RAD-seq in Roscoff

Matthieu Bruneaux

2015-03-10

Mini-workshop about ddRAD

Introduction about RAD-seq

- RAD? RAD-seq? ddRAD?
- Applications
- Workflow

Practicals

- One complete project, from raw reads to final results
- Cherry-picking of some analysis steps
- Open questions

Objectives

- Overview of RAD-seq
- Arouse curiosity
- Give useful pointers

Disclaimer about the speaker!

- Not a population geneticist, not a bioinformatician
- Evolutionary biologist who dropped into a RAD-seq project when he was a small post-doc
- Some things said here are probably incorrect or plainly wrong!

What are RAD markers?

Miller et al. 2007

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

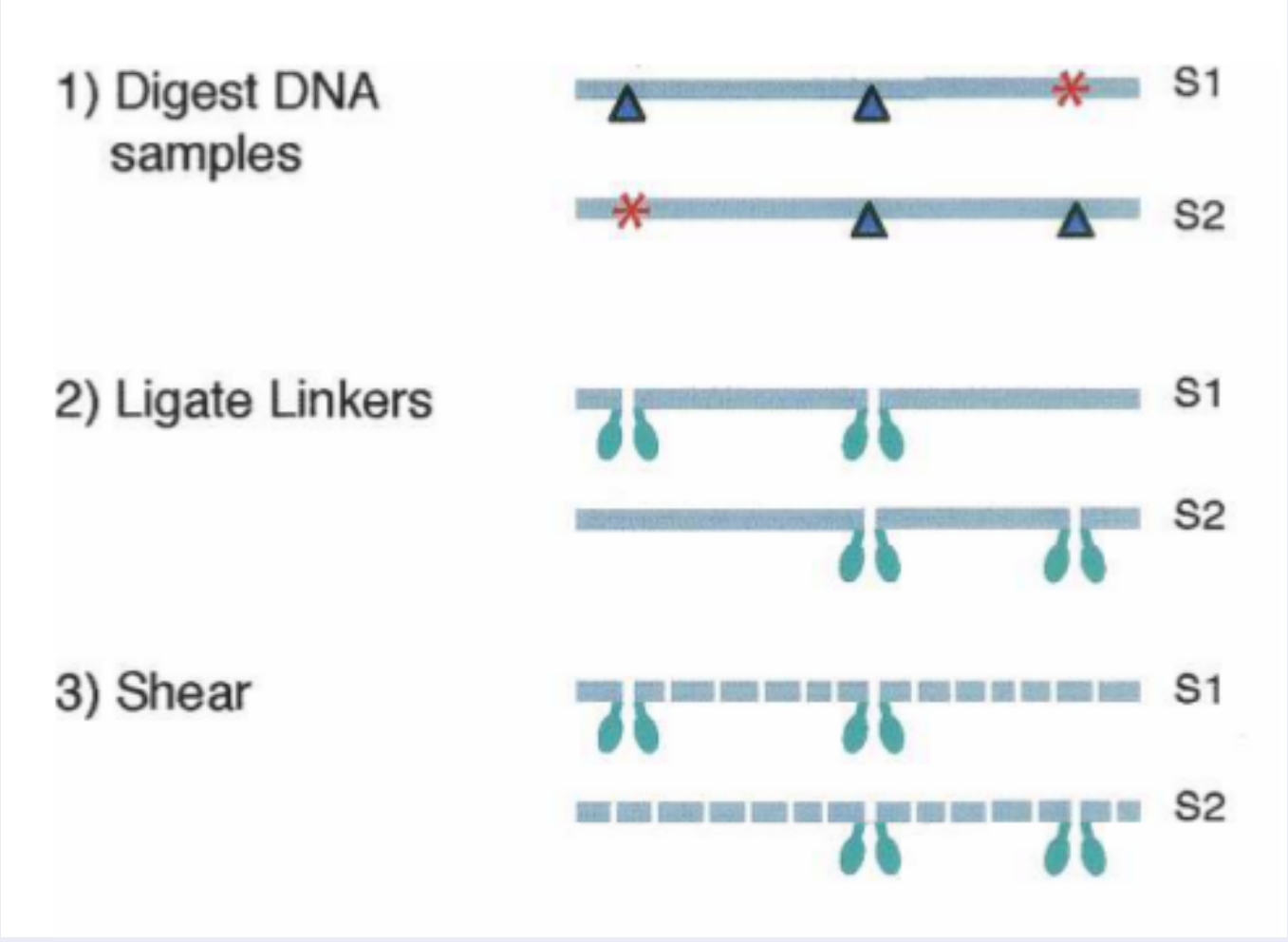
Michael R. Miller,[1] Joseph P. Dunham,[2] Angel Amores,[3] William A. Cresko,[2] and Eric A. Johnson[1,4]

Description of RAD markers

- Restriction site associated DNA fragments
- Used with micro-array systems
- Similar to RFLP or AFLP, but many more markers

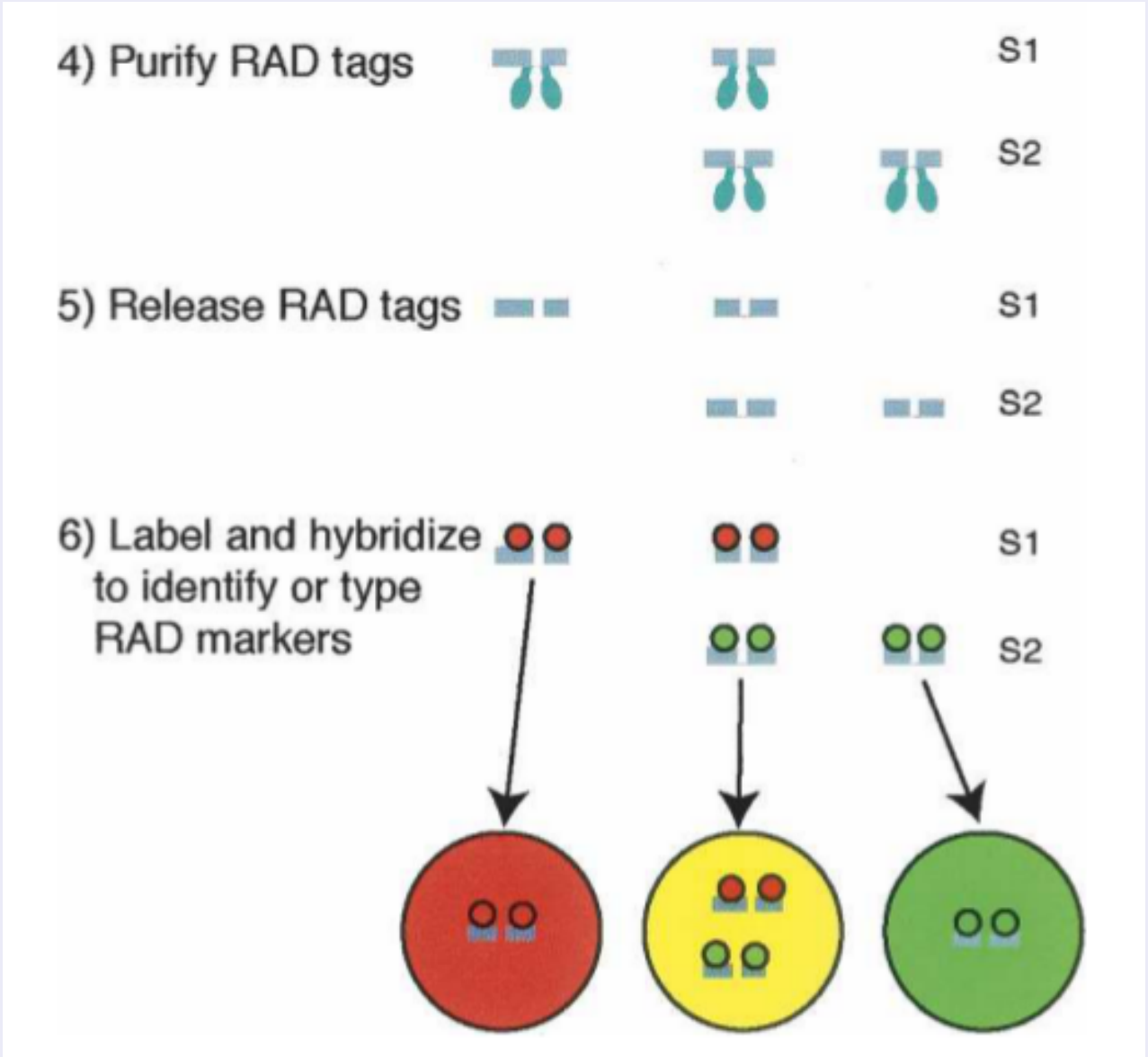
RAD - Miller et al. 2007 (6 steps)

Digest - tag - shear



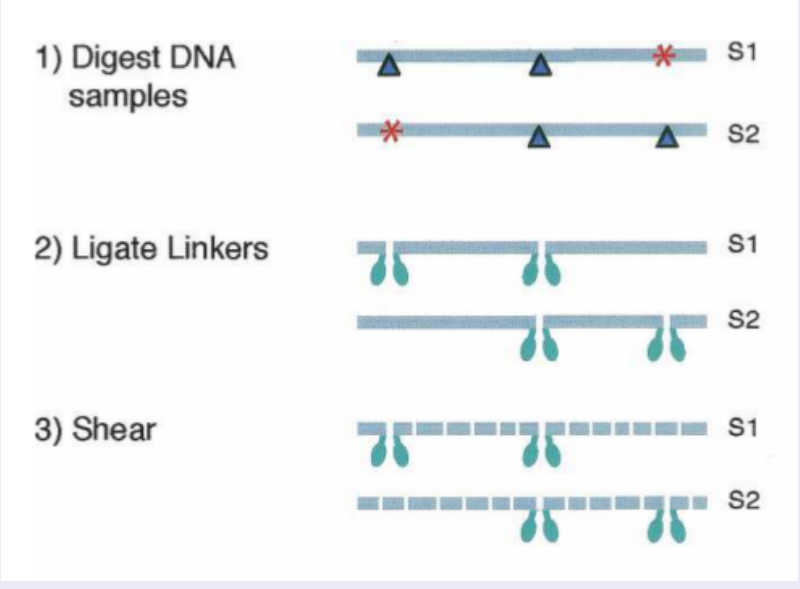
RAD - Miller et al. 2007 (6 steps)

Purify - release - type

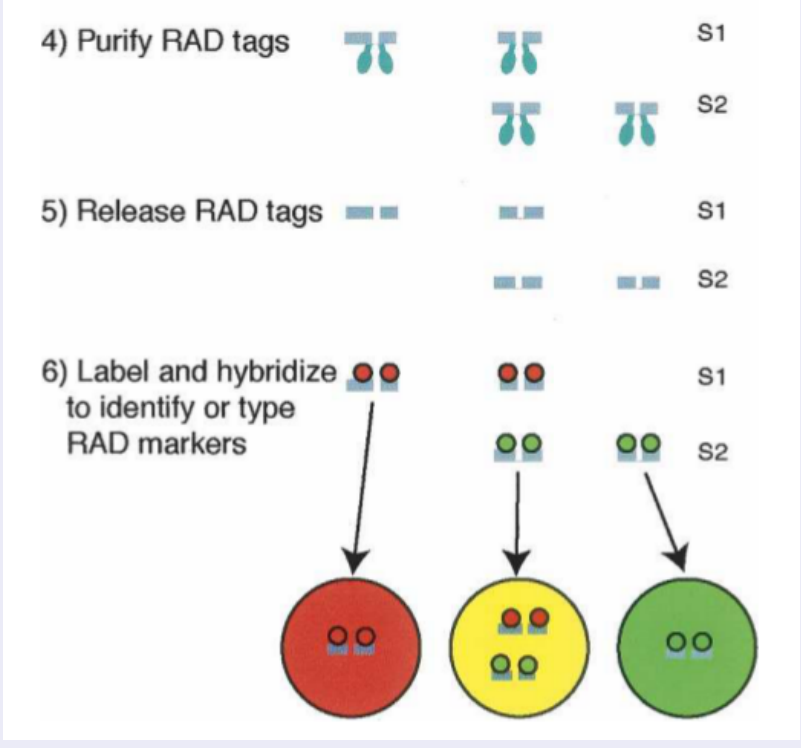


RAD - Miller et al. 2007 (method summary)

Digest - tag - shear



Purify - release - type



Demonstration

- Mapping breakpoint on a *Drosophila* chromosome
- Identification of the lateral plate locus in threespine stickleback

RAD - Miller et al. 2007

Advantage of the method

- Easy-to-produce genotyping resource for non-model species
- Moderate cost
- Genetic mapping possible (if markers location known)
- Bulk genotyping possible

But note that...

