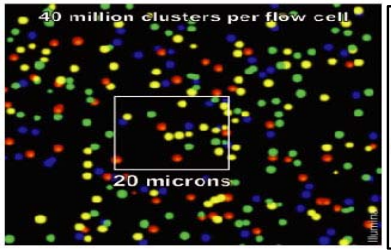
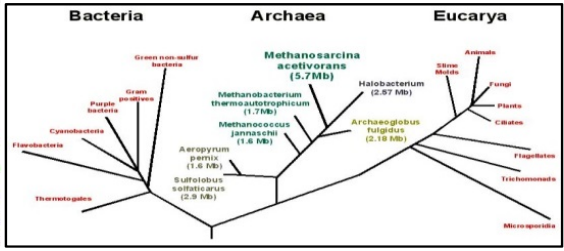
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA

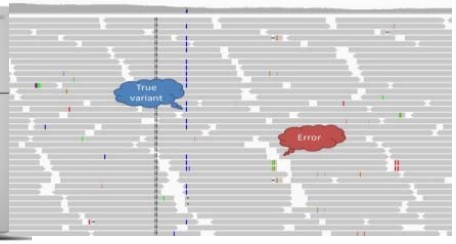


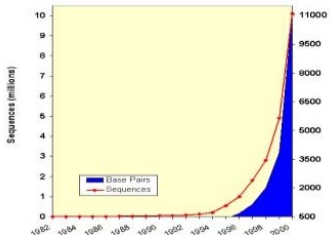


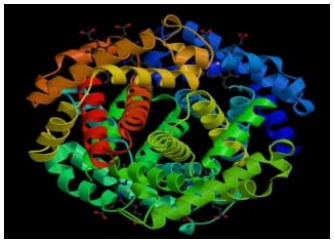
Case Study I: Origination of New Genes

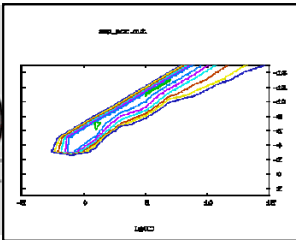
美国芝加哥大学生态与演化生物学系 龙漫远
北京大学生物信息学中心 魏丽萍

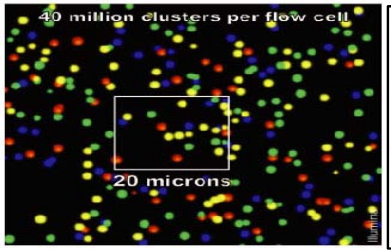
Manyuan Long, Department of Ecology and Evolutionary Biology, University of Chicago
Liping Wei, Center for Bioinformatics, Peking University

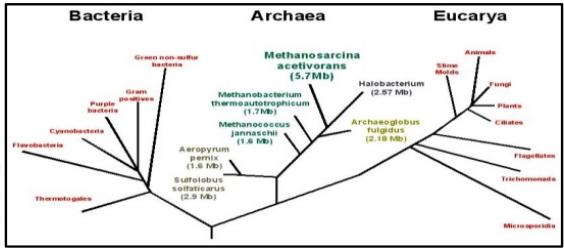










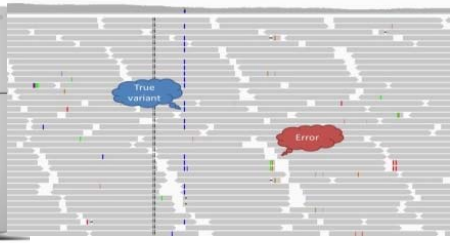


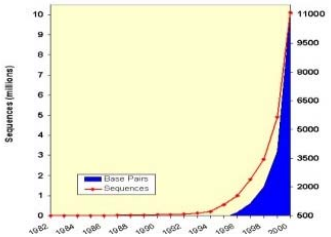
Unit 4: Origination of *de novo* Genes from Noncoding RNAs

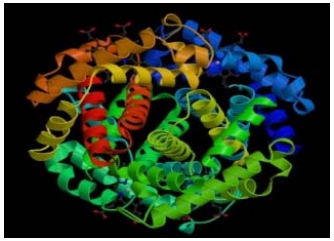
北京大学生物信息学中心 魏丽萍

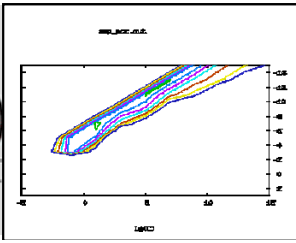
Liping Wei, Ph.D.

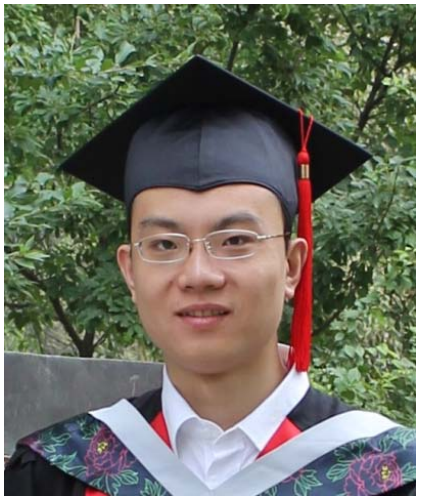
Center for Bioinformatics, Peking University











