

**Bioinformatics
approaches in
animal breeding**

*Summer
School*

July 9-11, 2025
Zagreb

University of Zagreb



University of Ljubljana



THE USE OF MOLECULAR INFORMATION IN LIVESTOCK SELECTION

Brajković Vladimir, Shihabi Mario

**Wednesday
9th July**

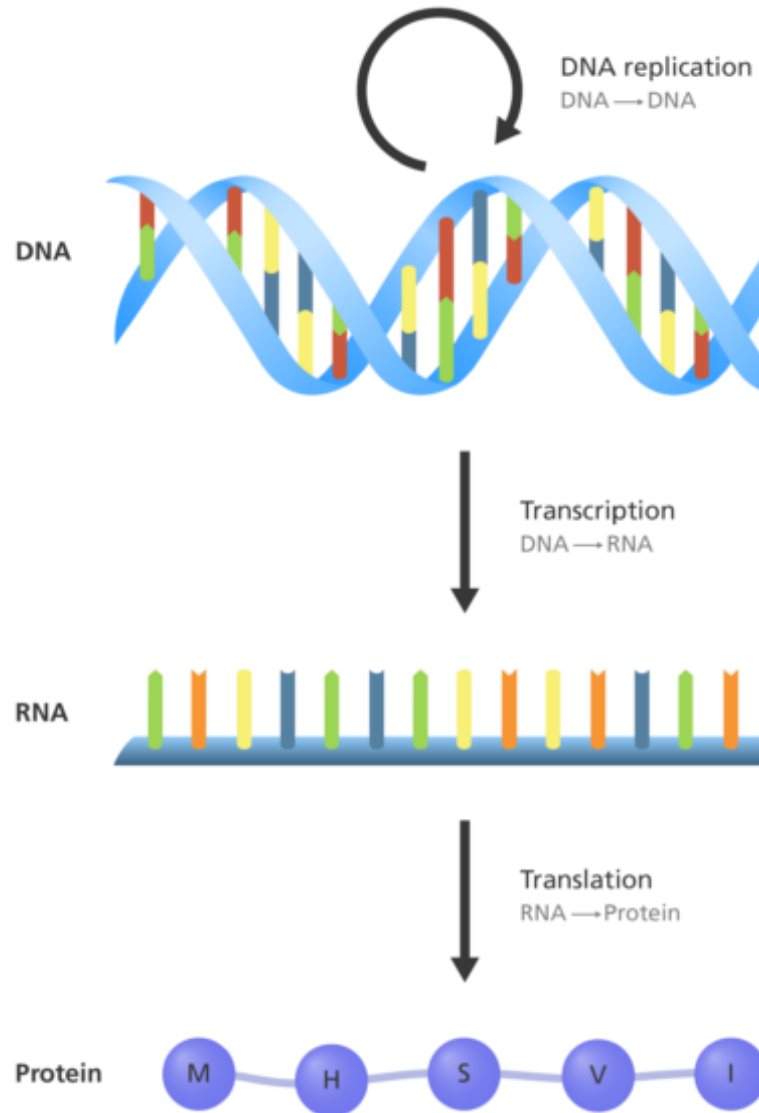
Contents

- **What is molecular information?**
 - About DNA and genome*
 - Genome composition*
 - Single Nucleotide Polymorphisms (SNPs)*
 - Associations with Phenotype*
- **Usage of molecular information in livestock selection**
- **Case study**
 - Identification of Selection Signals on the X chromosome in East Adriatic Sheep: A New Complementary Approach*

Summer
School

Bioinformatics
approaches in
animal
breeding

Central dogma of molecular biology



DNA – Deoxyribonucleic Acid

- Double-stranded molecule
- Stores entire genetic blueprint of individual

RNA – Ribonucleic Acid

- Single-stranded molecule transcribed from DNA
- messenger RNA (mRNA) - coding sequences
- regulatory (miRNA) and structural forms (rRNA, tRNA)

Proteins

- Structural and functional complex macromolecules
- Synthesized by ribosomes based on mRNA templates

**Molecular
information**

Data at
DNA
RNA
or
protein level

**Bioinformatics
approaches in
animal
breeding**

*Summer
School*

About DNA and genome

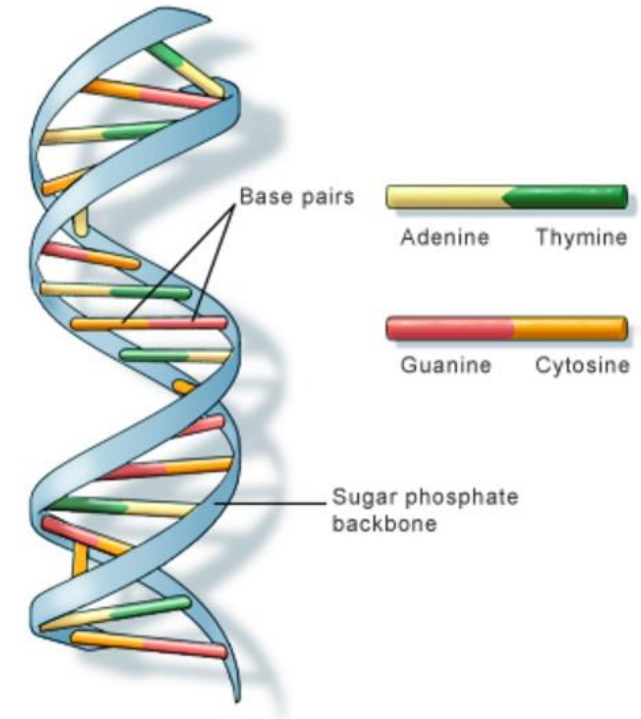
DNA (Deoxyribonucleic Acid) - hereditary material in all living organisms

- **It consists of:** Deoxyribose sugar
Phosphate group
Nitrogenous bases:
 - Adenine (A)
 - Thymine (T)
 - Cytosine (C)
 - Guanine (G)

🔑 **Only the nitrogenous bases vary** among individuals

These **nucleotide differences** are the **basis of genetic variation** and thus the focus of **genomics**

Genome - full set of DNA sequences in an organism



Simpler vs. Complex genome

Simpler organisms – Prokaryotes (e.g. Bacteria)

- Typically single circular chromosome in cytoplasm
- One copy of DNA (haploid)
- Genomes are compact – fast replication, efficient adaptation
- Genetic variation arises only through mutations

Complex organisms – Eukaryotes (e.g. Livestock)

- Multiple linear chromosomes enclosed in nucleus
- Two copies of DNA (diploid)
- Genomes are large – allow complex regulation and specialization
- Genetic variation arises from both mutations and recombination
- Recombination: exchange of genetic material during meiosis
 - Enhances genetic diversity
 - Breaks linkage between harmful and beneficial mutations

Trade-off between
genome simplicity (reproduction)
&
complexity (survival)

