Ecological & evolutionary genomic analyses using RAD-seq

2015 Workshop on Genomics
Český Krumlov

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

**RAD-seq for evolutionary genomic analyses**

**Evolutionary genomics of stickleback fish**

**RAD-seq experimental and statistical considerations**

**Genomically enabling pipefish**

Stacks software pipeline *(this afternoon & evening)*
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
Vertebrate zygotes or embryos

28 day human

19h zebrafish

Dr. Catchen in his ‘following Phish Phase’
Functional Genomics
Modified from Gilbert 1998
Conditions of Existence

Evolution
  - Systematics
  - Ecology

Genetics
  - Population Genetics
  - Experimental Embryology

Population Genetics
  - Modern Synthesis
  - Evolutionary Genetics

Molecular Genetics
  - Developmental Genetics

Functional evolutionary genomics

Unity of Type
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process
Studying cancer as an evolutionary process

Wild Type

Mutant
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?

Where does the basis for evolutionary novelties reside in genomes?

Four fundamental processes in evolution

Origin of genetic variation;
mutation
migration

Sorting of variation;
genetic drift
natural selection
Genetic drift is a null model
Population genomics

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

Genome-wide background provides more precise estimates:
- Demographic processes (e.g. $N_e$)
- Phylogeography

Outliers from background indicate:
- Selective sweeps
- Local adaptation
Population genomics of unordered markers

Population genomics of ordered markers

Hohenlohe et al. 2010. Int J Plant Sciences
Population genomics of ordered markers

(A) and (B) show different graphical representations of genetic data.

Hohenlohe et al. 2010. Int J Plant Sciences
Population genomics of ordered markers

(A) [Diagram]

(B) [Colorful sequence alignment]

(C) [Tree diagrams]

π, F_{ST}

i, ii, iii, iv, v, vi

Hohenlohe et al. 2010. Int J Plant Sciences
Population genomics of ordered markers

(A) 

(B) 

(C) 

(D) 

(E) 

position along genome

Hohenlohe et al. 2010. Int J Plant Sciences
How do we ‘genomically enable’ research studies of non-model organisms?

1. Genetic Markers & Genetic Maps
2. Physical Maps (genomes)
3. Transcriptomes
4. Gene Expression Analyses
5. Epigenetic analyses
In the field and in the lab until a few years ago....
Shouldn’t we just sequence everything?

(note - the answer to this question may be ‘yes’ soon, and if so I will stop at this slide. But until then....)
Why not sequence the entire genome??

- Still prohibitively expensive for many studies
  - Human height GWAS; over 15,000 individuals assayed
  - Identified many new regions contributing to the variation
  - Still only identified a fraction of the heritability

- For many studies 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 a lot of structural variation
Alternative approach -
Reduced representation NGS for genotyping

• Focus sequencing on homologous regions across the genome

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

• The cost will be a fraction of the cost of resequencing the genome
  • i.e. 1% genome coverage will be less than 1% the cost
  • often coverage is more even than whole genome sequencing

• Thousands of genomes to be assayed in just a few weeks

• WHY NOT - complete genomic sequence is often useful
  • when linkage disequilibrium blocks (LD) are very short
  • Inferring patterns of LD may be easiest with full sequences
Different flavors of Reduced Representation Library (RRL) Sequencing for genotyping

- **Common acronyms**
  - **RRL** - Reduced Representation Library
  - **GBS** - Genotyping By Sequencing
  - **CRoPS** - Complexity Reduction of Polymorphic Sequences
  - **RAD** - Restriction site Associated DNA

- All rely on restriction enzyme digestion
- RRL, CRoPS, and GBS use one or two restriction enzymes only
- RAD-seq uses a shearing step to more efficiently capture all restriction sites
- Incorporation of barcodes on adaptors for multiplexing
- **Aligned** against a reference genome or **assembled de novo**

- **Statistical issues**
  - new level of sampling variation (sequencing in addition to biological)
  - sequencing error and problems for aligning or clustering
What is RAD-seq?
(Restriction-site Associated DNA)

2007
Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers
Michael R. Miller,1 Joseph P. Dunham,2 Angel Amores,3 William A. Cresko,2 and Eric A. Johnson1,4

1Institute for Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA; 2Center for Ecology & Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, USA; 3Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA

2008
Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers
Nathan A. Baird1,3, Paul D. Etter1,3, Tressa S. Atwood3, Mark C. Currey3, Anthony L. Shiver1, Zachary A. Lewis1, Eric U. Selker1, William A. Cresko2, Eric A. Johnson1

1Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America; 2Institute of Neuroscience, University of Oregon, Eugene, Oregon, United States of America; 3Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, United States of America
What is RAD-seq?
(Restriction-site Associated DNA)

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

HiSeq2500 Illumina Lane:
160 million reads
HiSeq4000 Illumina Lane:
350 million reads
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

Diagram showing base pair distances and the TTAA motif.
Restriction Enzyme (RE) digestion and first adaptor ligation

