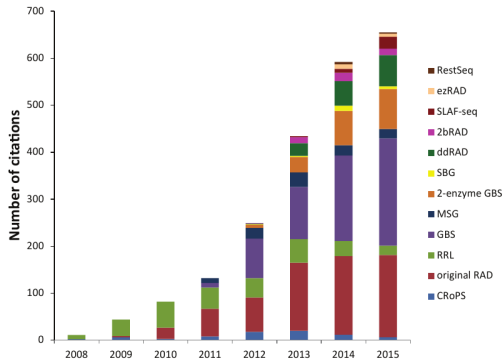


# Non-model species and RAD-sequencing

Alexander Jueterbock

June 2016

# RAD-Seq - young and successful NGS methods



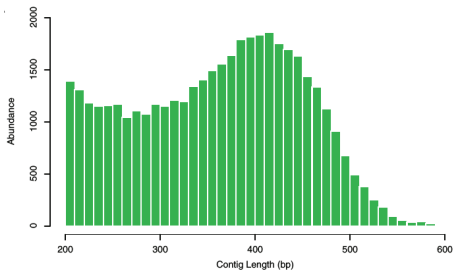
(Andrews et al., 2016)

# Purpose of RAD-seq

- Genome-reduction method to fragments adjacent to restriction enzyme recognition sites.
- Increases depth of coverage per locus compared to whole genome sequencing
- High-throughput genotyping of populations (multiplexing using barcoding) at relatively low cost.
- Makes genome-scale population genetic studies possible for non-model species lacking a reference genome.

# Reductive *de novo* genome sequencing and SNP identification

- RAD-Seq of the sunflower genome (Illumina)
  - 44.7M reads (PE:40bp x 80bp)
- *De novo* assembly of ca. 15.2 Mb in >42,000 contigs
- Identified >94,000 putative SNPs across six lines



(Pegadaraju et al., 2013)

# Genome-wide association study (GWAS)

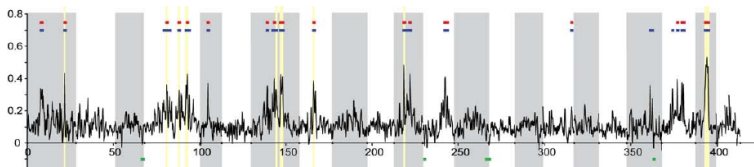
- No reference genome previously available
- identified >100,000 SNPs across 138 genotypes
- Related SNPs to 17 phenotypic traits in a field trial
- Increasing flexibility and speed of crop breeding



Figure: *Miscanthus sinensis*

# Population genomics and parallel adaptive differentiation in threespine sticklebacks

- Reference genome available
- >45,000 SNPs across 100 individuals ('genotyping by sequencing')
- Consistent signatures of selection between two oceanic and three freshwater populations
- Identified 31 candidate genes of evolutionary significance



**Figure:**  $F_{ST}$  for SNPs in sliding windows across the genome between oceanic and freshwater populations

# Original RAD-Seq protocol

- Developed by (Baird et al., n.d.; Miller et al., 2007).
- DNA fragments adjacent to restriction enzyme recognition sites

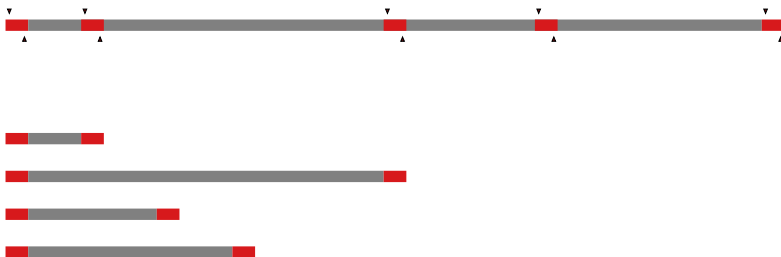


▼  
5' GAATTC 3'  
3' CTTAAG 5'