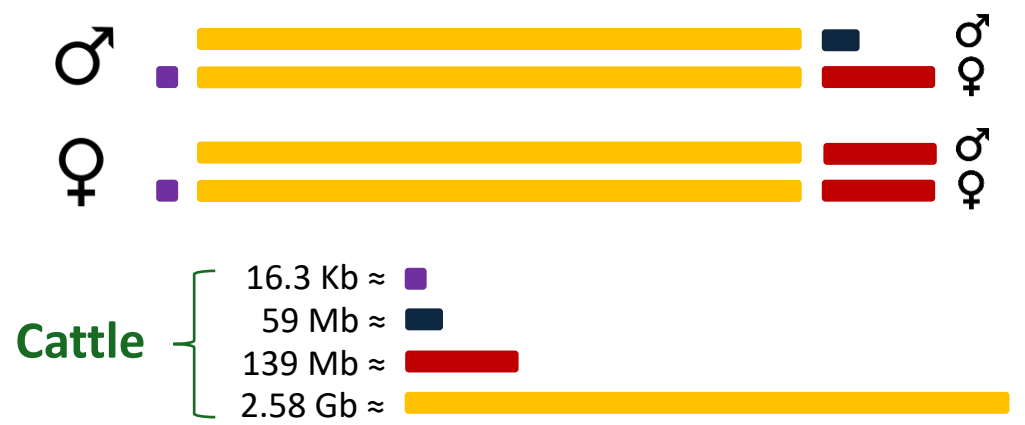
'Nothing in Nature happens accidentally'
|
genome design reflects evolutionary
strategy

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

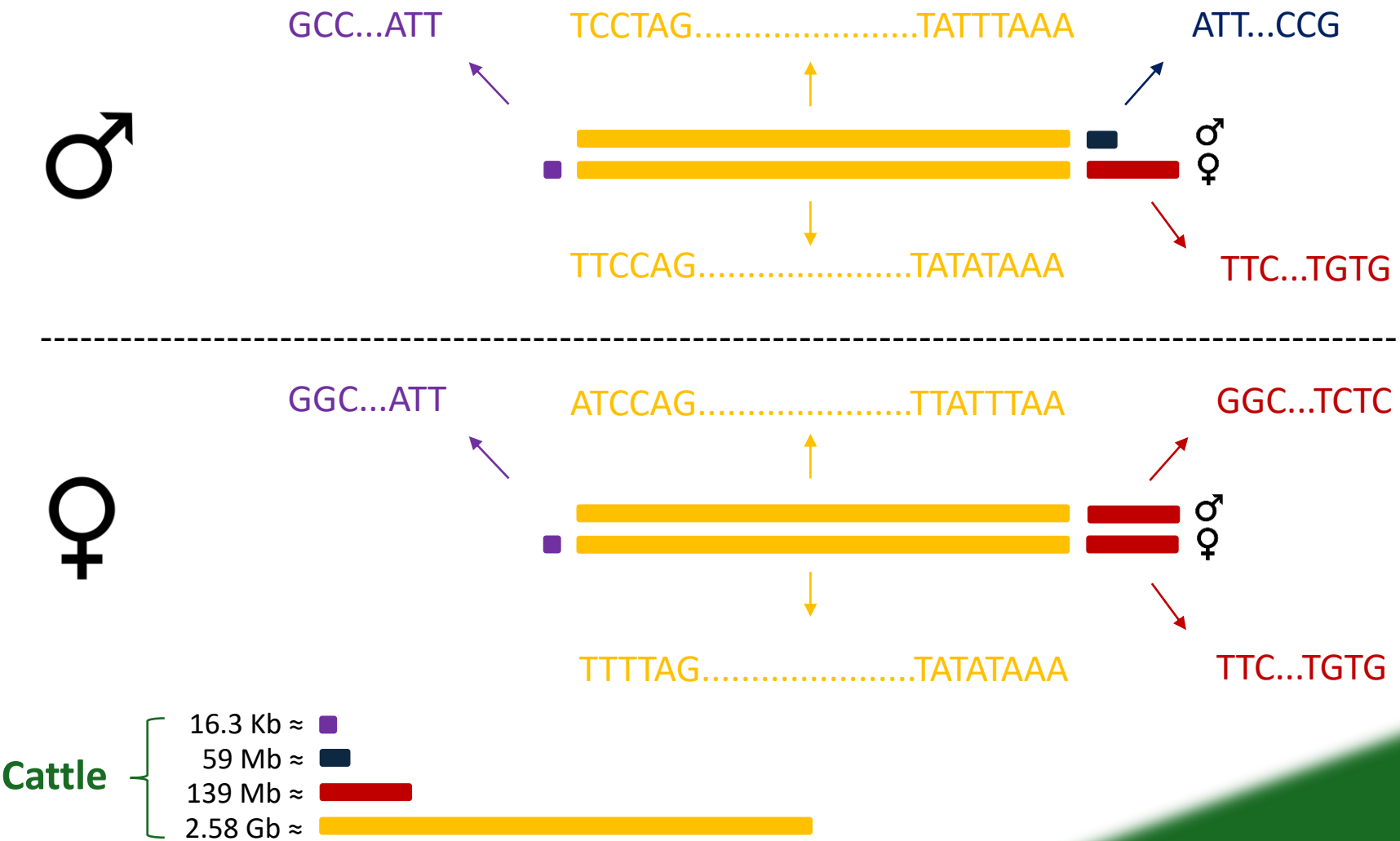
Livestock genome composition

- Consists of: **X autosomes** (non-sex chromosomes) *Depends on species (e.g. cattle = 29 pairs)*
- Sex chromosomes (**X** and **Y**)
- Mitochondrial DNA (**mtDNA**)



	Type	Inherited From	Recombines?
	Autosomes	Both parents	Yes
	mtDNA	Maternal only	No
without PAR	X chromosome	Both (♀) / Mother (♂)	Yes (♀), No (♂)
	Y chromosome	Paternal only (♂)	No

