



Recent de novo origin of human protein-coding genes

David G. Knowles and Aoife McLysaght

Genome Res. 2009 19: 1752-1759 originally published online September 2, 2009

Age Tats

Journal Club on 2 March 2010

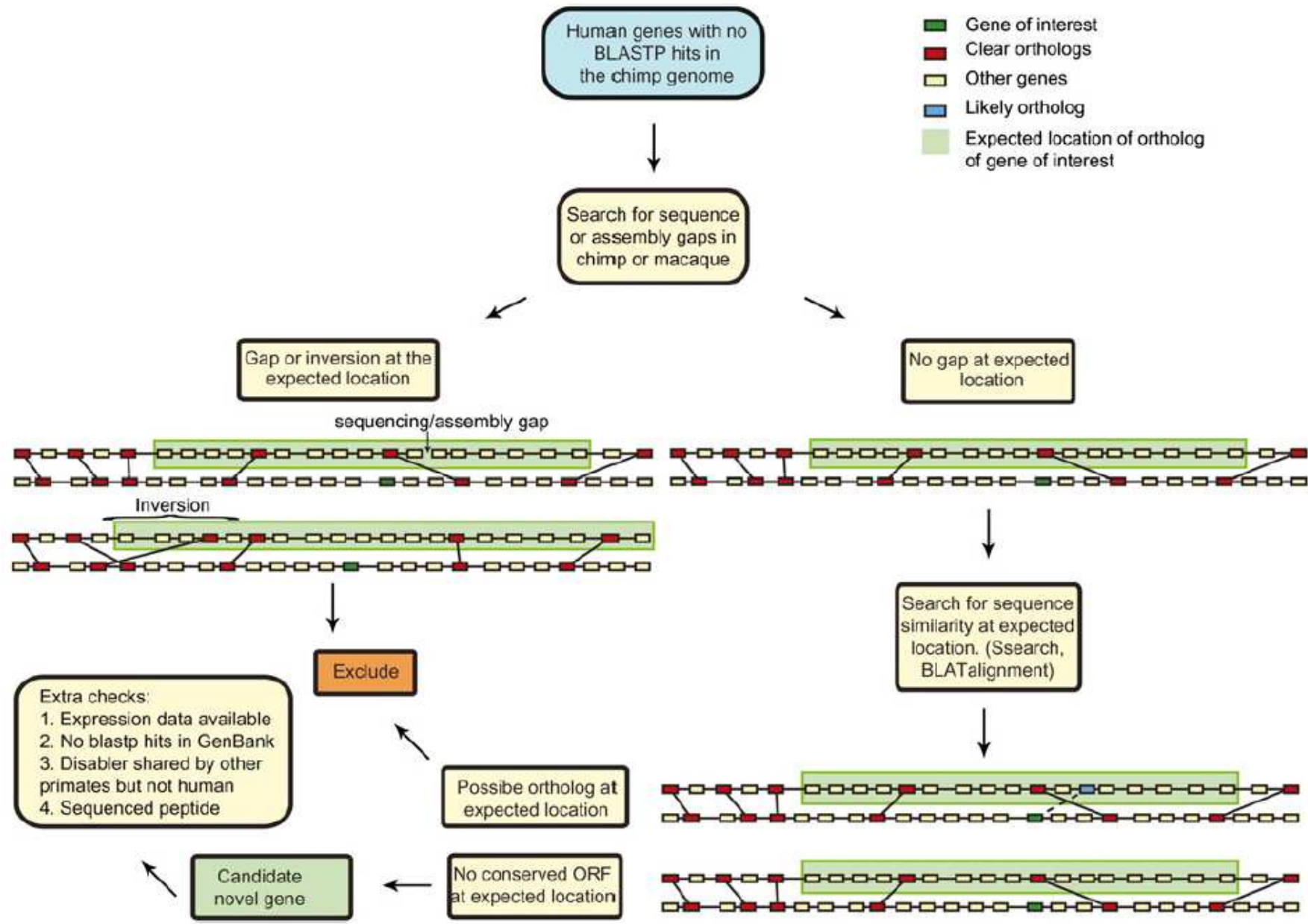
Table 1 | **Molecular mechanisms for creating new gene structures**

