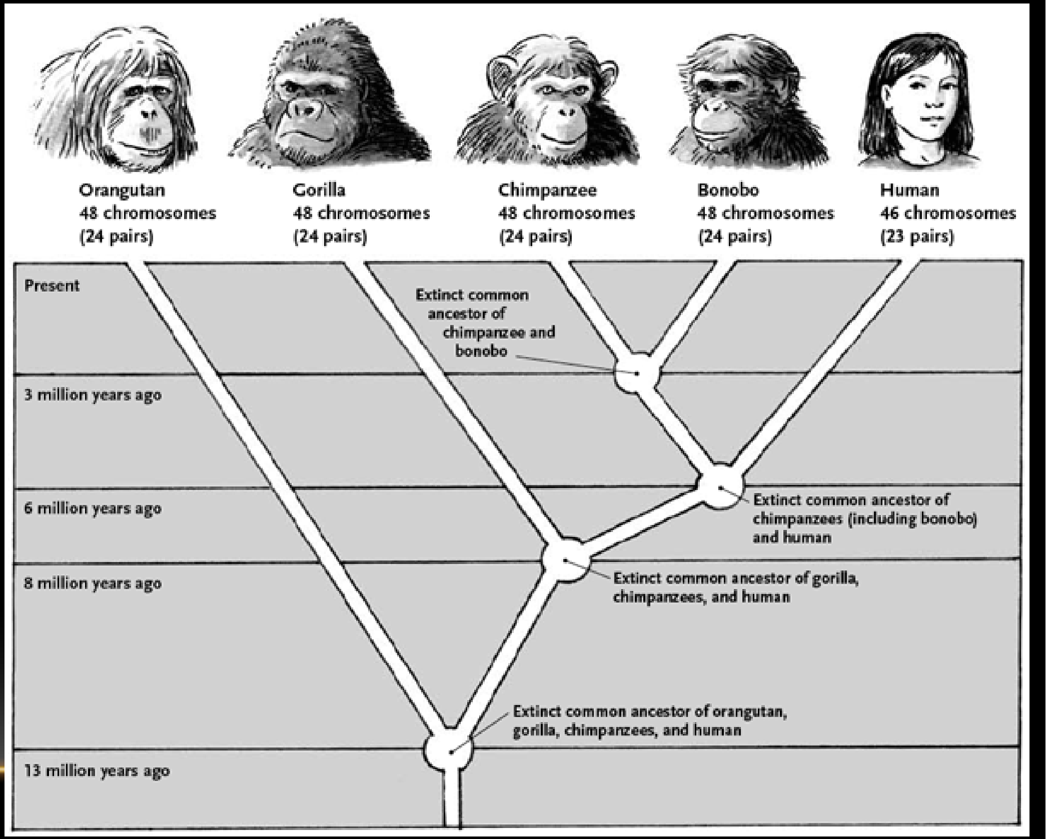
TOWARD DECIPHERING HOMINIDAE-SPECIFIC GENOMIC PROPERTIES

MORTEZA MAHMOUDISABER

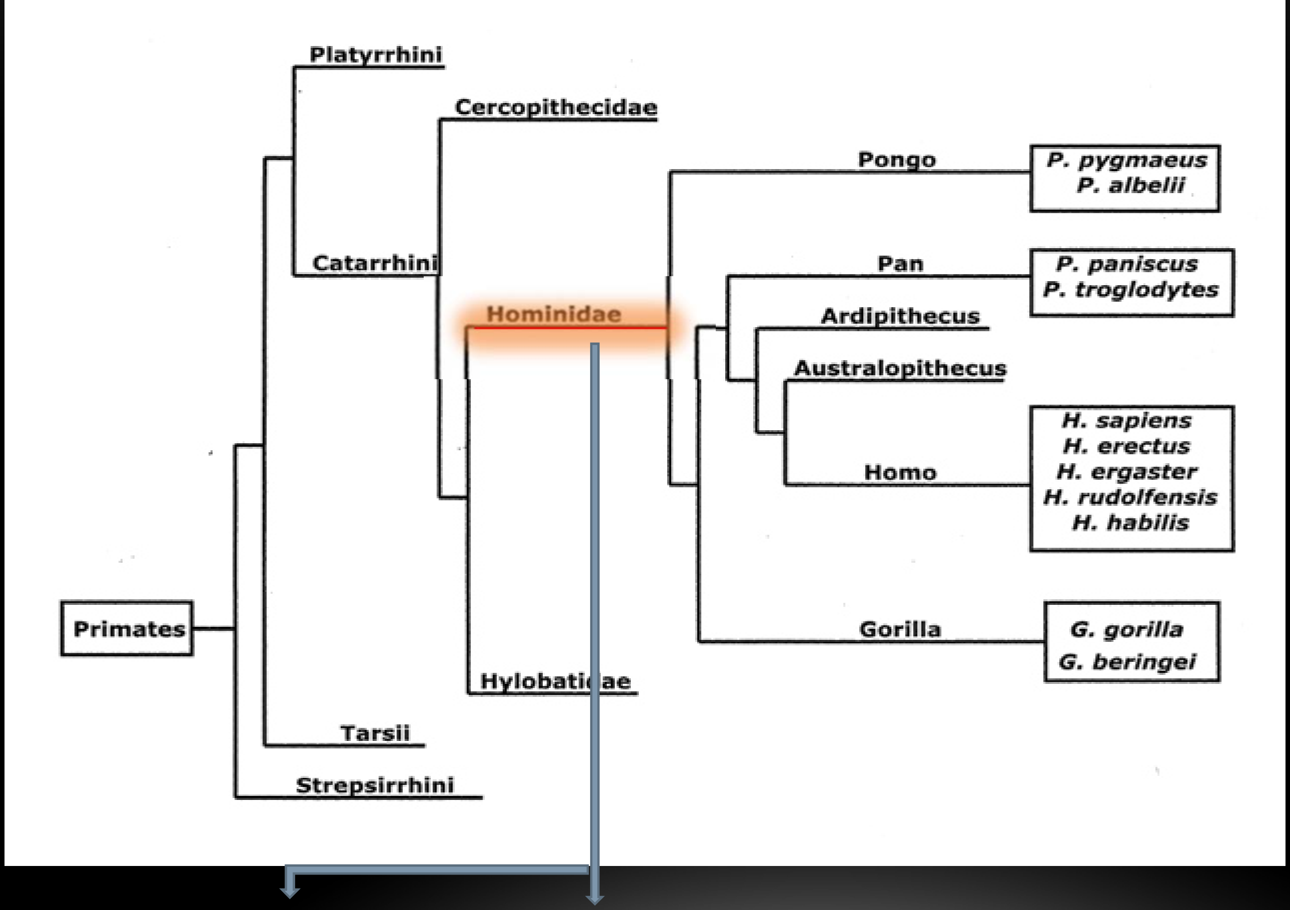
DEPARTMENT OF BIOLOGICAL SCIENCE
THE UNIVERSITY OF TOKYO
PROF. SAITOU NARUYA LABORATOY

HOMINIDAE FAMILY

- Hominidae family or Great Apes are large, tailless omnivore primates
- The hominids are probably the most advanced family in the world, well-known for the complexity of their social behavior and intellectual abilities.
- Even aside from human, other hominidae members can learn human sign language and also show innovation in solving problems



