RAD sequencing: from raw sequences to a RAD assembly

Mark Ravinet
Twitter: @Mark_Ravinet
Email: mark.ravinet@bioenv.gu.se
University of Gothenburg
CeMEB Marine Genomics Course

RAD sequencing

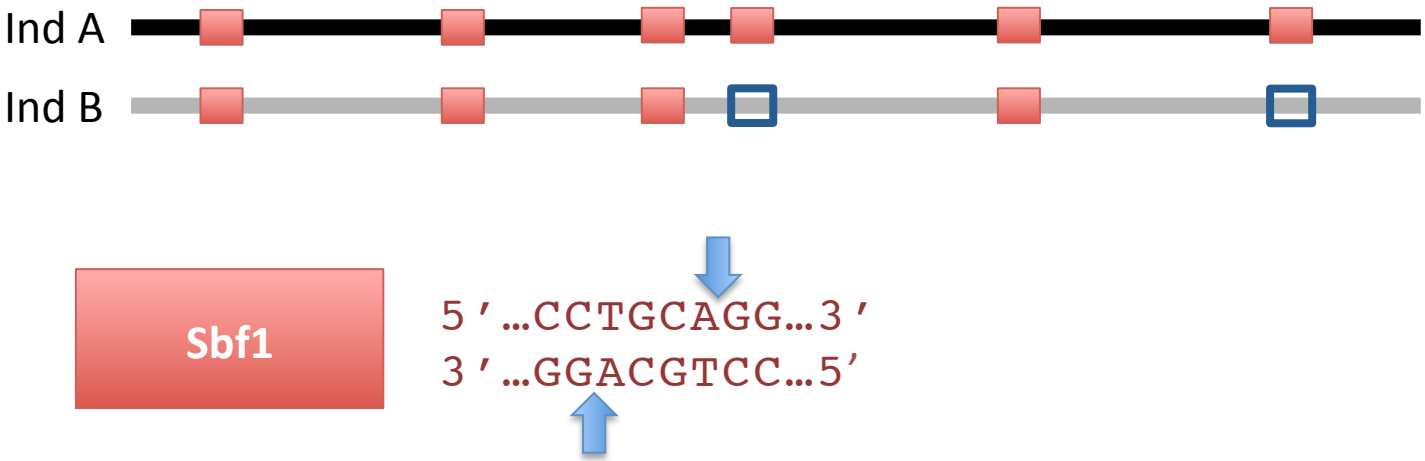
- Genome reduction method
- Genotyping-by-sequencing
- Simple protocol, huge numbers of loci
- Many uses
  - QTL analysis (Baird et al 2008)
  - Genome-wide differentiation (Nadeau et al 2012)
  - Phylogenomics (Emerson et al 2010, Wagner et al 2012)
  - Detecting regions under selection (Hohenlohe et al 2010)

RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - start with whole genome

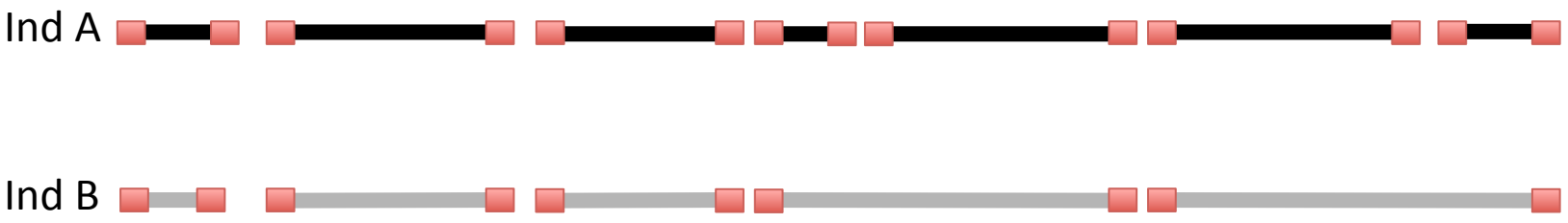
RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - introduce restriction enzyme



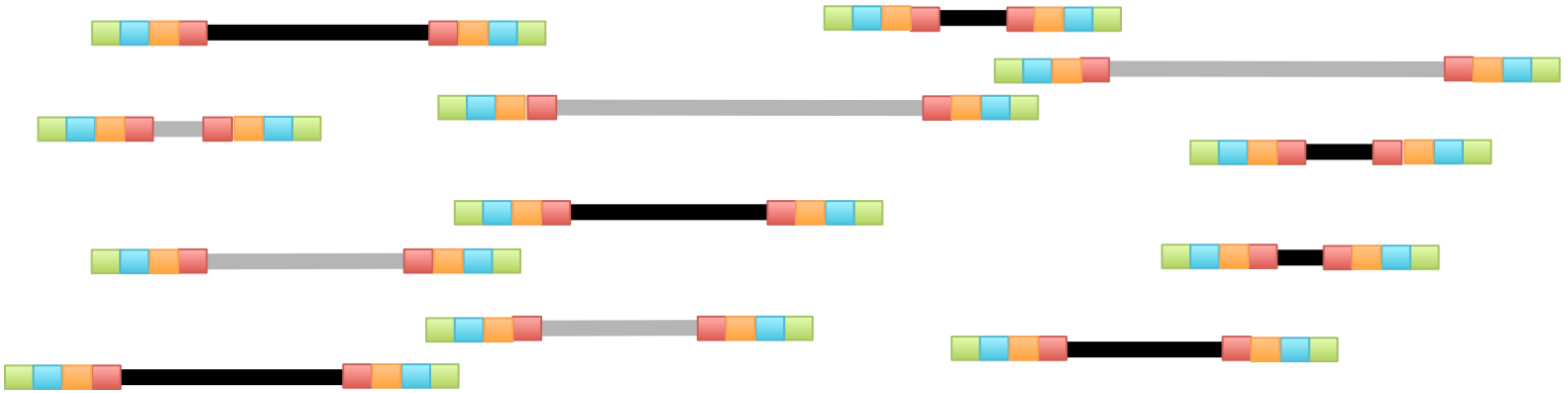
RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - genome is digested into smaller fragments



RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - adapters are ligated and individuals are pooled



