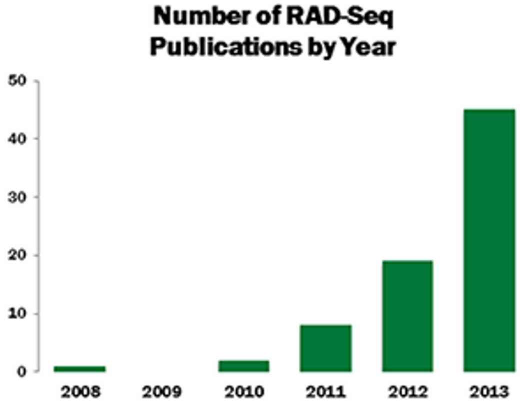
Non-model species and RAD-sequencing

Alexander Jueterbock

2015-05-30

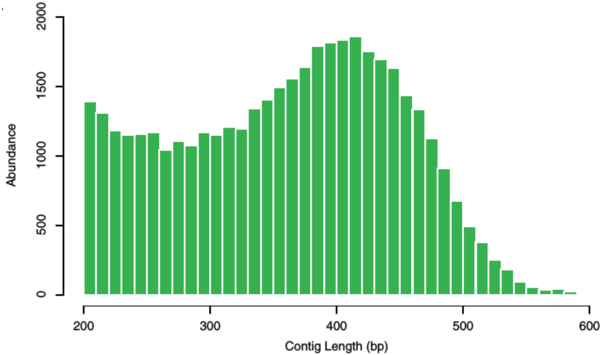
RAD-Seq is a young and successful NGS method



source: http://ngs-expert.com/2013/11/26/rad-seq-publications-in-2013/

Reductive *de novo* genome sequencing and SNP identification

- RAD-Seq of the sunflower genome (Illumina)
    - 44.7M reads (PE:40bpx80bp)
- *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*

source: http://ngs-expert.com/2013/11/26/rad-seq-publications-in-2013/ (Slavov et al., 2014)

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

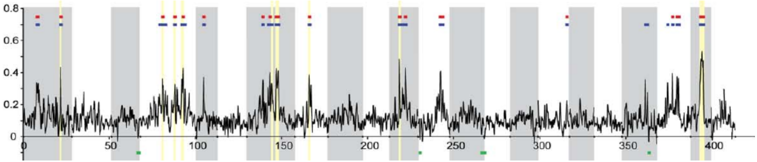


Figure : F_{ST} for SNPs in sliding windows across the genome between oceanic and freshwater populations

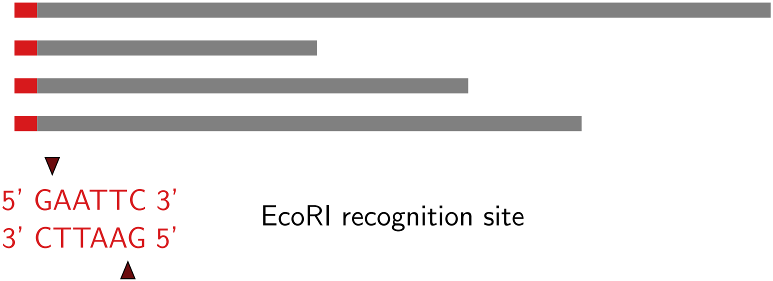
(Hohenlohe et al., 2010)

Purpose of RAD-seq

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

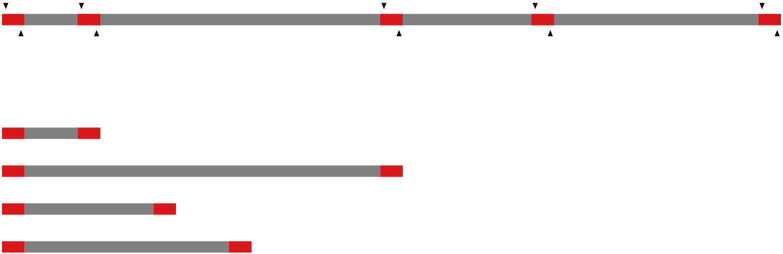
Restriction-site associated DNA (original mbRAD protocol)

- Developed by (Baird et al., 2008; Miller et al., 2007).
- DNA fragments adjacent to restriction enzyme recognition sites



