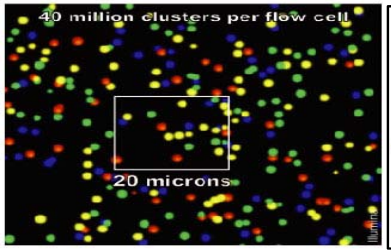
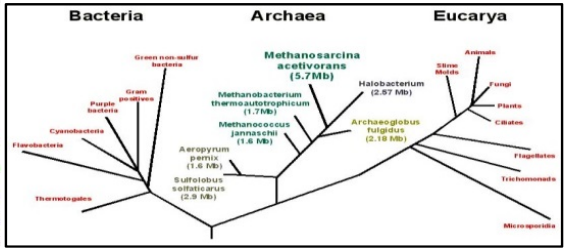
生物信息学：导论与方法
Bioinformatics: Introduction and Methods

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

TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA

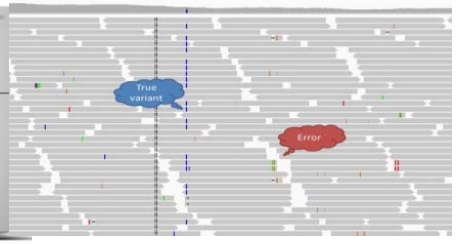


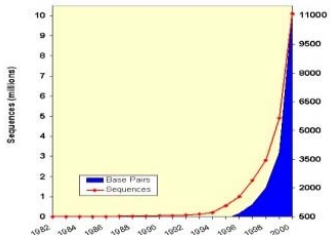


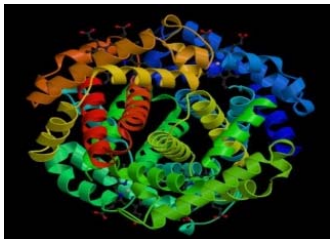
Case Study I: Origination of New Genes

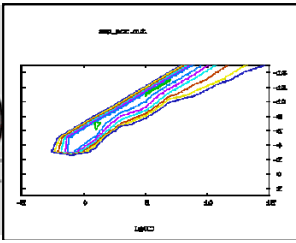
美国芝加哥大学生态与演化生物学系 龙漫远
北京大学生物信息学中心 魏丽萍

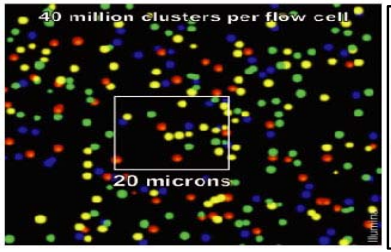
Manyuan Long, Department of Ecology and Evolutionary Biology, University of Chicago
Liping Wei, Center for Bioinformatics, Peking University

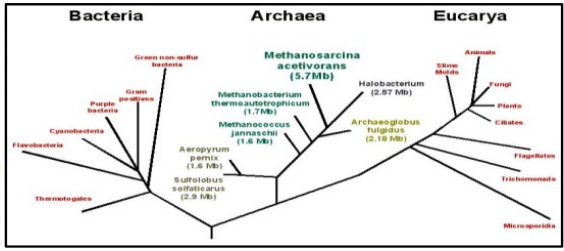










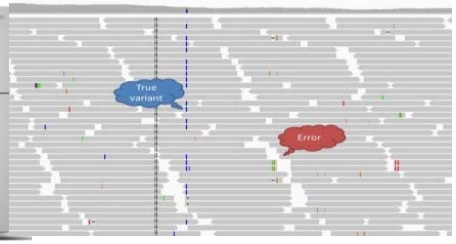


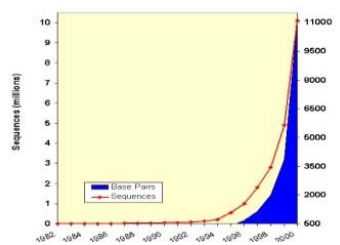
Unit 3:
A Human-Specific *de novo* Gene Associated with Addiction

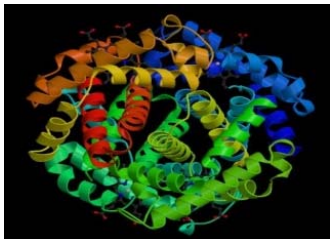
北京大学生物信息学中心 魏丽萍

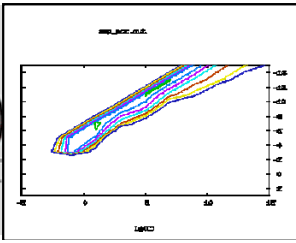
Liping Wei, Ph.D.

