

# Ensembl Gene Annotation (e!91)

## Primate Clade

### Table of Contents

<b>SECTION 1: GENOME PREPARATION</b>	<b>2</b>
REPEAT FINDING	3
LOW COMPLEXITY FEATURES, AB INITIO PREDICTIONS AND BLAST ANALYSES	3
<b>SECTION 2: PROTEIN-CODING MODEL GENERATION</b>	<b>5</b>
CDNA ALIGNMENT PIPELINE	5
PROJECTION MAPPING PIPELINE	6
PROTEIN-TO-GENOME PIPELINE	7
RNA-SEQ PIPELINE	8
<b>SECTION 3: FILTERING THE PROTEIN-CODING MODELS</b>	<b>9</b>
PRIORITISING MODELS AT EACH LOCUS	9
ADDITION OF UTR TO CODING MODELS	10
GENERATING MULTI-TRANSCRIPT GENES	10
PSEUDOGENES	10
IMMUNOGLOBULIN AND T-CELL RECEPTOR GENES	11
<b>SECTION 4: CREATING THE FINAL GENE SET</b>	<b>12</b>
SMALL NCRNAS	12
CROSS-REFERENCING	12
STABLE IDENTIFIERS	12
<b>SECTION 5: FINAL GENE SET SUMMARY</b>	<b>13</b>
<b>SECTION 6: APPENDIX - FURTHER INFORMATION</b>	<b>14</b>
ASSEMBLY INFORMATION	16
STATISTICS OF INTEREST	17
LAYERS IN DETAIL	18
MORE INFORMATION	19
<b>REFERENCES</b>	<b>20</b>

This document describes the annotation process of an assembly. The first stage is Assembly Loading where databases are prepared and the assembly loaded into the database.

