Restriction-based methods for population genomics

Genome-wide genetic marker discovery and genotyping using next-generation sequencing

John W. Davey, Paul A. Hohenlohe‡, Paul D. Etter§, Jason Q. Boone||, Julian M. Catchen‡ and Mark L. Blaxter*¶*

Abstract | The advent of next-generation sequencing (NGS) has revolutionized genomic and transcriptomic approaches to biology. These new sequencing tools are also valuable for the discovery, validation and assessment of genetic markers in populations. Here we review and discuss best practices for several NGS methods for genome-wide genetic marker development and genotyping that use restriction enzyme digestion of target genomes to reduce the complexity of the target. These new methods — which include reduced-representation sequencing using reduced-representation libraries (RRLs) or complexity reduction of polymorphic sequences (CRoPS), restriction-site-associated DNA sequencing (RAD-seq) and low coverage genotyping — are applicable to both model organisms with high-quality reference genome sequences and, excitingly, to non-model species with no existing genomic data.

Nature Reviews Genetics, 2011

CRoPS

Complexity Reduction of Polymorphic Sequences
van Orsouw et al. 2007

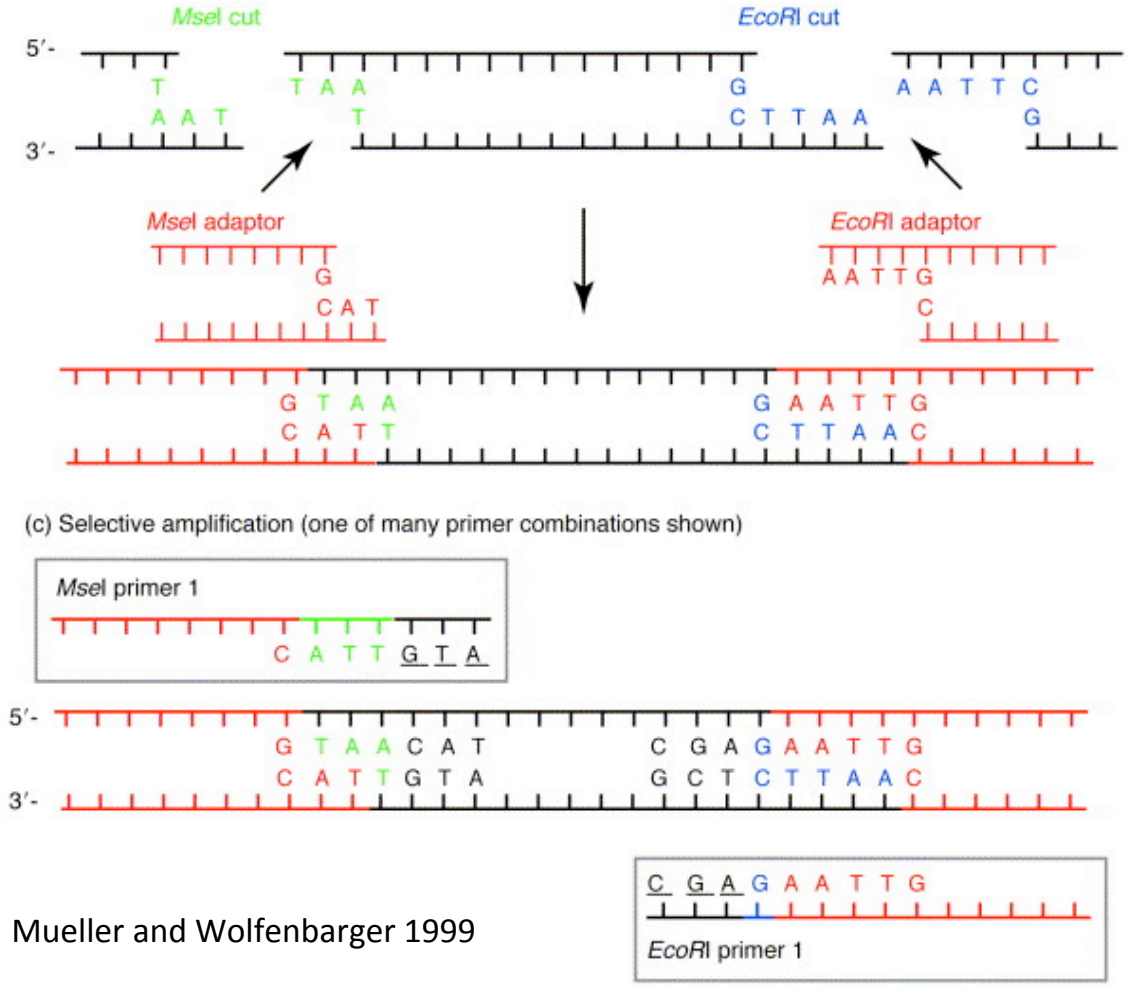
Pool samples

↓

Double digest and ligate

↓

+1 selective PCR

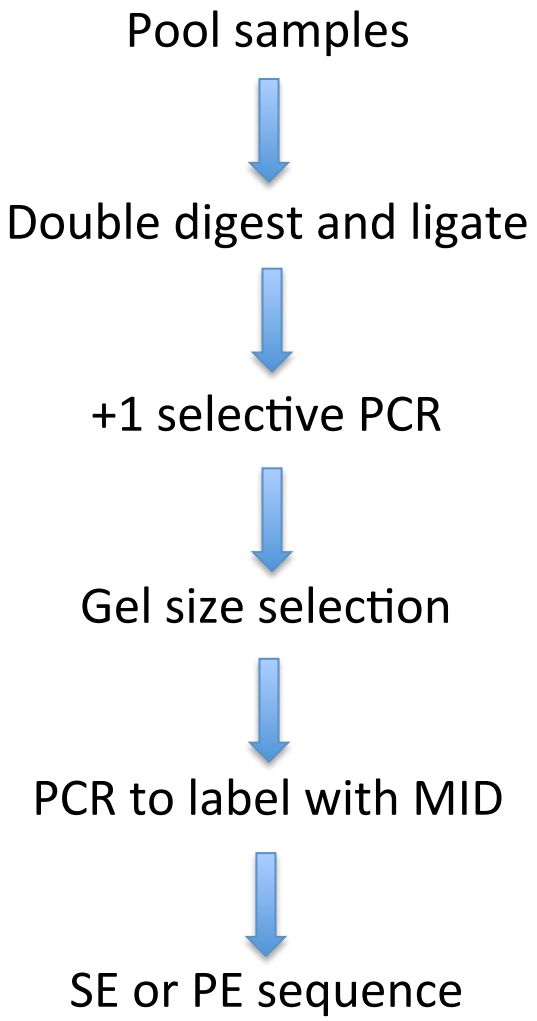


Mueller and Wolfenbarger 1999

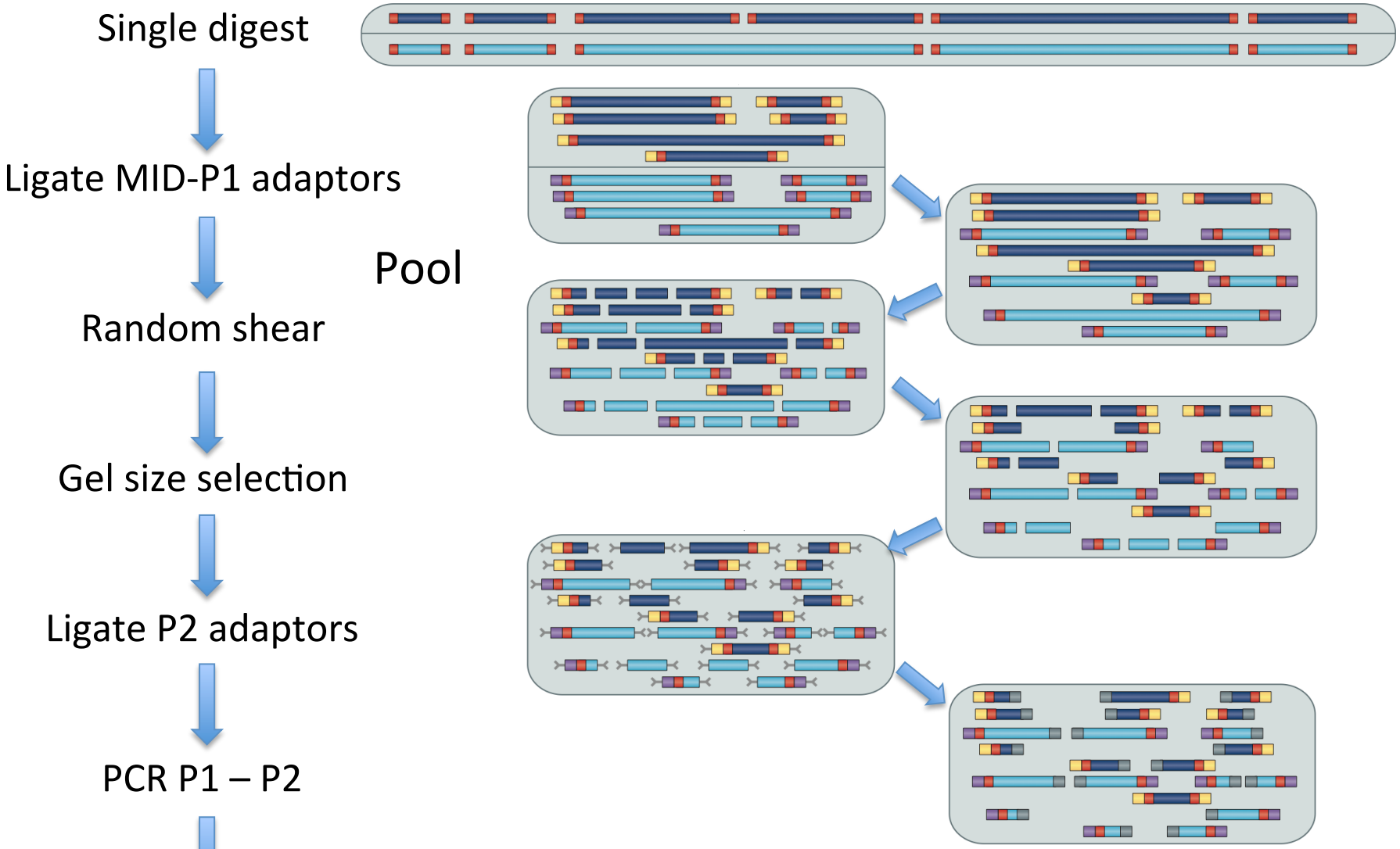
CRoPS

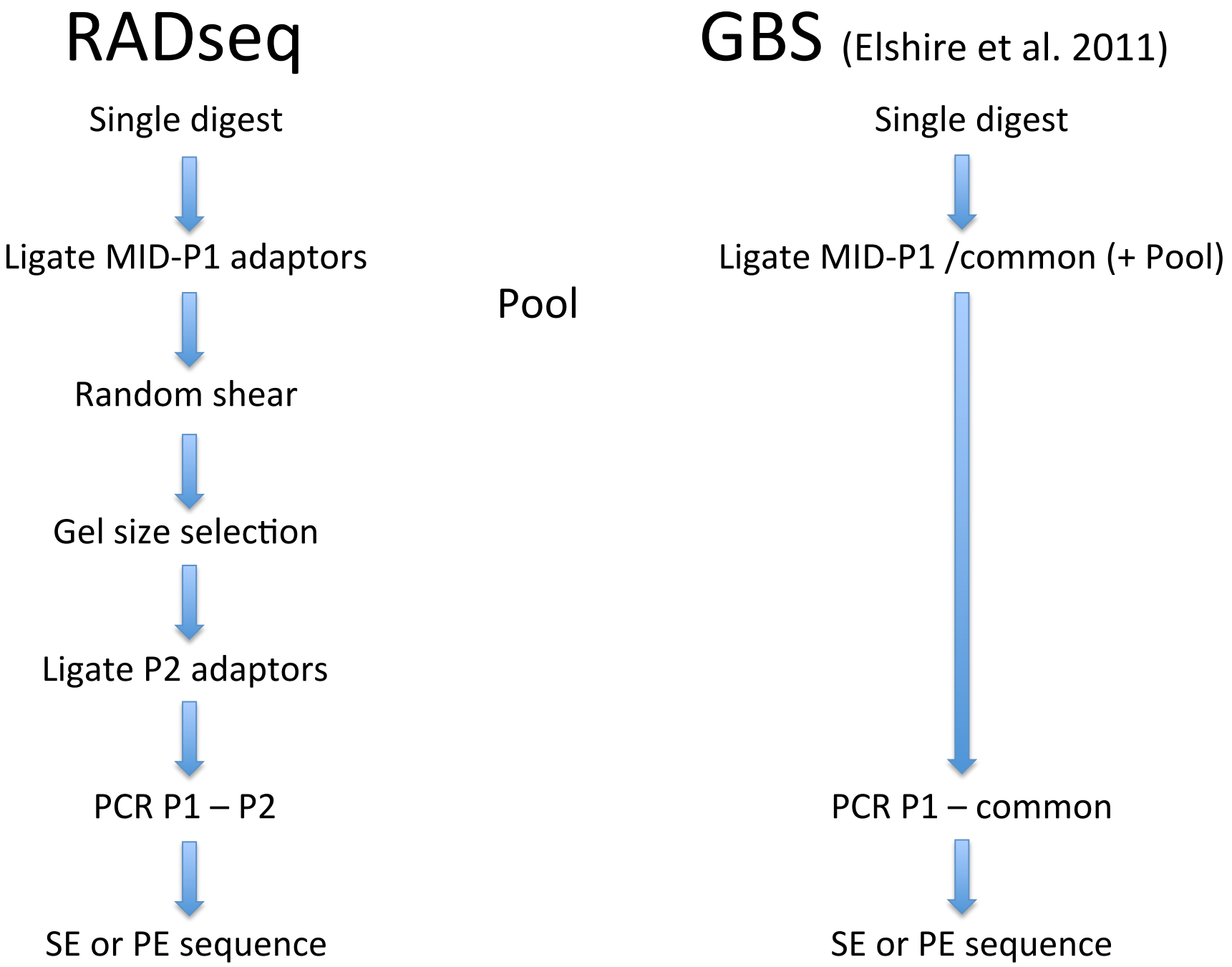
Complexity Reduction of Polymorphic Sequences
van Orsouw et al. 2007

