SIBE
SOCIETÀ ITALIANA DI BIOLOGIA EVOLUZIONISTICA

Conservation Genomics

Claudia Fontsere, PhD

MSCA postdoctoral fellow

University of Copenhagen

claudia.fontsere@sund.ku.dk

CEH | CENTER FOR EVOLUTIONARY HOLOGENOMICS

MSCA

UNIVERSITY OF COPENHAGEN

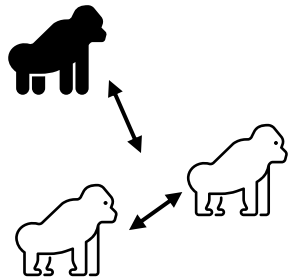
Outline

- Introduction to Conservation Genomics
  - Studying wild populations to preserve biodiversity
  - From conservation genetics to conservation genomics
- Why are genomes useful for conservation?
- Challenges: samples, data generation, assemblies and coverage
- Practical Case Study: Black Rhinos
  - Quality control
  - Population structure: PCA, ADMIXTURE
  - Quantifying temporal genomic erosion: heterozygosity, inbreeding
  - Technical aspects to consider
- Future perspectives and discussion

Studying populations

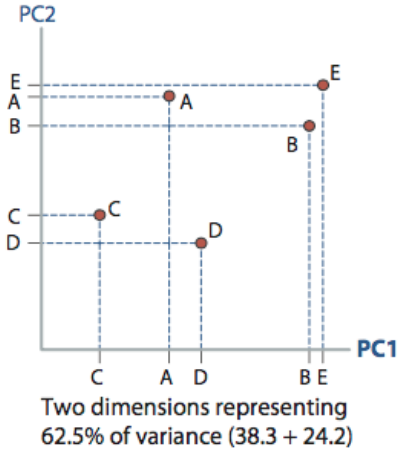
Population genetics:

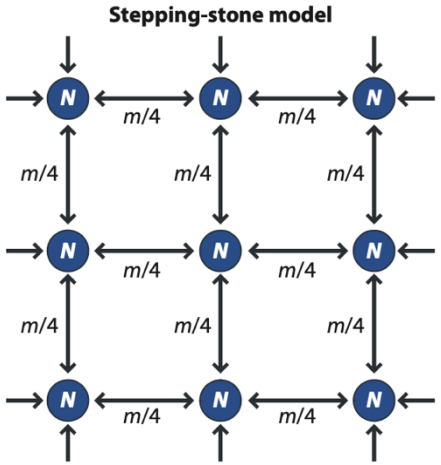
Genomics to understand population structure, migration, gene-flow… in wild populations

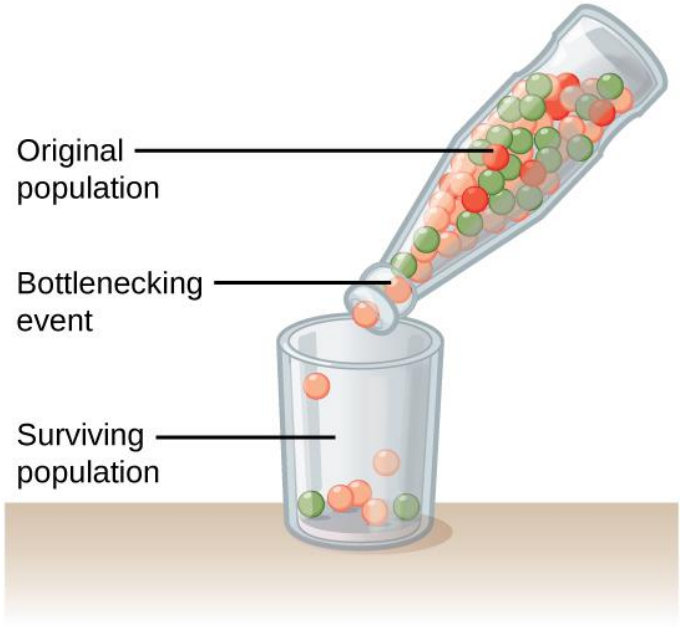


Conservation genetics:

Consequences of population decline



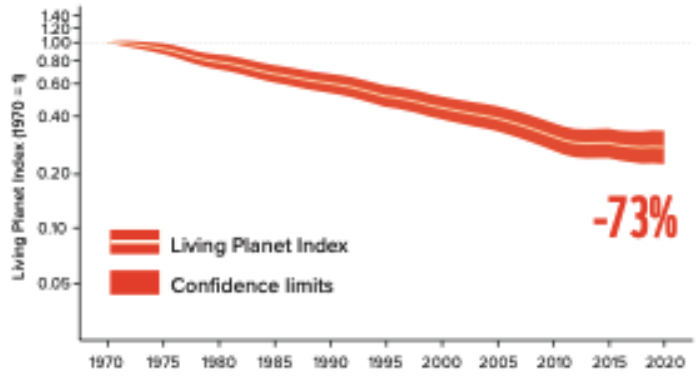




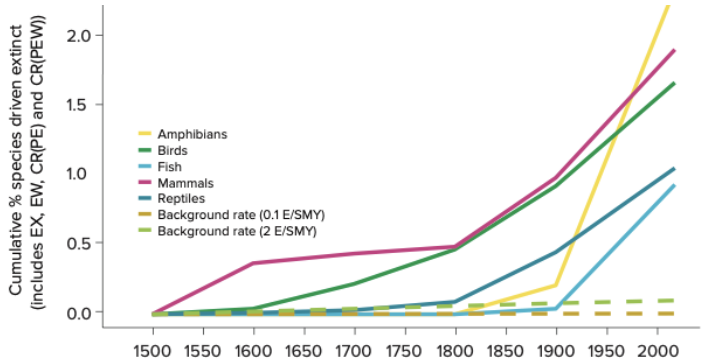
Preserving Biodiversity

The variability among living organisms including terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are a part.

Decline of wildlife and biodiversity with increase of rate of extinctions



a. Global Living Planet Index



d. Rate of extinctions

The Diversity of Biodiversity

Species Diversity

Population Diversity

Ecosystem Diversity

Ecosystem Functional Diversity

Why do populations decline?

Habitat loss, degradation and fragmentation

Overexploitation: Logging, mining, fishing, poaching, bycatch…

Climate change

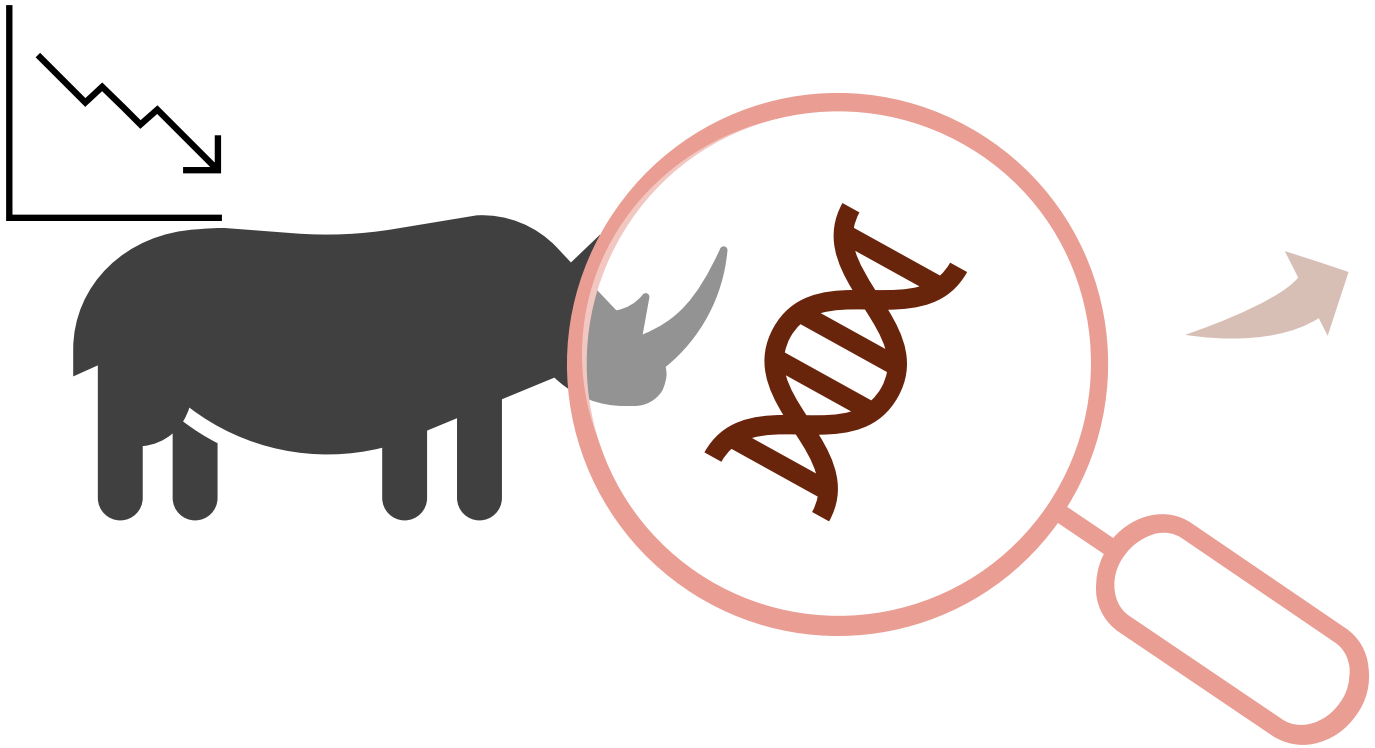
Pollution

Invasive species, genes

Disease

Conservation genetics

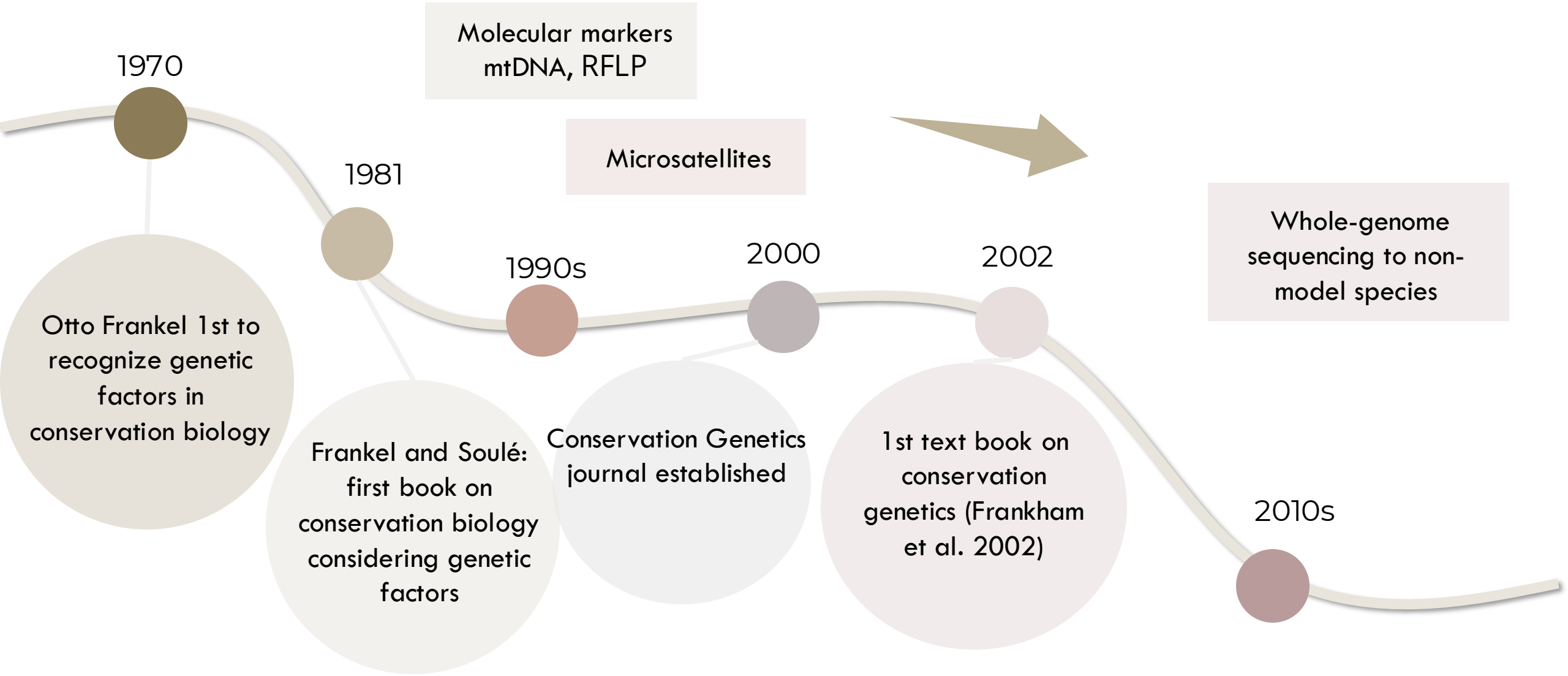
Application of genetics to understand and reduce the risk of population and species extinctions (Frankham, 2019)



Genetic analysis

Inform conservation plans

From early conservation genetics to genomics



Genetics vs Genomics

Genetic diversity has long been recognized as fundamental, but **genomics** is often neglected in biodiversity assessments and conservation efforts.

Genetic approaches

The study of individual genes and how they are passed down from generation to generation.

Use of small number of neutral markers

Genomic approaches

The study of genomes or the complete set of genetic material of an organisms (including non-coding regions)

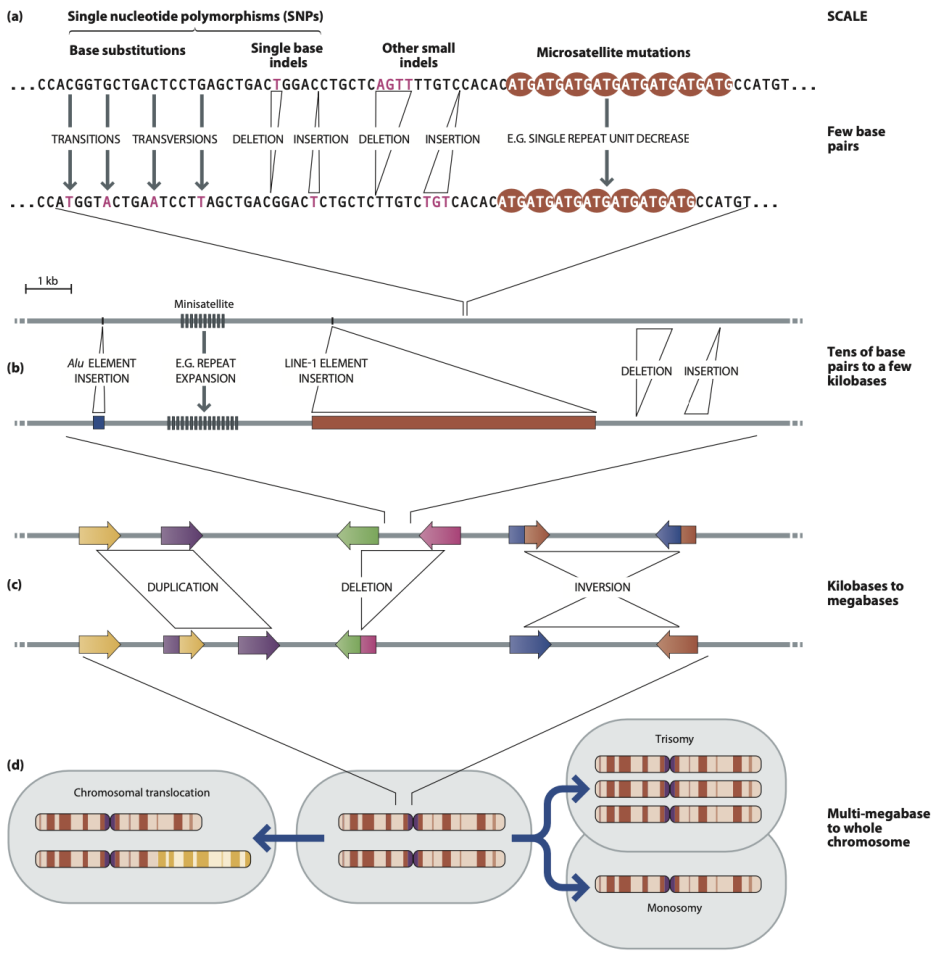
Use of complete genomes or genome-wide data

Principles of conservation genetics

Based on Population genetics:

- Change of allele frequencies through time and space to study evolutionary processes.
- Generate mathematical models than explain the reality.
- Population genetic models can be applied to data from any species.
- Population is a central concept.
  - Define what constitutes a population
  - Sometimes a population is more interesting than a single individual
  - But sometimes it is complicated to delineate.

Types of Genomic Variation



Single Nucleotide Variant or Polymorphism

Indels

Microsatellites

Duplications / Deletions / Inversions

Chromosomal translocations, duplications

The Era of Conservation Genomics

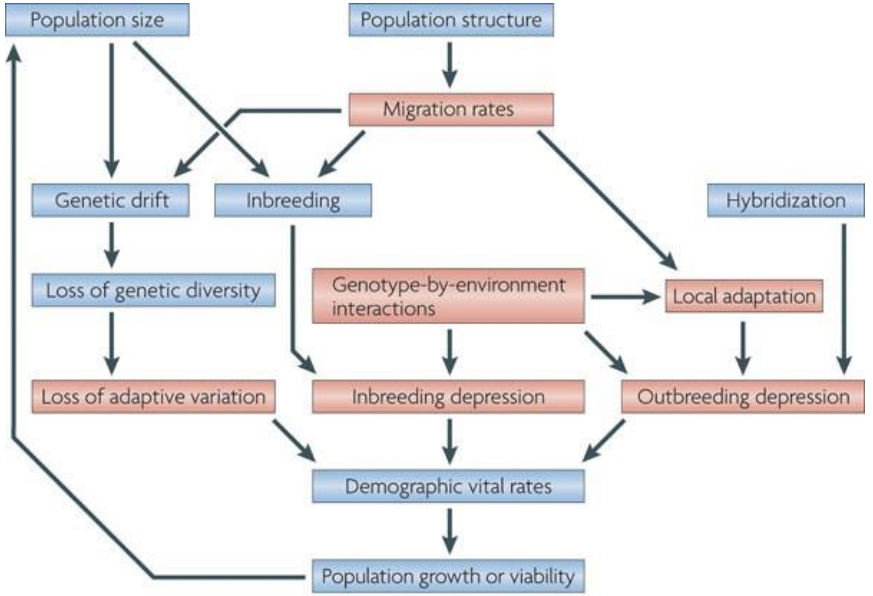
Use of complete genomes or genome-wide data

Increased density of markers across the genome

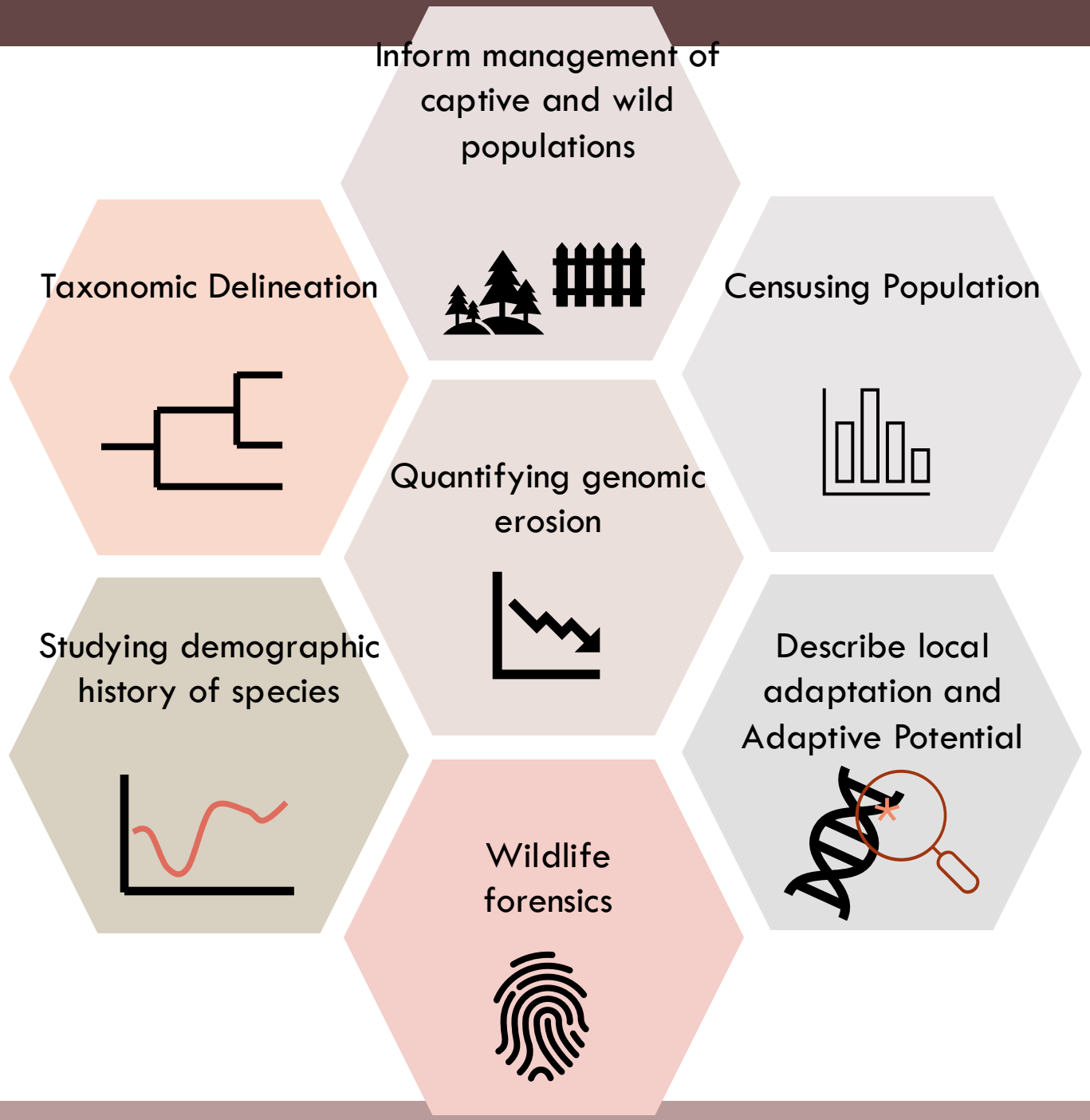
More accurate estimators of population parameters

Better inform conservation measures

- Examine different evolutionary mechanisms.
  - Neutral loci
  - Protein-coding regions
  - Non-coding regulatory regions that control gene expression
  - Whole-transcriptome sequencing allows the quantification of gene expression differences

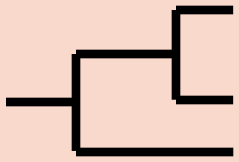


Why are genomes useful for conservation?



Aiding in Taxonomic Delineation

Taxonomic Delineation



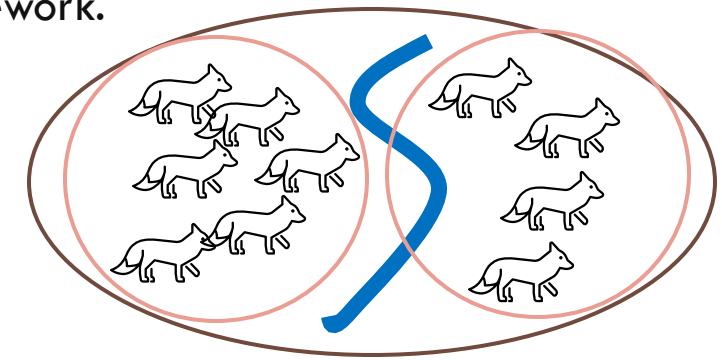
Necessary for the current conservation regulatory framework.

What is a species (and subspecies, population…)?

Biological Species Concept

Phylogenetic Species Concept

Many more…

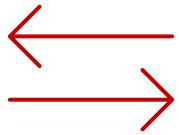


Outbreeding depression vs geneti
rescue blocked

Evolutionarily significant unit (ESU):

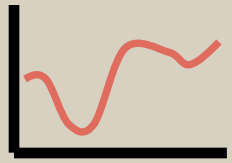
- substantially reproductively isolated from other conspecific population units
- represents an important component in the evolutionary legacy of the species

Admixture & Introgression



Demographic history

Studying demographic history of species



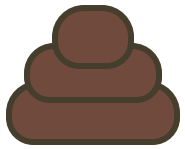
Understanding some key demographic traits that are unknown or difficult to obtain with other methods.

- Estimate effective population size
- Migration and Gene Flow
- Drift
- Detecting Introgression and Genetic Swamping
- Mating Systems
- Paternity

Censusing populations

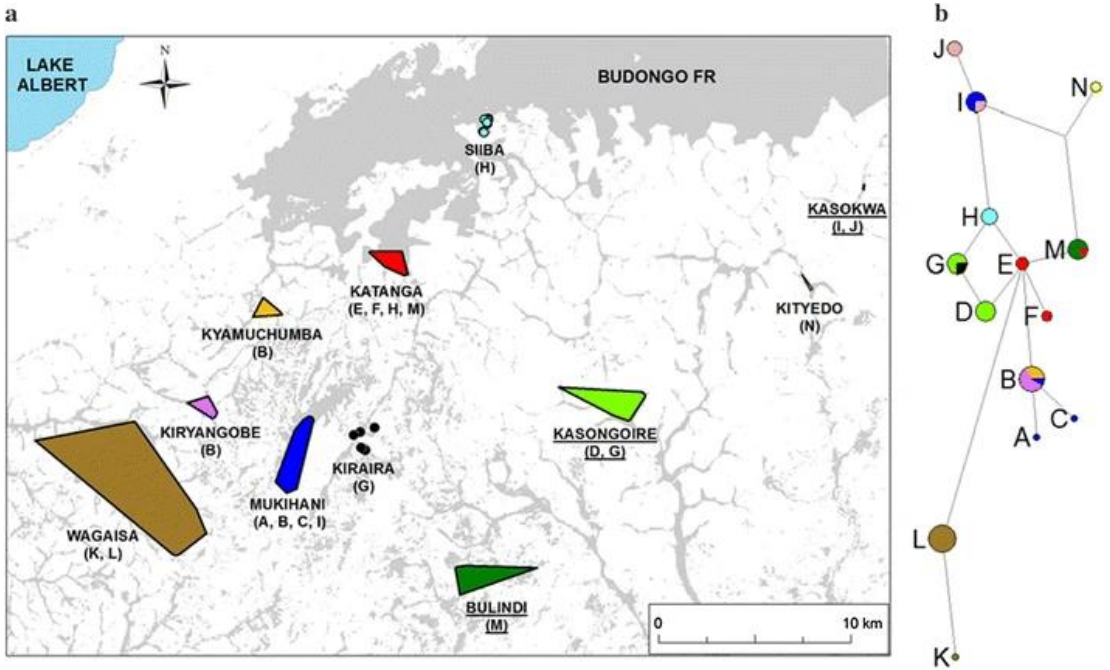
Censusing Population

Non-Invasive Sampling



Microsatellite genotyping

- Neutral
  - Number of individuals
  - Population belonging
  - Degree of relatedness
- Sex-linked
  - Determine sex
  - ChrY haplotypes



McCarthy et al (2015)

Genomic erosion

Quantifying genomic erosion

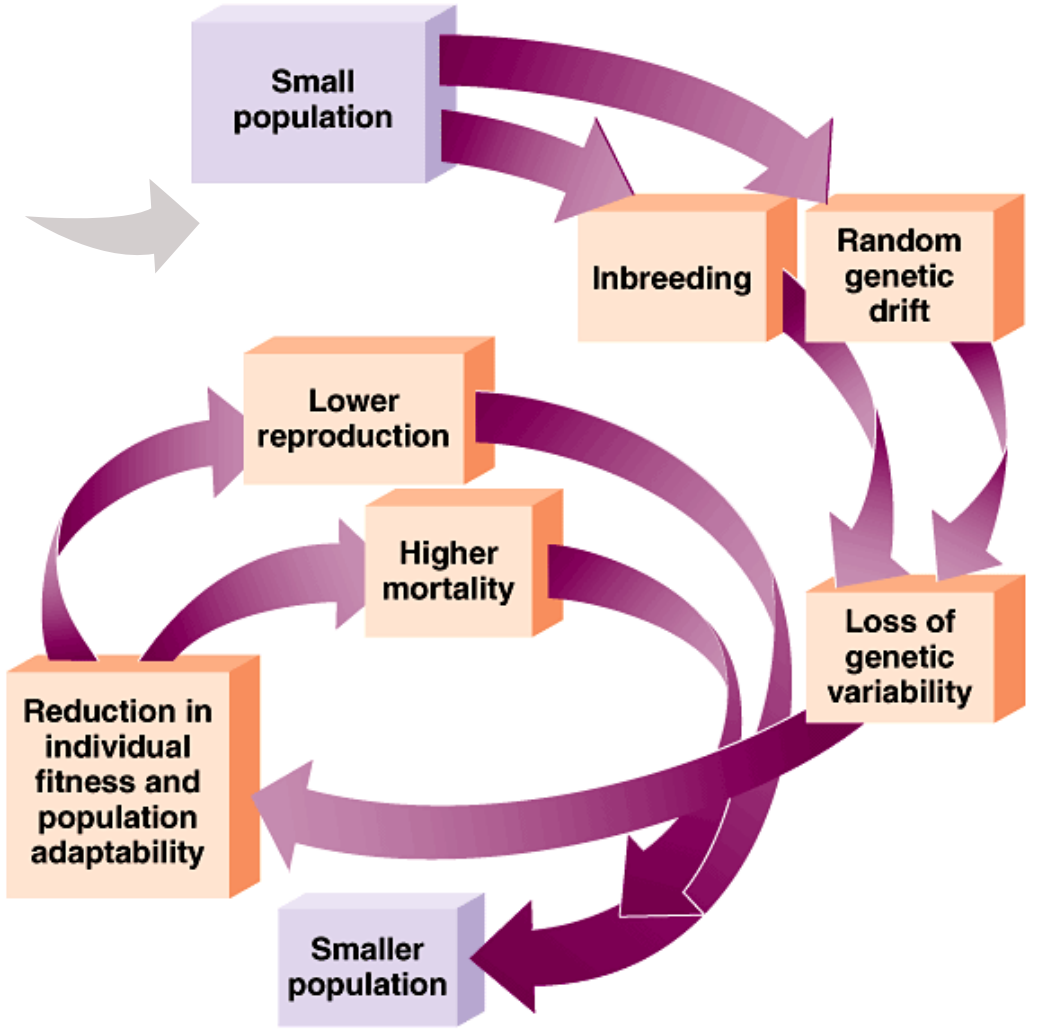
- Extinction risk linked to genetic factors:
  - Inbreeding
  - Loss of genetic diversity
  - Accumulation of harmful mutations → Genetic load

Genomic erosion is often only noticeable many generations after the onset of the immediate threats that lead to population decline

Time lag between demographic and genomic impact

Genomic erosion – Extinction Vortex

Environmental/anthropogenic factors:
- Reduction in population size due to
- Fragmentation without migration

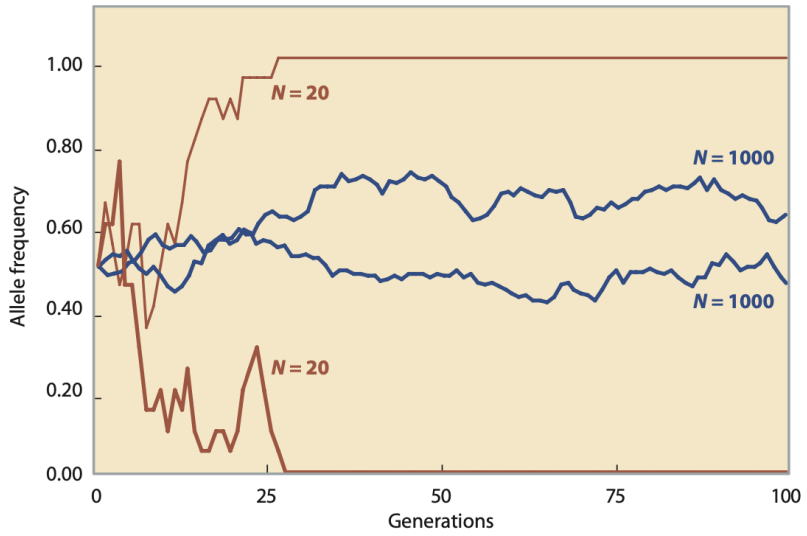


Small effective population siz (Ne) are more vulnerable to drift and inbreeding

Inbreeding depression

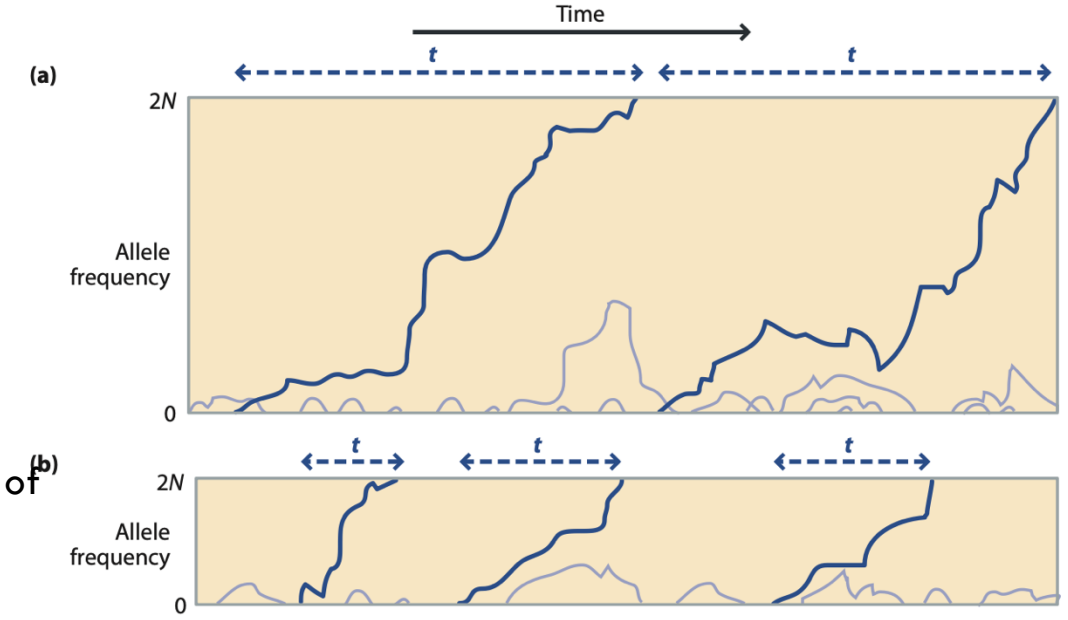
Expression of harmful mutations

Genomic erosion – Drift



- Random process
- Outcome depends on population size

- Fixation:
  - At low size, the deviation increases
  - Loss of one allele by chance
  - Lost is lost
  - t=4Ne
    - The concept of effective population size allows us to calculate the probability and rate of fixation in absence of selection and mutation



Genomic erosion – Genetic Diversity

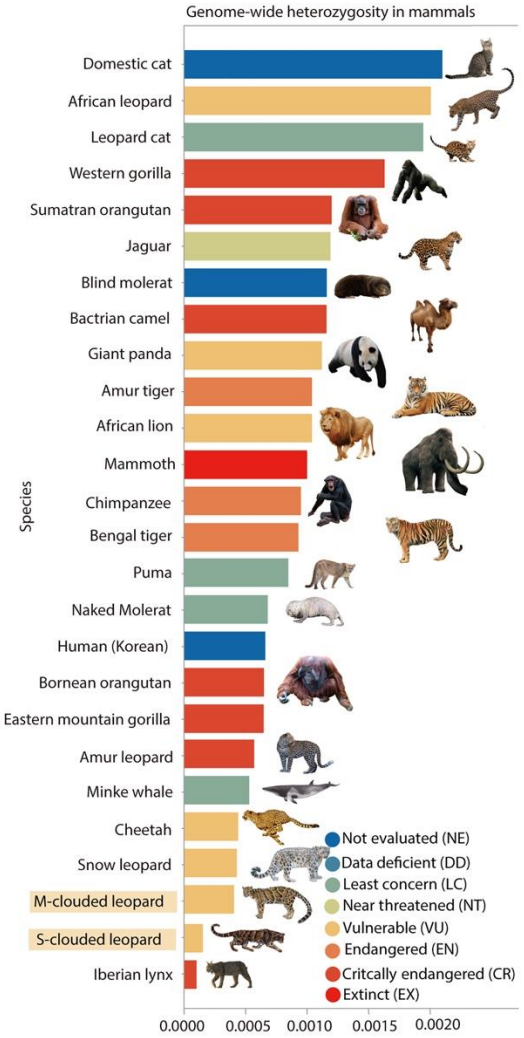
Genetic Diversity: amount of genetic variation (e.g. heterozygosity, nucleotide diversity) within a population, species…

Nucleotide diversity (π) $= 4\mu Ne$

Necessary for long-term survival and a key component of preserving biodiversity. Until now, it has rarely been integrated into conservation policies.

