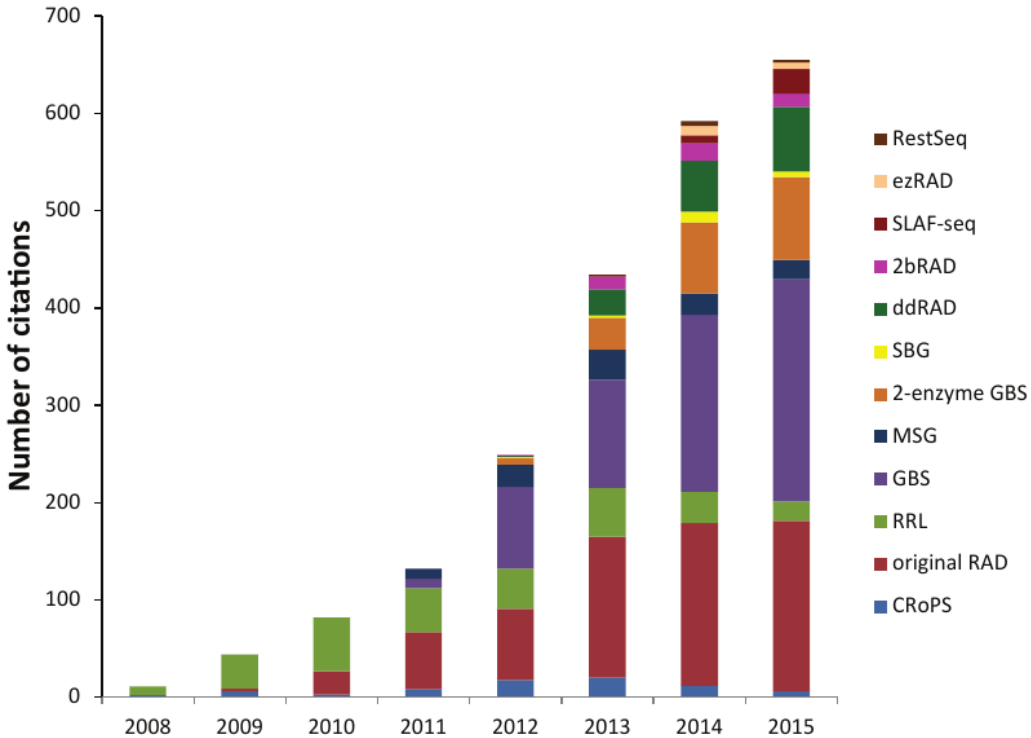
Non-model species and RAD-sequencing

Alexander Jueterbock

May 2017

RAD-Seq - young and successful NGS methods



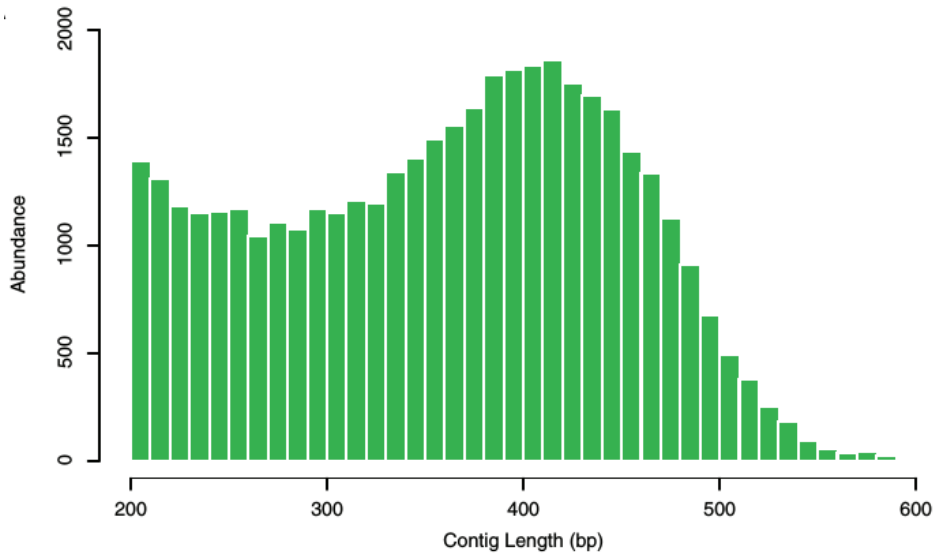
(Andrews2016)

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:40bpx80bp)
- *De novo* assembly of ca. 15.2 Mb in >42,000 contigs
- Identified >94,000 putative SNPs across six lines



(Pegadaraju2013)

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/ (**Slavov2014**)

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

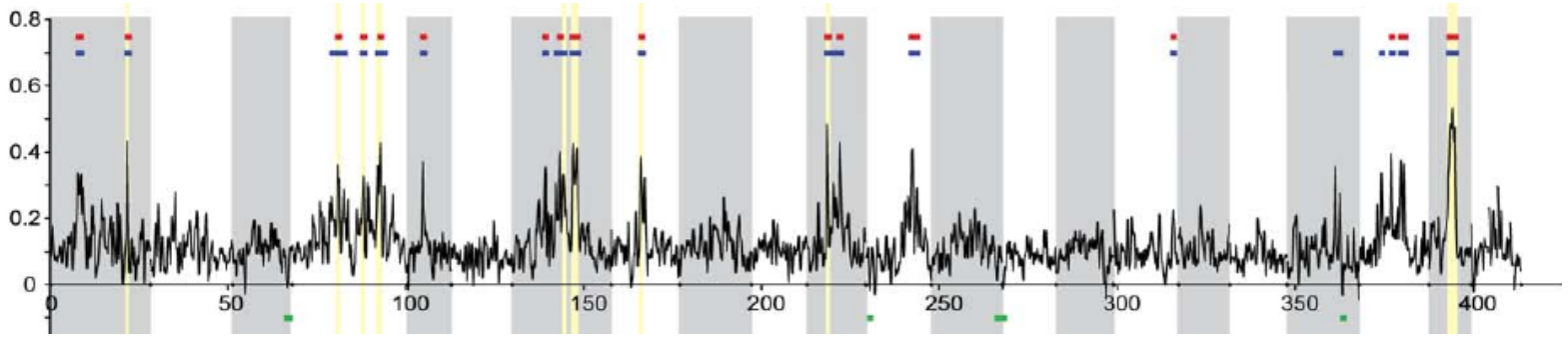
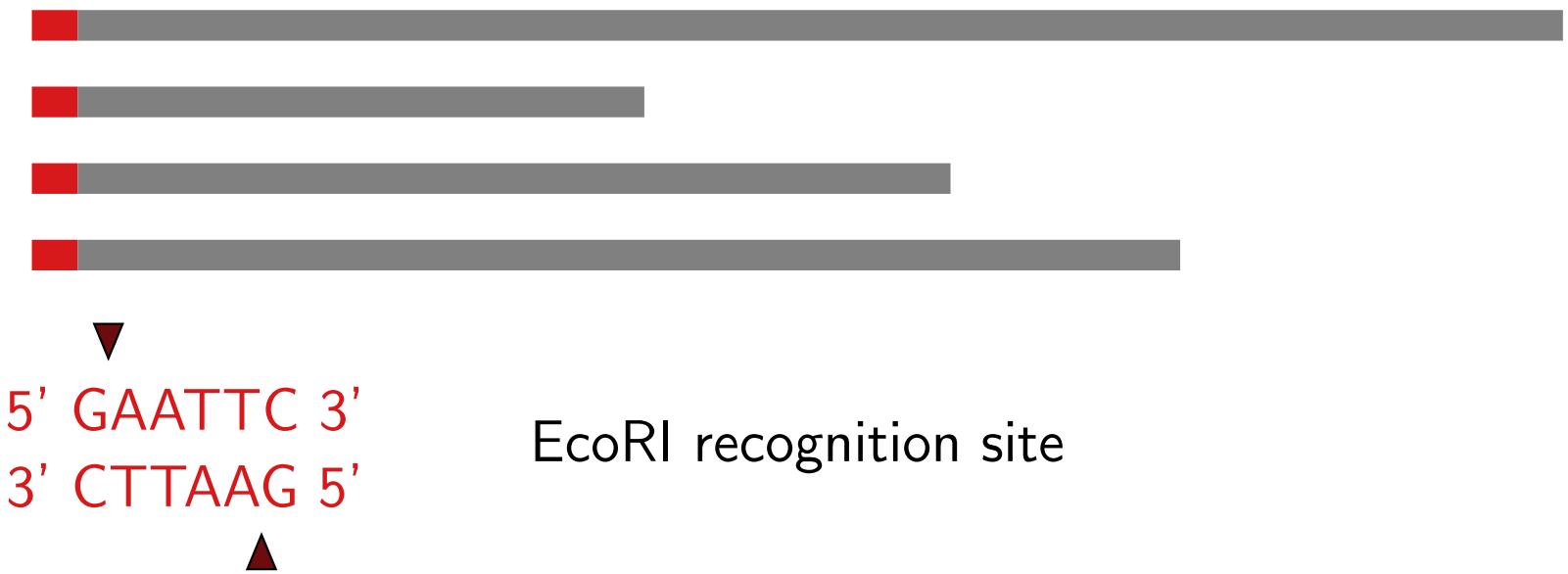


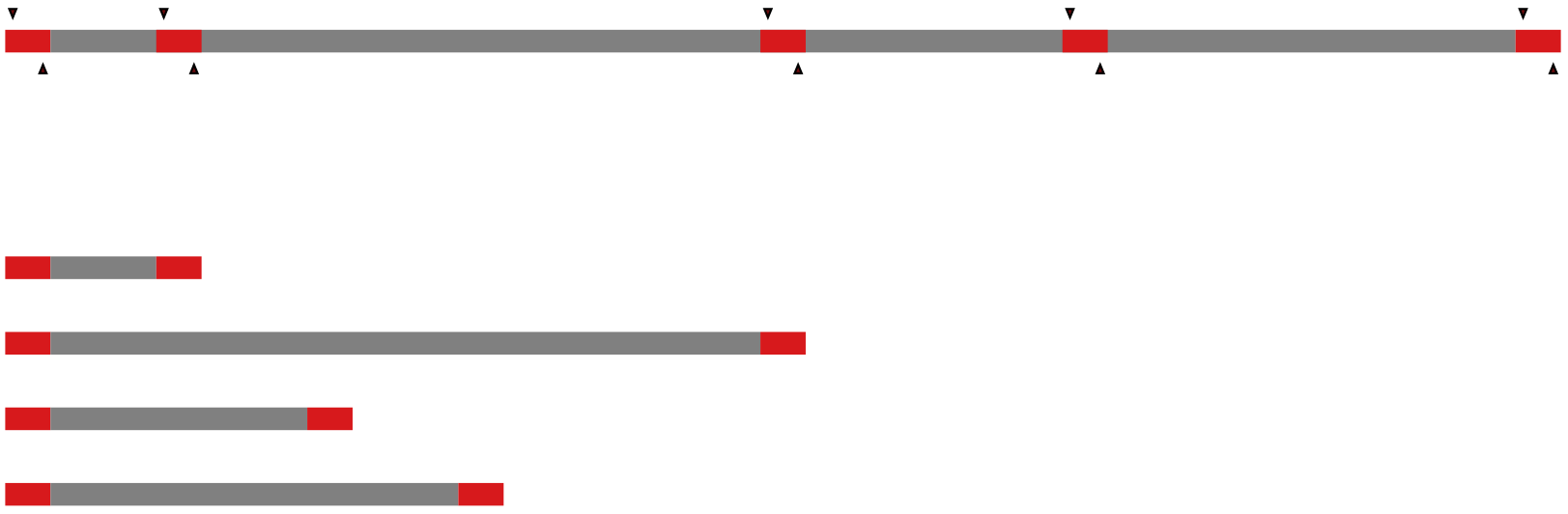
Figure: F_{ST} for SNPs in sliding windows across the genome between oceanic and freshwater populations

