Reduced representation libraries/RAD

Introduction

Niklaus Zemp

29 June 2022

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Non-model organisms

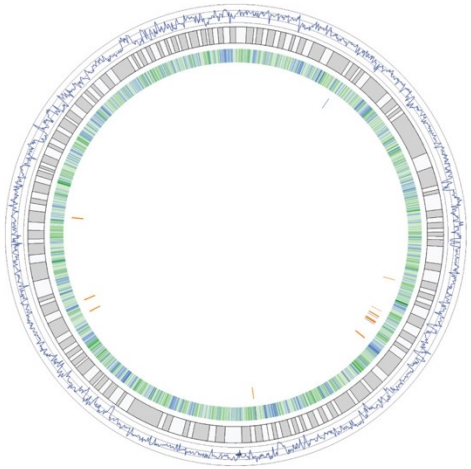







Genomics of large and unexplored genomes

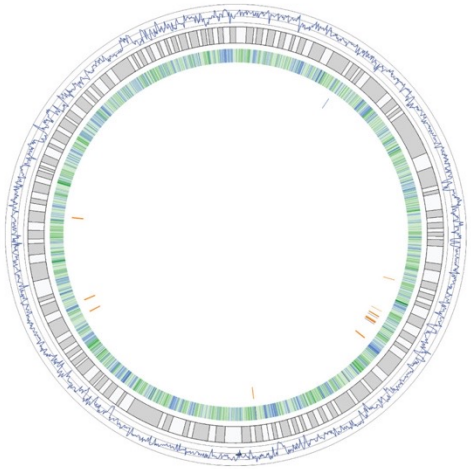
Assembly

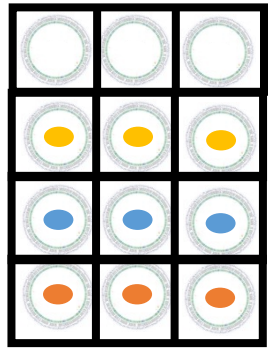


draft genome

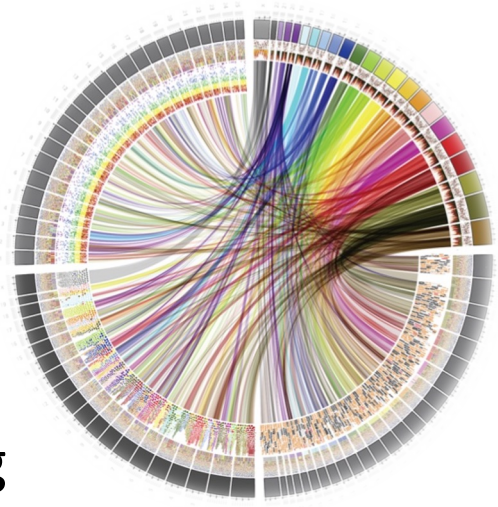
Genomics of large and unexplored genomes

