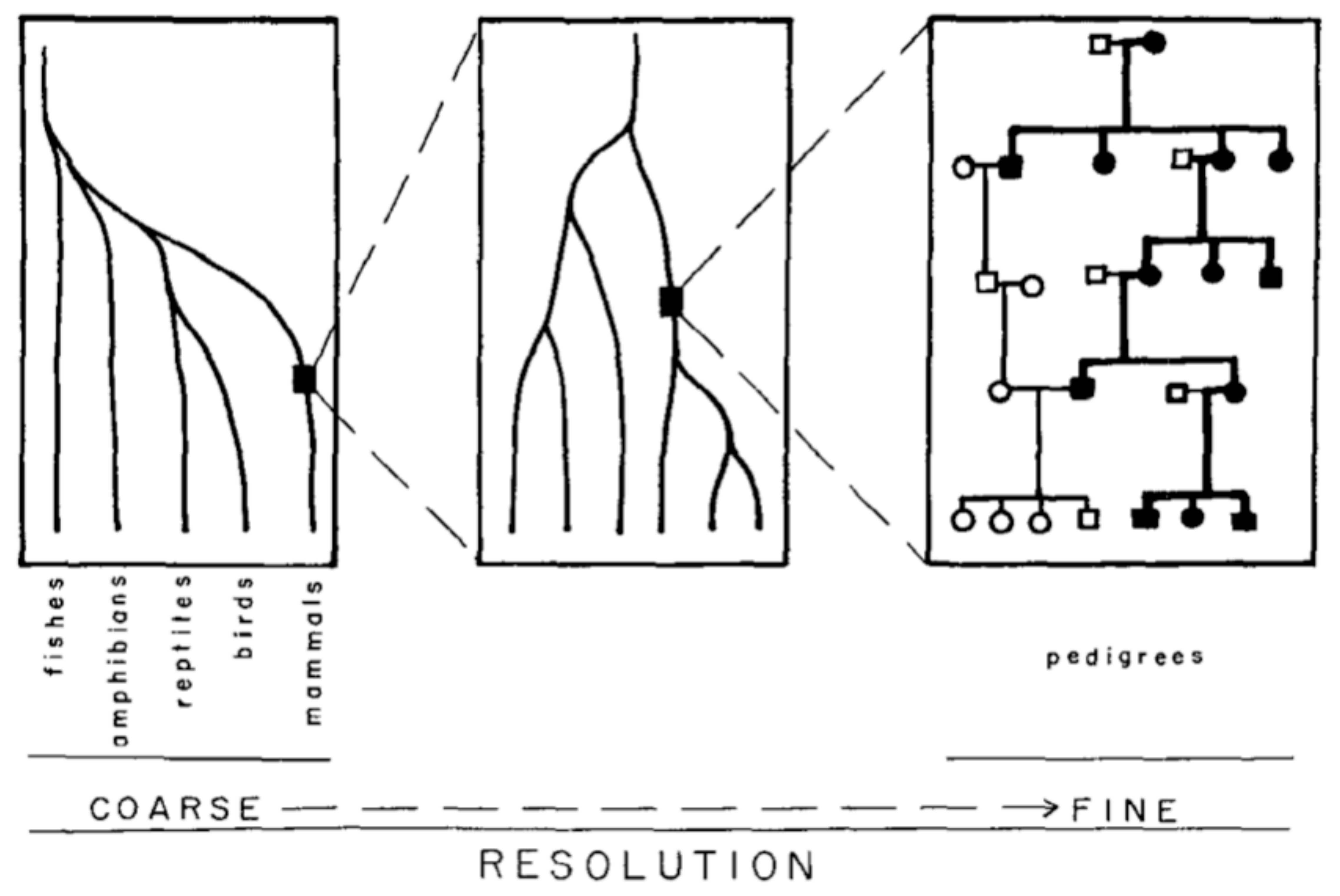
Genomic analyses of non-model organisms with RAD-seq and Stacks

Julian Catchen
jcatchen@uoregon.edu

MDCCCLXXVI

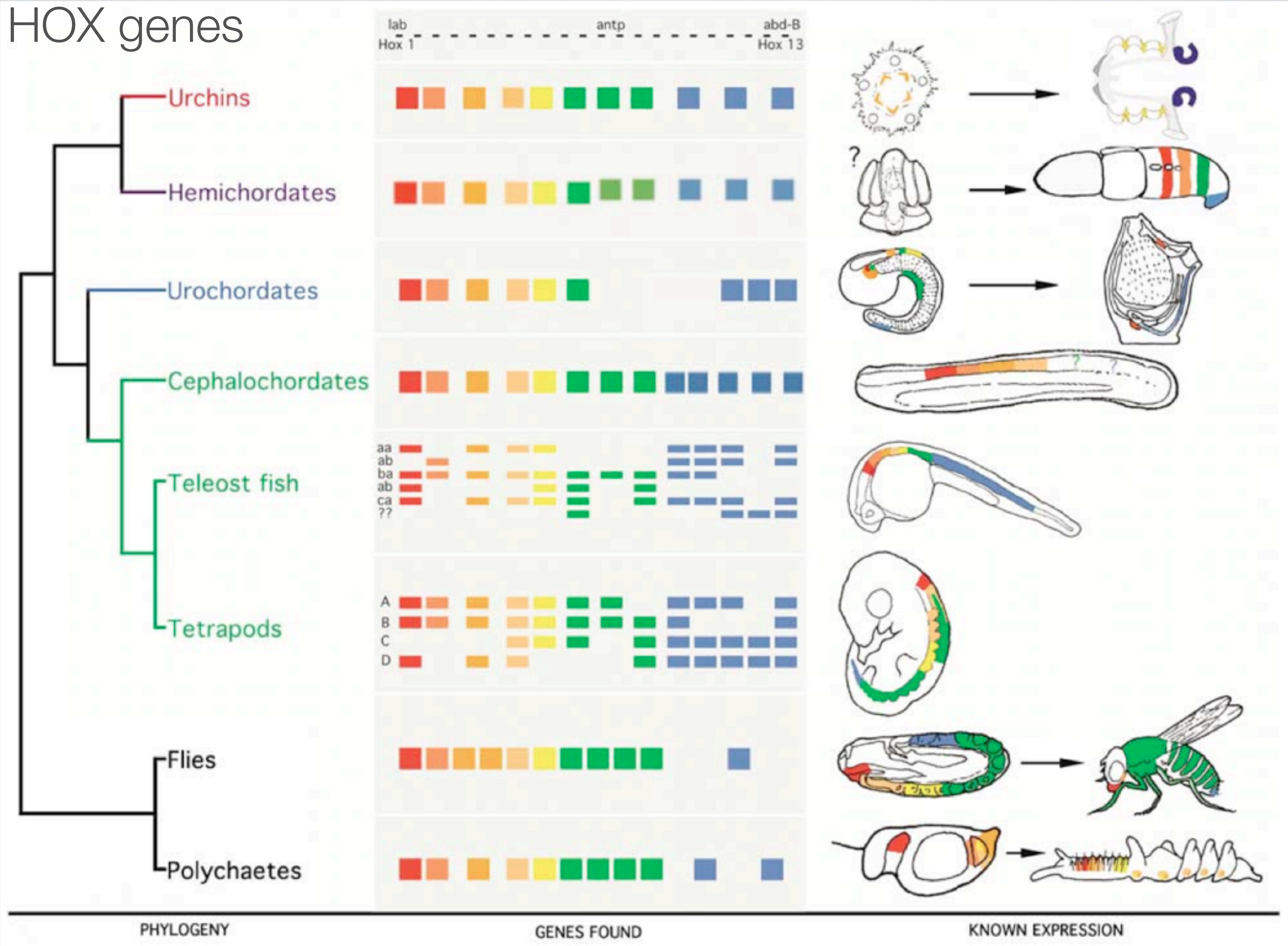
Cresko Lab
Institute of Ecology and Evolution
University of Oregon

"A bridge between population genetics and systematics"



Avise, et al, 1987

A common genetic toolkit

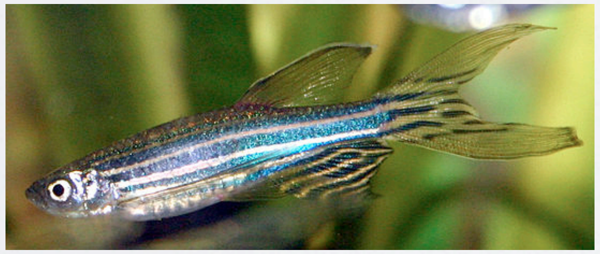


Swalla, 2006

What makes an organism a model?

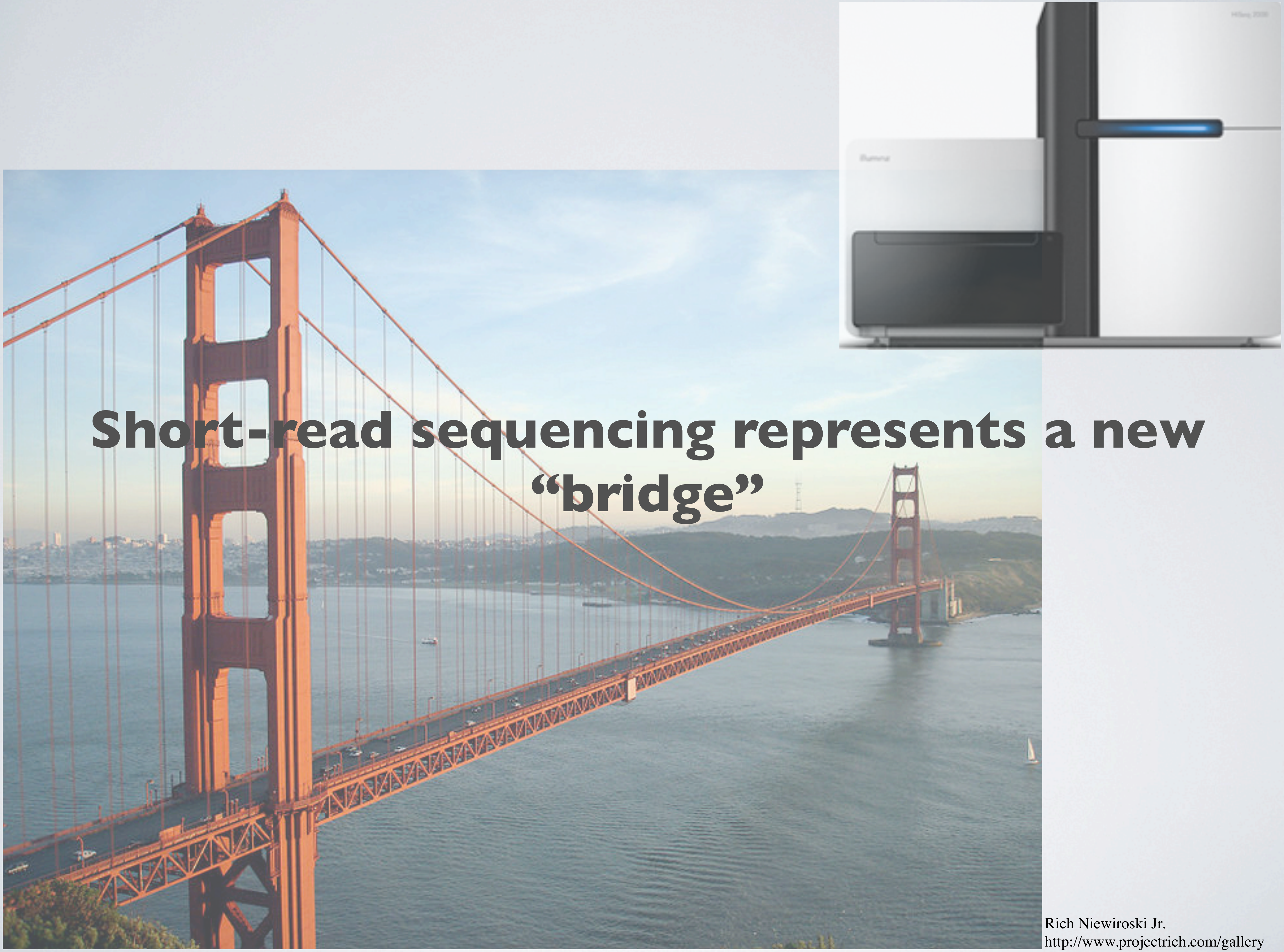
classic models: fruit fly, mouse, zebrafish

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?

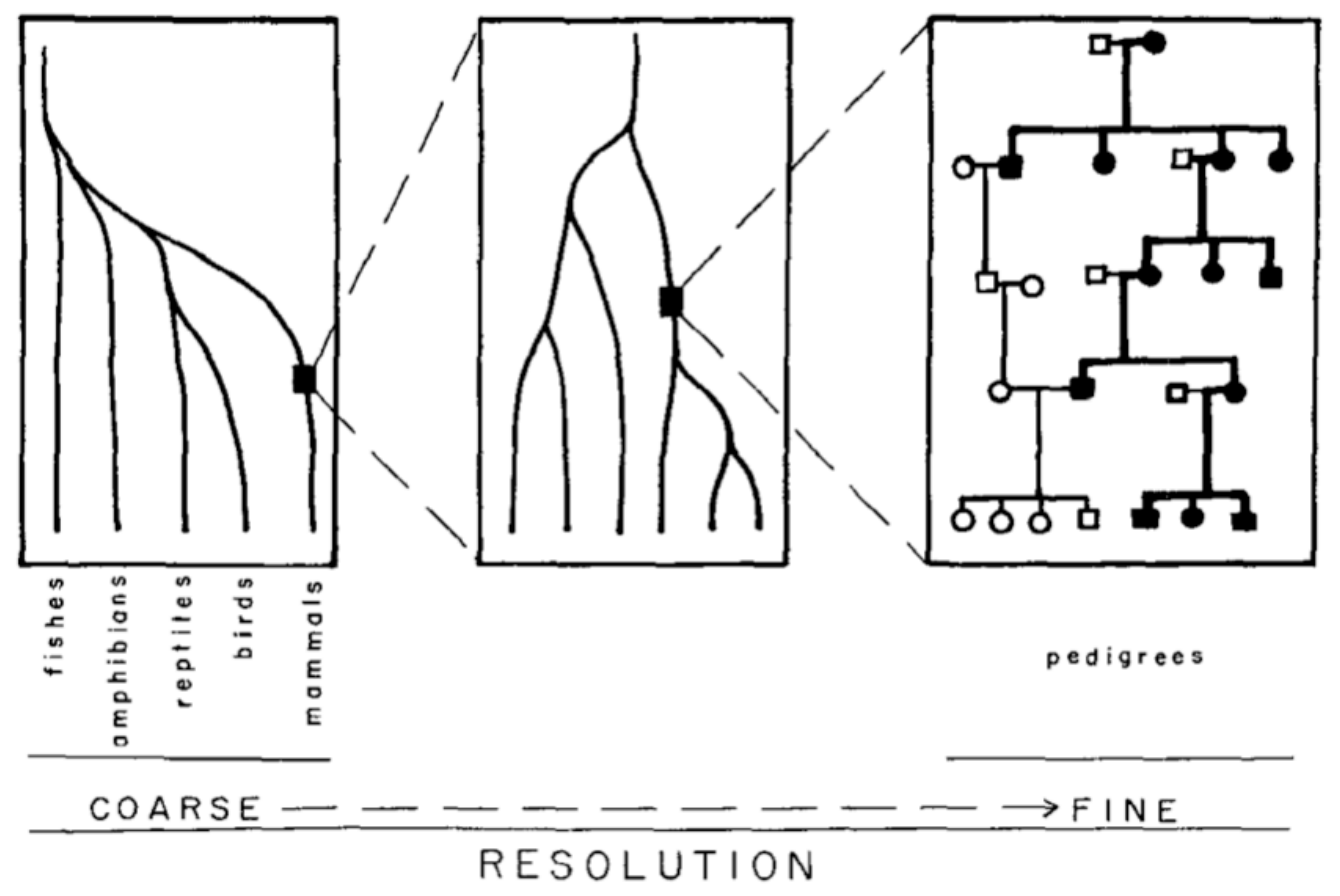
Organism	Markers	Organism	Markers
Bighorn Sheep	147	Silver carp	483
Shrimp	418	Guppy	790
Coral	420	Barramundi	240
Indian mustard	1,029	Catfish	331
Oilseed rape	13,551	Sea bass	368
Black spruce	1,111	Cichlid	204
Barley	2,111	Platyfish	290
Flounder	1,375	Halibut	604
Turbot	242	Sea bream	204
Human	**250,000**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



Short-read sequencing represents a new "bridge"

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

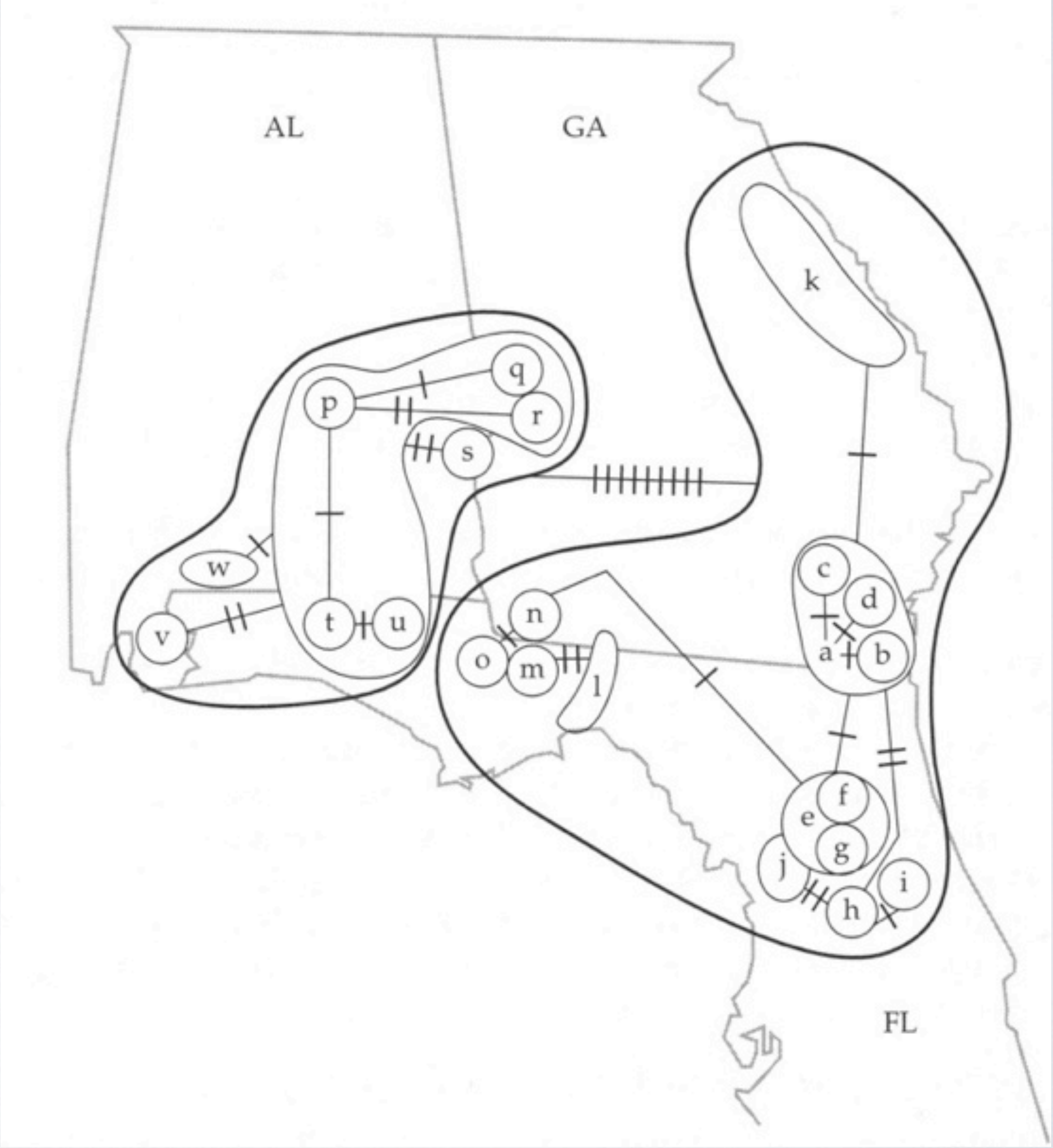
“A bridge between population genetics and systematics”



Avise, et al, 1987

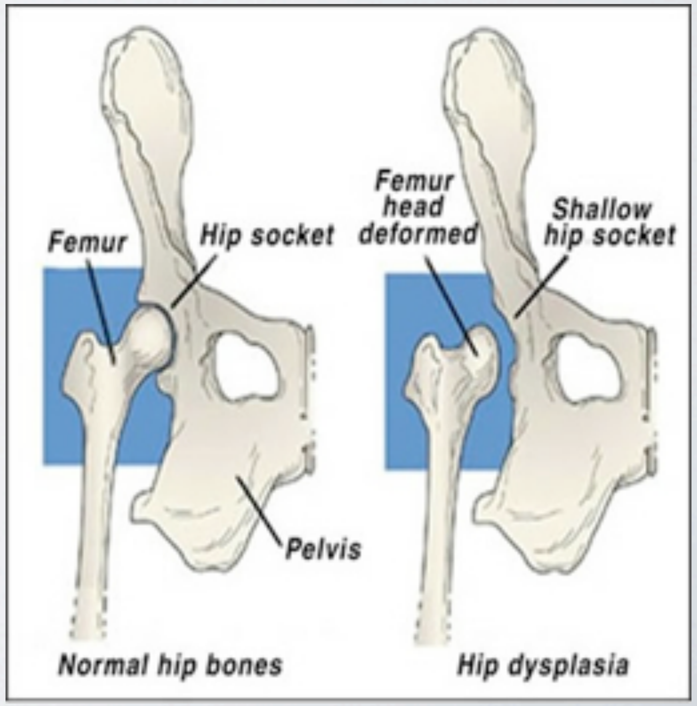
Fine scale estimates of population structure
Estimates of genome-region specific gene flow

Phylogeography of pocket gophers using mtDNA

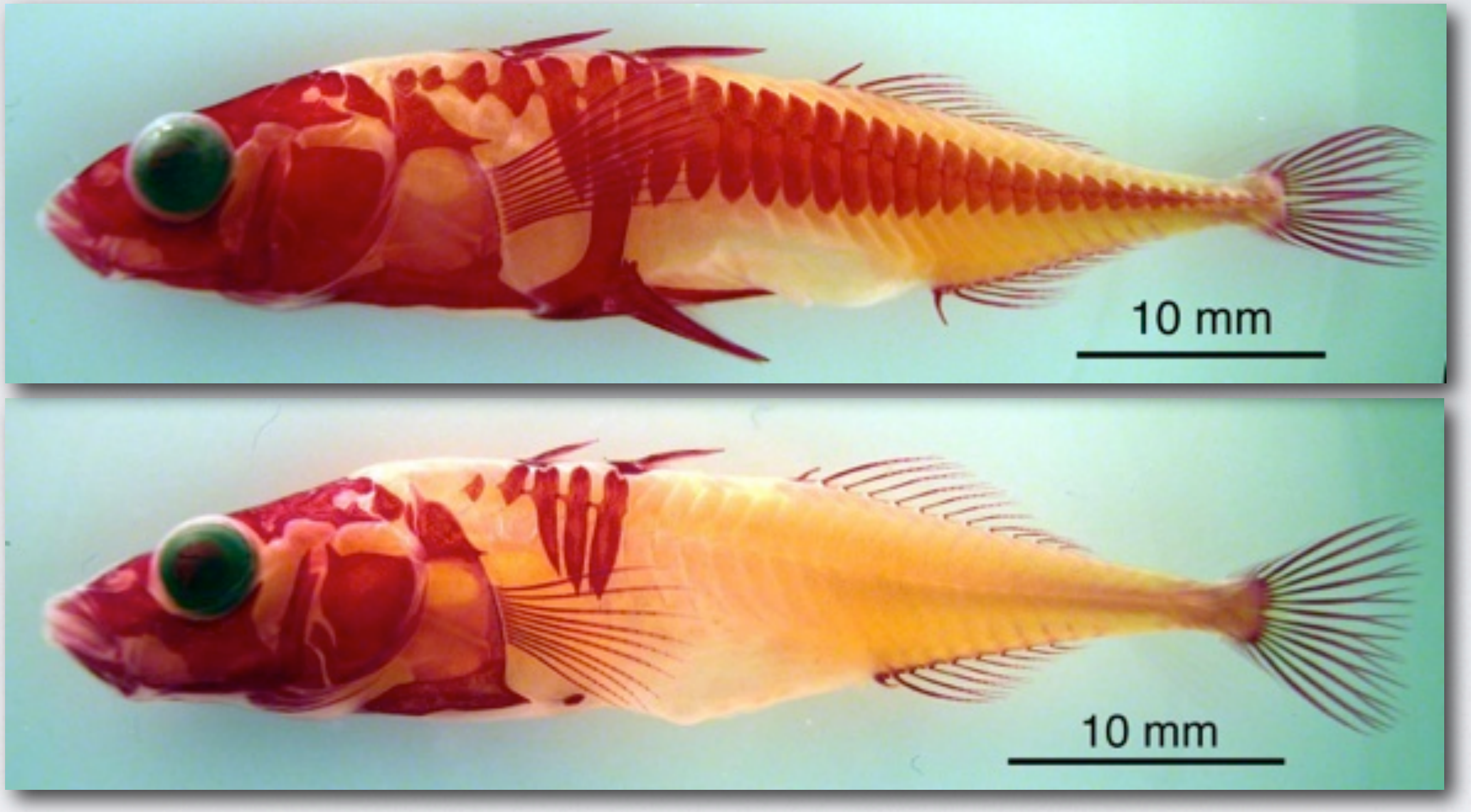


Avise, 1979

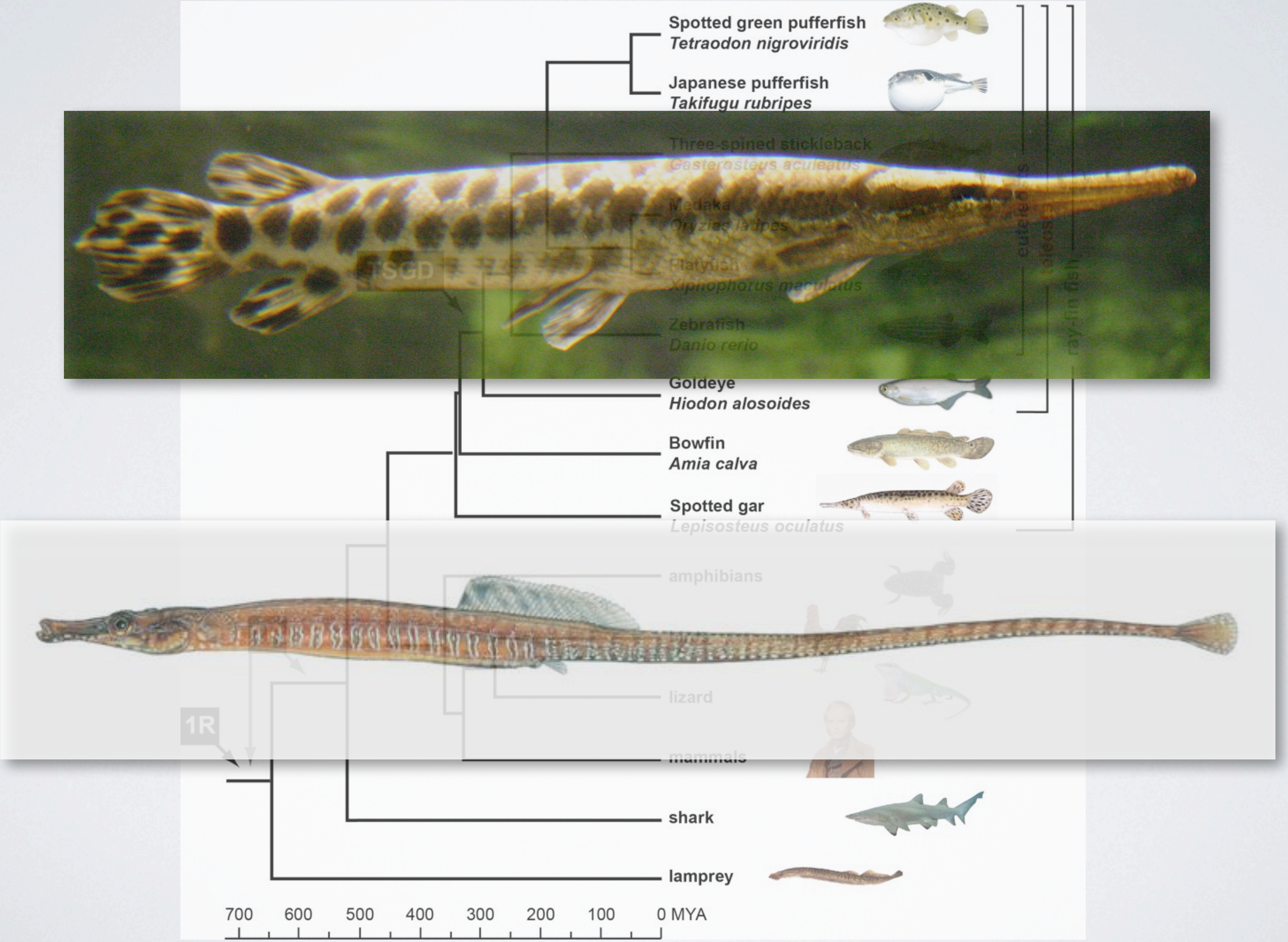
Precisely quantify the amount of inbreeding in wild and captive populations



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

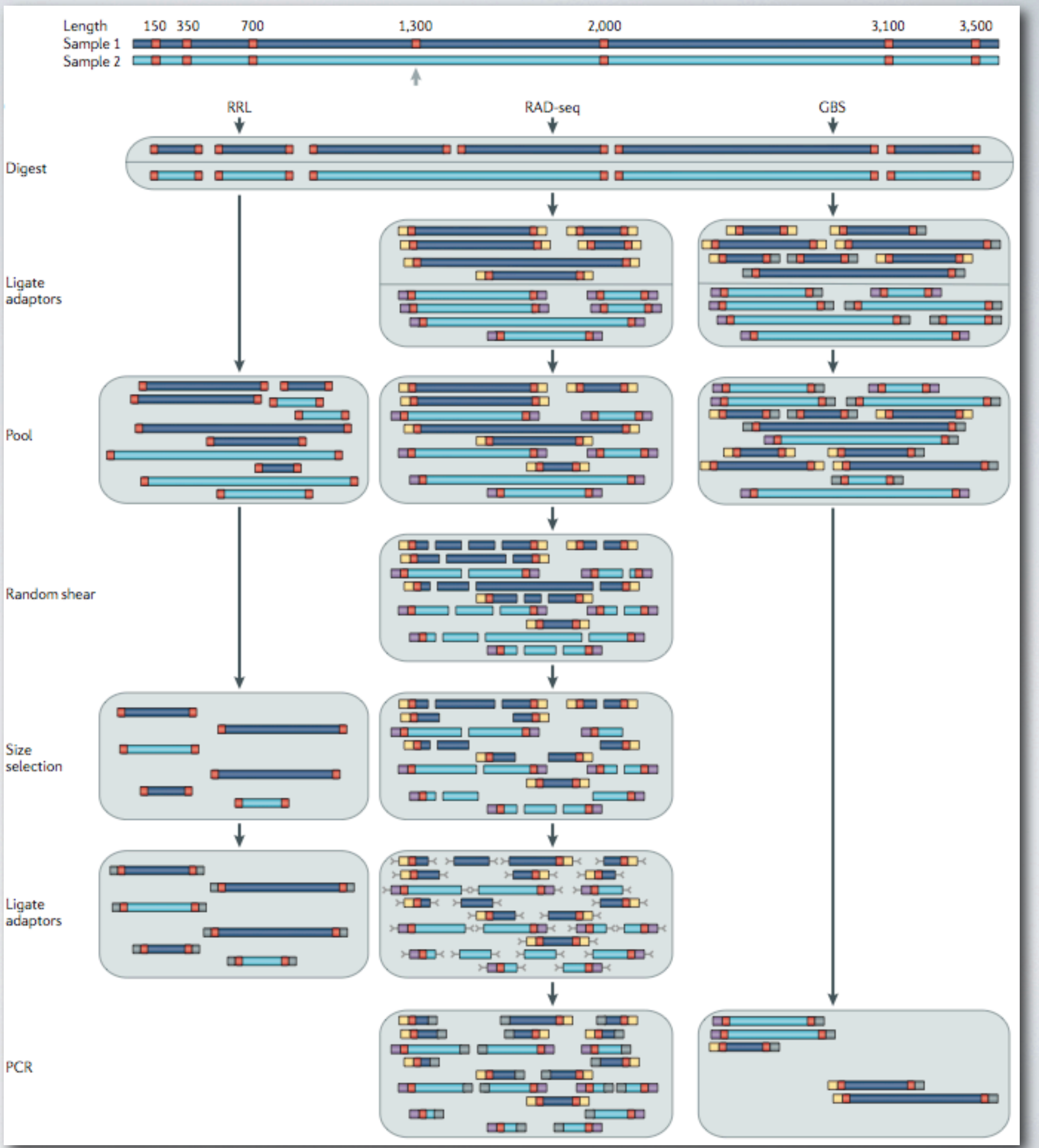
- Can allow thousands of genomes to be assayed in just a few weeks

- 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

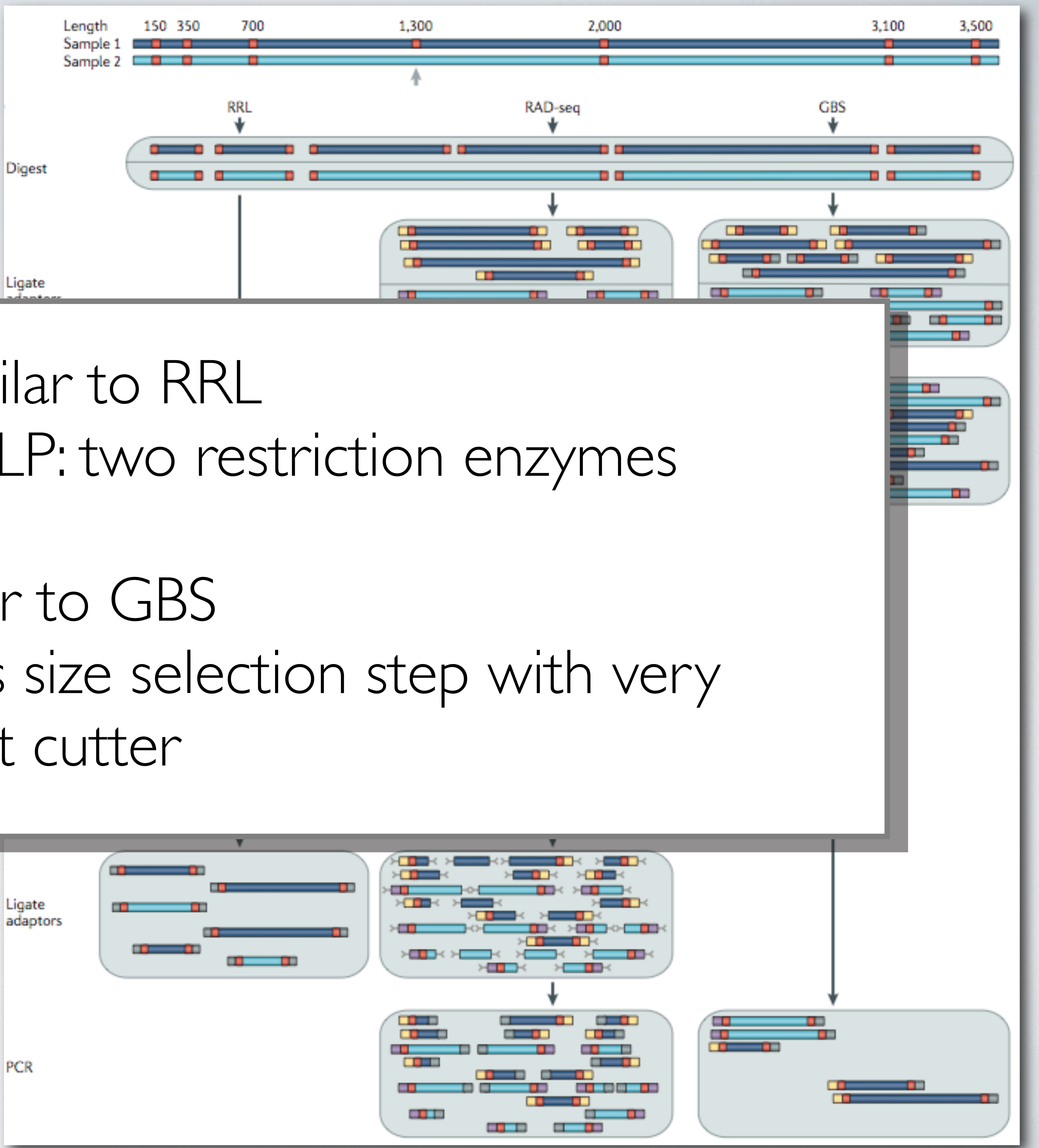


Nature Reviews Genetics, 2011

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

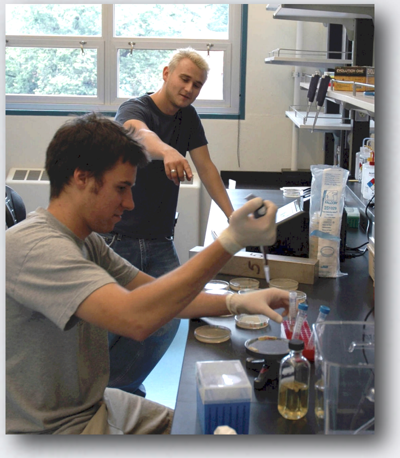
John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone||, Julian M. Catchen‡ and Mark L. Blaxter*¶*

What is RAD-seq?

(Restriction-site Associated DNA)

Eric Johnson

Joe Dunham



Mike Miller

Illumina

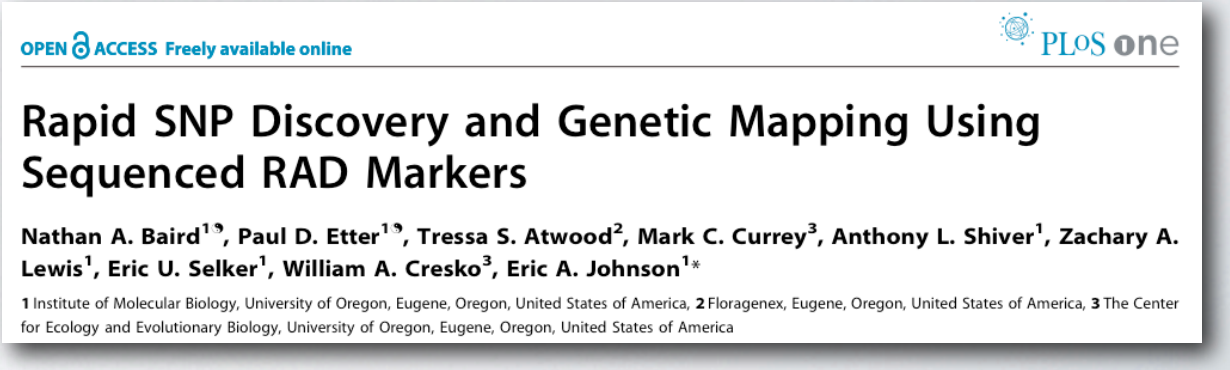
2007

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

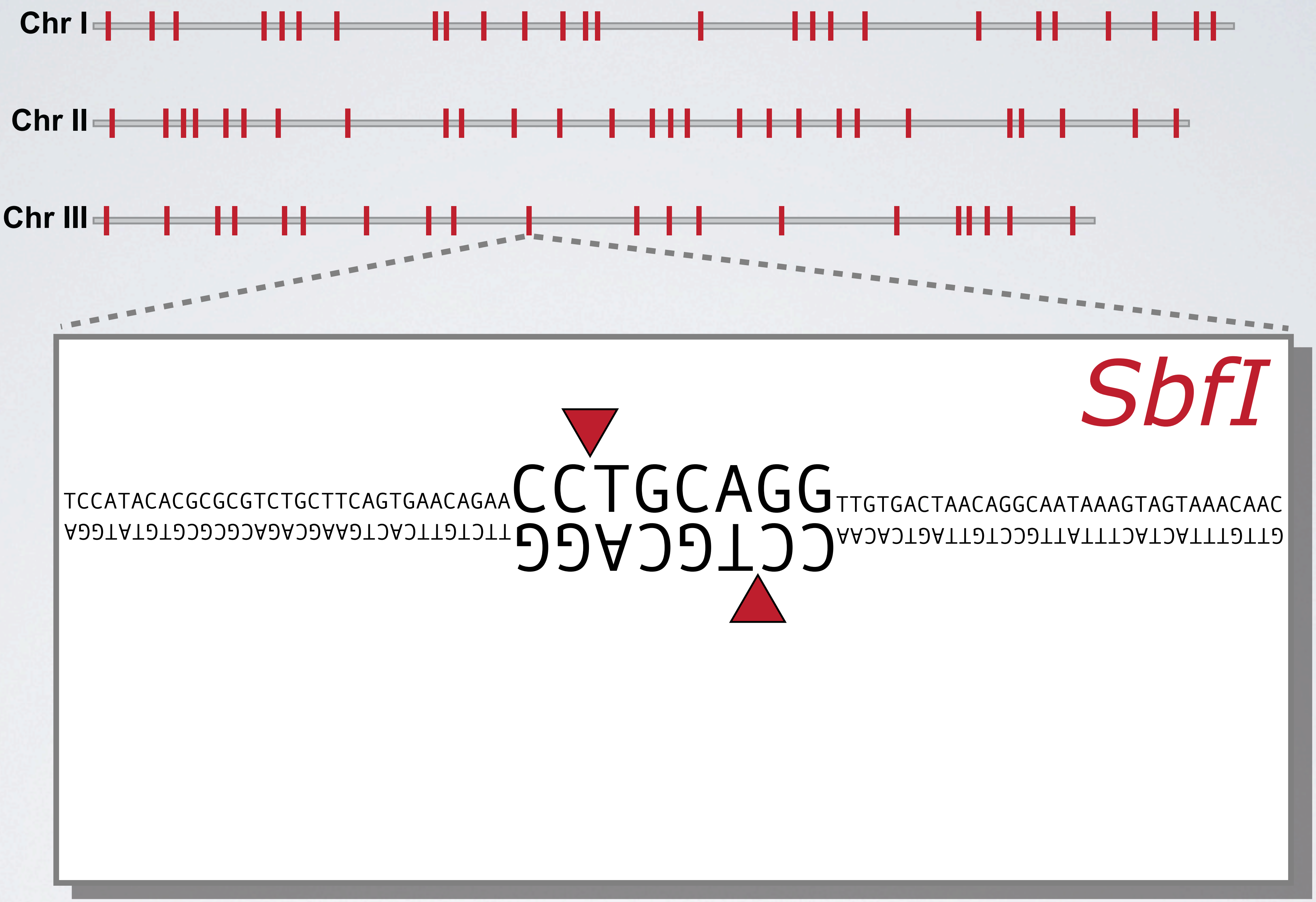
Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

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

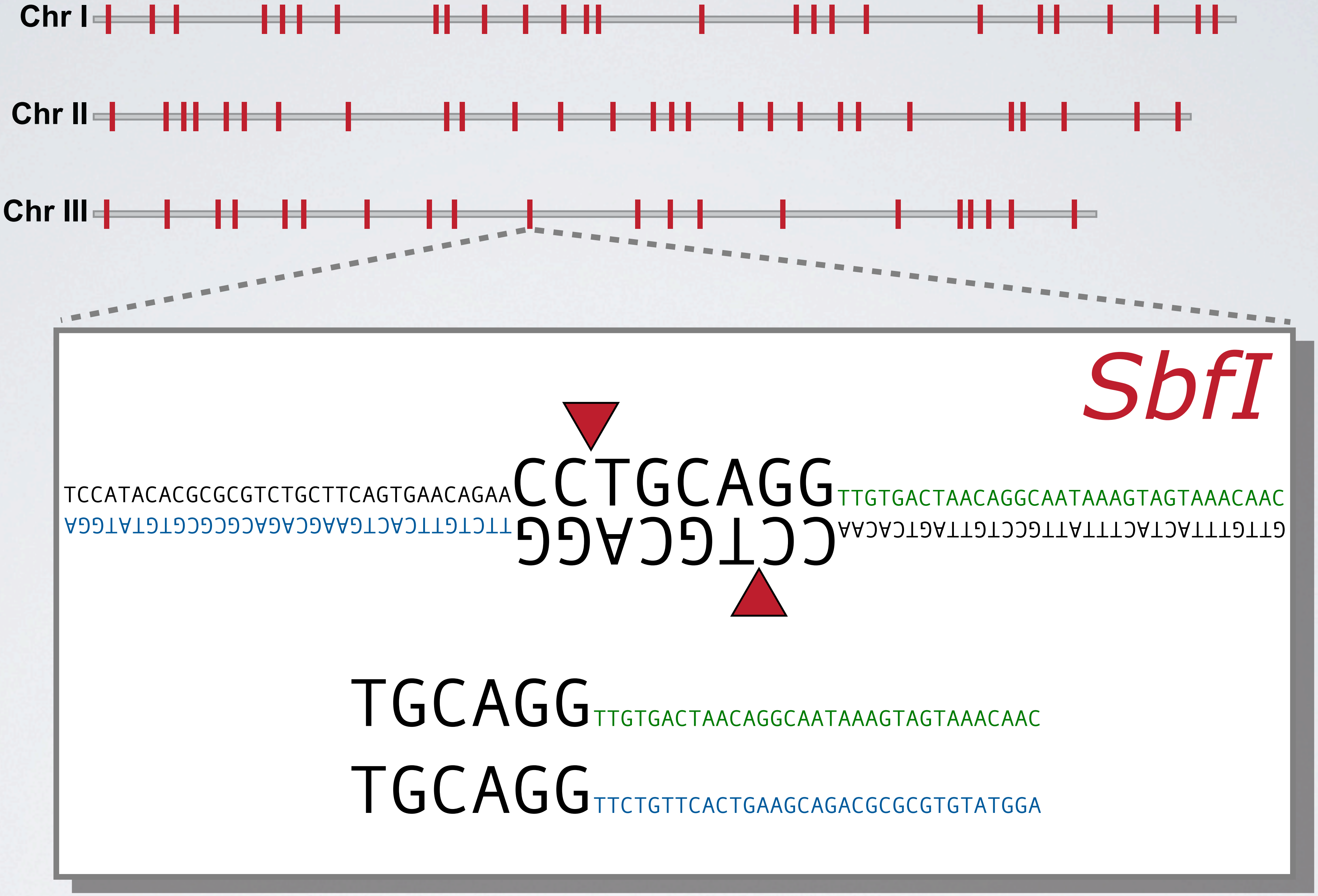
What is RAD-seq?