Livestock genome composition



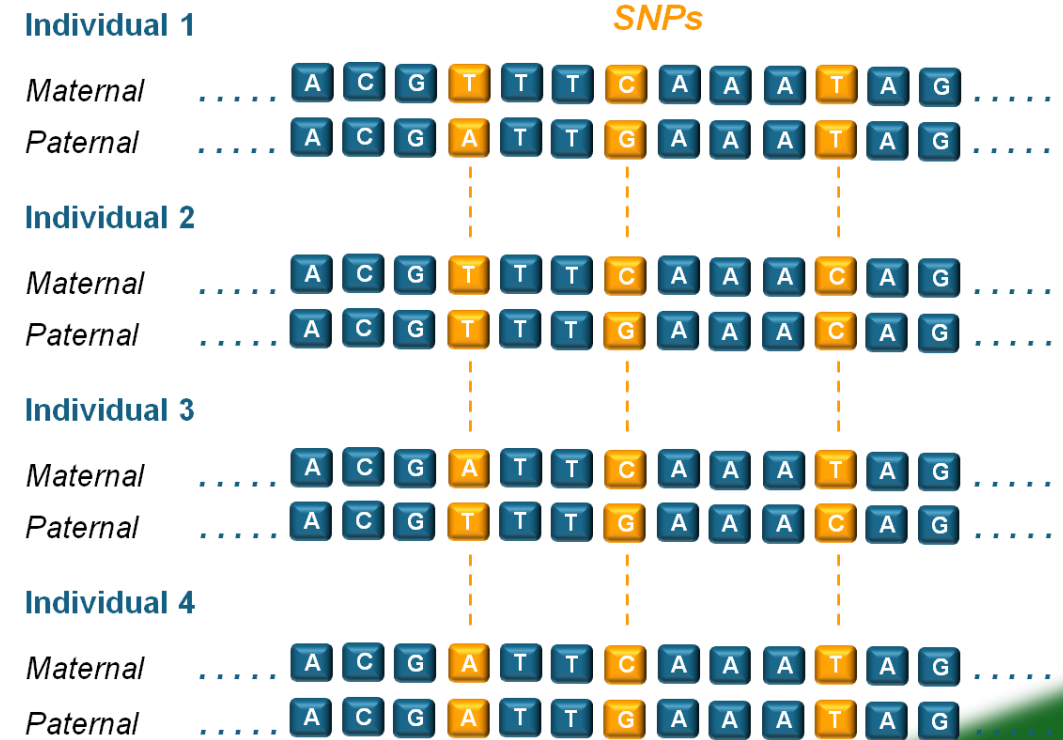
Obtaining molecular information

- In the past, molecular information was not accessible
- With technological advancement, **molecular markers** were developed
- Many types were used, with microsatellites once being dominant
- Today, **Single Nucleotide Polymorphisms (SNPs)** are the most widely applied markers
- Obtaining complete genome data (WGS: Whole Genome Sequence) is now possible, but still costly if high accuracy is required

Single Nucleotide Polymorphisms (SNPs)

Most used molecular marker

- Single base pair variations at specific genomic positions
- Widespread occurrence and informative nature
- Result from mutations spread by drift or selection
- Found across whole genome, in both coding and non-coding regions
- Most commonly obtained from SNP array data
- SNP arrays genotype thousands of informative variants reliably and cost-effectively
- Alternatively, SNPs can be extracted from WGS data (higher number, but lower per-SNP accuracy and often more noise than advantage)



Associations of molecular information & phenotype

Basic model that explains phenotype:

$$P = G + E + G \times E$$

P – Phenotype

G – Genetic effects (*mol. info. provides insights*)

E – Environmental effects

G×E – Genotype by Environment Interaction

Genetic Component (G): different genetic models

- a) Infinitesimal model – many genes with tiny effects
- b) Oligogenic model – few genes with large effects
- c) Mixed model – mostly small effects + few major genes

Different trait scenarios

$P = G + E + G \times E$ → most complex traits (e.g. milk yield, growth, fertility)

$P \approx G$ → traits mostly determined by genetics (e.g. eye color, monogenic diseases)

$P \approx E$ → traits strongly affected by environment (e.g. weight gain under extreme diet)

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Associations of molecular information & phenotype

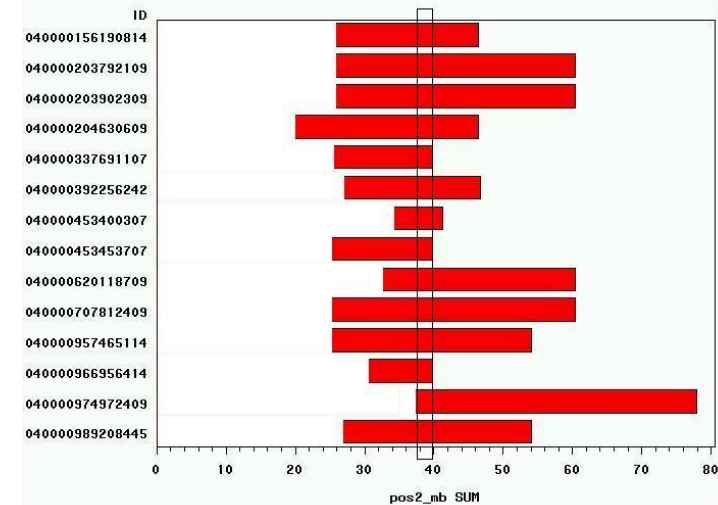
Example 1: Recessive disorder in Tyrol Grey

P ≈ G (monogenic trait)

- Occured in 2003
- Neuromuscular disorder
- Symptoms:
 - Disease breaks out at ~3 months of age
 - Loss of control over hind part of body
- Gusti (*1972) in the pedigrees of all diseased animals



Regions of chromosome 16 where all diseased animals are homozygous



1 locus at *MFN2* gene -> T/T disease
T/C & C/C healthy

Bioinformatics
approaches in
animal
breeding

Summer
School

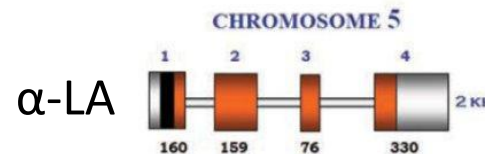
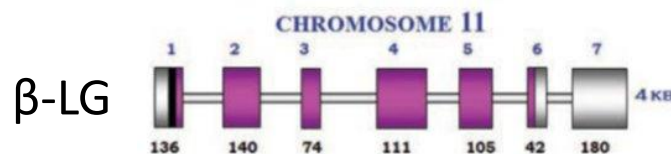
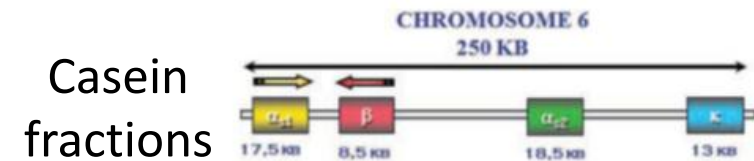
Associations of molecular information & phenotype