Center for Bioinformatics, Peking University

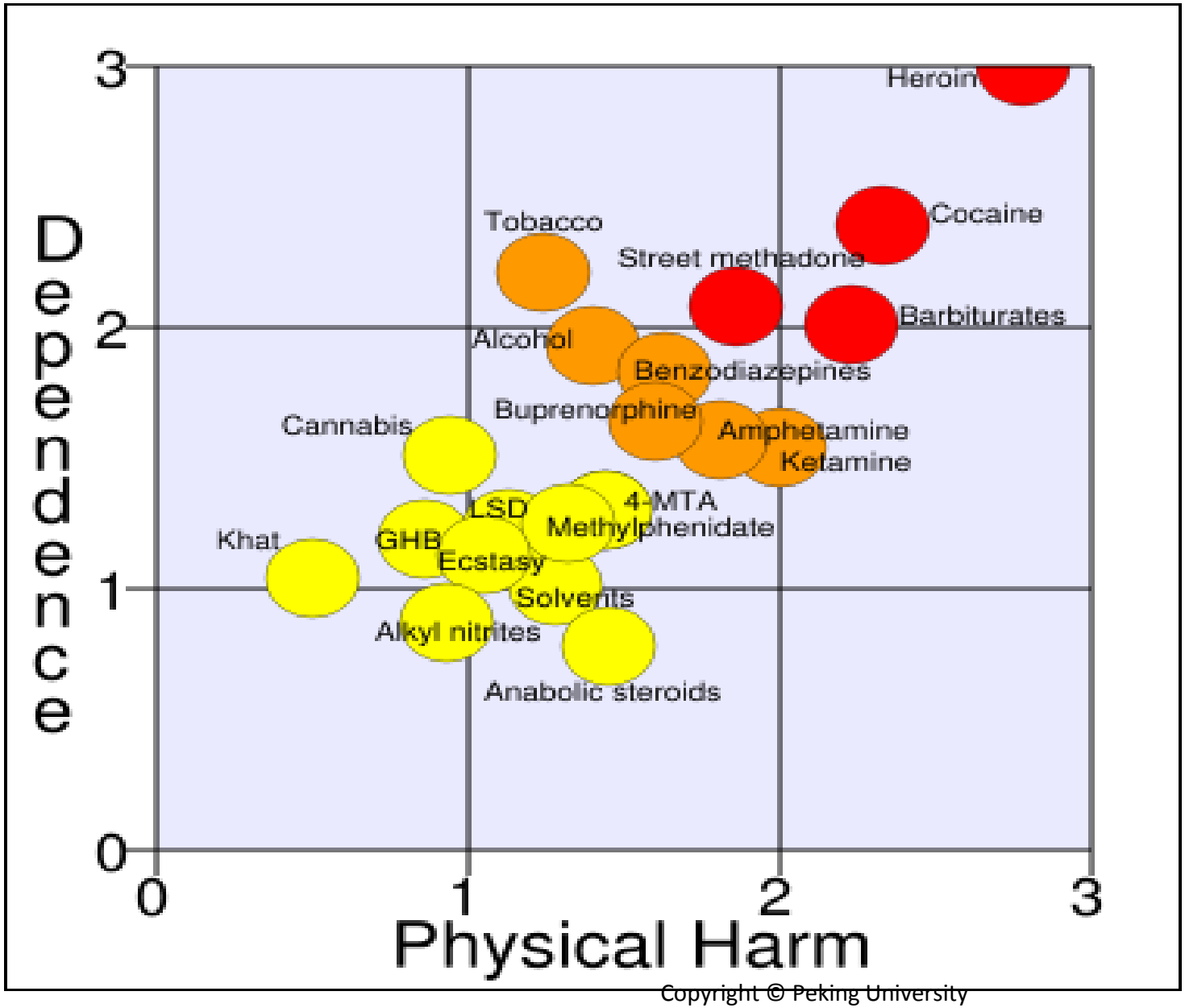






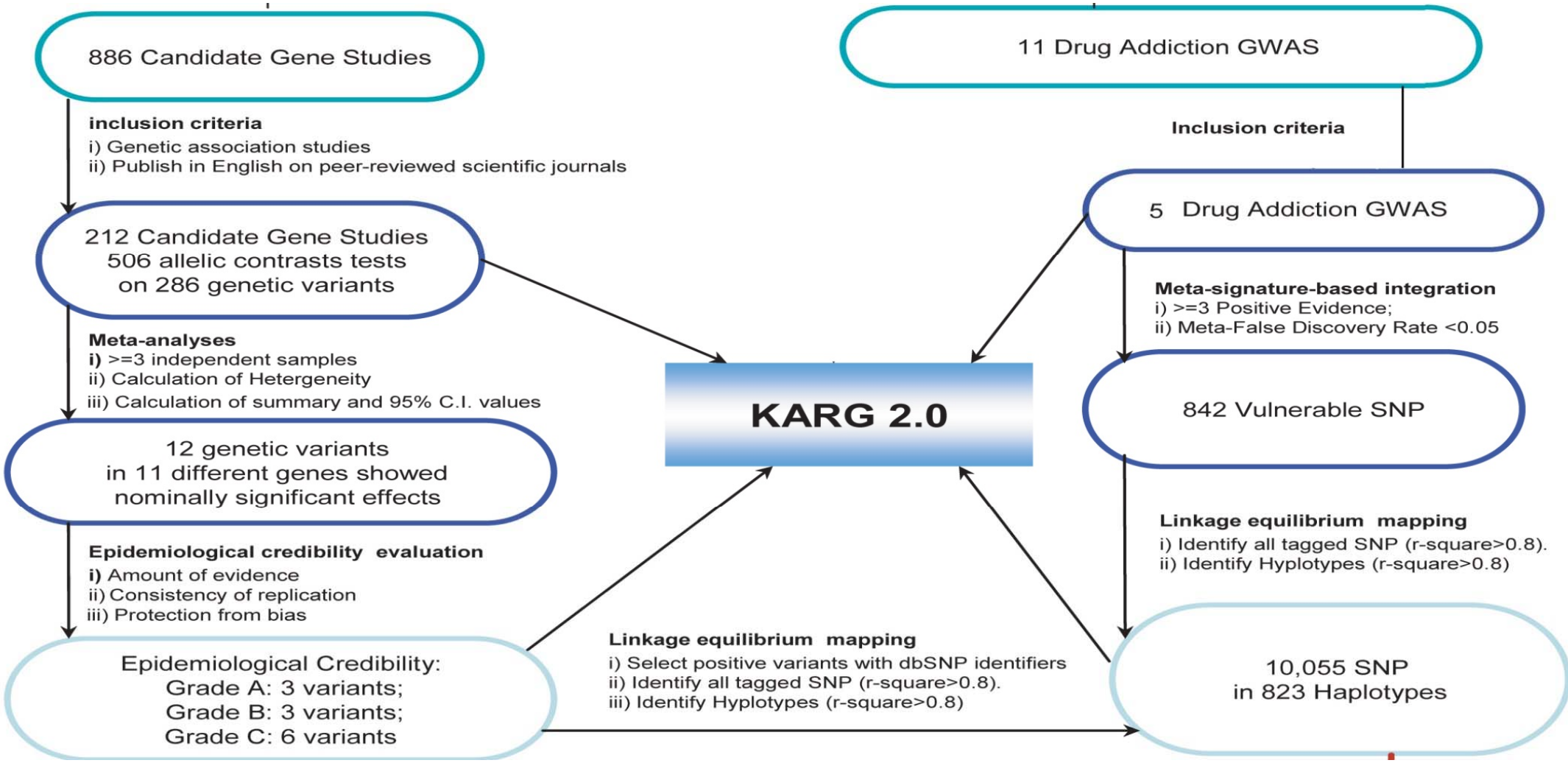


Addiction



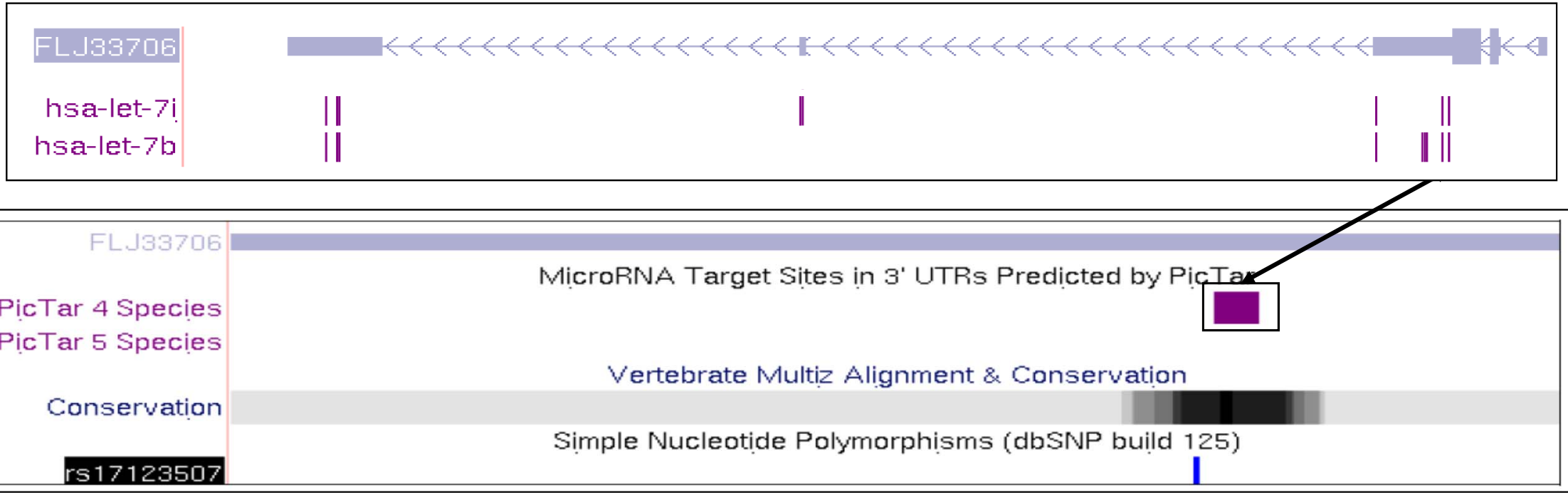
Nutt, *NEJM*, ’07

Collaboration with NIH/NIDA to analyze addiction GWAS data

