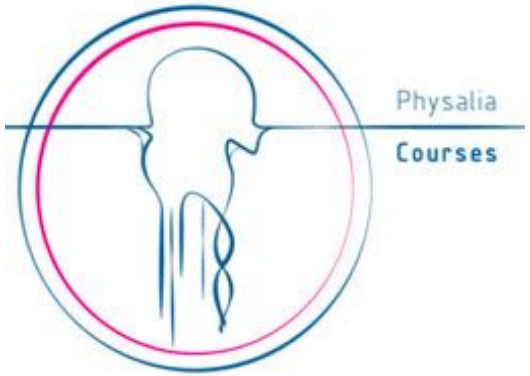


Population genomics for adaptation

Day 1 - Lecture 2

Analytical approaches

GWAS

Comparative genomics

Transcriptomics

Experimental evolution

QTL mapping

Epigenetics

Population genomics

Analytical approaches

GWAS

Comparative genomics

Transcriptomics

Experimental evolution

QTL mapping

Epigenetics

Population genomics

Population genomics

Population genetics studies the genetic differences within and between populations and the dynamics of how populations evolve.

Population genomics

Population genetics studies the genetic differences within and between populations and the dynamics of how populations evolve.

Population genomics studies these genetic differences using many markers to get a better sense of how evolutionary forces shape different parts of the genome.

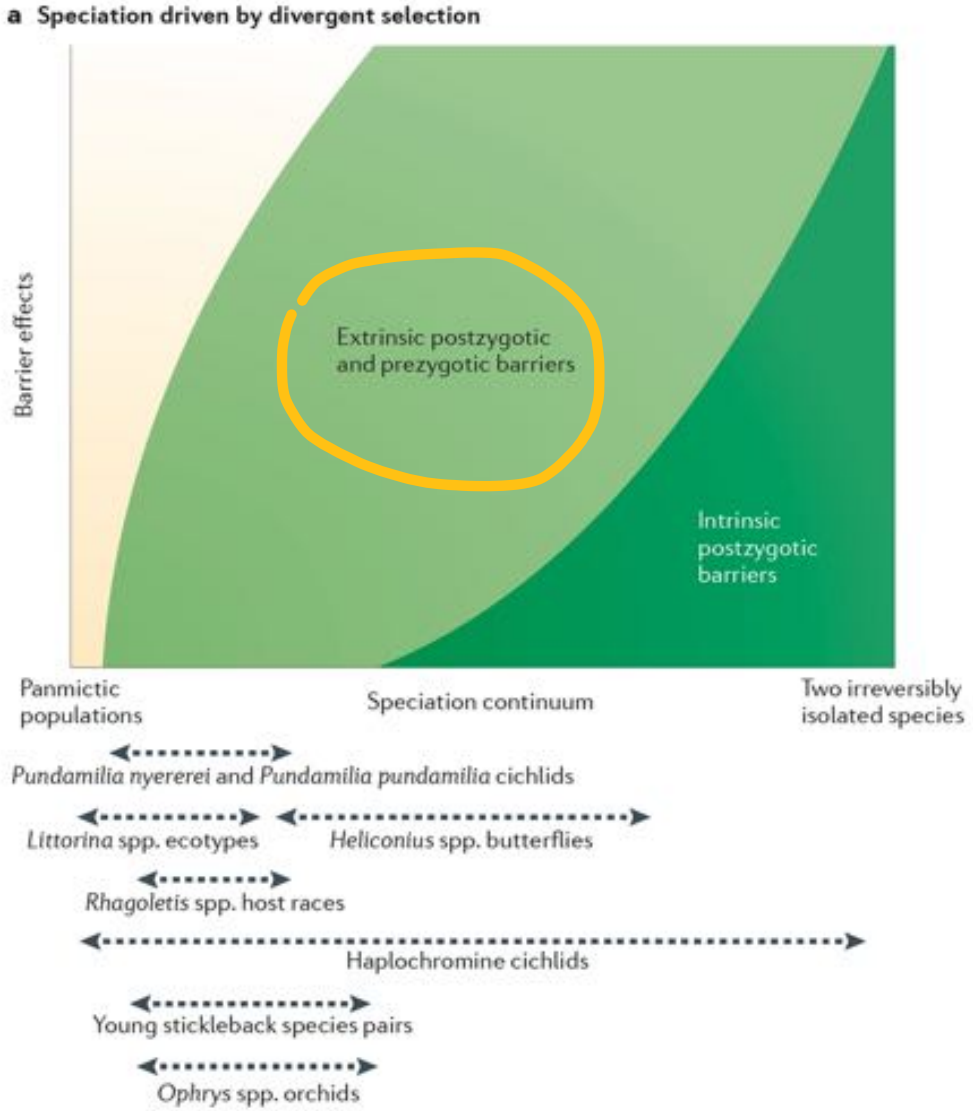
Population genomics

Population genetics studies the genetic differences within and between populations and the dynamics of how populations evolve.

Population genomics studies these genetic differences using many markers to get a better sense of how evolutionary forces shape different parts of the genome.

By comparing differences in genetic diversity and differentiation within species we can study population structure, speciation and adaptation.