Mechanism	Process	Examples	Comments	References
Exon shuffling: ectopic recombination of exons and domains from distinct genes		<i>fucosyltransferase, jingwei, Tre2</i>	~19% of exons in eukaryotic genes have been formed by exon shuffling	8,32,40,62,65-68,105
Gene duplication: classic model of duplication with divergence		<i>CGβ, Cid, RNASE1B</i>	Many duplicates have probably evolved new functions	9-11,29,35,39,47,48,106
Retroposition: new gene duplicates are created in new genomic positions by reverse transcription or other processes		<i>PGAM3, Pgk2, PMCHL1, PMCHL2, Sphinx</i>	1% of human DNA is retroposed to new genomic locations	23,43,61,76,80-82,107-110
Mobile element: a mobile element, also known as a transposable element (TE), sequence is directly recruited by host genes		<i>HLA-DR-1, human DAF, lungerkine mRNA, mNSC1 mRNA</i>	Generates 4% of new exons in human protein-coding genes	16,78,111,112
Lateral gene transfer: a gene is laterally (horizontally) transmitted among organisms		<i>acetylneuraminate lysase, Escherichia coli mutU and mutS</i>	Most often reported in prokaryotes and recently reported in plants	18-20,113
Gene fusion/fission: two adjacent genes fuse into a single gene, or a single gene splits into two genes		Fatty-acid synthesis enzymes, <i>Kua-UEV, Sdic</i>	Involved in the formation of ~0.5% of prokaryotic genes	21,22,42,114,115
De novo origination: a coding region originates from a previously non-coding genomic region		<i>AFGPs, BC1RNA, BC200RNA</i>	Rare for whole gene origination; might not be rare for partial gene origination	52-53,116,117

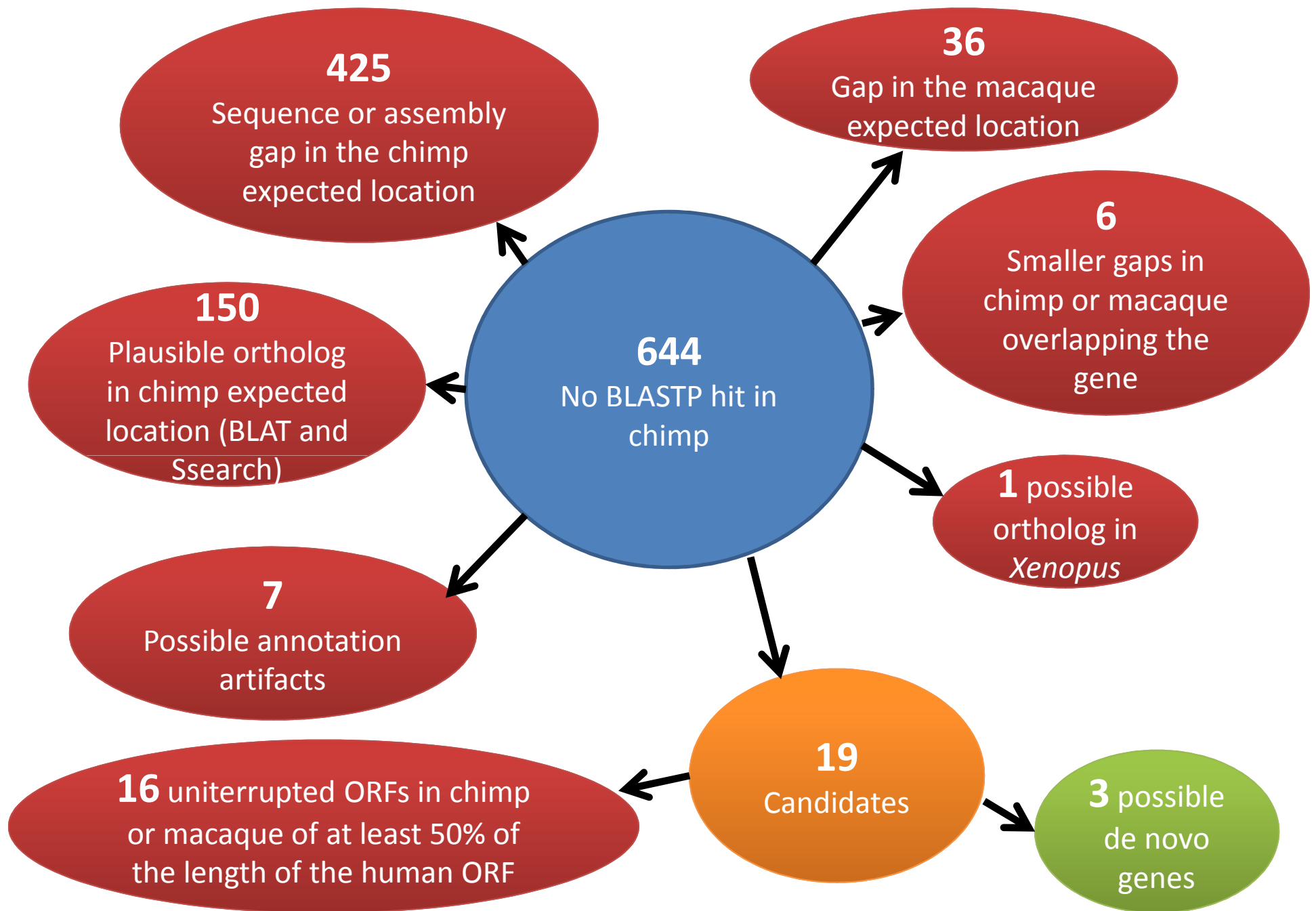
AFGP, antifreeze glycoprotein; *CGβ*, chorionic gonadotropin β polypeptide; *Cid*, centromere identifier; *DAF*, decay-accelerating factor; *HLA-DR-1*, major histocompatibility complex DR1; *PGAM3*, phosphoglycerate mutase 3; *Pgk2*, phosphoglycerate kinase 2; *PMCHL*, pro-melanin-concentrating hormone-like; *RNASE1B*, ribonuclease; *Sdic*, sperm-specific dynein intermediate chain; *UEV*, tumour susceptibility gene.