Assembly

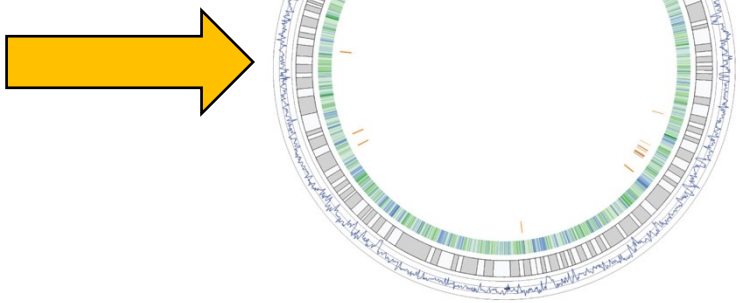


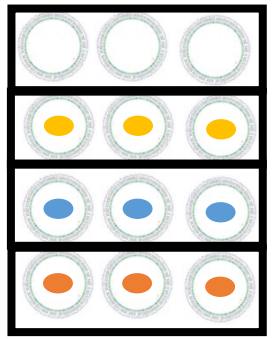


(low coverage) resequencing

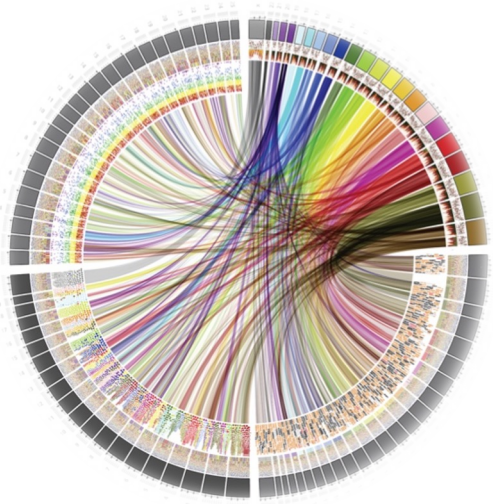


Genomics of large and unexplored genomes





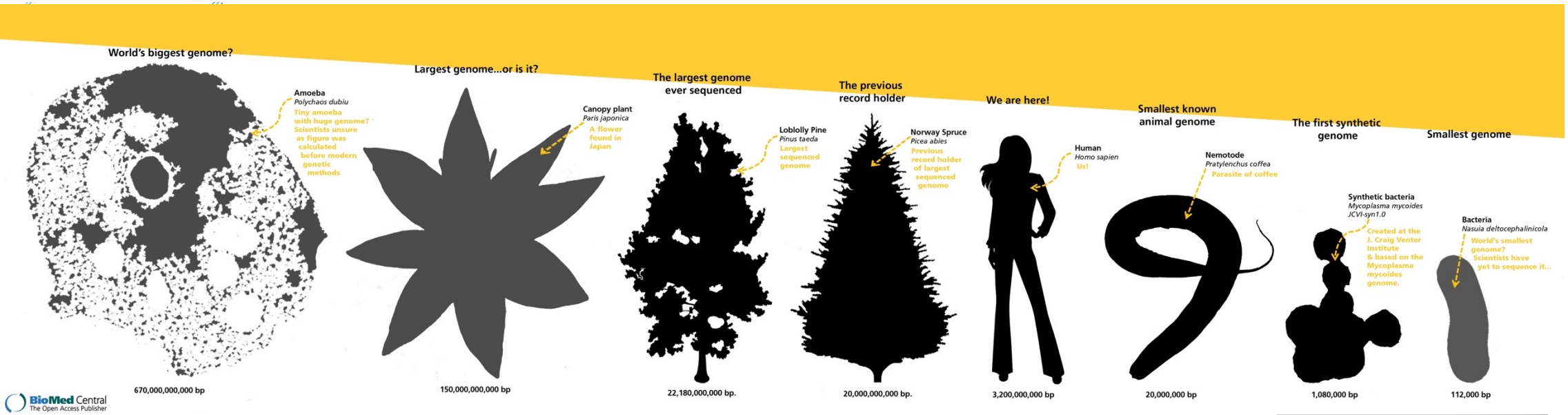
Pool-seq



Genomes can be large

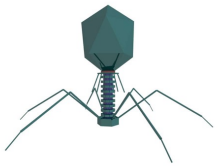
670 150 22.1 3.2

Genomes can be large



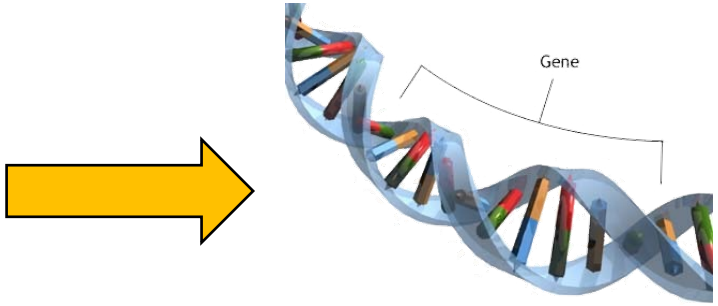
>1 year

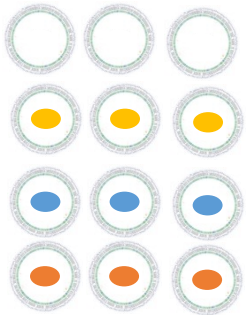
2 weeks



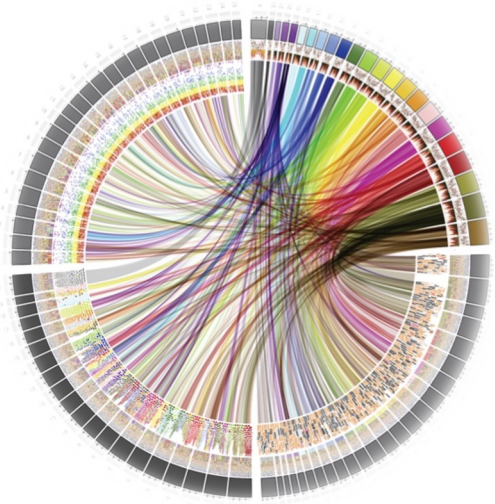
1 million bases is 1 minute

Genomics of large and unexplored genomes

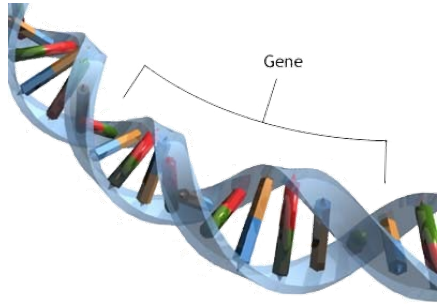


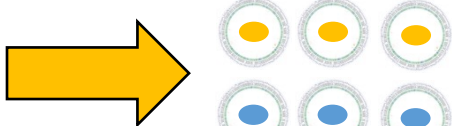


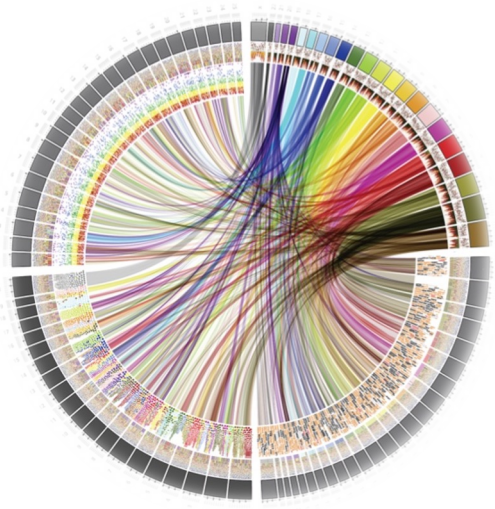
RNA-seq



Genomics of large and unexplored genomes



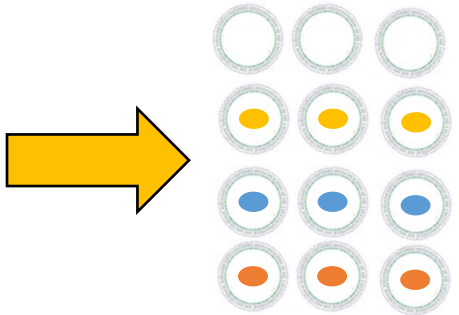


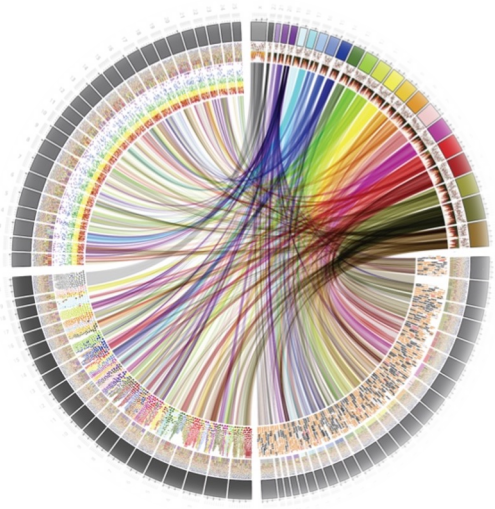


Target einrichtend

Resequencing

Genomics of large and unexplored genomes

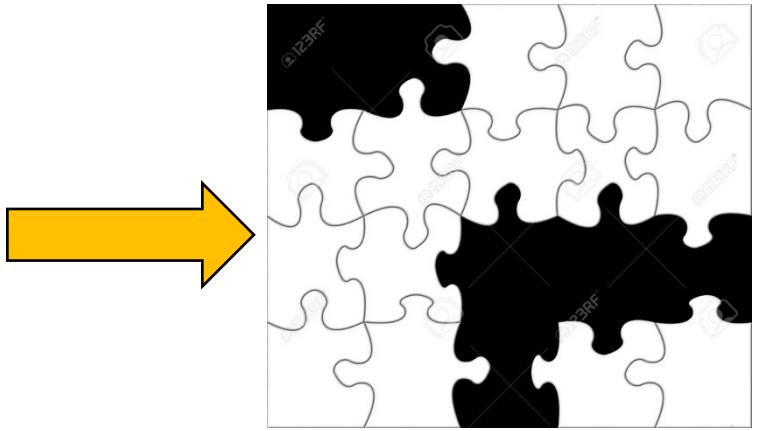


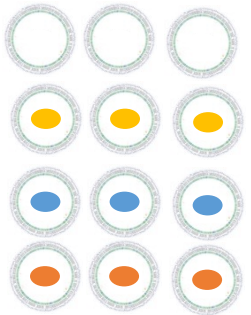


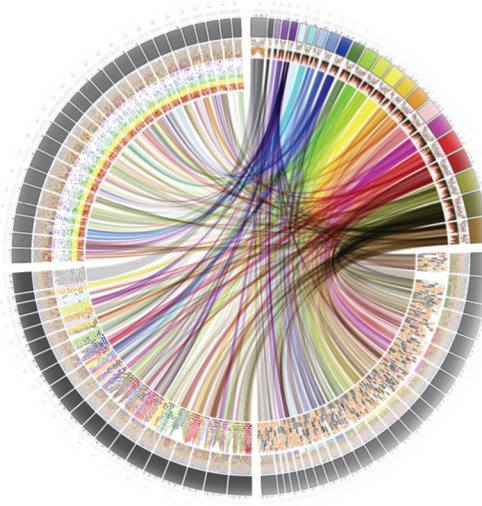
Target einrichtend

Resequencing

Genomics of large and unexplored genomes



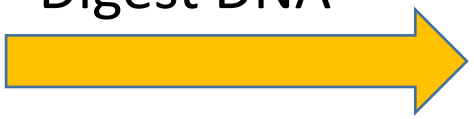


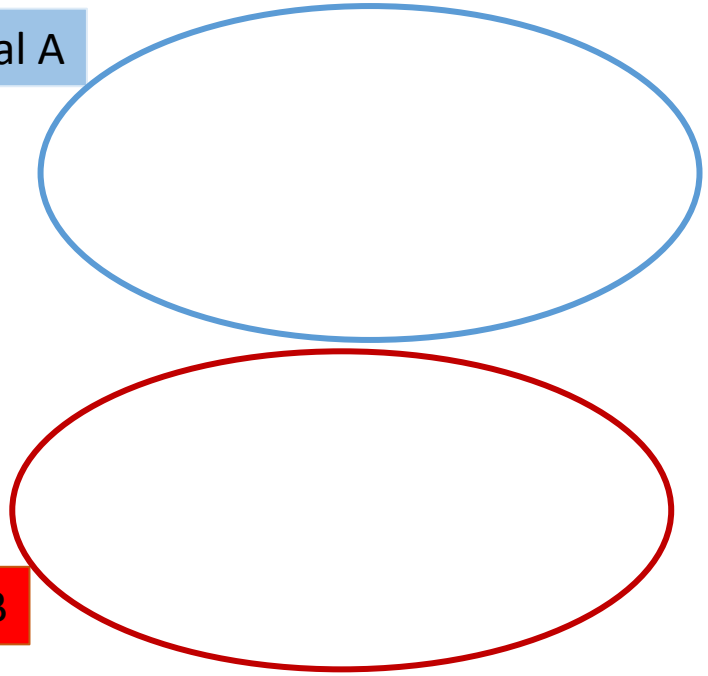


Reduced representation libraries/
Restriction site **a**ssociated **D**NA (RAD)

Reduced representation libraries (RAD)

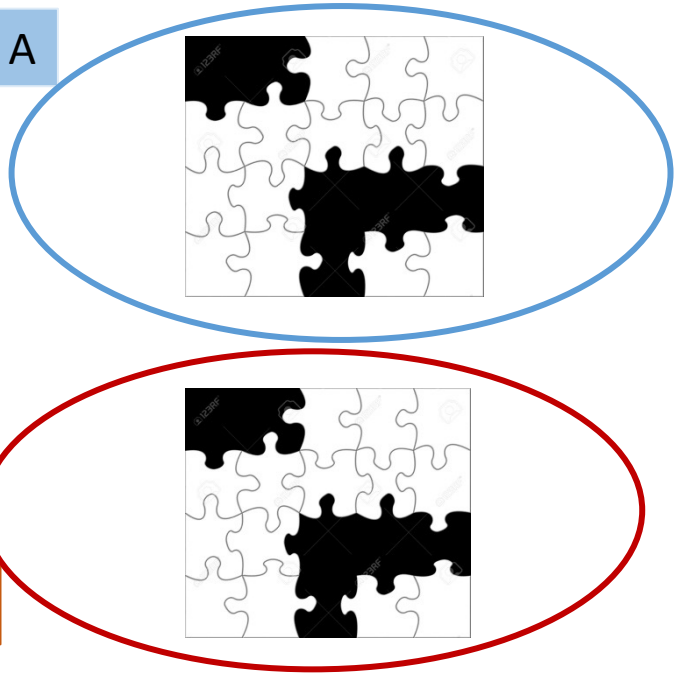
Digest DNA





Reduced representation libraries (RAD)

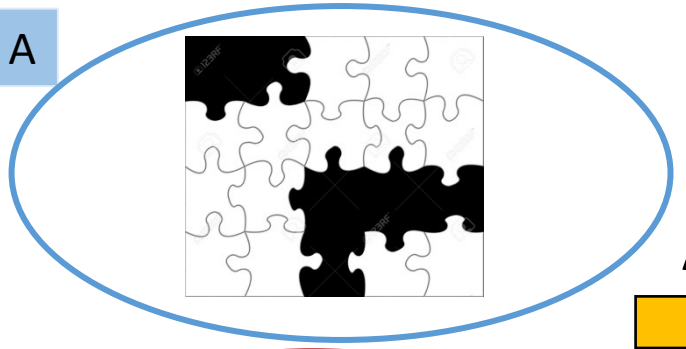
Digest DNA

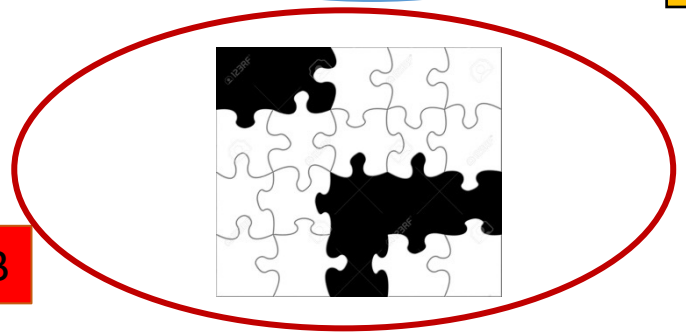


Reduced representation libraries (RAD)

Individual A

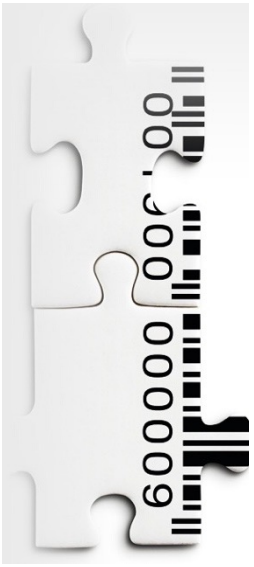
Digest DNA





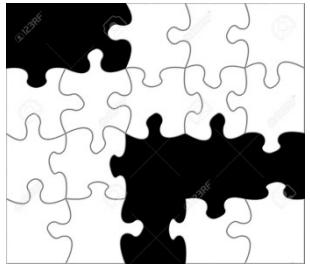
Individual B

Add barcodes



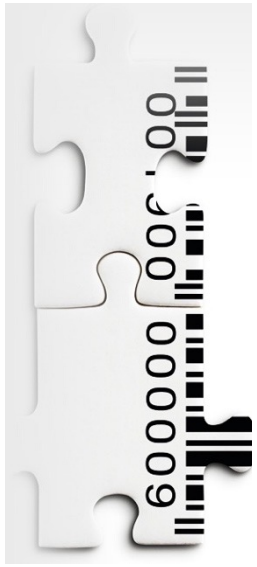
Reduced representation libraries (RAD)

Individual A

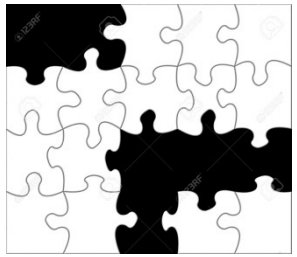


Digest DNA

Add barcodes



Individual B



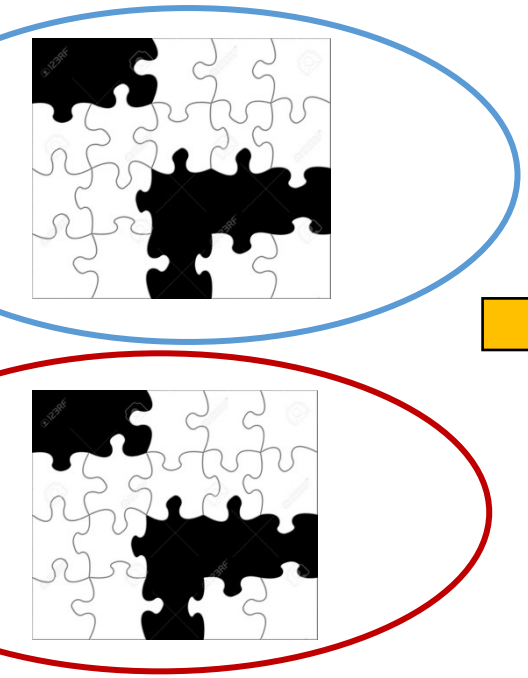
Sequence pooled libraries

Data analysis

Reduced representation libraries (RAD)

