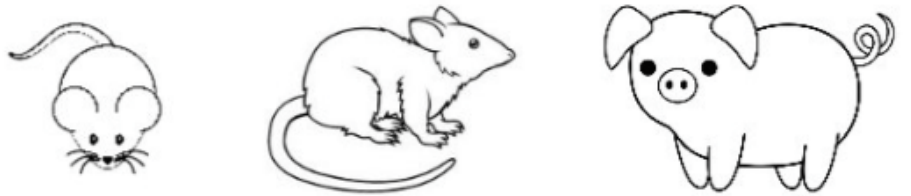
Integration of tissue expression datasets in model organisms



Oana Palasca

31st TBI Winterseminar
Bled, February 20th, 2016

RTH
CENTER FOR NON-CODING RNA IN TECHNOLOGY AND HEALTH

Motivation

Expression atlas available as a web resource

- Expression atlas for human and model organisms (mouse, rat, pig)
- Confidence scores, comparable across experiments and organisms

Model organisms (long-term goal)

- Enable expression comparisons between organisms
- Help in selecting a suitable model for a given disease context

Tissues (Santos et al, 2015)

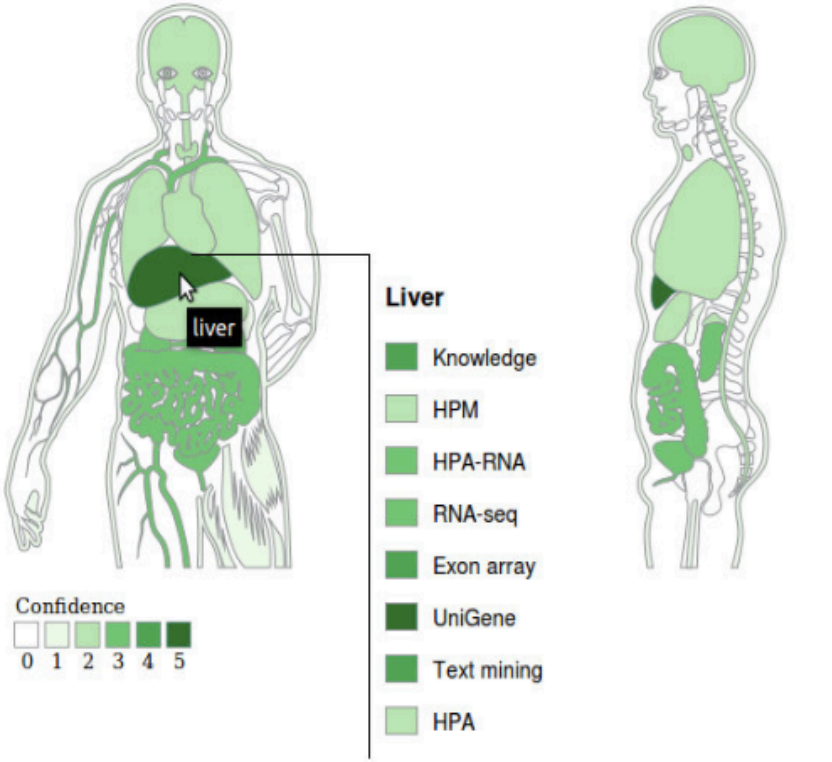
http://tissues.jensenlab.org/

CYP3A4 [ENSP00000337915]

Cytochrome P450, family 3, subfamily A, polypeptide 4

Synonyms: CYP3A4, B4DPQ5_HUMAN, C9JBD2_HUMAN, CP3A4_HUMAN, CYP3A4p ...

Linkouts: UniProt #1 #2 #3 #4 #5 #6 #7 OMIM



Knowledge

Name	Source	Evidence	Confidence
Liver	UniProtKB-RC	CURATED	★★★★☆

Experiments Next >

Name	Source	Evidence	Confidence
Liver	UniGene	212 ESTs	★★★★★
Liver	Exon array	2526 intensity units	★★★★☆
Liver	HPA-RNA	1318.3 FPKM	★★★☆☆
Duodenum	HPA-RNA	726.8 FPKM	★★★☆☆
Liver	RNA-seq	100.551 RPKM	★★★☆☆
Liver	HPA	High: 1 antibody	★★☆☆☆
Duodenum	HPA	High: 1 antibody	★★☆☆☆
Liver	HPM	17 peptides	★★☆☆☆
Brain	HPM	7 peptides	★☆☆☆☆
Adrenal gland	HPA-RNA	6.6 FPKM	★☆☆☆☆

