Genome wide recognition and analysis of long non-coding RNAs in the transcriptome of Arabian horses

Kacper Żukowski[1], Monika Stefaniuk-Szmukier[2],
Katarzyna Ropka-Molik[1], Monika Bugno-Poniewierska[1,2]

[1]National Research Institute of Animal Production, Cracow, Poland
[2]University of Agriculture in Cracow, Poland

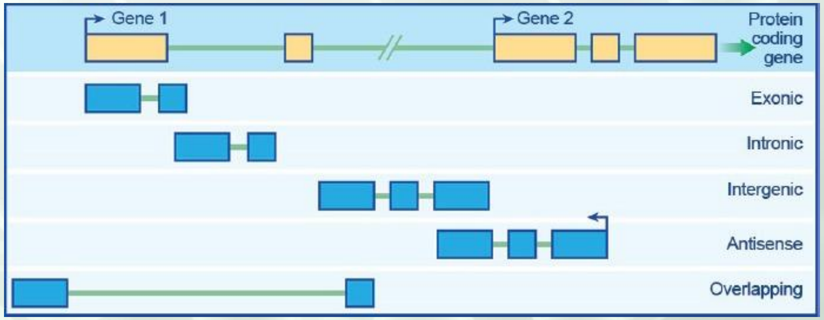
kacper.zukowski@izoo.krakow.pl

INSTYTUT ZOOTECHNIKI
PAŃSTWOWY INSTYTUT BADAWCZY

long non-coding RNA (lncRNA)

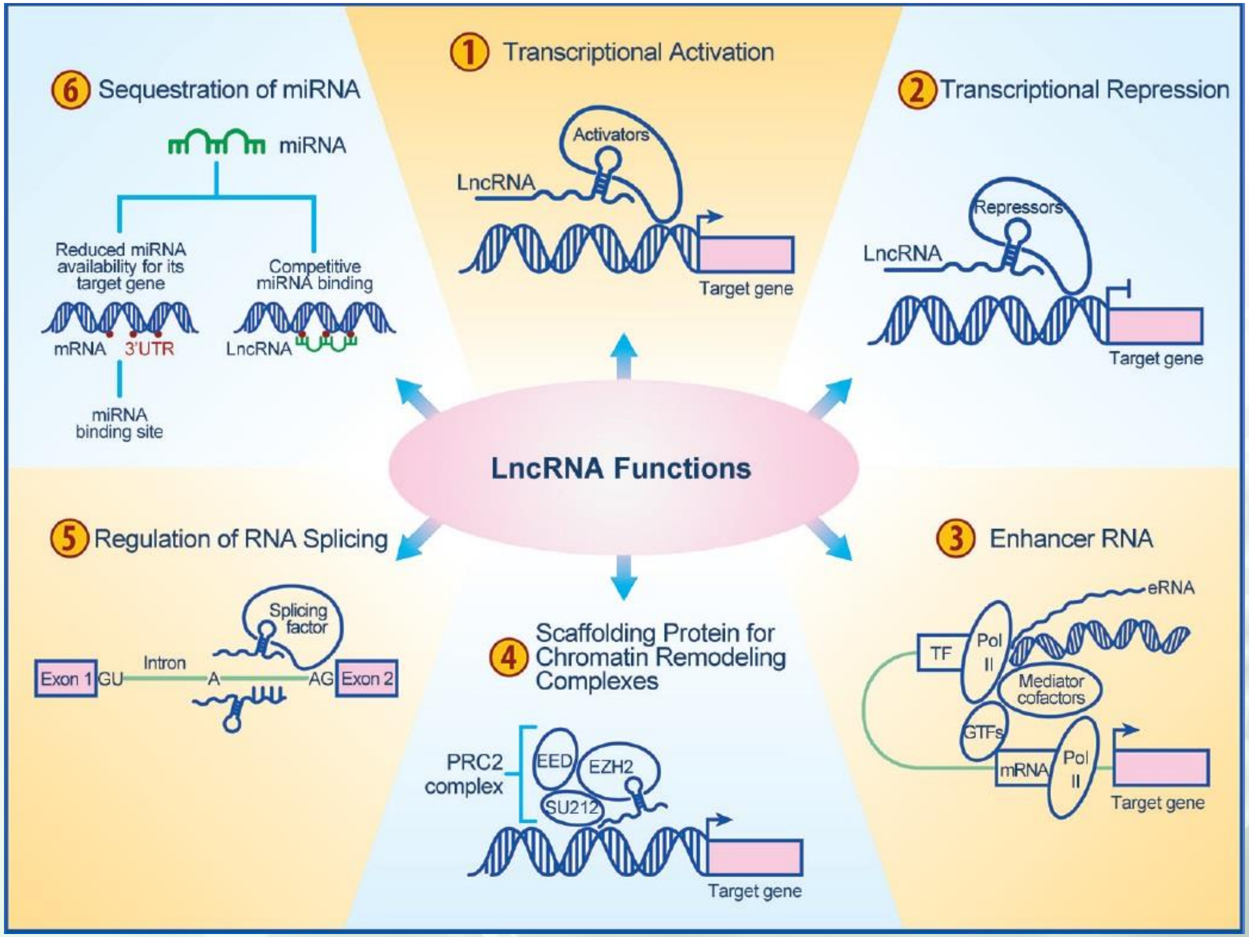
***lncRNA* is defined as all transcribed RNA molecules greater than 200 bp in length**

- the 200 bp cutoff is based on RNA isolation protocols.
- very broad category, so lncRNAs are a very large and functionally heterogeneous group.



Classification of long noncoding RNAs, *Malik and Feng, 2016*

The major molecular mechanisms of lncRNAs



Malik and Feng, 2016

The aim of the study

recognition and analysis of previously identified and novel *lncRNAs* in Arabian horses tissues (blood and muscles) across different genome releases.

