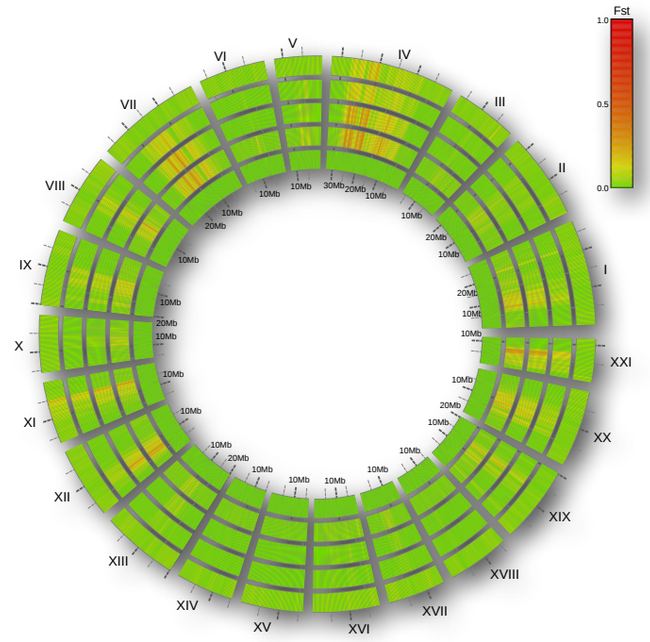


Ecological & evolutionary genomic analyses using RAD-seq

2016 Workshop on Genomics
Český Krumlov

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

UNIVERSITY OF OREGON

Outline for today's lecture

RAD-seq for ecological and evolutionary genomics

Evolutionary genomics of stickleback fish

RAD-seq experimental and statistical considerations

Genomically enabling pipefish

Stacks *software pipeline (this afternoon & evening)*

January 2016 Issue

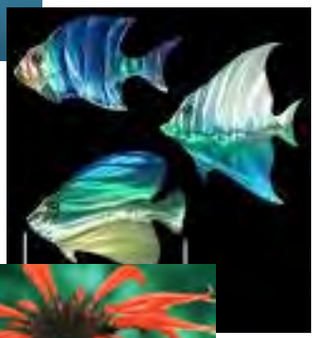
STUDY DESIGNS

Harnessing the power of RADseq for ecological and evolutionary genomics

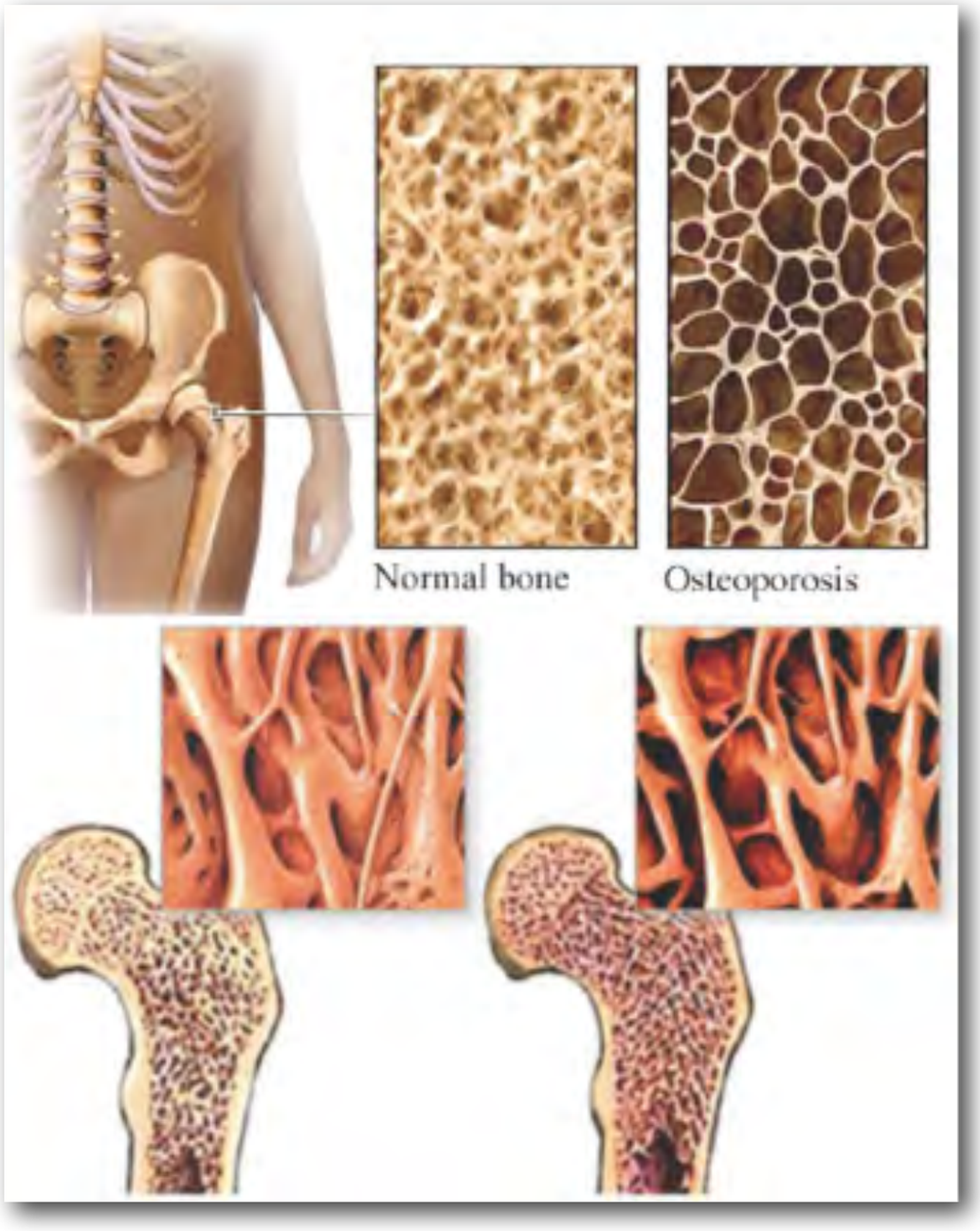
Kimberly R. Andrews[1], Jeffrey M. Good[2], Michael R. Miller[3], Gordon Luikart[4] and Paul A. Hohenlohe[5]

nature REVIEWS GENETICS

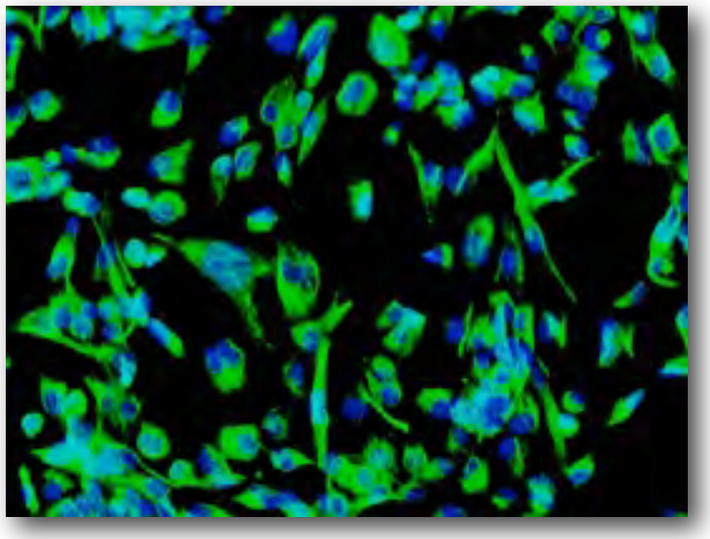
Why do species look the way that they do?

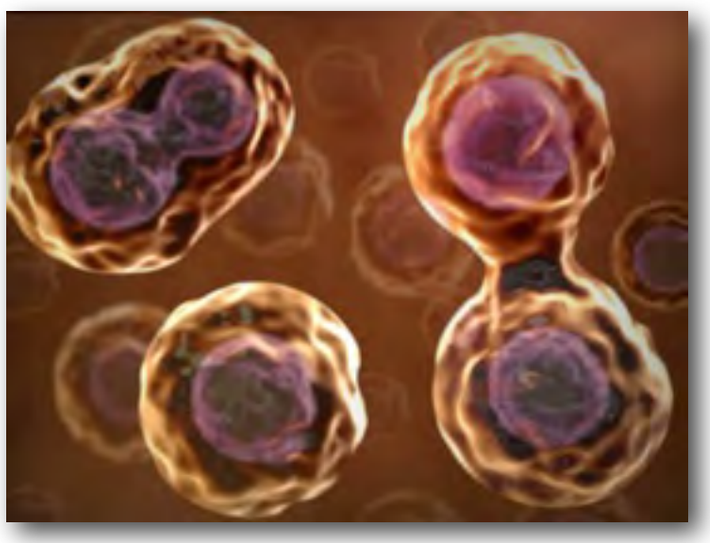


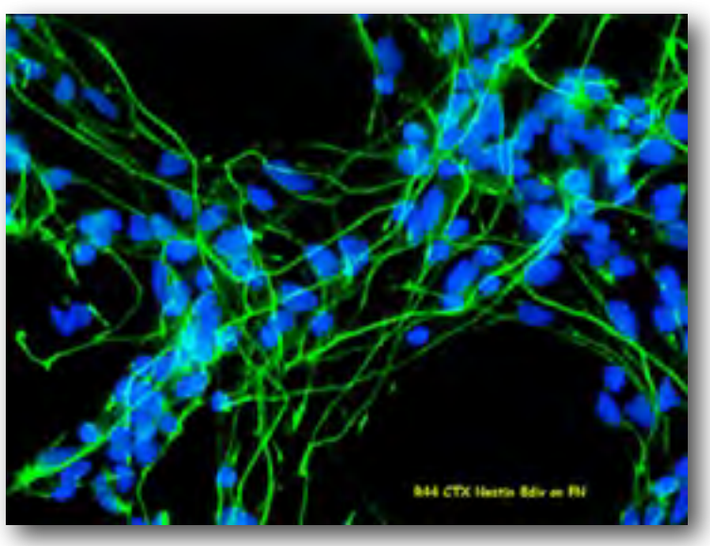
Why do organisms vary?

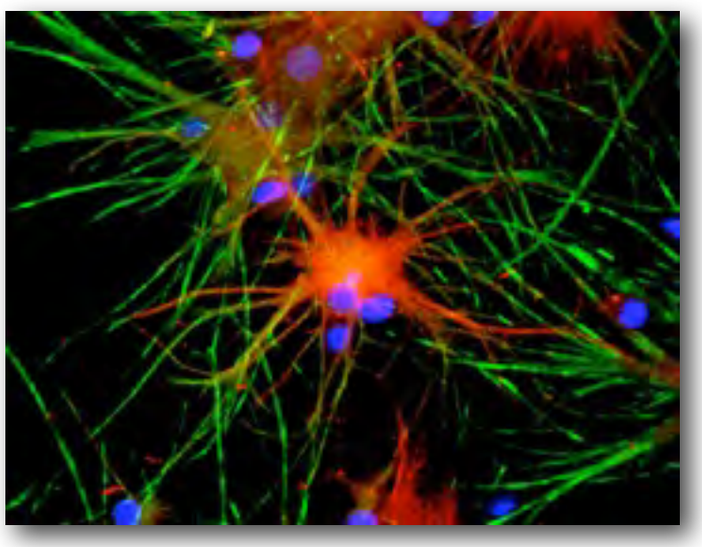


How is cellular functional diversity created?

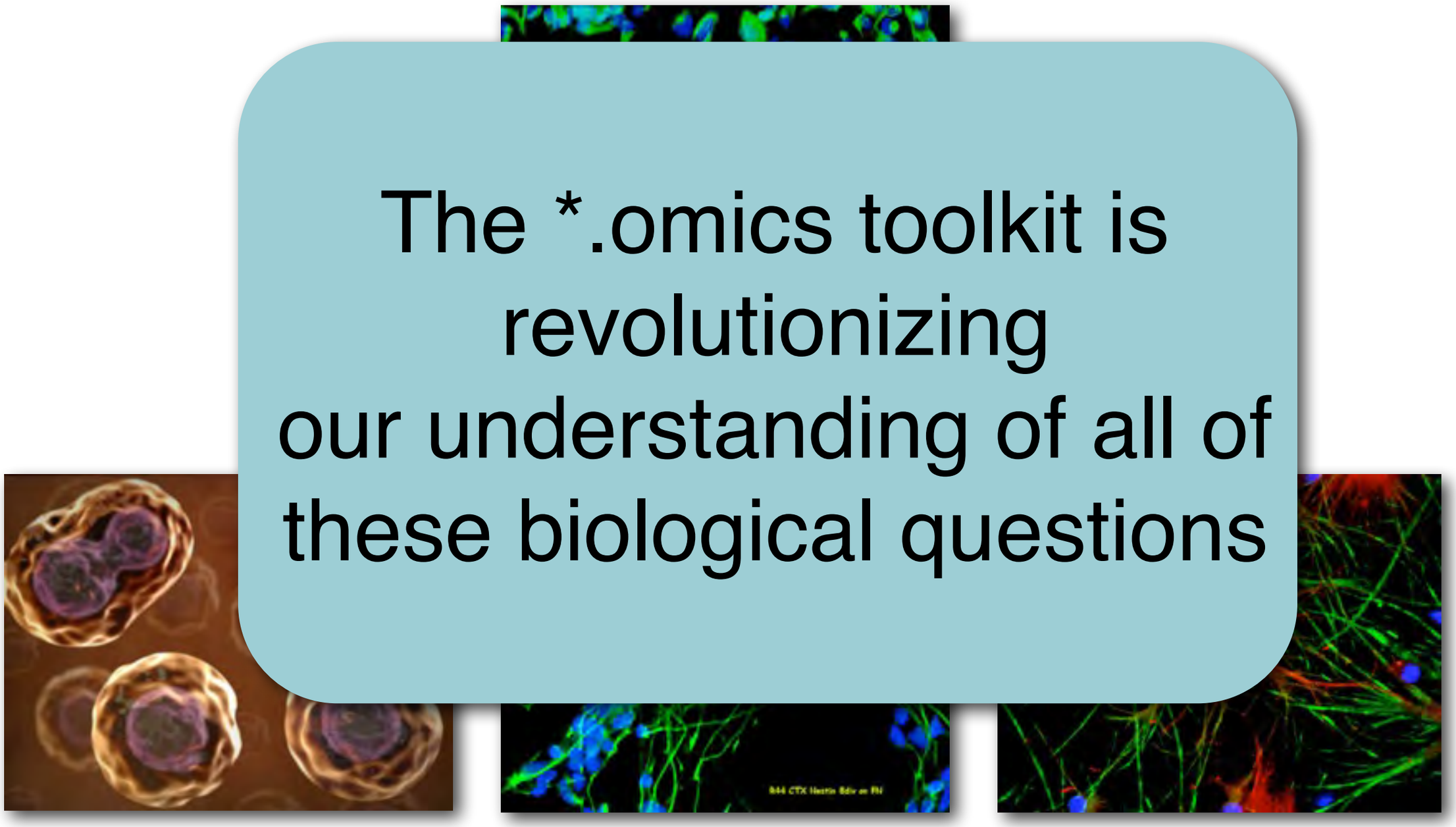








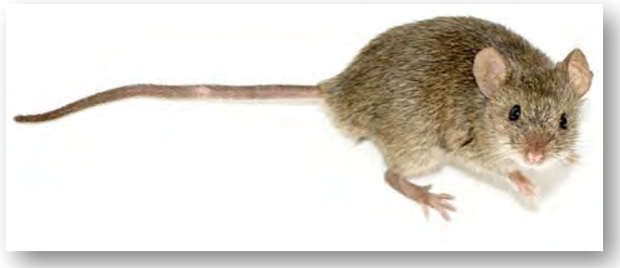
How is cellular functional diversity created?



The *.omics toolkit is revolutionizing our understanding of all of these biological questions

(1998)

(2000)



(2002)

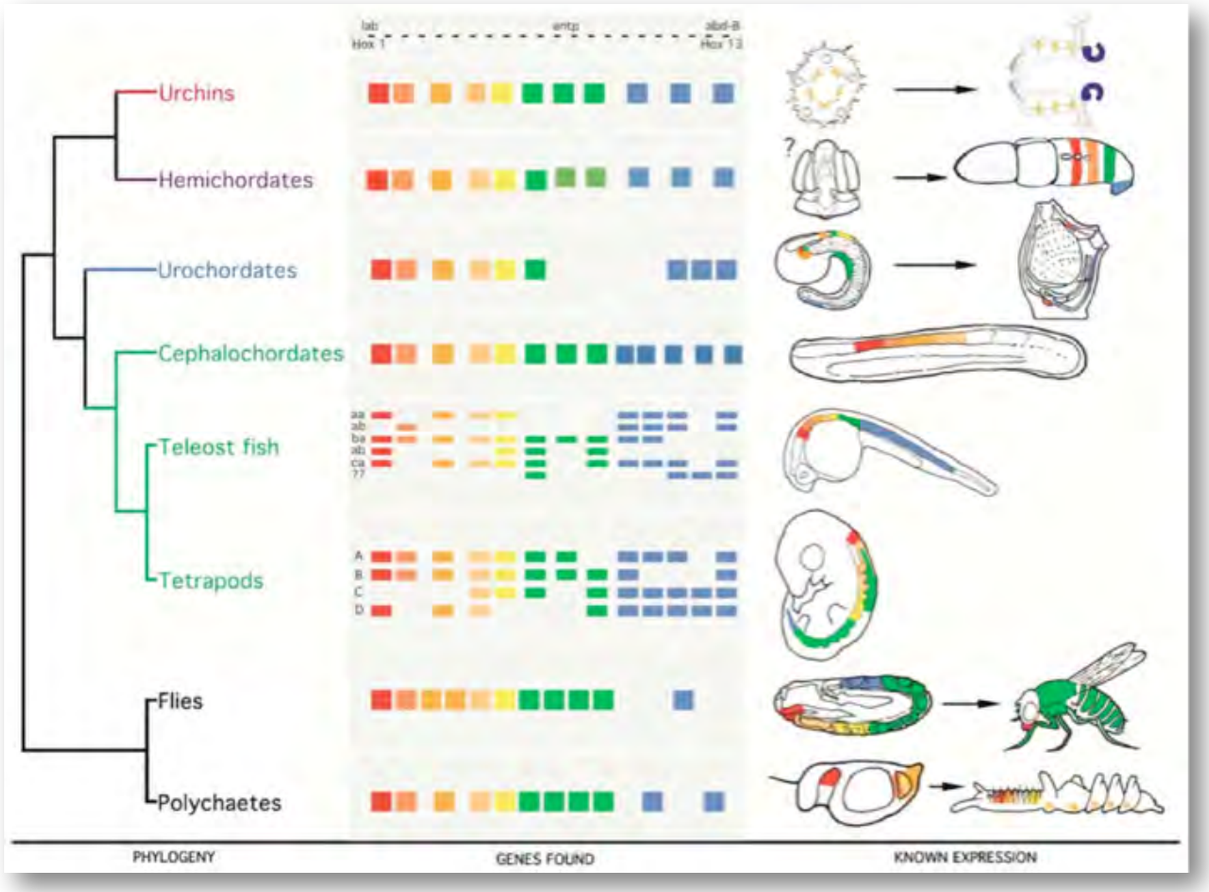
(2006)

(2006)

(2002)

(2007)

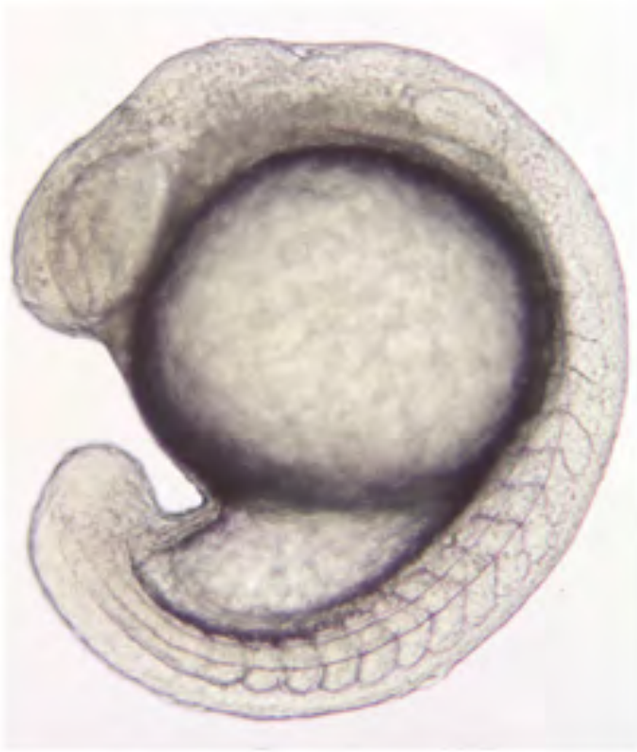
(2006)



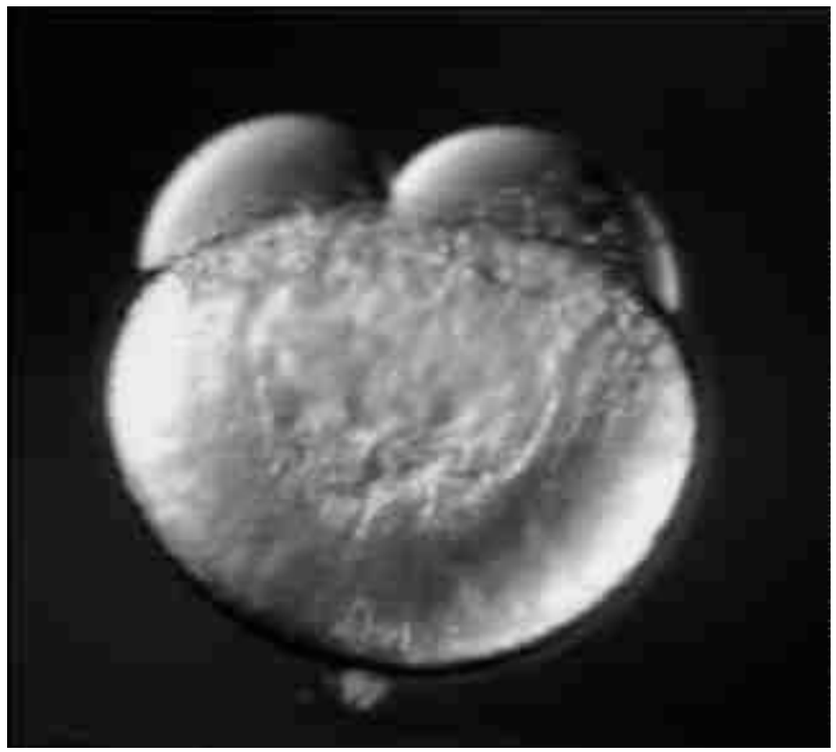
Comparative Genomics

Vertebrate **zygotes** or embryos

28 day human

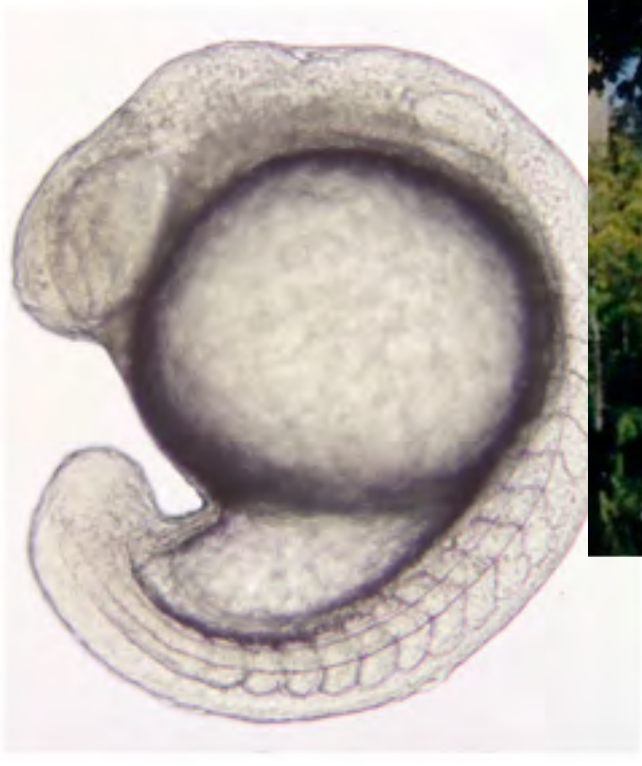


19h zebrafish



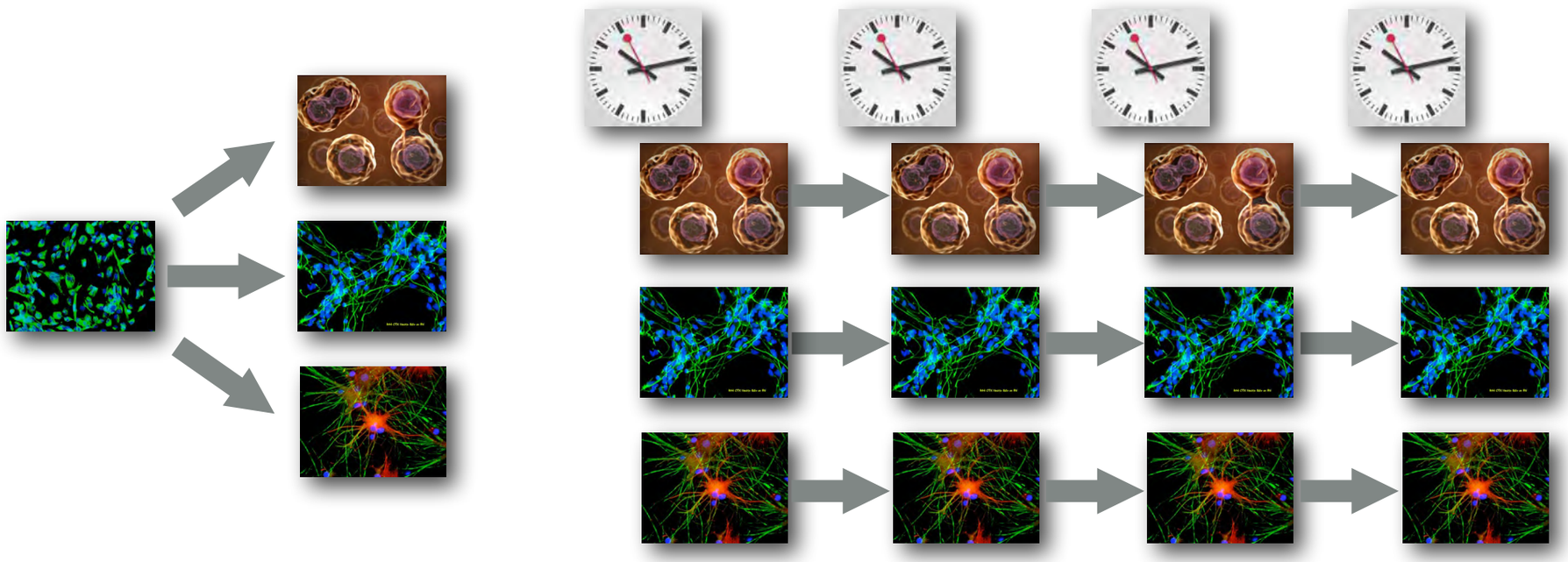
Vertebrate **zygotes** or embryos

28 day human



19h zebrafish

Dr. Catchen in his ‘following Phish Phase’

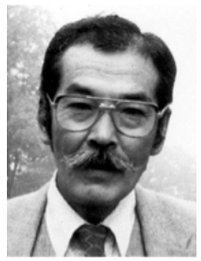


Functional Genomics







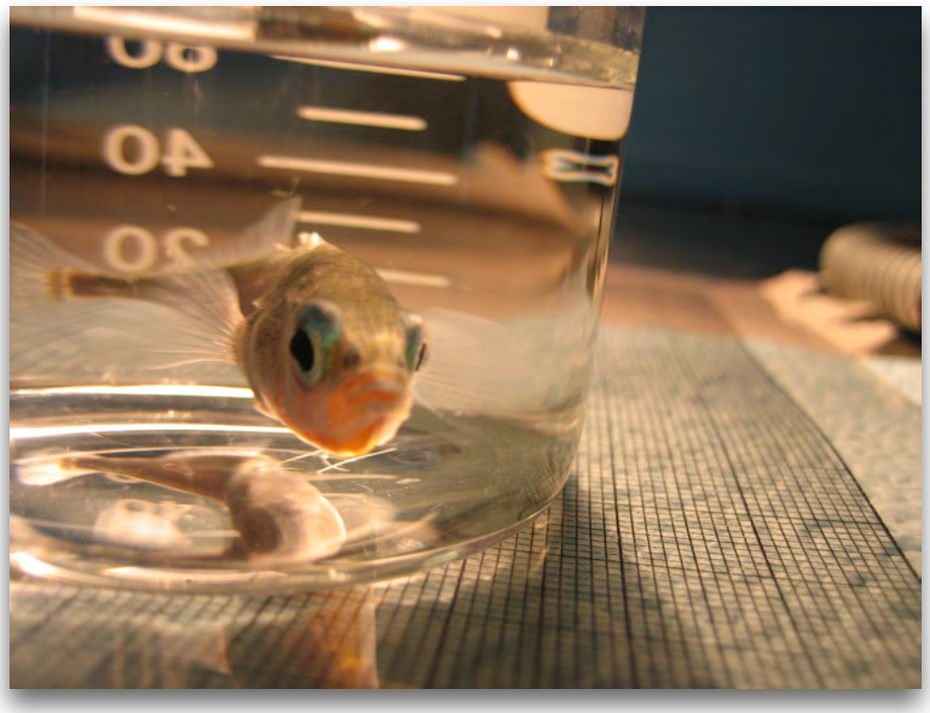




Population Genomics

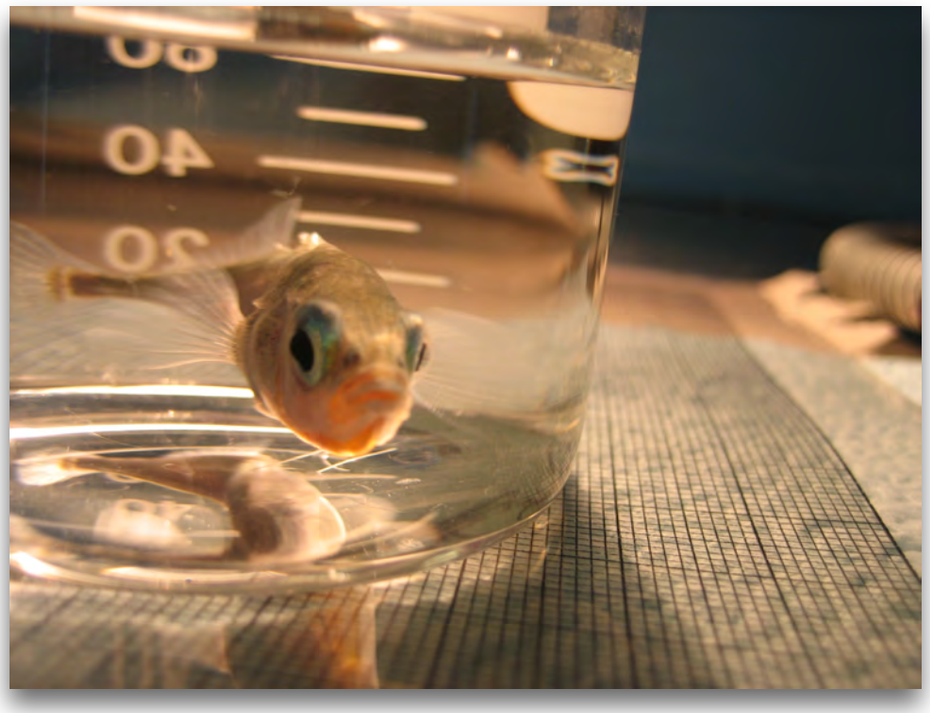
How do we ‘genomically enable’ research studies of non-model organisms?

1. Genetic markers & genetic maps
2. Physical maps (genomes)
3. Transcriptomes
4. Gene expression analyses
5. Epigenetic analyses



How do we ‘genomically enable’ research studies of non-model organisms?

1. **Genetic markers & genetic maps**
2. Physical maps (genomes)
3. Transcriptomes
4. Gene expression analyses
5. Epigenetic analyses



Assaying genetic variation:
Shouldn't we just sequence everything?

Why not just sequence entire genomes??

- Still prohibitively expensive for many studies

- For many studies a full sequence isn’t necessary
    - genomes of many organisms are organized in linkage blocks
    - well spaced markers will provide the necessary coverage

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

Alternative approach - Reduced representation NGS for genotyping

- Restriction-site Associated DNA Sequencing (RADseq)
- Focus sequencing on homologous regions across the genome
- Simultaneous identification and typing of single nucleotide polymorphisms (SNPs) and haplotypes
- The cost will be a fraction of the cost of resequencing the genome
- Thousands of genomes can be assayed in just a few weeks

What is RADseq?

(Restriction-site Associated DNA)

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 [illegible], 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)

22,830 *SbfI* sites in threespine stickleback
~ 45,000 RAD-Tags

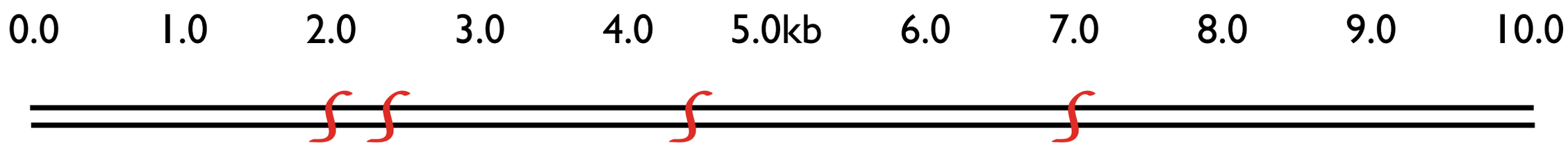
HiSeq2500 Illumina Lane:
160 million reads

HiSeq4000 Illumina Lane:
350 million reads

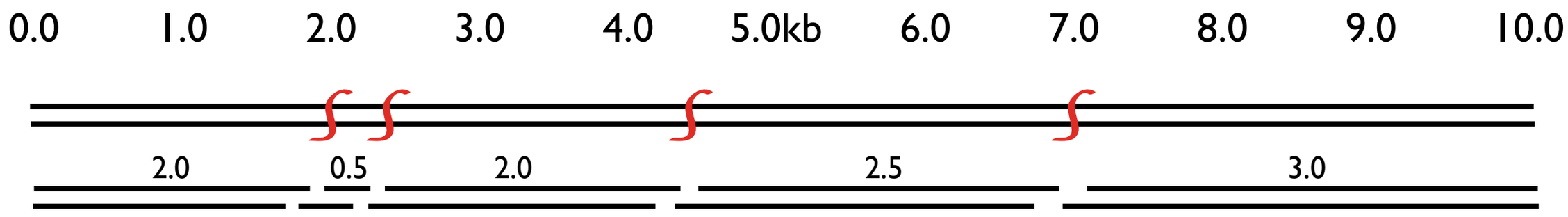
Let’s go into a little more detail

Restriction Enzyme (RE) digestion and first adaptor ligation

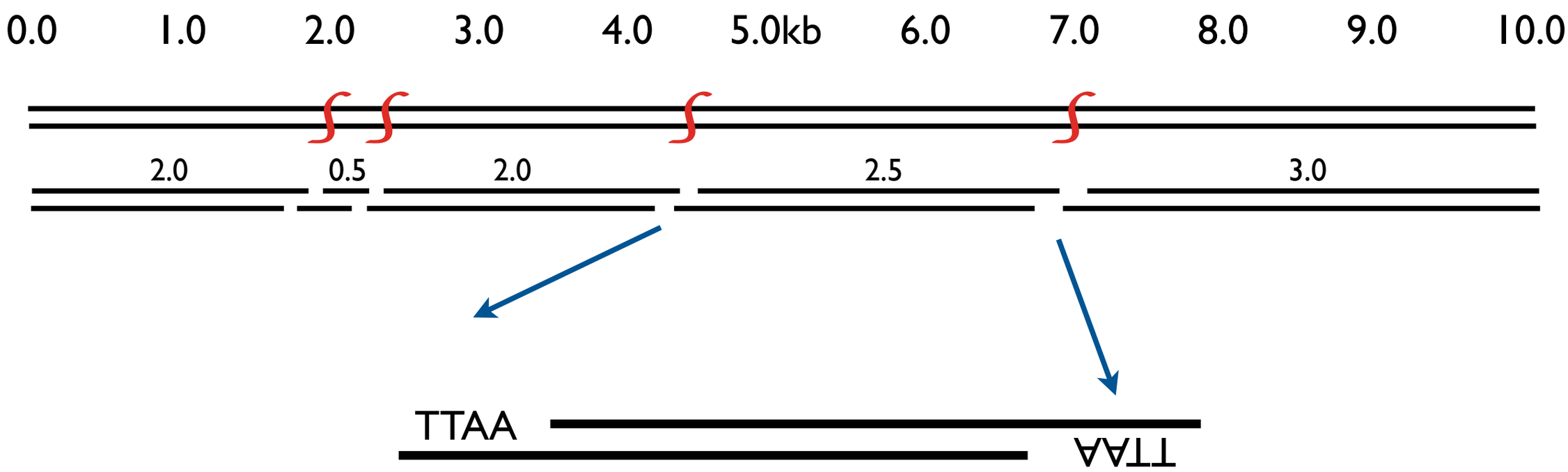
Restriction Enzyme (RE) digestion and first adaptor ligation



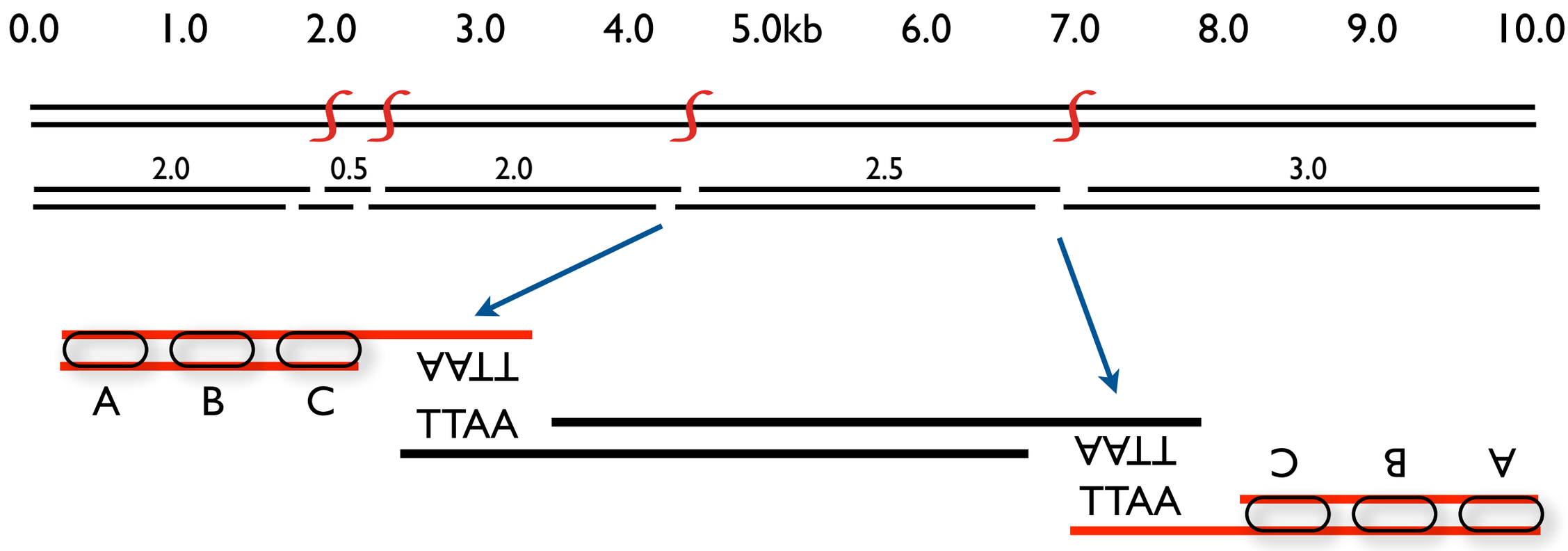
Restriction Enzyme (RE) digestion and first adaptor ligation



Restriction Enzyme (RE) digestion and first adaptor ligation

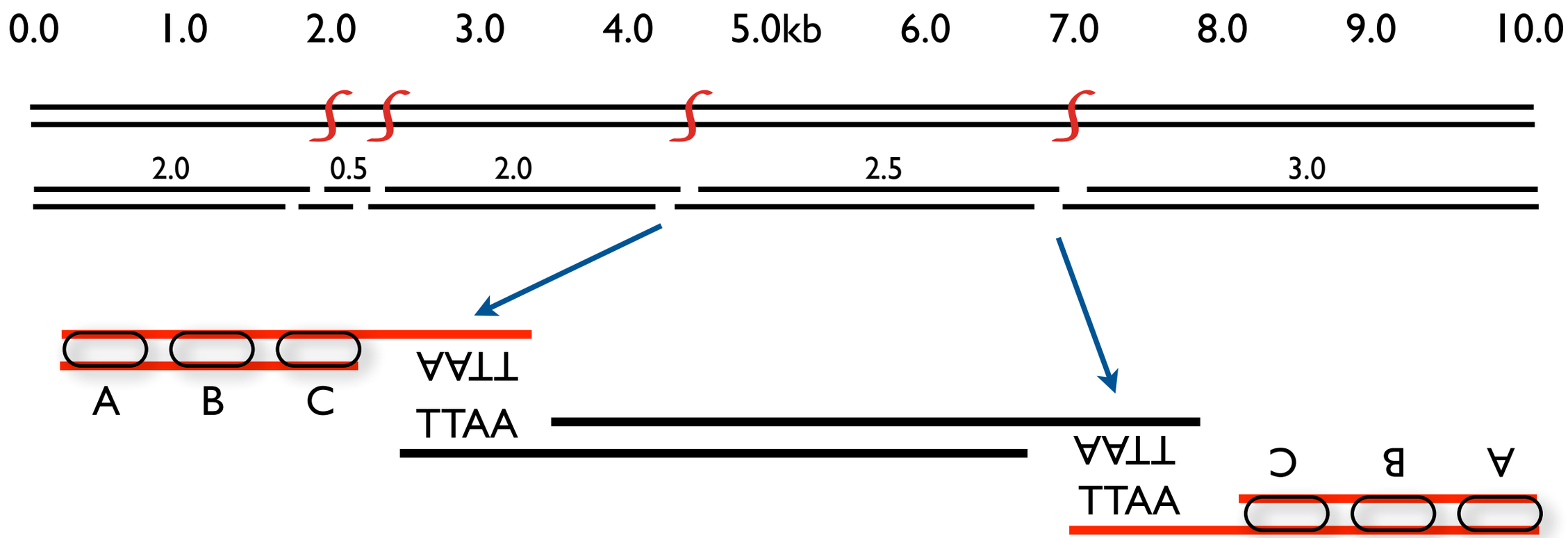


Restriction Enzyme (RE) digestion and first adaptor ligation



A = Amplification primer
B = Sequencing primer
C = Barcode

Restriction Enzyme (RE) digestion and first adaptor ligation

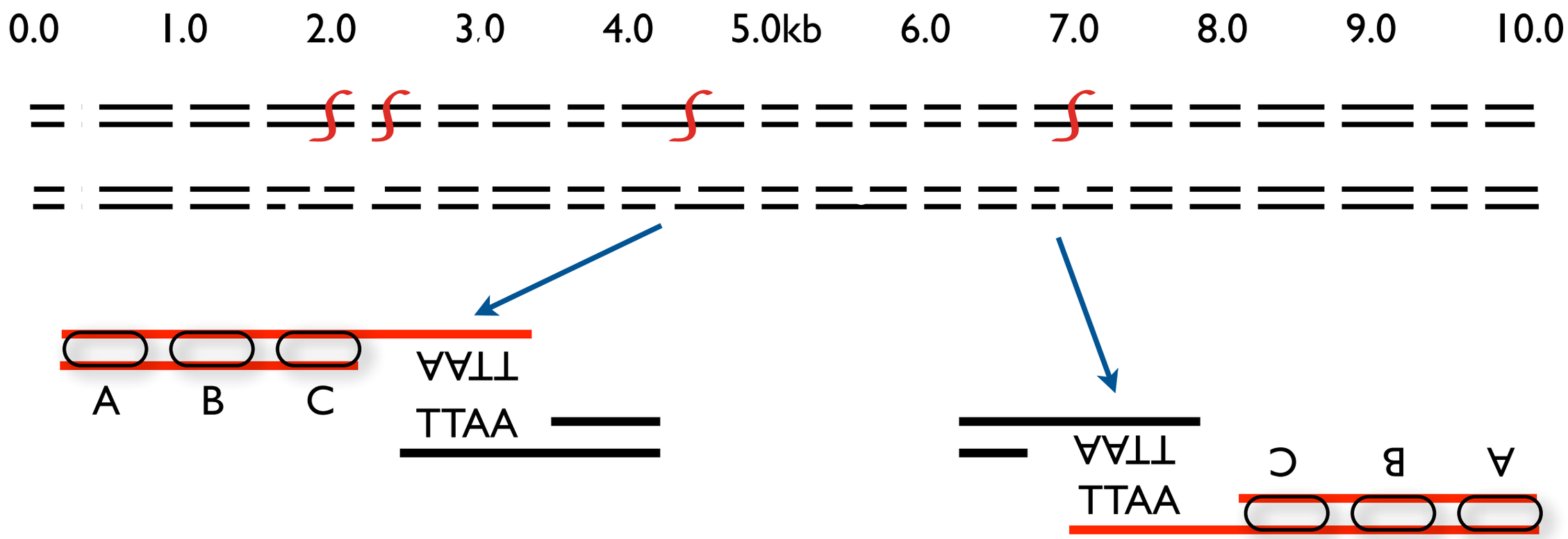


A = Amplification primer
B = Sequencing primer
C = Barcode

Note - there are now commonly two levels of barcodes used:

Sample Barcodes and
Molecular Identification Barcodes (MIPs)

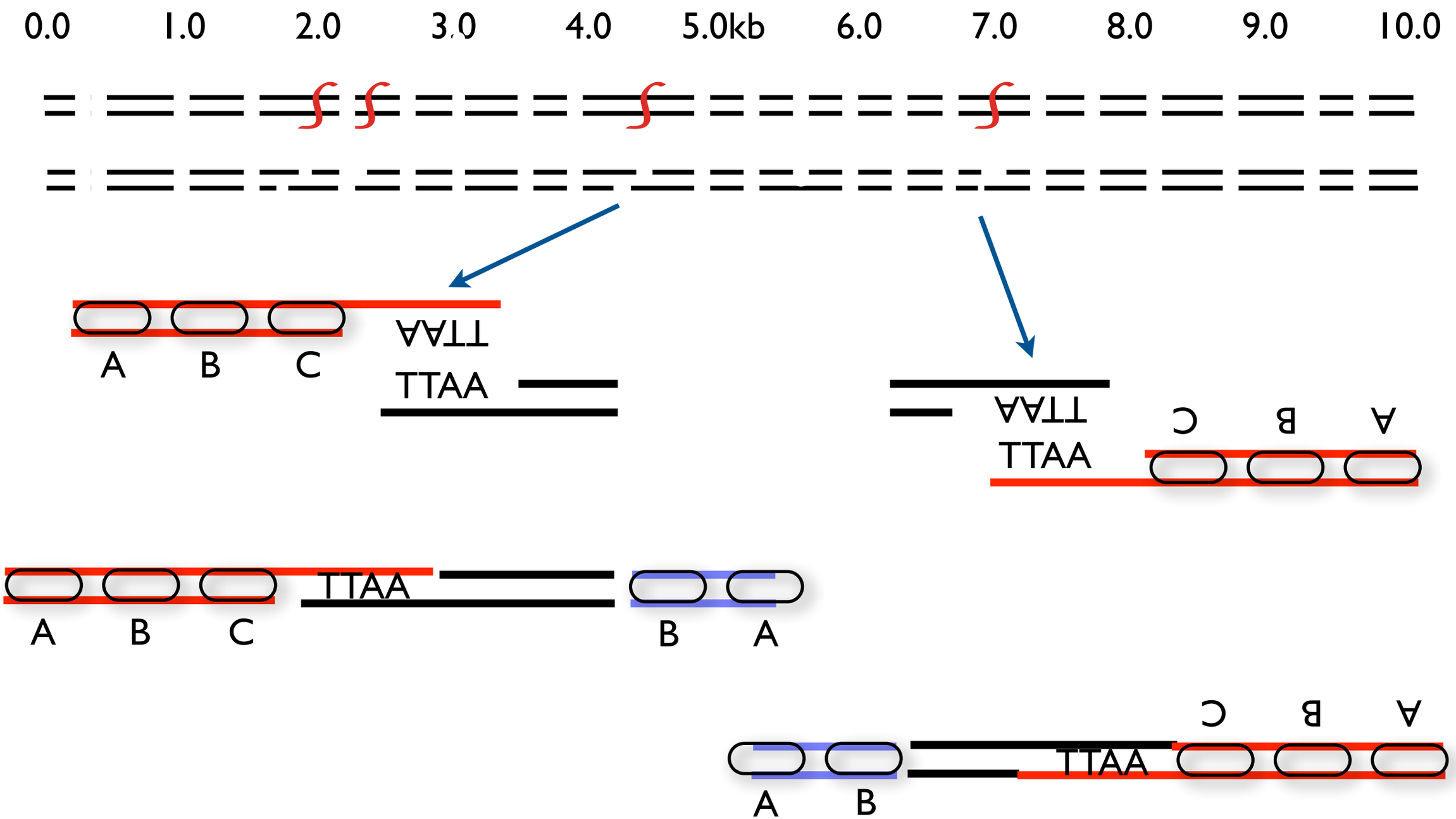
Shearing and second adaptor ligation



* Defining difference between original RAD and other approaches*

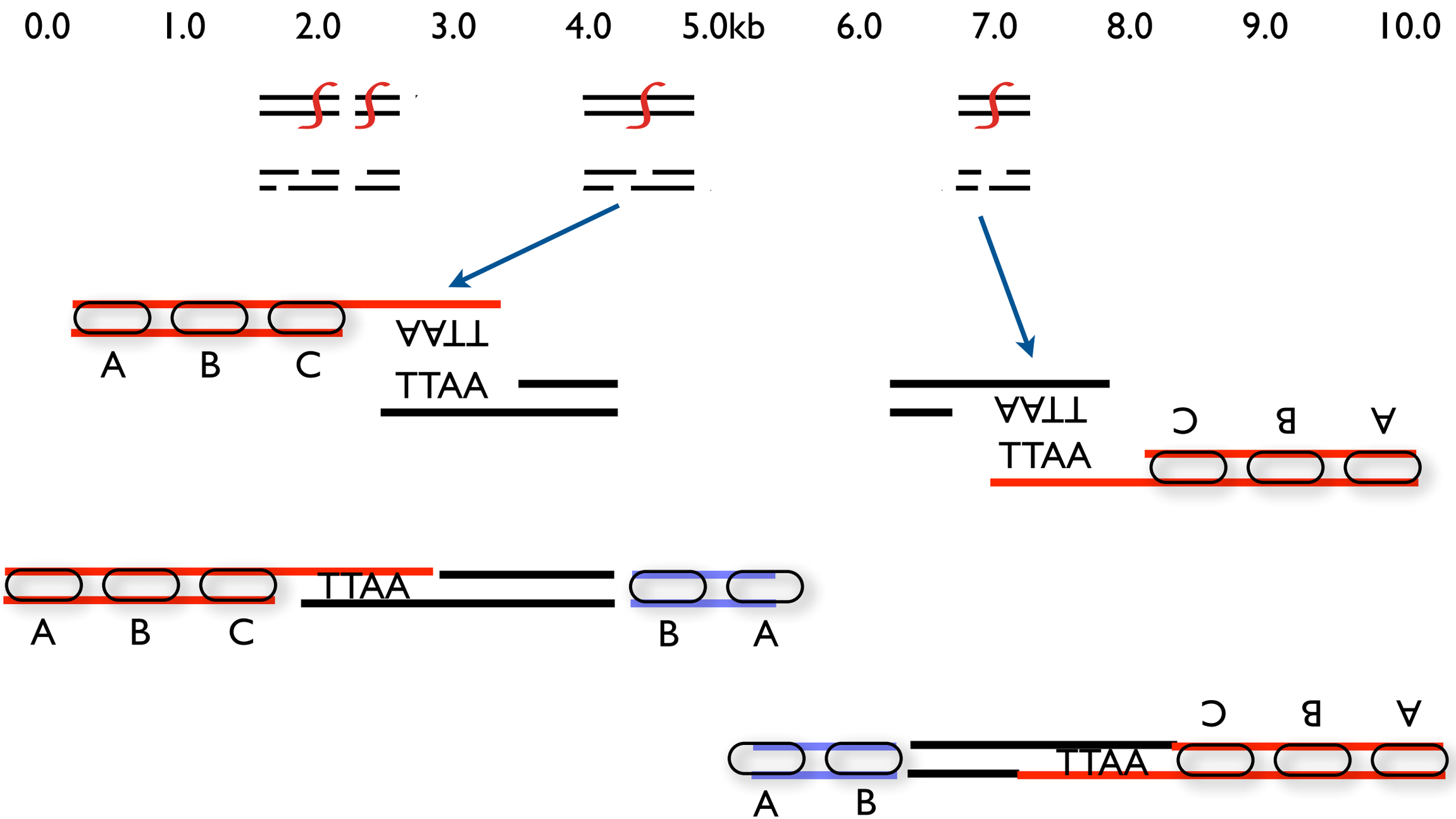
A = Amplification primer
B = Sequencing primer
C = Barcode

Shearing and second adaptor ligation



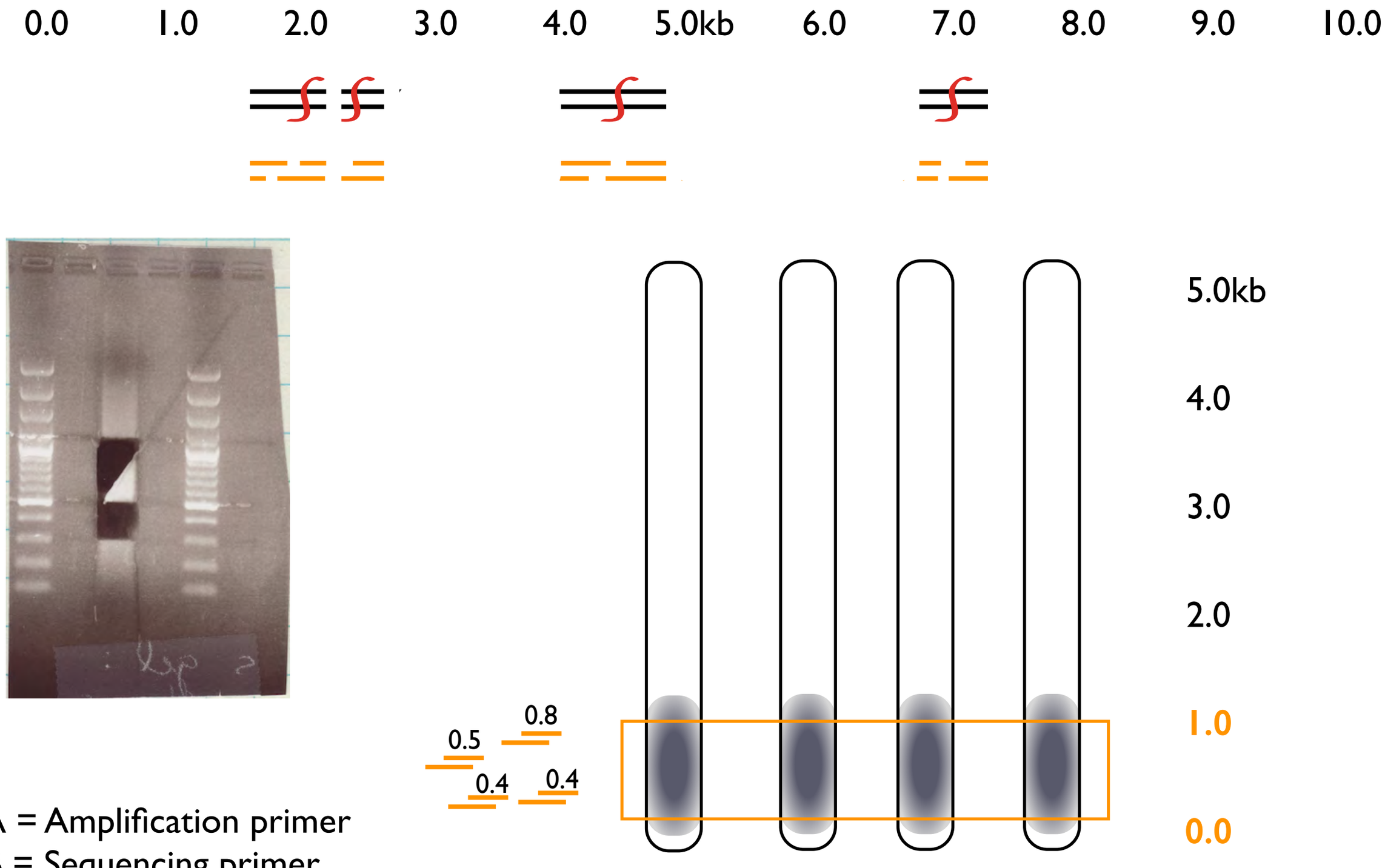
A = Amplification primer
B = Sequencing primer
C = Barcode

Shearing and second adaptor ligation



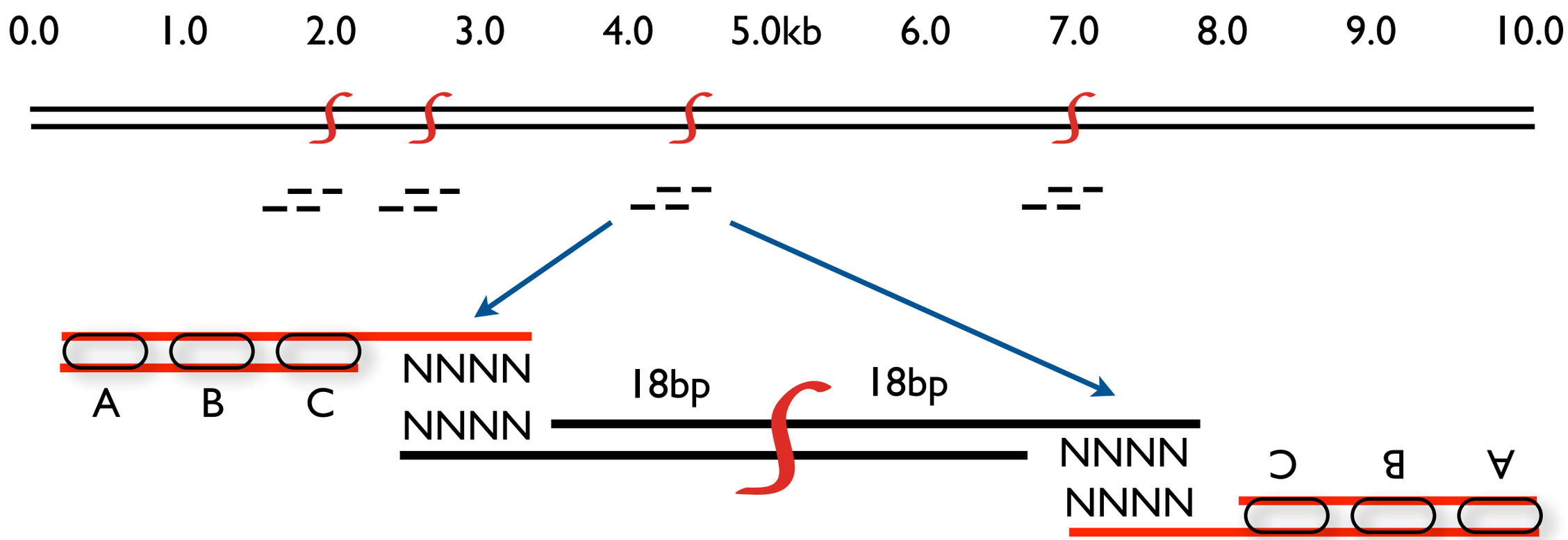
A = Amplification primer
B = Sequencing primer
C = Barcode

Shearing makes consistent fragments for sequencing



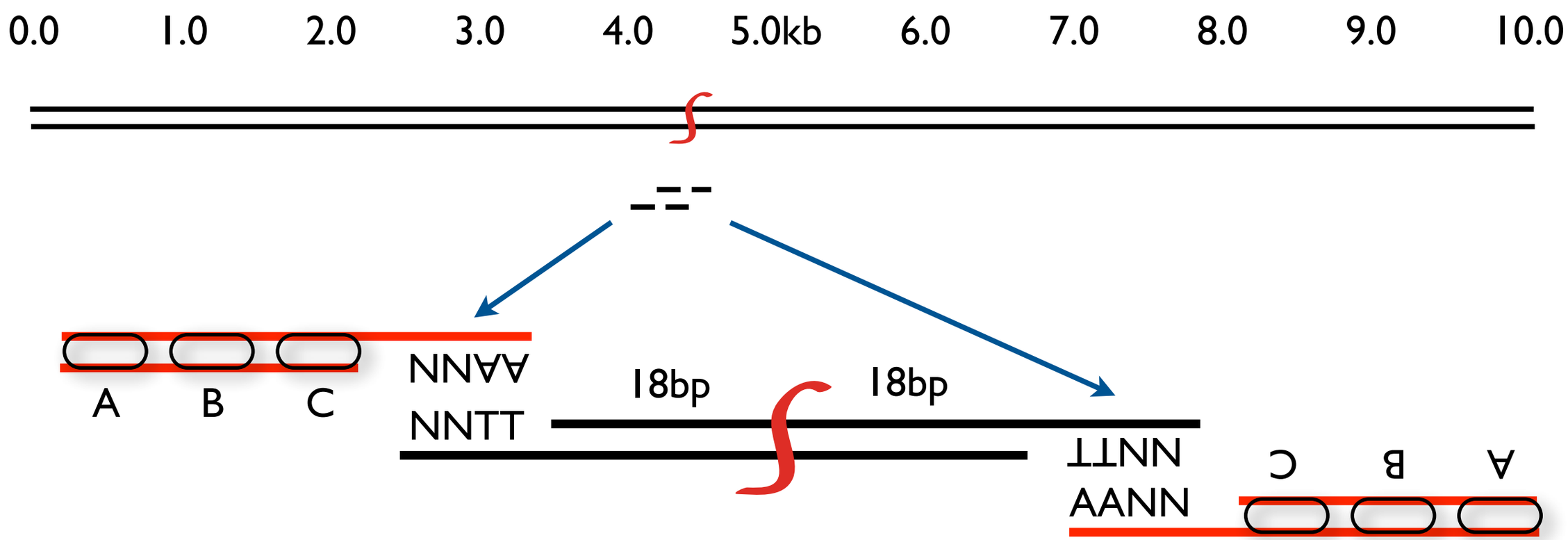
A = Amplification primer
B = Sequencing primer
C = Barcode

