



Reduced representation libraries/RAD Introduction

Niklaus Zemp
30 June 2021

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

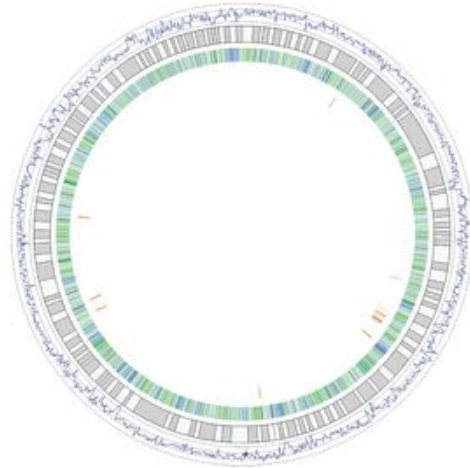
Non-model organisms



Genomics of large and unexplored genomes



Assembly

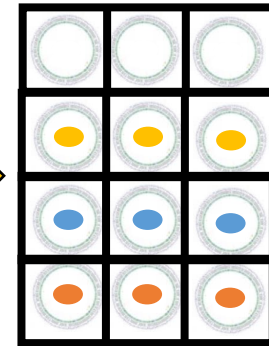
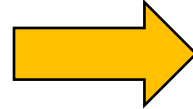
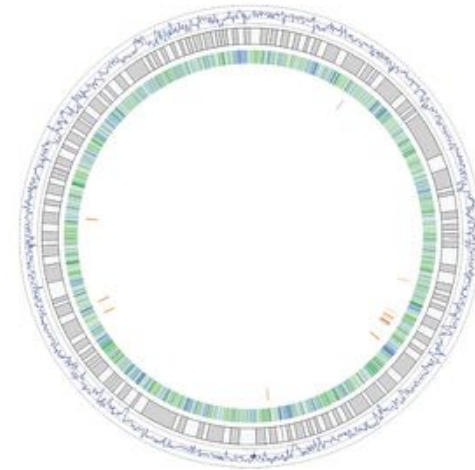


draft genome

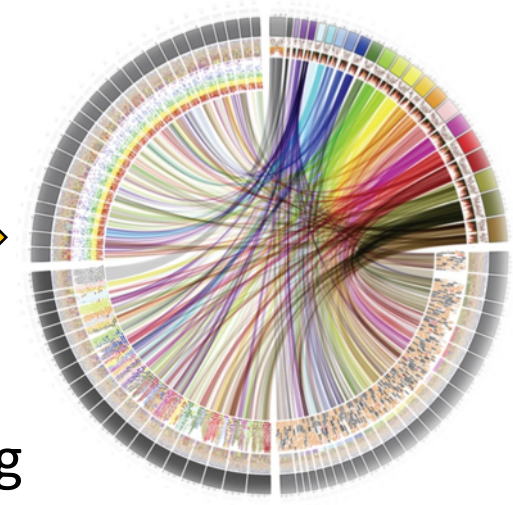
Genomics of large and unexplored genomes



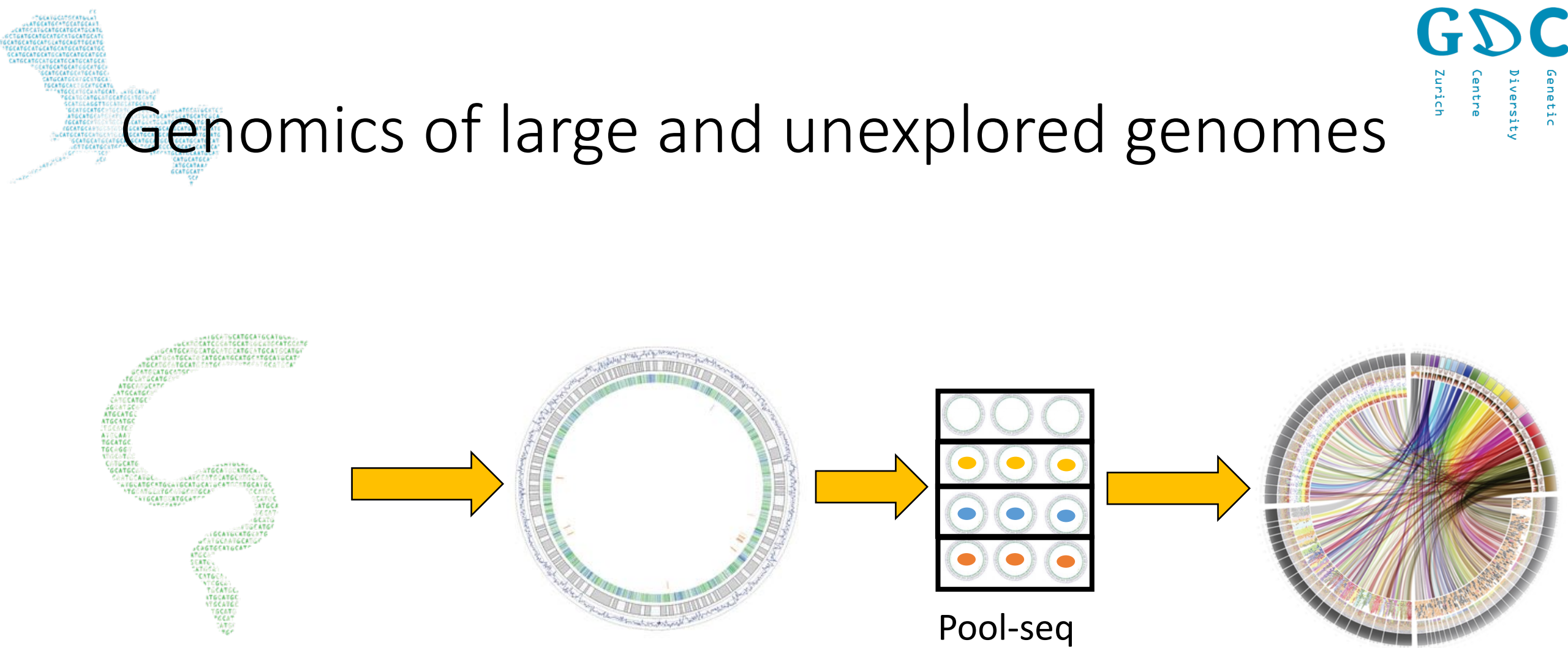
Assembly



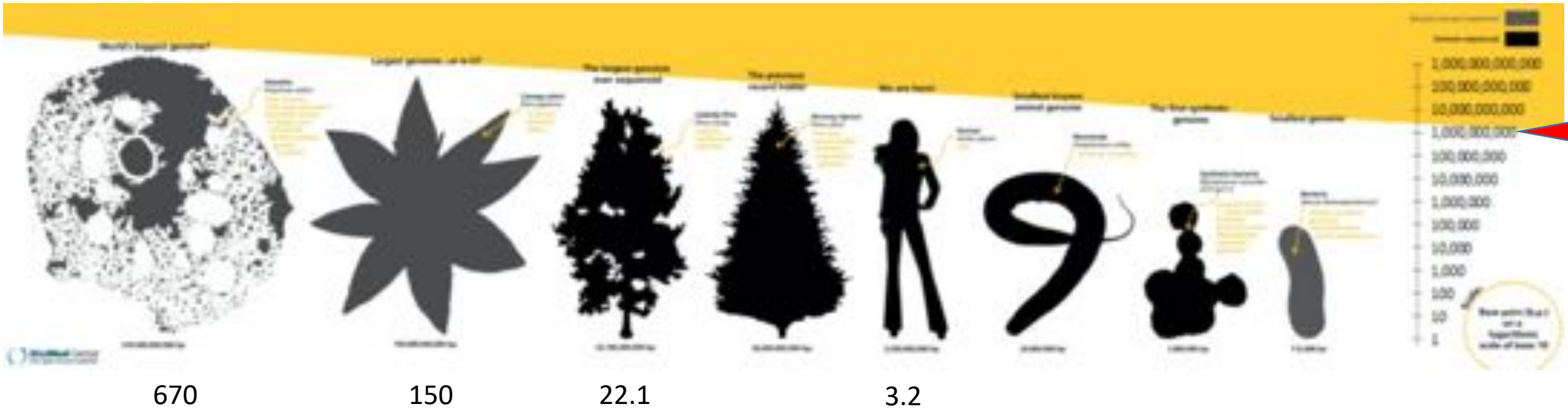
(low coverage) resequencing



Genomics of large and unexplored genomes



Genomes can be large

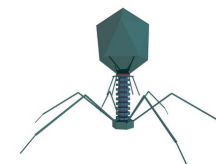
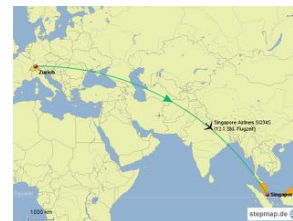


Genomes can be large



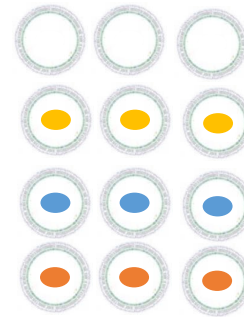
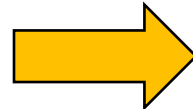
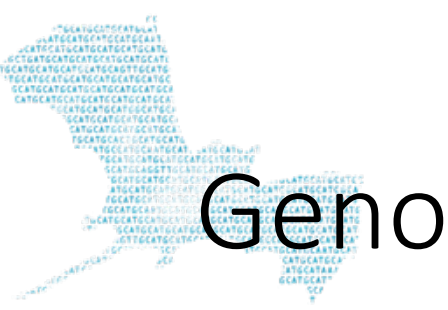
>1 year

2 weeks

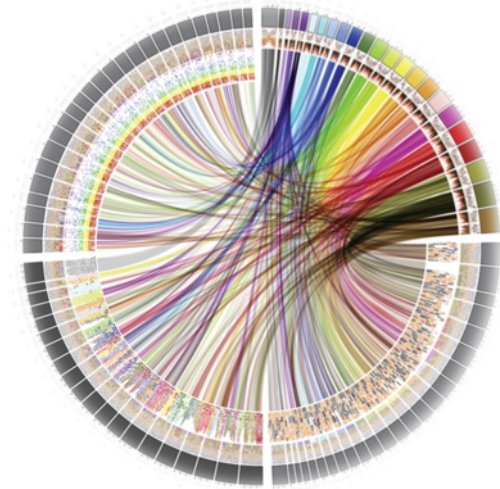


1 million bases is 1 minute

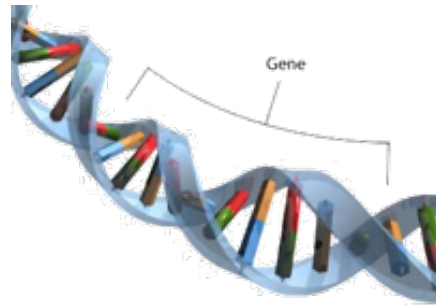
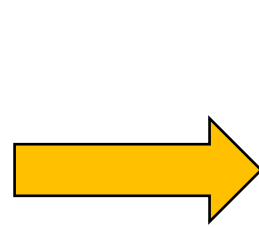
Genomics of large and unexplored genomes



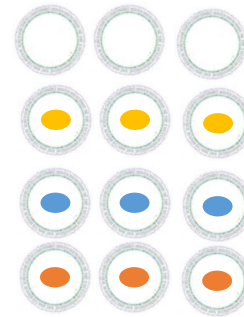
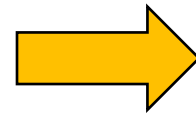
RNA-seq



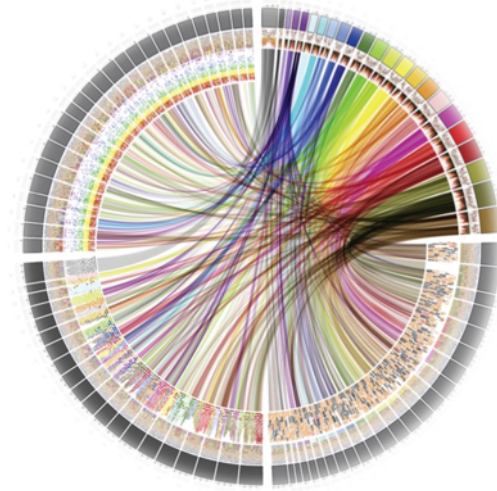
Genomics of large and unexplored genomes