Original RAD-Seq protocol

- Developed by (**Miller2007**; **Baird2008**).
- DNA fragments adjacent to restriction enzyme recognition sites

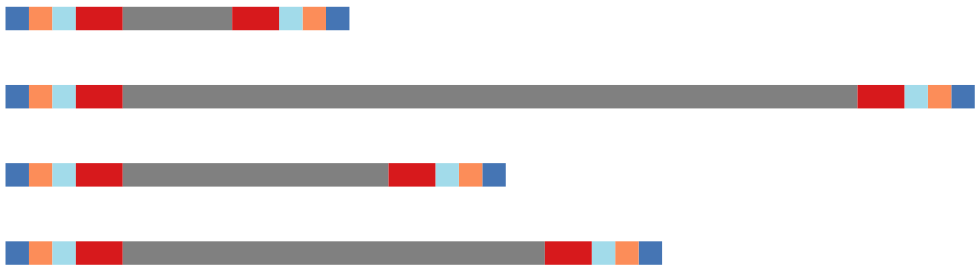


Step 1: cut DNA



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

Step 2: ligate P1 adapter

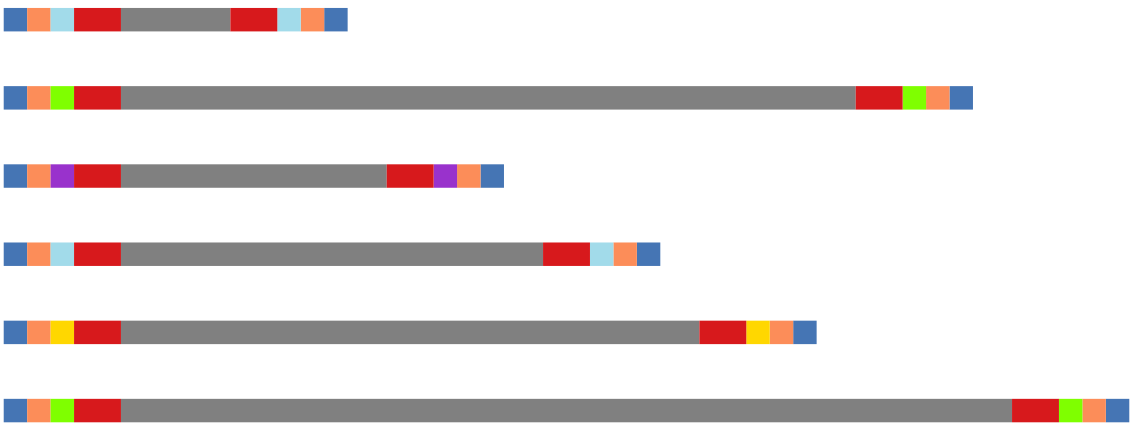


Amplification primer site

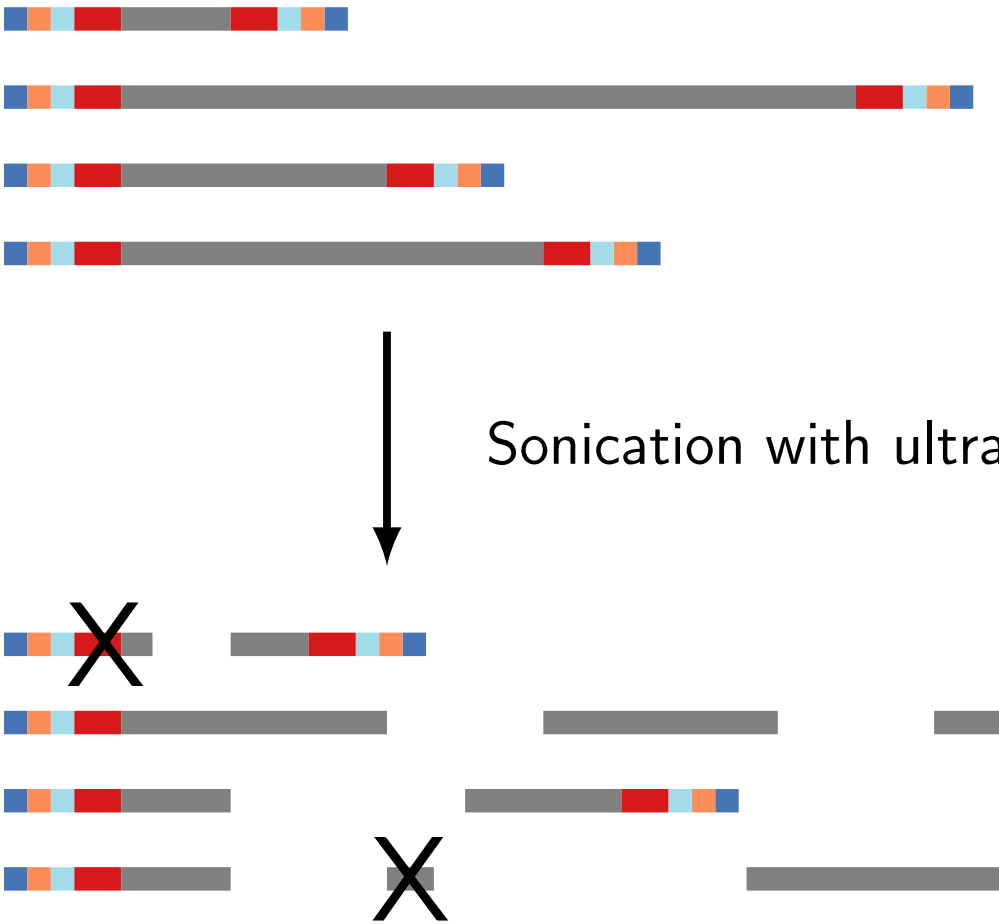
Sequencing primer site (Illumina-specific)

Barcode

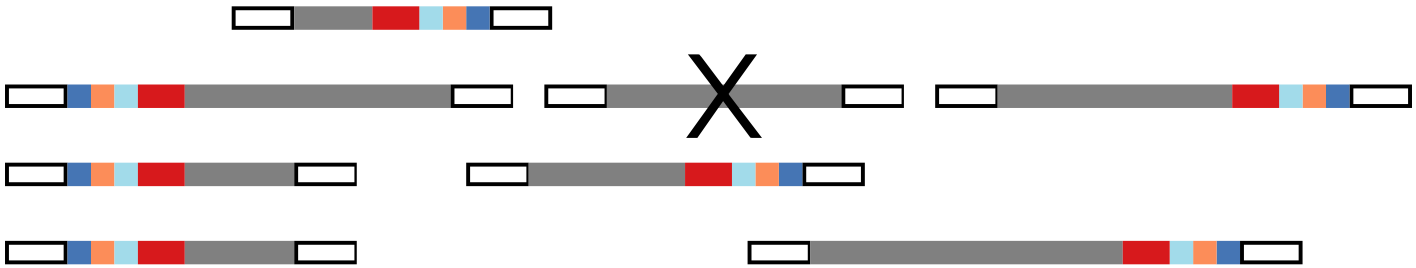
Barcoding allows to pool samples



Step 3: Shearing and size selection



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

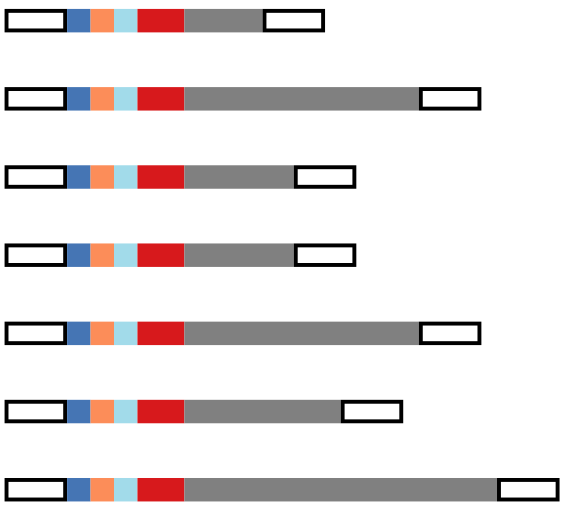


P2 adapter: AGATCGGTCCGA
TCTAGCGTCCT

P2 primer: TCTAGCGTCCT

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

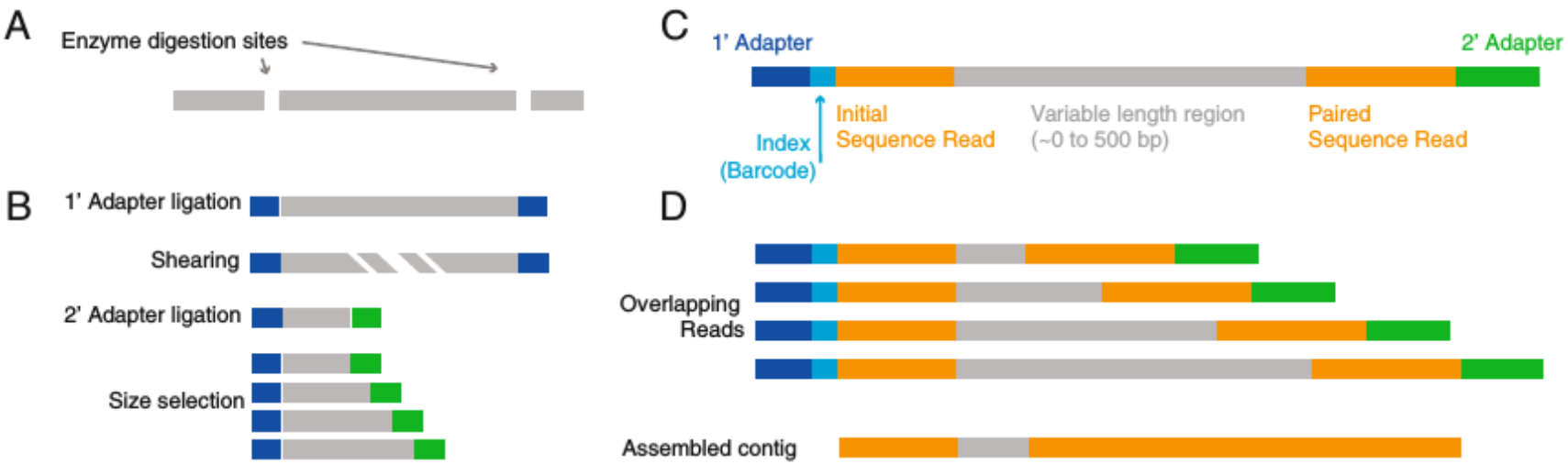
Step 5: Sequence amplified reads on Illumina



