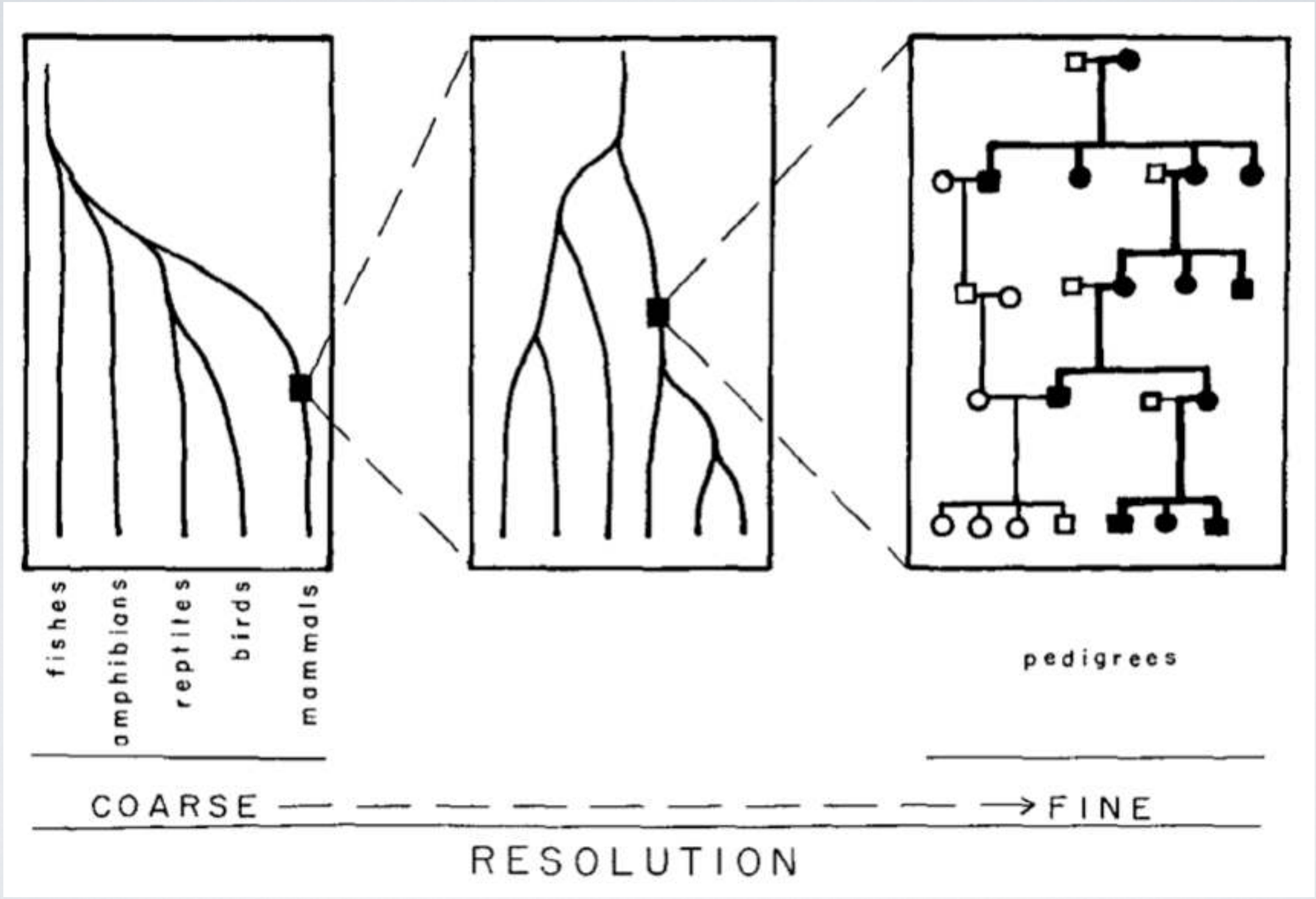
Genomic analyses of non-model organisms with RAD-seq and Stacks

Julian Catchen
jcatchen@uoregon.edu

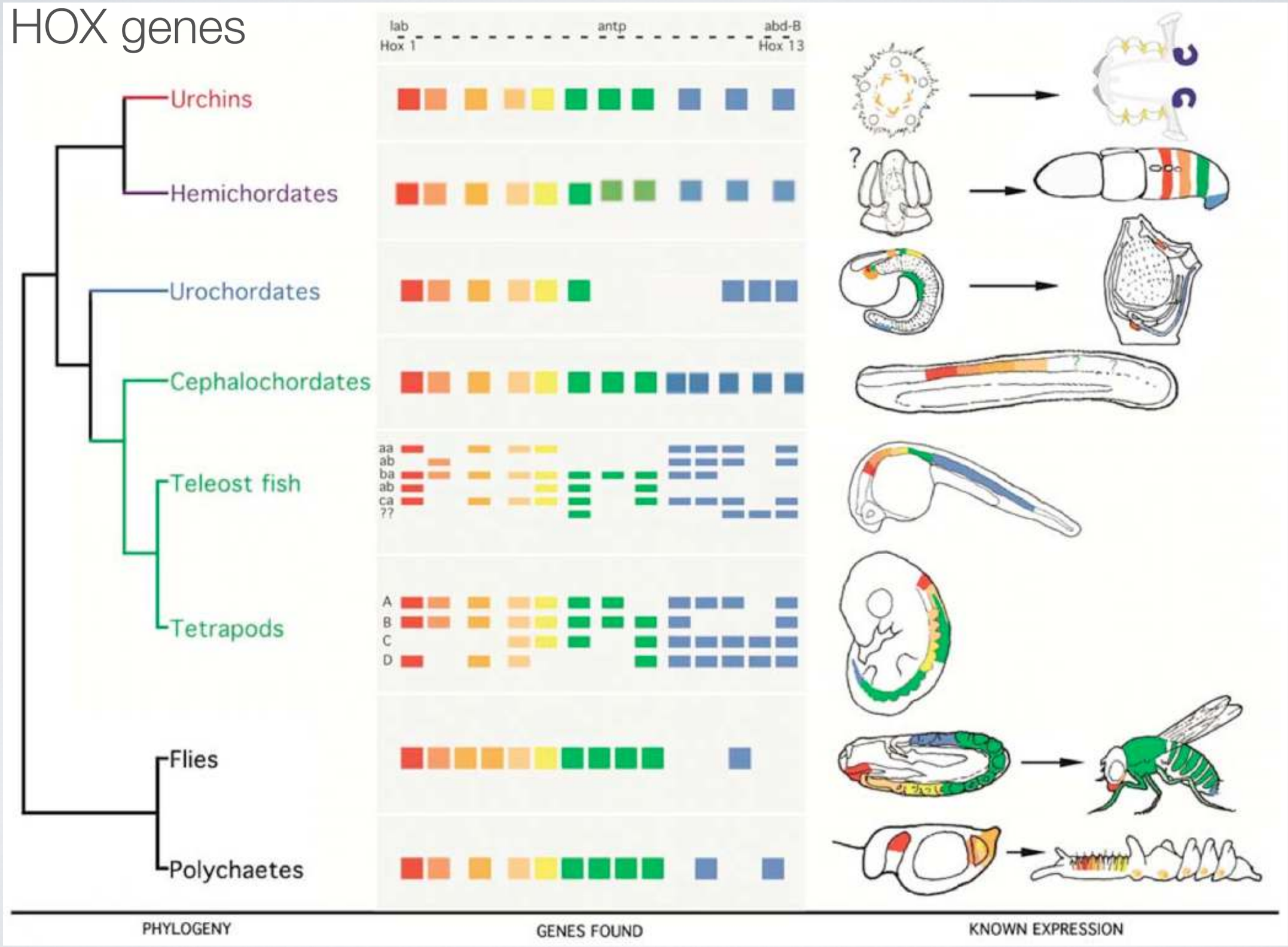
Cresko Lab
Institute of Ecology and Evolution
University of Oregon

"A bridge between population genetics and systematics"



Avise, et al, 1987

A common genetic toolkit



Swalla, 2006

What makes an organism a model?

classic models: fruit fly, mouse, zebrafish

1. Genetic Maps (phenotypes, allozymes, microsats)
2. Physical Maps (random, shotgun Sanger sequencing)
3. Transcriptomic Maps (EST sequencing)
4. Gene Expression Analyses (microarrays, RNA-seq)

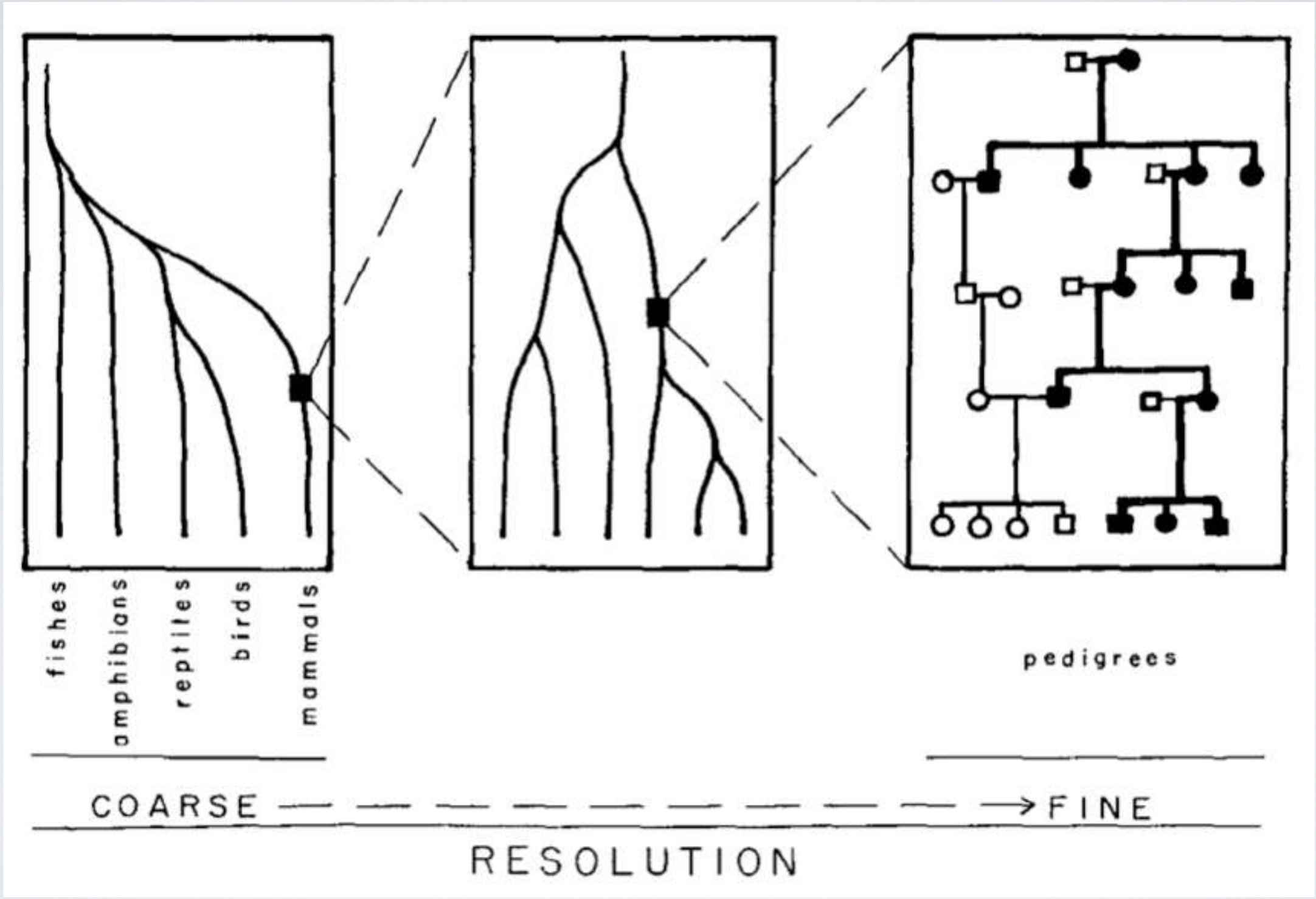
What makes an organism a model?

Organism	Markers	Organism	Markers
Bighorn Sheep	147	Silver carp	483
Shrimp	418	Guppy	790
Coral	420	Barramundi	240
Indian mustard	1,029	Catfish	331
Oilseed rape	13,551	Sea bass	368
Black spruce	1,111	Cichlid	204
Barley	2,111	Platyfish	290
Flounder	1,375	Halibut	604
Turbot	242	Sea bream	204
Human	**250,000**	**Mouse**	**10,000**

Short-read sequencing represents a new “bridge”

Rich Niewiroski Jr.
http://www.projectrich.com/gallery

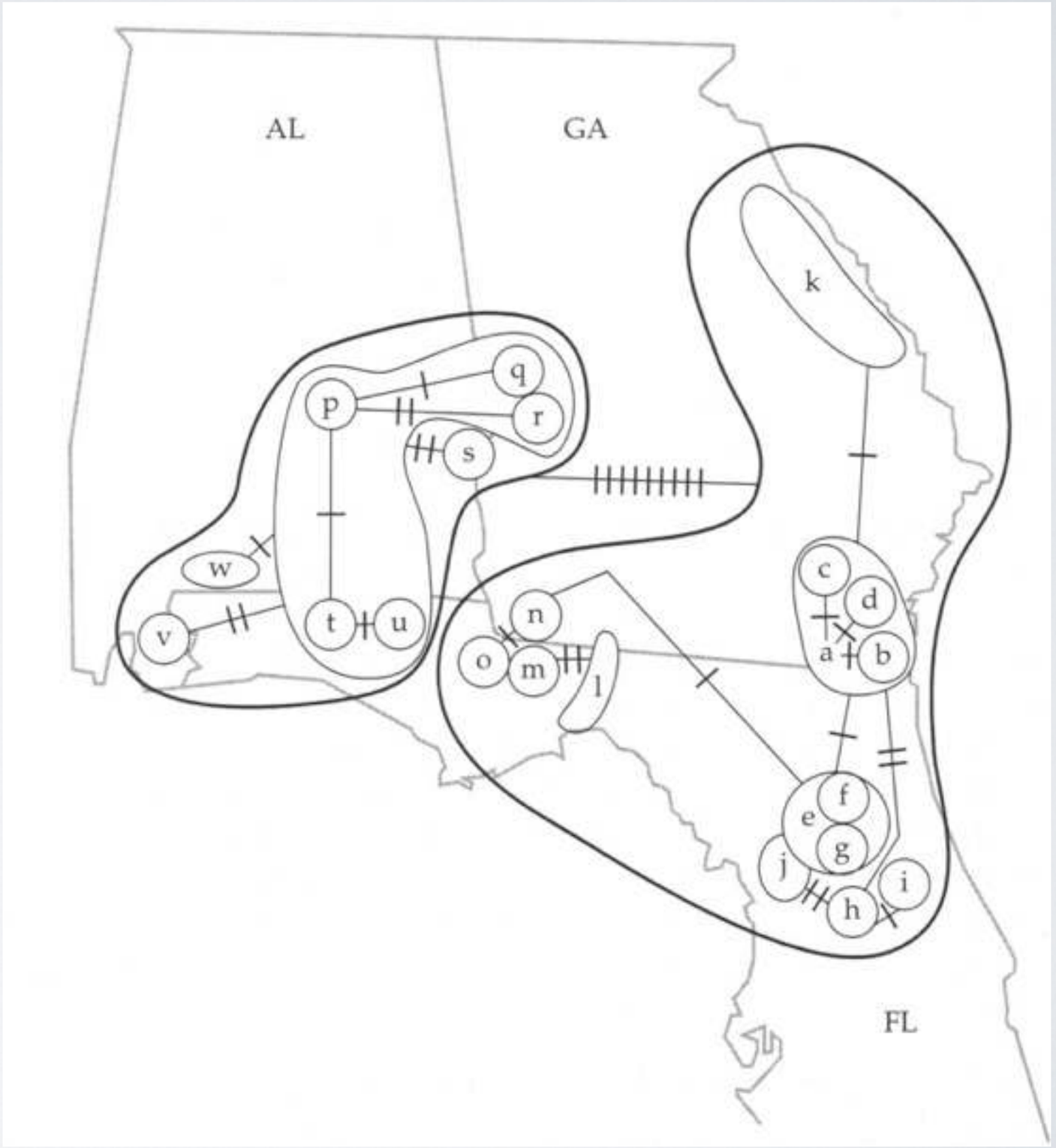
"A bridge between population genetics and systematics"



Avise, et al, 1987

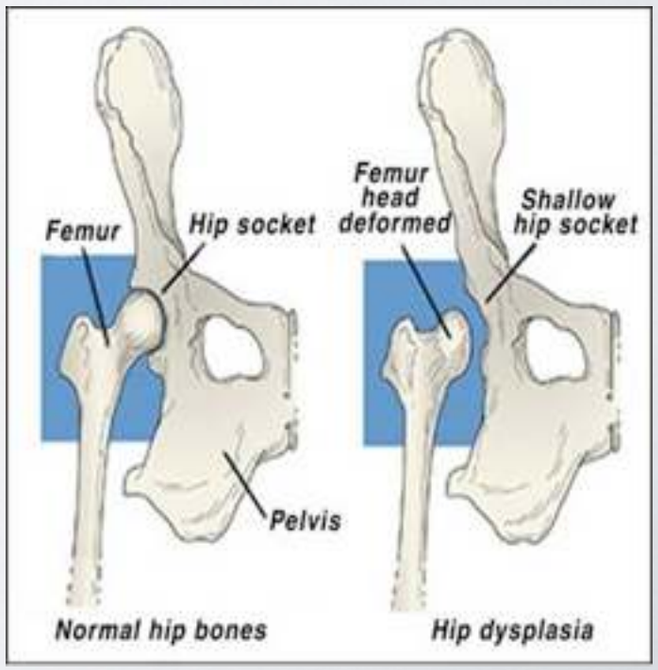
Fine scale estimates of population structure
Estimates of genome-region specific gene flow

Phylogeography of pocket gophers using mtDNA



Avise, 1979

Precisely quantify the amount of inbreeding in wild and captive populations



Identify signatures of selection in natural populations
Associate genotype to phenotype using GWAS and QTL

Fill important taxonomic nodes with genetic maps
Quantify genome evolution and structural variation

Sequence everything?

Why not sequence the entire genome?

Sample space has scaled with sequencing capacity

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

Why not sequence the entire genome?

Sample space has scaled with sequencing capacity

- 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, full sequence isn't necessary

- the genomes of many organisms are organized in linkage blocks
- well spaced markers will provide all the necessary coverage
- the cost of genotyping will almost always be a fraction

Why not sequence the entire genome?

Sample space has scaled with sequencing capacity

- 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, full sequence isn't necessary

- the genomes of many organisms are organized in linkage blocks
- well spaced markers will provide all the necessary coverage
- the cost of genotyping will almost always be a fraction

Genetic maps are relevant again

- a high density genetic map can facilitate genome assembly
- genomes may be segregating a lot of structural variation
- assaying this variation can presently be done best with genetic maps

Alternative approach - Reduced representation sequencing for genotyping

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

Alternative approach - Reduced representation sequencing for genotyping

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

- Simultaneous identification and typing of single nucleotide polymorphisms (SNPs)

Alternative approach - Reduced representation sequencing for genotyping

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

Alternative approach - Reduced representation sequencing for genotyping

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

- 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

Alternative approach - Reduced representation sequencing for genotyping

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

- 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
  - If the genetic architecture of a trait involves many rare variants
  - Inferring patterns of LD may be easiest with full sequences

Outline

1. RAD-seq and other RRL methods
    - Experimental design considerations
2. Anatomy of a PstI RAD-seq Analysis
3. Signatures of natural selection across the genome with the threespine stickleback
    - SNP detection and kernel Smoothing
4. Building a genetic map of the spotted gar
    - Stacks: the gory details
5. Assembling the genome of the platyfish
6. Phylogeography of the pitcher plant mosquito
7. RAD Synthesis with the gulf pipefish
8. Gene expression quantification with eRAD

Different types of reduced representation sequencing for genotyping

Nature Reviews Genetics, 2011

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone‖, Julian M. Catchen‡ and Mark L. Blaxter*¶*

Different types of reduced representation sequencing for genot

CRoPS: similar to RRL

uses AFLP: two restriction enzymes

MSG: similar to GBS

replaces size selection step with very frequent cutter

Nature Reviews Genetics, 2011

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone‖, Julian M. Catchen‡ and Mark L. Blaxter*¶*

What is RAD-seq?

(Restriction-site Associated DNA)

Eric Johnson

[redacted]

Joe Dunham

[redacted]

Mike Miller

Illumina

2007

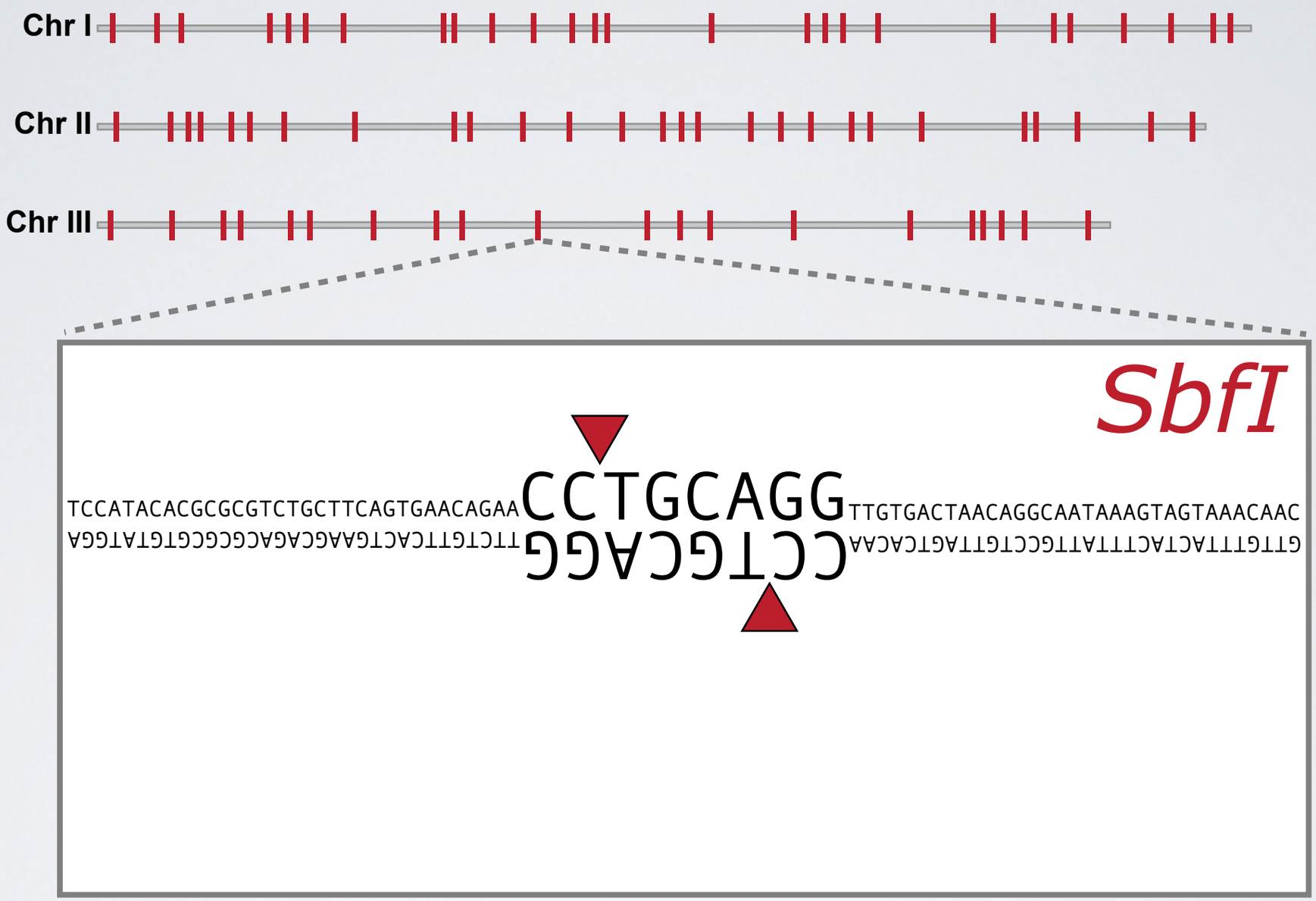
[redacted]

2008

[redacted]

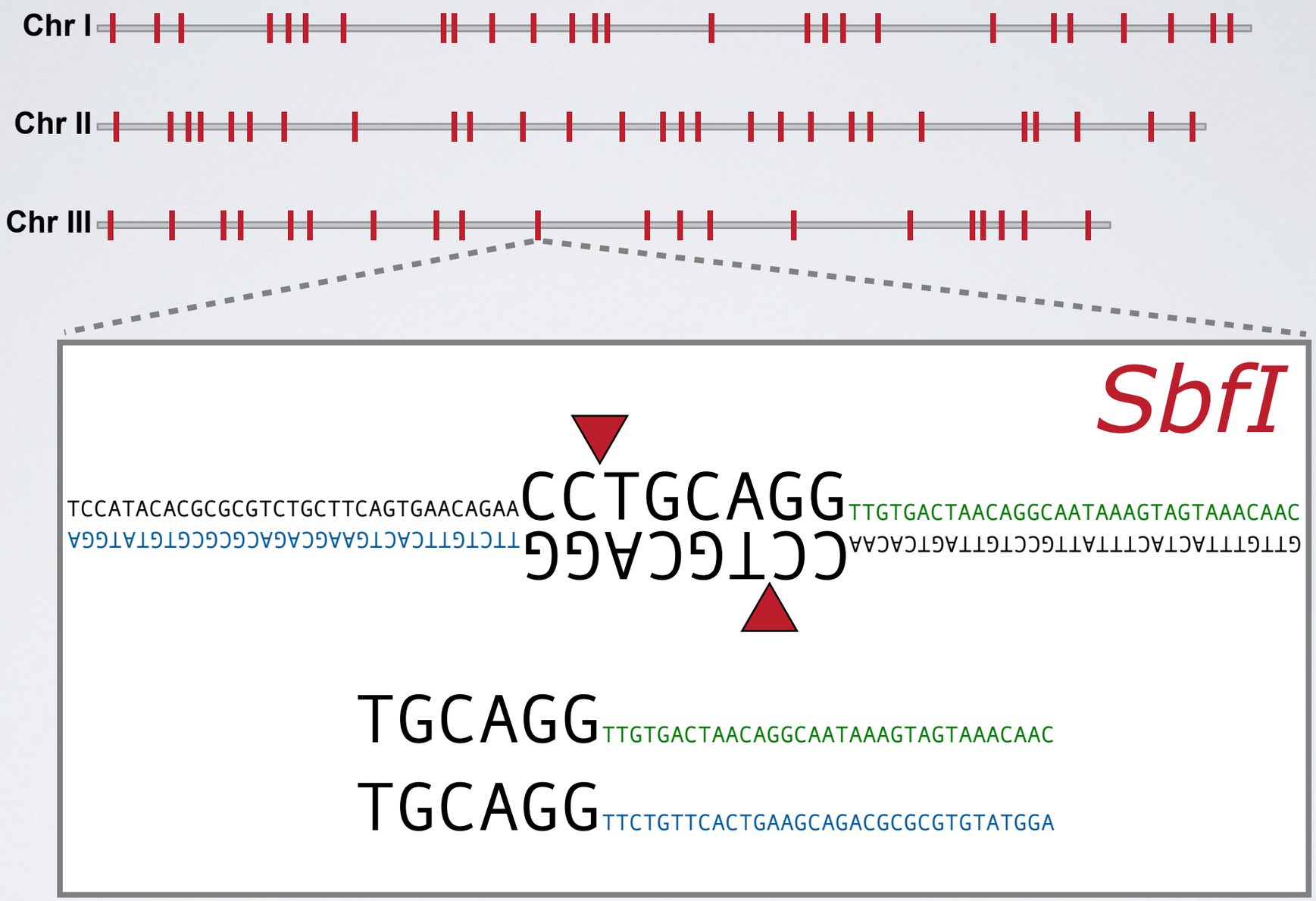
What is RAD-seq?

(Restriction-site Associated DNA)



What is RAD-seq?

(Restriction-site Associated DNA)

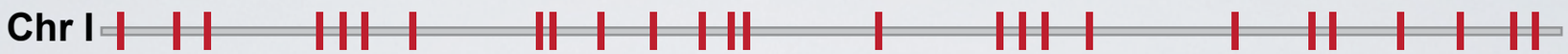


What is RAD-seq?

(Restriction-site Associated DNA)

What is RAD-seq?

(Restriction-site Associated DNA)



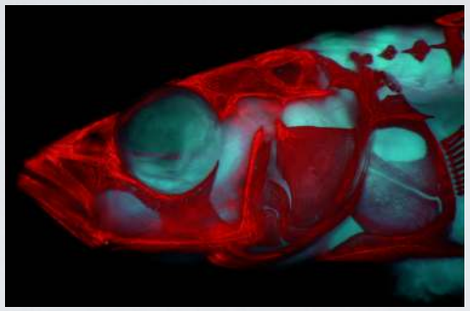
22,830 *SbfI* sites in Stickleback

~ 45,000 RAD-Tags

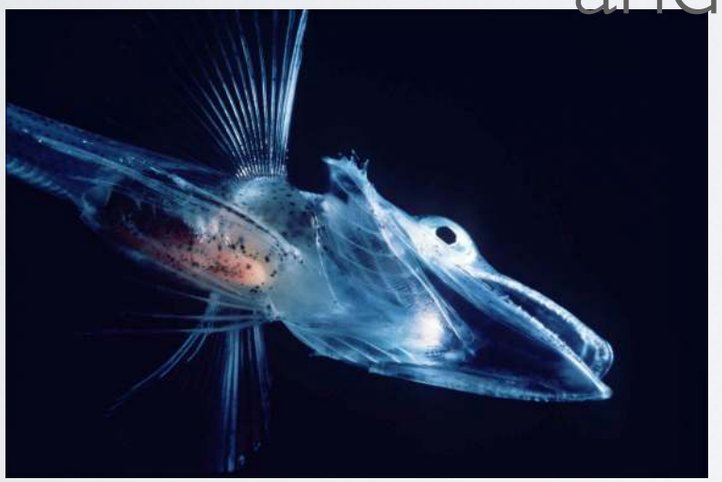
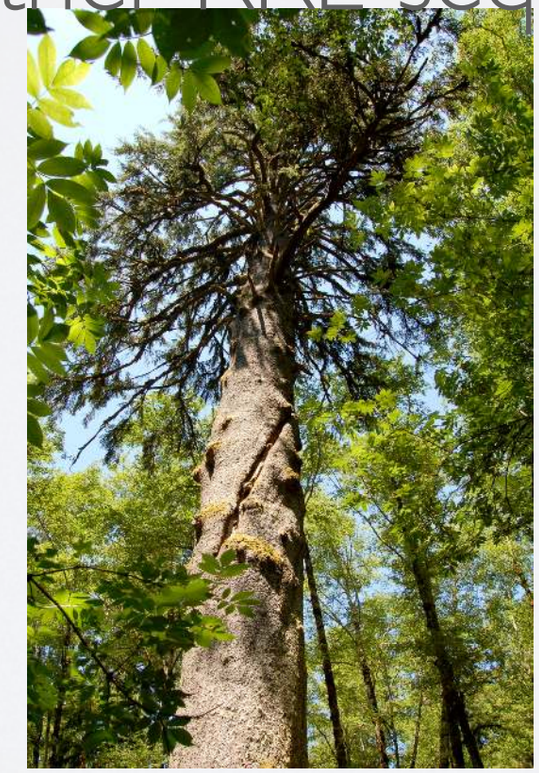
HiSeq Illumina Lane:
100 million reads, 96 barcoded individuals

1

2


Wet lab and statistical considerations for RAD-seq and other RRL-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**