EcoRI recognition site



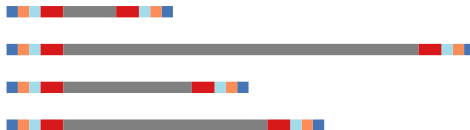
# Step 1: cut DNA



- Note: Bias in GC content of restriction site samples the genome non-randomly



## Step 2: ligate P1 adapter

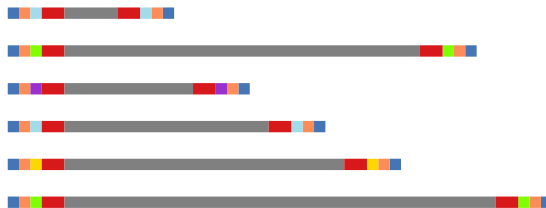


Amplification primer site

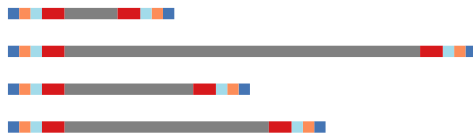
Sequencing primer site (Illumina-specific)

Barcode

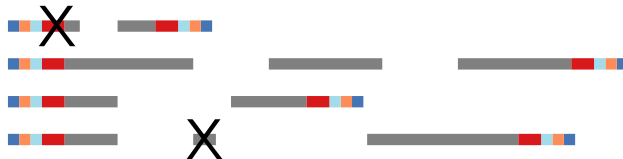
# Barcoding allows to pool samples



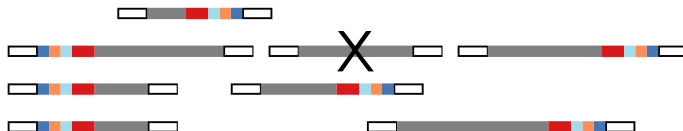
## Step 3: Shearing and size selection



Sonication with ultrasonic frequencies ( $>20$  kHz)



# Step 4: Ligation of P2 adapter with 'Y' structure

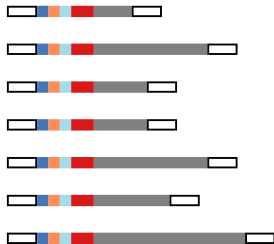


P2 adapter: AGATCGTCCGA  
TCTAGCGTCCT

P2 primer: TCTAGCGTCCT

P2 primer binds only when P2 primer site was completed by amplification starting from the P1 adapter (removes Y-structure)

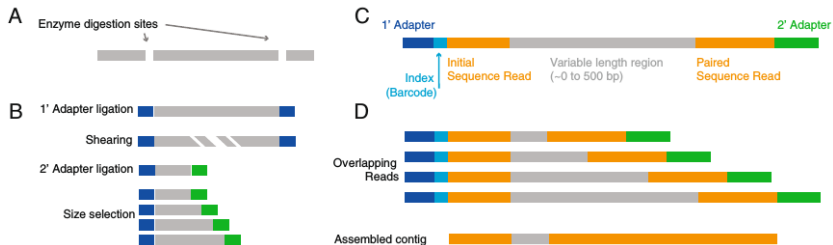
## Step 5: Sequence amplified reads on Illumina



Sequence 100 or so bp on Illumina

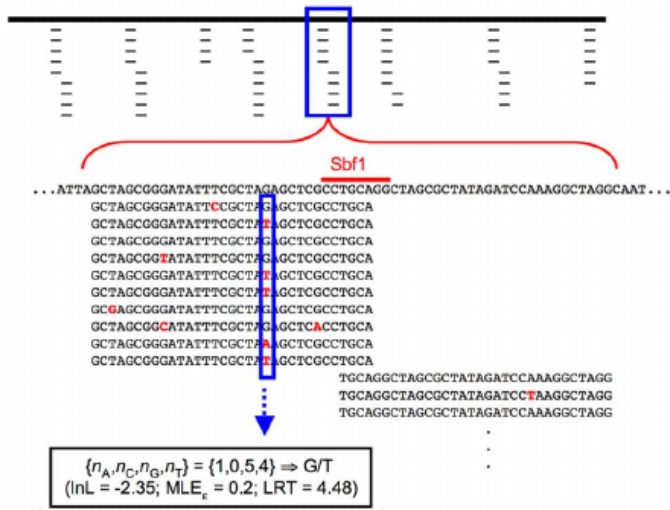
Random shearing of 3'ends helps to detect PCR duplicates