Sequence 100 or so bp on Illumina

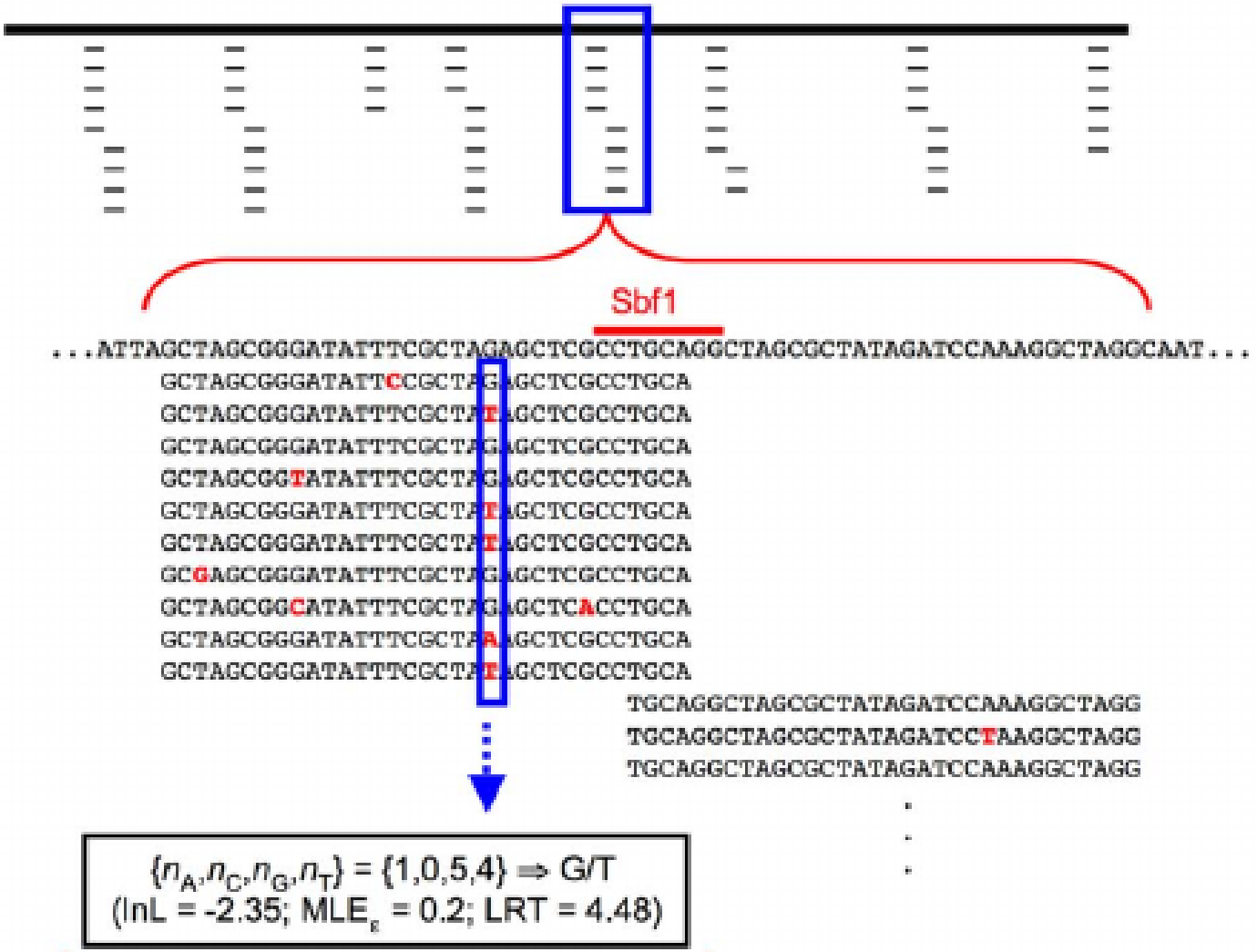
Random shearing of 3'ends helps to detect PCR duplicates

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



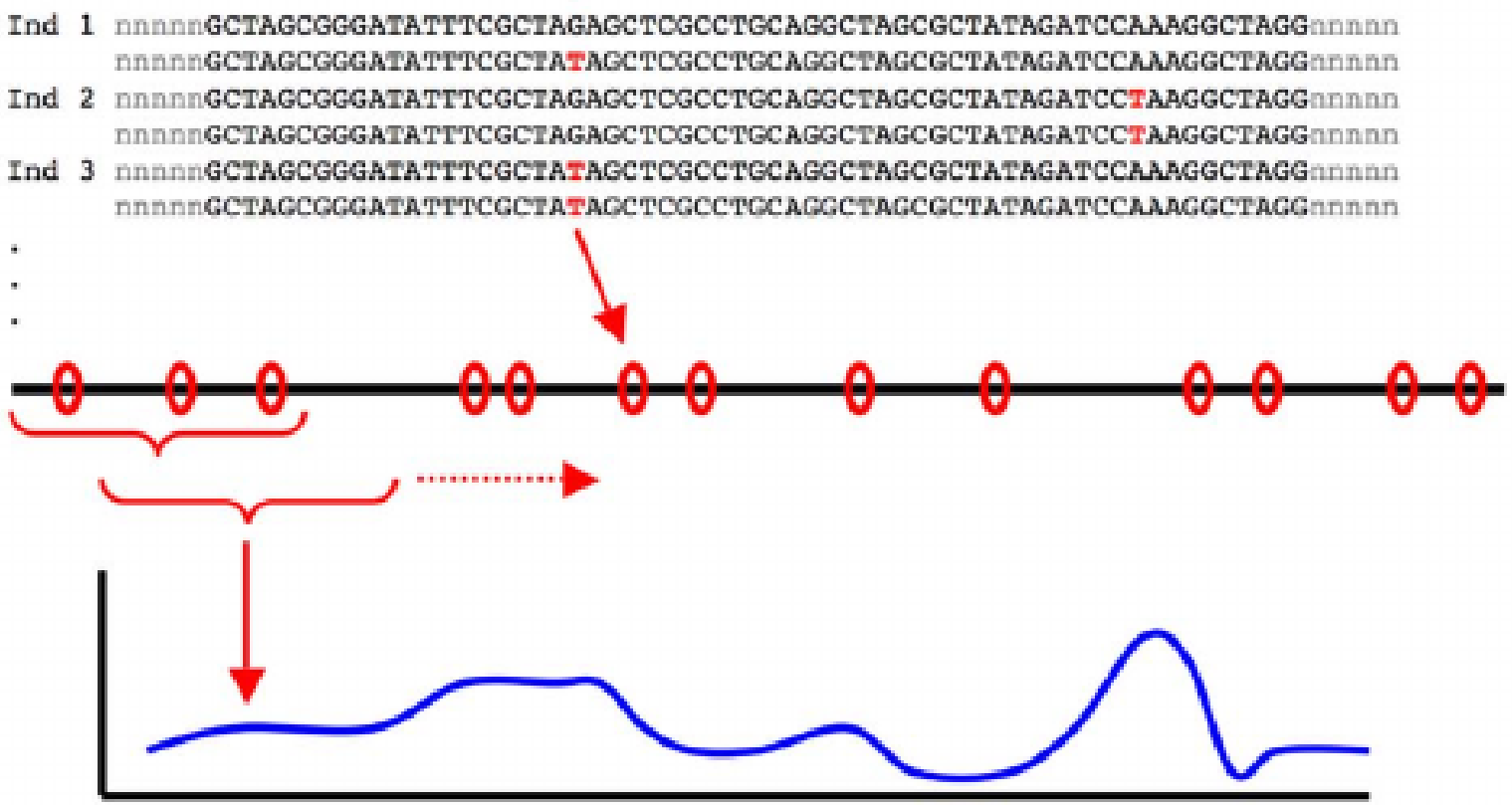
(Pegadaraju2013)

Calling SNPs from RAD-tags



(Hohenlohe2010)

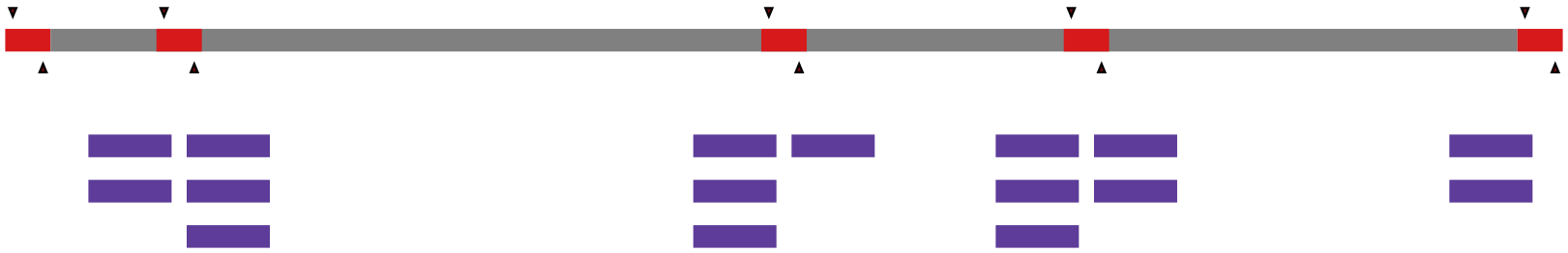
Summary statistics (e.g. population differentiation) along sliding windows



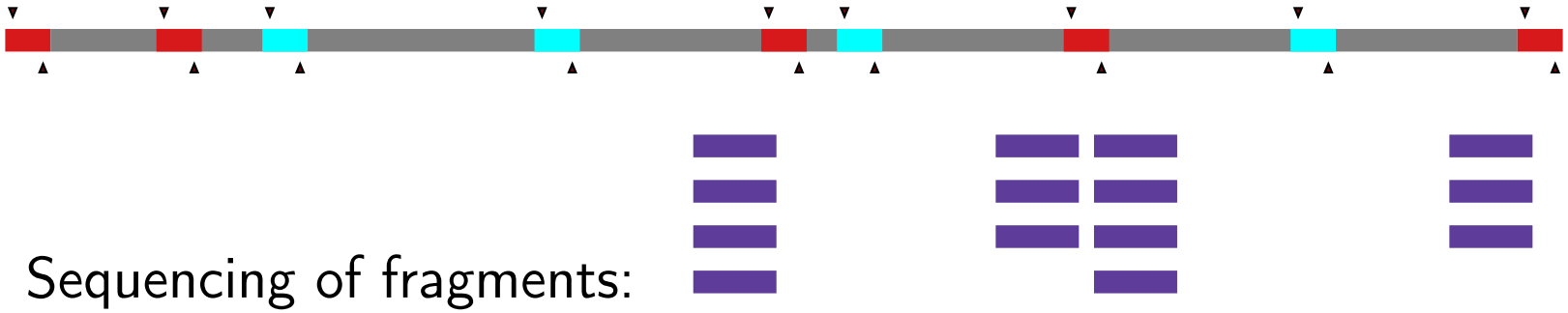
(Hohenlohe2010)

Double-digest RAD-seq (Peterson2012)

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.