2bRAD - type 2b restriction enzyme



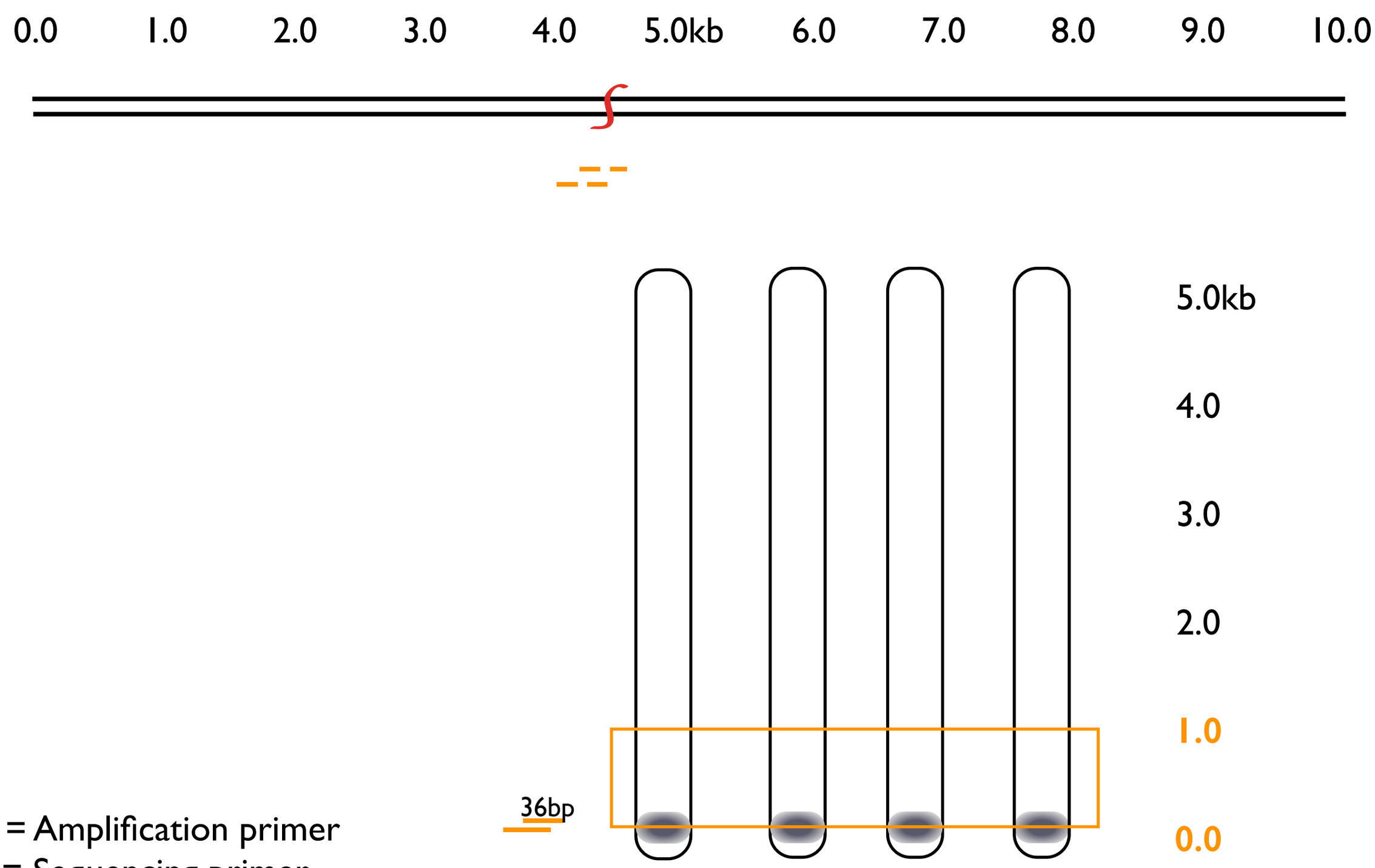
A = Amplification primer
B = Sequencing primer
C = Barcode

2bRAD - can scale number of markers easily



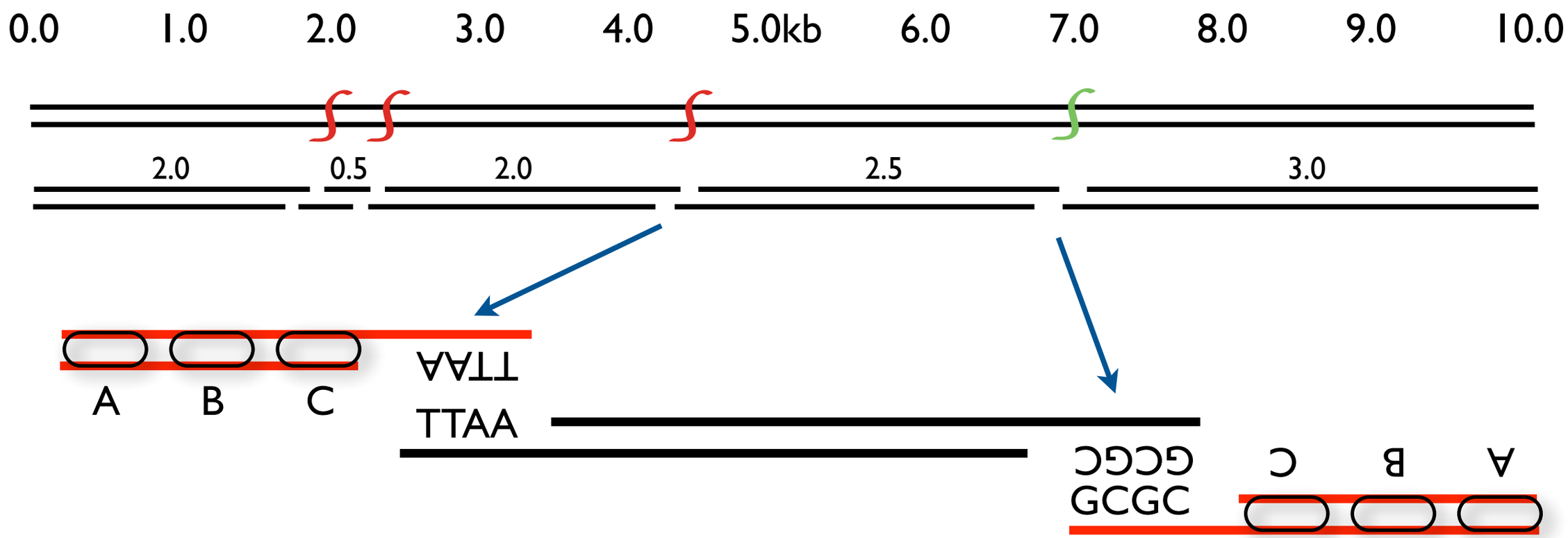
A = Amplification primer
B = Sequencing primer
C = Barcode

2bRAD - size selection is difficult



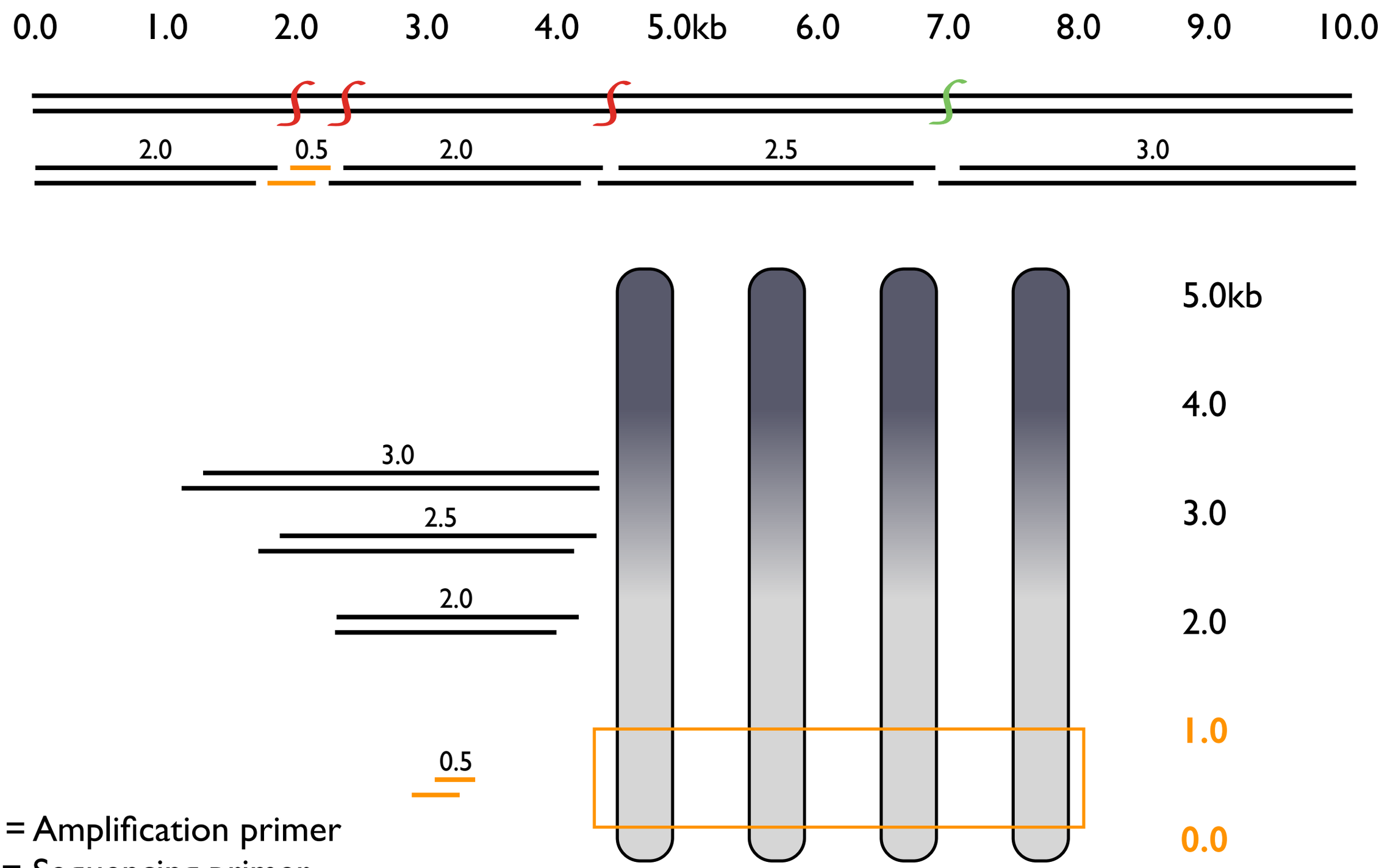
A = Amplification primer
B = Sequencing primer
C = Barcode

Single (GBS) or Double Digest RAD (ddRAD)



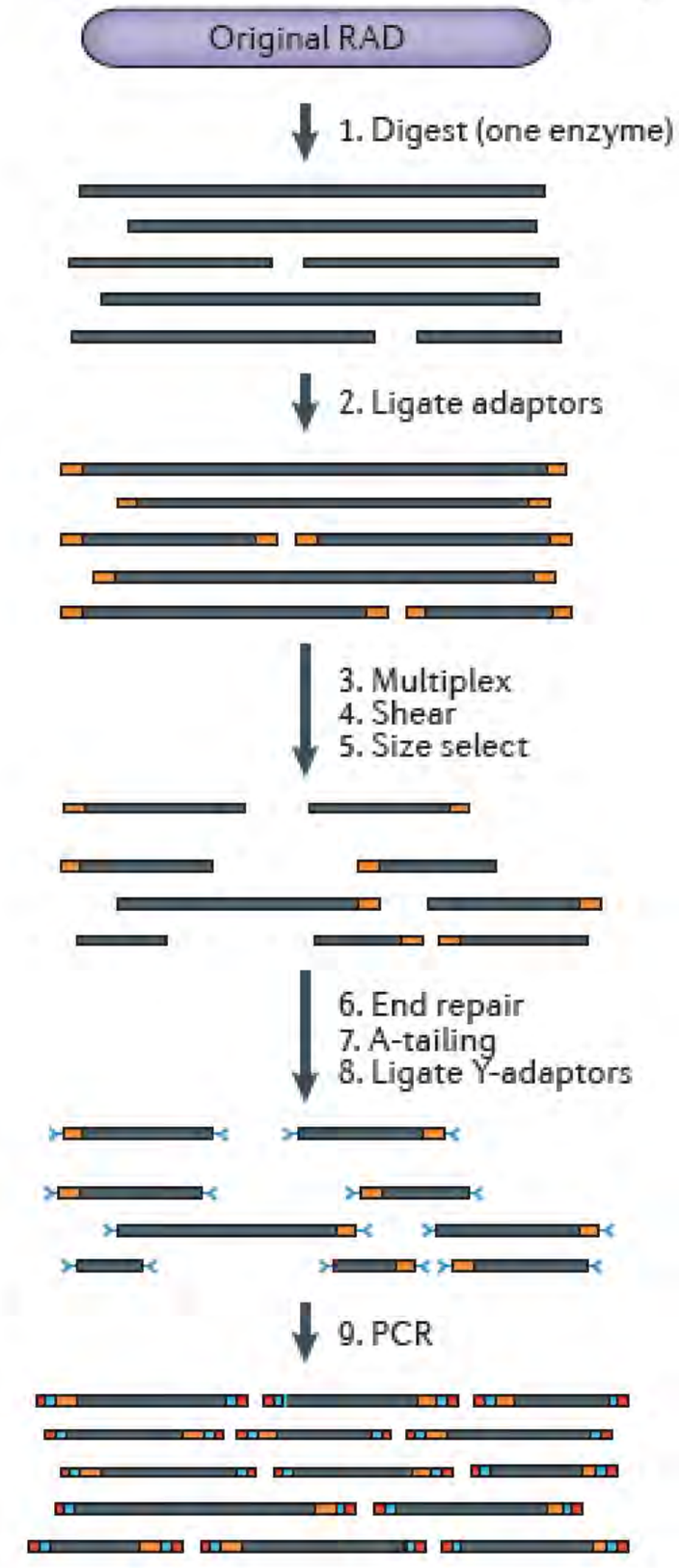
A = Amplification primer
B = Sequencing primer
C = Barcode

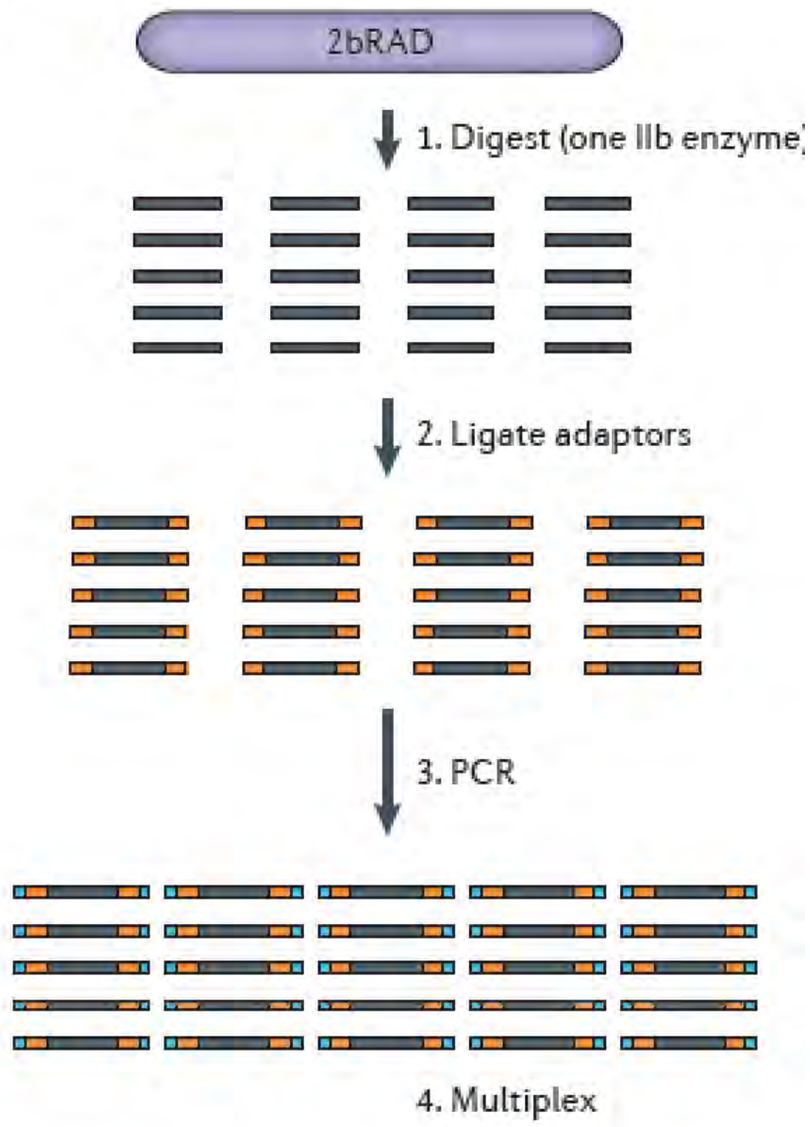
Size selection is more problematic without shearing



A = Amplification primer
B = Sequencing primer
C = Barcode

RADseq with one enzyme digestion & shearing



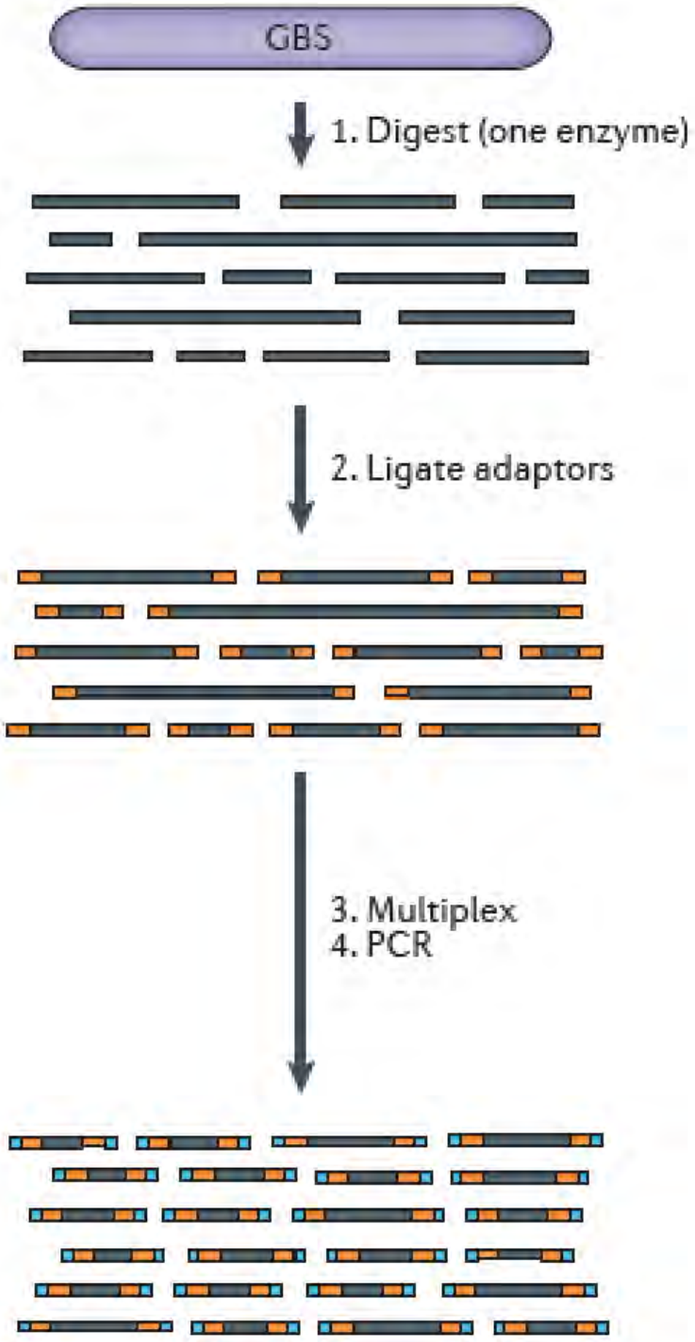


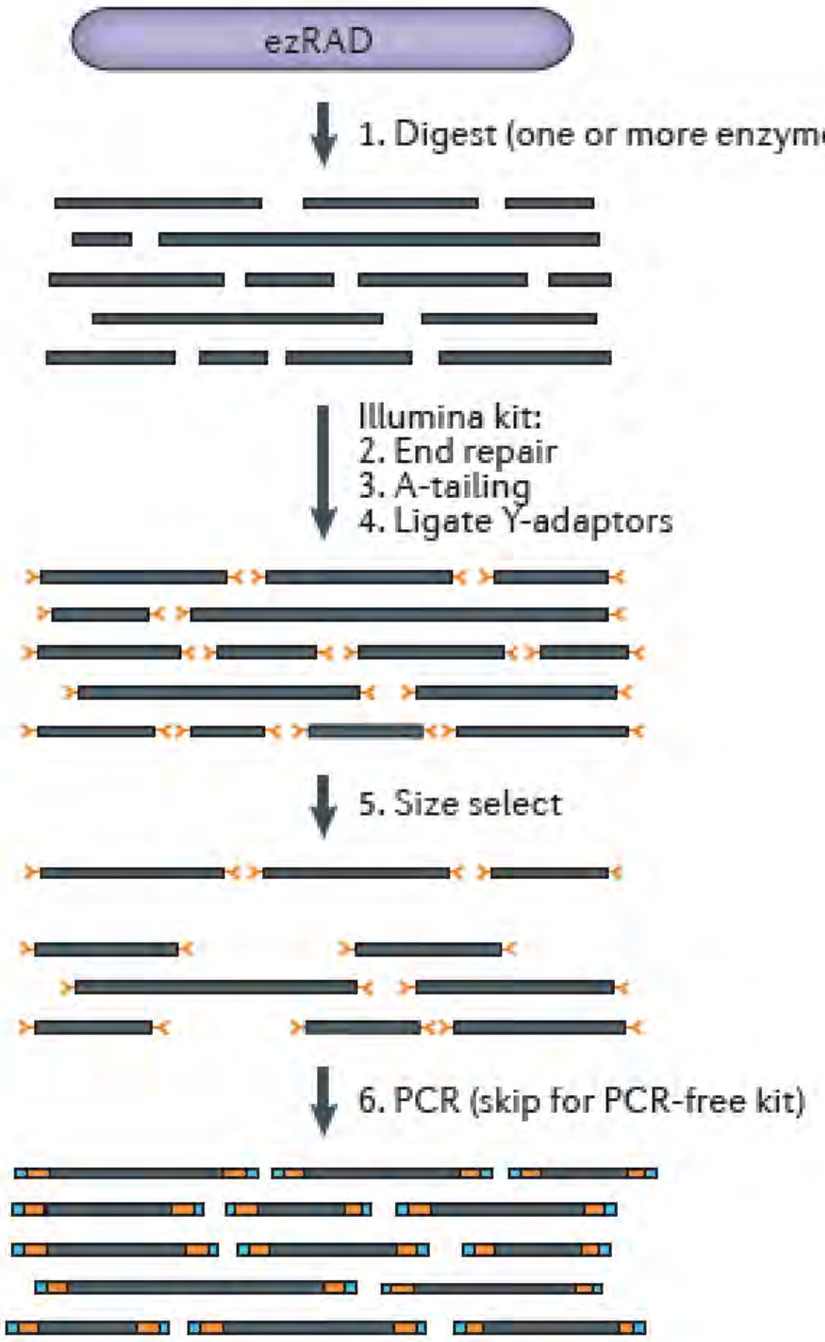
STUDY DESIGNS

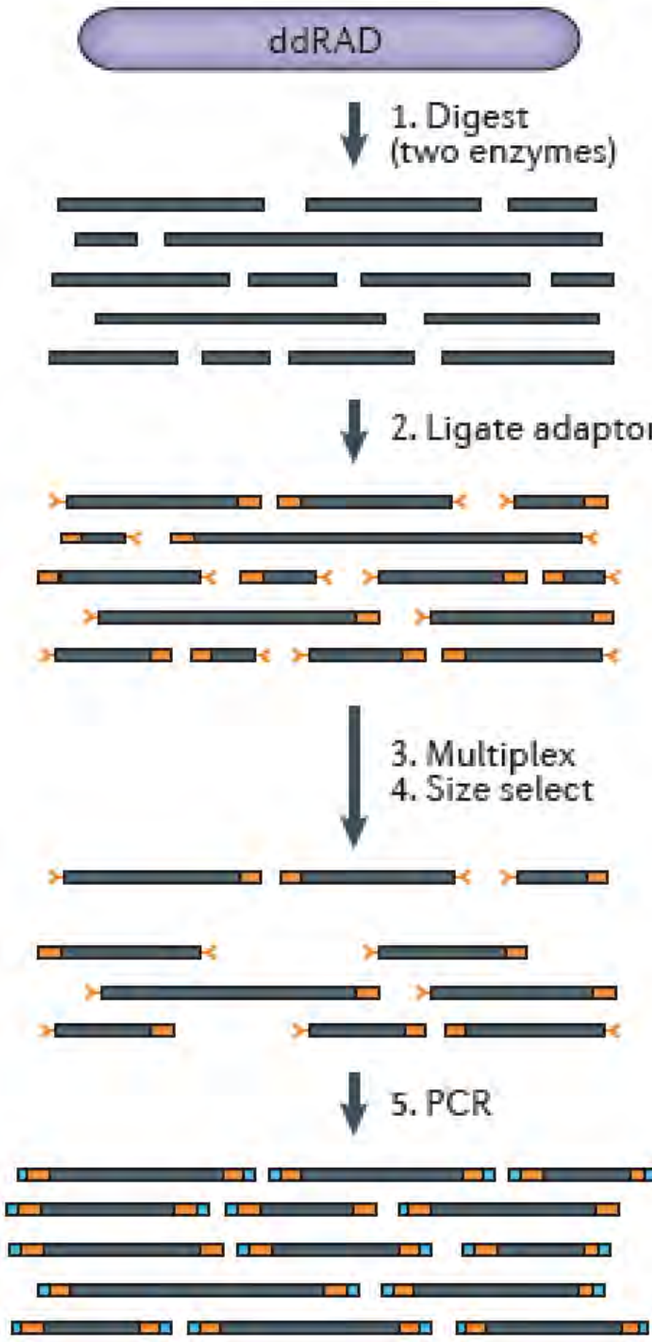
Harnessing the power of RADseq for ecological and evolutionary genomics

Kimberly R. Andrews[1], Jeffrey M. Good[2], Michael R. Miller[3], Gordon Luikart[4] and Paul A. Hohenlohe[5]

RADseq with one or two enzyme digestion







Summary of plusses and minuses of RAD family

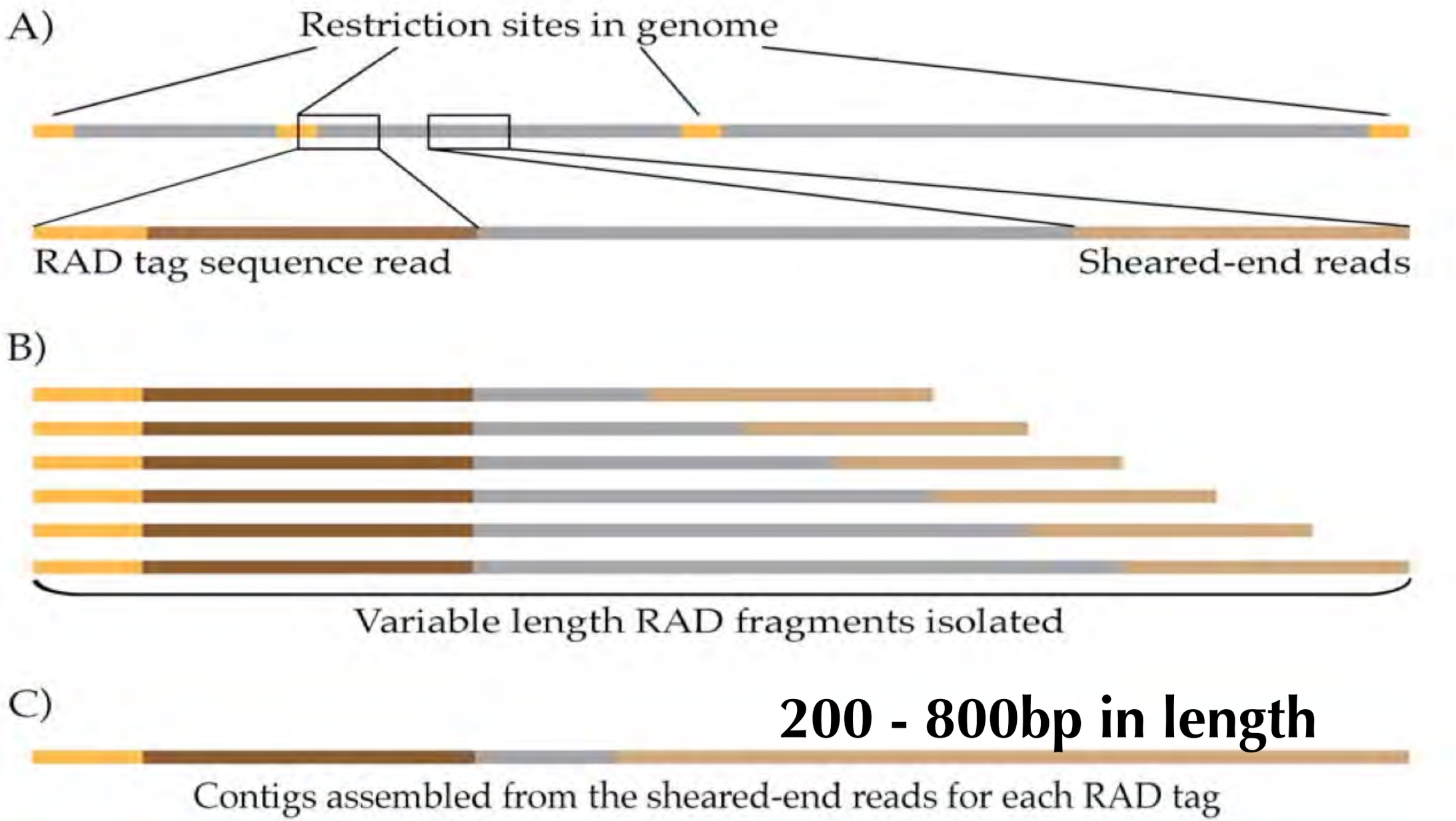
STUDY DESIGNS

Harnessing the power of RADseq for ecological and evolutionary genomics

Kimberly R. Andrews[1], Jeffrey M. Good[2], Michael R. Miller[3], Gordon Luikart[4] and Paul A. Hohenlohe[5]

	Original RAD	2bRAD	GBS	ddRAD	ezRAD
Options for tailoring number of loci	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme	Change restriction enzyme or size selection window	Change restriction enzyme or size selection window
Number of loci per 1 Mb of genome size*	30–500	50–1,000	5–40	0.3–200	10–800
Length of loci	≤1kb if building contigs; otherwise ≤300 bp‡	33–36 bp	<300 bp‡	≤300 bp‡	≤300 bp‡
Cost per barcoded or indexed sample	Low	Low	Low	Low	High
Effort per barcoded or indexed sample§	Medium	Low	Low	Low	High
Use of proprietary kit	No	No	No	No	Yes
Identification of PCR duplicates	With paired-end sequencing	No	With degenerate barcodes	With degenerate barcodes	No
Specialized equipment needed	Sonicator	None	None	Pippin Prep‖	Pippin Prep‖
Suitability for large or complex genomes¶	Good	Poor	Moderate	Good	Good
Suitability for *de novo* locus identification (no reference genome)#	Good	Poor	Moderate	Moderate	Moderate
Available from commercial companies	Yes	No	Yes	Yes	No

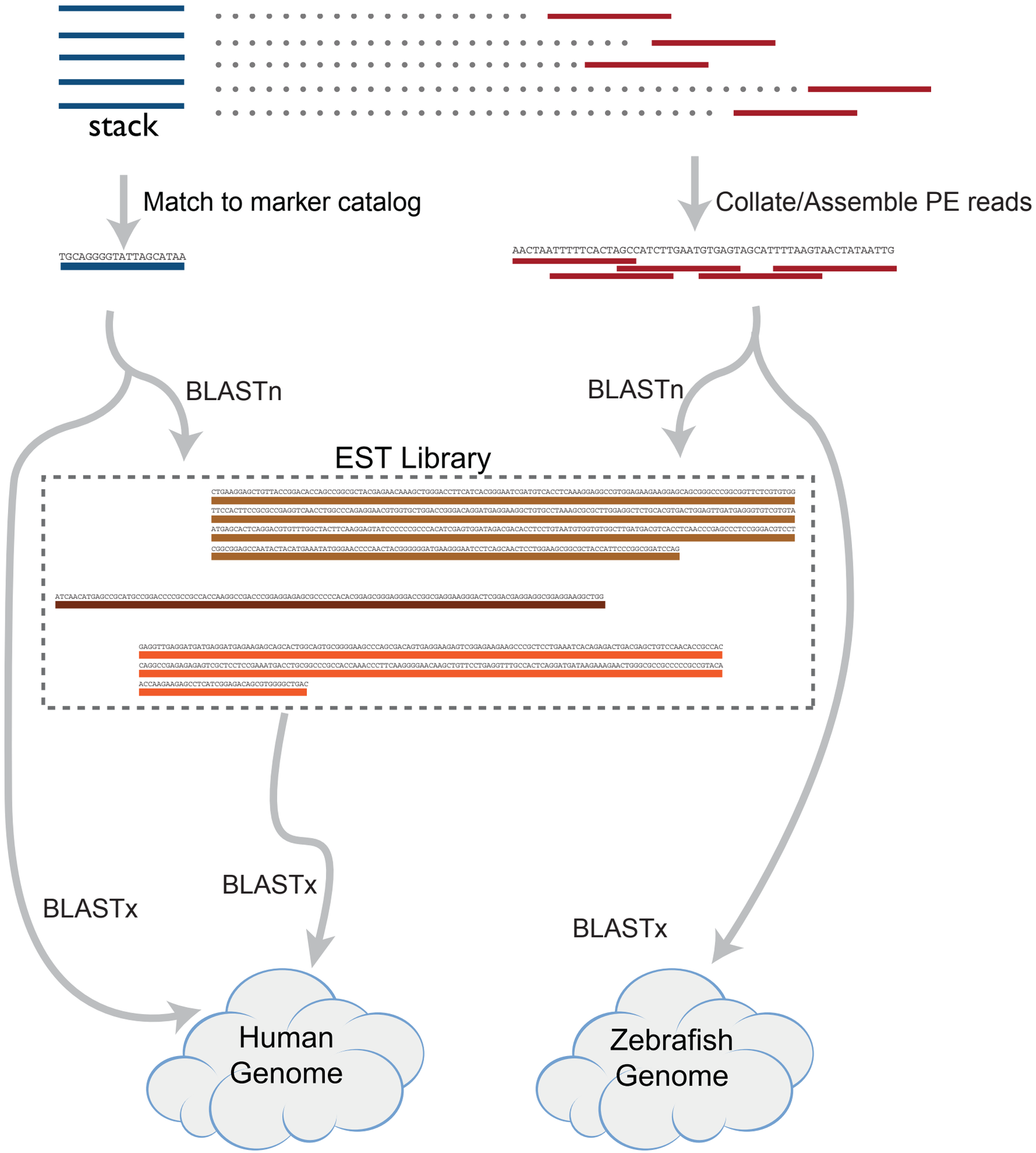
Random shearing in original RAD - Local Paired End (PE) Assemblies



Acquire paired-end sequence

Associate markers / PE contigs with ESTs

Assign orthology to:
- markers
- PE contigs
- ESTs

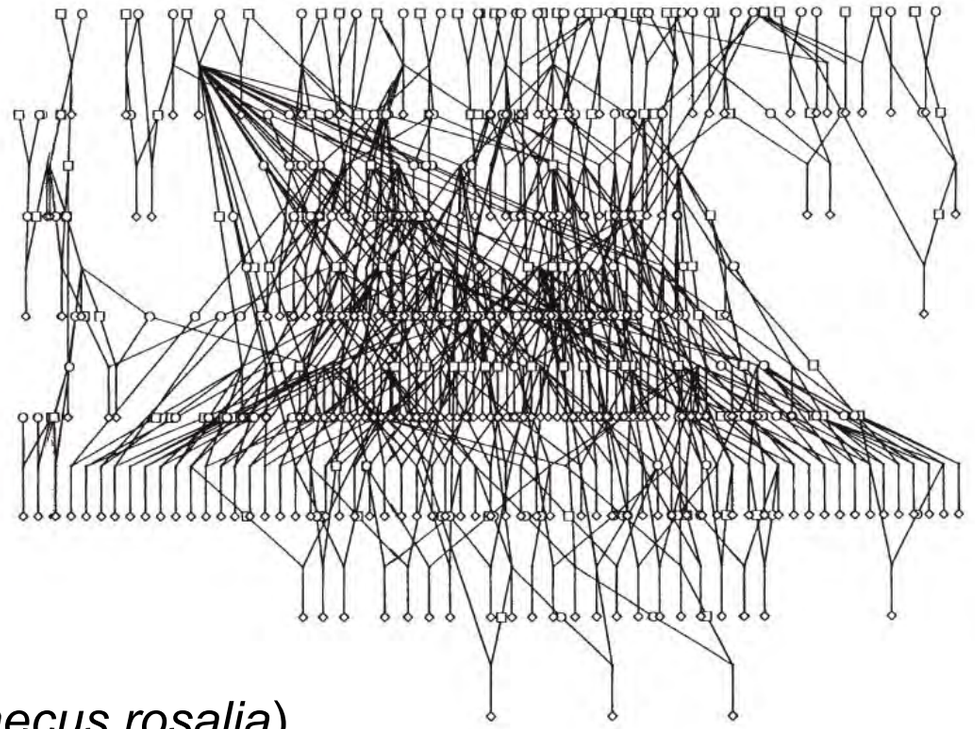


For what types of studies can RADseq be useful?

Identifying genetically distinct individuals and estimating genetic diversity

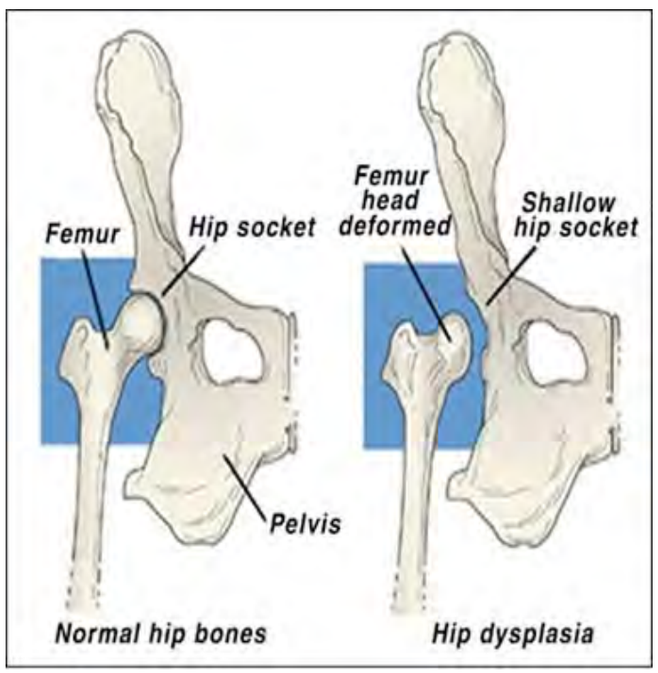
Quaking Aspens

Defining the relationships among individuals and populations

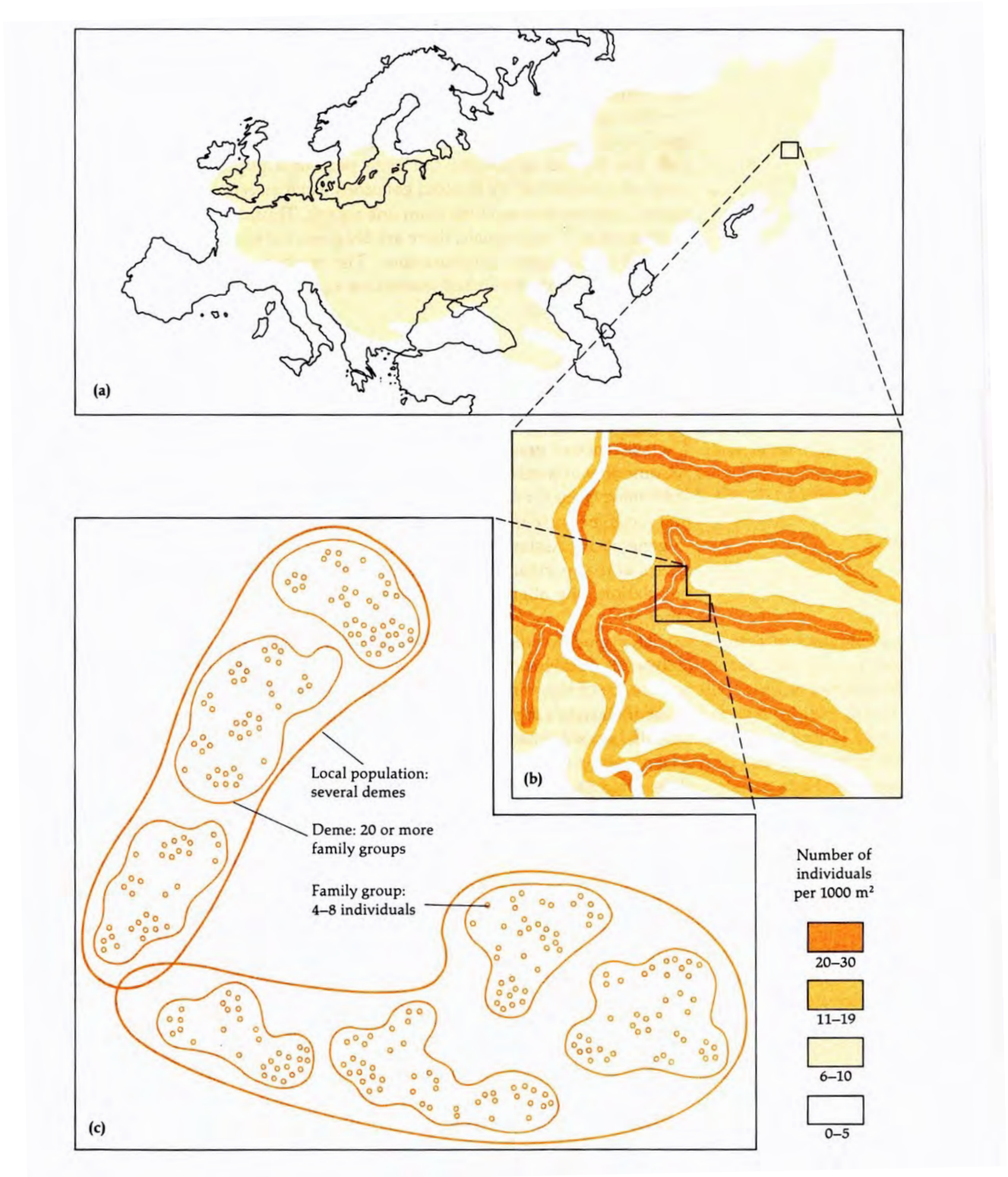


The golden lion tamarin (*Leontopithecus rosalia*).

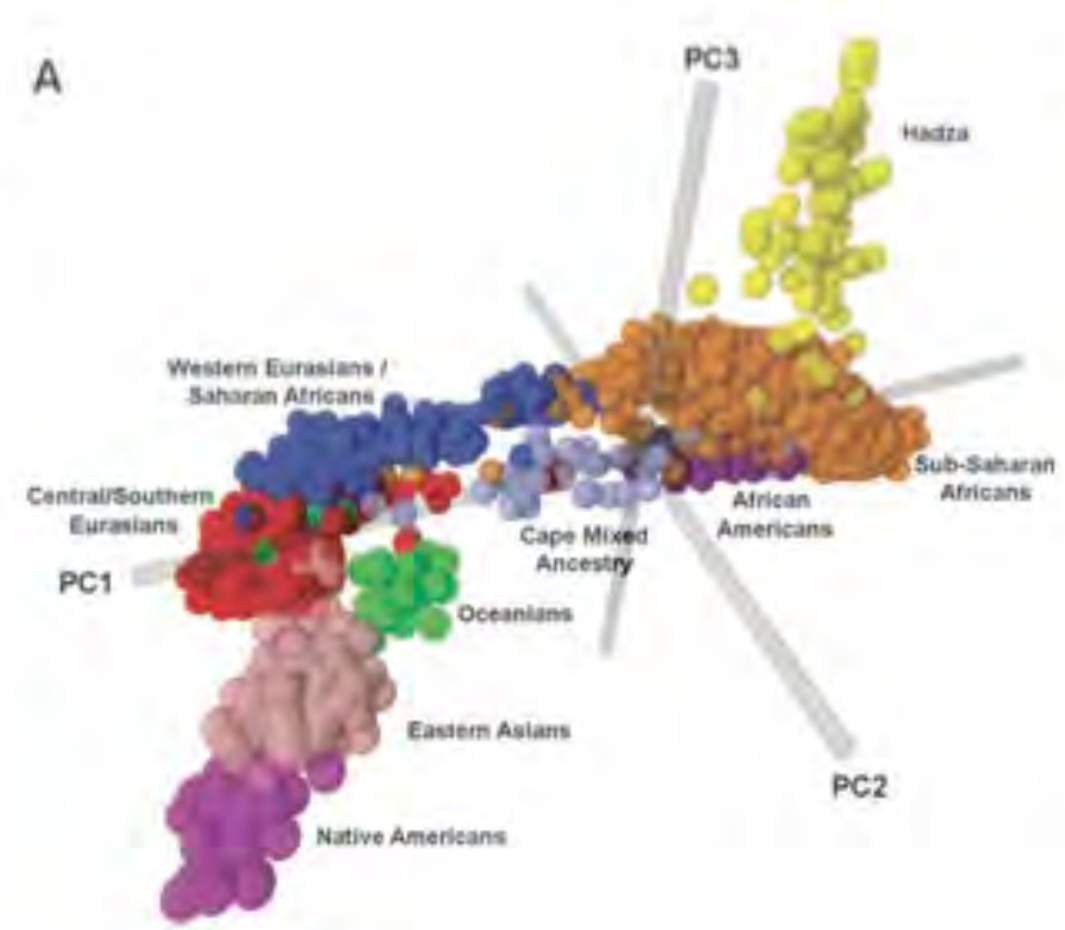
Precisely quantifying the amount of inbreeding in wild and captive populations

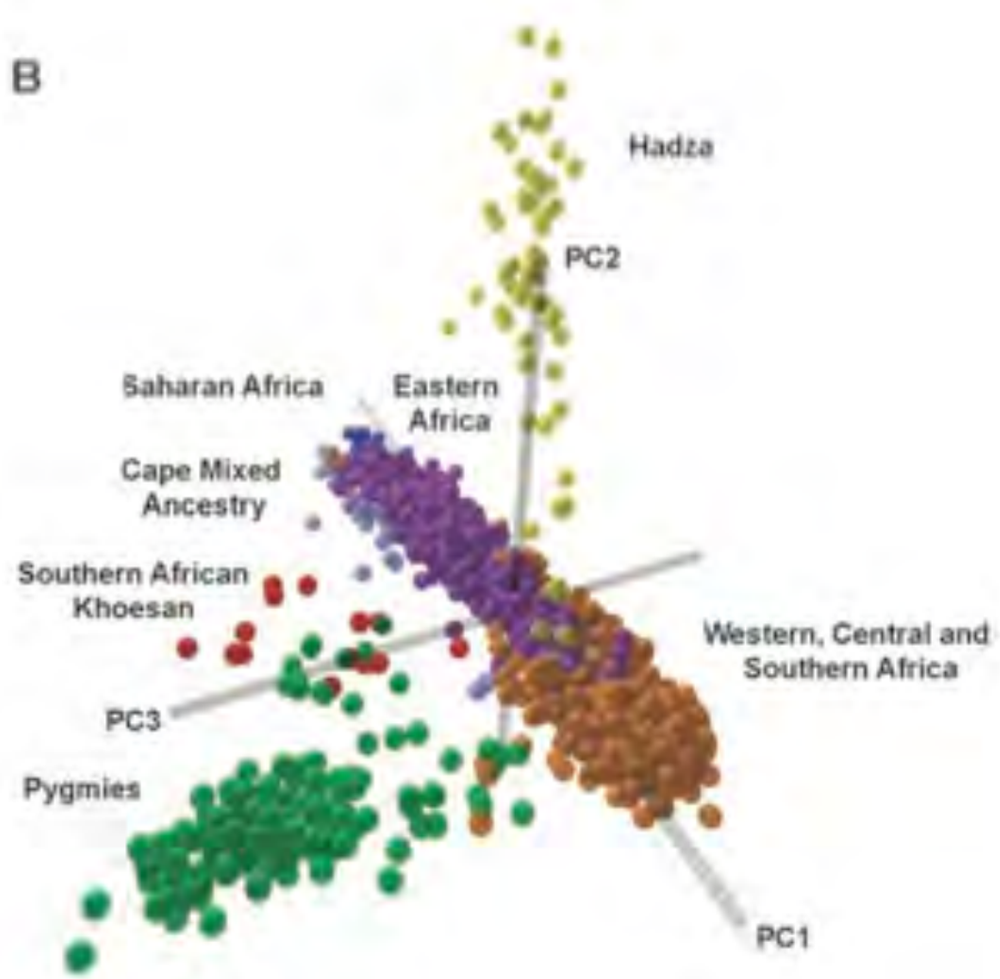


Defining the relationships among individuals and populations

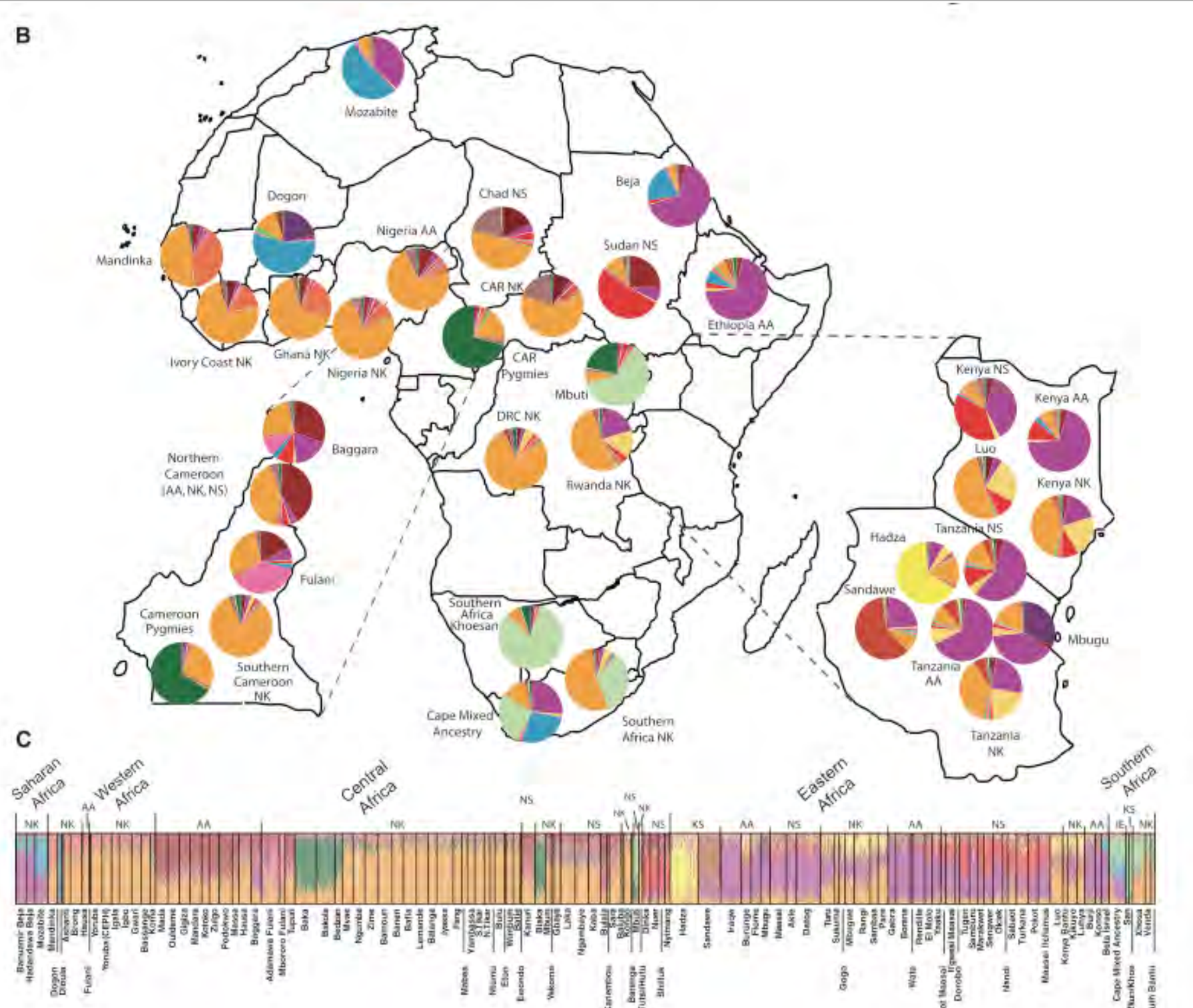


Multivariate Analysis of Genetic Data - Tiskoff et al. Structure of African Human populations

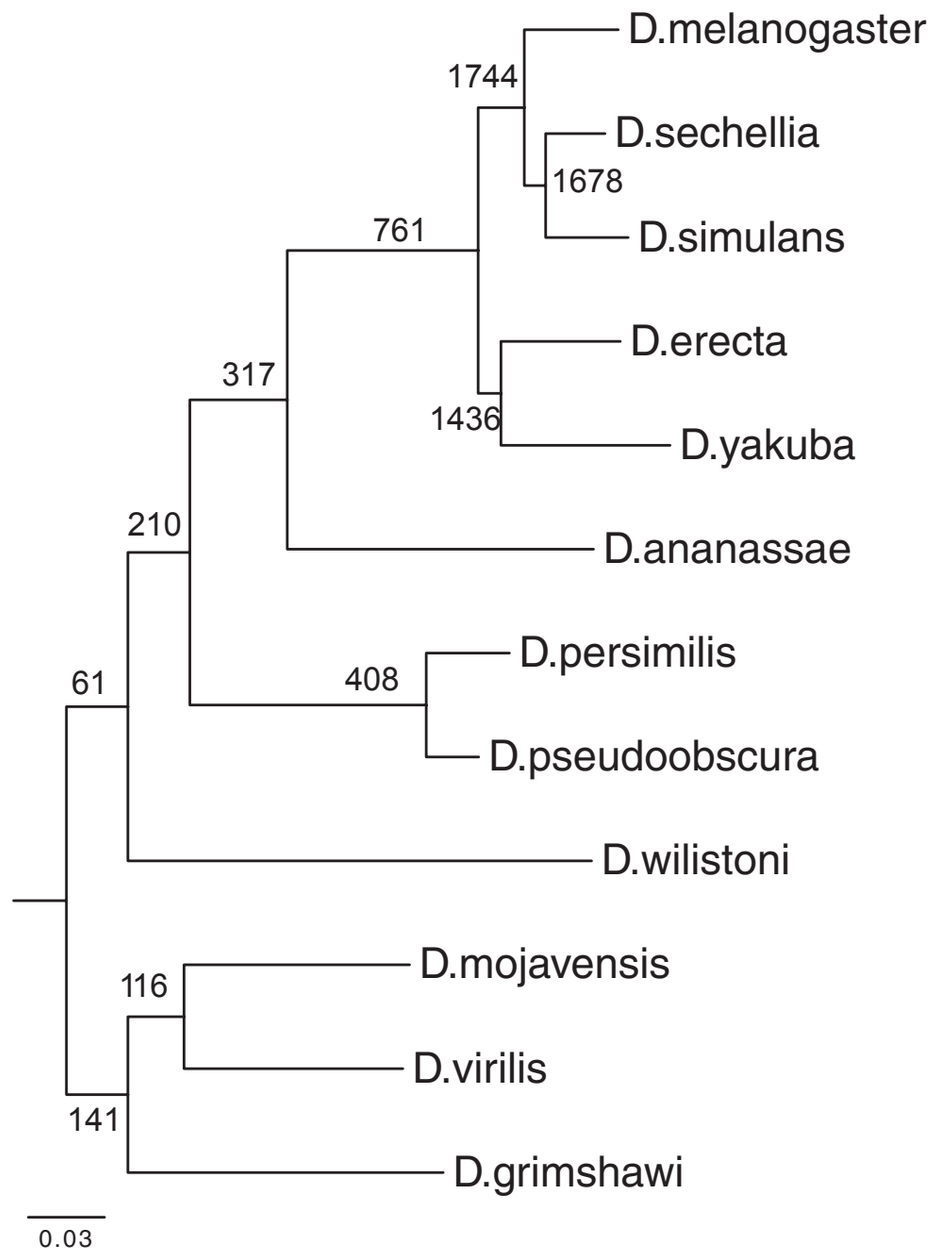




Multivariate Analysis of Genetic Data - Tiskoff et al. Structure of African Human populations



Phylogenetic relationships



Ecology and Evolution

Open Access

Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization

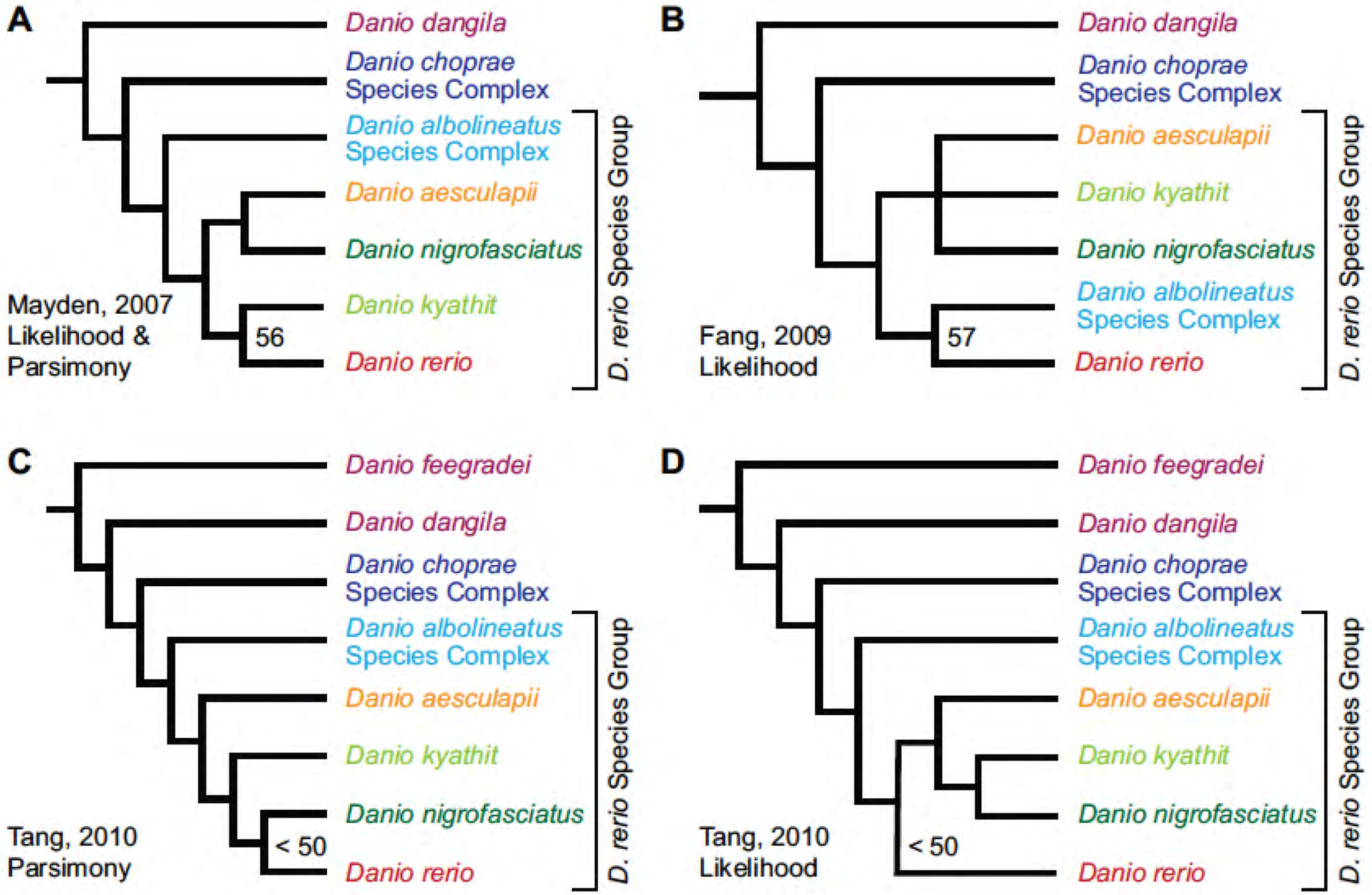
Marie Cariou, Laurent Duret & Sylvain Charlat

Phylogenetic relationships

Species pair *D. melanogaster*	Node depth (My)	Orthologous tags	Retrieved orthologous tags (%)	In clusters including paralogs (%)
D.sechellia	5.4	2978	99	5
D.simulans	5.4	2892	99	4
D.erecta	12.6	2390	97	3
D.yakuba	12.8	2314	97	8
D.ananassae	44.2	916	68	9
D.persimilis	54.9	648	65	9
D.pseudoobscura	54.9	648	66	9
D.wilistoni	62.2	242	49	6
D.grimshawi	62.9	290	60	8
D.virilis	62.9	286	59	5
D.mojavensis	62.9	298	59	8

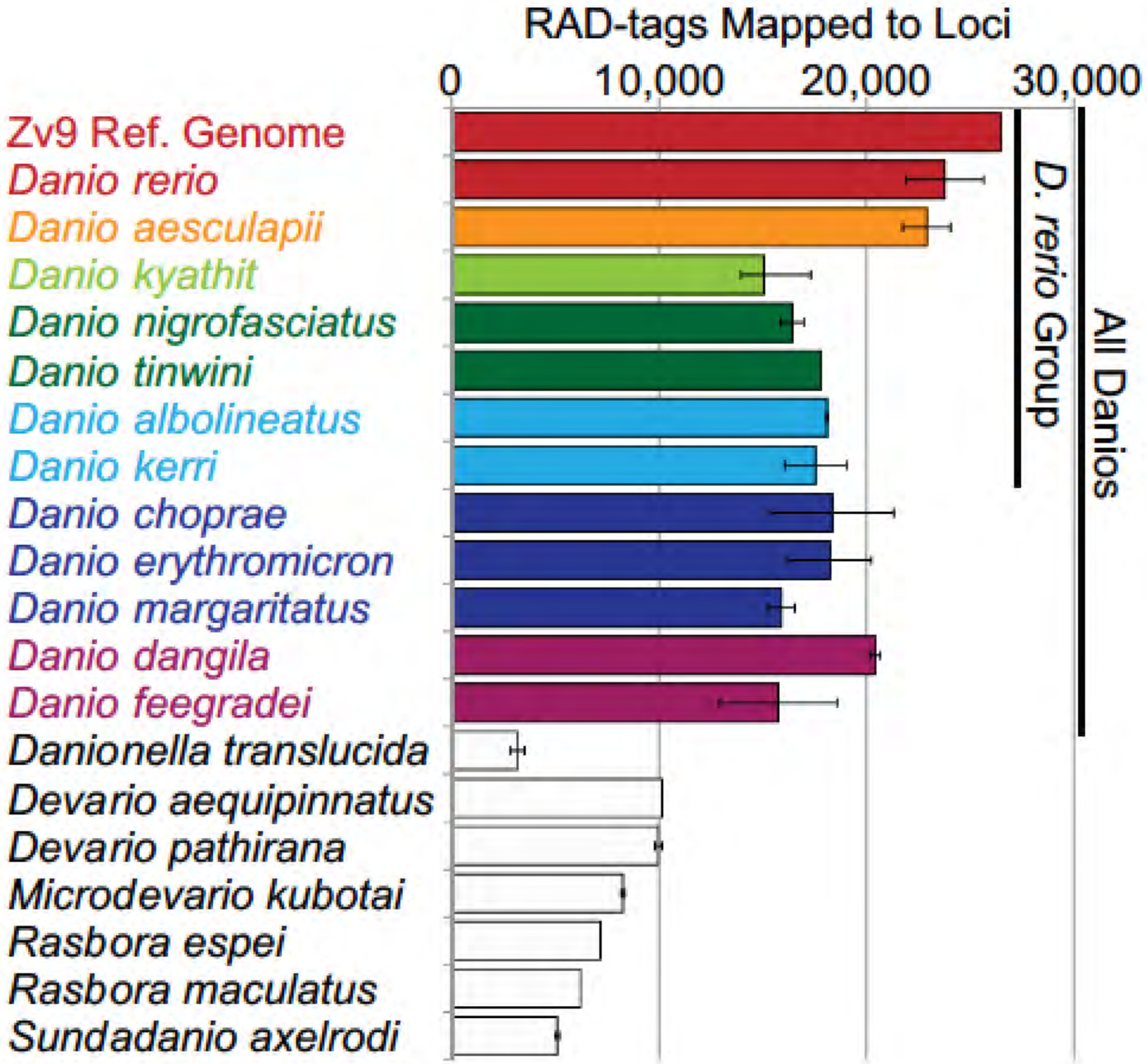
Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus"

