Detecting phenotype-associated genes by combining gene expression and GWAS summary statistics

Emily Clark
James Prendergast
Siddharth Jayaraman
(The Roslin Institute, University of Edinburgh)

BovReg Final Conference - Brussels
(14-15 February 2024)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 815668

BovReg WP4 - **Introduction**

Understanding Traits – Beyond Genome-wide association studies (**GWAS**)

GWAS links genetic regions to traits, but pinpointing exact causal genes is tricky. TWAS bridges this gap, helping uncover how genes contribute to diseases and characteristics.

Why Transcriptome-Wide Association Studies (TWAS)? A Powerful Tool for Genetic Research:

- Uncovers functionally relevant genes linked to traits.
- Offers tissue-specific insights into biological mechanisms.
- Analyses large datasets efficiently by predicting gene expression.
- Complements GWAS for deeper biological understanding.

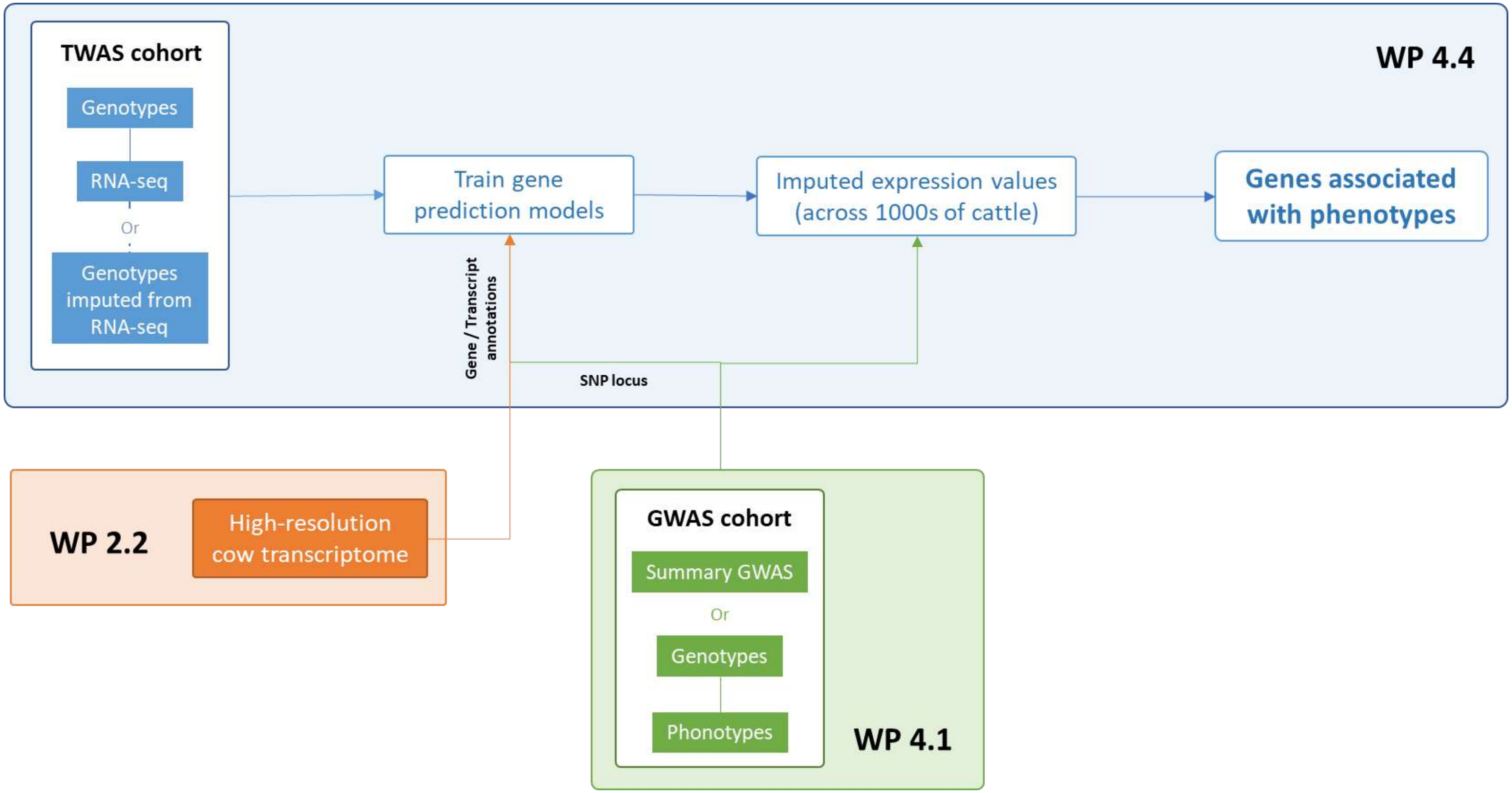
How TWAS Works

- Combines genetic data (GWAS) with gene expression data (eQTLs).
- Builds models to see how genetic variations influence gene activity.
- Predicts which genes are likely involved in the studied trait.

BovReg WP4 - Transcriptome-wide association studies (TWAS)

Integrating genome-wide association studies (GWAS) and gene expression datasets to identify **gene-trait associations**

- A key component of this approach is to build a model to impute gene expression levels from genotypes by using samples with matched genotypes and gene expression data in each tissue or cell type.