Material and methods

- **pure breed Arabian horses**
- **16 blood samples**
    - Ropka-Molik K, Stefaniuk-Szmukier M, Żukowski K, Piórkowska K, Gurgul A, Bugno-Poniewierska M (2017) *Transcriptome profiling of Arabian horses blood during training regime* BMC Genetics 18:31
    - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=**GSE83404**
- **15 muscles samples**
    - Ropka-Molik K, Stefaniuk-Szmukier M, Żukowski K, Piórkowska K, Bugno-Poniewierska M (2017) *Exercise-induced modification of skeletal muscle transcriptome in Arabian horses* Physiological Genomics 49: 318–326
    - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=**GSE88951**

- **RNA isolation**
- **construction of cDNA libraries with the use of the TruSeq RNA Kit v2 (Illumina)**
- **sequencing on the HiScanSQ platform (Illumina) - 75 single-end cycles**

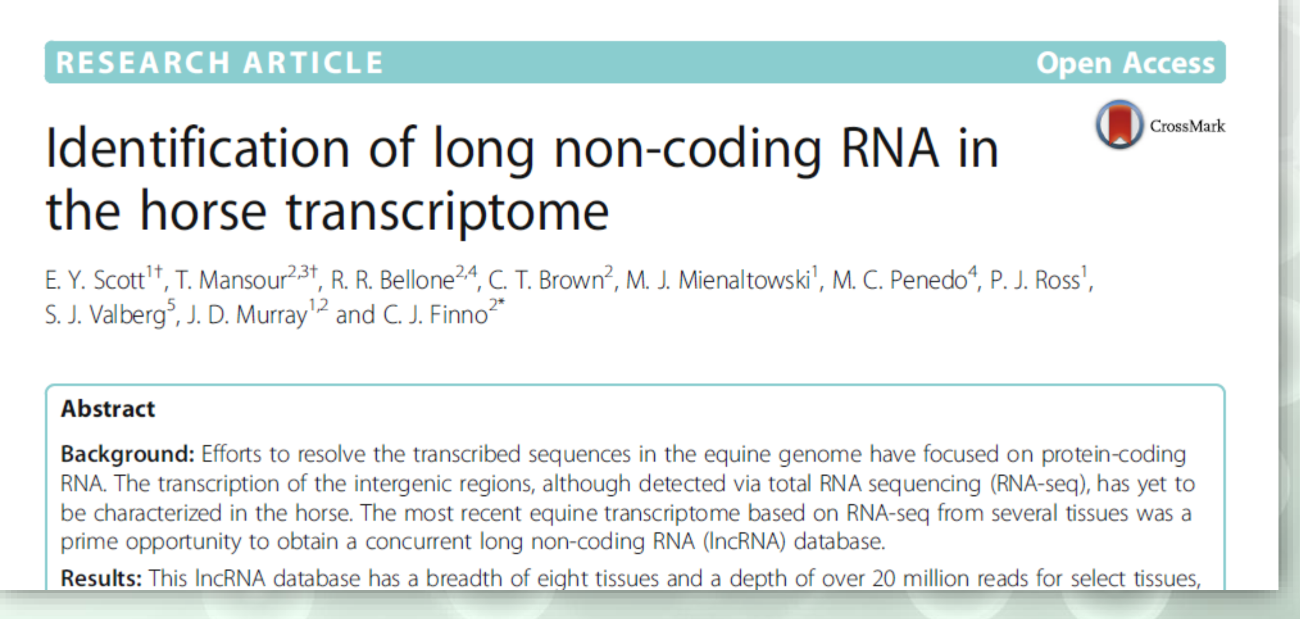
Material, horse genome assemblies

- 31 autosomes, two sex chromosomes and (un) scaffolds
- EquCab3.0 assembly; GCA_002863925.1
  - https://www.ncbi.nlm.nih.gov/genome?term=equus%20caballus
  - NCBI Annotation Release: 103
  - median total length (Mb): 2474.93
- EquCab2.93 assembly; INSDC Assembly GCA_000002305.1
  - http://www.ensembl.org/Equus_caballus/Info/Annotation
  - Ensembl Annotation Release: 93.2
  - median total length (Mb): 2428.79

Structure	EquCab2.93	EquCab3.0
Coding	20449	32847
Non coding	2142	16584
Small non coding	1967	
Misc non coding	175	
Pseudogenes	4400	2746
Genscan predictions	107701	113077

Annotated lncRNA

Scott EY, Mansour T, Bellone RR, Brown CT, Mienaltowski MJ, Penedo MC, Ross PJ, Valberg SJ, Murray JD, Finno CJ (2017) *Identification of long non-coding RNA in the horse transcriptome* BMC Genomics. 18(1):511.



