GALAXY Community Conference
Hosted by Indiana University June 28–29, 2016

RADSeq Data Analysis

Through STACKS on Galaxy

Yvan Le Bras
Anthony Bretaudeau
Cyril Monjeaud
Gildas Le Corguillé

RAD sequencing: next-generation tools for an old problem

INTRODUCTION

source: Karim Gharbi - edinburgh genomics / University of Edinburgh

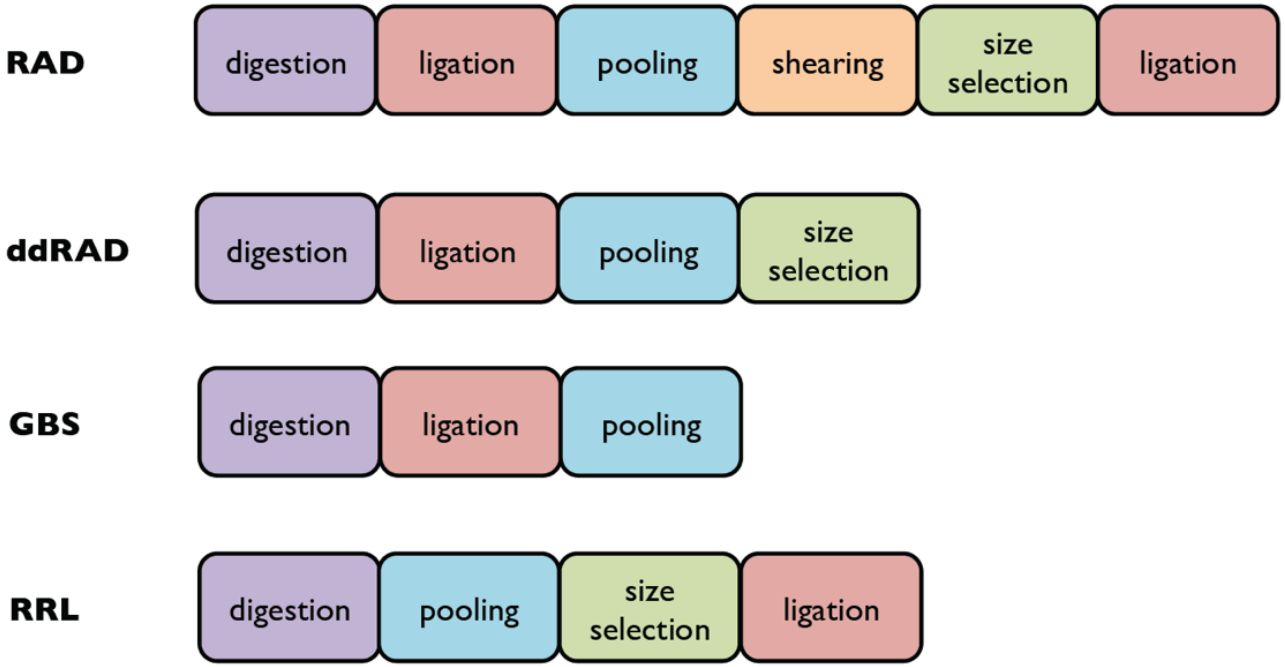
The NGS revolution in the GBS world

- Alloenzymes, RAPD, AFLP, Microsatellites, SNP array, […], **NGS**

- **NGS:** Low cost sequencing …
  - But it's **still expensive** to get enough markers on enough samples
  - Solution: sampling the genome

  - BEWARE: the analysis is not cheap!

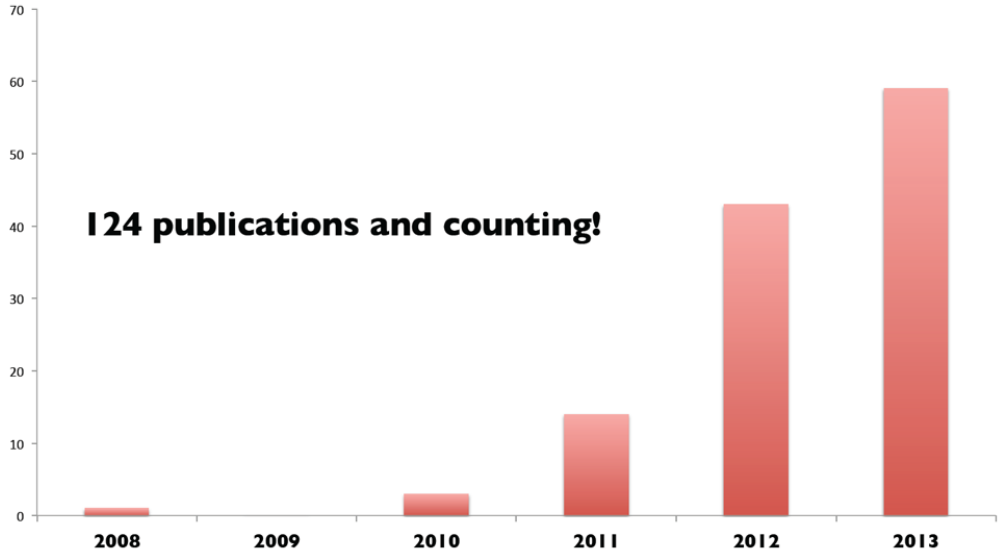
Sampling the genome



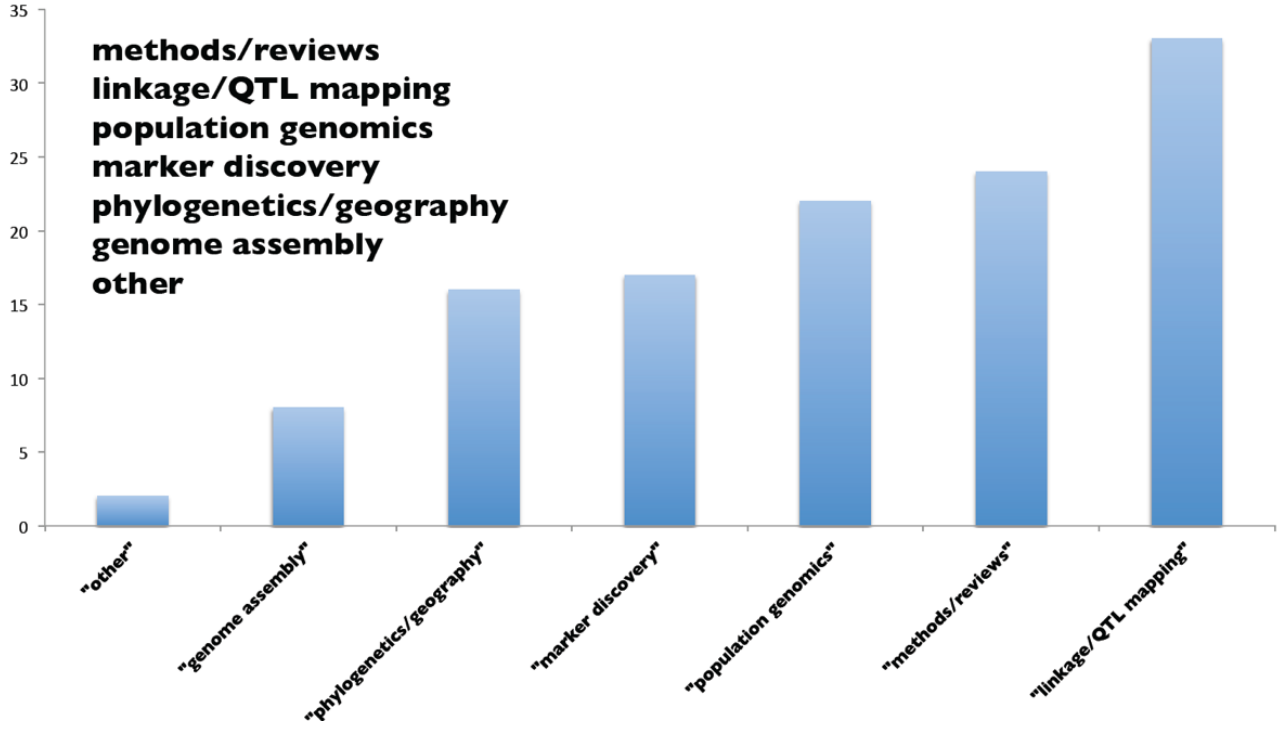
Used by the Roslin institute for their SNP arrays