Genetic diversity has been recognized as one of three important levels of biological diversity by the Convention on Biological Diversity (2022)

Debate wheter if conservation should focus on loci with adaptive potential, or overall genomic diversity

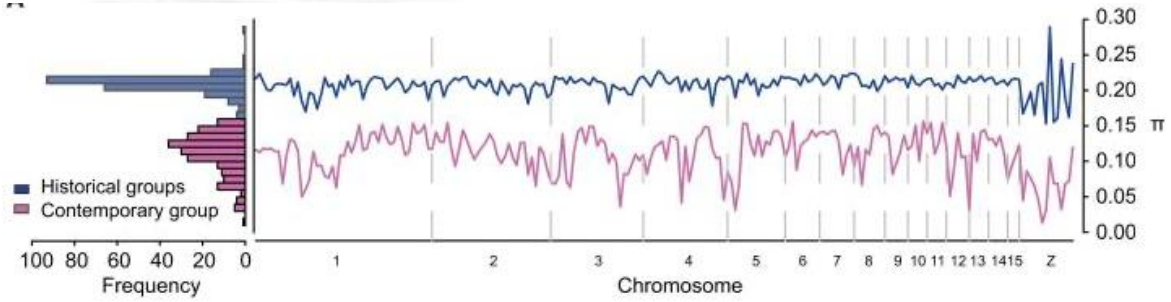


Yuan et al (2023)

Genomic erosion – Genetic Diversity

Genetic Diversity: amount of genetic variation (e.g. heterozygosity, nucleotide diversity) within a population, species…

Crested Ibis



Feng et al (2019)

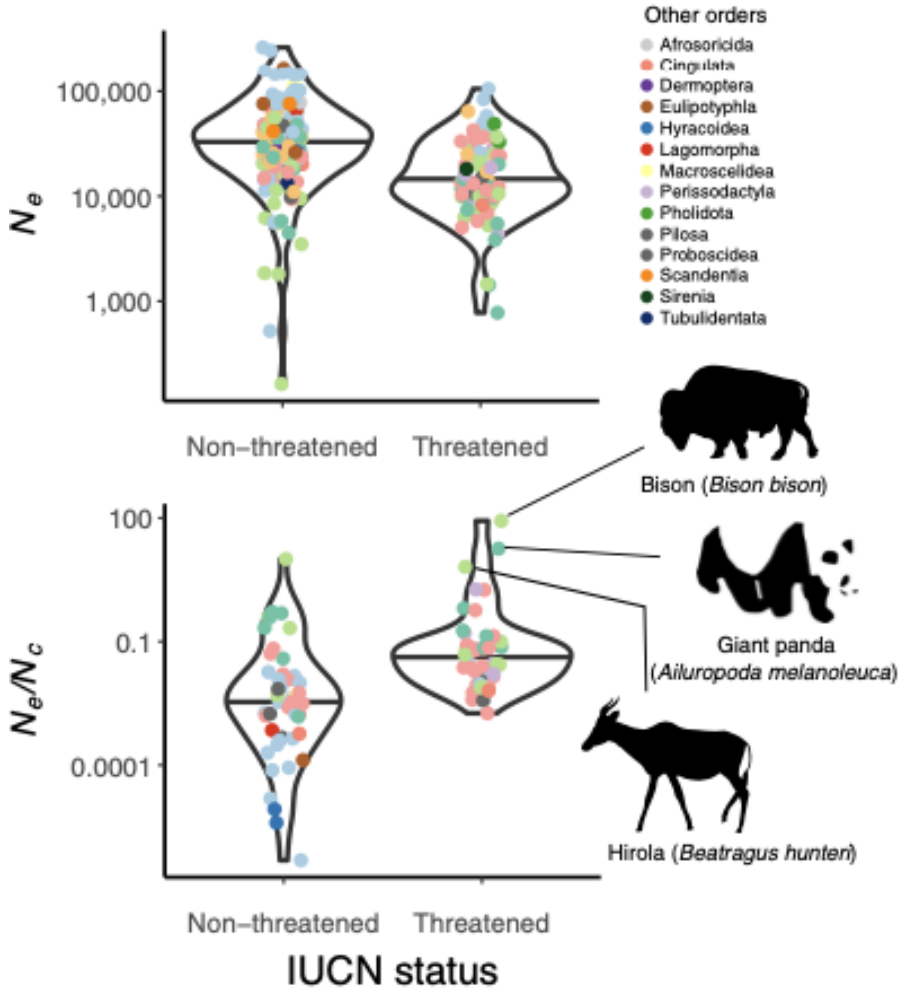
Loss of genetic diversity through time

Causes:

- habitat fragmentation
- founder effect
- population bottleneck

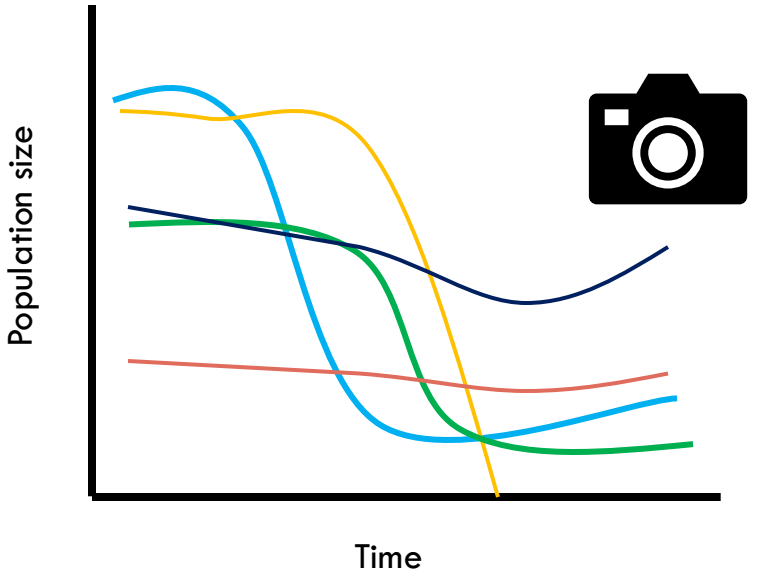
Directly related to effective population size (Ne)

Extinction Risk linked to Genetic Factors



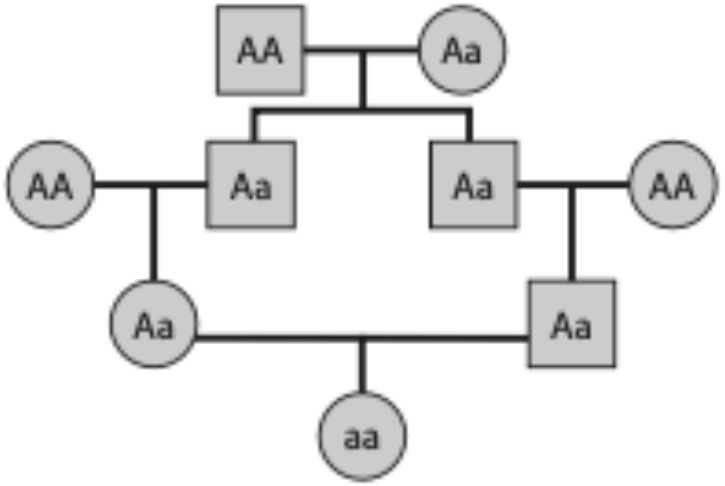
Wilder et al (2023)

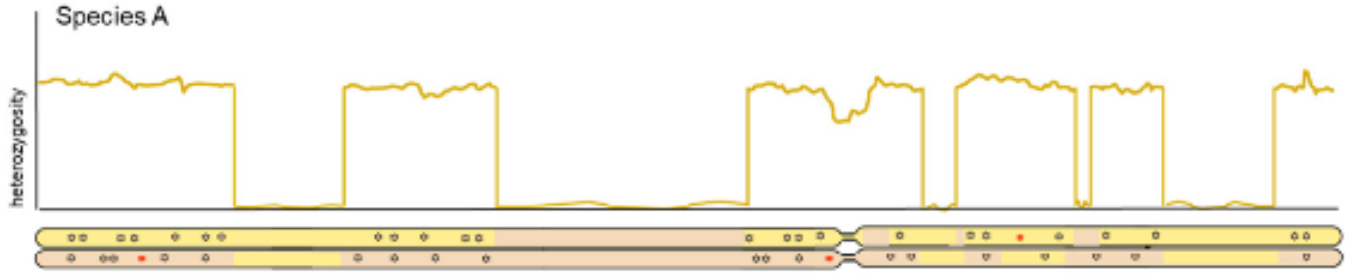
Effective population size: size of an idealized population that would experience the same rate of genetic drift or inbreeding as in the real population.



Genomic erosion - Inbreeding

Inbreeding: Inbreeding is the production of offspring from individuals that are related by (recent) descent from a common ancestor of their parents.
Inbreeding depression: Reduction in fitness for a quantitative trait due to inbreeding, especially manifest in reproductive fitness traits.





Bosse & van Loon (2022)

Genomic erosion - Genetic load

- **Fitness effects** of a particular genetic variant in a protein-coding sequence can be predicted from models of protein structure and by comparing the level of sequence conservation across species.
- Combining this information with patterns of inbreeding across the genome can identify candidate loci underlying inbreeding depression.

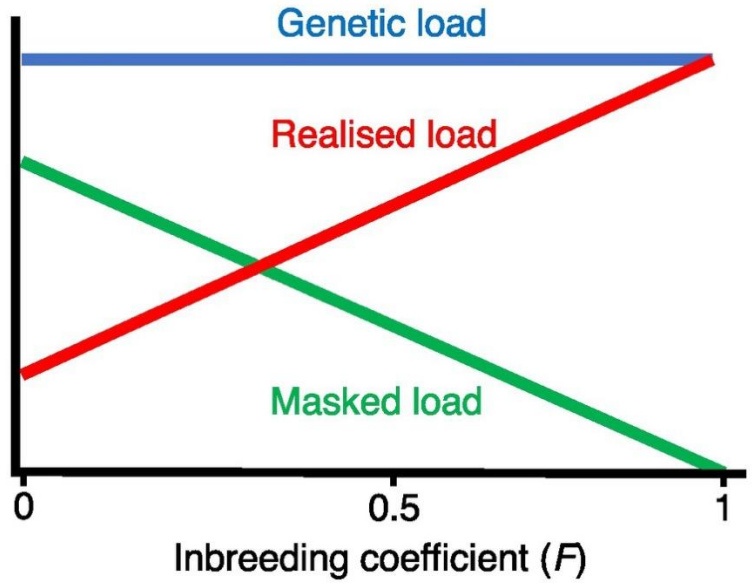
Genetic load: Masked Load + Realized Load

Genetic load is a major threat to small populations

- Drift Debt
  - Most deleterious mutations are initially rare and it can take many generations of drift for them to increase in frequency.
  - When they rise in frequency, they become homozygous and reduce the mean population fitness

Inbreeding can accelerate this process, resulting in **inbreeding depression**

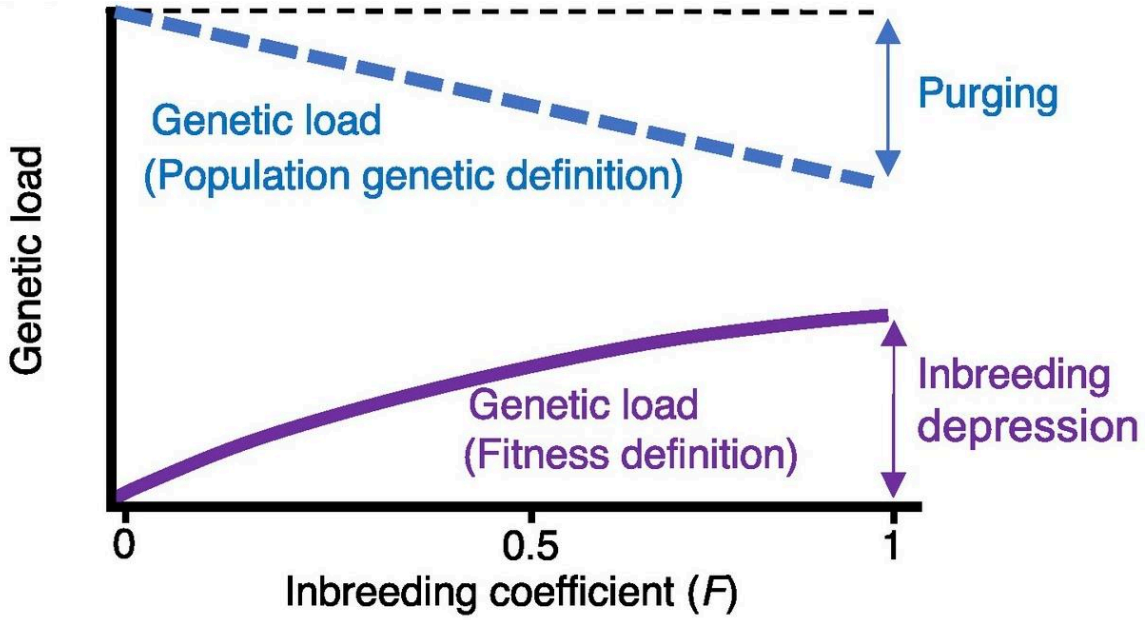
Conversion from Masked to Realized load



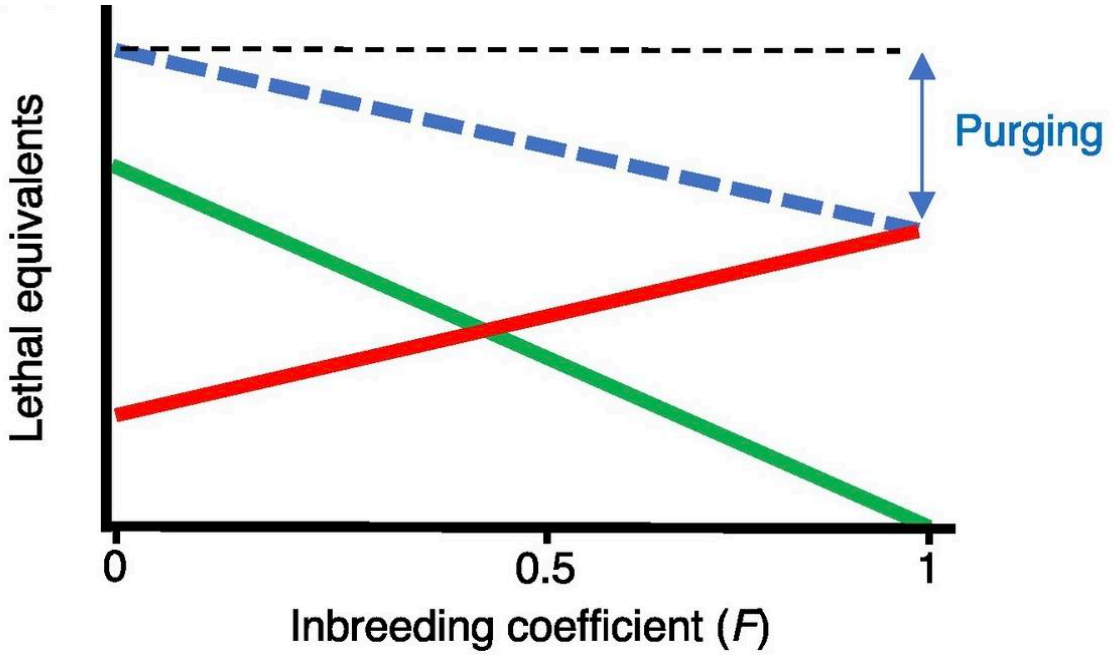
Dussex et al (2023)

Genomic erosion - Genetic load

CAUTION! Different definitions



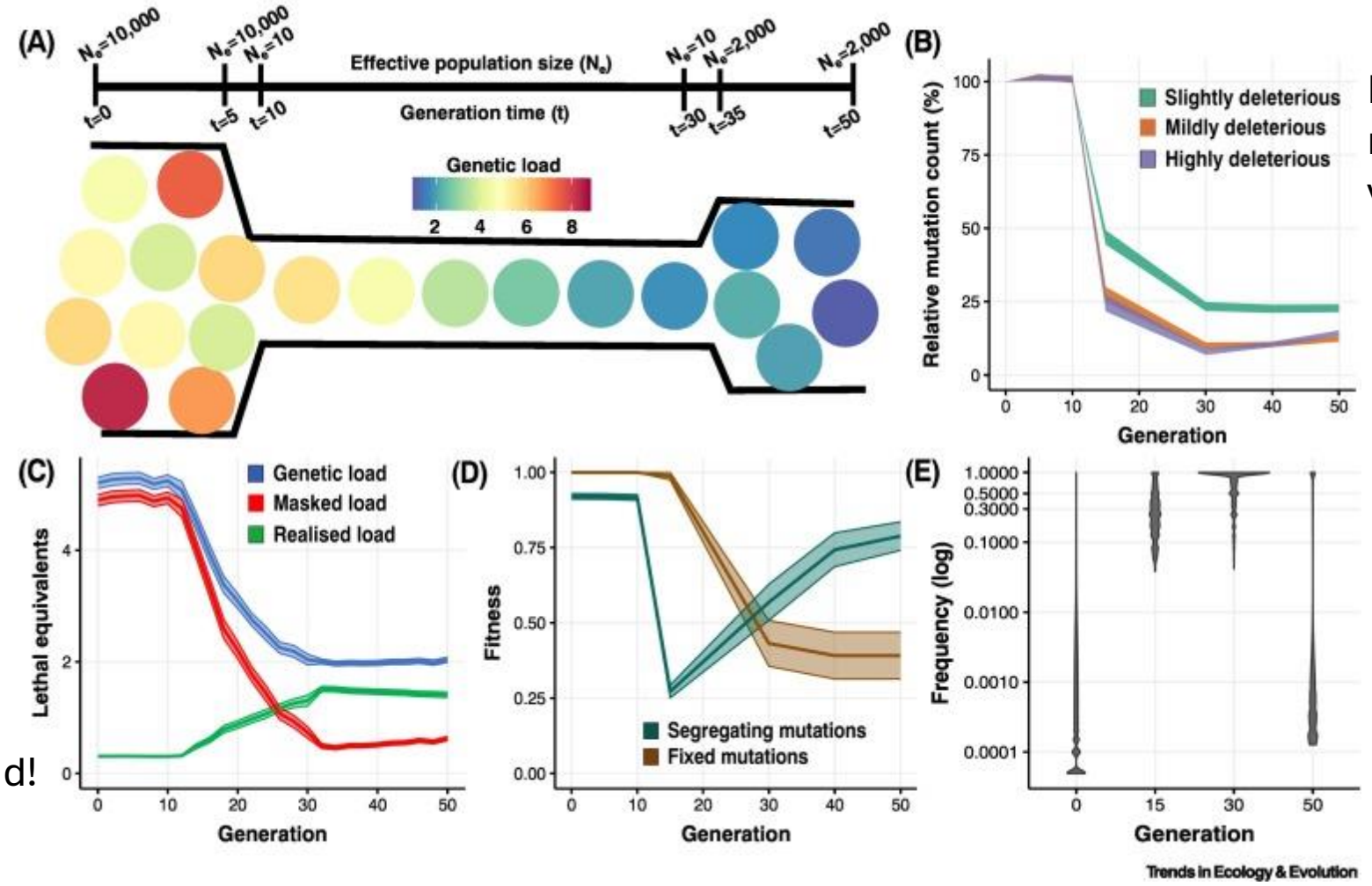
When selection happens....



Dussex et al (2023)

Genomic erosion - Genetic load

Dynamics of purging and accumulation of the genetic load



Drift and purging reduces number of deleterious variants

But… drift makes **purifying selection** less efficient, so some of the less deleterious mutations increase in frequency

Increase realized load!

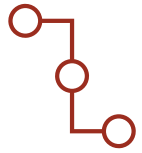
Dussex et al (2023)

Local adaptation and Adaptive Potential

Describe local adaptation and Adaptive Potential

- **Adaptation** is a genetic process that allows a species to persist for generations in a changing habitat.

Genomic region



Ecologically important trait

- Genetic diversity is the substrate of evolution: The ability to evolve depends on genetic diversity for fitness, selective forces, and genetically effective population size of populations

Threatened species

Local adaptation and Adaptive Potential

Local genetic adaptation to habitat in wild chimpanzees

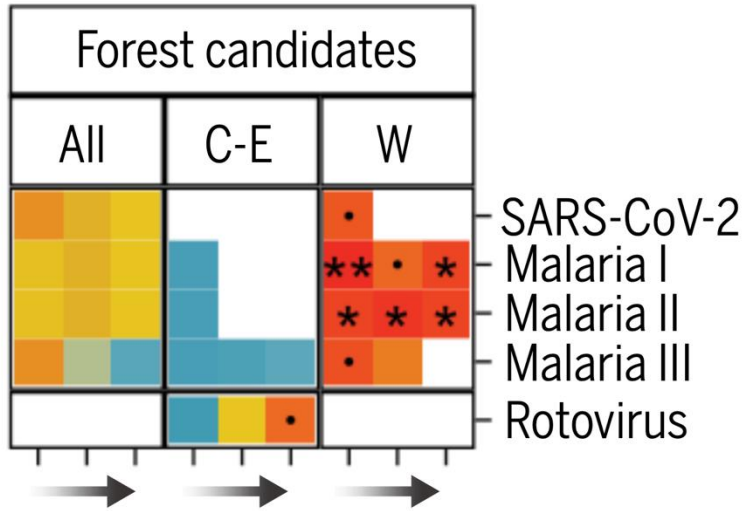
HARRISON J. OSTRIDGE, CLAUDIA FONTSERE, ESTHER LIZANO, DANIELA C. SOTO, JOSHUA M. SCHMIDT, VRISHTI SAXENA, MARINA ALVAREZ-ESTAPE, CHRISTOPHER D. BARRATT, PAOLO GRATTON, [...], AND AIDA M. ANDRÉS +74 authors Authors Info & Affiliations

SCIENCE • 10 Jan 2025 • Vol 387, Issue 6730 • DOI: 10.1126/science.adn7954

Exome sequencing from 388 fecal samples

Forest-dwelling populations → genetic adaptation to malaria resistance, similar to human

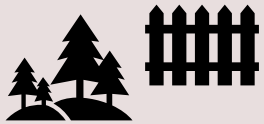
Enriched for pathogen-related genes in forests



Preserving adaptive genetic variation to support resilience against diseases and environmental change

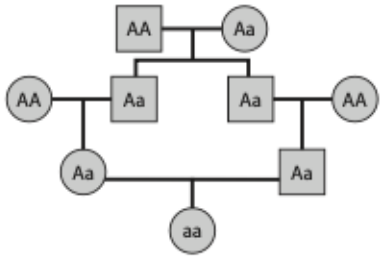
Genetic management of captive populations

Inform management of captive and wild populations



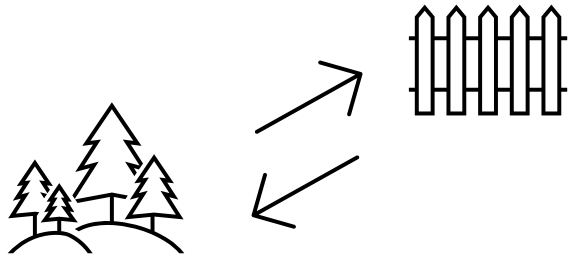
- Genetic issues in management of endangered captive populations:
  - inbreeding depression
  - loss of genetic diversity
  - genetic adaptations to captivity

Kinship assessement



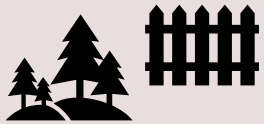
Founder relationships

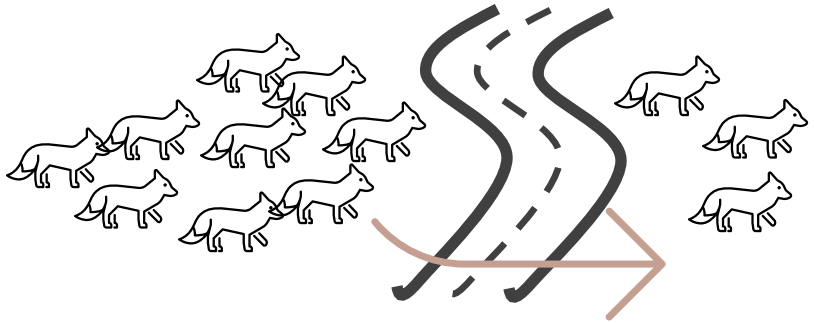
Population structure



Genetic management of wild populations

Inform management of captive and wild populations





Genetic Rescue

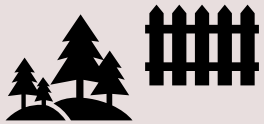
- Genetic management of wild population has been slow to take off (e.g. IUCN does not use genetics).
- Fragmented populations → genetically isolated

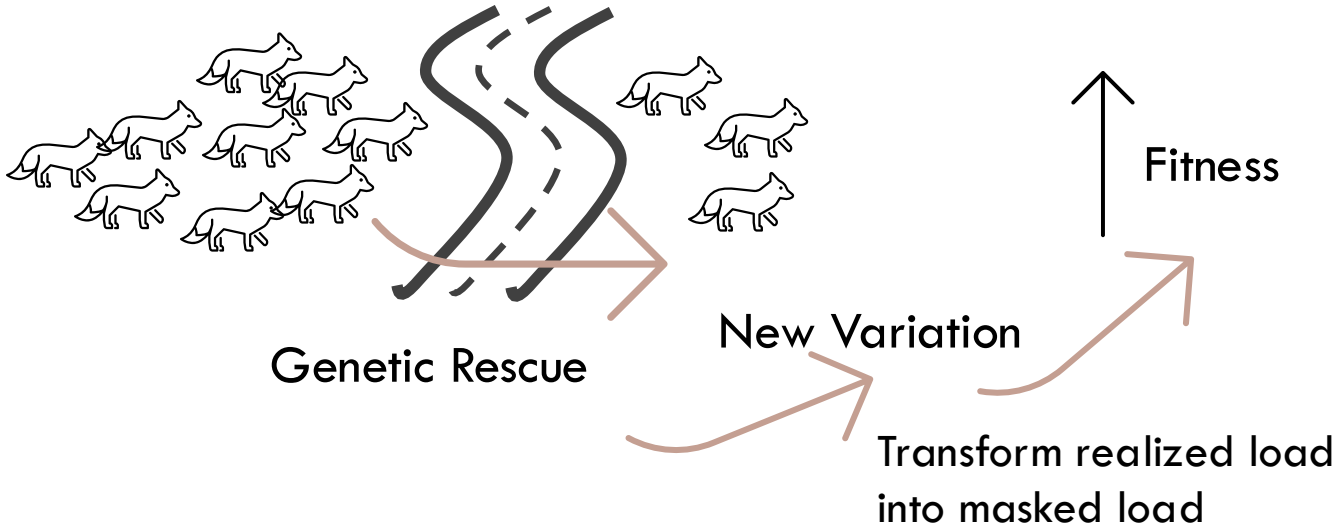
Adverse genetic impacts in these occur at a rate that is much more rapid than in the species as a whole.

- Rescued by gene flow from other population segments (genetic rescue).

Genetic management of wild populations

Inform management of captive and wild populations





- **Genetic rescue** is expected to be most useful for small, isolated populations that suffer from inbreeding

GENOMICS

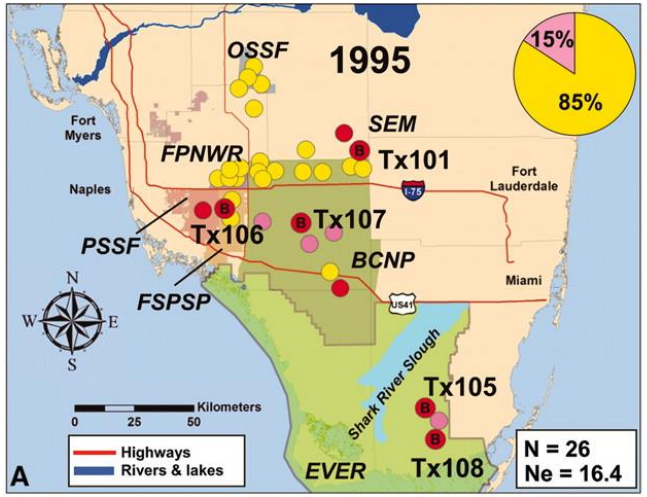
- Concerns over outbreeding depression
  - Important to understand the level of inbreeding in the population, which depends on the size of the population and its demographic history.