Text mining Next >

Name	Z-score	Confidence
Liver	7.3	★★★★☆
Blood plasma	6.1	★★★★☆
Juice	6.1	★★★★☆
Intestine	5.5	★★★☆☆
Urine	5.0	★★★☆☆
Adult	4.4	★★★☆☆
Kidney	4.4	★★★☆☆
Lung	3.4	★★☆☆☆

Tissues resource

Which transcripts are expressed in which tissues?

Idea: **Combine tissue expression datasets** obtained from different experiments and technologies, in order to **improve quality and tissue coverage**

Tissues resource

Which transcripts are expressed in which tissues?

- Evaluate the quality of each dataset in terms of the **fold enrichment** of correct transcript-tissue associations compared to a **gold standard dataset** (UniprotKB manual tissue annotation)

- Assign **confidence scores**, comparable across experiments/ datasets, to each transcript-tissue association.

Datasets

Mouse:
- GNF Expression Atlas – microarray (Su et al., 2004)
- GNF Expression Atlas v3 – microarray (Lattin et al., 2008)
- RNA-seq Atlas – polyA RNA, 9 tissues (Merkin et al., 2012)
- Mouse ENCODE RNA-Seq – polyA RNA, 22 tissues (ENCODE/CSHL, 2012)

Rat
- RGU34A Gene Atlas (Walker et al., 2004) (only ~7,000 protein IDs)
- Rat transcriptomic BodyMap – total RNA (Lattin et al., 2008)
- RNA-seq Atlas – polyA RNA, 9 tissues (Merkin et al., 2012)

Pig:
- Pig Array Atlas – microarray, 62 tissues (Freeman et al., 2012)
- RNA-seq Atlas – polyA RNA, 10 tissues (Farajzadeh et al., 2013)
- RNA-seq Atlas – total RNA, 8 tissues (Groenen lab/FAANG*, 2015)

*Functional annotation of animal genomes

Transcript quantification

RNA-Seq data processing

- Read mapping with **STAR** (Dobin et al, 2013)
- Transcript quantification with **cuffnorm** (Trapnell et al, 2010)

Gold standard datasets – UniprotKB tissue annotation

	Proteins in UniprotKB	Protein-tissue pairs in UniprotKB
Human	17,000	60,000
Mouse	7,700	13,000
Rat	4,600	6,000
Pig	650	800

Pig and rat UniprotKB datasets are too small!

➡ **orthology transfer** from human;
one-to one orthologs from eggNOG (Powell et al, 2013)