# Paired-end sequencing of RAD-tags allows for *de novo* genome sequencing



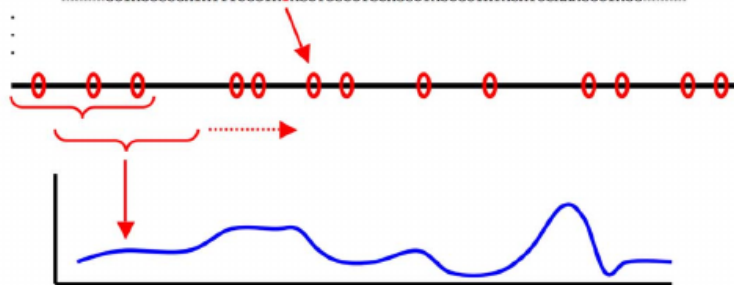
(Pegadaraju et al., 2013)

# Calling SNPs from RAD-tags



(Hohenlohe et al., 2010)

Ind 1 nnnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCTGCGAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnnn  
nnnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCGAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnnn  
Ind 2 nnnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCGAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn  
nnnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCGAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnnn  
Ind 3 nnnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCGAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnnn  
nnnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCGAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnnn

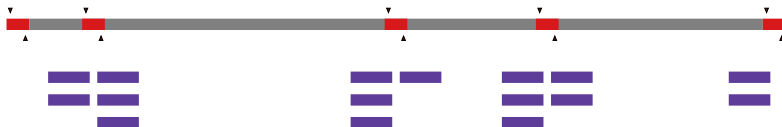


◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡



# Double-digest RAD-seq (Peterson et al., 2012)

## Single digest RAD-Seq



## Double digest RAD-seq



Sequencing of fragments:

- within a specific size range
- flanked by two different cutting sites

- EcoRI recognition site
- SbfI recognition site

# ddRAD compared to single-digest RAD sequencing

- 1 Rapid and 'cheap' protocol (8 hrs hands-on): Doesn't require difficult and high cost of shearing and enzymatic end-repair.

# ddRAD compared to single-digest RAD sequencing

- 
- 2 Lower number of loci but increased coverage and, thus, higher chance to target the same loci in different individuals.

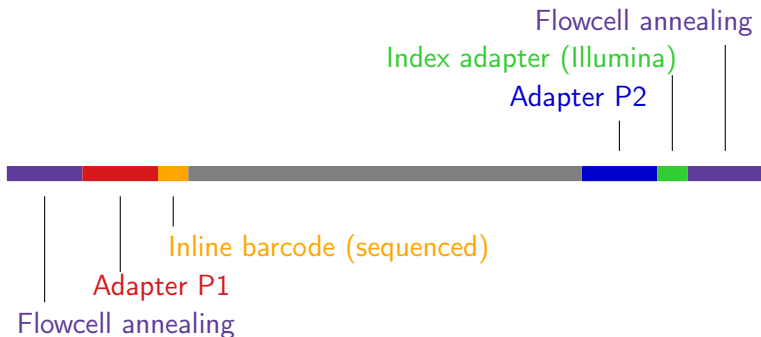
# ddRAD compared to single-digest RAD sequencing

- 
- 
- 3 Coverage expected to be equal among individuals and highest for fragment lengths targeted by size selection.

# ddRAD compared to single-digest RAD sequencing

- Combinatorial indexing allows to multiplex more individuals (up to 12 barcodes were affordable for single-digest RAD-Seq).

# Combinatorial indexing allows for high multiplexing levels in ddRAD-Seq



48 × 12 = 576 (multiplexing level)

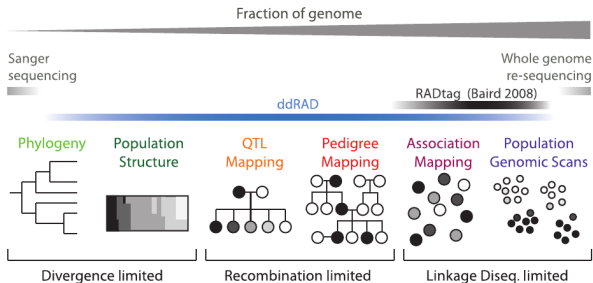
added first, with ligation of adapters, allows to pool samples

added second, with PCR primer, allows to combine multiple pools

# Great adjustability of the number of markers makes ddRAD suitable for a broader range of approaches than RAD-Seq

Number of markers adjusted by:

- Cutting frequency of restriction enzymes
- Size selection



# How to predict the number of fragments

Based on our own study on Guppy

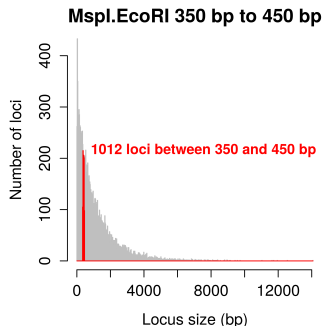
- Targeted coverage: 20x per individual
- Pooling: 60 individuals
- Sequencing output: 24M reads (12M fragments, minimum for Illumina v2 paired-end kits)
- Fragments per individual:  $12\text{M}/60 = 200,000$
- Target: **10,000** fragments (to reach a 20x coverage)

What combination of restriction enzymes to use to obtain the appropriate cutting frequency?



