Leveraging NGS for Population Studies: RAD-seq style

Genomic Tools Workshop

June 29, 2015

Tammy R Wilbert

Today's Talk

Why am I interested in this?

Restriction-site Associated DNA Sequencing

Going from raw data to analysis

Population Analysis Tools

Source-Sink Dynamics of a migratory bird: wood thrush (*Hylocichla mustelina*)

Project Design

- Three types of sampling sites
  - Big Oaks National Wildlife Refuge (4)
  - CRANE Naval Base (4)
  - Indiana Dept. Natural Resources (4)
- Intensive surveys for 4 years
- Data collection:
  - Demographics
  - Reproductive rates
  - Blood samples → Isotopes & Genetics

SERDP

Source-Sink Dynamics of a migratory bird: wood thrush (*Hylocichla mustelina*)

Connectivity

1. Genetic diversity of "residents"
2. Population structure
3. Gene flow & bottlenecks
4. Compare between sites

Source-Sink Dynamics of a migratory bird: wood thrush (*Hylocichla mustelina*)

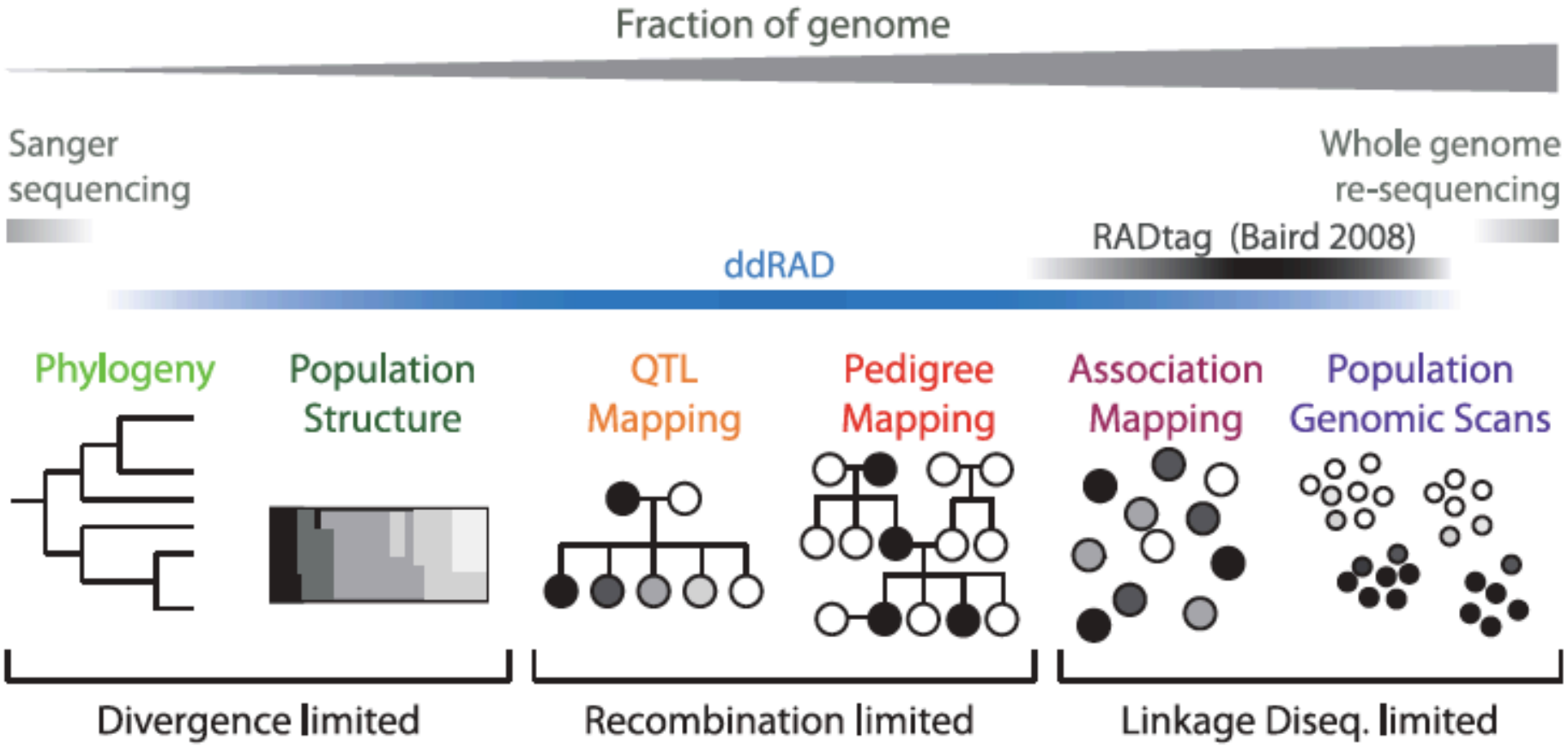
Connectivity

1. Genetic diversity of "residents"
2. Population structure
3. Gene flow & bottlenecks
4. Compare between sites

Dispersal

1. Genetic diversity of Second Year birds
2. Genetic assignment to natal site
3. Identify dispersal events
4. Compare between sites

