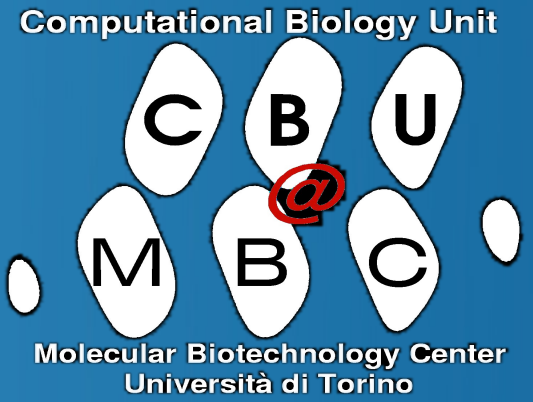
In silico prediction of lncRNA function

Umberto Perron, Paolo Provero and Ivan Molineris

Università degli Studi di Torino – Molecular Biotechnology Center

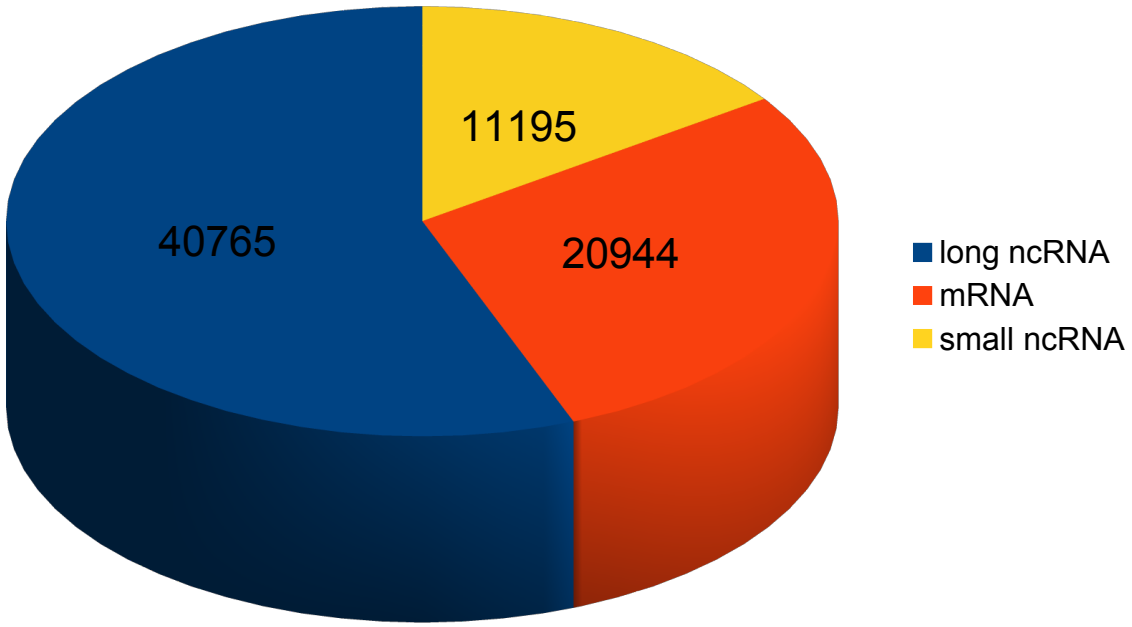


Introducing lncRNAs

- lncRNAs are defined as non-coding transcripts longer than ~200 nucleotides
- They act functionally as RNAs
- Brockdorff N, et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. **1992** Oct 30;71(3):515-26.
- Brannan CI, et al. The product of the H19 gene may function as an RNA. Mol Cell Biol. **1990** Jan;10(1):28-36.
- Okazaki Y, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, Nature 420, 563-573 (5 December **2002**)

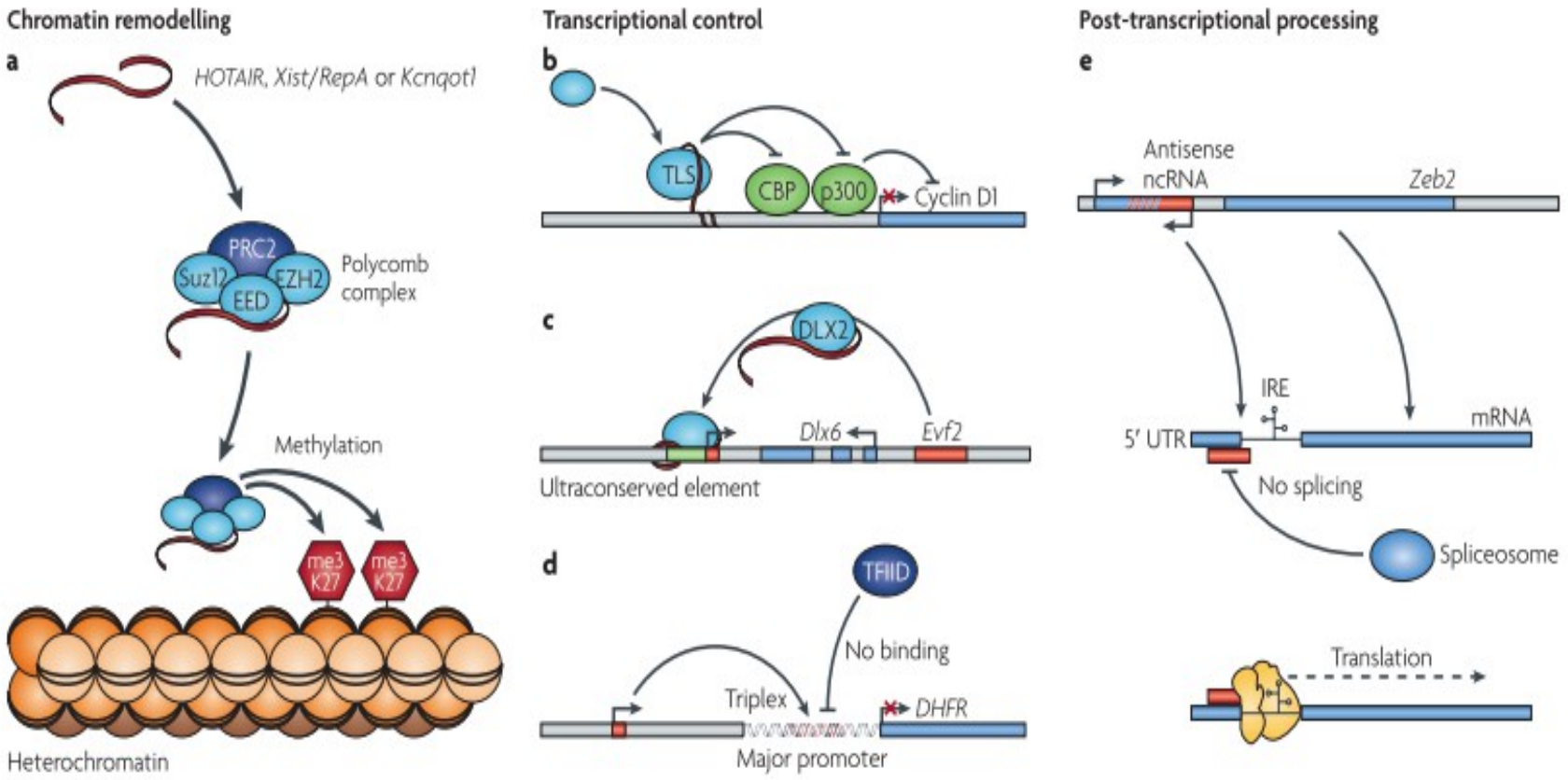
Introducing lncRNAs

- Combined human gene annotations from Ensembl (release 64), NCBI's RefSeq, UCSC Genome Browser and lncRNAs catalogued by Cabili et al. , 2011.