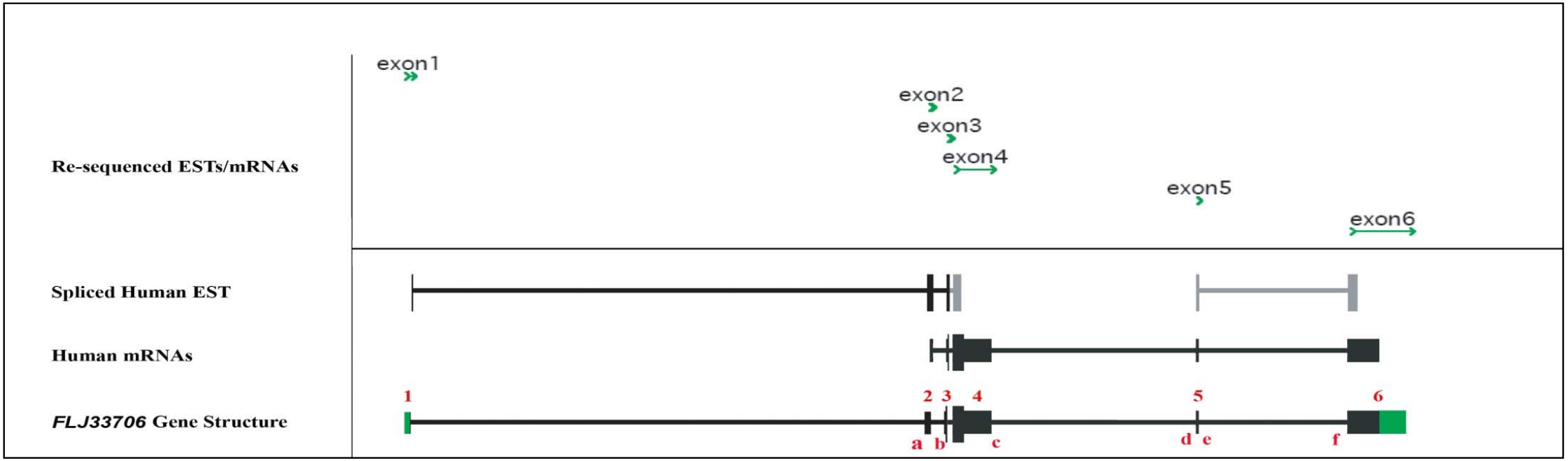


An SNP on the 3'UTR of *FLJ33706* is statistically significant in two GWAS of nicotine addiction and implicated in two linkage analyses.

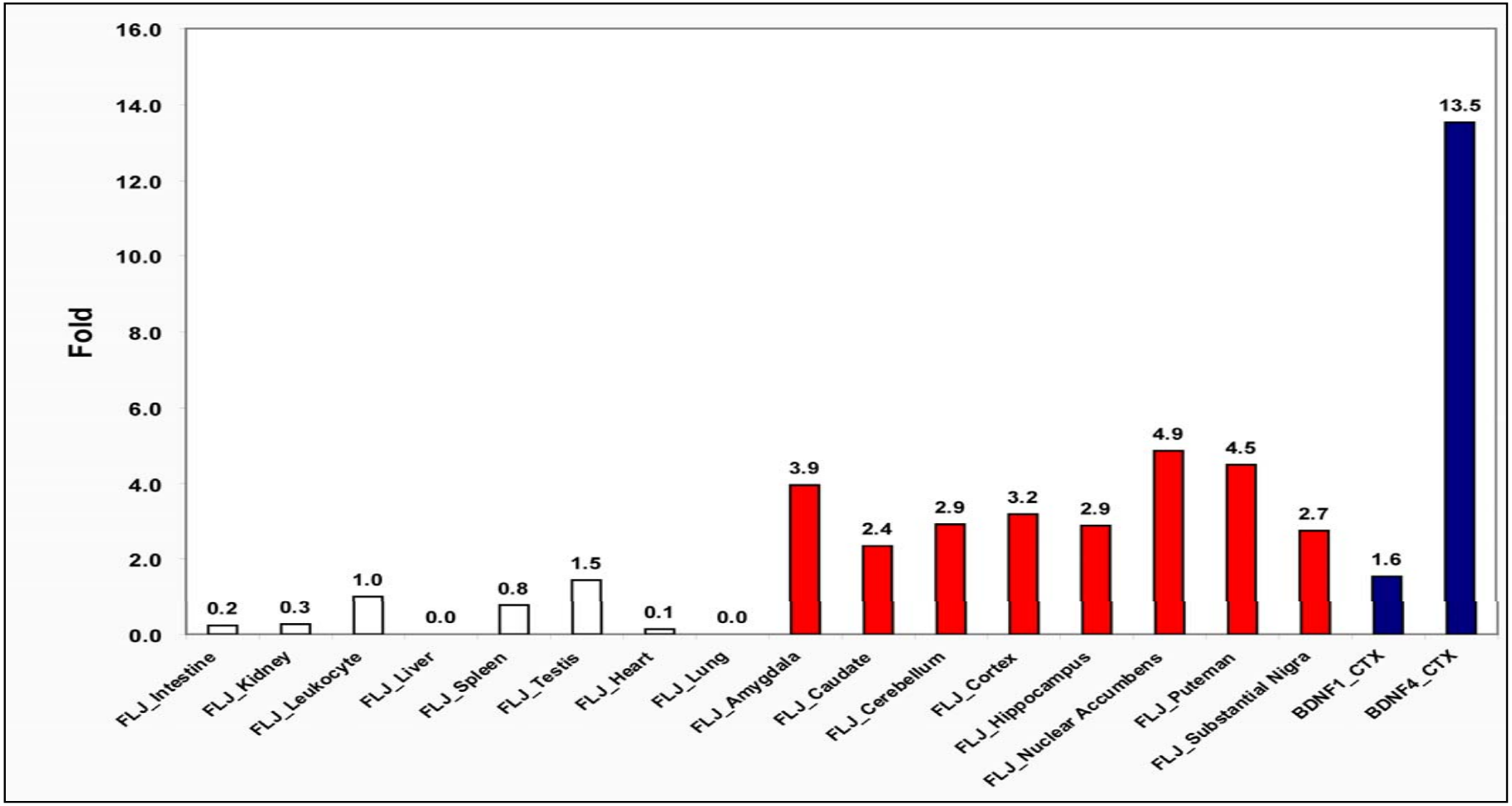
It is located in the middle of 12 binding sites of miRNA *let-7.*