Braedan M. McCluskey[1] and John H. Postlethwait*,[1]
[1]Institute of Neuroscience, University of Oregon



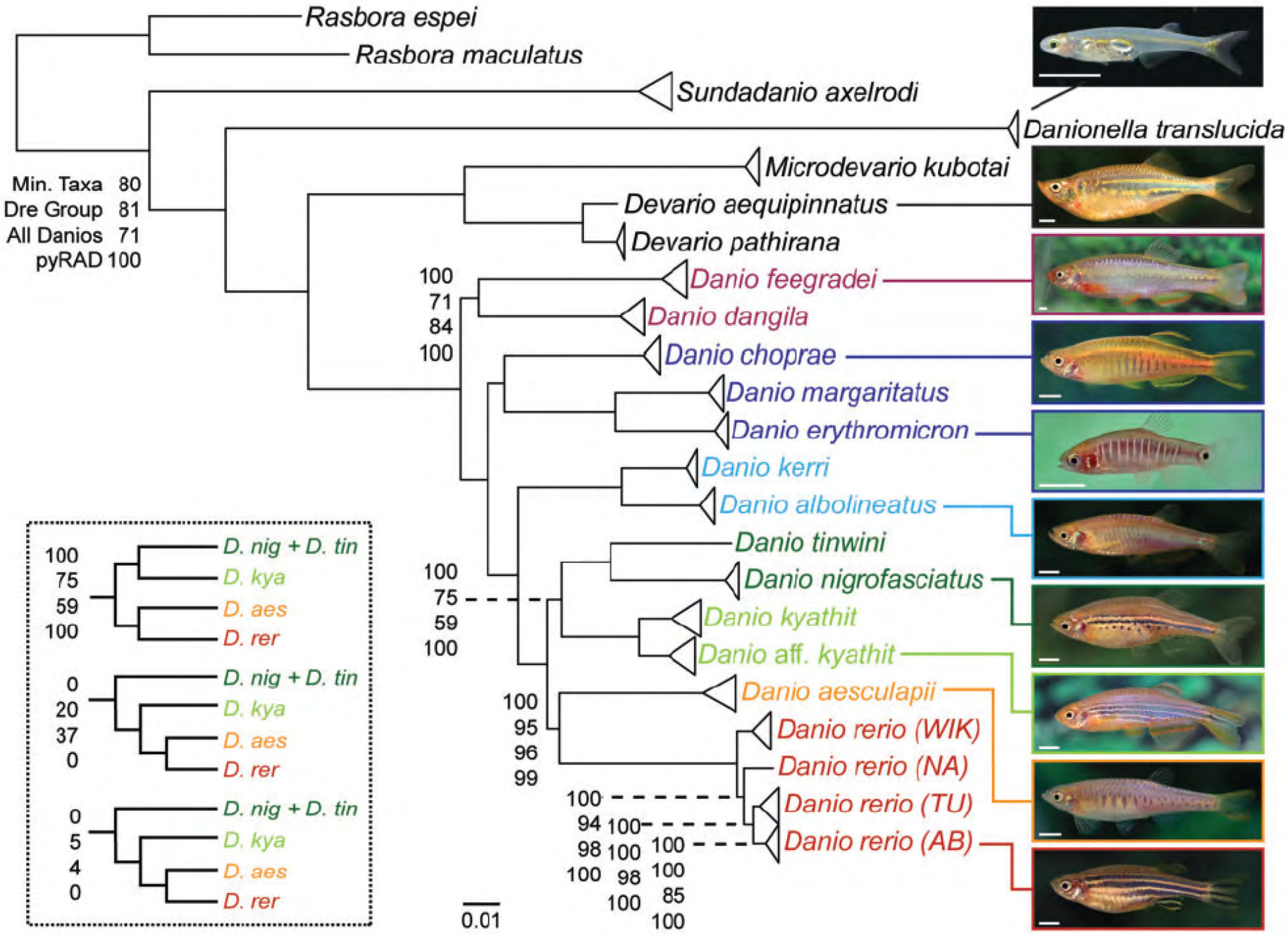
Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus"

Braedan M. McCluskey[1] and John H. Postlethwait*,[1]
[1]Institute of Neuroscience, University of Oregon

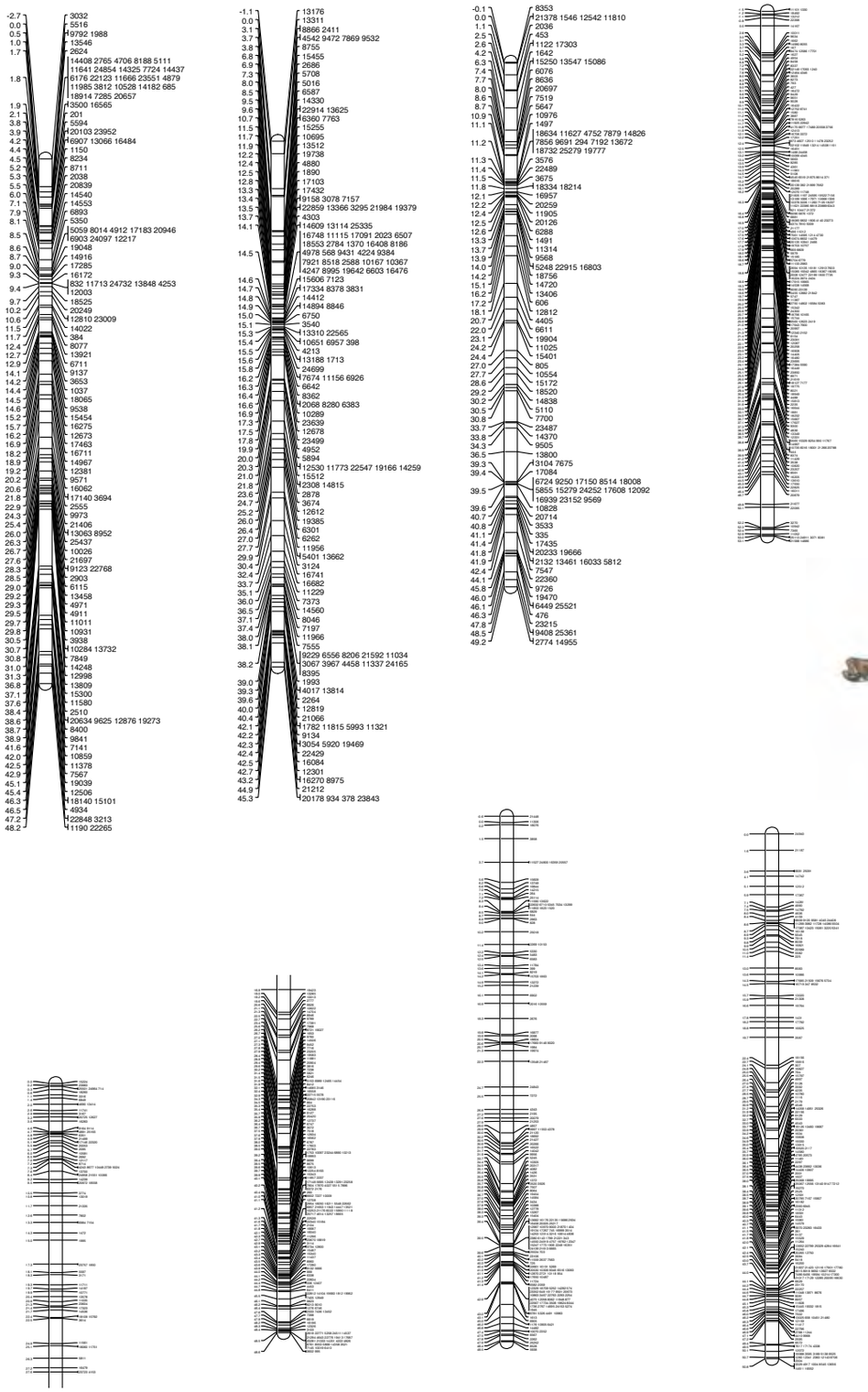


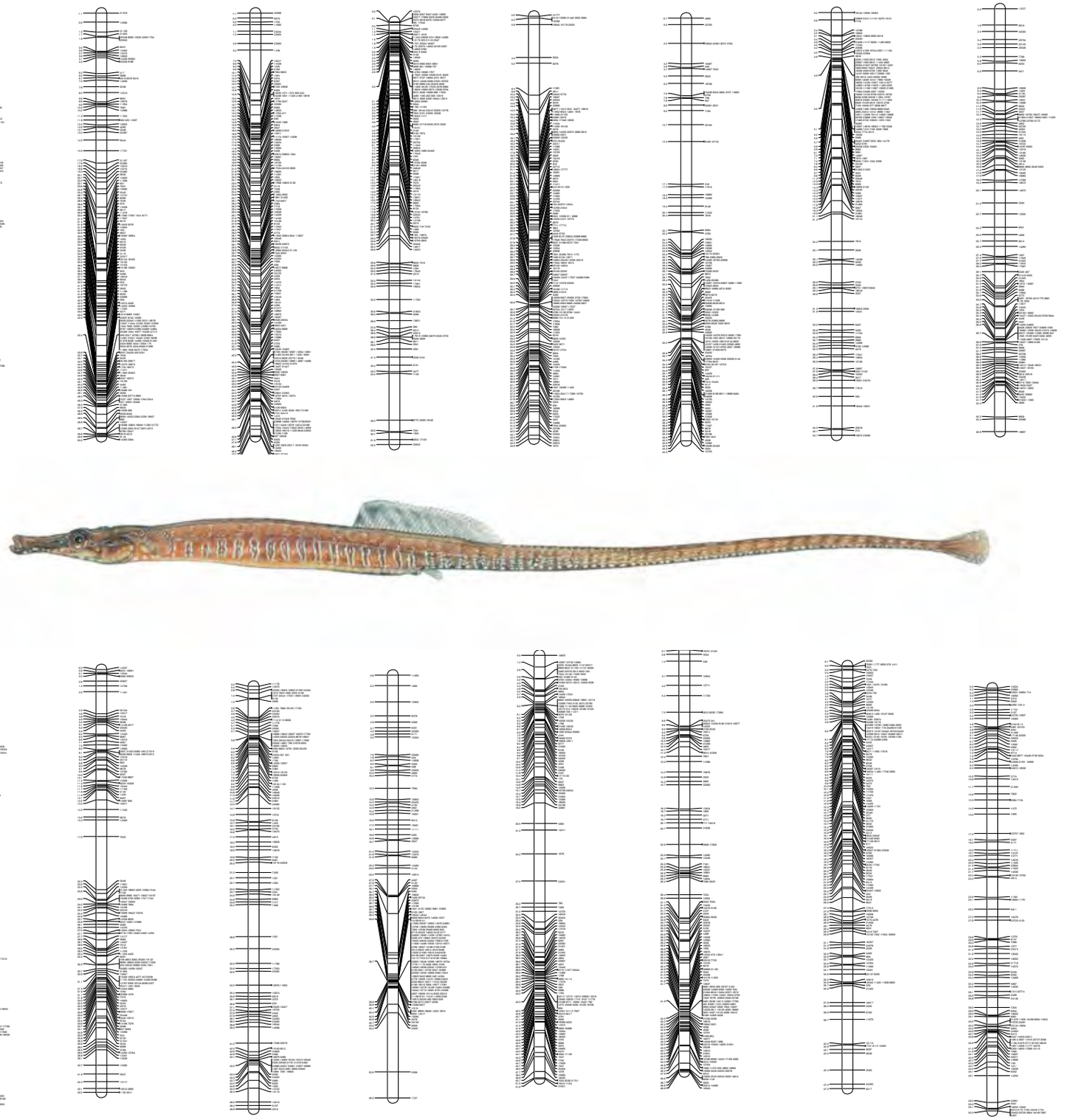
Phylogeny of Zebrafish, a "Model Species," within *Danio*, a "Model Genus"

Braedan M. McCluskey[1] and John H. Postlethwait*,[1]
[1]Institute of Neuroscience, University of Oregon

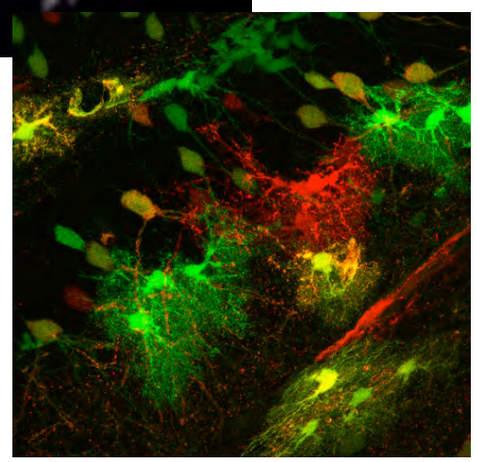


Genetically enable organisms with genetic maps

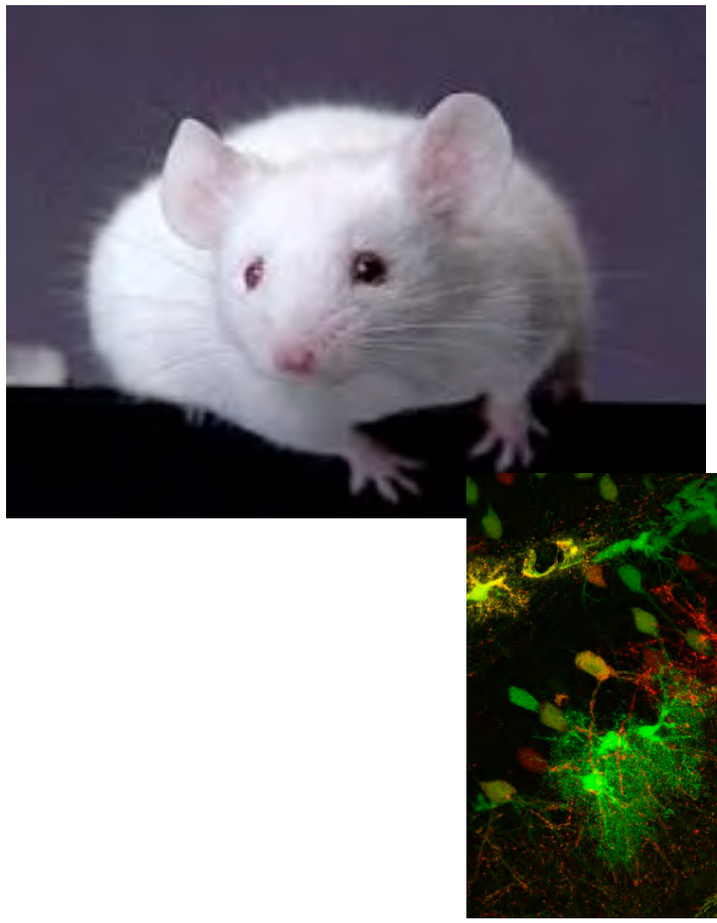


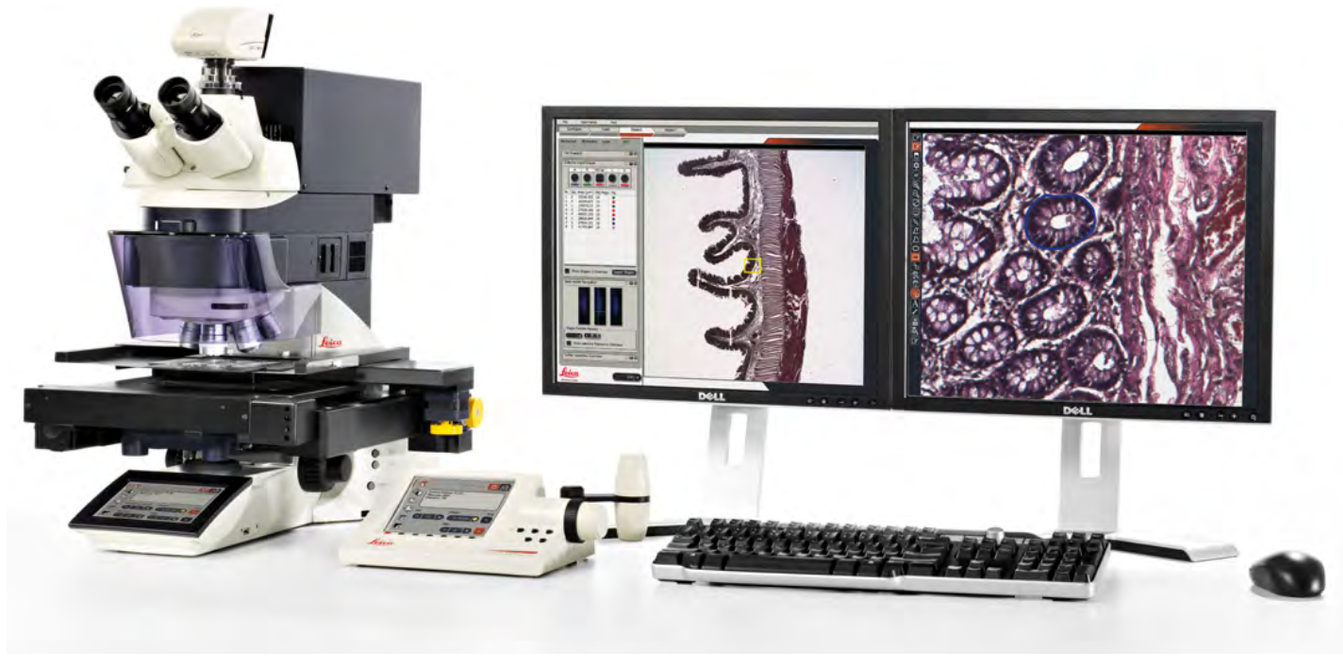


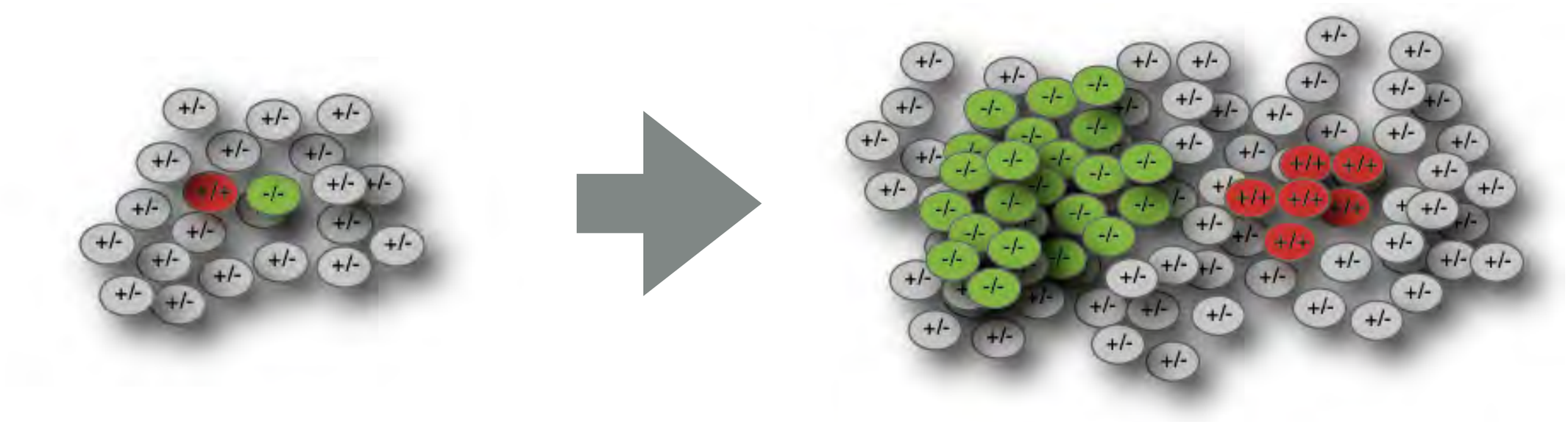
Studying cancer as an evolutionary process



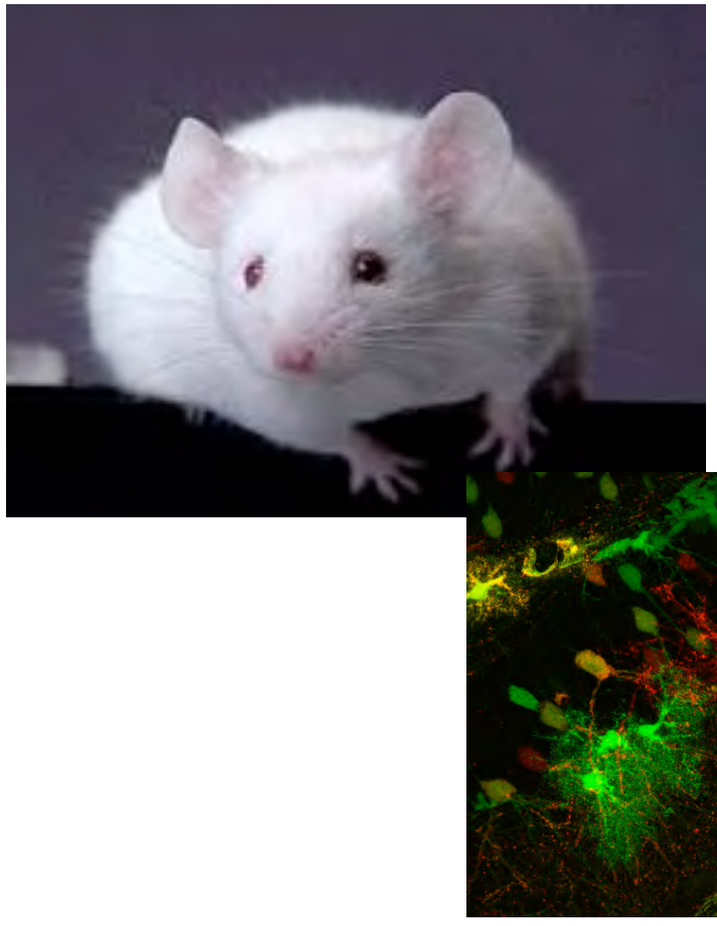
Studying cancer as an evolutionary process



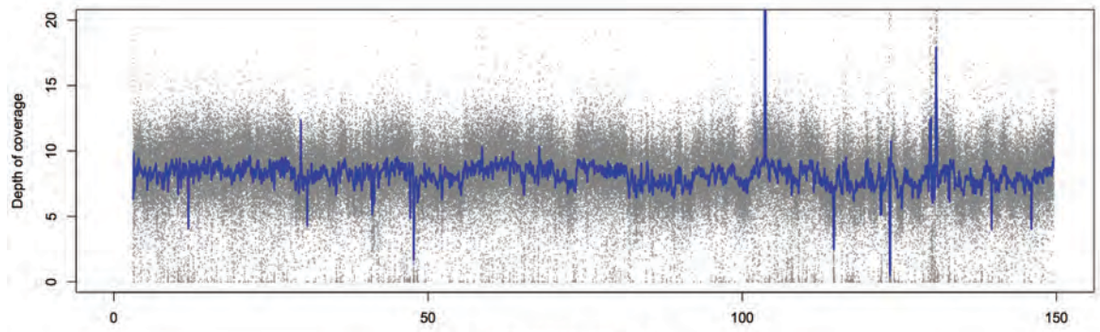




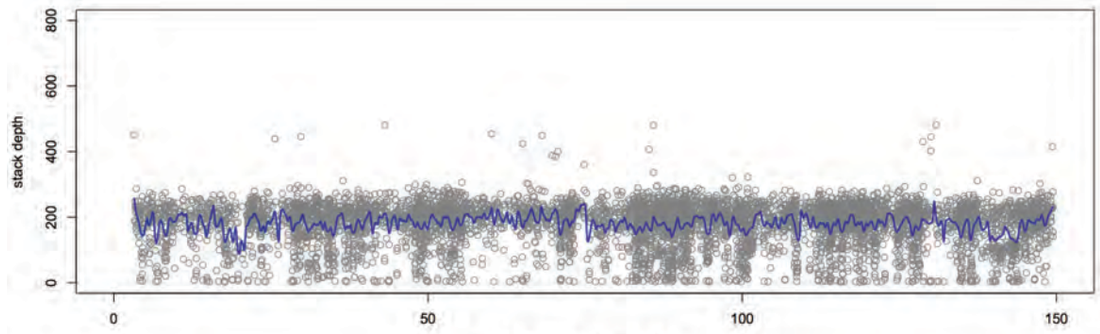
Studying cancer as an evolutionary process



Wild Type

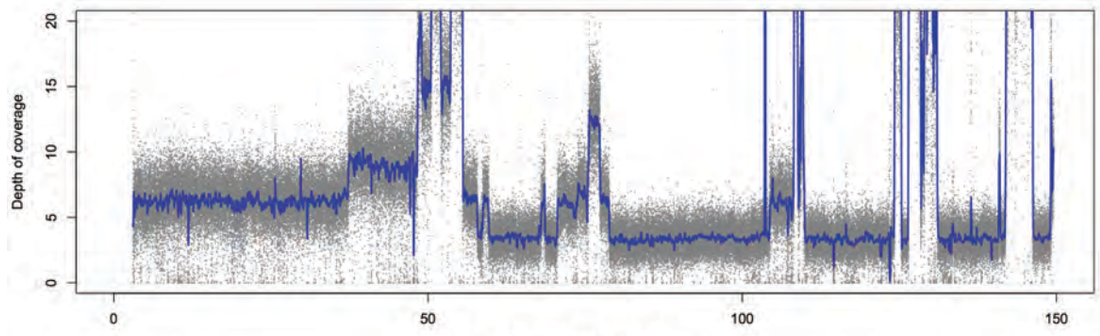


WGS

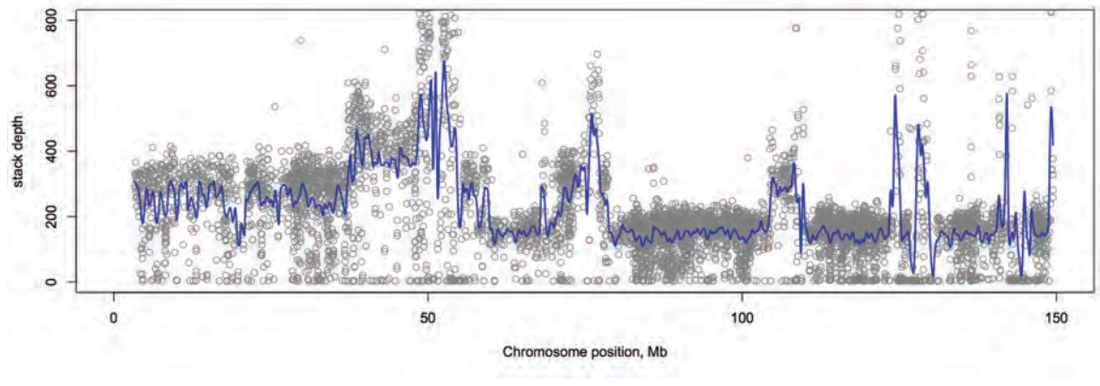


RAD

Mutant



WGS



RAD

Biological studies that could benefit from whole genome approaches

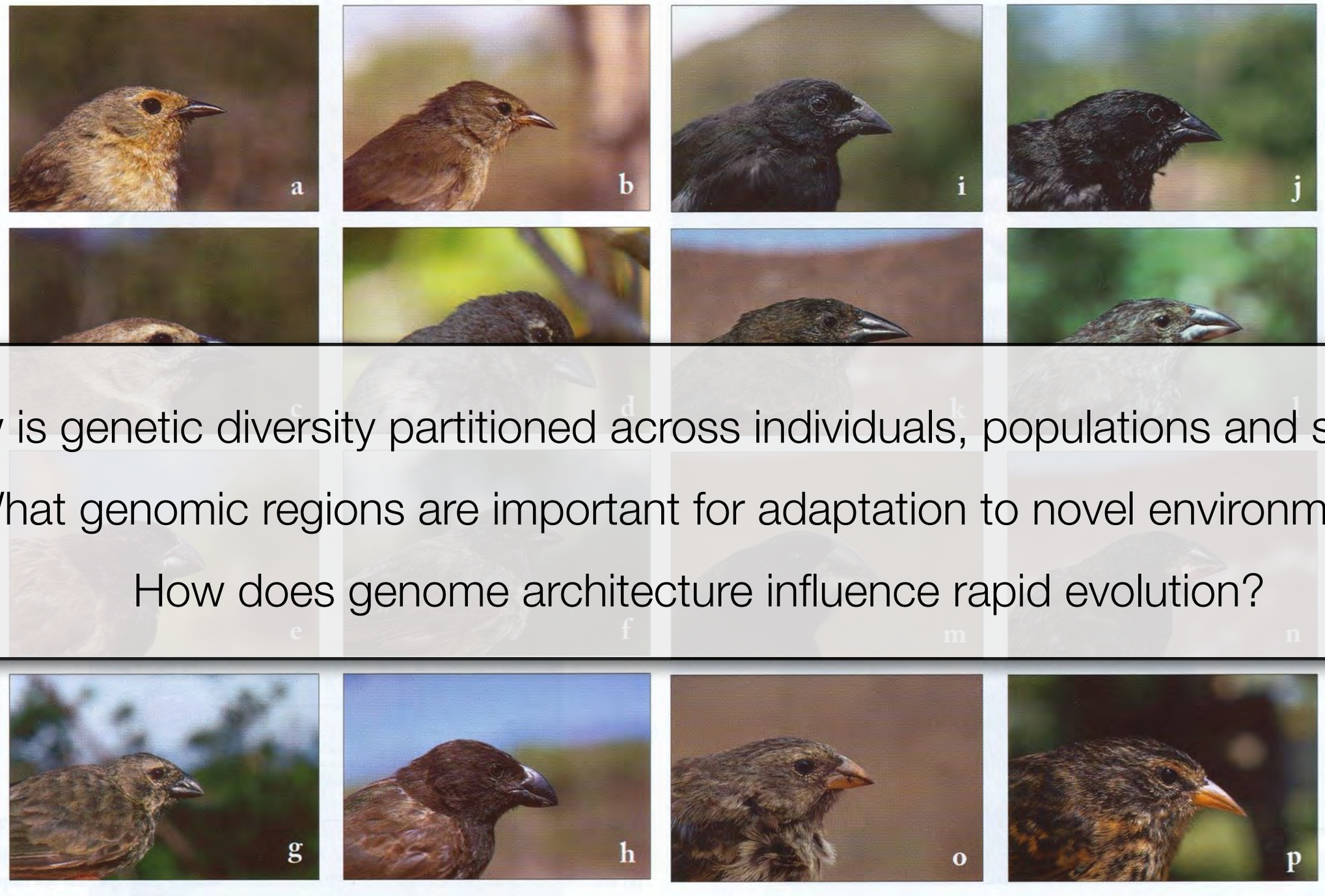
- Defining individuality, parentage and pedigrees
- Performing quantitative genetic studies in outbred populations
- Fine scale estimates of population structure
- Identifying the genetic basis of inbreeding depression, and making management decisions
- Genome Wide Association Studies (GWAS) studies of traits
- Building genetic maps to genetically enable non model organisms
- Estimating species and higher level phylogenetic relationships
- **Population genomics - identifying the signatures of natural selection**

Population Genomics

How do organisms adapt to novel environments?

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

How do organisms adapt to novel environments?



How is genetic diversity partitioned across individuals, populations and species?

What genomic regions are important for adaptation to novel environments?

How does genome architecture influence rapid evolution?

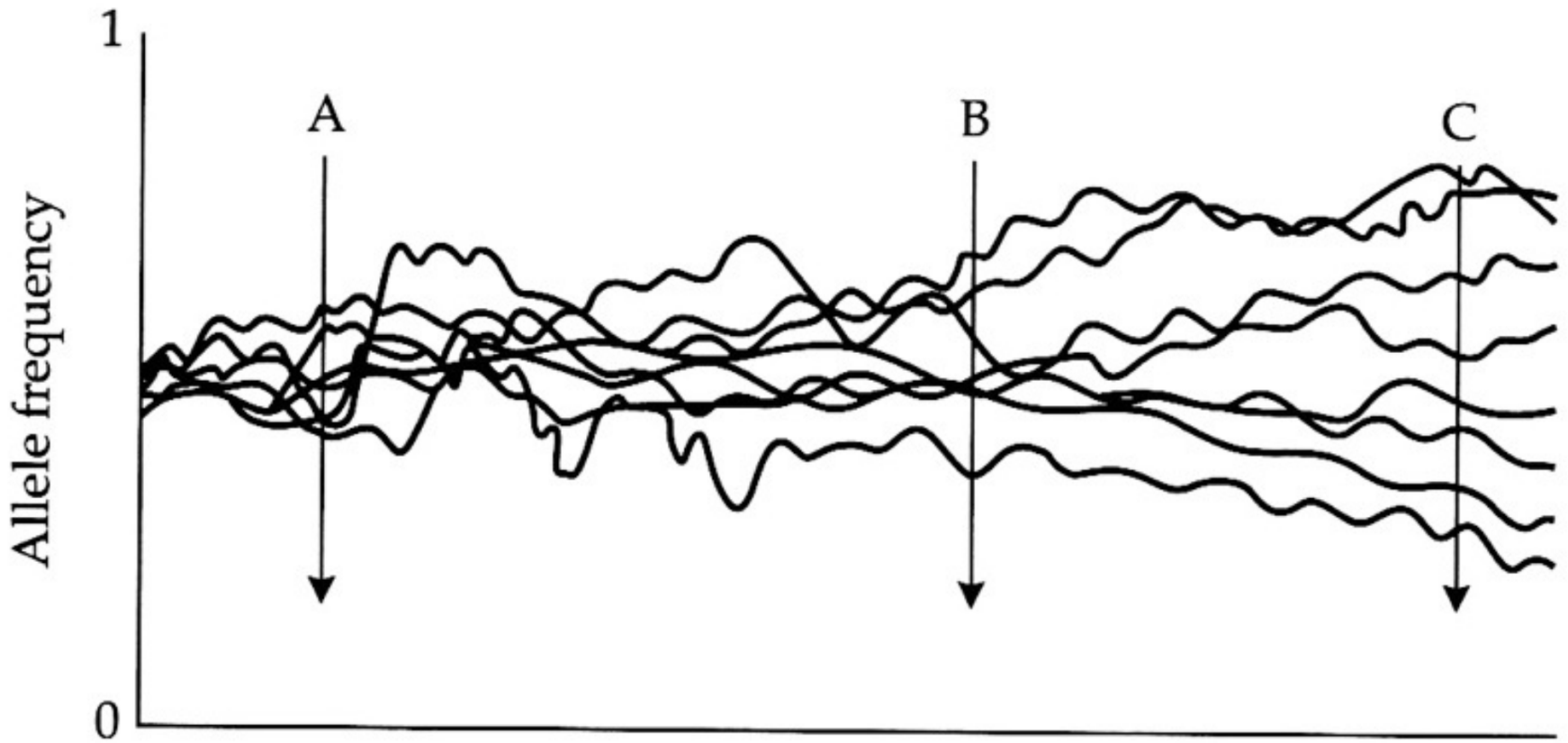
from Grant and Grant. 2007. How and why species multiply: The radiation of Darwin's finches. Princeton University Press

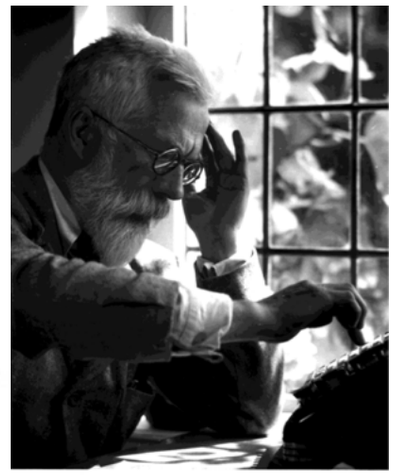
Four fundamental processes in evolution

Origin of genetic variation;
mutation
migration

Sorting of variation;
genetic drift
natural selection

Genetic drift is a null model





R. A. Fisher

Sewall Wright

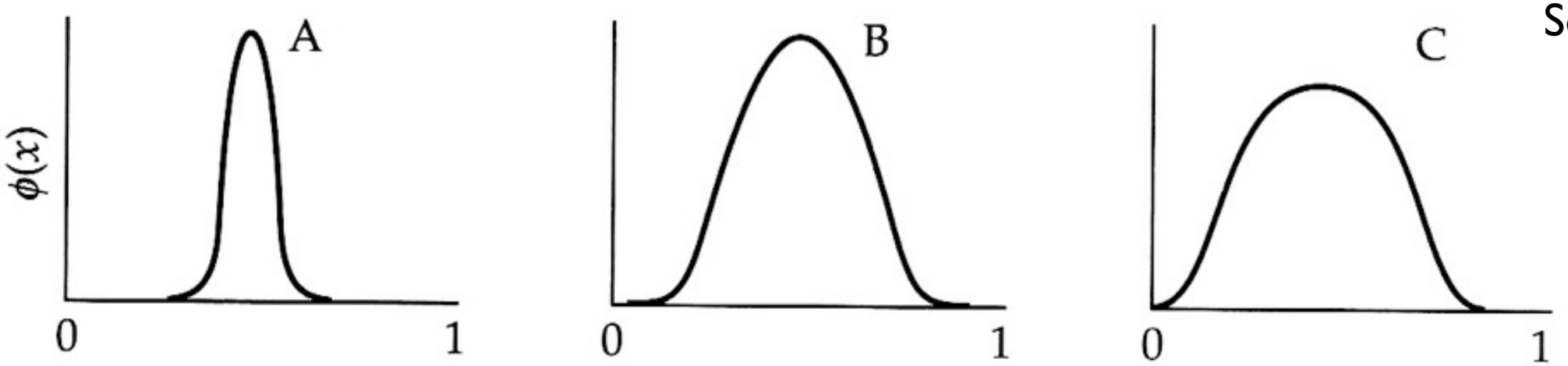
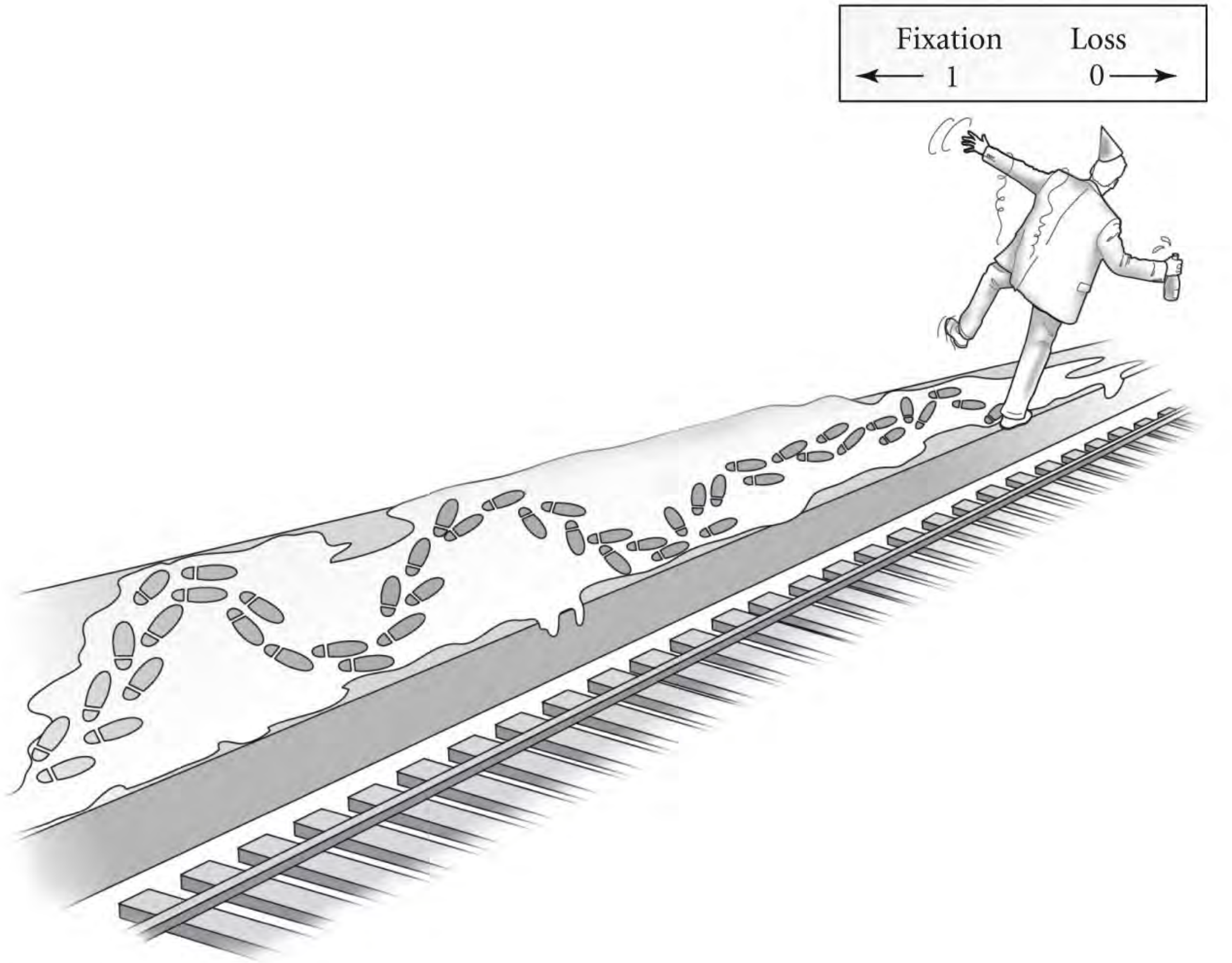


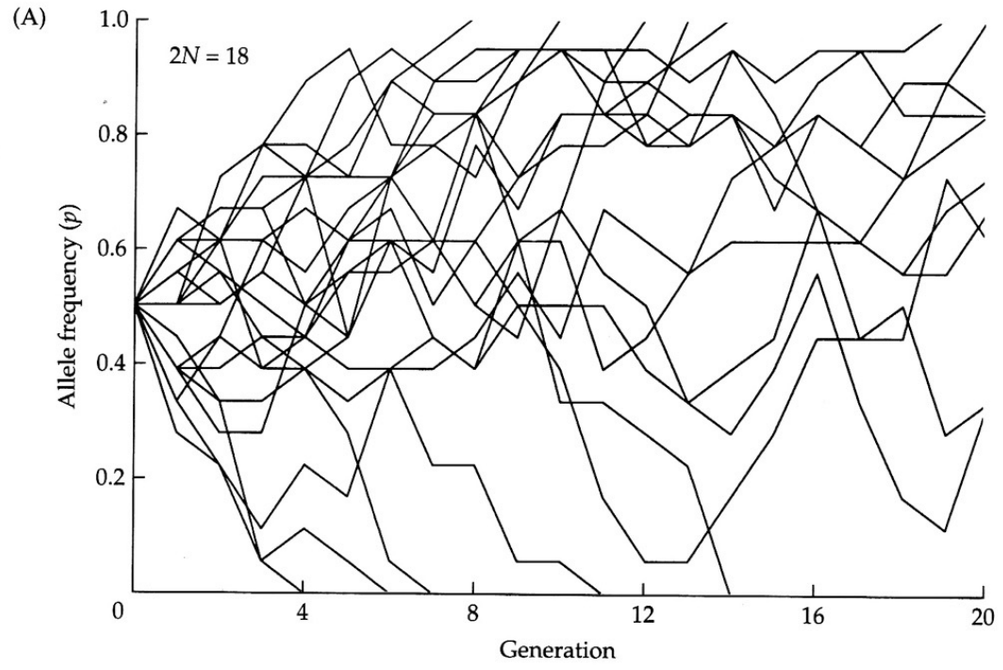
Figure 10.2 A “random walk” (or “drunkard’s walk”)

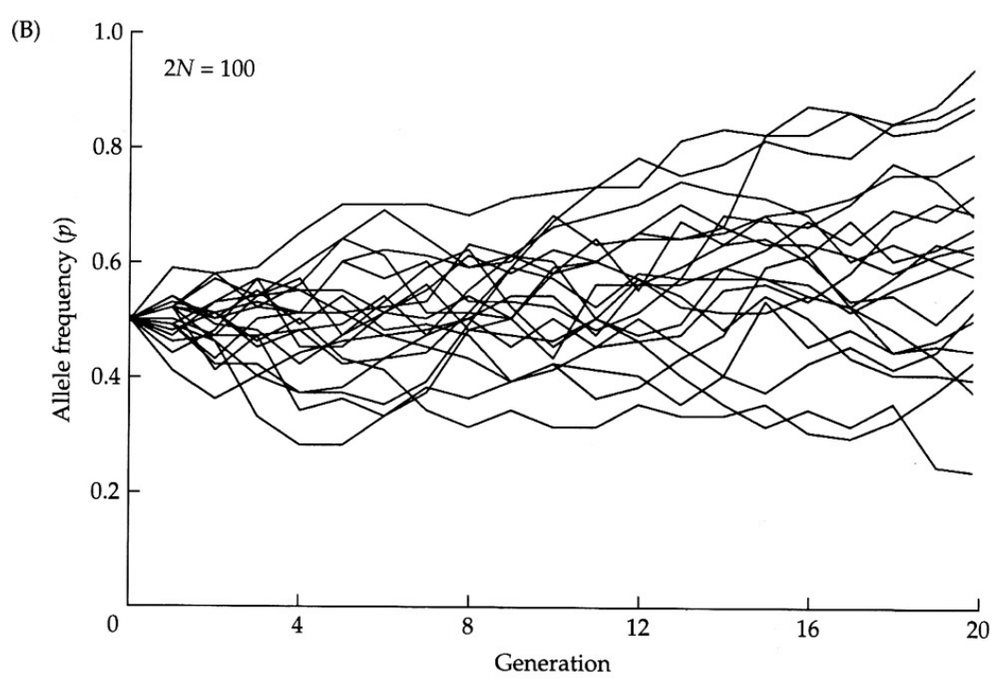


***EVOLUTION 2e*, Figure 10.2**

© 2009 Sinauer Associates, Inc.

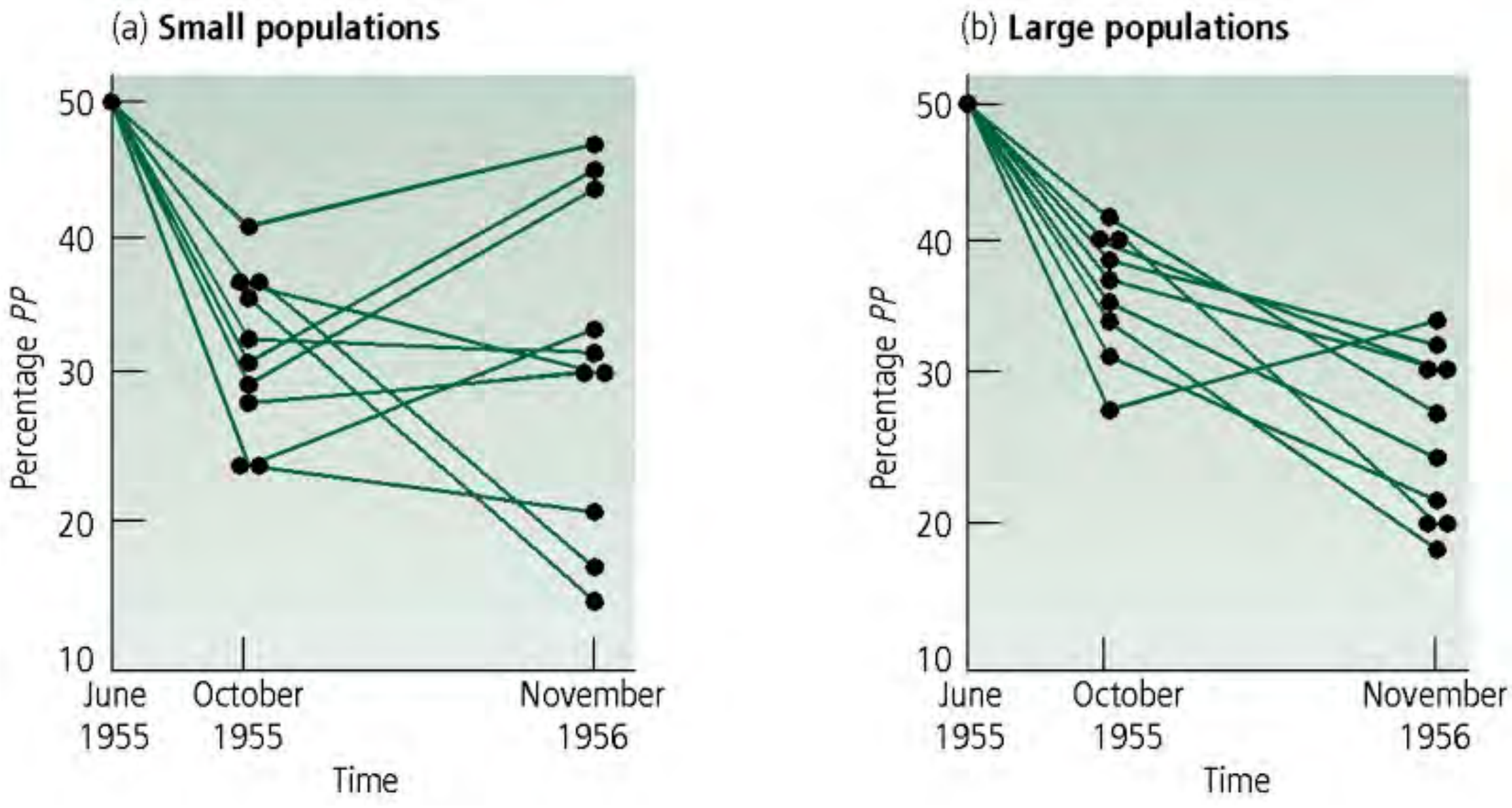
Size of population (Ne) affects rate of spread (diffusion)





*** Affects the entire genome equally (on average) ***

Natural selection biases the allele frequency change, but drift is still occurring



*** The effects of selection can be genomically localized ***

Population genomics

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

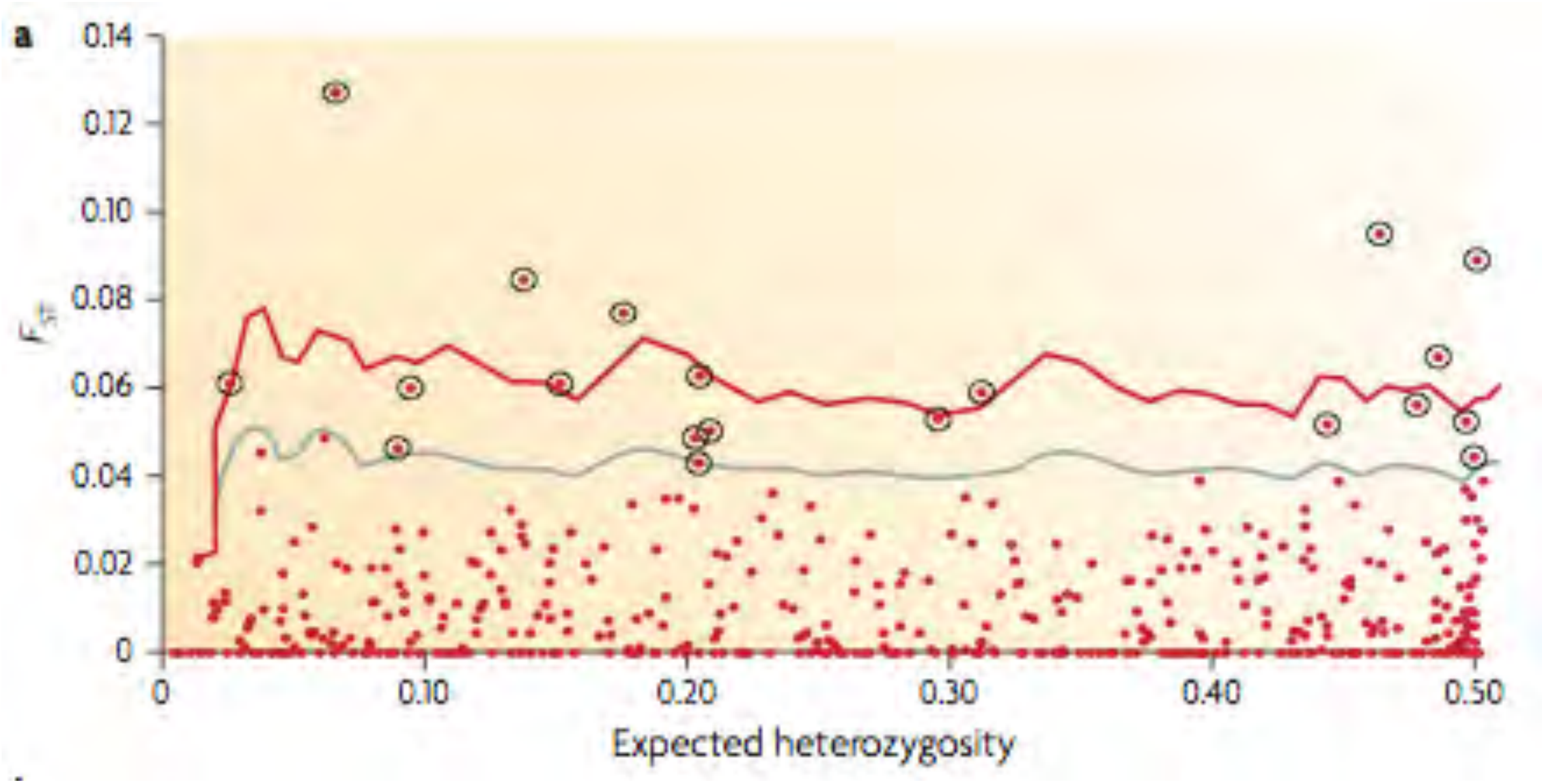
Genome-wide background provides more precise estimates:

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

Outliers from background indicate:

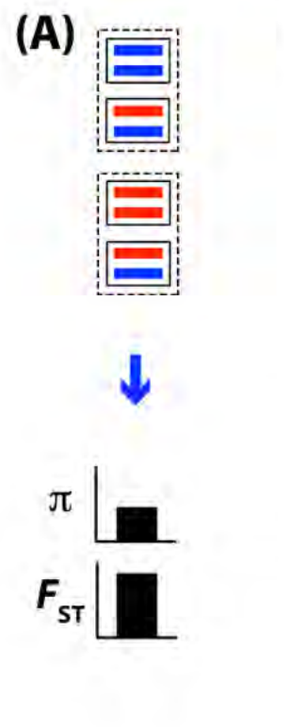
- Selective sweeps
- Local adaptation

Population genomics of unordered markers



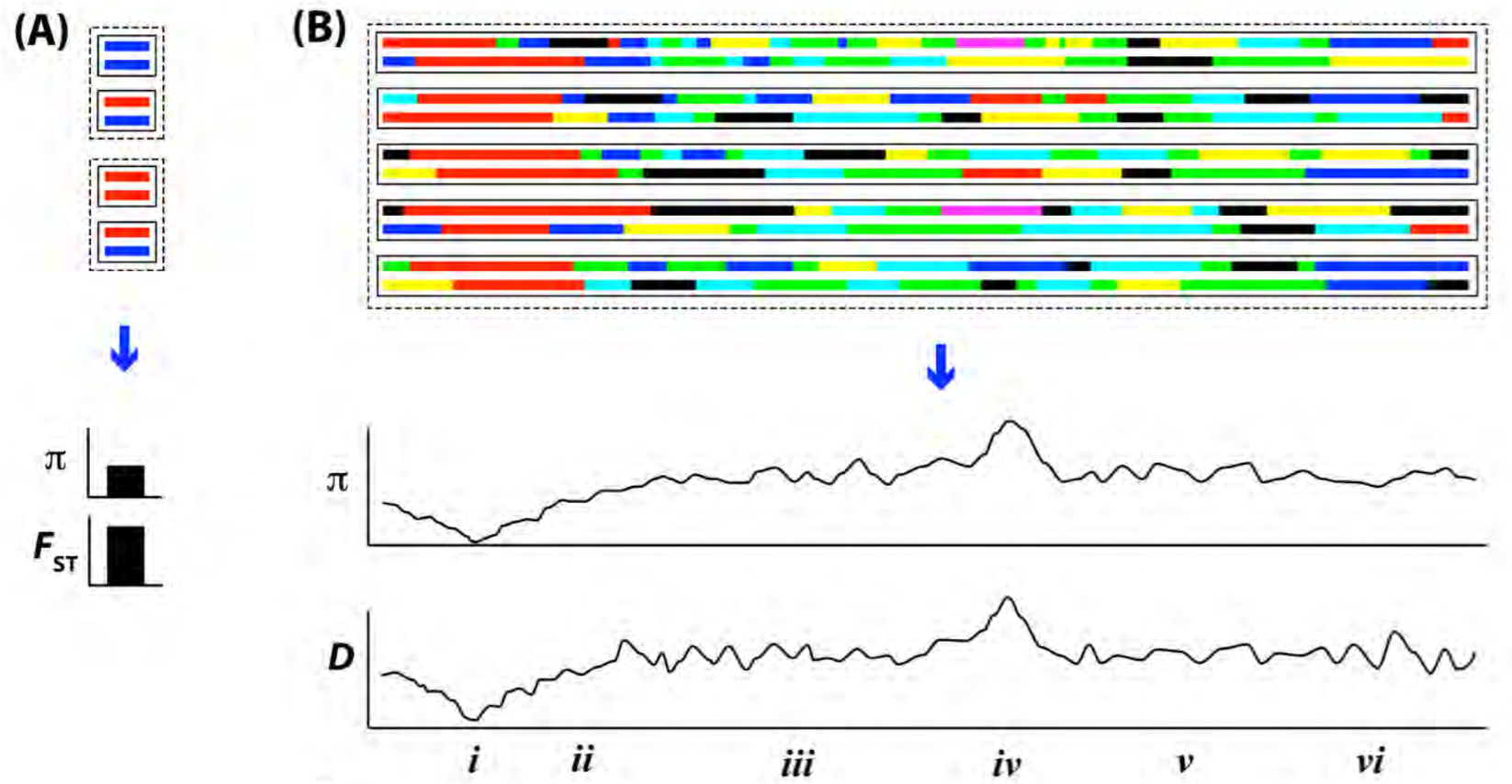
Namroud et al (2008) *Mol Ecol* 17:3599

Population genomics of ordered markers



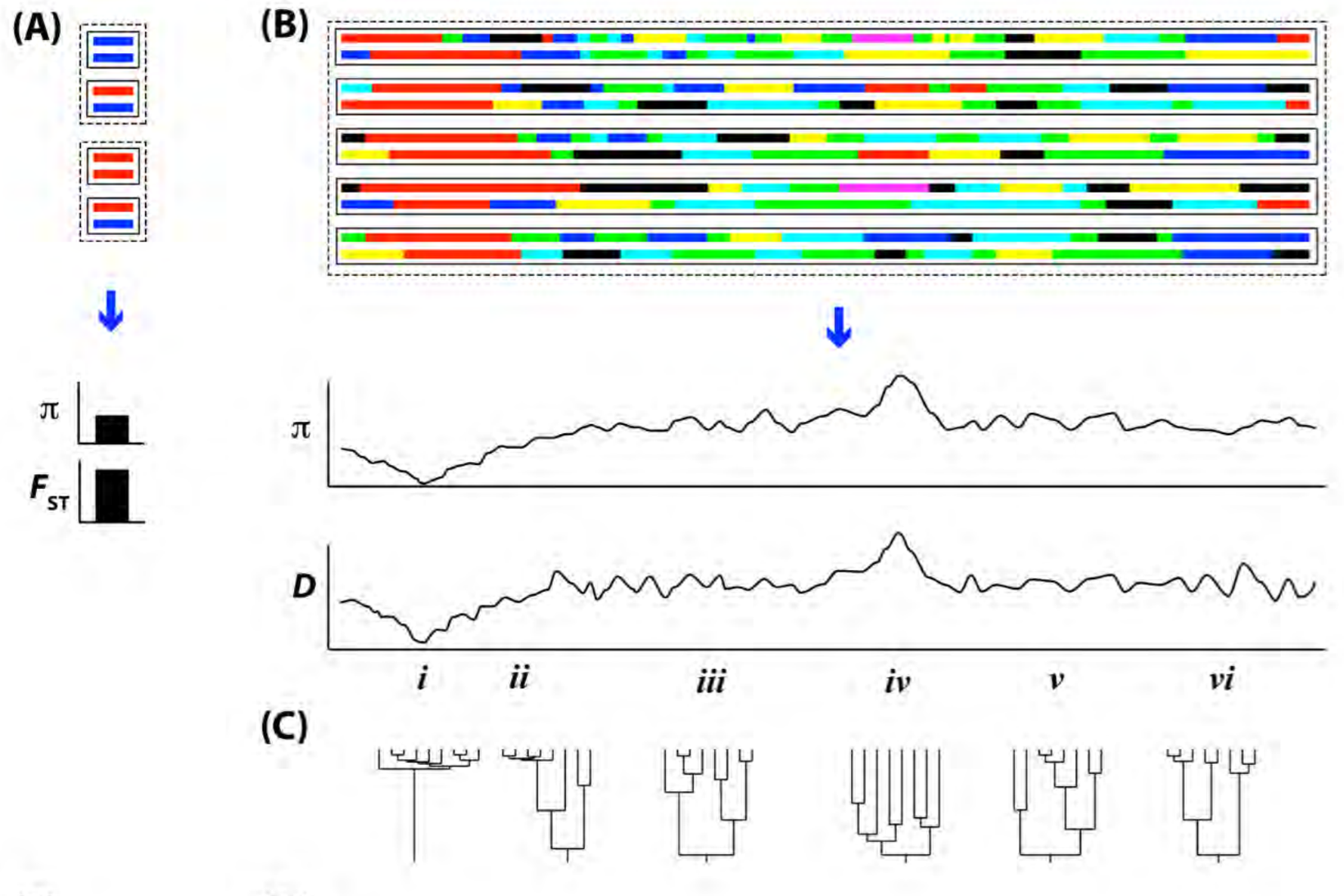
Hohenlohe et al. 2010. Int J Plant Sciences

