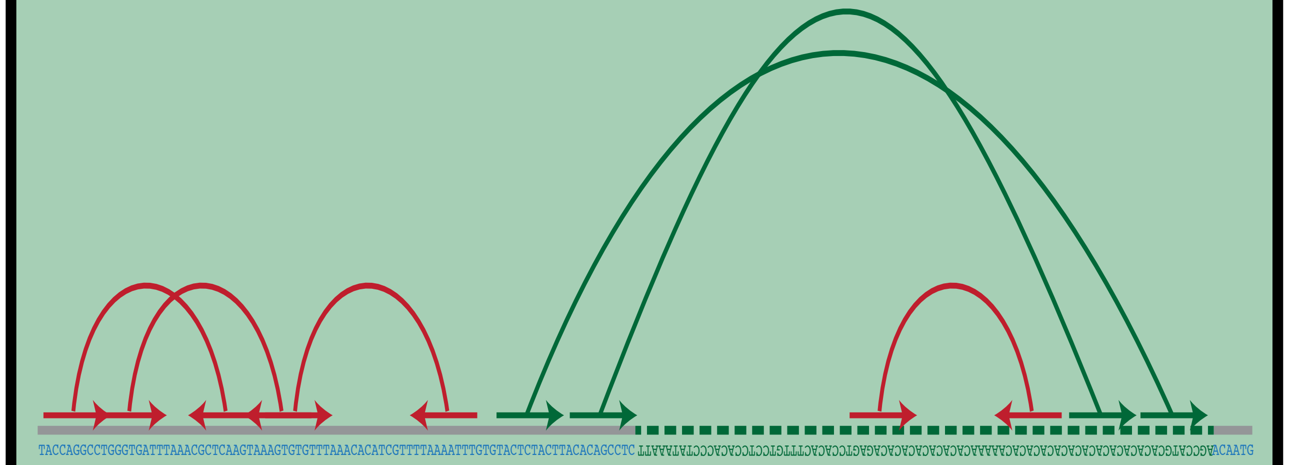


| | | | |
|---|---|---|---|
| 161,305,595 pairs | 144,396,898 pairs | 131,471,548 pairs | 150,786,462 pairs |

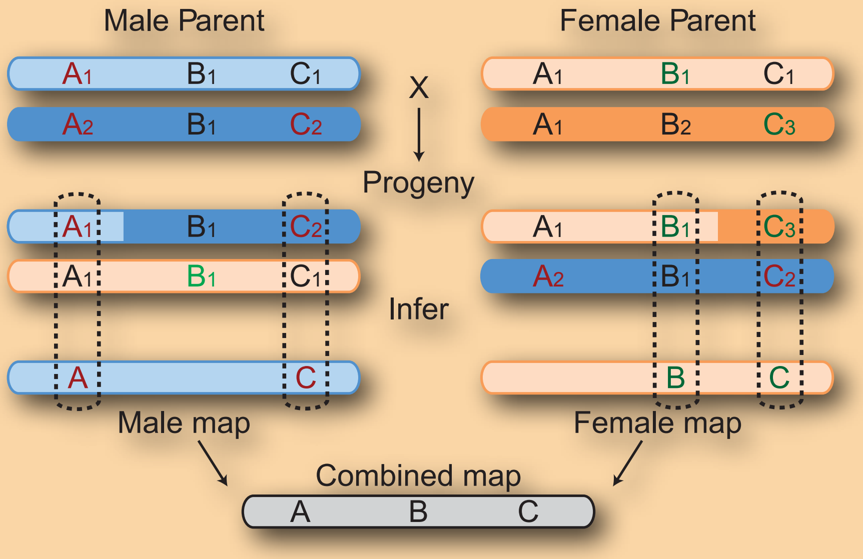
## Genome Assembly



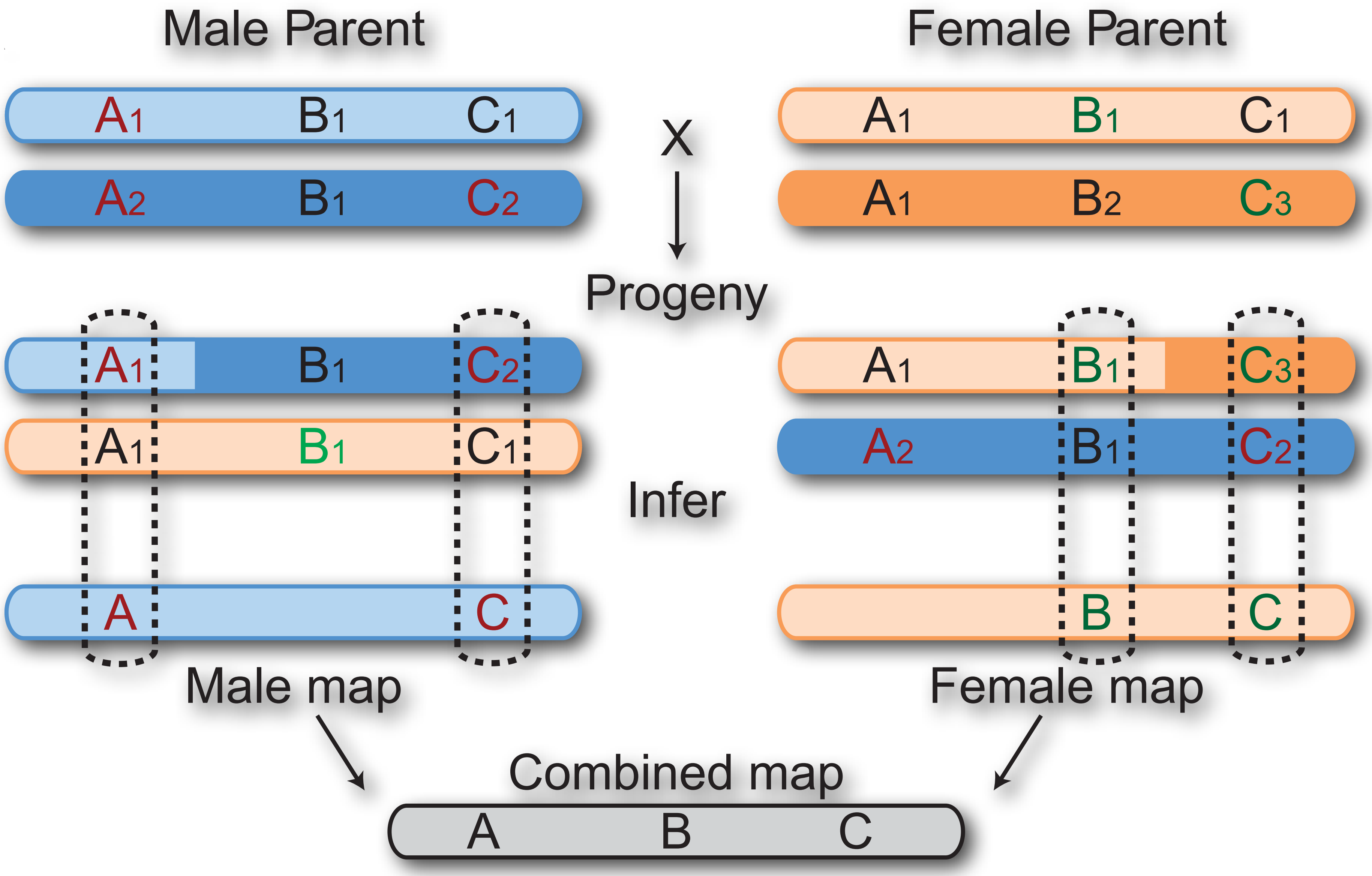
## Paired-end Alignments


TACCAGGCCTGGGTGATTTAAACGCTCAAGTAAAGTGTGTTTAAACACATCGTTTTAAAATTTGTGTACTCTACTTACACAGCCTC
TTAAATATCCCACACCTCCTGTTTCACTCCTGAGACACACACACACAAAACACACACACACACACACACACACGTACCGA
ACAATG

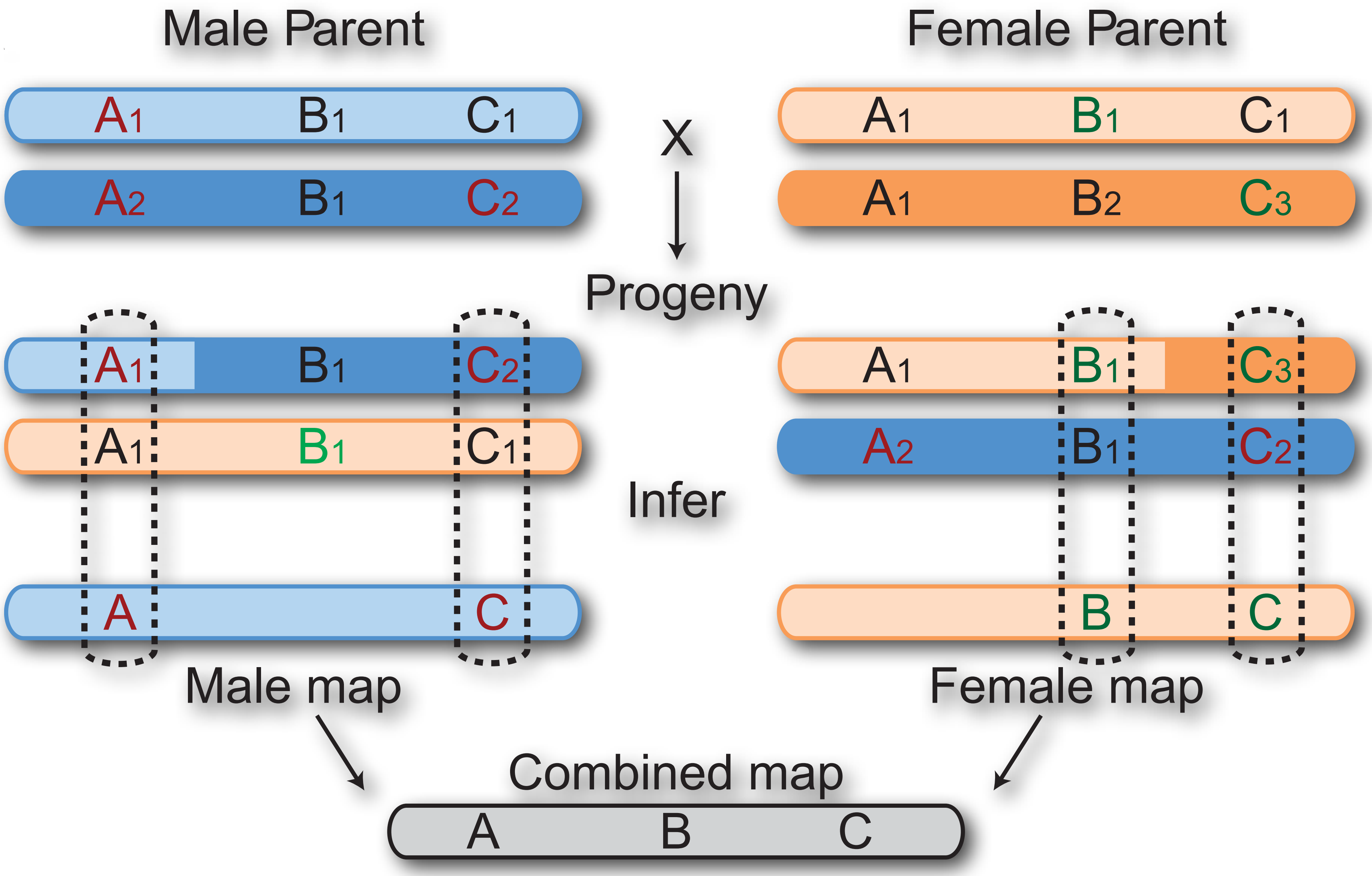
## Genetic Map Construction

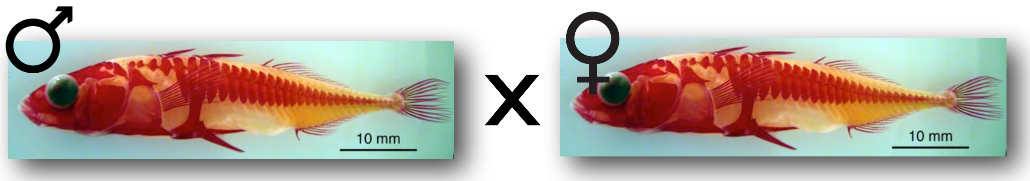


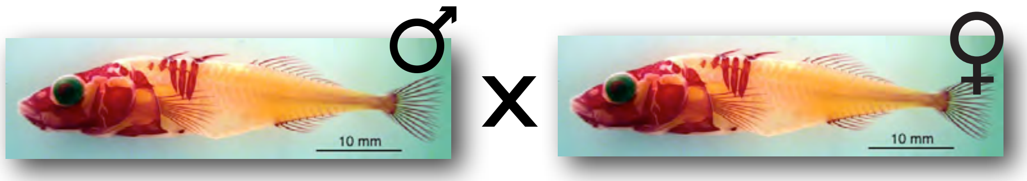
# F1 Pseudo-testcross



# F1 Pseudo-testcross



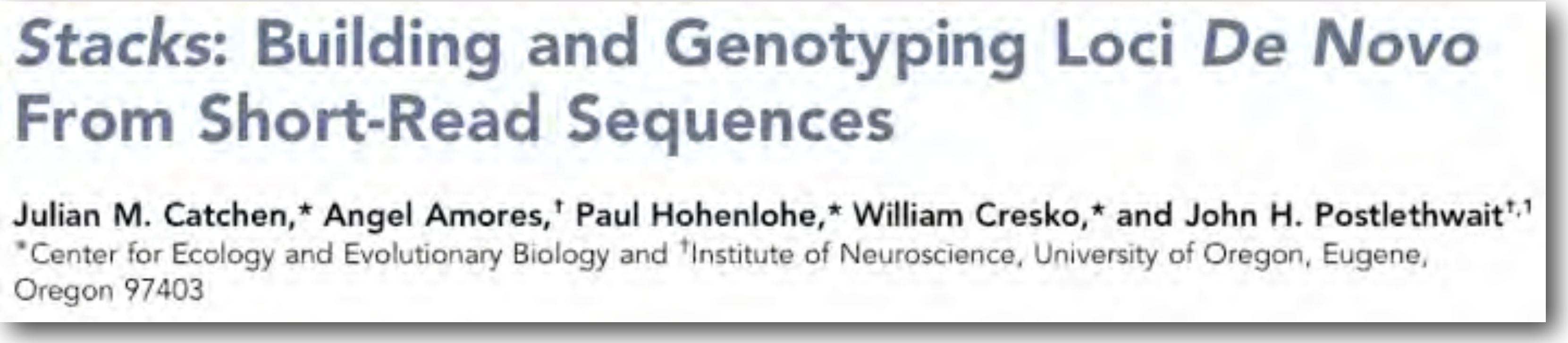




| |
|---|
| 93 progeny |
| 66,071 loci |
| 5,351 markers |

| |
|---|
| 93 progeny |
| 45,301 loci |
| 3,927 markers |

G3: Genes, Genomes, Genetics

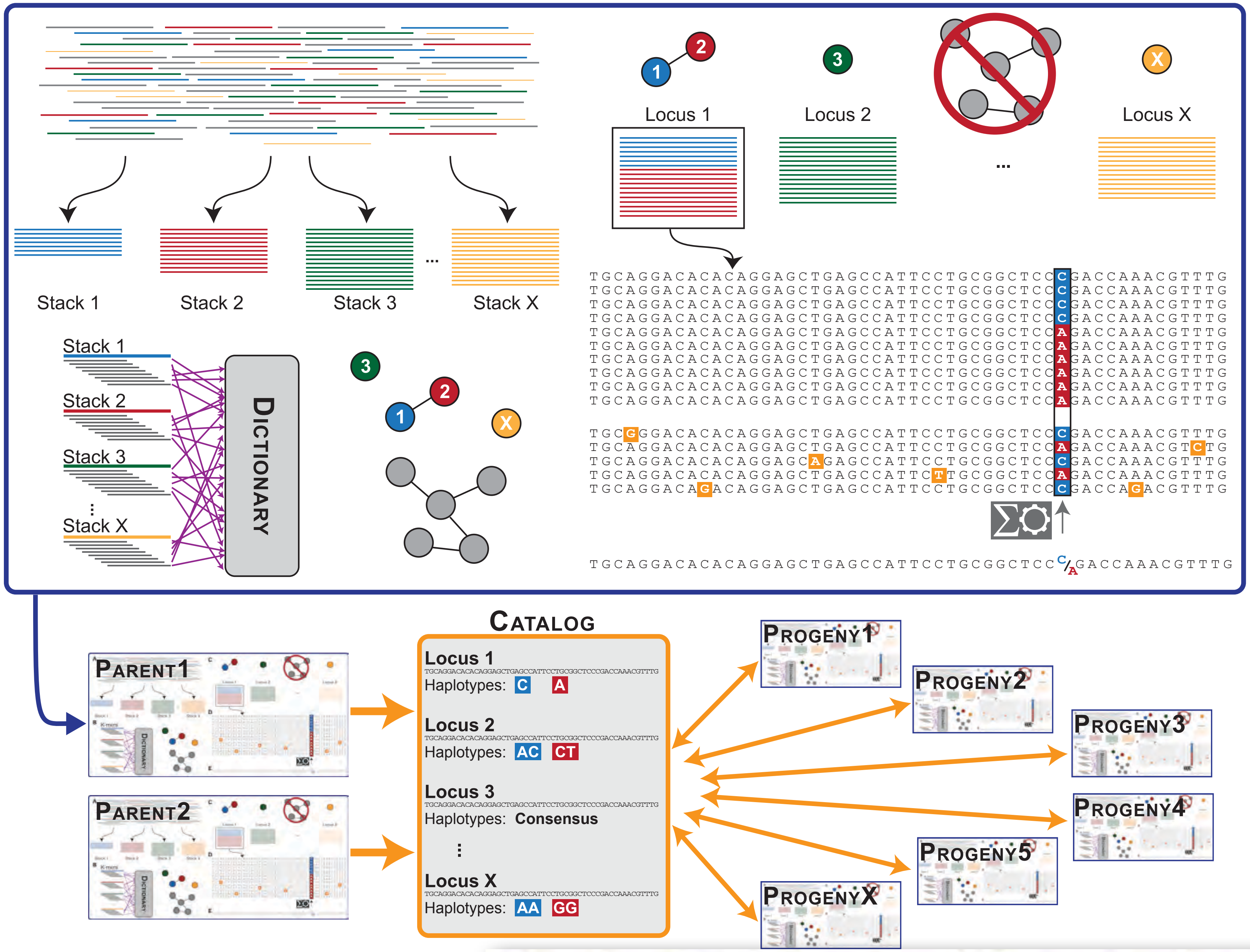


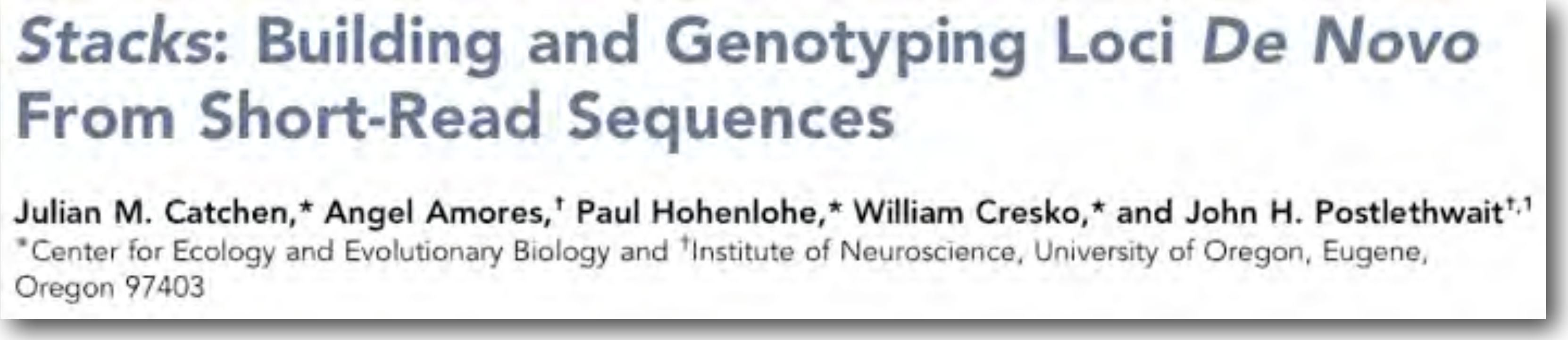
# *Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences

**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

# Stacks





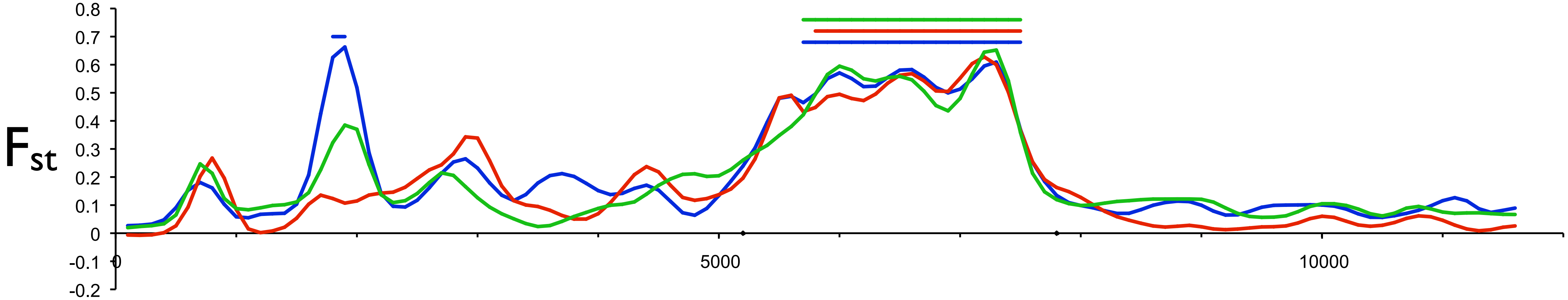
## *Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences

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

G3: Genes, Genomes, Genetics

♀

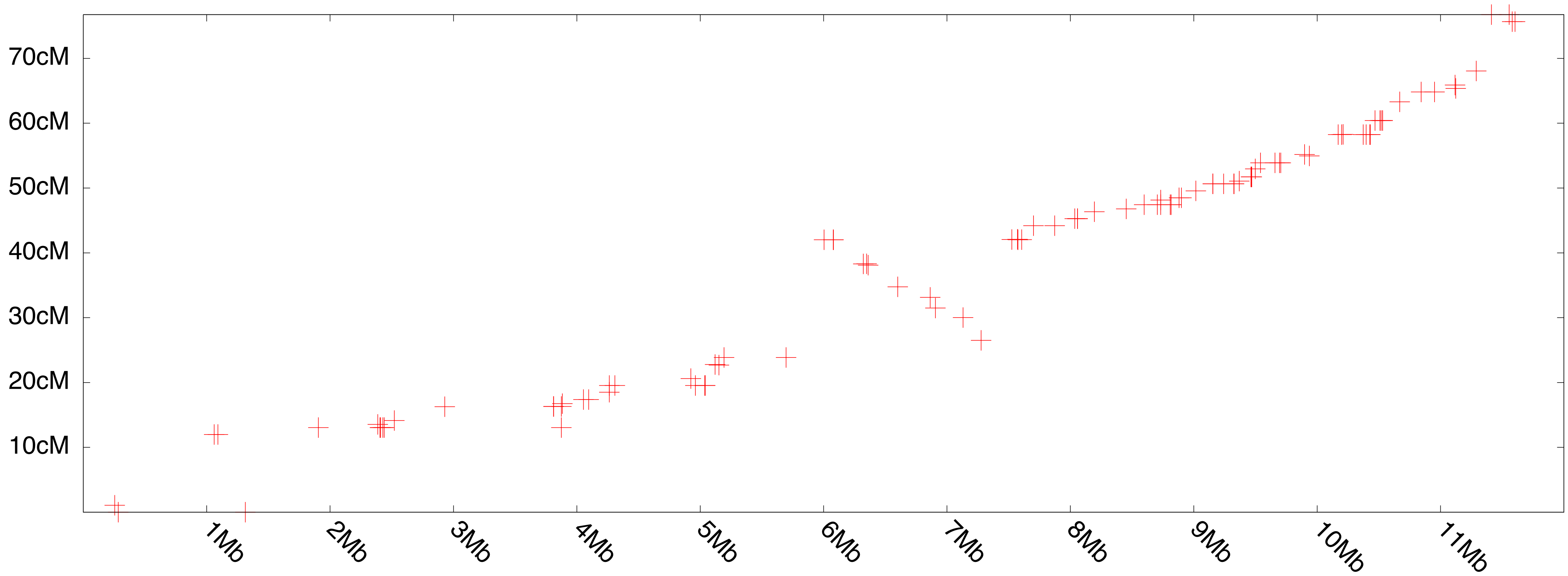
# LG21

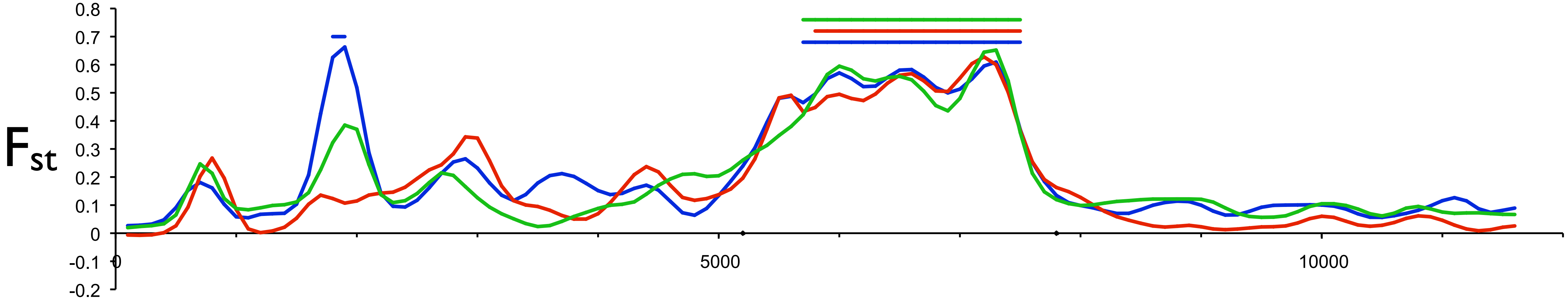


10 mm

♀

LG21

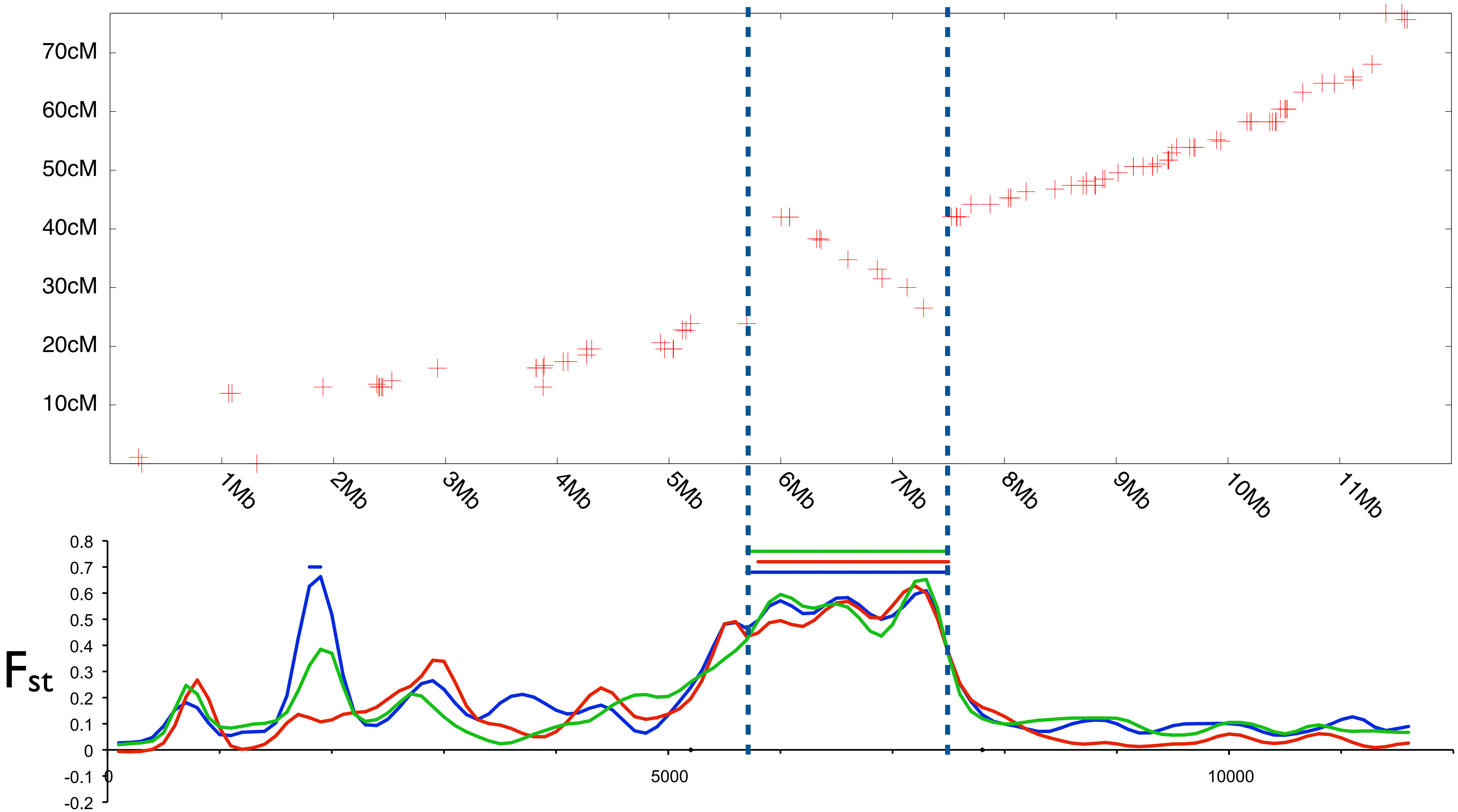

70cM
60cM
50cM
40cM
30cM
20cM
10cM
1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
9Mb
10Mb
11Mb