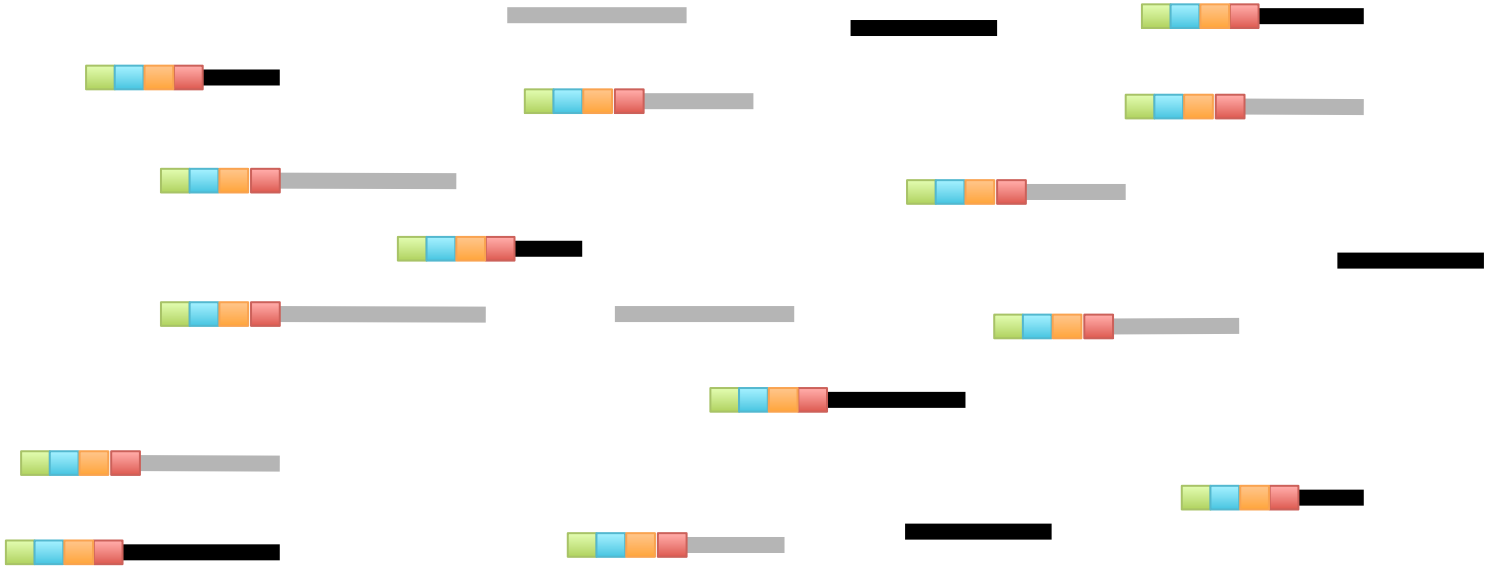
Illumina sequencing primer

Forward amplification primer

Unique barcode

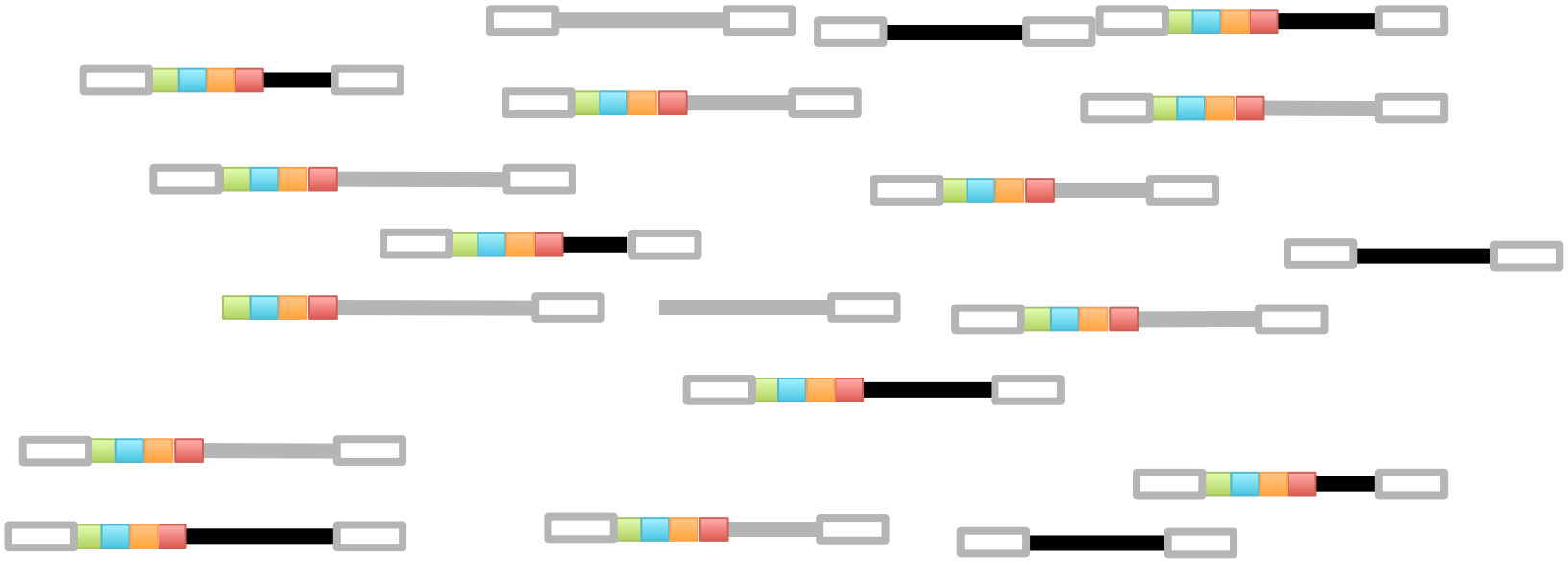
RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - fragments sheared and size selected on gel



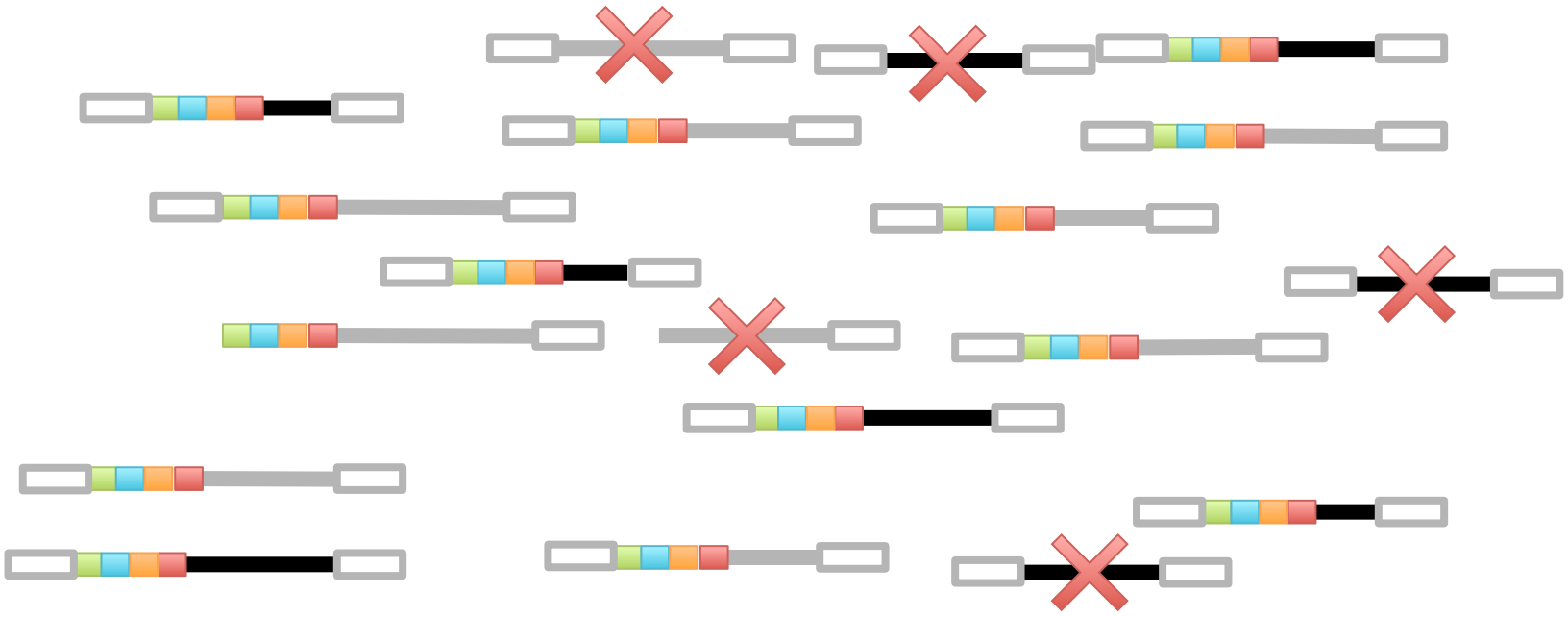
RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - second adapter is ligated



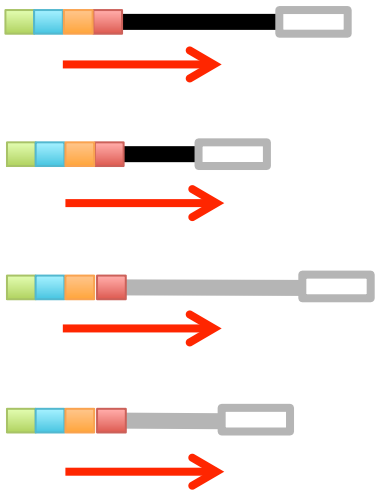
RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - selective amplification of tags with both adapters



RAD-sequencing protocols

- Single digest RADs – Baird et al (2008)
  - simple example with two individuals
  - sequence amplified reads