- Frequent cutters (small frags)
- Often sequenced with 454
- Pros:
  - Easy, start with minimal DNA
  - Flexible extent of complexity reduction
  - Selective PCR could enrich for genomic regions with distinct GC content
- Cons:
  - Lots of PCR (error, biases)
  - Polymorphic restriction sites produce null alleles (allelic dropout)



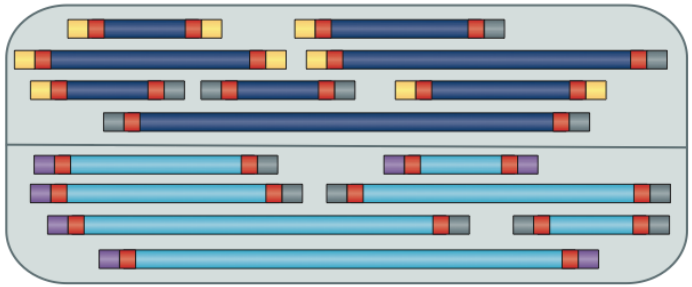
RADseq - Baird et al. 2008



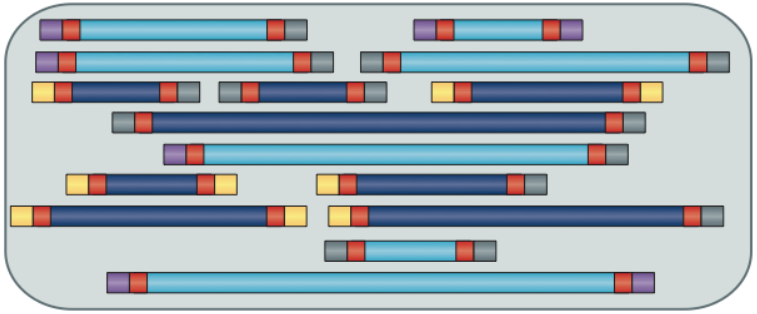

RADseq
Single digest
Ligate MID-P1 adaptors
Random shear
Gel size selection
Ligate P2 adaptors
PCR P1 – P2
SE or PE sequence
Pool
GBS (Elshire et al. 2011)
Single digest
Ligate MID-P1 /common (+ Pool)
PCR P1 – common
SE or PE sequence

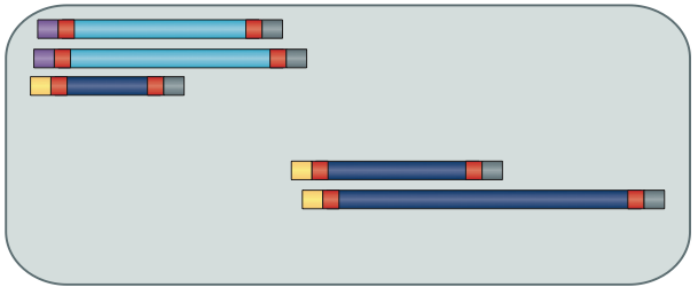
GBS (Elshire et al. 2011)

Single digest



Ligate MID-P1 /common (+ Pool)