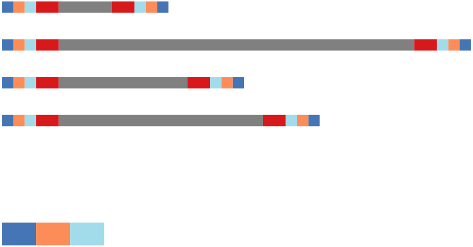
EcoRI recognition site

Step 1: cut DNA



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

Step 2: ligate P1 adapter

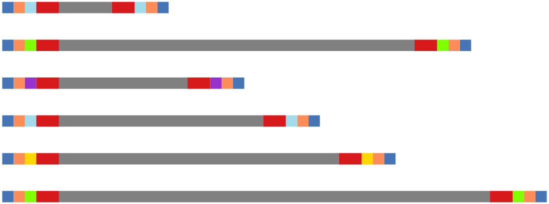


Amplification primer site

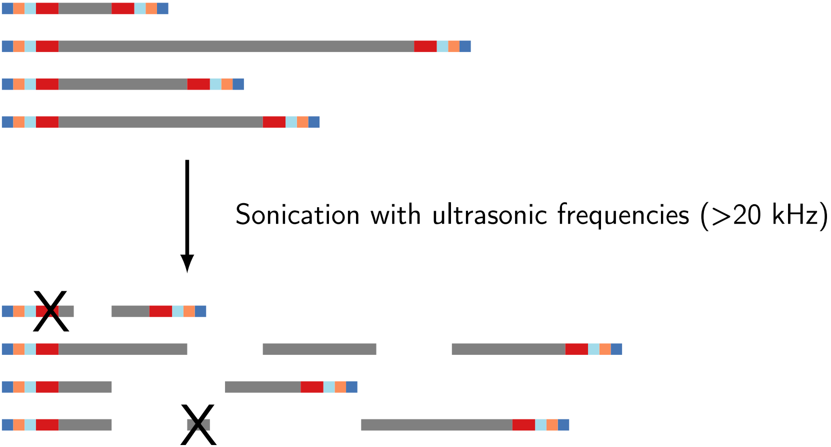
Sequencing primer site (Illumina-specific)

Barcode

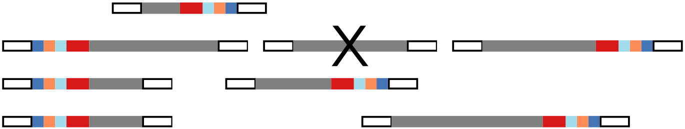
Barcoding allows to pool samples



Step 3: Shearing and size selection



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

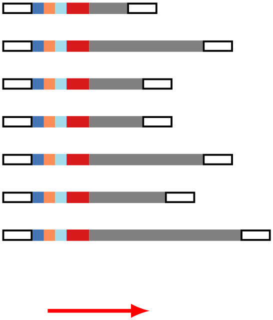


P2 adapter: $\text{AGATCG}^{\text{TCCGA}}$
TCTAGCGTCCT

P2 primer: TCTAGCGTCCT

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

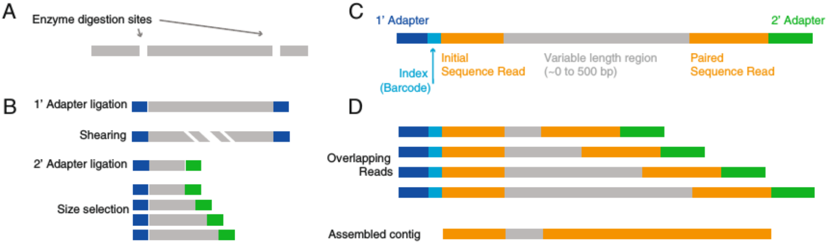
Step 5: Sequence amplified reads on Illumina



Sequence 100 or so bp on Illumina

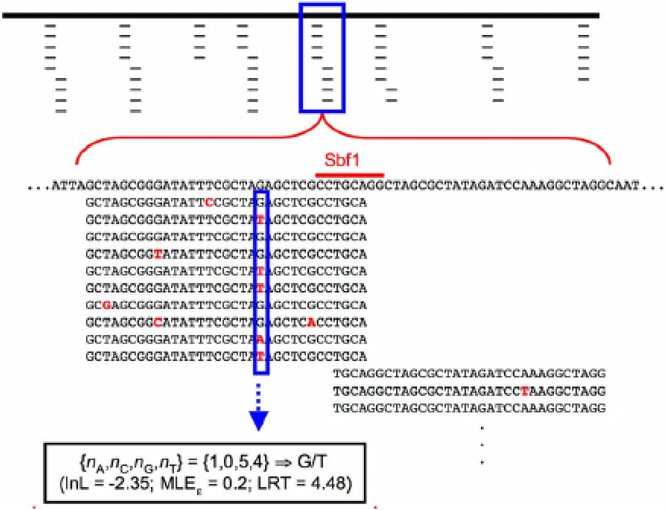
Random sharing of 3'ends helps to detect PCR duplicates

