Physalia
Courses

Population genomics for adaptation

Day 1 - Lecture 2

(adapted from Claire Mérot & Anna Tigano’s slides)

Analytical approaches

GWAS

Comparative genomics

Transcriptomics

Population genomics

Experimental evolution

Epigenetics

QTL mapping

Analytical approaches

GWAS

Comparative genomics

Transcriptomics

Population genomics

Experimental evolution

Epigenetics

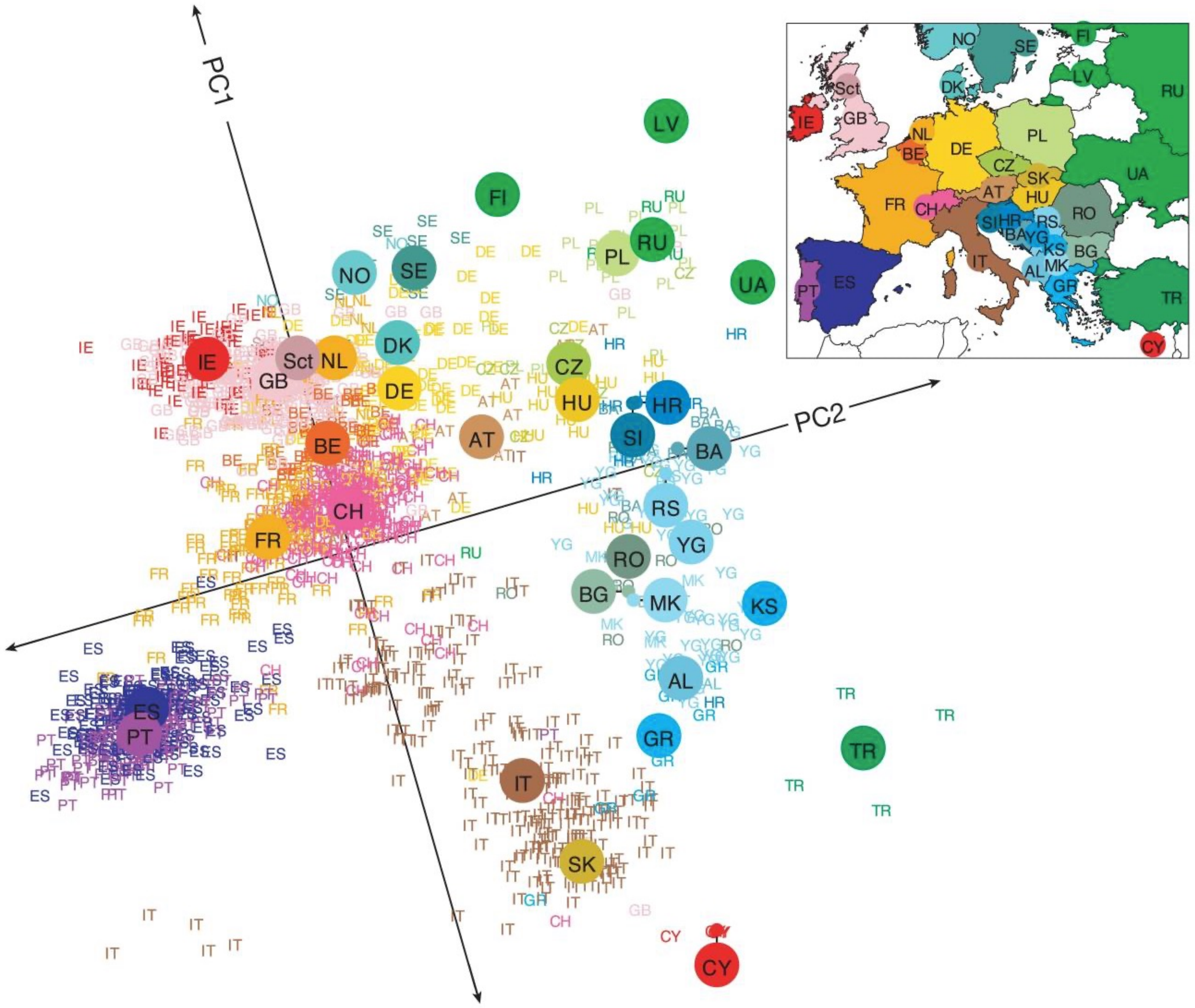
QTL mapping

Population genomics

- Studies the genetic differences within and between populations and the dynamics of how populations evolve

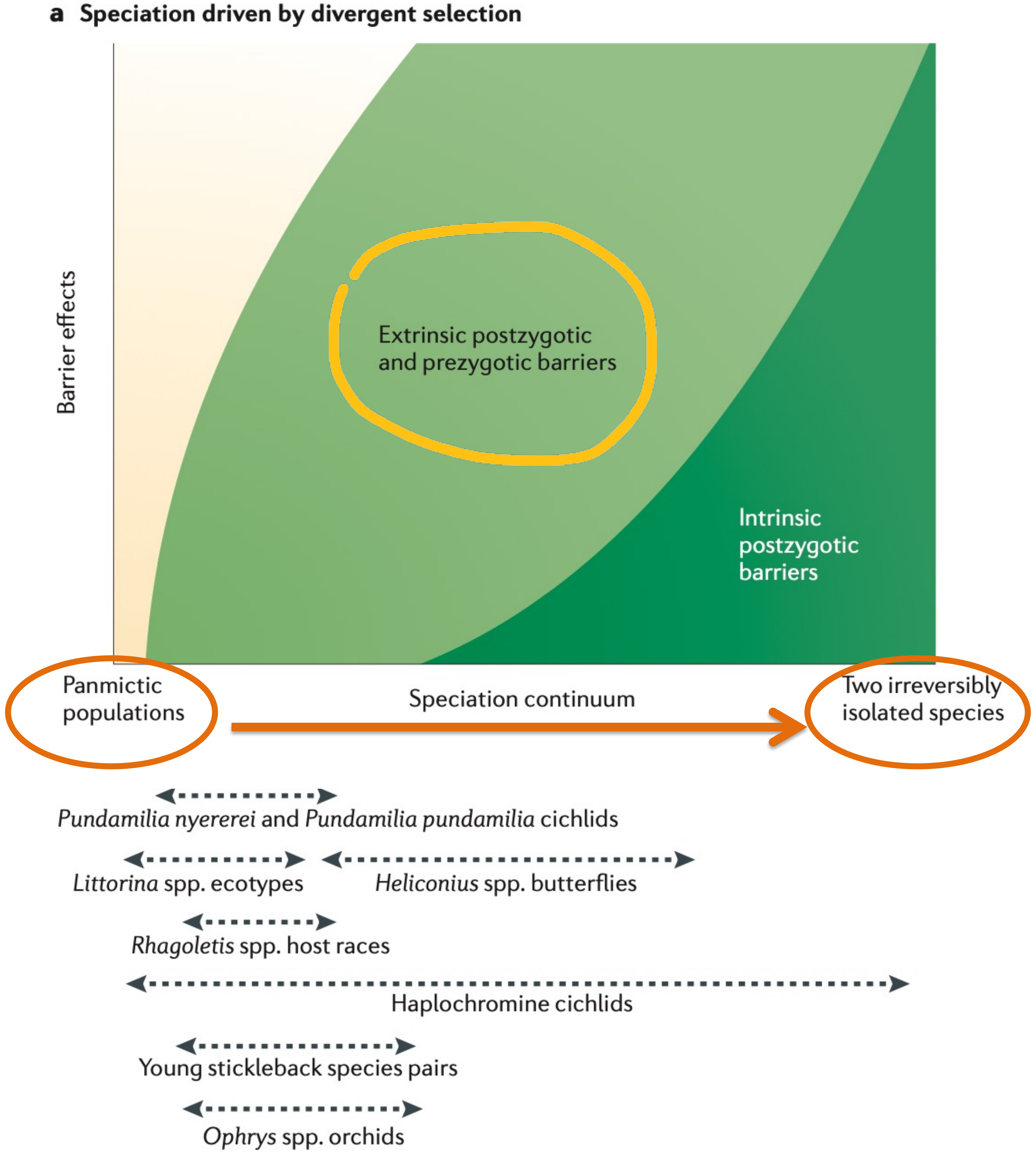
- Genetic differences are investigated using genetic markers that allow to assess how evolutionary forces shape different parts of the genome