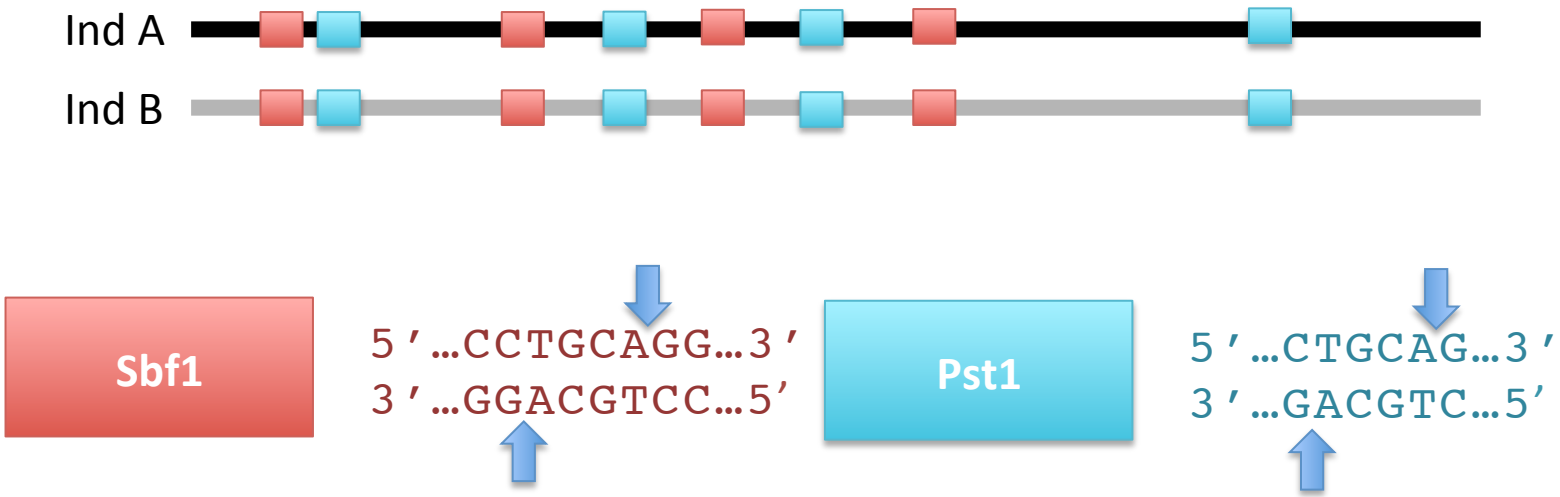


Illumina Hi-Seq 2000
100 bp read length

Sequence 100 bp including barcode and restriction site

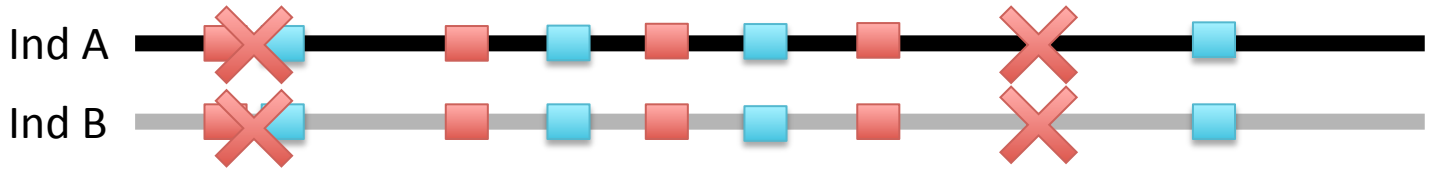
RAD-sequencing protocols

- Double digest RADs – Peterson et al (2012)
  - simple example with two individuals
  - introduce two restriction enzymes



RAD-sequencing protocols

- Double digest RADs – Peterson et al (2012)
  - simple example with two individuals
  - only fragments with each cut retained



  - size selection – too close or too distant removed

RAD protocol comparisons

- How do the two protocols compare?
  - depends on your question!

Single digest

- Single-end or paired-end protocols
- Most established method
- Greater genome coverage
- Shearing is difficult and expensive – requires sonification
- Thousands of loci ideal for genome scans and association studies

Double digest

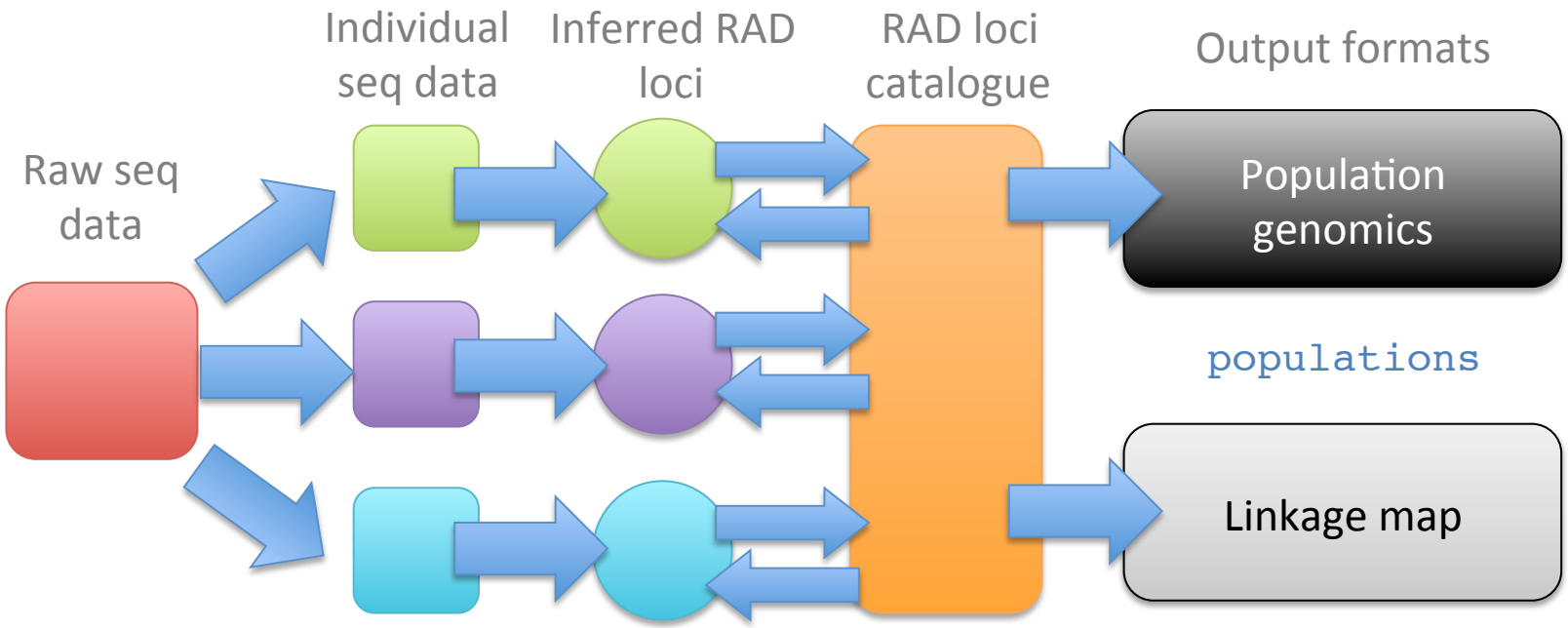
- Lower cost and rapid protocol
- Reduced duplicate sequencing (i.e. neighbouring loci)
- Bias towards recovering same loci in individuals
- Increased coverage at targeted sites
- Lower numbers of loci
- More individuals
- Maybe not suitable for association studies or genome scans

Computational challenges

- How do you go from raw sequences to RAD loci for population genomics?
  1. Identify sequences from individuals (demultiplex)
  2. Identify alleles and loci within individuals
  3. Identify homologous loci amongst individuals
  4. Generate population genomic datasets for downstream analysis
- Align to a reference genome
- Or create a *de novo* RAD loci assembly

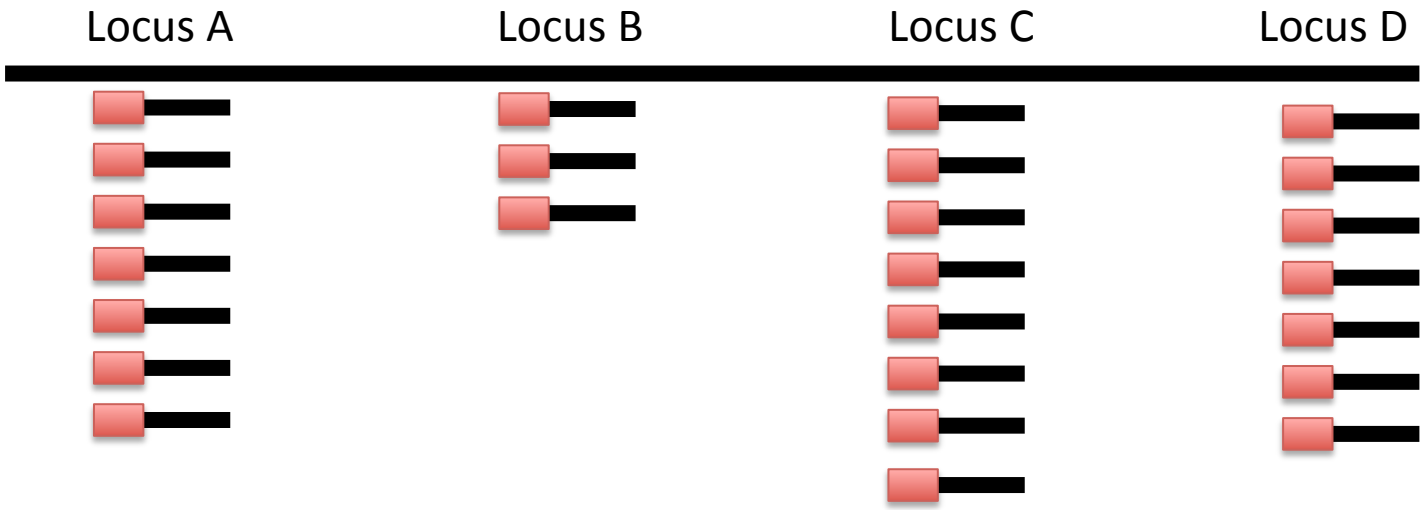
Stacks pipeline

- Julian Catchen and the Cresko laboratory at University of Oregon, USA
- Perl and C based pipeline for RAD-seq analysis



