



Towards an Integrated Functional Annotation of the Bovine Genome

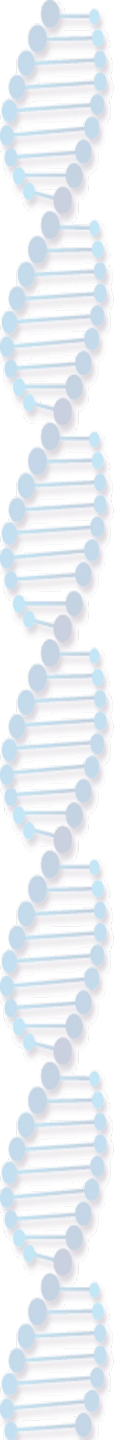
Emily Clark on behalf of Christa Kühn and the BovReg consortium



PAG 2024, Functional Annotation of Animal Genomes (FAANG) workshop
January 16th, 2024

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

Disclaimer: the sole responsibility of this presentation lies with the authors. The Research Executive Agency is not responsible for any use that may be made of the information contained therein.



Overview



- ▶ **Introduction to the BovReg Project**
- ▶ **Annotating the bovine genome using data for tissues from three populations of cattle**
- ▶ **Description of genomic structure and functional features in the five cell lines**
- ▶ **Reproducible, scalable and shareable analysis pipelines in nf-core**
- ▶ **Summary**



Introduction to the BovReg Project



An introduction to BovReg



- EU H2020 – funded consortium as contribution to global FAANG with focus on biology-informed breeding (Sep 2019 – Feb 2024)



Objectives:

- ▶ Establish **new laboratory and bioinformatics tools**
- ▶ **Annotate functionally active genomic regions** for tissues that are highly relevant for the BovReg target traits across ontogenetic stages, gender, metabolic breed types and environments
- ▶ **Map molecular and whole-animal phenotypes** (ranging from transcriptome to intermediate phenotype to performance/health) to newly annotated genomic features
- ▶ Develop **prototype models for integrating biological knowledge** on regulatory variation in genomic selection
- ▶ Provide **targeted training, dissemination and communication** activities and assess **social perception**

Aims:

- Annotation of the *Bos taurus* genome
- Key traits: robustness (fertility), health (mastitis susceptibility), biological efficiency (feed efficiency)



BovReg Consortium

- 20 (19) partners
- 13 countries & 2 international organisations



- Belgium (2)
- Denmark
- Finland
- France (2)
- Germany (2)
- Netherlands (2)
- Poland
- Portugal
- Spain
- Switzerland
- United Kingdom (2)
- Canada (Australia)
- EAAP
- EMBL-EBI



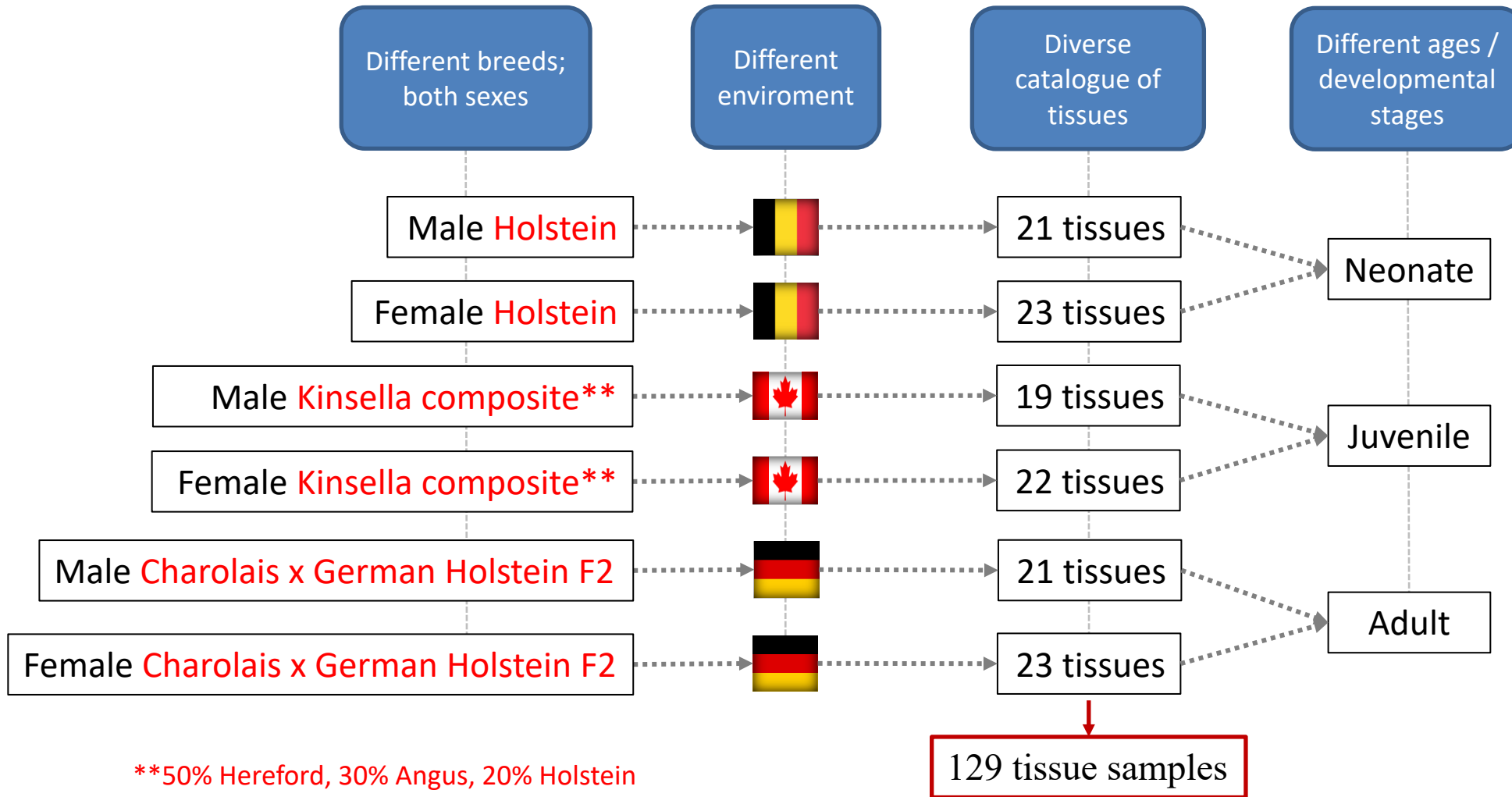
Annotating the bovine genome using data for tissues from three populations of cattle