WP 4.4

TWAS cohort

Genotypes

RNA-seq

Or

Genotypes imputed from RNA-seq

Train gene prediction models

Imputed expression values (across 1000s of cattle)

Genes associated with phenotypes

Gene / Transcript annotations

SNP locus

WP 2.2

High-resolution cow transcriptome

GWAS cohort

Summary GWAS

Or

Genotypes

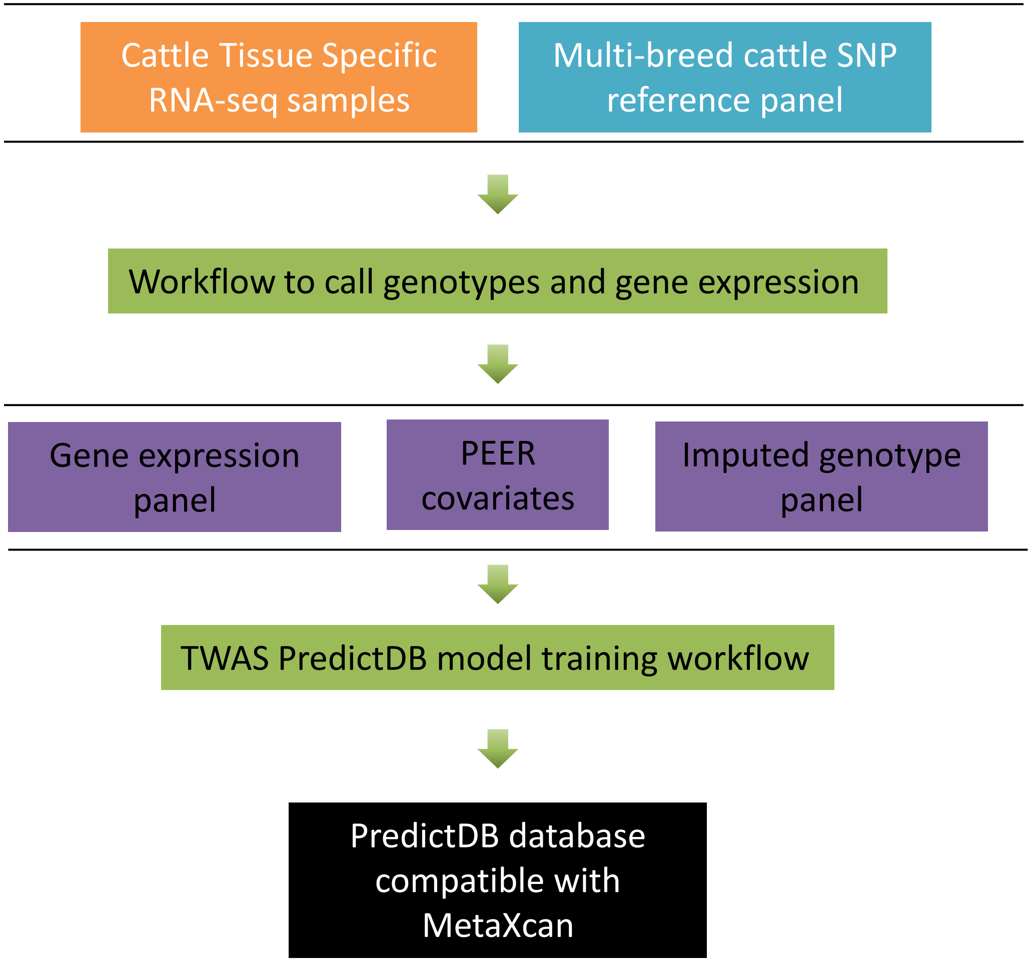
Phonotypes

WP 4.1

BovReg WP4 - Nextflow workflow for TWAS

Integrating genome-wide association studies (GWAS) and gene expression datasets to identify **gene-trait associations**

- We developed a Nextflow workflow that spans the whole analyses from raw fastq data download, through read mapping, variant calling, expression quantification, TWAS model generation and finally TWAS analyses.