Sampling the genome

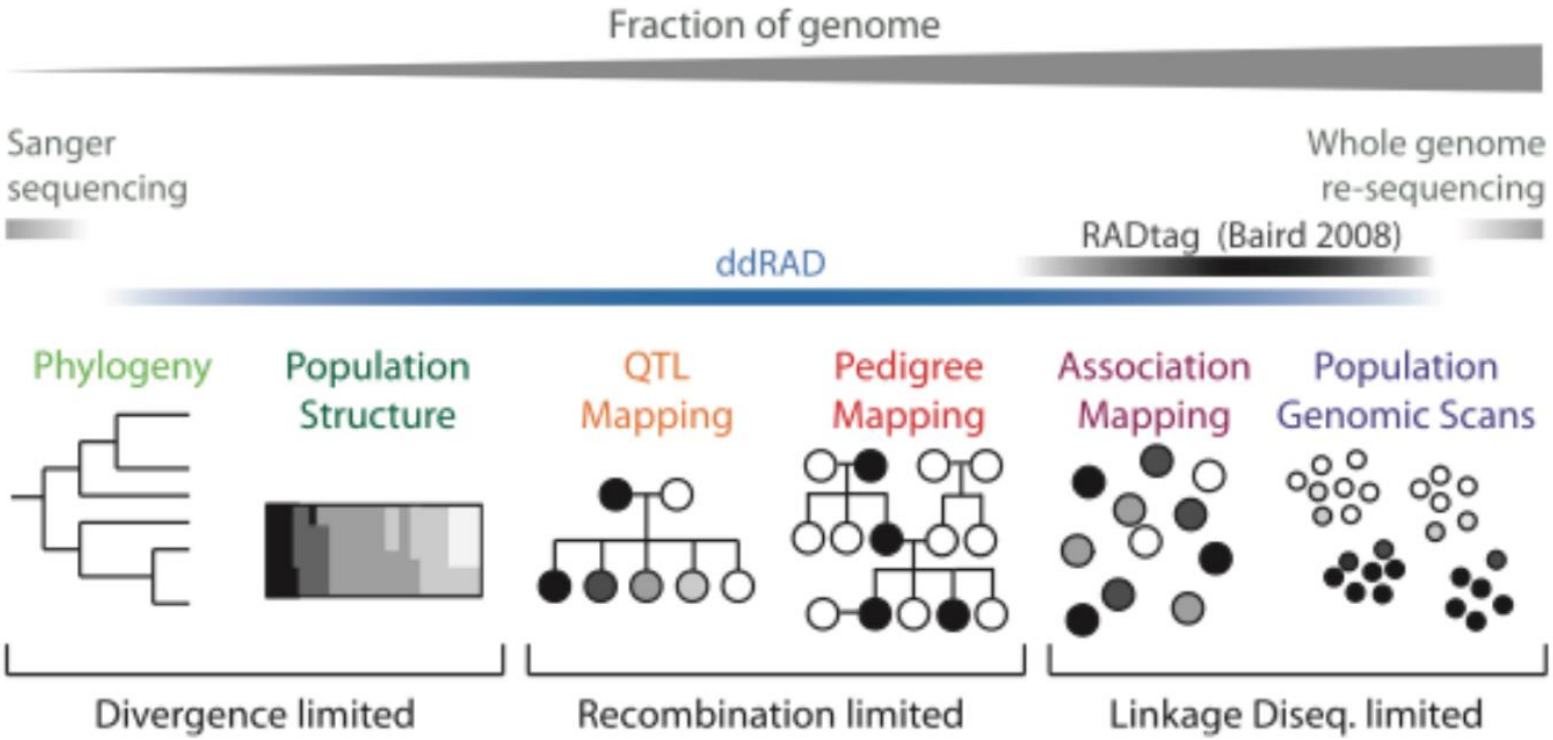
- Original paper: Eric Johnsson, 2008
- Hohenlohe is in the authorship of the first five



Applications



Applications



Peterson et al PLOS One 2012

Classic RAD

Protocols

Single-end RAD

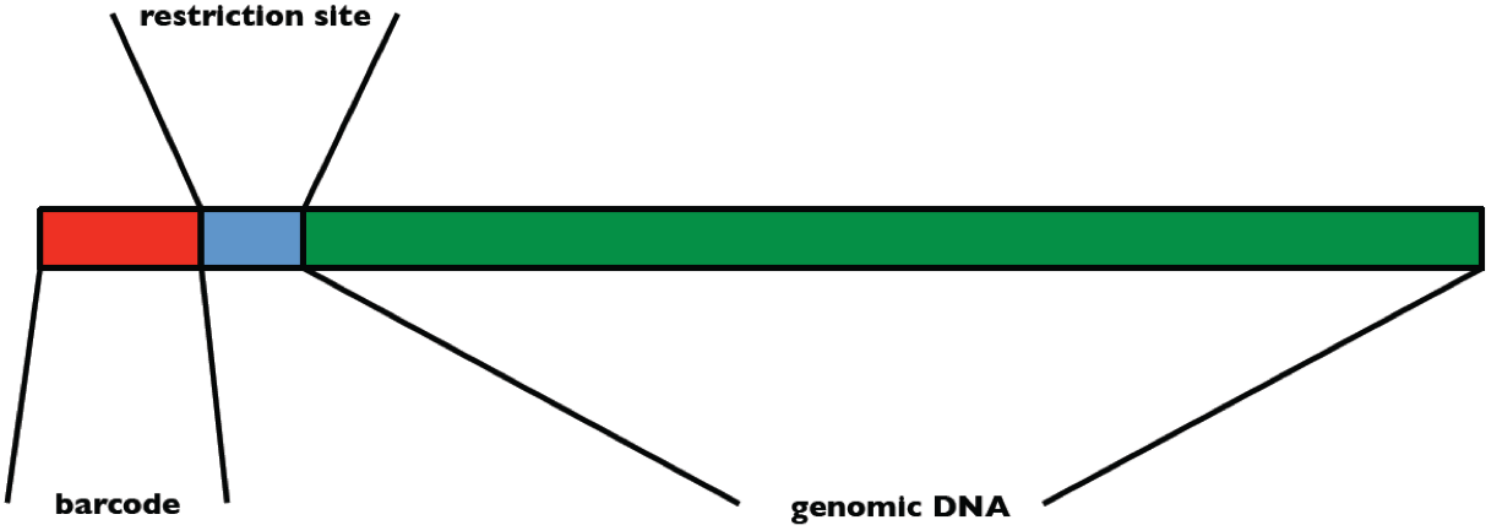
restriction site

Illumina read (30-300 bp)

