Identification of long non-coding RNAs in livestock species

Oana Palasca

February 15th, 2018

RTH
CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

Overview

- http://www.faang-europe.org
- Community effort to establish a database of ncRNAs in livestock species
- Annotation based on experimental evidence approach (RNA-Seq)
- Initiated as a hackathon in October 2016

Background

McMullen et al, Clinical Science, 2016

Long non coding RNAs

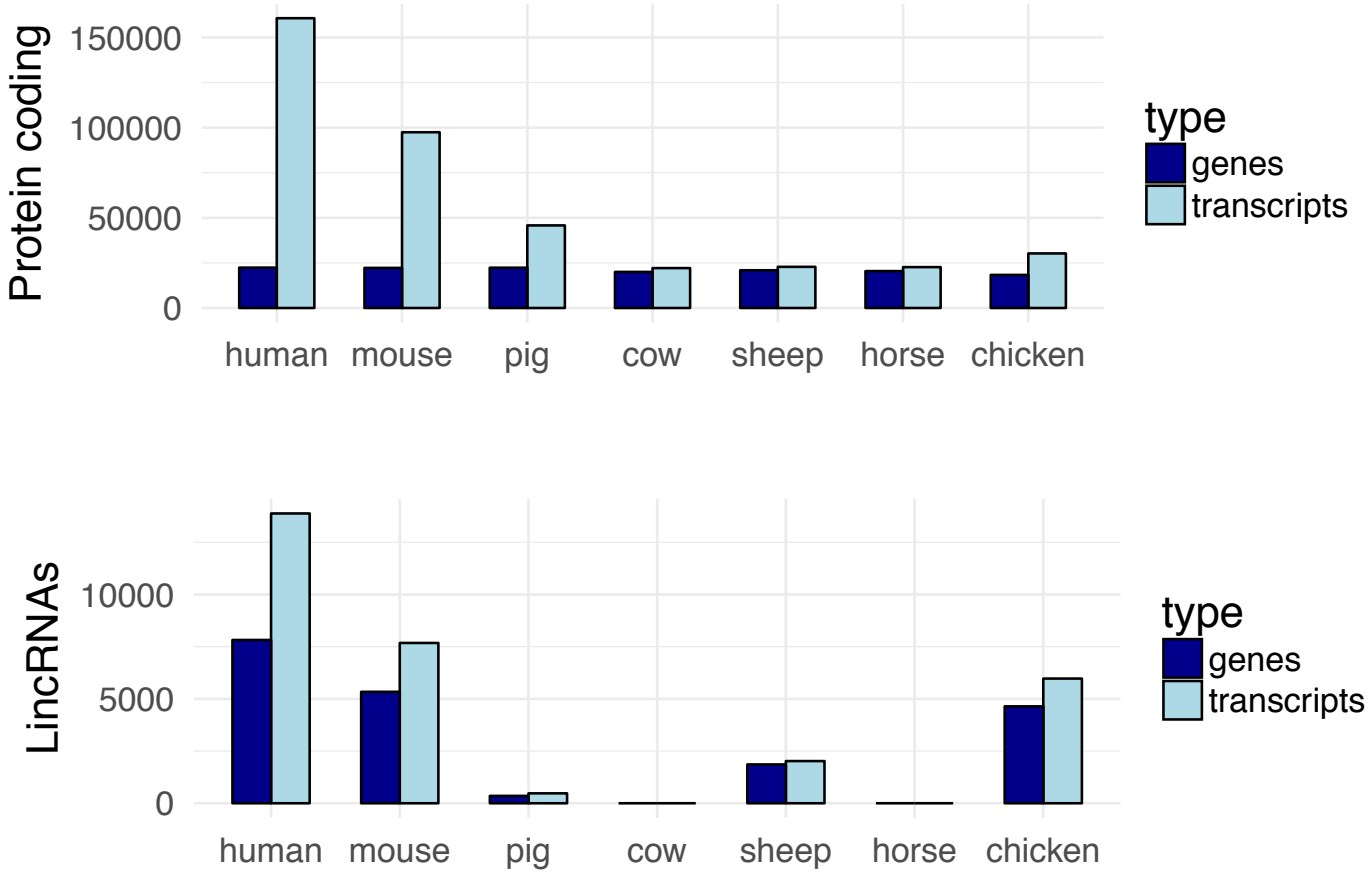
- Capped, polyadenylated, alternatively spliced transcripts
- Roles in regulation of transcription/translation, chromatin modification etc