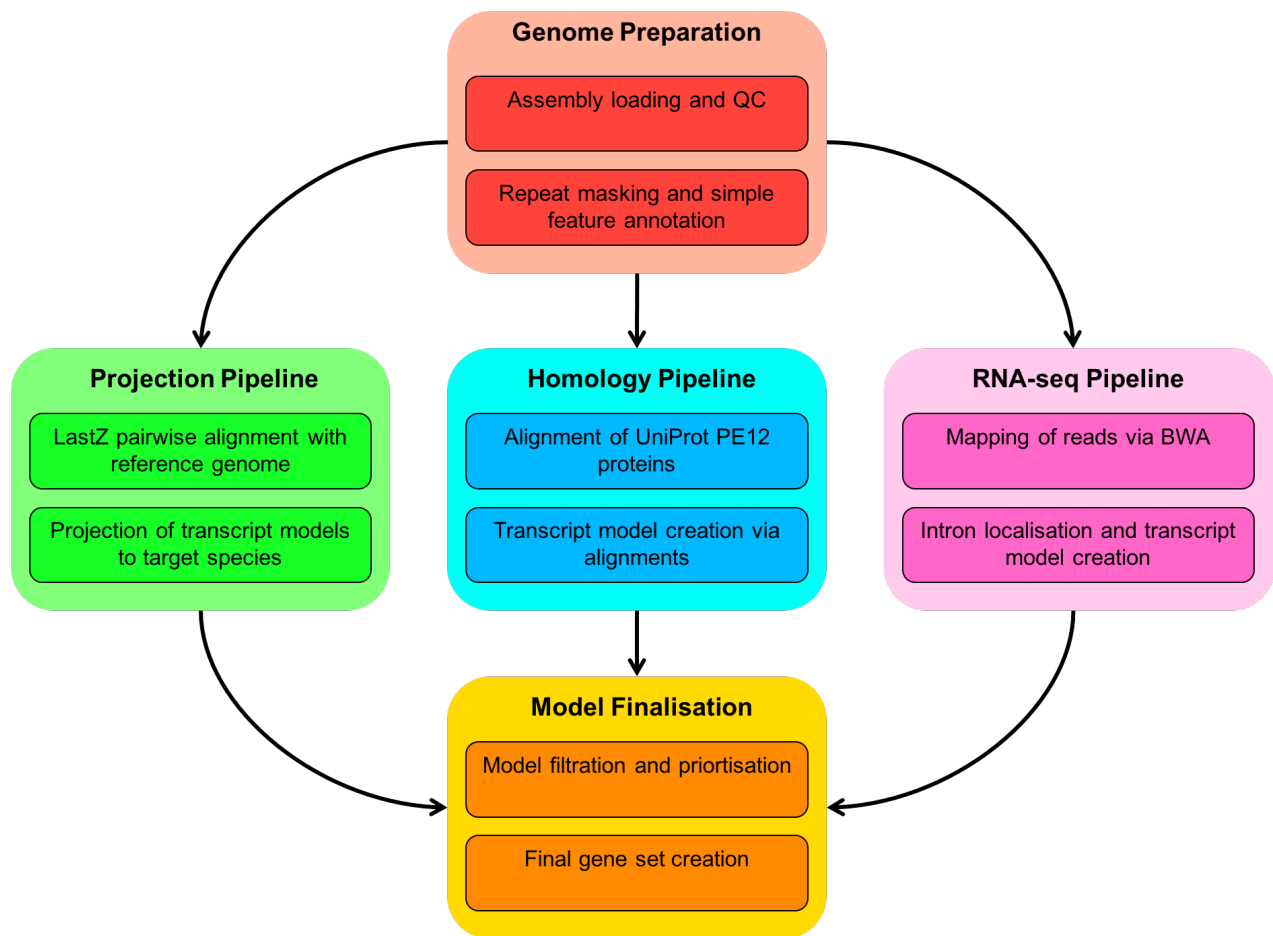


Fig. 1: Flowchart of the protein-coding annotation pipeline. Small ncRNAs, Ig genes, TR genes, and pseudogenes are computed using separate pipelines.

## Section 1: Genome Preparation

The genome phase of the Ensembl gene annotation pipeline involves loading an assembly into the Ensembl core database schema and then running a series of analyses on the loaded assembly to identify an initial set of genomic features.

The most important aspect of this phase is identifying repeat features (primarily through RepeatMasker) as soft masking of the genome is used extensively later in the annotation process.

### Repeat Finding

After the genomic sequence has been loaded into a database, it is screened for sequence patterns including repeats using RepeatMasker [1] (version 4.0.5 with parameters, using as the search engine), Dust [2] and TRF [3].

For the primate clade annotation, the Repbase primate library was used with RepeatMasker.

### Low complexity features, ab initio predictions and BLAST analyses

Transcription start sites are predicted using Eponine-scan [4]. CpG islands longer than 400 bases and tRNAs are also predicted. The results of Eponine-scan, CpG, and tRNAscan [5] are for display purposes only; they are not used in the gene annotation process.

Genscan [6] is run across repeat-masked sequence to identify ab initio gene predictions. The results of the Genscan analyses are also used as input for UniProt [7], UniGene [8]

and Vertebrate RNA alignments by NCBI-BLAST [9]. Passing only Genscan results to BLAST is an effective way of reducing the search space and therefore the computational resources required.

Genscan predictions are for display purposes only and are not used in the model generation phase.

## Section 2: Protein-Coding Model Generation

Various sources of transcript and protein data are investigated and used to generate gene models using a variety of techniques. The data and techniques employed to generate models are outlined here. The numbers of gene models generated are described in gene summary.

### cDNA alignment pipeline

cDNAs are downloaded from RefSeq [10] and aligned to the genome using Exonerate [11]. The cDNAs are mainly used for display purposes, but can be used to add UTR to the protein coding transcript models if they have a matching set of introns.

For the primate clade annotation, a minimal sequence length of 60bp and a cut-off of 95% identity and 50% coverage were required for an alignment to be kept.

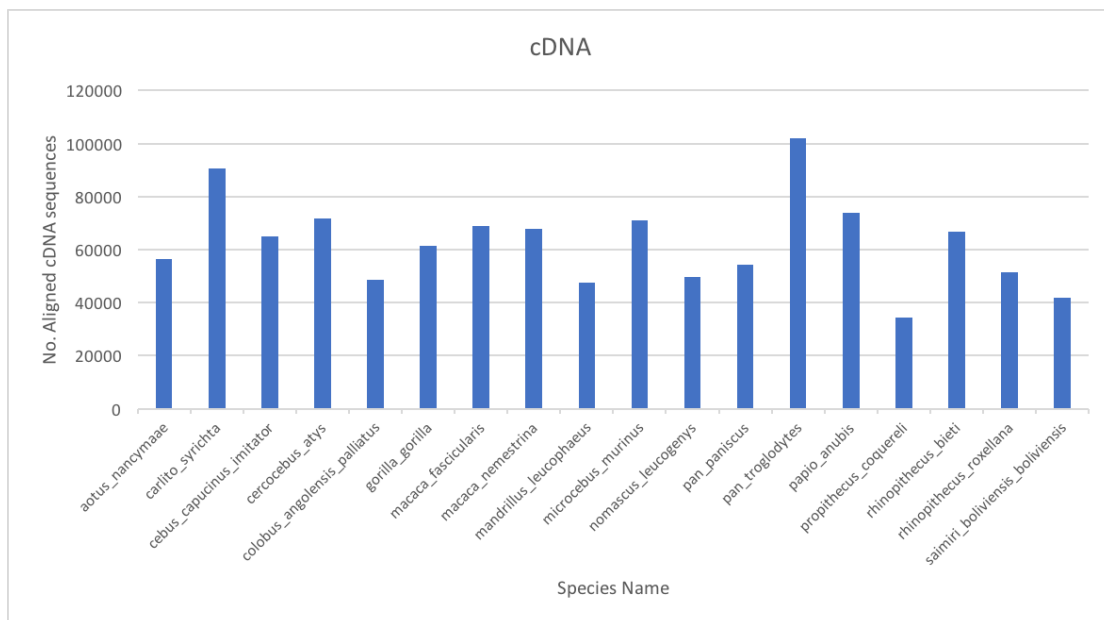


Fig. 2: Counts of cDNA genes in each species

## Projection mapping pipeline

For all species a whole genome alignment is generated against a suitable reference assembly using LastZ [12]. Syntenic regions identified using this alignment are then used to map protein coding annotation from the most recent GENCODE [13] gene set.