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

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

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 locus
```

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

Don't forget about sequencing error and sample bias

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?

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 $(0.25)^n$ x 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 $(0.25)^n$ x 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

Coding regions are GC rich - restriction recognition sequence can bias where tags fall.

e.g., in stickleback coding regions are 10% of the genome, but:
50% of SbfI sites (CCTGCAGG) fall in exons
14% of EcoRI sites (GAATTC) fall in exons

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone‖, Julian M. Catchen‡ and Mark L. Blaxter*¶*

Nature Reviews Genetics 2011

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

60 nt single end: sample 22,800 SbfI sites, 2 tags per site, 49 nt = 2.23 Mb = 22,300 SNPs
If N = 100, about 48% of tags have a SNP

80 nt single end: 45,600 tags, 69 nt = 3.15 Mb; 60% of tags have a SNP

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone‖, Julian M. Catchen‡ and Mark L. Blaxter*¶*

Nature Reviews Genetics 2011

Experimental design considerations for RAD

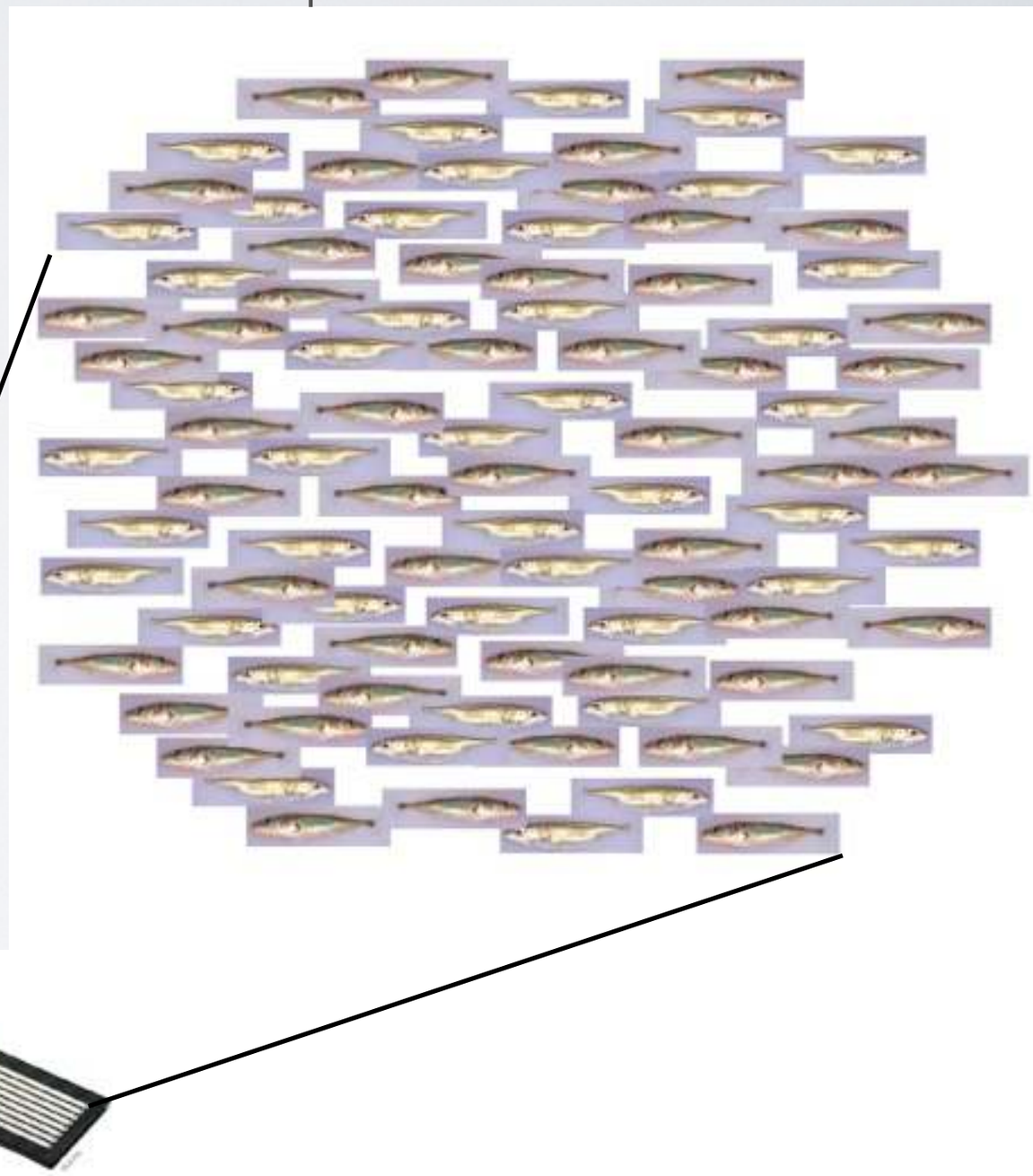
Tradeoffs:
Number of sites versus **Depth** of sequencing per site versus **Number of samples**

How many samples should be multiplexed?

Things to consider

Barcoded adapters
- 5 nt barcodes
- Variable length barcodes
- Combinatorial barcodes (PE)

Barcode distance - read recovery



Experimental design considerations for RAD

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

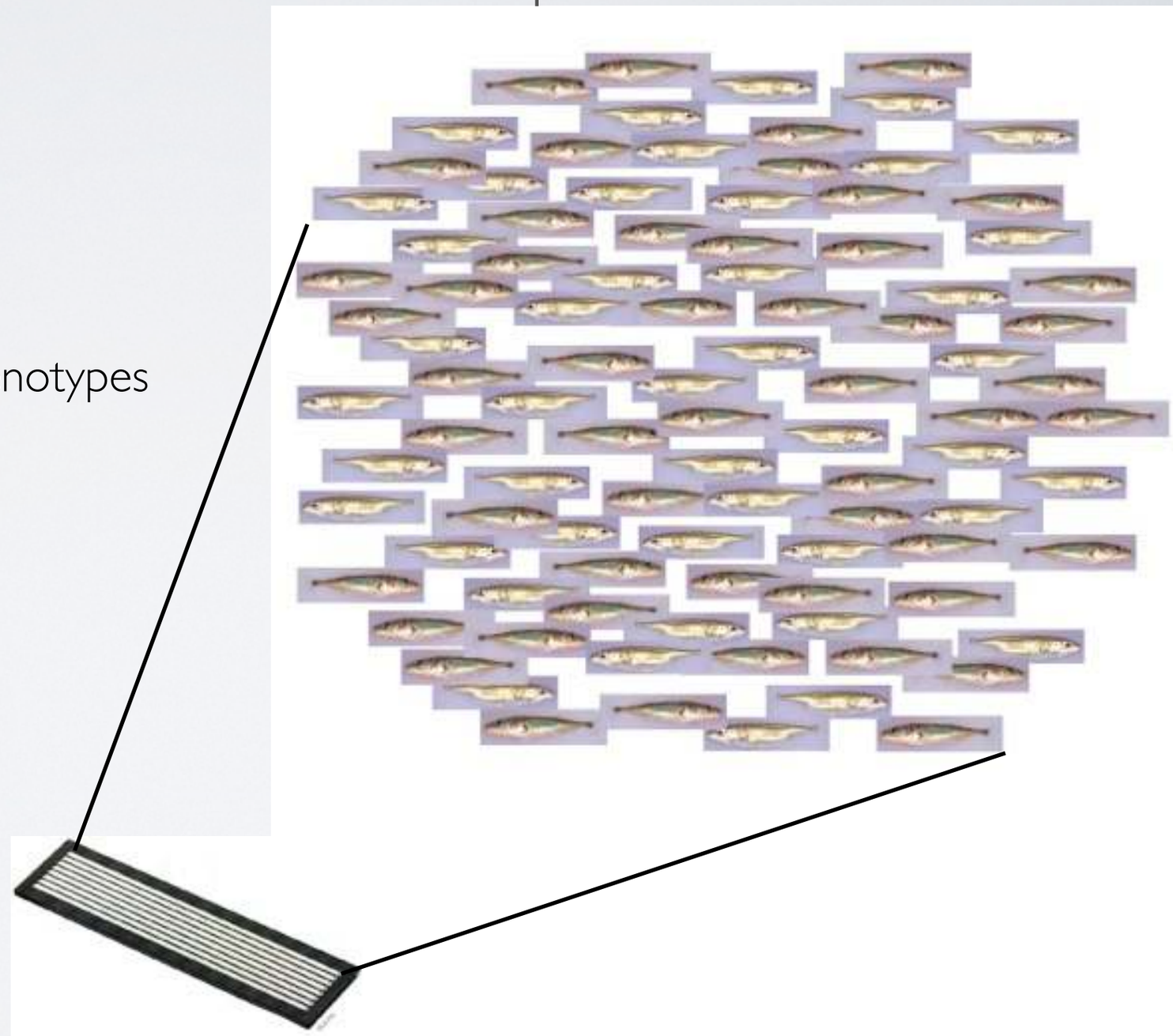
How many samples should be multiplexed?

Things to consider

Barcoded adapters

Depth

Mean depth of 20 to 40 to call genotypes



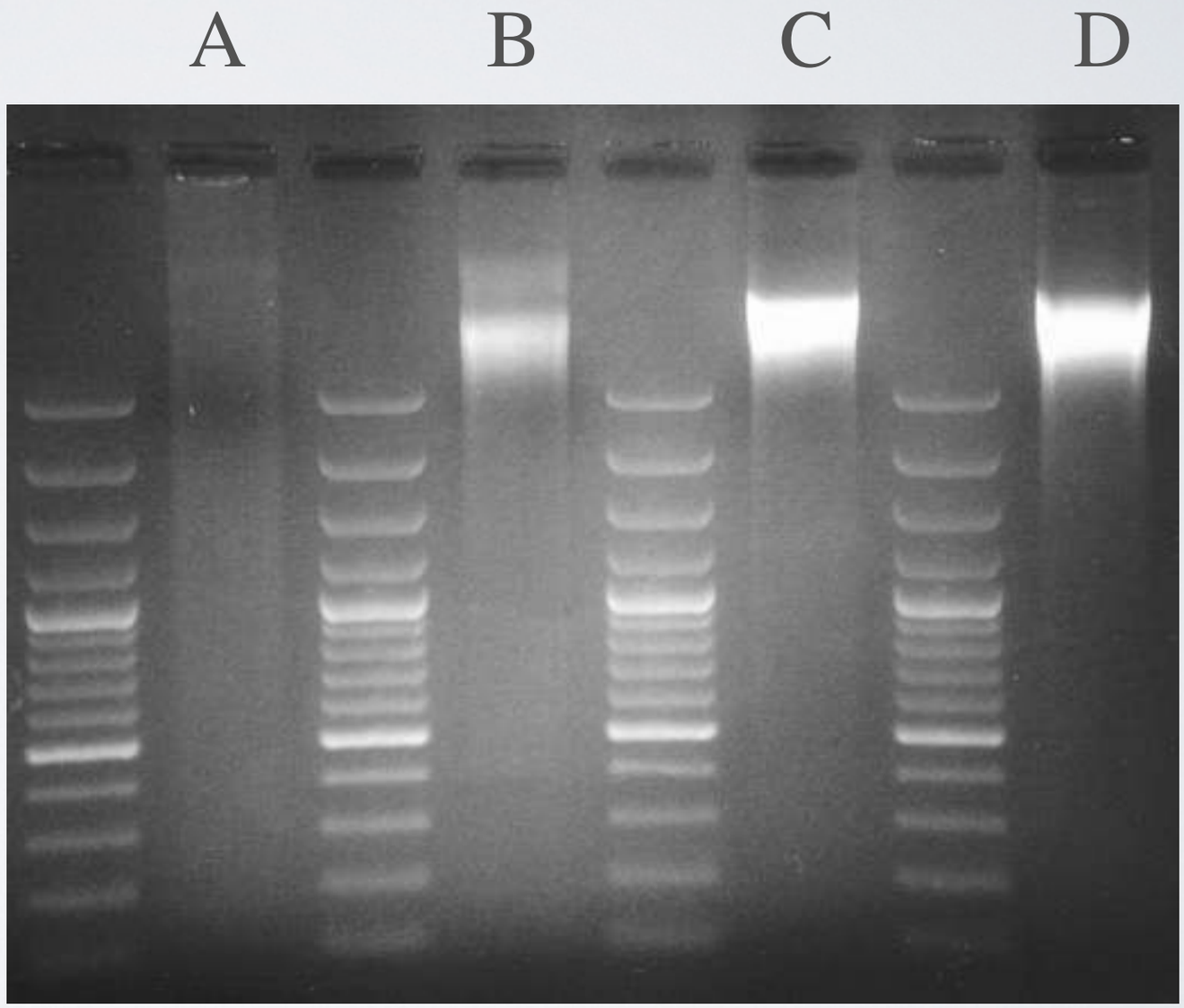
Molecular considerations in library building

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

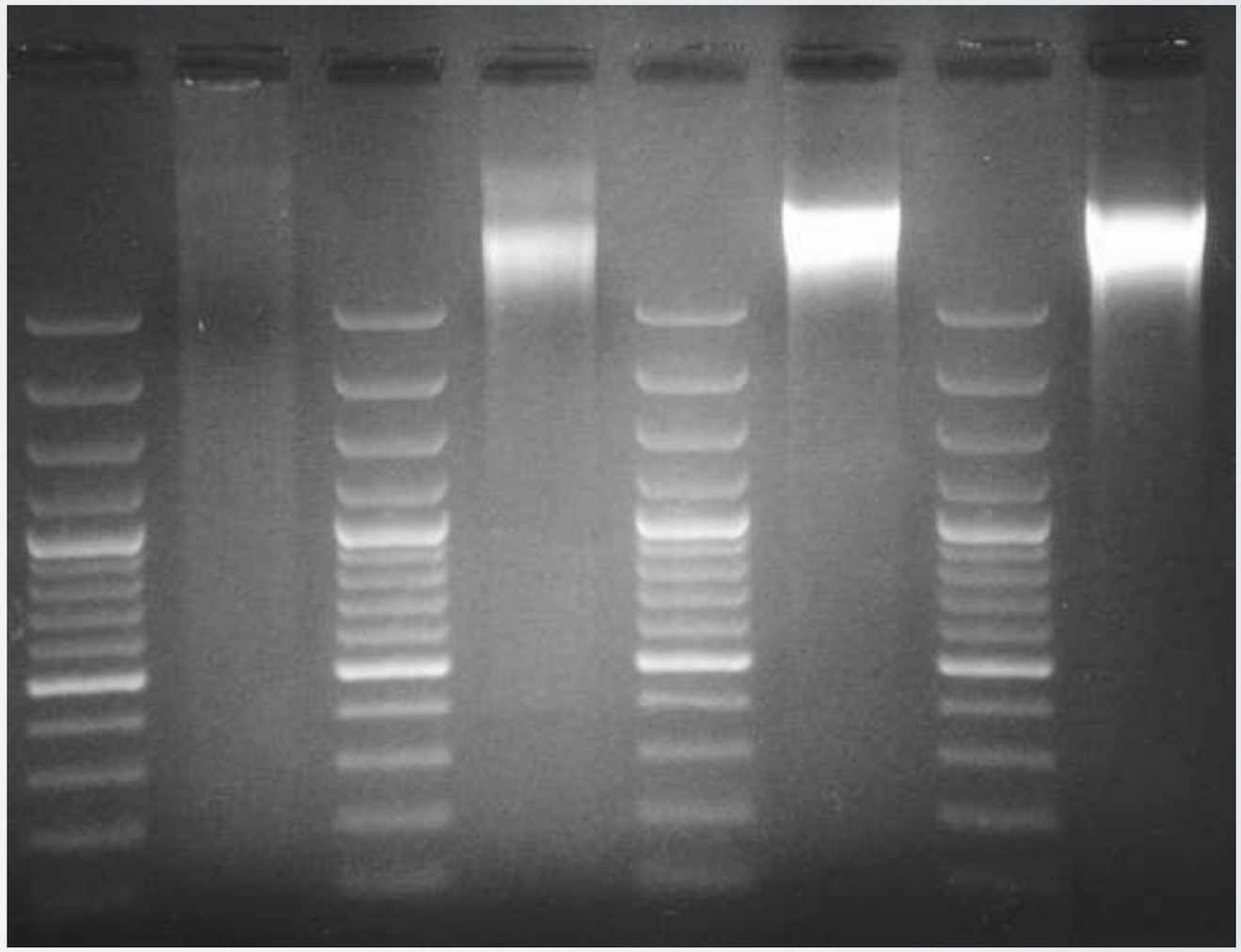
How many samples should be multiplexed?

Things to consider

DNA Quality

Diversify barcodes

- Illumina Genome Analyzer is confused by repetition in first 4 bases



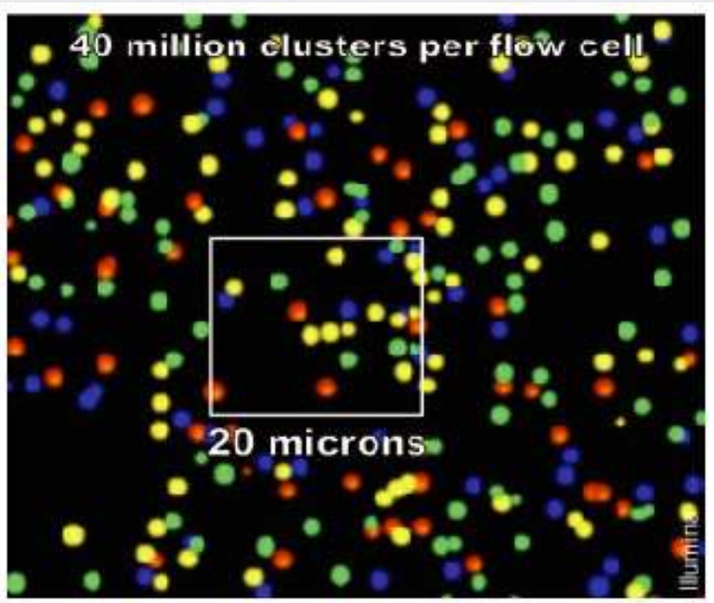
Molecular considerations in library building

How can I get the best depth of coverage?

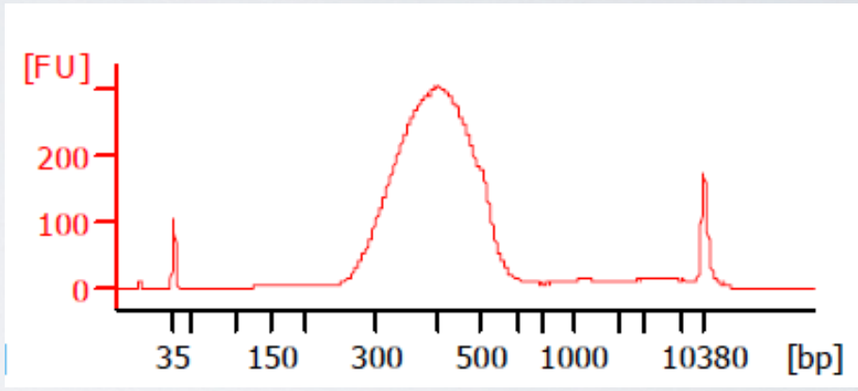
Things to consider

Fragment size

Smaller/tighter is better



Agilent Bioanalyzer



Molecular considerations in library building

How can I get the best depth of coverage?

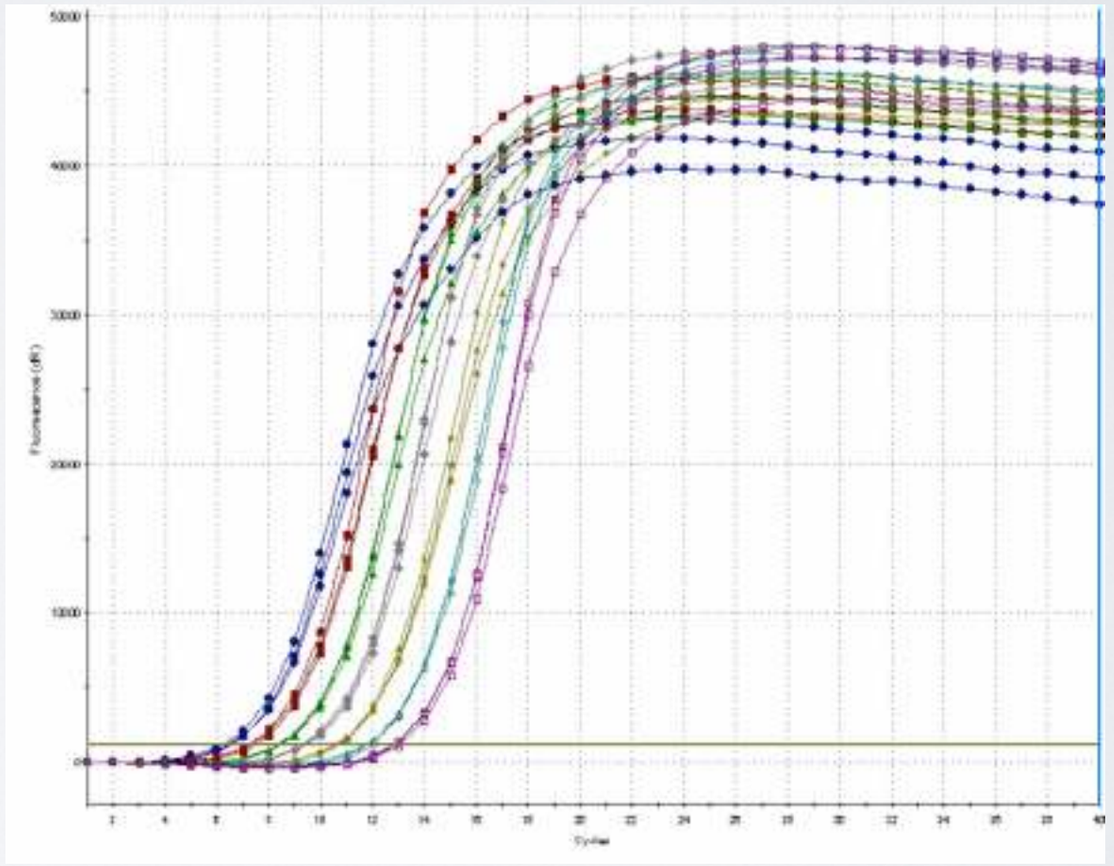
Things to consider

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

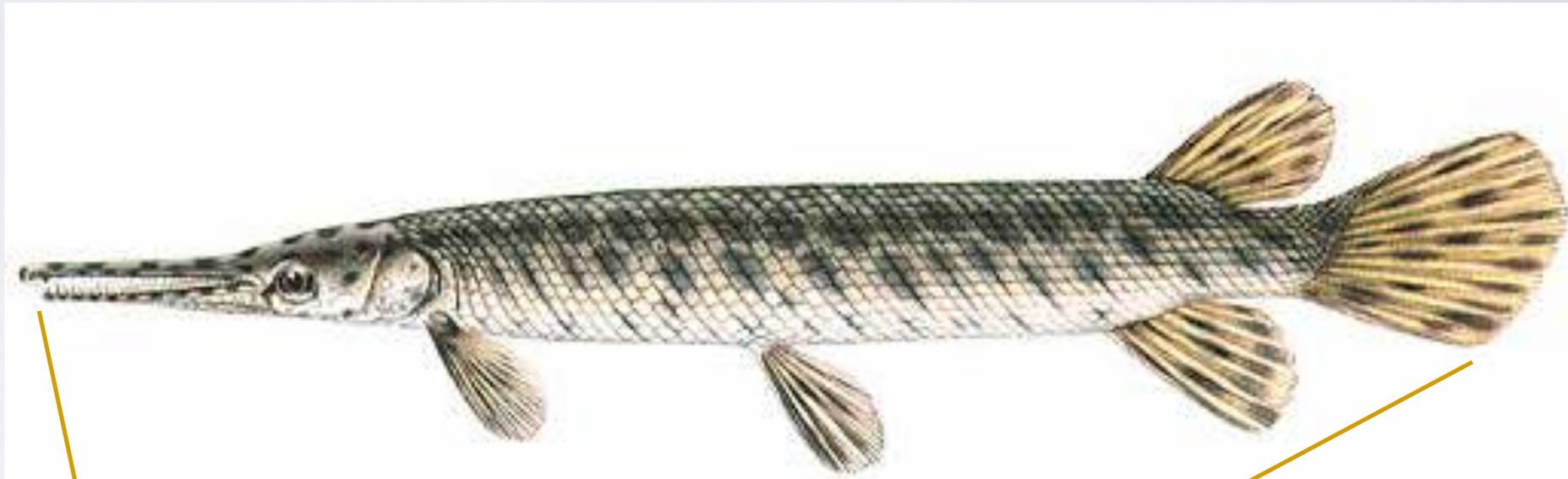
Fragment size

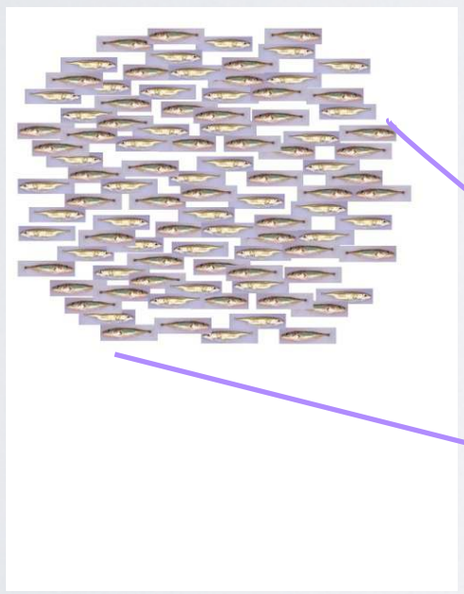
Library quality

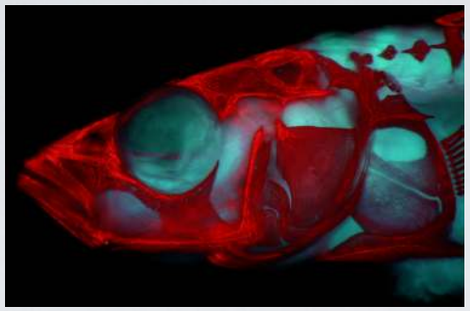
qPCR

Pilot Experiment:

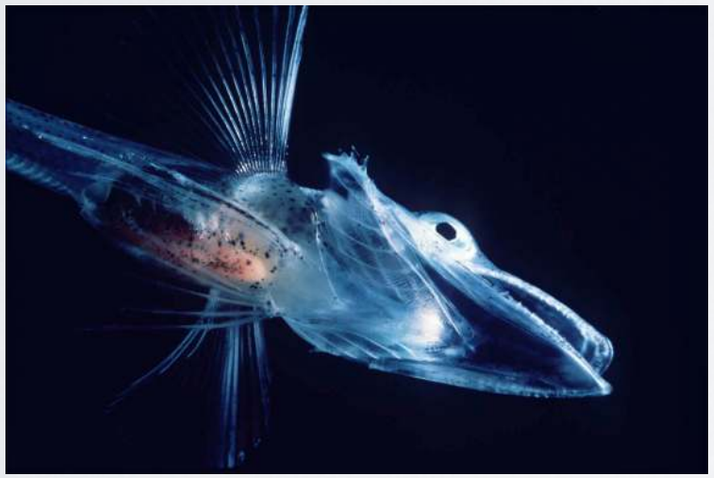
Spike or split a lane






What if you have a reference genome?

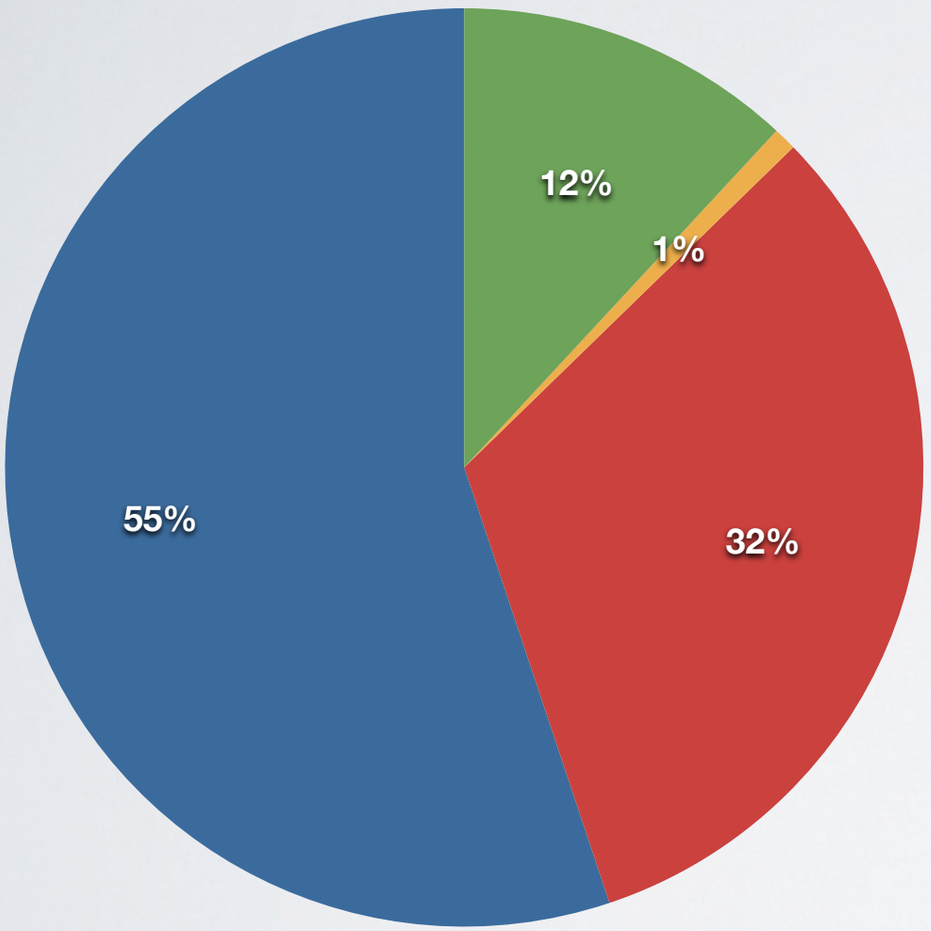




Anatomy of a PstI RAD-seq Analysis

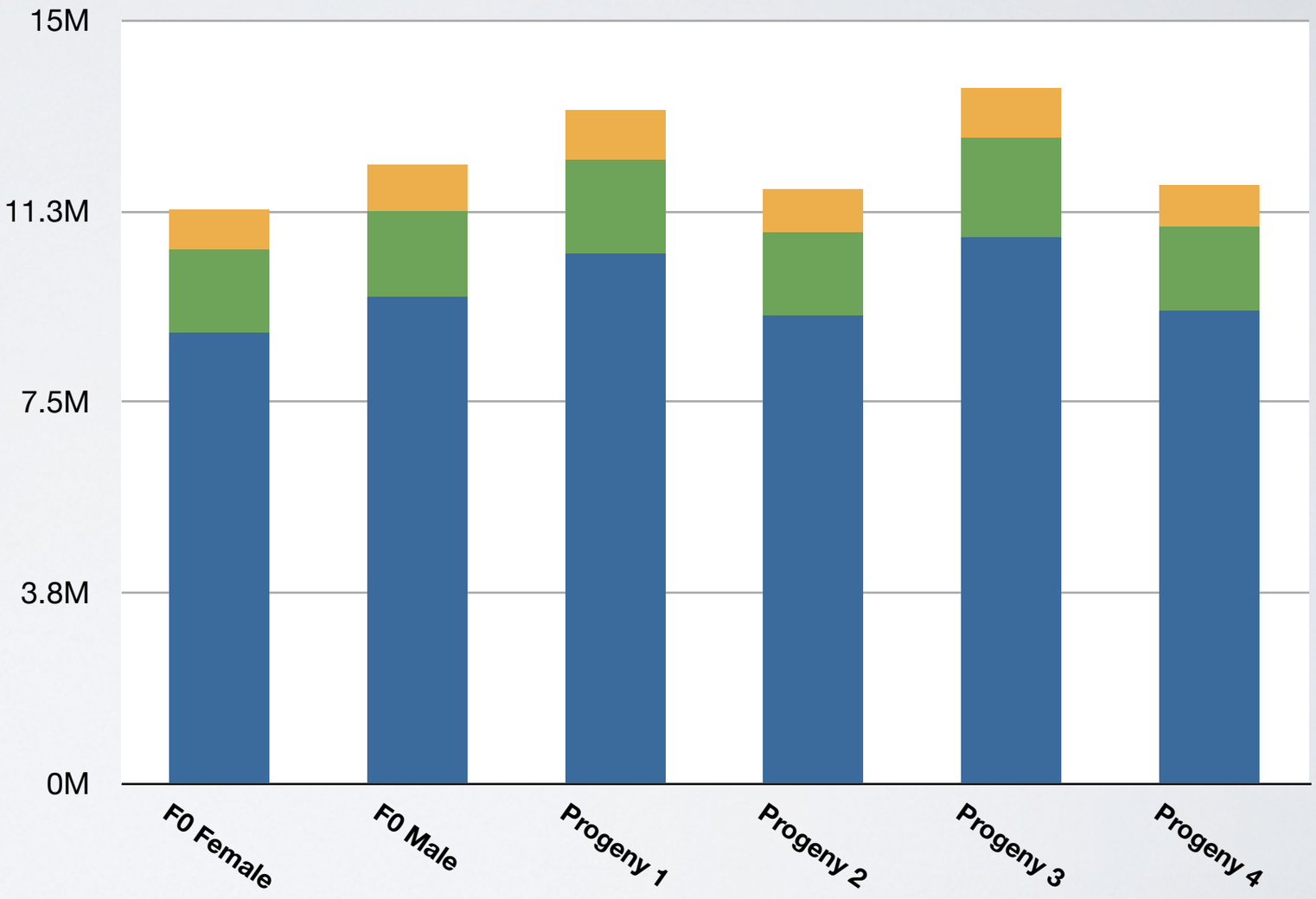
CTGCAG

133M Illumina Raw Reads



Aligned to Reference

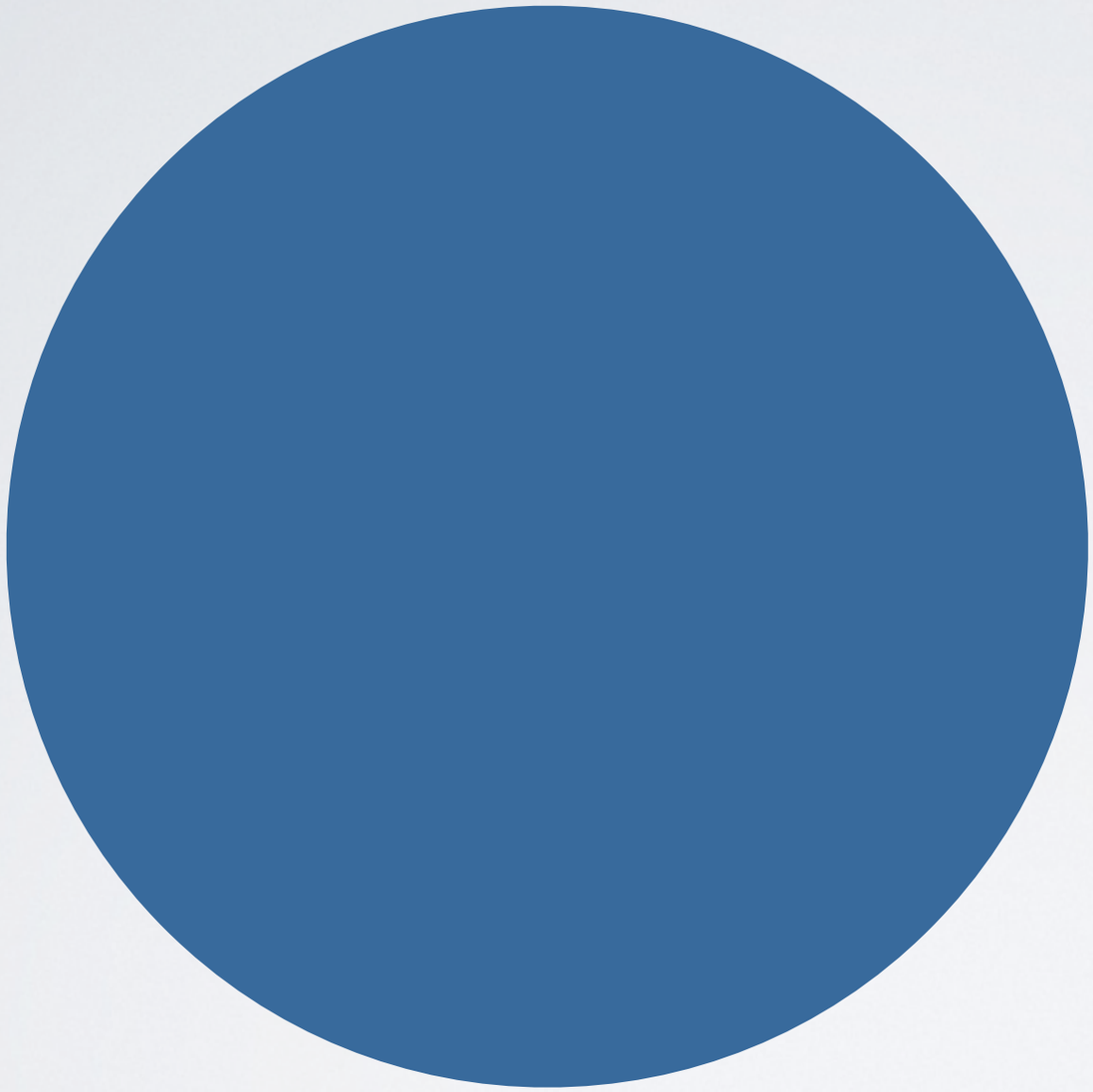