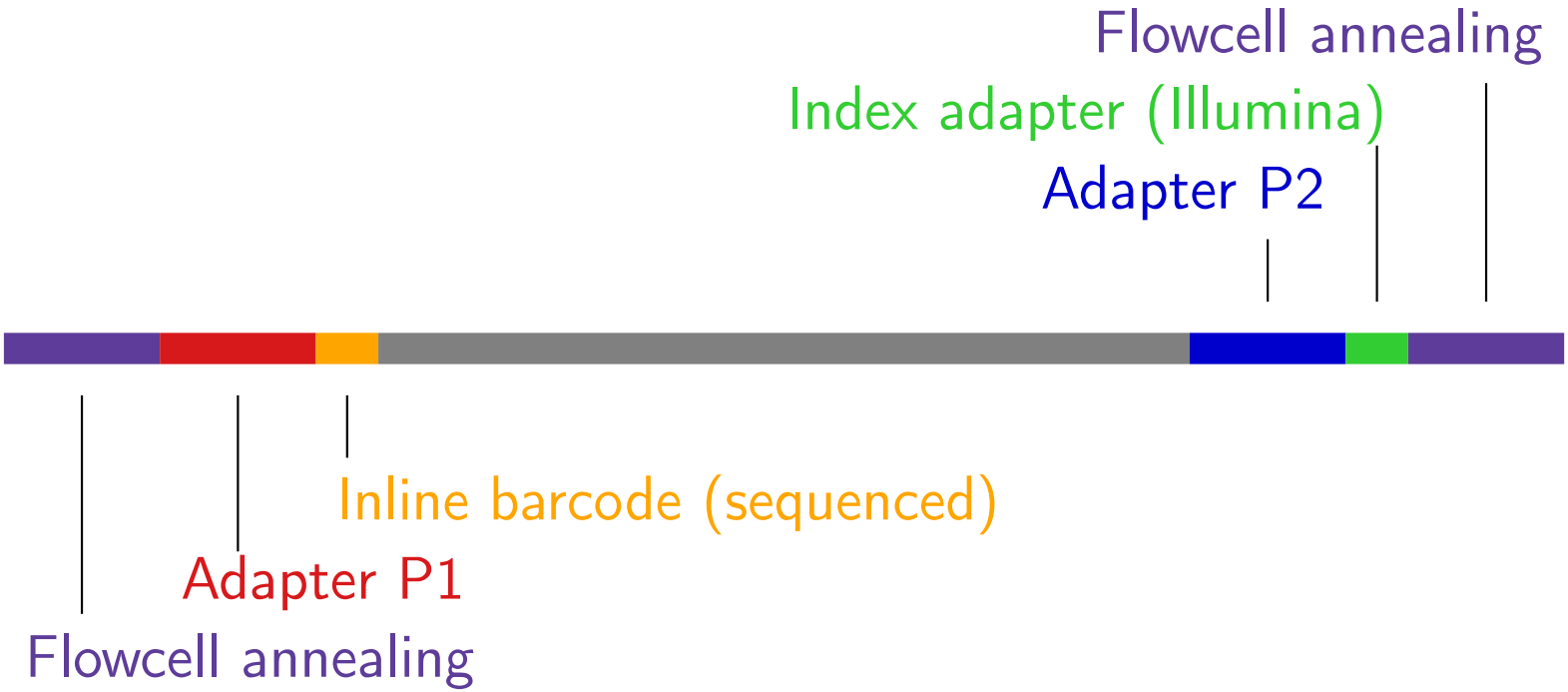
ddRAD compared to single-digest RAD sequencing

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).

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

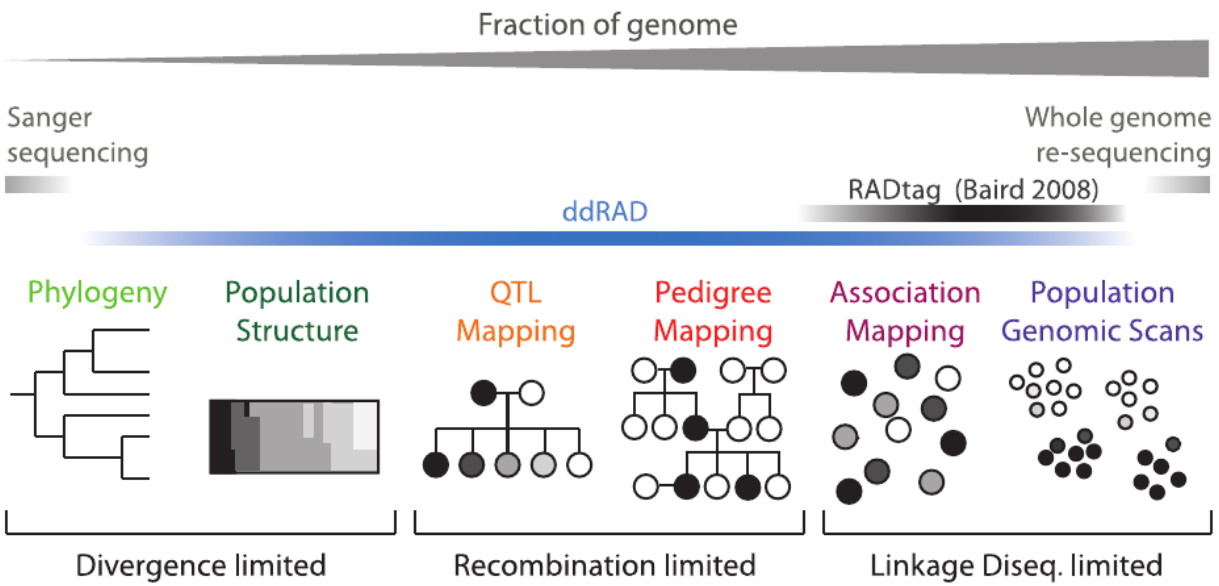
Pooling recommendations

- 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

Number of markers adjusted by:

- Cutting frequency of restriction enzymes
- Size selection



(Peterson2012)

How to predict the number of fragments

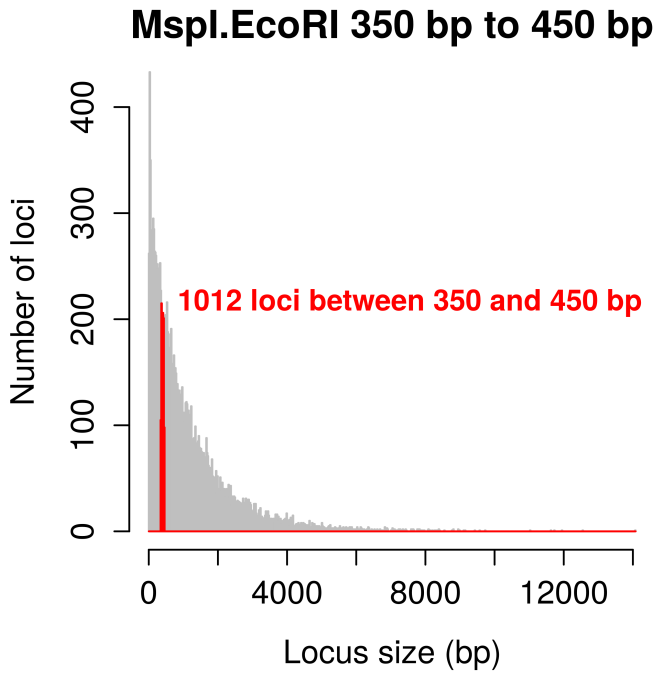
Based on our own study on Guppy

- 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

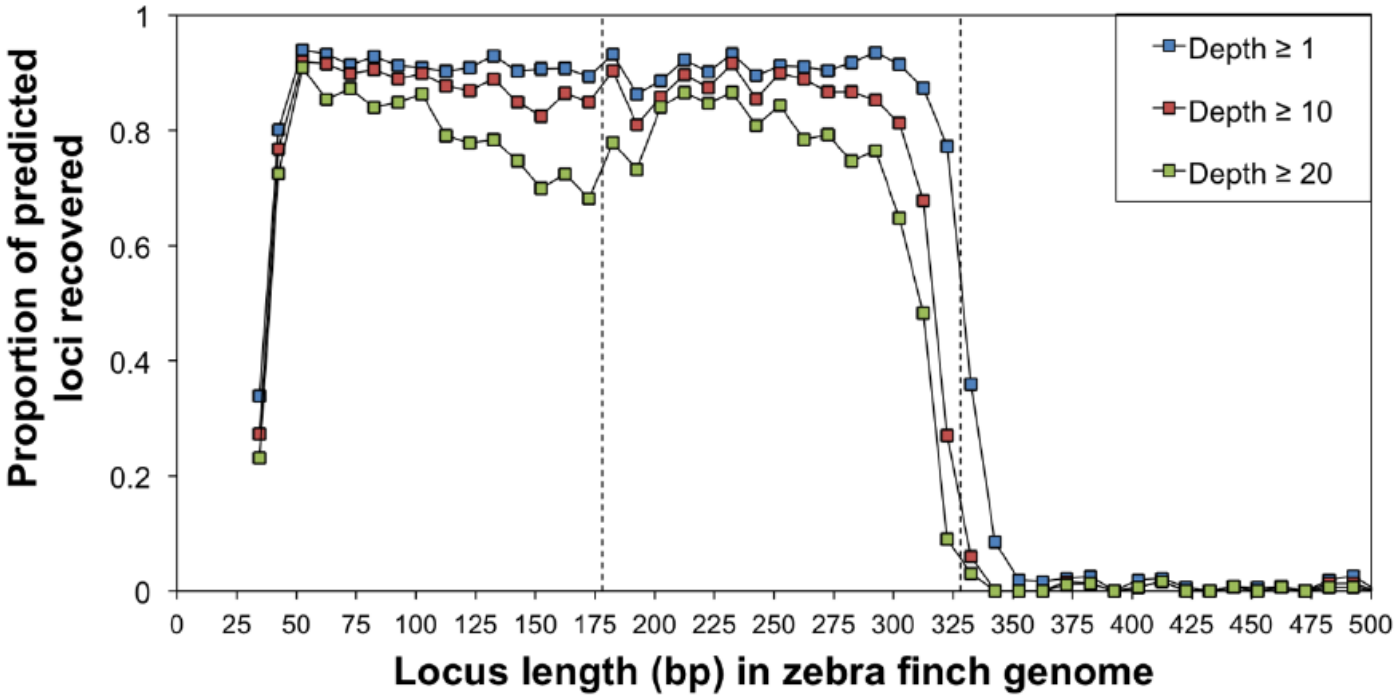
Simulate restriction enzyme digestion with the R package simRAD (**Lepais2014**)



Based on 10% of the entire genome size

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

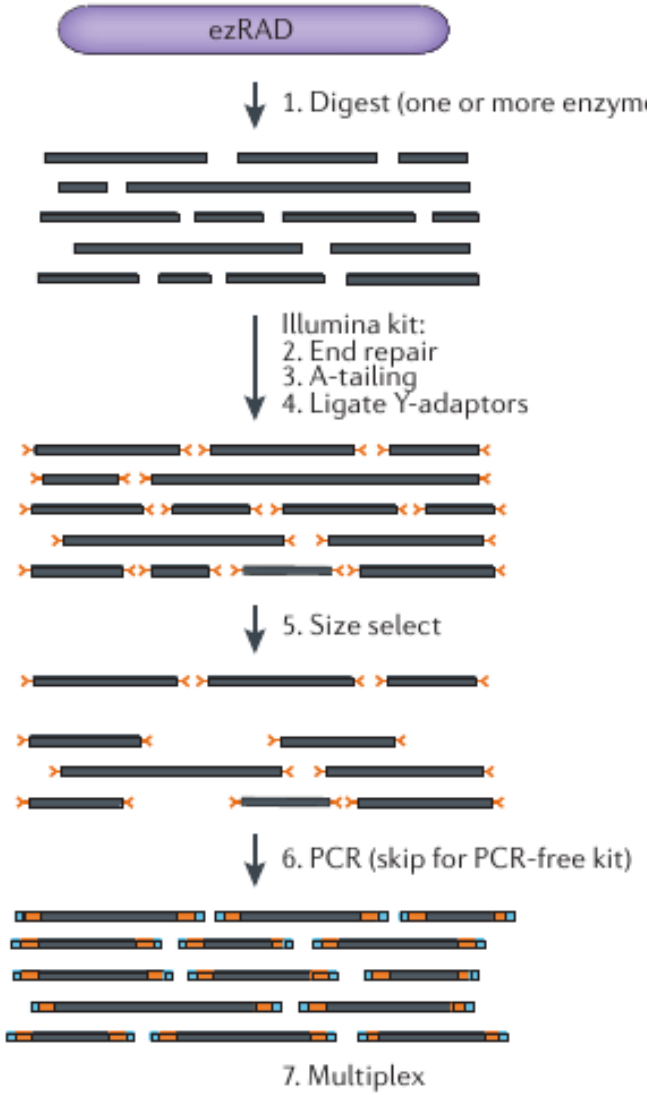
Recovery of *in silico* predicted loci



(Dacosta2014)