Restriction-site Associated DNA Sequencing

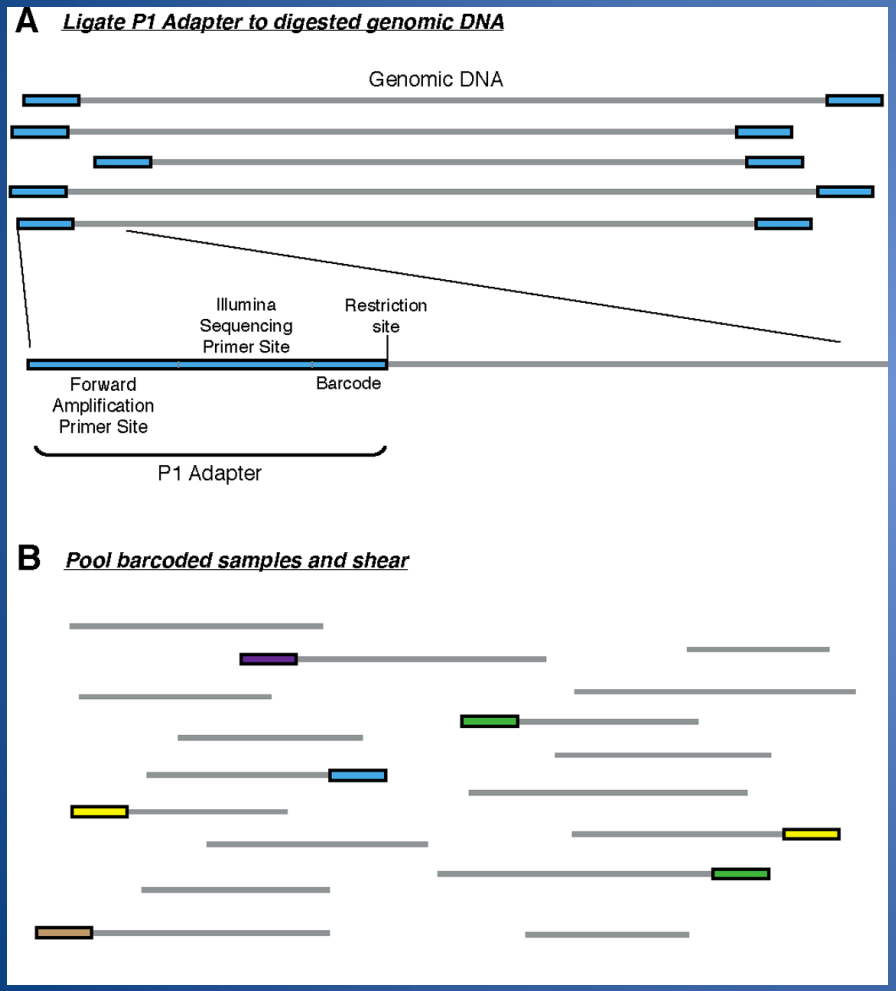


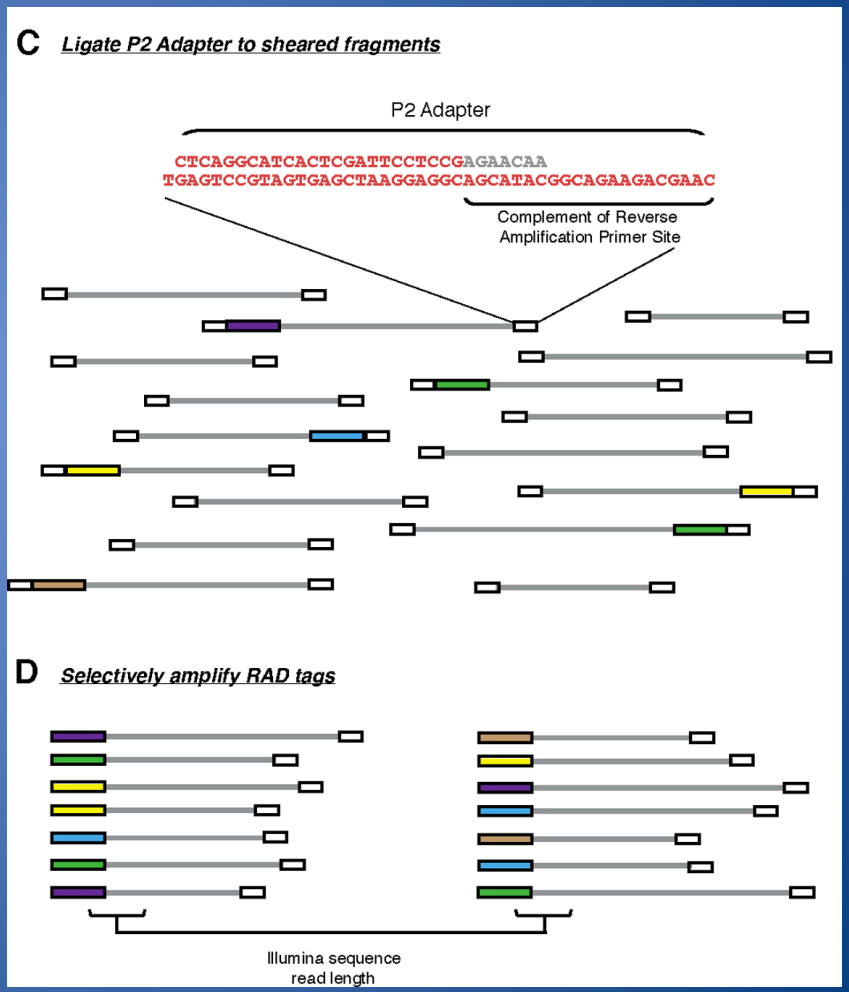
Peterson *et al.* 2012

Restriction-site Associated DNA Sequencing

- RAD-seq (Miller *et al.* 2007, Baird *et al.* 2008)
- ddRAD (Peterson *et al.* 2012)
- GBS (Elshire *et al.* 2011)
- 2bRAD (Wang *et al.* 2012)
- ezRAD (Toonen *et al.* 2013)