Cattle Tissue Specific RNA-seq samples

Multi-breed cattle SNP reference panel

Workflow to call genotypes and gene expression

Gene expression panel

PEER covariates

Imputed genotype panel

TWAS PredictDB model training workflow

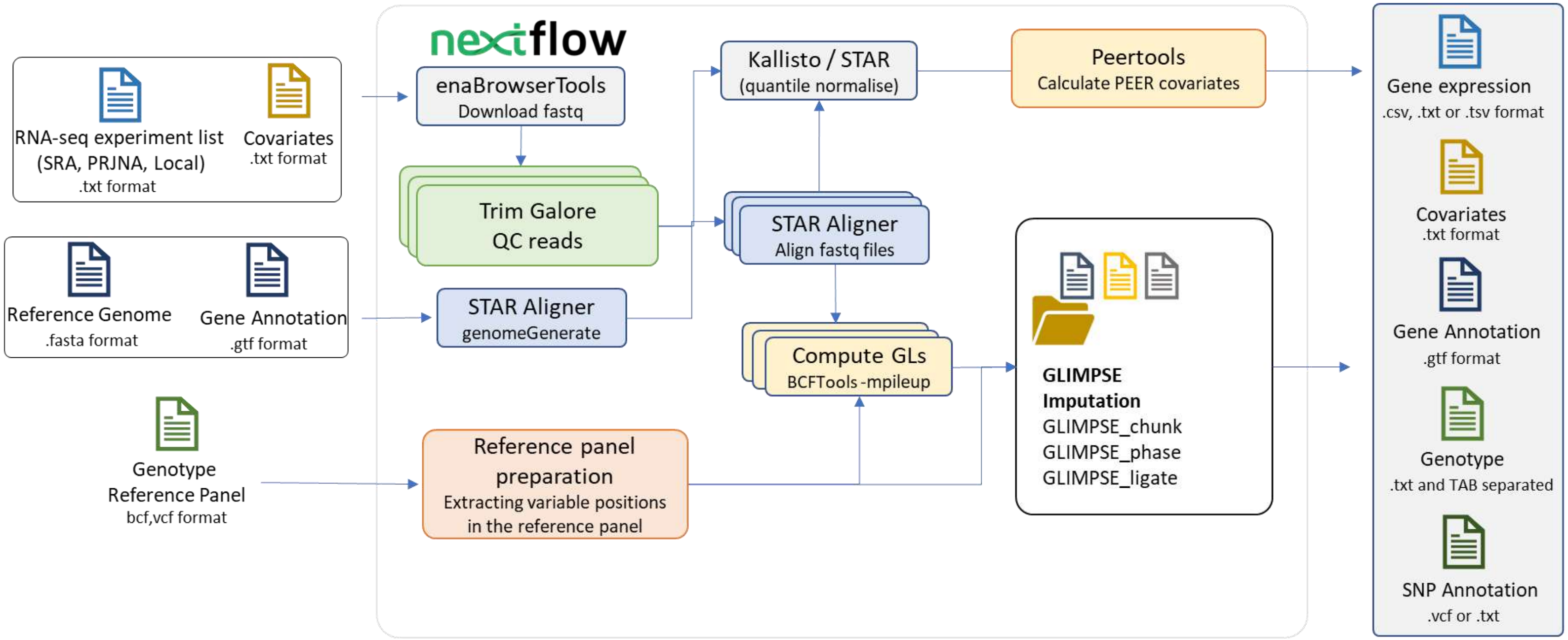
PredictDB database compatible with MetaXcan

BovReg WP4 - Nextflow workflow for TWAS

Integrating genome-wide association studies (GWAS) and gene expression datasets to identify **gene-trait associations**

- We developed a Nextflow workflow that spans the whole analyses from raw fastq data download, through read mapping, variant calling, expression quantification, TWAS model generation and finally TWAS analyses.

Workflow to call genotypes and gene expression



RNA-seq experiment list
(SRA, PRJNA, Local)
.txt format

Covariates
.txt format

Reference Genome
.fasta format

Gene Annotation
.gtf format

Genotype
Reference Panel
bcf,vcf format

nextflow

enaBrowserTools
Download fastq

Trim Galore
QC reads

STAR Aligner
genomeGenerate

Reference panel preparation
Extracting variable positions in the reference panel

Kallisto / STAR
(quantile normalise)