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

A = Amplification primer
B = Sequencing primer
C = Barcode

* Important step here*
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
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
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
## Summary of plusses and minuses of RAD family

<table>
<thead>
<tr>
<th></th>
<th>Sheared RAD</th>
<th>Single or ddRAD</th>
<th>2b-RAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>plusses</strong></td>
<td>- Consistent reads</td>
<td>- Fewer steps</td>
<td>- Fewest steps</td>
</tr>
<tr>
<td></td>
<td>- Local assemblies</td>
<td>- Easy marker scaling</td>
<td>- Easy marker scaling</td>
</tr>
<tr>
<td></td>
<td>- Identify PCR duplicates</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>minuses</strong></td>
<td>- Shearing step</td>
<td>- Multiple enzymes</td>
<td>- Very short reads</td>
</tr>
<tr>
<td></td>
<td>- Scaling requires different enzymes</td>
<td>- Poor consistency</td>
<td>- PCR duplicates</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- PCR duplicates</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Additional benefits of random shearing in RAD

A) Restriction sites in genome

RAD tag sequence read

Sheared-end reads

B) Variable length RAD fragments isolated

C) 200-1200bp in length

Contigs assembled from the sheared-end reads for each RAD tag
Acquire paired-end sequence

Match to marker catalog

Collate/Assemble PE reads

Associate markers / PE contigs with ESTs

Assign orthology to: markers PE contigs ESTs

BlASTn

EST Library

Human Genome

Zebrafish Genome
Case studies of using RAD for an organism with a reference genome: population genomics of threespine stickleback
Threespine stickleback, *Gasterosteus aculeatus*
Threespine stickleback, *Gasterosteus aculeatus*
Threespine stickleback, *Gasterosteus aculeatus*

Pelvic Structure  
Lateral Plates

Rundle and McKinnon 2002
Threespine stickleback, *Gasterosteus aculeatus*
Laboratory mapping of large effect loci

Cresko et al. 2004. PNAS
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 problem is that laboratory mapping approaches are under-powered in stickleback

A question is whether population genomics studies can provide complementary and more complete information.
Population genomic structure of Oregon stickleback

590 Individuals
115,000 SNPs each

Catchen et al. 2013, Molecular Ecology
Stacks: an analysis tool set for population genomics

Julian M. Catchen, Paul A. Hohenlohe,† Susan Bassham,‡ Angel Amores,‡ and William A. Cresko*  
*Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA, †Biological Sciences, University of Idaho, Moscow, ID 83844-3611, USA, ‡Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254, USA

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

Julian M. Catchen,* Angel Amores,† Paul Hohenlohe,* William Cresko,* and John H. Postlethwait††  
*Center for Ecology and Evolutionary Biology and †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403
Population structure using PCA

PC 1 explains 89% of the overall variance
Phylogenetic relationship among populations

Winchester Creek

Pony Creek Reservoir

South Jetty

Millport Slough

Cushman Slough

Crooked River

South Twin Lake

Riverbend

Paulina
Population structure using Bayesian analysis (*Structure*)

Oregon Coast Oceanic

Willamette Valley

Oregon Coast

Central Oregon
Genome-Wide Association Study (GWAS) using RAD

Kristin Alligood and Mark Currey
Genome-Wide Association Study (GWAS) using RAD
CRISPR gene editing in stickleback to test associations
What genomic regions are associated with the different habitats?

How quickly can the allele frequencies change?
Signatures of natural selection in 13,000 years

Hohenlohe, Bassham et al. 2010. PLoS Genetics
Signatures of natural selection in 13,000 years

Bear Paw
(mean $F_{ST} = 0.121$)

Boot
(mean $F_{ST} = 0.112$)

Mud
(mean $F_{ST} = 0.117$)

Genomic location (mBases)

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

Different alleles

More often the same alleles

LGI

LGXXI
Global analysis of complete sequencing consistent with the Alaskan results

Global analysis of complete sequencing consistent with the Alaskan results
Shake rattle and evolve in 50 years
team earthquake

Susan Bassham

Julian Catchen

Emily Lescak

Mary Sherbick

Frank von Hippel
Middleton Island

1955  2008

2km

Photo Credits: BLM, E-Terra
Middleton Island

1955  2008

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

1955

2008

2km

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

110,000 SNPs per individual
>1000 Individuals
20 million genotypes
The majority of the genetic variation is partitioned between oceanic and freshwater fish.
The majority of the genetic variation is partitioned between oceanic and freshwater fish.
Extensive phenotypic change occurs even in populations with little overall differentiation.
Extensive phenotypic change occurs even in populations with little overall differentiation.
Structure analysis shows independent evolution even among populations on a single island.
Structure analysis shows independent evolution even among populations on a single island.
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?
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

FST
MoSe vs. Ocean
MoSW vs. Ocean
FW2 vs. Ocean
FW1 vs. Ocean
Ocean vs. Ocean
Haplotype diversity increases in freshwater

FW1 vs. Ocean
FW2 vs. Ocean

Middleton Island
Linkage disequilibrium is extensive between oceanic and freshwater stickleback, but not within.
Linkage disequilibrium is extensive between oceanic and freshwater stickleback, but not within certain regions. The images depict the linkage disequilibrium binned by SNP distance (20Kb buckets) for different chromosomes (LG 4, LG 7, LG 15, LG 21) and polymorphisms (mi08, mi12, mi06, mi11, mi17) for marine and freshwater conditions.