# *In silico* genome digestion

Simulate restriction enzyme digestion with the R package simRAD  
(Lepais and Weir, 2014)

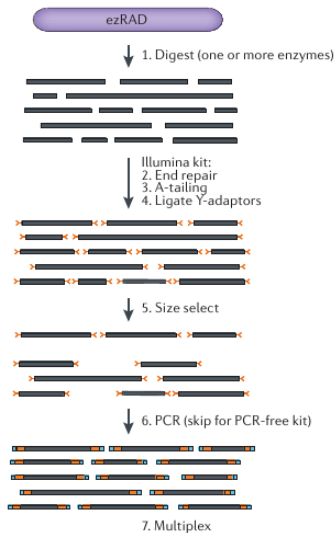


Based on 10% of the entire genome size

Without reference genome: evaluate double-digest fragments on  
Tape station

## 23 / 52

# ezRAD (Toonen et al., 2013)



(Andrews et al., 2016)

# ezRAD (Toonen et al., 2013)

## Advantages

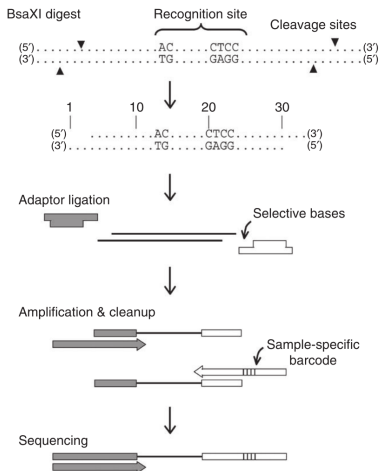
- non-PCR kits can avoid PCR duplication and bypass any PCR bias.

## Disadvantages

- All reads start with the same four bases (GATC).
  - Low diversity libraries can lead to poor read quality on Illumina sequencers. Use e.g. PhiX spiking or dark-cycling.

# 2bRAD (Wang et al., 2012)

- Type IIb restriction endonuclease to excise 36-bp fragments.
- Number of loci customized by base-selective adapters.



# 2bRAD (Wang et al., 2012)

## Advantages

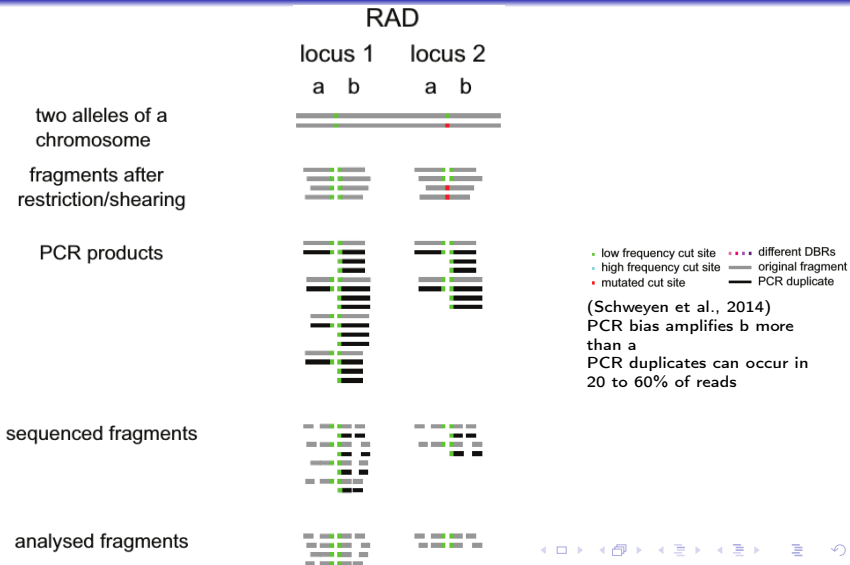
- Extremely simple and cost-effective: no purification or size selection.
- No biases due to fragment size selection.
- Sequencing either strand of the restriction fragments allows for the use of strand bias as a quality filtering criteria.