STAR Aligner
Align fastq files

Compute GLs
BCFTools -mpileup

Peertools
Calculate PEER covariates

GLIMPSE Imputation
GLIMPSE_chunk
GLIMPSE_phase
GLIMPSE_ligate

Gene expression
.csv, .txt or .tsv format

Covariates
.txt format

Gene Annotation
.gtf format

Genotype
.txt and TAB separated

SNP Annotation
.vcf or .txt

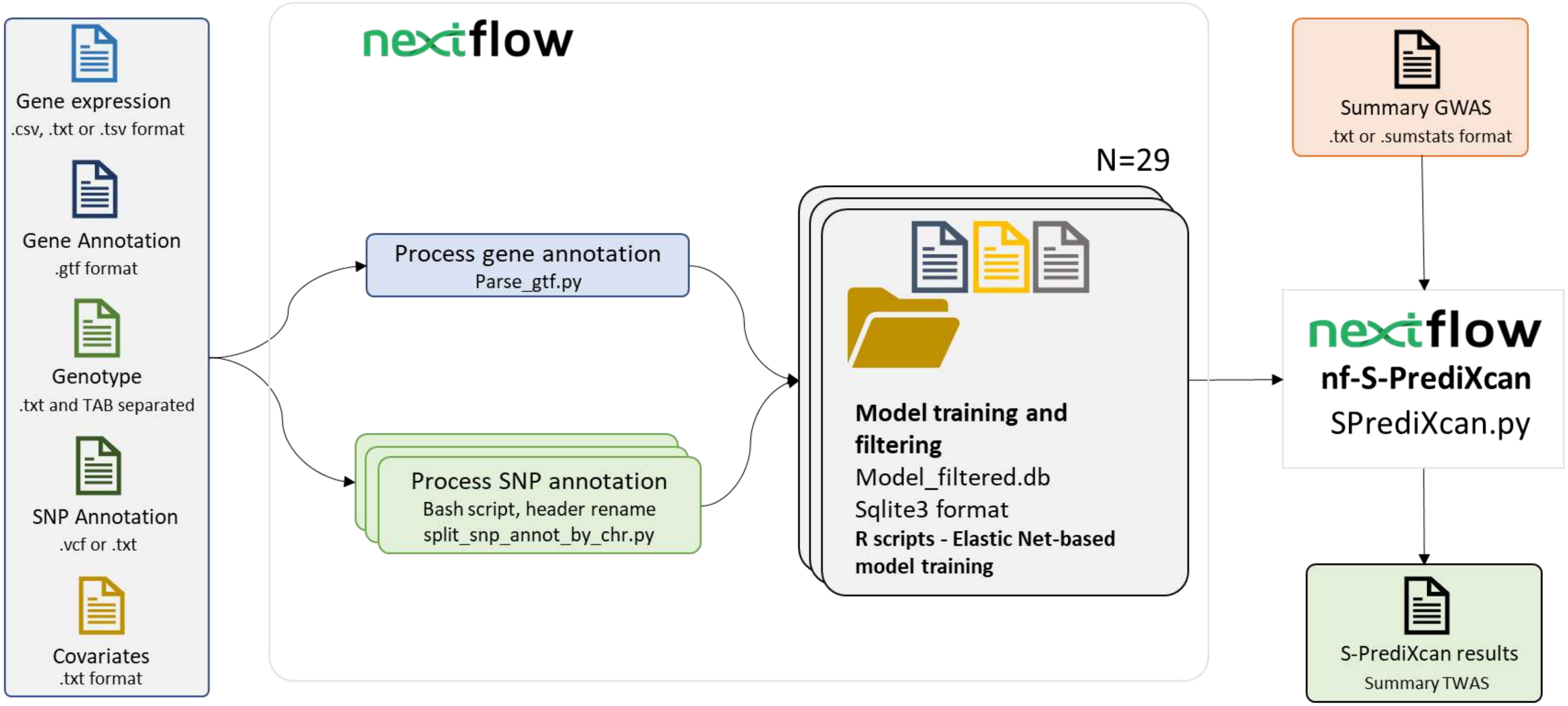
BovReg WP4 - Nextflow workflow for TWAS

Integrating genome-wide association studies (GWAS) and gene expression datasets to identify **gene-trait associations**

- We developed a Nextflow workflow that spans the whole analyses from raw fastq data download, through read mapping, variant calling, expression quantification, TWAS model generation and finally TWAS analyses.

TWAS PredictDB model training workflow

MetaXcan nextflow



Gene expression
.csv, .txt or .tsv format

Gene Annotation
.gtf format

Genotype
.txt and TAB separated

SNP Annotation
.vcf or .txt

Covariates
.txt format

nextflow

Process gene annotation
Parse_gtf.py

Process SNP annotation
Bash script, header rename
split_snp_annot_by_chr.py

N=29

Model training and filtering
Model_filtered.db
Sqlite3 format
R scripts - Elastic Net-based model training

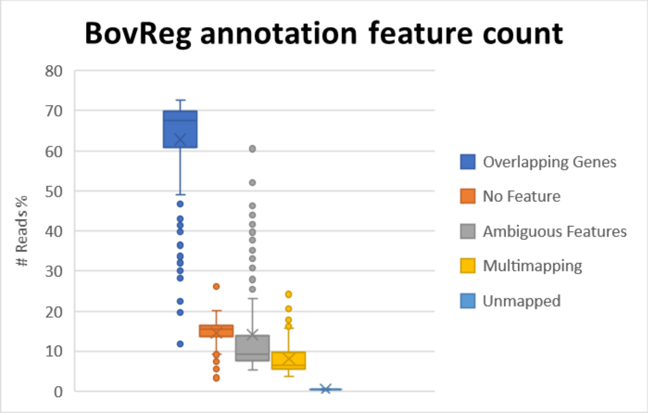
Summary GWAS
.txt or .sumstats format

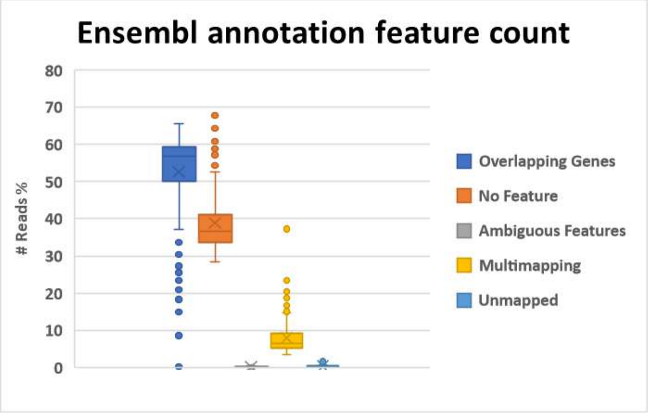
nextflow
nf-S-PrediXcan
SPrediXcan.py

S-PrediXcan results
Summary TWAS

BovReg WP4 - Processing RNA-seq using an improved transcriptome annotation

For TWAS model training we collected RNA-seq data for five distinct, trait-relevant tissues: muscle, rumen, mammary gland, liver and jejunum; size of these RNA-seq cohorts ranged from 48 to 188 samples which minimised batch effects.