Introducing lncRNAs

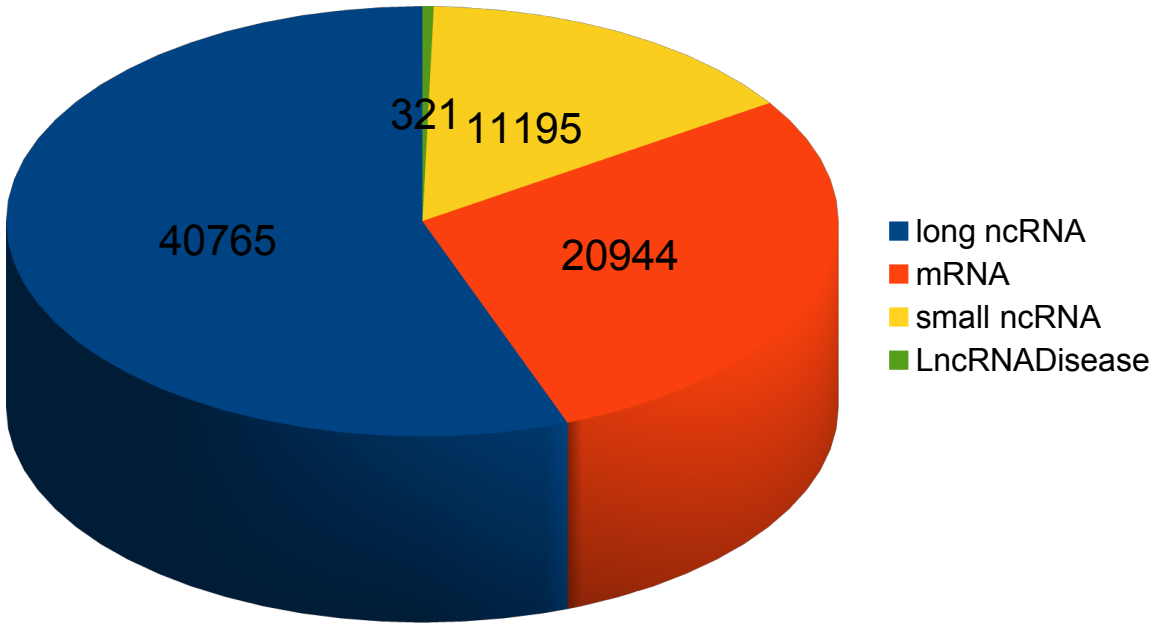
- The existence of a large class of lncRNAs with potential regulatory function is now widely accepted.



Mercer T, 2009

The annotation problem

- At the moment the most autoritative database of lncRNA annotations is LncRNADisease.
- It comprehends 321 lncRNA experimentally linked to pathological conditions.



Genome-wide annotation of lncRNAs

- We know very little about the function of the vast majority of lncRNAs.
- Therefore the need for a genome-wide, systematic approach to infer putative functions for a large number of lncRNAs.

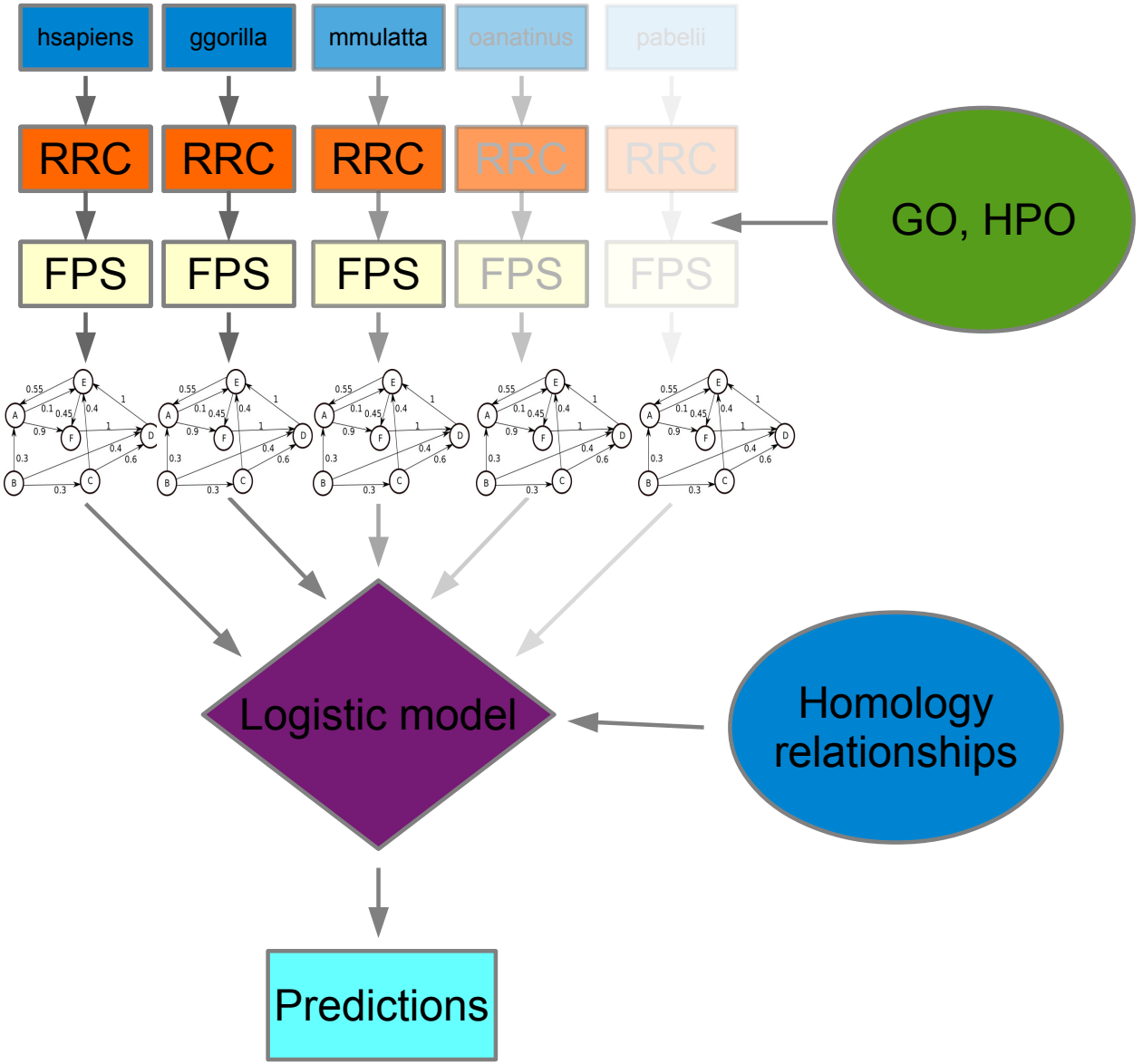
Coexpression networks for functional prediction

- Biological functions are related to network modules rather than single genes.
- In the past our group has found that coexpression networks are a powerful tool to make functional predictions and identify candidate disease-genes among coding genes.
- This is even more true with the integration of human-mouse conserved coexpression (Ala et al. 2008 ; Piro et al. 2011).