## Disadvantages

- 36-bp tags could be too short to be non-ambiguously mapped in highly duplicated genomes.

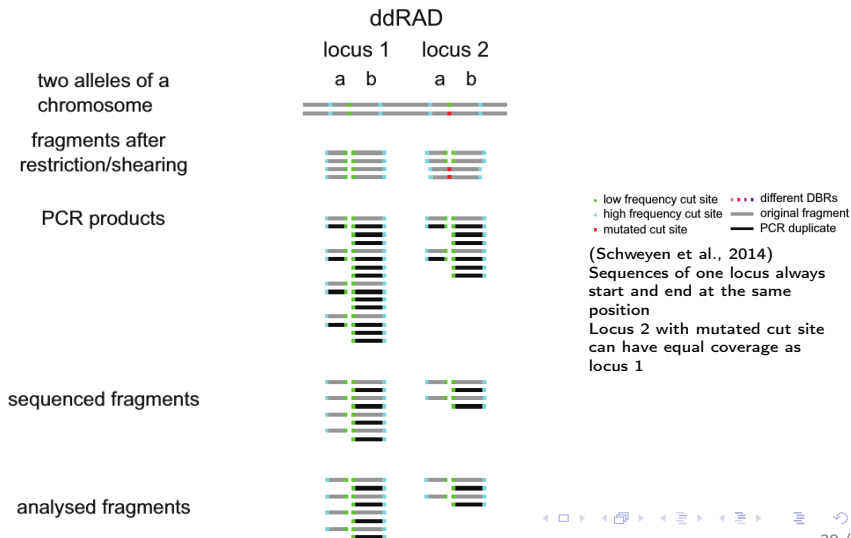


# Detect PCR duplicates by length in paired-end RAD sequencing





# PCR duplicates in ddRAD and ezRAD - not detectable



# Degenerate base regions detect PCR duplicates in ddRAD (Schweyen et al., 2014; Tin et al., 2014)

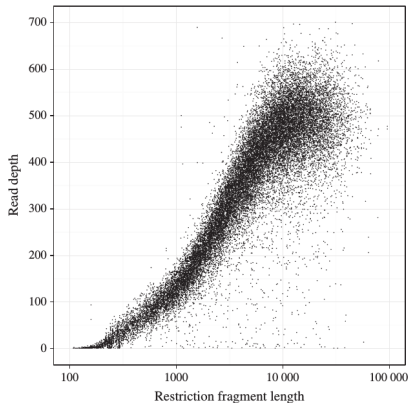


# Avoid PCR duplicates

- Reduce occurrence by lowering PCR steps
- Avoid PCR duplicates in ezRAD with Illumina PCR-free kits

# Shearing introduces bias in coverage

Bias in sequencing depth towards larger fragment sizes

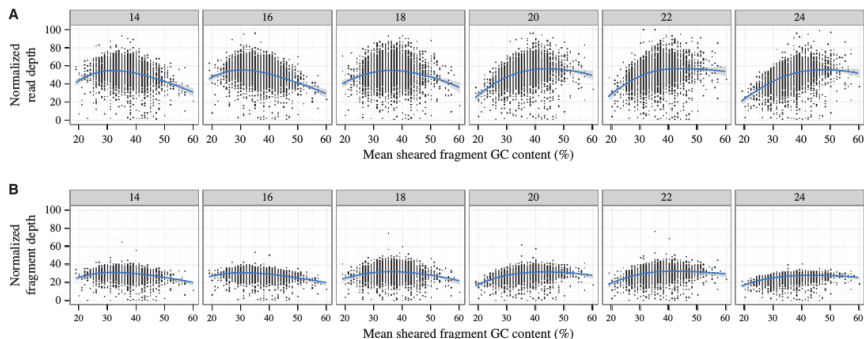


(Davey et al., 2013)