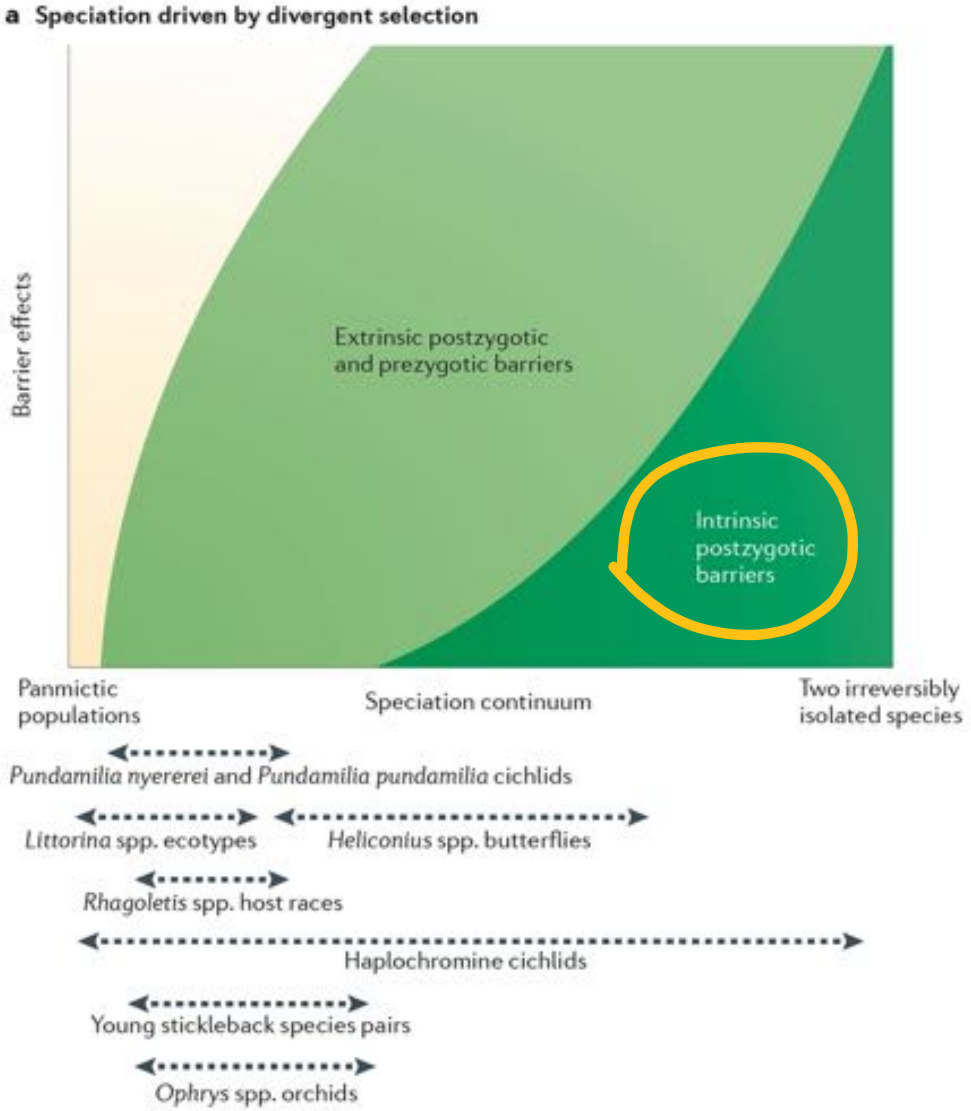
Population genomics for adaptation



Population genomics study populations early in the speciation continuum.

Seehausen et al. 2014, Nat. Gen. Rev.

Population genomics for adaptation

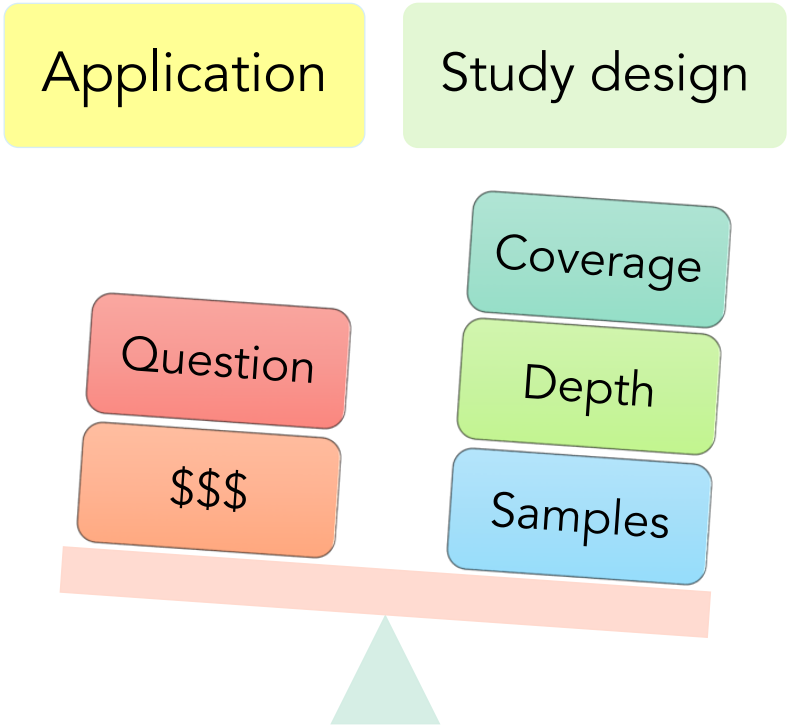


Population genomics study populations early in the speciation continuum.

Later on in the continuum, differentiation builds up and it becomes more and more difficult to distinguish whether genetic differentiation is due to ecological divergence and adaptation or to other factors.

Seehausen et al. 2014, Nat. Gen. Rev.

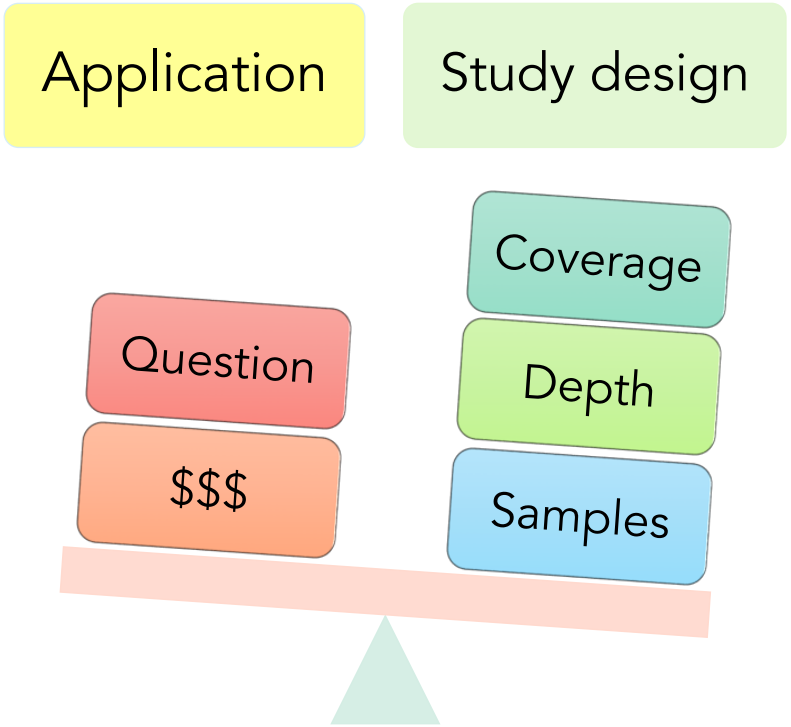
Sequencing methods for population genomics