Gabriel Costa Moreira, Sébastien Dupont and Carole Charlier, University of Liege





Diversity in sample selection for a comprehensive annotation





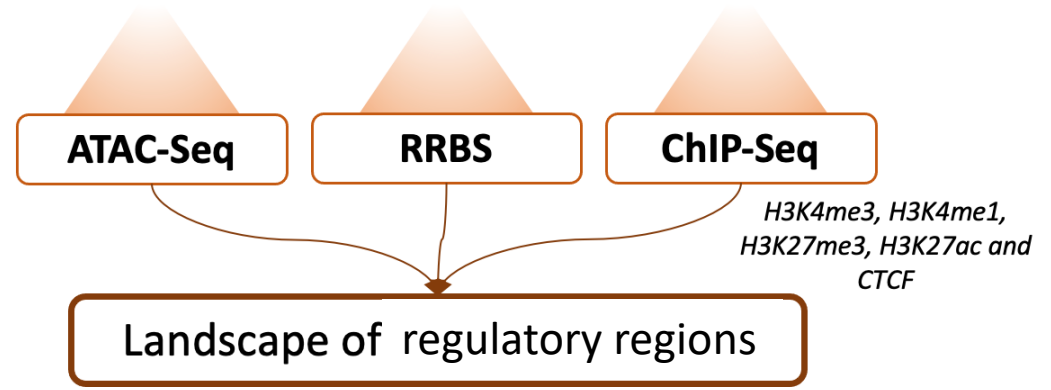
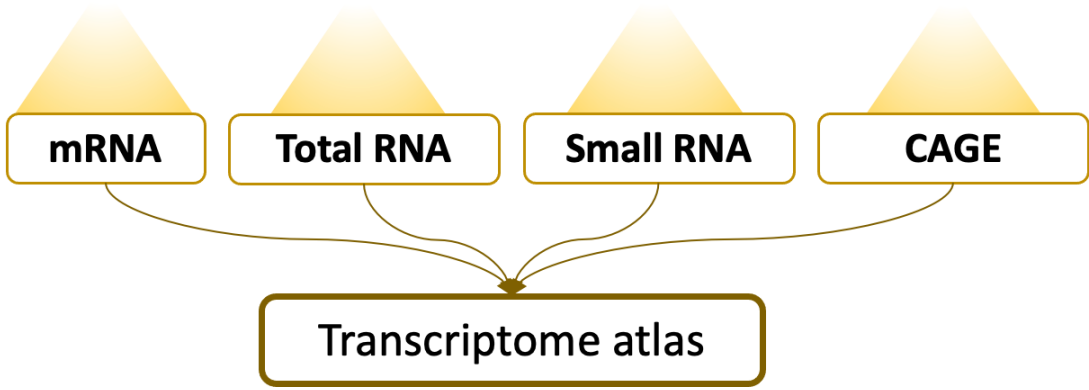
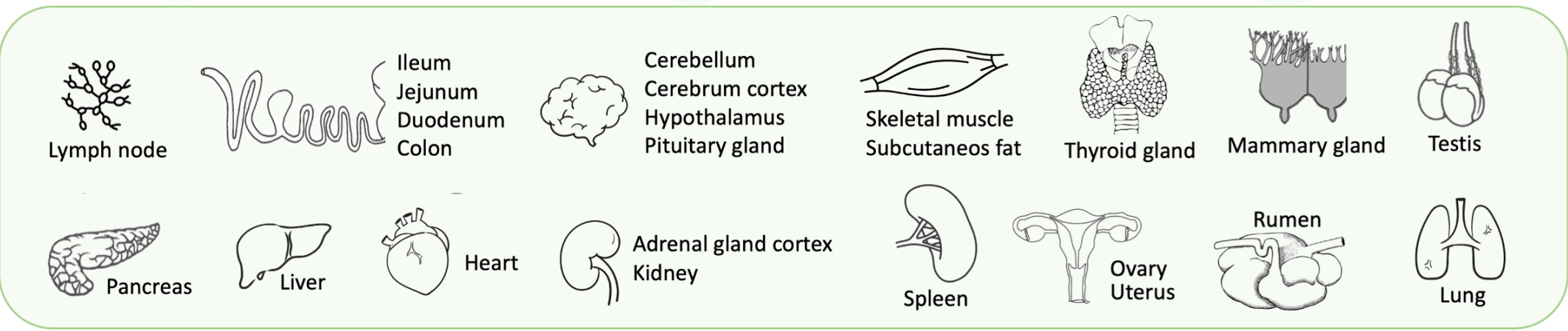
Holstein neonate



Kinsella composite juvenile

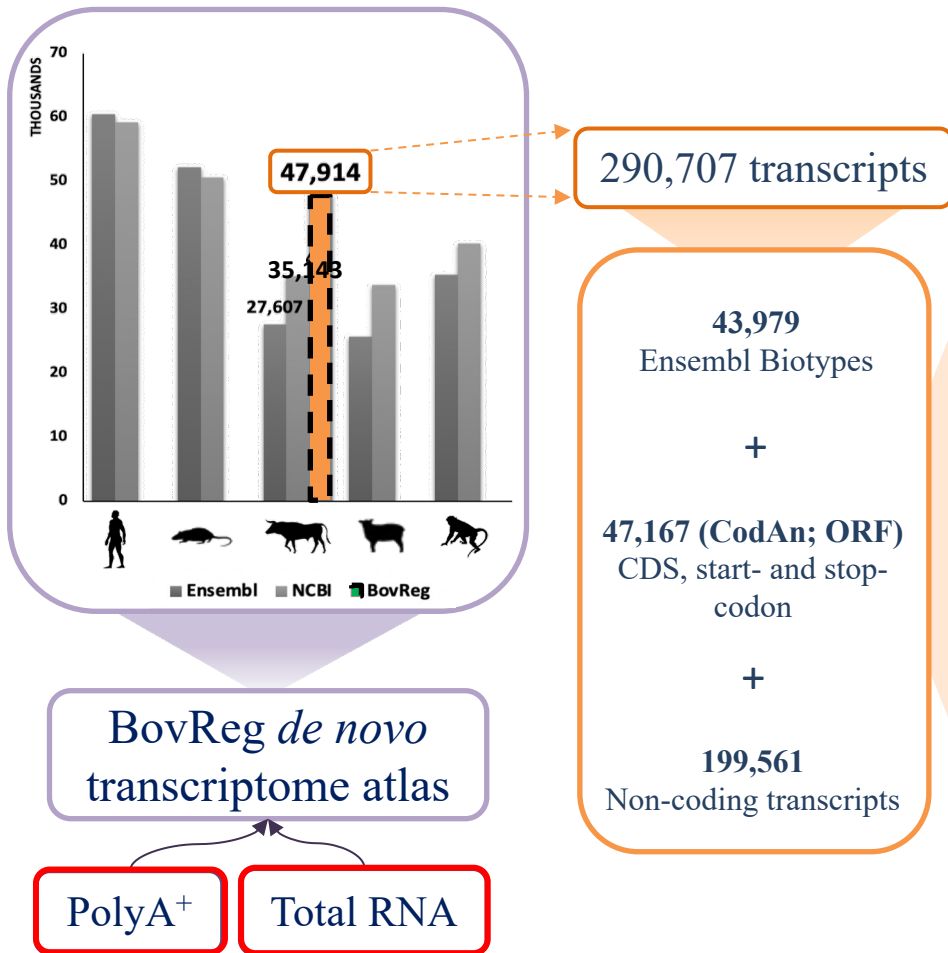


Charolais x German Holstein F2 adult



BovReg - Transcriptome Atlas

- BovReg expanded the catalog of bovine non-coding RNAs by including non-polyadenylated transcripts.
- 48k genes models including $\geq 15k$ potentially novel transcripts!