RESEARCH ARTICLE **Open Access**

CrossMark

Identification of long non-coding RNA in the horse transcriptome

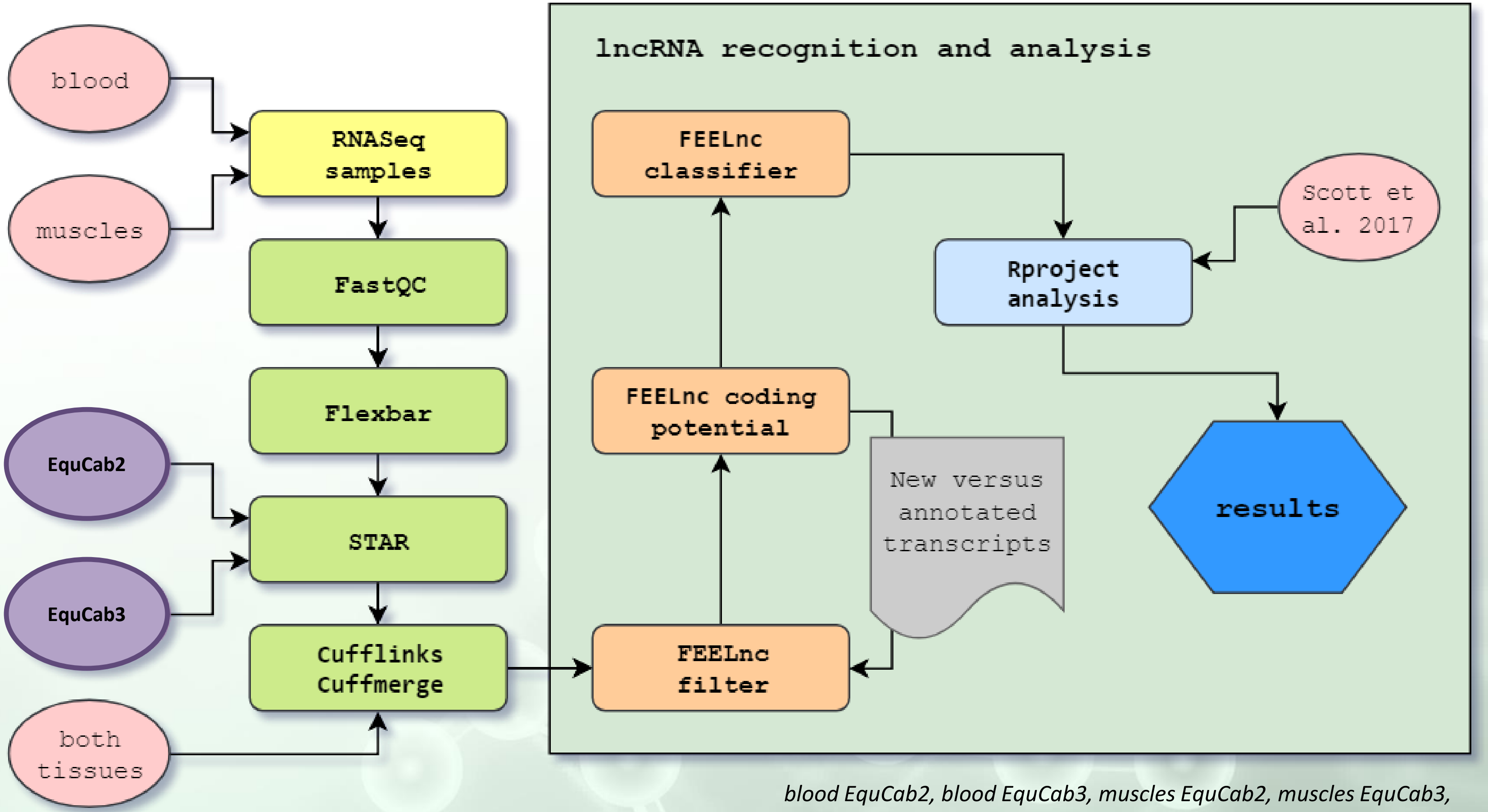
E. Y. Scott[1†], T. Mansour[2,3†], R. R. Bellone[2,4], C. T. Brown[2], M. J. Mienaltowski[1], M. C. Penedo[4], P. J. Ross[1], S. J. Valberg[5], J. D. Murray[1,2] and C. J. Finno[2*]

Abstract

Background: Efforts to resolve the transcribed sequences in the equine genome have focused on protein-coding RNA. The transcription of the intergenic regions, although detected via total RNA sequencing (RNA-seq), has yet to be characterized in the horse. The most recent equine transcriptome based on RNA-seq from several tissues was a prime opportunity to obtain a concurrent long non-coding RNA (lncRNA) database.

Results: This lncRNA database has a breadth of eight tissues and a depth of over 20 million reads for select tissues,