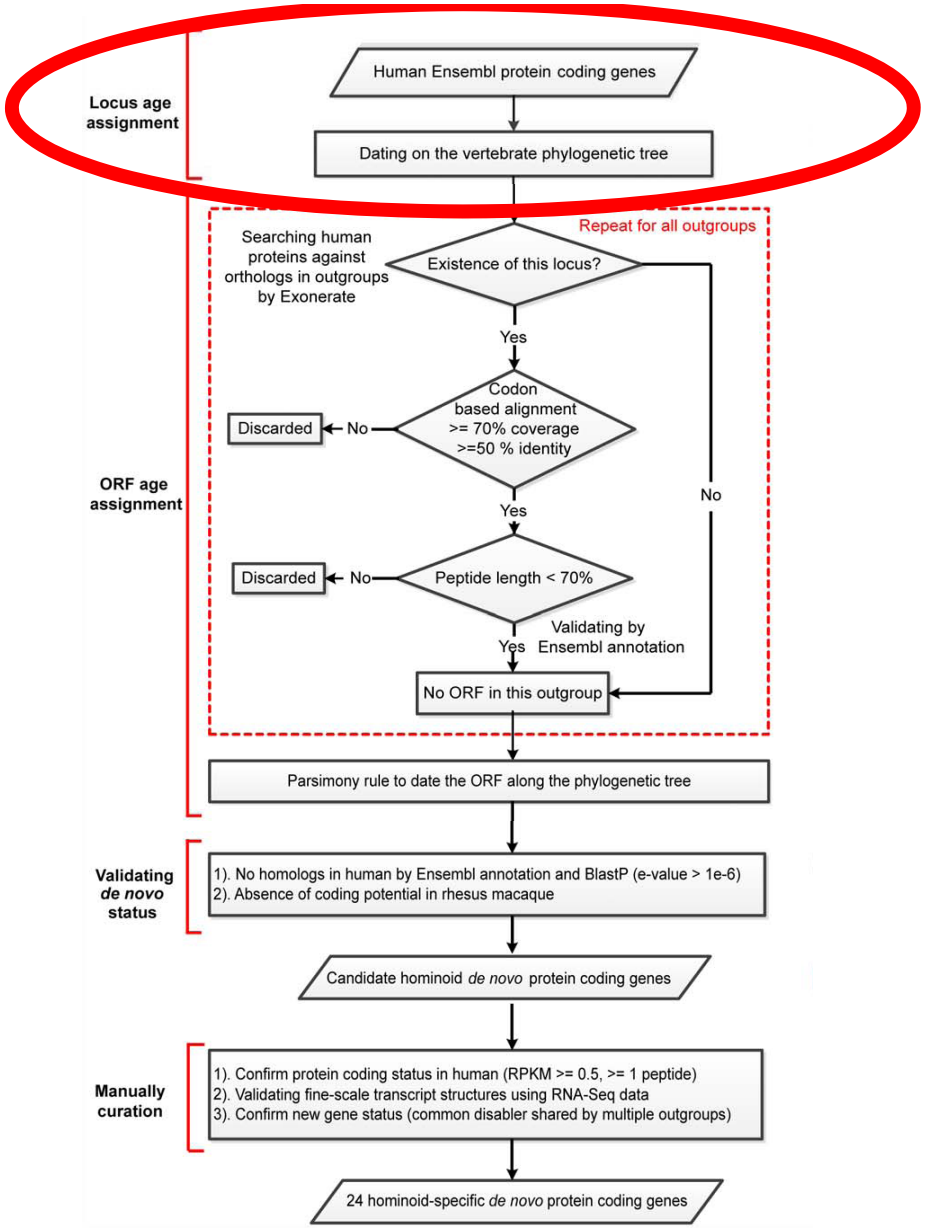
Chen Xie

Yong Zhang

Chuan-Yun Li

How many other human-specific *de novo* genes are there?

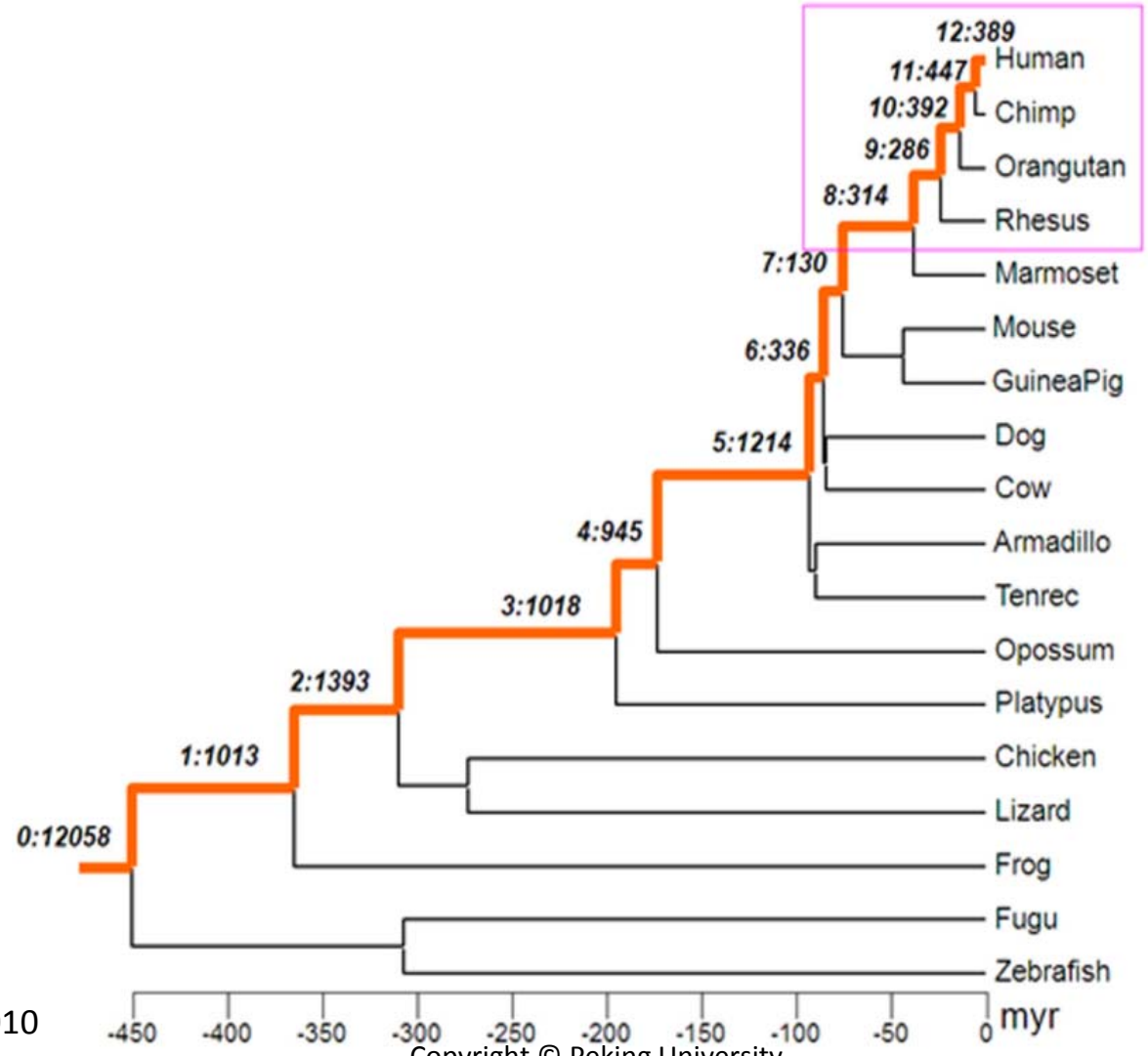
Where did they originate from?