ORIGIN OF HOMINIDAE SPECIFIC PHENOTYPES

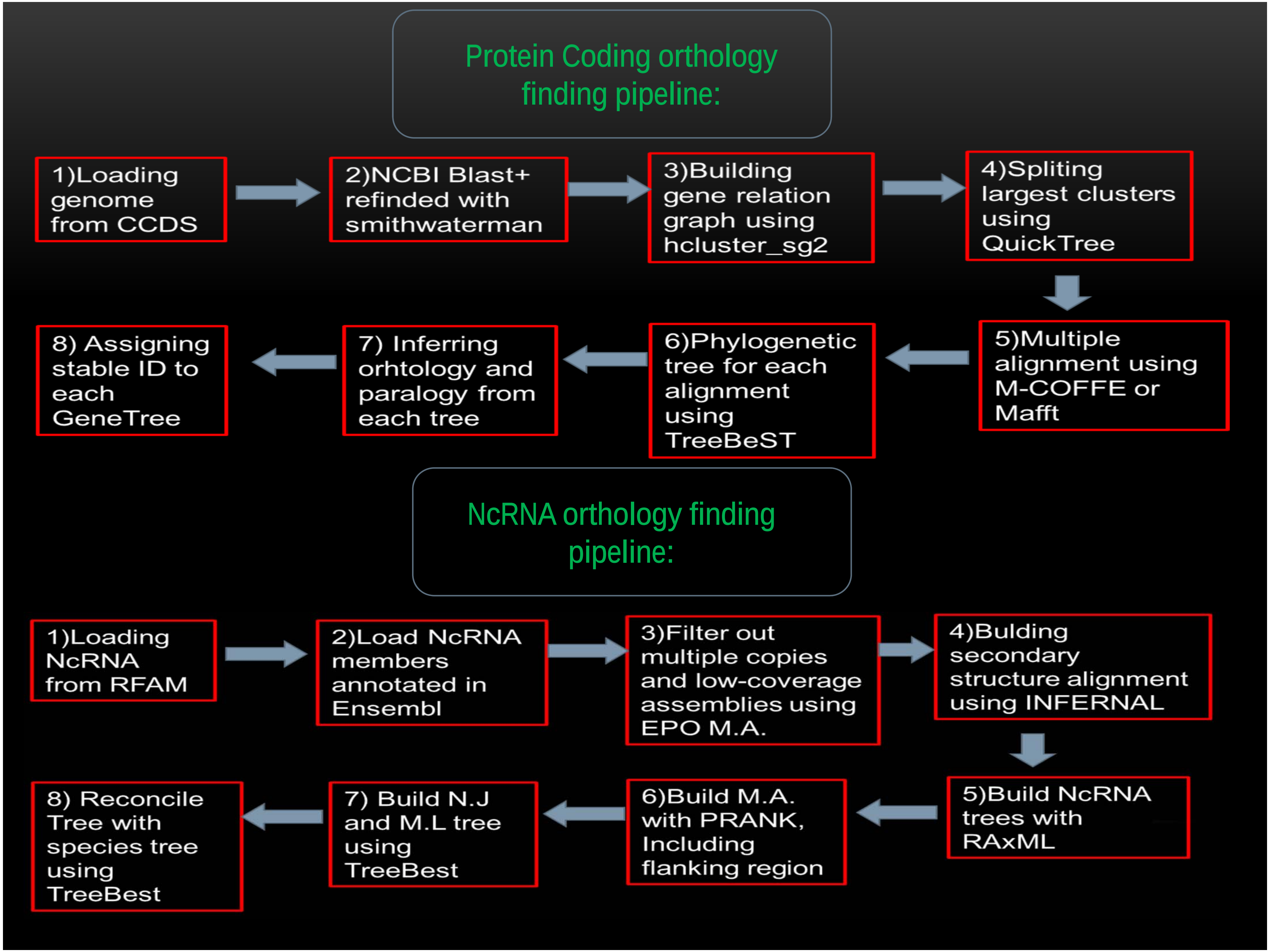


Advent of Novel Hominidae-specific genes

Hominidae-specific conserved non coding sequence

GENCODE.V19

Gene Biotype	Number	Gene Biotype	Number	Gene Biotype	Number
Protein_Coding	20345	3prime-overlapping-ncRNA	21	IG_J_gene	18
Sense_overlapping	202	Misc_RNA	2034	snRNA	1916
TR_C_gene	5	TR_D_gene	3	Pseudogene	13931
miRNA	3055	IG_V_gene	138	Processed_transcript	515
IG_C_gene	14	TR_V_gene	97	Polymorphic_pseudogene	45
rRNA	527	IG_D_gene	37	IG_C_psudogene	9
snoRNA	1457	Sense_intronic	742	IG_V_pseudogene	187
Mt-RNA	2	Mt_tRNA	22	TR_G_pseudogene	4
IG_J_Pseudogene	3	lincRNA	7114	TR_J_gene	74
TR_V_Pseudogene	27	Antisense	5276	*TOTAL*	*57820*



Protein Coding orthology finding pipeline:

1)Loading genome from CCDS

2)NCBI Blast+ refinded with smithwaterman

3)Building gene relation graph using hcluster_sg2

4)Spliting largest clusters using QuickTree

5)Multiple alignment using M-COFFE or Mafft

6)Phylogenetic tree for each alignment using TreeBeST

7) Inferring orhtology and paralogy from each tree

8) Assigning stable ID to each GeneTree

NcRNA orthology finding pipeline:

1)Loading NcRNA from RFAM

2)Load NcRNA members annotated in Ensembl

3)Filter out multiple copies and low-coverage assemblies using EPO M.A.

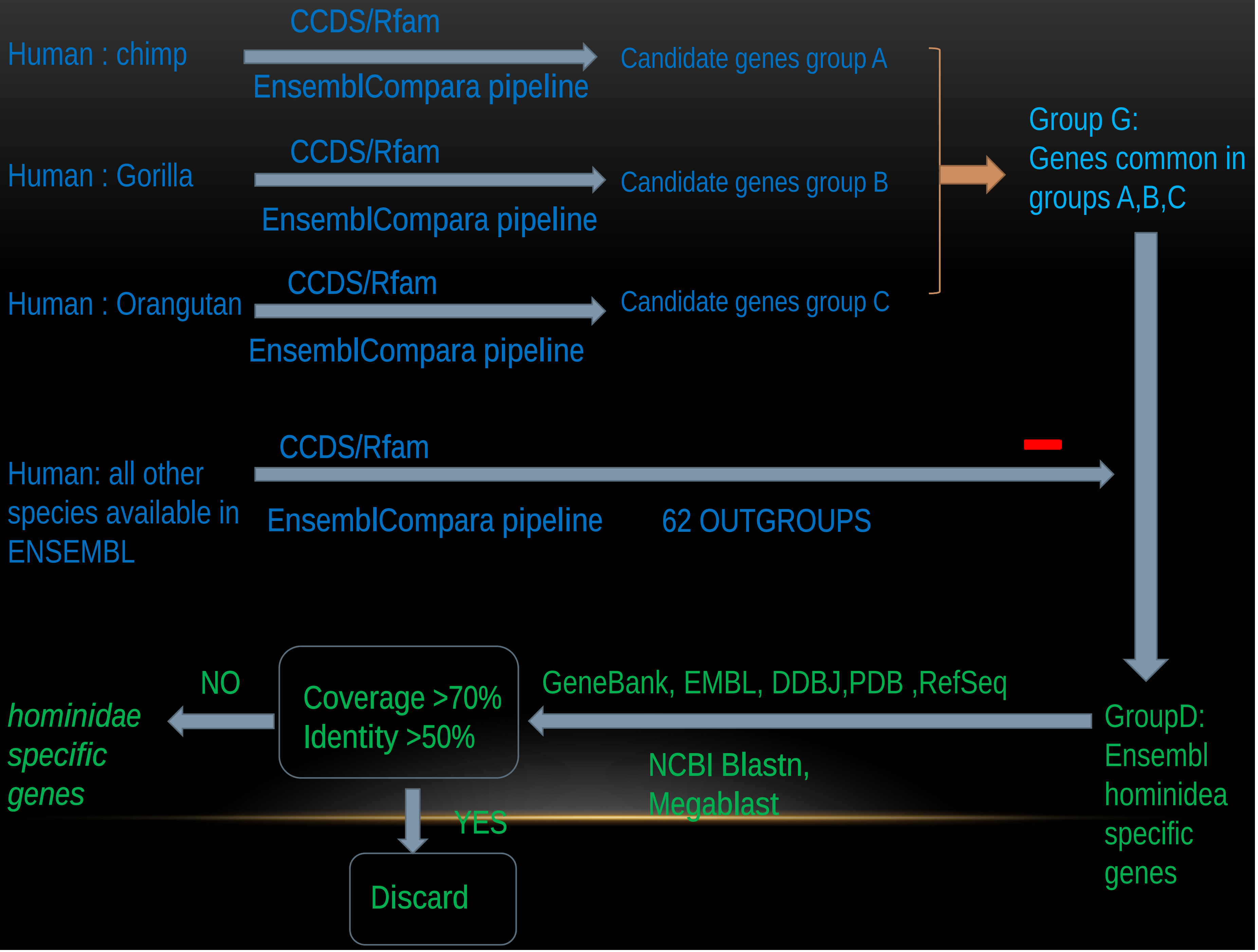
4)Bulding secondary structure alignment using INFERNAL

