Monopoly ANR project :

Integrated gene network analyses between mono- and poly-ovulating species

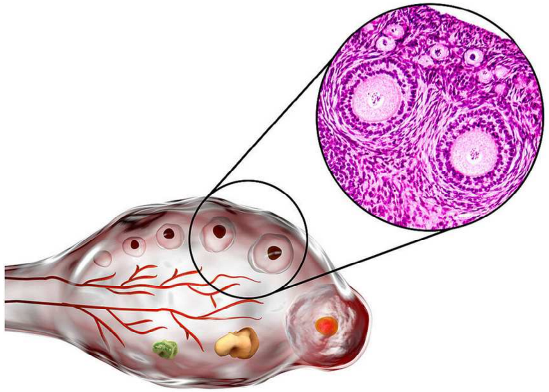
S. Maman, G. Agoutin, J. Sarry, F. Plisson-Petit, S. Fabre, C. Genêt

GenROC team

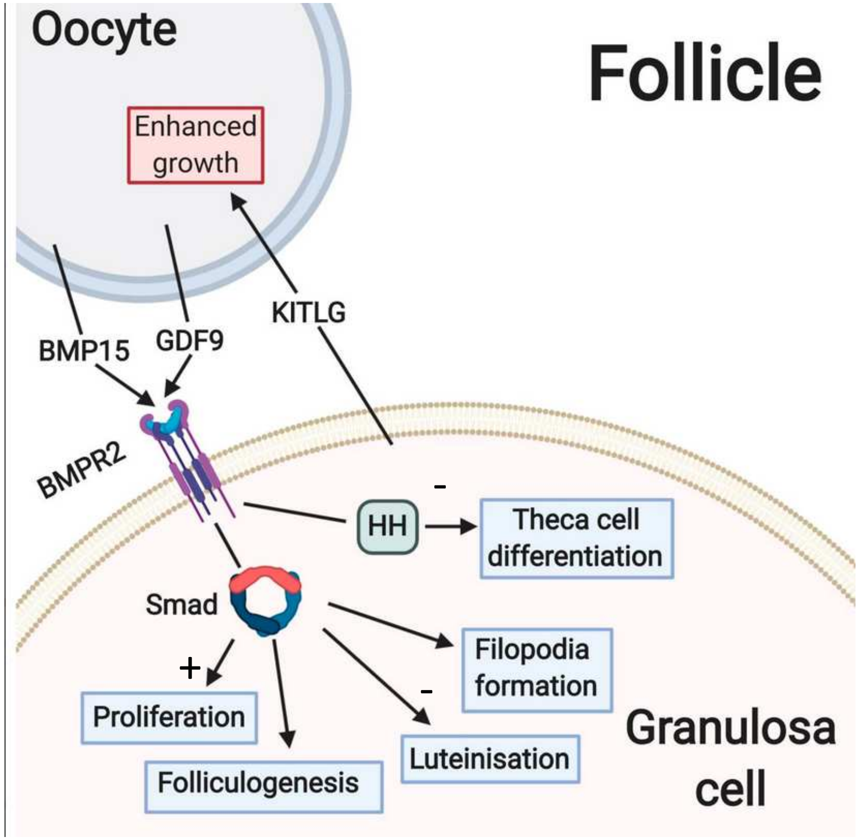
Context

Folliculogenesis is important for the development and maintenance of fertility

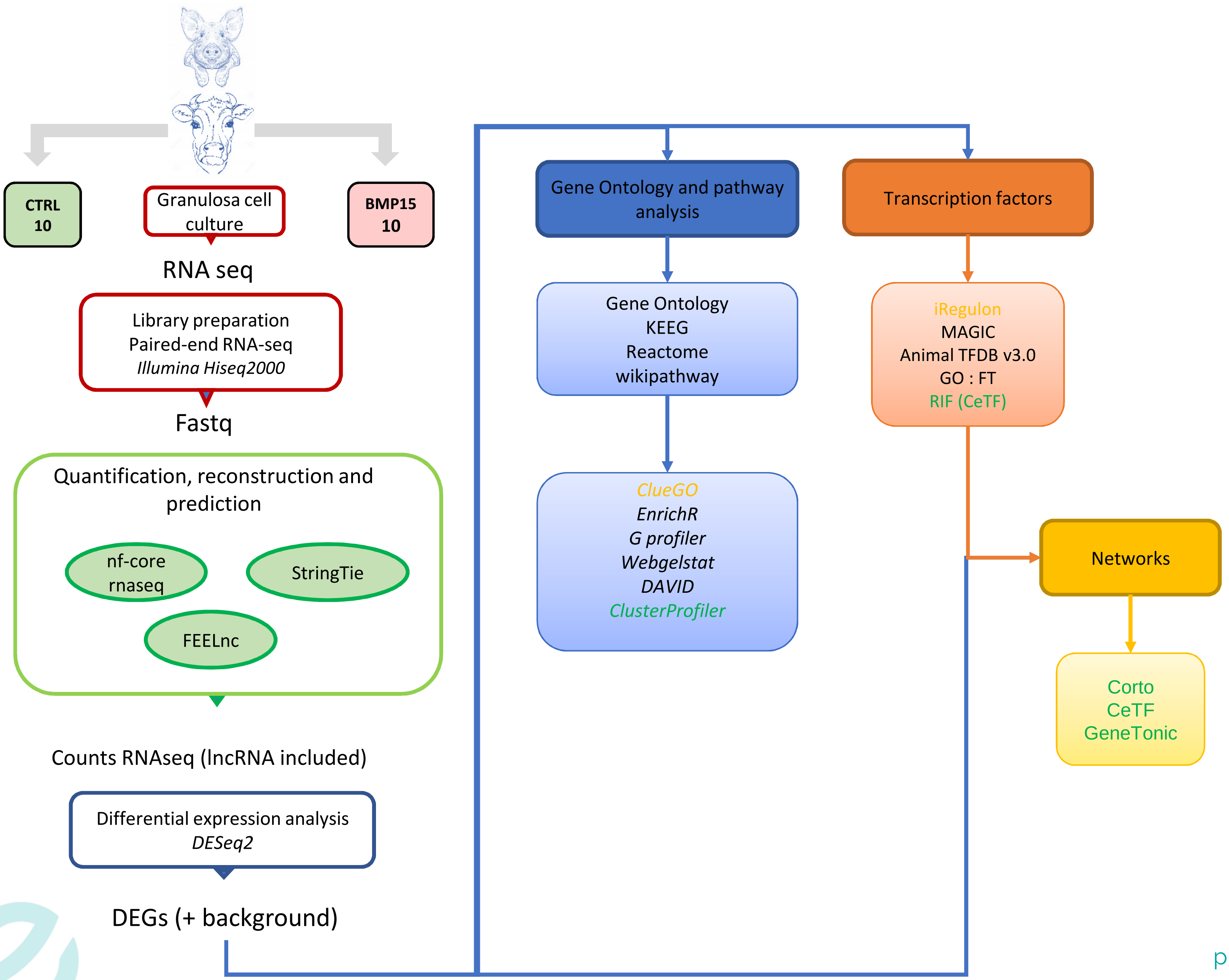
Ovarian microcosmos.



- BMP15 protein
- Secreted ligand of the TGF-β superfamily
- Involved in folliculogenesis
- Contributed to paracrine dialog



From Jozkowiak et al., 2020



CTRL 10

Granulosa cell culture

BMP15 10

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

FEELnc

Counts RNAseq (lncRNA included)

Differential expression analysis
DESeq2

DEGs (+ background)

Gene Ontology and pathway analysis

Gene Ontology
KEEG
Reactome
wikipathway

ClueGO
EnrichR
G profiler
Webgelstat
DAVID
ClusterProfiler

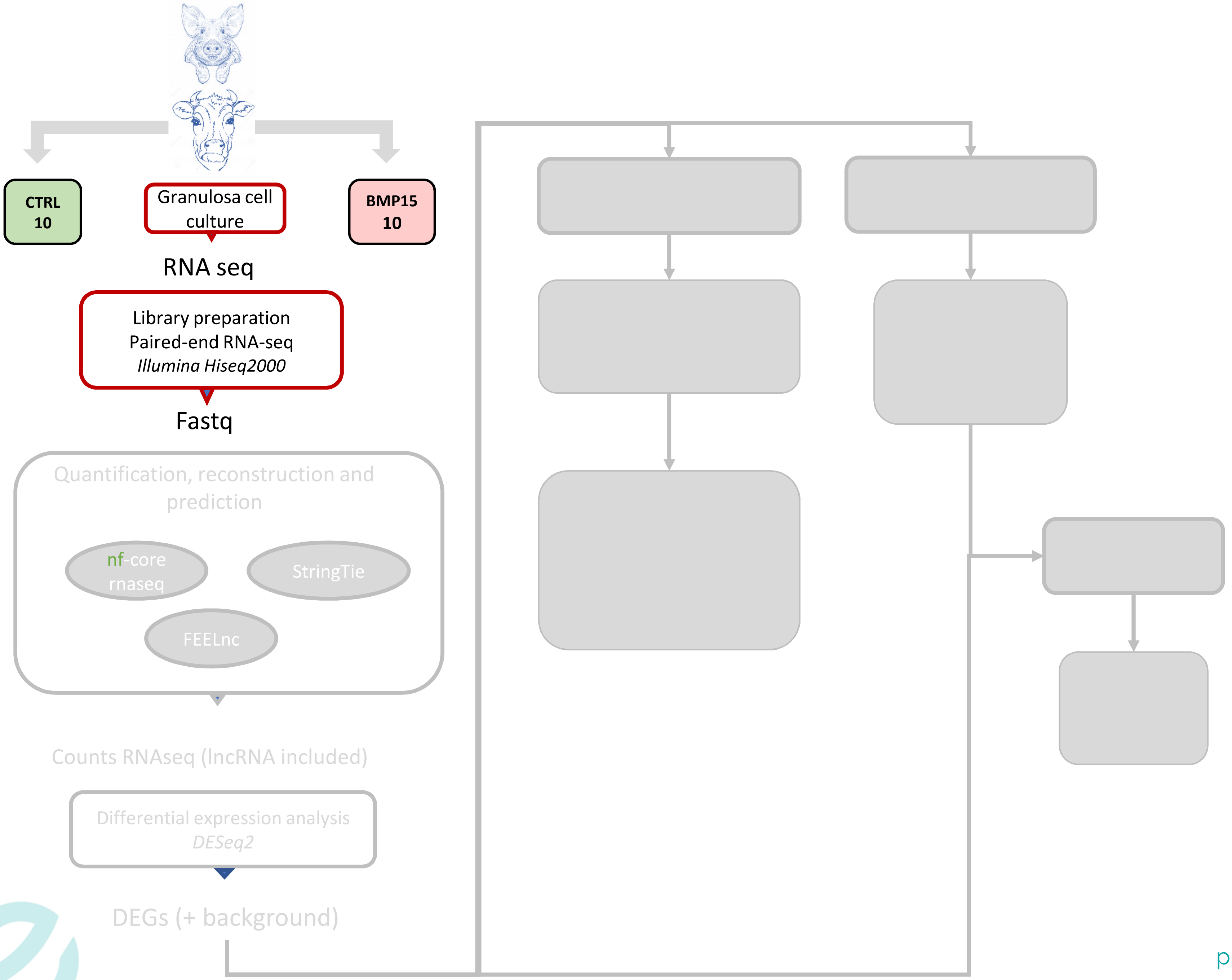
Transcription factors

iRegulon
MAGIC
Animal TFDB v3.0
GO : FT
RIF (CeTF)

Networks

Corto
CeTF
GeneTonic

Web ; Cytoscape ; R package



CTRL
10

Granulosa cell culture

BMP15
10

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

FEELnc

Counts RNAseq (lncRNA included)

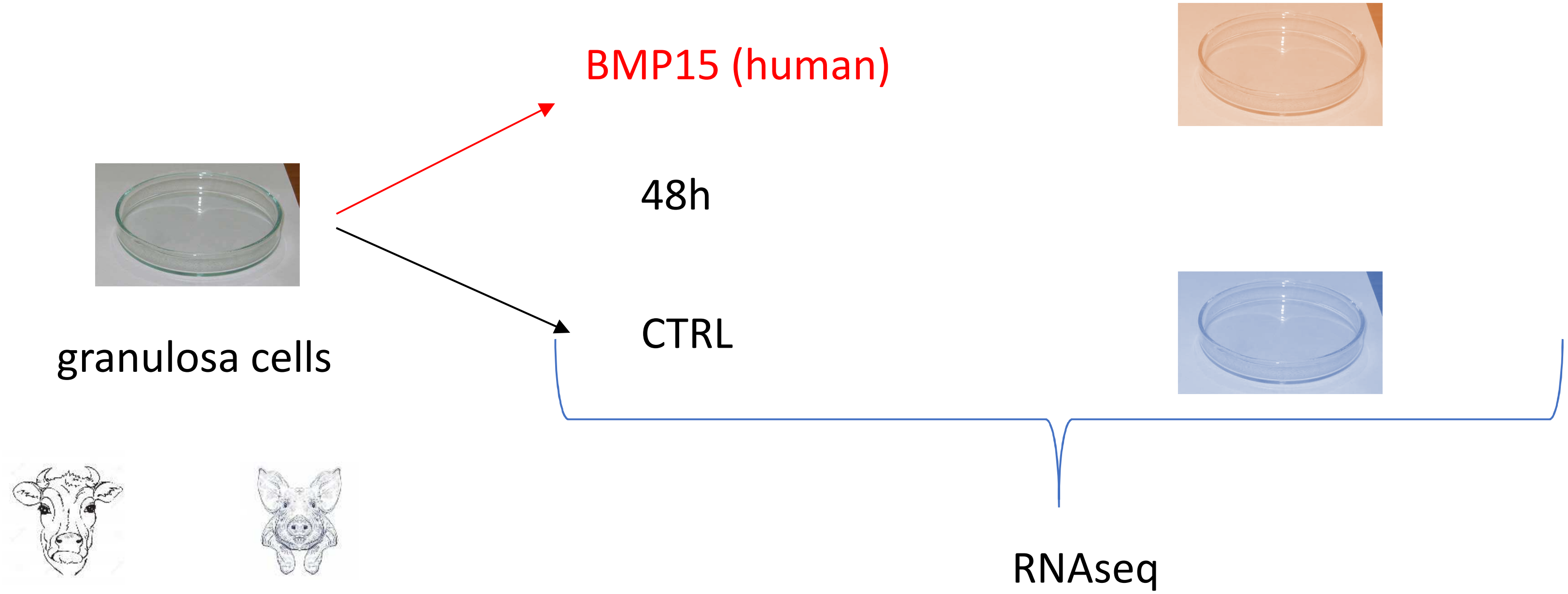
Differential expression analysis
DESeq2

DEGs (+ background)

Web ; Cytoscape ; R package

Monopoly ANR project (2010 S. Fabre)

Decipher BMP15 response in granulosa cells of cow and sow (mono vs poly-ovulating species)



BMP15 (human)