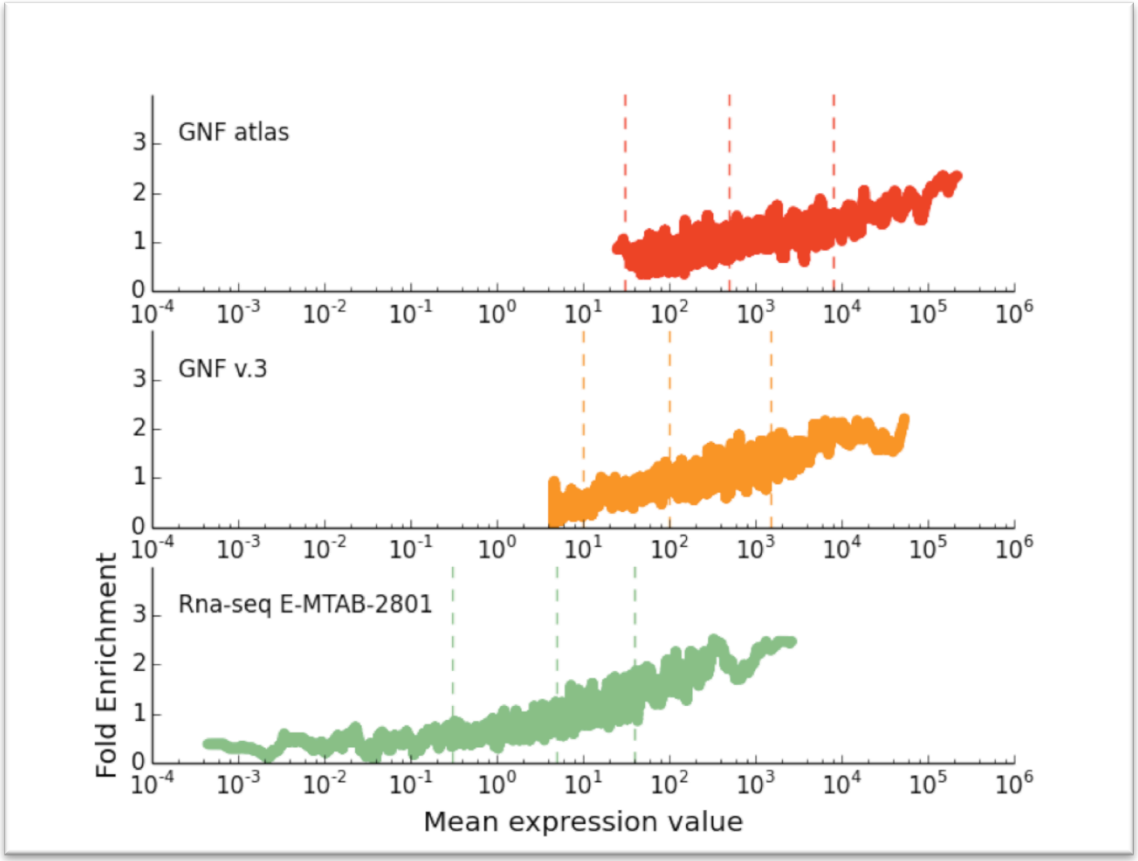
Orthology transfer gold standard datasets

One-to-one orthologs extracted from eggNOG with tissue annotation in the Human UniProtKB dataset

	Proteins	Protein-tissue pairs
Mouse	10,600	36,000
Rat	9,200	31,000
Pig	9,150	31,000

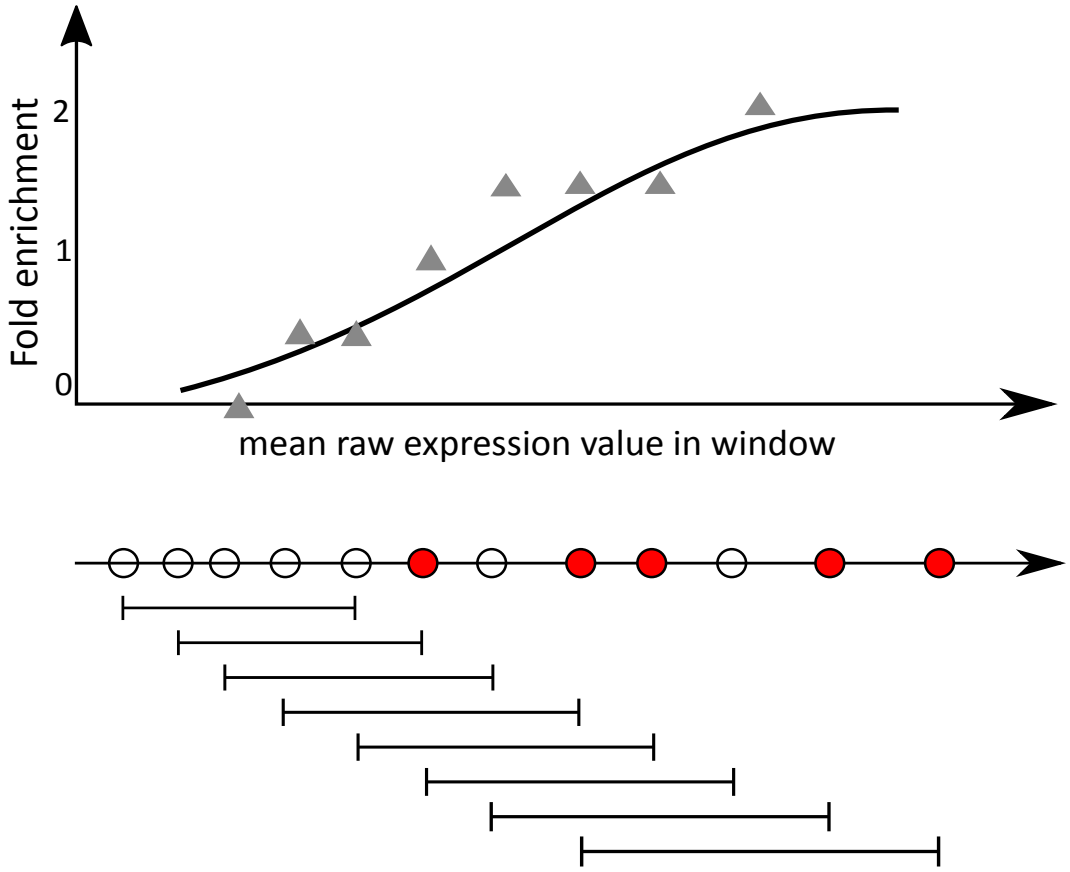
Fold enrichment of transcript-tissue associations found in the gold standard dataset compared to random chance, calculated over sliding windows.