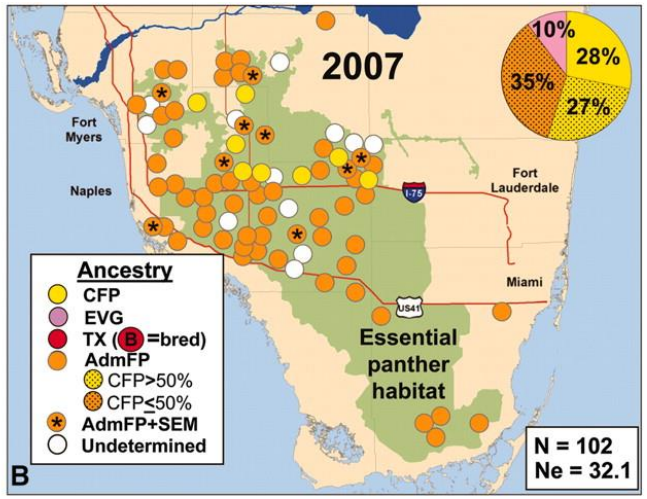
Genetic management of wild populations

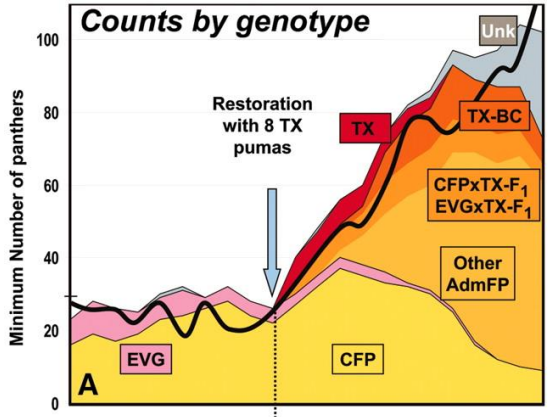
- Florida Panther – genetic rescue

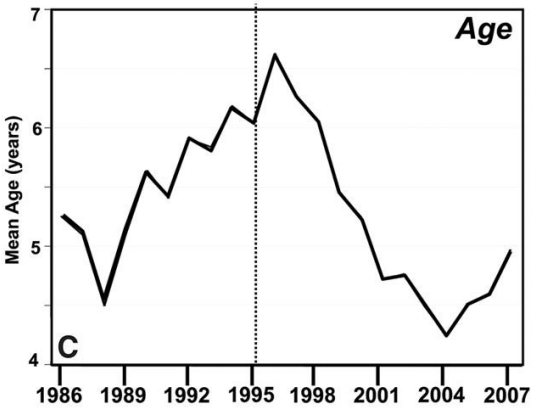
In the 1990s, <30 individuals suffering from inbreeding depression

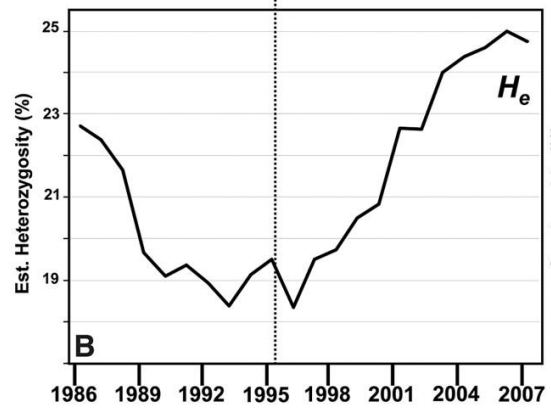
8 individuals from Texas translocated to Florida in 1995

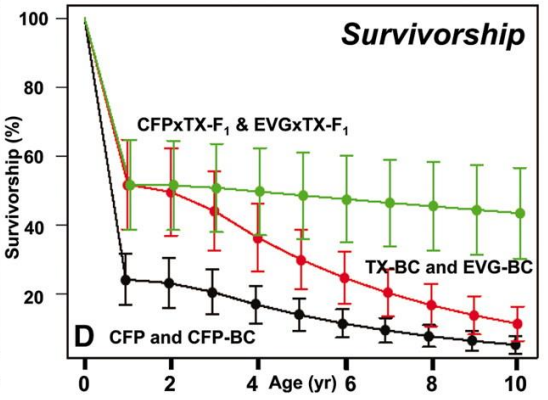










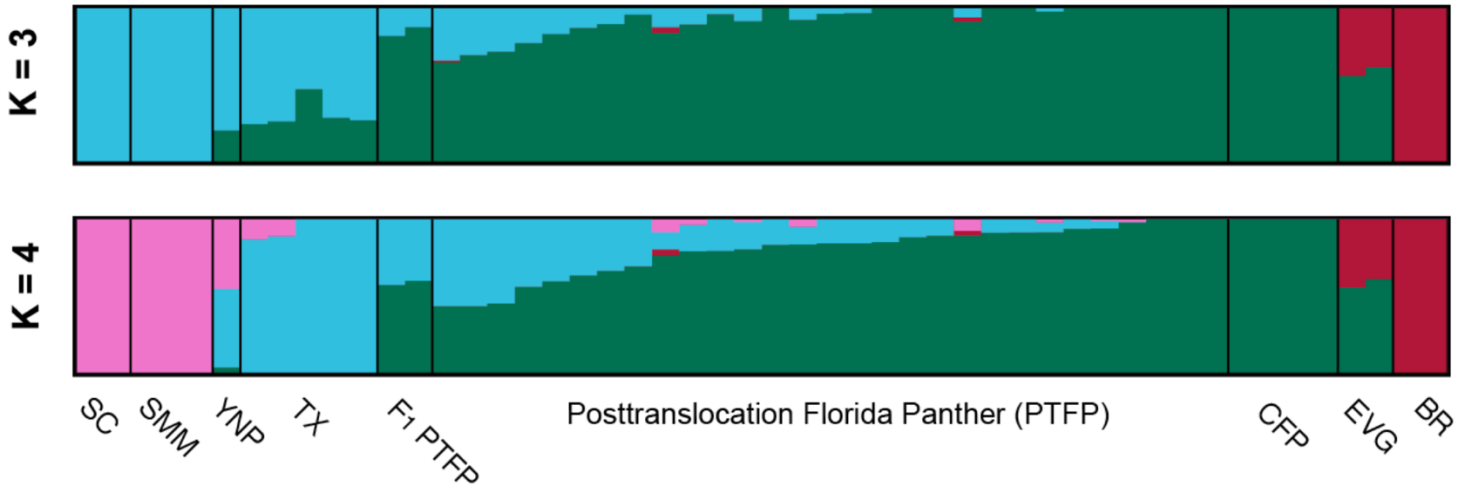


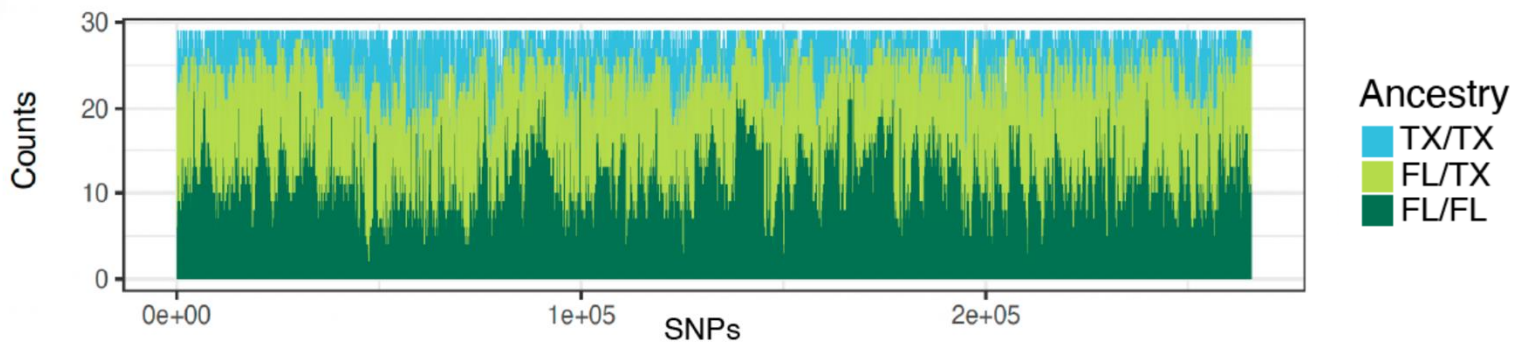
Johnson et al (2010)

Genetic management of wild populations

- Florida Panther – genetic rescue

Now with WG!



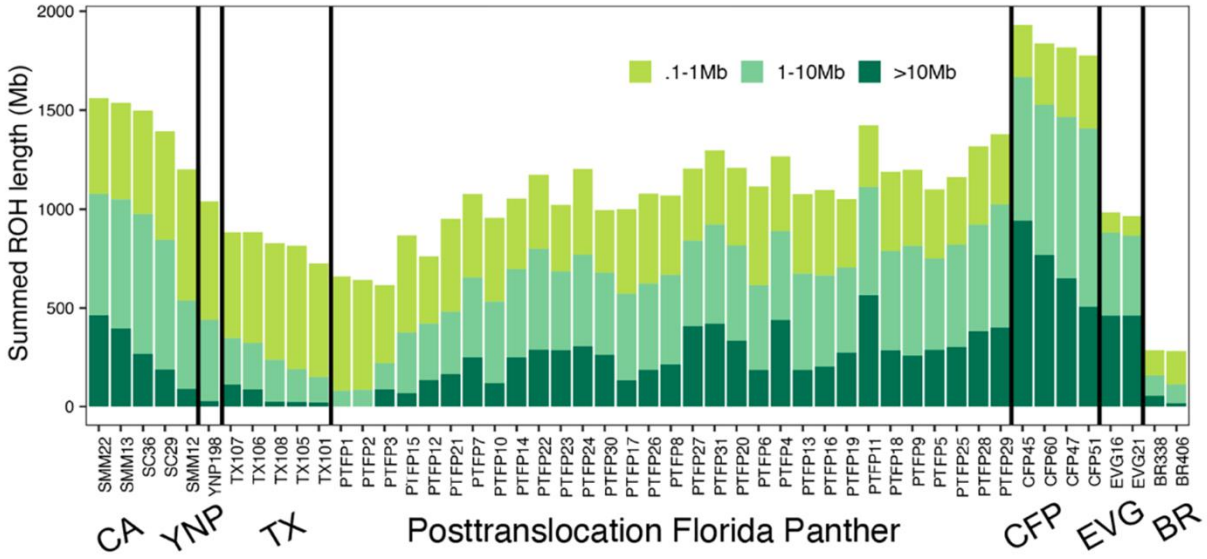


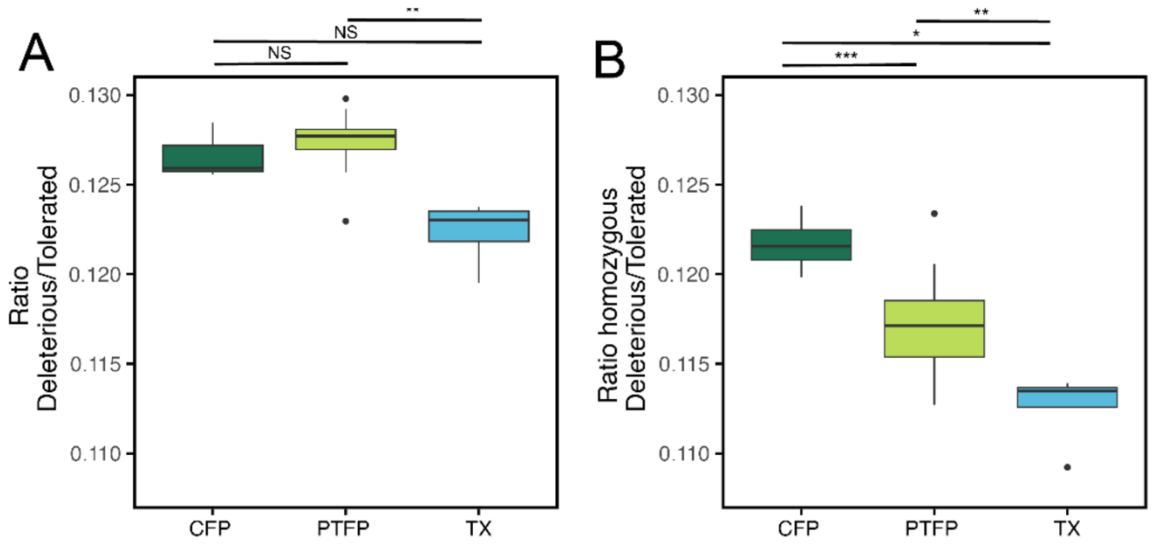
Aguilar-Gómez, et al. (2024)

Genetic management of wild populations

- Florida Panther – genetic rescue

Now with WG!





Reduction of Homozygosity

→ alleviate recessive deleterious load

→ not a reduction in number of deleterious variants

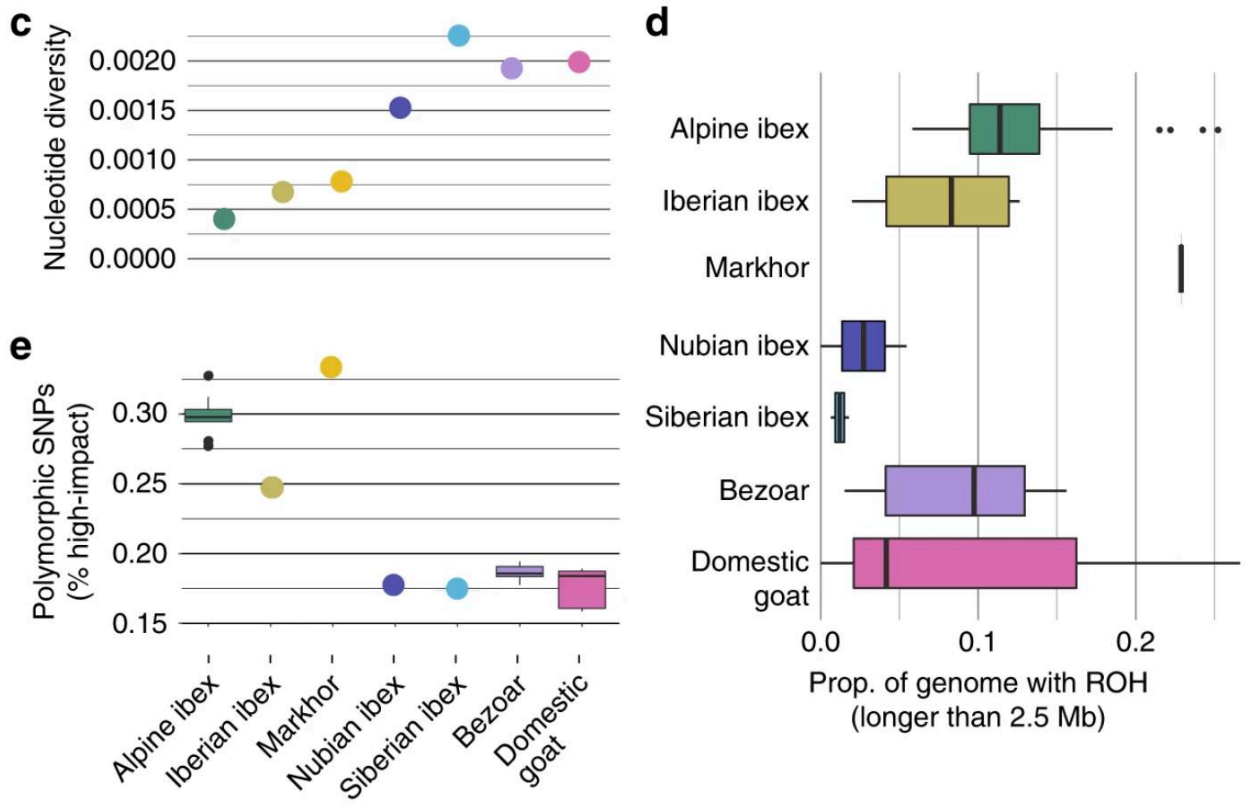
Aguilar-Gómez, et al. (2024)

Genetic management of wild populations

Alpine Ibex

~100 individuals in the 19th century in a single population in the Gran Paradiso region of Northern Italy.

Recovered to 50,000 individuals



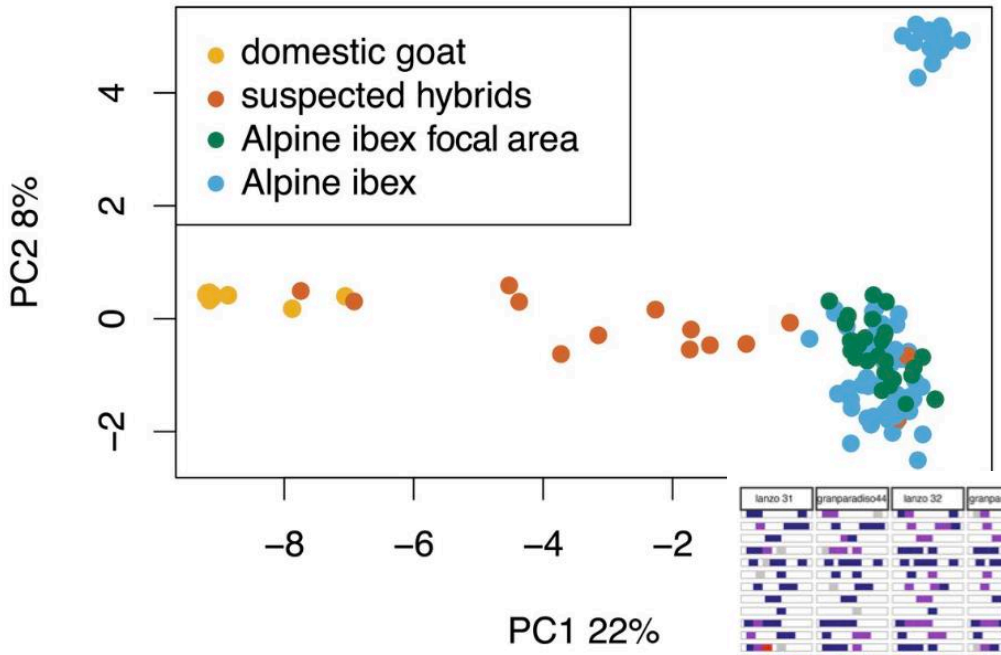
Grossen et al. (2020)

Genetic management of wild populations

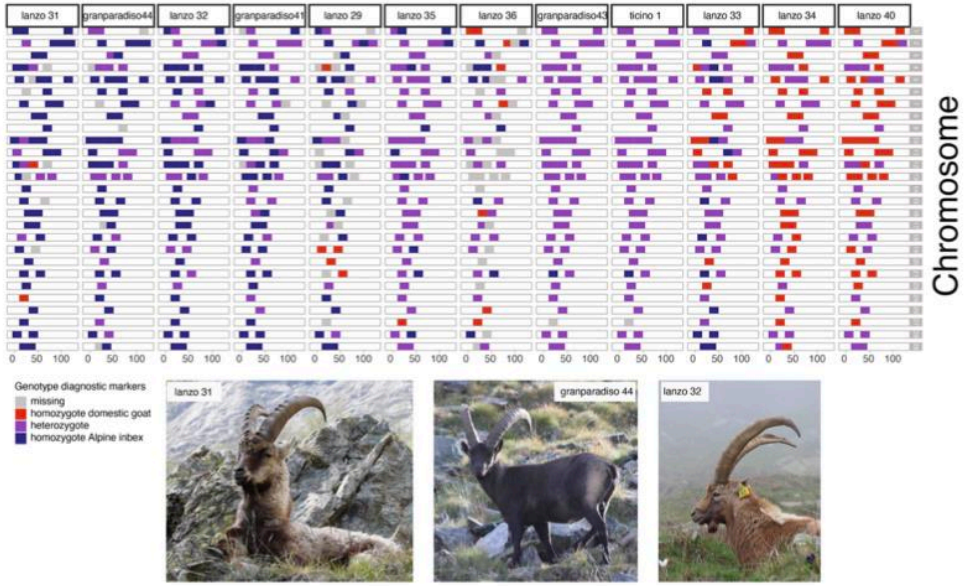
Alpine Ibex

~100 individuals in the 19th century in a single population in the Gran Paradiso region of Northern Italy.

Recovered to 50,000 individuals

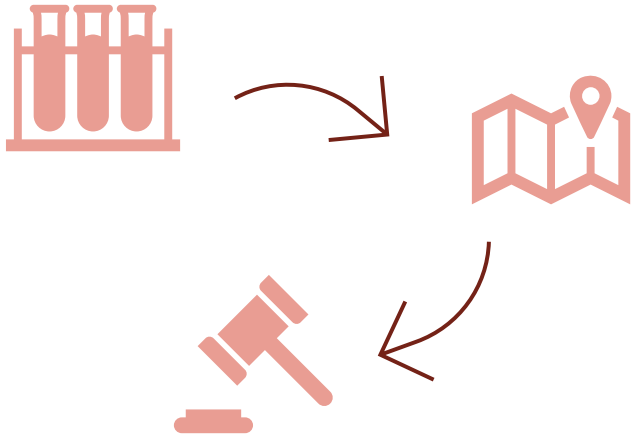


Hybrid swarm between Alpine ibex (*Capra ibex*) and domestic goat (*C. hircus*)



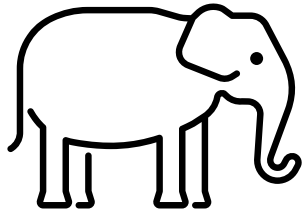
Grossen et al. (2020)

Wildlife Forensics



Wildlife forensics

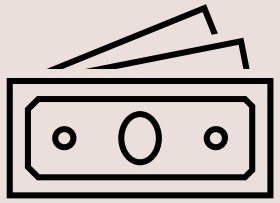
- Poaching and wildlife trade
- Species can be identified from DNA: hair, horns, ivory, meat, eggs, turtle shells and plant material.



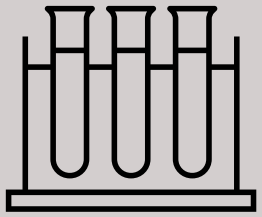
Wasser et al (2008)

Challenges in the implementation

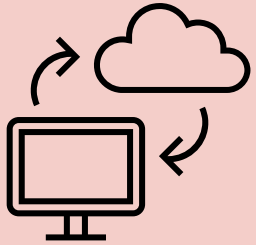
Limited budget



Sample availability

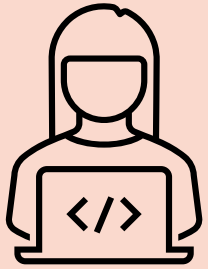


Data analysis

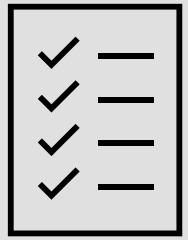


Reference genome

Non-user friendly software



Interpretation and translation



Samples

- The issue is always… sampling!

Modern samples

Invasive Samples

Blood
Tissue

High

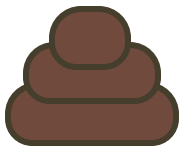
Endogenous DNA
DNA quality
DNA quantity

Reference genome
High Coverage
Long-read sequencing
Multi-omics
Low contamination

Ethics / Not allowed
Disturbance of the animals
Sometimes only for captive populations
Preservation in the field is crucial

Samples

- The issue is always… sampling!



Modern samples

Non-Invasive Samples

Feces
Hair
Feathers
Saliva

Low

Endogenous DNA
DNA quality
DNA quantity

Increase local resolution

Plenty of samples for wild
Geolocalized
No disturbance

Low coverage
Damage patterns

PCR-based studies

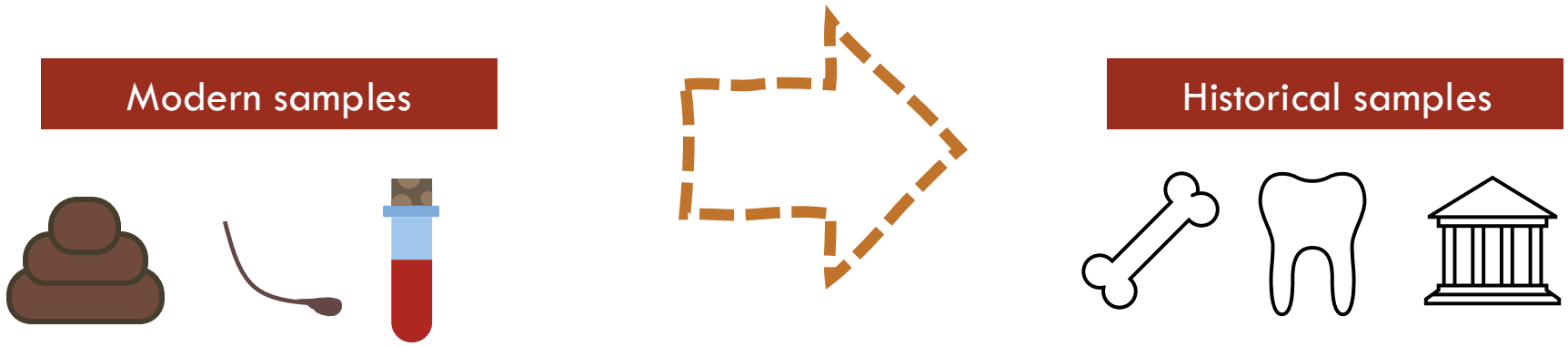
Autosomal markers
Microsatellites
mtDNA

Target capture methods

Whole genome
Exome
Specific target regions

Samples

- From spatial sampling to temporal sampling.



- Generate a genetic baseline before loss of population.

Samples

- From only spatial sampling to temporal sampling.

Historical samples

Formalin-fixed
Ethanol-preserved
Skin
Bone
Teeth

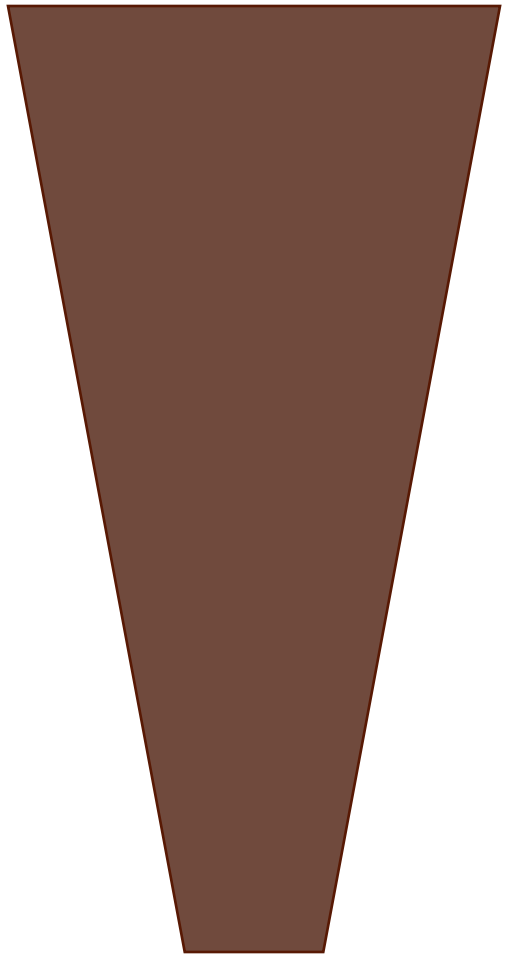
Low

Endogenous DNA
DNA quality
DNA quantity

Sampling from already extinct populations or difficult to access
Temporal comparisons

Damage patterns
Fragmented DNA
Usually low coverage
Risk of human contamination
Requires aDNA protocols

Data Generation for Genome Sequencing



Whole-genome (high vs low coverage)

Reduced representation (RAD-seq, ddRad)

Target-Capture approaches:

- Exome
- Region-specific
- SNP arrays

Not covering mtDNA seq, eDNA, metagenomics, genome assembly

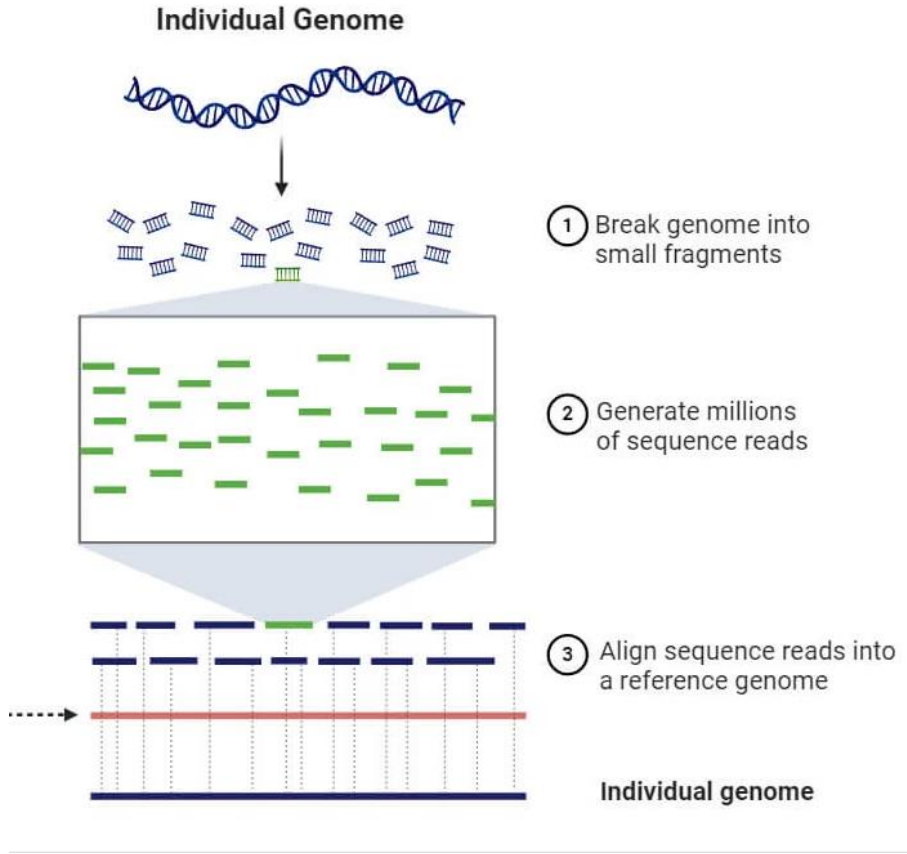
Whole-genome sequencing

Sequencing the entire nuclear genome with high-thoughput platforms (e.g. Illumina).

Reads are mapped to a Reference Genome

Coding and non-coding regions
Enables many analysis

High cost (but… low coverage WGS is an option)
Requires high quality samples
Computation resources



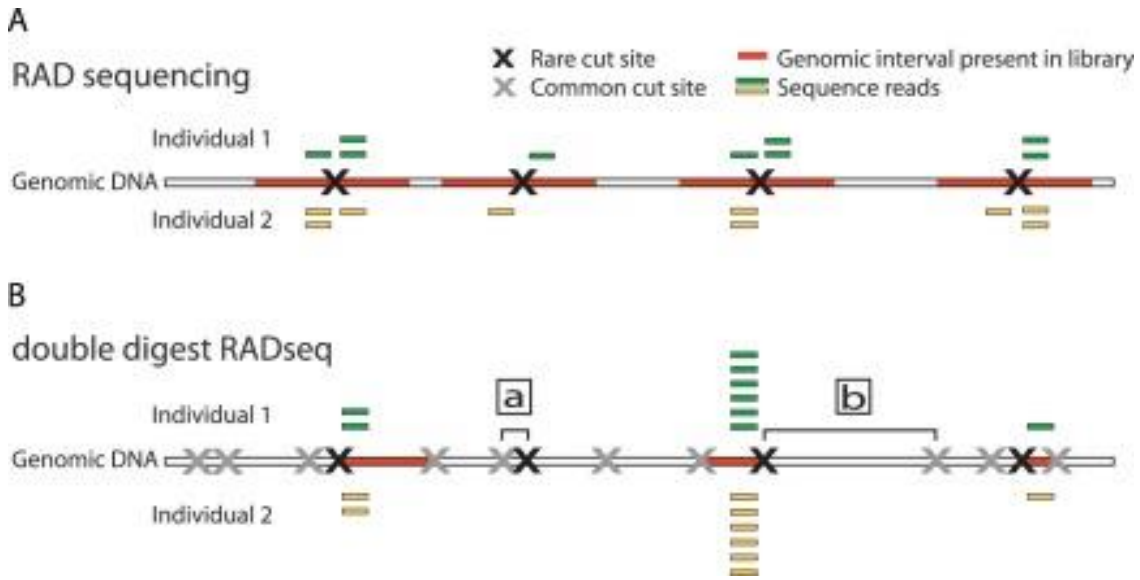
Reduced Representation (RAD-seq, ddRAD)

Sequencing a subset of the genome using restriction enzyms (fragments are size-selected).

Sequencing reduced representation libraries

Cost-effective if many individuals
No reference genome required
Pop structure and diversity analysis

Not ideal for ROH, adaptation
Just a small representation of the genome



Target Capture Approaches

Design of probes (baits) that hybridize with the interested region(s). Then DNA hybridized fragments are captured with magnetic beats and sequenced.

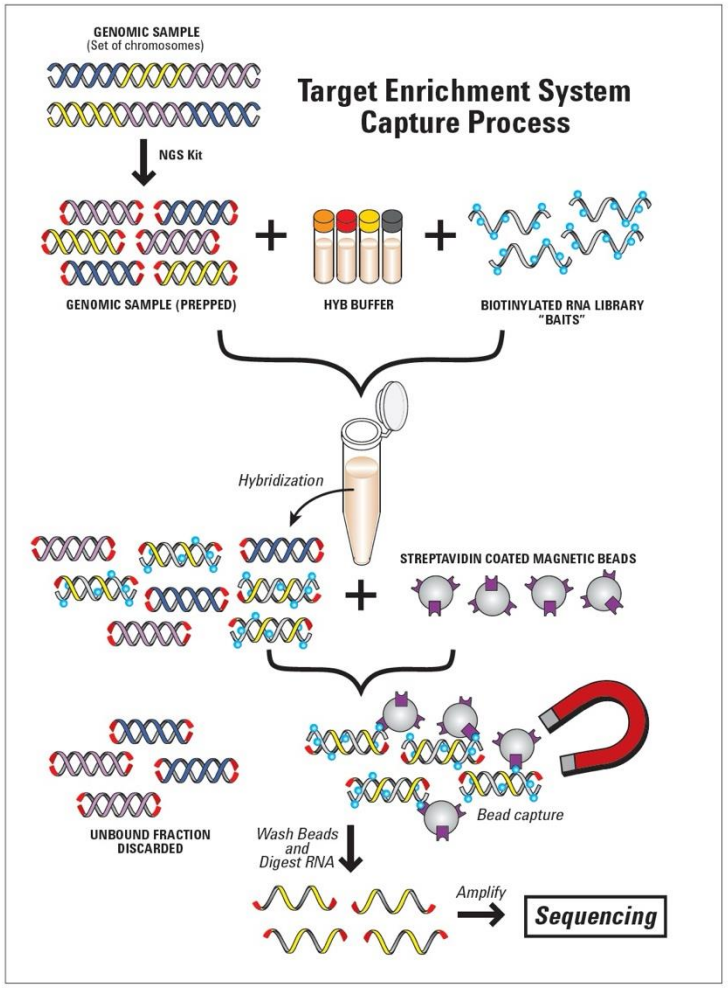


Figure 1 Target enrichment system workflow

- Focus on functional regions (exome), regions or SNPs
- Cheaper than WGS, and still genome-wide
- Works with degraded DNA

- Requires good reference genome, prior information and annotations
- Ascertainment bias in SNP arrays

Assemblies for non-model species

Early Assemblies (~2000s–2015)

De novo short read with Illumina

- Highly fragmented genomes
- Tens to hundreds of thousands of scaffolds
- Low N50 values
- Difficult to resolve repetitive regions, structural variants, annotations…

Current Assemblies (2016-present)

Combination of Long-read sequencing (PacBio, ONT), short-read Illumina, Hi-C or Bionano

- Chromosome-level assemblies, high continuity
- BUSCO completenes > 95%

Global iniciatives supporting the generation of reference genomes from non-model species

EARTH BIOGENOME PROJECT
sequencing life for the future of life

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

ERGA
EUROPEAN REFERENCE GENOME ATLAS

Darwin TREE of LIFE

B10K

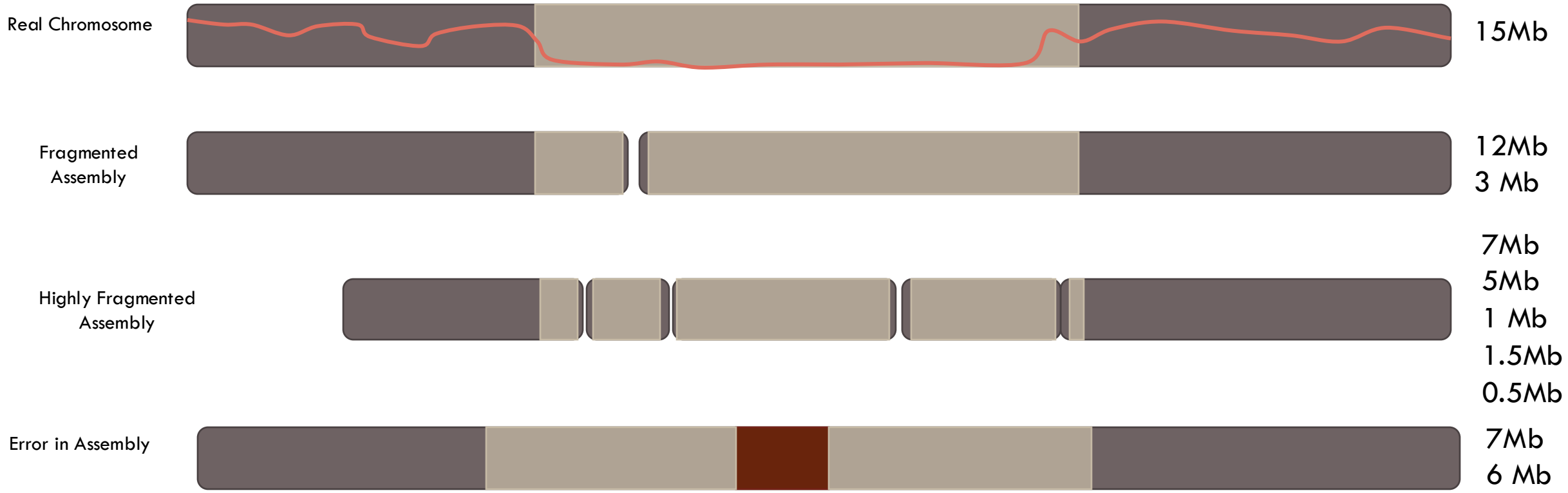
Reference Genome Selection

Ideally:

- High N50/L50, high BUSCO scores, and low number of scaffolds:
  - Chromosome-level assembly is key to provide good estimates of:
    - ROH
    - Structural variation – synteny analysis
    - Selection scans
  - Scaffold-level assembly may be ok:
    - Genome-wide heterozygosity estimates
    - Broad popualtion structure analysis

Reference Genome

Runs of Homozygosity (ROH)



Reference G

- Quality assessment
  - Year
  - Contiguity
  - Annotation

Download | Select columns | 19 Genomes | Rows per page 20 | 1-19 of 19

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
NHGRI_mGorGor1-v2.1_pri	GCA_029281585.3	GCF_029281585.2	Gorilla gorilla gorilla (western lo...	KB3781 (isolate)	NCBI RefSeq	
ASM4964050v1	GCA_049640505.1		Gorilla beringei beringei (easter...	Igicumbi (isolate)		
NHGRI_mGorGor1-v2.0_mat	GCA_028885495.2		Gorilla gorilla gorilla (western lo...	KB3781 (isolate)		
Susie3	GCA_900006655.3		Gorilla gorilla gorilla (western lo...	female (sex)		
GT22 assembly	GCA_965153075.1		Gorilla gorilla gorilla (western lo...			
GT43 assembly	GCA_965153065.1		Gorilla gorilla gorilla (western lo...			
GT21 assembly	GCA_965153055.1		Gorilla gorilla gorilla (western lo...			
Kamilah_GGO_hifiasm-v0.15.2....	GCA_030174185.1		Gorilla gorilla gorilla (western lo...	Kamilah (isolate)		
Kamilah_GGO_hifiasm-v0.15.2....	GCA_030174155.1		Gorilla gorilla gorilla (western lo...	Kamilah (isolate)		
ASM4964058v1	GCA_049640585.1		Gorilla beringei beringei (easter...	Igicumbi (isolate)		
ASM4964060v1	GCA_049640605.1		Gorilla beringei beringei (easter...	Igicumbi (isolate)		
PGDP_GorBer	GCA_963575185.1		Gorilla beringei (eastern gorilla)			
Kamilah_GGO_v0	GCA_008122165.1	GCF_008122165.1	Gorilla gorilla gorilla (western lo...	western lowland go...		
gorGor4	GCA_000151905.3	GCF_000151905.2	Gorilla gorilla gorilla (western lo...			
NHGRI_mGorGor1-v2.0_pat	GCA_028885475.2		Gorilla gorilla gorilla (western lo...	KB3781 (isolate)		
gorGor.msY.makovalab.ver3	GCA_015021865.1		Gorilla gorilla gorilla (western lo...	KB3781 (isolate)		
GorgorY_ver1.0	GCA_001484535.2		Gorilla gorilla gorilla (western lo...	KB3781 (isolate)		
GorY_WUR	GCA_900199665.1		Gorilla gorilla (western gorilla)			
ASM16751v2	GCA_000167515.2		Gorilla gorilla (western gorilla)	male (sex)		

Rows per page 20 | 1-19 of 19

Reference Genome

Genome assembly NHGRI_mGorGor1-v2.1_pri reference

Assembly statistics

	RefSeq
Genome size	3.5 Gb
Total ungapped length	3.5 Gb
Number of chromosomes	25
Number of organelles	1
Number of scaffolds	25
Scaffold N50	150.8 Mb
Scaffold L50	10
Number of contigs	27
Contig N50	150.8 Mb
Contig L50	10
GC percent	40.5
Genome coverage	109x
Assembly level	Chromosome
View sequences	view RefSeq sequences

Genome assembly ASM4964050v1 reference

Assembly statistics

	GenBank
Genome size	3.5 Gb
Total ungapped length	3.5 Gb
Number of scaffolds	350
Scaffold N50	94.8 Mb
Scaffold L50	14
Number of contigs	350
Contig N50	94.8 Mb
Contig L50	14
GC percent	40.5
Genome coverage	80x
Assembly level	Contig
View sequences	view GenBank sequences

Reference Genome

Same Species

Mapped to

- Most accurate mapping and variant calling
- More accurate structural variation analysis and gene annotation

Closely-related Species

Mapped to

- Allows genomic analysis when reference genome is lacking
- Avoids reference bias

Things to consider?