1 restriction site = 2 RAD tags

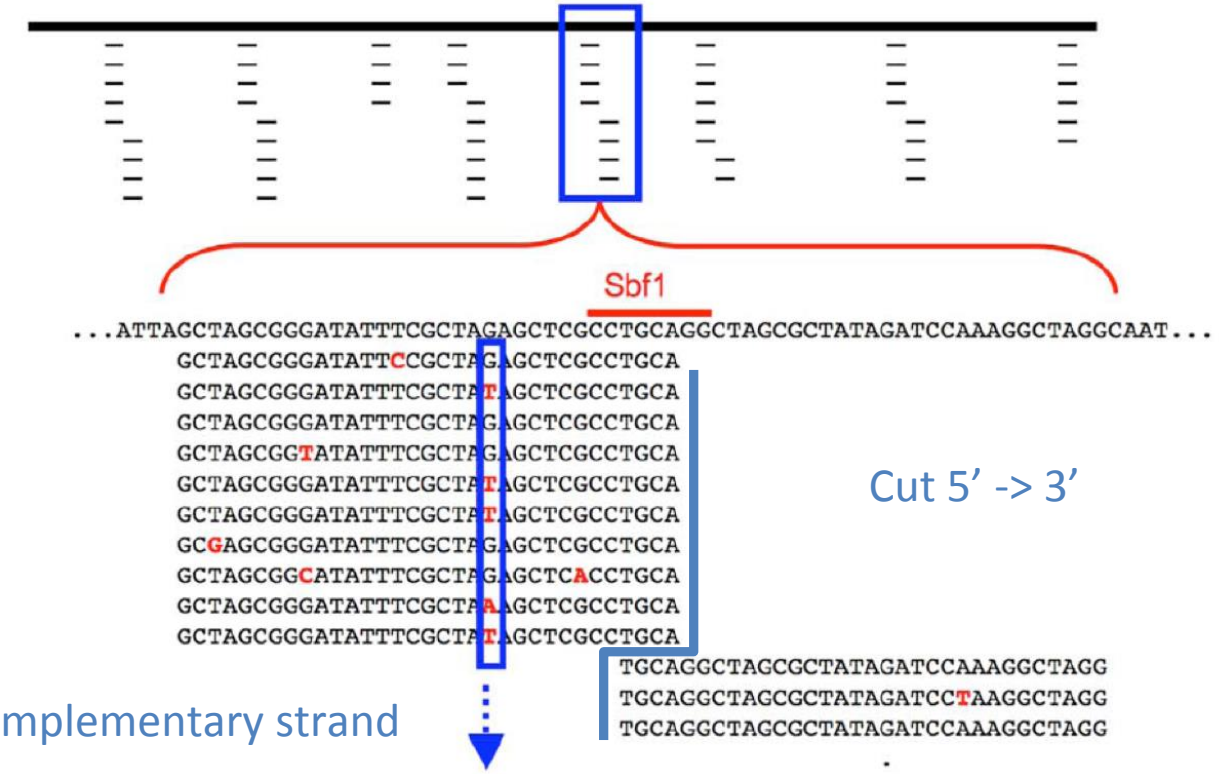
Baird et al 2008

Single-end RAD



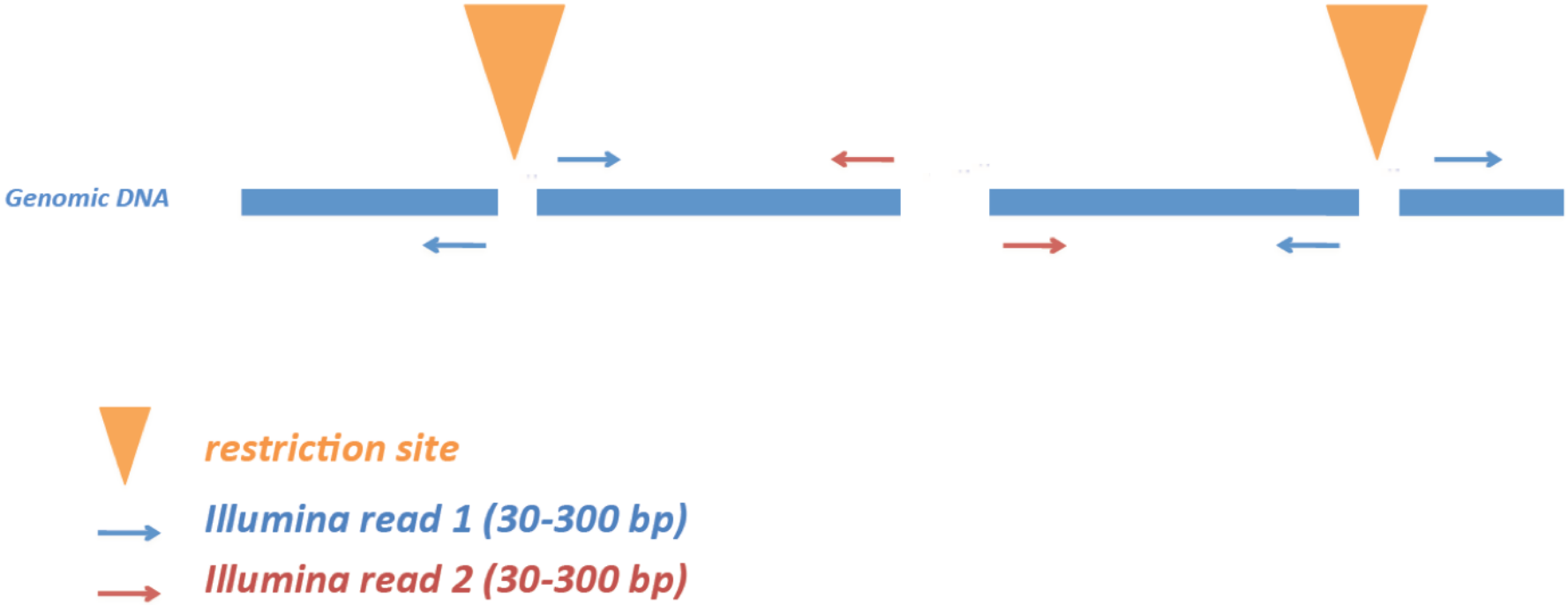
@M00689:44:000000000-A1N97:1:1101:11642:2590 1:N:0:1
CTGATGCTTGCAGGACGCACCTCCCCGCGGTGCGGCTAATGTCCTCGCAGC
+
AAAAABBBDDDDDDDDGGGGGGIIIHHHHHEHHHHHBHHIIIIIIIHHH@E