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



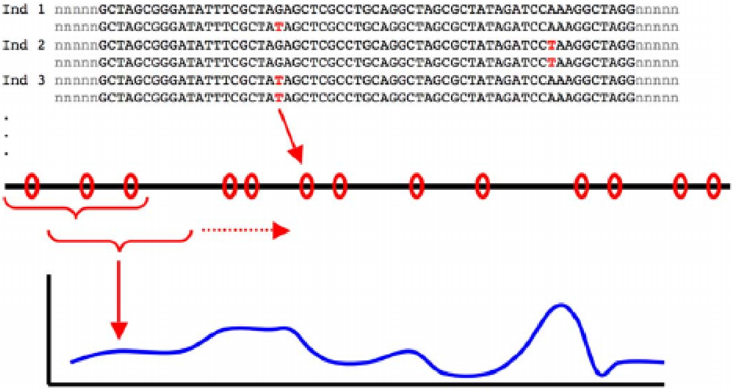
(Pegadaraju et al., 2013)

Calling SNPs from RAD-tags


Sbf1
...ATTAGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGCAAT...
GCTAGCGGGATATTCCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGTATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
GCGAGCGGGATATTTCGCTAGAGCTCGCCTGCA
GCTAGCGGCATATTTCGCTAGAGCTCACCTGCA
GCTAGCGGGATATTTCGCTAAAGCTCGCCTGCA
GCTAGCGGGATATTTCGCTATAGCTCGCCTGCA
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCTAAGGCTAGG
TGCAGGCTAGCGCTATAGATCCAAAGGCTAGG
$\{n_A,n_C,n_G,n_T\} = \{1,0,5,4\} \Rightarrow$ G/T
($\ln L = -2.35$; $\mathrm{MLE}_\varepsilon = 0.2$; LRT = 4.48)

(Hohenlohe et al., 2010)

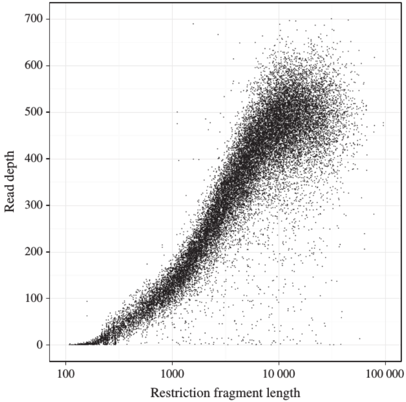
Summary statistics (e.g. population differentiation) along sliding windows


Ind 1 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
Ind 2 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
Ind 3 nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
.
.
.

(Hohenlohe et al., 2010)

Sharing introduces bias

Bias in sequencing depth towards larger fragment sizes

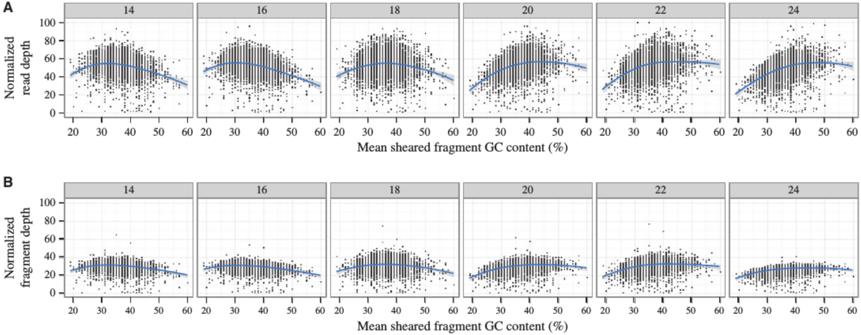


(Davey et al., 2013)

Potential reason: Sonicators shear fragments of different lengths with varying efficiencies

Bias of read depth due to GC content

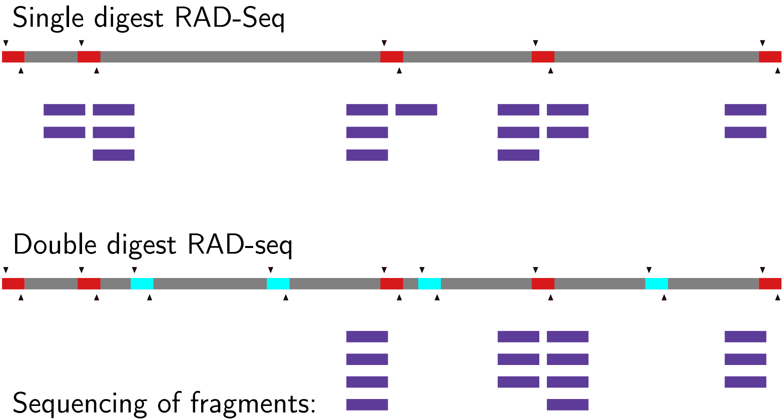
Read depths are influenced by GC content and number of PCR cycles, with (A) or without PCR duplicates (B).