Example 2: Milk protein composition in dairy cattle

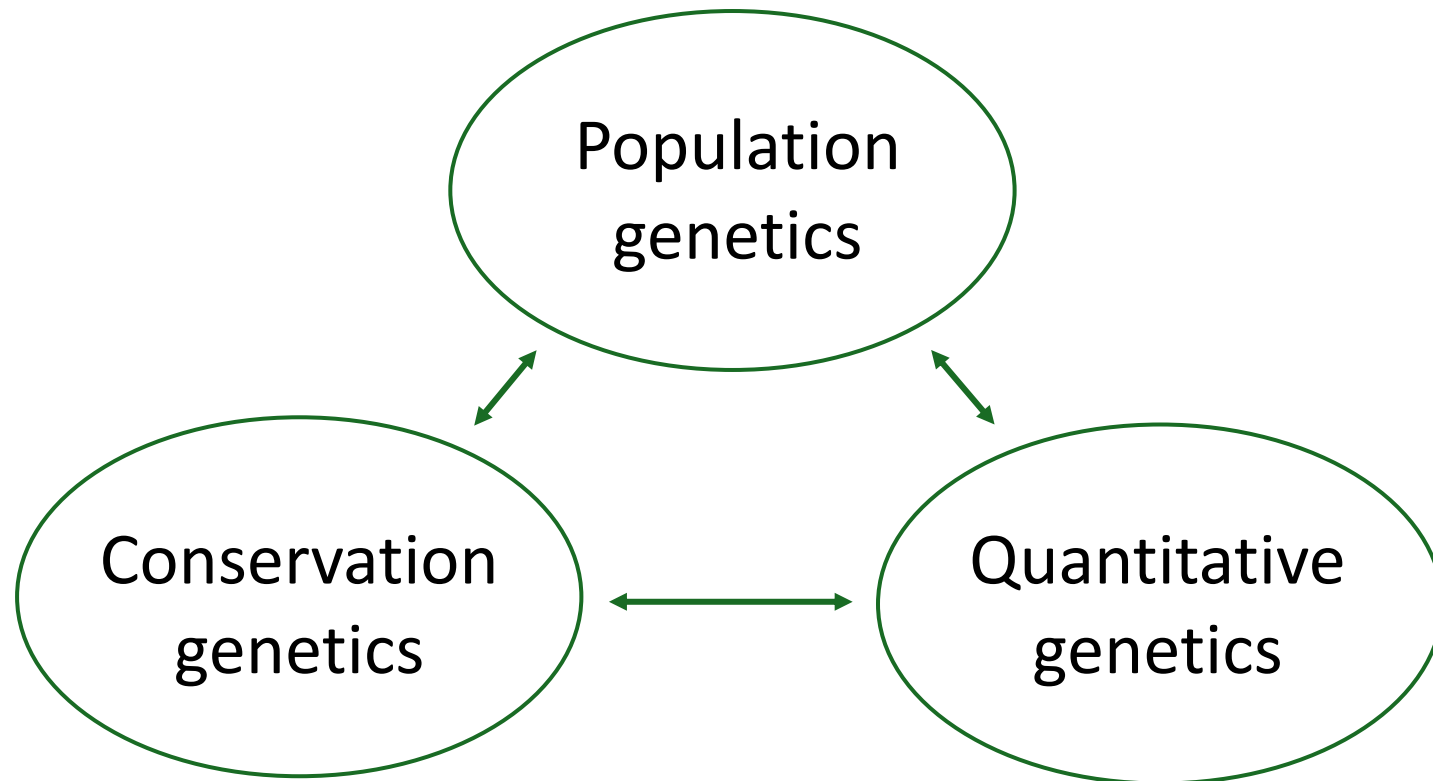
$P = G + E + G \times E$ (Polygenic trait with major genes)

Six genes are responsible for the synthesis of 95% of milk proteins

- 4 Casein fractions:
- α_{s1} -CN, β -CN, α_{s2} -CN > calcium sensitive proteins are phylogenetically related
- κ -Cn > phylogenetically related to fibrinogen
- α -lactalbumin (α -LA) & β -lactoglobulin (β -LG)
- α -LA > responsible for biosynthesis of lactose
- β -LG > way protein that binds with retinol



General applications of molecular information in livestock



Applications of molecular information in livestock selection

1. Genetic improvement of economically important traits

- Milk yield, growth, fertility, carcass quality, disease resistance

2. Detection and management of genetic defects

- Identification and elimination of carriers of harmful recessive alleles

3. Parentage verification and breed composition analysis

- Ensures accuracy in breeding programs and conservation management

4. Selection for robustness and adaptability

- Heat tolerance, disease resistance, stress resilience

5. Optimized mating strategies

- Genomic information supports mating plans that maximize gain and minimize inbreeding

....

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Main categories of molecular information use

1. Evaluation of animals

- Predicting breeding potential (e.g. GEBVs)
- Used in: MAS, GS, mate selection

2. Evaluation of markers (SNPs)

- Identifying SNPs associated with traits or populations
- Used in: GWAS, selection signals

Core methods used in selection

1. Marker-Assisted Selection (MAS)

- Uses markers linked to major genes or QTLs
- Suitable for traits mainly controlled by major genes
- Simple, but limited to known loci with large effects

Example: selection against recessive genetic disorders or for hornless (polled) allele

Core methods used in selection

2. Genomic Selection (GS)

- Uses genome-wide SNP data to predict Genomic Estimated Breeding Values (GEBVs)
- Enables early evaluation of young animals
- Powerful for complex (polygenic) traits
- Reduces generation interval and increases selection accuracy

Uses statistical models like GBLUP, BayesA/B/C, LASSO

Core methods used in selection

3. Introgression of beneficial alleles

- Controlled transfer of favorable alleles from one breed/population to another
- Supported by molecular tracking using markers
- Used in crossbreeding or gene introgression programs

Example: Booroola gene in sheep significantly increases ovulation rate and litter size (homozygous)

Core methods used in selection

4. Genomic mate selection

- Matches animals based on genomic profiles
- Helps maximize genetic gain while minimizing inbreeding
- Crucial in small populations or conservation breeding

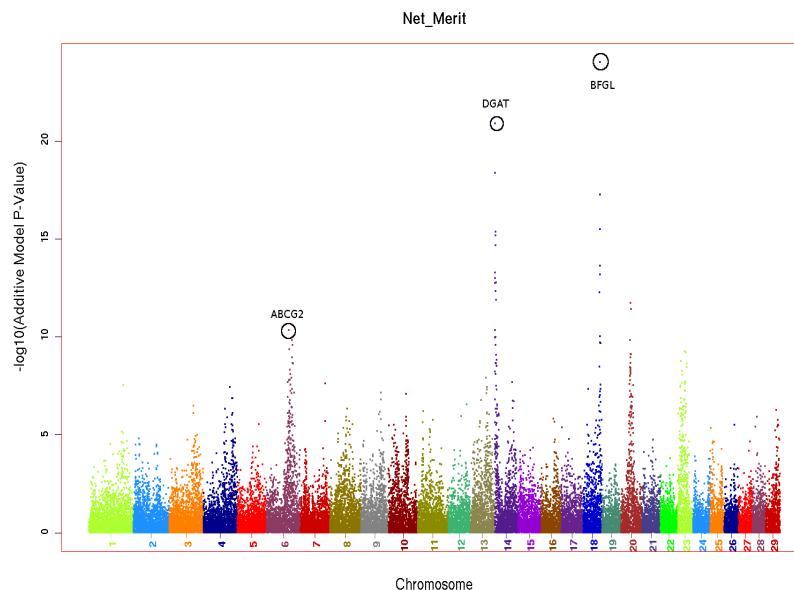
*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Core methods for exploring trait architecture and selection

1. Genome-Wide Association Studies (GWAS)

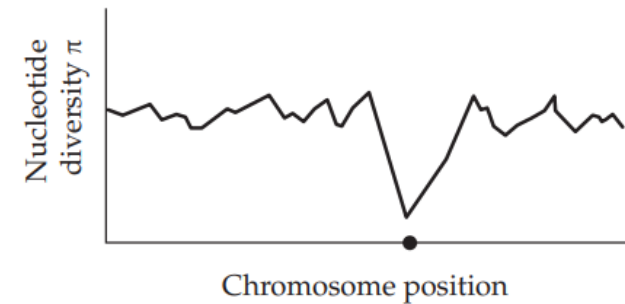
- Statistical method that associates phenotypic traits with SNPs across genome
- Identifies QTLs influencing traits (e.g. milk yield, fertility)
- Supports development of better MAS and GS models
- Findings often lead to discovery of candidate genes with functional relevance



Core methods for exploring trait architecture and selection

2. Selection Signals

- Genomic patterns indicating past selection events
- Identify loci under natural and/or artificial selection
- Reveal adaptation and breeding history
- Useful in conservation and breed characterization



Modern Livestock Genomics – From Data to Application



Big data challenge

- Molecular datasets (e.g. SNP arrays, WGS) contain a high number of variables per sample
- Complex trait analyses require advanced bioinformatics approaches



End of the Genomics-centric era?