(Bowtie)



Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

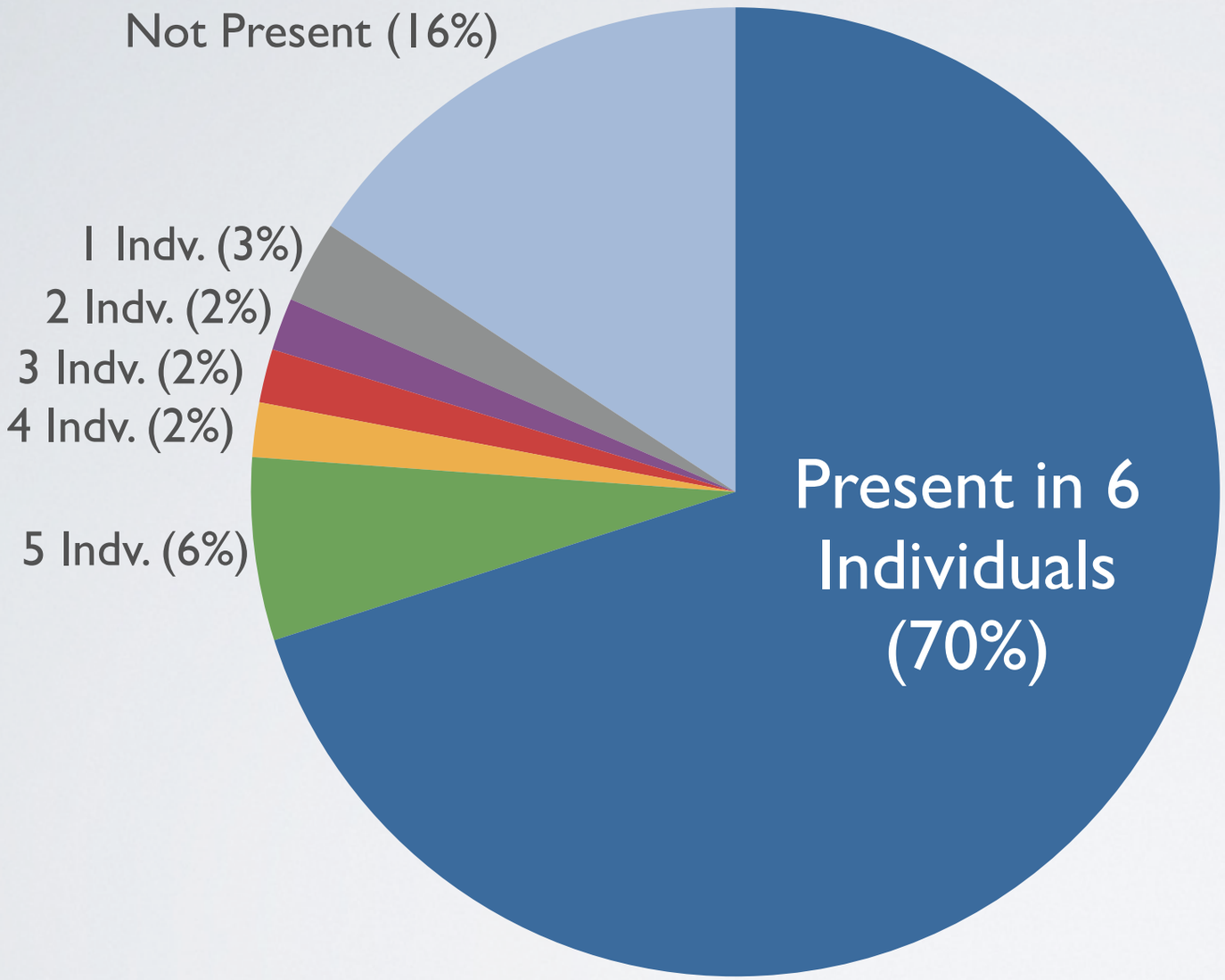
(224K Sites)



Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

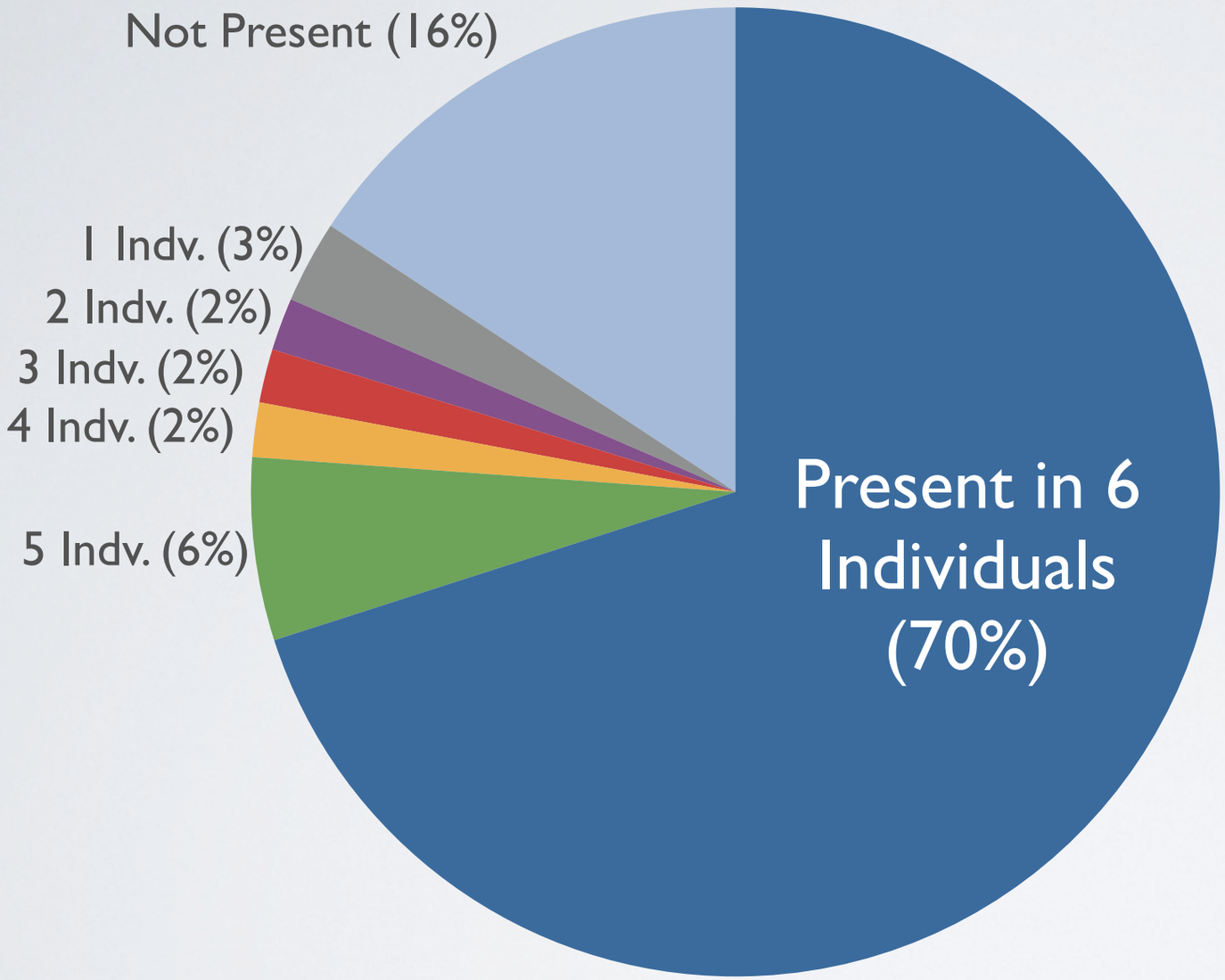
(224K Sites)



Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)

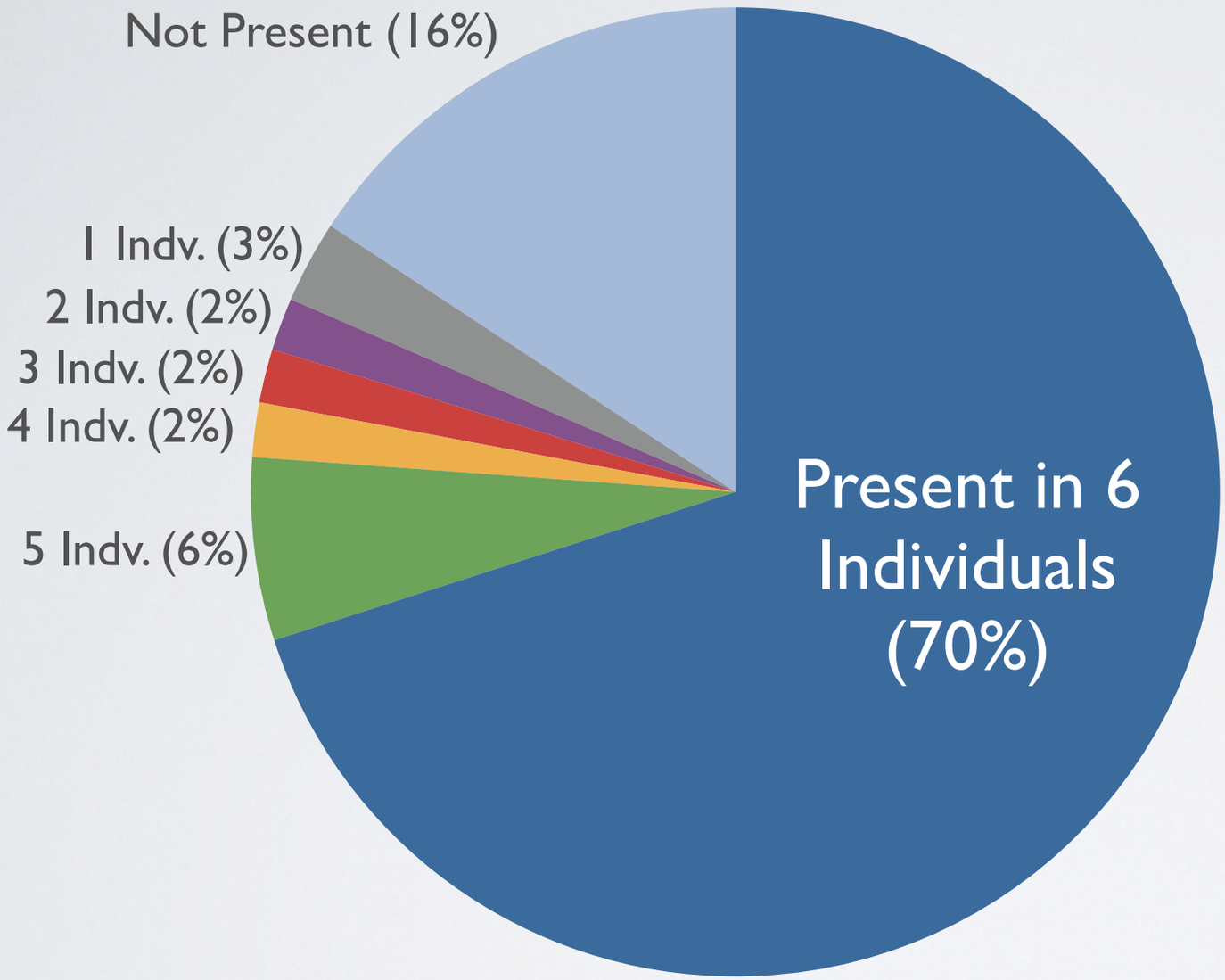


15K Non-Reference Sequences

Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)



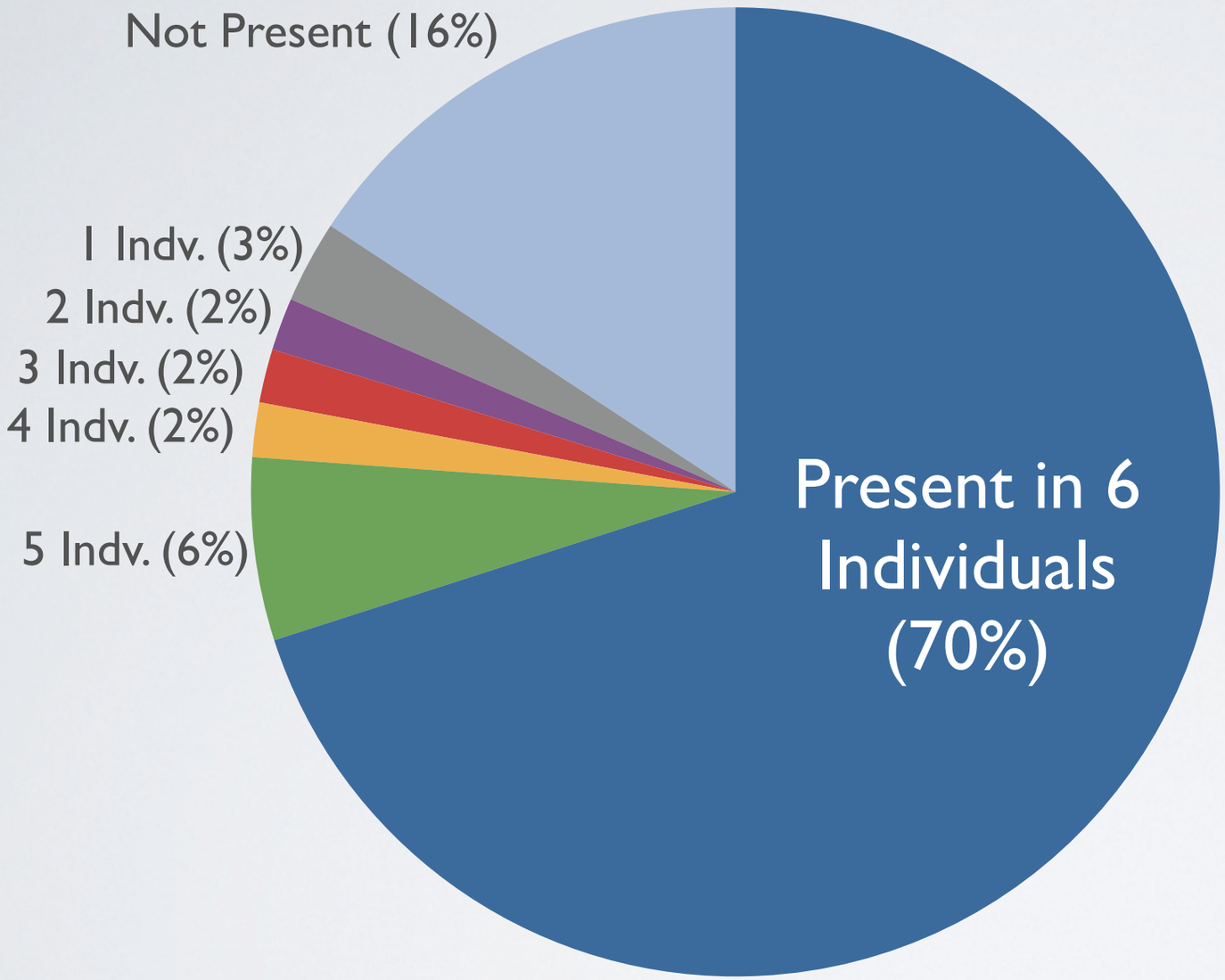
15K Non-Reference Sequences

97K *de novo* Stacks

Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)

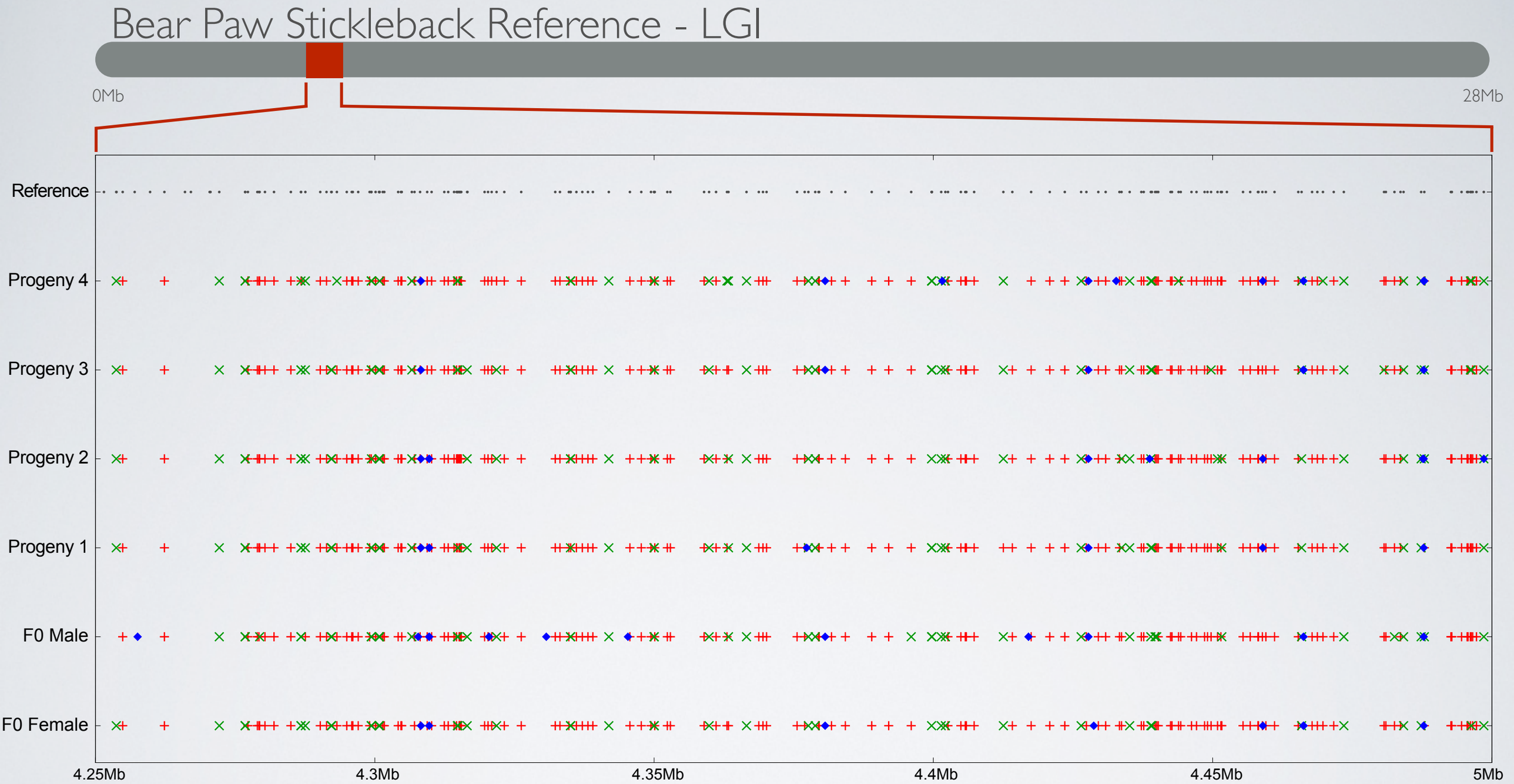


15K Non-Reference Sequences

97K *de novo* Stacks

Genome Size	Method	Cut Sites	Frequency
442Mb	Naive	84K	5.3K
	Empirical	224K	1.5K

Anatomy of a PstI RAD-seq Analysis

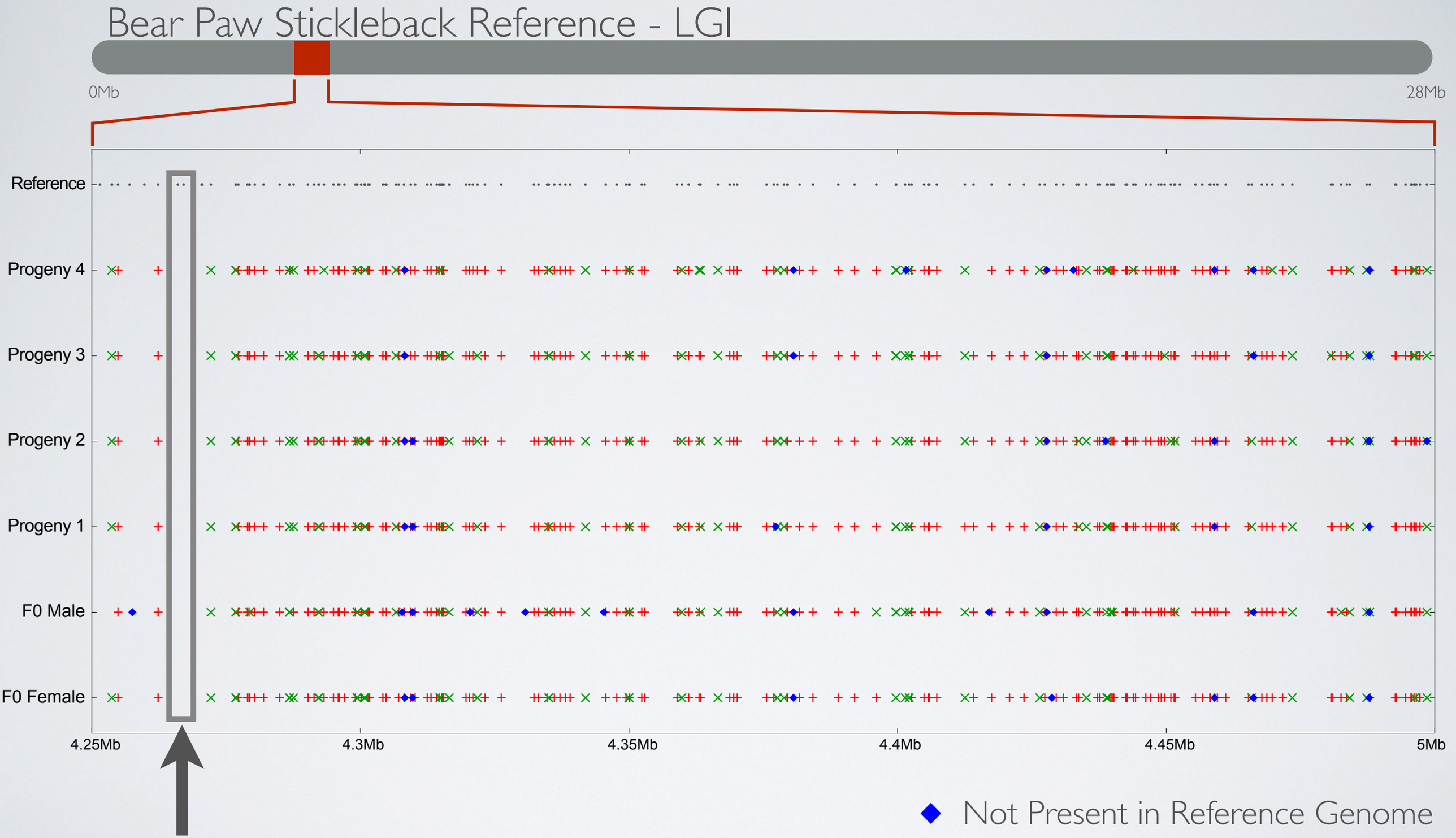


◆ Not Present in Reference Genome

× Detected on a Single Strand

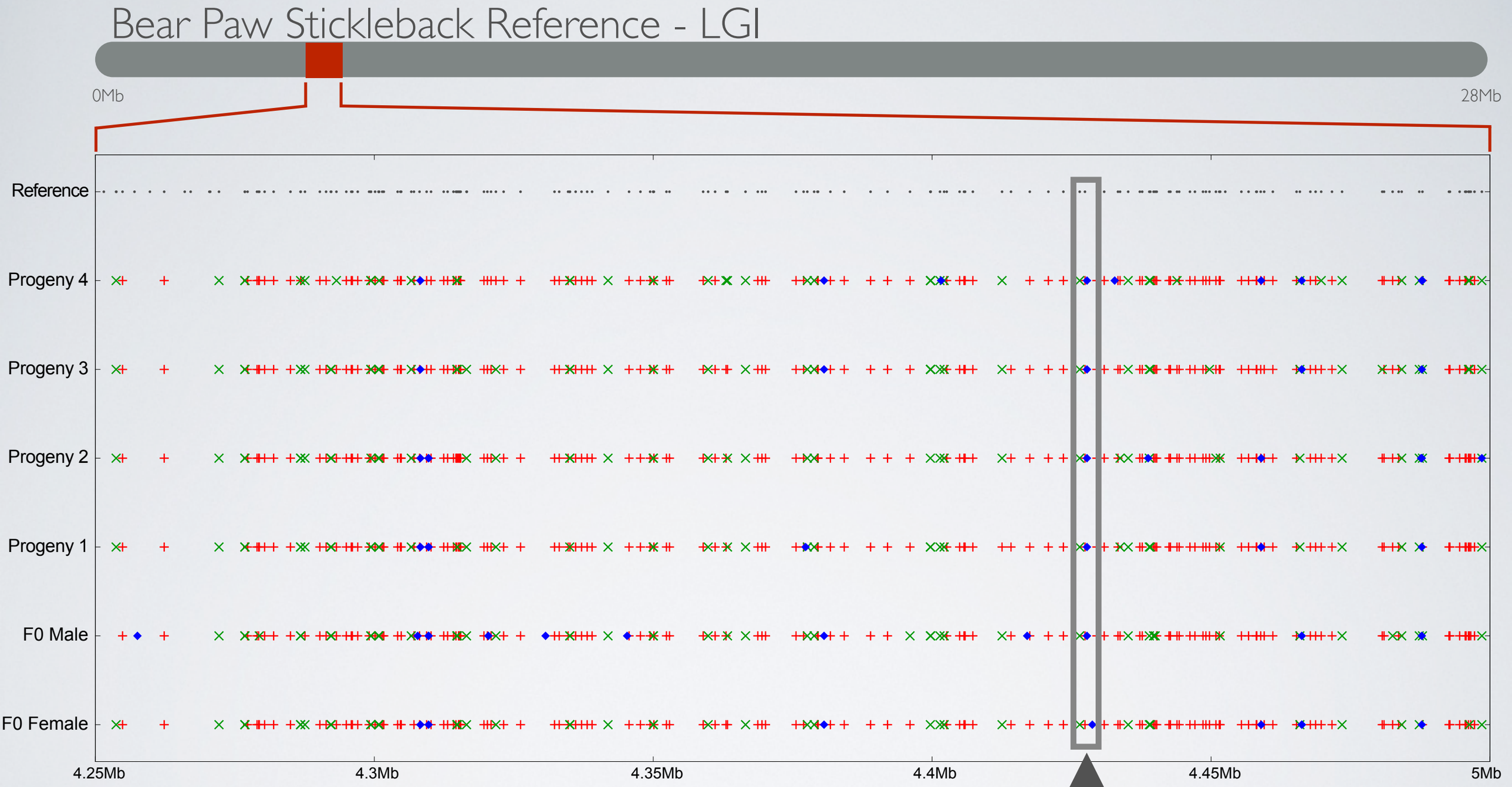
\+ Detected on Both Strands

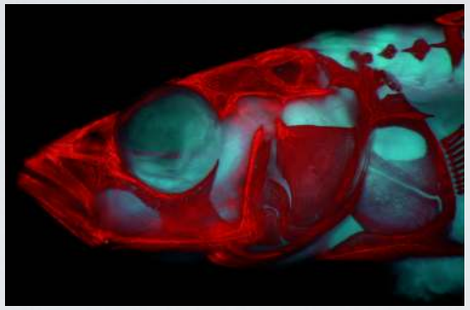
Anatomy of a PstI RAD-seq Analysis



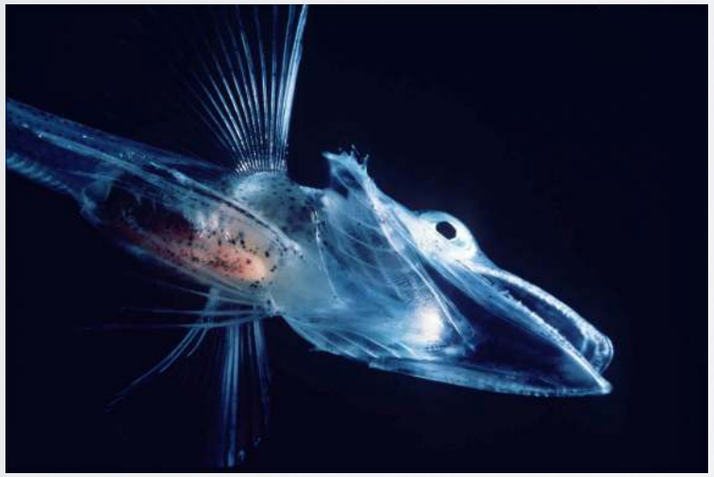
Anatomy of a PstI RAD-seq Analysis

Anatomy of a PstI RAD-seq Analysis





Signatures of natural selection across the genome



Threespine Stickleback, *Gasterosteus aculeatus*

- ***Ancestral Oceanic Populations***
  Marine and Anadromous
  Old (> 10 million years)

- ***Derived Freshwater Populations***
  Lake and stream
  Young (<15,000 years)

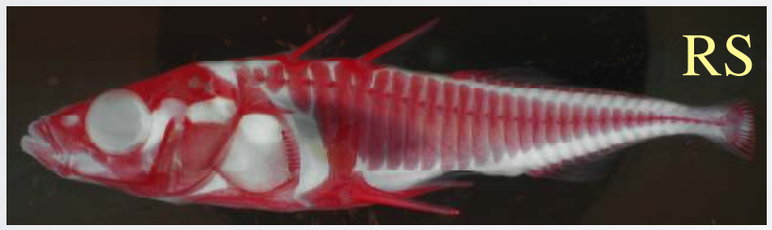
Ocean form

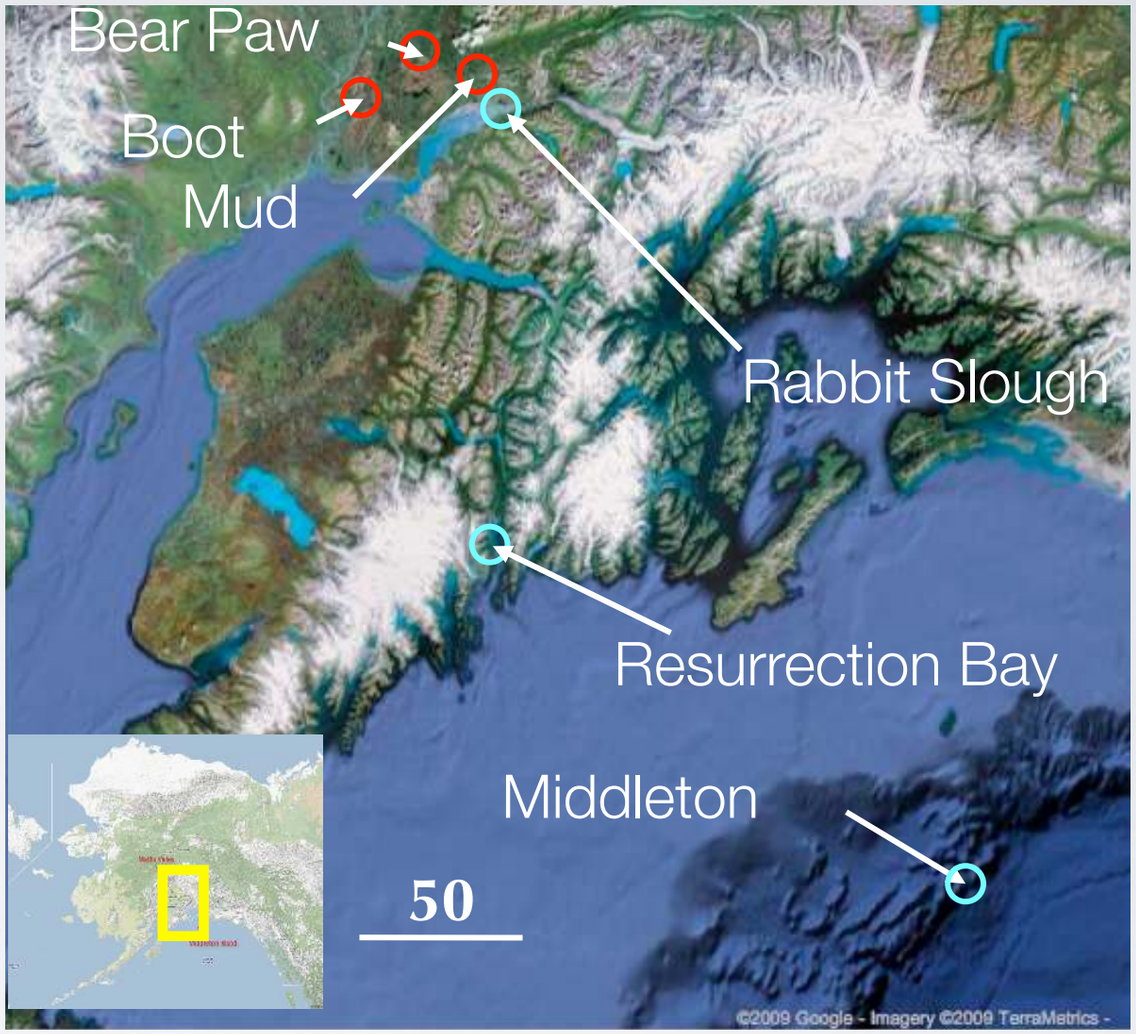
Freshwater form

Susan Bassham

Paul Hohenlohe

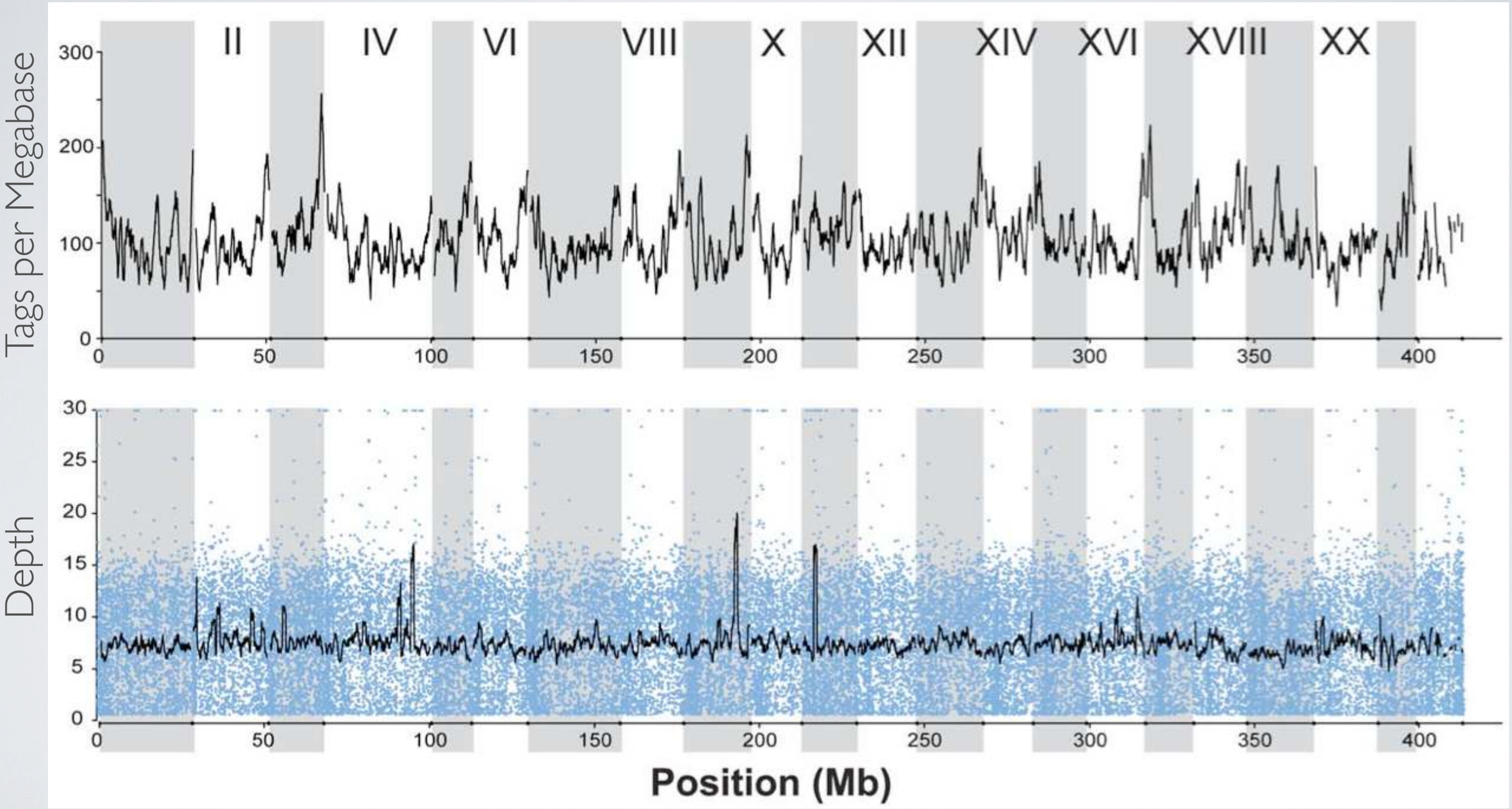
Signatures of natural selection across the genome





- 20 individuals in each of 5 populations
- 2 Ocean & 3 Freshwater
- 45,000 SNPs in each individual

Even coverage of RAD tags across the genome



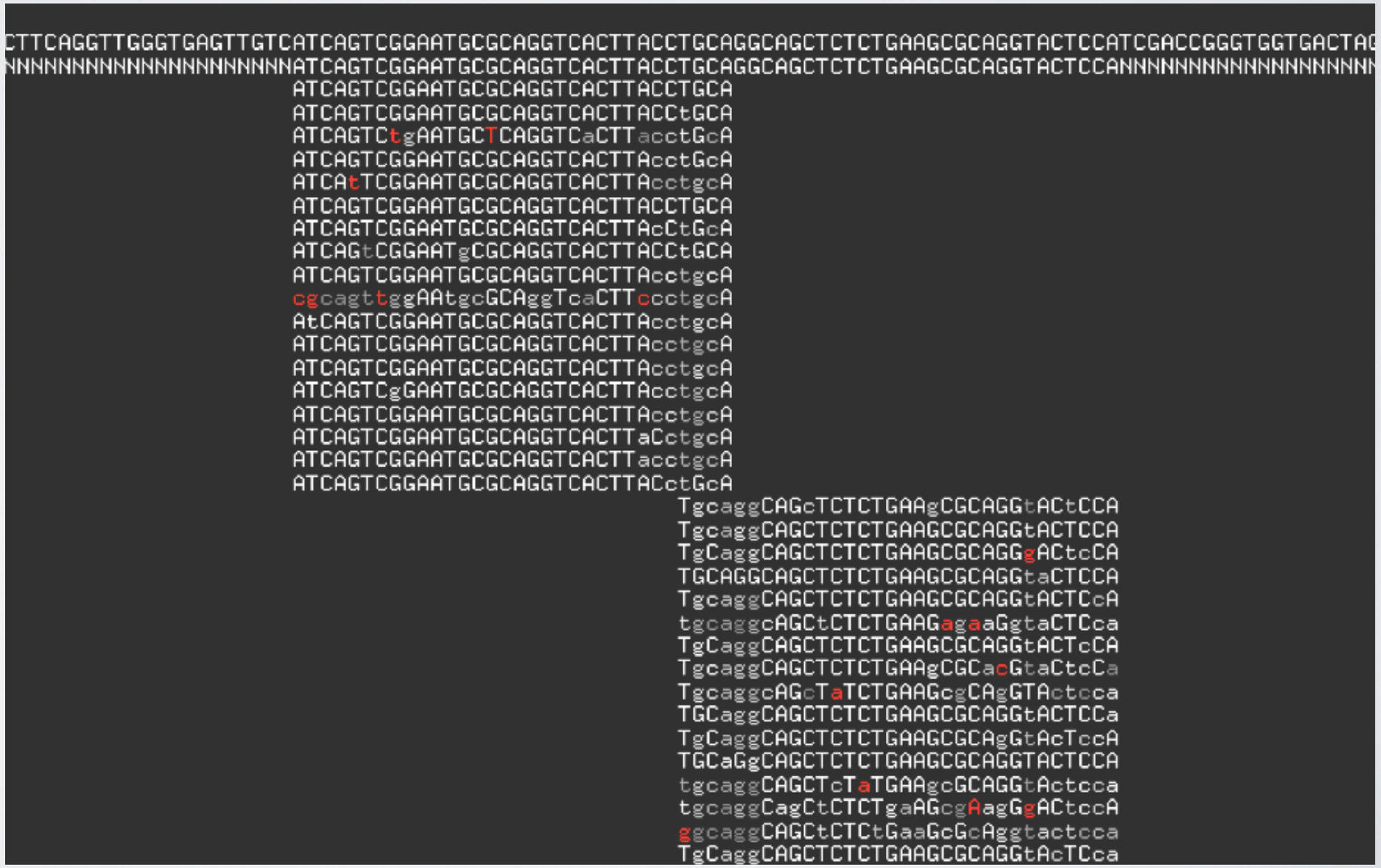
Concatenated stickleback genome

Differentiating SNPs from error

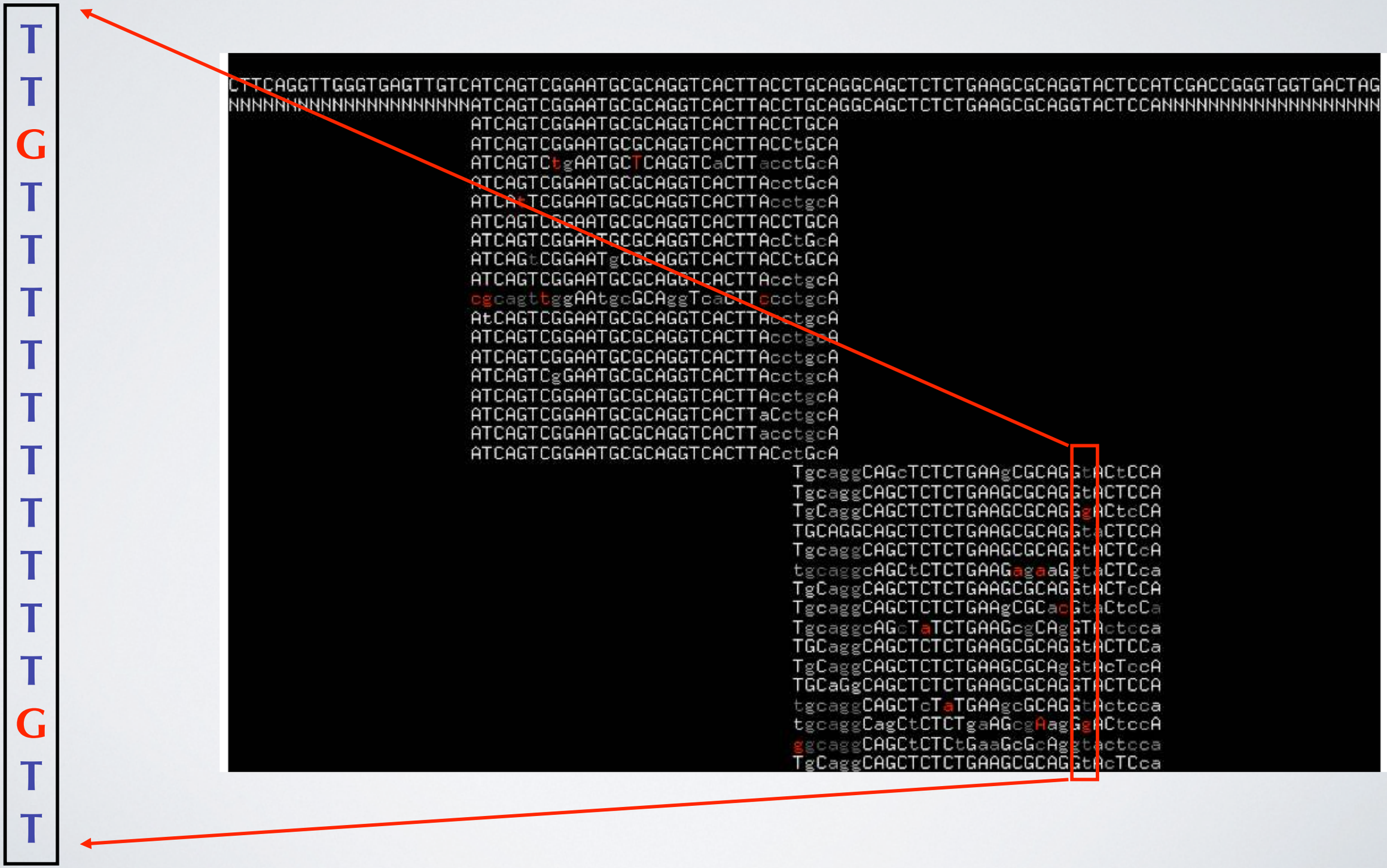
Restriction enzyme recognition site

Reference genome sequence

sequence reads



Differentiating SNPs from error



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

T
T
G
T
T
T
T
T
T
T
T
T
T
T
T
T
G
T
T

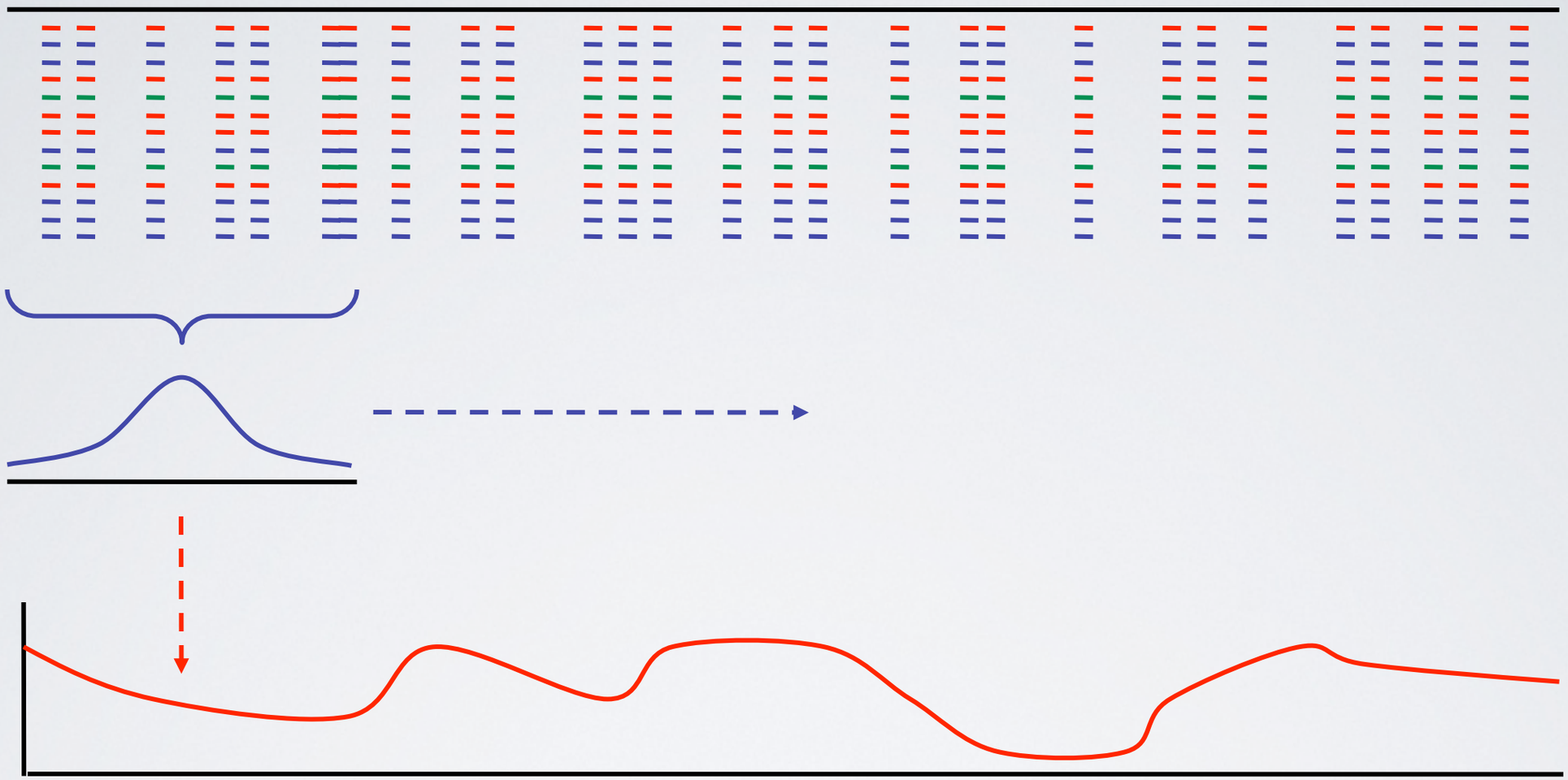
$$L(n_1\,\text{hom}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

$$L(n_1n_2\text{het}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(0.5-\frac{\varepsilon}{4}\right)^{n_1}\left(0.5-\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

Maximum likelihood genotyping based on multinomial distribution of nucleotide reads

Making statistics continuous across the genome

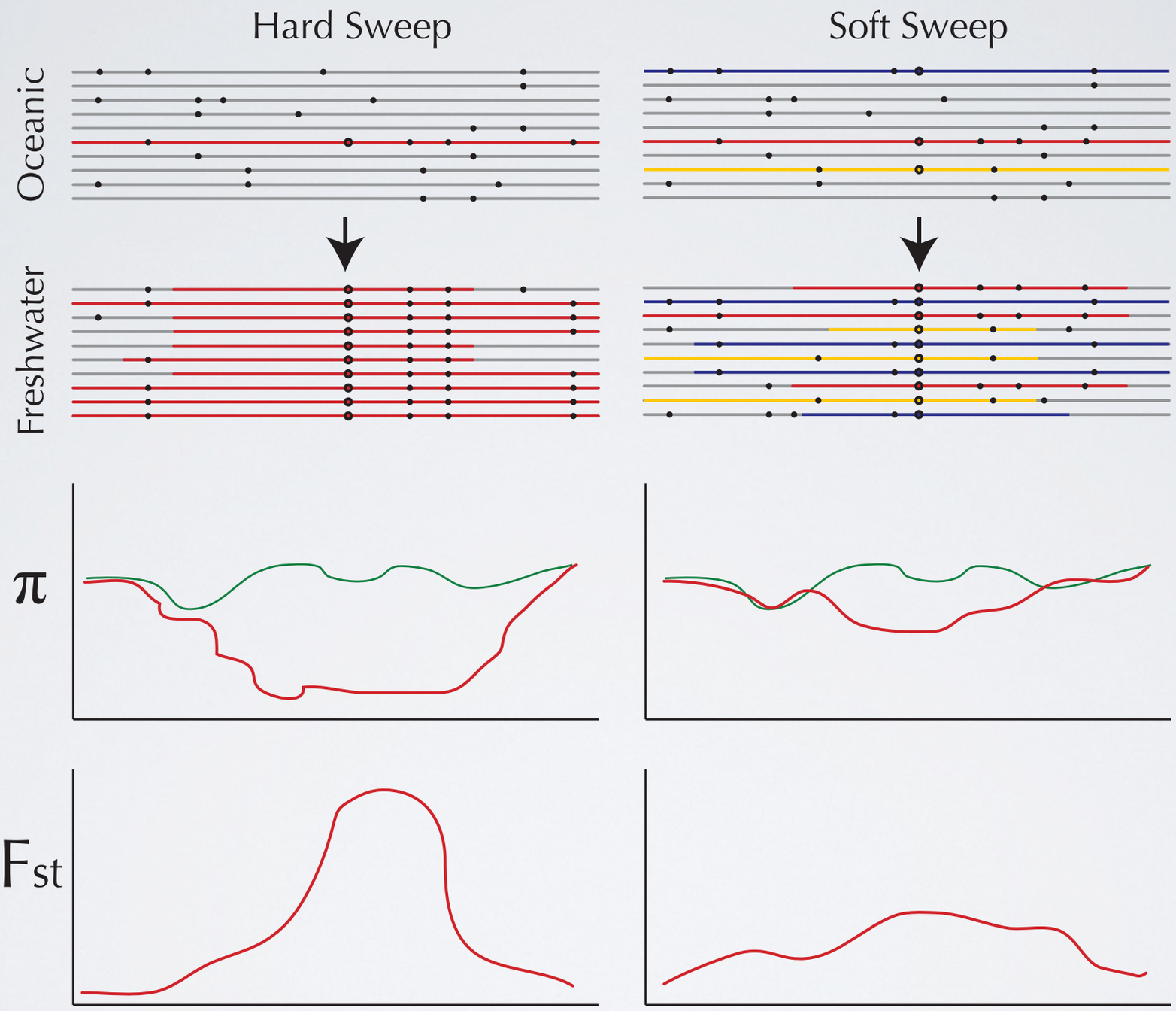
Kernel-smoothing average of summary statistics along genome



Bootstrap re-sampling to estimate significance of moving average

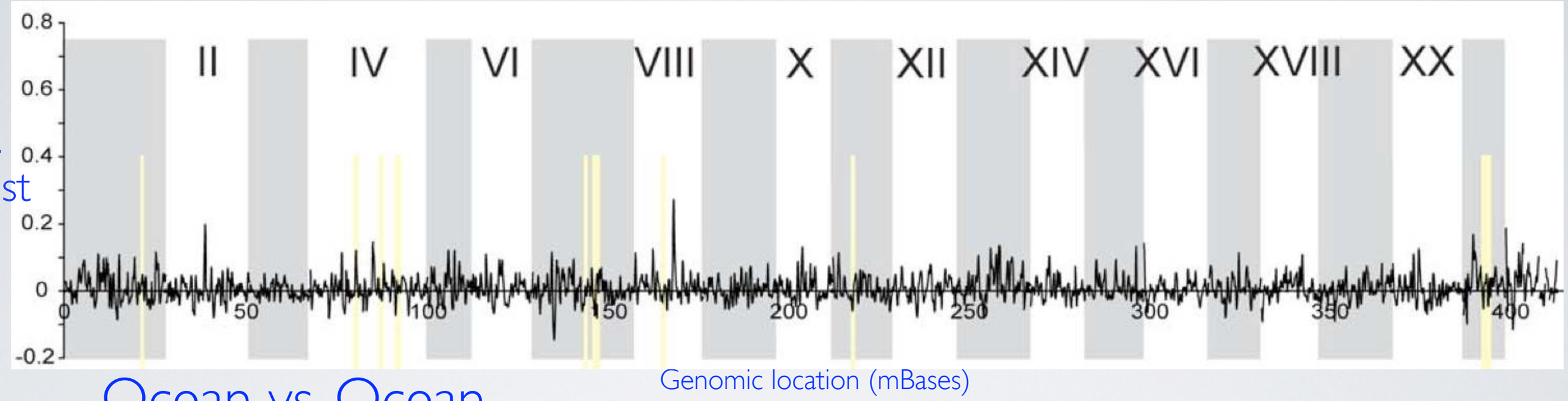
- Re-sample from across genome within the same population
- Significance is relative to genome-wide average, not based on a null model

Signatures of natural selection across the genome



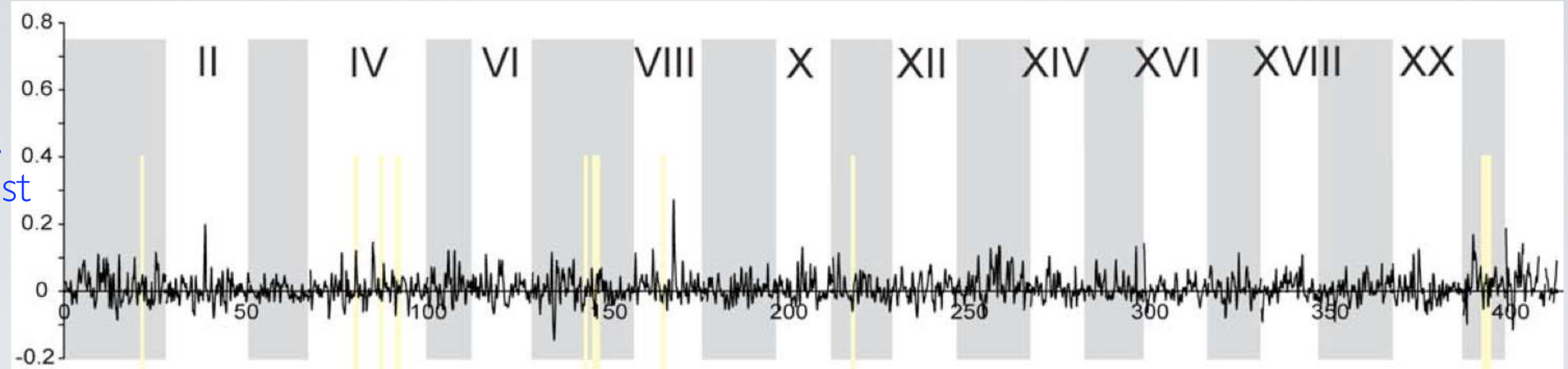
requires a very dense set of markers or complete sequences
assays need to be performed in dozens or hundreds of individuals
hard to develop for non-model organisms

Signatures of natural selection across the genome