- At this point the restriction site is the polymorphic marker
- One restriction enzyme only is used

What is RAD-seq?

Baird et al. 2008

OPEN ACCESS Freely available online

PLoS one

Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

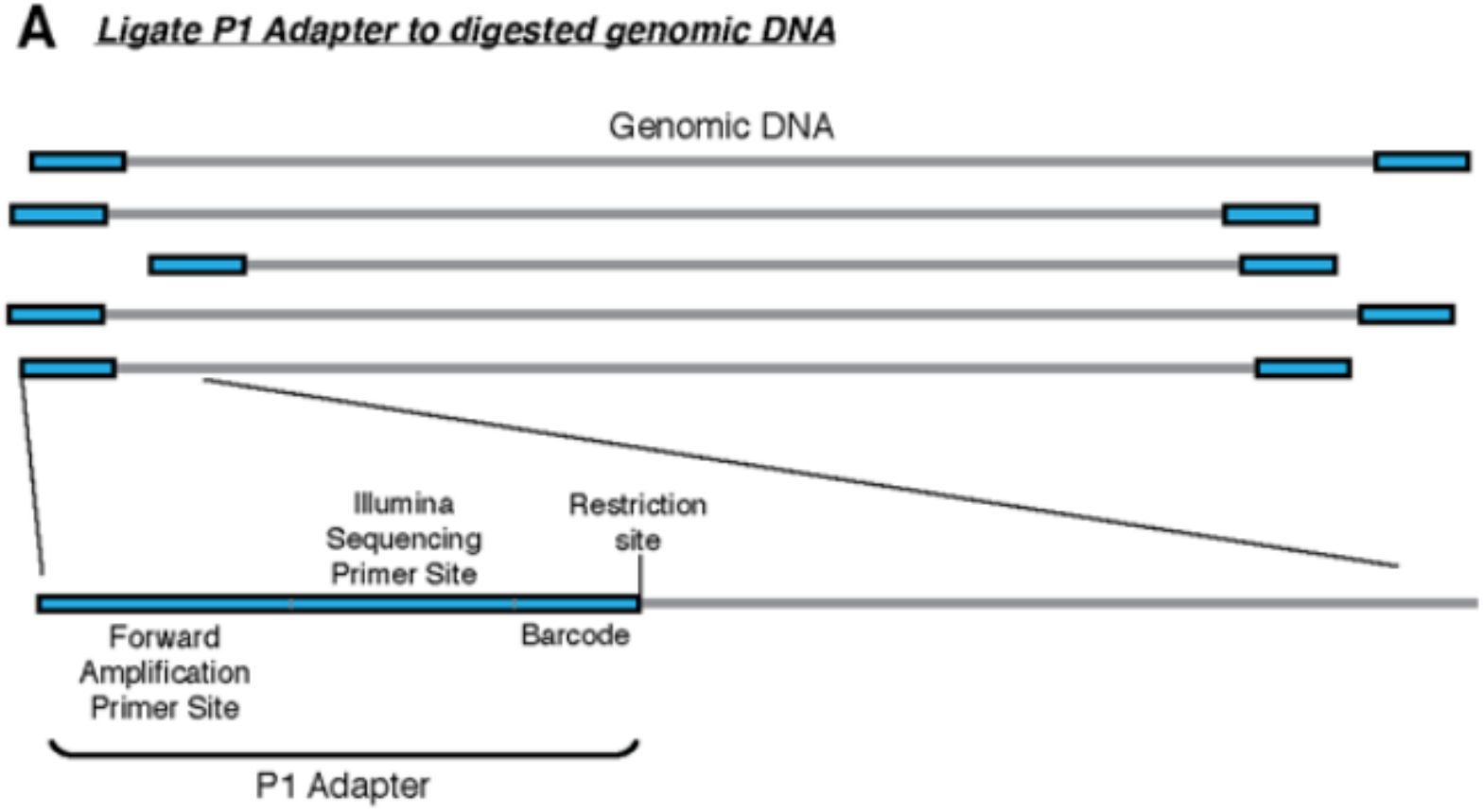
Nathan A. Baird[1☯], Paul D. Etter[1☯], Tressa S. Atwood[2], Mark C. Currey[3], Anthony L. Shiver[1], Zachary A. Lewis[1], Eric U. Selker[1], William A. Cresko[3], Eric A. Johnson[1*]

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

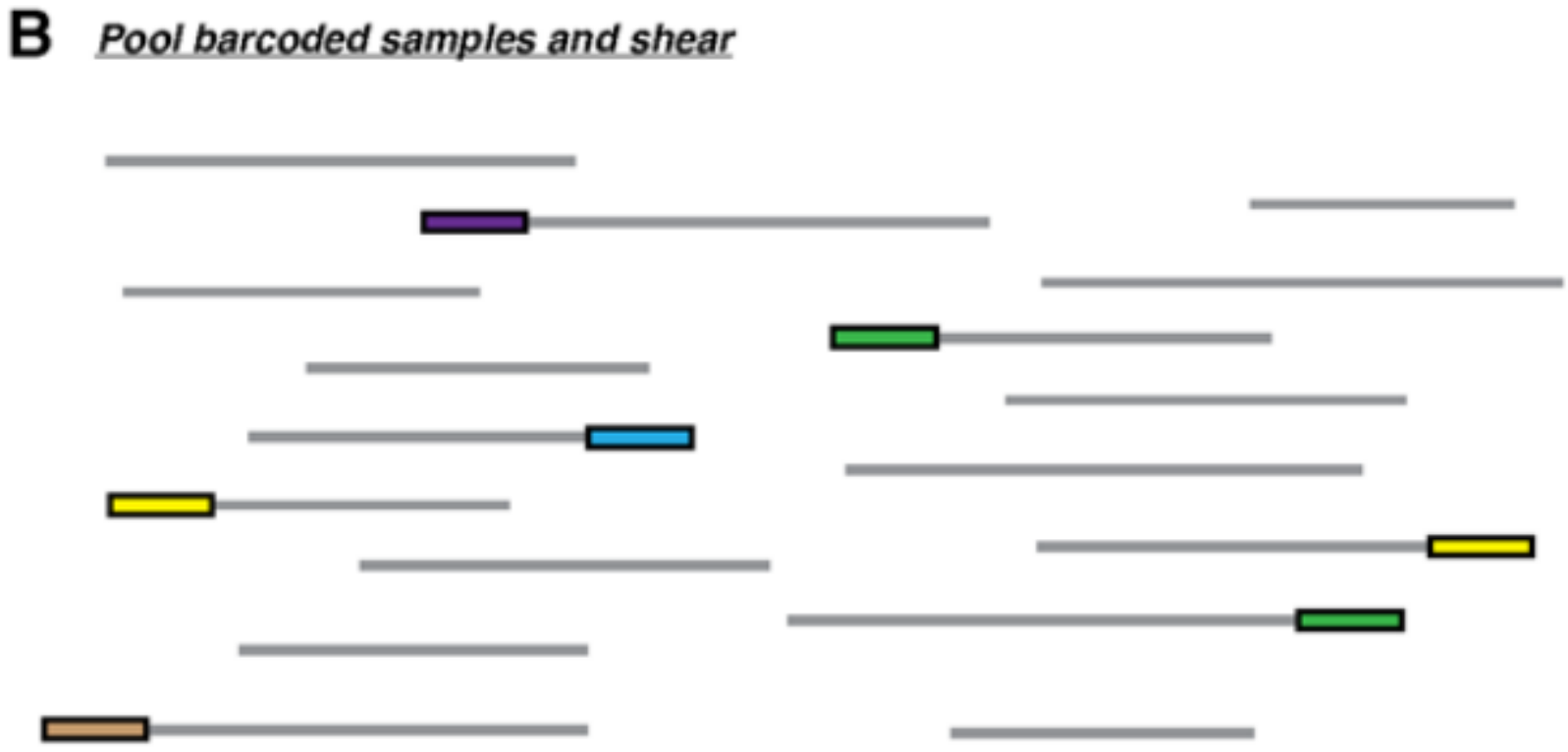
RAD-seq

- RAD fragments with high-throughput sequencing (Illumina)
- SNP identified by sequence polymorphism and site disruption
- Can be used with or without reference genome