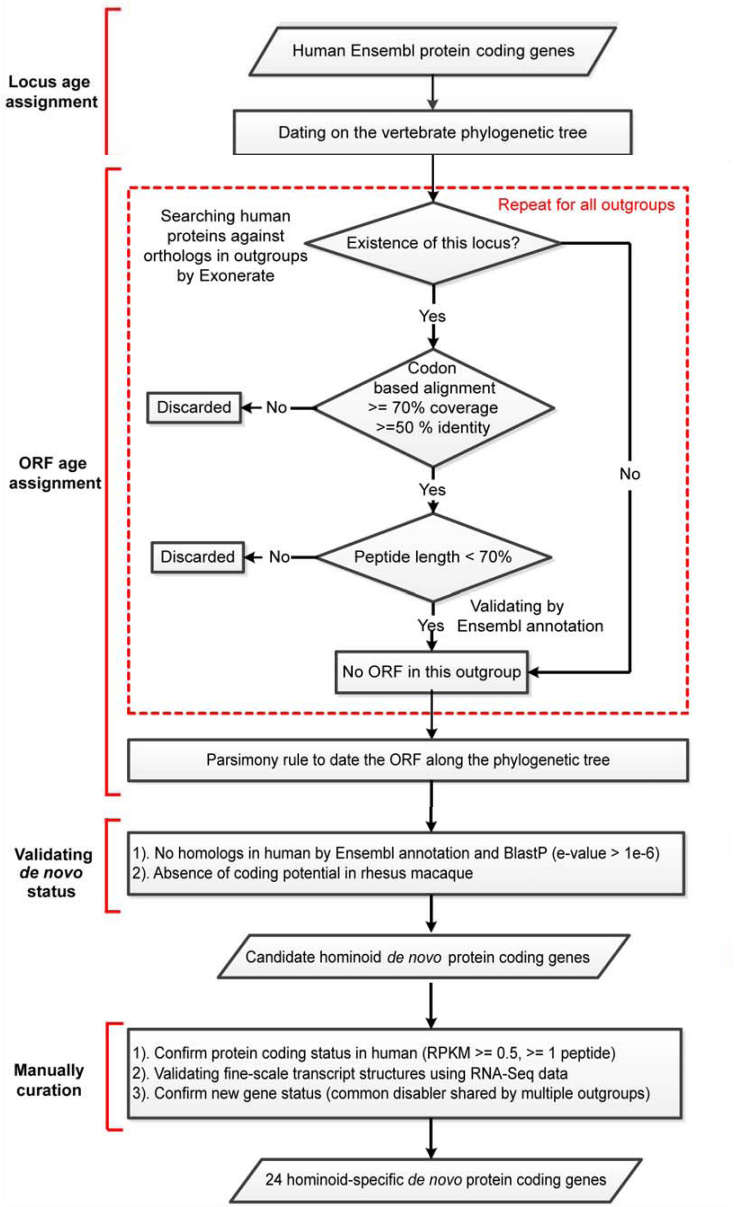
Genome-wide identification of human- and human-chimpanzee-specific *de novo* genes

Xie *et al.*, *PLoS Genet.*, 2012

Inferring the origination times of human gene loci



Zhang *et al.*, *PLoS Biol.*, 2010



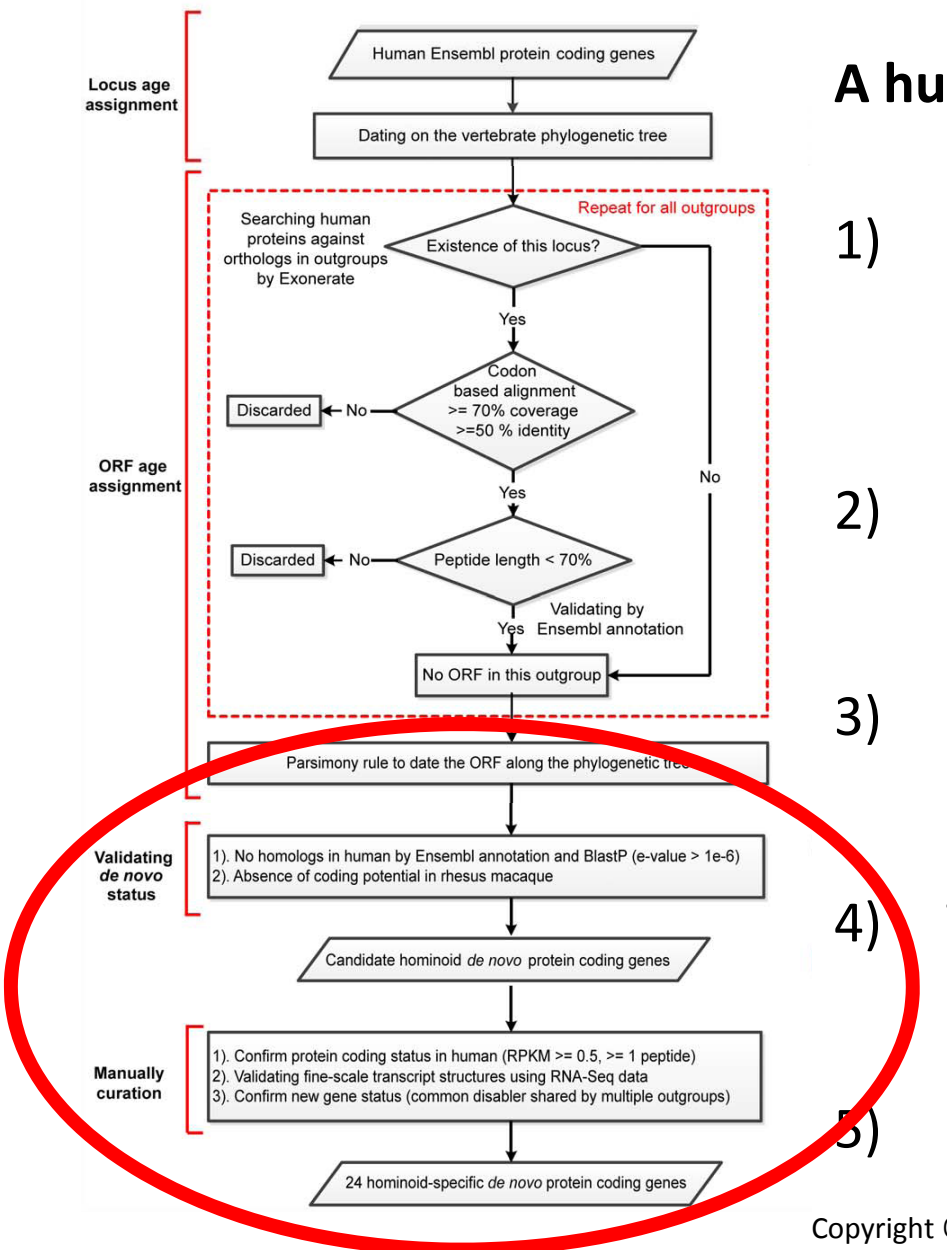
Inference of age of ORF

For each locus in each outgroup species, an ORF is considered absent if

(1) Reliable codon-based alignment (i.e., >=70% coverage and >=50% identity) shows that the maximum continuous peptide before the first ORF disabler was shorter than 70% of the human ORF;