Restriction-site Associated DNA Sequencing

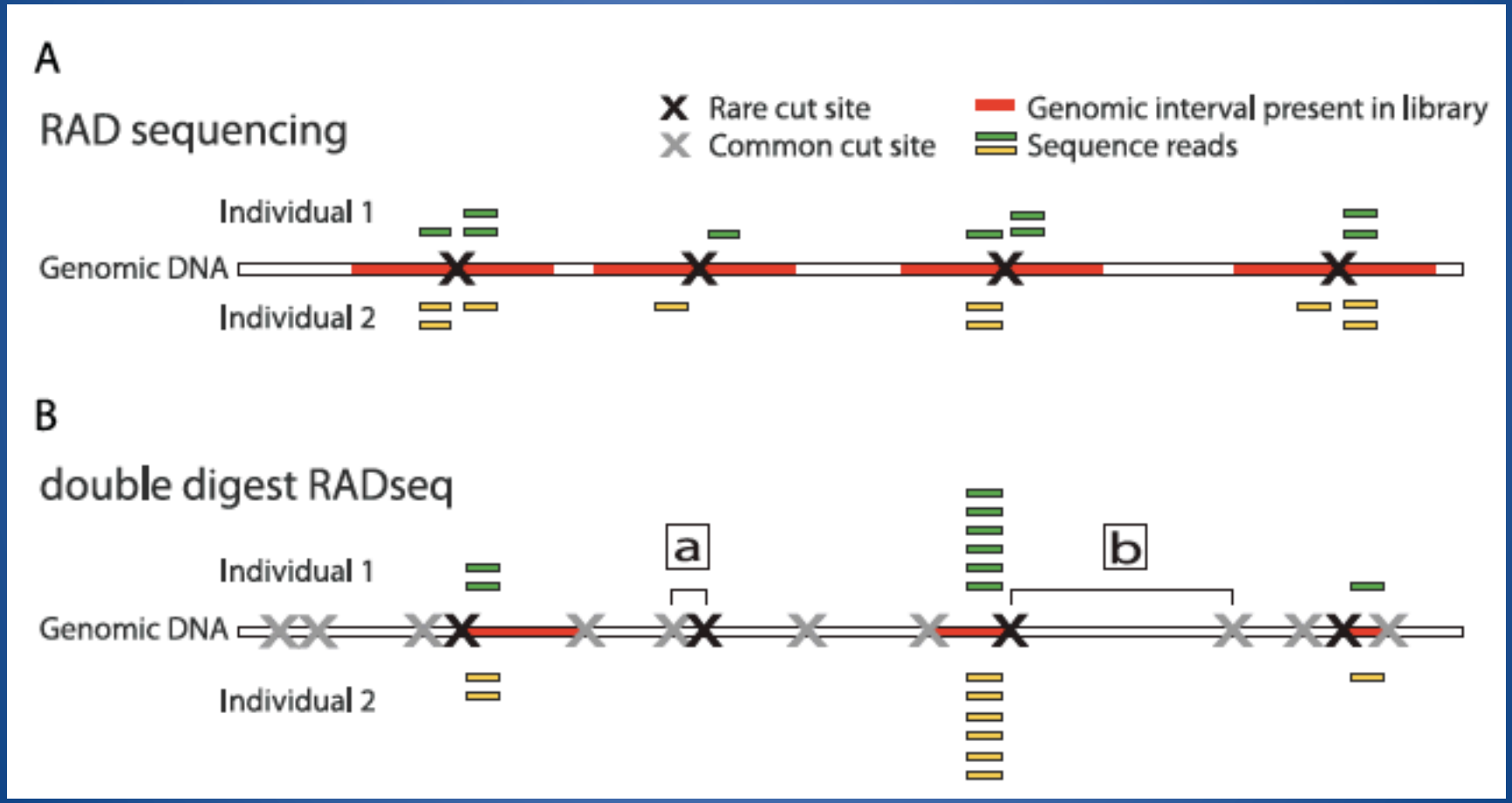




Selective PCR amplification of fragments with both P1 & P2 adapters

Baird *et al.* 2008

ddRAD: double digest RAD



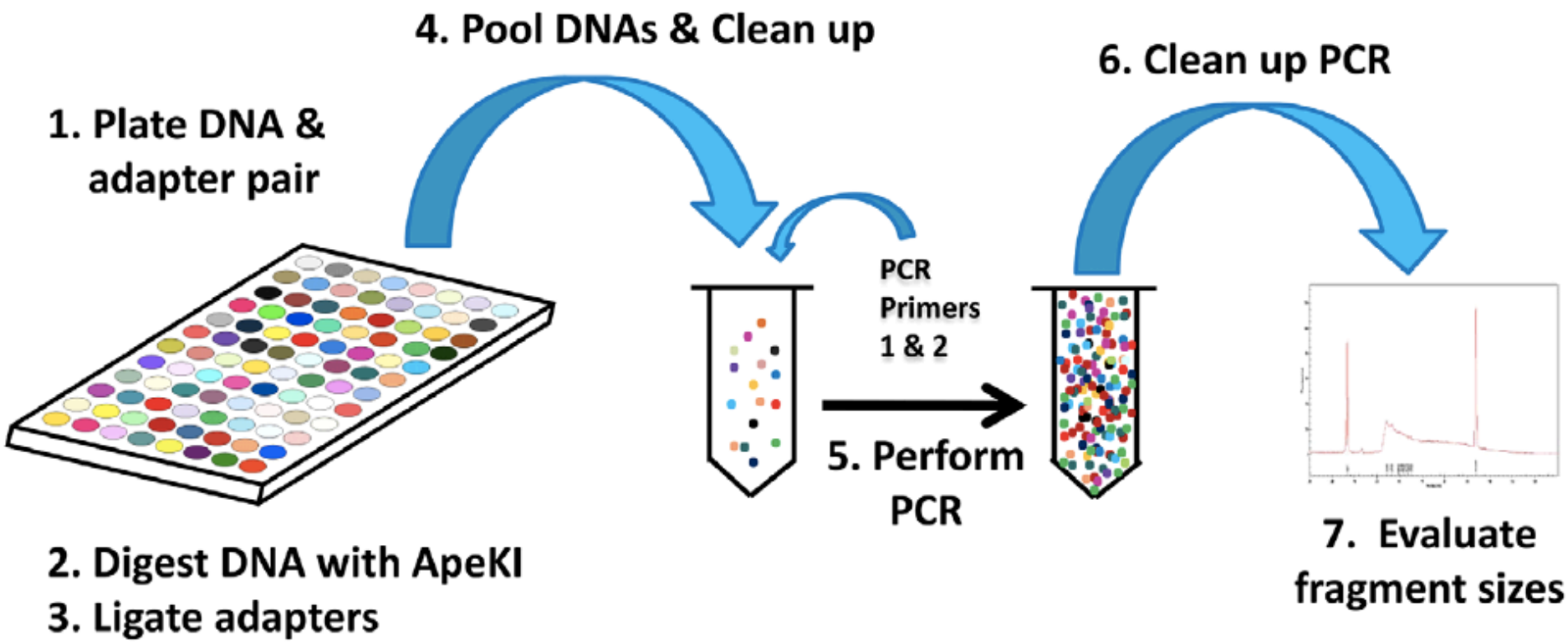
P1 Adapter = restriction site 1
P2 Adapter = restriction site 2