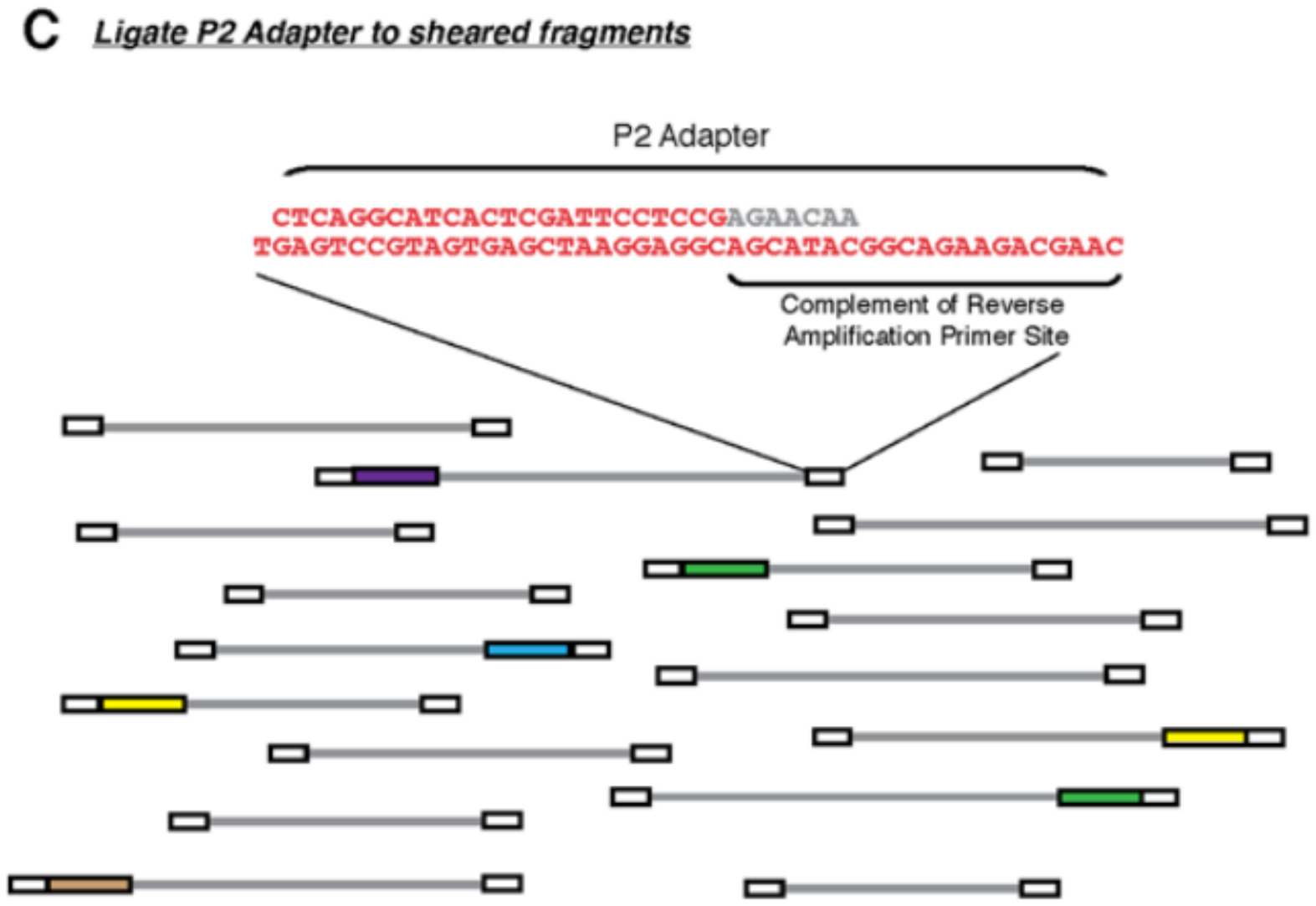
RAD-seq - Baird 2008



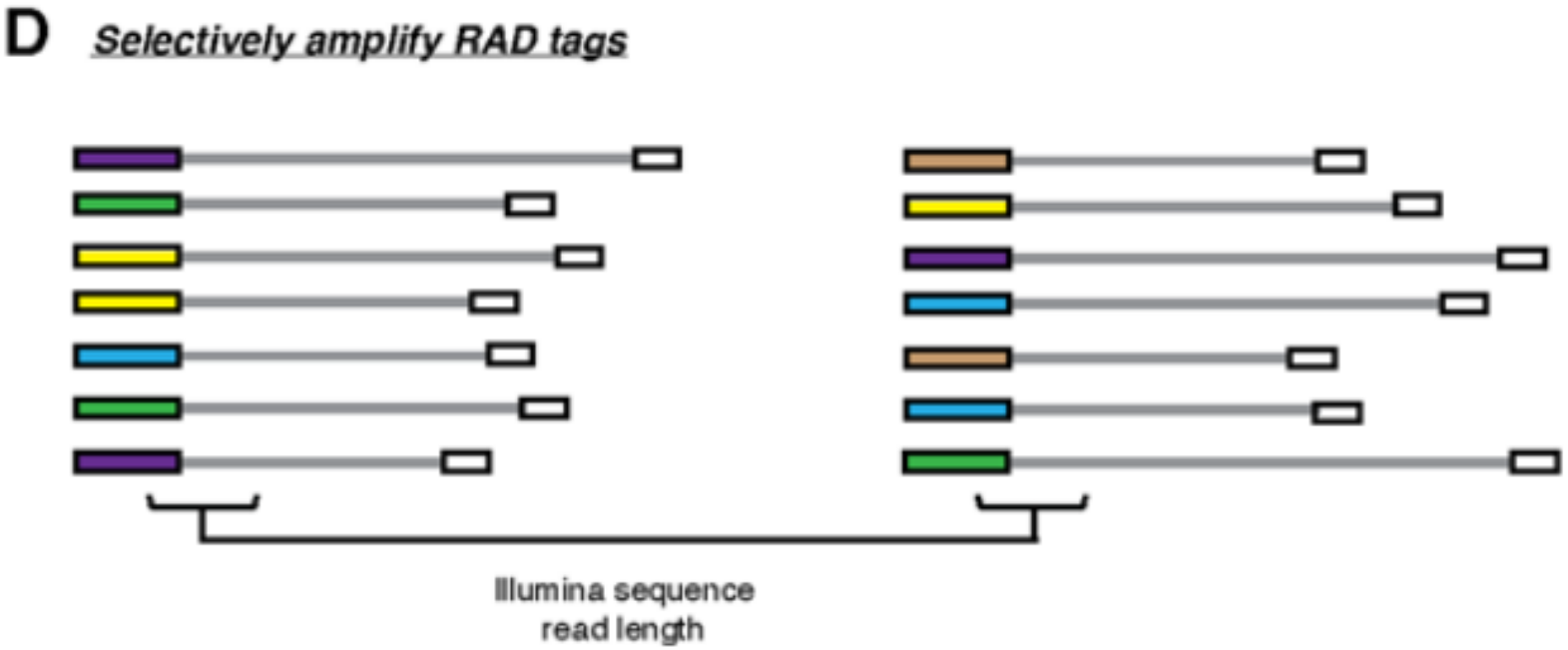
RAD-seq - Baird 2008



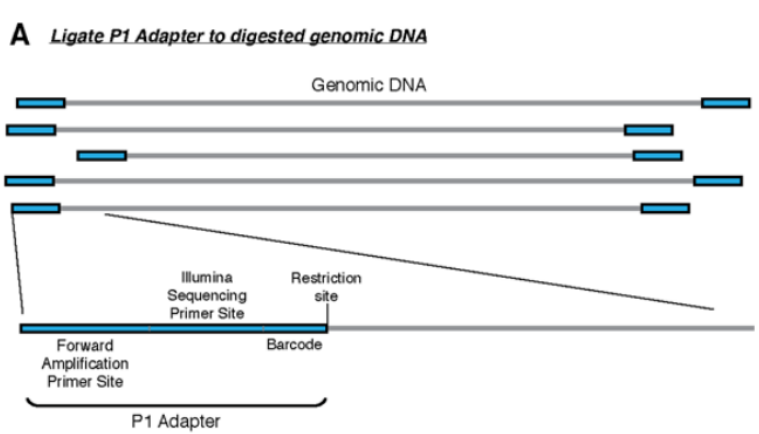
RAD-seq - Baird 2008

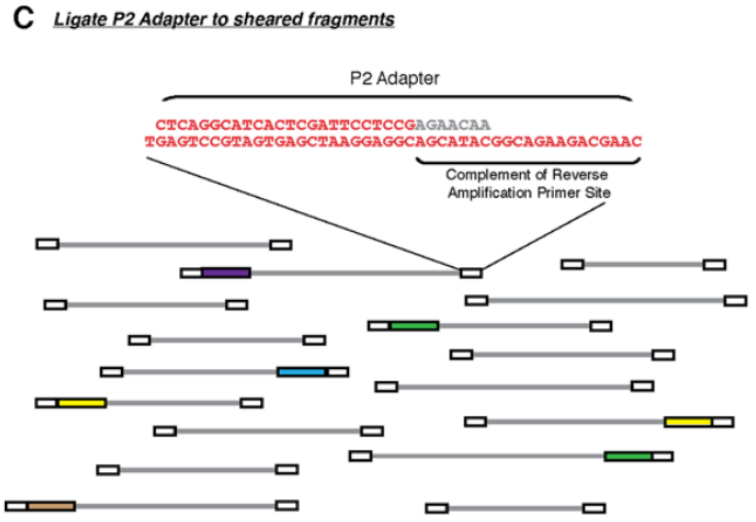


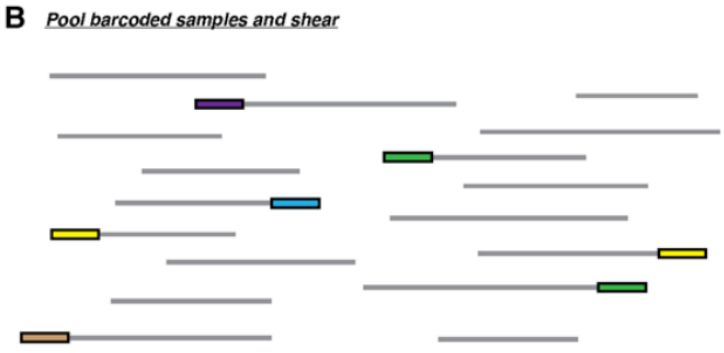
RAD-seq - Baird 2008

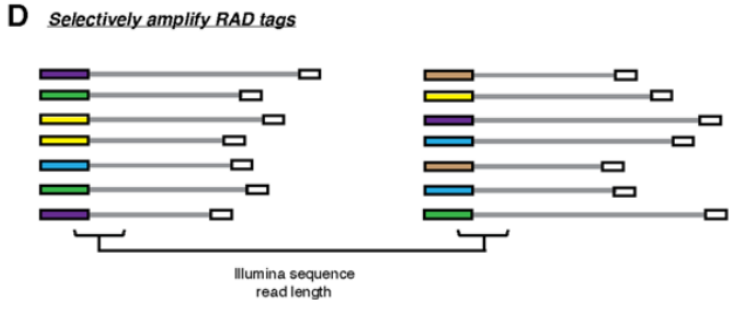


RAD-seq - Baird 2008







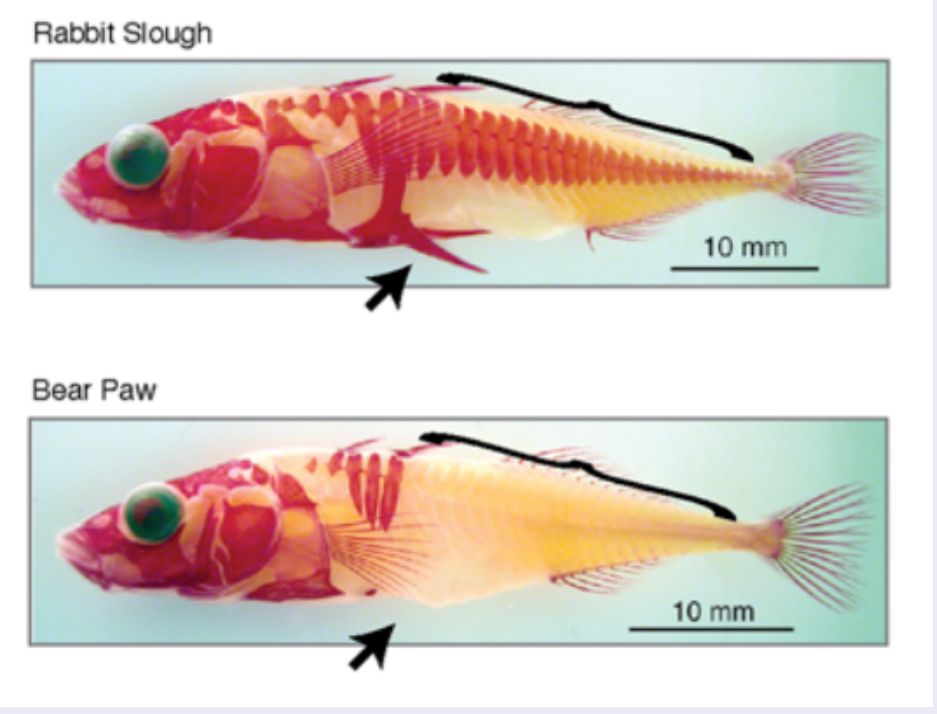


RAD-seq - Baird 2008

Demonstration

- Discover 13000 SNP in threespine stickleback and in *Neurospora*
- Barcoding system for multiplexing
- Marker density can be tuned by the choice of restriction enzyme