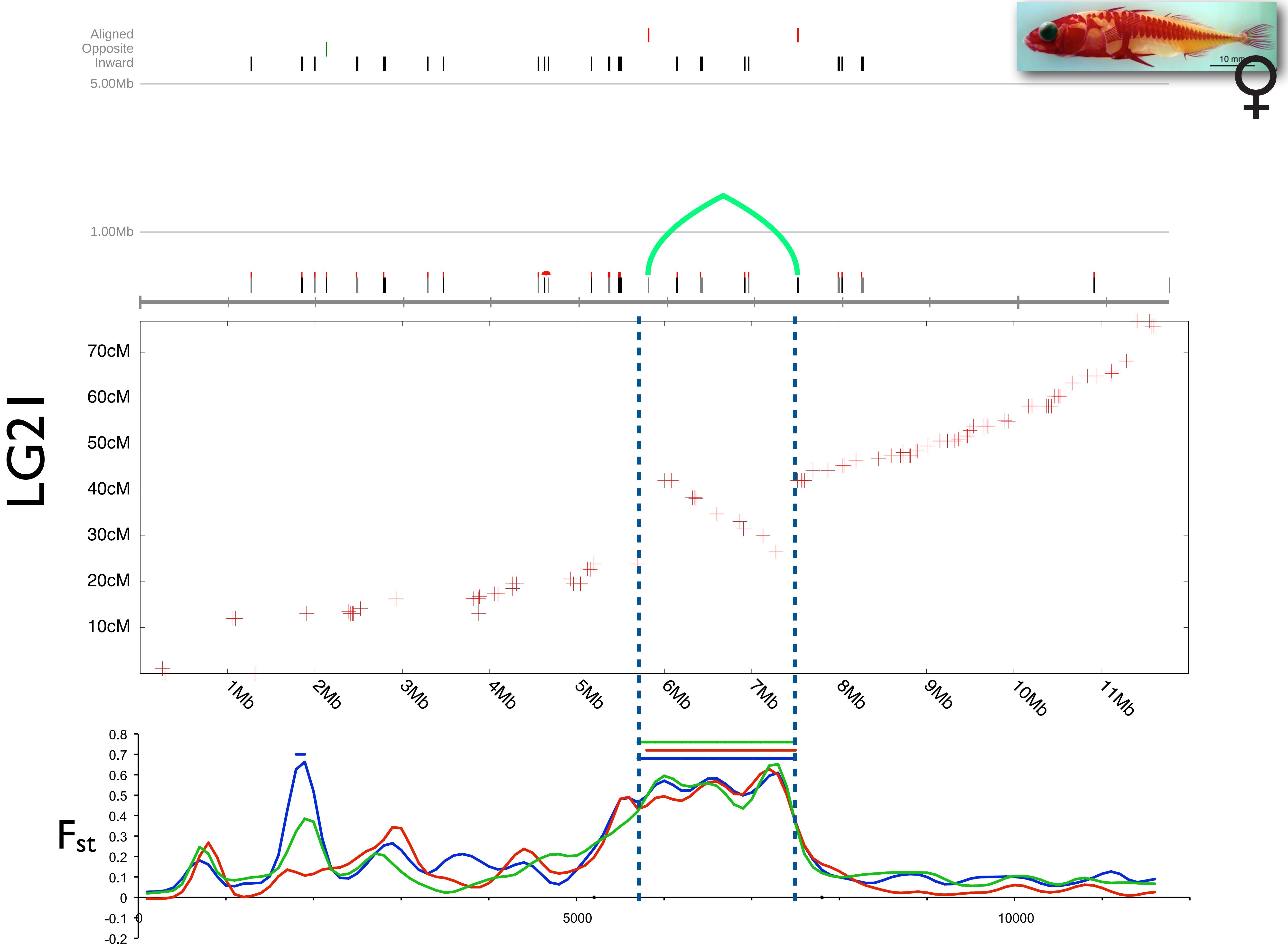

$F_{st}$
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
5000
10000

10 mm

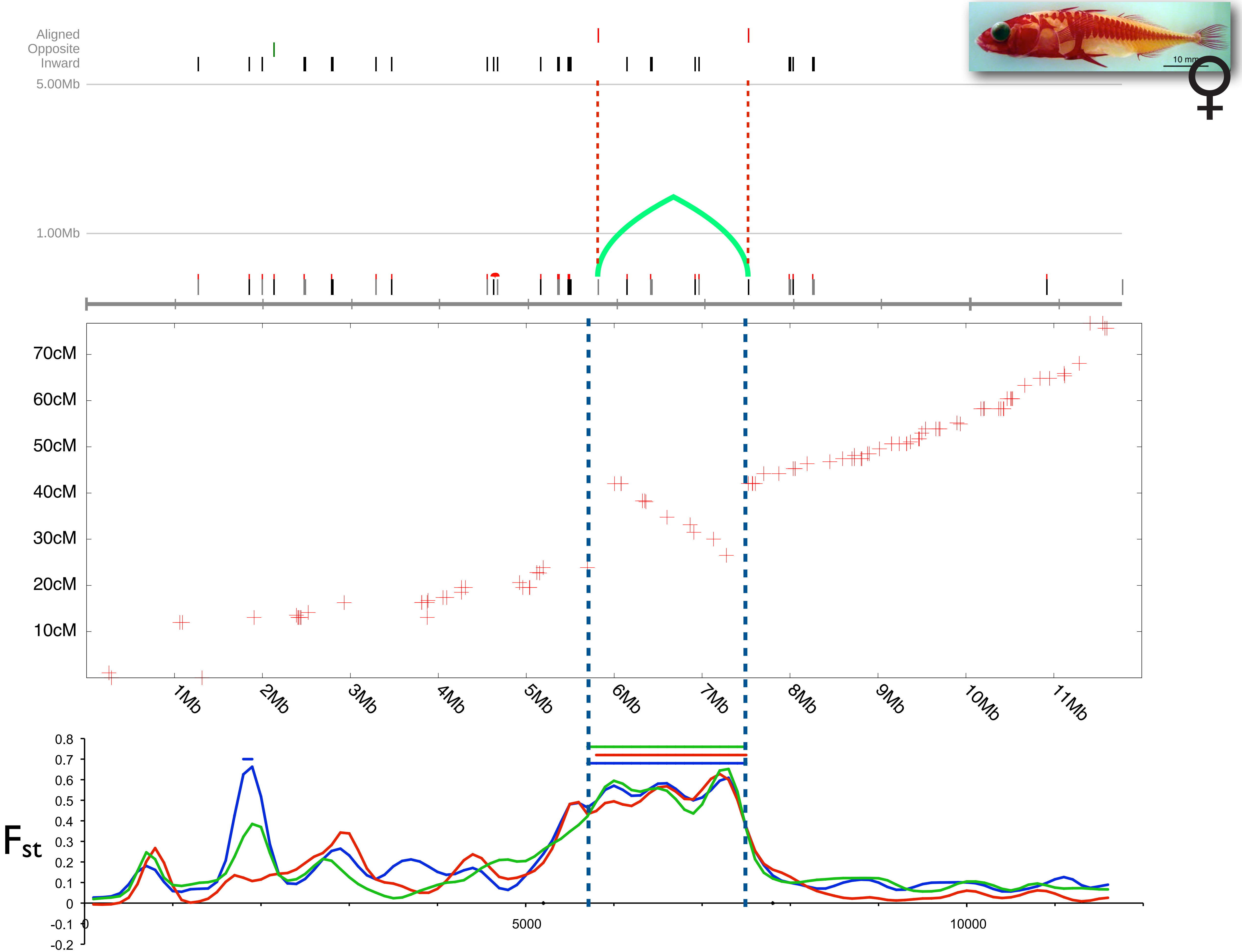
♀

LG21

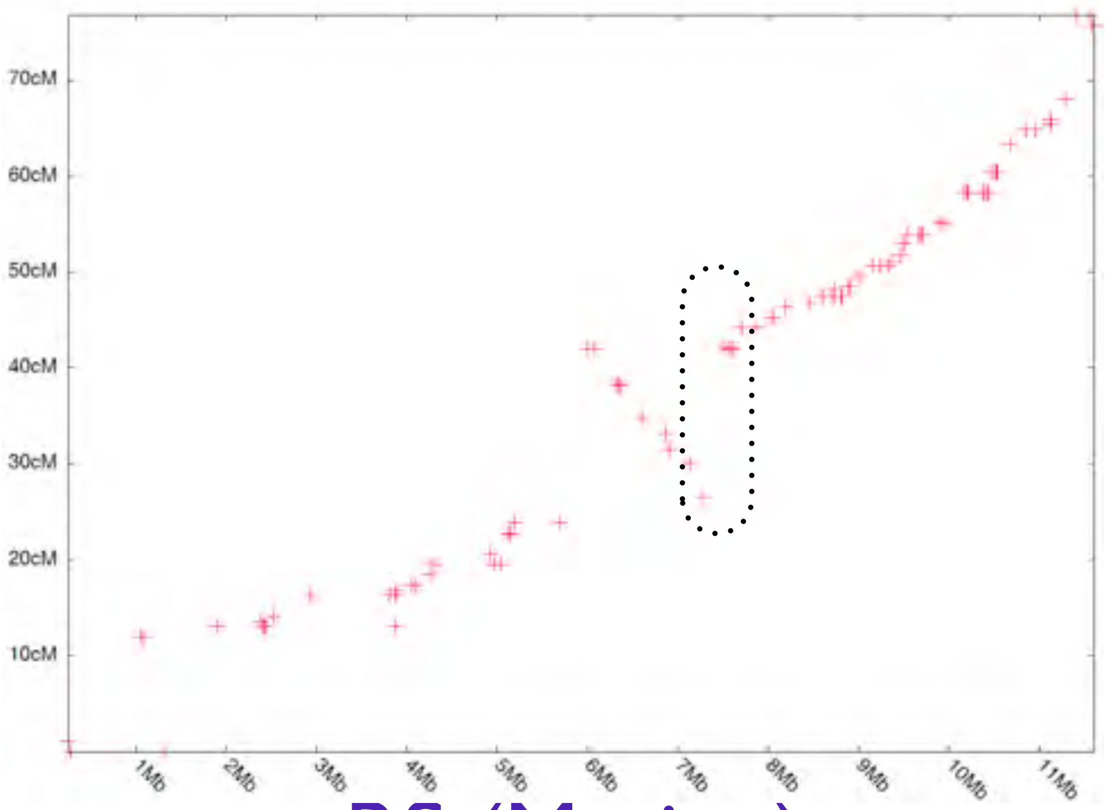

70cM
60cM
50cM
40cM
30cM
20cM
10cM
1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
9Mb
10Mb
11Mb
Fst
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
5000
10000


Aligned
Opposite
Inward
5.00Mb
10 mm
♀
1.00Mb
LG21
70cM
60cM
50cM
40cM
30cM
20cM
10cM
1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
9Mb
10Mb
11Mb
$F_{st}$
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
5000
10000

LG21


Aligned
Opposite
Inward
5.00Mb
1.00Mb
10 mm
♀
70cM
60cM
50cM
40cM
30cM
20cM
10cM
1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
9Mb
10Mb
11Mb
$F_{st}$
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
5000
10000

# Linkage Group XXI

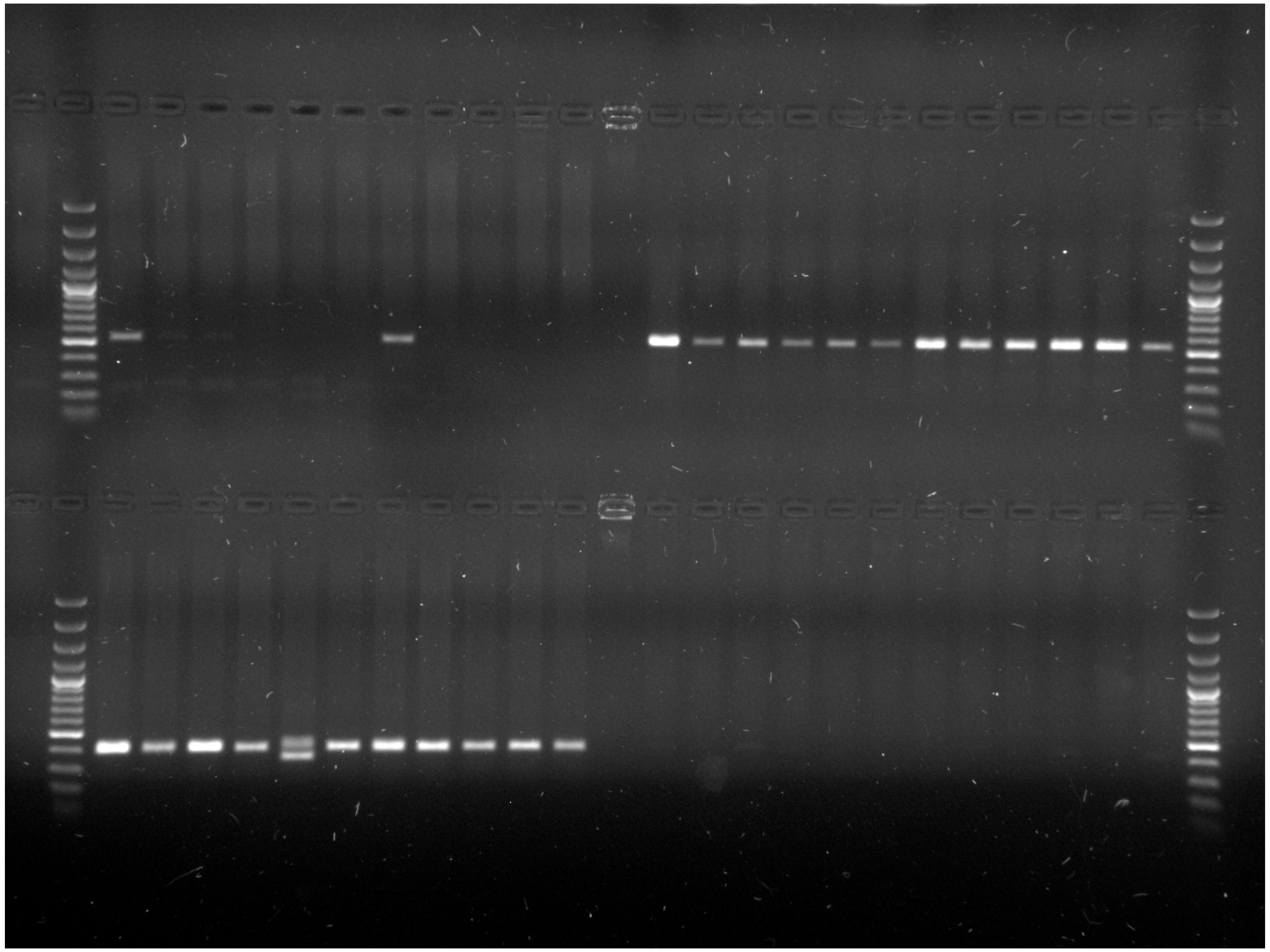


RS (Marine)

Boot

Like Bear Paw

Inverted



# How quickly does stickleback evolution occur?

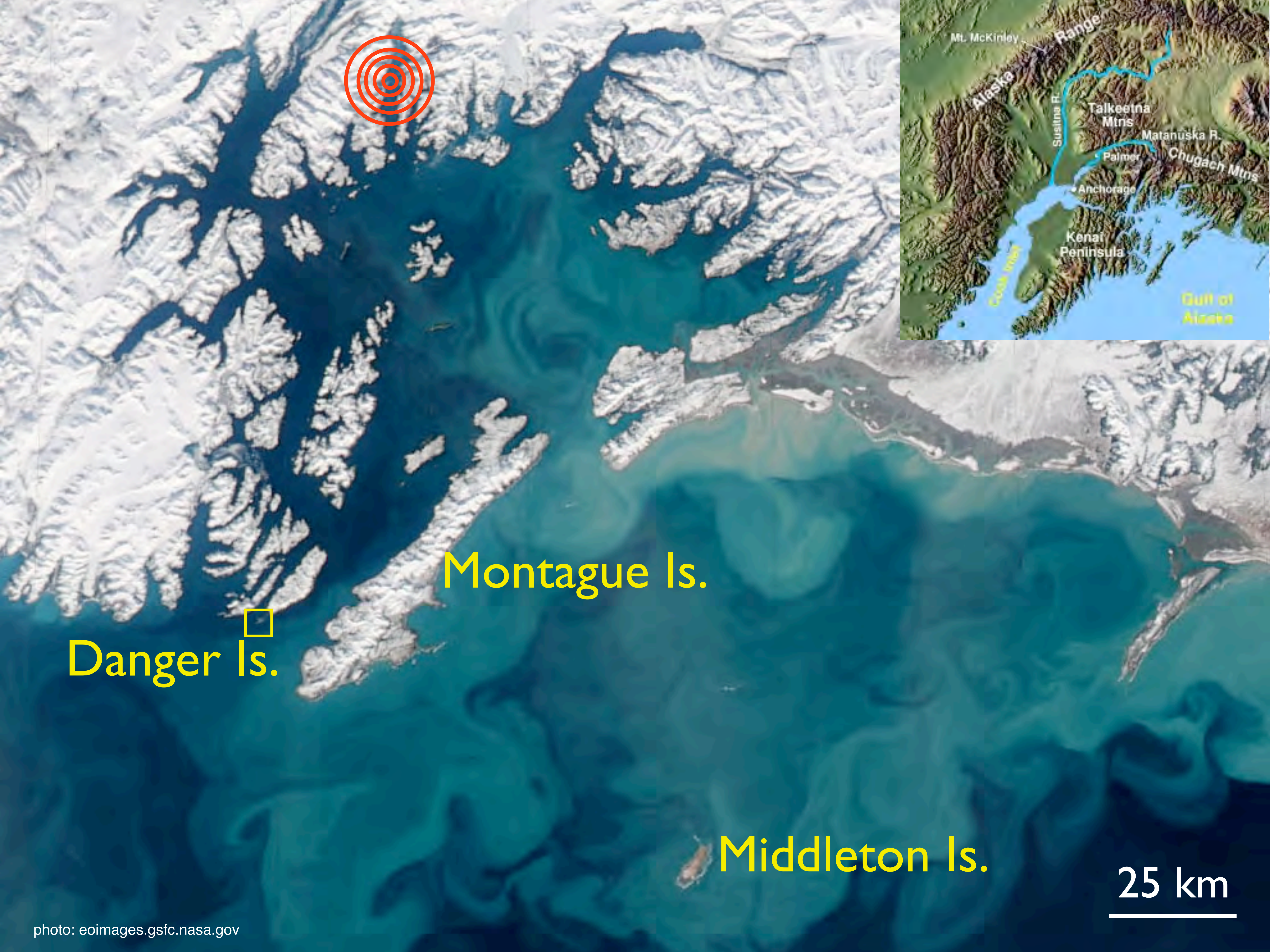


photo: eoimages.gsfc.nasa.gov

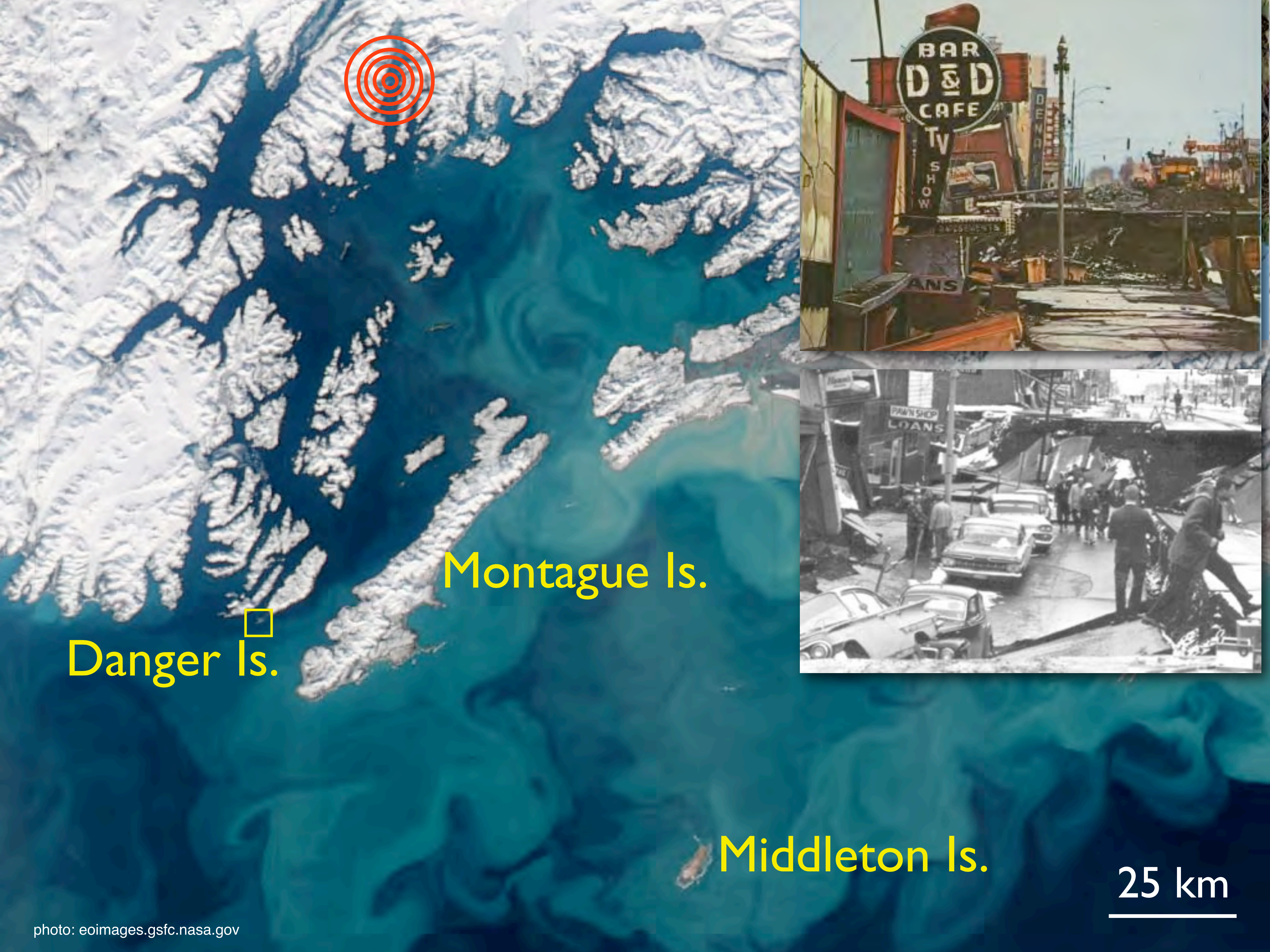
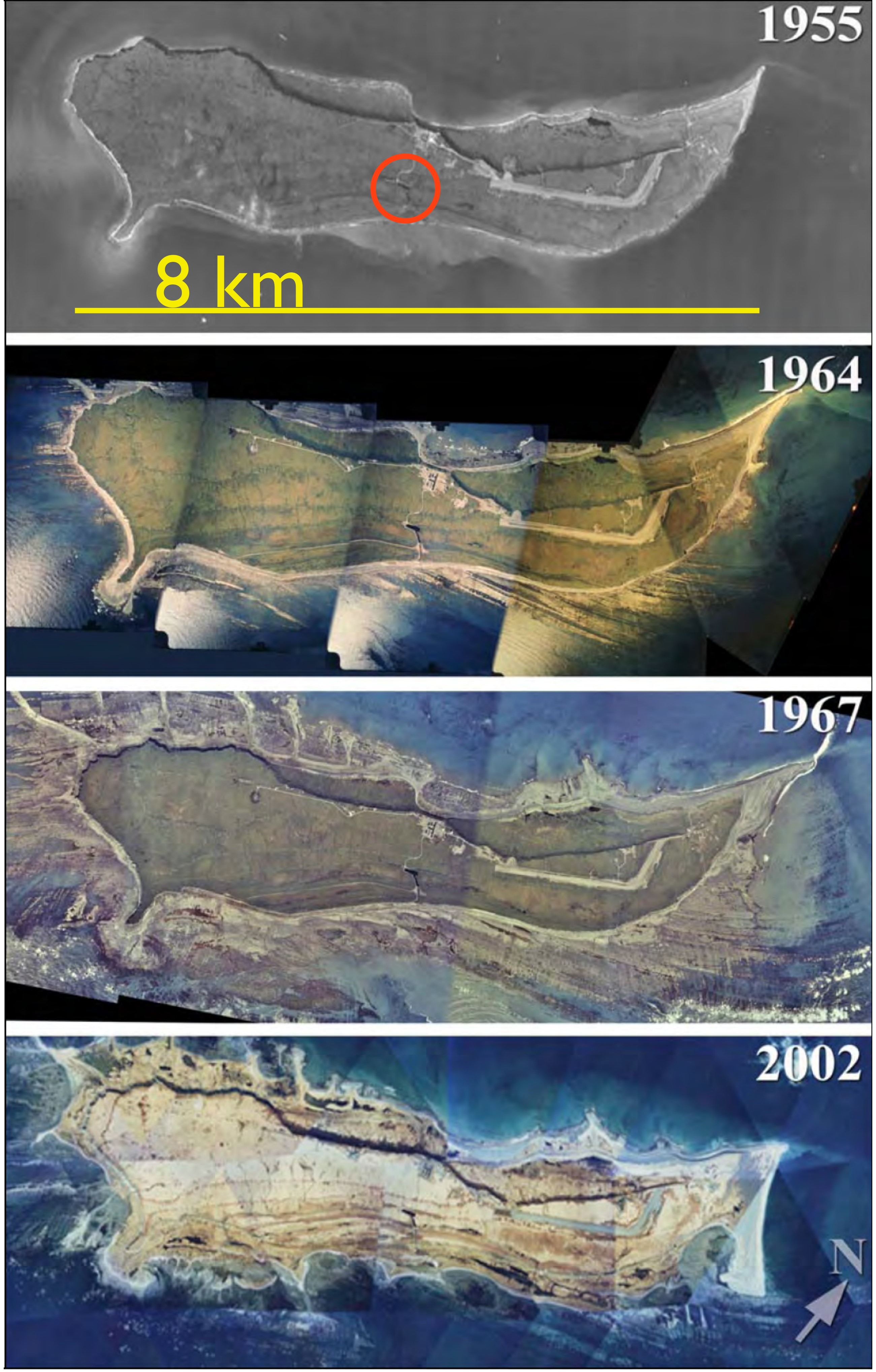


photo: eoimages.gsfc.nasa.gov



photos: BLM and Aero-Metric, Inc., Anchorage, AK.