AND

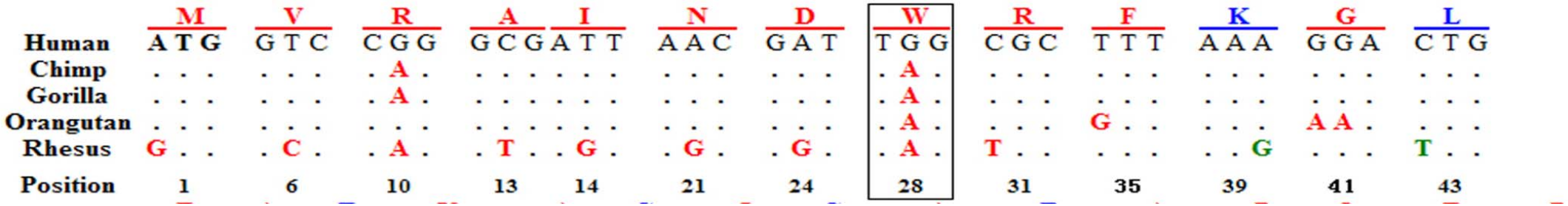
(2) Ensembl annotation did not identify any ortholog.



A human gene is considered *de novo* if

1) Intact ORF with RNA-Seq RPKM score larger than 0.5 in at least one of the nine human tissues; standard start and stop codons and intron lengths no less than 18 nucleotides
2) At least one unique supporting peptide from mass spectrometry data in PeptideAtlas or PRIDE
3) BLASTP and Ensembl found no homologous proteins in other species and no paralogous proteins in human (E-value cutoff of 10^{-6})
4) The outgroup species have no intact ORF. (Genes with the stop codon-containing exon spliced out in rhesus macaque were discarded.)
5) Multiple outgroups share a common disabler

Using common disablers to rule out the possibility of gene loss



	M	V	R	A	I	N	D	W	R	F	K	G	L
Human	A T G	G T C	C G G	G C G	A T T	A A C	G A T	T G G	C G C	T T T	A A A	G G A	C T G
Chimp	. . .	. . .	. A .	. . .	. . .	. . .	. . .	. A .	. . .	. . .	. . .	. . .	. . .
Gorilla	. . .	. . .	. A .	. . .	. . .	. . .	. . .	. A .	. . .	. . .	. . .	. . .	. . .
Orangutan	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. A .	. . .	G . .	. . .	A A .	. . .
Rhesus	G . .	. C .	. A .	. T .	. G .	. G .	. G .	. A .	T . .	. . .	. . G	. . .	T . .
Position	1	6	10	13	14	21	24	28	31	35	39	41	43

Li, *et al*, PLoS *Comp Biol*, '10

24 hominoid-specific *de novo* originated new protein-coding genes were identified