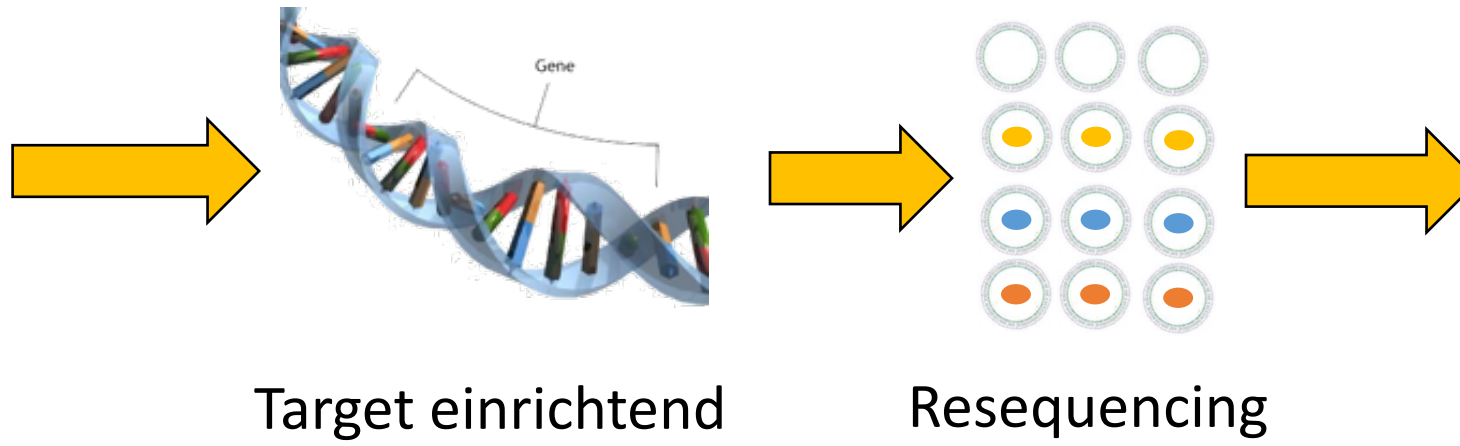
Target einrichtend



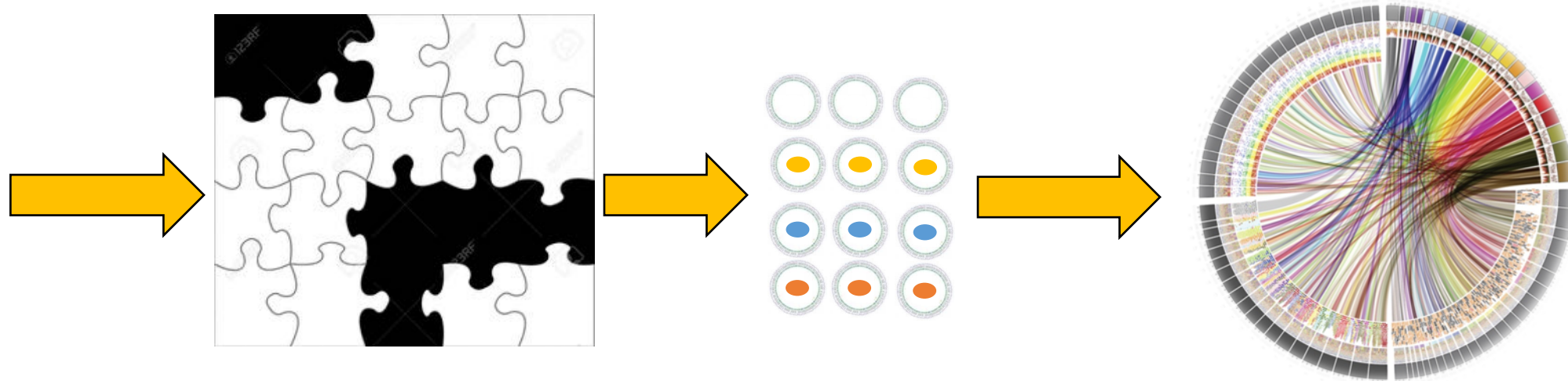
Resequencing



Genomics of large and unexplored genomes

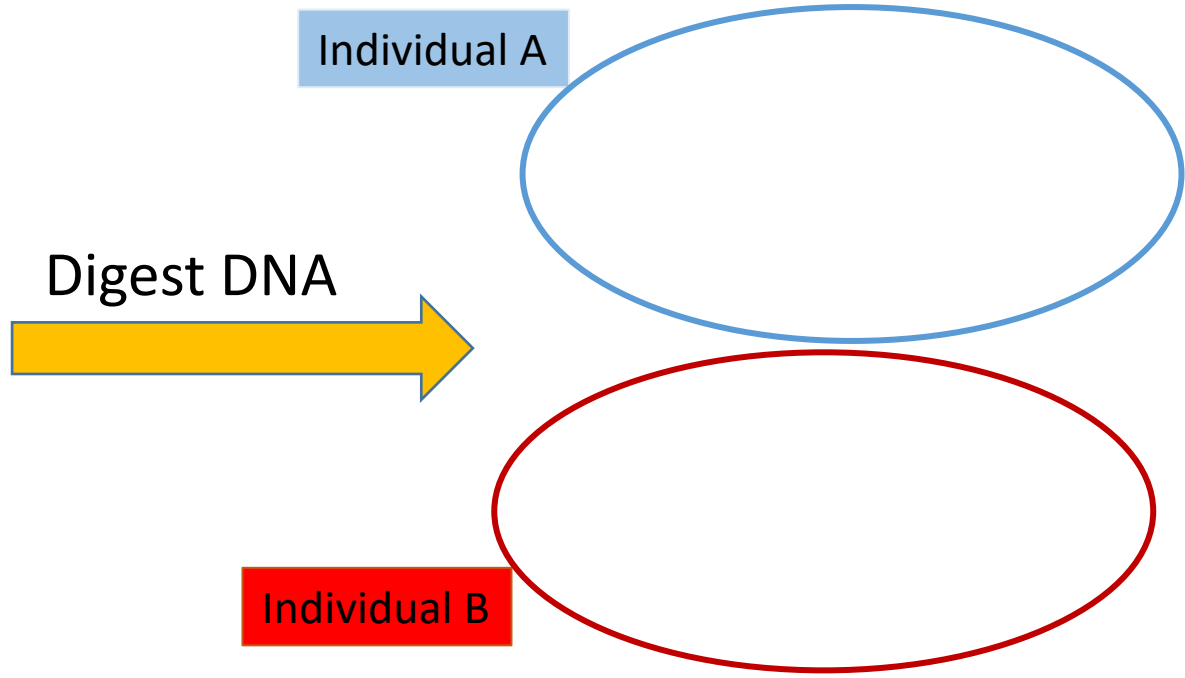


Genomics of large and unexplored genomes

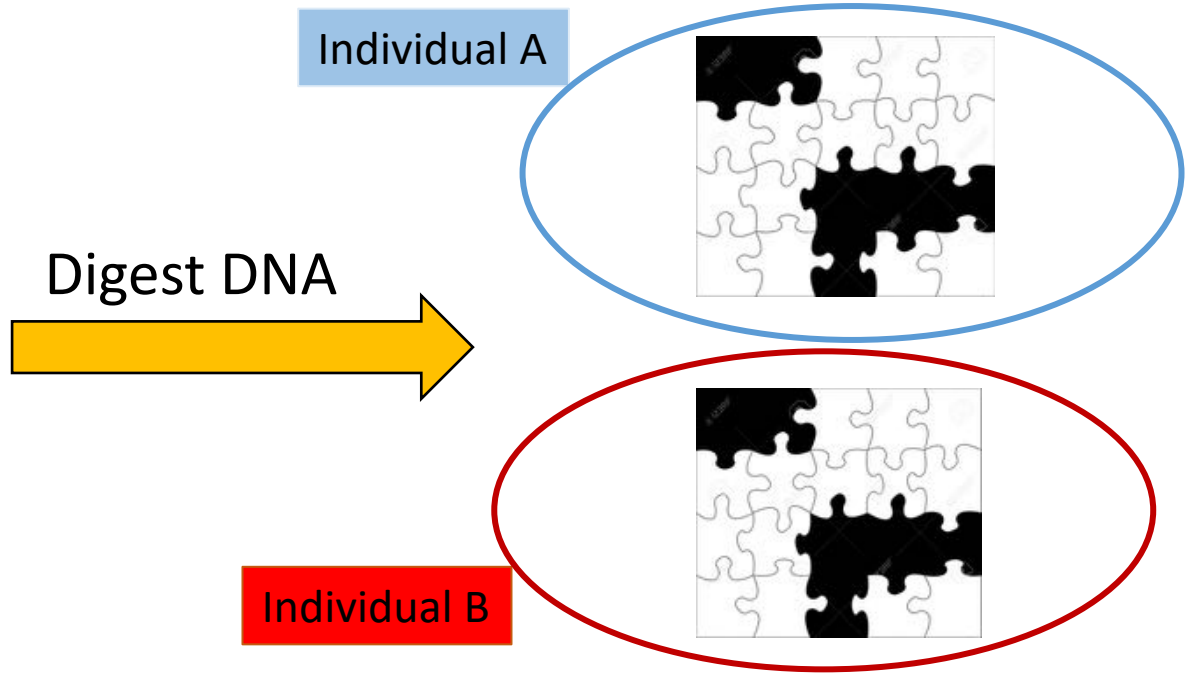


Reduced representation libraries/
Restriction site associated **D**NA (RAD)

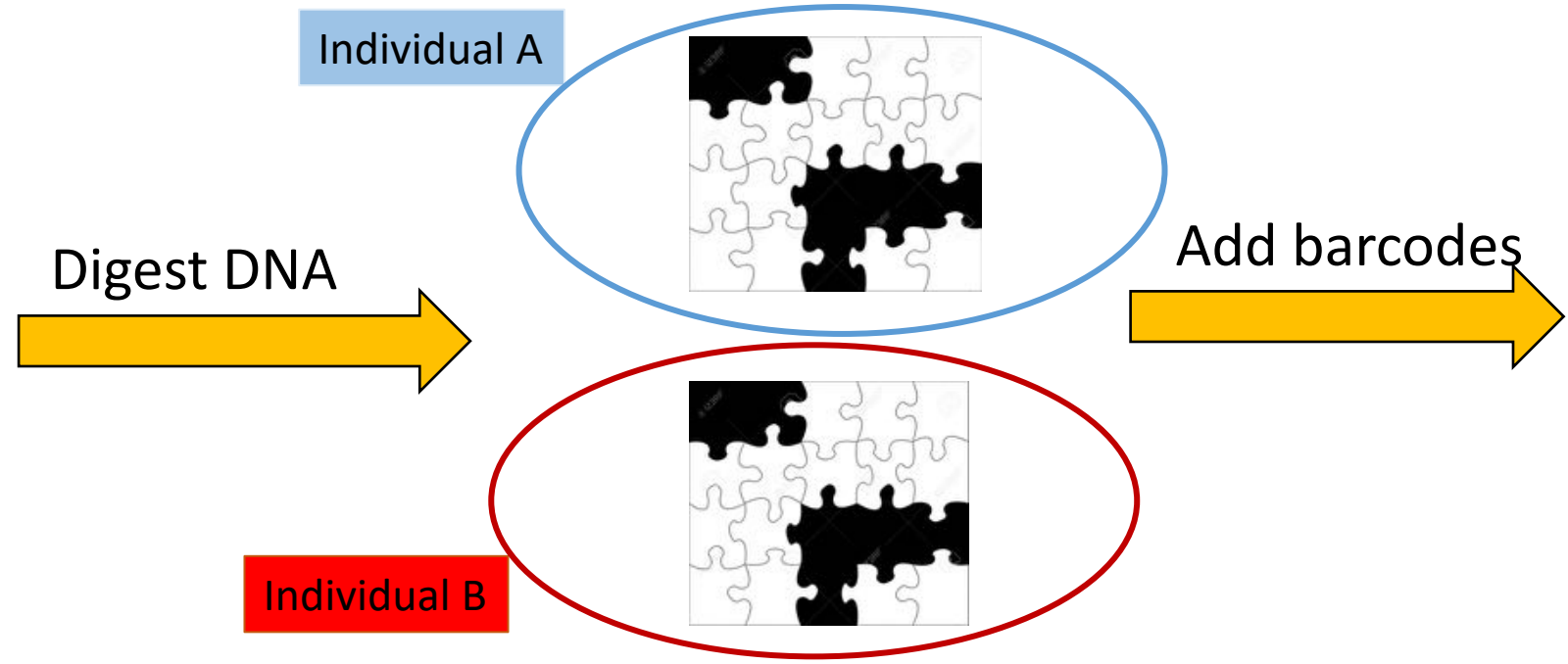
Reduced representation libraries (RAD)