The diagrams show the distribution of linkage disequilibrium across the genome, with red indicating higher values and blue indicating lower values. The mean D' values are binned by SNP distance, indicating the degree of association between different genetic markers.

- Chromosome 4
- Chromosome 7
- Chromosome 15
- Chromosome 21
Are these genomic blocks habitat specific?
From SNPs to haplotypes

- SNPs can be ordered into haplotypes
- Haplotypes provide deep & shallow evolutionary information
- Phasing genotypes within and among RAD sites
- Genotype imputation for missing SNPs
Coalescent analysis using RAD-seq data
Coalescent analysis using RAD-seq data

Noah A. Rosenberg & Magnus Nordborg
Neutral coalescent expectations
Natural selection and the coalescent

Divergent Selection

Balancing Selection

Noah A. Rosenberg & Magnus Nordborg
RAD-seq coalescent in stickleback

Thom Nelson & Julian Catchen
RAD-seq coalescent in stickleback

Ocean
Freshwater
RAD-seq coalescent in stickleback - UNIFRAC

- FW-specific branch
- OC-specific branch
- Shared branch

UNIFRAC distance $\approx 0.25$

UNIFRAC distance $= 1.0$
RAD-seq coalescent in stickleback - UNIFRAC
Many haplotypes are habitat specific
And correspond with signatures of divergent selection.
Many haplotypes in regions subject to divergent section are quite old
What can explain such rapid evolution and haplotype structure?

Is the stickleback genome architecture partly responsible?

Julian Catchen, Susie Bassham and Thom Nelson
Genome Assembly

Paired-end Alignments

Genetic Map Construction

PCR Screening of Breakpoints

Paired-end Alignments

Genetic Map Construction

PCR Screening of Breakpoints
Genome Assembly

<table>
<thead>
<tr>
<th></th>
<th>N50</th>
<th>Max</th>
<th>Total</th>
<th>Median Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bp</td>
<td>bp</td>
<td>Mb</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>17,417</td>
<td>199,905</td>
<td>488.8</td>
<td>24.6x</td>
</tr>
<tr>
<td>M3</td>
<td>18,982</td>
<td>192,283</td>
<td>472.5</td>
<td>26.5x</td>
</tr>
<tr>
<td>M5</td>
<td>15,555</td>
<td>238,768</td>
<td>456.4</td>
<td>24.1x</td>
</tr>
<tr>
<td>M7</td>
<td>15,534</td>
<td>254,734</td>
<td>473.4</td>
<td>25.8x</td>
</tr>
</tbody>
</table>
Illumina Paired-end Reads

Reference Genome

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>161,305,595 pairs</td>
<td>144,396,898 pairs</td>
<td>131,471,548 pairs</td>
<td>150,786,462 pairs</td>
</tr>
</tbody>
</table>
**F1 Pseudo-testcross using RAD-seq**

**Male Parent**
- **A₁** B₁ C₁
- **A₂** B₁ C₂

**Female Parent**
- **A₁** B₁ C₁
- **A₁** B₂ C₃

**Progeny**
- **A₁** B₁ C₂
- **A₁** B₁ C₁
- **A₂** B₁ C₃
- **A₁** B₁ C₁
- **A₁** B₂ C₁
- **A₂** B₁ C₂

**Combined map**
- **A** B C

**Female map**
- **B** C

**Male map**
- **A** C

**Infer**
- **CA** B C

**93 progeny**
- 66,071 loci
- 5,351 markers

**93 progeny**
- 45,301 loci
- 3,927 markers
Linkage Group XXI

RS (Marine)       Boot (Freshwater)

Genome Arrangement

Inverted
1. Previous work has shown that the freshwater genomes evolve in 13,000 years.

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

3. Much of the divergence involves soft sweeps.

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

5. Many haplotypes are habitat specific and are quite ancient.

6. Haplotypes often coincide with structural variation across the stickleback genome
Hypothesis: Old genomic architecture variation is a product of the metapopulation structure of stickleback, and this architecture strongly influences subsequent rapid evolution.
Considerations for RAD-seq studies
Experimental design considerations for RAD

Tradeoffs:

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

Number of sites versus Depth of sequencing per site versus Number of samples

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

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

25 to 50x average coverage per RAD locus is a good goal
Experimental design considerations for RAD

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

How many tags do I need?

