Recent de novo origin of human protein-coding genes

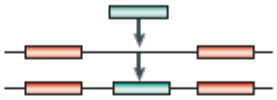
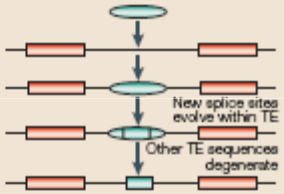
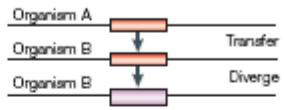
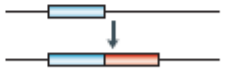
David G. Knowles and Aoife McLysaght

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		*fucosyltransferase*, *jingwei*, *Tre2*	~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		*CGβ*, *Cid* , *RNASE1B*	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		*PGAM3*, *Pgk2*, *PMCHL1*, *PMCHL2*, *Sphinx*	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		*HLA-DR-1*, human *DAF*, *lungerkine* mRNA, *mNSC1* mRNA	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		*acytylneuraminate lysase*, *Escherichia coli* *mutU* and *mutS*	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, *Kua-UEV*, *Sdic*	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		*AFGPs*, *BC1RNA*, *BC200RNA*	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; *RNASE*, ribonuclease; *Sdic*, sperm-specific dynein intermediate chain; *UEV*, tumour susceptibility gene.

Long *et al* (2003) *Nature Reviews*

- 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*).

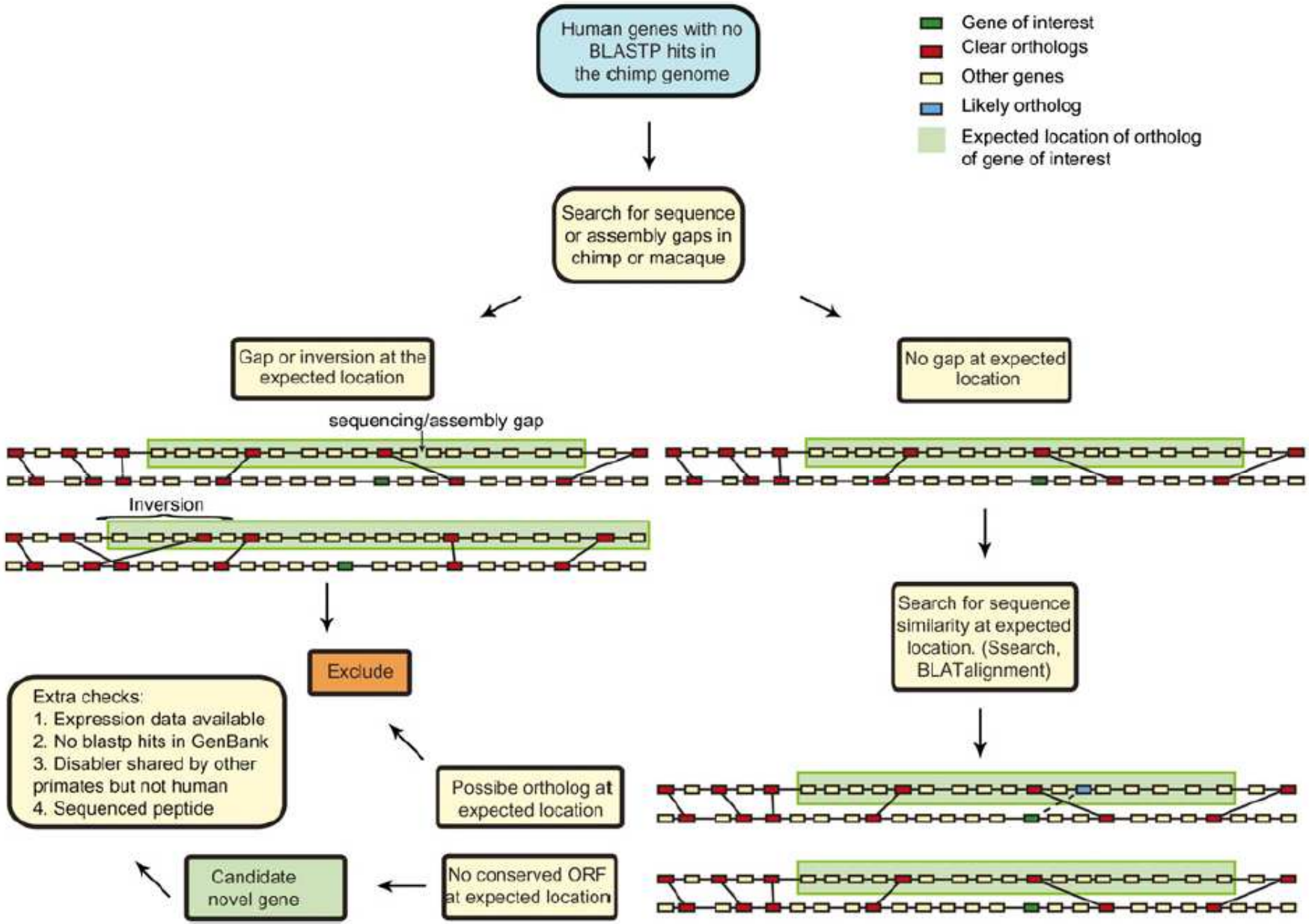
- All-against-all BLASTP search between human, chimp and macaque proteins from Ensembl v46 (E-value < 1x10E-4).