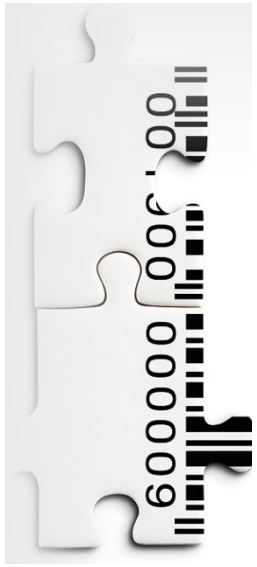
Individual A

Individual B



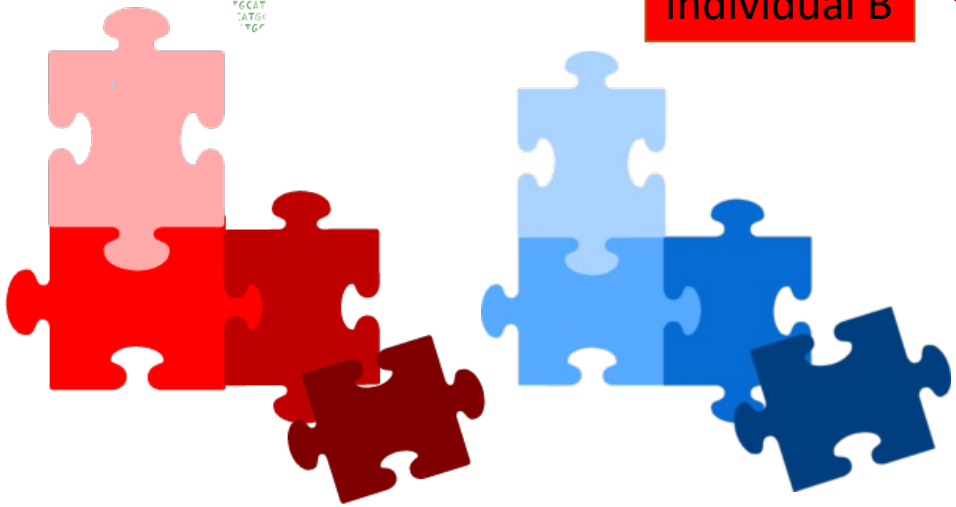
Digest DNA

Add barcodes



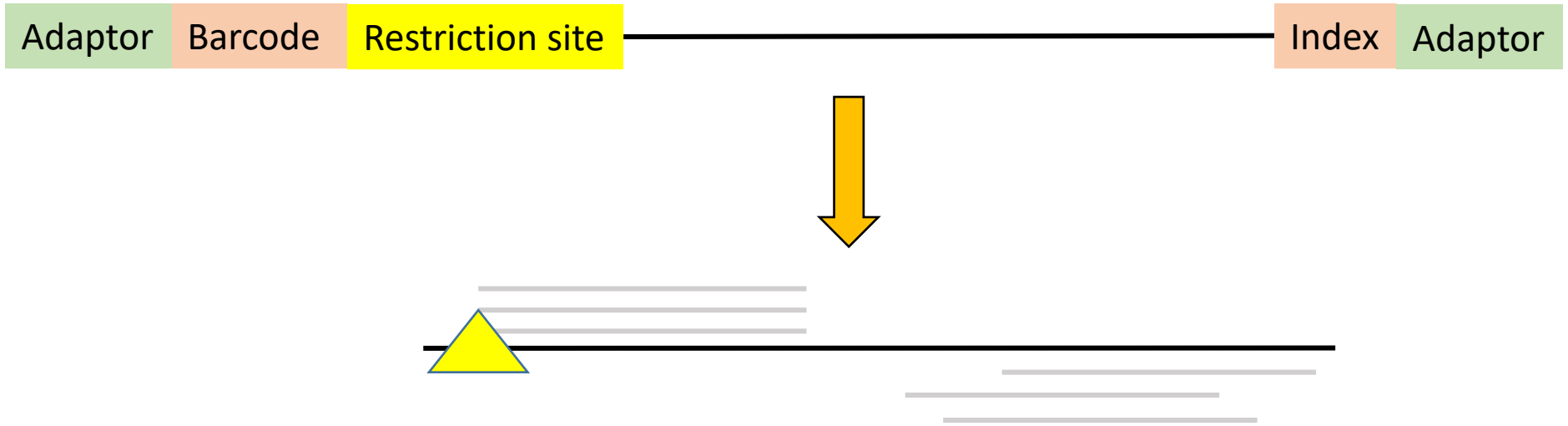
Sequence pooled libraries

Data analysis



mbRAD, ddRAD, GBS, 2bRAD, ezRAD

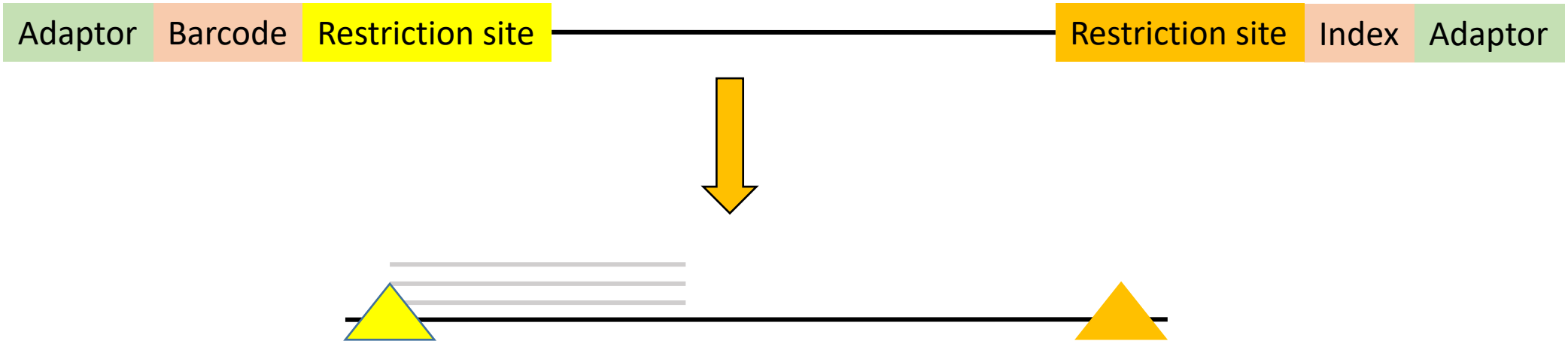
mbRAD



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

Miller et al. 2007, Baird et al. 2008

ddRAD



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

Peterson et al. 2012

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
- ∝ Genome size
- ∝ Sequencing depth

Genome available:

In silco digest (simRAD)

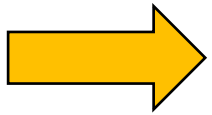
***de novo*:**

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

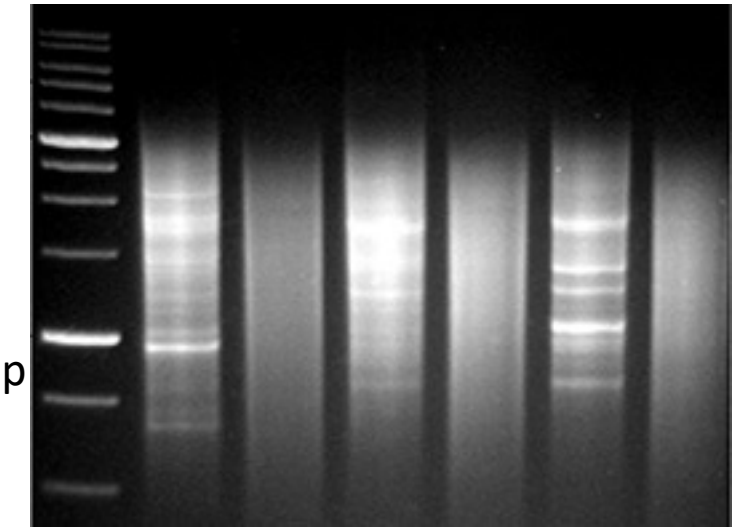
Can I produced these libraries at the GDC?

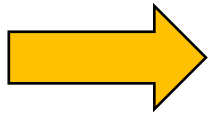
Primers sets and protocols for producing RAD and ddRAD libraries are available

