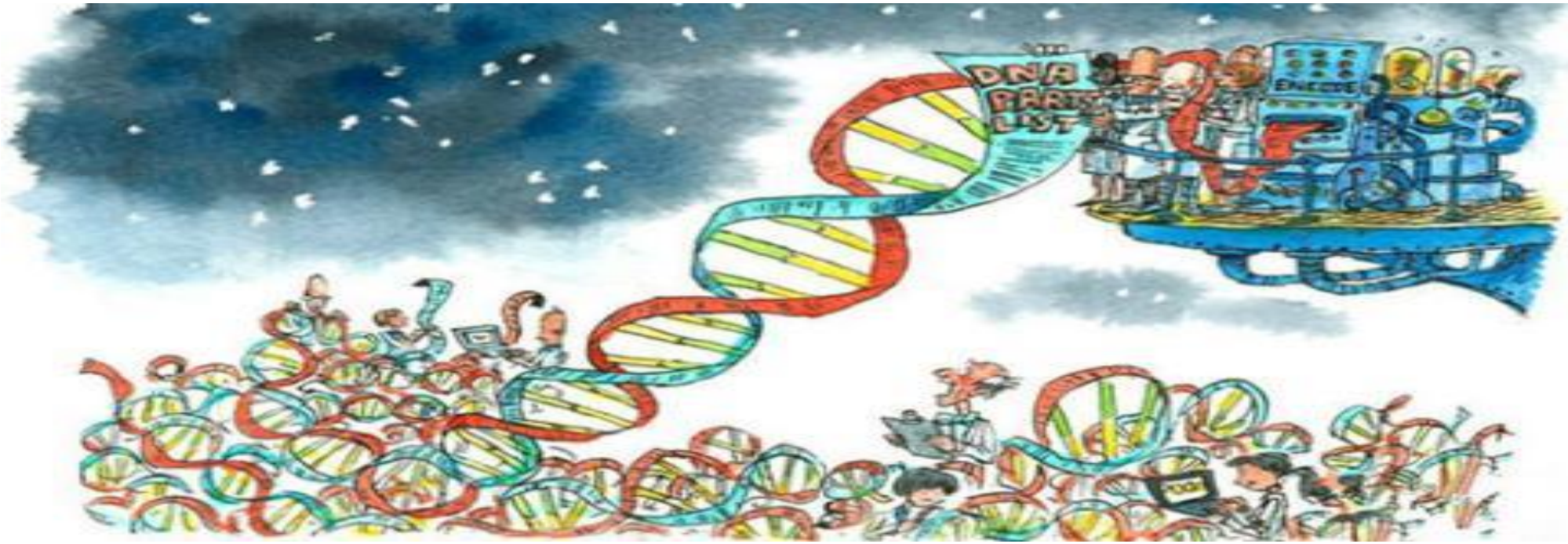
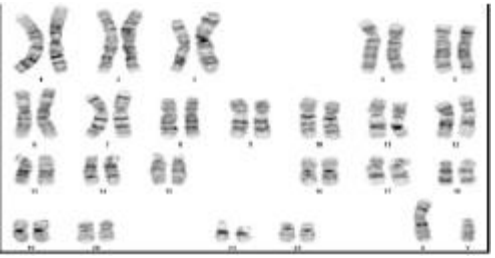
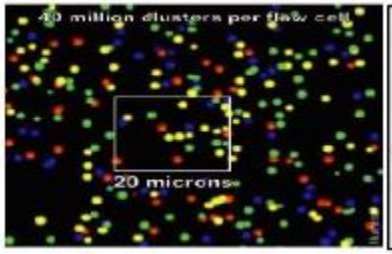
Bioinformatics: Introduction and Methods

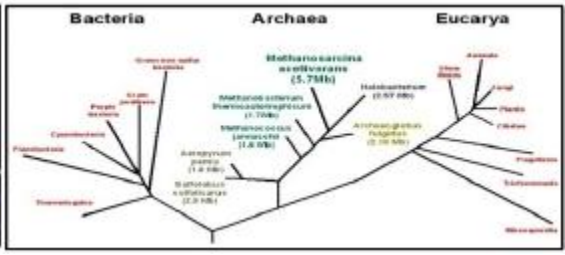
Le Zhang

Computer Science Department, Southwest University







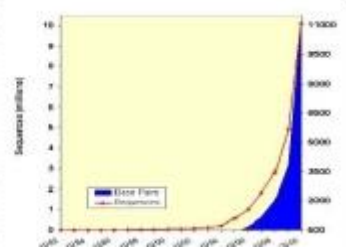


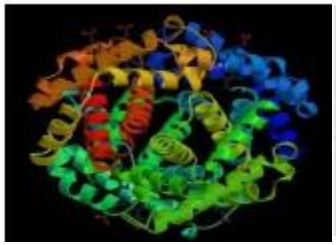
Case Study I: Origination of New Genes

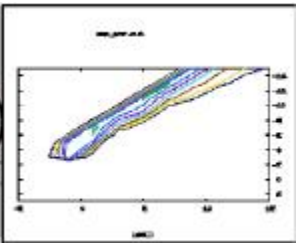
Le Zhang, Ph. D.

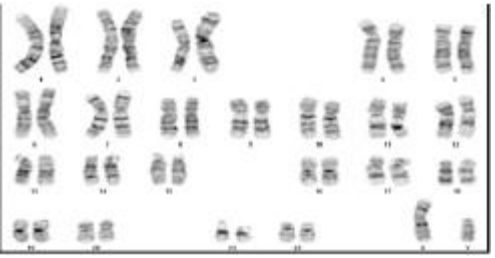
Computer Science Department

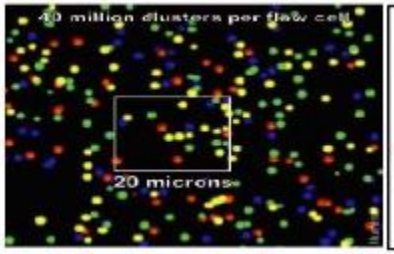
Southwest University

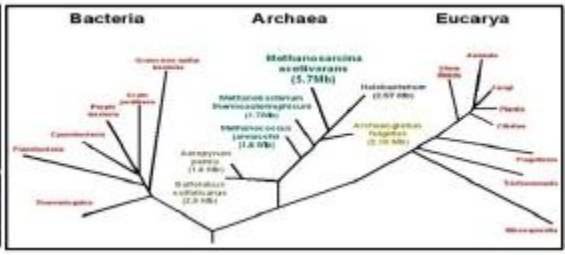










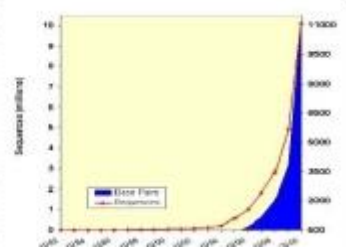


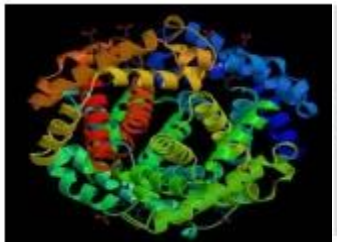
Unit 1:
Basic Concepts and Examples

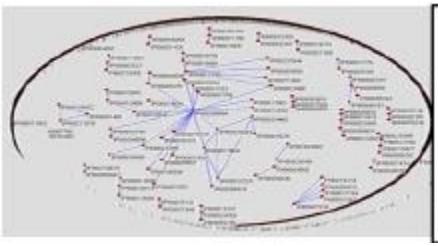
Le Zhang, Ph. D.

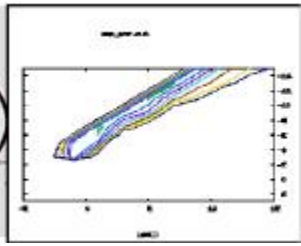
Computer Science Department

Southwest University









The role bioinformatics can play in biology and medicine

We emphasize a role of common bioinformatics can play in biology and medicine

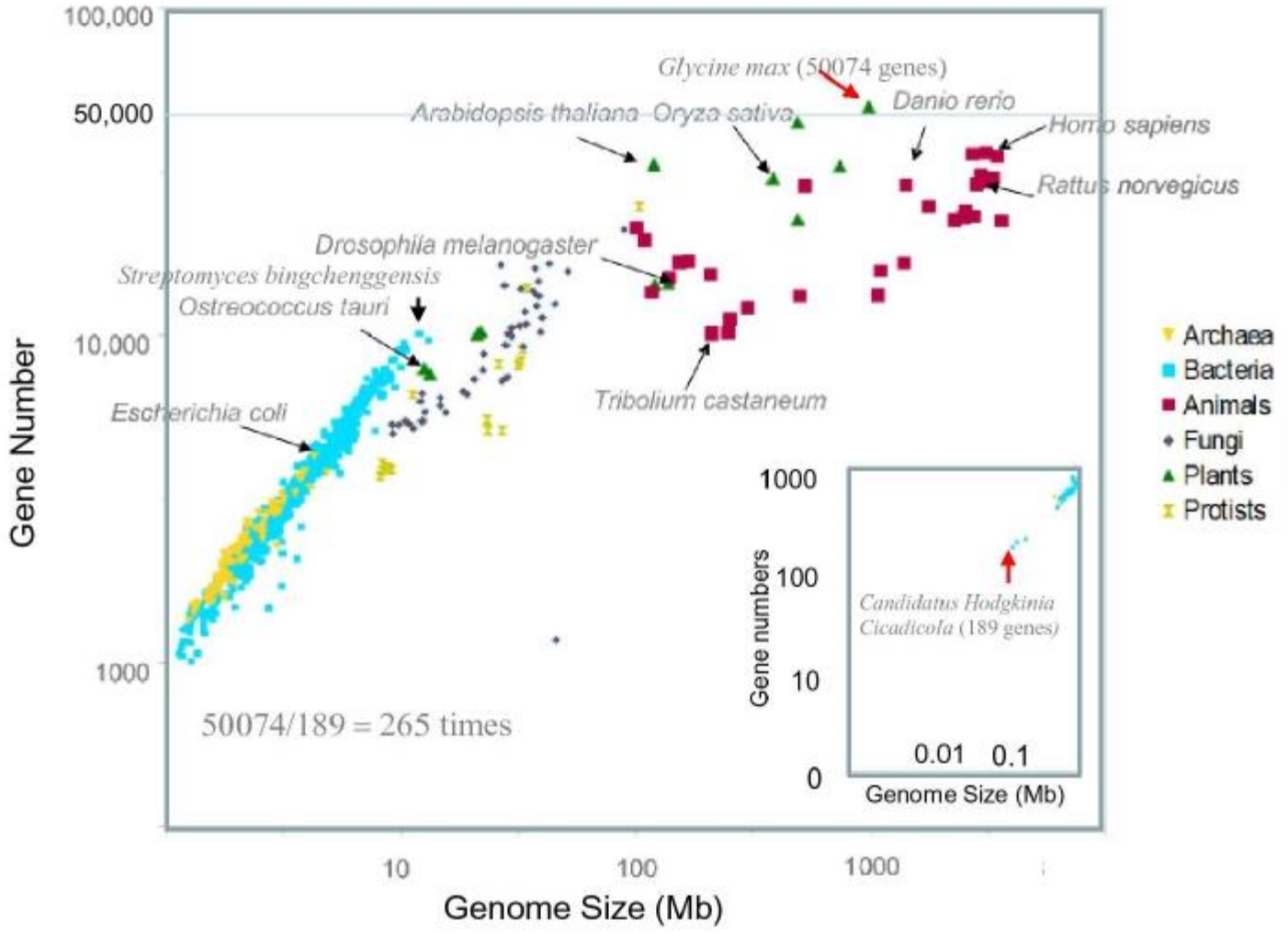
1. Present-day biological and medical studies are in a quick paradigm transition toward genomic analyses in gene identification and expression analysis that created astronomical-scale data.

2. The bioinformatics is a must for data analyses in various levels from preliminary data presentation to advanced interpretation for various scientific problems, with an unprecedented power to detect natural phenomena with the underlying mechanisms.

3. The biological rules and various correlations among the involved factors detected by the bioinformatic analysis from biological and medical studies are illuminating in the progress of understanding basic biological and medical problems.

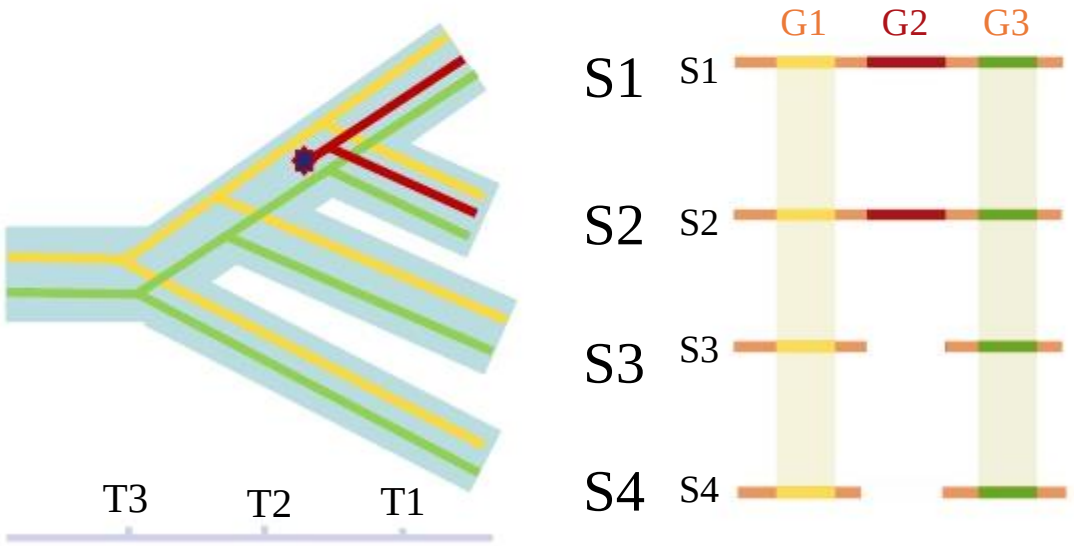
In this class, we are going to apply the bioinformatic analyses to a basic biological problem: the origin and evolution of new genes in a general concept and our understanding of evolution of humans and other mammals. These results are valuable for solving relevant biological and medical problems, exemplified by the case analyses.

New Gene Evolution Added to Genomic Diversity of Organisms



Organisms evolved in number of genes and size of genomes, suggesting a general process of birth and death of genes in evolution

New Gene: Definition for Synten-based Computational Identification



The new gene, G2, that originated in the most recent common ancestor of species S1 and S2 is located between two older genes G1 and G3. Because the divergence time of S1 and S2 is T1, the age of G2 is longer than T1 but shorter than T2, whereas G1 and G3 are older than T3. In general, the units of divergence time are often measured in units of million years ago (MYA).

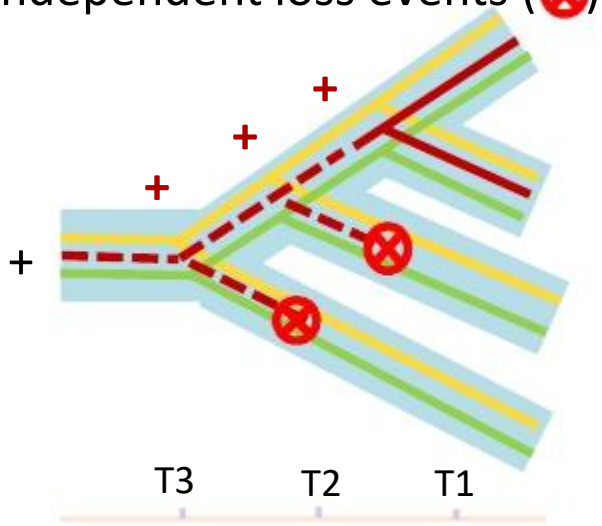
New Gene: Definition for Synten-based Computational Identification

Question: Why do we not define G2 in S3 and S4 as the consequence of gene loss that may have occurred in the ancestor before the divergence of S1 and S2, which may lead to the absence of G2 in S3 and S4?

Solution: the parsimony principle in evolutionary analysis.
The principle of accounting for observations by the hypothesis requiring the fewest or simplest assumptions that lack evidence; in evolution, the principle of invoking the minimal number of evolutionary changes to infer the more likely possibility.

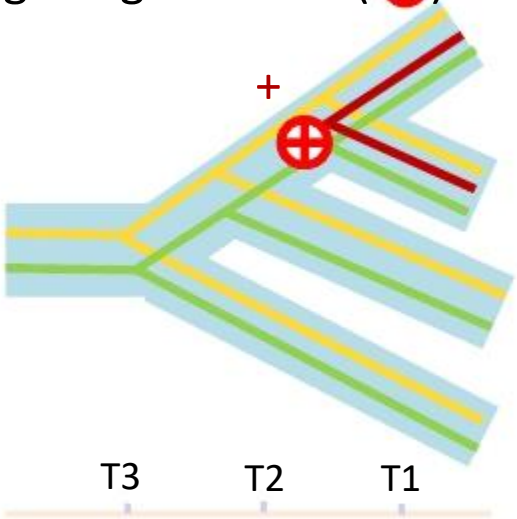
-- Revised from Dauglas J. Futuyma, 2009, Evolution.

2 independent loss events (⊗)



Gene Loss

1 gene gain event (⊕)

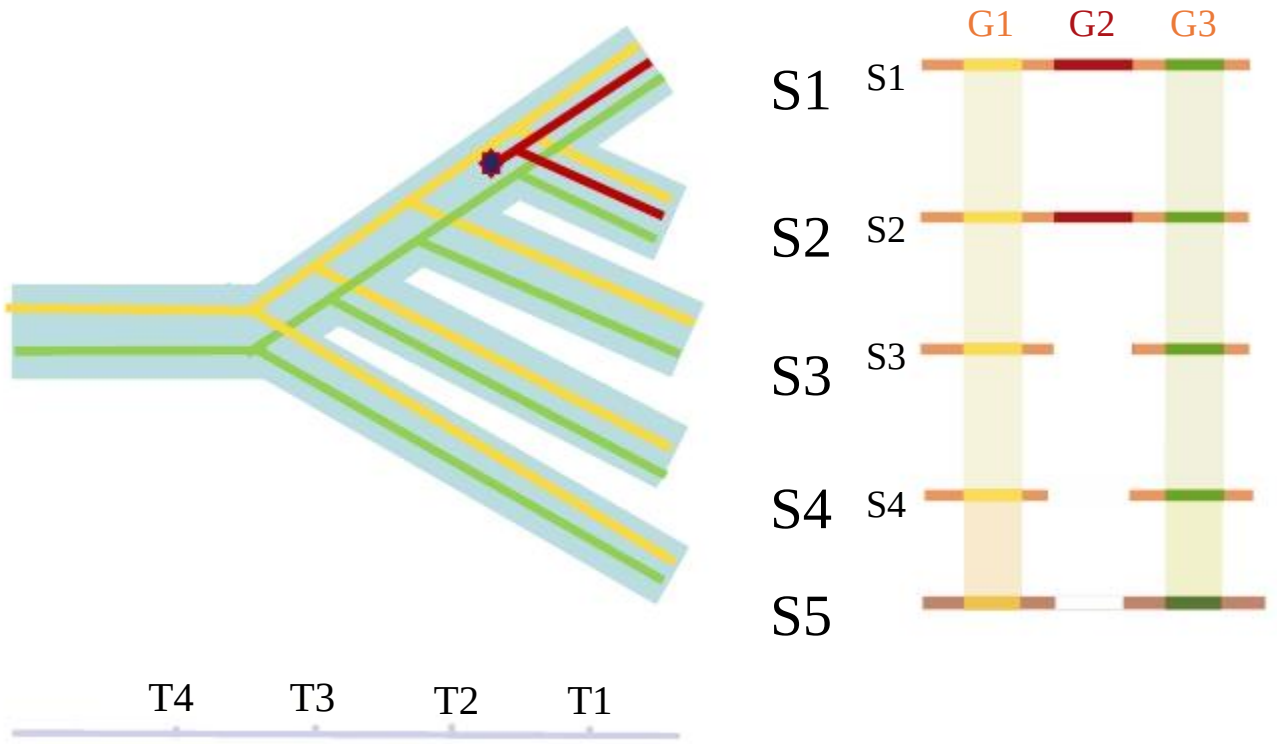


Gene Gain

New Gene: Definition for Synten-based Computational Identification

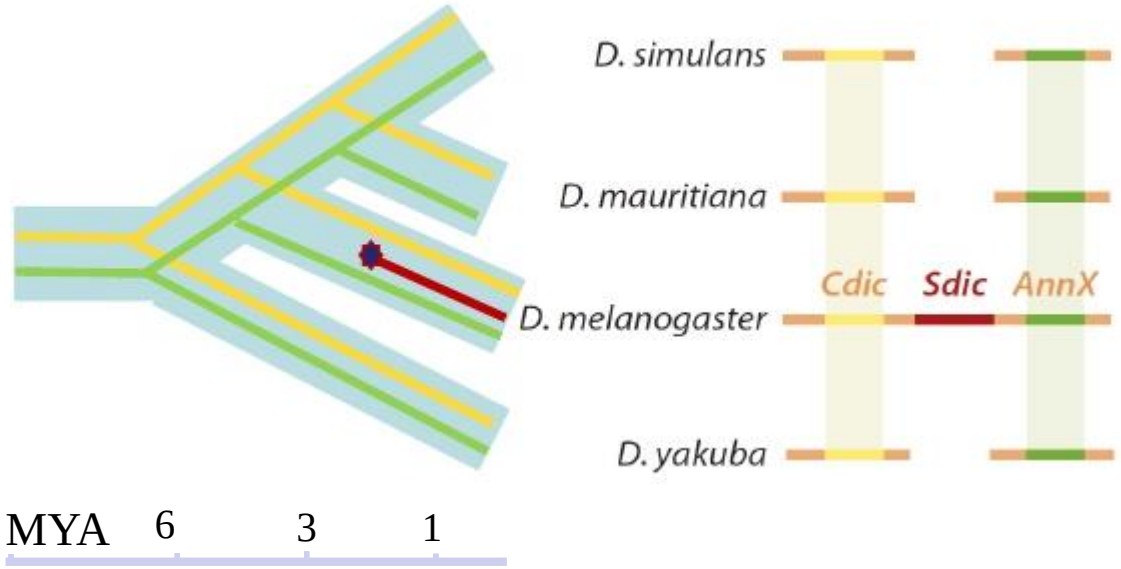
Exercise: Assuming the equal probability of gene gain and loss in each evolutionary change in the process, infer the ancestral state of presence or absence of the gene in T1, T2 and T3 in the two hypotheses of new gene gain or ancestral loss of an old gene. Then, choose the most parsimony hypothesis by calculating the total numbers of evolutionary changes required by the two hypotheses.