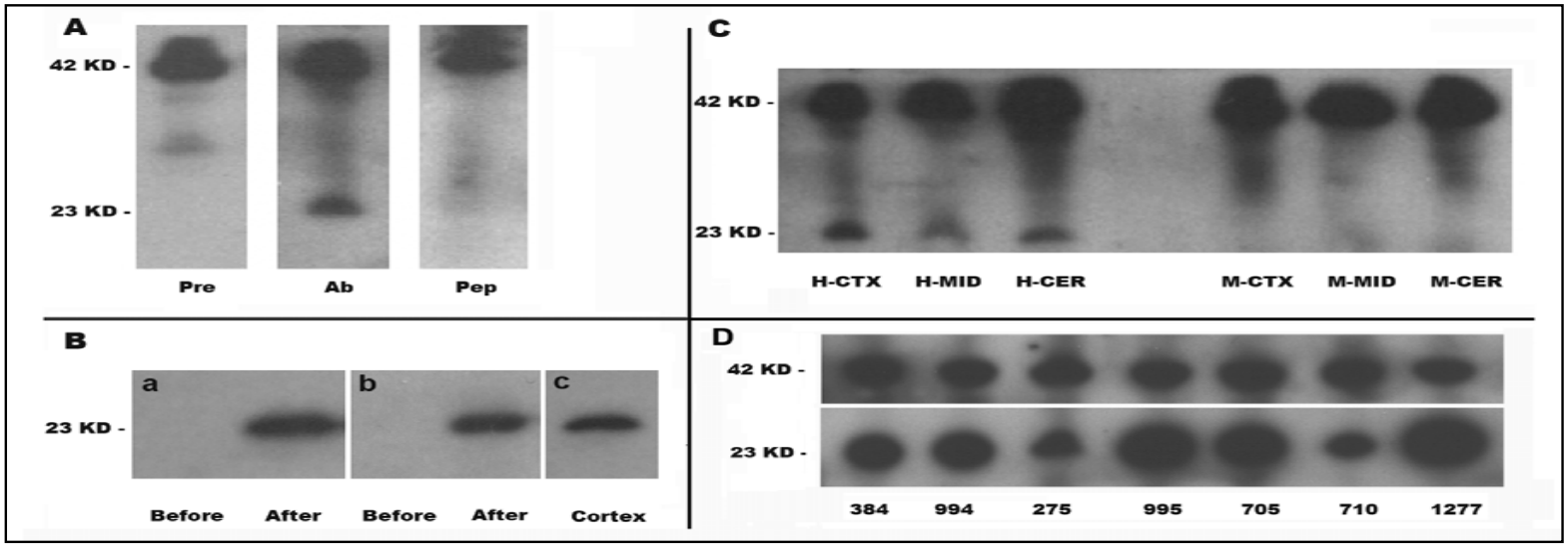
FLJ33706 is a human-specific *de novo* protein-coding gene



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



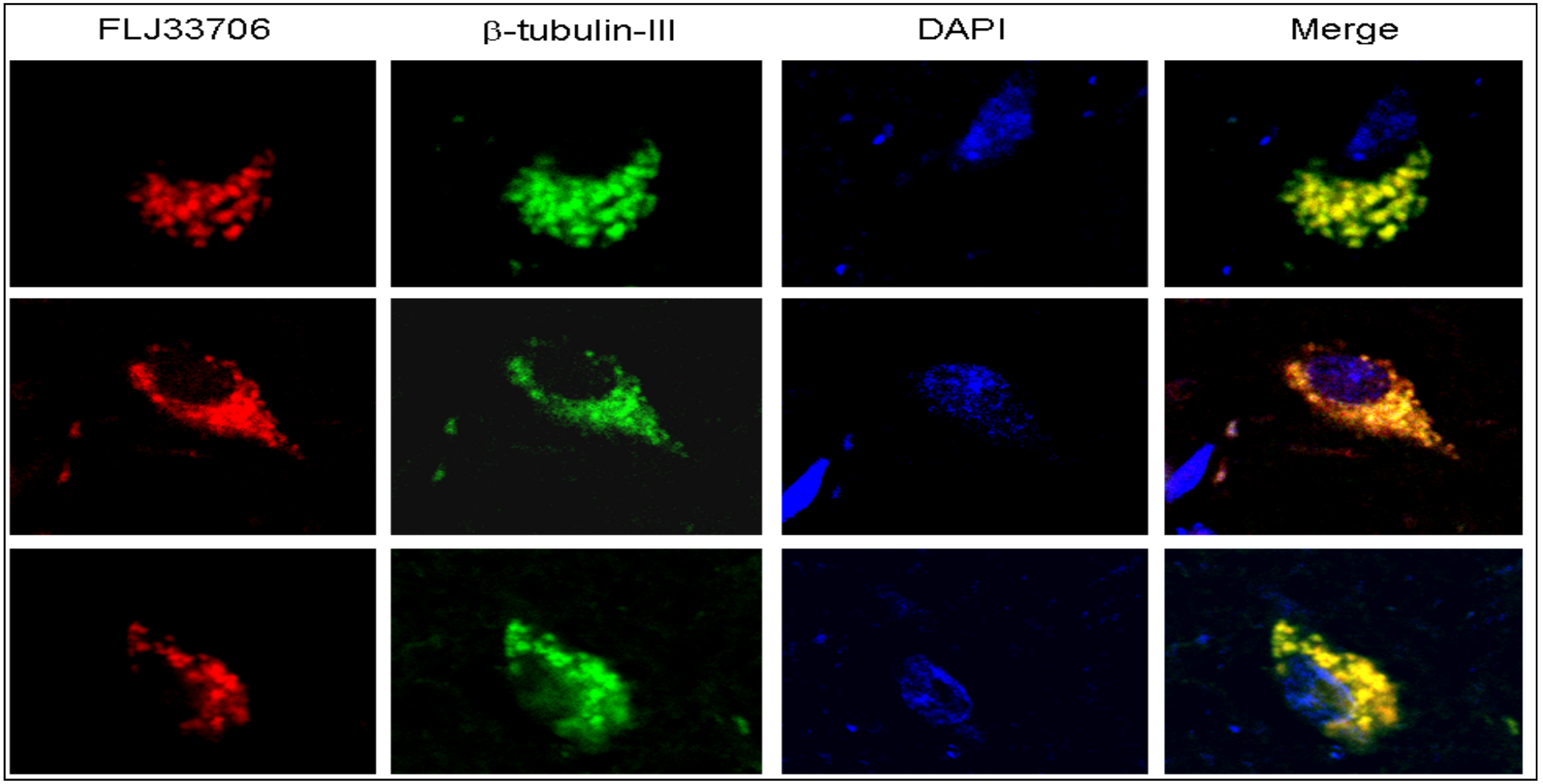
Western blot assay using an antibody designed against a 17-amino-acid peptide confirmed expression of FLJ33706 protein



- (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.