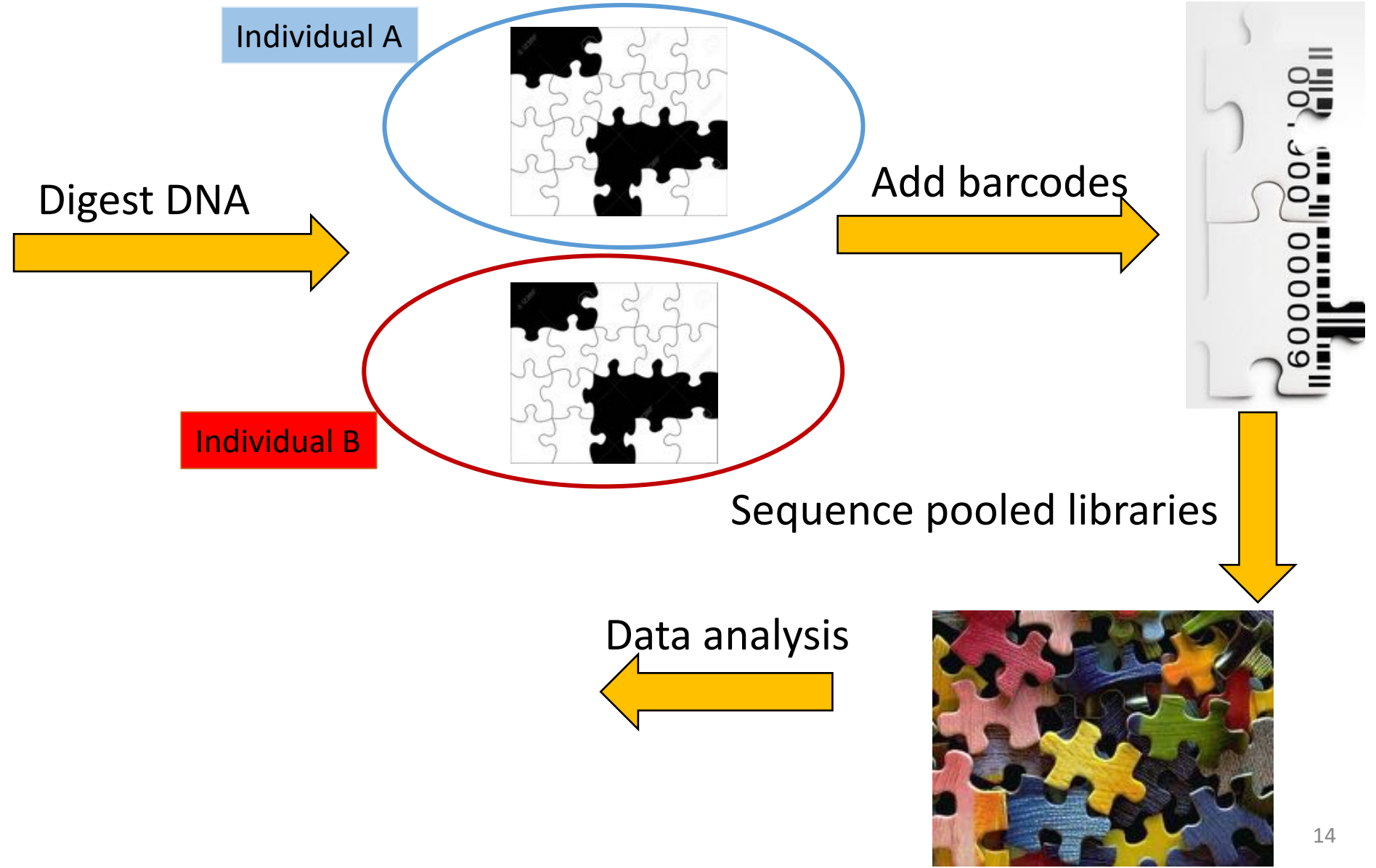
Reduced representation libraries (RAD)



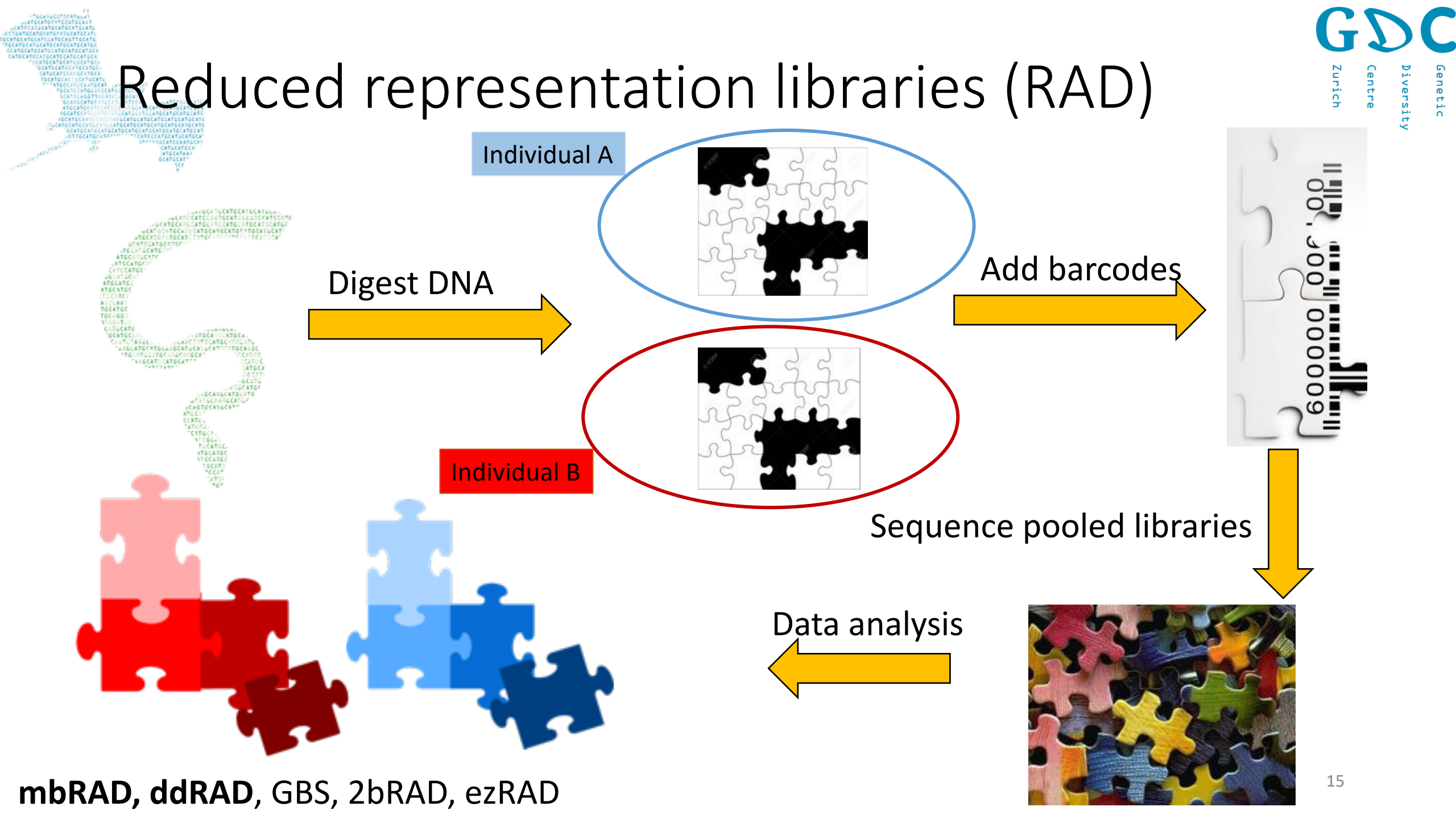
Reduced representation libraries (RAD)



Reduced representation libraries (RAD)



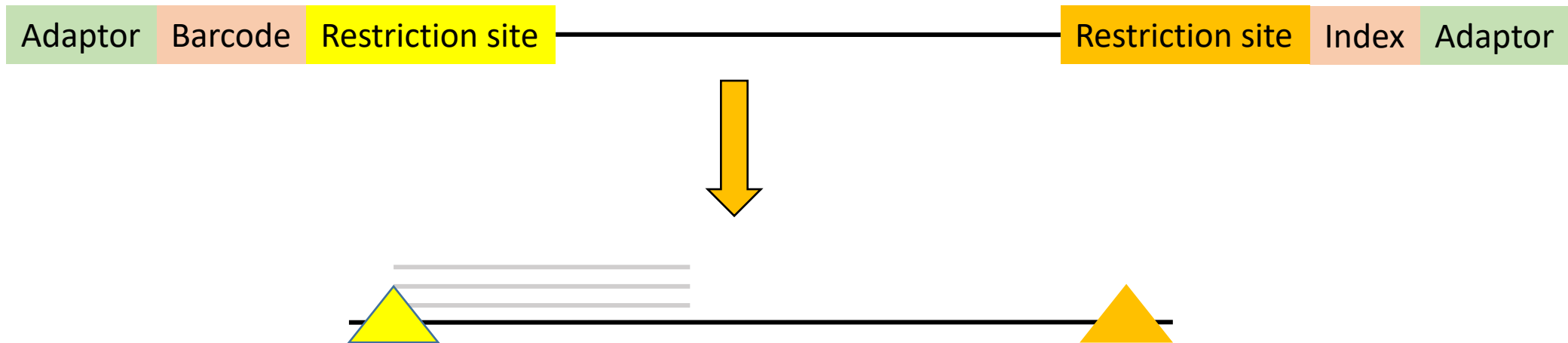
Reduced representation libraries (RAD)






- Digestion with one rare cutting enzyme
- Barcoded adapters ligated to fragments
- Ligated fragments are then sonicated
- Size selection is used to reduce sampled genome
- Paired-end reads

ddRAD



- Digestion with one rare and one common cutter
- Barcoded adapters ligated to fragments
- Size selection is used to reduce sampled genome
- Often single-end reads



mbRAD versus ddRAD

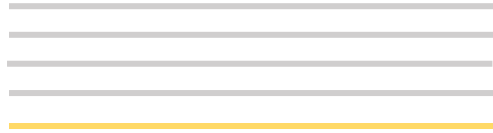
mbRAD

ddRAD

Costs

higher

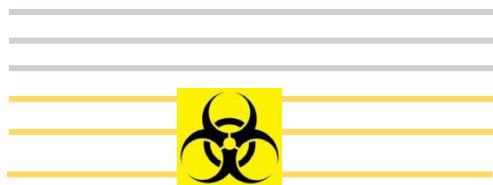
lower



PCR duplicates

can be detected

not possible



Allele dropouts

reduced

increased



How many fragments do I get?

- Restriction enzyme
- \propto Genome size
- \propto Sequencing depth

Genome available:

In silico digest (simRAD)

de novo:

Predictions are possible based on the concentration but a test run is often needed

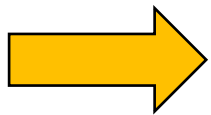




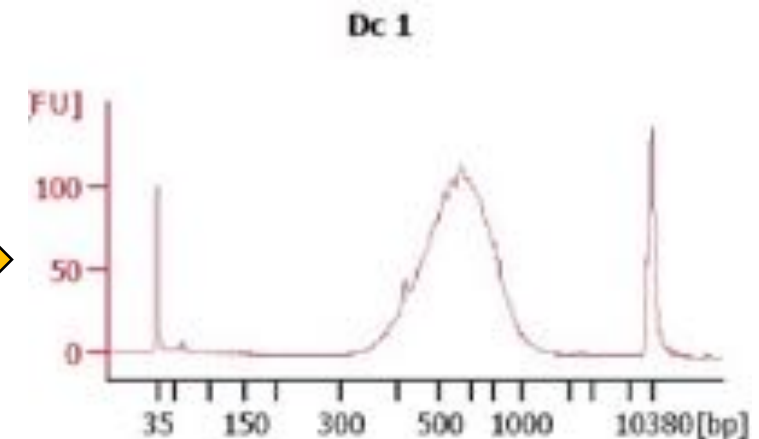
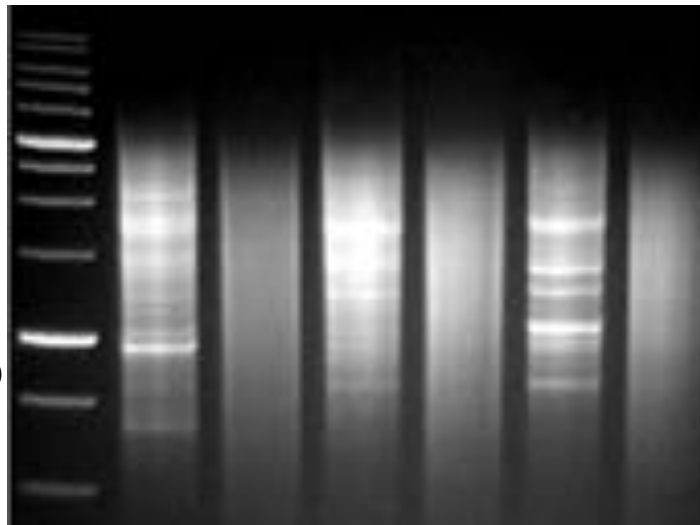
Can I produced these libraries at the GDC?