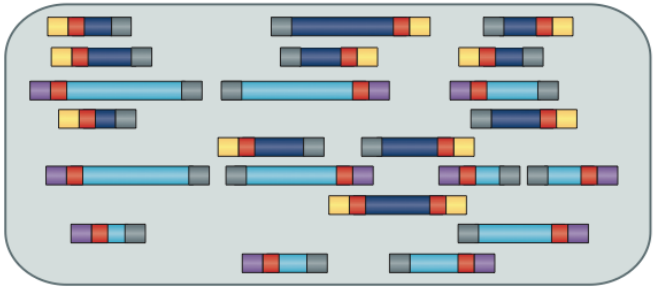


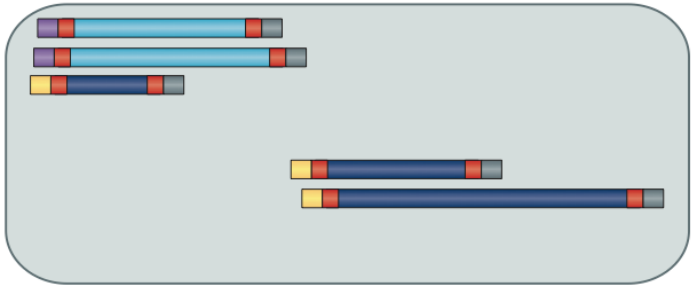


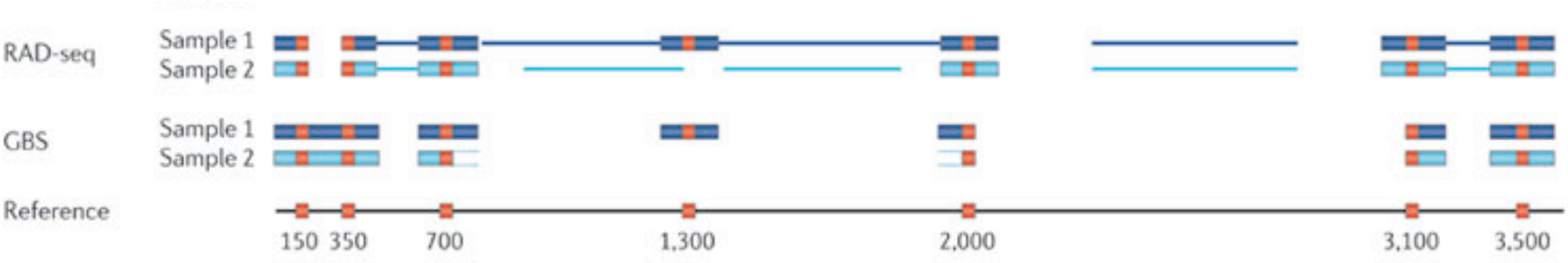
PCR P1 – common

SE or PE sequence

RADseq vs GBS sequencing tags







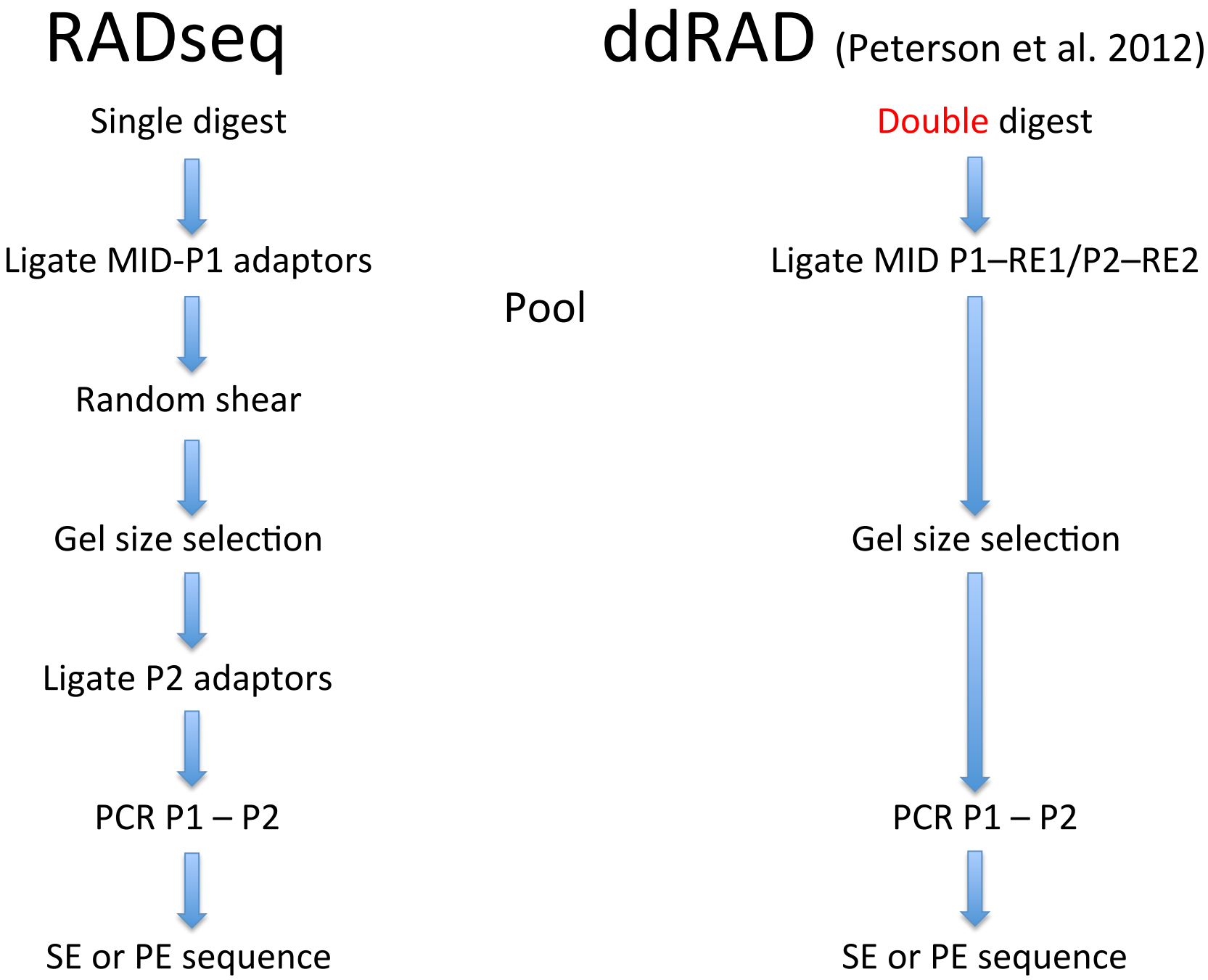
Davey et al. 2011

RADseq

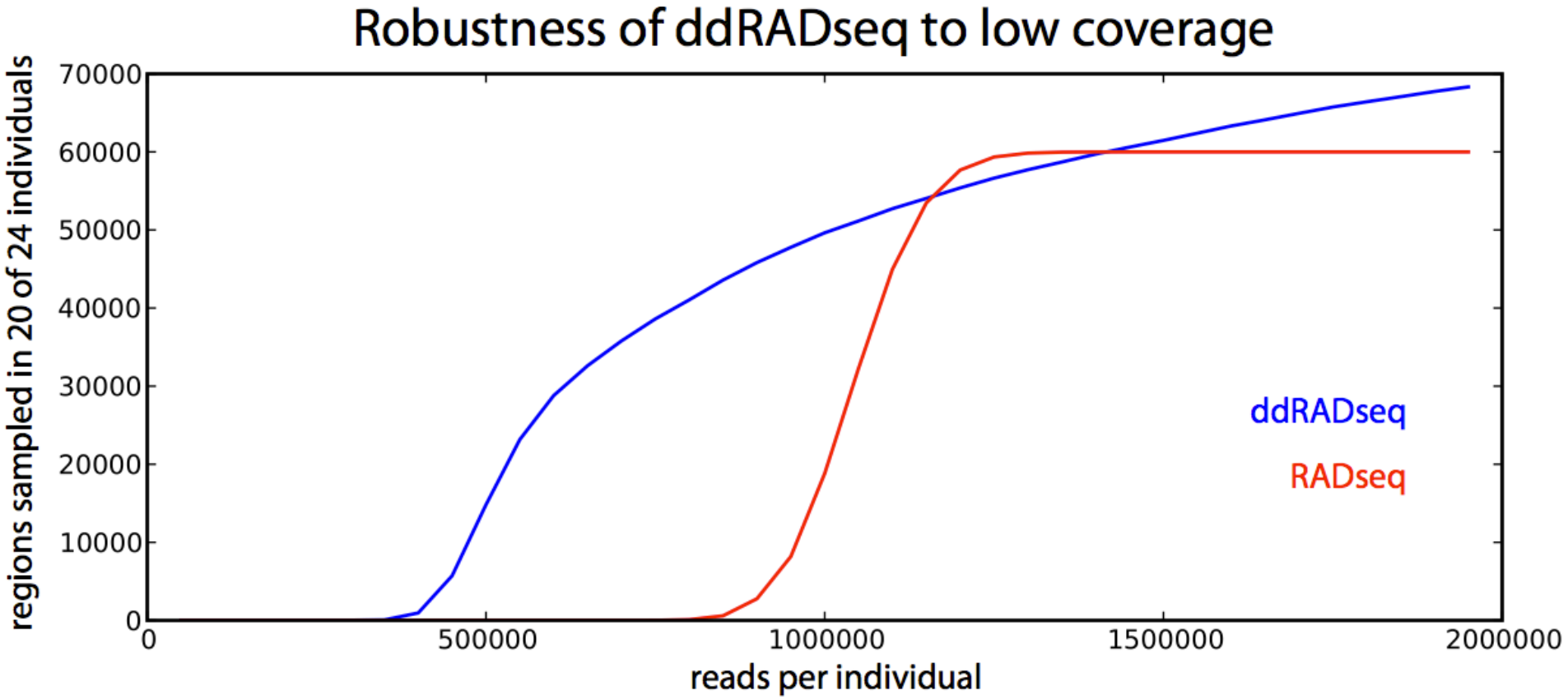
- Pros
  - Sequence almost every restriction fragment from every individual
- Cons
  - Lots of steps
  - Each restriction site has both flanks sequenced; 2- fold redundancy
  - Coverage biases from shearing (Davey et al. 2013 Mol. Ecol.)
  - Polymorphic restriction sites produce null alleles