RAD loci: reference genome

- Align reads to a reference genome
- pstacks extracts aligned stacks, identifies RAD loci and calls SNPs



RAD *de novo* assembly

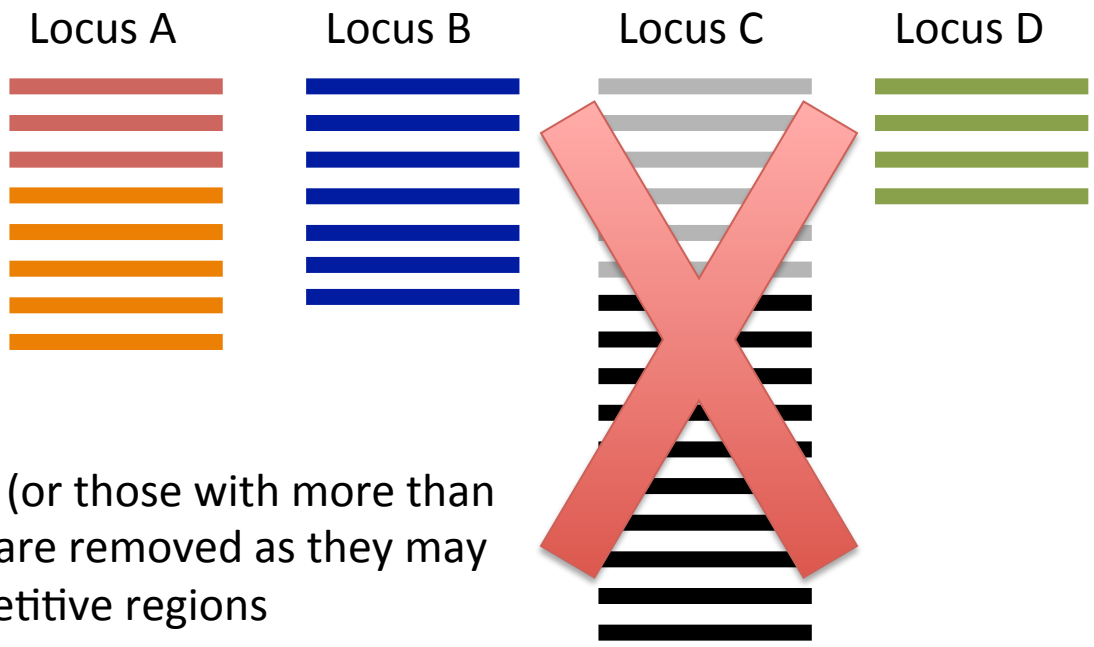
- *de novo* assembly is more complicated
  - identical reads from an individual are combined into stacks in ustacks

RAD *de novo* assembly

- *de novo* assembly is more complicated
  - identical reads from an individual are combined into stacks in ustacks

RAD *de novo* assembly

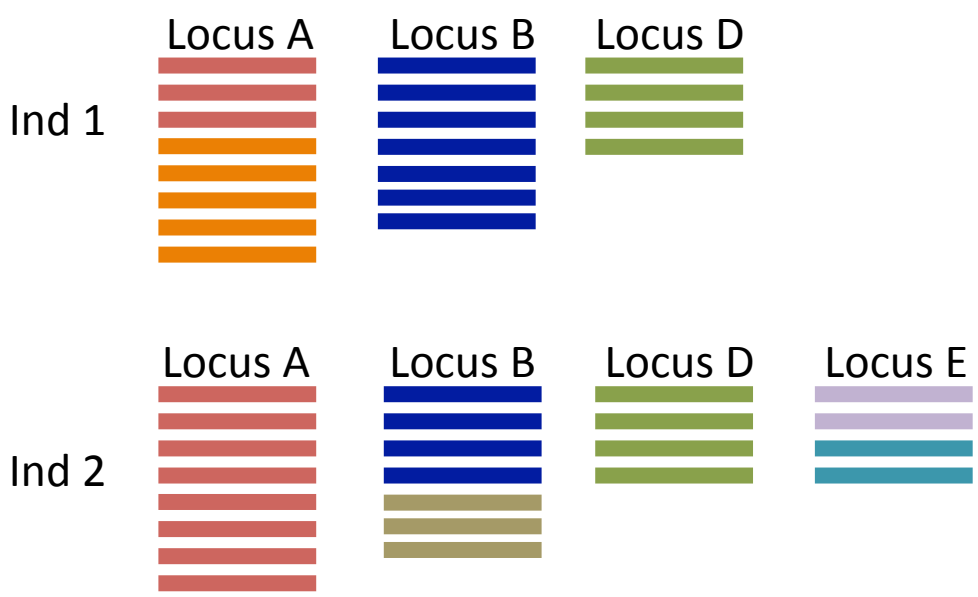
- *de novo* assembly is more complicated
  - ustacks then clusters similar stacks as loci and calls SNPS



Large stacks (or those with more than two alleles) are removed as they may be from repetitive regions

RAD *de novo* assembly

- *de novo* assembly is more complicated
  - cstacks and sstacks produce a RAD loci catalogue by matching loci from different individuals



	Ind1	Ind2
RAD A	A/B	A/A
RAD B	A/A	A/B
RAD D	A/A	A/A
RAD E	-/-	A/B

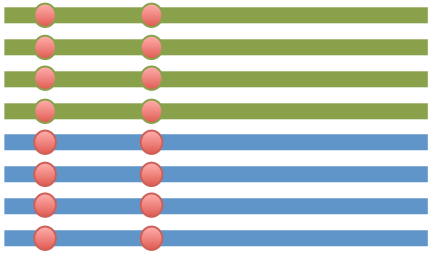
N.B. This example is based on haplotypes at RAD loci not SNPs

Dealing with heterozygosity

- A word of warning – heterozygosity can make *de novo* assembly difficult

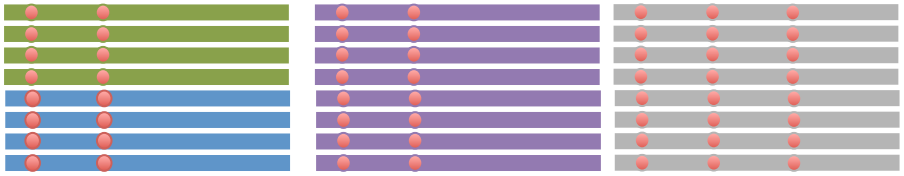
Simple case with 2 nucleotide mismatches allowed

Within an individual

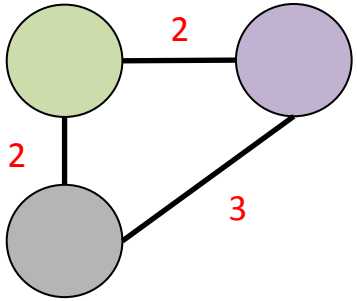


Matching loci among individuals:

Ind 1 – Locus A Ind 2 – Locus M Ind 3 – Locus X



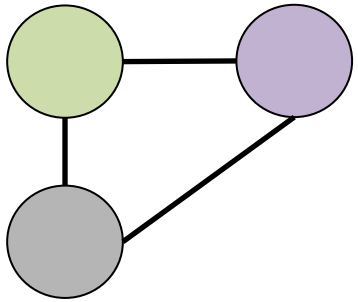
`cstacks and sstacks`



Stacks merges the 3 loci. 4 haplotypes, a total of 3 SNPs at this catalogue locus