Coexpression networks for functional prediction

- We expand and improve on those methods in a multi-species conservation context and we apply them to lncRNA functional prediction.
- Because coding genes are much better annotated than lncRNAs, we aim to project known functional information regarding proteins onto non coding genes.
- We assumed that, if a lncRNA shows an expression profile that correlates with those of a set of coding genes, that lncRNA is probably involved in the same function (guilt by association principle).

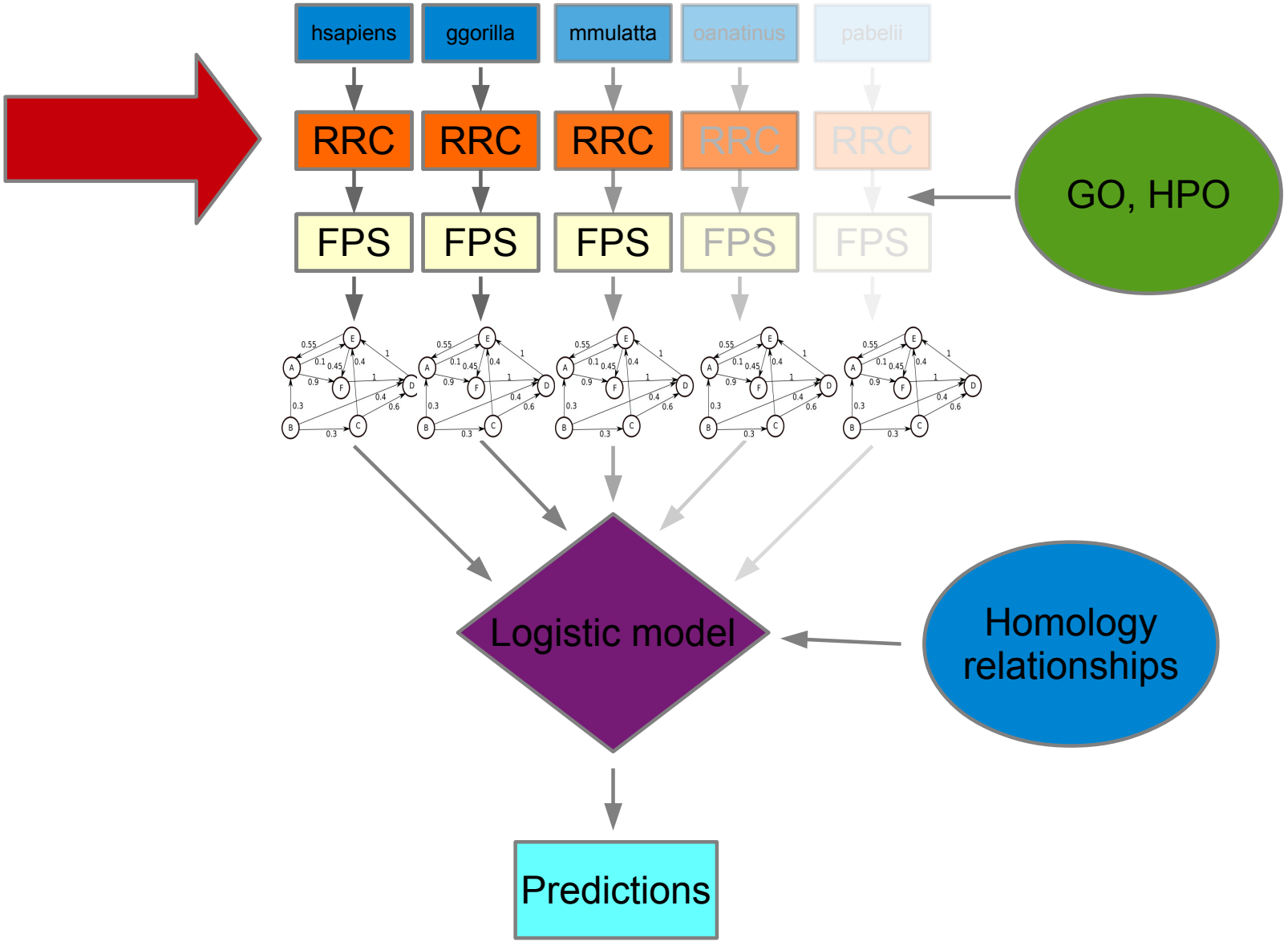
The pipeline



RNA-Seq dataset

- Curated by Necsulea and colleagues in 2014.
- 185 RNA-seq samples
- 10 species (human, chimpanzee, gorilla, orangutan, macaque, mouse, opossum, platypus, chicken and frog)
- 8 organs (cortex or whole brain, cerebellum, heart, kidney, liver, placenta, ovary and testes)
- The expression of 22000 protein-coding genes and 5400 lncRNAs was measured.