- Unambiguous 1:1 orthologs with no other similarly strong hits. Lineage specific segmental duplications were excluded.

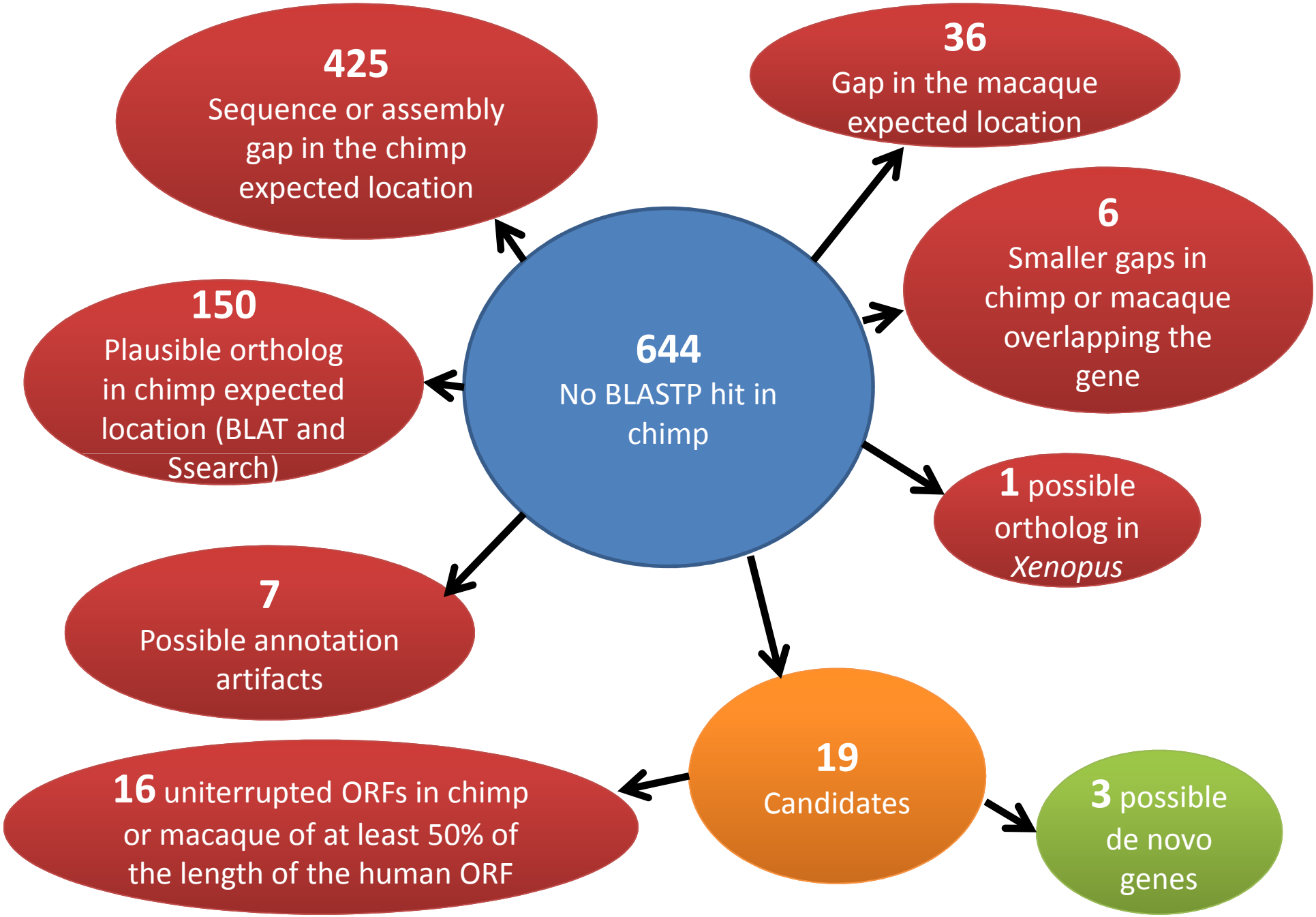
- Synteny blocks were constructed, anchored on these unambiguous orthologs, where the gap between anchors was no more than 10 genes in either genome. Local differences in gene order were permitted.