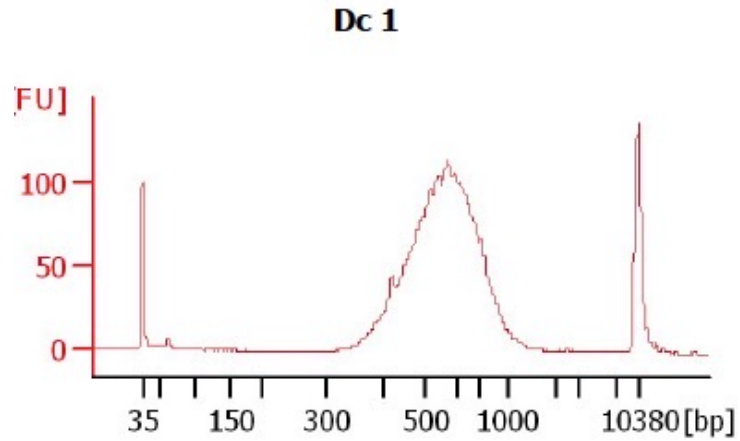
200-500 ng high quality DNA



500 bp

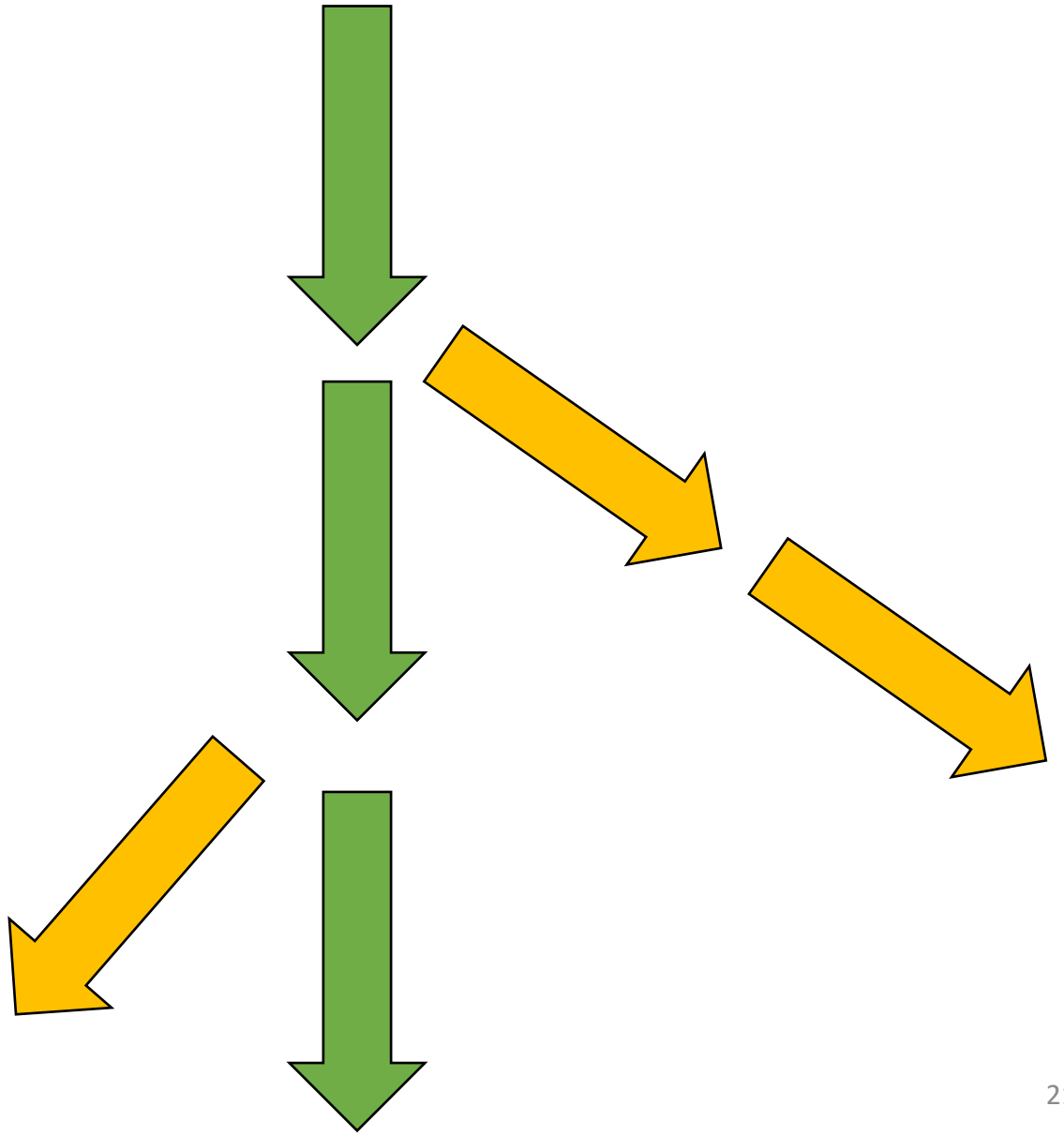




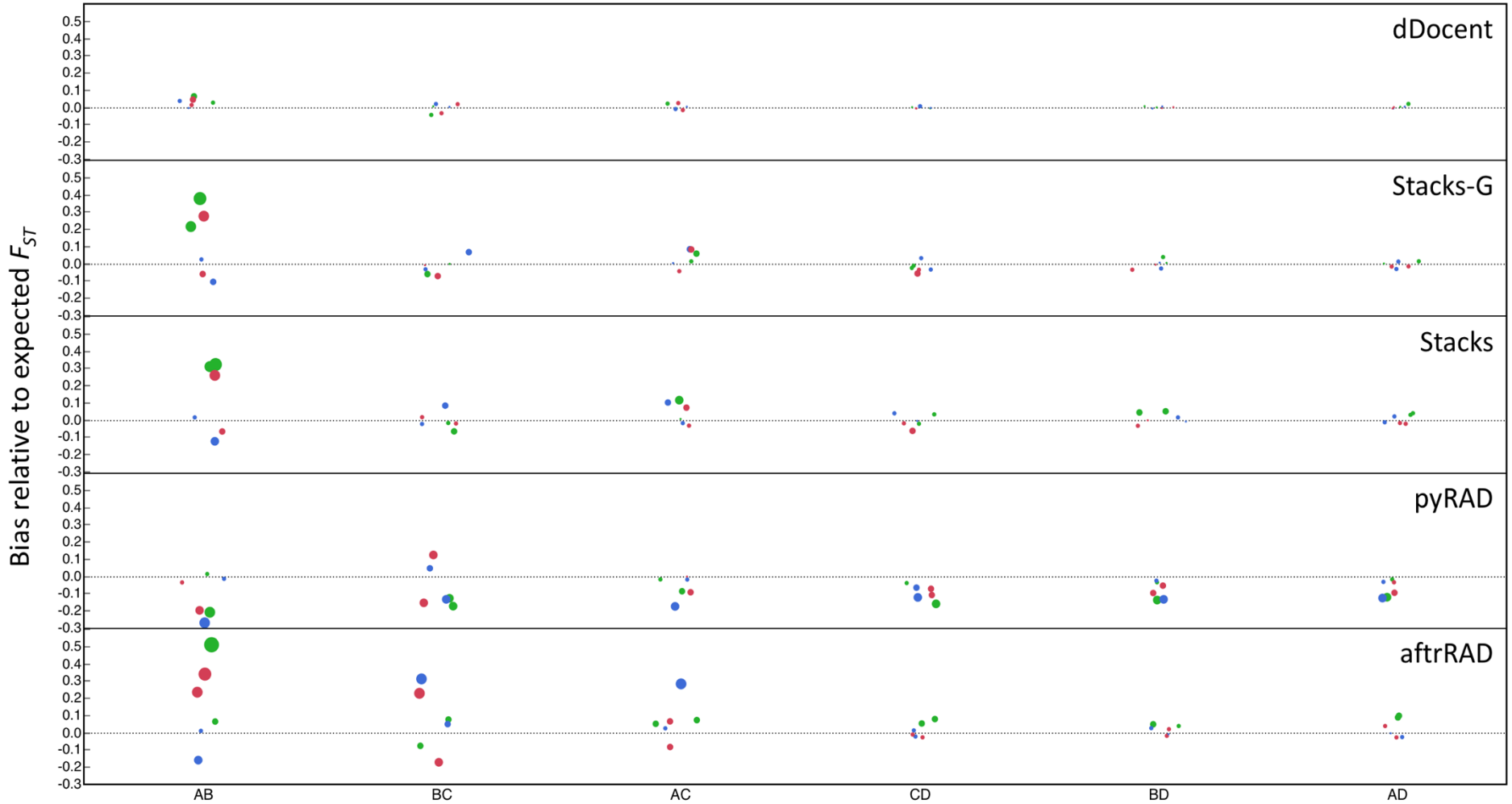


Pipelines

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



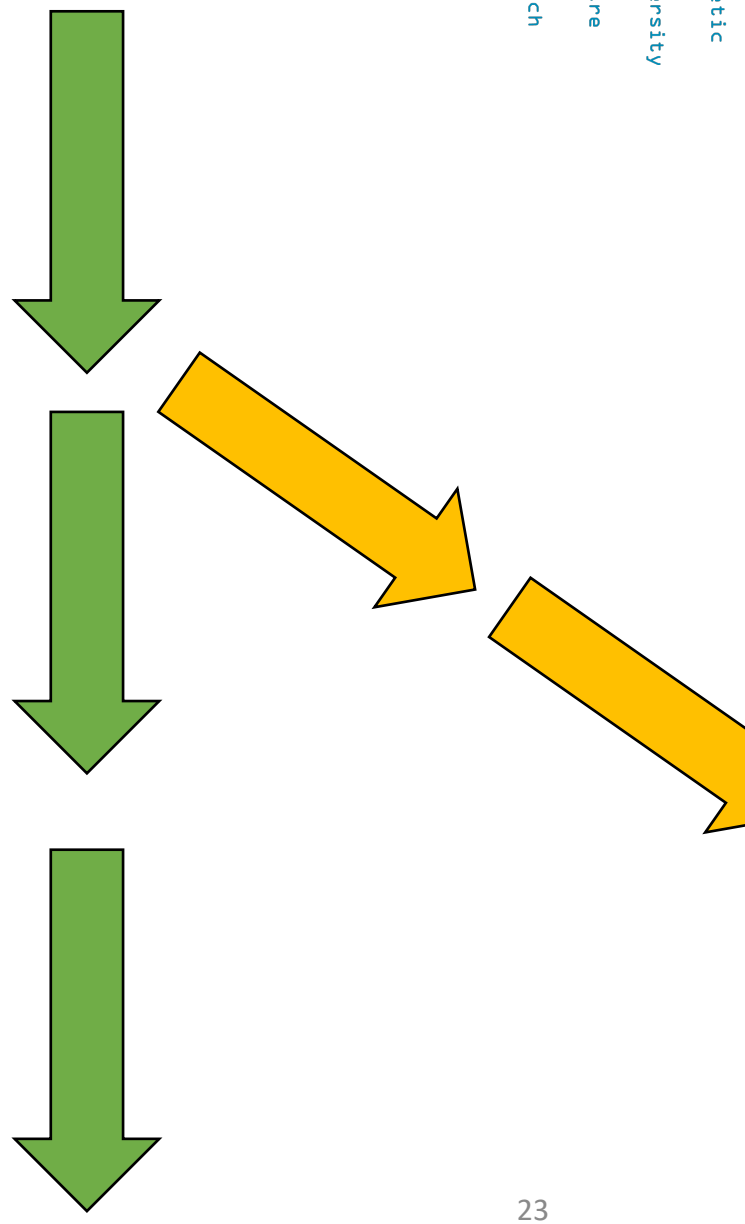
Population differentiation bias of different pipelines