48h

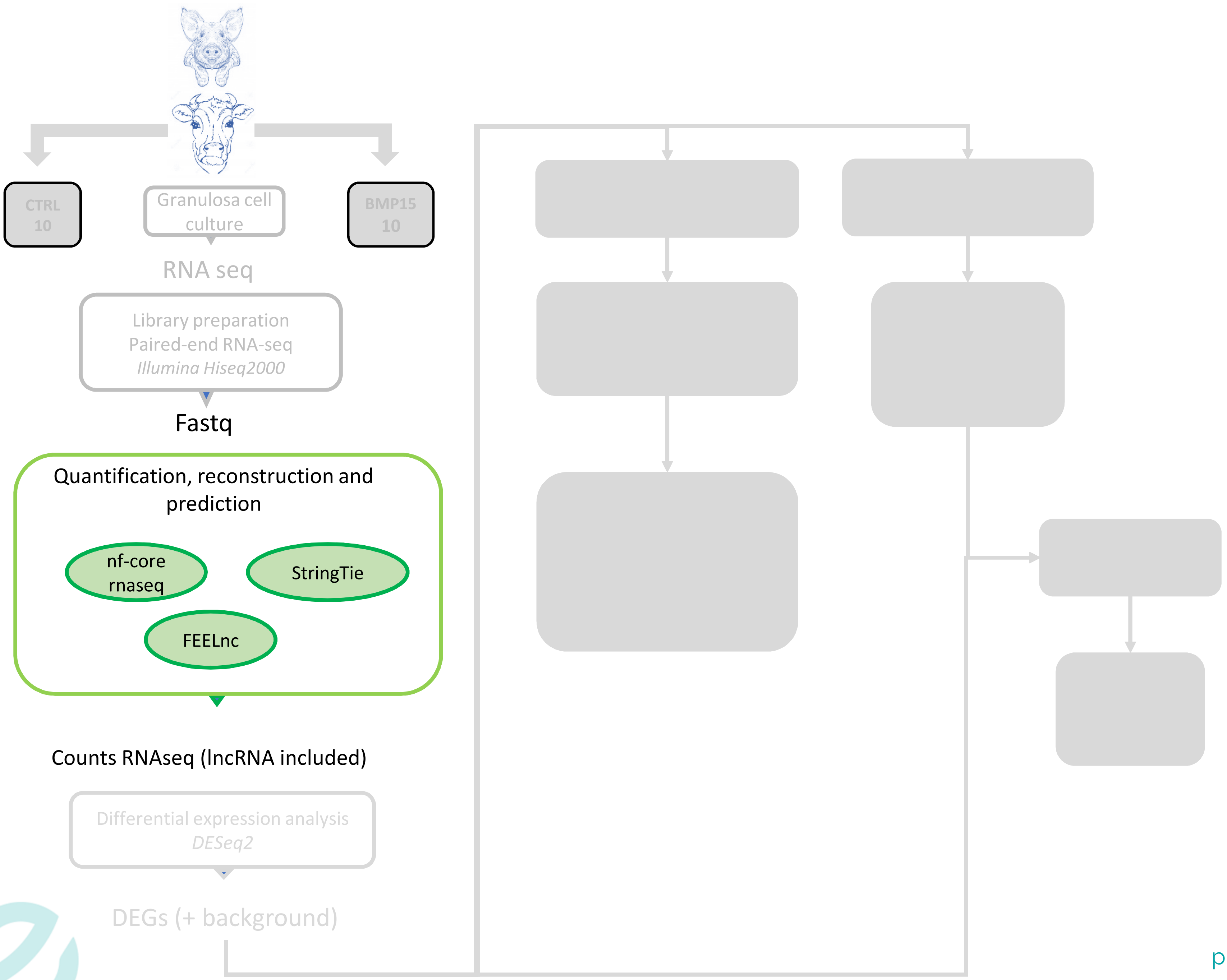
CTRL

granulosa cells

RNAseq

20 samples by species (10 CTRL vs 10 BMP15)

40 sequencing lanes (0,94<Pearson Correlation<0,98)



CTRL
10

Granulosa cell culture

BMP15
10

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

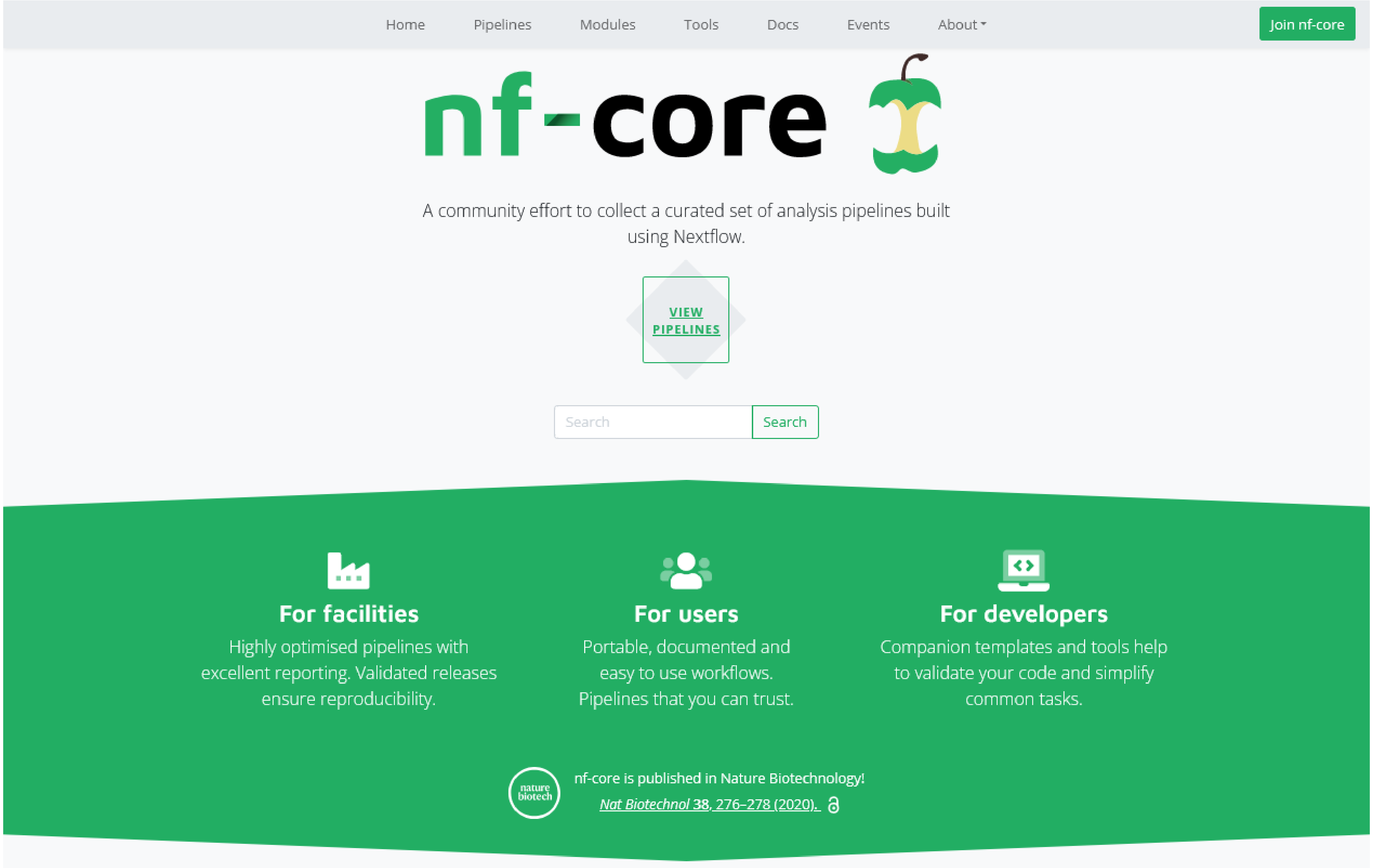
FEELnc

Counts RNAseq (lncRNA included)

Differential expression analysis
DESeq2

DEGs (+ background)

Web ; Cytoscape ; R package



A community effort to collect a curated set of analysis pipelines built using Nextflow.

For facilities

Highly optimised pipelines with excellent reporting. Validated releases ensure reproducibility.

For users

Portable, documented and easy to use workflows. Pipelines that you can trust.

For developers

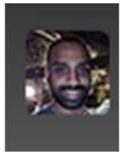
Companion templates and tools help to validate your code and simplify common tasks.

nf-core is published in Nature Biotechnology!

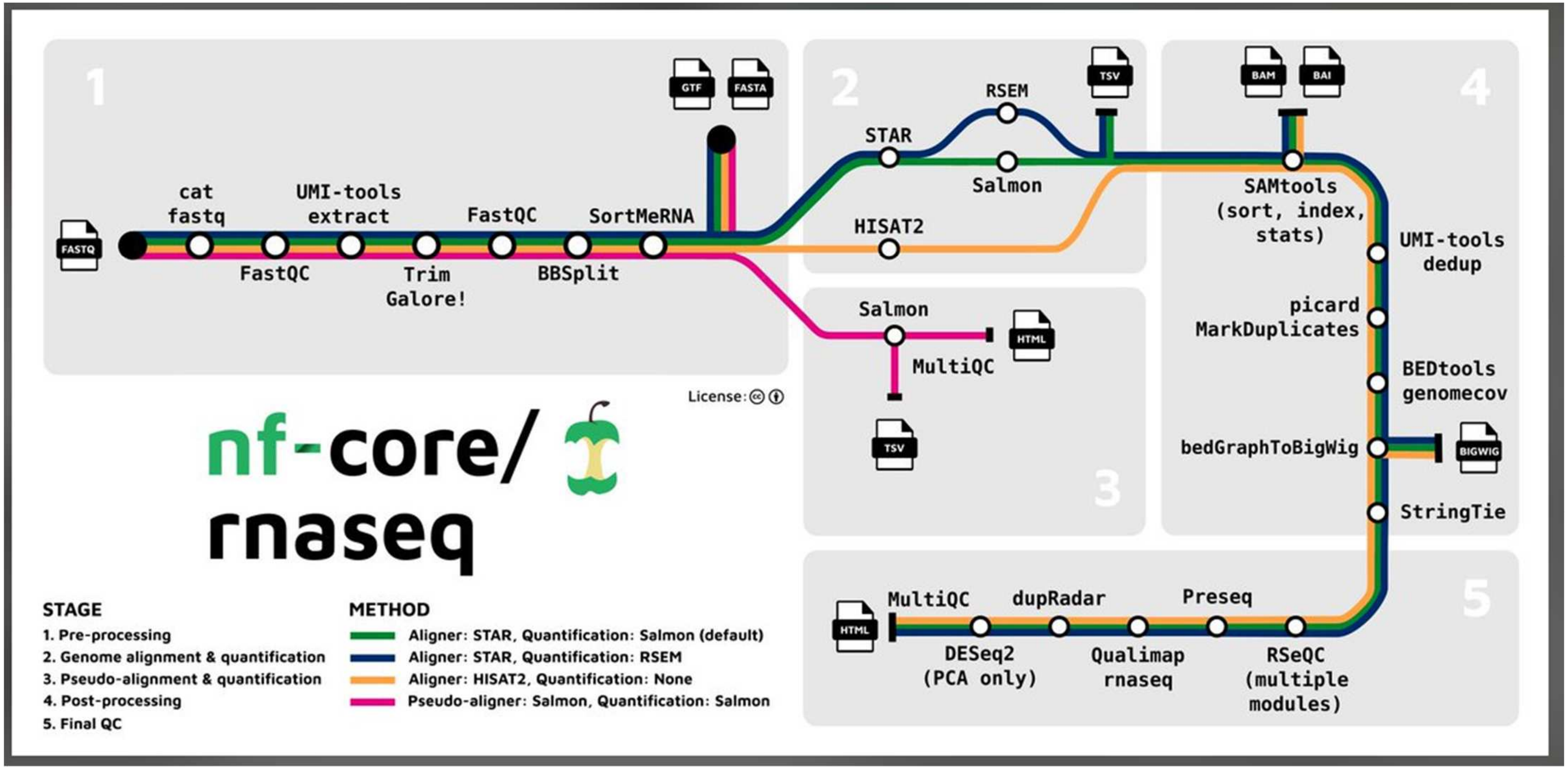
Nat Biotechnol 38, 276–278 (2020).

Pipeline nf-core rnaseq

nf-core



hpatel



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

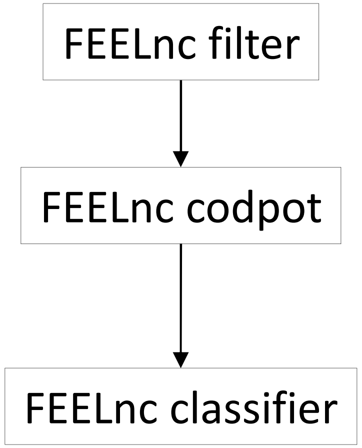
METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟪 Pseudo-aligner: Salmon, Quantification: Salmon

MONOPOLY : BLUE method

Feelnc (prediction only !!)