Dealing with heterozygosity

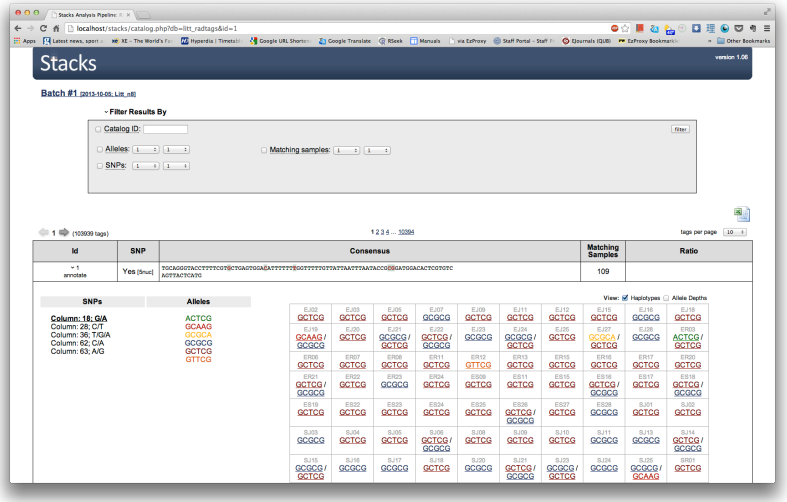
- A word of warning – heterozygosity can make *de novo* assembly difficult
- When choosing nucleotide mismatch params:
  - what is average genome-wide H_O? (individuals)
  - what is H_O amongst populations?
- Always a balance:
  - too low – alleles called as separate loci
  - too high – loci are merged

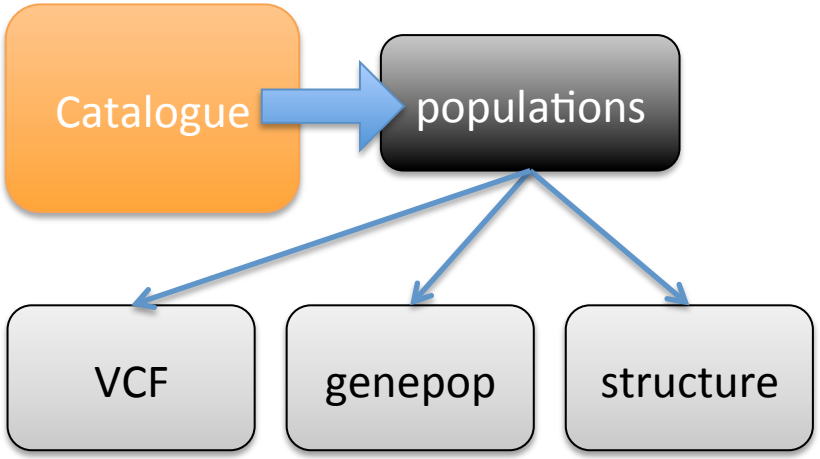


Stacks catalogue

- "At the core of Stacks is the catalogue…" Catchen et al (2013)
- All loci and alleles identified in all individuals

PHP interface – useful for checking loci





An internal reference - flexible

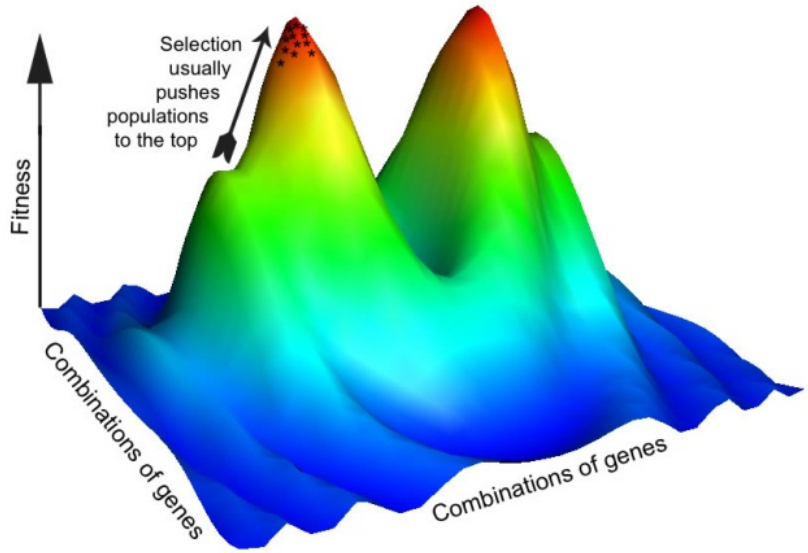
Outlier analysis
Tools for detecting selection and adaptation at the genomic level

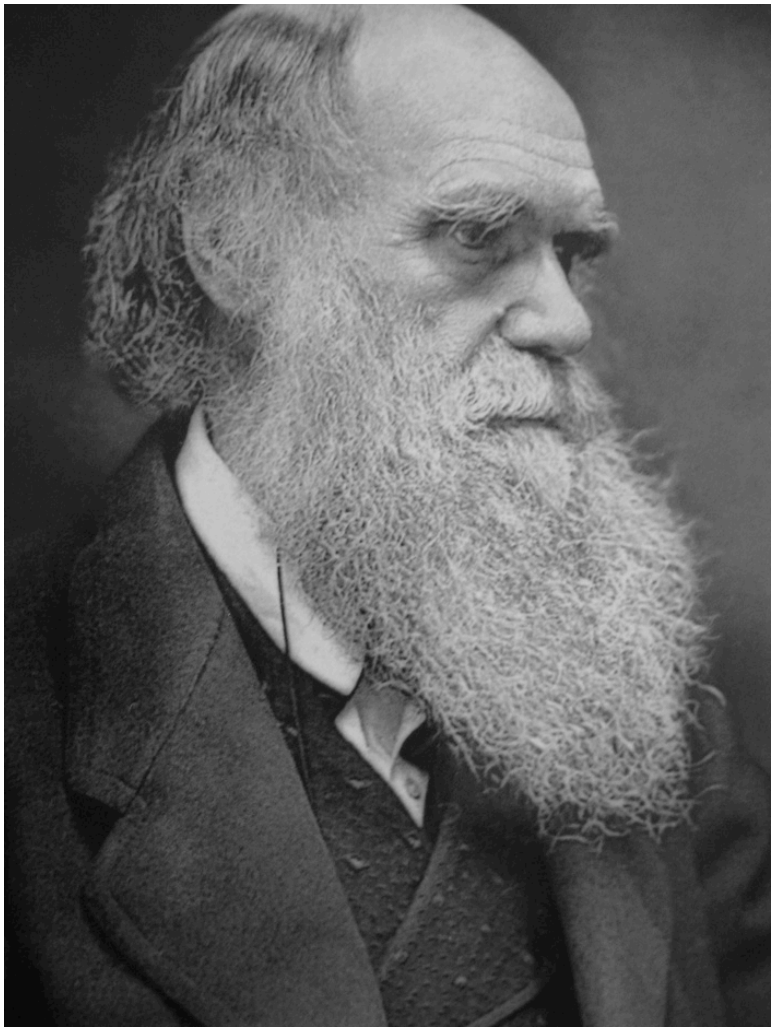
Mark Ravinet
Twitter: @Mark_Ravinet
Email: mark.ravinet@bioenv.gu.se
University of Gothenburg
CeMEB Marine Genomics Course

Adaptation and selection

"It is… of the highest importance to gain insight into the means of modification and co-adaptation."

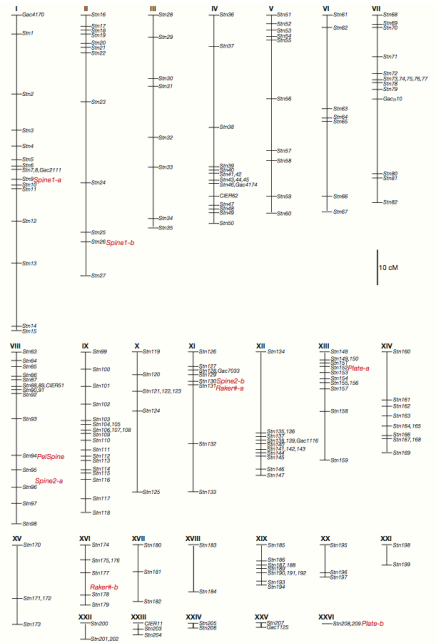
Darwin (1859)

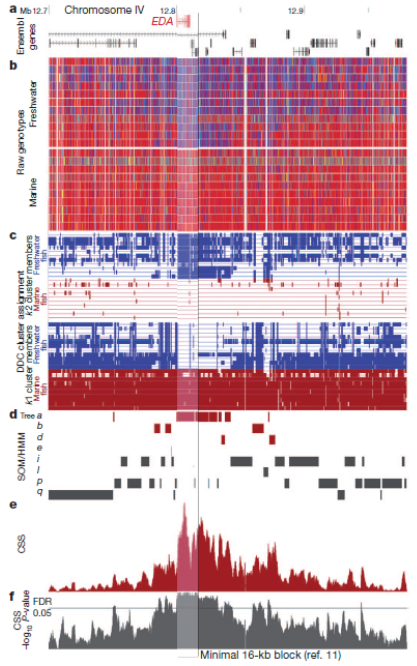




Genetic basis of adaptation

- How to find the genetic architecture underlying adaptive traits?