Mouse datasets, UniprotKB mouse annotation.



Fold enrichment analysis

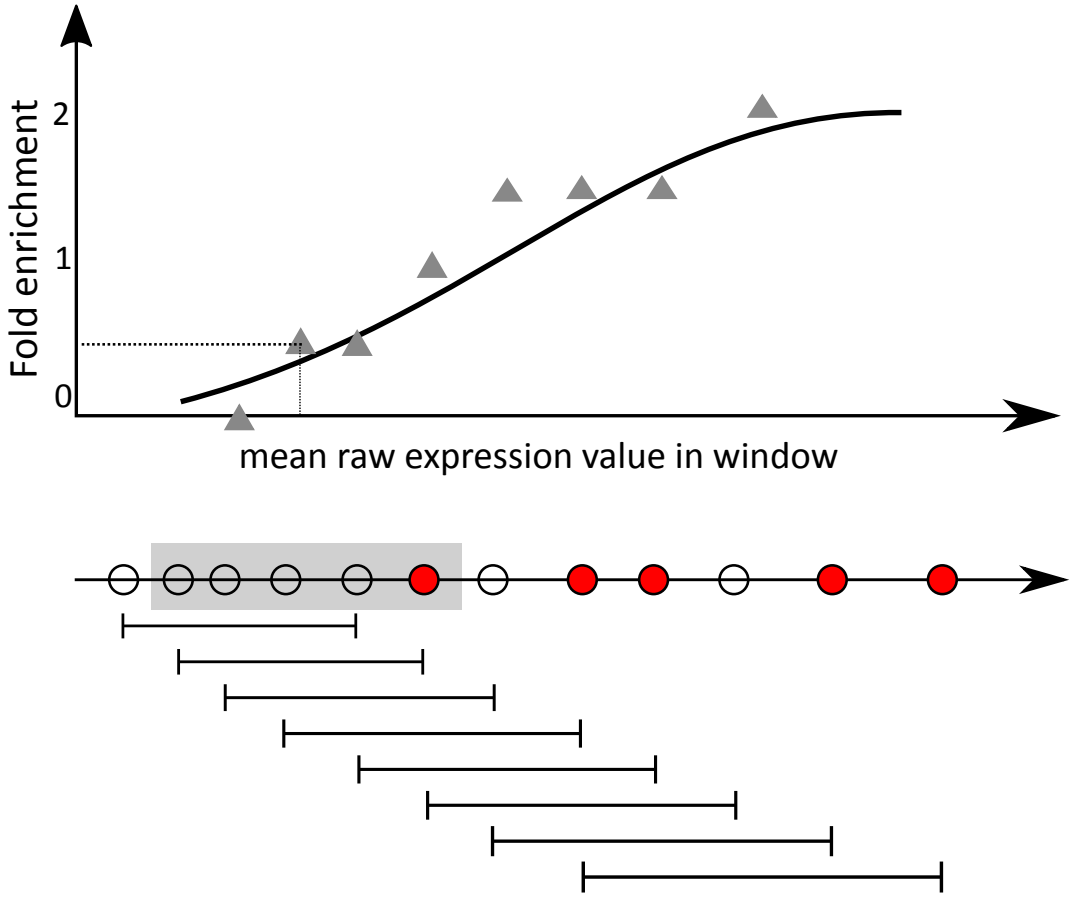
- **4** transcripts, **3** tissues
- **12** transcript-tissue pairs in the expression dataset
- **5** transcript-tissue pairs in the gold standard dataset
- random_chance = **5/12**