Challenges in identification of lncRNA genes

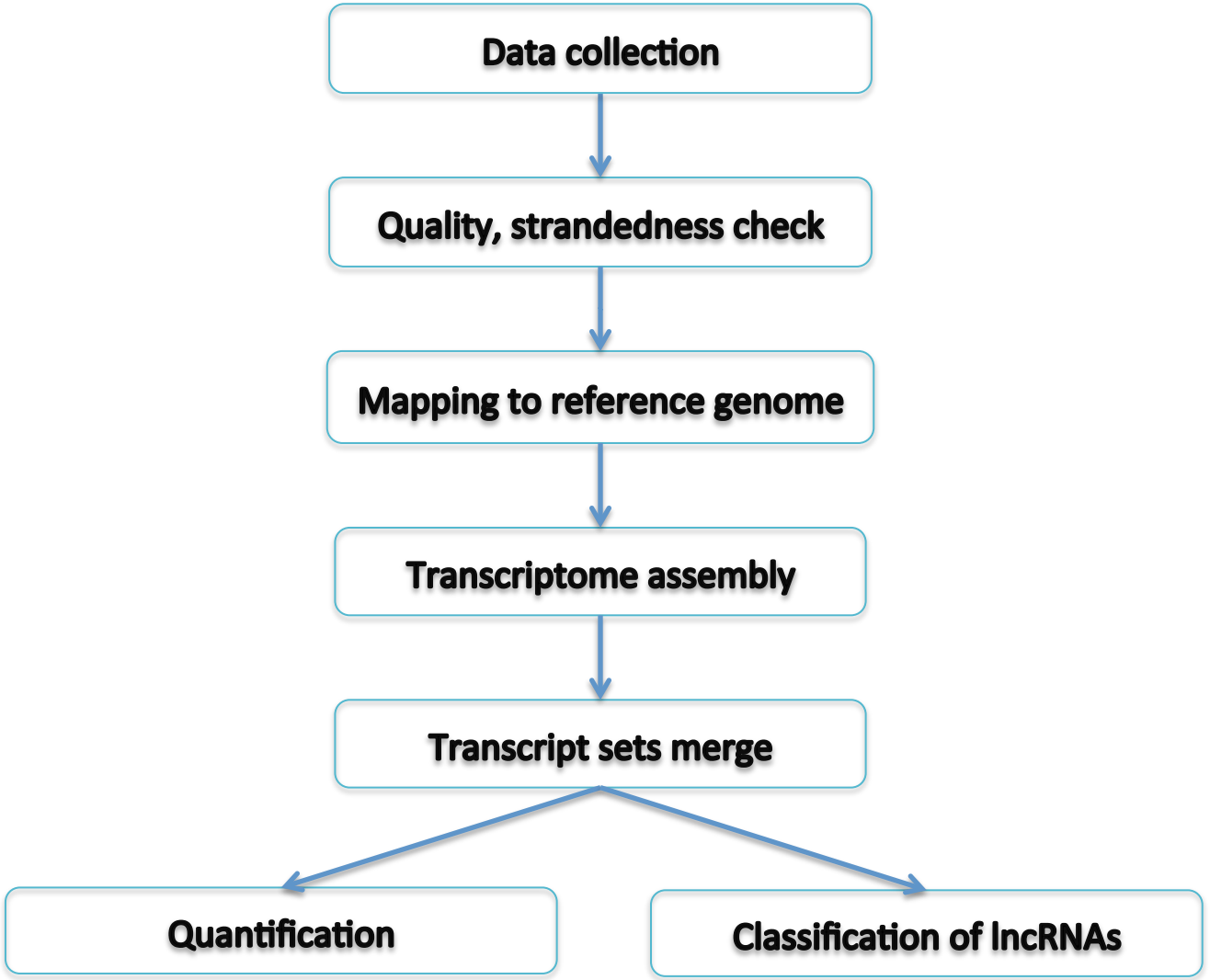
- Little sequence conservation – difficult to predict "de novo"
- Low expression levels – difficult to distinguish from transcriptional noise
- Low consistency between biological replicates