Primers sets and protocols for producing RAD and ddRAD libraries are available

200-500 ng high
quality DNA



500 bp

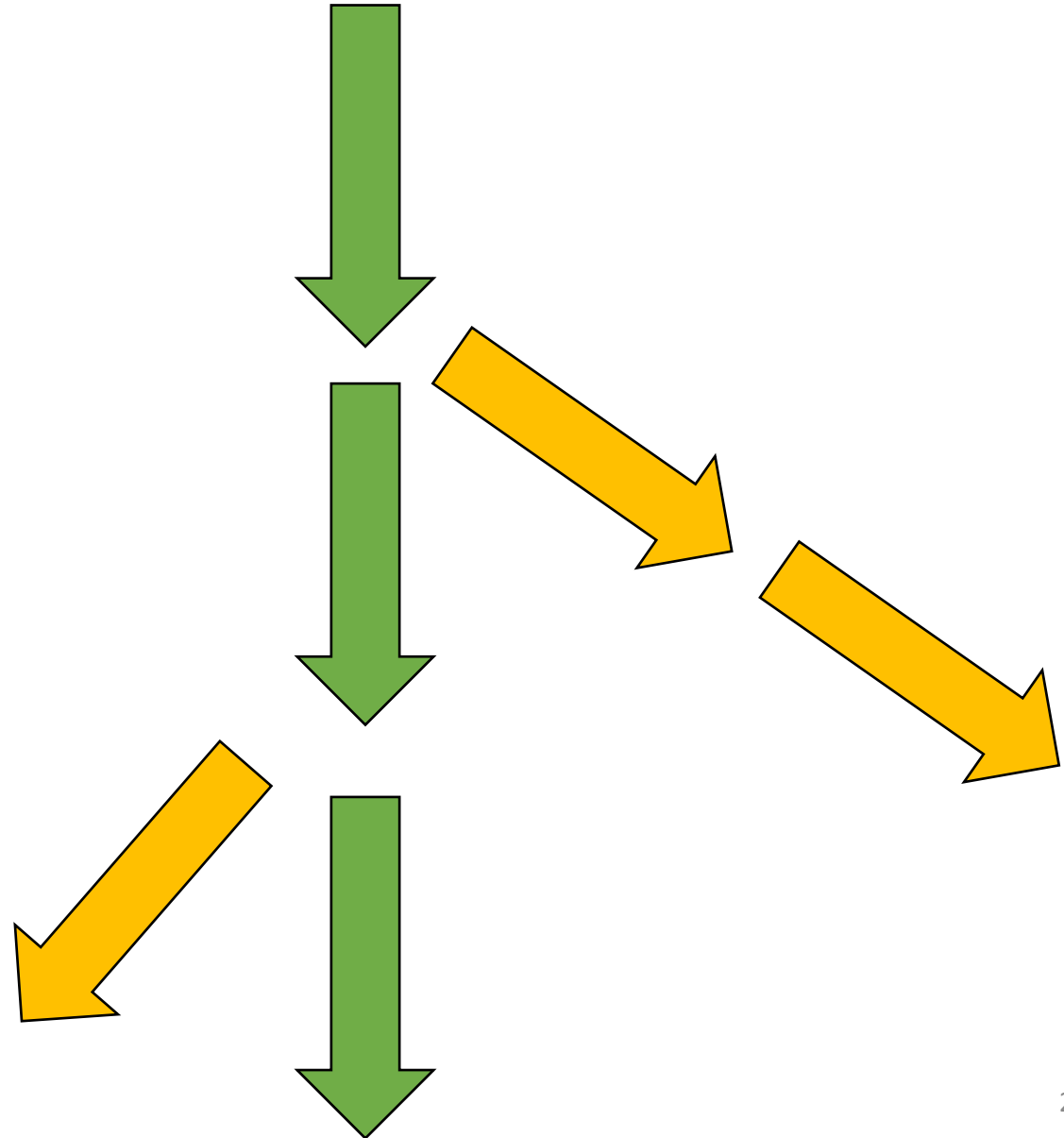




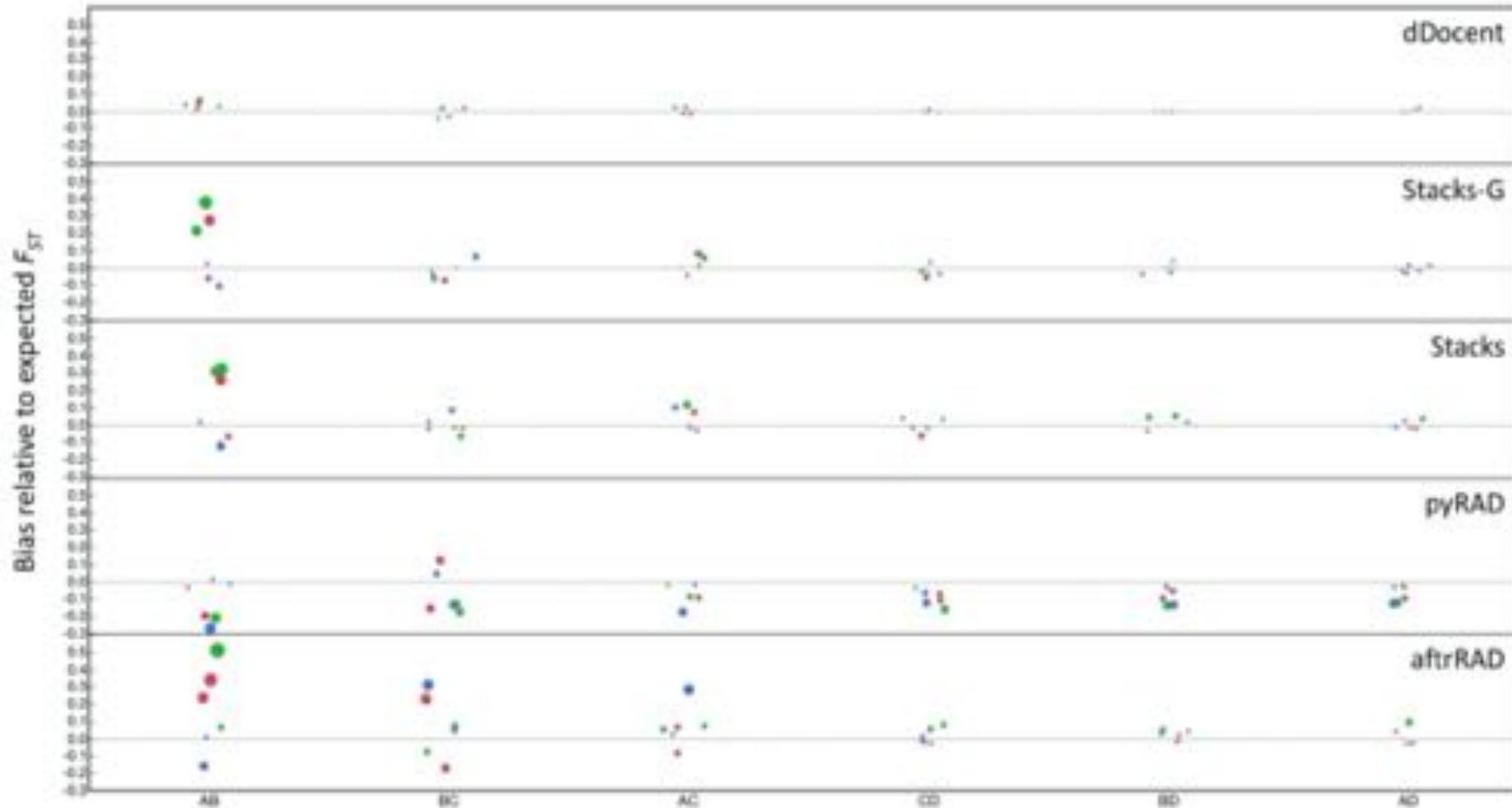
RAD Analysis

Pipelines

- Stacks (Catchen et al. 2013)
- dDocent (Puritz et al. 2014)
- pyRAD (Eaton 2014)
- aftrRAD (Sovic et al. 2015)

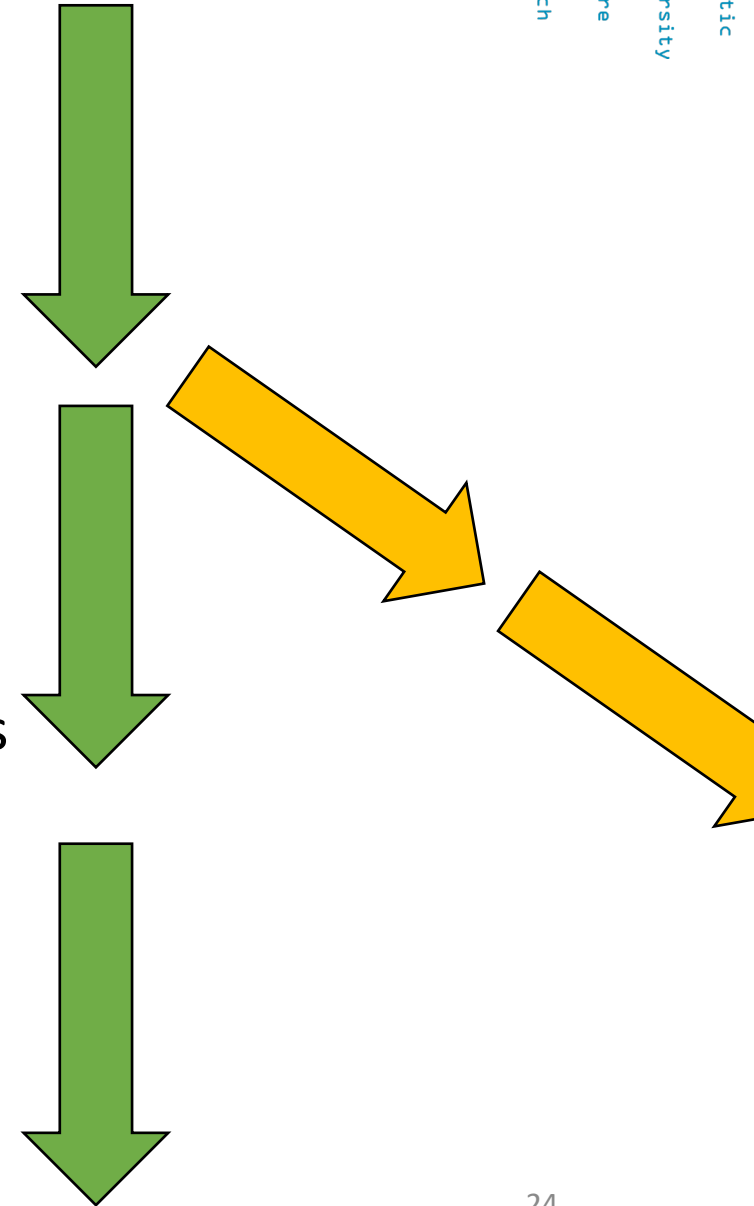


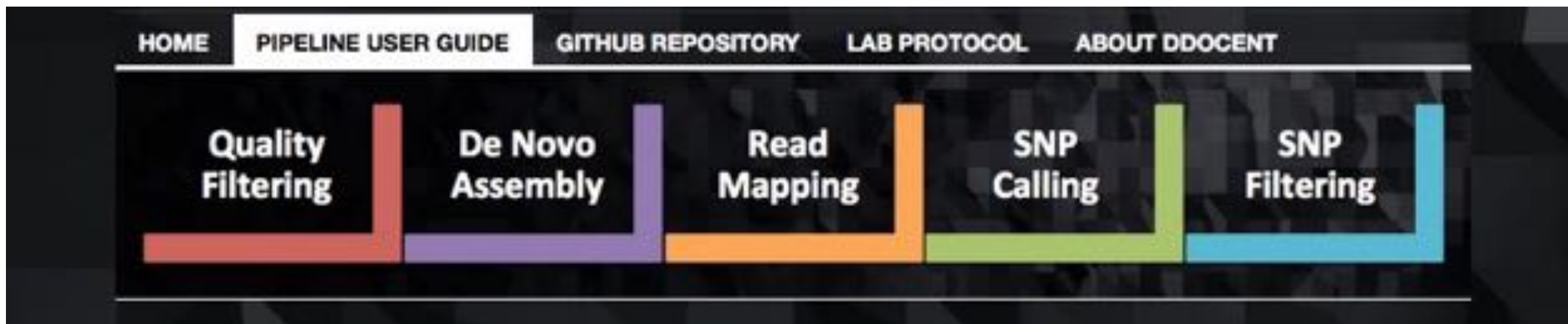
Population differentiation bias of different pipelines



Pipelines

- Stacks (Catchen et al. 2013)
- **dDocent (Puritz et al. 2014)**
 - Can handle Indels
 - Simple customizable backbone for bioinformatics
- pyRAD (Eaton 2014)
 - Can handle many RADseq types, focused on phylogenetics