Population genomics of ordered markers



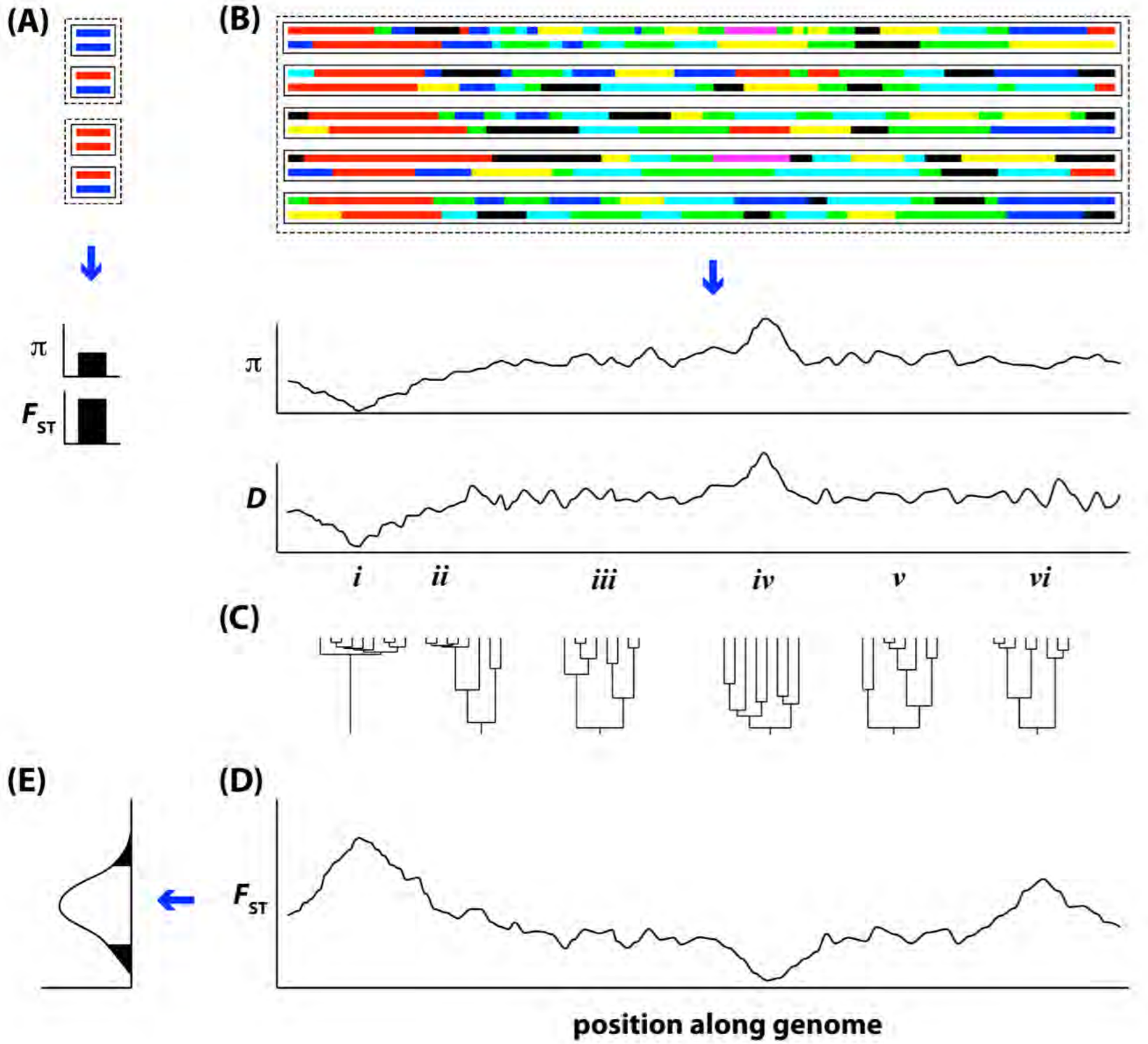
Hohenlohe et al. 2010. Int J Plant Sciences

Population genomics of ordered markers



Hohenlohe et al. 2010. Int J Plant Sciences

Population genomics of ordered markers

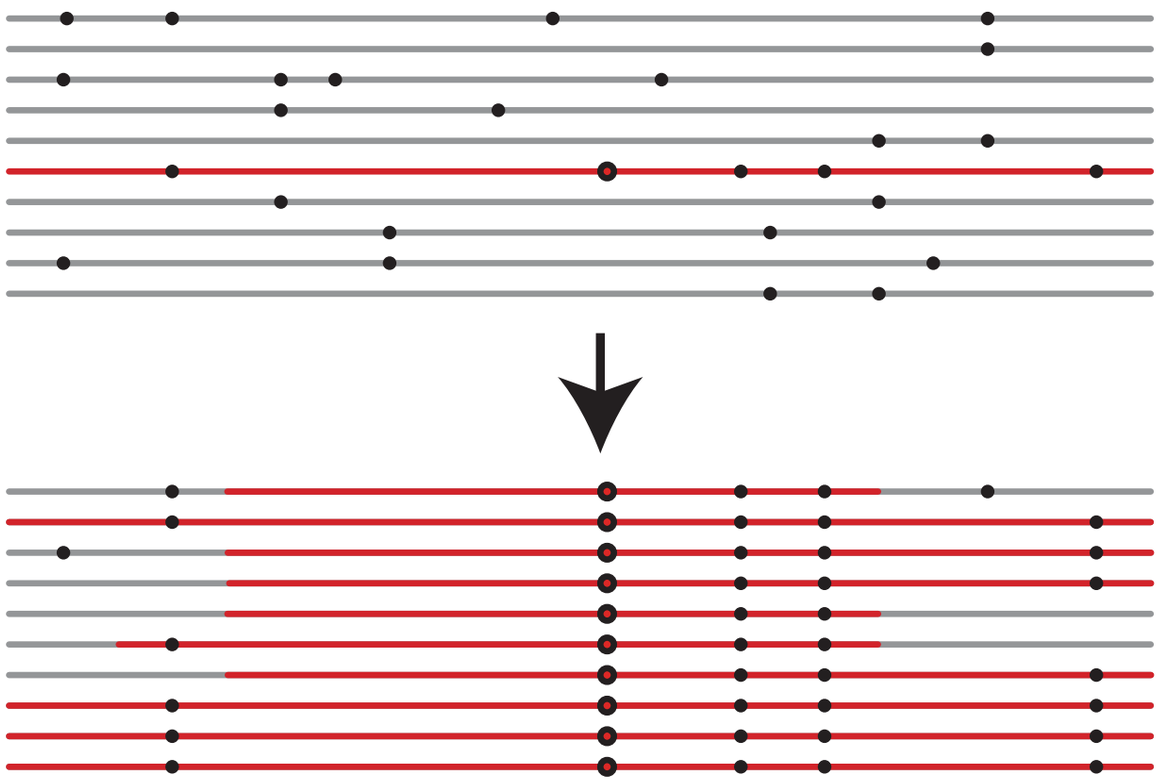


Hohenlohe et al. 2010. Int J Plant Sciences

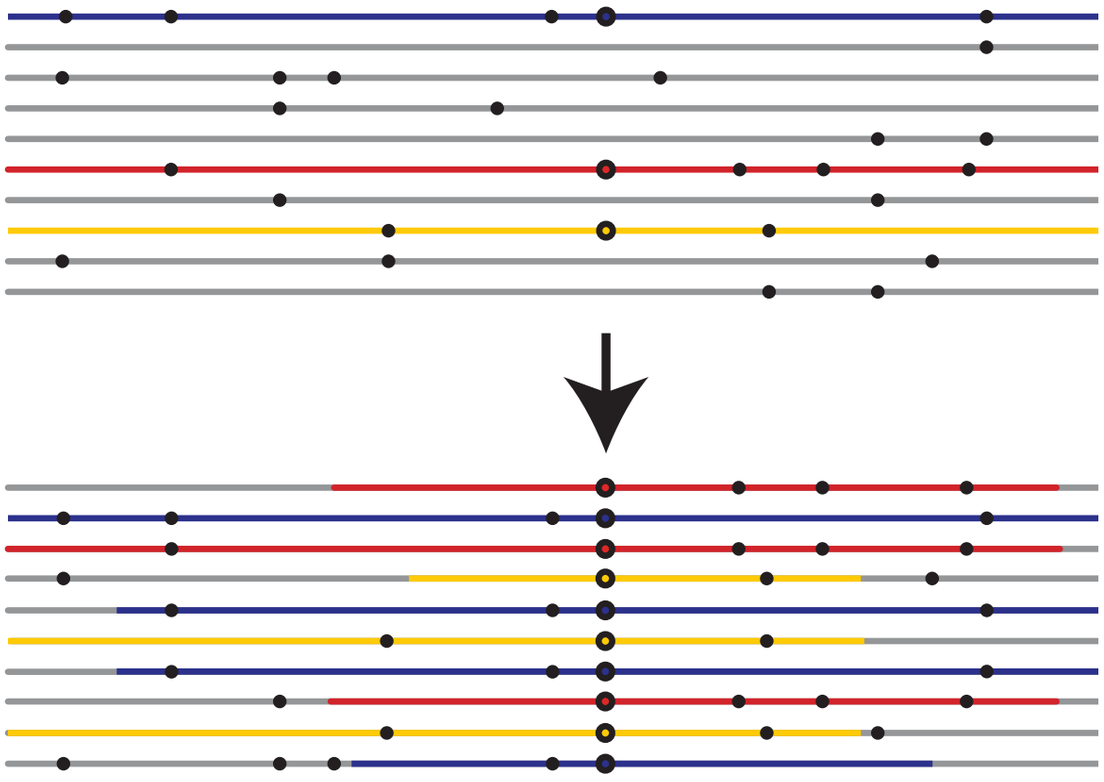
Sweeps of directional selection across genomes

Past

Hard Sweep

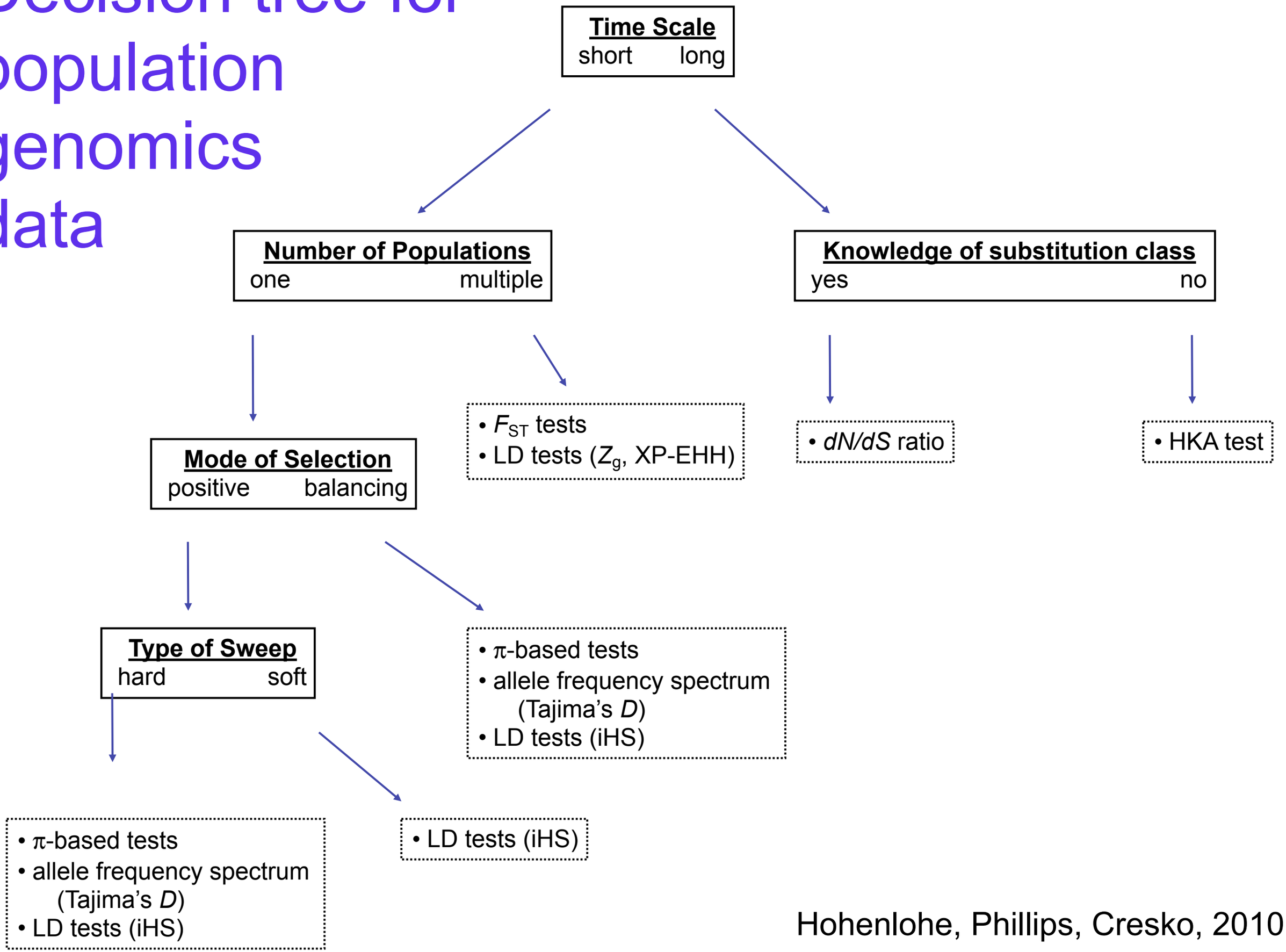


Soft Sweep



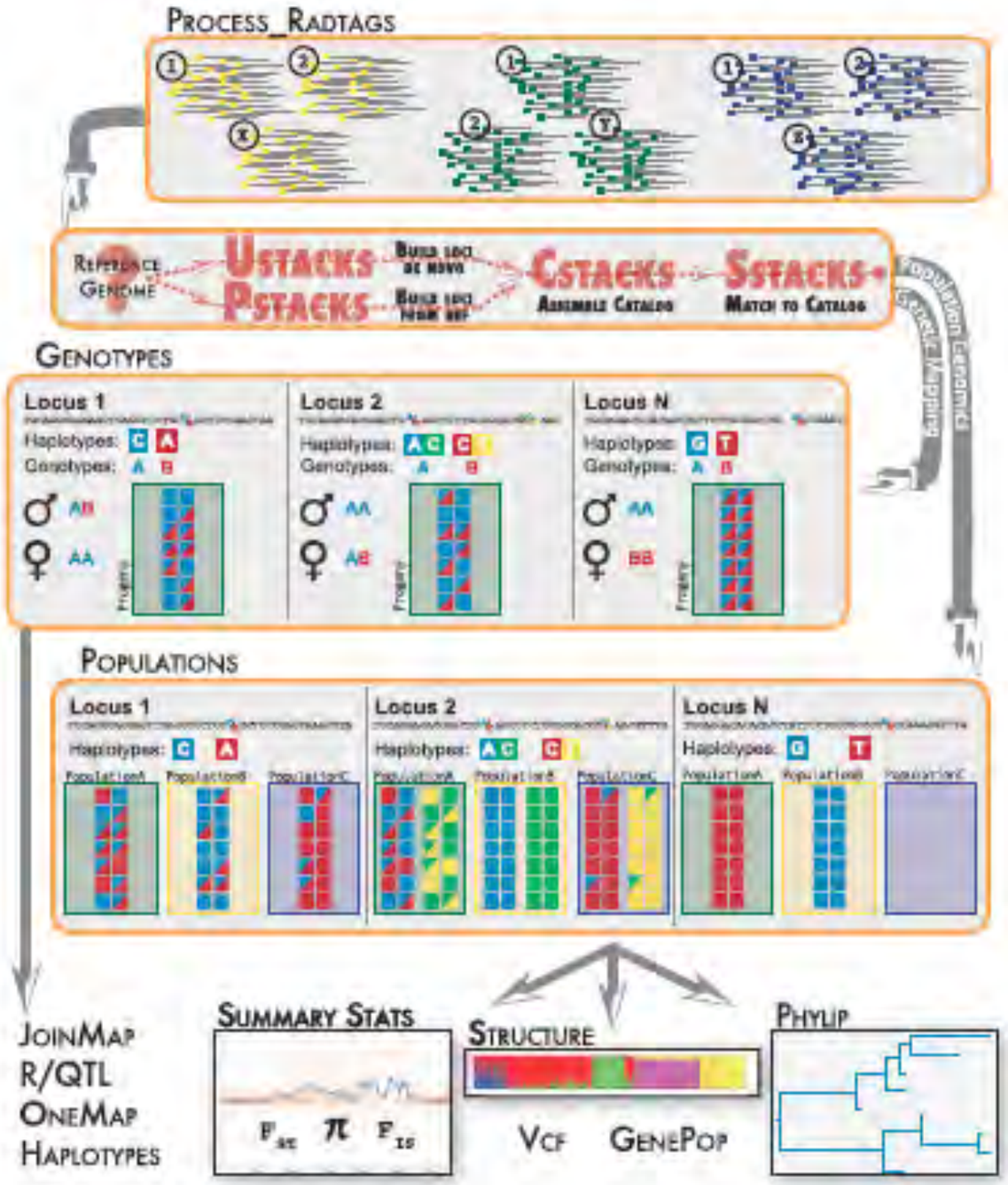
Present

Decision tree for population genomics data



Hohenlohe, Phillips, Cresko, 2010

Stacks analysis pipeline for RAD-seq



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

Julian M. Catchen,* Angel Amores,† Paul Hohenlohe,* William Cresko,* and John H. Postlethwait†,1

*Center for Ecology and Evolutionary Biology and †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403

Stacks: an analysis tool set for population genomics

JULIAN CATCHEN,* PAUL A. HOHENLOHE,*† SUSAN BASSHAM,* ANGEL AMORES‡ and WILLIAM A. CRESKO*

*Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403-5289, USA, †Biological Sciences, University of Idaho, Moscow, ID 83844-3051, USA, ‡Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254, USA

Stacks

7.4.2 batch_X.sumstats_summary.tsv: Summary of summary statistics for each population

Column	Name	Description
1	Pop ID	Population ID as defined in the Population Map file.
2	Private	Number of private alleles in this population.
3	Number of Individuals	Mean number of individuals per locus in this population.
4	Variance	
5	Standard Error	
6	P	Mean frequency of the most frequent allele at each locus in this population.
7	Variance	
8	Standard Error	
9	Observed Heterozygosity	Mean obsered heterozygosity in this population.
10	Variance	
11	Standard Error	
12	Observed Homozygosity	Mean observed homozygosity in this population.
13	Variance	
14	Standard Error	
15	Expected Heterozygosity	Mean expected heterozygosity in this population.
16	Variance	
17	Standard Error	
18	Expected Homozygosity	Mean expected homozygosity in this population.
19	Variance	
20	Standard Error	
21	Π	Mean value of π in this population.
22	Π Variance	
23	Π Standard Error	
24	F_{IS}	Mean measure of F_{IS} in this population.
25	F_{IS} Variance	
26	F_{IS} Standard Error	

Stacks

7.4.4 batch_X.hapstats.tsv: Haplotype-based summary statistics for each locus in each population

Column	Name	Description
1	Batch ID	The batch identifier for this data set.
2	Locus ID	Catalog locus identifier.
3	Chromosome	If aligned to a reference genome.
4	Basepair	If aligned to a reference genome.
5	Population ID	The ID supplied to the populations program, as written in the population map file.
6	N	Number of alleles/haplotypes present at this locus.
7	Haplotype count	
8	Gene Diversity	
9	Smoothed Gene Diversity	
10	Smoothed Gene Diversity P-value	
11	Haplotype Diversity	
12	Smoothed Haplotype Diversity	
13	Smoothed Haplotype Diversity P-value	
14	Haplotypes	A semicolon-separated list of haplotypes/haplotype counts in the population.

Stacks

7.4.3 batch_X.fst_Y-Z.tsv: F_{ST} calculations for each pair of populations

Column	Name	Description
1	Batch ID	The batch identifier for this data set.
2	Locus ID	Catalog locus identifier.
3	Population ID 1	The ID supplied to the populations program, as written in the population map file.
4	Population ID 2	The ID supplied to the populations program, as written in the population map file.
5	Chromosome	If aligned to a reference genome.
6	Basepair	If aligned to a reference genome.
7	Column	The nucleotide site within the catalog locus, reported using a zero-based offset (first nucleotide is enumerated as 0).
8	Overall π	An estimate of nucleotide diversity across the two populations.
9	F_{ST}	A measure of population differentiation.
10	FET p-value	P-value describing if the F_{ST} measure is statistically significant according to Fisher's Exact Test.
11	Odds Ratio	Fisher's Exact Test odds ratio.
12	CI High	Fisher's Exact Test confidence interval.
13	CI Low	Fisher's Exact Test confidence interval.
14	LOD Score	Logarithm of odds score.
15	Corrected F_{ST}	F_{ST} with either the FET p-value, or a window-size or genome size Bonferroni correction.
16	Smoothed F_{ST}	A weighted average of F_{ST} depending on the surrounding 3σ of sequence in both directions.
17	AMOVA F_{ST}	Analysis of Molecular Variance alternative F_{ST} calculation. Derived from Weir, Genetic Data Analysis II, chapter 5, "F Statistics," pp166-167.
18	Corrected AMOVA F_{ST}	AMOVA F_{ST} with either the FET p-value, or a window-size or genome size Bonferroni correction.
19	Smoothed AMOVA F_{ST}	A weighted average of AMOVA F_{ST} depending on the surrounding 3σ of sequence in both directions.
20	Smoothed AMOVA F_{ST} P-value	If bootstrap resampling is enabled, a p-value ranking the significance of F_{ST} within this pair of populations.
21	Window SNP Count	Number of SNPs found in the sliding window centered on this nucleotide position.

Notes: The preferred version of F_{ST} is the **AMOVA F_{ST}** in column 17, or the corrected version in column 18 if you have specified a correction to the **`populations`** program (option `-f`)

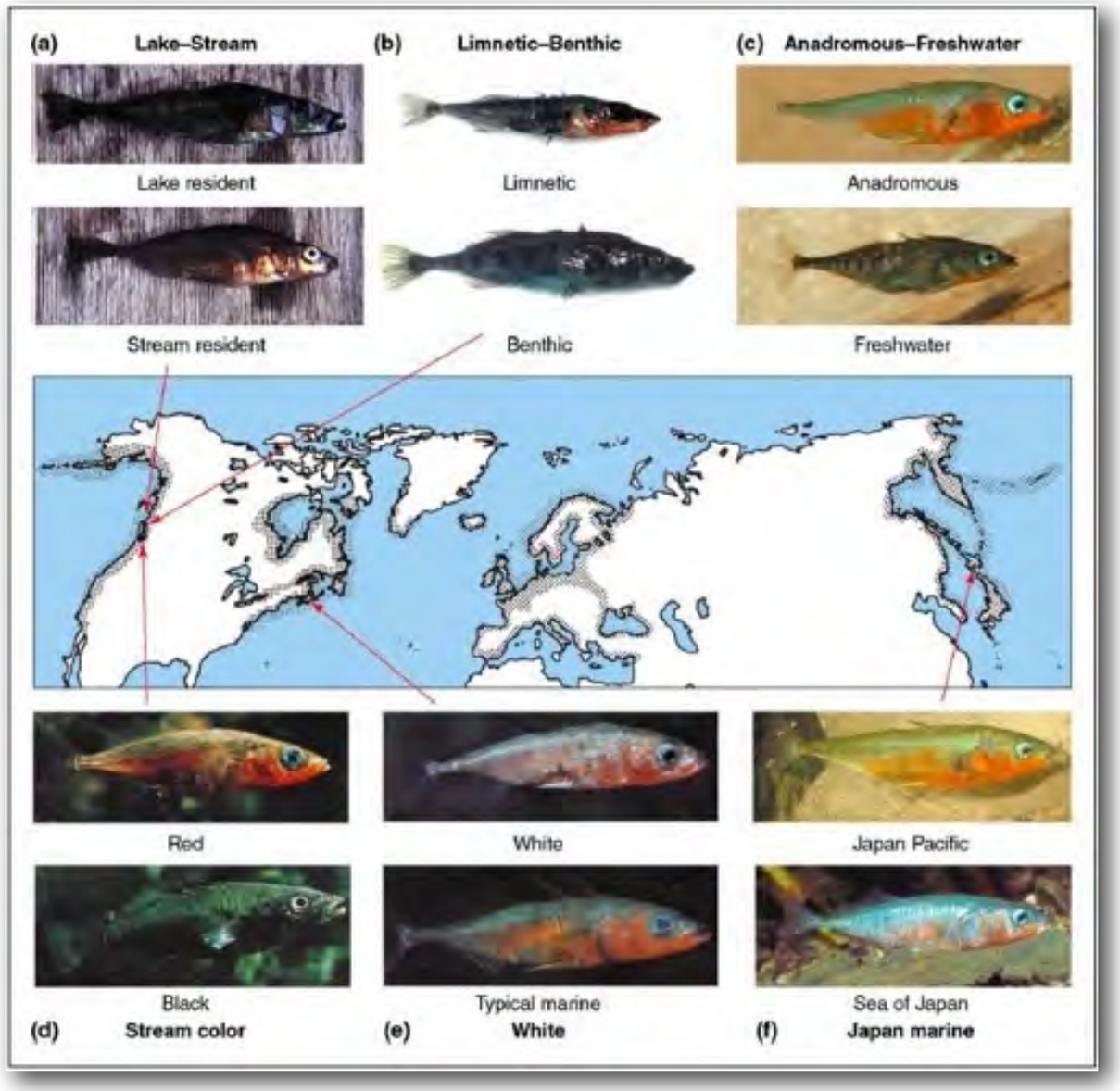
7.4.6 batch_X.phistats_Y-Z.tsv: Haplotype-based F_{ST} calculations for each pair of populations

Column	Name	Description
1	Batch ID	The batch identifier for this data set.
2	Locus ID	Catalog locus identifier.
3	Population ID 1	The ID supplied to the populations program, as written in the population map file.
4	Population ID 2	The ID supplied to the populations program, as written in the population map file.
5	Chromosome	If aligned to a reference genome.
6	Basepair	If aligned to a reference genome.
7	Φ_{ST}	
8	Smoothed Φ_{ST}	
9	Smoothed Φ_{ST} P-value	
10	F_{ST}'	
11	Smoothed F_{ST}'	
12	Smoothed F_{ST}' P-value	
13	D_{EST}	
14	Smoothed D_{EST}	
15	Smoothed D_{EST} P-value	

Case studies of using RAD for an organism with a reference genome: population genomics of threespine stickleback

Threespine stickleback, *Gasterosteus aculeatus*

Threespine stickleback, *Gasterosteus aculeatus*

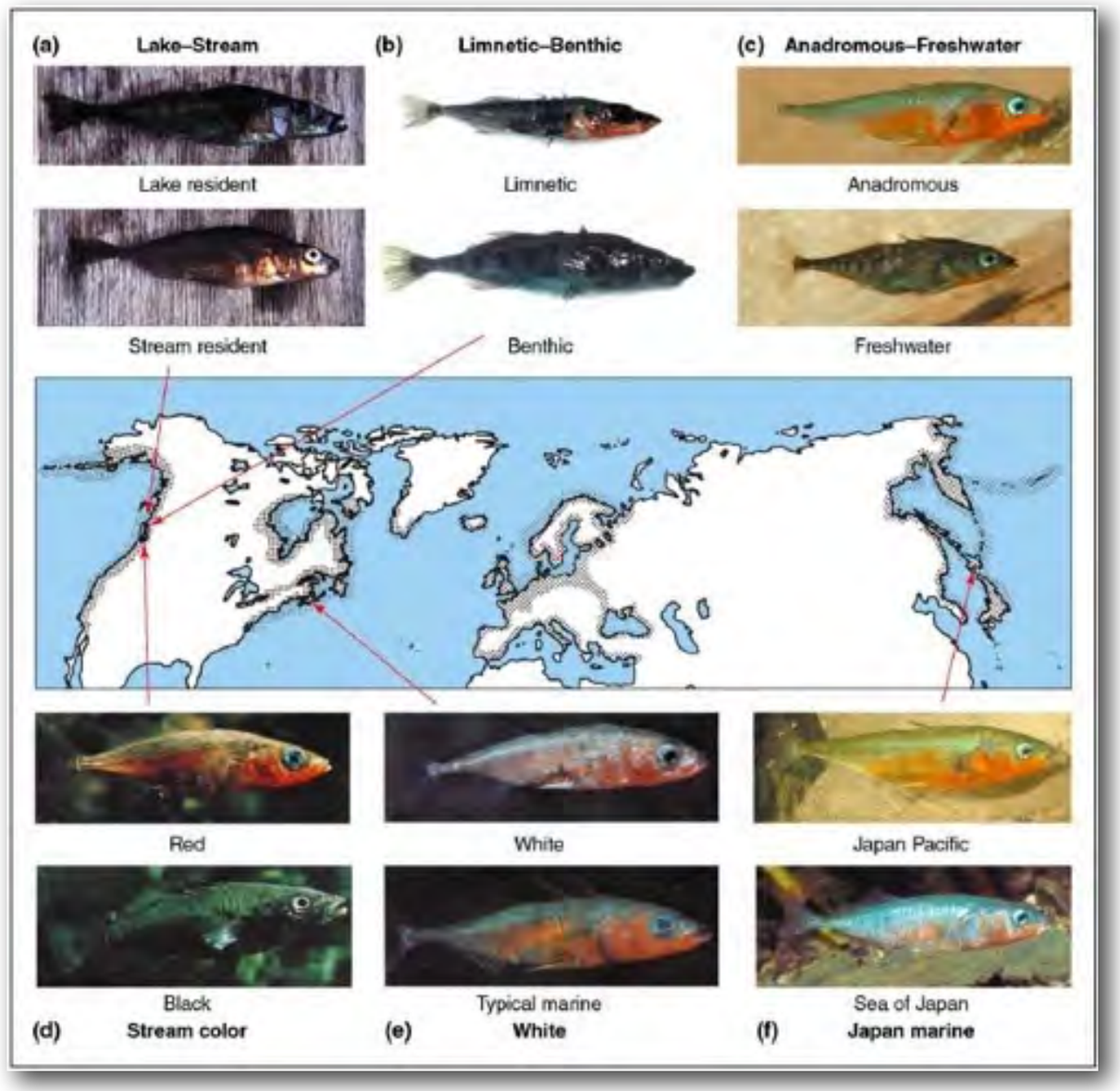


Rundle and McKinnon 2002

Threespine stickleback, *Gasterosteus aculeatus*

Pelvic Structure

Lateral Plates

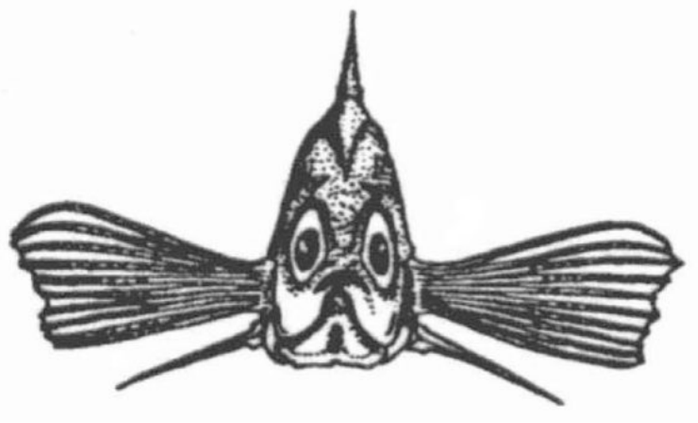


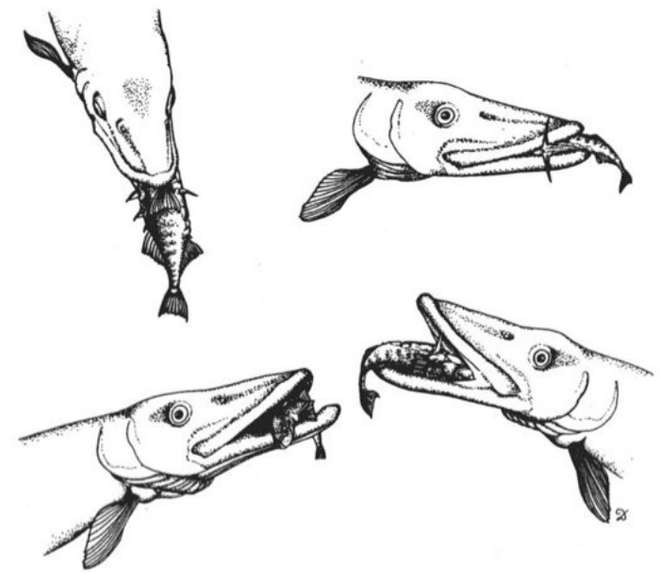
Rundle and McKinnon 2002

Threespine stickleback, *Gasterosteus aculeatus*

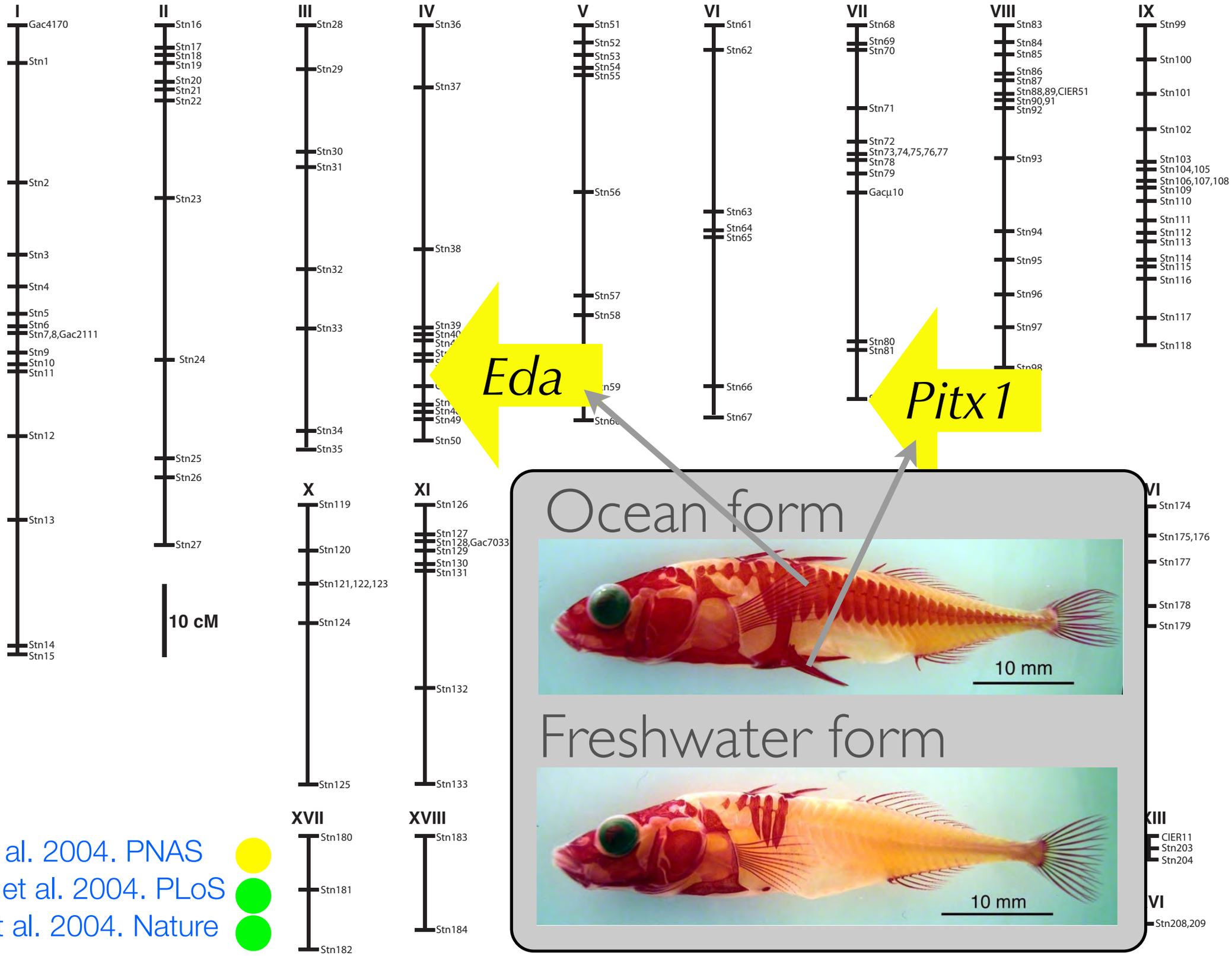
Pelvic Structure

Lateral Plates





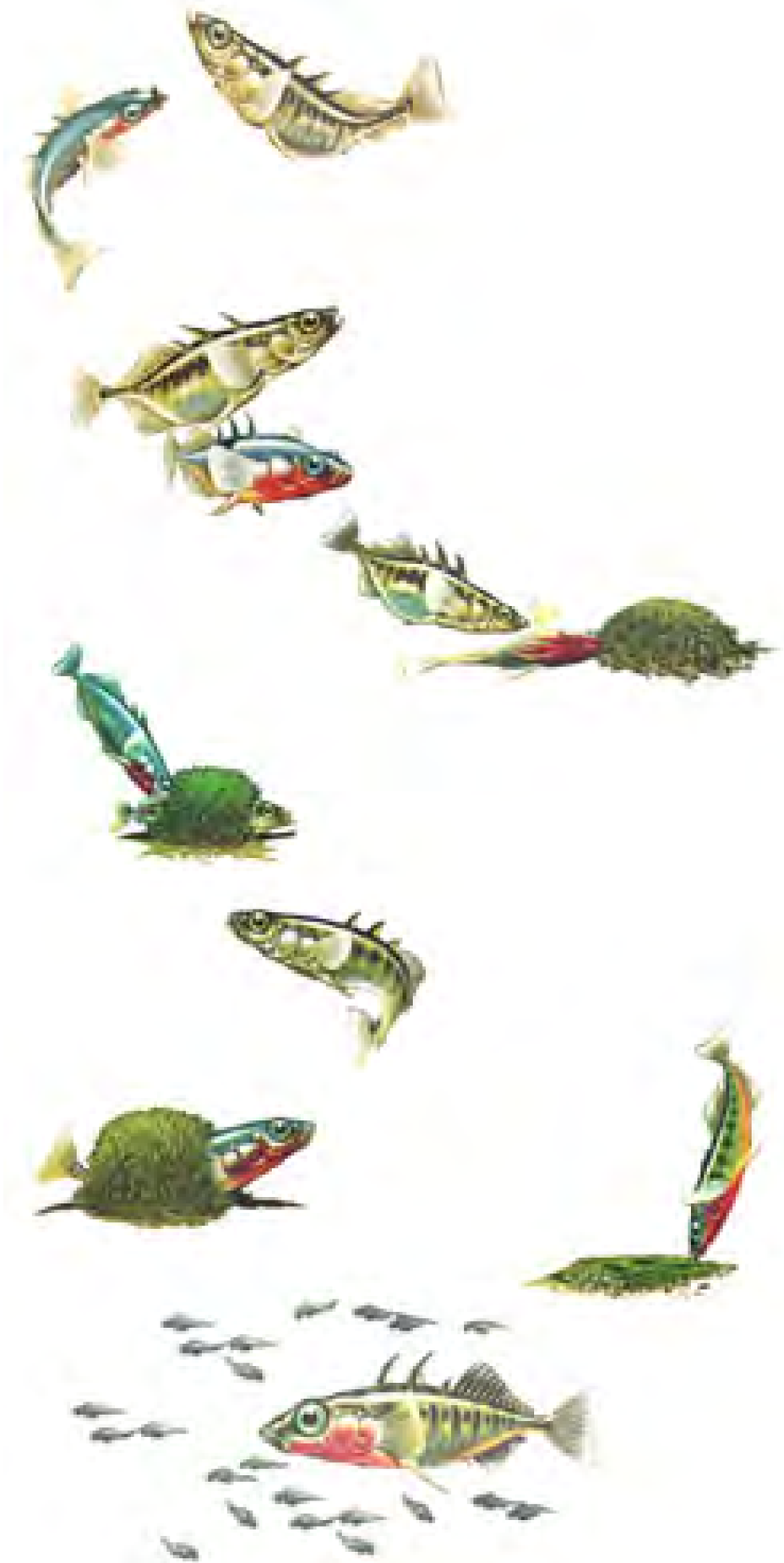
Laboratory mapping of large effect loci

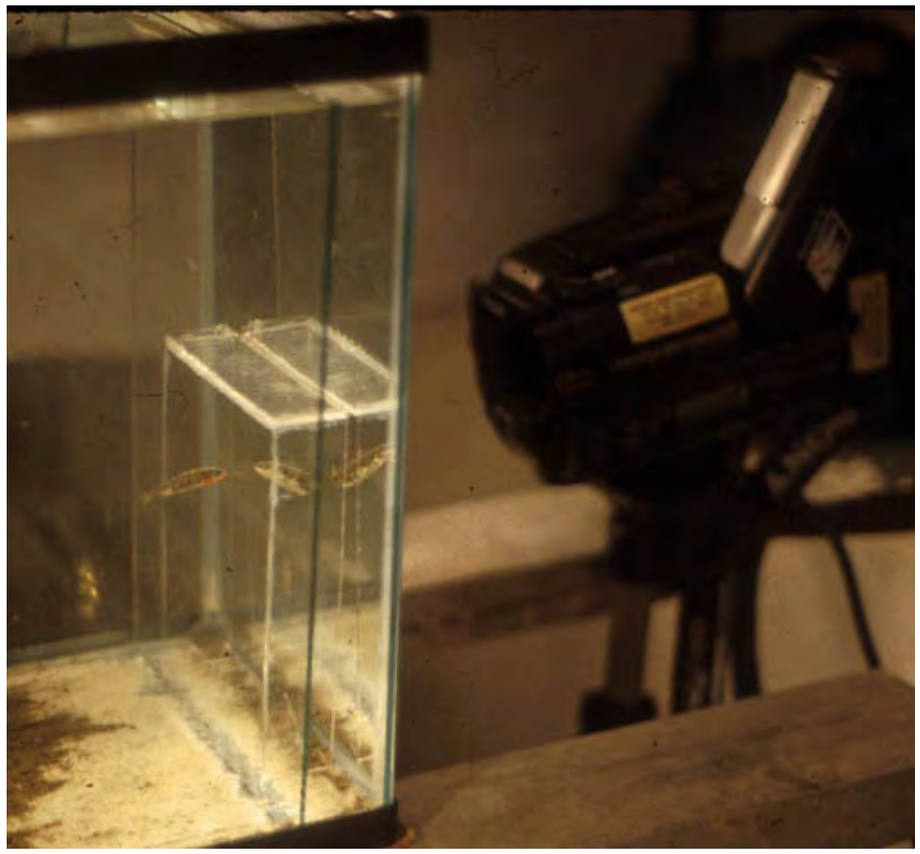


Cresko et al. 2004. PNAS

Colosimo et al. 2004. PLoS

Shapiro et al. 2004. Nature





Stickleback phenotypes mapped in the lab so far.....

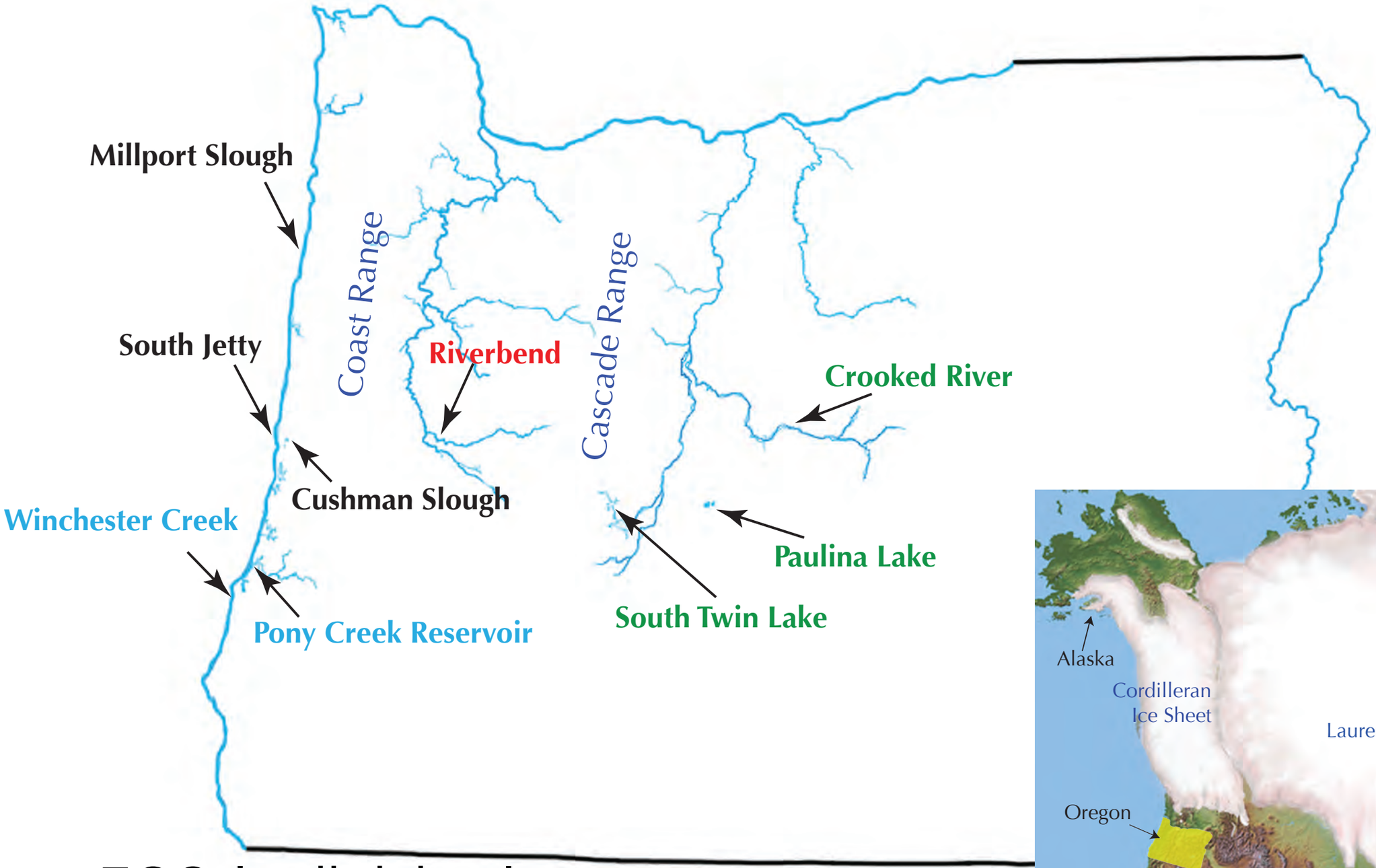
Pelvic structure size and shape *** (*Eda*)
Lateral plate number *** (*Pitx1*)
Body coloration *** (*KitL*)
Opercle bone shape
Pelvic spine length
Body shape
Courtship behavior
Gill raker size
Dorsal spine length

- A trend of large effect loci identified in the laboratory

- Similar genomic regions and sometimes alleles mapped in independent populations

- A question is whether population genomics studies can provide complementary and more complete information.

Population genomic structure of Oregon stickleback

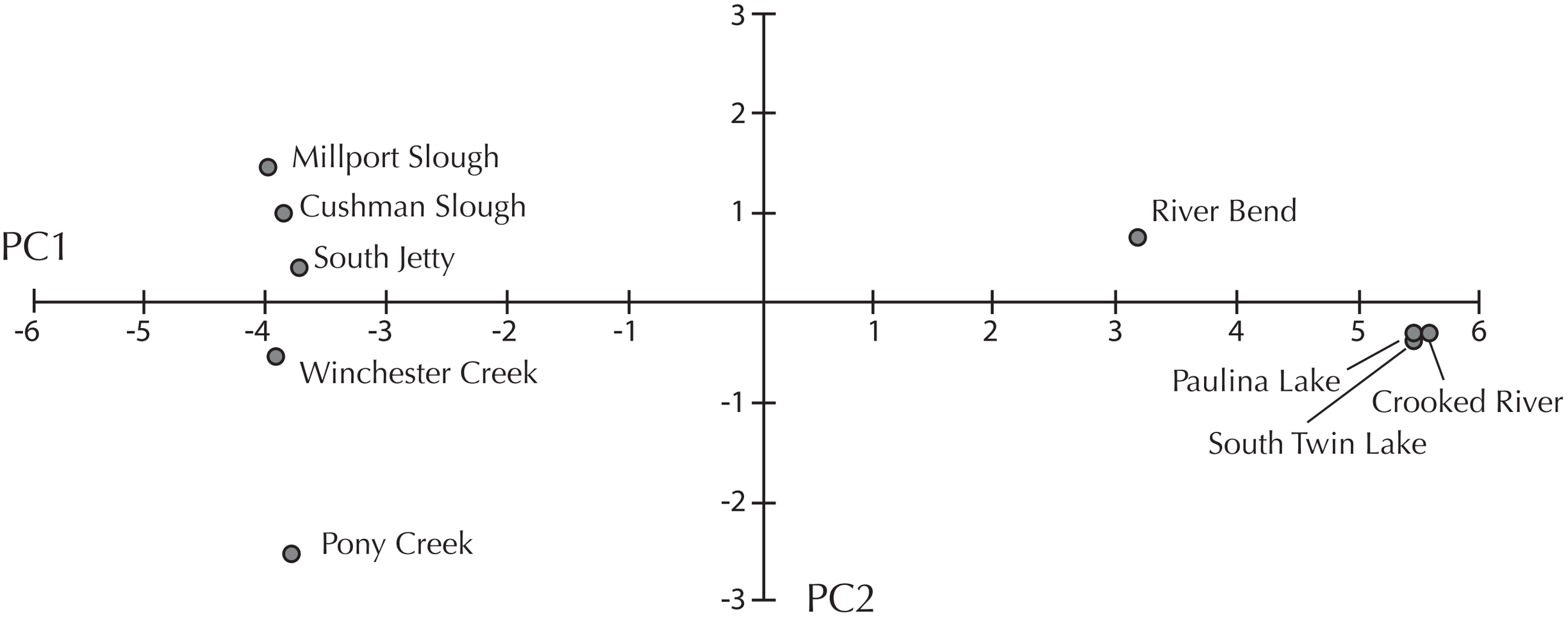


590 Individuals

115,000 SNPs each

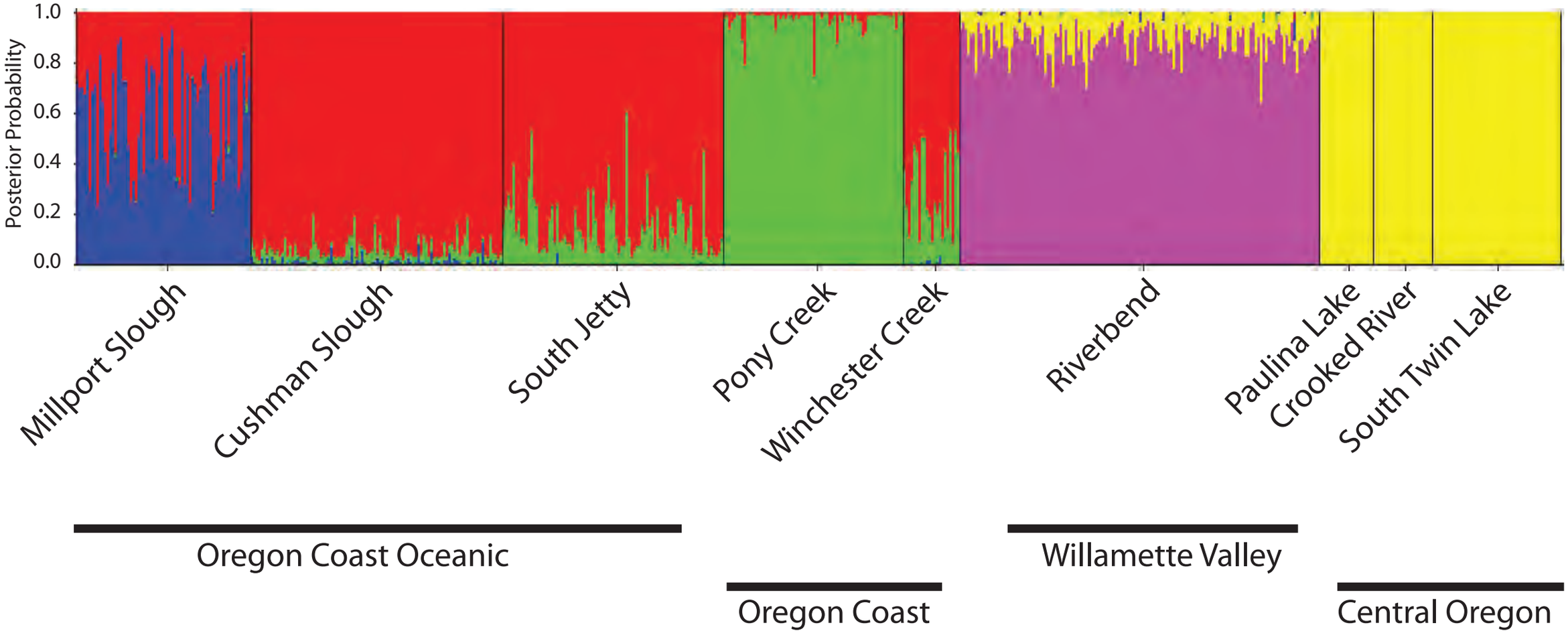
Catchen et al. 2013, Molecular Ecology

Population structure using multivariate statistics (PCA)



PC 1 explains 89% of the overall variance

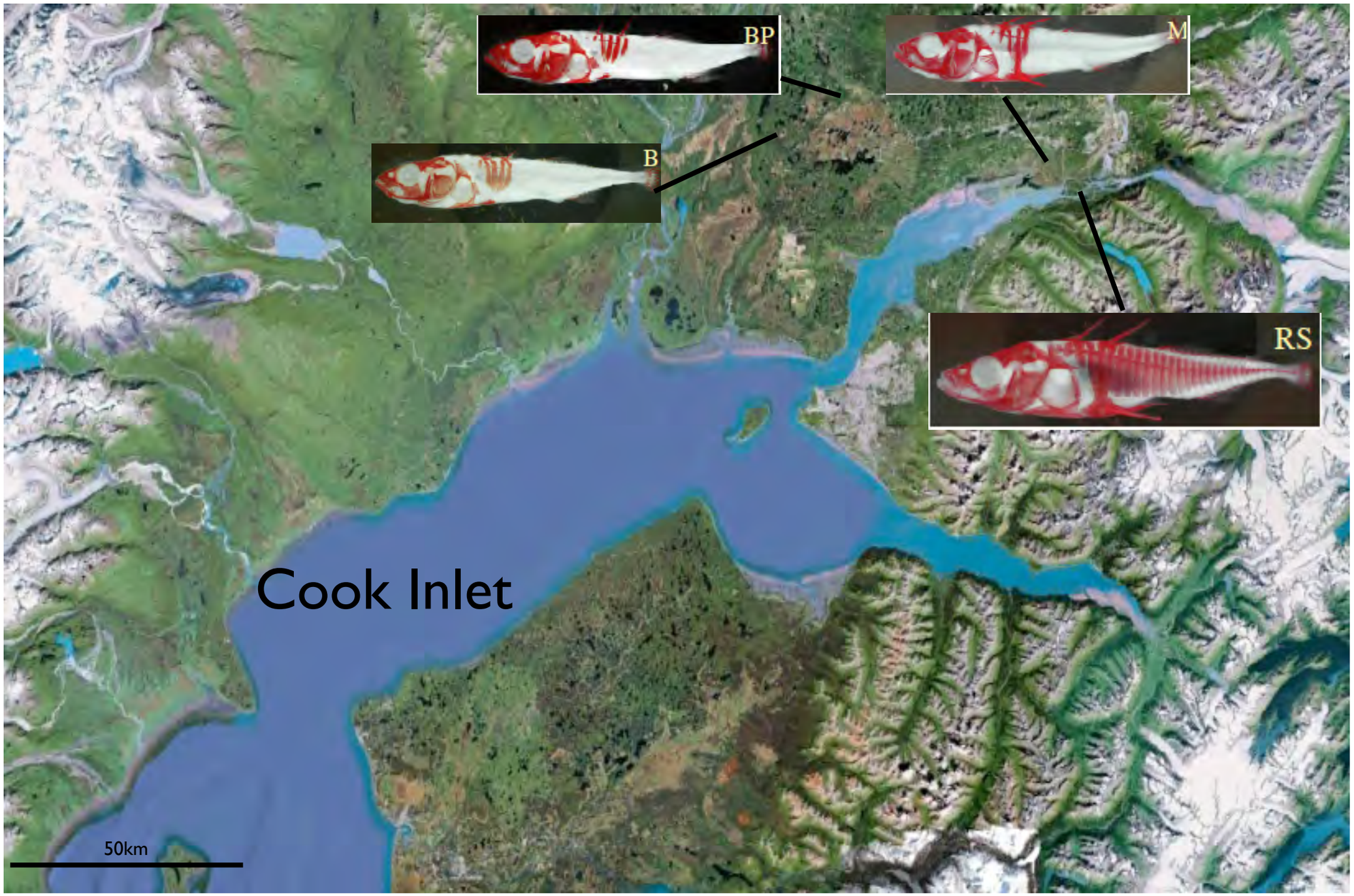
Population structure using Bayesian analysis (*Structure*)



What genomic regions are associated with the different habitats?

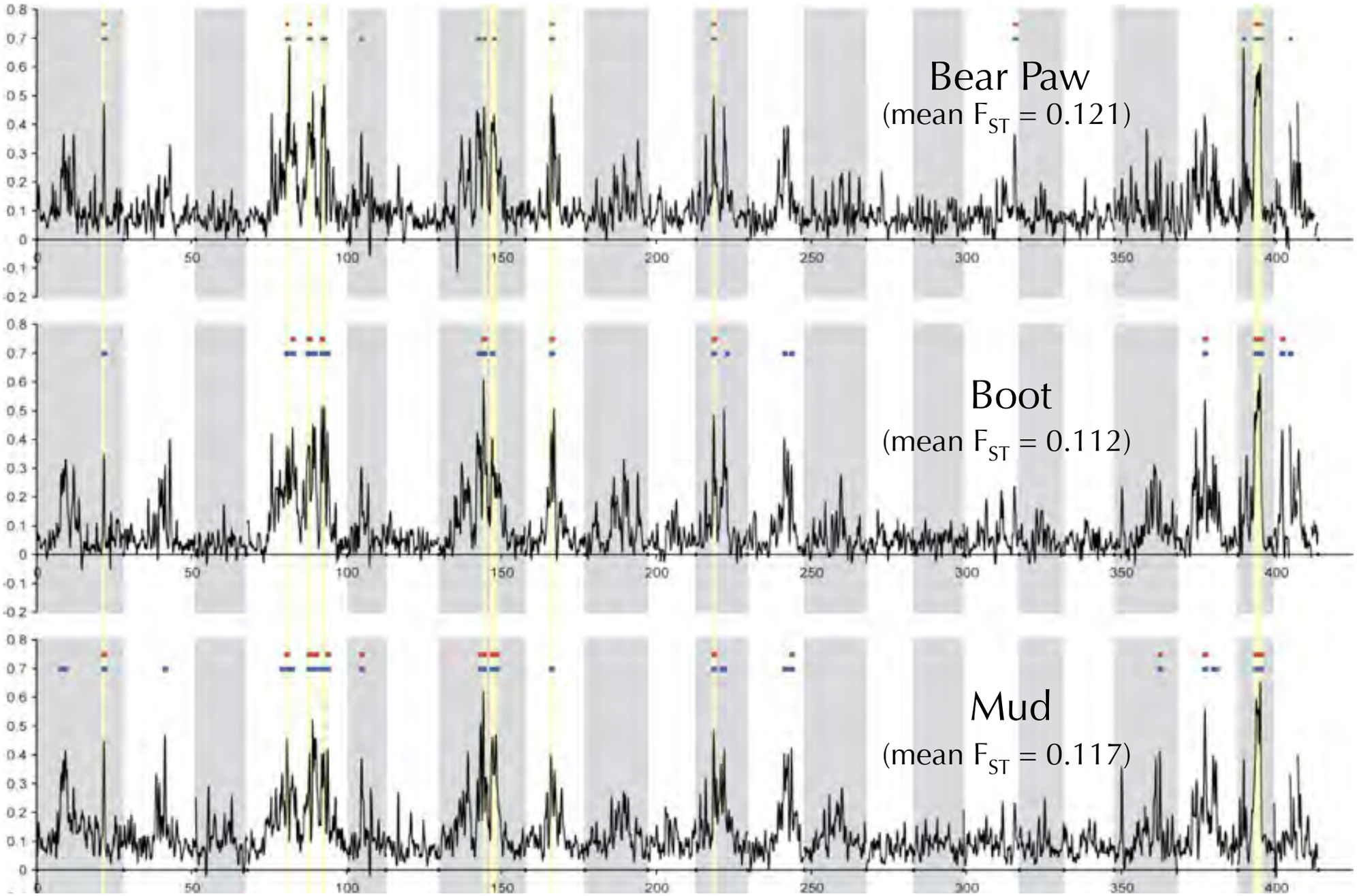
How quickly can the allele frequencies change?