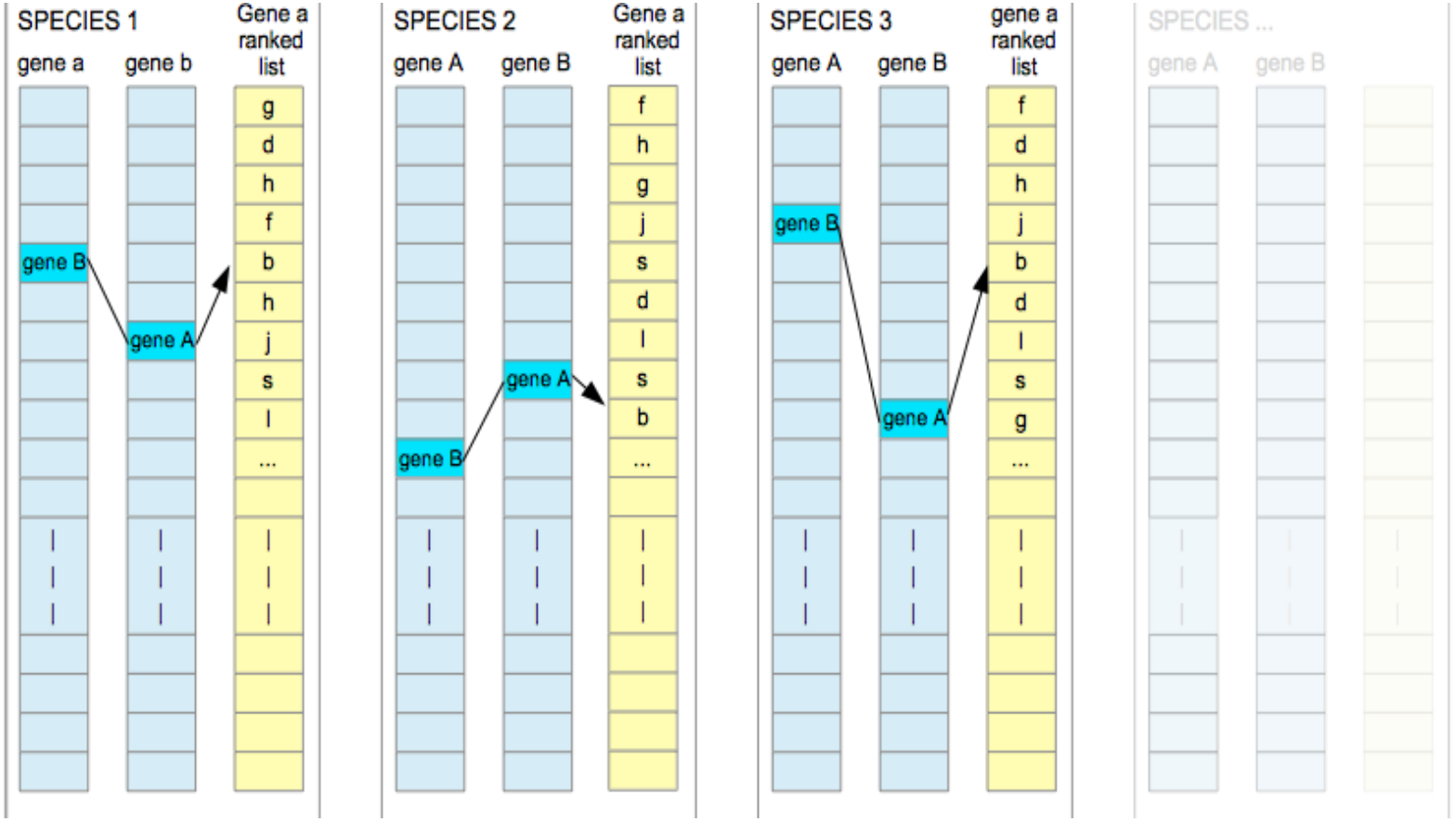
The pipeline



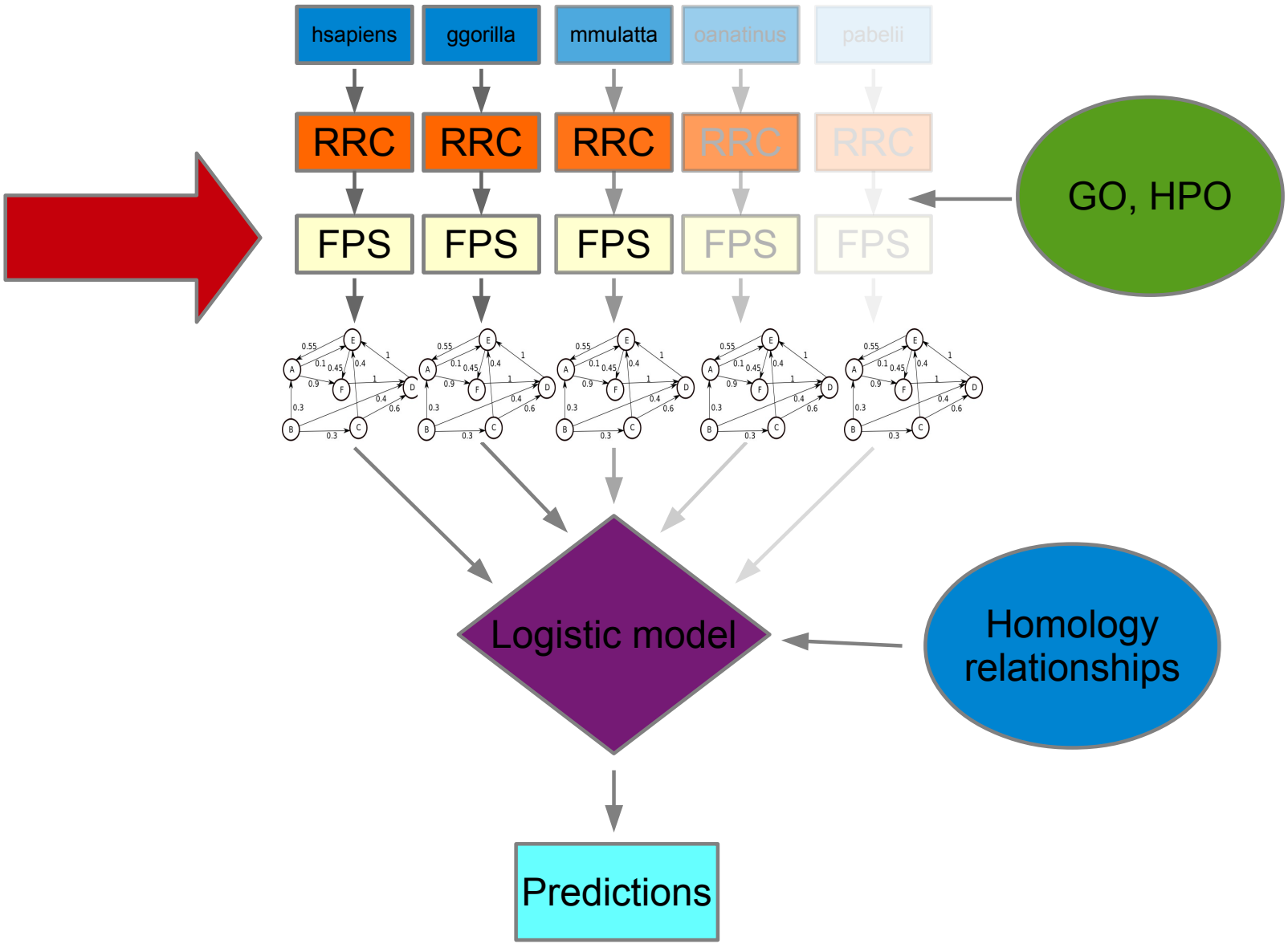
Reciprocal rank correlation

- We defined *w(a, b)* as the RRC between gene a and gene b.

$$w(a,b)=min(r(a,b),r(b,a))$$



The pipeline



Ontologies

- We used two controlled vocabularies to annotate genes:
  - Human Phenotype Ontology (HP) version 67
  - Gene Ontology (GO) downloaded from ensembl version 75
- For both GO and HP we only used terms that had between 4 and 2000 genes.