(Restriction-site Associated DNA)



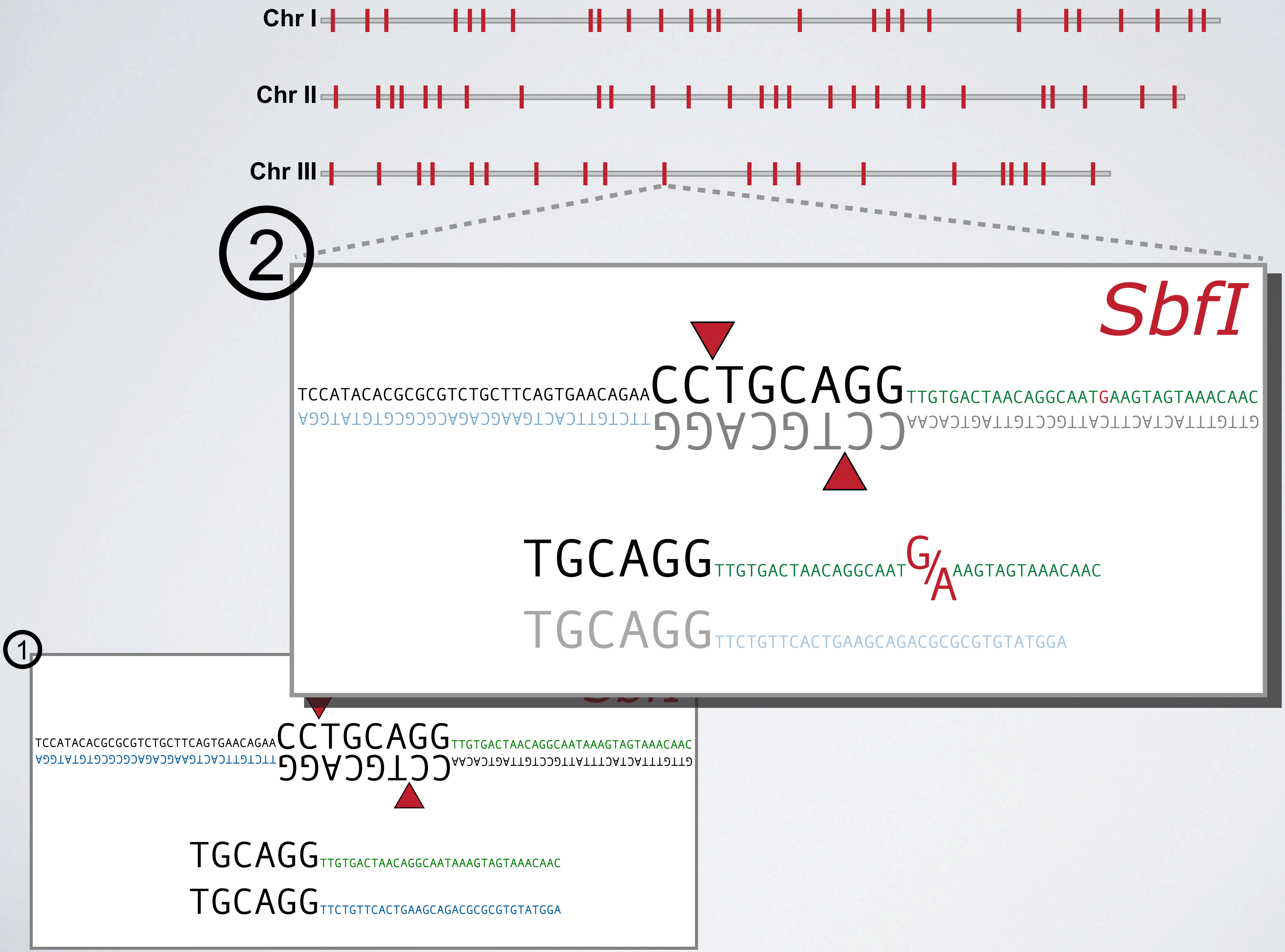
What is RAD-seq?

(Restriction-site Associated DNA)



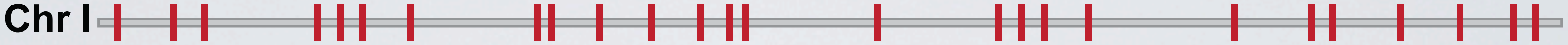
What is RAD-seq?

(Restriction-site Associated DNA)



What is RAD-seq?

(Restriction-site Associated DNA)



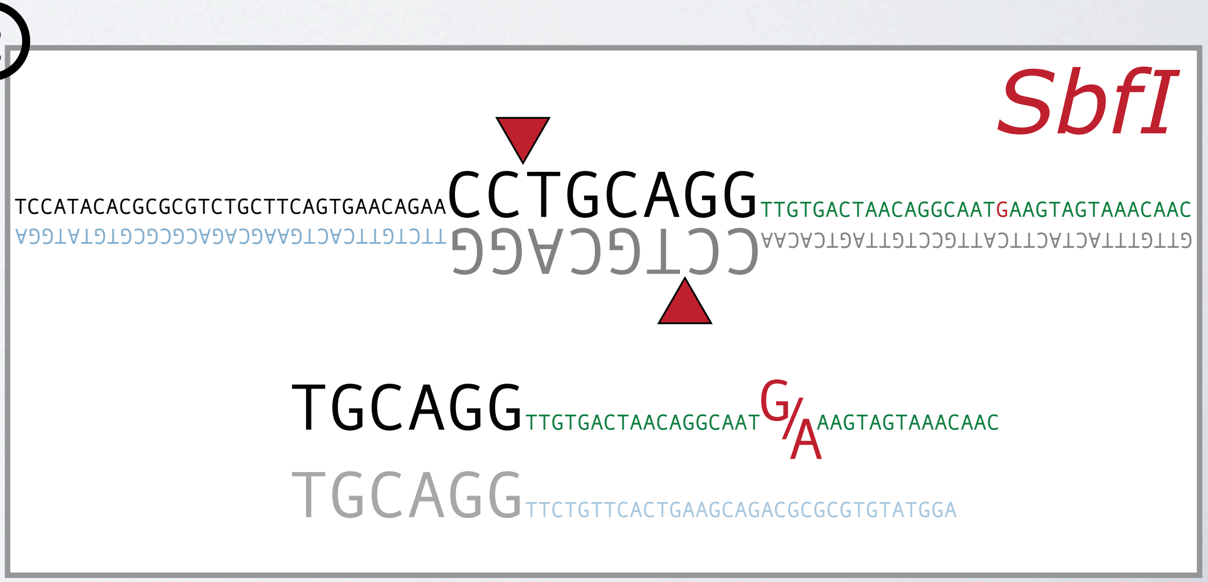
22,830 *SbfI* sites in Stickleback

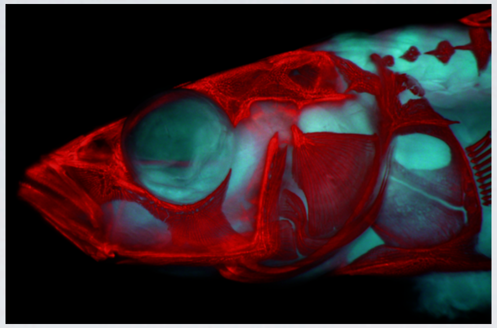
~ 45,000 RAD-Tags

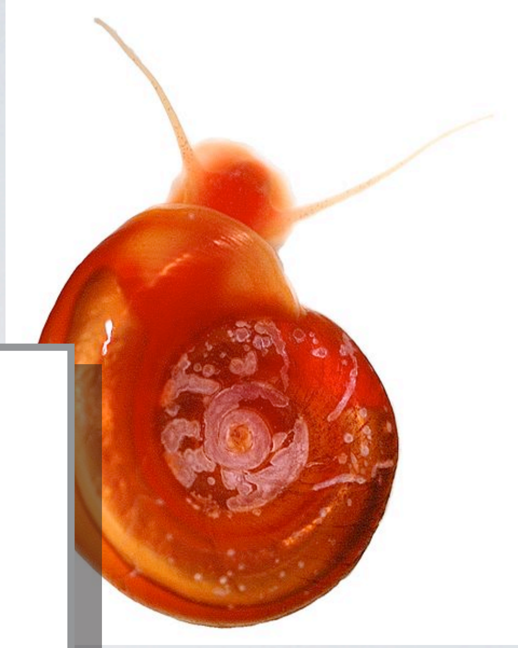
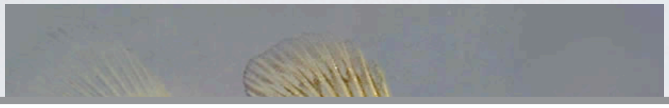
HiSeq Illumina Lane:
100 million reads, 96 barcoded individuals

①

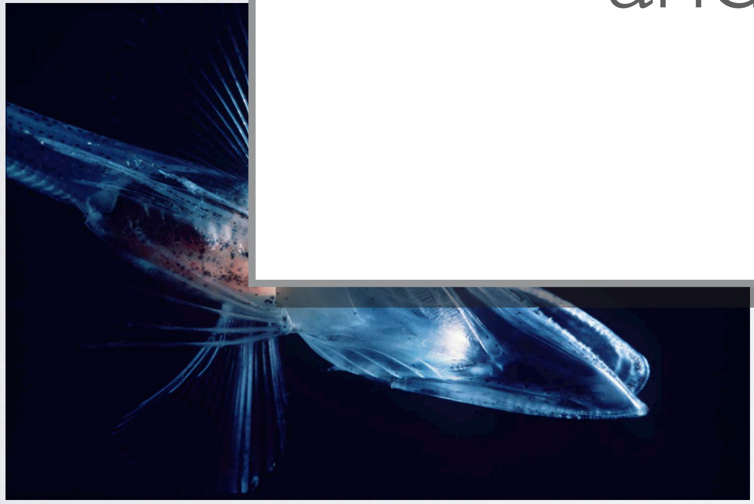
②

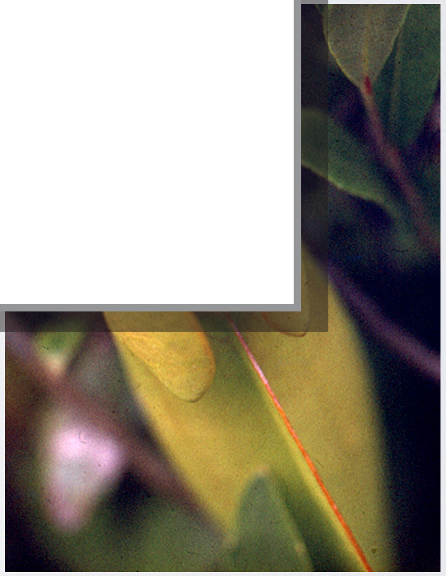


Wet lab and statistical considerations for RAD-seq and other RRL-seq studies





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

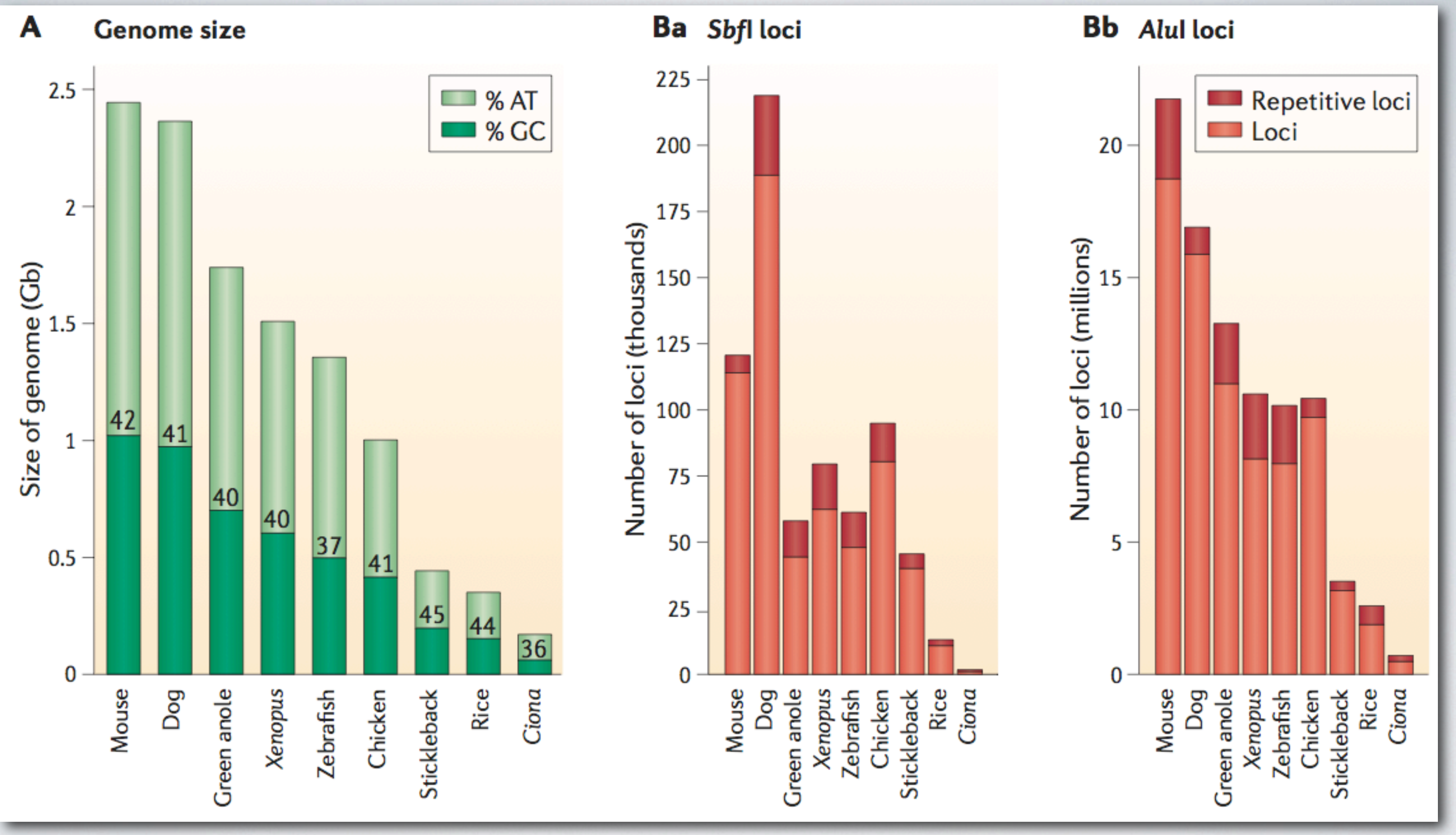
Choice of enzyme and genome size $(0.25)^n \times$ genome size = expected # sites

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

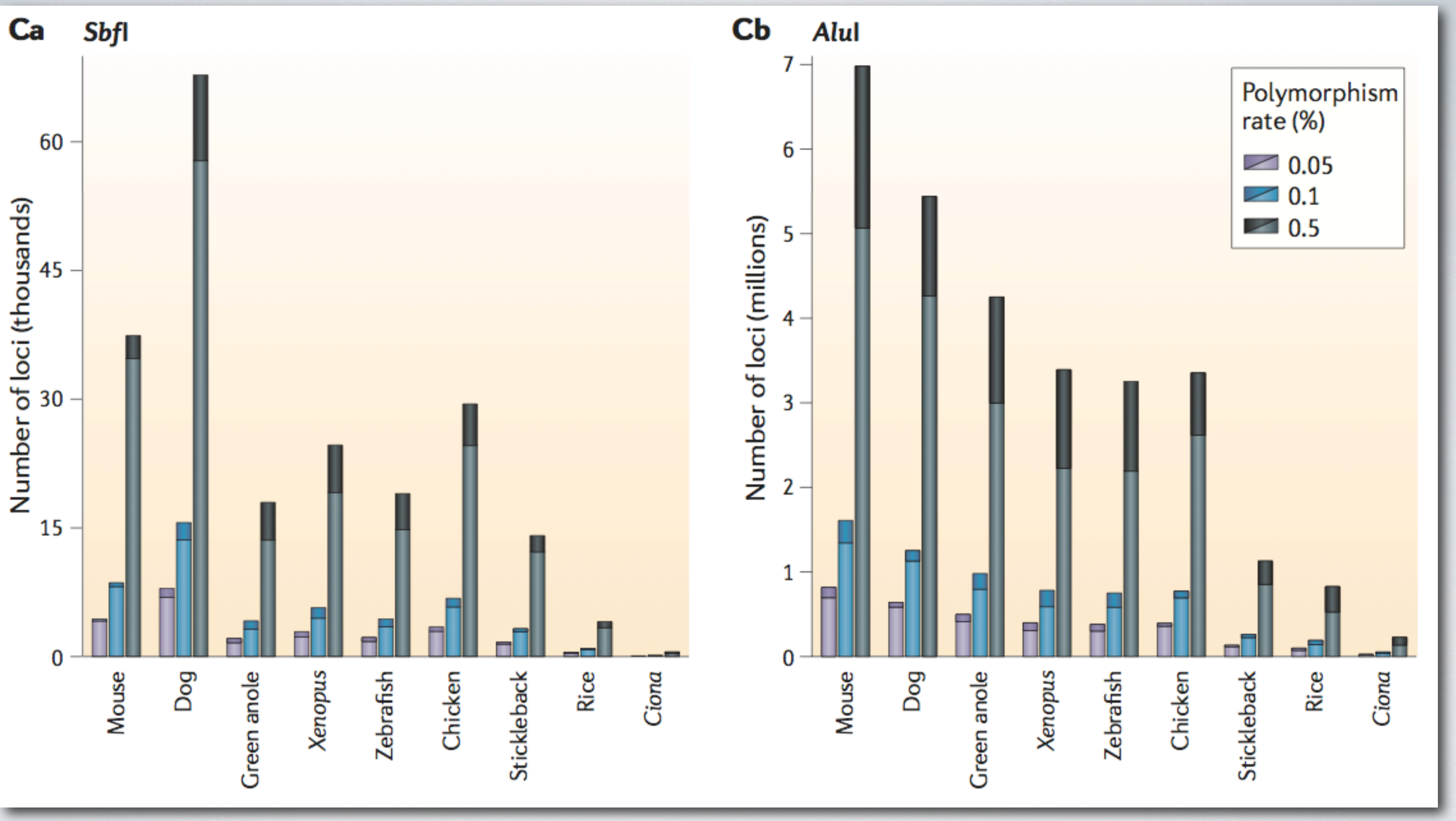
Choice of enzyme and genome size
Polymorphism and read length

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

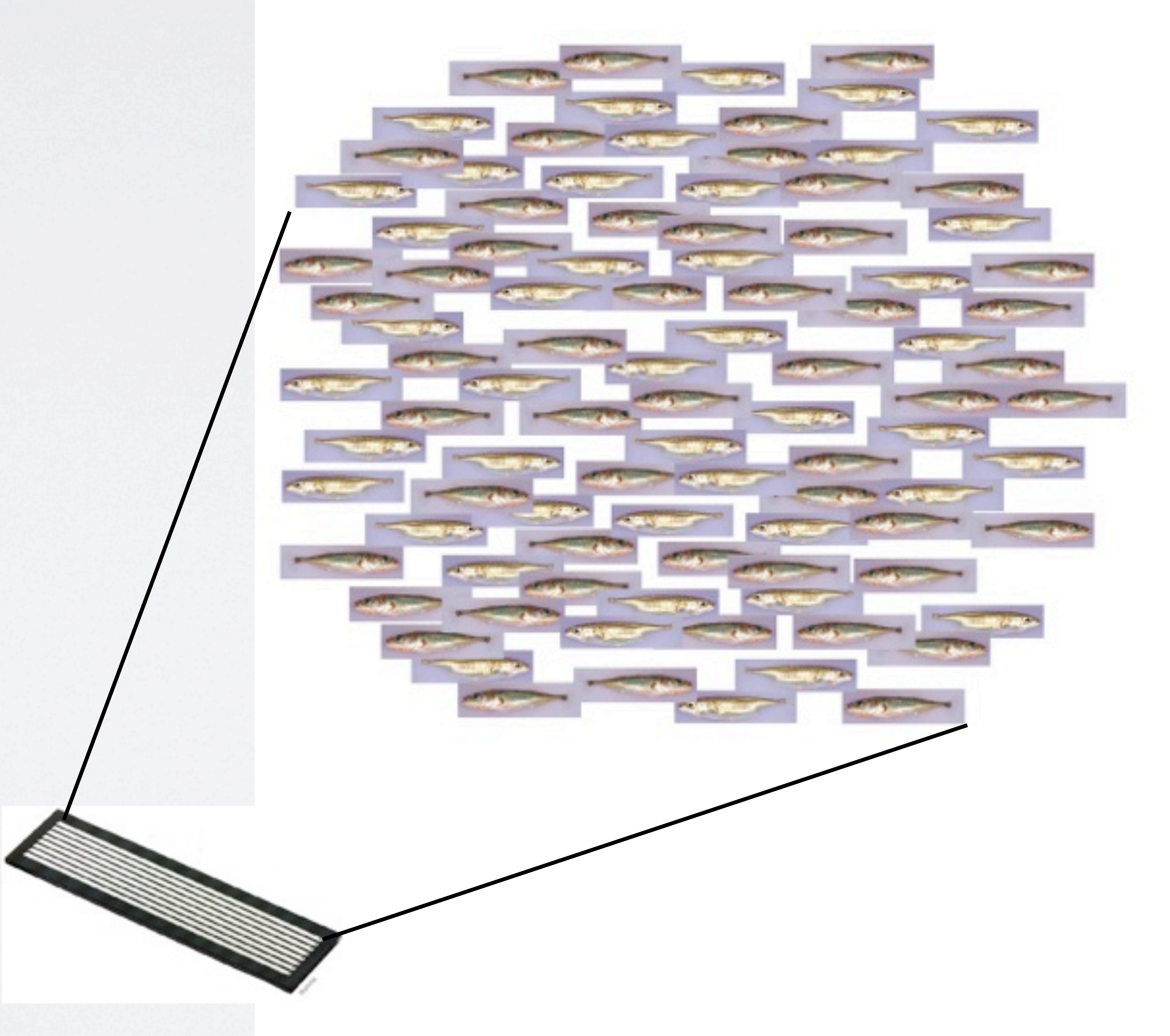
John W. Davey[*], Paul A. Hohenlohe[‡], Paul D. Etter[§], Jason Q. Boone[||], Julian M. Catchen[‡] and Mark L. Blaxter[*¶]*

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

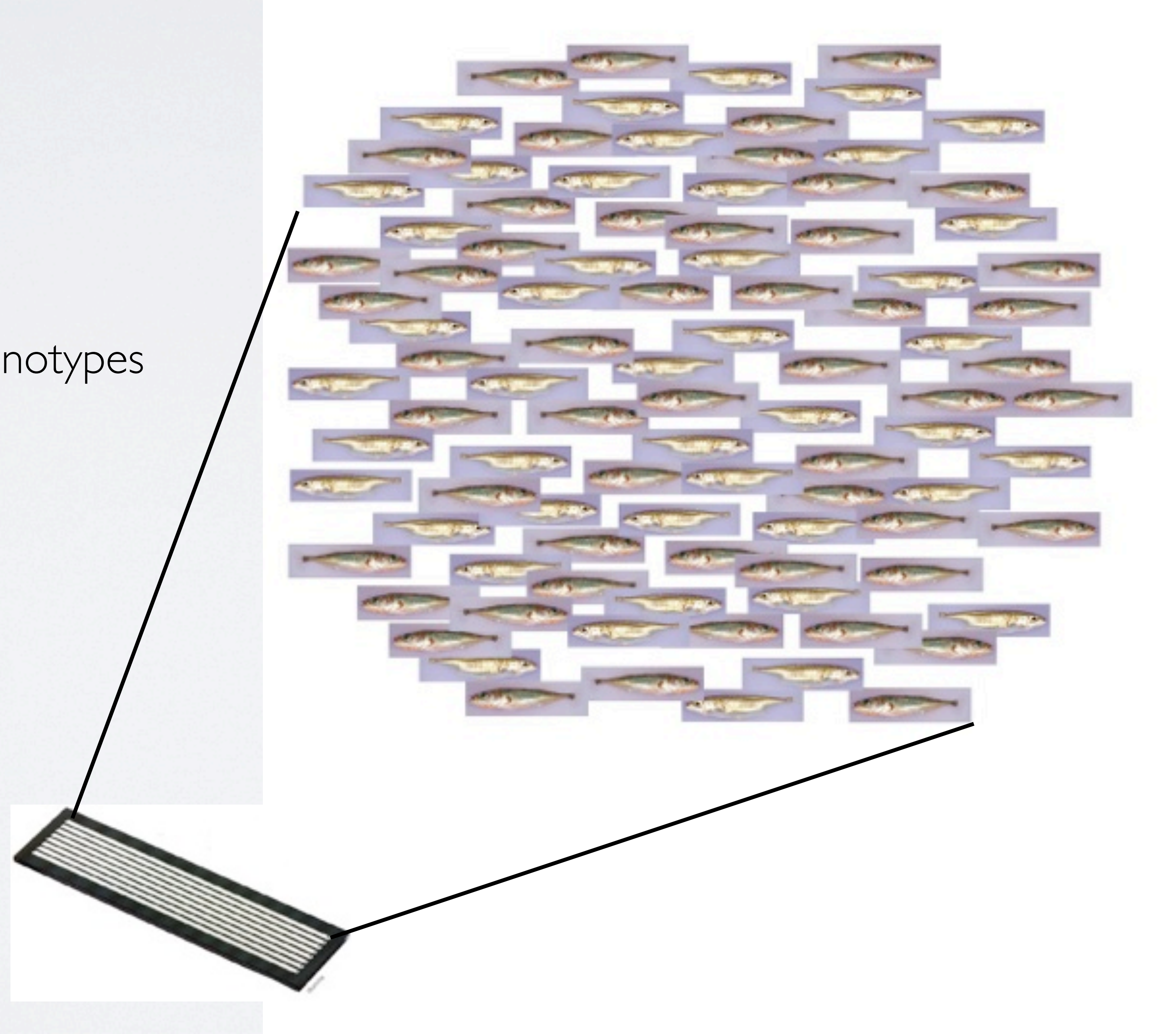
Experimental design considerations for RAD

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



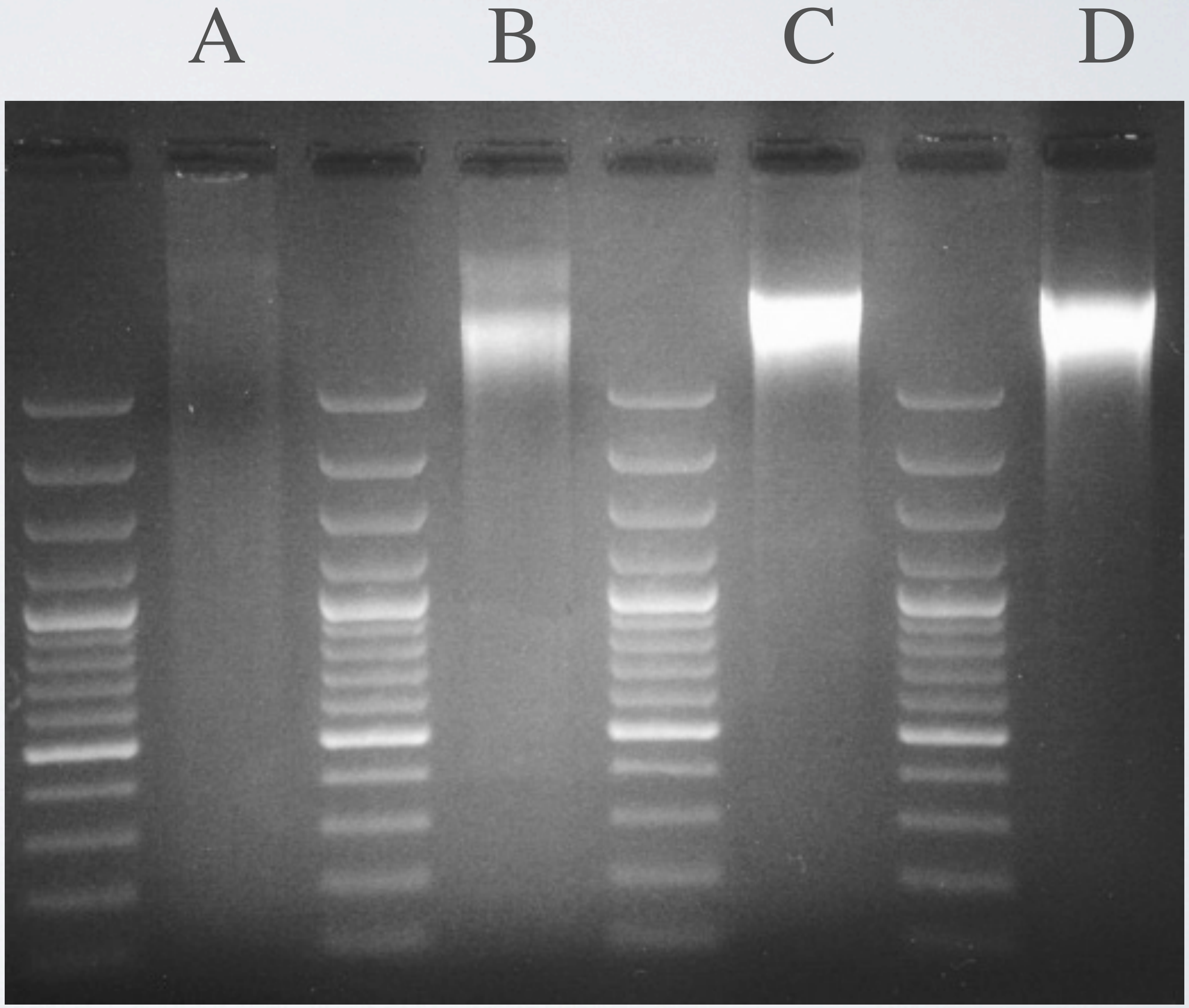
Molecular considerations in library building

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

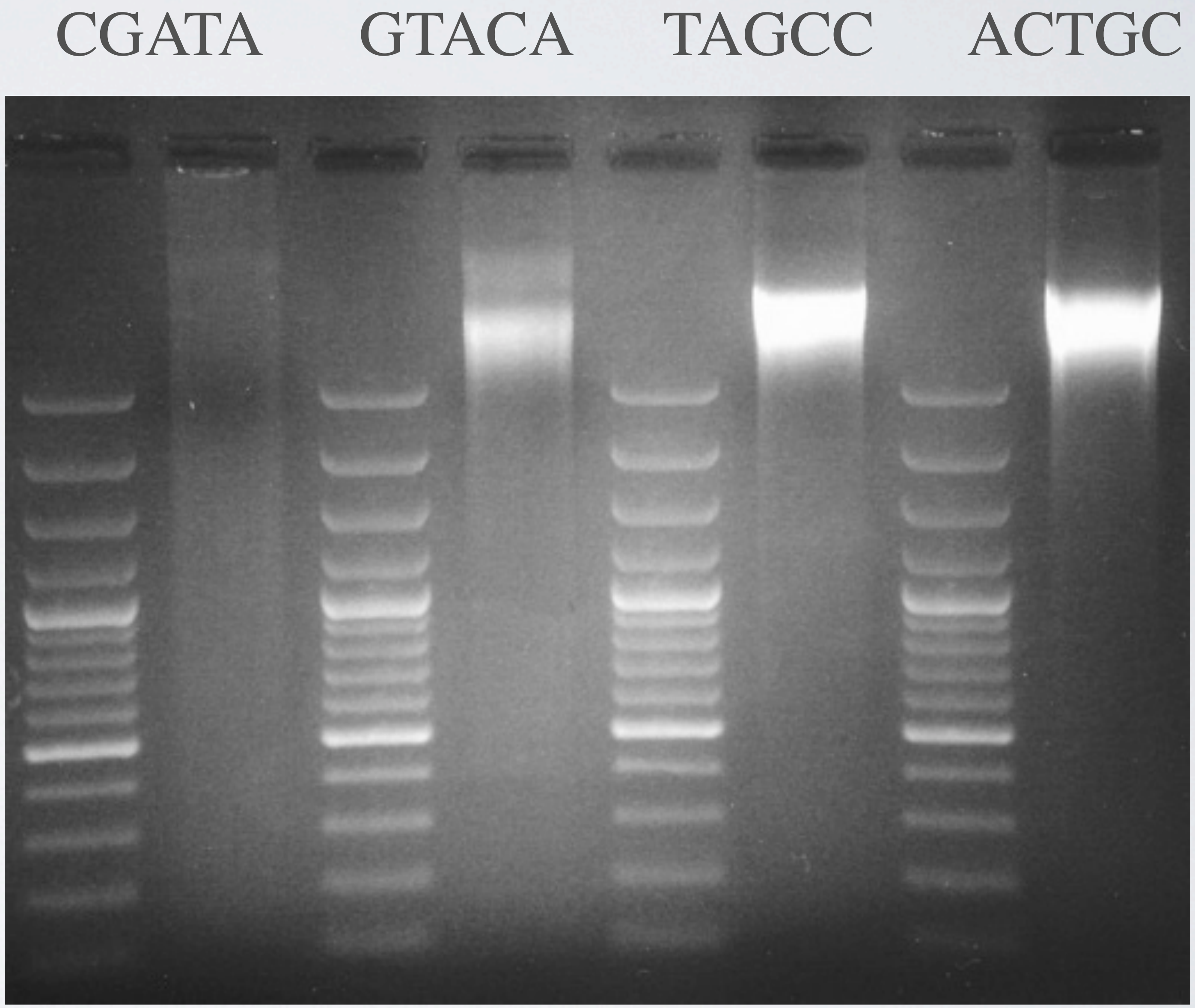
How many samples should be multiplexed?

Things to consider

DNA Quality

Diversify barcodes

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

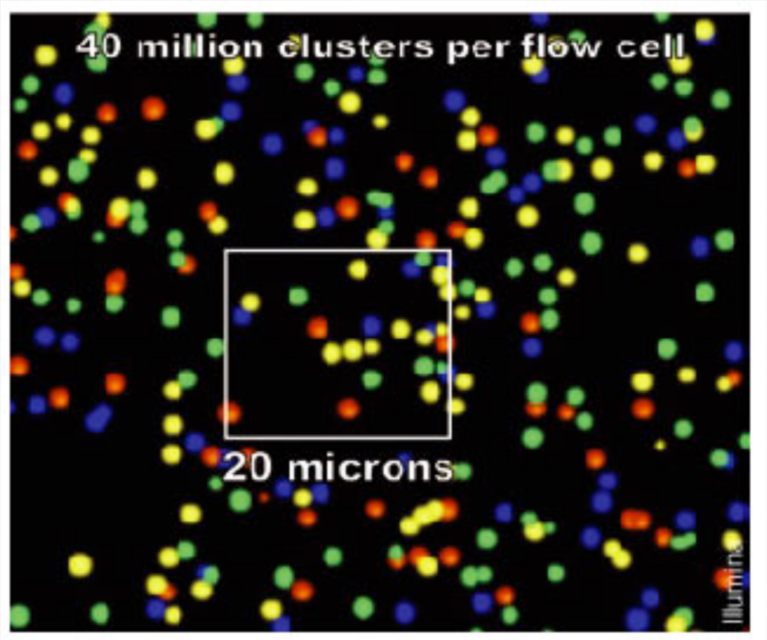


Molecular considerations in library building

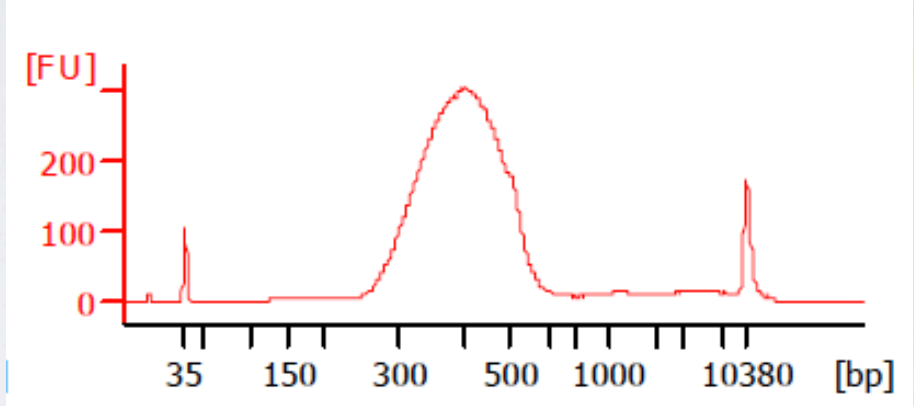
How can I get the best depth of coverage?

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?

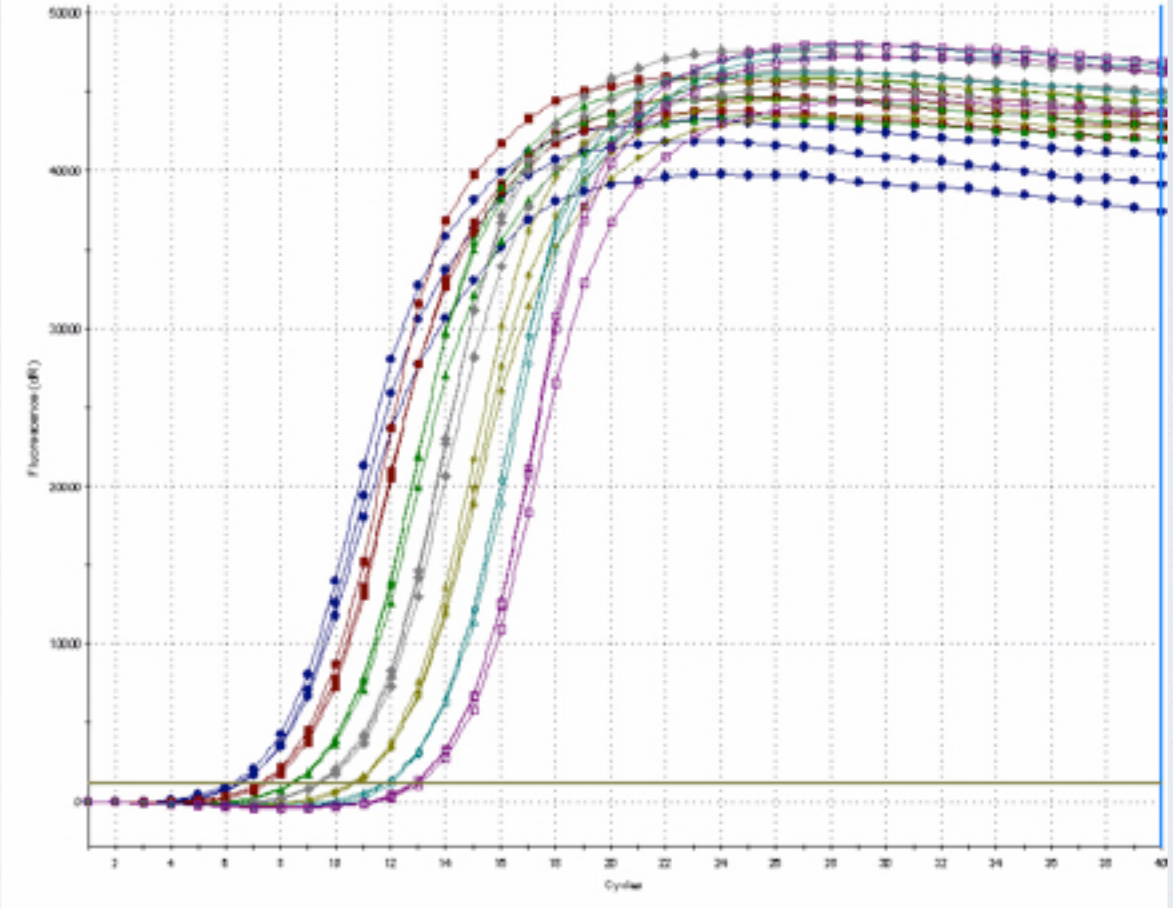
Things to consider

Fragment size

Library quality

qPCR

qPCR control should be similar to measured sample:

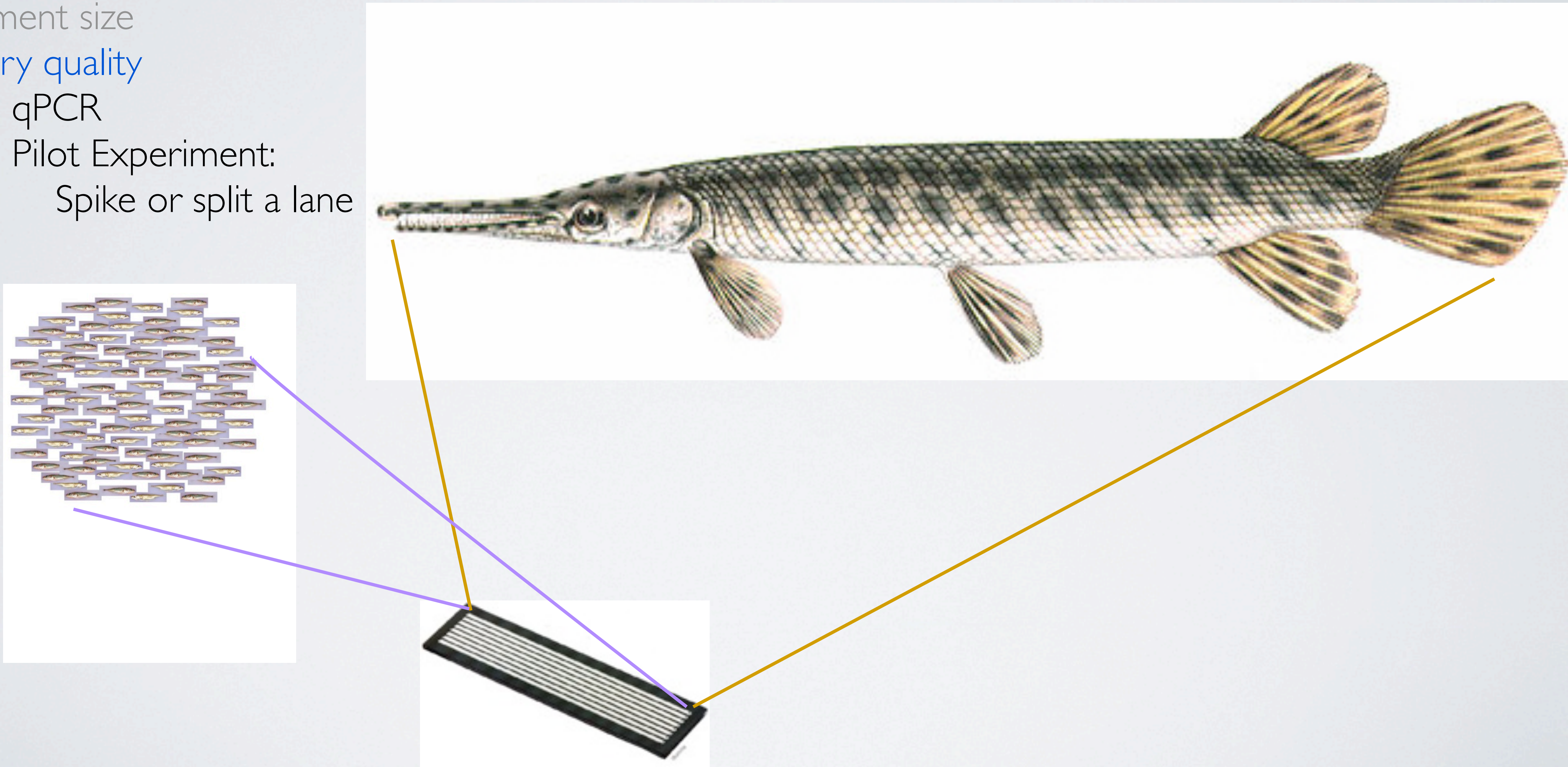


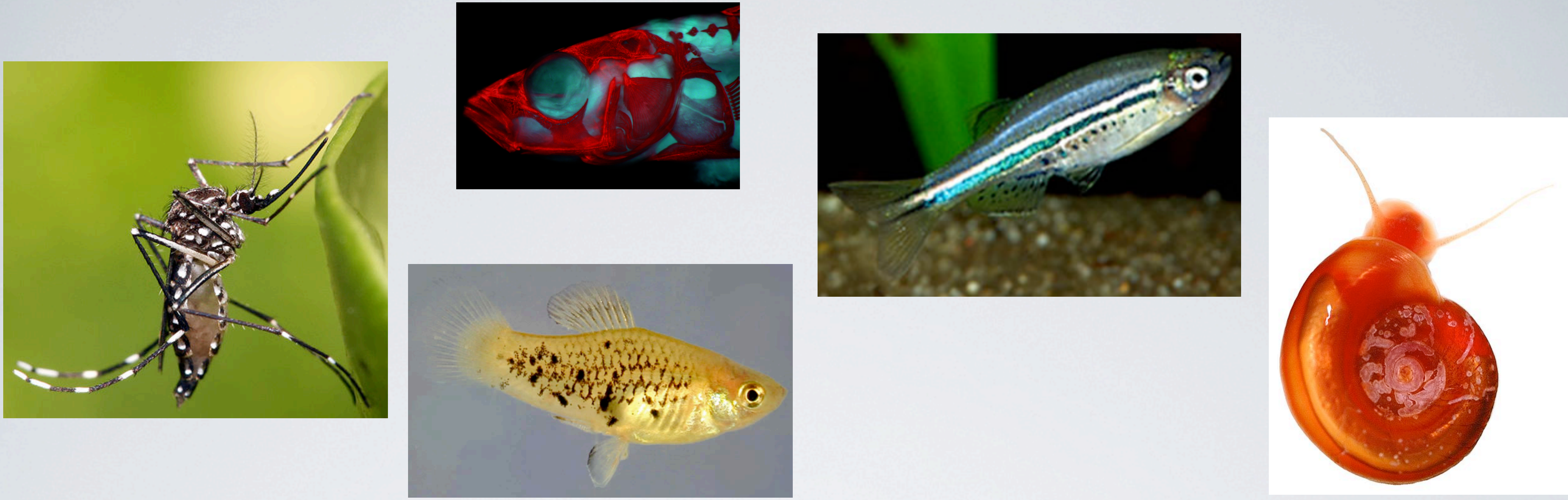
Molecular considerations in library building

How can I get the best depth of coverage?