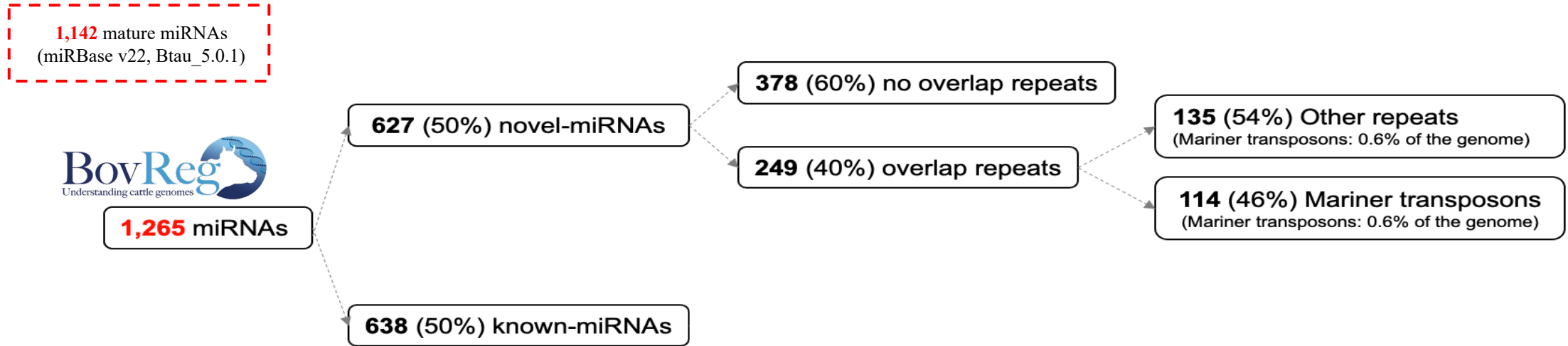


| Class code ¹ | Features ¹ | Compared with bovine annotation | |
|-----------------------------|--|---------------------------------|-----------------|
| | | Ensembl r.105 | NCBI r.106 |
| = | complete transcripts with exact match of intron chain | 42873 (14.75%) | 36660 (12.61%) |
| c | transcripts contained in the reference (intron compatible) | 4 (0.00%) | 3082 (1.06%) |
| k | containment of reference (reverse containment) | 25403 (8.74%) | 24814 (8.54%) |
| m | transcripts with retained introns (all introns matched or retained) | 4555 (1.57%) | 6321 (2.17%) |
| n | transcripts with retained introns (not all introns matched or retained) | 13812 (4.75%) | 14760 (5.08%) |
| j | multi-exon transcripts with at least one junction match | 142987 (49.19%) | 150476 (51.76%) |
| e | single exon transcript partially covering an intron (possible pre-mRNA fragment) | 0 (0.00%) | 170 (0.06%) |
| o | other same strand transcripts overlap with reference exons | 6422 (2.21%) | 9458 (3.25%) |
| s | transcripts with intron match on the opposite strand (likely a mapping error) | 63 (0.02%) | 81 (0.03%) |
| x | exonic transcripts overlap on the opposite strand (like o or e but on the opposite strand) | 10569 (3.64%) | 8817 (3.03%) |
| i | transcripts fully contained within a reference intron | 10961 (3.77%) | 12419 (4.27%) |
| y | transcripts with a reference transcript within its intron | 1871 (0.64%) | 999 (0.34%) |
| p | possible polymerase run-on (no actual overlap) | 357 (0.12%) | 342 (0.12%) |
| r | repeat (at least 50% bases soft-masked) | 0 (0.00%) | 0 (0.00%) |
| u | none of the above (unknown, intergenic) | 30830 (10.61%) | 22308 (7.67%) |
| Total number of transcripts | | | 290,707 |

¹ Transcript classification code from gffcompare v.0.12.2¹⁰.

BovReg - Transcriptome Atlas – mature miRNAs

➤ BovReg expanded the repertoire of mature miRNAs annotated in bovine.



58% non-canonical or false entries

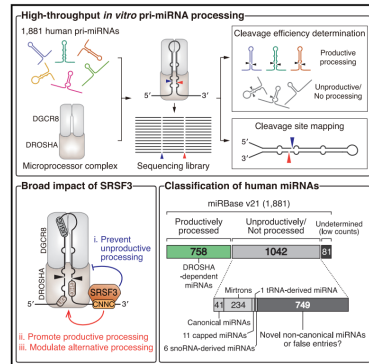
Are they all processed through canonical pathways?
What about mirtrons?

Molecular Cell

Resource

A quantitative map of human primary microRNA processing sites

Graphical abstract



Authors

Kijun Kim, S. Chan Baek, Young-Yoon Lee, Carolien Bastiaanssen, Jeesoo Kim, Haedong Kim, V. Narry Kim

Correspondence

narrykim@snu.ac.kr

In brief

More than 1,800 human miRNAs have been annotated, yet it remains unknown how many are authentic and canonical miRNAs. Kim et al. perform high-throughput *in vitro* processing to provide the first quantitative map of primary miRNA processing sites, which reveals their DROSHA dependence and enables miRNA classification.