Things to consider

**Choice of enzyme and genome size**

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

Genomes are biased:

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

Restriction enzyme recognition site

Reference genome sequence

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

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

\[
L(n_1n_2\text{het}) = P(n_1, n_2, n_3, n_4) = \frac{n!}{n_1!n_2!n_3!n_4!} \left( 0.5 - \frac{\varepsilon}{4} \right)^{n_1} \left( 0.5 - \frac{\varepsilon}{4} \right)^{n_2} \left( \frac{\varepsilon}{4} \right)^{n_3} \left( \frac{\varepsilon}{4} \right)^{n_4}
\]
Making statistics continuous across the genome

Kernel-smoothing average of summary statistics along genome

Bootstrap resampling to estimate significance of moving average
Overall pipeline
‘Bias’ in RAD-sequencing

INTRODUCTORY STATISTICAL PRINCIPLES

Fig 3.1 Fictitious histogram (a) and (b) normal and (c-d) log-normal probability distributions.

curve represent the probability of values falling in the associated interval. 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 accumulated values follows a "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.

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

\[e = 2.7182....\]

\[\pi = 3.1415....\]
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 accumulated values follows a ‘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

Abstract

Reduced representation genome-sequencing approaches based on restriction digestion are enabling large-scale marker generation and facilitating genomic studies in a wide range of model and nonmodel systems. However, sampling chromosomes based on restriction digestion may introduce a bias in allele frequency estimation due to nonrandom cleavage that ensures primarily the same random DNA shearing. Enzyme digestion results in genome-wide coverage, but is highly sensitive to the choice of restriction enzyme, as opposed to random DNA shearing. RADseq has an additional potential problem, caused by PCR bias in library construction, sequencing errors, and sample pooling. These all are known to distort allele frequency estimation. Nonrandom DNA shearing or pooling is expected to introduce biases. Here we study the potential impact of nonrandom missing data by developing a coalescent-simulation framework to mimic the biased recovery of chromosomes in restriction-based short-read sequencing experiment if mutations occur within the recognition sequence in the most divergent haplotype; the genealogy is thus truncated to point ‘X’ and results in incomplete allele frequency spectrum. We use simulations to study how levels and patterns of genomic variation necessitate nonrandom sampling ascertainment bias whose effects have not been re-sequencing projects. RADseq has an additional potential problem, caused by PCR bias in library construction, sequencing errors, and sample pooling. These all are known to distort allele frequency estimation. Nonrandom DNA shearing or pooling is expected to introduce biases.

Methods

We used Hudson's ms (Hudson 2002) to simulate 10 kb lifespans of genomes with a restriction enzyme, as opposed to random DNA shearing. Enzyme digestion results in genome-wide coverage, but is highly sensitive to the choice of restriction enzyme, as opposed to random DNA shearing. RADseq has an additional potential problem, caused by PCR bias in library construction, sequencing errors, and sample pooling. These all are known to distort allele frequency estimation. Nonrandom DNA shearing or pooling is expected to introduce biases.

Results

Overall, we find that RADseq nonrandomly subsamples the genomes, yielding results similar to what might be expected from re-sequencing projects. RADseq differs from other genome-sequencing approaches in that DNA fragments for construction of a library of sequences are generated by digesting genomes with a restriction enzyme, as opposed to random DNA shearing. Enzyme digestion results in genome-wide coverage, but is highly sensitive to the choice of restriction enzyme, as opposed to random DNA shearing. RADseq has an additional potential problem, caused by PCR bias in library construction, sequencing errors, and sample pooling. These all are known to distort allele frequency estimation. Nonrandom DNA shearing or pooling is expected to introduce biases.

Discussion

While powerful and widely used, RADseq is still underdeveloped compared to whole-genome sequencing, largely uncharacterized biases. Potential problems arising from PCR bias in library construction, sequencing errors, and sample pooling are known to distort allele frequency estimation. Nonrandom DNA shearing or pooling is expected to introduce biases. Here we study the potential impact of nonrandom missing data by developing a coalescent-simulation framework to mimic the biased recovery of chromosomes in restriction-based short-read sequencing experiment if mutations occur within the recognition sequence in the most divergent haplotype; the genealogy is thus truncated to point ‘X’ and results in incomplete allele frequency spectrum. We show that RADseq nonrandomly subsamples the genomes, yielding results similar to what might be expected from re-sequencing projects.
Bias in RAD-sequencing; genetic diversity

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.

![Graph showing the relationship between chromosome sampling depth and AFS](image)

**Fig. 2** (A) True and estimated values of $p$ (red) and $\theta_w$ (blue) from in silico recovery of chromosomes in restriction-based short-read sequencing (RAD)seq as a function of chromosome sampling depth for $h = 0.01$ per bp without recombination. Here, the simulation average of $h$ is 1 per 100 bp sequence read. Shaded regions show the 95% bootstrap percentile confidence intervals (1000 simulations) for the mean of true values of $p$ (solid red) and $\theta_w$ (solid blue) and estimated values of $p$ (shaded red) and $\theta_w$ (shaded blue) from in silico RADseq. ‘Chromosome sampling depth’ refers to the number of chromosomes that are actually sampled (have intact restriction sites) in the in silico experiment, and ‘true’ values are those calculated using the complete data for the same markers. The histograms in A (no recombination) and B (with recombination, $q = h$) show the proportion of each chromosome sampling depth in the data and indicate that most markers are highly sampled with these simulation parameters, especially for lower values of $h$ (B).