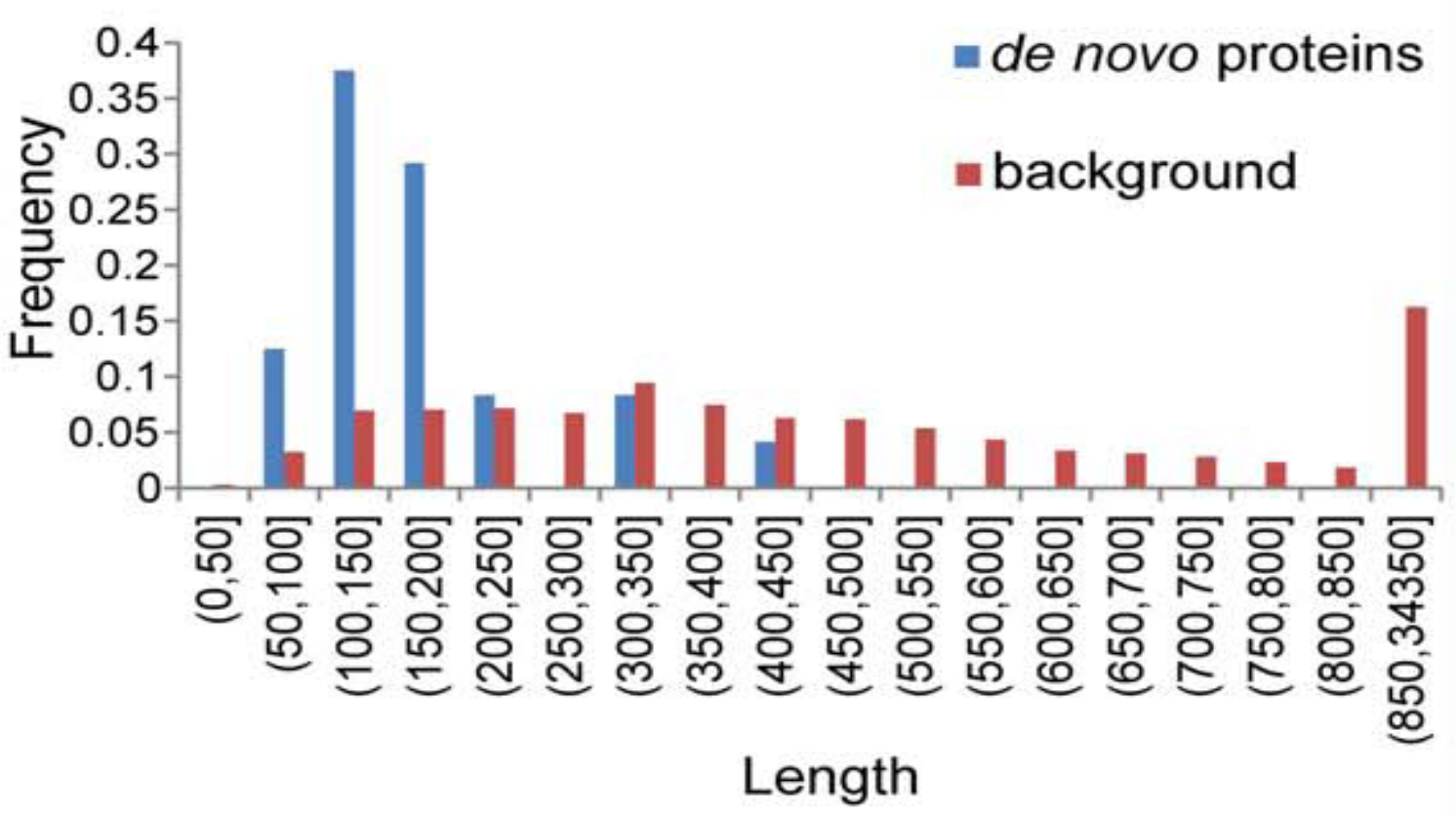
11 encode proteins only in human

7 encode proteins in both human and chimpanzee

6 encode proteins in human, chimpanzee and orangutan

All of them do not encode proteins in rhesus macaque and other out-group species

The gene products are generally smaller



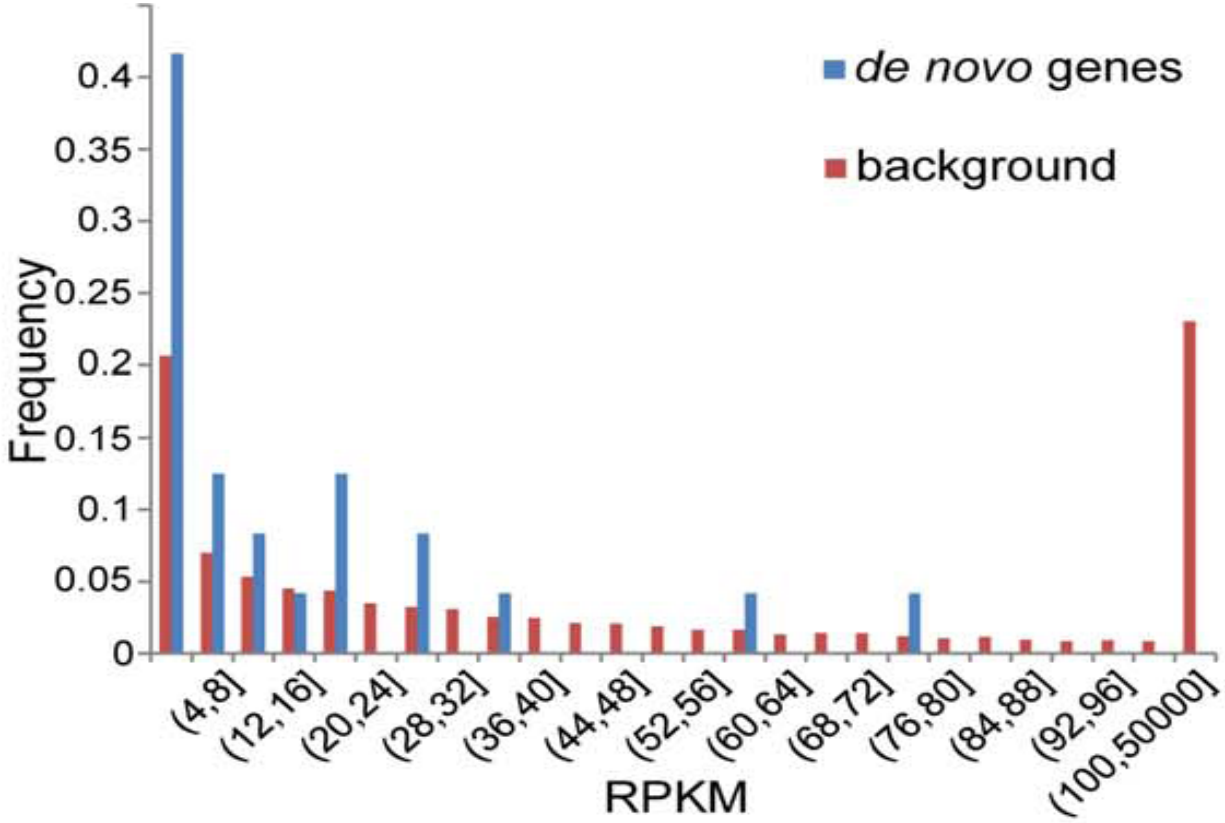
Xie *et al.*, *PLoS Genet.*, 2012

Median = 150.5, P-Value = 4.1×10^{-10}

18/24 have single coding exon

Alu elements contribute to exons of 8 genes and splicing sites of 2 genes

The transcripts are expressed at relatively lower levels



P-Value = 0.037

Xie *et al.*, *PLoS Genet.*, 2012

19 of the 24 de novo genes showed evidence to co-opt the transcriptional context such as antisense and bi-directional promoters.

How did hominoid-specific *de novo* protein-coding genes originate from ancestral non-coding DNAs?

ORF-first or transcription-first?

origination of ORF→transcription→translation

versus

transcription of noncoding RNA→acquisition of ORF→translation

We integrated and analyzed RNA-Seq data from 19 tissues from human, chimpanzee, and rhesus macaque

	Prefrontal cortex	Cerebellum	Testis	Liver	Heart	Skeletal muscle	Adipose
Human	√	√	√	√	√	√	√
Chimp	√	√	√	√	√	×	×
Rhesus	√	√	√	√	√	√	√

Wang *et al.*, *Nature*, 2008 Blekhman *et al.*, *Genome Res.*, 2010 Brawand *et al.*, *Nature*, 2011

20 out of the 24 hominoid-specific *de novo* protein coding genes exist as noncoding RNA in outgroup species