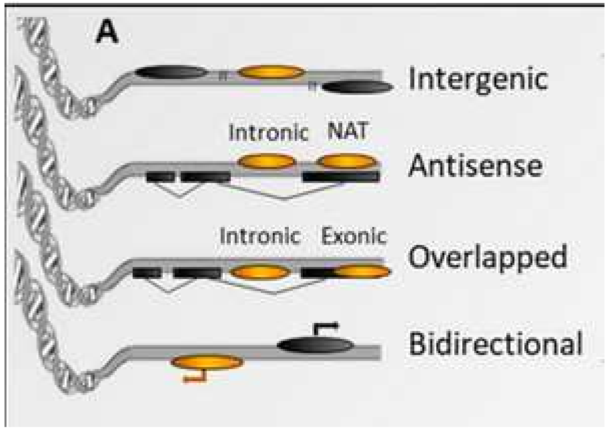
https://github.com/tderrien/FEELnc



Extract, filter candidate transcripts

Compute the coding potential of candidates

Classify lncRNAs based on their genomic locations

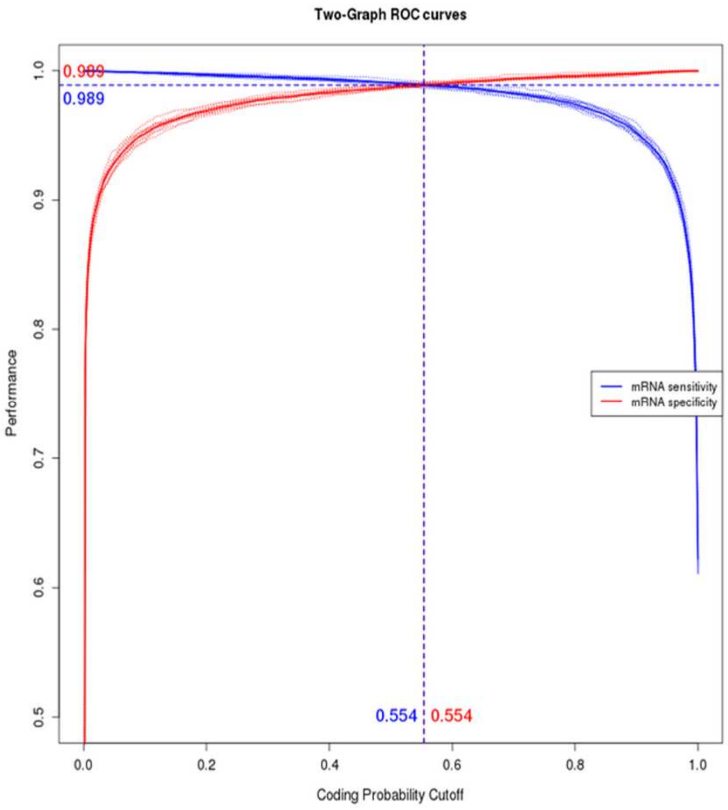


Salviano-Silva et al., 2018, *Non-coding RNA*

FEELnc Wucher et al., 2017, *NAR*

Results FEELnc

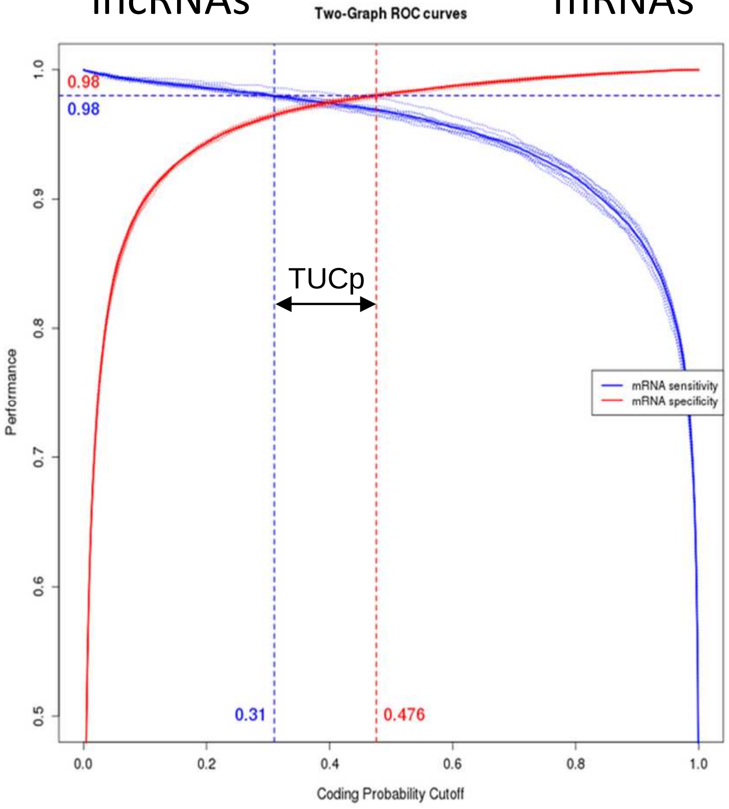
intergenic

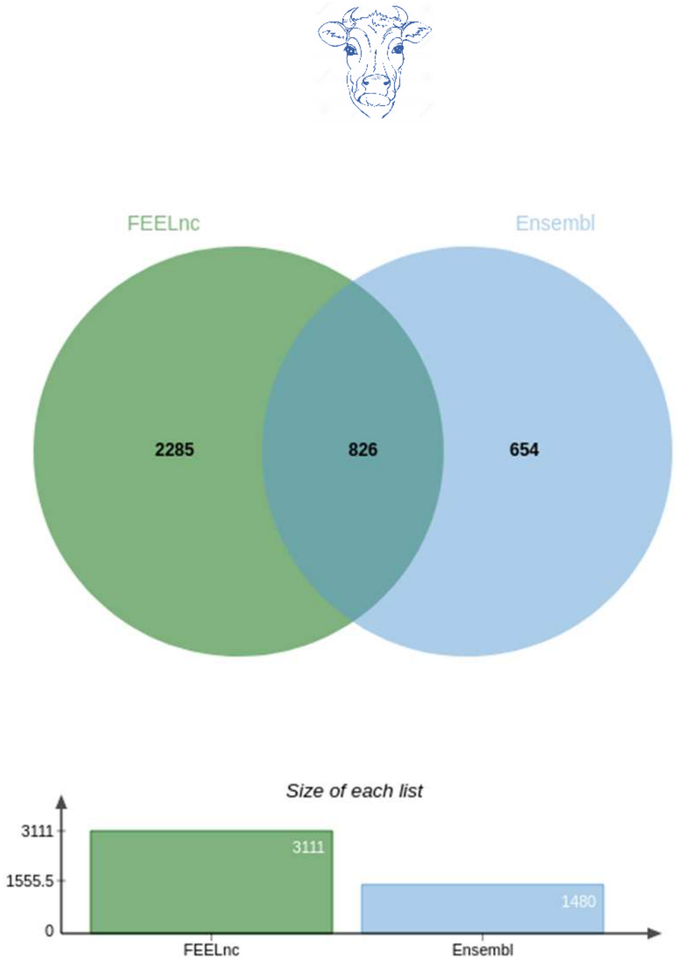


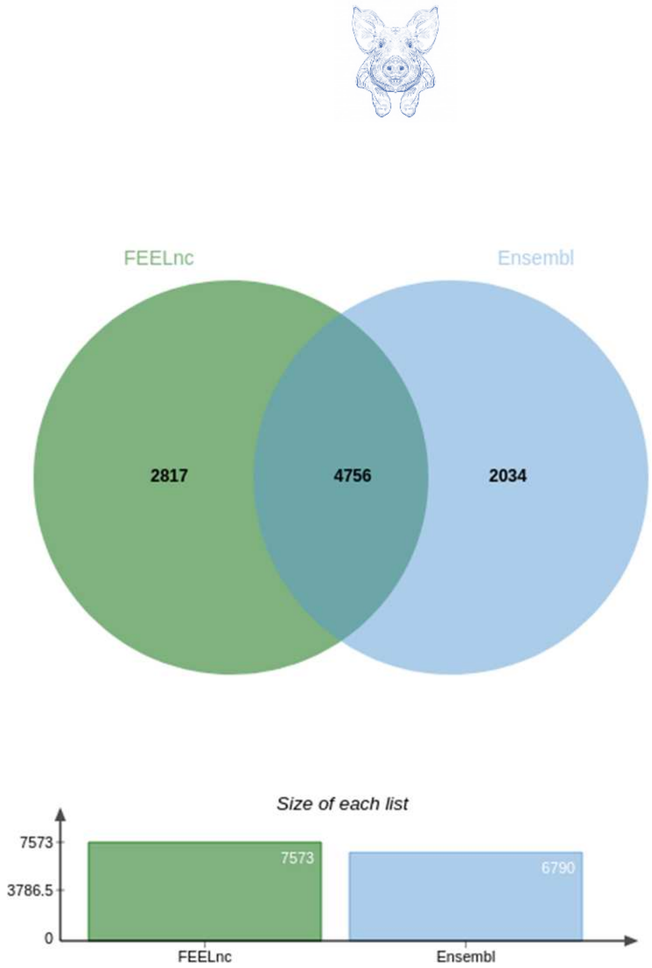
shuffle

lncRNAs

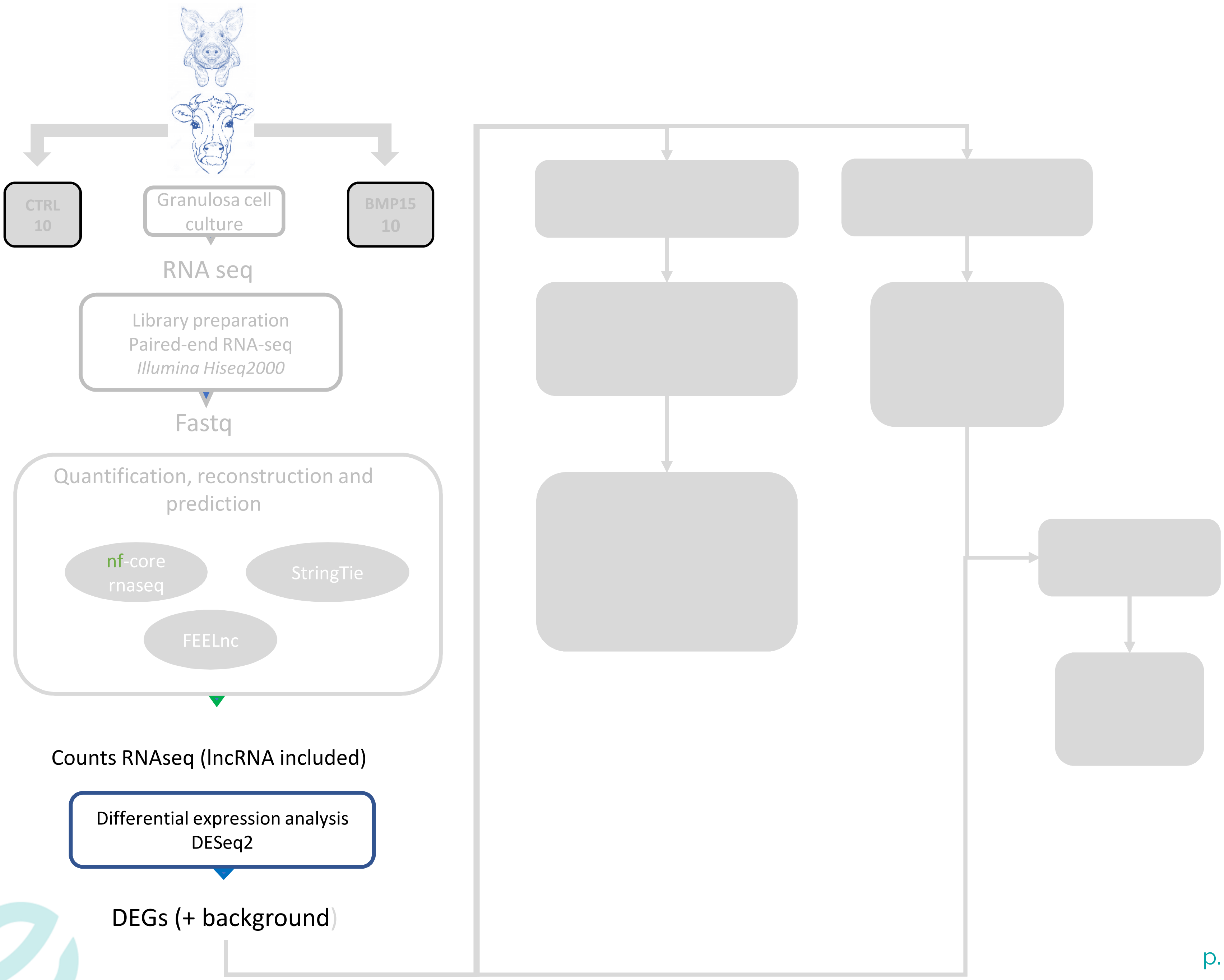
mRNAs







Venn diagram comparing the lncRNAs already known by ENSEMBL and those predicted by FEELnc



CTRL
10

Granulosa cell culture

BMP15
10

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

FEELnc

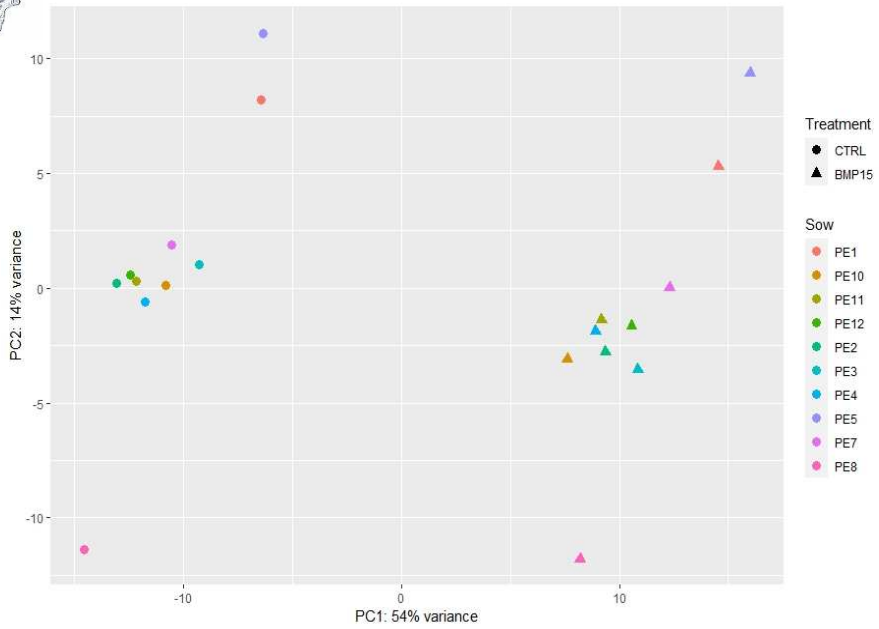
Counts RNAseq (lncRNA included)

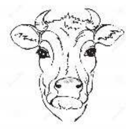
Differential expression analysis
DESeq2

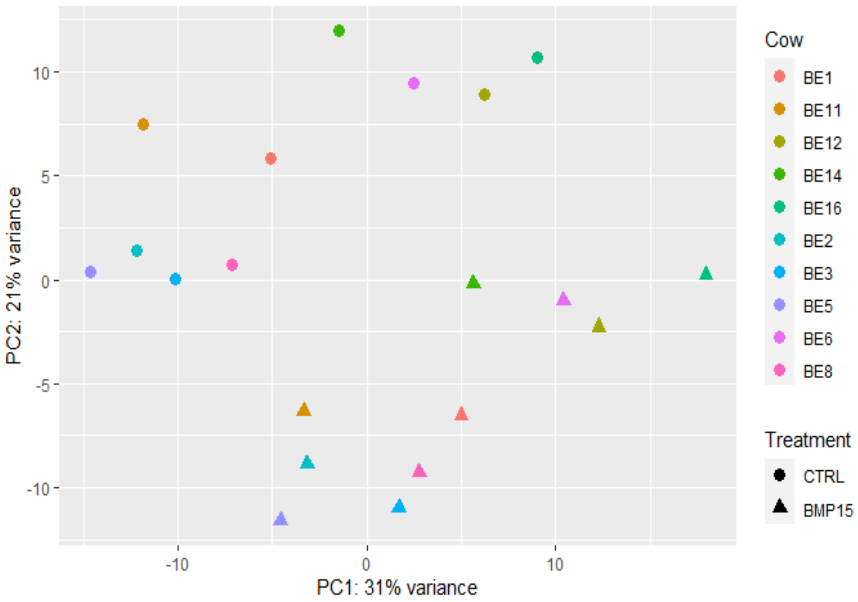
DEGs (+ background)

Web ; Cytoscape ; R package

PCA : paired samples

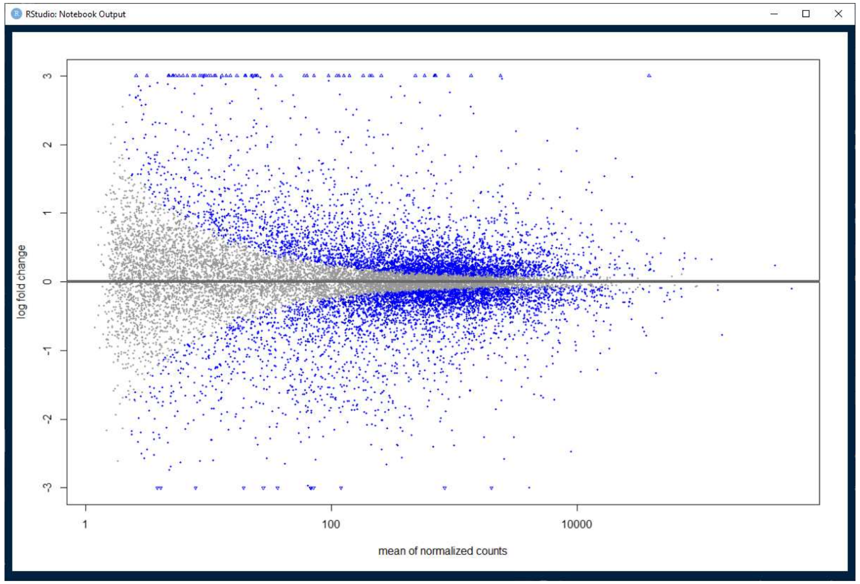




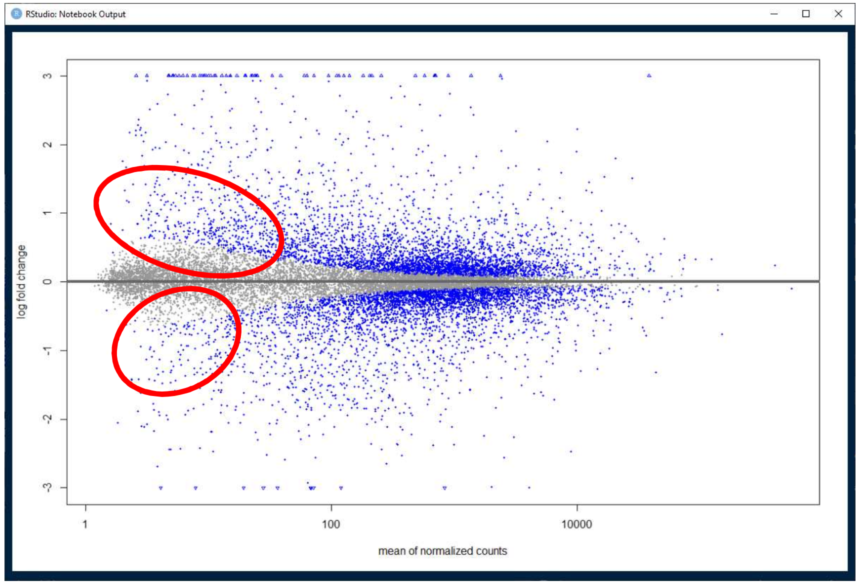


Differential analysis

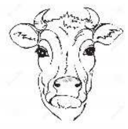
DESeq2 Shrunken log2FC, paired samples



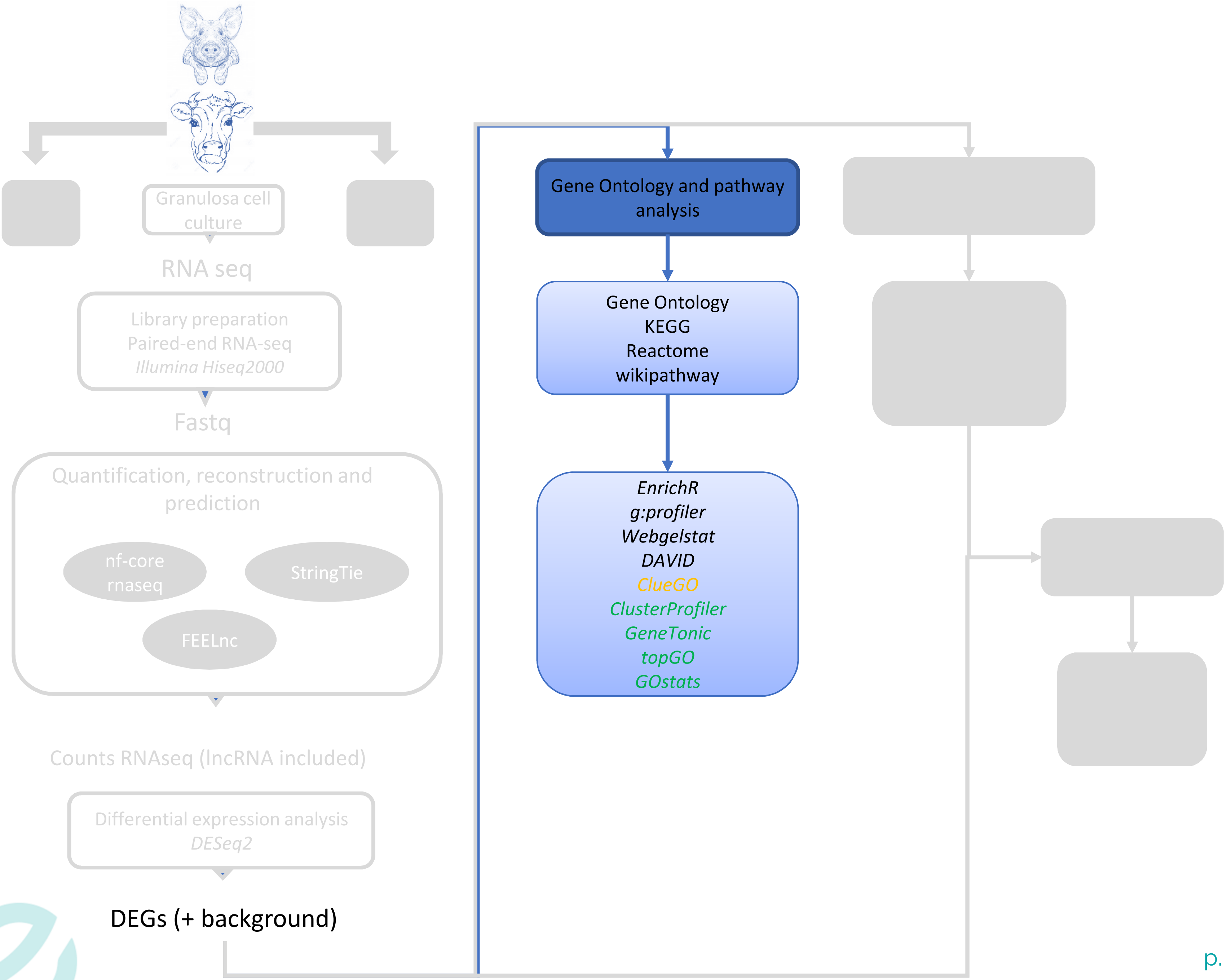
Apeglm package



Identification of DEGs



<table>
<tr><th></th><th></th><th></th></tr>
<tr><td>15 164 expressed genes</td><td>(rowsum(counts(dds)>1)>10</td><td>16 135 expressed genes</td></tr>
<tr><td>4 521 DEGs</td><td>padj<0,05</td><td>7 909 DEGs</td></tr>
<tr><td>240 up ; 203 down</td><td>padj<0,05 et |Log2FC|>1</td><td>487 up ; 521 down</td></tr>
<tr><td>443 DEGs</td><td></td><td>1 008 DEGs</td></tr>
</table>



Granulosa cell culture

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

FEELnc

Counts RNAseq (lncRNA included)

Differential expression analysis
DESeq2

DEGs (+ background)

Gene Ontology and pathway analysis

Gene Ontology
KEGG
Reactome
wikipathway

EnrichR
g:profiler
Webgelstat
DAVID
ClueGO
ClusterProfiler
GeneTonic
topGO
GOstats

Web ; Cytoscape ; R package

GeneOntology – functional enrichment

ORA : Over representation analysis (input = DEGs), is based on a hypergeometric test (Fisher's exact test) (Boyle et al., 2004).