<table>
<thead>
<tr>
<th>Protocol</th>
<th>$h_{per \text{bp}}$</th>
<th>Mean Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$h_{wa}$</td>
</tr>
<tr>
<td>Standard</td>
<td>0.0001</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
</tr>
<tr>
<td>Double digest</td>
<td>0.0001</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
</tr>
</tbody>
</table>

Results from two different simulation parameters of $h$ 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
Bias in RAD-sequencing; genetic diversity

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.

Chromosome sampling depth

<table>
<thead>
<tr>
<th>Protocol</th>
<th>$h_w$ per bp</th>
<th>$\text{Mean}$</th>
<th>$\text{Variance}$</th>
<th>Recombination</th>
<th>No recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.994</td>
<td>0.995</td>
<td>0.991</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.987</td>
<td>0.982</td>
<td>0.988</td>
<td>0.984</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.956</td>
<td>0.933</td>
<td>0.940</td>
<td>0.909</td>
</tr>
<tr>
<td>Standard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
<td>0.838</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
<td>0.830</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
<td>0.833</td>
<td>0.772</td>
</tr>
<tr>
<td>Double digest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.835</td>
<td>0.836</td>
<td>0.838</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>0.858</td>
<td>0.851</td>
<td>0.830</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.829</td>
<td>0.797</td>
<td>0.833</td>
<td>0.772</td>
</tr>
</tbody>
</table>

Results from two different simulation parameters of $h$ 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
Bias in RAD-sequencing; Fst

Although the levels and patterns of genetic variation in neutral loci that are linked to locally adapted alleles will depend on demographic and selective circumstances, it is interesting to consider outliers in the distributions of summary statistics as potential metrics for detecting positive selection and local adaptation. In particular, high $F_{ST}$ may indicate that a locus is in linkage disequilibrium with locally adapted alleles. However, we show that missing data may inflate $F_{ST}$ values, and rates of false positives quickly increase as the chromosome sampling depth cutoff decreases, especially when chromosome sample sizes amongst populations are allowed to vary as little as 20% (Fig. 6). Thus, it may be wise to constrain analyses to loci with complete chromosome sampling, but of loci in the upper 5% tail of true $F_{ST}$ distribution, only 13%, 11% and 5% have complete chromosome sampling in both populations for $Nm=10$, $Nm=1$ and $Nm=0.1$ respectively.

Within a population, genomic regions with low nucleotide diversity and left-skewed site frequency spectra may indicate the presence of a recent selective sweep via the hitchhiking effect (Maynard-Smith & Haigh 1974), or strong purifying selection (Charlesworth et al. 1993). We explored the effect of missing data on outlier analyses involving the commonly used diversity statistics $h_w$ and $p$. Specifically, we examined the lower 5% tail of the distributions of these statistics to assess how missing data affects false positive and false negative rates. Using different sampling depth cutoffs, rates of false positives and false negatives increase with the inclusion of loci with missing data for both the standard RADseq protocol (Fig. 7) and the double-digest protocol (Fig. S7, Supporting information). Similar analyses with lower values of $h_w$ (0.001 per bp and lower) were not possible since the 5% quantile of summary statistics contained the majority of loci due to low levels of polymorphism.

Since loci with missing data have more false positives and negatives, a possible solution is to limit outlier analyses to loci with complete chromosome sampling. If outliers were evenly distributed across loci irrespective of missing data, 5% of loci in each sampling depth category would be outliers. However, in agreement with the results presented in Table 1, loci with complete sampling have slightly decreased diversity and are

$\begin{align*}
\text{(A)} & \quad \text{Nm} = 10 \\
\text{(B)} & \quad \text{Nm} = 1 \\
\text{(C)} & \quad \text{Nm} = 0.1
\end{align*}$

$\begin{align*}
\text{Estimated } F_{ST} & \text{ as a function of chromosome sampling depth cutoff per population (each consisting of 50 chromosome total) for three different migration rates: } Nm = 10 \text{ (A), } Nm = 1 \text{ (B) and } Nm = 0.1 \text{ (C). The dashed line is the true simulation average. Here, we condition on sample sizes being the same in both populations to avoid inflated estimates of } F_{ST}. \text{ Note that the Y-axes do not start at zero to more clearly illustrate differences between true and estimated values.}
\end{align*}$
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 \( p_t \), 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
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 \( p_t \), 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>
<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.
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 $p_t$, 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.

![Image of table]

Table 1: Comparison of estimated values of summary statistics ($\theta_{we}$ or $\pi_e$) when all chromosomes are sampled to true simulation averages ($\theta_{wa}$ or $\pi_a$).