743 fish
27,878 RAD loci
110,000 SNPs

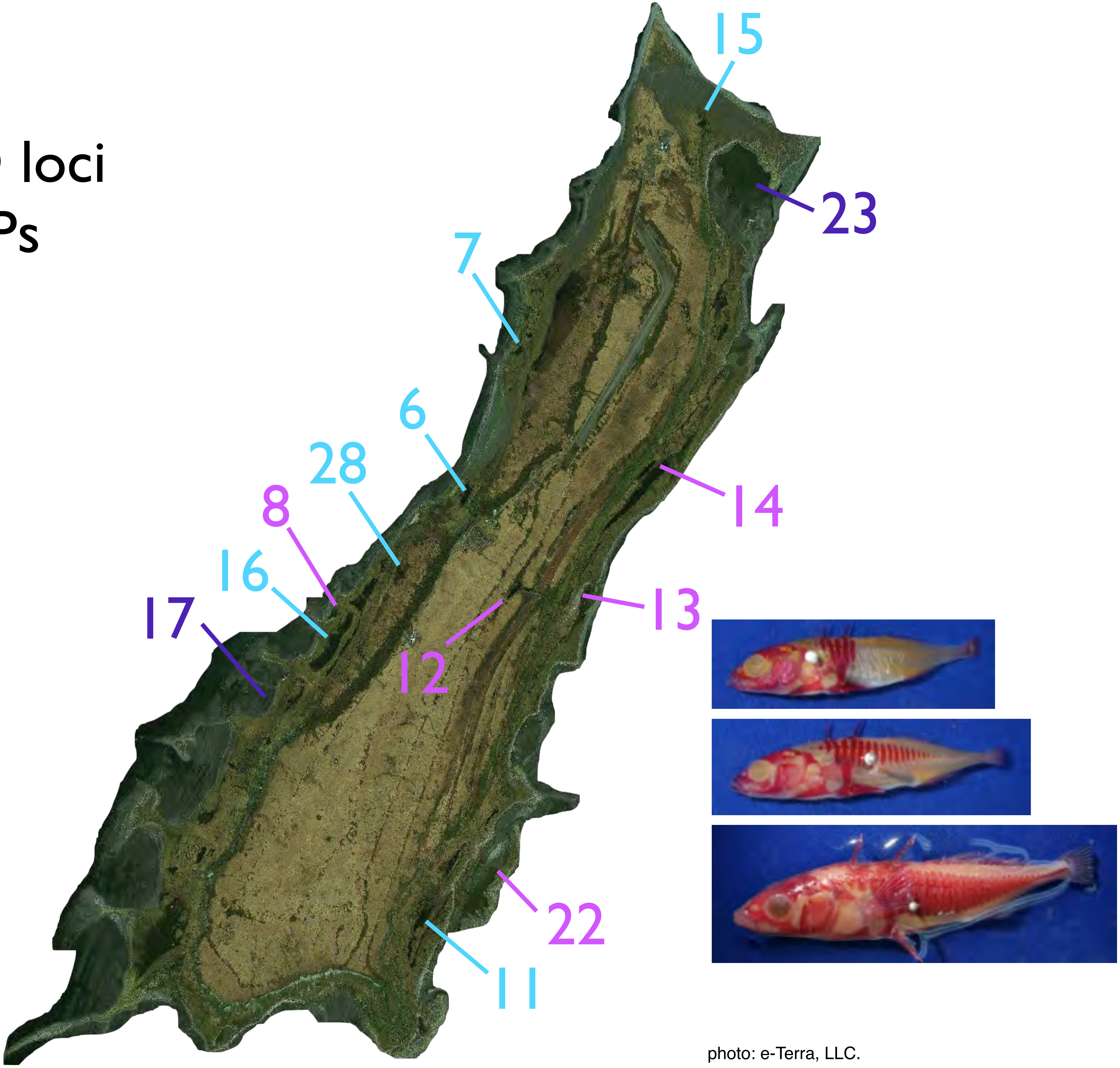


photo: e-Terra, LLC.

Marine x Marine $F_{ST}$ 0.001

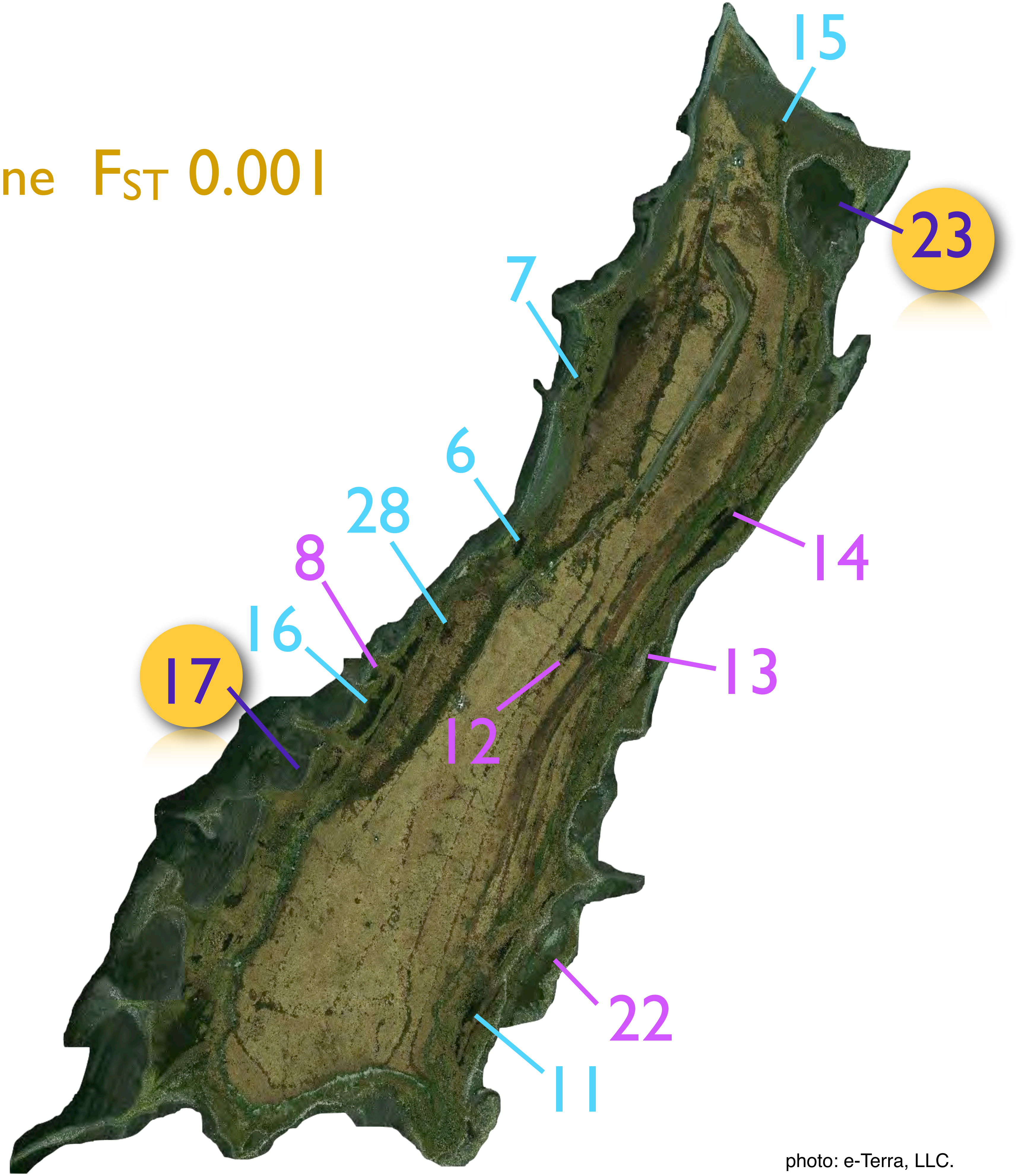


photo: e-Terra, LLC.

Marine x Marine $F_{ST}$ 0.001

Fresh x Fresh $F_{ST}$ 0.052

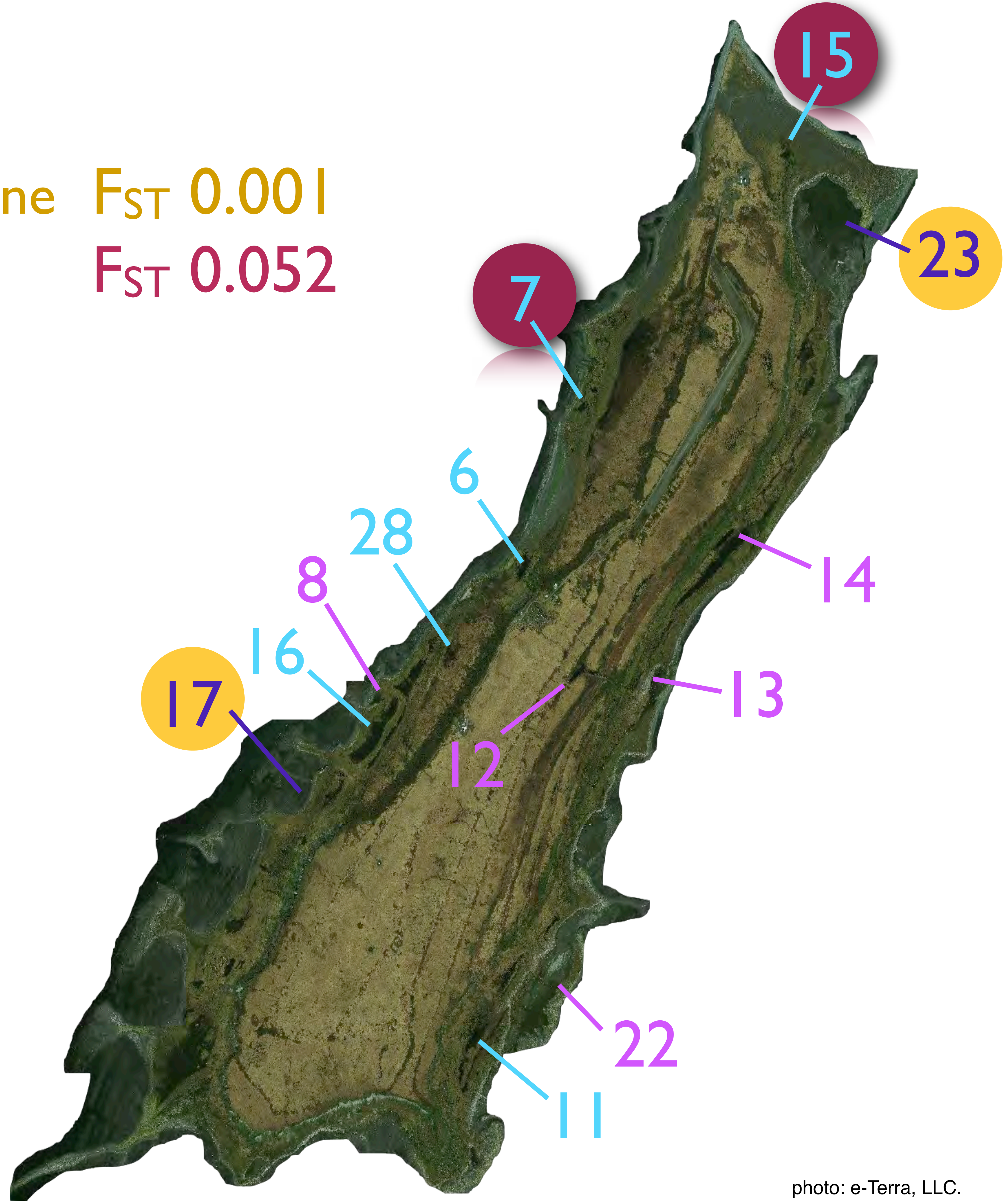


photo: e-Terra, LLC.

# NJ $F_{ST}$ tree

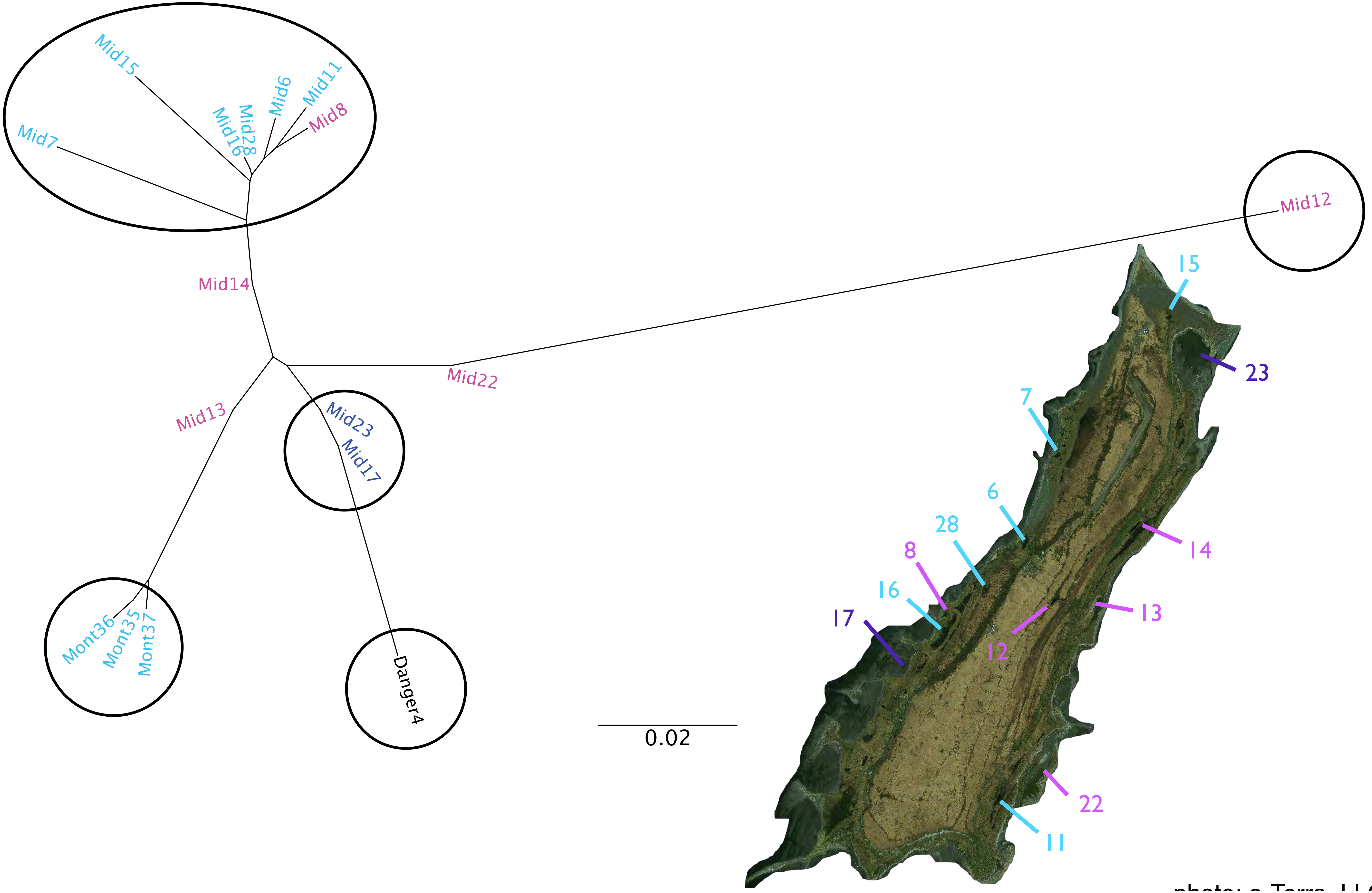
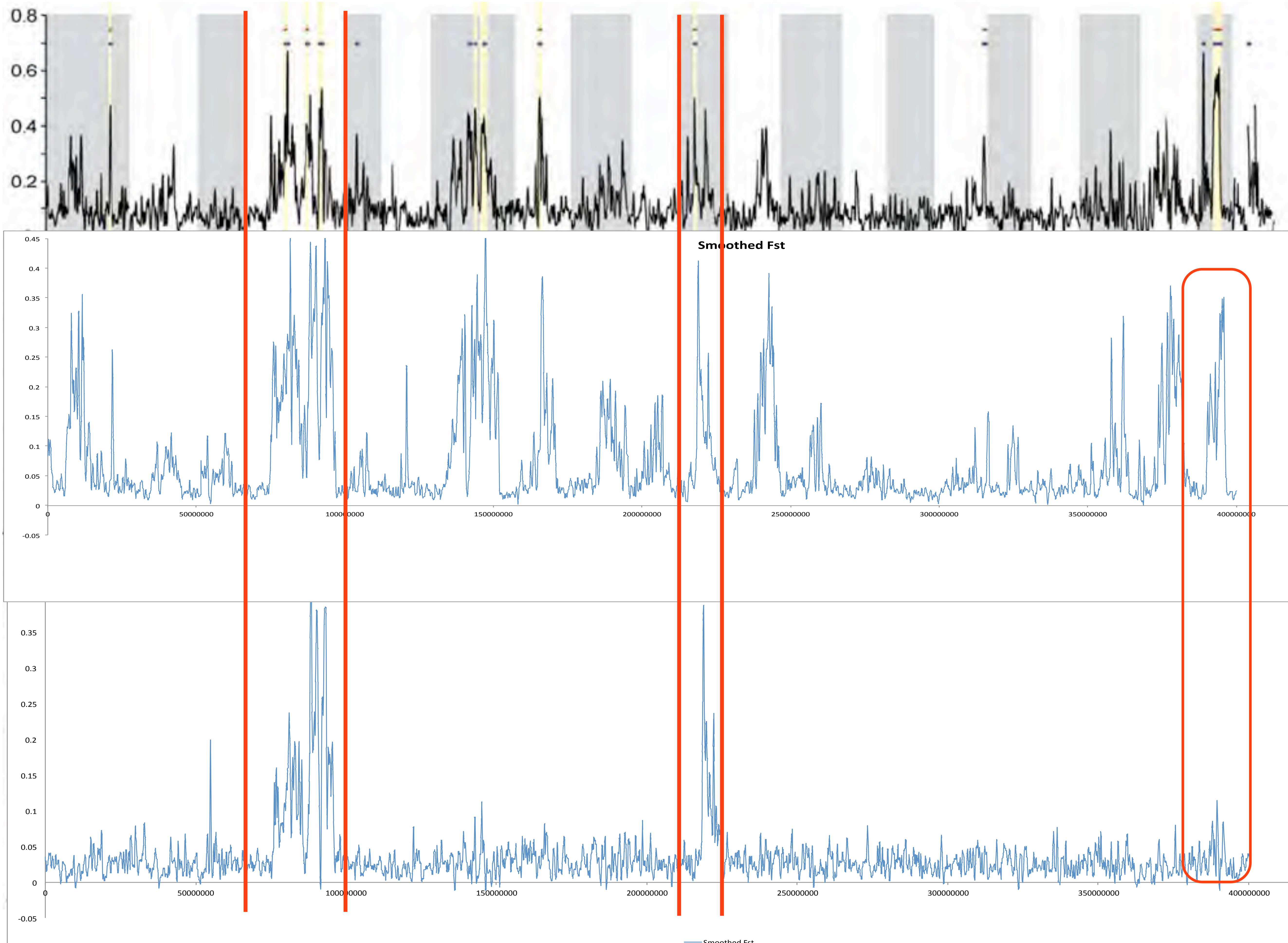
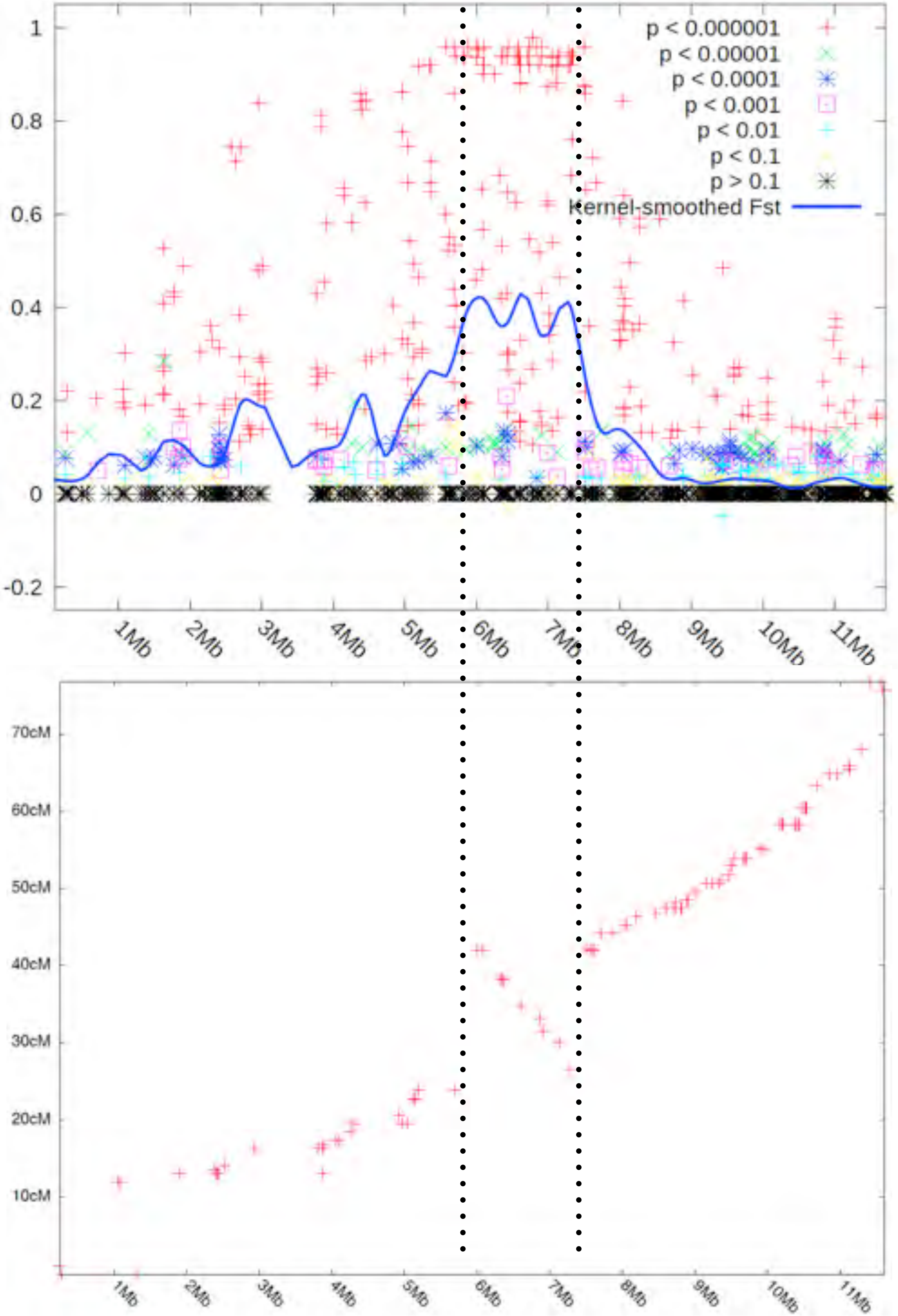


photo: e Terra, LL

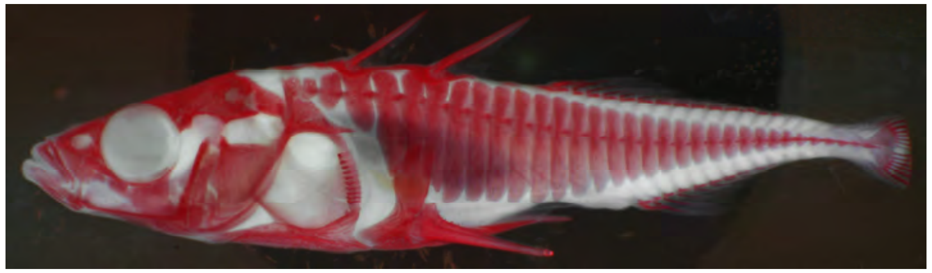

0.8
0.6
0.4
0.2
Smoothed Fst
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
50000000
100000000
150000000
200000000
250000000
300000000
350000000
400000000
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
50000000
100000000
150000000
200000000
250000000
300000000
350000000
400000000
Smoothed Fst

# Linkage Group XXI

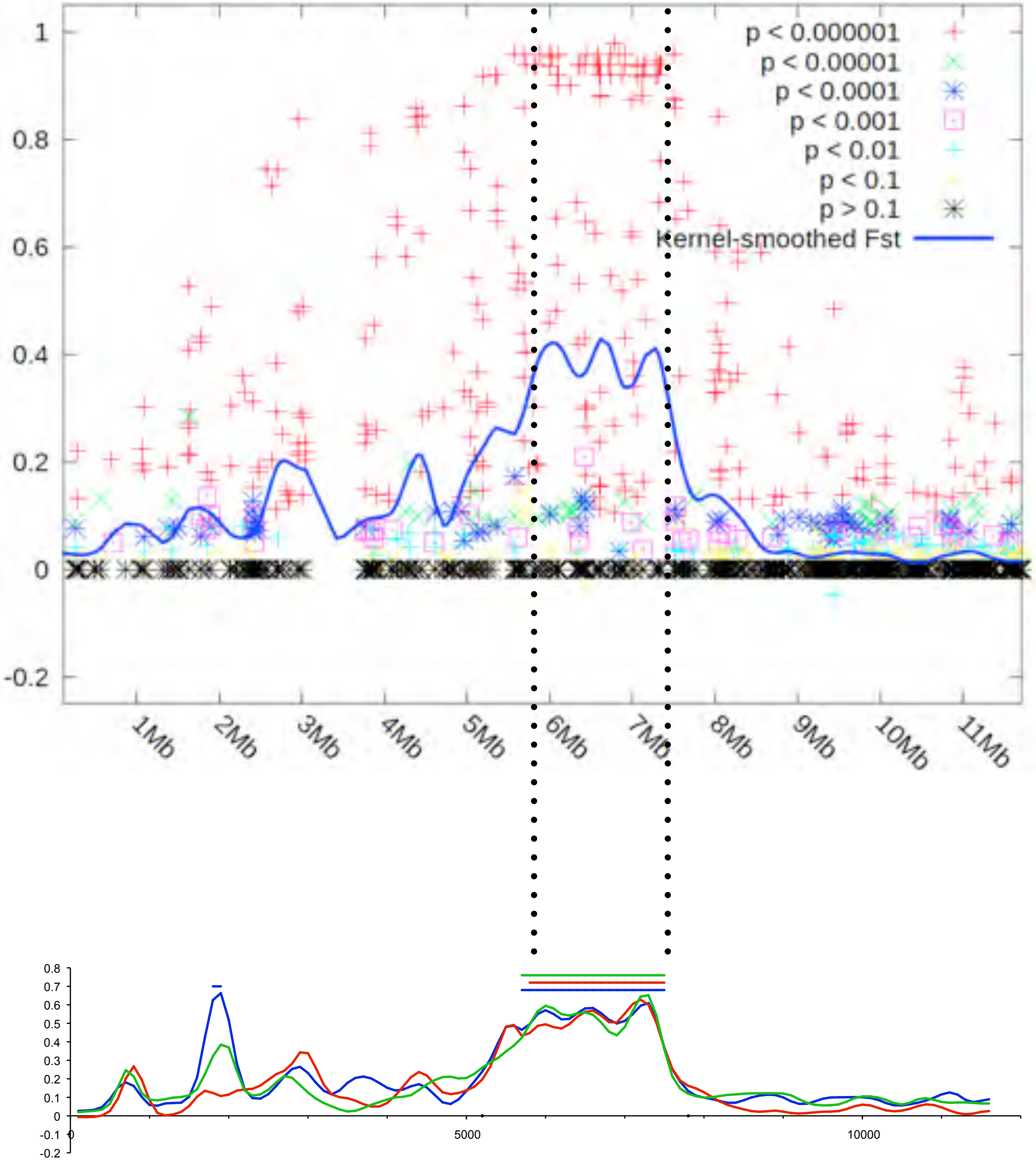


Fresh vs Marine

## Rabbit Sl x Rabbit Sl



# Linkage Group XXI



Fresh vs Marine

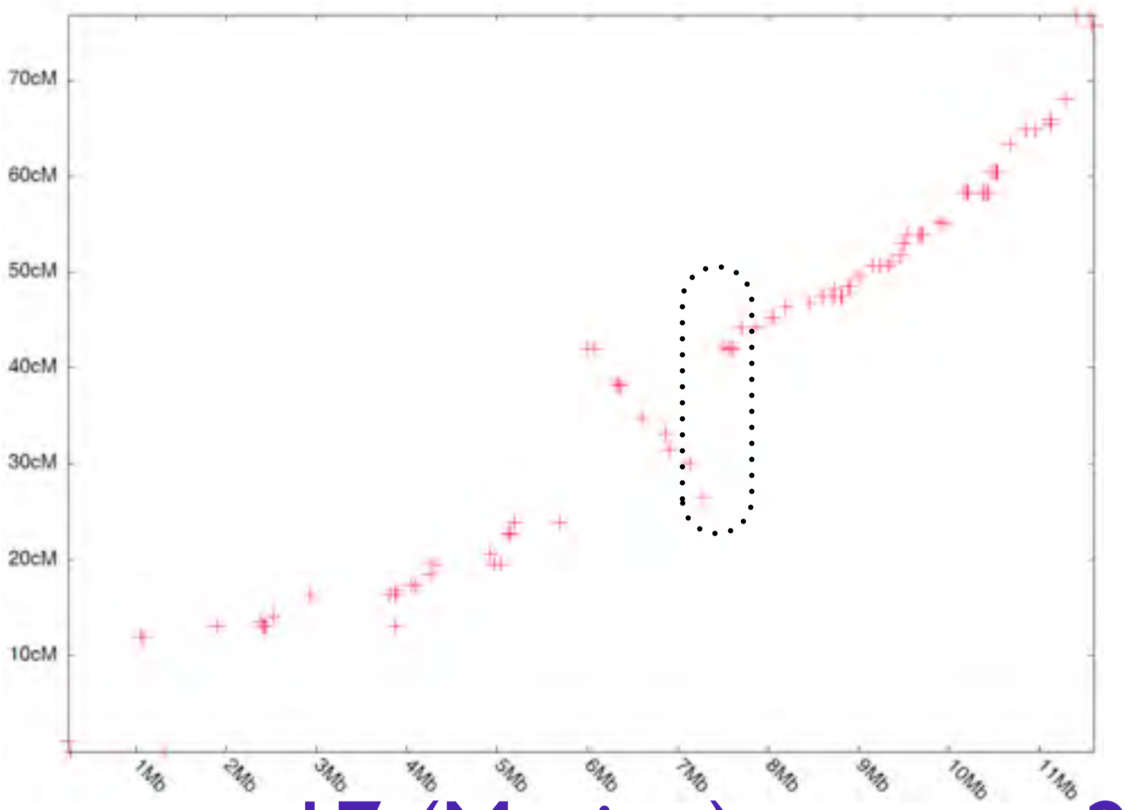
Bear Paw Lk

Boot Lk vs Marine

Mud Lk

(Hohenlohe, Bassham et al. 2010)