Fold enrichment analysis

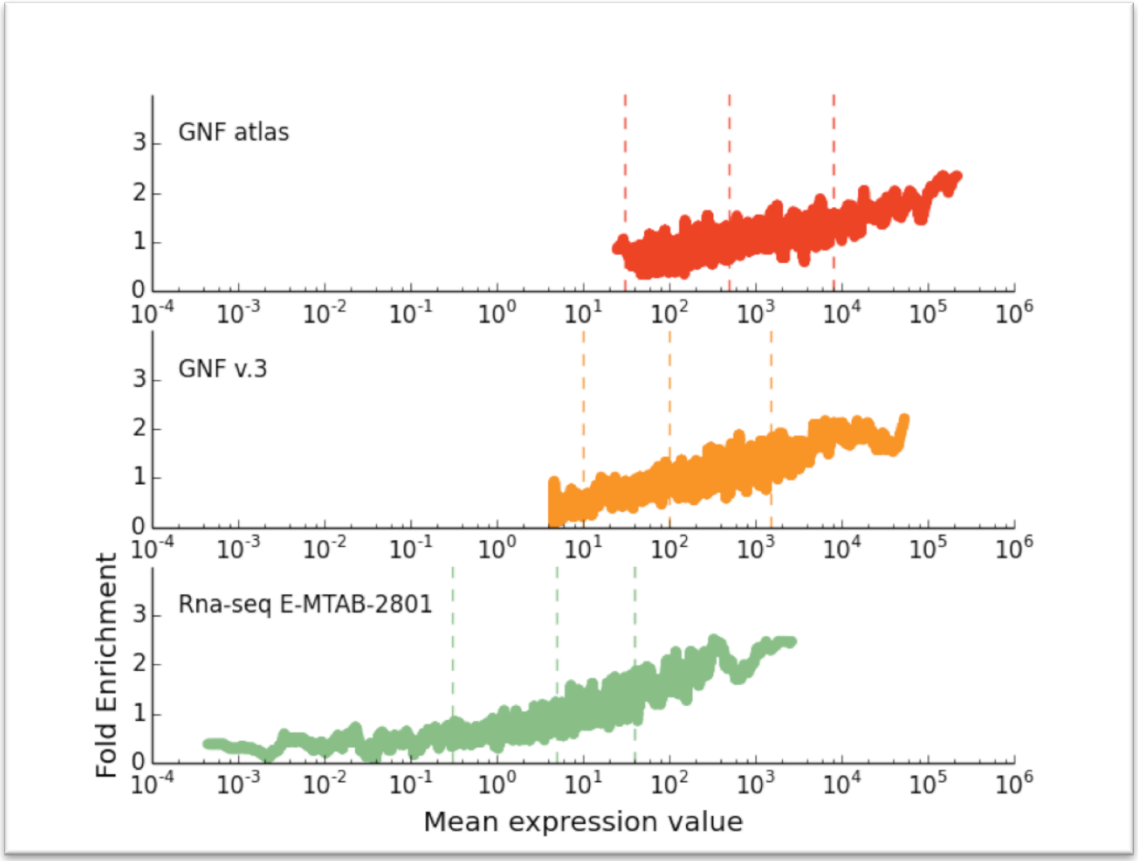
- **4** transcripts, **3** tissues
- **12** transcript-tissue pairs in the expression dataset
- **5** transcript-tissue pairs in the gold standard dataset
- random_chance = **5/12**