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



Novembre, J., Johnson, T., Bryc, K. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008). https://doi.org/10.1038/nature07331

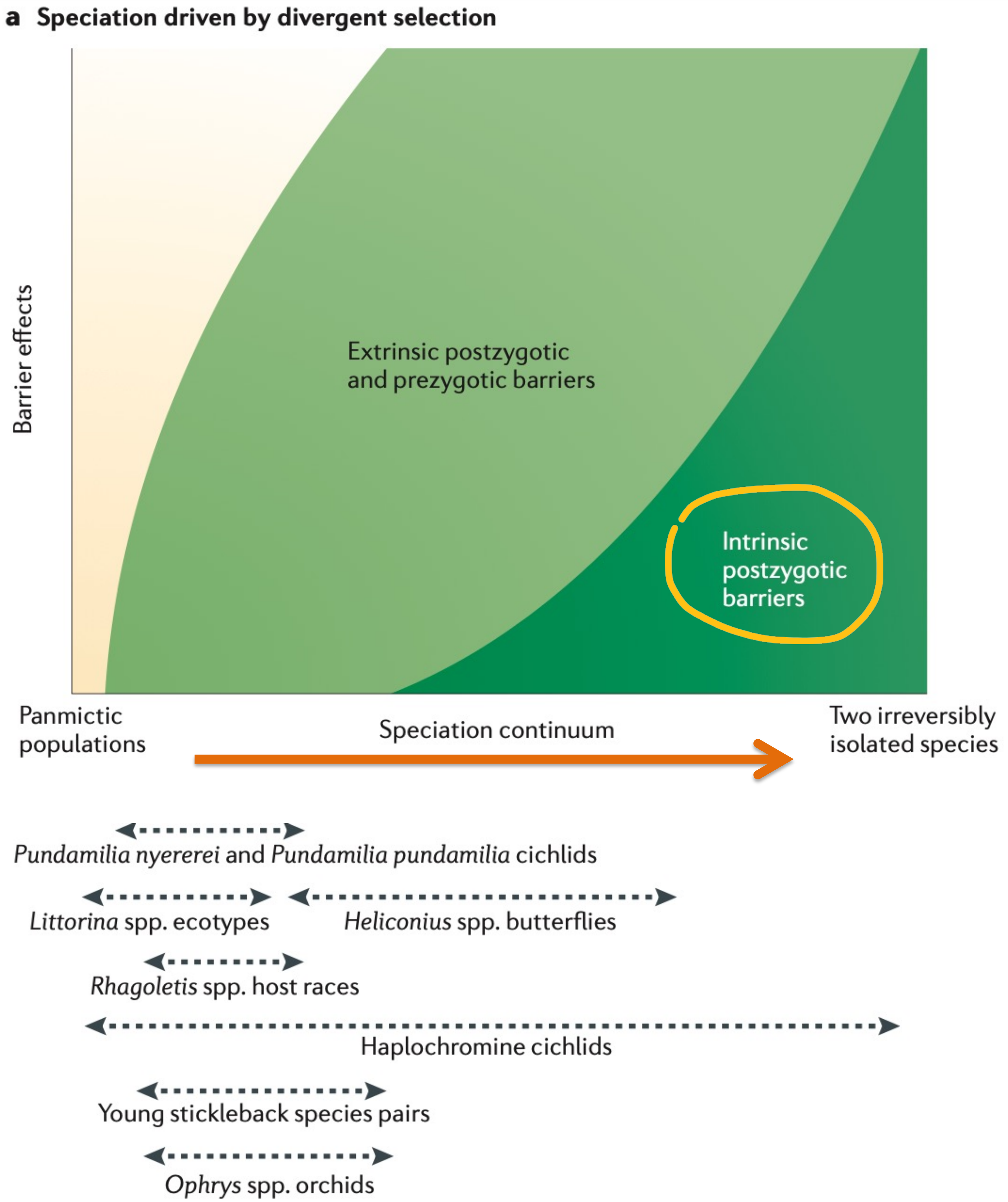
Population genomics for adaptation


a Speciation driven by divergent selection
Barrier effects
Extrinsic postzygotic and prezygotic barriers
Intrinsic postzygotic barriers
Panmictic populations
Speciation continuum
Two irreversibly isolated species
Pundamilia nyererei and *Pundamilia pundamilia* cichlids
Littorina spp. ecotypes
Heliconius spp. butterflies
Rhagoletis spp. host races
Haplochromine cichlids
Young stickleback species pairs
Ophrys spp. orchids

- Population genomics study populations early in the speciation continuum

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

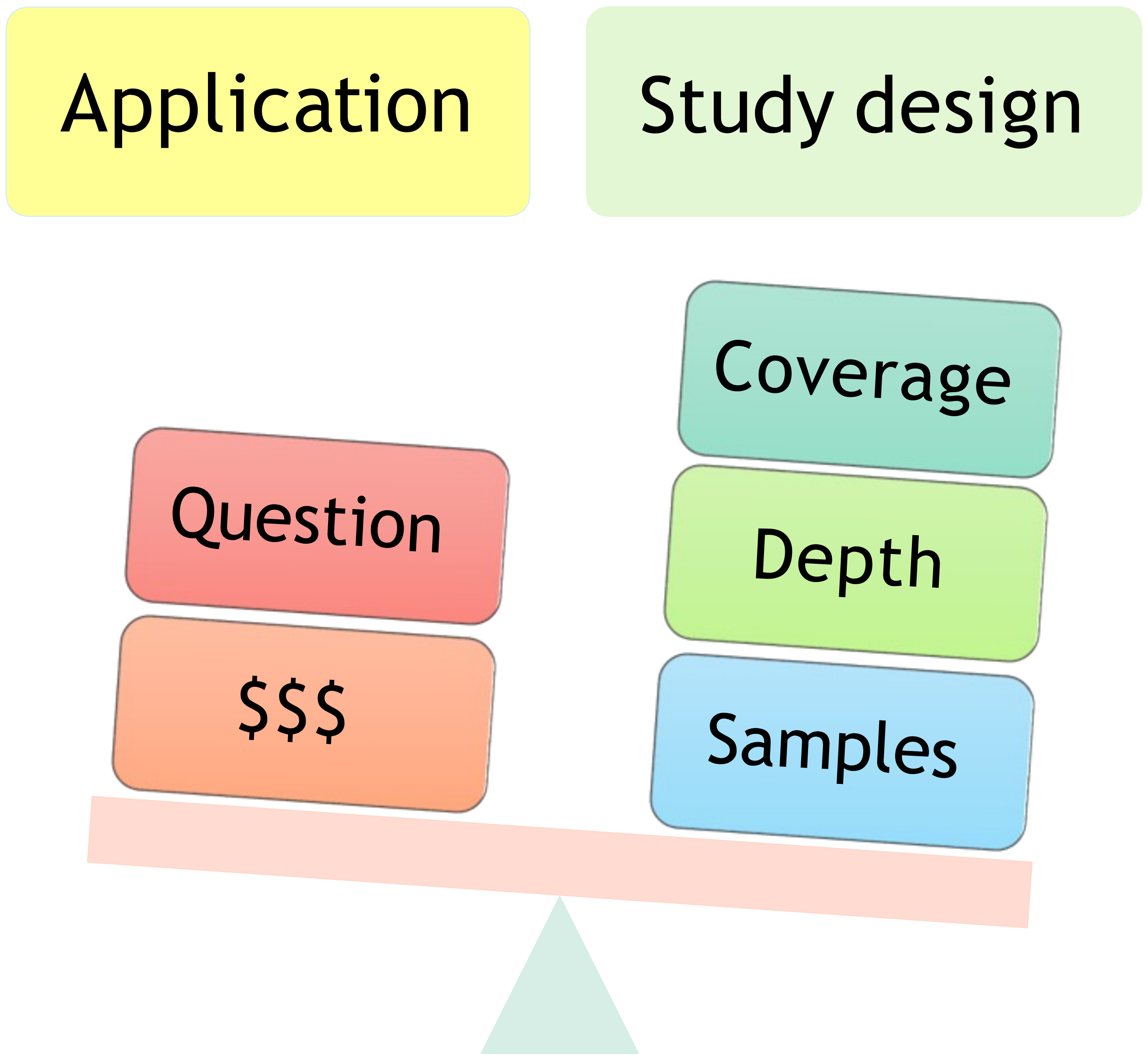
Population genomics for adaptation


a Speciation driven by divergent selection
Barrier effects
Extrinsic postzygotic and prezygotic barriers
Intrinsic postzygotic barriers
Panmictic populations
Speciation continuum
Two irreversibly isolated species
Pundamilia nyererei and *Pundamilia pundamilia* cichlids
Littorina spp. ecotypes
Heliconius spp. butterflies
Rhagoletis spp. host races
Haplochromine cichlids
Young stickleback species pairs
Ophrys spp. orchids