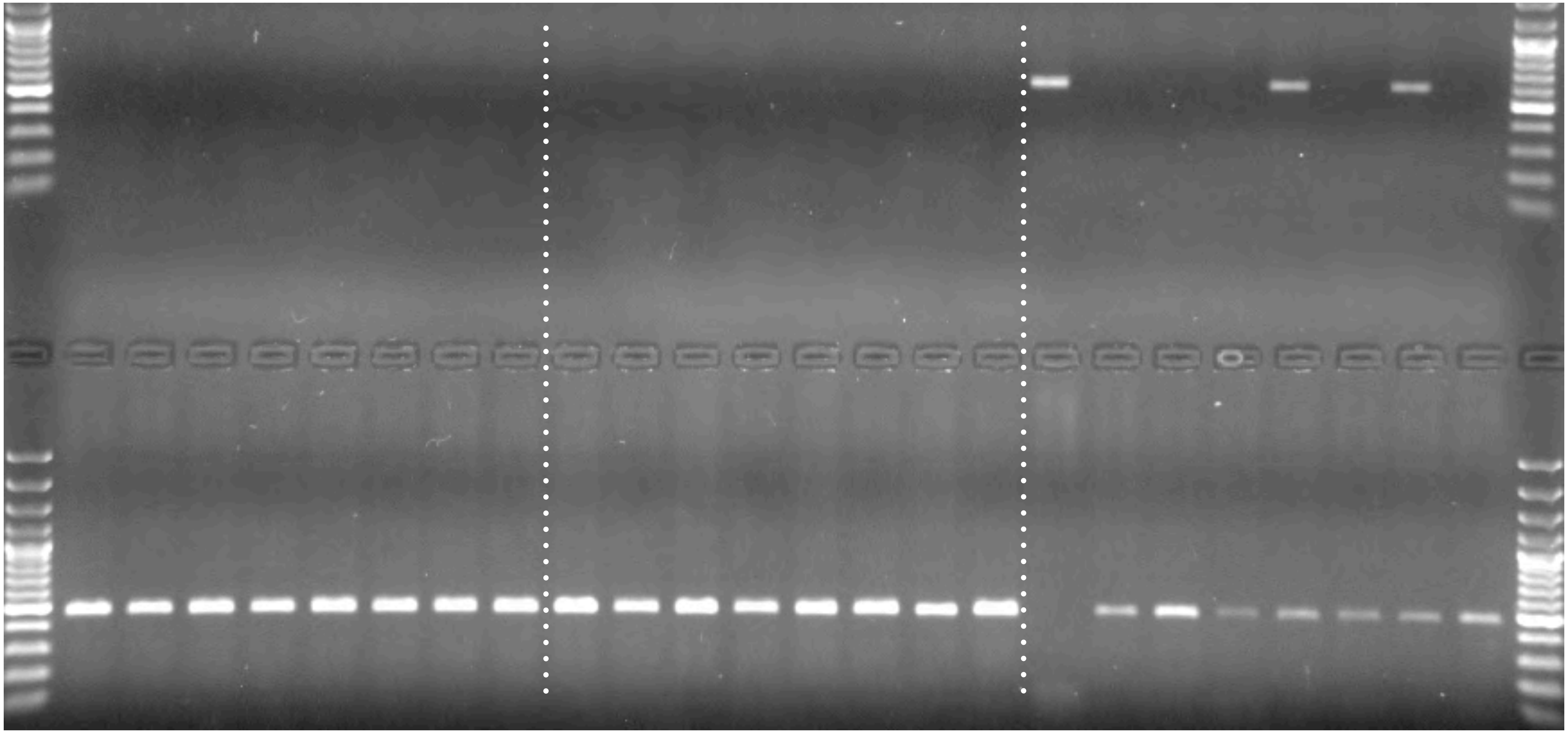
# Linkage Group XXI



17 (Marine)

23 (Marine)

8 (Both)



Like
Bear Paw

Inverted

# Other recent uses of RAD-seq

Quantitative Trait Loci (QTL) mapping

Population genomics & Genome Wide Association Studies (GWAS)

Phylogenetics and phylogeography

Genetic mapping, comparative genomics

*de novo* Genome assembly

Identifying signatures of selection in natural populations

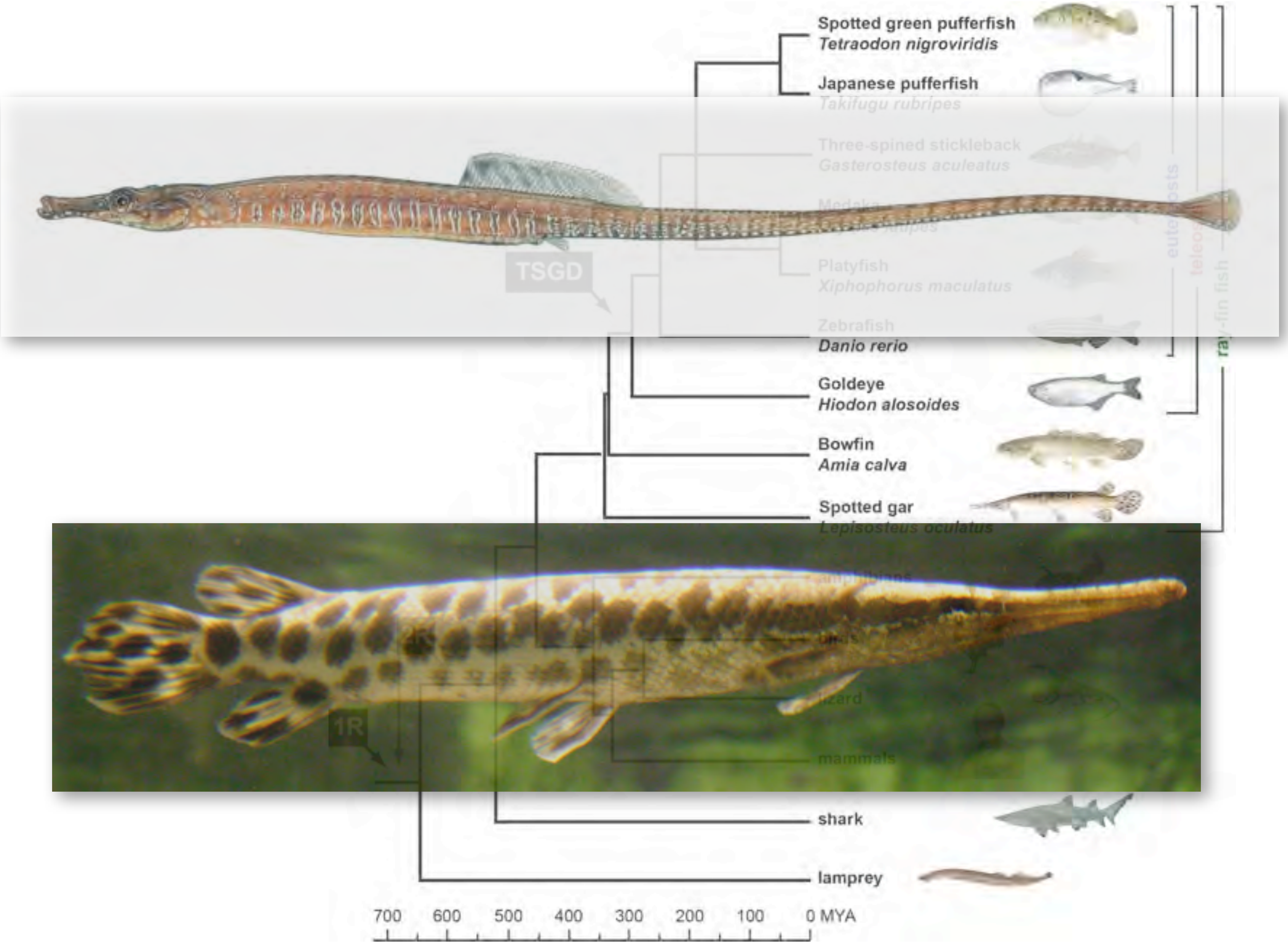
Inferring parentage and pedigrees in the wild

Quantitative genetics in outbred populations

Allele specific transcriptional profiling using RNA-seq

What if you don’t have a genome sequence?

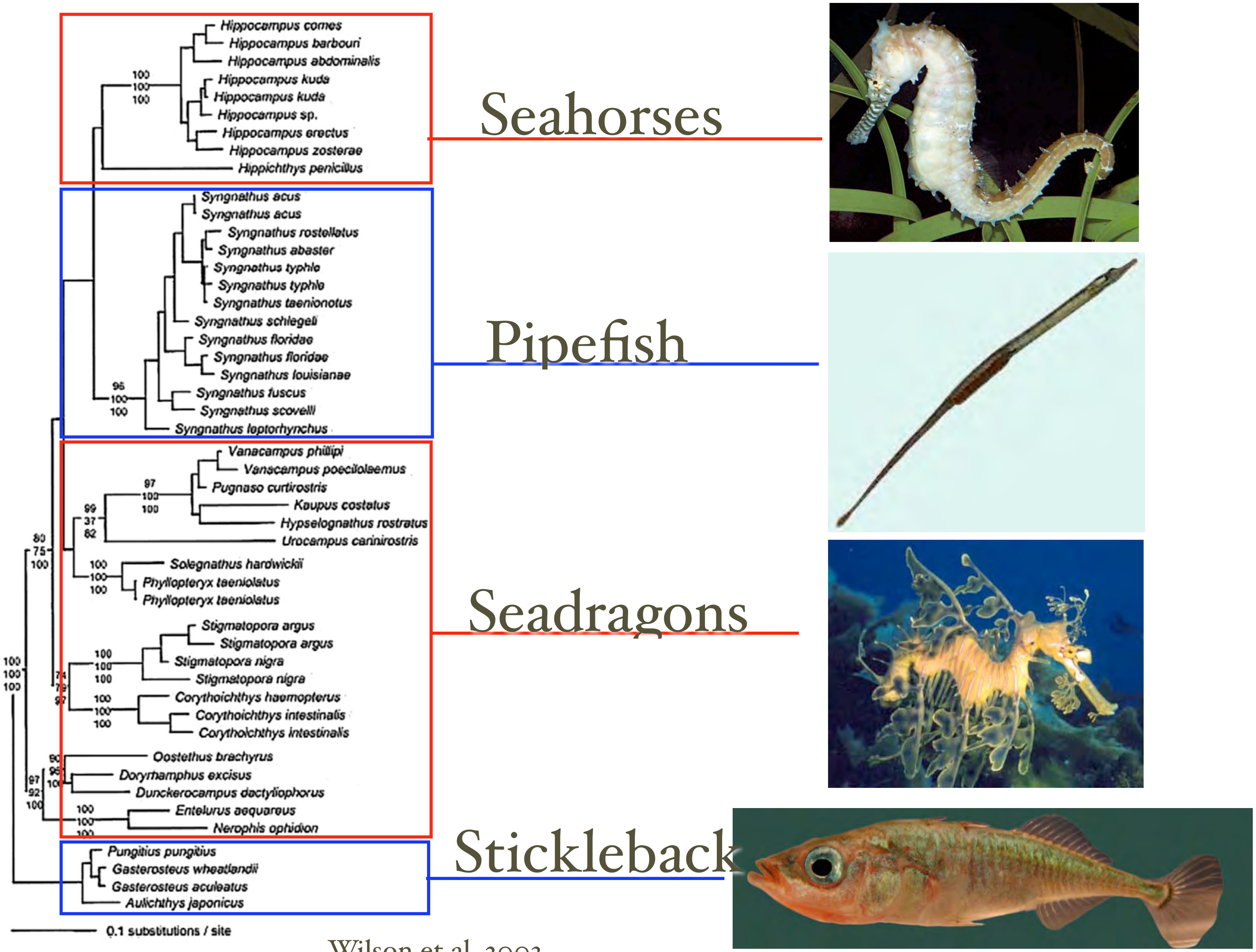
Genomically enabling very non-model organisms



Andrew Nishida, Julian Catchen, Susie Bassham,
Clay Small and Adam Jones

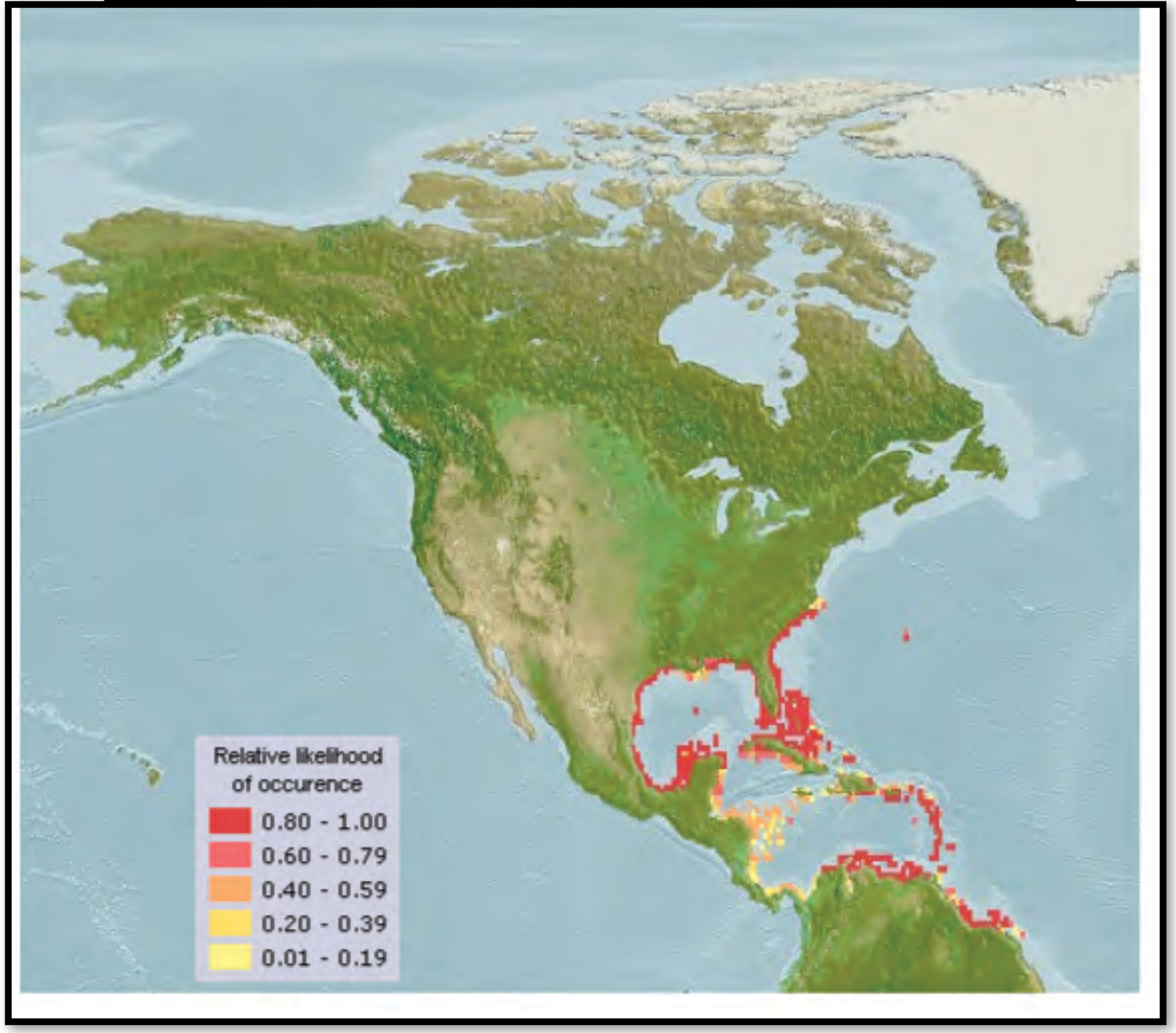
# Seahorses, sea dragons and pipefishes

# Gasterosteidae and Syngnathidae are historically considered to be closely related



Wilson et al. 2003

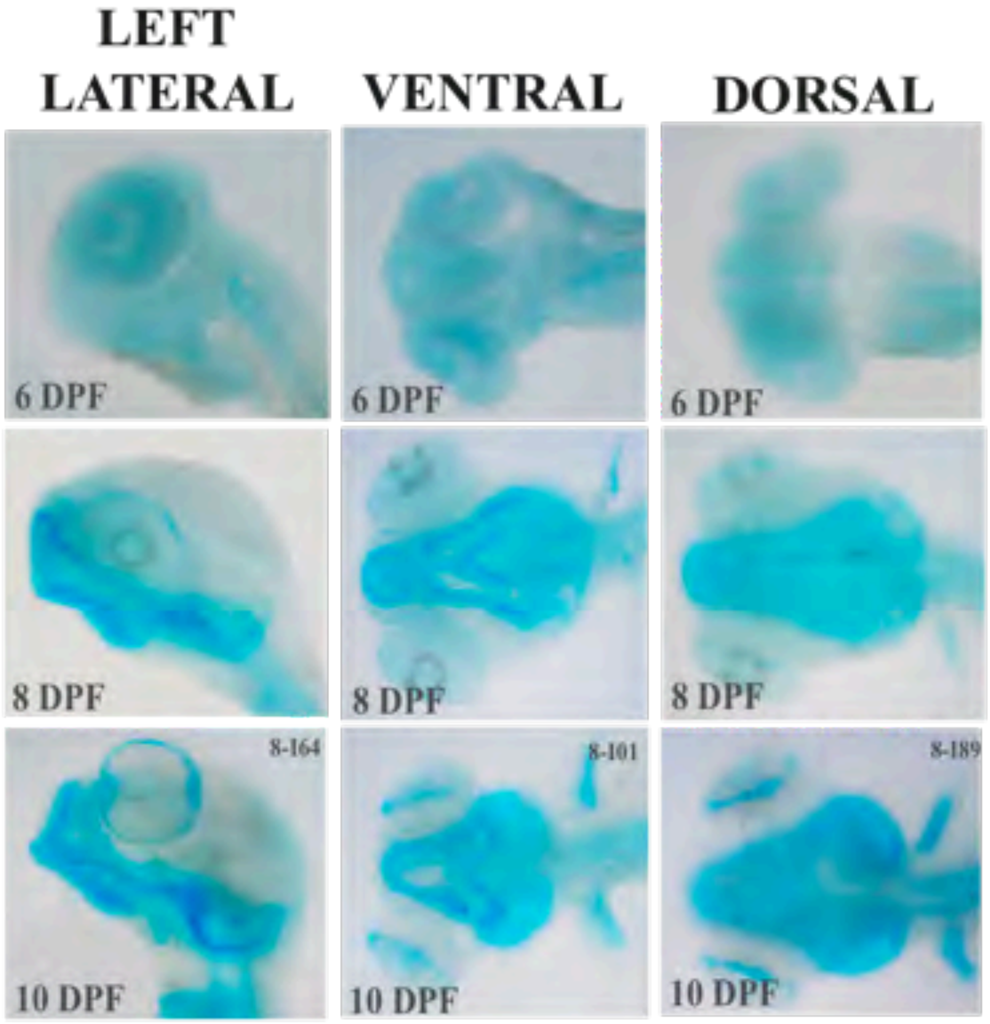
# Gulf Pipefish
# Syngnathus scovel

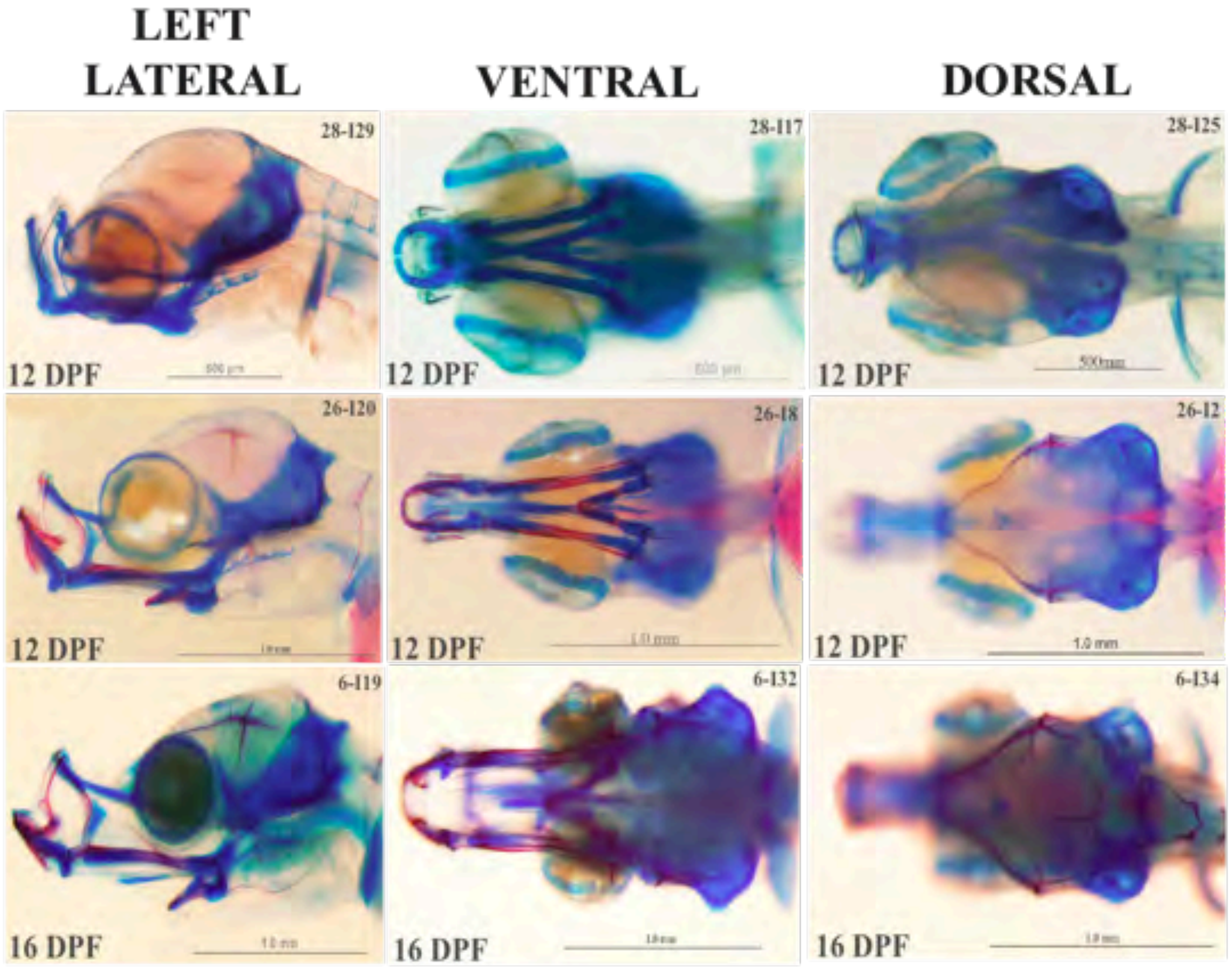


-160 mm (6.3”)
-reversed sex roles
-sexual dimorphism
-specialized suction feeding
-no sequences in international databases

www.bio.tamu.edu/USERS/ajones/charlyn.html
Fishbase.org

# We’re really interested in the head and body axis





# *Solution: ‘genomically enable’ pipefish*

1) A high quality transcriptome

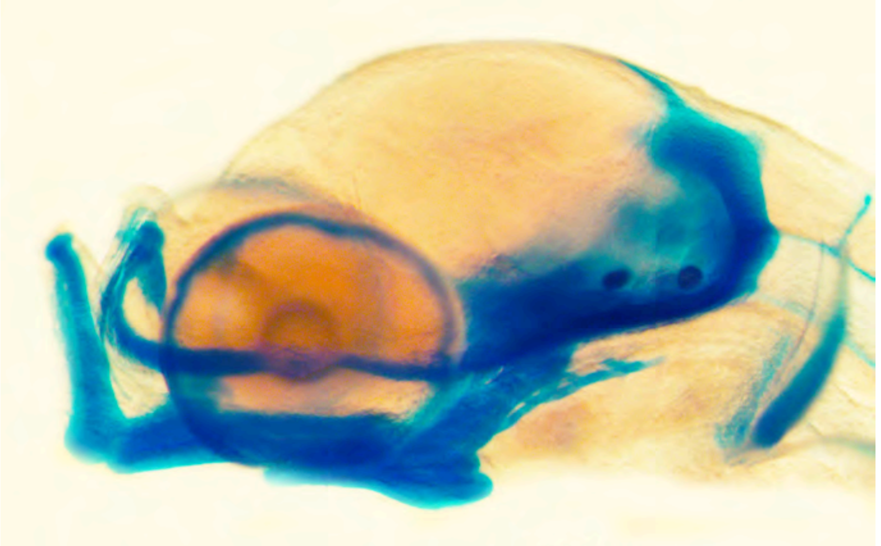
2) Very dense RAD genetic map

3) Deep coverage shotgun sequencing of genome

4) Order genomic and transcriptomic contigs against the RAD reference map

# Pipefish Transcriptome

# Building an EST database in pipefish



Pipefish embryonic mRNA

↓

Illumina sequencing:
100 nt, paired-end

↓

200 million reads (two lanes)