(Davey et al., 2013)

Modifications of PCR enrichment can help (see (Puritz et al., 2014b))

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

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

ddRAD compared to single-digest RAD sequencing

5. PCR duplicates can only be detected with specific adapters (Schweyen et al., 2014; Tin et al., 2014)

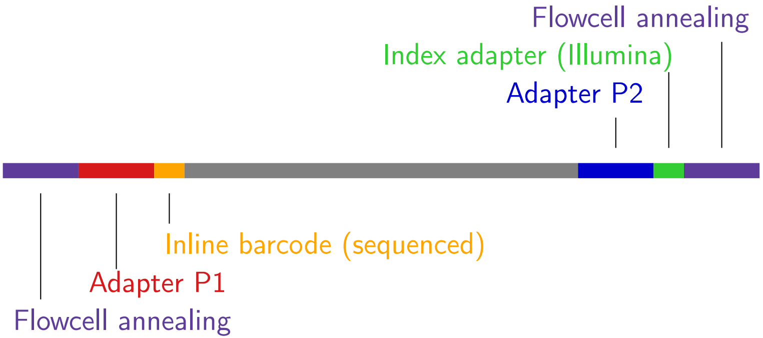
ddRAD compared to single-digest RAD sequencing

6. Precise size selection reduces amplification bias (Pippin Prep instrument - Sage Science) (DaCosta and Sorenson, 2014).

ddRAD compared to single-digest RAD sequencing

7. Null alleles, which can inflate homozygosity (underestimate diversity) by allele-dropout, are more frequent in ddRAD (two recognition sites) (Arnold et al., 2013).

Combinatorial indexing allows for high multiplexing levels in ddRAD-Seq



48 x 12 = 576 (multiplexing level)

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

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

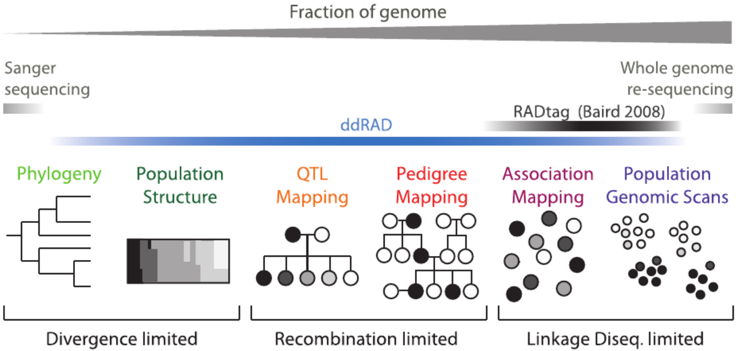
Pooling recommendations

- Critical: equimolar concentrations of individuals expected
- Recommended: >40 individuals/pool
  - Higher numbers
    - + decrease unequal representation of individuals in the pool
    - - make it more more difficult to discriminate minor allele frequencies from sequencing errors

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

Number of markers adjusted by:

- Cutting frequency of restriction enzymes
- Size selection



(Peterson et al., 2012)

How to predict the number of fragments

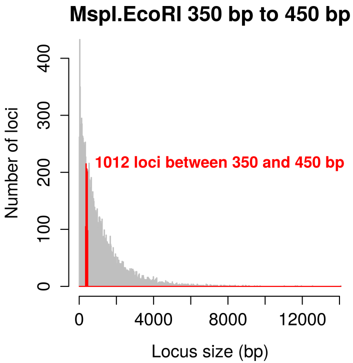
Based on our own study on Guppy

- Targeted coverage: 20x per individual
- Pooling: 60 individuals
- Sequencing output: 24M reads (12M fragments, minimum for Illumina v2 paired-end kits)
- Fragments per individual: 12M/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

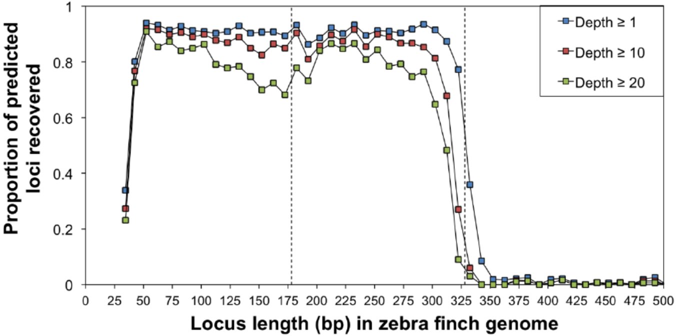
Simulate restriction enzyme digestion with the R package simRAD (Lepais and Weir, 2014)