- 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.


Human genes with no BLASTP hits in the chimp genome
Gene of interest
Clear orthologs
Other genes
Likely ortholog
Expected location of ortholog of gene of interest
Search for sequence or assembly gaps in chimp or macaque
Gap or inversion at the expected location
No gap at expected location
sequencing/assembly gap
Inversion
Search for sequence similarity at expected location. (Ssearch, BLATalignment)
Exclude
Extra checks:
1. Expression data available
2. No blastp hits in GenBank
3. Disabler shared by other primates but not human
4. Sequenced peptide
Possibe ortholog at expected location
Candidate novel gene
No conserved ORF at expected location

- BLAT and Ssearch sequence matches criteria:
  - translated sequence had ≥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


644
No BLASTP hit in chimp
425
Sequence or assembly gap in the chimp expected location
36
Gap in the macaque expected location
6
Smaller gaps in chimp or macaque overlapping the gene
150
Plausible ortholog in chimp expected location (BLAT and Ssearch)
1 possible ortholog in Xenopus
7
Possible annotation artifacts
19
Candidates
16 uniterrupted ORFs in chimp or macaque of at least 50% of the length of the human ORF
3 possible de novo genes

CLLU1
C22orf45
DNAH10OS

- 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
CLLU1	ENSG00000205056	121	42	EST/cDNA: Blood (AJ845165, AJ845166); 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.
C22orf45	ENSG00000178803	159	87 (25 amino acids align with human sequence)	EST/cDNA: Kidney, other (AX747284, AK091970, DA635985); 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.
DNAH10OS	ENSG00000204626	163	90 (75 amino acids align with human sequence)	EST/cDNA: Hippocampus (AK127211); 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.