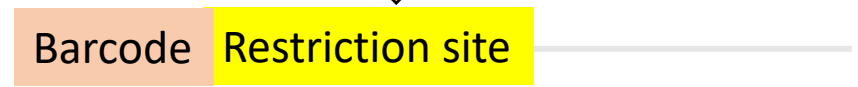
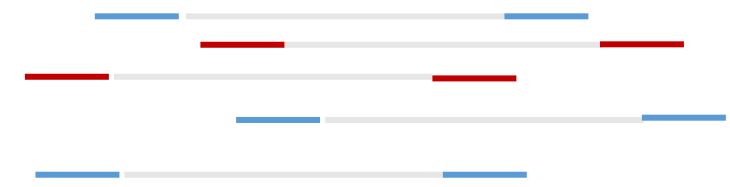
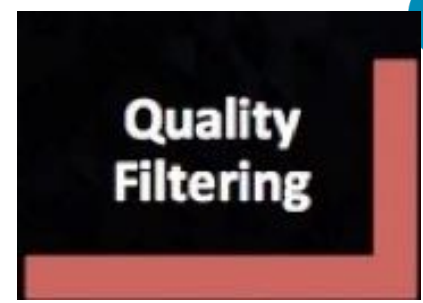


Tutorials: <https://github.com/jpuritz/dDocent>

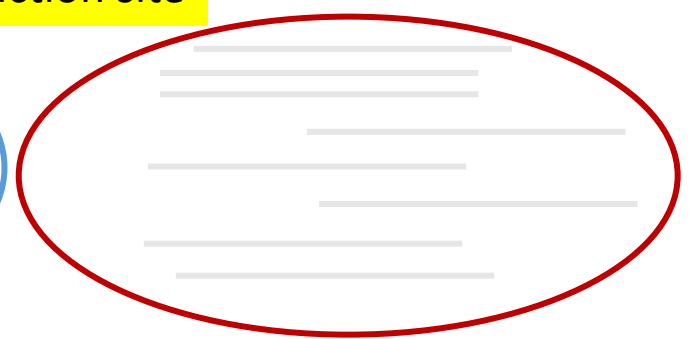
Demultiplexing and quality filtering

- Demultiplex reads
process_radtags (Stacks)

- Remove adaptors and low quality bases
Trimmomatic/fastp

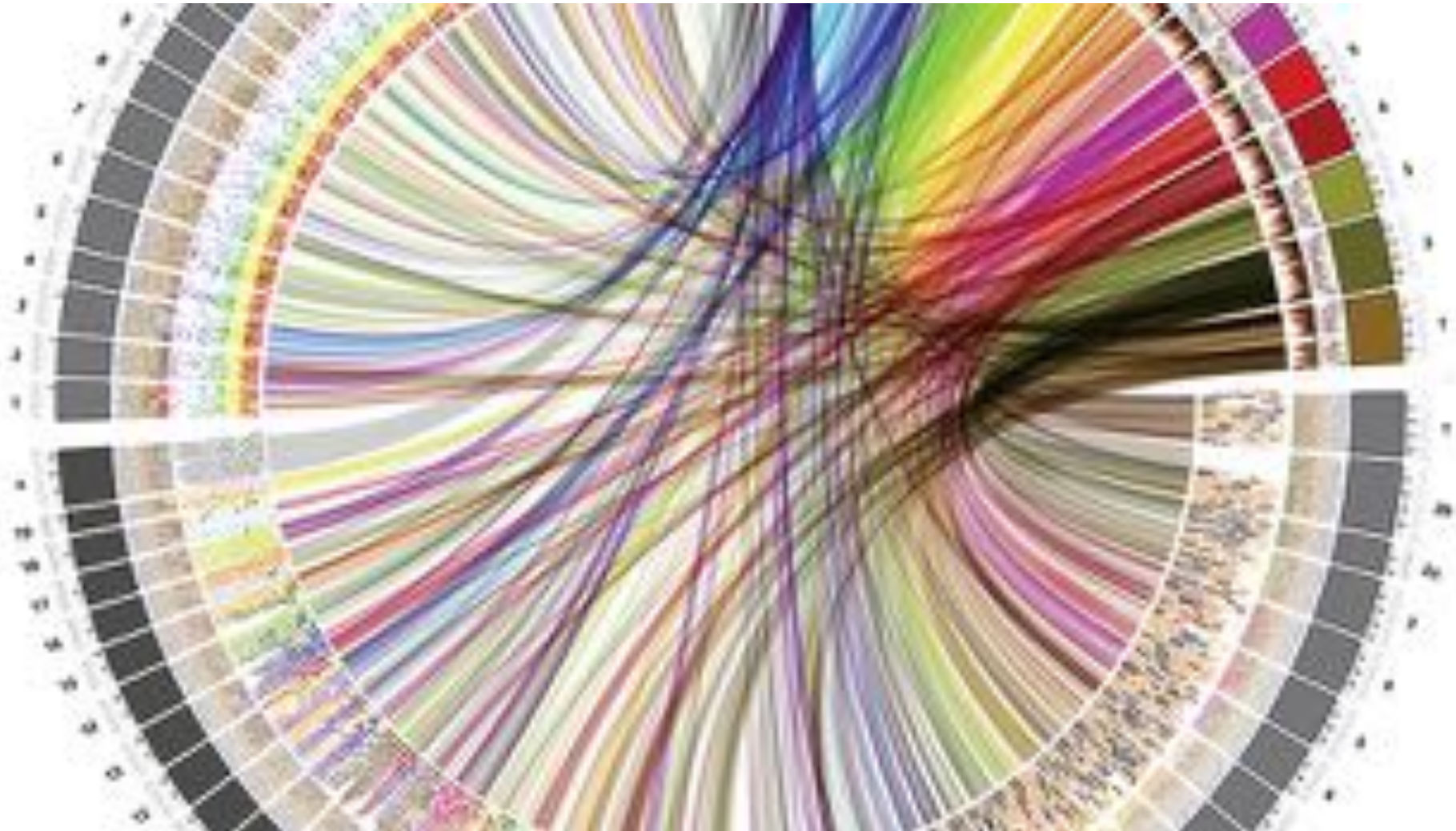


Individual A



Individual B

Reference assembly



De novo assembly

Merge reads in case of overlaps

PEAR

Remove all identical reads

Pool all individuals together

customized scripts

Single-end:

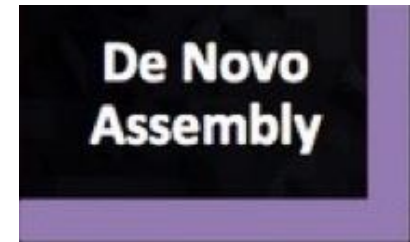
Cluster the non-redundant sequences based on similarity

cd-hit-est

Paired-end:

Assembly the non-redundant sequences and then using paired-end information

rainbow, cd-hit-est



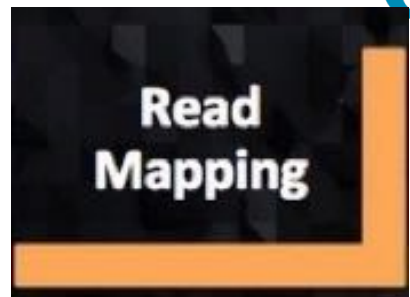
Locus 1

Locus 2

Locus 3

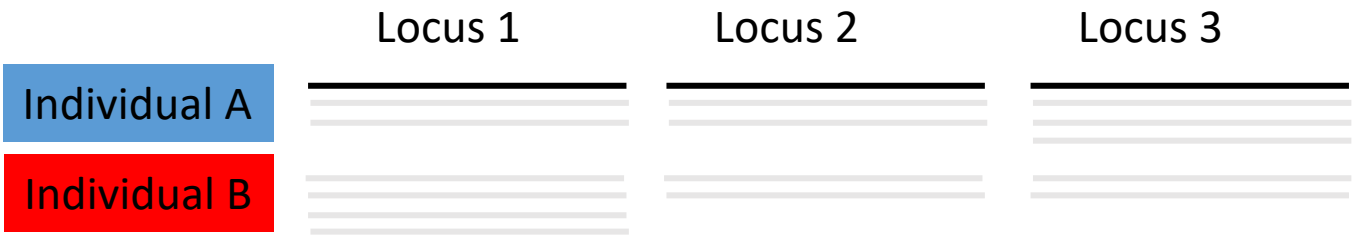


Read mapping

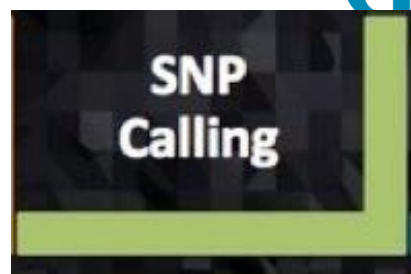


- Mapping reads against the reference catalogue

BWA



SNP calling



FreeBayes

Locus 1

Locus 2

Locus 3

Individual A

AATGCAGGG
AATGCAGGG
AATGCAGGG

AATGCTGGGA
AATGCAGGGA
AATGCTGGGA

AATGCTTGGGA
AATGCTAGGGA
AATGCTTGGGA

Individual B

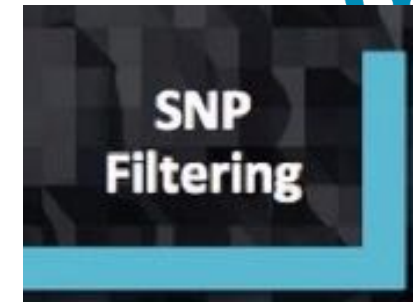
AATGCTGGGA
AATGCTGGGA
AATGCTGGGA

AATGCTGGGA
AATGCTGGGA
AATGCTGGGA

AATGCT GGGA
AATGCT GGGA
AATGCT GGGA



SNP filtering



Filter only for good SNPs
VCFTools, vcfliib

Criteria:

Mapping quality

Coverage

Missing genotypes

Minor allele frequency

Balanced alleles



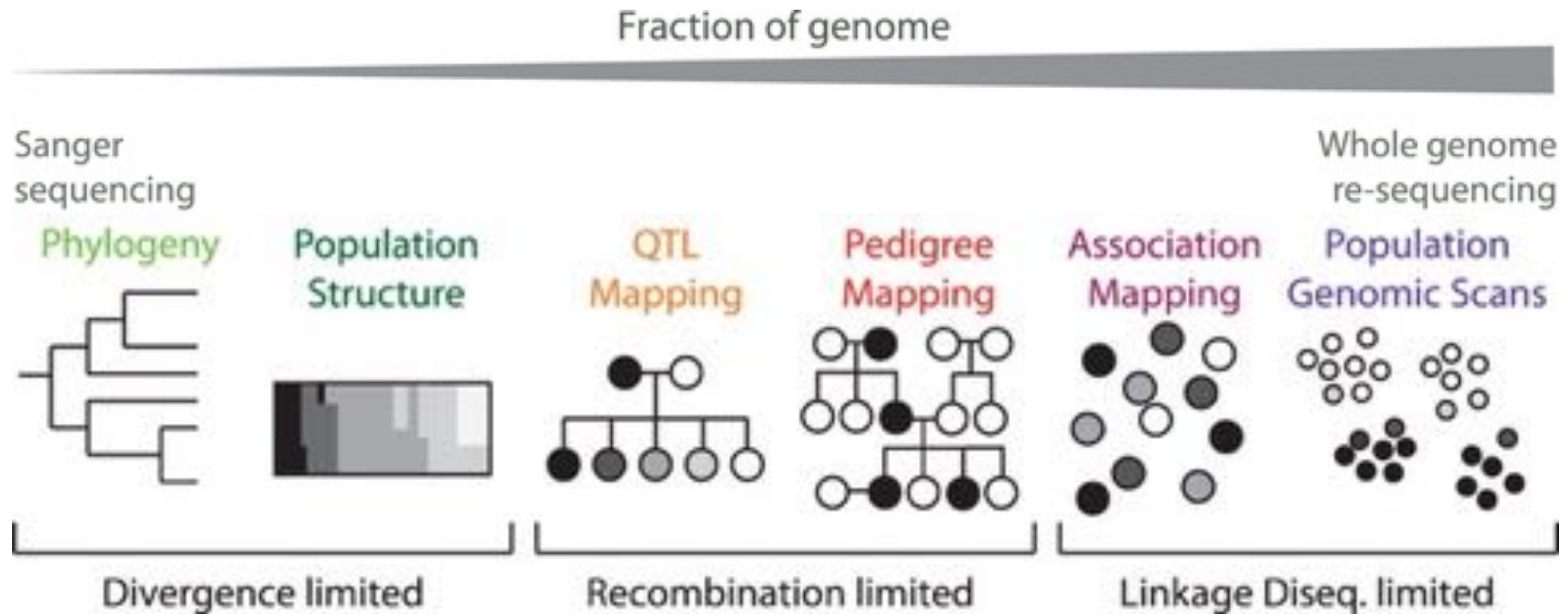


RAD

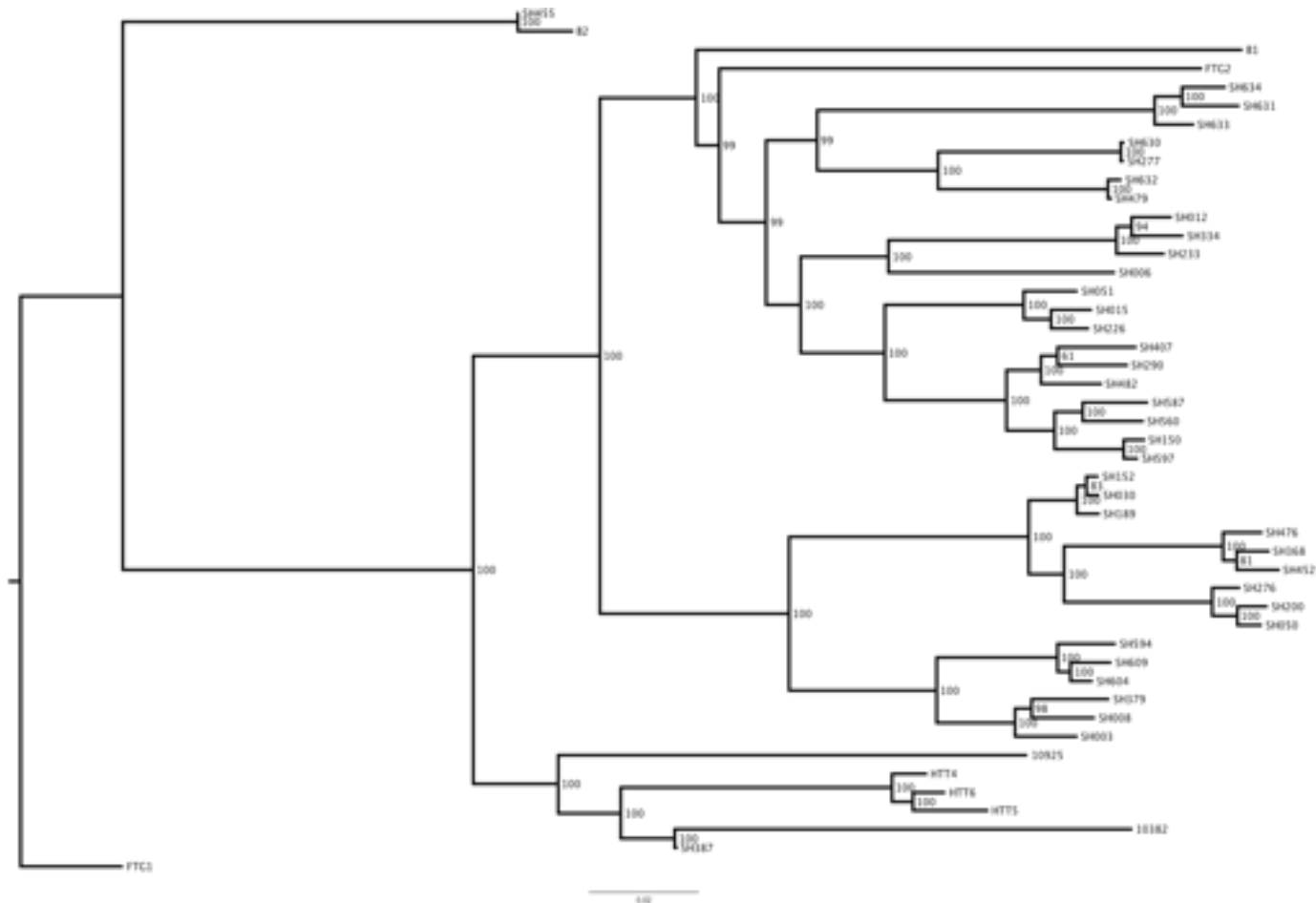
Applications



What is it good for?

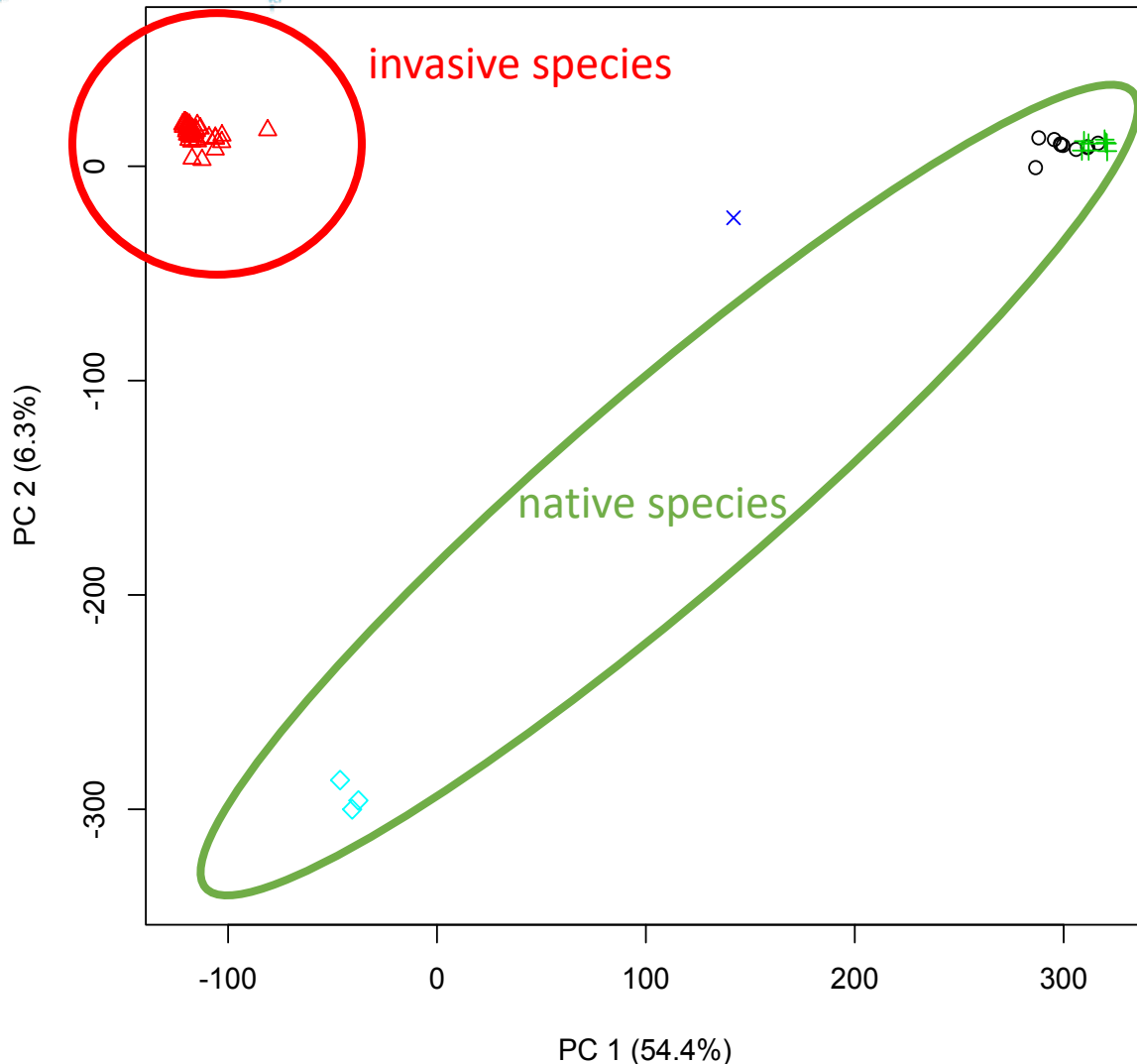


Phylogeny of rosewood species from Madagascar



Better resolution with ddRAD compared to traditional barcoding markers

Population structure

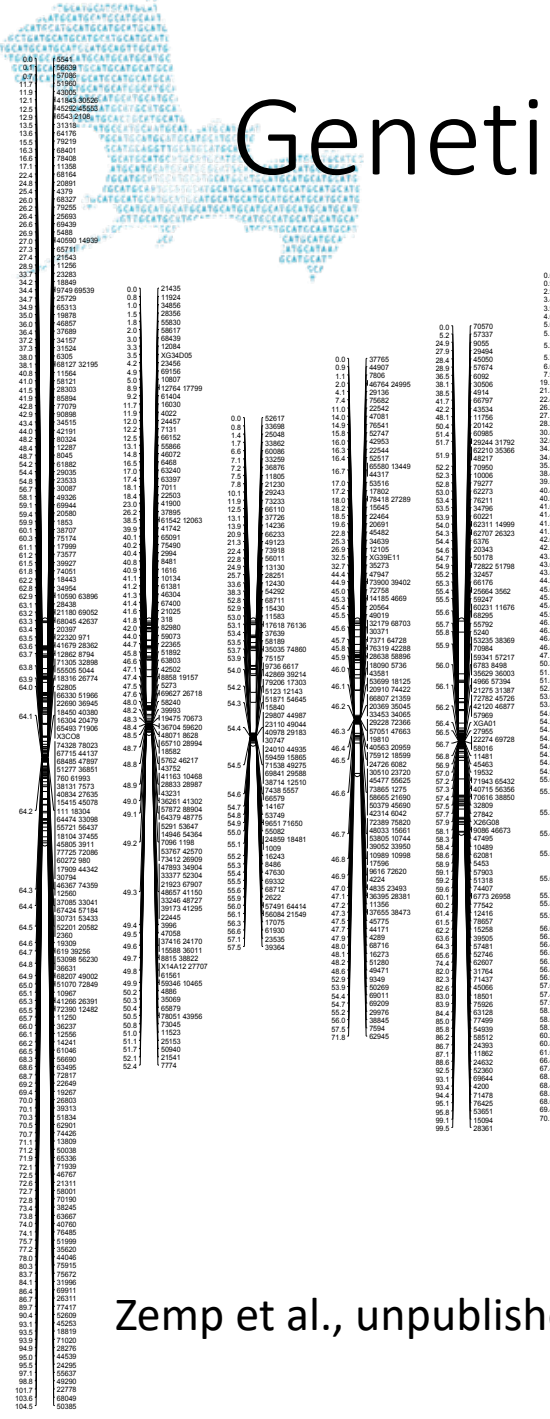


No hybridization between invasive and native groundsel species



- △ inaequidens
- erucifolius
- × not identified
- ◇ vulgaris
- + jacobaea

Genetic linkage mapping



A More recombination events between A and C than A and B

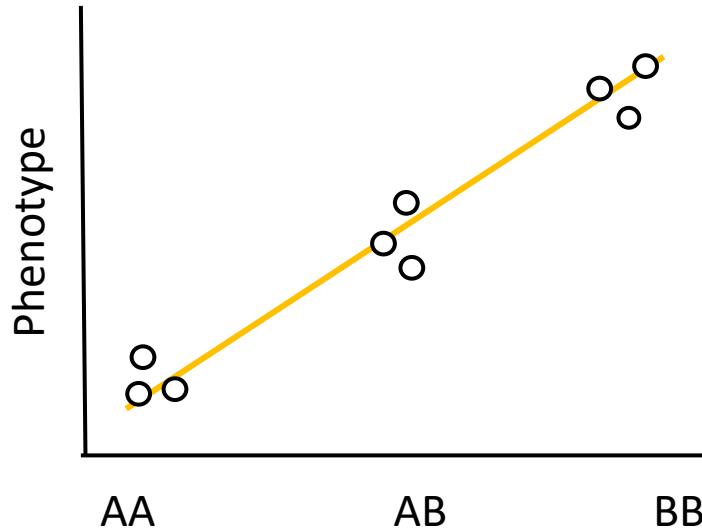
Fior et al., unpublished

Ordering the fragments using recombination rates in the offspring

- Distribution of fragments across the genome
- Scaffolding and comparative analyses of draft genomes
- Estimate recombination rate across the genome

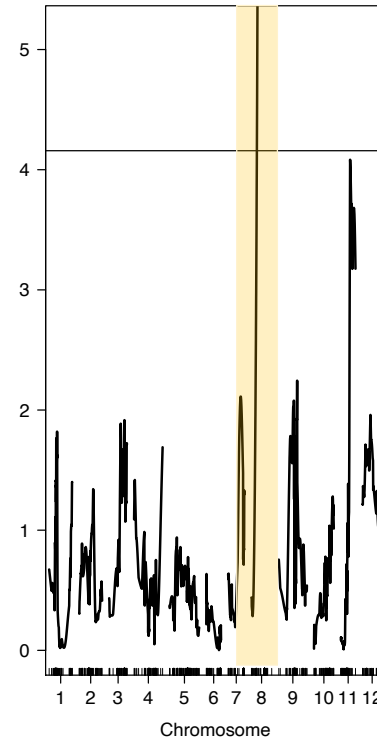
Zemp et al., unpublished

Correlations between phenotype and genotype (QTL, GWAS)



