Ecological & evolutionary genomic analyses using RADseq

2017 Workshop on Genomics
Český Krumlov

Bill Cresko
Institute of Ecology and Evolution
Department of Biology
University of Oregon
Outline for today’s lecture

RADseq for ecological and evolutionary genomics

Primer on Population Genomics

Evolutionary genomics of stickleback fish
  - Population genomics of rapid adaptation
  - Using long read RAD-seq for coalescent analyses
  - Genome Wide Association Studies using RAD-seq

Genomically enabling the Gulf pipefish

RAD-seq experimental and statistical considerations

Stacks software pipeline (this afternoon & evening)
Harnessing the power of RADseq for ecological and evolutionary genomics

Kimberly R. Andrews¹, Jeffrey M. Good², Michael R. Miller³, Gordon Luikart⁴ and Paul A. Hohenlohe⁵
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.
Comparative Genomics
Vertebrate **zygotes** or embryos

- **28 day human**
- **19h zebrafish**
Vertebrate zygotes or embryos

28 day human

19h zebrafish
Population Genomics
How do we ‘genomically enable’ research studies of non-model (and model) organisms?

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

• Use restriction enzyme digestion to focus sequencing of multiple samples on homologous regions across the genome

• Simultaneous identification and typing of single nucleotide polymorphisms (SNPs) and haplotypes

• The cost is a fraction of the cost of re-sequencing the genome

• Thousands of genomes can be assayed in just a few weeks
What is RADseq?

(Restriction-site Associated DNA)
What is RADseq?

(Restriction-site Associated DNA)

22,830 SbfI sites in threespine stickleback genome

~ 45,000 RAD-Tags

HiSeq2500 Illumina Lane:
160 million reads

HiSeq4000 Illumina Lane:
350 million reads
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
Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization

Marie Cariou, Laurent Duret & Sylvain Charlat

Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 bou Villeurbanne F-69622, France

---

**Fig. 1. Phylogenetic relationships among the species used in this study.**

**Fig. 2. Results of SiLiX clustering of RAD sequences from the 12 Drosophila genomes.**

<table>
<thead>
<tr>
<th>Species</th>
<th>Percentage of loci sequenced at least</th>
<th>Bootstrap value (100 replicates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.melanogaster</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>D.sechellia</td>
<td>91.8</td>
<td></td>
</tr>
<tr>
<td>D.simulans</td>
<td>94.9</td>
<td></td>
</tr>
<tr>
<td>D.persimilis</td>
<td>80.4</td>
<td></td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td>D.yakuba</td>
<td>19.2</td>
<td></td>
</tr>
<tr>
<td>D.sechellia</td>
<td>24.6</td>
<td></td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>80.4</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 2.**

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of clusters with at least 5 sequences</th>
<th>Number of clusters with at least 4 sequences</th>
<th>Number of clusters with at least 3 sequences</th>
<th>Number of clusters with at least 2 sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.simulans</td>
<td>1744</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.simulans</td>
<td>1744</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.sechellia</td>
<td>1678</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.melanogaster</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.persimilis</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.pseudoobscura</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
<tr>
<td>D.yakuba</td>
<td>1436</td>
<td>761</td>
<td>317</td>
<td>210</td>
</tr>
</tbody>
</table>

---

**Funding Information**

Marie Cariou, Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 bou Villeurbanne F-69622, France.

E-mail: marie.cariou@univ-lyon1.fr

Villeurbanne F-69622, France.

doi: 10.1002/ece3.512
Phylogenetic relationships from diploid genomes, with a 10^5 sampling reads from the list of all possible RAD loci. RAD sequencing was then simulated by randomly selecting alleles (the upper bound of realistic polymorphism values). RAD sequencing was then simulated by randomly selecting alleles (the upper bound of realistic polymorphism values). Imaging random mutations of the sequenced genome, to produce a 5% average distance between homologous sequences (see main text). Node depth from Tamura et al. (2004).