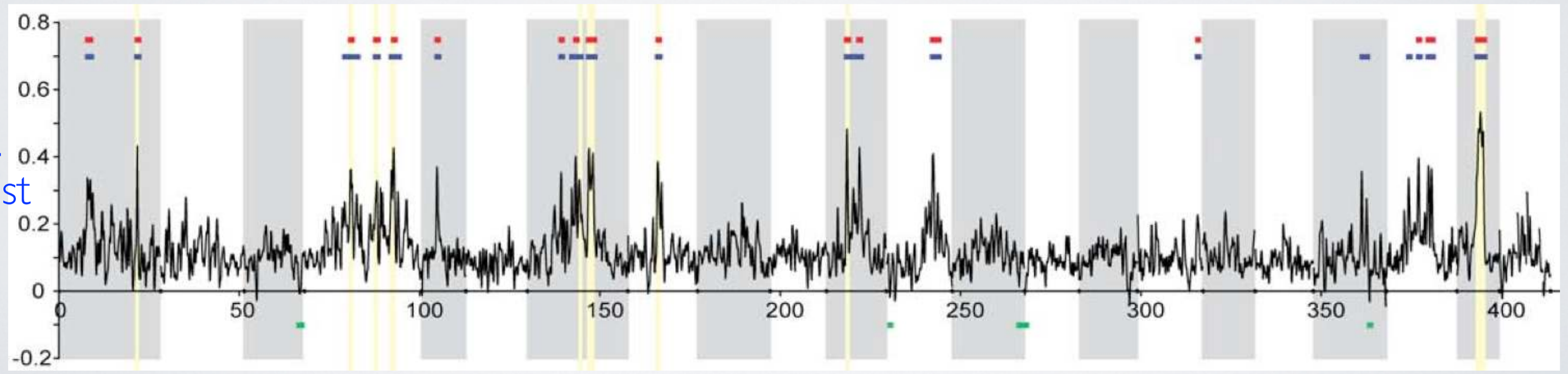


Ocean vs. Ocean

Signatures of natural selection across the genome

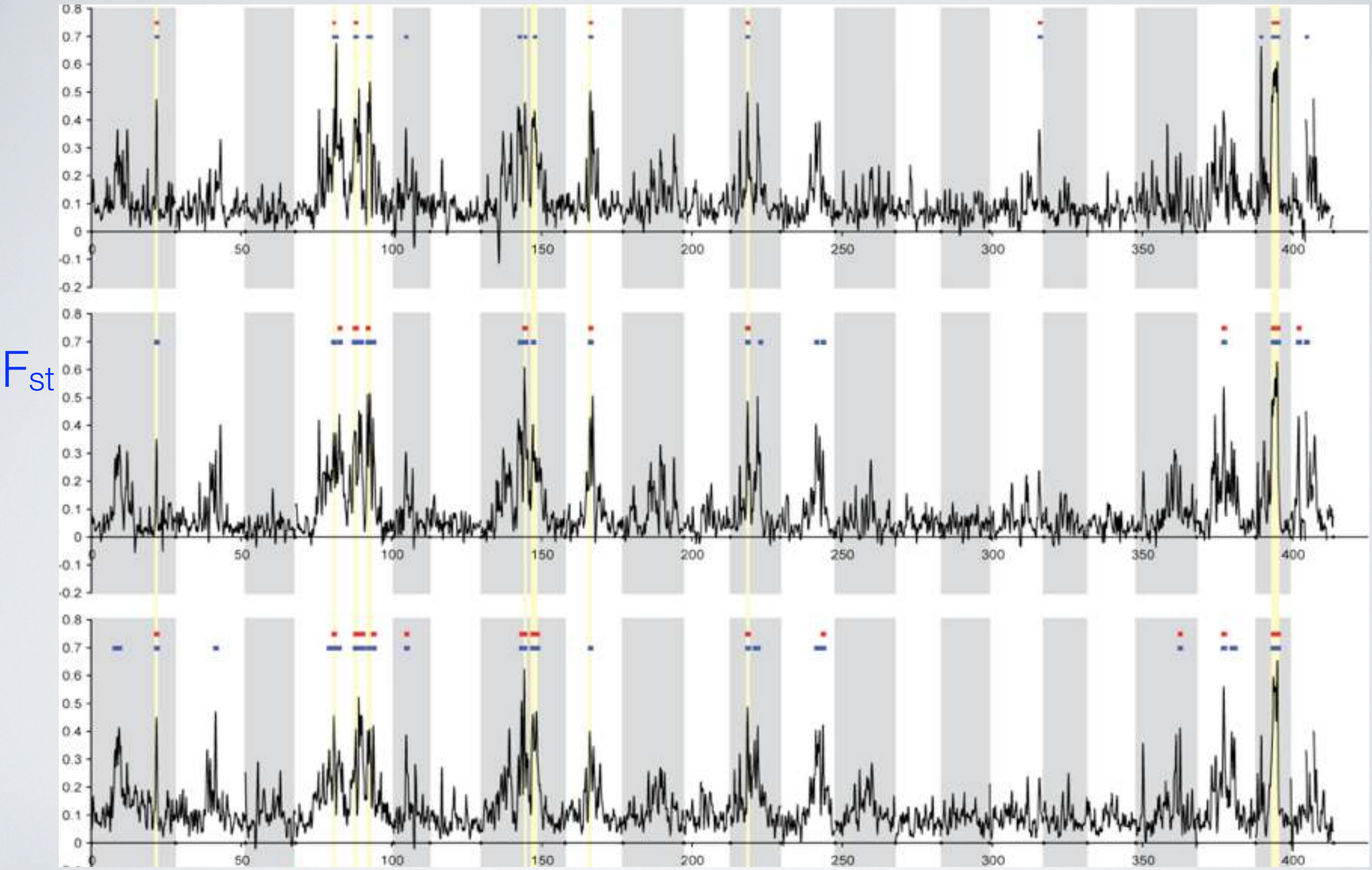


Ocean vs. Ocean

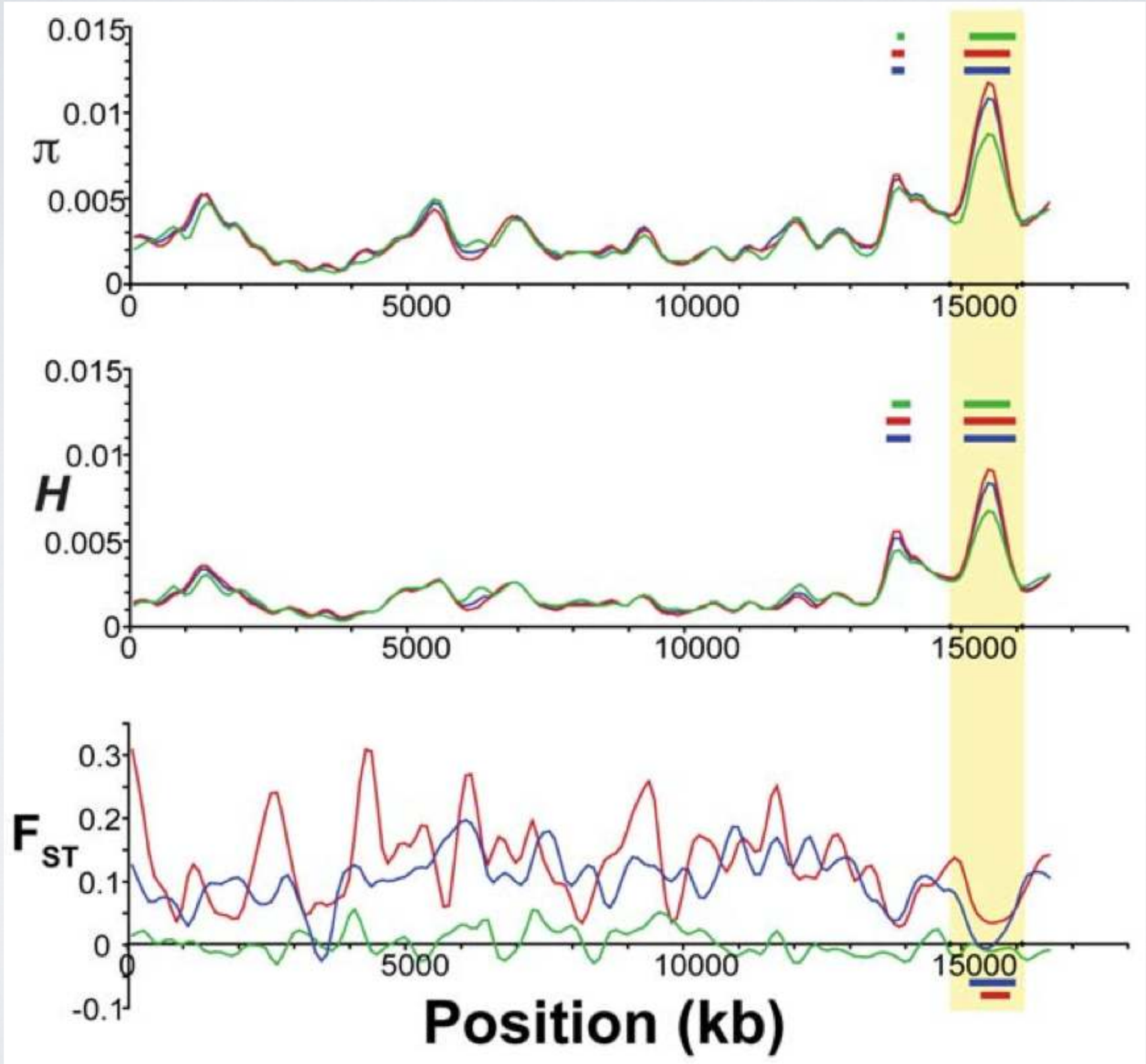


Freshwater Populations (grouped) vs. Ocean

Consistent pattern of parallel selection

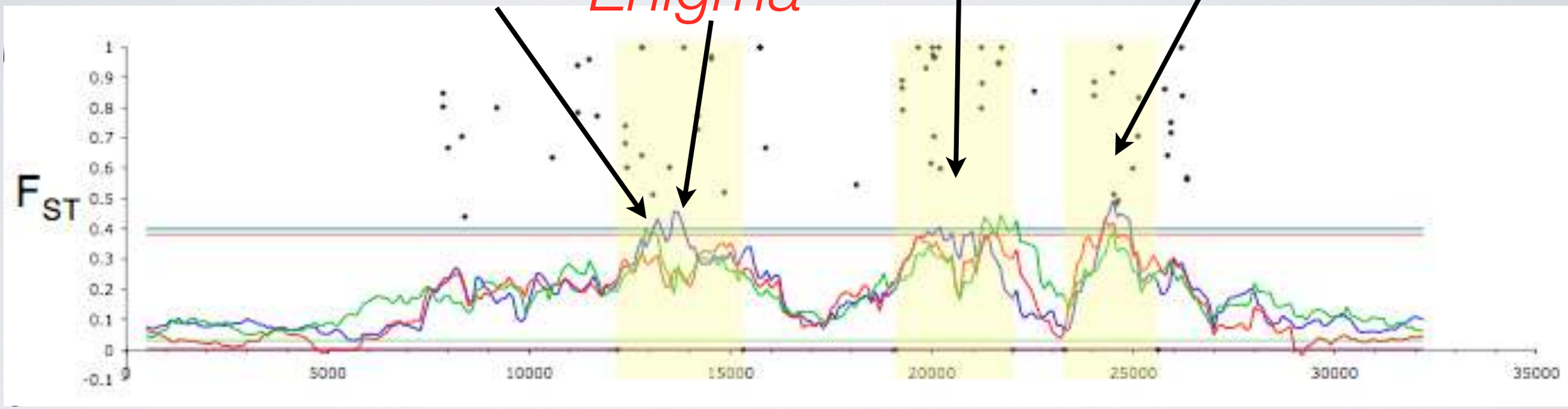


Evidence for balancing selection on Linkage Group III

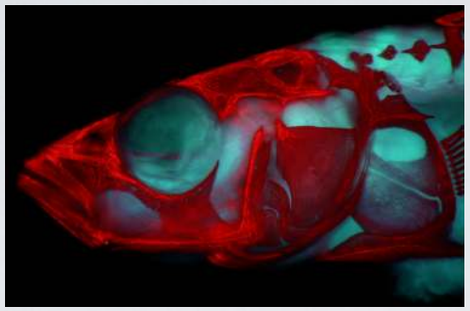


Previously identify quantitative trait loci (QTLs) are under selection

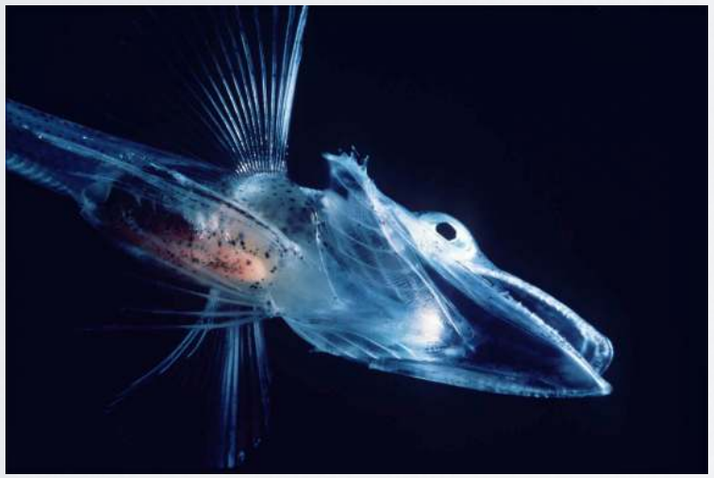
Natural populations



Lateral plate major locus on LGIV (4000 SNPs)

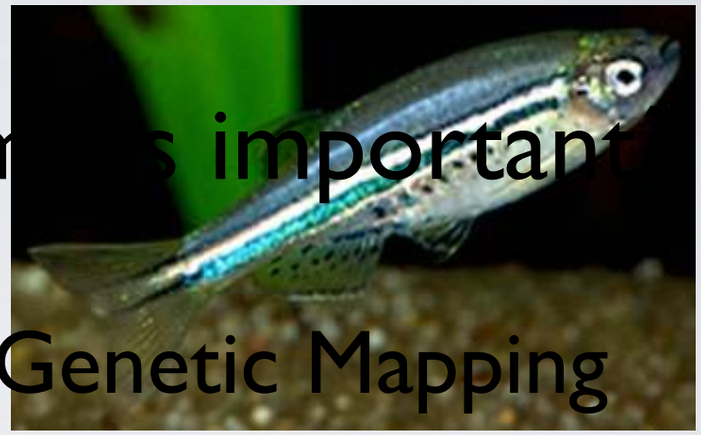




What if you **don't** have a reference genome?

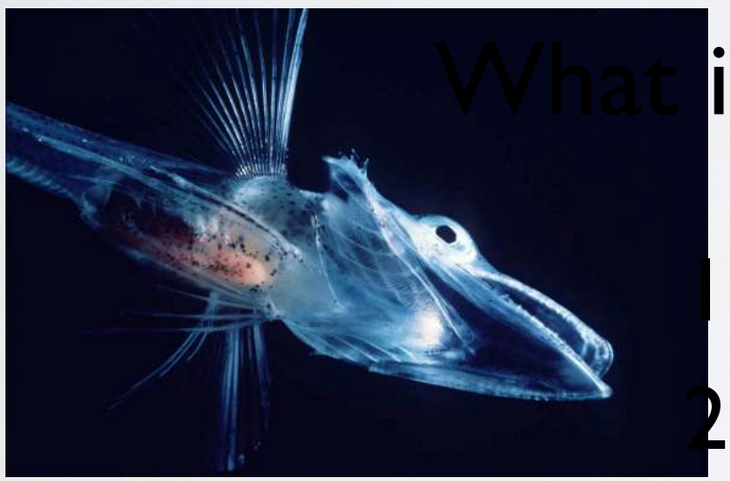




Why are genetic maps important?



- Forward/Reverse Genetic Mapping
- QTL Mapping
- Physical Genome Assembly



What is required to build a map?

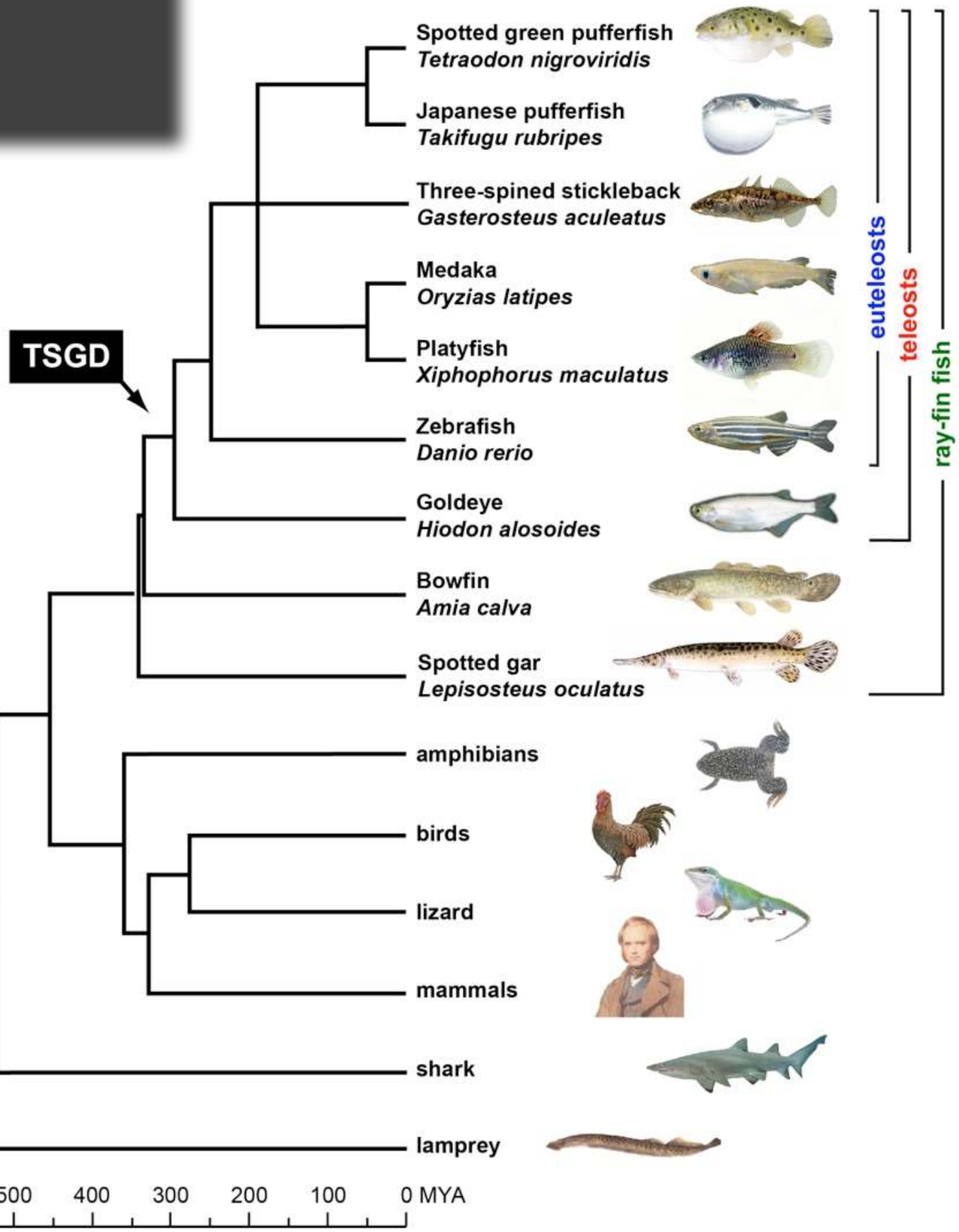
1. Generate progeny
2. Recombination
3. Detect polymorphism

Genetics

Spotted Gar

Angel Amores

John Postlethwait

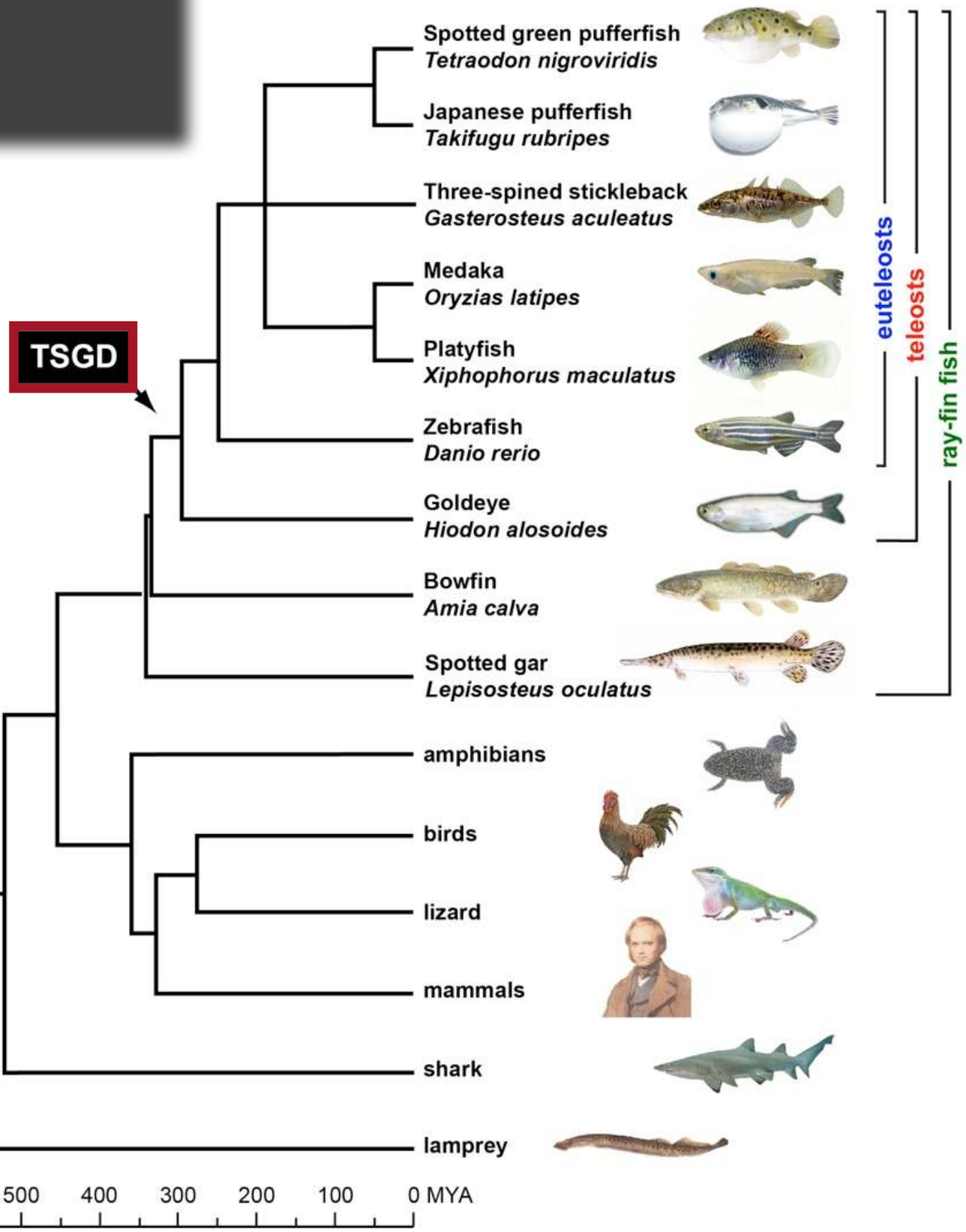


Genetics

Spotted Gar

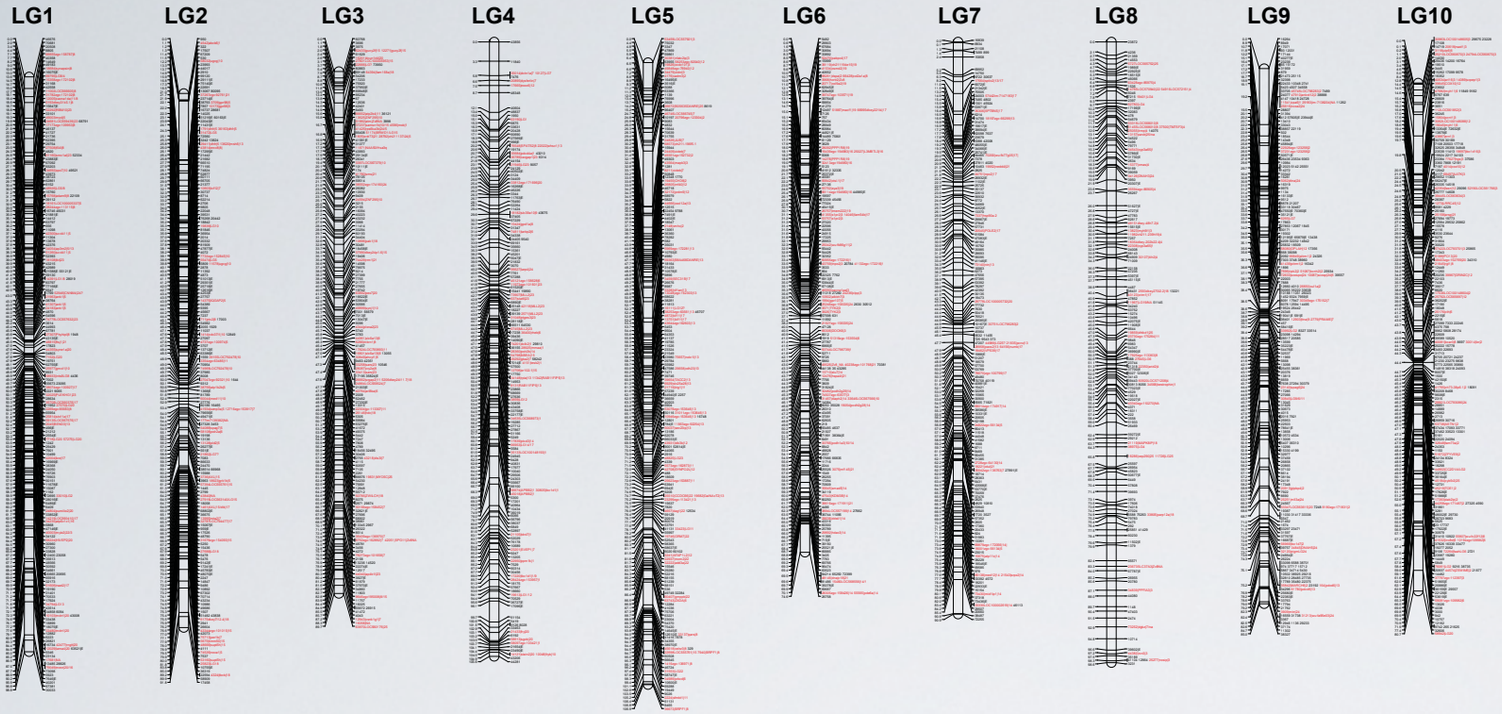
Angel Amores

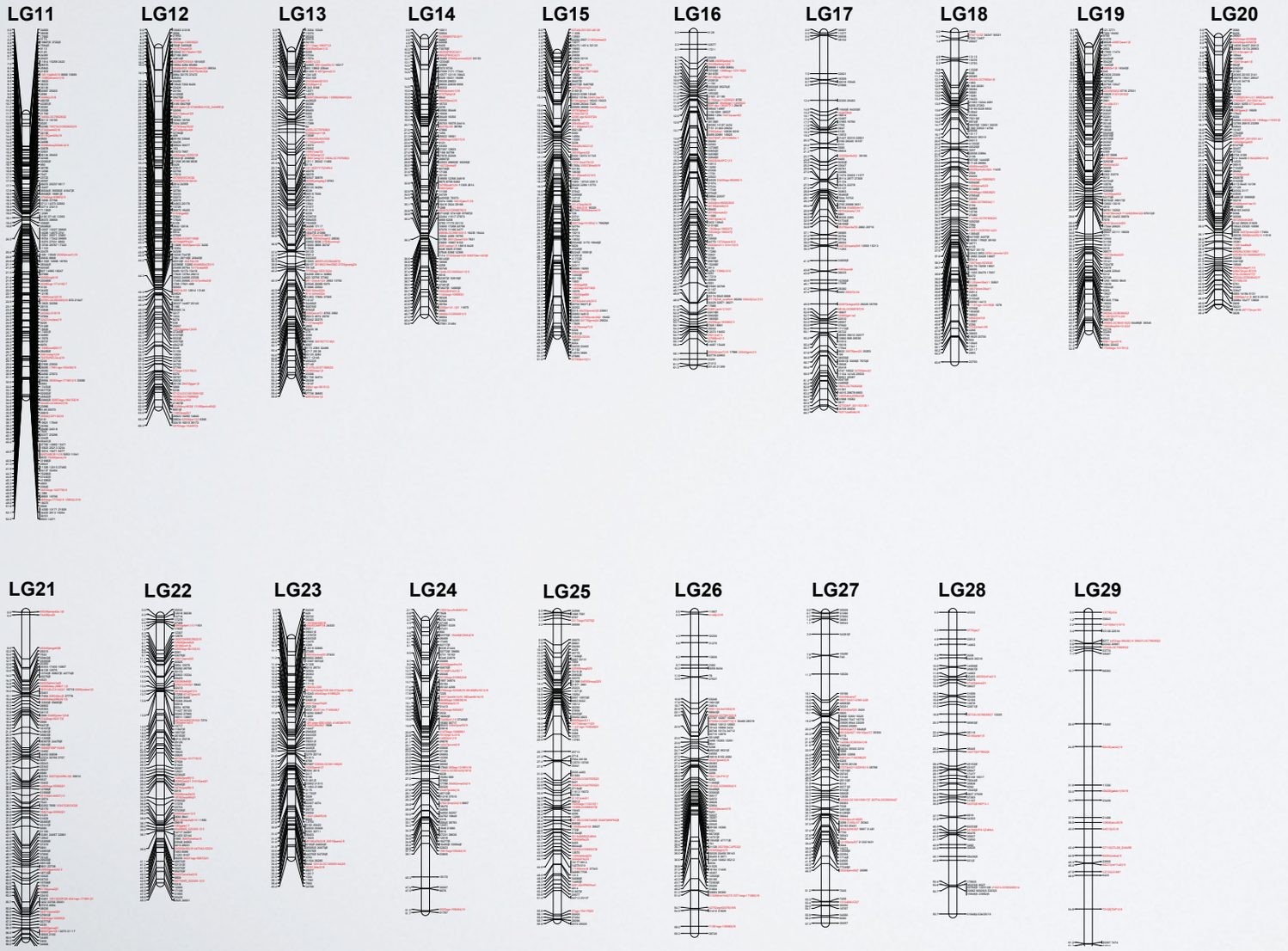
John Postlethwait



F1 Pseudo-Test Cross

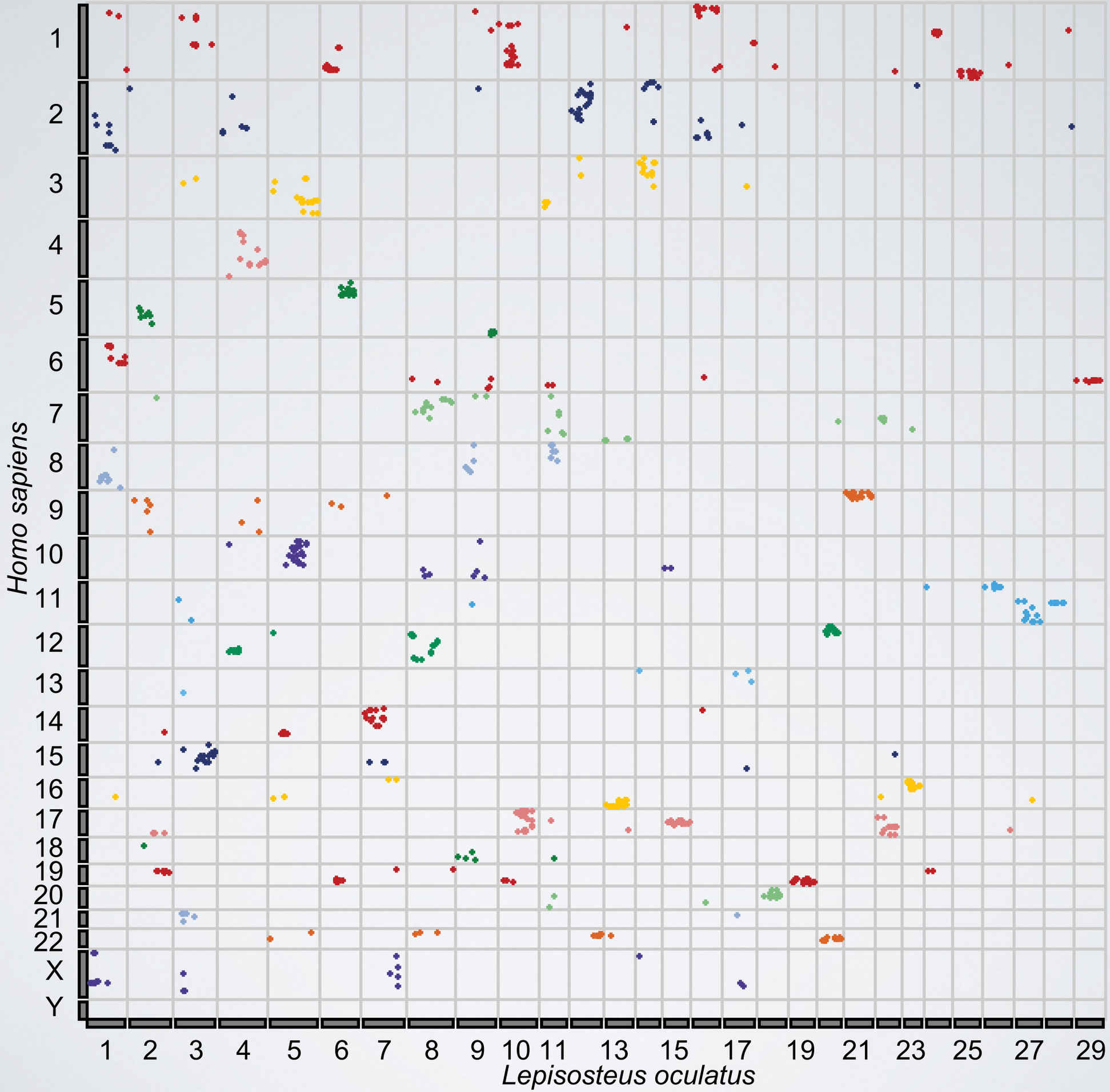
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

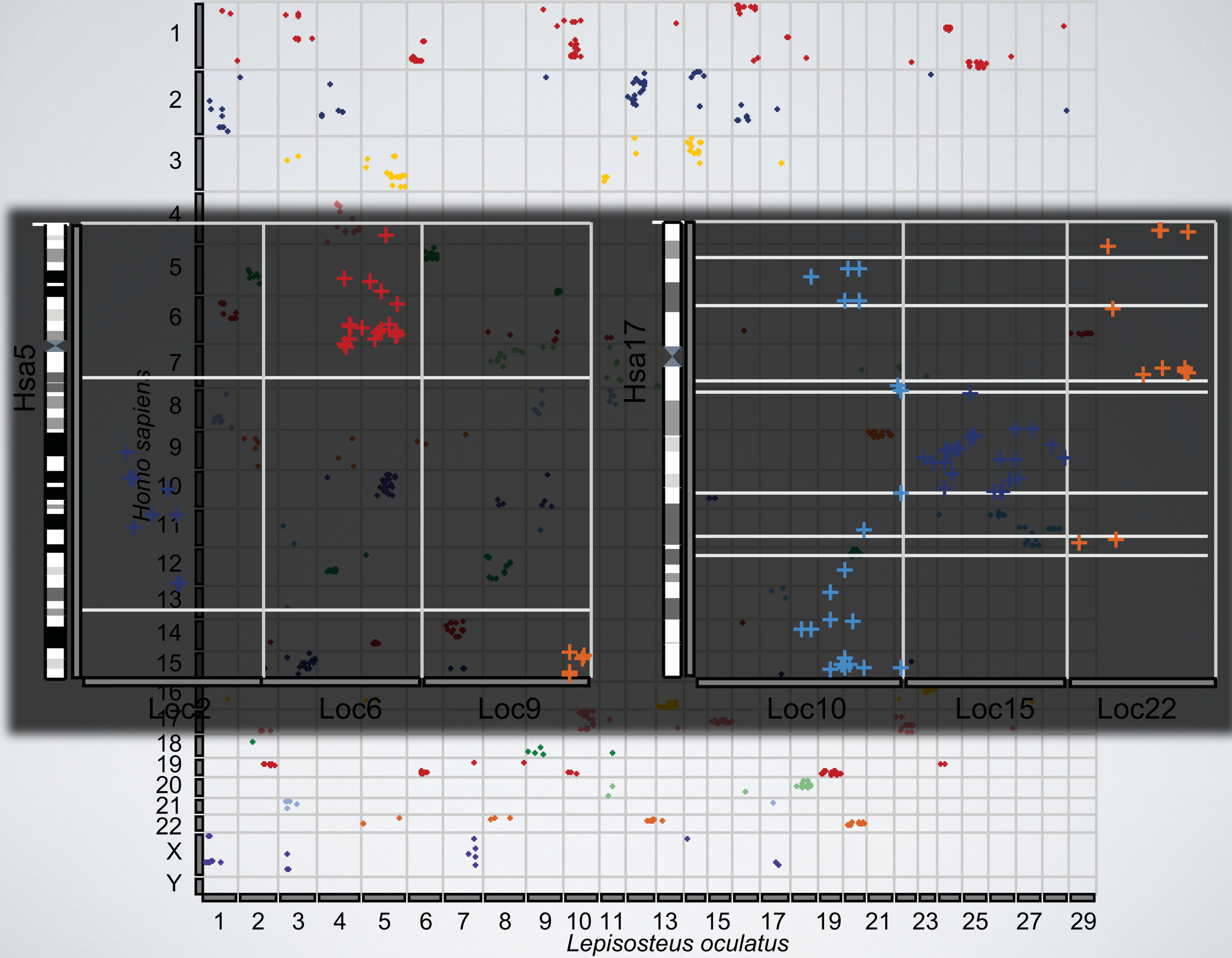


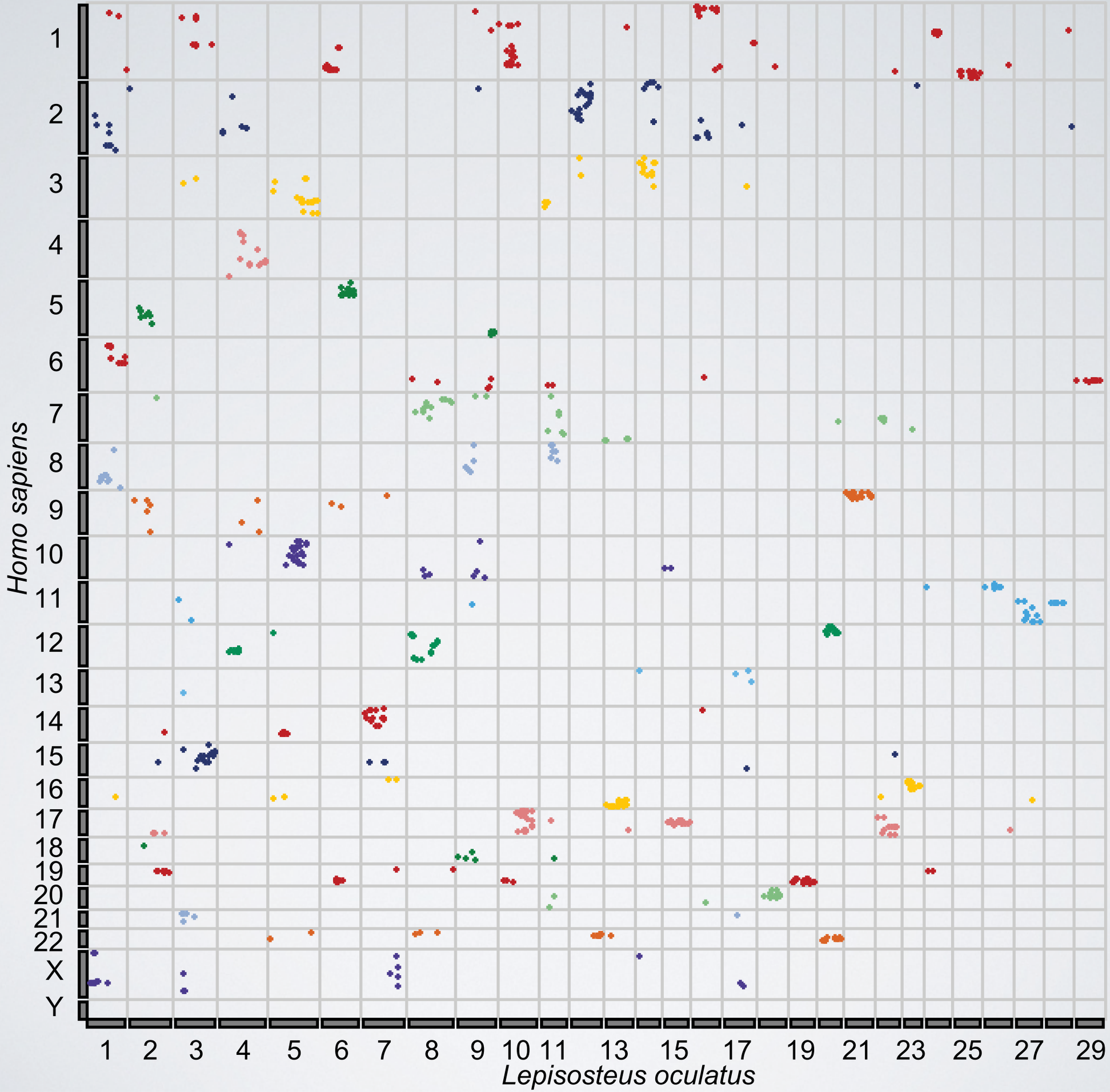


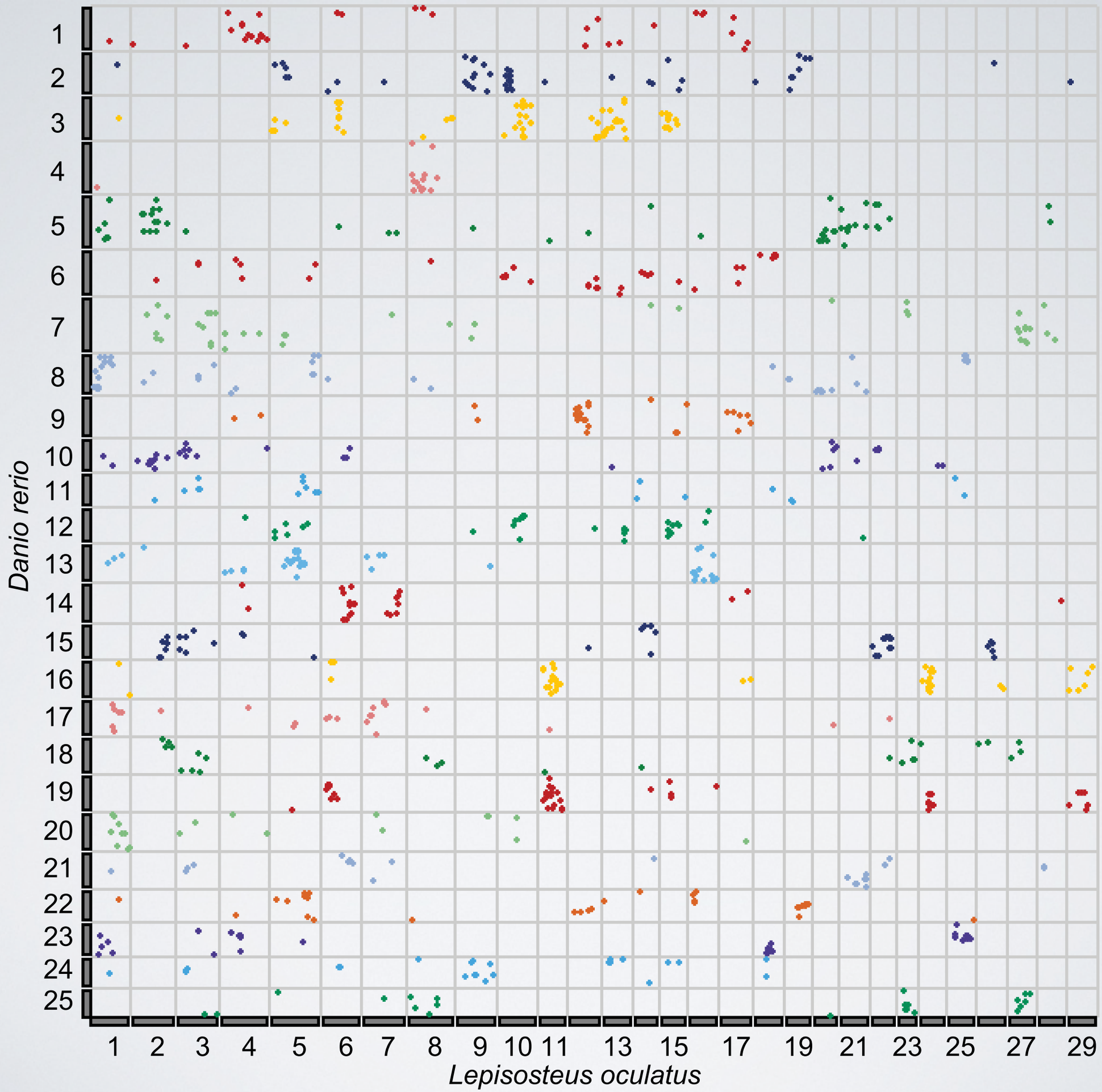
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

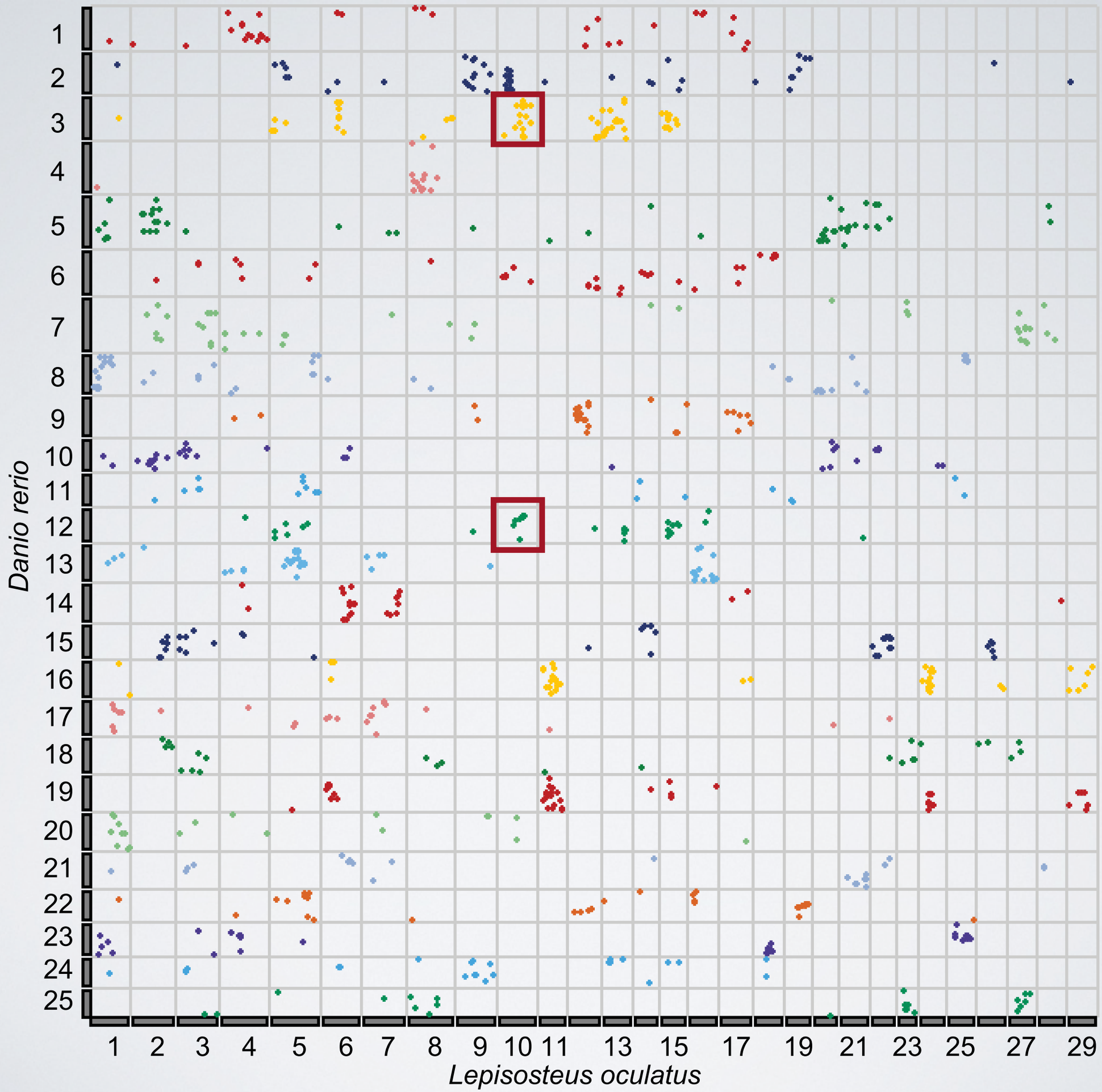
Organism	Markers
Silver carp	483
Guppy	790
Barramundi	240
Catfish	331
Sea bass	368
Cichlid	204
Platyfish	290
Halibut	604
Sea bream	204


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

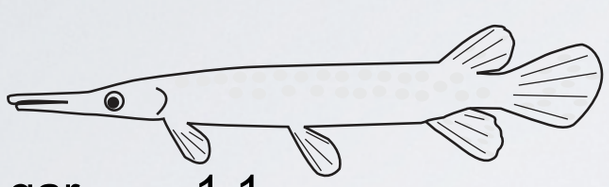

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
1 2 3 4 5 6 7 8 9 10 11 13 15 17 19 21 23 25 27 29
Lepisosteus oculatus
Hsa5
Hsa17
Loc2
Loc6
Loc9
Loc10
Loc15
Loc22


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


Danio rerio
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
1 2 3 4 5 6 7 8 9 10 11 13 15 17 19 21 23 25 27 29
Lepisosteus oculatus


Danio rerio
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
1 2 3 4 5 6 7 8 9 10 11 13 15 17 19 21 23 25 27 29
Lepisosteus oculatus

Hsa17
Loc10
Dre12
Dre3
1:1
1:2
Homo sapiens
Lepisosteus oculatus
Danio rerio
Lepisosteus oculatus



gar

1.1

7.2

2.2

human

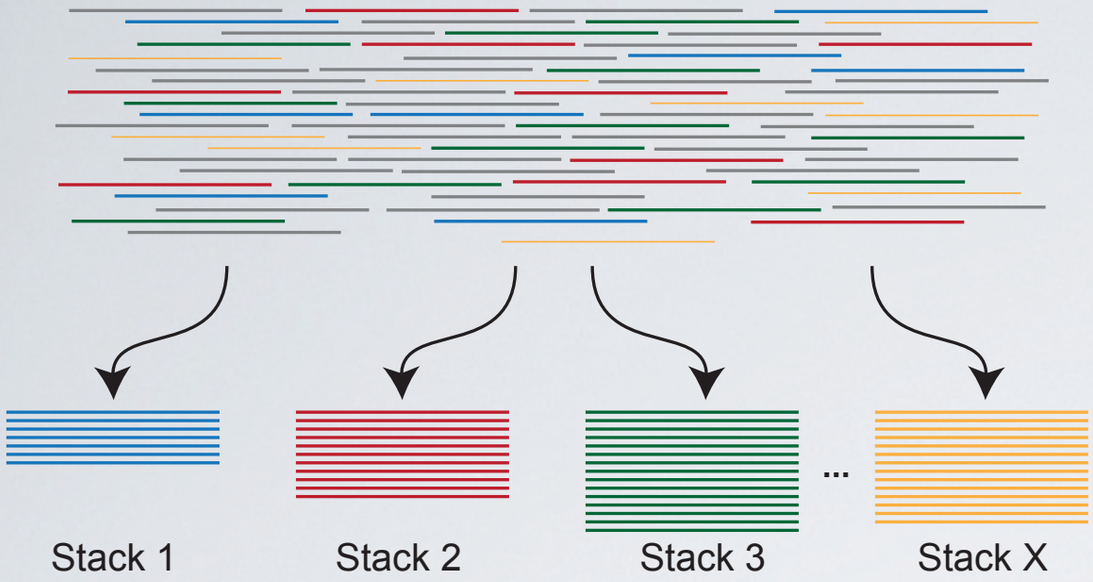
5.4

zebrafish

stickleback

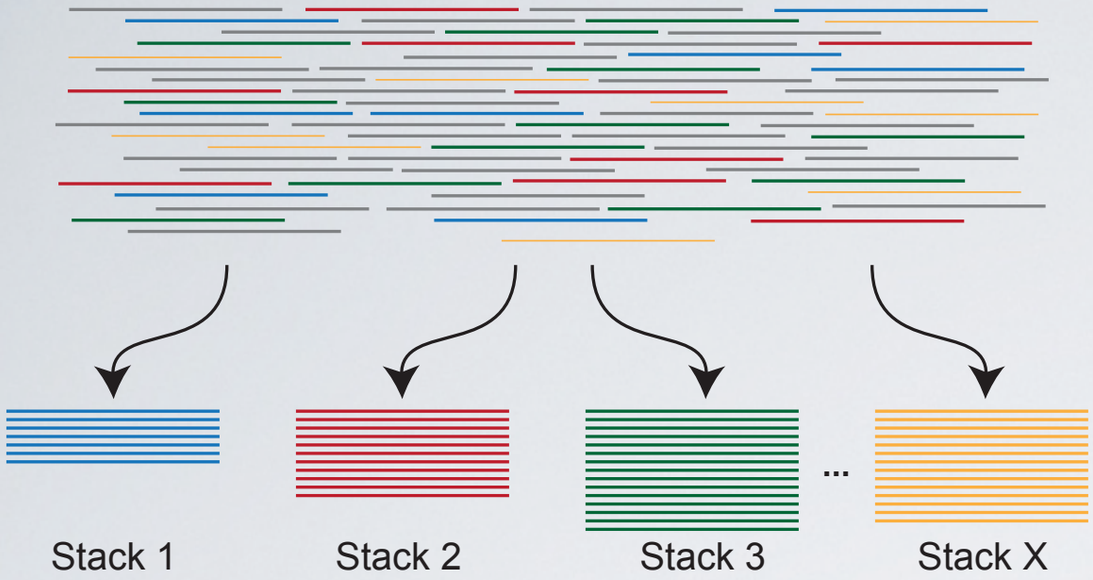
4.8

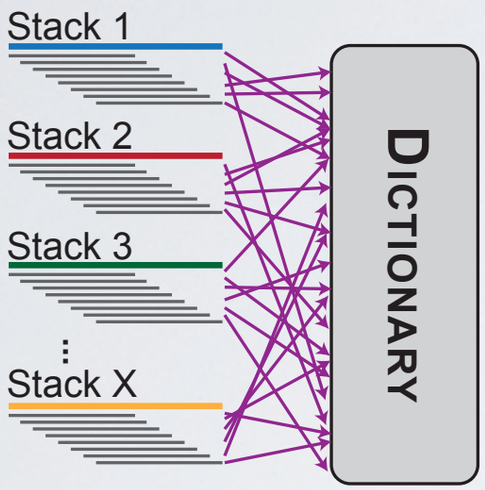
G3: Genes, Genomes, Genetics

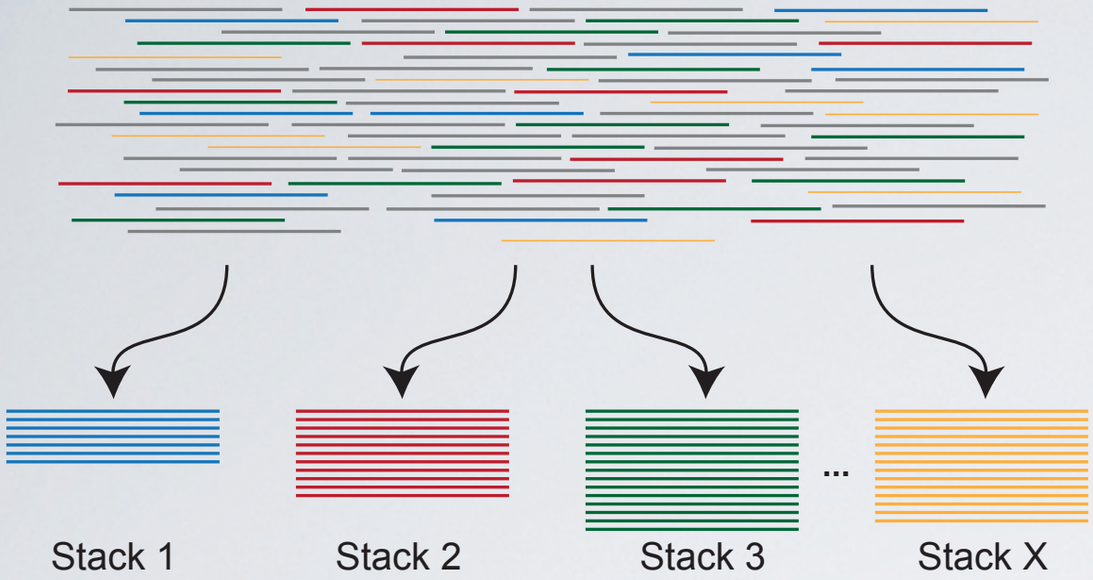


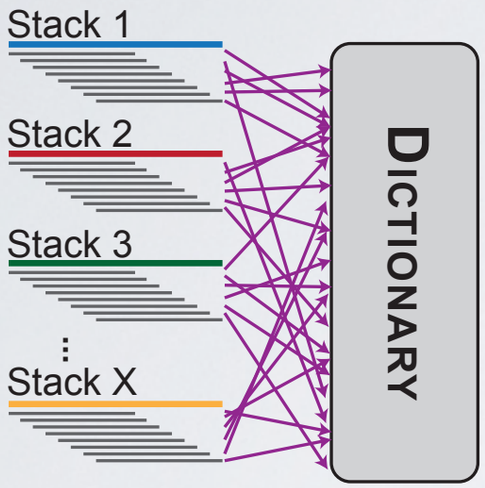
C

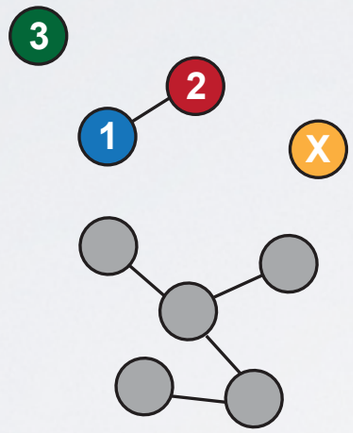
CT

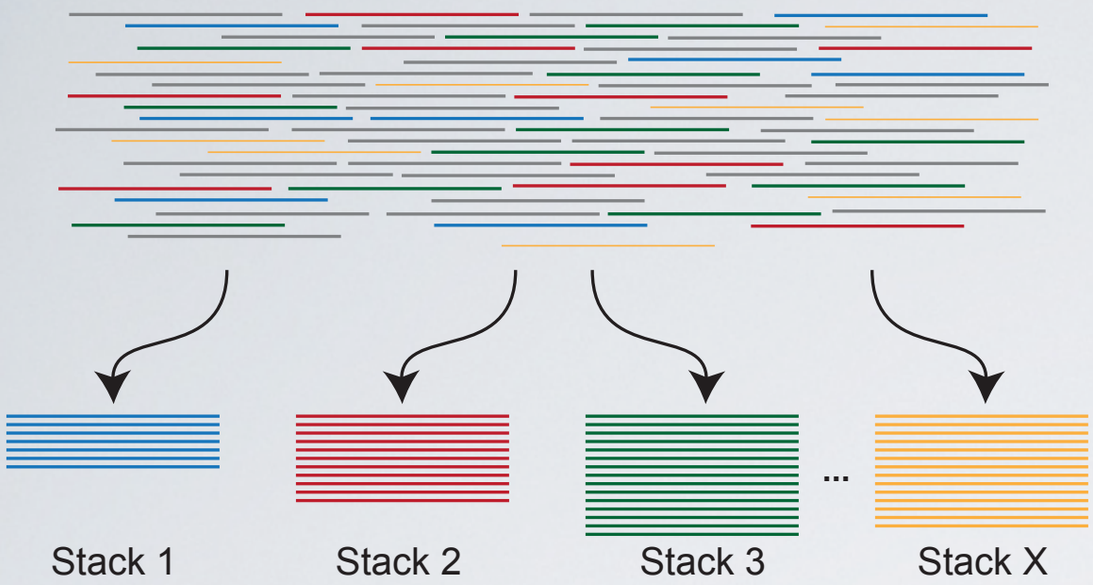

Stack 1
Stack 2
Stack 3
...
Stack X

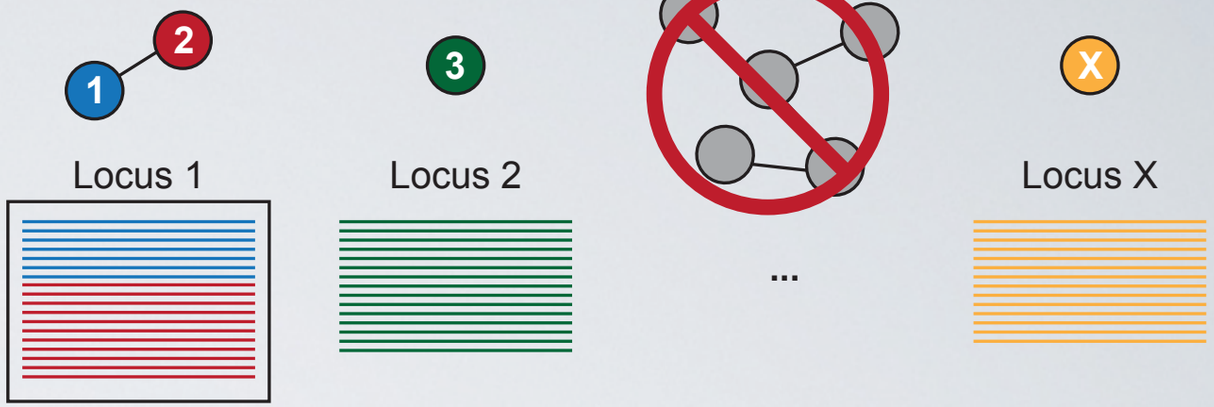

Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

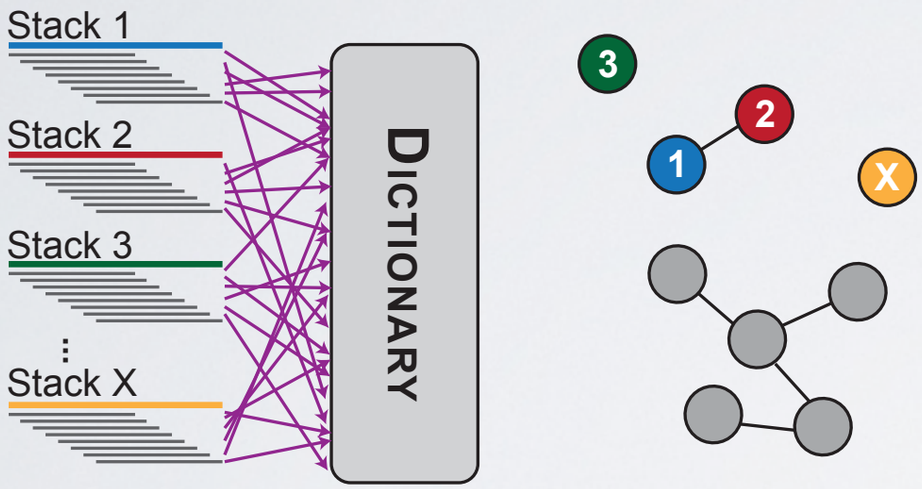

Stack 1
Stack 2
Stack 3
...
Stack X


Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

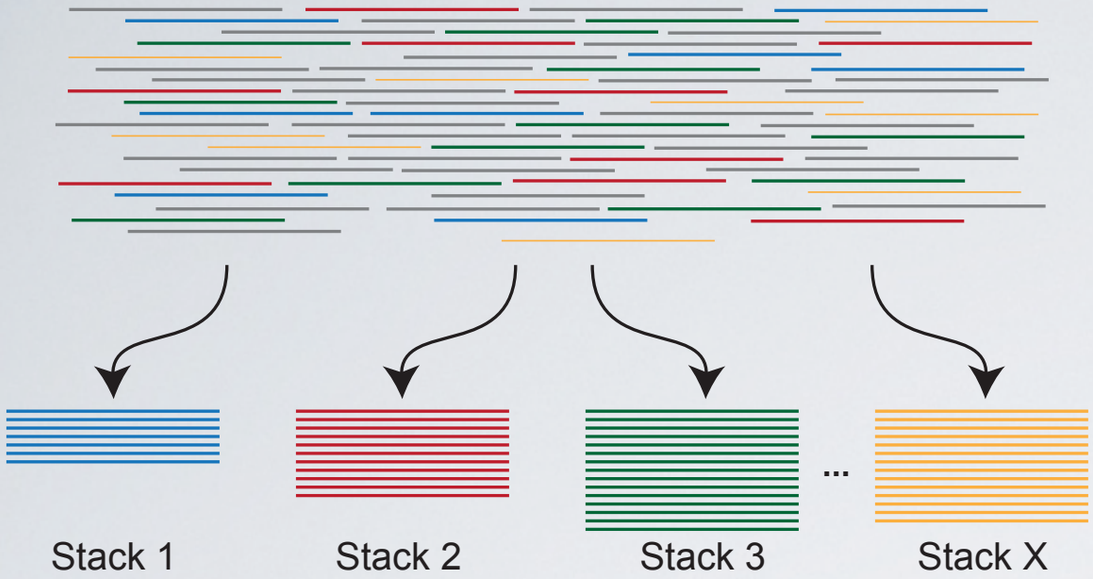

3
2
1
X

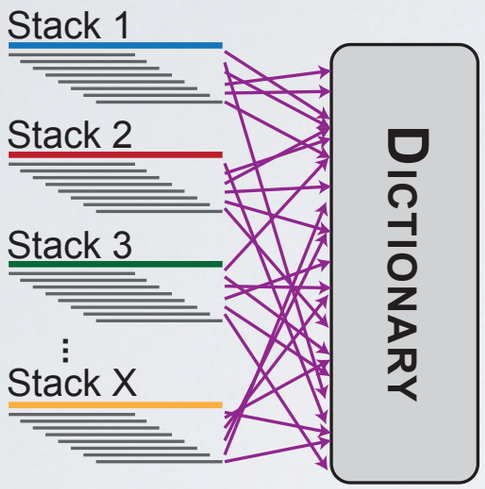

Stack 1
Stack 2
Stack 3
...
Stack X