- Population genomics study populations early in the speciation continuum
- Later on in the continuum, differentiation builds up and it becomes more challenging to distinguish whether genetic differentiation is due to ecological divergence and adaptation, or to other factors

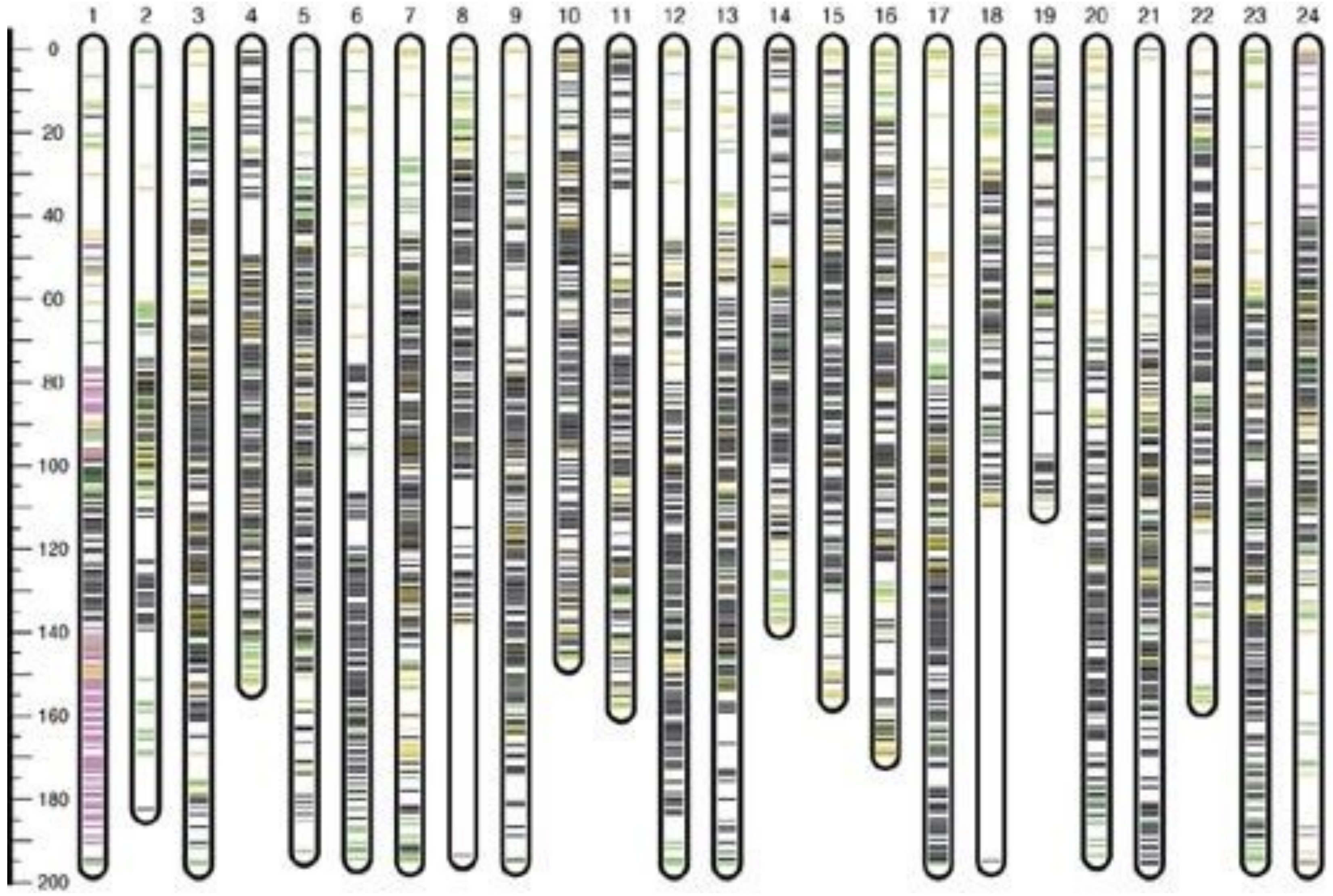
Seehausen et al. 2014, Nat. Gen. Rev.

Sequencing methods for population genomics



For the course, we will analyze genomic data obtained with **RAD-seq**, as it provides:

- a manageable amount of data for quick analyses (short compute runs)
- genotype of thousands of loci for many individuals at a reasonable cost
- power to address diversity of research questions (need fine-tunning)
- data analysis skills easily transferable to other genomic data types (WGS, targeted sequencing)



Example of potential genomic coverage of RADseq (fined-tuned)

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 fine-tune coverage of the genome and depth of sequencing
- It samples random loci across the genome, both putative neutral and adaptive loci

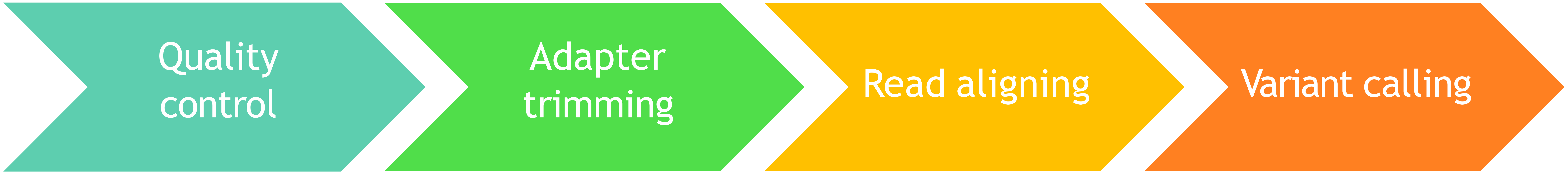
Pros of RADseq

- 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 fine-tune coverage of the genome and depth of sequencing
- It samples random loci across the genome, both putative neutral and adaptive loci

Cons of RAD-seq

- Because coverage of the genome is not full, there is a risk of missing 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