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

ezRAD (Toonen2013)



(Andrews2016)

ezRAD (Toonen2013)

Advantages

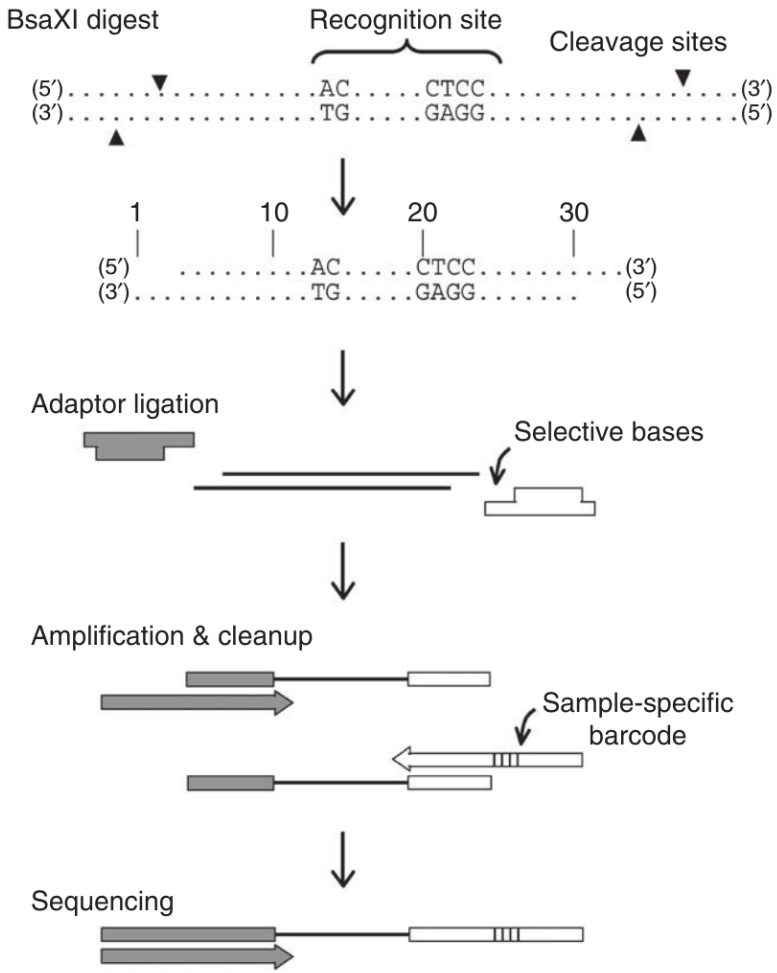
- non-PCR kits can avoid PCR duplication and bypass any PCR bias.

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 (Wang2012)

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



2bRAD (Wang2012)

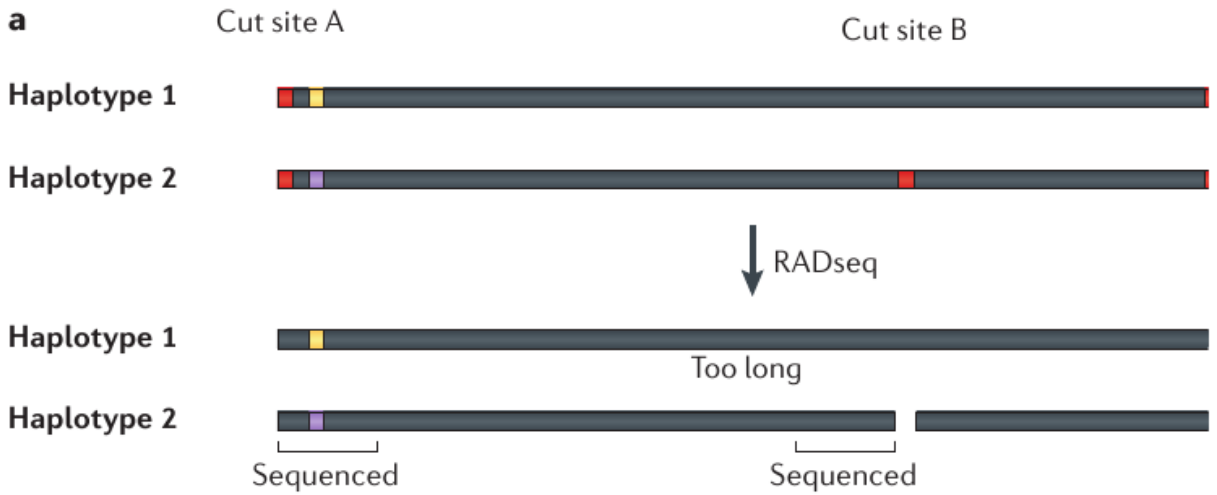
Advantages

- 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.

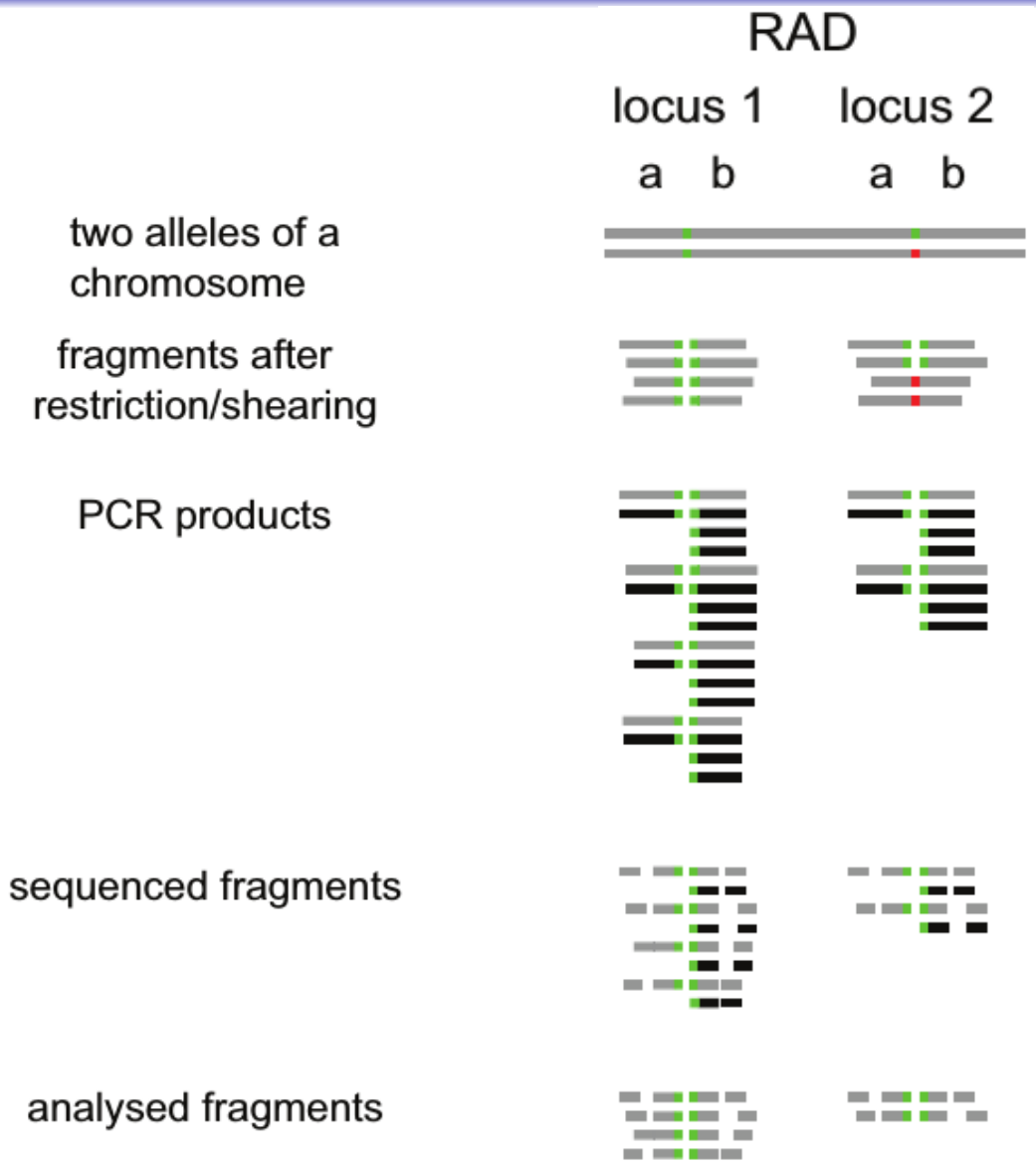
Null alleles



Null allelles due to allele dropbout
(**Andrews2016**)

- Inflate homozygosity
- Underestimate genetic differentiation
- More frequent in longer cutting enzyme recognition sites
- More frequent in ddRAD than in original RADseq
- Identifiable by high variance in coverage among samples

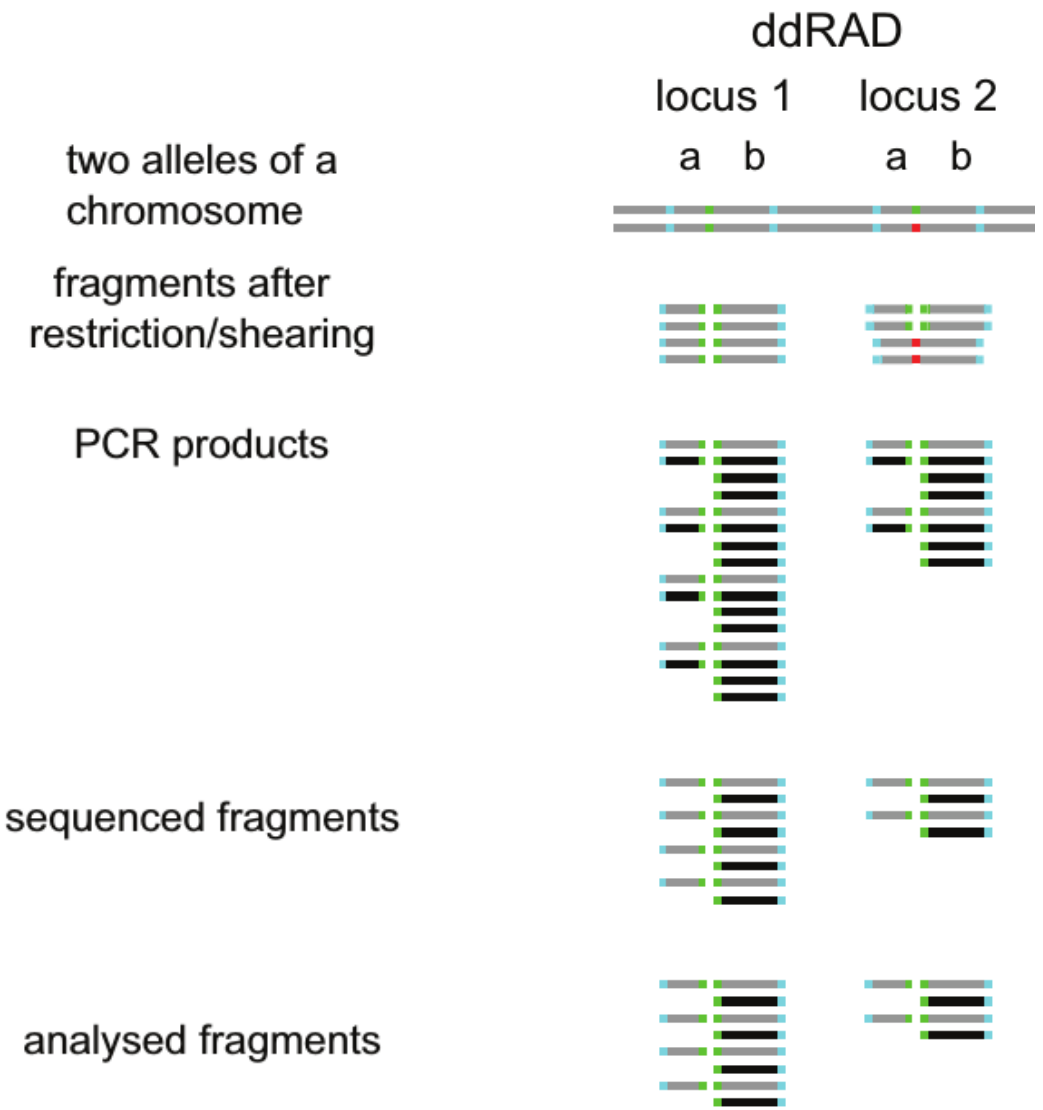
Detect PCR duplicates by length in paired-end RAD sequencing



low frequency cut site
high frequency cut site
mutated cut site
different DBRs
original fragment
PCR duplicate