5)Build NcRNA trees with RAxML

6)Build M.A. with PRANK, Including flanking region

7) Build N.J and M.L tree using TreeBest

8) Reconcile Tree with species tree using TreeBest



Human : chimp

CCDS/Rfam

EnsemblCompara pipeline

Candidate genes group A

Human : Gorilla

CCDS/Rfam

EnsemblCompara pipeline

Candidate genes group B

Human : Orangutan

CCDS/Rfam

EnsemblCompara pipeline

Candidate genes group C

Group G:
Genes common in groups A,B,C

Human: all other species available in ENSEMBL

CCDS/Rfam

EnsemblCompara pipeline

62 OUTGROUPS

−

GroupD:
Ensembl hominidea specific genes

GeneBank, EMBL, DDBJ,PDB ,RefSeq

NCBI Blastn, Megablast

Coverage >70%
Identity >50%

NO

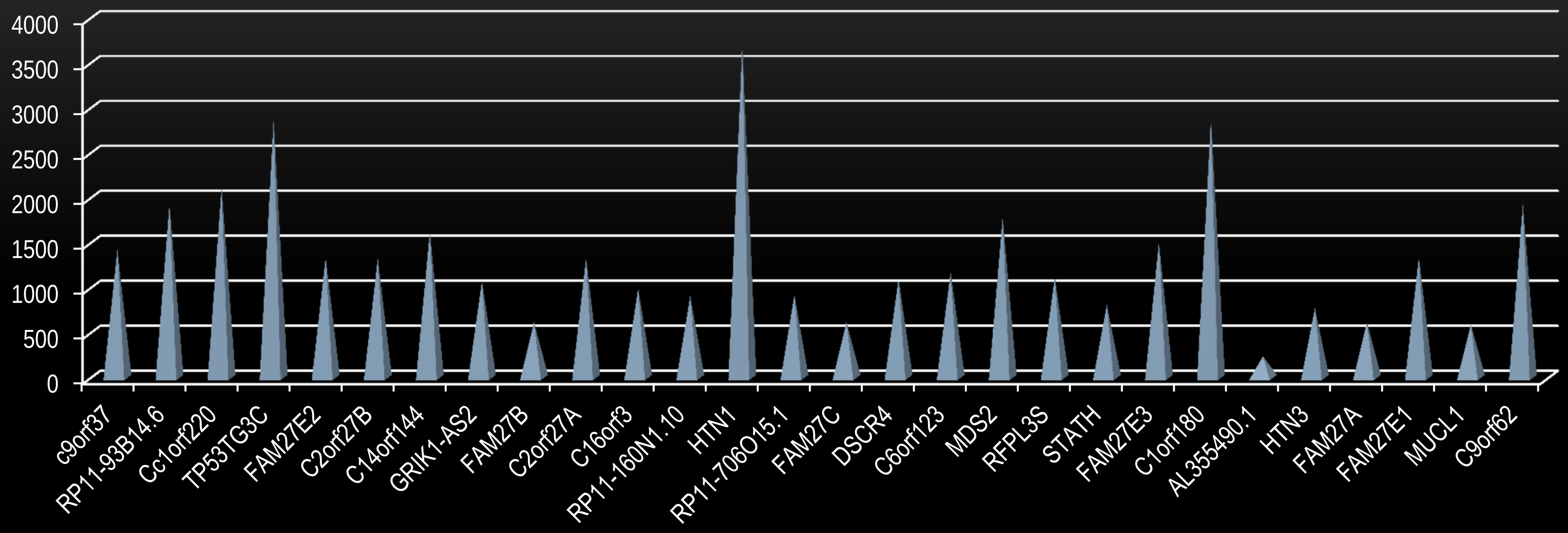
hominidae specific genes

YES

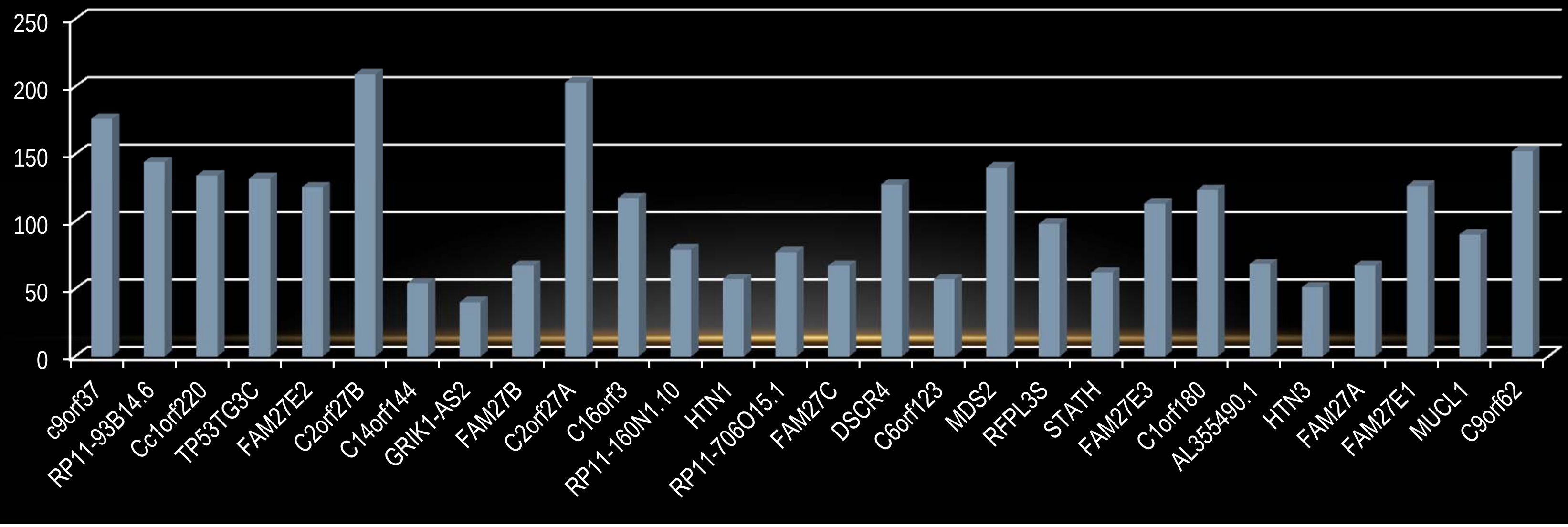
Discard

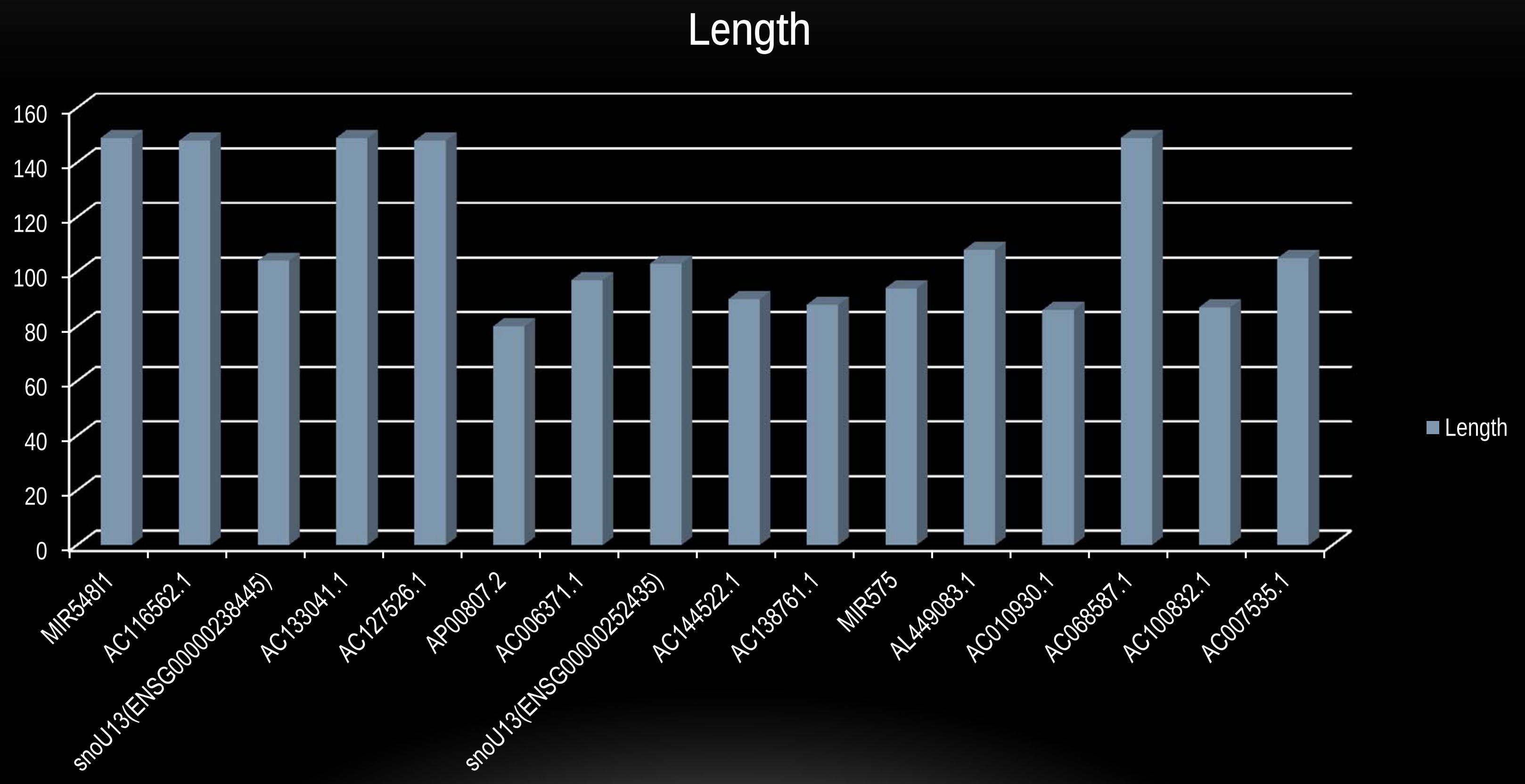
GENE BIOTYPE	NUMBER	GENE NAMES
Protein_coding	27	C9orf37,RP11-93B14.6, C1orf220, TP53TG3C, FAM27E2, C2orf27B,C14orf144,GRIK1-AS2, FAM27B, C2orf27A, C16orf3, RP11-160N1.10, HTN1, FAM27C, DSCR4, C6orf123, MDS2, RFPL3S, STATH, FAM27E3, C1orf180, AL355490.1, HTN3,FAM27A,FAM27E1,MUCL1,C9orf62
miRNA	14	MIR548I1,AC116562.1,AC133041.1,AC127526.1,AP00807.2,AC006371.1,AC144522.1,AC138761.1,MIR575,AL449083.1,AC010930.1,AC068587.1,AC100832.1,AC007535.1
SnoRNA	2	snoU13(ENSG00000238445), snoU13(ENSG00000252435)