Reference bias (if the reference genome is from one population)

Might not exist or be of very low quality

Is there a genome annotation?

How far away (in terms of divergence) can one go?

There can be rearrengments that bias stuctural analysis or ROH

Genome annotation might be off

Loss of specific variation

How deep should I sequence?

It depends!

Low Coverage

<10x

Limited budget → possiblity to sequence more samples

The only option for degraded samples

Reduced genotype certainty + missigness so some analysis might not be feasible

High Coverage

>20x

When you need accurated individual-level genotypes

ROH, inbreeding or rare variants

Demographic modelling

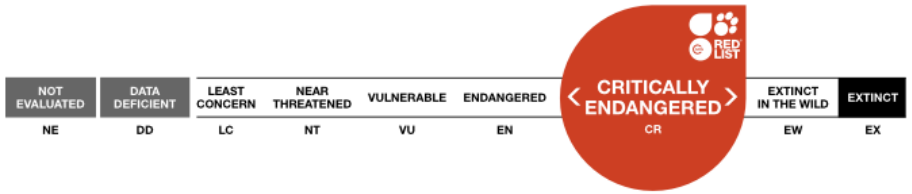
Low vs High Coverage

	Low/medium Coverage	High Coverage
Variant Calling	Genotype Likelihoods ANGSD	Hard Calls: GATK; bcftools
Pop structure	ngsAdmix, PCAngsd, NgsRelate, ngsLD, ngsFST, realSFS	vcftools, plink, ADMIXTURE, eigenstrat
ROH	ROHAN (medium)	bcftools roh, plink
Heterozygosity	ANGSD, realSFS, winSFS	vcftools, plink
Demography	StairwayPlot	PSMC, smc++, MSMC
Genetic Load	Important to normalize	Higher accuracy

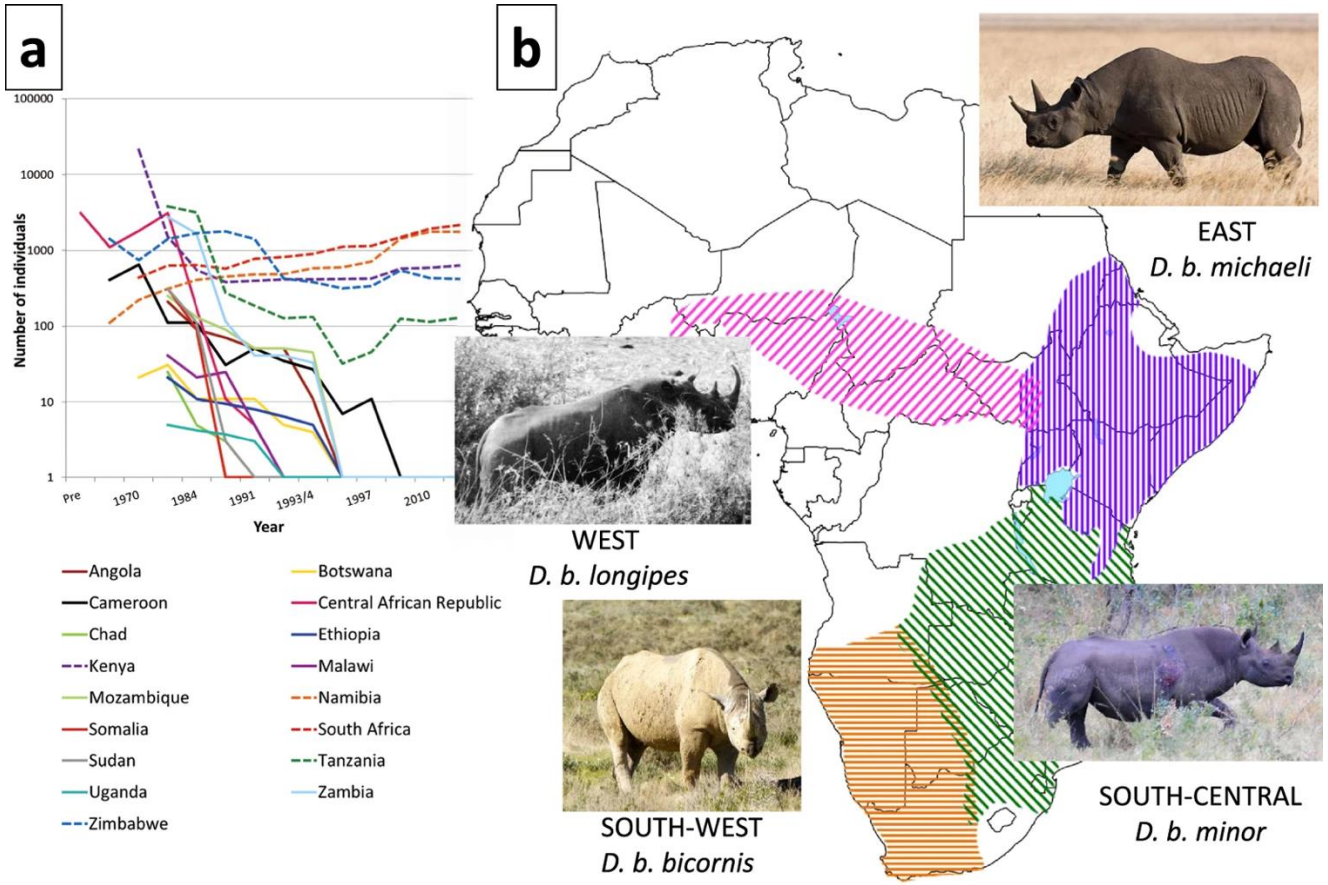
If extremely low coverage (<1x) then it becomes more challenging and only broad populations structure analysis might be possible.

Practical Case – Black Rhinos

Prior to 1960 it was abundant



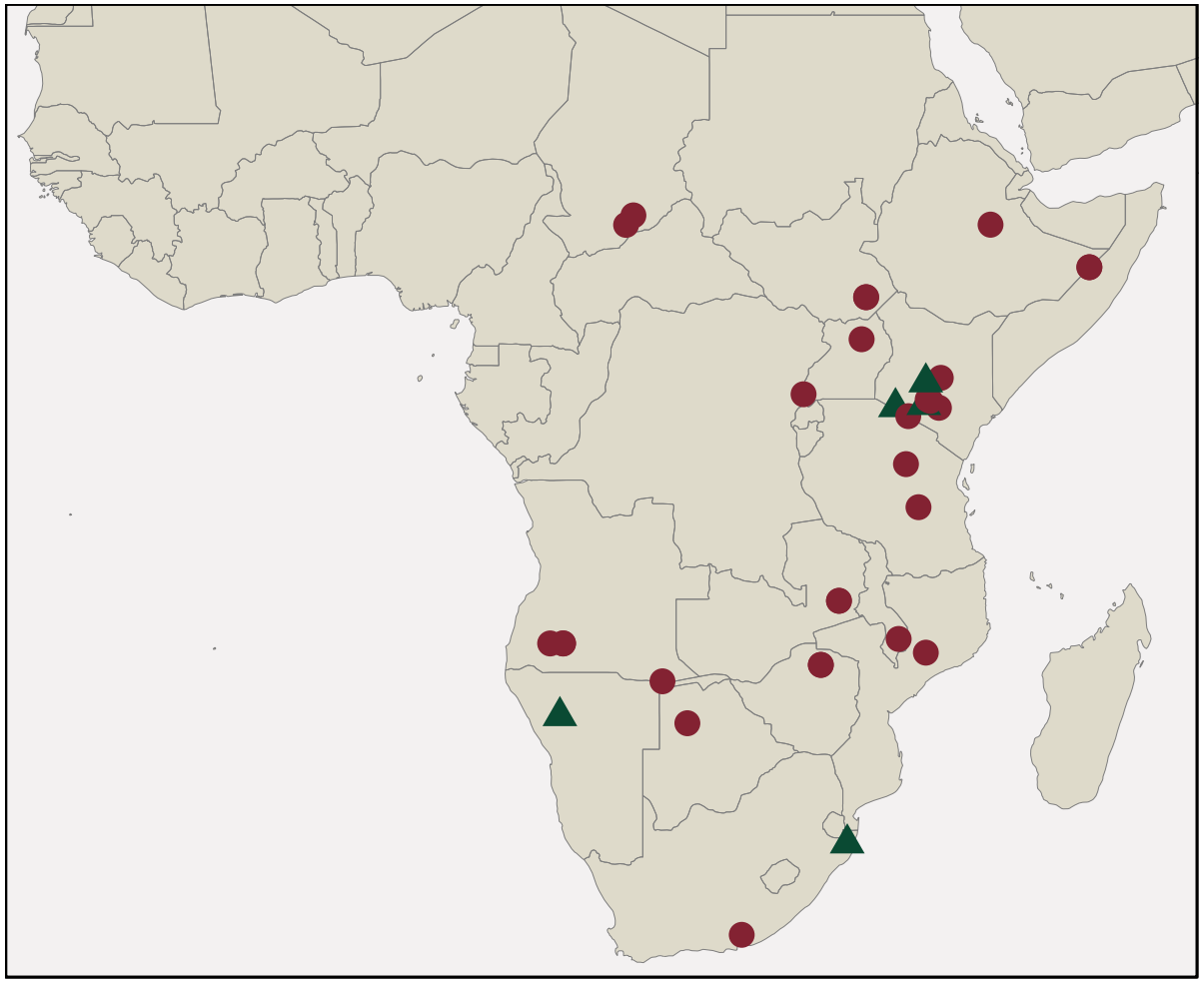
1990s: census of 2,354
2023: census of 6,195



Since late 80s: 4 subspecies - No genetic evidence but managed separately

Moodley et al (2017)

Practical Case – Black Rhinos



wild samples

invasive

non-invasive

museum samples

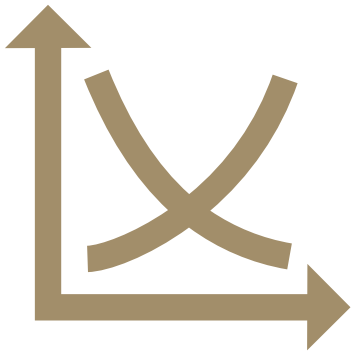
What was the genetic before massive population a[illegible]e?

- Is there a continuum of populations through time? To which historical population do current wild population relate to?
- Has any of the populations gone extinct?
- Do we see loss of genetic diversity?
- How should they be managed? Separetely/together?
- Can we determine the current population structure of wild black rhino populations?

But first… Quality Control!

- Technical aspects to considers according to different sample types
  - FastQC
  - Human contamination
  - Mapping success
  - Damage patterns
  - Relatedness
  - Population structure

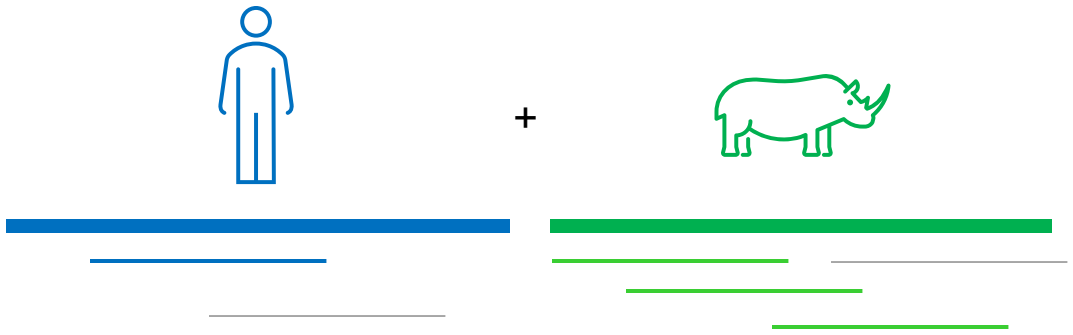
Exercises



Human Contamination

- **Q:** Which samples have higher human contamination?

```
bbsplit.sh -Xmx10g ref_human=GCF_000001405.40_GRCh38.p14_genomic.fna ref_blackrhino=/black_rhino_19Dec2016_S9zPn.
fasta qin=33 in=${dirBB}/FASTQs/${sample}.fastq.gz basename=${dirBB}${sample}_%.fq refstats=${dirBB}${sample}.sta
ts  outu=${dirBB}${sample}_unmapped.fq
```



Proxy of human contamination

% Unambiguous mappings

	Human	Rhino
BW1908.1_museum	0.10	39.52
CD1925.1_museum	0.00	99.96
MA3_blood	0.27	74.27
→ **MA3_fecal**	31.36	10.09
NA1_blood	0.01	99.48
→ **NA1_fecal**	4.92	2.09
NNP1_blood	0.01	98.16
→ **NNP1_fecal**	87.86	0.99

Q: Why doesn't it sum up to 100%?

- **Q:** How would you proceed?

Mapping Success

- **Q:** How many reads map to the Black Rhino reference genome?
- **Q:** What is the amount of PCR duplicates per sample?

Q: Why are we counting only the 1st in pair in the PE reads? Why for some samples we only have one pair?

```
# Count total number of sequenced reads (first in pair)
while read sample; do zcat ${path2FASTQ}/${sample}_1.fastq.gz | wc -l; done < <(head -n 4 list_stats)
while read sample; do zcat ${path2FASTQ}/${sample}.fastq.gz | wc -l; done < <(tail -n 2 list_stats)

#Divide this value by 4 (as each read has 4 lines in a FASTQ file) and safe this information in a excel file
```

```
@SRR31694952.265 265/1
GTTAAATTTAGGGCTAGGGTTAGTGTTCCGGTACGAGTTAGTGTTAGGGTTAGGCCTAGGGTTAG
+
FEFFEDAFEFFFFFFAFFF7F9FEFFFFFFFEFFFFF=FFFFFBBGFEF=EFEFFF?=FEFDDFF
@SRR31694952.266 266/1
GGCTTCAGGTTAGGATTAGGGTTAGGGTACGGGTTAAATTTAGGGCTAGGGTTAGTGTTCCGGTACGAGTTAGTGTTAGGGTTAGGCCTAGGGTTAG
+
FEFFFFFFFFFFFFFF?FFFFF:FFFFFFFFFFDFFFFFFFFFEFFGFFFFFGGF?FFEFFFFFGFFFFFEEF?FFBGFFFFDGFGFDFEFEFGFGF
@SRR31694952.267 267/1
TAACCCTAGGCCTAACCCTAACAGTCACTCGTACCGGAACACTAACCCTAGCCCTAAACTTAACCTGTACCCTAACCCTAATCCTAACCTGAACTCTAAC
+
FD:B>E>EFEE@=?FEFDEECEDEAFDEFCEDEDEFB4DDCDFEFEEFEFFFEFCFE=DFFF<AEEACFFDFFEFAFFEFDEFFEE@EFFF3CFFFFGGF
@SRR31694952.268 268/1
TAACCCTAGGCCTAACCCTAACACTAACTCGTACCGGAACACTAAC
+
=G7G@EBFEG=BB=G=FEFE>FFB@;FABF<G<F9FG;FD@DGF>E
```

Mapping Success

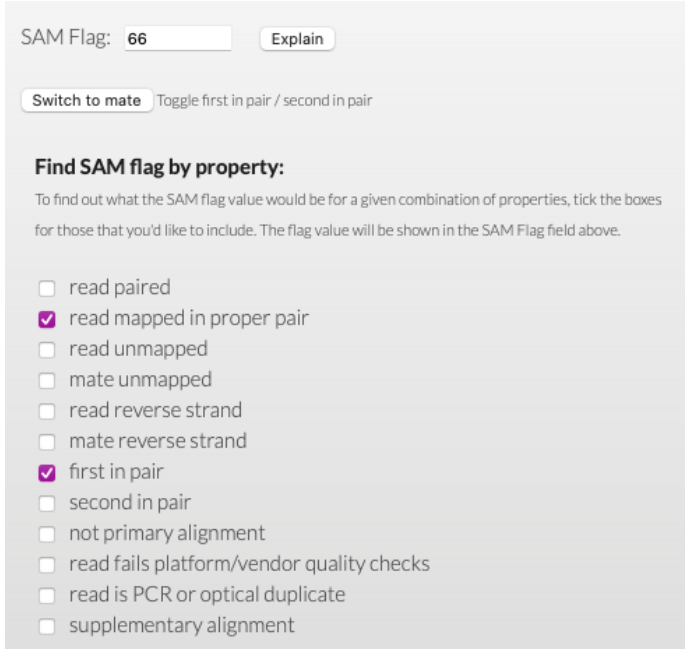
- **Q:** How many reads map to the Black Rhino reference genome
- **Q:** What is the amount of PCR duplicates per sample?

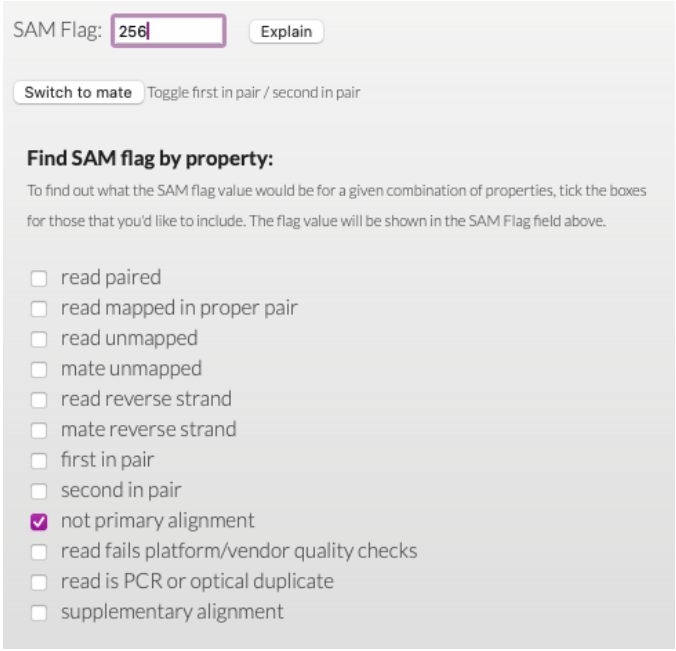
```
# Now, count the reads in the bam file, there are many different ways to do so, we will use samtools and SAM flag
s (https://broadinstitute.github.io/picard/explain-flags.html).
while read sample; do samtools view -f 66 -F 256 -c ${path2BAM}/${sample}.bam ; done < <(head -n 4 list_stats)
while read sample; do samtools view -F 260 -c ${path2BAM}/${sample}.bam ; done < <(tail -n 2 list_stats)


#Now let's count the reads after removing PCR duplicates
while read sample; do samtools view -f 66 -F 1280 -c ${path2BAM}/${sample}_rmDups.bam ; done < <(head -n 4 list_s
tats)
while read sample; do samtools view -F 1284 -c ${path2BAM}/${sample}_rmDups.bam ; done < <(tail -n 2 list_stats)
```

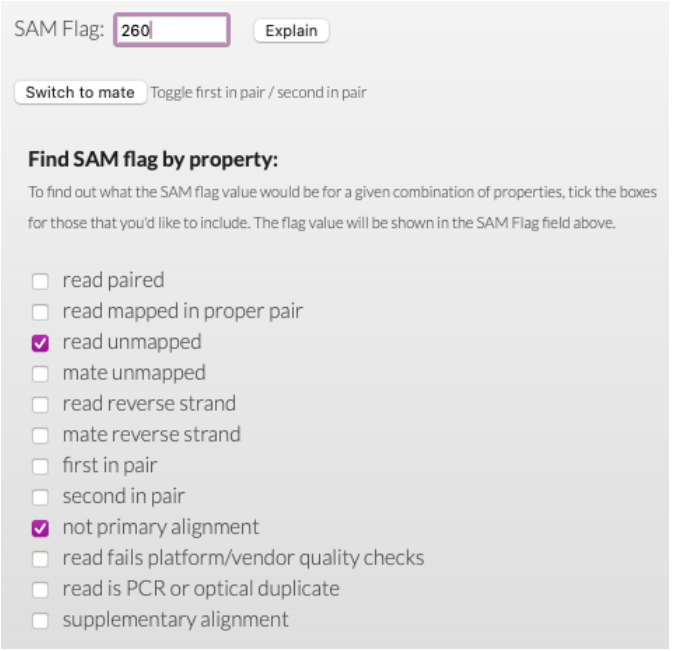
Q: What do the -f and -F SAM flags stand for? What do the number mean?

TIP! Always keep the intermediate files when mapping to count the reads!

Mapping Success







Mapping Success

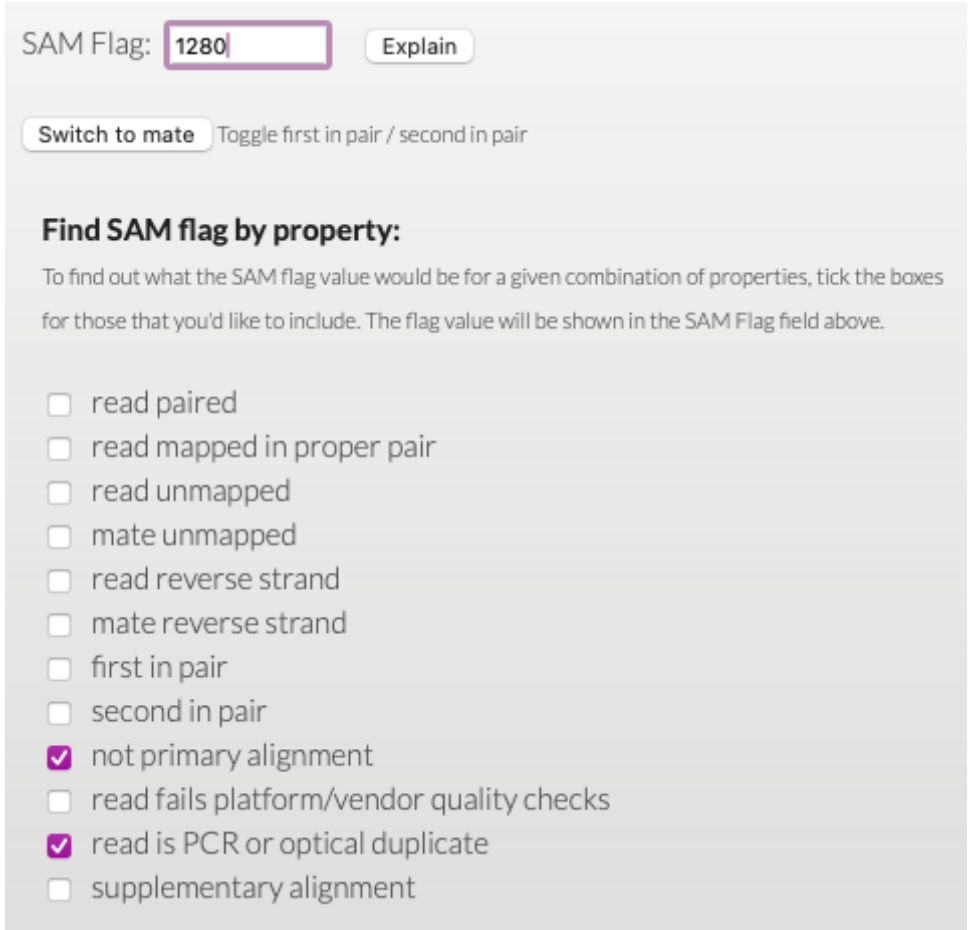
- **Q:** How many reads map to the Black Rhino reference genome
- **Q:** What is the amount of PCR duplicates per sample?

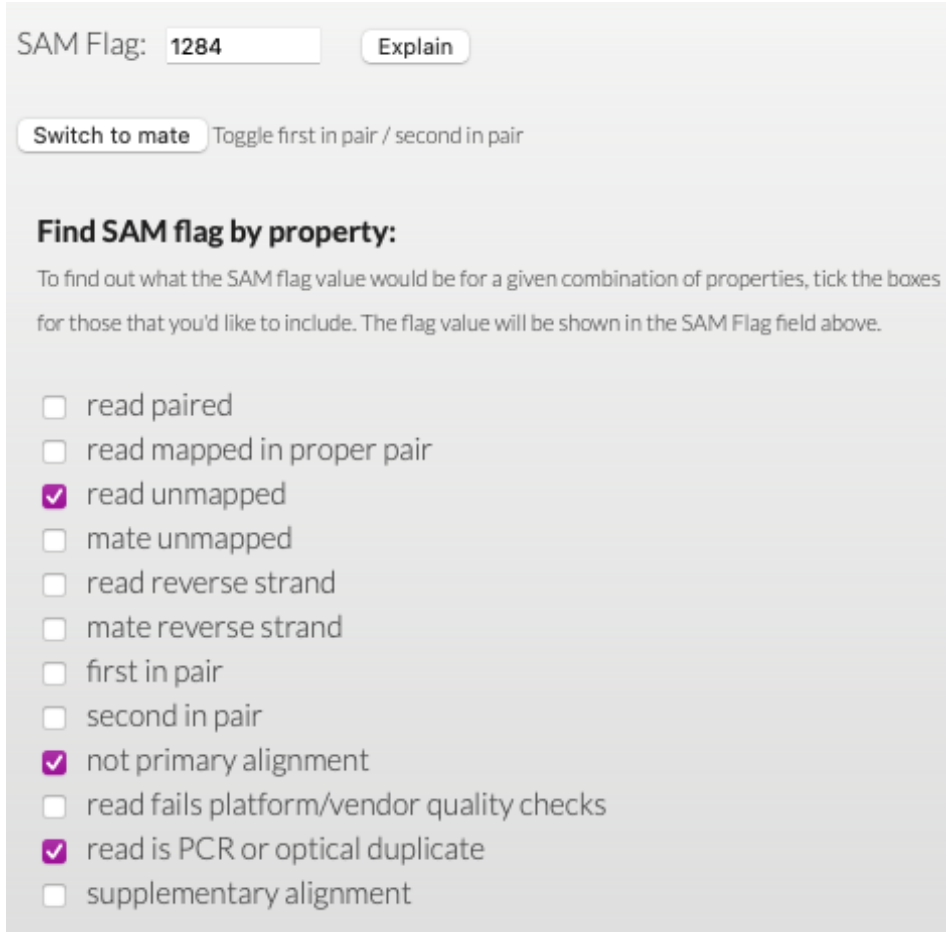
```
# Now, count the reads in the bam file, there are many different ways to do so, we will use samtools and SAM flag
s (https://broadinstitute.github.io/picard/explain-flags.html).
while read sample; do samtools view -f 66 -F 256 -c ${path2BAM}/${sample}.bam ; done < <(head -n 4 list_stats)
while read sample; do samtools view -F 260 -c ${path2BAM}/${sample}.bam ; done < <(tail -n 2 list_stats)


#Now let's count the reads after removing PCR duplicates
while read sample; do samtools view -f 66 -F 1280 -c ${path2BAM}/${sample}_rmDups.bam ; done < <(head -n 4 list_s
tats)
while read sample; do samtools view -F 1284 -c ${path2BAM}/${sample}_rmDups.bam ; done < <(tail -n 2 list_stats)
```

Q: What do the -f and -F SAM flags stand for? What do the number mean?

TIP! Always keep the intermediate files when mapping to count the reads!

Mapping Success



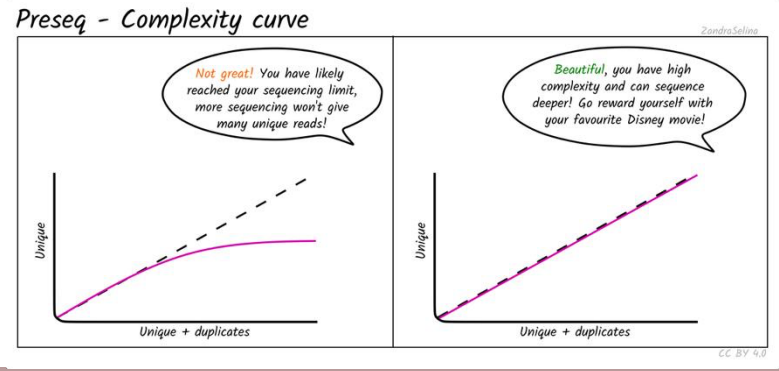


Mapping Success

- **Q:** How many reads map to the Black Rhino reference genome
- **Q:** What is the amount of PCR duplicates per sample?

ID	Type	Raw Reads	Mapped Reads	**% Mapped Reads**	Unique Reads	**% Unique Reads**	Duplicates	**Duplicate rate**
MA3	Blood	100000	88819	88,82%	87741	87,74%	1078	1,08%
NA1	Blood	100000	94952	94,95%	86855	86,86%	8097	8,10%
NNP1	Blood	100000	97563	97,56%	96062	96,06%	1501	1,50%
ZA1	Blood	100000	76719	76,72%	74975	74,98%	1744	1,74%
BW1908.1	Bone	100000	11539	11,54%	10669	10,67%	870	0,87%
CD1925.1	Bone	100000	98913	98,91%	8205	8,21%	90708	90,71%

With the BAM files (including PCR duplicates), you could now run **Preseq** software to estimate library complexity and if it makes sense or not to sequenced deeper!