For our adaptation genomics course, we'll analyze data obtained with **RAD-seq:**

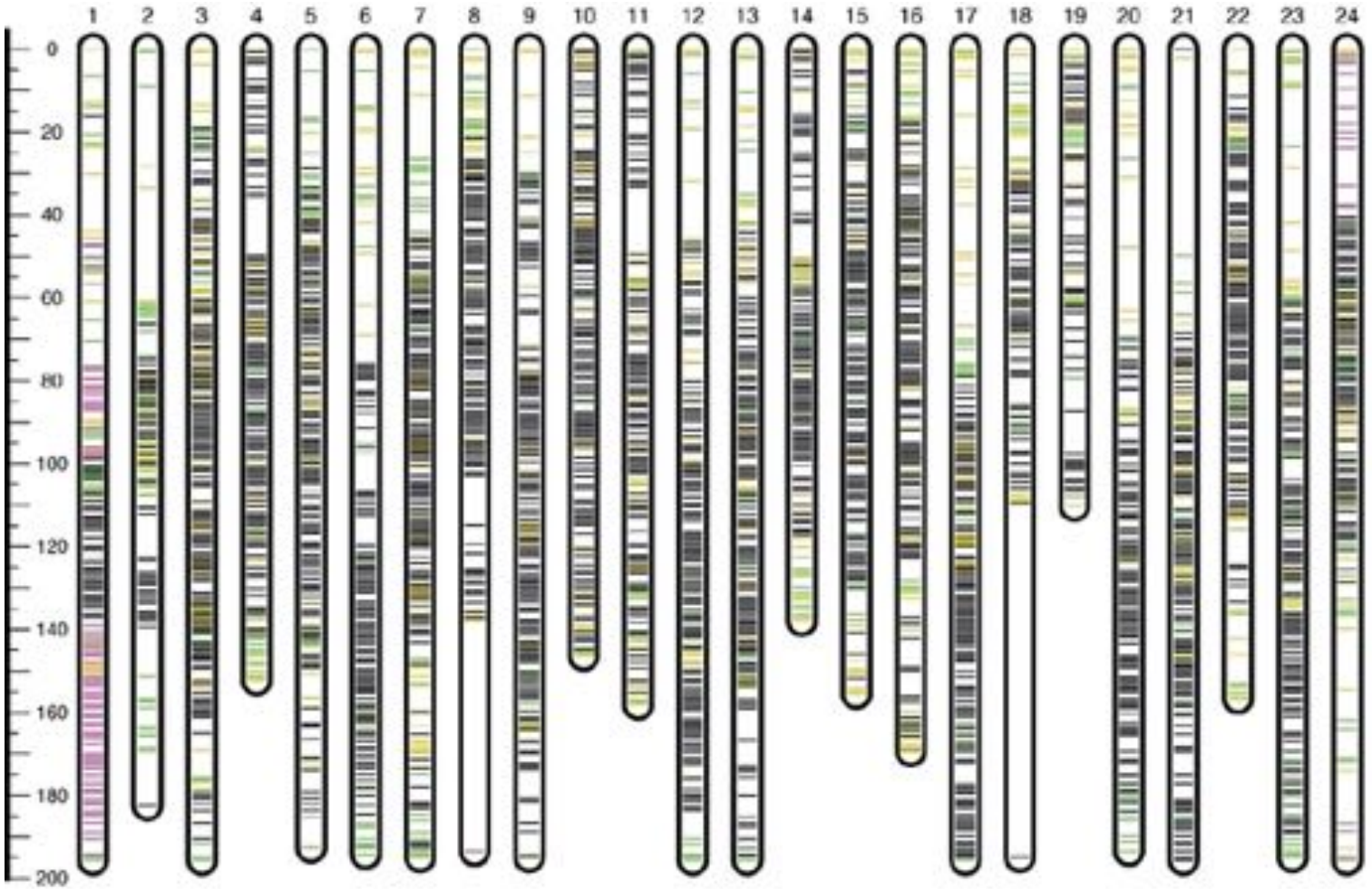
- It provides a manageable amount of data that allows quick analyses.
- It provides skills that are easily transferable for the analysis of other data type (targeted sequencing or WGS)

Sequencing methods for population genomics



RAD-seq

RADseq allows to genotype thousands of loci across many individuals at a reasonable cost and can be tuned to address many different questions



Example of genomic coverage of RAD-seq

Akopyan et al. 2022, Molecular Ecology

Pros of RADseq

- It doesn’t require extensive genomic resources: no need of a high-quality reference genome (though it helps)
- It is customizable: through choice of restriction enzyme and sequencing volumes you can tune coverage of the genome and depth of sequencing
- It samples random loci across the genome, both putative neutral and adaptive loci.

Cons of RADseq

- Because coverage of the genome is not full, there is a risk of missing the locus of interest
- It’s hard to investigate the genomic architecture of adaptive traits
- We have limited information for the characterization of structural variants that could be involved in adaptation (i.e. genomic basis or recombination suppressant)