Longest Transcript



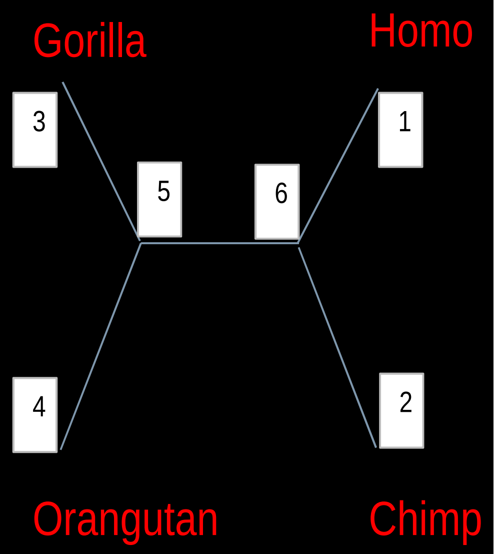
Longest Protein



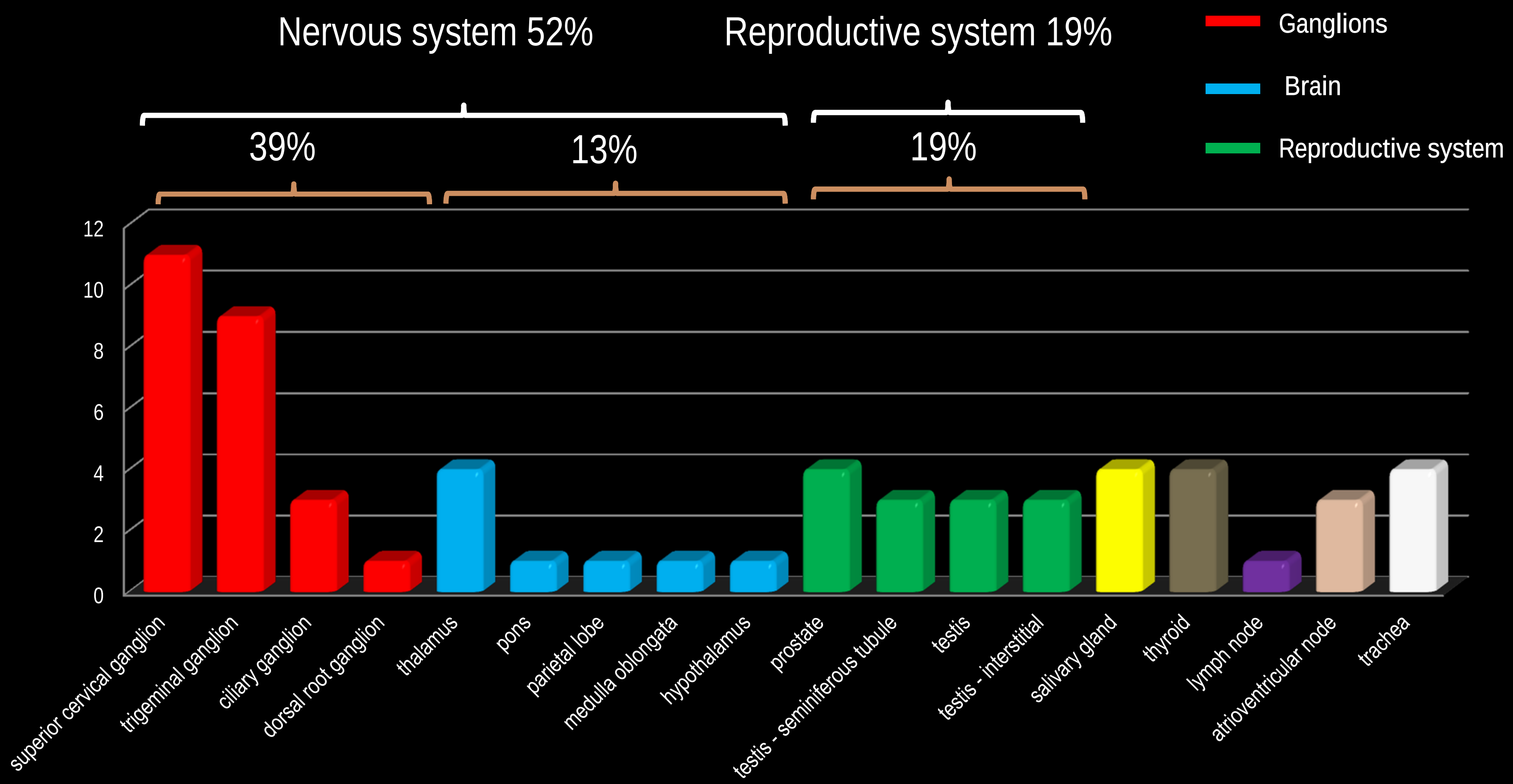

Length
160
140
120
100
80
60
40
20
0
MIR548I1
AC116562.1
snoU13(ENSG00000238445)
AC133041.1
AC127526.1
AP00807.2
AC006371.1
snoU13(ENSG00000252435)
AC144522.1
AC138761.1
MIR575
AL449083.1
AC010930.1
AC068587.1
AC100832.1
AC007535.1
■ Length

DN/DS RATIO FOR HOMINIDAE SPECIFIC GENES

Gene Name	5..3	5..6	6..2	6..1	5..4	homo_pan	homo_gorilla	homo_pongo
STATH	0.0001	NA(dn=0)	NA(dn=0.0	NA(dn=0)	N.D	NA(dn=0)	0	NA(dn=0)
HTN1	NA(dn=0)	NA(dn=0.0	NA(dn=0)	0.45	N.D	0.5	0.78	0.51
C6ORF123	NA(dn=0)	NA(dn=0)	0.45	NA(dn=0.1	N.D	1	NA(dn=0)	0.5
FAM27C	NA(dn=0.0	NA(dn=0.0	4.8	0.38	N.D	1.38	0.56	1.9
RP1193B14.6	1.02	0.0001	0.774	1.173	N.D	0.76	1.8	1.3
FAM27B	NA(dn=0.0	NA(dn=0.0	4.8	0.79	N.D	1.37	0.56	1.9
MUCL1	0.0001	NA(dn=0.1	0.0001	0.49	N.D	0.37	0.27	2.16
C9orf62	1.2	NA(dn=0.0	0.403	0.537	N.D	1	1.8	4
C1orf180	0.547	0.0001	0.548	NA(dn=0.0	N.D	0.58	0.58	0.45
FAM27A	NA(dn=0)	NA(dn=0)	4.8	0.79	N.D	1.37	0.56	1.9
GRIK1-AS2	NA(dn=0.0	NA(dn=0)	0.93	4.46	N.D	NA(dn=0)	NA(dn=0.014)	NA(dn=0.027)
DSCR4	NA(dn=0.0	0.586	0.0001	0.0001	N.D	0	NA(dn=0.02)	0.91
C1ORF220	0.282	NA(dn=0.0	0.0001	NA(dn=0.0	N.D	0.14	0.38	1.62
C2orf27B	0.537	4.306	0.734	NA(dn=0.0	N.D	3.4	1	2.57
MDS2	NA(dn=0.0	0.76	0.35	1.09	N.D	0.35	1	0.75
C2orf27A	0.668	NA(dn=0)	1.06	NA(dn=0.0	N.D	3.4	1	0.38
C9orf37	3.12	0.44	NA(dn=0.0	1.45	N.D	2.57	1.02	0.63
FAM27E2	1.32	NS(dn=0.0	0.86	2.17	N.D	0.89	1.24	1.1
RP11-160N1	NA(dn=0)	0.18	NA(dn=0.0	NA(dn=0.0	N.D	NA(dn=0.024	0.5	0.48
TP53TG3C	0.573	0.986	0.189	0.0001	N.D	0.09	0.24	1.1
HTN3	0.456	NA(dn=0.0	0.0001	NA(dn=0.0	N.D	NA(dn=0.008	NA(dn=0.017)	1.7
RP11-706O1	0.0001	0.331	NA(dn=0.0	0.24	N.D	1.08	0.79	1.1
RFPL3S	0.361	NA(dn=0)	1.12	1.56	N.D	NA(dn=0.031	0.66	0.8
C16orf3	NA(dn=0.0	NA(dn=0)	NA(dn=0.0	NA(dn=0.0	N.D	NA(dn=0.01)	NA(dn=0.015)	0.86
AL355490.1	0.83	0.0001	NA(dn=0.0	NA(dn=0)	N.D	NA(dn=0)	0.89	0.38
C14orf144	NA(dn=0.0	NA(dn=0.0	NA(dn=0.0	0.0001	N.D	0	0.7	0.46
FAM27E3	1.25	NA(dn=0.0	0.75	1.01	N.D	0.74	0.76	0.87
FAM27E1	1.327	NA(dn=0.0	1.09	1.38	N.D	0.8	0.98	1.04