Background

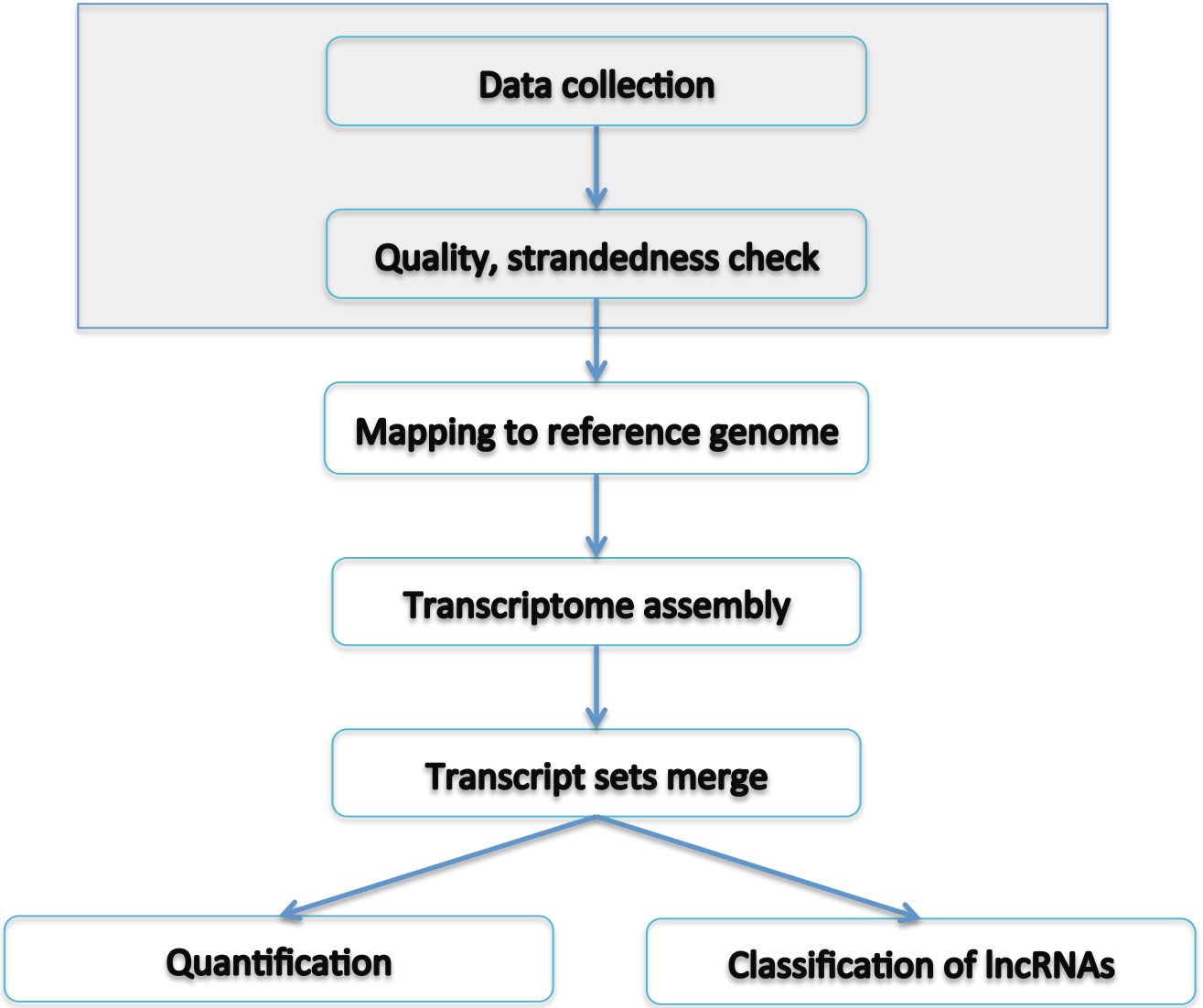
Current annotation status, Ensembl 91



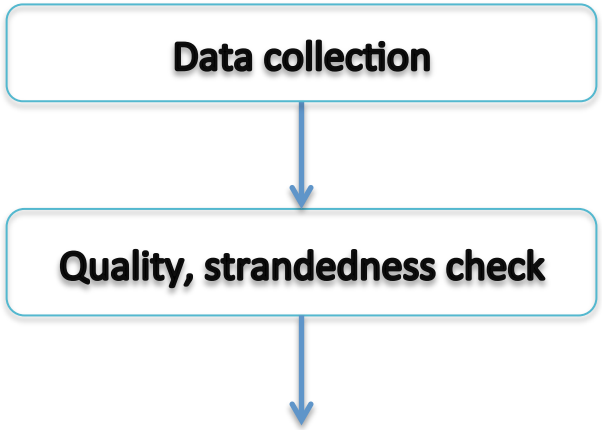
Workflow


Data collection
Quality, strandedness check
Mapping to reference genome
Transcriptome assembly
Transcript sets merge
Quantification
Classification of lncRNAs