<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.
Why is ddRAD so much more biased?
RAD-seq and phylogenetics of divergent species

- 5.1% of its restriction sites with D. melanogaster, and 50.5% of which are conserved in its two closest relatives.

- Conserved target sites of the restriction enzyme Sbf1 were counted in each pairwise alignment.

- Notably, these estimates are conservative because its 8 bp, GC-rich recognition site is seq studies because its 8 bp, GC-rich recognition site is.

- Divergence time and RAD loci ranging from 5.4 to 63 My (Tamura et al. 2004).

- We further show here that sequence clustering based on the 12 genomes, potentially from hundreds of specimens in a genome with.

- Figure 1 shows the relationship between restriction site and restriction sites conservation. Complete genomes and genomes, with divergence times ranging from 5 to 63 million years. These simulations show that RAD-seq allows the recovery of the known (whole-genome-based) phylogeny.

- Thus, we identified clusters containing more than 50 sites conserved between species; (2) the flanking regions must be sufficient for RAD-seq to be suitable for phylogenetic inference.

- Most nuclear markers lack resolution at short evolution- incomplete lineage sorting and introgression, different loci contain enough phylogenetic signal.

- The RAD-seq technique was initially designed to generate code sequence identifying a sample (up to 8 bp long) and therefore the number of loci usable for phylogenetic inference. Here, we assess the suitability of RAD-seq for phylogeny using a simulated experiment on.

- The procedures for data analysis were previously described (Rubin et al. 2012; Emerson et al. 2010; Hohenlohe et al. 2010). However, not all fragments code for amino acids, and therefore the number of loci usable for phylogenetic inference. Here, we assess the suitability of RAD-seq for phylogeny using a simulated experiment on.

- We further show that clustering based on these results with the SiLiX software (Miele et al. 2011), maximize the potential phylogenetic signal (see below for a discussion

- Restriction sites located in repeated regions may be problematic for phylogenetic analysis and should be problematic for phylogenetic analysis and should.

- The 8 bp restriction site followed by 85 bp of usable data.

- The 8 bp restriction site followed by 85 bp of usable data.

- Of each restriction.

- Numbers next to divergence time indicate divergence time.

- 5 MY

- 13 MY

- 44 MY

- 55 MY

- 62–63 MY

- Numbers next to divergence time indicate divergence time.

- 5 MY

- 13 MY

- 44 MY

- 55 MY

- 62–63 MY

- Number of sites shared with D. melanogaster

- Number of mutations/site at four fold degenerate sites
RAD-seq and phylogenetics of divergent species

### Table 1: Orthologous tags in pairwise alignments

<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>
RAD-seq and phylogenetics

Abstract

Recent studies, published as this article was in preparation, have repeatedly proved the efficiency of RAD-seq for phylogenetic inference (Hohenlohe et al. 2010; Emerson et al. 2010). However, as it does not require preliminary knowledge on the taxa under study (Baxter et al. 2011), RAD-seq can be considered as a promising technique for large-scale phylogenetic studies.

The RAD-seq approach consists in designing restriction enzymes that cut DNA at sites present in evolutionary conserved restriction sites (Rubin et al. 2012; Wagner et al. 2012). This technique is particularly attractive for phylogenetic studies on Drosophila species, in which a large number of restriction enzymes is available. However, RAD-seq requires several conditions to be fulfilled: (1) enough restriction sites must be conserved between species; (2) the flanking regions must be sufficiently conserved for RAD-seq to be suitable for phylogenetic inference; (3) the resulting alignments must contain enough phylogenetic signal; and (4) the resulting alignments must be sufficiently conserved for homology to be detectable by sequence similarity; and (3) the resulting alignments must be sufficiently conserved for homology to be detectable by sequence similarity.

Caveats. Specifically, the following conditions must be fulfilled for the utilization of RAD-seq to compare genomes from different species: (1) enough restriction sites must be conserved for RAD-seq to be suitable for phylogenetic inference; (2) the flanking regions must be sufficiently conserved for RAD-seq to be suitable for phylogenetic inference; (3) the resulting alignments must contain enough phylogenetic signal; and (4) the resulting alignments must be sufficiently conserved for homology to be detectable by sequence similarity.

Introduction

Here, we compare the results obtained with the standard RAD-seq approach with those obtained with a modification of the approach, namely the introduction of orthologs-only clusters (Emerson et al. 2010; Morris et al. 2012). This approach improves the recovery of orthologous RAD loci compared with previously proposed methods (Baxter et al. 2011). The orthologs-only clusters approach is particularly useful for phylogenetic studies on Drosophila species, in which a large number of restriction enzymes is available.

Methods

In a typical RAD-seq experiment, DNA samples from several individuals are tagged with molecular identifiers and sequencing coverage per locus per individual is given by the following formula:

\[ \text{Coverage} = \frac{\text{Number of reads}}{\text{Number of bases}} \]

where \(i\) is the total number of reads,

\[ \text{Number of bases} \]

and \(n\) is the number of sequenced bases.