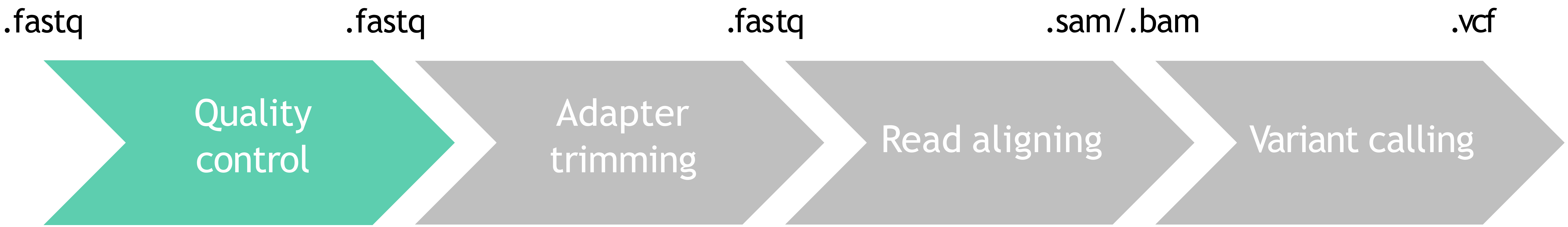
Quality control

Adapter trimming

Read aligning

Variant calling

Bioinformatic pipeline



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

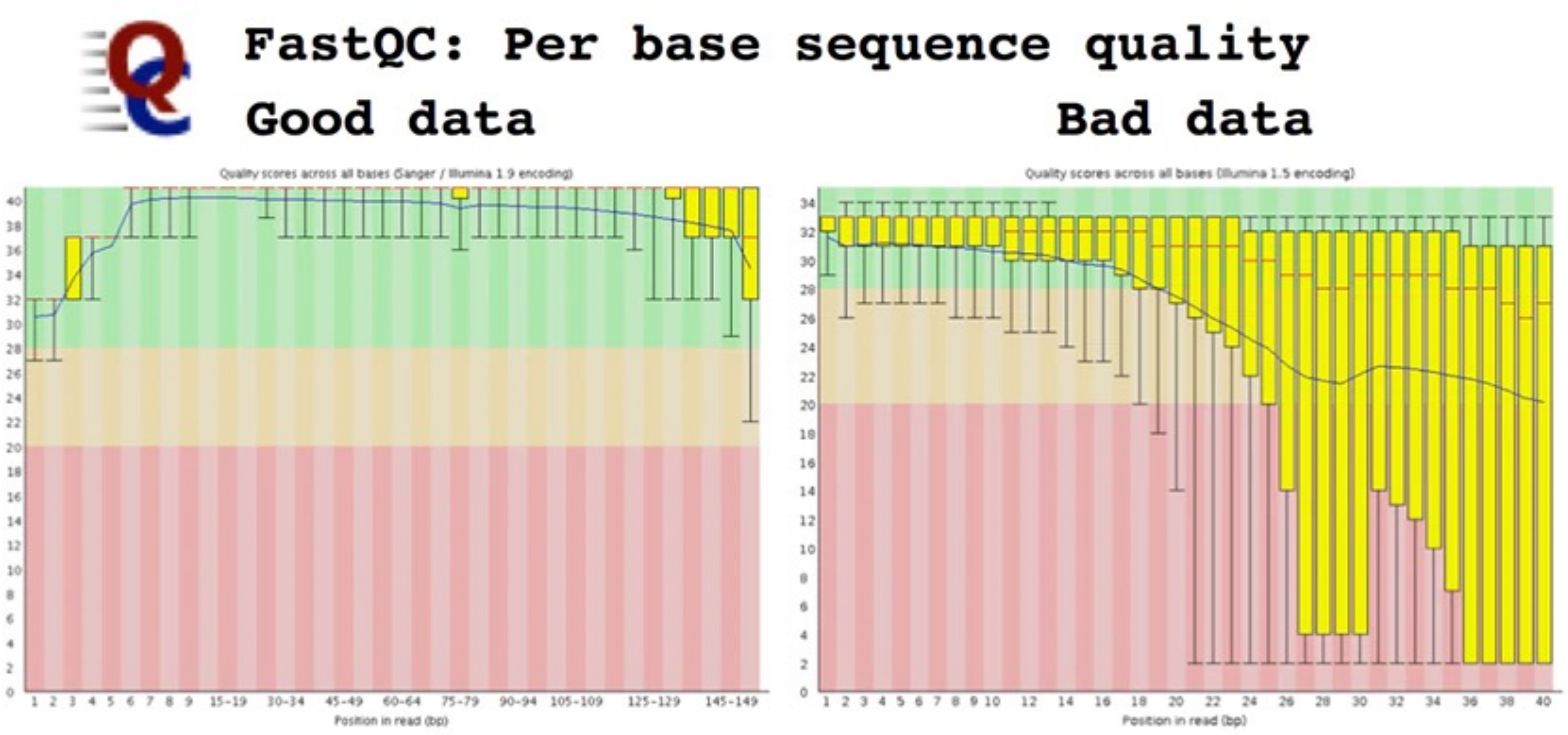
Quality control

Adapter trimming

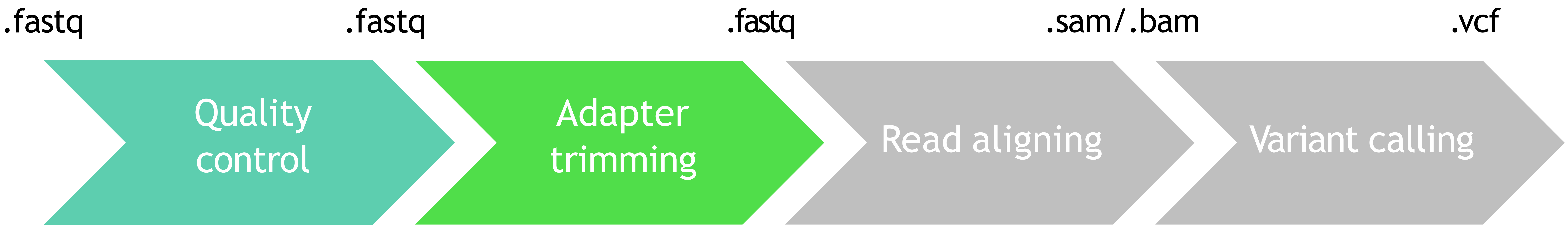
Read aligning

Variant calling

FastQC



Bioinformatic pipeline



.fastq

Quality control

.fastq

Adapter trimming

.fastq

Read aligning

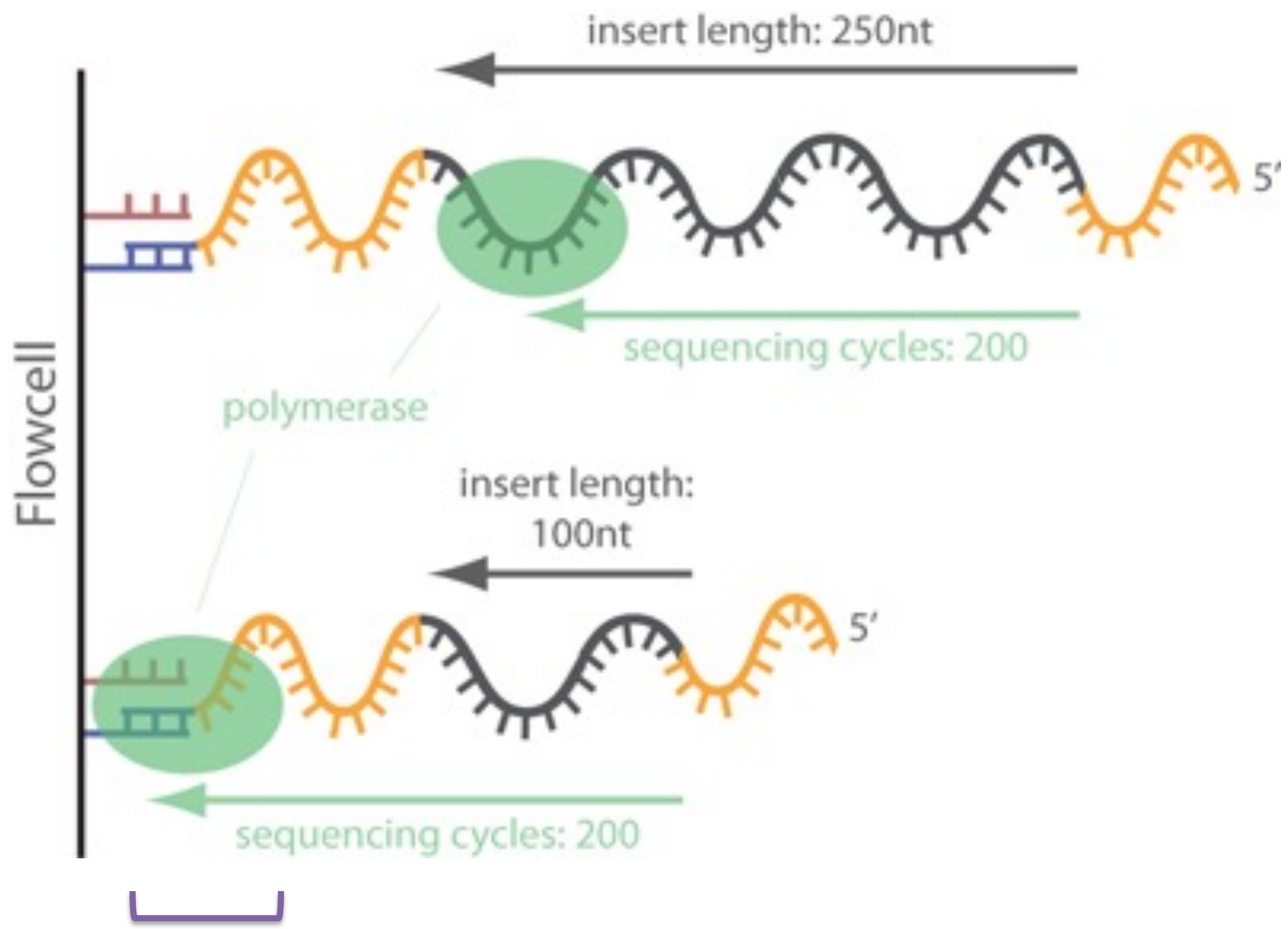
.sam/.bam