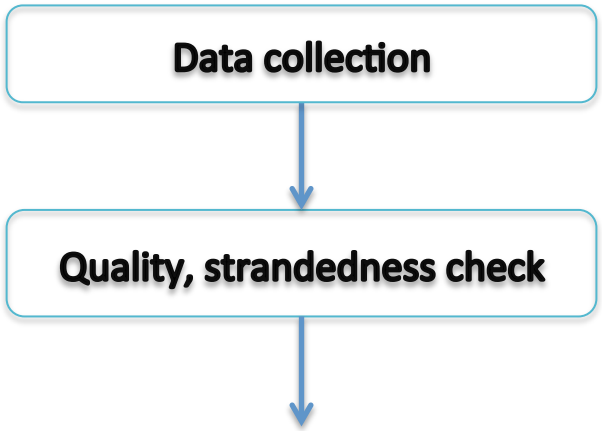
Workflow


Data collection
Quality, strandedness check
Mapping to reference genome
Transcriptome assembly
Transcript sets merge
Quantification
Classification of lncRNAs

Data collection and filtering



Data collection and filtering

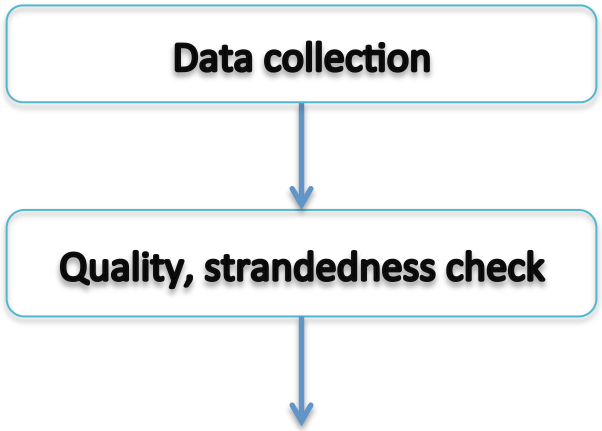


- Data obtained from the European Nucleotide Archive
- Cow, horse, pig, sheep, chicken
- Selection criteria: Illumina, paired-ended, stranded, >100 bp

[1]

[1] Patro et. Al, Nature Methods, 2017

Data collection and filtering

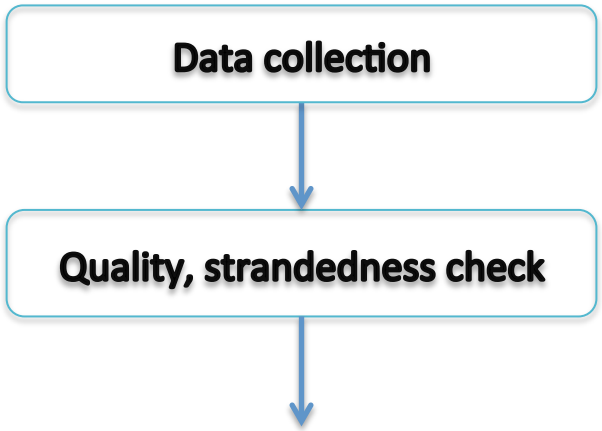


- Data obtained from the European Nucleotide Archive
- Cow, horse, pig, sheep, chicken
- Selection criteria: Illumina, paired-ended, stranded, >100 bp

- Quality check – FastQC
- Strandedness check – Salmon[1]
  - 15% samples unstranded
  - 10% samples with first read mapping to the forward strand