Pooled & Size selection
PCR enrichment

Peterson *et al.* 2012

GBS: Genotype by Sequencing

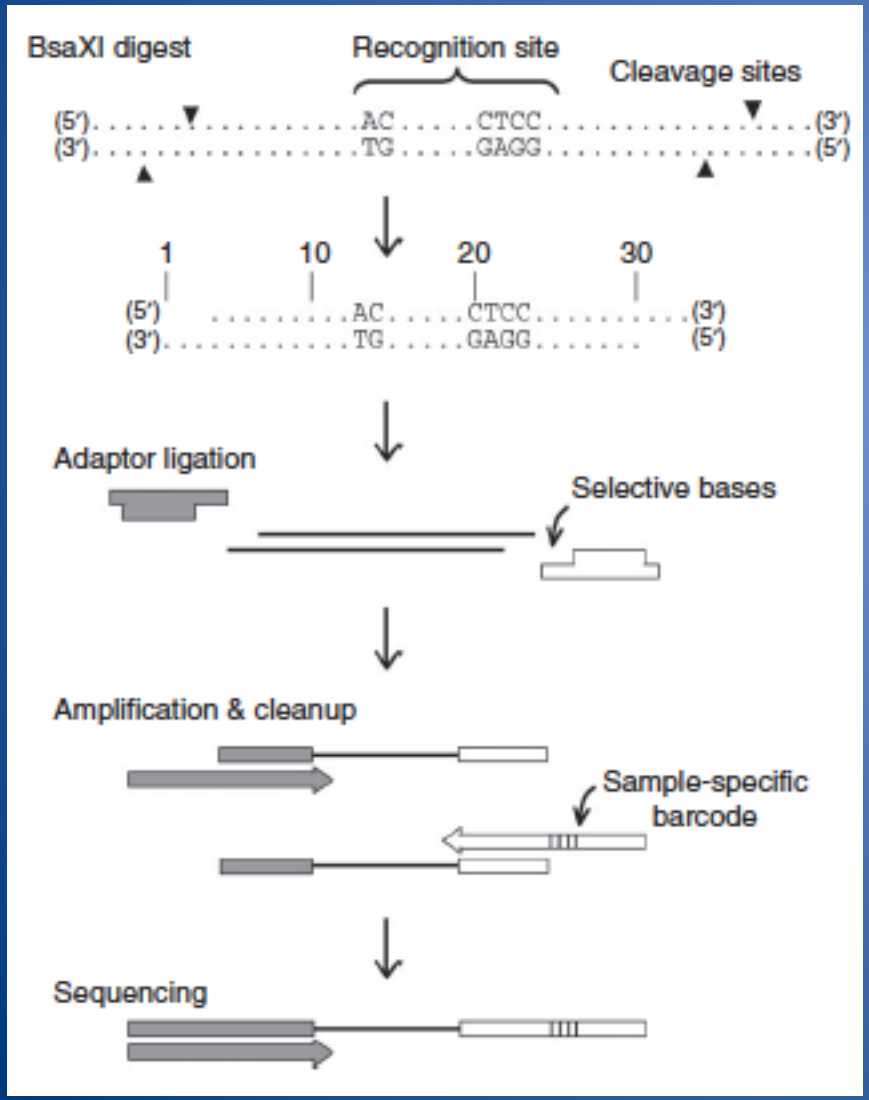
http://www.biotech.cornell.edu/brc/genomic-diversity-facility



Adapters → specific to overhang produced by enzyme digestion
→ include barcode and common adapter

Elshire *et al.* 2011

2bRAD - type IIB restriction endonucleases



Enzymes cleave genomic regions upstream & downstream of recognition site (examples: BsaXI, AlfI)

Adaptors with degenerate cohesive ends
5’-NNN-3’

Modify to sub-select fragments with more specific overhangs
5’-NNG-3’

PCR amplification with barcoding

Wang *et al.* 2012

ezRAD: Isoschizomer Enzymes

Cleave genomic regions with 2 high-frequency isoschizomer enzymes
example: MboI and Sau3AI both cut /GATC

Clean reaction

TruSeq DNA Kit:
- End repair
- Add A to 3' end
- Ligate TruSeq adapters
- Size Selection

Toonen *et al.* 2013

Pros/Cons of RAD-seq methods

Method	Pros	Cons
RAD	Shearing allows identification of PCR duplicates & creation of longer contigs	Technically challenging, more equipment, Sequencing depth biased to fragment length
ddRAD	Customization – number SNPs or fragment sizes by enzymes & size selection	Allelic dropout from size selection, need high-quality DNA
GBS	Simple protocol, Cost-effective, Customization	Optimization can be difficult, ADO, need high-quality DNA
2bRAD	Simple protocol, Cost-effective, No fragment size biases	SHORT sequences may be hard to map to genome, cannot build large contigs
ezRAD	Easy, Illumina support, with Illumina PCR-free TruSeq kit -> no PCR bias	ADO, Expensive, Sequencing can fail or create errors from starting bases

Puritz *et al.* 2014

Comparing Cost & Scalability

Table 3 **Comparison of most commonly used RAD sequencing methodologies and associated costs.**