Single-end RAD

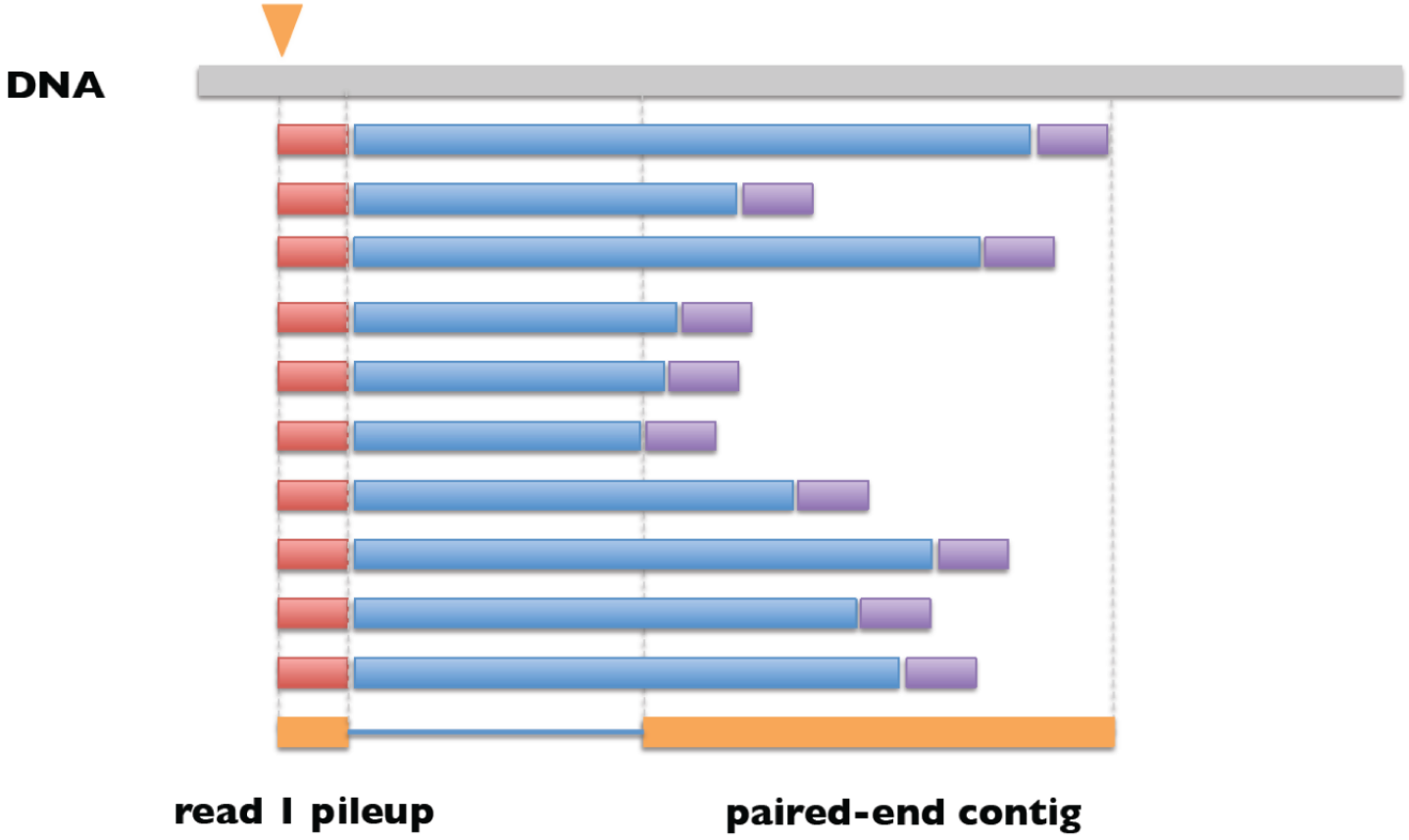


Hohenlohe et al., PLoS Genetics 2010

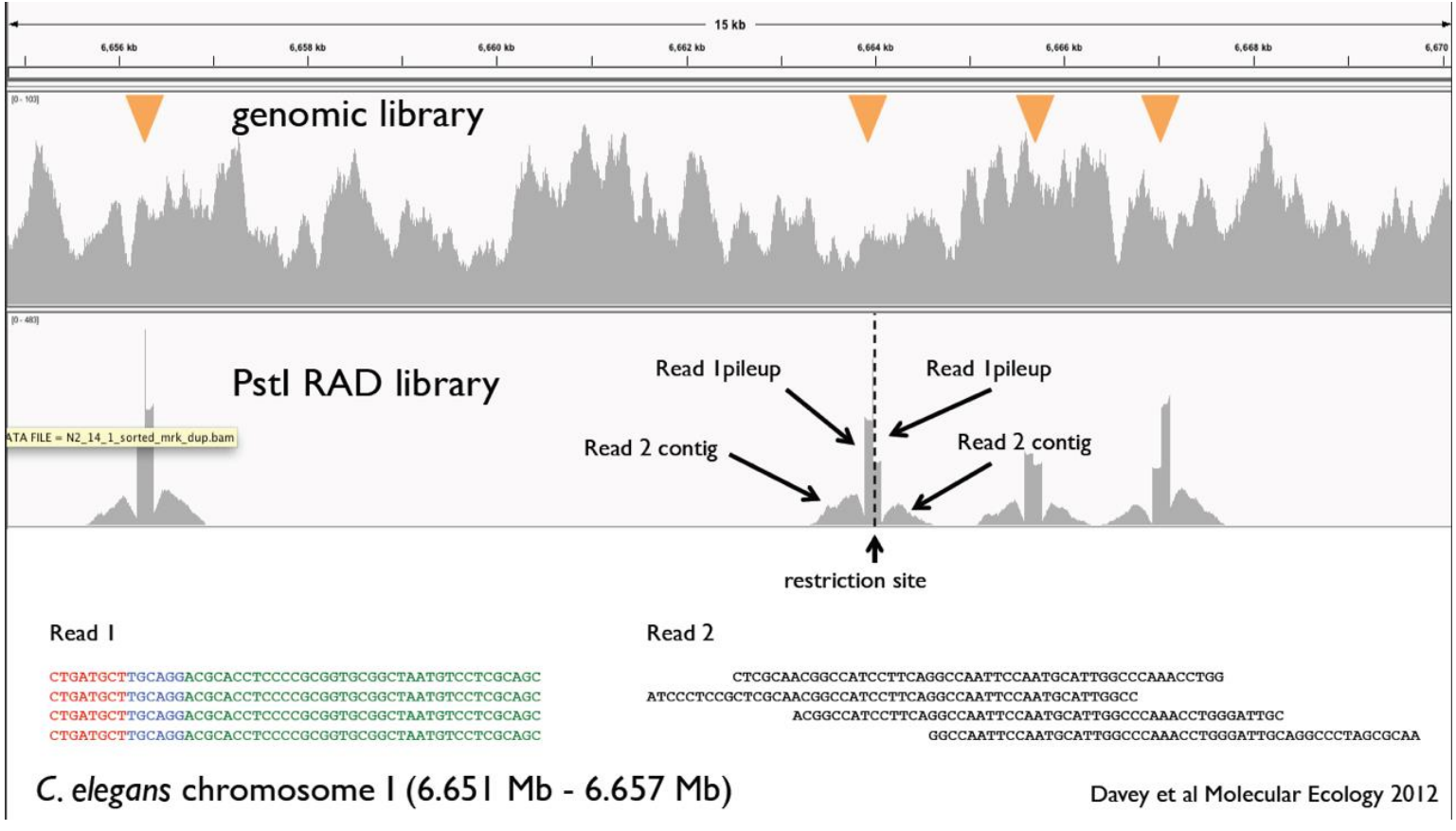
Paired-end RAD



Paired-end RAD



Paired-end RAD



Single vs Paired-end RAD

	single-end	paired-end
Library preparation	=	=
Costs	+	++
Bioinformatics	+	++
Bases per tag	up to 300	up to 300-500
Design of genotyping assays	limited	good
Filtering of duplicate reads	no	yes
Paralog resolution	+	++

ddRAD

Protocols

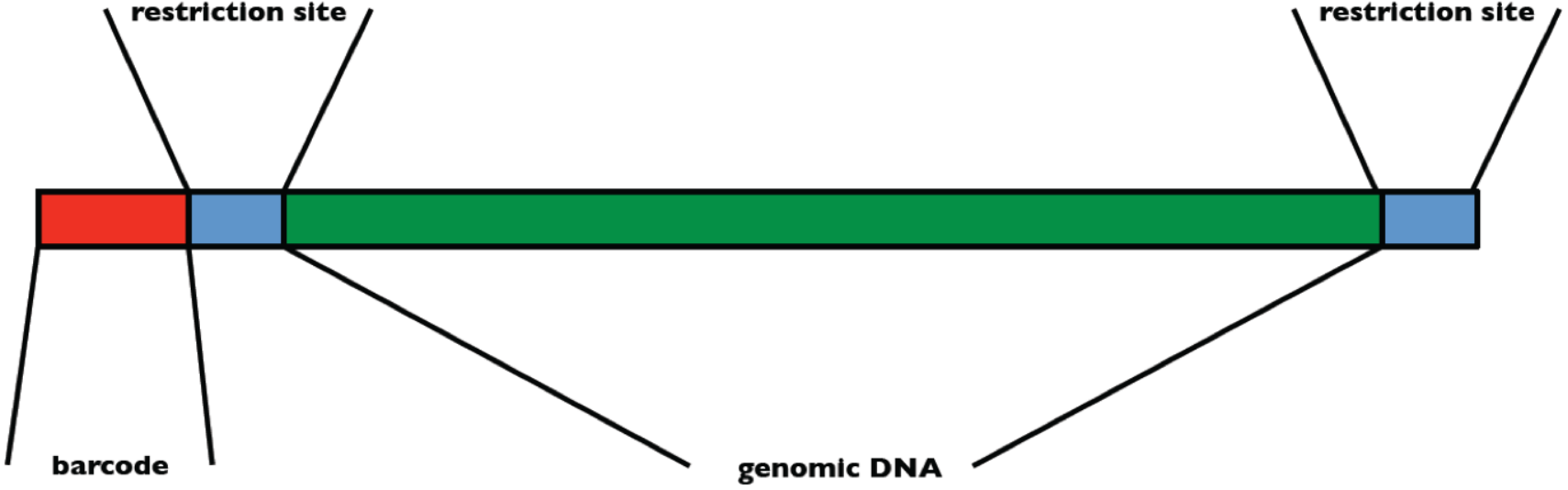
ddRAD

restriction site

Illumina read 1 (30-250 bp)