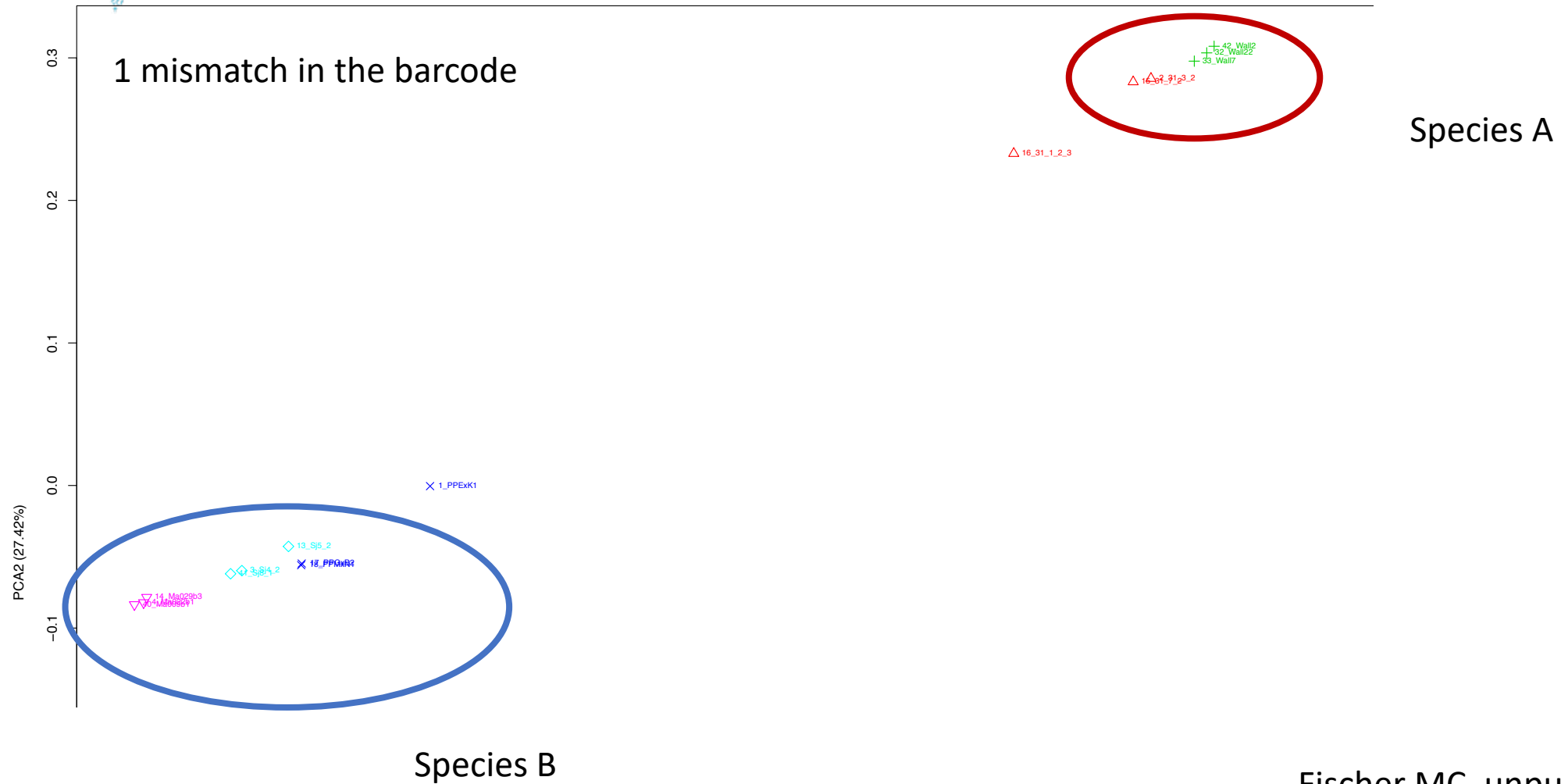
Looking for loci that are highly correlated with a certain phenotype

Plant height



Crameri, Nenadic and Zemp, unpublished

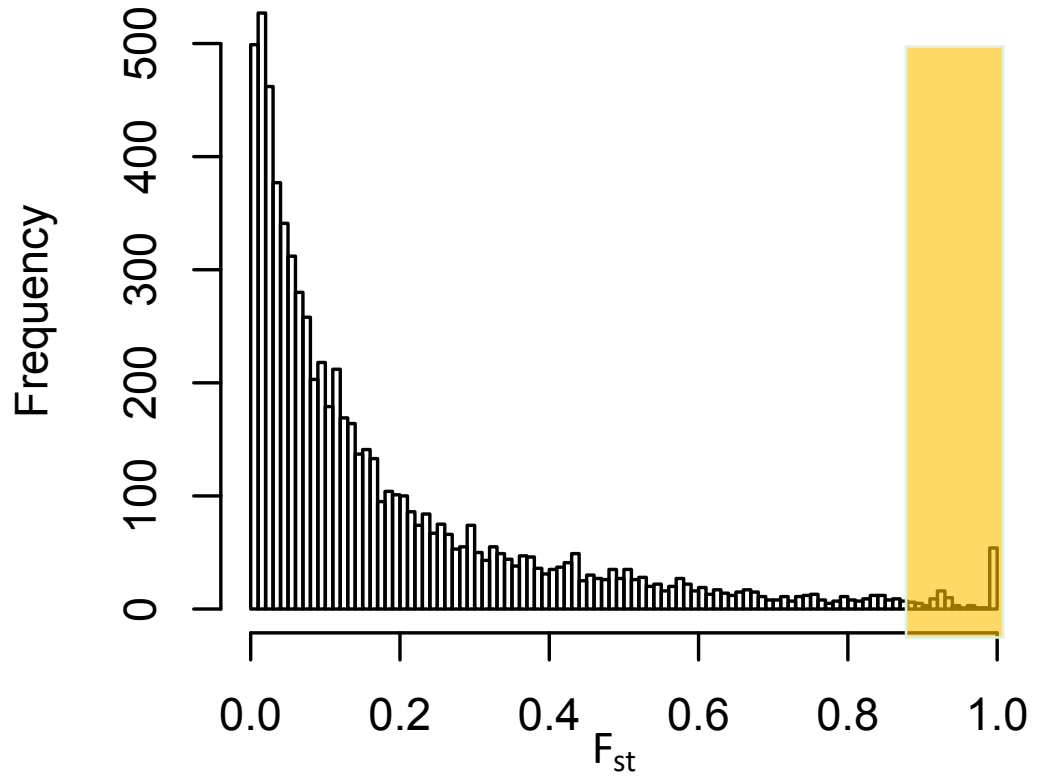
ddRAD data from single individuals



Fischer MC, unpublished



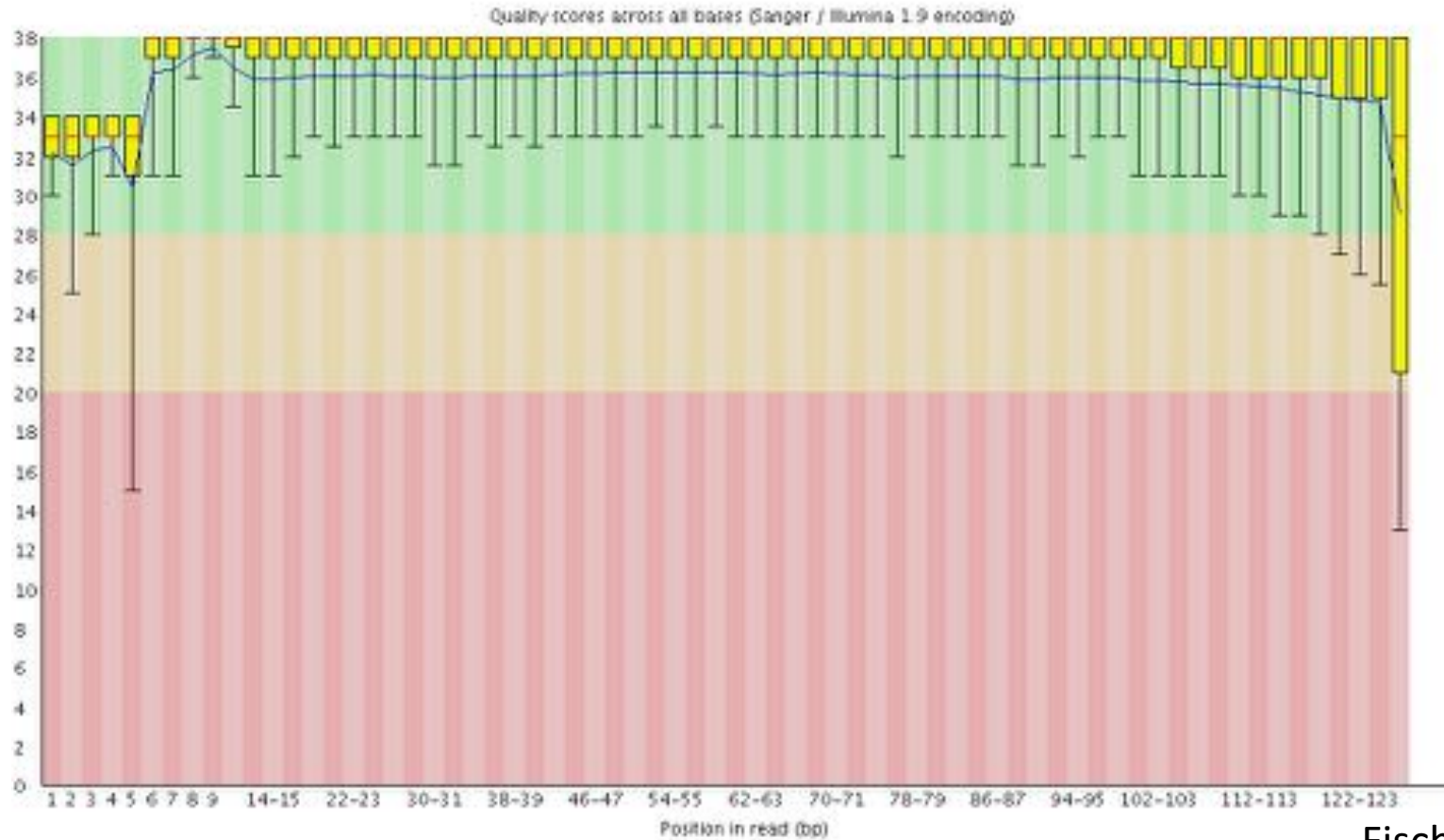
Genome scans



no differentiation \longrightarrow complete differentiation

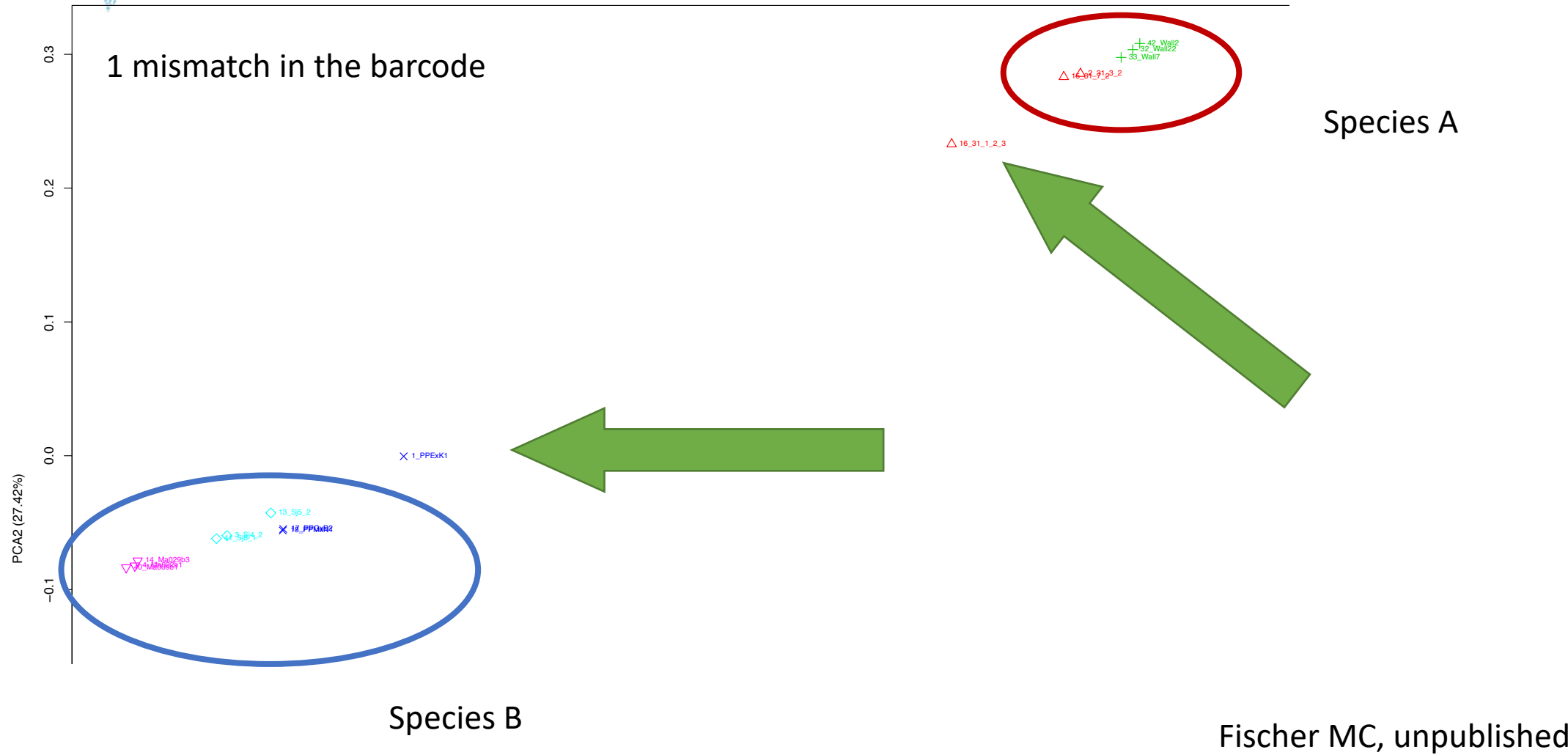
Looking for loci with increased differentiation (F_{st}) between species