	No. of enzymes	Cut frequency	Shearing required	Size selection	Library prep time & required expertise	Initial outlay cost	Subsequent library cost per sample	Scalability to reduce overall cost per sample
ezRAD	1 or more	Frequent	No	Yes	Low	Very Low	Moderate	Low
RAD tags	1	Rare	Yes	Yes	High	High	Low	Low
GBS	1	Rare or frequent	No	No	Moderate	High	Moderate to very low	Low
2-enzyme GBS	2	Rare + frequent	No	No	Moderate	High	Moderate to very low	Low
ddRAD	2	Frequent	No	Yes	Moderate	High	Very low	Moderate
2b-RAD	1	Frequent	No	No	Moderate	High	Low	Moderate

Library preparation of samples can range from $5 – $60 per sample

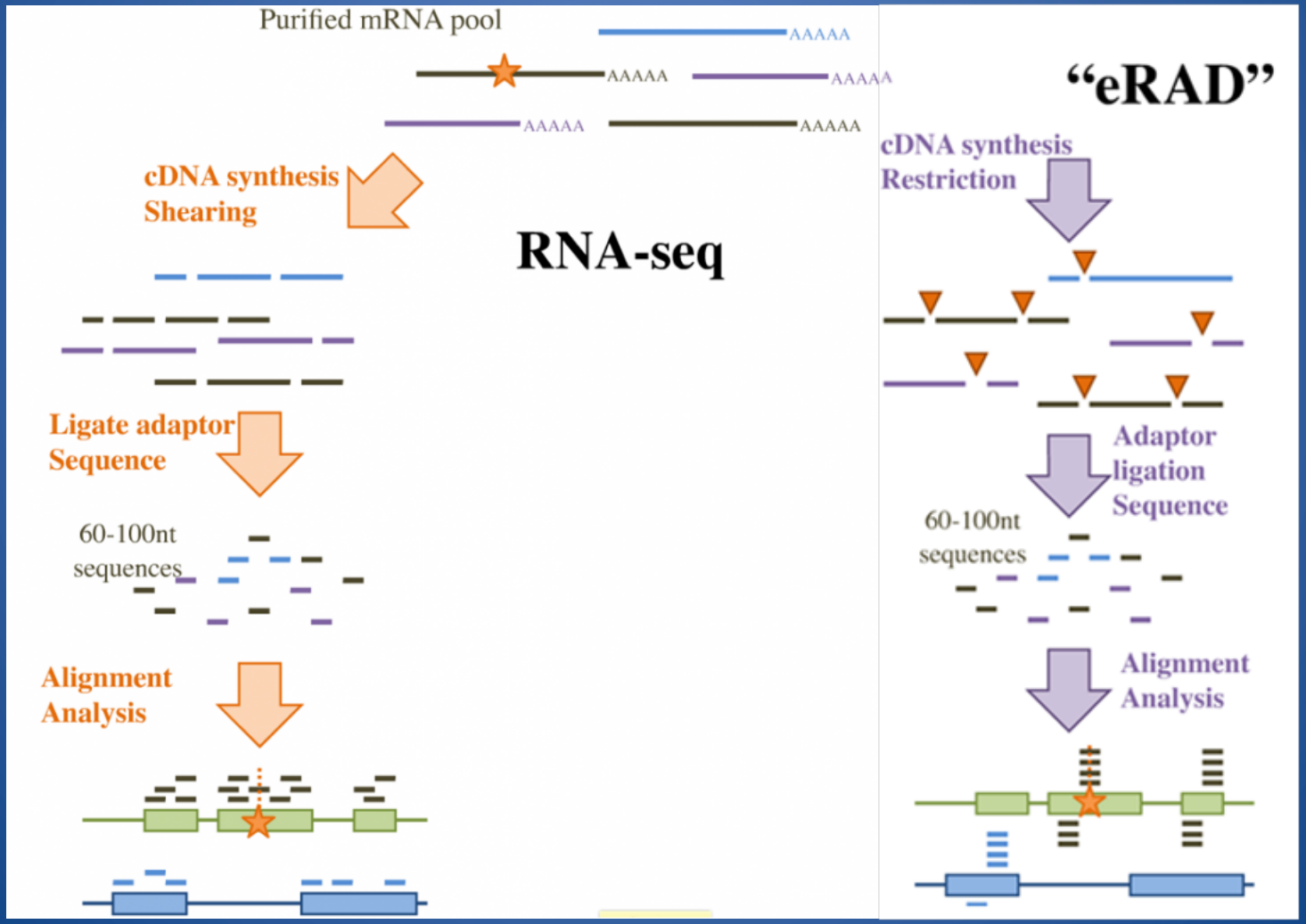
Illumina Sequencing costs:

MiSeq	$900 in house	$1300 at facility	10 million reads
HiSeq	$2000?	$2500 at facility	100 million reads