For the primate clade annotation, the human assembly, GRCh38, was used as a reference and The GENCODE 27 gene set was used to map protein coding annotation.

The mapped transcripts are then assessed for non-canonical splice sites and frameshifts; this can happen when mapping coordinates from one assembly to another. Mapped transcripts featuring two or more non-canonical splice sites/frameshifts are passed into a realignment pipeline. Here they are re-aligned to the original sequence in the region they are mapped to. If possible, a model with canonical splicing is built otherwise the transcript model is discarded.

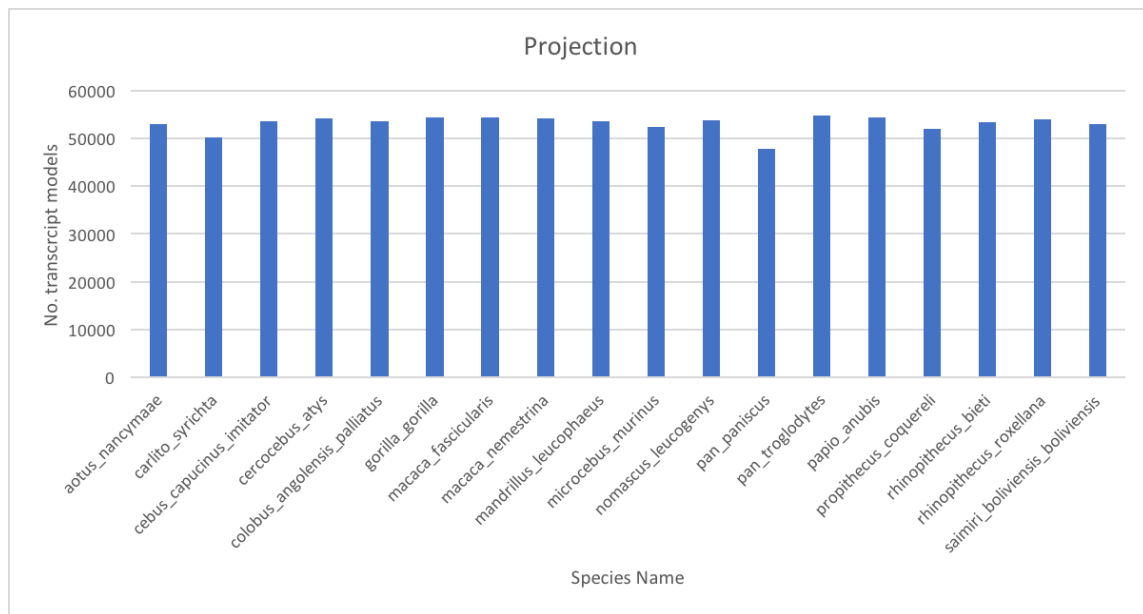


Fig 3: Counts of transcript models build by the projection pipeline for each species

## Protein-to-genome pipeline

Protein sequences are downloaded from UniProt and aligned to the genome in a splice aware manner using GenBlast [14]. The set of proteins aligned to the genome is a subset of UniProt proteins used to provide a broad, targeted coverage of the primate proteome. The set consists of the following:

- Self SwissProt/TrEMBL PE 1 & 2
- Human SwissProt/TrEMBL PE 1 & 2
- Other primates SwissProt/TrEMBL PE 1 & 2
- Other mammals SwissProt/TrEMBL PE 1 & 2

Note: PE level = protein existence level

For the primate clade annotation, a cut-off of 50 percent coverage and identity and an e-value of  $e^{-1}$  were used for GenBlast with the exon repair option turned on. The top 5 transcript models built by GenBlast for each protein passing the cut-offs are kept.