Things to consider

Fragment size
Library quality
- qPCR
- Pilot Experiment:
  - Spike or split a lane



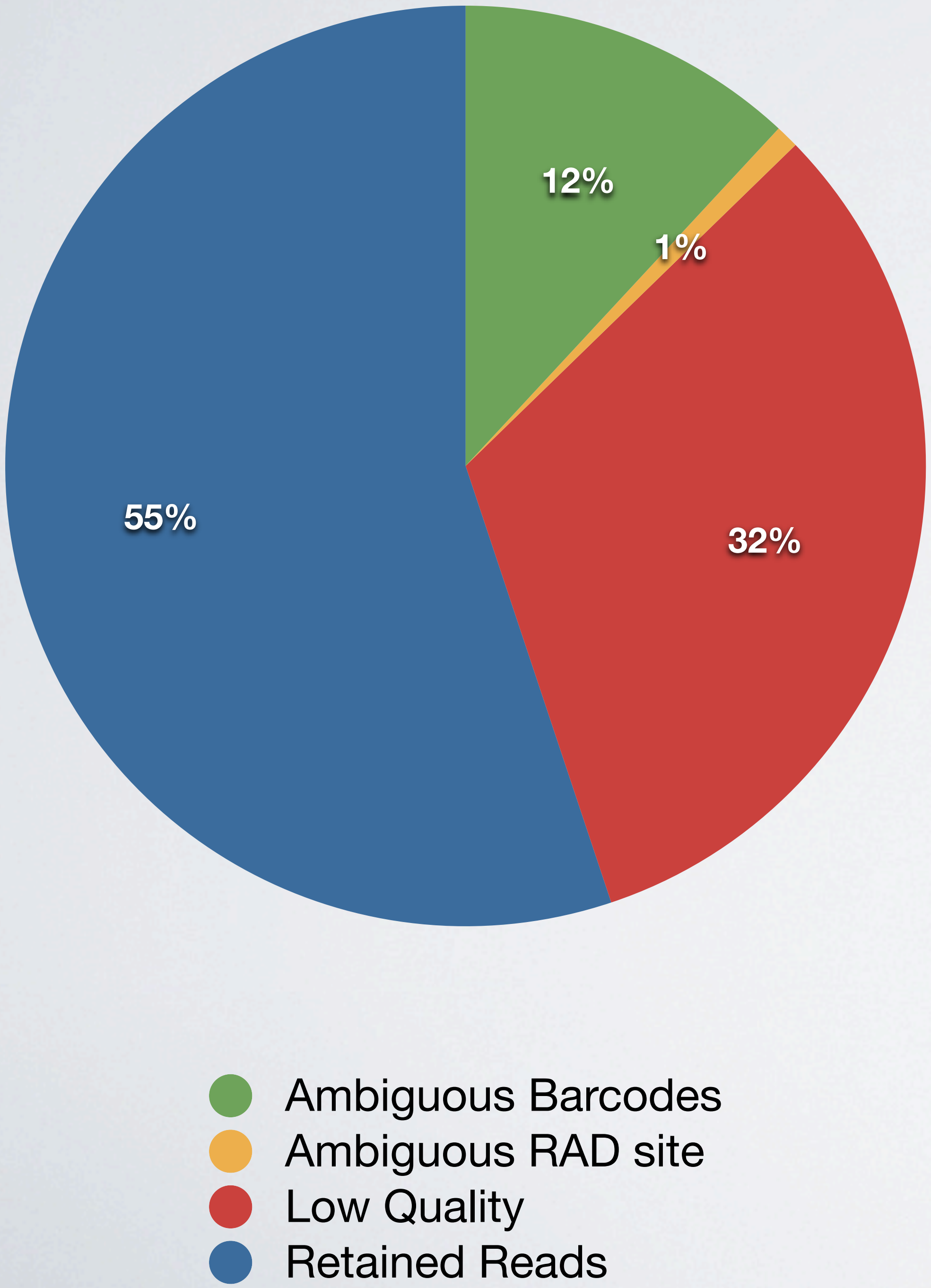


What if you have a reference genome?

Anatomy of a PstI RAD-seq Analysis

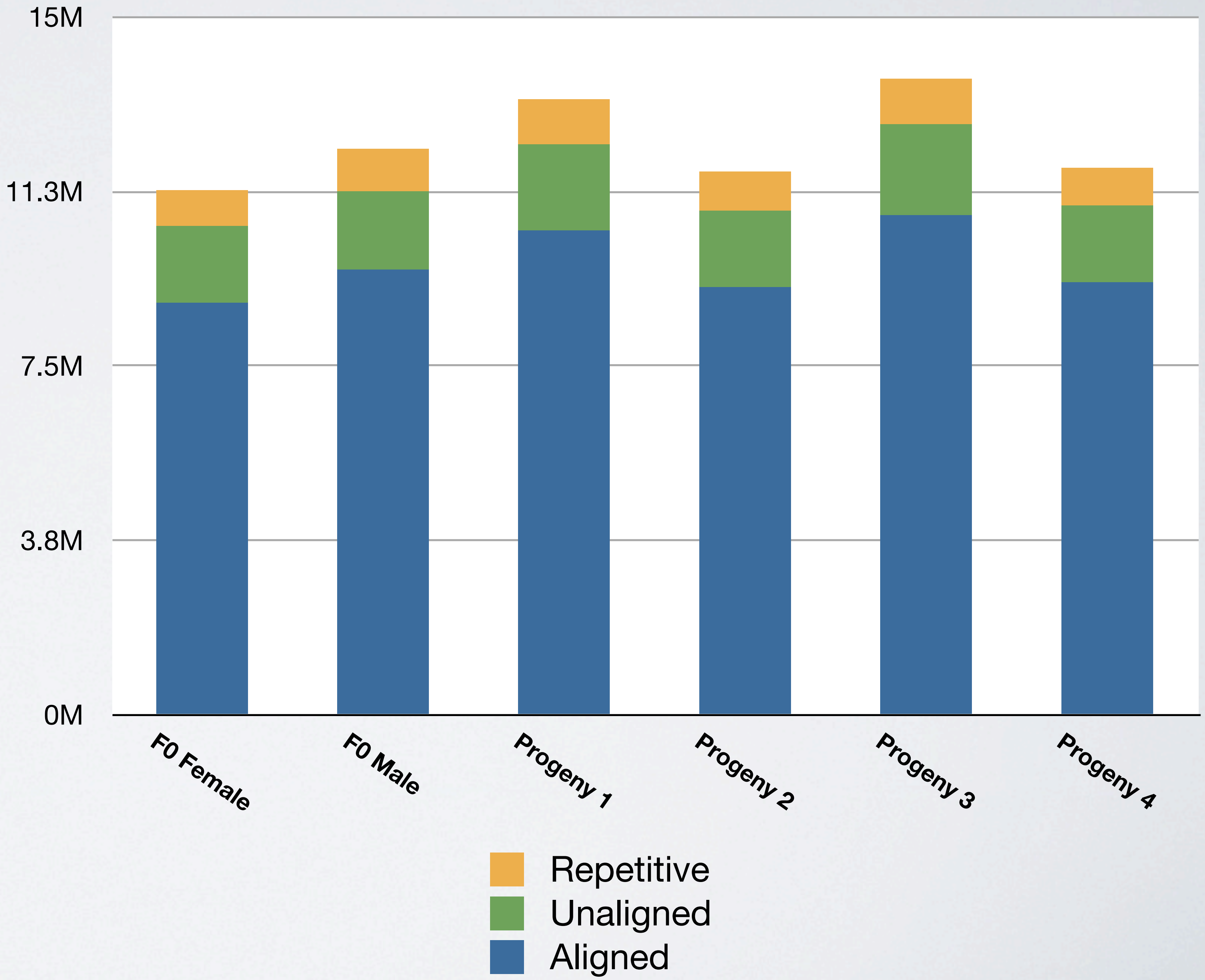
CTGCAG

133M Illumina Raw Reads



Aligned to Reference

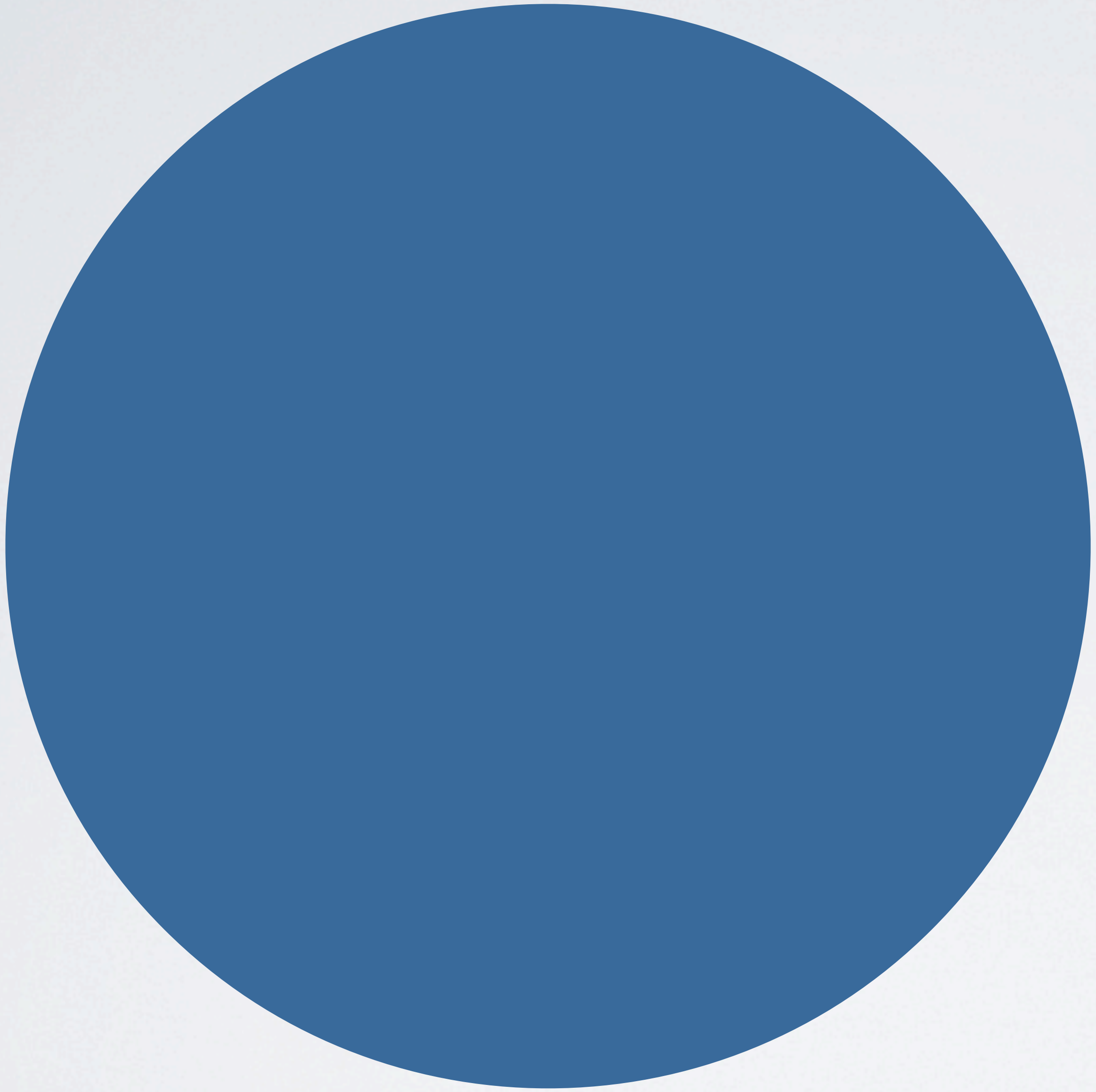
(Bowtie)



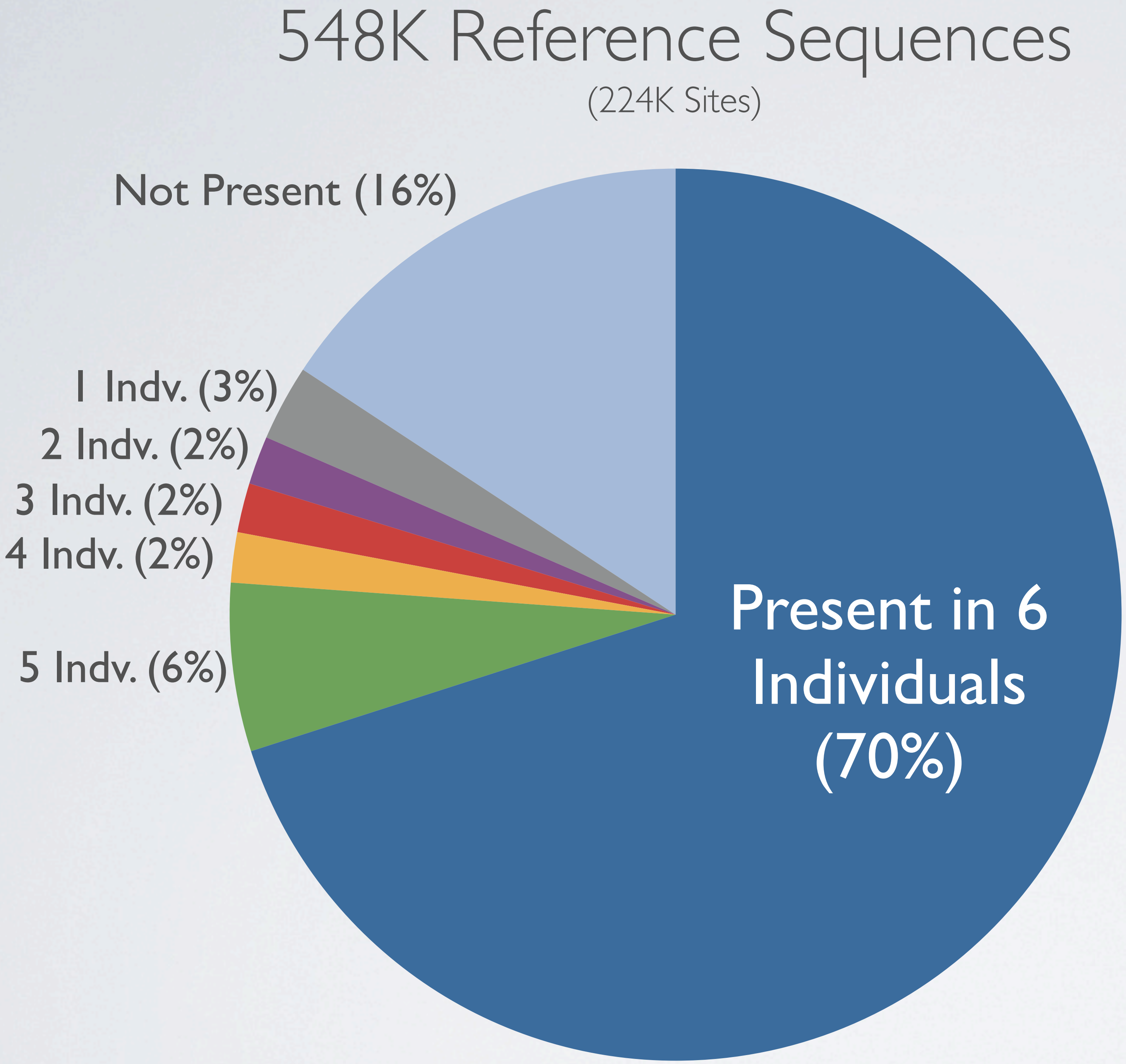
Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)



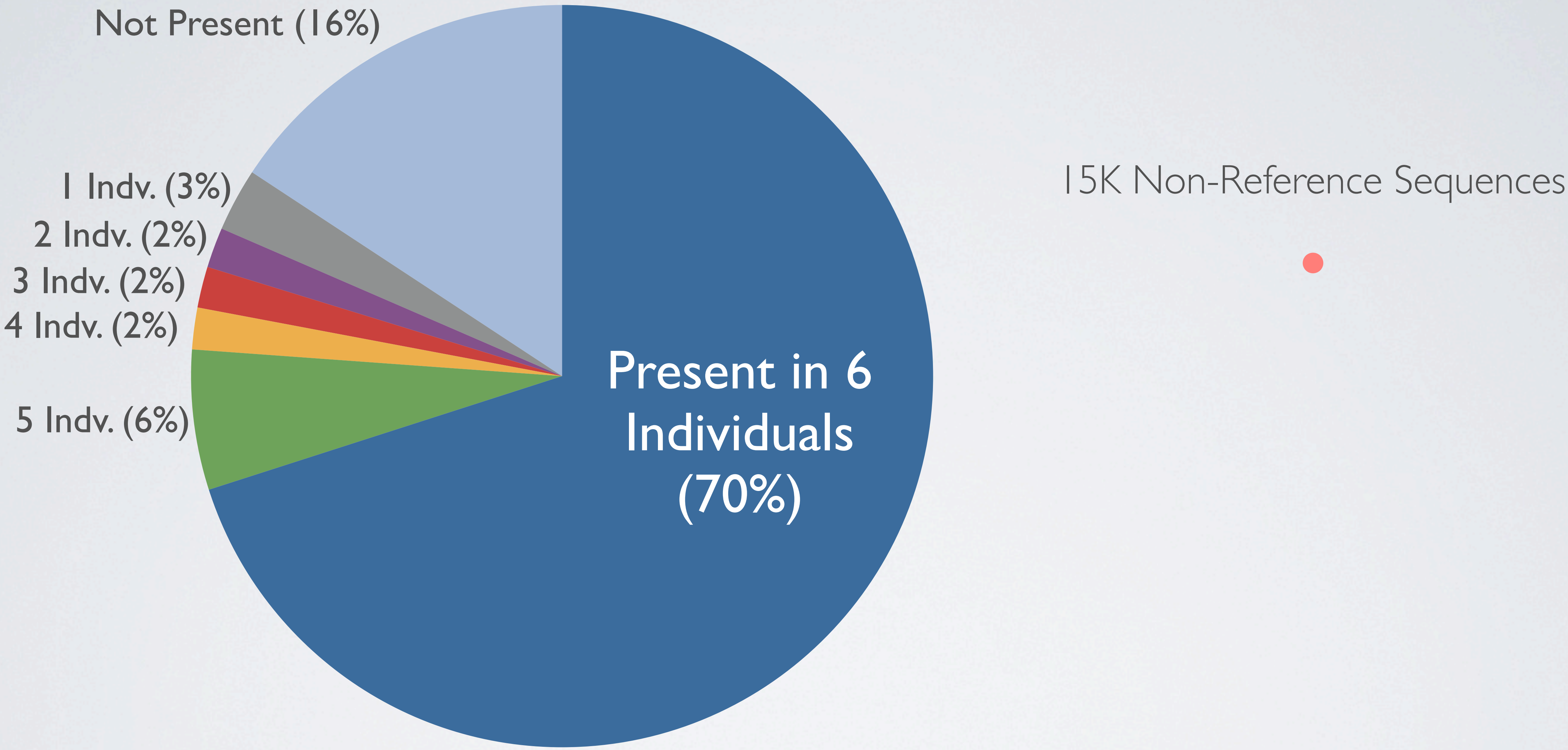
Anatomy of a PstI RAD-seq Analysis



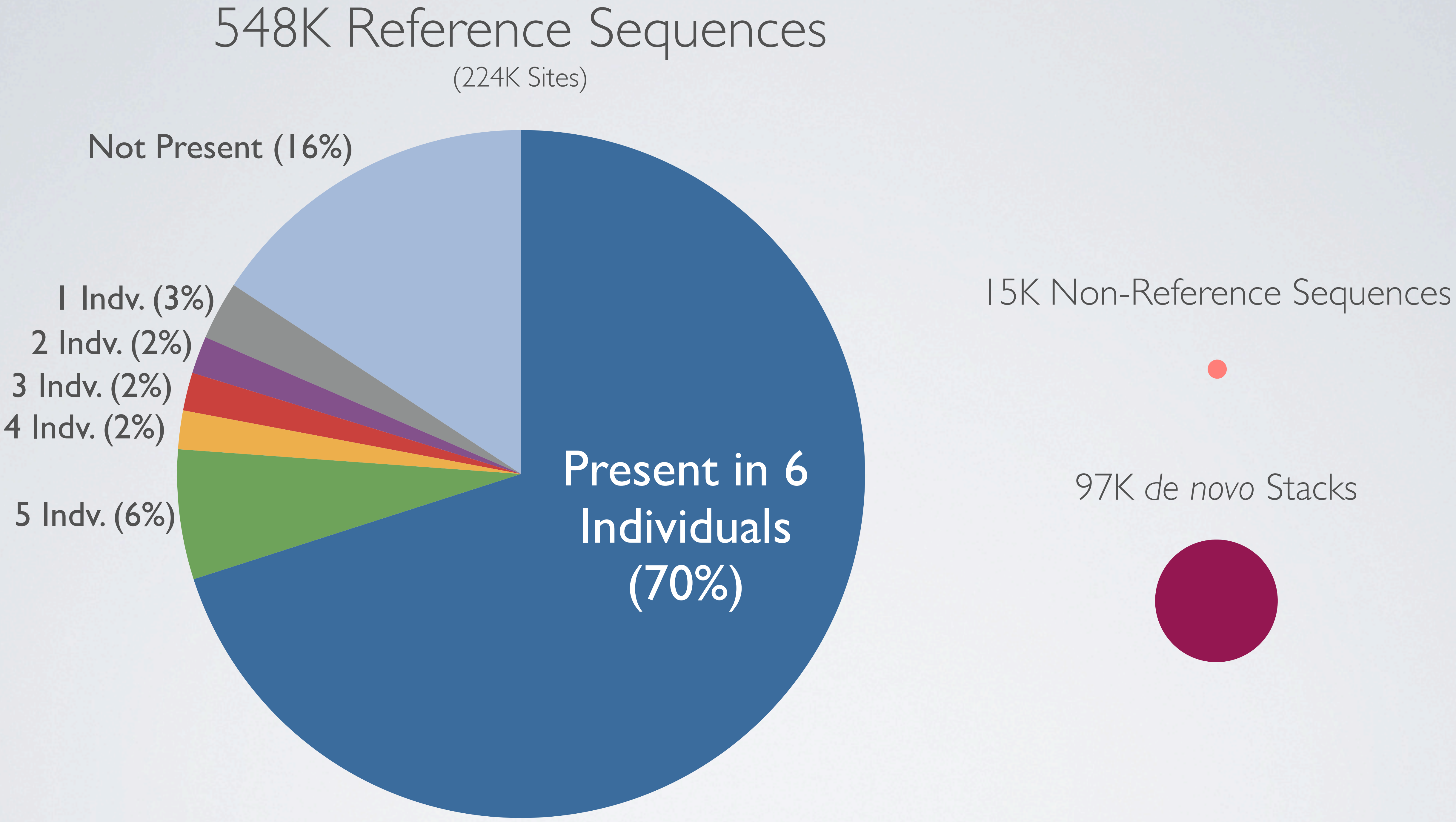
Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)



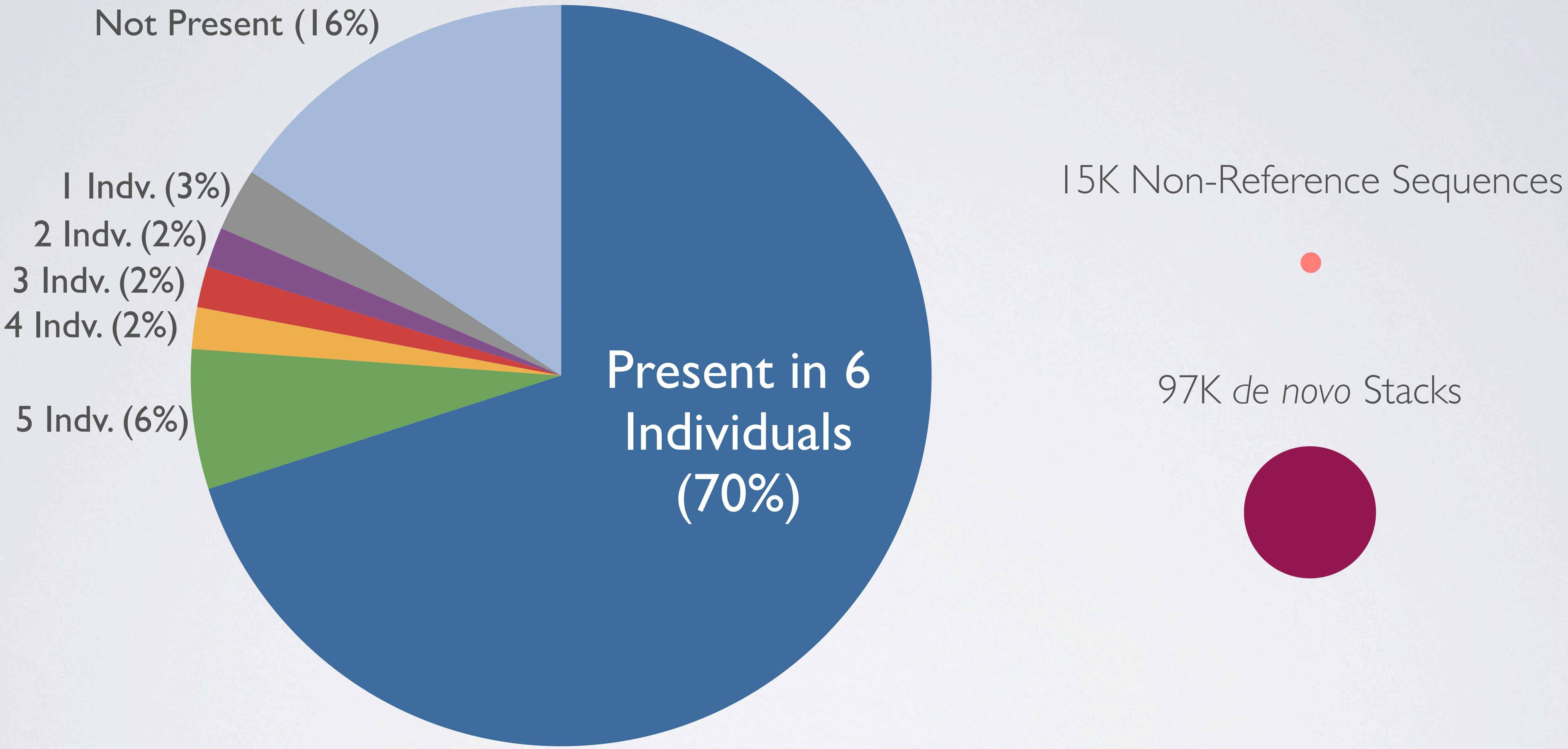
Anatomy of a PstI RAD-seq Analysis



Anatomy of a PstI RAD-seq Analysis

548K Reference Sequences

(224K Sites)



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

Anatomy of a PstI RAD-seq Analysis

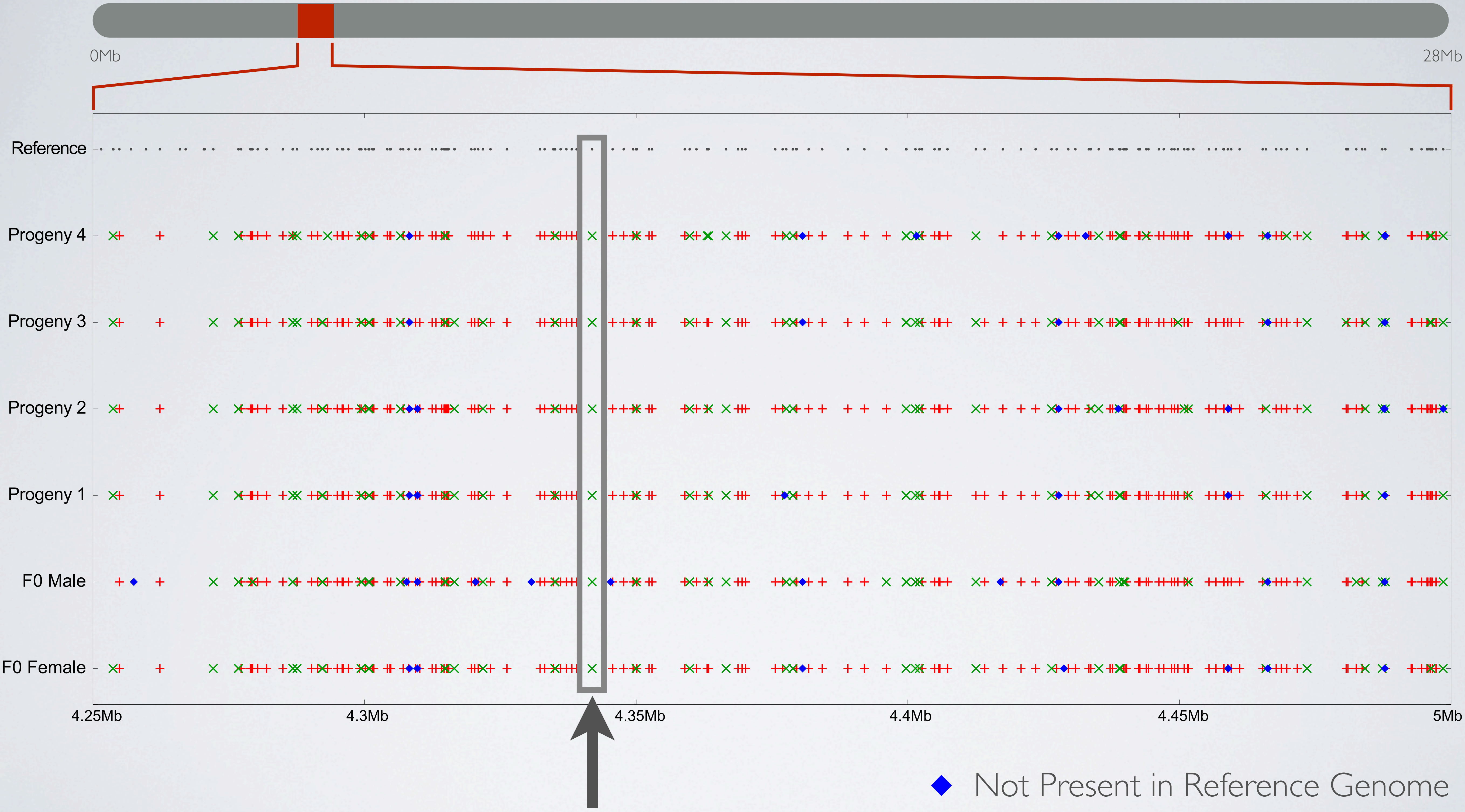
Bear Paw Stickleback Reference - LGI

Anatomy of a PstI RAD-seq Analysis

Bear Paw Stickleback Reference - LGI

Anatomy of a PstI RAD-seq Analysis

Bear Paw Stickleback Reference - LGI

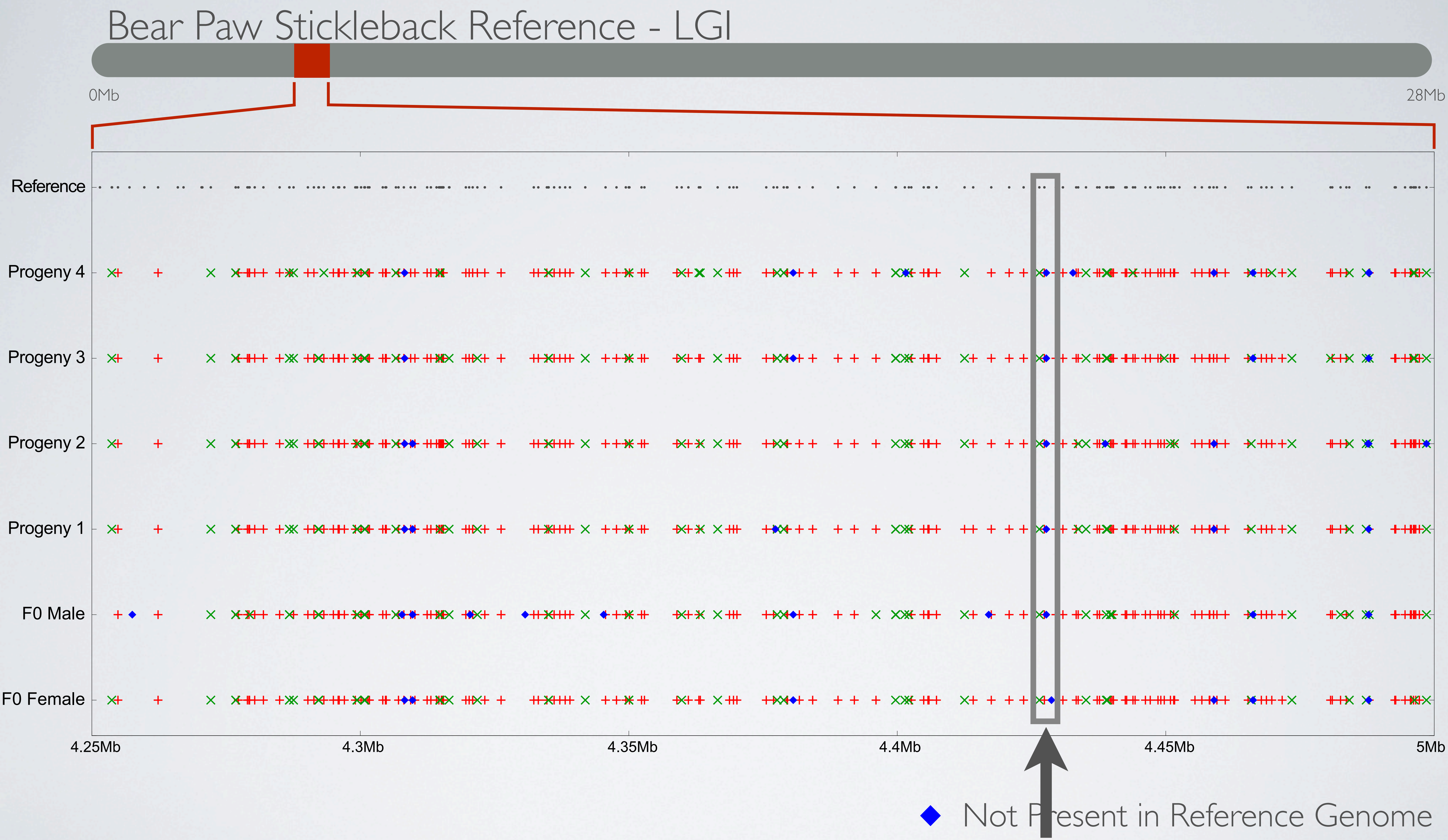


◆ Not Present in Reference Genome

× Detected on a Single Strand

+ Detected on Both Strands

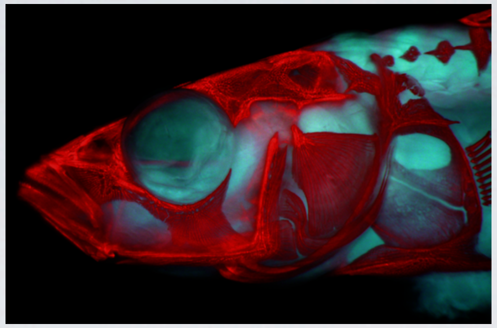
Anatomy of a PstI RAD-seq Analysis

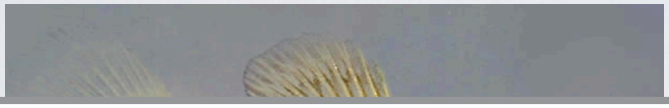
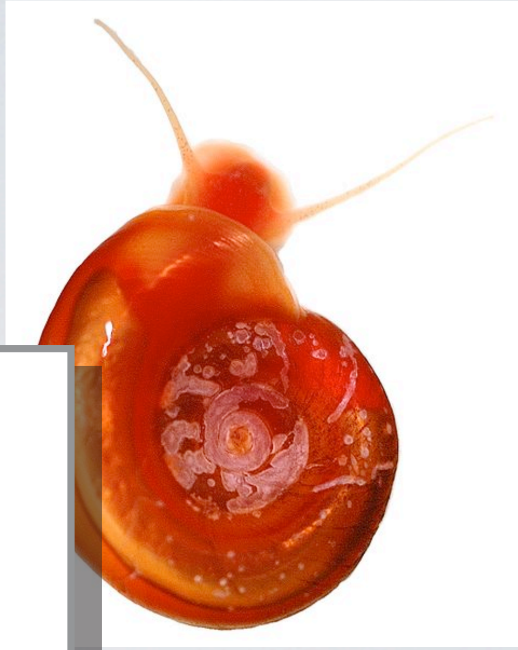