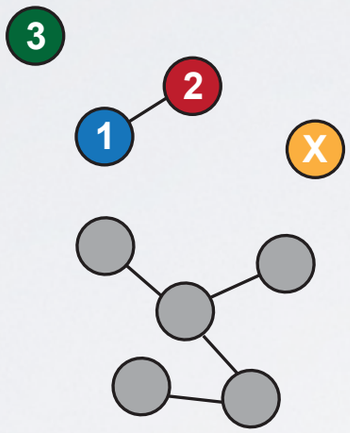

1
2
Locus 1
3
Locus 2
...
X
Locus X

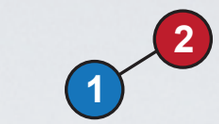

Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

3
2
1
X

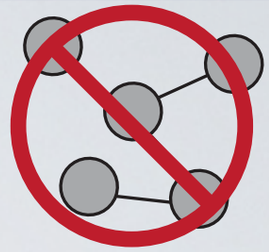

Stack 1
Stack 2
Stack 3
...
Stack X


Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

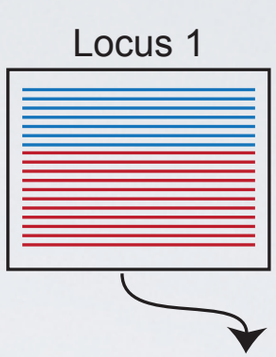

3
2
1
X


2
1

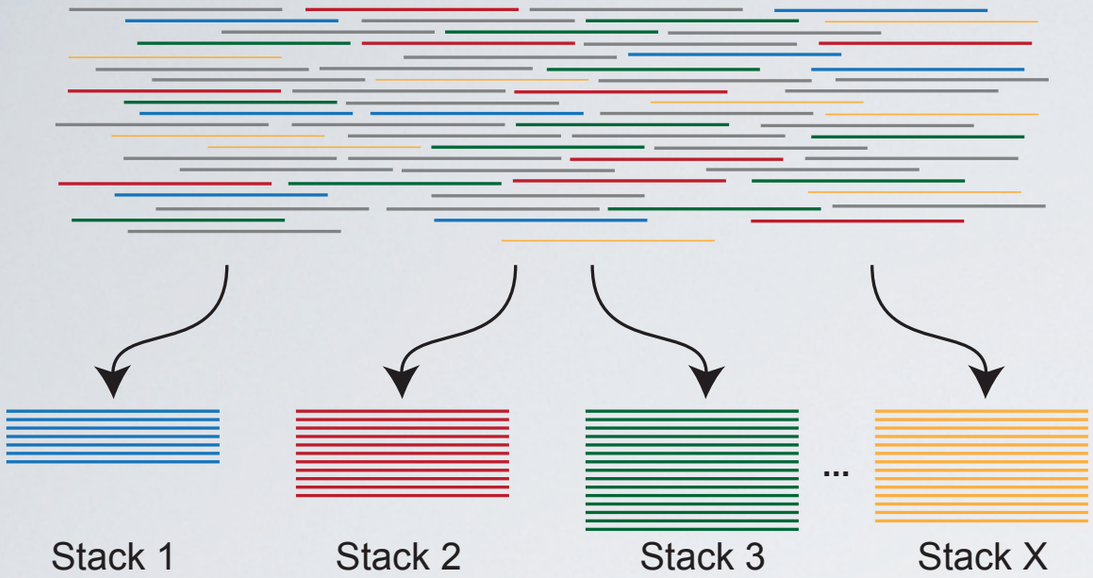

3

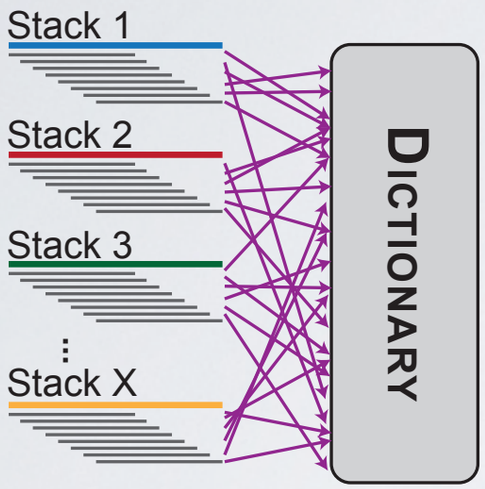


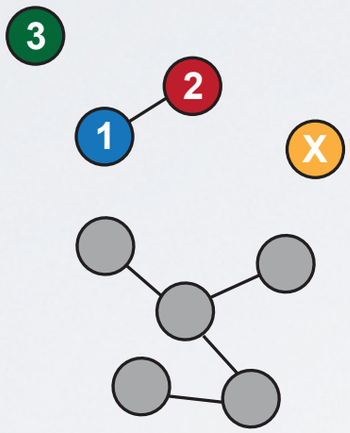

X

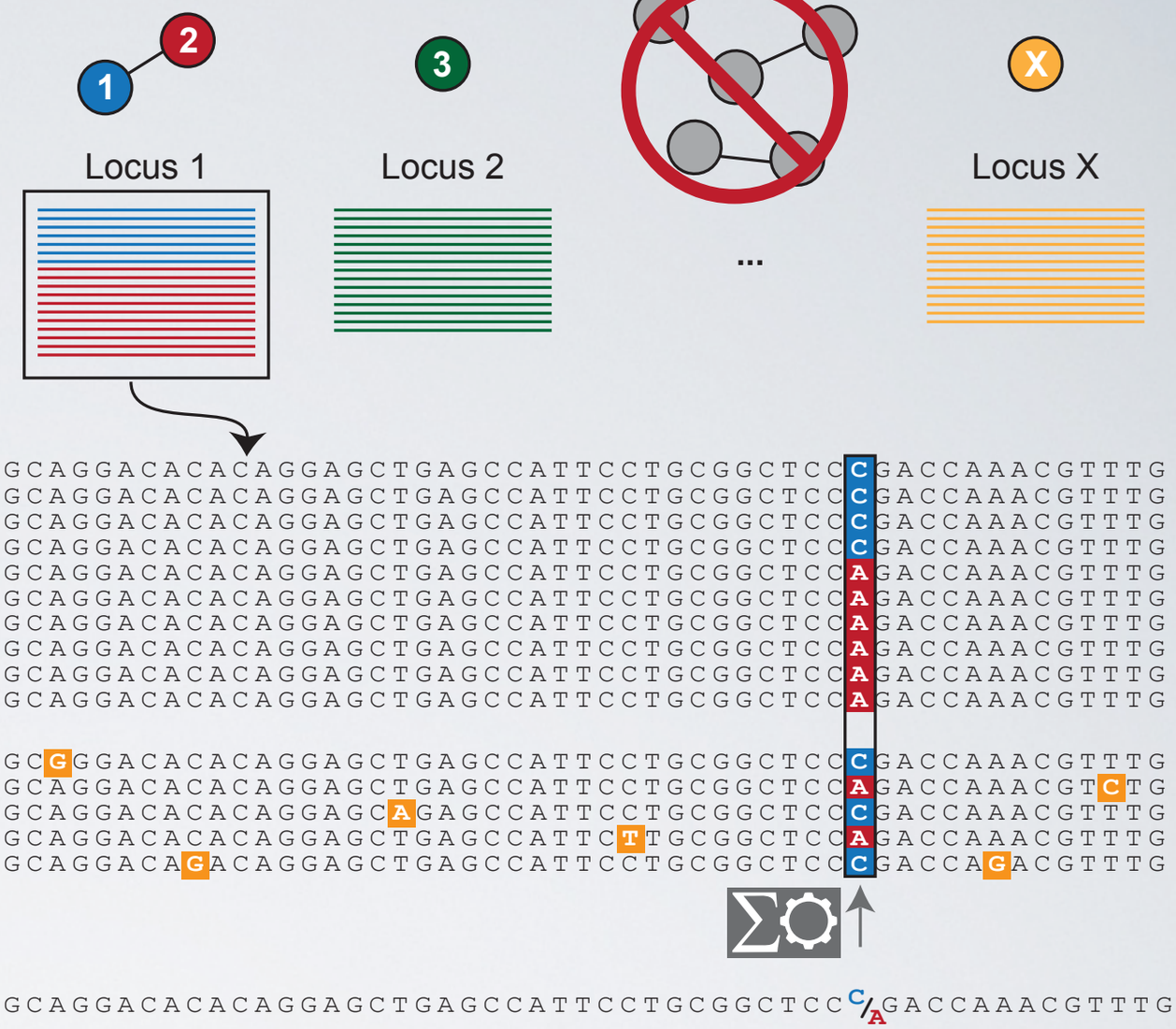

Locus 1
Locus 2
...
Locus X

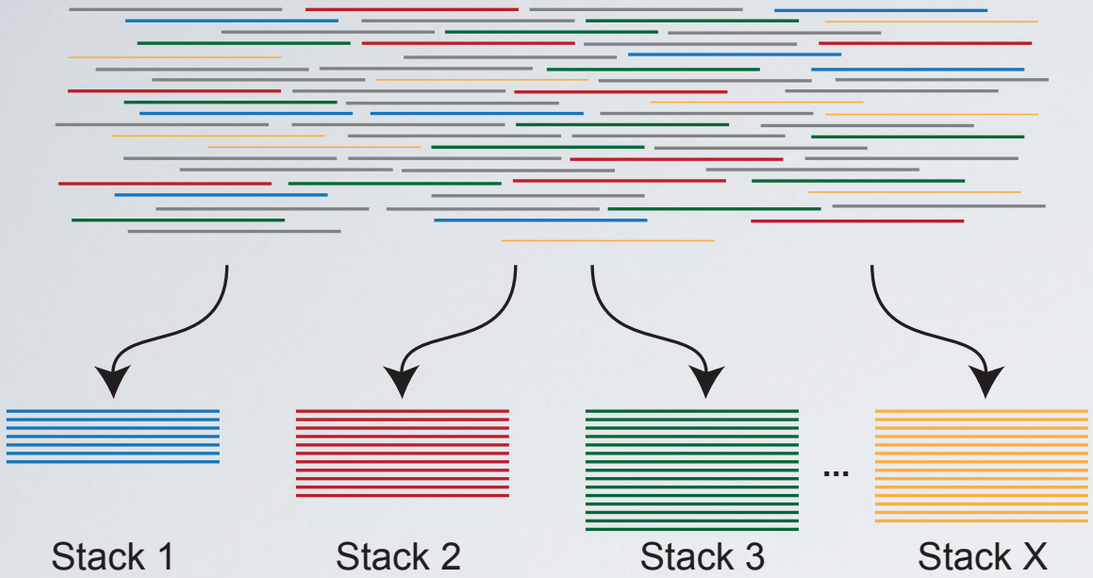

TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG

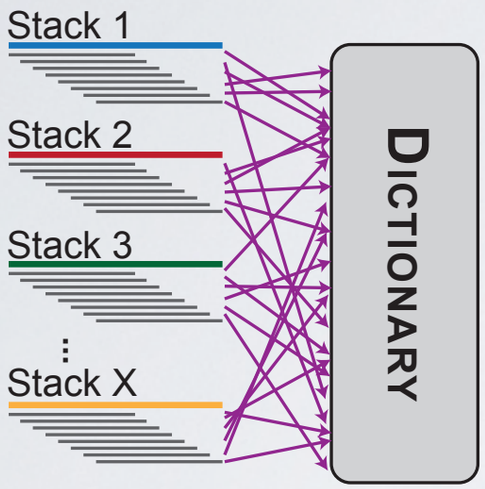

Stack 1
Stack 2
Stack 3
...
Stack X

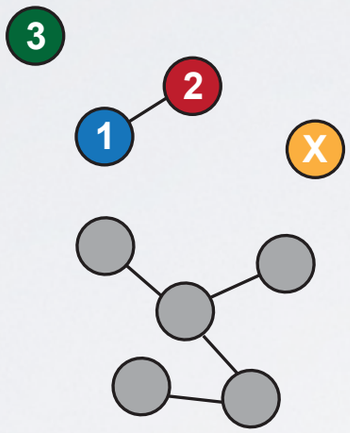

Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

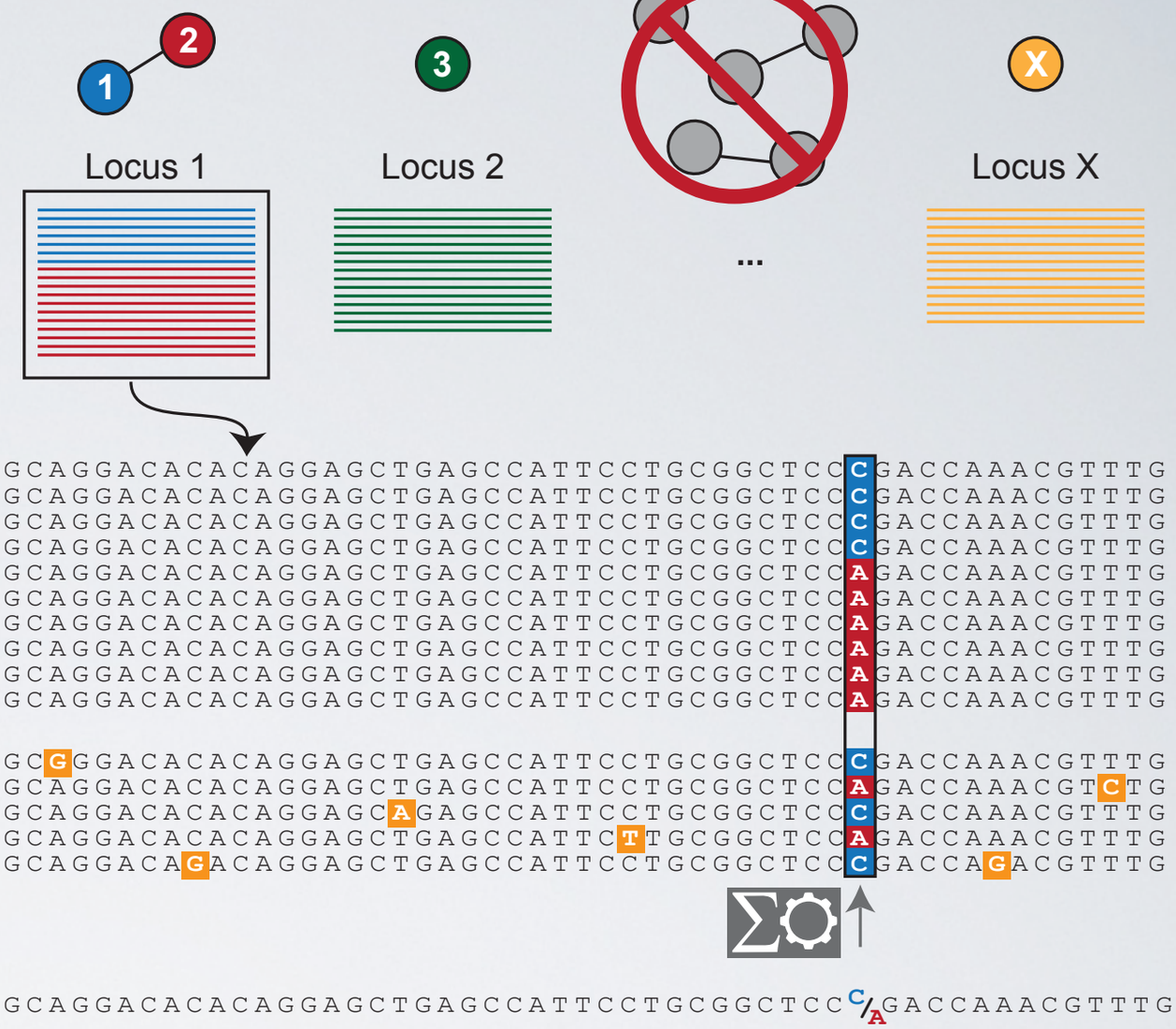

3
2
1
X


2
1
3
X
Locus 1
Locus 2
...
Locus X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCC/AGACCAAACGTTTG


Stack 1
Stack 2
Stack 3
...
Stack X


Stack 1
Stack 2
Stack 3
Stack X
DICTIONARY

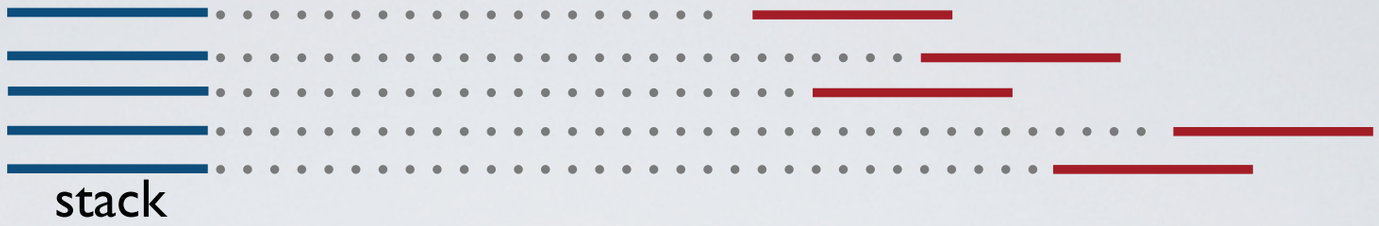

3
2
1
X


2
1
3
X
Locus 1
Locus 2
...
Locus X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCC/AGACCAAACGTTTG

Stacks

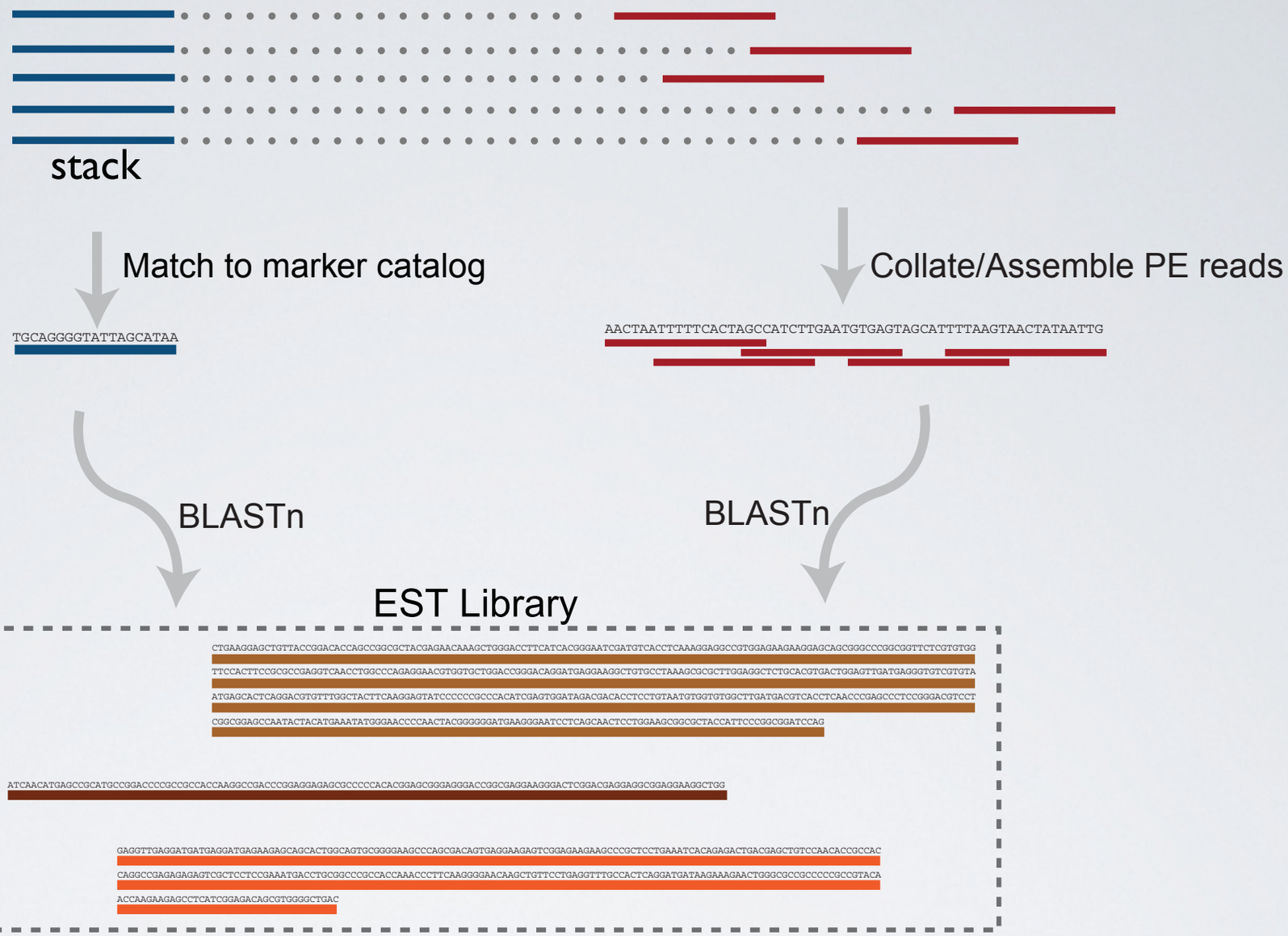

stack

Acquire paired-end sequence



Acquire paired-end sequence

Acquire paired-end sequence

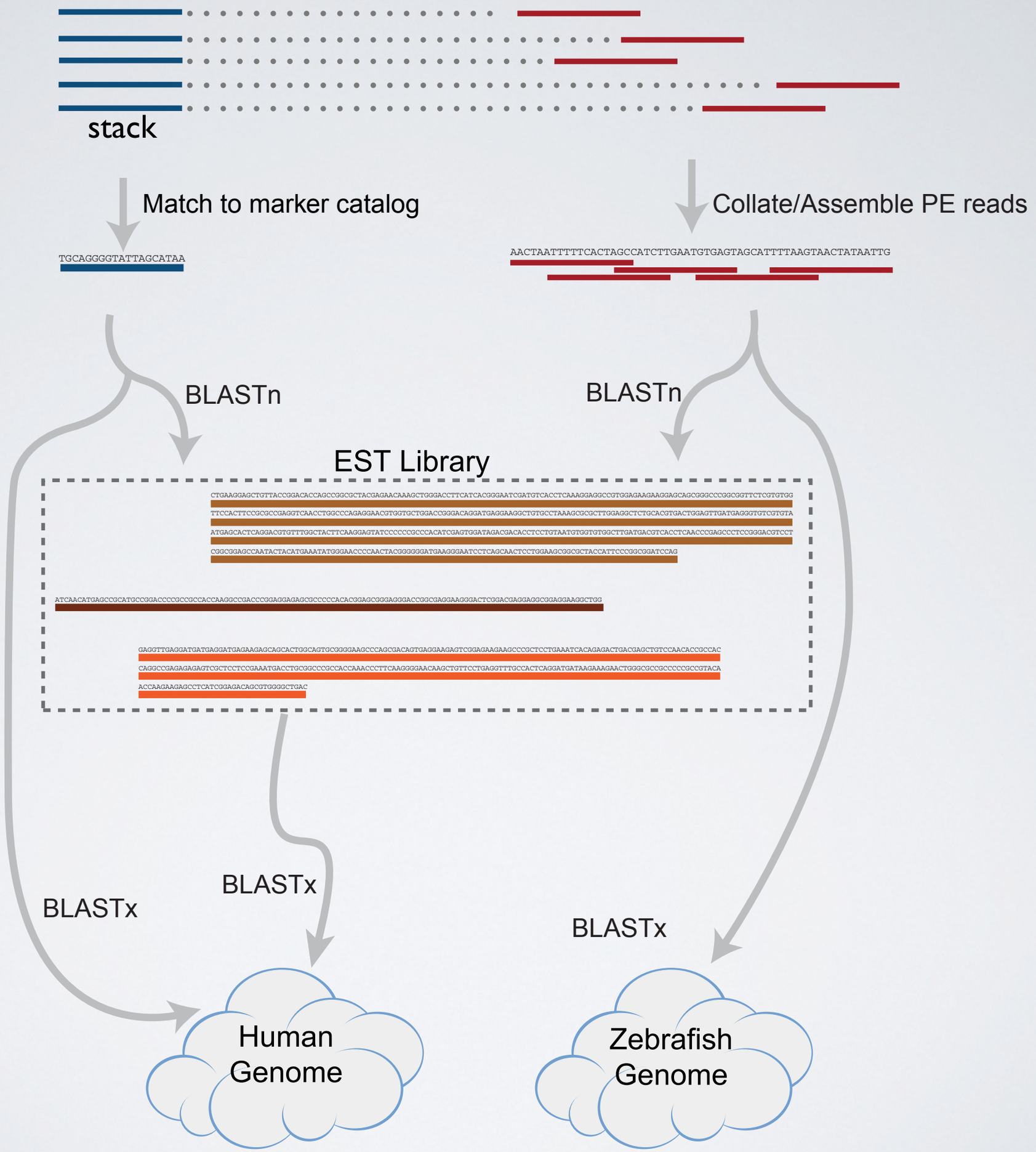


Associate markers / PE contigs with ESTs

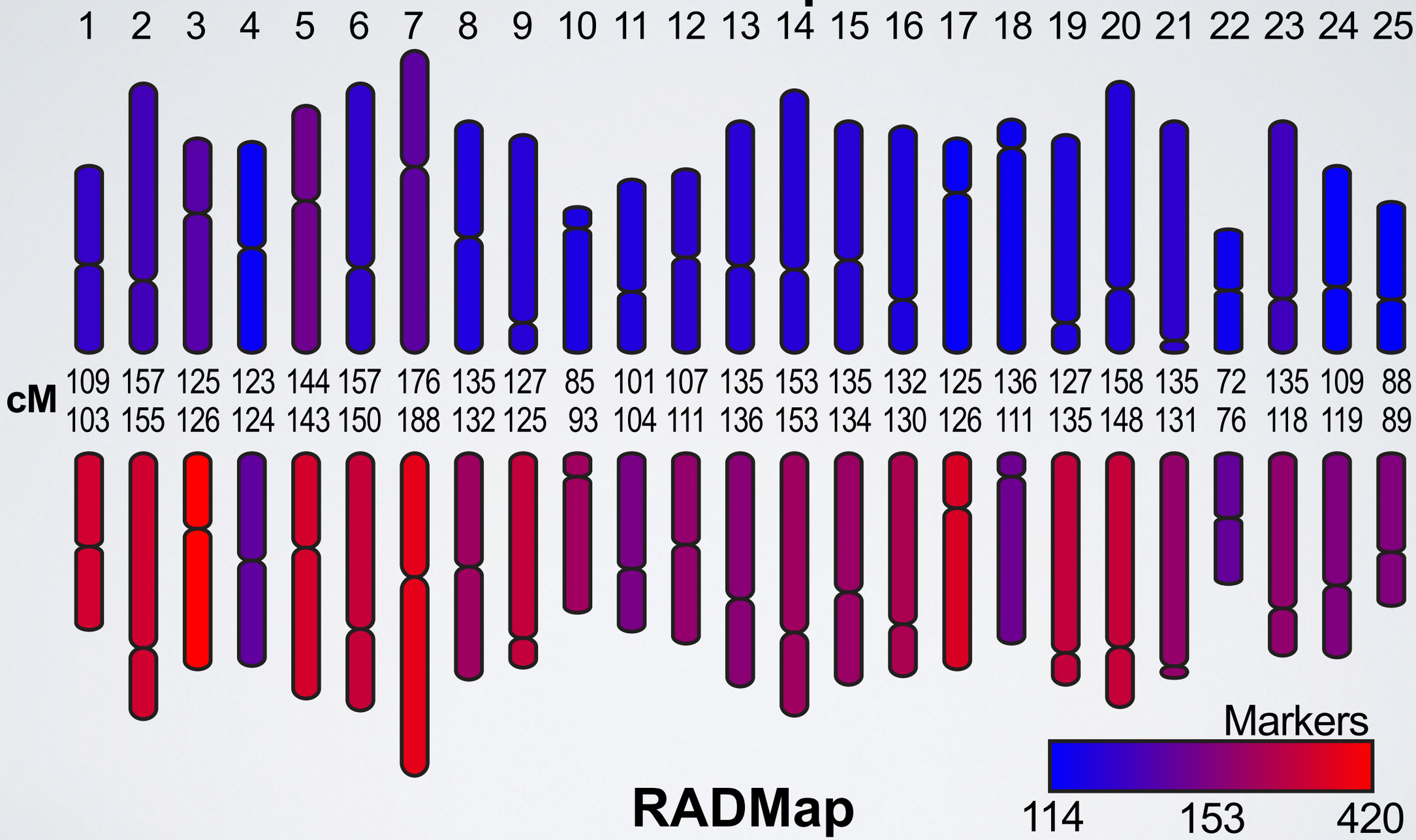
Acquire paired-end sequence

Associate markers / PE contigs with ESTs

Assign orthology to:
- markers
- PE contigs
- ESTs



HSMap



RADMap

DreLG20

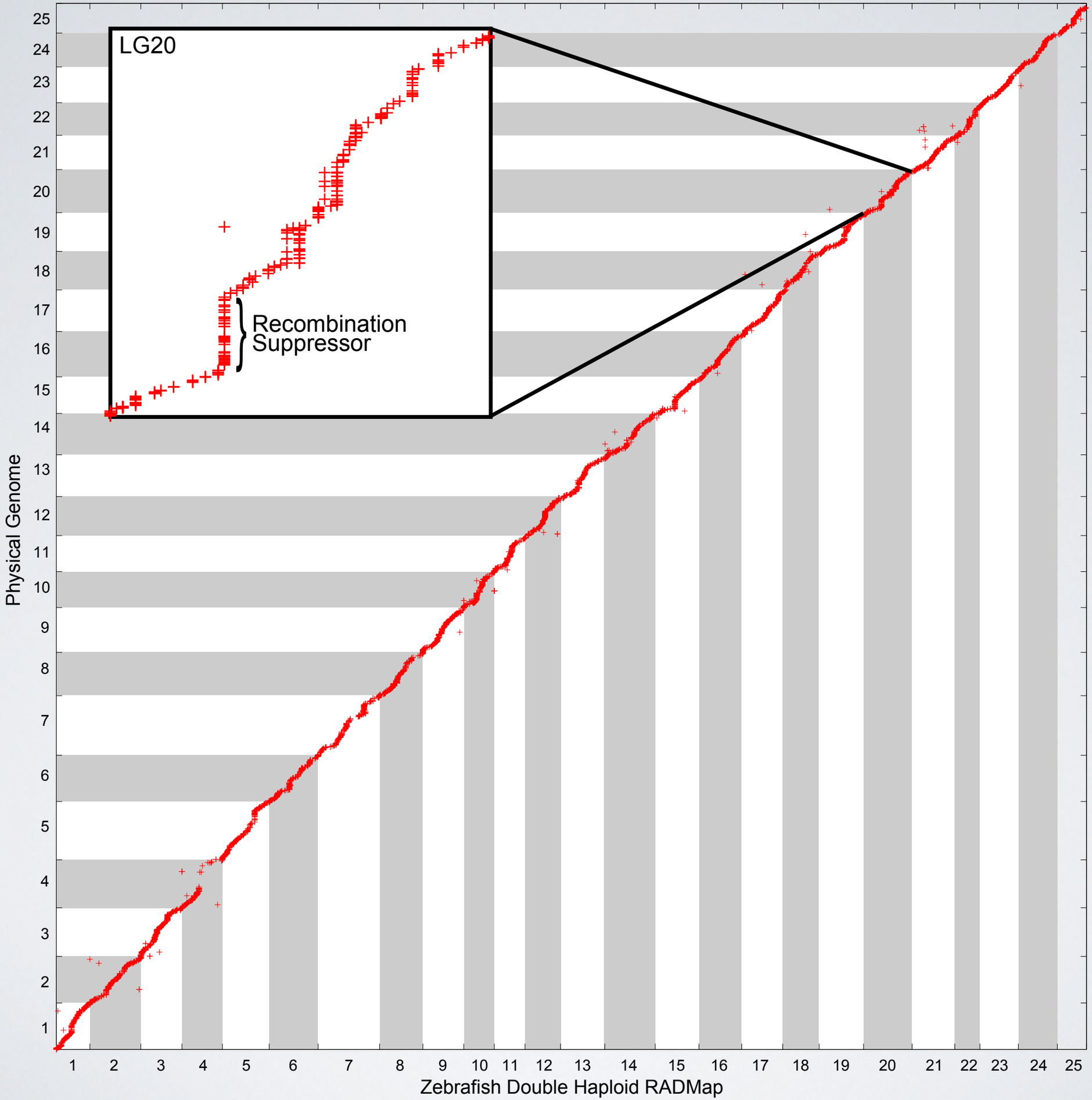
157 EST markers

0.0 HSP20_Z9962
8.2 HSP20_BI474288 HSP20_BC065672
13.2 HSP20_AI657956
15.6 HSP20_pcmt
20.6 HSP20_BC045294 HSP20_BI882204 HSP20_AI497198 HSP20_Z7933 HSP20_Z9708 HSP20_AI331733 HSP20_syncrip
23.1 HSP20_BM861561 HSP20_BC056814
28.1 HSP20_Z10056 HSP20_AI588245
33.1 HSP20_AA605749
35.5 HSP20_AI476962
40.6 HSP20_BG728944
43.0 HSP20_AI397048 HSP20_AI544701
45.5 HSP20_Z11841
56.0 HSP20_AI641655 HSP20_Z3824 HSP20_dtl HSP20_tbr1 HSP20_BC059659 HSP20_fmo5 HSP20_Z21067 HSP20_BI880068 HSP20_AA497290 HSP20_pigc HSP20_AI658305 HSP20_nek2 HSP20_serpina1 HSP20_AW202887 HSP20_BC047862 HSP20_AI959361 HSP20_Z7803 HSP20_BC071538 HSP20_AI584412 HSP20_BC045515 HSP20_Z13214 HSP20_AI959532 HSP20_BC086711 HSP20_nfkbiaa HSP20_AI585033 HSP20_AI641705
63.7 HSP20_AY394945 HSP20_AI667501 HSP20_AI958075
66.2 HSP20_AI477407 HSP20_Z22659 HSP20_AI558899 HSP20_AI959629
71.2 HSP20_six4.1
73.6 HSP20_rtn1
76.0 HSP20_jag2
78.5 HSP20_BC044390
79.8 HSP20_paics
80.9 HSP20_CF550564 HSP20_BC078239 HSP20_Z3211 HSP20_esr1 HSP20_AI878681 HSP20_kit HSP20_BC076242 HSP20_AW171519 HSP20_AI658298 HSP20_pdgfra HSP20_CD596570 HSP20_AI957729 HSP20_AI943246 HSP20_BC065446 HSP20_AA495418 HSP20_cp450 HSP20_AI584331 HSP20_AF370884 HSP20_AI584766 HSP20_AL919819 HSP20_AI667152
88.7 HSP20_AI601548 HSP20_AI721928 HSP20_AI877511
89.7 HSP20_BI671369
91.1 HSP20_AW174998 HSP20_Z8150 HSP20_BC044195 HSP20_lft1 HSP20_cpsf3 HSP20_rpl7l1 HSP20_AW420838 HSP20_rab32 HSP20_lbr HSP20_CA474314 HSP20_BC056589 HSP20_acat2 HSP20_pdip5 HSP20_ivns1abpa HSP20_CK025611 HSP20_wtap HSP20_snap25a HSP20_AI584357