Relatedness

```
# This will take around 6 min
angsd -bam  $listBAM_wildRhinos -ref $ref -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -rmTrans 1 -C 50 -ski
pTriallelic 1 -GL 2 -minMapQ 30 -minQ 20 -doGlf 3 -doMajorMinor 1 -doMaf 1 -SNP_pval 1e-6 -r ScS9zPn_17: -out ${d
irREL}/BR_wild_related
```

Q: Which files result from running the previous command?

```
ls BR_wild_related*
```

```
BR_wild_related.arg BR_wild_related.glf.gz  BR_wild_related.glf.pos.gz  BR_wild_related.mafs.gz
```

With this we have the files needed to calculated the relatedness:

```
# Obtain the frequencies of each SNP
zcat ${dirREL}/BR_wild_related.mafs.gz | cut -f6 | sed 1d > ${dirREL}/BR_wild_related.freq

# Calculate relatedness with ngsRelate
ngsRelate -g ${dirREL}BR_wild_related.glf.gz -n 17 -f ${dirREL}BR_wild_related.freq -O ${dirREL}BR_wild_related.m
l
```

Relatedness

Q: Can you make sense of the output file?
Check this link: https://github.com/ANGSD/NgsRelate

a	b	nSites	J9	J8	J7	J6	J5	J4	J3	J2
0	1	99927	0.384487	0.360978	0.001416	0.178610	0.071681	0.000617	0.002172	0.000034

The first two columns contain indices of the two individuals used for the analysis. The third column is the number of genomic sites considered. The following nine columns are the maximum likelihood (ML) estimates of the nine jacquard coefficients, where K0==J9; K1==J8; K2==J7 in absence of inbreeding. Based on these Jacquard coefficients, NgsRelate calculates 11 summary statistics:

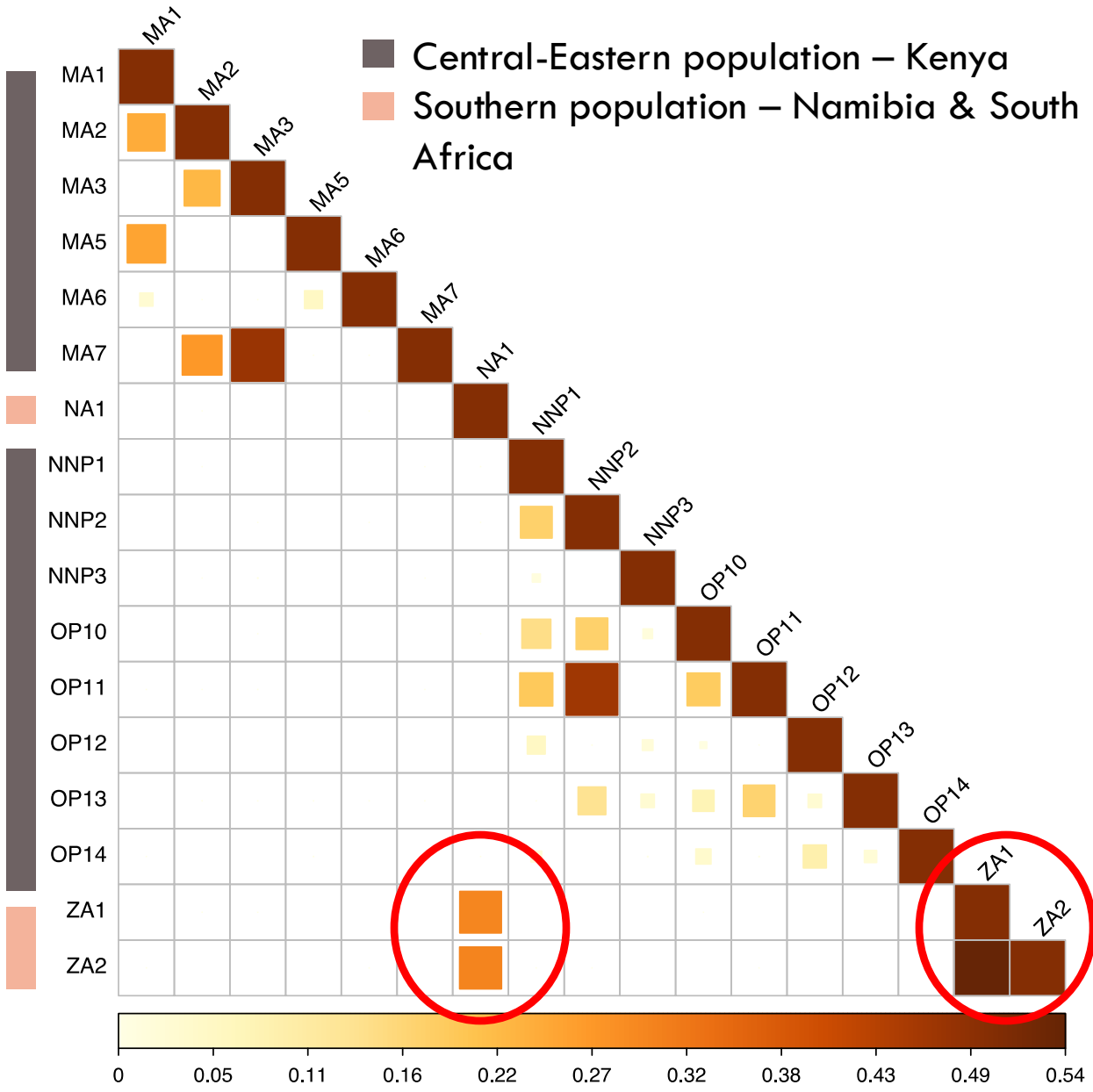
13. rab is the pairwise relatedness `(J1+J7+0.75*(J3+J5)+.5*J8)` Hedrick et al
14. Fa is the inbreeding coefficient of individual a `J1+J2+J3+J4` Jacquard
15. Fb is the inbreeding coefficient of individual b `J1+J2+J5+J6` Jacquard
16. theta is the coefficient of kinship `J1 + 0.5*(J3+J5+J7) + 0.25*J8)` Jacquard
17. inbred_relatedness_1_2 `J1+0.5*J3` Ackerman et al
18. inbred_relatedness_2_1 `J1+0.5*J5` Ackerman et al
19. fraternity `J2+J7` Ackerman et al
20. identity `J1` Ackerman et al
21. zygosity `J1+J2+J7` Ackerman et al
22. Two-out-of-three IBD `J1+J2+J3+J5+J7+0.5*(J4+J6+J8)` Miklos csuros
23. Inbreeding difference `0.5*(J4-J6)` Miklos csuros
24. the log-likelihood of the ML estimate.
25. number of EM iterations. If a `-1` is displayed. A boundary estimate had a higher likelihood.
26. If differs from `-1`, a boundary estimate had a higher likelihood. Reported loglikelihood should be highly similar to the corresponding value reported in `loglh`
27. fraction of sites used for the ML estimate

Relatedness

Q: Do you see any concerning results? How do you interpret the results?

The **kinship coefficient** is the probability that a pair of randomly sampled homologous alleles are identical by descent

Relationship	Kinship Coefficient
Individual-Self / Twins	**0.5**
Full Siblings Parent-Offspring	**0.25** (0.1875-0.375)
Grandparent - Grandoffspring Avuncular (aunt-niece)	**0.125** (0.09375-0.1875)
First cousins	**0.0625** (0.046875-0.09375)



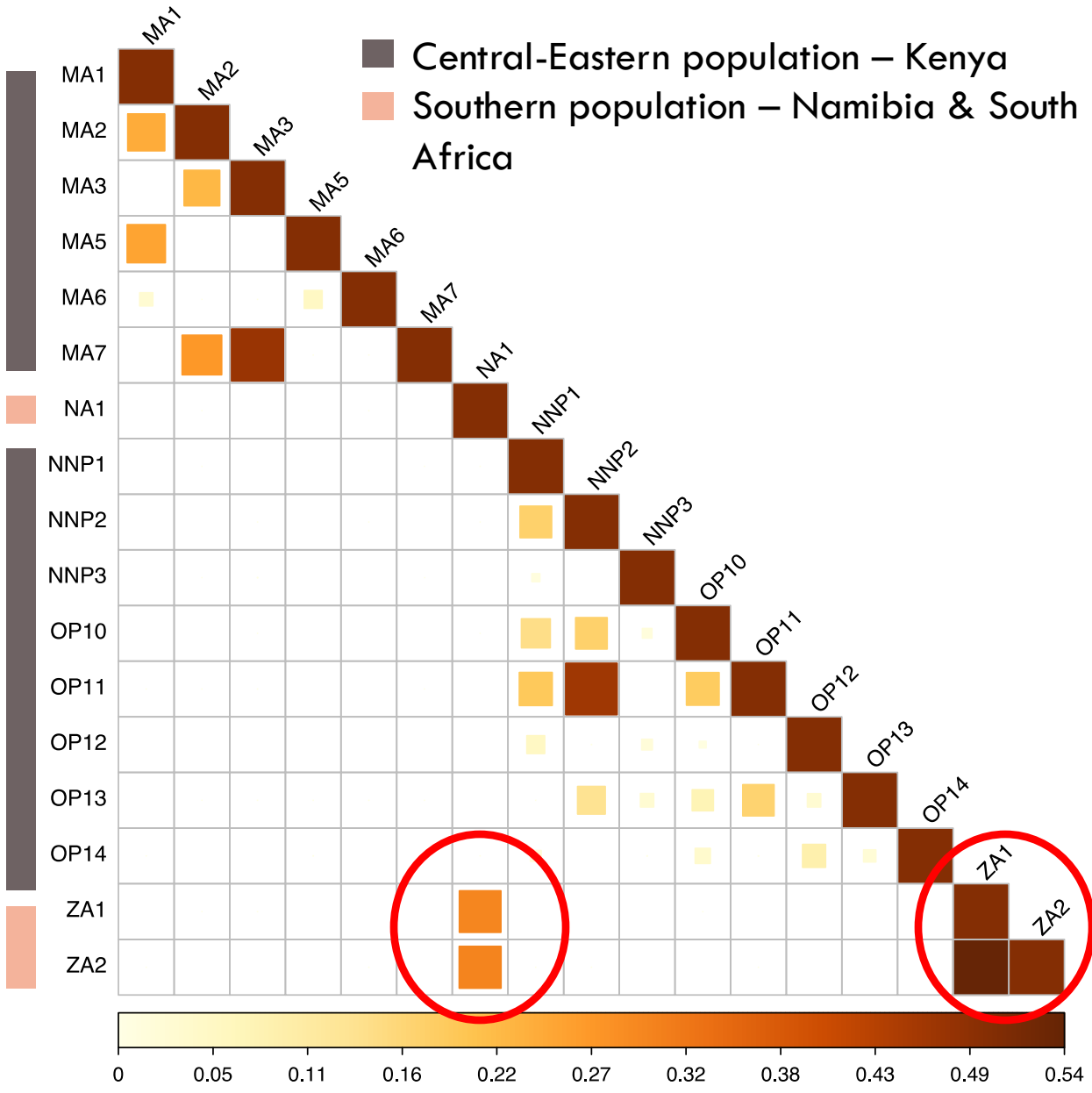
Relatedness

Q: Do you see any concerning results? How do you interpret the results?

- Population structure affects relatedness estimates.

Important to know the populations structure before

Central-Eastern population – Kenya	Southern population – Namibia & South Africa



Population structure

```
# create a directory to run this analysis:
dirPCA=${directory_day4}/E4_PCA/
mkdir -p $dirPCA

ref=${data}/RefGenome/black_rhino_19Dec2016_S9zPn.fasta
listBAM_wildRhinos=${directory_day4}/bams_wildBR.txt

# Run ANGSD to obtain GL.
angsd -bam $listBAM_wildRhinos -ref $ref -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -rmTrans 1 -C 50 -minI
nd 15 -skipTriallelic 1 -GL 2 -minMapQ 30 -minQ 20 -minMaf 0.05 -doGlf 2 -doMajorMinor 1 -doMaf 1 -SNP_pval 1e-6
-r ScS9zPn_17: -nThreads 4 -out ${dirPCA}BR_wild
```

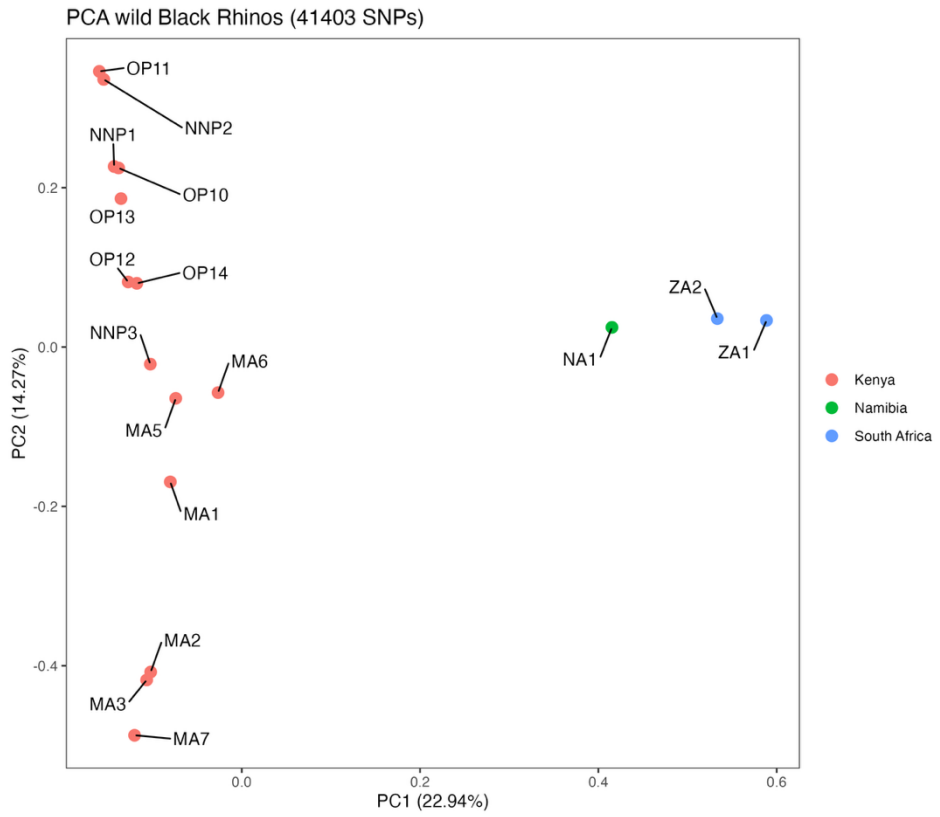
Beagle file format

marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1	Ind2	Ind2	Ind2	Ind3	Ind3	Ind3	Ind
ScS9zPn_17_142	1	0	0.998050			0.001950			0.000000	0.999939			0.000061			0.000000	
ScS9zPn_17_507	0	1	0.992246			0.007754			0.000000	0.999755			0.000245			0.000000	
ScS9zPn_17_528	2	3	0.996108			0.003892			0.000000	0.999878			0.000122			0.000000	
ScS9zPn_17_1145	3	0	0.941171			0.058829			0.000000	0.998043			0.001957			0.000000	
ScS9zPn_17_2495	1	0	1.000000			0.000000			0.000000	0.999756			0.000244			0.000000	
ScS9zPn_17_3254	3	0	0.999996			0.000004			0.000000	0.999939			0.000061			0.000000	
ScS9zPn_17_4419	1	0	0.000000			1.000000			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_5238	1	0	0.999512			0.000488			0.000000	0.998050			0.001950			0.000000	
ScS9zPn_17_5549	1	0	0.999992			0.000008			0.000000	0.999878			0.000122			0.000000	
ScS9zPn_17_5977	2	1	0.998050			0.001950			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_6219	2	3	0.999024			0.000976			0.000000	0.999878			0.000122			0.000000	
ScS9zPn_17_6998	0	3	0.999511			0.000489			0.000000	0.999512			0.000488			0.000000	
ScS9zPn_17_7085	0	1	0.000000			1.000000			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_7373	2	1	0.000000			1.000000			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_8030	1	2	0.000000			1.000000			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_8150	1	2	0.999939			0.000061			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_8319	2	3	0.000000			1.000000			0.000000	0.000000			1.000000			0.000000	
ScS9zPn_17_8336	1	2	0.999878			0.000122			0.000000	0.999992			0.000008			0.000000	

Population structure

```
pcangsd -b ${dirPCA}BR_wild.beagle.gz -o ${dirPCA}BR_wild
```

You get a 17x17 matrix → transform to eigenvectors and eigenvalues



Q: Would you say there is population structure?
Q: How many populations would you describe there are?

Two-population structure

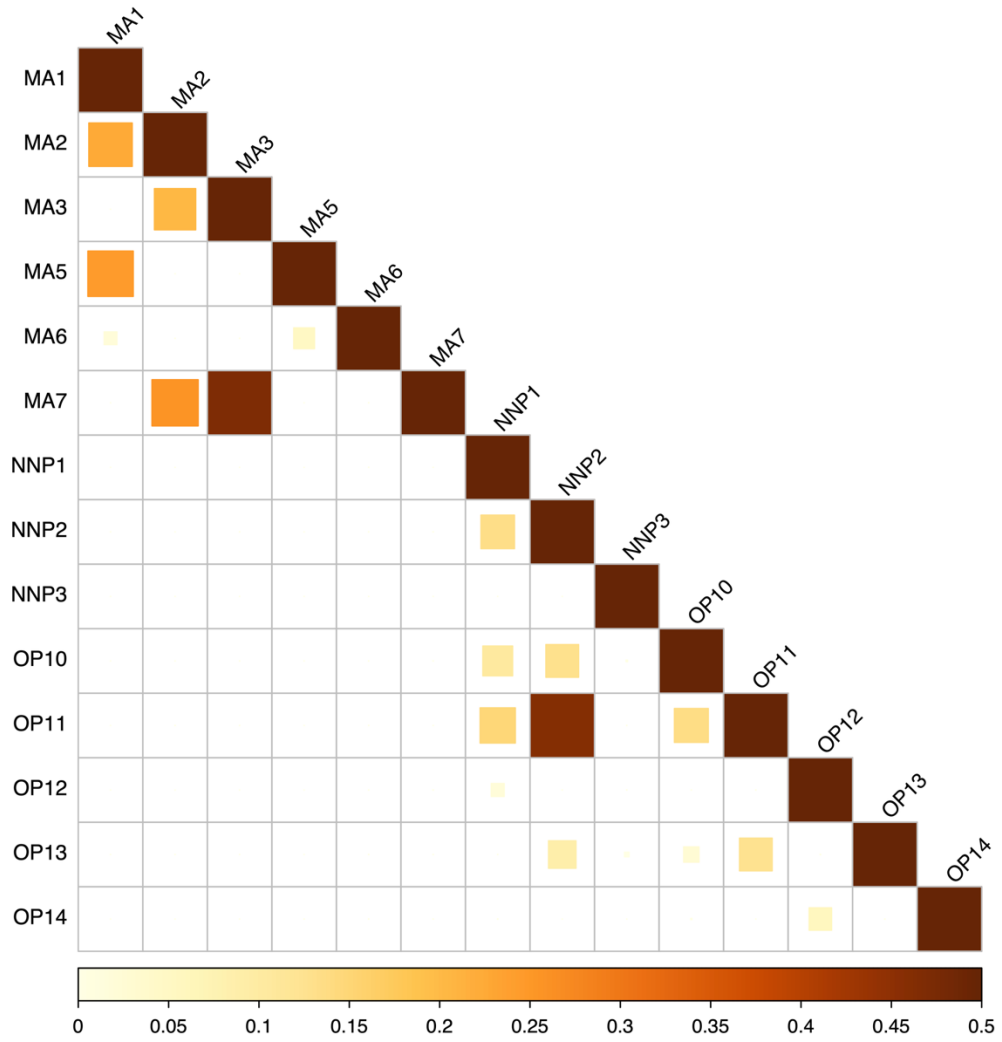
Central-Eastern population – Kenya
Southern population – Namibia & South Africa

Relatedness per group

Q: How many highly related samples are there in the northern group? And in the southern?

Central-Eastern population – Kenya	6: 1st degree or higher related pairs 6: 2nd degree related pairs
outhern population – Namibia & South Africa	No first or 2nd degree related pairs

Relationship	Kinship Coefficient
Individual-Self / Twins	**0.5**
Full Siblings Parent-Offspring	**0.25** (0.1875-0.375)
Grandparent - Grandoffspring Avuncular (aunt-niece)	**0.125** (0.09375-0.1875)
First cousins	**0.0625** (0.046875-0.09375)

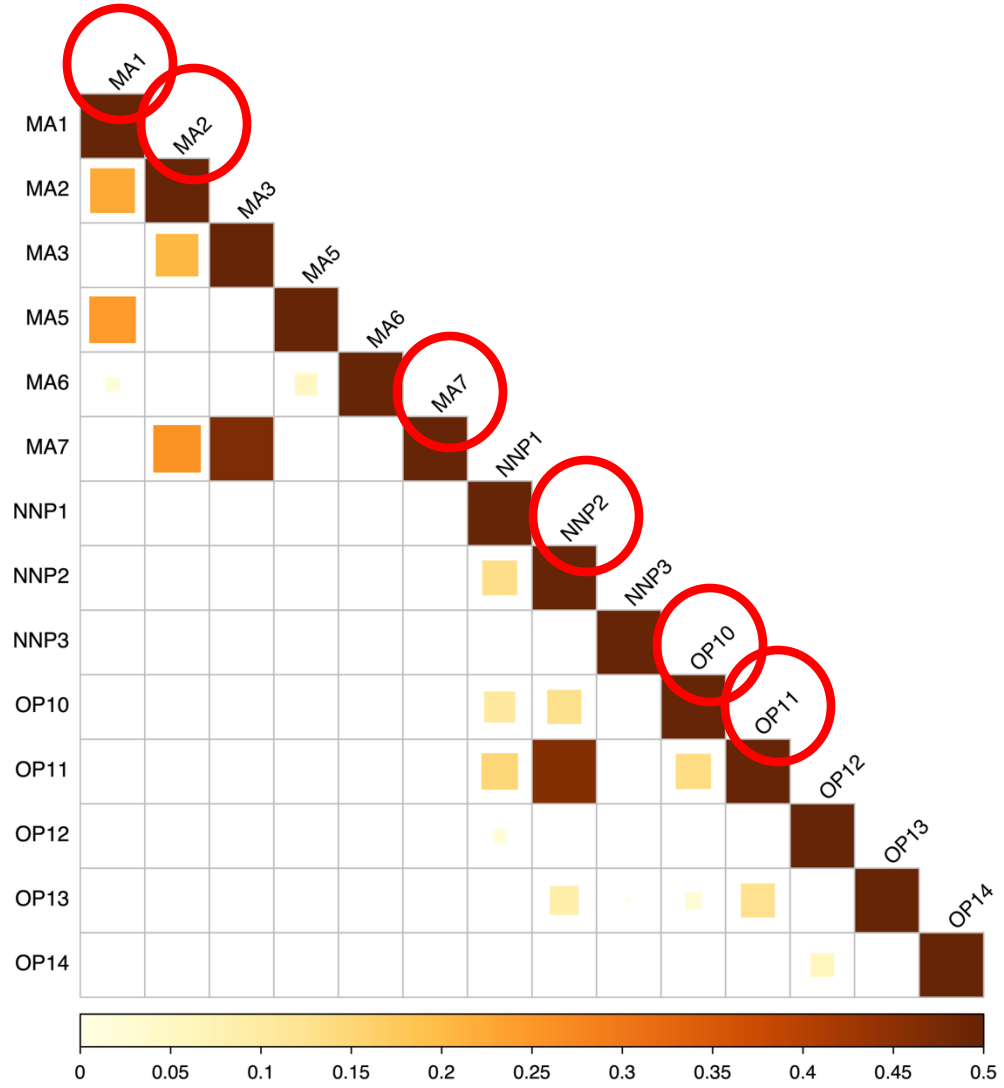


Relatedness per group

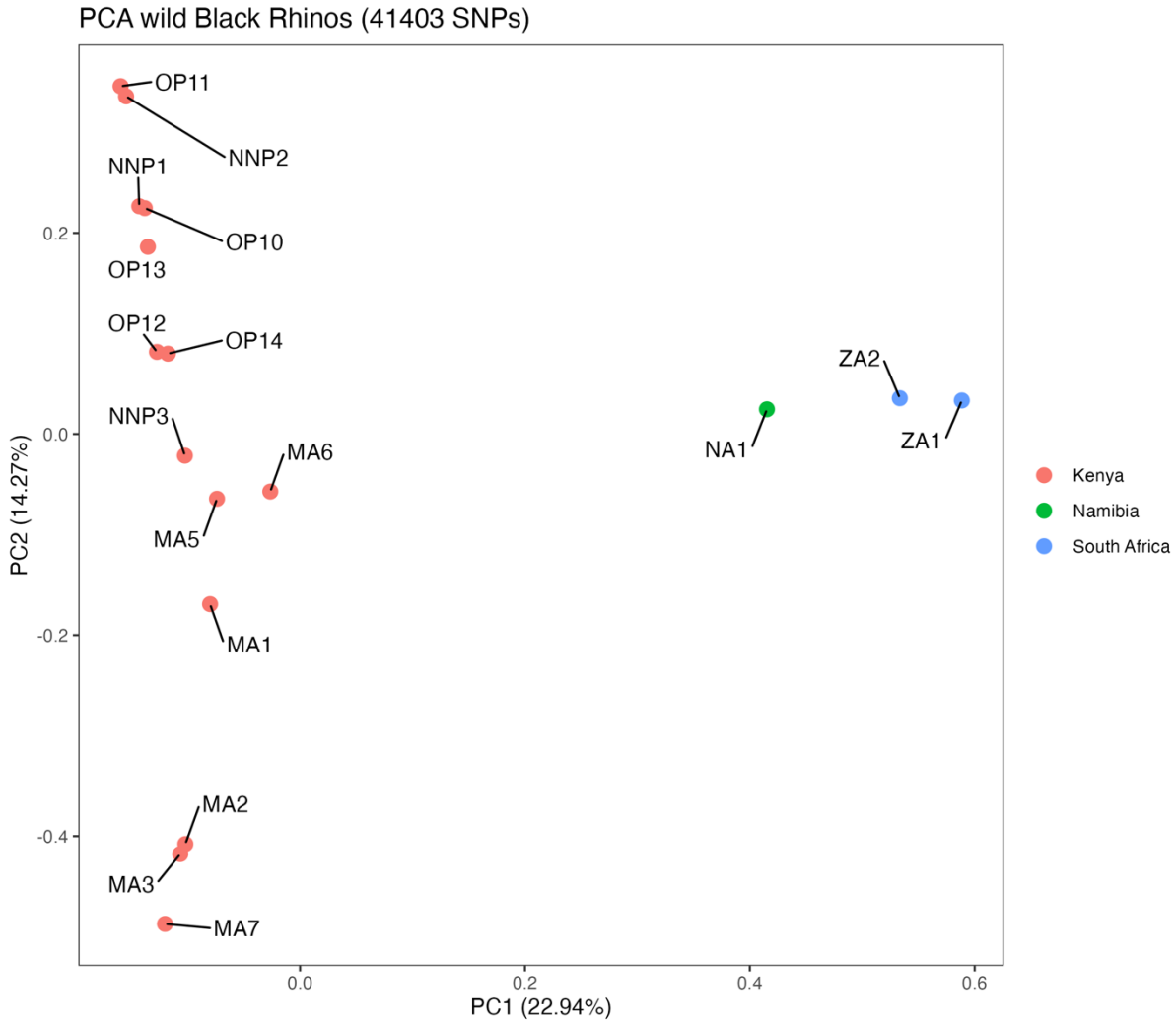
Q: Which samples are you going to exclude for any further analysis because are related?

MA1, MA2, MA7, NNP2, OP10, OP11

Relationship	Kinship Coefficient
Individual-Self / Twins	**0.5**
Full Siblings Parent-Offspring	**0.25** (0.1875-0.375)
Grandparent - Grandoffspring Avuncular (aunt-niece)	**0.125** (0.09375-0.1875)
First cousins	**0.0625** (0.046875-0.09375)



Population structure



Q: Are they representative of the historical distribution?

Our sampling is VERY LIMITED:

- Difficulties to get wild samples
- Already extinct wild populations

We need a genetic baseline

What can museum samples bring to this field?

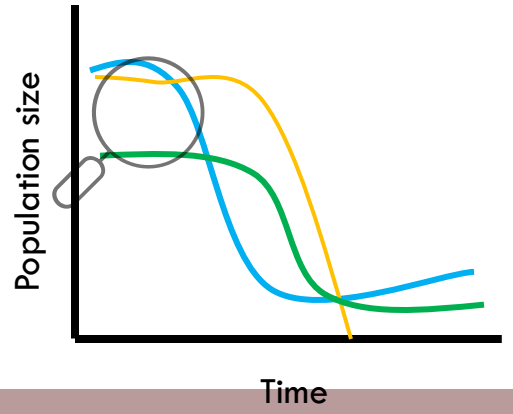
Sampling wild (endangered) populations

Getting good quality samples for the wild is challenging.

Historical samples

- Temporal sampling
  - Obtain a genetic baseline
  - Has any (sub)population gone extinct?
  - Has it been a loss of genetic diversity through time?

Before anthropogenic destruction of habitat and population bottleneck → Good representation of historical population diversity.



1. Extinction of a population
2. Loss of genetic diversity
3. Increase of inbreeding
4. Accumulation of deleterious variation

Quantifying Temporal Genomic Erosion

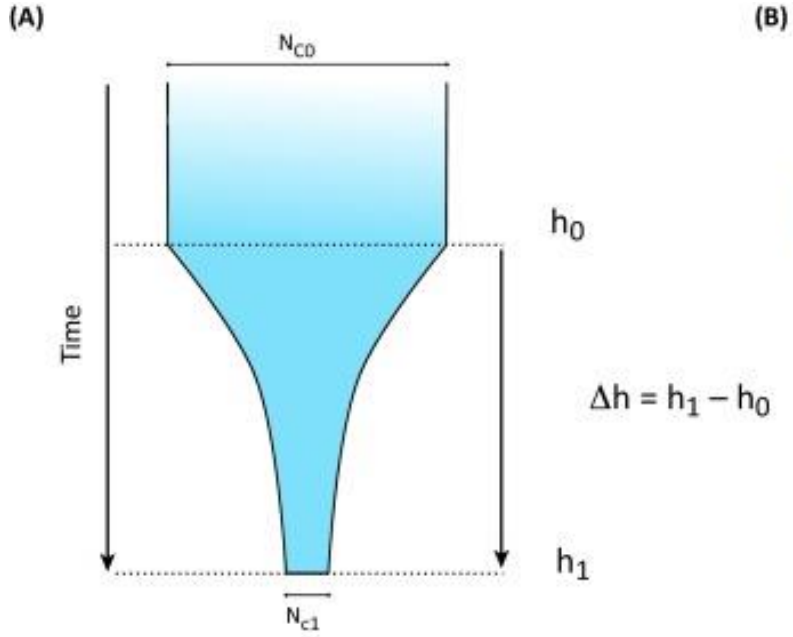
- Reference genome
  - Ideally same species
  - Closely related species

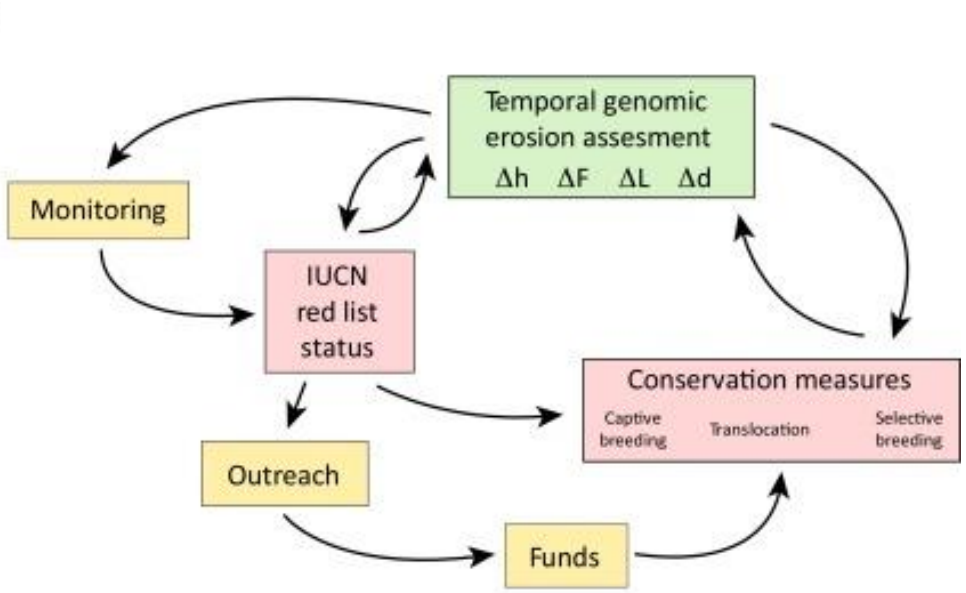
Calculation of temporal changes

Heterozygosity

Inbreeding

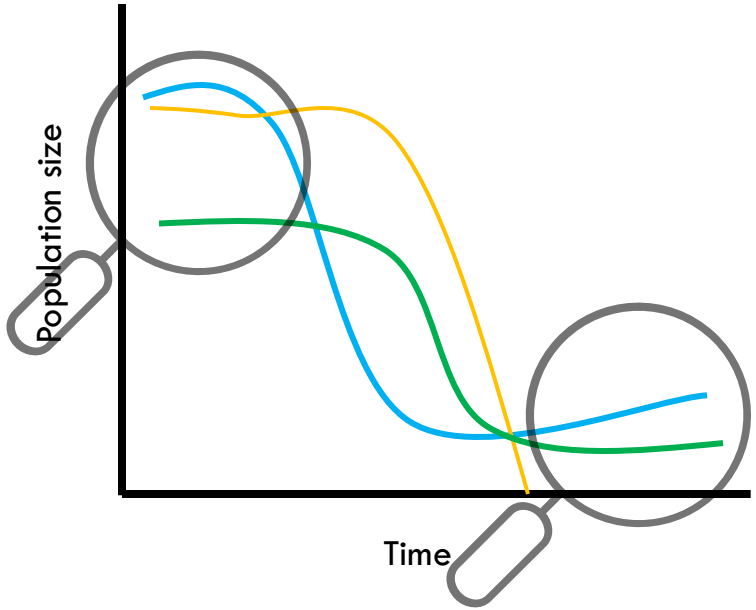
Genetic Load

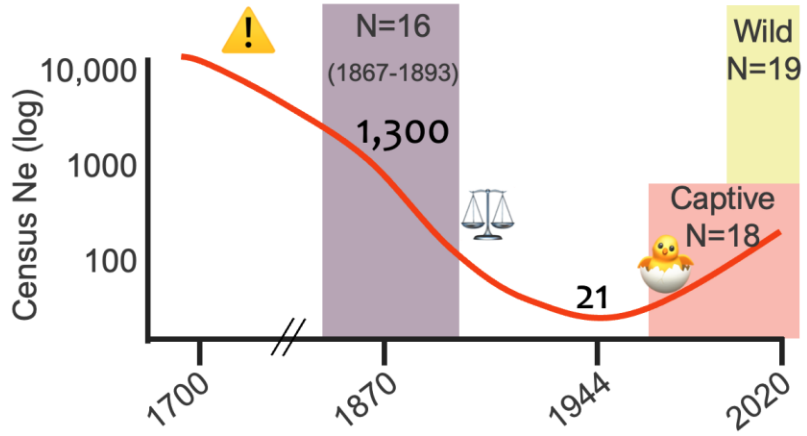


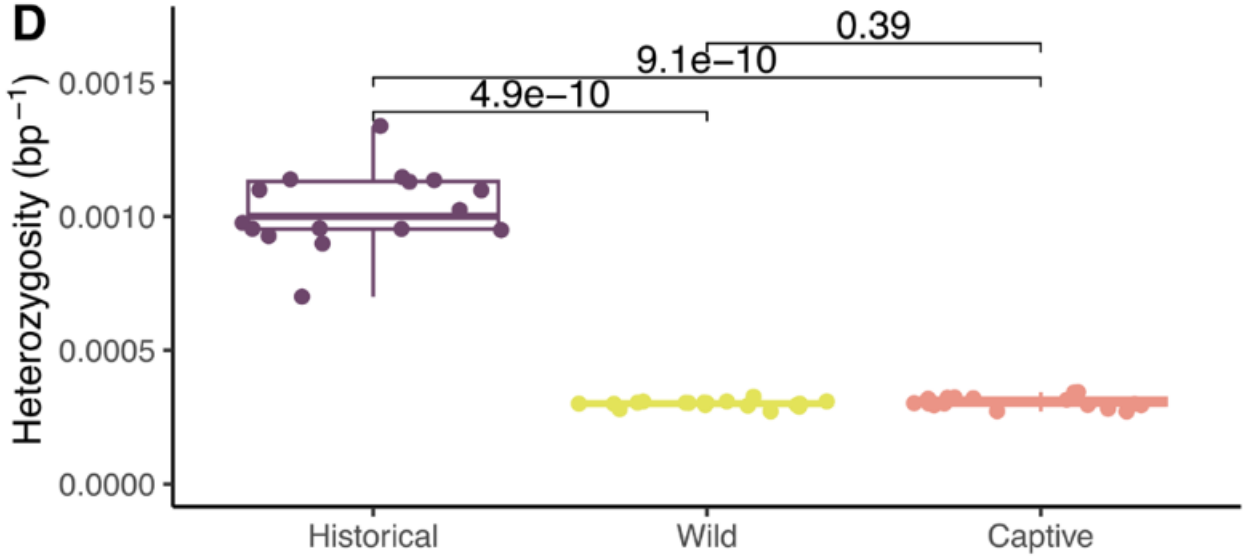


Díez-del-Molino et al (2018)