Question: in evolutionary analysis, S4 is called the outgroup species that can be used to help infer the ancestral state of G2 at time T2. Repeat the exercise when you add one more outgroup species that also has no G2 and find if you are more confident for our previous inference that G2 is a new gene that originated between T1 and T2, as is shown below:

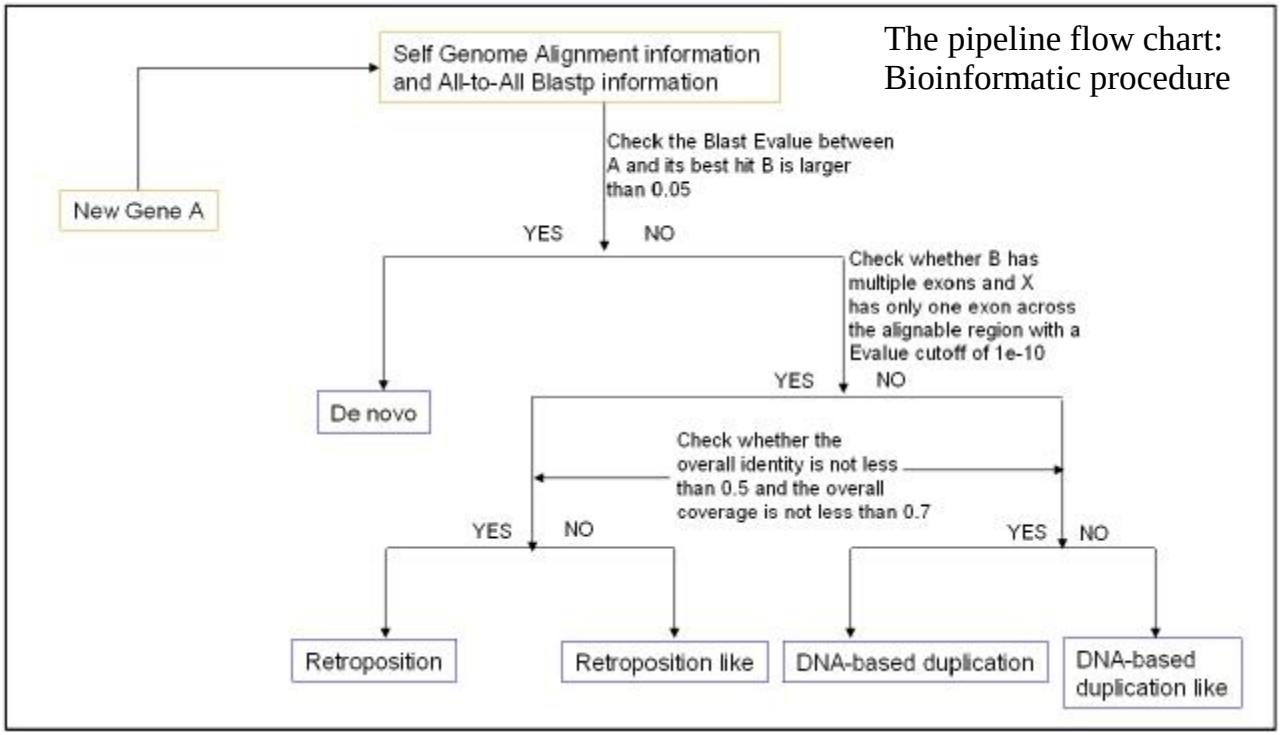


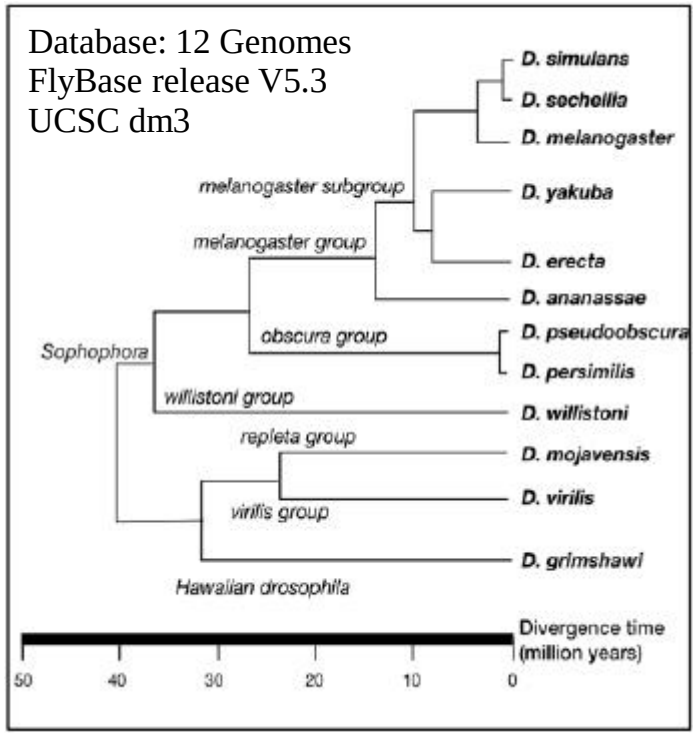
New Gene: An Example for the definition

Sdic is a new gene in D. melanogaster that codes for a sperm-specific axonemal dynein subunit, which is immediately flanked by two parental genes, Cdic and Annx.

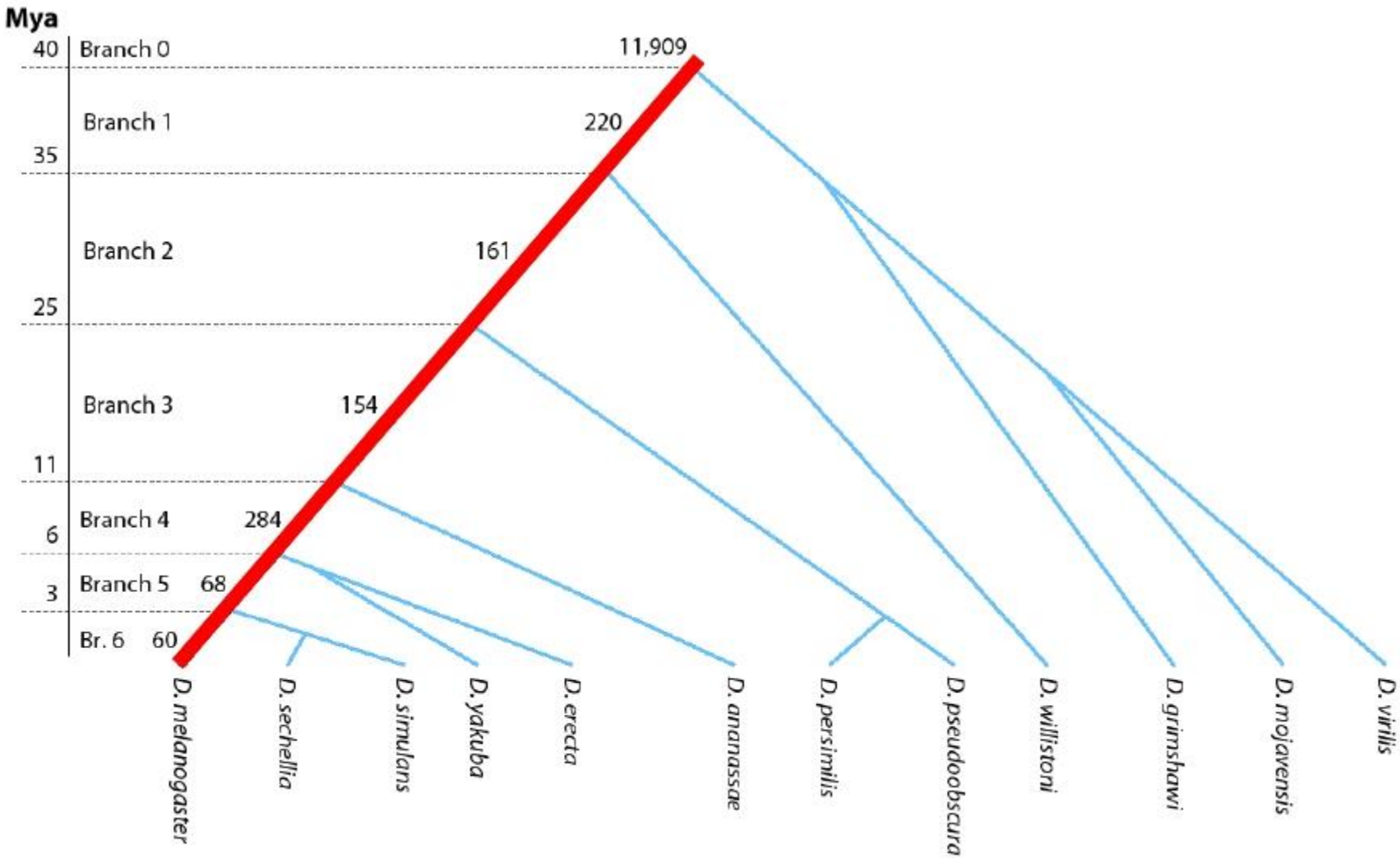


Computational identification of new genes from 12 Drosophila genome sequences

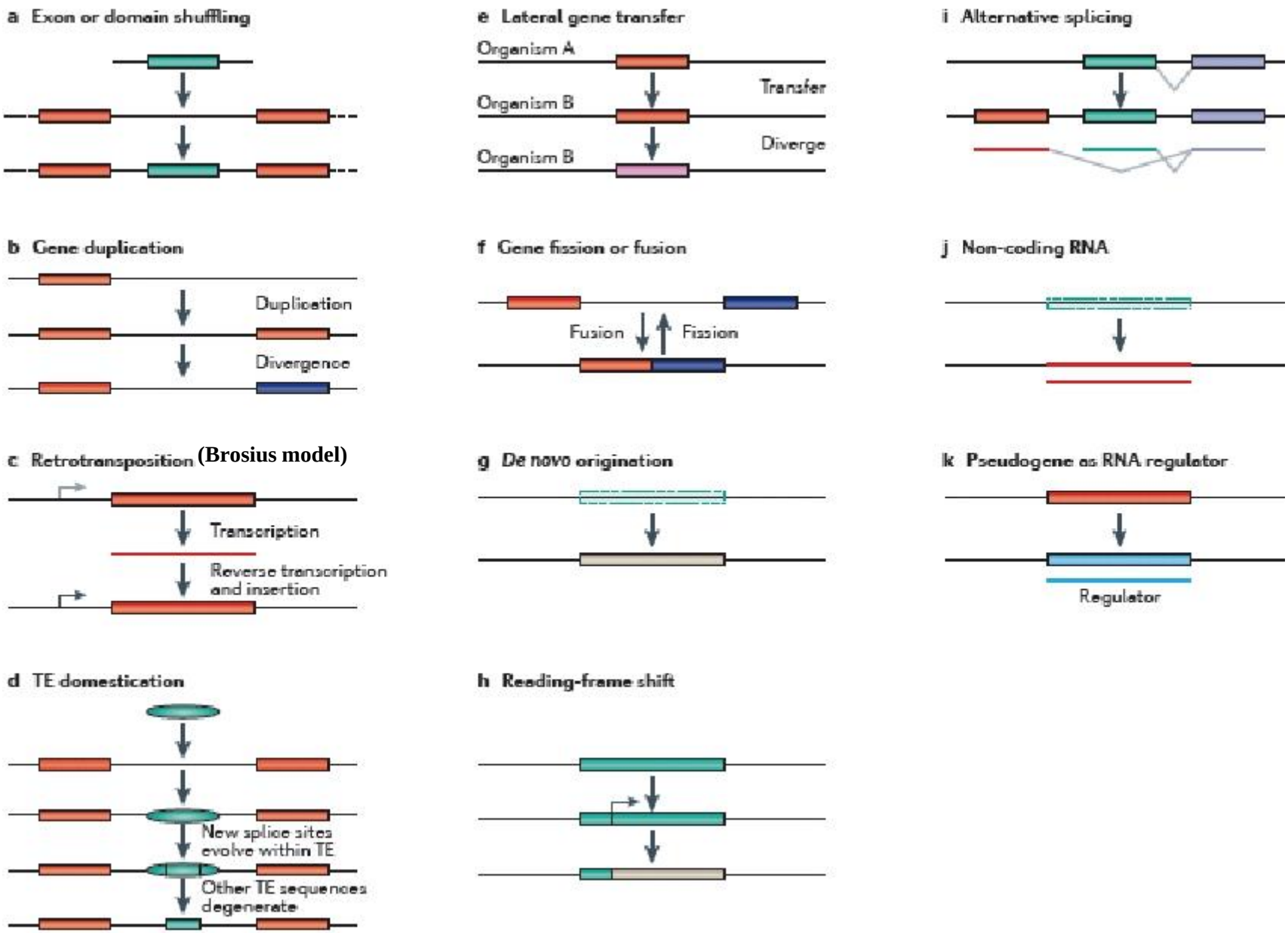




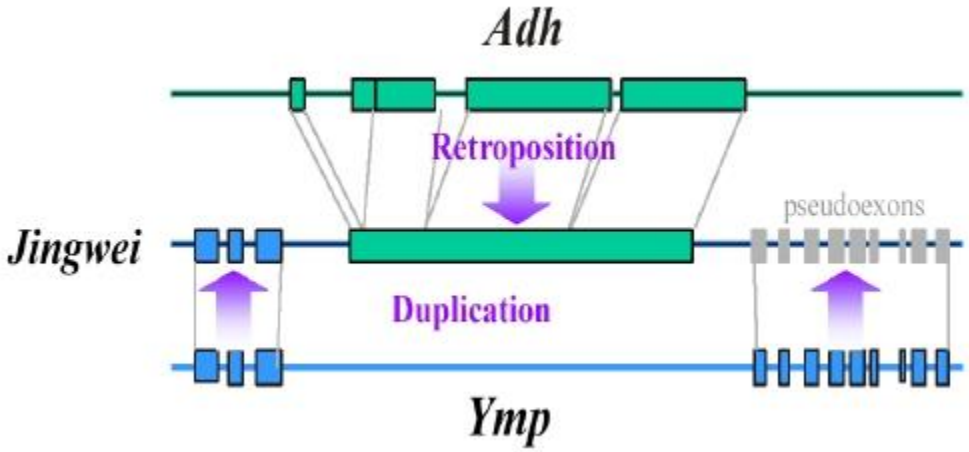
New genes distribution mapped in the evolutionary tree of Drosophila

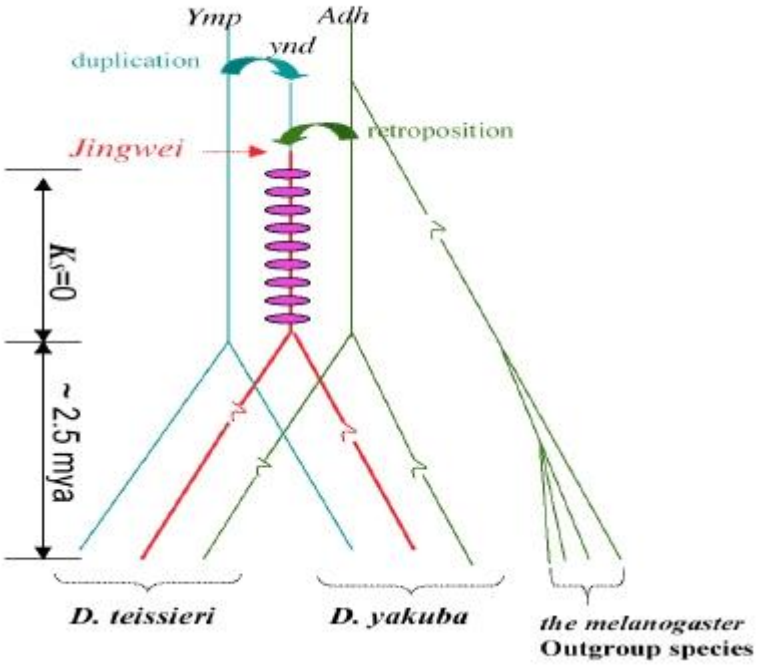


Mechanisms of New Gene Origination

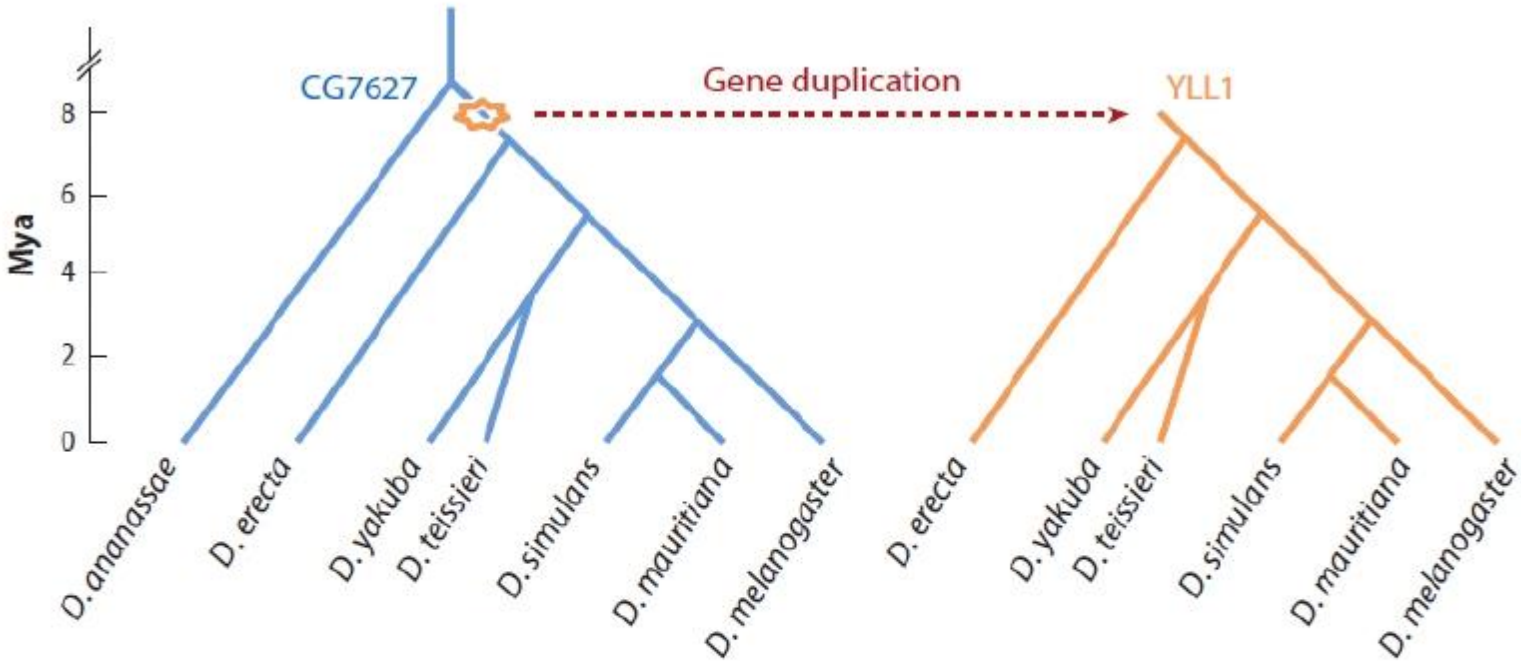


Example from the first observed new gene: the Jingwei in Drosophila that reveal several mechanisms can be involved to generate a new gene





Biological importance of new genes



LINES	METHOD/MUTATION	PHENOTYPES
11343	P-element insertion	Lethal, pupal stage
SH2-0995	EMS/G717S	Lethal, pupal stage
SH2-1101	EMS/T765I	Lethal, pupal stage
SH2-0504	EMS/synonymous	Viable
V39539	RNAi/constitutive Gal4	Lethal, pupal stage
V39540	RNAi/constitutive Gal4	Lethal, pupal stage

Biological importance of new genes: Examples for Published New Genes

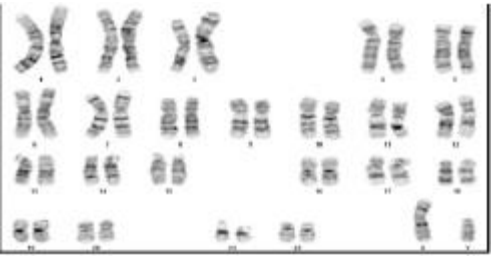
New genes	Age (m illion years)	O rigination m echanism	Expression	Phenotype	Function	Refs
Drosophilaspp.						
Sdicfam ily	0–3	DNA duplication	Testis	Sperm com petition	Cytoskeleton	69–71
sphinx	0–3	Retrotransposition	Neuronaland reproductivetissue	M alecourtship	ncRNA	77,111
jingwei	0–3	Retrotransposition	Testis	Recruitm entpherom one and horm one	Alcoholm etabolism	7,17
p24-2	0–3	DNA duplication	M ultiplestages and tissues	Developm ent,m ale reproduction	Proteintrafficking	46,47
Xcbp1	3–6	Retrotransposition	Neuronaltissue	Foragingbehaviour	Chaperone	76
Zeusm als	3–6	Retrotransposition	M alereproductive tissues	M alereproduction	Generegulation	74,122
FGF4	~0.01	Retrotransposition	Distalhum erus	Hum erusdevelopm ent	FGFsignalling	66
SRGAP2C	1.0–3.4	PartialDNA duplication	Brain	Predicted toaffect cortexdevelopm entin am ousem odel	Unknown	109, 110
CDC14C	7–12	Retrotransposition	Brainand testis	Unknown	Cellcycle	21
CYPA	<10	Retrotransposition	Unknown	Viralinfection	HIV-1resistance	29
PO LDI	2.5–3.5	Denovoorigination	Testis	Knockoutreduced testisweightand sperm m otility	Unknown	44
TBC1D3	<35	Segm ental duplication	Prostate	Insulinm odulation	IGFsignalling	128
Plants						
CYP98A8	<28	Retrotransposition	Vasculartissue, pistilis,roottip,etc.	Pollendevelopm ent	Phenolicsynthesis	19
CYP84A4	<8	Geneduplication	Stem and seedling	Unknown	Arabiodopyrone biosynthesis	20
CYP98A9	<28	Retrotransposition	Vasculartissue, pistilis,roottip,etc.	Pollendevelopm ent	Phenolicsynthesis	19

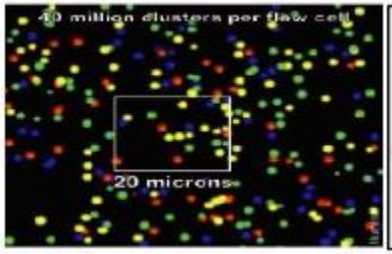
SUMMARY

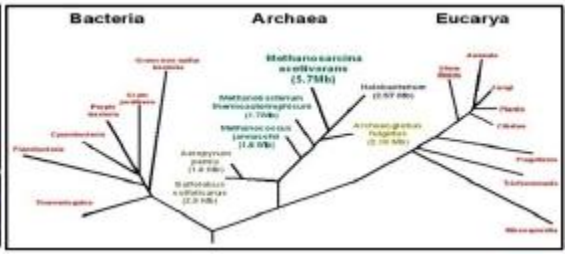
1. A new gene is a gene that originated recently in a genome and can be identified by syntenic alignment of genomic sequences from a group of closely species.

2. A number of molecular mechanisms can generate new genes and more than one mechanism can be involved in making one new gene.

3. New genes can be biologically important as old or ancient genes. In fruitflies, essential functions can evolve rapidly any time in evolution.





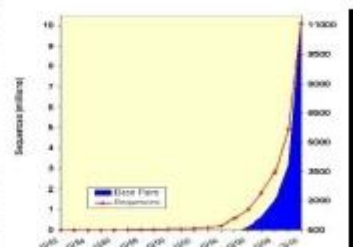


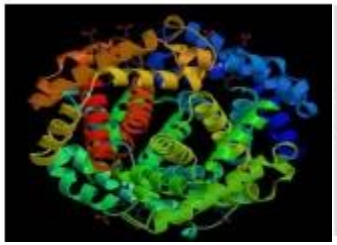
Unit 2:
A Driver for Human Brain Evolution

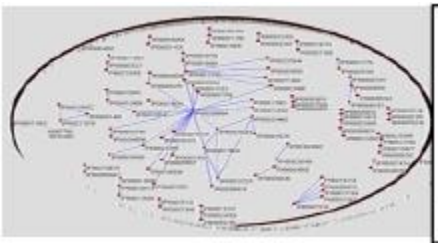
Le Zhang, Ph. D.