Bioinformatic pipeline

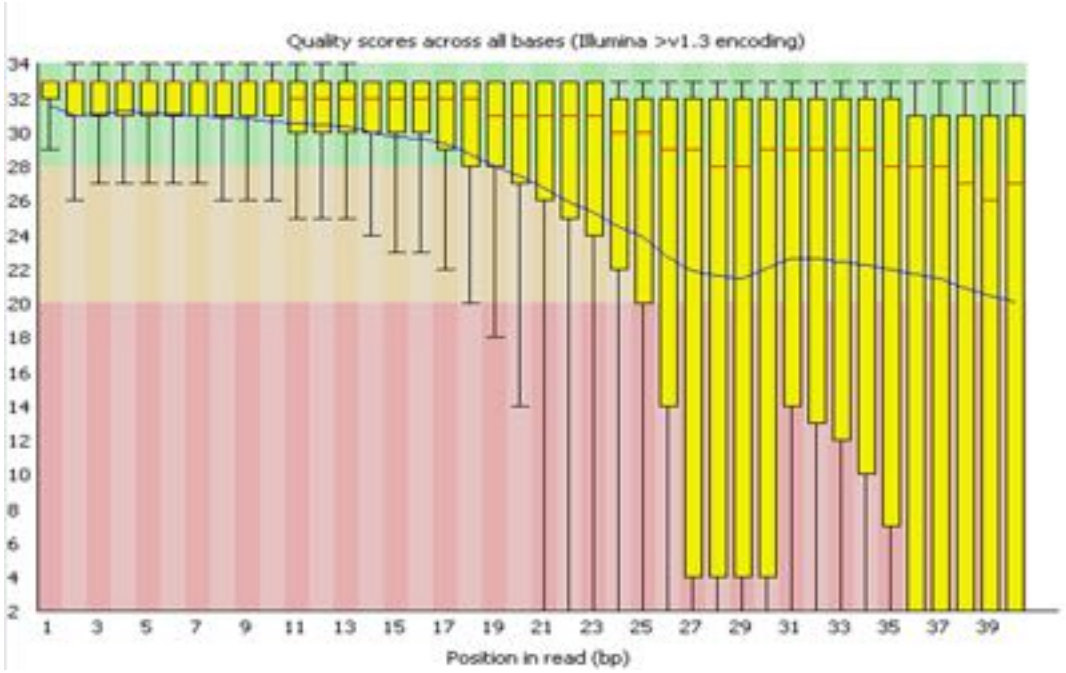
Bioinformatic pipelines

.fastq .fastq

Quality control → Adapter trimming → Read aligning → Variant calling

FASTQC

file:///Users/anna/Dropbox/genomes%20analyses/murre_hunt/completedataset/fastqc_results/lane1.Tig1_R1_fastqc.html