Reason: Fragments of <10 kb shear with lower efficiency

# Amplification bias in favor of high GC content

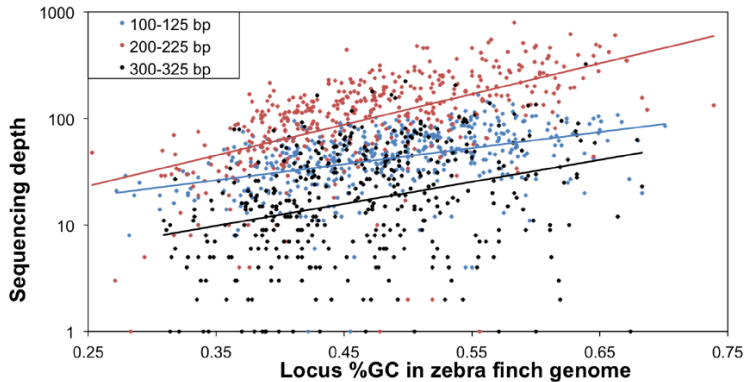
Read depths are influenced by GC content and number of PCR cycles, with (A) or without PCR duplicates (B).



(Davey et al., 2013)

Modifications of PCR enrichment can help (see (Benjamini and Speed, 2012; Puritz et al., 2014a))

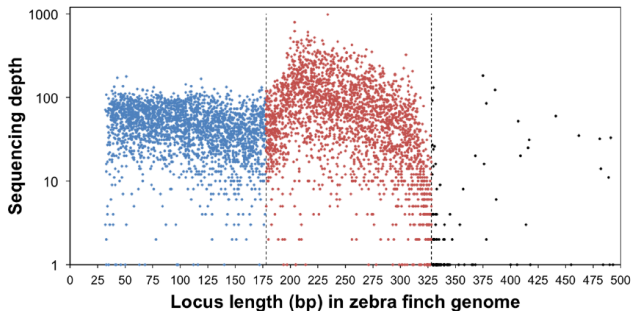
# Sequencing depth bias in favor of loci with high GC content



(DaCosta and Sorenson, 2014)

- Combined with a GC-rich recognition sequence, this can result in an overrepresentation of GC-rich portions of the genome

# Amplification and, thus, depth decreases with fragment length



(DaCosta and Sorenson, 2014)

- Affects ddRAD more than RAD-seq (each locus different fragment lengths) or 2bRAD (all loci same fragment length)
- Bias reduced by precise size selection (Pippin Prep instrument) (DaCosta and Sorenson, 2014).

# STACKS (Puritz et al., 2014b)

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

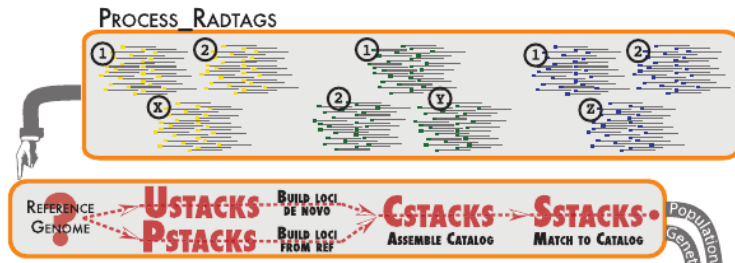
Julian M. Catchen,<sup>\*</sup> Angel Amores,<sup>†</sup> Paul Hohenlohe,<sup>\*</sup> William Cresko,<sup>\*</sup> and John H. Postlethwait<sup>†,1</sup>

<sup>\*</sup>Center for Ecology and Evolutionary Biology and <sup>†</sup>Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403



# STACKS - basic pipeline for RAD-Seq

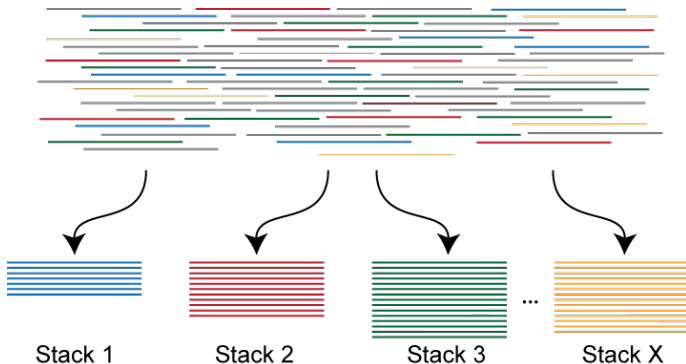
STACKS - software pipeline to build loci from RADseq reads and use them for population genomics and phylogeographic analyses.