### Table 1: Number of known and retrieved orthologous RAD tags in each species pair

<table>
<thead>
<tr>
<th>Species pair D. melanogaster</th>
<th>Node depth (My)</th>
<th>Orthologous tags</th>
<th>Retrieved orthologous tags (%)</th>
<th>In clusters including paralogs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. sechellia</td>
<td>5.4</td>
<td>2978</td>
<td>99</td>
<td>5</td>
</tr>
<tr>
<td>D. simulans</td>
<td>5.4</td>
<td>2892</td>
<td>99</td>
<td>4</td>
</tr>
<tr>
<td>D. erecta</td>
<td>12.6</td>
<td>2390</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>D. yakuba</td>
<td>12.8</td>
<td>2314</td>
<td>97</td>
<td>8</td>
</tr>
<tr>
<td>D. ananassae</td>
<td>44.2</td>
<td>916</td>
<td>68</td>
<td>9</td>
</tr>
<tr>
<td>D. persimilis</td>
<td>54.9</td>
<td>648</td>
<td>65</td>
<td>9</td>
</tr>
<tr>
<td>D. pseudoobscura</td>
<td>54.9</td>
<td>648</td>
<td>66</td>
<td>9</td>
</tr>
<tr>
<td>D. willistoni</td>
<td>62.2</td>
<td>242</td>
<td>49</td>
<td>6</td>
</tr>
<tr>
<td>D. grimshawi</td>
<td>62.9</td>
<td>290</td>
<td>60</td>
<td>8</td>
</tr>
<tr>
<td>D. virilis</td>
<td>62.9</td>
<td>286</td>
<td>59</td>
<td>5</td>
</tr>
<tr>
<td>D. mojavensis</td>
<td>62.9</td>
<td>298</td>
<td>59</td>
<td>8</td>
</tr>
</tbody>
</table>
Empirical Assessment of RAD Sequencing for Interspecific Phylogeny

Astrid Cruaud, Mathieu Gautier, Maxime Galan, Julien Foucaud, Laure Sauné, Gwenaëlle Genson, Emeric Dubois, Sabine Nidelet, Thierry Deuve, and Jean-Yves Rasplus

Abstract

Next-generation sequencing opened up new possibilities in phylogenetics; however, choosing an appropriate method of sample preparation remains challenging. Here, we demonstrate that restriction-site-associated DNA sequencing (RAD-seq) generates useful data for phylogenomics. Analysis of our RAD library using current bioinformatic and phylogenetic tools produced 400× more sites than our Sanger approach (2,262,825 nt/species), fully resolving relationships between 18 species of ground beetles (divergences up to 17 My). This suggests that RAD-seq is promising to infer phylogeny of eukaryotic species, though potential biases need to be evaluated and new methodologies developed to take full advantage of such data.
Improve genome assemblies with genetic maps
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process
Let’s go into a little more detail
Restriction Enzyme (RE) digestion and first adaptor ligation

<table>
<thead>
<tr>
<th>0.0</th>
<th>1.0</th>
<th>2.0</th>
<th>3.0</th>
<th>4.0</th>
<th>5.0kb</th>
<th>6.0</th>
<th>7.0</th>
<th>8.0</th>
<th>9.0</th>
<th>10.0</th>
</tr>
</thead>
</table>


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

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

Sample Barcodes and Molecular Identification Barcodes (MIPs)

A = Amplification primer
B = Sequencing primer
C = Barcode
Shearing and second adaptor ligation

A = Amplification primer
B = Sequencing primer
C = Barcode

* Defining difference between original RAD and other approaches*
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

Original RAD
1. Digest (one enzyme)
2. Ligate adaptors
3. Multiplex
4. Shear
5. Size select
6. End repair
7. A-tailing
8. Ligate Y-adaptors
9. PCR

2bRAD
1. Digest (one llb enzyme)
2. Ligate adaptors
3. PCR
4. Multiplex

Harnessing the power of RADseq for ecological and evolutionary genomics
Kimberly R. Andrews¹, Jeffrey M. Good², Michael R. Miller³, Gordon Luikart⁴ and Paul A. Hohenlohe⁵
RADseq with one or two enzyme digestion

**GBS**
1. Digest (one enzyme)
2. Ligate adaptors
3. Multiplex
4. PCR

**ezRAD**
1. Digest (one or more enzymes)
   - Illumina kit:
     2. End repair
     3. A-tailing
     4. Ligate Y-adaptors
5. Size select
6. PCR (skip for PCR-free kit)
7. Multiplex

**ddRAD**
1. Digest (two enzymes)
2. Ligate adaptors
3. Multiplex
4. Size select
5. PCR
### Summary of plusses and minuses of RAD family