Bioinformatic pipeline


lncRNA recognition and analysis

blood

muscles

RNASeq samples

FastQC

Flexbar

EquCab2

EquCab3

STAR

Cufflinks Cuffmerge

both tissues

FEELnc filter

FEELnc coding potential

New versus annotated transcripts

FEELnc classifier

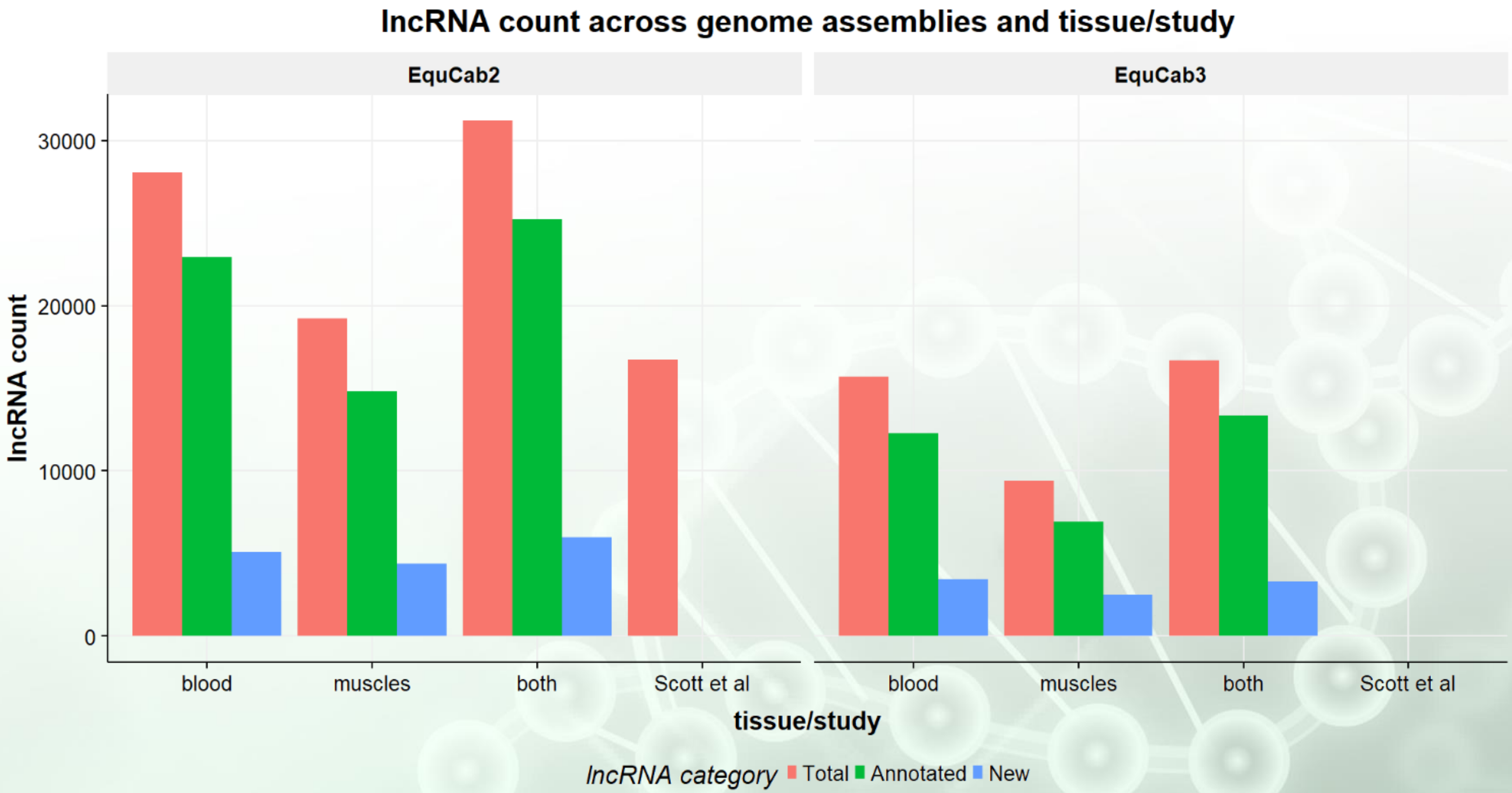
Scott et al. 2017

Rproject analysis

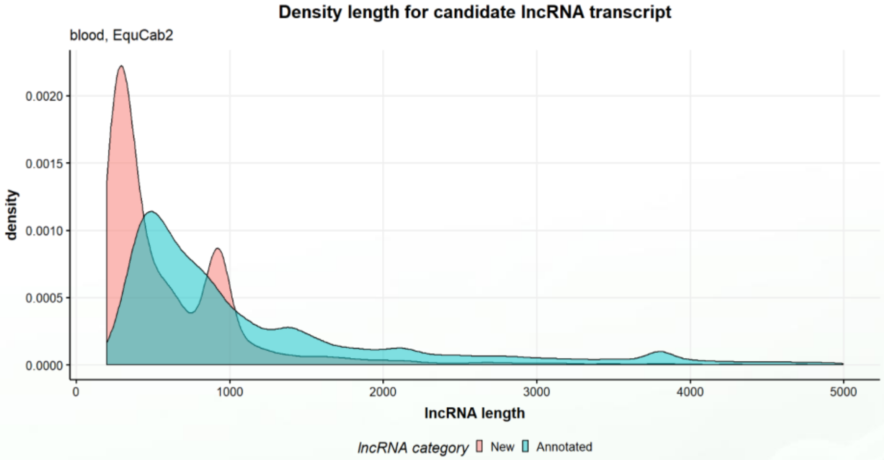
results

blood EquCab2, blood EquCab3, muscles EquCab2, muscles EquCab3, both EquCab2, both EquCab3 and Scott et al.