Variant calling

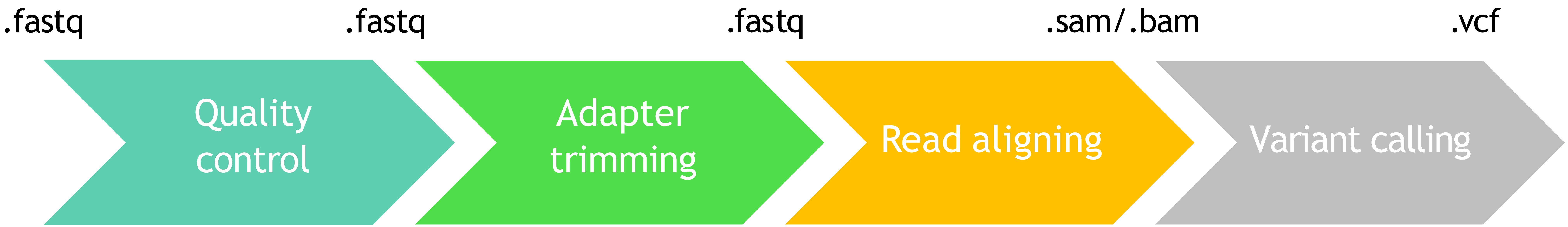
.vcf

FastQC

Trimmomatic
Cutadapt
Fastp



Bioinformatic pipeline



.fastq

Quality control

.fastq

Adapter trimming

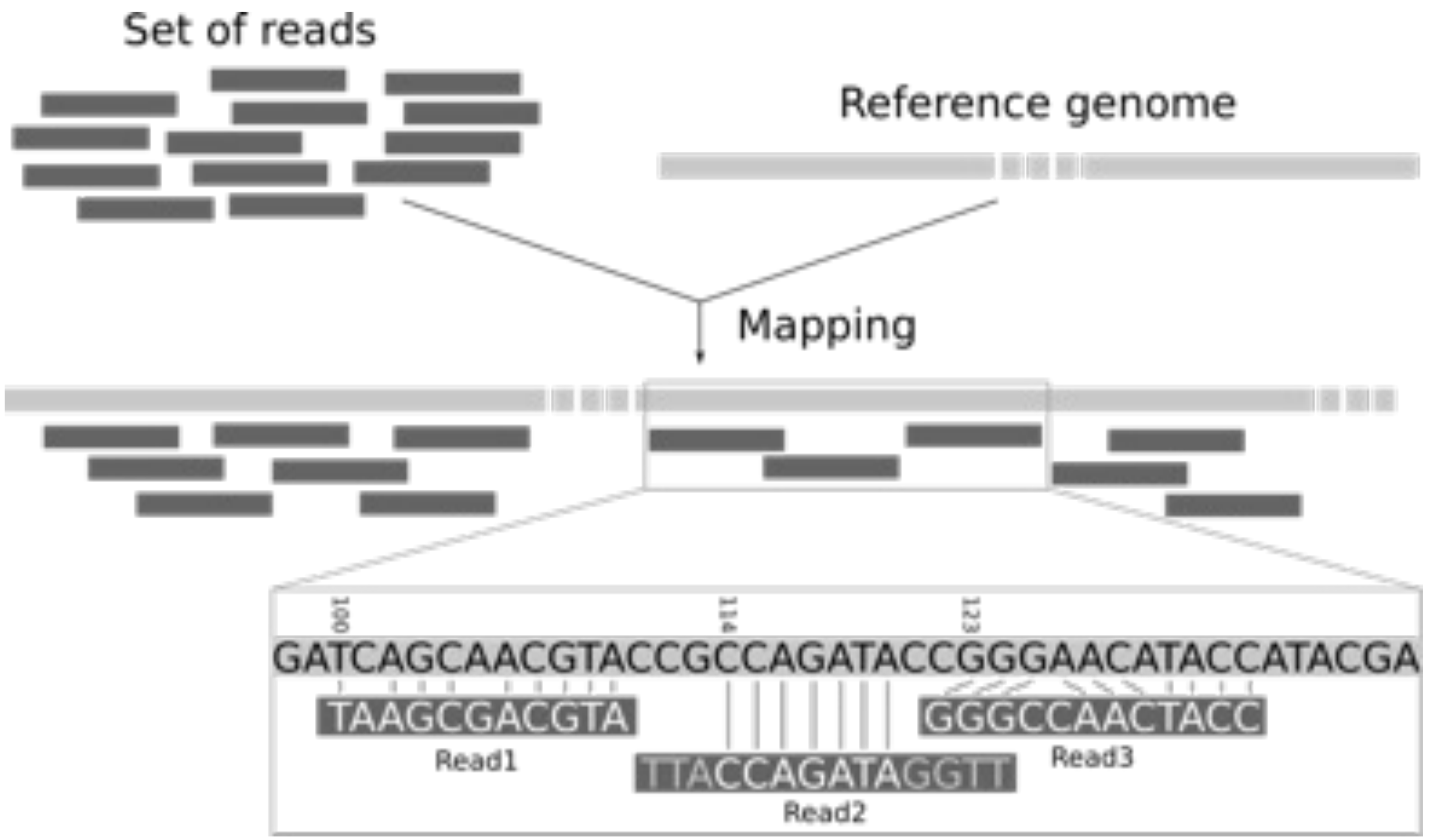
.fastq

Read aligning

.sam/.bam

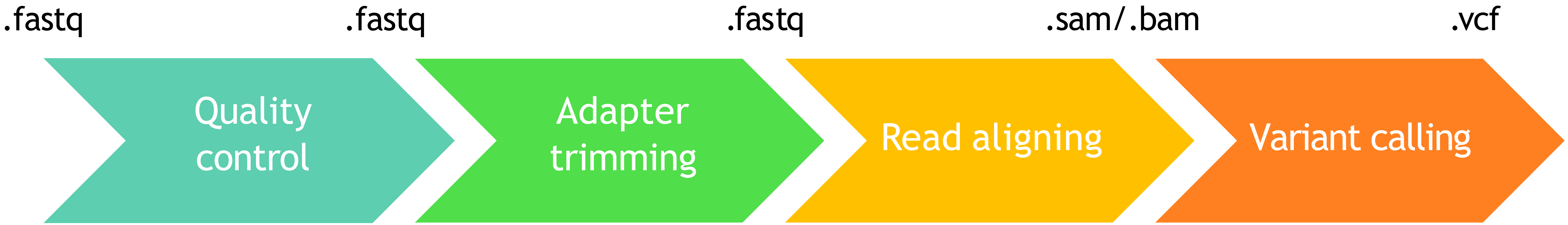
Variant calling

.vcf



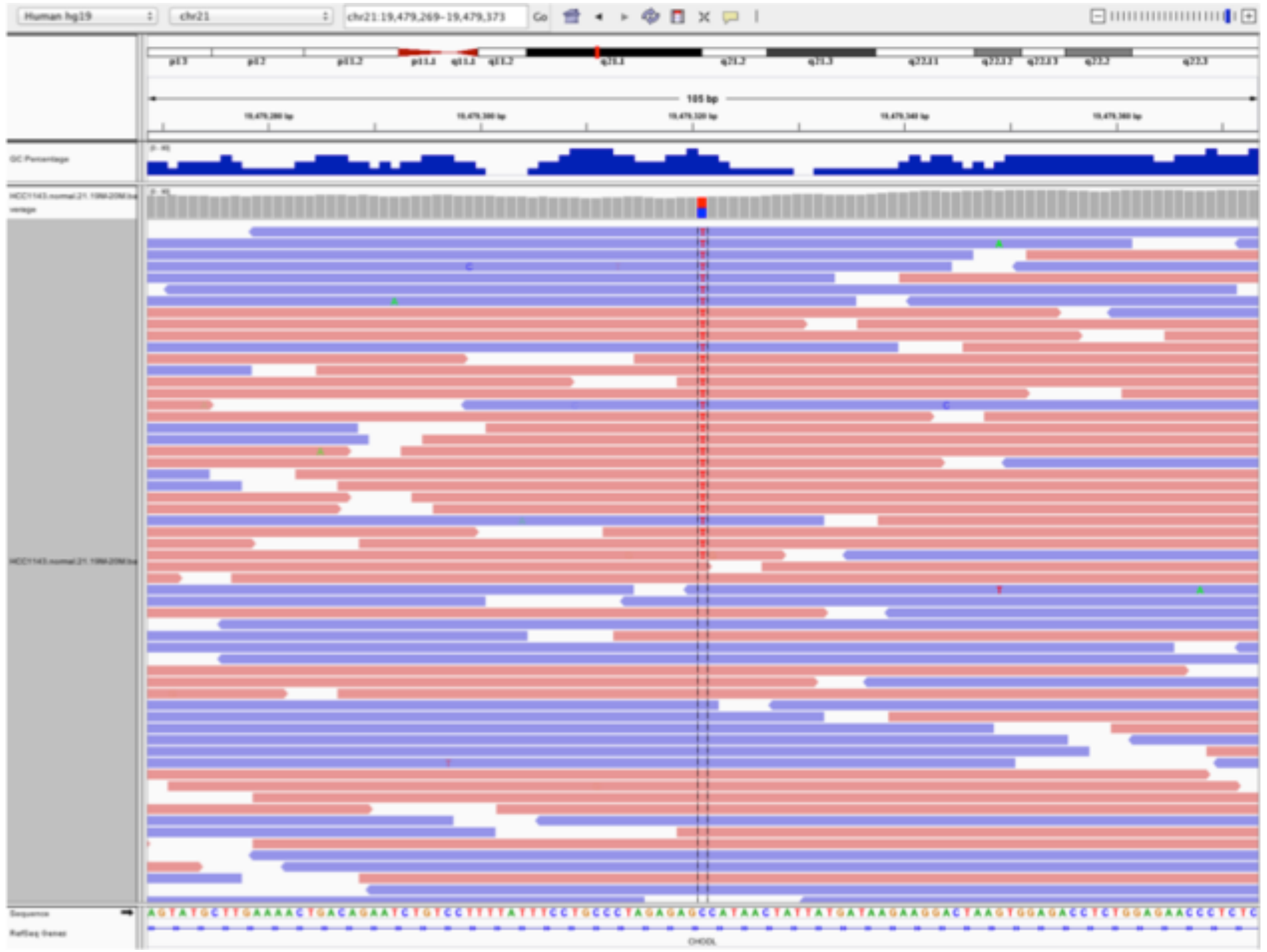
Bowtie2
BWA

Bioinformatic pipeline



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

Quality control → Adapter trimming → Read aligning → Variant calling