Peichel et al (2001) Colosimo et al (2004, 2005) Jones et al (2012)

Genetic basis of adaptation

- Second generation sequencing is revolutionizing the field

Genotyping by sequencing

- SNP discovery
- QTL mapping
- GWAS
- Admixture mapping
- Genome scans

NGS

Whole genome sequencing

- Resequencing
- SNP discovery
- Population history
- Gene annotation

Targeted sequencing

- Expression analysis
- Exon capture
- Candidate gene discovery

Why perform outlier analysis?

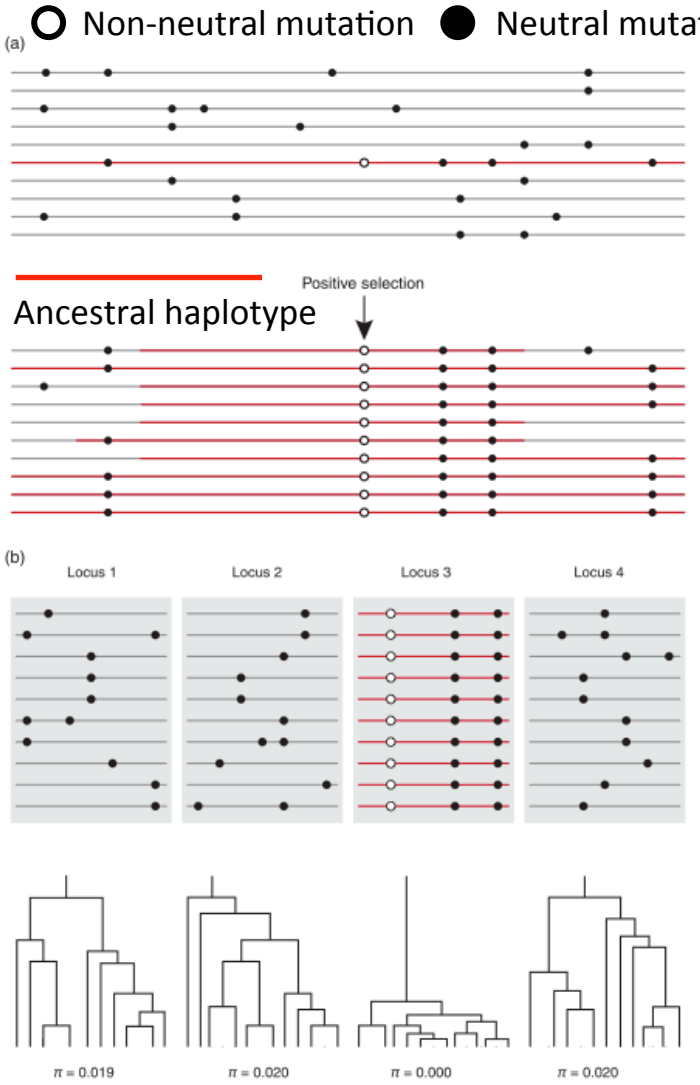
- Determine whether selection is occurring:
  - positive (directional)
  - balancing
- Identify genomic regions involved in adaptation
  - a 'bottom-up' approach with no prior knowledge of genes or traits involved
  - may not identify a candidate gene but a step in the right direction

Genotyping-by-sequencing

- Reduced representation – i.e. RAD sequencing
- Rapid, cost-effective and produces many thousands of loci scattered across the genome
- More loci means greater power:
  - to estimate genome-wide population parameters
  - with enough individuals and coverage also locus specific parameters

e.g $F_{ST,}$ H_O

Detecting outliers: basics



New mutation arises in a population with genetic variation and captures ancestral haplotype

Positive selection drives to locus and part of ancestral haplotype to fixation (some recombination breaks down)

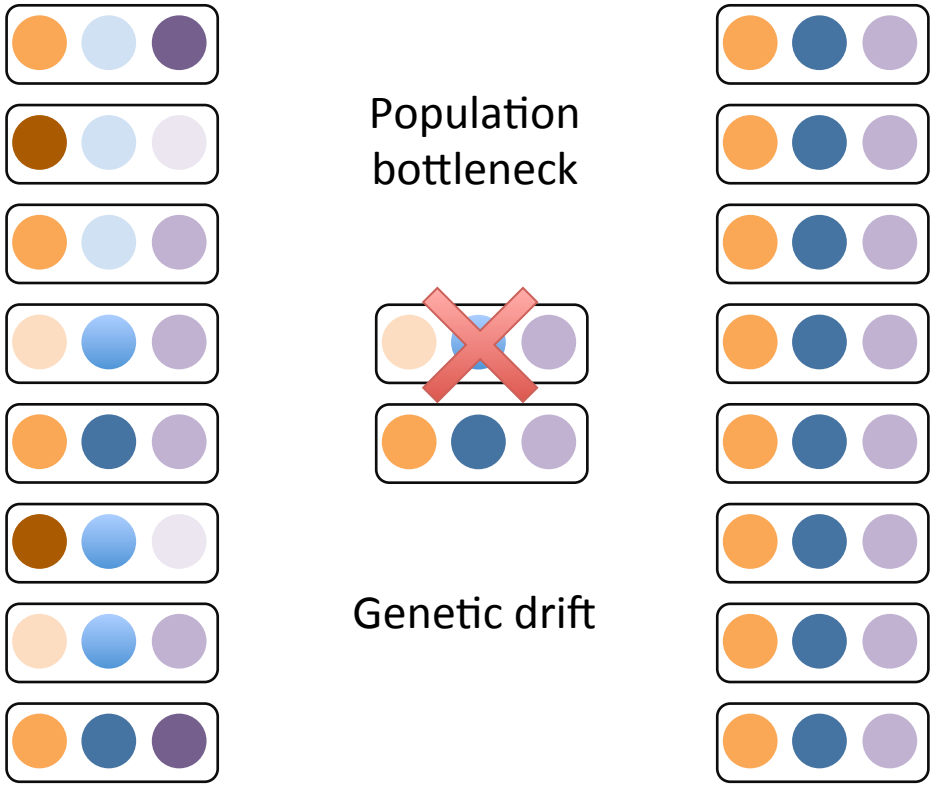
Locus 3 is now fixed, as are linked non-neutral mutations (genetic hitch-hiking)
Other loci show no signature of fixation

Alleles at this locus share a very recent common ancestor, low nucleotide diversity (π)