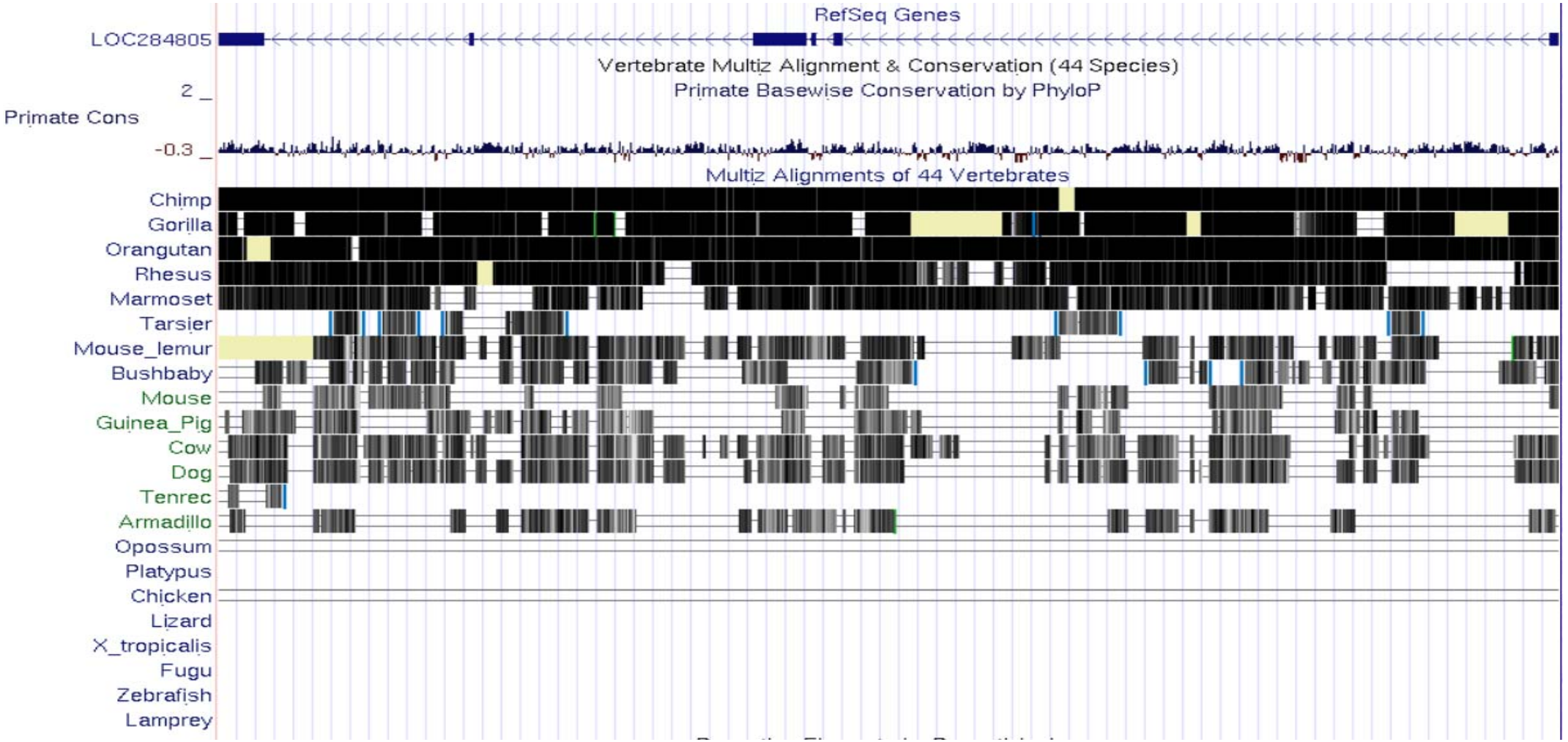
Li, *et al*, PLoS *Comp Biol*, '10

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



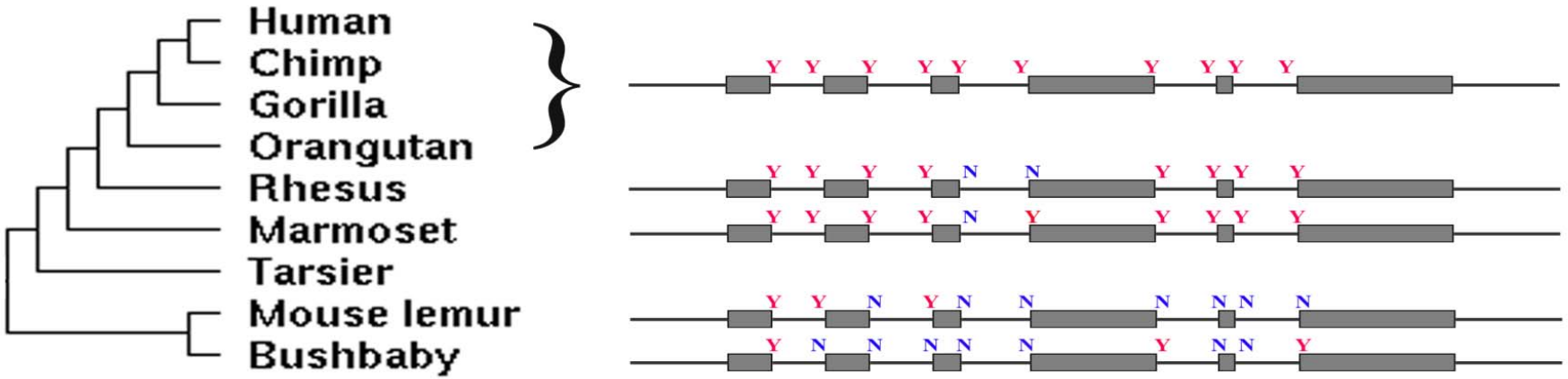
Li, *et al*, PLoS *Comp Biol*, '10

The DNA segment emerged in eutherian mammals



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

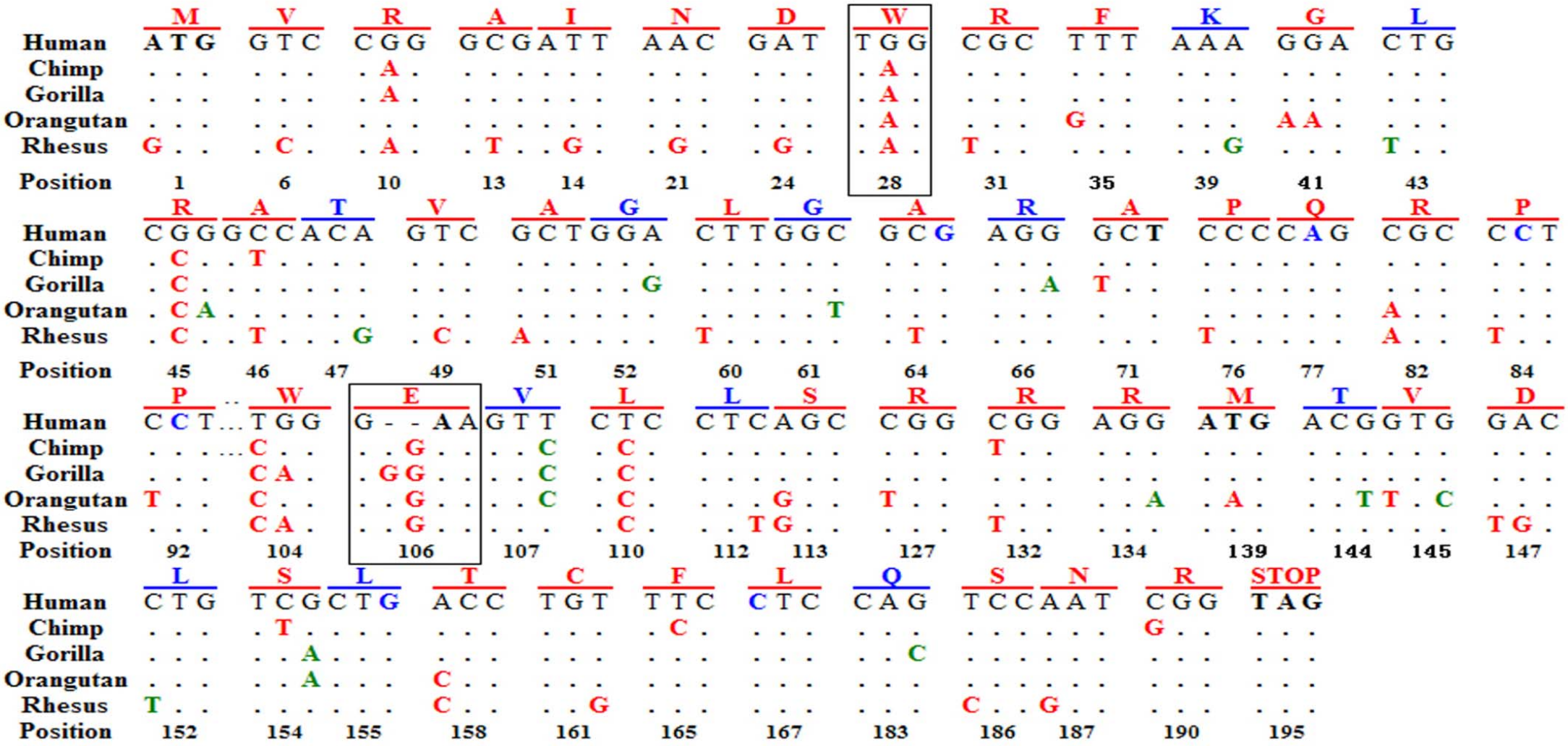
Insertion of *Alu* elements generated splicing sites