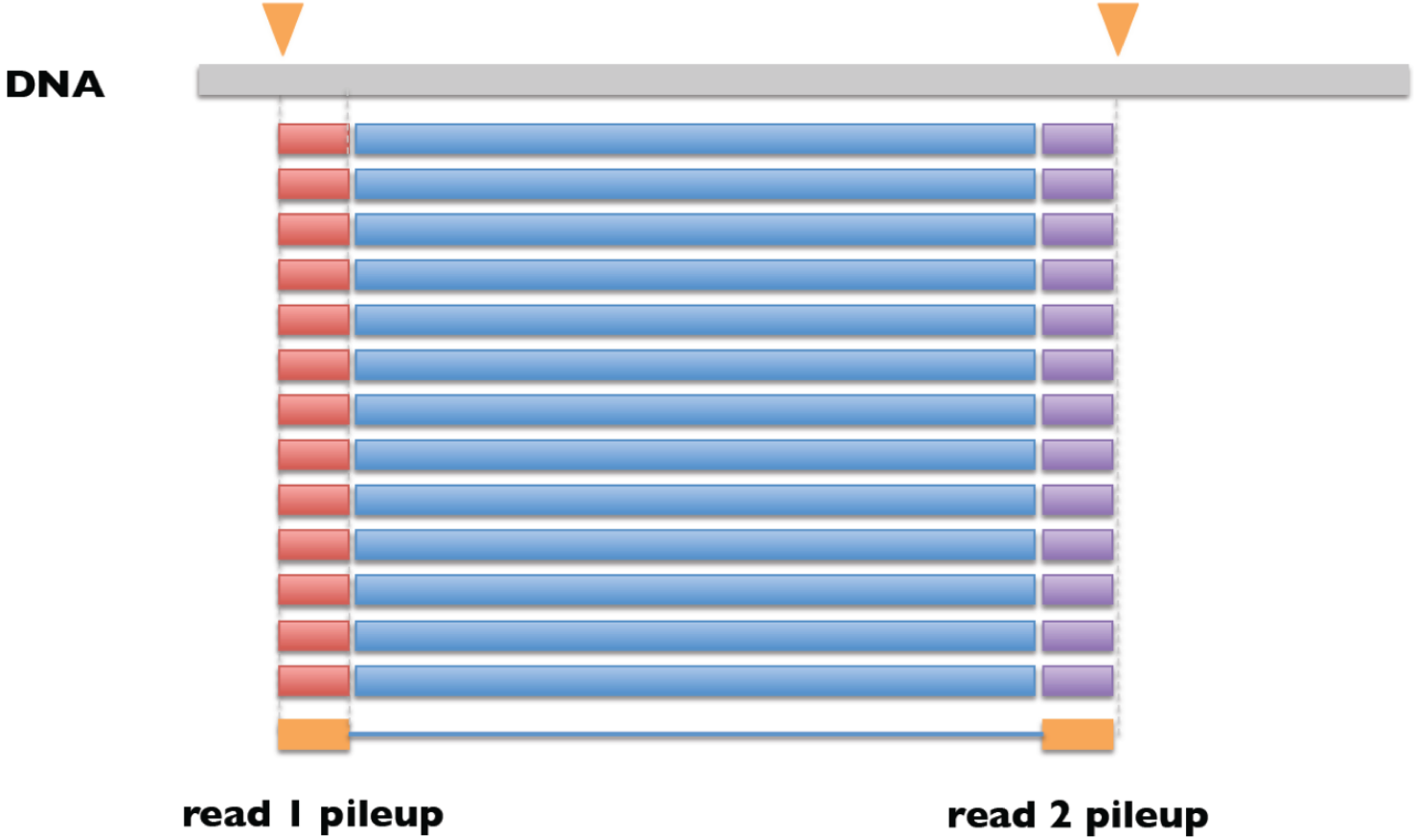
Illumina read 2 (30-250 bp)

ddRAD

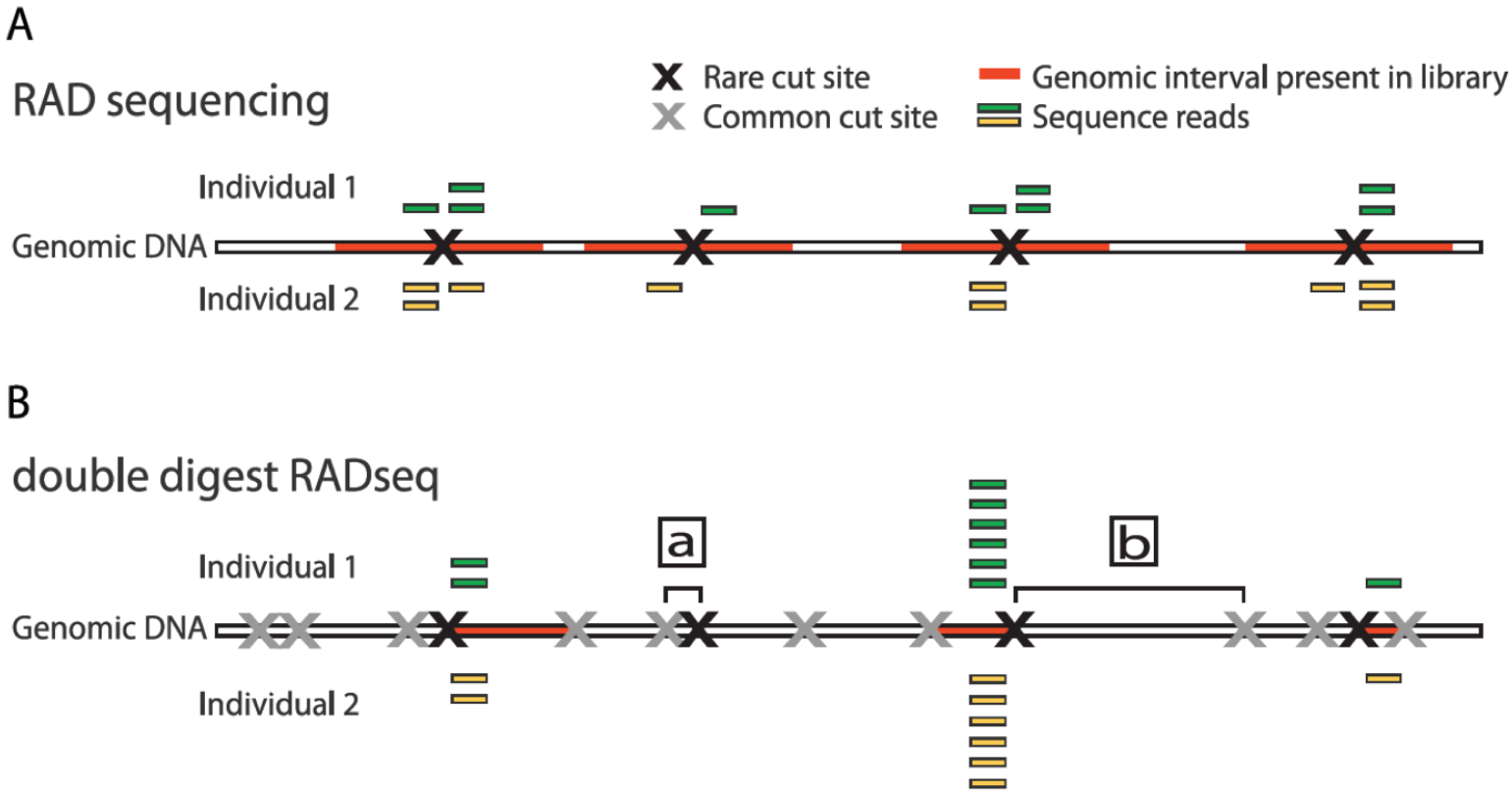


~ 500 pb

Paired-end ddRAD



RAD vs ddRAD



Peterson et al PLOS One 2012

RAD vs ddRAD

- **classic RAD**: reads between the restriction site and a random site (shearing/sonication)

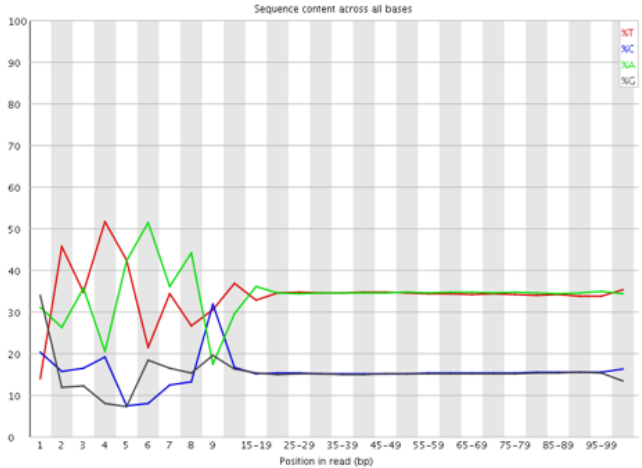
- **ddRAD**: reads between the 2 restriction sites. So more flexibility on the balance coverage / depth of covarage