<table>
<thead>
<tr>
<th></th>
<th>Original RAD</th>
<th>2bRAD</th>
<th>GBS</th>
<th>ddRAD</th>
<th>ezRAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Options for tailoring number of loci</strong></td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme</td>
<td>Change restriction enzyme or size selection window</td>
<td>Change restriction enzyme or size selection window</td>
</tr>
<tr>
<td><strong>Number of loci per 1 Mb of genome size</strong></td>
<td>30–500</td>
<td>50–1,000</td>
<td>5–40</td>
<td>0.3–200</td>
<td>10–800</td>
</tr>
<tr>
<td><strong>Length of loci</strong></td>
<td>≤1 kb if building contigs; otherwise ≤300 bp</td>
<td>33–36 bp</td>
<td>&lt;300 bp</td>
<td>≤300 bp</td>
<td>≤300 bp</td>
</tr>
<tr>
<td><strong>Cost per barcoded or indexed sample</strong></td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td><strong>Effort per barcoded or indexed sample</strong></td>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td><strong>Use of proprietary kit</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Identification of PCR duplicates</strong></td>
<td>With paired-end sequencing</td>
<td>No</td>
<td>With degenerate barcodes</td>
<td>With degenerate barcodes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Specialized equipment needed</strong></td>
<td>Sonicator</td>
<td>None</td>
<td>None</td>
<td>Pippin Prep</td>
<td>Pippin Prep</td>
</tr>
<tr>
<td><strong>Suitability for large or complex genomes</strong></td>
<td>Good</td>
<td>Poor</td>
<td>Moderate</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td><strong>Suitability for de novo locus identification (no reference genome)</strong></td>
<td>Good</td>
<td>Poor</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td><strong>Available from commercial companies</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Random shearing in original RAD - Local Paired End (PE) Assemblies

A) Restriction sites in genome

RAD tag sequence read

Sheared-end reads

B) Variable length RAD fragments isolated

C) Contigs assembled from the sheared-end reads for each RAD tag

200 - 800bp in length
‘Bias’ in RADseq

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
INTRODUCTORY STATISTICAL PRINCIPLES

Fig 3.1  Fictitious histogram (a) and (b) normal and (c-d) log-normal probability distributions. Note, it is not possible to determine the probability of discrete events (such as the probability of encountering a koala weighing 12.183 kg) only ranges of values.

3.1.1 The normal distribution
It has been a long observed mathematical phenomenon that the accumulation of a set of independent random influences tend to converge upon a central value (central limit theorem) and that the distribution of such accumulate values follows a specific 'bell shaped' curve called a normal or Gaussian distribution (see Figure 3.1b). The normal distribution is a symmetrical distribution in which values close to the center of the distribution are more likely and that progressively larger and smaller values are less commonly encountered.

Many biological measurements (such as the weight of a Victorian male koala) are likewise influenced by an almost infinite number of factors (many of which can be considered independent and random) and thus many biological variables also follow a normal distribution. Since any scientific variable behaves according to the central limit theorem, many of the common statistical procedures have been specifically derived for (and thus assume) normally distributed data. In fact, the reliability of inferences based on such procedures is directly related to the degree of conformity to this assumption of normality. Likewise, many other statistical elements rely on normal distributions, and thus the normal distribution (or variants thereof) is one of the most important mathematical distributions.

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

\[
e = 2.7182....
\]

\[
\pi = 3.1415....
\]
INTRODUCTORY STATISTICAL PRINCIPLES

3.1.1 The normal distribution

It has been a long observed mathematical phenomenon that the accumulation of a set of independent random influences tend to converge upon a central value (central limit theorem) and that the distribution of such accumulations follows a specific 'bell shaped' curve called a normal or Gaussian distribution (see Figure 3.1b). The normal distribution is a symmetrical distribution in which values close to the center of the distribution are more likely and that progressively larger and smaller values are less commonly encountered.

Many biological measurements (such as the weight of a Victorian male koala) are likewise influenced by an almost infinite number of factors (many of which can be considered independent and random) and thus many biological variables also follow a normal distribution. Since many scientific variables behave according to the central limit theorem, many of the common statistical procedures have been specifically derived for (and thus assume) normally distributed data. In fact, the reliability of inferences based on such procedures is directly related to the degree of conformity to this assumption of normality. Likewise, many other statistical elements rely on normal distributions, and thus the normal distribution (or variants thereof) is one of the most important mathematical distributions.