↓

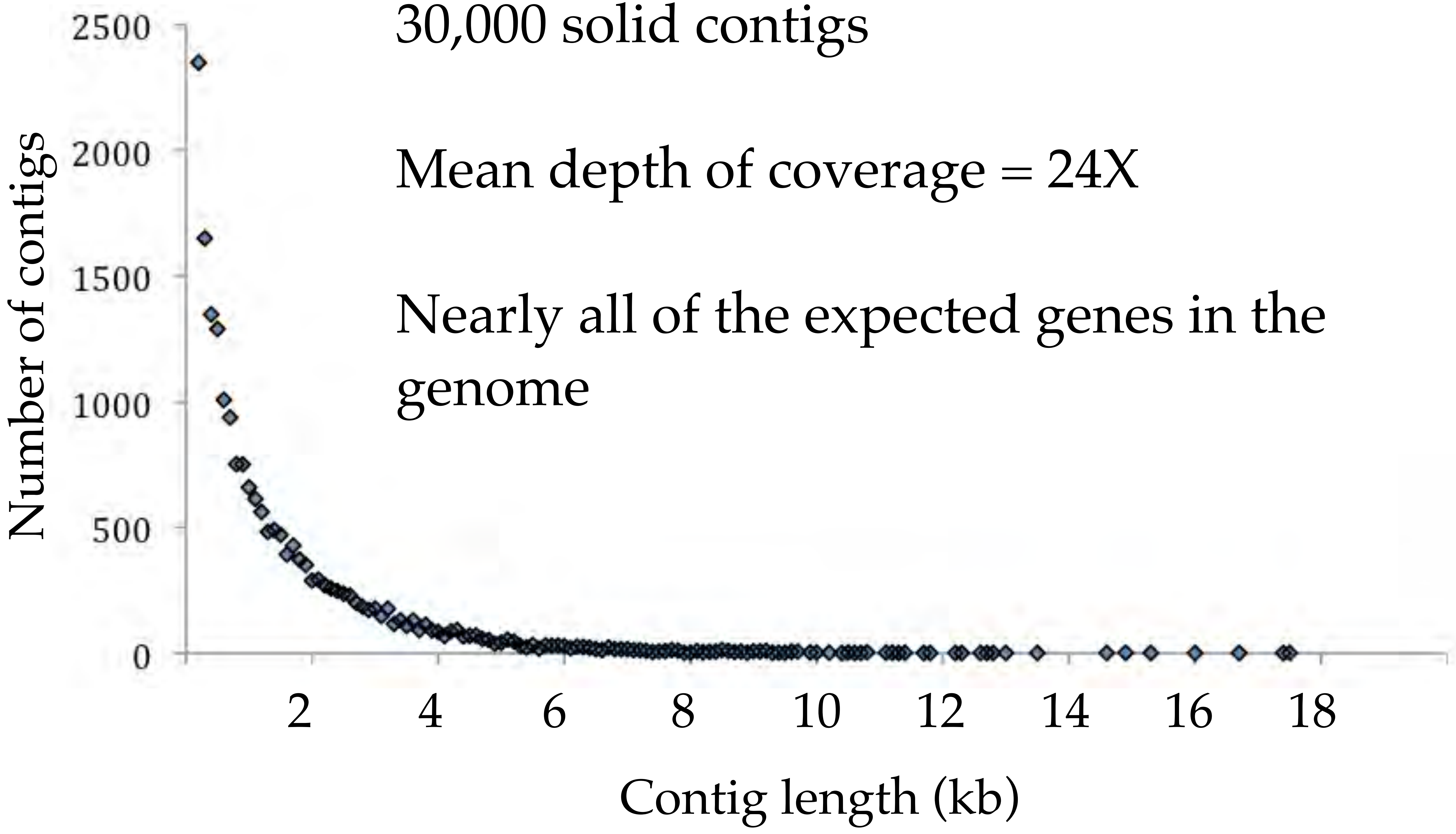
Assembly of transcripts

# Transcriptome

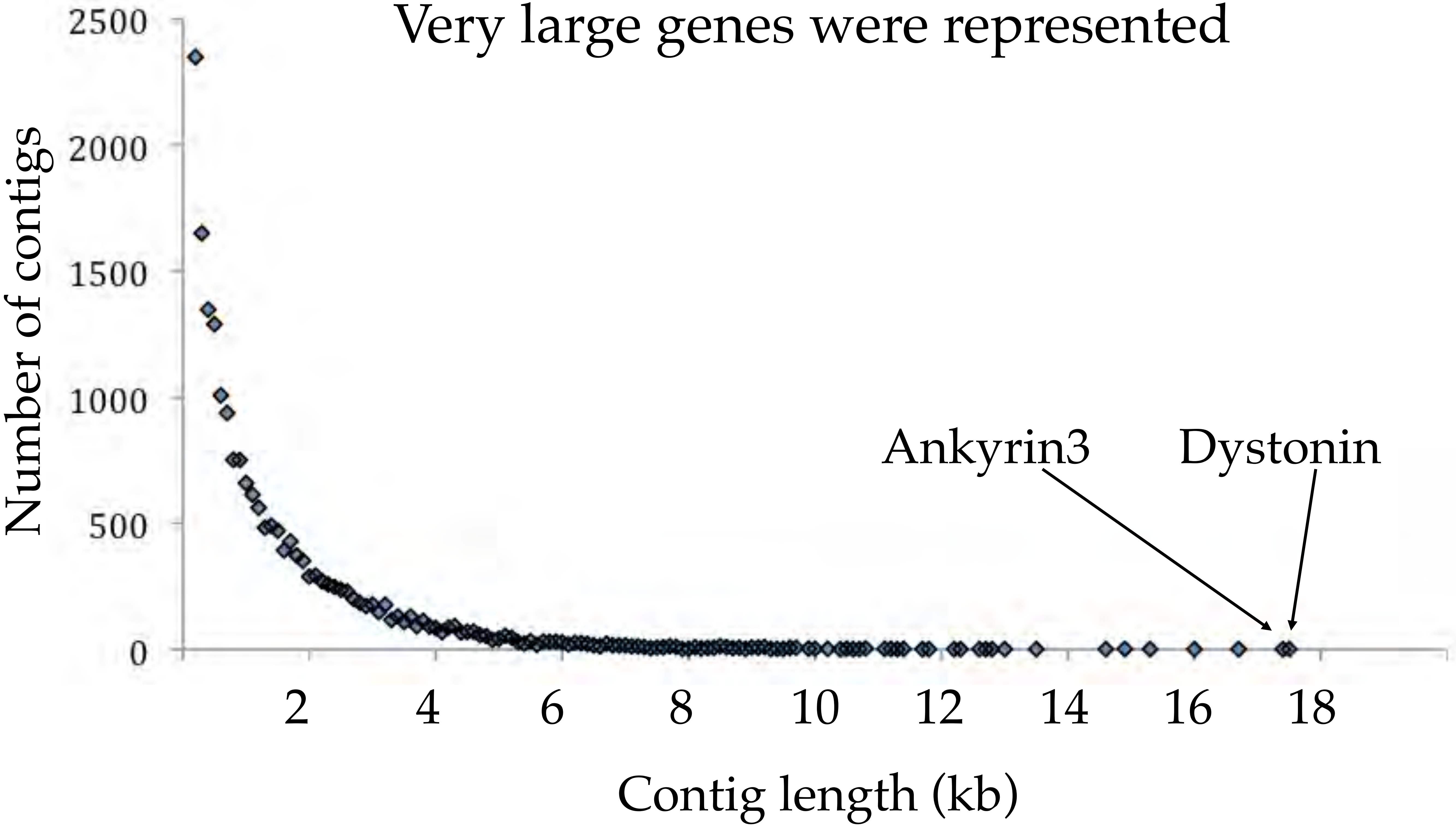
30,000 solid contigs

Mean depth of coverage = 24X

Nearly all of the expected genes in the genome

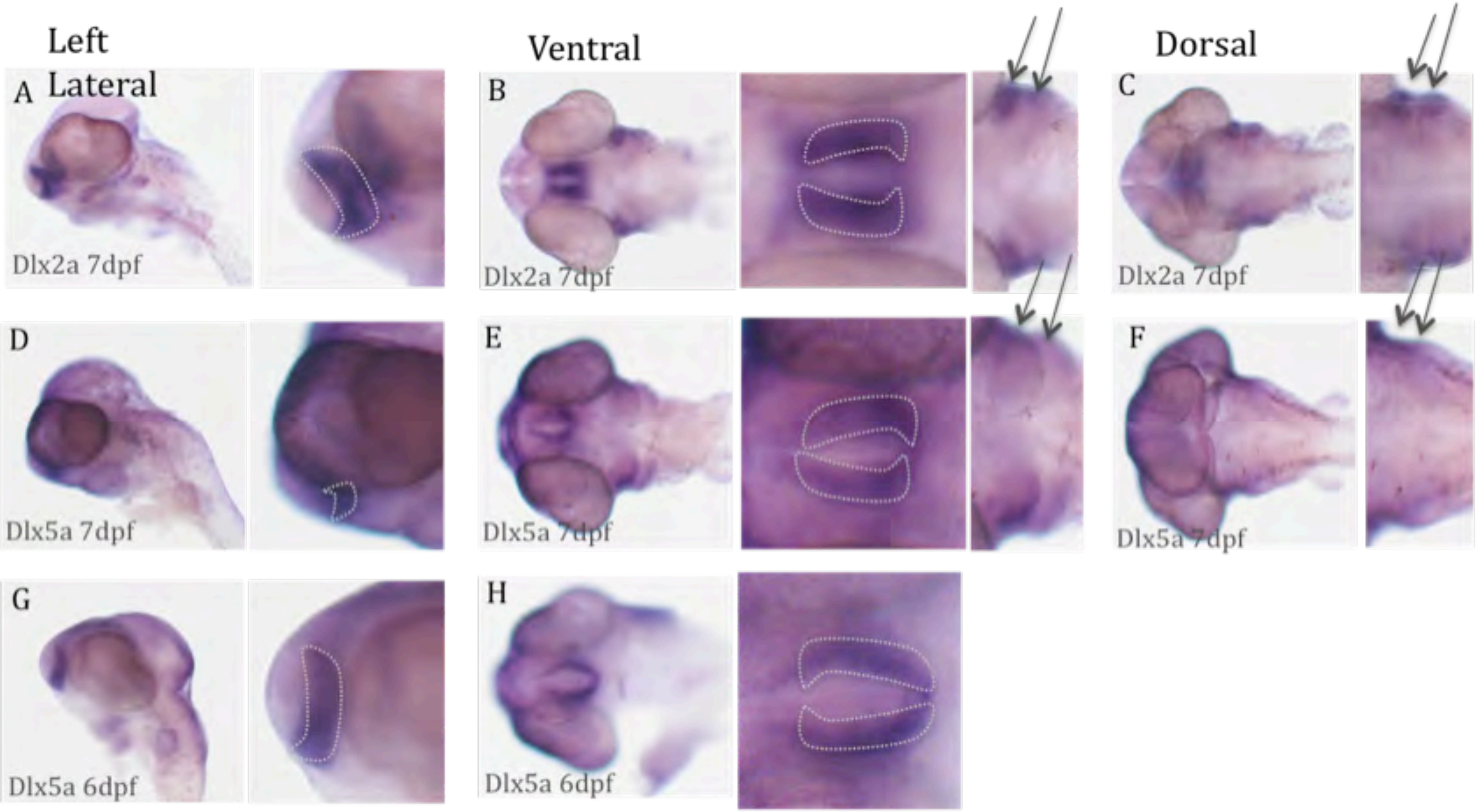


# Transcriptome



Very large genes were represented

# We could use these genes right away *Dlx2a* and *Dlx5a* expression in pipefish



# Pipefish Genetic Map

# Genetic map workflow

Generated an F1 family of 103 individuals

RAD sequenced the parents and offspring

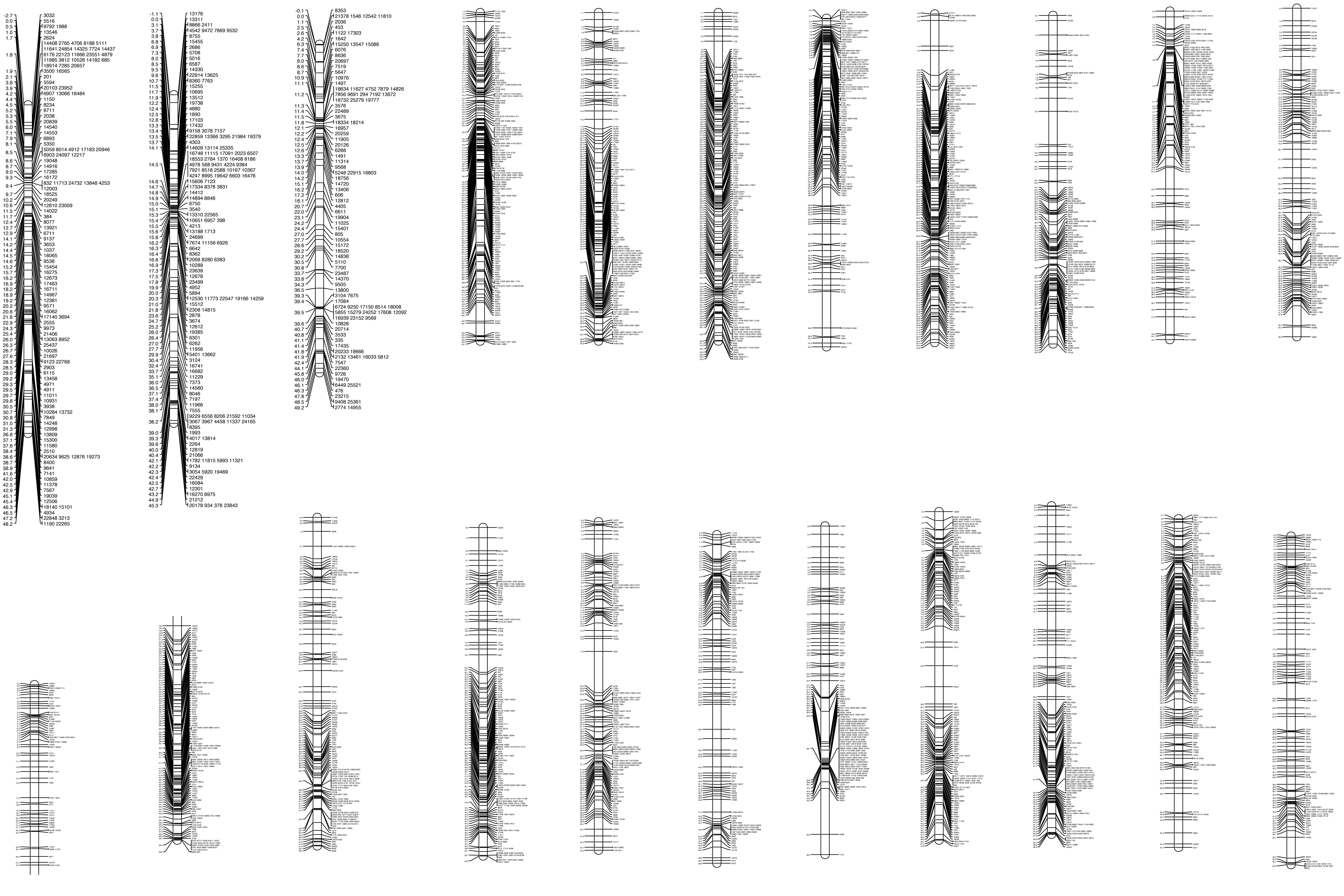
Analyzed the data using *Stacks*

Paired end local assemblies

Output to JoinMap format

Created Linkage map

# The pipefish genetic map is closed; 22 LGs 6000 segregating SNPs; 30,000 RAD sites



# Pipefish Genome Project

# Genome workflow

Generated DNA from a single individual

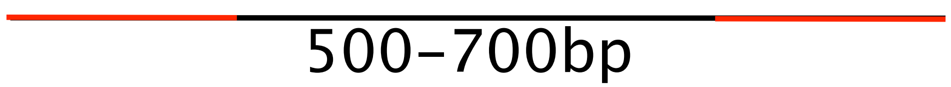
Random Illumina shotgun sequencing

Removed highly repetitive kmers

Produced *several* different genome assemblies

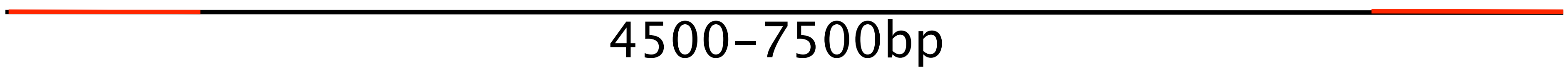
# Illumina genomic libraries for pipefish genome

paired end 101bp



25x

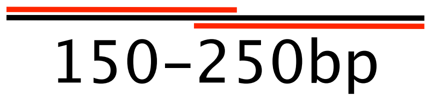
mate pair



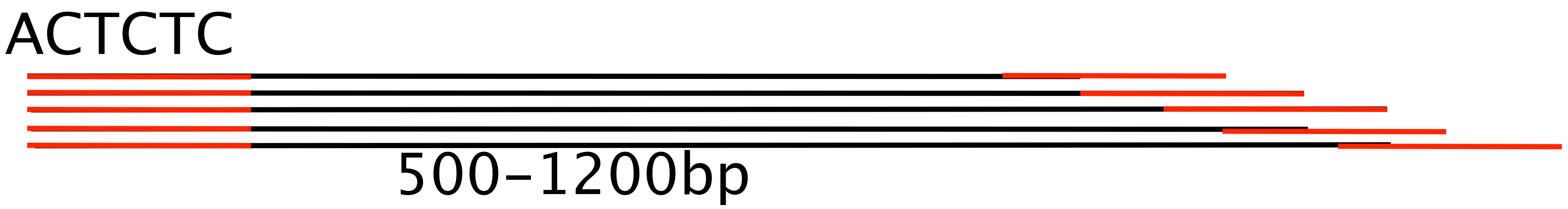
2x

overlapping

40x



paired end RAD



15–25x of
3% of the
genome

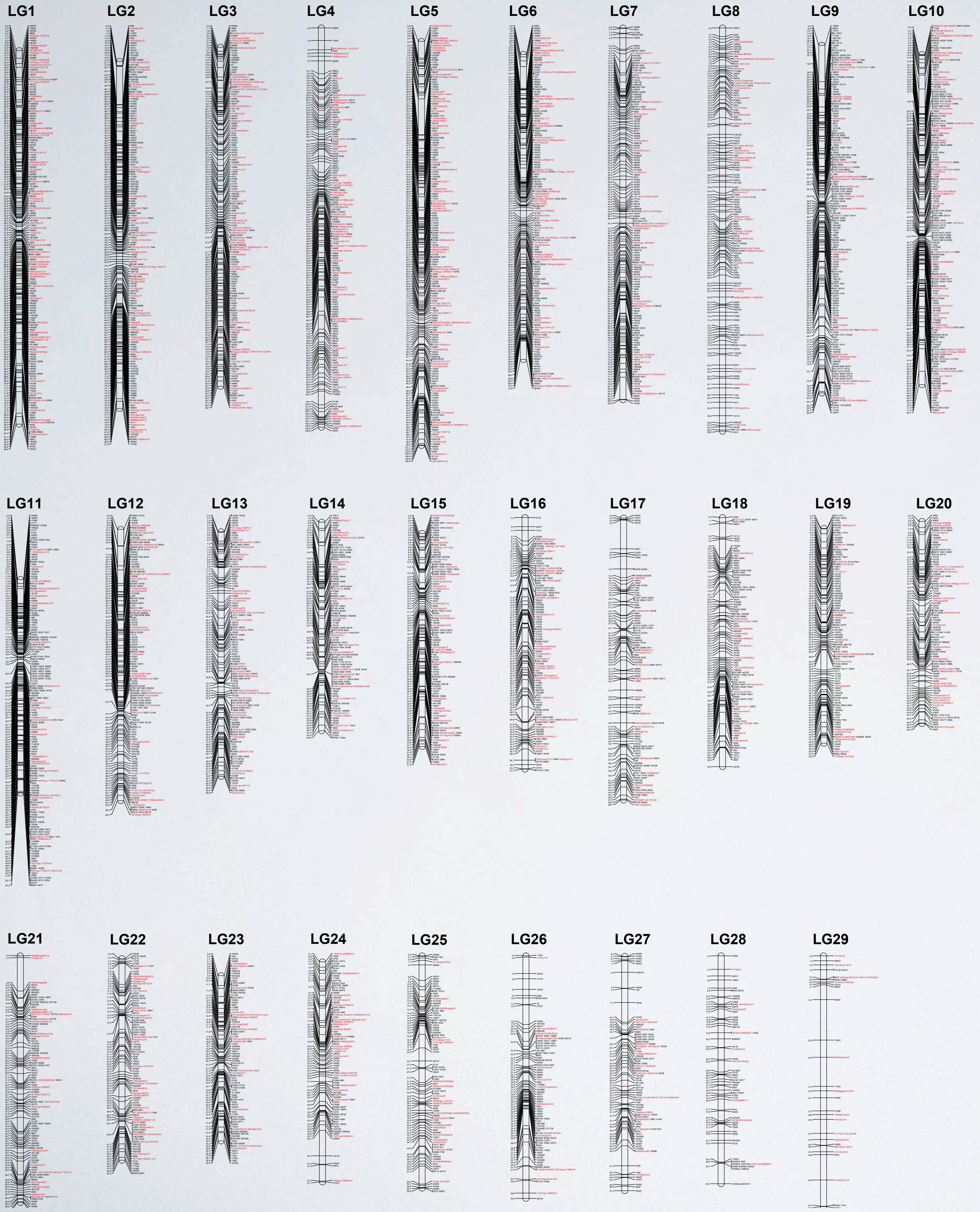
# Pipefish genome assembly version 0.99
Nearly the whole genome is covered

| Coverage | Scaffolds | Contigs | Scaffold N50 | Contig N50 |
|---|---|---|---|---|
| All (66.6x) | 33,911 | 307,317 | 26,109 | 1,840 |

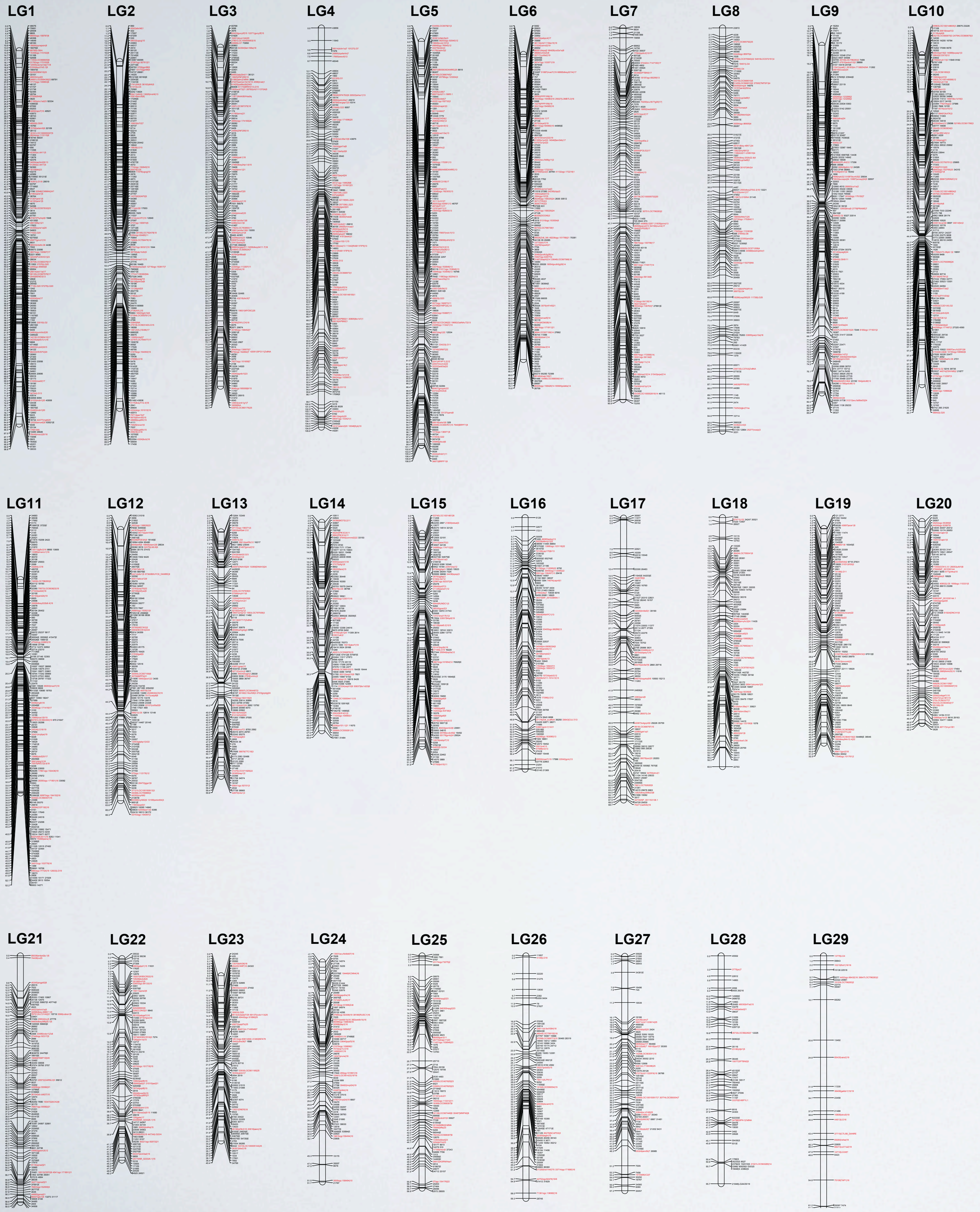
| Max | Average Length | Total Length | Gap Length | % |
|---|---|---|---|---|
| 198,155 | 9,916.35 | 336,273,415 | 38,303,839 | (11.39%) |

# Bringing it all together; the spotted gar

Amores, Catchen et al. 2011. Genetics

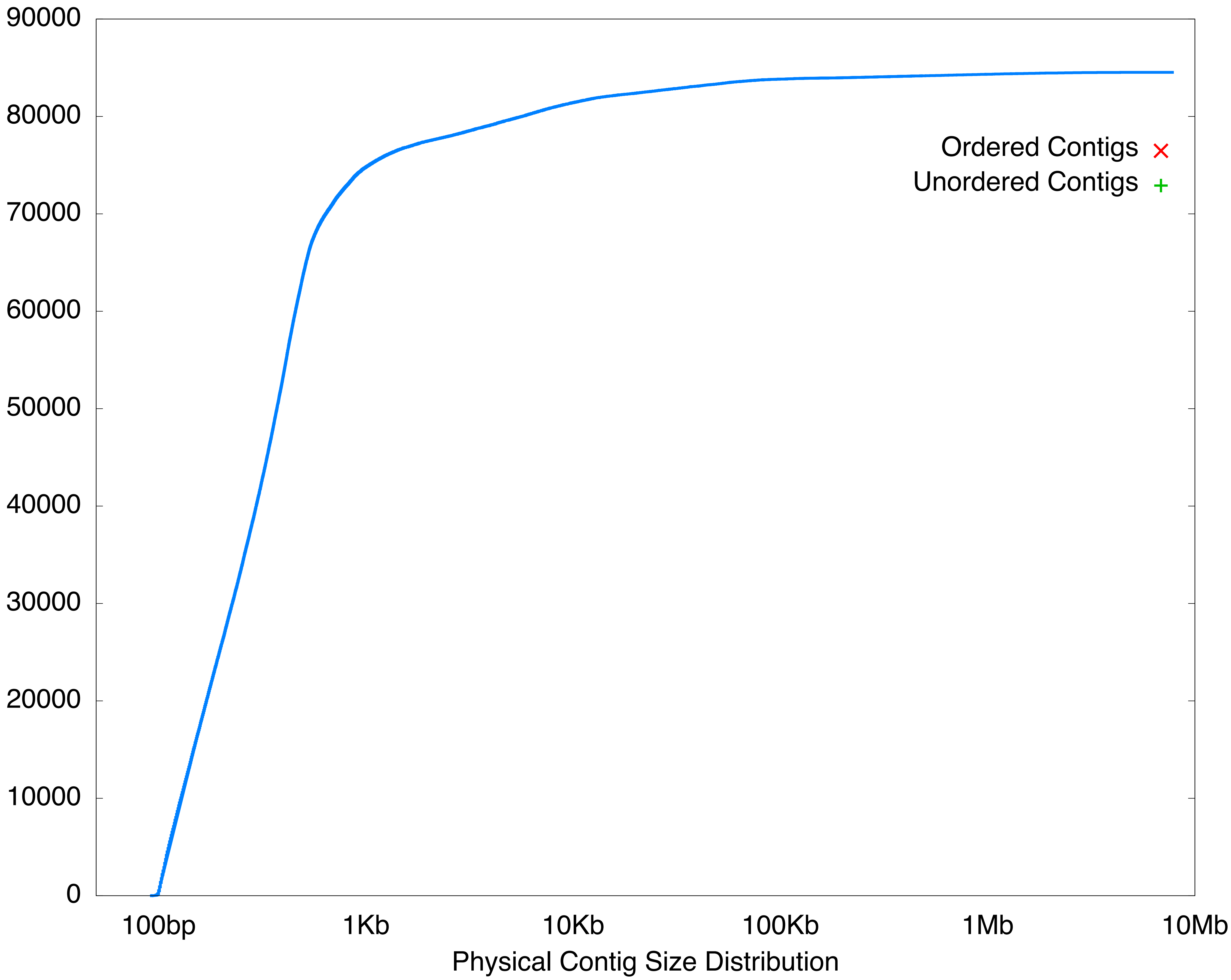


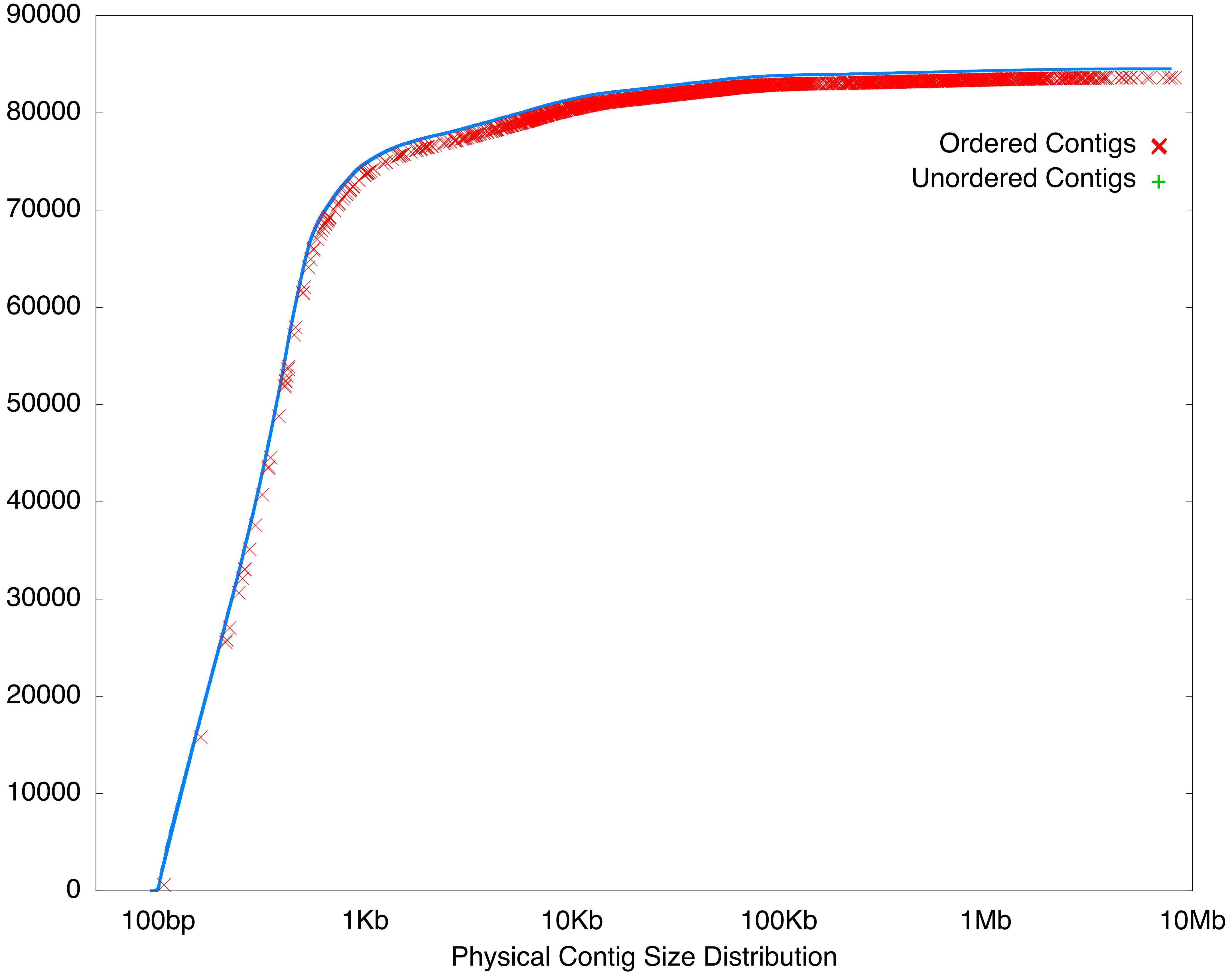
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