- Genomics is reaching maturity; SNP arrays and WGS are widely adopted
- Yet, added value now comes from integration with other omics (transcriptomics, epigenomics)



Rise of the Phenomics era

- High-throughput phenotyping (e.g. sensors, imaging, metabolomics) enables large-scale, accurate phenotype collection
- Accurate phenotypes = improved selection accuracy

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Case study



Identification of Selection Signals on the X-Chromosome in East Adriatic Sheep: A New Complementary Approach

Mario Shihabi^{1*}, Boris Lukic², Vlatka Cubric-Curik¹, Vladimir Brajkovic¹, Milan Oršanić³, Damir Ugarković³, Luboš Vostry⁴ and Ino Curik^{1*}

¹Department of Animal Science, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia, ²Department for Animal Production and Biotechnology, Faculty of Agrobiotechnical Sciences Osijek, J.J. Strossmayer University of Osijek, Osijek, Croatia,

³Department of Forest Ecology and Silviculture, Faculty of Forestry and Wood Technology, University of Zagreb, Zagreb, Croatia,

⁴Department of Genetics and Breeding, Faculty Agrobiology, Food and Natural Resources, Czech University of Life Sciences, Prague, Czechia

<https://doi.org/10.3389/fgene.2022.887582>

Summer
School

**Bioinformatics
approaches in
animal
breeding**



Identification of Selection Signals on the X-chromosome in East Adriatic Sheep: A New Complementary Approach

M. Shihabi, B. Lukic, V. Cubric-Curik, V. Brajkovic, M. Oršanić, D. Ugarković, L. Vostry, I. Curik

✉ mshihabi@agr.hr; icurik@agr.hr

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

East Adriatic sheep

- One of the most important livestock species in Croatia
- Mainly in the coastal and mountainous regions (mostly kept extensively)
- Eight indigenous breeds (Cres Island Sheep, Dalmatian Pramenka, Dubrovnik Ruda, Istrian Sheep, Krk Island Sheep, Lika Pramenka, Pag Island Sheep and Rab Island Sheep)

Intensive artificial
selection has never
been practiced

Genomic composition → environmental adaptation + sustainable production



Summer
School

**Bioinformatics
approaches in
animal
breeding**

The origin of European sheep

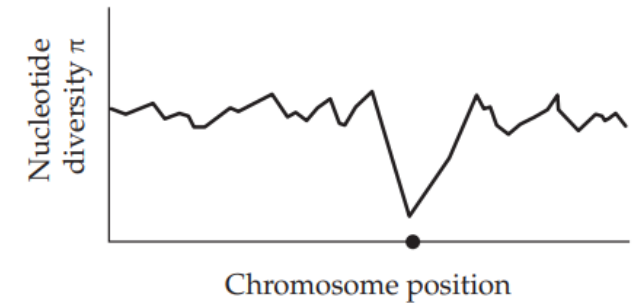
Ciani *et al.* (2020)



Balkan sheep breeds form a genetically specific cluster that is distinct from other European sheep breeds

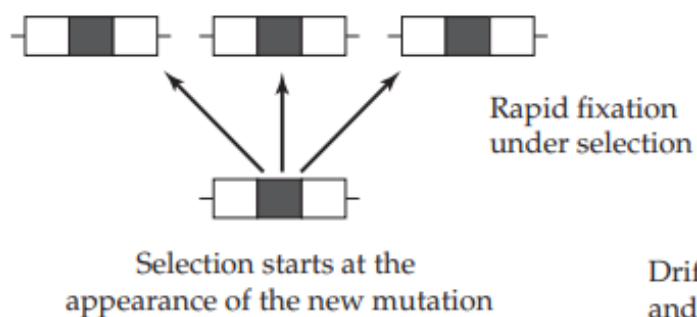
Bioinformatics
approaches in
animal
breeding

Positive selection signals

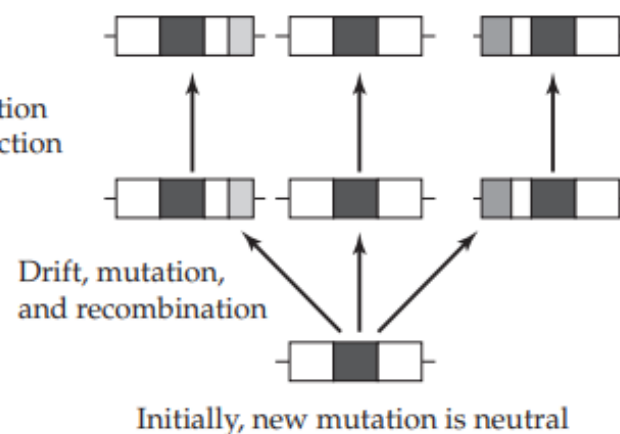


- ❖ Increase of the frequency of certain alleles depending on the selection pressure
- ❖ Regions with increased homozygosity (hitchhiking)
 - a) Hard sweep - Emerging desirable mutation
 - b) Soft sweep - Existing allele (previously neutral)
 - Multiple mutations whose frequency gradually increases

(A) Hard sweep



(B) Soft sweep



X-Chromosome

♀ XX

♂ XY

- ❖ 135 Mb; ~5 % genome size
- ❖ Pseudo-Autosomal Region (PAR)

Chr X vs autosome

- ↓ Mutation rate (0.015 mutations/Mb/generation)
- ↓ Recombination rate (2/3)
- ↓ Effective population size (3/4)
- ↑ Linkage disequilibrium

Understudied in the analyses identifying positive selection signals

- ❖ Better understanding of selection behaviour on the sex chromosome
- ❖ Good potential for methodological improvements

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Main objective

to identify signals of positive selection on the X chromosome in East Adriatic sheep

Inter-population analyses

Intra-population analyses

1. **extreme Runs Of Homozygosity islands (eROHi)**
2. **integrated Haplotype Score (iHS)**
3. **number of Segregating Sites by Length (nSL)**
4. **Haplotype Richness Drop (HRiD) → NEW APPROACH!**