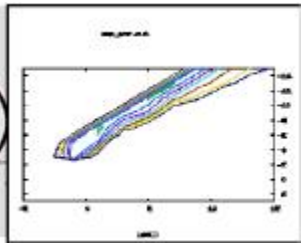
Computer Science Department

Southwest University

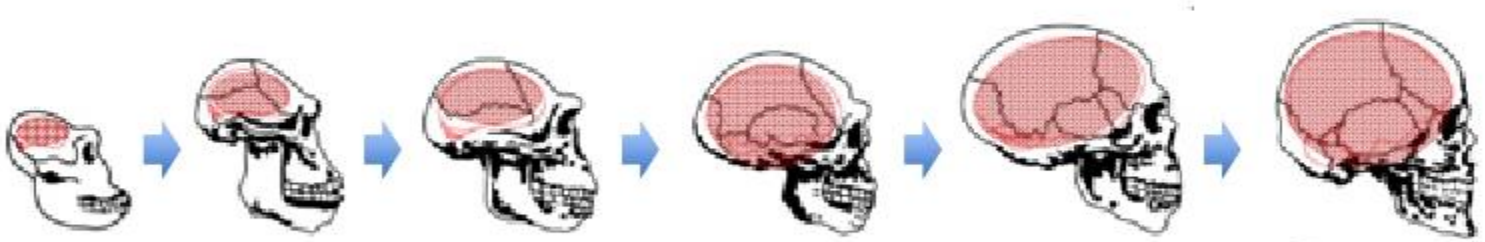








The evolution and wonders of our brains

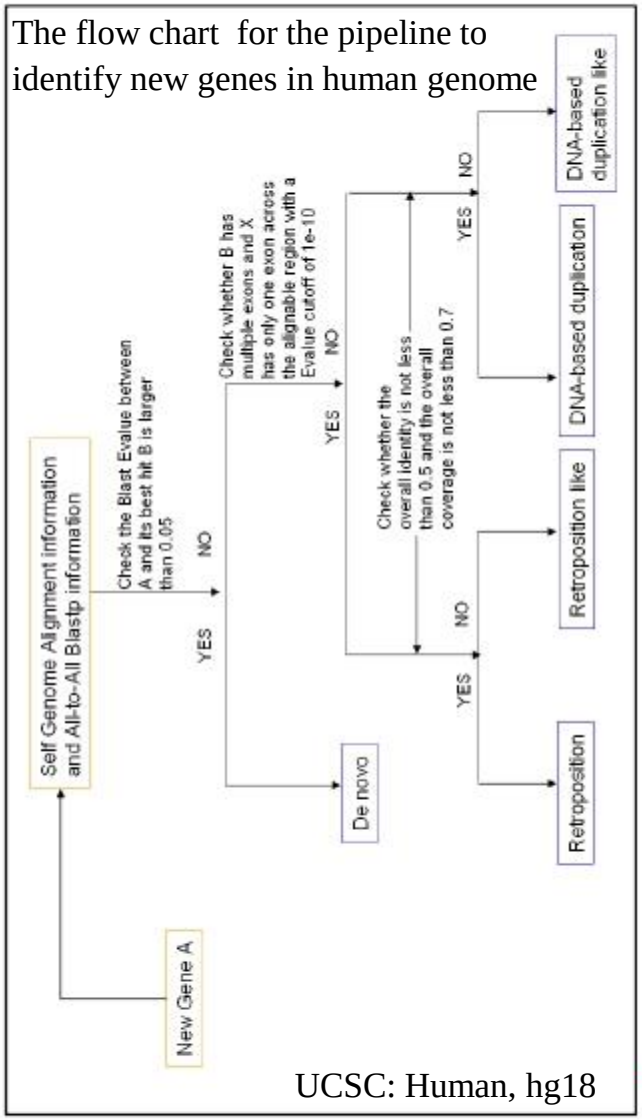


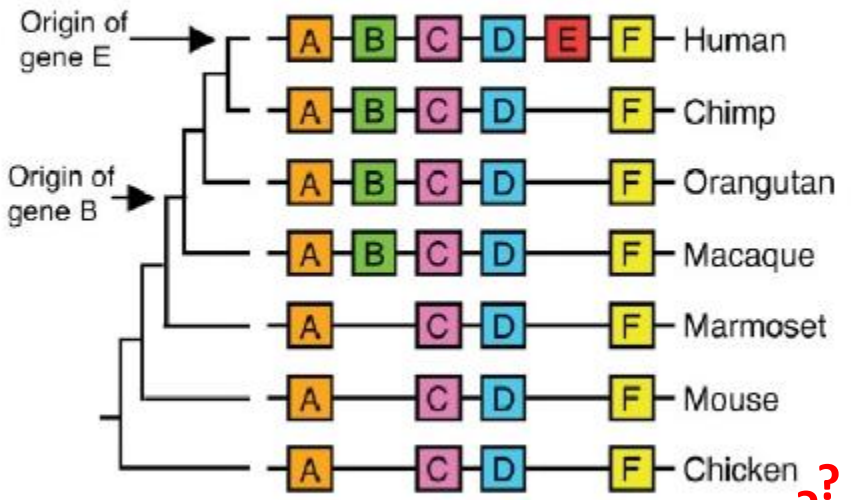
What genetic changes occurred in our ancestors drove evolution?

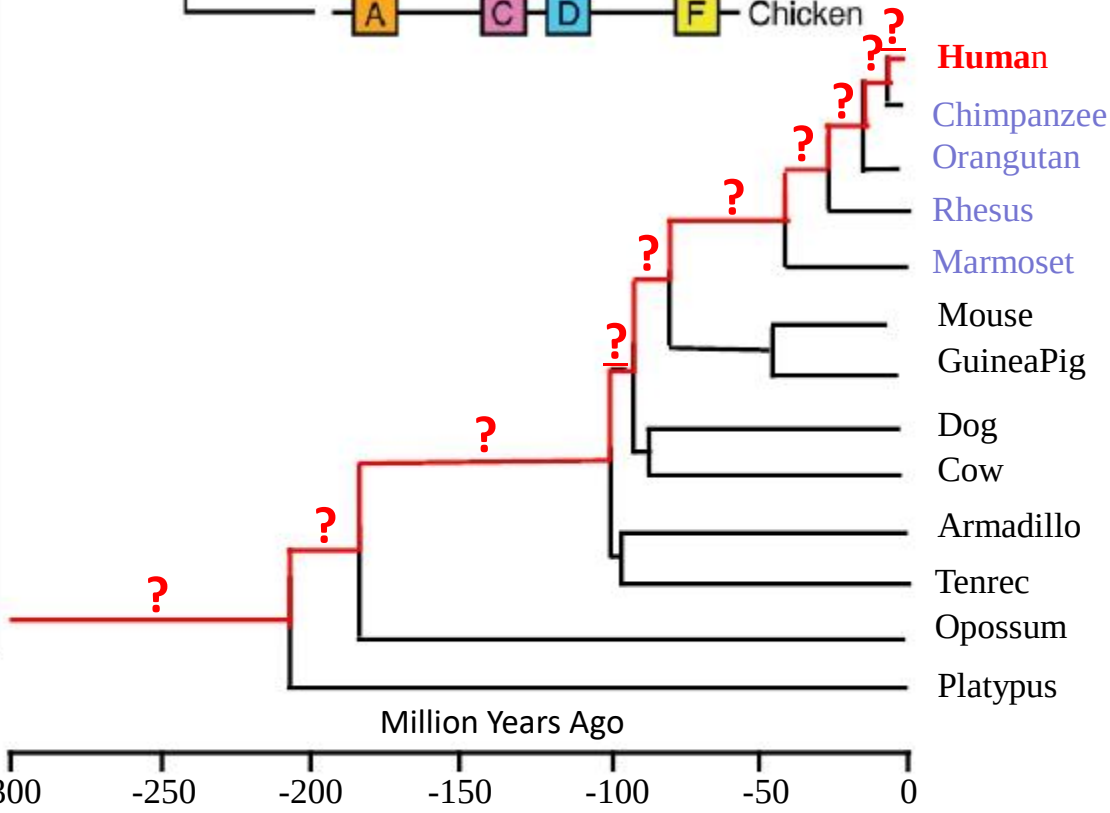
-- The role of new genes in brain evolution

Computational identification of new genes in vertebrate genomes

The flow chart for the pipeline to identify new genes in human genome







Distribution of identified new genes mapped to the lineage toward humans

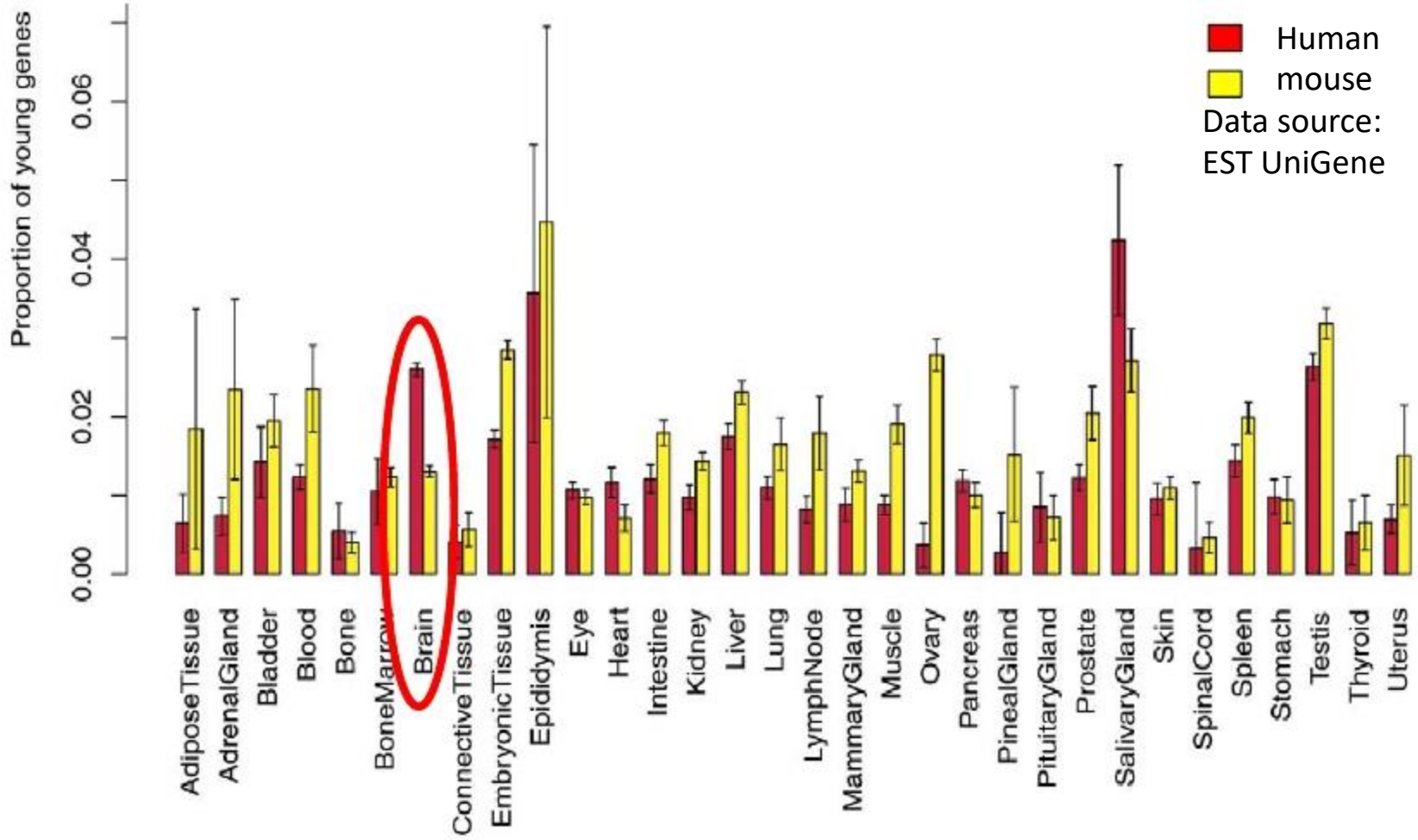
Average Rate = 20 genes / million years

Primate Rate = 30 genes / million years

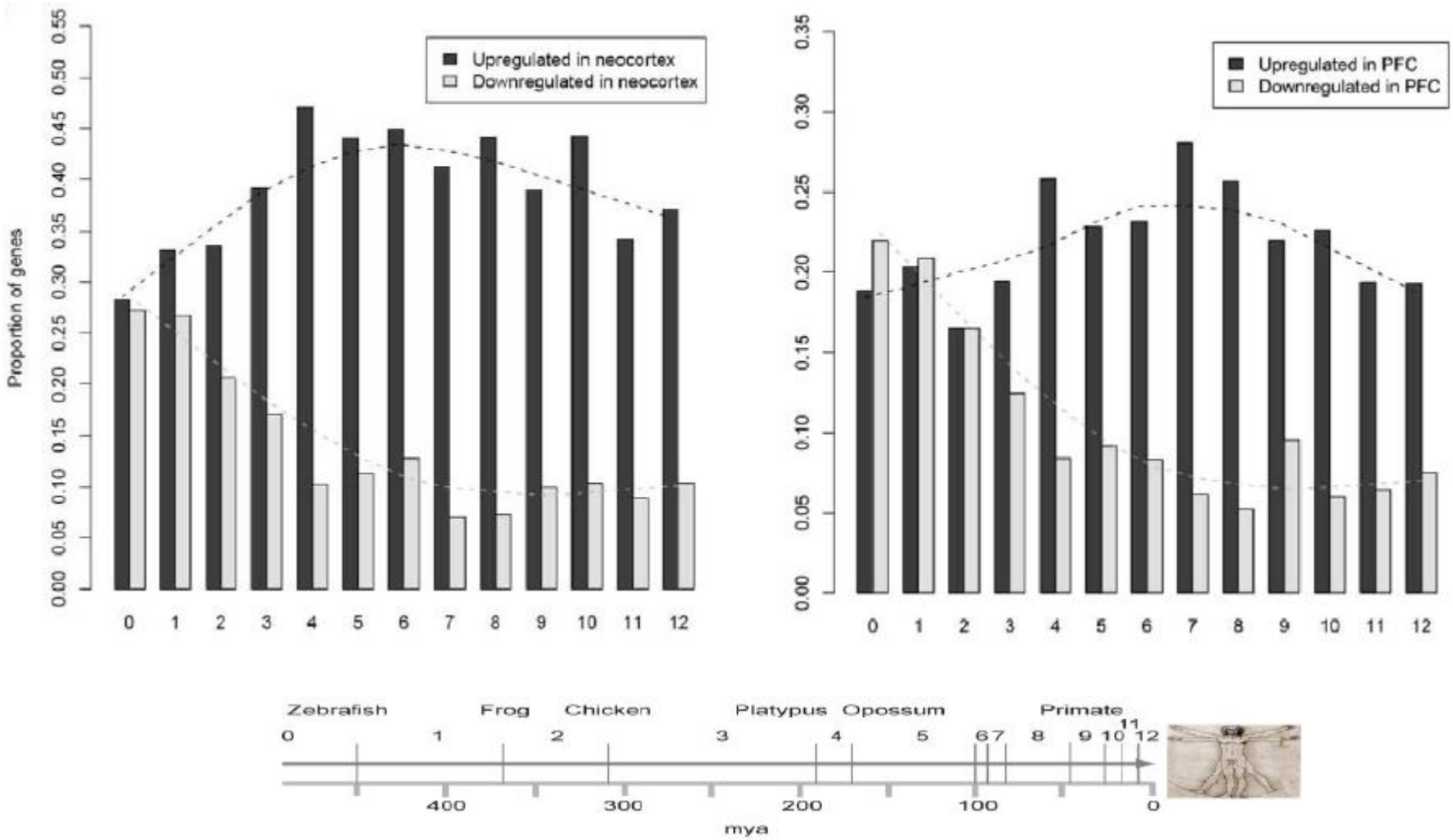
Human Rate = 80 genes / million years

The lineage toward the human: 5500 mammalian new genes; 1800 primate new genes

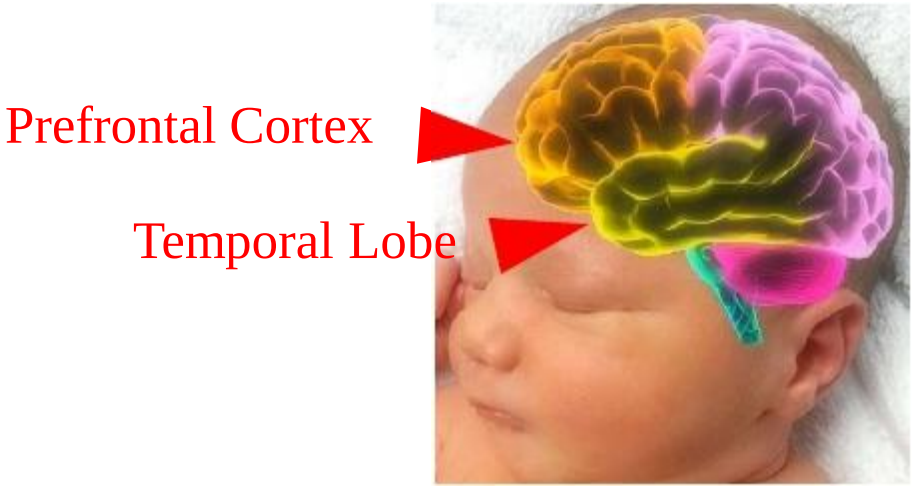
Looking for new genes that are specifically expressed in human brain



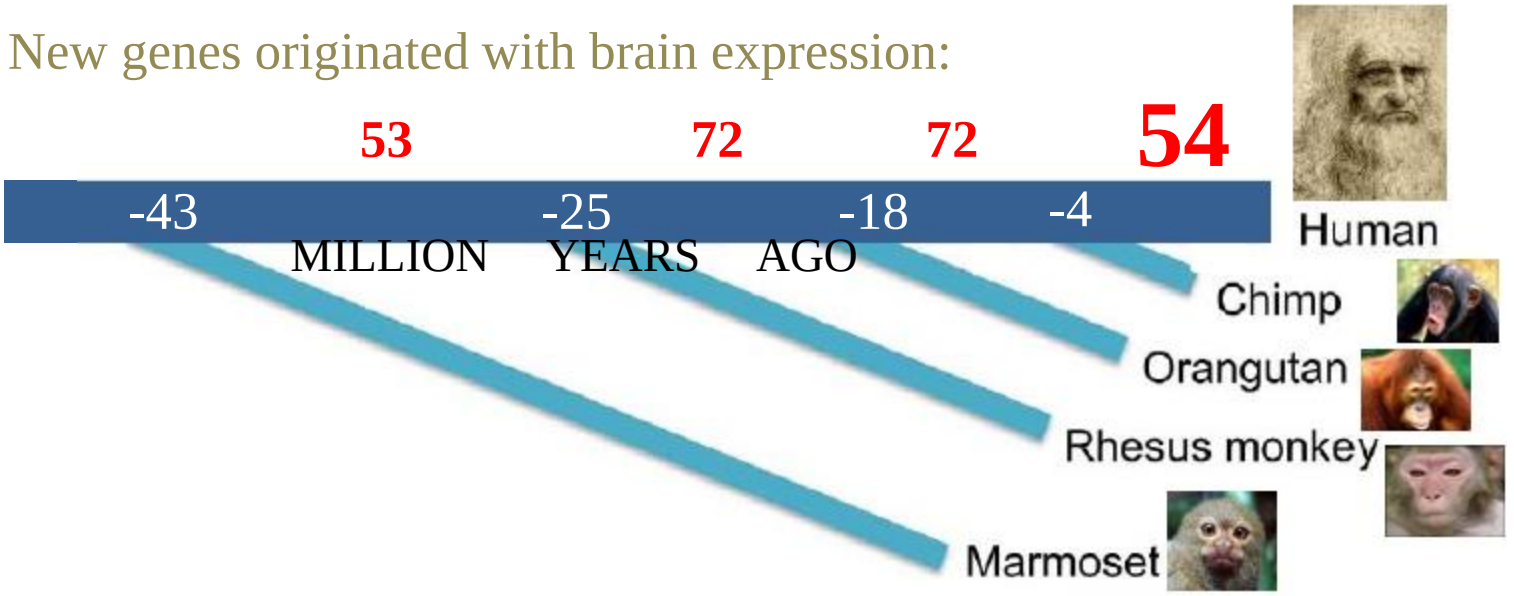
Expression of new genes that originated in various evolutionary stages



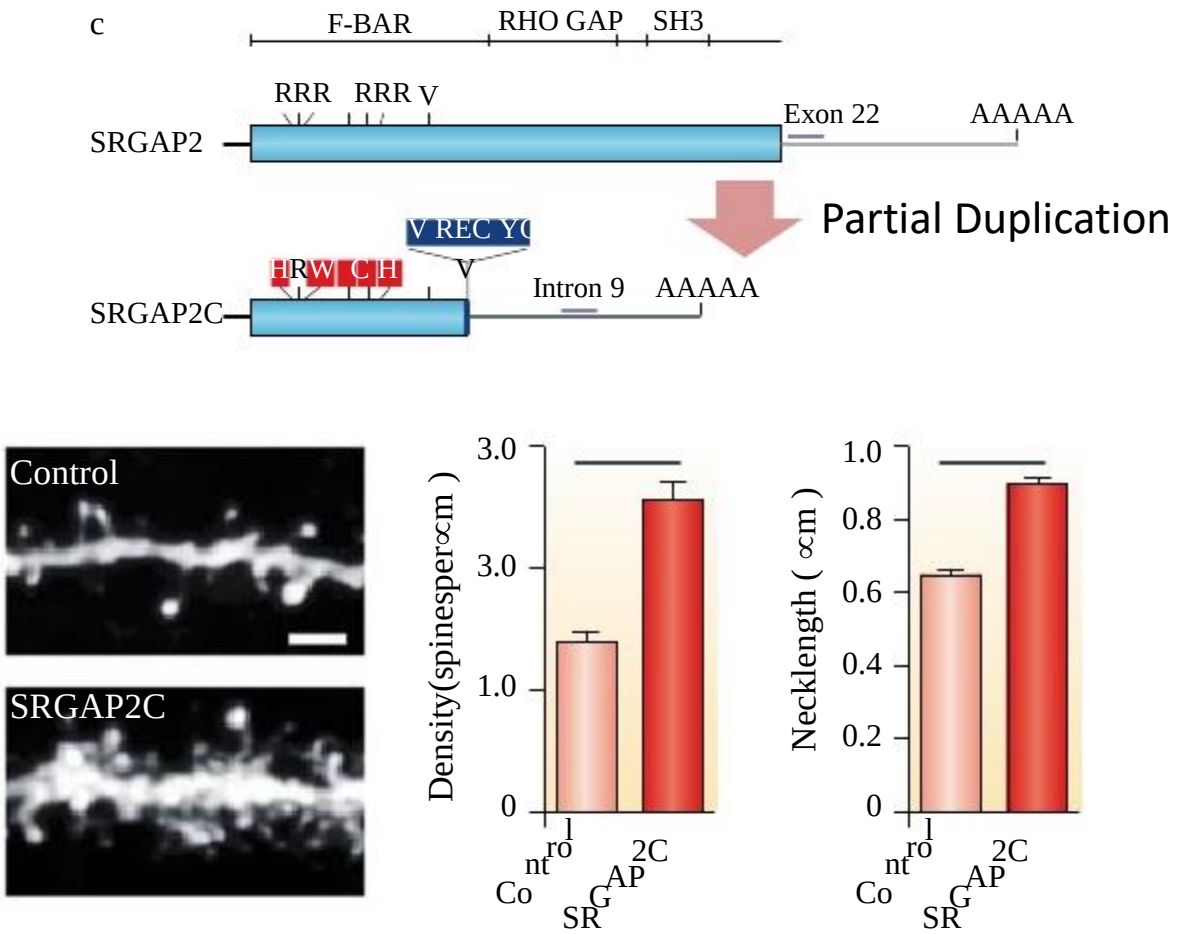
New genes are expressed in early developing brain



New genes originated with brain expression:



Possible functions of the human-specific genes: the example of SRGAP2C



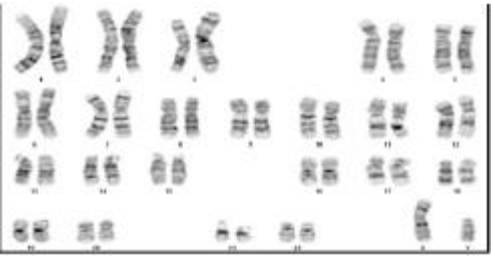
The transgenic expression of SRGAP2C in cultured mouse cortical neurons detected a higher proportion of the nerve cells growing denser dendritic spines with longer necks to connect with neighbouring neurons better, which may enhance the ‘computing power’ of brains.