Signatures of natural selection in 13,000 years



Hohenlohe, Bassham et al. 2010. PLoS Genetics

Signatures of natural selection in 13,000 years



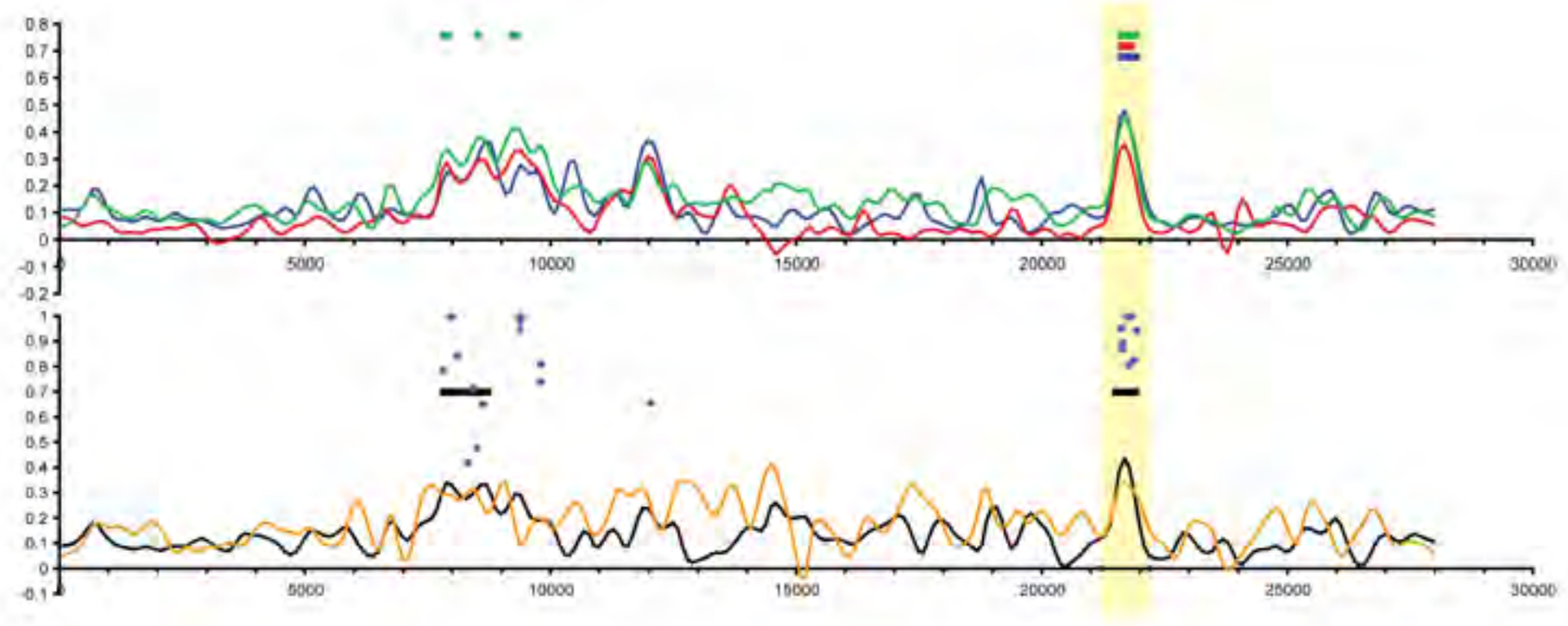
F_{st}

Genomic location (mBases)

Hohenlohe, Bassham et al. 2010. PLoS Genetics

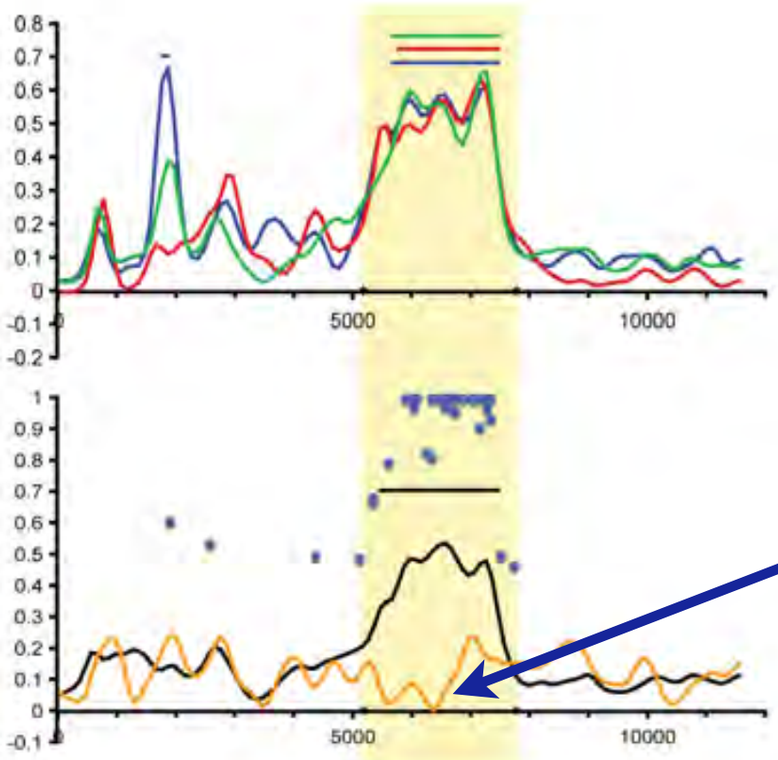
Numerous novel regions identified

LGI



Different alleles

LGXXI



More often the same alleles

Emily Lescak

Julian Catchen

Susan Bassham

Mary Sherbick

Frank von Hippel

Evolution of stickleback in 50 years on earthquake-uplifted islands

Emily A. Lescak[a,b], Susan L. Bassham[c], Julian Catchen[c,d], Ofer Gelmond[b,1], Mary L. Sherbick[b], Frank A. von Hippel[b], and William A. Cresko[c,2]

[a]School of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, AK 99775; [b]Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK 99508; [c]Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403; and [d]Department o University of Illinois at Urbana–Champaign, Urbana, IL 61801

Edited by John C. Avise, University of California, Irvine, CA, and approved November 9, 2015 (received for review June 19, 2015)

Middleton Island - 50 year old populations

1955

2008


Independent

Giant Quake Smashes Alaska Cities; 100 Feared Dead, Tidal Waves Roll

Photo Credits: BLM, E-Terra

Middleton Island - 50 year old populations

1955

2008

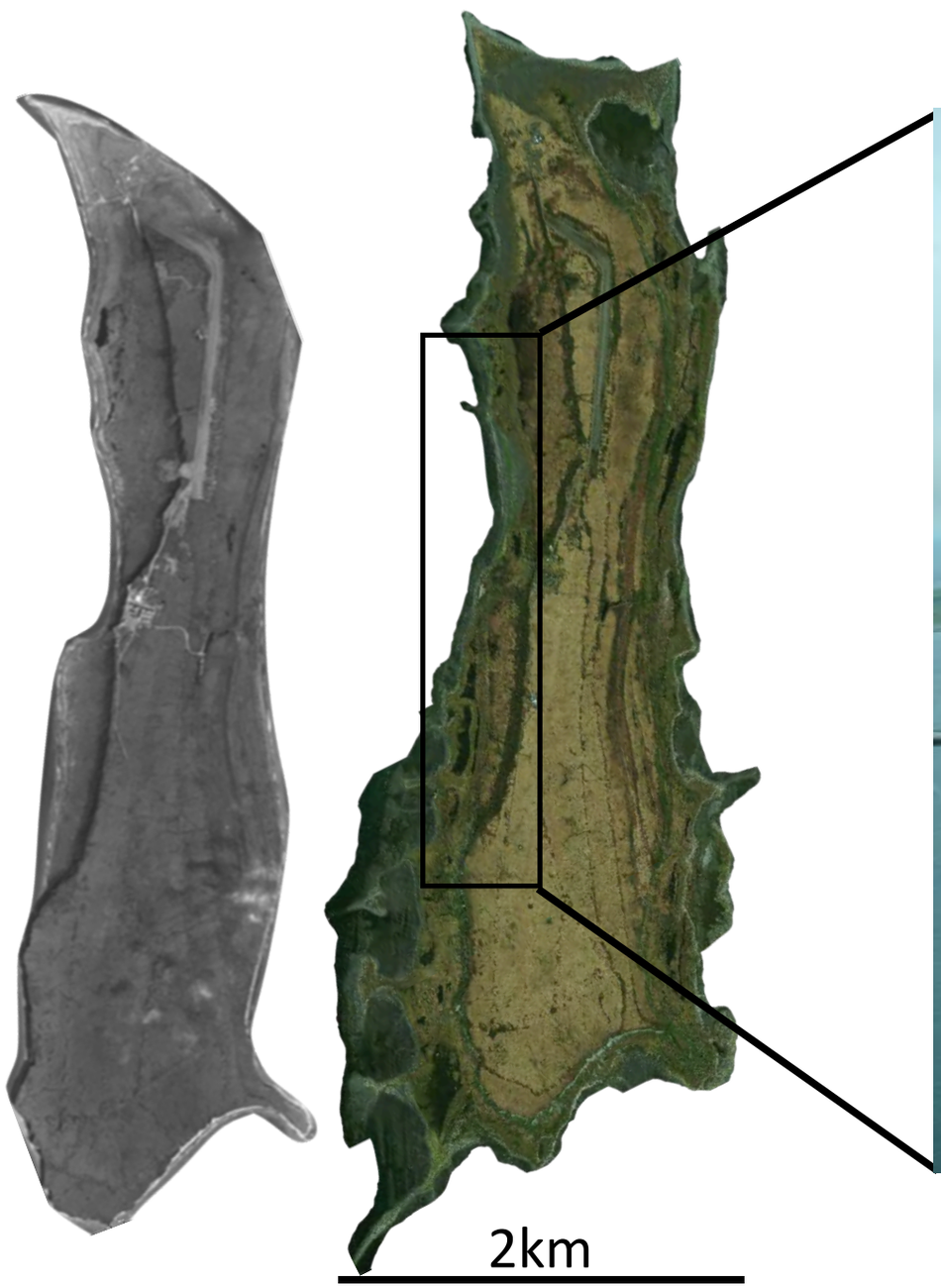


Photo Credits: BLM, E-Terra

Middleton Island - 50 year old locations

1955

2008

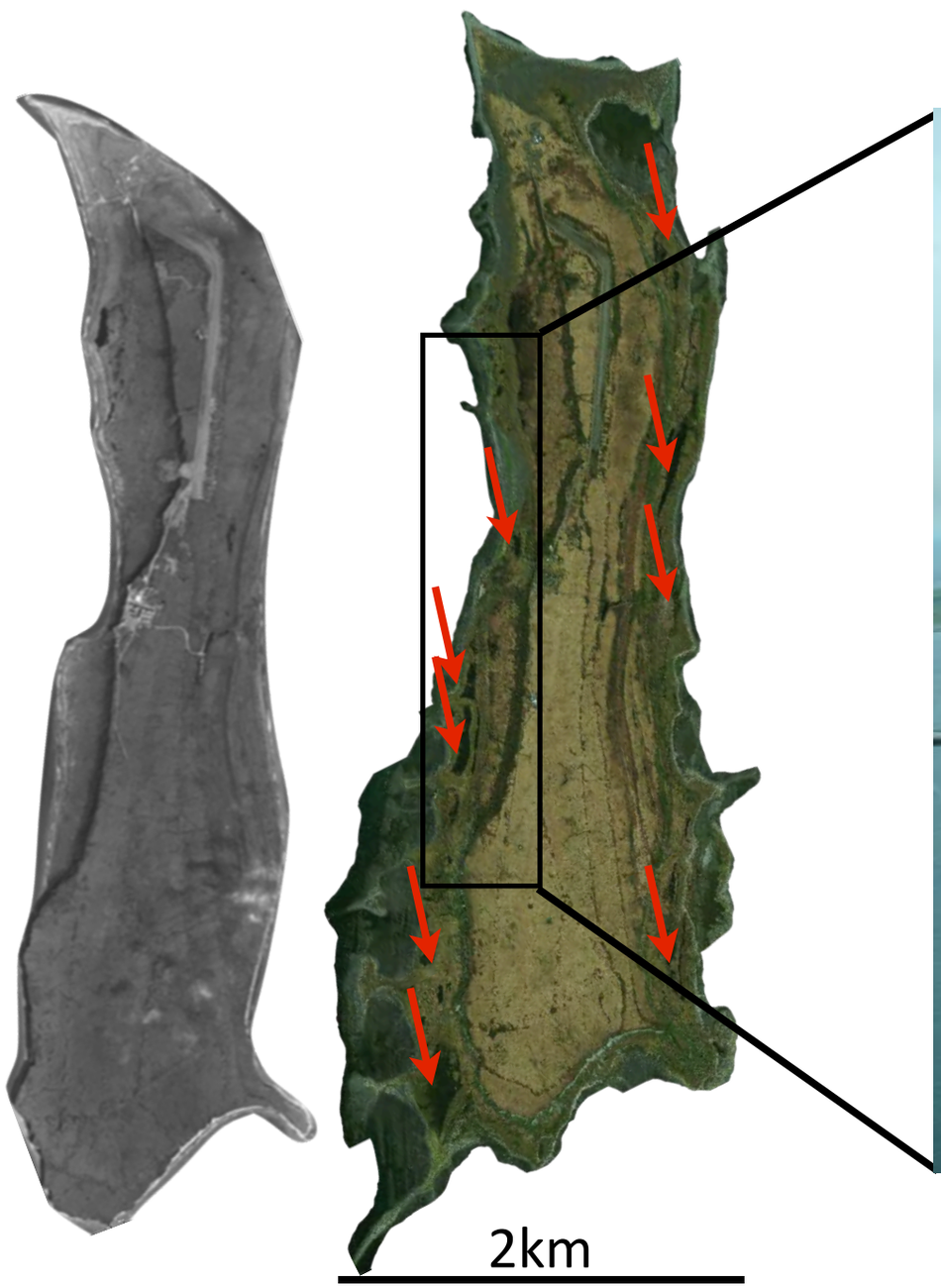


Photo Credits: BLM, E-Terra


MPS SITE 13 BAR


N34WM


DANGER
HIGH
VOLTAGE

B)

Danger Island

Da02

Da04

0.5 km

Oceanic Habitat

Freshwater Habitat

Completely Plated

Low Plated

C)

Montague Island

Mo30

Mo33

Mo35

Mo36

Mo37

40 km

D)

N

Middleton Island

Mi15

Mi07

Mi23

Mi06

Mi16

Mi28

Mi13

Mi09

Mi08

Mi12

Mi14

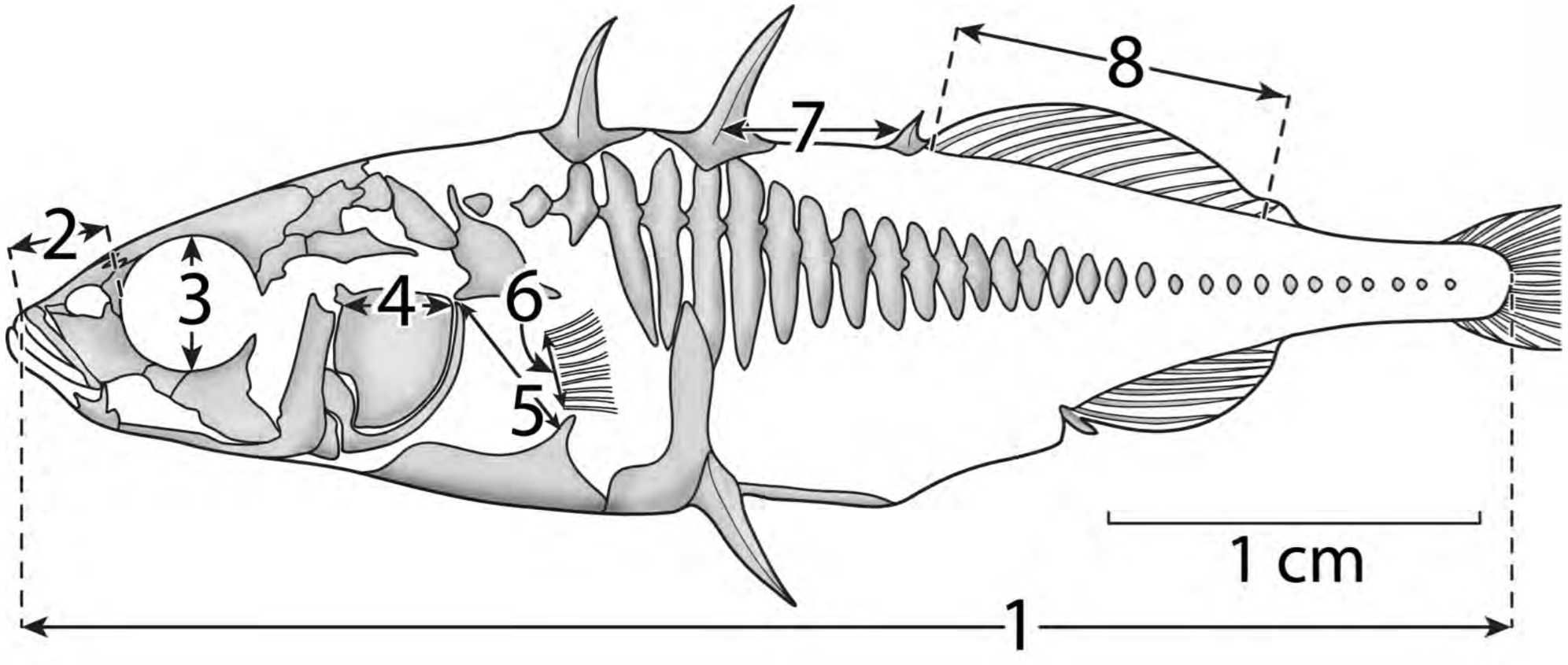
Mi17

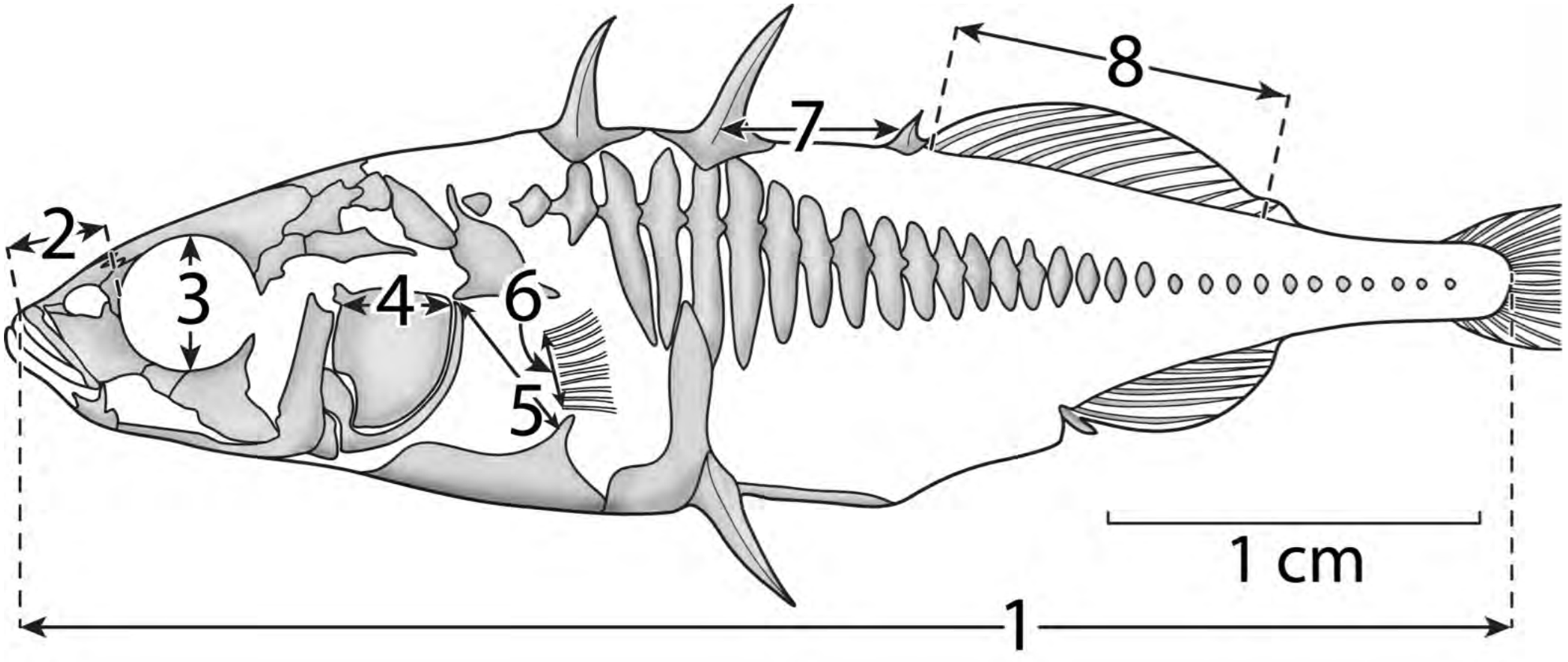
Mi22

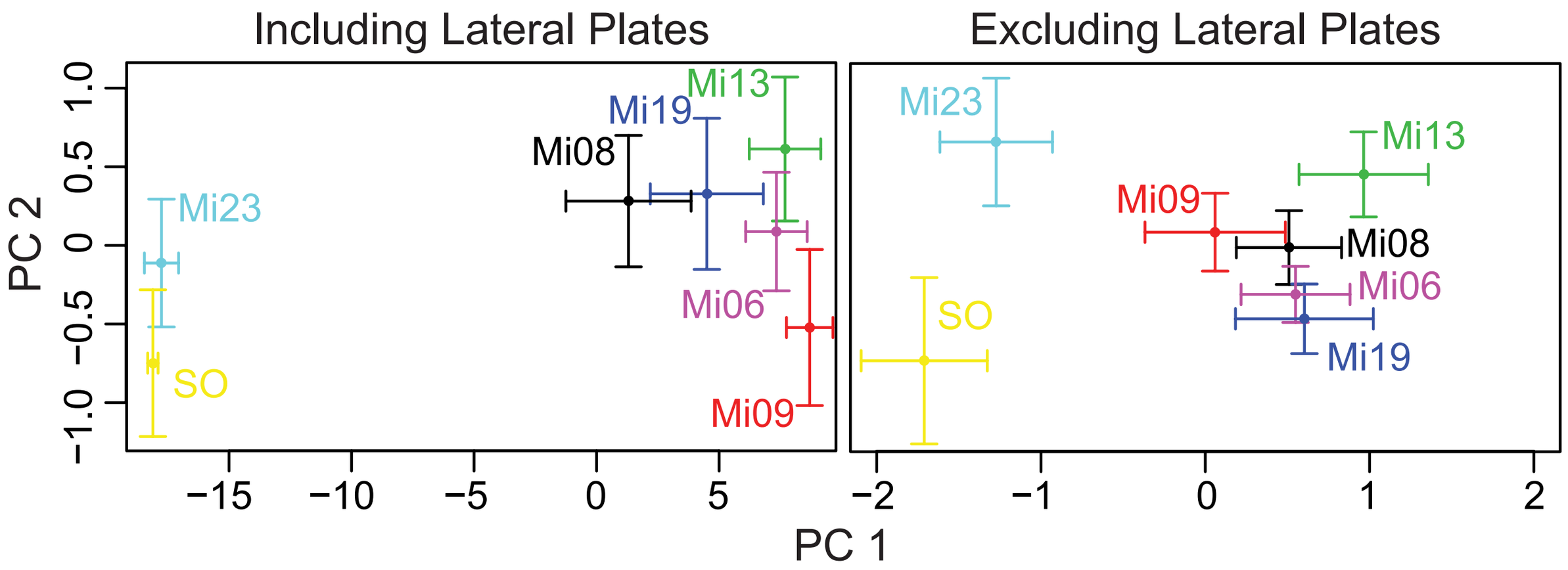
Mi19

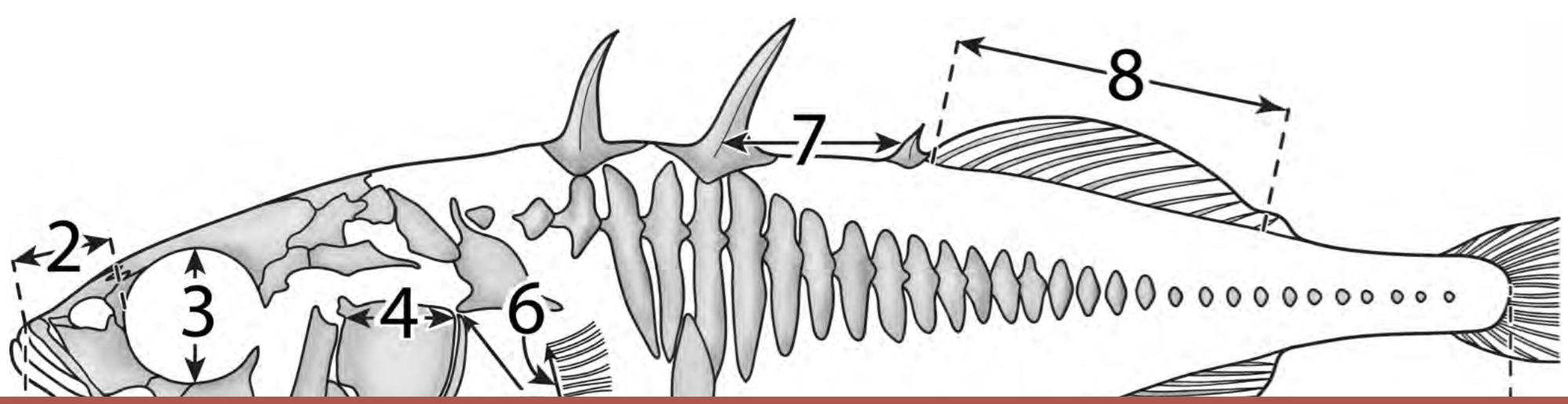
Mi11

2 km


8
7
2
3
4
6
5
1 cm
1


8
7
2
3
4
6
5
1 cm
1


Including Lateral Plates
Excluding Lateral Plates
PC 2
PC 1
1.0
0.5
0
−0.5
−1.0
−15
−10
−5
0
5
−2
−1
0
1
2
Mi23
SO
Mi08
Mi19
Mi13
Mi06
Mi09

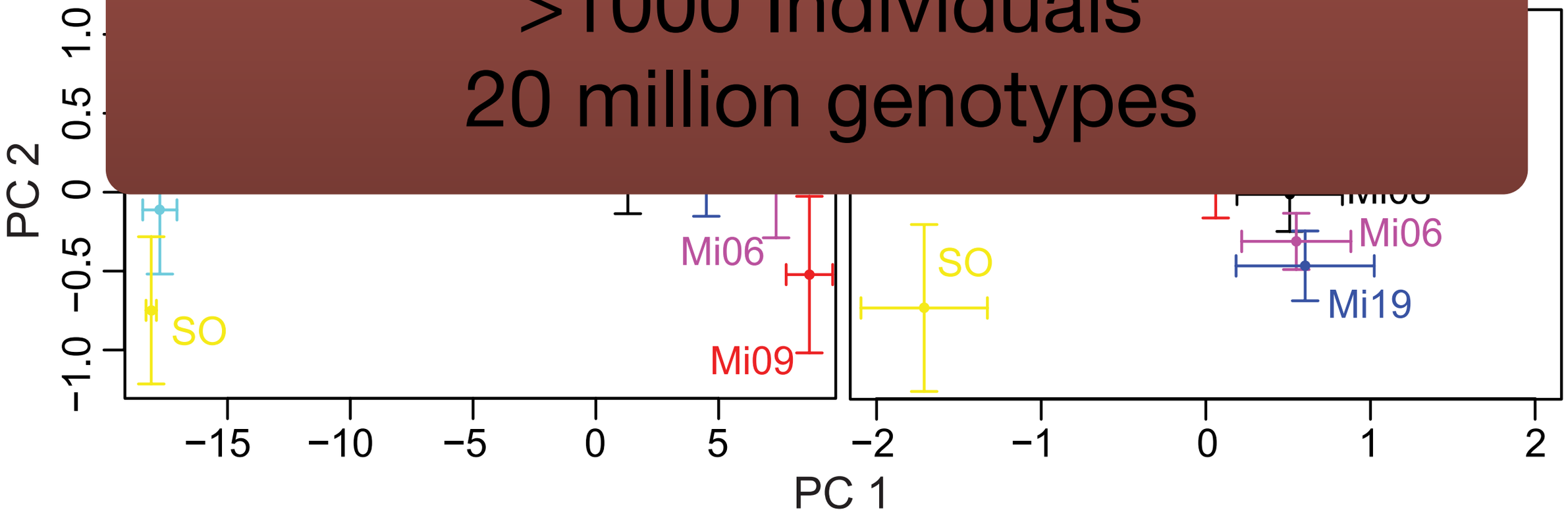


RAD-seq analysis

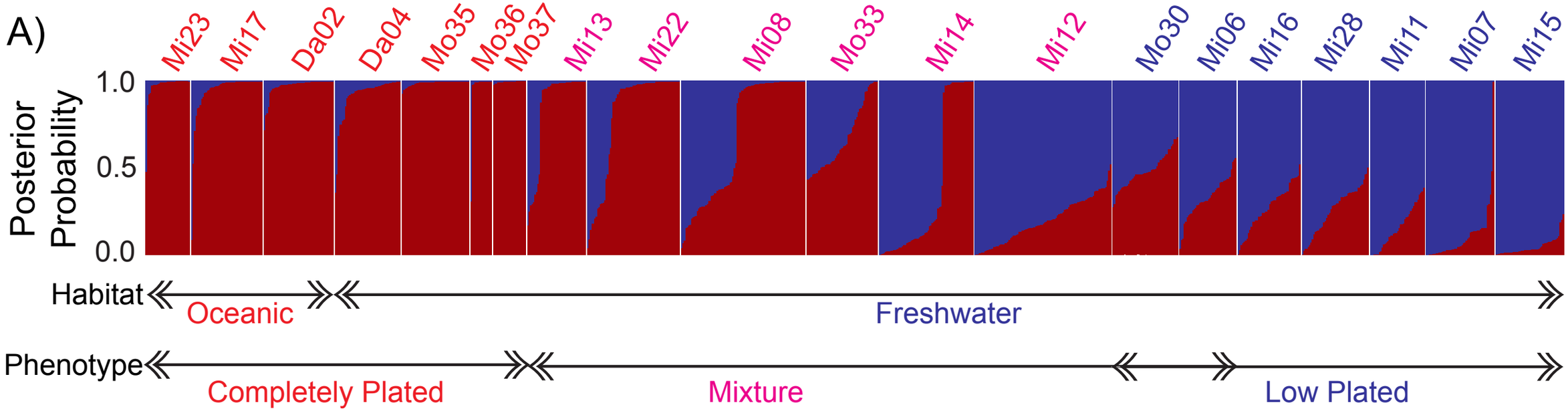
110,000 SNPs per individual

>1000 Individuals

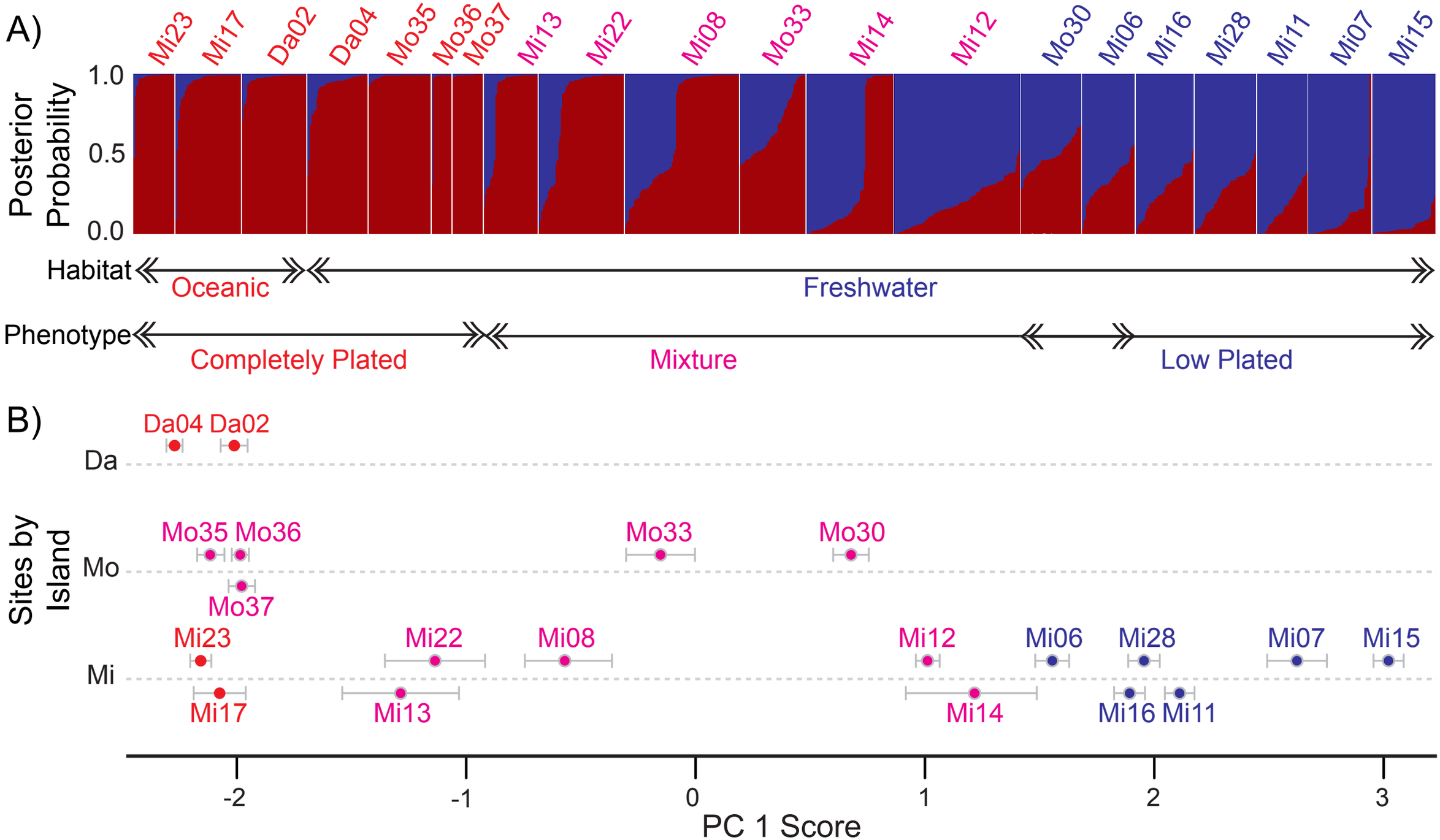
20 million genotypes



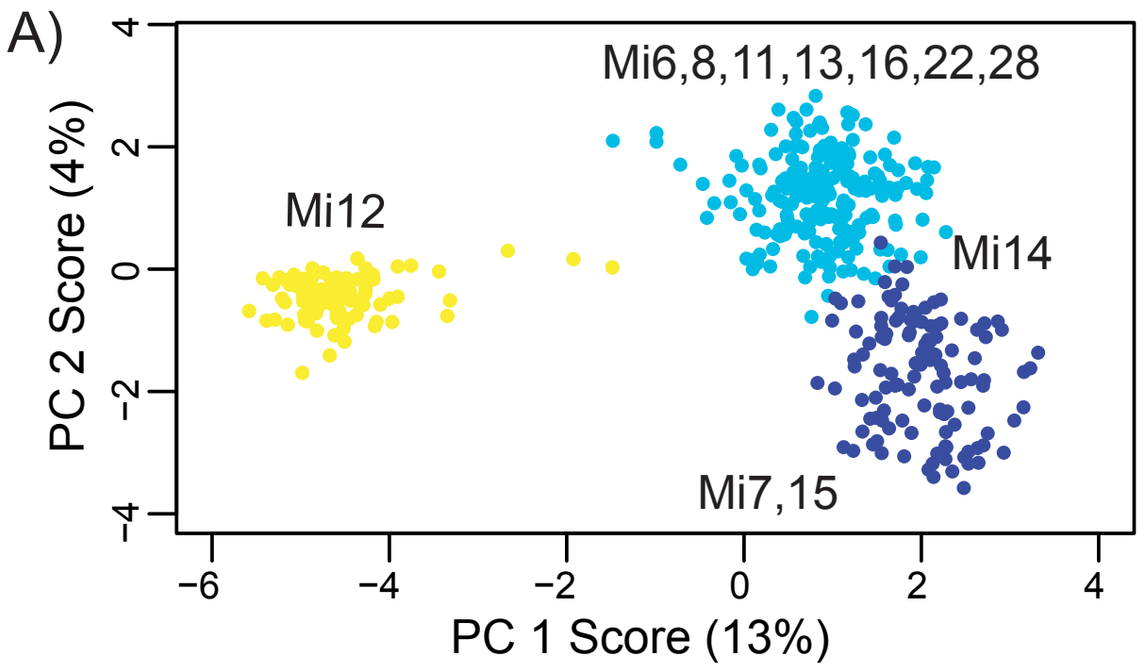
The majority of the genetic variation is partitioned between oceanic and freshwater fish

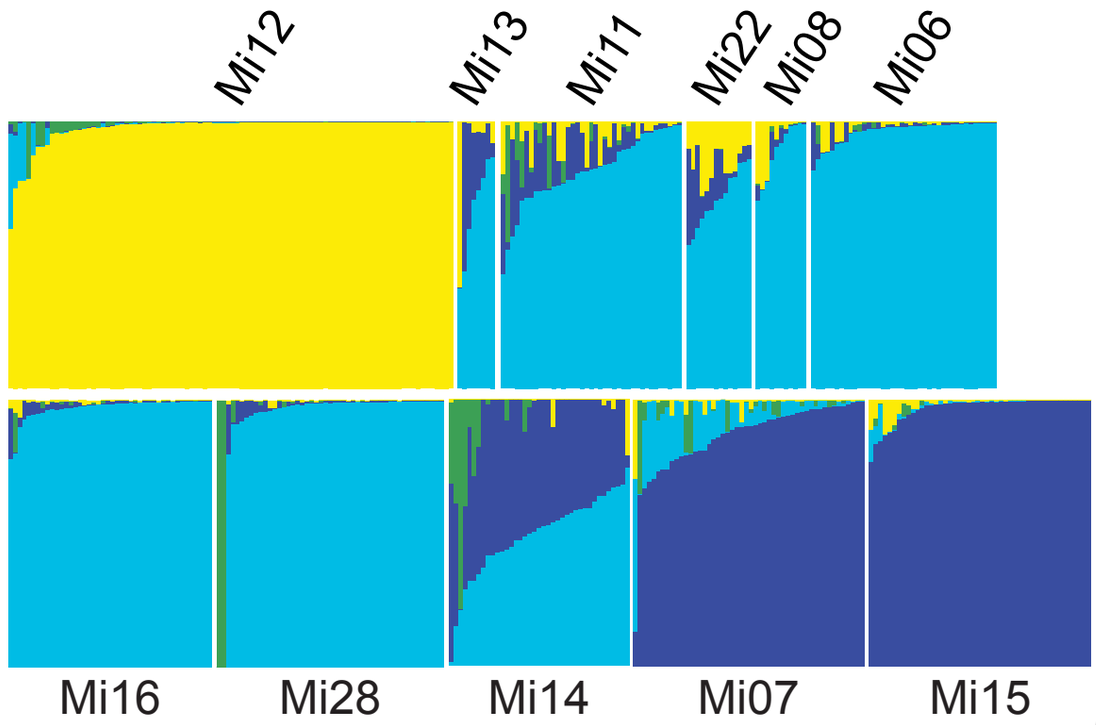


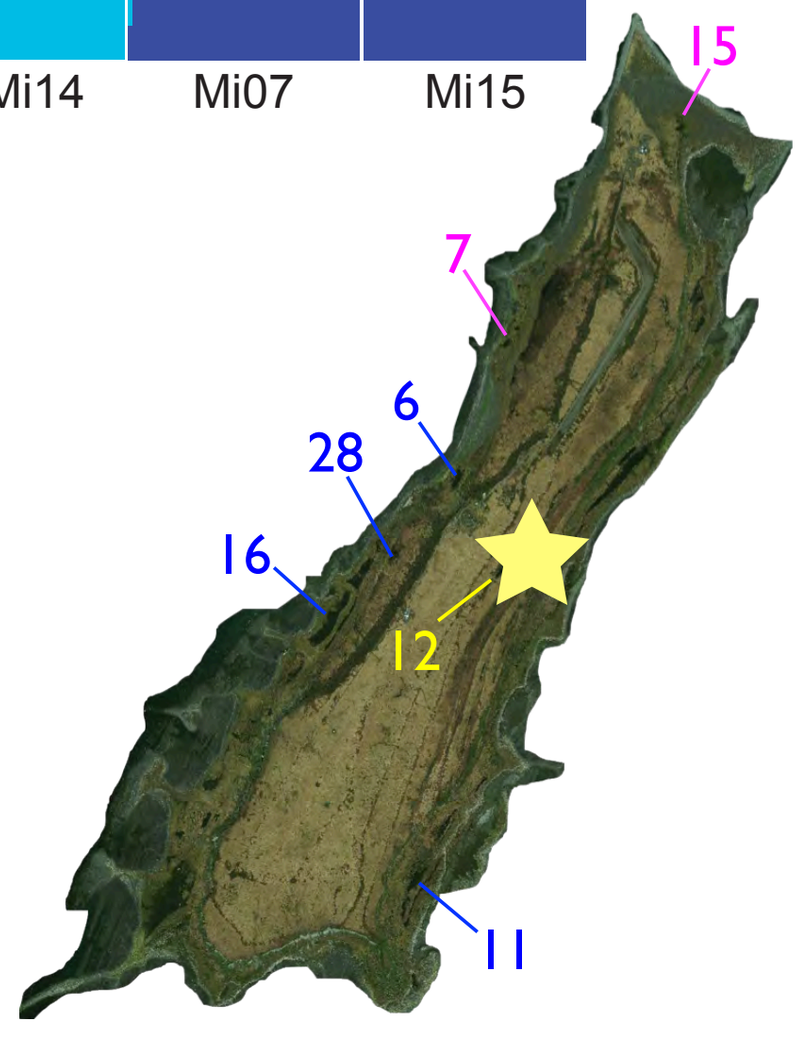
The majority of the genetic variation is partitioned between oceanic and freshwater fish



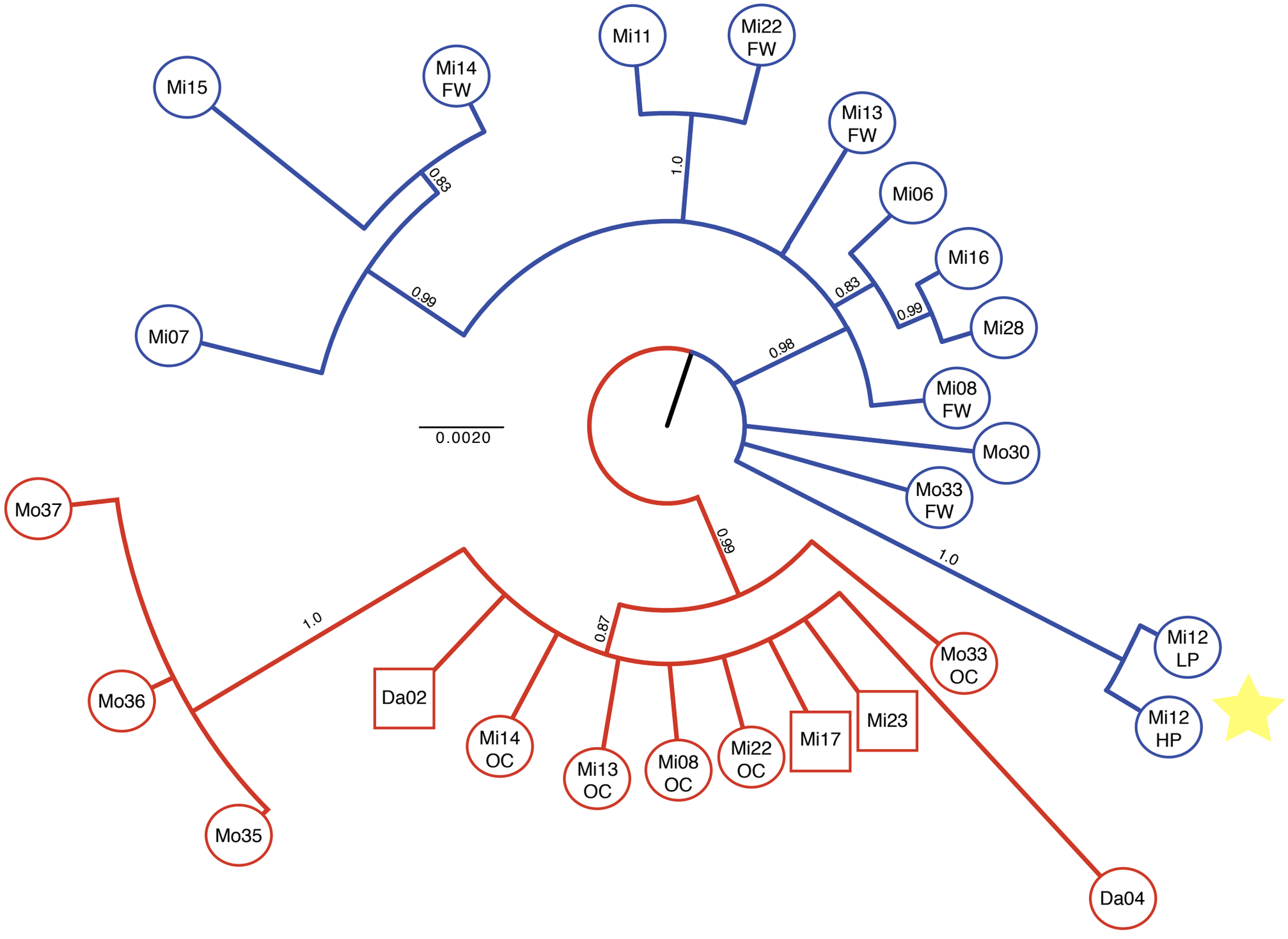
Structure analysis shows independent evolution even among populations on a single island



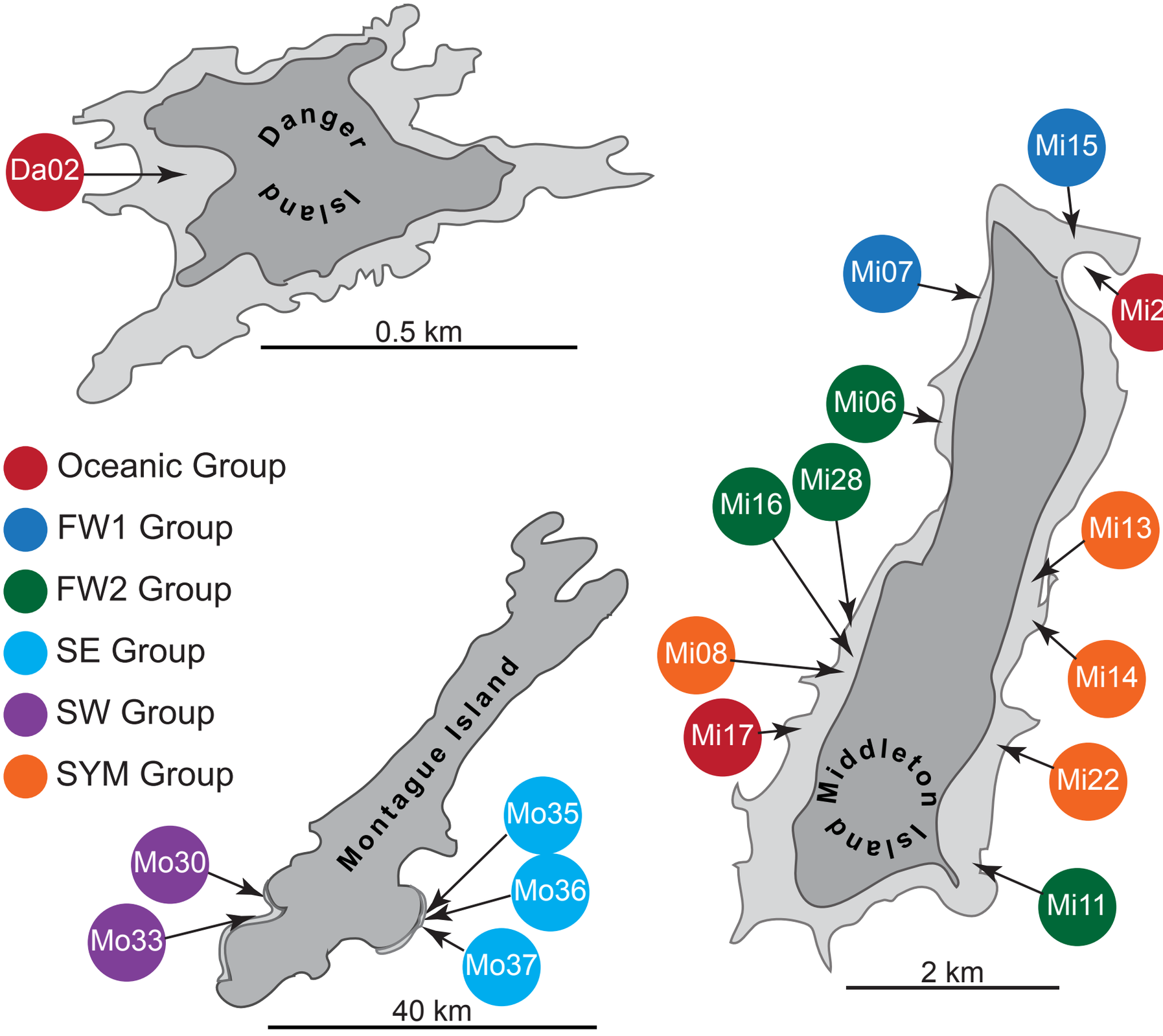




The overall population phylogeny supports independent evolution as well



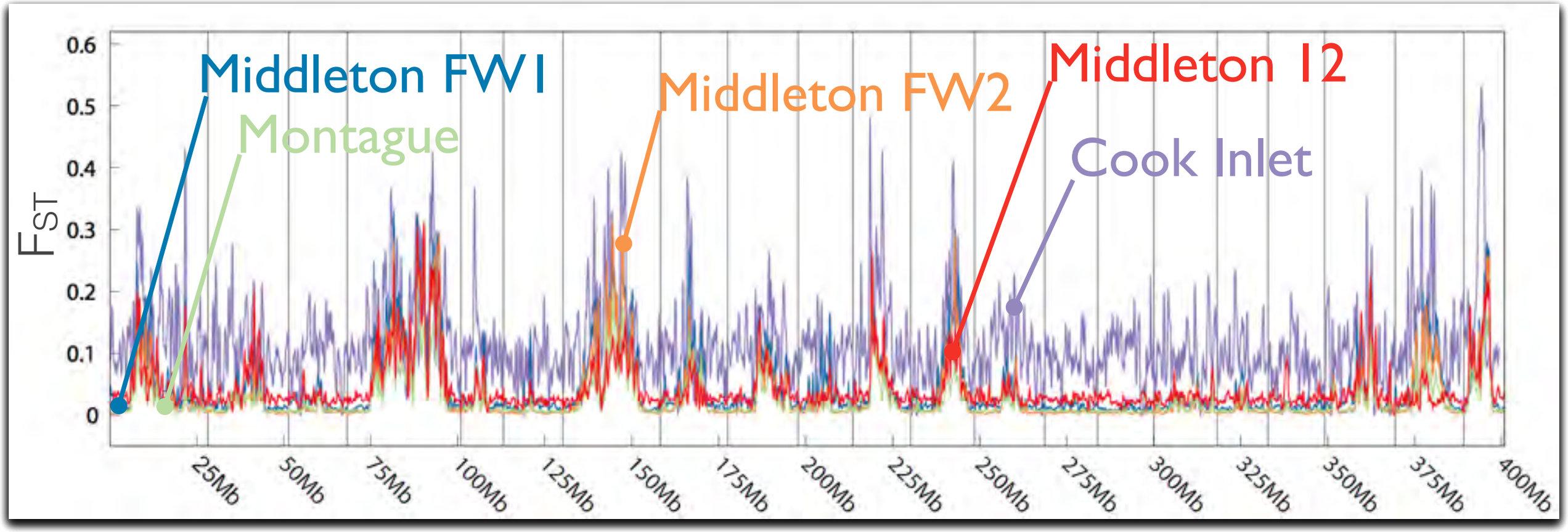
At least six independent evolutionary events in freshwater in the last 50 years



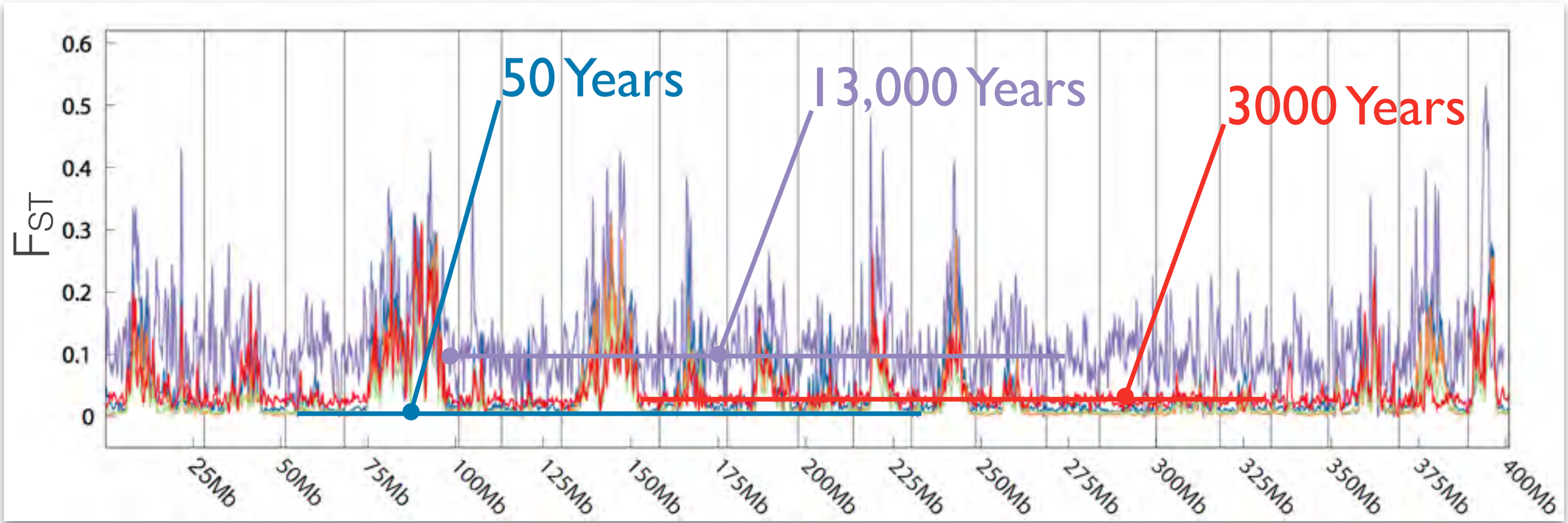
How much of the genome is differentiated?

How similar are the genomic patterns of differentiation?

Bassham, Catchen et al. *in prep*

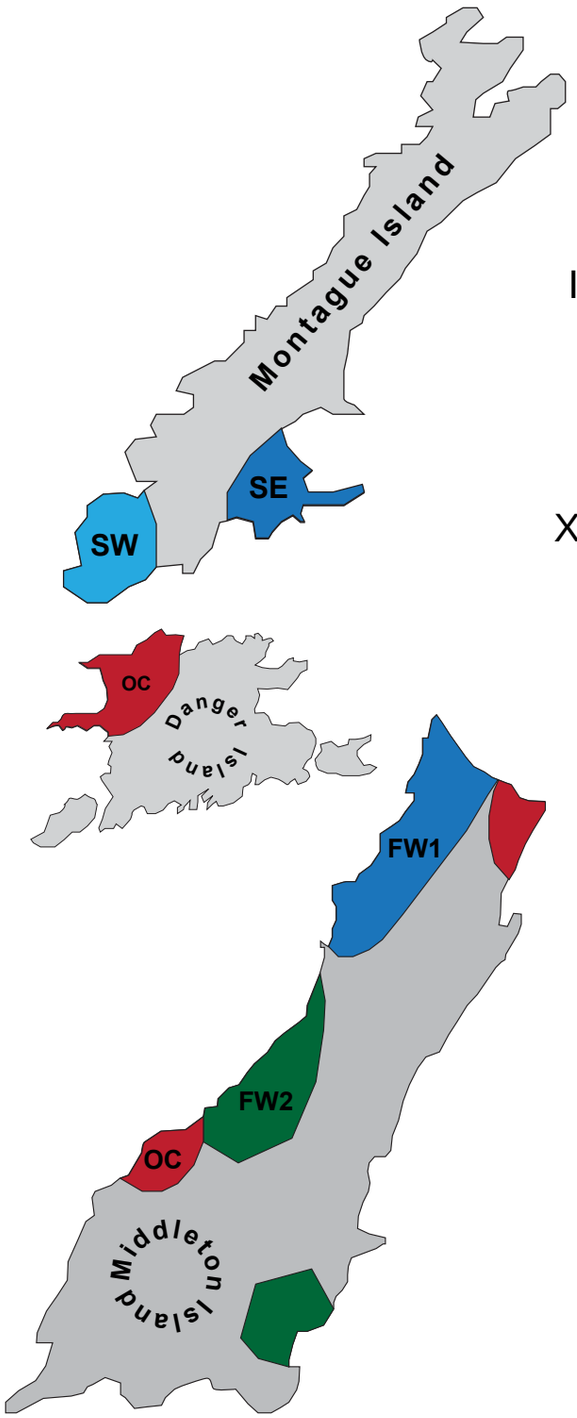


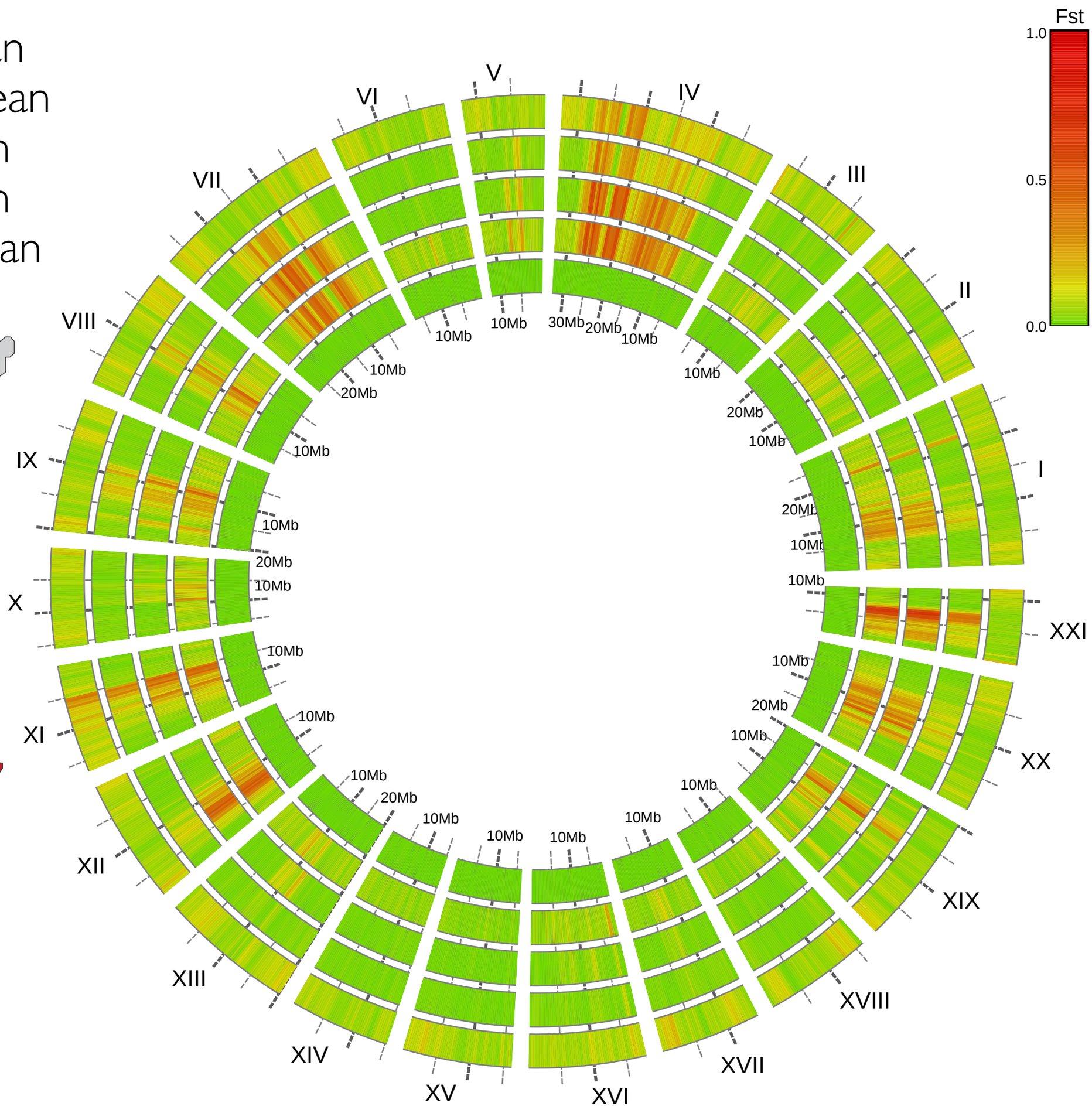
Freshwater Populations (grouped) vs. Marine



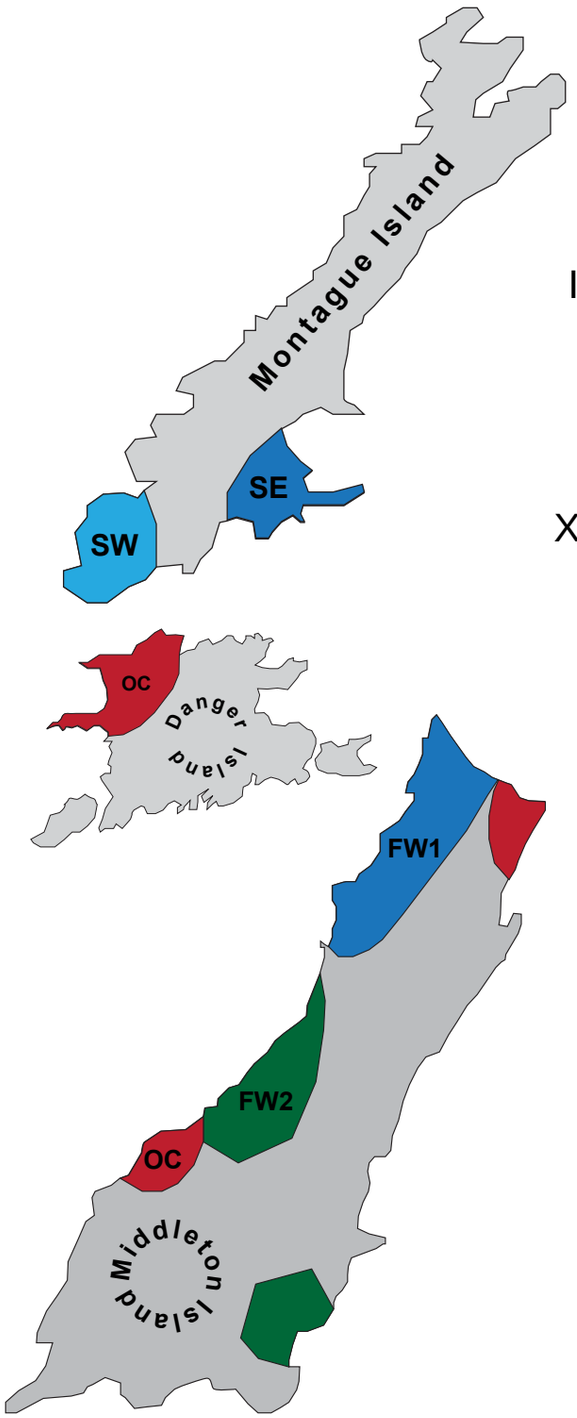
Freshwater Populations (grouped) vs. Marine

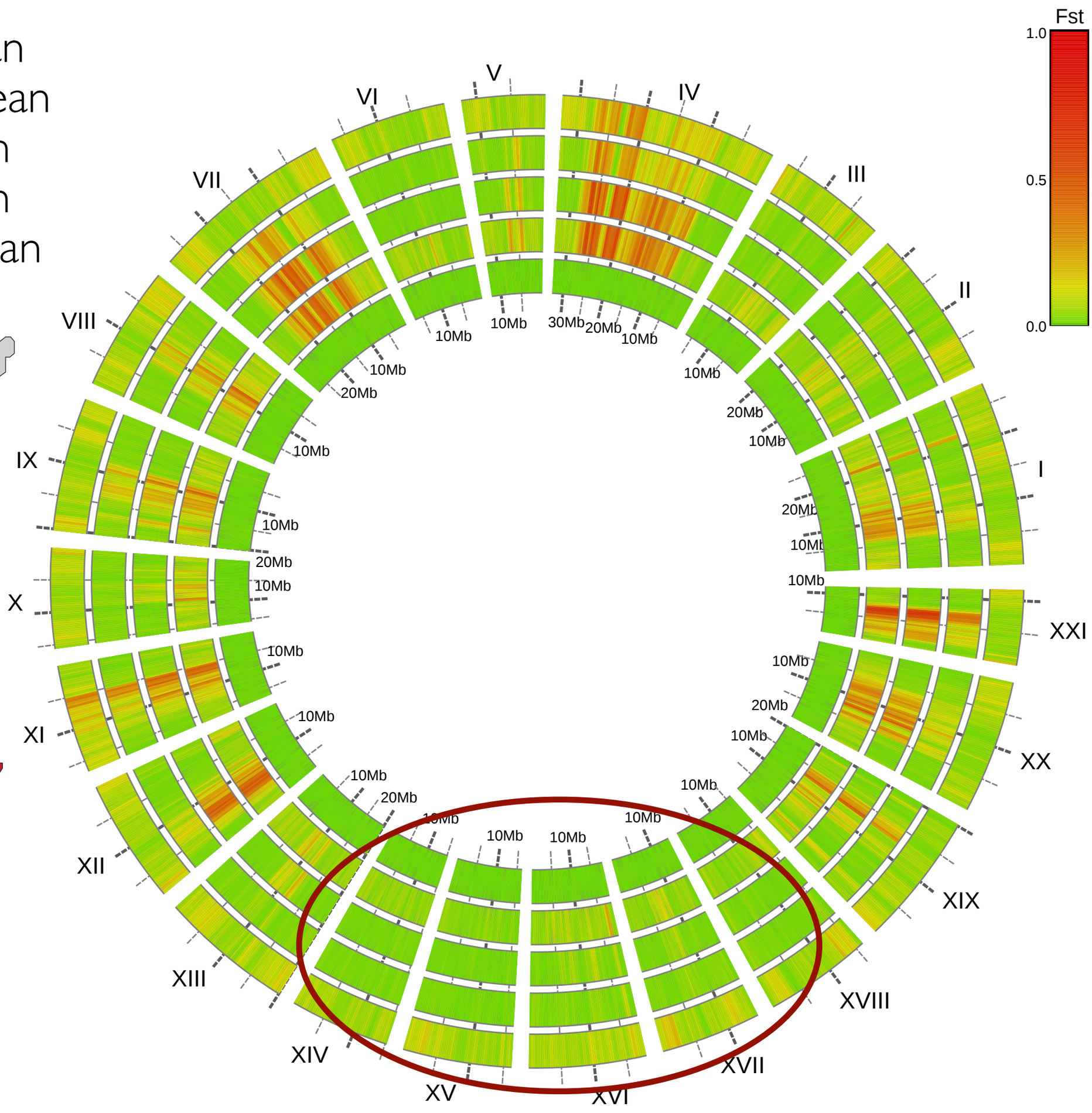
MoSe vs. Ocean
MoSW vs. Ocean
FW2 vs. Ocean
FW1 vs. Ocean
Ocean vs. Ocean