91.9 HSP20_AI618559
93.5 HSP20_AL722719 HSP20_fntb HSP20_AI959609 HSP20_sox11b HSP20_AI882839 HSP20_BC059690 HSP20_max HSP20_vps18
93.7 HSP20_BC071385
98.2 HSP20_Z21123
98.5 HSP20_pro2013 HSP20_BC076168 HSP20_BI878481
100.9 HSP20_Z536 HSP20_AY751464 HSP20_ift172
103.4 HSP20_CK027003 HSP20_BM889954 HSP20_tde2 HSP20_cx43 HSP20_Z20046
113.9 HSP20_AW202791 HSP20_AW173950 HSP20_laptm4a
116.4 HSP20_AF509492 HSP20_matn3a
124.1 HSP20_AI657979
126.5 HSP20_AI641589 HSP20_Z22041 HSP20_bmp2b HSP20_AI397344 HSP20_BC059496 HSP20_Z7171 HSP20_BQ263921 HSP20_AI722296 HSP20_dtnb
127.5 HSP20_BI887534
128.9 HSP20_extl3
136.6 HSP20_disp1 HSP20_AI618255 HSP20_AI330882 HSP20_hsp90b
144.3 HSP20_BC078246 HSP20_BC081476 HSP20_Z8554 HSP20_AI546038
149.3 HSP20_Z4329 HSP20_AI883979
151.8 HSP20_AI959738
156.8 HSP20_zp2.4 HSP20_Z21485 HSP20_zp2.2
162.5 HSP20_prpf39

DreLG20

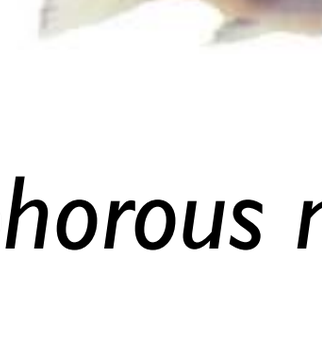
311 RAD markers

0.0 HSP20_Z9962
5.3 HSP20_BI474288 189499 88203 108274 HSP20_BC065672 161400|20|0.38 83039|20|0.53 113294|20|0.28 52538|20|0.29 45411|20|0.08 37987
7.7 18209|20|1.30 74007 HSP20_AI657956
10.1 155098|20|1.52 3841|20|1.46 104611|20|1.49 178331|20|1.54 196186|20|1.33 HSP20_pcmt 78779 131615|20|1.39 135773 42193 163435|20|1.52 206518|20|1.54
15.2 190886|20|2.23 281308 HSP20_Z7933 85877 103668|20|2.56 HSP20_Z9708 104895 61591 199183 101135 57921 72038|20|2.23 72048 HSP20_BC045294 HSP20_AI331733 29302|20|2.53 203308|20|1.71
17.7 107272
22.7 89906|20|2.51
24.0 88499
25.2 40139|20|2.23 HSP20_BI882204 HSP20_AI497198 98054|20|2.94 HSP20_syncrip
27.6 HSP20_BC056814 HSP20_BM861561
32.6 60206 HSP20_Z10056 138380|20|3.47 HSP20_AI588245 241241|20|3.56
35.0 56070 43864
40.0 HSP20_AI476962 106173 36003 207956 5854 81593
42.5 HSP20_AA605749
50.2 HSP20_BG728944 70020
52.6 HSP20_AI544701 HSP20_AI397048 207768|20|4.79
55.1 HSP20_Z11841
57.5 67359|5|73.22
62.5 19472|20|4.10 214420
67.5 106737 205995 97203 114037 106784
68.8 28102|20|12.88 38630 6201|6|35.03 HSP20_Z3824 17415|20|11.41 90234 182057|20|10.23 72615|20|6.29 HSP20_fmo5 HSP20_BI880068 25698|20|11.72 101621|20|13.04 HSP20_nek2 94344|20|7.40 HSP20_AA497290 HSP20_AI641655 HSP20_AI959532 HSP20_serpina1 HSP20_AI584412 HSP20_dtl 137481|20|6.60 62834 84084|20|12.21 21332|20|13.56 HSP20_AI658305 80346 71031 15475|20|12.15 46690 43401|20|15.74 235653|20|8.00 17036 HSP20_BC045515 54666 175327|20|11.72 HSP20_Z7803
69.9 HSP20_pigc 181424 215418|20|10.15 95215 12402|20|7.17 34382|20|6.80 HSP20_BC059659 107321|20|14.60 HSP20_AI959361 HSP20_tbr1 HSP20_BC047862 102373|20|12.88 125203 187620|20|7.40 179589|20|14.13 HSP20_nfkbiaa HSP20_Z21067 17783|20|13.79 132796|20|7.57 211660 53374 18865|20|14.13 25371 HSP20_Z13214 HSP20_AI585033 752|20|10.41 162412|20|8.18 HSP20_BC071538 58483|20|9.65 102984|20|6.45 11527|4|26.71 104135 6341|20|9.32 62476 51222 30079 81880|20|15.53 HSP20_AW202887 HSP20_BC086711 59446|20|10.42
71.1 HSP20_AI641705
72.3 276232 44873
74.8 22609|20|17.20
77.2 61370|20|18.00 HSP20_AY394945 HSP20_AI958075 94341 HSP20_AI667501 16196 68692|20|17.71 115759|20|17.56
79.7 HSP20_AI959629 13413 HSP20_AI477407 58651 HSP20_Z22659 25362 HSP20_AI558899 30986
80.8 50273|20|18.43
82.1 61405
84.5 HSP20_six4.1