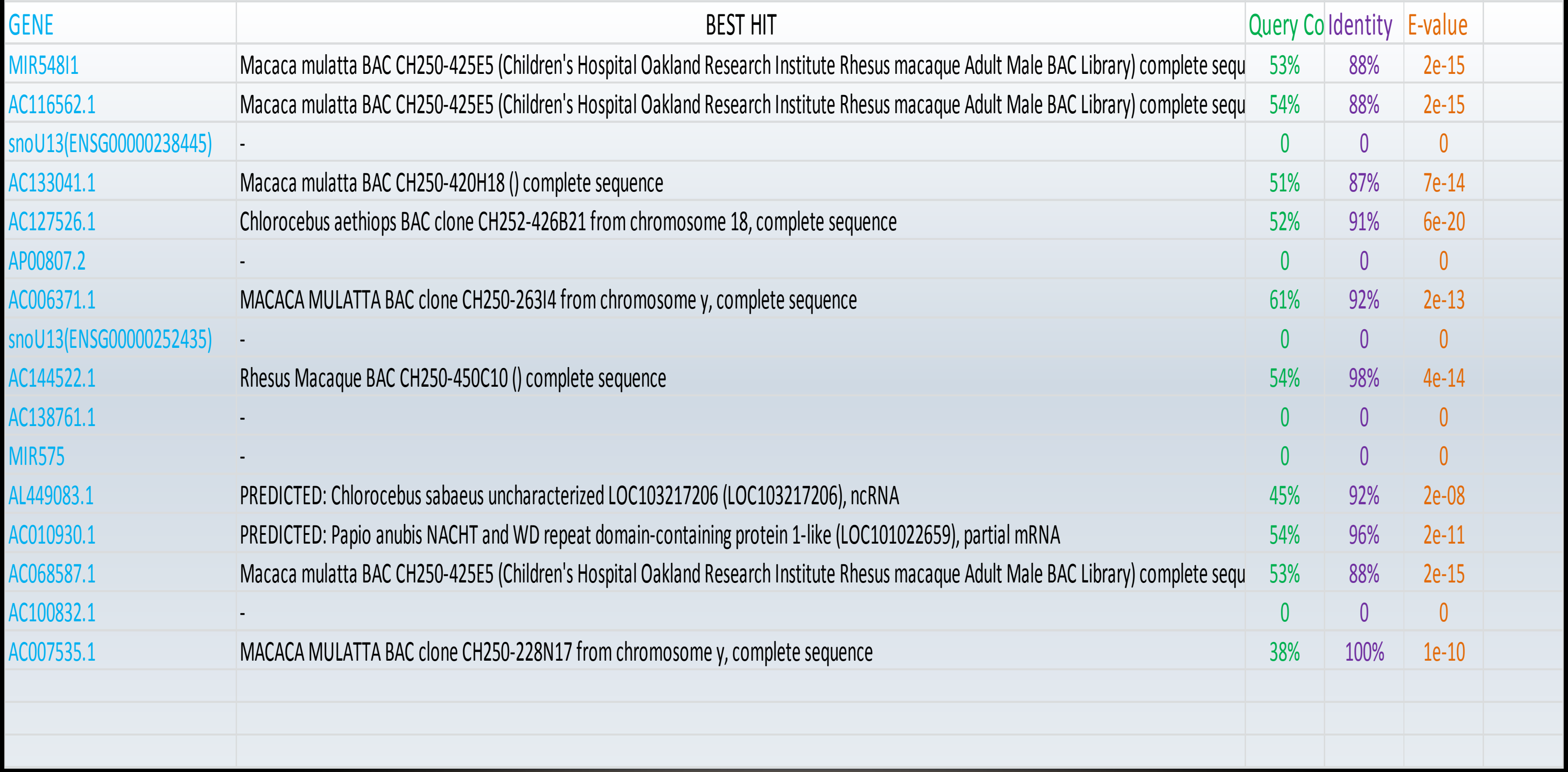
GNF EXPRESSION ATLAS



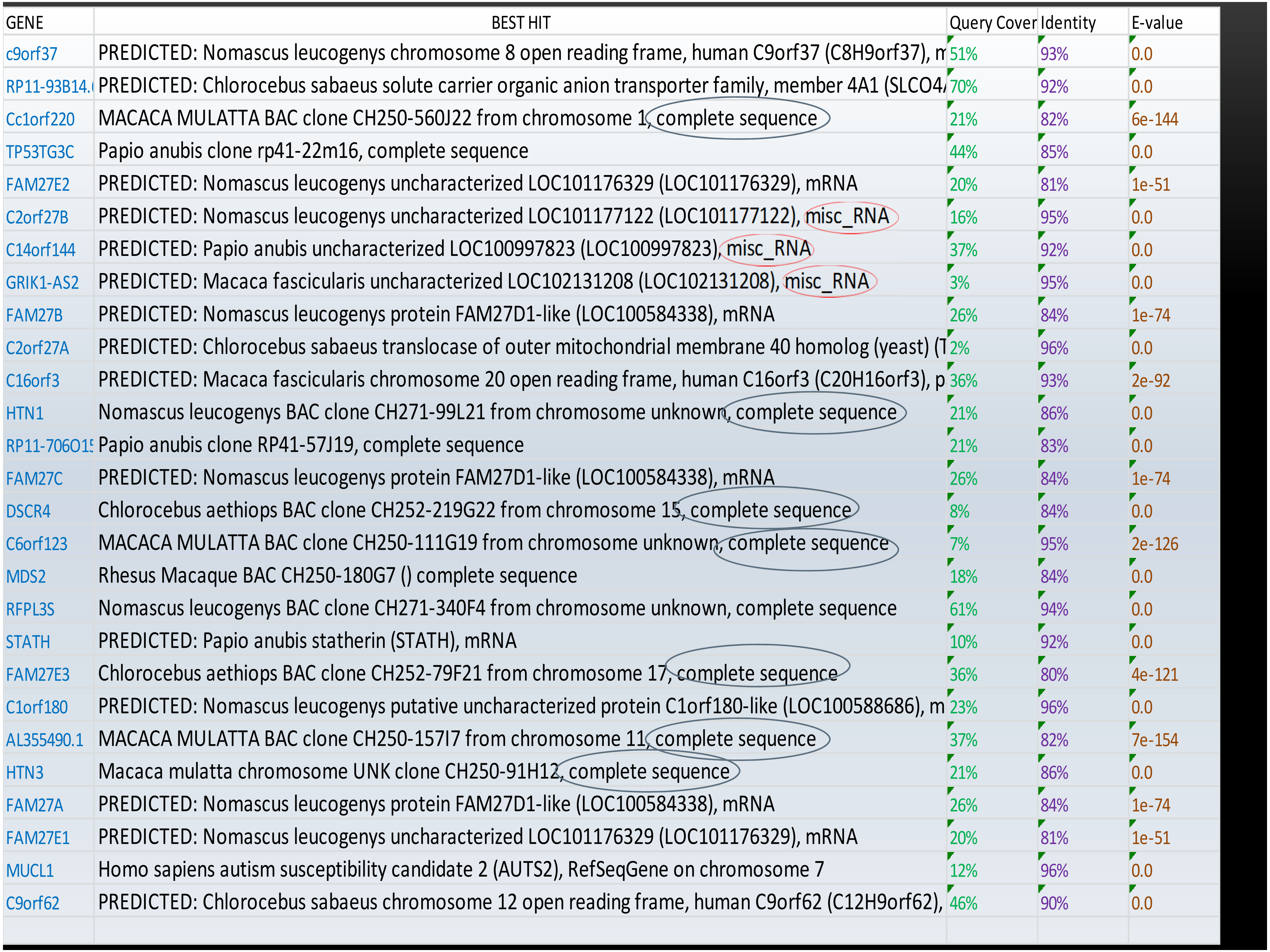
OMIM

- DSCR4(DOWN SYNDROME CRITICAL REGION 4)
- The region of chromosome 21 between genes CBR and ERG (CBR-ERG region), which spans 2.5 Mb on 21q22.2, has been defined by analysis of patients with partial trisomy 21. It contributes significantly to the pathogenesis of many characteristics of Down syndrome, including morphological features, hypotonia, and mental retardation. This gene is found in this region and multiple transcripts may exist. It is mainly expressed in the placenta.
- *MDS2(Myelodysplasia syndrome gene 2)*
- fused with ETV6 in myelodysplastic syndrome