Results

The phylogenetic tree of the 12 Drosophila species included in this study is shown in Figure 1. The tree was constructed using the concatenated alignments of all RAD loci and the bootstrap values were estimated using PhyML 3.0 (Guindon et al. 2010) with 100 replicates.

Discussion

The phylogenetic tree constructed using the RAD-seq approach is consistent with the species tree. The bootstrap values were high for all nodes, indicating that the tree is robust to the potentially confounding effects of sequencing errors, heterozygosity, and low coverage. We further show that clusters with at least 3 sequences are filled (Rubin et al. 2012; Wagner et al. 2012). Here, we show that clusters with at least 2 sequences are also filled.

Conclusions

In conclusion, the RAD-seq approach is a promising technique for phylogenetic studies on Drosophila species. The orthologs-only clusters approach improves the recovery of orthologous RAD loci compared with previously proposed methods.

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

Marie Cariou, Laurent Duret & Sylvain Charlat

Ecology and Evolution

Open Access
What if you don’t have a genome sequence?

Genomically enabling very non-model organisms: RAD-seq can help
Julian Catchen, Allison Fuiten, Susie Bassham, Clay Small and Adam Jones
Seahorses, sea dragons and pipefishes
Gasterosteidae and Syngnathidae are historically considered to be closely related

Wilson et al. 2003
Gulf Pipefish

*Syngnathus scovelli*

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

www.bio.tamu.edu/USERS/ajones/charlyn.html
Fishbase.org
We’re really interested in the head and body axis.
Few teleost genomes are available

Gasterosteiformes: only stickleback

- Zebrfish
- Medaka
- Stickleback
- Two pufferfish
Solution: ‘genomically enable’ pipefish

1) A high quality transcriptome

2) Very dense RAD genetic map

3) Deep coverage shotgun sequencing of genome

4) Order genomic and transcriptomic contigs against the RAD reference map
Pipefish Transcriptome
Building an EST database in pipefish

Pipefish embryonic mRNA

Illumina sequencing:
100 nt, paired-end

200 million reads (two lanes)

Assembly of transcripts
Transcriptome

30,000 solid contigs

Mean depth of coverage = 24X

Nearly all of the expected genes in the genome
Very large genes were represented.

Ankyrin3  Dystonin
Pipefish Genetic Map
Genetic map workflow

Generated an F1 family of 103 individuals
RAD sequenced the parents and offspring
Analyzed the data using Stacks
Paired end local assemblies
Output to JoinMap format
Created Linkage map
The pipefish genetic map is closed; 22 LGs
6000 segregating SNPs; 30,000 RAD sites
Pipefish Genome Project
Genome workflow

Generated DNA from a single individual

Random Illumina shotgun sequencing

Removed highly repetitive kmers

Produced several different genome assemblies
Illumina genomic libraries for pipefish genome

paired end 101bp

500-700bp

500-1200bp

mate pair

4500-7500bp

150-250bp

overlapping

paired end RAD

ACTCTC

500-1200bp

ACTCTC

150-250bp

25x

2x

40x

15-25x of 3% of the genome
de novo Genome Sequencing

Single-end Sequencing

TAAAGAAAAACATTCCGTTCCCATGGCGATGGCTCGGTGGCCTGAGGCGGCTCTGAGATGGCTGCCGGGAGTGCTGACAGGCCTGTGTCAGAGCAGAATTTCCACCCGGCCATTAAGGATCACTCCGTCTCTTCACCCCTTGA
de novo Genome Sequencing

Paired-end Sequencing
Mate-pair Sequencing
Pipefish genome assembly version 0.99
Nearly the whole genome is covered

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Scaffolds</th>
<th>Contigs</th>
<th>Scaffold N50</th>
<th>Contig N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>All (66.6x)</td>
<td>1,820</td>
<td>25,075</td>
<td>796,183</td>
<td>23,012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Max</th>
<th>Average Length</th>
<th>Total Length</th>
<th>Gap Length</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,676,603</td>
<td>162,928</td>
<td>296,529,585</td>
<td>27,303,839</td>
<td>(8.39%)</td>
</tr>
</tbody>
</table>
Overall Conclusions

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

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

Open Source Genomics provides a suite of breakthrough technologies
- the molecular approaches are not as daunting as they first appear
- analytical and computational approaches are challenging
  - New software tools can help, but knowledge of Unix and Scripting is essential
  - Also essential to be comfortable with classical and modern statistics
Acknowledgments

• Past and present lab members **Paul Hohenlohe, Thom Nelson**, Joe Dunham, Nicole Nishimura & **Mark Currey**
• Collaborators **Eric Johnson**, Patrick Phillips, **Chuck Kimmel**, **John Postlethwait**
• Funding from NSF & NIH, as well as Keck & Murdock Foundations
Stacks is a software pipeline for building loci out of a set of short-read sequenced samples. Stacks was developed for the purpose of building genetic maps from RAD-Tag Illumina sequence data, but can also be readily applied to population studies, and phylogeography.