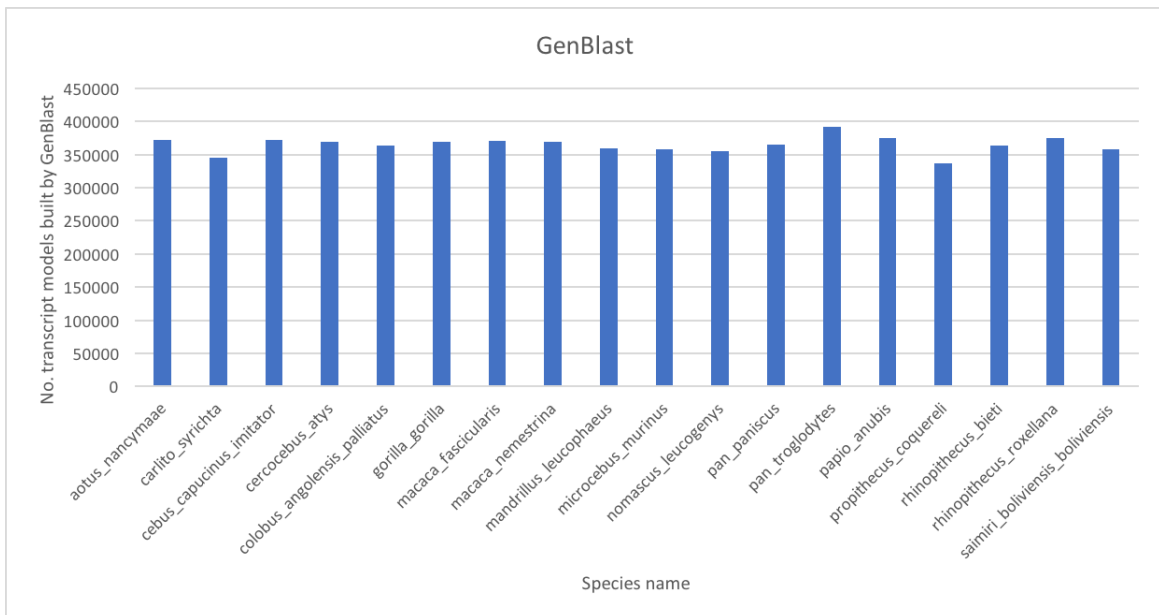


Fig. 4: Counts of transcript models built by GenBlast for each species

## RNA-seq pipeline

RNA-seq data is downloaded from ENA (<https://www.ebi.ac.uk/ena/>) and used in the annotation. A merged file containing reads from all tissues/samples is created. The merged data is less likely to suffer from model fragmentation due to read depth. The available reads are aligned to the genome using BWA [15], with a tolerance of 50 percent mismatch to allow for intron identification via split read alignment. Initial models generated from the BWA alignments are further refined via exonerate. Protein coding models are identified via a BLAST alignment of the longest ORF against the UniProt vertebrate PE 1 & 2 data set.

In the case where multiple tissues/samples are available we create a gene track for each such tissue/sample that can be viewed in the Ensembl browser and queried via the API.

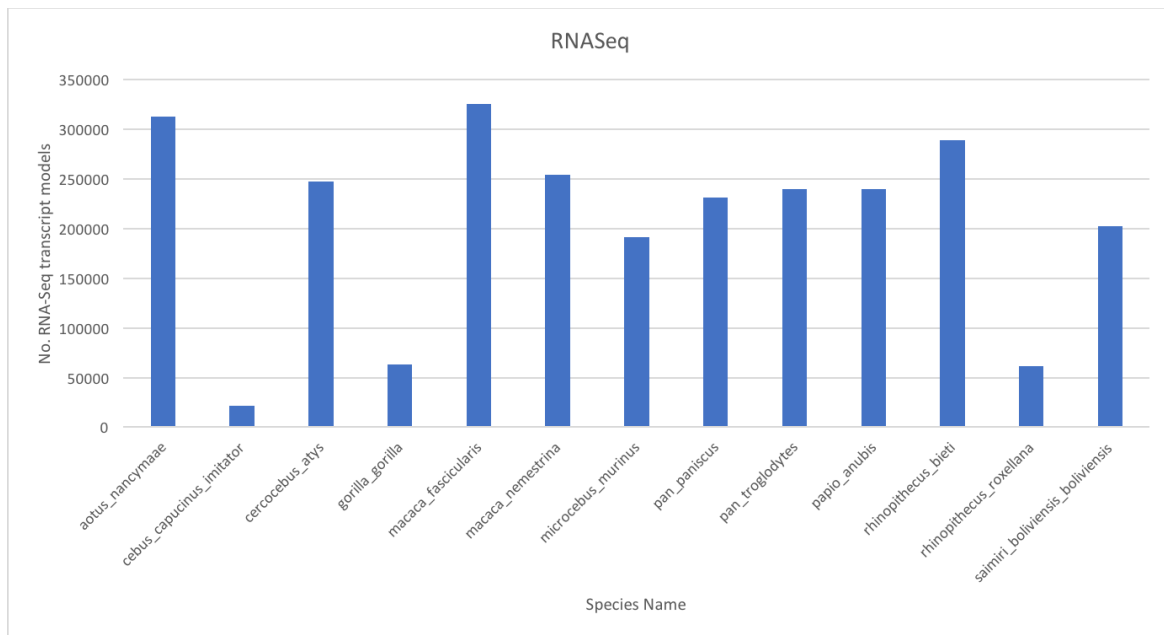


Fig 5: Counts of RNA-Seq transcript models for each species with publically available RNA-Seq data



## Section 3: Filtering the Protein-Coding Models

The filtering phase decides the subset of protein-coding transcript models, generated from the model-building pipelines, that comprise the final protein-coding gene set. Models are filtered based on information such as what pipeline was used to generate them, how closely related the data are to the target species and how good the alignment coverage and percent identity to the original data are.