Storz (2005) Mol Ecol

Detecting outliers: basics

- But… demographic processes can cause a similar pattern



Fixation at multiple loci…

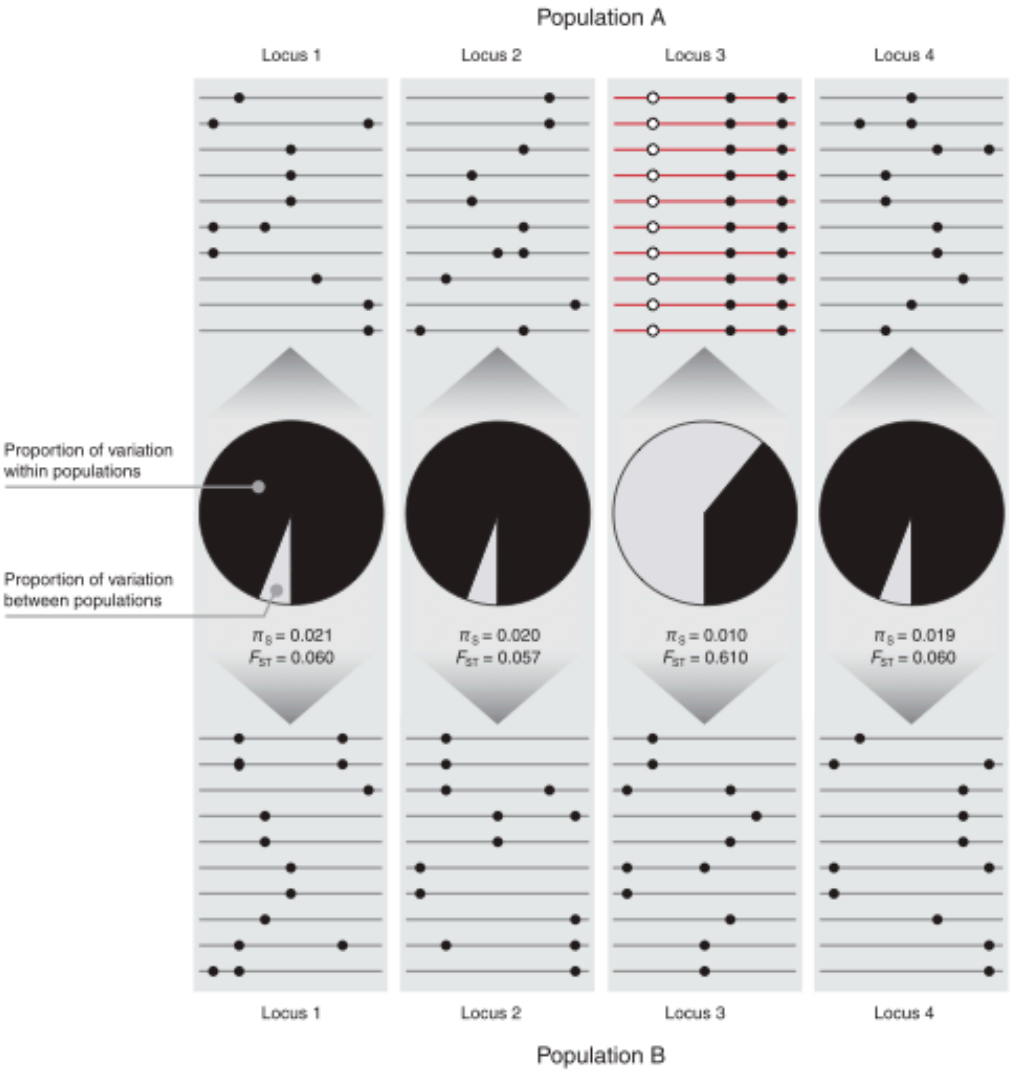
…but with a uniform effect across the genome

Compare multiple loci to distinguish population history and selection

Detecting outliers: F_{ST}

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2}$$

- Reduction of within-population diversity relative to increase in between-population diversity



Storz (2005) Mol Ecol

F_{ST} methods for outlier detection

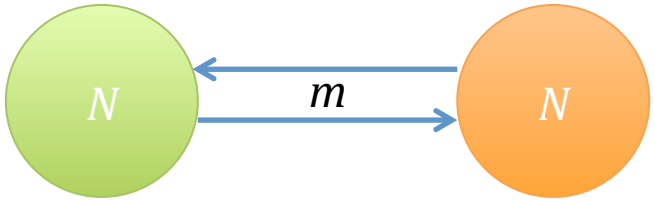
- Lewontin & Krakuer (1973) – ratio of expected variance in vs observed variance in F_{ST}
- if variance is greater than expected – selection

$$E\left[\mathrm{var}\left(F_{ST}\right)\right] : \mathrm{var}\left(F_{FST}\right)$$

- However, E[F_{ST}] is too low and can be distorted by population structure and demographic history

F_{ST} methods for outlier detection

- FDIST - Beaumont & Nichols (1996):
  1. Estimate F_{ST} for each locus
  2. Estimated $E[F_{ST}]$ from average of loci weighted by heterozygosity
  3. Use coalescent simulations under island model to estimate distribution of F_{ST} as a function of heterozygosity
  4. Estimate quantiles of distribution – i.e. 99%



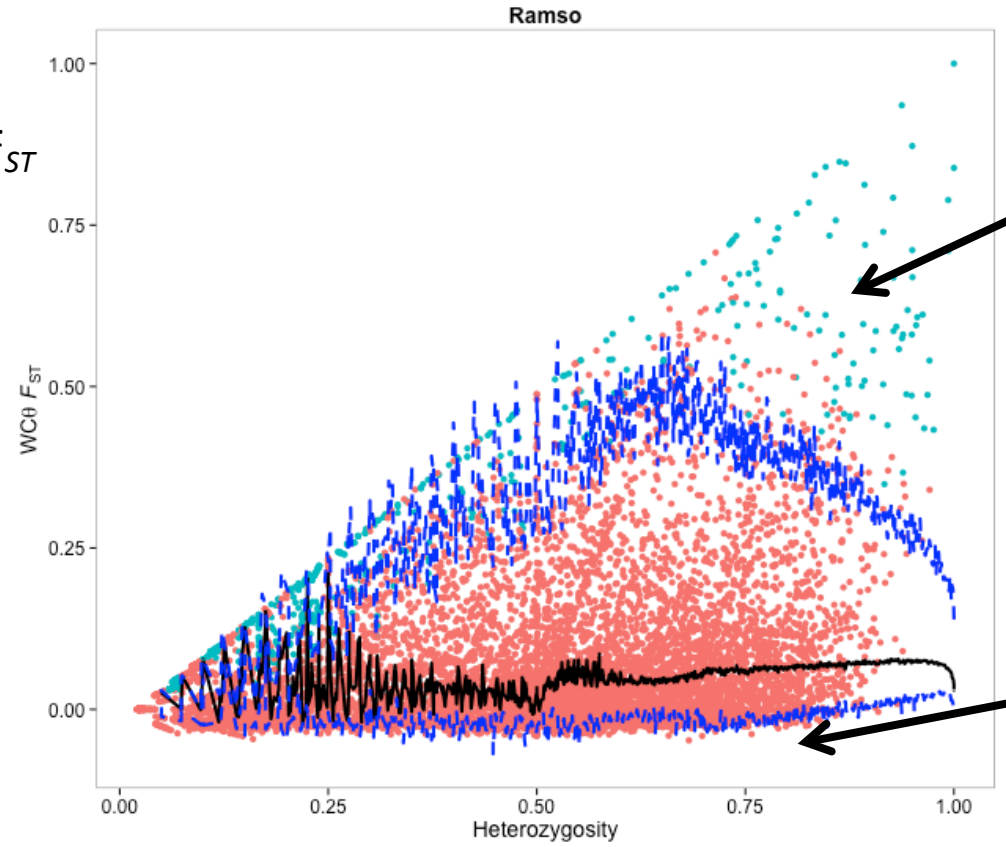
N = deme size
m = migration rate
μ = mutation rate
$\theta = N\mu$

F_{ST} methods for outlier detection

- FDIST on *L.saxatilis* ecotypes (Ramsö)
- 99% CI on joint F_{ST}/H_E distribution

99% CI

Median F_{ST}



Positive selection

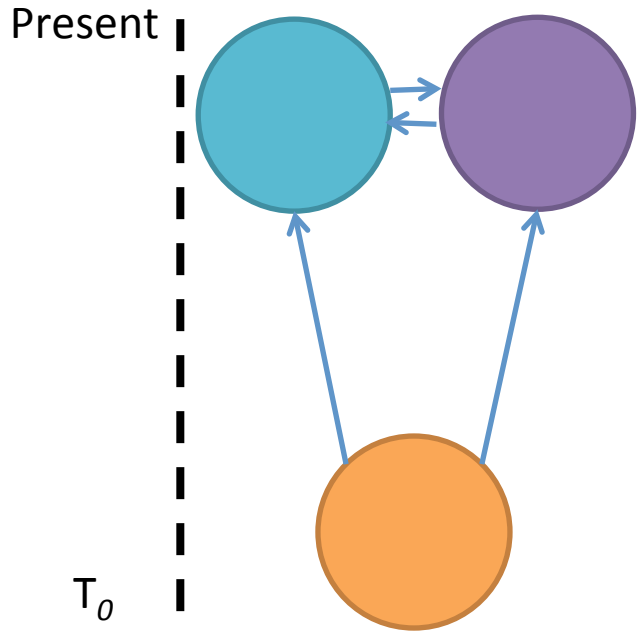
Balancing selection

FDIST – pros and cons

- Advantages:
  - robust to a wide range of scenarios and parameters
  - fast and straightforward to run
- Disadvantages
  - Simple model is easily violated and high false positive rate for balancing selection (Narum 2011)
  - Distribution of E[F_{ST}] is dependent on input loci which includes outliers – should use neutral loci

F_{ST} methods for outlier detection

- Bayescan – Foll & Gaggioitti (2008)
  1. Bayesian estimation of F_{ST}



Differences in allele frequencies at *i* loci due to drift and migration

j populations split from ancestral

$$F_{ST}^{ij}$$

- i.e. difference at locus *i* in population *j* from ancestral
- due to demographic history

F_{ST} methods for outlier detection

- Bayescan – Foll & Gaggioitti (2008)

  2. Determine locus and population specific effects

$$\log\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right)=\alpha_i+\beta_j$$