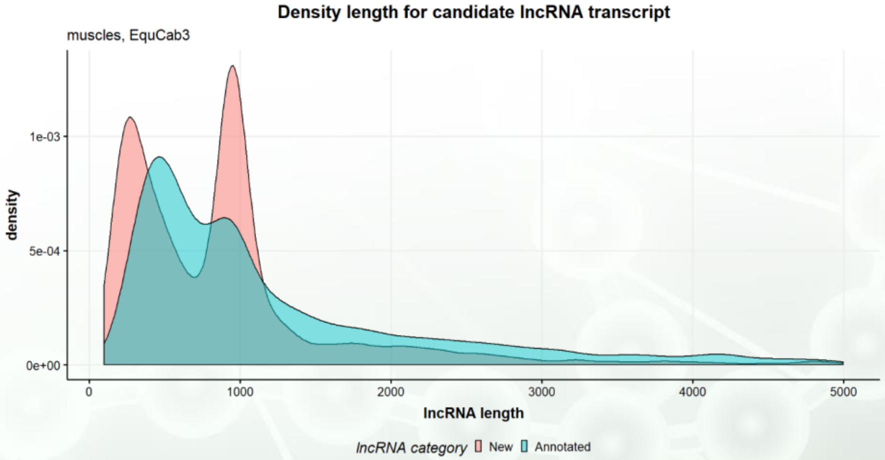
Results, lncRNAs count



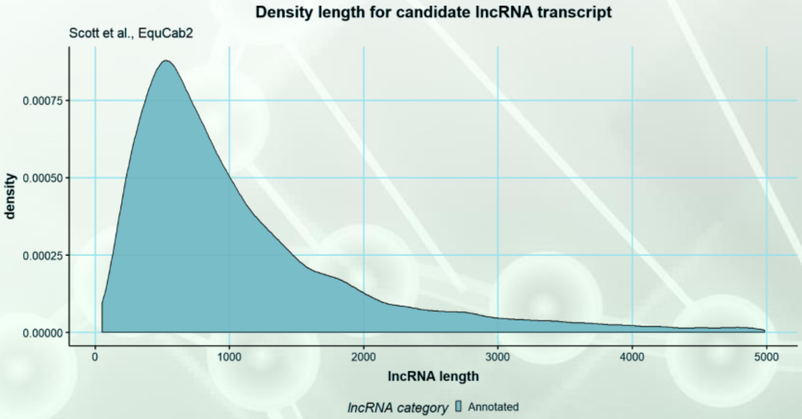
Results, lncRNA length distribution


Density length for candidate lncRNA transcript

blood, EquCab2

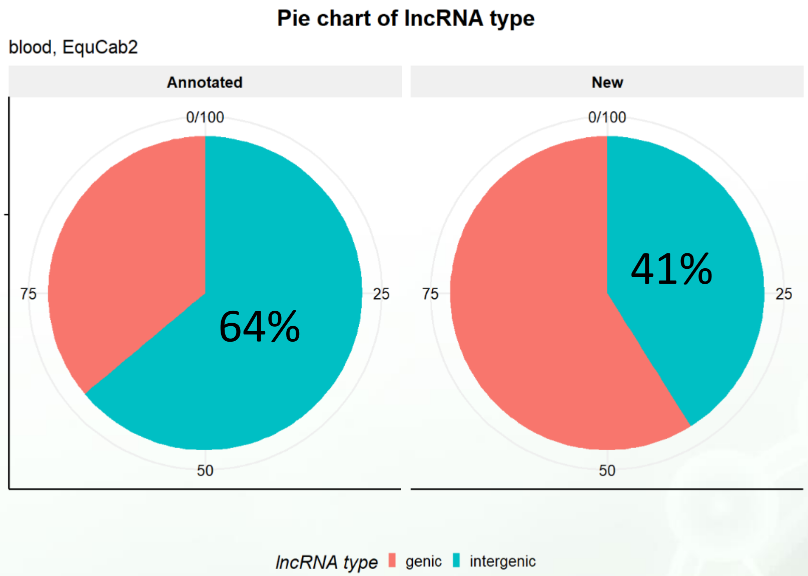

Density length for candidate lncRNA transcript

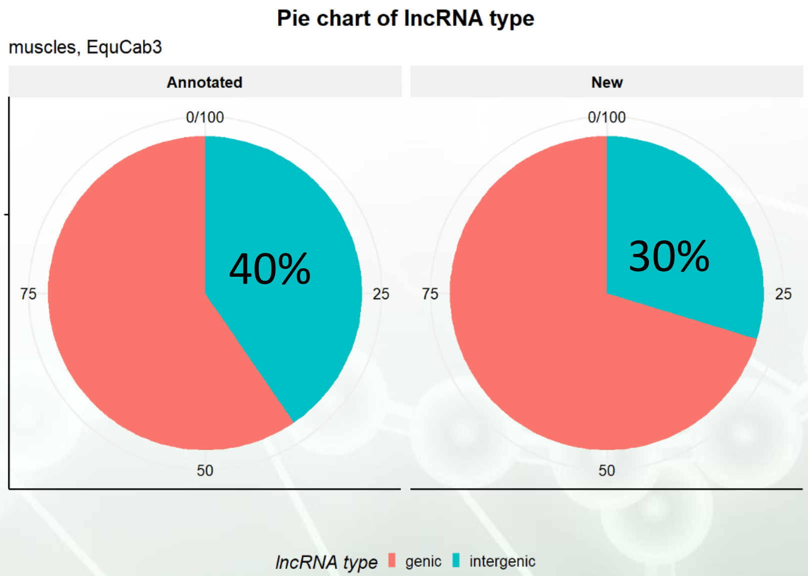
muscles, EquCab3

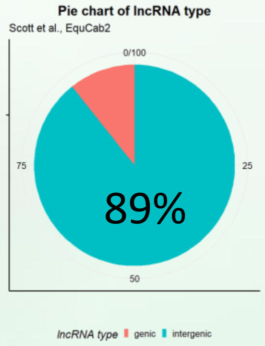

Density length for candidate lncRNA transcript

Scott et al., EquCab2

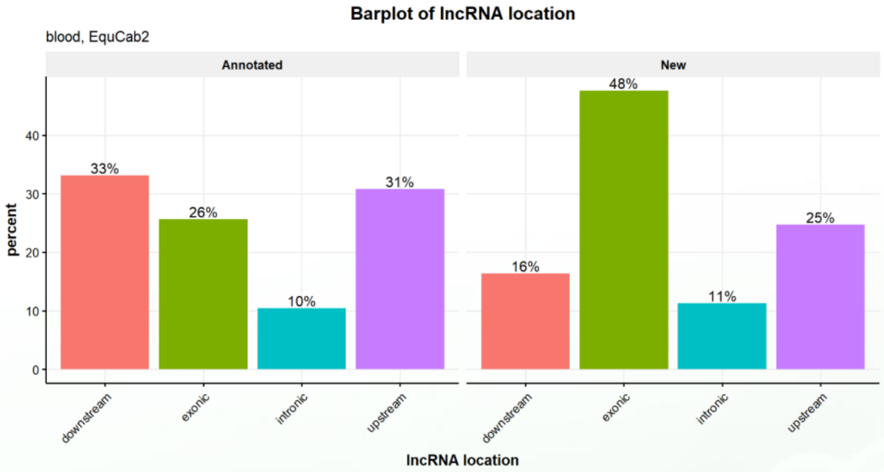
Results, type of interactions with nearest transcripts