Threespine stickleback



Population genomics of parallel adaptation - Hohenlohe 2010

A major paper

OPEN ACCESS Freely available online

PLoS GENETICS

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

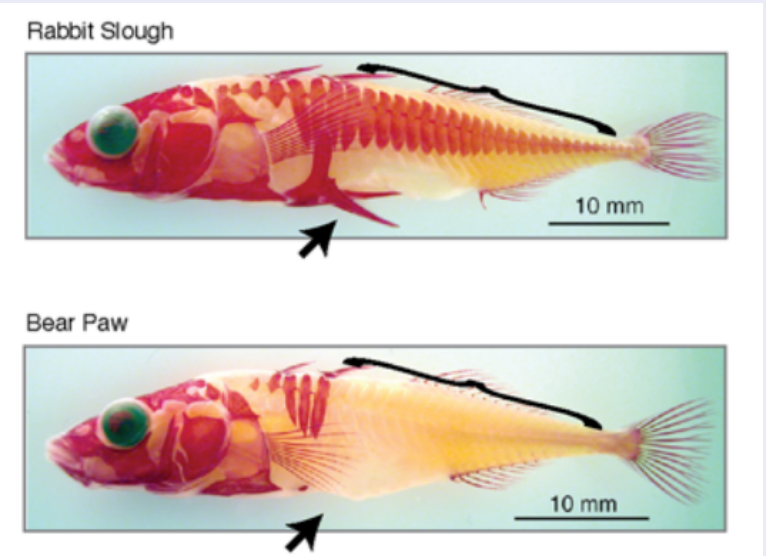
Paul A. Hohenlohe[1☯], Susan Bassham[1☯], Paul D. Etter[2], Nicholas Stiffler[3], Eric A. Johnson[2], William A. Cresko[1*]

Method

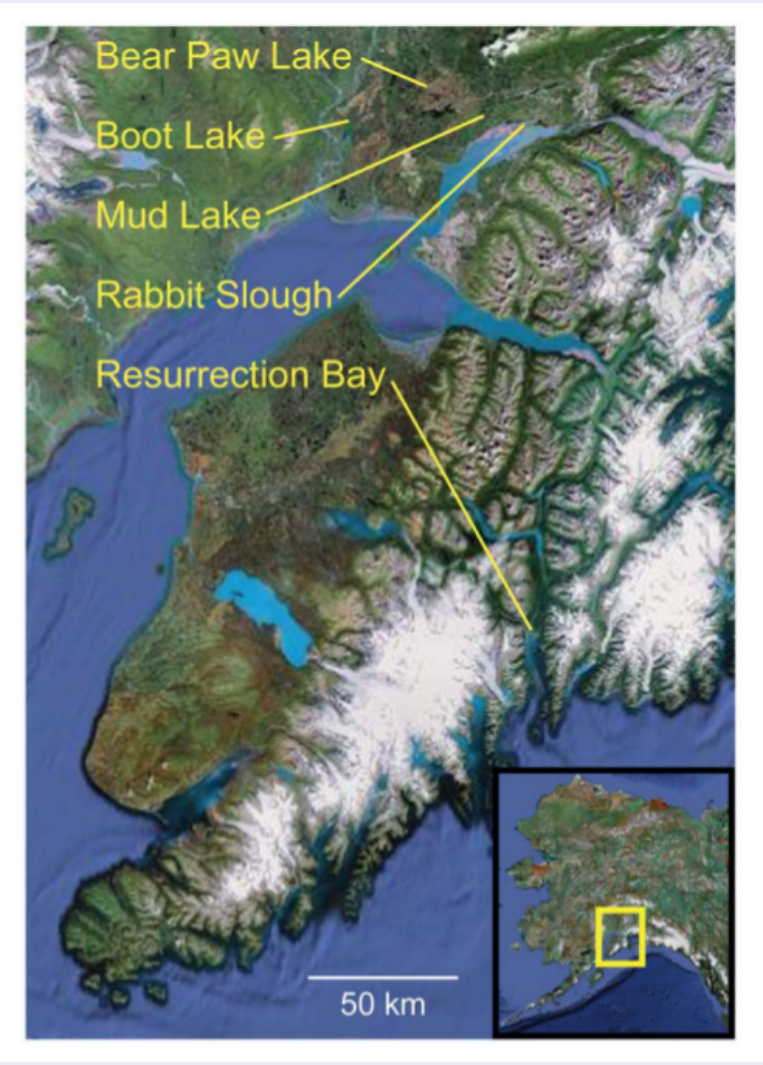
- Model: threespine stickleback
- Comparison of 3 freshwater and 2 marine populations
- 20 individuals per population, individual barcodes
- Single reads (not paired ends)