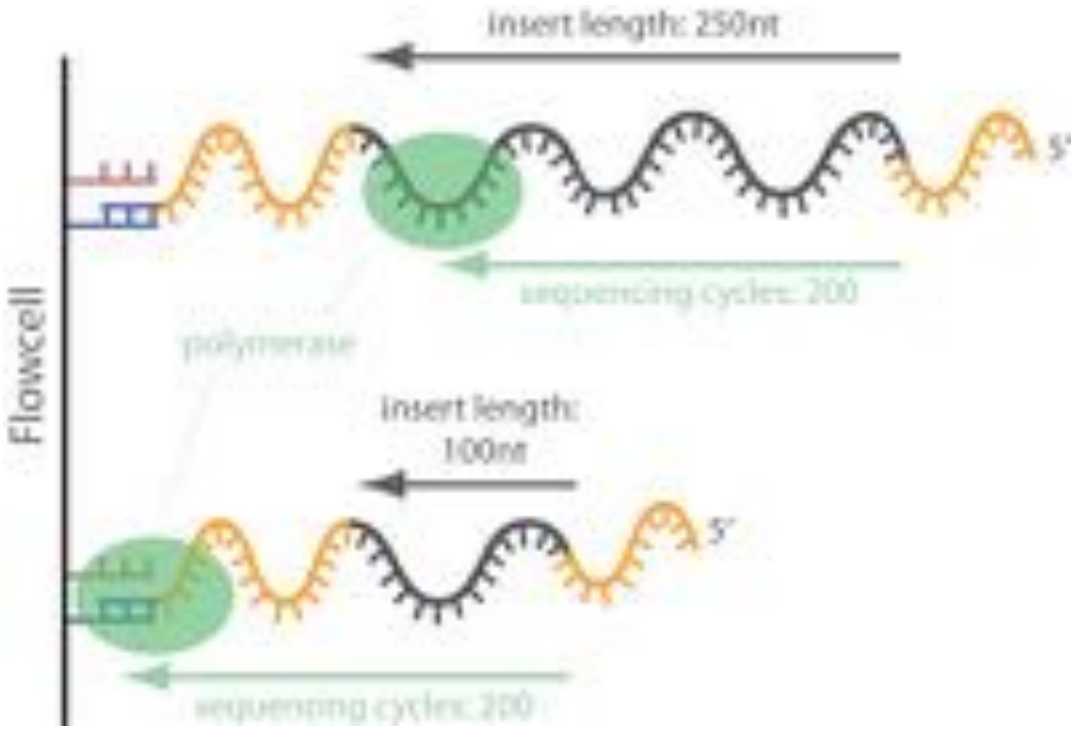
Bioinformatic pipeline

.fastq .fastq .fastq

Quality control → Adapter trimming → Read aligning → Variant calling

FASTQC

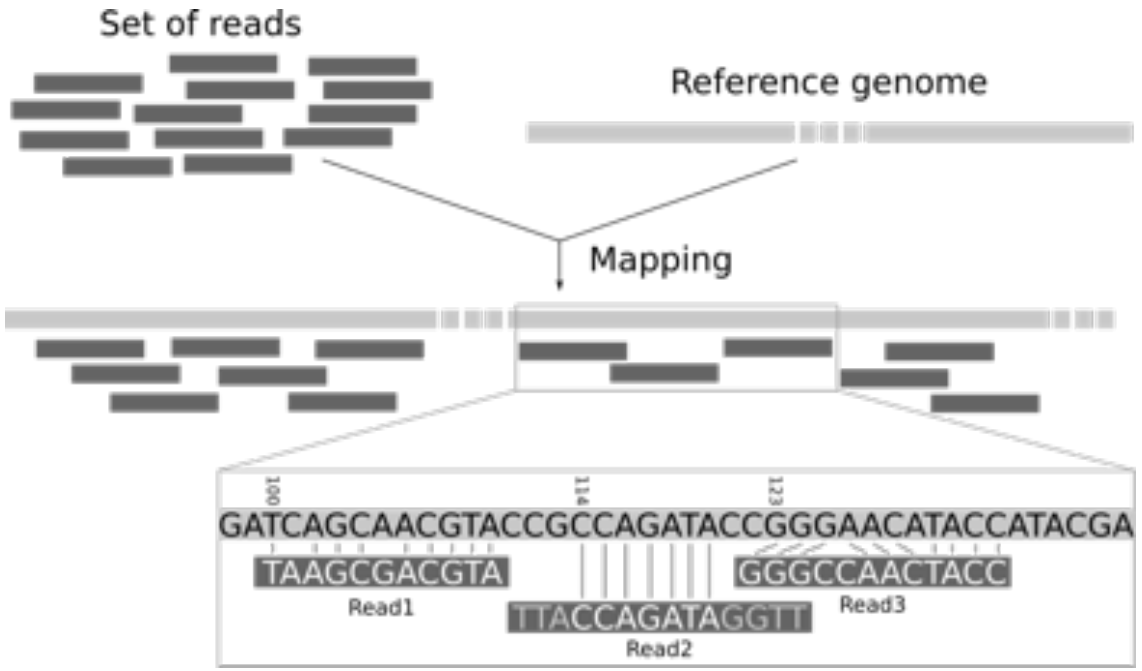
Trimmomatic
Cutadapt
Fastp



Bioinformatic pipeline

.fastq .fastq .fastq .sam/.bam

Quality control → Adapter trimming → Read aligning → Variant calling

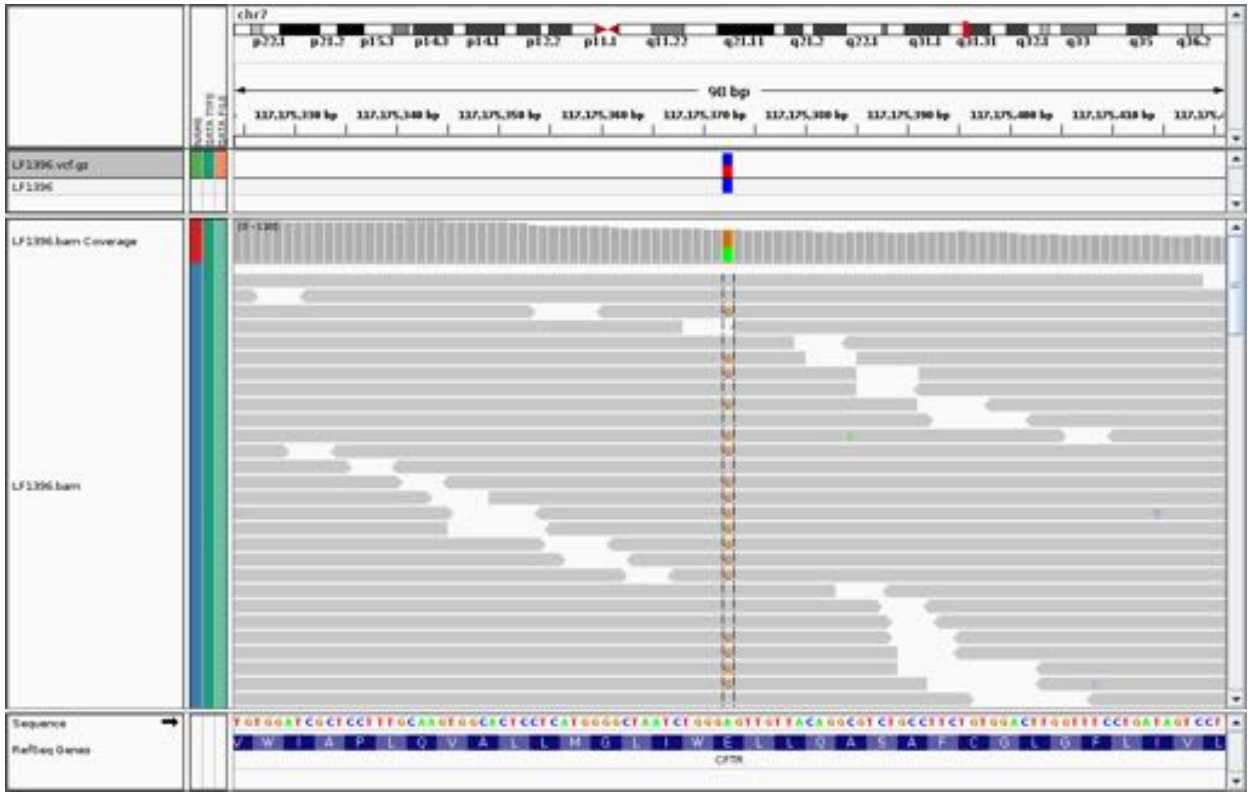


Bowtie2
BWA

Bioinformatic pipeline

.fastq .fastq .fastq .sam/.bam .vcf

Quality control → Adapter trimming → Read aligning → Variant calling



STACKS
ANGSD
GATK
SAMtools
bcftools
…

A VCF is a VCF is a VCF

Header – commands + contigs/chromosomes

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.11+htslib-1.11
##bcftoolsCommand=mpileup -Ou -f reference/onerka_chr.fa -b sample_lists/bams_allmgi.txt -q 5 -Q 30 -r NC_042535.1:1-10000000 -I -a AD,DP,SP,ADF,ADR -d 200
##reference=file://reference/onerka_chr.fa
##contig=<ID=NC_042535.1,length=41065921>
##contig=<ID=NC_042536.1,length=61175412>
##contig=<ID=NC_042537.1,length=59001101>
```

Header – info fields

```
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of raw reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of raw reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
```

Header – columns names

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  goodbam/ALOL_DP_0187.bam        goodbam/ALOL_DP_2757.bam        goodbam/ALOL_DP_2780.bam
```

Variant information

```
NC_042535.1     801     .       G       A       988     PASS    AN=976;AC=39    GT:PL:DP:SP:ADF:ADR:AD  0/0:0,27,239:9:0:2,0:7,0:9,0    0/0:0,45,255:15:0:7,0:8,0:15,0  0/0:0,36,255:12:0:5,0:7,0:12,0
```