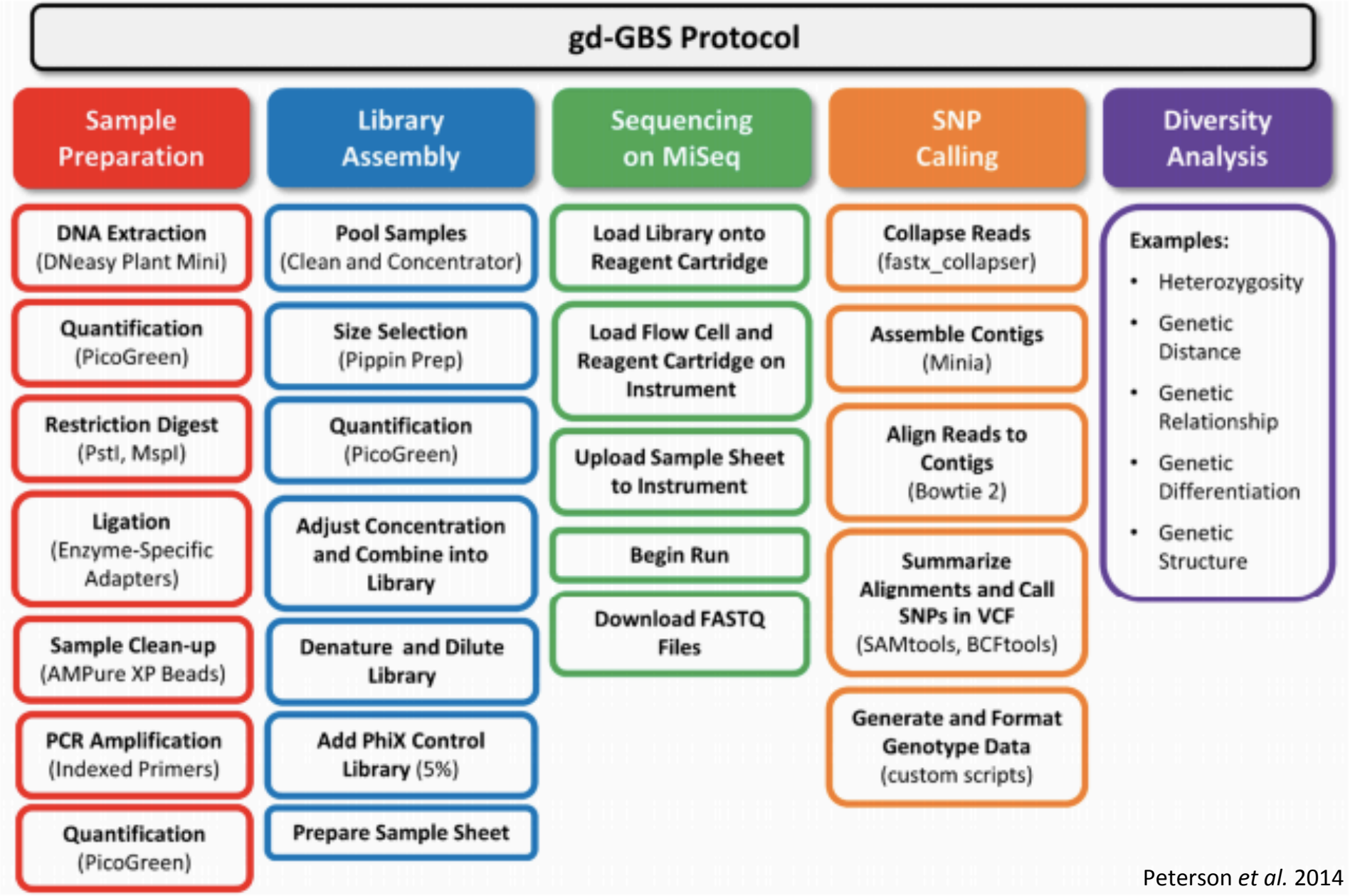
SEQUENCING DEPTH IS CRUCIAL!!