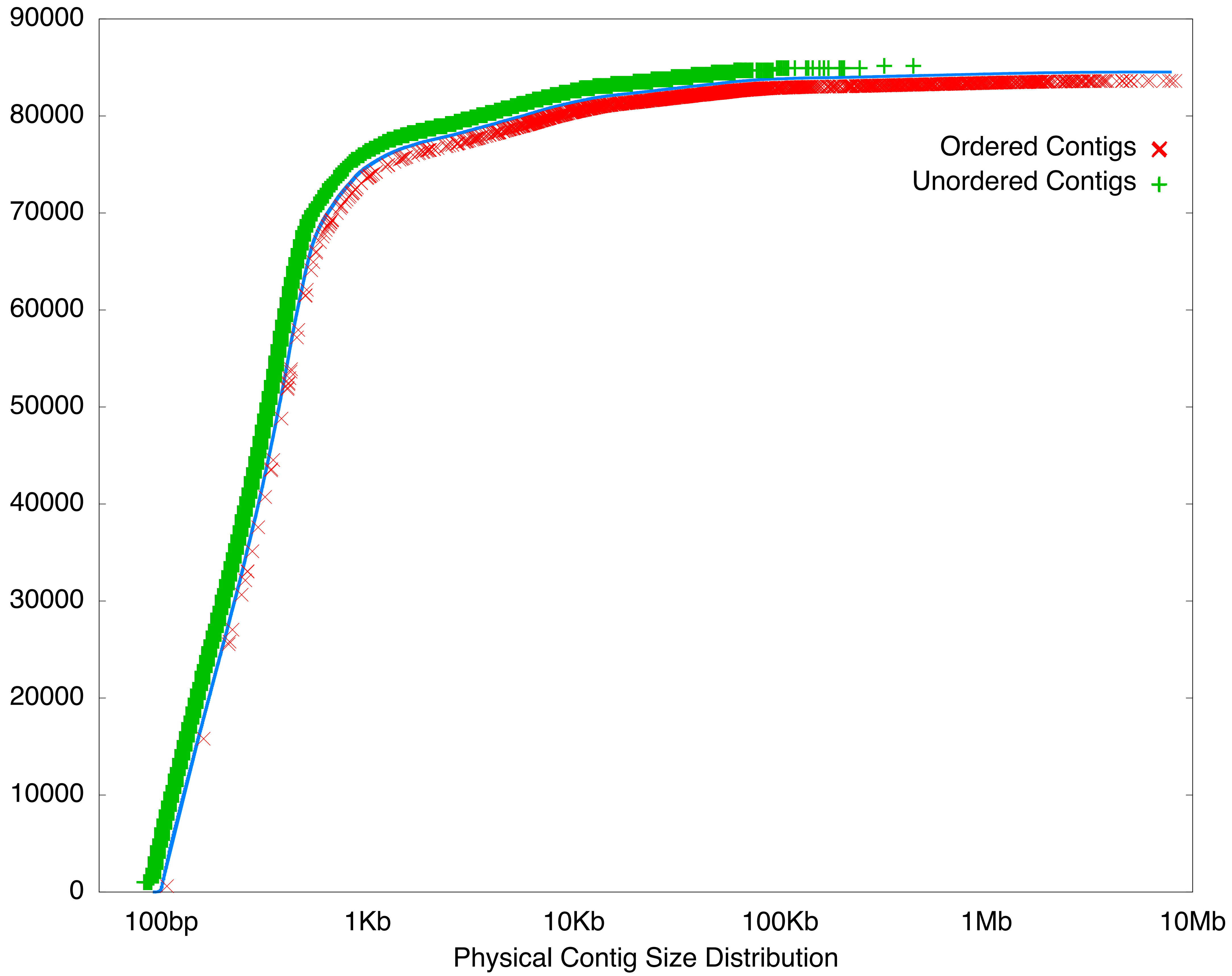


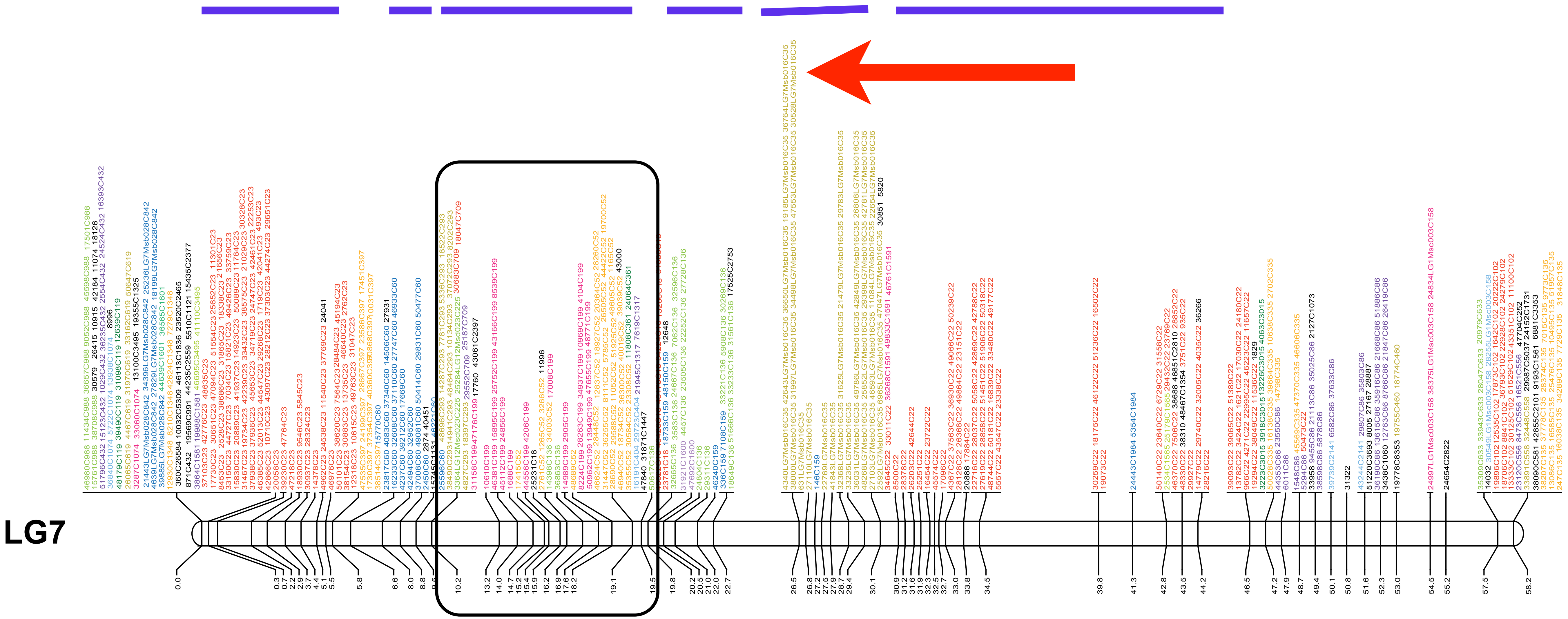
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

| Organism | Markers |
|---|---|
| Silver carp | 483 |
| Guppy | 790 |
| Barramundi | 240 |
| Catfish | 331 |
| Sea bass | 368 |
| Cichlid | 204 |
| Platyfish | 290 |
| Halibut | 604 |
| Sea bream | 204 |


90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Ordered Contigs ×
Unordered Contigs +
100bp
1Kb
10Kb
100Kb
1Mb
10Mb
Physical Contig Size Distribution


90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Ordered Contigs ×
Unordered Contigs +
100bp
1Kb
10Kb
100Kb
1Mb
10Mb
Physical Contig Size Distribution

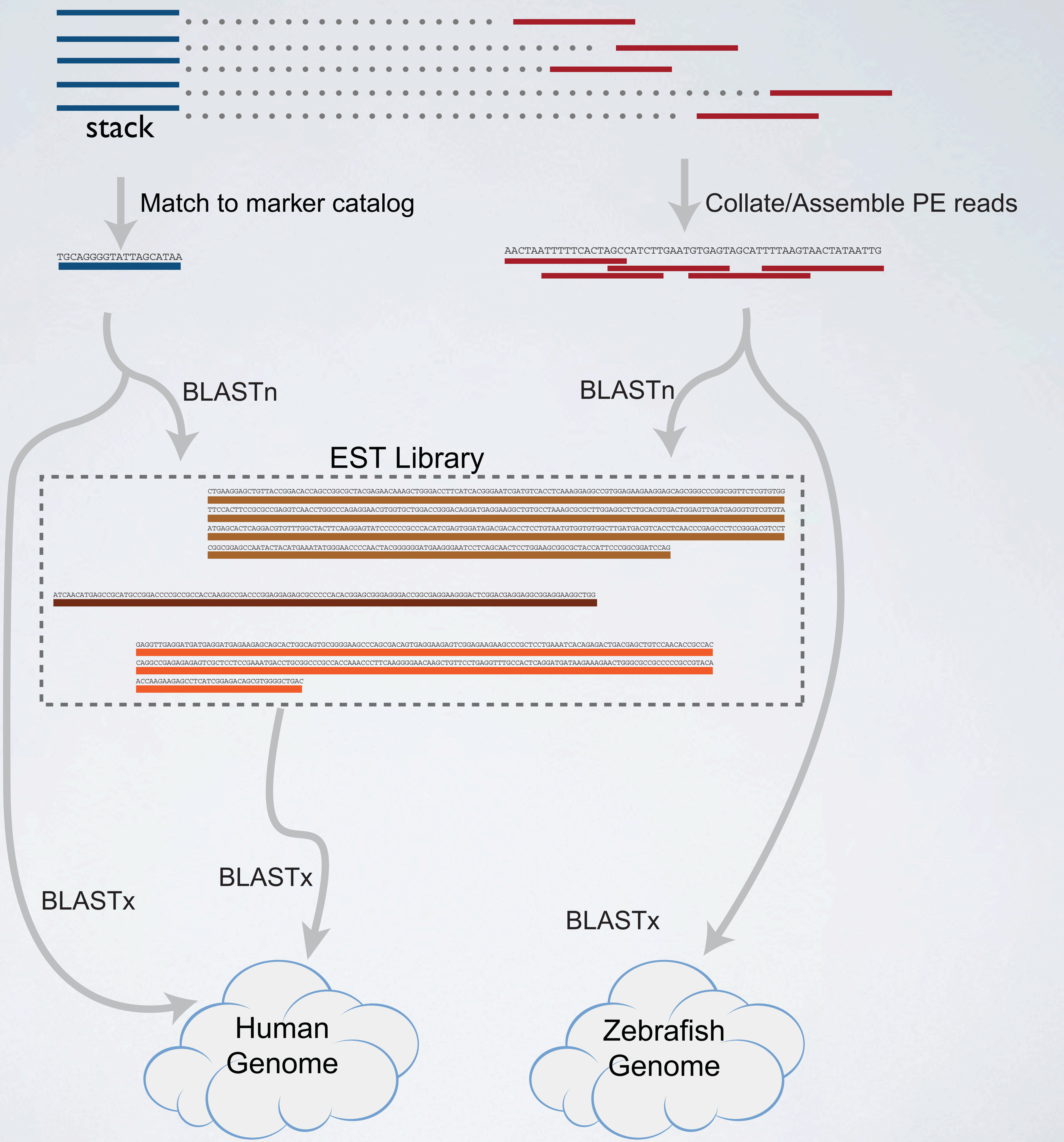

90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Ordered Contigs ×
Unordered Contigs +
100bp
1Kb
10Kb
100Kb
1Mb
10Mb
Physical Contig Size Distribution



- multiple RAD sites per segregating marker means that more of the genome can be tiled

- Mis-assemblies are easily identified

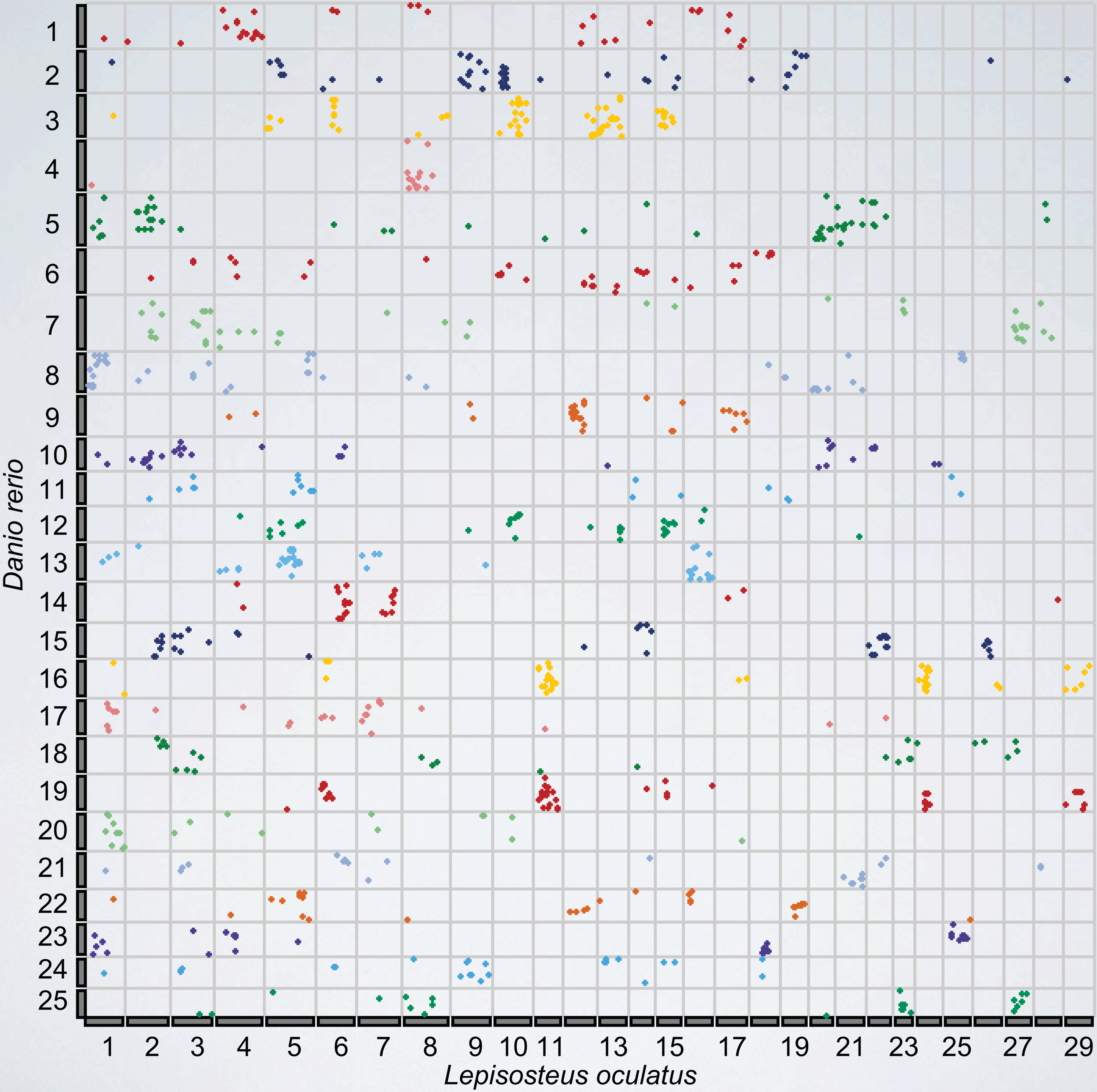
# Using PE RAD for local assembly

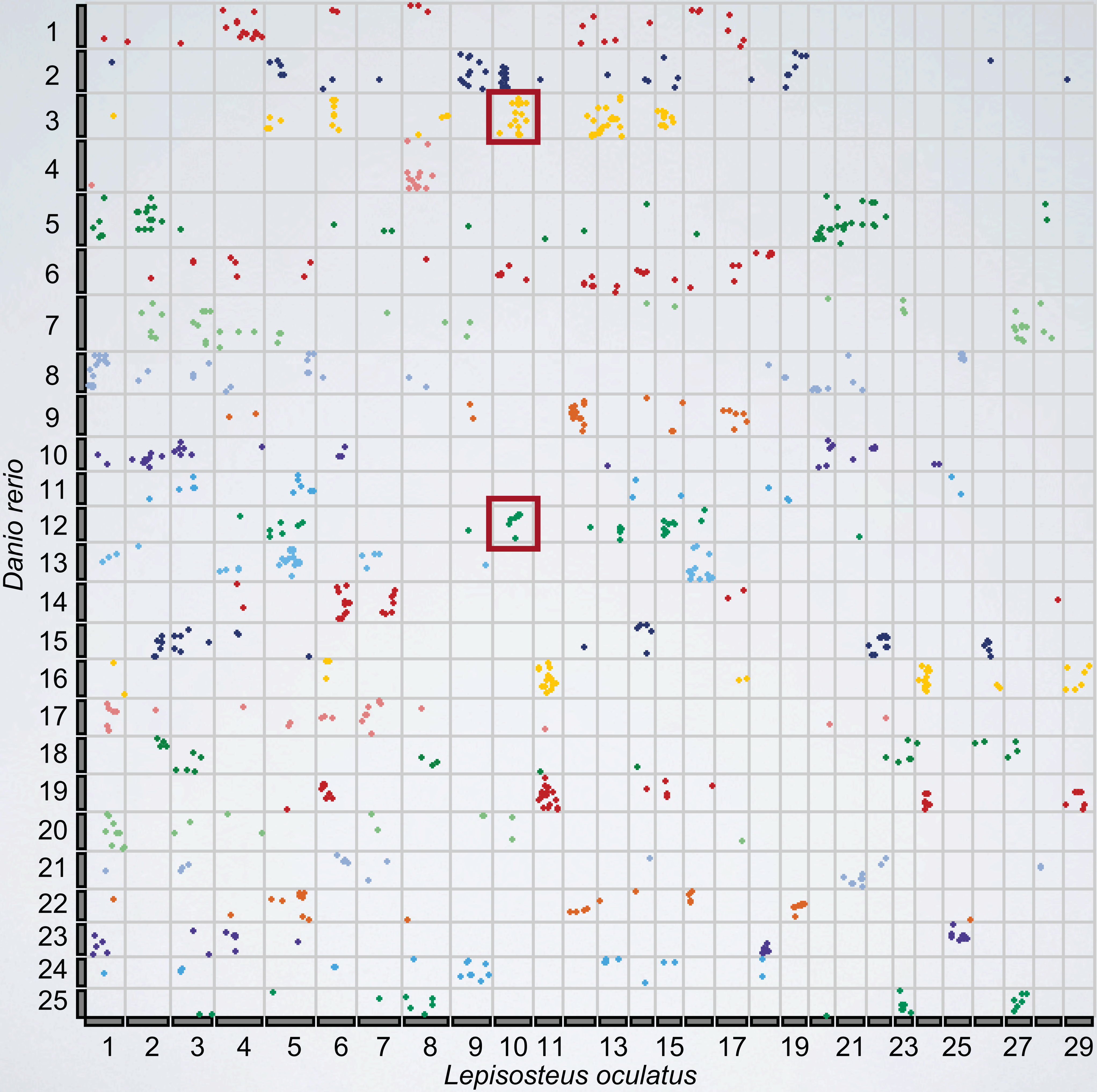

stack
Match to marker catalog
Collate/Assemble PE reads
TGCAGGGGTATTAGCATAA
AACTAATTTTTCACTAGCCATCTTGAATGTGAGTAGCATTTTAAGTAACTATAATTG
BLASTn
BLASTn
EST Library
CTGAAGGAGCTGTTACCGGACACCAGCCGGCGCTACGAGAACAAAGCTGGGACCTTCATCACGGGAATCGATGTCACCTCAAAGGAGGCCGTGGAGAAGAAGGAGCAGCGGGCCCGGCGGTTCTCGTGTGG
TTCCACTTCCGCGCCGAGGTCAACCTGGCCCAGAGGAACGTGGTGCTGGACCGGGACAGGATGAGGAAGGCTGTGCCTAAAGCGCGCTTGGAGGCTCTGCACGTGACTGGAGTTGATGAGGGTGTCGTGTA
ATGAGCACTCAGGACGTGTTTGGCTACTTCAAGGAGTATCCCCCCGCCCACATCGAGTGGATAGACGACACCTCCTGTAATGTGGTGTGGCTTGATGACGTCACCTCAACCCGAGCCCTCCGGGACGTCCT
CGGCGGAGCCAATACTACATGAAATATGGGAACCCCAACTACGGGGGGATGAAGGGAATCCTCAGCAACTCCTGGAAGCGGCGCTACCATTCCCGGCGGATCCAG
ATCAACATGAGCCGCATGCCGGACCCCGCCGCCACCAAGGCCGACCCGGAGGAGAGCGCCCCCACACGGAGCGGGAGGGACCGGCGAGGAAGGGACTCGGACGAGGAGGCGGAGGAAGGCTGG
GAGGTTGAGGATGATGAGGATGAGAAGAGCAGCACTGGCAGTGCGGGGAAGCCCAGCGACAGTGAGGAAGAGTCGGAGAAGAAGCCCGCTCCTGAAATCACAGAGACTGACGAGCTGTCCAACACCGCCAC
CAGGCCGAGAGAGAGTCGCTCCTCCGAAATGACCTGCGGCCCGCCACCAAACCCTTCAAGGGGAACAAGCTGTTCCTGAGGTTTGCCACTCAGGATGATAAGAAAGAACTGGGCGCCGCCCCCGCCGTACA
ACCAAGAAGAGCCTCATCGGAGACAGCGTGGGGCTGAC
BLASTx
BLASTx
BLASTx
Human Genome
Zebrafish Genome

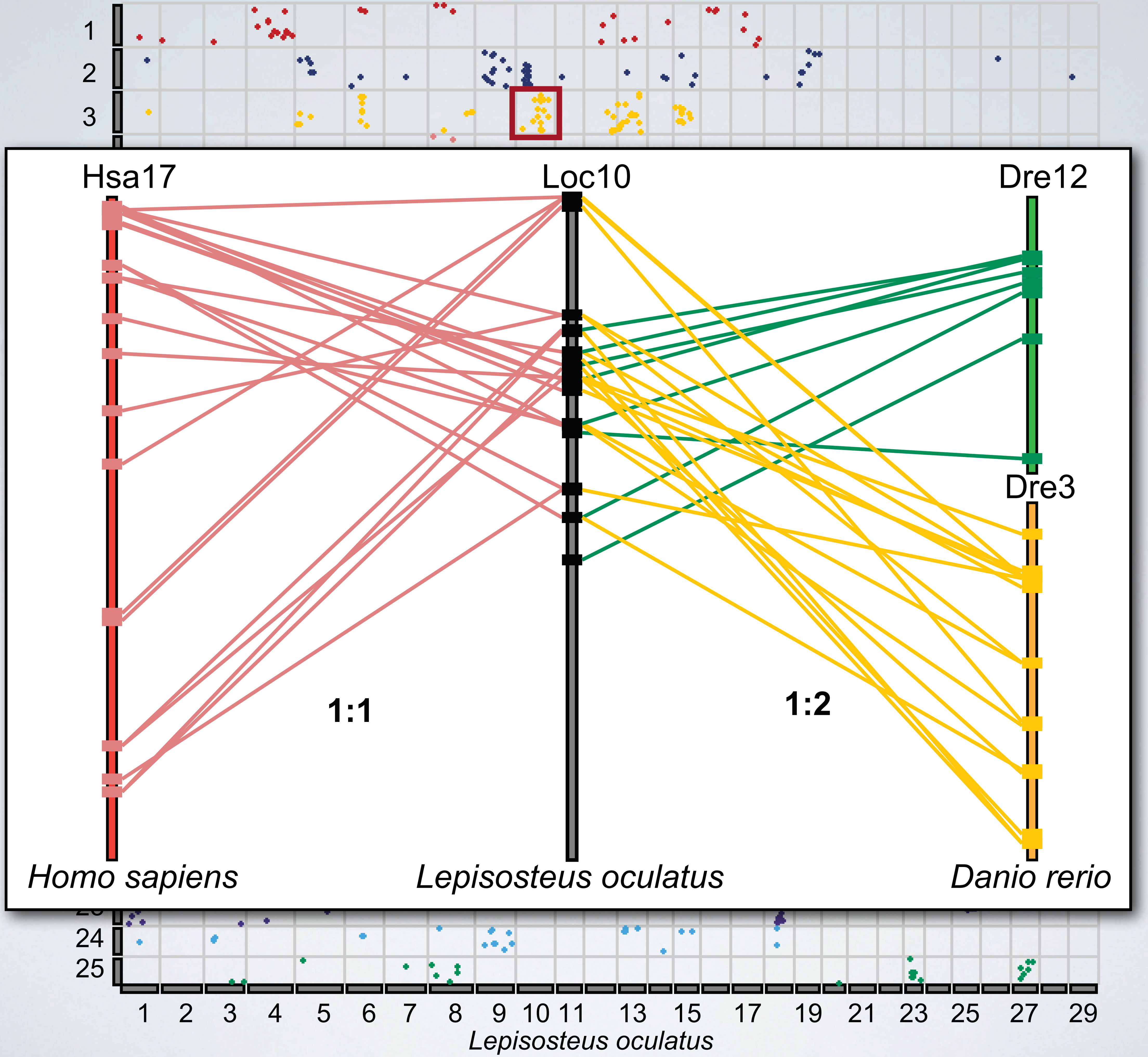
Associate markers / PE contigs with ESTs & genomic contigs