Based on 10% of the entire genome size

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

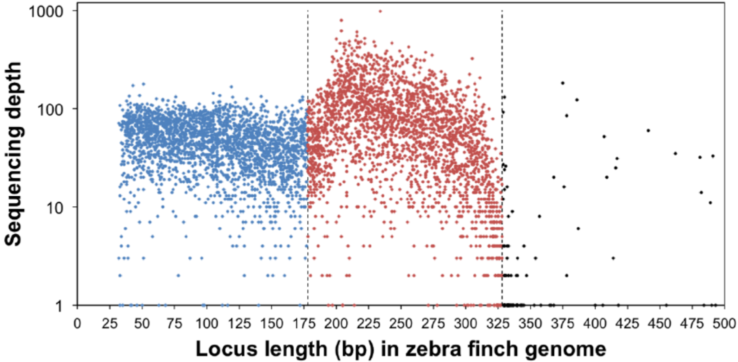
Recovery of *in silico* predicted loci



(DaCosta and Sorenson, 2014)

Targeted: 178-328bp, but short restriction fragments (38–178 bp) were carried through the agarose gel size selection step

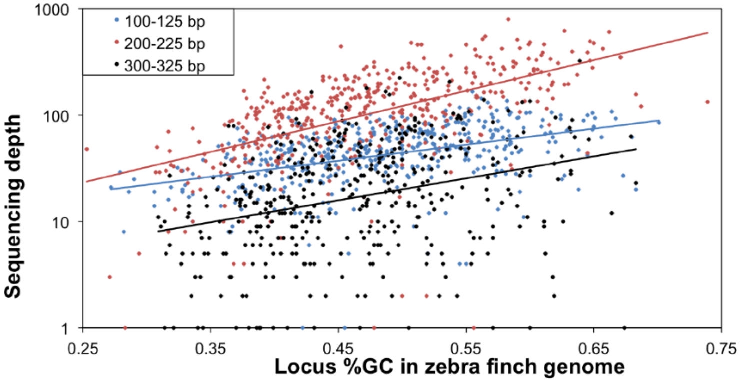
Sequencing depth decreases with fragment lenth



(DaCosta and Sorenson, 2014)

- Negative correlation between depth and fragment length in the 178–200 bp range, not for smaller loci.
- Among-locus variation in sequencing depth was consistent among samples.

Sequencing depth bias in favor of loci with high GC content



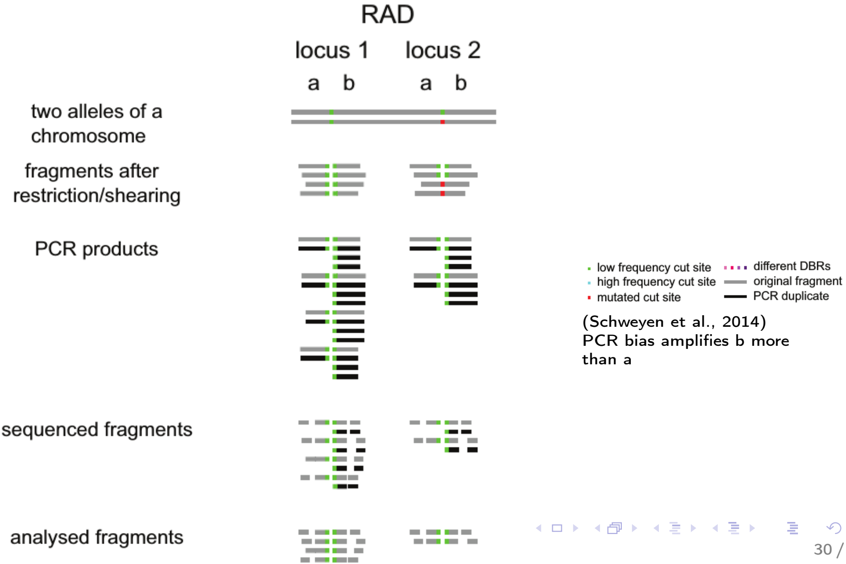
(DaCosta and Sorenson, 2014)