Common biases

Biases

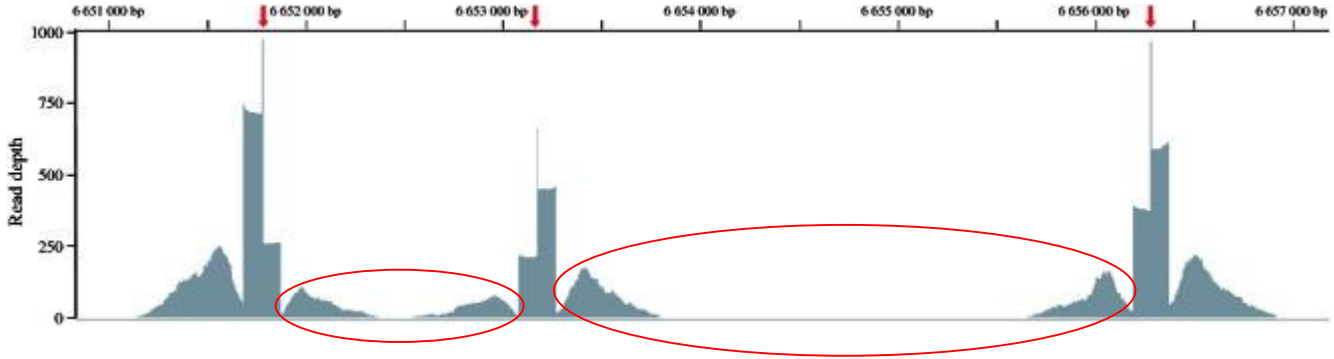
Because all reads begin with [half of] the restriction site

- Consequence:
  - The Illumina sequencer have difficulty separating polonies/clusters during the first cycles imaging step
- Solution:
  - use a set barcodes with different sizes
  - mix different experiences which use different restriction enzymes



Biases

Restriction fragment length biases read depth



source: Special features of RAD Sequencing data: implications for genotyping (2013)

Biases

Mutations within the recognition sequence of the restriction enzyme

- Consequence:
  - Allele dropout (ADO)
    - → overestimates genetic variation both within and between populations
- Solution:
  - Filter any loci that are not represented in all genotyped individuals

source: The effect of RAD allele dropout on the estimation of genetic variation within and between populations (2012)

Advice/Information from the edinburgh genomics

- 250 ng of DNA is needed, 1 µg is asked by the Edinburgh genomics
- High quality DNA if not from 30 to 40% of data can be useless
- PCR: 12 to 14 cycles to reduce the PCR duplicates
- Warning: QiaGen Licence / Patent

PROS / CONS

THE RAD FAMILLY

BATTLE

MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 1661–1667

NEWS AND VIEWS

MEETING REVIEW

Recent novel approaches for population genomics data analysis

KIMBERLY R. ANDREWS* and GORDON LUIKART†

School of Biological & Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK; †Flathead Lake Biological Station, Fish and Wildlife Genomics Group, University of Montana, Polson, MT 59860, USA

MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 5937–5942

NEWS AND VIEWS

COMMENT

Demystifying the RAD fad

JONATHAN B. PURITZ,* MIKHAIL V. MATZ,† ROBERT J. TOONEN,‡ JESSE N. WEBER,§ DANIEL I. BOLNICK§ and CHRISTOPHER E. BIRD¶

Marine Genomics Laboratory, Harte Research Institute, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, TX 78412-5869, USA; †Department of Integrative Biology, University of Texas at Austin, 205 W 24th ST C0990, Austin, TX 78712, USA; ‡Hawai'i Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawai'i at Mānoa, PO Box 1346, Kāne'ohe, HI 96744, USA; §Department of Integrative Biology, Howard Hughes Medical Institute, University of Texas at Austin, Austin, TX 78712, USA; ¶Department of Life Sciences, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, TX 78412-5869, USA

mbRAD

- Original RAD (mbRAD Miller *et al.* 2007 and Baird *et al.* 2008)
  - Genomic DNA digestion by 1 restriction enzyme (low frequency cutter)
  - Ligation of barcode containing adapters onto digested 5’ ends
  - Ligated genomic DNA sonication
  - Ligation of a 3’ adapter to the sonicated end
  - Pooling of the samples
  - Size-selection of the library
  - RAD fragments PCR enrichment

mbRAD - PROS

- Random shearing of the 3’ end helps to identify putative PCR duplicates
  - If identical starting position of the paired-end read: duplicate
- Random shearing improves the distribution of coverage
- Random shearing + larger insert size ranges: *de novo* assembled RAD loci are of greater length
  - Critical for identifying function & Gene ontology
- Coverage and quality are fundamental!!!
  - Distinguishing true SNP from sequencing error: if coverage is low, your statistical test will not yield significant results!

mbRAD - CONS

- The most technically challenging and complex protocol!
- Requires non standard lab equipment: sonicator
- Restriction fragment length bias (due to the shearing)
  - Sequencing at different depth
- Strand bias
  - Different genotypes from forward & reverse reads
  - Solution: Filter any loci in this case… only possible in 2bRAD

ddRAD

- Double digest RAD protocol (Peterson *et al.* 2012)
  - Genomic DNA digestion by 2 restriction enzymes (low + high frequency cutter)
  - Ligation of barcode P1 adapters (matching the first restriction site) and P2 adapters (matching second restriction site)
  - Pooling of the samples
  - Size-selection of the library
  - RAD fragments PCR enrichment + second barcode introduction to increase multiplexing potential

  - Extremely simila
  - Pros & cons asso

Tseq (Stolle & Moritz 2013)

ddRAD - PROS

- Greatest degree of customization
  - Depending on the chosen enzymes & the selected range of fragment sizes
  - Allow to have hundreds of SNPs per individual at very low cost or thousands for QTL mapping experiments at moderate cost
  - Flexibility on the balance coverage / depth of covarage

- Examine histograms of digested samples early
  - Identify / exclude excessively frequent fragments (i.e. transposons)

ddRAD - CONS

- Using fragment size selection to tune the quantity of loci can lead to variable representation of some loci
  - This can be minimized using precise selection tool (i.e. Pippin Prep)
- Particularly susceptible to ADO (Arnold *et al.* 2013)
  - To be considered when performing sensitive population genetic analyses
- Requires the highest quality genomic DNA of all RAD methods
  - Proper fragment ligation relies on completely intact 5’ & 3’ overhangs!
- PCR duplicates cannot be detected

ezRAD

- ezRAD protocol (Toonen *et al.* 2013)
- Genomic DNA digestion by 2 restriction enzymes (high frequency cutter on the same cut site)

- Commercially available Illumina TruSeq library preparation kit
  - DNA end reparation
  - Ligation of single or dual indexing adapters onto genomic fragments
  - Pooling of samples
  - Size selection of the library

- RAD fragments PCR enrichment, or not, depending on the Illumina kit

ezRAD - PROS

- Illumina TruSeq kit
  - Extensive manual, customer support & guarantee
  - Probably the simplest path to obtain RAD data for small lab without experience / equipment / resources to develop in-house RAD capability

- Combined with an Illumina PCR-Free TruSeq kit, ezRAD is the only RAD protocol that can bypass all potential PCR bias

ezRAD - CONS

- Illumina TruSeq kit
  - Simplicity & uniformity but expensive
  - However can be used with ½ & 1/3 reaction volumes

- All ezRAD reads start with the same four GATC bases
  - The first 4-5 nucleotides of Read 1 are used to discriminate between adjacent clusters
  - If always the same 4 first bases, difficulty to discriminate the different samples

2bRAD

- 2bRAD protocol (Wang *et al.* 2012)
  - Genomic DNA digestion by 1 restriction enzyme (36-bp fragments excision recognition site + adjacent 5’ & 3’ base pairs)
  - Ligation of dual barcode adapters
  - Agarose gel target band excision after PCR enrichment
  - No intermediate purification stages
  - No size-selection