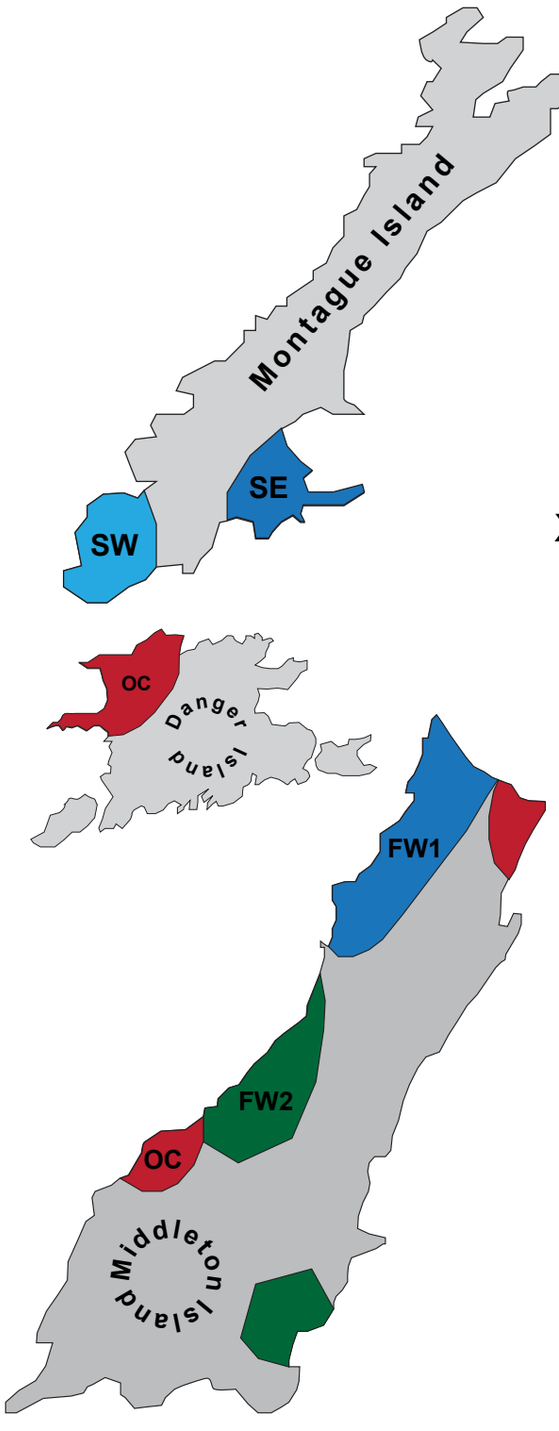


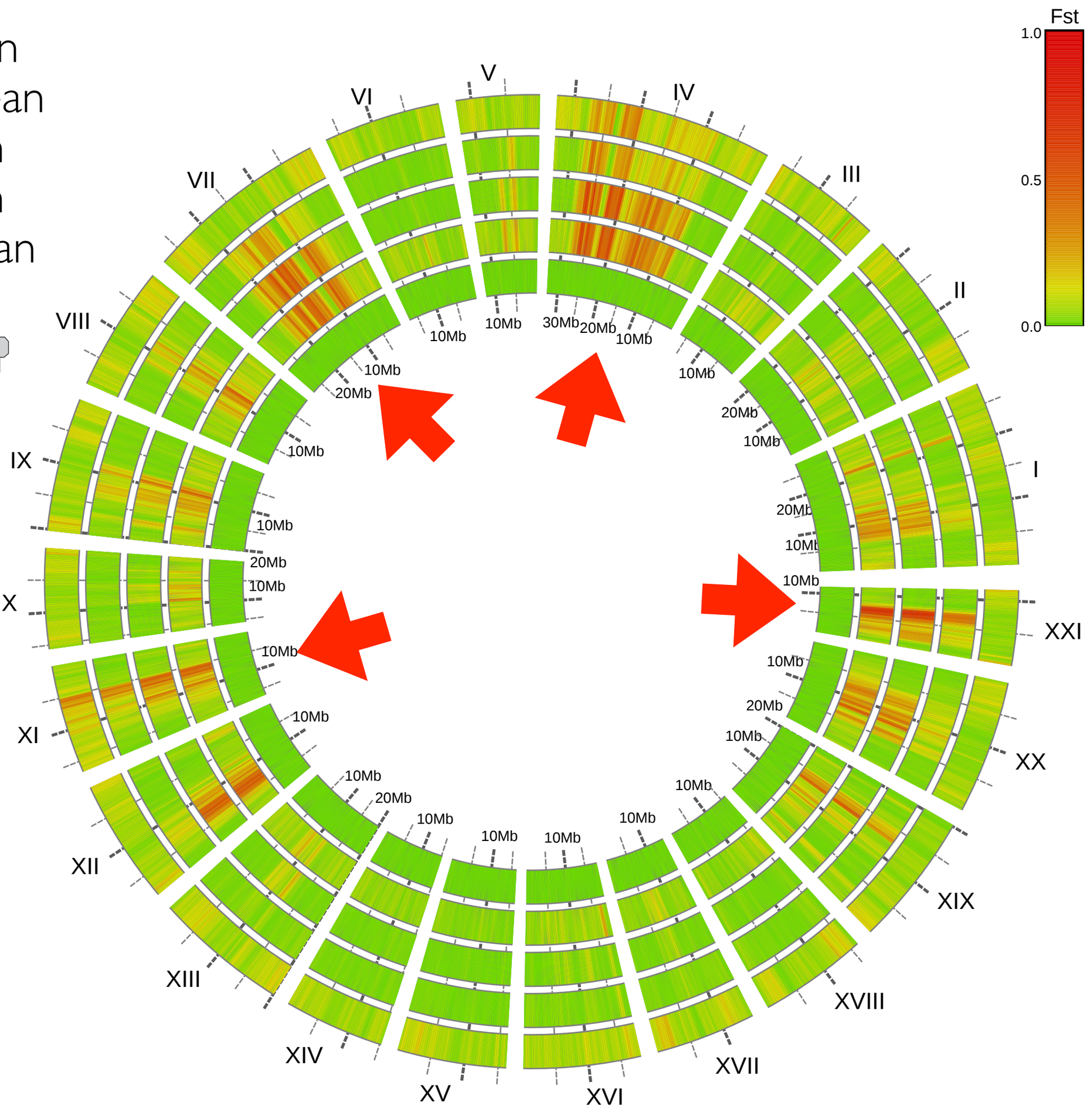
MoSe vs. Ocean
MoSW vs. Ocean
FW2 vs. Ocean
FW1 vs. Ocean
Ocean vs. Ocean

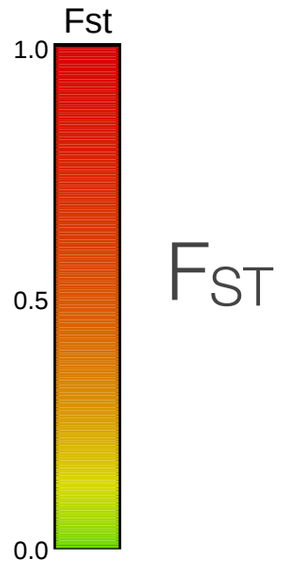




MoSe vs. Ocean
MoSW vs. Ocean
FW2 vs. Ocean
FW1 vs. Ocean
Ocean vs. Ocean

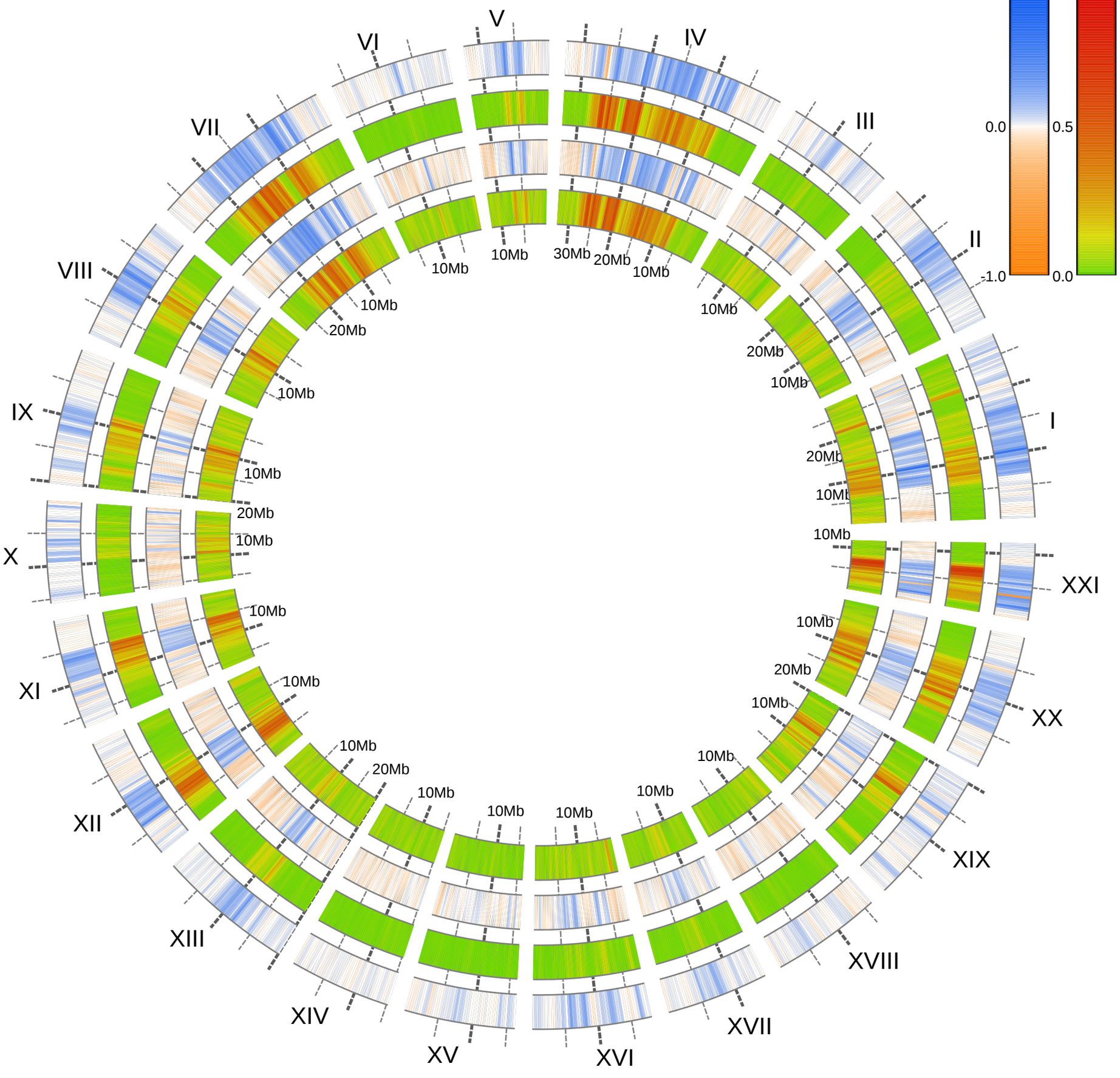


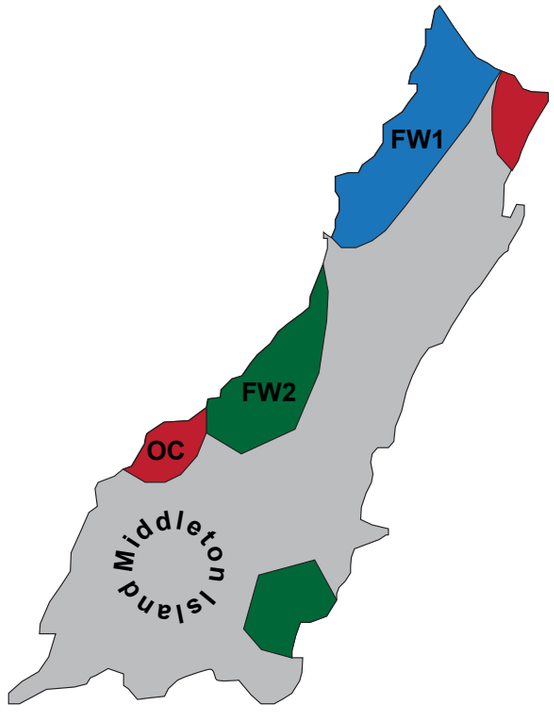




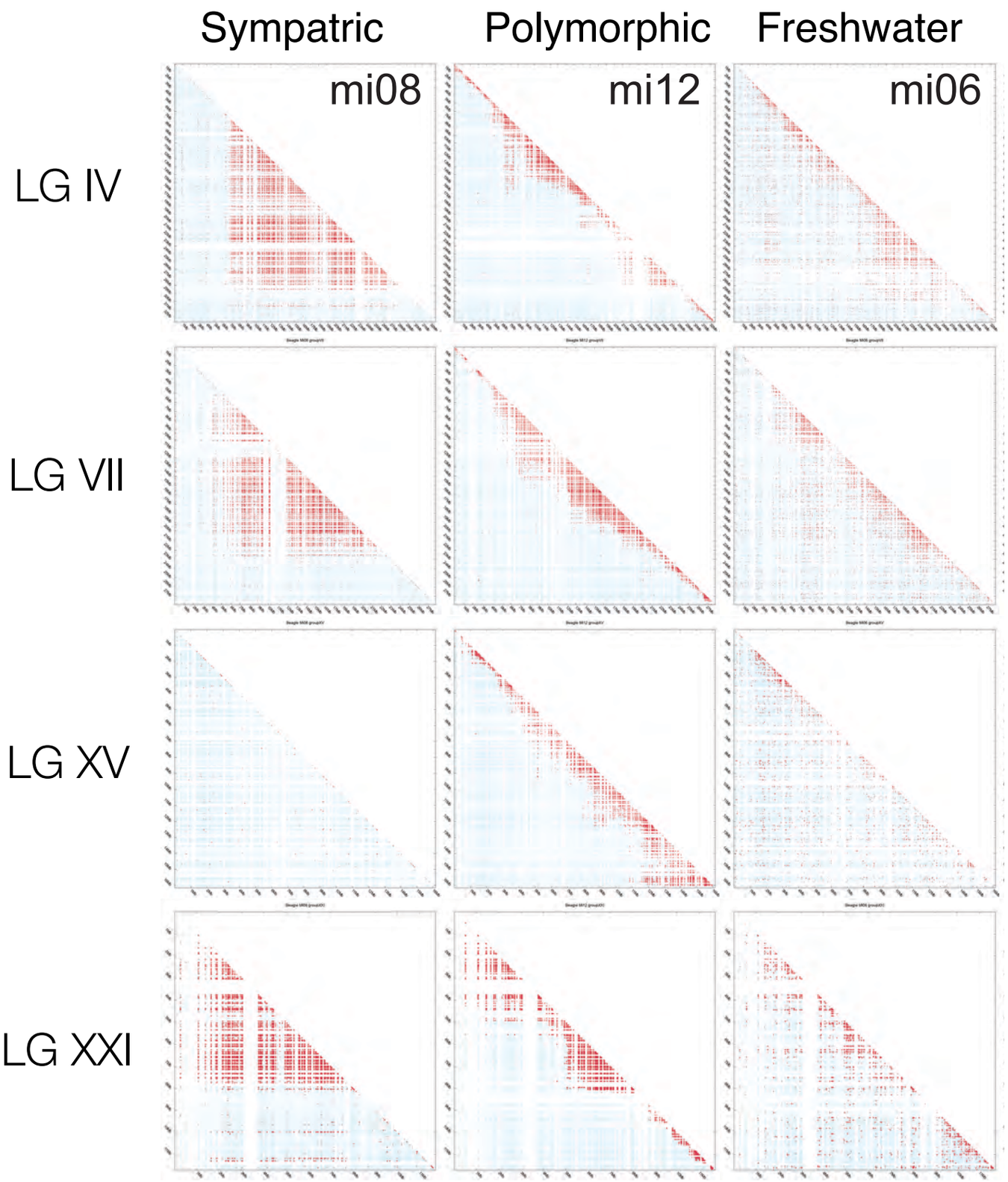
Haplotype diversity increases in freshwater

FW1 vs. Ocean
FW2 vs. Ocean



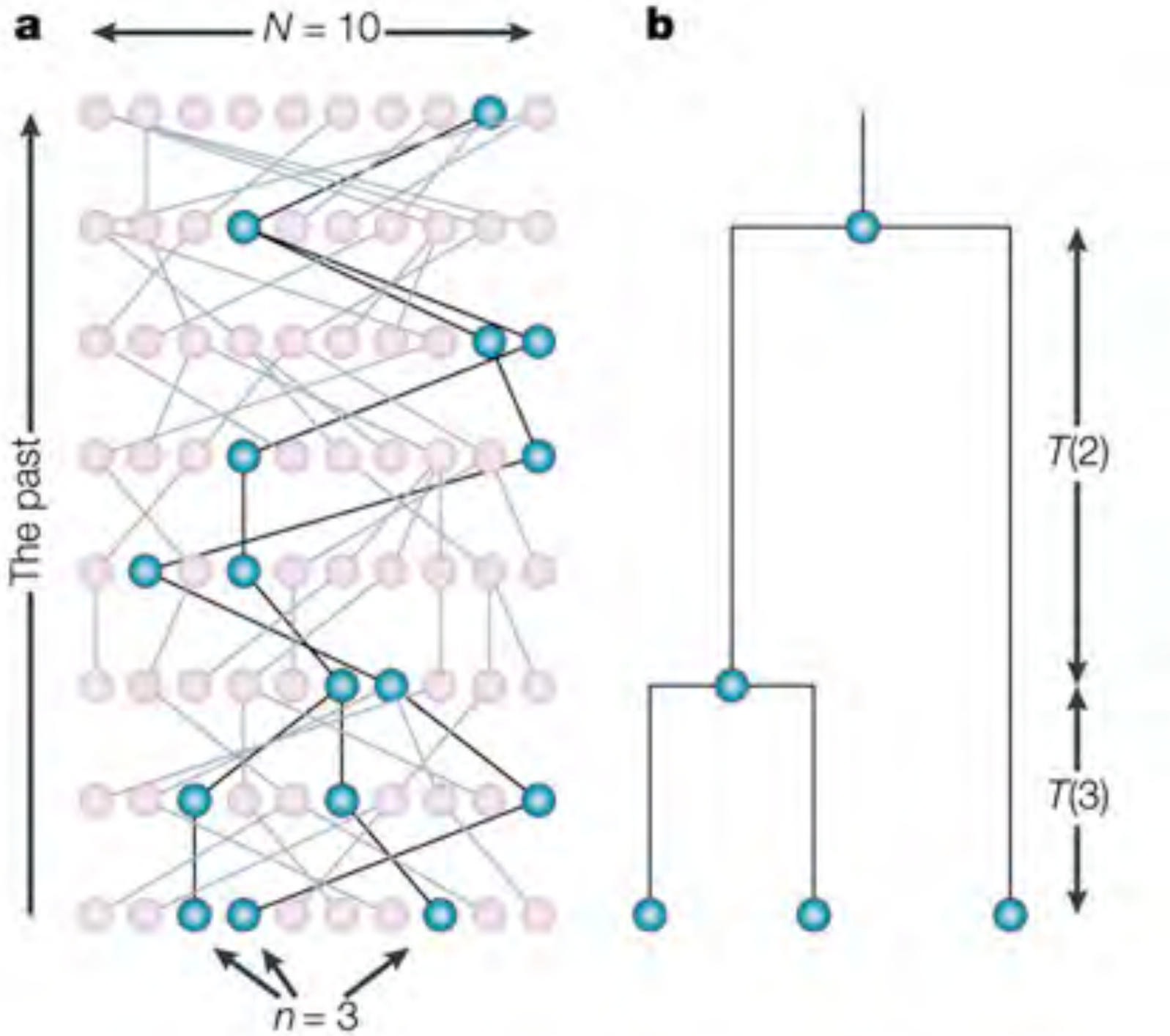


Linkage disequilibrium between oceanic and freshwater stickleback in sympatric sites



Coalescent analyses using RADseq

Coalescent analysis using RAD-seq data

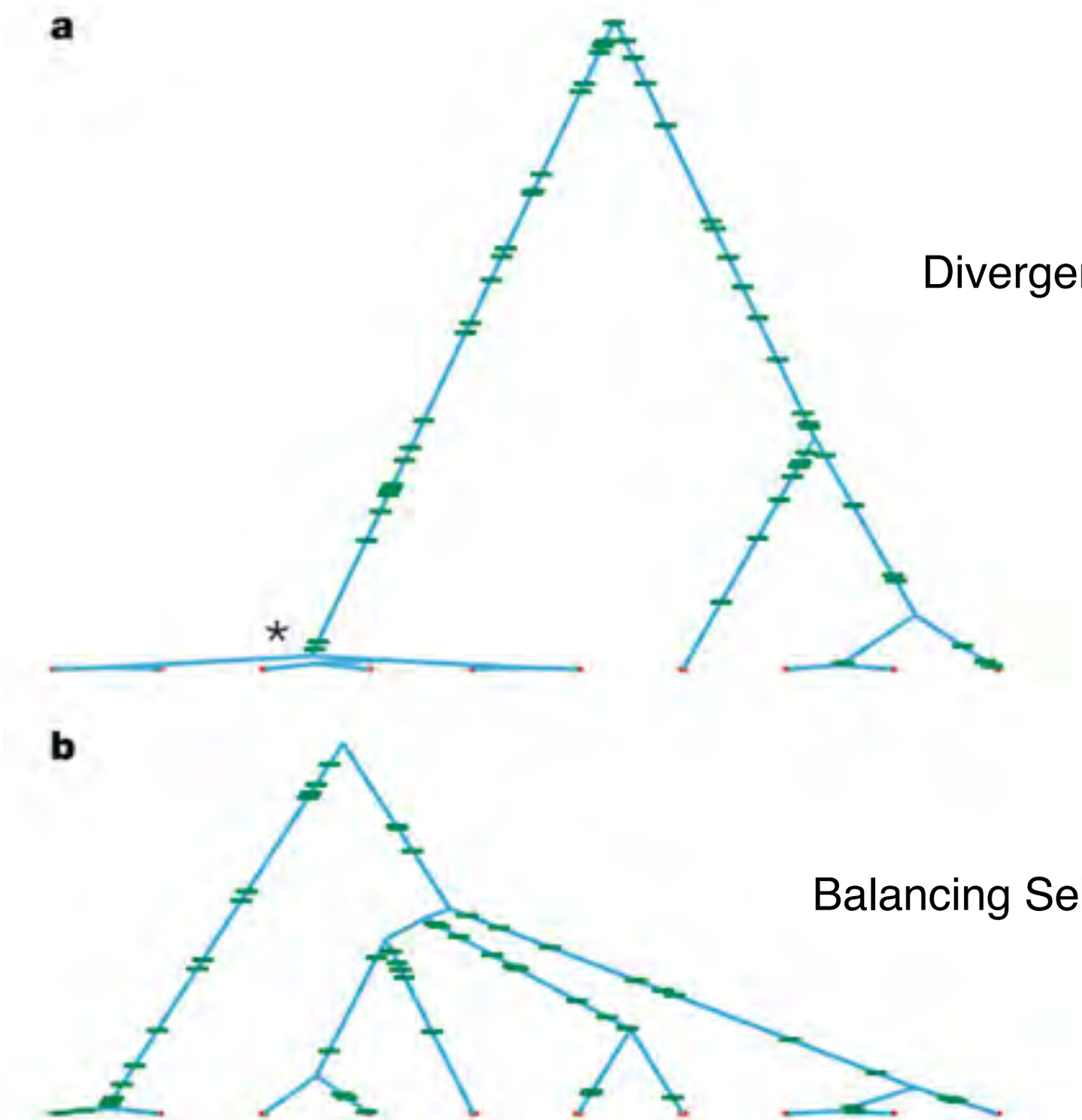


Noah A. Rosenberg & Magnus Nordborg

Nature Reviews | Genetics

Neutral coalescent expectations

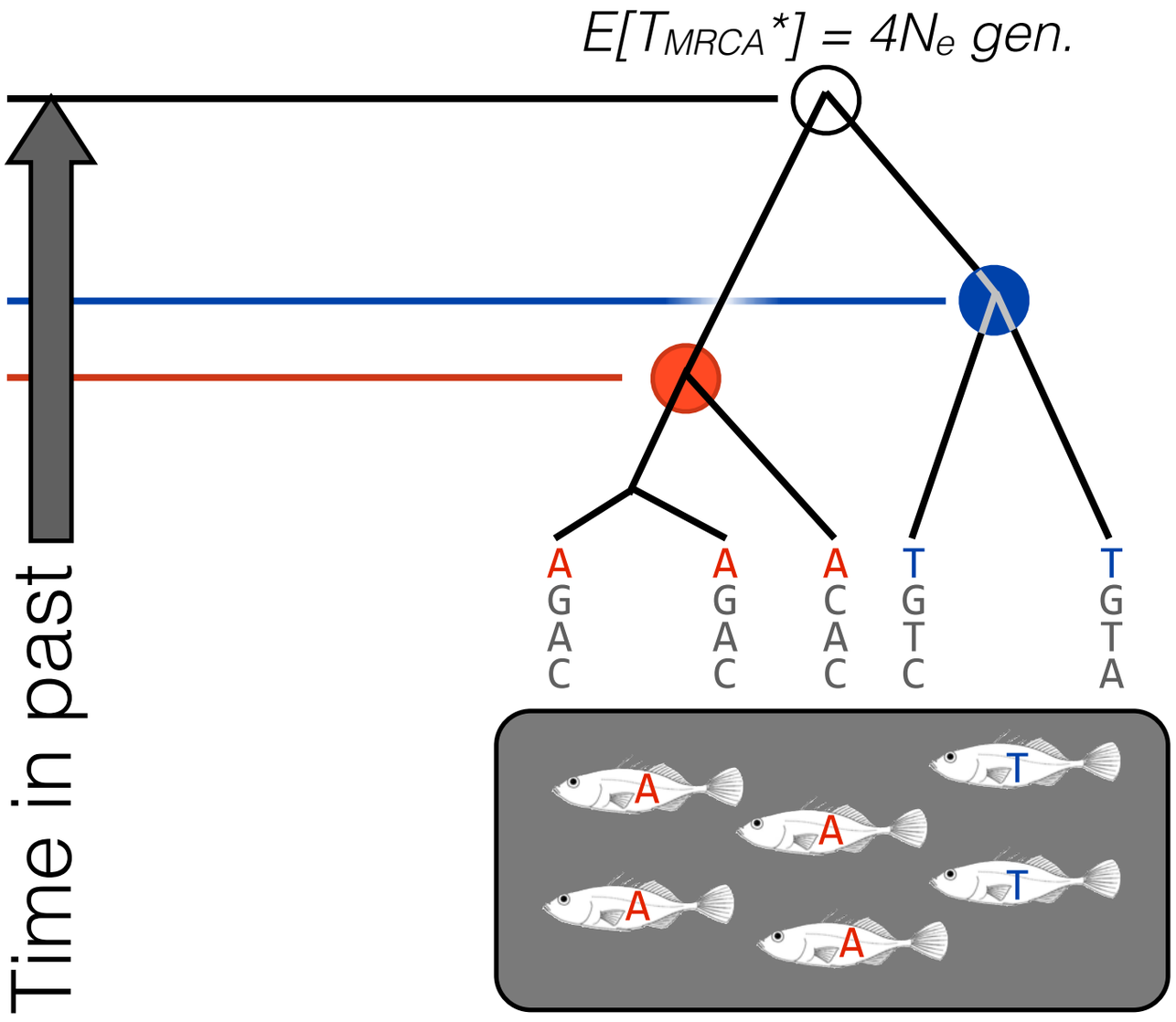
Natural selection and the coalescent



Divergent Selection

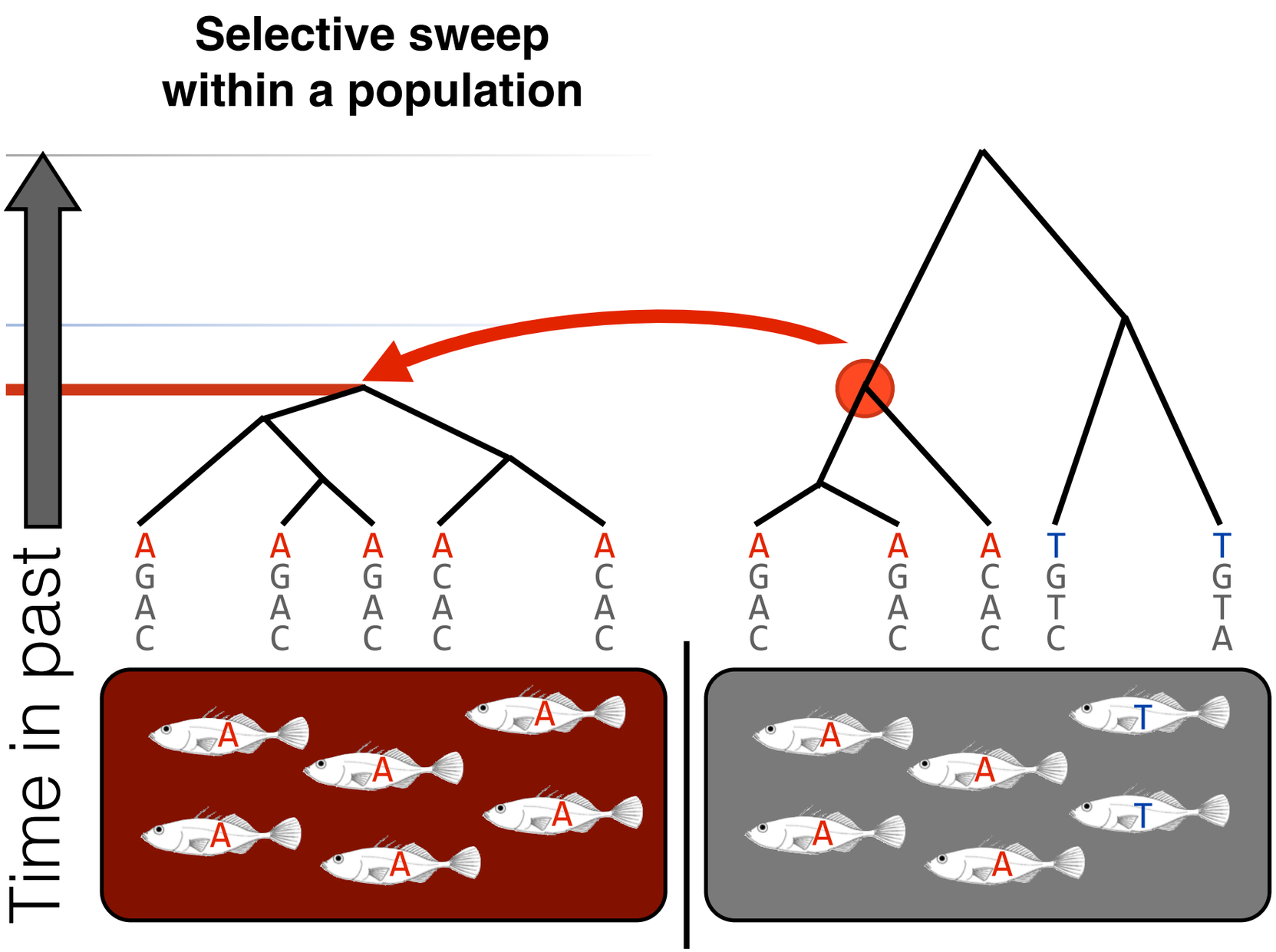
Balancing Selection

Coalescent analyses in stickleback

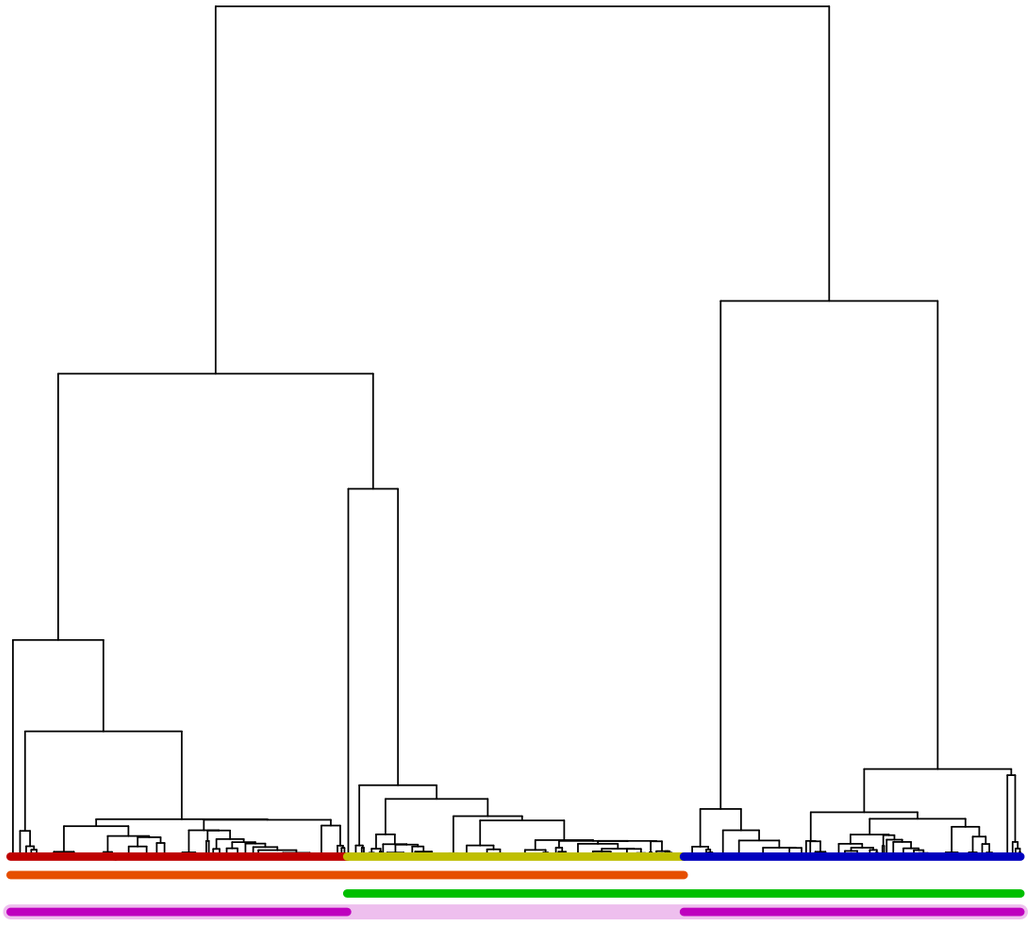


Thom Nelson

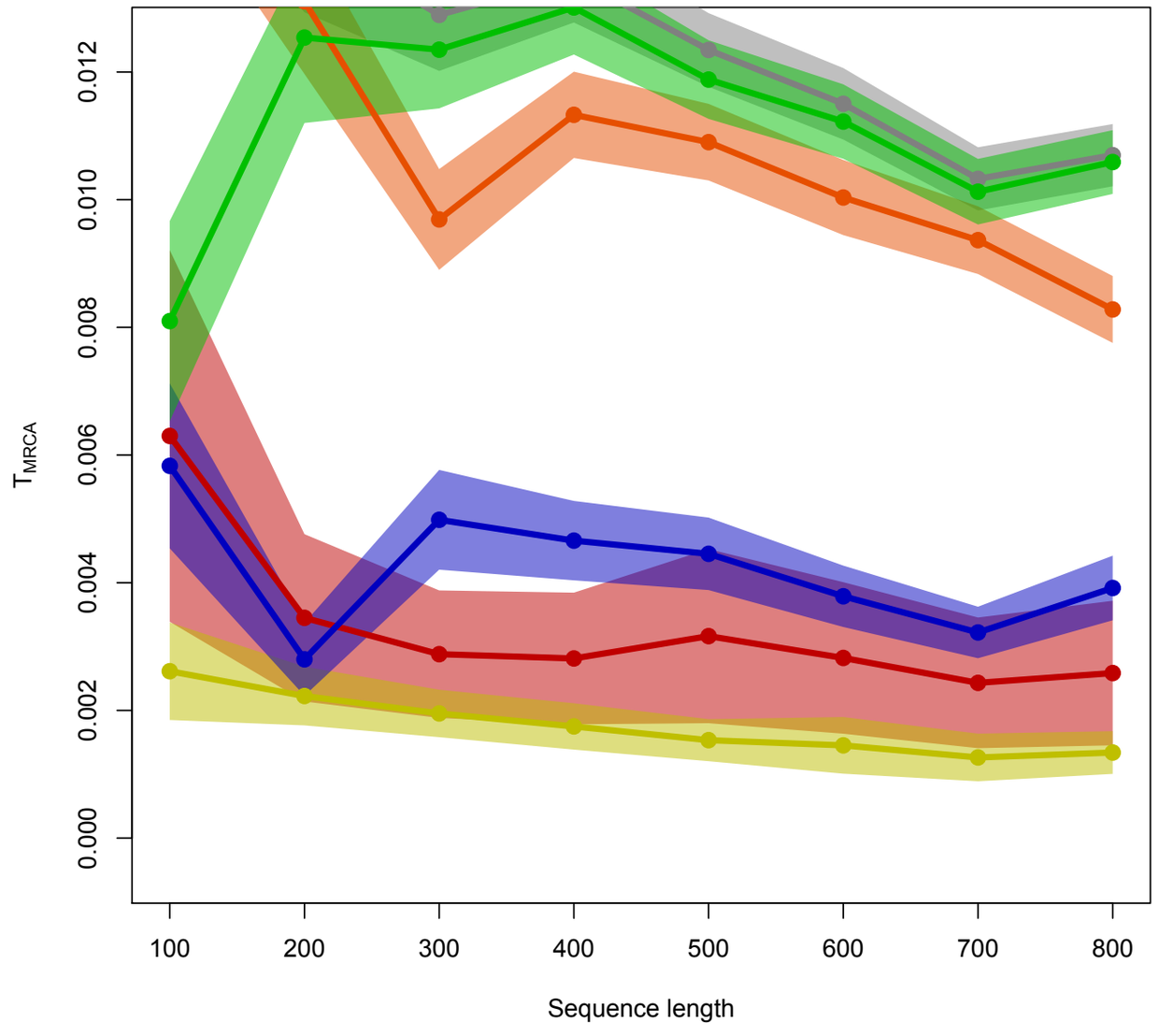
Coalescent analyses in stickleback



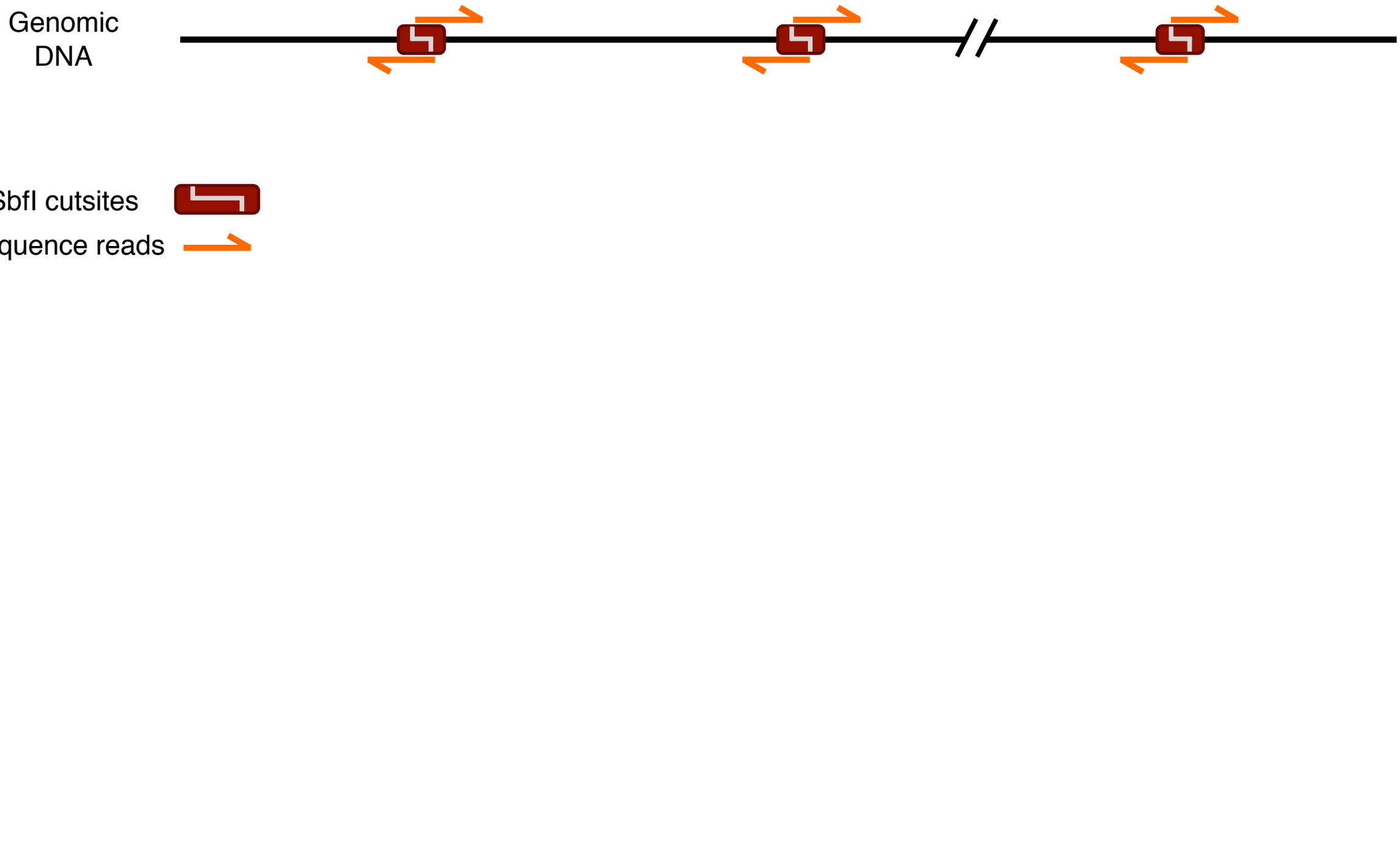
Simulation of sequence length needed for coalescent analyses



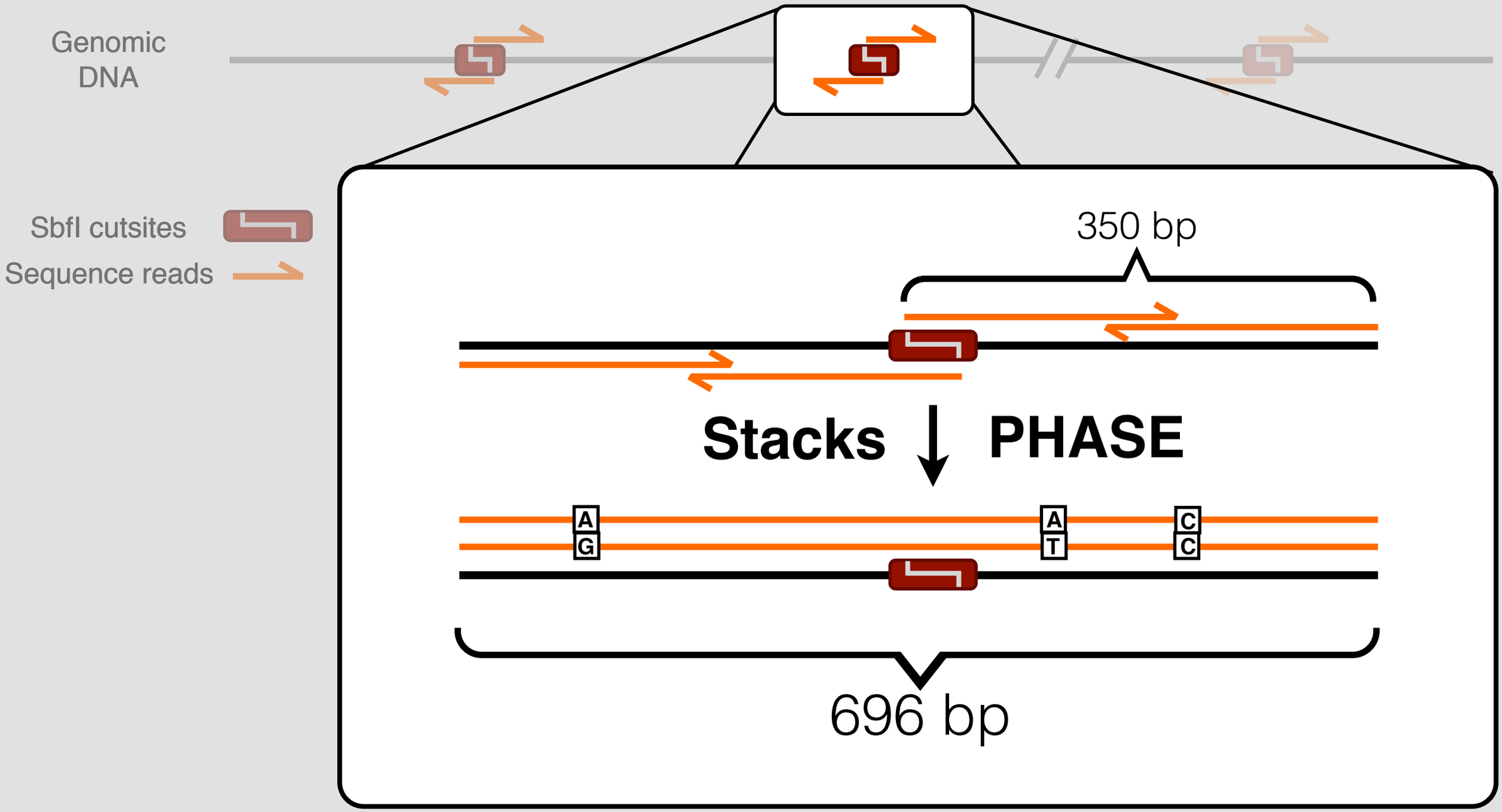
Simulated genealogy



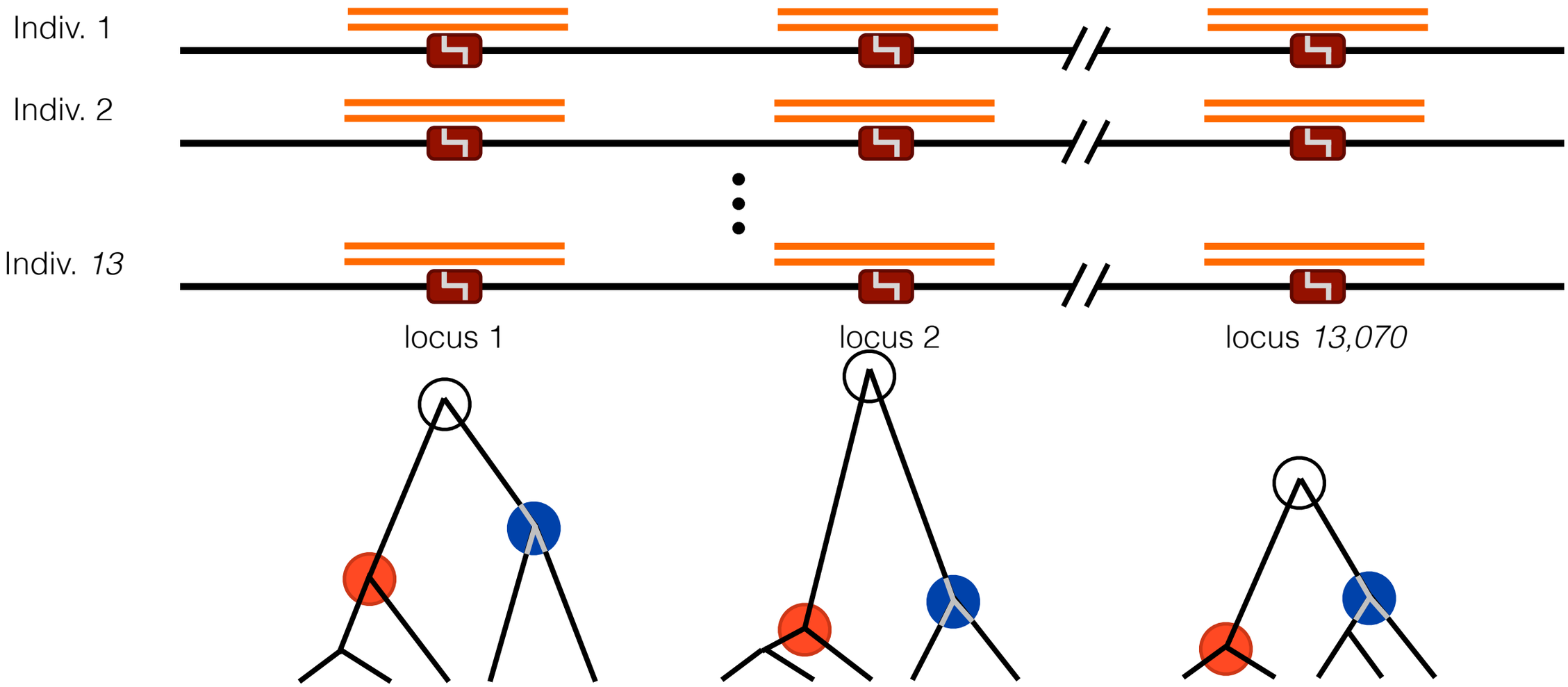
Coalescent analyses with RAD-seq



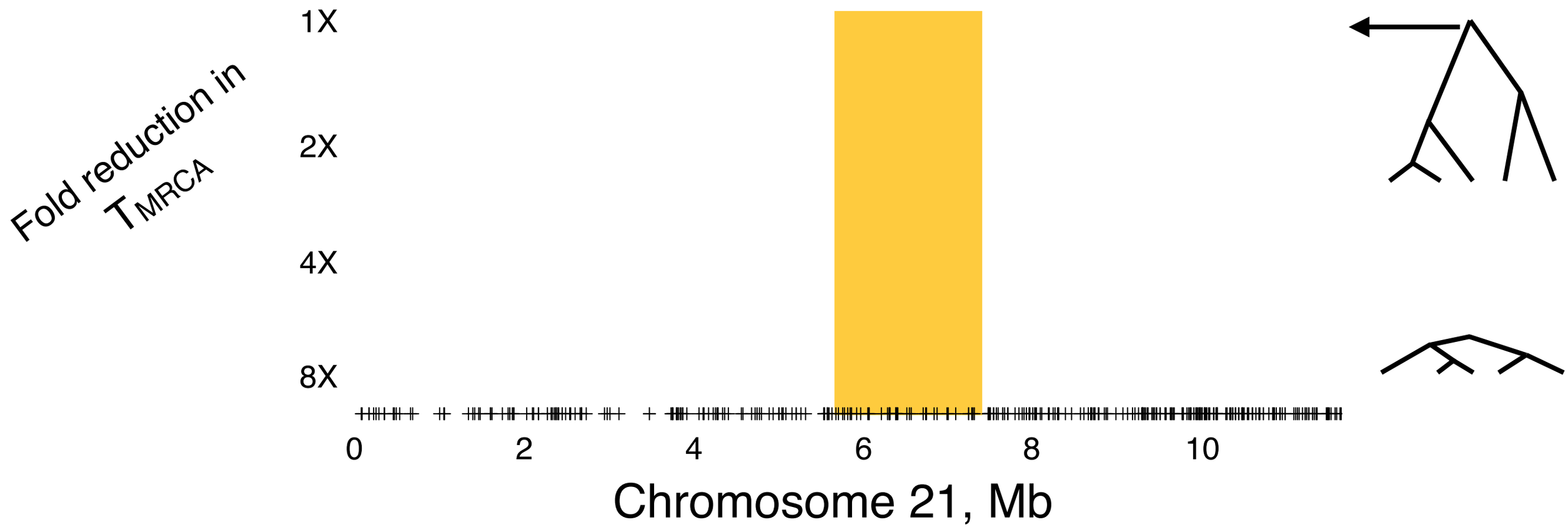
Coalescent analyses with RAD-seq



Coalescent analyses with RAD-seq

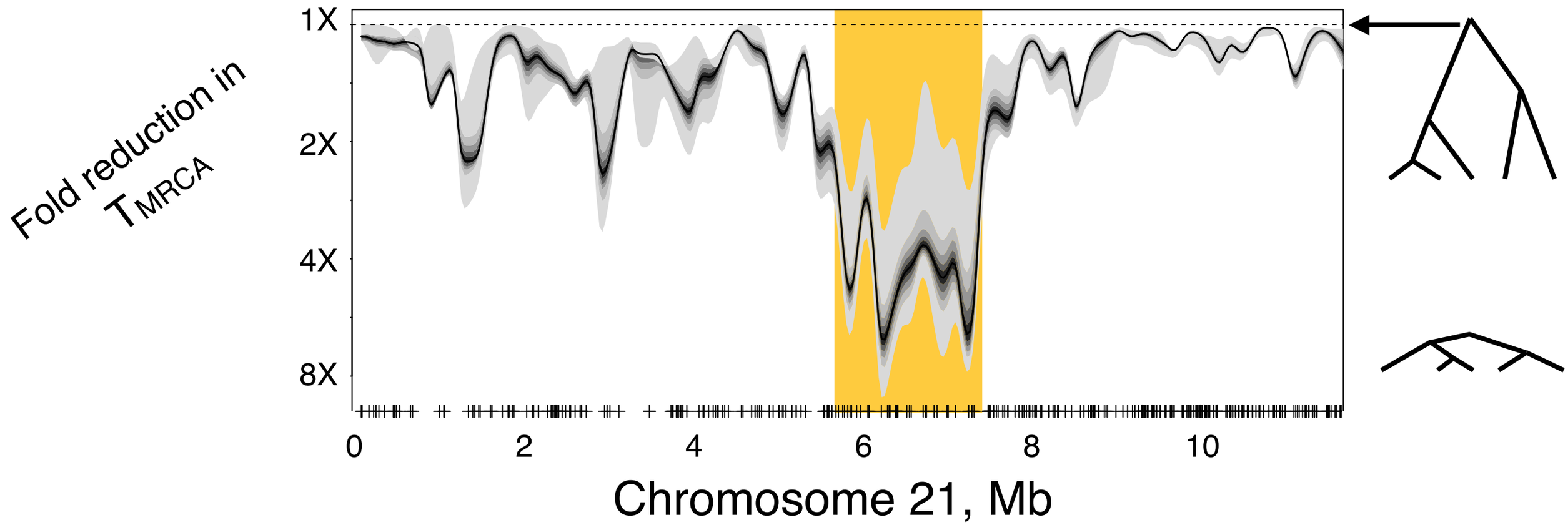


Selection in one population reduces coalescence time



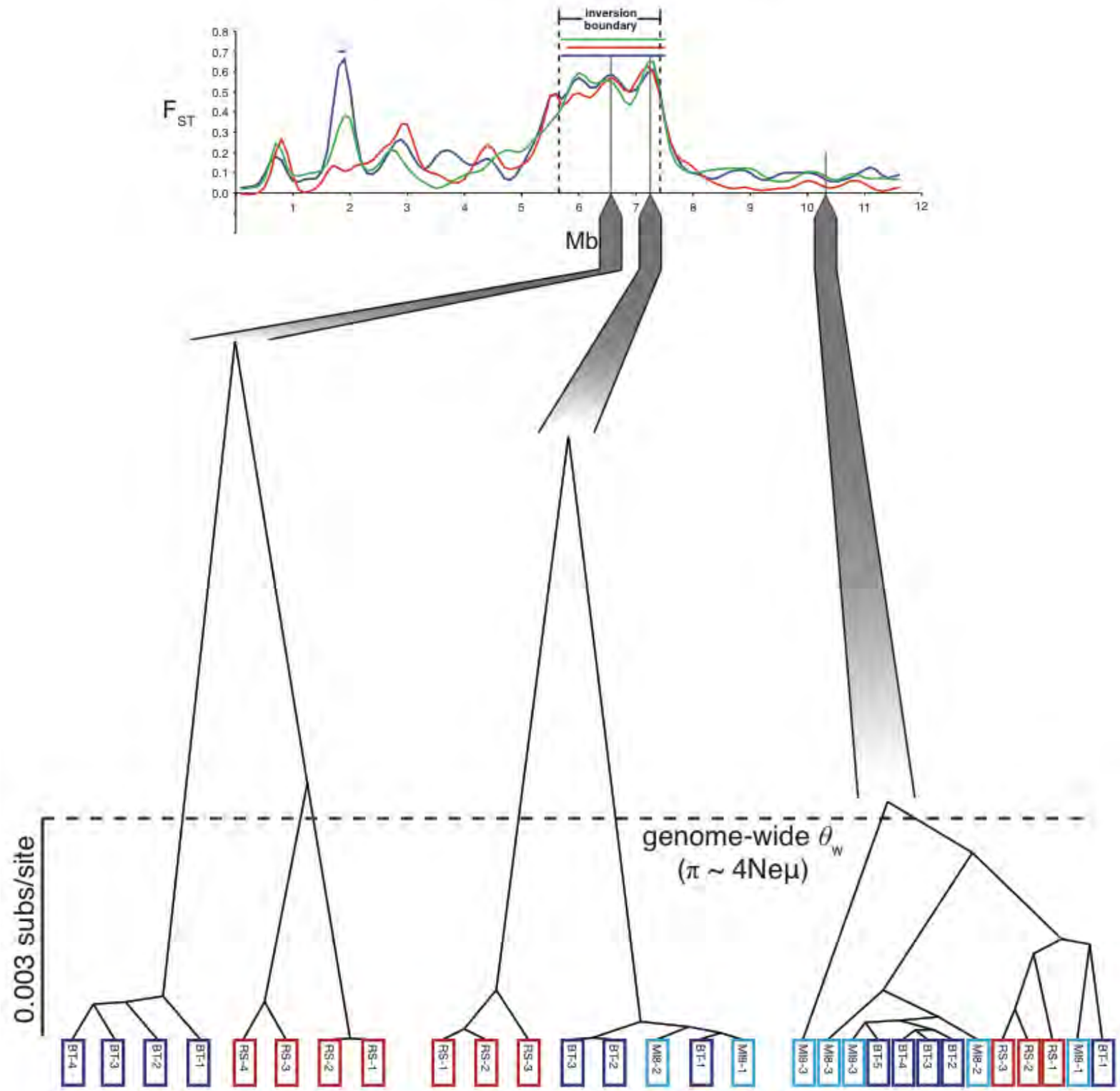
$$\log_2\left(\frac{T_{MRCA}\mathit{FW}}{T_{MRCA}\mathit{ALL}}\right)$$

Selection in one population reduces coalescence time

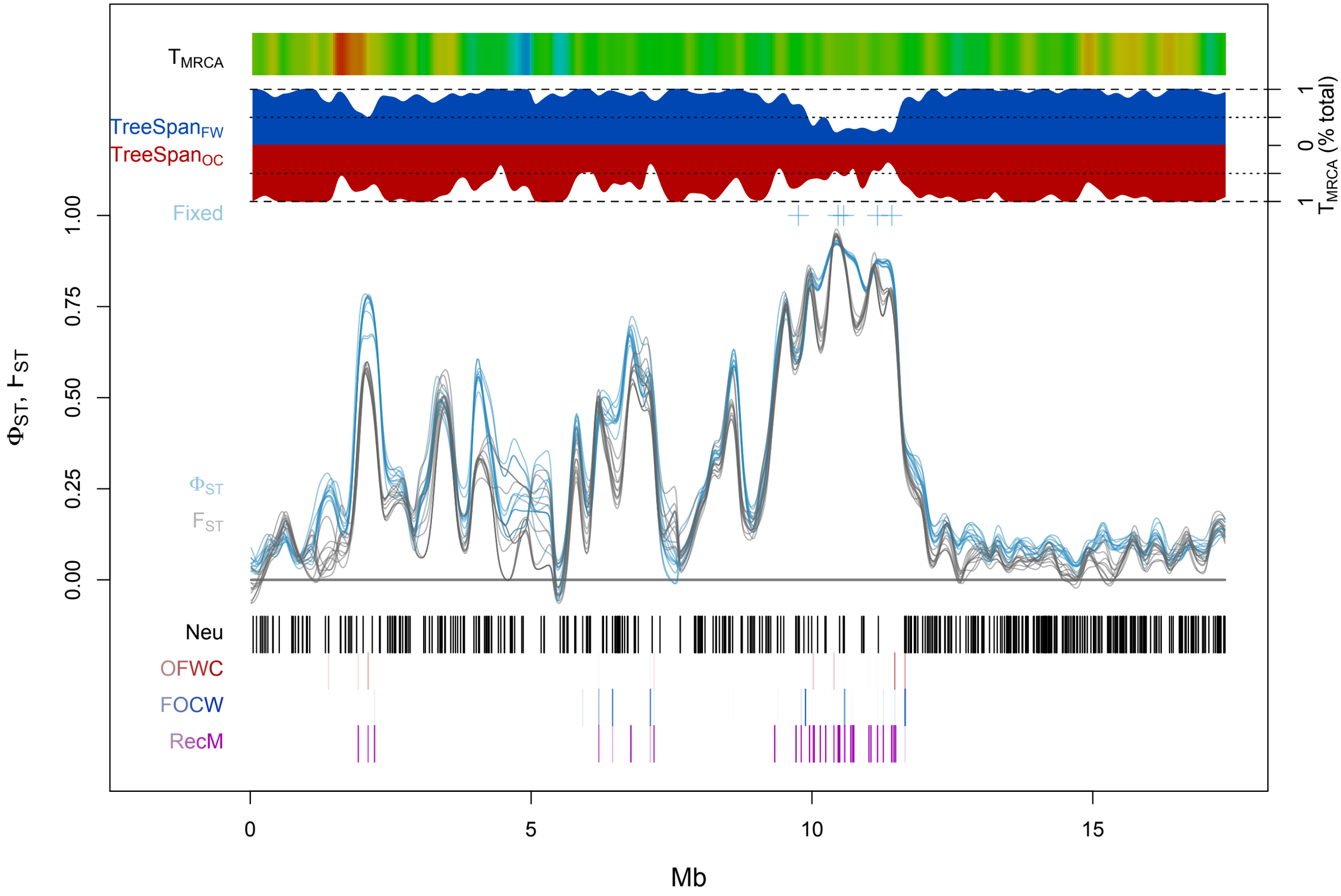


$$\log_2\left(\frac{T_{MRCA}\mathit{FW}}{T_{MRCA}\mathit{ALL}}\right)$$

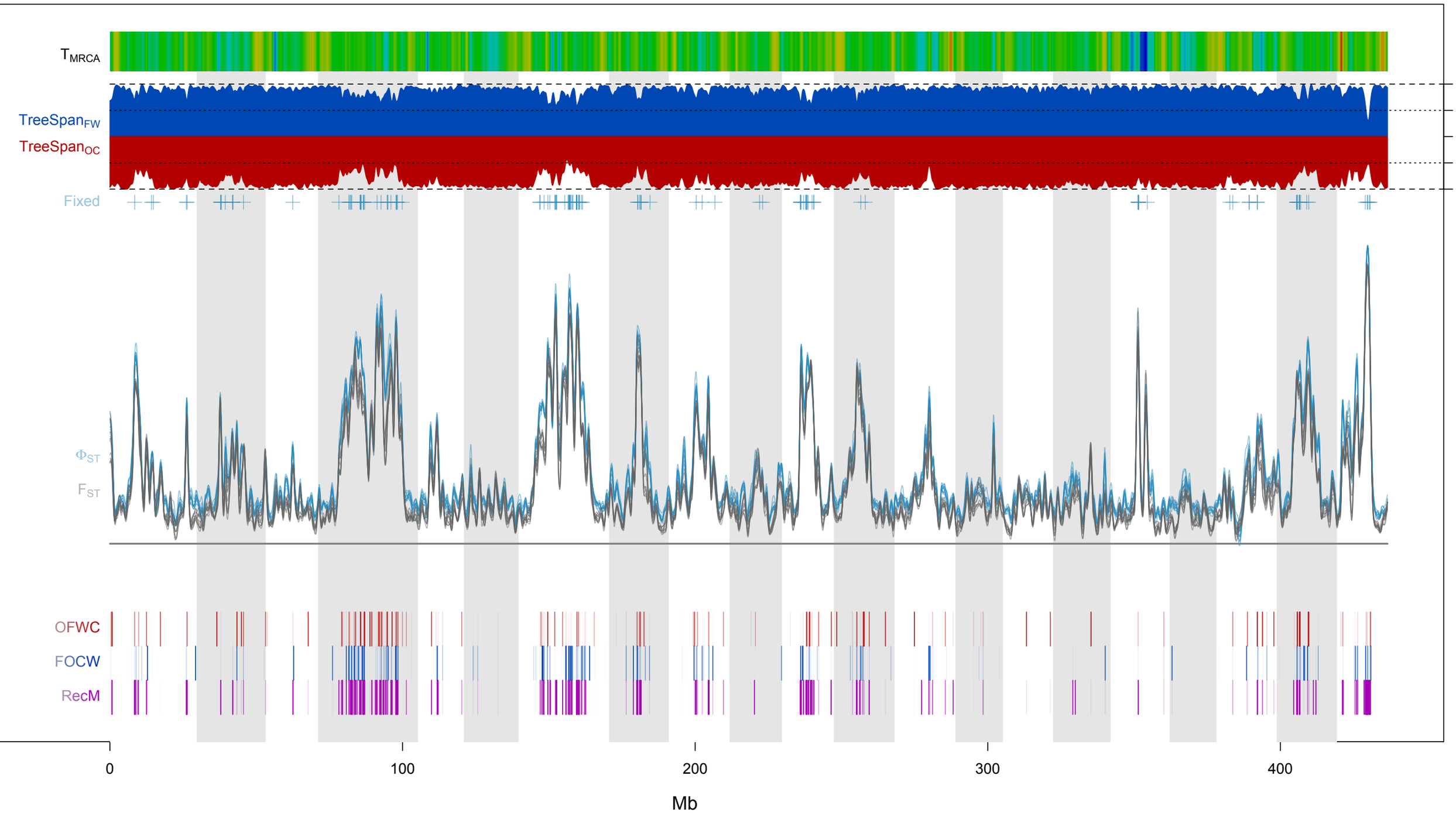
However, across populations the coalescence time can increases significantly in a genomic region



Relative proportion of T_{mrca} across different habitats LG21



Relative proportion of T_{mrca} across different habitats
Entire genome

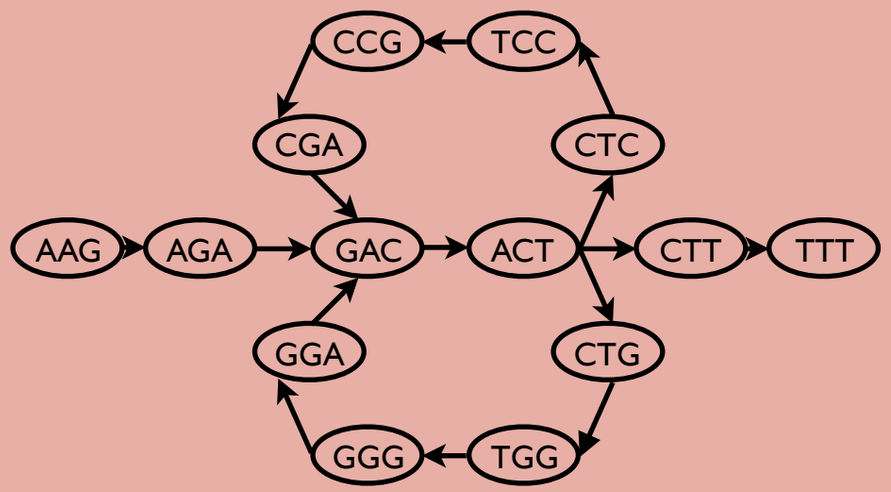


What can explain such rapid evolution and haplotype structure?