GBS

- Pros
  - Much easier library prep
- Cons
  - Complexity reduction less controlled, potential for more missing data across individuals
  - Polymorphic restriction sites produce null alleles


RADseq
Single digest
Ligate MID-P1 adaptors
Random shear
Gel size selection
Ligate P2 adaptors
PCR P1 – P2
SE or PE sequence
Pool
ddRAD (Peterson et al. 2012)
Double digest
Ligate MID P1–RE1/P2–RE2
Gel size selection
PCR P1 – P2
SE or PE sequence

More Even Representation of Loci Across Individuals at Low Coverage



Peterson et al. 2012

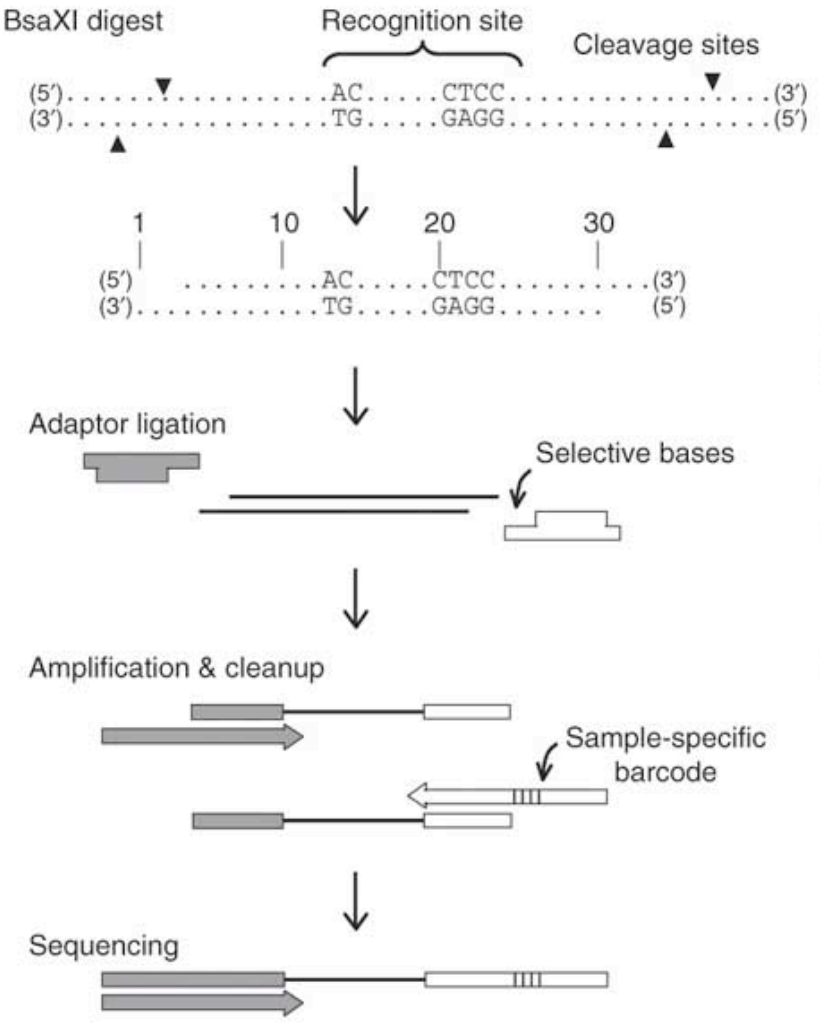
ddRAD

- Pros
  - Fine tuning of complexity reduction by enzyme choice and size selection window
  - At low coverage, greater uniformity across individuals (less missing data)
  - Fewer duplicate (adjacent) restriction site tags

- Cons – both scale with heterozygosity
  - Longer combined length of restriction site sequences exacerbates allele dropout
  - Allele dropout also produced by novel restriction sites, shortening fragment below minimum size

2b-RAD

- Type IIB restriction enzymes produce short uniform-length fragments
- Sequence all fragments or use selective amplification to subsample
- De-novo clustering is feasible
- Pros:
  - 4 h library prep
  - Sampling density easily adjusted
  - Cheaper 35 bp sequencing
- Cons:
  - Polymorphic restriction sites cause allelic dropout