HOMINIDAE-SPECIFIC NCRNA GENES BEST HIT IN GENEBANK, EMBL, DDBJ,PDB ,REFSEQ



GENE	BEST HIT	Query Co	Identity	E-value
MIR548I1	Macaca mulatta BAC CH250-425E5 (Children's Hospital Oakland Research Institute Rhesus macaque Adult Male BAC Library) complete sequ	53%	88%	2e-15
AC116562.1	Macaca mulatta BAC CH250-425E5 (Children's Hospital Oakland Research Institute Rhesus macaque Adult Male BAC Library) complete sequ	54%	88%	2e-15
snoU13(ENSG00000238445)	-	0	0	0
AC133041.1	Macaca mulatta BAC CH250-420H18 () complete sequence	51%	87%	7e-14
AC127526.1	Chlorocebus aethiops BAC clone CH252-426B21 from chromosome 18, complete sequence	52%	91%	6e-20
AP00807.2	-	0	0	0
AC006371.1	MACACA MULATTA BAC clone CH250-263I4 from chromosome y, complete sequence	61%	92%	2e-13
snoU13(ENSG00000252435)	-	0	0	0
AC144522.1	Rhesus Macaque BAC CH250-450C10 () complete sequence	54%	98%	4e-14
AC138761.1	-	0	0	0
MIR575	-	0	0	0
AL449083.1	PREDICTED: Chlorocebus sabaeus uncharacterized LOC103217206 (LOC103217206), ncRNA	45%	92%	2e-08
AC010930.1	PREDICTED: Papio anubis NACHT and WD repeat domain-containing protein 1-like (LOC101022659), partial mRNA	54%	96%	2e-11
AC068587.1	Macaca mulatta BAC CH250-425E5 (Children's Hospital Oakland Research Institute Rhesus macaque Adult Male BAC Library) complete sequ	53%	88%	2e-15
AC100832.1	-	0	0	0
AC007535.1	MACACA MULATTA BAC clone CH250-228N17 from chromosome y, complete sequence	38%	100%	1e-10



GENE	BEST HIT	Query Cover	Identity	E-value
c9orf37	PREDICTED: Nomascus leucogenys chromosome 8 open reading frame, human C9orf37 (C8H9orf37), m	51%	93%	0.0
RP11-93B14.	PREDICTED: Chlorocebus sabaeus solute carrier organic anion transporter family, member 4A1 (SLCO4	70%	92%	0.0
Cc1orf220	MACACA MULATTA BAC clone CH250-560J22 from chromosome 1, complete sequence	21%	82%	6e-144
TP53TG3C	Papio anubis clone rp41-22m16, complete sequence	44%	85%	0.0
FAM27E2	PREDICTED: Nomascus leucogenys uncharacterized LOC101176329 (LOC101176329), mRNA	20%	81%	1e-51
C2orf27B	PREDICTED: Nomascus leucogenys uncharacterized LOC101177122 (LOC101177122), misc_RNA	16%	95%	0.0
C14orf144	PREDICTED: Papio anubis uncharacterized LOC100997823 (LOC100997823), misc_RNA	37%	92%	0.0
GRIK1-AS2	PREDICTED: Macaca fascicularis uncharacterized LOC102131208 (LOC102131208), misc_RNA	3%	95%	0.0
FAM27B	PREDICTED: Nomascus leucogenys protein FAM27D1-like (LOC100584338), mRNA	26%	84%	1e-74
C2orf27A	PREDICTED: Chlorocebus sabaeus translocase of outer mitochondrial membrane 40 homolog (yeast) (T	2%	96%	0.0
C16orf3	PREDICTED: Macaca fascicularis chromosome 20 open reading frame, human C16orf3 (C20H16orf3), p	36%	93%	2e-92
HTN1	Nomascus leucogenys BAC clone CH271-99L21 from chromosome unknown, complete sequence	21%	86%	0.0
RP11-706O15	Papio anubis clone RP41-57J19, complete sequence	21%	83%	0.0
FAM27C	PREDICTED: Nomascus leucogenys protein FAM27D1-like (LOC100584338), mRNA	26%	84%	1e-74
DSCR4	Chlorocebus aethiops BAC clone CH252-219G22 from chromosome 15, complete sequence	8%	84%	0.0
C6orf123	MACACA MULATTA BAC clone CH250-111G19 from chromosome unknown, complete sequence	7%	95%	2e-126
MDS2	Rhesus Macaque BAC CH250-180G7 () complete sequence	18%	84%	0.0
RFPL3S	Nomascus leucogenys BAC clone CH271-340F4 from chromosome unknown, complete sequence	61%	94%	0.0
STATH	PREDICTED: Papio anubis statherin (STATH), mRNA	10%	92%	0.0
FAM27E3	Chlorocebus aethiops BAC clone CH252-79F21 from chromosome 17, complete sequence	36%	80%	4e-121
C1orf180	PREDICTED: Nomascus leucogenys putative uncharacterized protein C1orf180-like (LOC100588686), m	23%	96%	0.0
AL355490.1	MACACA MULATTA BAC clone CH250-157I7 from chromosome 11, complete sequence	37%	82%	7e-154
HTN3	Macaca mulatta chromosome UNK clone CH250-91H12, complete sequence	21%	86%	0.0
FAM27A	PREDICTED: Nomascus leucogenys protein FAM27D1-like (LOC100584338), mRNA	26%	84%	1e-74
FAM27E1	PREDICTED: Nomascus leucogenys uncharacterized LOC101176329 (LOC101176329), mRNA	20%	81%	1e-51
MUCL1	Homo sapiens autism susceptibility candidate 2 (AUTS2), RefSeqGene on chromosome 7	12%	96%	0.0
C9orf62	PREDICTED: Chlorocebus sabaeus chromosome 12 open reading frame, human C9orf62 (C12H9orf62),	46%	90%	0.0