(Schweyen2014)
PCR bias amplifies b more than a
PCR duplicates can occur in 20 to 60% of reads

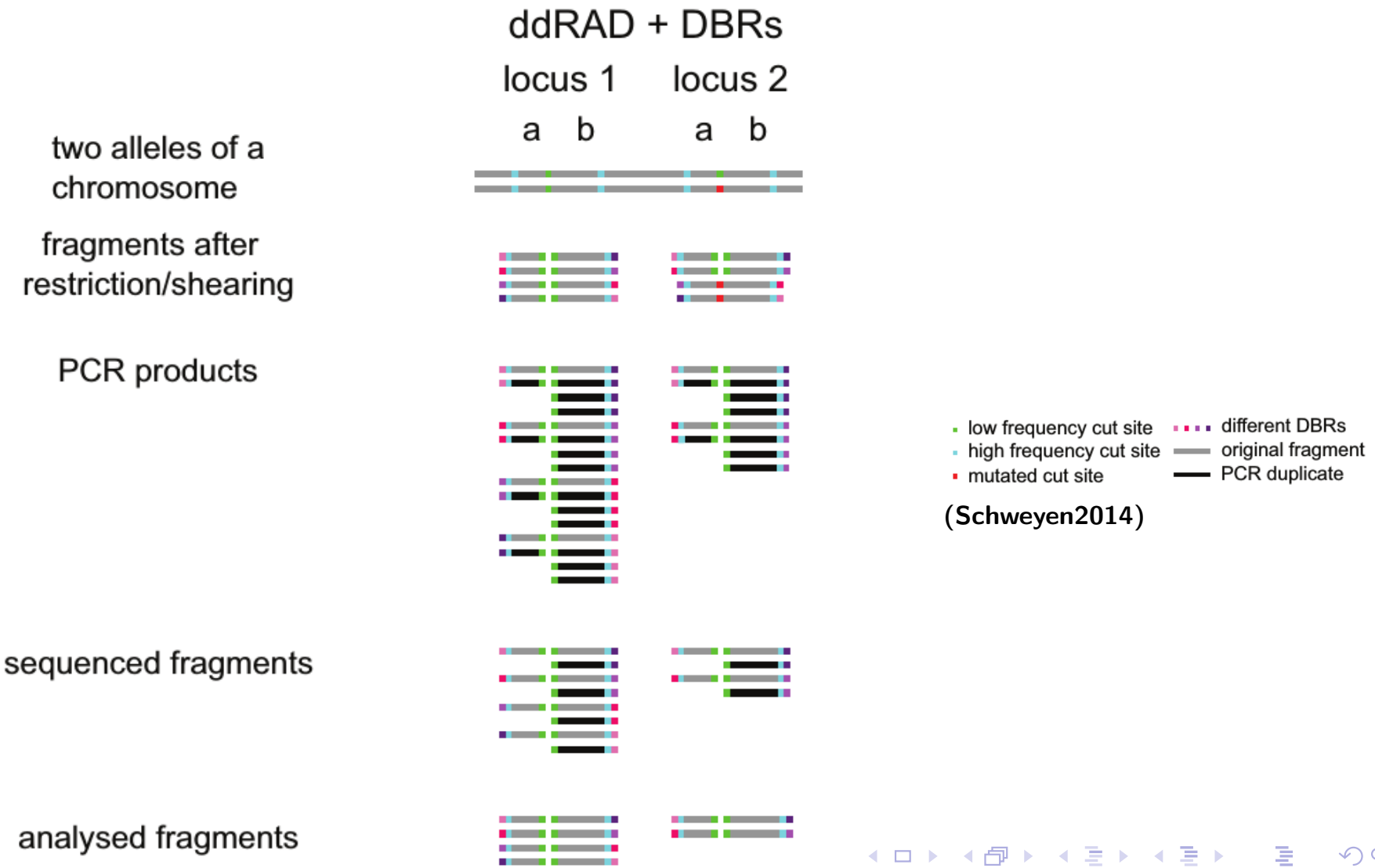
PCR duplicates in ddRAD and ezRAD - not detectable



low frequency cut site
high frequency cut site
mutated cut site
different DBRs
original fragment
PCR duplicate

(Schweyen2014)
Sequences of one locus always start and end at the same position
Locus 2 with mutated cut site can have equal coverage as locus 1

Degenerate base regions detect PCR duplicates in ddRAD ((Tin2014; Schweyen2014))



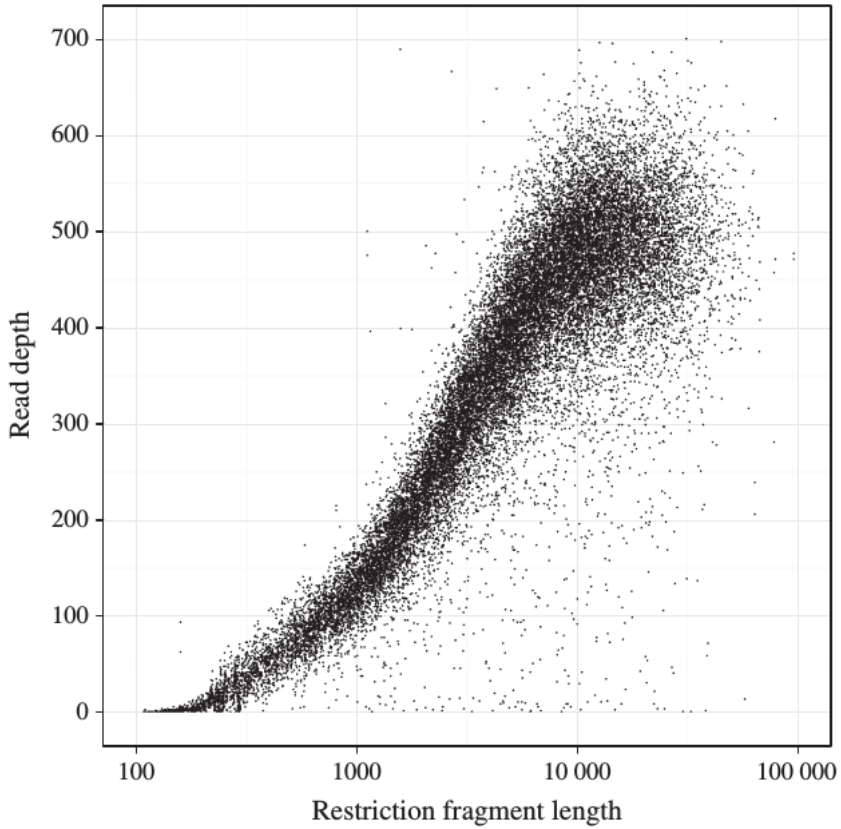
(Schweyen2014)

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

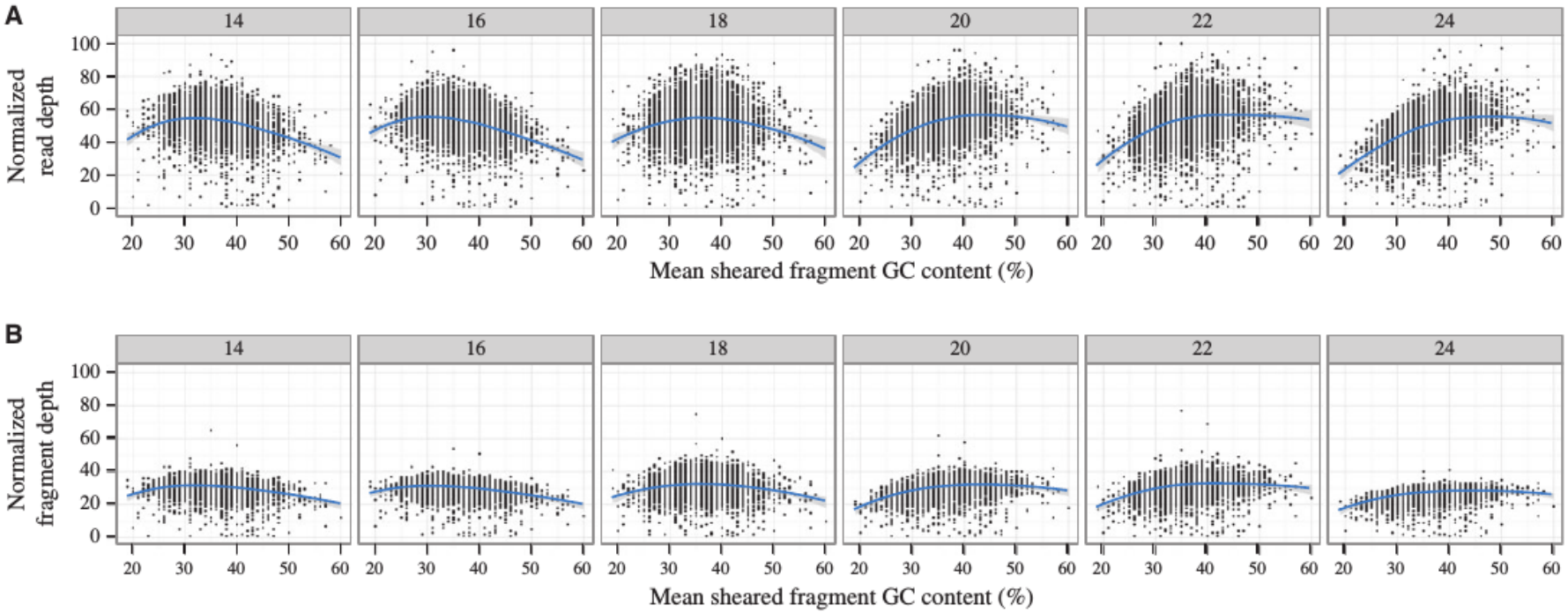


(Davey2013)