Loss of Genetic Diversity







Compute heterozygosity (h) from whole-genomes from historical and modern samples

Het (bp^{-1})=Heterozygous positions/genome size (bp)

How:

- ANGSD + realSFS
- Counting "0/1" in VCF

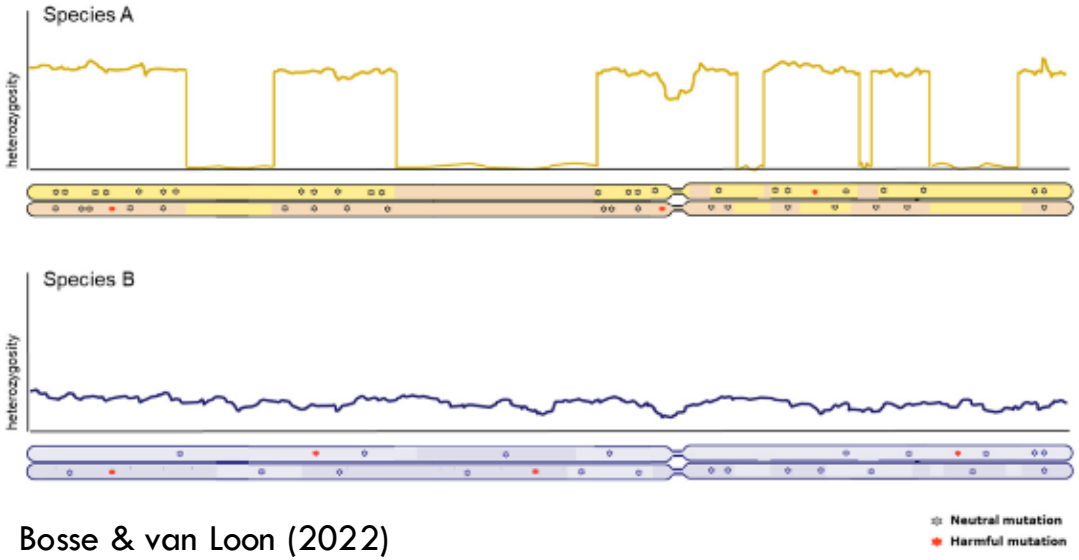
Challenges:

- Uneven coverage
- Errors and postmortem damage

Fontsere et al. (2025)

Inbreeding and ROHs

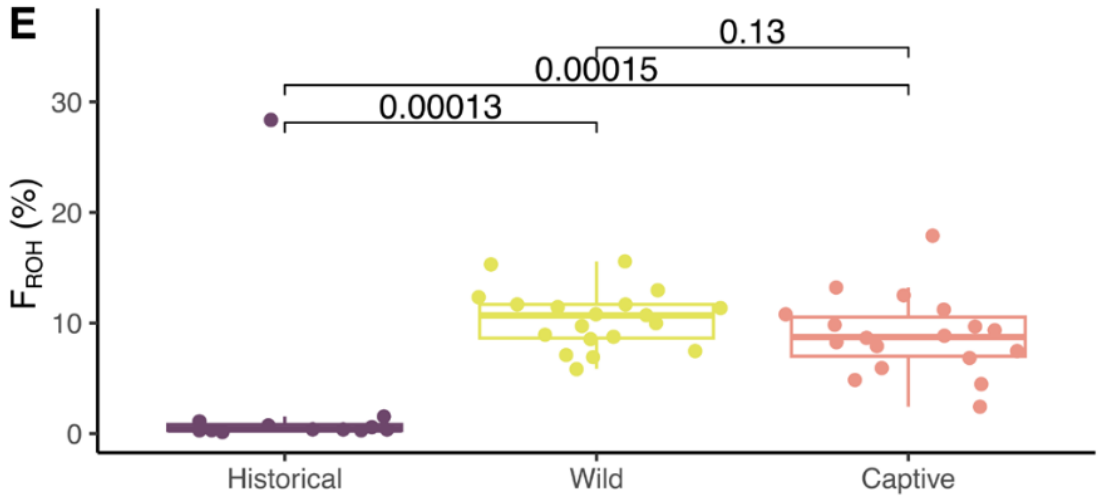
Runs of homozygosity (ROH): chromosomal stretches of a diploid genome that are in a homozygous state.



Bosse & van Loon (2022)

Inbreeding and ROHs

Runs of homozygosity (ROH): chromosomal stretches of a diploid genome that are in a homozygous state.



How:

- Plink
- Bcftools roh
- Rohan

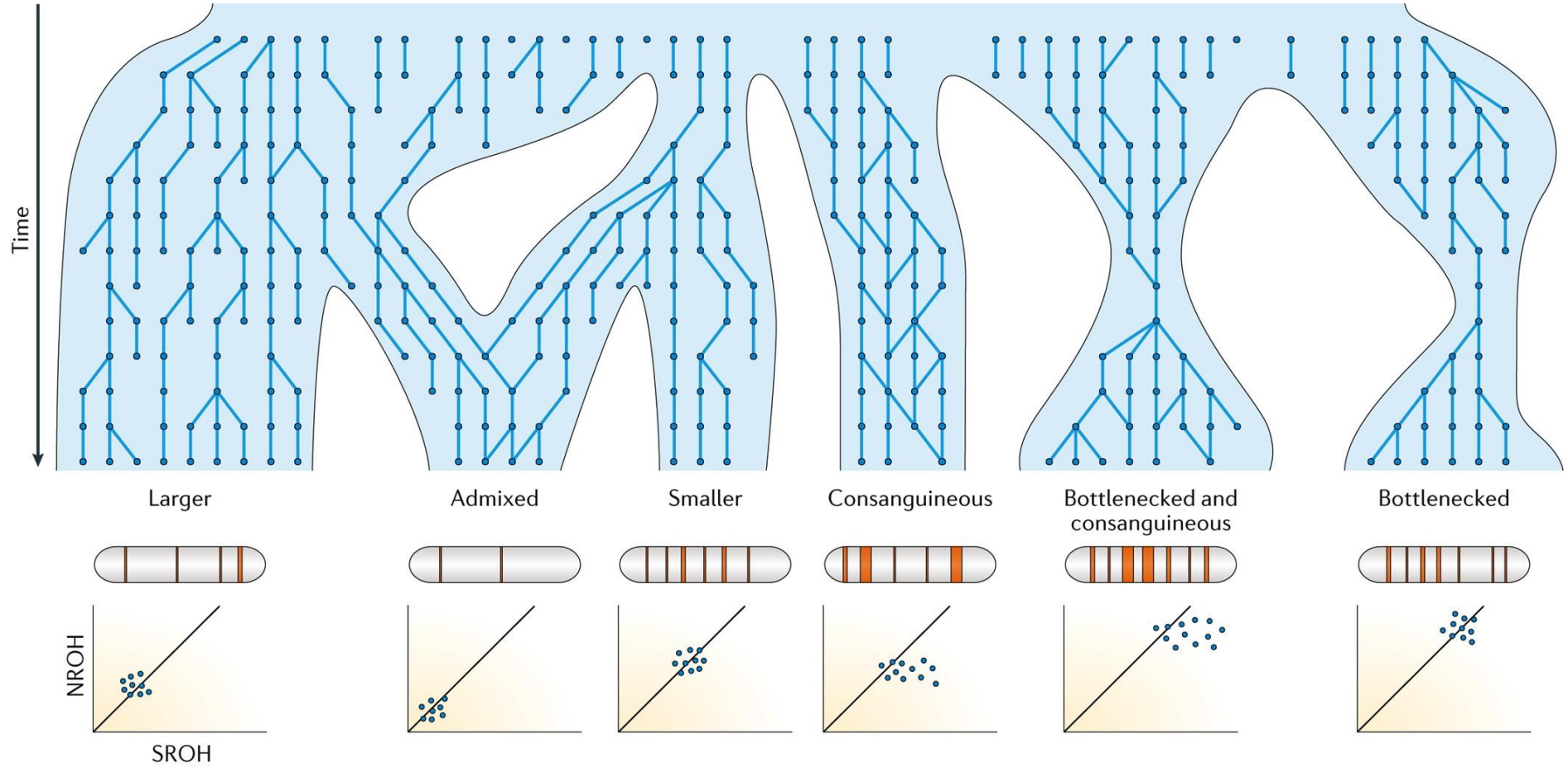
Challenges:

- There is no concensus
- Uneven coverage
- Errors and postmortem damage
- Fragmented genome
- Low diversity and ROHs

Inbreeding Coefficient (F_{ROH})

Fontsere et al. (2025)

Demographic history from ROH



Ceballos *et al.* (2018)

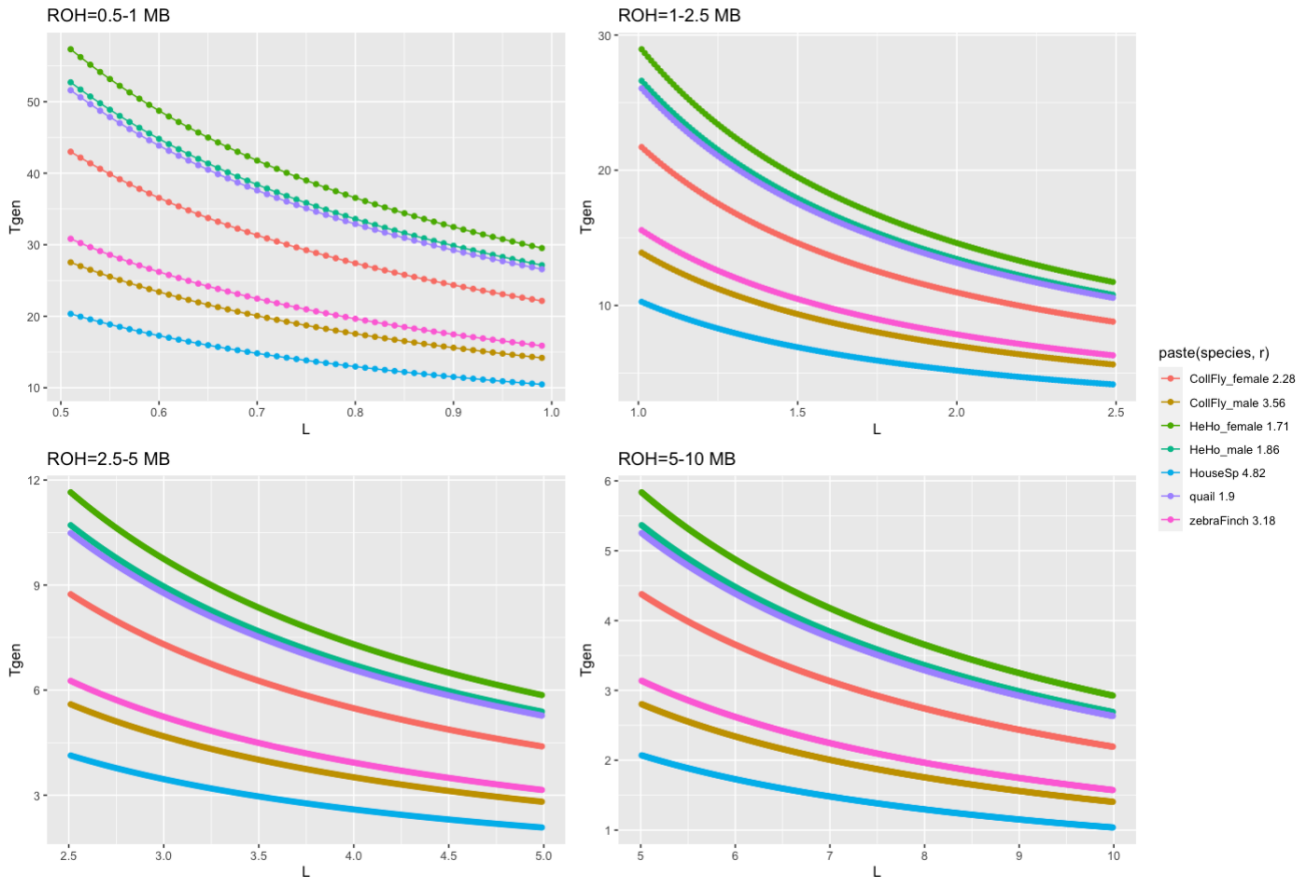
Timing inbreeding

Length(cM)=100/(2*generations)

Size of ROH link to generations ago

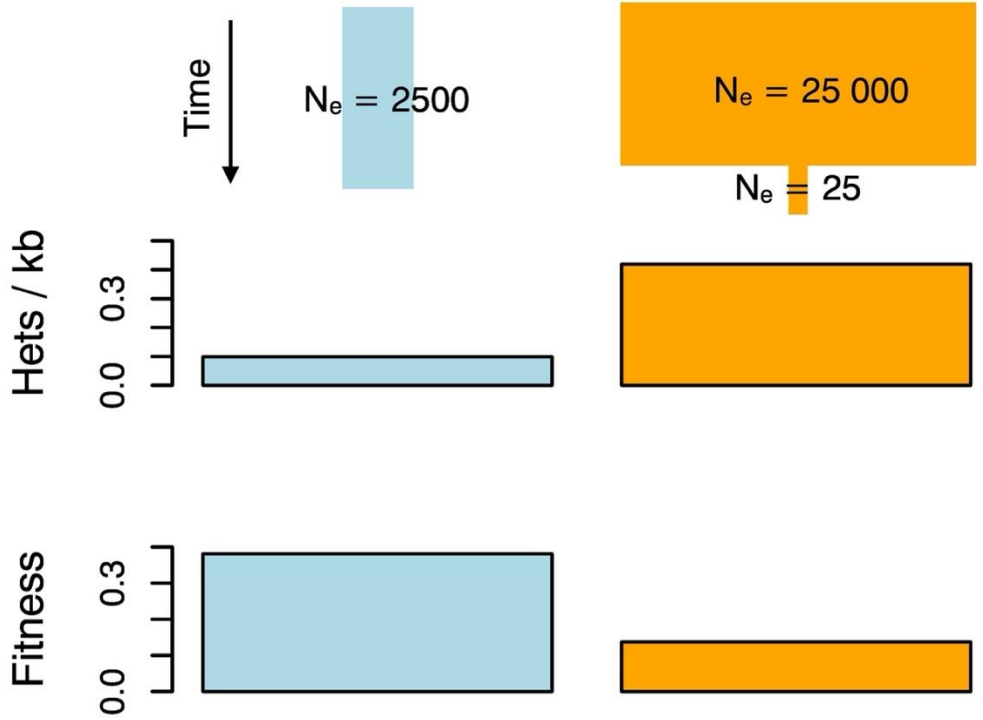
Small: remnants of ancient inbreeding events where the resulting ROHs have been fragmented due to recombination

Long: stretches of DNA that are directly identical by descent and can be used to assess the levels of recent inbreeding in the population



Diversity, inbreeding and fitness

- Demographic history:
  - confounds the link between genomic diversity and fitness

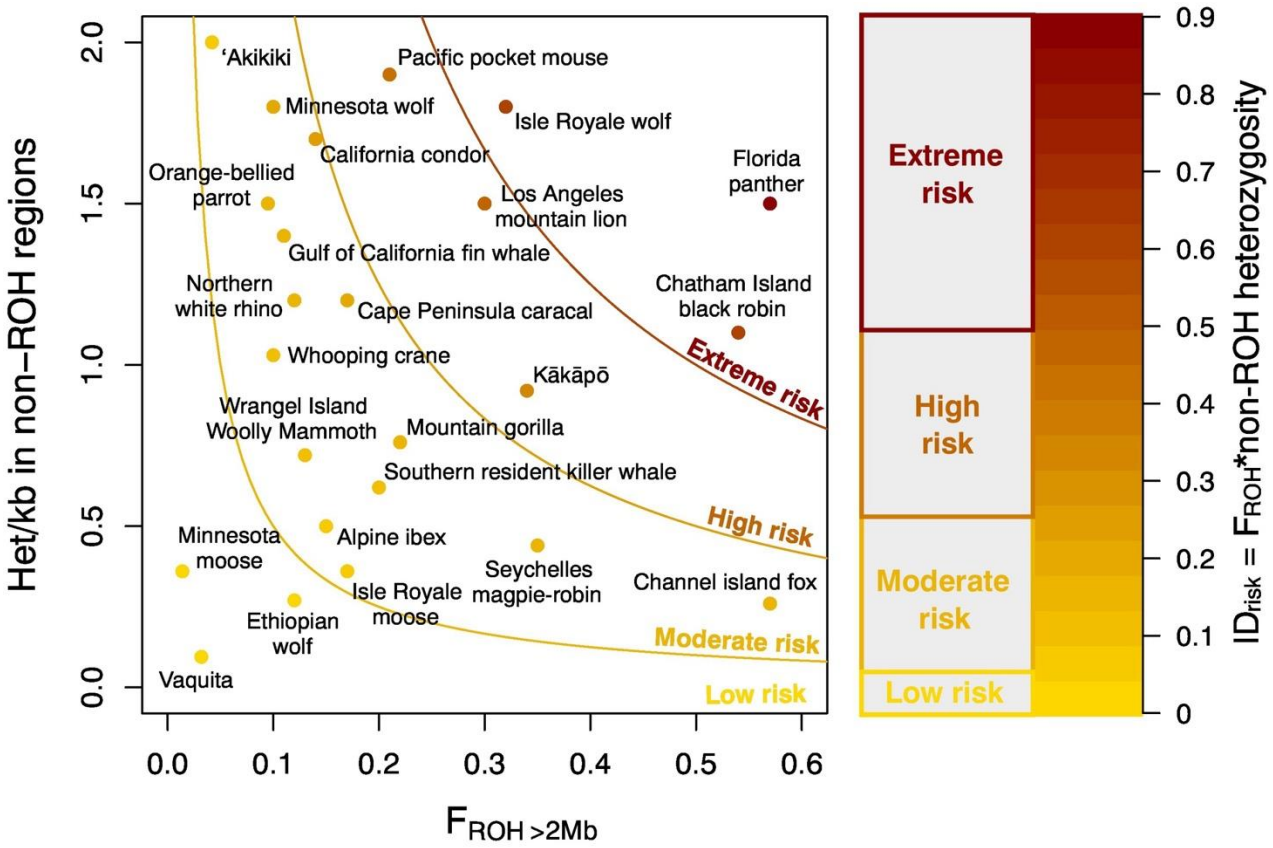


- But… long ROHs are minimally impacted
  - almost entirely a product of recent inbreeding among closely related individuals

Kyriazis et al. (2025)

Long ROHs and Inbreeding Depression

The integration of long ROH into conservation strategies provides a powerful tool for assessing population viability.



Trends in Ecology & Evolution

ID_{risk} statistic

- Increases: inbreeding
- Decreases: admixture or genetic rescue

Kyriazis et al. (2025)

Inbreeding and outbreeding

Inbreeding and outbreeding (genetic rescue or outbreeding depression)

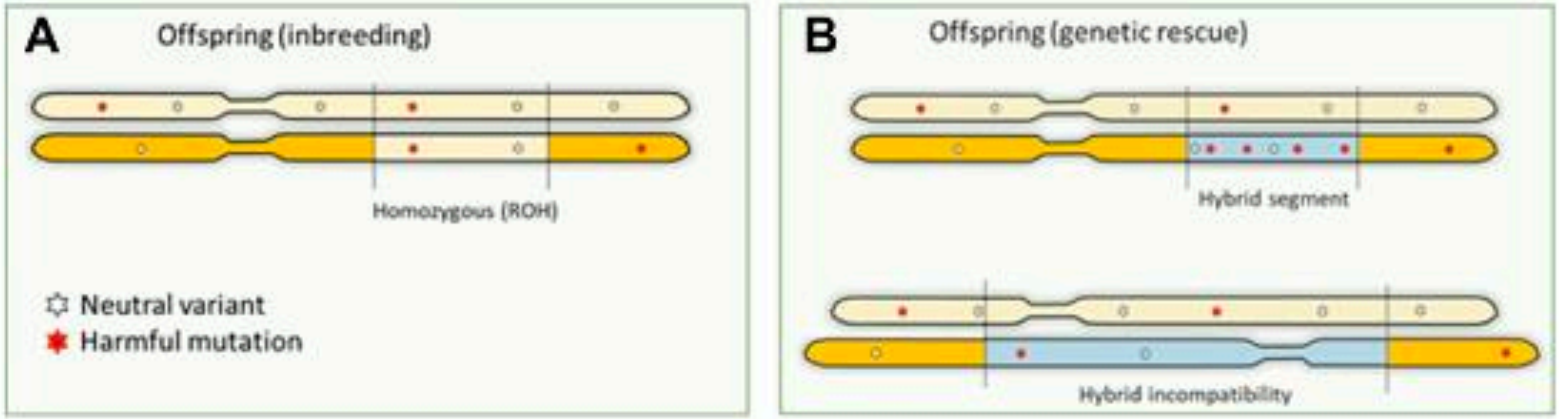
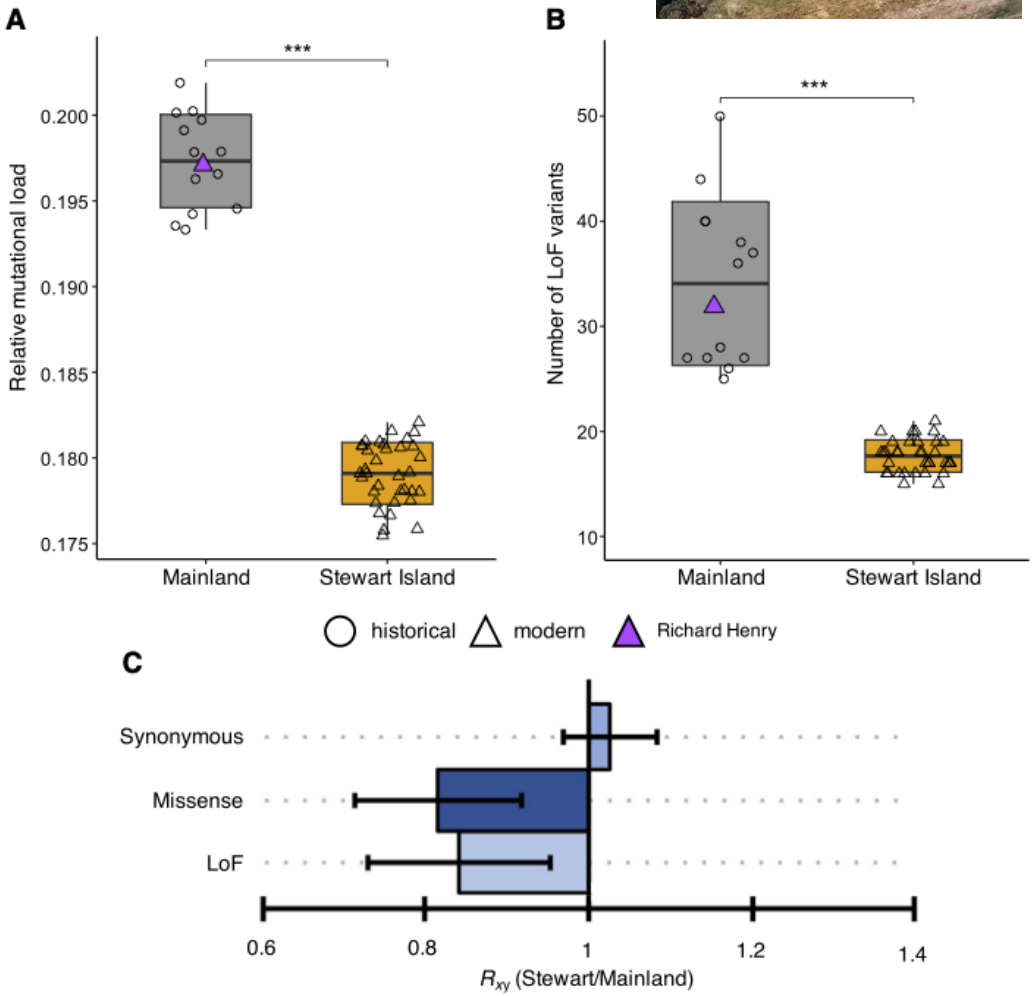


FIGURE 3
Genomic consequences of inbreeding and outbreeding. **(A)**. Offspring inherit identical DNA twice when parents are related, which means that harmful recessive mutations become homozygous and expressed. This is the primary mechanism causing inbreeding depression. **(B)**. Outbreeding (genetic rescue) introduces DNA from another source into a population, increasing genetic variation. However, this can also increase the number of harmful mutations as a masked load. These mutations could become expressed in homozygotes in future generations, as inbreeding continues to convert the masked load into a realized load. Also, large structural differences between the donor and source will result in hybrid incompatibilities.

Bosse et al (2022)

Quantifying genetic load

- How the genetic load of a population has changed through time:
  - Sum of all derived (hom and het) alleles multiplied by their conservation score (GERP score > 2)/total number of derived alleles.
  - Number of LoF variants
  - Rxy ratio of derived allele frequencies for synonymous, missense, and LoF variants

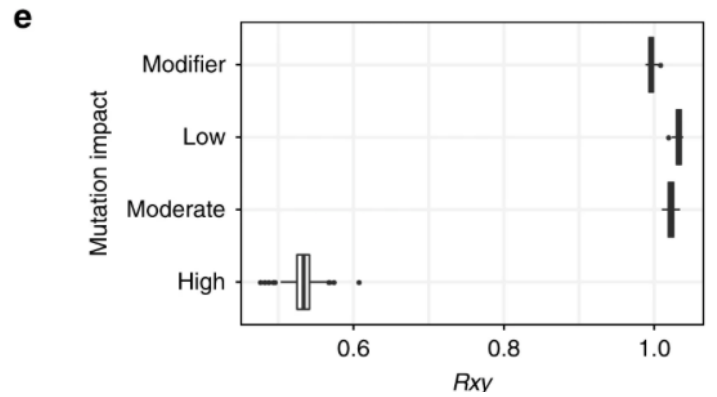


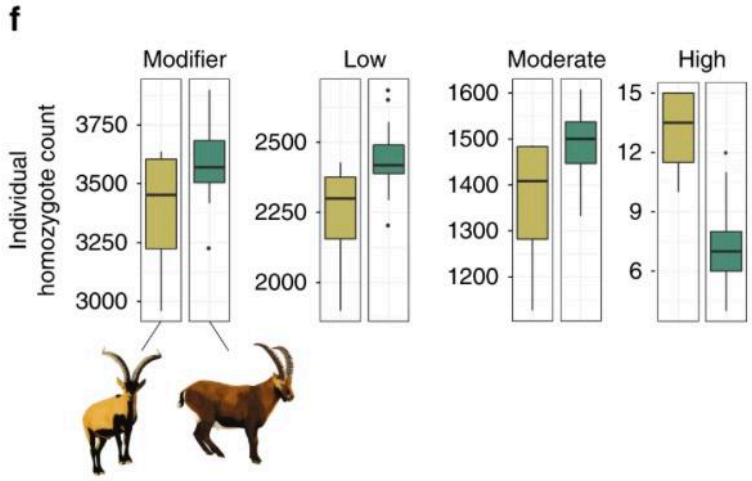
- Purging vs accumulation?
  - Reduced mutational load is due a combination of genetic drift and purging of deleterious mutations, through increased inbreeding and purifying selection.

Dussex et al (2021)

Quantifying genetic load

- How the genetic load of a population compares to sister species:
  - Rxy ratio of derived allele frequencies for different SNPeff categories
  - Homozygous count





- Purging vs accumulation?
- Even if there is purging of a subset of the most deleterious mutations, moderately deleterious variants might become fixed as a result of genetic drift in most of the cases

Grossen et al (2020)

Quantifying genetic load

Scores related to deleteriousness		
Categorical basic (e.g. synonymous/non-synonymous substitution)	**Categorical, deleteriousness** (e.g. high or moderate effect)	**Numerical** (e.g. GERP score)

Proxies of the genetic load based on the deleteriousness scores (Most of the indices below can be separately computed at homozygous or heterozygous sites, within different classes of allelic frequencies and within or outside runs of homozygosity		
Number or proportions of: • non-synonymous sites • derived alleles **Ratio between:** • variation at non-synonymous (or zerofold degenerate) sites and variation at synonymous (or fourfold degenerate) sites • number of private mutations in different populations	**Number or proportions of:** • fixed deleterious alleles • segregating deleterious alleles **Ratio between:** • number of high-damage sites and number of low-damage sites • number of sites with deleterious mutations and number of synonymous variants • number of private deleterious mutations in different populations • number of deleterious alleles in homozygotes versus heterozygotes genotypes **SFS only at deleterious alleles**	**Sum of scores** **Averages across:** • all derived alleles • only homozygous loci • only highly conserved sites • only coding sites

Bertorelle et al. (2022)

Quantifying genetic load

Table 1 | **Main approaches to predict deleteriousness from genomic data**

Approach	Principle	Pros	Cons	Examples
Evolutionary conservation	Mutations at sites with a reduced number of substitutions compared with neutral expectations in multiple alignments are likely harmful	Annotation-free; allows direct comparisons between many species	Alignments among distant species could imply errors and missing regions; computationally intensive	GERP[31], PhyloP[28]
Basic annotation, physicochemical properties	Well-known functional effects of mutations in coding regions across species are used to classify a mutation	Alignment-free; simple to use and robust in classifying major classes of deleteriousness (for example, non-synonymous, stop codons, loss of function)	Requires basic annotation; focused on coding regions; numerical scores are not always predicted	SnpEff[32], Grantham scores[34]
Extended annotation	Diverse information (for example, experimental evidence, physicochemical properties and evolutionary conservation) on the predicted harmful effects of variants is weighted and integrated into one metric	Exploits multiple information types	Still largely limited to humans, model and domesticated organisms where multiple data sources are available	PolyPhen[39], SIFT[38], ANNOVAR[33], VEP[40], CADD[42], GWAVA[43]

These approaches are designed principally to quantify the load due to SNPs. Other types of polymorphisms (for example, copy number variation, short tandem repeats, transposable elements) are not necessarily captured with these methods. CADD, combined annotation-dependent depletion; GERP, genomic evolutionary rate profiling.

Challenges:

- There is no concensus
- Link to fitness
- Uneven coverage
- Need High Coverage - VCF
- Errors and postmortem damage

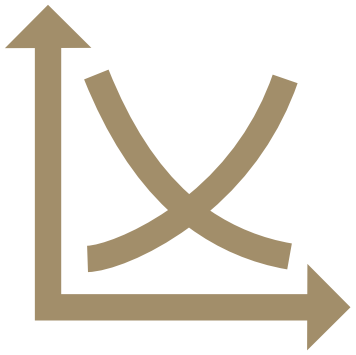
Bertorelle et al. (2022)

Demography shapes genetic variation

Demographic scenario	Predictions for neutral heterozygosity	Predictions for weakly deleterious additive mutations	Predictions for recessive strongly deleterious mutations
Long-term small population	Low across the genome	May accumulate due to drift	Few segregating due to increased drift and purifying selection
Long-term large population	High across the genome	Efficiently removed by selection	Many segregating due to being masked as heterozygotes
Recent population contraction	Slight decrease across the genome	Slight increase due to drift	Increased homozygosity; decreased total number of alleles after selection
Recent population expansion/growth	Little effect	Depends on parameters; likely little effect	Possible increases on timescale of hundreds of generations
Recent inbreeding in small population	Little effect	Little effect	Little effect because few are segregating
Recent inbreeding in large population	Regions of low heterozygosity mixed with regions of high heterozygosity	Little effect	Can become exposed due to inbreeding and decrease fitness
Gene flow from very heterozygous population	Increase	Increase in population size should decrease effects of drift, preventing fixation	Can help mask existing recessive deleterious mutations but also introduce new recessive deleterious mutations
Gene flow from moderately heterozygous population	Slight increase	Increase in population size should decrease effects of drift, preventing fixation	Can help mask existing recessive deleterious mutations without introducing many new recessive deleterious mutations

Robinson et al. (2023)

Exercises



Depth of coverage with samtools

```
dirCoverage=${directory_day4}/E5_Coverage/

samtools coverage -r ScS9zPn_17 ${data}/BAMs/AO1921.2.Black_rhino.rmdups.bam > ${dirCoverage}/AO1921.2_coverage.t
xt
samtools coverage -r ScS9zPn_17 ${data}/BAMs/TZ1909.1.Black_rhino.rmdups.bam > ${dirCoverage}/TZ1909.1_coverage.t
xt
```

TIP! I also like Mosdepth a lot!