CONSERVED NON CODING SEQUENCE

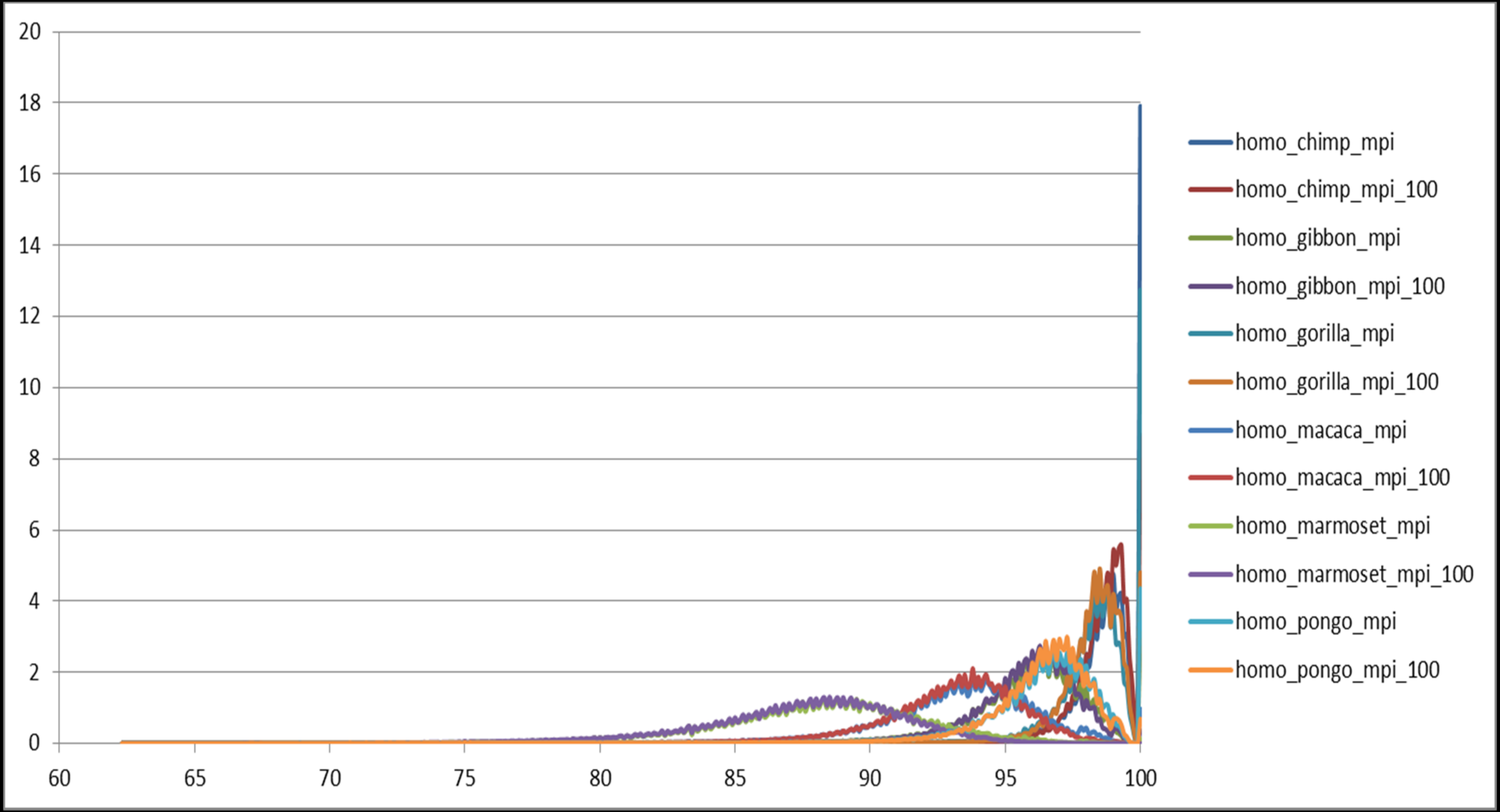
- King and Wilson suggested that the protein sequence divergence is insufficient to account for the difference between human and chimpanzee

- Conserved non-coding sequence (CNS) is a DNA sequence of noncoding DNA that is evolutionary conserved in certain group of species.

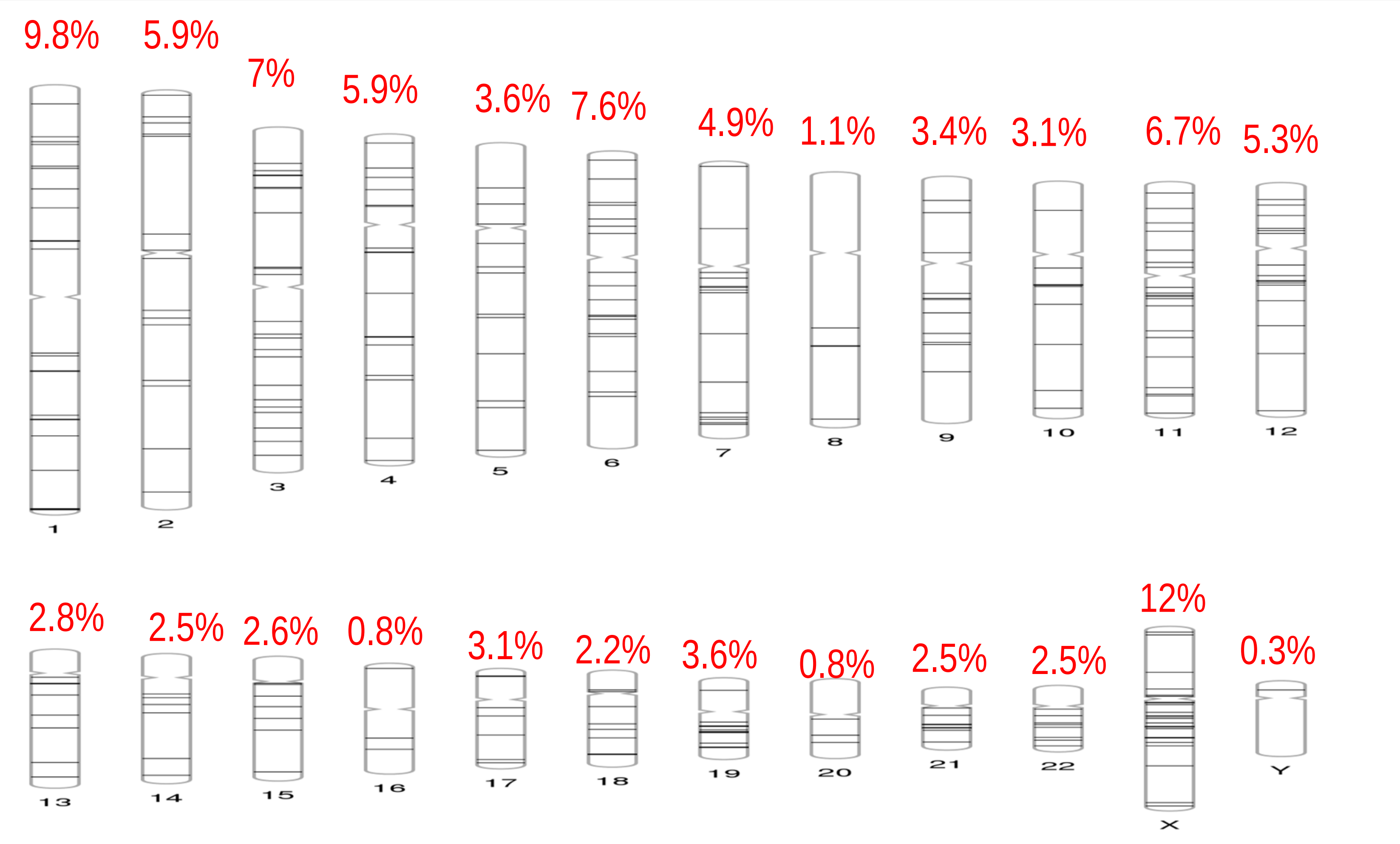
- CNSs in animals and plants are strongly associated with CRMs with diverse array of candidate functions such as matrix attachment regions ,transcription factor (TF) binding sites, chromosome-level regulatory regions , DNase I hypersensitive sites , enhancers, and prevention of ectopic expression

1.King M, Wilson A. Evolution at two levels in humans and chimpanzees. Science. 1975;188(4184):107-16.

IDENTITY DISTRIBUTION



HOMINIDAE SPECIFIC CNS DISTRIBUTION



1.Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet. 2006;7(8):645-53.

GENE ONTOLOGY

- 26% inside protein coding genes
- 152Kb average distance from closest protein coding genes