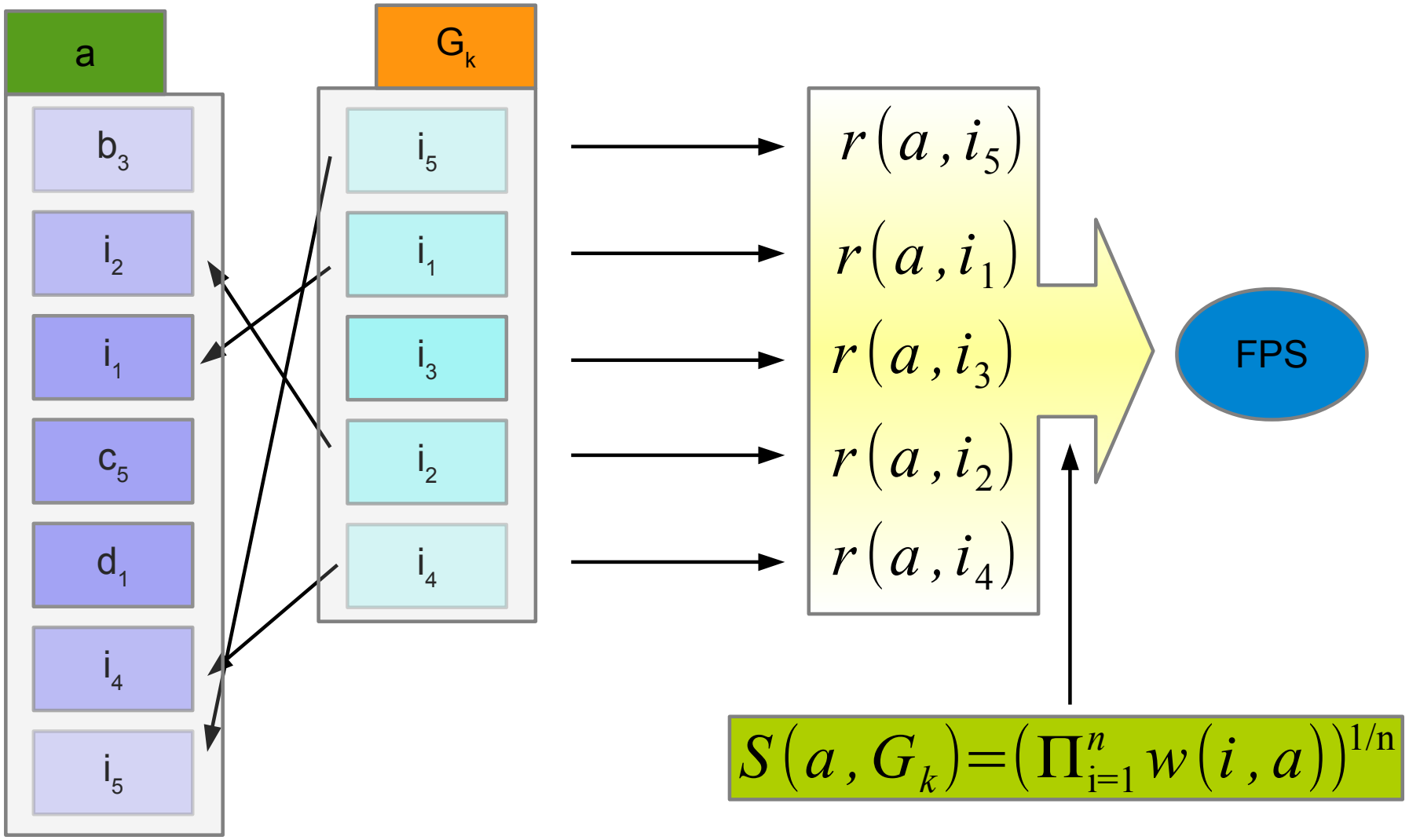
Functional prediction score

- Given a gene *a* and a set of *n* genes G_k annotated to some ontology keyword *k*, we computed:

$$S(a, G_k) = \left(\Pi_{i=1}^{n} w(i, a)\right)^{1/n}$$

- This score has the properties of the rank product described in Breitling et Al. 2006.
- It was developed for differential expression analysis.
- It provides a straightforward and statistically stringent way to identify biologically relevant expression changes.
- RP results are reliable in highly noisy data.

Functional prediction score



Leave-one-out validation

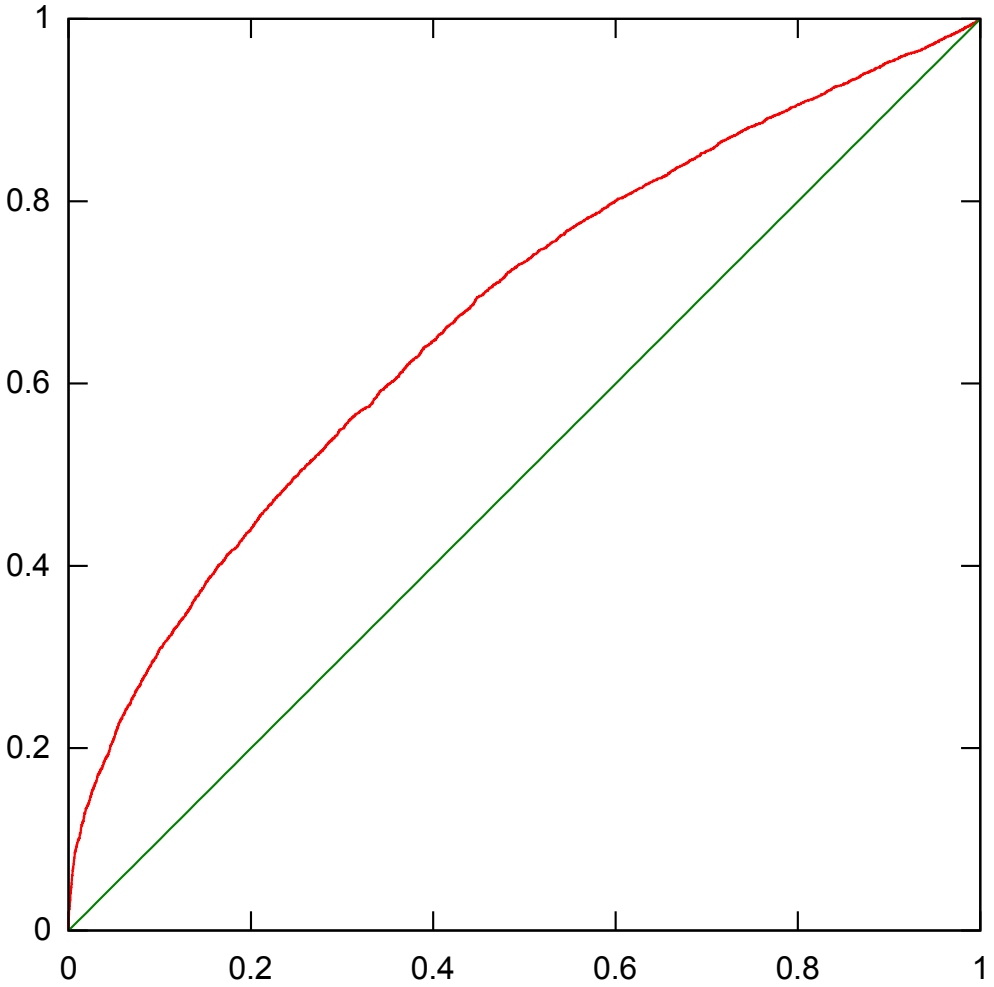
- We considered only genes already annotated with GO terms (mostly coding genes)
- we performed a ROC analysis:
  - If G_k is the set of all genes annotated to k; each gene $a \notin G_k$ is considered as negative.
  - The curve is computed based on this binary classif cations and the FPS
  - if $a \in G_k$ we use instead $S\left(a, G_k - \{a\}\right)$
  - for k the number of positives is much less than the number of negatives, therefore we considered, for each k independently, only a randomly chosen subset of size G_k.

Non-redundant Gene Ontology

- For ROC analysis we experimented with a “non redundant” version of GO where:
  - we consider each domain (BP, CC, MF) separately
  - we annotate gene *a* with the most specific term among those it is associated with.