◆ Not Present in Reference Genome

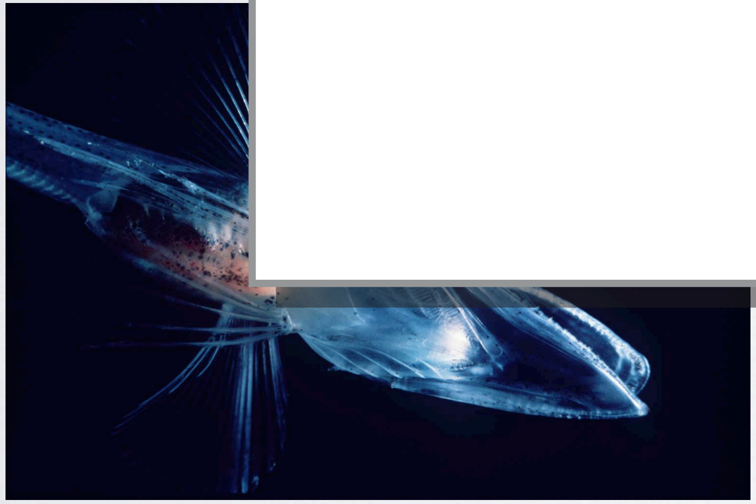
✕ Detected on a Single Strand

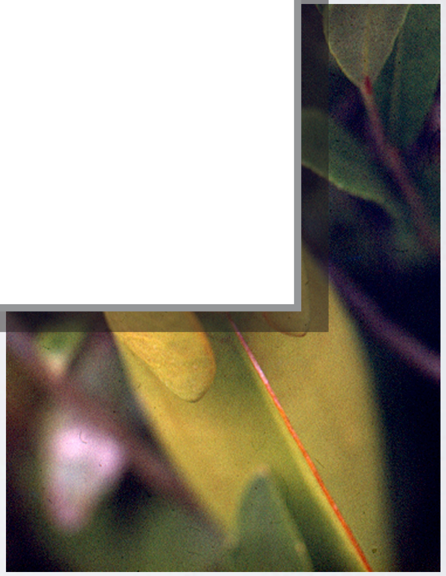
\+ Detected on Both Strands

Signatures of natural selection across the genome



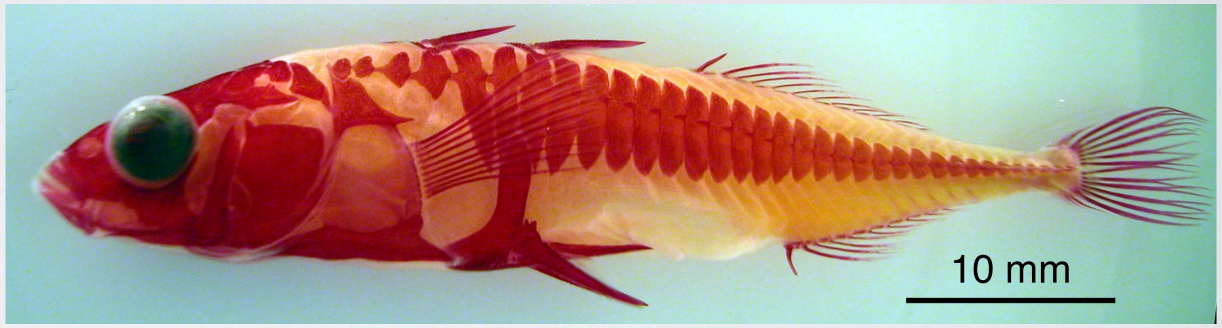


Threespine Stickleback, *Gasterosteus aculeatus*

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

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

Ocean form

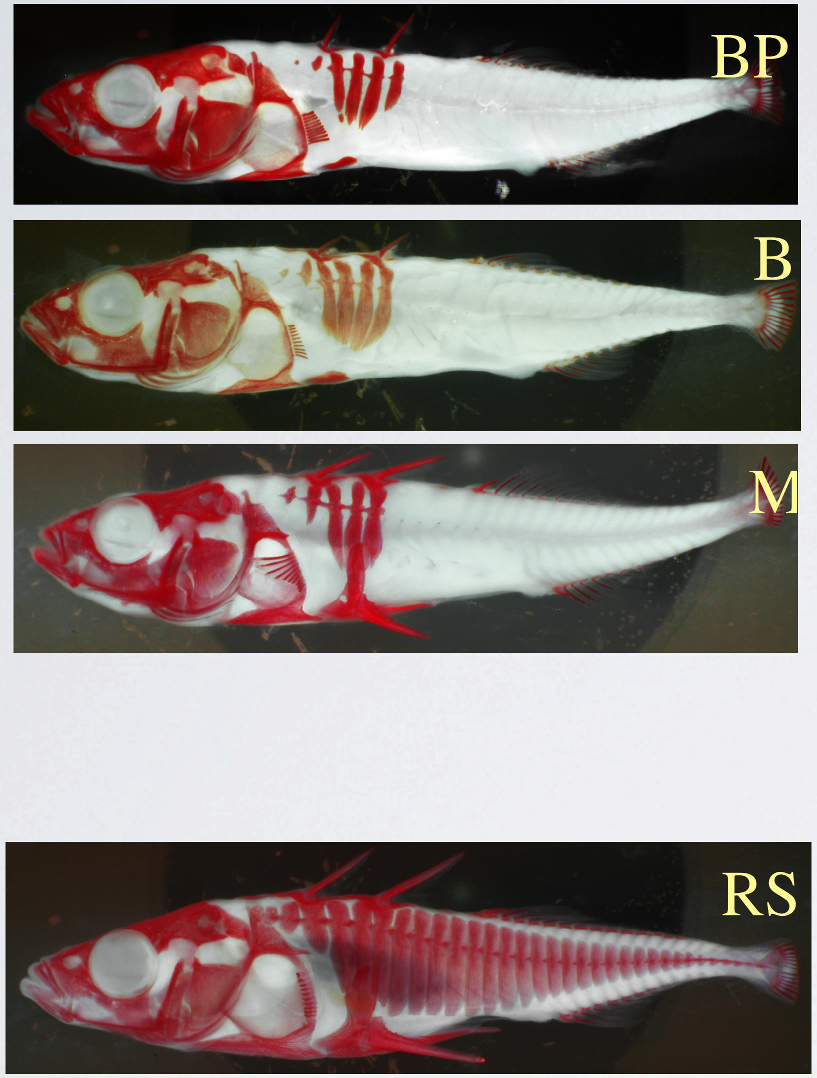


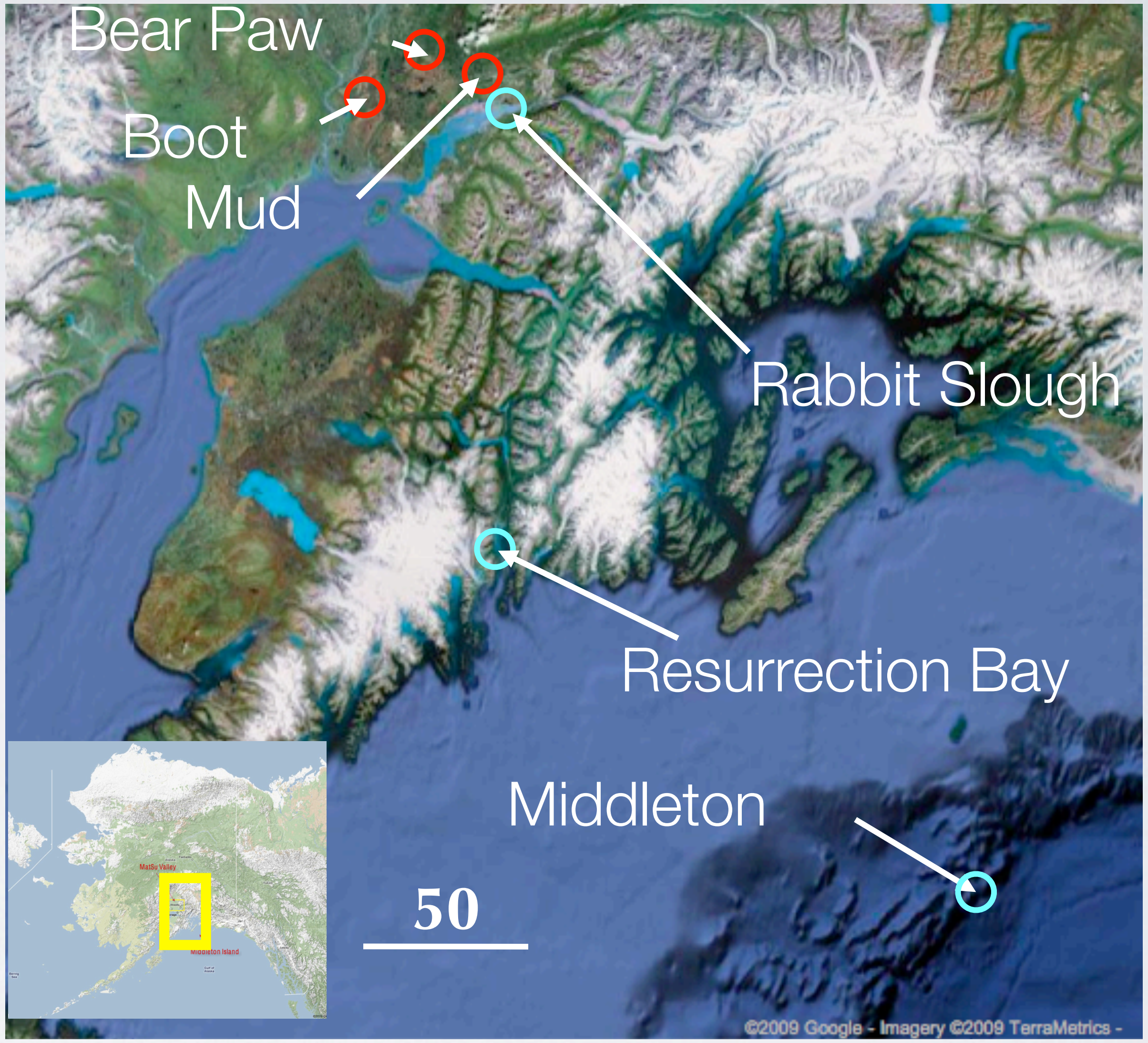
Freshwater form


Susan Bassham


Paul Hohenlohe

Signatures of natural selection across the genome

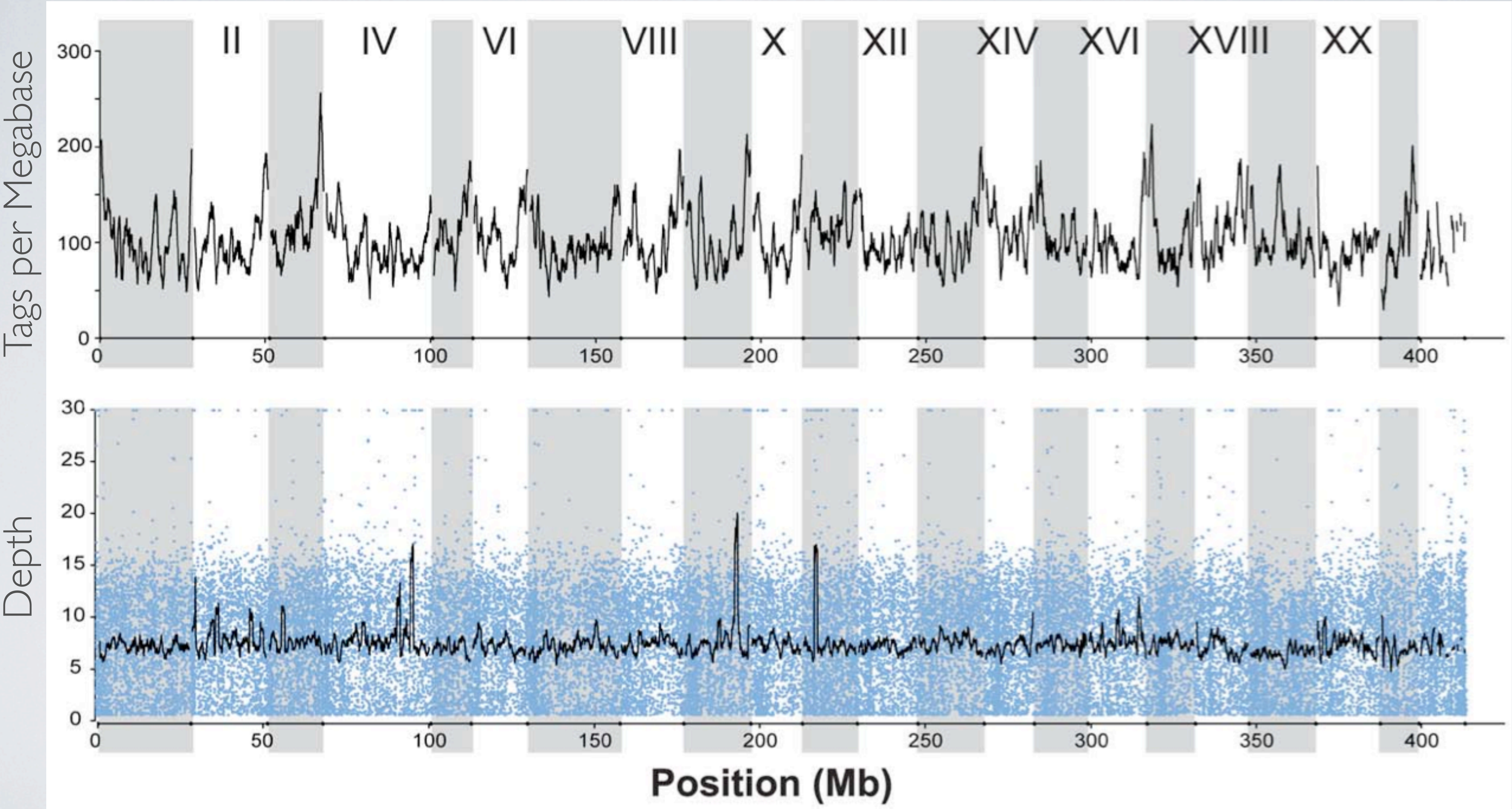




©2009 Google - Imagery ©2009 TerraMetrics -

- 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

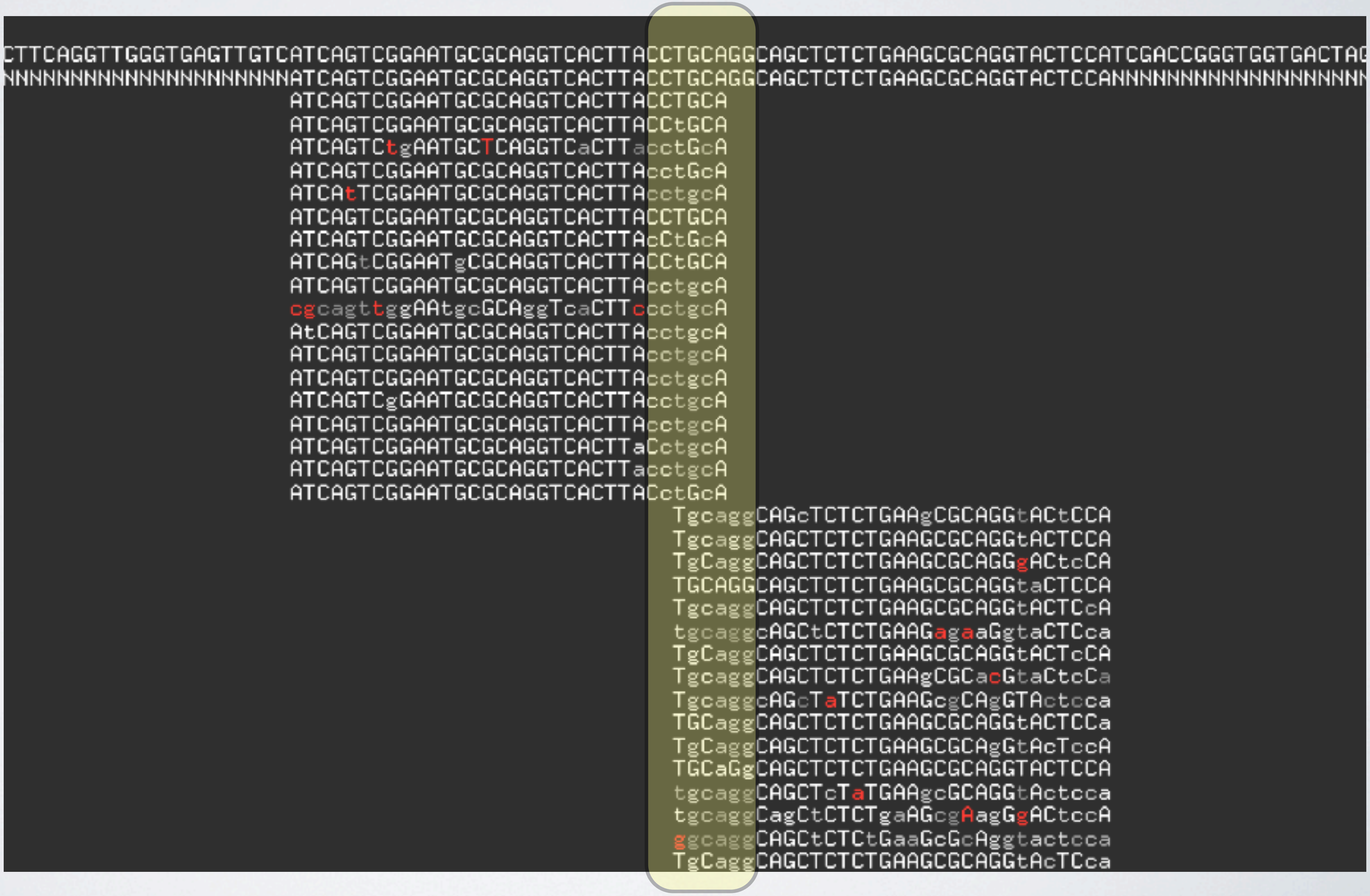


Concatenated stickleback genome

Differentiating SNPs from error

Restriction enzyme recognition site

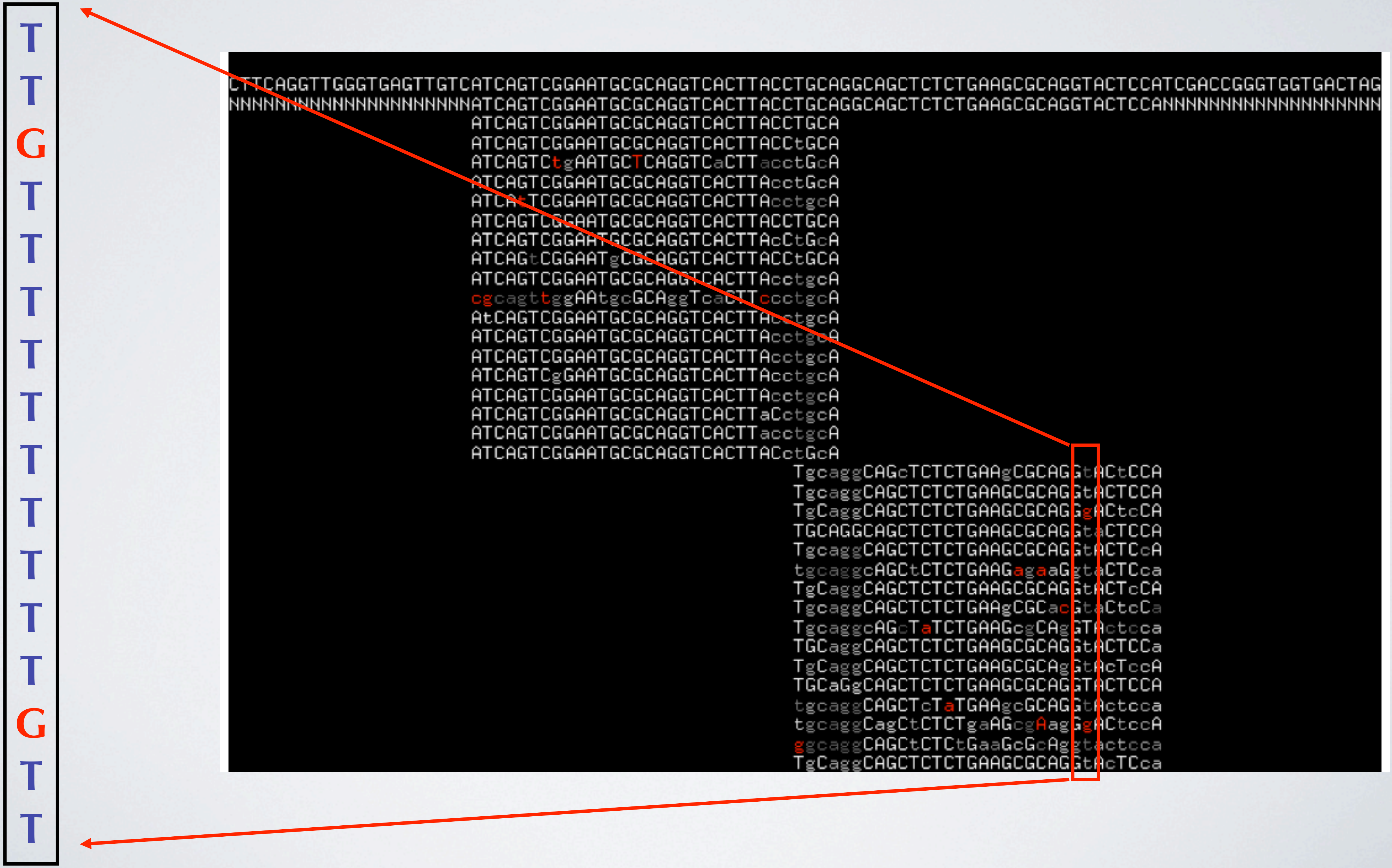
Reference genome sequence



```
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
```