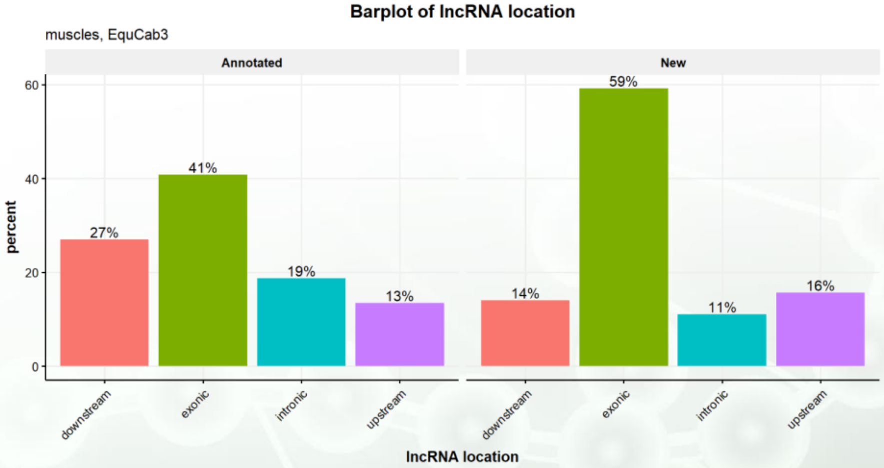


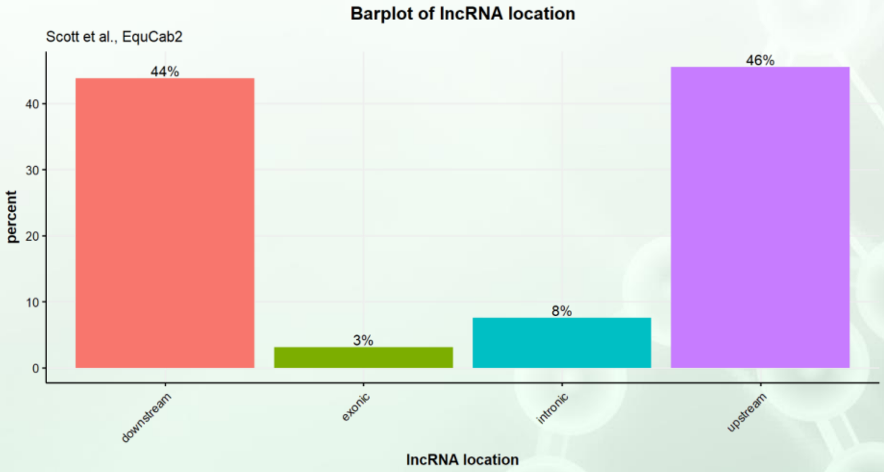




Results, localization of the interactions with nearest transcripts

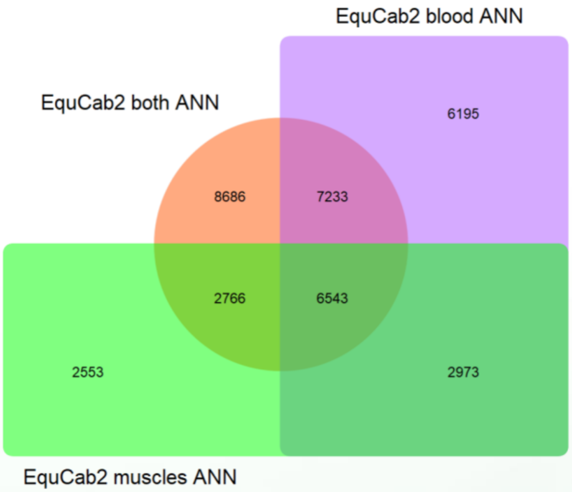




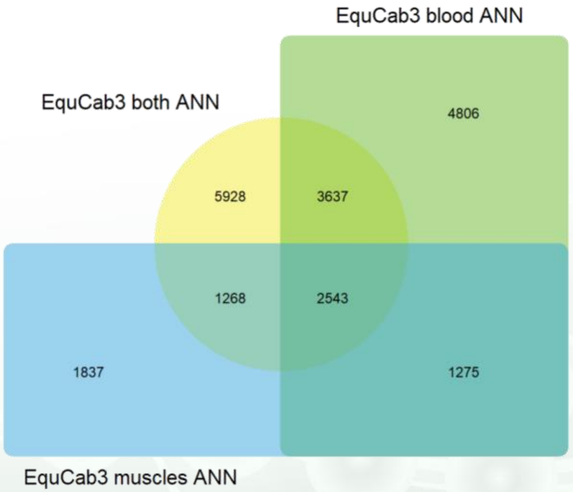


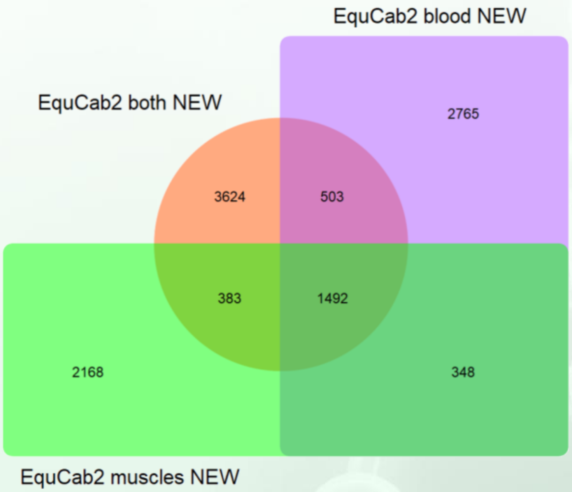
downstream
exonic
intronic
upstream

Results, Comparison of within lncRNAs group

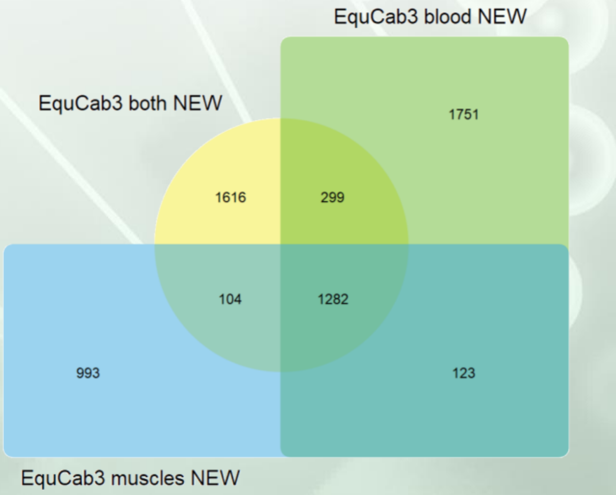


Annotated



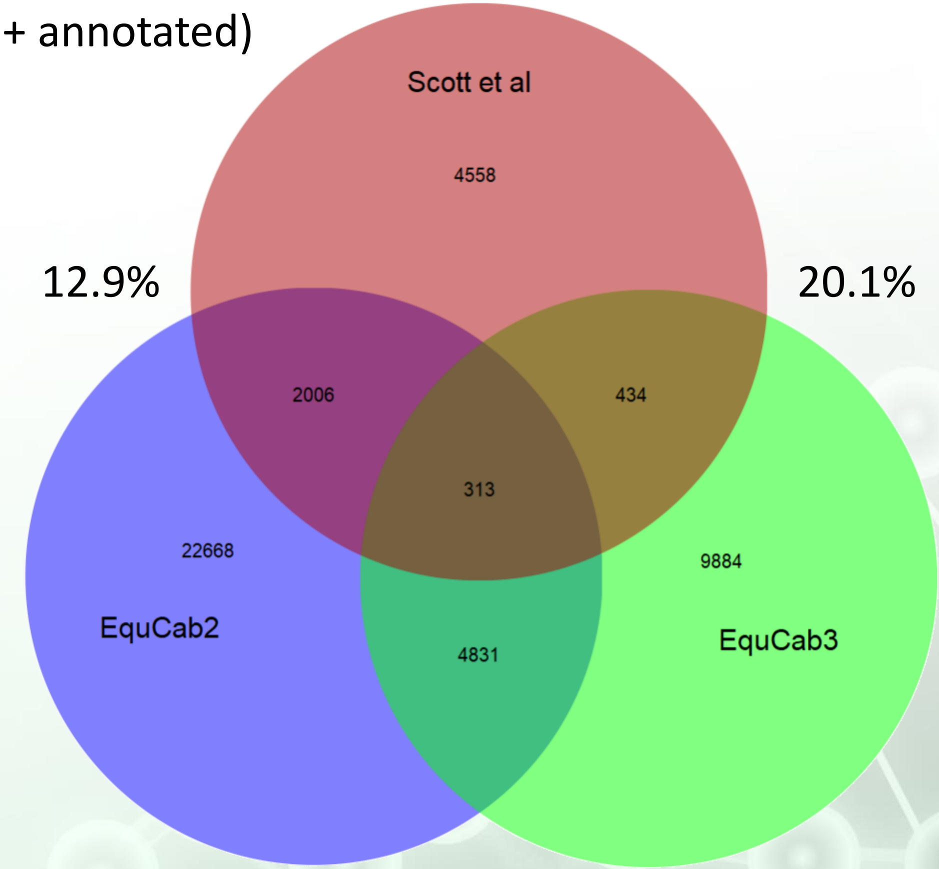


New



Results, Comparison with Scott et al.

whole set (new + annotated)



Summary and further plans

Summary:

- we identified more than 31K *lncRNAs* for *EquCab2* and more than 17K for *EquCab3* for both tissues,
- we recognized more lncRNAs for blood than for muscles across both assemblies,
- we observed patterns across tissues types more connected with the number of *lncRNAs* and interactions with nearest transcripts,
- quite low concordance in the number of *lncRNAs* with *Scott et al.* study.

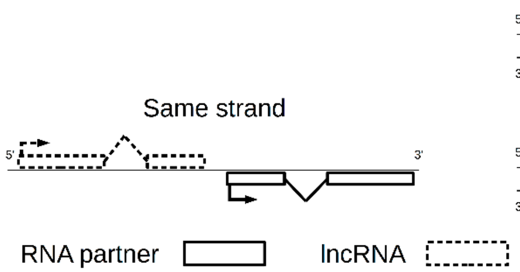
Further plans:

- evaluate expression level of identified *lncRNAs*,
- reanalysis with official *EquCab3*,
- better assemblies comparison (BLASTN),
- extend study with number of samples and tissues,
- release annotated *lncRNAs* within NCBI project page.