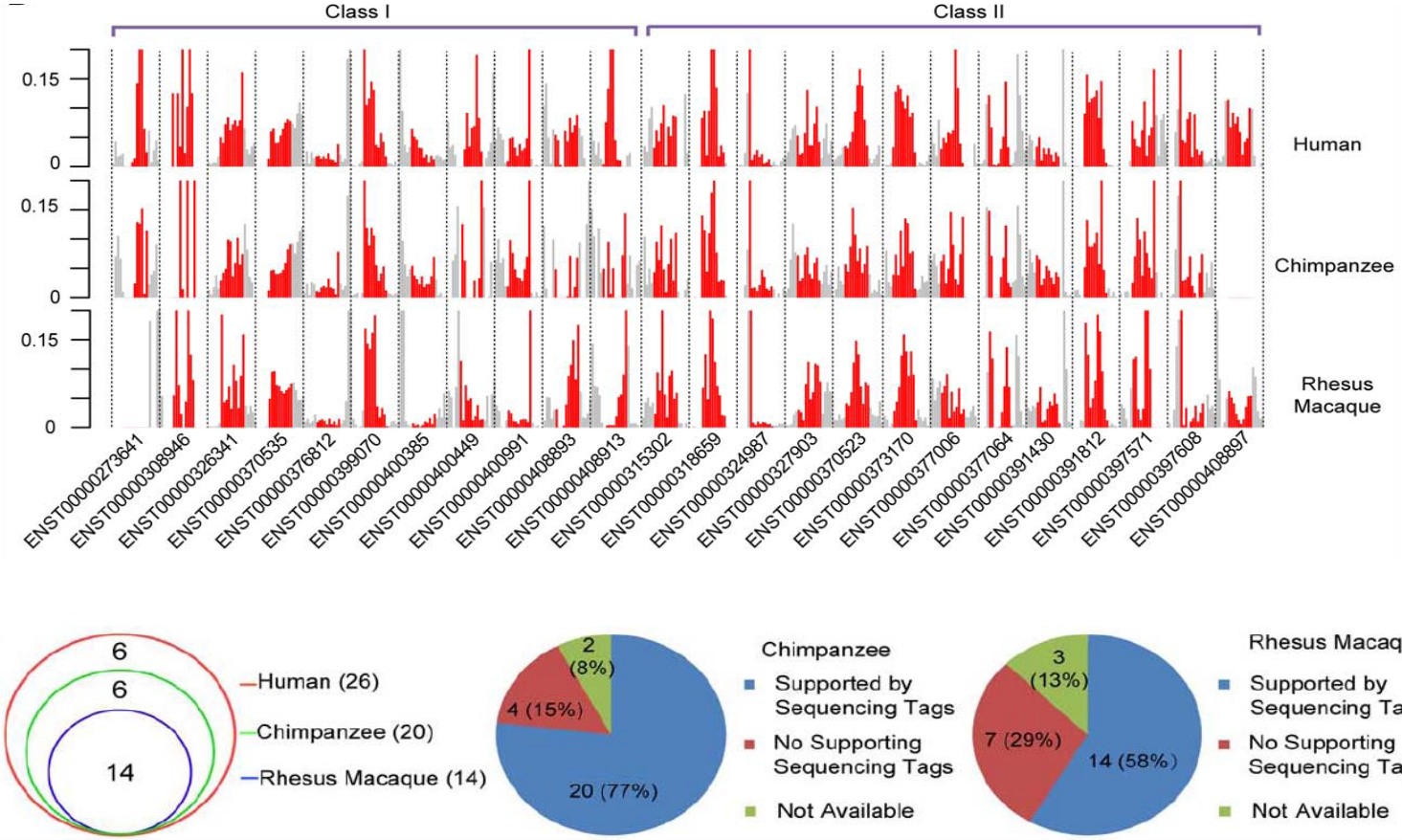
ORF first or regulated transcription first?

transcription leakage/noise until ORF

versus

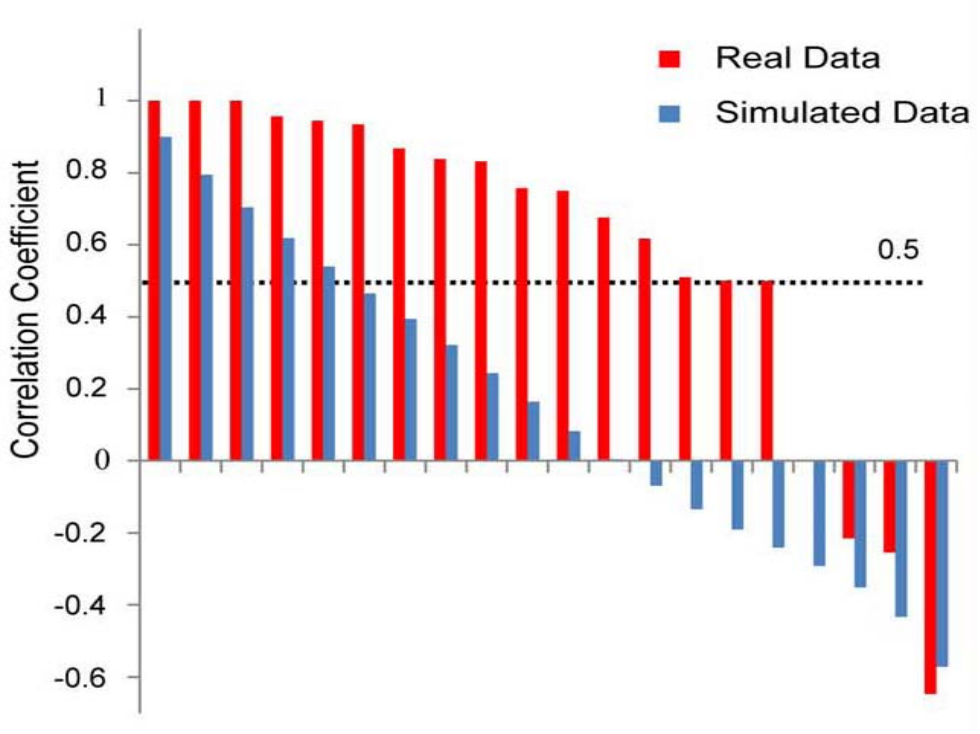
regulated transcriptional profile and structure of ncRNA

Non-coding genes tend to have similar gene structure with their protein-coding orthologs