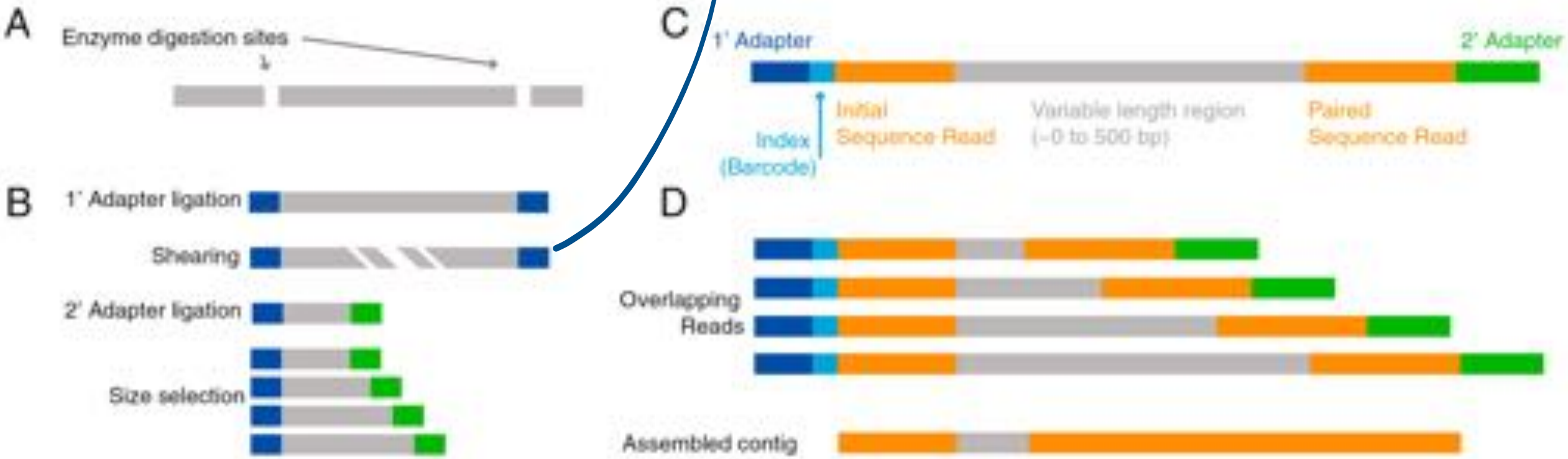
Library preparation and sequencing

Knowing the technical aspects of library preparation and sequencing is important to properly handle and analyze the data and identify potential biases/problems

- Type of library preparation: method, enzymes used, insert size, input DNA quantity and quality, etc…
- Sequencing: technology, platform, read length, single vs. paired-end, depth, etc…

RADseq pipeline

Or double digestion

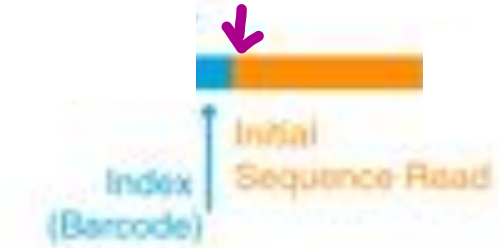


RADseq pipeline

Raw reads

READ 1

Restriction site



READ 2

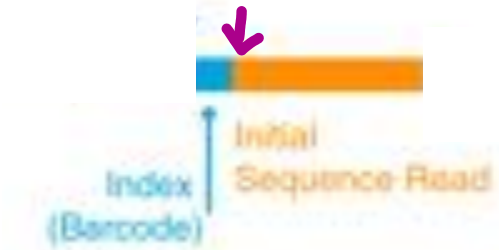
Restriction site

RADseq pipeline

Raw reads

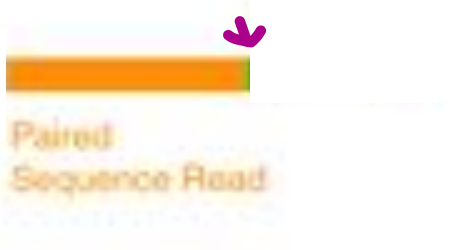
READ 1

Restriction site



READ 2

Restriction site



In addition to potential adapter contamination,
we need to demultiplex RADseq libraries

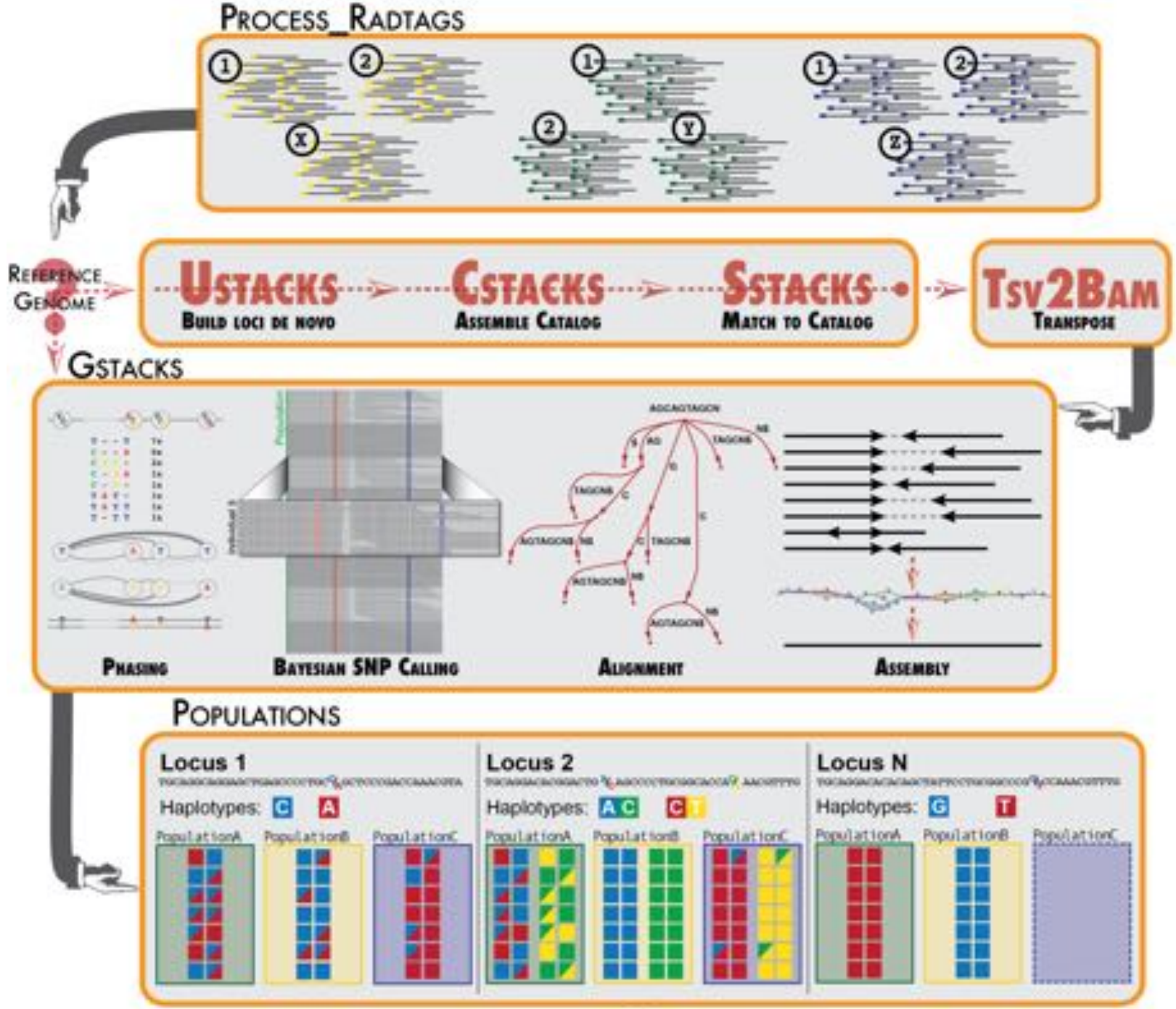
RADseq pipelines

- STACKS (Catchen et al. 2013, Molecular Ecology)
- dDocent (Puritz et al. 2014, PeerJ)
- PyRAD (Eaton 2014, Bioinformatics)
- AftrRAD (Sovic et al. 2015, Molecular Ecology Resources)
- ANGSD (Korneliussen et al. 2014)
- GATK (McKenna et al. 2010, Genome Research)