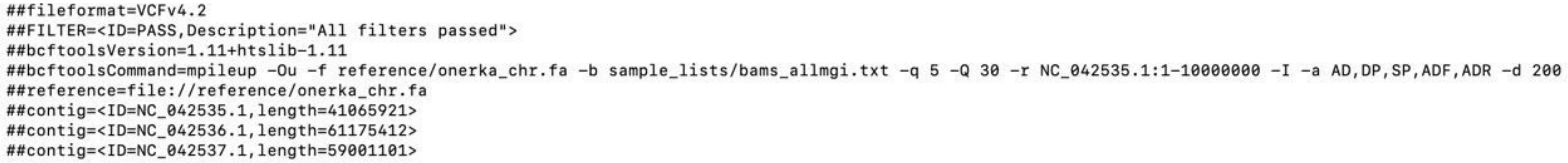


IGV screenshot of a SNP

Stacks
ANGSD
GATK
SAMtools
bcftools
…

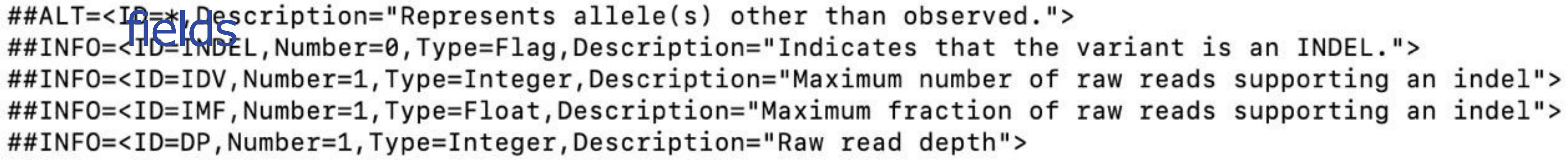
The VCF file

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

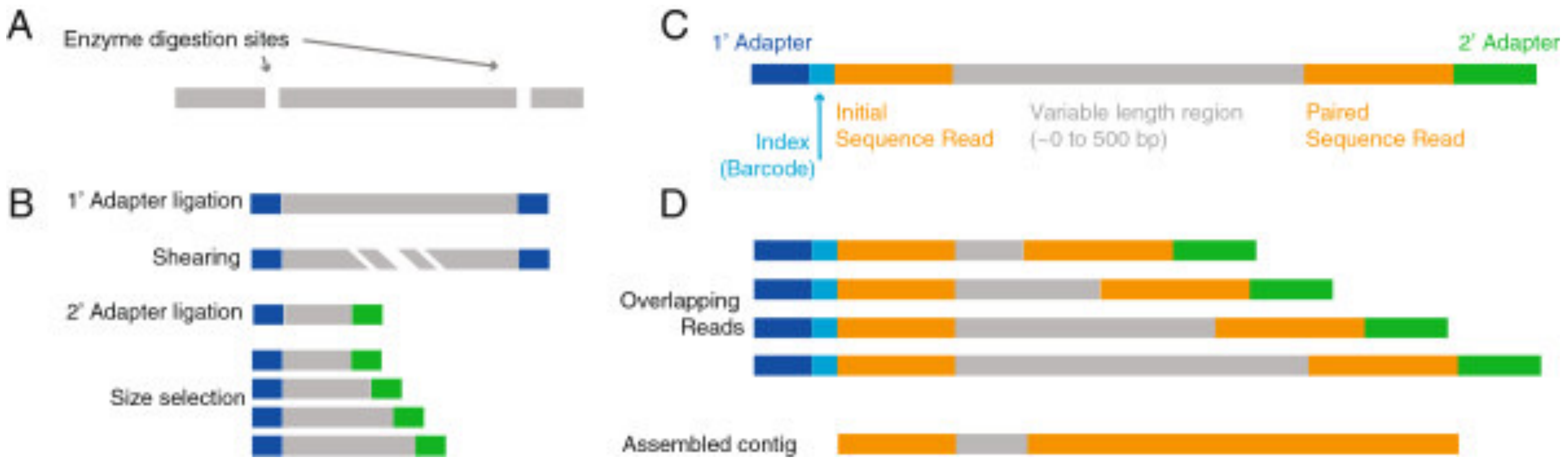
```
NC_042535.1     801     .       G       A       988     PASS    AN=976;AC=39    GT:PL:DP:SP:ADF:ADR:AD  0/0:0,27,239:9:0:2,0:7,0:9,0    0/0:0,45,255:15:0:7,0:8,0:15,0  0/0:0,36,255:12:0:5,0:7,0:12,0
```