		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 agc…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 agc…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 agc…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 agc…===	= =	===…===	= = ===…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…---	- - ---…===	= =	===…===	= = ===…===	= =	===…

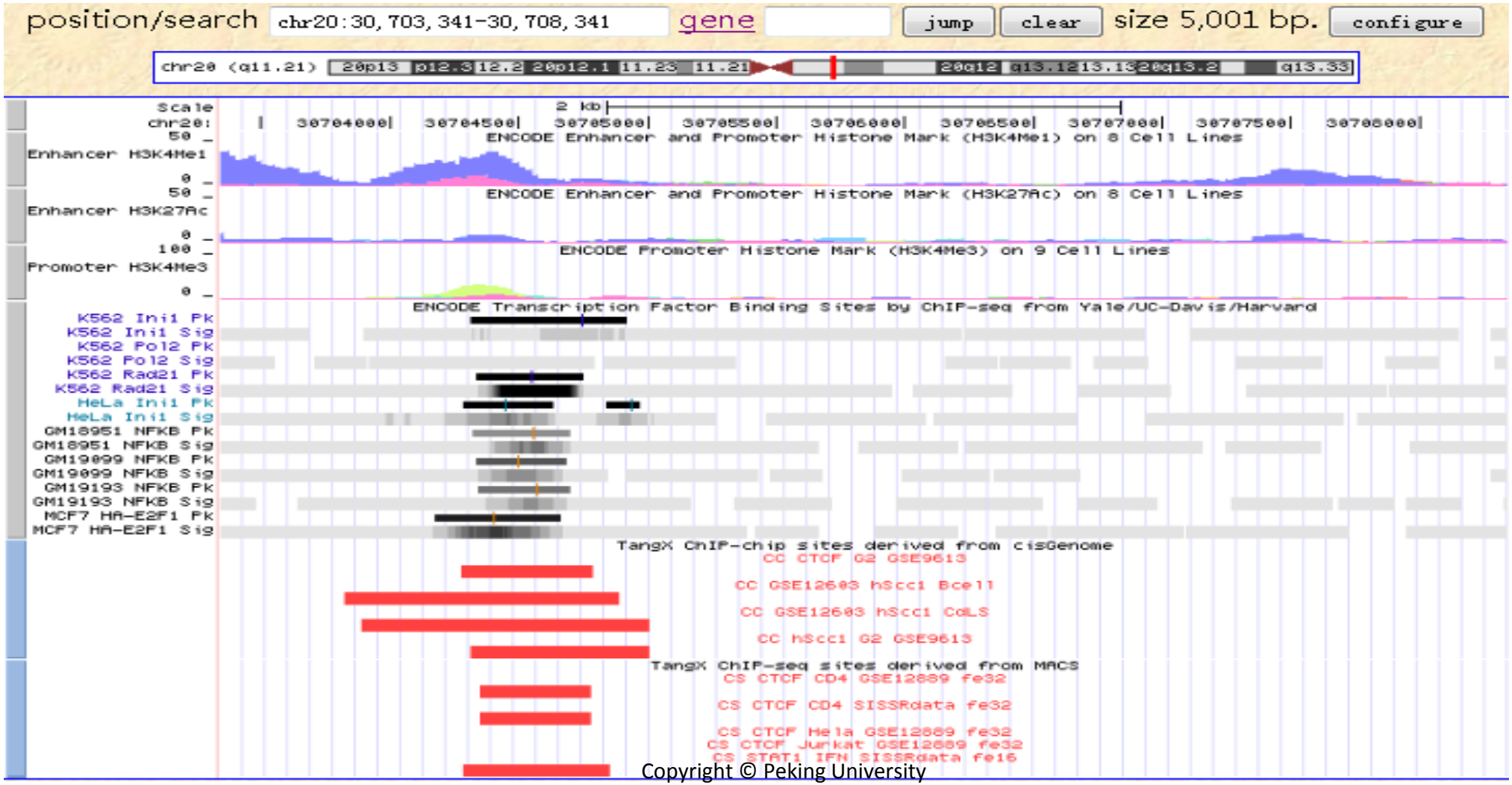
Li, *et al*, PLoS *Comp Biol*, '10