G3: Genes, Genomes, Genetics

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

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

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

**PROCESSED RADTAGS**

1. Locus 1
   - Haplotypes: C A
   - Genotypes: A B

2. Locus 2
   - Haplotypes: A C
   - Genotypes: A B

3. Locus N
   - Haplotypes: C G
   - Genotypes: A B

**GENOTYPES**

**POPULATIONS**

**REFERENCE GENOME**

**U STACKS**

**P STACKS**

**C STACKS**

**S STACKS**

**GENETIC MAPPING**
Stacks workflow
<table>
<thead>
<tr>
<th>Id</th>
<th>SNP</th>
<th>Consensus</th>
<th>Matching Parents</th>
<th>Progeny</th>
<th>Marker</th>
<th>Ratio</th>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>~103</td>
<td>Yes</td>
<td>2</td>
<td>92/91</td>
<td>ab/ac</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

### SNPs

- **Column:** 52; G/A
- **Column:** 70; T/G

### Alleles

- a: GT
- b: GG
- c: AG

### Matching Samples

View: □ Haplotypes  □ Allele Depths  □ Genotypes
<table>
<thead>
<tr>
<th>Id</th>
<th>SNP</th>
<th>Consensus</th>
<th>Matching Parents</th>
<th>Progeny</th>
<th>Marker</th>
<th>Ratio</th>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><code>^103 annotate</code></td>
<td>TGGAGGGAGCTCGCAGTGGGAGCTCCACTTCAGTGAGAGGCACCCACATTTTACACATCAATTTCTCCC</td>
<td>2</td>
<td>92 / 91</td>
<td>ab/ac</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

**SNPs**
- Column: 52; G/A
- Column: 70; T/G

**Alleles**
- a: GT
- b: GG
- c: AG

**Matching Samples**
View: Haplotypes, Allele Depths, Genotypes
### Batch #1 [2011-08-10; 80bp Lepisosteus oculatus F1 Genetic Map RAD-Tag Samples]

**RAD-Tag Sample #2 [female]**

#### Sequence #73

<table>
<thead>
<tr>
<th>Catalog ID</th>
<th>Depth</th>
<th>Column:</th>
<th>SNPs</th>
<th>Alleles</th>
<th>Deleveraged?</th>
<th>Lumberjackstack?</th>
<th>Blacklisted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>#103</td>
<td>26x</td>
<td>52</td>
<td>G/A</td>
<td>AG</td>
<td>46.15%</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70</td>
<td>T/G</td>
<td>GT</td>
<td>53.85%</td>
<td>False</td>
<td>False</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Seq ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>consensus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 primary</td>
<td>CAGTC_2_0018_768_1365_1</td>
<td>...</td>
</tr>
<tr>
<td>2 primary</td>
<td>CAGTC_2_0029_1628_1195_1</td>
<td>...</td>
</tr>
<tr>
<td>3 primary</td>
<td>CAGTC_2_0053_1692_1388_1</td>
<td>...</td>
</tr>
<tr>
<td>4 primary</td>
<td>CAGTC_2_0054_1692_1388_1</td>
<td>...</td>
</tr>
<tr>
<td>5 primary</td>
<td>CAGTC_2_0055_1524_1186_1</td>
<td>...</td>
</tr>
<tr>
<td>6 primary</td>
<td>CAGTC_2_0064_1356_1854_1</td>
<td>...</td>
</tr>
<tr>
<td>7 primary</td>
<td>CAGTC_2_0078_1976_1542_1</td>
<td>...</td>
</tr>
<tr>
<td>8 primary</td>
<td>CAGTC_2_0081_1658_1374_1</td>
<td>...</td>
</tr>
<tr>
<td>9 primary</td>
<td>CAGTC_2_0082_1658_1374_1</td>
<td>...</td>
</tr>
<tr>
<td>10 primary</td>
<td>CAGTC_2_0083_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>11 primary</td>
<td>CAGTC_2_0084_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>12 primary</td>
<td>CAGTC_2_0085_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>13 primary</td>
<td>CAGTC_2_0086_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>14 primary</td>
<td>CAGTC_2_0087_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>15 primary</td>
<td>CAGTC_2_0088_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>16 primary</td>
<td>CAGTC_2_0089_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>17 primary</td>
<td>CAGTC_2_0090_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>18 primary</td>
<td>CAGTC_2_0091_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>19 primary</td>
<td>CAGTC_2_0092_1484_1904_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0016_86_1022_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0017_86_1022_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0018_86_1022_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0019_86_1022_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0020_86_1022_1</td>
<td>...</td>
</tr>
<tr>
<td>secondary</td>
<td>CAGTC_2_0021_86_1022_1</td>
<td>...</td>
</tr>
</tbody>
</table>

last updated: Sun Jan 8 08:59:35 PST 2012