GSEA : Gene Set Enrichment Analysis (All expressed genes, ranked), is based on an enrichment score (Subramanian et al., 2005)

DAVID
EnrichR
g:profiler
Webgelstat
ClueGO
GeneTonic
topGO
Gostats
ClusterProfiler

DAVID : https://david.ncifcrf.gov/

Enrichr: https://amp.pharm.mssm.edu/Enrichr/

g:profiler: https://biit.cs.ut.ee/gprofiler/

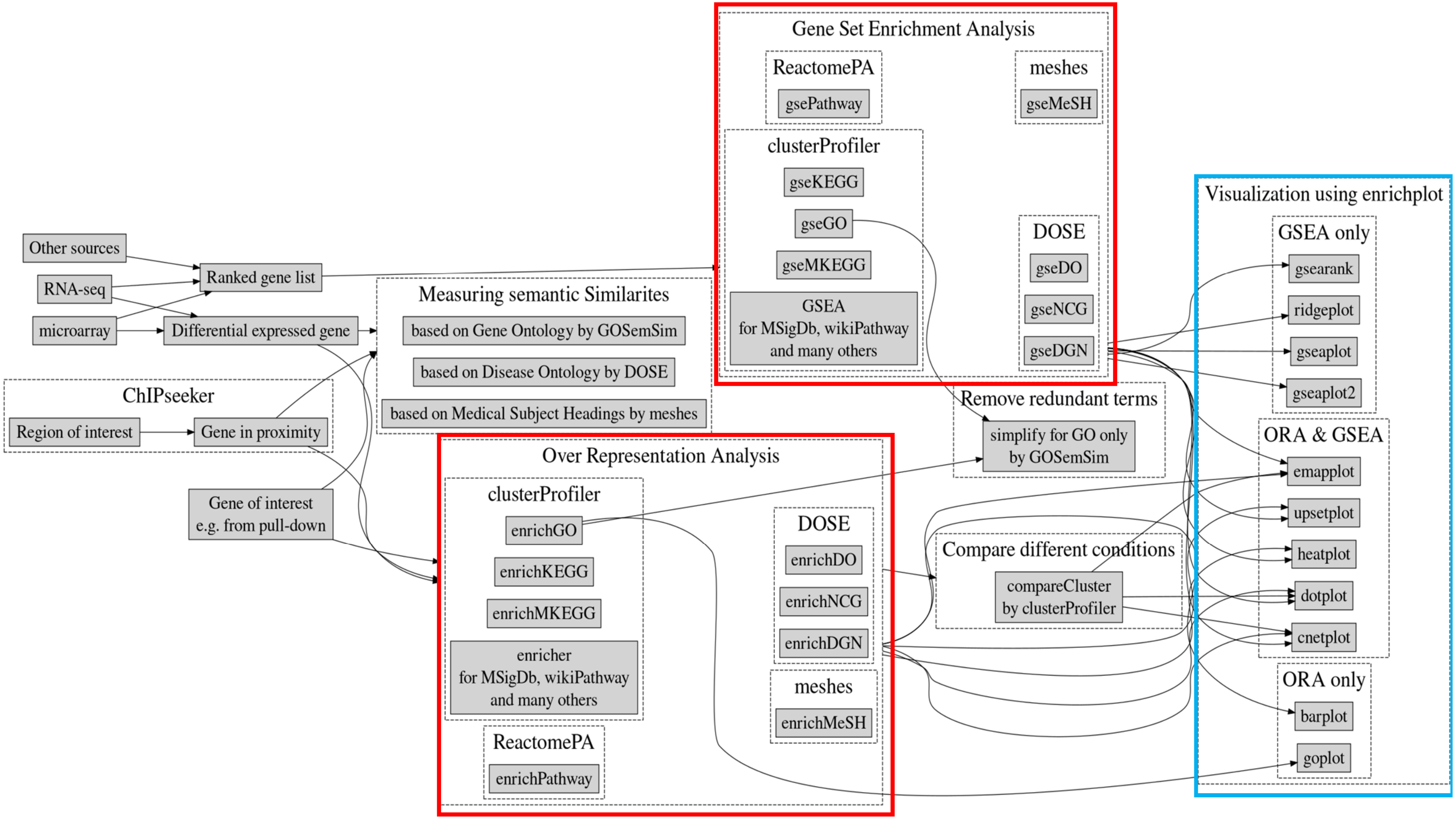
Webgelstat : http://www.webgestalt.org/

TopGO and fgsea R packages

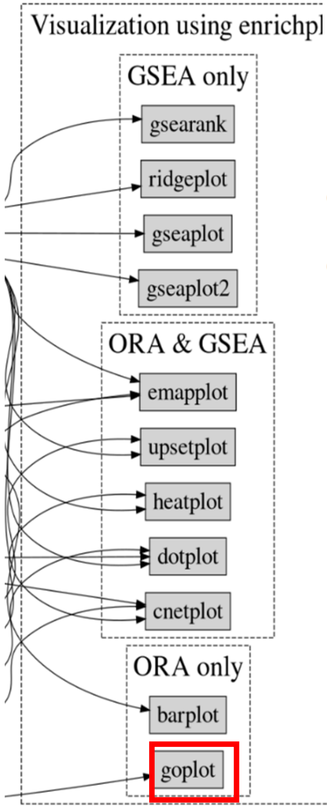
Package *GeneTonic* (Marini et al., 2021)

Package *ClusterProfiler* (Wu et al, 2021)

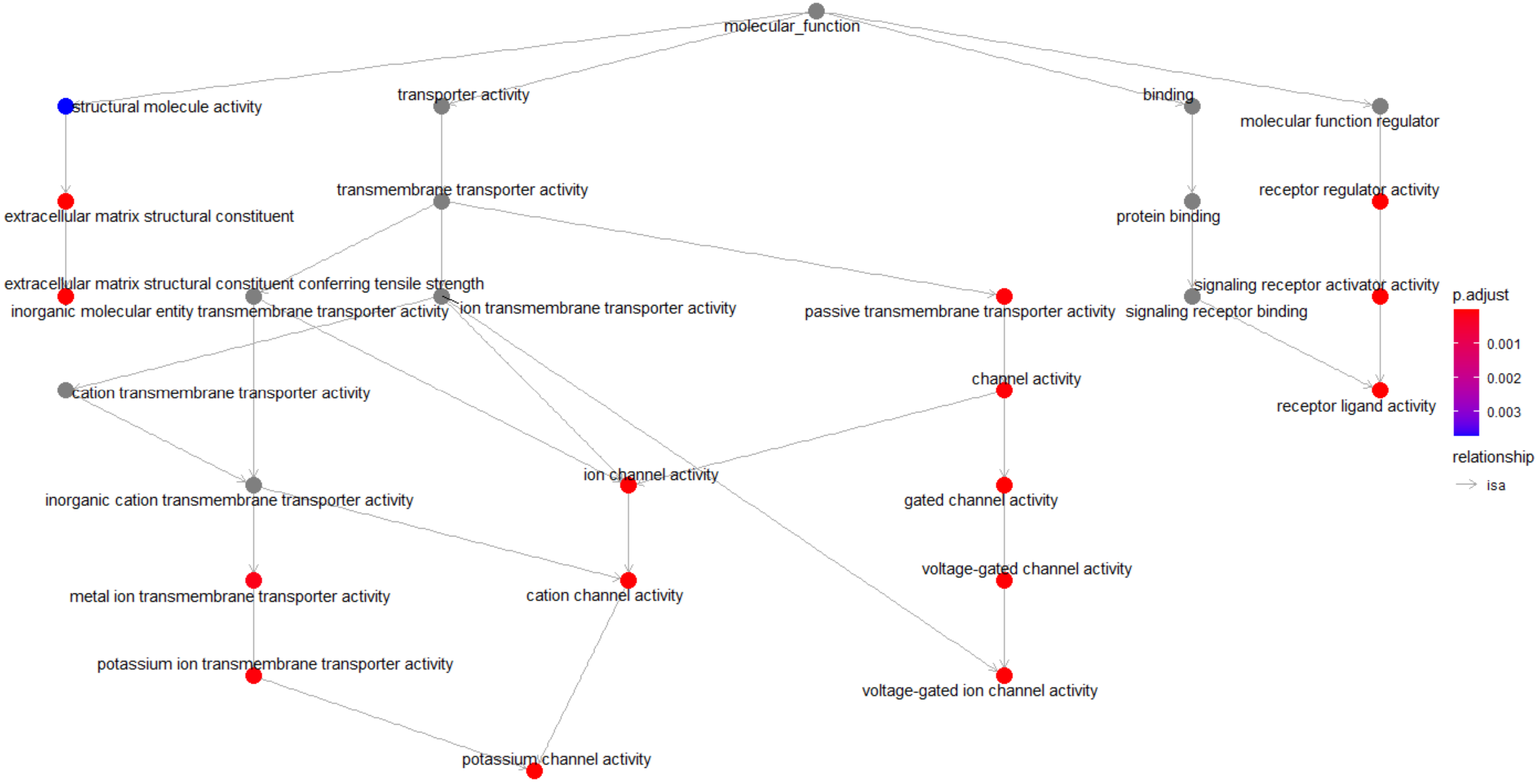
ClusterProfiler package Guangchuang Yu



GO plot

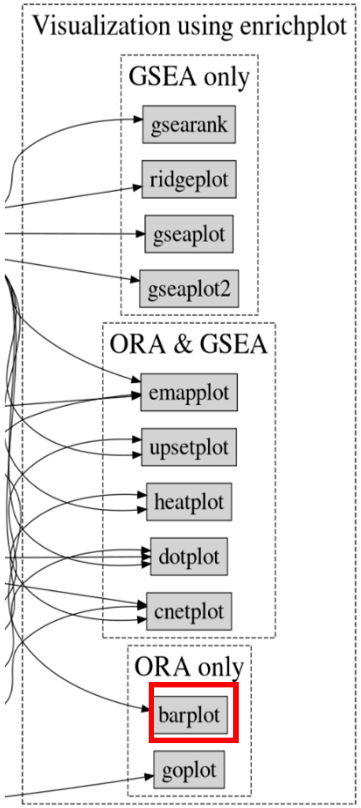


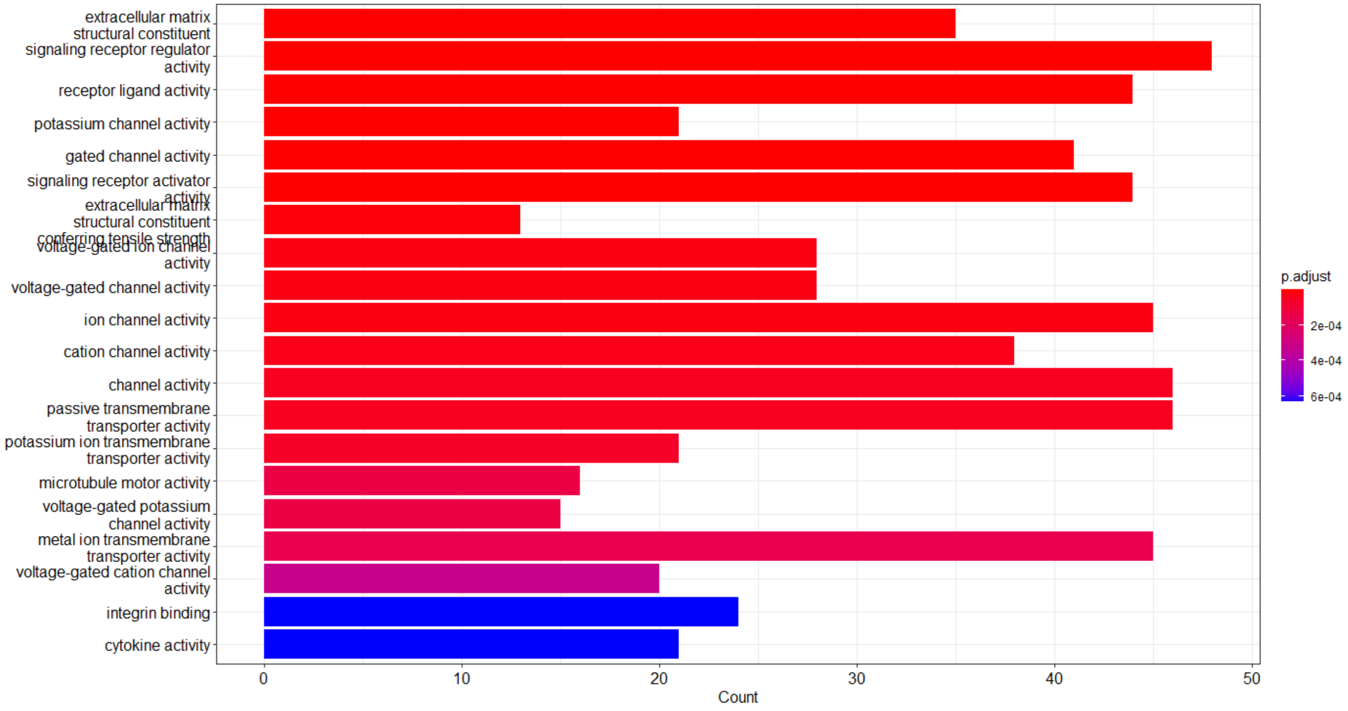
Visualization using enrichpl
GSEA only
gsearank
ridgeplot
gseaplot
gseaplot2
ORA & GSEA
emapplot
upsetplot
heatplot
dotplot
cnetplot
ORA only
barplot
goplot



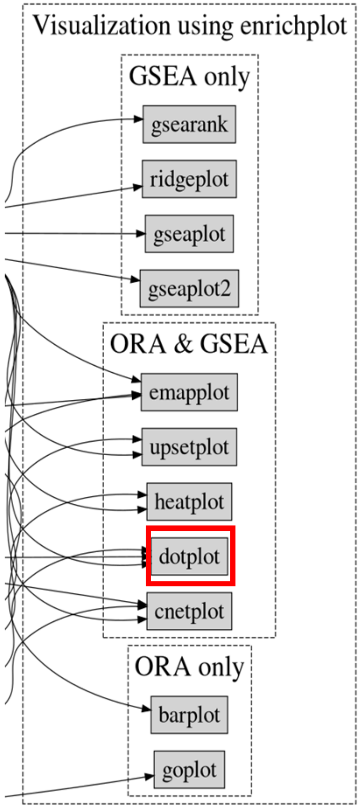
molecular_function
structural molecule activity
transporter activity
binding
molecular function regulator
extracellular matrix structural constituent
transmembrane transporter activity
protein binding
receptor regulator activity
extracellular matrix structural constituent conferring tensile strength
inorganic molecular entity transmembrane transporter activity
ion transmembrane transporter activity
passive transmembrane transporter activity
signaling receptor binding
signaling receptor activator activity
cation transmembrane transporter activity
channel activity
receptor ligand activity
inorganic cation transmembrane transporter activity
ion channel activity
gated channel activity
metal ion transmembrane transporter activity
cation channel activity
voltage-gated channel activity
potassium ion transmembrane transporter activity
voltage-gated ion channel activity
potassium channel activity
p.adjust
0.001
0.002
0.003
relationship
isa