Two changes in human escaped two stop codons

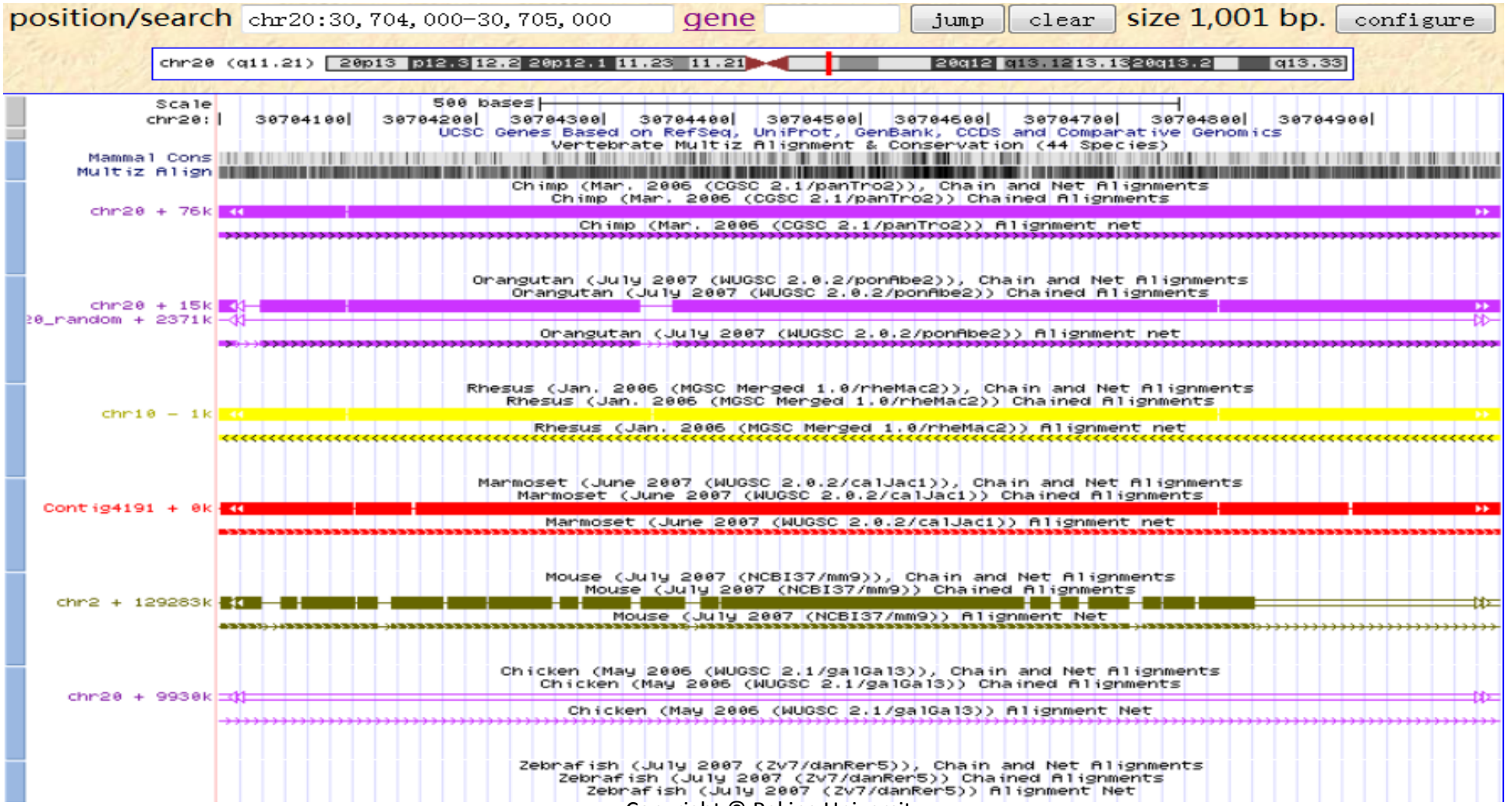


Li, *et al*, PLoS *Comp Biol*, '10

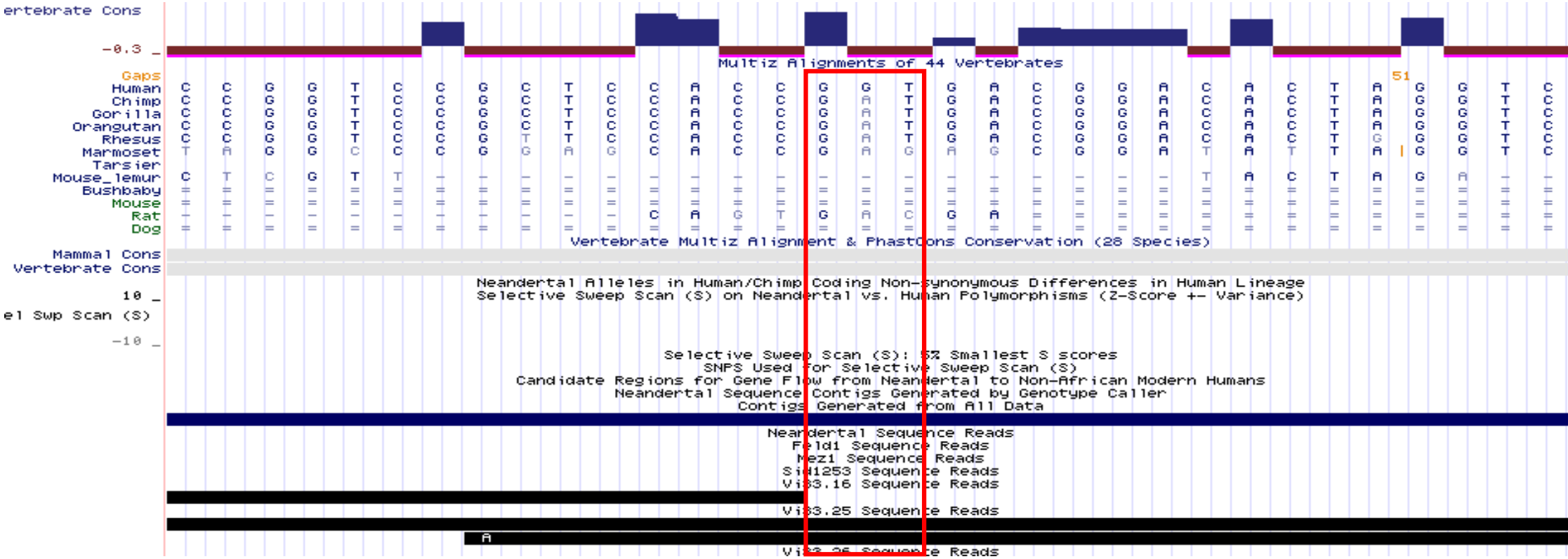
There are signals of enhancer and transcription factor binding sites in the 5kb upstream regions



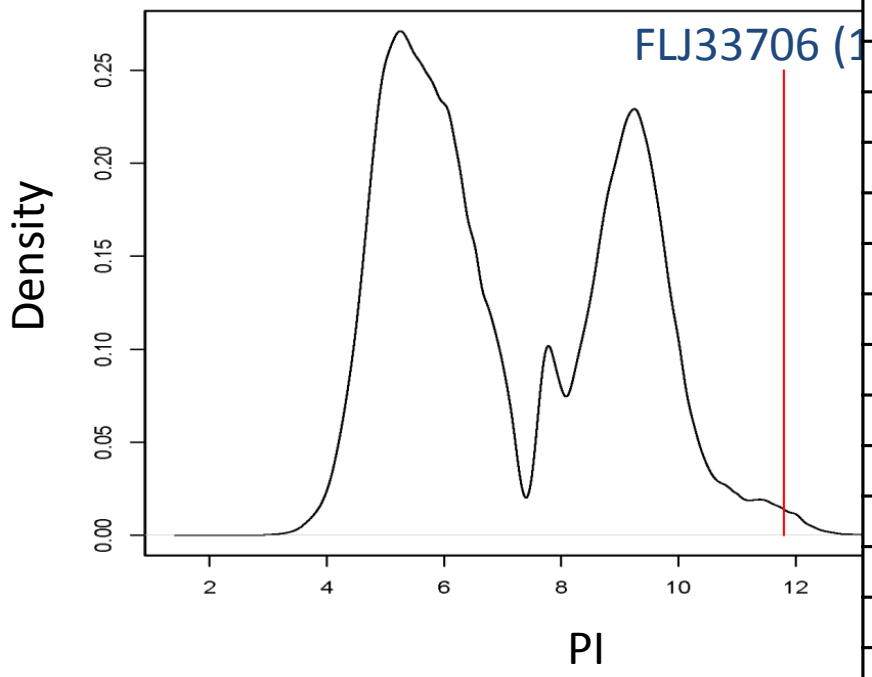
Promoter region is absent in chicken/zebrafish, emerged in mouse, and is similar in rhesus and chimpanzee