Toonen *et al.* 2013

One step further: RNA-seq and eRAD

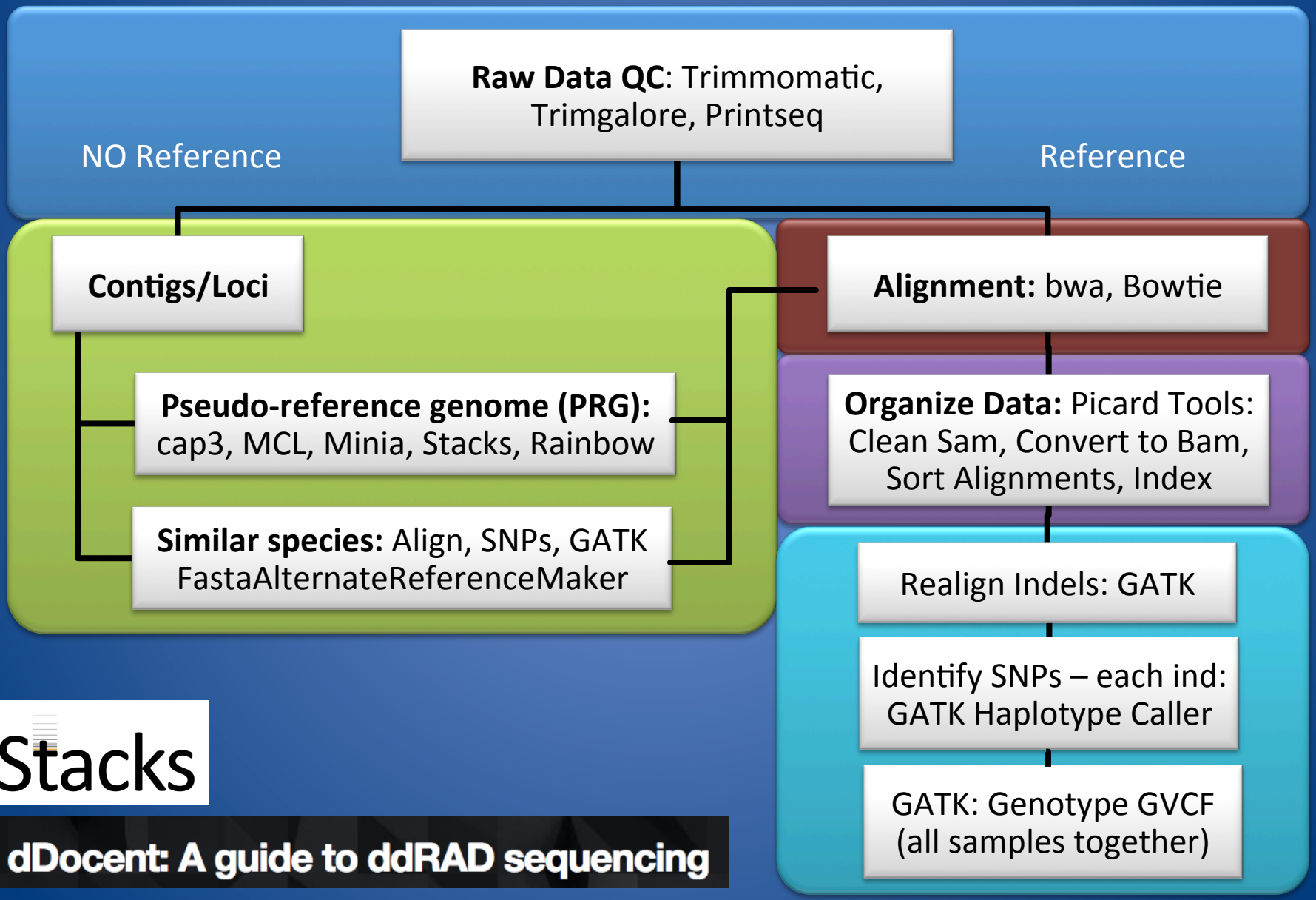


Genetic-diversity GBS Protocol

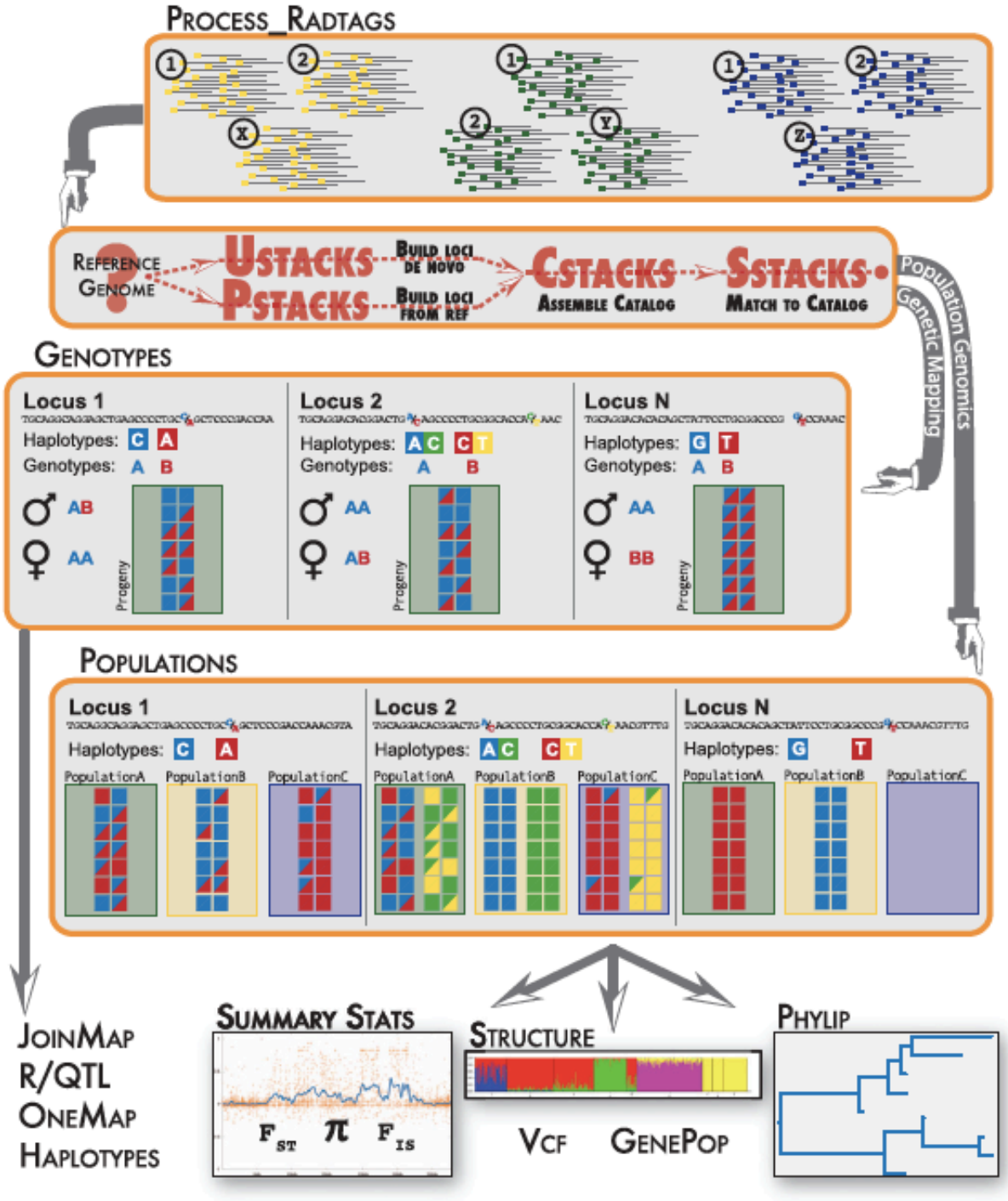


Peterson *et al.* 2014

Going from raw data to analysis



Stacks



Stacks

Julian Catchen

1. Process_radtags
2. Ustacks – de novo loci

   Pstacks – alignment
3. Cstacks
4. Sstacks
5. Individual/Population

   files for next analyses

- perl scripts
- mySQL database

dDocent: A guide to ddRAD sequencing

FreeBayes	https://github.com/ekg/freebayes
STACKS	http://creskolab.uoregon.edu/stacks/
PEAR	http://sco.h-its.org/exelixis/web/software/pear/
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
Mawk	http://invisible-island.net/mawk/
BWA	http://bio-bwa.sourceforge.net
SAMtools	http://samtools.sourceforge.net
VCFtools v.1.11**	http://vcftools.sourceforge.net/index.html
Rainbow	http://sourceforge.net/projects/bio-rainbow/files/
seqtk	https://github.com/lh3/seqtk
CD-HIT	http://weizhong-lab.ucsd.edu/cd-hit/
gnu-parallel	http://www.gnu.org/software/parallel/
bedtools	https://code.google.com/p/bedtools/
vcflib	https://github.com/ekg/vcflib
gnuplot	http://www.gnuplot.info/

Going from raw data to analysis

VCF for all samples (from GATK)

FILTER SNPs: MAF, position, quality, depth

FILTER INDIVIDUALS: % missing, Mendelian inheritance

Add info: age, location, habitat

→ Formatting

PGDSpider

Heidi Lischer

GREAT MANUAL!!

plink...