1 Patro et. Al, Nature Methods, 2017

Data collection and filtering



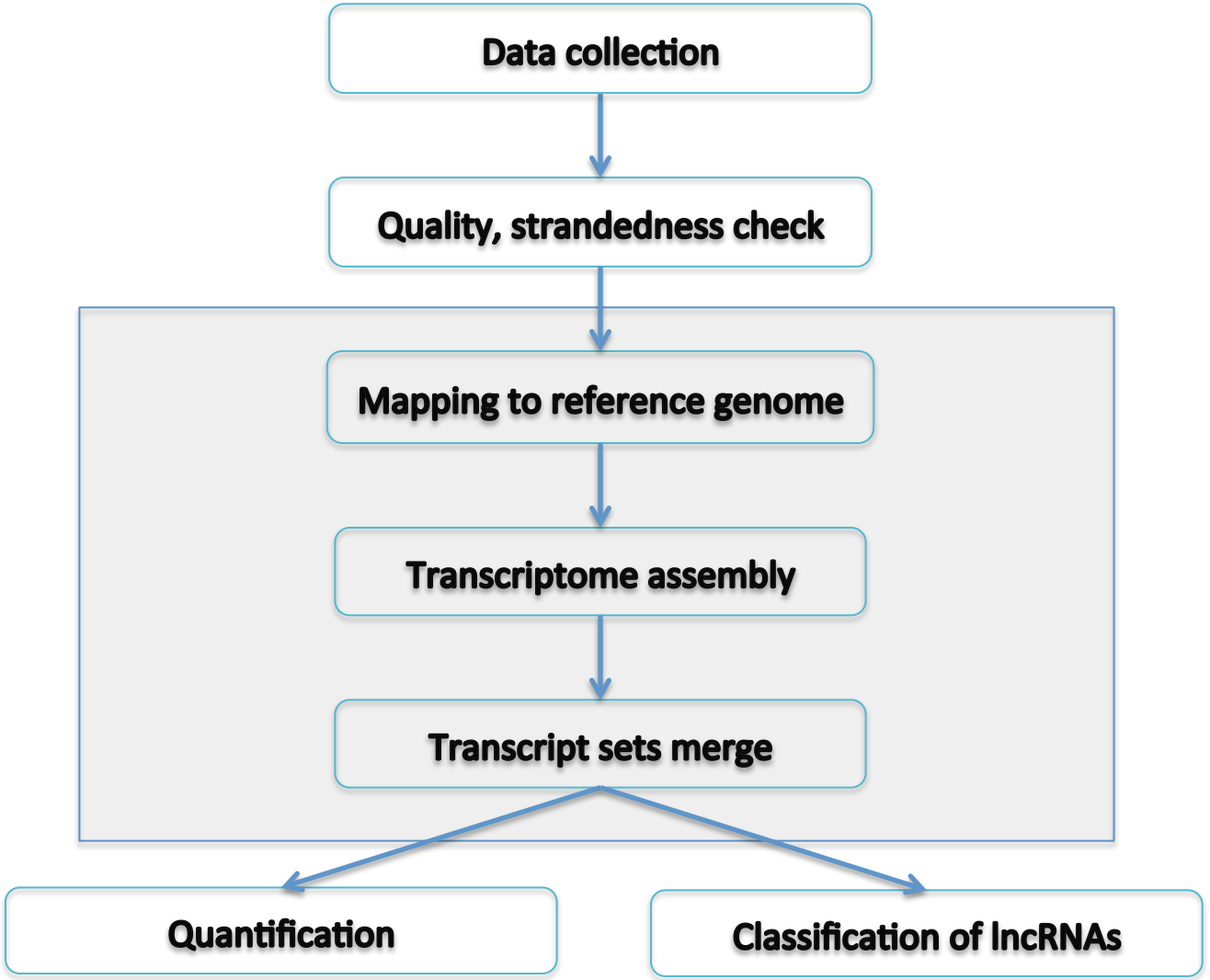
- Data obtained from the European Nucleotide Archive
- Cow, horse, pig, sheep, chicken
- Selection criteria: Illumina, paired-ended, stranded, >100 bp

- Quality check – FastQC
- Strandedness check – Salmon [1]
  - 15% samples unstranded
  - 10% samples with first read mapping to the forward strand

- ~ 900 samples, 32 Brenda tissue ontology terms
- Muscle and brain available for all species