SCIENTIFIC REPORTS

OPEN

Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods

22 November 2017
13 April 2018
15 May 2018

Grzegorz Rorbach¹, Olgierd Unold¹ & Bogumil M. Konopka²

PLOS COMPUTATIONAL BIOLOGY

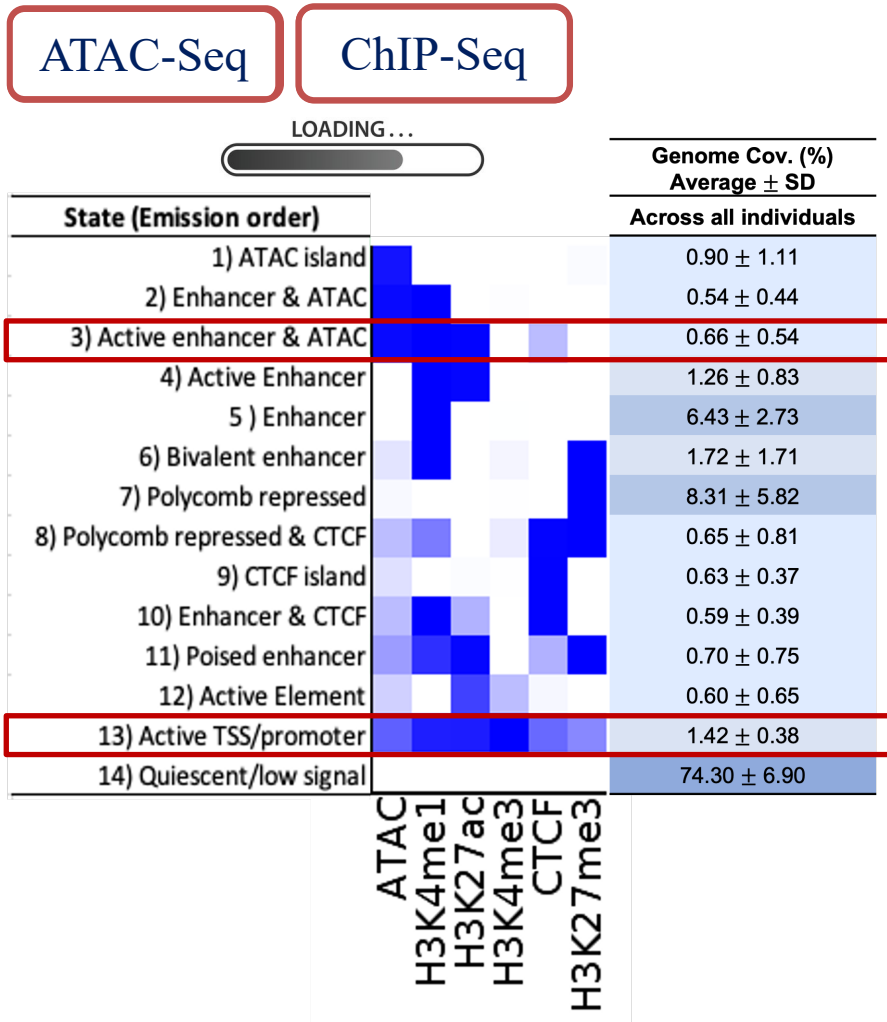
RESEARCH ARTICLE

Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates

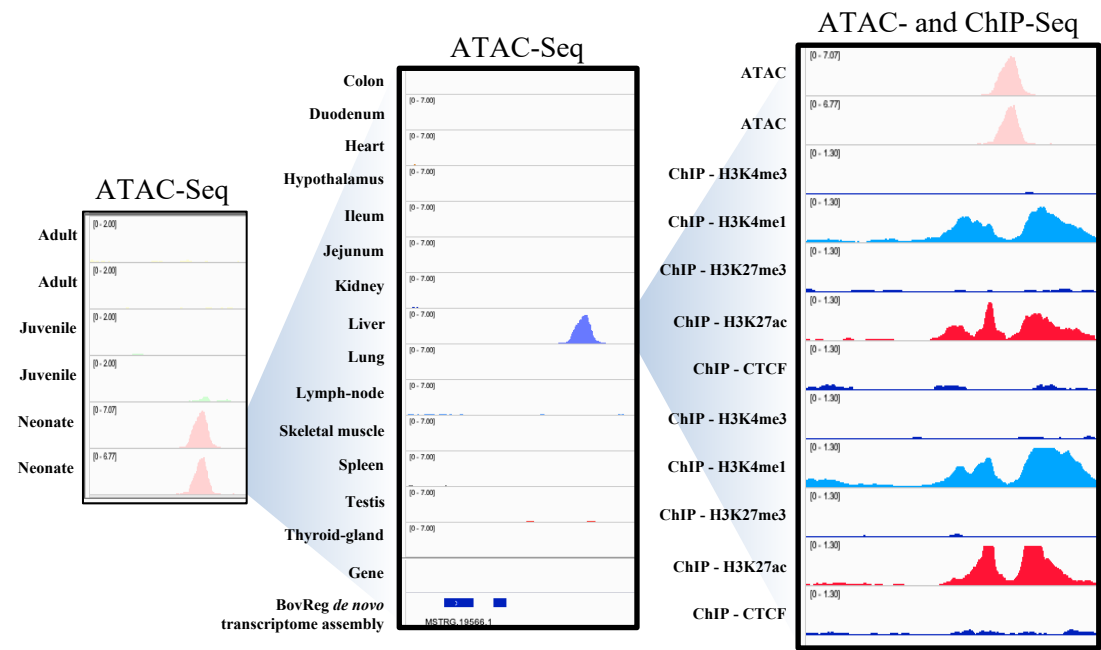
Jiayu Wen^{1*}, Erik Ladewig^{1*}, Sol Shenker^{1,2}, Jaaved Mohammed^{1,2}, Eric C. Lai^{1*}

BovReg - Regulatory Regions

➤ Around 25% of the genome was covered by different chromatin states (except quiescent)



Active enhancer (liver specific) only detected in neonate animals!!



79% TSS annotated on RNA-Seq were validated by chromatin states

BovReg - Regulatory Regions

➤ We recaptured known enhancers and annotated new ones!



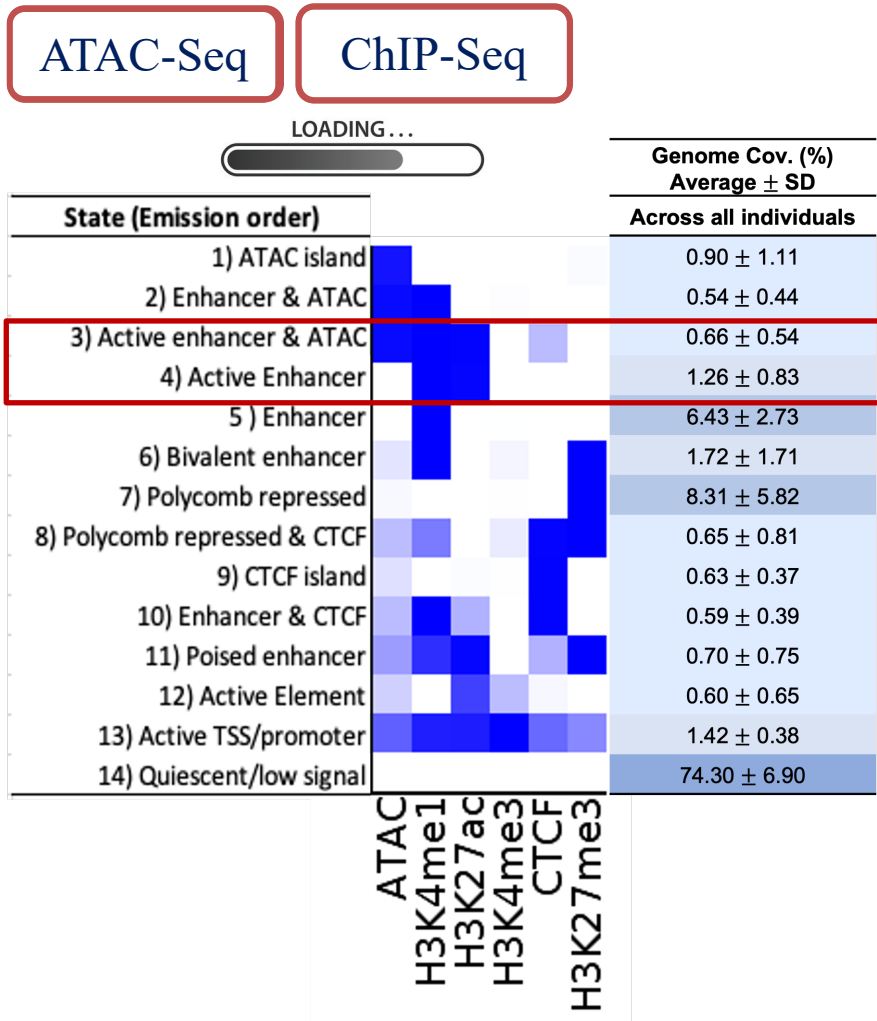
Line 1 Herefords



Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research

Colin Kern¹, Ying Wang¹, Xiaoqin Xu¹, Zhangyuan Pan¹, Michelle Halstead¹, Ganrea Chanthavixay¹, Perot Saelao¹, Susan Waters¹, Ruidong Xiang^{2,3}, Amanda Chamberlain³, Ian Kort⁴, Mary E. Delany¹, Hans H. Cheng⁵, Juan F. Medrano¹, Alison L. Van Eenennaam¹, Chris K. Tuggle⁶, Catherine Ernst⁷, Paul Flicek⁸, Gerald Quon⁹, Pablo Ross¹⁰ & Huaijun Zhou¹⁰