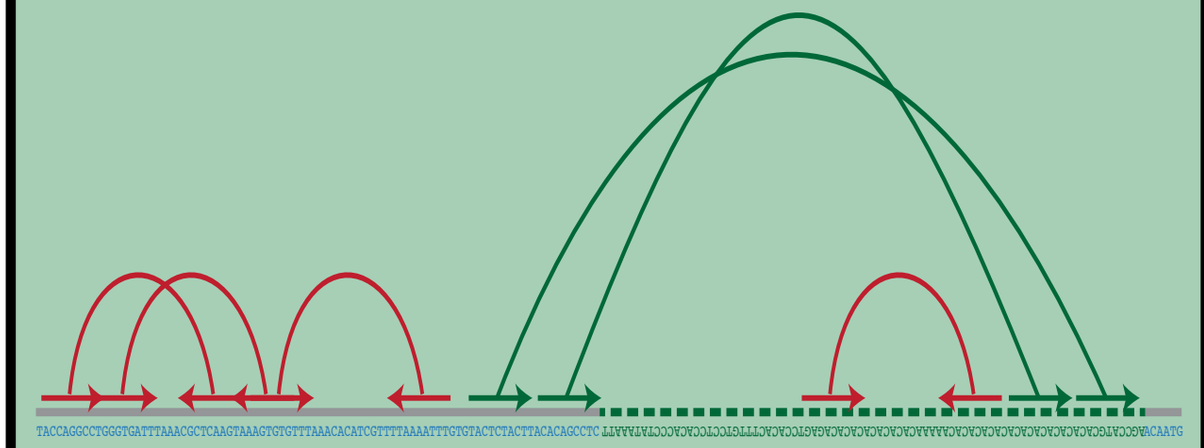
Is the stickleback genome architecture partly responsible?

Catchen, Bassham et al. *in prep*

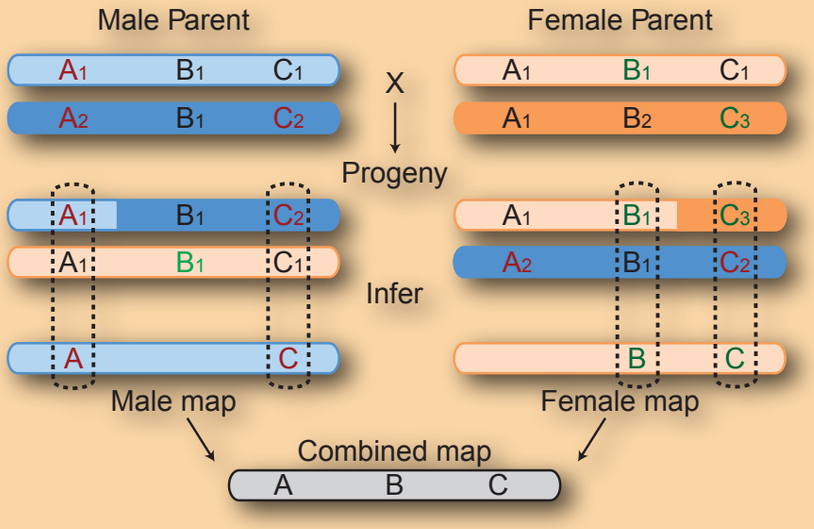
Genome Assembly



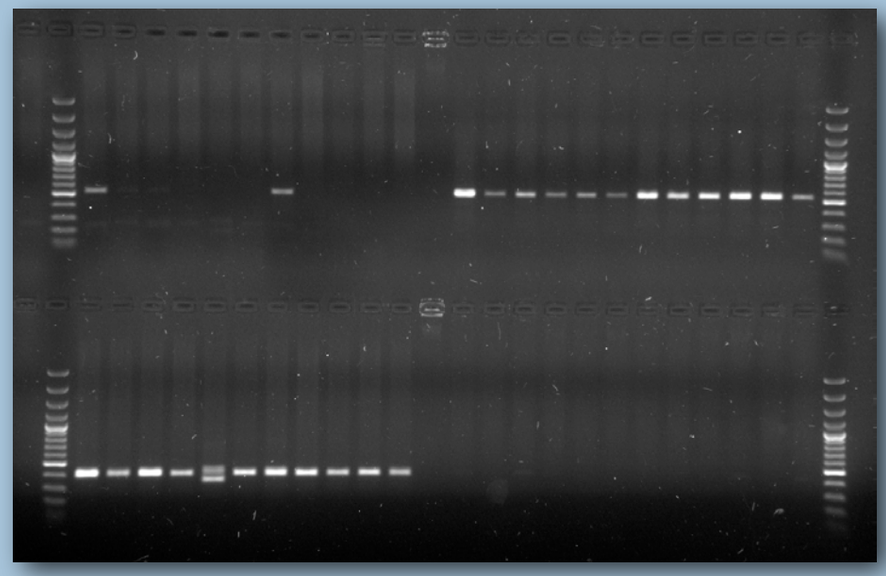
Paired-end Alignments



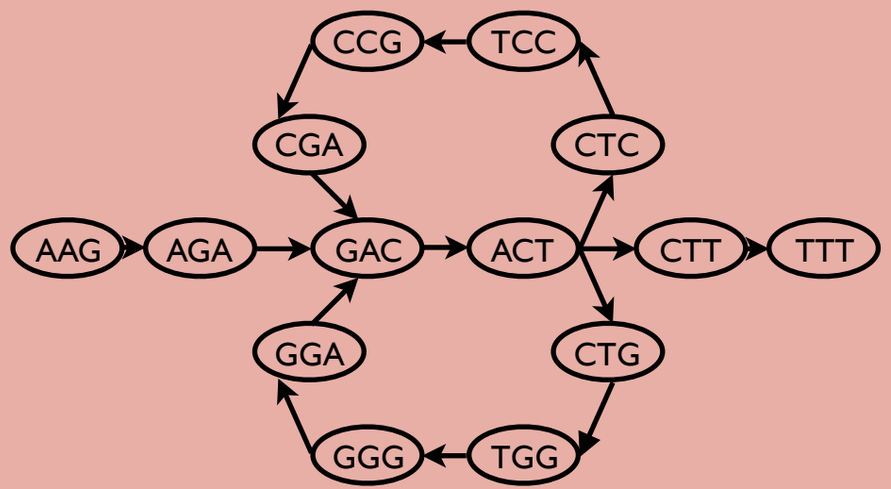
Genetic Map Construction



PCR Screening of Breakpoints

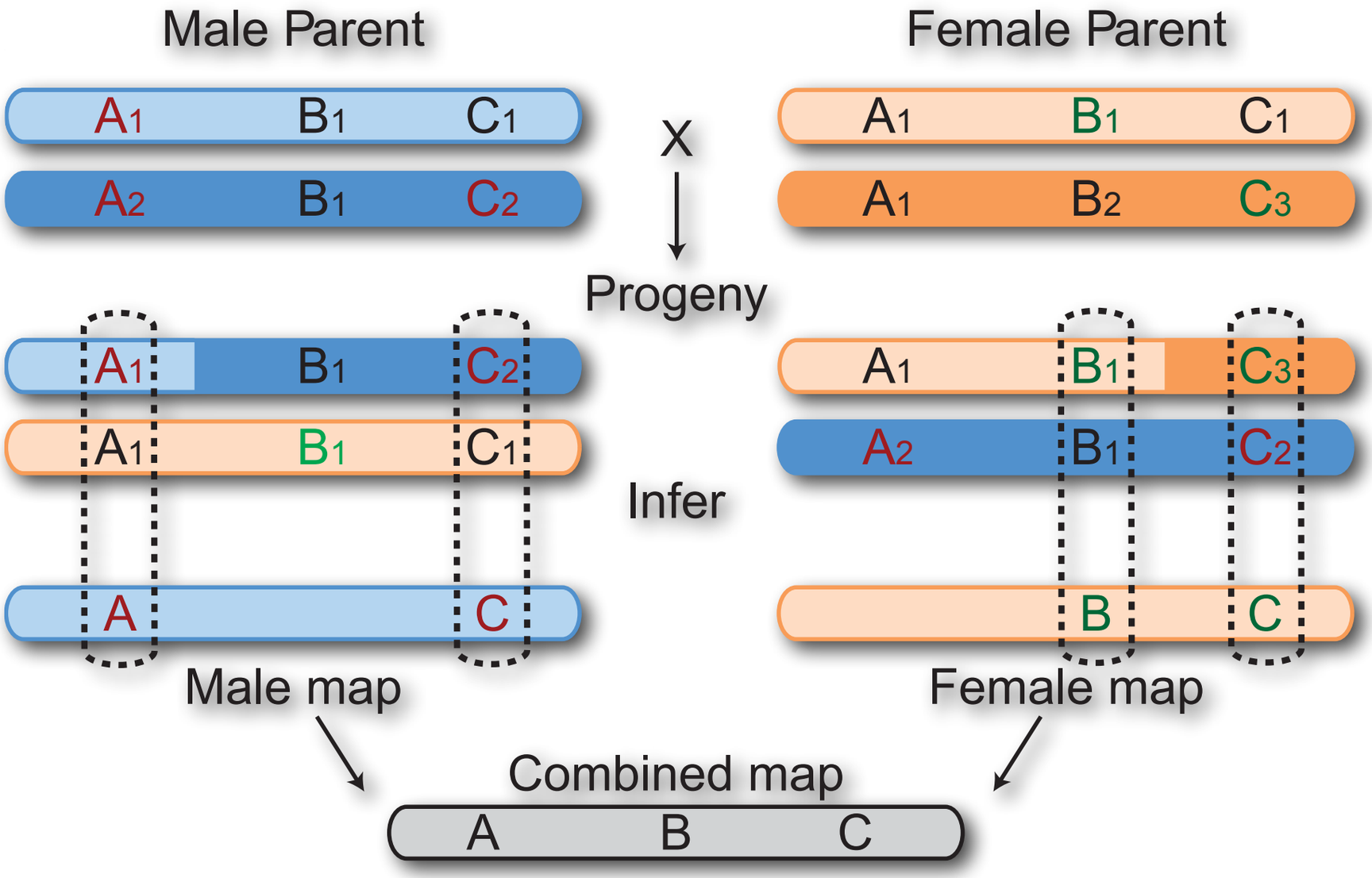


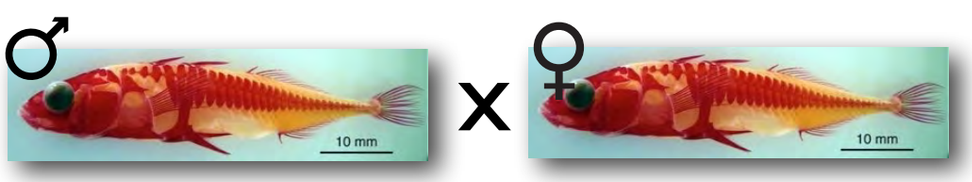
Genome Assembly

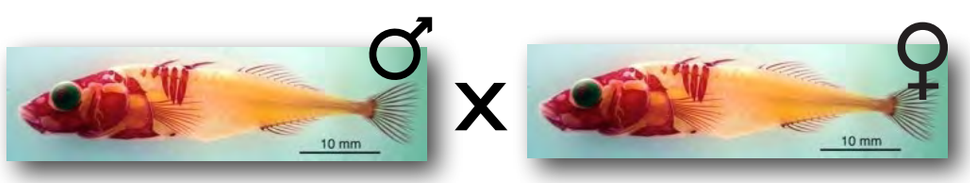


	♂	♀	♂	♀
N50	17,417 bp	18,982 bp	15,555 bp	15,534 bp
Max	199,905 bp	192,283 bp	238,768 bp	254,734 bp
Total	488.8 Mb	472.5 Mb	456.4 Mb	473.4 Mb
Median Coverage	24.6x	26.5x	24.1x	25.8x

F1 Pseudo-testcross using RAD-seq







93 progeny
66,071 loci
5,351 markers

93 progeny
45,301 loci
3,927 markers

Aligned
Opposite
Inward
5.00Mb
1.00Mb

LGXI

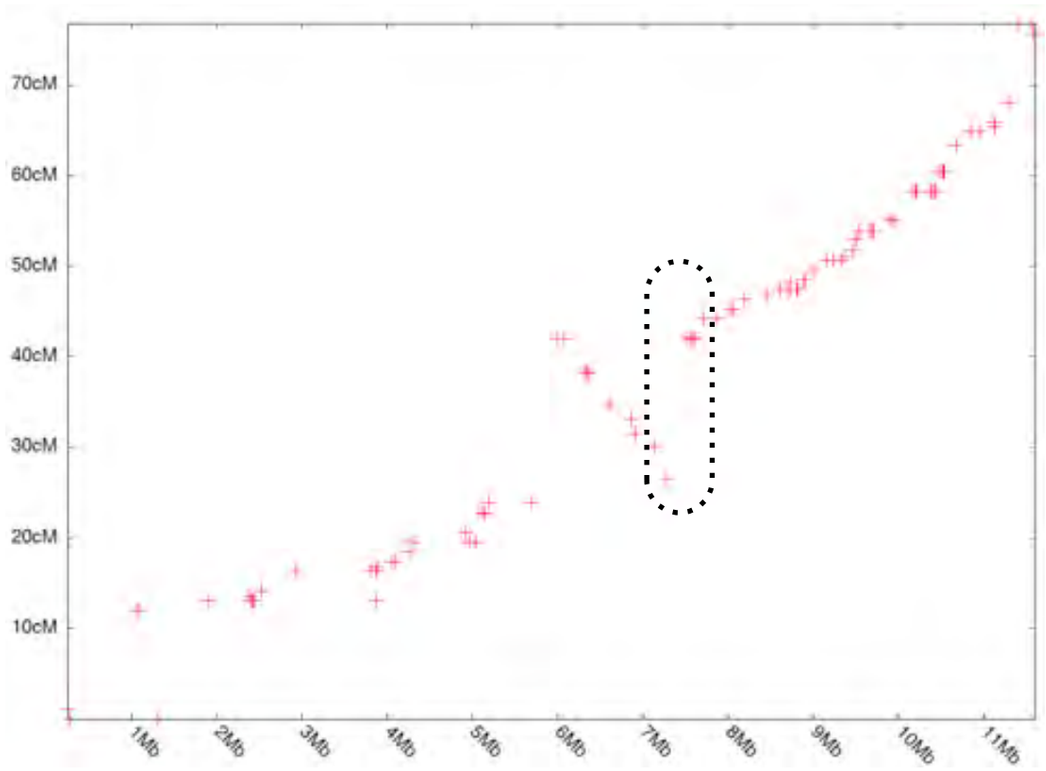
70cM
60cM
50cM
40cM
30cM
20cM
10cM

1Mb
2Mb
3Mb
4Mb
5Mb
6Mb
7Mb
8Mb
9Mb
10Mb
11Mb

F_{st}

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2

0
5000
10000

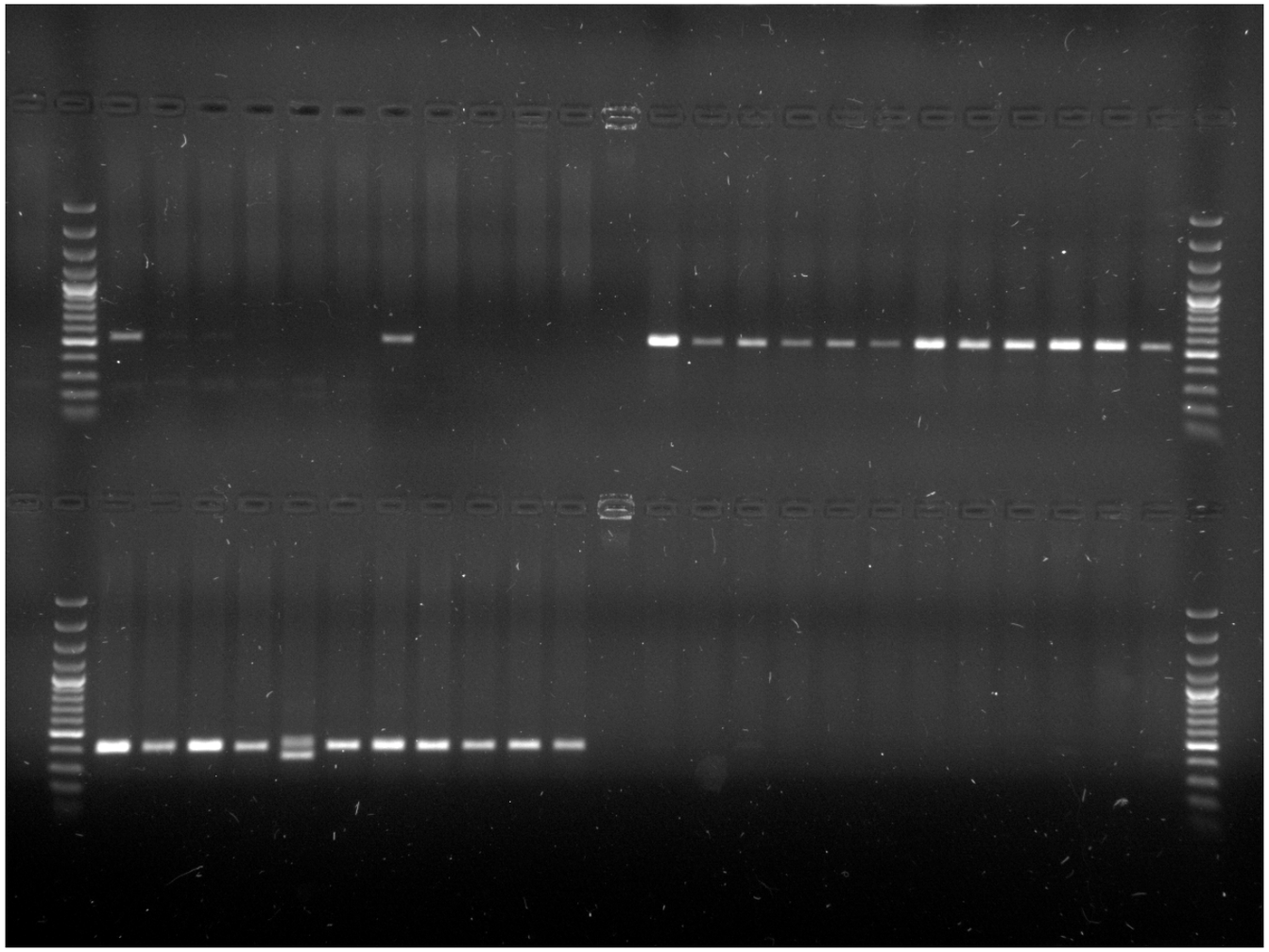


Linkage Group XXI

RS (Marine) Boot (Freshwater)

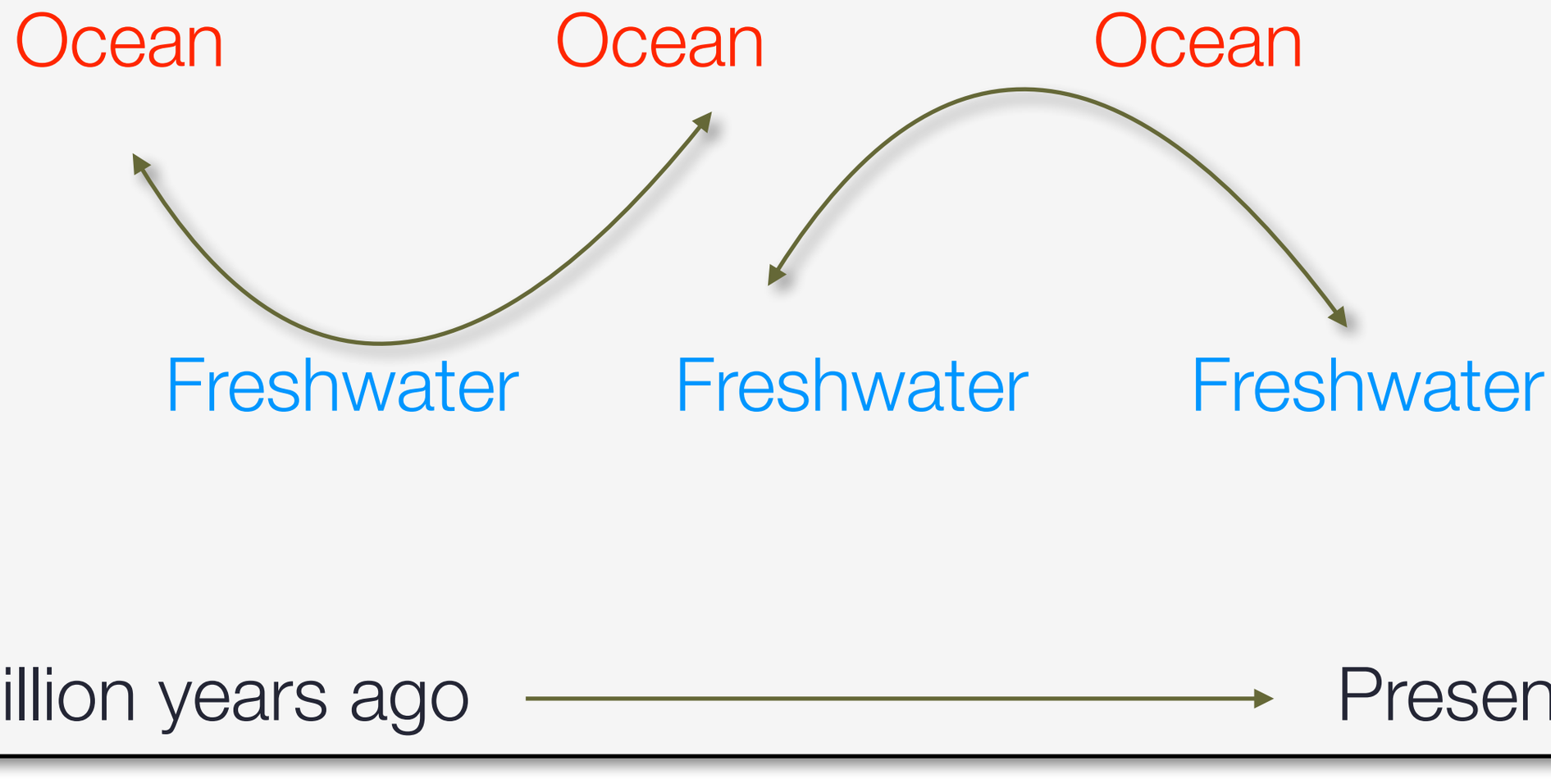
Genome Arrangement

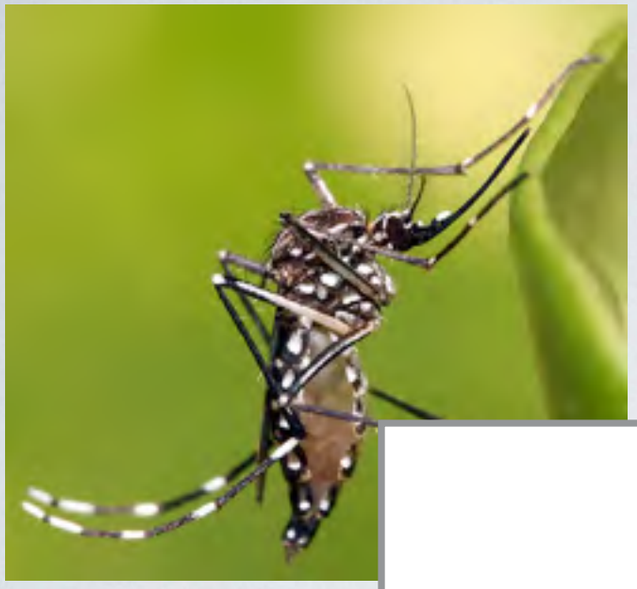
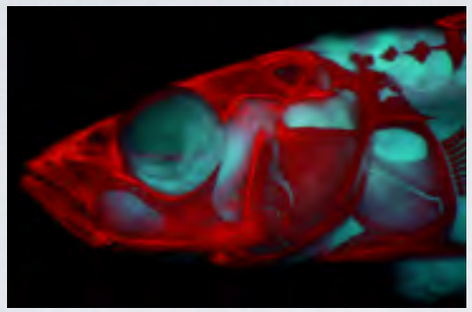
Inverted

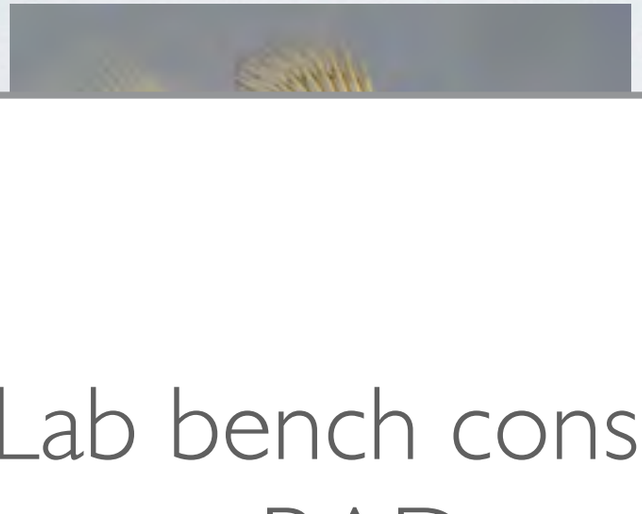


1. Previous work has shown that the freshwater genomes evolve in 13,000 years.
2. These new Middleton Island data shows that the phenotype can appear in as little as *50 years*.
3. Much of the divergence involves soft sweeps.
4. This could represent thousands of haplotypes reassembling, but the genome appears *chunkier*.
5. Many haplotypes are habitat specific and are quite ancient.
6. Haplotypes often coincide with structural variation across the stickleback genome

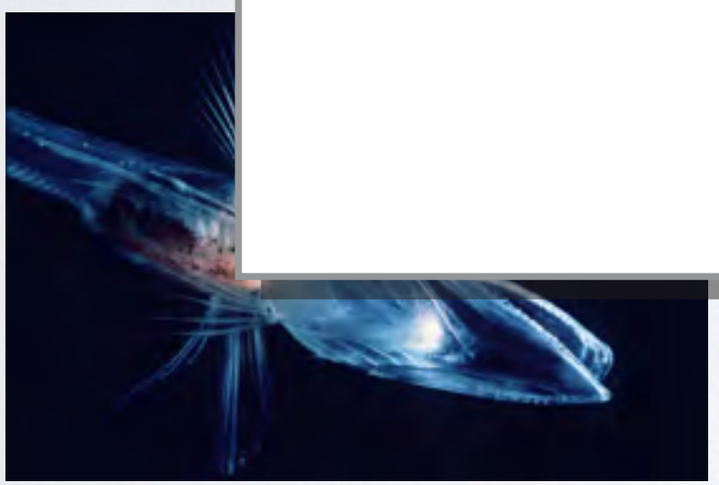
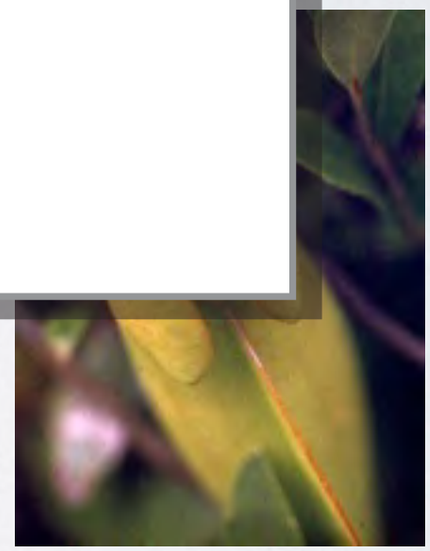
Hypothesis: Old genomic architecture variation is a product of the metapopulation structure of stickleback, and this architecture strongly influences subsequent rapid evolution.



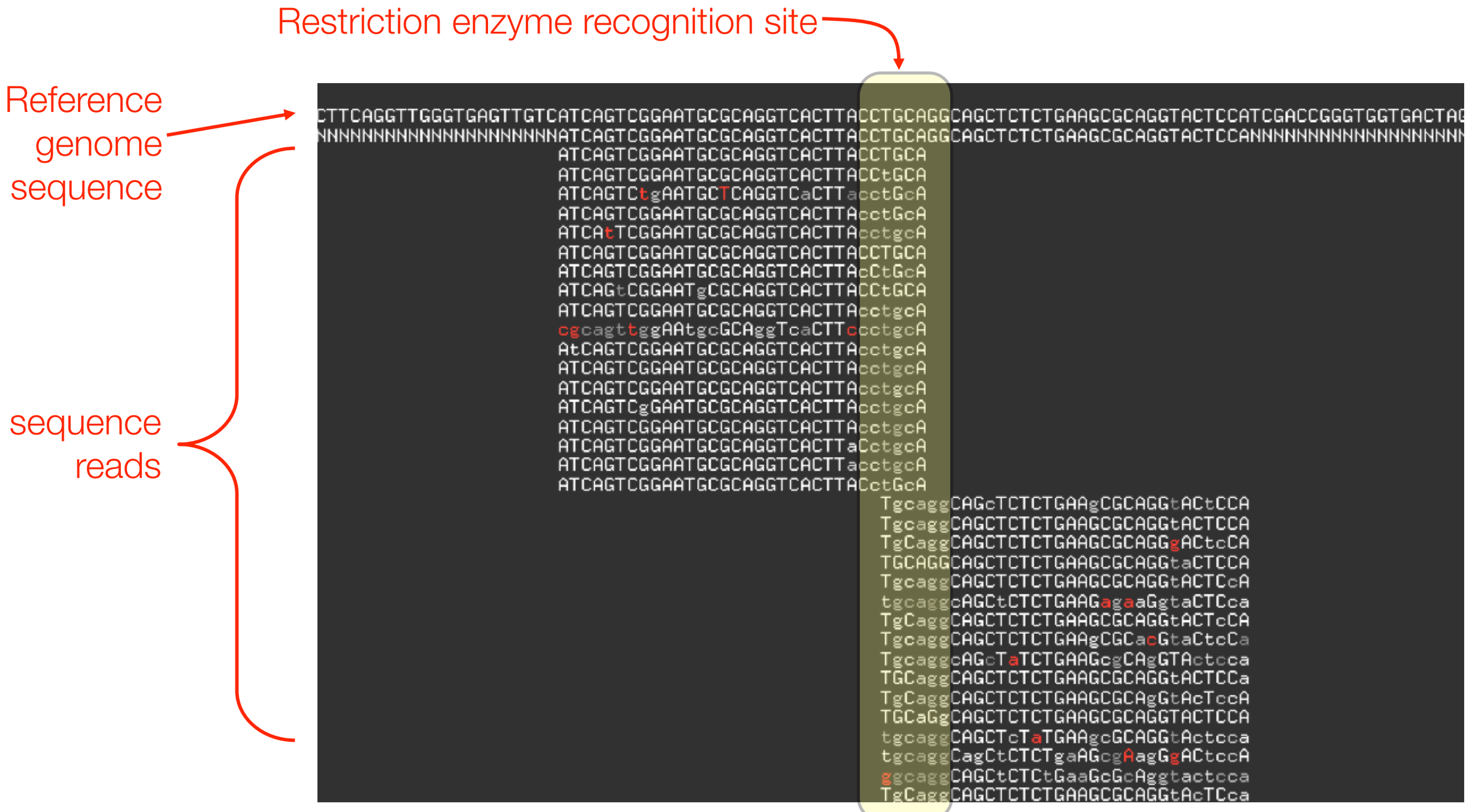



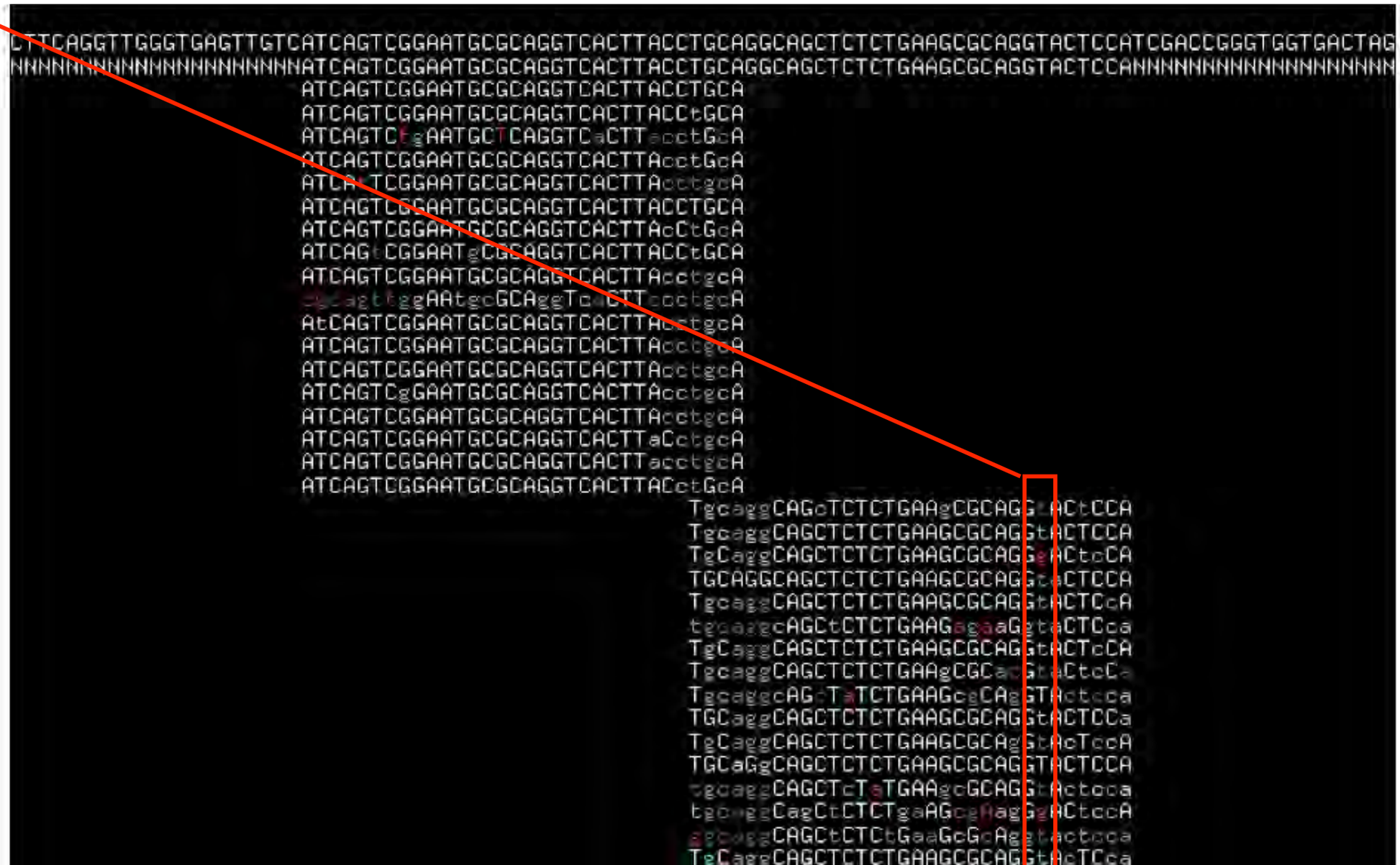
Lab bench considerations for RADseq studies

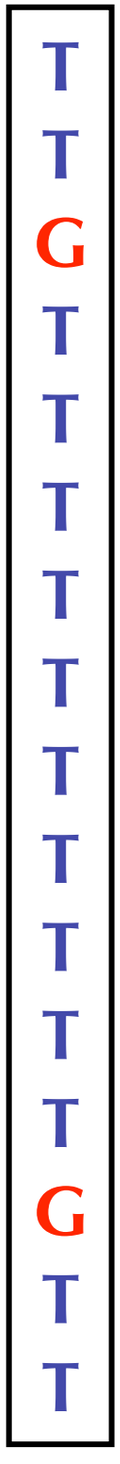




Statistical considerations in RAD-seq



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





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

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

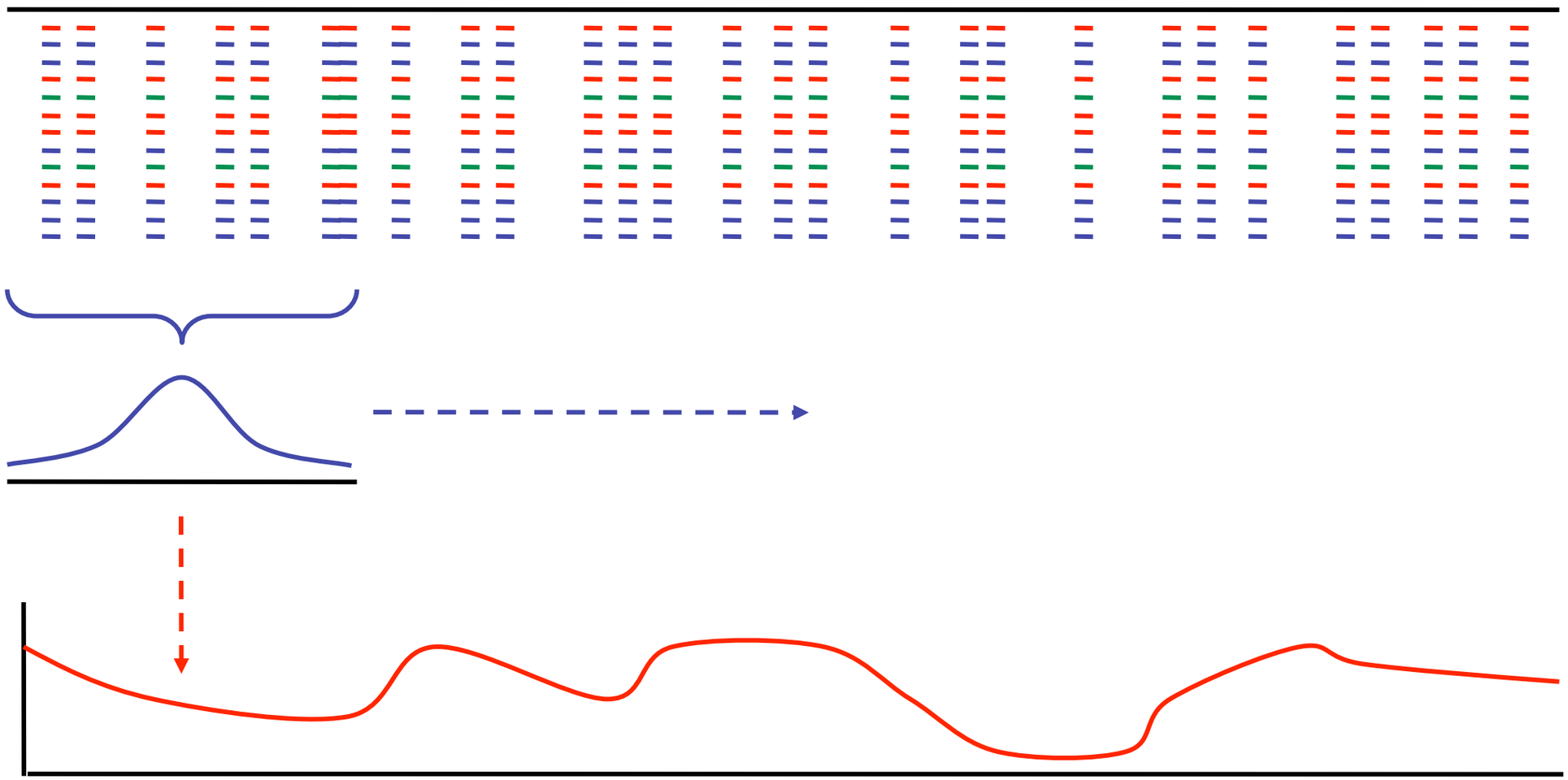
$$L(n_1 \text{hom}) = P(n_1,n_2,n_3,n_4) = \frac{n!}{n_1!n_2!n_3!n_4!}\left(1-\frac{3\varepsilon}{4}\right)^{n_1}\left(\frac{\varepsilon}{4}\right)^{n_2}\left(\frac{\varepsilon}{4}\right)^{n_3}\left(\frac{\varepsilon}{4}\right)^{n_4}$$

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

Maximum likelihood genotyping based on multinomial distribution of nucleotide reads

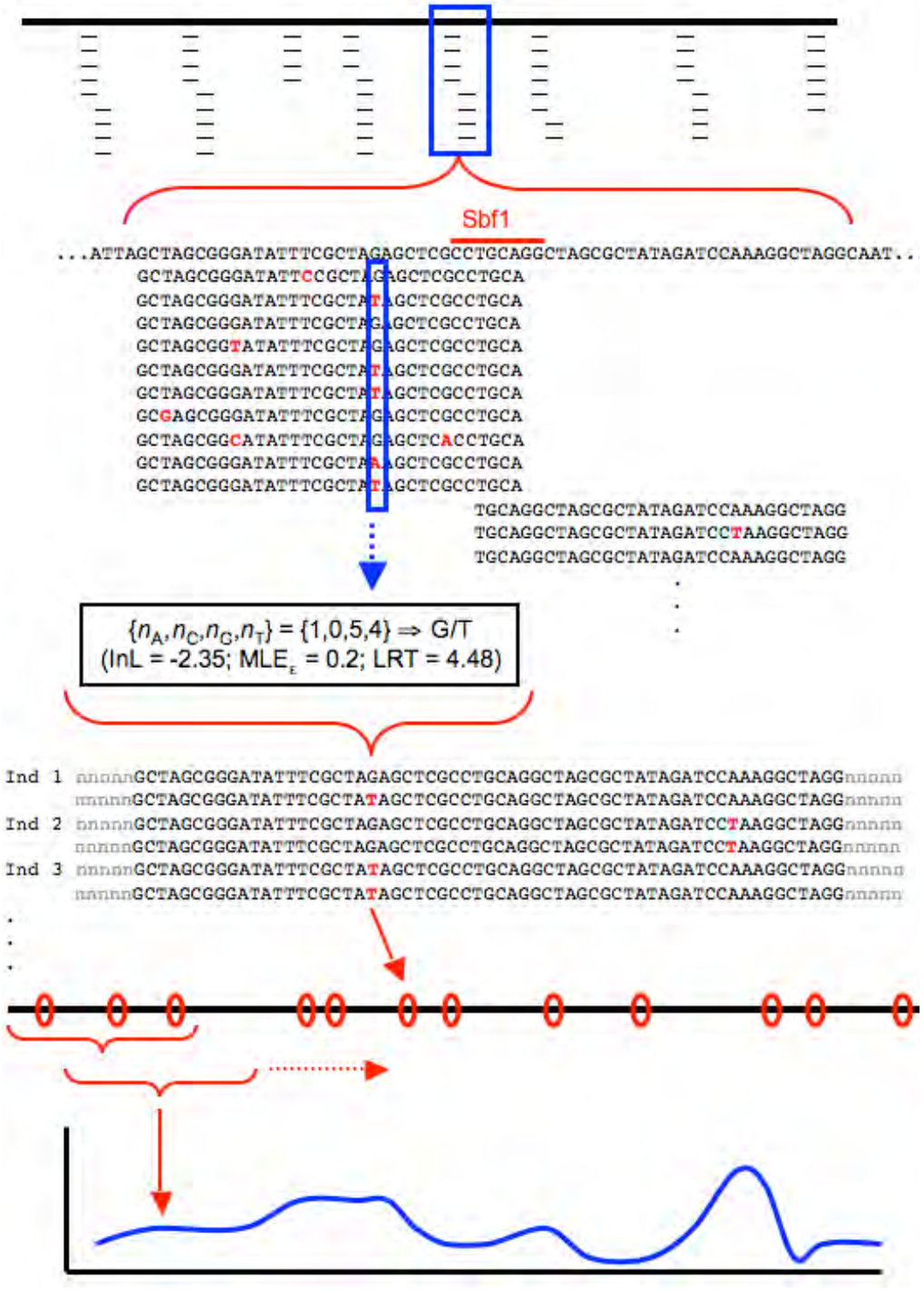
Making statistics continuous across the genome

Kernel-smoothing average of summary statistics along genome

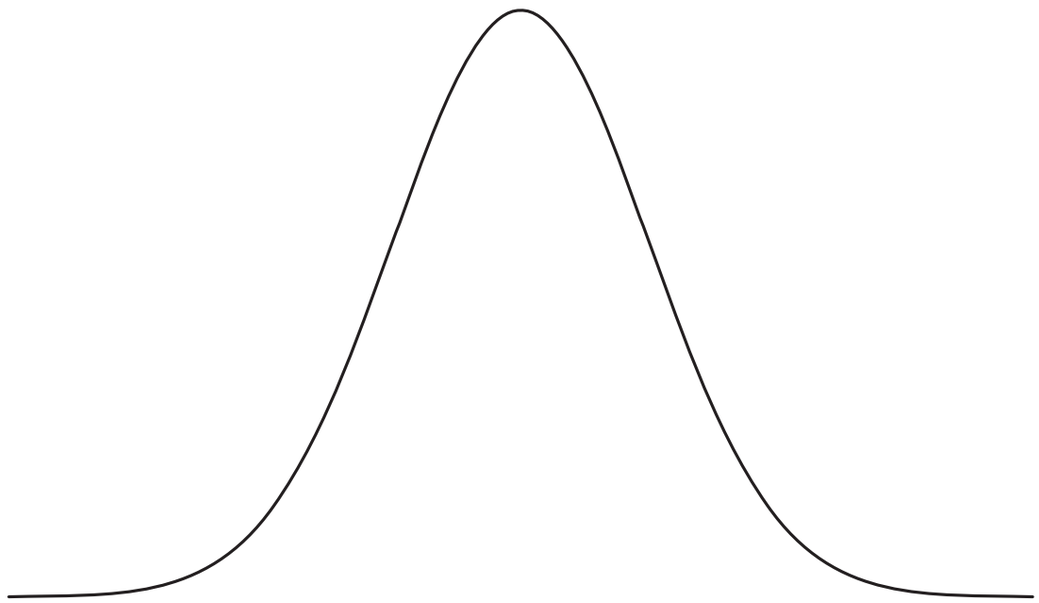


Bootstrap re-sampling to estimate significance of moving average

Overall pipeline



‘Bias’ in RAD-sequencing

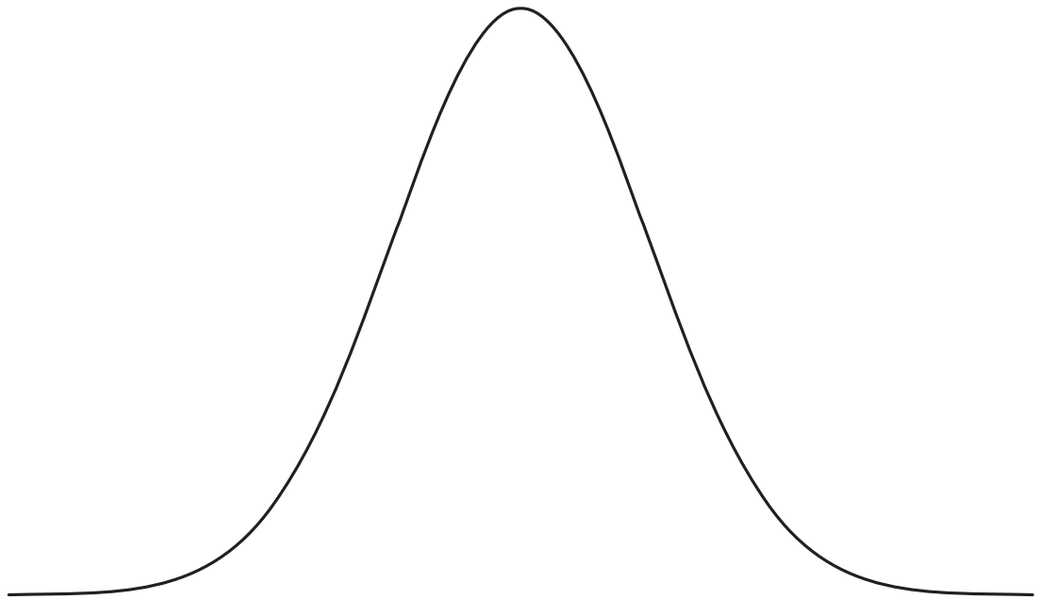


$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$e = 2.7182....$

$\pi = 3.1415....$

‘Bias’ in RAD-sequencing



$$\overline{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$$

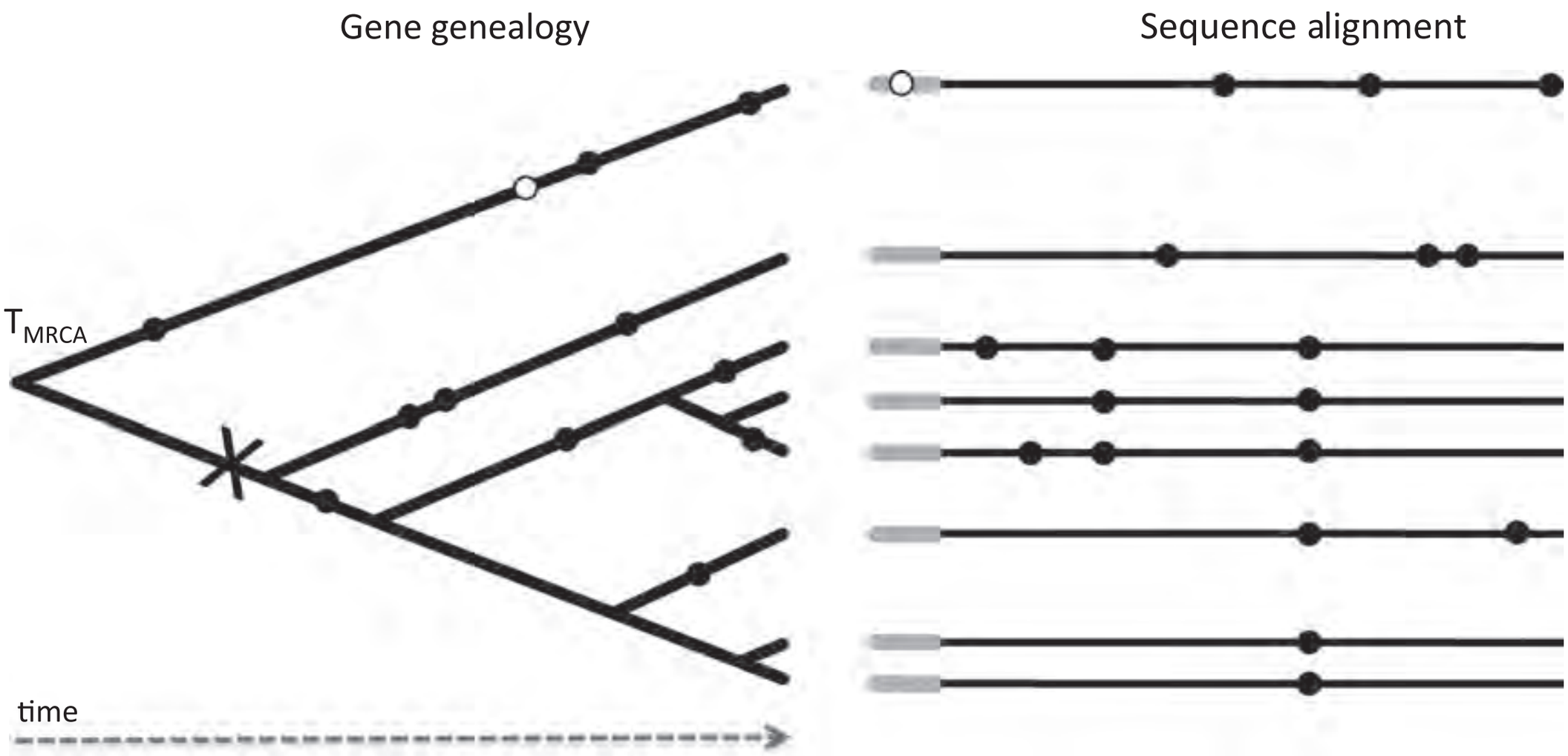
$$s^2 = \frac{1}{n-1}\sum_{i-1}^{n} (y_i - \overline{y})^2$$

Bias in RAD-sequencing

MOLECULAR ECOLOGY
Molecular Ecology (2013) **22**, 3179–3190 doi: 10.1111/mec.12

RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling

B. ARNOLD,[1] R. B. CORBETT-DETIG,[1] D. HARTL and K. BOMBLIES
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA



Bias in RAD-sequencing summary

		Mean	
		Recombination	
Protocol	θ per bp	θ_{we}/θ_{wa}	π_e/π_a
Standard	0.0001	0.994	0.995
	0.001	0.987	0.982
	0.01	0.956	0.933
Double digest	0.0001	0.835	0.836
	0.001	0.858	0.851
	0.01	0.829	0.797

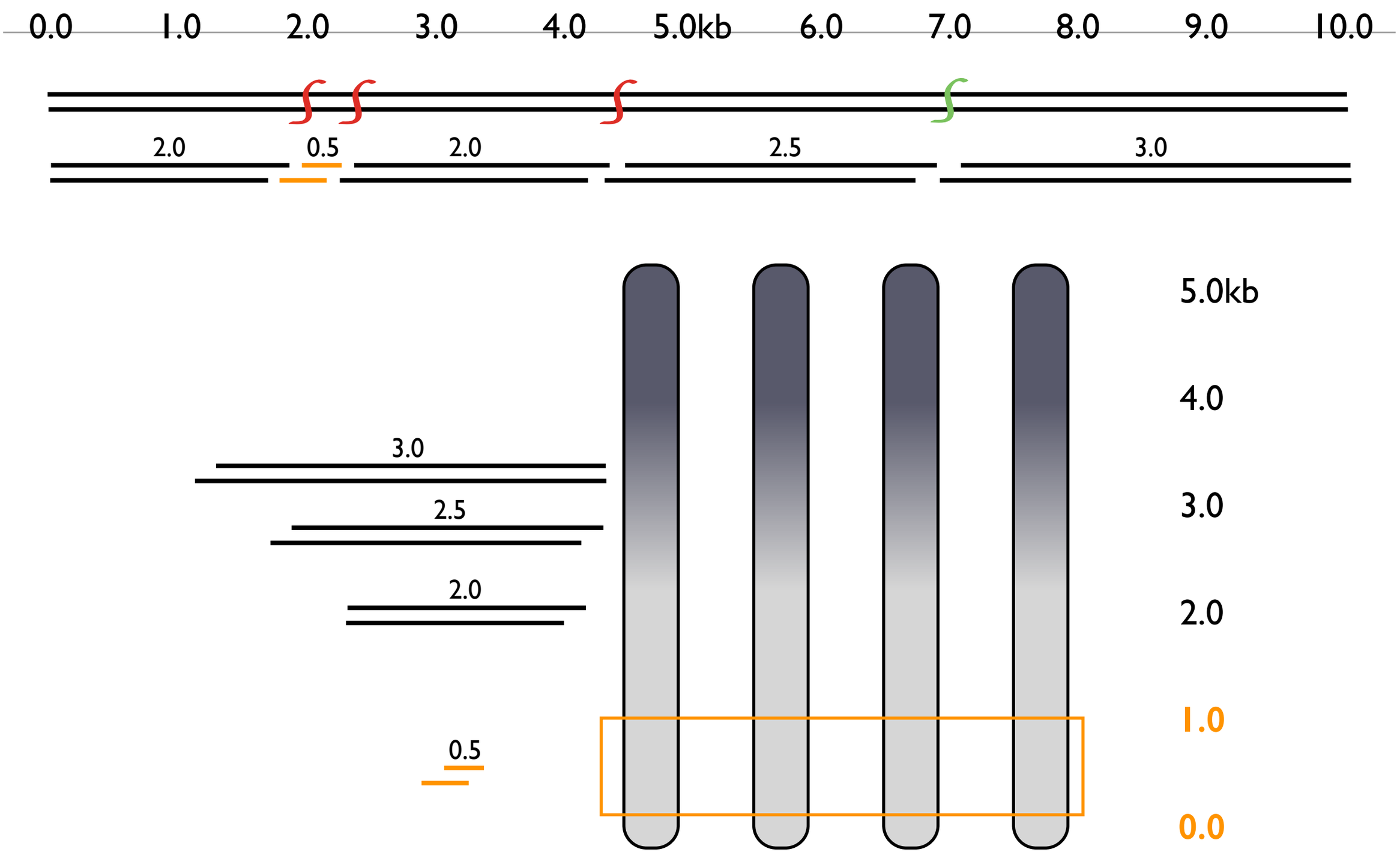
Bias in RAD-sequencing summary

		Mean Recombination	
Protocol	θ per bp	θ_{we}/θ_{wa}	π_e/π_a
Standard	0.0001	0.994	0.995
	0.001	0.987	0.982
	0.01	0.956	0.933
Double digest	0.0001	0.835	0.836
	0.001	0.858	0.851
	0.01	0.829	0.797

Bias in RAD-sequencing summary

		Mean Recombination	
Protocol	θ per bp	θ_{we}/θ_{wa}	π_e/π_a
Standard	0.0001	0.994	0.995
	0.001	0.987	0.982
	0.01	0.956	0.933
Double digest	0.0001	0.835	0.836
	0.001	0.858	0.851
	0.01	0.829	0.797

Why is ddRAD so much more biased?