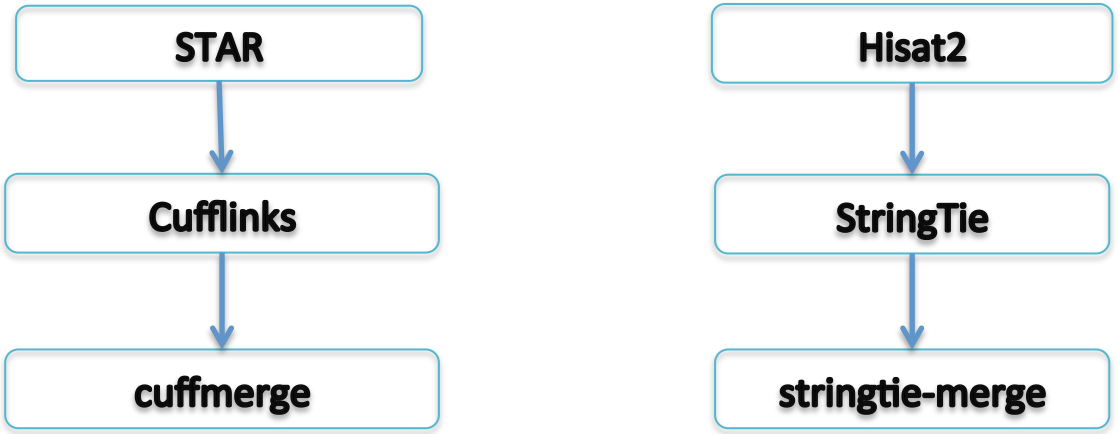
[1] Patro et. Al, Nature Methods, 2017

Workflow


Data collection
Quality, strandedness check
Mapping to reference genome
Transcriptome assembly
Transcript sets merge
Quantification
Classification of lncRNAs

Comparison of pipelines

Star[1]/Cufflinks[2](SC) vs. Hisat2/Stringtie[3](HS)



Input:

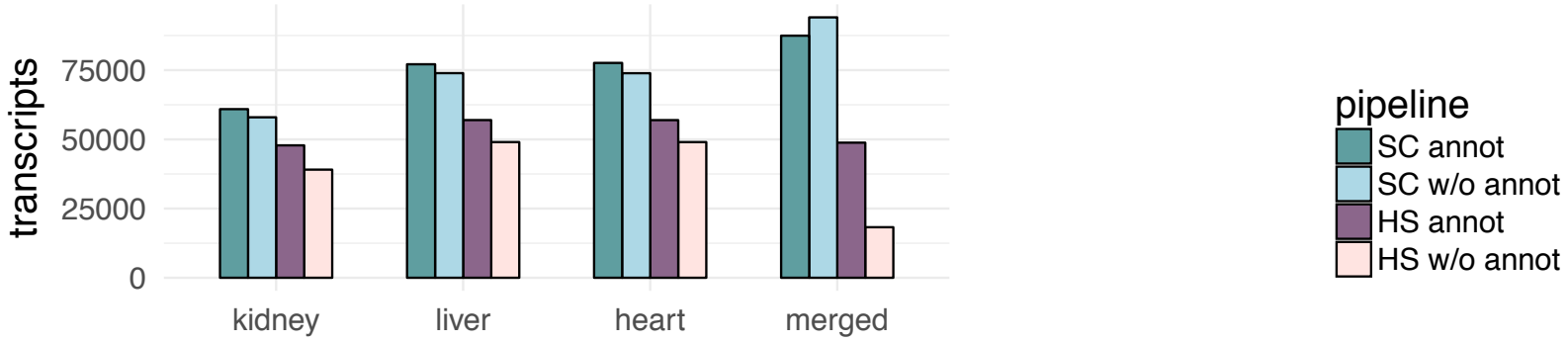
- 3 samples from chicken (kidney, liver, heart – pooled)
- 5 samples from cow (heart, cerbral cortex, spleen, liver, kidney)
- With/without annotation