Assign orthology to:

- markers
- PE contigs
- ESTs


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Danio rerio
1
2
3
4
5
6
7
8
9
10
11
13
15
17
19
21
23
25
27
29
Lepisosteus oculatus


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Danio rerio
1
2
3
4
5
6
7
8
9
10
11
13
15
17
19
21
23
25
27
29
Lepisosteus oculatus


1
2
3
Hsa17
Loc10
Dre12
Dre3
1:1
1:2
Homo sapiens
Lepisosteus oculatus
Danio rerio
24
25
1
2
3
4
5
6
7
8
9
10
11
13
15
17
19
21
23
25
27
29
Lepisosteus oculatus

# Overall Conclusions

Genomics can be a tool for enabling new ecology and evolution research

- documenting patterns of genetic variation
- identifying the molecular genetic basis of important phenotypic variation
- assessing how ecological processes structure this genetic variation in genomes
- RAD-seq is a powerful tool for SNP identification and genotyping
- analytical and computational approaches are challenging but

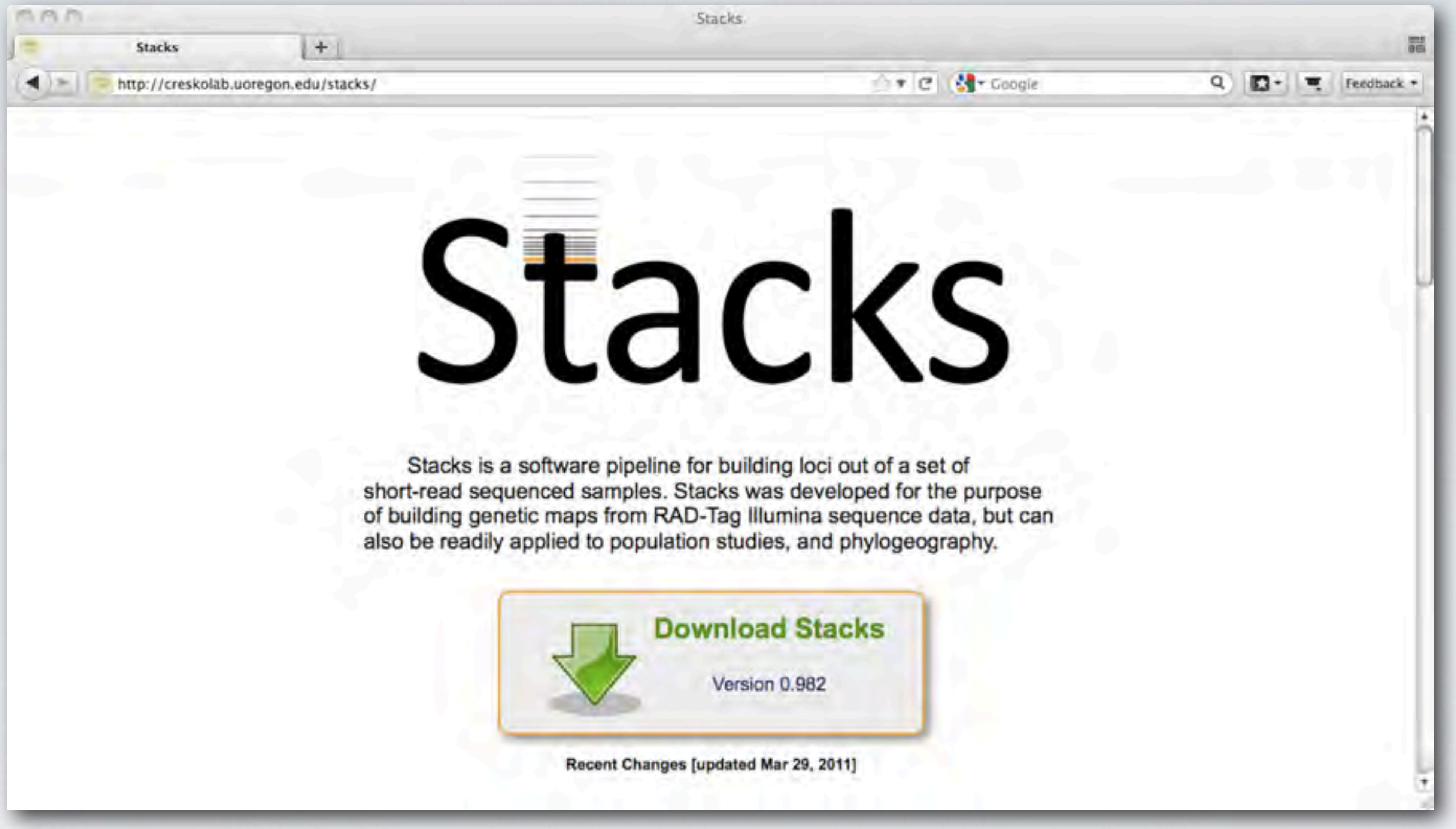
Not your father's genome assembly

- a mixture of data types can be efficiently combined
- a genetic map is extremely useful for pulling it all together
- having a tiled genome is good enough - it doesn't have to be completely closed

*Open Source Genomics* provides a suite of breakthrough technologies

- the molecular approaches are not as daunting as they first appear
- analytical and computational approaches are challenging
- **New software tools can help, but knowledge of Unix and Python is essential**

# TUTORIAL - USING *STACKS*



Stacks

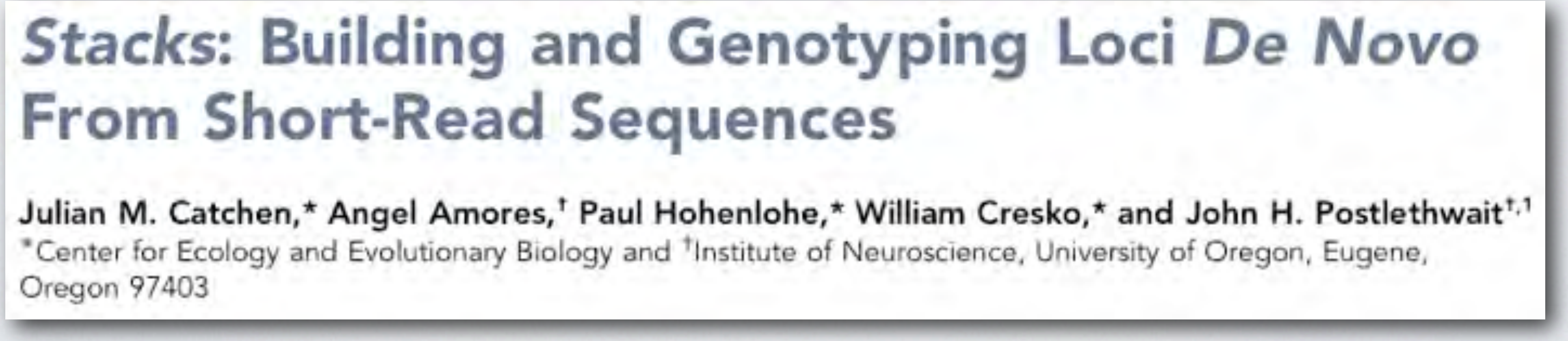
Stacks is a software pipeline for building loci out of a set of short-read sequenced samples. Stacks was developed for the purpose of building genetic maps from RAD-Tag Illumina sequence data, but can also be readily applied to population studies, and phylogeography.

Download Stacks

Version 0.982

Recent Changes [updated Mar 29, 2011]

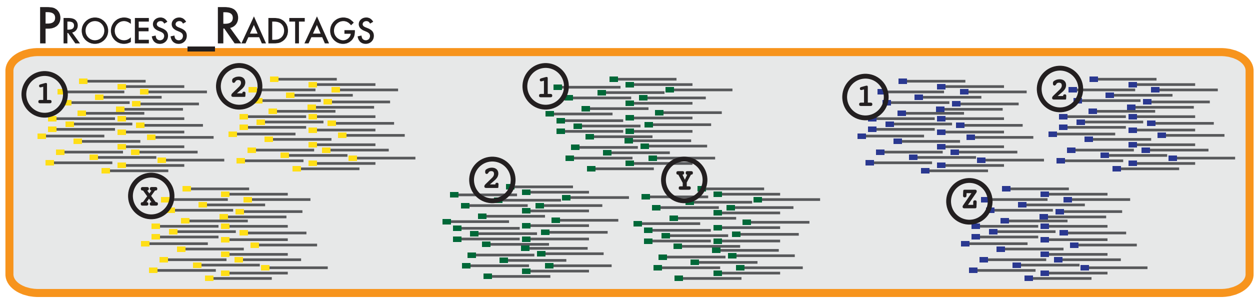
G3: Genes, Genomes, Genetics



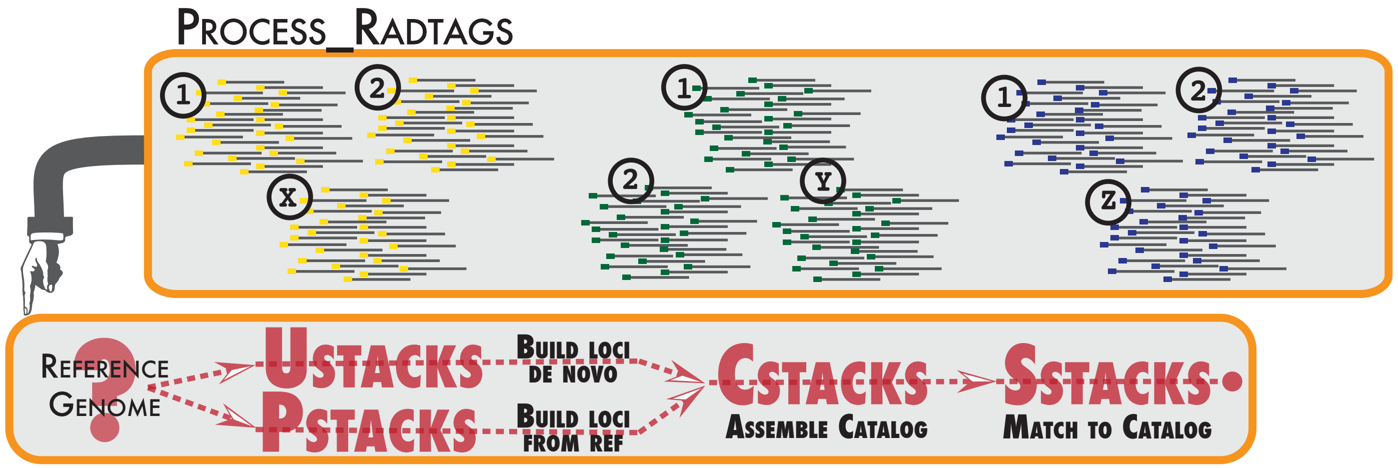
## *Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences

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

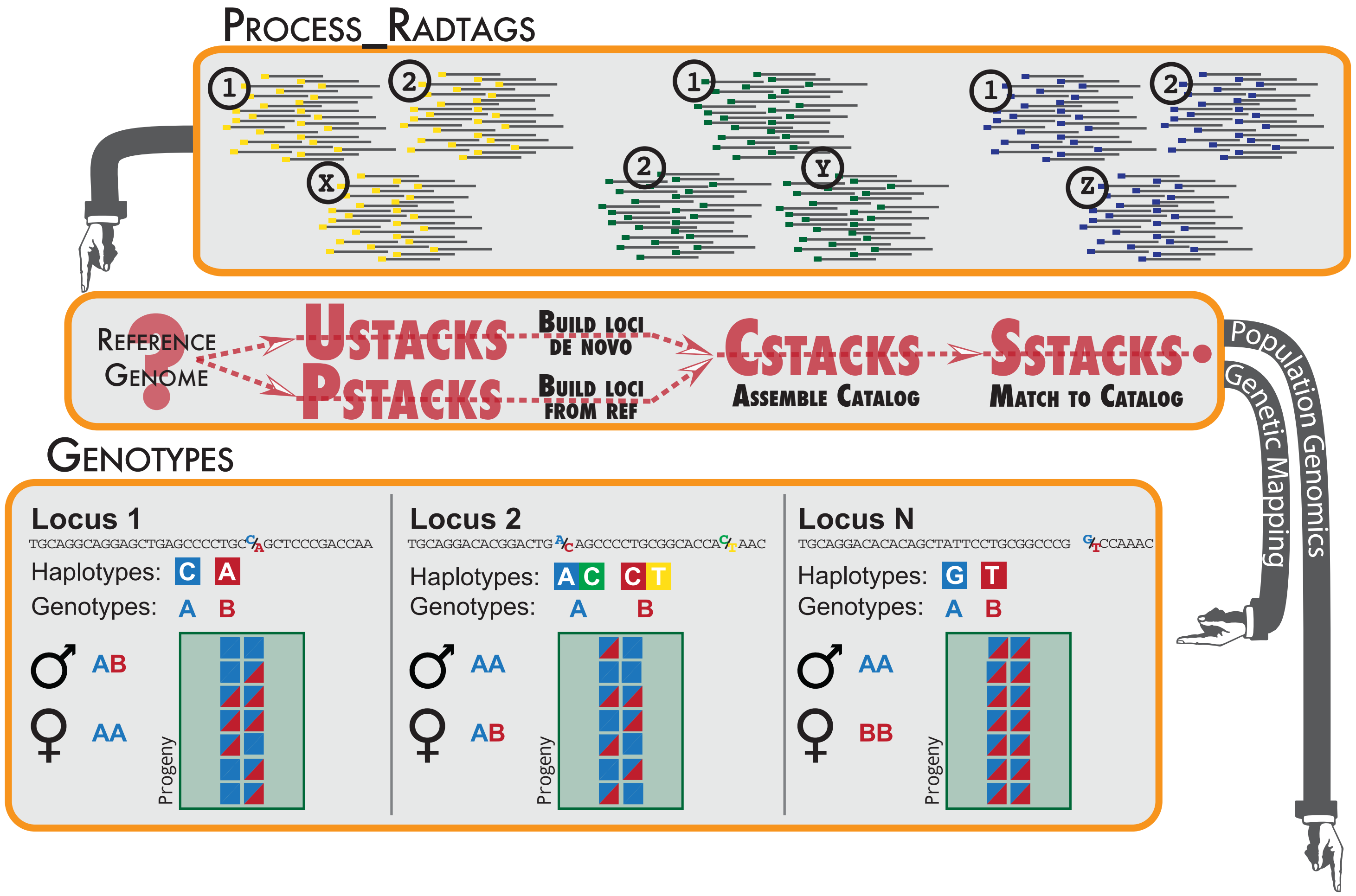
# *Stacks* workflow



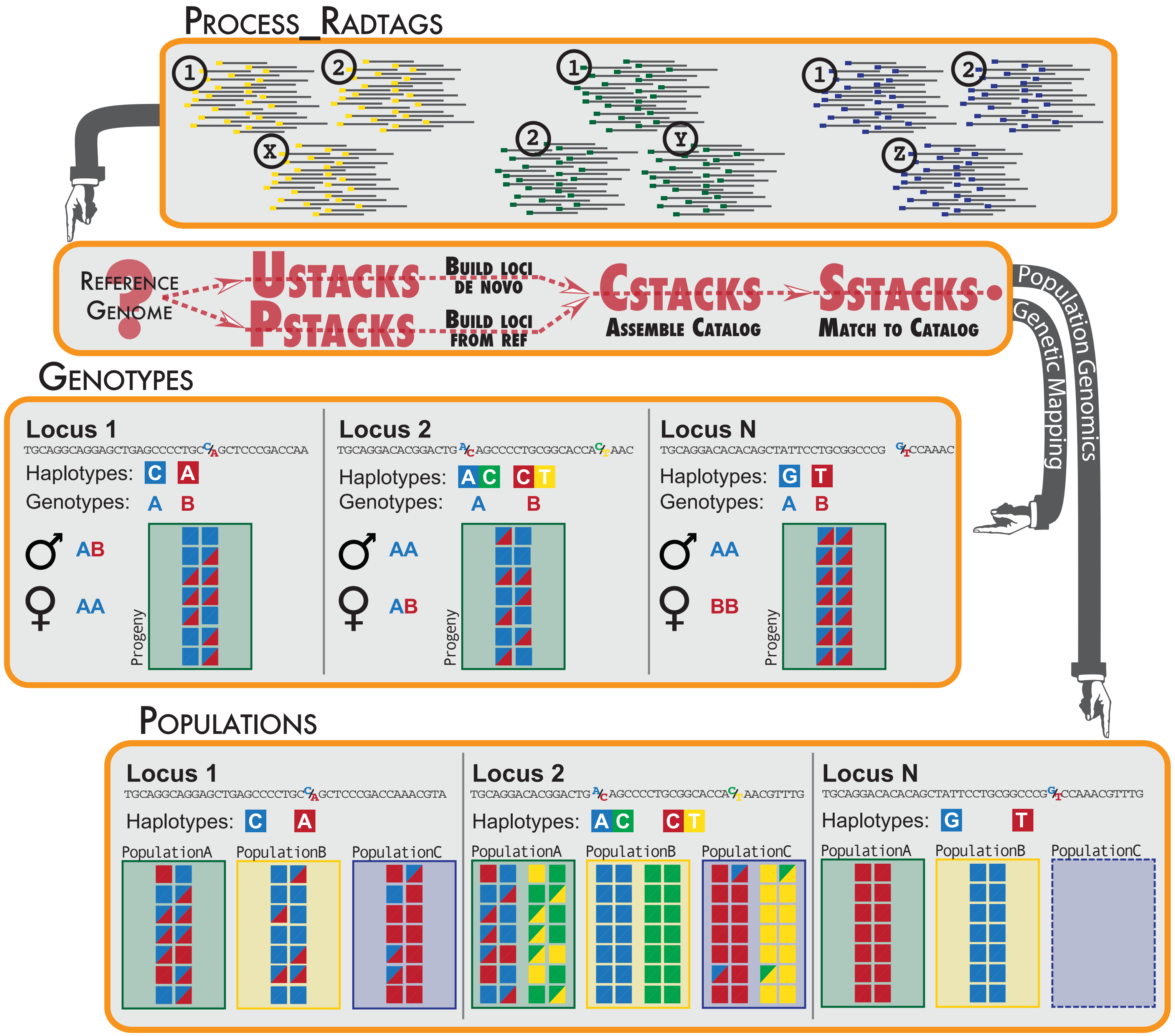
# *Stacks* workflow



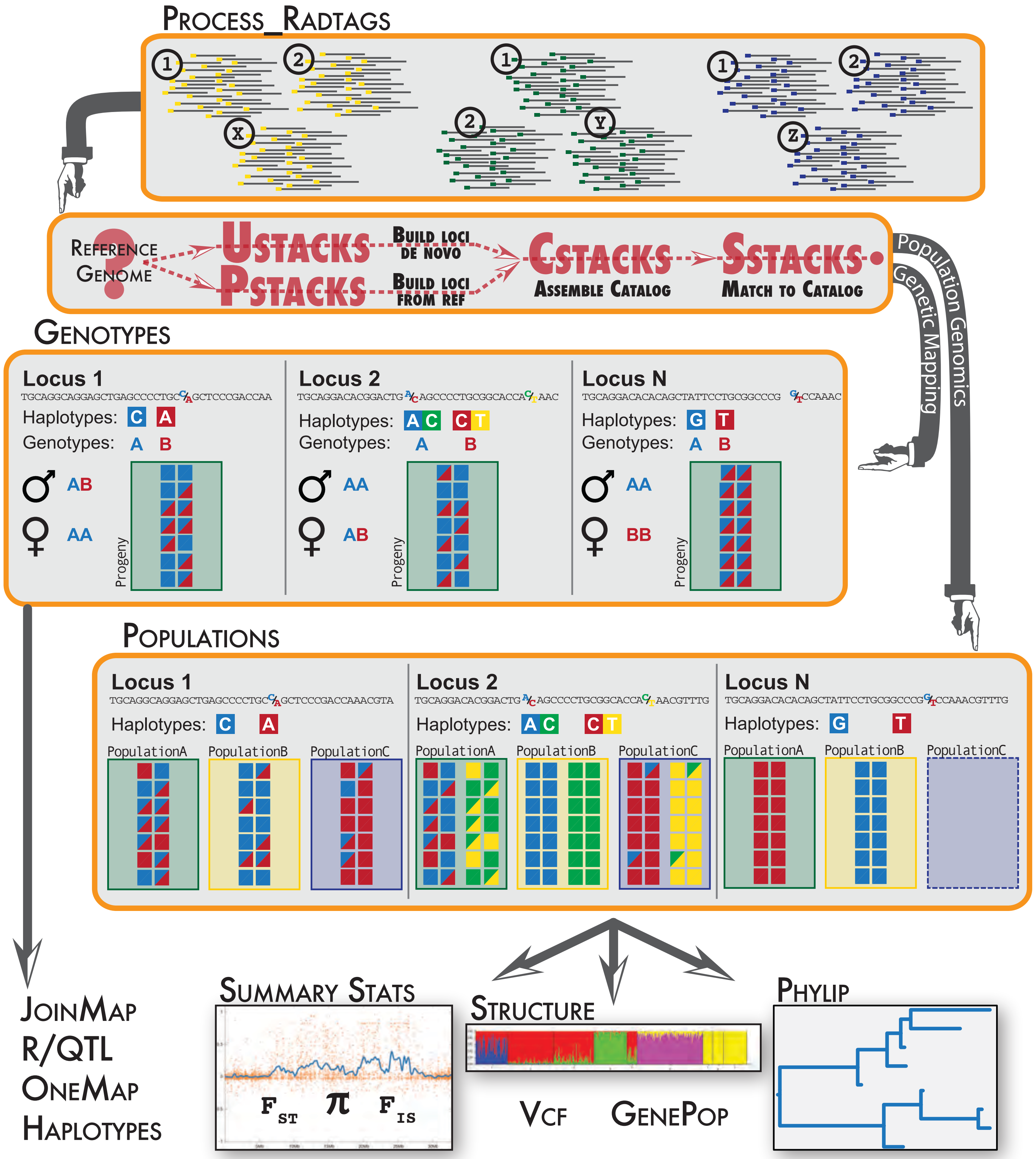
# *Stacks* workflow

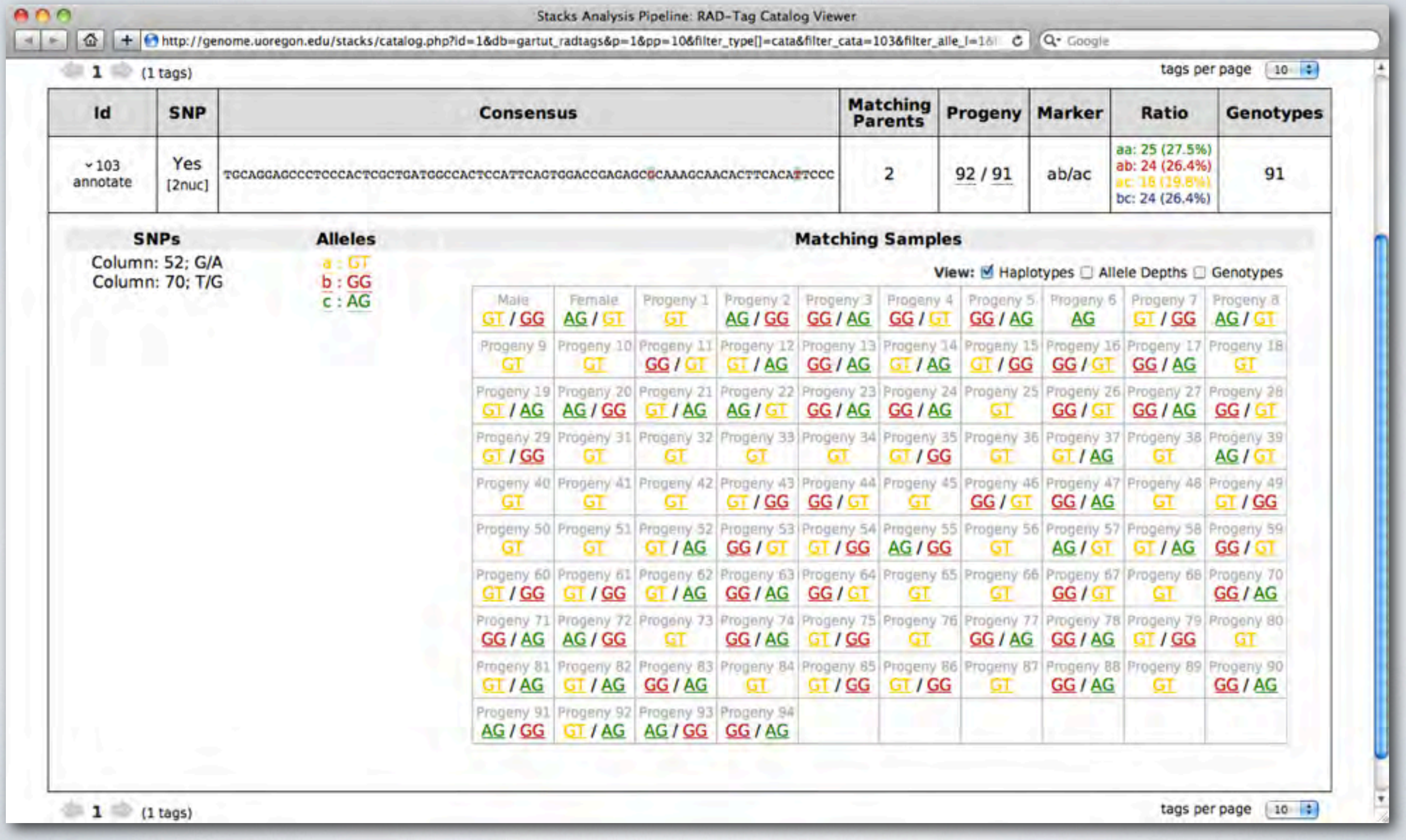


# *Stacks* workflow



# *Stacks* workflow





Stacks Analysis Pipeline: RAD-Tag Catalog Viewer
http://genome.uoregon.edu/stacks/catalog.php?id=1&db=gartut_radtags&p=1&pp=10&filter_type[]=cata&filter_cata=103&filter_alle_l=1&
Google
1 (1 tags)
tags per page 10