- Since 2006 several reports of de novo gene origins from *Drosophila* and yeast.
- ~12% of newly emerged genes in the *Drosophila melanogaster* subgroup may have arisen de novo from noncoding DNA, independently of transposable elements (Zhou *et al* (2008) *Genome Res*).
- 15 de novo genes identified in primate ancestor (Toll-Riera *et al* (2009) *Mol Biol Evol*).

- All-against-all BLASTP search between human, chimp and macaque proteins from Ensembl v46 (E-value < 1×10^{-4}).
- Unambiguous 1:1 orthologs with no other similarly strong hits. Lineage specific segmental duplications were excluded.
- Synteny blocks were constructed, anchored on these unambiguous orthologs, where the gap between anchors was no more than 10 genes in either genome. Local differences in gene order were permitted.
- Synteny blocks span 91% of the human and 85% of the chimp genomes.
- 94% of human protein-coding genes annotated by Ensembl are located within these blocks.



- BLAT and Ssearch sequence matches criteria:
 - translated sequence had $\geq 90\%$ identity with the human protein in each of the exons;
 - no in-frame stop codons in the first half of the alignment;
 - any inferred introns were at least 18 nt long



CLLU1

C22orf45

DNAH100S

- In chimp and macaque no potential ORF from the same start codon or in the same reading frame aligning to at least half of the human protein.
- BLASTP search against all of GenBank – absence of paralogs or orthologs of these genes in any sequenced genome.

- Length 121-163 amino acids
- No introns in coding sequence
- Introns in UTRs
- No complex protein domains annotated
- Overlapping other genes on the opposite strand

Table 1. Novel human protein-coding genes and supporting evidence.

Gene name	Ensembl ID	Length (codons)	Longest chimp ORF ^a	Expression support and tissue ^b	Primate shared disablers ^c	Other major sequence differences	Presence of enabler in other human complete genome sequences ^d	HapMap SNPs
<i>CLLU1</i>	ENSG00000205056	121	42	EST/cDNA: Blood (<u>AJ845165</u> , <u>AJ845166</u>); UniGene: Blood, embryonic tissue, eye, lymph, lymph node, muscle, pharynx, tonsil (Hs.339918)	1-bp indel ^e	Macaque: 4- and 1-bp indels	Sequence available and enabler conserved in all	1 syn.; 1 nonsyn.
<i>C22orf45</i>	ENSG00000178803	159	87 (25 amino acids align with human sequence)	EST/cDNA: Kidney, other (<u>AX747284</u> , <u>AK091970</u> , <u>DA635985</u>); ArrayExpress: Sperm, lung (E-GEOD-6872, E-GEOD-3020)	Premature stop codon	Chimp: 1-bp indel; Macaque: lacks ATG start codon; 4-bp indel	Reverse strand is available and conserved in Venter	1 nonsyn.
<i>DNAH1005</i>	ENSG00000204626	163	90 (75 amino acids align with human sequence)	EST/cDNA: Hippocampus (<u>AK127211</u>); UniGene: Blood, embryonic tissue, eye, lymph, lymph node, muscle, pharynx, tonsil (Hs.339918)	10-bp indel	Chimp: 2- and 1-bp indels; Macaque: lacks ATG start codon; 13-, 8-, 1-, and 1-bp indels	Reverse strand is available and conserved in Venter, Watson and HuAA	1 syn.; 1 nonsyn.