The Open Reading Frame is intact in Neadertal genome



FLJ33706 has high PI

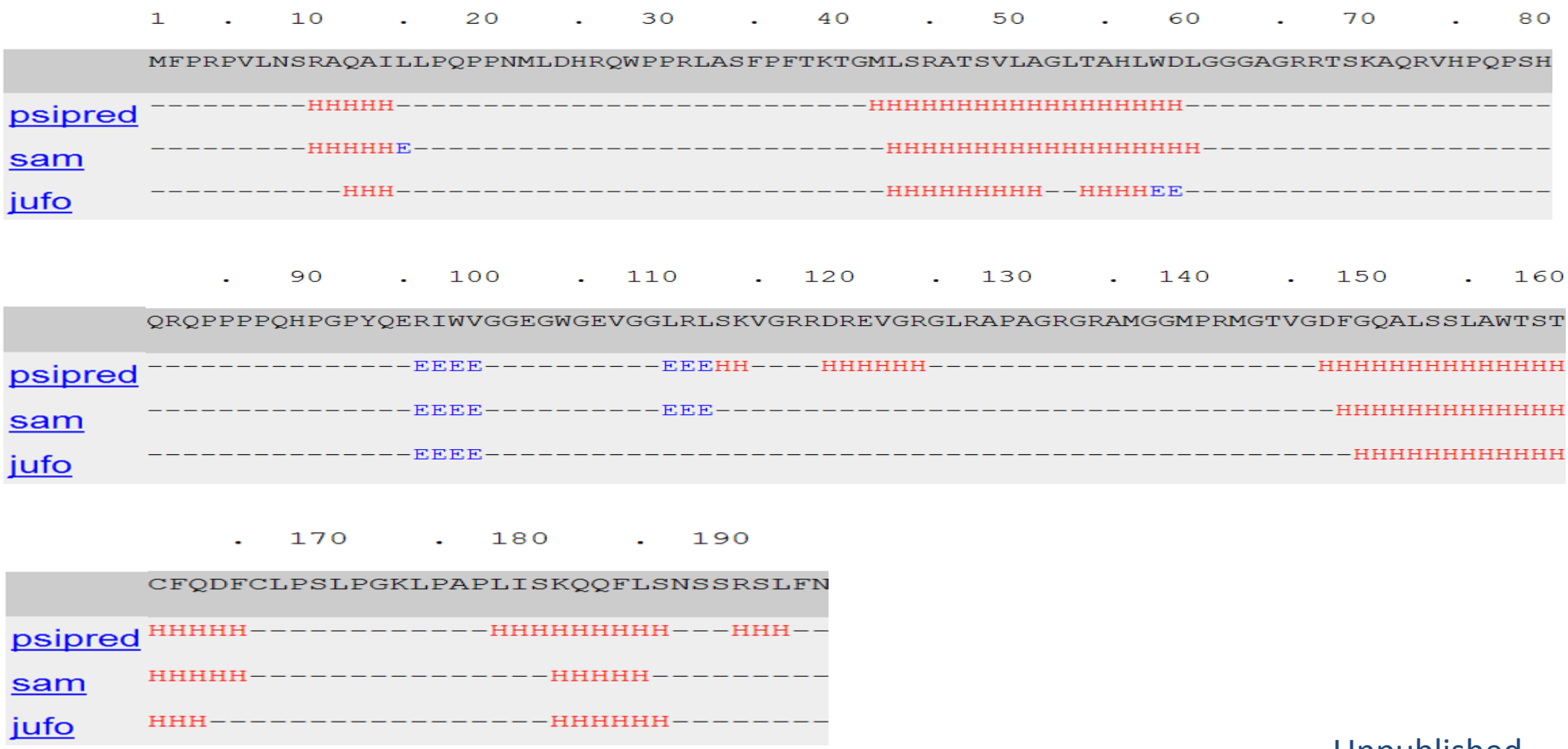


GO Term	FDR q-value
RNA binding	5.50E-08
cytosolic ribosome	3.68E-07
macromolecular complex	1.63E-06
cytosolic large ribosomal subunit	4.61E-05
RNA splicing	6.71E-05
cytosolic part	7.73E-05
ribosomal subunit	4.54E-04
large ribosomal subunit	7.89E-04
intracellular organelle part	9.99E-04
organelle part	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

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

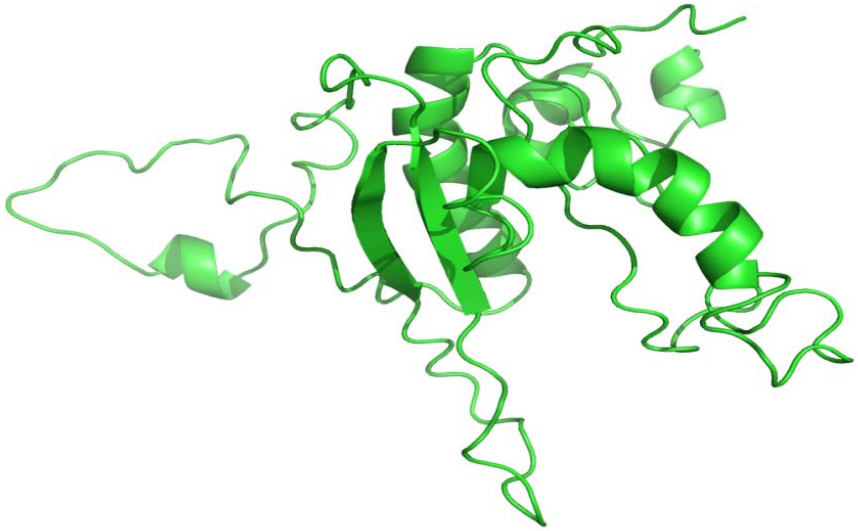
Unpublished

Predicted Secondary Structure include four helices & one beta strand

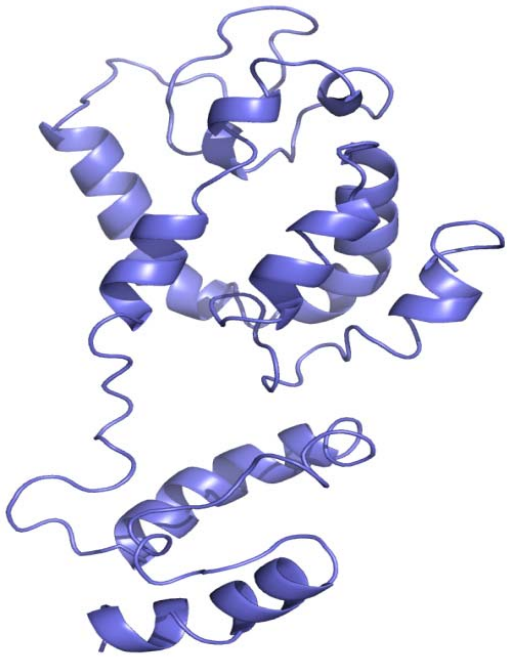


Unpublished

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?

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