[1] Dobin et. al, 2013 [2] Trapnell et.al, 2010 [3] Pertea et.al, 2016

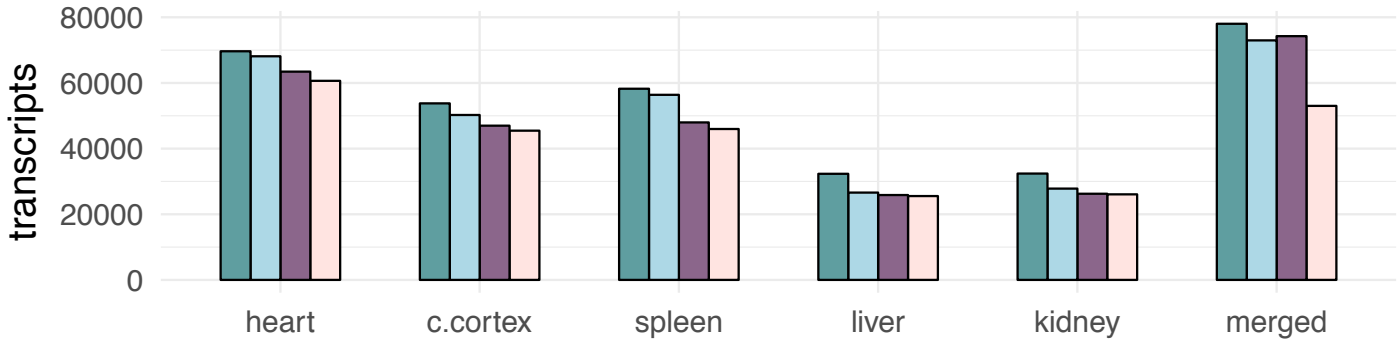
Comparison of pipelines

Total number of transcripts, per tissue and by merging all tissues

Size of the transcript sets obtained by running the two pipelines, CHICKEN



Size of the transcript sets obtained by running the two pipelines, COW

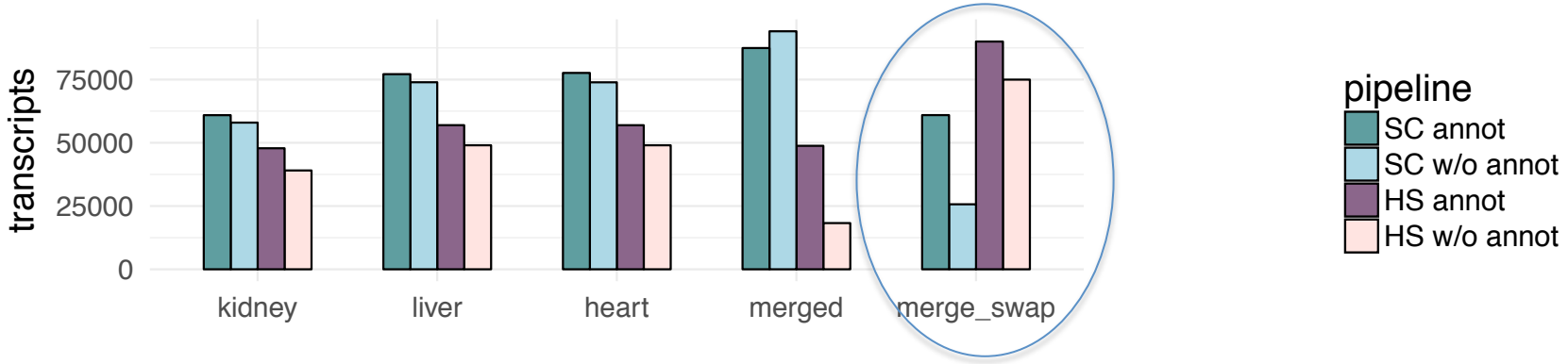


Comparison of pipelines

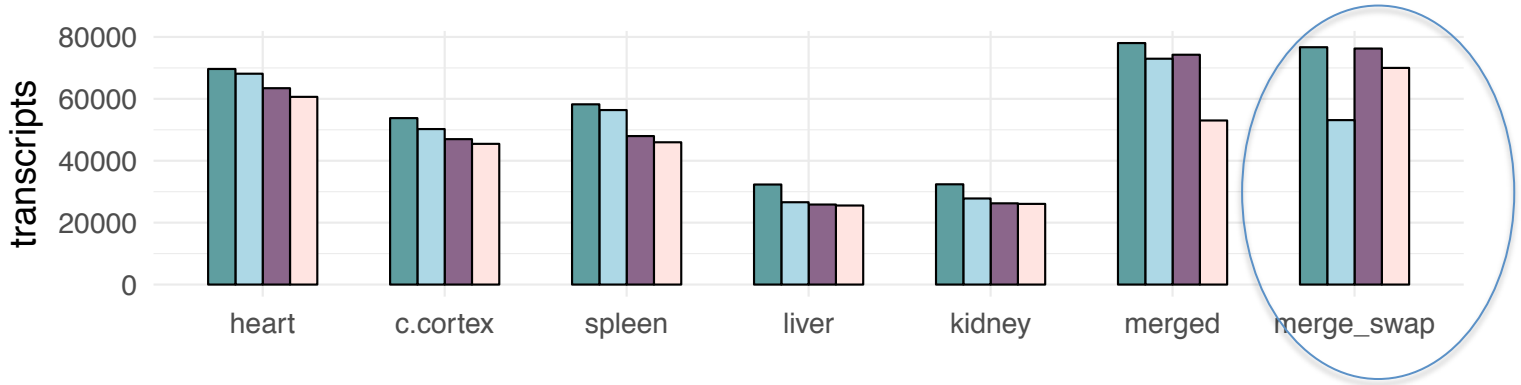
Impact of the merging step

stringtie-merge on the SC output and cuffmerge on the HS output

Size of the transcript sets obtained by running the two pipelines, CHICKEN



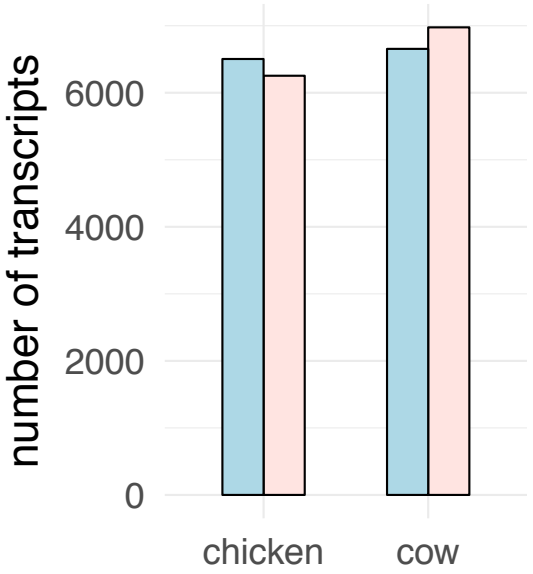
Size of the transcript sets obtained by running the two pipelines, COW



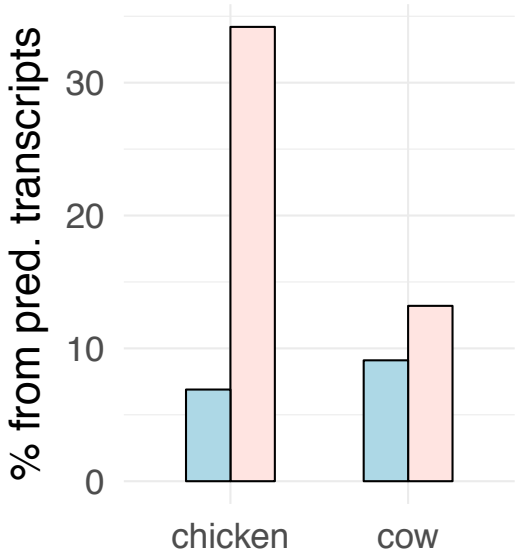
Comparison of pipelines

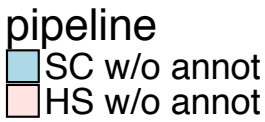
Recovery rate compared to annotation

Number of transcripts exactly overlapping the annotation



Percentage of transcripts exactly overlapping the annotation from the total transcripts predicted





Comparison of pipelines

Assessment of capacity of reconstructing lncRNAs in mouse

Data

- 30 RNA-Seq mouse samples from ENCODE/CSHL, illumina, paired-ended, 100bp
- Gencode mouse annotation down-sampled such as to contain a set of genes and transcripts similar to e.g. cow (e.g. 20.000 PCGs orthologous with cow + 3000 randomly selected PCGs + 2000 miRNas/snoRNAs)

Comparison of pipelines