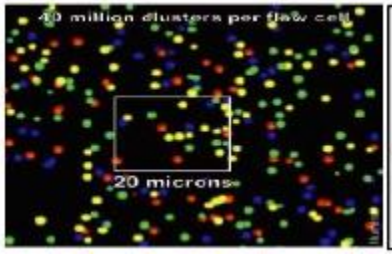
SUMMARY

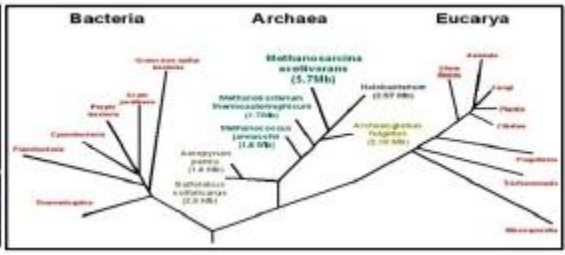
1. Evolution of brain was accompanied with origin of new genes.

2. New genes are upregulated in the neocortex, in particular the prefrontal cortex regions, throughout evolution of vertebrates.

3. Many new genes, in particular human-specific, new genes expressed in the prefrontal cortex and temporal lobe, the brain structure involved for cognitive functions.





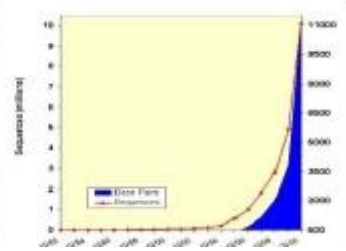


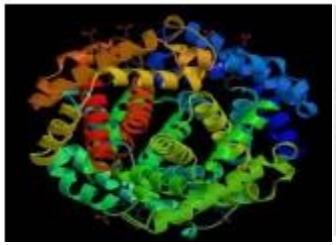
Unit 3:
A Human-Specific de novo Gene Associated with Addiction

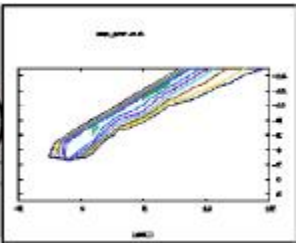
Le Zhang, Ph. D.

Computer Science Department

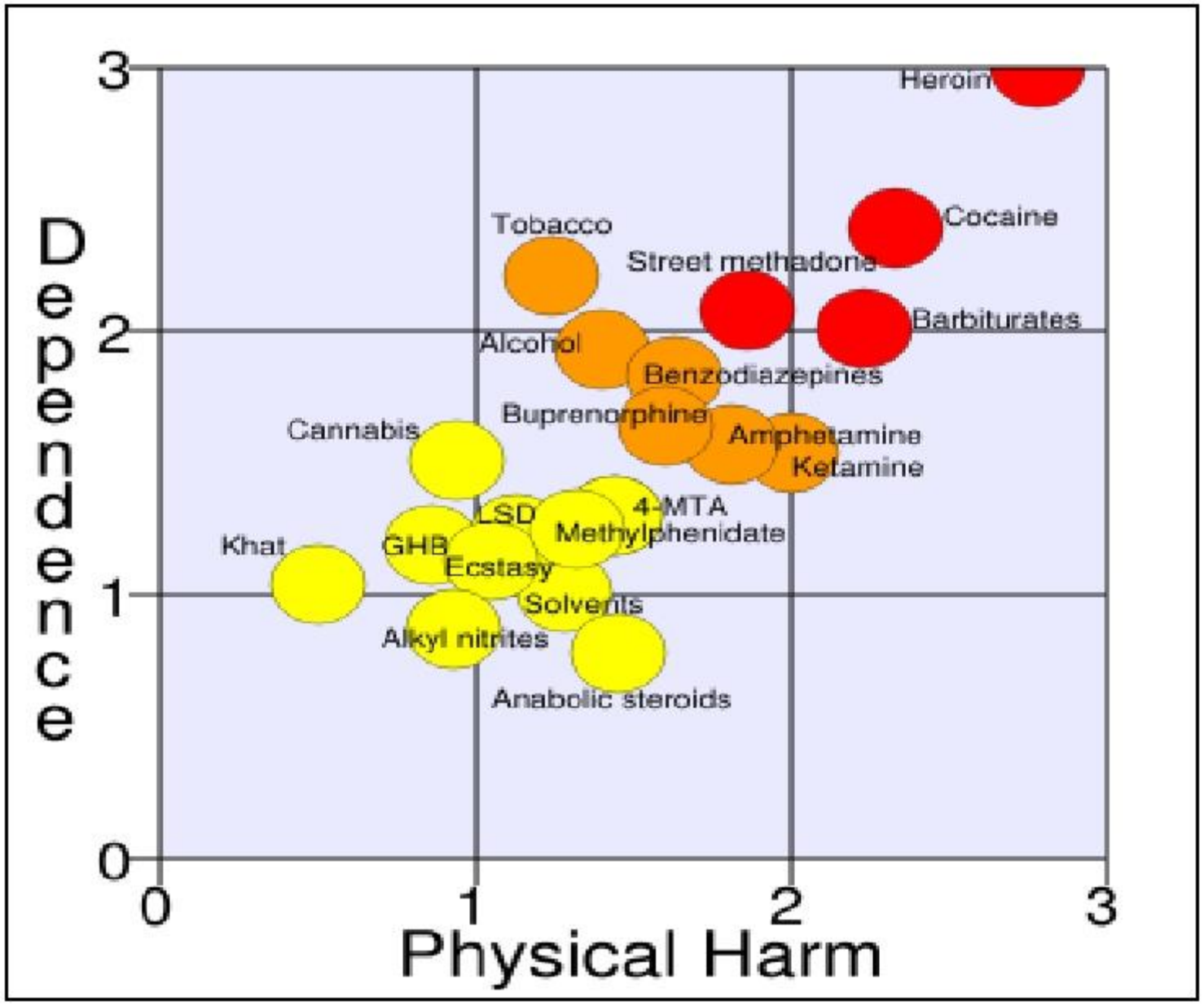
Southwest University





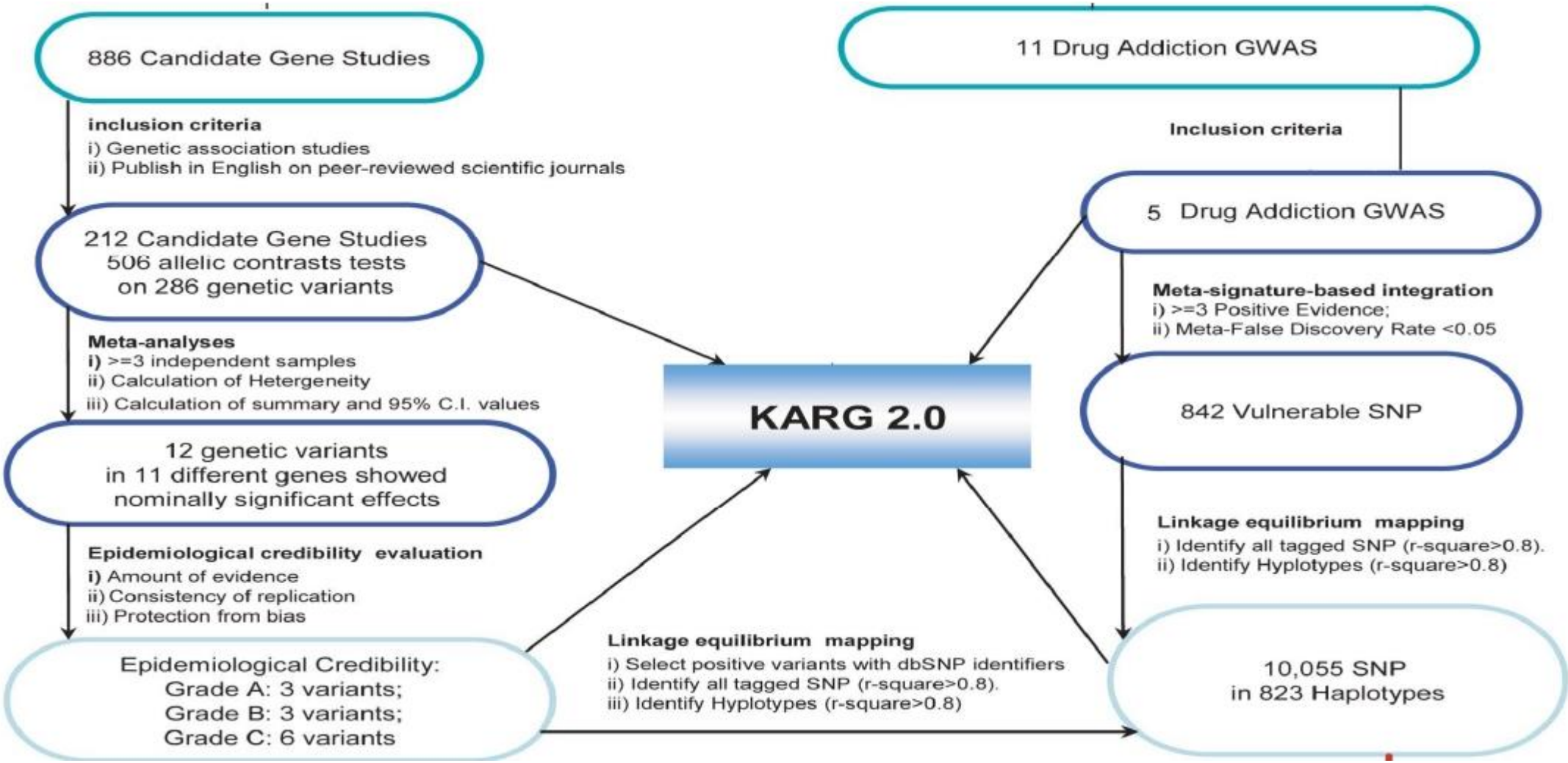


Addiction



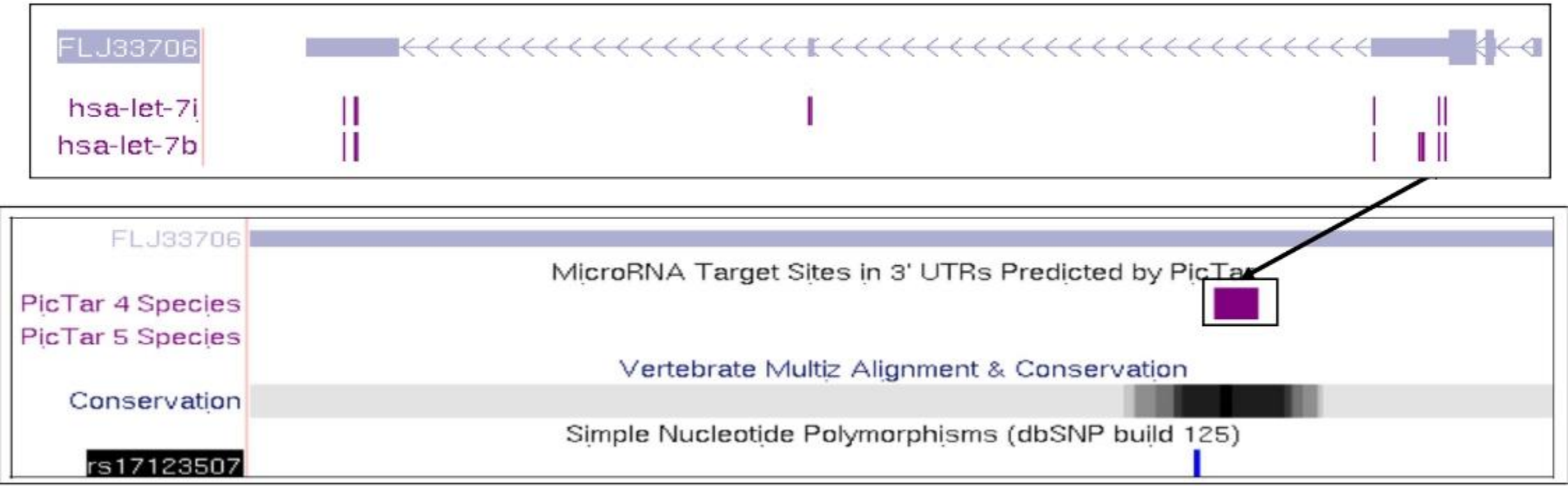
Nutt, *NEJM*, ’07

Collaboration with NIH/NIDA to analyze addiction GWAS data

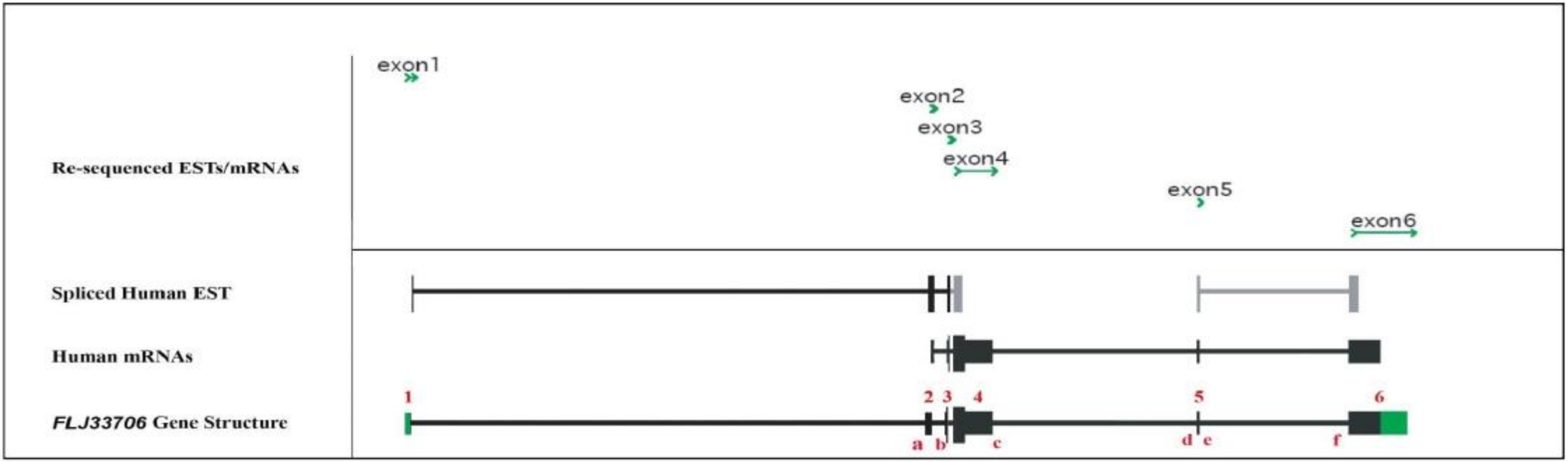


An SNP on the 3'UTR of *FLJ33706* is statistically significant in two GWAS of nicotine addiction and implicated in two linkage analyses.

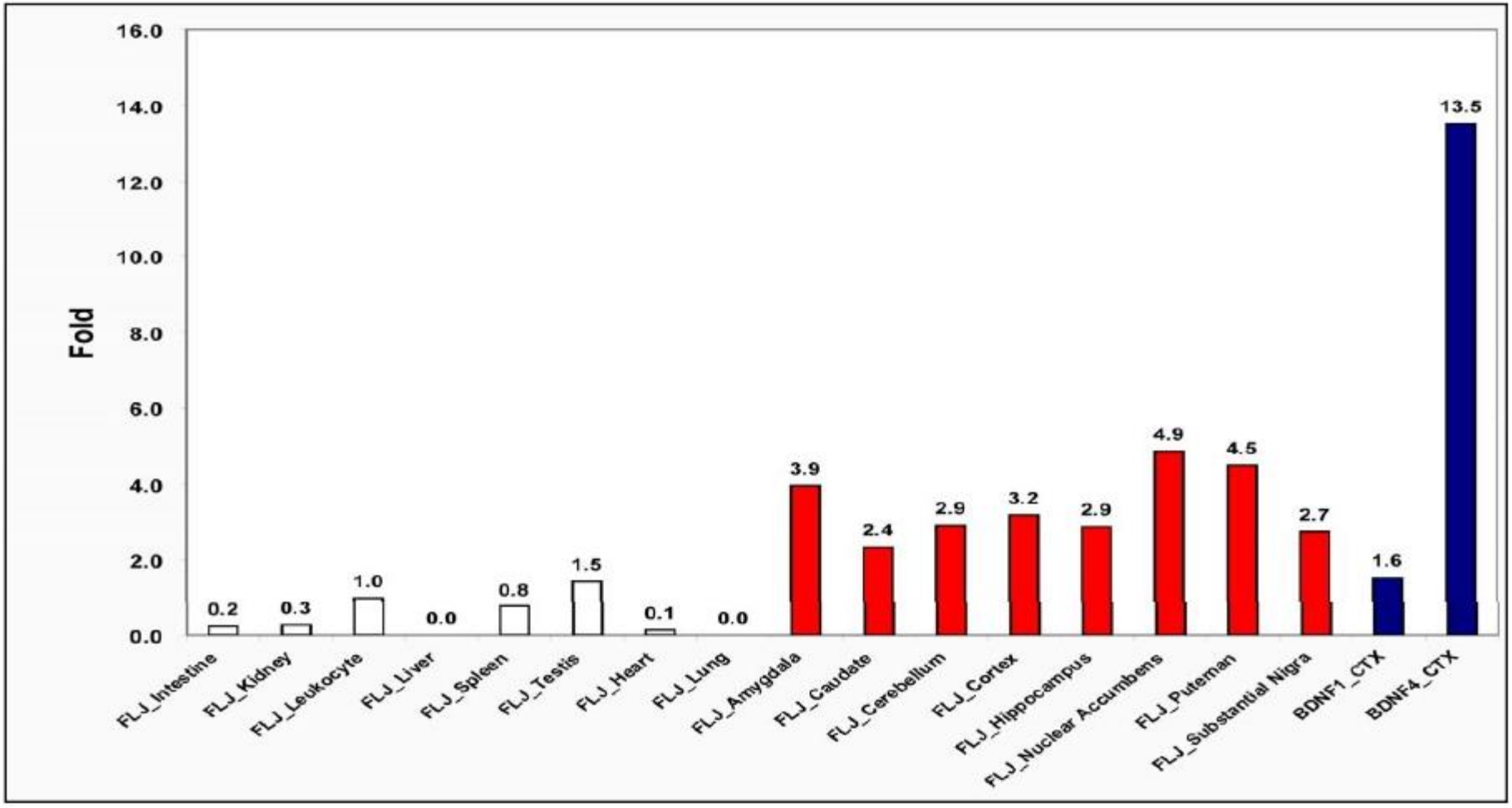
It is located in the middle of 12 binding sites of miRNA *let-7.*



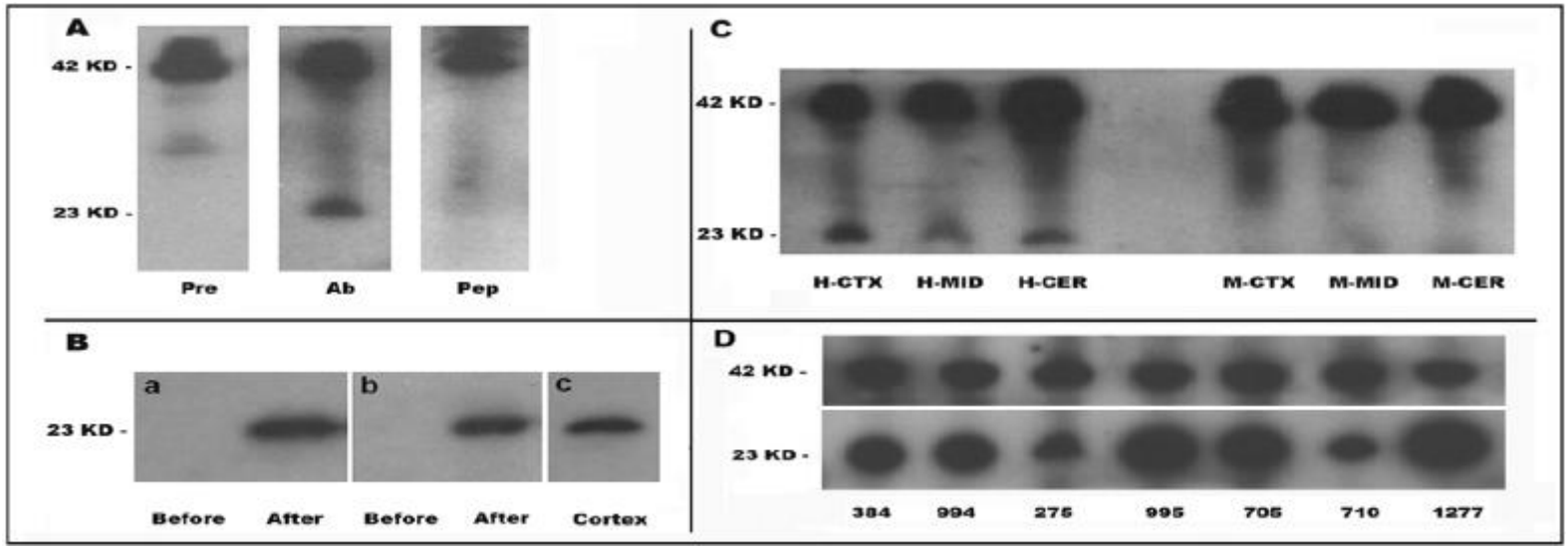
FLJ33706 is a human-specific *de novo* protein-coding gene



TaqMan-based Real-Time PCR showed that *FLJ33706* mRNA is enriched in human brain regions

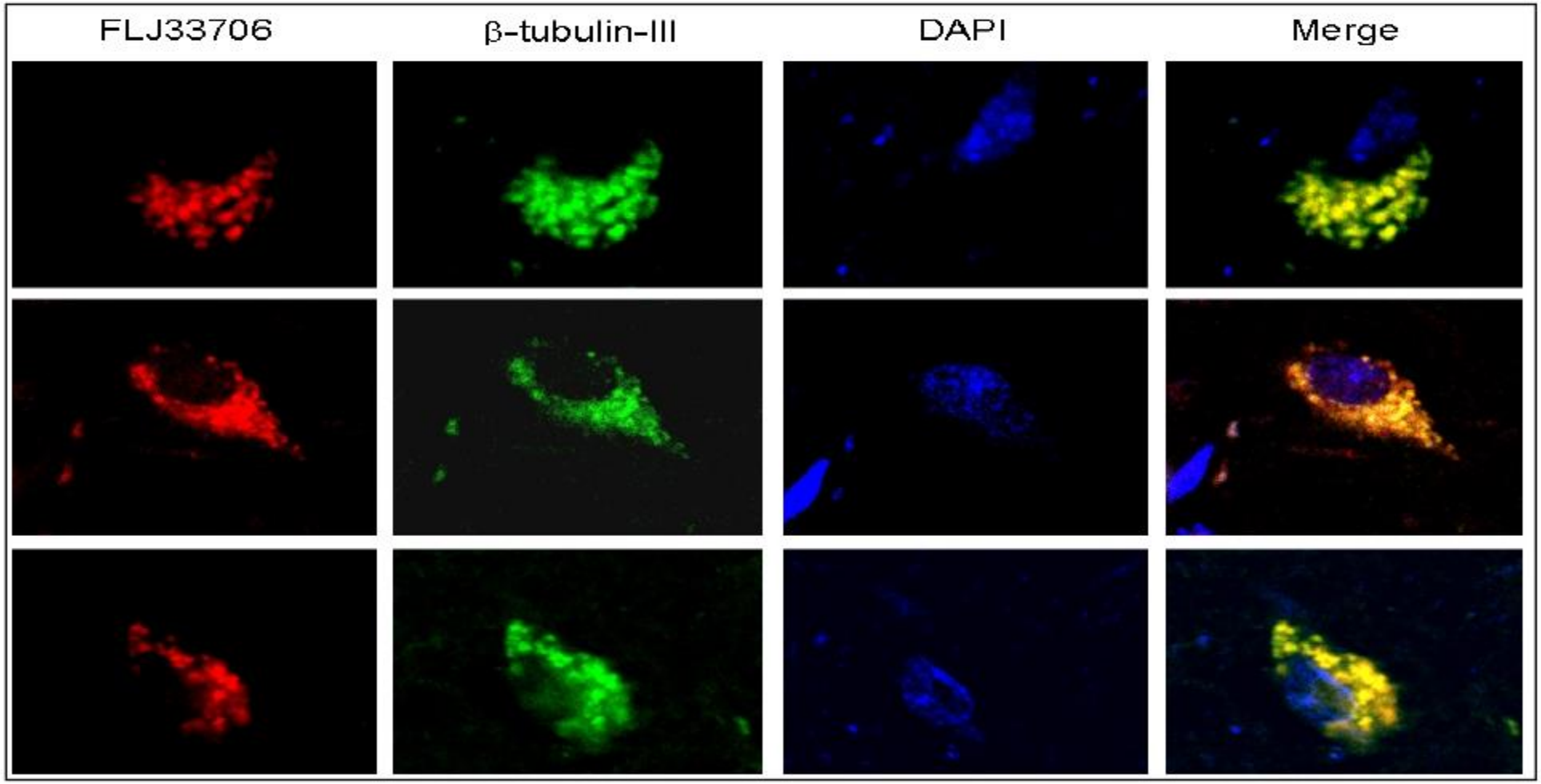


Western blot assay using an antibody designed against a 17-amino-acid peptide confirmed expression of FLJ33706 protein