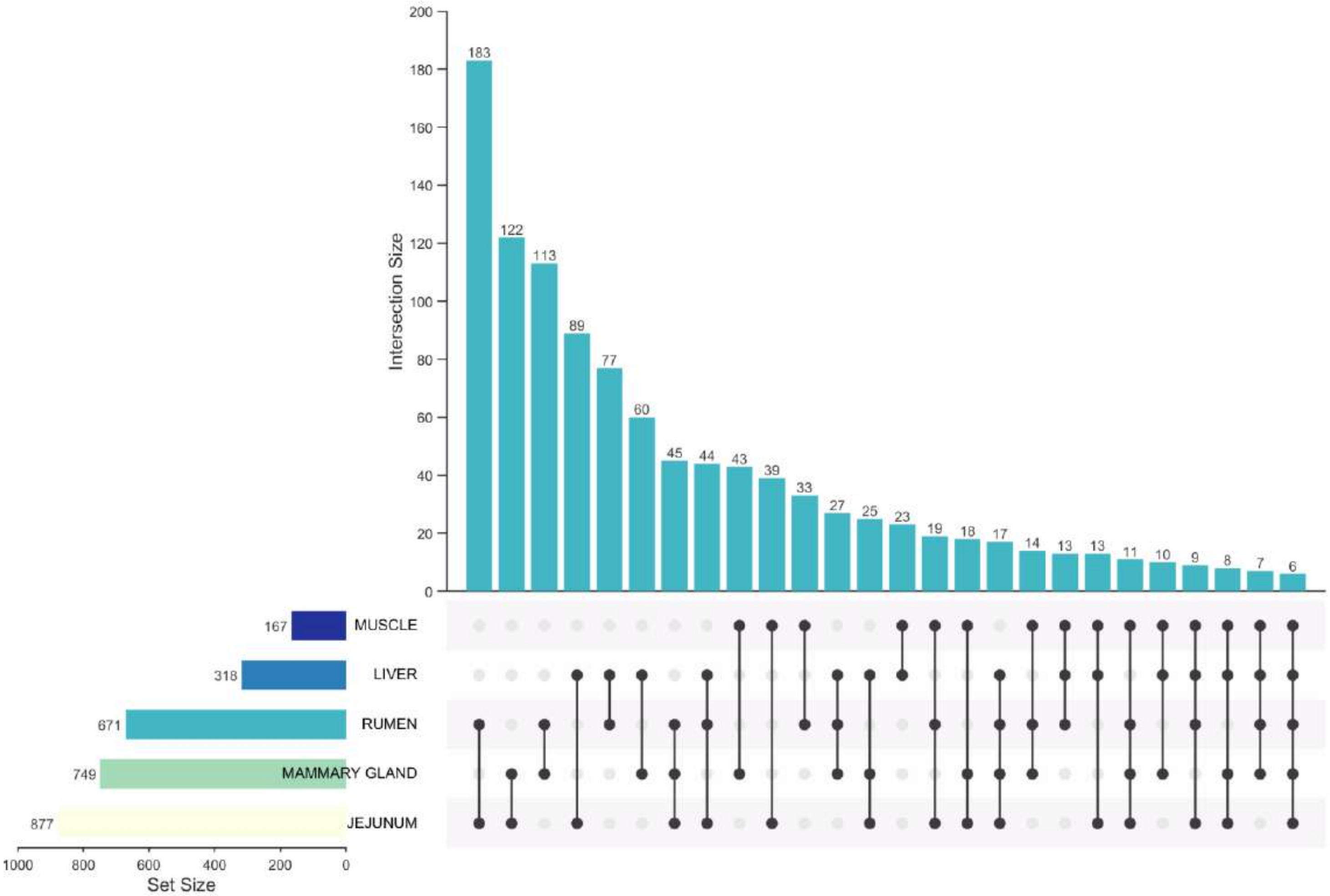

BovReg annotation feature count


Ensembl annotation feature count

BovReg WP4 - TWAS model training

- Using GLIMPSE derived genotypes and our workflow we trained PredictDB TWAS models for predicting gene and transcript levels from cis-genotypes

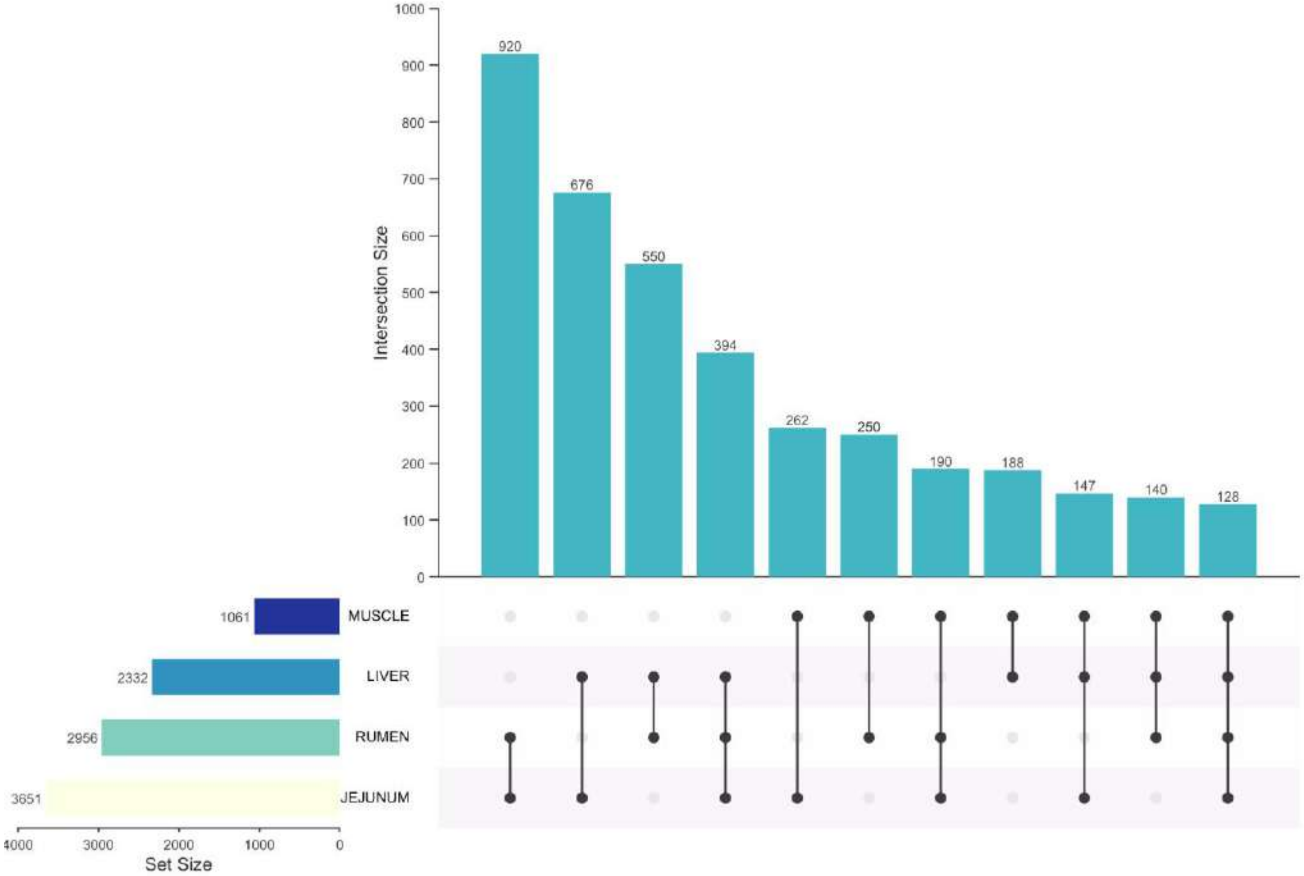
Gene model counts and tissue overlap for PredictDB trained models (zscore pval <0.05 and rho avg.>0.1) with SNPs filtered on ETH GWAS study