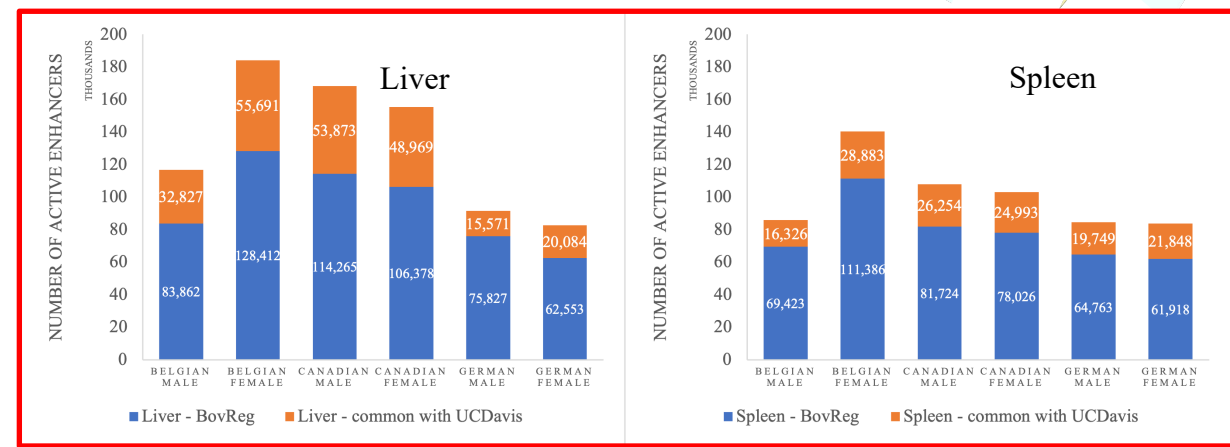
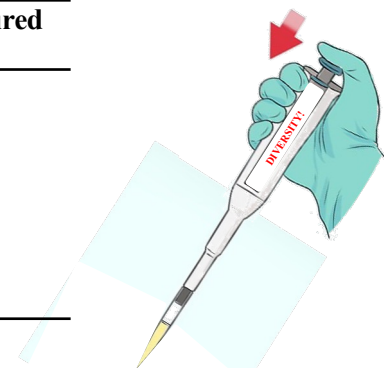
- 8 tissues from 2 Herefords male cattle (14 months old)



BovReg active enhancers

vs

| Tissues | Active enhancers (Kern et al., 2021) | Enhancers recaptured BovReg |
|--------------|--------------------------------------|-----------------------------|
| Cerebellum | 16,612 | 12,993 (78.21%) |
| Cortex | 28,235 | 24,408 (86.45%) |
| Hypothalamus | 27,796 | 17,277 (62.16%) |
| Liver | 57,457 | 46,312 (80.60%) |
| Lung | 49,861 | 38,537 (77.29%) |
| Muscle | 28,023 | 19,442 (69.38%) |
| Spleen | 29,707 | 24,520 (82.54%) |





Description of genomic structure and functional features in the five cell lines

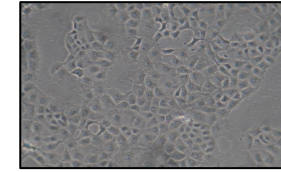
D. Becker, F. Hadlich, C. Mörke, A. Sharma, J. Vanselow, R. Weikard, Ch. Kühn (FBN),
D. Rocha (INRAE), G. Costa, Ch. Charlier (ULIEGE)



Description of genomic structure and functional features in the five cell lines

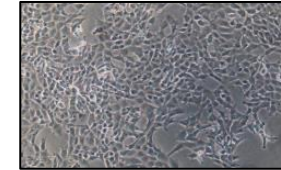
❖ EBL

- Embryonic bovine lung cells: established from a **lung** of a 7-month old bovine fetus, spontaneously immortalized



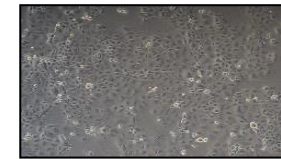
❖ MDBK

- Madin-Darby Bovine Kidney cells: derived from a **kidney** of an adult steer, spontaneously immortalized



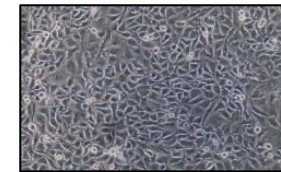
❖ F3 trophoblast cells:

- isolated from a bovine cotyledon of a male fetus (approx. 5 month of gestation), spontaneously immortalized



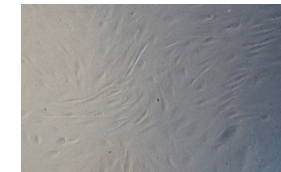
❖ Mac-T mammary alveolar cells:

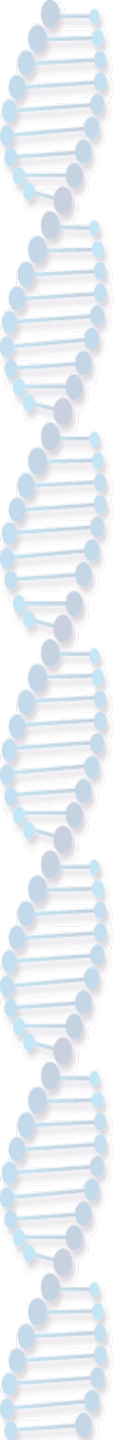
- generated from mammary alveolar cells using the large T antigen of SV40



❖ Luteal cells:

- generated from ovary tissue from lactating Holstein cow (A. Sharma, FBN)



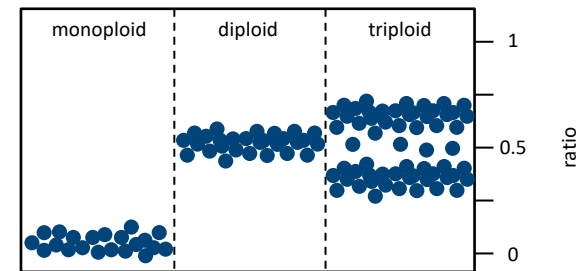
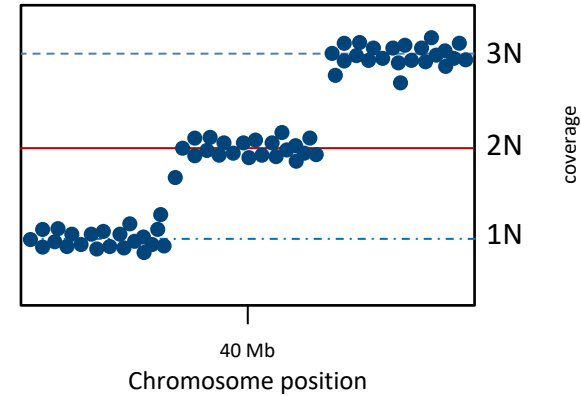