ddRAD data from single individuals

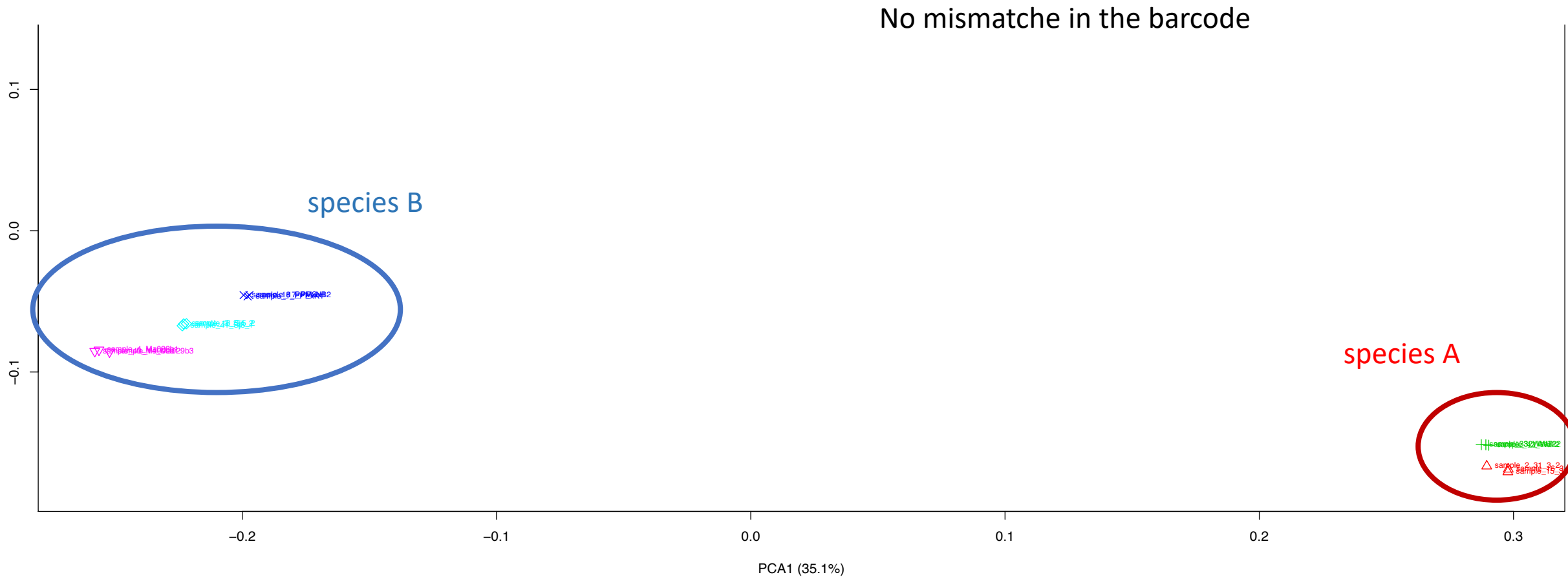


Fischer MC, unpublished

ddRAD data from single individuals

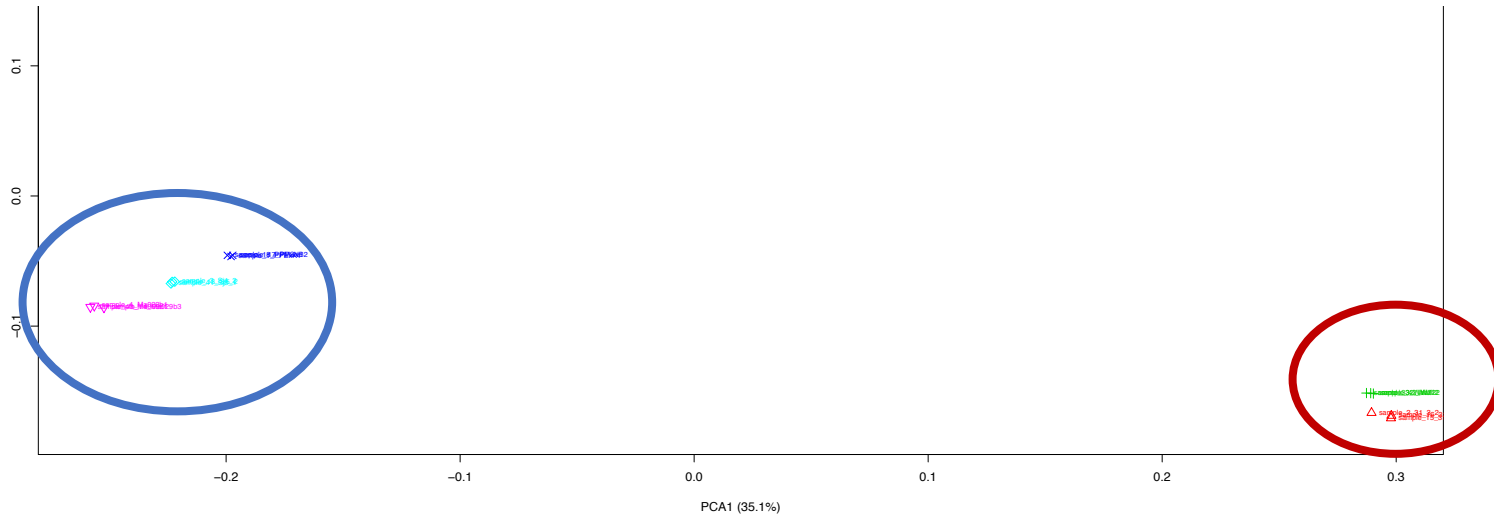


ddRAD data from single individuals



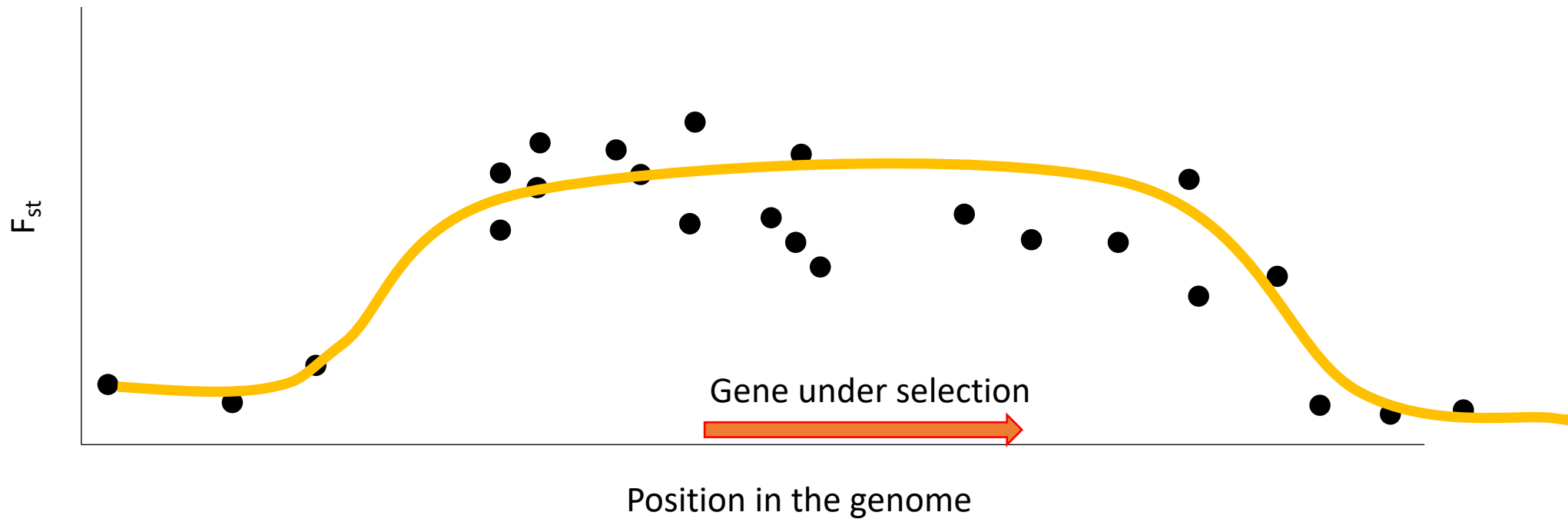
Fischer MC, unpublished

ddRAD data of single individuals

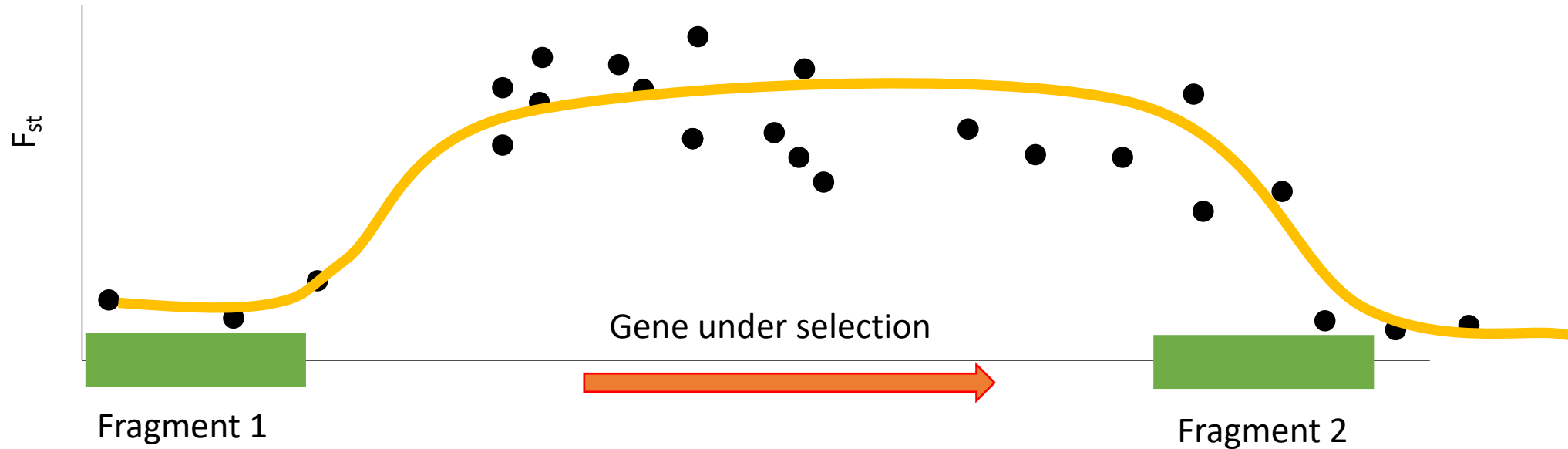


stringent demultiplexing
biological replicates

Can RAD be used to detect genes under selection?



Can RAD detect every genomic signal?



- Linkage decay (LD)
- Fragment density
- Island size



Potential challenges

High stochasticity in read coverage

Consistent library preparation, appropriate filtering

Reconstruction of loci can be difficult because of short single-end reads

Long paired-end reads can improve the de novo assembly

Wrongly inferred SNPs due to PCR duplicates

Modifications of the protocol, few PCR cycles during library preparation

Single individuals

Use replicates or apply stringent demultiplexing



Potential challenges

Allele dropouts

Biased population genetic estimators

Often not in gene regions

Possible signals could not be detected in genome scans or GWAS

Single loci

The reconstruction of single loci can be challenging even with a genome



Take home message

- Useful and inexpensive approach for a wide range of genomic questions
- There are limitations