BovReg WP4 - TWAS model training

- Using GLIMPSE derived genotypes and our workflow we trained PredictDB TWAS models for predicting gene and transcript levels from cis-genotypes

Transcript model counts and tissue overlap for PredictDB trained models (zscore pval <0.05 and rho avg.>0.1) with SNPs filtered on ETH GWAS study



BovReg WP4 - Identification of novel trait-gene associations linked to important cattle phenotypes

To demonstrate the utility of these TWAS models at providing novel insights into the genes potentially underlying cattle phenotypes we collated summary statistics from four GWAS cohorts that together had tested genetic associations with following phenotypes:

USDA GWAS traits (35 Dairy and Beef traits; Holstein bulls, ~25K cohort size)

1. Milk yield
2. Fat yield
3. Protein yield
4. Fat percentage
5. Protein percentage
6. Net merit
7. Productive life
8. Somatic cell score
9. Age at first calving
10. Days to first breeding
11. Daughter pregnancy rate
12. Heifer conception rate
13. Cow conception rate
14. Sire calving ease
15. Daughter calving ease
16. Sire stillbirth
17. Daughter stillbirth
18. Final score
19. Stature
20. Strength
21. Dairy form
22. Foot angle
23. Rear legs (side view)
24. Body depth
25. Rump angle
26. Rump width
27. Fore udder attachment
28. Rear udder height
29. Udder depth
30. Udder cleft
31. Front teat placement
32. Teat length
33. Rear legs (rear view)
34. Feet and legs composite
35. Rear teat placement

ETH GWAS traits (11 Dairy traits; Brown Swiss cattle, ~3-12K cohort size)

1. Calving Ease direct
2. Calving Ease Maternal
3. Cell count
4. Fertility
5. Linear conformation trait
6. Milk production
7. Milking ability
8. Persistence of milk yield
9. Productive Lifetime
10. Slaughtering characteristics Adults
11. Slaughtering traits Calves

INRAE GWAS traits (5 Dairy traits; Hol, Mon and Nor breeds ~2.5-10K cohort size)

1. Fat Content
2. Fat Yield
3. Milk Yield
4. Protein Content
5. Protein Yield

FBN GWAS traits (4 Beef traits; Charolais x Holstein cross)

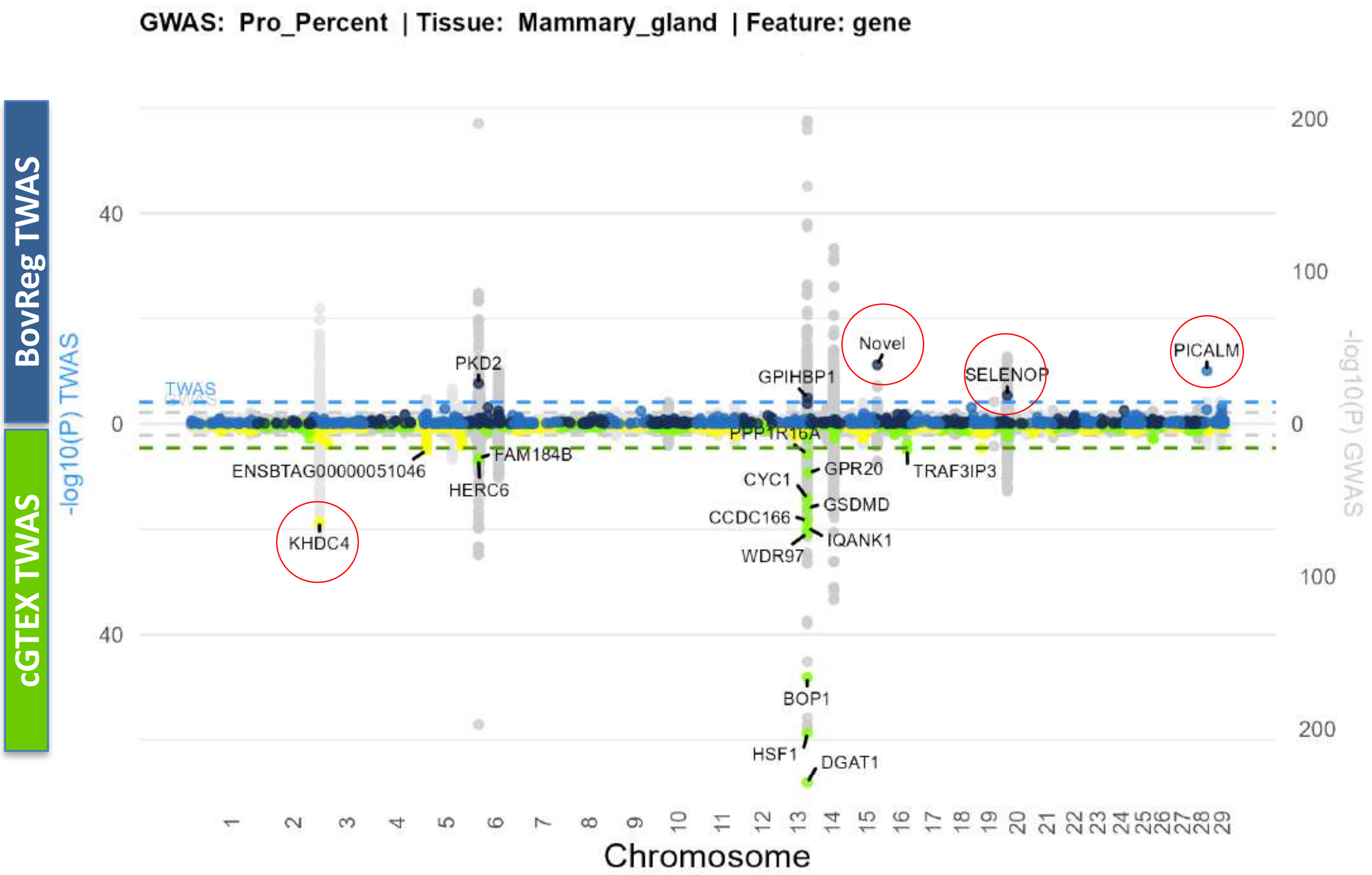
1. Birth Weight
2. Carcass Yield Bulls
3. Intramuscular lipid
4. Residual Feed Intake