❖ Genomic features – large structural aberrations

Criteria

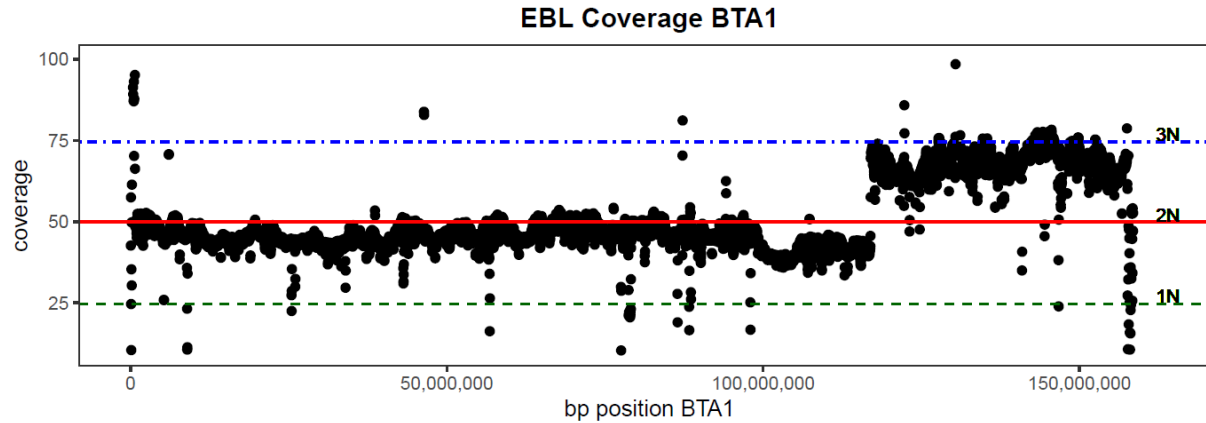
- **Read depth**
 - Coverage of each chromosome should be similar to the average coverage

- **Allele support for heterozygous SNPs**
 - ratio of reads that support the alternative allele for heterozygous SNPs should be around 0.5 for diploid chromosomes

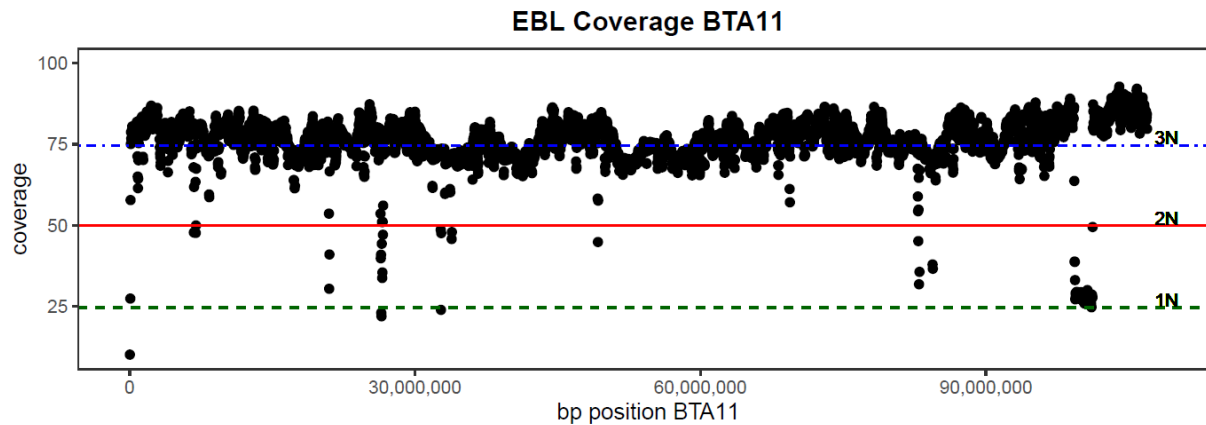




Example chromosomal aberration EBL



→ BTA1 partly triploid



→ BTA11 triploid



Description of genomic structure and functional features in the five cell lines

| Cell Line | Aneuploidy (whole chromosome) | Structural variation (part of the chromosome) |
|-----------|----------------------------------|---|
| EBL | 1 | 11 |
| F3 | 6 | 2 |
| MacT | 5 | 11 |
| MDBK | 1 | 7 |
| Luteal | 0 | 7 |



Aneuploidy and structural variations are common in all characterized cell lines



Overview expression activity

| | TPM > 1 ^a | TPM > 0.1 ^b | Median TPM ^c |
|---------------|----------------------|------------------------|-------------------------|
| EBL | 12,205 | 14,423 | 12.26 |
| F3 | 11,348 | 13,280 | 8.28 |
| MacT | 11,478 | 13,709 | 9.90 |
| MDBK | 12,273 | 14,568 | 11.88 |
| Luteal | 11,780 | 14,186 | 8.98 |



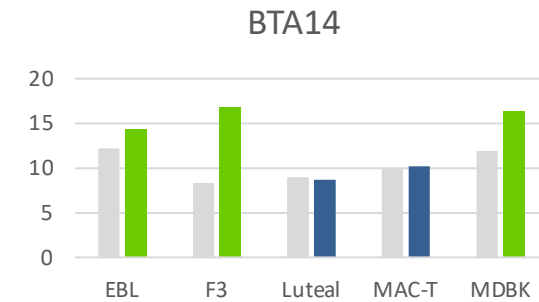
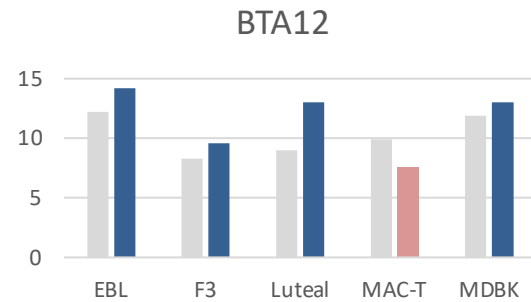
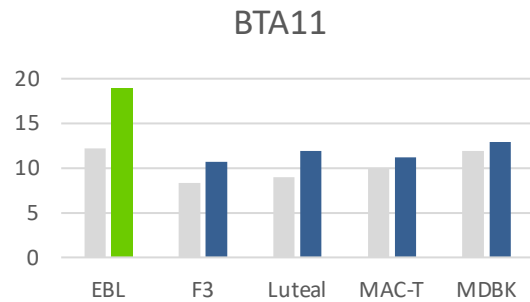
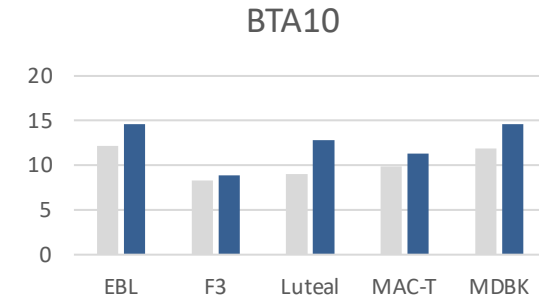
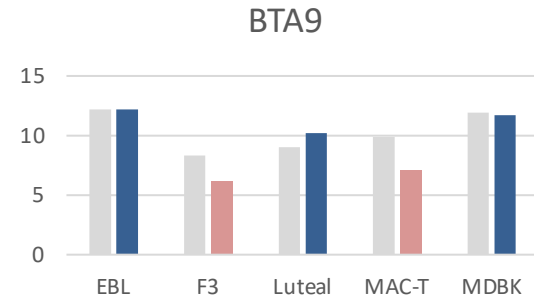
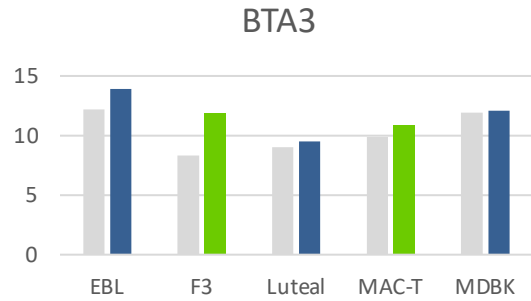
Between 13,280 (F3) to 14,568 (MDBK) genes were expressed per cell line