2bRAD - PROS

- Extreme protocol simplicity & cost-efficiency
  - No intermediate purification stages
  - No need for special instrumentation (only PCR + standard agarose gel)

- Lack of biases due to fragment size selection
  - All endonuclease recognition sites can be sampled

2bRAD - CONS

- Difficulties to map 36 bp tags in a unambiguously way
  - But works well in no or moderately duplicated genomes (i.e. Wang *et al* 2012 on *Arabidopsis*)

- 2bRAD fragments cannot be used to build genome contigs

- 2bRAD fragments are less likely to be cross-mappable across large genetic distances, such as across different species

Conclusions

- Most important considerations when selecting a particular RAD protocol are
  - The facilities & the molecular experience of the researcher applying the approach
  - The biology of the organisms
  - The hypotheses being tested

- All RAD protocols are powerful tools for SNP discovery & genotyping of nonmodel species

- It is important to learn about pitfalls inherent to each method & how to adress them

SOFTWARES

Main Bioinformatics pipelines

- STACKS
  - Website: http://catchenlab.life.illinois.edu/stacks/
  - mbRAD, ddRAD, ezRAD & 2bRAD?
  - ~~STACKS does not handle INDELS, so any loci near an INDEL is lost~~
  - STACKS does not call SNPs from paired end reads natively, and does especially poorly with paired end fragments that are not of a random length (e.g., ddRAD and ezRAD)

- dDocent
  - Website: https://ddocent.wordpress.com/ddocent-pipeline-user-guide/
  - ddRAD & ezRAD

- PyRAD
  - Website: http://dereneaton.com/software/pyrad/
  - mbRAD, ddRAD, PE-ddRAD, GBS, PE-GBS, EzRAD, PE-EzRAD, 2B-RAD
  - use of an alignment-clustering method (*vsearch*)

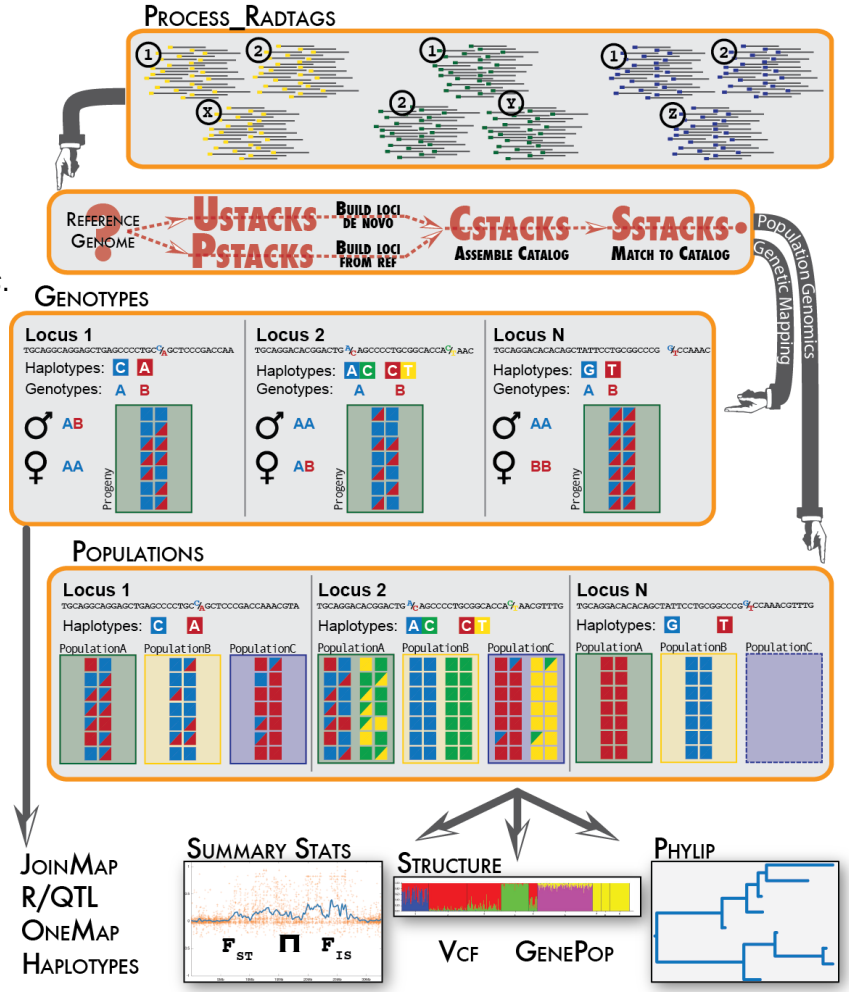
- 2bRAD (Wang *et al* 2012)
  - *de novo:* https://github.com/z0on/2bRAD_denovo
  - With reference genome: https://github.com/z0on/2bRAD_GATK
  - 2bRAD

SOFTWARES

Stacks

http://catchenlab.life.illinois.edu/stacks

J. Catchen, A. Amores, P. Hohenlohe, W. Cresko, and J. Postlethwait. *Stacks: building and genotyping loci de novo from short-read sequences*. **G3: Genes, Genomes, Genetics**, 1:171-182, 2011.