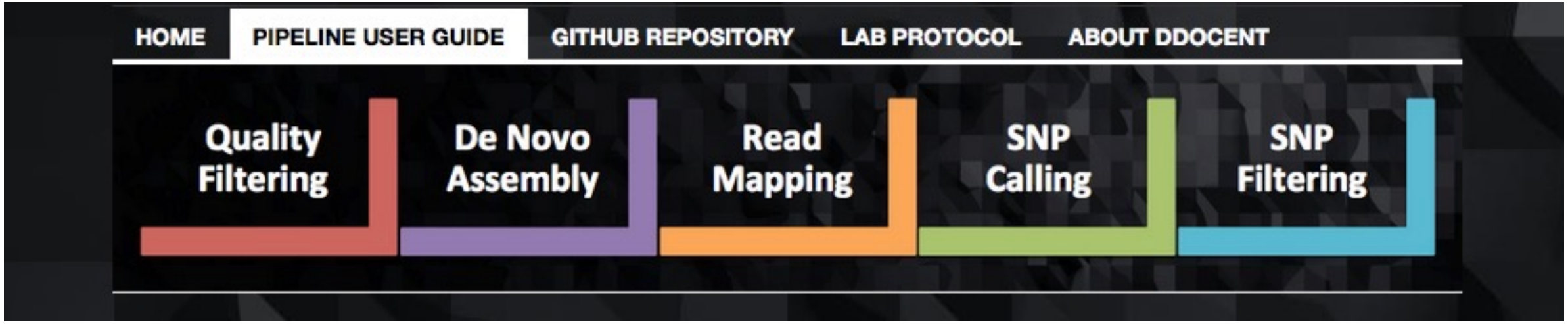
https://github.com/jpuritz/dDocent

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 phylogentics



dDocent



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

Puritz et al. 2014

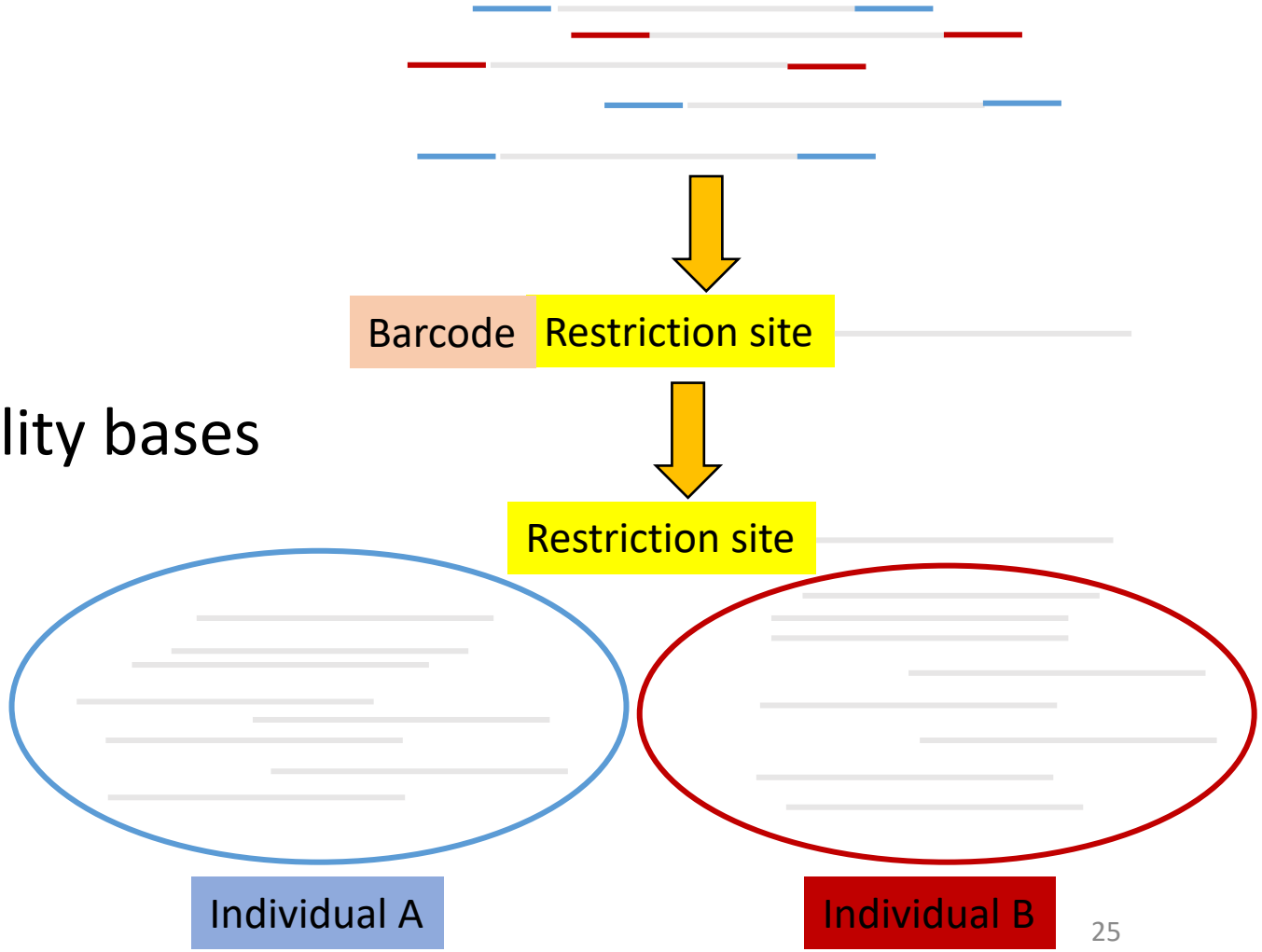
Demuliplexing and quality filtering

- Demuliplex reads

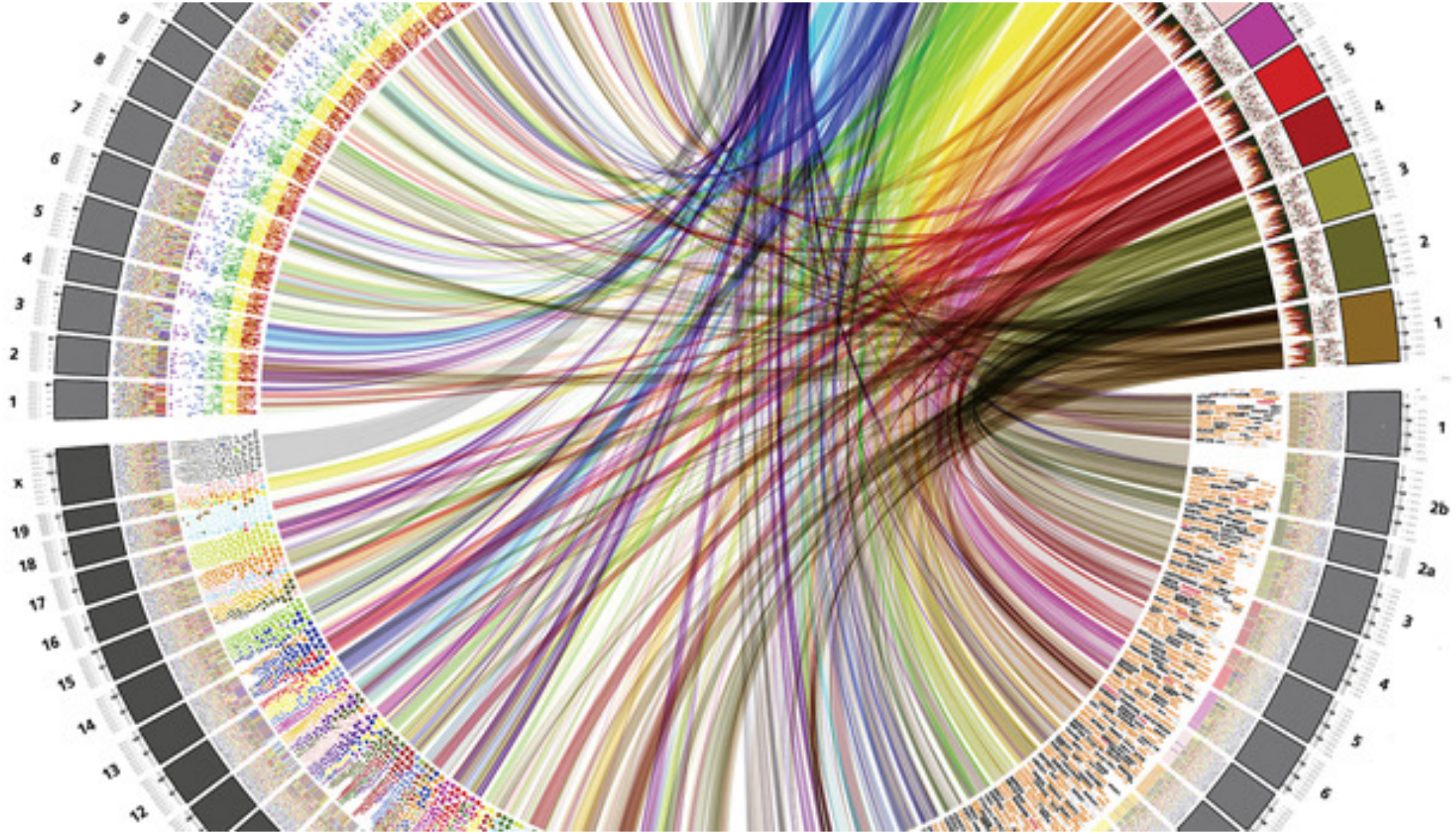
  process_radtags (Stacks)

- Remove adaptors and low quality bases

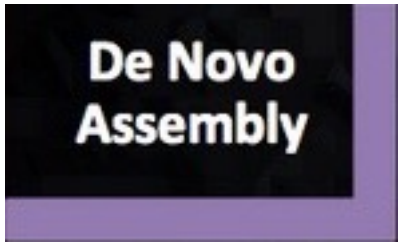
  Trimmomatic/fastp



Reference assembly

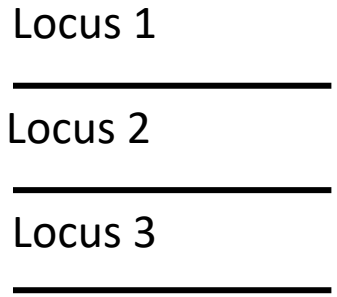


De novo assembly



Merge reads in case of overlaps
PEAR

Remove all identical reads
Pool all individuals together
customized scripts



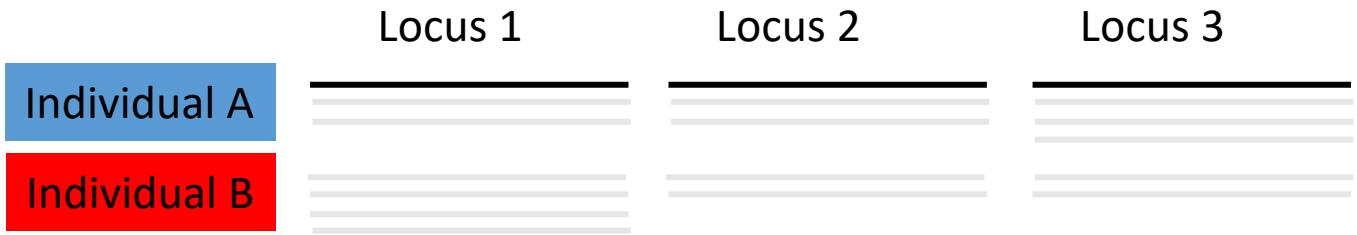
Single-end:
Cluster the non-redundant sequences based on similarity
cd-hit-est

Paired-end:
Assembly the non-redundant sequences and than using paired-end information
rainbow, cd-hit-est

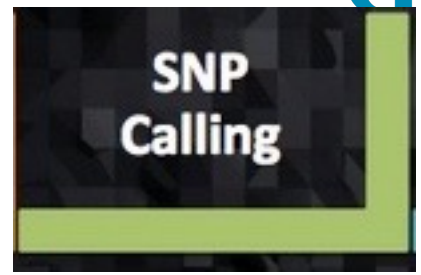
Read mapping

- Mapping reads against the reference catalogue

  BWA



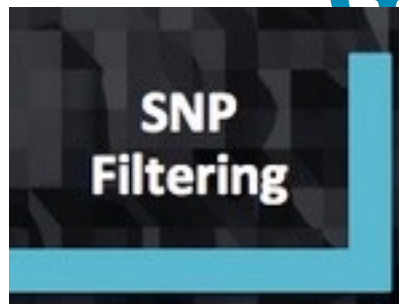
SNP calling



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

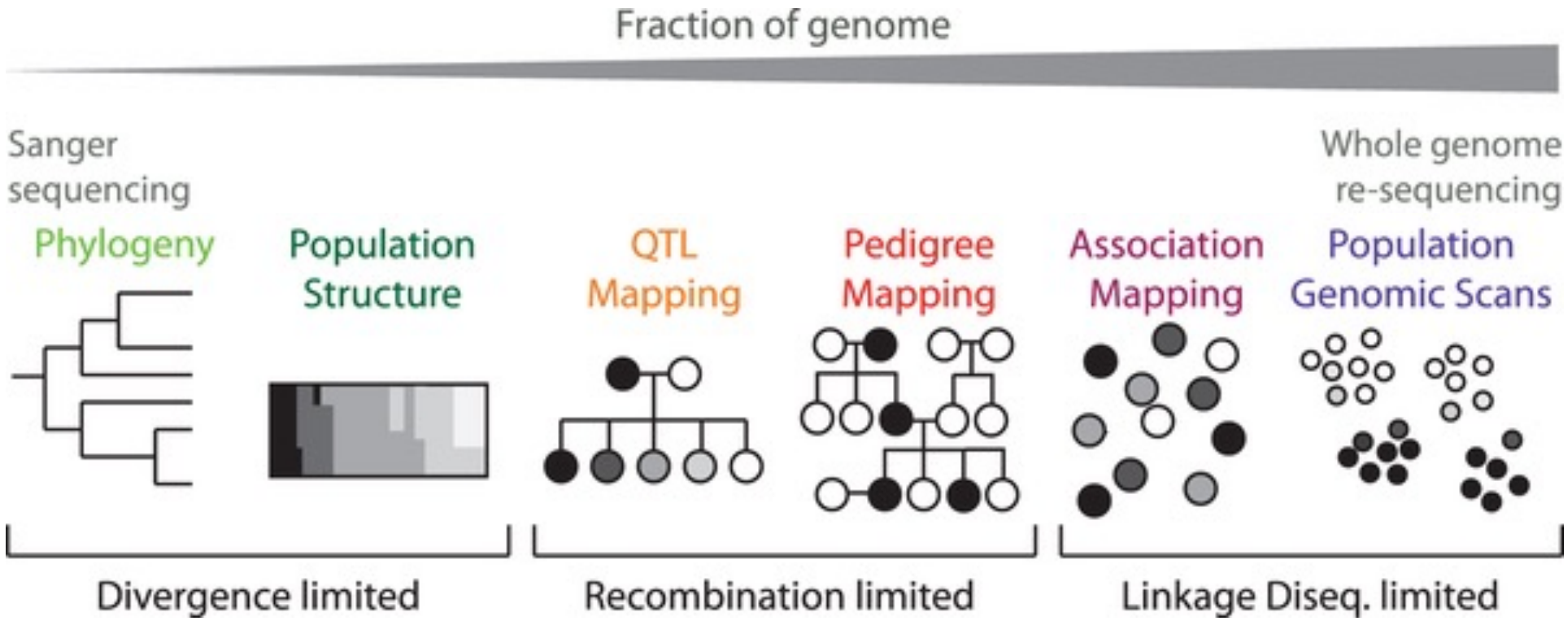
SNP filtering



Filter only for good SNPs
VCFtools, vcflib

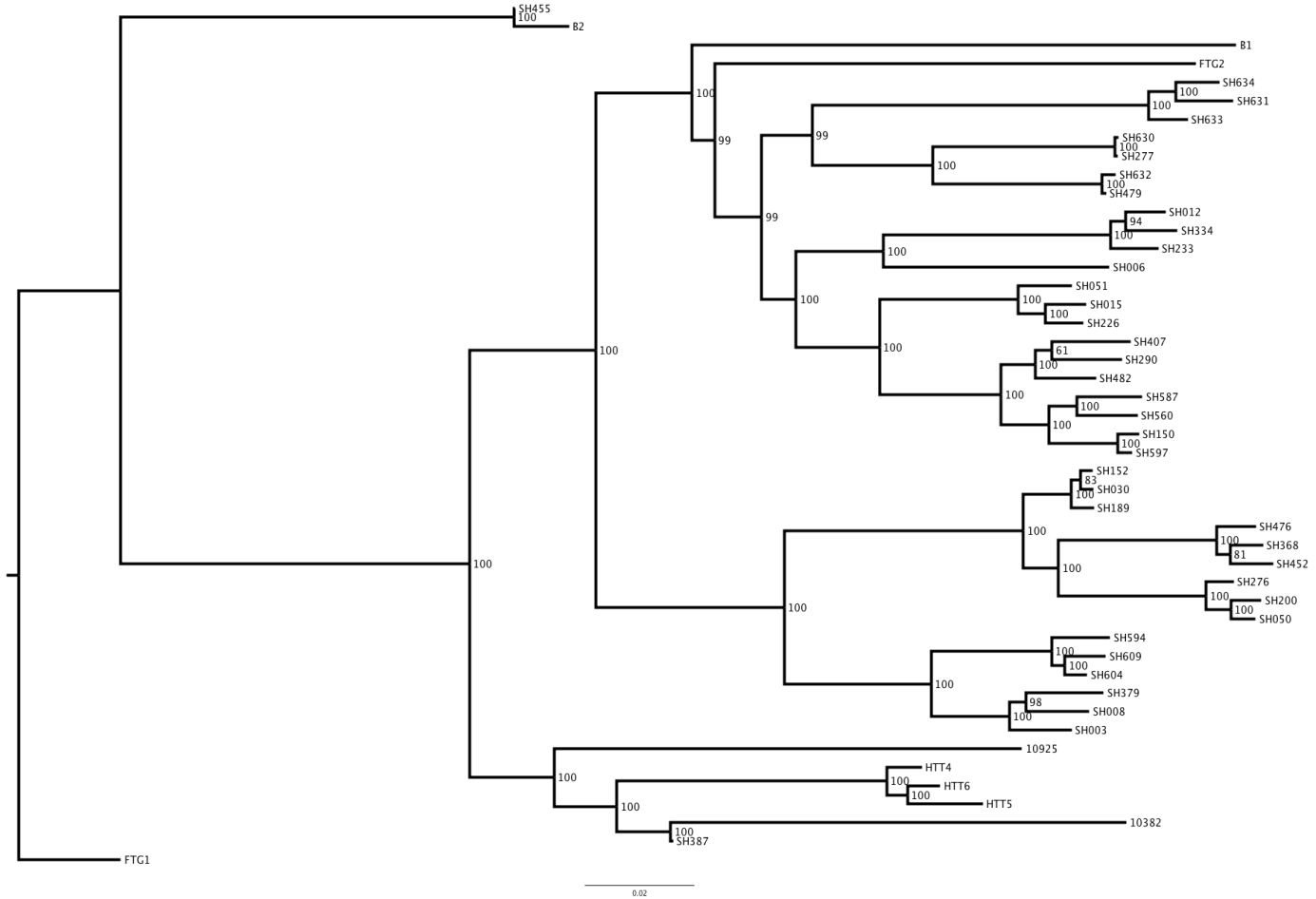
Criteria:
Mapping quality
Coverage
Missing genotypes
Minor allele frequency
Balanced alleles

What is it good for?