^aLength in codons of longest in-frame (alignable) ORF starting from any ATG in the region.

^bType of data/database is listed followed by tissue information with database identifiers in parentheses. Underlined accession numbers are full-length, spliced cDNA.

^cShared disablers are sequence differences shared by chimp, gorilla, orangutan, gibbon, and macaque that eliminate the capacity to produce a protein similar to the human protein.

^dIndependently sequenced whole genomes: Venter, Watson, HuAA, HuBB, HuCC, HuDD, and HuFF. All data are listed where available.

^eNot shared with orangutan.

Table 2. Peptide support for genes

Gene name	Codon position of shared disabler	Peptide match	Peptide database references ^a	Location in protein seq	BLASTP hits ^b	TBLASTN hits ^c
<i>CLU1</i>	41	HIIYSTFLSK	PeptideAtlas: PAp00140670	101	Self (0.41;10)	—
<i>C22orf45</i>	115	PCSNGGPAAAGEGR	PRIDE: 69; 73; 74; 75; 76; 8653; 8667	102	Self (9e-04; 14)	—
<i>DNAH10OS</i>	76-79	WQGCTRPALLAPSLATLK	PRIDE: 8668; 8672	137	Self (2e-08; 18)	Self (0.069)
		NPHSWGIKAHGLR	PRIDE: 8670a	75	^d	Self (8.8)
		LERCMVPESEWAPWQPQLPCEPK	PRIDE: 8670b	94	^d	Self (3e-05)

^aDatabase name and experiment numbers or identifiers.

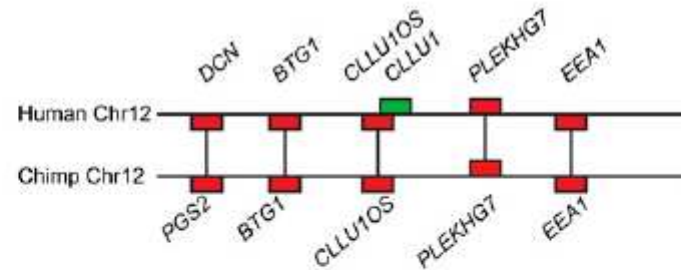
^bBLASTP search (with *E*-values < 10) against the GenBank nonredundant protein database (*E*-value and number of identities of the match are shown in parentheses).

^cTBLASTN search against the human genome (*E*-value is shown in parentheses).

^dNot in NCBI nonredundant database.