BovReg **WP4 - Identification of novel trait-gene associations linked to important cattle phenotypes**

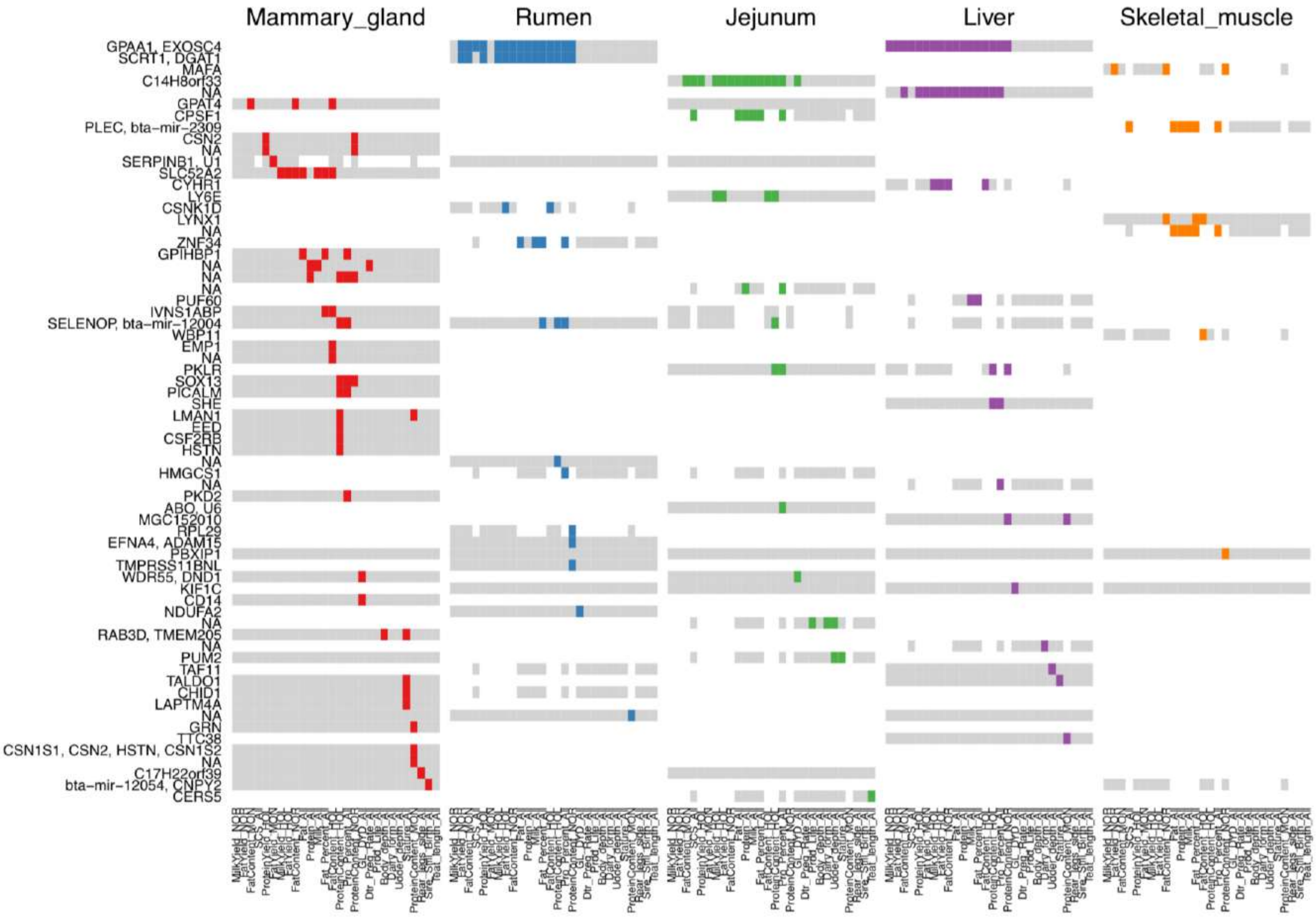
- The figure illustrates the results for milk protein percentage traits using the BovReg lactating mammary gland PredictDB models.