Thank you for your attention

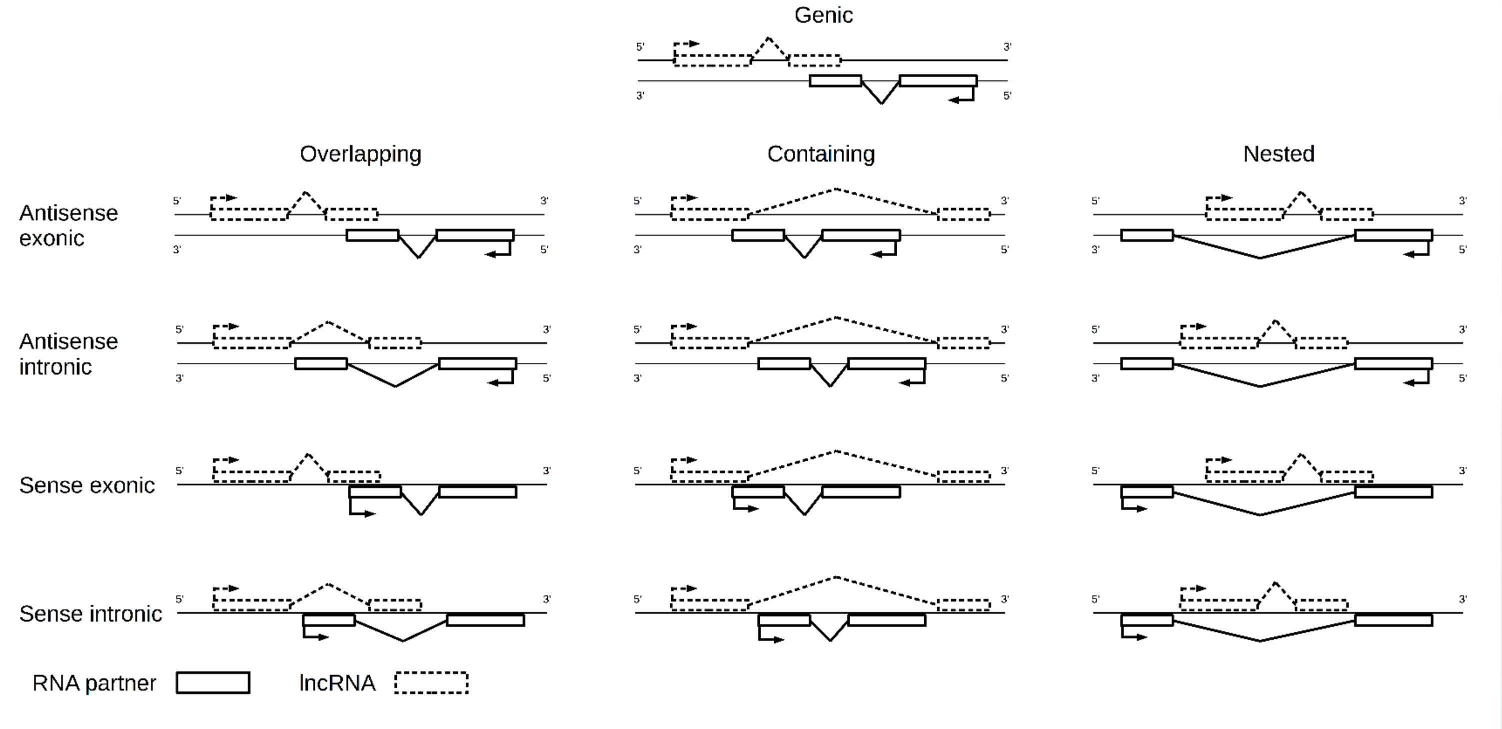
National Science Centre in Cracow, Poland,
project no. 2014/15/D/NZ9/05256

FEELnc Illustration of the classification



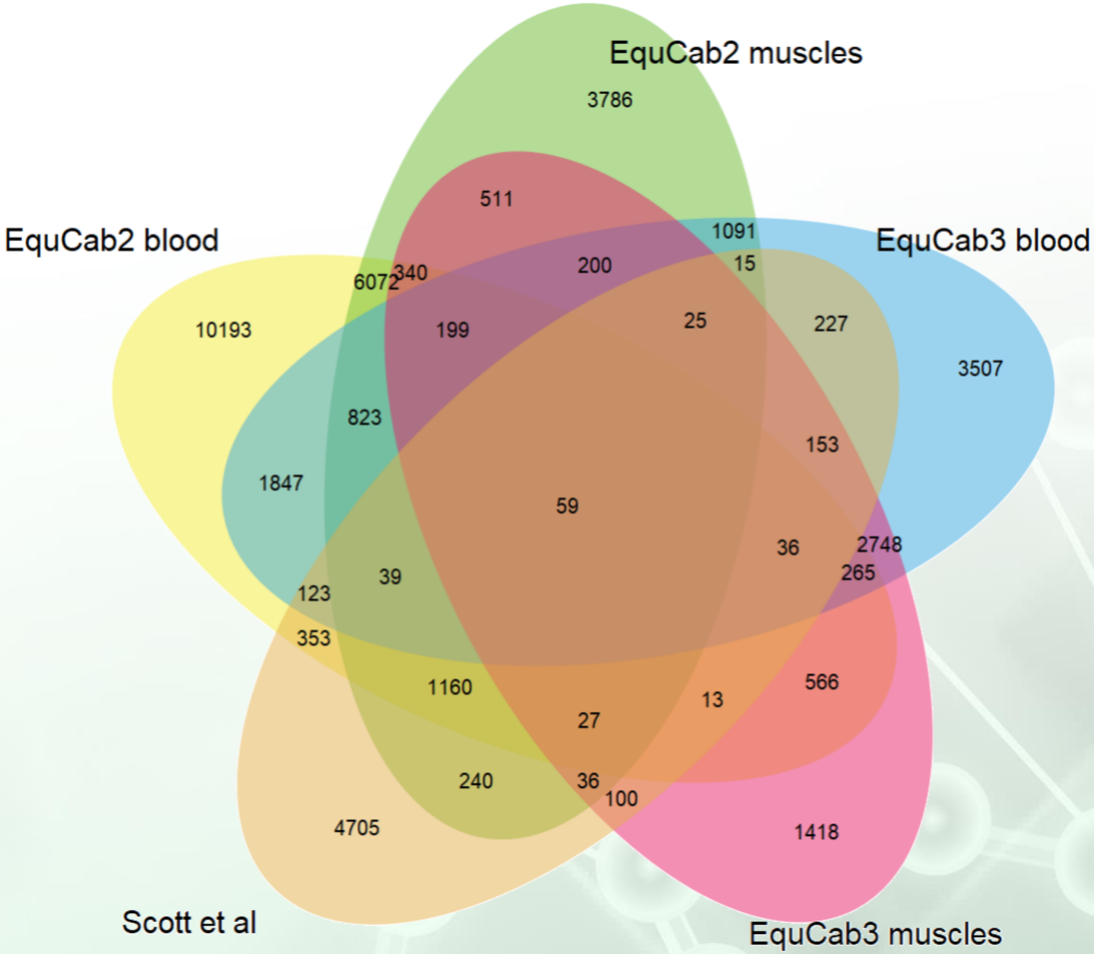
https://github.com/tderrien/FEELnc

FEELnc *Illustration of the classification*



https://github.com/tderrien/FEELnc

Results, Comparison of ANNOTATED lncRNA



Results, Comparison of NEW lncRNA