Population genomics of parallel adaptation - Hohenlohe 2010

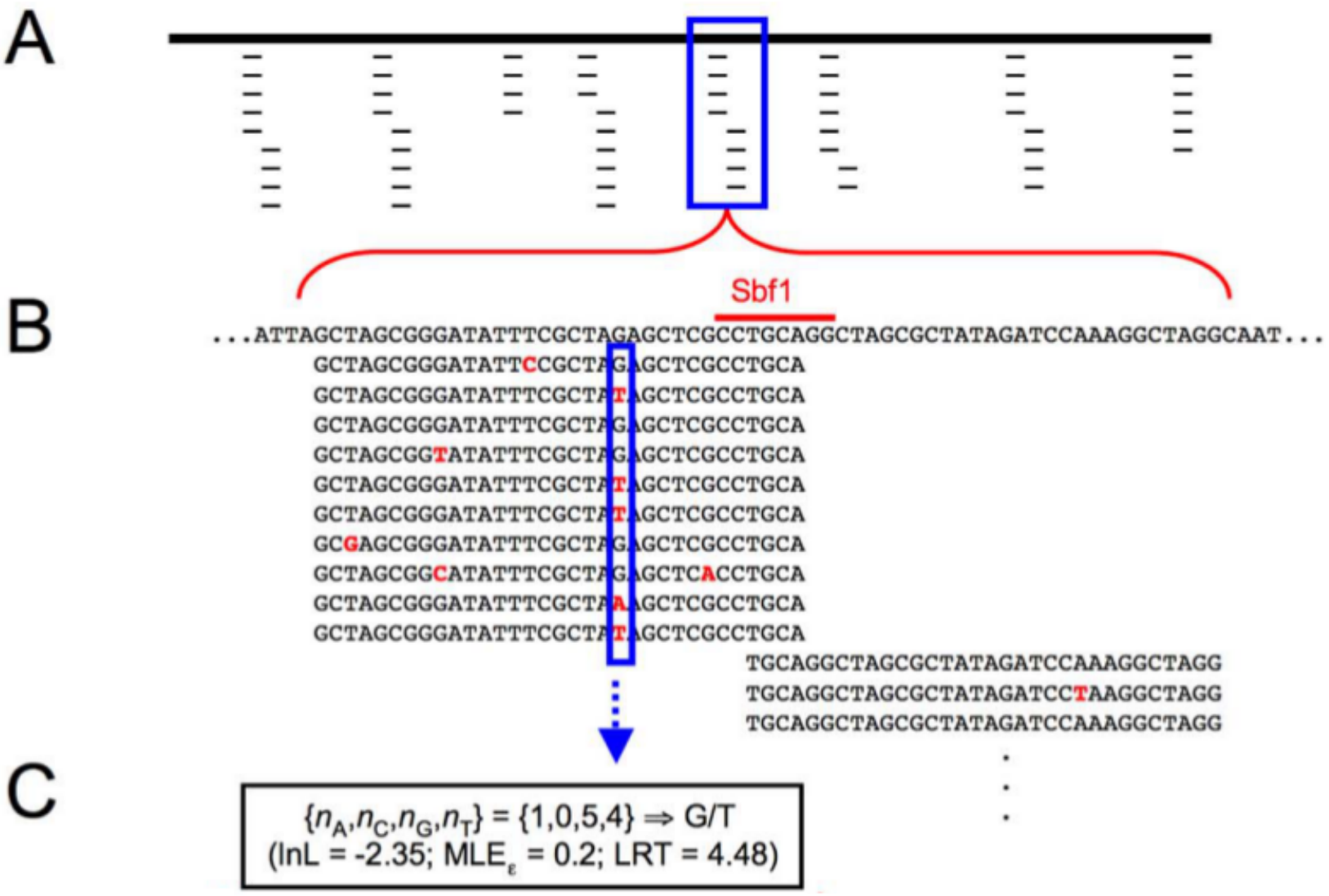
Gasterosteus aculeatus



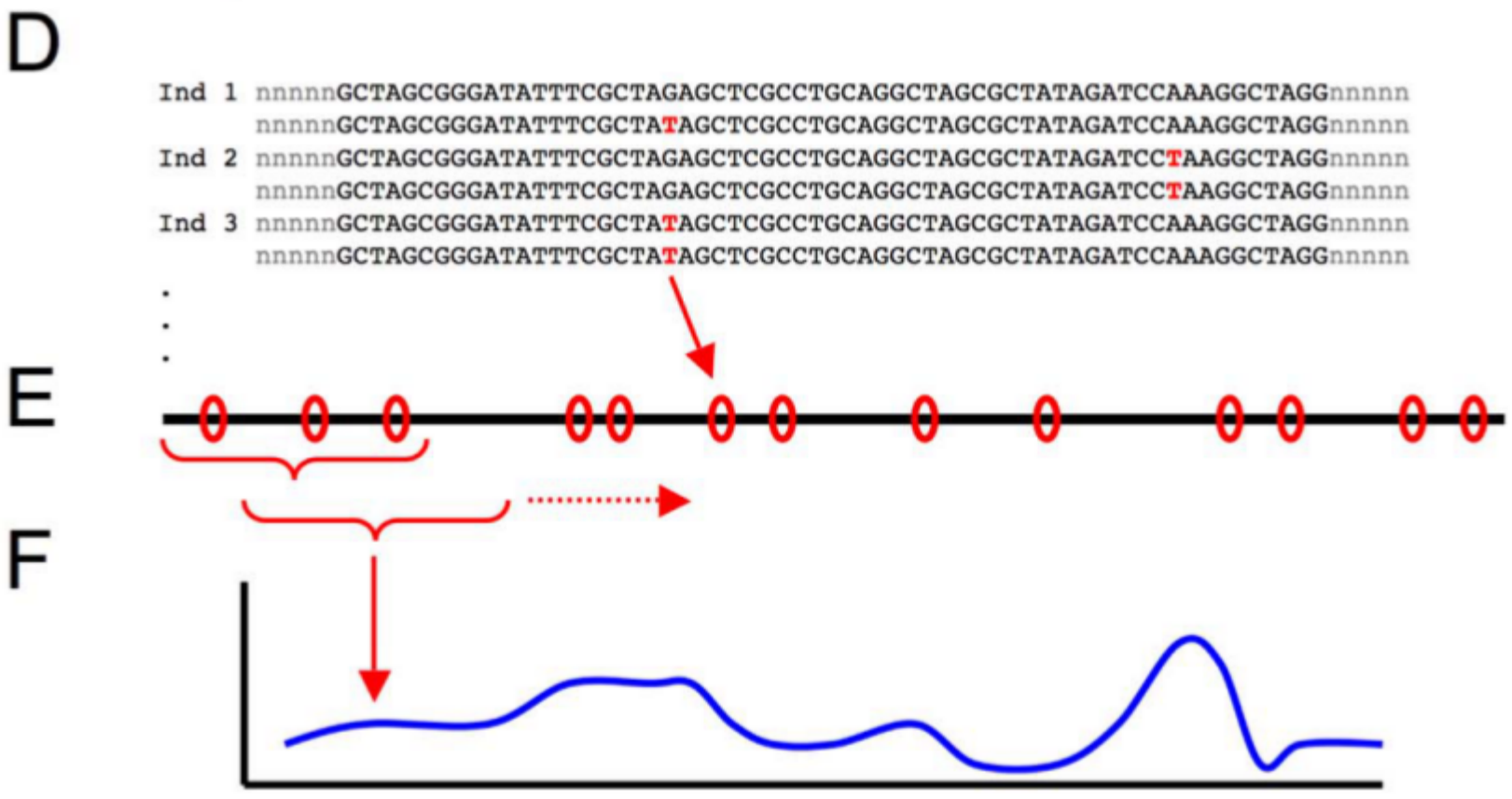
Locations



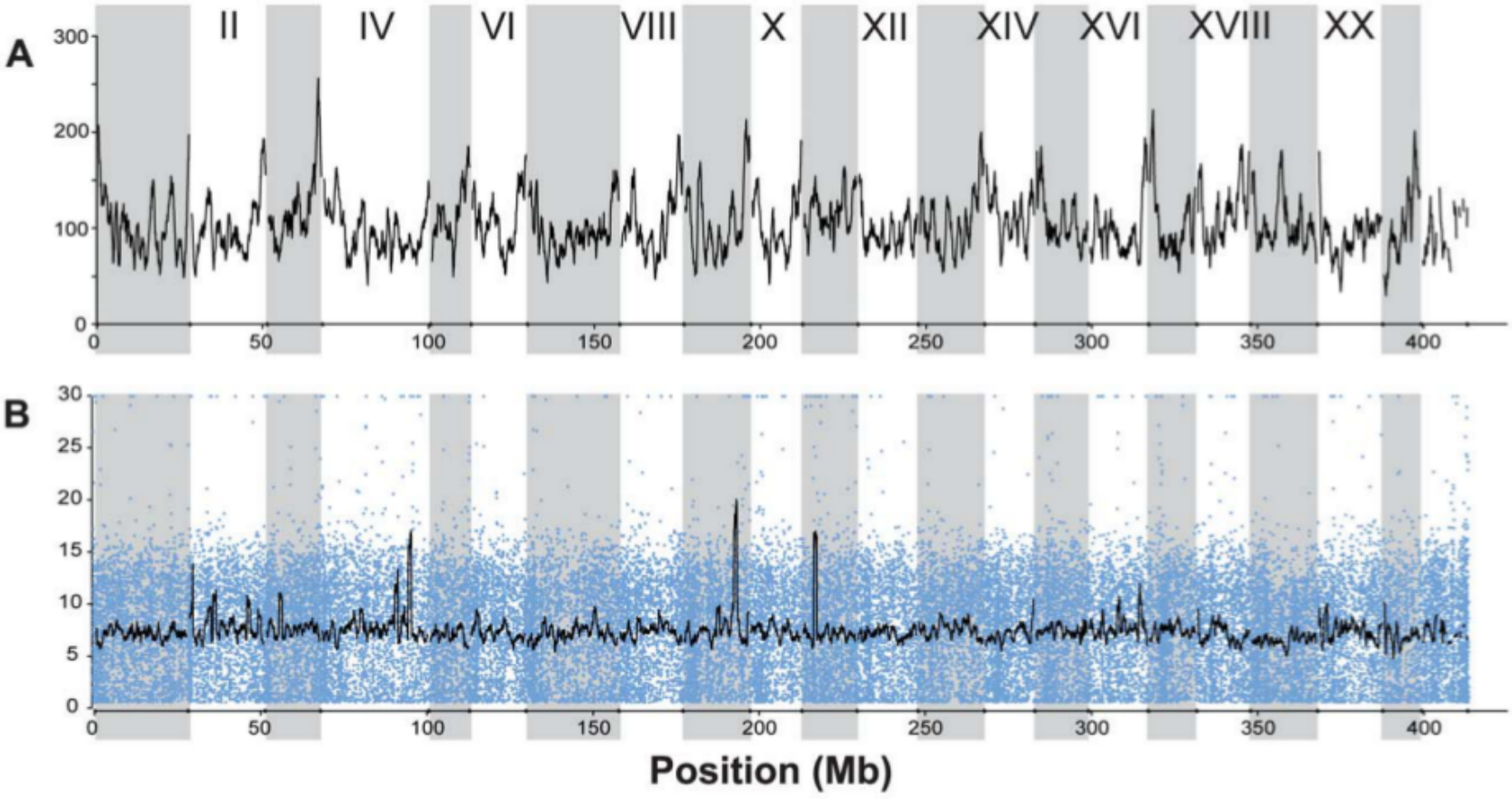
Hohenlohe 2010



Hohenlohe 2010

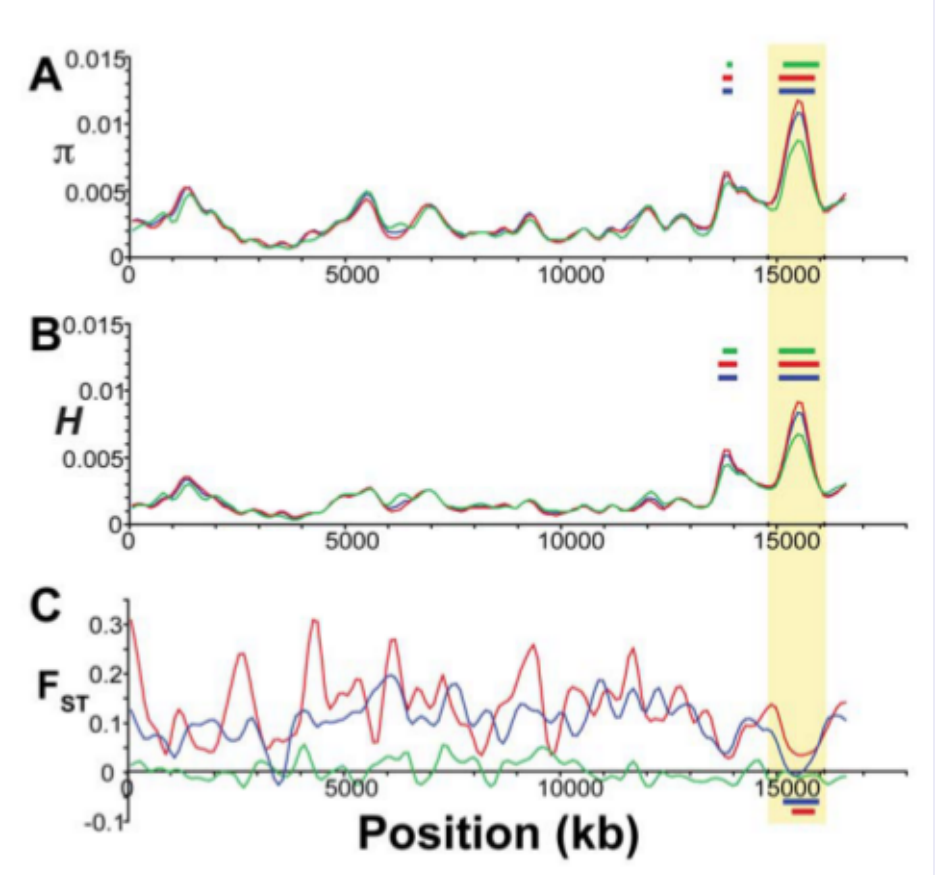


Hohenlohe 2010 - Genome profiles



- A: number of RAD tags per 1Mb
- B: Coverage per RAD per individual in one run (16 individuals - black line is average)