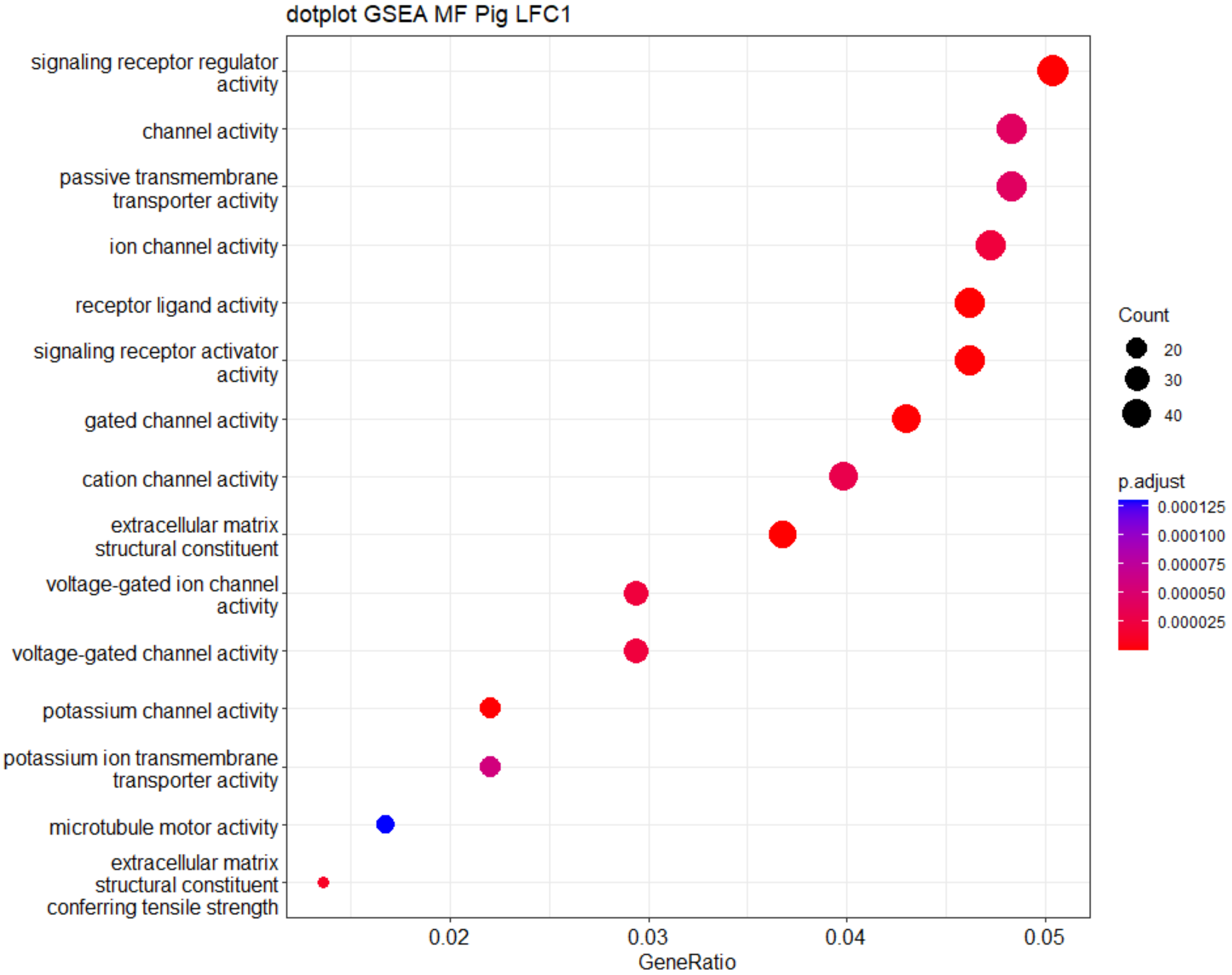
Bar plot


Visualization using enrichplot
GSEA only
gsearank
ridgeplot
gseaplot
gseaplot2
ORA & GSEA
emapplot
upsetplot
heatplot
dotplot
cnetplot
ORA only
barplot
goplot

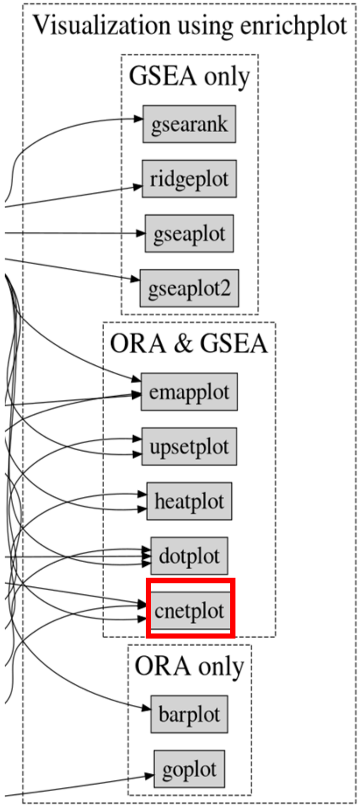

extracellular matrix structural constituent
signaling receptor regulator activity
receptor ligand activity
potassium channel activity
gated channel activity
signaling receptor activator activity
extracellular matrix structural constituent conferring tensile strength
voltage-gated ion channel activity
voltage-gated channel activity
ion channel activity
cation channel activity
channel activity
passive transmembrane transporter activity
potassium ion transmembrane transporter activity
microtubule motor activity
voltage-gated potassium channel activity
metal ion transmembrane transporter activity
voltage-gated cation channel activity
integrin binding
cytokine activity
0
10
20
30
40
50
Count
p.adjust
2e-04
4e-04
6e-04

Dot plot


Visualization using enrichplot
GSEA only
gsearank
ridgeplot
gseaplot
gseaplot2
ORA & GSEA
emapplot
upsetplot
heatplot
dotplot
cnetplot
ORA only
barplot
goplot


dotplot GSEA MF Pig LFC1
signaling receptor regulator activity
channel activity
passive transmembrane transporter activity
ion channel activity
receptor ligand activity
signaling receptor activator activity
gated channel activity
cation channel activity
extracellular matrix structural constituent
voltage-gated ion channel activity
voltage-gated channel activity
potassium channel activity
potassium ion transmembrane transporter activity
microtubule motor activity
extracellular matrix structural constituent conferring tensile strength
0.02
0.03
0.04
0.05
GeneRatio
Count
20
30
40
p.adjust
0.000125
0.000100
0.000075
0.000050
0.000025

Cnet plot (circular)