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

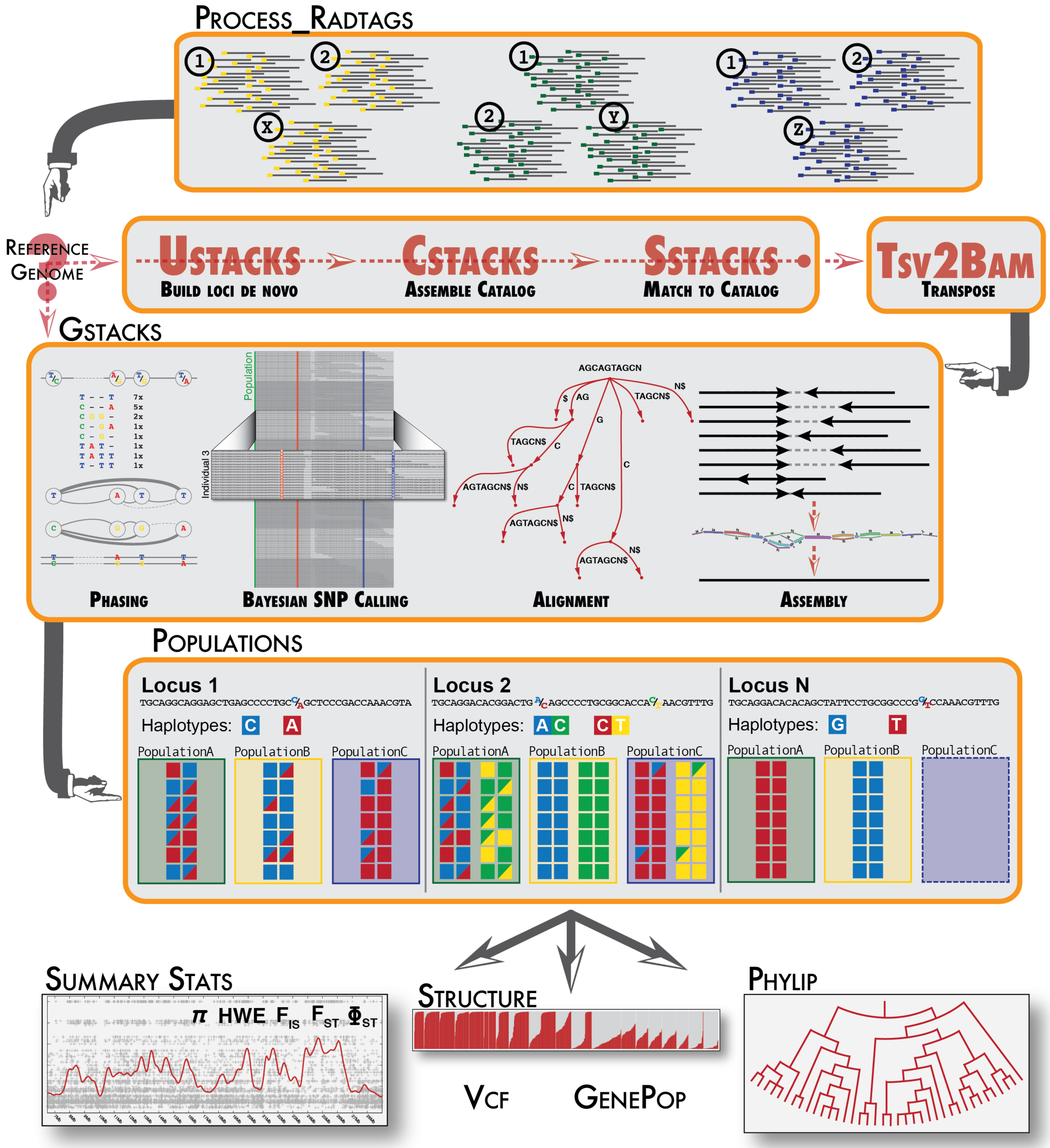
RADseq pipelines

- 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

Summary Stats

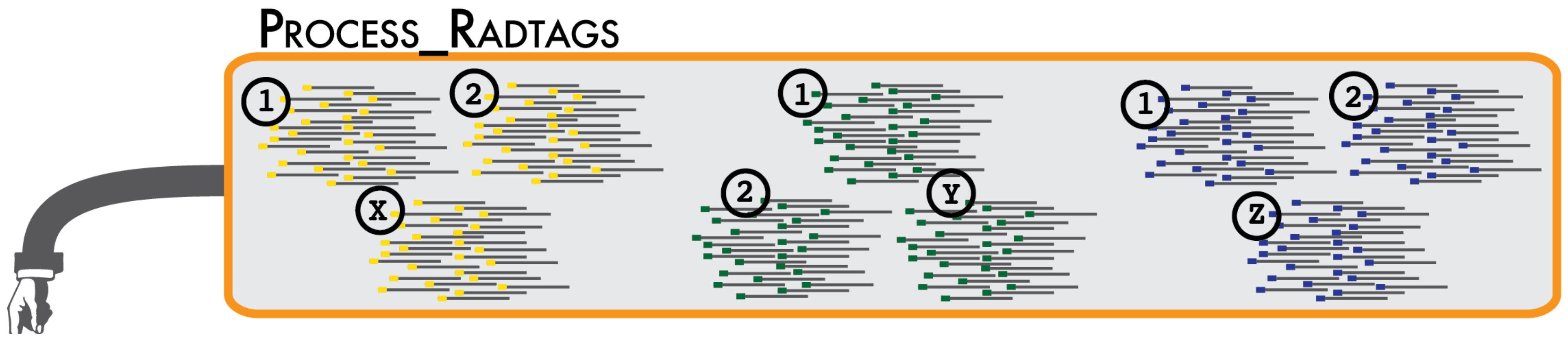
Structure

Phylip

Vcf

GenePop

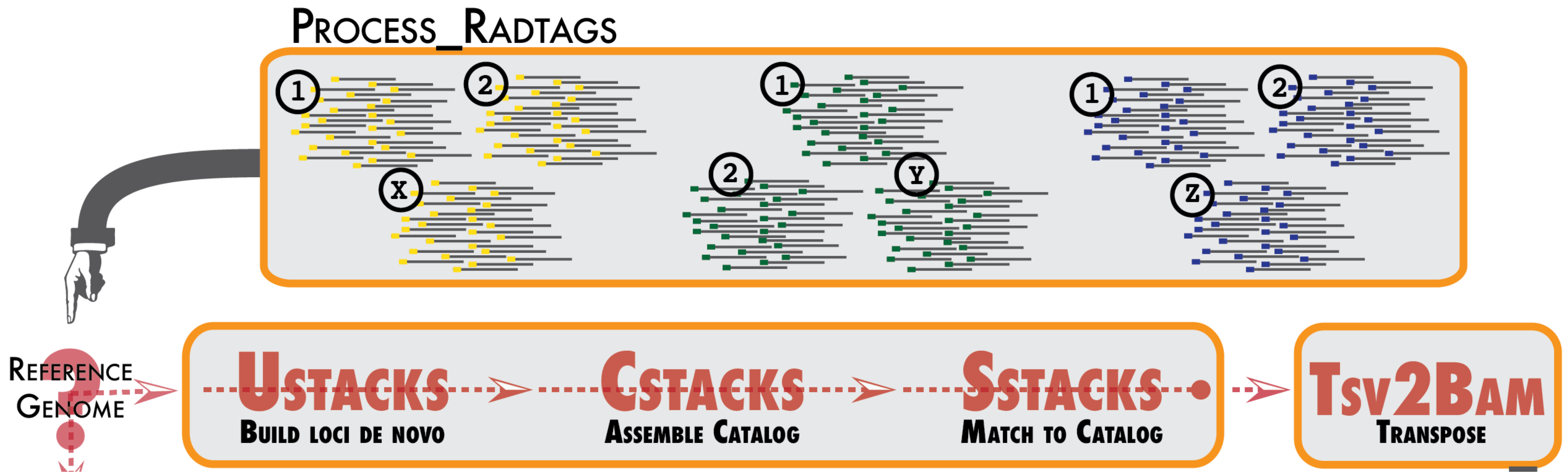
Stacks



To preprocess raw data

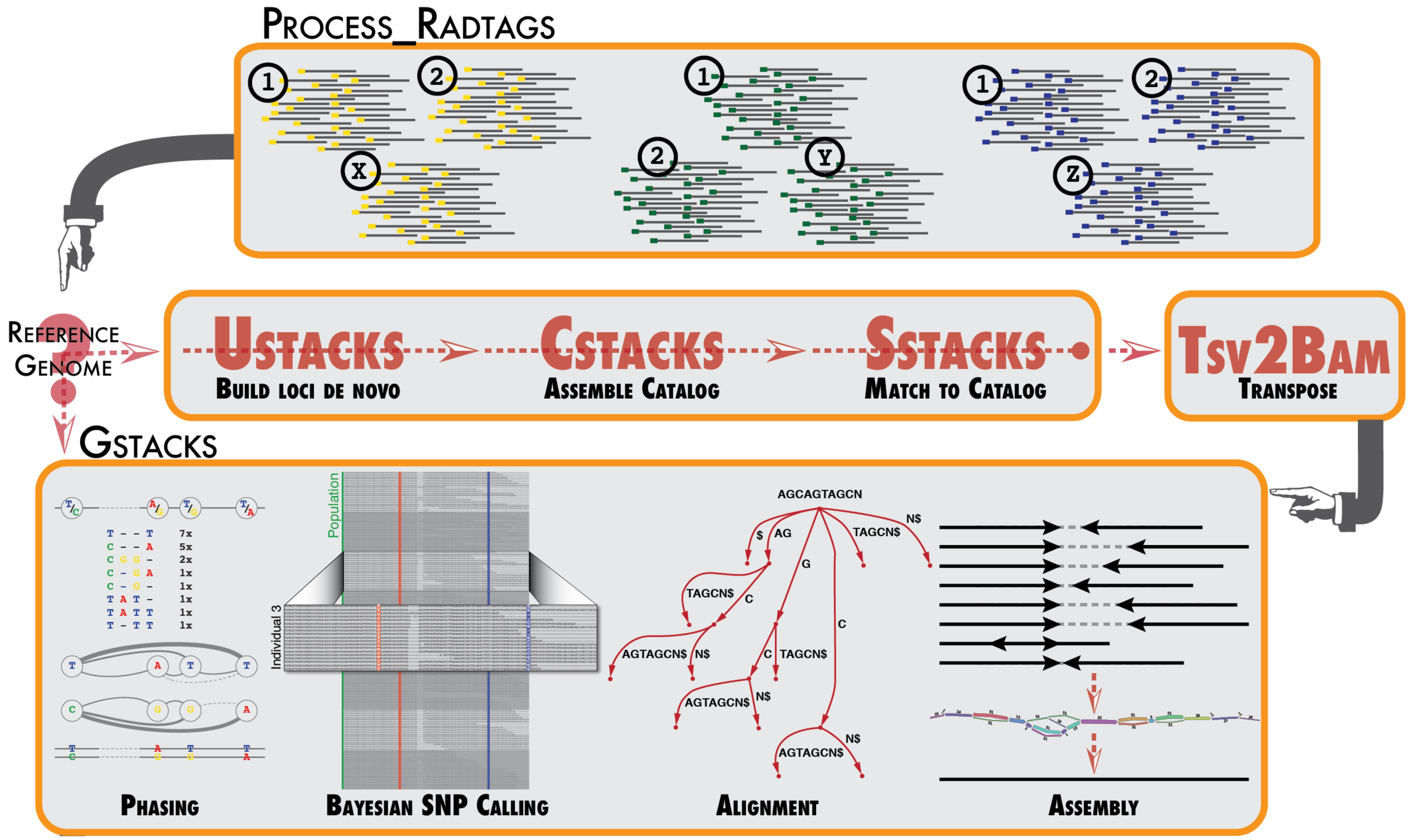
- Demultiplexing
- Adapter removal
- Quality filtering

Stacks



Loci assembly without reference genome

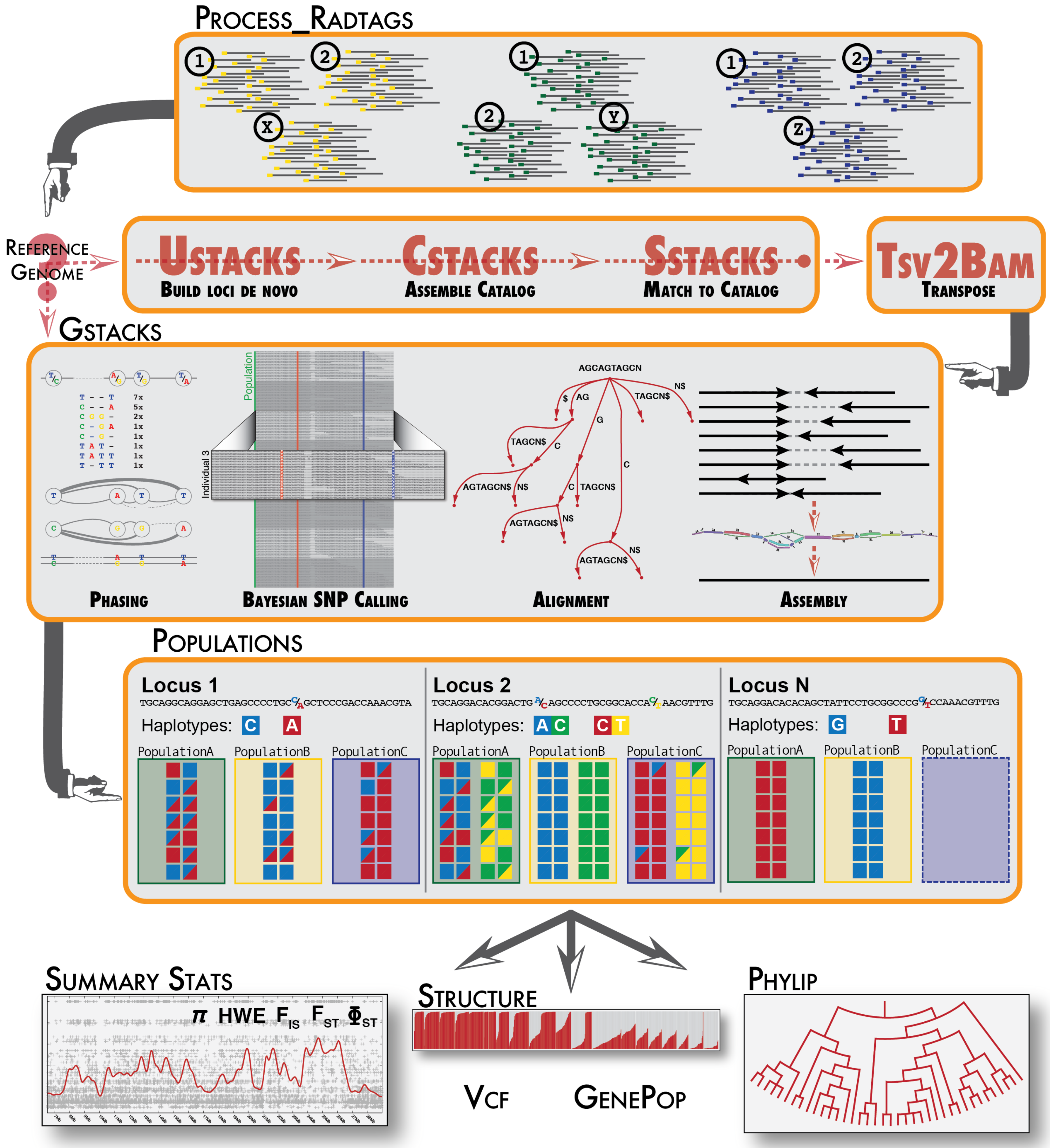
Stacks



If you have a reference genome, align RADseq data with external software

GSTACKS does different things according to data input but at the end it calls variants from assembled loci

Stacks


Process_Radtags
1
2
X
1
2
Y
1
2
Z
Reference Genome
Ustacks
Build loci de novo
Cstacks
Assemble Catalog
Sstacks
Match to Catalog
Tsv2Bam
Transpose
Gstacks
Population
Individual 3
AGCAGTAGCN
N$
$
AG
TAGCN$
G
TAGCN$
C
C
AGTAGCN$
N$
C
TAGCN$
AGTAGCN$
N$
N$
AGTAGCN$
Phasing