Peterson et al. 2012

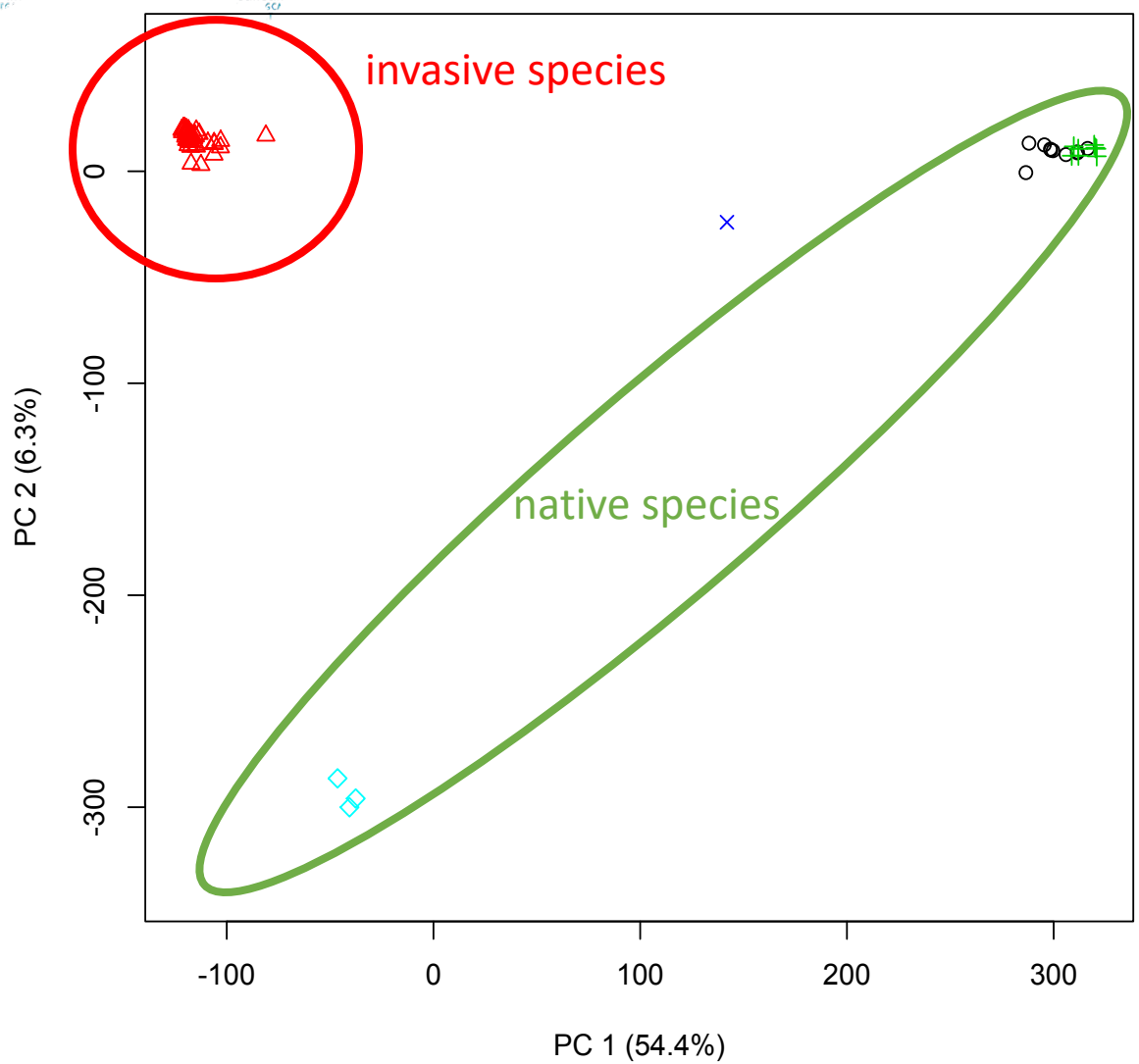
Phylogeny of rosewood species from Madagascar