CCLU1

A

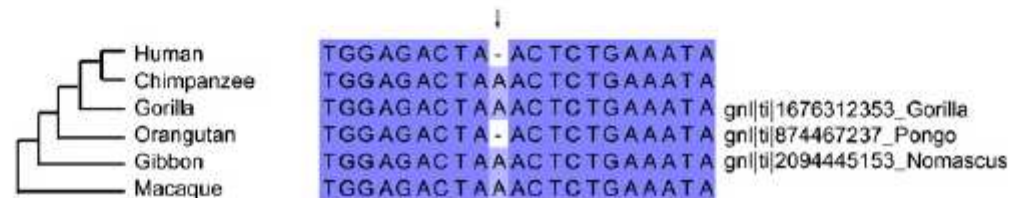


B

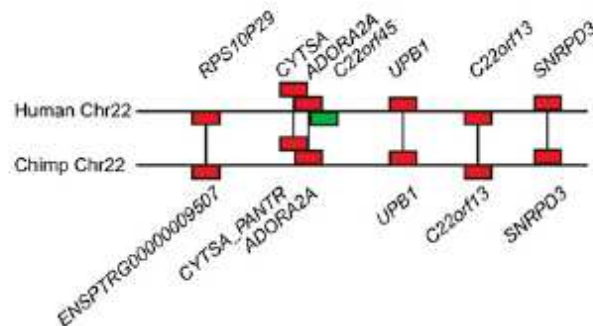
Start

Human	GTTTGGAGG - - - ATGTTCAAC AAATGCTCCTTTCATTTCCTCTATTTACAGACC TGCCGCA
Chimpanzee	GTTTGGAGG - - - ATGTTCAATAAATGCTGCCTTTCAC TCCTCTATTTACAGACC TGCCGCA
Macaque	GTTTGGAGG - - - ATGCTCAATAAATGCTCCTTTCATTTCCTCATTTTACAAACT TGCCGCA
Human	GACAATTC TGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee	GACAATTC TGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque	GACAATTC TGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human	GATCTGGAGACTAA - CTC TGA AATAAATAAGCTGATTATTTATTTATTTTCTCAAACAA
Chimpanzee	GATCTGGAGACTAAACTCTGA AATAAATAAGCTGATTATTTATTTATTTTCTCAAACAA
Macaque	TATCTGGAGACTAAACTCTGA AATAAATAAGCTGATTATTTATTTATTTTCTCAAACAA
Human	CAGAATACGATTTAGCAAATTACTTCTTAAGATATTATTTTACATTTCTATATTCTCCTA
Chimpanzee	CAGAATACGATTTAGCAAATTACTTCTTAAGATAC TATTTTACATTTCTATATTCTCCTA
Macaque	CAGAATA TGATTTAGCAAATTACTTCTTAAGATATTATTTTGCAC TTC TATATTCTCCTA
Human	CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTTCATAAAGCCAGGTATACA - - - TTATG
Chimpanzee	CCCTGAGTTGATGTGTGAGCCGATGTCAC TTTTCATAAAGCCAGGTATACA - - - TTATG
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCAC TTTCCACA AAGCCAGGTATATATACATTACG
	H I I Y S T F L S K
Human	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCA AAAATTTTAAATTTCT
Chimpanzee	GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAG AATTTTAAATTTCT
Macaque	GACAGGTAAGTAAAAA - CATATTATTTATTCTAGGTTTTTGTCCAAGAGTTTTTAAATTTCT
Human	AAC TGT TGC GCG TGT GT TGG TAA - - - TG TAAAACAAACTCAGTACA
Chimpanzee	AAC TGT TGC GCG TGT GT TGG TAA - - - TG TAAAACAAACTCAGTACA
Macaque	AAC TGT TGT TGC A TGT GT TGG TAA - - - CG TAAAACAAATTCAGTACG