(Catchen et al., 2013)

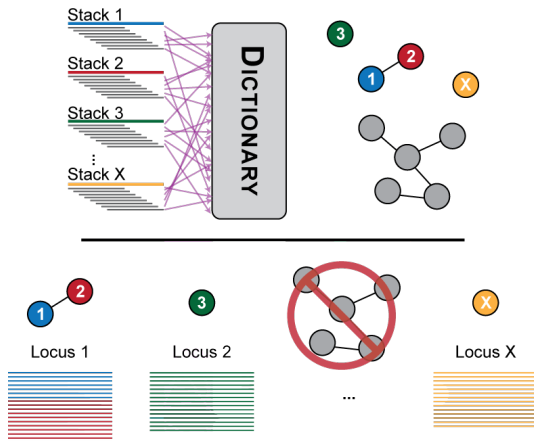
# STACKS - Ustacks *de novo* assembly step 1

- Only exact matches are assembled
- Secondary reads are set aside
- The minimum stack depth parameter controls the number of raw reads required to form an initial stack



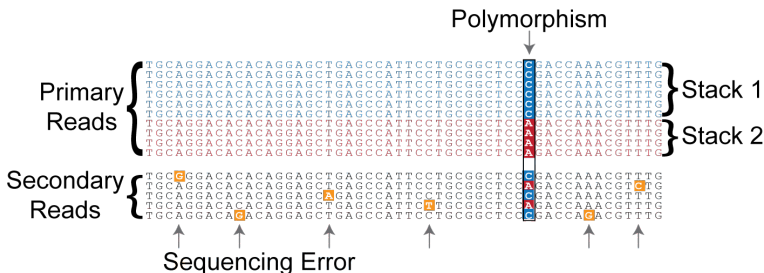
# STACKS - Ustacks *de novo* assembly step 2

- Stacks with few nucleotide differences are merged.
- Repetitive sequences with many alleles are excluded



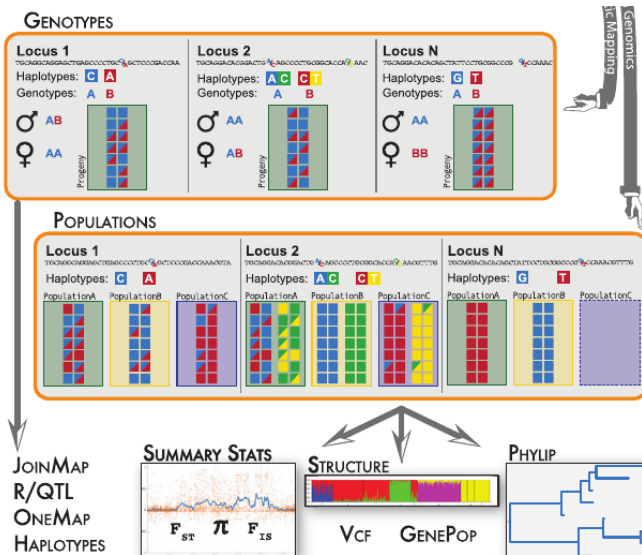
# STACKS - Ustacks *de novo* assembly step 3

- Alignment of secondary reads (those not included in stacks) against stacks.
- Alleles are discriminated from sequencing errors by their frequency.



(Catchen et al., 2013)

# STACKS - populations or genotypes pipeline



# DDocent (Puritz et al., 2014b)



## *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms

Jonathan B. Puritz, Christopher M. Hollenbeck and John R. Gold

Marine Genomics Laboratory, Harte Research Institute, Texas A&M University-Corpus Christi, Corpus Christi, TX, USA

# DDocent (Puritz et al., 2014b)

Uses stand-alone software packages to perform

- quality trimming
- adapter removal
- *de novo* assembly of RAD loci
- read mapping
- SNP and InDel calling
- data filtering.

Identifies more SNPs at a higher coverage than STACKS, due to

- simultaneous use of forward and reverse reads during alignment to reference instead of clustering
- quality trimming instead of removing entire reads

# AftrRAD

## MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2015) 15, 1163–1171

doi: 10.1111/1755-0998.12378

## AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data

MICHAEL G. SOVIC,\*† ANTHONY C. FRIES\* and H. LISLE GIBBS\*†

\*Department of Evolution, Ecology, and Organismal Biology, Aronoff Laboratory, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210, USA, †Ohio Biodiversity Conservation Partnership, Aronoff Laboratory, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210, USA