ROC analysis

- Human network, AUC 0,6474

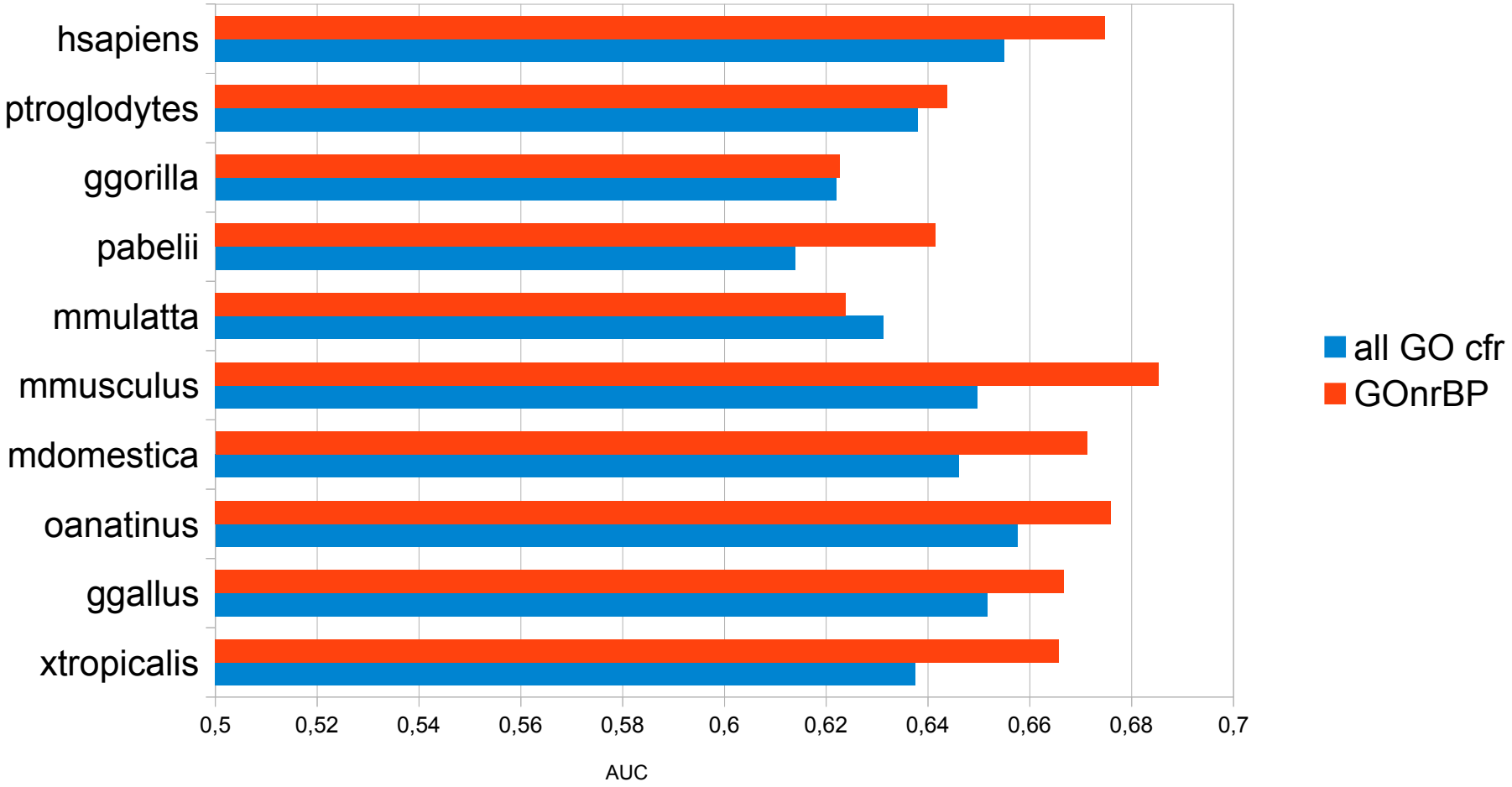


Homology relatioships

- We need to project the GO annotations available for humans into other species
- Therefore to translate non-human genes into human genes.
- We used:
  - orthologous genes downloaded from biomart.org [ensembl 75] for coding genes
  - Necsulea et al.'s lncRNAs families.
- We also took into account one-to-many homology relationships.

ROC analysis

- All single-species networks with non-redundant GO

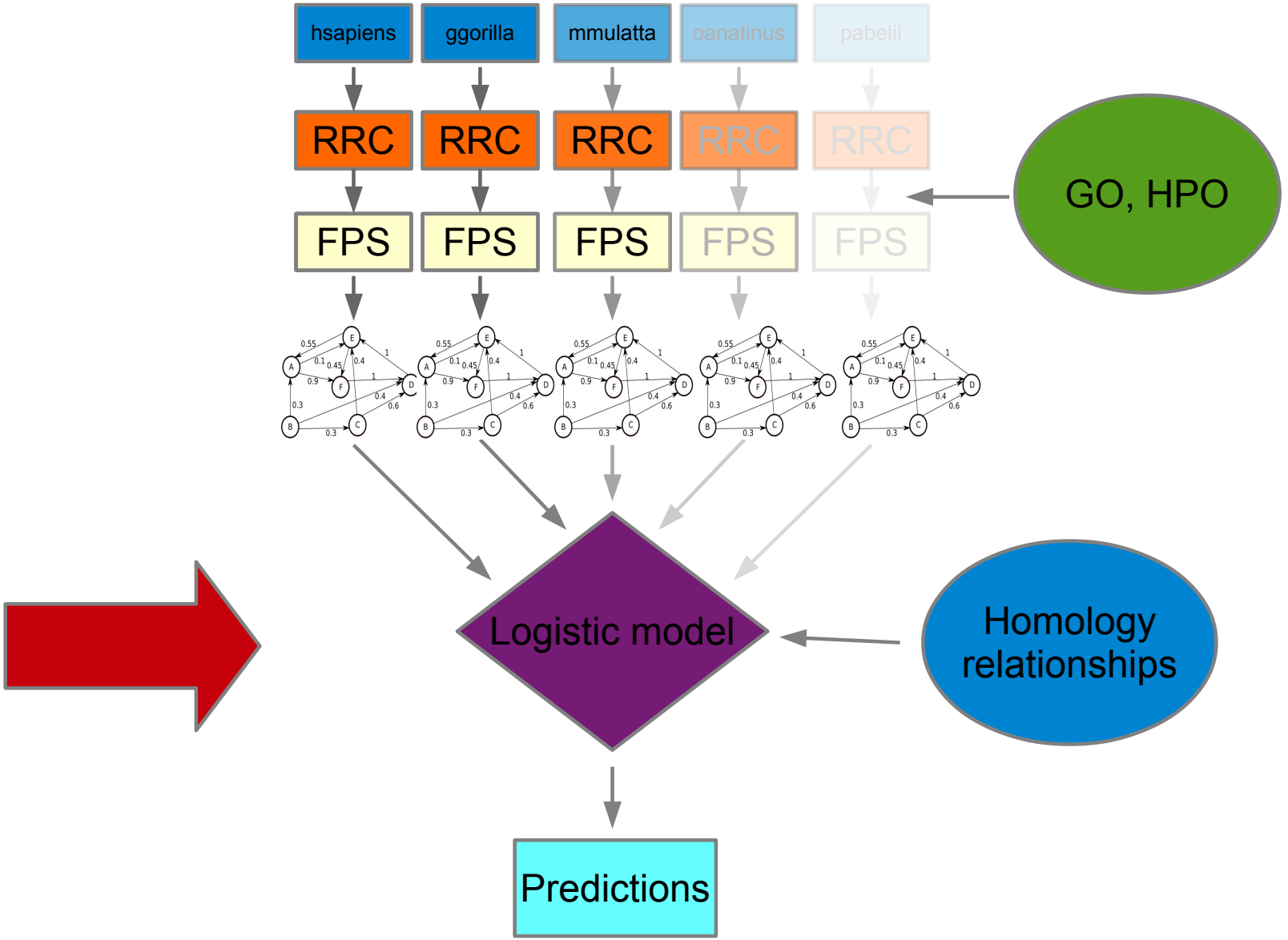


Hsapiens Predictions

- Here the hsapiens single-species FPS is used to rank our predictions.