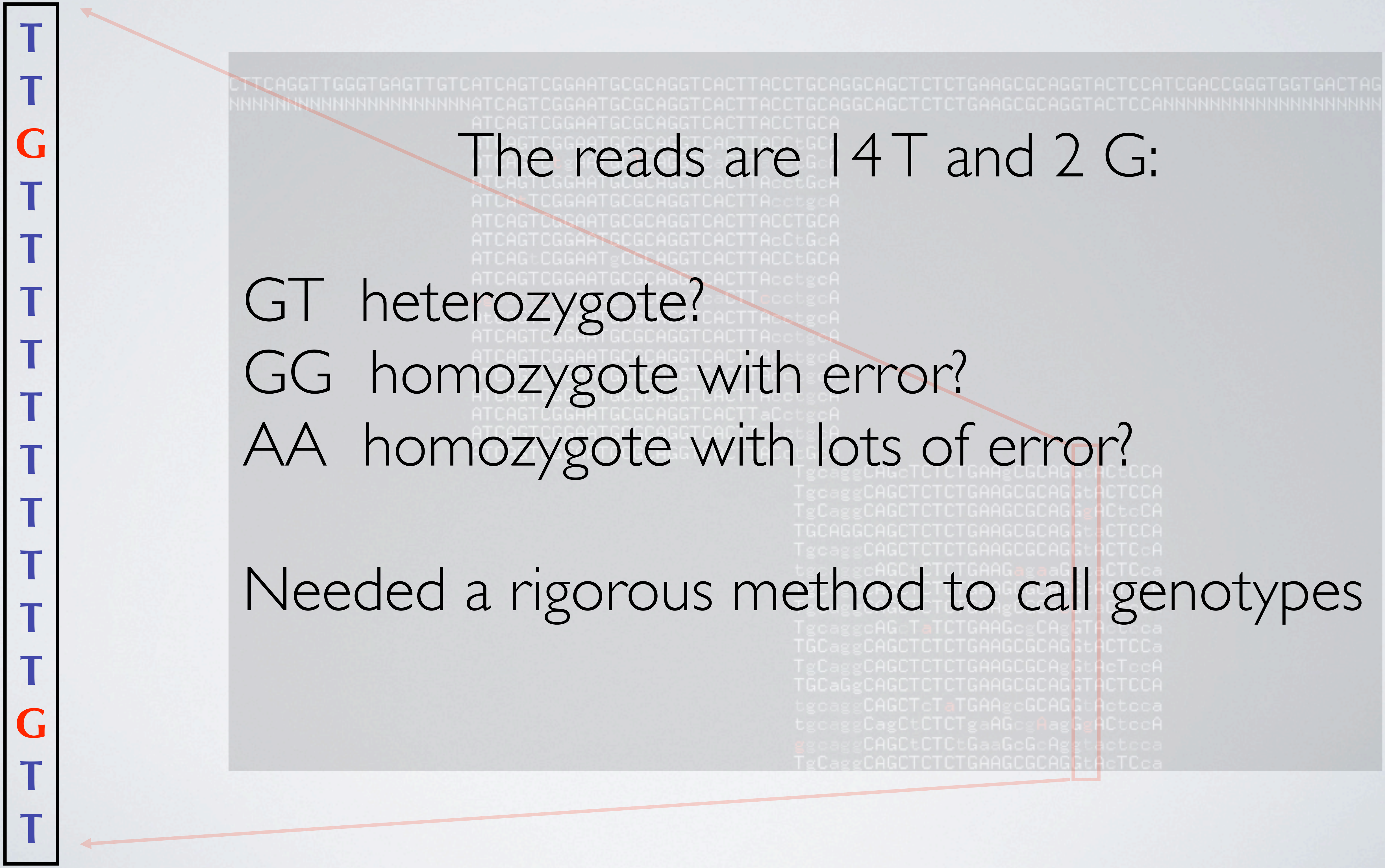
sequence reads

Differentiating SNPs from error



```
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



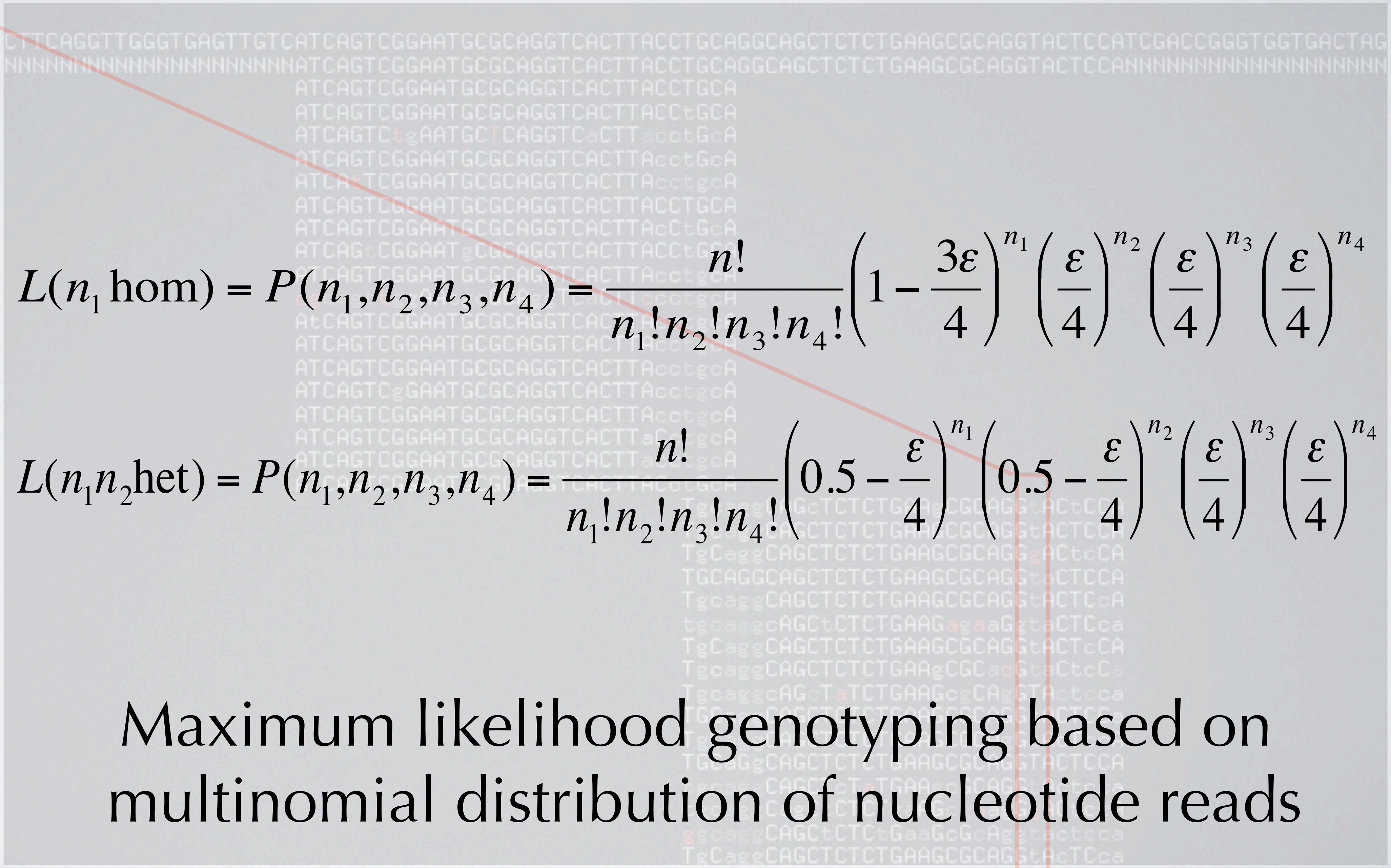
The reads are 14 T and 2 G:

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
G
T
T



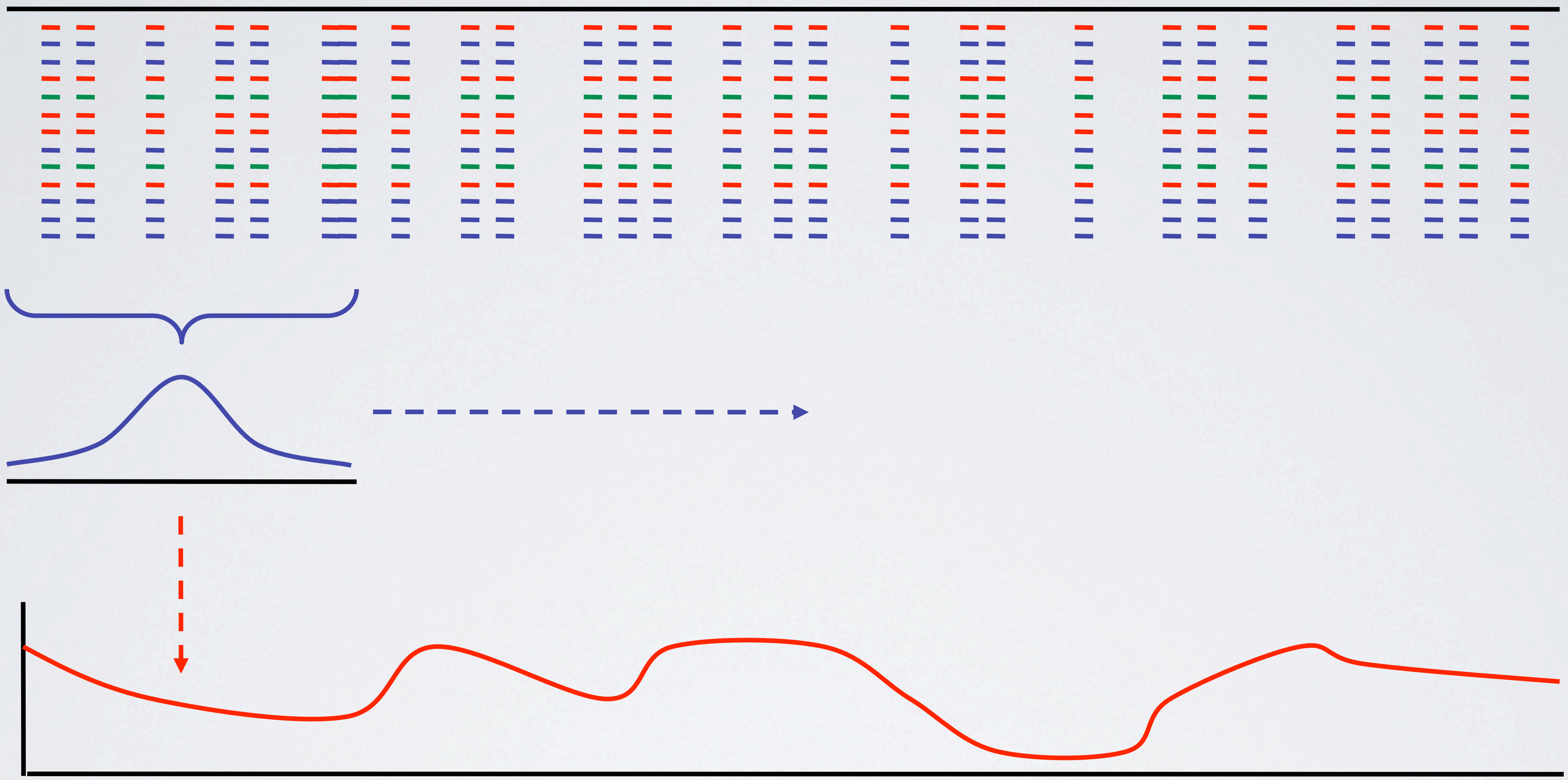
$$L(n_1\,\mathrm{hom}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

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

Maximum likelihood genotyping based on multinomial distribution of nucleotide reads

Making statistics continuous across the genome

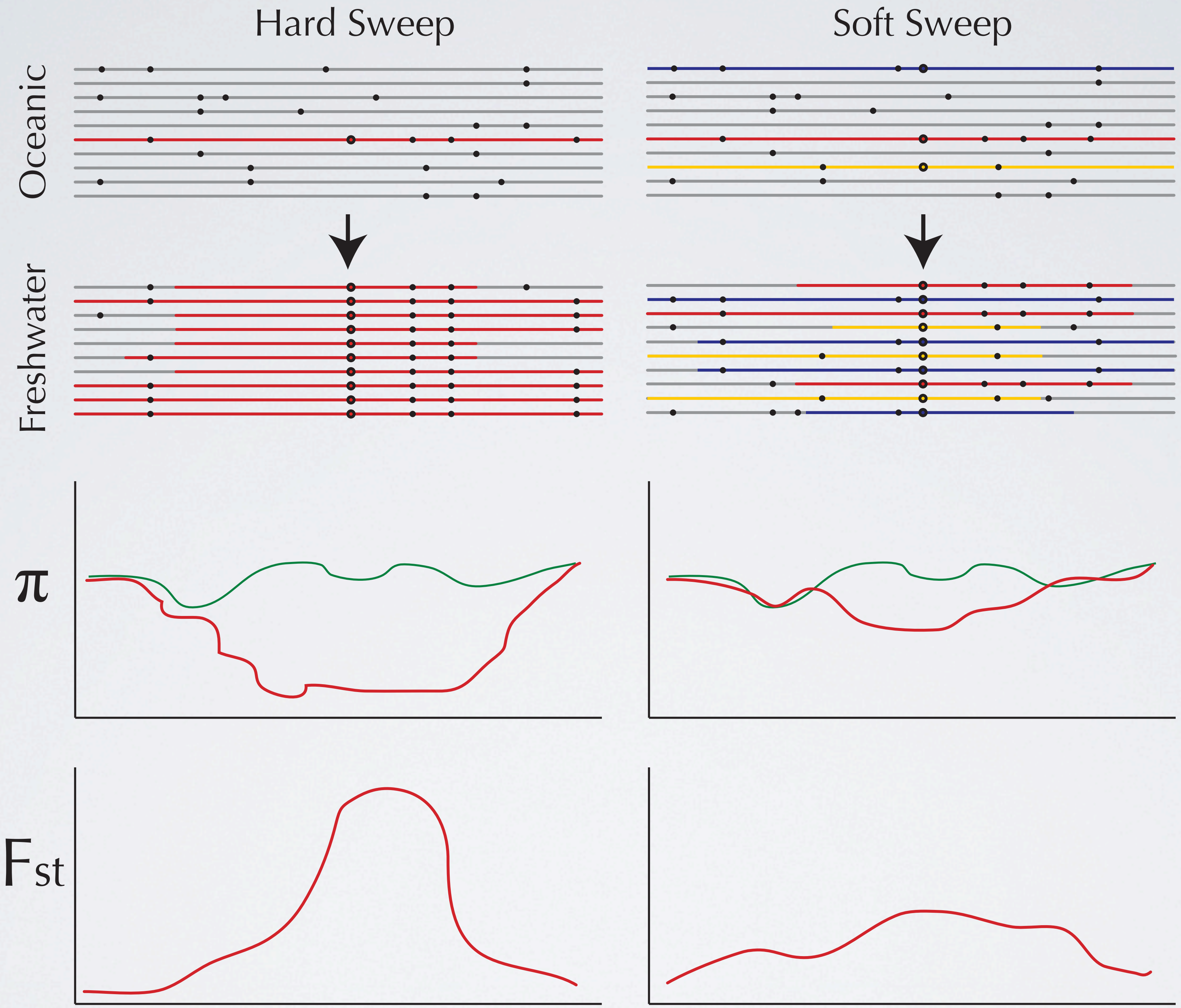
Kernel-smoothing average of summary statistics along genome



Bootstrap re-sampling to estimate significance of moving average

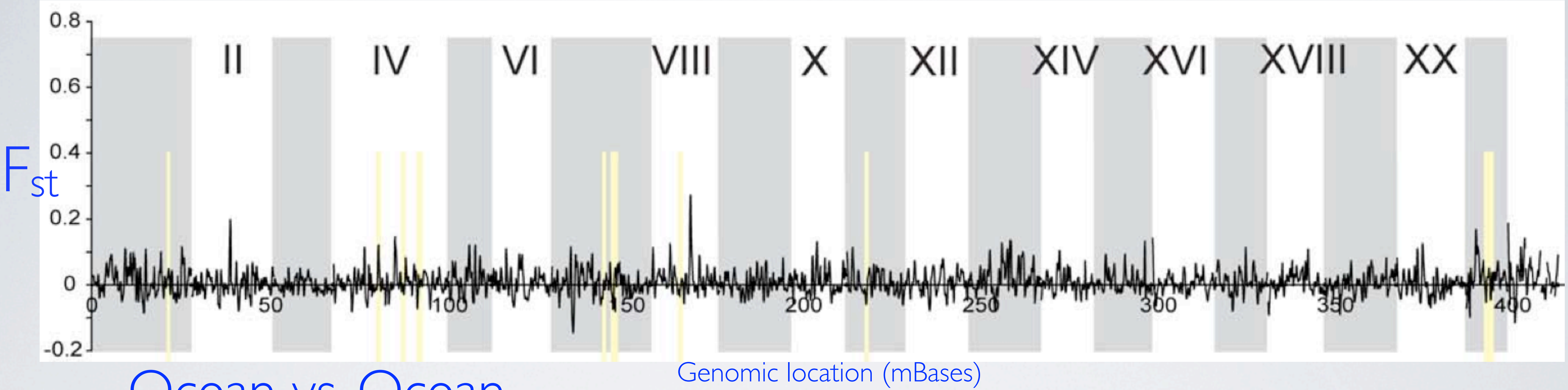
- Re-sample from across genome within the same population
- Significance is relative to genome-wide average, not based on a null model

Signatures of natural selection across the genome



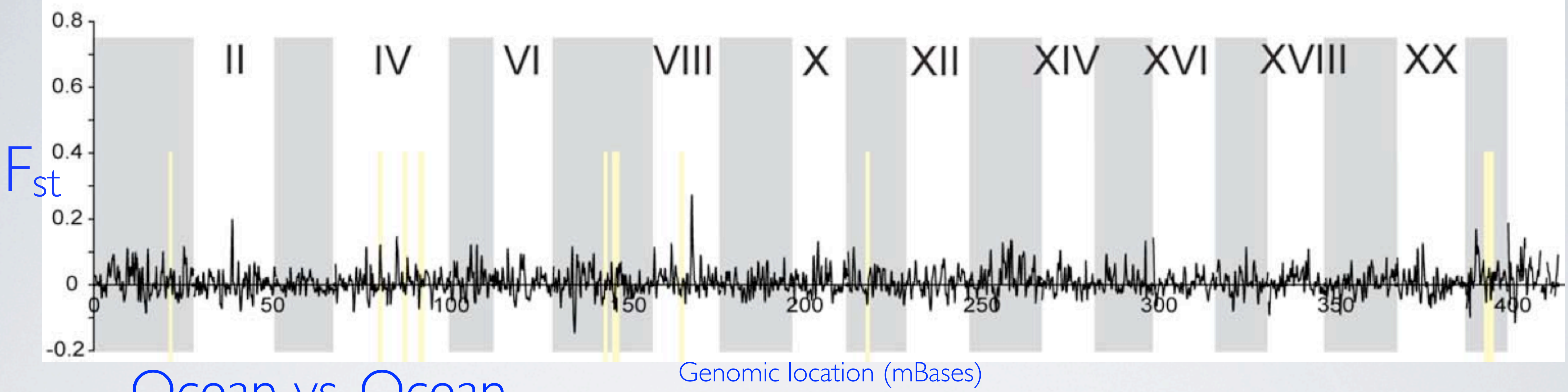
requires a very dense set of markers or complete sequences
assays need to be performed in dozens or hundreds of individuals
hard to develop for non-model organisms

Signatures of natural selection across the genome

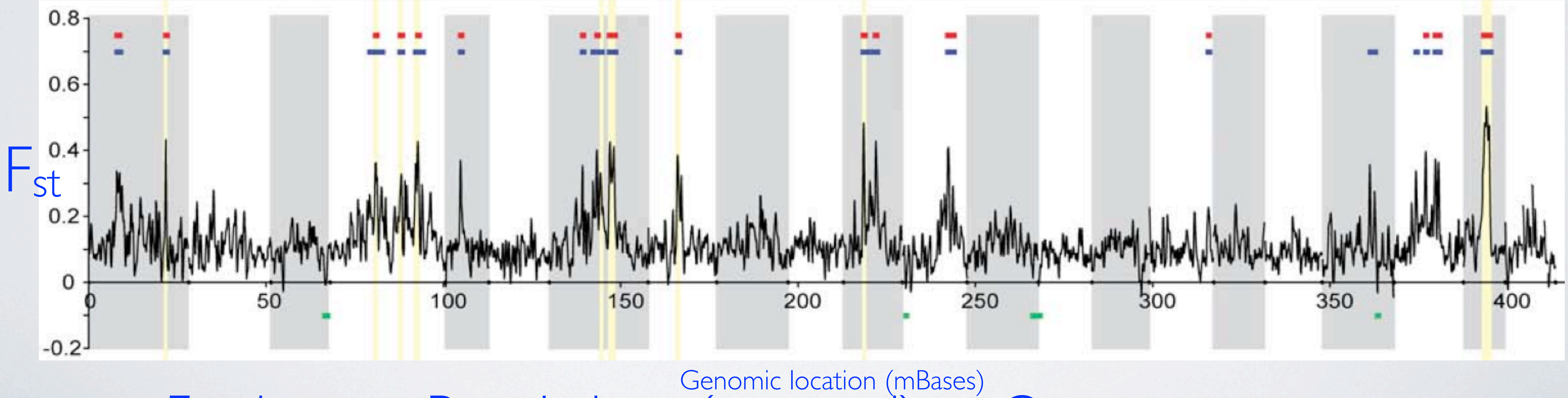


Ocean vs. Ocean

Signatures of natural selection across the genome

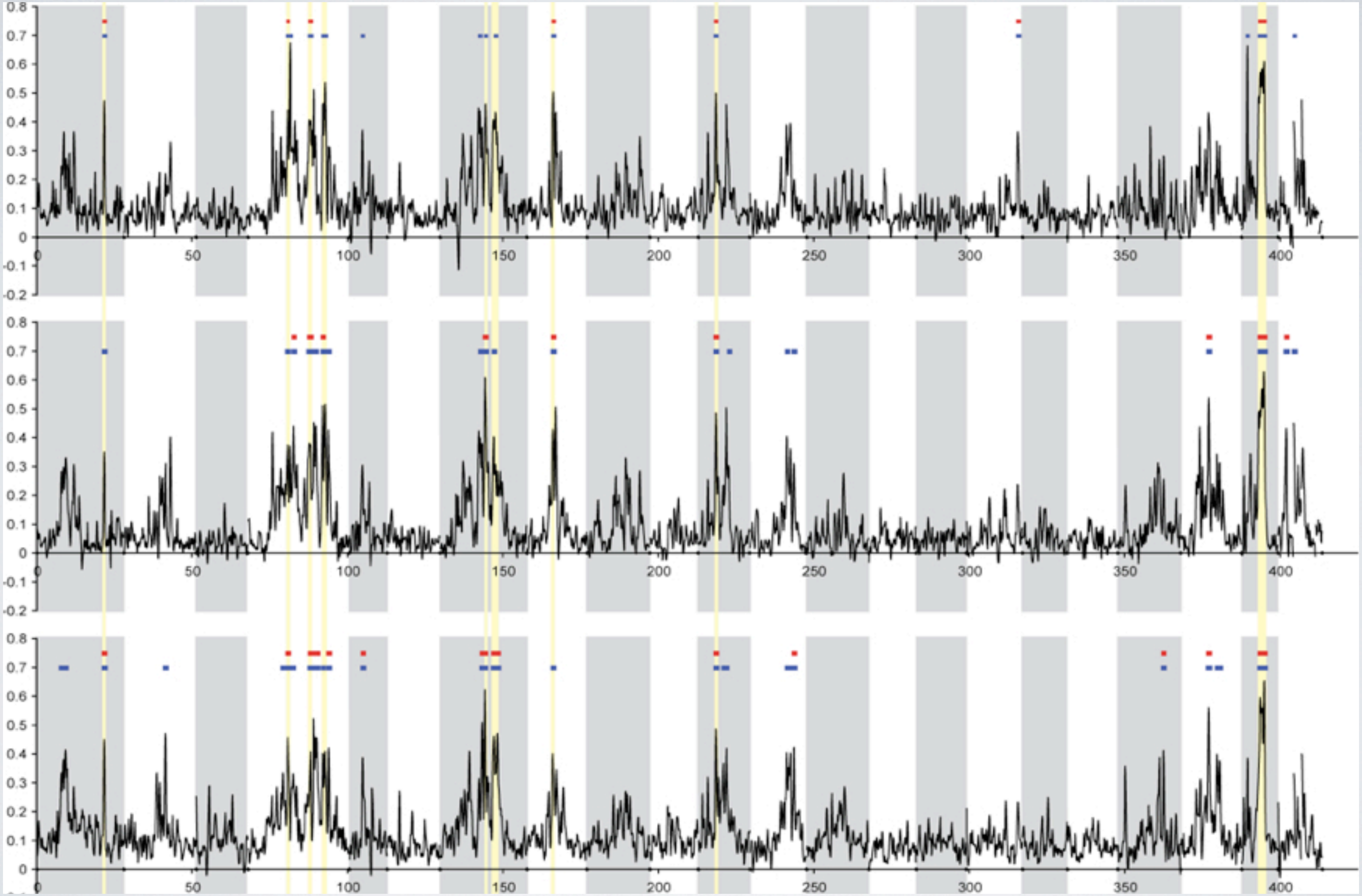


Ocean vs. Ocean



Freshwater Populations (grouped) vs. Ocean

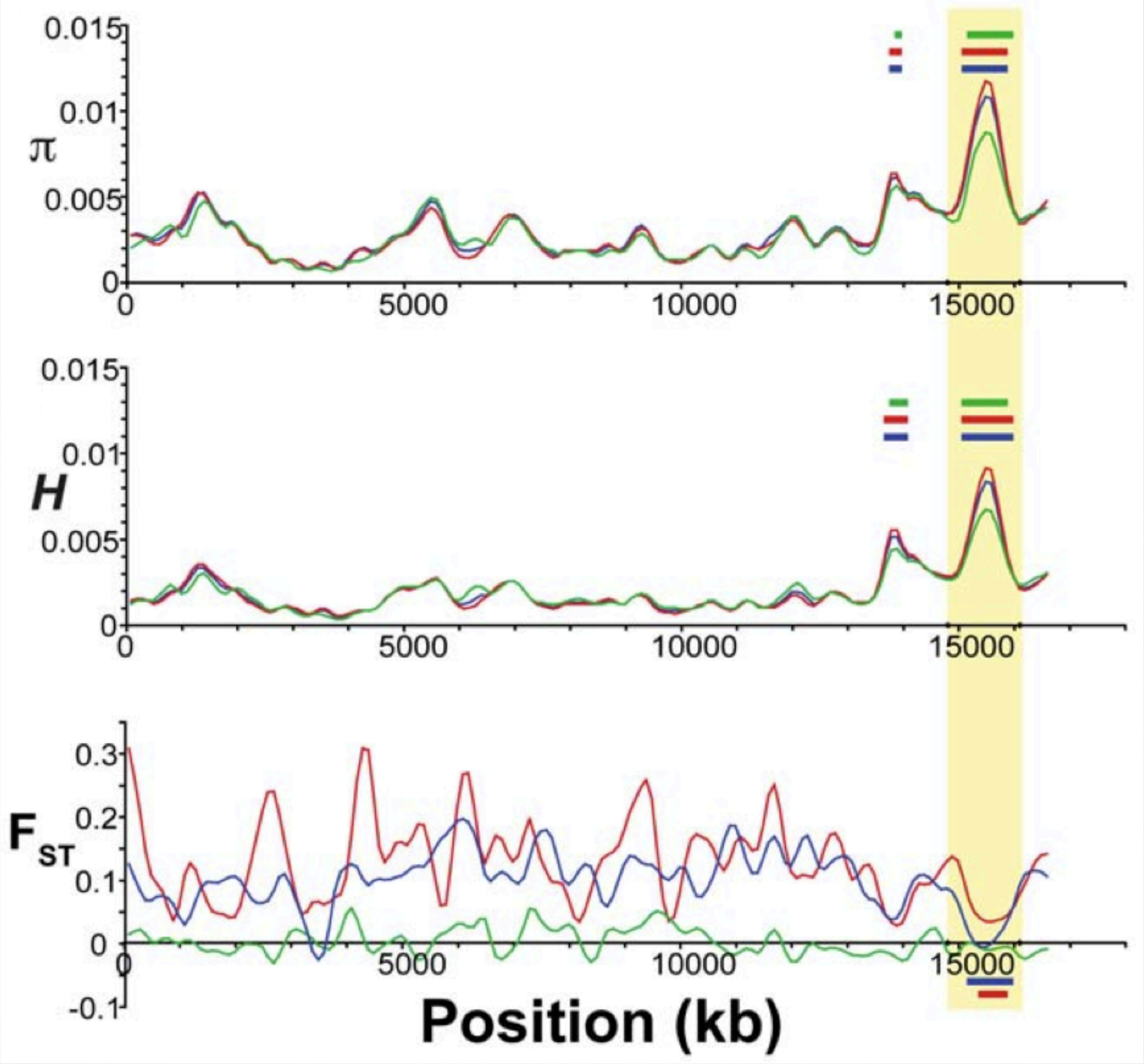
Consistent pattern of parallel selection



F_{st}

Genomic location (mBases)

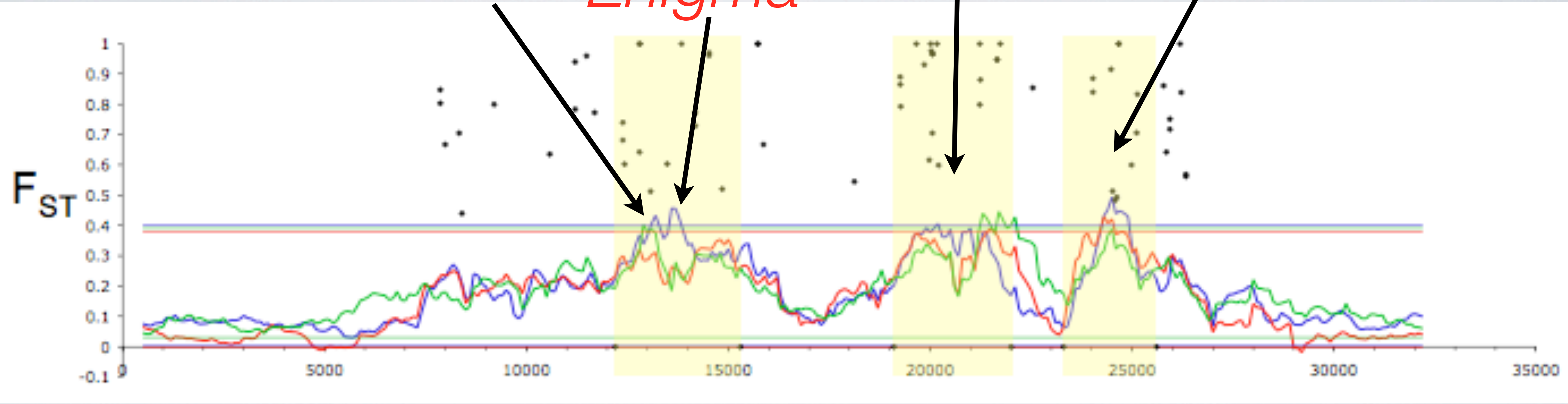
Evidence for balancing selection on Linkage Group III



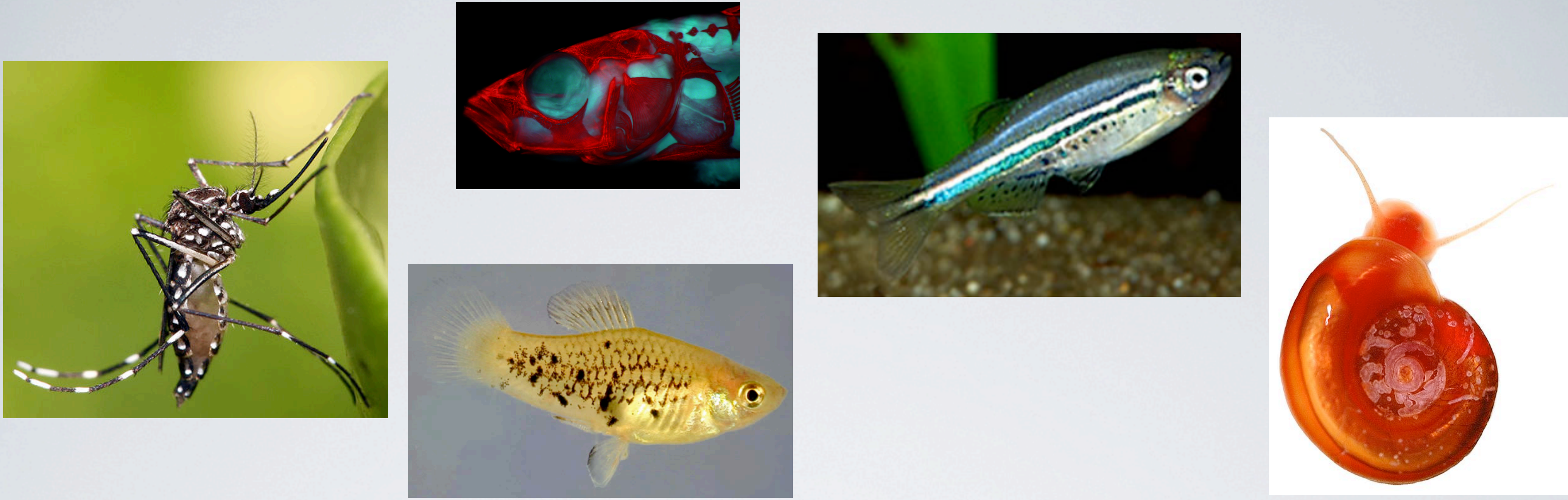
Previously identify quantitative trait loci (QTLs) are under selection

Natural populations

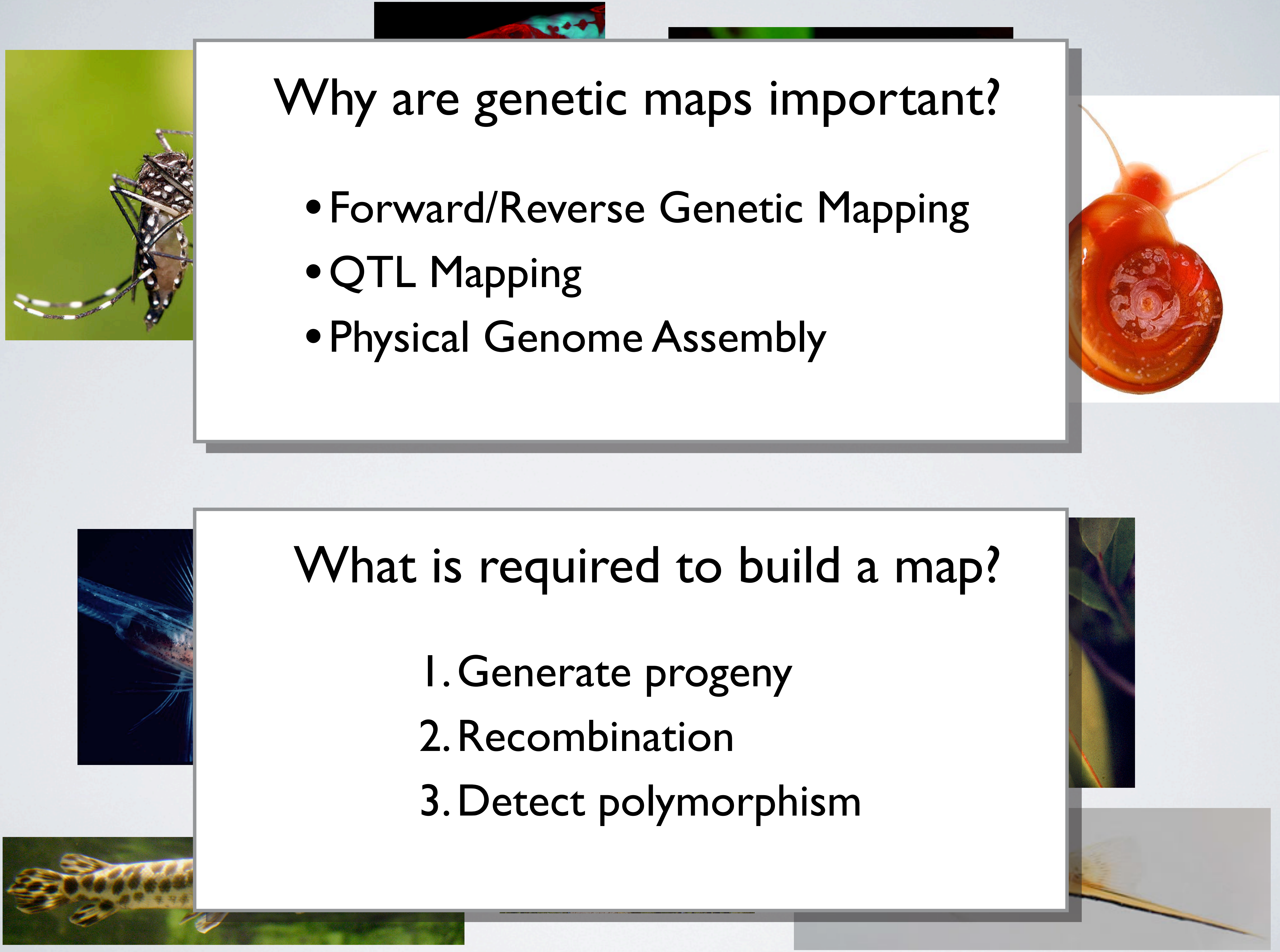
Eda *Enigma* *Foxi3b* *FGFR*



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

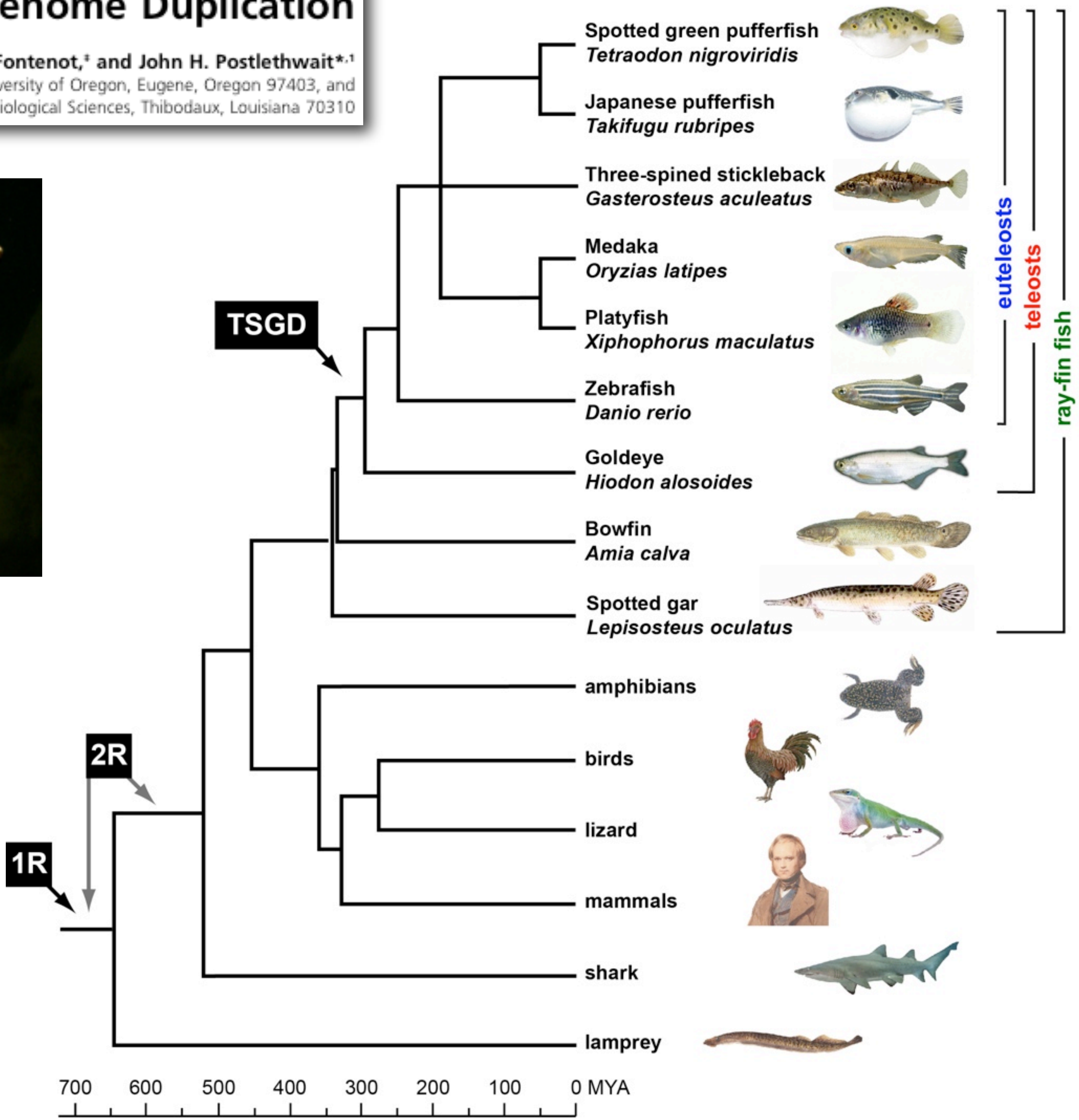
Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication

Angel Amores,* Julian Catchen,† Allyse Ferrara,‡ Quenton Fontenot,‡ and John H. Postlethwait*,1
*Institute of Neuroscience and †Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, and
‡Nicholls State University, Department of Biological Sciences, Thibodaux, Louisiana 70310

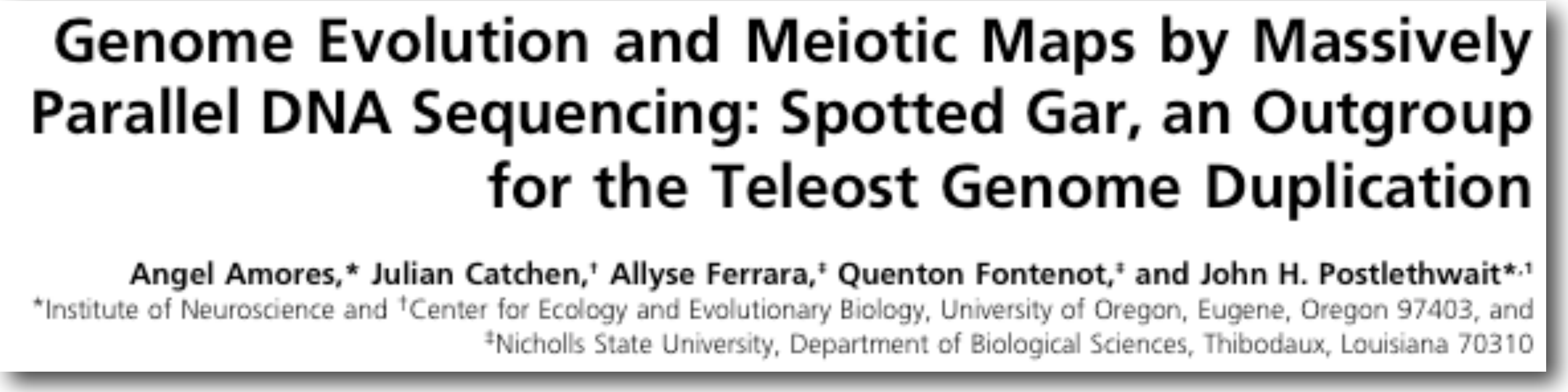
Spotted Gar

Angel Amores

John Postlethwait



Genetics



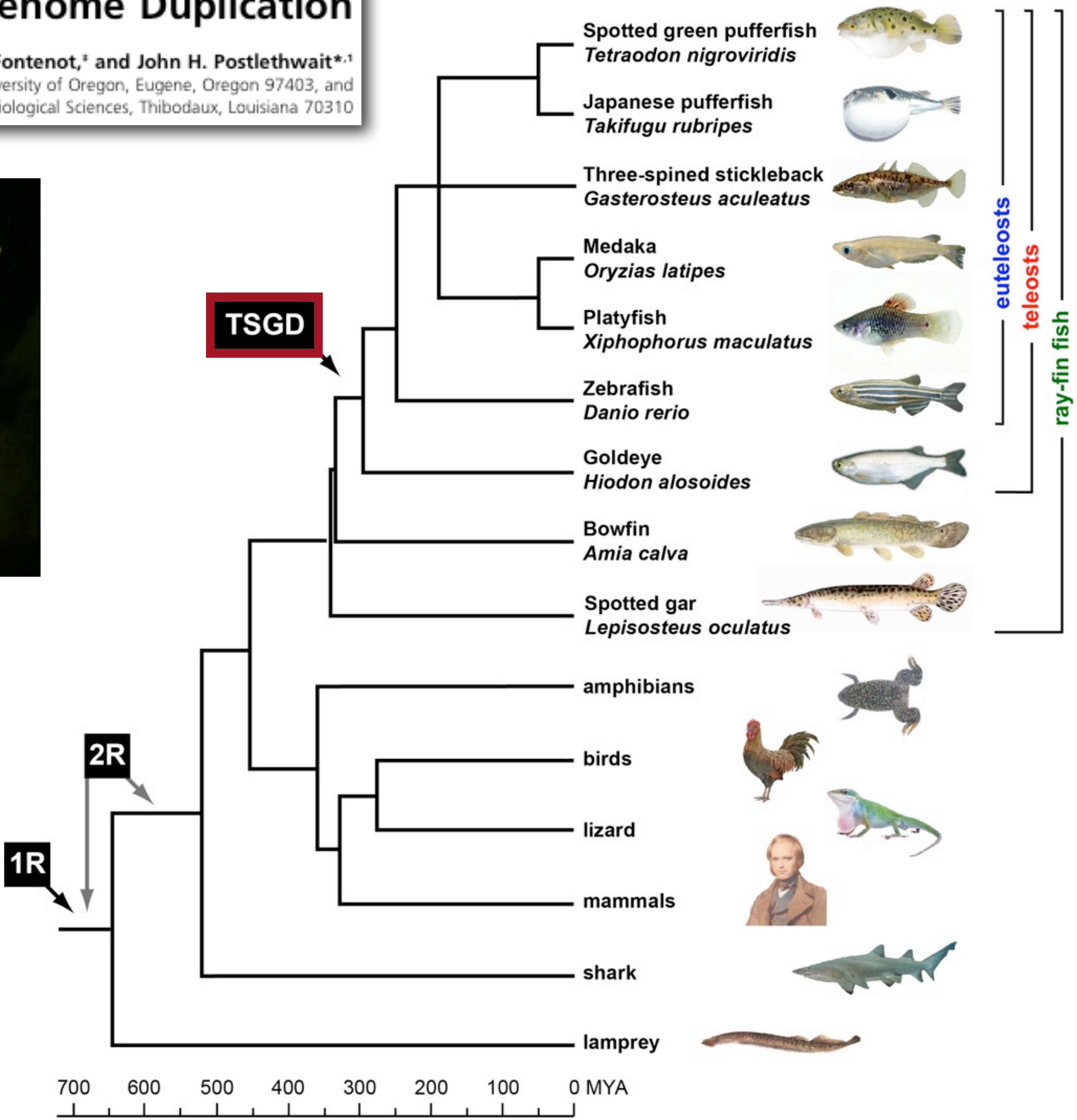
Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication

Angel Amores,[*] Julian Catchen,[†] Allyse Ferrara,[‡] Quenton Fontenot,[‡] and John H. Postlethwait[*,1]
[*]Institute of Neuroscience and [†]Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, and
[‡]Nicholls State University, Department of Biological Sciences, Thibodaux, Louisiana 70310

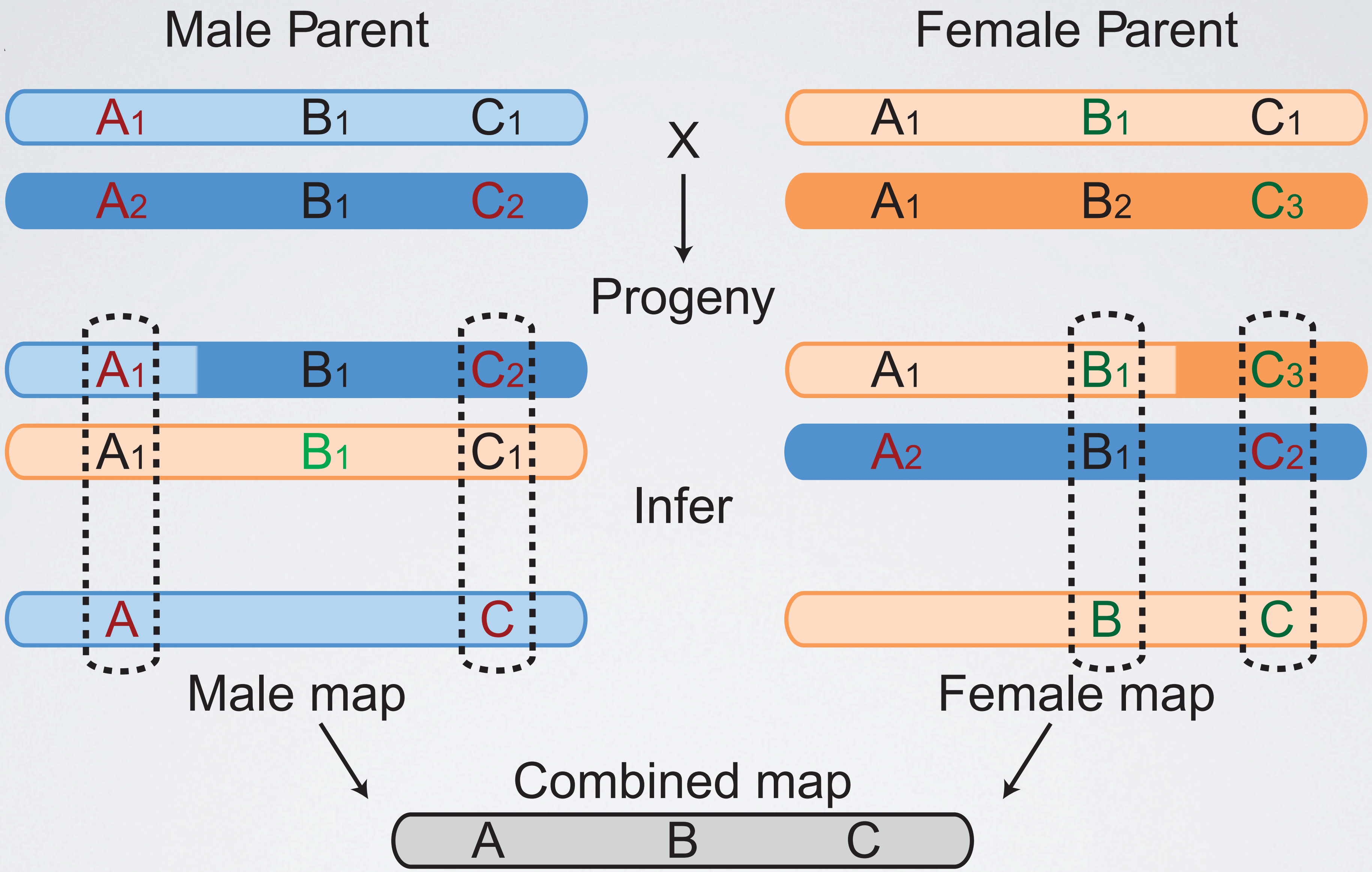
Spotted Gar

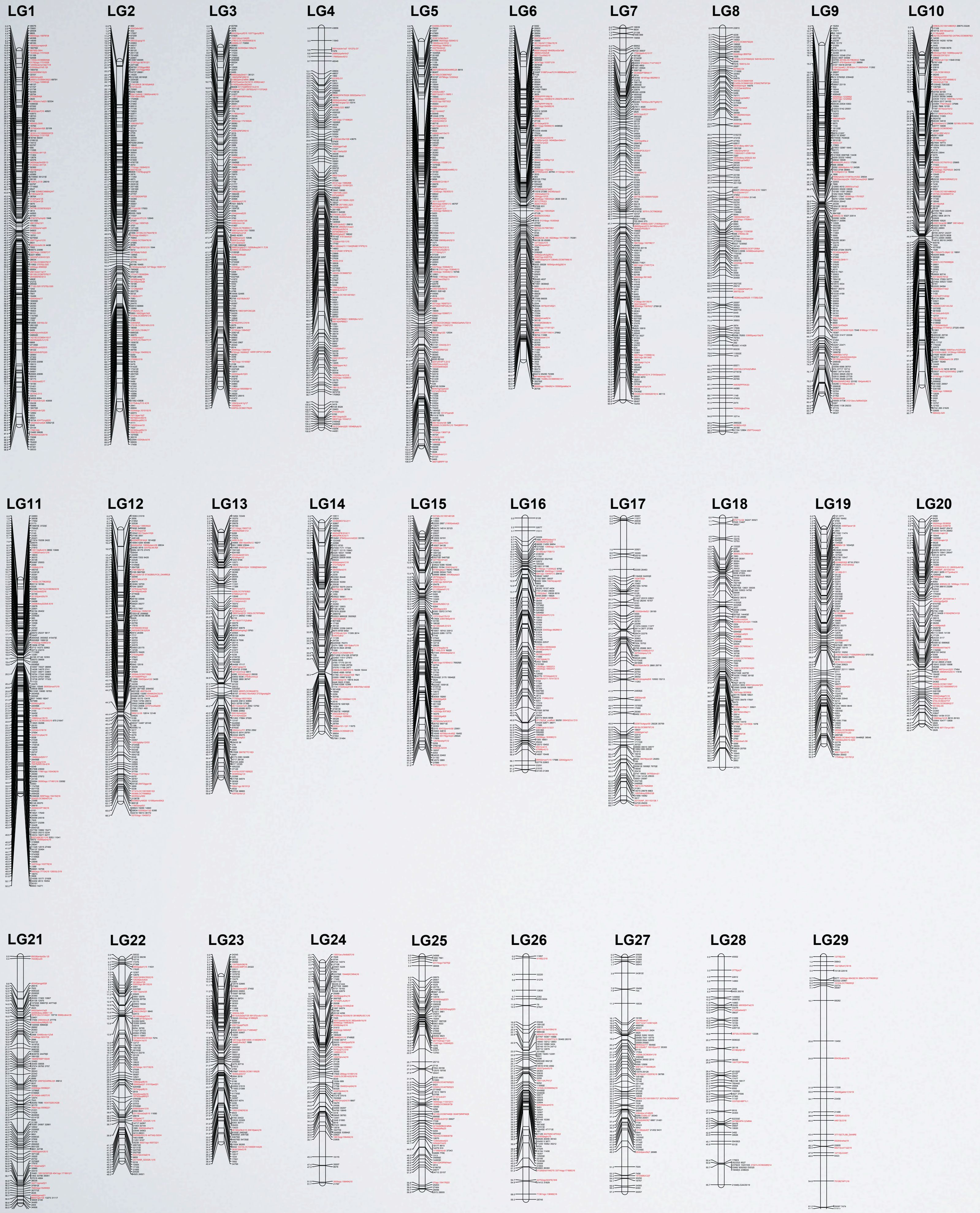
Angel Amores

John Postlethwait

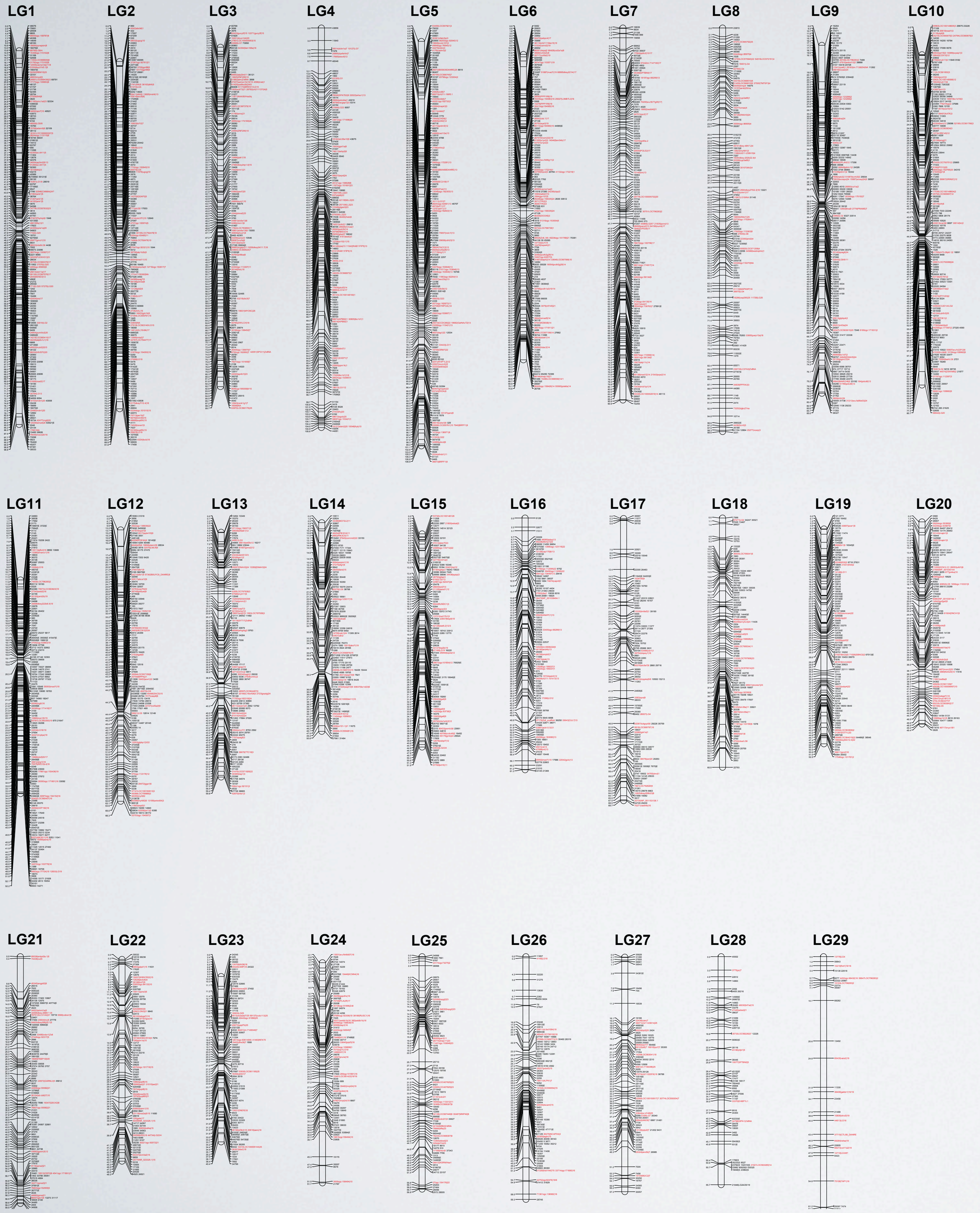


F1 Pseudo-Test Cross



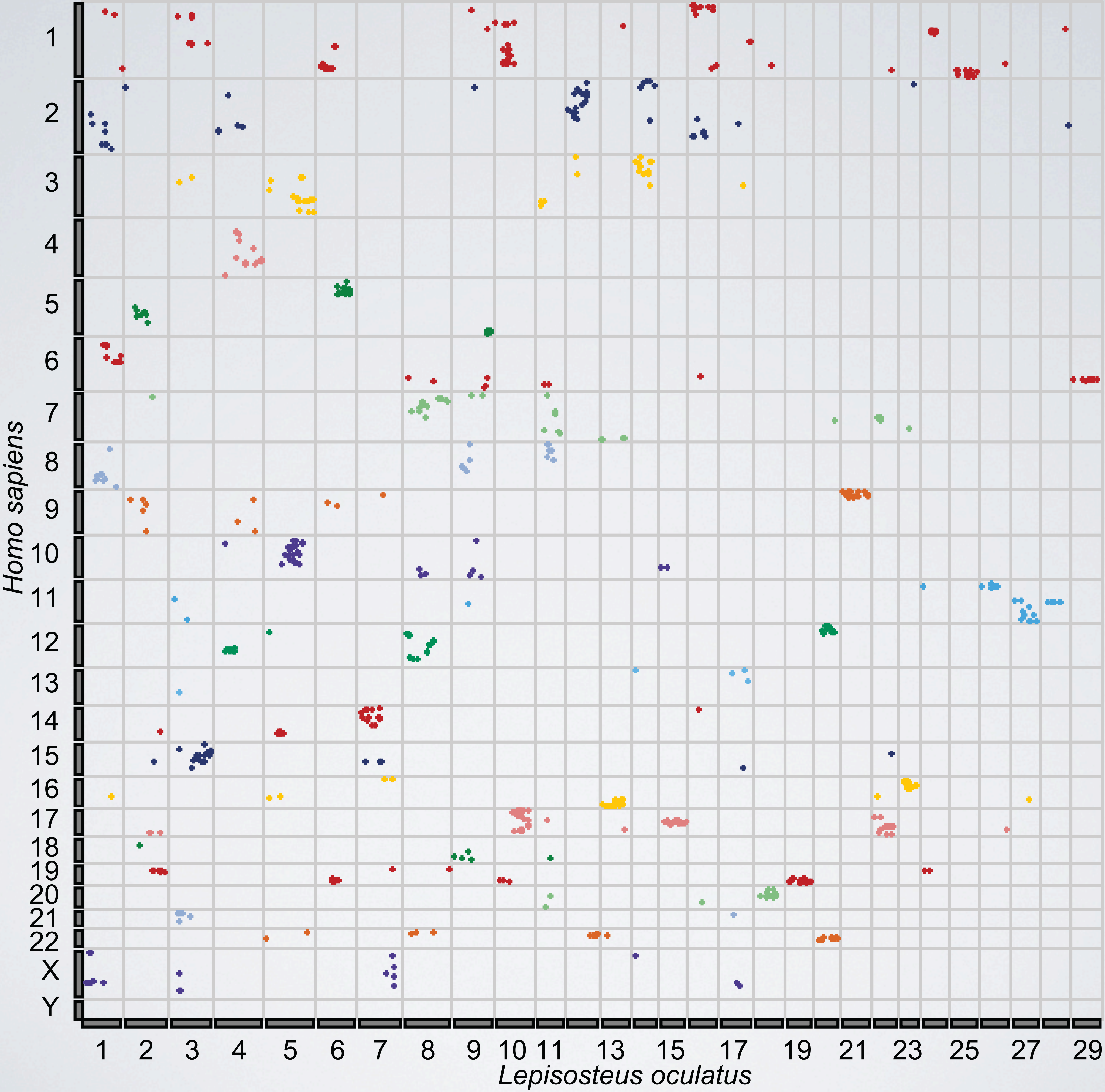


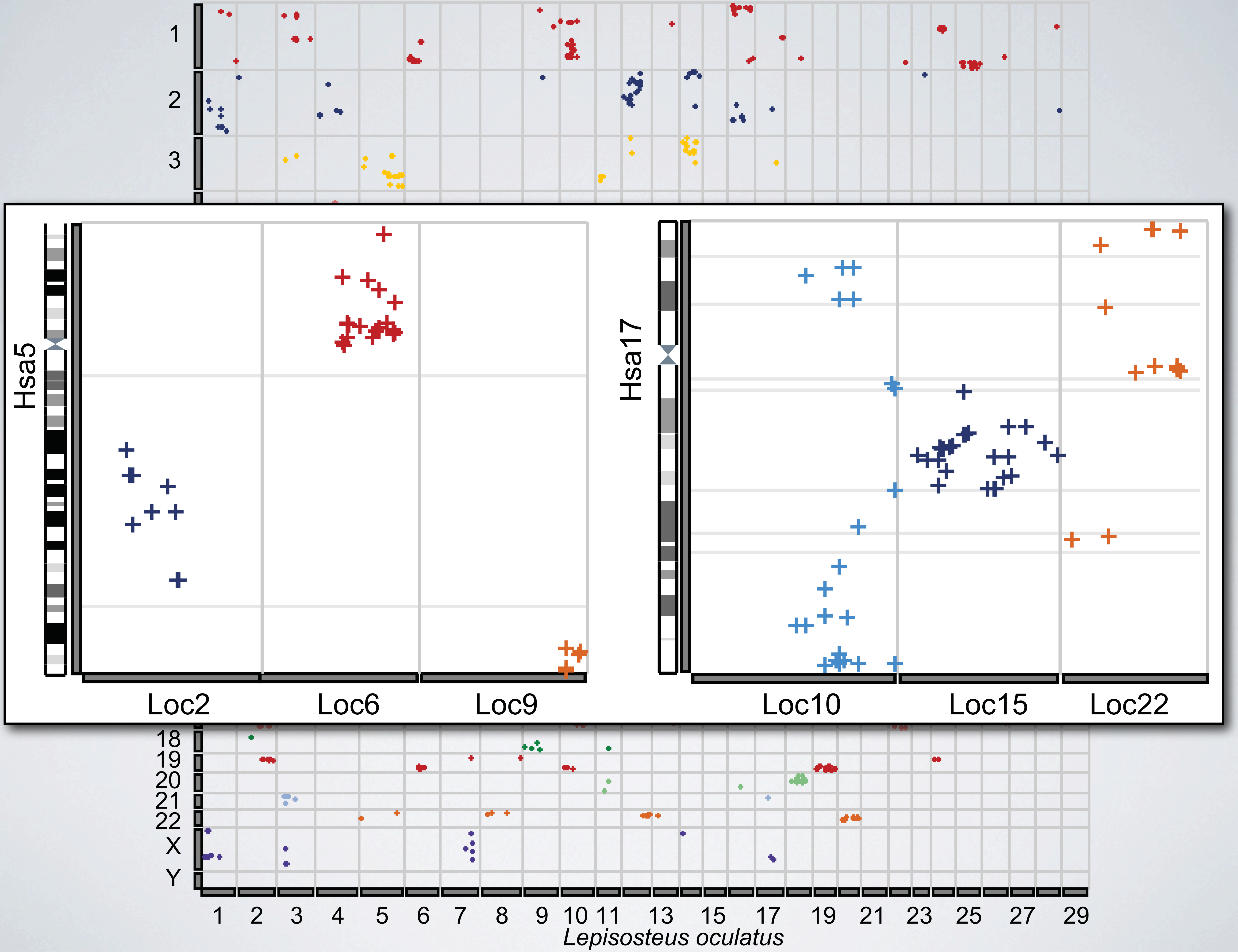
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

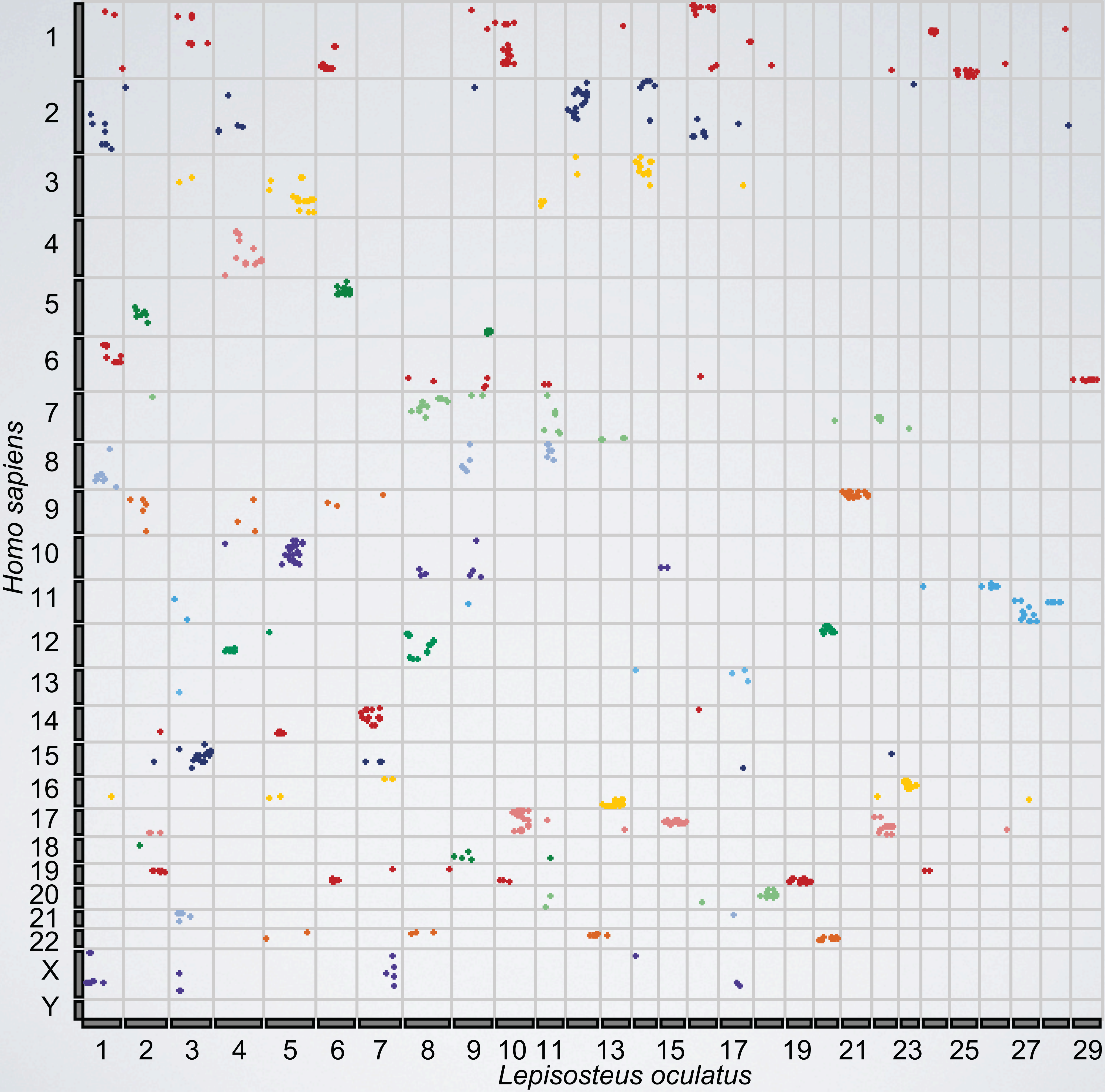


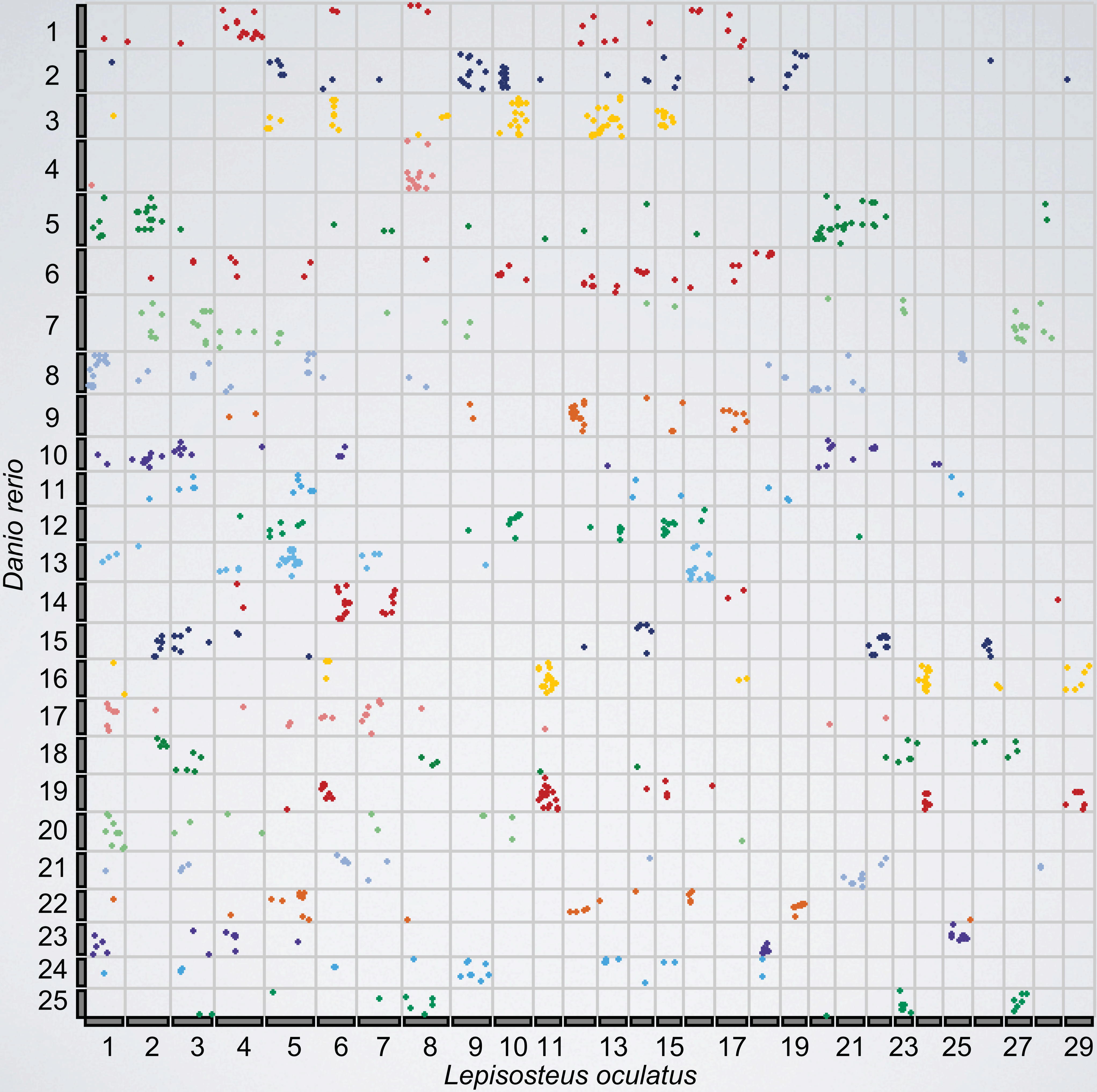
94 Individuals
15,076 Markers
8,046 Mapped
974 In Genes

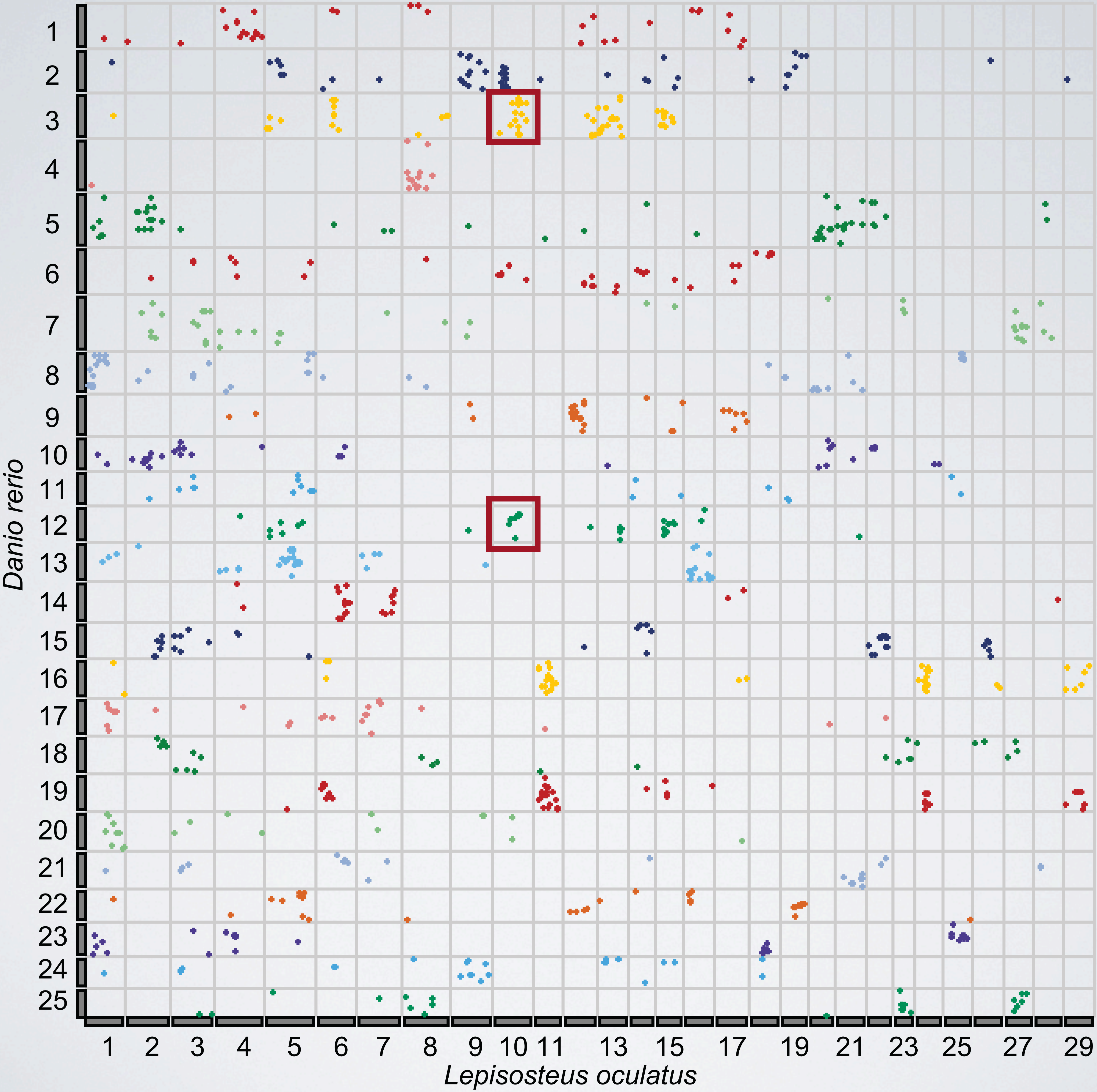
Organism	Markers
Silver carp	483
Guppy	790
Barramundi	240
Catfish	331
Sea bass	368
Cichlid	204
Platyfish	290
Halibut	604
Sea bream	204

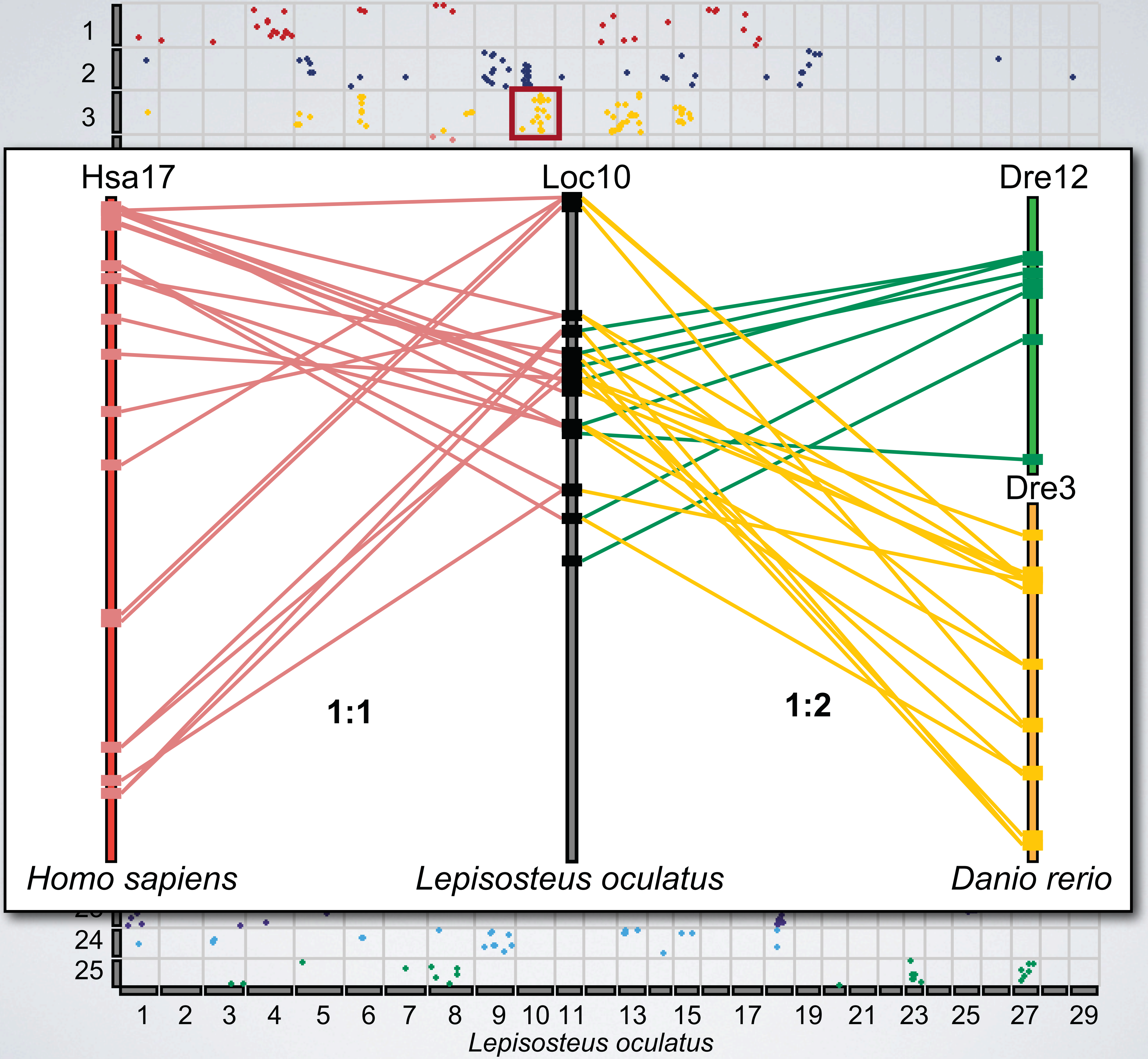

Homo sapiens
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
1
2
3
4
5
6
7
8
9
10
11
13
15
17
19
21
23
25
27
29
Lepisosteus oculatus

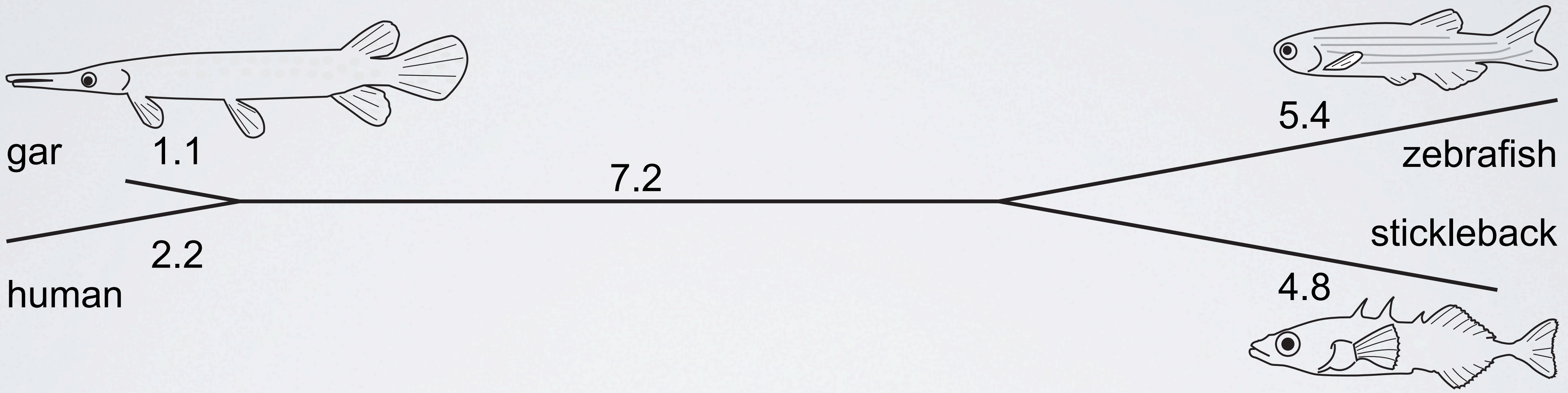

1
2
3
Hsa5
Loc2
Loc6
Loc9
Hsa17
Loc10
Loc15
Loc22
18
19
20
21
22
X
Y
1
2
3
4
5
6
7
8
9
10
11
13
15
17
19
21
23
25
27
29
Lepisosteus oculatus


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