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]
Chromosome sampling depth is correlated with particular genealogies

Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the true genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of $p_t$, loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA's, not shown).

### Table 1

<table>
<thead>
<tr>
<th>Protocol</th>
<th>$\theta$ per bp</th>
<th>$\theta_{we}/\theta_{wa}$</th>
<th>$\pi_{e}/\pi_{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
</tr>
</tbody>
</table>
Chromosome sampling depth is correlated with particular genealogies. Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the 'true' genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of \( p_t \), loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA's, not shown) relative to the simulation average.

### Table 1

<table>
<thead>
<tr>
<th>Protocol</th>
<th>( \theta ) per bp</th>
<th>( \theta_{we}/\theta_{wa} )</th>
<th>( \pi_e/\pi_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
<td></td>
</tr>
<tr>
<td><strong>Double digest</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
<td></td>
</tr>
</tbody>
</table>

Results from two different simulation parameters of \( \theta \) are shown. When recombination is present, \( q = \theta \). Results are given for both the standard and double digest RADseq protocols.
Chromosome sampling depth is correlated with particular genealogies.

Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the 'true' genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of \( p_t \), loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA’s, not shown) relative to the simulation average.

Table 1

<table>
<thead>
<tr>
<th>Protocol</th>
<th>( \theta ) per bp</th>
<th>( \theta_{we}/\theta_{wa} )</th>
<th>( \pi_e/\pi_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
</tr>
</tbody>
</table>

‘Bias’ in RADseq is increased in some RAD protocols.
Biological studies that benefit from whole genome approaches powered by RAD-seq

- 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
- Making management decisions for biological populations
- 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
Biological studies that benefit from whole genome approaches powered by RAD-seq

- 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
- Making management decisions for biological populations
- 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

Any study where improving biological sample size would be beneficial
Outline for today’s lecture

RAD-seq for ecological and evolutionary genomics

Primer on Population Genomics

Evolutionary genomics of stickleback fish

- Population genomics of rapid adaptation
- Using long read RAD-seq for coalescent analyses
- Genome Wide Association Studies using RAD-seq

Genomically enabling the Gulf pipefish
How do organisms adapt to novel environments?

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?

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
Fixation  Loss
1           0
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 ordered markers
Population genomics of ordered markers
Population genomics of ordered markers
Population genomics of ordered markers
Sweeps of directional selection across genomes

Past

Hard Sweep

Present

Soft Sweep
Decision tree for population genomics data

- **Time Scale**
  - short
  - long

- **Number of Populations**
  - one
  - multiple

- **Mode of Selection**
  - positive
  - balancing

- **Type of Sweep**
  - hard
  - soft

- **Knowledge of substitution class**
  - yes
  - no

- **FST tests**
- LD tests (Zg, XP-EHH)

- **π-based tests**
- allele frequency spectrum (Tajima’s D)
- LD tests (iHS)

- **dN/dS ratio**

- **HKA test**

Hohenlohe, Phillips, Cresko, 2010
Stacks analysis pipeline for RAD-seq

Stacks: an analysis tool set for population genomics

Julian M. Catchen,* Paul A. Hohenlohe,† Susan Bsham,* Angel Amores‡ and William A. Cresko*"
Stacks
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
### 7.4.2 batch_X.sumstats_summary.tsv: Summary of summary statistics for each population

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pop ID</td>
<td>Population ID as defined in the Population Map file.</td>
</tr>
<tr>
<td>2</td>
<td>Private</td>
<td>Number of private alleles in this population.</td>
</tr>
<tr>
<td>3</td>
<td>Number of Individuals</td>
<td>Mean number of individuals per locus in this population.</td>
</tr>
<tr>
<td>4</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>Mean frequency of the most frequent allele at each locus in this population.</td>
</tr>
<tr>
<td>7</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Observed Heterozygosity</td>
<td>Mean observed heterozygosity in this population.</td>
</tr>
<tr>
<td>10</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Observed Homozygosity</td>
<td>Mean observed homozygosity in this population.</td>
</tr>
<tr>
<td>13</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Expected Heterozygosity</td>
<td>Mean expected heterozygosity in this population.</td>
</tr>
<tr>
<td>16</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Expected Homozygosity</td>
<td>Mean expected homozygosity in this population.</td>
</tr>
<tr>
<td>19</td>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Standard Error</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>( \pi )</td>
<td>Mean value of ( \pi ) in this population.</td>
</tr>
<tr>
<td>22</td>
<td>( \pi ) Variance</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>( \pi ) Standard Error</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>( F_{IS} )</td>
<td>Mean measure of ( F_{IS} ) in this population.</td>
</tr>
<tr>
<td>25</td>
<td>( F_{IS} ) Variance</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>( F_{IS} ) Standard Error</td>
<td></td>
</tr>
</tbody>
</table>
### 7.4.4 batch_X.hapstats.tsv: Haplotype-based summary statistics for each locus in each population