### Prioritising models at each locus

The LayerAnnotation module is used to define a hierarchy of input data sets, from most preferred to least preferred. The output of this pipeline includes all transcript models from the highest ranked input set. Models from lower ranked input sets are included only if their exons do not overlap a model from an input set higher in the hierarchy.

Note that models cannot exist in more than one layer. For UniProt proteins, models are also separate into clades, to help selection during the layering process. Each UniProt protein is in one clade only, for example mammal proteins are present in the mammal clade and are not present in the vertebrate clade to avoid aligning the proteins multiple times.

When selecting the model or models kept at each position, we prioritise based on the highest layer with available evidence. In general, the highest layers contain the set of evidence containing the most trustworthy evidence in terms of both alignment/mapping quality, and also in terms of relevance to the species being annotated. So, for example, when a primate is being annotated, well aligned evidence from either the species itself or other closely related vertebrates would be chosen over evidence from more distant species. Regardless of what species is being annotated, well-aligned human proteins are usually included in the top layer as human is the current most complete vertebrate

annotation. For further details on the exact layering used please refer to section 6.

### Addition of UTR to coding models

The set of coding models is extended into the untranslated regions (UTRs) using RNA-seq data (if available) and alignments of species-specific RefSeq cDNA sequences. The criteria for adding UTR from cDNA or RNA-seq alignments to protein models lacking UTR (such as the projection models or the protein-to-genome alignment models) is that the intron coordinates from the model missing UTR exactly match a subset of the coordinates from the UTR donor model.

### Generating multi-transcript genes

The above steps generate a large set of potential transcript models, many of which overlap one another. Redundant transcript models are collapsed and the remaining unique set of transcript models are clustered into multi-transcript genes where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene.

### Pseudogenes

Pseudogenes are annotated by looking for genes with evidence of frame-shifting or lying in repeat heavy regions. Single exon retrotransposed pseudogenes are identified by searching for a multi-exon equivalent elsewhere in the genome. A total number of genes that are labelled as pseudogenes or processed pseudogenes will be included in the core db, please check Final Gene set Summary.

## Immunoglobulin and T-cell Receptor genes

Translations of different human IG gene segments are downloaded from the IMGT database [16] and aligned to the genome using GenBlast.

For the primate clade annotation, a cut-off of 80 percent coverage, 70 percent identity and an e-value of  $e^{-1}$  were used for GenBlast with the exon repair option turned on. The top 10 transcript models built by GenBlast for each protein passing the cut-offs are kept. In cases where multiple sequences aligned to the same locus, we selected the alignment with the highest combined percent identity and coverage.

## Section 4: Creating the Final Gene Set

### Small ncRNAs

Small structured non-coding genes are added using annotations taken from RFAM [17] and miRBase [18]. Rfam and miRBase annotations were searched against the genomic sequence using NCBI-BLAST. The resulting alignments were then filtered using RNA-fold (miRBase hits) and Infernal [19] (Rfam hits).

### Cross-referencing

Before public release the transcripts and translations are given external references (cross-references to external databases). Translations are searched for signatures of interest and labelled where appropriate.

### Stable Identifiers

Stable identifiers are assigned to each gene, transcript, exon and translation. When annotating a species for the first time, these identifiers are auto-generated. In all subsequent annotations for a species, the stable identifiers are propagated based on comparison of the new gene set to the previous gene set.

## Section 5: Final Gene Set Summary

Species	Protein coding	pseudogenes	RNAs	IG genes	TR genes
<i>aotus_nancymae</i>	20320	396	7110	55	37
<i>carlito_syrichtha</i>	18304	537	5986	80	36
<i>cebus_capucinus_imitator</i>	20200	550	6973	86	31
<i>cercocebus_atys</i>	20746	540	6455	124	56
<i>colobus_angolensis_palliatus</i>	20467	497	6138	101	49
<i>gorilla_gorilla</i>	21588	522	7286	133	96
<i>macaca_fascicularis</i>	21404	302	6706	127	75
<i>macaca_nemestrina</i>	20872	584	6588	132	56
<i>mandrillus_leucophaeus</i>	20660	414	6448	120	83
<i>microcebus_murinus</i>	18103	428	6774	68	33
<i>nomascus_leucogenys</i>	20648	567	6465	75	71
<i>pan_paniscus</i>	21041	549	7010	98	93
<i>pan_troglodytes</i>	23302	485	7932	138	116
<i>papio_anubis</i>	21464	423	6699	127	56
<i>propithecus_coquereli</i>	17884	416	5294	31	32
<i>rhinopithecus_bieti</i>	20824	563	6575	90	74
<i>rhinopithecus_roxellana</i>	21132	648	6664	104	75
<i>saimiri_boliviensis_boliviensis</i>	19290	439	7306	59	53