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


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


gar
1.1
human
2.2
7.2
5.4
zebrafish
stickleback
4.8

Stacks

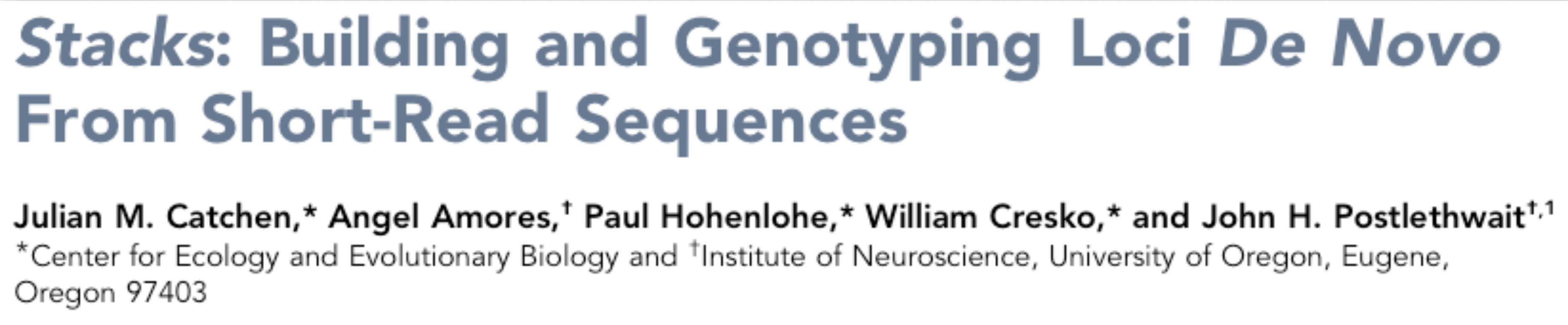
Stacks is a software pipeline for building loci out of a set of short-read sequenced samples. Stacks was developed for the purpose of building genetic maps from RAD-Tag Illumina sequence data, but can also be readily applied to population studies, and phylogeography.

Download Stacks

Version 0.982

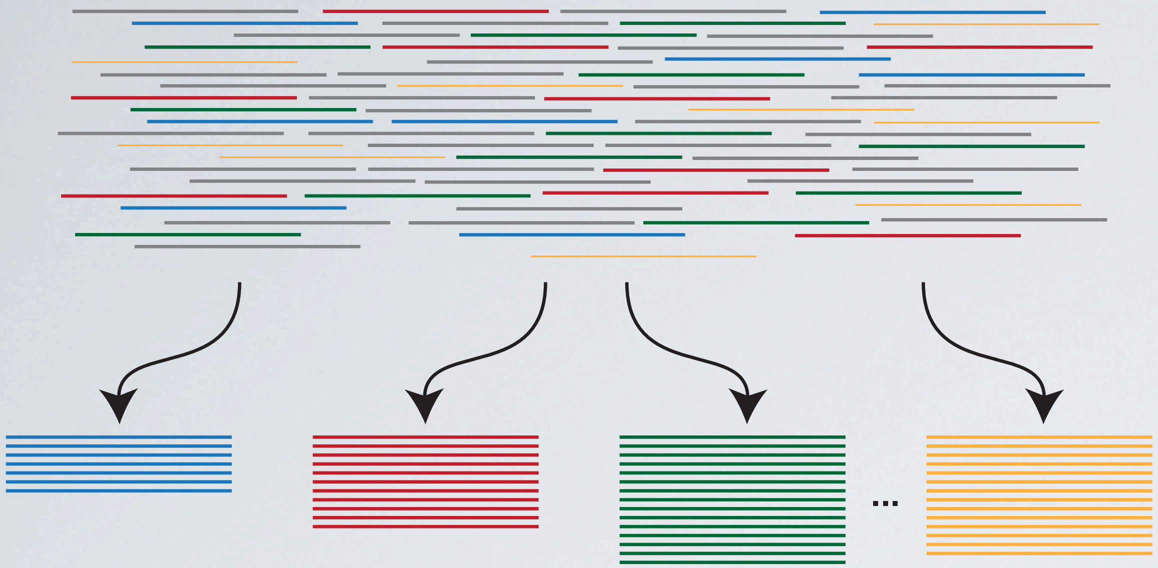
Recent Changes [updated Mar 29, 2011]

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



Stack 1

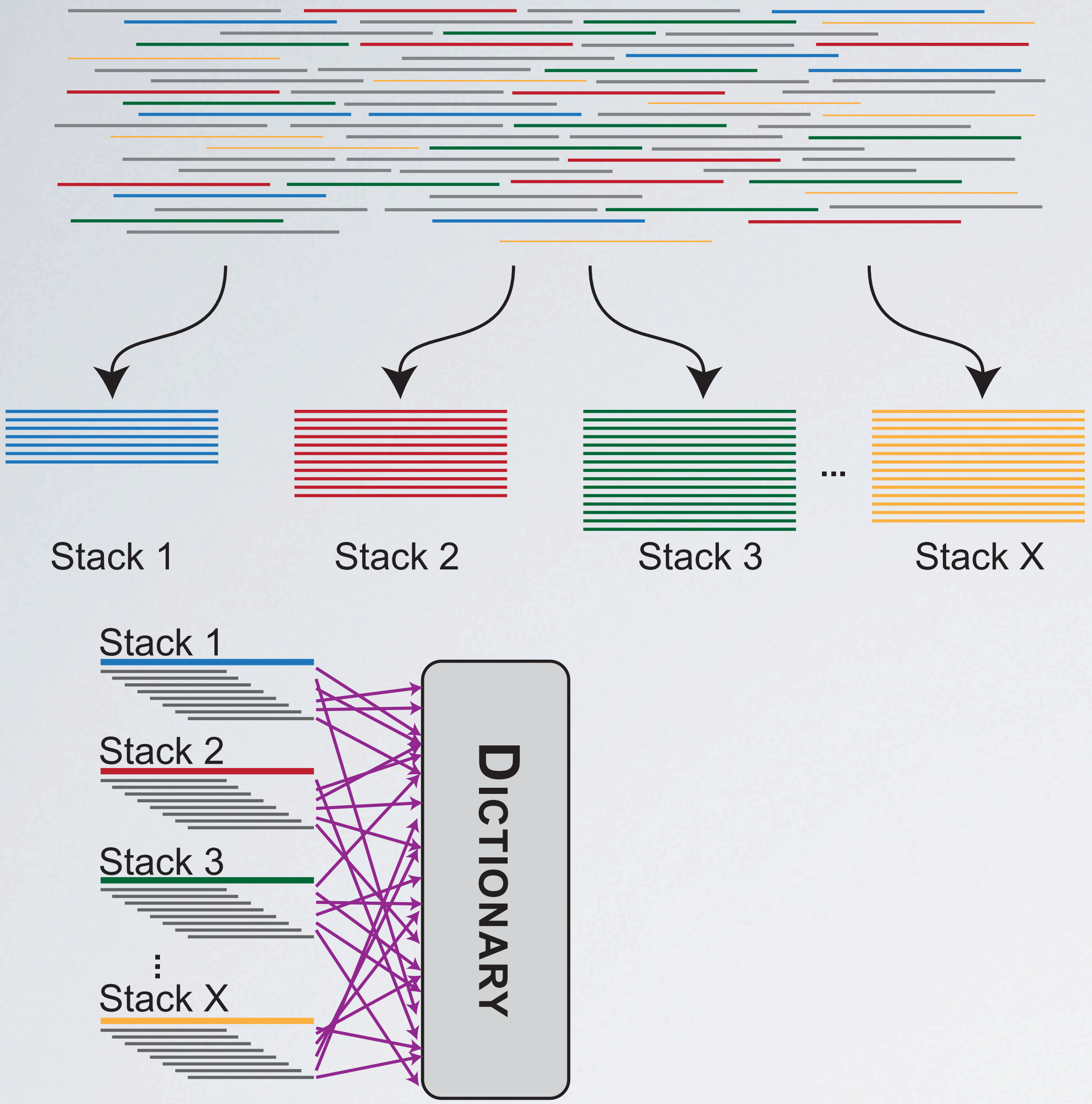
Stack 2

Stack 3

Stack X

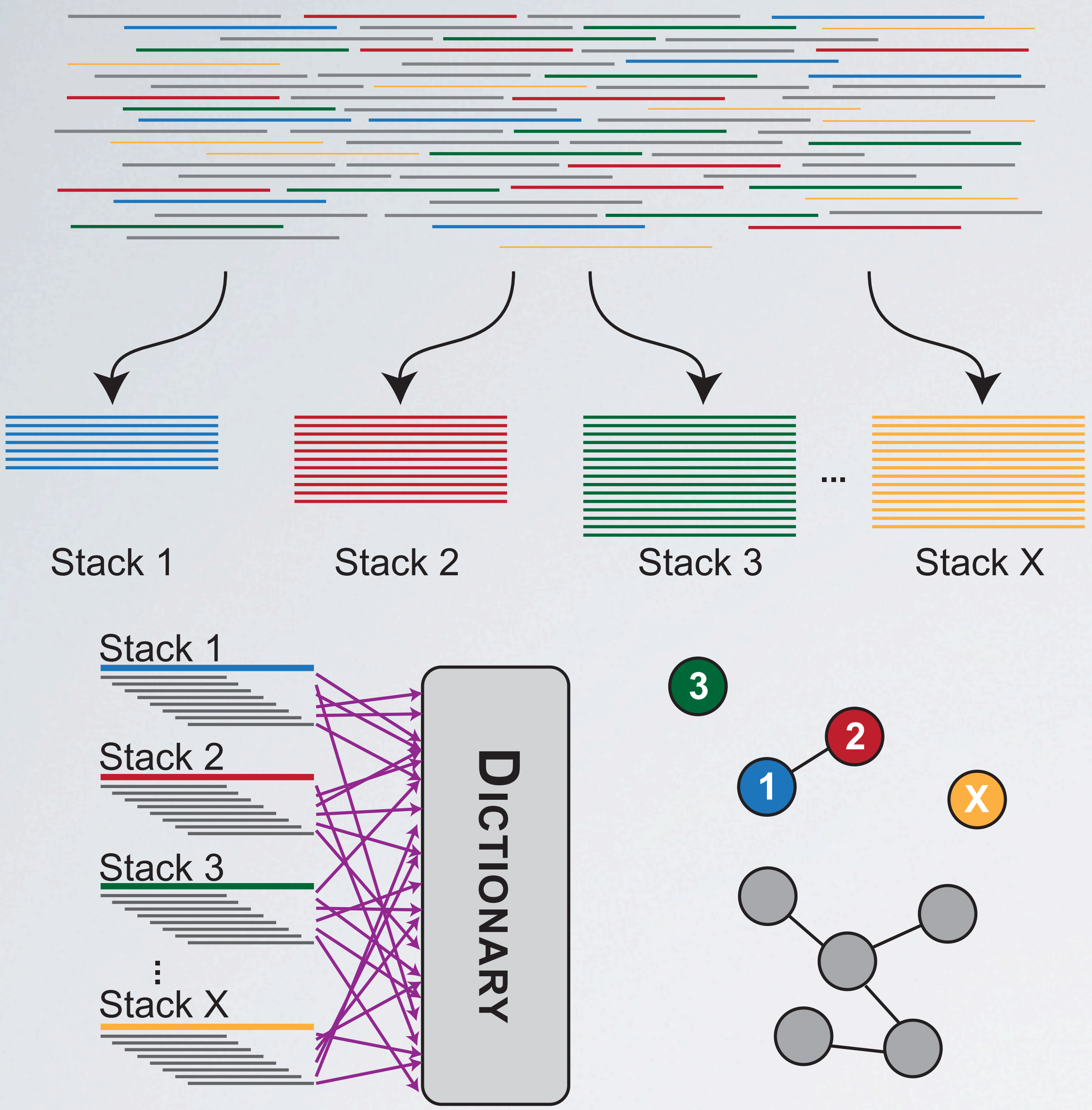
C

CT


Stack 1
Stack 2
Stack 3
...
Stack X
Stack 1
Stack 2
Stack 3
⋮
Stack X
DICTIONARY

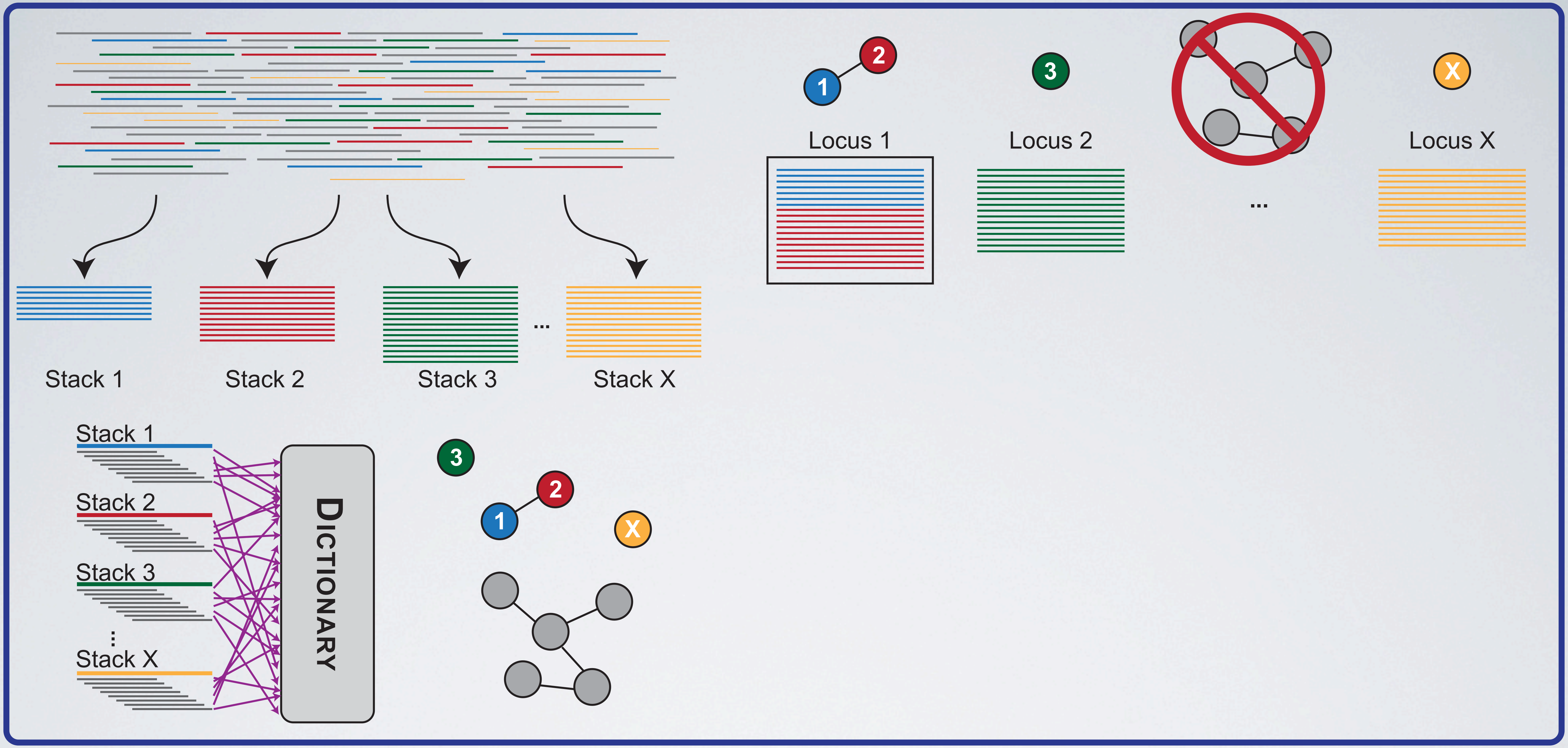
C

CT


Stack 1
Stack 2
Stack 3
...
Stack X
Stack 1
Stack 2
Stack 3
⋮
Stack X
DICTIONARY
3
2
1
X

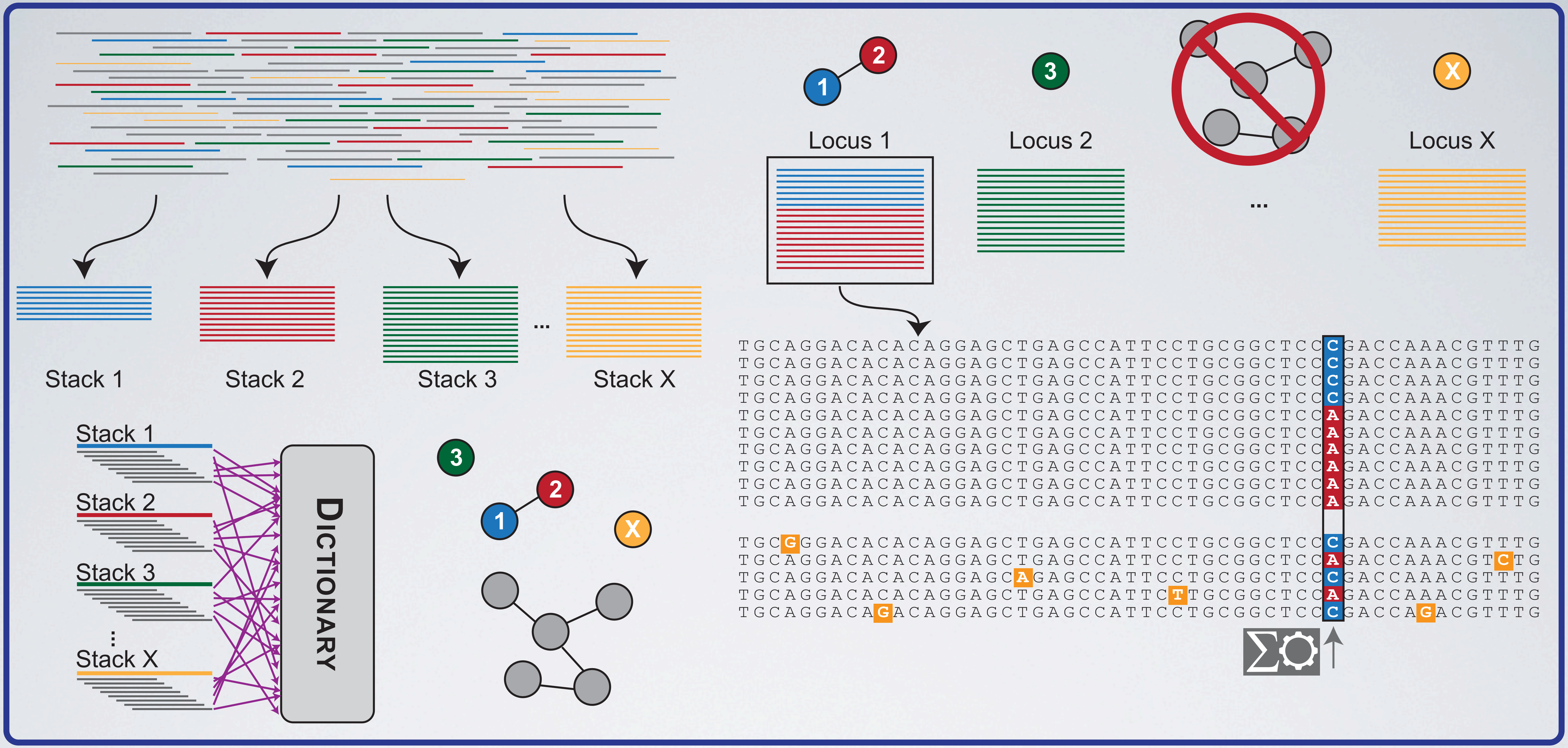
C

CT


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

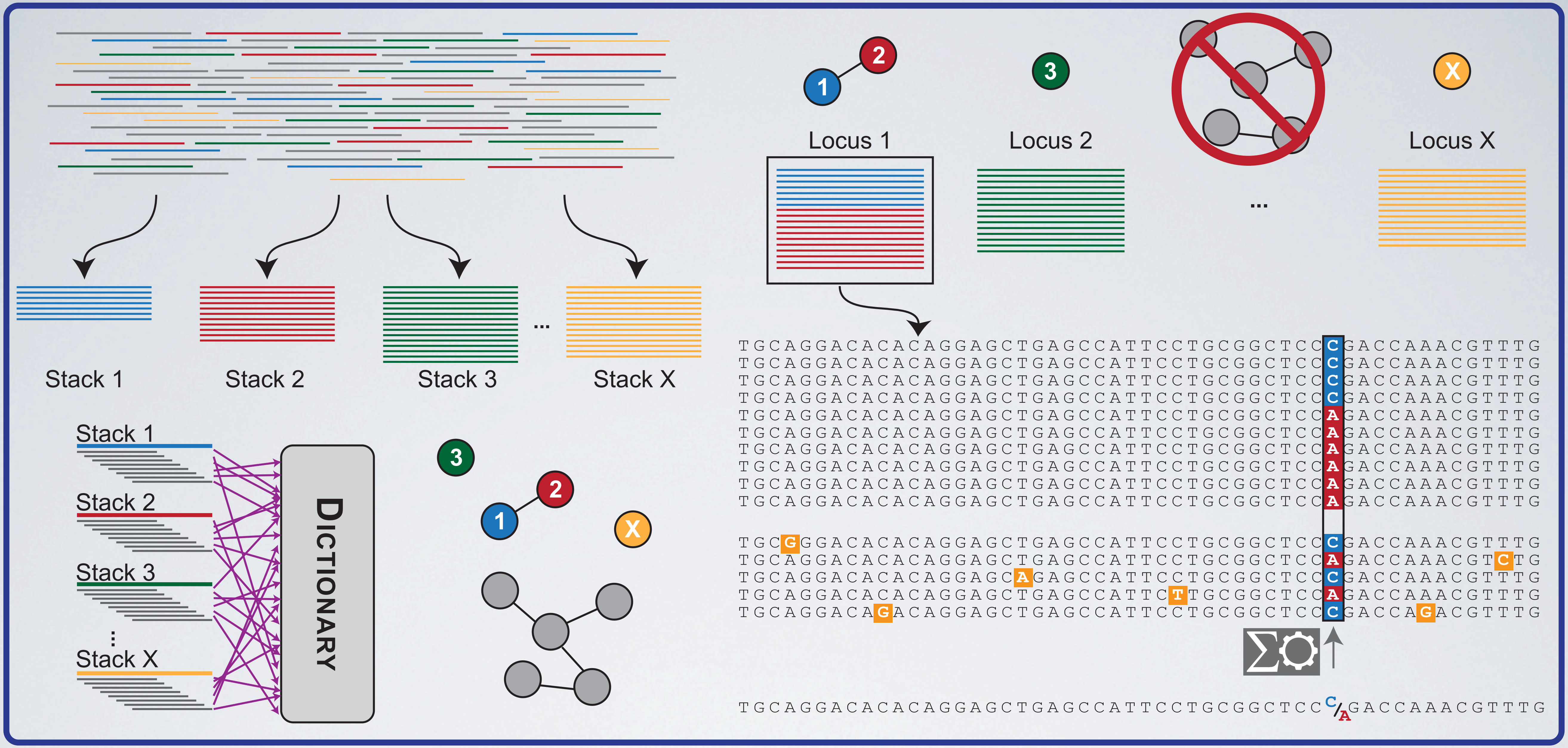
C

CT


Stack 1
Stack 2
Stack 3
...
Stack X
Locus 1
Locus 2
Locus X
1
2
3
X
DICTIONARY
Stack 1
Stack 2
Stack 3
⋮
Stack X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
Σ

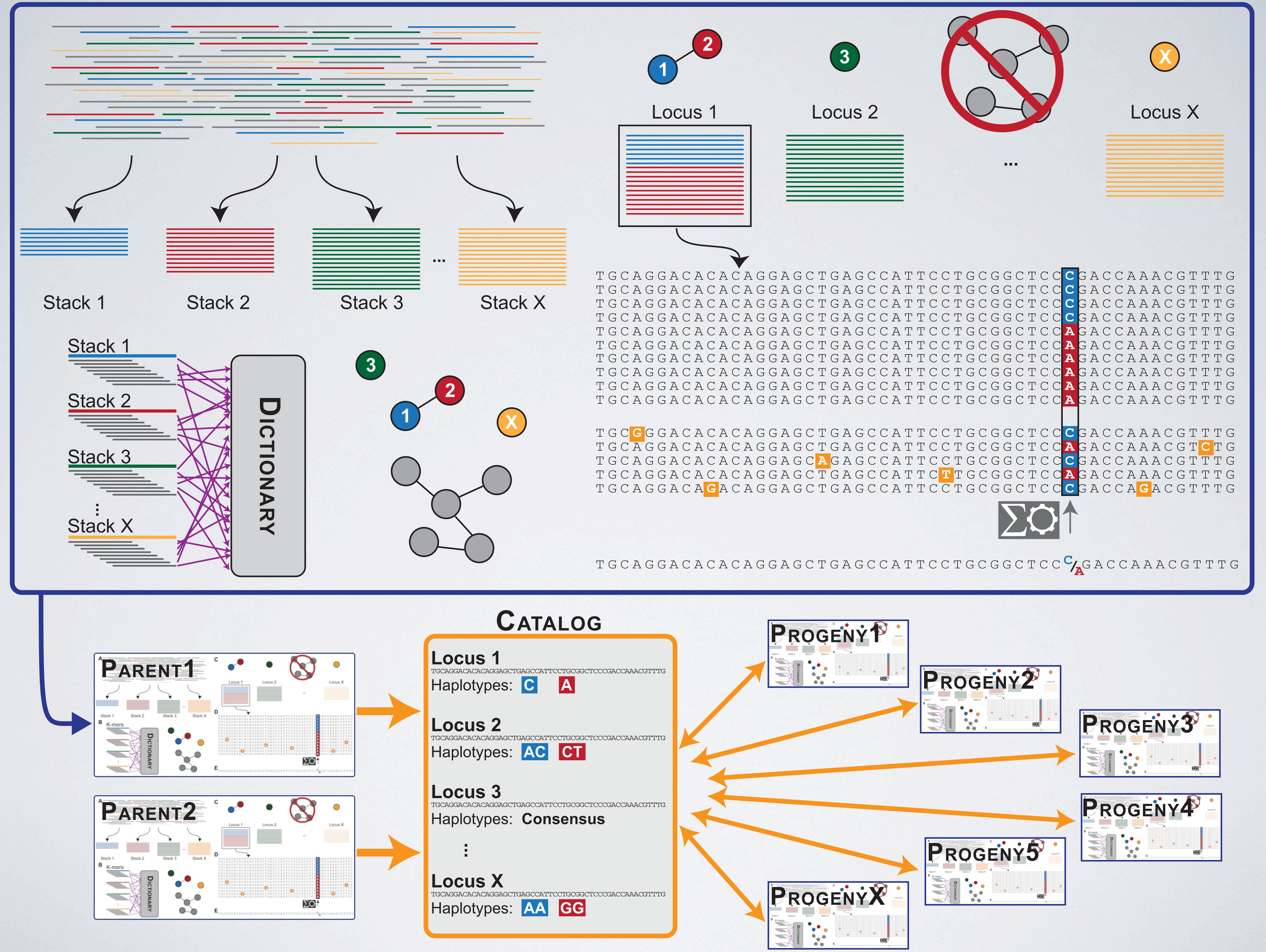
C

CT

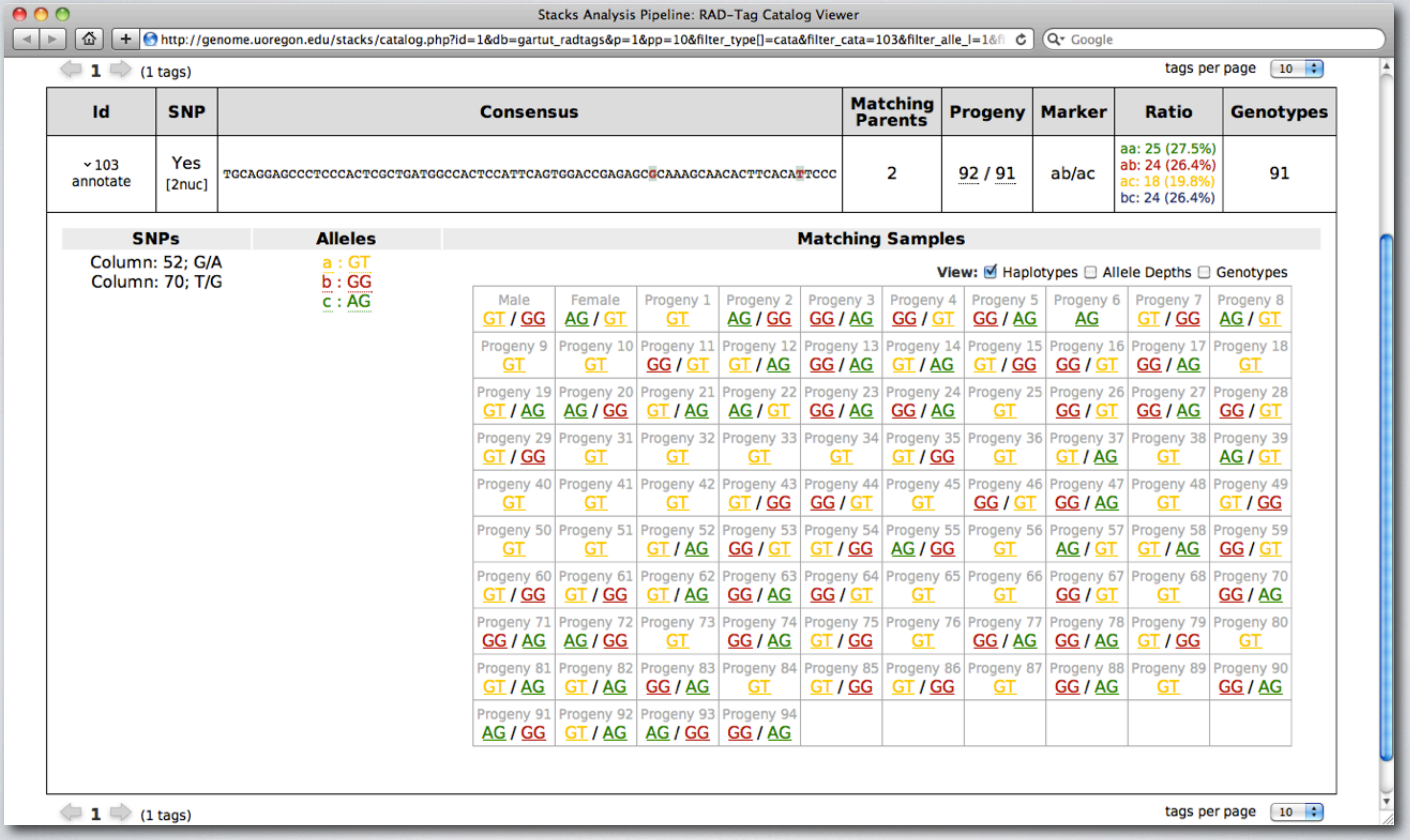

Stack 1
Stack 2
Stack 3
...
Stack X
Stack 1
Stack 2
Stack 3
⋮
Stack X
DICTIONARY
3
2
1
X
1
2
Locus 1
3
Locus 2
...
X
Locus X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCC C/A GACCAAACGTTTG

C

CT


Stack 1
Stack 2
Stack 3
...
Stack X
Stack 1
Stack 2
Stack 3
⋮
Stack X
DICTIONARY
3
2
1
X
1
2
Locus 1
3
Locus 2
...
X
Locus X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
Σ
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCC C/A GACCAAACGTTTG
CATALOG
Locus 1
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
Haplotypes: C A
Locus 2
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
Haplotypes: AC CT
Locus 3
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
Haplotypes: Consensus
⋮
Locus X
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
Haplotypes: AA GG
PARENT1
PARENT2
PROGENY1
PROGENY2
PROGENY3
PROGENY4
PROGENY5
PROGENYX

Stacks



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&f

1 (1 tags) tags per page 10

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

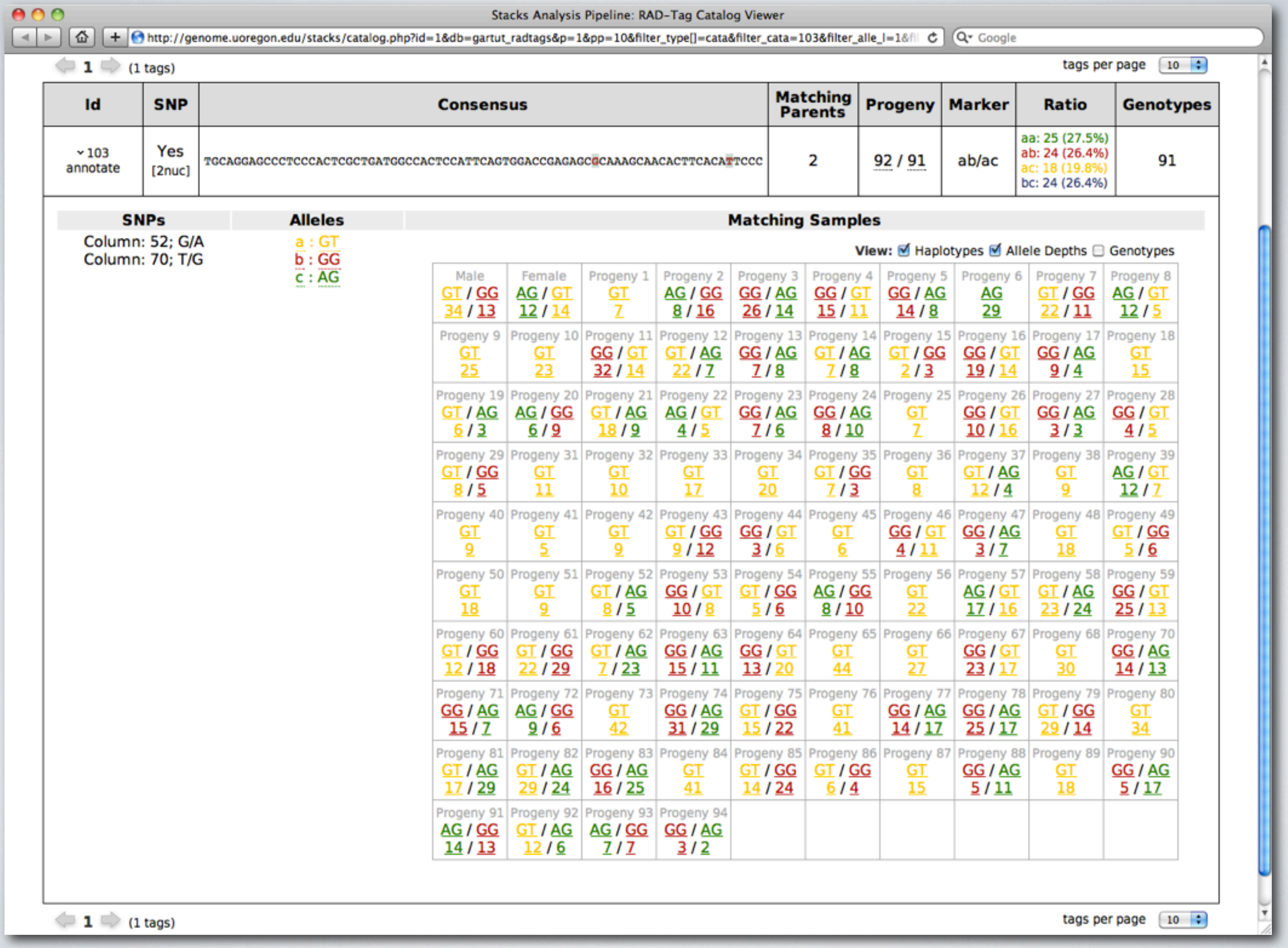
SNPs	Alleles
Column: 52; G/A Column: 70; T/G	a : GT b : GG c : AG

Matching Samples

View: ☑ Haplotypes ☐ Allele Depths ☐ Genotypes

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

1 (1 tags) tags per page 10



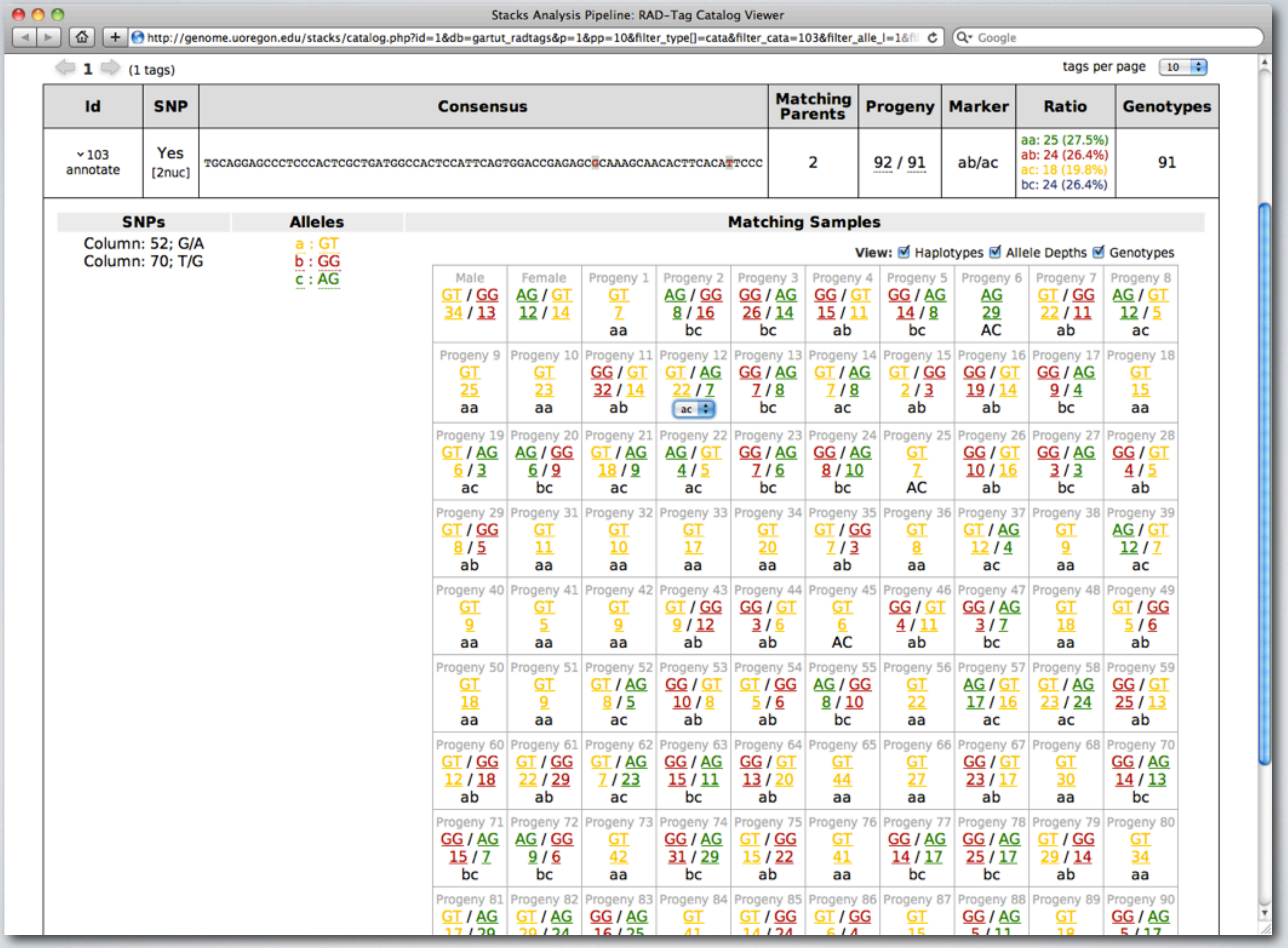
Id	SNP	Consensus	Matching Parents	Progeny	Marker	Ratio	Genotypes
103 annotate	Yes [2nuc]	TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC	2	92 / 91	ab/ac	aa: 25 (27.5%) ab: 24 (26.4%) ac: 18 (19.8%) bc: 24 (26.4%)	91

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

Matching Samples

View: ☑ Haplotypes ☑ Allele Depths ☐ Genotypes

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



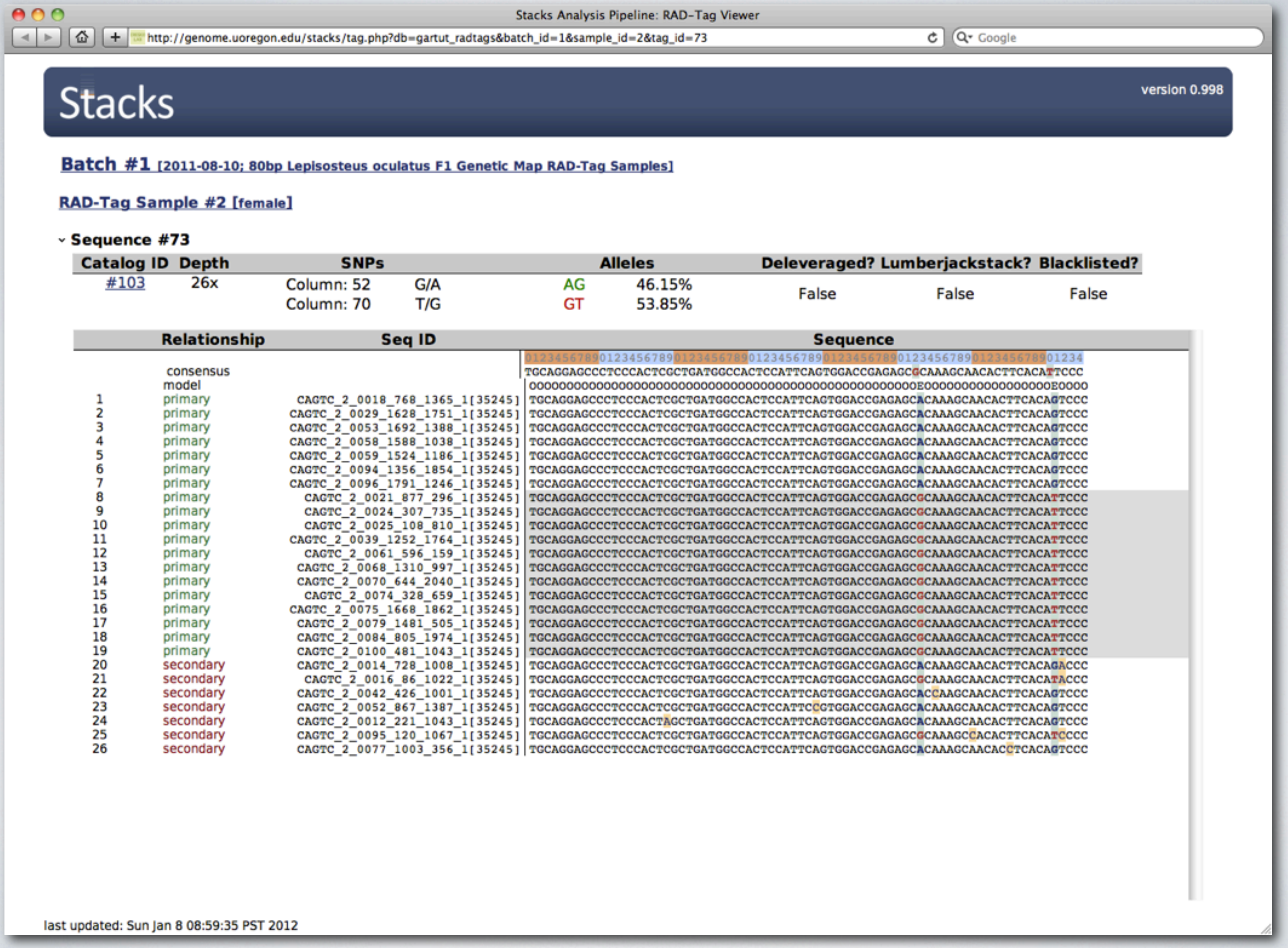
Id	SNP	Consensus	Matching Parents	Progeny	Marker	Ratio	Genotypes
103 annotate	Yes [2nuc]	TGCAGGAGCCCTCCCACTCGCTGATGGCCACTCCATTCAGTGGACCGAGAGCGCAAAGCAACACTTCACATTCCC	2	92 / 91	ab/ac	aa: 25 (27.5%) ab: 24 (26.4%) ac: 18 (19.8%) bc: 24 (26.4%)	91

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

Matching Samples

View: ☑ Haplotypes ☑ Allele Depths ☑ Genotypes

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



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

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 Column: 70	G/A T/G	AG GT	46.15% 53.85%	False	False	False

	Relationship	Seq ID	Sequence
			01234567890123456789012345678901234567890123456789012345678901234567890123456789
	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