Hohenlohe 2010

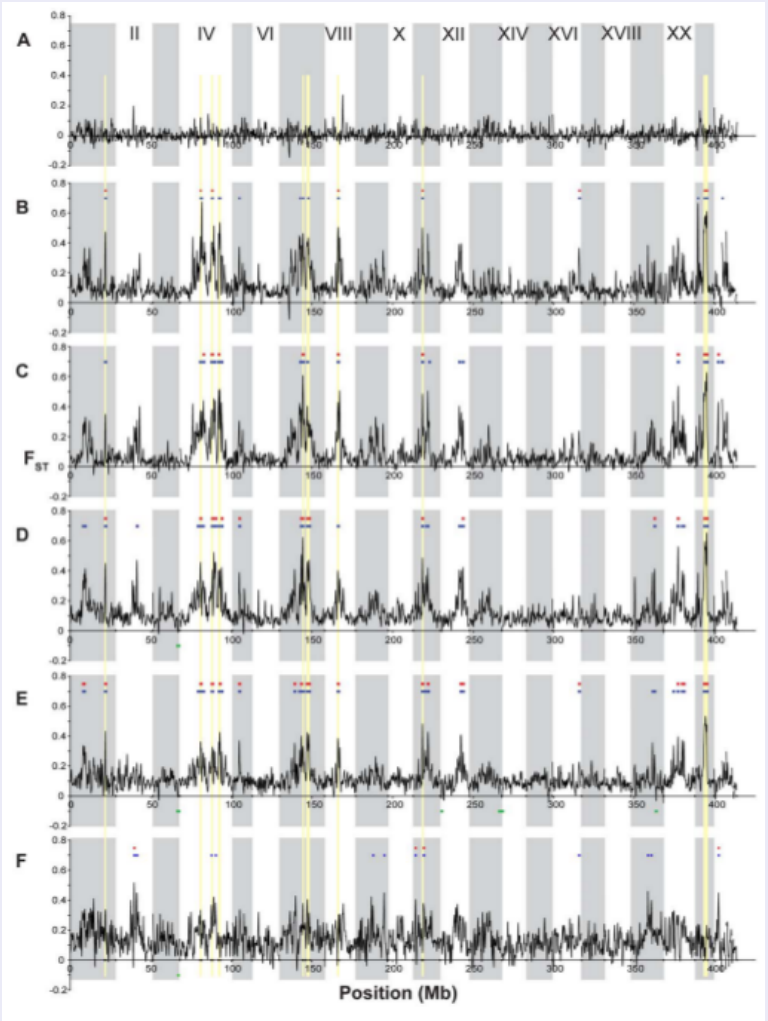


Evidence for balancing selection

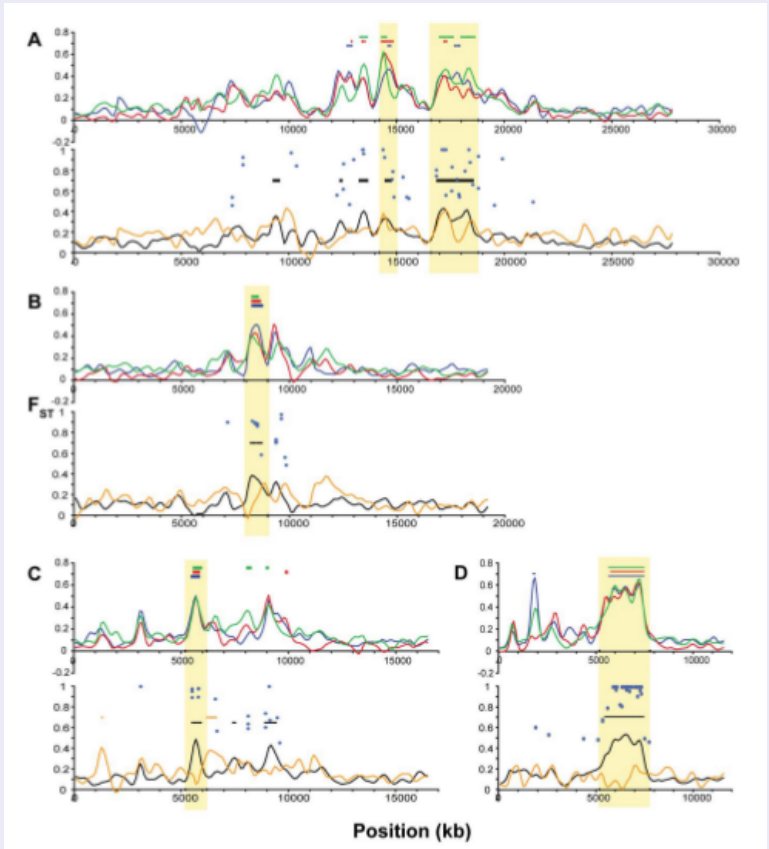
- A: Nucleotide diversity, B: heterozygosity across all five populations (blue), three FW (red) or two SW (green)
- C: Fst between FW and SW (blue), among FW (red) and among SW (green)
- Horizontal bars shows regions of significantly elevated or reduced values on the profile

Hohenlohe 2010

Genome-wide differentiation among populations



Differentiation among SW and FW, zoom on LG



Hohenlohe 2010

Highlights

- RAD-seq on natural populations, 45000 SNPs in 100 individuals
- Barcoded samples
- Genome profiling, kernel smoothing and permutation testing

But note that...

- Genome available
- Single reads

What is paired-end RAD-seq?

Etter 2011

OPEN ACCESS Freely available online

PLOS one

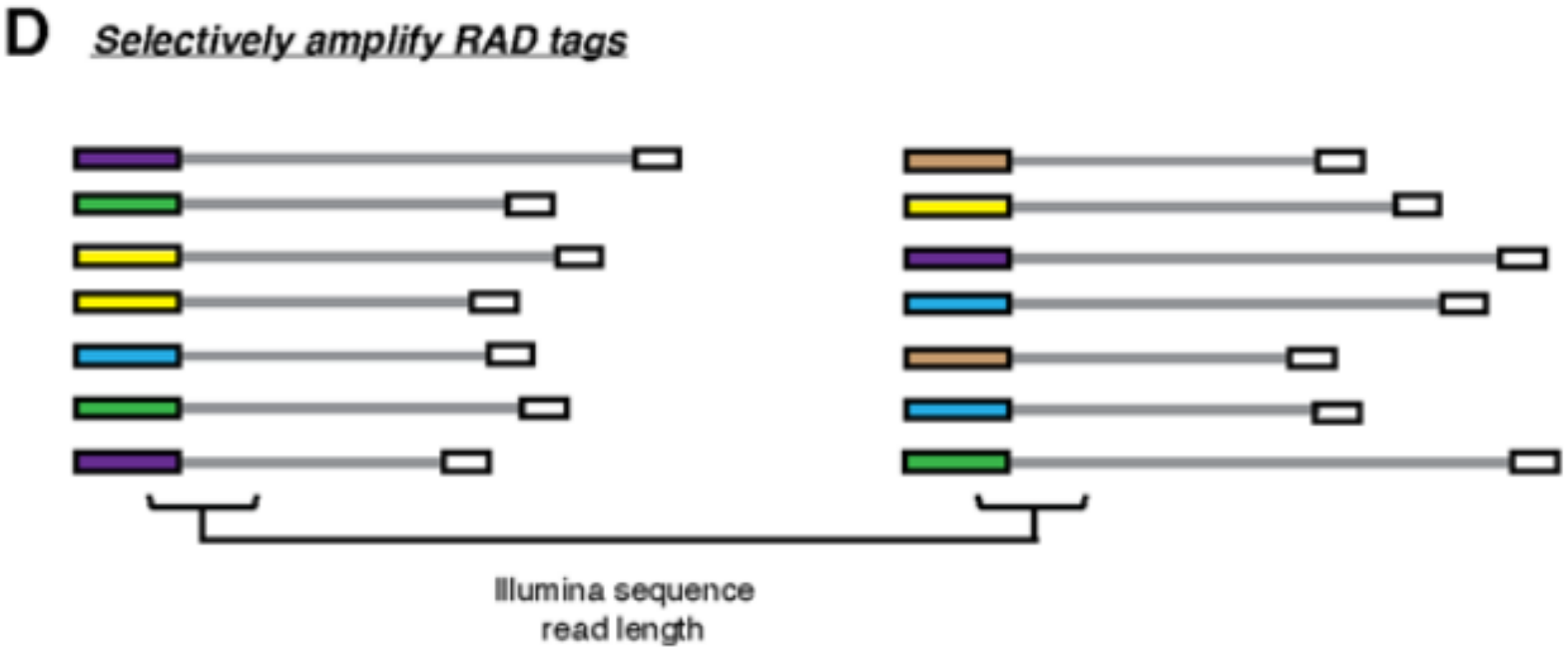
Local *De Novo* Assembly of RAD Paired-End Contigs Using Short Sequencing Reads

Paul D. Etter[1], Jessica L. Preston[1], Susan Bassham[2], William A. Cresko[2], Eric A. Johnson[1]*

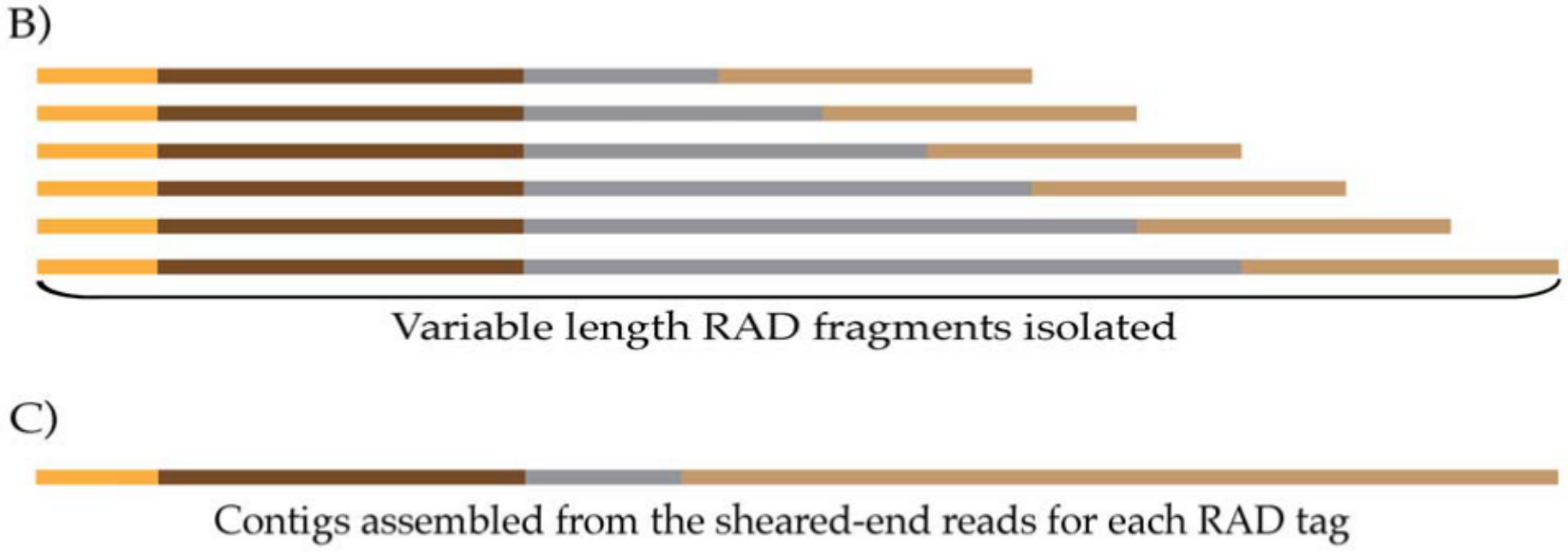
Method

- Paired-end sequencing of RAD fragments to build contigs on the randomly sheared side
- Demonstration with threespine and *E. coli* sequencing
- Up to 5kb contigs with circularization step

Single-reads RAD-seq



Paired-ends RAD-seq



Notes

- The stacked end is useful for high coverage work (SNP calling, allele frequency estimates)
- The echelon end is useful for contig building, but base coverage is lower

What is double-digest RAD-seq?

Peterson et al. 2012

OPEN ACCESS Freely available online

PLoS one

Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species

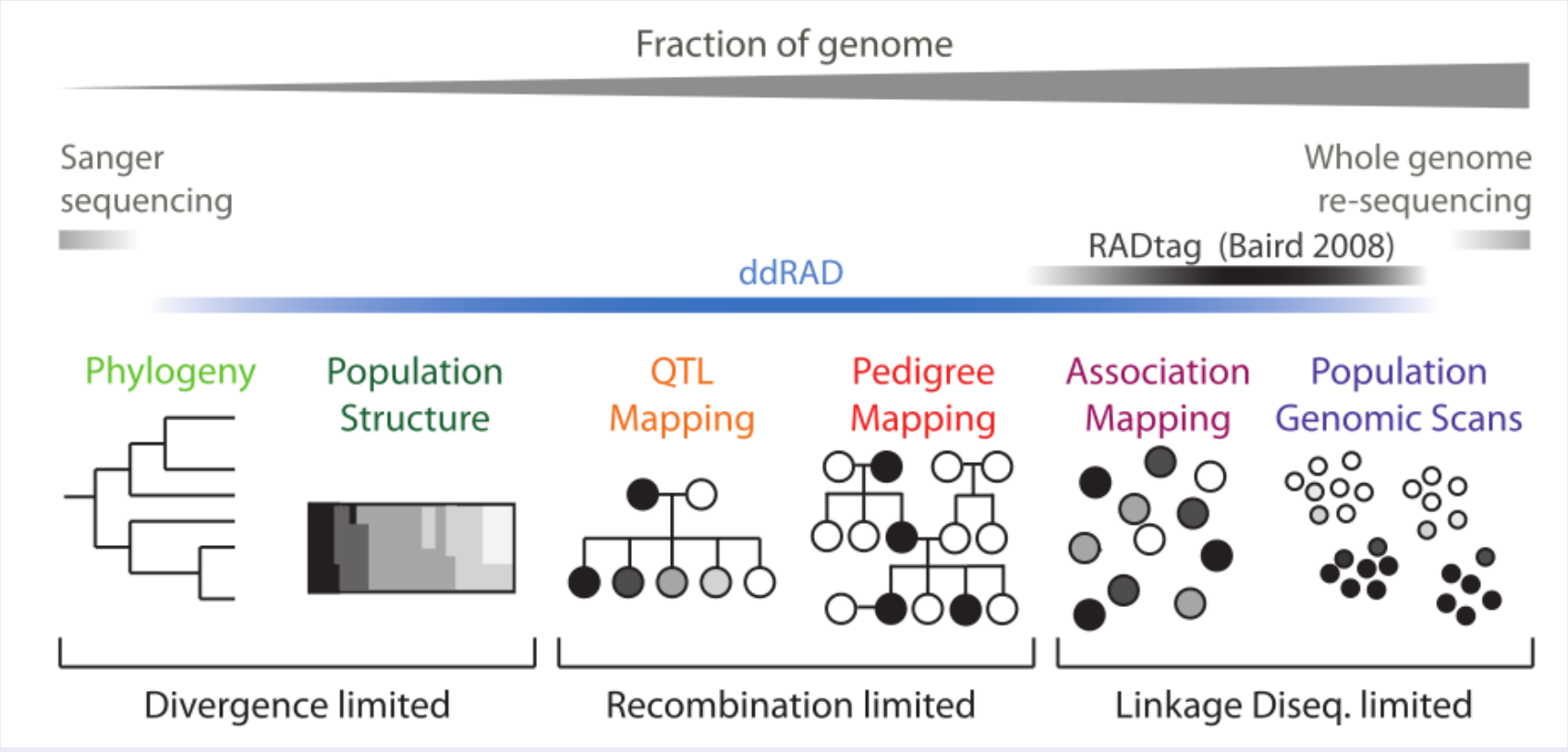
Brant K. Peterson*, Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, Hopi E. Hoekstra