Visualization using enrichplot

GSEA only

gsearank

ridgeplot

gseaplot

gseaplot2

ORA & GSEA

emapplot

upsetplot

heatplot

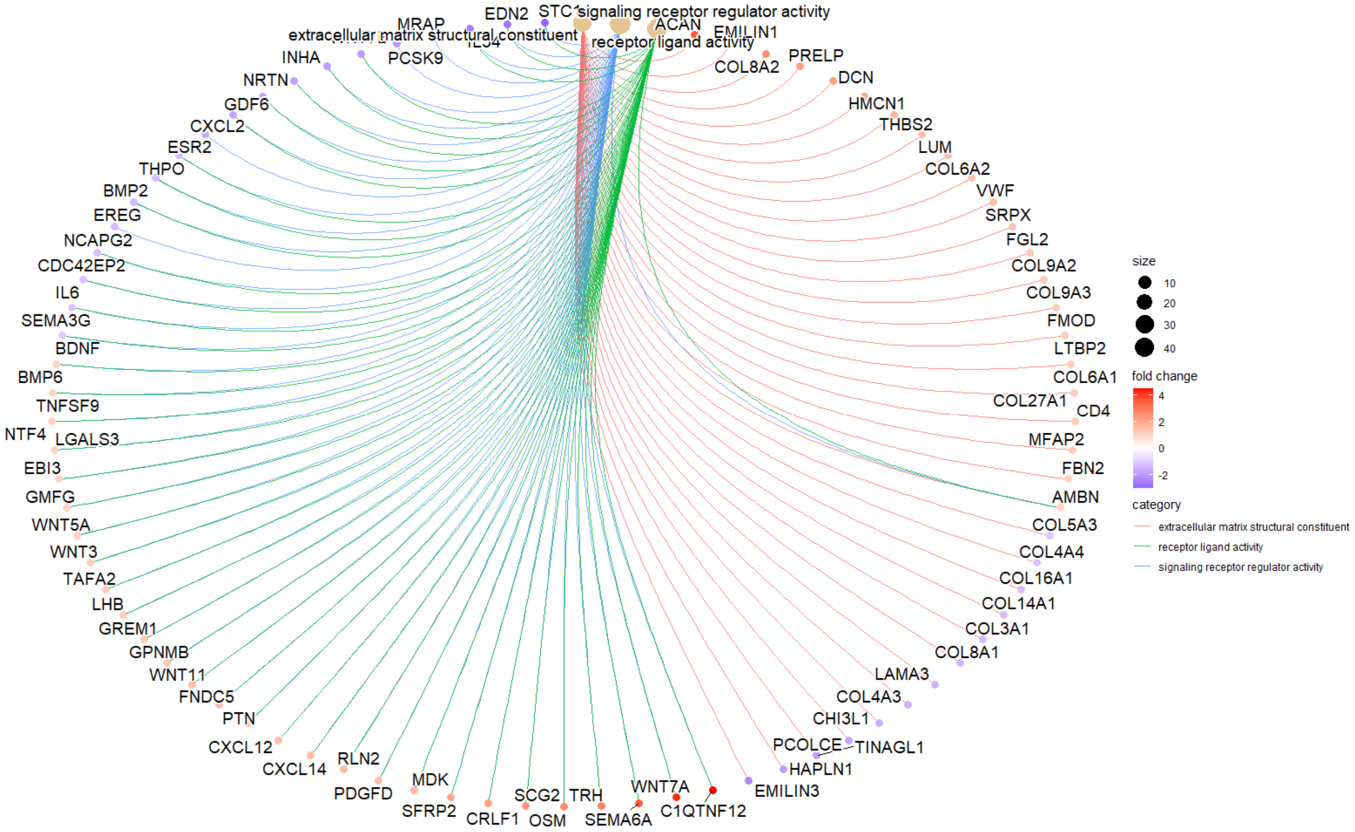
dotplot

cnetplot

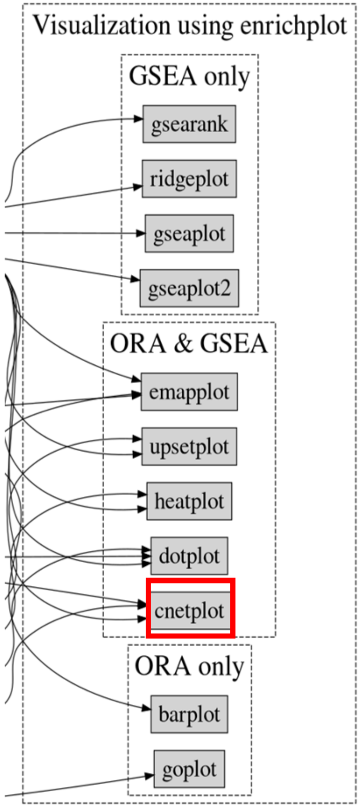
ORA only

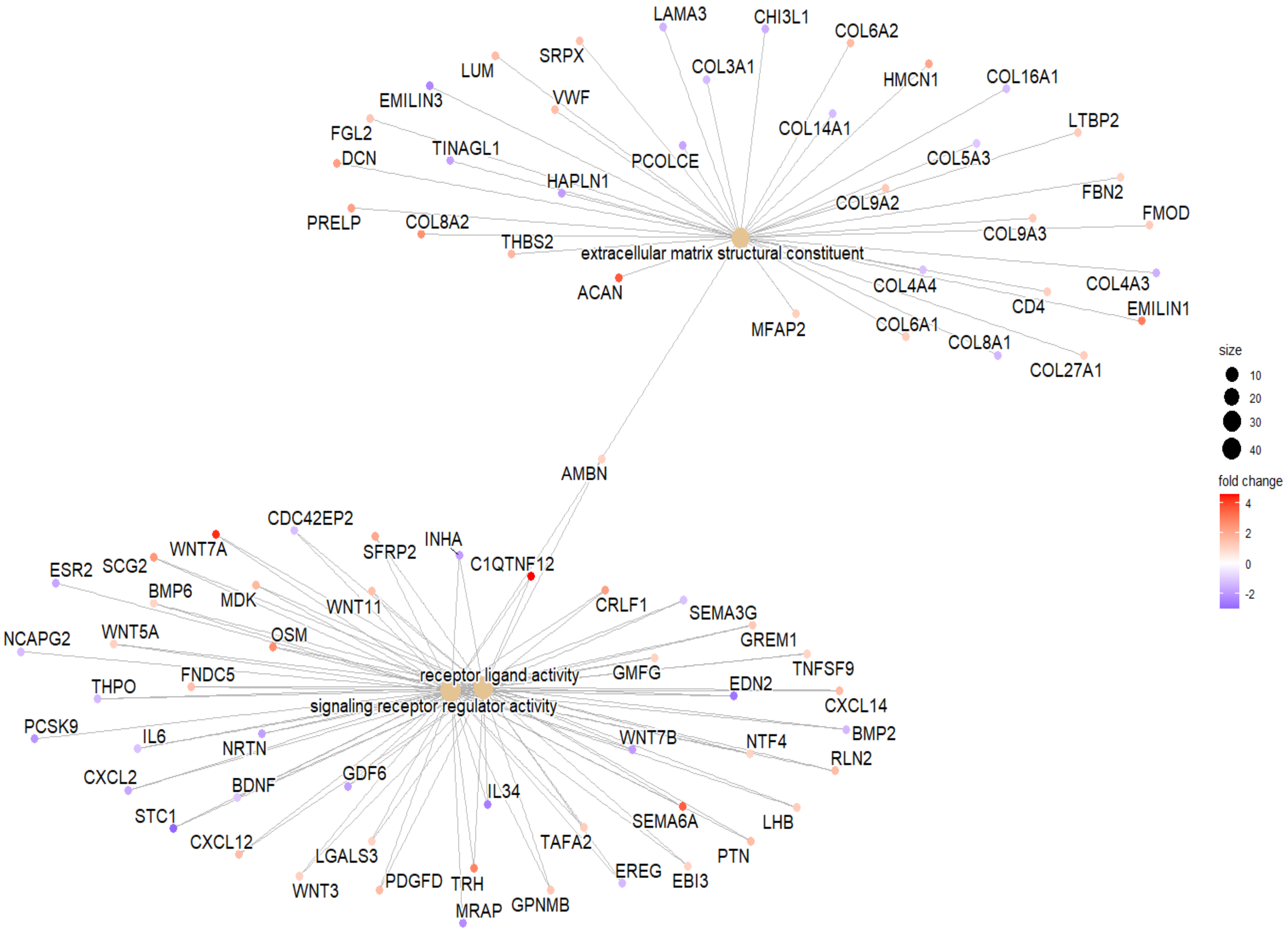
barplot

goplot

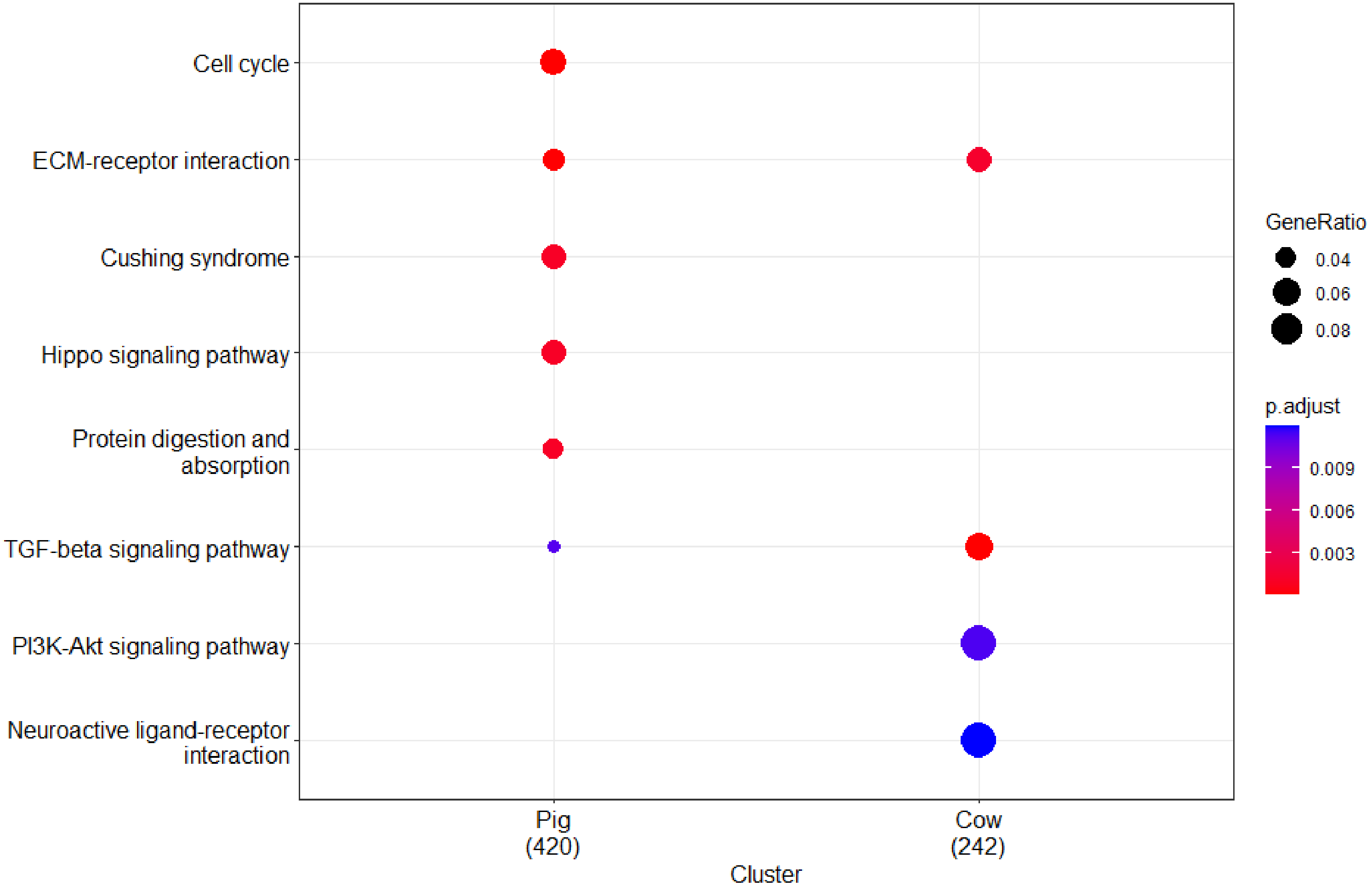


Cnet plot


Visualization using enrichplot
GSEA only
gsearank
ridgeplot
gseaplot
gseaplot2
ORA & GSEA
emapplot
upsetplot
heatplot
dotplot
cnetplot
ORA only
barplot
goplot

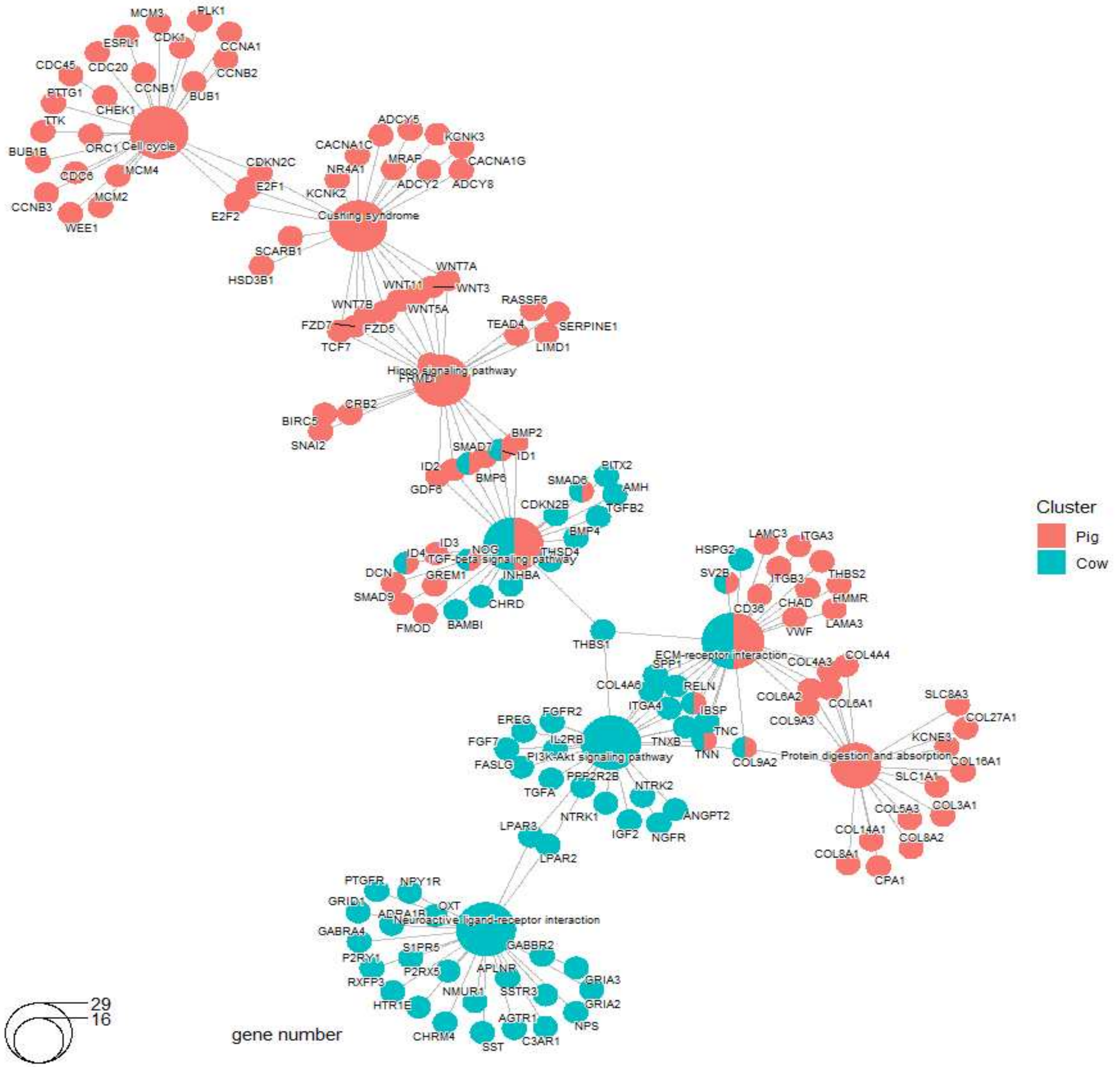

LAMA3
CHI3L1
COL6A2
SRPX
COL3A1
LUM
HMCN1
COL16A1
VWF
EMILIN3
COL14A1
LTBP2
FGL2
TINAGL1
PCOLCE
COL5A3
DCN
HAPLN1
FBN2
COL9A2
FMOD
PRELP
COL8A2
COL9A3
THBS2
extracellular matrix structural constituent
COL4A3
COL4A4
ACAN
CD4
EMILIN1
MFAP2
COL6A1
COL8A1
COL27A1
AMBN
CDC42EP2
INHA
WNT7A
SFRP2
C1QTNF12
ESR2
SCG2
BMP6
MDK
WNT11
CRLF1
SEMA3G
NCAPG2
WNT5A
OSM
GREM1
TNFSF9
receptor ligand activity
GMFG
THPO
FNDC5
EDN2
signaling receptor regulator activity
CXCL14
PCSK9
BMP2
IL6
NRTN
WNT7B
NTF4
RLN2
CXCL2
GDF6
BDNF
IL34
STC1
SEMA6A
LHB
CXCL12
TAFA2
LGALS3
PTN
EREG
EBI3
PDGFD
TRH
WNT3
GPNMB
MRAP
size
10
20
30
40
fold change
4
2
0
-2

Compare Cluster



KEGG pathway

Compare cluster

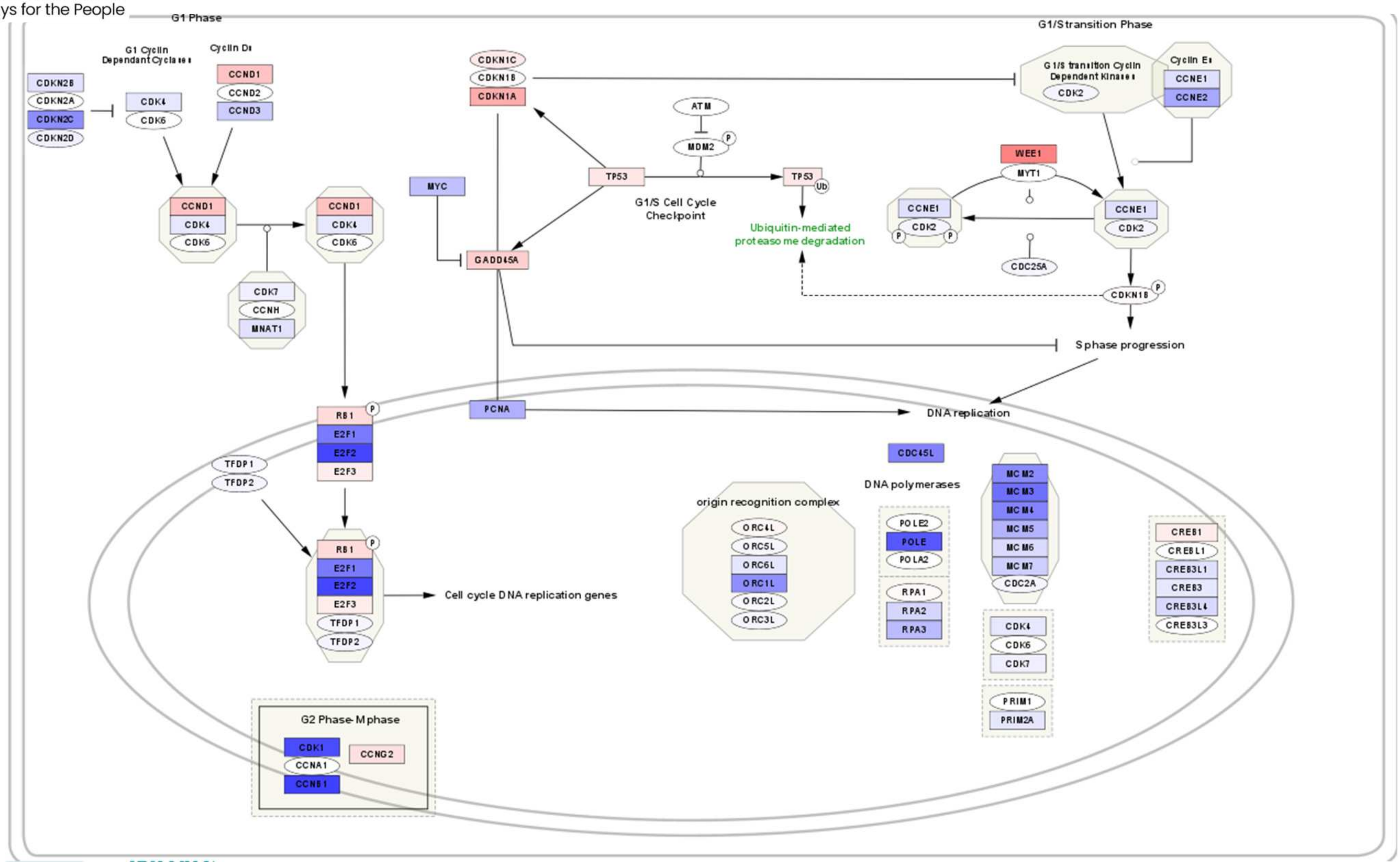


BP GSEA

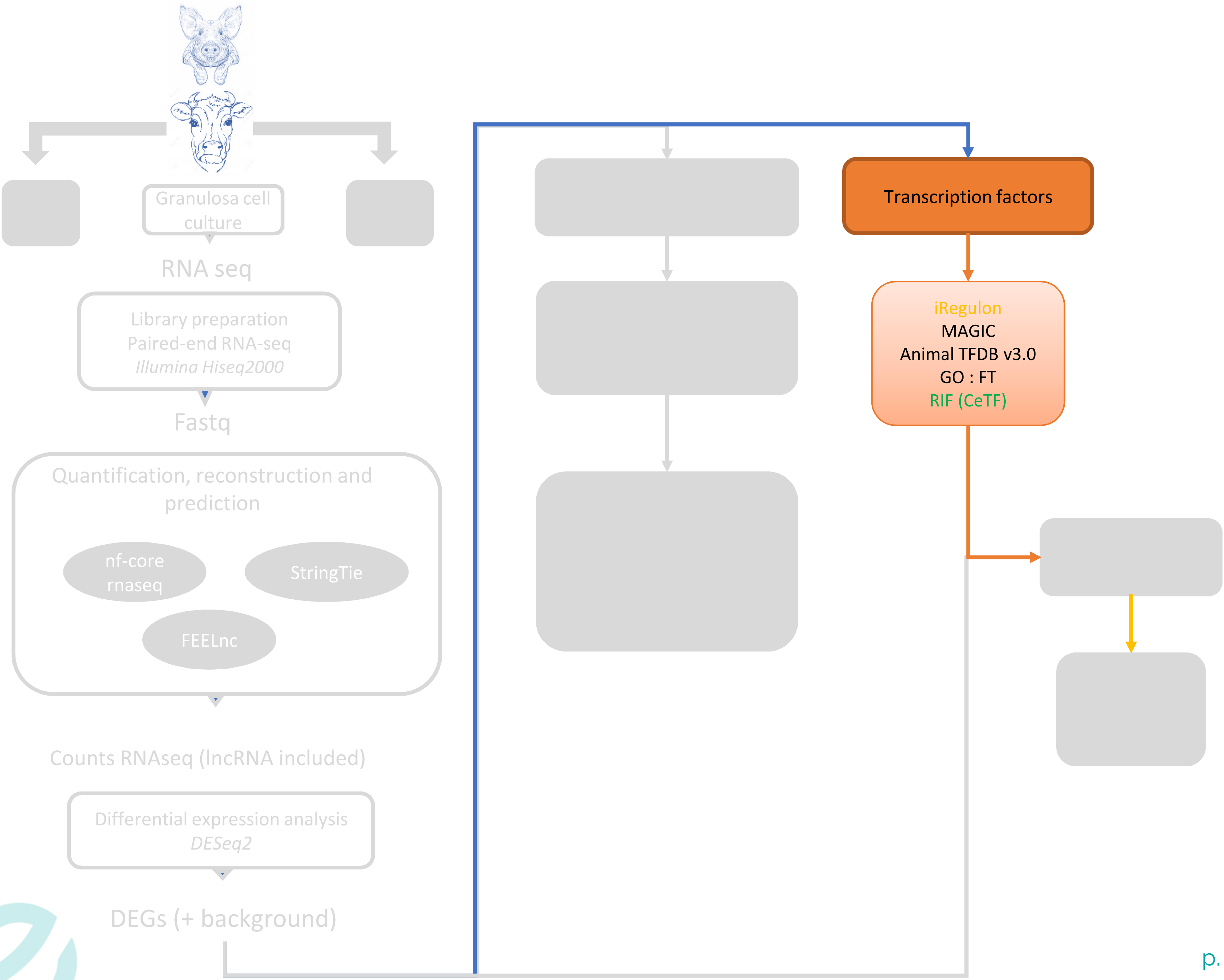
WikiPathways
Pathways for the People

WP45 G1 to S cell cycle

Cytoscape



Inhibition cell cycle and tissue development,
inhibition cell survival; increased apoptosis



Granulosa cell culture

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

nf-core rnaseq

StringTie

FEELnc

Counts RNAseq (lncRNA included)

Differential expression analysis
DESeq2

DEGs (+ background)

Transcription factors

iRegulon
MAGIC
Animal TFDB v3.0
GO : FT
RIF (CeTF)

Web ; Cytoscape ; R package

Transcriptions factors :

iRegulon
MAGIC
Animal TFDB v3.0
GO : FT
RIF (CeTF)

Raw view :

Iregulon : Database de 10 000 FTs + 1 120 CHiP-seq datasets (Janky et al., 2014)