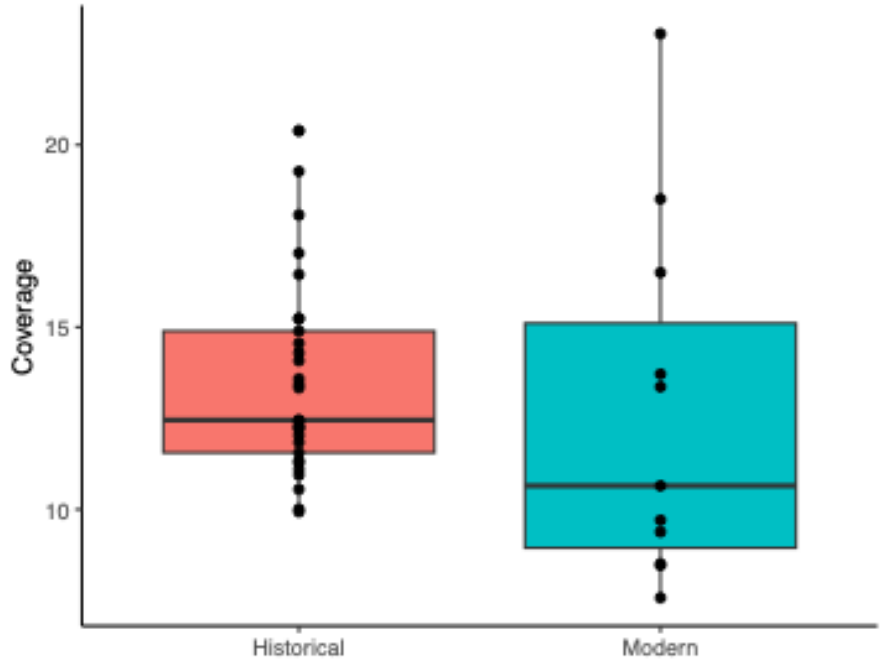
Q: How does the output from *samtools coverage* looks like?

#rname	startpos	endpos	numreads	covbases	coverage	meandepth	meanbaseq	meanmapq
ScS9zPn_17	1	30335633	4381937	29379658	96.8487	13.4314	35.3	36.4

- **Q:** Why is it important?
- **Q:** Which samples are below 5x coverage?
  - CD1925.1
  - TZ1909.1
  - ZA1775.1
  - ZA1845.2

Depth of coverage

- **Q:** Do you see differences in coverage between historical and modern samples? If so, how would you proceed?



Average Historical = ~13,5

Average Modern = ~12.5

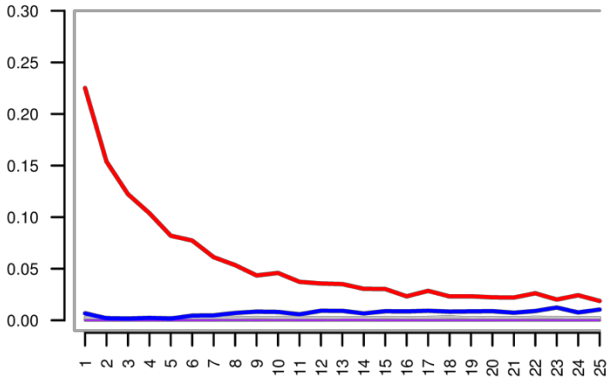
If your dataset has very different depth of coverage between samples (historical vs modern or between sites):
→ Downsample bam files to same coverage
samtools view –s

Damage Patterns

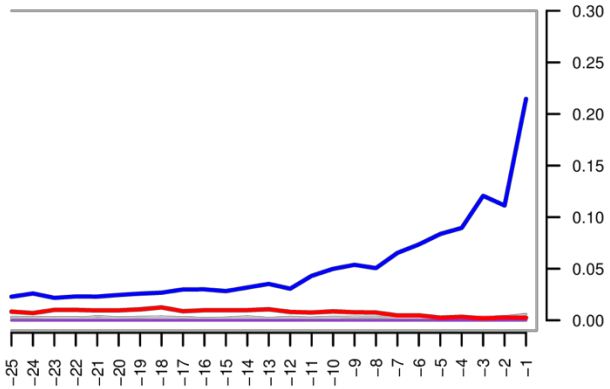
Why at the ends?

dsDNA fragments are “mostly” protected to damage but hanging ends can accumulate damage, that gets incorporated when building libraries.

5’: C->T



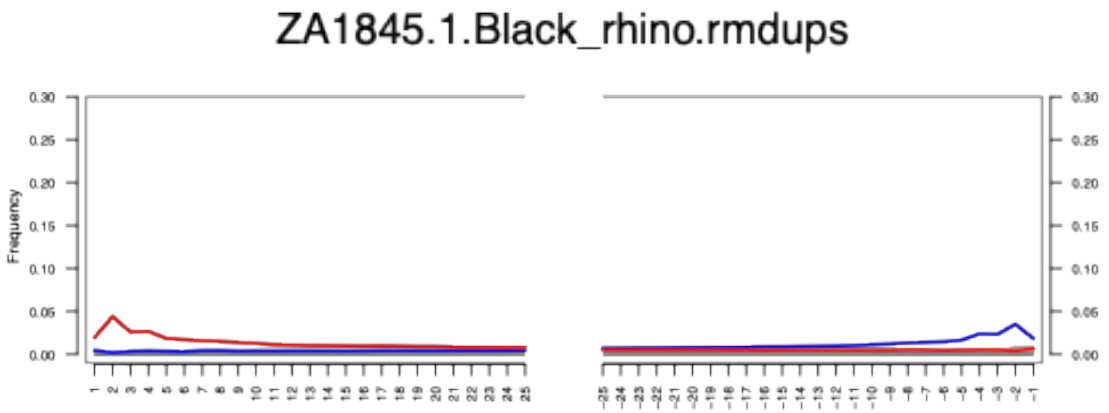
3’ G->A

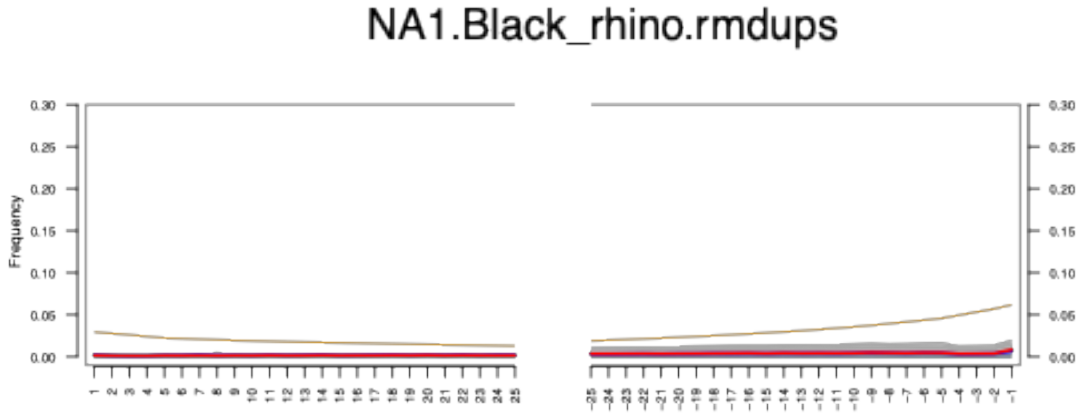


“very old” aDNA sample or “very badly” preserved sample

Damage Patterns

- **Q:** Can you interpret the plots? Do you see damage in any sample?
- **Q:** How do this findings affect downstream analysis?



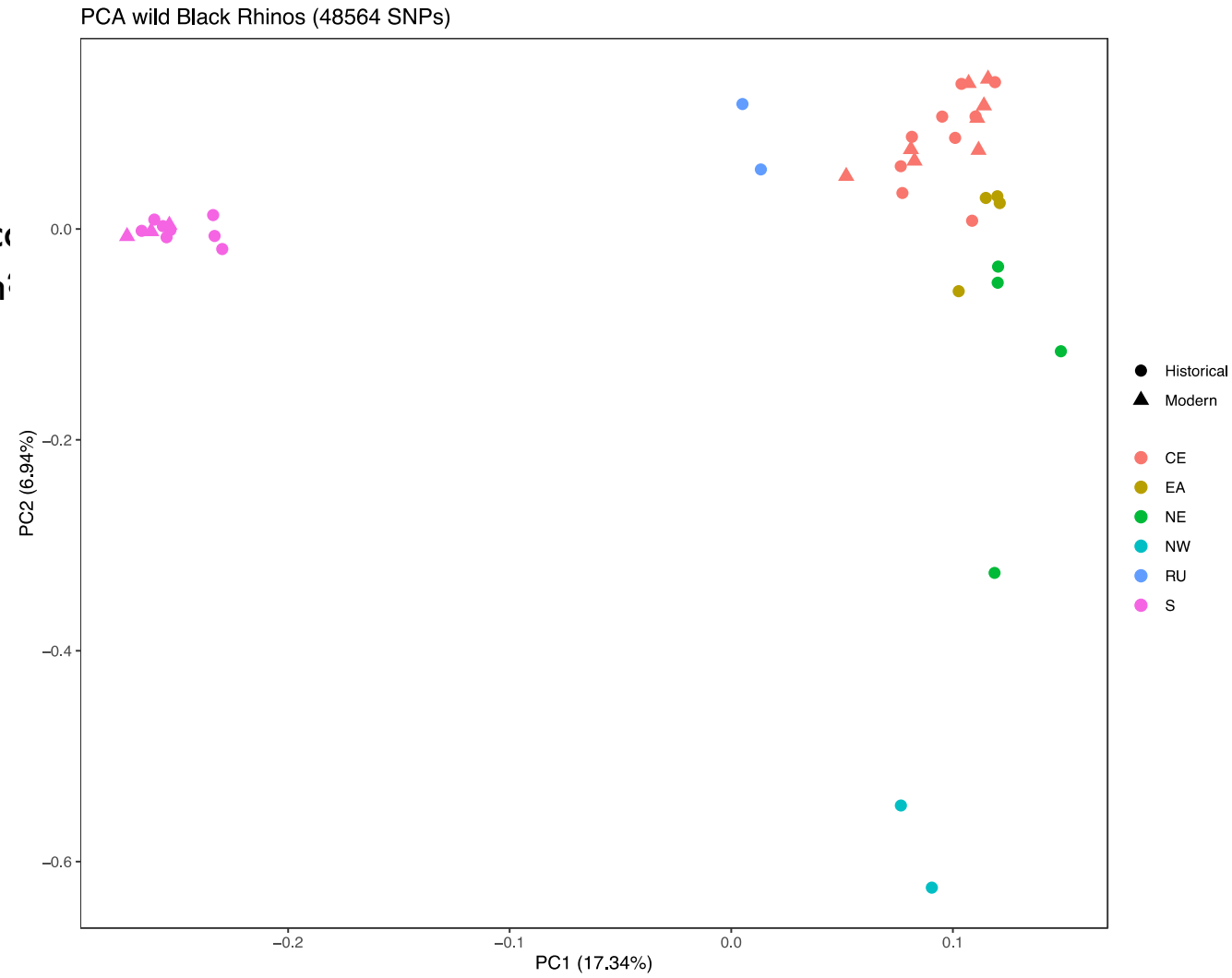


Historical population structure

Q: Are the results similar to the ones obtained from only current wild rhinos?

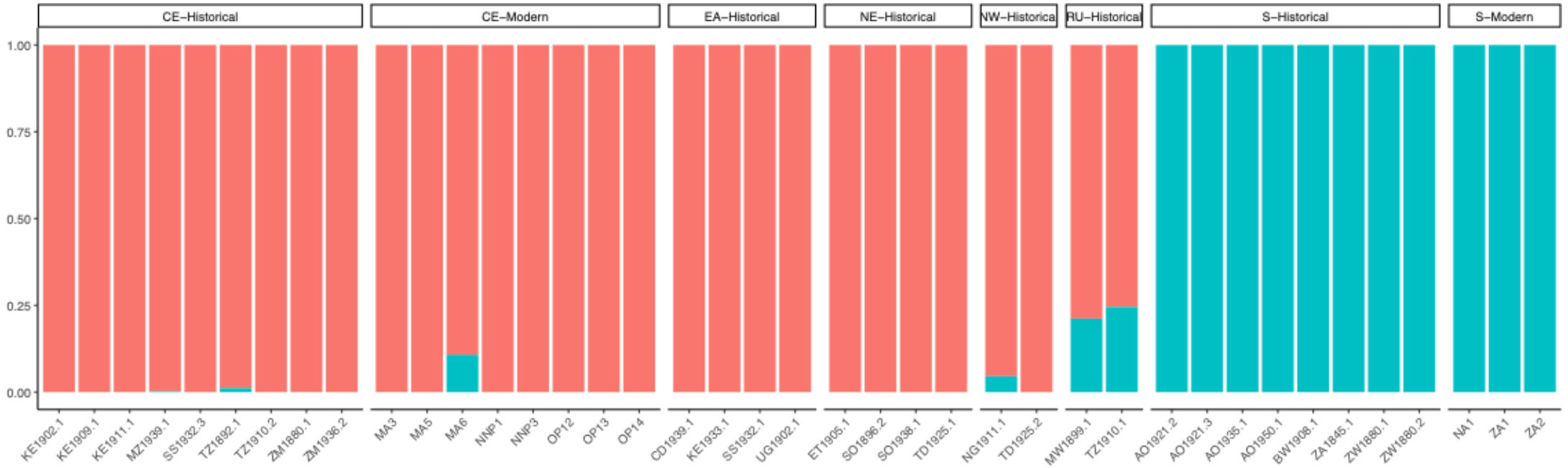
Q: How do the modern samples relate to the historic ones? Do you detect any currently extinct population

Only CE and S have historical and present-day samples.



Historical population structure

K=2

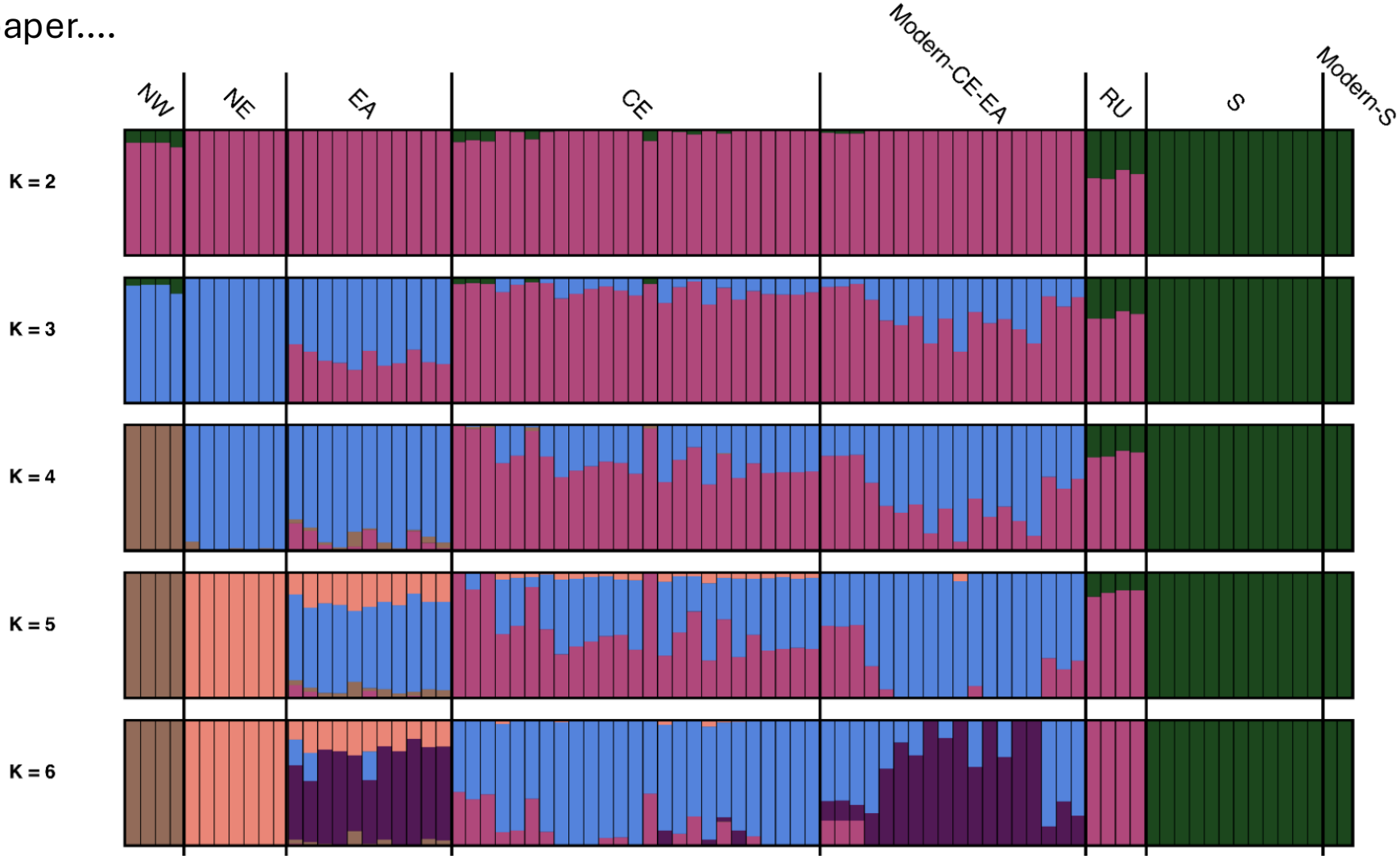


Historical population structure

K=3

Historical population structure

From the paper....



Heterozygosity estimation with ANGSD

- Site Frequency Spectrum (SFS)
  - Distribution of allele frequencies throughout the genome in a population or sample
    - Number of observed sites with derived allele frequency of x

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8
Sample 1	0	1	0	0	0	0	1	0
Sample 2	1	0	1	0	0	0	1	0
Sample 3	0	1	1	0	0	1	0	0
Sample 4	0	0	0	0	1	0	1	1
Sample 5	0	0	1	0	0	0	1	0
Sample 6	0	0	0	1	0	1	1	0
Total	1	2	3	1	1	2	5	1

n=6 individuals with eight observed variable sites

The observed allele frequency spectrum is (4,2,1,0,1)

Heterozygosity estimation with ANGSD

- Site Frequency Spectrum (SFS)
  - For diploid single samples, the heterozygous sites are simply the second value in the SFS.

- How? Two step:
  - Generate a ".saf" file (site allele frequency likelihood)

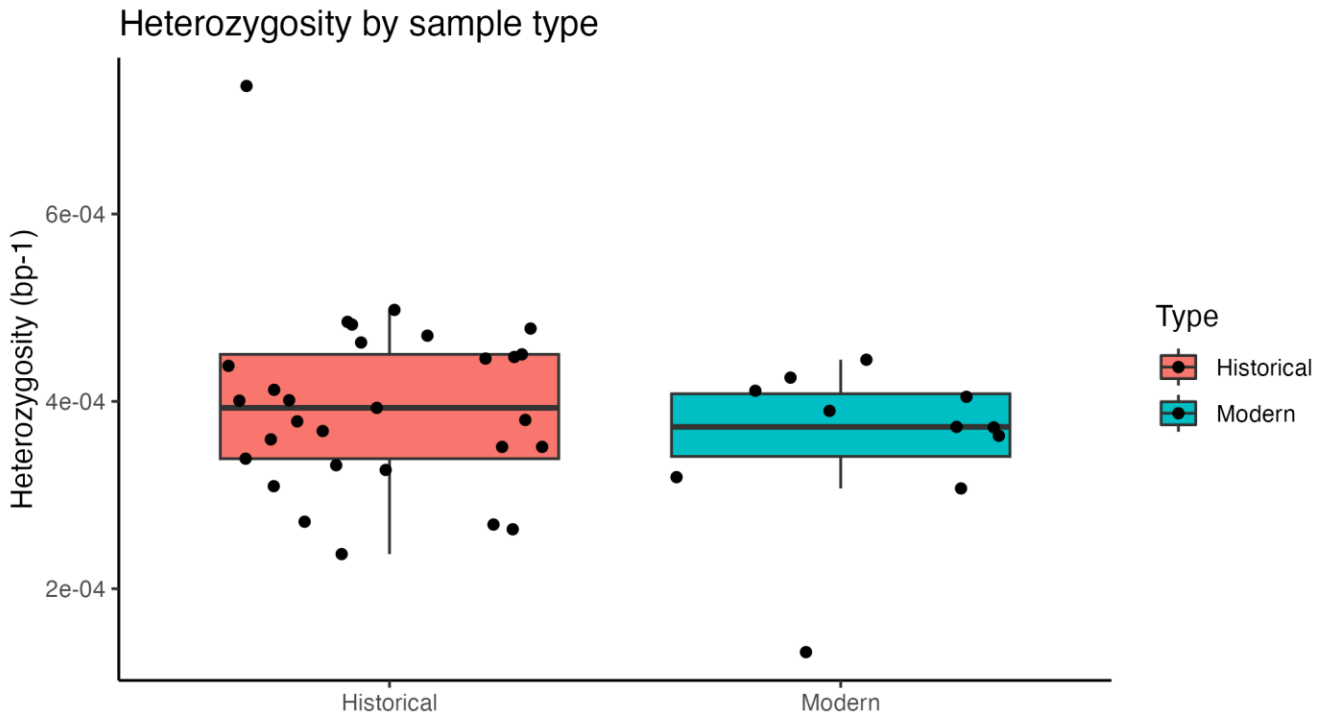
```
angsd -i $bam -ref $ref -r ScS9zPn_17: -anc $ref -out ${dirHet}/${sample} -uniqueOnly 1 -remove_bads 1 -only_prop
er_pairs 1 -noTrans 1 -C 50 -minMapQ 30 -minQ 20 -setMinDepth 4 -setMaxDepth 60 -doCounts 1 -GL 2 -doSaf 1
```

  - Optimization of the .saf file which will estimate the Site frequency spectrum (SFS)

```
realSFS ${dirHet}/${sample}.saf.idx -fold 1 > ${dirHet}/${sample}.ml
```

Heterozygosity

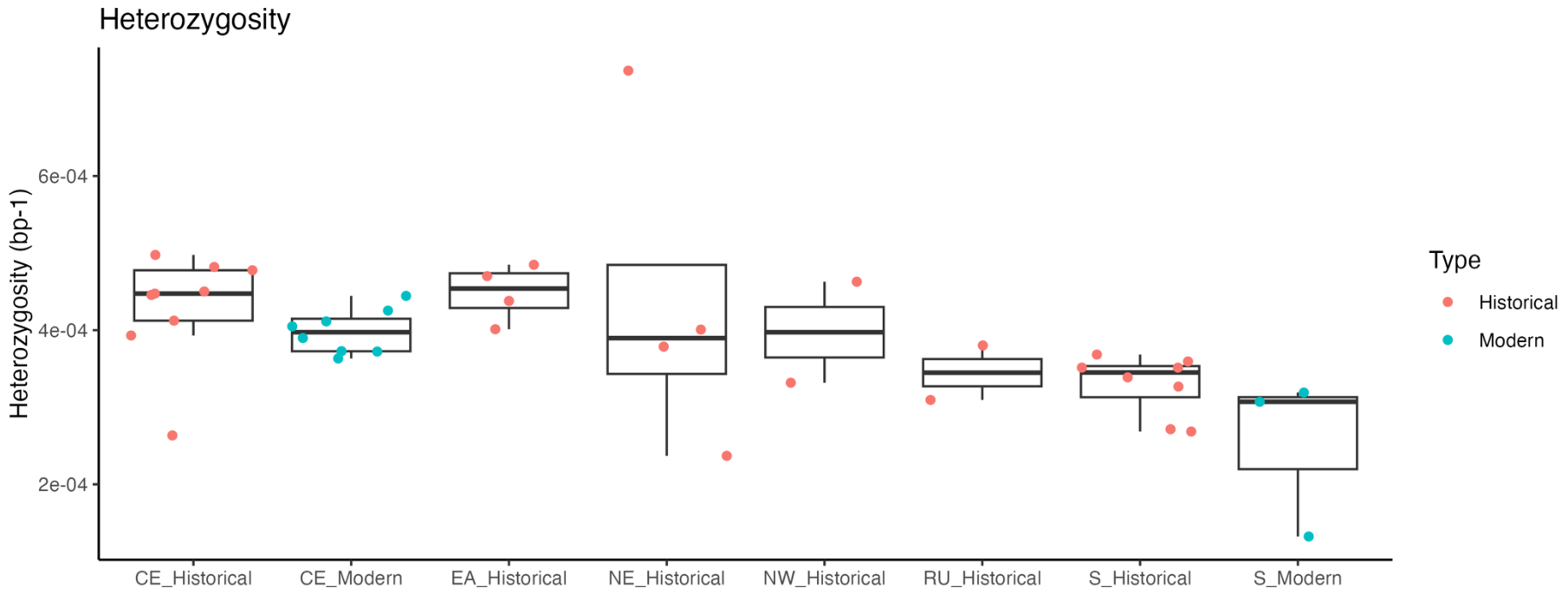
Q: Which conclusions can we take from this? Can you think of a better way to visualize it?



No difference in heterozygosity when considered all populations together

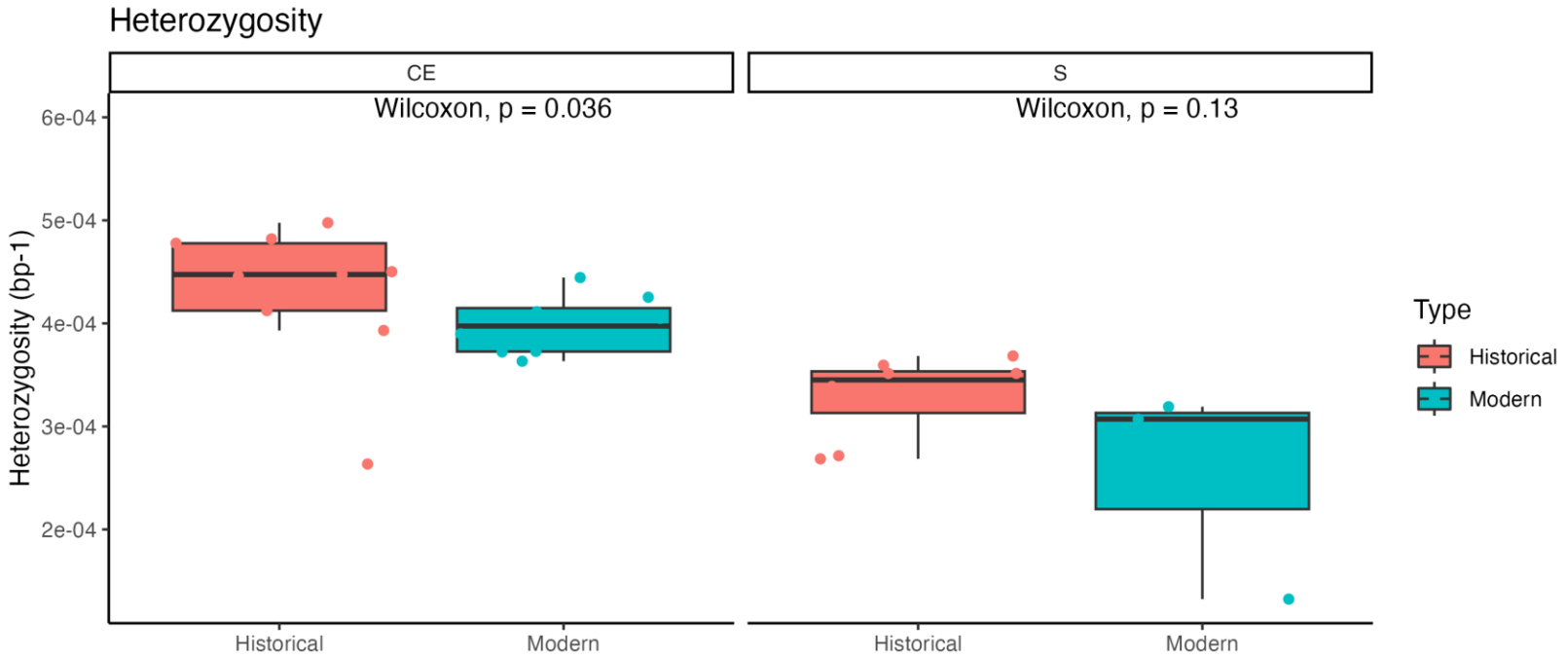
Heterozygosity

Q: Which conclusions can we take from this? Can you think of a better way to visualize it?



Heterozygosity

Q: Which conclusions can we take from this? Can you think of a better way to visualize it?



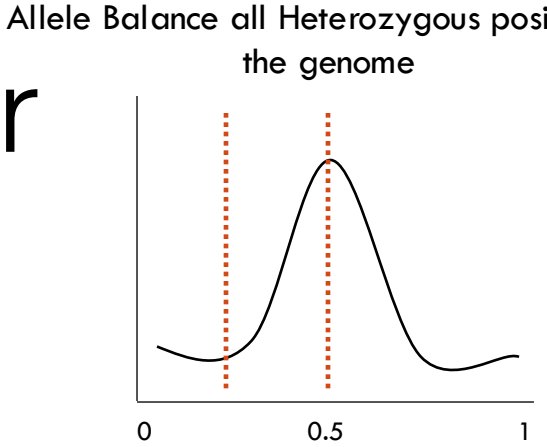
Heterozygosity - considerations

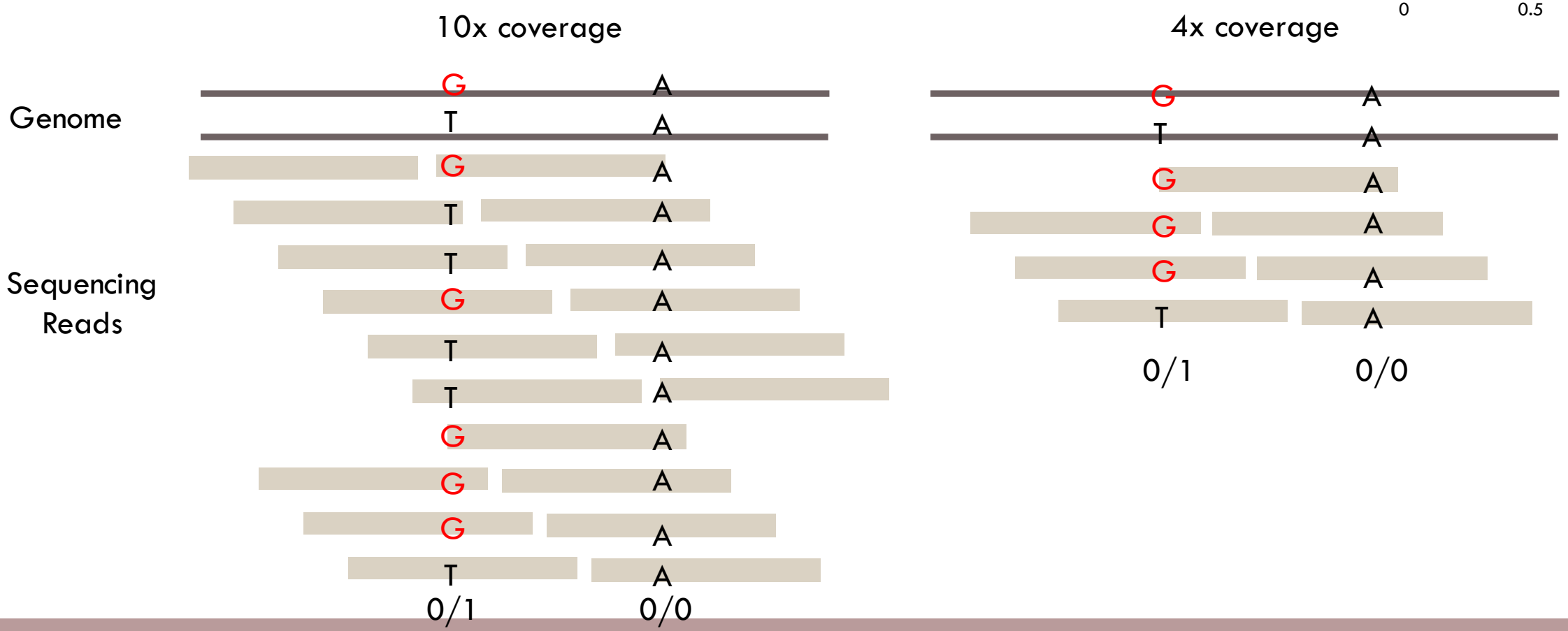
Q: What would have happened if we had used the samples with very low coverage here?

Q: What would have happened if we had used all sites and kept the transitions?