- a: number of genes with TPM > 1
- b: number of genes with TPM > 0.01
- c: median TPM (only genes with TPM>0)



Description of genomic structure and functional features in the five cell lines

Median TPM values



across all chromosomes

No strong signal of aneuploidy

Indication of triploidy

Indication of haploidy



Reproducible, scalable and shareable analysis pipelines in nf-core

Jose Espinosa-Carrasco, Björn Langer, Cedric Notredame



Perspective | [Published: 23 September 2021](#)

Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers

[Laura Wratten](#), [Andreas Wilm](#) & [Jonathan Göke](#)

[Nature Methods](#) **18**, 1161–1168 (2021) | [Cite this article](#)

Table 1 | Overview of workflow managers for bioinformatics (top, editable version; bottom, image version)

| Tool | Class | Ease of use ^a | Expressiveness ^b | Portability ^c | Scalability ^d | Learning resources ^e | Pipeline initiatives ^f |
|-----------------------|------------------------------------|--------------------------|-----------------------------|--------------------------|--------------------------|---------------------------------|-----------------------------------|
| Galaxy | Graphical | ●●● | ●○○ | ●●● | ●●● | ●●● | ●●○ |
| KNIME | Graphical | ●●● | ●○○ | ○○○ | ●●○ | ●●● | ●●○ |
| Nextflow | DSL | ●○○ | ●●● | ●●● | ●●● | ●●● | ●●● |
| Snakemake | DSL | ●○○ | ●●● | ●●○ | ●●● | ●●○ | ●●● |
| GenPipes | DSL | ●○○ | ●●● | ●●○ | ●●○ | ●●○ | ●●○ |
| bPipe | DSL | ●○○ | ●●● | ●●○ | ●●○ | ●●○ | ●○○ |
| Pachyderm | DSL | ●○○ | ●●● | ●○○ | ●●○ | ●●● | ○○○ |
| SciPipe | Library | ●○○ | ●●● | ○○○ | ○○○ | ●●○ | ○○○ |
| Luigi | Library | ●○○ | ●●● | ●○○ | ●●○ | ●●○ | ○○○ |
| Cromwell + WDL | Execution + workflow specification | ●○○ | ●●○ | ●●● | ●●○ | ●●○ | ●●○ |
| cwltool + CWL | Execution + workflow specification | ●○○ | ●●○ | ●●○ | ○○○ | ●●● | ●●○ |
| Toil + CWL/WDL/Python | Execution + workflow specification | ●○○ | ●●● | ●●○ | ●●● | ●●○ | ●●○ |



- A community
- A set of guidelines (standard)
- A curated set of analysis pipelines built using Nextflow
- Helper tools

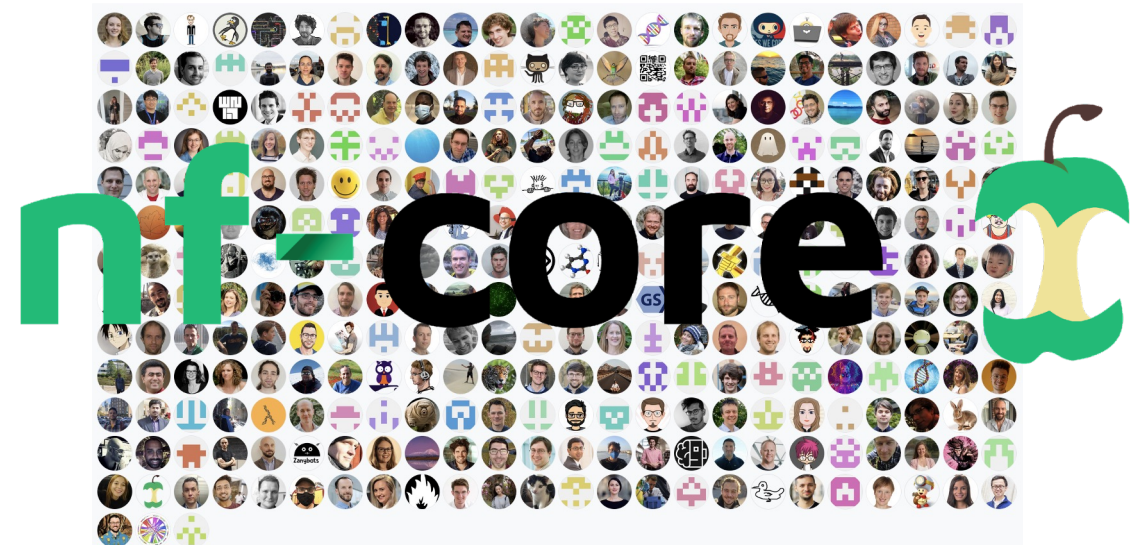
**nature
biotechnology**

Correspondence | Published: 13 February 2020

The nf-core framework for community-curated bioinformatics pipelines

Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen

Nature Biotechnology (2020) | [Cite this article](#)





The EuroFAANG pipelines working group



Who: EuroFAANG projects (BovReg, Gene-SWitCH, AQUA-FAANG, GEroNIMO, Rumigen, HoloRuminant) + nf-core core members

What: Collect existing resources, list good practices, identify gaps and overlaps