86.9 43227|20|20.39 HSP20_rtn1 39399 44279|20|20.19
89.4 149183 HSP20_jag2 24690 256761|20|20.76
91.8 28888 HSP20_BC044390
92.6 HSP20_paics HSP20_AI878681 HSP20_AI584766 65907 HSP20_AL919819 21769 49257 HSP20_CF550564 HSP20_AI943246 25143 10723|20|23.19 27111 HSP20_pdgfra 148055 14086 187914|20|21.29 70312|20|25.28 HSP20_AI584331 73227|20|25.29 HSP20_AI957729 95866|20|24.68 HSP20_CD596570 76978|20|25.03 154261|20|26.87 200798|20|24.76 HSP20_AA495418
94.2 170680|20|21.29 HSP20_AF370884 HSP20_Z3211 52953|20|25.88 117549|20|21.99 HSP20_AW171519 HSP20_BC065446 84436 174175|20|26.49 HSP20_esr1 10117 27821|20|23.74 176162|20|25.28 80478 HSP20_cp450 283097|20|23.72 118666 111788 HSP20_BC078239 74725|20|21.28 45730|20|26.42 56103|20|25.85 HSP20_BC076242 134715|20|26.13 HSP20_kit HSP20_AI667152 HSP20_AI658298 164776|20|26.63
96.7 40177|20|27.17 20907
101.7 HSP20_AI601548 HSP20_vps18 HSP20_fntb HSP20_AI721928 94829 226493 88264|20|29.44 28378|20|29.39 HSP20_AI959609 HSP20_BC059690 HSP20_sox11b HSP20_AI877511 HSP20_AL722719 1830 HSP20_AI882839 74974|20|28.54 41637|20|28.37 73269 HSP20_max 184471|20|29.92 28998 HSP20_AI618559 68582
102.8 HSP20_BC071385 HSP20_ivns1abpa 3819 HSP20_CA474314 51438 43683 HSP20_lbr 41913|20|33.57 HSP20_AI584357 HSP20_AW420838 18033 32472 233533|20|33.30 HSP20_BC044195 35403 71103|20|31.07 HSP20_BI671369 58737 HSP20_Z8150 120123|20|33.81 76093 87588 27349 HSP20_AW174998 HSP20_rpl7l1 HSP20_snap25a
104.1 HSP20_acat2 HSP20_lft1 HSP20_CK025611 57254 108462 38186 39408 282846|20|36.55 HSP20_wtap 28847|20|34.52 22260 64895 27713 8218 105963 HSP20_pdip5 HSP20_cpsf3 17492|20|33.85 47962 HSP20_rab32 63389 86456 HSP20_BC056589 193161
105.5 HSP20_Z21123
106.5 84650 24109|20|36.92 21140|20|37.53 52024 1108
108.9 HSP20_BI878481 HSP20_BC076168 218376|20|36.61 89406 HSP20_pro2013 42874|20|39.20 HSP20_AY751464 HSP20_ift172 236660|20|38.99 190618 HSP20_Z536 59181
111.4 HSP20_cx43 175748 HSP20_Z20046 16015 83079 HSP20_CK027003 12012 9769 125267|20|39.46 25128 55332|20|41.47 HSP20_BM889954 49759 40131 108672 99723 216467 HSP20_tde2 100206|20|40.27 30268
113.8 9673
116.3 126311|20|42.34
118.7 186467|20|42.44 7535|20|42.44 HSP20_laptm4a HSP20_AW202791 106327 24848 12125
121.1 16581 29683 10569|20|42.92 HSP20_AW173950 53384|20|42.66 91984|20|43.46 70258 42912 HSP20_AF509492 87856 167397 85997|3|52.28 43254
123.6 HSP20_matn3a 27727|20|43.70 1036 96244
126.0 76239
128.5 83566
130.9 HSP20_AI657979
132.2 33762 HSP20_BC059496 HSP20_AI397344 101638 105727 HSP20_BQ263921 111083|20|45.92 HSP20_bmp2b 106911 79520|20|44.90 129396|20|45.03 HSP20_AI641589 11184 98425 18987 103353
133.3 185091 48283|20|47.83 110005 171886 130382 HSP20_Z7171 61982 HSP20_AI722296 HSP20_Z22041 HSP20_dtnb 65839 213105
134.4 HSP20_BI887534
135.7 10842 74260 HSP20_extl3
143.4 157786|16|47.67 43547 115741|20|2.75 HSP20_hsp90b 29204|11|0.73 HSP20_AI330882 8055|11|0.73 39049|20|49.81 234816 5191|20|49.58 227868 19531 HSP20_AI618255 HSP20_disp1
148.4 35240|20|50.09 65980
150.9 HSP20_BC081476 106774 137089 HSP20_Z8554 HSP20_BC078246 HSP20_AI546038
153.3 42094 83718
155.7 HSP20_Z4329 HSP20_AI883979
158.2 HSP20_AI959738 74047|20|49.26 35474 54939
160.6 53260 142239|20|51.02 5667
161.7 18116
163.0 HSP20_zp2.2 HSP20_Z21485 18327 36903 HSP20_zp2.4 37064 140570 45291|20|51.85 26667 76344 39382 102396 55296 87608