- We compared the results of our TWAS to those from the cattle GTEx cohort; despite overlaps these results are not redundant, with some GWAS peaks only being linked to a candidate gene in one or other dataset.

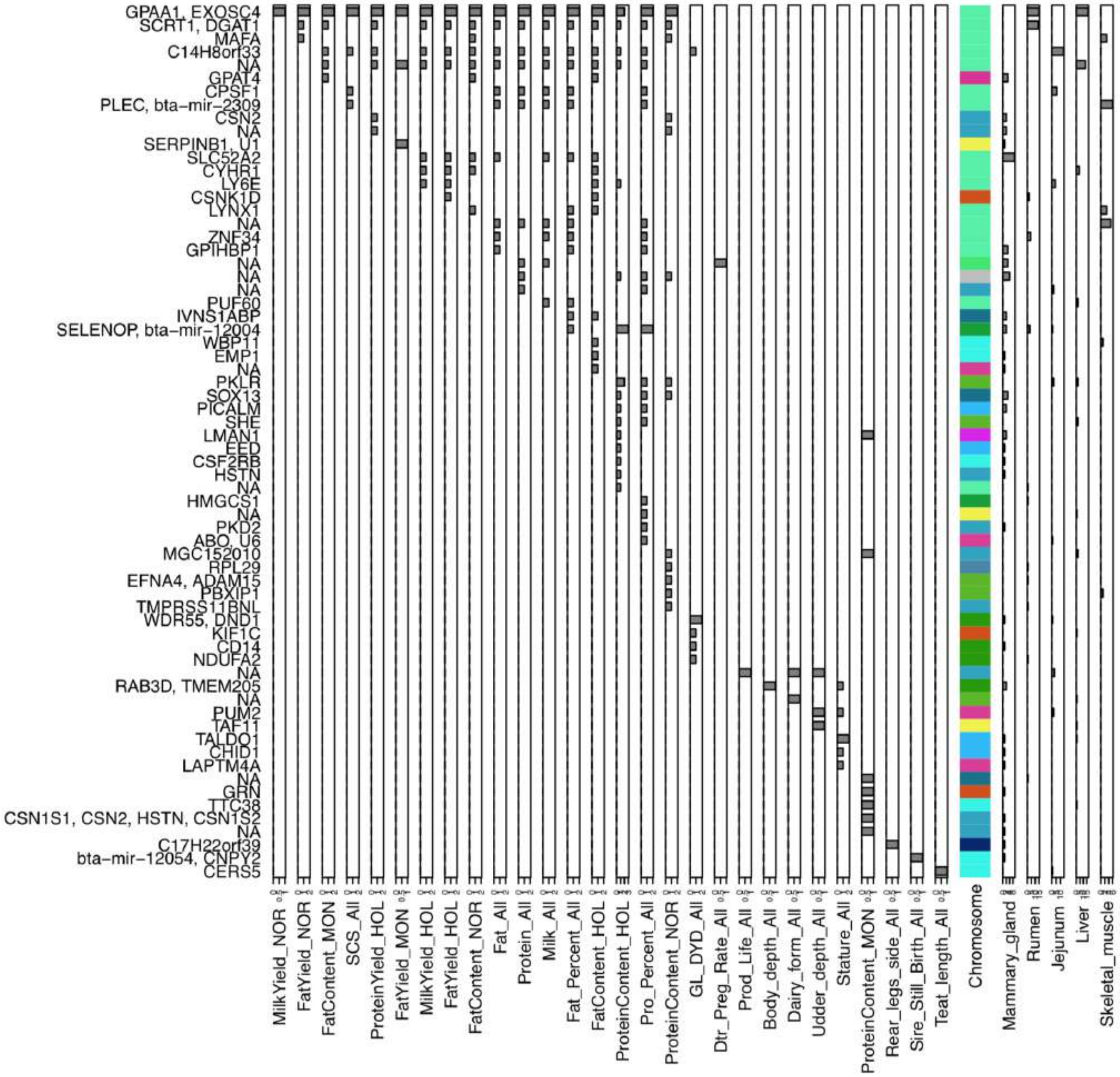
- GWAS in **grey**, cGTEX TWAS in **yellow/green** and BovReg TWAS in **blue**, <u>KHDC4</u> locus was only identified in the cattle GTEx dataset, whereas the <u>SELENOP</u> and <u>PICALM</u> genes was only identified in our analysis:



BovReg WP4 - Identification of novel trait-gene associations linked to important cattle phenotypes



BovReg WP4 - Identification of novel trait-gene associations linked to important cattle phenotypes



BovReg *PARTNERS*

AARHUS UNIVERSITY

CRG Centre for Genomic Regulation

diagenode Innovating Epigenetics Solutions

edinethics

EMBL-EBI

ETH zürich

FBN

Faculdade de Medicina Veterinária Universidade de Lisboa

Instytut Genetyki i Hodowli Zwierząt PAN

INRAE

LIÈGE université GIGA Medical Genomics

Université de Limoges

Luke Luonnonvarakeskus

MARTIN-LUTHER UNIVERSITÄT HALLE-WITTENBERG

ROSLIN

UNIVERSITY OF ALBERTA

The University of Edinburgh

Universiteit Utrecht

WAGENINGEN UNIVERSITY & RESEARCH

Thank you for your attention

www.bovregproject.eu

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 815668