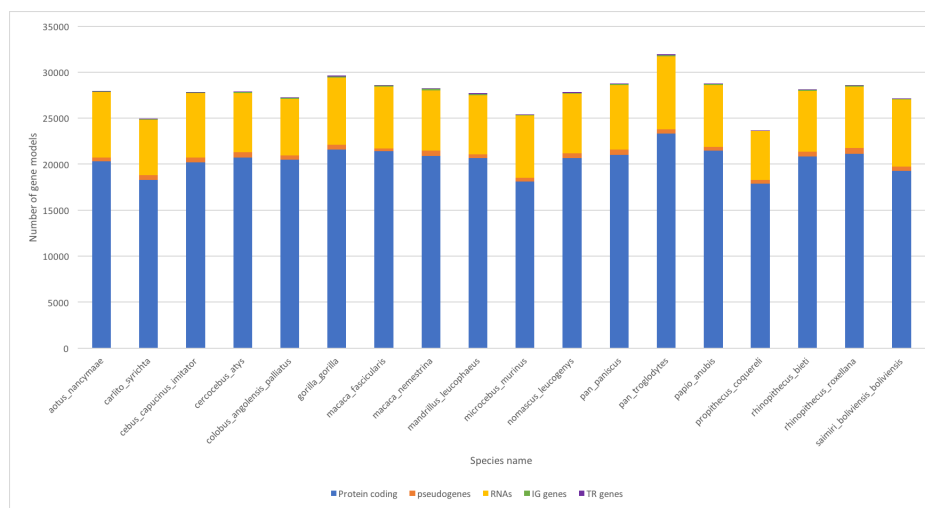


Table 1 and Fig. 6: Counts of the major gene classes in each species

## Section 6: Appendix - Further information

The Ensembl gene set is generated automatically, meaning that gene models are annotated using the Ensembl gene annotation pipeline. The main focus of this pipeline is to generate a conservative set of protein-coding gene models, although non-coding genes and pseudogenes may also be annotated.

Every gene model produced by the Ensembl gene annotation pipeline is supported by biological sequence evidence (see the “Supporting evidence” link on the left-hand menu of a Gene page or Transcript page); ab initio models are not included in our gene set. Ab initio predictions and the full set of cDNA and EST alignments to the genome are available on our website.

The quality of a gene set is dependent on the quality of the genome assembly. Genome assembly can be assessed in a number of ways, including:

### 1. Coverage estimates

- A higher coverage usually indicates a more complete assembly.
- Using Sanger sequencing only, a coverage of at least 2x is preferred.

### 2. N50 of contigs and scaffolds

- A longer N50 usually indicates a more complete genome assembly.
- Bearing in mind that an average human gene may be 10-15 kb in length, contigs shorter than this length will be unlikely to hold full-length gene models.

### 3. Number of contigs and scaffolds

- A lower number top level sequences usually indicates a more complete

genome assembly.

#### 4. Alignment of cDNAs and ESTs to the genome

- A higher number of alignments, using stringent thresholds, usually indicates a more complete genome assembly.

## Assembly Information

Species name	Common name	Assembly name	Genbank accession ID	Assembly level
<i>aotus_nancymae</i>	Mas night monkey	Anan_2.0	GCA_000952055.2	Scaffold
<i>carlito_syrichta</i>	Tarsier	Tarsius_syrichta-2.0.1	GCA_000164805.2	Scaffold
<i>cebus_capucinus_imitator</i>	Capuchin	Cebus_imitator-1.0	GCA_001604975.1	Scaffold
<i>cercocebus_atys</i>	Sooty mangabey	Caty_1.0	GCA_001604975.1	Scaffold
<i>colobus_angolensis_palliatus</i>	Angola colobus	Cang.pa_1.0	GCA_000951035.1	Scaffold
<i>gorilla_gorilla</i>	Gorilla	gorGor4	GCA_000151905.3	Chromosome
<i>macaca_fascicularis</i>	Crab-eating macaque	Macaca_fascicularis_5.0	GCA_000364345.1	Chromosome
<i>macaca_nemestrina</i>	Pig-tailed macaque	Mnem_1.0	GCA_000956065.1	Scaffold
<i>mandrillus_leucophaeus</i>	Drill	Mleu.le_1.0	GCA_000951045.1	Scaffold
<i>microcebus_murinus</i>	Mouse lemur	Mmur_3.0	GCA_000165445.3	Chromosome
<i>nomascus_leucogenys</i>	Gibbon	Nleu_3.0	GCA_000146795.3	Chromosome
<i>pan_paniscus</i>	Bonobo	panpan1.1	GCA_000258655.2	Chromosome
<i>pan_troglodytes</i>	Chimpanzee	Pan_tro_3.0	GCA_000001515.5	Chromosome
<i>papio_anubis</i>	Olive baboon	Panu_3.0	GCA_000264685.2	Chromosome
<i>propithecus_coquereli</i>	Coquerels sifaka	Pcoq_1.0	GCA_000956105.1	Scaffold
<i>rhinopithecus_bieti</i>	Black snub-nosed monkey	ASM169854v1	GCA_001698545.1	Scaffold
<i>rhinopithecus_roxellana</i>	Golden snub-nosed monkey	Rrox_v1	GCA_000769185.1	Scaffold
<i>saimiri_boliviensis_boliviensis</i>	Bolivian squirrel monkey	SaiBol1.0	GCA_000235385.1	Scaffold