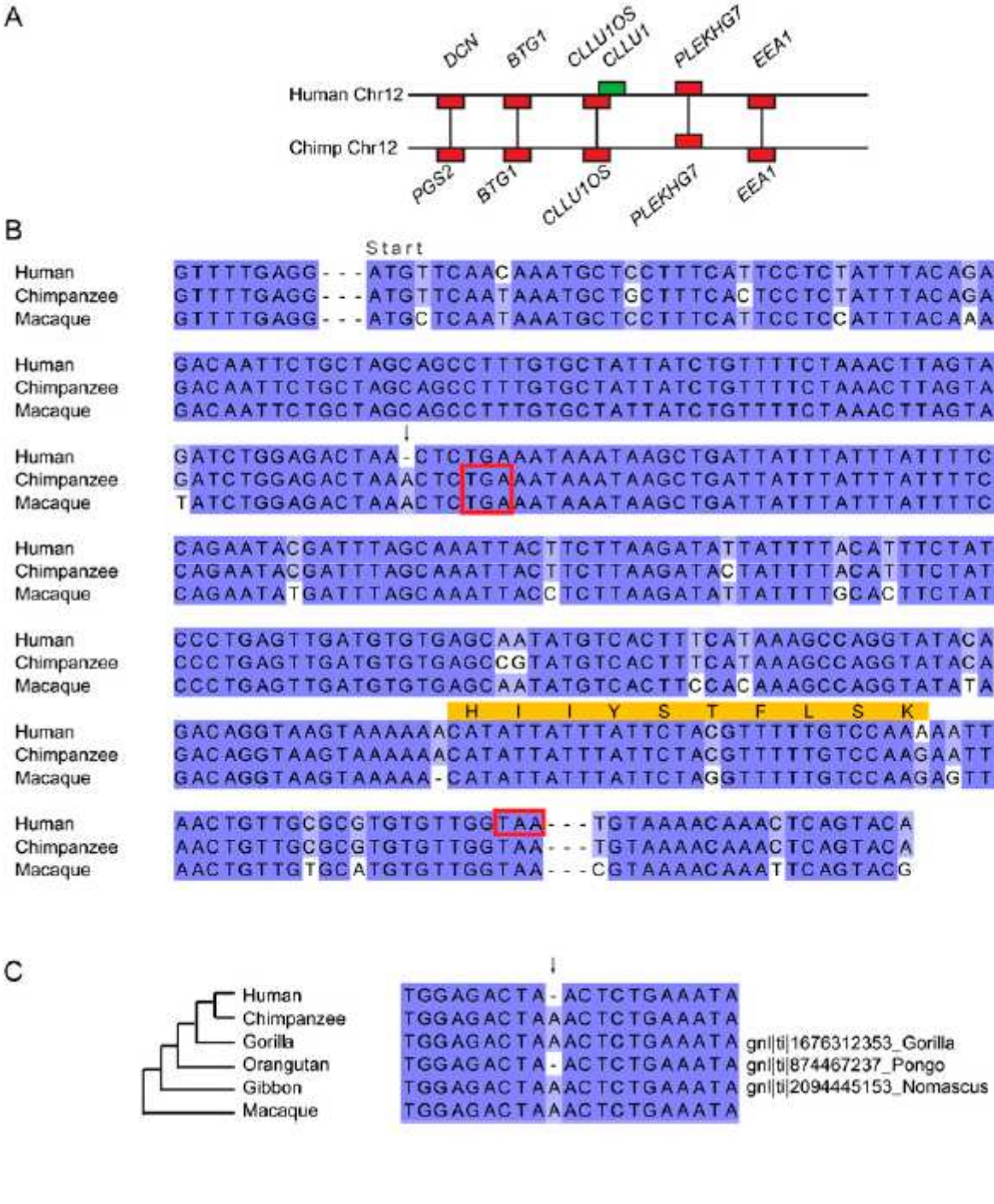
[a]Length in codons of longest in-frame (alignable) ORF starting from any ATG in the region.
[b]Type of data/database is listed followed by tissue information with database identifiers in parentheses. Underlined accession numbers are full-length, spliced cDNA.
[c]Shared disablers are sequence differences shared by chimp, gorilla, orangutan, gibbon, and macaque that eliminate the capacity to produce a protein similar to the human protein.
[d]Independently sequenced whole genomes: Venter, Watson, HuAA, HuBB, HuCC, HuDD, and HuFF. All data are listed where available.
[e]Not shared with orangutan.