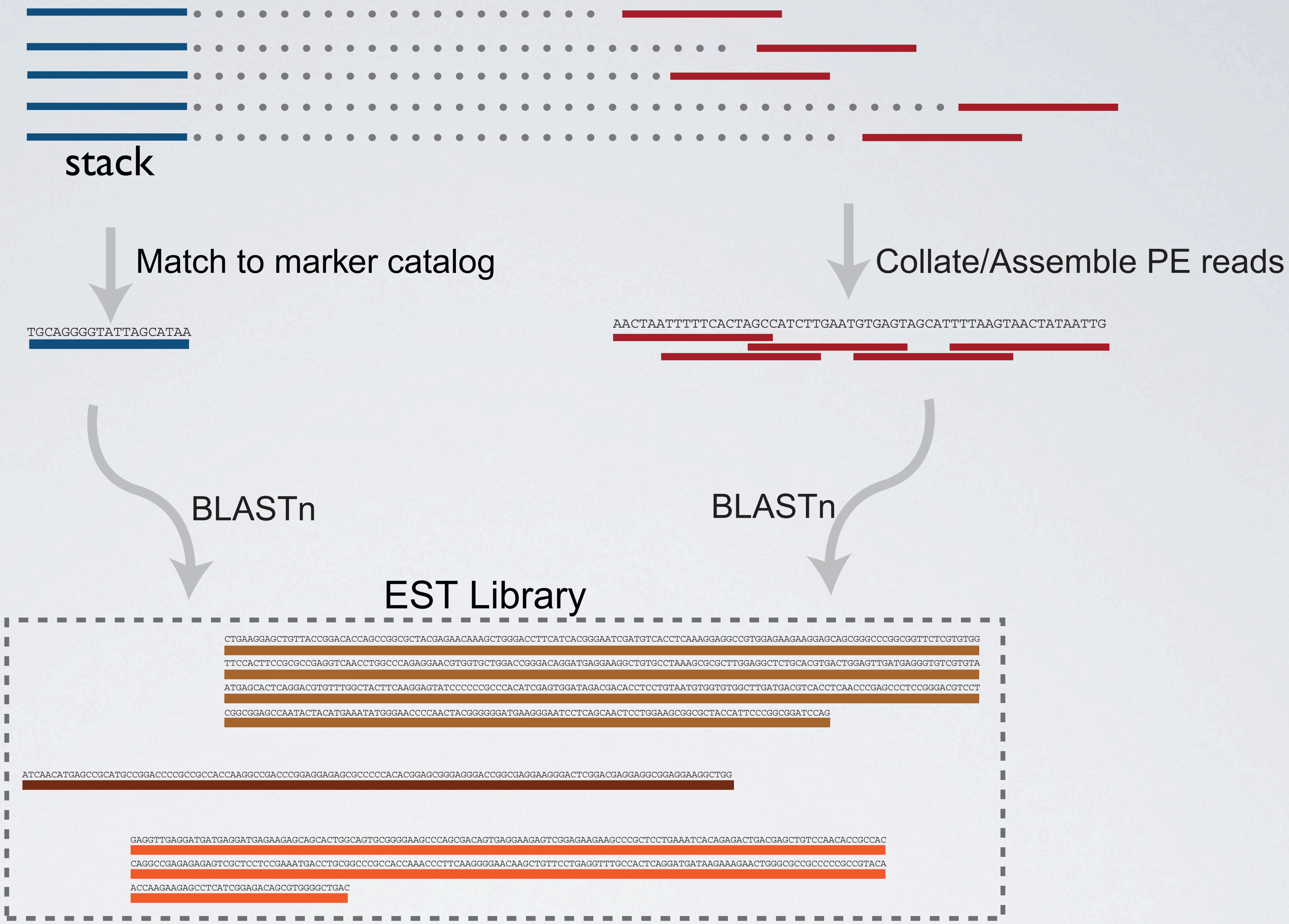
stack

Acquire paired-end sequence

Acquire paired-end sequence

Acquire paired-end sequence

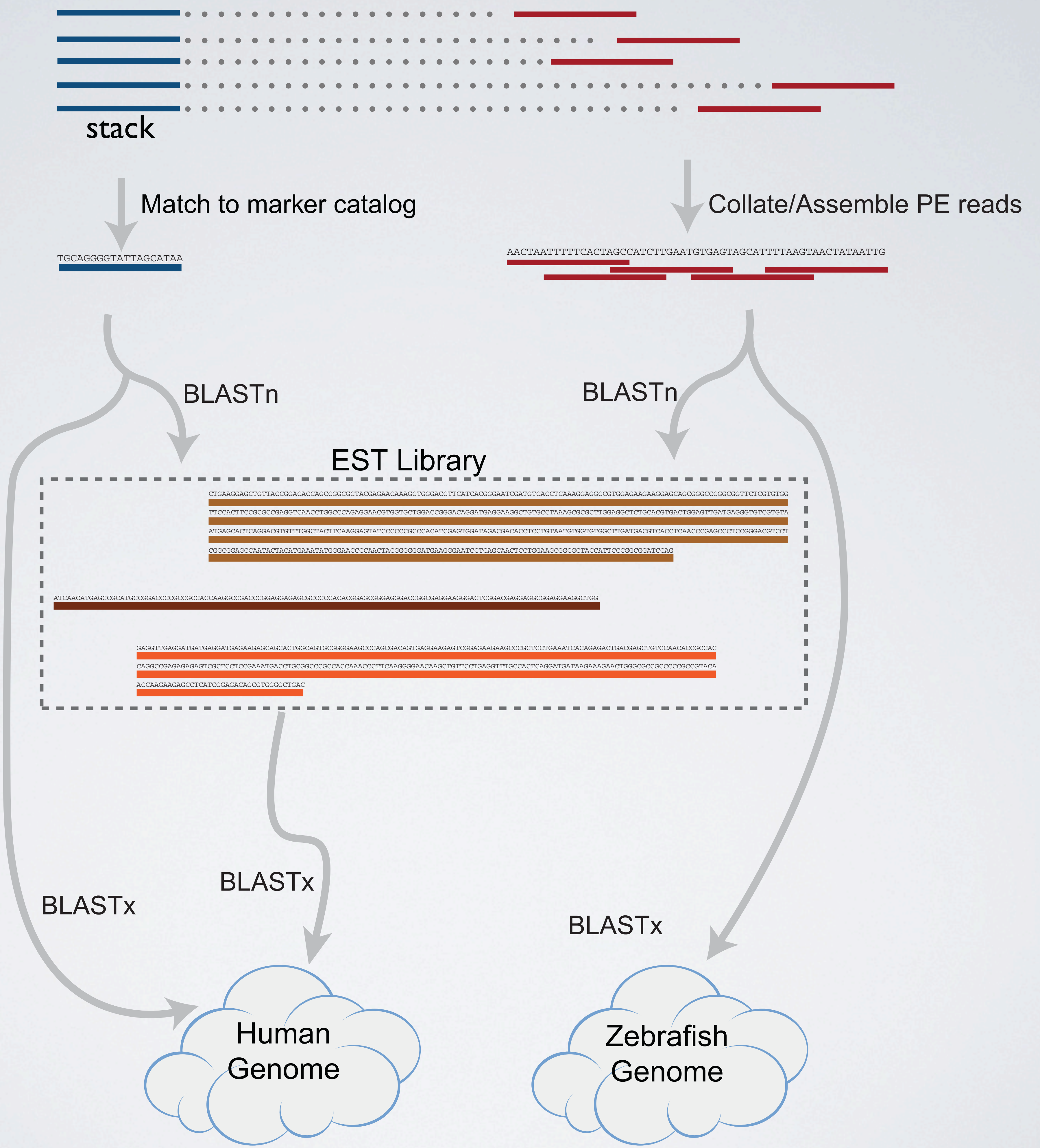
Associate markers / PE contigs with ESTs

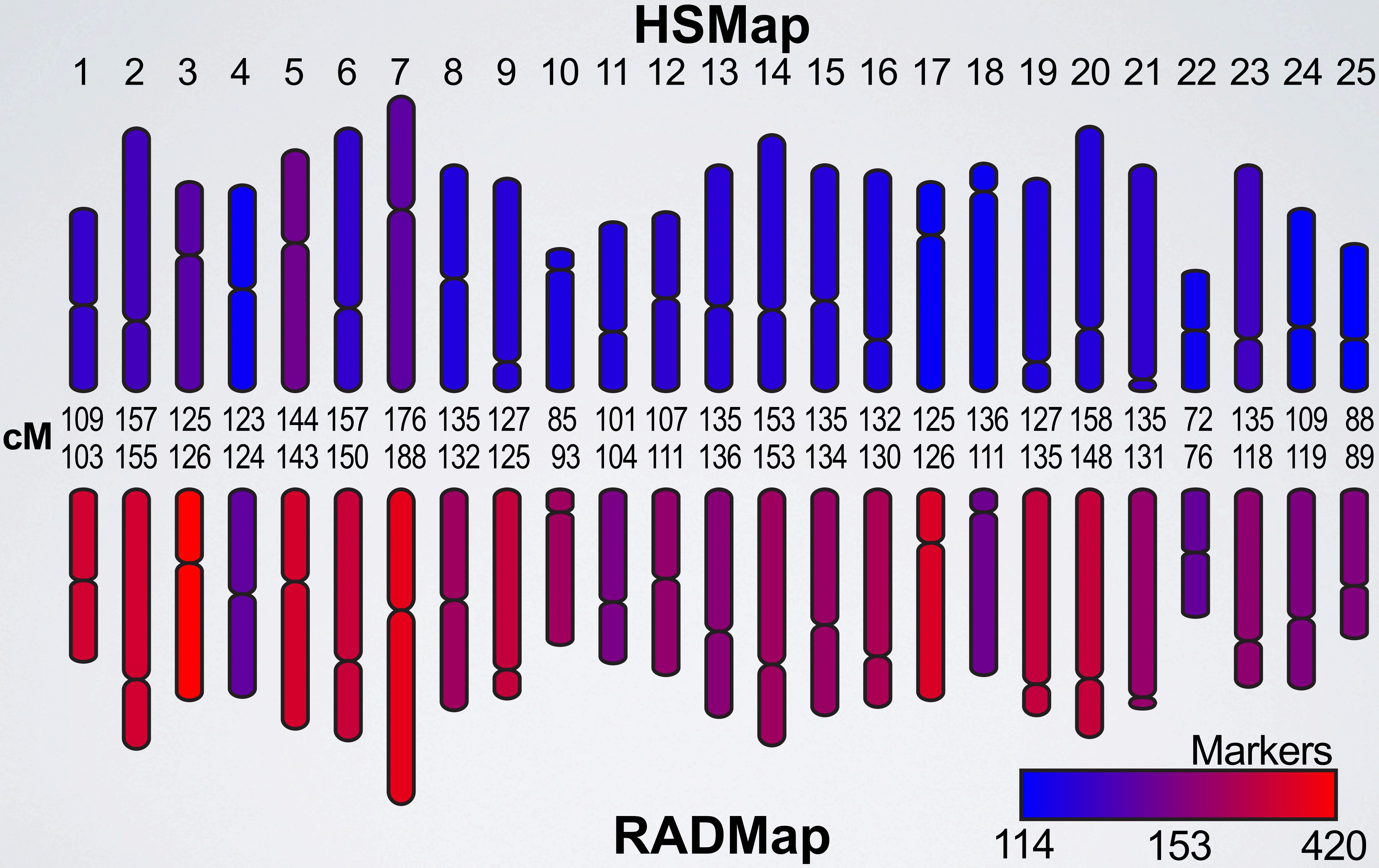


Acquire paired-end sequence

Associate markers / PE contigs with ESTs

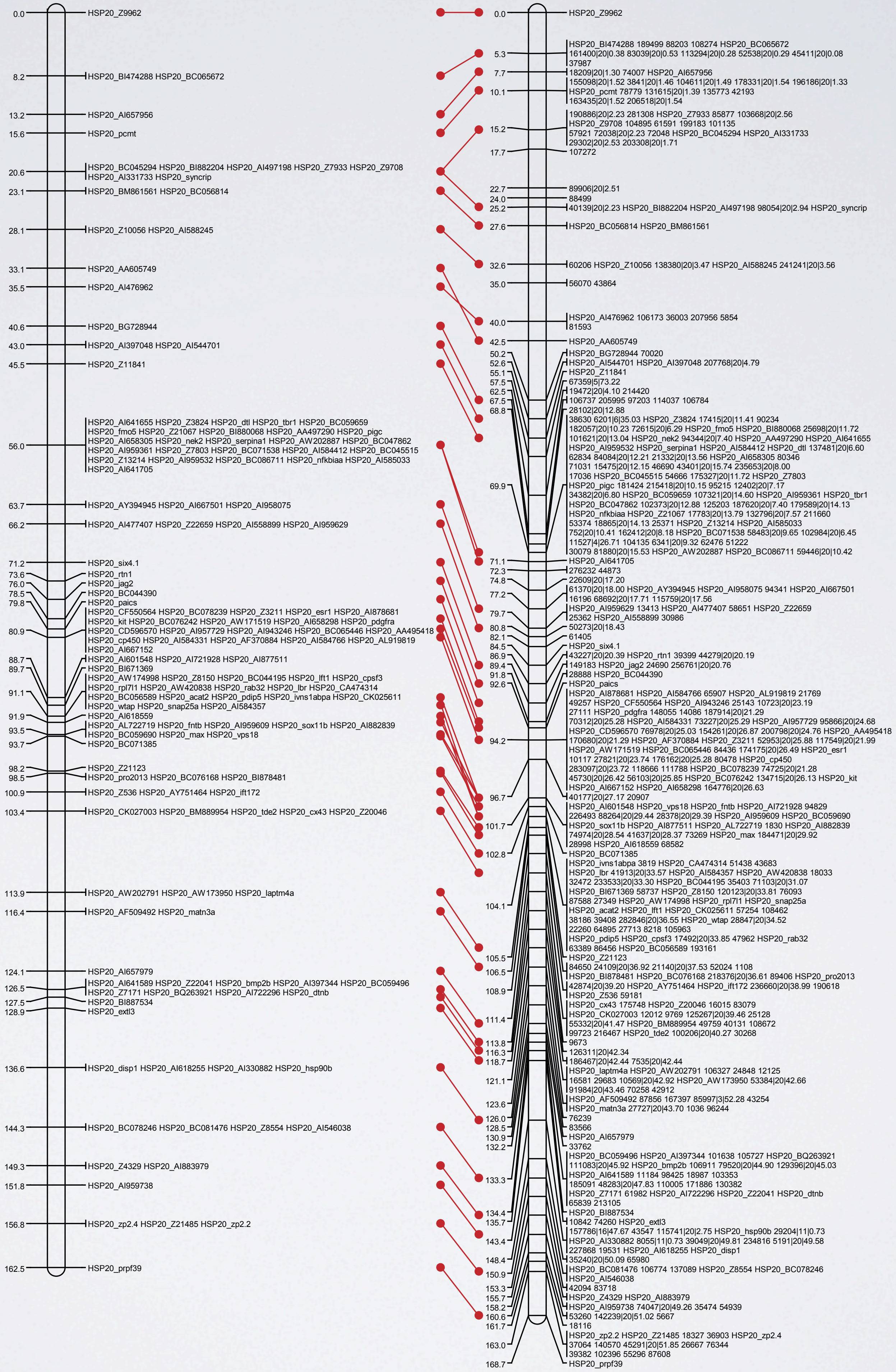
Assign orthology to:
  markers
  PE contigs
  ESTs

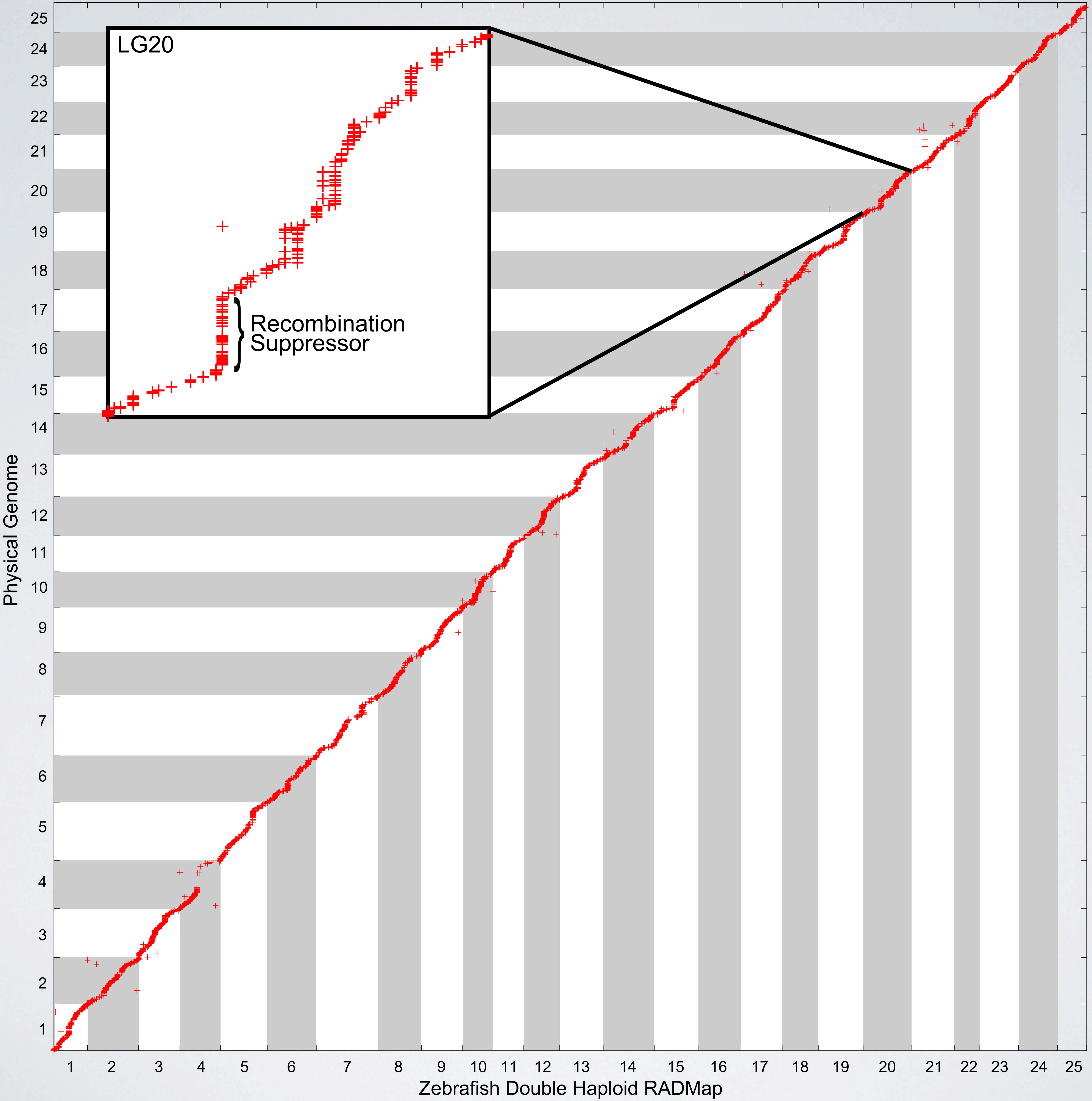



HSMap
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
cM
109 157 125 123 144 157 176 135 127 85 101 107 135 153 135 132 125 136 127 158 135 72 135 109 88
103 155 126 124 143 150 188 132 125 93 104 111 136 153 134 130 126 111 135 148 131 76 118 119 89
RADMap
Markers
114
153
420

DreLG20
157 EST markers

DreLG20
311 RAD markers

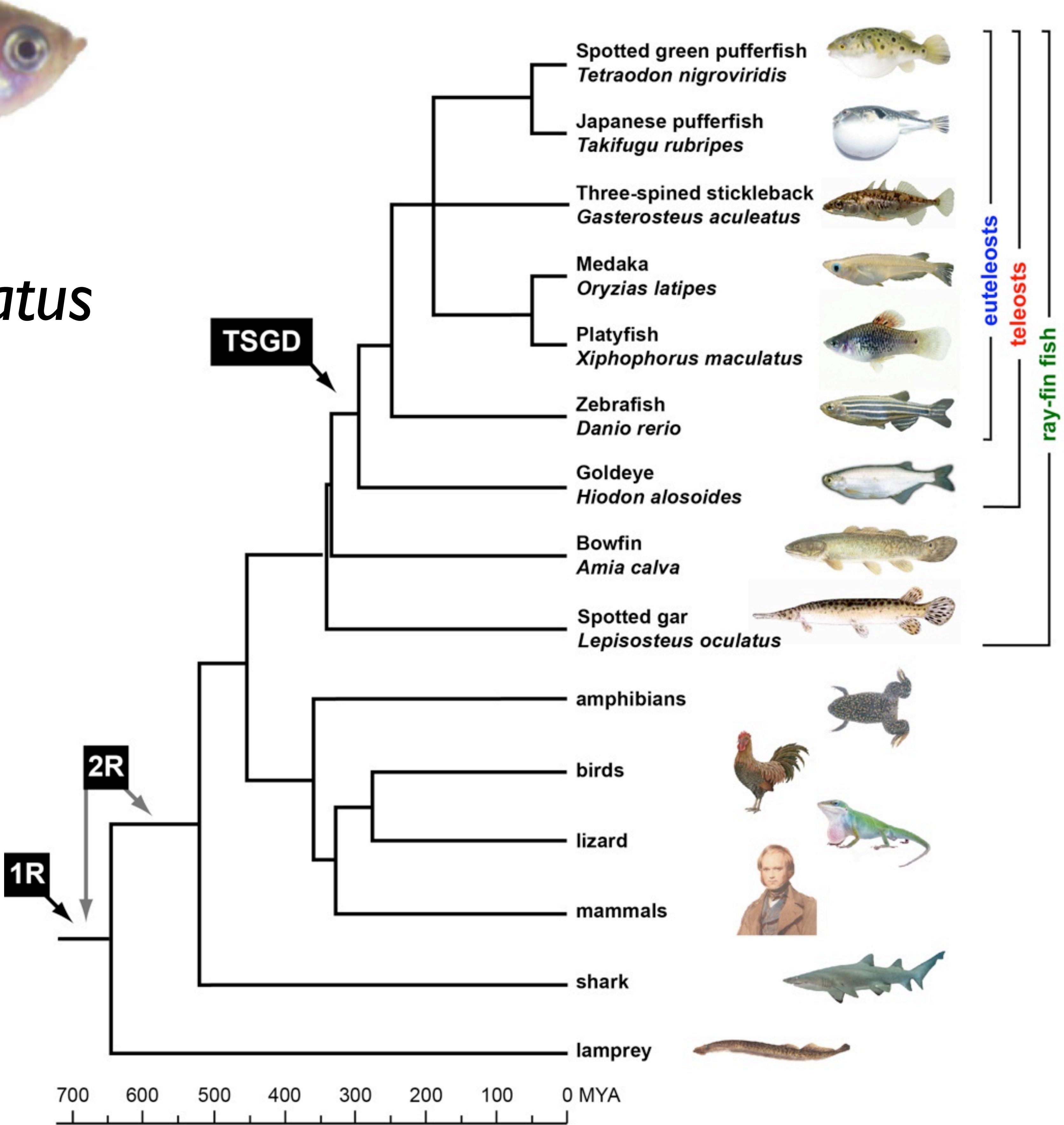



LG20
Recombination Suppressor
Physical Genome
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Zebrafish Double Haploid RADMap

Xiphophorous maculatus

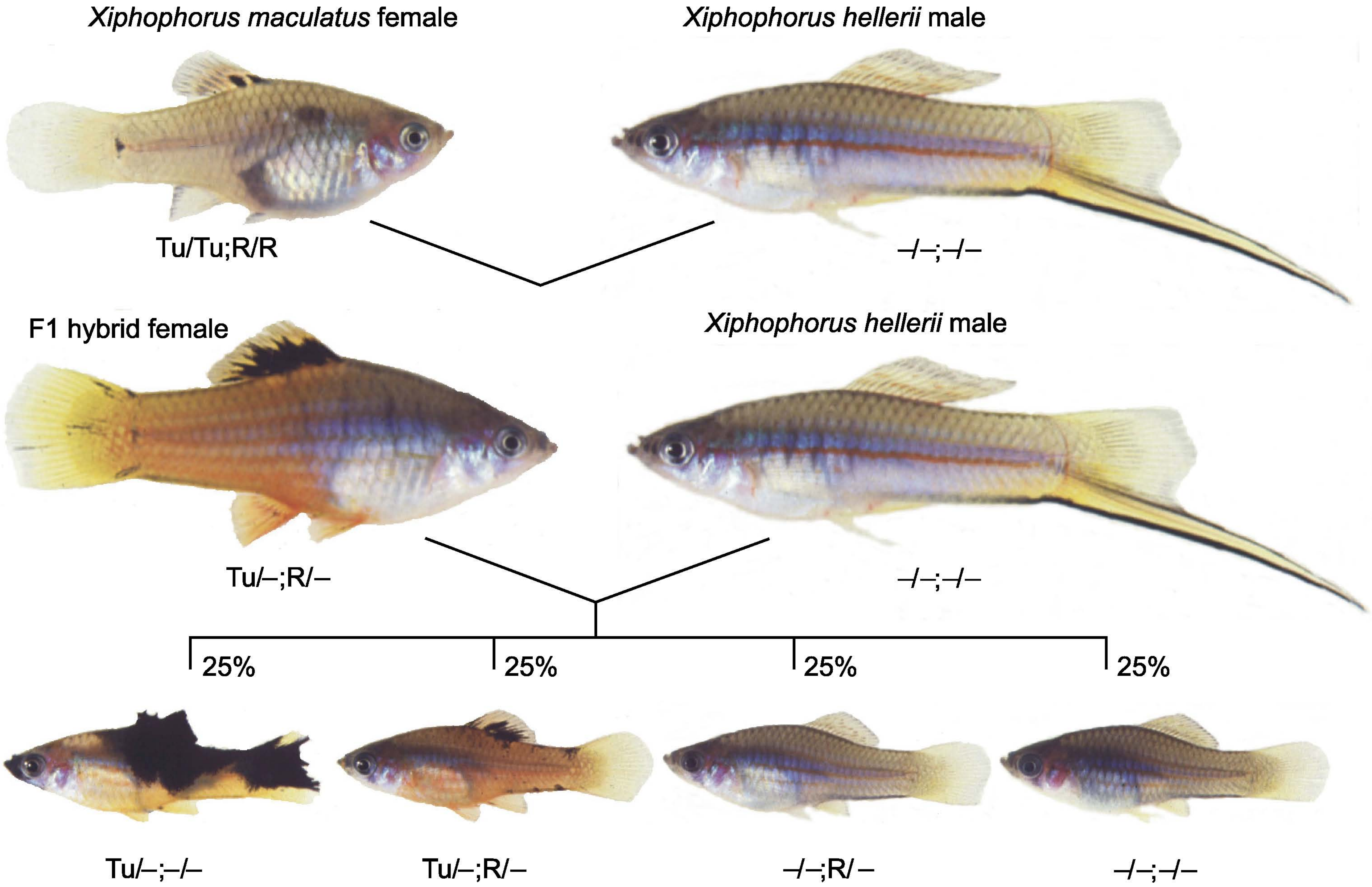
Angel Amores

John Postlethwait



Ingo Braasch

Genetic Mapping for Genome Assembly



Xiphophorus maculatus female

Tu/Tu;R/R

Xiphophorus hellerii male

–/–;–/–

F1 hybrid female

Tu/–;R/–

Xiphophorus hellerii male

–/–;–/–

25%

Tu/–;–/–

25%

Tu/–;R/–

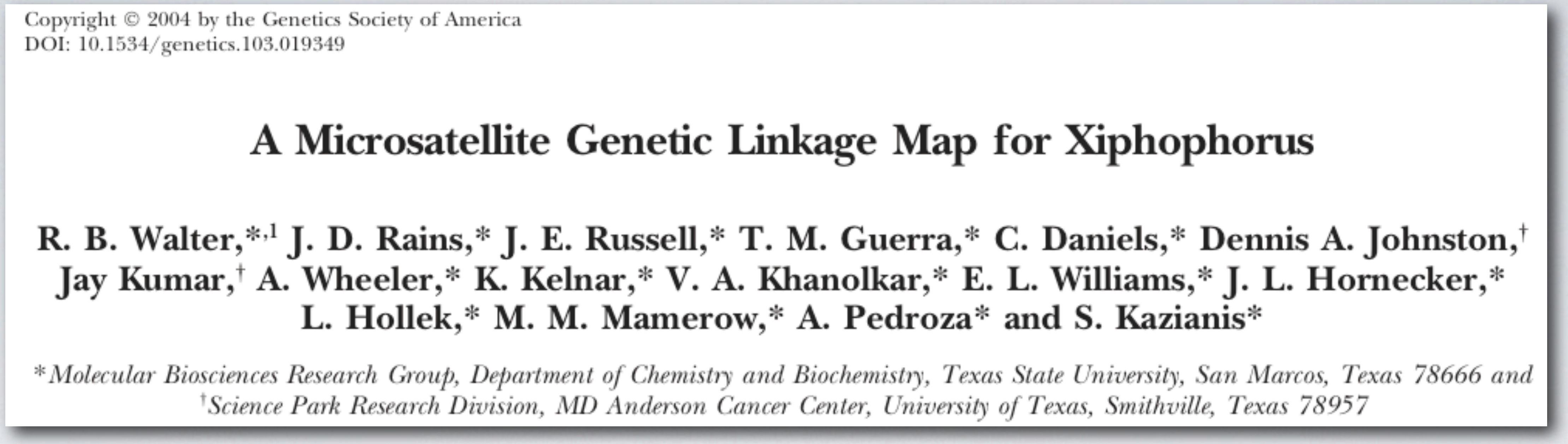
25%

–/–;R/ –

25%

–/–;–/–

Version 1.0



DOI: 10.1534/genetics.103.019349

A Microsatellite Genetic Linkage Map for Xiphophorus

R. B. Walter,*[,1] J. D. Rains,* J. E. Russell,* T. M. Guerra,* C. Daniels,* Dennis A. Johnston,† Jay Kumar,† A. Wheeler,* K. Kelnar,* V. A. Khanolkar,* E. L. Williams,* J. L. Hornecker,* L. Hollek,* M. M. Mamerow,* A. Pedroza* and S. Kazianis*

**Molecular Biosciences Research Group, Department of Chemistry and Biochemistry, Texas State University, San Marcos, Texas 78666 and* †*Science Park Research Division, MD Anderson Cancer Center, University of Texas, Smithville, Texas 78957*

290 markers

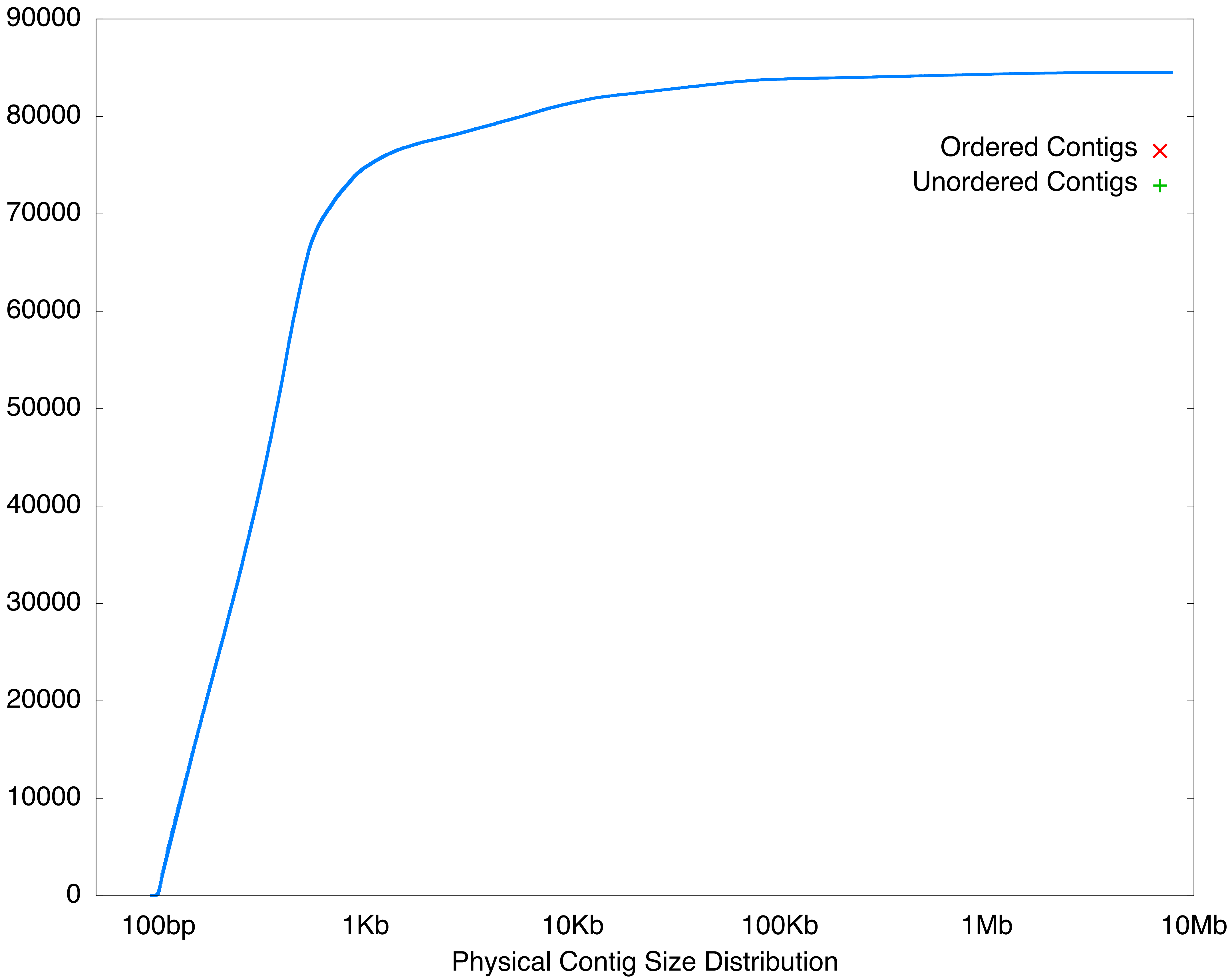
Version 2.0

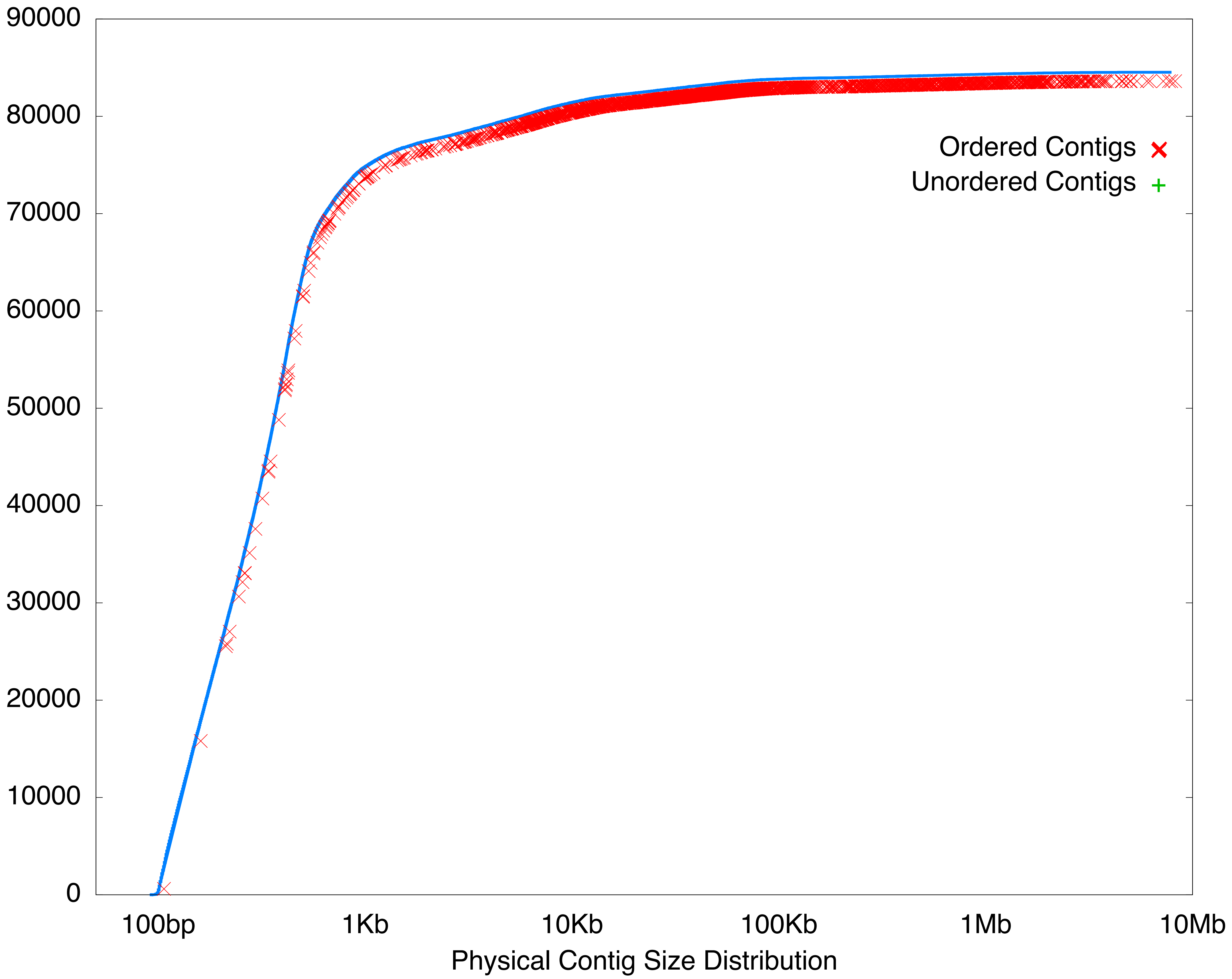
269 total fish

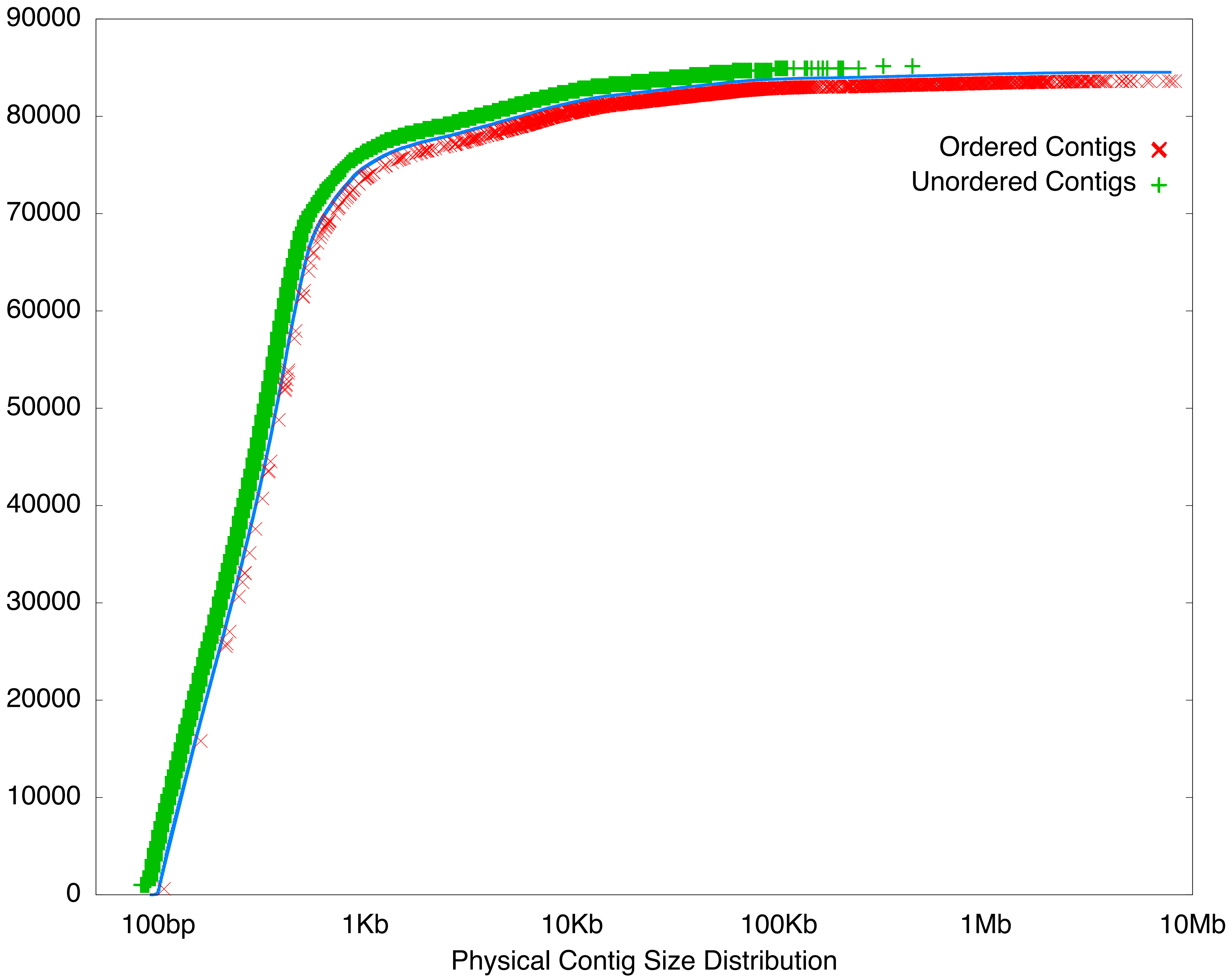
828,373,617 raw reads

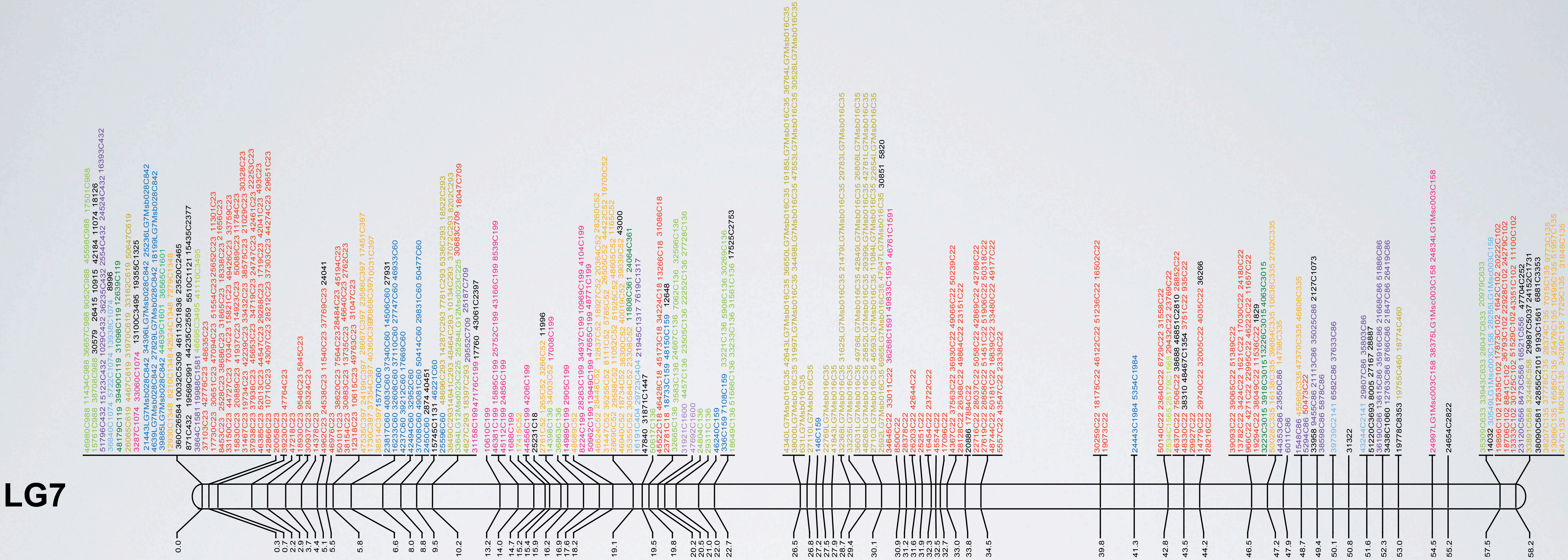
613,162,521 incorporated reads

22,144 markers in at least 90 progeny


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

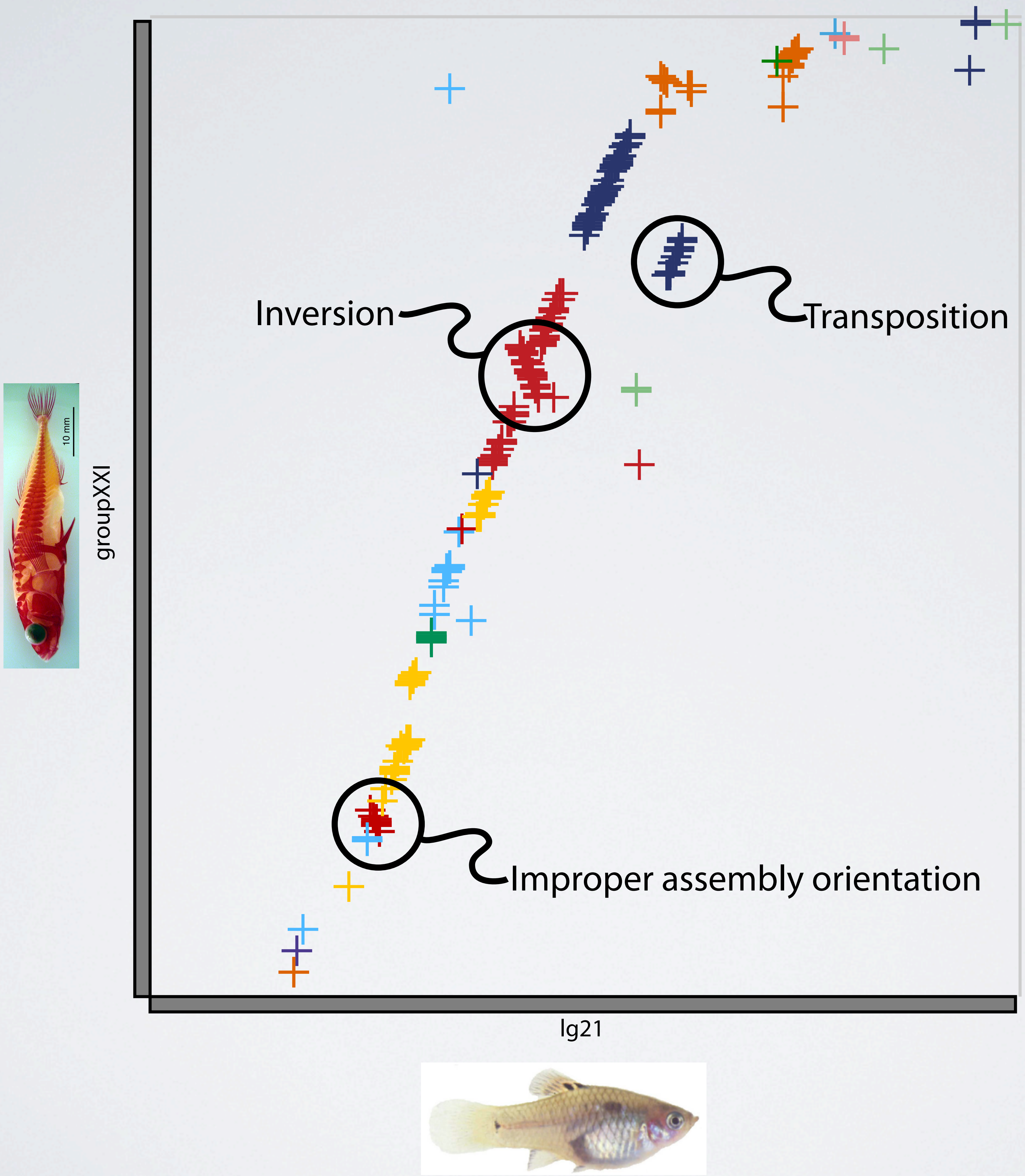

90000
80000
70000
60000
50000
40000
30000
20000
10000
0
100bp
1Kb
10Kb
100Kb
1Mb
10Mb
Physical Contig Size Distribution
Ordered Contigs ×
Unordered Contigs +


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



Given 80,000 contigs: 86% of genome ordered

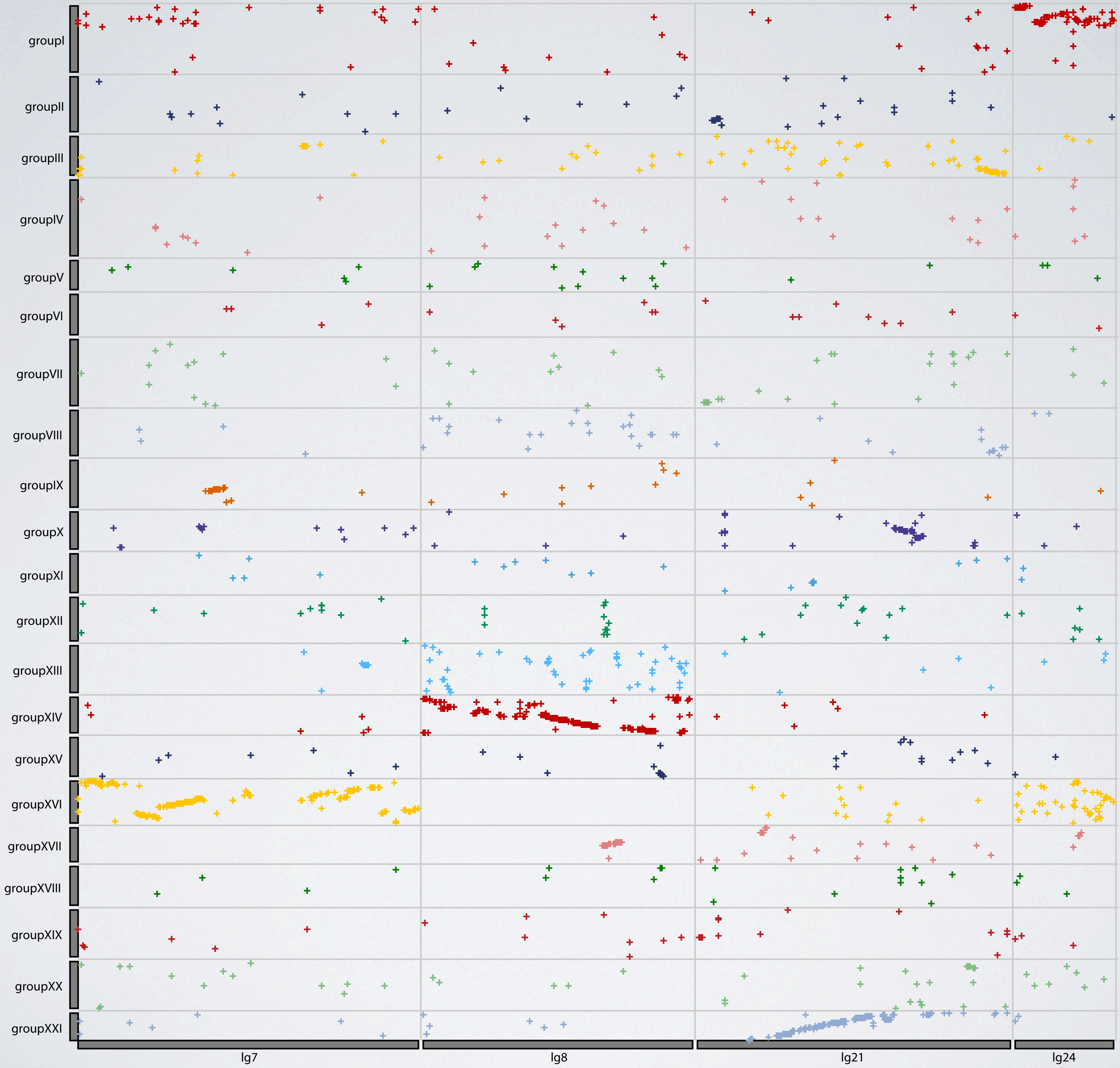

10 mm
groupXXI
lg21

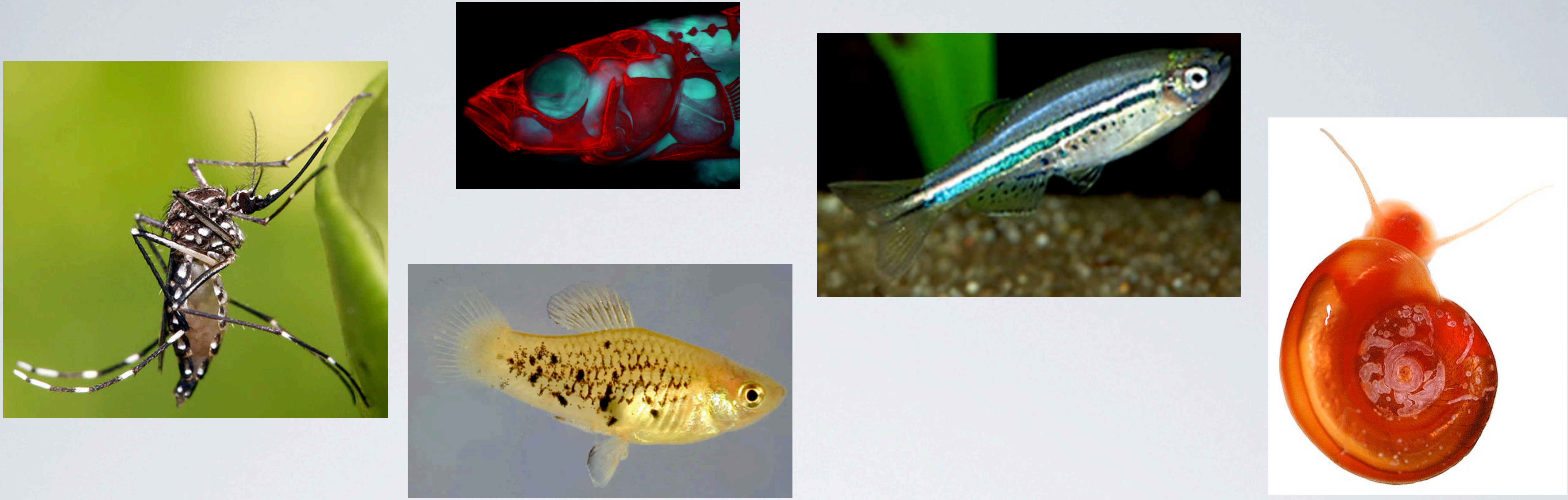


Inversion

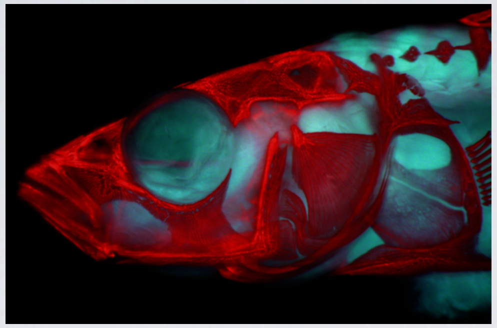
Transposition

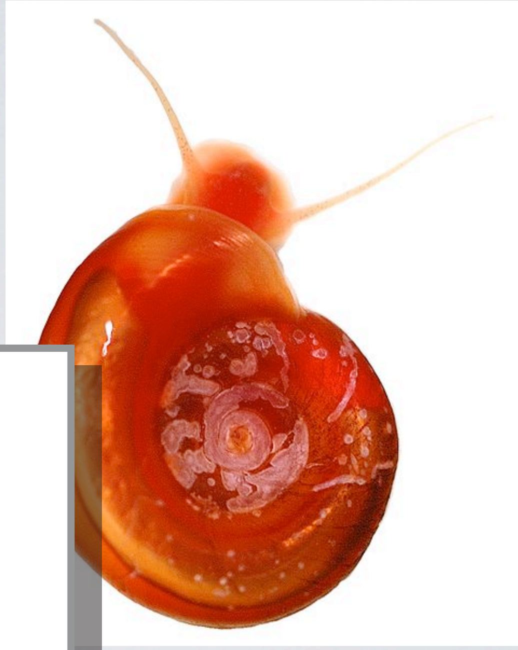
Improper assembly orientation

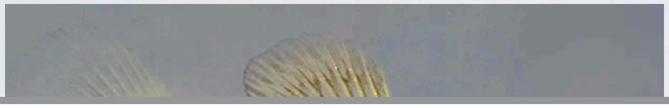

groupI
groupII
groupIII
groupIV
groupV
groupVI
groupVII
groupVIII
groupIX
groupX
groupXI
groupXII
groupXIII
groupXIV
groupXV
groupXVI
groupXVII
groupXVIII
groupXIX
groupXX
groupXXI
lg7
lg8
lg21
lg24



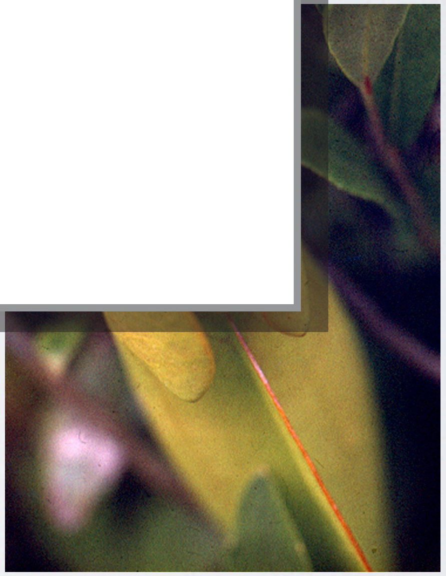
What if you **don't** have a reference genome?

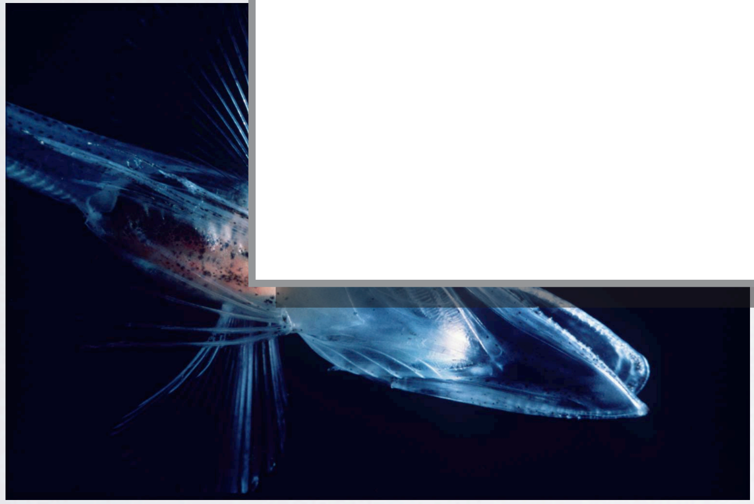




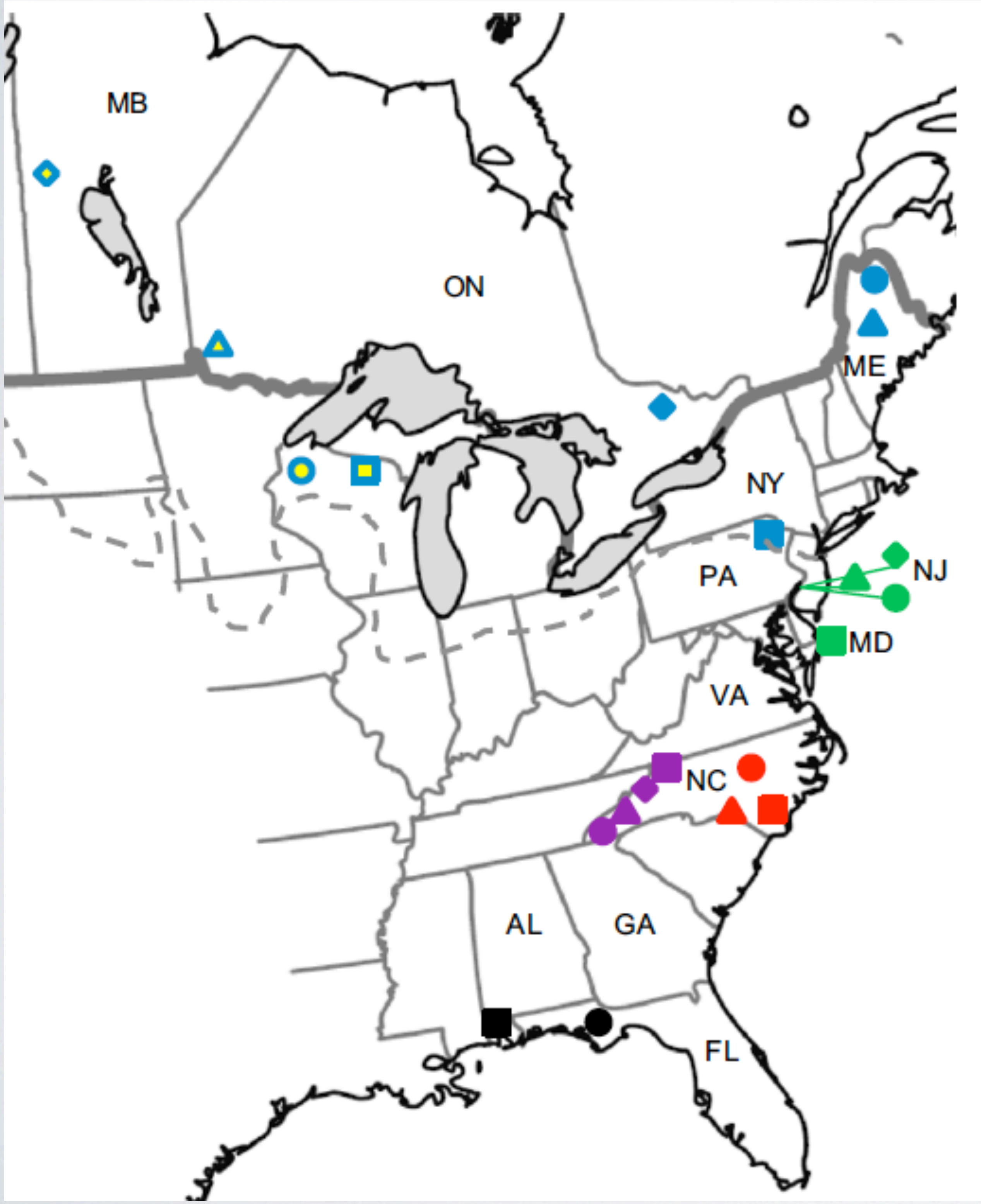


Phylogeography

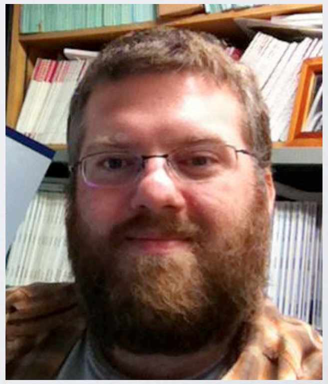




The pitcher plant mosquito, *Wyeomyia smithii* a model for evolution due to climate change



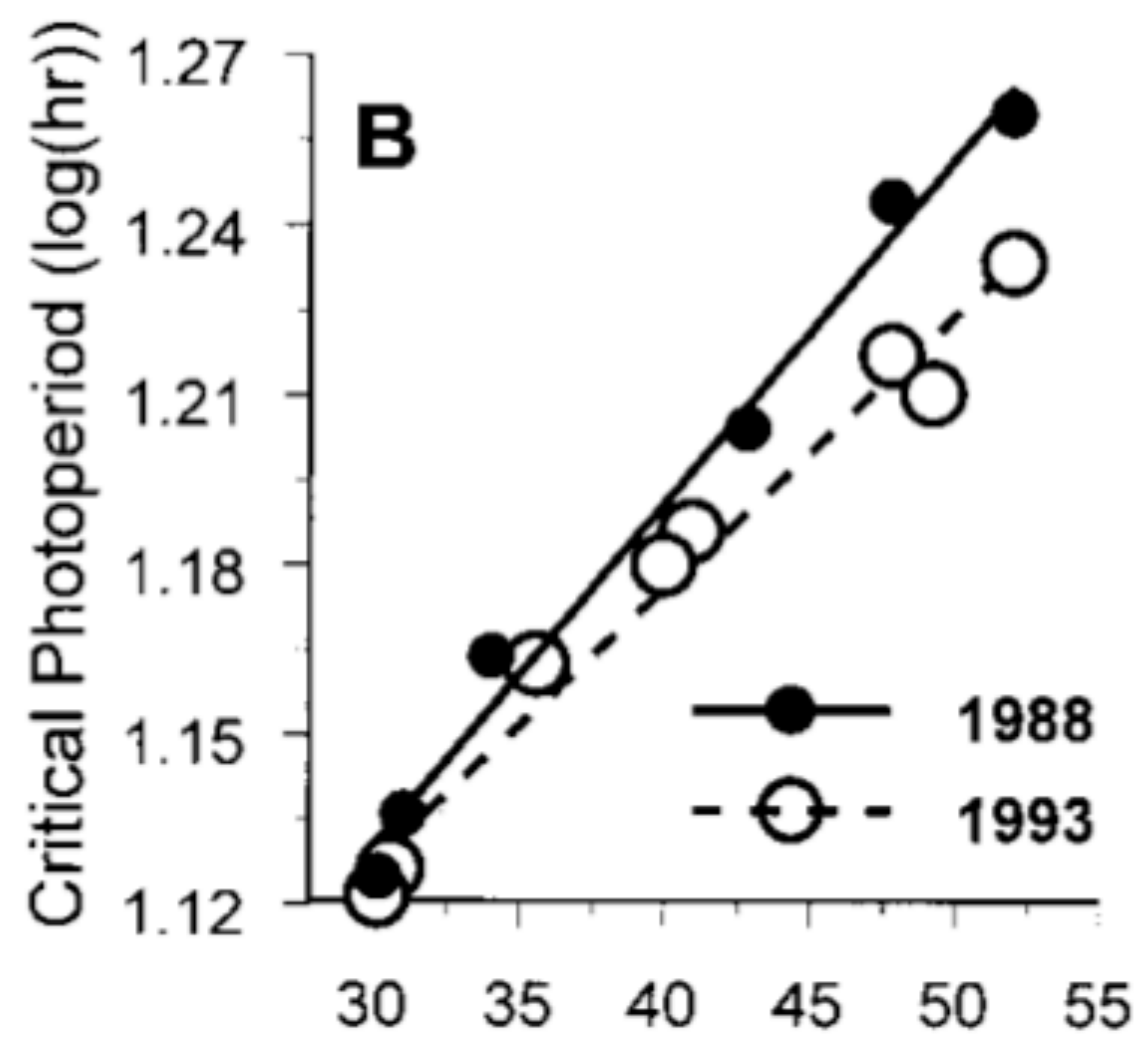
Paul Hohenlohe



Kevin Emerson

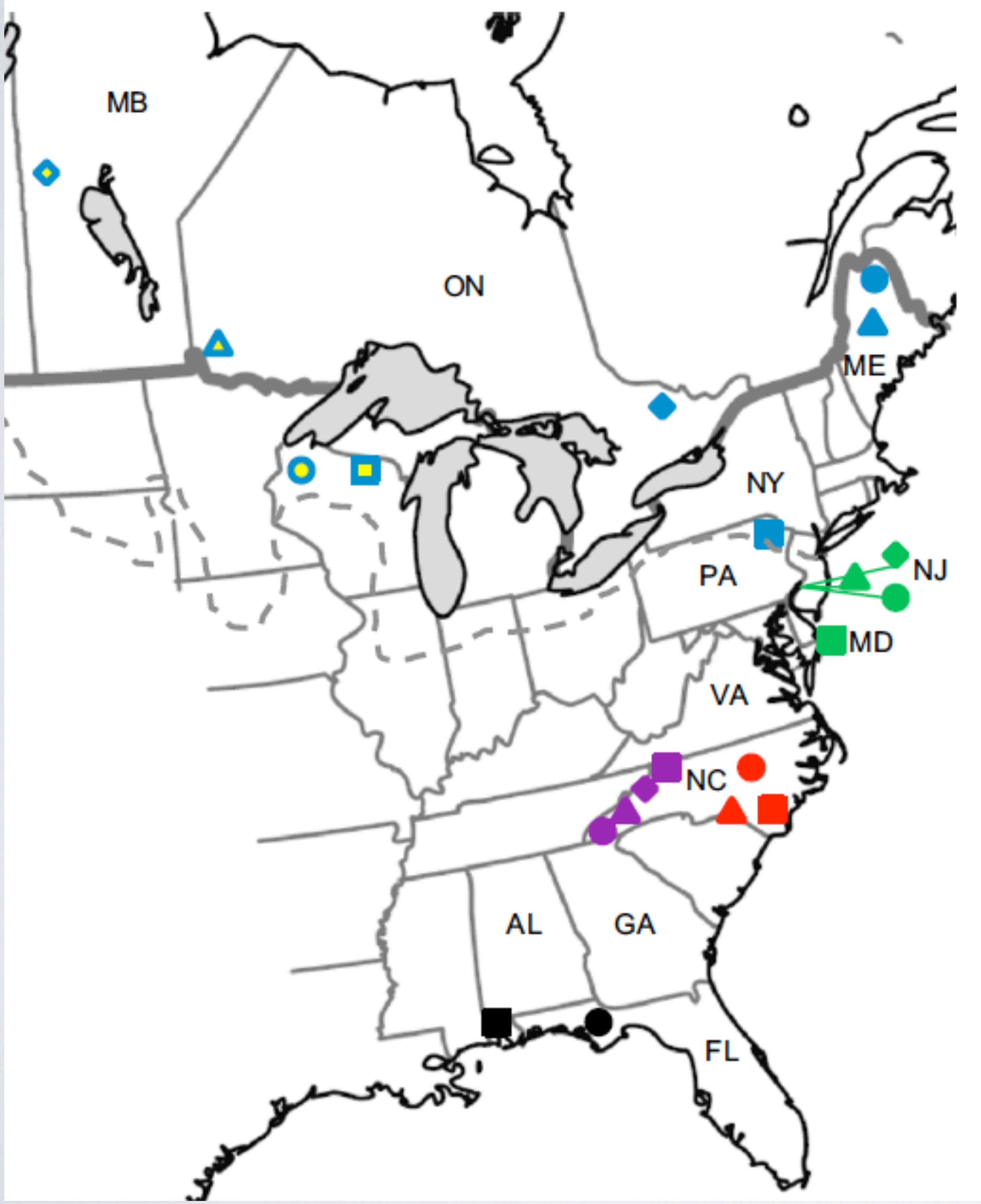
The pitcher plant mosquito, *Wyeomyia smithii* a model for evolution due to climate change

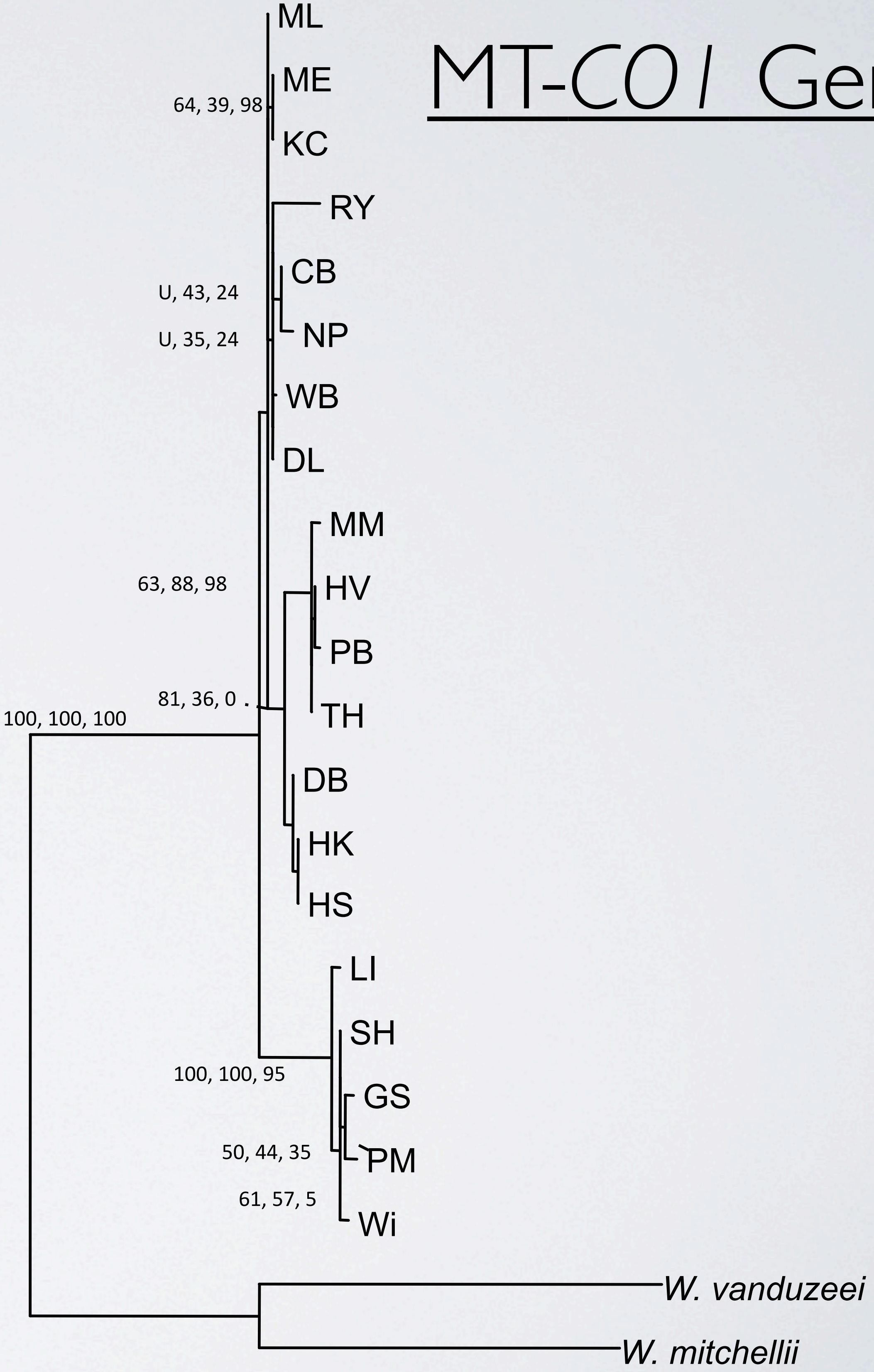
Bradshaw et al (2001) PNAS



Latitude

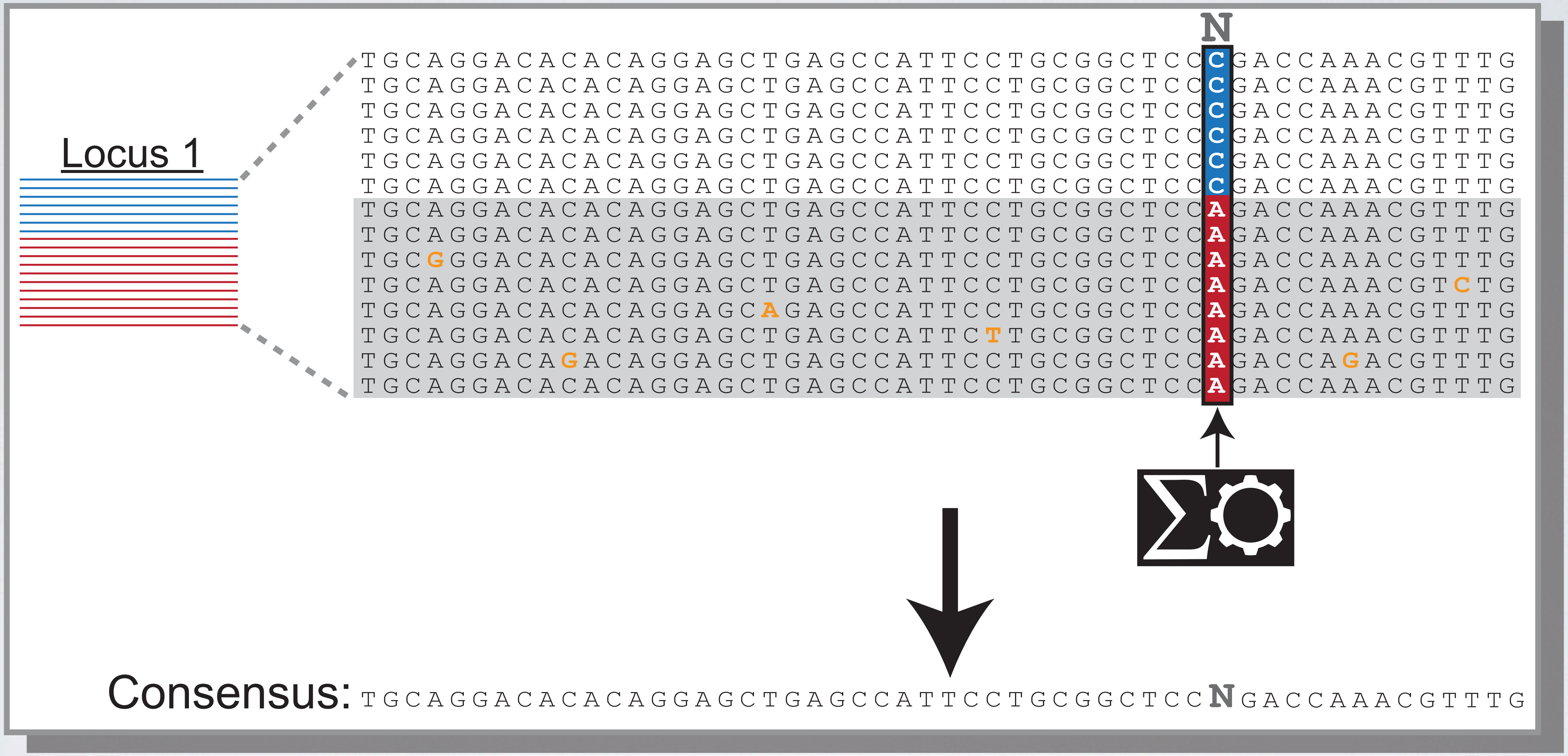
Resolving post-glacial phylogeography in *Wyeomyia smithii*





MT-*CO1* Gene

Sampling RAD-Tag markers for population-level phylogenetic analysis