Table 2: Assembly info



## Statistics of Interest

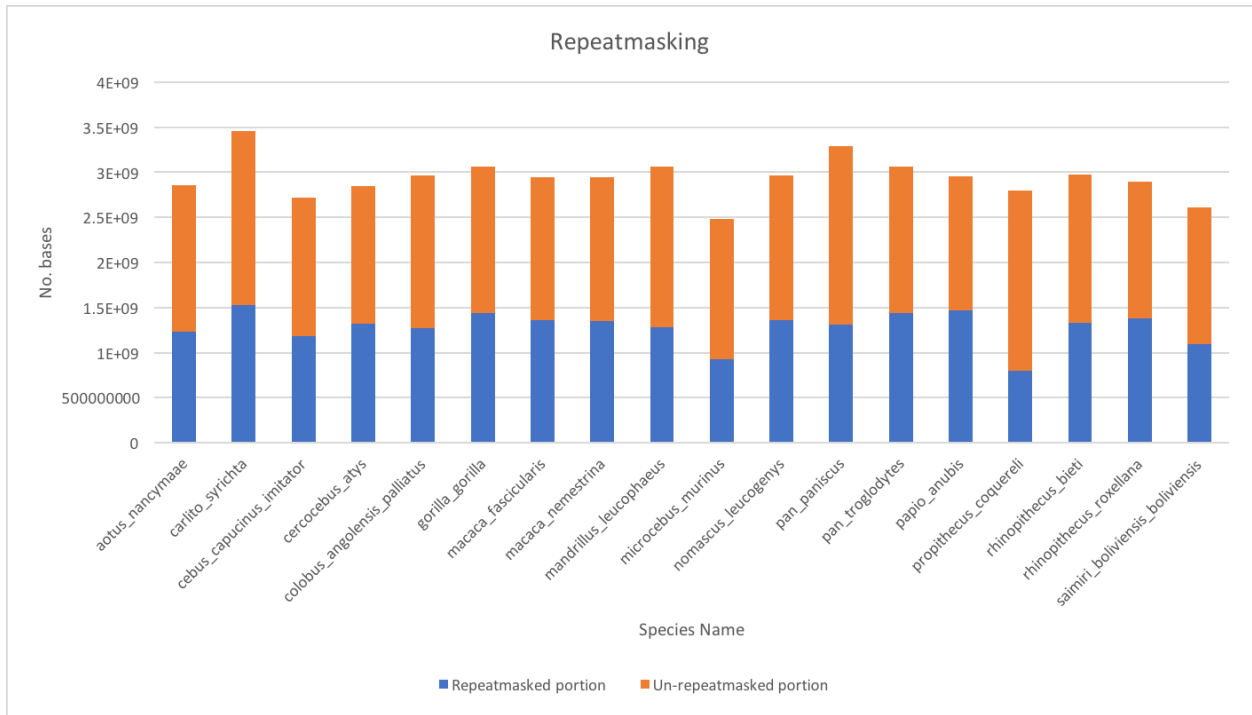


Fig 7: Number of bases unmasked (orange) and repeat masked (blue) per species

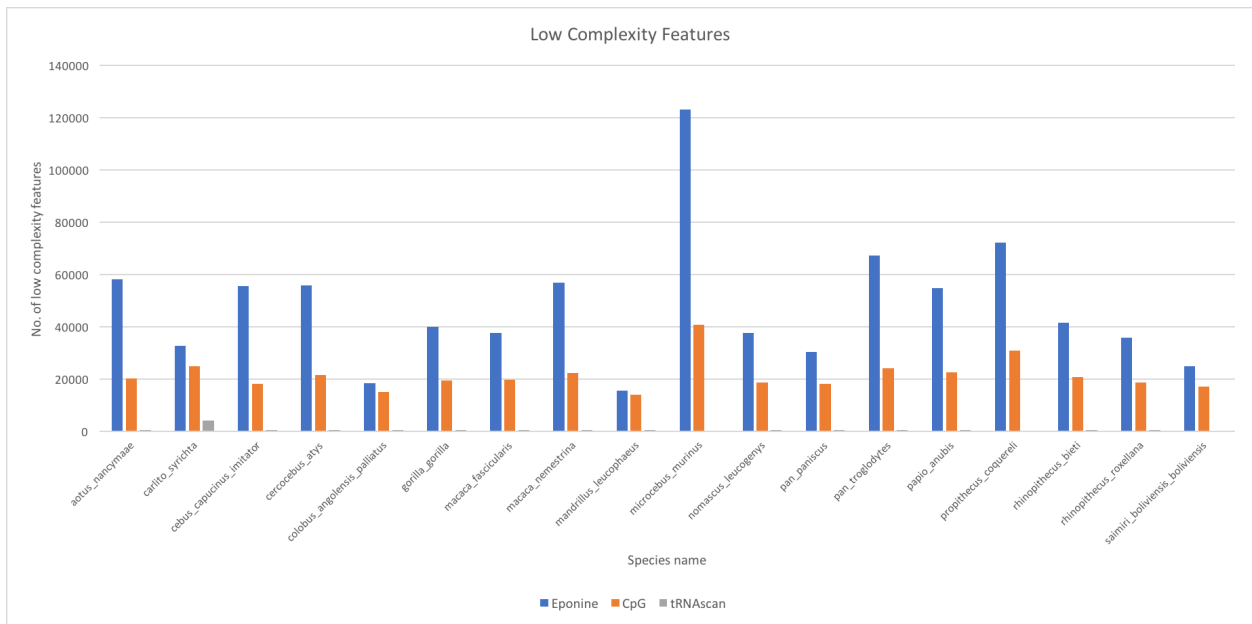


Fig 8: Counts of low complexity features in each species

## Layers in detail

### Layer 1

IG\_C\_gene, IG\_J\_gene, IG\_V\_gene, IG\_D\_gene, TR\_C\_gene, TR\_J\_gene,  
TR\_V\_gene, TR\_D\_gene

### Layer 2

realign\_95, realign\_80, rnaseq\_merged\_95, rnaseq\_merged\_80, self\_pe12\_sp\_95,  
self\_pe12\_tr\_95, self\_pe12\_sp\_80, self\_pe12\_tr\_80, human\_pe12\_sp\_95,  
human\_pe12\_tr\_95, primates\_pe12\_sp\_95, primates\_pe12\_tr\_95,  
mammals\_pe12\_sp\_95, mammals\_pe12\_tr\_95

### Layer 3

rnaseq\_tissue\_95, human\_pe12\_sp\_80, human\_pe12\_tr\_80, primates\_pe12\_sp\_80,  
primates\_pe12\_tr\_80, mammals\_pe12\_sp\_80, mammals\_pe12\_tr\_80

### Layer 4

rnaseq\_tissue\_80, realign\_50

### Layer 5

human\_pe12\_sp\_50, human\_pe12\_tr\_50

## More information

More information on the Ensembl automatic gene annotation process can be found at:

- Publication

Aken B et al.: The Ensembl gene annotation system. Database 2016.

- Web

[Link to Ensembl gene annotation documentation](#)

## References

1. Smit, A., R. Hubley, and P. Green, <http://www.repeatmasker.org/> RepeatMasker Open, 1996. **3**: p. 1996-2004.
2. Kuzio, J., R. Tatusov, and D. Lipman, *Dust*. Unpublished but briefly described in: Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology*, 2006. **13**(5): p. 1028-1040.
3. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences*. *Nucleic acids research*, 1999. **27**(2): p. 573.
4. Down, T.A. and T.J. Hubbard, *Computational detection and location of transcription start sites in mammalian genomic DNA*. *Genome research*, 2002. **12**(3): p. 458-461.
5. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence*. *Nucleic acids research*, 1997. **25**(5): p. 955-964.
6. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA*. *Journal of molecular biology*, 1997. **268**(1): p. 78-94.
7. Consortium, U., *UniProt: the universal protein knowledgebase*. *Nucleic acids research*, 2017. **45**(D1): p. D158-D169.
8. Pontius, J.U., L. Wagner, and G.D. Schuler, *21. UniGene: A unified view of the transcriptome*. *The NCBI Handbook*. Bethesda, MD: National Library of Medicine (US), NCBI, 2003.

9. Altschul, S.F., et al., *Basic local alignment search tool*. Journal of molecular biology, 1990. **215**(3): p. 403-410.
10. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation*. Nucleic acids research, 2015. **44**(D1): p. D733-D745.
11. Slater, G.S.C. and E. Birney, *Automated generation of heuristics for biological sequence comparison*. BMC bioinformatics, 2005. **6**(1): p. 31.
12. Harris, R.S., *Improved pairwise alignment of genomic DNA*. 2007: The Pennsylvania State University.
13. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project*. Genome research, 2012. **22**(9): p. 1760-1774.
14. She, R., et al., *genBlastG: using BLAST searches to build homologous gene models*. Bioinformatics, 2011. **27**(15): p. 2141-2143.
15. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-1760.
16. Lefranc, M.-P., et al., *IMGT®, the international ImMunoGeneTics information system® 25 years on*. Nucleic acids research, 2014. **43**(D1): p. D413-D422.
17. Griffiths-Jones, S., et al., *Rfam: an RNA family database*. Nucleic acids research, 2003. **31**(1): p. 439-441.
18. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature*. Nucleic acids research, 2006. **34**(suppl\_1): p. D140-D144.
19. Nawrocki, E.P. and S.R. Eddy, *Infernal 1.1: 100-fold faster RNA homology searches*. Bioinformatics, 2013. **29**(22): p. 2933-2935.