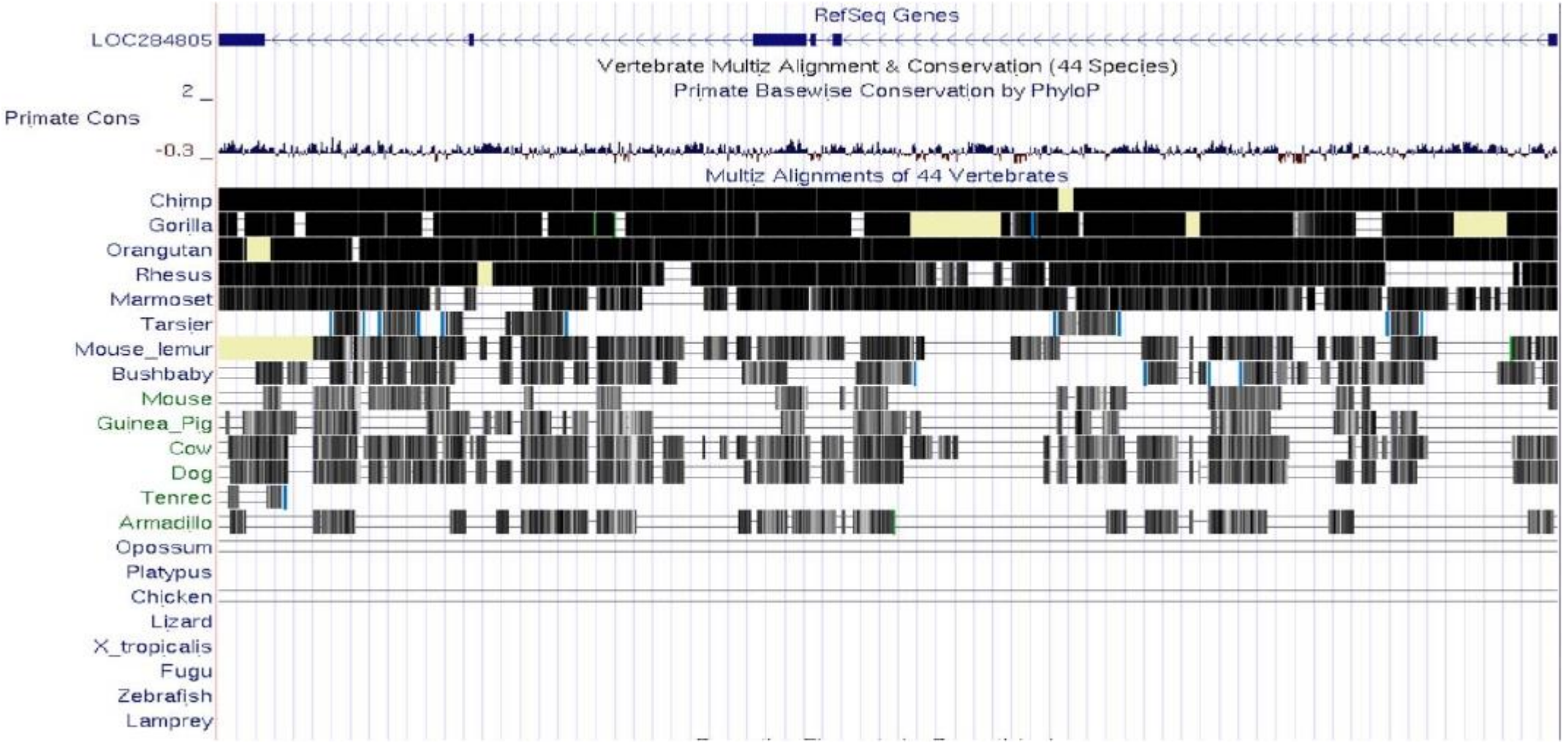


- (A) The band was not detected in pre-immune serum or in the presence of excess synthetic antigenic peptides
- (B) The band was detected only after transformation of FLJ33706 recombination plasmids in *E. coli* (a) His-tag specific antibody and (b) anti-FLJ33706.
- (C) The band was detected in human cortex, midbrain, and cerebellum, but not in mouse.
- (D) FLJ33706 expression can be detected in the cortext of seven different human individuals.

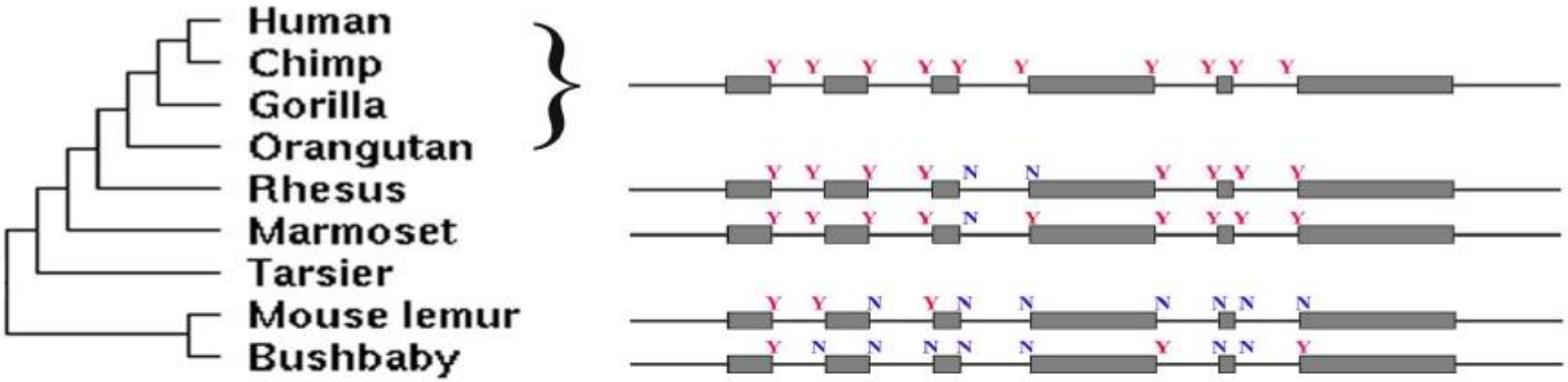
Immunohistochemistry studies of human cortex slides showed enrichment in cytoplasm of neuronal cells



The DNA segment emerged in eutherian mammals

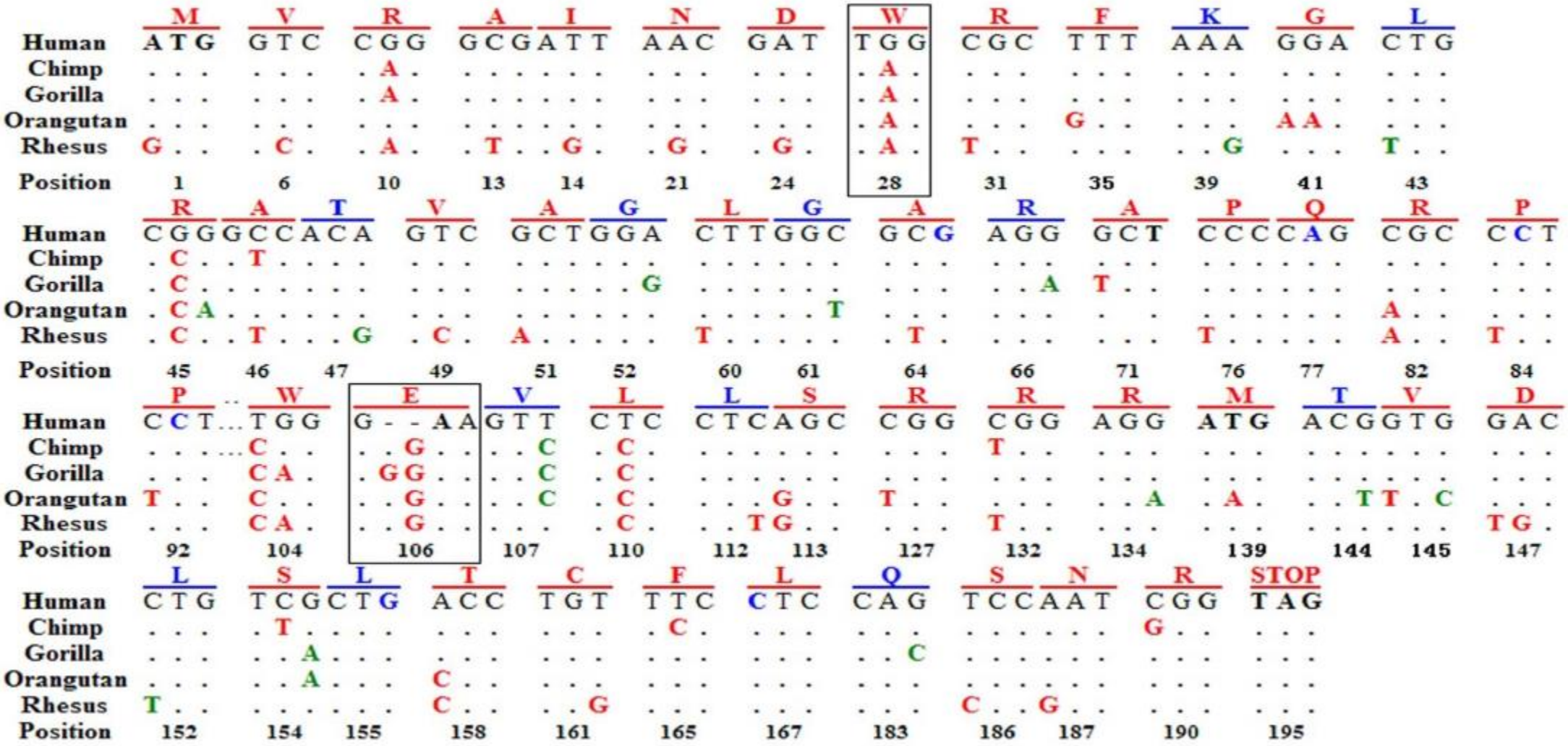


Insertion of *Alu* elements generated splicing sites

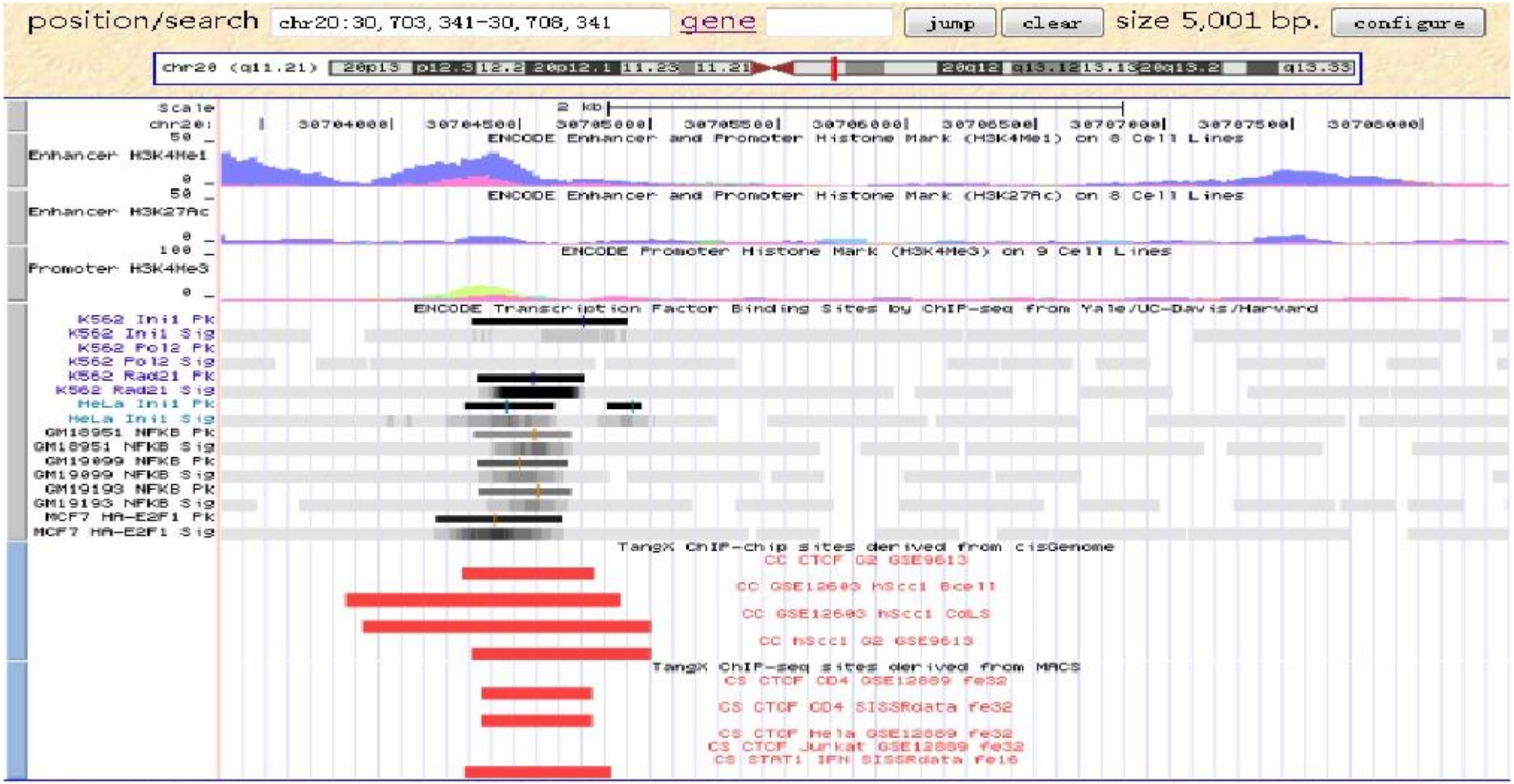


		Intron 1						Intron 2						Intron 3						Intron 4						Intron 5					
Human	...CAG	G	T	ggg...tcc	A	G	ACT...GAG	G	T	aag...ttt	A	G	AGA...CGG	G	T	aag...aac	A	G	GCC...CTG	G	T	agg...gac	A	G	AGT...CAG	G	T	acg...tcc	A	G	ACT...
Chimp	...CAG	G	T	ggg...tcc	A	G	ACT...GAG	G	T	aag...---	A	G	AGA...CCG	G	T	aag...aac	A	G	GTC...CTG	G	T	agg...gac	A	G	AGT...CAG	G	T	acg...tcc	A	G	ACT...
Gorilla	...CAG	G	T	ggg...tcc	A	G	ACT...GAG	G	T	aag...---	A	G	AGA...CCG	G	T	aag...aac	A	G	GCC...CTG	G	T	agg...gac	A	G	AGT...CAG	G	T	aca...tcc	A	G	ACT...
Orangutan	...CAG	G	T	ggg...tcc	A	G	ACT...GAG	G	T	aag...ttt	A	G	AGA...CCA	G	T	aag...aac	A	G	GCC...CTG	G	T	age...gac	A	G	--T...CAG	G	T	acg...nnn	N	N	NNN...
Rhesus	...CAG	G	T	ggg...tct	A	G	ACT...GAG	G	T	aag...---	A	G	AGA...CCG	G	A	aag...aat	A	A	GTC...CTG	G	T	age...gac	A	G	AGT...TAG	G	T	acg...tcc	A	G	ACT...
Marmoset	...CAG	G	T	gga...tcc	A	G	ACT...GAG	G	T	aag...---	A	G	AGA...CGG	C	T	aag...aac	A	G	GCC...CTG	G	T	age...tag	G	G	AGT...CAG	G	T	acg...tcc	A	G	ACT...
Mouse_lemur	...CAG	G	T	ggg...ttc	A	G	ACT...GGG	G	T	aag...---	A	A	---...CCA	G	A	gag...aac	T	G	CCA...---	-	-	---...---	-	-	---...---	-	-	---...nnn	N	N	NNN...
Bushbaby	...CAG	A	G	agg...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...CAA	G	T	age...---	-	-	---...---	-	-	---...tcc	A	G	ATT...
Mouse	...CAG	-	-	---...tcc	A	G	ACT...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...-tg	G	A	GCC...
Guinea_Pig	...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...-cc	G	C	ACT...
Cow	...GAG	-	C	tgg...tgc	A	G	ACC...GAG	G	T	cac...---	-	-	---...---	-	-	---...agc	A	G	CCA...---	-	-	---...---	-	-	---...---	-	-	---...tgc	A	G	ATG...
Dog	...CAG	-	T	cgg...---	-	-	---...GAG	G	T	cac...---	-	-	---...---	-	-	---...a-c	A	G	CCA...---	-	-	---...---	-	-	---...---	-	-	---...tcc	A	G	ATG...
Armadillo	...CAG	G	T	ggg...---	-	-	---...---	G	C	ctg...---	-	-	---...AAG	G	A	agg...aac	A	G	CCA...---	-	-	---...---	-	-	---...---	-	-	---...---	-	-	---...

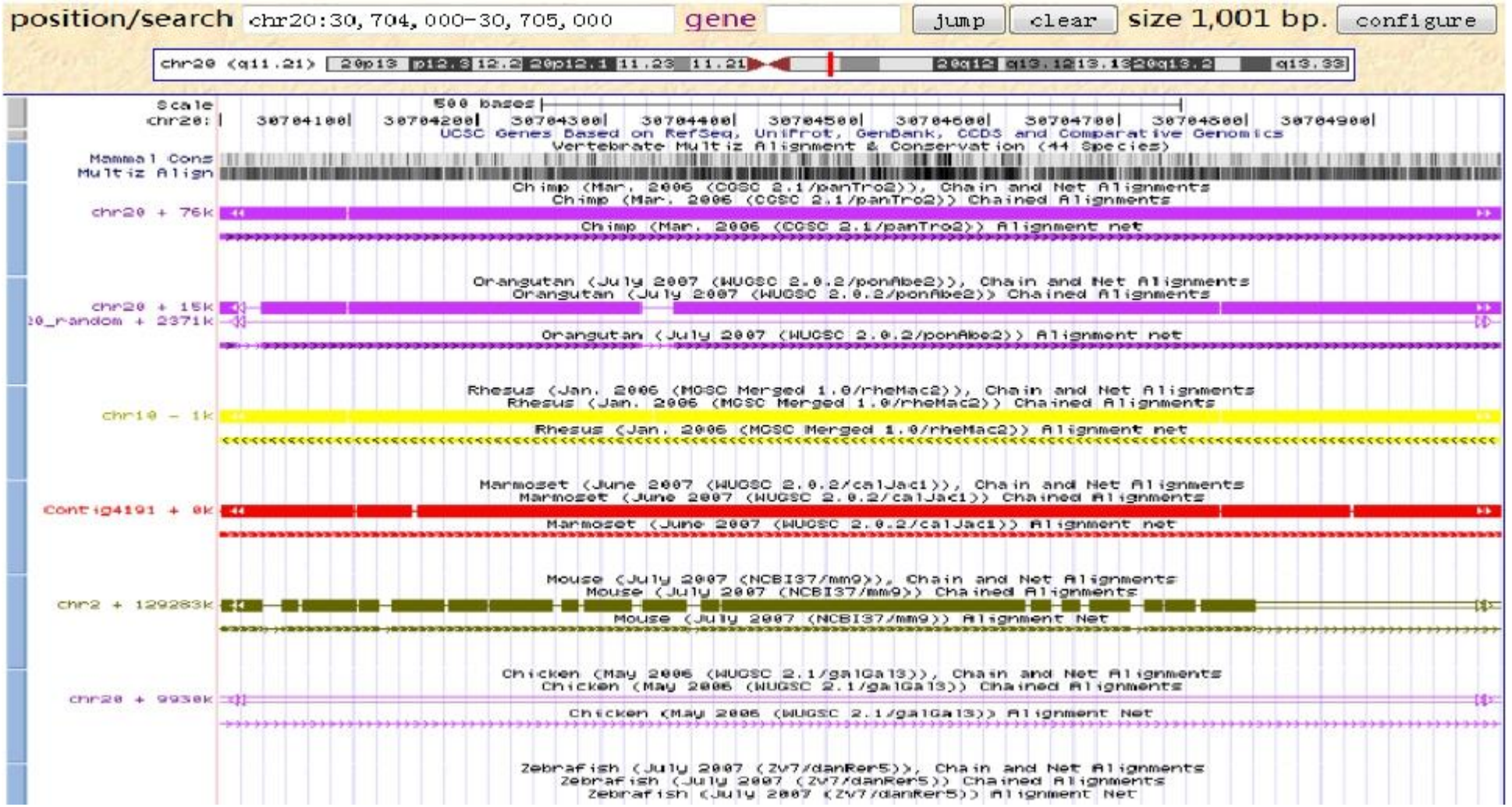
Two changes in human escaped two stop codons



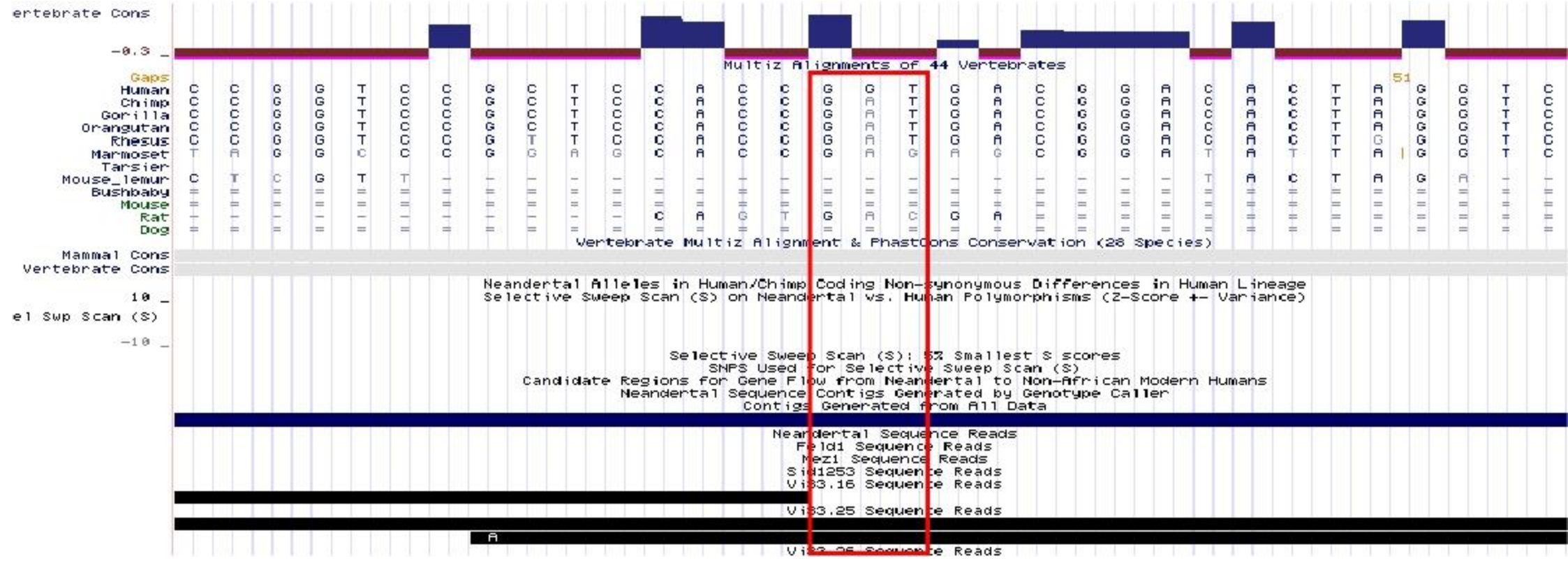
There are signals of enhancer and transcription factor binding sites in the 5kb upstream regions