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Batch ID</td>
<td>The batch identifier for this data set.</td>
</tr>
<tr>
<td>2</td>
<td>Locus ID</td>
<td>Catalog locus identifier.</td>
</tr>
<tr>
<td>3</td>
<td>Chromosome</td>
<td>If aligned to a reference genome.</td>
</tr>
<tr>
<td>4</td>
<td>Basepair</td>
<td>If aligned to a reference genome.</td>
</tr>
<tr>
<td>5</td>
<td>Population ID</td>
<td>The ID supplied to the populations program, as written in the population map file.</td>
</tr>
<tr>
<td>6</td>
<td>N</td>
<td>Number of alleles/haplotypes present at this locus.</td>
</tr>
<tr>
<td>7</td>
<td>Haplotype count</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Gene Diversity</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Smoothed Gene Diversity</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Smoothed Gene Diversity P-value</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Haplotype Diversity</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Smoothed Haplotype Diversity</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Smoothed Haplotype Diversity P-value</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Haplotypes</td>
<td>A semicolon-separated list of haplotypes/haplotype counts in the population.</td>
</tr>
</tbody>
</table>
### 7.4.3 batch_X.fst_Y-Z.tsv: \(F_{ST}\) calculations for each pair of populations

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

**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 -\(x\))
# Stacks

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

<table>
<thead>
<tr>
<th>Column</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Batch ID</td>
<td>The batch identifier for this data set.</td>
</tr>
<tr>
<td>2</td>
<td>Locus ID</td>
<td>Catalog locus identifier.</td>
</tr>
<tr>
<td>3</td>
<td>Population ID 1</td>
<td>The ID supplied to the populations program, as written in the population map file.</td>
</tr>
<tr>
<td>4</td>
<td>Population ID 2</td>
<td>The ID supplied to the populations program, as written in the population map file.</td>
</tr>
<tr>
<td>5</td>
<td>Chromosome</td>
<td>If aligned to a reference genome.</td>
</tr>
<tr>
<td>6</td>
<td>Basepair</td>
<td>If aligned to a reference genome.</td>
</tr>
<tr>
<td>7</td>
<td>$\Phi_{ST}$</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Smoothed $\Phi_{ST}$</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Smoothed $\Phi_{ST}$ P-value</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$F_{ST}'$</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Smoothed $F_{ST}'$</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Smoothed $F_{ST}'$ P-value</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>$D_{EST}$</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Smoothed $D_{EST}$</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Smoothed $D_{EST}$ P-value</td>
<td></td>
</tr>
</tbody>
</table>
Outline for today’s lecture

RAD-seq for ecological and evolutionary genomics

Primer on Population Genomics

**Evolutionary genomics of stickleback fish**
- Population genomics of rapid adaptation
- Using haplo-RAD-seq for coalescent analyses
- Genome Wide Association Studies using RAD-seq

Genomically enabling the Gulf pipefish
Threespine stickleback, *Gasterosteus aculeatus*
Threespine stickleback, *Gasterosteus aculeatus*

Rundle and McKinnon 2002
Threespine stickleback, *Gasterosteus aculeatus*

- Pelvic Structure
- Lateral Plates

Rundle and McKinnon 2002
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.
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} vs Genomic location (mBases)

Bear Paw (mean F_{ST} = 0.121)

Boot (mean F_{ST} = 0.112)

Mud (mean F_{ST} = 0.117)

Hohenlohe, Bassham et al. 2010. PLoS Genetics
Numerous novel regions identified

Different alleles

More often the same alleles

LGI

LGXXI
Evolution of stickleback in 50 years on earthquake-uplifted islands

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

\textsuperscript{a}School of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, AK 99775; \textsuperscript{b}Department of Biological Sciences, University of Alaska Anchorage, Anchorage, AK 99508; \textsuperscript{1}Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403; and \textsuperscript{c,2}Department of Ecology and Evolution, 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