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

PCR duplicates

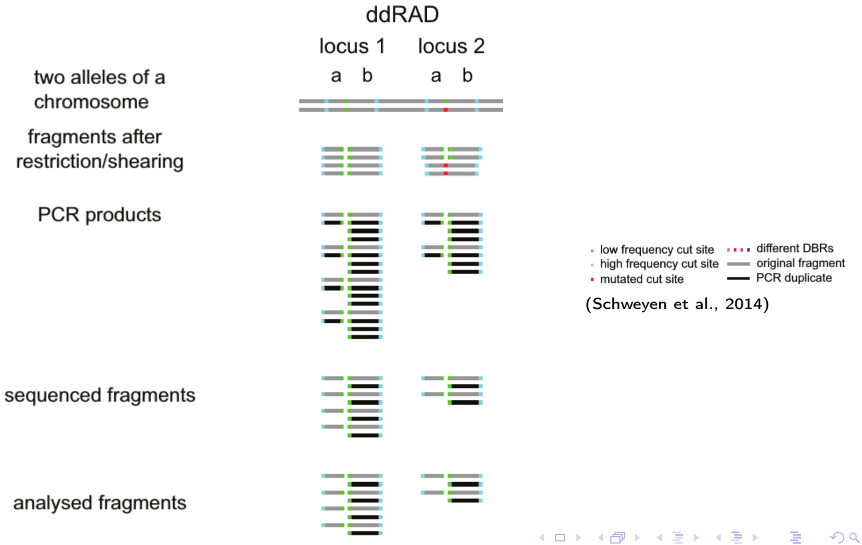
- PCR duplicates are statistically nonindependent and inflate the confidence of genotype calls at a site.
- Can inflate the proportion of homozygous loci (allele dropout) (Schweyen et al., 2014).
- RAD-tags: homologous sequences start at the same location and can not be discriminated from PCR duplicates if they have the same length. All are generally removed
- ddRAD-tags: Paired-end sequences always start and end at the same position
- Detection of duplicate reads only possible with specific adapters of random four bases that are ligated to the first index read of the template molecule before PCR. (Schweyen et al., 2014; Tin et al., 2014).

Detect PCR duplicates in paired-end RAD sequencing



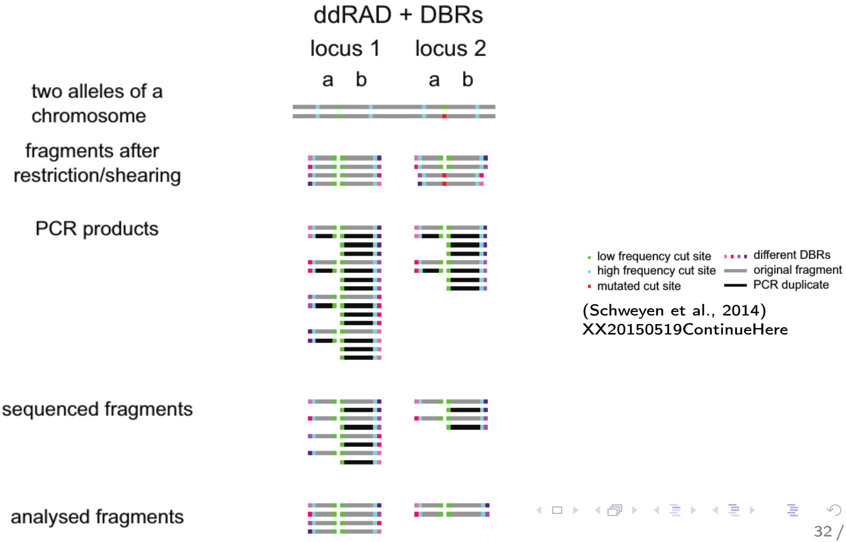
(Schweyen et al., 2014)
PCR bias amplifies b more than a

PCR duplicates in ddRAD - not detectable



(Schweyen et al., 2014)

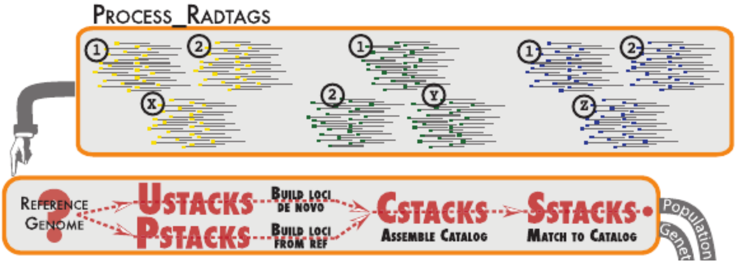
Degenerate base regions detect PCR duplicates in ddRAD



(Schweyen et al., 2014)
XX20150519ContinueHere

STACKS - basic pipeline for mbRAD

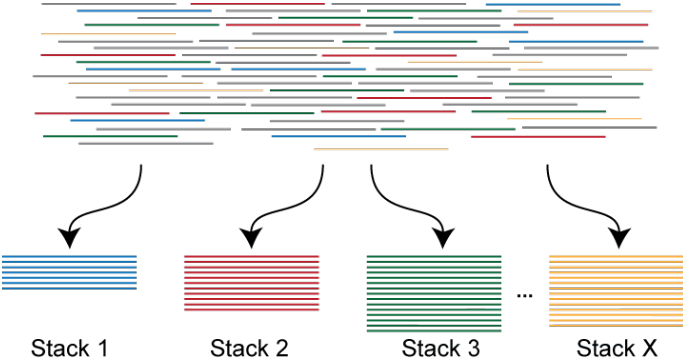
STACKS - software pipleine to build loci from RADseq reads and use them to population genomics and phylogeographic analyses.