Gene Annotation



Summer School

Bioinformatics approaches in animal breeding

Data

❖ East Adriatic sheep metapopulation (202 individuals) + 10 mouflons

100(101) ♂ ; 101 ♀ 5 ♂ ; 5 ♀

Ancestral information



❖ Ovine Infinium® HD SNP BeadChip 600K



Krk Island Sheep
n = 20



Pag Island Sheep
n = 45



Cres Island Sheep
n = 20



Rab Island Sheep
n = 20



Istrian sheep
n = 25



Lika Pramenka
n = 20



Dalmatian Pramenka
n = 25(26)



Dubrovnik Ruda
n = 26

Quality control:

- chrX SNPs only
- GenTrain score <0.4
- GenCall score <=0.8
- Genotype call rate <0.9
 - HWE 0.0000001
 - Duplicates
- Individual call rate <0.95
- 27 questionable SNPs

18 983 SNPs

201 individuals

10 mouflons

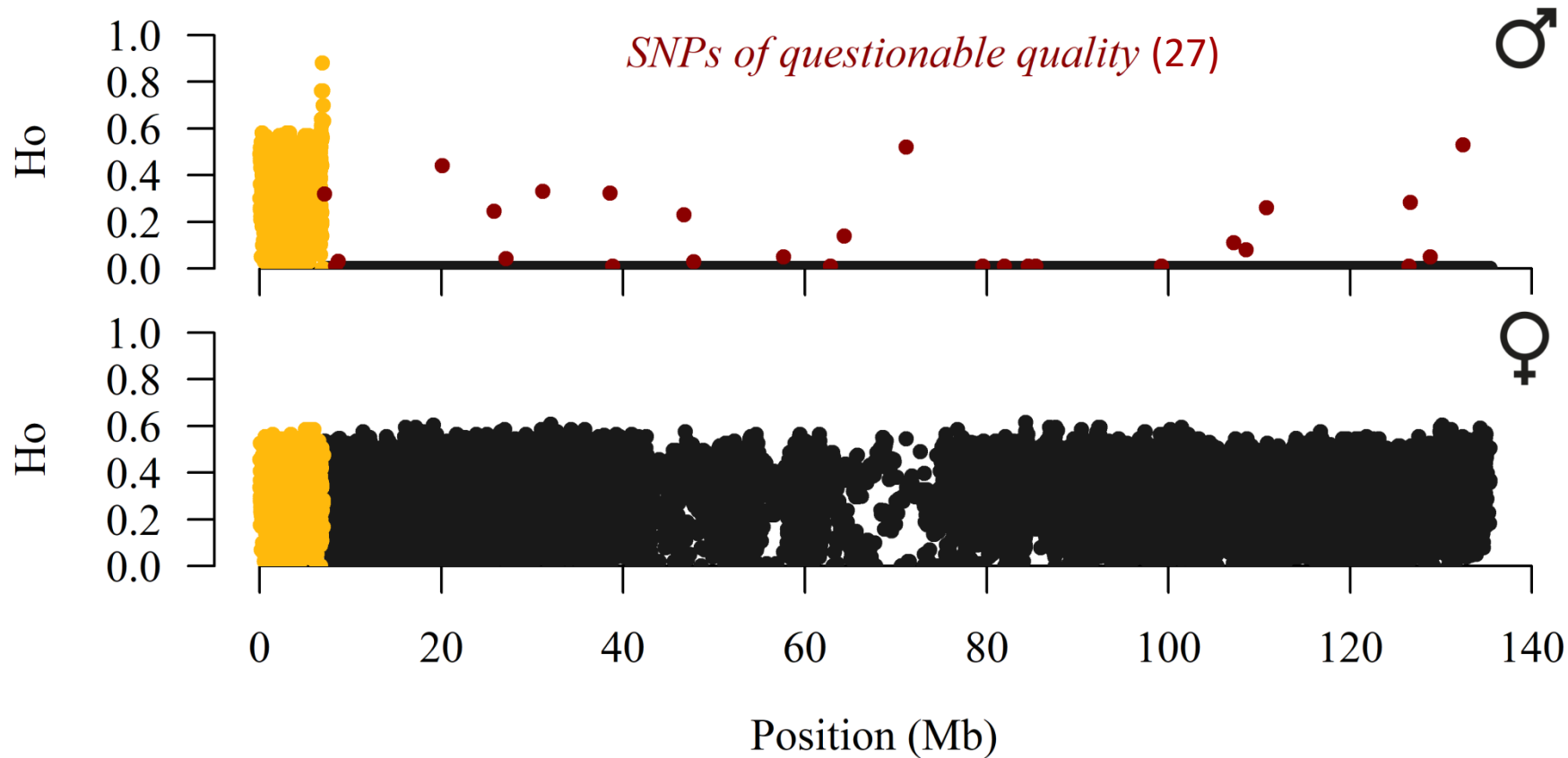
Max gap = 232 kb

Mean gap = 7.13 kb

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**

Pseudo-Autosomal Region (PAR)



0 – 7.04 Mb (1232 SNPs)

Summer
School

Bioinformatics
approaches in
animal
breeding

extreme ROH islands (eROHi)

♀ XX → 101 individuals

GOLDEN HELIX
Accelerating the Quest for Significance™

- Min SNP = 15
- Max gap = 250 kb
- Min density = 1 SNP / 20 kb

six different length classes:

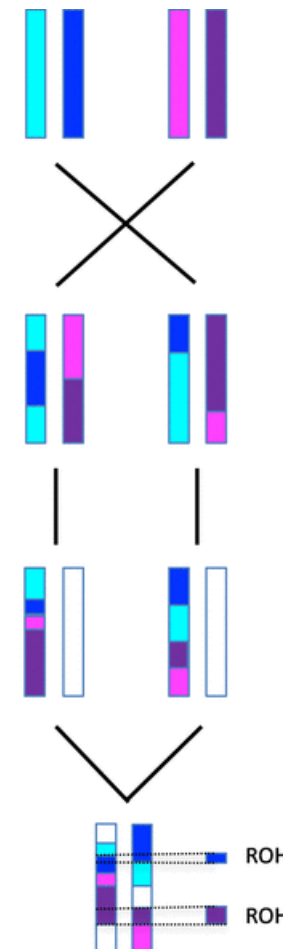
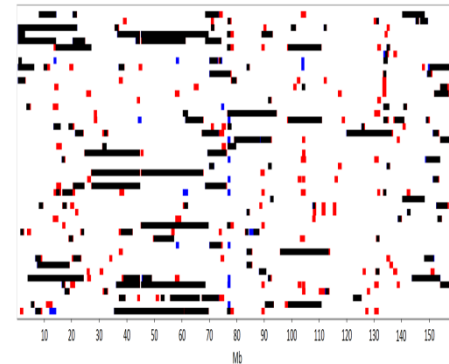
- **0.25 – 1 Mb**
(no het or missing)
- 1 – 2 Mb
- 2 – 4 Mb
- 4 – 8 Mb
- 8 – 16 Mb
- >16 Mb

het and missing according to
Ferenčaković *et al.* (2013)