# PyRAD

Bioinformatics Advance Access published March 20, 2014

BIOINFORMATICS

ORIGINAL PAPER

2014, pages 1–6  
doi:10.1093/bioinformatics/btu121

Phylogenetics

Advance Access publication March 5, 2014

## PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses

Deren A. R. Eaton<sup>1,2</sup>

<sup>1</sup>Committee on Evolutionary Biology, University of Chicago, 1025 E. 57th St. Chicago, IL 60637, USA and <sup>2</sup>Botany Department, Field Museum of Natural History, 1400 S. Lake Shore Dr. Chicago, IL 60605, USA

Associate Editor: David Posada

# Important considerations

- Degraded DNA interferes with cutted DNA in methods with enzyme-unspecific adaptors
- Higher amount of starting DNA can reduce number of PCR cycles and thus minimize PCR duplicates.
- RADseq libraries are low-diversity libraries as they all start with the same cutting site and can cause problems in cluster generation for Illumina sequencing.
  - Solution: Reduce cluster density and spike-in PhiX control

# References I



Andrews, KR, JM Good, MR Miller, G Luikart, and PA Hohenlohe (2016). “Harnessing the power of RADseq for ecological and evolutionary genomics”. In: *Nat Rev Genet* advance on.2, pp. 81–92.



Baird, NA, PD Etter, TS Atwood, MC Currey, AL Shiver, ZA Lewis, et al. “Rapid SNP discovery and genetic mapping using sequenced RAD markers”. In: *PLoS One* 3.10.



Benjamini, Y and T Speed (2012). “Summarizing and correcting the GC content bias in high-throughput sequencing”. In: *Nucleic acids research*.



Catchen, J, Pa Hohenlohe, S Bassham, A Amores, and Wa Cresko (2013). “Stacks: an analysis tool set for population genomics.” In: *Molecular ecology* 22.11, pp. 3124–3140.

# References II



DaCosta, JM and MD Sorenson (2014). “Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol”. In: *PloS one* 9.9, e106713.



Davey, JW, T Cezard, P Fuentes-Utrilla, C Eland, K Gharbi, and ML Blaxter (2013). “Special features of RAD Sequencing data: implications for genotyping”. In: *Molecular Ecology* 22.11, pp. 3151–3164.



Hohenlohe, PA, S Bassham, PD Etter, N Stiffler, EA Johnson, and WA Cresko (2010). “Population genomics of parallel adaptation in threespine stickleback using Sequenced RAD Tags”. In: *Plos Genetics* 6.2.



Lepais, O and JT Weir (2014). “SimRAD: a R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches.” In: *Molecular ecology resources* 33, n/a–n/a.

## References III



Miller, MR, JP Dunham, A Amores, WA Cresko, and EA Johnson (2007). “Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers”. In: *Genome Research* 17.2, pp. 240–248.



Pegadaraju, V, R Nipper, B Hulke, L Qi, and Q Schultz (2013). “De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach”. In: *BMC Genomics* 14.1, p. 556.



Peterson, BK, JN Weber, EH Kay, HS Fisher, and HE Hoekstra (2012). “Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species.” In: *PloS one* 7.5, e37135.



Puritz, JB, MV Matz, RJ Toonen, JN Weber, DI Bolnick, and CE Bird (2014a). “Demystifying the RAD fad”. In: *Molecular ecology* 23.24, pp. 5937–5942.

## References IV



Puritz, J, CM Hollenbeck, and JR Gold (2014b). “dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms”. In: *PeerJ*.



Schweyen, H, A Rozenberg, and F Leese (2014). “Detection and Removal of PCR Duplicates in Population Genomic ddRAD Studies by Addition of a Degenerate Base Region (DBR) in Sequencing Adapters”. In: *The Biological Bulletin* 227.2, pp. 146–160.



Slavov, GT, R Nipper, P Robson, K Farrar, GG Allison, M Bosch, et al. (2014). “Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*”. In: *New Phytologist* 201.4, pp. 1227–1239.

# References V



Tin, MMY, FE Rheindt, E Cros, and AS Mikheyev (2014).

“Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy”. In: *Molecular ecology resources*.



Toonen, RJ, JB Puritz, ZH Forsman, JL Whitney, I Fernandez-Silva, KR Andrews, et al. (2013). “ezRAD: a simplified method for genomic genotyping in non-model organisms”. In: *PeerJ* 1, e203.



Wang, S, E Meyer, JK McKay, and MV Matz (2012). “2b-RAD: a simple and flexible method for genome-wide genotyping”. In: *Nature methods* 9.8, pp. 808–810.