(Catchen et al., 2013)

STACKS - ustacks *de novo* assembly step 1

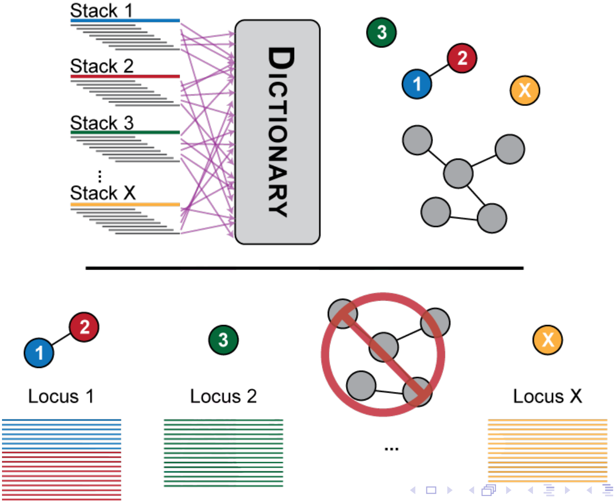
- Only exact matches are assembled
- Secondary reads are set aside
- The minimum stack depth parameter controls the number of raw reads required to form an initial stack



(Catchen et al., 2013)

STACKS - Ustacks *de novo* assembly step 2

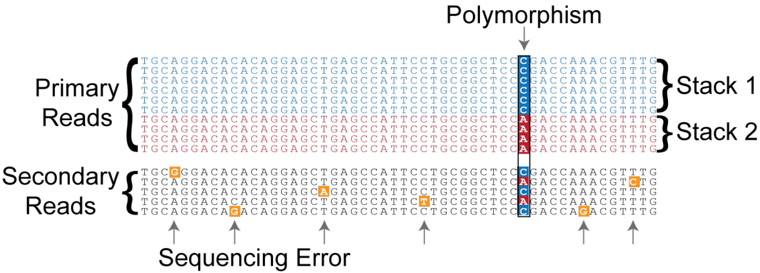
- Stacks with few nucleotide differences are merged.
- Repetitive sequences with many alleles are excluded



(Catchen et al., 2013)

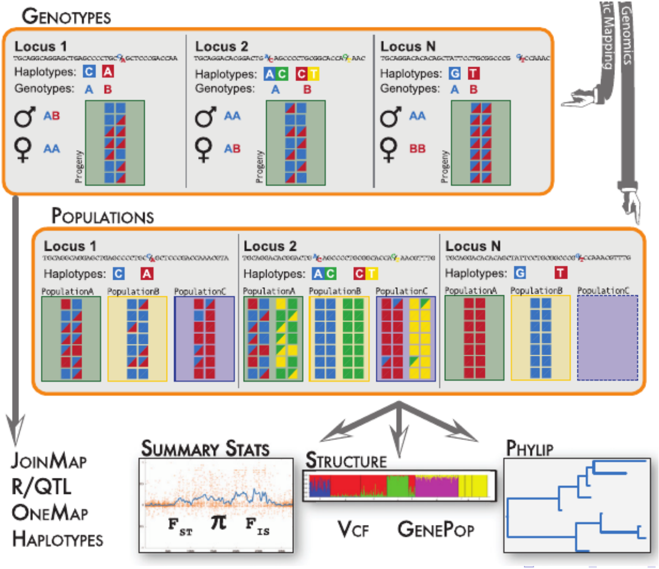
STACKS - Ustacks *de novo* assembly step 3

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


Polymorphism
Primary Reads
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
Stack 1
Stack 2
Secondary Reads
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
Sequencing Error

(Catchen et al., 2013)

STACKS - populations or genotypes pipeline



(Catchen et al., 2013)

DDocent (Puritz et al., 2014a)

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

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

ezRAD (Toonen et al., 2013)

- Uses 2 isoschizomers of restriction enzymes specific to the same recognition sequence (GATC)
- digested DNA is inserted in Illumina TruSeq library preparation kit.
- DNA is digested and single- or dual-indexed, then pooled and size-selected.

Advantages

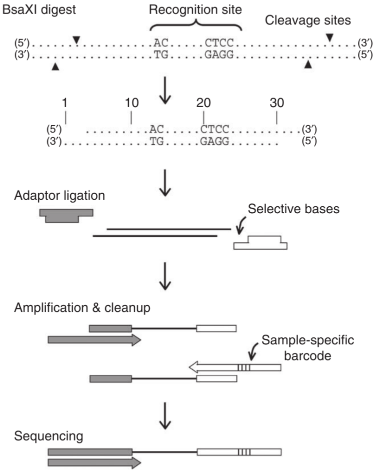
- non-PCR kits can avoid PCR duplication and bypass any PCR bias.