Method

- Two enzyme double digest followed by precise size selection
- Library contains only fragments close to target size
- Read counts across regions are expected to be correlated between individuals

Peterson 2012

Double digest RAD tag



What is paired-end double RAD?

Bruneaux et al. 2013

Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach

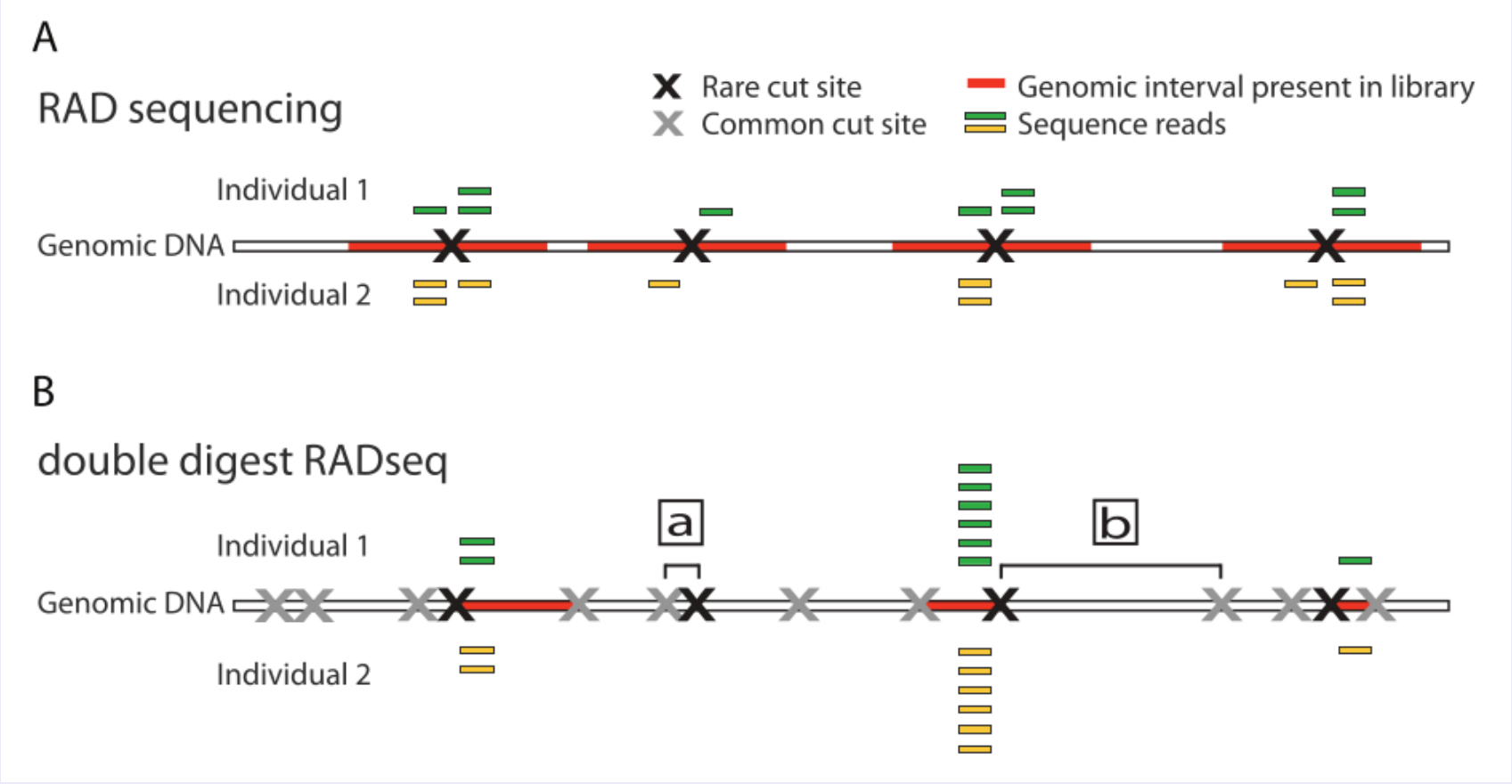
MATTHIEU BRUNEAUX,*[1] SUSAN E. JOHNSTON,*[1] GÁBOR HERCZEG,† JUHA MERILÄ,† CRAIG R. PRIMMER* and ANTI VASEMÄGI*‡

Method

- Two enzyme double digestion
- Paired-end sequencing after size-selection
- You will hear more about it soon (see practicals)

Uses of RAD tags

From Peterson 2012



There are also some potential issues...

Crucial to understand the potential biases of RAD tags

- PCR-duplicates
- Individual vs pool genotyping for allele frequencies
- Comparison SNP vs microsat

Needs for (bio)informatic analyses

- Specific pipelines have been developed (STACKS, Rainbow, dDocent)
- Usual NGS tools can be used
- Again, the most important is to understand what is going on

Conclusion

In a nutshell

- RAD tags: versatile method of genome complexity reduction
- RAD-seq: large scale discovery of SNPs, affordable
- Useful for both model and non-model organisms
- Just a tool: the downstream analyses are still your expertise

Before starting the practicals

Any questions ?

Practical plan

Complete analysis, from raw reads to results

- Reproduce results from Bruneaux et al. 2013
- From raw reads to final results
- Skipping some steps

Cherry picking some other analyses?

- If we have time
- You can tell me what you would be interested in

General workflow (1/2)

RAD-seq experiment

1. DNA extraction (pooling?)
2. Digestion and adapter ligation (simple or double RAD? Barcodes?)
3. Size selection
4. Sequencing (single reads? double reads?)

Read processing

- Demultiplexing and barcode removal
- Quality control / trimming

General workflow (2/2)

de novo assembly or mapping back

- Consensus sequences from *de novo* assembly
- Mapping back the reads to consensus (or to reference genome)

Variant calling and allelotyping

- Variant calling (filtering? likelihood? bayesian?)
- Genotyping / allelotyping

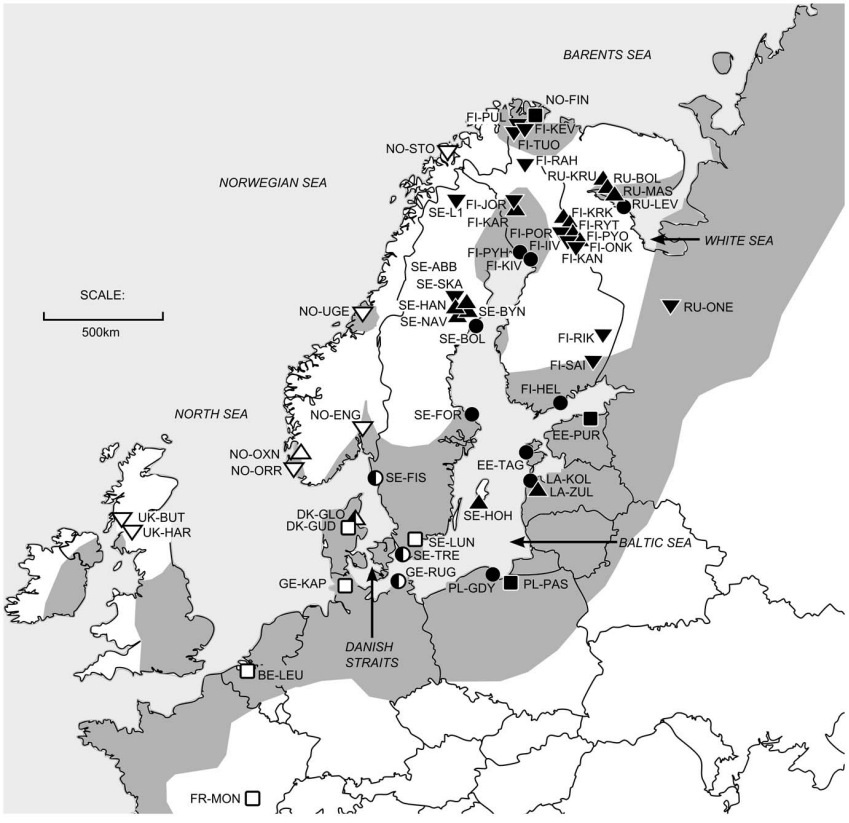
Downstream analysis

- Genome scans
- QTL mapping
- Phylogenies
- etc. . .