Intercept or population-specific effect

Slope or locus-specific effect

Makes it possible to derive two models – one with selection and one without

F_{ST} methods for outlier detection

- Bayescan – Foll & Gaggioitti (2008)

  2. Determine locus and population specific effects

Neutral model

$$\log\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right)=\beta_j$$

Selection model

$$\log\left(\frac{F_{ST}^{ij}}{1-F_{ST}^{ij}}\right)=\alpha_i+\beta_j$$

- Is α_i significantly different from zero? If so, then selection model supported
- Positive α_i suggests positive selection
- Negative α_i suggests balancing selection

F_{ST} methods for outlier detection

- Bayescan – Foll & Gaggioitti (2008)

  3. Use MCMC to estimate posterior probabilities and posterior odds

$$\frac{P(M_2|\theta)}{P(M_1|\theta)}$$

M_2 = selection model
M_1 = neutral model
θ = dataset

- Ratio of posterior probabilities – higher posterior odds means higher probability of selection model

F_{ST} methods for outlier detection

- Bayescan – Foll & Gaggioitti (2008)

  4. Control for false positives using q-values

  - Order p-values $p_1 \leq p_2 \ldots \leq p_k$ where k = no. of tests
  - Start at largest p-value and find first that satisfies:

$$p_i \leq (i/k)\alpha \quad \text{i = i}_{th} \text{ observation and } \alpha = \text{FDR}$$

  - This p-value is the new critical p-value

Benjamini & Hochberg (1995); Narum (2008)

- By defining a q-value, we determine an acceptable level of false positives in the final dataset – i.e. FDR = 0.05 means 5% of outliers are false positives

F_{ST} methods for outlier detection

- Bayescan on *L.saxatilis* ecotypes (Ramsö)
- FDR = 0.05

Bayescan – pros and cons

- Advantages:
  - Allows for uncertainty in allele frequencies
  - Tunes estimates of parameter distributions automatically
  - Lowest false positive rates (Narum 2011)
- Disadvantages
  - Like FDIST, relatively high false positive for loci under balancing selection
  - MCMC takes a long time to run for large datasets
  - Very low true positive rate with weak selection (De Mita 2013)

Some final points to consider

- Some methods better than others - i.e. F_{ST} differentiation methods do not appear to detect weak selection well (De Mita et al 2013)
- Methods specific to question – i.e. methods identifying associations with environmental clines can have high power
- A combination of two or more well chosen approaches is a robust means of identifying outlier loci

Recommended reading

Detecting selection - reviews:

- Storz (2005) Mol Ecol, 14, 671
- Nielsen (2005) Ann Rev. Gen. 39, 197

Outlier methods:

- Beaumont & Nichols (1996) Proc B, 263, 1619
- Askey et al (2002) Gen. Res. 12, 1805
- Foll & Gaggiotti (2008) Genetics, 180, 977

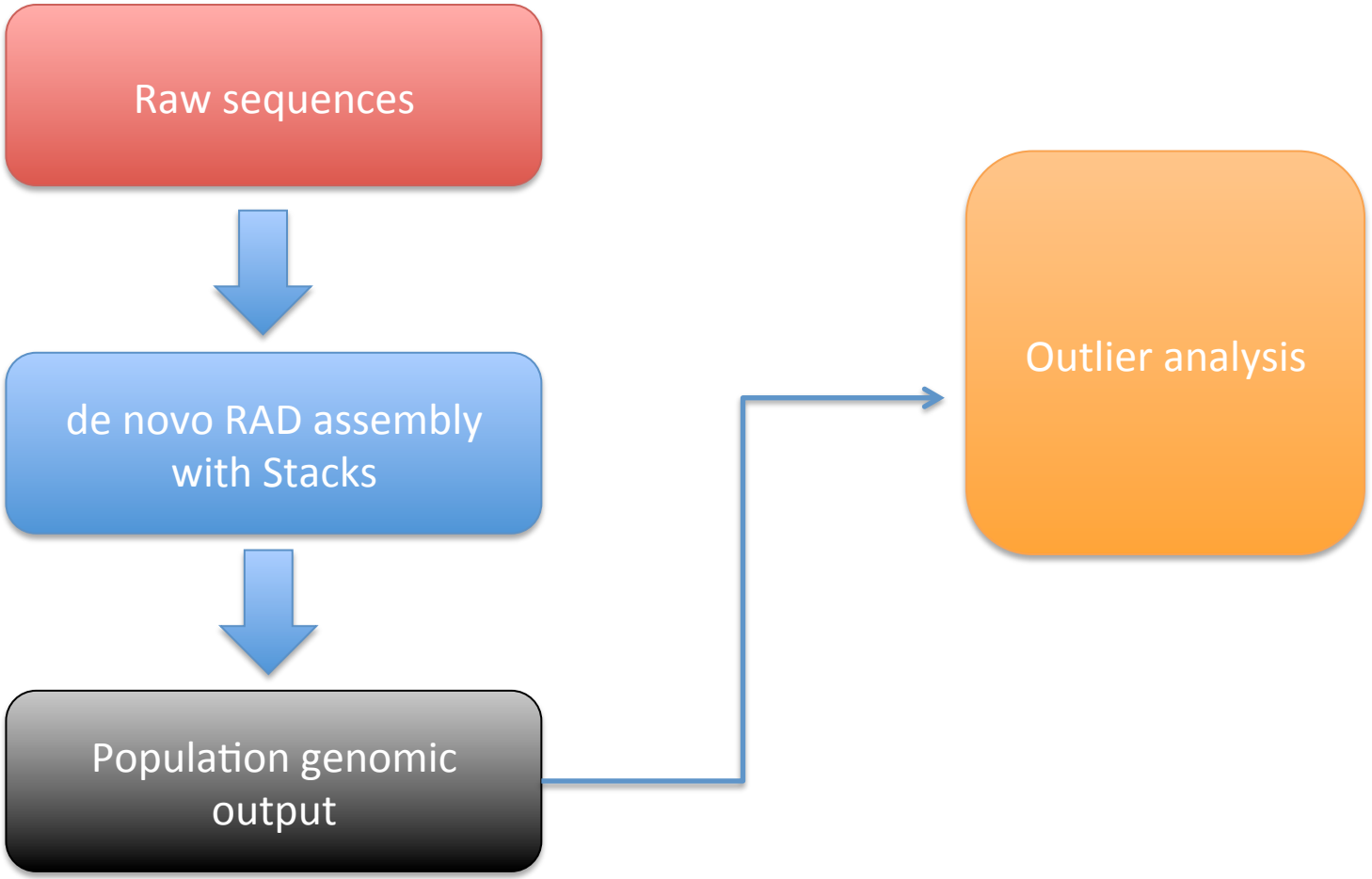
Outlier issues:

- Teshima et al (2006) Gen. Res. 16, 702
- Narum et al (2011) Mol Ecol, 11, 184
- De Mita et al (2013) Mol Ecol, 22, 1383

Recommend reading

- RAD sequencing
  - Baird et al (2008) PloS One, 3, e3376
  - Davey et al (2011) Nat. Rev. Gen., 12, p 499
  - Hohenlohe et al (2010) Plos Genetics, 6, e100862
  - Peterson et al (2012) PloS One, 7, e37135
- Stacks
  - Catchen et al (2011) G3, 3, p 171
  - Catchen et al (2013a) Mol Ecol, 22
  - Catchen et al (2013a) Mol Ecol, 22, 2864

RAD sequencing practical



Parallel selection – *Littorina saxatilis*

WAVE ECOTYPE
(AKA EXPOSED)

- Cliff habitat
- Wave action
- Smaller
- Thinner shell
- Bold

CRAB ECOTYPE
(AKA SHELTERED)

- Boulder habitat
- Crab predation
- Larger
- Thicker shell
- Less bold


Similar habitat clines occurring across the Western Swedish coast
Kosterhavets Nationalpark
Kosterhavets nationalpark
Sydkoster
Jutholmen
Saltö
A
Ramsö
Imagery ©2013 TerraMetri

Littorina genomic divergence

- Jutholmen, Western Sweden
- 48 individuals – sent to Floragenex for RAD library prep and sequencing

- 24 crab
- 24 wave

Now the sequences are back and it's time to get analysing!

RAD sequencing practical

Part 1: How to perform a *de novo* RAD assembly

- http://wp.me/p3efPk-1A

Part 2: Running the Stacks population module

- http://wp.me/p3efPk-1Q

Part 3: Outlier analysis with Bayescan

- http://wp.me/3efPk

- Overview
  - http://wp.me/p3efPk-2p