RADseq pipelines

- STACKS (Catchen et al. 2013, Molecular Ecology)
- dDocent (Puritz et al. 2014, PeerJ)
- PyRAD (Eaton 2014, Bioinformatics)
- AftrRAD (Sovic et al. 2015, Molecular Ecology Resources)
- ANGSD (Korneliussen et al. 2014)
- GATK (McKenna et al. 2010, Genome Research)

STACKS

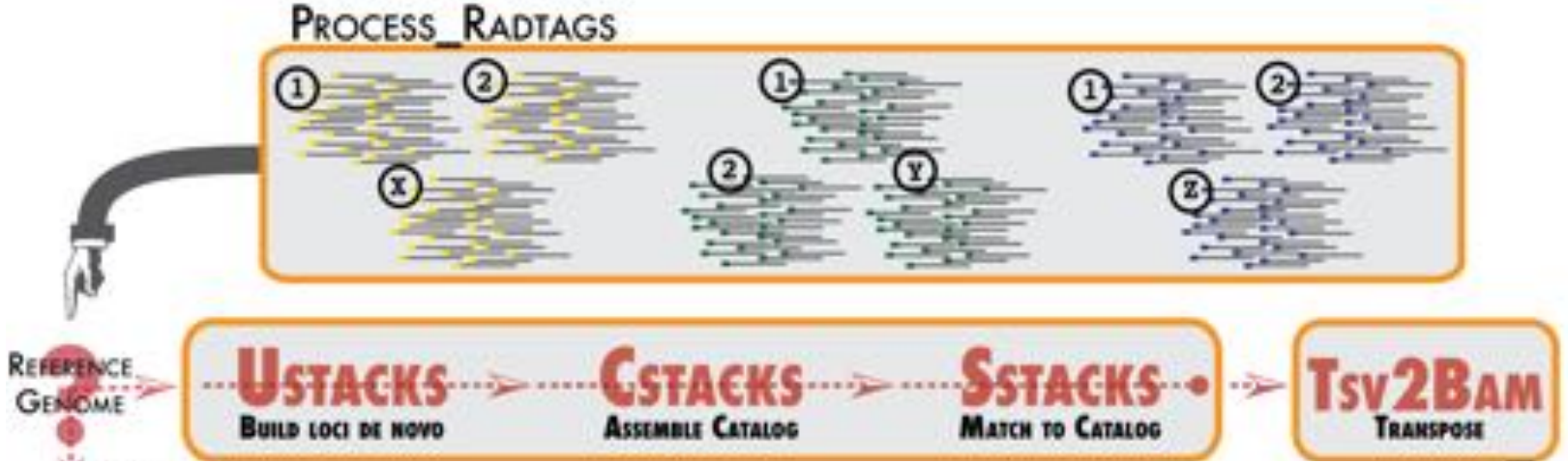

Process_Radtags
Reference Genome
Ustacks
Build loci de novo
Cstacks
Assemble Catalog
Sstacks
Match to Catalog
Tsv2Bam
Transpose
Gstacks
Phasing
Bayesian SNP Calling
Alignment
Assembly
Populations
Locus 1
Haplotypes: C A
Locus 2
Haplotypes: A C C T
Locus N
Haplotypes: G T
PopulationA
PopulationB
PopulationC

STACKS

To preprocess raw data

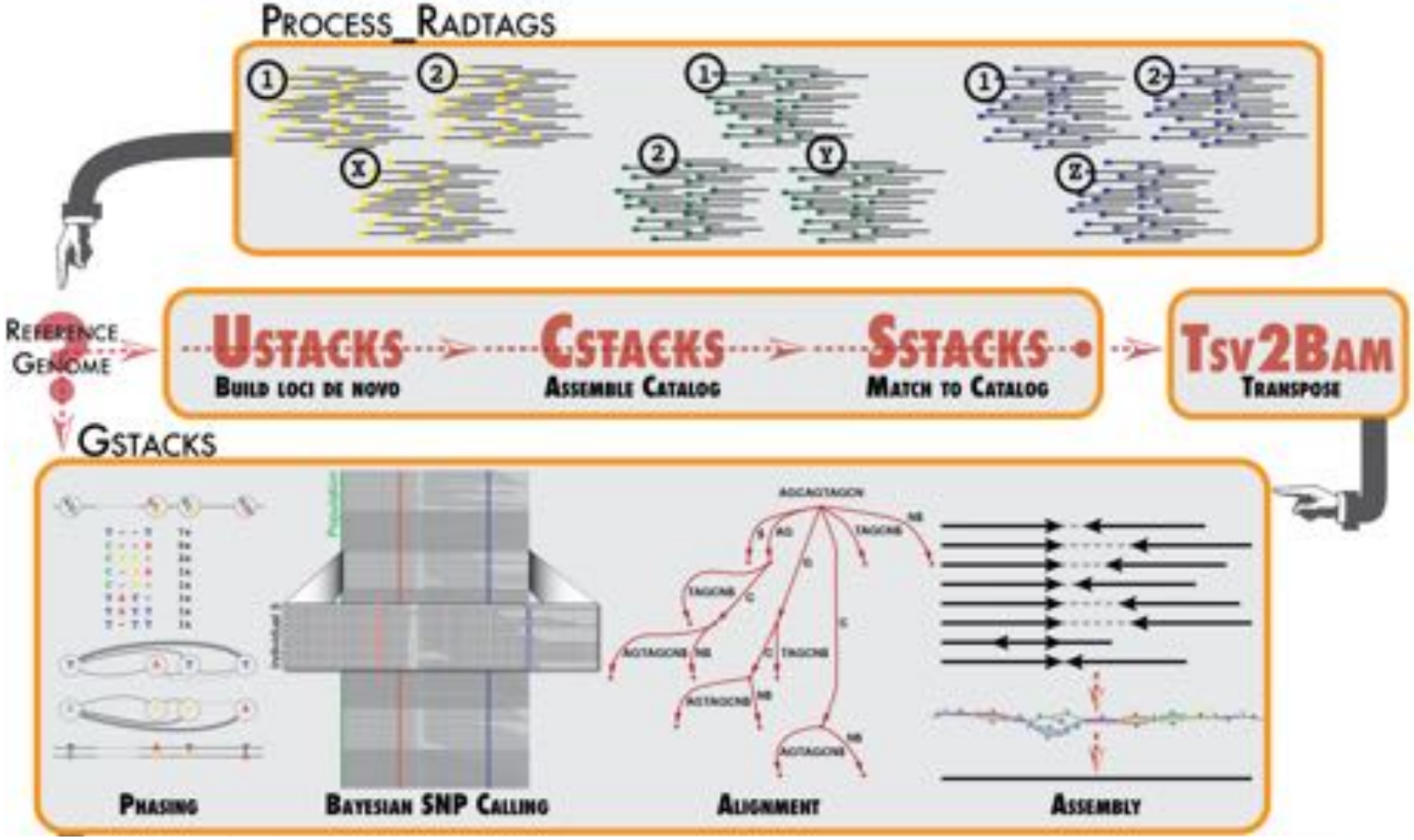
- Demultiplexing
- Adapter removal
- Quality filtering