Nine-spined stickleback in Fenno-Scandia

Nine-spined stickleback

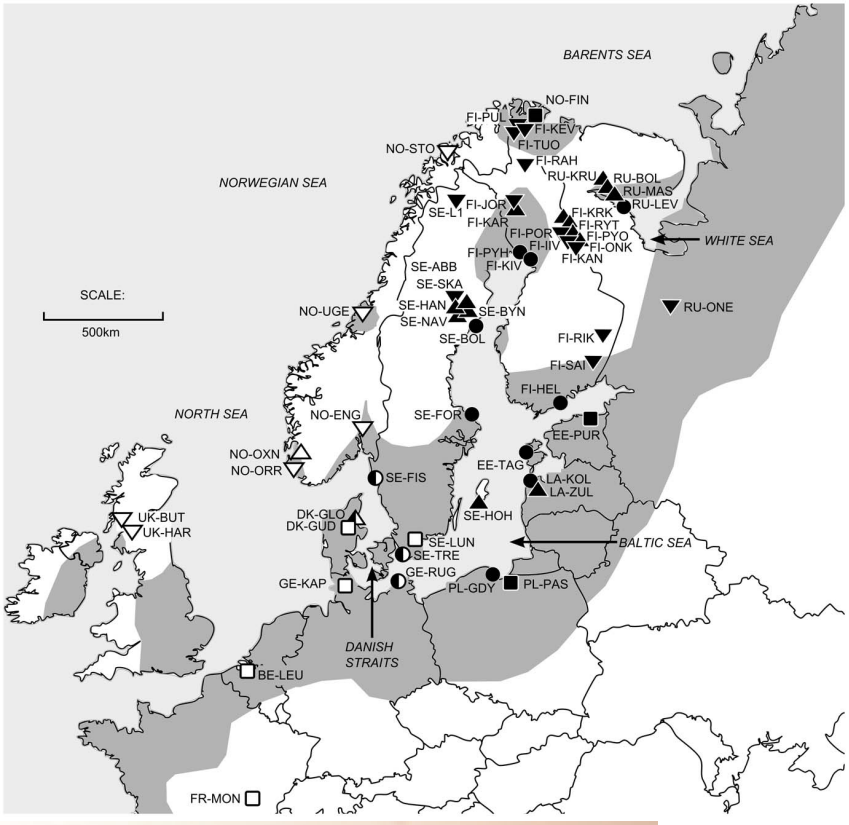
- Versatile fish species
- Recent history of recolonization (Teacher 2011)
- Evidences of local adaptation (Prof. Merilä's group)



Nine-spined stickleback in Fenno-Scandia

Nine-spined stickleback

- Versatile fish species
- Recent history of recolonization (Teacher 2011)
- Evidences of local adaptation (Prof. Merilä's group)



RAD tag experiments

Context and approach

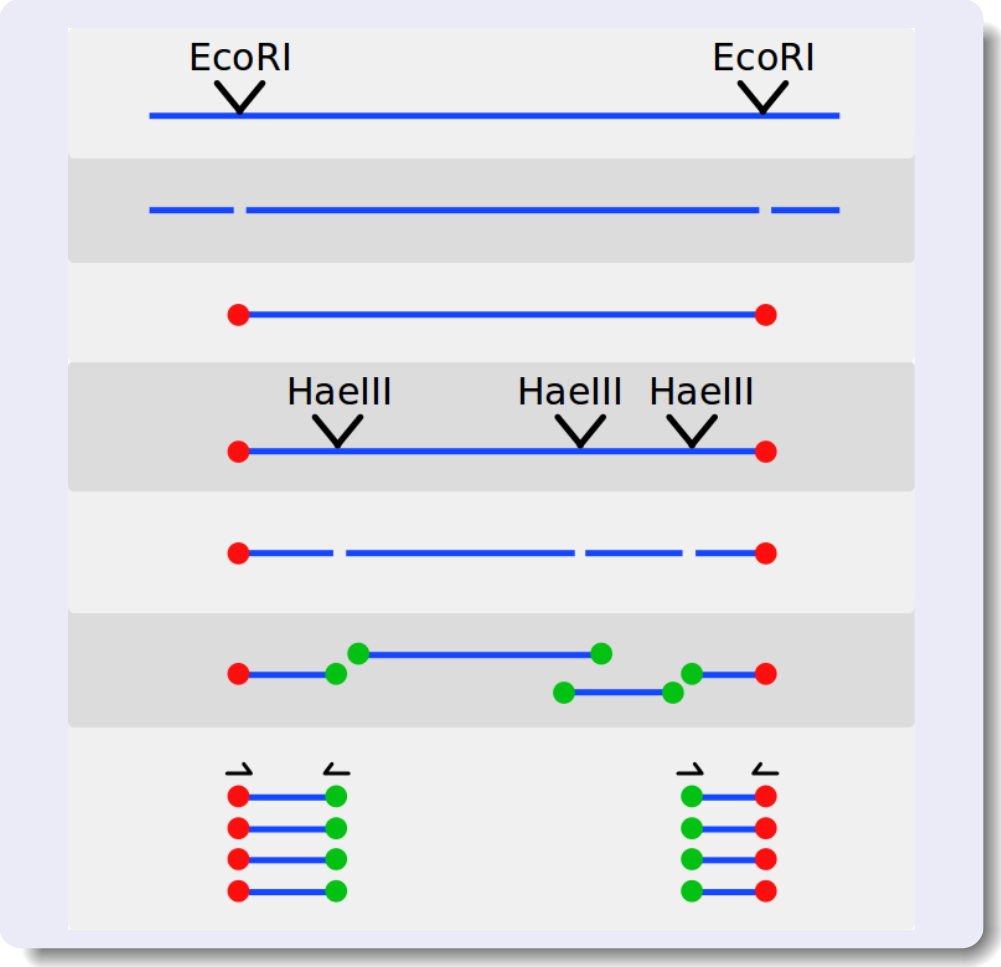
- No transcriptomic or genomic resources
- But three-spined stickleback genome available
- Aim: mapping the genetic differences associated with local adaptation

RAD tag experiments

Context and approach

- No transcriptomic or genomic resources
- But three-spined stickleback genome available
- Aim: mapping the genetic differences associated with local adaptation

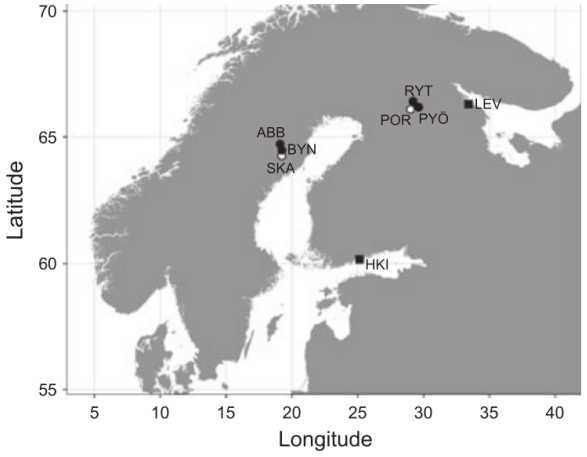
- paired-end, double RAD tag approach
  - DNA of 48 individuals pooled per population
  - Digestion by EcoRI and HaeIII
  - Purification, amplification and size-selection



Results (1/2)

Low coverage issues

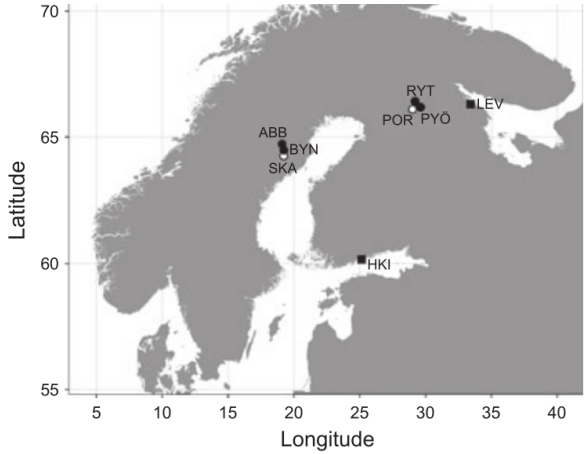
- SNP coverage lower than expected
- Populations pooled by habitat type



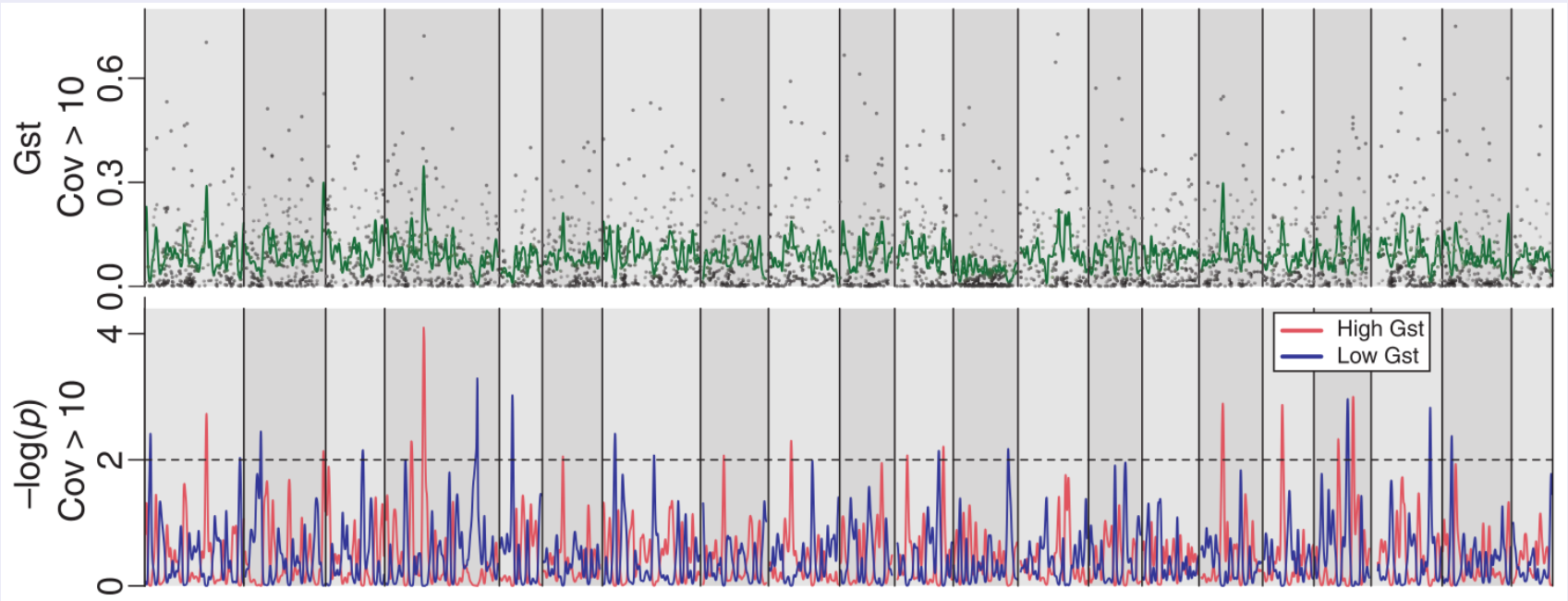
Results (1/2)

Low coverage issues

- SNP coverage lower than expected
- Populations pooled by habitat type



Kernel smoothing and permutation tests



Results (2/2)

Identification of candidate genes

- Annotations from the three-spined stickleback genome
- Gene Ontology information

Results (2/2)

Identification of candidate genes

- Annotations from the three-spined stickleback genome
- Gene Ontology information

GO enrichment tests

Global function	GO category	GO enrichment test *P*-value	Number of genes involved
Kidney development	Metanephric mesenchyme development	0.037	5
	Glomerular capillary formation	0.023	2
	Renal sodium ion absorption	0.05	1
Immunity	Immune system development	0.029	7
	Positive regulation of T-cell differentiation	0.039	3
	Cellular response to macrophage colony-stimulating factor stimulus	0.02	2
	Cytokine binding	0.038	3
MAP kinase regulation	Regulation of protein modification process	0.046	8
	Positive regulation of kinase activity	0.05	5
	Positive regulation of ERK1 and ERK2 cascade	0.033	3

During the first part of the practicals

Simple scripts can be used also

- This is one thing I want to show during the practical
- The objective is to get a good grip and a good feeling/understanding about the data with simple, straightforward methods
- Once we are comfortable, we can choose to apply more complex methods which rely on third-party scripts
- It is important to understand what the third-party scripts do!