❖ ROH frequency for each SNP

Consecutive SNPs with $-\log(P) > 3.3$

One-sided test
(40 individuals)



Summer
School

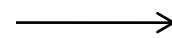
**Bioinformatics
approaches in
animal
breeding**

integrated Haplotype Score (iHS) number of Segregating Sites by Length (nSL)

SHAPEIT



--chrX Option



201 individuals



- ❖ The VCF file recoded - ancestral allele (mouflons) = reference
- derived allele = alternative

iHS

- ❖ rehh R package

nSL

- ❖ selscan software

- ❖ Normalisation inside frequency bin size of 0.025
- ❖ Sliding window approach (500 kb size; 100 kb slide); SNPs with $-\log(P) > 2$ = outliers

Two-sided
test

nonoverlapping windows with >10% outliers

Summer
School

Bioinformatics
approaches in
animal
breeding

Haplotype Richness Drop (HRiD); new approach

♂ XY → 100 individuals


- ❖ On the X-chromosome without PAR, male genotypes are hemizygous, making it easy to derive exact haplotypes of different lengths

the effective number of haplotypes of the i th sliding window under study

$$HRiD_{w_{i+1}} = \frac{n_{hw_i} + n_{hw_{i+2}}}{2n_{hw_{i+1}}}$$

(formula is modified for the first and the last window)

- ❖ If there is no selection, HRiD values should fluctuate around the value of one 

- ❖ The presence of positive selection leads to a sudden decrease in the effective number of haplotypes (high positive HRiD values) 

- ❖ Lower sensitivity to variation in recombination rate

Summer
School

Bioinformatics
approaches in
animal
breeding

Haplotype Richness Drop (HRiD); new approach

- ❖ Sliding window approach (size of 70 SNPs \approx 500 kb; slide of 35 SNPs \approx 250 kb)

windows with $-\log(P) > 3.3$

One-sided test

(HRiD = 2.8)

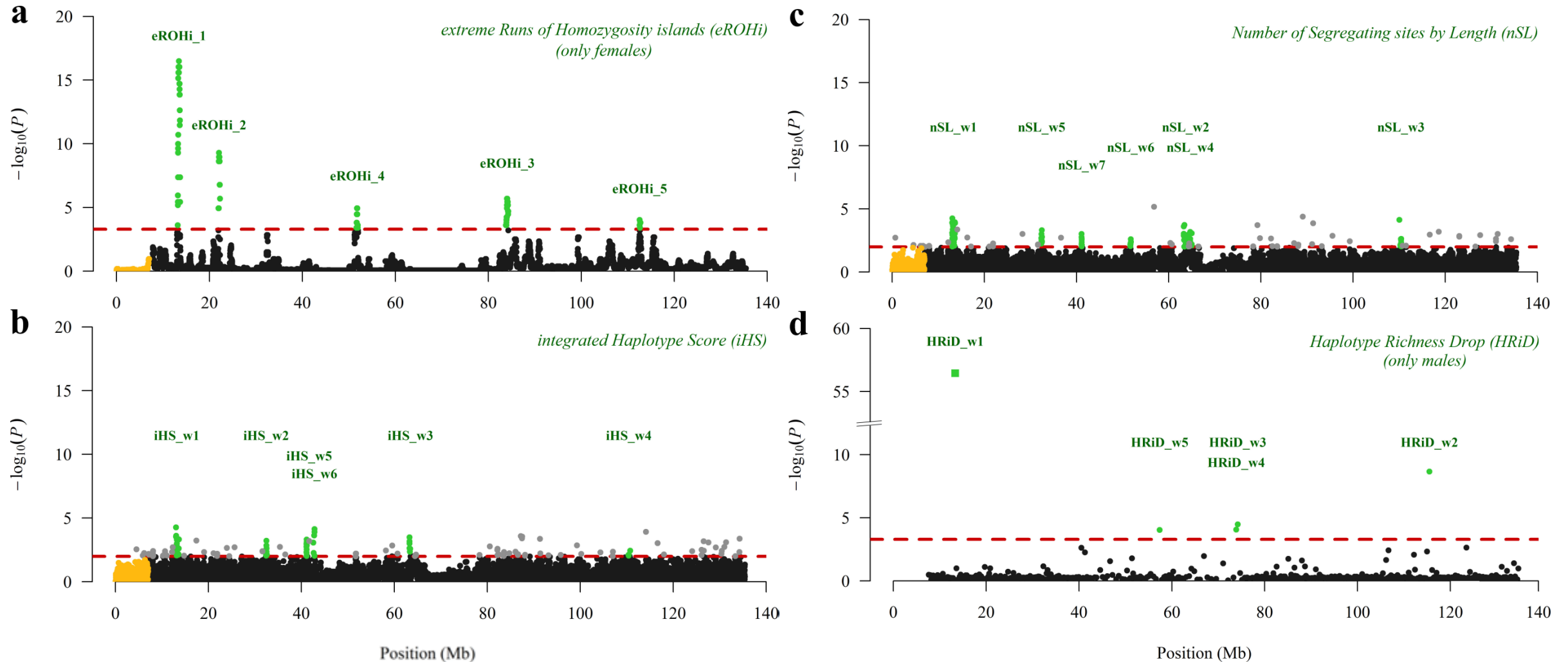
- ❖ **Median-joining network (MJN)** = to illustrate the phylogenetic relationship between ancestral and derived haplotypes in the each selection signal obtained

ancestral haplotypes = mouflons

Summer
School

Bioinformatics
approaches in
animal
breeding

Visualisation of positive selection signals in the Manhattan plot



Description of mapping statistics and annotation of genes in selection signals by three classical (eROHi, iHS and nSL) approaches

Signal name	Position (Mb)	SNPs	-log(P)	Candidate genes under selection
eROHi_1	13.17-13.69	59/59	16.5	<u>CA5B</u>, <u>ZRSR2</u>, <u>AP1S2</u>, <u>GRPR</u>
eROHi_2	21.96-22.26	35/35	9.3	<i>POLA1</i> , <i>ARX</i>
eROHi_3	83.78-84.28	73/73	5.7	No annotated genes found
eROHi_4	51.63-51.94	33/33	4.9	<i>DGKK</i>, <i>CCNB3</i>
eROHi_5	112.53-112.72	11/11	4.0	<i>PLS3</i>
iHS_w1	13.10-13.60	17/50	4.3	<i>TMEM27</i>, <i>CDC42</i>, <u>CA5B</u>, <u>ZRSR2</u>, <u>AP1S2</u>, <u>GRPR</u>
iHS_w2	32.20-32.70	13/55	3.2	No annotated genes found
iHS_w3	63.20-63.70	6/35	3.5	<i>RLIM</i> , <i>KIAA2022</i>, <i>ABCB7</i>
iHS_w4	110.30-110.80	5/36	2.4	<i>DOCK11</i>, <i>WDR44</i>, <i>KLHL13</i>
iHS_w5	41.00-41.50	9/65	3.3	<i>NDP</i>, <i>EFHC2</i>
iHS_w6	42.50-43.00	6/60	4.1	<i>MIR221</i>
nSL_w1	13.10-13.60	35/50	4.3	<i>TMEM27</i>, <i>CDC42</i>, <u>CA5B</u>, <u>ZRSR2</u>, <u>AP1S2</u>, <u>GRPR</u>
nSL_w2	63.30-64.30	19/44	3.7	<i>KIAA2022</i>, <i>ABCB7</i>, <i>UPRT</i>, <i>ZDHHC15</i>, <i>MAGEE2</i>
nSL_w3	110.10-110.60	12/39	4.1	<i>DOCK11</i>
nSL_w4	64.60-65.10	10/39	3.2	<i>MAGT1</i> , <i>ATRX</i> , <i>FGF16</i>
nSL_w5	32.30-32.80	14/61	3.3	No annotated genes found
nSL_w6	51.40-51.90	8/56	2.6	<i>SHROOM4</i> , <i>DGKK</i>, <i>CCNB3</i>
nSL_w7	41.00-41.50	9/65	3.0	<i>NDP</i>, <i>EFHC2</i>