Human network		
GO_Id	GO_term	Gene_Id
GO:0042738	exogenous drug catabolic process	ENSG00000228826
GO:0060968	regulation of gene silencing	blastnggorilla,Locus_369905
GO:0060397	JAK-STAT cascade involved in growth hormone signaling pathway	Locus_166952
GO:0042738	exogenous drug catabolic process	CUFF,57965
GO:0042773	ATP synthesis coupled electron transport	blastnptroglodytes,CUFF,383474
GO:0009820	alkaloid metabolic process	XLOC_002353
GO:0035886	vascular smooth muscle cell differentiation	Locus_175359
GO:0010873	positive regulation of cholesterol esterification	blastnggorilla,Locus_196608
GO:0045653	negative regulation of megakaryocyte differentiation	blastnmmusculus,ENSMUSG00000090913
GO:0051592	response to calcium ion	blastnmmulatta,CUFF,184706
GO:0010873	positive regulation of cholesterol esterification	CUFF,57965

The pipeline



Logistic model

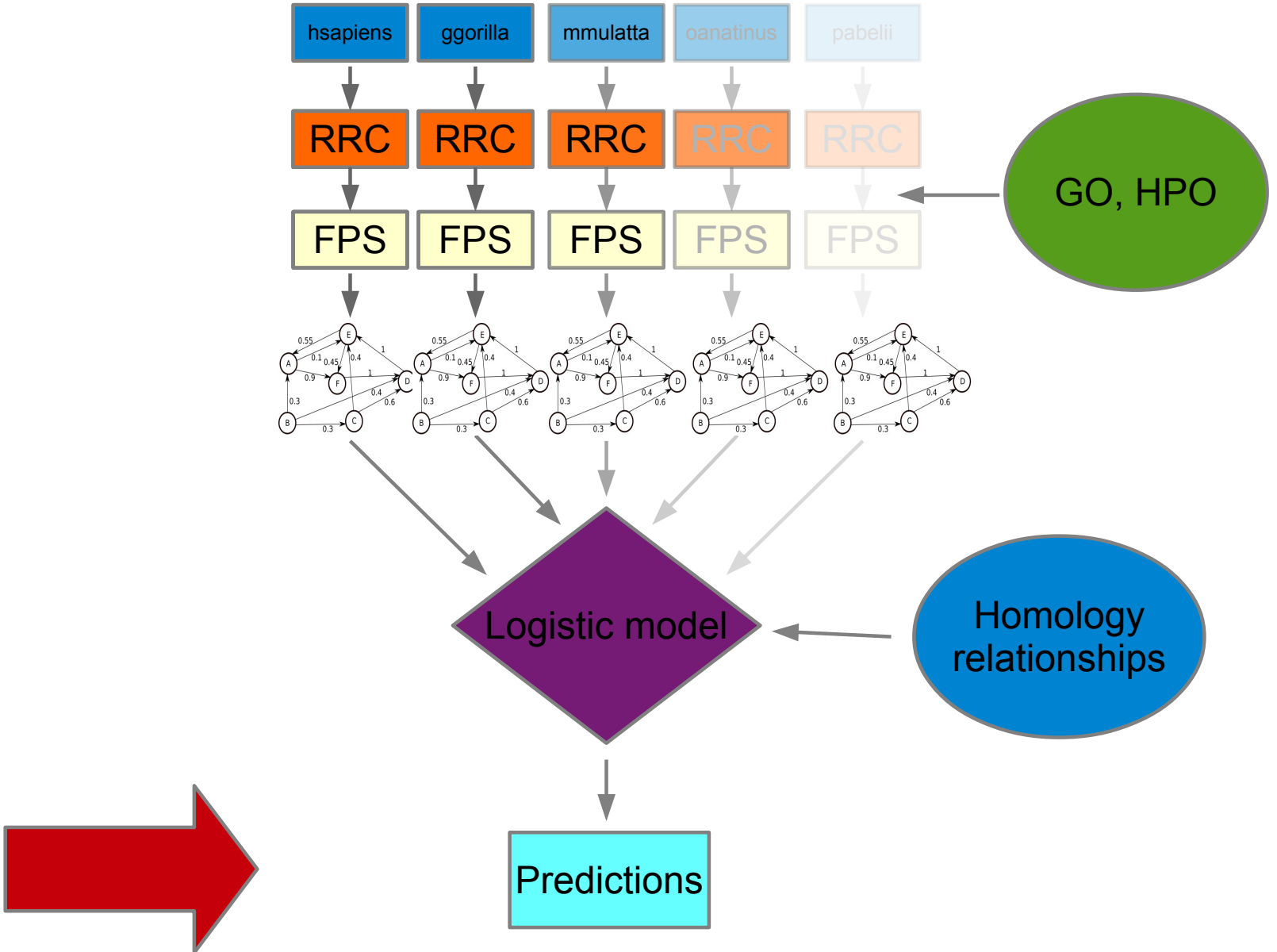
- A binary logistic model is simple way of integrating our FPSs into a single score.
- It is used to estimate the probability of a binary response based on one or more predictors.
  - Here the binary variable is 1 if gene *a* is annotated with a term *k,* 0 if otherwise.
  - The predictors are the single-species FPSs.

Validation

- Log odds ratios are a mesure of the extra information carried by each species on top of the information provided by all other species.