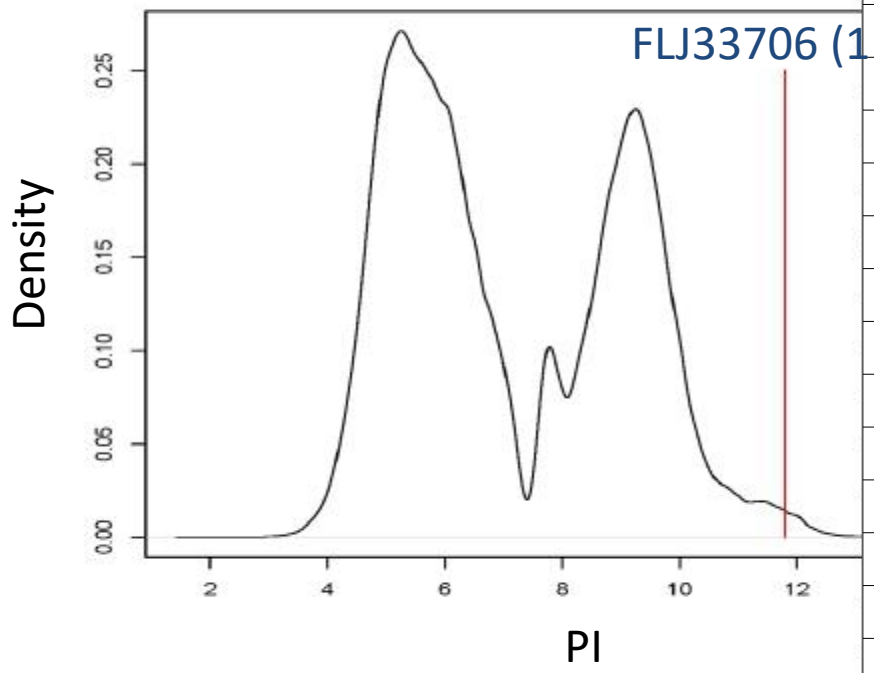
Promoter region is absent in chicken/zebrafish, emerged in mouse, and is similar in rhesus and chimpanzee



The Open Reading Frame is intact in Neadertal genome



FLJ33706 has high PI



GOTerm	FDRq-value
RNAbinding	5.50E-08
cytosolicribosome	3.68E-07
macromolecularcomplex	1.63E-06
cytosoliclargeribosomalsubunit	4.61E-05
RNAsplicing	6.71E-05
cytosolicpart	7.73E-05
ribosomalsubunit	4.54E-04
largeribosomalsubunit	7.89E-04
intracellularorganellepart	9.99E-04
organellepart	0.001136772
ribonucleoproteincomplex	0.003187642
cellularbiosyntheticprocess	0.007101674
MHCclassIIreceptoractivity	0.009220135
translation	0.010595406
mRNAprocessing	0.012153244
RNAprocessing	0.012167141
structuralconstituentofribosome	0.017365179
mRNAmetabolicprocess	0.020473341
macromoleculemetabolicprocess	0.021017467
intracellularnon-membrane-boundorganelle	0.024935299
non-membrane-boundorganelle	0.024935299
ribosome	0.036638186

GO enrichment of proteins with PI > 11 (FDR < 0.05)

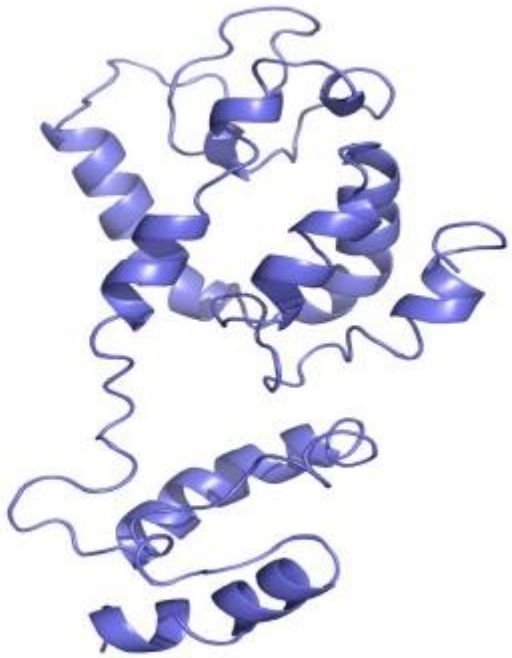
Unpublished

Predicted Secondary Structure include four helices & one beta strand

Unpublished

Predicted 3D Structure (probably not reliable)

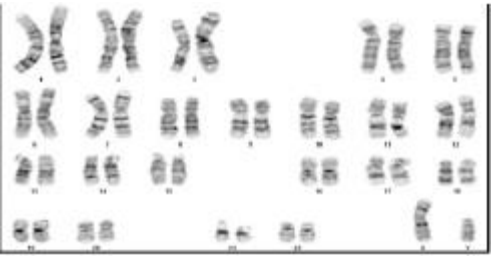
Scored as best by I-TASSER



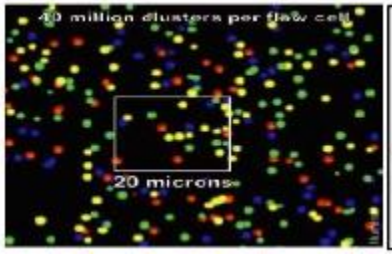
Another putative conformation

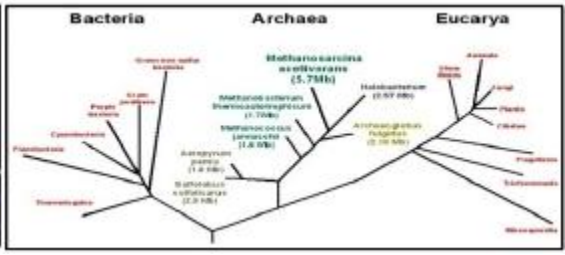
How many other human-specific *de novo* genes are there?

Where did they originate from?



TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



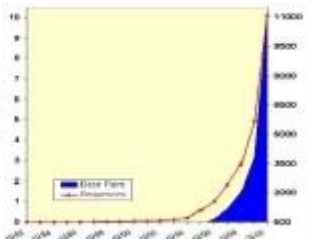


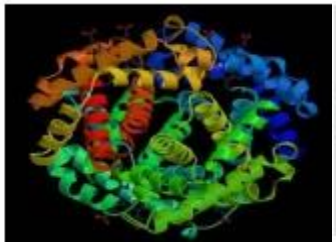
Unit 4: Origination of de novo Genes from Noncoding RNAs

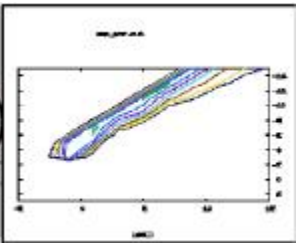
Le Zhang, Ph. D.

Computer Science Department

Southwest University

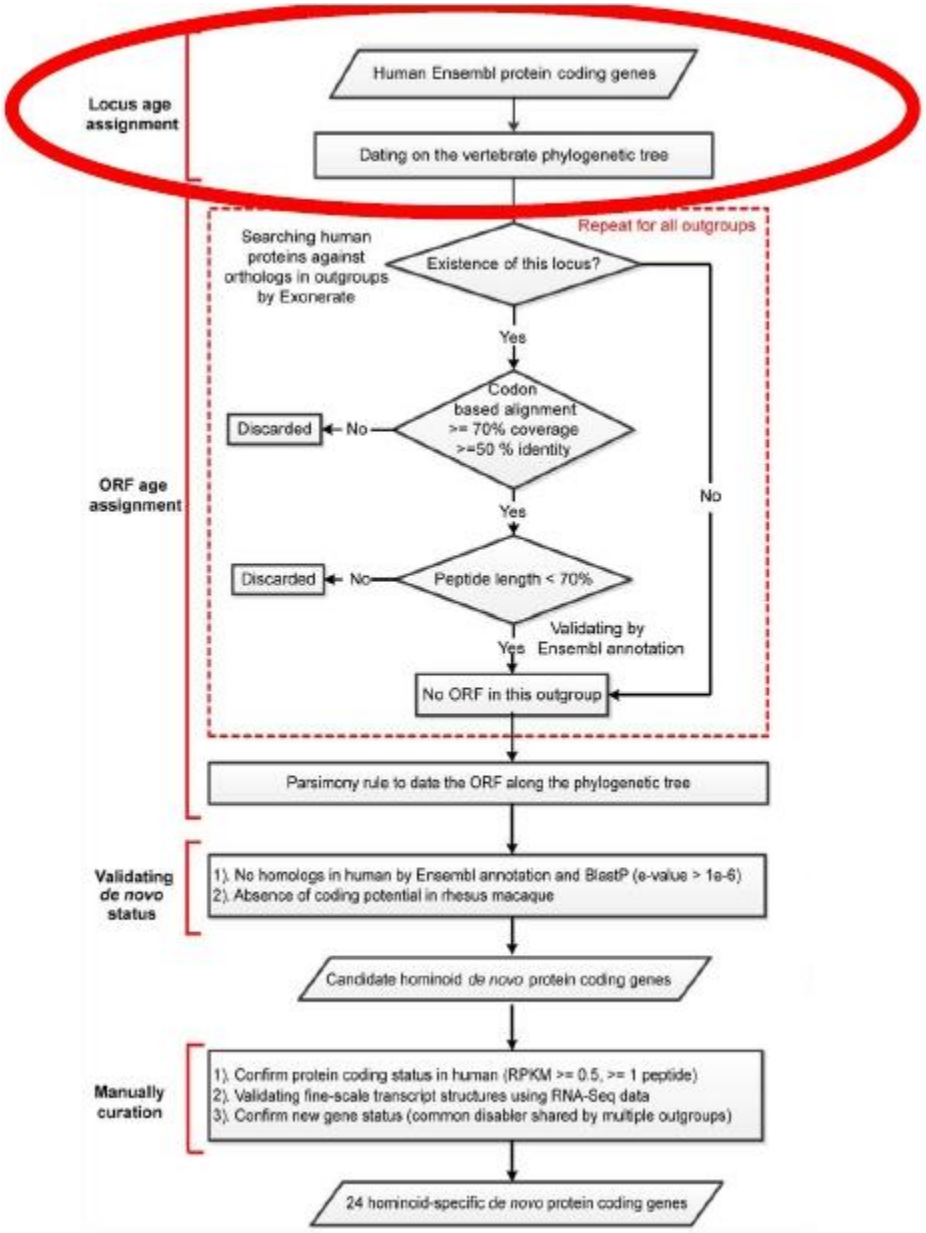






How many other human-specific *de novo* genes are there?

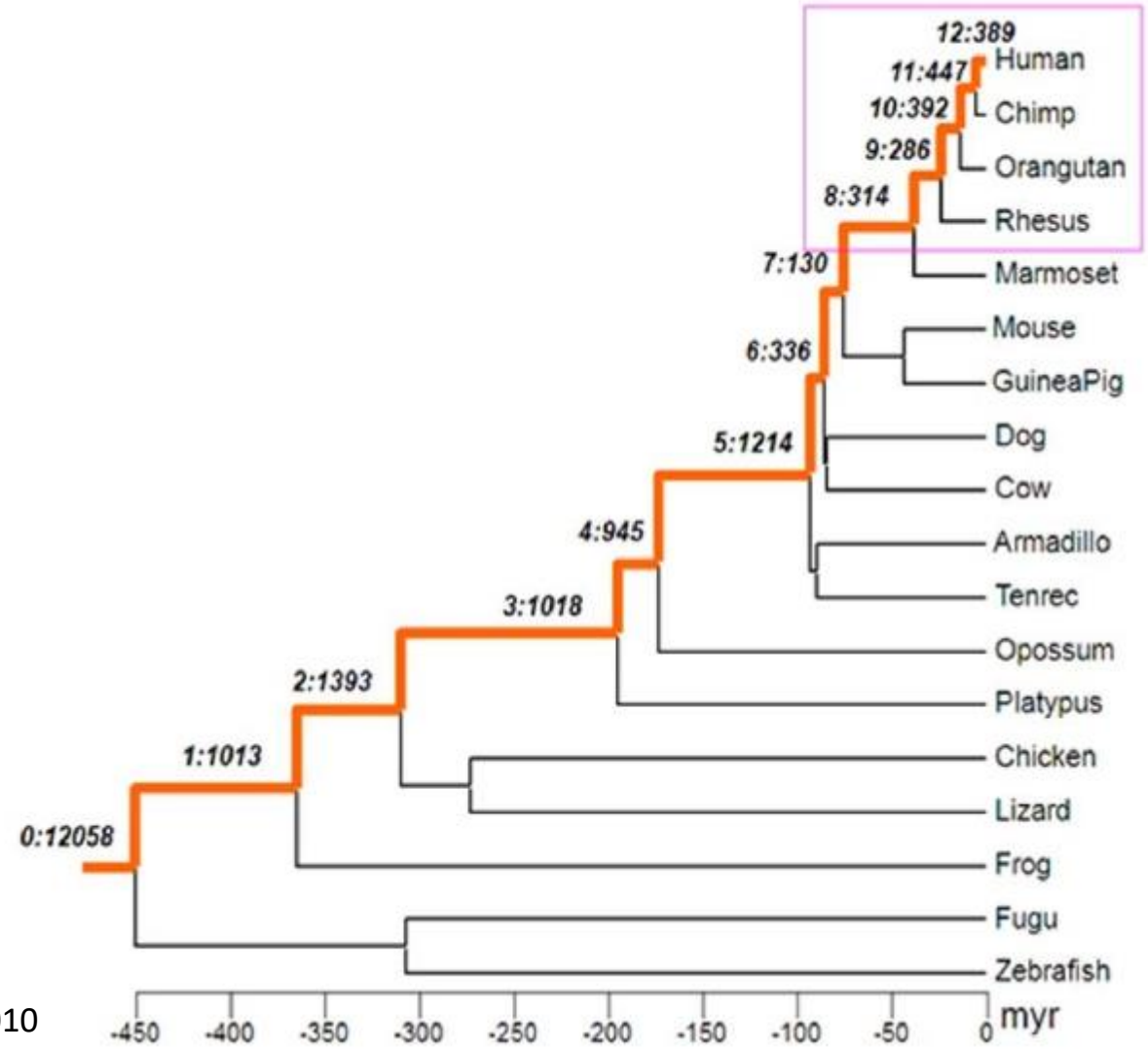
Where did they originate from?



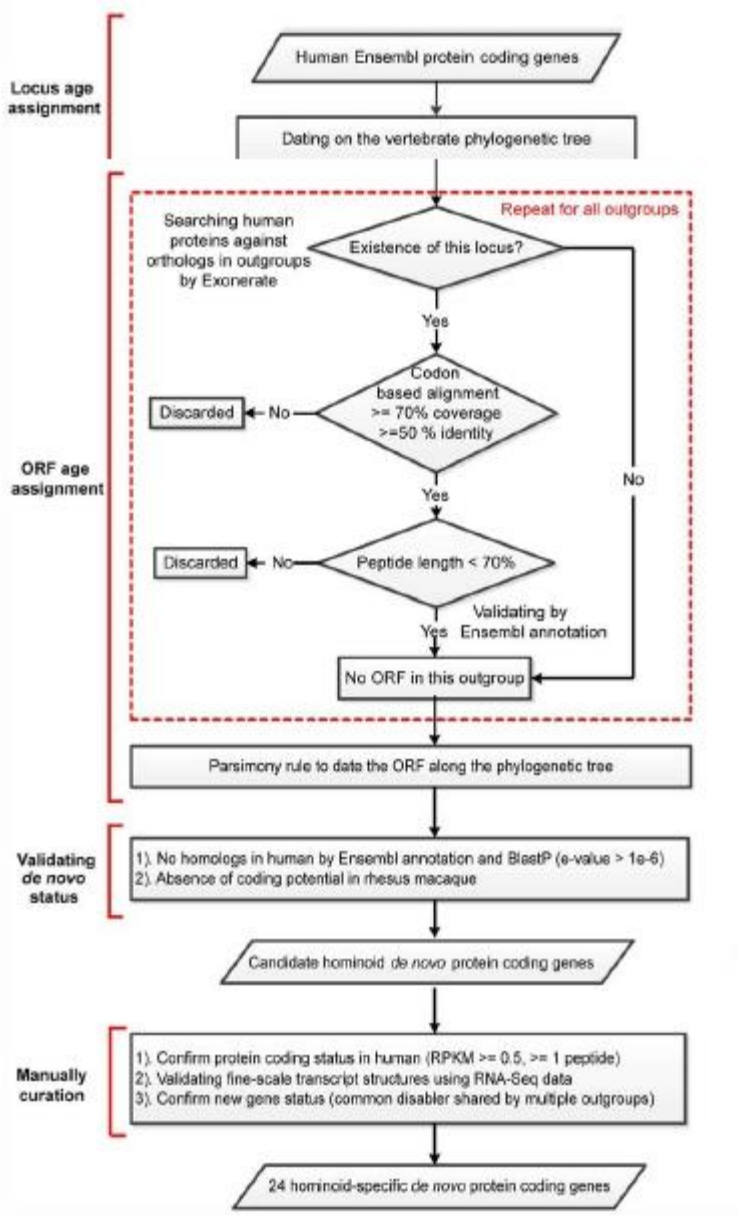
Genome-wide identification of human- and human-chimpanzee-specific *de novo* genes

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

Inferring the origination times of human gene loci



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



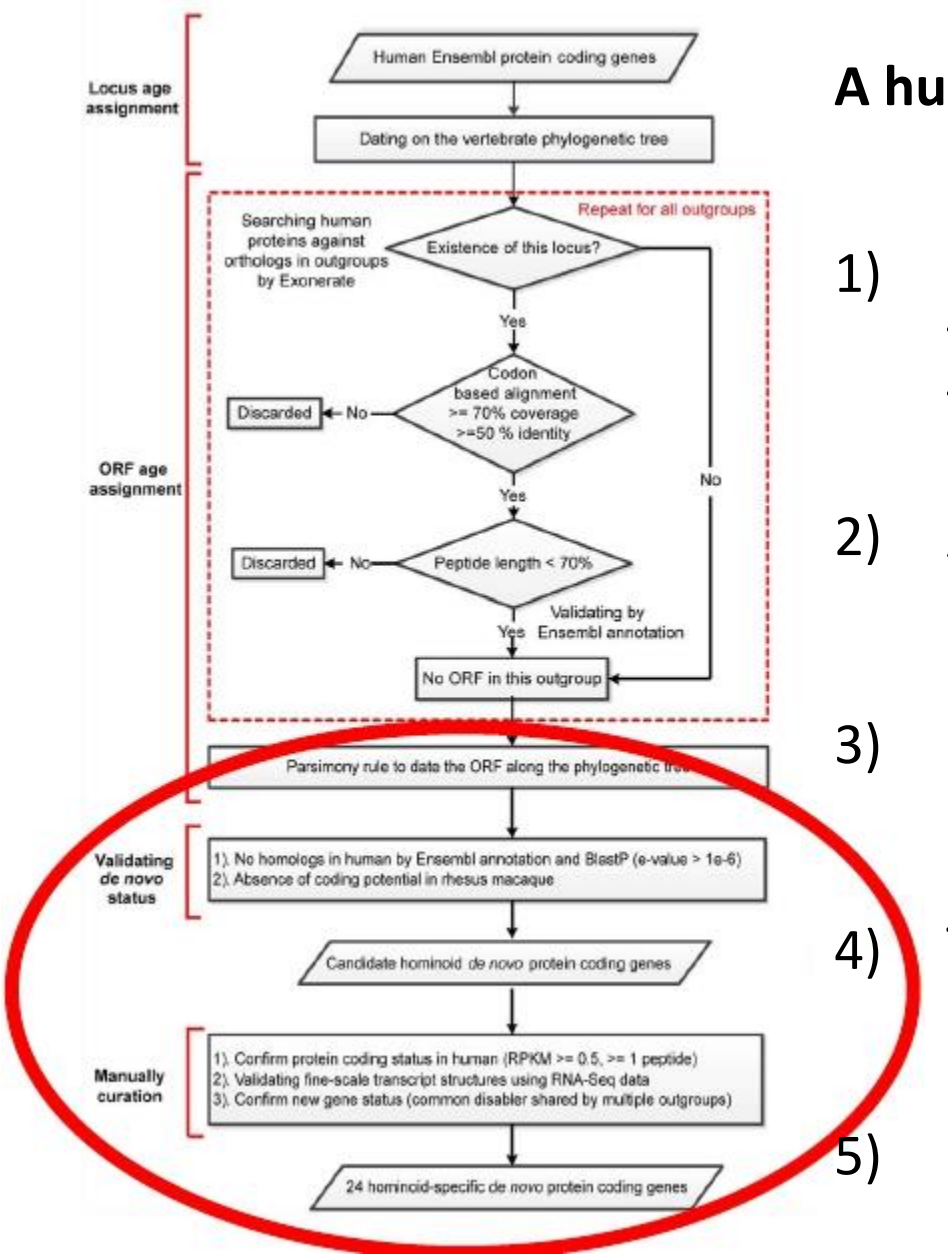
Inference of age of ORF

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

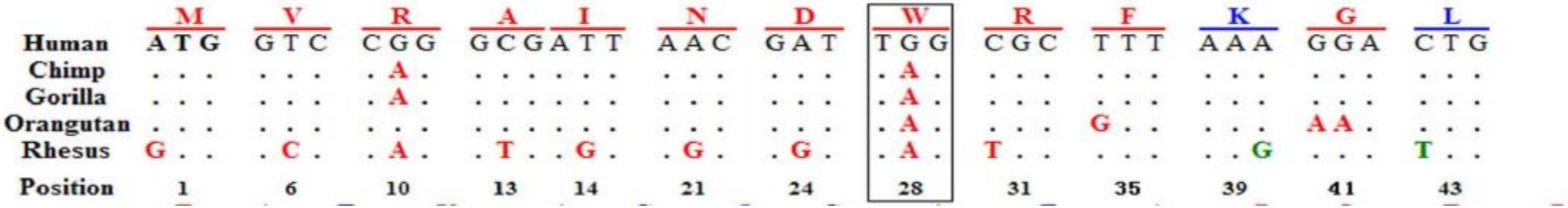
(2) Ensembl annotation did not identify any ortholog.