Better resolution with ddRAD compared to traditional barcoding markers

Crameri S, unpublished

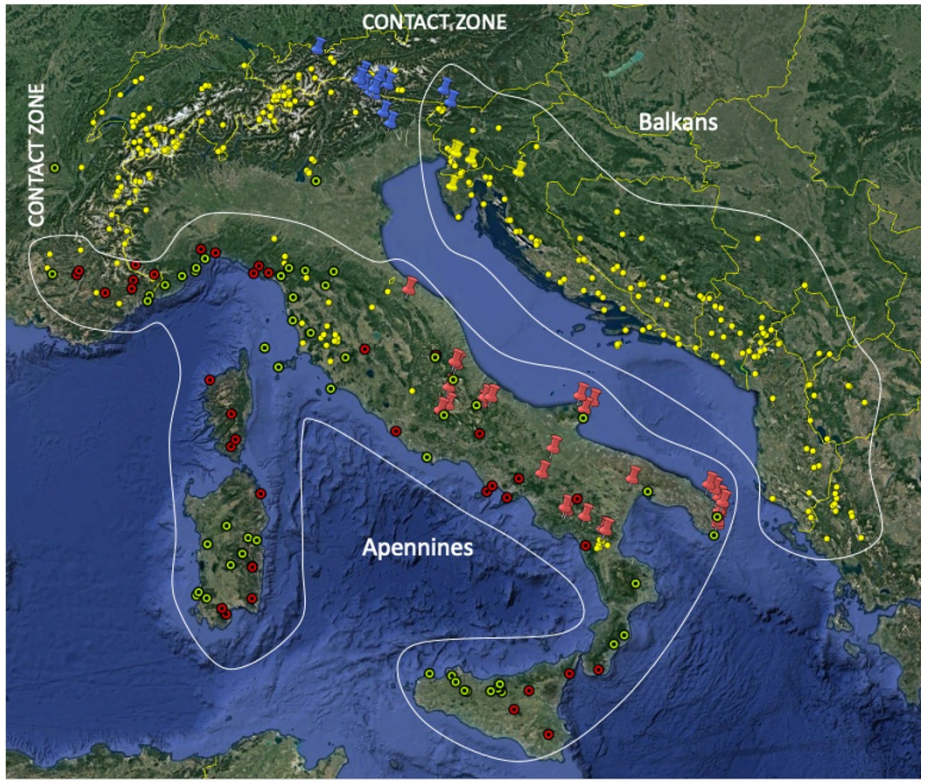
Population structure

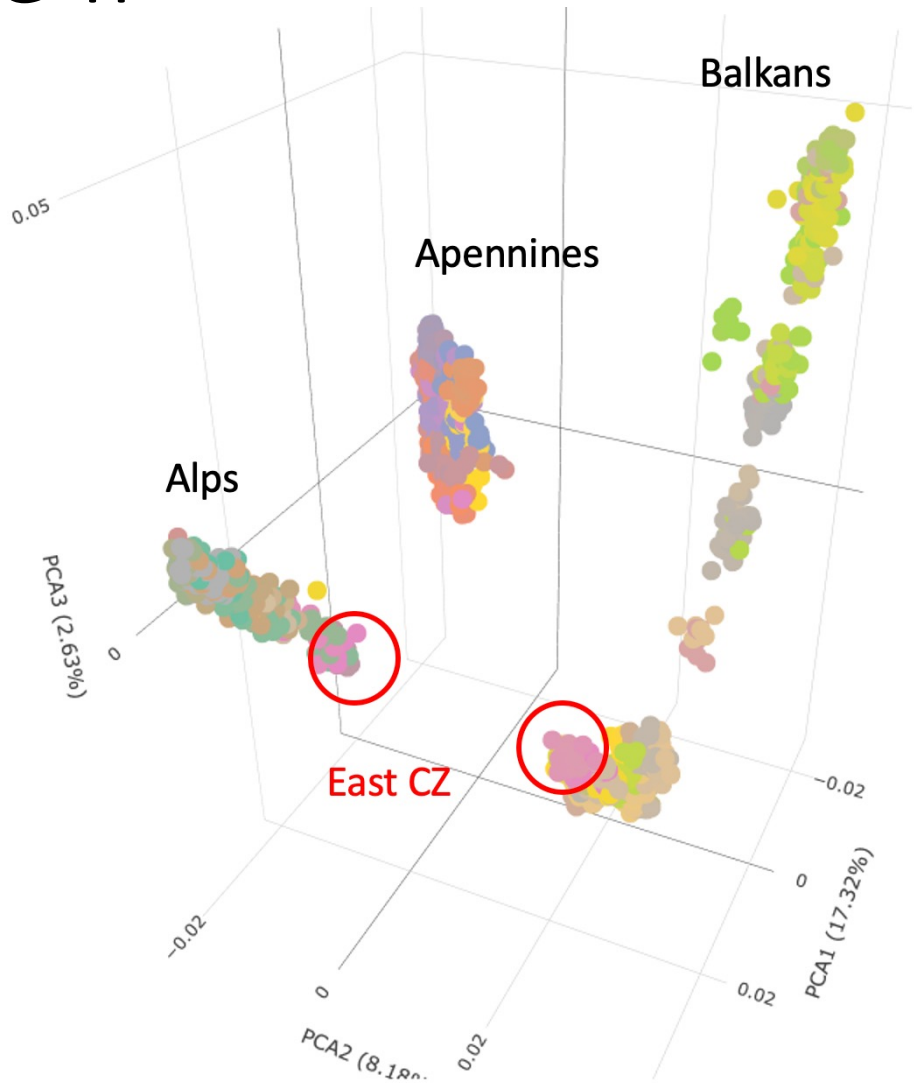


Widmer and Zemp, unpublished

No hybridization between invasive and native groundsel species

Population structure II

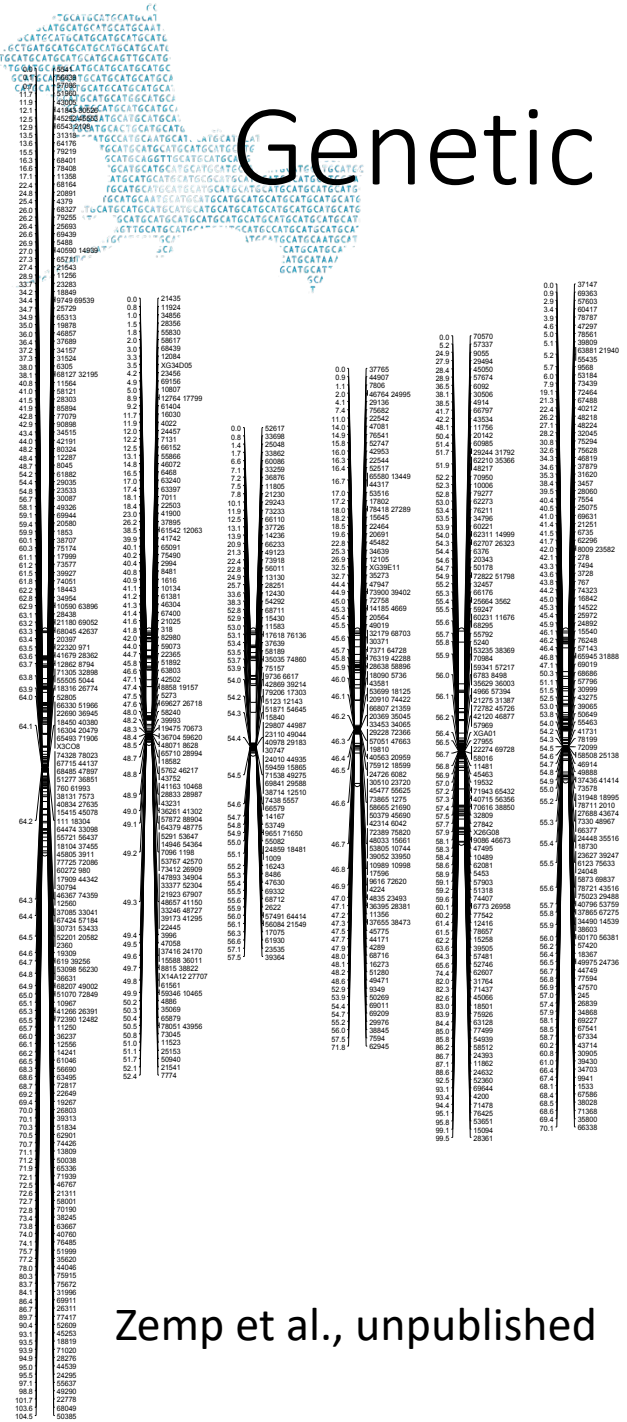




~2500 samples
SNP panel

Fior S, unpublished

Genetic linkage mapping



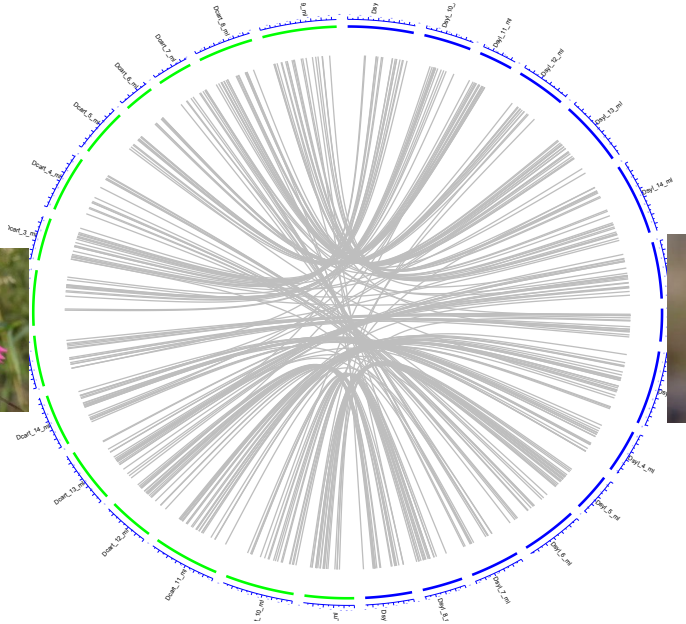
Zemp et al., unpublished

A

B

C

More recombination events between A and C than A and B

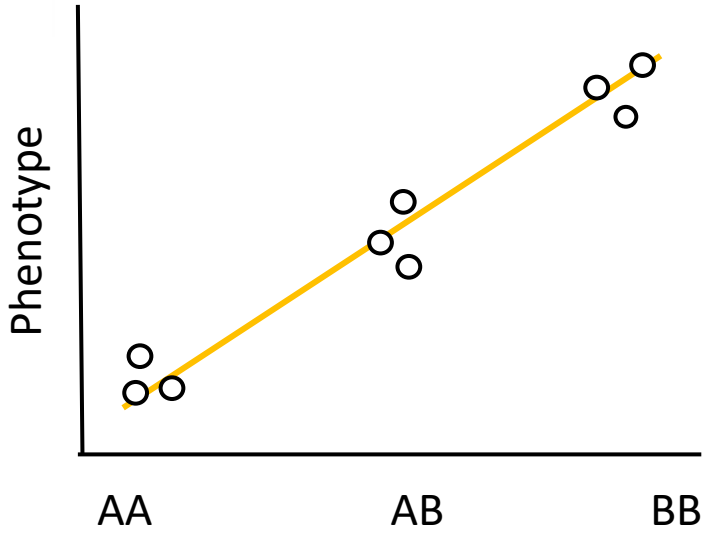


Fior et al., unpublished

Ordering the fragments using recombination rates in the offspring

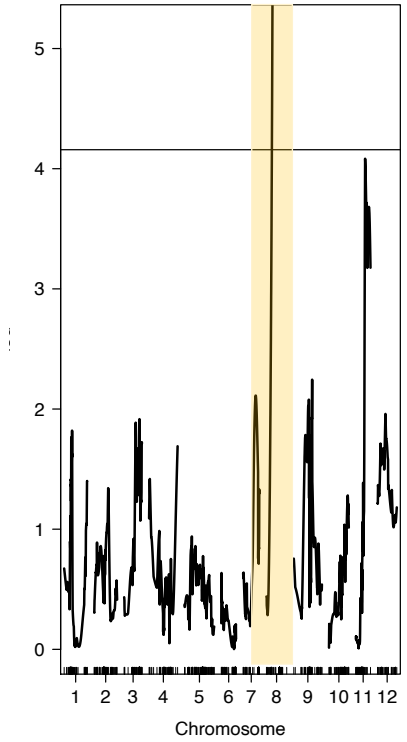
- Distribution of fragments across the genome
- Scaffolding and comparative analyses of draft genomes
- Estimate recombination rate across the genome

Correlations between phenotype and genotype (QTL, GWAS)



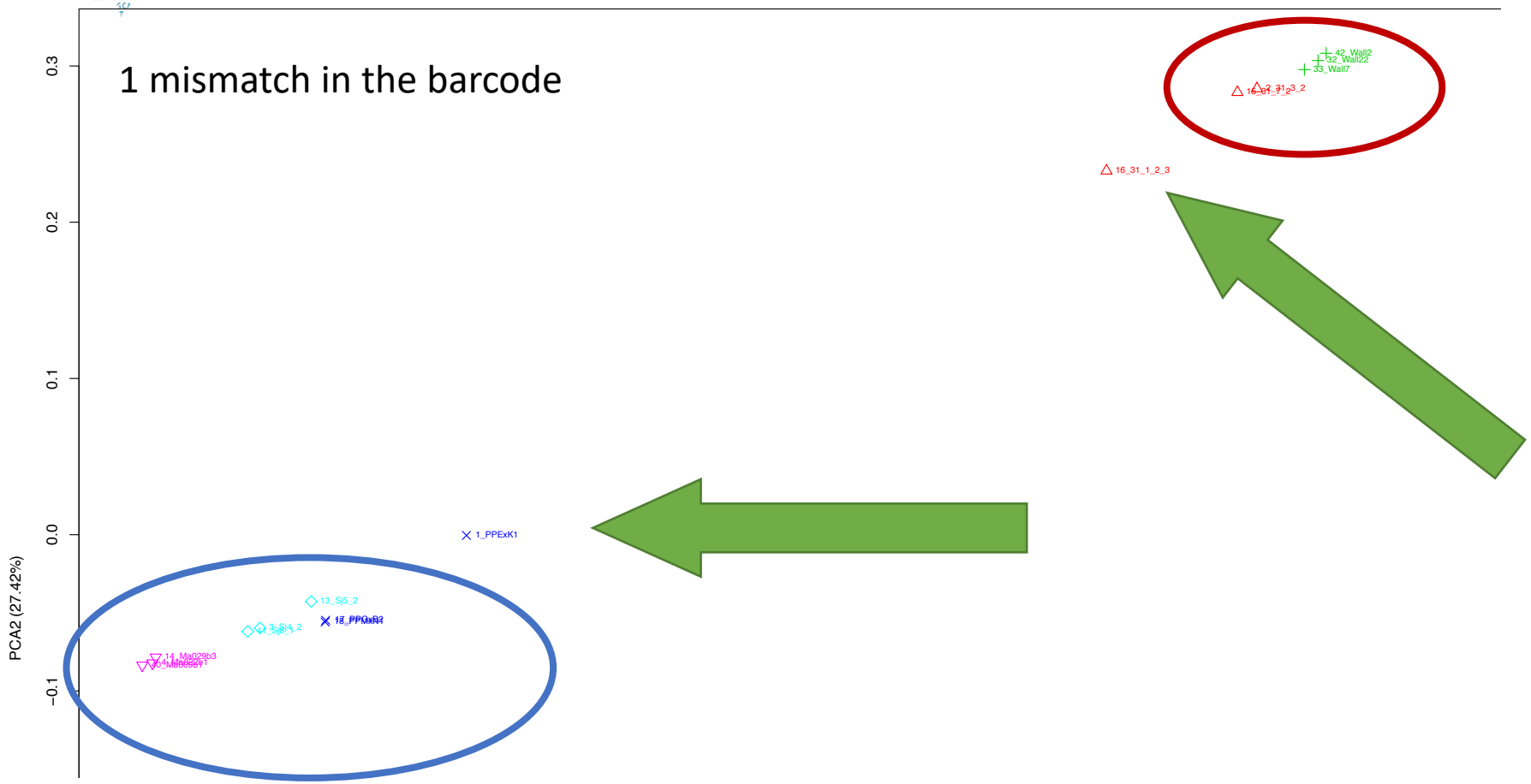
Looking for loci that are highly correlated with a certain phenotype

Plant height



Crameri, Nenadic and Zemp, unpublished

ddRAD data from single individuals



Species A

Species B

Fischer MC, unpublished

Cross-Contamination

Mapping statistics

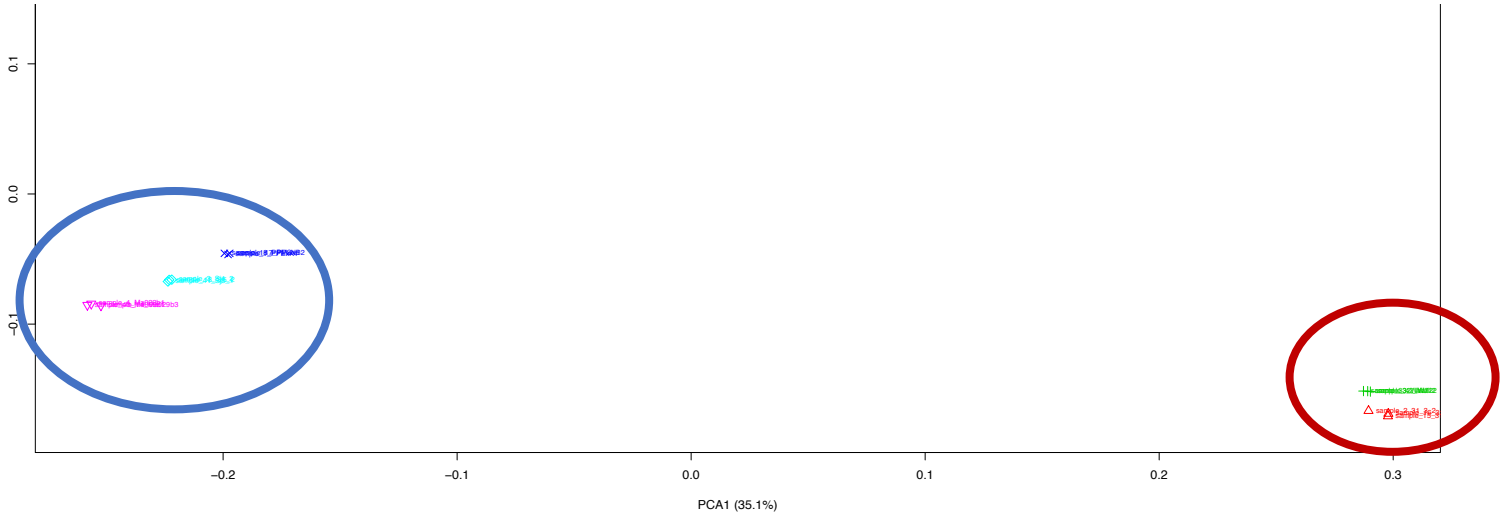
Sample	total_reads	mapped_reads	%mapped_reads	statsQ20_mapped_reads	%statsQ20_mapped_reads	properly_paired-reads	%properly_paired-reads	note
Apenn256_9	140975	68327	0.484675	49921	0.354112	35880	0.254513	Excluded
EMPTY_1	80575	67996	0.843885	51022	0.633224	38779	0.481278	Excluded
EMPTY_4	190232	152370	0.800969	113789	0.598159	82052	0.431326	Excluded
CZ16866_1	1546578	1259269	0.814229	1002826	0.648416	727511	0.4704	
D196_1	1470830	1343594	0.913494	1006629	0.684395	736645	0.500836	
Apenn280_9	1489948	1386199	0.930367	1006843	0.675757	669529	0.449364	