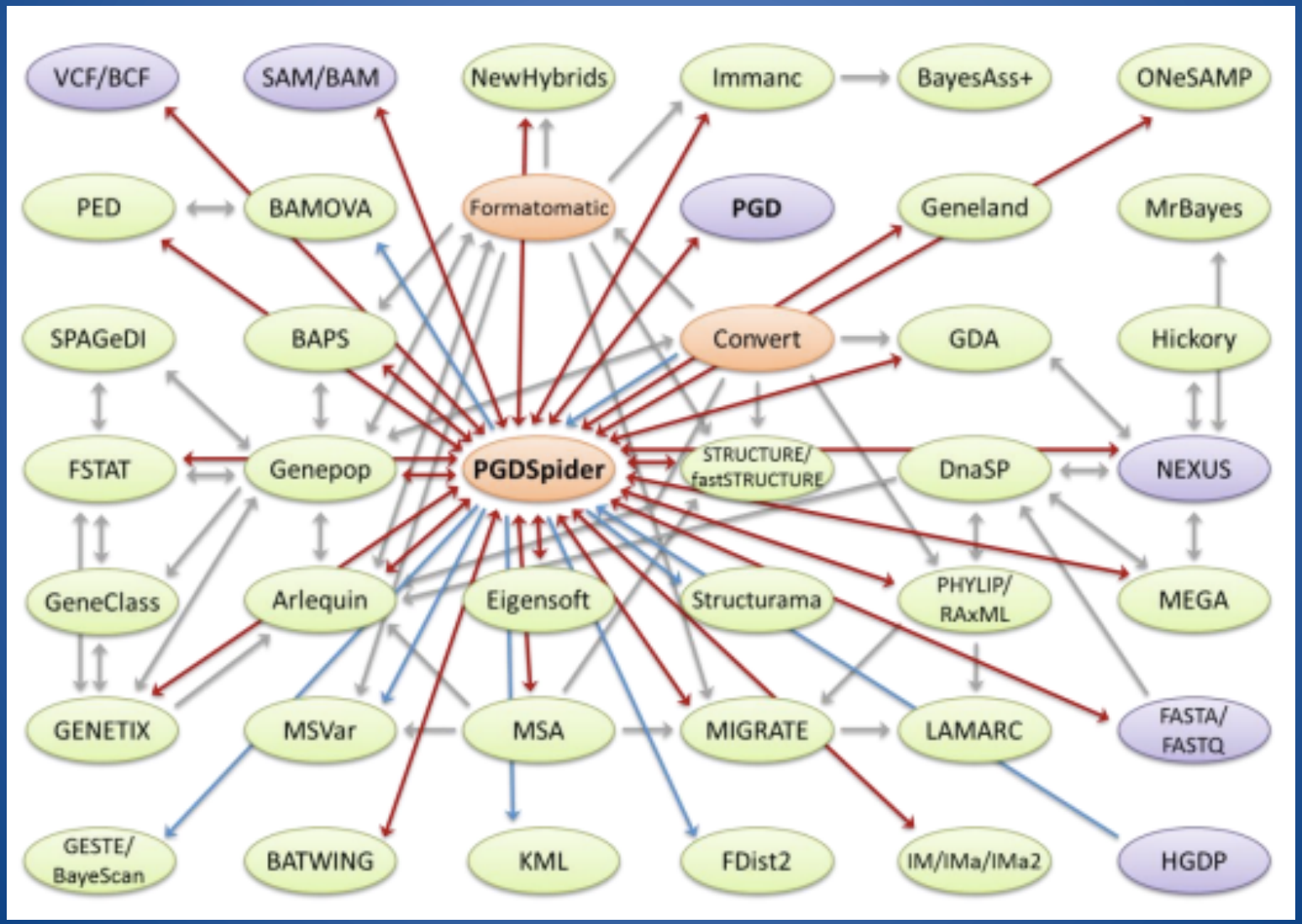
Shaun Purcell

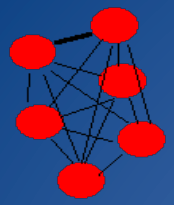
ANGSD

Thorfinn Sand Korneliussen
Anders Albrechtsen
Rasmus Nielsen

PGDSpider

Heidi Lischer





ANGSD

Korneliussen
Albrechtsen
Nielsen

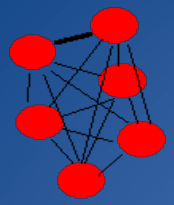
- Summary statistics (contamination, alleles, depth, quality)
- SNPs and genotypes (likelihoods, errors, allele freq)
- Population genetics (Site frequency spectrum, Theta, HWE)
- Population structure (admixture, Fst, PCA, ABBABABA, Relatedness)
- Medical genetics – association tests
- Outputs for Beagle & Plink

→ Input files: BAM/CRAM

Genotype likelihood file

Beagle files

VCF files



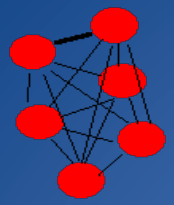
plink...

Shaun Purcell

- Data Management (SNPs, individuals, subsets, filtering)
- Summary Statistics (HWE, allele freq, linkage disequilibrium)
- Population stratification (clustering, IBS, IBD, inbreeding)
- Association & statistics (Heterogeneity, QTL, modeling)
- Family-based Association
- Permutation tests
- Multi-marker tests (Haplotype frequencies, association)
- Simulations
- R plugin

… and more

→ Merlin file – program KING: multi-dimensional scaling with IBS



plink...

Shaun Purcell

Chromosome	SNP ID	Genetic Distance	Base Pair Position	Allele 1	Allele 2

.map file

.bim file

Family ID	Ind ID	Pat ID	Mat ID	Sex	Phenotype	Allele 1	Allele 2

.fam file

.bed file

.ped file

http://pngu.mgh.harvard.edu/~purcell/plink2/formats.html

Questions?