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

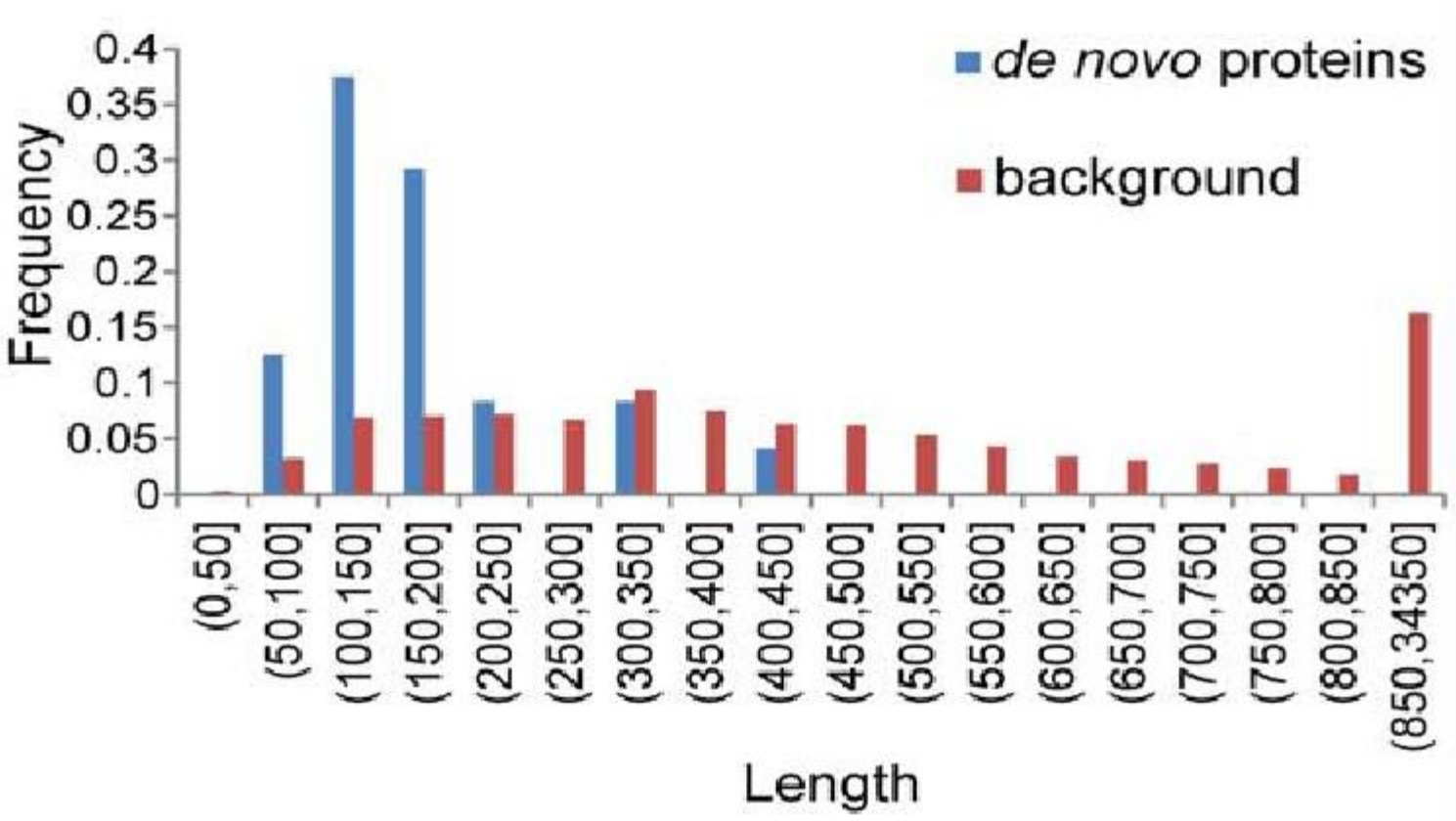
11 encode proteins only in human

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



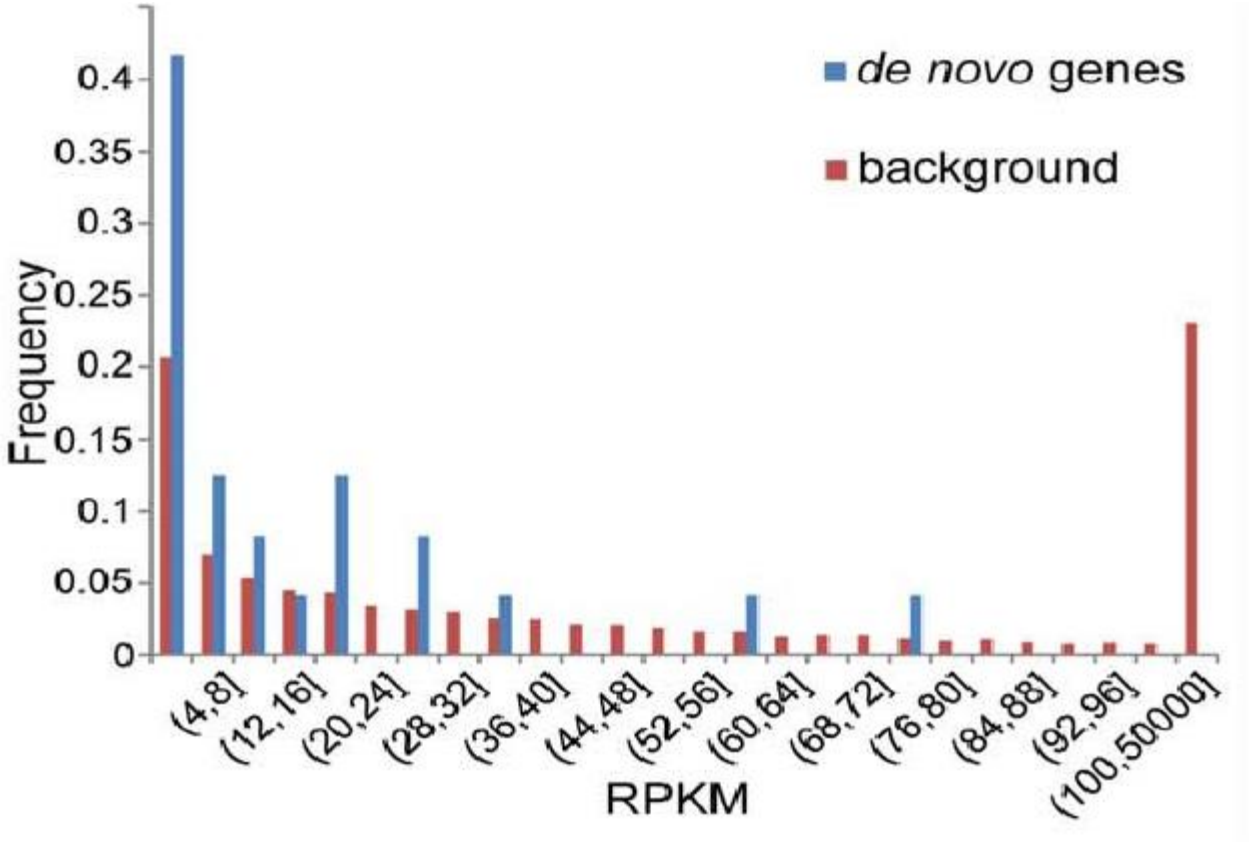
Xie *et al.*, *PLoS Genet.*, 2012

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

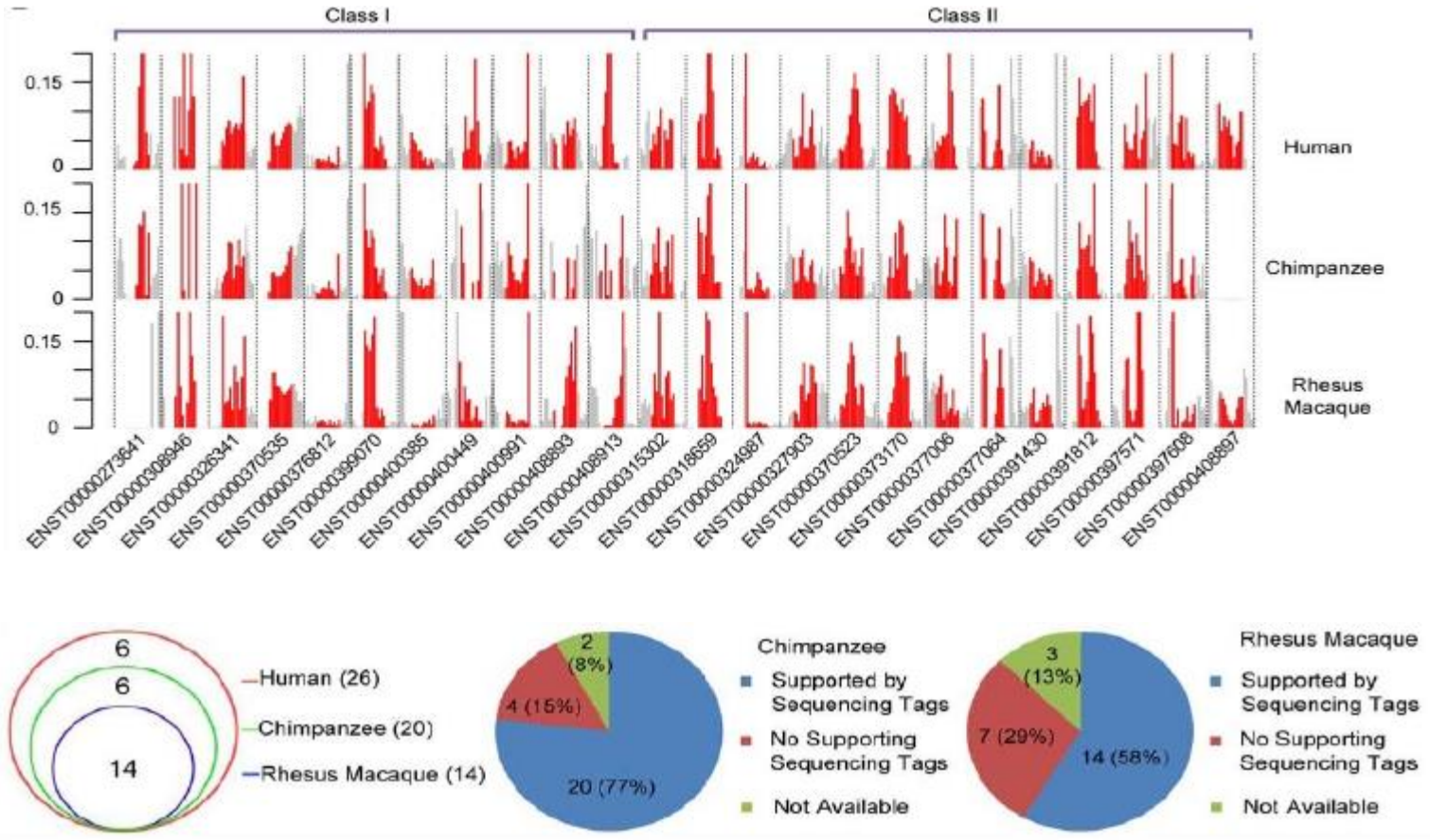
ORF first or regulated transcription first?

transcription leakage/noise until ORF

versus

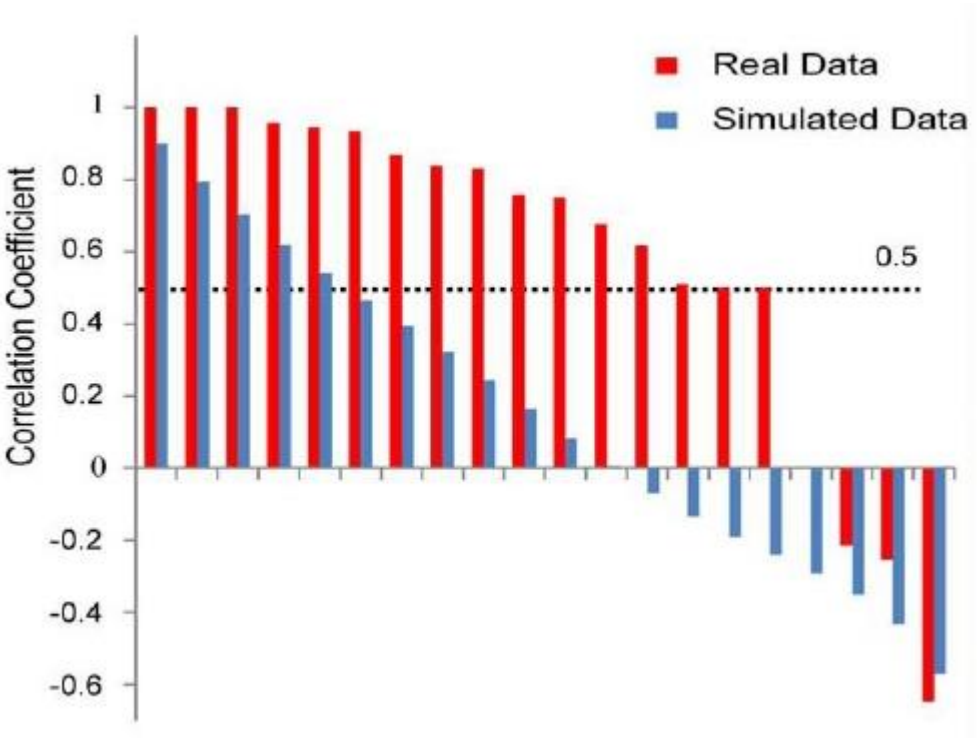
regulated transcriptional profile and structure of ncRNA

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

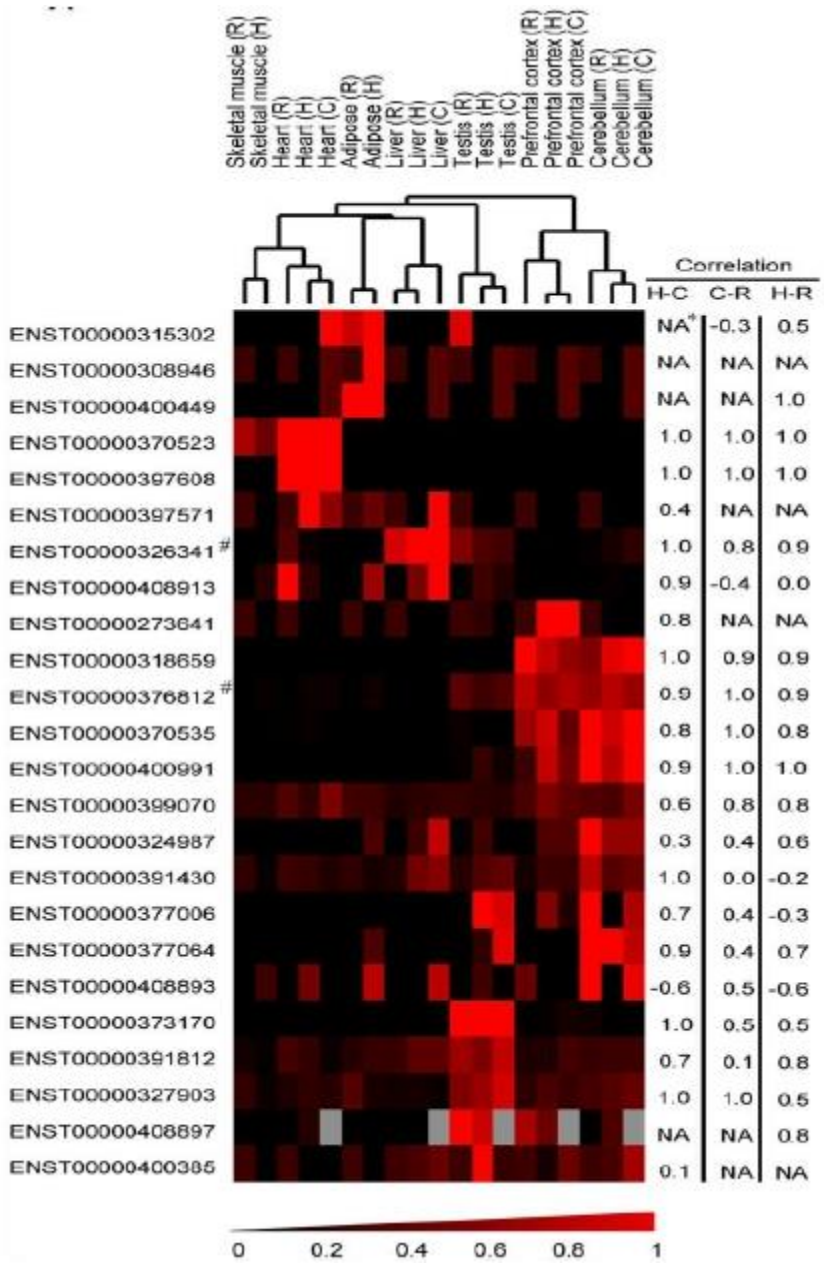


Xie *et al.*, *PLoS Genet.*, 2012

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

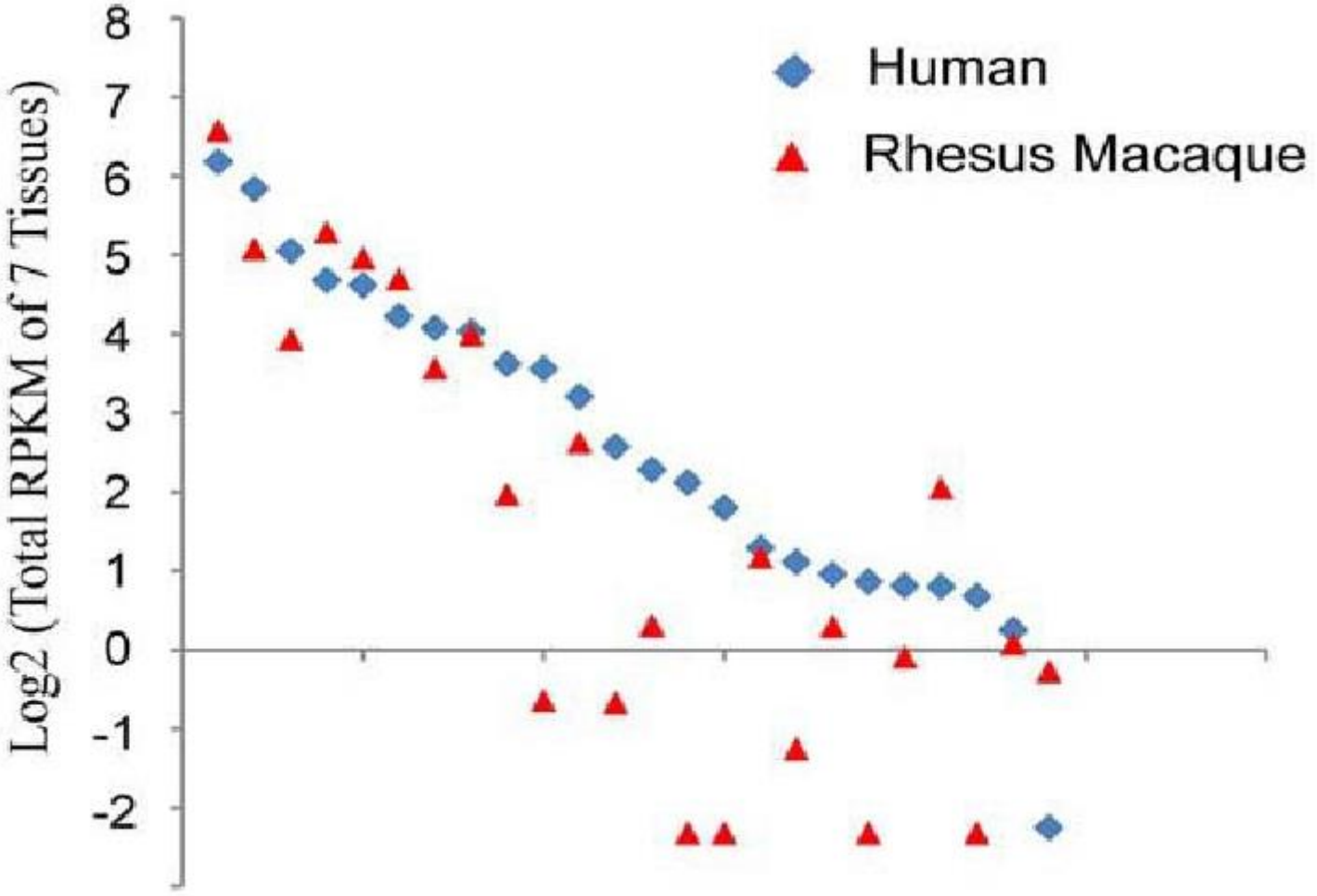


P-Value < 0.0001



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

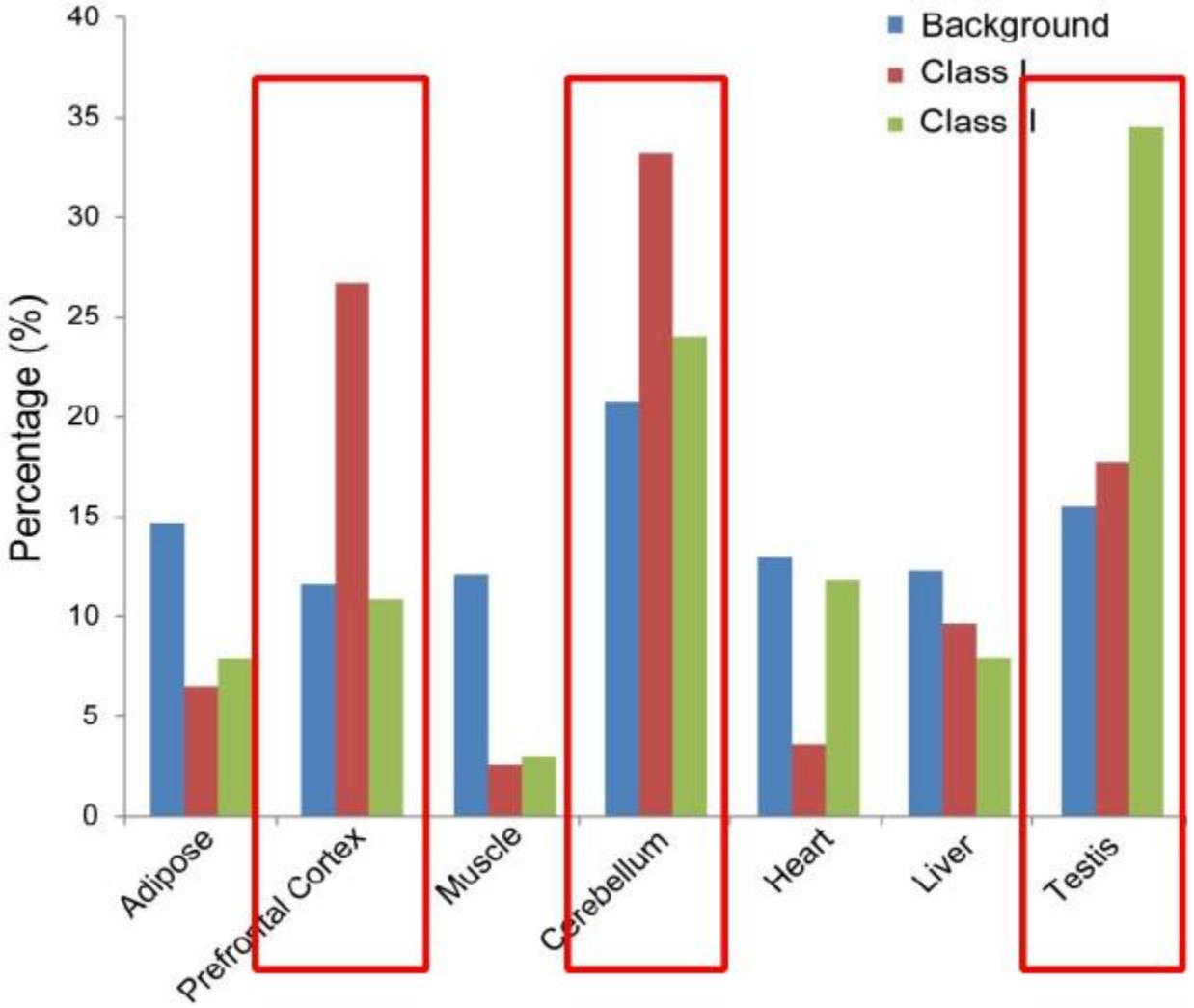
Non-coding genes tend to have correlated, but lower, transcription level than their protein-coding orthologs