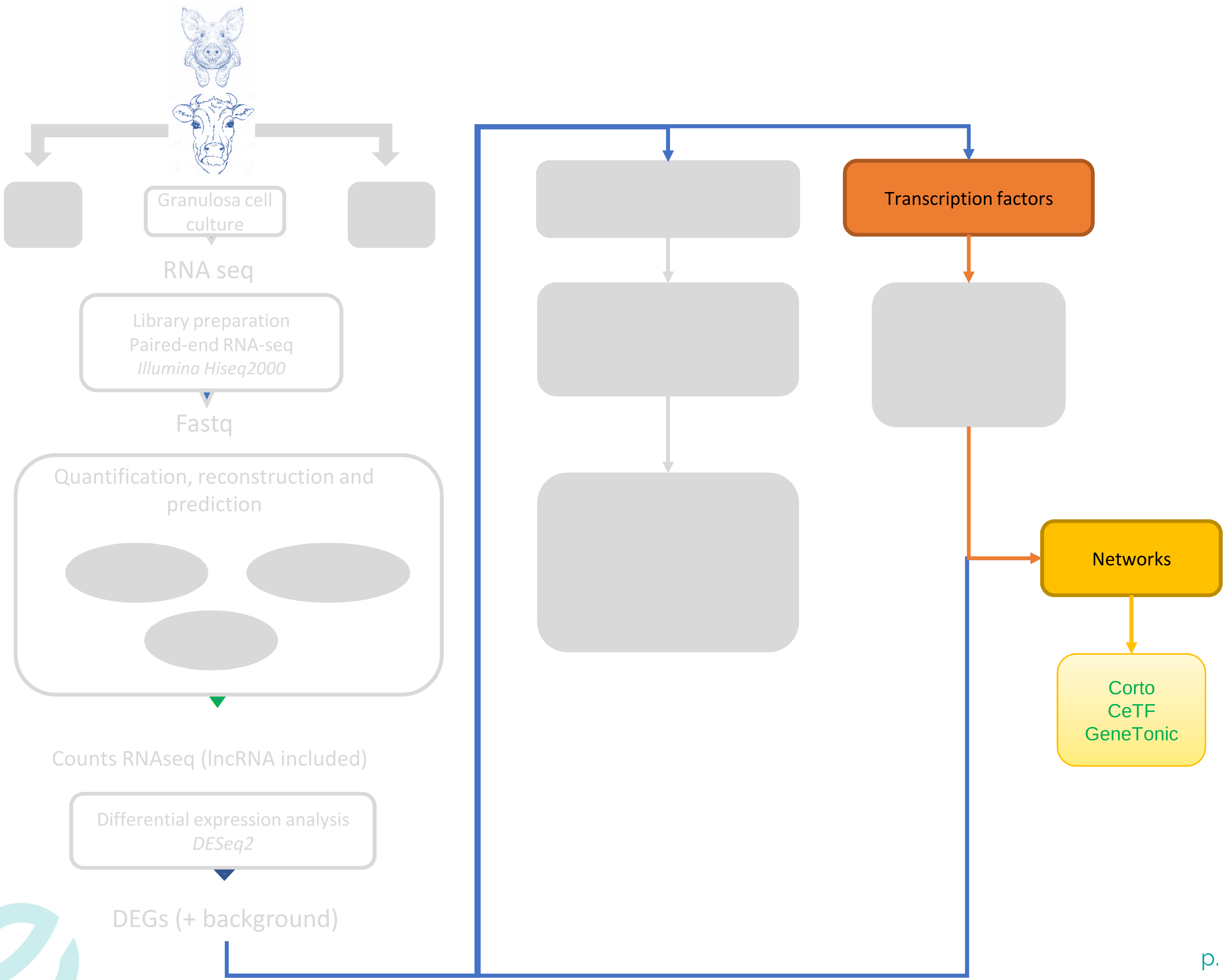
MAGIC Mining Algorithm for GenetIc Controllers: 2312 CHiP-seq tracks 684 FT in 588 cell lines (ENCODE data) (Roopra 2020 *Plos Comput Biol) (cluster)*

Gene Regulatory Networks (GRN) → Detailed view

Bos taurus: 1396 TFs and 935 TF Cofactors

Sus scrofa: 1490 TFs and 937 TF Cofactors

Gene Ontology : DNA-binding transcription factor activity



Granulosa cell culture

RNA seq

Library preparation
Paired-end RNA-seq
Illumina Hiseq2000

Fastq

Quantification, reconstruction and prediction

Counts RNAseq (lncRNA included)

Differential expression analysis
DESeq2

DEGs (+ background)

Transcription factors

Networks

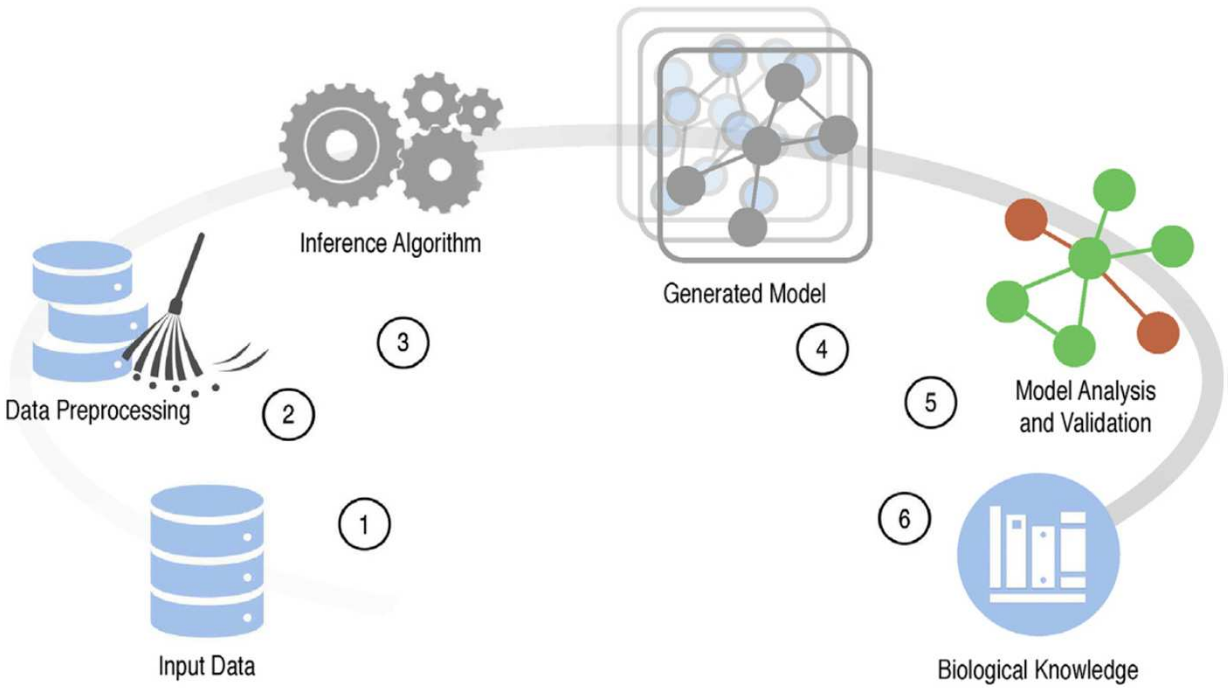
Corto
CeTF
GeneTonic

Web ; Cytoscape ; R package

Gene Regulatory Networks reconstruction

Several inference algorithms

- Information theory (co-expression)
- Boolean networks
- Differential equations models
- Bayesian models
- Neural models



Delgado et al., 2019, *Art Intelligence in medicine*

PCIT approach : Partial Correlation with Information Theory

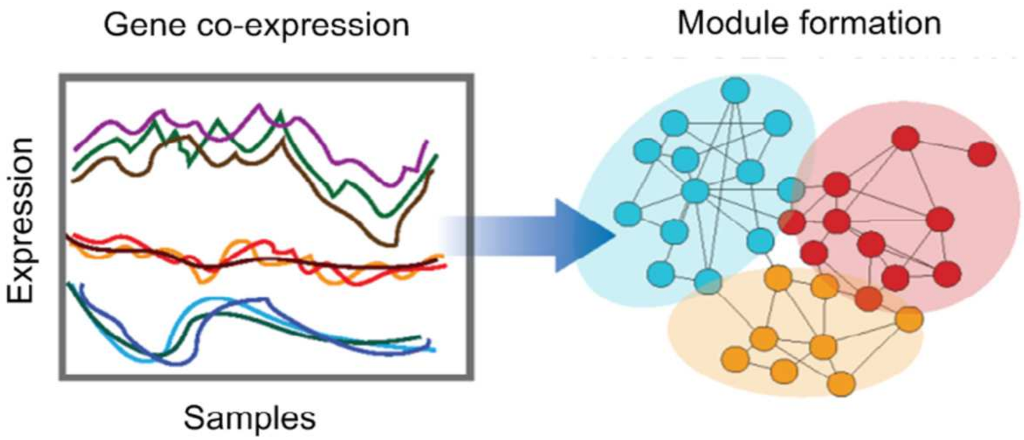
For every trio of genes in x, y and z

1) The three first-order partial correlation coefficients are computed -> the strength of the linear relationship between x and y that is independent of (uncorrelated with) z.

   Obtention d'un seuil « local » pour capturer les associations significatives

2) Data Processing Inequality (DPI) or theorem of Information. Theory which states that 'no clever manipulation of the data can improve the inference that can be made from the data' (Cover and Thomas 2012)

Packages R CeTF, Corto



Reverter et Chan, 2008 *Bioinformatics*

Package CeTF : identification of crucial FTs

RIF Regulatory Information Factors algorithm (Reverter et al., 2010 *Bioinformatics*) → MSTN

iRegulon
MAGIC
Animal TFDB v3.0
GO : FT
RIF (CeTF)

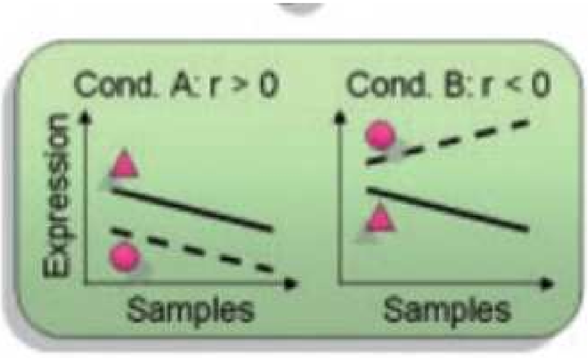
Counts normalisés de DEGs
Cibles

Liste de FTs
Régulateurs

Calcul de la corrélation de la co-expression de chaque TF avec les DEGs pour les 2 conditions

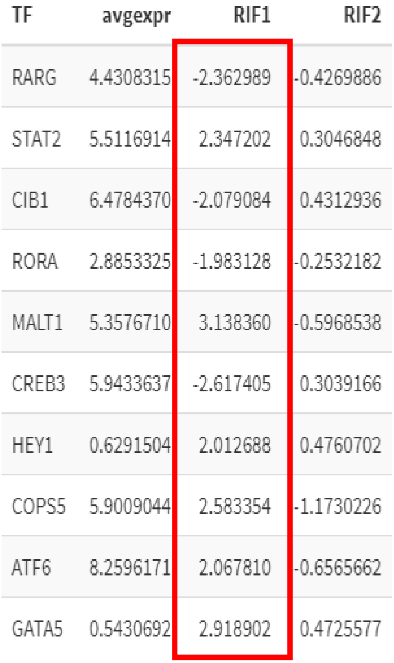
RIF1: high score to TFs that are most differentially co-expressed, highly abundant.
High RIF1 = FTs « constants »

RIF2: high score to TFs whose expression can predict better the abundance of DEGs



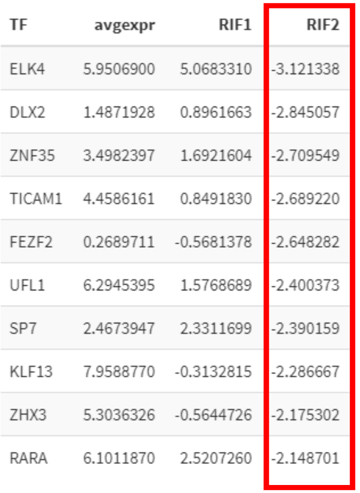
Reverter et al., 2010, *bioinformatics*

RIF Results



TF	avgexpr	RIF1	RIF2
RARG	4.4308315	-2.362989	-0.4269886
STAT2	5.5116914	2.347202	0.3046848
CIB1	6.4784370	-2.079084	0.4312936
RORA	2.8853325	-1.983128	-0.2532182
MALT1	5.3576710	3.138360	-0.5968538
CREB3	5.9433637	-2.617405	0.3039166
HEY1	0.6291504	2.012688	0.4760702
COPS5	5.9009044	2.583354	-1.1730226
ATF6	8.2596171	2.067810	-0.6565662
GATA5	0.5430692	2.918902	0.4725577

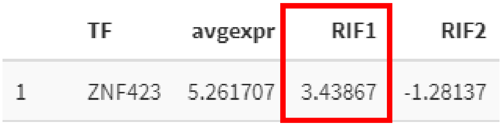
26 TFs RIF1



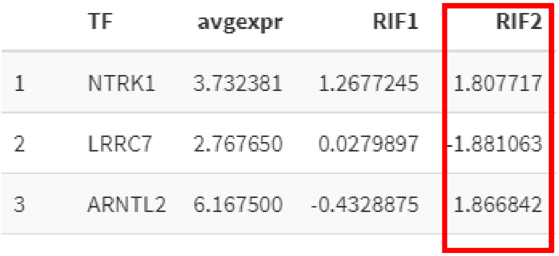
TF	avgexpr	RIF1	RIF2
ELK4	5.9506900	5.0683310	-3.121338
DLX2	1.4871928	0.8961663	-2.845057
ZNF35	3.4982397	1.6921604	-2.709549
TICAM1	4.4586161	0.8491830	-2.689220
FEZF2	0.2689711	-0.5681378	-2.648282
UFL1	6.2945395	1.5768689	-2.400373
SP7	2.4673947	2.3311699	-2.390159
KLF13	7.9588770	-0.3132815	-2.286667
ZHX3	5.3036326	-0.5644726	-2.175302
RARA	6.1011870	2.5207260	-2.148701

29 TFs RIF2

Zscore 1,96 pval<0,05



	TF	avgexpr	RIF1	RIF2
1	ZNF423	5.261707	3.43867	-1.28137

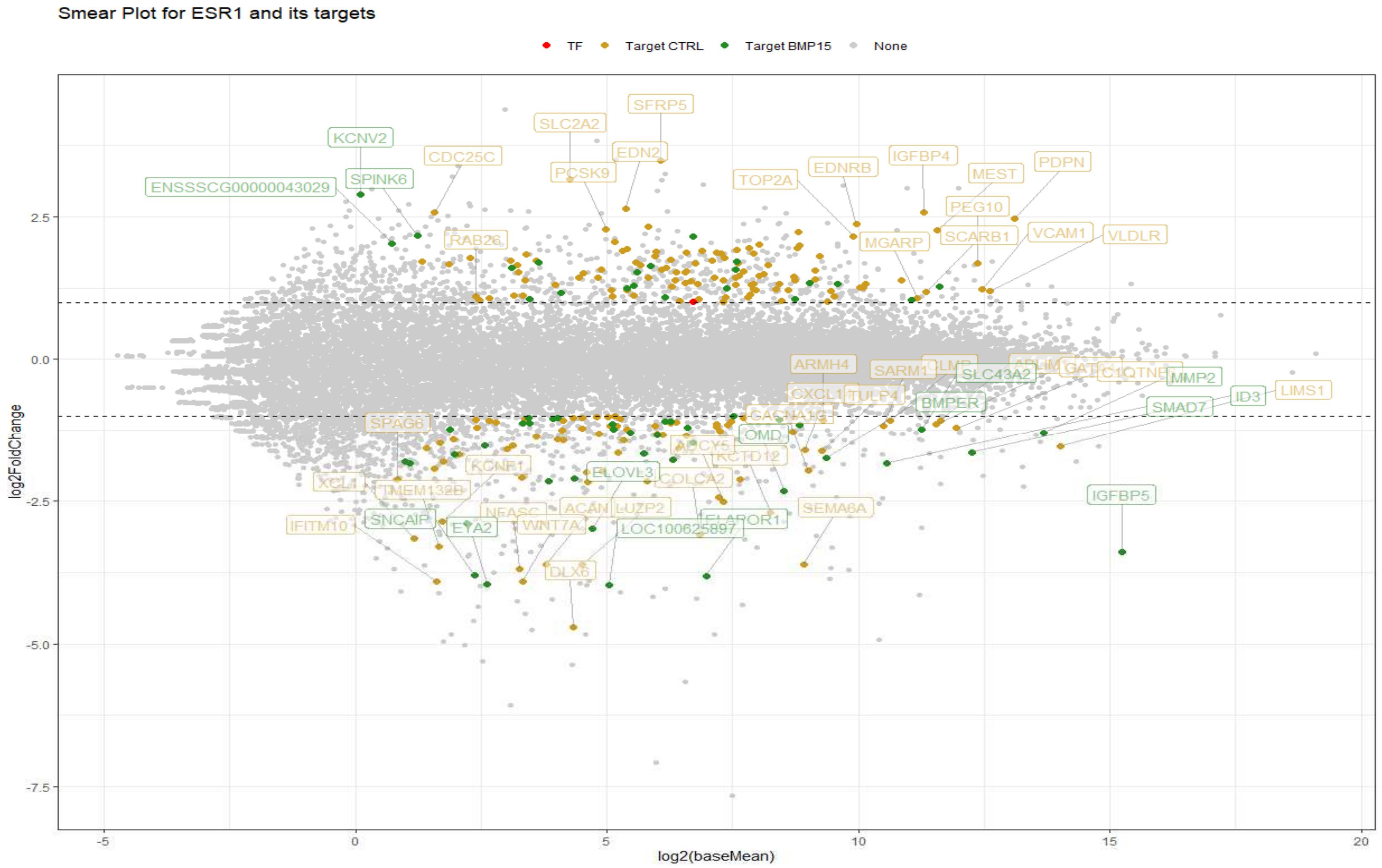


	TF	avgexpr	RIF1	RIF2
1	NTRK1	3.732381	1.2677245	1.807717
2	LRRC7	2.767650	0.0279897	-1.881063
3	ARNTL2	6.167500	-0.4328875	1.866842