fold_enr = (1/5)/(5/12) = 0.48



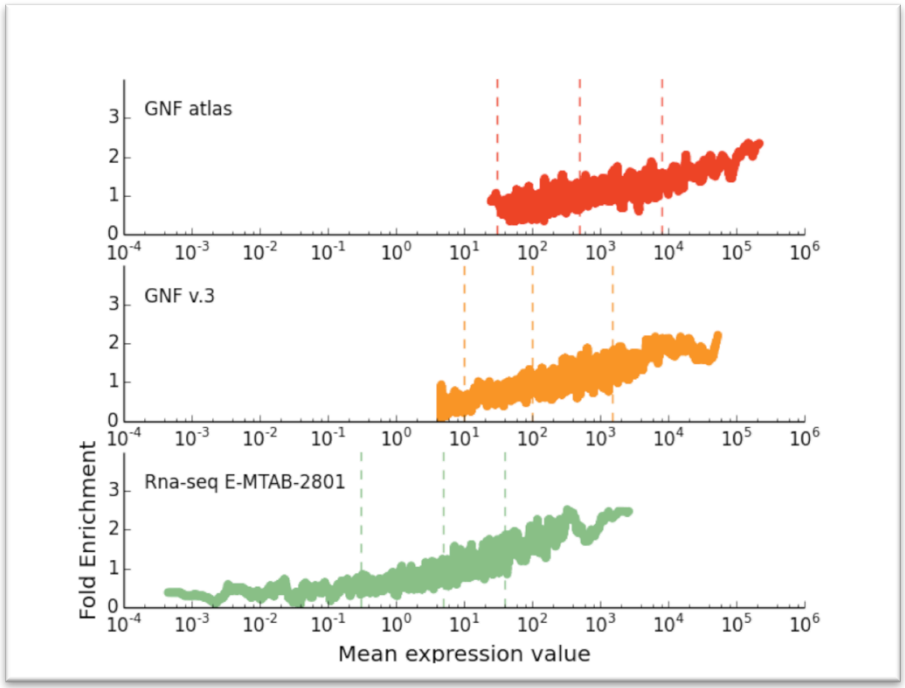
Fold enrichment of transcript-tissue associations found in the gold standard dataset compared to random chance, calculated over sliding windows.

Mouse datasets, UniprotKB mouse annotation.

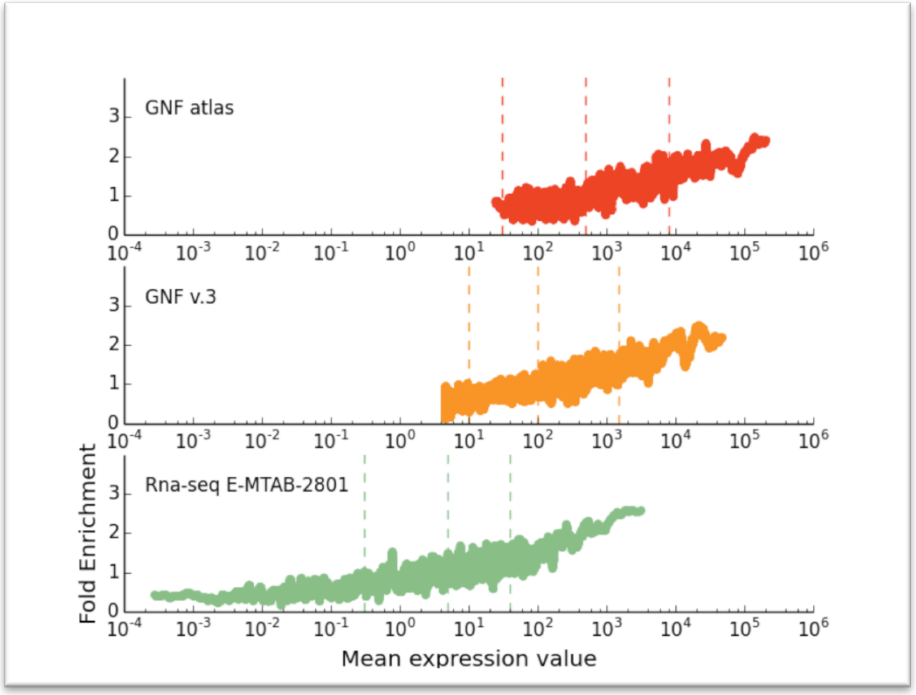


Testing orthology transfer for mouse datasets.

Mouse UniprotKB vs. orthology human as gold standard.