R-squared = 0.56
P-Value = 2.8×10^{-5}

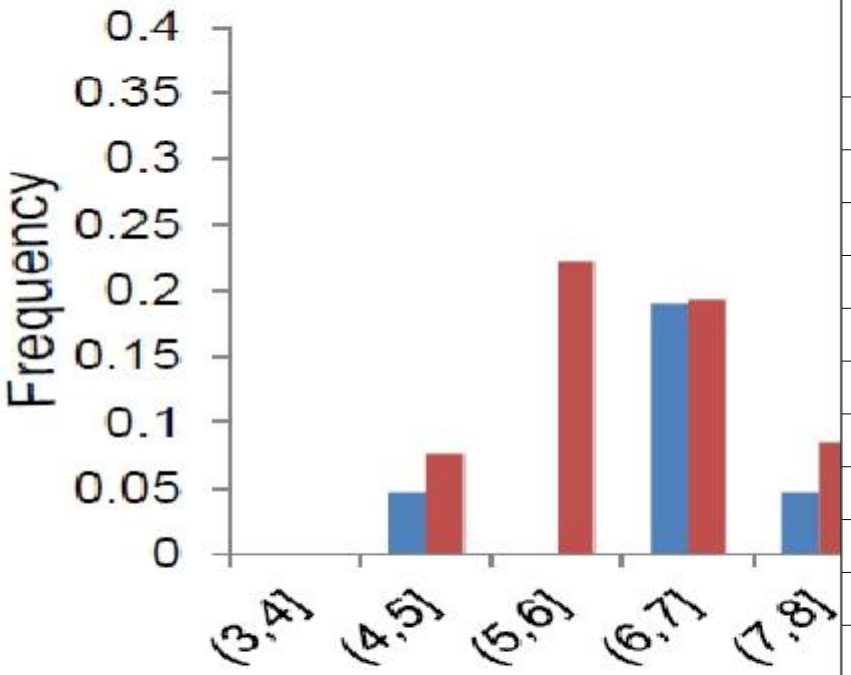
Xie *et al.*, *PLoS Genet.*, 2012

de novo genes have enriched expression in brain and testis



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

The pI values of



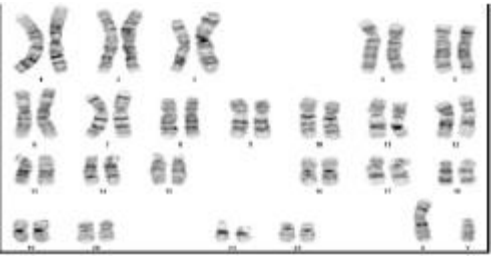
P-Value = 1.4×10^{-4}

GOTerm	FDRq-value
RNAbinding	5.50E-08
cytosolicribosome **are**	**higher** 3.68E-07
macromolecularcomplex	1.63E-06
cytosoliclargeribosomalsubunit	4.61E-05
RNAsplicing	6.71E-05
cytosolicpart	7.73E-05
ribosomalsubunit	4.54E-04
largeribosomalsubunit	7.89E-04
intracellularorganellepart	9.99E-04
organellepart	0.001136772
ribonucleoproteincomplex	0.003187642
cellularbiosyntheticprocess	0.007101674
MHCclassIIreceptoractivity	0.009220135
translation	0.010595406
mRNAprocessing	0.012153244
RNAprocessing	0.012167141
structuralconstituentofribosome	0.017365179
mRNAmetabolicprocess	0.020473341
macromoleculemetabolicprocess	0.021017467
intracellularnon-membrane-boundorganelle	0.024935299
non-membrane-boundorganelle	0.024935299
ribosome	0.036638186

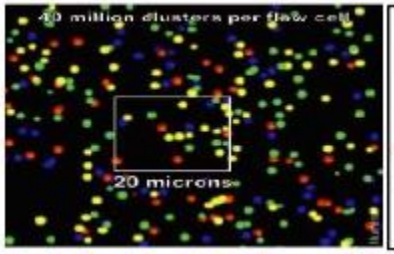
Summary

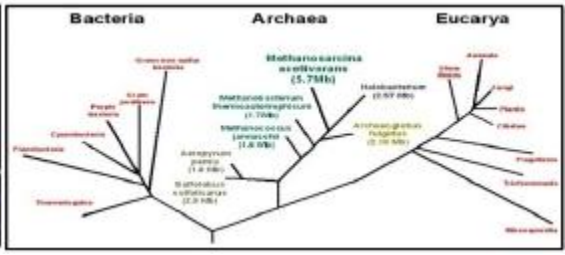
Bioinformatic methods and analyses can play key roles in evolutionary biology.

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



TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



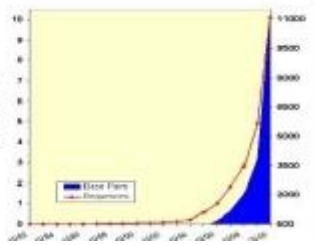


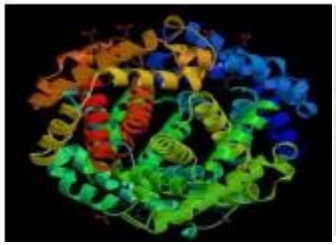
Unit 5: What is bioinformatics?

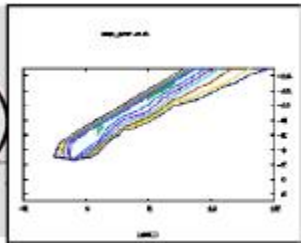
Le Zhang, Ph. D.

Computer Science Department

Southwest University







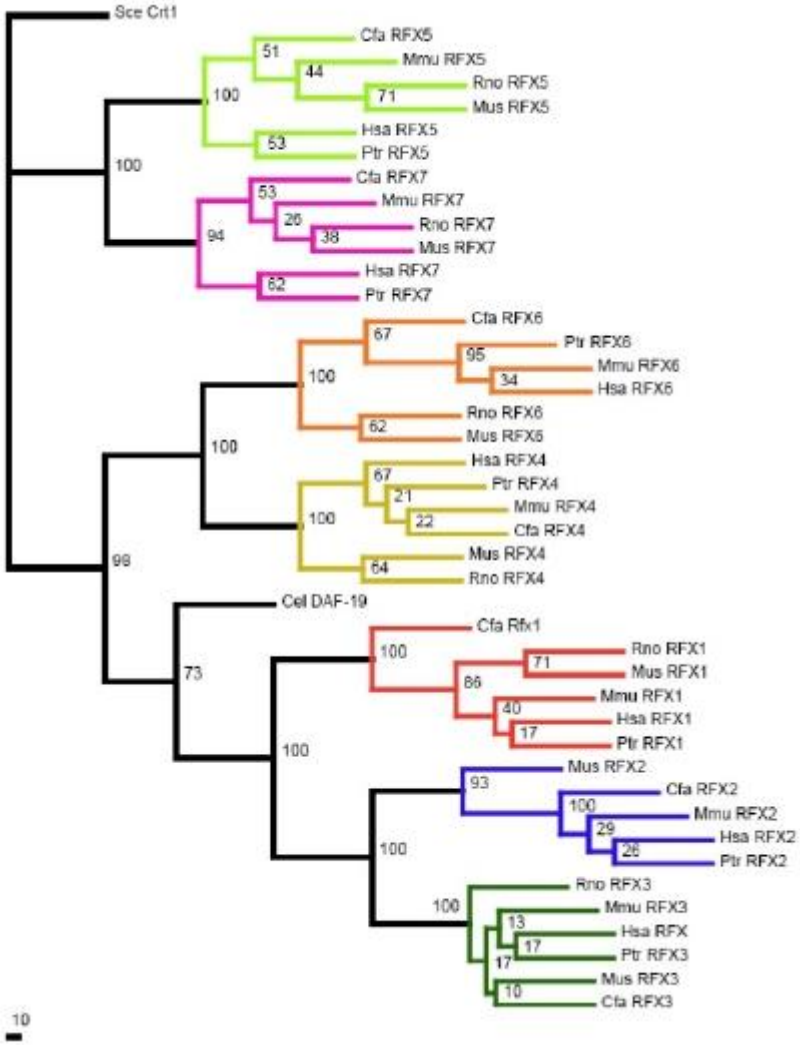
Outline

- What is phylogeny estimation?
- Why estimate phylogeny?
- How to estimate phylogeny?
  - Traditional approaches
  - Bayesian approaches

What is phylogeny?

- Phylogenetics: the study of evolutionary relationships among groups of organisms (e.g. species, populations) or genes, which are discovered through molecular sequencing data and morphological data matrices. (modified from Wikipedia)

- Phylogenetic tree: A graph depicting the ancestor–descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences.



Phylogenetic analysis of mammalian RFX genes.
The species names included in this figure are abbreviated. They are: Mus–mouse (*Mus musculus*); Rno–Rat (*Rattus norvegicus*); Cfa–dog (*Canis familiaris*);
Ptr–chimpanzee (*Pan troglodytes*); Mmu–monkey (*Macaca mulatta*) and Hsa–human (*Homo sapiens*).

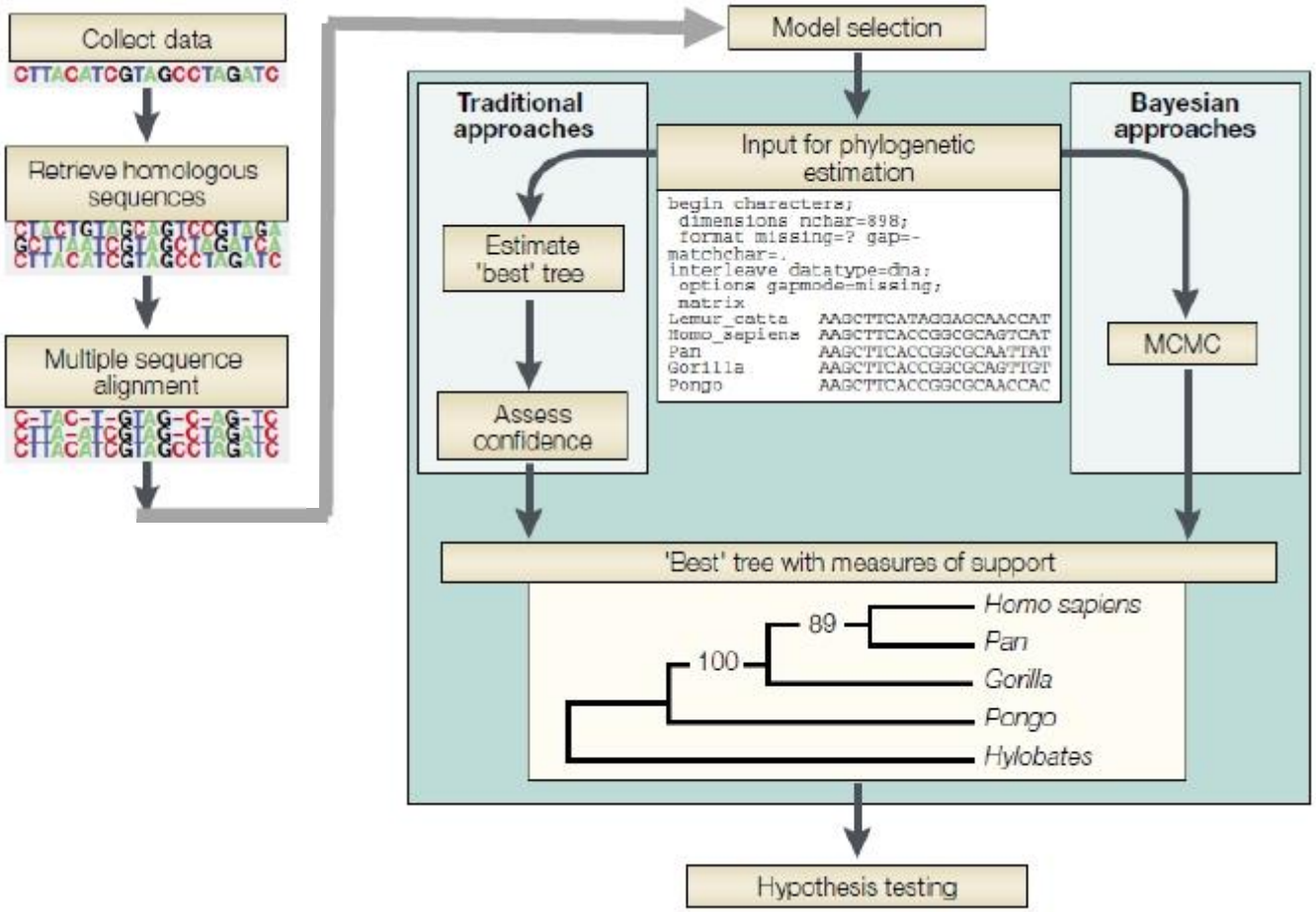
Why do phylogeny estimation?

- Detection of orthology and paralogy
- Estimating divergence times
- Reconstructing ancient proteins
- Finding the residues that are important to natural selection
- Detecting recombination points
- Identifying mutations likely to be associated with disease
- Determining the identity of new pathogens

How to estimate phylogeny?

- Assumption
  - As the time increases since two sequences diverged from their last common ancestor, so does the number of differences between them.
- Basic idea
  - Count the number of differences between sequences and group those that are most similar.
- Complexity
  - The rate of sequence evolution is not constant over time.
  - Natural selection or changing mutational biases exist.
  - Many of the sites in a DNA sequence are not helpful.

The phylogenetic inference process

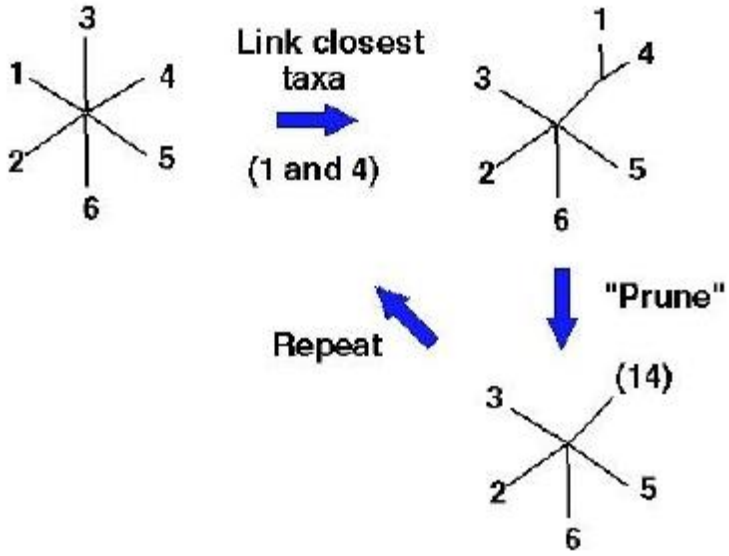


Traditional approaches

- Neighbour-joining (NJ) algorithm

- Tree searches that use an optimality criterion
  - Parsimony
  - maximum likelihood (ML)

Neighbour-joining



- Description
- Advantages
  - Fast
- Disadvantages
  - Information is lost in compressing sequences into distances.
  - Reliable estimates of pairwise distance can be hard to obtain for divergent sequences.
- Software
  - PAUP*
  - MEGA
  - PHYLIP

Parsimony

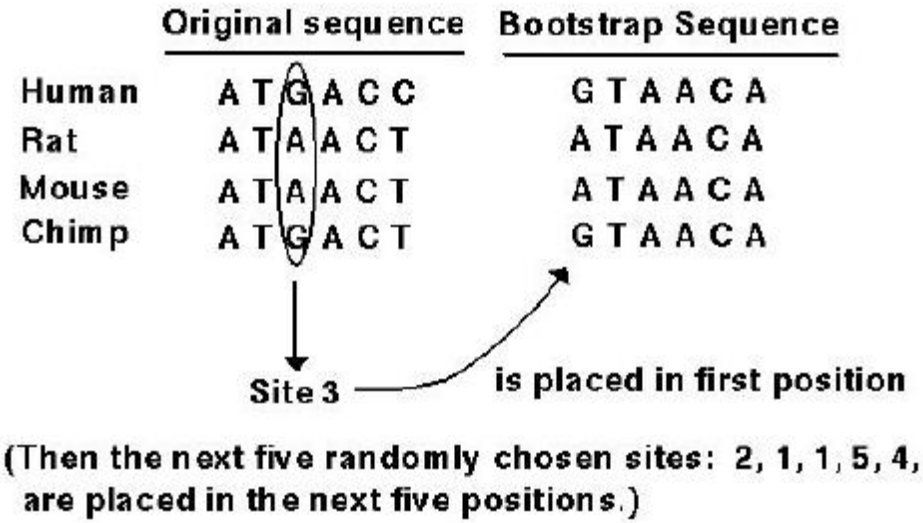
- Description
  - To determine the tree (or trees) that require the fewest number of mutations in order to explain the data that you have.
- Advantages
  - Fast enough for the analysis of hundreds of sequences.
  - Robust if branches are short (closely related sequences or dense sampling)
- Disadvantages
  - Can perform poorly if there is substantial variation in branch lengths.
- Software
  - PAUP*
  - NONA
  - MEGA
  - PHYLIP

Maximum likelihood

- Description
  - The tree that has the highest probability of producing the observed sequences $P(x_u^{\bullet}|T, t_{\bullet})$ is preferred.
- Advantages
  - The likelihood fully captures what the data tell us about the phylogeny under a given model.
- Disadvantages
  - Can be prohibitively slow (depending on the thoroughness of the search and access to computational resources)
- Software
  - PAUP*
  - PAML
  - PHYLIP

Assessing confidence — the bootstrap

- A high percentage of the bootstrap replicates implies that if another data set were collected, there is a good chance that the group would be recovered.
- Chief drawback: computational burden

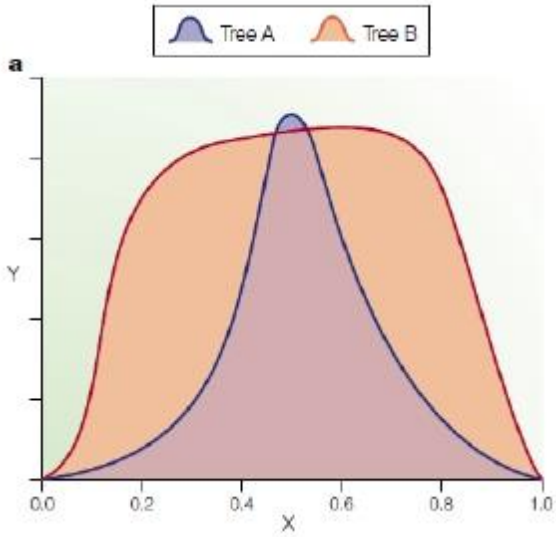


Hypothesis testing

- Use a phylogenetic analysis to determine whether an unknown virus belongs to ‘group A’ or ‘group B’.
- A tree with representatives of both candidate groups and the unknown sample is constructed, and the unknown sequence is intermingled with those from group A.
- The traditional approach involves finding the best tree in which the unknown sample clusters with the group B viruses, and then assessing how much worse this tree is compared to the best tree found in the original search.
- If the placement of the unknown with group B scores much worse than the optimal solution, then the data reject the possibility of the unknown sample actually belonging to group B.

Bayesian phylogenetics

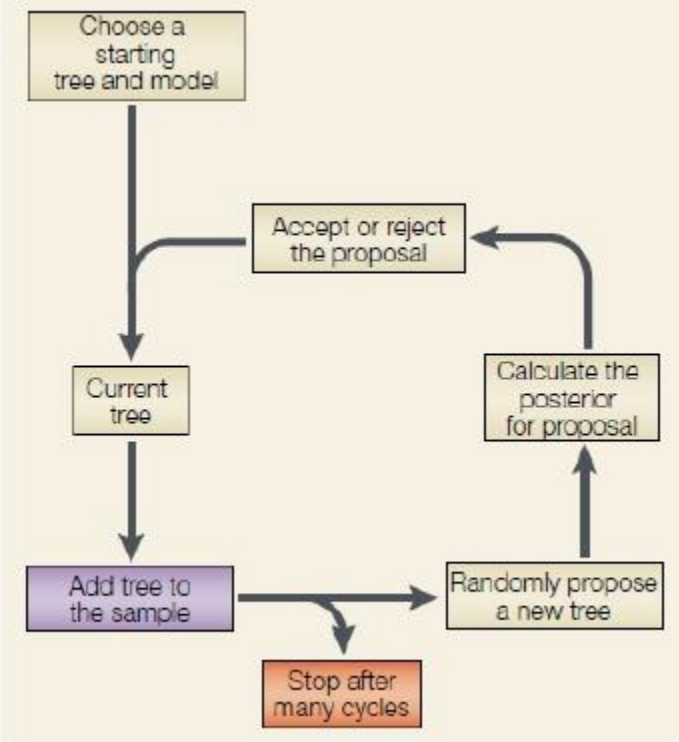
- Description
  - To maximize the posterior probability $P(\text{tree}|\text{data})$
  - $$P(T, t_\bullet|x^\bullet) = \frac{P(x^\bullet|T, t_\bullet)P(T, t_\bullet)}{P(x^\bullet)}$$
- Advantages
  - It has a strong connection to the maximum likelihood method.
  - The primary analysis produces measures of uncertainty.
  - It allows complex models of sequence evolution to be implemented.
  - It doesn't rely on the molecular clock assumption to estimate divergence times.
  - The nuisance parameters are integrated out(marginalized) to obtain the marginal posterior probability of a tree.

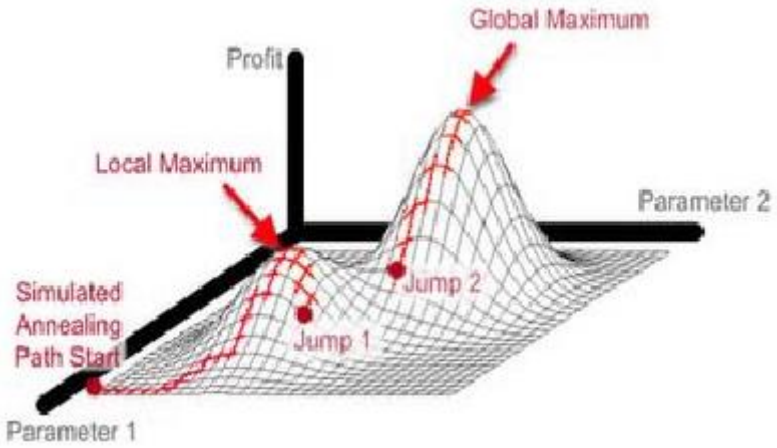


Bayesian phylogenetics

- Disadvantages
  - The prior distributions for parameters must be specified.
  - It can be difficult to determine whether the MCMC approximation has run for long enough.
- Software
  - MrBayes
  - BAMBE

Markov chain Monte Carlo





http://www.stanford.edu/~hwang41/

Conclusion

- The estimation of phylogenies has become a regular step in the analysis of new gene sequences.
- MCMC-based approaches are extending the field by answering previously intractable questions.
- These new techniques seem poised to teach us a great deal about the tree of life and molecular genetics.

References

- Holder M, Lewis P O. Phylogeny estimation: traditional and Bayesian approaches[J]. Nature reviews genetics, 2003, 4(4): 275-284.
- Aftab S, Semenec L, Chu J S C, et al. Identification and characterization of novel human tissue-specific RFX transcription factors[J]. BMC evolutionary biology, 2008, 8(1): 226.
- http://www.zoology.ubc.ca/~bio336/Bio336/Lectures/Lecture14/Overheads.html
- http://www.stanford.edu/~hwang41/
- Richard Durbin et al. 生物序列分析. 2010

Bioinformatics: Introduction and Methods

Computer Science Department, Southwest University

Thank you