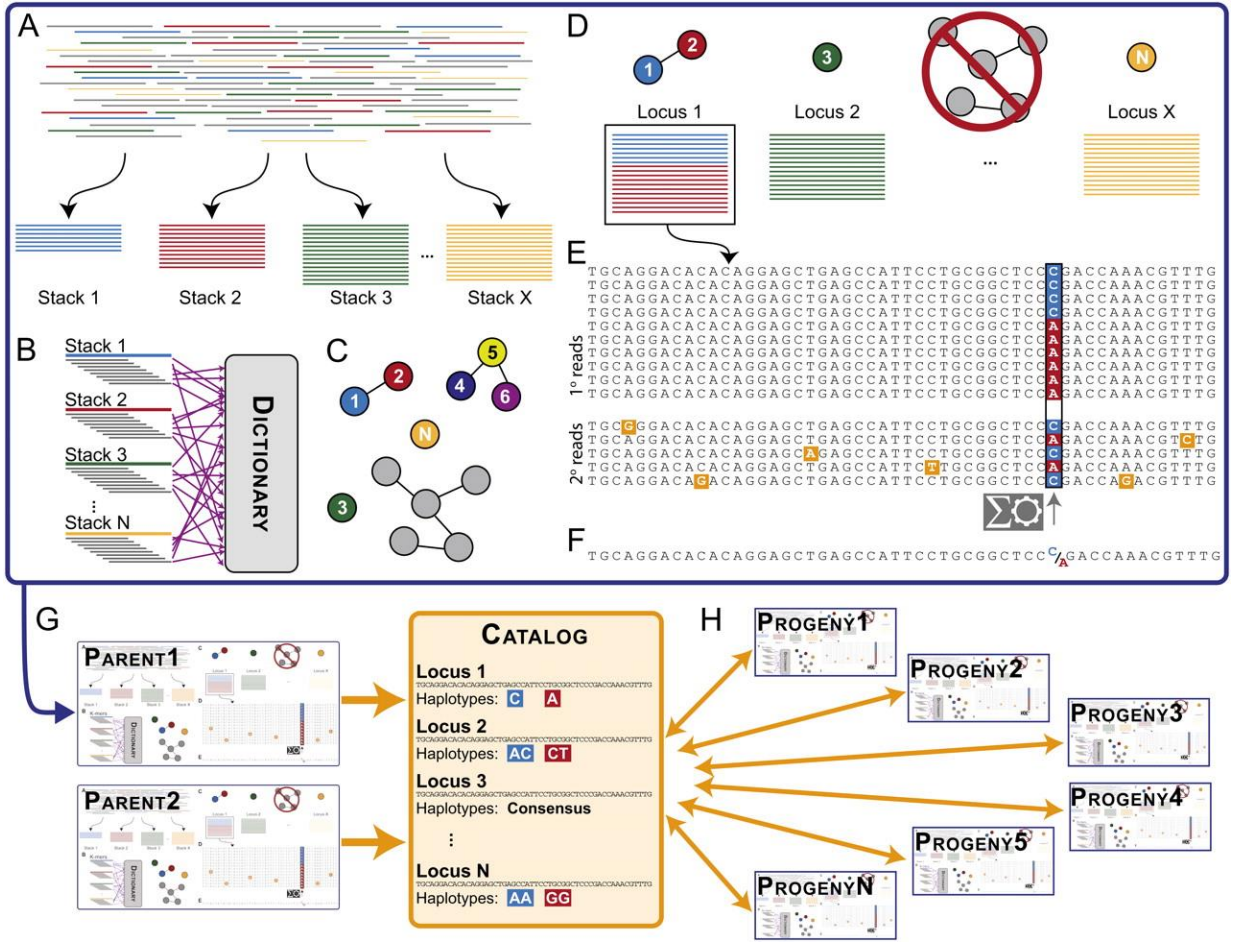
Stacks

denovo_map pipeline

ustacks

cstacks

sstacks



Stacks

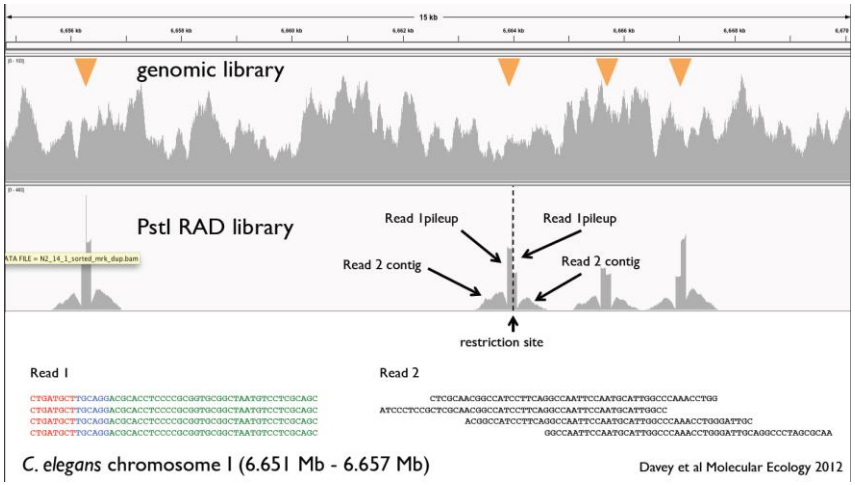
ref_map pipeline

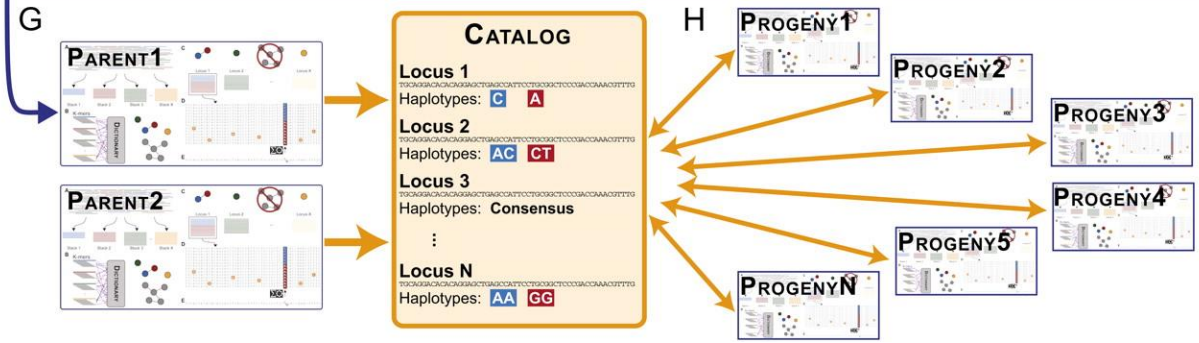
pstacks

cstacks

sstacks

.bam
.bam
.bam





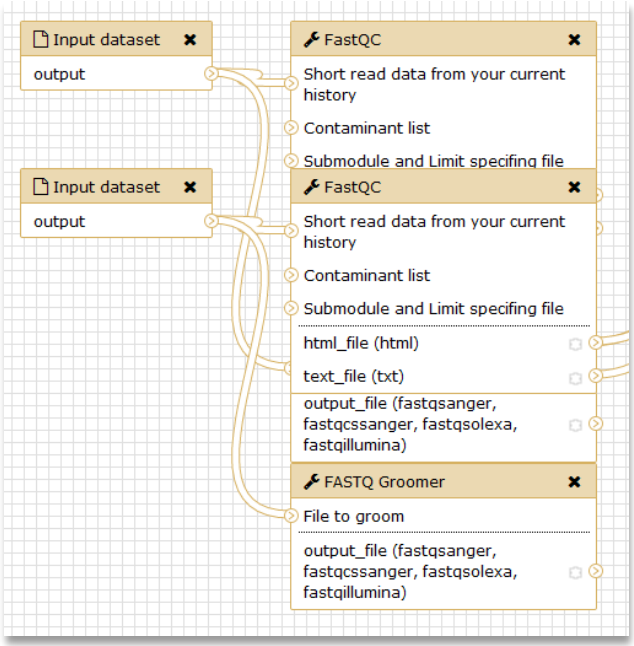
SOFTWARES

Today: hands on

- SNP detection
  - On 2 parents of a family
- Genetic map
  - On a family with 93 offsprings
- Mini-contig assembly
  - Paired end data
- Population genomics
  - Without reference genome
  - With reference genome

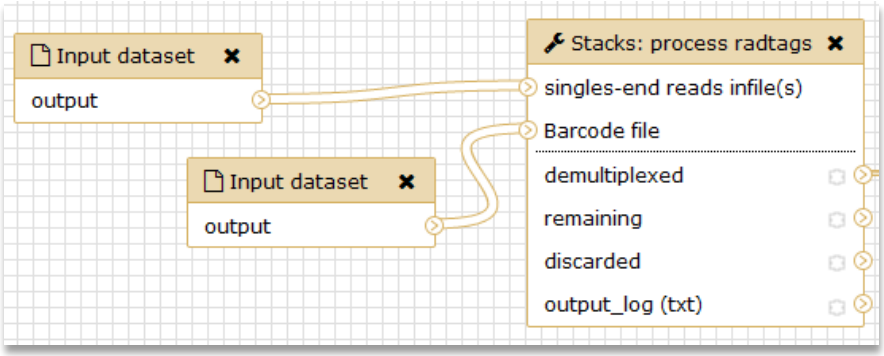
Today: hands on

- SNP detection
  - On 2 parents of a family
- Genetic map
  - On a family with 93 offsprings
- **Mini-contig assembly**
  - **Paired end data**
- Population genomics
  - Without reference genome
  - With reference genome



Today: hands on

- SNP detection
  - On 2 parents of a family
- Genetic map
  - On a family with 93 offsprings
- Mini-contig assembly
  - Paired end data
- **Population genomics**
  - **Without reference genome**
  - With reference genome



Goals

- Learning to analyse NGS data from Reduce-Representation Libraries (RRL)
- Learning to use
  - Galaxy
  - The STACKS pipeline
- Learning
  - Raw Illumina RAD preparation
  - Use a reference genome
  - Assembly of RAD loci
  - Detection of SNPs, genotypes and haplotypes determination
  - Population genetics statistics

Datasets and tools

- Datasets used during Julian Catchen training sessions
- Stickleback dataset from *Hohenlohe et al. 2010*
- Data cleaning and analyses with Galaxy, the Stacks pipeline and BWA.
- All data produced with Illumina GAII or HiSeq2000.
- Open Source software

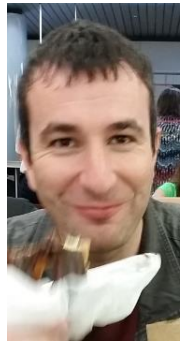
Merci de votre attention

Muséum national d'histoire naturelle
Station Marine de Concarneau

CESCO
Centre d'Écologie et des Sciences de la Conservation

yvan.le-bras@mnhn.fr

Merci de votre attention



RADseqGCC2016 page ***http://tinyurl.com/radseqgcc2016***

Toulouse Sigenae Galaxy server ***http://sigenae-workbench.toulouse.inra.fr/galaxy/***

GenOuest Galaxy instance ***http://galaxy.genouest.org***