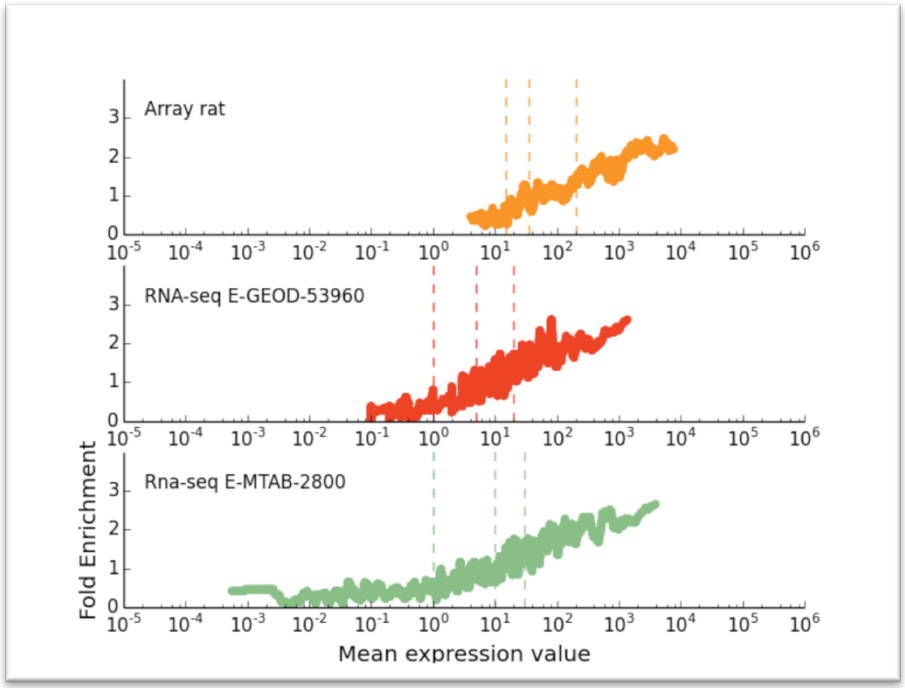
Mouse Uniprot



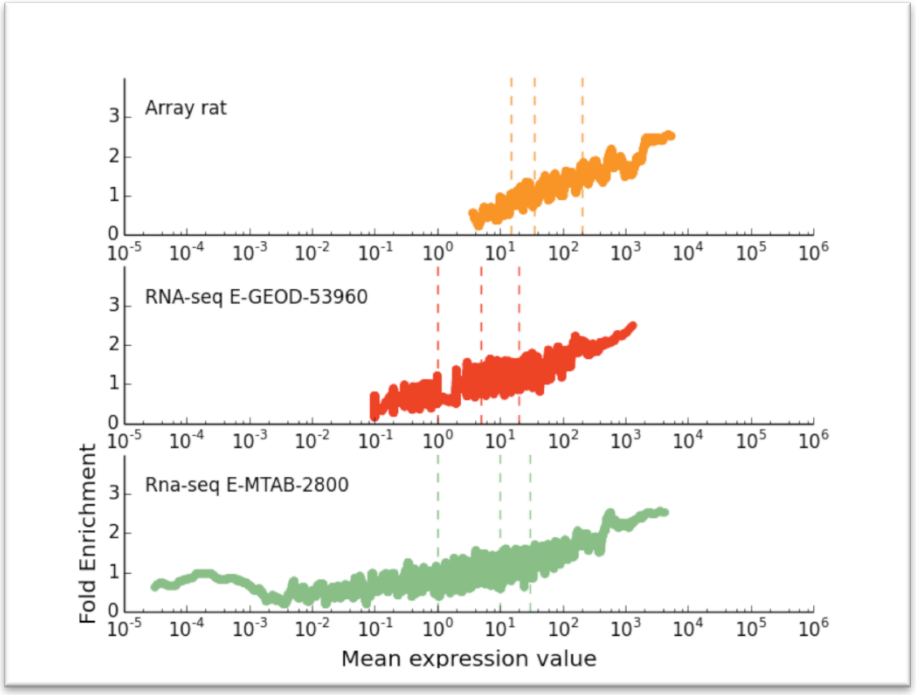
Orthology human

Orthology transfer for rat datasets

Rat UniprotKB vs. orthology human as gold standard.