2km

Photo Credits: BLM, E-Terra
Middleton Island - 50 year old populations

Photo Credits: BLM, E-Terra
Middleton Island - 50 year old locations

Photo Credits: BLM, E-Terra
RAD-seq analysis

110,000 SNPs per individual
>1000 Individuals
20 million genotypes
Structure analysis shows independent evolution even among populations on a single island
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

[Image of a circular genome map with annotations for different regions and FST values, including Montague Island, SW, SE, OC, FW1, and FW2.]
Coalescent analyses using RADseq
The coalescent in population genetics

Noah A. Rosenberg & Magnus Nordborg

Nature Reviews Genetics
Neutral coalescent expectations
Natural selection and the coalescent

Divergent Selection

Balancing Selection
Coalescent analyses in stickleback

\[ E[T_{MRCA^*}] = 4N_e \text{ gen.} \]
Coalescent analyses in stickleback

Selective sweep within a population

Time in past
Coalescent analyses with RAD-seq

Genomic DNA

SbfI cutsites

Sequence reads
Coalescent analyses with RAD-seq

- Genomic DNA
- SbfI cutsites
- Sequence reads

Stacks → PHASE

350 bp

696 bp
Coalescent analyses with RAD-seq

Indiv. 1

Indiv. 2

Indiv. 13

locus 1

locus 2

locus 13,070
Selection in one population reduces coalescence time

Fold reduction in $T_{MRCA}$

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

Chromosome 21, Mb
Selection in one population reduces coalescence time

\[
\log_2\left(\frac{T_{MRCA}^{FW}}{T_{MRCA}^{ALL}}\right)
\]
However, across populations the coalescence time can increases significantly in a genomic region.
Relative proportion of $T_{\text{mrca}}$ across different habitats

LG21
Relative proportion of $T_{\text{mrca}}$ across different habitats

Entire genome
Increased absolute divergence between habitats

Stickleback chr IV: 12.80 - 12.82 Mb

Eda
Increased absolute divergence between habitats
Absolute divergence co-localizes with inversions
How do the genomic patterns of divergence link to phenotypic diversification?

Kristin Alligood
Lateral plate and opercle shape co-vary in the wild

A

Anterior | Supporting | Posterior | Keel

Marine

Freshwater

Standard Length (SL)

1cm

B
An interesting stickleback hybrid population in Oregon
An interesting stickleback hybrid population in Oregon

A

Marine

Freshwater

Lateral Plate Count

B

Opercle PCI scores

61% variation
An interesting stickleback hybrid population in Oregon

**A**

Genome Wide Association: GEMMA

Zhou and Stephens 2012

**B**

Opercle PCI scores

61% variation
Pairwise relatedness among individuals
Lateral plate count:
A good hit to a novel locus on LGXX
Lateral plate count:
A good hit to a novel locus on LGXX
Opercle shape:
Good hits on LGVII and LGVIII
Opercle shape:
Good hits on LGVII and LGVIII
Can’t we just pool together young divergent populations and do GWAS?
Linkage disequilibrium created by natural selection
This sort of population structure is difficult to control.
• Previous work has shown that the freshwater genomes evolve in 13,000 years.

• These new Middleton Island data shows that the phenotype can appear in as little as 50 years.

• Much of the divergence involves soft sweeps.

• This could represent thousands of haplotypes reassembling, but the genome appears chunkier.

• Many haplotypes are habitat specific, quite ancient and coincide with structural variation

• Diverging phenotypes map to these same genomic regions
Hypothesis: Old genomic architecture variation is a product of the metapopulation structure of stickleback, and this architecture strongly influences subsequent rapid evolution.
Outline for today’s lecture

RAD-seq for ecological and evolutionary genomics

Primer on Population Genomics

Evolutionary genomics of stickleback fish
  - Population genomics of rapid adaptation
  - Using long read RAD-seq for coalescent analyses
  - Genome Wide Association Studies using RAD-seq

*Genomically enabling the Gulf pipefish*
What if you don’t have a genome sequence?