Experimental design considerations for RAD

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

Experimental design considerations for RAD

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

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

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

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

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

Polymorphism and read length

Nucleotide polymorphism rate = 0.01 to 0.001 for most vertebrates

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

Experimental design considerations for RAD

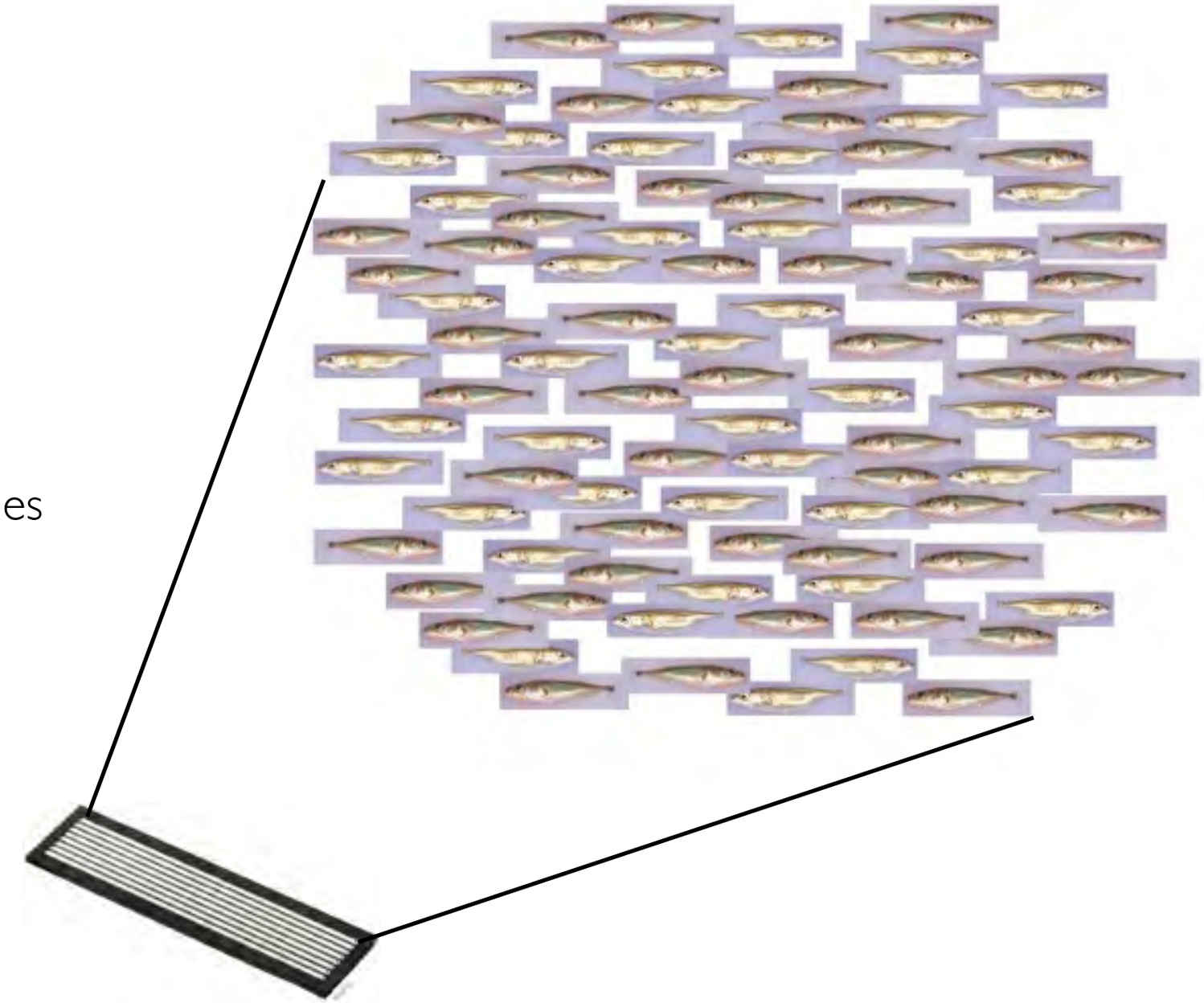
Tradeoffs:
Number of sites versus **Depth** of sequencing per site versus **Number of samples**

How many samples should be multiplexed?

Things to consider

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

Barcode distance - two mismatches



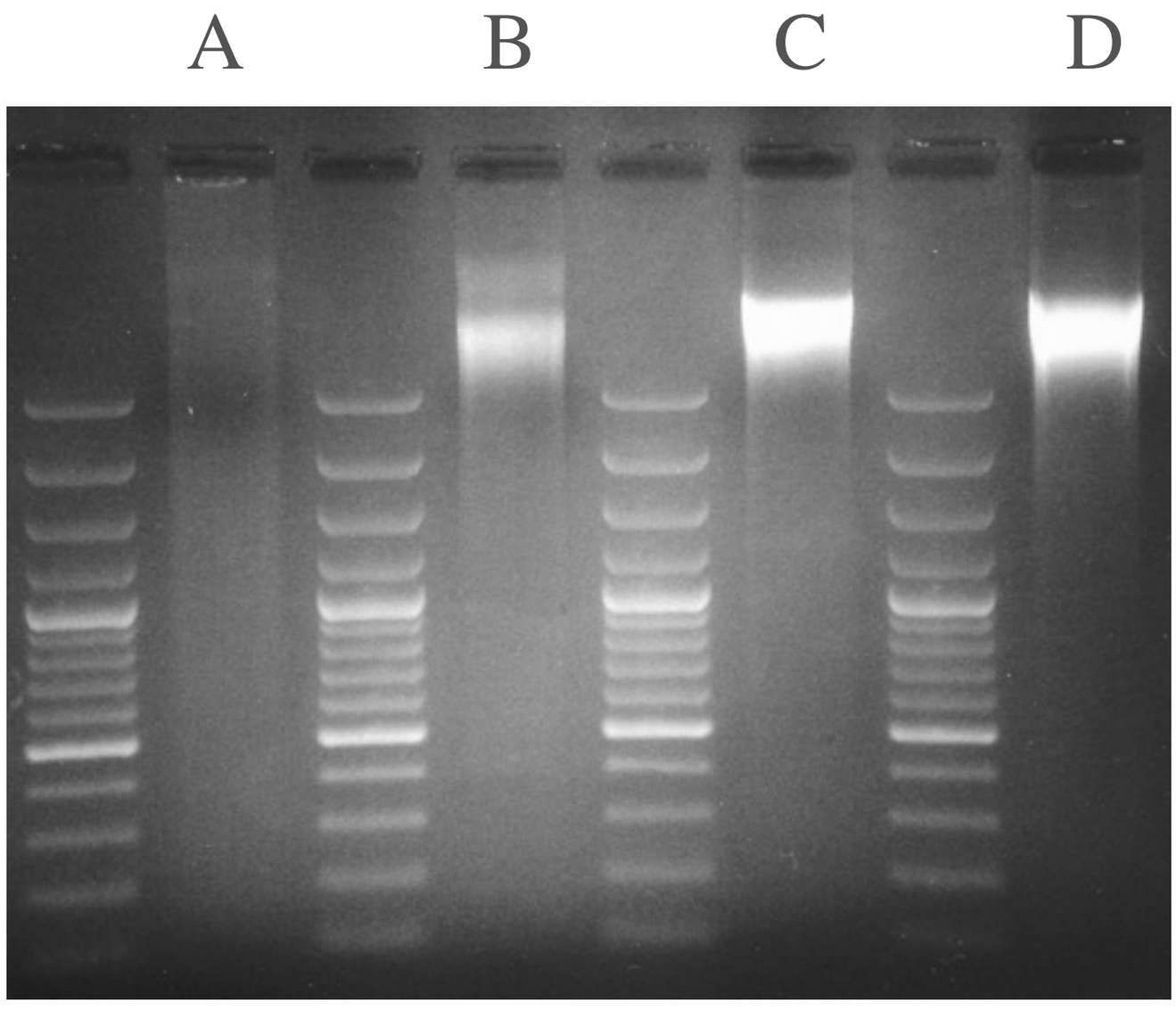
Molecular considerations in library building

How many samples should be multiplexed?

Things to consider

DNA Quality

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



Molecular considerations in library building

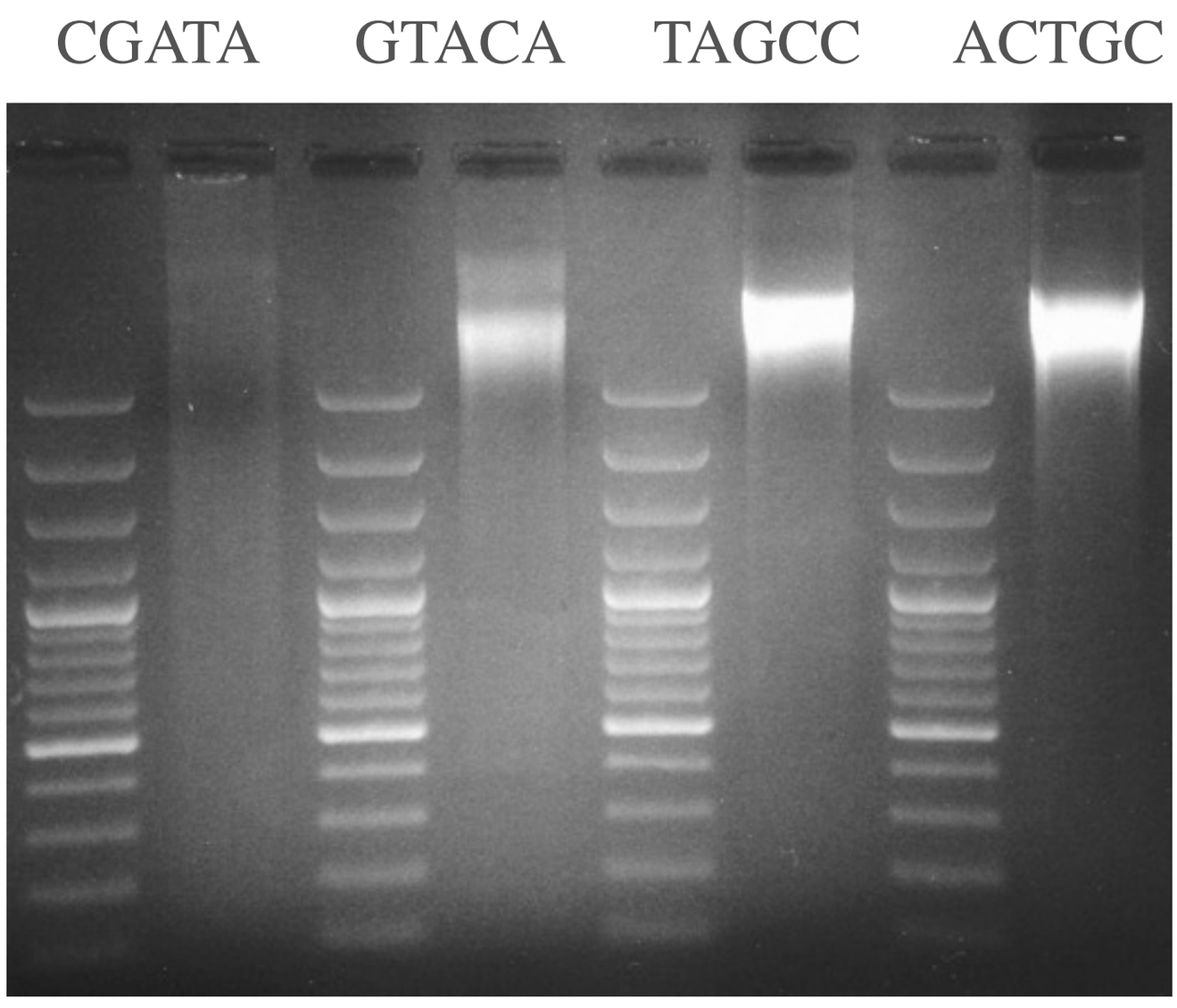
How many samples should be multiplexed?

Things to consider

DNA Quality

Diversify barcodes

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



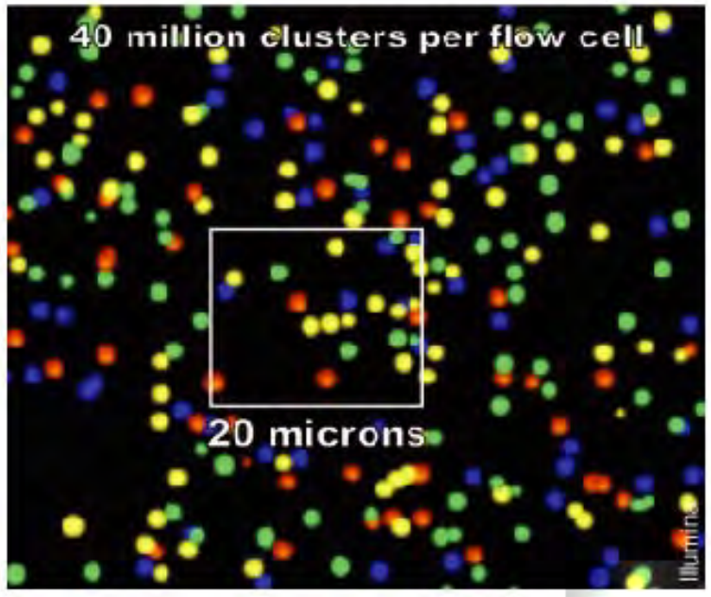
Molecular considerations in library building

How can I get the best depth of coverage?

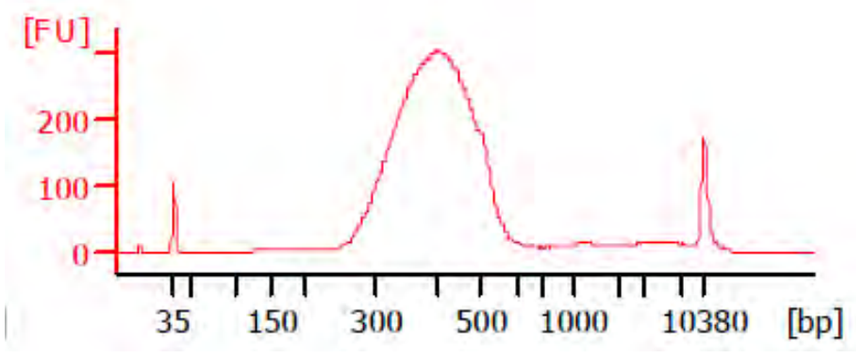
Things to consider

Fragment size

Smaller/tighter is better



Agilent Bioanalyzer



Molecular considerations in library building

How can I get the best depth of coverage?

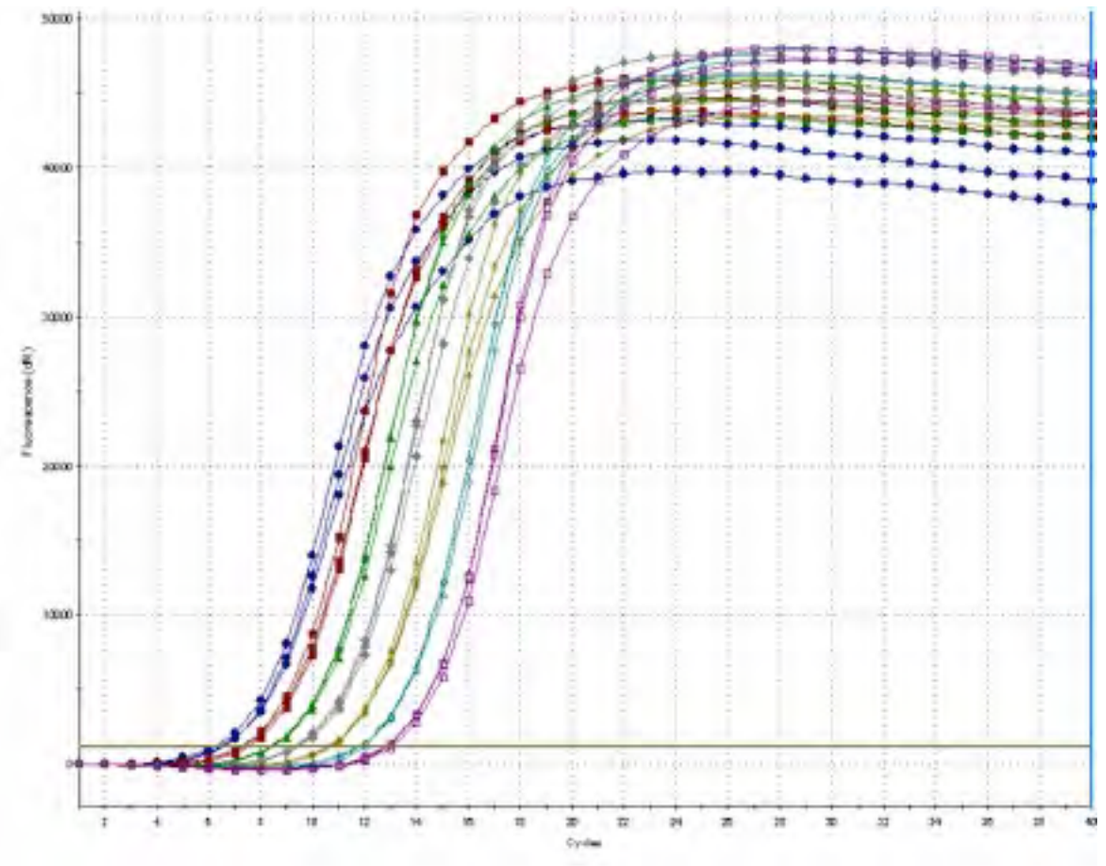
Things to consider

Fragment size

Library quality

qPCR

qPCR control should be similar to measured sample:



Molecular considerations in library building

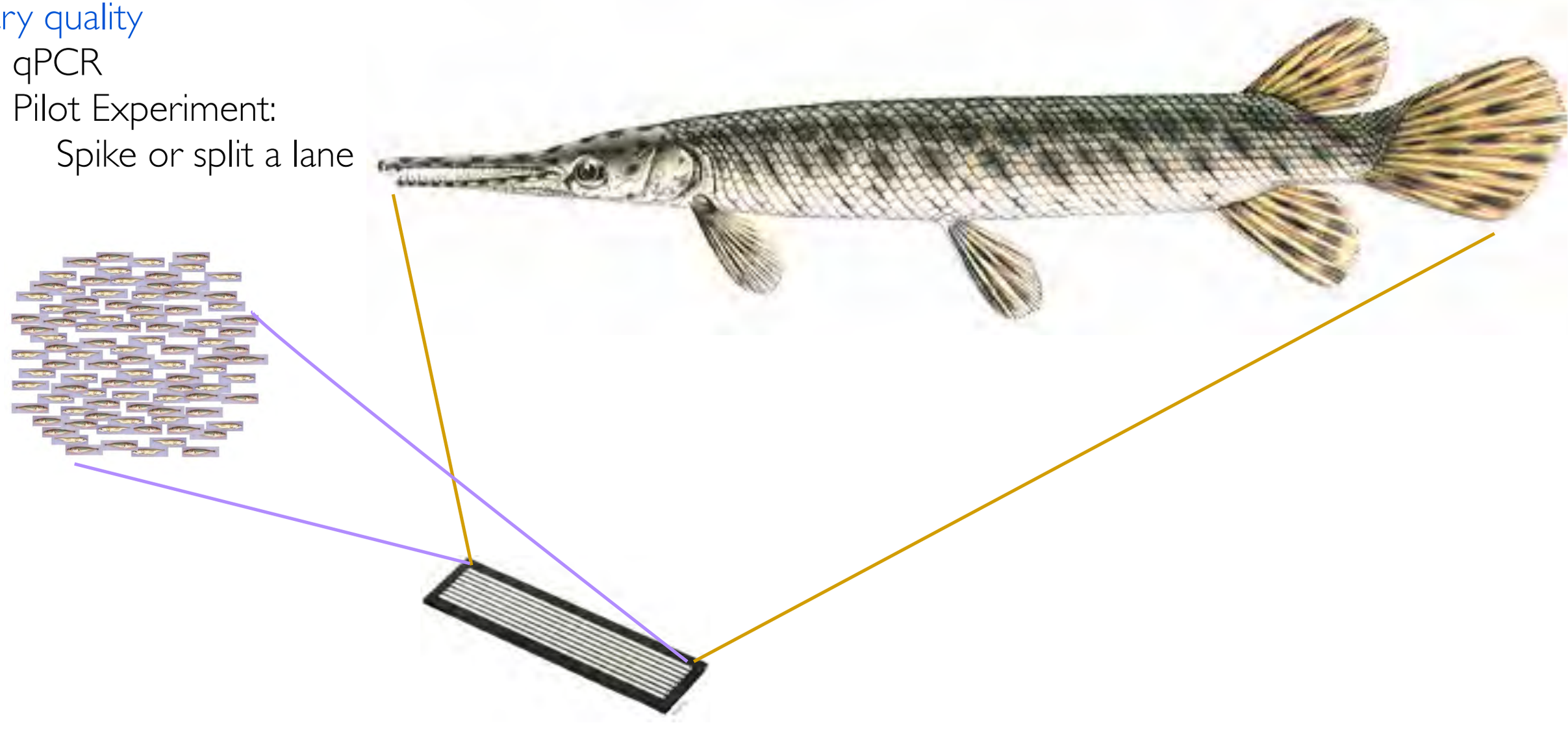
How can I get the best depth of coverage?

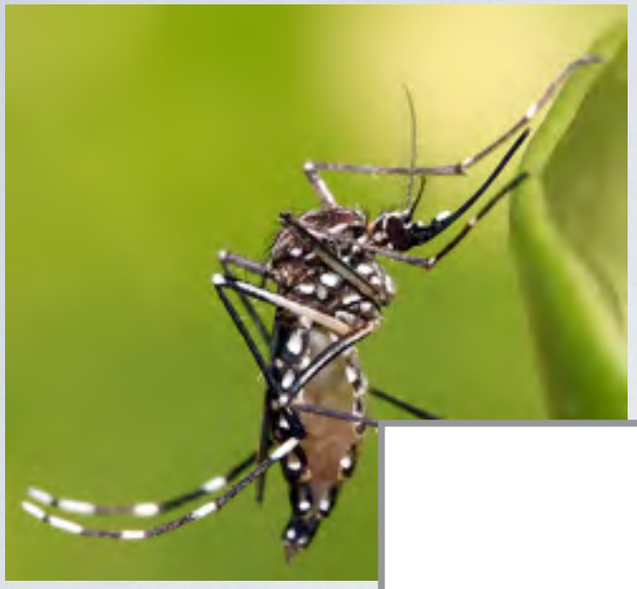
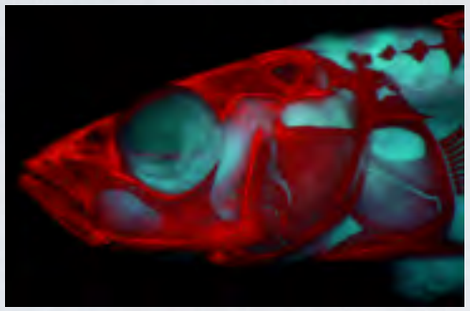
Things to consider

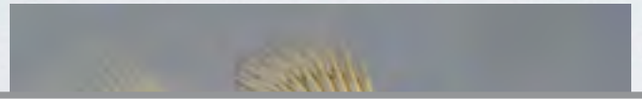
Fragment size

Library quality

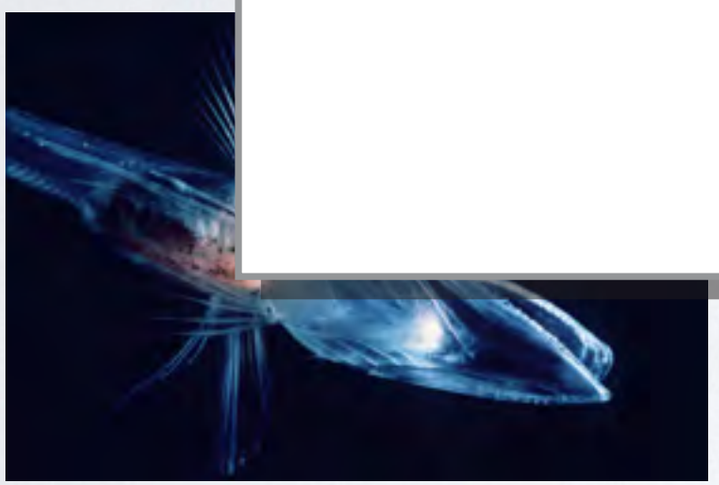
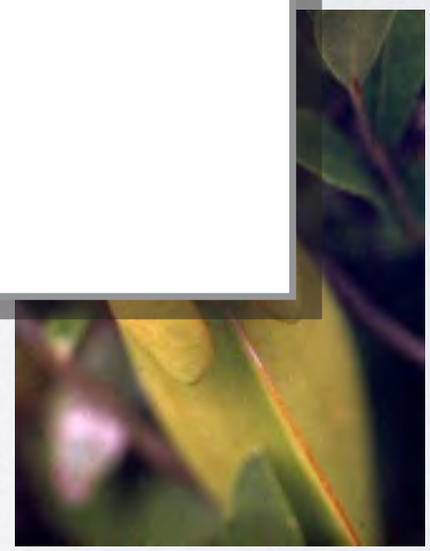
- qPCR
- Pilot Experiment:
  - Spike or split a lane



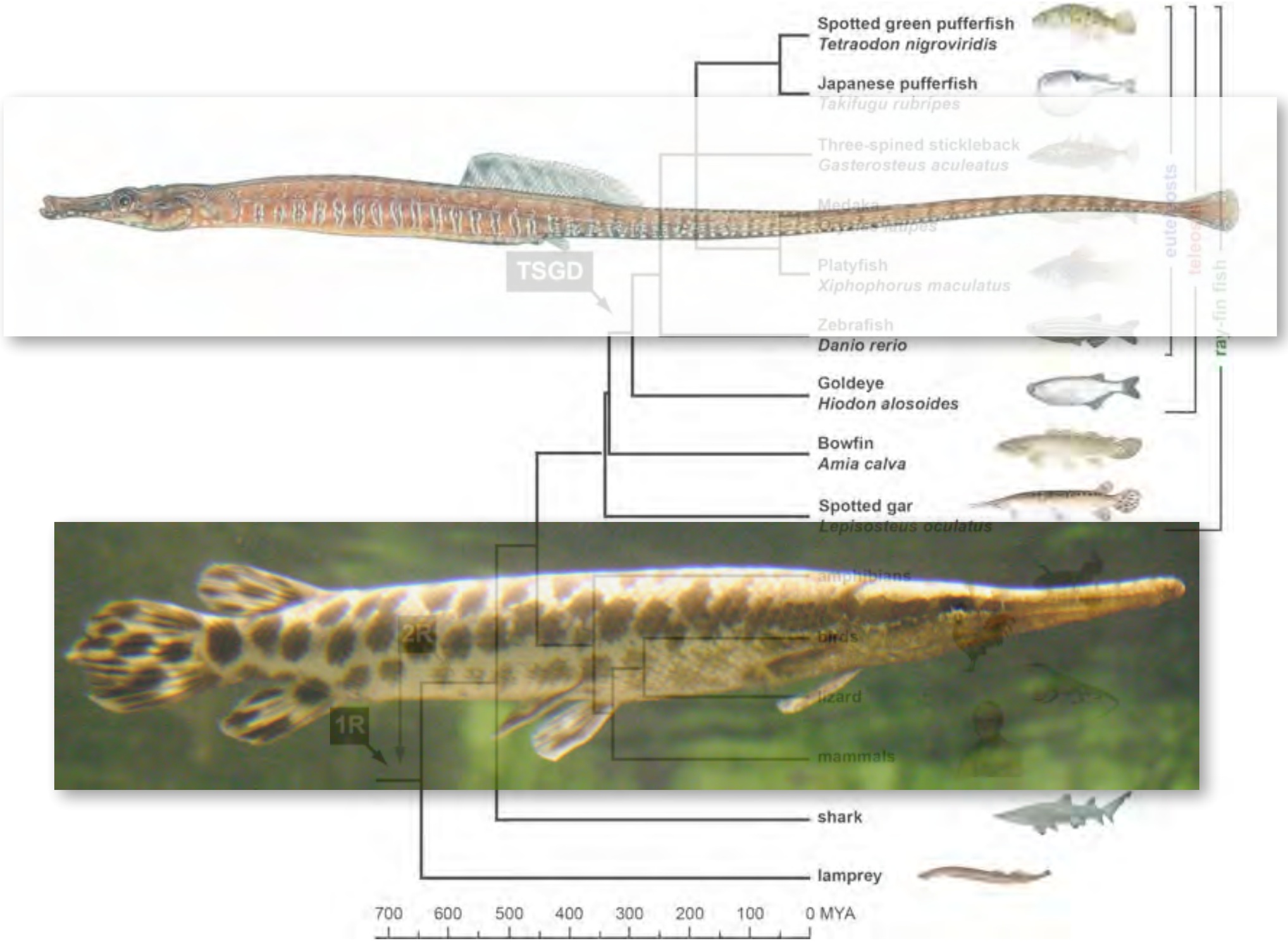



Statistical considerations for RADseq studies


What if you don’t have a genome sequence?

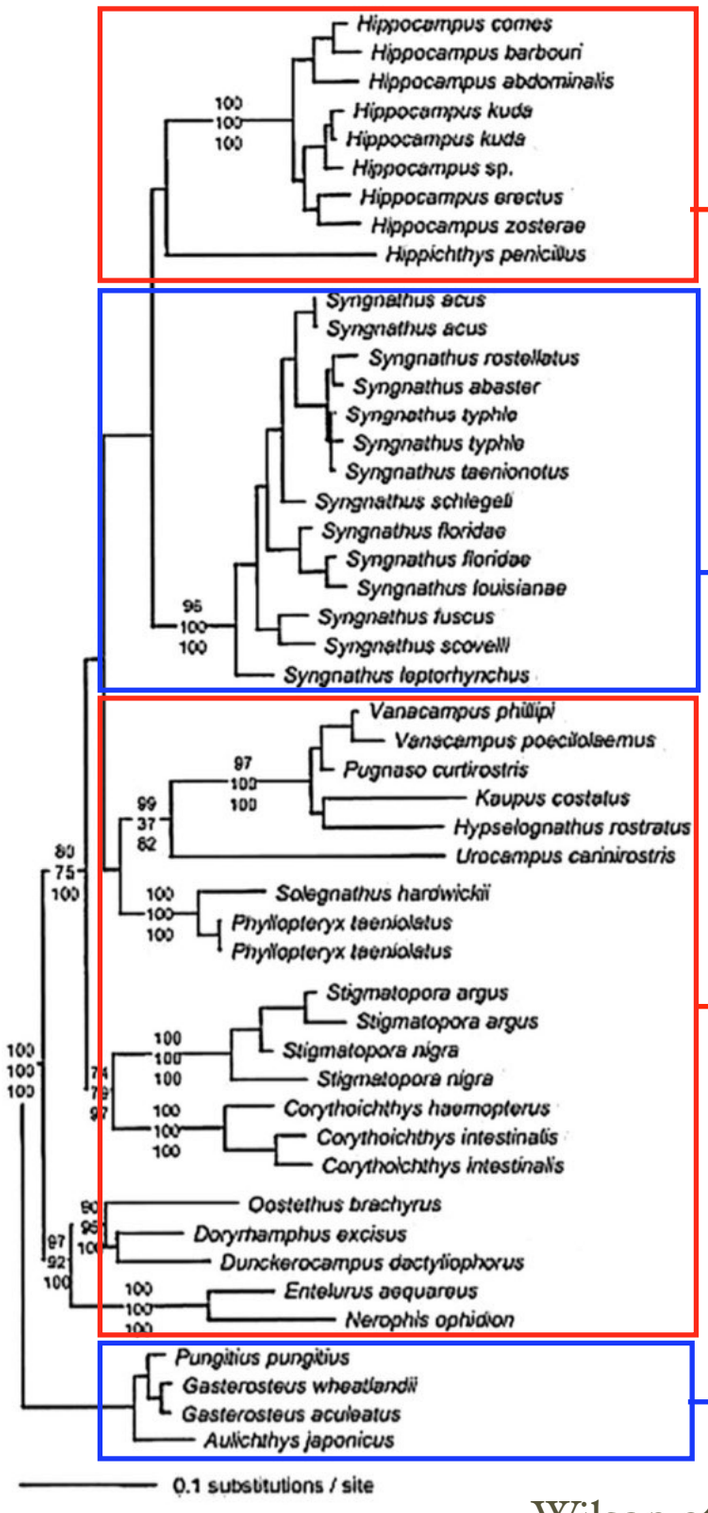
Genomically enabling a non-model organisms: RADseq can help



Julian Catchen, Allison Fuiten, Susie Bassham,
Clay Small and Adam Jones

Seahorses, sea dragons and pipefishes

Gasterosteidae and Syngnathidae are historically considered to be closely related



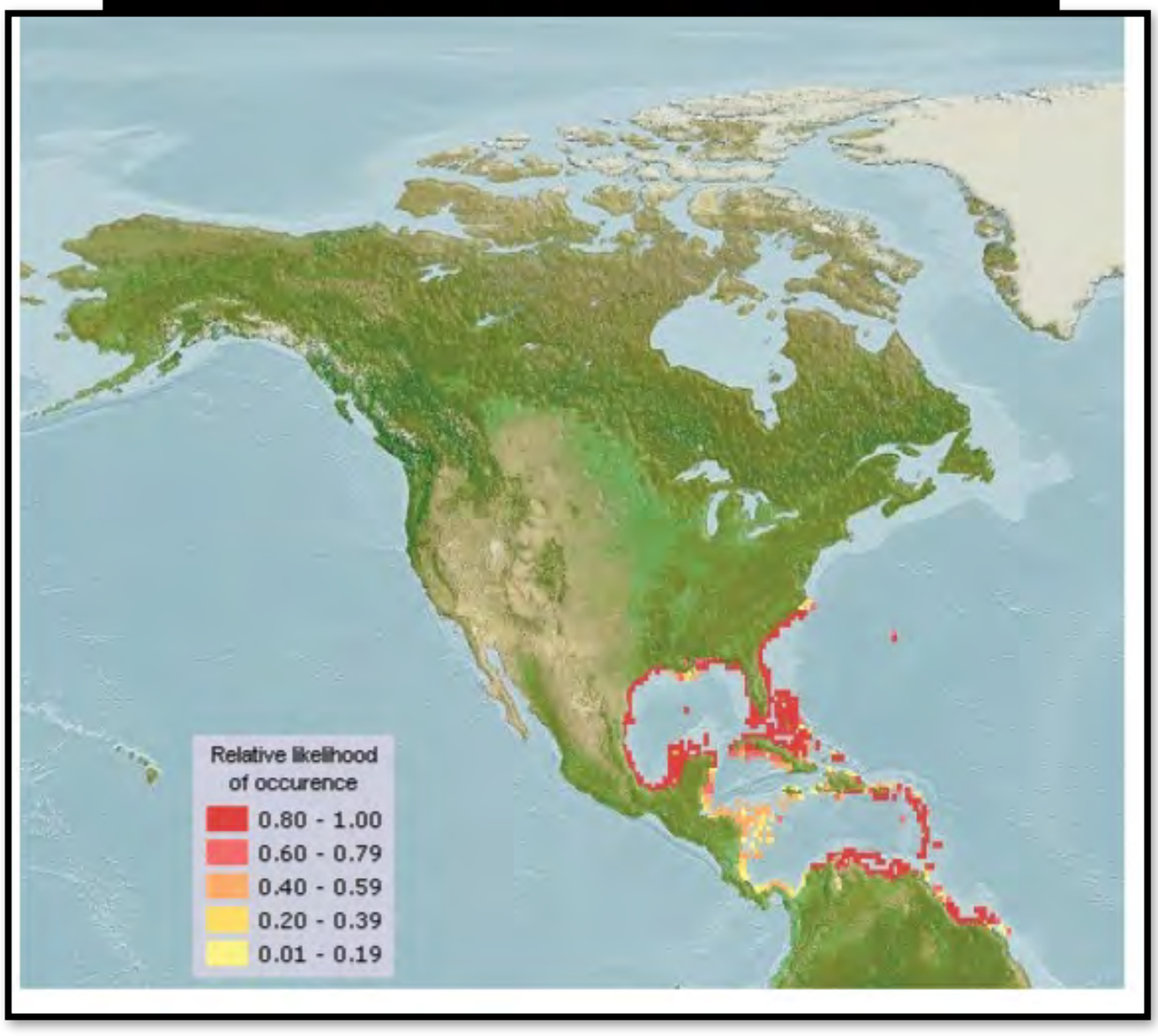
Seahorses

Pipefish

Seadragons

Stickleback

Wilson et al. 2003

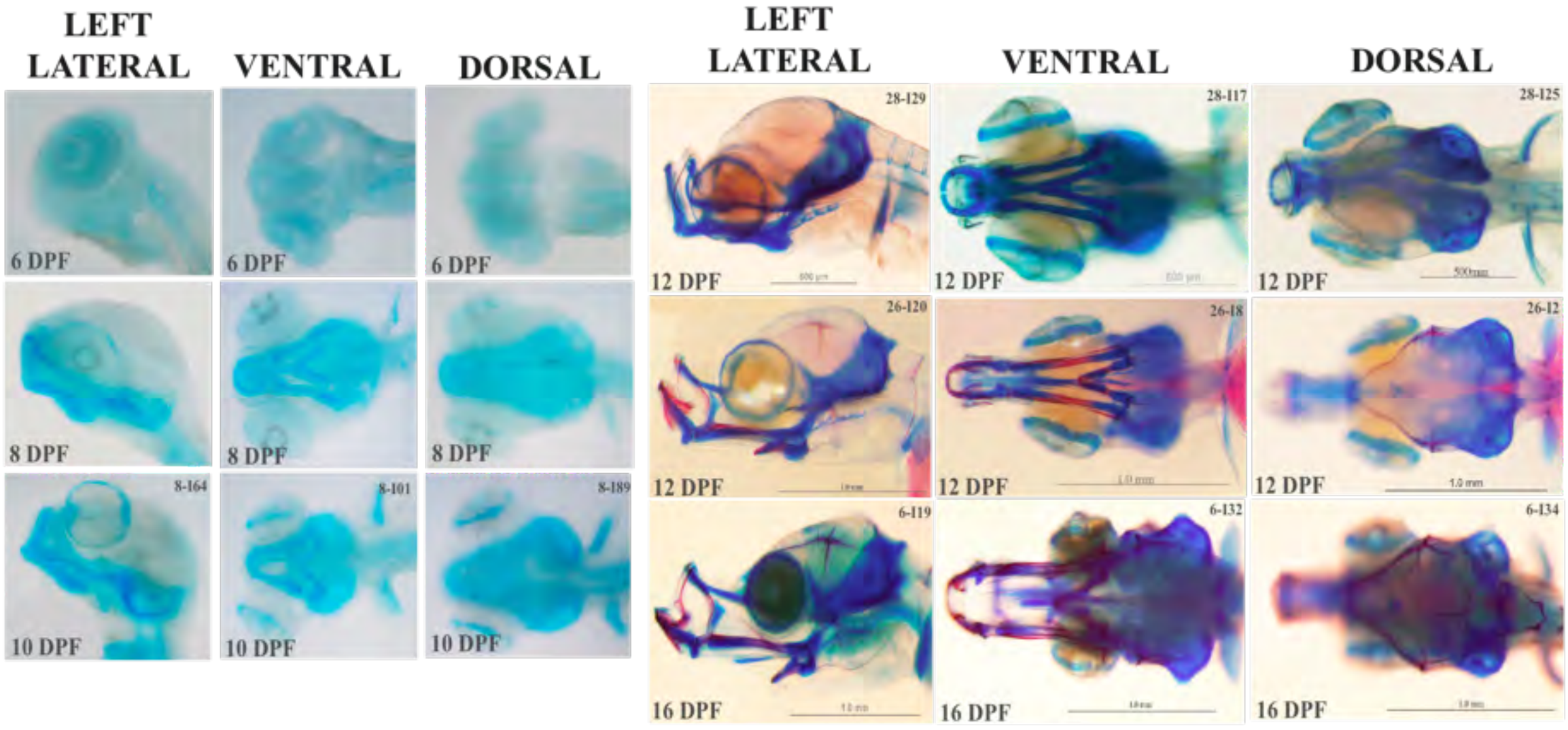


Gulf Pipefish
Syngnathus scovelli

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

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

We're really interested in the head and body axis



Solution: 'genomically enable' pipefish

1) A high quality transcriptome

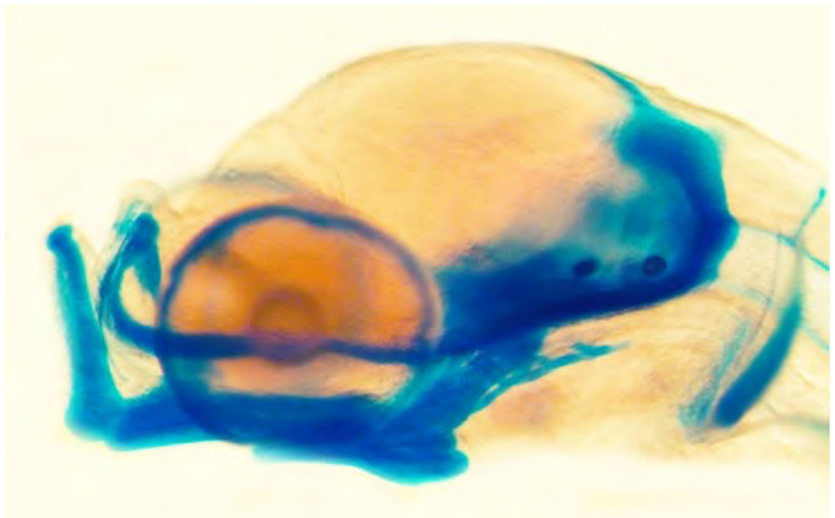
2) Very dense RAD genetic map

3) Deep coverage shotgun sequencing of genome

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

Pipefish Transcriptome

Building an EST database in pipefish



Pipefish embryonic mRNA

↓

Illumina sequencing:
100 nt, paired-end

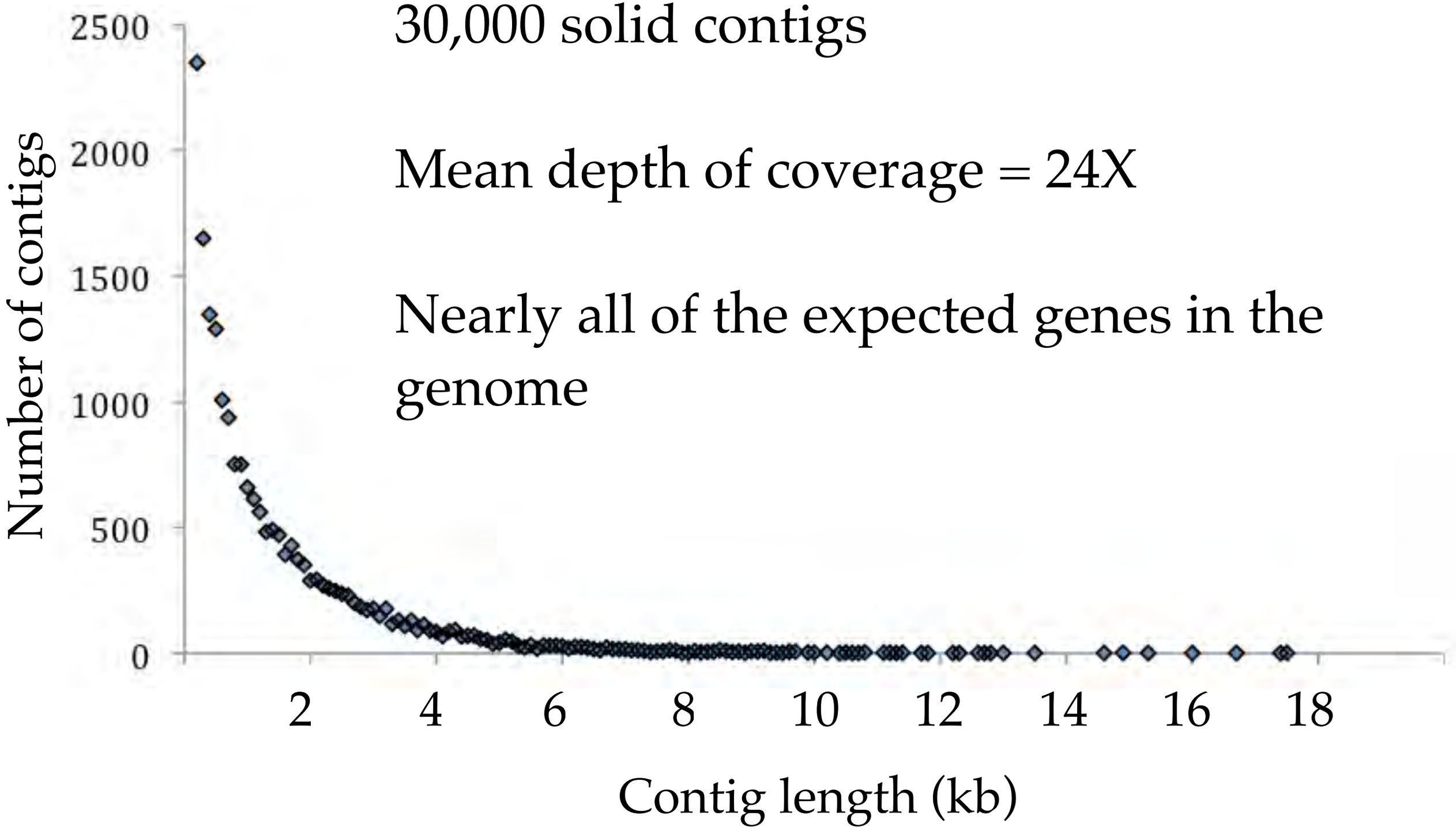
↓

200 million reads (two lanes)

↓

Assembly of transcripts

Transcriptome



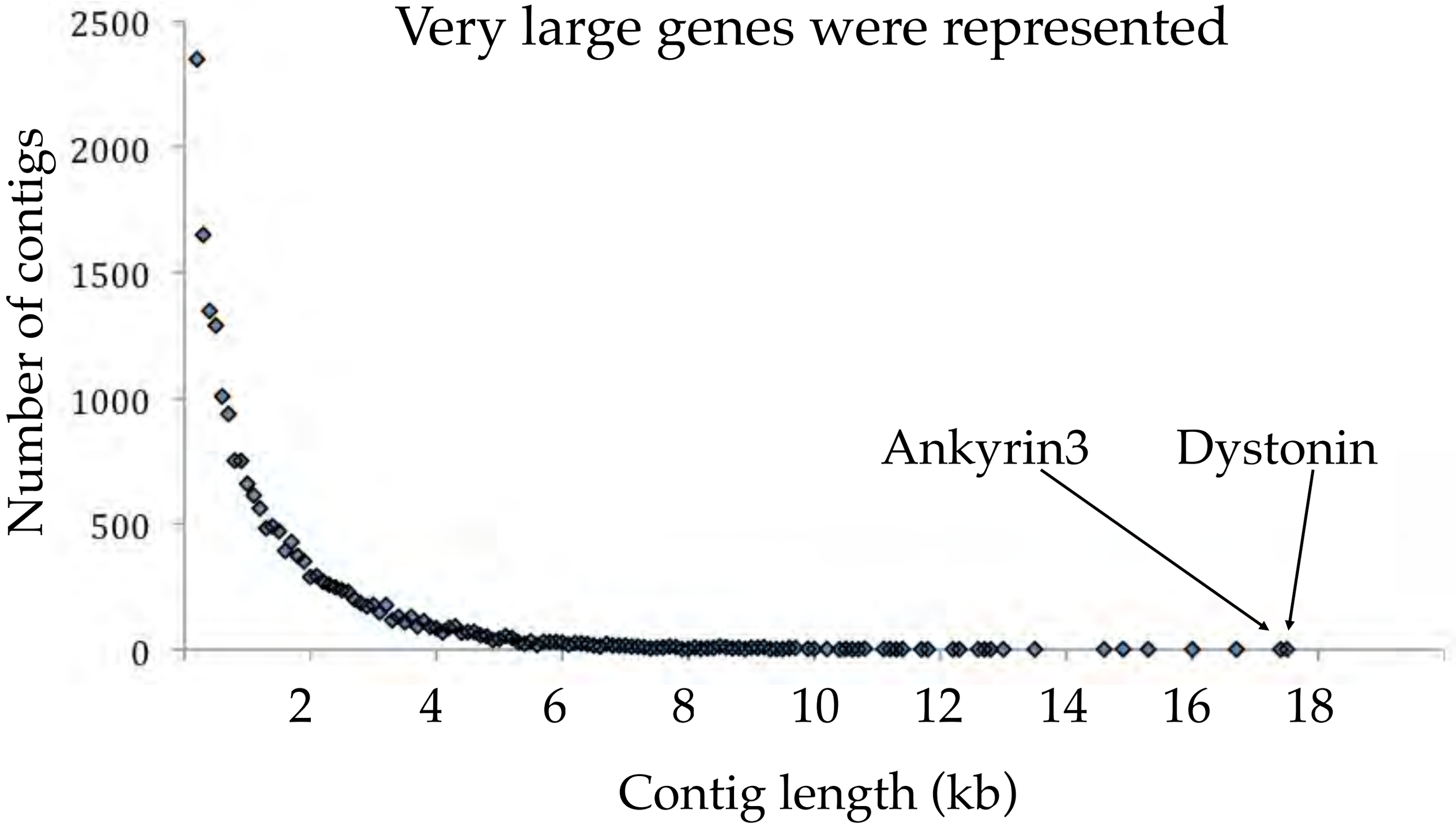
30,000 solid contigs

Mean depth of coverage = 24X

Nearly all of the expected genes in the genome

Transcriptome

Very large genes were represented



Pipefish Genetic Map

Genetic map workflow

Generated an F1 family of 103 individuals

RAD sequenced the parents and offspring

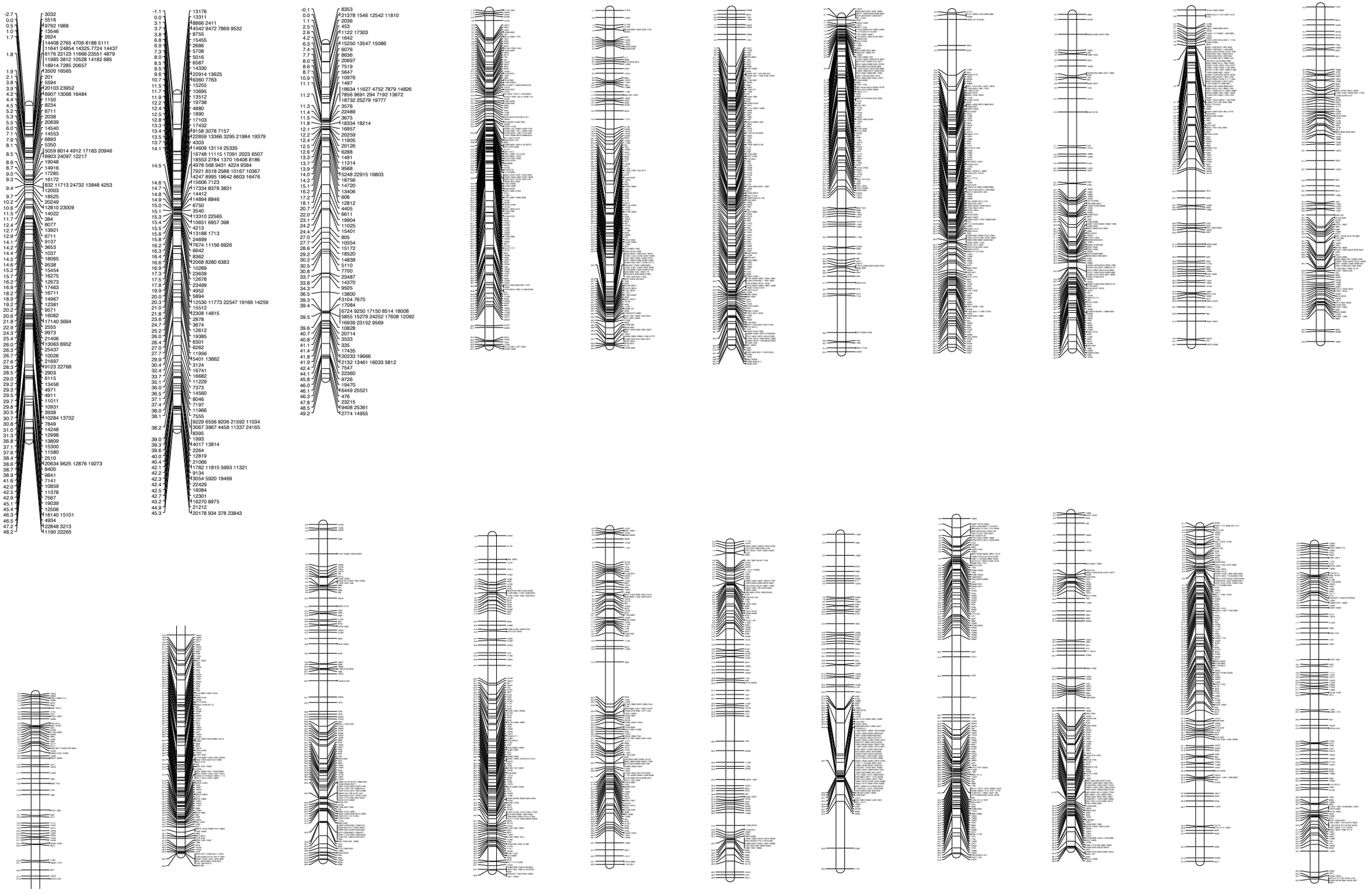
Analyzed the data using *Stacks*

Paired end local assemblies

Output to JoinMap format

Created Linkage map

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



Pipefish Genome Project

Genome workflow

Generated DNA from a single individual

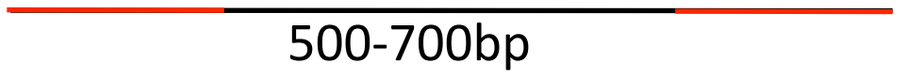
Random Illumina shotgun sequencing

Removed highly repetitive kmers

Produced *several* different genome assemblies

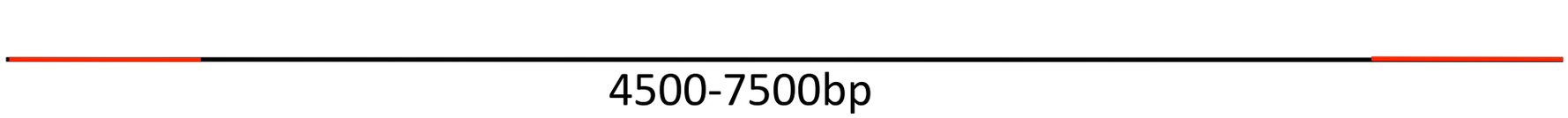
Illumina genomic libraries for pipefish genome

paired end 101bp



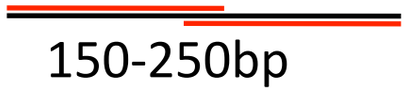
25x

mate pair



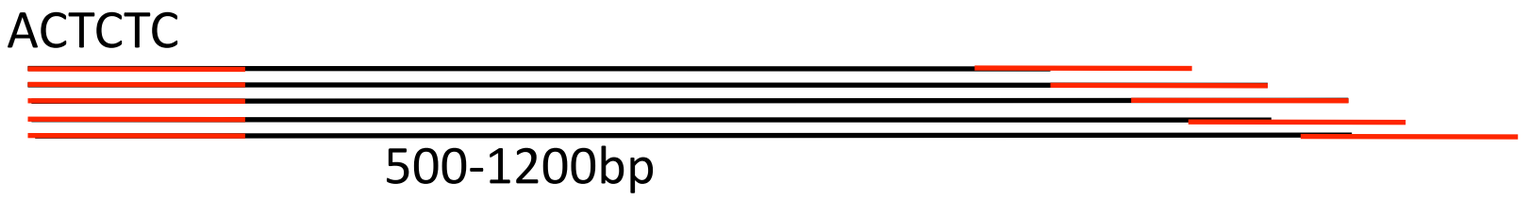
2x

overlapping



40x

paired end RAD



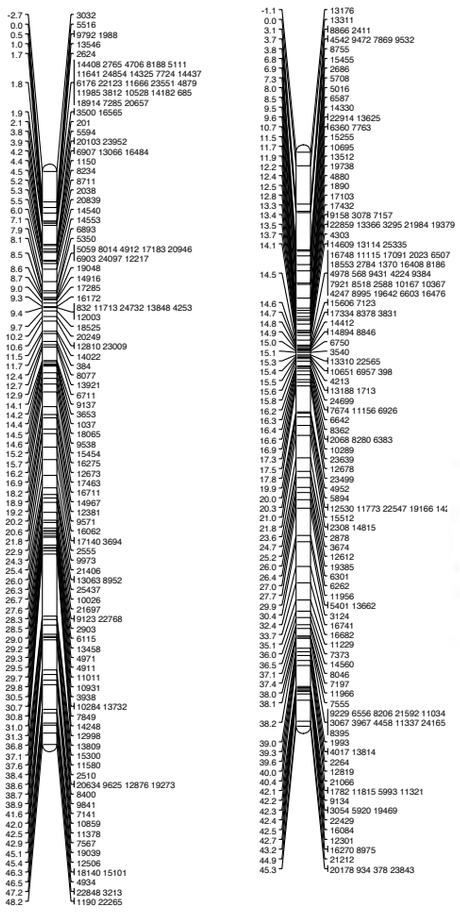
15-25x of 3% of the genome

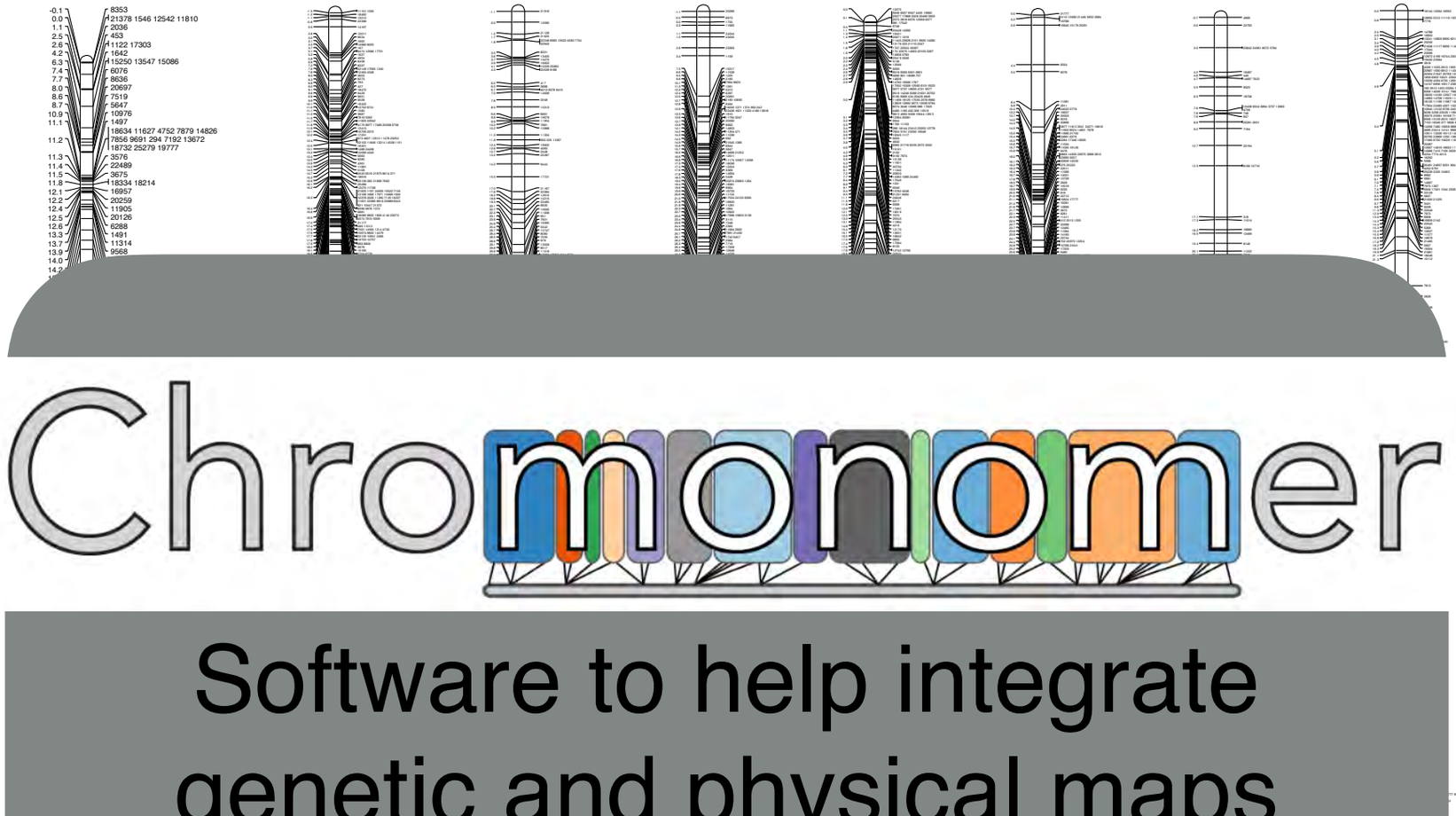
Pipefish genome assembly version 0.99
Nearly the whole genome is covered

Coverage	Scaffolds	Contigs	Scaffold N50	Contig N50
All (66.6x)	1,820	25,075	796,183	23,012

Max	Average Length	Total Length	Gap Length	%
5,676,603	162,928	296,529,585	27,303,839	(8.39%)

The pipefish genetic map and genome together

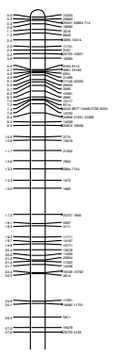


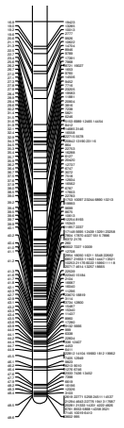


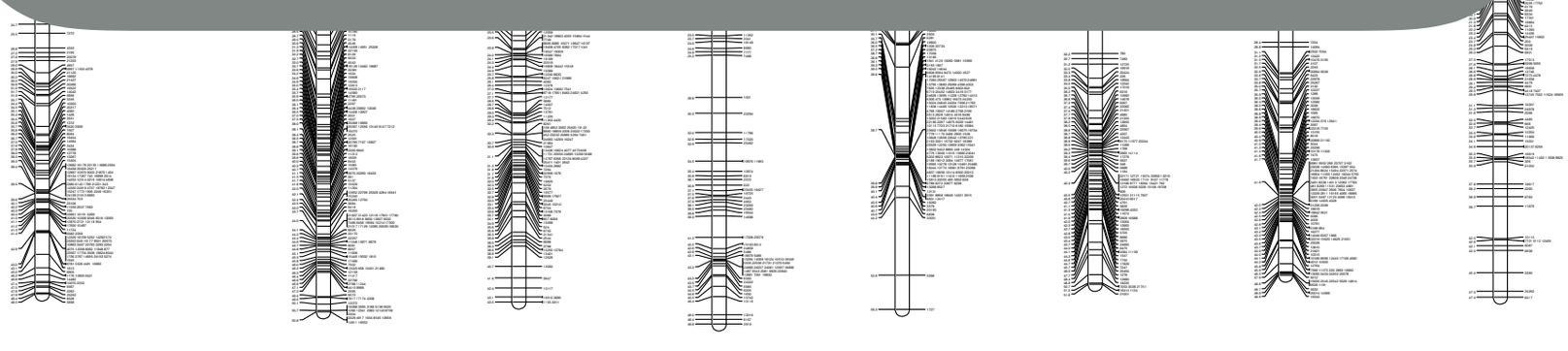
Chromonomer

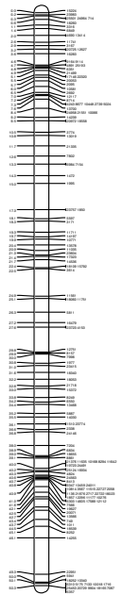
Software to help integrate genetic and physical maps

http://catchenlab.life.illinois.edu/chromonomer/









Overall Conclusions

Genomics can be a tool for enabling new research in models & nonmodels

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

Not your father's genome assembly

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

Open Source Genomics provides a suite of breakthrough technologies

- the molecular approaches are not as daunting as they first appear
- analytical and computational approaches are challenging
  - **New software tools can help, but knowledge of Unix and Scripting is essential**
  - **Also essential to be comfortable with classical and modern statistics**

Acknowledgments

- *Past and present lab members* **Paul Hohenlohe**, **Thom Nelson**, Joe Dunham, Nicole Nishimura & **Mark Currey**
- *Collaborators* **Eric Johnson**, Patrick Phillips, **Chuck Kimmel**, **John Postlethwait**
- *Funding* from NSF & NIH, as well as Keck & Murdock Foundations