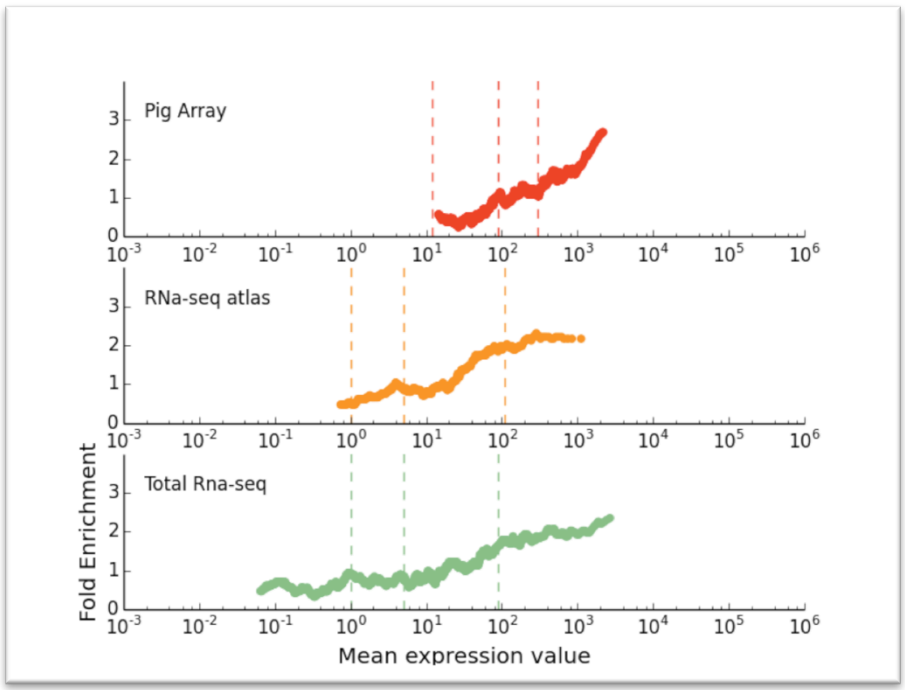
Rat Uniprot



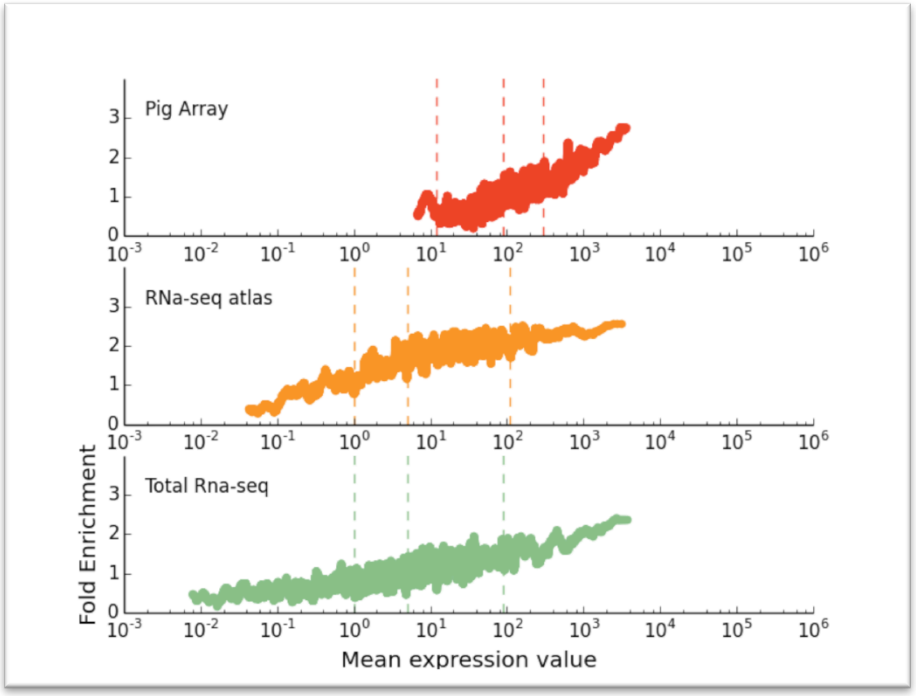
Orthology human

Orthology transfer for pig datasets

Pig UniprotKB vs. orthology human as gold standard.



Pig Uniprot



Orthology human

Outlook

- Include more datasets (e.g. Mouse ENCODE)
- Test confidence scoring scheme on ncRNAs
- Develop metrics for comparisons between organisms
- Web resource available soon
  - Enrichment analysis
  - Integration with EggNOG and STRING (Szklarczyk et al. 2015)

Acknowledgements

- Alberto Santos
- Christian Anthon
- Lars Juhl Jensen
- Jan Gorodkin