Table 2. Peptide support for genes

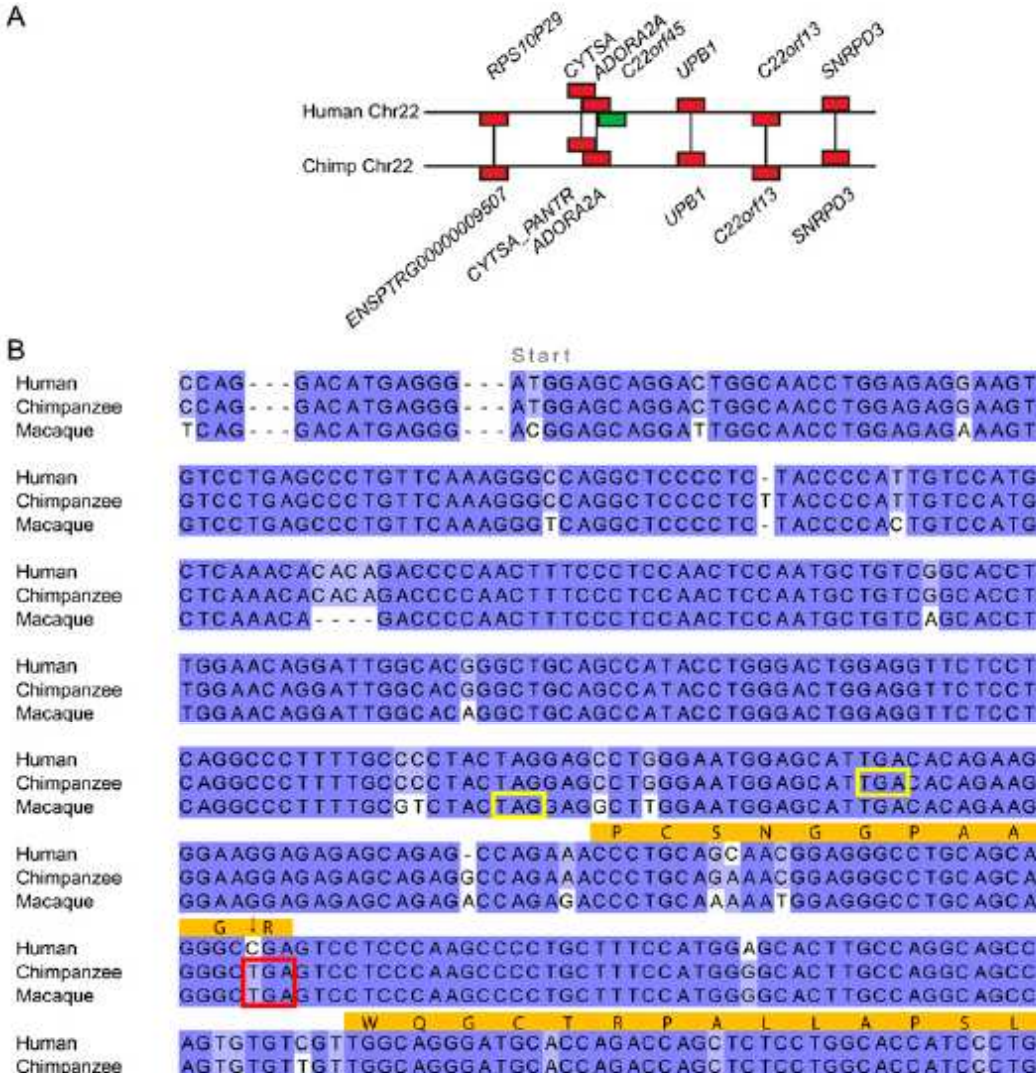
Gene name	Codon position of shared disabler	Peptide match	Peptide database references[a]	Location in protein seq	BLASTP hits[b]	TBLASTN hits[c]
CLLU1	41	HIIYSTFLSK	PeptideAtlas: PAp00140670	101	Self (0.41;10)	—
C22orf45	115	PCSNGGPAAAGEGR	PRIDE: 69; 73; 74; 75; 76; 8653; 8667	102	Self (9e-04; 14)	—
		WQGCTRPALLAPSLATLK	PRIDE: 8668; 8672	137	Self (2e-08; 18)	Self (0.069)
DNAH10OS	76-79	NPHSWGIKAHGLR	PRIDE: 8670a	75	[d]	Self (8.8)
		LERCMVPESEWAPWQPQLPCEPK	PRIDE: 8670b	94	[d]	Self (3e-05)

[a]Database name and experiment numbers or identifiers.
[b]BLASTP search (with *E*-values < 10) against the GenBank nonredundant protein database (*E*-value and number of identities of the match are shown in parentheses).
[c]TBLASTN search against the human genome (*E*-value is shown in parentheses).
[d]Not in NCBI nonredundant database.