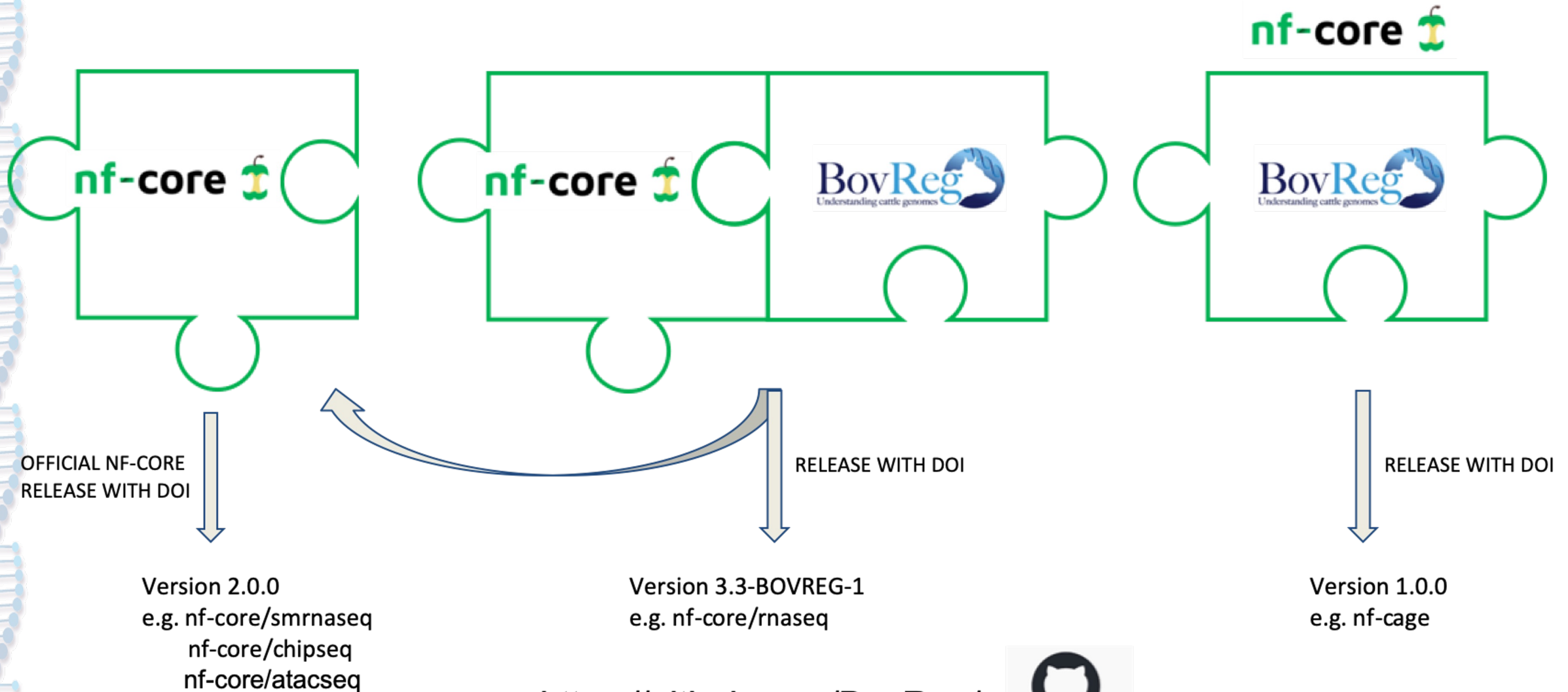
When: Monthly meetings

Lead: EMBL-EBI and CRG

BovReg partners involved: Jose Espinosa-Carrasco (CRG), Björn Langer (CRG), Cedric Notredame (CRG), Alexey Solokov, Peter Harrison (EMBL-EBI), Praveen Chitneedi (FBN), Andreia Amaral (FMV), Mathieu Charles (INRAe), Daniel Fischer (Luke), Mazdak Salavati (UEDIN), Gabriel Costa (Uliece)



BovReg pipelines examples



<https://github.com/BovReg/>

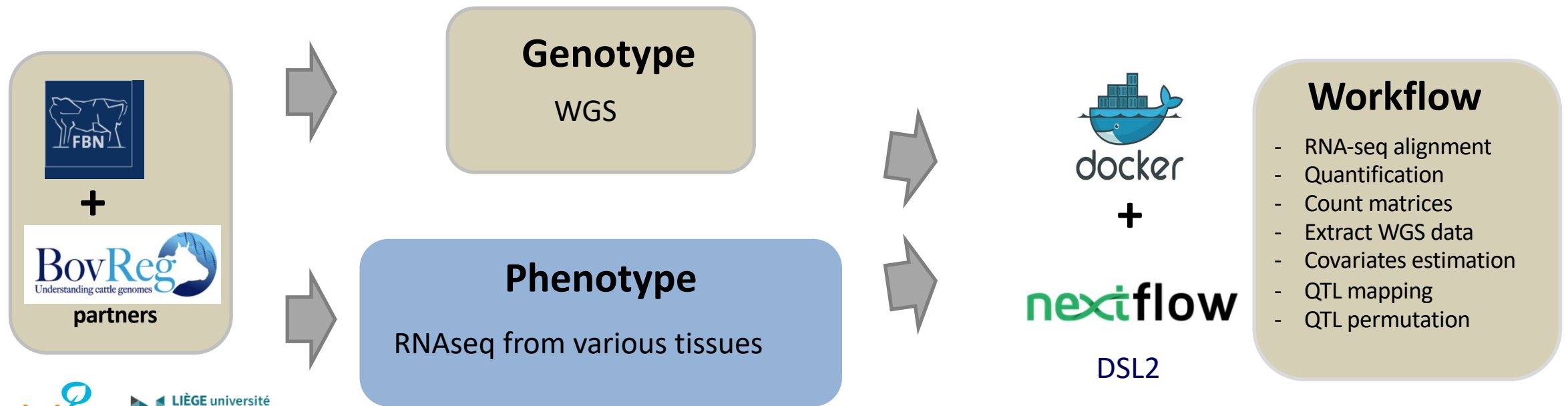




Example:



Development of a container- and workflow manager-based eQTL detection pipeline



FBN + BovReg partners



<https://github.com/BovReg/>



Summary



Summary



For the BovReg Project we have:

- ▶ **Annotated functionally active genomic regions for multiple tissues across three populations of cattle from different developmental stages.**
- ▶ **Identified novel coding and non-coding transcripts, and tissue- and developmental stage specific enhancers.**
- ▶ **Provided analysis of genomic structure and functional features in the five cell lines which suggests that aneuploidy and structural variations are common.**
- ▶ **Worked with the other EuroFAANG projects to provide reproducible, scalable and shareable analysis pipelines for the livestock genomics community pipelines in nf-core.**

The BovReg functional genome annotation team


FBN

D. Becker
P. Chitneedi
C. Mörke
R. Weikard
F. Hadlich
N. Trakooljul
F. Becker
J. Vanselow
Ch. Kühn

INRAE

M. Sanchez
D. Boichard
M. Charles
M. Bernard
P. Croiseau
D. Rocha

GIGA

G. Costa Moreira
C. Charlier

UEDIN

M. Salavati
J. Prendergast
S. Jayaraman
E. Clark

CRG

B. Langer
J. Espinosa-Carrasco
C. Notredame

EMBL-EBI

A. Sokolov
P. Harrison

AU

G. Sahana
L. Janss
E. Karaman

WR

A. Bouwman
B. Gredler
R. Veerkamp

LUKE

T. Iso-Touru
D. Fischer
J. Vilkki

ETH

N. Kadri
H. Pausch

FMV-Lisboa

A. Amaral

IGHZ PAN

N. Stojak
H. Taniguchi

ULIM

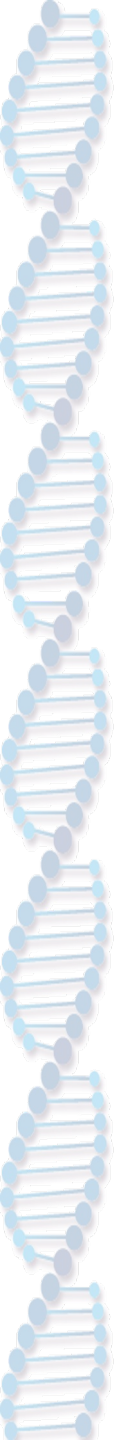
N. Duprat
V. Blanchet

UAL

C. Li
J. Li
G. Plastow

AgVic

A. Chamberlain



BovReg partners



www.bovreg.eu

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



BovReg Final conference
University Foundation - Room Felicien Cattier
Brussels
14-15 February 2024

<https://bovreg.eu/bovreg-final-conference/>