Wang et al. 2012

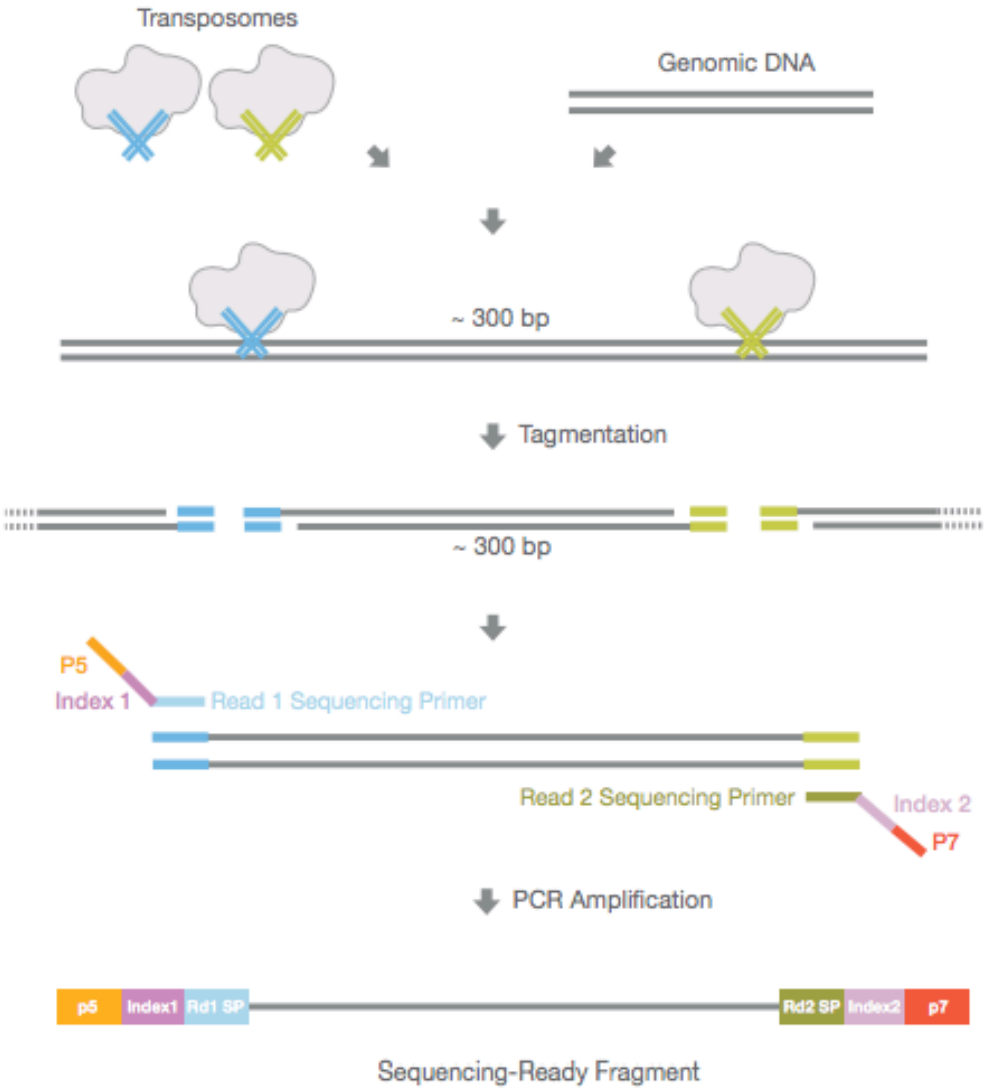
nextRAD (Russello et al. 2015)

"Nextera-tagmented reductively-amplified DNA"

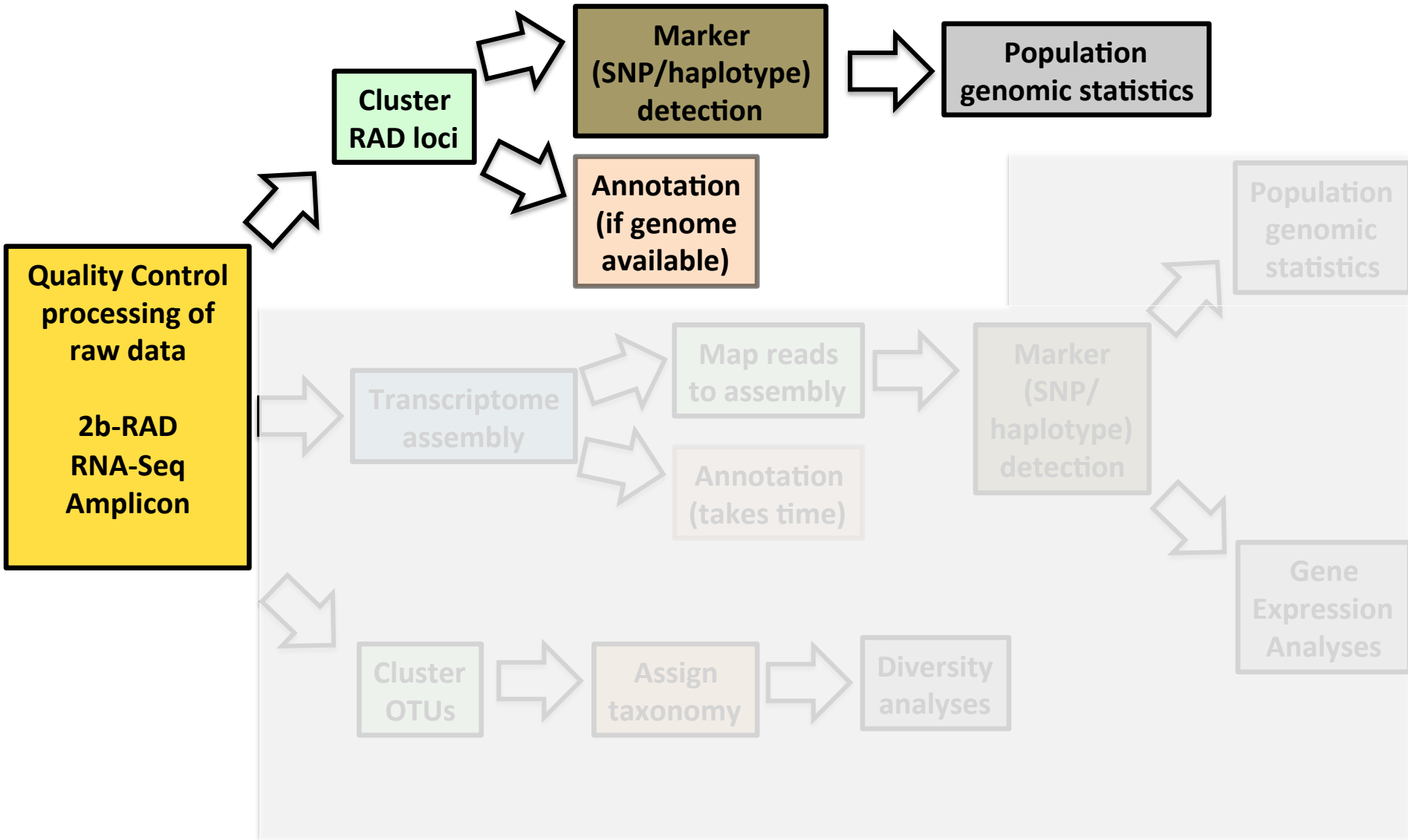
Engineered transposon simultaneously fragments and adds adapters.

Fragments can then be amplified Rd1-Rd2, adding barcoded index primers.