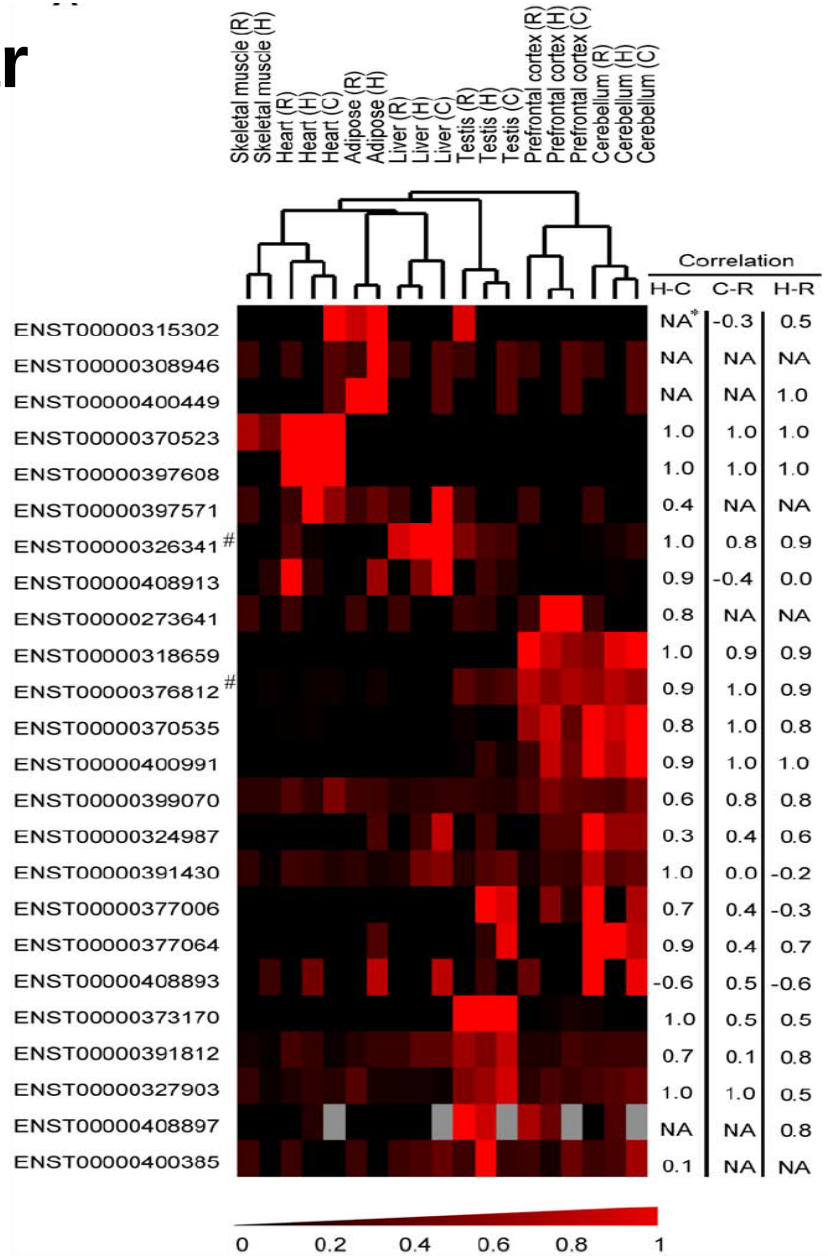


Xie *et al.*, *PLoS Genet.*, 2012

Non-coding genes tend to have similar tissue expression profile as their protein-coding orthologs

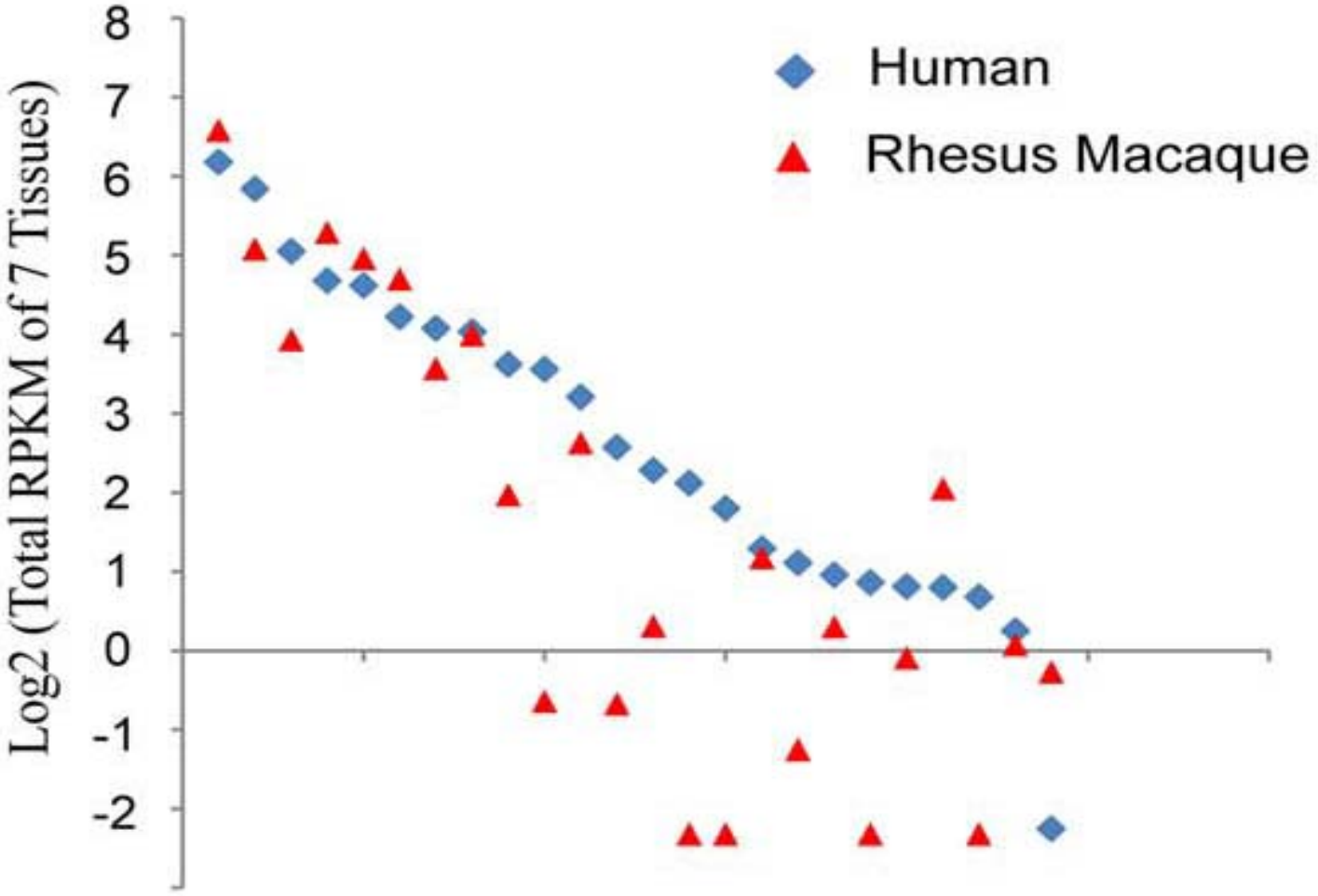


P-Value < 0.0001



Xie *et al.*, *PLoS Genet.*, 2012

Non-coding genes tend to have correlated, but lower, transcription level than their protein-coding orthologs