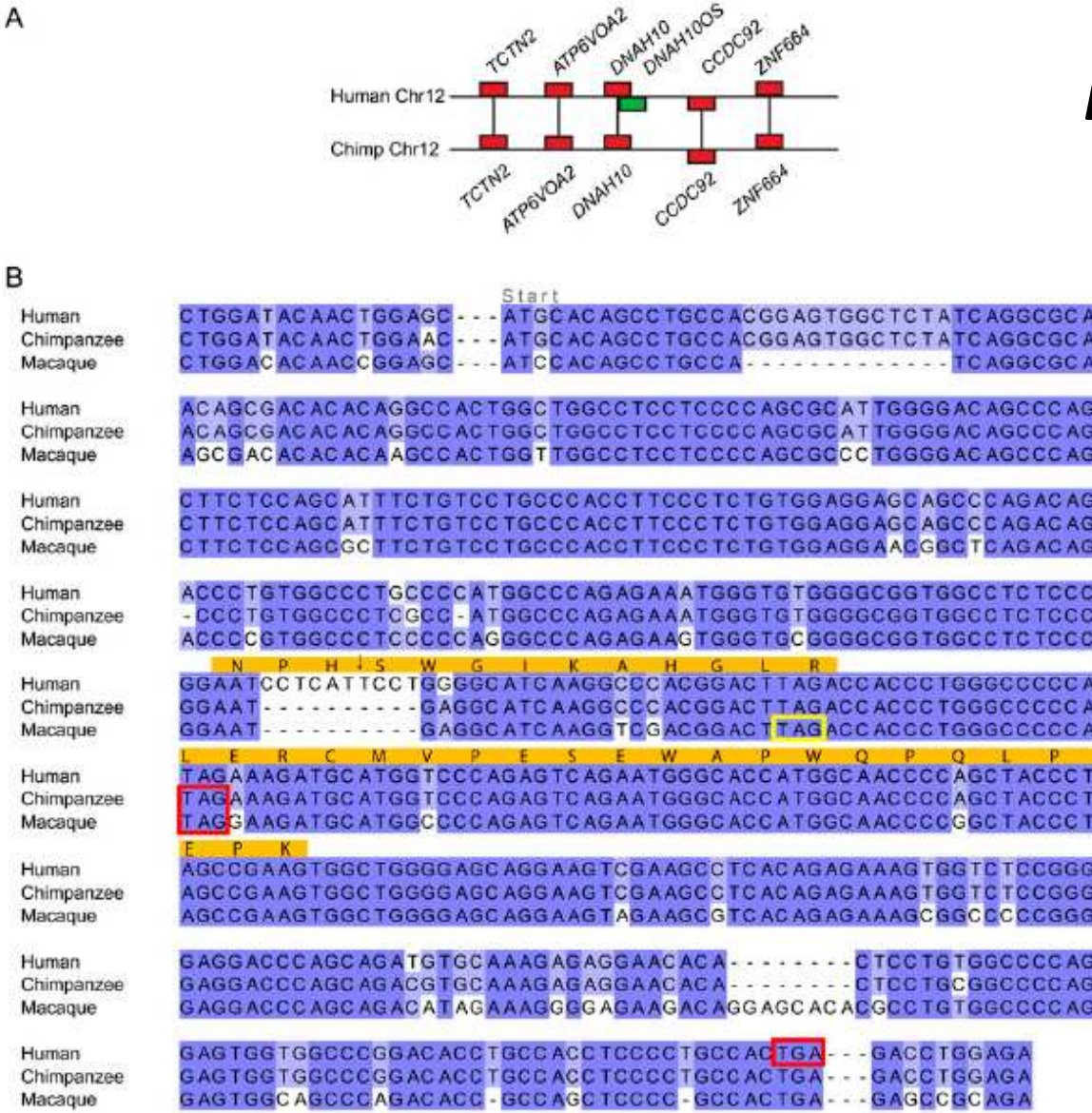
CCLU1


A
DCN
BTG1
CLLU1OS
CLLU1
PLEKHG7
EEA1
Human Chr12
Chimp Chr12
PGS2
BTG1
CLLU1OS
PLEKHG7
EEA1
B
Start
Human
Chimpanzee
Macaque
H I I Y S T F L S K
C
Human
Chimpanzee
Gorilla
Orangutan
Gibbon
Macaque
gnl|ti|1676312353_Gorilla
gnl|ti|874467237_Pongo
gnl|ti|2094445153_Nomascus

C22orf45


A
RPS10P29
CYTSA
ADORA2A
C22orf45
UPB1
C22orf13
SNRPD3
Human Chr22
Chimp Chr22
ENSPTRG00000009507
CYTSA_PANTR
ADORA2A
UPB1
C22orf13
SNRPD3
B
Start
Human
Chimpanzee
Macaque
C
Human
Chimpanzee
Gorilla
Orangutan
Gibbon
Macaque
gnl|ti|1676820278_Gorilla
gnl|ti|880352846_Pongo
gnl|ti|2060600787_Nomascus

DNAH10OS


A
TCTN2
ATP6V0A2
DNAH10
DNAH10OS
CCDC92
ZNF664
Human Chr12
Chimp Chr12
B
Start
Human
Chimpanzee
Macaque
C
Human
Chimpanzee
Gorilla
Orangutan
Gibbon
Macaque
gnl|ti|2037239993_Gorilla
gnl|ti|1012096583_Pongo
gnl|ti|1880790723_Nomascus

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