C



A



C22orf45

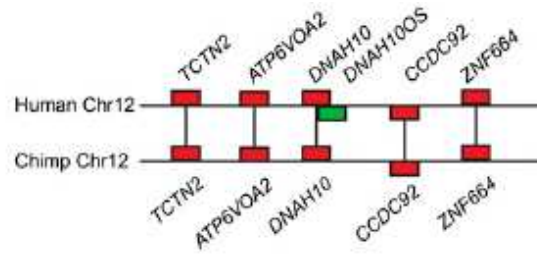
B

	Start
Human	CCAG - - - GACATGAGGG - - - ATGGAGCAGGACTGGCAACCTGGAGAGGAAGTCAC TCCTG
Chimpanzee	CCAG - - - GACATGAGGG - - - ATGGAGCAGGACTGGCAACCTGGAGAGGAAGTCAC TCCTG
Macaque	TCAG - - - GACATGAGGG - - - ACGGAGCAGGAT TGGCAACCTGGAGAGAAAGTCAG TCCTG
Human	GTCC TGAGCCCTGTTCAAAGGGCCAGGCTCCCC TC - TACCCCAT TGTCCATGTGAC AGAG
Chimpanzee	GTCC TGAGCCCTGTTCAAAGGGCCAGGCTCCCC TC T TACCCCAT TGTCCATGTGAC AGAG
Macaque	GTCC TGAGCCCTGTTCAAAGGG TCAGGCTCCCC TC - TACCCCAC TGTCCATGTGAC GGAG
Human	CTCAAACACACAGACCCCAACTTTCCCTCCAAC TCCAATGCTGTGCGGCACCTCAAGTGGC
Chimpanzee	CTCAAACACACAGACCCCAACTTTCCCTCCAAC TCCAATGCTGTGCGGCACCTCAAGTGGC
Macaque	CTCAAACA - - - GACCCCAACTTTCCCTCCAAC TCCAATGCTGTGCAAGCACC TCAAGTGGC
Human	TGGAACAGGATTGGCACGGGC TGCAGCCATACC TGGGACTGGAGGTTC TCC TGCACCCAG
Chimpanzee	TGGAACAGGATTGGCACGGGC TGCAGCCATACC TGGGACTGGAGGTTC TCC TGCACCCAG
Macaque	TGGAACAGGATTGGCACAGGC TGCAGCCATACC TGGGACTGGAGGTTC TCC TGCACCCAA
Human	CAGGCCCTTTTGCCCTAC TAGGAGCCTGGGAATGGAGCAT TGACACAGAAGCAGGAGGA
Chimpanzee	CAGGCCCTTTTGCCCTAC TAGGAGCCTGGGAATGGAGCAT TGA ACAGAAGCAGGAGGA
Macaque	CAGGCCCTTTTGCGTCTAC TAGGAGCCTGGGAATGGAGCAT TGACACAGAAGCAGGAGGA
	P C S N G G P A A A G E
Human	GGAAGGAGAGAGCAGAG - CCAGAAACCC TGCAGCAACGGAGGGCC TGCAGCAGC TGGAGA
Chimpanzee	GGAAGGAGAGAGCAGAGCCAGAAACCC TGCAGAAACCC TGCAGAAACGGAGGGCC TGCAGCAGC TGGAGA
Macaque	GGAAGGAGAGAGCAGAGACCAGAGACCC TGCAAAAATGGAGGGCC TGCAGCAGC TGGAGA
	G I R
Human	GGGCGGAGTCC TCCC AAGCCCTGCTTTCCATGGAGC ACTTGCCAGGCAGCCATTCACAA
Chimpanzee	GGGCGGAGTCC TCCC AAGCCCTGCTTTCCATGGGGC ACTTGCCAGGCAGCCATTCACAA
Macaque	GGGCGGAGTCC TCCC AAGCCCTGCTTTCCATGGGGC ACTTGCCAGGCAGCCATTCACAA
	W Q G C T R P A L L A P S L A T L
Human	AGTGTGTCGTTGGCAGGGATGCC ACCAGACCAGCTCTCC TGGCACCATCC TGGCCACACT
Chimpanzee	AGTGTGTTGTTGGCAGGGATGCC ACCAGACCAGCTCTCC TGGCACCATCC TGGCCACACT
Macaque	AGCATGTGCTGGCAGGGATGCC ACCAGACCAGCTCTCC TGGCACCATECT TGGCCACACT
	K
Human	CAAGGAACACAGTTATCCC TGA - - - TGCTCTTGGC
Chimpanzee	CAAGGAACACAGTTATCCC TGA - - - TGCTCTTGGC
Macaque	CAAGGAACACAGTTATCCC TGA - - - TGTTCCTGGC