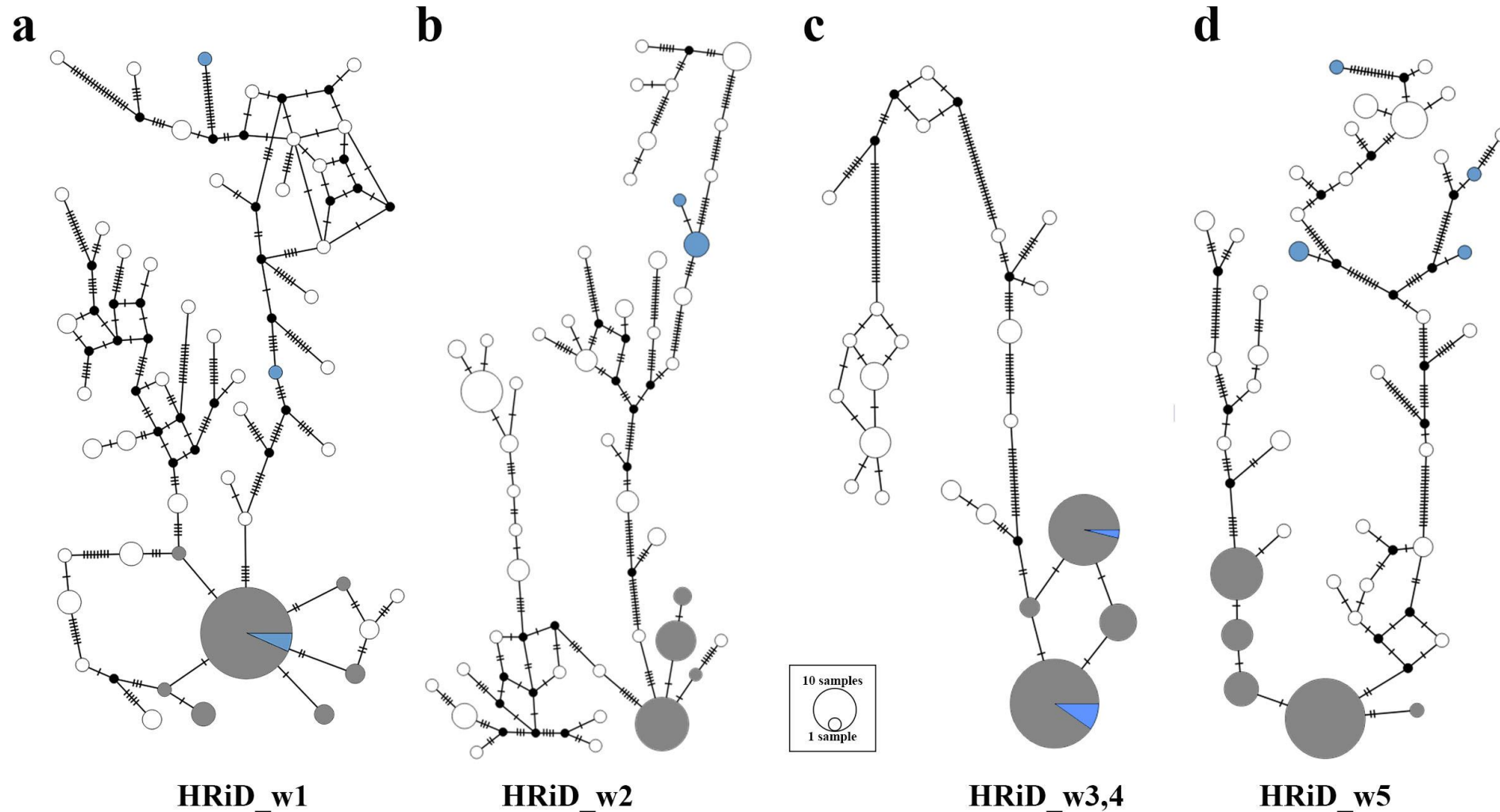
Summer
School

**Bioinformatics
approaches in
animal
breeding**

Description of mapping statistics and annotation of genes in selection signals by new HRiD approach

Signal name	Position (Mb)	n_a	n_h	HRiD	$-\log(P)$	Candidate genes under selection
HRiD_w1	13.04-13.62	42	5.4	9.6	56.5	<i>TMEM27, CDC42, <u>CA5B</u>, <u>ZRSR2</u>, <u>AP1S2</u>, <u>GRPR</u></i>
HRiD_w2	115.30-115.73	36	13.3	4.2	8.7	<i>AMOT, LHFPL1</i>
HRiD_w3	73.90-74.54	13	4.3	3.2	4.5	<i>DACH2</i>
HRiD_w4	73.57-74.20	10	1.9	3.1	4.1	<i>CHM, DACH2</i>
HRiD_w5	56.64-58.09	33	6.9	3.1	4.0	<i>AR, OPHN1, YIPF6</i>

Median-joining network showing the phylogenetic relationship between ancestral and derived haplotypes inside selection signals identified by HRiD



Conclusions

- ❖ 14 positive selection signals with a total of 34 annotated genes were identified
- ❖ Our results show that **HRiD** offers an interesting possibility to be used **complementary** to the eROHi, iHS and nSL approaches **or when only males are genotyped**, which is often the case in livestock where genomic breeding value estimates are routinely performed for males
- ❖ Furthermore, we have shown that **phylogenetic analyses** can provide additional information when performed within the selection signals identified by HRiD, particularly with respect to the **ancestral or derived status** of the advantageous selected haplotypes
- ❖ Overall, our results highlight the importance of the X-chromosome in the adaptive architecture of domestic ruminants, while our novel HRiD approach opens new avenues for research

*Summer
School*

**Bioinformatics
approaches in
animal
breeding**



Thank you for your attention!

Croatian Science Foundation grant ANAGRAMS- IP-2018-01-7317 („Application of NGS in the Assessment of Genomic vaRiAbility in ruMinantS”).