168.7 HSP20_prpf39

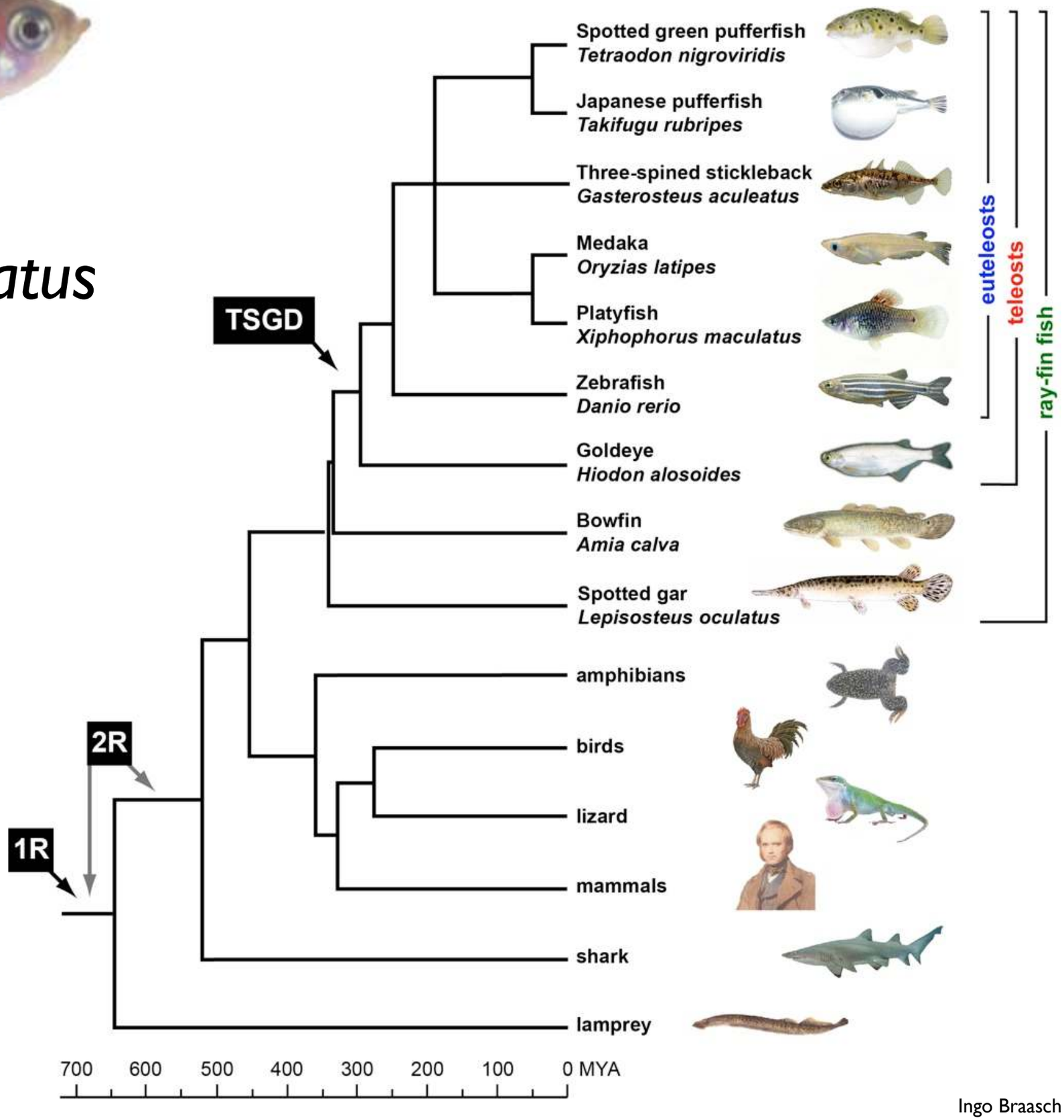

LG20
Recombination
Suppressor
Physical Genome
Zebrafish Double Haploid RADMap
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Xiphophorous maculatus

Angel Amores

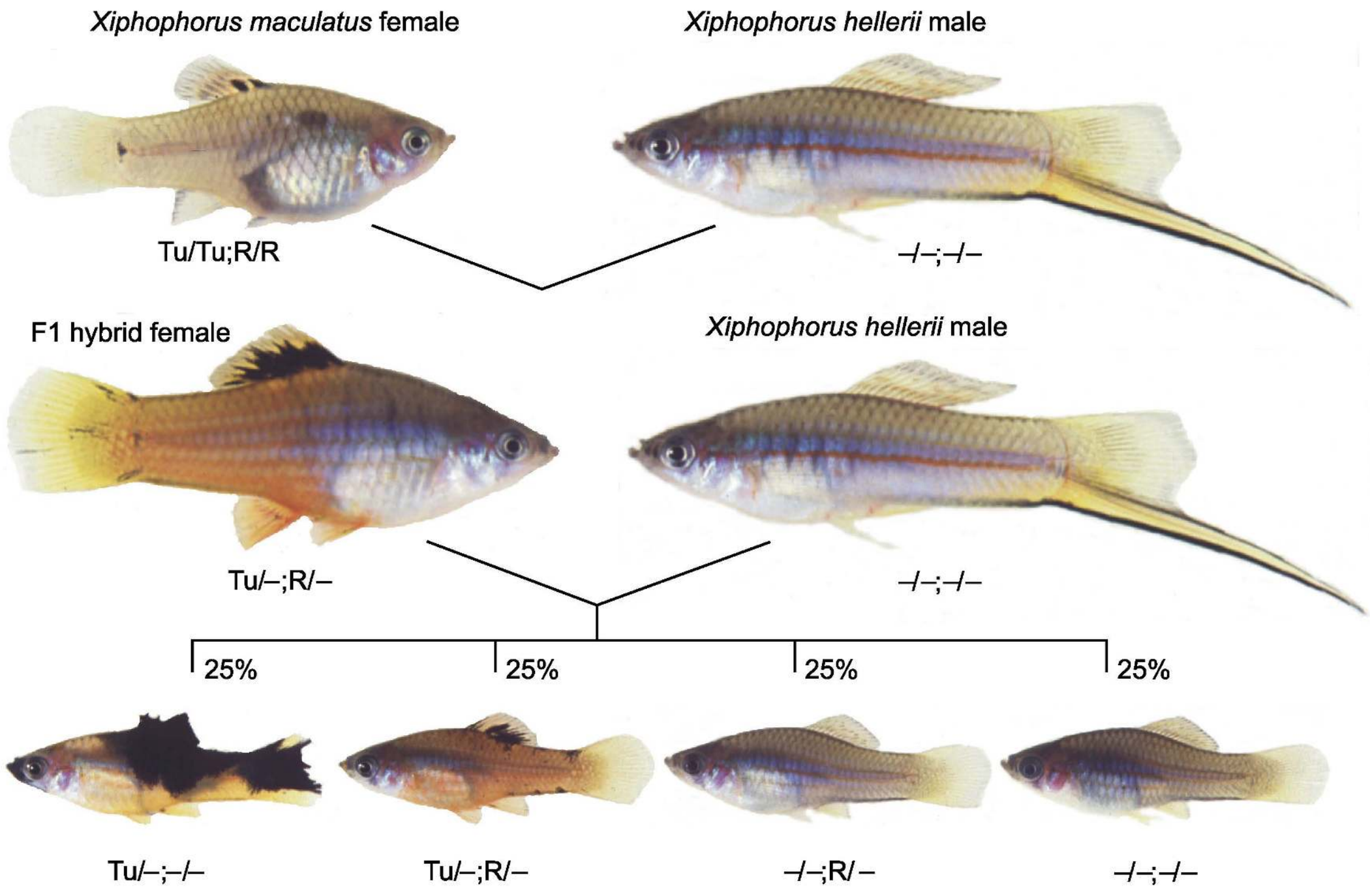


John Postlethwait



Ingo Braasch

Genetic Mapping for Genome Assembly



Version 1.0

290 markers

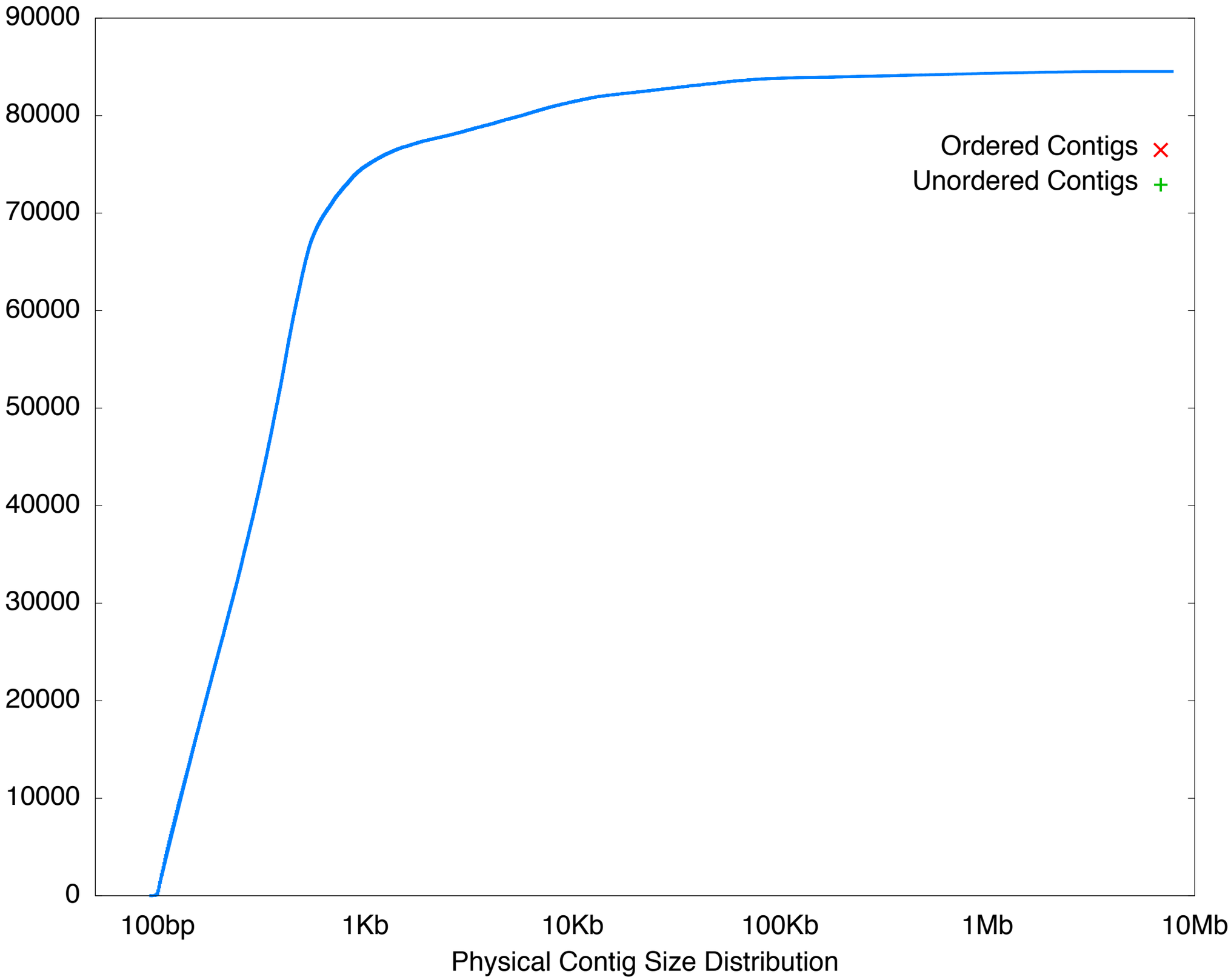
Version 2.0

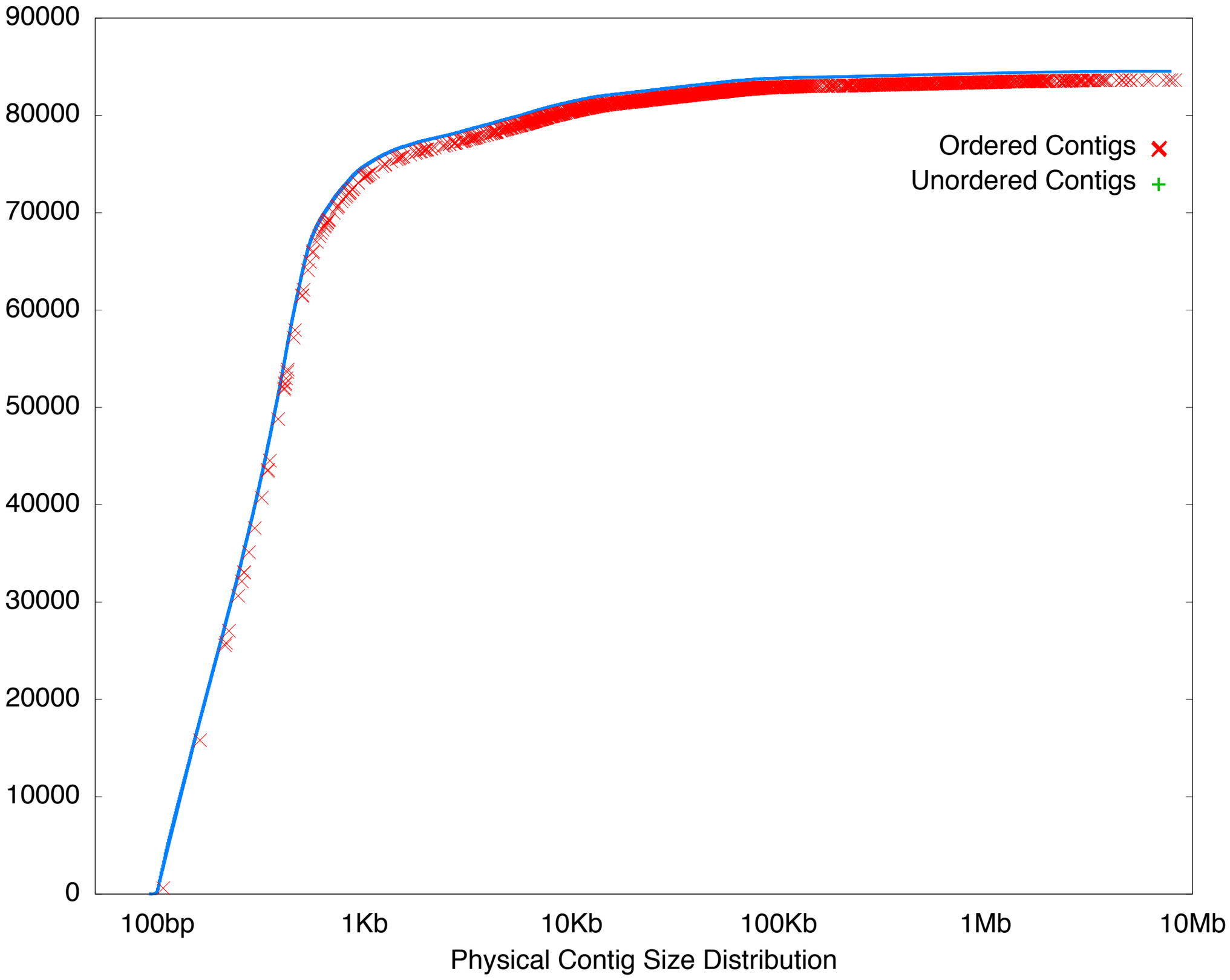
269 total fish

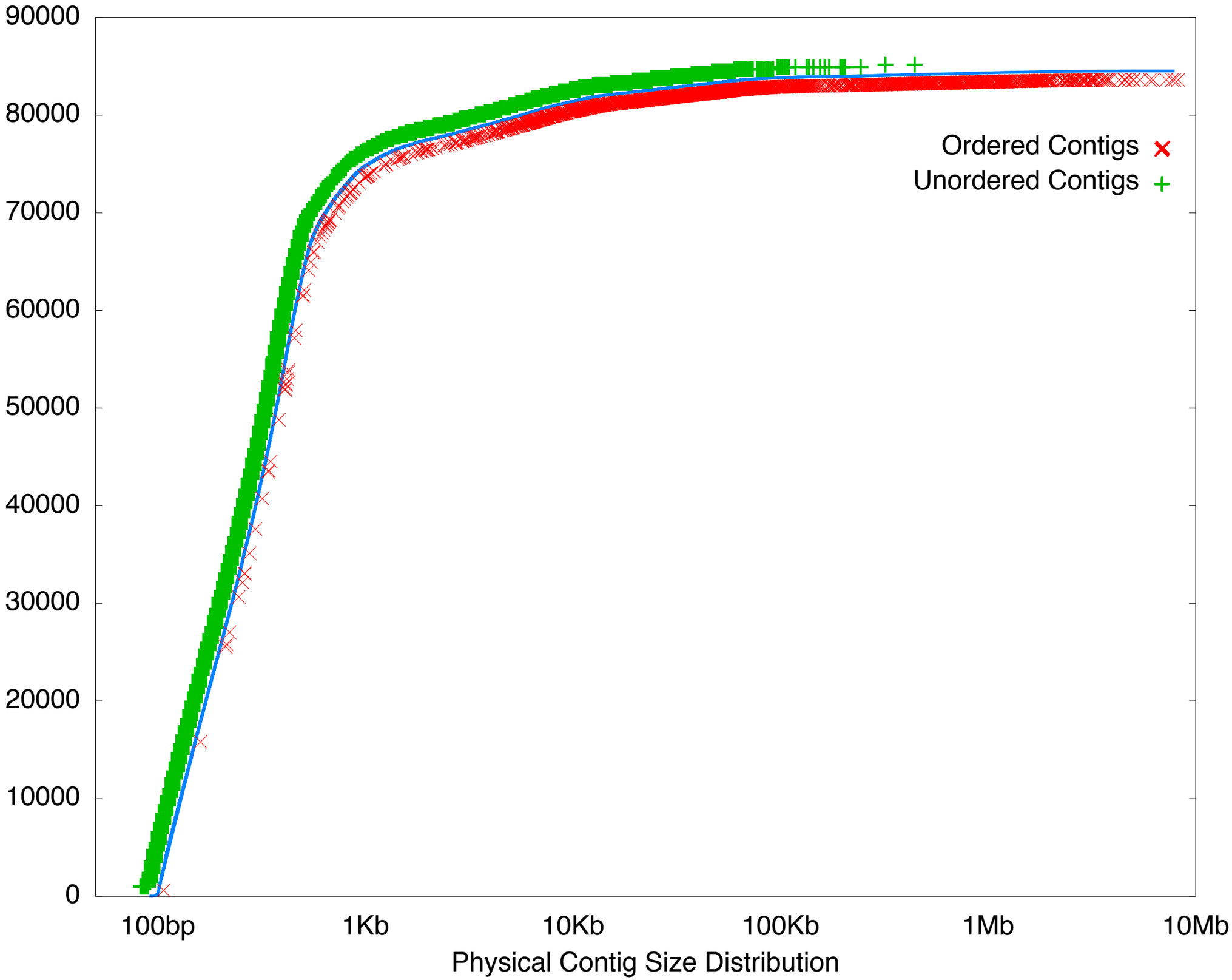
828,373,617 raw reads

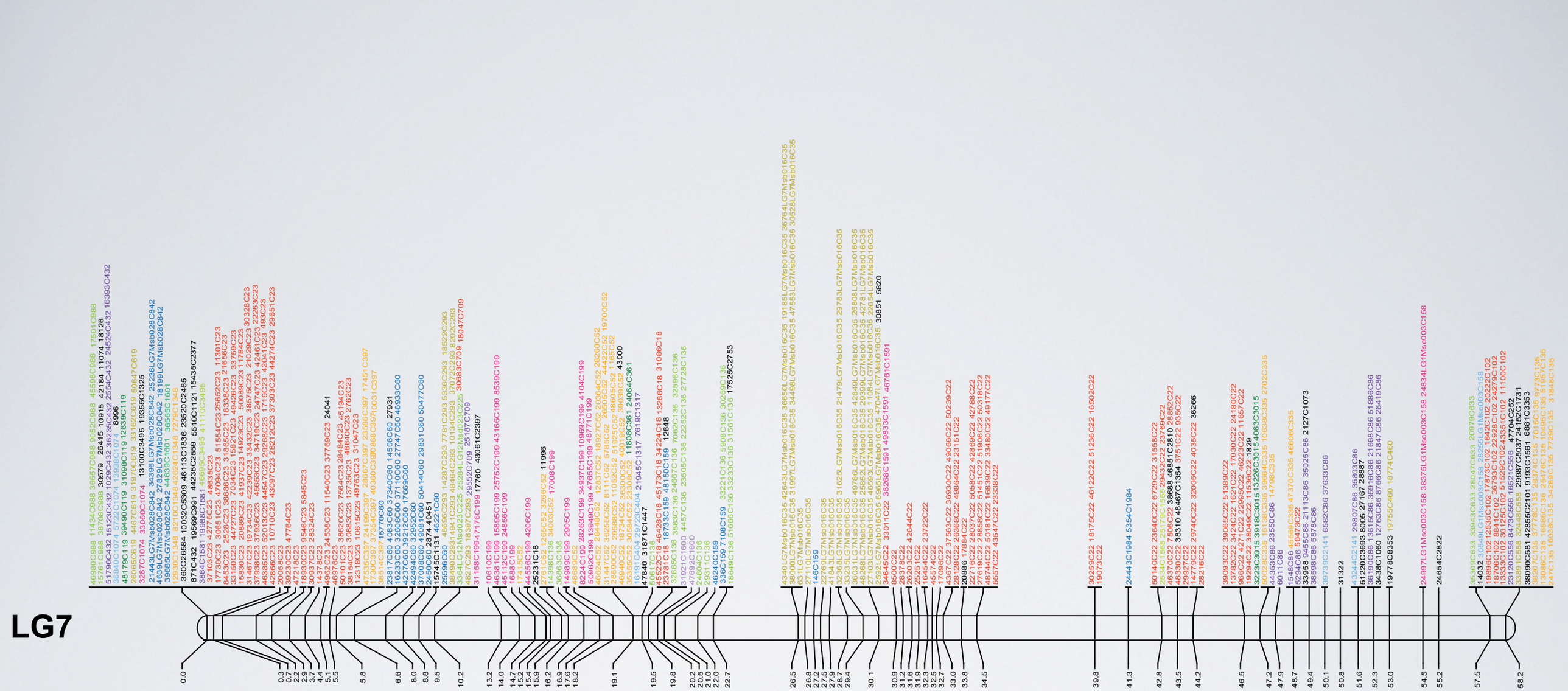
613,162,521 incorporated reads

22,144 markers in at least 90 progeny


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

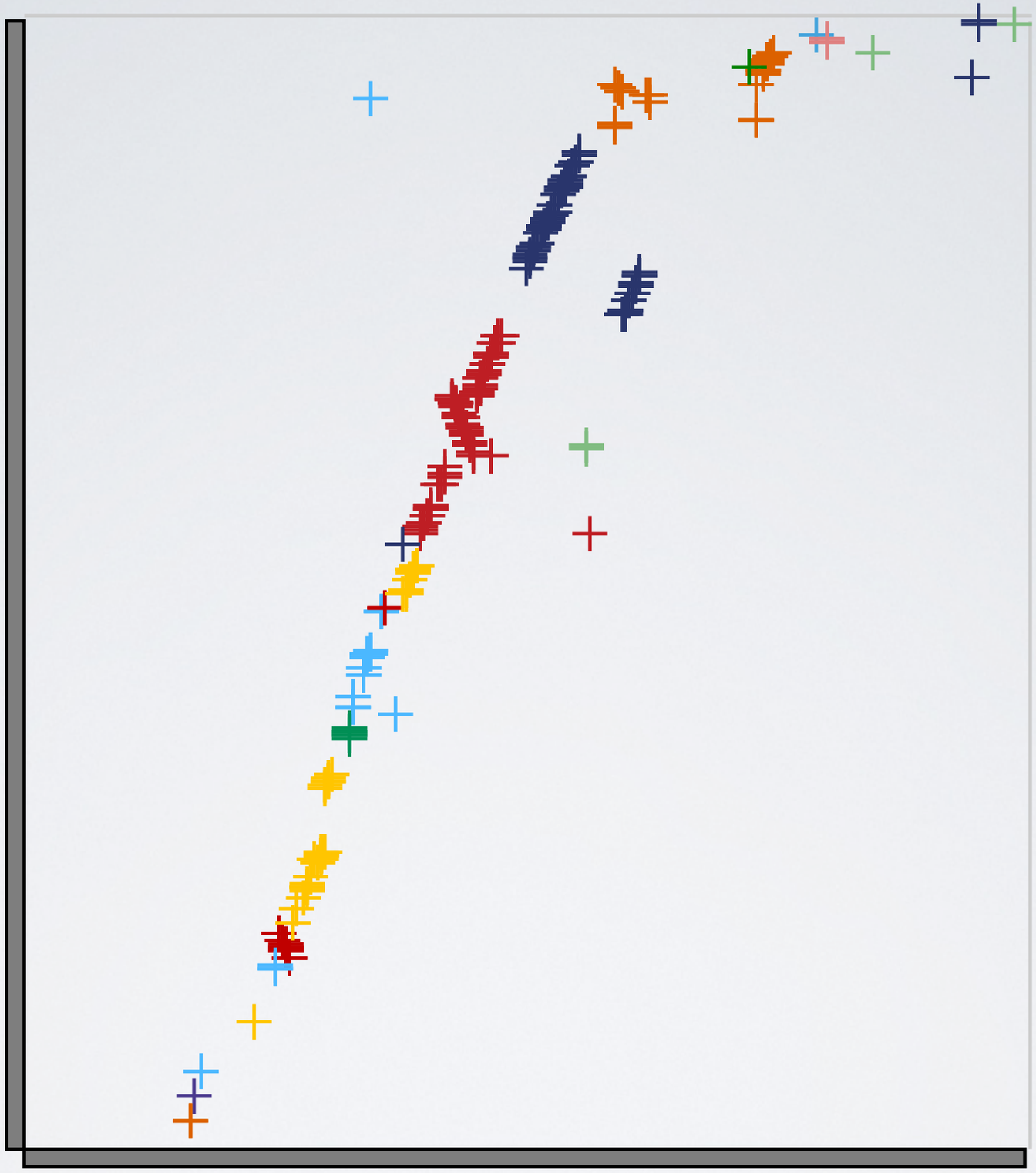

90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Ordered Contigs
Unordered Contigs
100bp
1Kb
10Kb
100Kb
1Mb
10Mb
Physical Contig Size Distribution


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

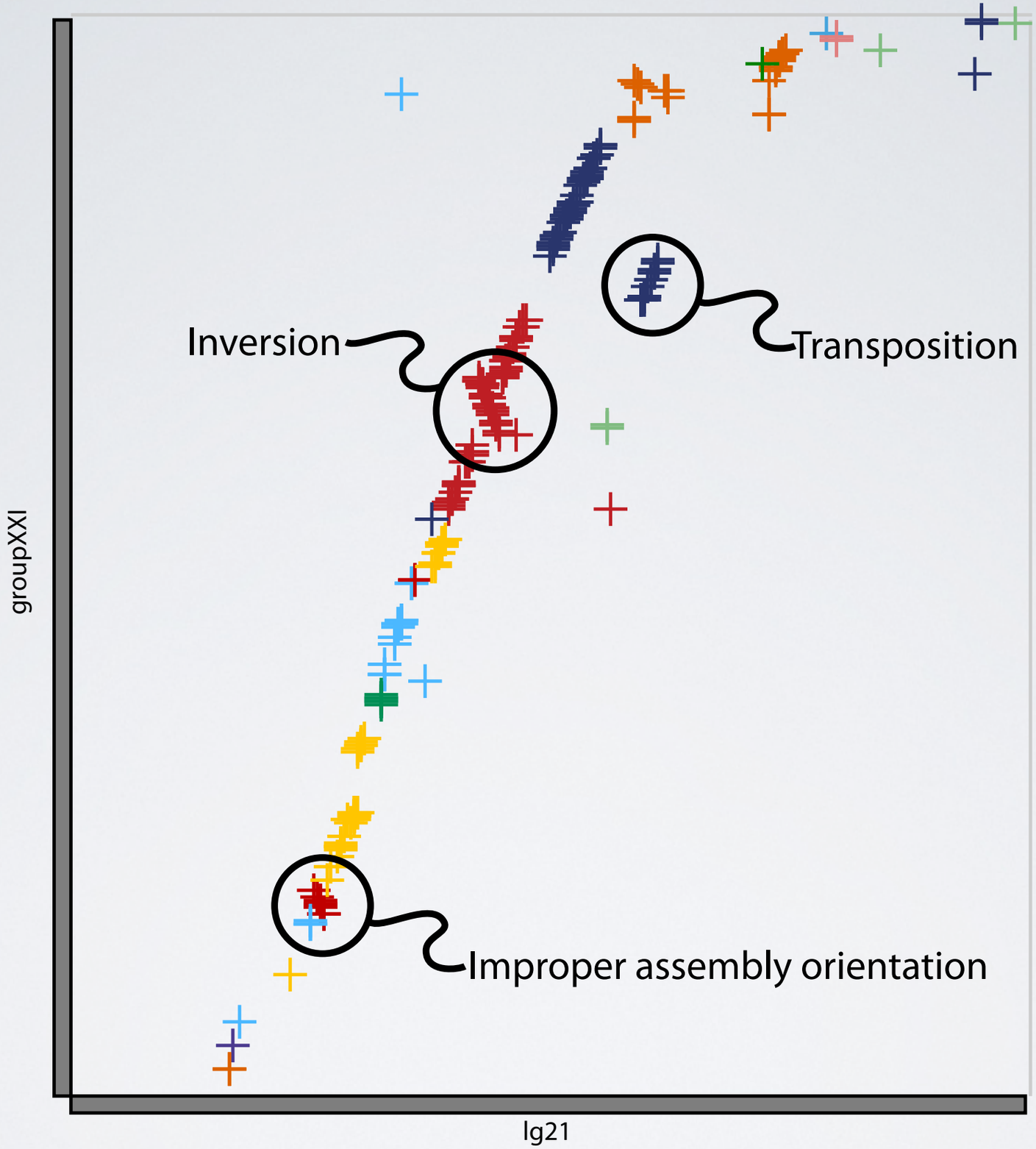


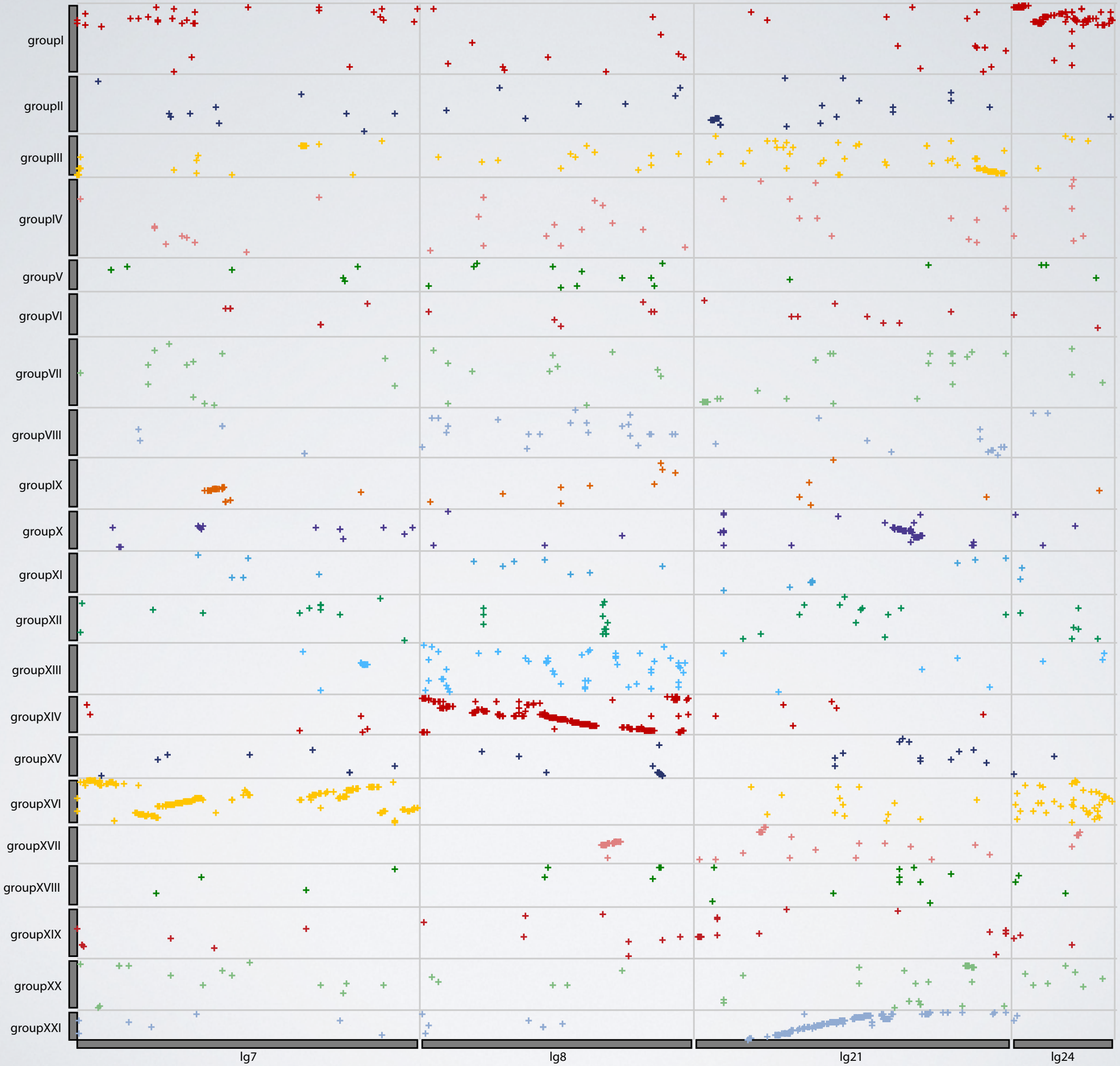
Given 80,000 contigs: 86% of genome ordered

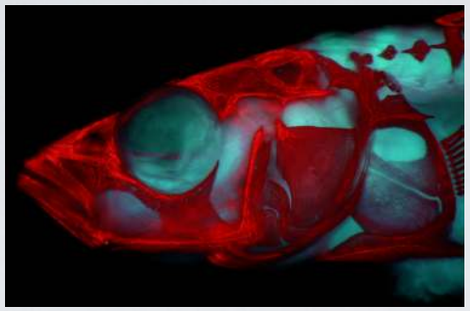

10 mm


groupXXI
lg21

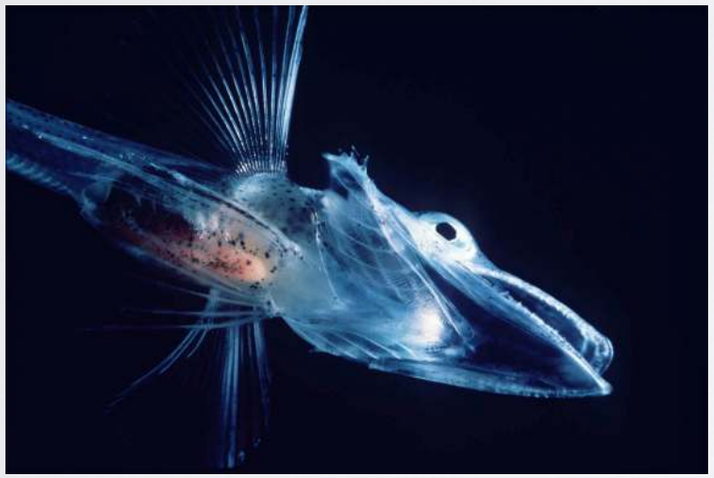

10 mm


groupXXI
Inversion
Transposition
Improper assembly orientation
lg21

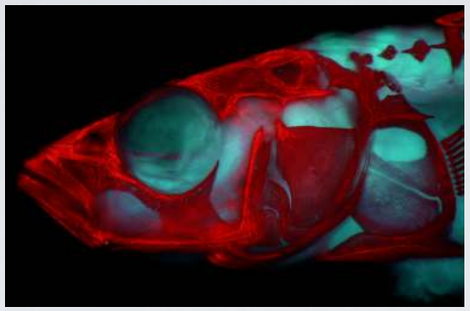

groupI
groupII
groupIII
groupIV
groupV
groupVI
groupVII
groupVIII
groupIX
groupX
groupXI
groupXII
groupXIII
groupXIV
groupXV
groupXVI
groupXVII
groupXVIII
groupXIX
groupXX
groupXXI
lg7
lg8
lg21
lg24

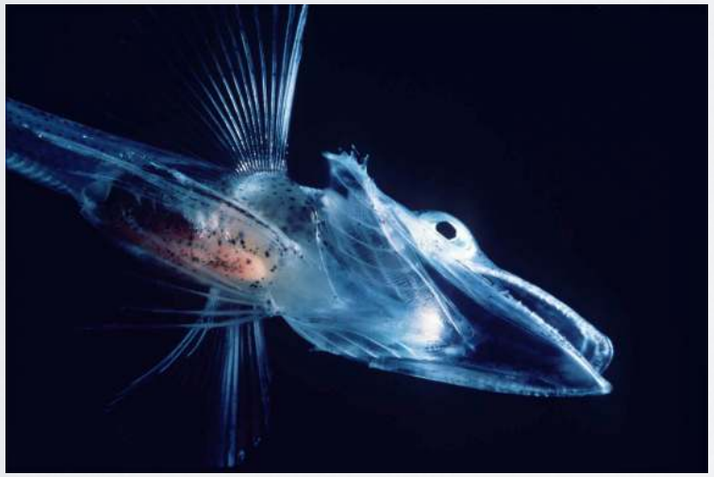

What if you **don't** have a reference genome?





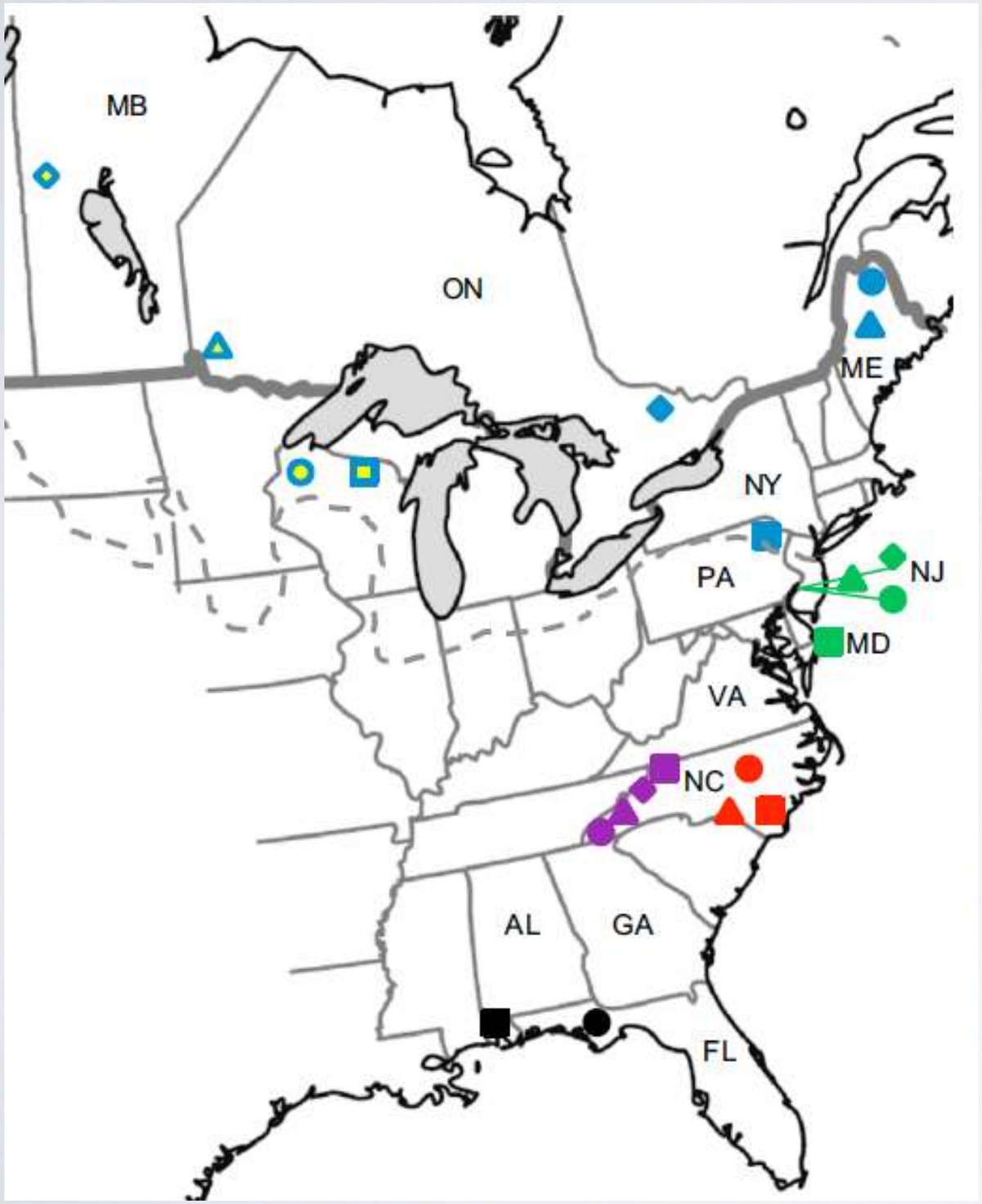




Phylogeography





The pitcher plant mosquito, *Wyeomyia smithii* a model for evolution due to climate change

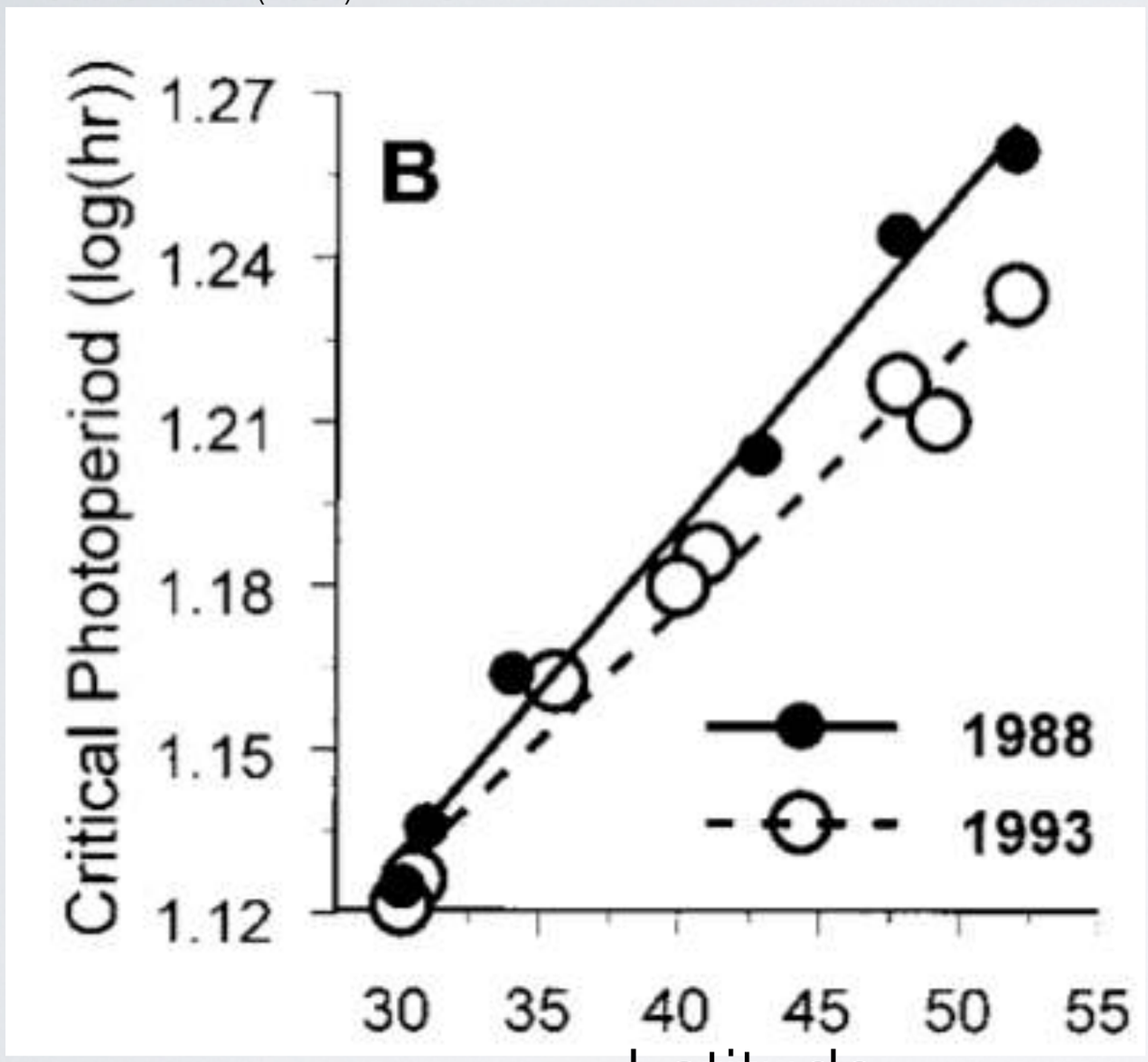


Paul Hohenlohe

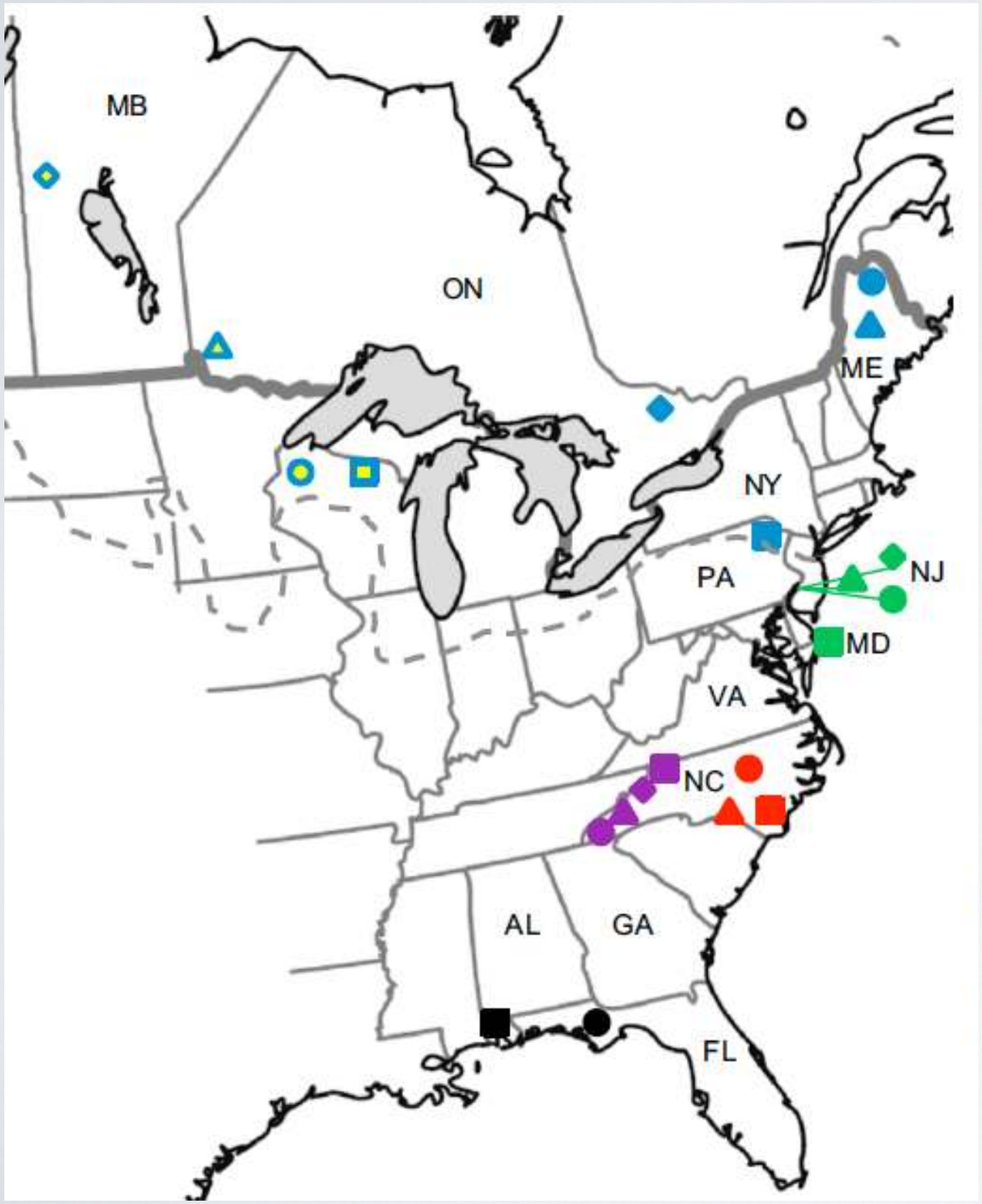
Kevin Emerson

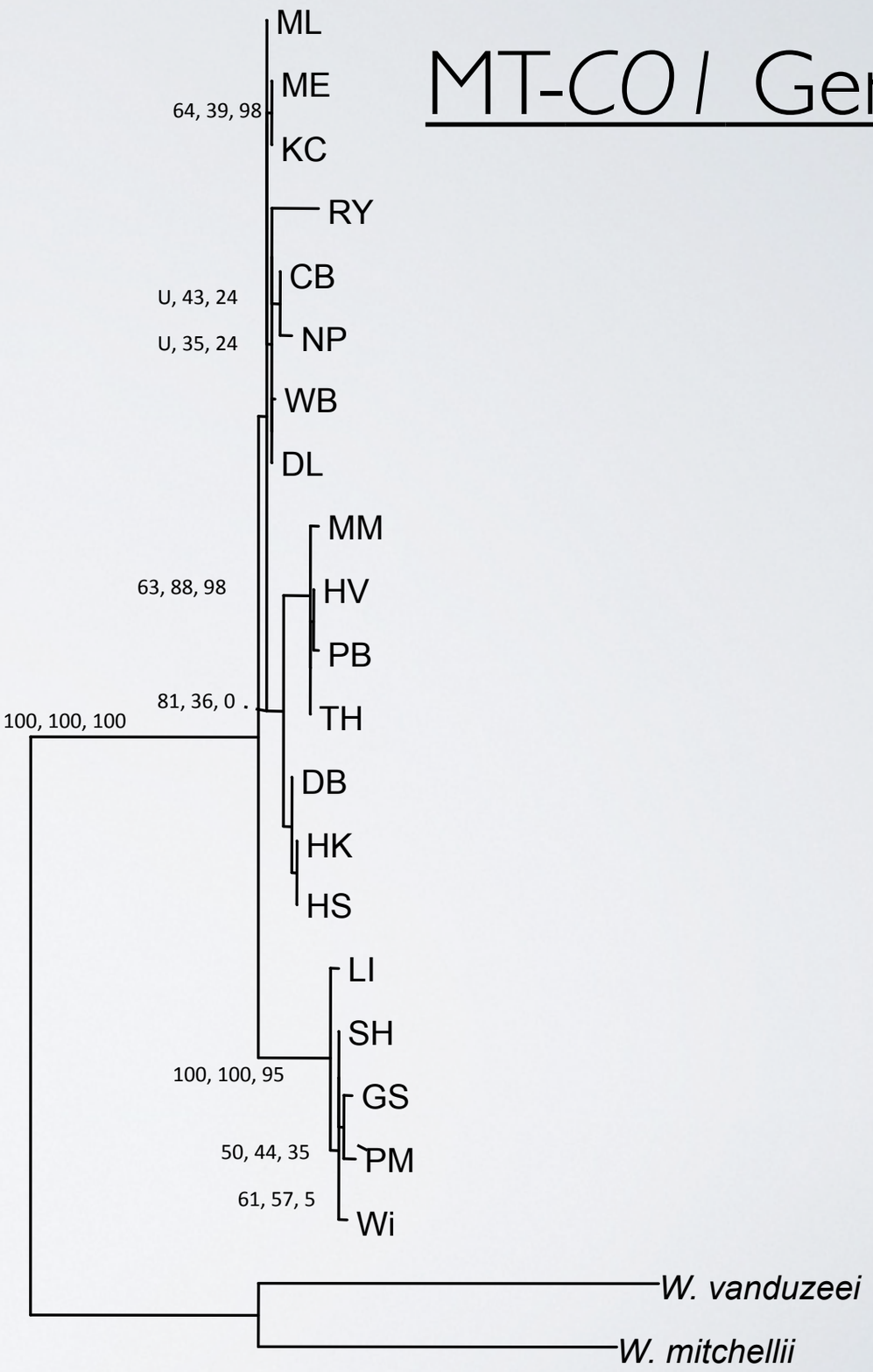
The pitcher plant mosquito, *Wyeomyia smithii* a model for evolution due to climate change

Bradshaw et al (2001) PNAS



Resolving post-glacial phylogeography in *Wyeomyia smithii*



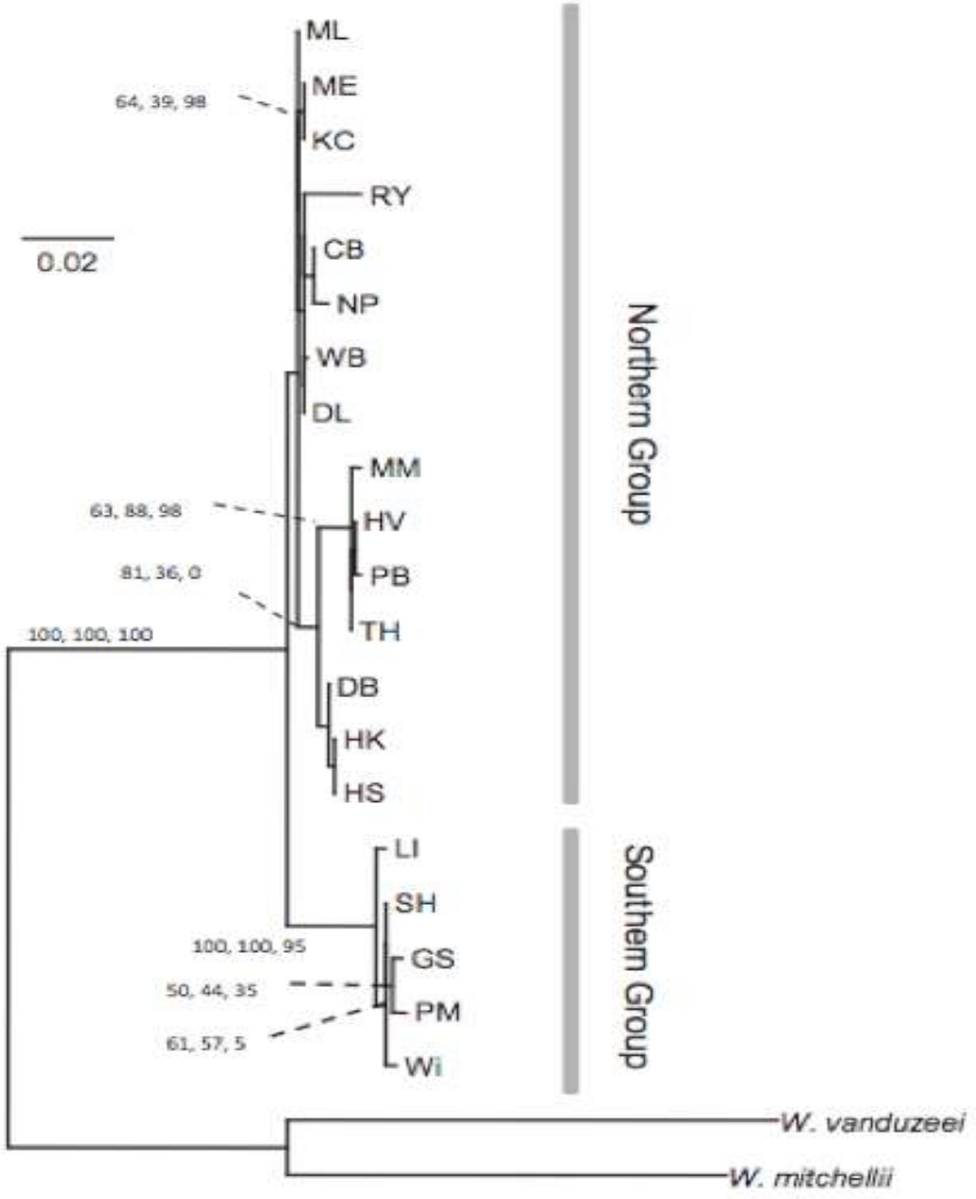


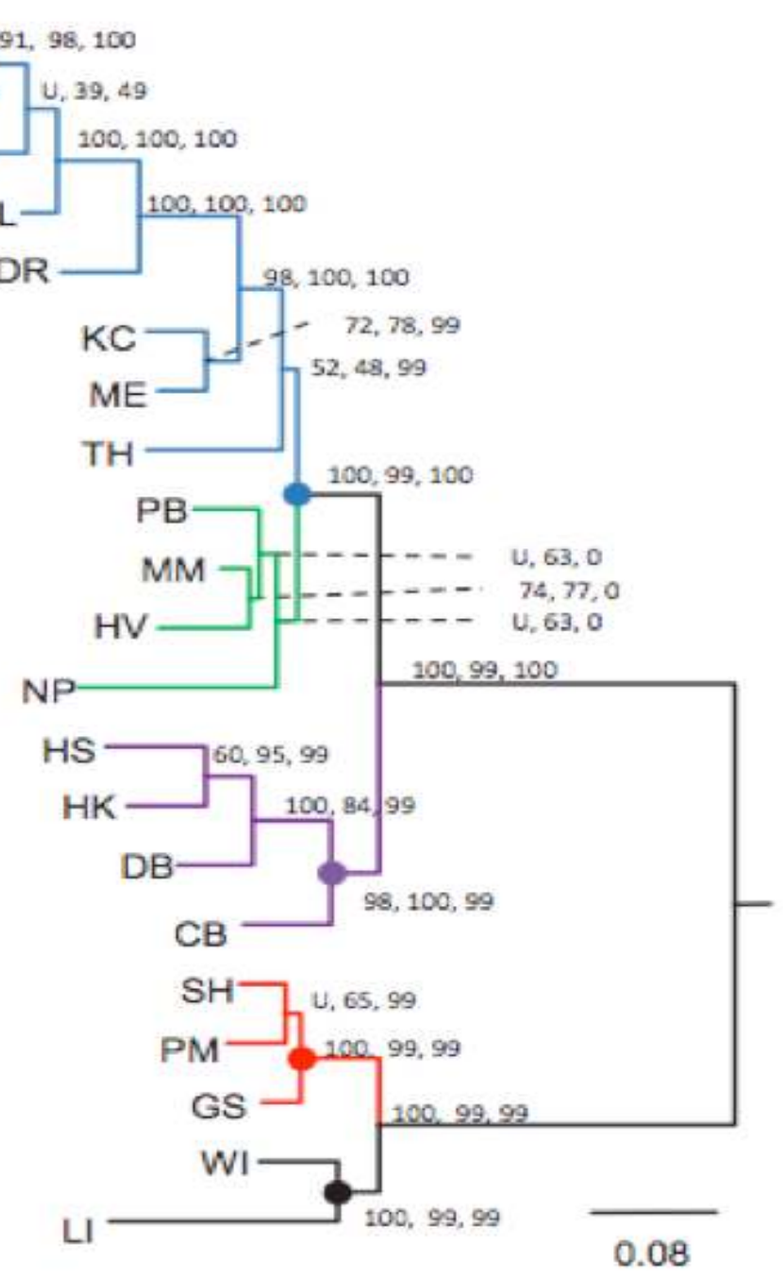
MT-*CO1* Gene

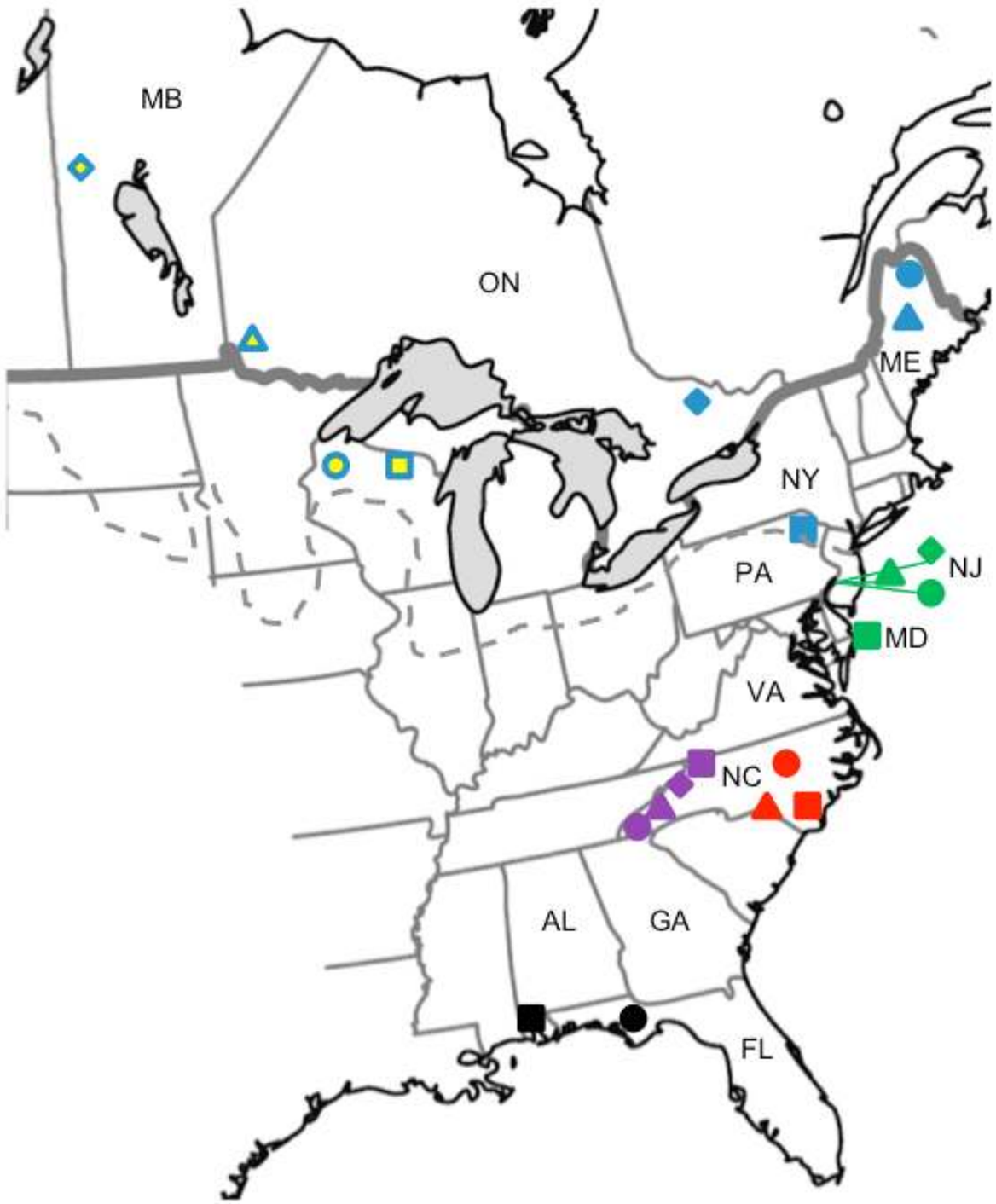
Sampling RAD-Tag markers for population-level phylogenetic analysis

Sampling RAD-Tag markers for population-level phylogenetic analysis

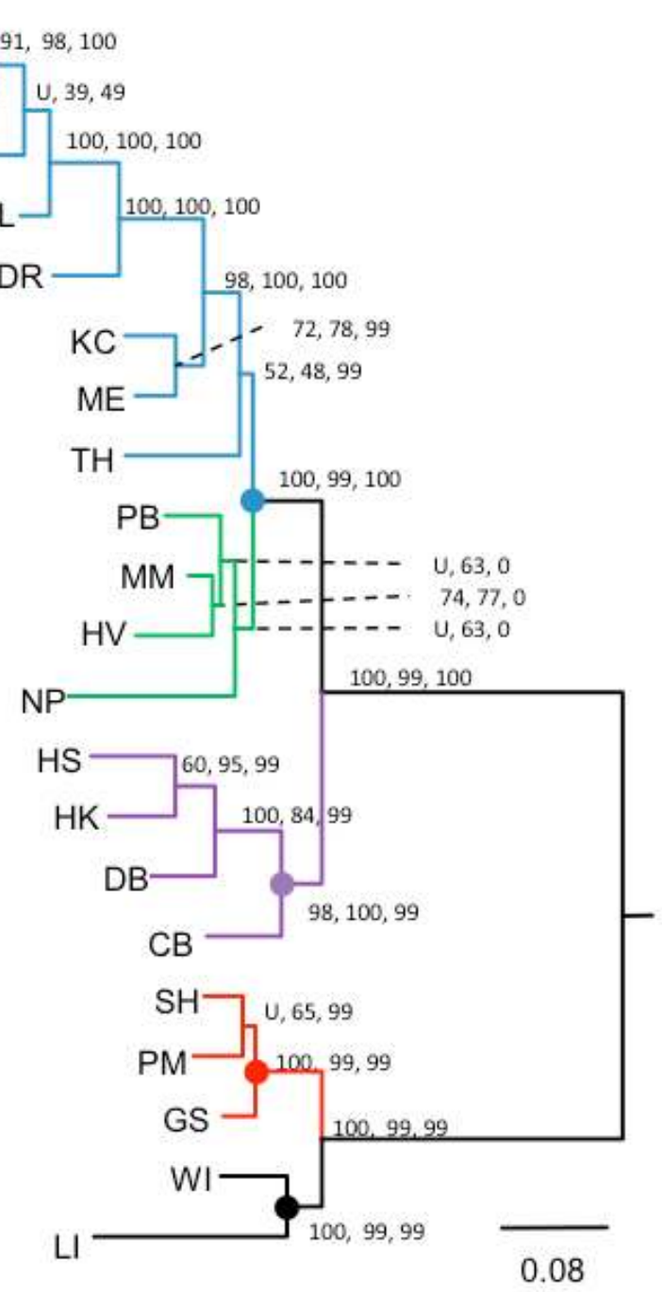
Genome-wide view improves resolution






MB
ON
ME
NY
PA
NJ
MD
VA
NC
AL
GA
FL


Northern
Appalachian
NC Coast
Gulf Coast


WB
DL
RY
ML
DR
KC
ME
TH
PB
MM
HV
NP
HS
HK
DB
CB
SH
PM
GS
WI
LI
91, 98, 100
U, 39, 49
100, 100, 100
100, 100, 100
98, 100, 100
72, 78, 99
52, 48, 99
100, 99, 100
U, 63, 0
74, 77, 0
U, 63, 0
100, 99, 100
60, 95, 99
100, 84, 99
98, 100, 99
U, 65, 99
100, 99, 99
100, 99, 99
100, 99, 99
0.08

SPECIAL THANKS TO ....

- Cresko Lab members Susan Bassham, Paul Hohenlohe, and Kristin Sikkink

- Collaborators Angel Amores, John Postlethwait, Kevin Emerson, Paul Etter, William Bradshaw, Christina Holzapfel, Eric Johnson, Patrick Phillips

- Funding from NSF & NIH, as well as Keck & Murdock Foundations

NATIONAL INSTITUTES OF HEALTH

NSF