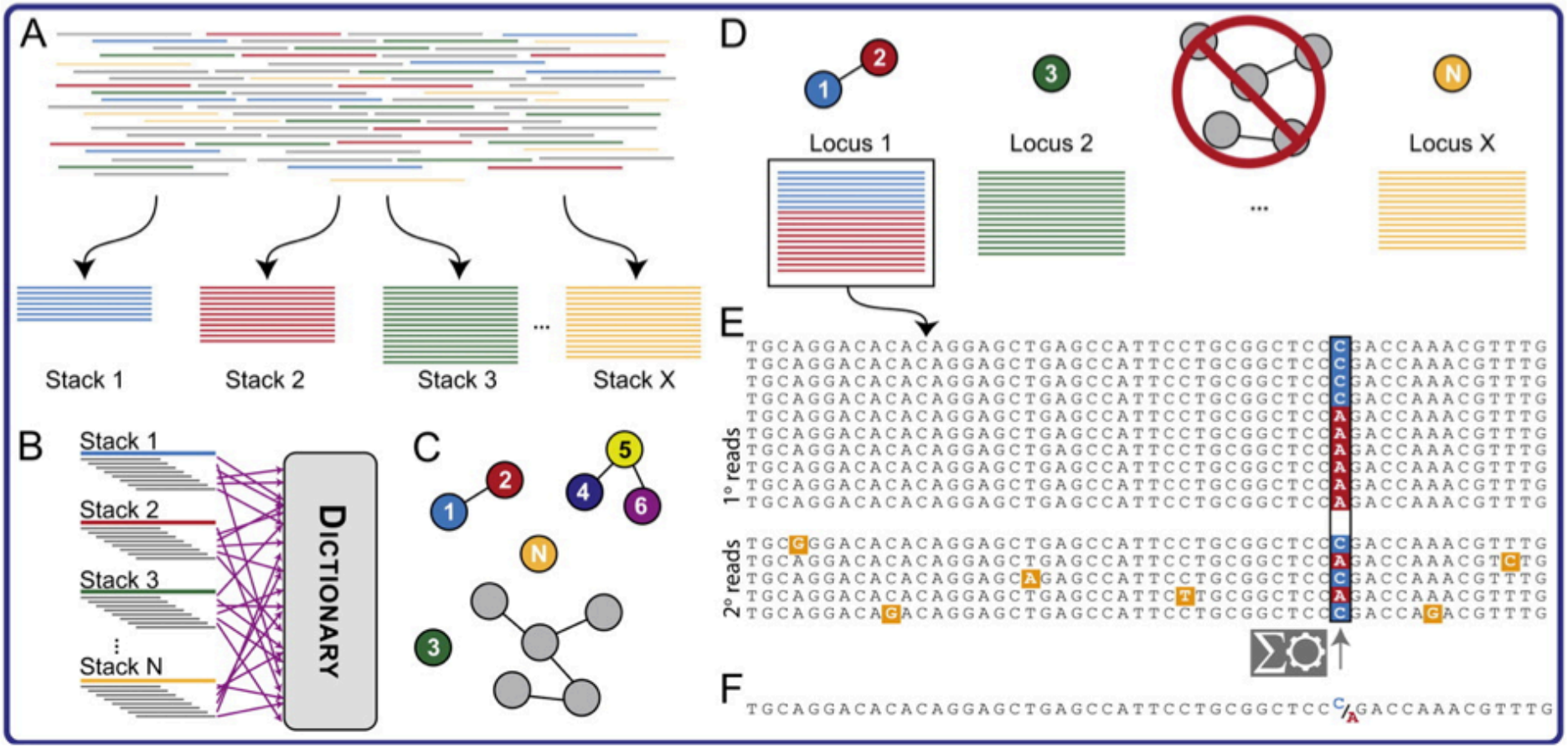
Generates fragments of about 300 bp length