Technical aspects to consider

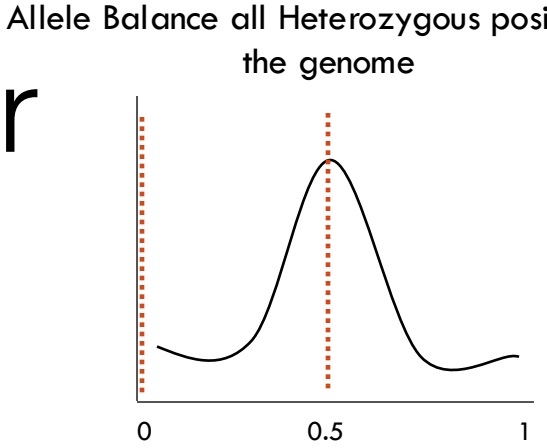
- Effect of coverage
  - Limits the capacity to detect heterozygous positions

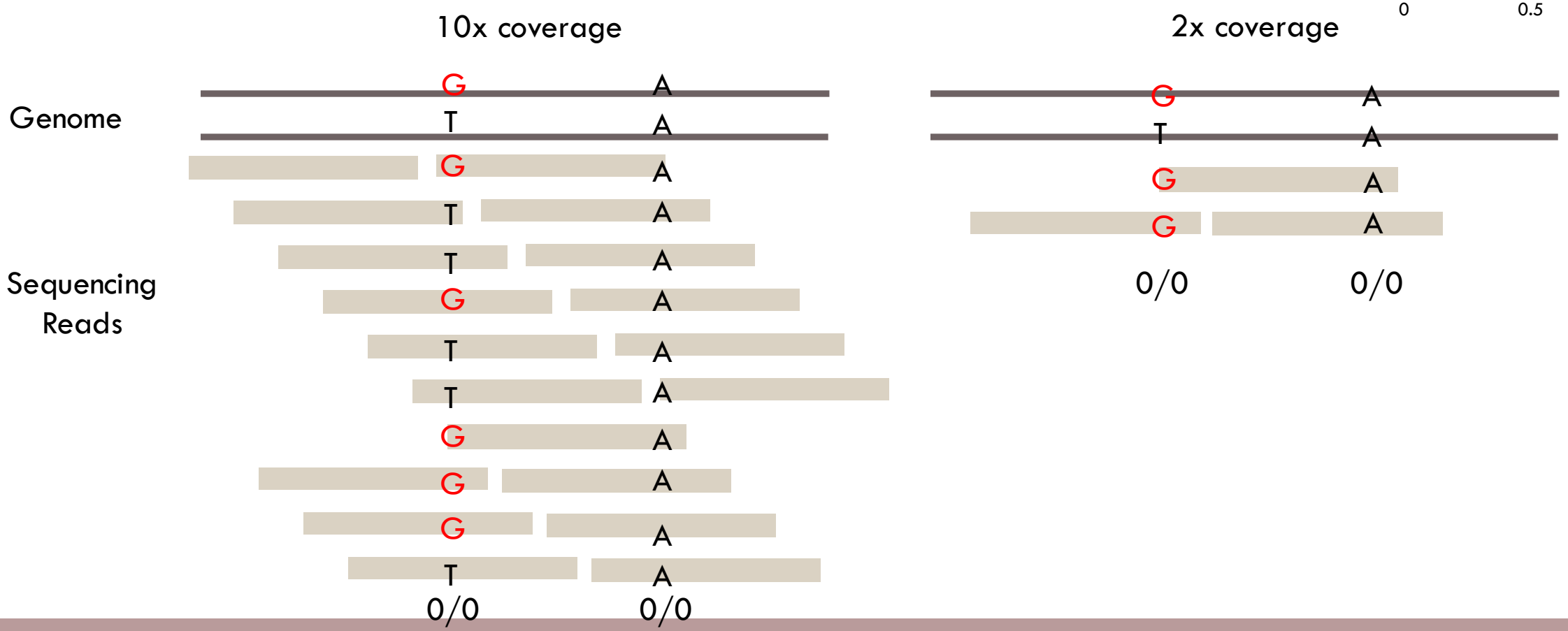




Technical aspects to consider

- Effect of coverage
  - Limits the capacity to detect heterozygous positions





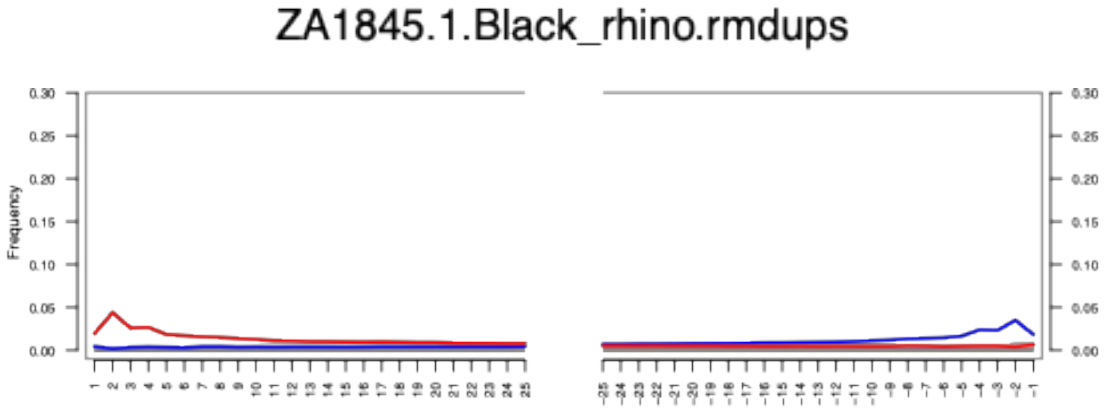
Technical aspects to consider

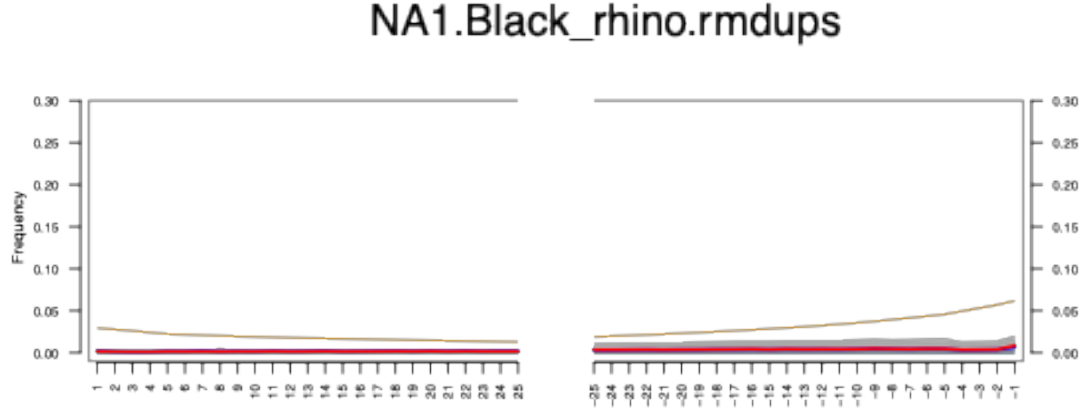
- Effect of coverage
  - How to detect this?
    - Correlation between coverage and heterozygosity values.
  - How to solve this?
    - Limit the study to samples with higher ~8-10x coverage
    - Downsample to the same coverage (even if you infraestimate the absolute value)
    - Always compare heterozygosity estimates calculated the same way
    - ANGSD as it uses GL is more resilient to coverage

Technical aspects to consider

- Effect of transitions
  - Estimate damage with Mapdamage plots in modern and in historical

Although it is moderate compared to ancient samples it is not negligable

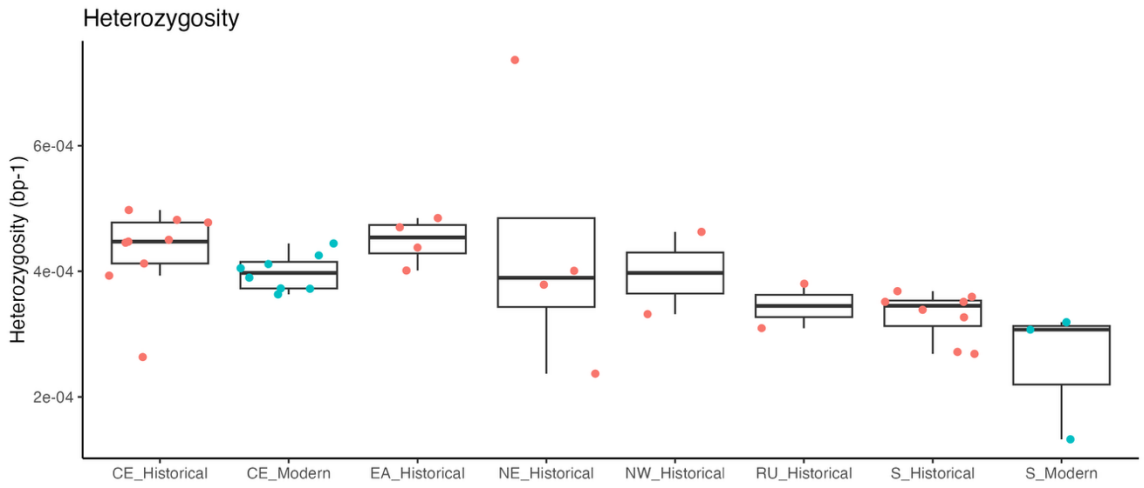


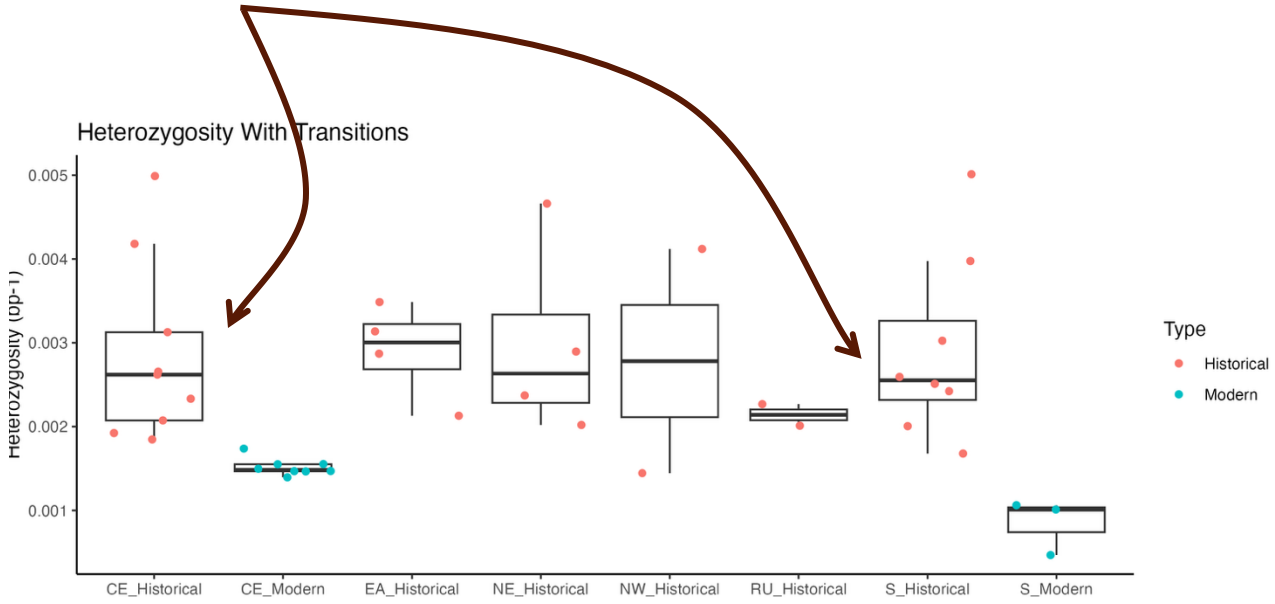


Technical aspects to consider

- Effect of transitions

Inflated heterozygosity in historical samples due to damage





Inbreeding (F_{ROH})

Estimate proportion of the genome in ROHs

Rohan when working with aDNA:

- Estimate damage profile:

```
estimateDamage.pl --length 50 --threads 16 -o $dirRohan/${sample} $ref ${bam}
```

- Run Rohan:

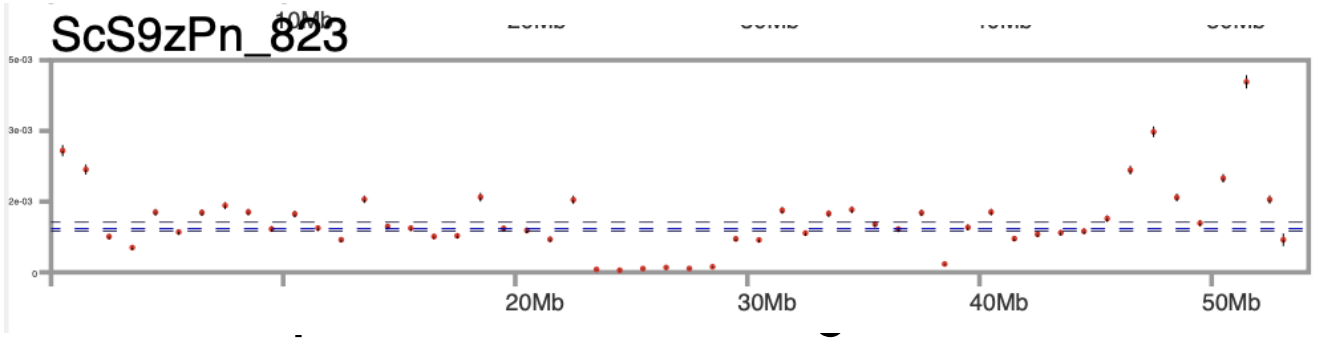
```
rohan --size 1000000 --rohmu 5e-5 --deam5p $dirRohan/${sample}_1.5p.prof --deam3p $dirRohan/${sample}_1.3p.prof --auto  ${dirRohan}/scaffold.txt -o ${dirRohan}/${sample}_aDNA_5e5_1Mb $ref ${bam} # do not run!
```

Inbreeding (F_{ROH})

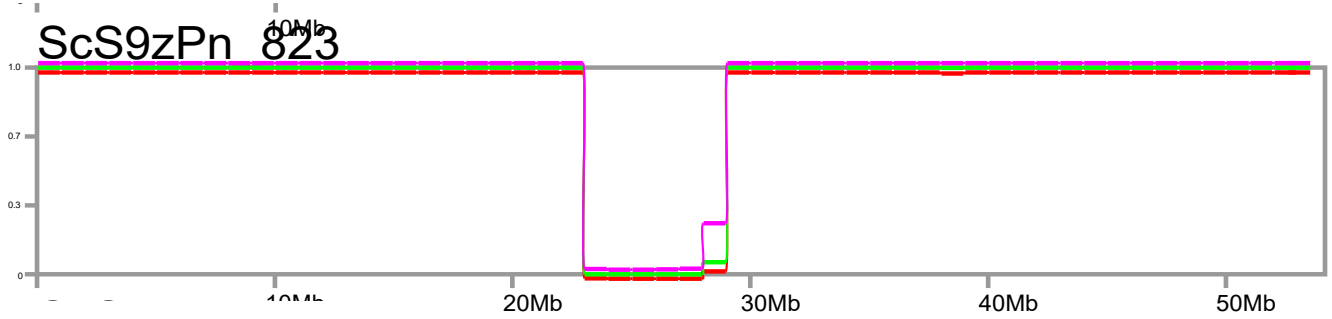
```
Command line: /home/qvw641/bin/rohan/src/rohan -t 16 --size 1000000 --rohmu 5e-5 --auto /projects/mjolnir
1/people/qvw641/NY_aDNA_course/scripts/BigScaffolds_rohan.txt -o /projects/mjolnir1/people/qvw641/NY_aDNA_cours
e/Rohan_course//NA1_modern_5e5_1Mb /projects/mjolnir1/data/genomes/black_rhino/black_rhino_19Dec2016_S9zPn.fasta
/projects/mjolnir1/people/qvw641/NY_aDNA_course/Dataset/NA1.Black_rhino.rmdups.bam
Github version: abb23a9fe1287a8ea5f2f8084fa83e22fb5776c4
Genome-wide theta outside ROH:  0.00110719  0.00105115  0.0012613
Genome-wide theta inc. ROH: 0.000956613 0.000908679 0.00111573
Segments unclassified    :  38000000 (36000000,52000000)
Segments unclassified (%):  2.09302 (2.09302,3.02326)
Segments in ROH          :  222000000 (195000000,224000000)
Segments in ROH(%)       :  13.1829 (11.5933,13.5511)
Segments in non-ROH      :  1462000000 (1458000000,1473000000)
Segments in non-ROH (%)  :  86.8171 (86.6825,89.1107)
Avg. length of ROH       :  3.97959e+06 (3.58065e+06,4.14815e+06)
```

Inbreeding (F_{ROH})

- Heterozygosity: Plot of the local estimates of heterozygosity 1 of NA1

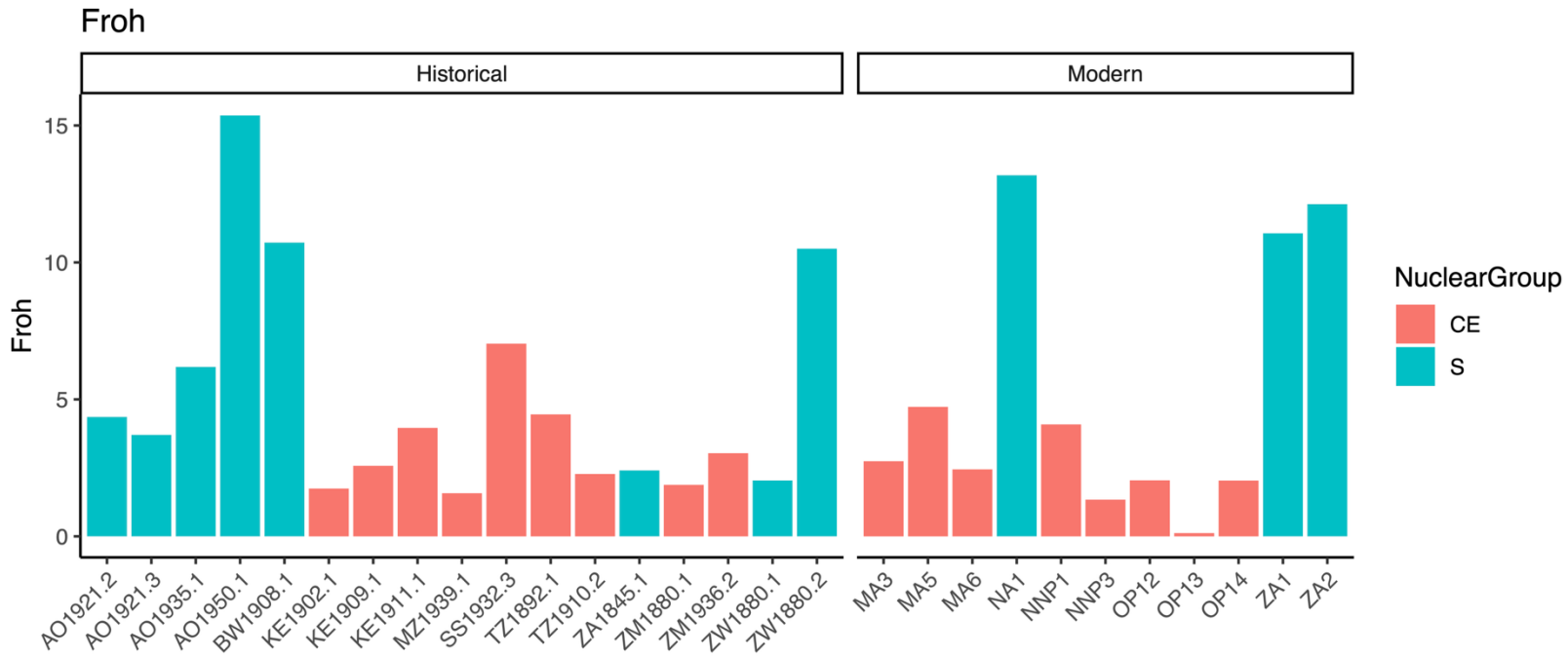


- ROH: P A1



Inbreeding (F_{ROH})

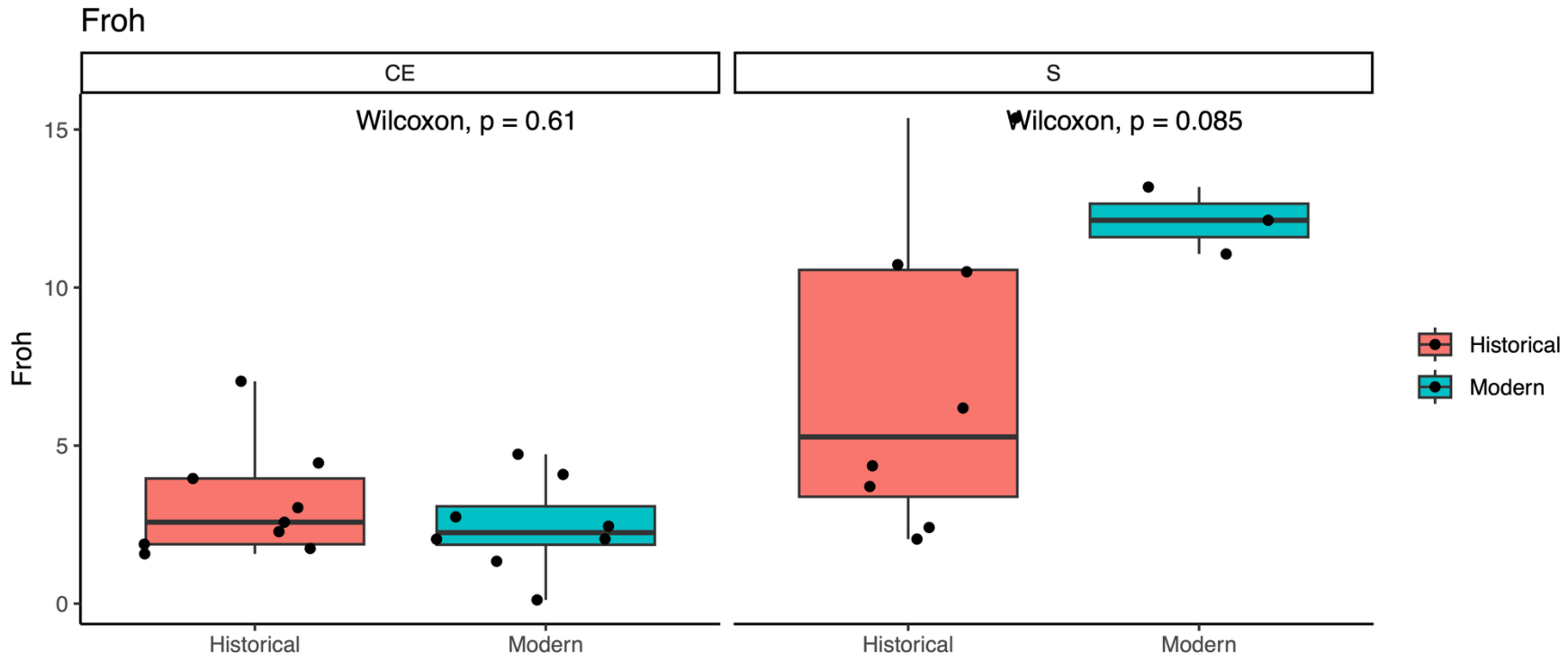
Q: Which is the sample with the highest F_{ROH} (Inbreeding coefficient)? And the lowest?



Inbreeding (F_{ROH})

Q: Do you see any significant difference between F_{ROH} (Inbreeding coefficient) between groups?

Q : How do you interpret this results? Do you have some hypothesis?



Conclusions

- Genomics applied to a dense spatial and temporal sampling can provide insights into
  - past demographic history of endangered species
  - delimitate important conservation unites – taxonomic delineation
  - quantify genetic erosion.
- Genetic erosion occurs in declining populations.
- Only observed when the decline is not extremely fast, since it would lead to extinction even before the first genetic effects establish in the genome.
  - Long generation time

Conclusions

We have analized a dataset of Black Rhino genomes and determined their:

- Historical and current population structure
- Described genetic erosion
- Highlighted key quality control steps: relatedness, coverage, damage patterns. And their impact if not dealt appropriatly

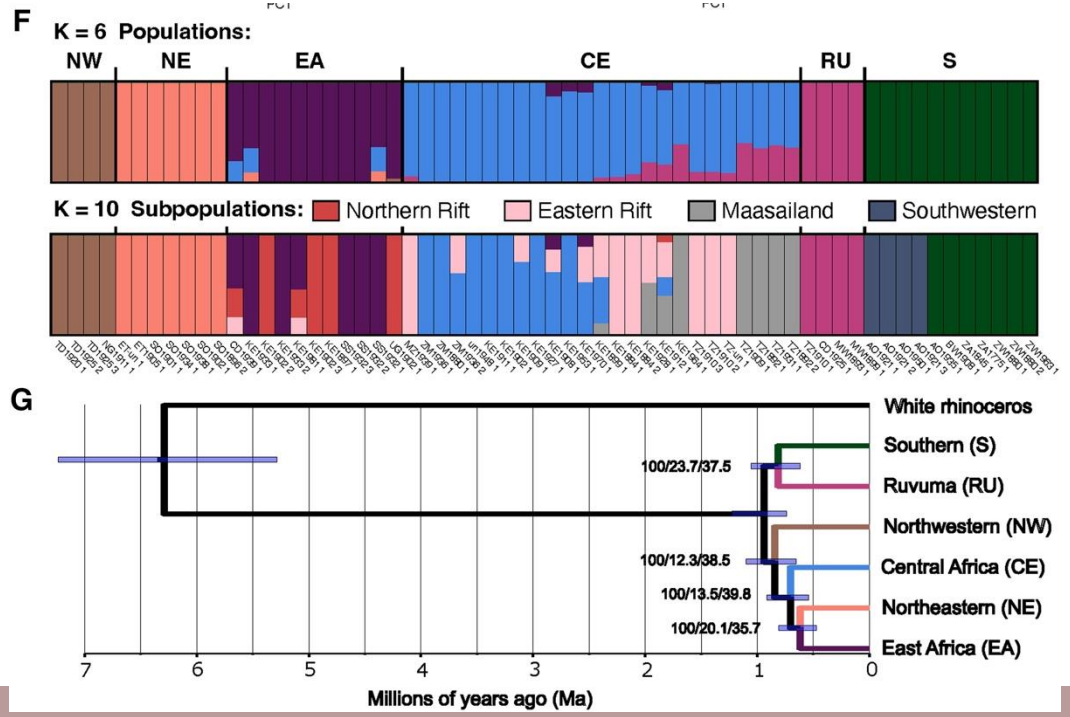
JOURNAL ARTICLE

Historic Sampling of a Vanishing Beast: Population Structure and Diversity in the Black Rhinoceros

Fátima Sánchez-Barreiro, Binia De Cahsan ✉, Michael V Westbury, Xin Sun, Ashot Margaryan, Claudia Fontsere, Michael W Bruford, Isa-Rita M Russo, Daniela C Kalthoff, Thomas Sicheritz-Pontén ... Show more

Author Notes

Molecular Biology and Evolution, Volume 40, Issue 9, September 2023, msad180,
https://doi.org/10.1093/molbev/msad180
Published: 10 August 2023



Future prospects

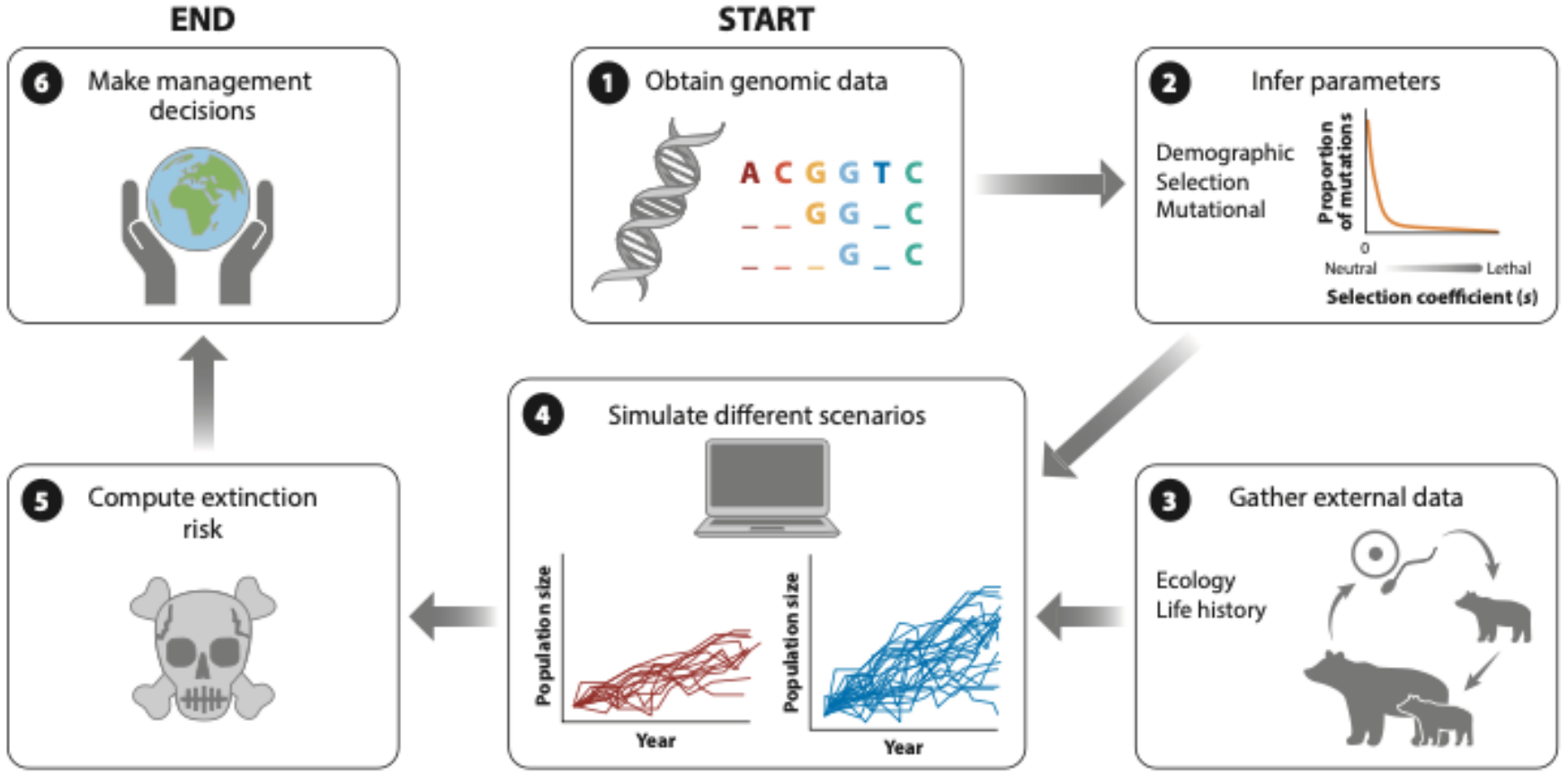


Figure 3

Simulation workflow for genomic forecasting including genomic data and deleterious mutations. This framework was used to show that the vaquita is not doomed to extinction from inbreeding depression (80).

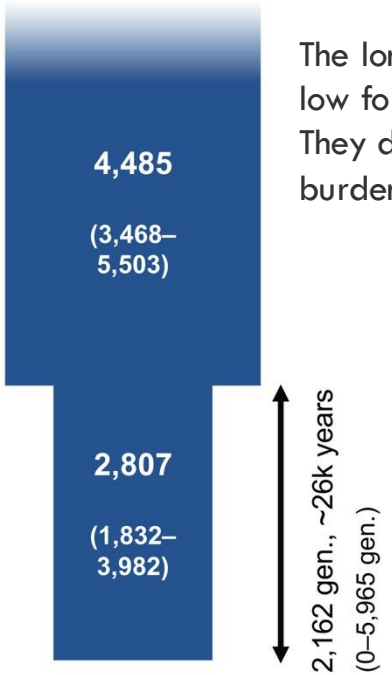
Robinson et al (2023)

REPORT | CONSERVATION

The critically endangered vaquita is not doomed to extinction by inbreeding depression

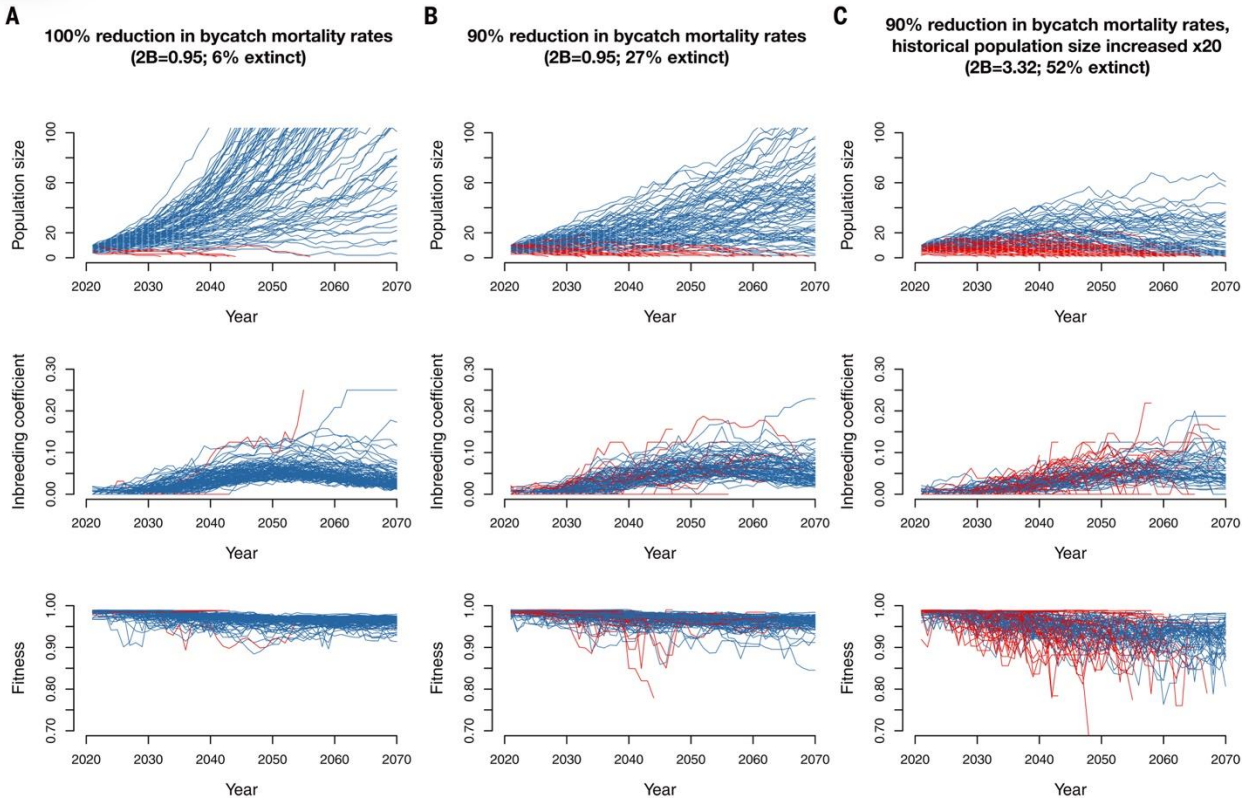
JACQUELINE A. ROBINSON, CHRISTOPHER C. KYRIAZIS, SERGIO F. NIGENDA-MORALES, ANNABEL C. BEICHMAN, LORENZO ROJAS-BRACHO, KELLY M. ROBERTSON, MICHAEL C. FONTAINE, ROBERT K. WAYNE, KIRK E. LOHMUELLER, [...], AND PHILLIP A. MORIN +1 authors Authors Info & Affiliations

SCIENCE • 5 May 2022 • Vol 376, Issue 6593 • pp. 635-639 • DOI: 10.1126/science.abm1742



The long-term population size of vaquitas has been low for a marine mammal.
They do not suffer from inbreeding depression (low burden of deleterious variants).

Genome-informed simulations suggest that the vaquita can recover if bycatch mortality is immediately halted



Future prospects

- Access to thousands of genomes → genome consortia
- Explore other types of variation
  - Structural Variation
    - CNVs
  - Regulation – epigenetics

- Link between computational variant effect preditions (deleterious and adaptive) to real-world fitness consequences
- Function of microbes within microbiome of endangered species.
- Improve technology to use low-quality, non-invasive samples in portable and cost-effective devise.
- From vulnerability genomics to restoration genomics (van der Valk & Dalén (2024)

Future prospects

- From vulnerability genomics to restoration genomics (van der Valk & Dalén (2024)