A case study of RAD-seq and genome assembly
Seahorses, sea dragons and pipefishes
The genome of the Gulf pipefish enables understanding of evolutionary innovations

C. M. Small¹‡, S. Bassham¹‡, J. Catchen¹²‡, A. Amores³, A. M. Fuiten¹, R. S. Brown¹⁴, A. G. Jones⁵ and W. A. Cresko¹∗
We’re really interested in head development

Pipefish

Stickleback
How did we genomically enable pipefish

1) A high quality transcriptome

2) Very dense RAD genetic map

3) Deep shotgun sequencing of the genome

4) Order contigs against the RAD reference map
Illumina genomic libraries for pipefish genome

**paired end 101bp**

- 500-700bp
- 25x

**mate pair**

- 4500-7500bp
- 2x

**overlapping**

- 150-250bp
- 40x

**paired end RAD**

- ACTCTC
- 500-1200bp
- 15-25x of 3% of the genome
Nearly the whole genome is covered

Table 1: Scaffold-level assembly statistics for the Gulf pipefish genome

<table>
<thead>
<tr>
<th>Genome</th>
<th>Scaffolds (n)</th>
<th>Longest scaffold</th>
<th>Scaffold N50</th>
<th>Contig N50</th>
<th>Assembly length</th>
<th>Gaps in assembly (%)</th>
<th>CEGs complete (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulf pipefish (Syngnathus scovelli)</td>
<td>2104</td>
<td>6.7 Mb</td>
<td>640.4 kb</td>
<td>32.2 kb</td>
<td>307.0 Mb</td>
<td>6.6</td>
<td>98.8</td>
</tr>
<tr>
<td>African turquoise killifish (Nothobranchius furzeri)</td>
<td>29,054</td>
<td>0.7 Mb</td>
<td>119.7 kb</td>
<td>8.7 kb</td>
<td>1010.9 Mb</td>
<td>7.7</td>
<td>94.8</td>
</tr>
<tr>
<td>Blind cave fish (Astyanax mexicanus)</td>
<td>10,542</td>
<td>9.8 Mb</td>
<td>1775.3 kb</td>
<td>14.7 kb</td>
<td>1191.1 Mb</td>
<td>19.1</td>
<td>87.9</td>
</tr>
<tr>
<td>Spotted gar (Lepisosteus oculatus)</td>
<td>2105</td>
<td>21.3 Mb</td>
<td>6928.1 kb</td>
<td>68.3 kb</td>
<td>945.8 Mb</td>
<td>8.1</td>
<td>90.7</td>
</tr>
</tbody>
</table>
Created a genetic map

Generated an F1 family of 103 individuals

RAD sequenced the parents and offspring

Analyzed the data using Stacks

Output to JoinMap format

Created Linkage map
22 LGs; 6000 segregating SNPs; 30,000 RAD sites
The pipefish genetic map and genome together

Software to help integrate genetic and physical maps
http://catchenlab.life.illinois.edu/chromonomer/
Two instances of chromosome fusion in pipefish
Two instances of chromosome fusion in pipefish
Key losses in *Hox* gene clusters
Key losses in conserved non-coding elements
Disruption of core hind fin patterning program

No evidence of *Tbx4* in the assembly
Disruption of core hind fin patterning program

Disruption in coding sequence of *Pitx1*
Expansion of male pregnancy specific gene family
Ecological & evolutionary genomic analyses using RAD-seq - what have we learned?
RAD-seq can be a tool for enabling new research in models & nonmodels
- 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 (or your mother’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 of smaller contigs is often good enough

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
Lab bench considerations for RADseq studies
Statistical considerations in RAD-seq

Reference genome sequence

Restriction enzyme recognition site

Sequence reads
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
Maximum likelihood genotyping based on multinomial distribution of nucleotide reads

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

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

3.1.1 The normal distribution

It has been long observed that the accumulation of a set of independent random influences tends to converge upon a central value (central limit theorem) and that the distribution of such accumulated values follows a specific 'bell shaped' curve called a normal or Gaussian distribution (see Figure 3.1b). The normal distribution is a symmetrical distribution in which values close to the center of the distribution are more likely and that progressively larger and smaller values are less commonly encountered.

Many biological measurements (such as the weight of a Victorian male koala) are likewise influenced by an almost infinite number of factors (many of which can be considered independent and random) and thus many biological variables also follow an normal distribution. Since many scientific variables behave according to the central limit theorem, many of the common statistical procedures have been specifically derived for (and thus assume) normally distributed data. In fact, the reliability of inferences based on such procedures is directly related to the degree of conformity to this assumption of normality. Likewise, many other statistical elements rely on normal distributions, and thus the normal distribution (or variants thereof) is one of the most important mathematical distributions.