Bayesian SNP Calling
Alignment
Assembly
Populations
Locus 1
TGCAGGCAGGAGCTGAGCCCCTGCC/AGCTCCCGACCAAACGTA
Haplotypes: C A
PopulationA
PopulationB
PopulationC
Locus 2
TGCAGGACACGGACTGA/CAGCCCCTGCGGCACCAC/TAACGTTTG
Haplotypes: AC CT
PopulationA
PopulationB
PopulationC
Locus N
TGCAGGACACACAGCTATTCCTGCGGCCCGG/TCCAAACGTTTG
Haplotypes: G T
PopulationA
PopulationB
PopulationC
Summary Stats
π HWE F_{IS} F_{ST} Φ_{ST}
Structure
Vcf
GenePop
Phylip

Variant calling from whole genome data

Most commonly used software for variant calling

Low coverage whole genome data (<= 5X)

- Genotype likelihoods
- ANGSD (genotype uncertainty)

Moderate to high coverage data (> 5X)

- Genotype quality
- bcftools mpileup
- GATK

Variant calling from whole genome data

Most commonly used software for variant calling

Low coverage whole genome data (<= 5X)

- Genotype likelihoods
- ANGSD (genotype uncertainty)

Moderate to high coverage data (> 5X)

- Genotype quality
- bcftools mpileup -> we will use in the tutorial
- GATK