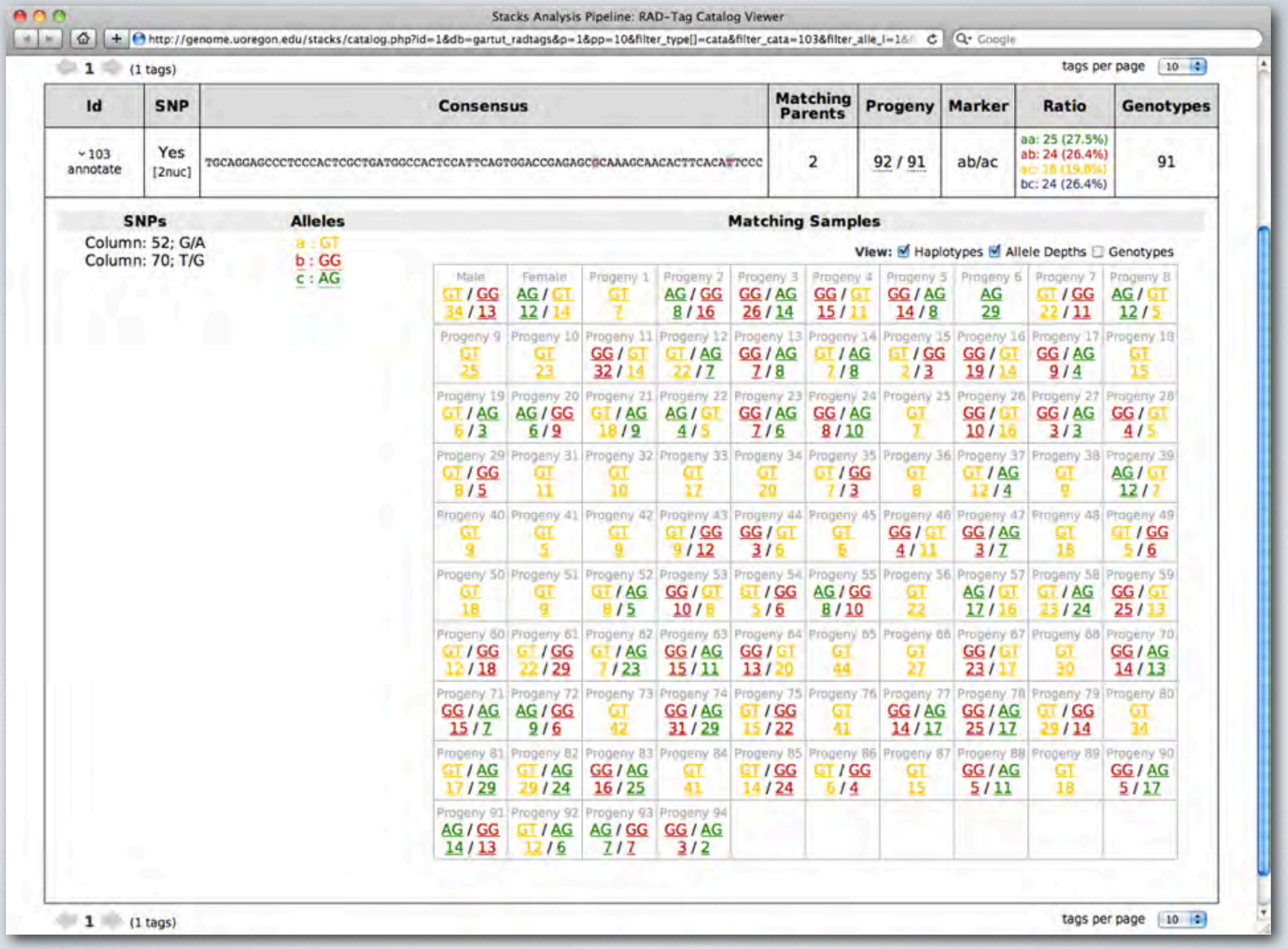
| Id | SNP | Consensus | Matching Parents | Progeny | Marker | Ratio | Genotypes |
|---|---|---|---|---|---|---|---|
| 103<br>annotate | Yes<br>[2nuc] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC | 2 | 92 / 91 | ab/ac | aa: 25 (27.5%)<br>ab: 24 (26.4%)<br>ac: 18 (19.8%)<br>bc: 24 (26.4%) | 91 |

| SNPs | Alleles |
|---|---|
| Column: 52; G/A | a : GT |
| Column: 70; T/G | b : GG |
| | c : AG |

**Matching Samples**

View: ☑ Haplotypes □ Allele Depths □ Genotypes

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Male<br>GT / GG | Female<br>AG / GT | Progeny 1<br>GT | Progeny 2<br>AG / GG | Progeny 3<br>GG / AG | Progeny 4<br>GG / GT | Progeny 5<br>GG / AG | Progeny 6<br>AG | Progeny 7<br>GT / GG | Progeny 8<br>AG / GT |
| Progeny 9<br>GT | Progeny 10<br>GT | Progeny 11<br>GG / GT | Progeny 12<br>GT / AG | Progeny 13<br>GG / AG | Progeny 14<br>GT / AG | Progeny 15<br>GT / GG | Progeny 16<br>GG / GT | Progeny 17<br>GG / AG | Progeny 18<br>GT |
| Progeny 19<br>GT / AG | Progeny 20<br>AG / GG | Progeny 21<br>GT / AG | Progeny 22<br>AG / GT | Progeny 23<br>GG / AG | Progeny 24<br>GG / AG | Progeny 25<br>GT | Progeny 26<br>GG / GT | Progeny 27<br>GG / AG | Progeny 28<br>GG / GT |
| Progeny 29<br>GT / GG | Progeny 31<br>GT | Progeny 32<br>GT | Progeny 33<br>GT | Progeny 34<br>GT | Progeny 35<br>GT / GG | Progeny 36<br>GT | Progeny 37<br>GT / AG | Progeny 38<br>GT | Progeny 39<br>AG / GT |
| Progeny 40<br>GT | Progeny 41<br>GT | Progeny 42<br>GT | Progeny 43<br>GT / GG | Progeny 44<br>GG / GT | Progeny 45<br>GT | Progeny 46<br>GG / GT | Progeny 47<br>GG / AG | Progeny 48<br>GT | Progeny 49<br>GT / GG |
| Progeny 50<br>GT | Progeny 51<br>GT | Progeny 52<br>GT / AG | Progeny 53<br>GG / GT | Progeny 54<br>GT / GG | Progeny 55<br>AG / GG | Progeny 56<br>GT | Progeny 57<br>AG / GT | Progeny 58<br>GT / AG | Progeny 59<br>GG / GT |
| Progeny 60<br>GT / GG | Progeny 61<br>GT / GG | Progeny 62<br>GT / AG | Progeny 63<br>GG / AG | Progeny 64<br>GG / GT | Progeny 65<br>GT | Progeny 66<br>GT | Progeny 67<br>GG / GT | Progeny 68<br>GT | Progeny 70<br>GG / AG |
| Progeny 71<br>GG / AG | Progeny 72<br>AG / GG | Progeny 73<br>GT | Progeny 74<br>GG / AG | Progeny 75<br>GT / GG | Progeny 76<br>GT | Progeny 77<br>GG / AG | Progeny 78<br>GG / AG | Progeny 79<br>GT / GG | Progeny 80<br>GT |
| Progeny 81<br>GT / AG | Progeny 82<br>GT / AG | Progeny 83<br>GG / AG | Progeny 84<br>GT | Progeny 85<br>GT / GG | Progeny 86<br>GT / GG | Progeny 87<br>GT | Progeny 88<br>GG / AG | Progeny 89<br>GT | Progeny 90<br>GG / AG |
| Progeny 91<br>AG / GG | Progeny 92<br>GT / AG | Progeny 93<br>AG / GG | Progeny 94<br>GG / AG | | | | | | |



Stacks Analysis Pipeline: RAD-Tag Catalog Viewer

http://genome.uoregon.edu/stacks/catalog.php?id=1&db=gartut_radtags&p=1&pp=10&filter_type[]=cata&filter_cata=103&filter_alle_l=1&

1 (1 tags) tags per page 10

| Id | SNP | Consensus | Matching Parents | Progeny | Marker | Ratio | Genotypes |
|---|---|---|---|---|---|---|---|
| 103<br>annotate | Yes<br>[2nuc] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC | 2 | 92 / 91 | ab/ac | aa: 25 (27.5%)<br>ab: 24 (26.4%)<br>ac: 18 (19.8%)<br>bc: 24 (26.4%) | 91 |

**SNPs**

Column: 52; G/A
Column: 70; T/G

**Alleles**

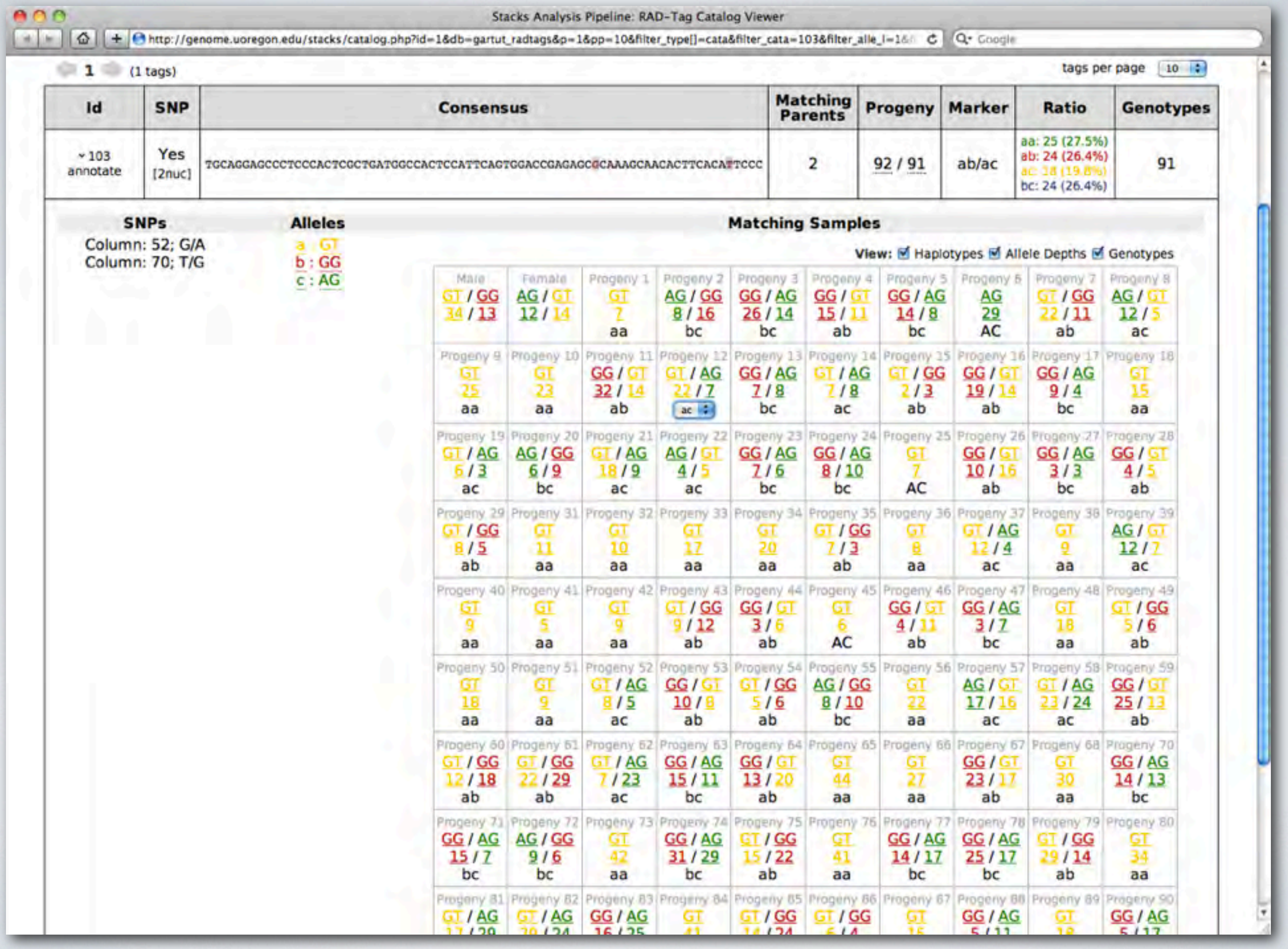
a : GT
b : GG
c : AG

**Matching Samples**

View: ☑ Haplotypes ☑ Allele Depths ☐ Genotypes

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Male<br>GT / GG<br>34 / 13 | Female<br>AG / GT<br>12 / 14 | Progeny 1<br>GT<br>7 | Progeny 2<br>AG / GG<br>8 / 16 | Progeny 3<br>GG / AG<br>26 / 14 | Progeny 4<br>GG / GT<br>15 / 11 | Progeny 5<br>GG / AG<br>14 / 8 | Progeny 6<br>AG<br>29 | Progeny 7<br>GT / GG<br>22 / 11 | Progeny 8<br>AG / GT<br>12 / 5 |
| Progeny 9<br>GT<br>25 | Progeny 10<br>GT<br>23 | Progeny 11<br>GG / GT<br>32 / 14 | Progeny 12<br>GT / AG<br>22 / 7 | Progeny 13<br>GG / AG<br>7 / 8 | Progeny 14<br>GT / AG<br>7 / 8 | Progeny 15<br>GT / GG<br>2 / 3 | Progeny 16<br>GG / GT<br>19 / 14 | Progeny 17<br>GG / AG<br>9 / 4 | Progeny 18<br>GT<br>15 |
| Progeny 19<br>GT / AG<br>6 / 3 | Progeny 20<br>AG / GG<br>6 / 9 | Progeny 21<br>GT / AG<br>18 / 9 | Progeny 22<br>AG / GT<br>4 / 5 | Progeny 23<br>GG / AG<br>7 / 6 | Progeny 24<br>GG / AG<br>8 / 10 | Progeny 25<br>GT<br>7 | Progeny 26<br>GG / GT<br>10 / 16 | Progeny 27<br>GG / AG<br>3 / 3 | Progeny 28<br>GG / GT<br>4 / 5 |
| Progeny 29<br>GT / GG<br>8 / 5 | Progeny 31<br>GT<br>11 | Progeny 32<br>GT<br>10 | Progeny 33<br>GT<br>17 | Progeny 34<br>GT<br>20 | Progeny 35<br>GT / GG<br>7 / 3 | Progeny 36<br>GT<br>8 | Progeny 37<br>GT / AG<br>12 / 4 | Progeny 38<br>GT<br>9 | Progeny 39<br>AG / GT<br>12 / 7 |
| Progeny 40<br>GT<br>9 | Progeny 41<br>GT<br>5 | Progeny 42<br>GT<br>9 | Progeny 43<br>GT / GG<br>9 / 12 | Progeny 44<br>GG / GT<br>3 / 6 | Progeny 45<br>GT<br>6 | Progeny 46<br>GG / GT<br>4 / 11 | Progeny 47<br>GG / AG<br>3 / 7 | Progeny 48<br>GT<br>18 | Progeny 49<br>GT / GG<br>5 / 6 |
| Progeny 50<br>GT<br>18 | Progeny 51<br>GT<br>9 | Progeny 52<br>GT / AG<br>8 / 5 | Progeny 53<br>GG / GT<br>10 / 8 | Progeny 54<br>GT / GG<br>5 / 6 | Progeny 55<br>AG / GG<br>8 / 10 | Progeny 56<br>GT<br>22 | Progeny 57<br>AG / GT<br>17 / 16 | Progeny 58<br>GT / AG<br>23 / 24 | Progeny 59<br>GG / GT<br>25 / 13 |
| Progeny 60<br>GT / GG<br>12 / 18 | Progeny 61<br>GT / GG<br>22 / 29 | Progeny 62<br>GT / AG<br>7 / 23 | Progeny 63<br>GG / AG<br>15 / 11 | Progeny 64<br>GG / GT<br>13 / 20 | Progeny 65<br>GT<br>44 | Progeny 66<br>GT<br>27 | Progeny 67<br>GG / GT<br>23 / 17 | Progeny 68<br>GT<br>30 | Progeny 70<br>GG / AG<br>14 / 13 |
| Progeny 71<br>GG / AG<br>15 / 7 | Progeny 72<br>AG / GG<br>9 / 6 | Progeny 73<br>GT<br>42 | Progeny 74<br>GG / AG<br>31 / 29 | Progeny 75<br>GT / GG<br>15 / 22 | Progeny 76<br>GT<br>41 | Progeny 77<br>GG / AG<br>14 / 17 | Progeny 78<br>GG / AG<br>25 / 17 | Progeny 79<br>GT / GG<br>29 / 14 | Progeny 80<br>GT<br>34 |
| Progeny 81<br>GT / AG<br>17 / 29 | Progeny 82<br>GT / AG<br>29 / 24 | Progeny 83<br>GG / AG<br>16 / 25 | Progeny 84<br>GT<br>41 | Progeny 85<br>GT / GG<br>14 / 24 | Progeny 86<br>GT / GG<br>6 / 4 | Progeny 87<br>GT<br>15 | Progeny 88<br>GG / AG<br>5 / 11 | Progeny 89<br>GT<br>18 | Progeny 90<br>GG / AG<br>5 / 17 |
| Progeny 91<br>AG / GG<br>14 / 13 | Progeny 92<br>GT / AG<br>12 / 6 | Progeny 93<br>AG / GG<br>7 / 7 | Progeny 94<br>GG / AG<br>3 / 2 | | | | | | |

1 (1 tags) tags per page 10


Stacks Analysis Pipeline: RAD-Tag Catalog Viewer
http://genome.uoregon.edu/stacks/catalog.php?id=1&db=gartut_radtags&p=1&pp=10&filter_type[]=cata&filter_cata=103&filter_alle_l=1&
Google
1 (1 tags)
tags per page 10

| Id | SNP | Consensus | Matching Parents | Progeny | Marker | Ratio | Genotypes |
|---|---|---|---|---|---|---|---|
| 103<br>annotate | Yes<br>[2nuc] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC | 2 | 92 / 91 | ab/ac | aa: 25 (27.5%)<br>ab: 24 (26.4%)<br>ac: 18 (19.8%)<br>bc: 24 (26.4%) | 91 |

**SNPs**

Column: 52; G/A
Column: 70; T/G

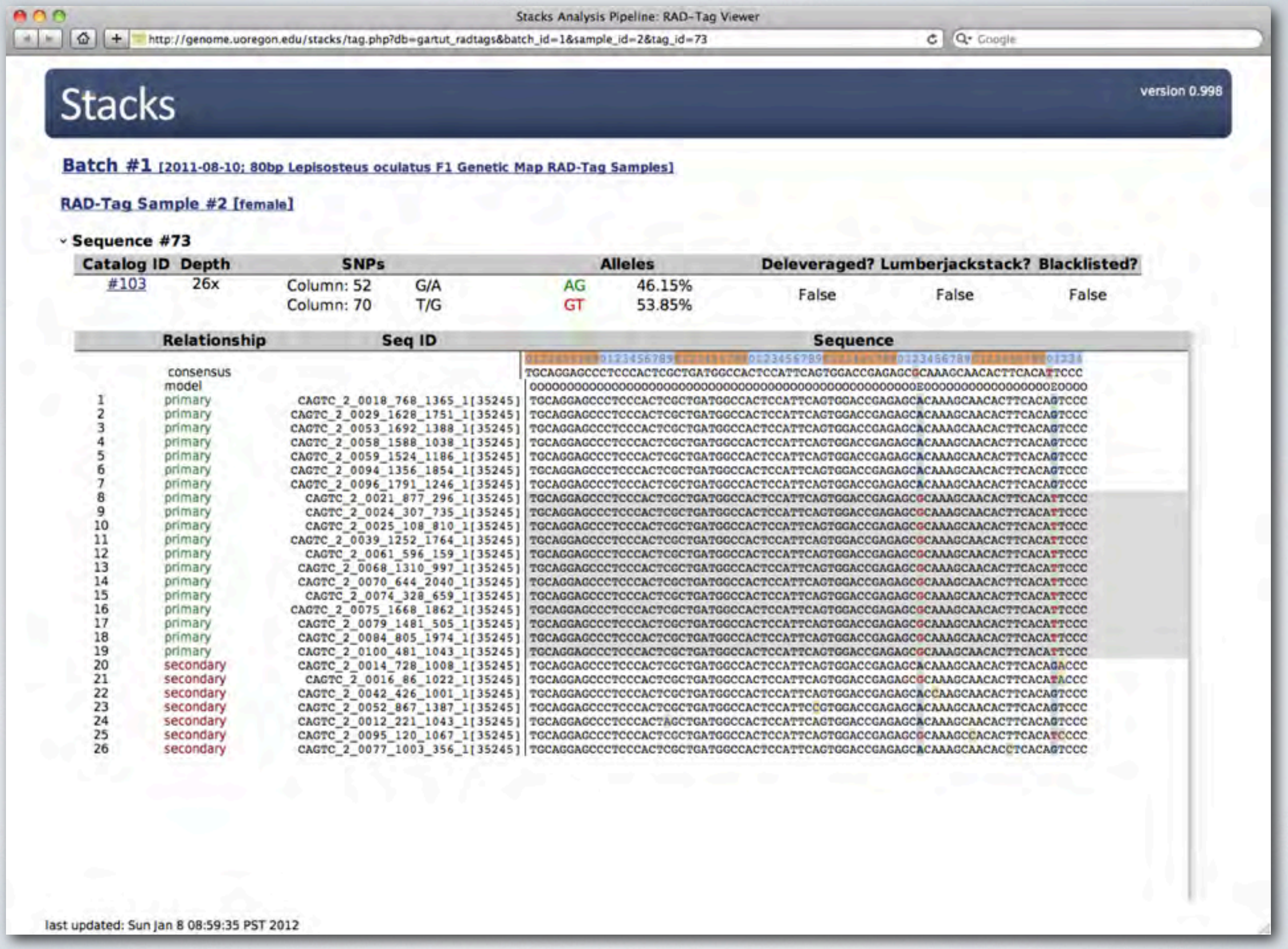
**Alleles**

a : GT
b : GG
c : AG

**Matching Samples**

View: ☑ Haplotypes ☑ Allele Depths ☑ Genotypes

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Male<br>GT / GG<br>34 / 13 | Female<br>AG / GT<br>12 / 14 | Progeny 1<br>GT<br>7<br>aa | Progeny 2<br>AG / GG<br>8 / 16<br>bc | Progeny 3<br>GG / AG<br>26 / 14<br>bc | Progeny 4<br>GG / GT<br>15 / 11<br>ab | Progeny 5<br>GG / AG<br>14 / 8<br>bc | Progeny 6<br>AG<br>29<br>AC | Progeny 7<br>GT / GG<br>22 / 11<br>ab | Progeny 8<br>AG / GT<br>12 / 5<br>ac |
| Progeny 9<br>GT<br>25<br>aa | Progeny 10<br>GT<br>23<br>aa | Progeny 11<br>GG / GT<br>32 / 14<br>ab | Progeny 12<br>GT / AG<br>22 / 7<br>ac | Progeny 13<br>GG / AG<br>7 / 8<br>bc | Progeny 14<br>GT / AG<br>7 / 8<br>ac | Progeny 15<br>GT / GG<br>2 / 3<br>ab | Progeny 16<br>GG / GT<br>19 / 14<br>ab | Progeny 17<br>GG / AG<br>9 / 4<br>bc | Progeny 18<br>GT<br>15<br>aa |
| Progeny 19<br>GT / AG<br>6 / 3<br>ac | Progeny 20<br>AG / GG<br>6 / 9<br>bc | Progeny 21<br>GT / AG<br>18 / 9<br>ac | Progeny 22<br>AG / GT<br>4 / 5<br>ac | Progeny 23<br>GG / AG<br>7 / 6<br>bc | Progeny 24<br>GG / AG<br>8 / 10<br>bc | Progeny 25<br>GT<br>7<br>AC | Progeny 26<br>GG / GT<br>10 / 16<br>ab | Progeny 27<br>GG / AG<br>3 / 3<br>bc | Progeny 28<br>GG / GT<br>4 / 5<br>ab |
| Progeny 29<br>GT / GG<br>8 / 5<br>ab | Progeny 31<br>GT<br>11<br>aa | Progeny 32<br>GT<br>10<br>aa | Progeny 33<br>GT<br>17<br>aa | Progeny 34<br>GT<br>20<br>aa | Progeny 35<br>GT / GG<br>7 / 3<br>ab | Progeny 36<br>GT<br>8<br>aa | Progeny 37<br>GT / AG<br>12 / 4<br>ac | Progeny 38<br>GT<br>9<br>aa | Progeny 39<br>AG / GT<br>12 / 7<br>ac |
| Progeny 40<br>GT<br>9<br>aa | Progeny 41<br>GT<br>5<br>aa | Progeny 42<br>GT<br>9<br>aa | Progeny 43<br>GT / GG<br>9 / 12<br>ab | Progeny 44<br>GG / GT<br>3 / 6<br>ab | Progeny 45<br>GT<br>6<br>AC | Progeny 46<br>GG / GT<br>4 / 11<br>ab | Progeny 47<br>GG / AG<br>3 / 7<br>bc | Progeny 48<br>GT<br>18<br>aa | Progeny 49<br>GT / GG<br>5 / 6<br>ab |
| Progeny 50<br>GT<br>18<br>aa | Progeny 51<br>GT<br>9<br>aa | Progeny 52<br>GT / AG<br>8 / 5<br>ac | Progeny 53<br>GG / GT<br>10 / 8<br>ab | Progeny 54<br>GT / GG<br>5 / 6<br>ab | Progeny 55<br>AG / GG<br>8 / 10<br>bc | Progeny 56<br>GT<br>22<br>aa | Progeny 57<br>AG / GT<br>17 / 16<br>ac | Progeny 58<br>GT / AG<br>23 / 24<br>ac | Progeny 59<br>GG / GT<br>25 / 13<br>ab |
| Progeny 60<br>GT / GG<br>12 / 18<br>ab | Progeny 61<br>GT / GG<br>22 / 29<br>ab | Progeny 62<br>GT / AG<br>7 / 23<br>ac | Progeny 63<br>GG / AG<br>15 / 11<br>bc | Progeny 64<br>GG / GT<br>13 / 20<br>ab | Progeny 65<br>GT<br>44<br>aa | Progeny 66<br>GT<br>27<br>aa | Progeny 67<br>GG / GT<br>23 / 17<br>ab | Progeny 68<br>GT<br>30<br>aa | Progeny 70<br>GG / AG<br>14 / 13<br>bc |
| Progeny 71<br>GG / AG<br>15 / 7<br>bc | Progeny 72<br>AG / GG<br>9 / 6<br>bc | Progeny 73<br>GT<br>42<br>aa | Progeny 74<br>GG / AG<br>31 / 29<br>bc | Progeny 75<br>GT / GG<br>15 / 22<br>ab | Progeny 76<br>GT<br>41<br>aa | Progeny 77<br>GG / AG<br>14 / 17<br>bc | Progeny 78<br>GG / AG<br>25 / 17<br>bc | Progeny 79<br>GT / GG<br>29 / 14<br>ab | Progeny 80<br>GT<br>34<br>aa |
| Progeny 81<br>GT / AG<br>17 / 29 | Progeny 82<br>GT / AG<br>20 / 24 | Progeny 83<br>GG / AG<br>16 / 25 | Progeny 84<br>GT<br>41 | Progeny 85<br>GT / GG<br>14 / 24 | Progeny 86<br>GT / GG<br>6 / 4 | Progeny 87<br>GT<br>15 | Progeny 88<br>GG / AG<br>5 / 11 | Progeny 89<br>GT<br>18 | Progeny 90<br>GG / AG<br>5 / 17 |


Stacks Analysis Pipeline: RAD-Tag Viewer
http://genome.uoregon.edu/stacks/tag.php?db=gartut_radtags&batch_id=1&sample_id=2&tag_id=73
Google

# Stacks

version 0.998

## Batch #1 [2011-08-10; 80bp Lepisosteus oculatus F1 Genetic Map RAD-Tag Samples]

### RAD-Tag Sample #2 [female]

#### Sequence #73

| Catalog ID | Depth | SNPs | | Alleles | | Deleveraged? | Lumberjackstack? | Blacklisted? |
|---|---|---|---|---|---|---|---|---|
| #103 | 26x | Column: 52<br>Column: 70 | G/A<br>T/G | AG<br>GT | 46.15%<br>53.85% | False | False | False |

| | Relationship | Seq ID | Sequence |
|---|---|---|---|
| | | | 012345678901234567890123456789012345678901234567890123456789012345678901234 |
| | consensus | | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| | model | | 0000000000000000000000000000000000000000000000000000E00000000000000000E0000 |
| 1 | primary | CAGTC_2_0018_768_1365_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 2 | primary | CAGTC_2_0029_1628_1751_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 3 | primary | CAGTC_2_0053_1692_1388_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 4 | primary | CAGTC_2_0058_1588_1038_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 5 | primary | CAGTC_2_0059_1524_1186_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 6 | primary | CAGTC_2_0094_1356_1854_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 7 | primary | CAGTC_2_0096_1791_1246_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 8 | primary | CAGTC_2_0021_877_296_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 9 | primary | CAGTC_2_0024_307_735_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 10 | primary | CAGTC_2_0025_108_810_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 11 | primary | CAGTC_2_0039_1252_1764_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 12 | primary | CAGTC_2_0061_596_159_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 13 | primary | CAGTC_2_0068_1310_997_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 14 | primary | CAGTC_2_0070_644_2040_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 15 | primary | CAGTC_2_0074_328_659_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 16 | primary | CAGTC_2_0075_1668_1862_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 17 | primary | CAGTC_2_0079_1481_505_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 18 | primary | CAGTC_2_0084_805_1974_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 19 | primary | CAGTC_2_0100_481_1043_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC |
| 20 | secondary | CAGTC_2_0014_728_1008_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGACCC |
| 21 | secondary | CAGTC_2_0016_86_1022_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATACCC |
| 22 | secondary | CAGTC_2_0042_426_1001_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACCAAGCAACACTTCACAGTCCC |
| 23 | secondary | CAGTC_2_0052_867_1387_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCCGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 24 | secondary | CAGTC_2_0012_221_1043_1[35245] | TGCAGGAGCCCTCCCACTAGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACTTCACAGTCCC |
| 25 | secondary | CAGTC_2_0095_120_1067_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCCACACTTCACATCCCC |
| 26 | secondary | CAGTC_2_0077_1003_356_1[35245] | TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCACAAAGCAACACCTCACAGTCCC |

last updated: Sun Jan 8 08:59:35 PST 2012