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.


BsaXI digest
Recognition site
Cleavage sites
(5′)......▼..........AC.....CTCC..............▼....(3′)
(3′)...............TG.....GAGG........................(5′)
1
10
20
30
(5′) .........AC.....CTCC.........(3′)
(3′)...........TG.....GAGG......(5′)
Adaptor ligation
Selective bases
Amplification & cleanup
Sample-specific barcode
Sequencing

(Wang et al., 2012)

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.
- Likely not cross-mappable across large genetic distances.

Demultiplexing and trimming

- 'process radtags' from STACKS did not work well on our data
- DDemux (Rasic et al., 2014) used for demultiplexing
- Remove barcodes if this is not done during demultiplexing
- Discard unpaired (orphan) reads
- Cut off the Restriction enzyme overhangs (5bp from read 1 and 3bp from read2)

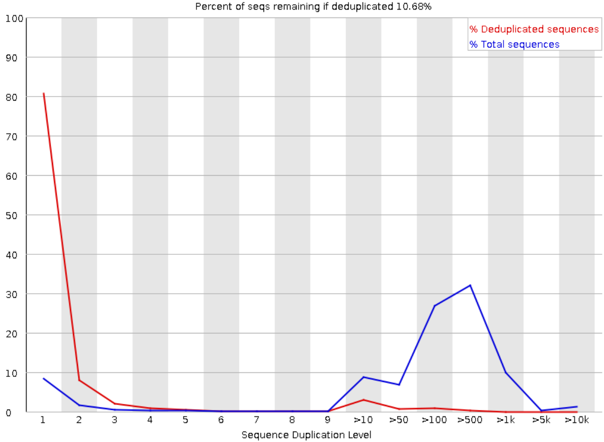
Paired read:

ADAPTER1 AATTAAATTCNNNNCCG ADAPTER2
ADAPTER1 TTAATTTAAGNNNNGGC ADAPTER2

First read: AATTCNNN
Second read: CGGNNN

Barcode
Restriction enzyme overhang (EcoRI and MspI)
Target sequence

Don't remove duplicates from conventional ddRAD data



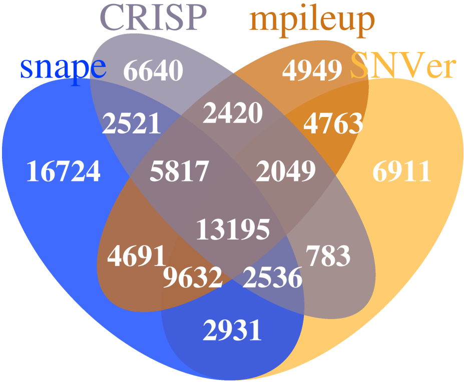
Only preferential amplification of one allele will result in a biased allele frequency estimate

Mapping - Recommendations in Schlötterer et al. (2014)

- Use semi-global alignment (local alignment with soft-clipping of terminal bases can lead to biased allele frequency estimates)
- Allow gaps to avoid false positives
- Realign around indels (misalignment in these regions can lead to false positives)
- Filtering
  - remove broken pairs (increases mapping precision)
  - remove reads with a mapping quality <20
- Disregard regions of too high coverage (potential copy number variations)

SNP calling

Use a consensus approach to call SNPs that were independently identified by different SNP callers



References I

- Arnold, B, R Corbett-Detig, D Hartl, and K Bomblies (2013). "RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling". In: *Molecular ecology* 22.11, pp. 3179–3190.
- Baird, NA, PD Etter, TS Atwood, MC Currey, AL Shiver, ZA Lewis, et al. (2008). "Rapid SNP discovery and genetic mapping using sequenced RAD markers". In: *PLoS One* 3.10.
- 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. 3124–40.
- 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.

References II

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

References IV

- Puritz, JB, MV Matz, RJ Toonen, JN Weber, DI Bolnick, and CE Bird (2014b). "Demystifying the RAD fad". In: *Molecular ecology* 23.24, pp. 5937–5942.
- Rasic, G, I Filipovi, AR Weeks, AA Hoffmann, et al. (2014). "Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*". In: *BMC genomics* 15.1, p. 275.
- Schlötterer, C, R Tobler, R Kofler, and V Nolte (2014). "Sequencing pools of individuals [mdash] mining genome-wide polymorphism data without big funding". In: *Nature Reviews Genetics*.
- 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.

References V

- 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.
- Tin, M, F Rheindt, E Cros, and A 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.

References VI

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.