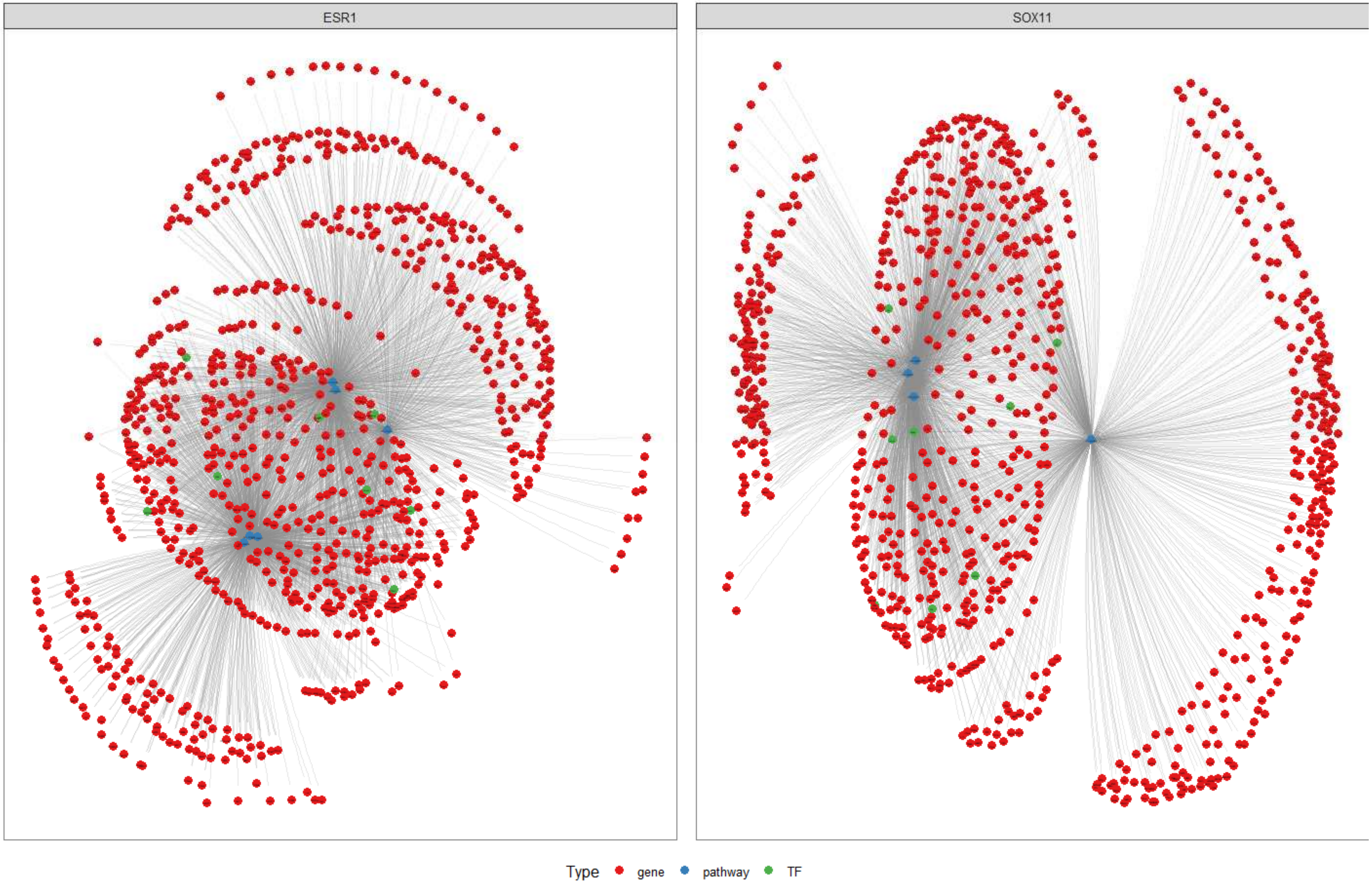
Zscore 1,65 pval<0,10

Packages R CeTF

FT ESR1 and its targets



INRAE


ESR1
SOX11
Type
● gene
● pathway
● TF

Package CeTF : Pros and cons

😃 RIF algorithm in R

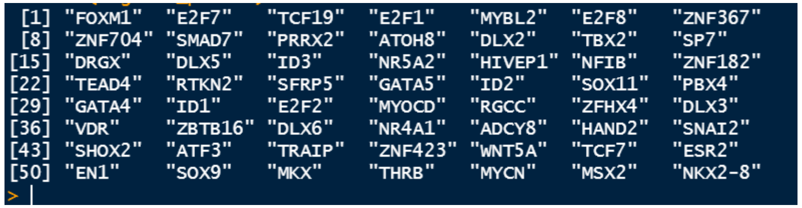
😢 2 conditions, little used, maintainer not responding, graphic interface.

Corto : (Correlation Tool): an R package to generate correlation-based DPI networks. Mercatelli et al., 2020 Bioinformatics

😃 Maintainer ++

Corto

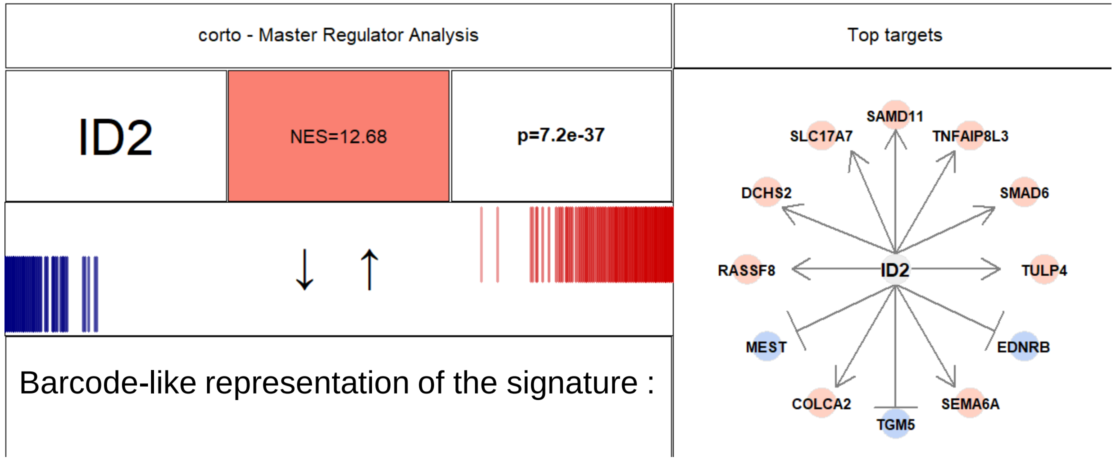
1) Identification of TFs expressed in the dataset



```
 [1] "FOXM1"  "E2F7"   "TCF19"  "E2F1"   "MYBL2"  "E2F8"   "ZNF367"
 [8] "ZNF704" "SMAD7"  "PRRX2"  "ATOH8"  "DLX2"   "TBX2"   "SP7"
[15] "DRGX"   "DLX5"   "ID3"    "NR5A2"  "HIVEP1" "NFIB"   "ZNF182"
[22] "TEAD4"  "RTKN2"  "SFRP5"  "GATA5"  "ID2"    "SOX11"  "PBX4"
[29] "GATA4"  "ID1"    "E2F2"   "MYOCD"  "RGCC"   "ZFHX4"  "DLX3"
[36] "VDR"    "ZBTB16" "DLX6"   "NR4A1"  "ADCY8"  "HAND2"  "SNAI2"
[43] "SHOX2"  "ATF3"   "TRAIP"  "ZNF423" "WNT5A"  "TCF7"   "ESR2"
[50] "EN1"    "SOX9"   "MKX"    "THRB"   "MYCN"   "MSX2"   "NKX2-8"
> |
```

2) Gene network inference (optimized pairwise correlation, DPI and bootstrapping)

Master Regulator Analysis : TF networks vs a signature (= 2 gene expression matrices)



Barcode-like representation of the signature :

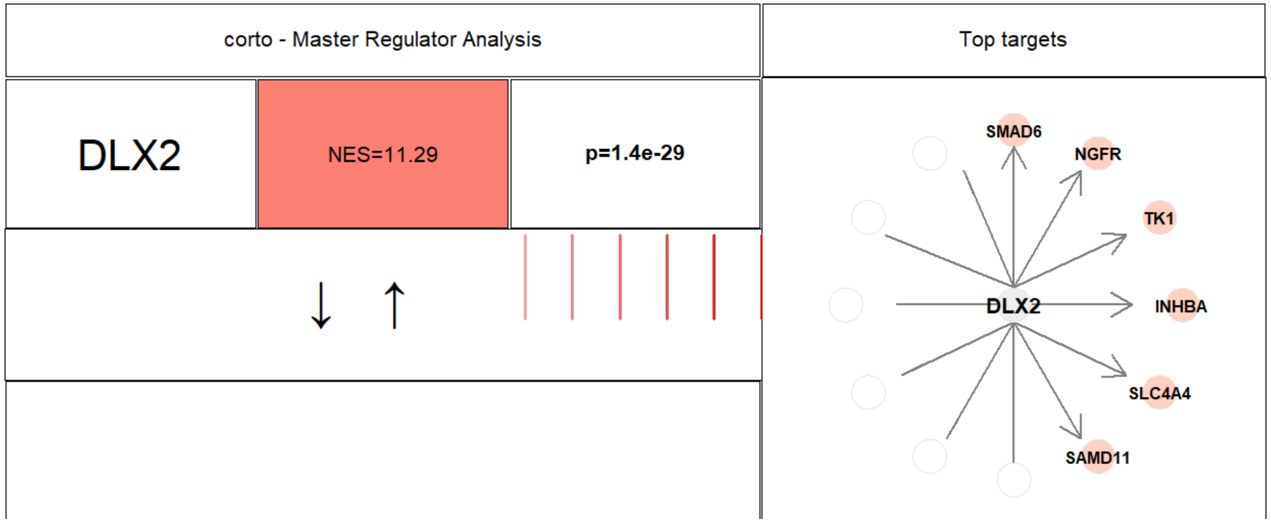
Expression levels in the signature

Individual GR network

corresponding P value (based on 10 000 permutation tests and signature sample shuffling)

Corto

Corto
CeTF
GeneTonic



The transparency of each bar is associated to the value of the target in the signature.

What else ?

- From R to Cytoscape or Gephi :

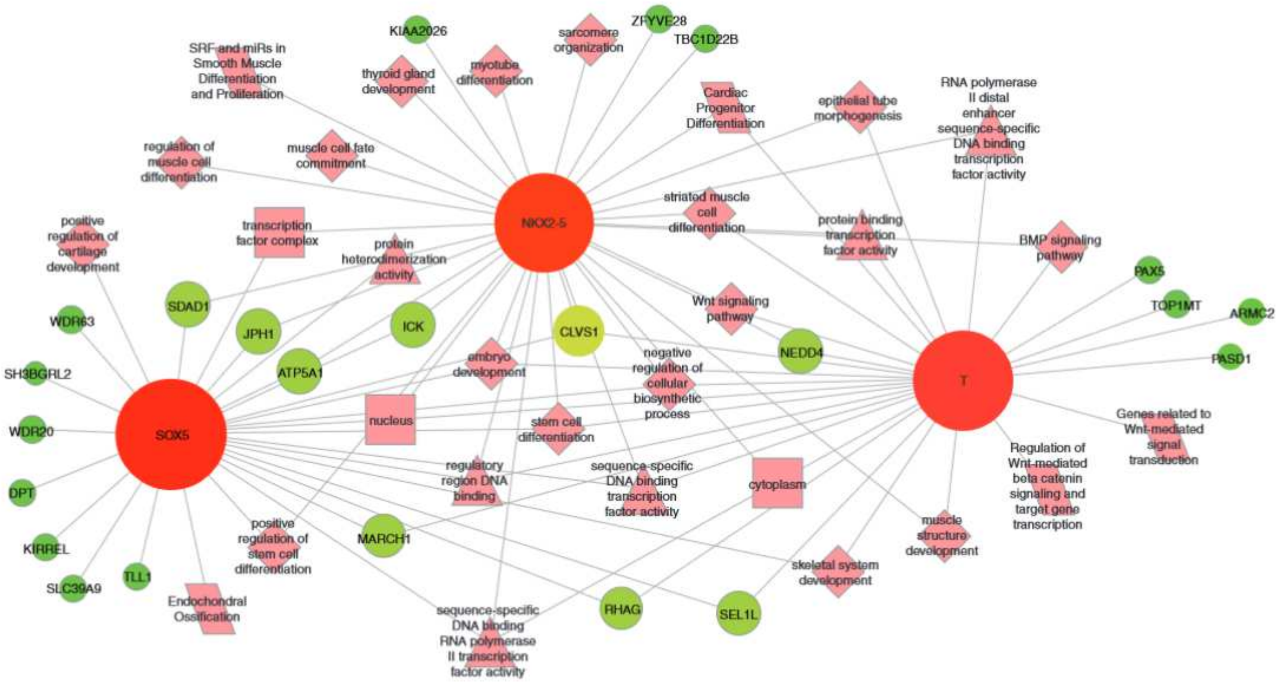

KIAA2026
sarcomere organization
ZFYVE28
TBC1D22B
SRF and miRs in Smooth Muscle Differentiation and Proliferation
thyroid gland development
myotube differentiation
Cardiac Progenitor Differentiation
epithelial tube morphogenesis
RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity
regulation of muscle cell differentiation
muscle cell fate commitment
striated muscle cell differentiation
positive regulation of cartilage development
transcription factor complex
NKX2-5
protein binding transcription factor activity
BMP signaling pathway
protein heterodimerization activity
PAX5
TOP1MT
ARMC2
SDAD1
WDR63
JPH1
ICK
CLVS1
Wnt signaling pathway
NEDD4
PASD1
SH3BGRL2
ATP5A1
embryo development
negative regulation of cellular biosynthetic process
T
WDR20
SOX5
nucleus
stem cell differentiation
Genes related to Wnt-mediated signal transduction
regulatory region DNA binding
sequence-specific DNA binding transcription factor activity
cytoplasm
Regulation of Wnt-mediated beta catenin signaling and target gene transcription
DPT
positive regulation of stem cell differentiation
MARCH1
muscle structure development
KIRREL
TLL1
SLC39A9
Endochondral Ossification
skeletal system development
RHAG
SEL1L
sequence-specific DNA binding RNA polymerase II transcription factor activity

Fig. 3. Transcription factors network. Transcription factors (in red), with related genes (green), pathways (parallelogram) and gene ontology: molecular functions (triangle), biological process (diamond) and cellular component (rectangle).

- Add lncRNAs : Test TAGADA pipeline

Perspectives : TAGADA pipeline (GeneSwitch project)

Another pipeline using Nextflow for count matrix generation and lncRNA prediction

	Nextflow nf-core rnaseq + FEELnc + StringTie	Nextflow TAGADA
Quantification	RSEM	Stringtie
Merge of GTFs by sample	Outside, StringTie Merge	Inside,T-merge
transcript quantification	Outside, Stringtie	Inside, Stringtie
LncRNA prediction	Outside, FEELnc Choice of the method intra WF	Inside, FEELnc Choice of the methode upstream WF

+ Filters on transcripts « de novo » are also different.

What else ?

- GRN with lncRNAs
- Use co-expression networks to identify potential trans-targeted genes.
- Validate the trans target with lncTar (Li et al., 2014, *brief. Bioinform) :*

  Predict lncRNA-RNA interactions based on secondary structure

Thanks

TAGADA pipeline

Sarah Djebali — Inserm

Cervin Guyomar — SIGENAE

Cyril Kurylo

Sylvain Foissac

Genotoul Bioinfo

Genotoul GeT