STACKS



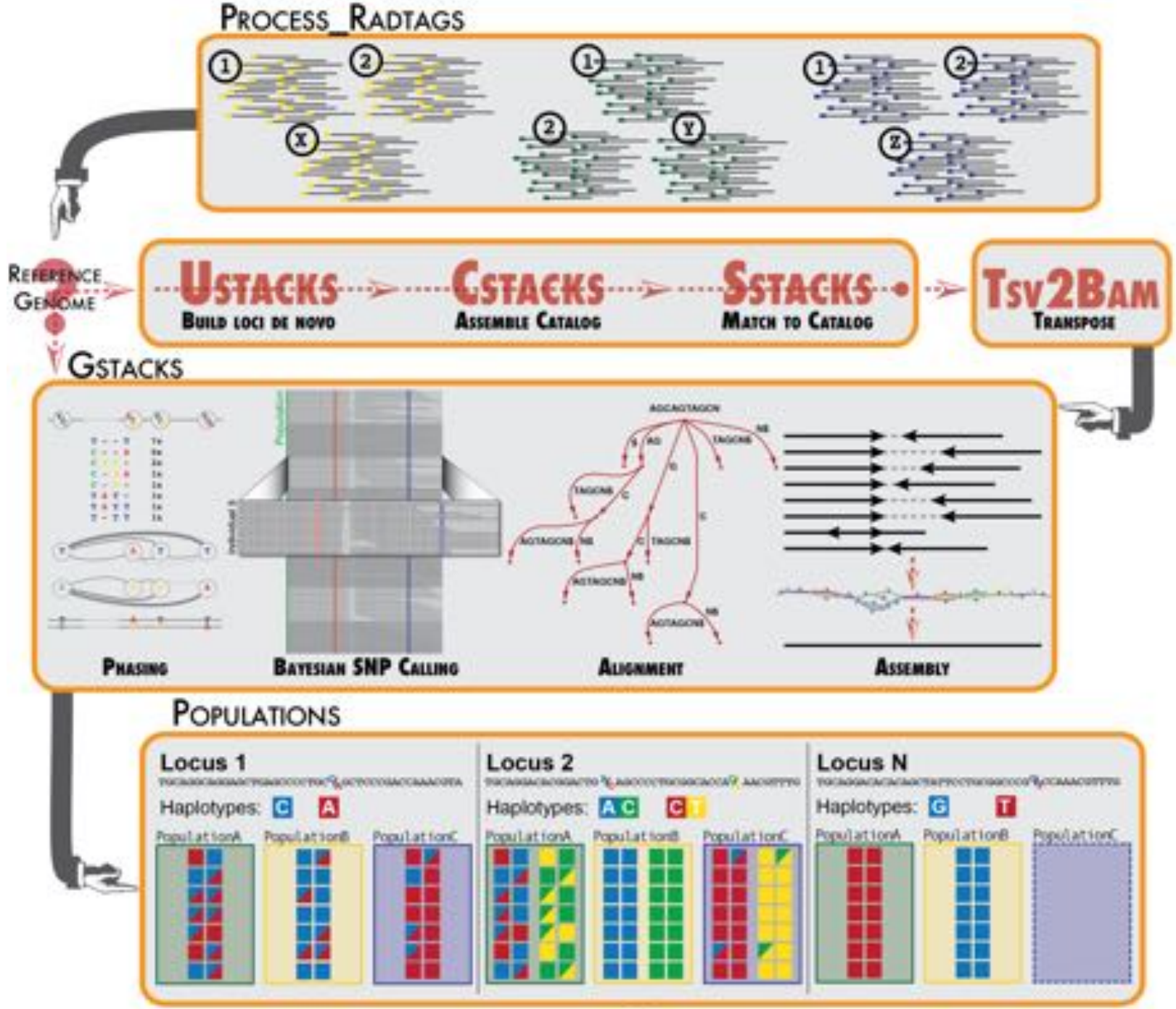
Loci assembly without reference genome

STACKS



If you have a reference genome, align RAD data with external software.
GSTACKS does different things according to data input but at end it calls variants from assembled loci.

STACKS


Process_Radtags
Reference Genome
Ustacks
Build loci de novo
Cstacks
Assemble Catalog
Sstacks
Match to Catalog
Tsv2Bam
Transpose
Gstacks
Phasing
Bayesian SNP Calling
Alignment
Assembly
Populations
Locus 1
Haplotypes: C A
Locus 2
Haplotypes: A C C T
Locus N
Haplotypes: G T
PopulationA
PopulationB
PopulationC

Variant calling from whole genome data

Most commonly used software for variant calling

Low coverage WGS

→ ANGSD to keep into account genotype uncertainty

→ may require specific software to work with genotype likelihoods rather than genotypes

Moderate to high coverage

→ bcftools mpileup

→ GATK

Variant calling from whole genome data

Most commonly used software for variant calling

Low coverage WGS

→ ANGSD to keep into account genotype uncertainty

→ may require specific software to work with genotype likelihoods rather than genotypes

Moderate to high coverage

→ bcftools mpileup → example in the tutorial later

→ GATK