Analysis pipeline for RAD-data



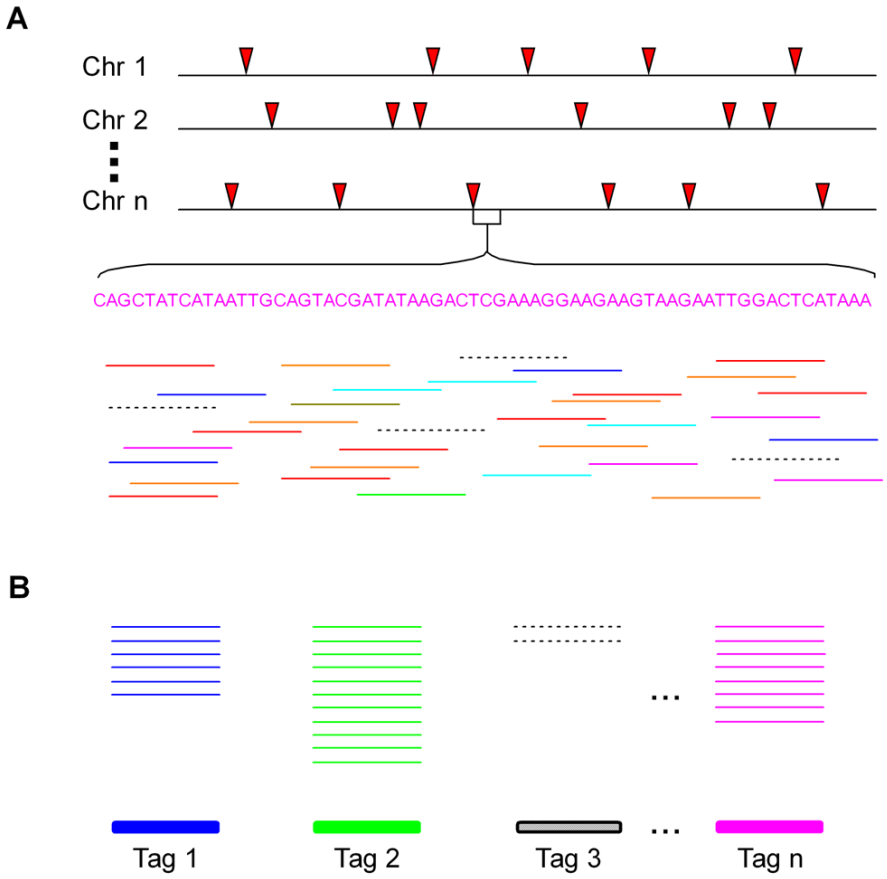
STACKS pipeline

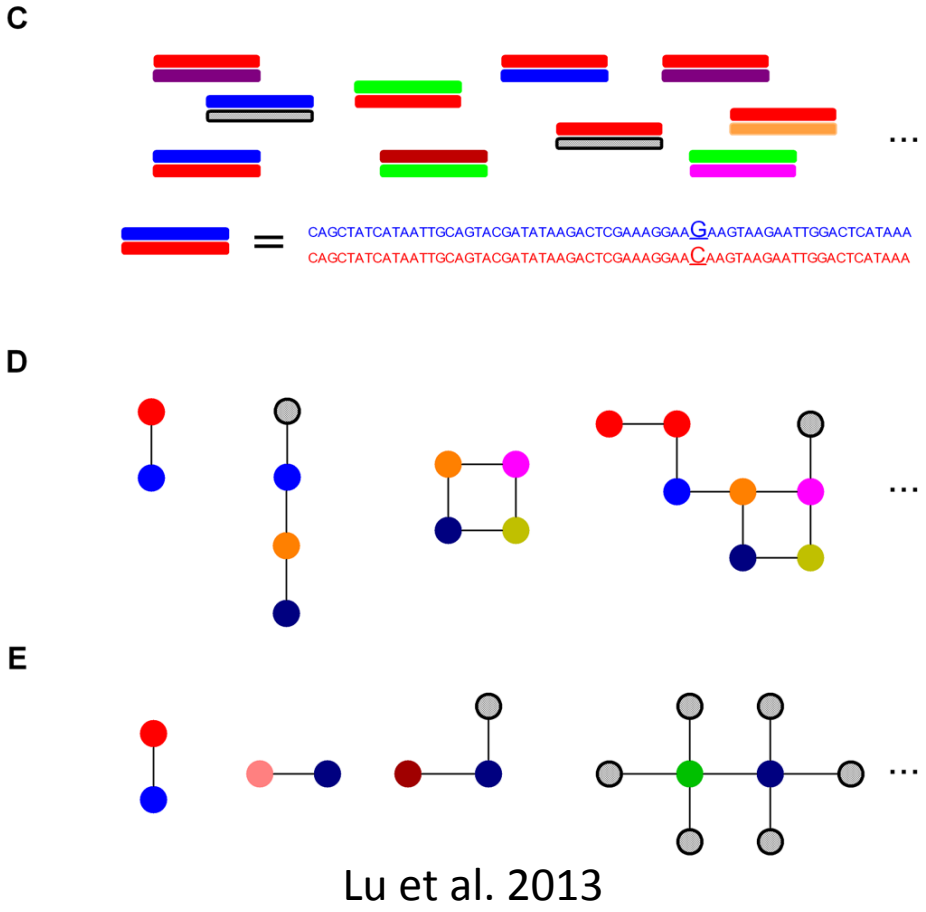


Catchen et al. 2011

UNEAK

Network clustering,
uses ploidy level to remove paralogs, repeats, sequencing errors



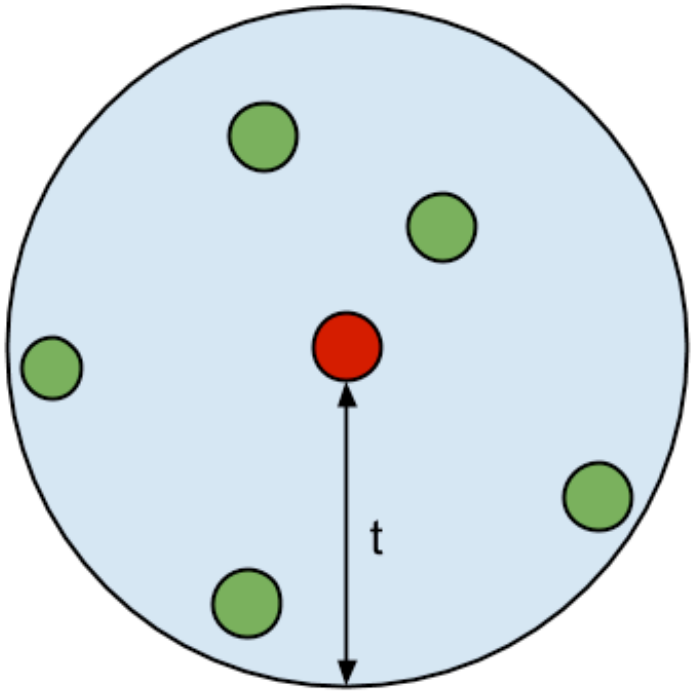


Lu et al. 2013

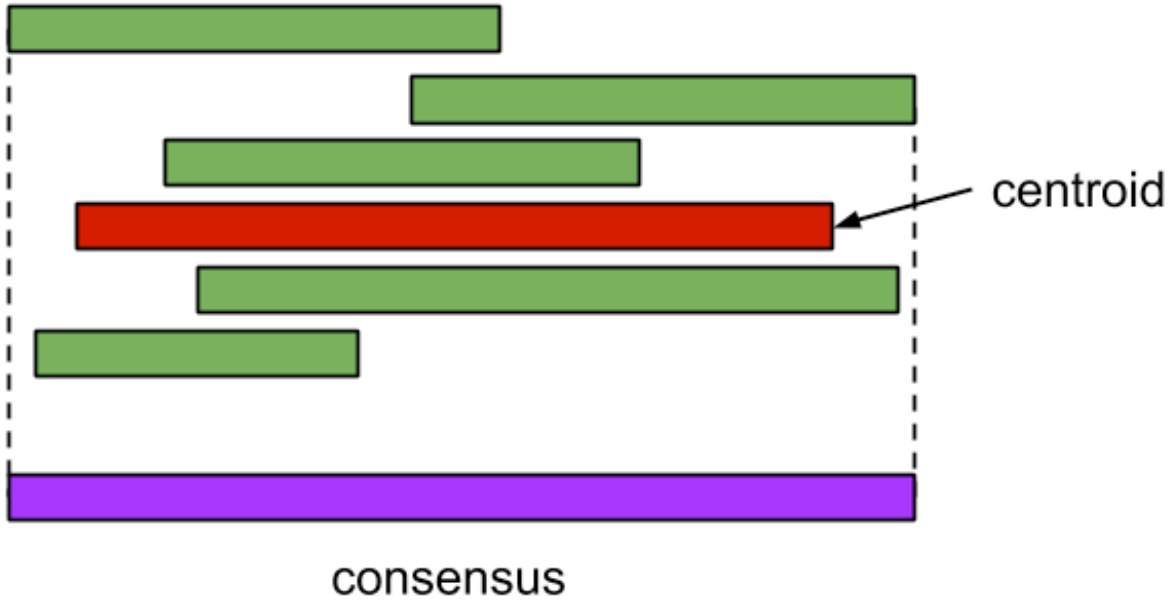
CD-HIT-EST

Uses very fast clustering algorithm, optimized for short sequences
Greedy, grabs everything within the specified identity threshold

This is what we will use today



t = identity threshold



Genotype filtering

- We will use custom-made scripts to call genotypes from our cluster database.
- We then need to filter this set based on depth, fraction of heterozygotes (to remove paralogs) and evenness.
- We also use technical replicates for filtering out potential contaminants or PCR artifacts.
- Finally, extract data for high-quality genotypes for downstream analyses.

Beyond the pipeline

What to do with the output data?

- Check proportion of loci out of HWE (estimation of allele dropout).
- Population structure (F_{ST}, PCA, STRUCTURE etc.)
- Scan for outlier loci
- Align to genome if available and annotate.

Exercise layout

- Each exercise has a Introductory section to read through, followed by a step-by-step protocol to follow.
- All commands to type in the Terminal window are in colored boxes.
- Commands can be copy-pasted or typed into the Terminal (make sure that they are on one line only before hitting enter!)
- Names of files are typically in `courier` font.
- In general, bash scripts (.sh) are to be opened and examined using nano (or another Text Editor)

Bash scripts are just a list of commands that could be typed directly into the Terminal

- Details of Python (.py) and Perl (.pl) scripts are beyond the scope of this class, but feel free to open them if you are interested.