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

\[ e = 2.7182 \ldots \]
\[ \pi = 3.1415 \ldots \]
INTRODUCTORY STATISTICAL PRINCIPLES

3.1.1 The normal distribution

It has been a long observed mathematical phenomenon that the accumulation of a set of independent random influences tend to converge upon a central value (central limit theorem) and that the distribution of such accumulated values follows a specific "bell shaped" curve called a normal or Gaussian distribution (see Figure 3.1b). The normal distribution is a symmetrical distribution in which values close to the center of the distribution are more likely and that progressively larger and smaller values are less commonly encountered.

Many biological measurements (such as the weight of a Victorian male koala) are likewise influenced by an almost infinite number of factors (many of which can be considered independent and random) and thus many biological variables also follow a normal distribution. Since any scientific variable behaves according to the central limit theorem, many of the common statistical procedures have been specifically derived for (and thus assume) normally distributed data. In fact, the reliability of inferences based on such procedures is directly related to the degree of conformity to this assumption of normality. Likewise, many other statistical elements rely on normal distributions, and thus the normal distribution (or variants thereof) is one of the most important mathematical distributions.

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]
Bias in RAD-sequencing

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

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

Gene genealogy

Sequence alignment

$T_{MRCA}$

time

RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling.
Bias in RAD-sequencing summary

Chromosome sampling depth is correlated with particular genealogies. Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the ‘true’ genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of \( \pi_t \), loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA’s, not shown) relative to the simulation average.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>( \theta ) per bp</th>
<th>( \theta_{we}/\theta_{wa} )</th>
<th>( \pi_{e}/\pi_{a} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
</tr>
</tbody>
</table>
Bias in RAD-sequencing summary

Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the 'true' genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of \( p_t \), loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA's, not shown) relative to the simulation average.

### Table 1

<table>
<thead>
<tr>
<th>Protocol</th>
<th>( \theta ) per bp</th>
<th>( \theta_{we}/\theta_{wa} )</th>
<th>( \pi_e/\pi_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Results from two different simulation parameters of \( \theta \) are shown. When recombination is present, \( q = \theta \). Results are given for both the standard and double digest RADseq protocols.

© 2013 John Wiley & Sons Ltd
Chromosome sampling depth is correlated with particular genealogies. Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the 'true' genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe et al. 2012; Langley et al. 2012; Pool et al. 2012). Lastly, in agreement with their higher values of \( p_t \), loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (TMRCA's, not shown) relative to the simulation average.

### Table 1

<table>
<thead>
<tr>
<th>Protocol</th>
<th>( \theta ) per bp</th>
<th>( \theta_{we}/\theta_{wa} )</th>
<th>( \pi_e/\pi_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Results from two different simulation parameters of \( \theta \) are shown. When recombination is present, \( q = h \). Results are given for both the standard and double digest RADseq protocols.

© 2013 John Wiley & Sons Ltd
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

\[ \frac{1,000,000}{100} \div 1,000 = 10x \text{ coverage} \]

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**

\[(0.25)^n \times \text{genome size} = \text{expected # sites}\]

Genomes are biased:

<table>
<thead>
<tr>
<th>Expected Sites</th>
<th>Actual Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>112,300 six-cutter sites in stickleback (460 Mb)</td>
<td>actual <strong>EcoRI</strong> sites = 90,000</td>
</tr>
<tr>
<td>7000 eight-cutter sites in stickleback</td>
<td>actual <strong>SbfI</strong> sites = 22,800</td>
</tr>
<tr>
<td>32,900 six-cutter sites in <em>C. remanei</em> (135 Mb)</td>
<td>actual <strong>EcoRI</strong> sites = 73,200</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>CGATA</th>
<th>GTACA</th>
<th>TAGCC</th>
<th>ACTGC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Image of gel electrophoresis showing DNA fragment patterns]
Molecular considerations in library building

How can I get the best depth of coverage?

Things to consider

Fragment size
Smaller/tighter is better
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
The pipefish genetic map is closed; 22 LGs
6000 segregating SNPs; 30,000 RAD sites