Reason: Fragments of <10 kb shear with lower efficiency

Amplification bias in favor of high GC content

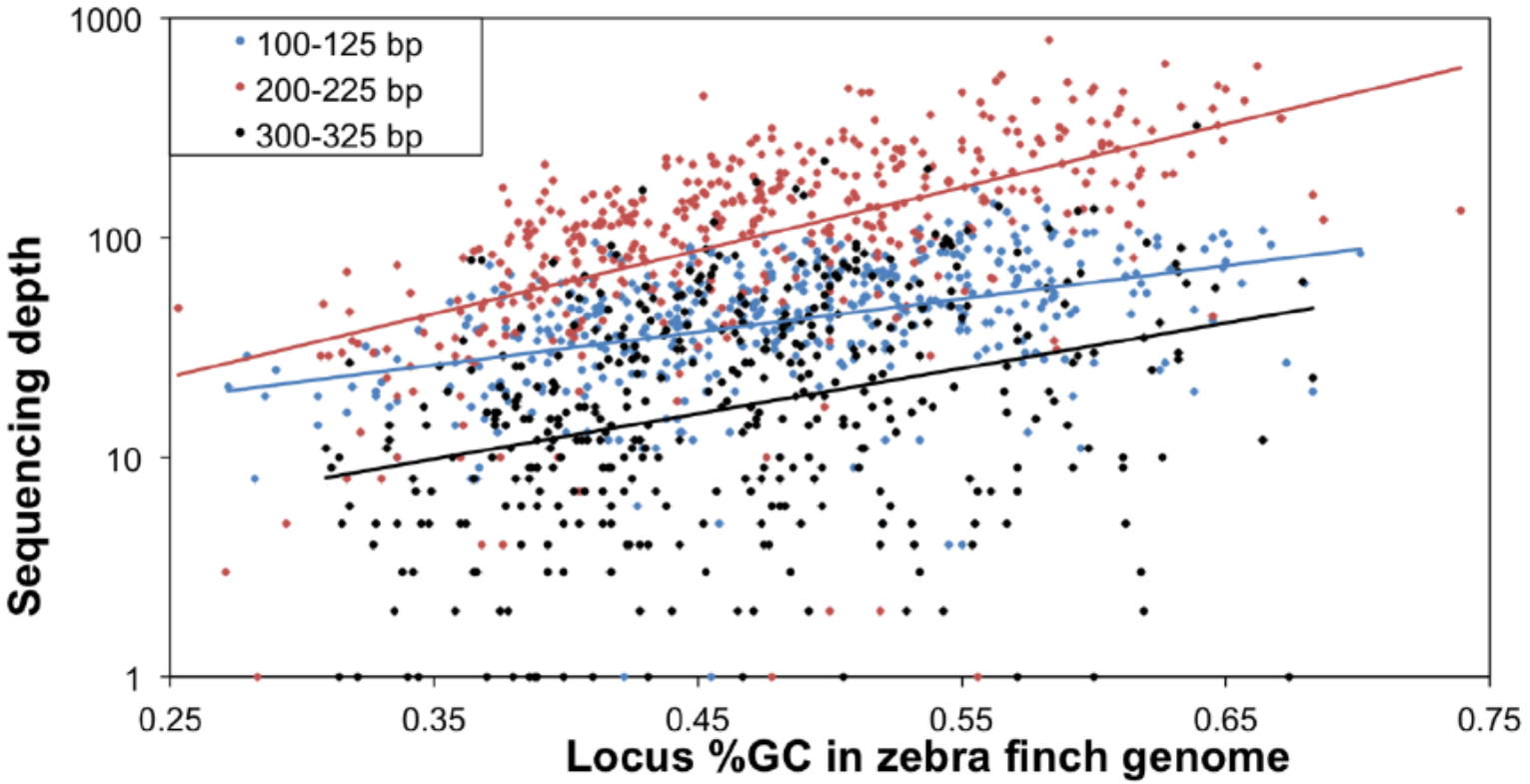
Read depths are influenced by GC content and number of PCR cycles, with (A) or without PCR duplicates (B).



(Davey2013)

Modifications of PCR enrichment can help (see **(Puritz2014b; Benjamini2012)**)

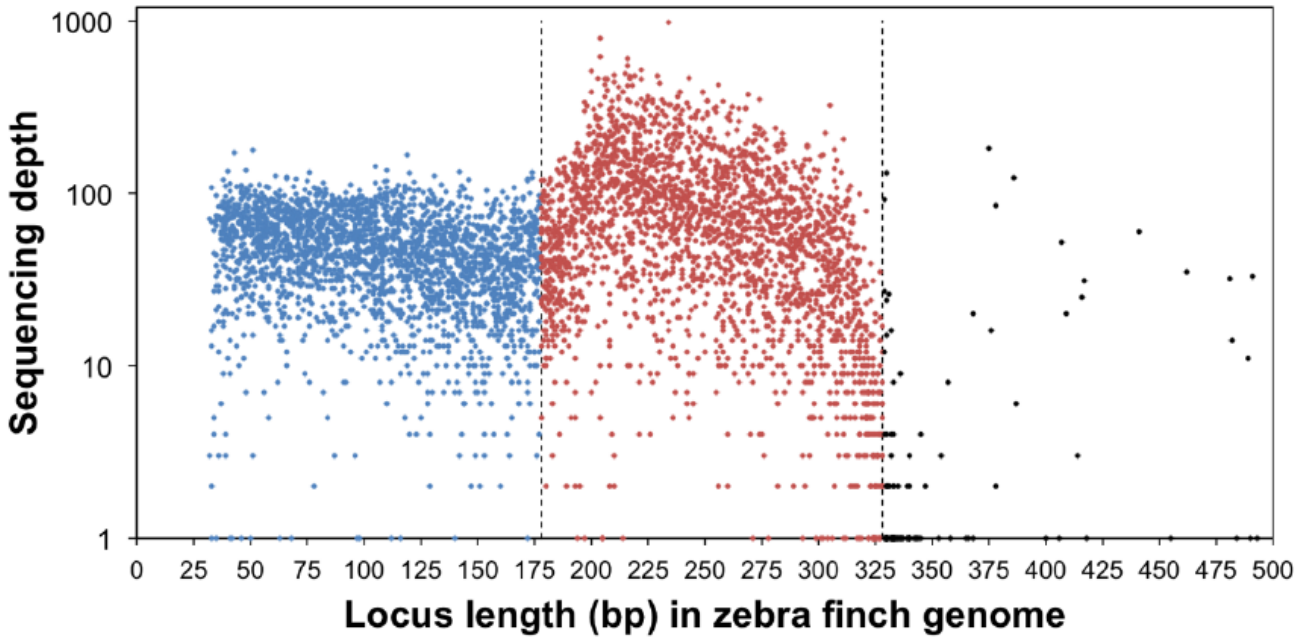
Sequencing depth bias in favor of loci with high GC content



(Dacosta2014)

- 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



(Dacosta2014)

- 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) (**Dacosta2014**).

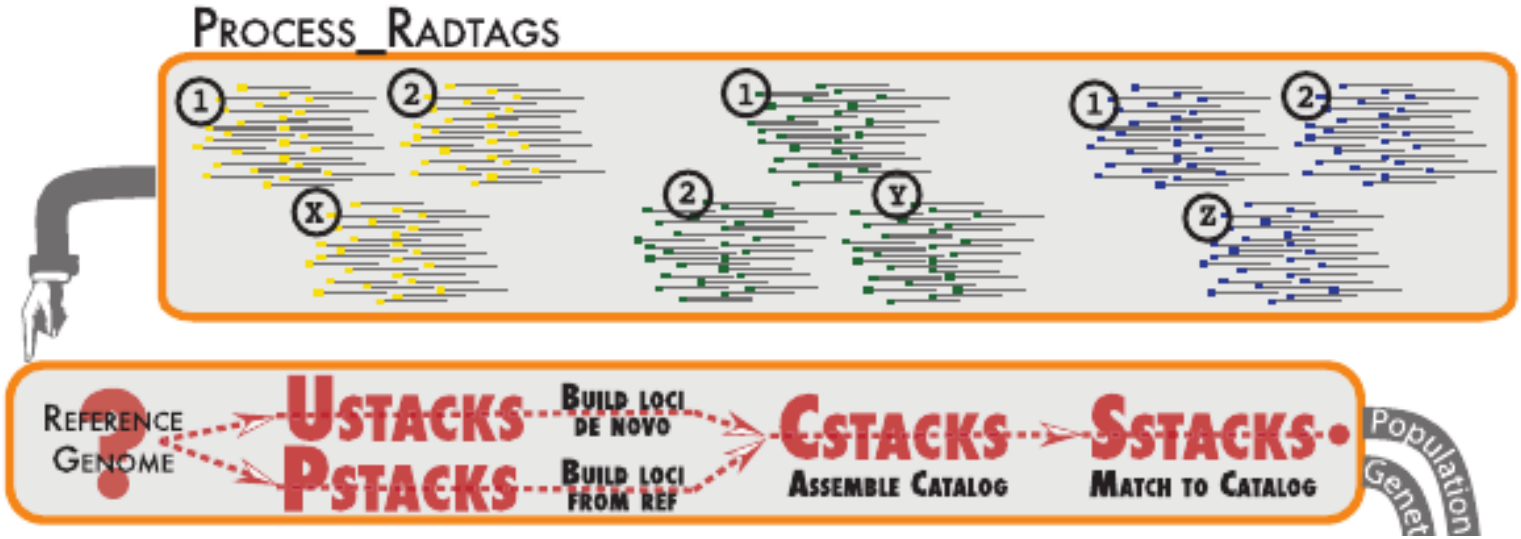
STACKS (Puritz2014)

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

Julian M. Catchen,* Angel Amores,† Paul Hohenlohe,* William Cresko,* and John H. Postlethwait†,1
*Center for Ecology and Evolutionary Biology and †Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403

STACKS - basic pipeline for RAD-Seq

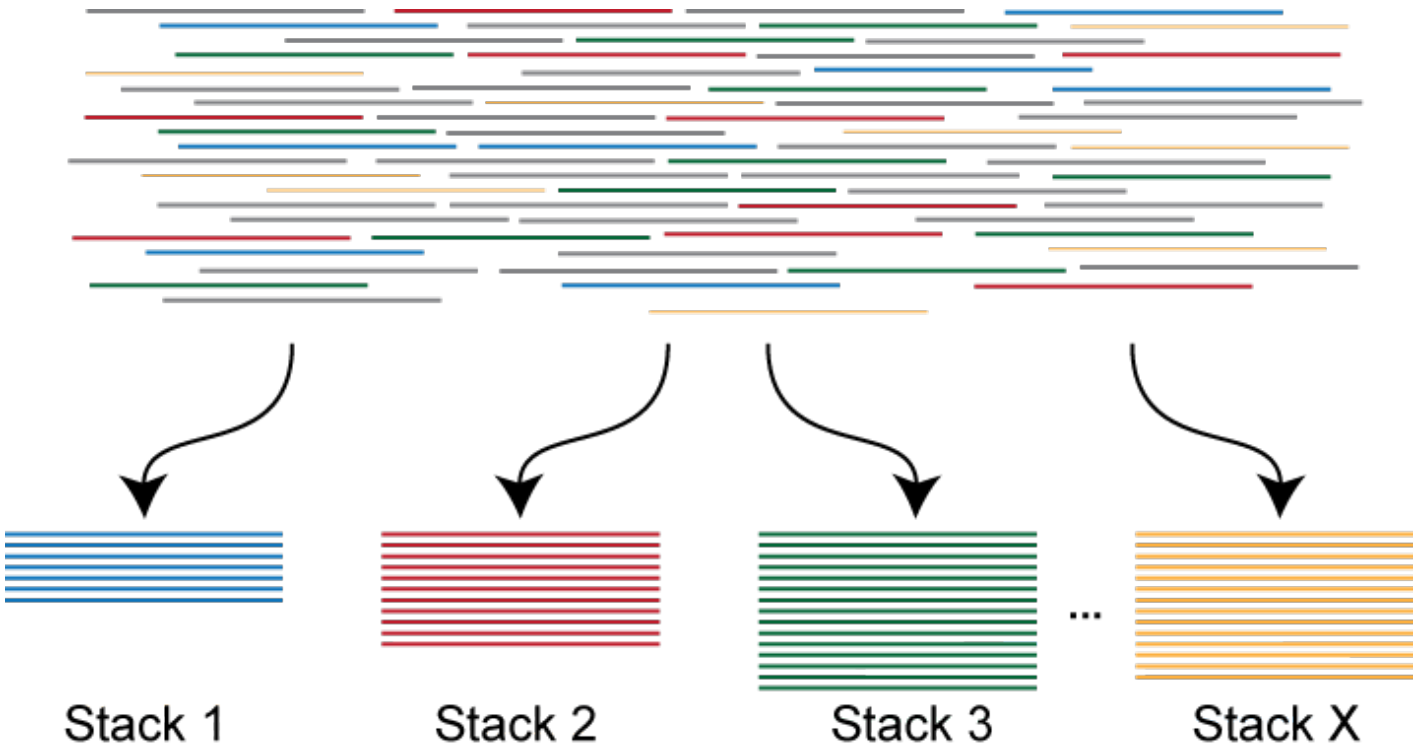
STACKS - software pipleine to build loci from RADseq reads and use them for population genomics and phylogeographic analyses.