C



A

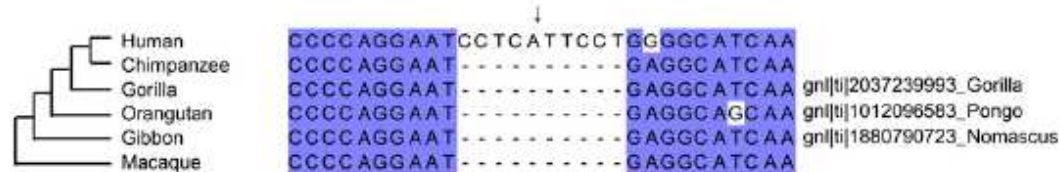


DNAH10OS

B

	Start
Human	CTGGATACAAC TGGAGC --- ATGCACAGCC TGCCACGGAGTGGCTCTATCAGGCCGCACAC
Chimpanzee	CTGGATACAAC TGGAAC --- ATGCACAGCC TGCCACGGAGTGGCTCTATCAGGCCGCACAC
Macaque	CTGGACACAACCGGAGC --- ATCCACAGCC TGCCA ----- TCAGGCCGCACAC
Human	ACAGCGACACACAGGCCACTGGCTGGCCCTCC TCCCCAGCGCAT TGGGGACAGCCCAGGCC
Chimpanzee	ACAGCGACACACAGGCCACTGGCTGGCCCTCC TCCCCAGCGCAT TGGGGACAGCCCAGGCC
Macaque	AGCGACACACACAAGCCACTGGT TGGCCCTCC TCCCCAGCGCC TGGGGACAGCCCAGGCC
Human	CTTC TCCAGCATTCTGTGCTGCCCCACCTTCCCCTCTGTGGAGGAGCAGCCCAGACAGGAG
Chimpanzee	CTTC TCCAGCATTCTGTGCTGCCCCACCTTCCCCTCTGTGGAGGAGCAGCCCAGACAGGA-
Macaque	CTTC TCCAGCGCTTCTGTGCTGCCCCACCTTCCCCTCTGTGGAGGAACGGCTCAGACAGGAC
Human	ACCC TGTGGCCCTGCCCATGGCCCCAGAGAAATGGGTGTGGGGCGGTGGCCCTC TCCCCCA
Chimpanzee	- DCC TGTGGCCCTGGCC - ATGGCCCCAGAGAAATGGGTGTGGGGCGGTGGCCCTC TCCCCCA
Macaque	ACCC TGTGGCCCTC TCCCCAGGGCCCCAGAGAAATGGGTGTGGGGCGGTGGCCCTC TCCCCCA
	N P H S W G I K A H G L R
Human	GGAATCCTCATTCTTGGGGCATCAAGGCCACGGACTTAGACCACCC TGGGCCCCAGGC
Chimpanzee	GGAAT-----GAGGCATCAAGGCCACGGACTTAGACCACCC TGGGCCCCAGGC
Macaque	GGAAT-----GAGGCATCAAGGTCGACGGACTTAGACCACCC TGGGCCCCAGGC
	L E R C M V P E S E W A P W Q P Q L P C
Human	TAGAAAGATGCATGGTCCCAGAGTCAGAAATGGGCACCATGGCAACCCCAGCTACCC TGTG
Chimpanzee	TAGAAAGATGCATGGTCCCAGAGTCAGAAATGGGCACCATGGCAACCCCAGCTACCC TGTG
Macaque	TAGAAAGATGCATGGCCCCAGAGTCAGAAATGGGCACCATGGCAACCCCAGCTACCC TGGG
	E P K
Human	AGCCGAAGTGGCTGGGGAGCAGGAAGTCGAAGCCTCACAGAGAAAGTGGTCTCCGGGGAG
Chimpanzee	AGCCGAAGTGGCTGGGGAGCAGGAAGTCGAAGCCTCACAGAGAAAGTGGTCTCCGGGGAG
Macaque	AGCCGAAGTGGCTGGGGAGCAGGAAGTAGAAGCGTCACAGAGAAAGCGGCCCCGGGGAG
Human	GAGGACCCAGCAGATGTGCAAGAGAGGGAACACA-----CTCCTGTGGCCCCAGAGA
Chimpanzee	GAGGACCCAGCAGACGTGCAAGAGAGGGAACACA-----CTCCTGTGGCCCCAGAGA
Macaque	GAGGACCCAGCAGACATAGAAAGGGAGAAAGACAGGAGCACAGCCTGTGGCCCCAGAGA
Human	GAGTGGTGGCCCGGACACCTGCCACCTCCCC TGGC ACTGA --- GACCTGGAGA
Chimpanzee	GAGTGGTGGCCCGGACACCTGCCACCTCCCC TGGC ACTGA --- GACCTGGAGA
Macaque	GAGTGGCAGCCAGACACCTGCCACCTCCCC -GCCAGCTCCCC -GCCACTGA --- GAGCCGCAGA

C



Additional proof

- Resequencing three orthologous regions DNA in one chimp individual verified the critical sequence differences.
- Each of the disablers shared by chimp and macaque is also shared with the gorilla and gibbon; two are also shared with orangutan.

- For novel chimp genes no reliable cases were identified.
- Estimation: the frequency of novel protein-coding genes in human genome is about 0.075% → ~18 genes

- Knowles and McLysaght (2009) Recent de novo origin of human protein-coding genes. *Genome Res.*, **19**, 1752-1759
- Siepel (2009) Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.*, **19**, 1693-1695
- Long et al (2003) The origin of new genes: Glimpses from the young and old. *Nature Reviews*, **4**, 865-875