```
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAGACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
```

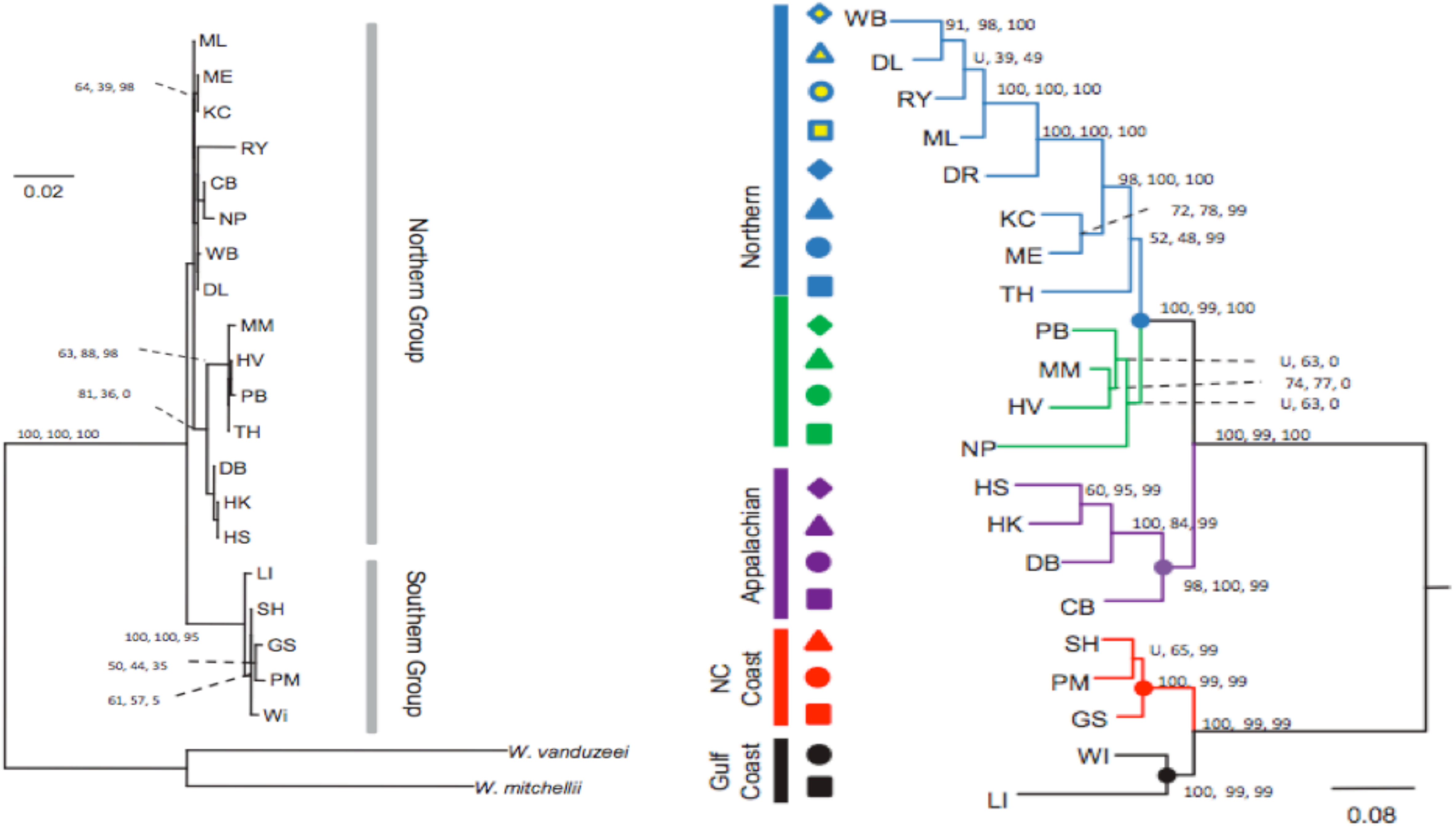
Consensus: TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCNGACCAAACGTTTG

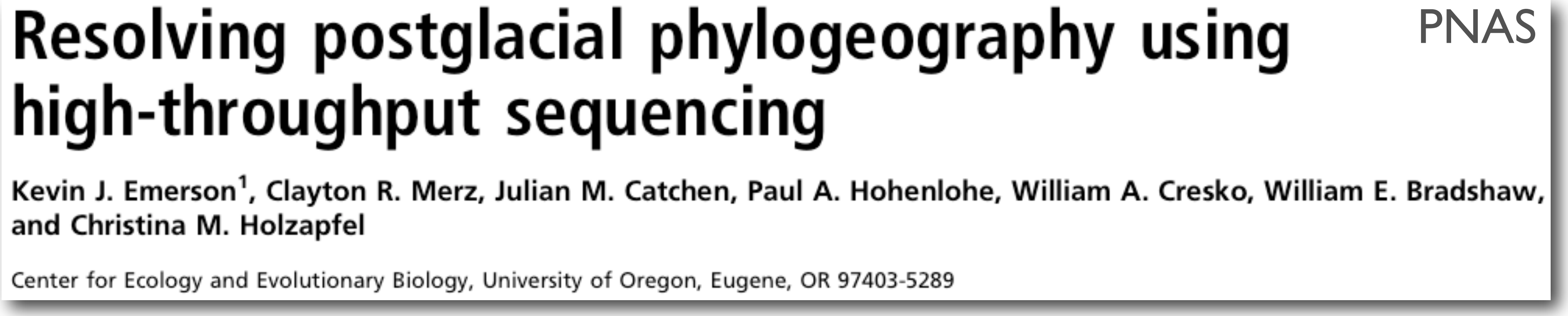
Sampling RAD-Tag markers for population-level phylogenetic analysis

TGCAGGA**N**ACACAGGAGCTGAGCCATTCCTGCGGCTCC**N**GACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGT**N**TG
TGCAGGACACACAGGAGCTG**N**GCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACAC**N**CAGTAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG

TGCAGGACACACAGTAGCTGAGCCATTCCTGCGGCTCCCGACCAA**N**CGTTTG

Genome-wide view improves resolution



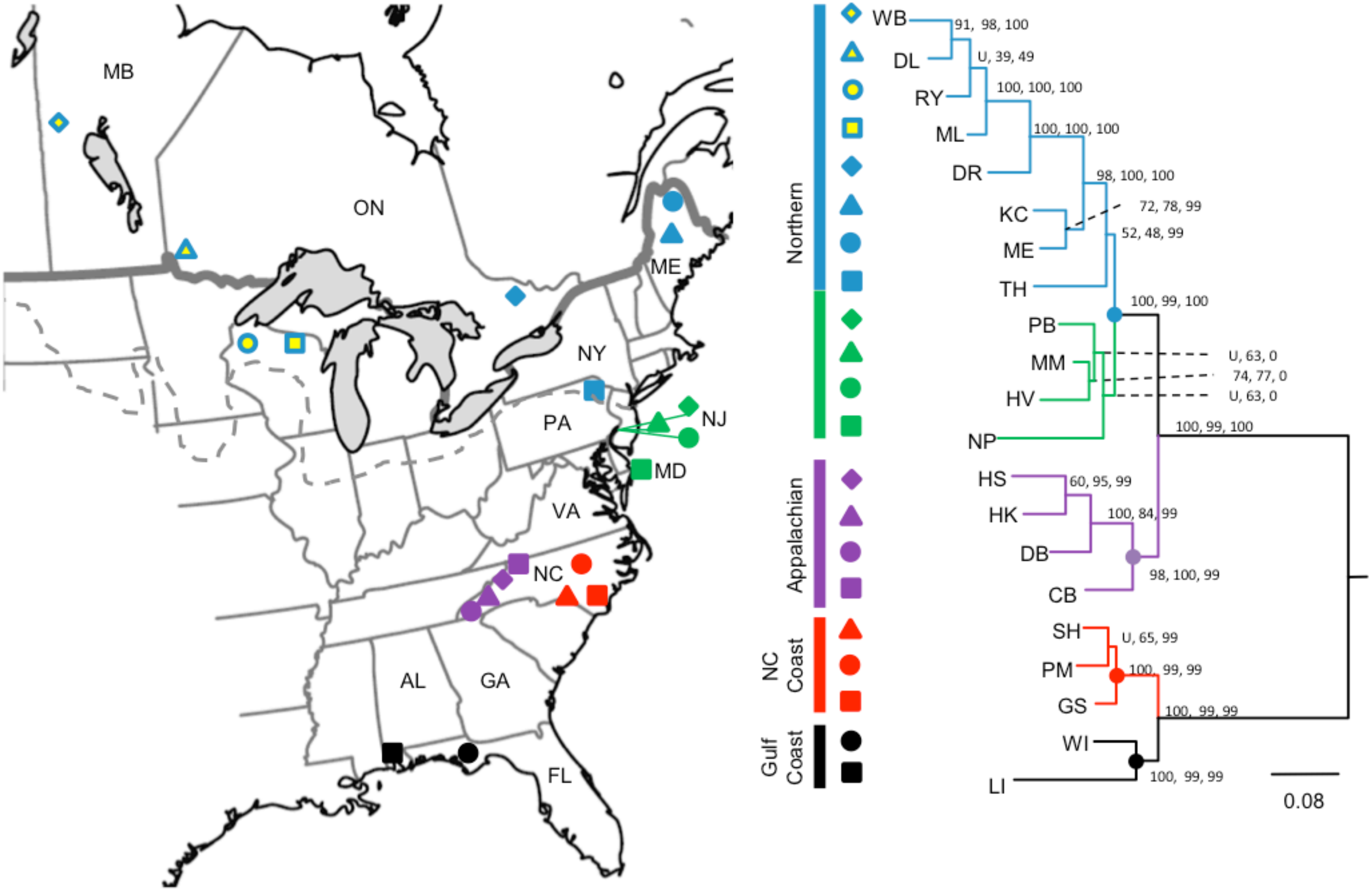


PNAS

Resolving postglacial phylogeography using high-throughput sequencing

Kevin J. Emerson[1], Clayton R. Merz, Julian M. Catchen, Paul A. Hohenlohe, William A. Cresko, William E. Bradshaw, and Christina M. Holzapfel

Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403-5289



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