(Catchen2013a)

STACKS - Ustacks *de novo* assembly step 1

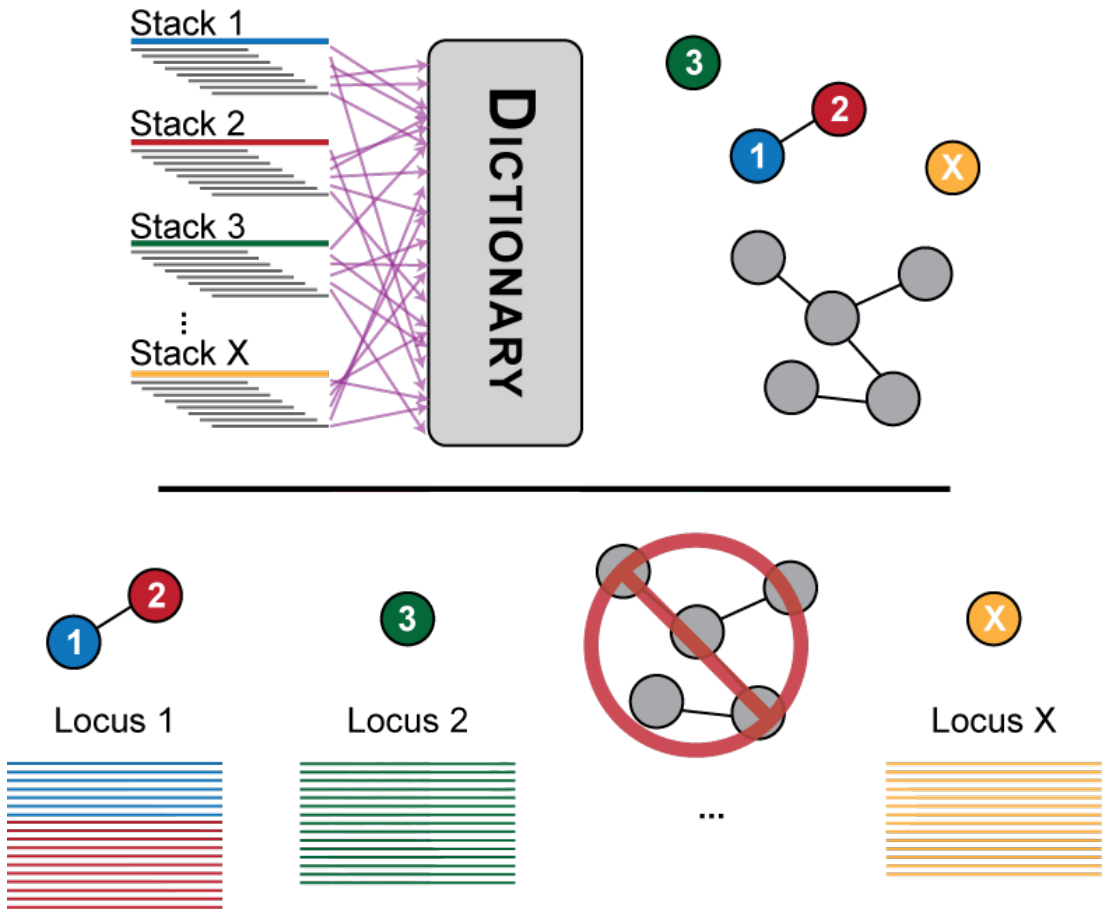
- Only exact matches are assembled
- Secondary reads are set aside
- The minimum stack depth parameter controls the number of raw reads required to form an initial stack



(Catchen2013a)

STACKS - Ustacks *de novo* assembly step 2

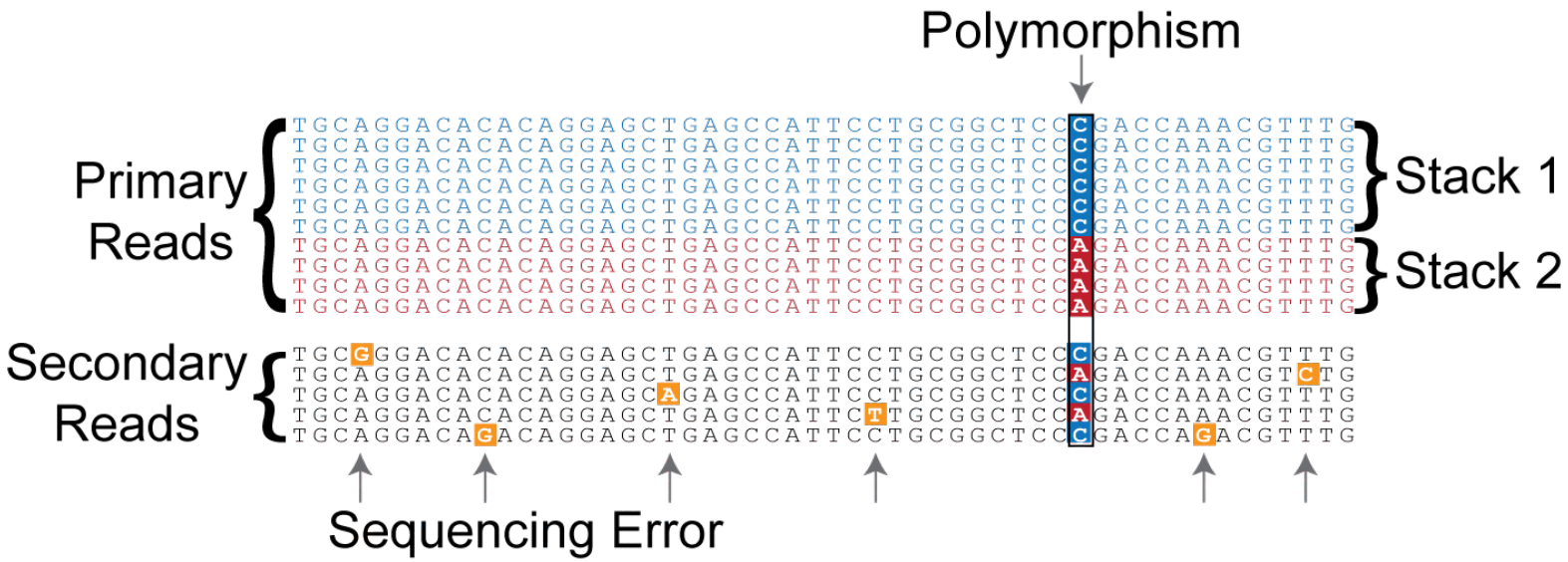
- Stacks with few nucleotide differences are merged.
- Repetitive sequences with many alleles are excluded



(Catchen2013a)

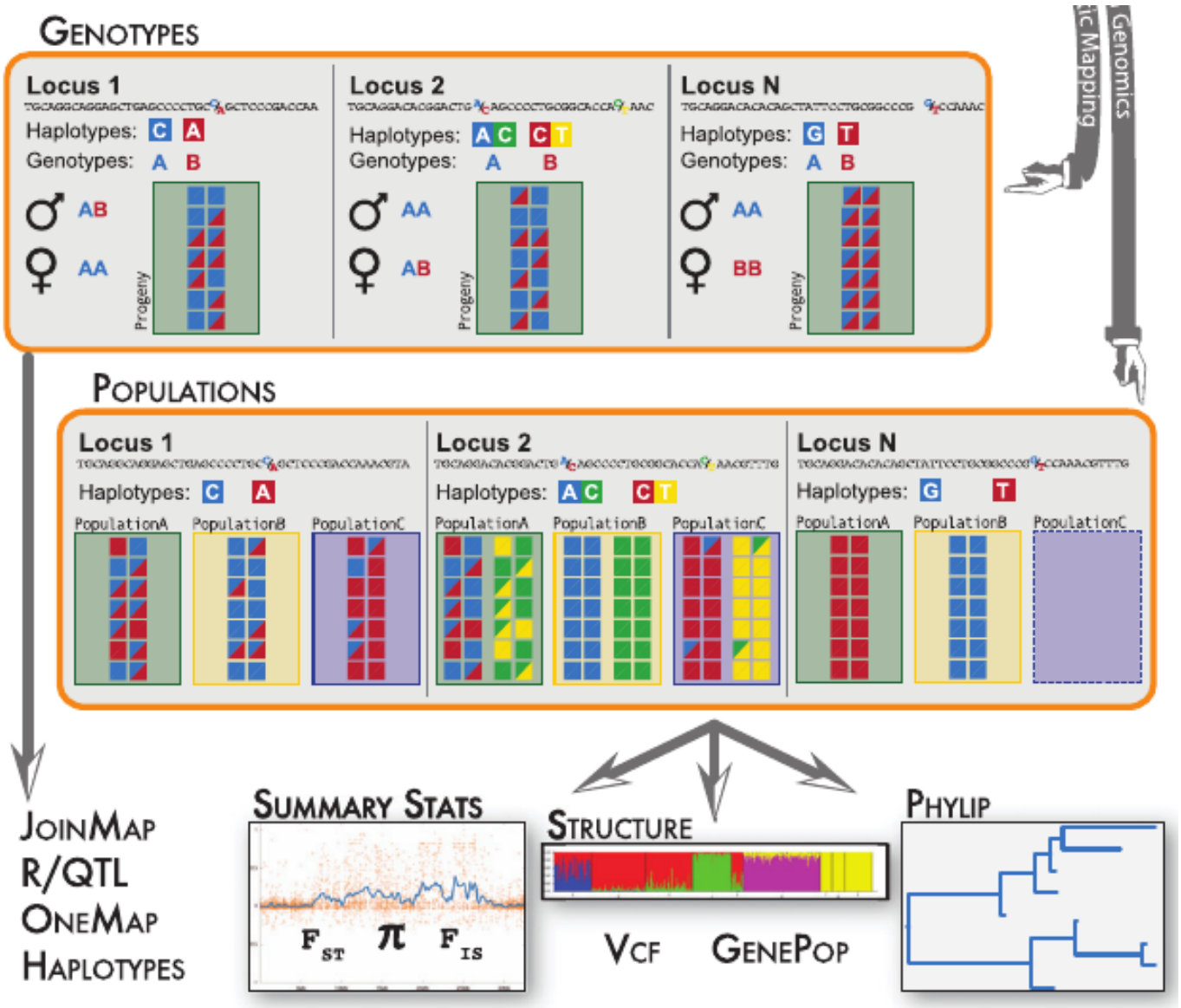
STACKS - Ustacks *de novo* assembly step 3

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



(Catchen2013a)

STACKS - populations or genotypes pipeline



(Catchen2013a)

DDocent (Puritz2014)

PeerJ

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 (Puritz2014)

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

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[1,2]

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

Associate Editor: David Posada

Important considerations

- Degraded DNA interfers 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