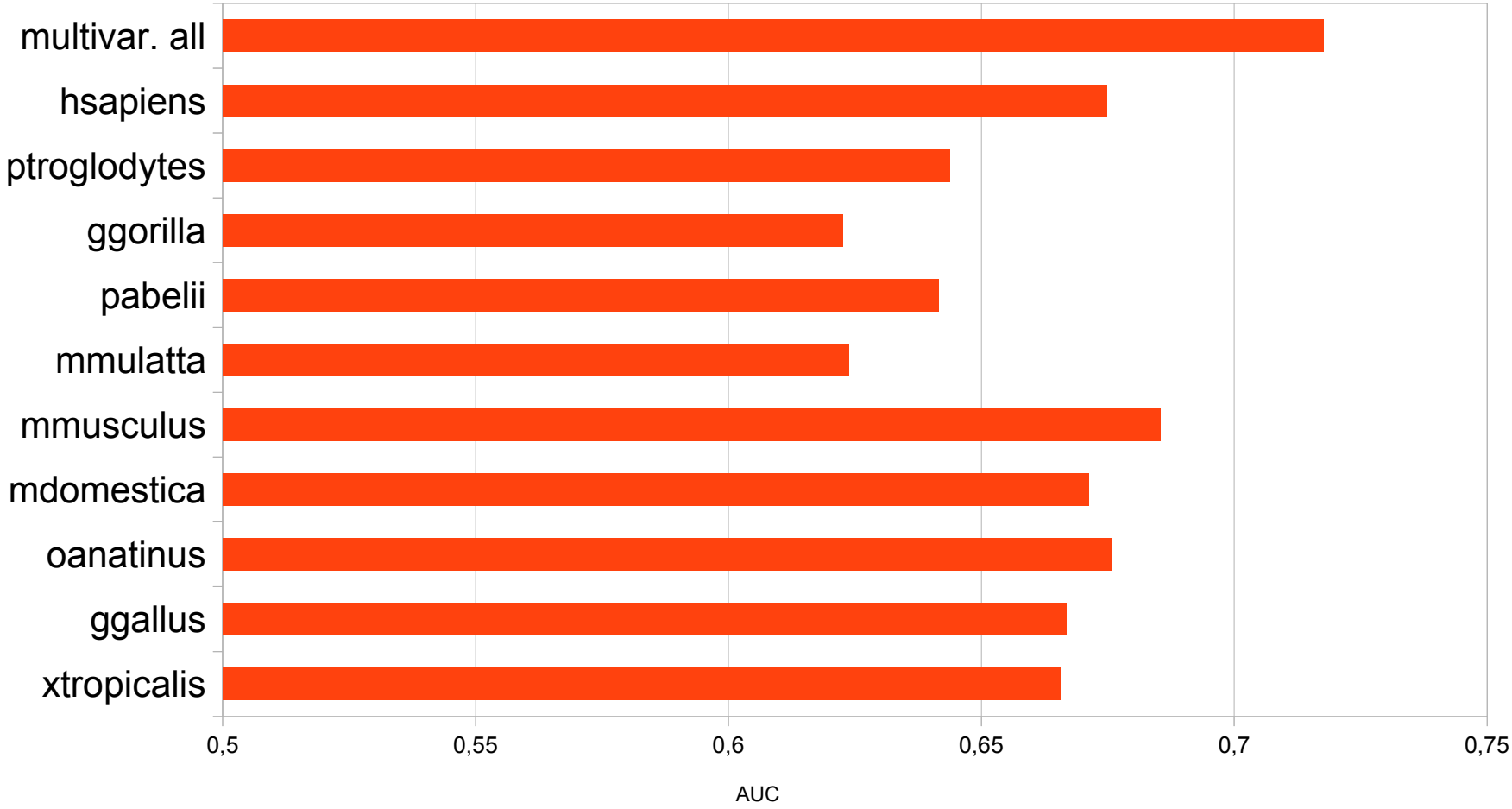
	Full model		Primates only		Univariate	
	Log_odd	Std_err	Log_odd	Std_err	Log_odd	Std_err
hsapiens	0,7477	0,0092	1,5412	0,0105	3,5867	0,0122
ptroglodytes	0,1598	0,0108	0,6391	0,0113	2,0450	0,0102
ggorilla	-0,2212	0,0112	0,1023	0,0126	2,3578	0,0122
pabelii	0,0842	0,0094	0,1480	0,0091	1,1600	0,0089
mmulatta	-0,1000	0,0102	0,0518	0,0098	1,6657	0,0090
mmusculus	0,5060	0,0091	NA	NA	1,8166	0,0084
mdomestica	0,3181	0,0083	NA	NA	0,6954	0,0058
oanatinus	0,3137	0,0067	NA	NA	0,5339	0,0045
ggallus	0,1974	0,0066	NA	NA	0,4852	0,0046
xtropicalis	0,2753	0,0068	NA	NA	0,6022	0,0047

The pipeline



Validation

- Logistic Model ROC analysis with non-redundant GO.



Predictions

- The output of the logistic model is used as a score to rank our predictions.

Primates network		
GO_Id	GO_term	Gene_Id
GO:0042738	exogenous drug catabolic process	ENSG00000248740
GO:0030299	intestinal cholesterol absorption	ENSG00000254235
GO:0009820	alkaloid metabolic process	XLOC_002353
GO:0055003	cardiac myofibril assembly	Locus_175215
GO:0055003	cardiac myofibril assembly	ENSG00000236208
GO:0030449	regulation of complement activation	ENSG00000223956
GO:0060968	regulation of gene silencing	ENSG00000232131
GO:0038003	opioid receptor signaling pathway	ENSG00000231533
GO:0032490	detection of molecule of bacterial origin	H19X
GO:0034587	piRNA metabolic process	ENSG00000235740

Predictions

- The output of the logistic model is used as a score to rank our predictions.

Primates network		
GO_Id	HP_term	Gene_Id
HP:0001563	Fetal polyuria	ENSG00000251151
HP:0005305	Cerebral venous thrombosis	ENSG00000223956
HP:0003715	Myofibrillar myopathy	ENSG00000236208
HP:0002914	Hyperchloridura	ENSG00000223392
HP:0011858	Reduced factor IX activity	ENSG00000248740
HP:0004431	Complement deficiency	XLOC_002353
HP:0010444	Pulmonary insufficiency	H19X
HP:0008151	Prolonged prothrombin time	ENSG00000249364
HP:0011713	Left bundle branch block	Locus_175215
HP:0000206	Glossitis	ENSG00000234506

Conclusions

- We suggest that phylogenetically conserved coexpression is a valuable tool to investigate transcript function.
- Our leave-one out analysis on coding genes conf rms the predictive power of the guilt by association principle.
- We are therefore able to project onto lncRNAs the abundant functional annotations available on protein-coding genes using gene networks built upon gene expression data.

Perspectives

- As we have pointed out in the past (Piro et Al. 2011) it is advantageous to take into consderation tissue specif ɔity alongside coexpression when considering the problem of functional annotation.
- Threfore we want to apply our pipeline to the GTeX dataset.
- We also plan to integrate other sources of information such as miRNA-lncRNA interactions in order to approach a gold-standard functional prediction tool.

Acknowledgements

Paolo Provero

Ivan Molineris

Elena Grassi

Ugo Ala

Mattia Forneris

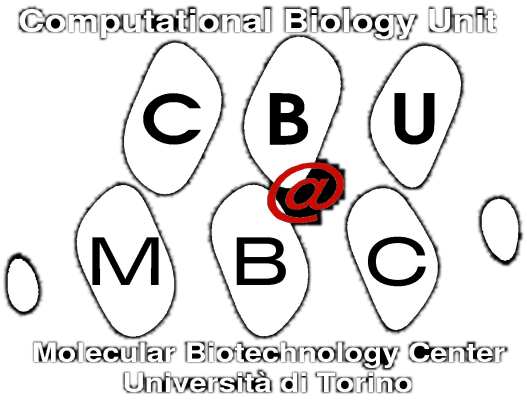
Davide Marnetto

Simona Baghai

Elisa Mariella

Marika Catapano

for their advice, mentoring and invaluable support.



The pipeline