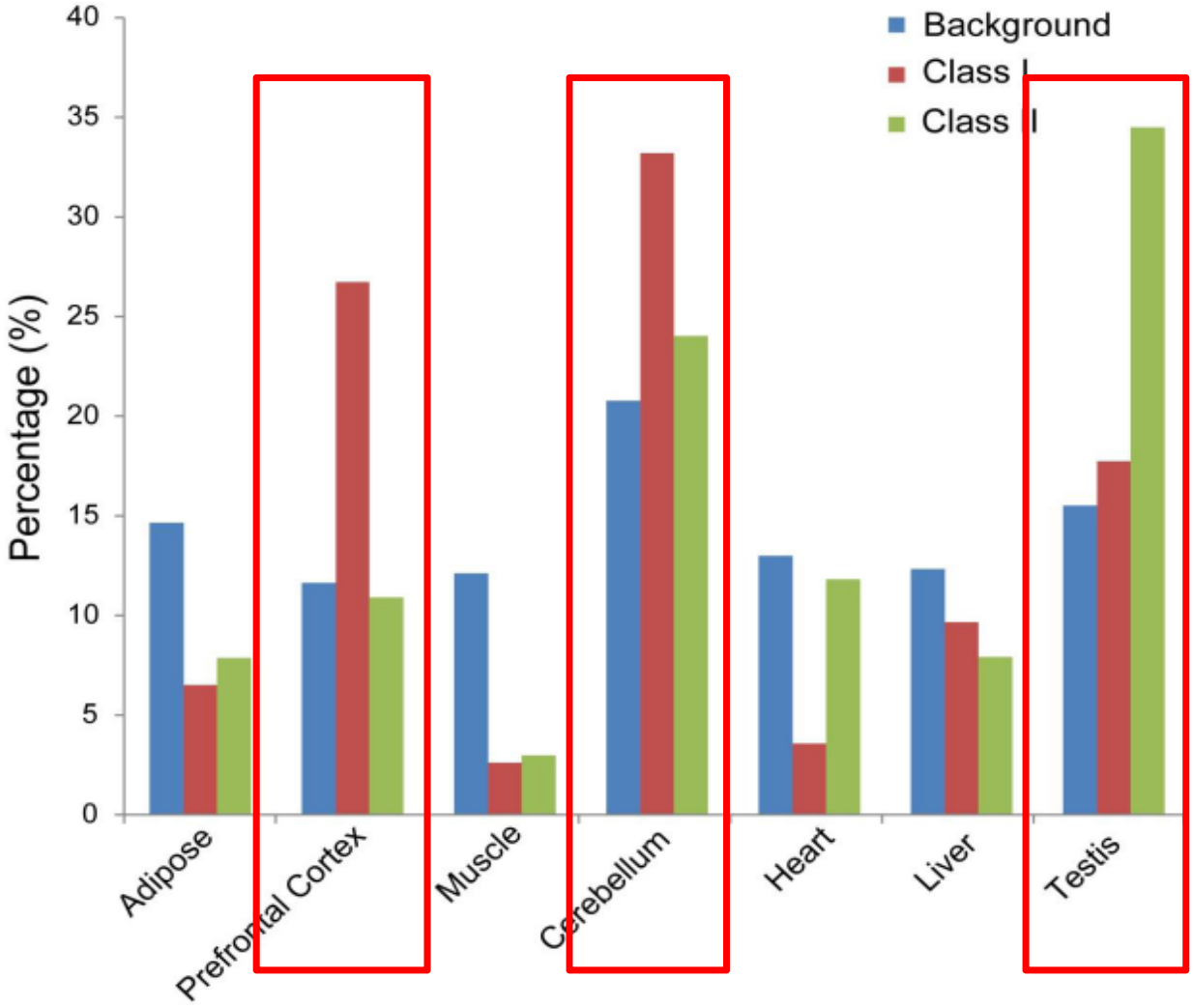


R-squared = 0.56

P-Value = 2.8×10^{-5}

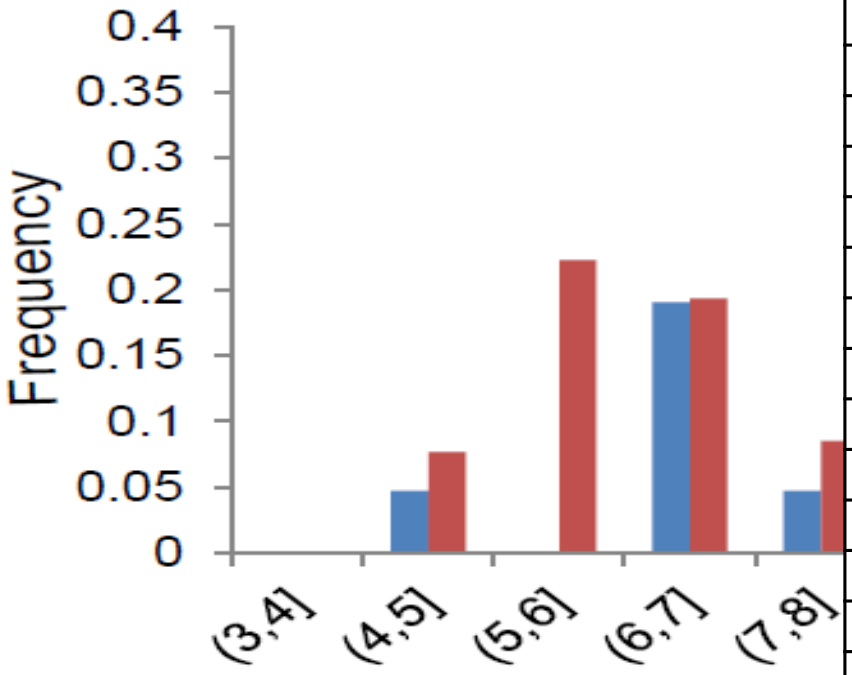
Xie *et al.*, *PLoS Genet.*, 2012

de novo genes have enriched expression in brain and testis



Xie *et al.*, *PLoS Genet.*, 2012

The pI values of



P-Value = 1.4×10^{-4}

GO Term	FDR q-value
RNA binding	5.50E-08
cytosolic ribosome	3.68E-07
macromolecular complex	1.63E-06
cytosolic large ribosomal subunit	4.61E-05
RNA splicing	6.71E-05
cytosolic part	7.73E-05
ribosomal subunit	4.54E-04
large ribosomal subunit	7.89E-04
intracellular organelle part	9.99E-04
organelle part	0.001136772
ribonucleoprotein complex	0.003187642
cellular biosynthetic process	0.007101674
MHC class II receptor activity	0.009220135
translation	0.010595406
mRNA processing	0.012153244
RNA processing	0.012167141
structural constituent of ribosome	0.017365179
mRNA metabolic process	0.020473341
macromolecule metabolic process	0.021017467
intracellular non-membrane-bound organelle	0.024935299
non-membrane-bound organelle	0.024935299
ribosome	0.036638186

Summary

Bioinformatic methods and analyses can play key roles in evolutionary biology.

- Identify interesting novel candidates at genome scale
- Discover genome-wide patterns
- Discover cross-species patterns

生物信息学：导论与方法
Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University