Cross-Contamination

Population genetic estimator

#Sample	Total_Sites	Het_Sites	Het_Sites_%
Apenn275_6	17260	10521	0.60956
D233_3	16805	9542	0.567807
Apenn282_3	17337	11382	0.656515
Apenn284_5	17259	10154	0.588331
Apenn284_3	16870	8932	0.529461
Apenn277_3	16769	9235	0.550719
Apenn272_10	17165	8515	0.496068
D236_1	17168	5193	0.302481
D230_4	17052	4200	0.246305
Apenn261_5	17121	4741	0.276911
Apenn281_7	16882	4665	0.27633
D223_3	17136	4937	0.288107
Apenn279_6	17226	5451	0.31644
Apenn273_5	16967	4560	0.268757

ddRAD data of single individuals

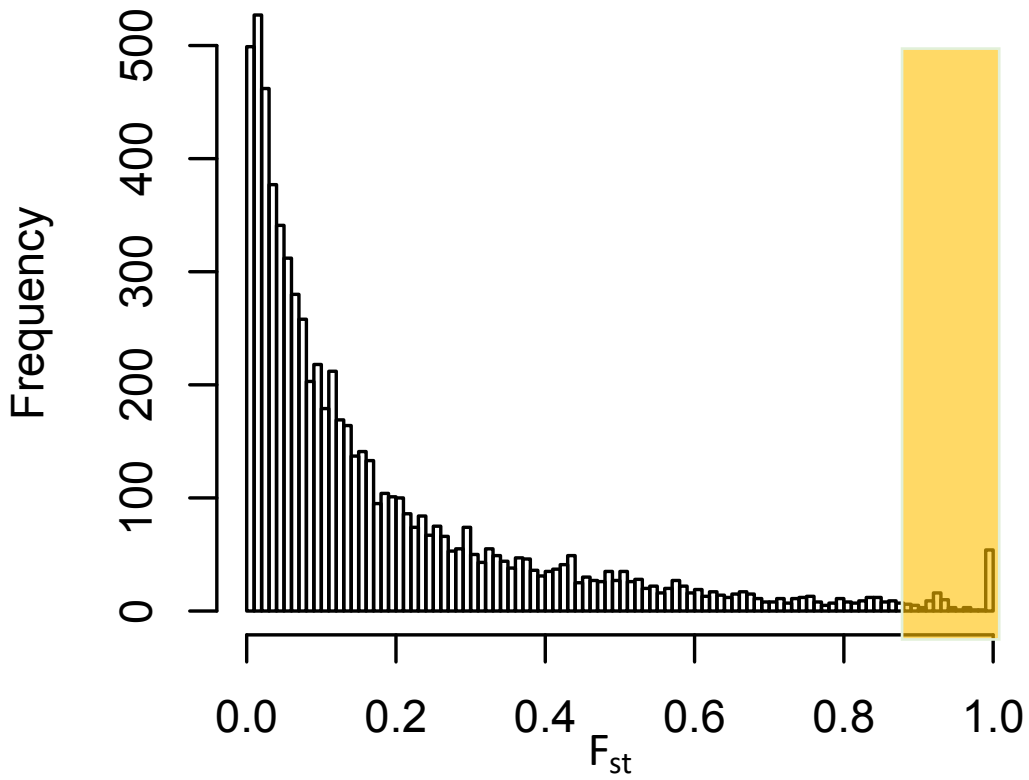


Biological replicates

Positive controls (e.g. estimate genotyping error)

Negative controls

Genome scans

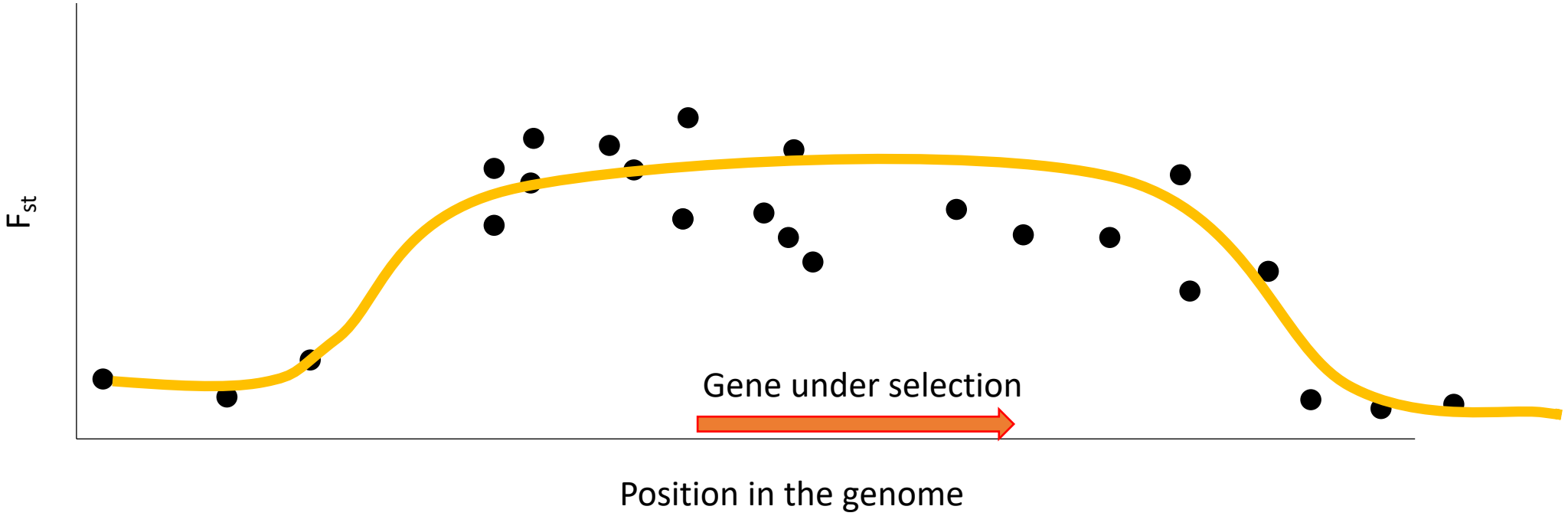


no differentiation → complete differentiation

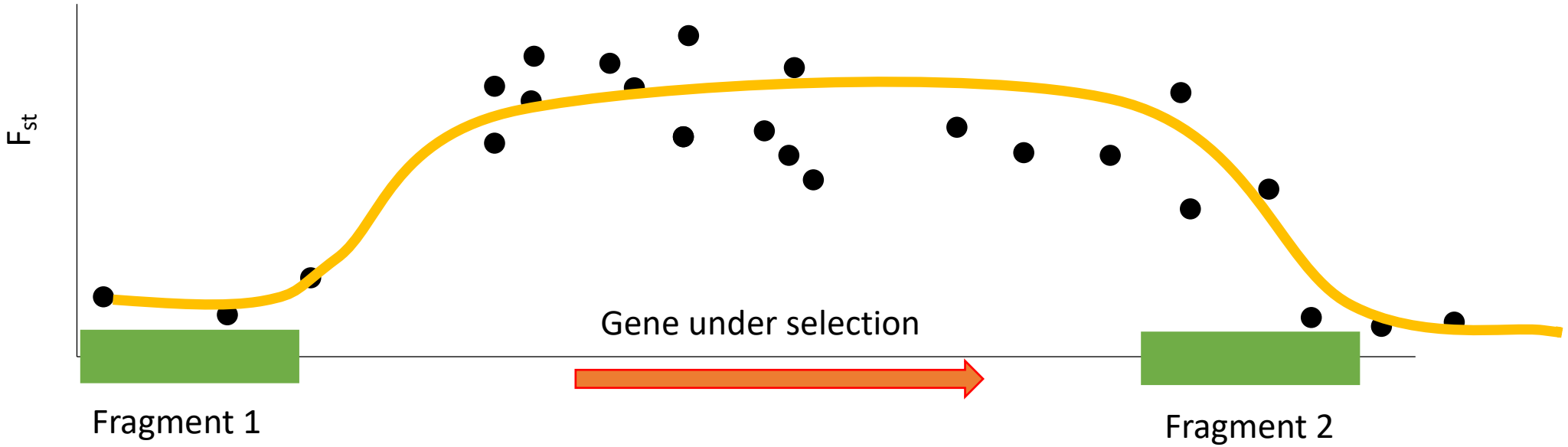
Looking for loci with increased differentiation (F_{st}) between species

Zemp et al., unpublished

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

Many samples

Use replicates, negative and positive controls.

Potential challenges

Allele dropouts

Biased population genetic estimators

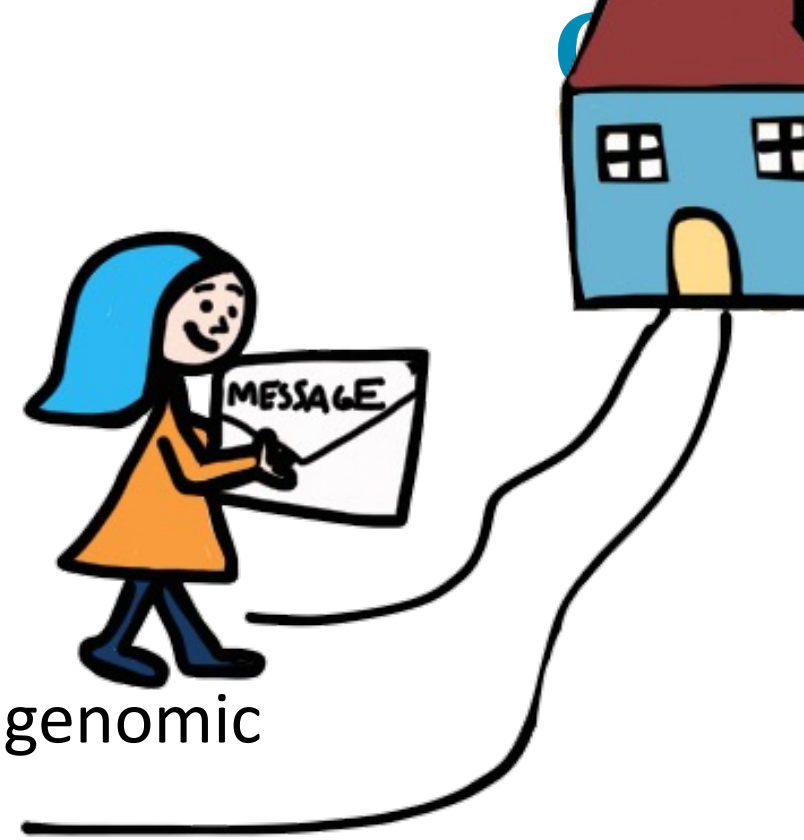
Often not in generich regions

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