Assessment of capacity of reconstructing lncRNAs in mouse

Data

- 30 RNA-Seq mouse samples from ENCODE/CSHL, illumina, paired-ended, 100bp
- Gencode mouse annotation down-sampled such as to contain a set of genes and transcripts similar to e.g. cow (e.g. 20.000 PCGs orthologous with cow + 3000 randomly selected PCGs + 2000 miRNas/snoRNAs)

Pipelines to test

- STAR>Cufflinks->cuffmerge/stringtie-merge->FEELnc
- HiSat2->Stringtie->stringtie-merge/cuffmerge->FEELnc

Comparison of pipelines

Assessment of capacity of reconstructing lncRNAs in mouse

Data

- 30 RNA-Seq mouse samples from ENCODE/CSHL, illumina, paired-ended, 100bp
- Gencode mouse annotation down-sampled such as to contain a set of genes and transcripts similar to e.g. cow (e.g. 20.000 PCGs orthologous with cow + 3000 randomly selected PCGs + 2000 miRNas/snoRNAs)

Pipelines to test

- STAR>Cufflinks->cuffmerge/stringtie-merge->FEELnc
- HiSat2->Stringtie->stringtie-merge/cuffmerge->FEELnc

Output

- Sensitivity and specificity for predicting lncRNAs and PCGs, based on the exact overlap with the remaining set of the annotation

Further work

- Synteny analysis
- Correlation of expression levels across organisms
- Structural predictions

- Public database, connected to existing repositories, e.g. EBI

Acknowledgements

- Daniel Fischer
- Sarah Djebali
- Christian Anton
- Alicja Pacholewska
- Nadezhda Doncheva
- Thomas Derrien
- Lel Eory
- Frank Panitz
- Magda Mielczarek
- Christa Kuhn
- Alan Archibald
- Jan Gorodkin